Dissertation zur Erlangung des Doktorgrades
der Fakultät für Chemie und Pharmazie
der Ludwig-Maximilians-Universität München

Transient transcriptome sequencing captures enhancer landscapes immediately after T-cell stimulation

Carina Elke Demel

aus

München, Deutschland

2017

Erklärung

Diese Dissertation wurde im Sinne von § 7 der Promotionsordnung vom 28.November 2011 von Herrn Prof. Dr. Patrick Cramer betreut.

Eidesstattliche Versicherung

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

München, den 30.03.2017

Carina Elke Demel

Dissertation eingereicht am: 03.04.2017

1. Gutachter: Prof. Dr. Patrick Cramer

2. Gutachter: PD Dr. Dietmar Martin

Mündliche Prüfung am: 29.05.2017

Acknowledgements

I am truly grateful to Prof. Dr. Patrick Cramer for giving me the opportunity to work in his lab, for his supervision and guidance on this fascinating project. I appreciate his positive attitude despite high workload and I am thankful for all the encouraging and supportive discussions. It has been a pleasure to work in this outstanding scientific environment.

I would like to express my deepest gratitude to Prof. Dr. Julien Gagneur, from whom I have learned a lot. I am deeply grateful for his help on mathematical and statistical methods that became a substantial part of this thesis. He always had great ideas and gave very valueable input on data analysis and the manuscript.

Special thanks goes to Prof. Dr. Achim Tresch, who introduced me to the field of systems biology and offered me a collaborative project with Patrick Cramer, thereby paving the way for my future.

I would like to thank the other members of my thesis committee: PD Dr. Dietmar Martin, Dr. Fabiana Perocchi, and Prof. Dr. Förstemann for their time and support.

Moreover, I want to especially thank Margaux Michel, who generated all the high-quality data that formed the basis of my work. Together we have learned a lot about sequening protocols, biases, and quality control. I am really happy I could work with such a great biochemist and wonderful person, and I am happy to see that all our efforts turned into a nice manuscript.

Likewise, I am deeply grateful to Björn Schwalb, who shared the office, ideas, and code with me for more than 8 years now. It has been a pleasure to work with him.

Furthermore, I am particularly grateful to Benedikt Zacher, Katja Frühauf, Michael Lidschreiber, Philipp Eser, Leonhard Wachutka, Kerstin Maier, Ania Sawicka, and Katharina Hofmann, for fruitful scientific discussions and collaborations.

I owe my gratitute to all present and former members of the Cramer, Gagneur, Tresch, and Söding labs, for the extraordinary working atmosphere that made me enjoy coming to work every single day. I am happy that I can count many of my collegues to my friends and want to thank Margaux, Björn, Merle, Michi, Katja, Saskia, Sofia, Wolfgang, Katharina, Ania, Jinmi, Livia, and Kristina for fun times

outside the lab and especially our weekend trips to Stockholm, Hannover, Krakow, Bremen, Barcelona, Berlin, Hamburg, Goslar, and Cologne. Special thanks goes to the Gagneur lab and the remaining Söding lab for hosting me every time I came to Munich. I always felt warmly welcomed and I am happy to count many of them to my friends. I also want to thank my graduate school QBM for the financial support, interesting workshops, and social activities. In this context, I would like to extend my thanks to all the people who helped me in bureaucratic issues and kept the labs running both at the Gene Center in Munich and at the MPI in Göttingen.

Many thanks to Verena Link, Marco Leitwein, Katharina Hofmann, Thomas Demel, and Björn Schwalb for critical reading (parts) of this thesis.

I am very grateful to Marco, who supports me in everything I do and makes my life better, easier, and funnier.

Last but not least, I want to thank my whole family. They were always interested in my work and its progress. Above all, I am grateful to my parents for their continuous support, unconditional love, and always believing in me.

Summary

Transcription regulation is poorly understood. Transcriptional enhancers produce enhancer RNAs (eRNAs), a class of transient RNAs, whose function remains mainly unclear.

To monitor transcriptional regulation in human cells, rapid changes in enhancer and promoter activity must be captured with high sensitivity and temporal resolution. Here I show that the recently established protocol TT-seq ('transient transcriptome sequencing') can monitor rapid changes in transcription from enhancers and promoters during the immediate response of T-cells to ionomycin and phorbol 12-myristate 13-acetate (PMA). Transient transcriptome sequencing (TT-seq) maps eRNAs and mRNAs every 5 minutes after T-cell stimulation with high sensitivity, and identifies many new primary response genes. TT-seq reveals that the synthesis of 1,601 eRNAs and 650 mRNAs changes significantly within only 15 minutes after stimulation, when standard RNA-seq does not detect differentially expressed genes. Transcription of enhancers that are primed for activation by nucleosome depletion can occur immediately and simultaneously with transcription of target gene promoters. My results indicate that enhancer transcription is a good proxy for enhancer regulatory activity in target gene activation, and establish TT-seq as a tool for monitoring the dynamics of enhancer landscapes and transcription programs during cellular responses and differentiation.

Additionally, I developed a normalization method for TT-seq that scales labeled and total RNA-seq samples relative to each other, allowing to determine absolute half-lives. The method provides a powerful tool to normalize various samples relative to each other on a global scale, and therefore allows to observe global changes in RNA synthesis and degradation.

Taken together, metabolical labeling of RNA followed by kinetic modeling enables to quantify RNA metabolism rates and to detect dynamic changes in enhancer landscapes and RNA expression levels.

Publications

Part of this work has been published or is currently in the process of being published.

2017 **rCube - RNA Rates in R: Applications note for bioconductor package**
C. Demel*, L. Wachutka*, P. Cramer, and J. Gagneur
(* These authors contributed equally to this work)
Manuscript in preparation

2017 **TT-seq captures enhancer landscapes immediately after T-cell stimulation**
M. Michel*, **C. Demel***, B. Zacher, B. Schwalb, S. Krebs, H. Blum, J. Gagneur, and P. Cramer (* These authors contributed equally to this work)
Molecular Systems Biology, 13(3), 920.

Author contributions: MM carried out experiments. CD performed bioinformatics analyses. BZ provided scripts and assistance with transcriptome annotation. BS provided scripts and advice on data analysis. SK performed sequencing, supervised by HB. JG supervised bioinformatics. PC designed and supervised research. MM, CD, JG, and PC prepared the manuscript.

2017 **Spt5 plays vital roles in the control of sense and antisense transcription elongation**
A. Shetty*, S.P. Kallgren*, **C. Demel**, K.C. Maier, D. Spatt, B.H. Alver, P. Cramer, P.J. Park, and F. Winston (* These authors contributed equally to this work)
Molecular Cell, 66, 1-12.

Author contributions: AS performed the ChIP-seq, NET-seq, and RNA-seq experiments. SPK, BHA, and PJP performed and interpreted the computational analysis for the ChIP-seq, NET-seq, and RNA-seq experiments. CD, KCM, and PC performed the 4tU-seq experiments and computational analysis, and DS performed the Northerns and RT-qPCR splicing analysis. FW and AS wrote the paper.

2016 **TT-seq maps the human transient transcriptome**
B. Schwalb*, M. Michel*, B. Zacher*, K. Frühauf, **C. Demel**, A. Tresch, J. Gagneur, and P. Cramer (* These authors contributed equally to this work)
Science, 352(6290), 1225-1228.

Author contributions: MM designed and carried out all experiments. BS carried out all bioinformatics analysis except transcript calling, RNA classification, analysis of U1 sequence motifs, and prediction of RNA secondary structure, which were carried out by BZ. BZ, JG and AT developed the chromatin state annotation. KF designed RNA spike-in probes. CD established the spike-in normalization method. BS, JG and PC designed research. JG and PC supervised research. BS, MM, JG, and PC prepared the manuscript, with input from all authors.

2016 **Determinants of RNA metabolism in the *Schizosaccharomyces pombe* genome**
P. Eser*, L. Wachutka*, K.C. Maier, **C. Demel**, M. Boroni, S. Iyer, P. Cramer, and J. Gagneur (* These authors contributed equally to this work)
Molecular Systems Biology, 12(2), 857.

Author contributions: PC and JG designed and supervised the research. PC conceived and designed the experiments. KCM performed the experiments. PE, LW, MB, JG, CD, and SI analyzed the data. PE, JG, PC, LW, and MB wrote the manuscript.

Table of Contents

Part I

Introduction

Parts of this section have been published in:

TT-seq captures enhancer landscapes immediately after T-cell stimulation

M. Michel*, **C. Demel***, B. Zacher, B. Schwalb, S. Krebs, H. Blum, J. Gagneur, and P. Cramer

Molecular Systems Biology (2017)

For detailed author contributions see page ix.

The central dogma of molecular biology postulated by Francis Crick (Crick 1970) explains the transfer of genetic information among the three classes of biopolymers: the deoxyribonucleic acid (DNA), the ribonucleic acid (RNA) and proteins. DNA encodes the genetic information that gets replicated during cell division (replication). DNA is transcribed into RNA (transcription), which serves as template for protein synthesis by ribosomes during translation.

During transcription the genetic information encoded in the DNA is transcribed into RNA by DNA-dependent RNA polymerases (Roeder et al. 1969). Eukaryotes have three nuclear RNA polymerases (Pols): Pol I, Pol II and Pol III (Cramer et al. 2008). The three RNA polymerases synthesize different classes of transcripts: Pol I produces most of the ribosomal RNAs (rRNAs), Pol II is responsible for the synthesis of messenger RNAs (mRNAs) and several classes of non-coding RNAs (ncRNAs), such as small nucleolar RNAs (snoRNAs), small nuclear RNAs (snRNAs), long non-coding RNAs (lncRNAs), and enhancer RNAs (eRNAs), and Pol III synthesizes transfer RNAs (tRNAs), 5s rRNA, and other small RNAs (Cramer et al. 2008).

Transcription is a highly studied process (Section 1). The complexity of eukaryotic transcriptomes is not only created by the genome size and the number of transcripts and transcript isoforms, but also by highly dynamic transcription regulation mechanisms. In eukaryotes, the spatio-temporal patterns of gene expression are established by enhancers (Section 2). Proper regulation of transcription is required for development, growth, cellular differentiation, and responses to environmental stimuli. One well-studied model system for the response to a stimulus is T-cell

activation (Section 3).

In recent years, high-resolution methods such as deep-sequencing of RNA (RNA-seq and its derivatives) or chromatin-immunoprecipitated samples (ChIP-seq) have been developed (Johnson et al. 2007; Wang et al. 2009b). The direct sequencing of transcription products or regions of active transcription marked by specific transcription factors led to improved transcriptome profiling, as these methods are not restricted to known transcript annotations (Mortazavi et al. 2008; Nagalakshmi et al. 2008) (Section 4). The development of these high-resolution profiling methods revealed that the majority of the genome is pervasively transcribed (Djebali et al. 2012; Jacquier 2009). This resulted in the identification of new transcript classes, such as eRNAs, which are short, usually unstable transcripts originating from enhancer sequences (Kim et al. 2010; Ren 2010; Wang et al. 2011). The function of eRNAs is not completely understood, but it is possible they influence transcription regulation (Section 2).

The combination of metabolic RNA labeling (Section 4.2) and RNA-sequencing creates new challenges for sample normalization (Section 5), but also presents new opportunities to model RNA synthesis and degradation rates (Section 6).

1 Transcription by RNA Polymerase II

In eukaryotes, RNA polymerase II (Pol II) transcribes protein-coding genes into mRNAs as well as several classes of non-coding RNAs (Cramer et al. 2008). The synthesis of mRNAs is completed in a cyclic process that can be divided into three major steps: initiation and promoter clearance (Section 1.1), elongation (Section 1.2), and termination (Section 1.3). Transcription is tightly coupled with mRNA processing to ensure the maturation of mRNA precursors (pre-mRNA). Correctly transcribed and processed RNAs are exported into the cytoplasm, where they get translated into proteins by ribosomes.

Figure 1 shows an overview of the transcription cycle of protein-coding genes, which is explained in detail below.

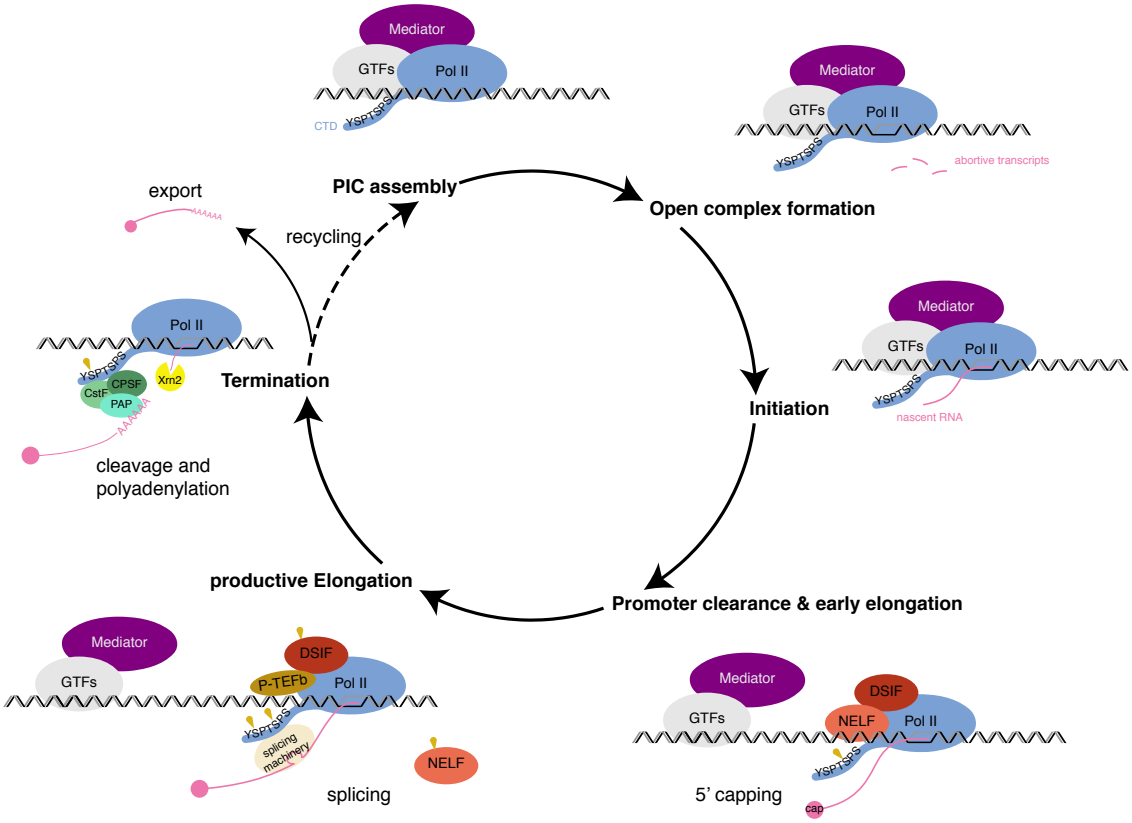


Figure 1: Schematic representation of the eukaryotic transcription process. Adapted from (Shandilya et al. 2012; Parker et al. 2007; Svejstrup 2004).

1.1 Transcription initiation and promoter clearance

The first crucial step of transcription initiation is the formation of the transcription-competent pre-initiation complex (PIC) on the promoter sequence. DNA is wrapped around nucleosomes, and therefore not accessible for the transcription machinery. Chromatin remodellers and histone modifying enzymes are required for transcription initiation, as they provide access for the transcription machinery to the template DNA (Li et al. 2007; Svejstrup 2004).

The PIC contains the general transcription factors (GTFs) TFIIA, -B, -D, -E, -F, and -H, and Pol II. PIC assembly starts by TFIID binding to defined promoter elements. At TATA-containing promoters, the TATA-binding protein (TBP) and TBP associated factors (TAFs) recognize the TATAA sequence (Buratowski et al. 1989; Buratowski 1994). The majority of eurkaryotic promoters lacks a canonical TATA box and other core promoter elements (CPEs), such as the initiator (Inr) and

downstream promoter elements (DPE), act as promoter recognition elements for the transcription machinery (Shandilya et al. 2012). TBP associates with co-activator complexes such as TFIID or SAGA (Basehoar et al. 2004). After TFIID binding, the other GTFs and Pol II are either recruited sequentially (Buratowski et al. 1989) or in the form of a pre-assembled holoenzyme to the promoter sequence to form the PIC (Koleske et al. 1994). The co-activator Mediator, which is recruited by transcriptional activators bound to upstream activating sequences (UASs), promotes PIC formation by facilitating the recruitment of Pol II to the emerging PIC (Kornberg 2005; Malik et al. 2005).

ATP-dependent promoter melting around the gene's transcription start site (TSS) allows the open complex formation by the helicase activity of TFIIH (Hahn 2004; Wang et al. 1992). This step, also called PIC activation, provides access for Pol II to the template strand via the transcription 'bubble' and allows the formation of the first phosphodiester bond (Roeder 1996; Wang et al. 1992). The beginning synthesis of RNA often results in the release of short (2-3 nt) abortive transcripts while Pol II is still associated with the promoter (Margeat et al. 2006). The growing nascent RNA increases the stability of the transcription complex and results in promoter clearance (Kugel et al. 2002). When the nascent transcript reaches a length of 8-9 nt, the RNA:DNA hybrid is thermodynamically stable, elongation continues and the likelihood of premature RNA release is reduced (Roeder 1996; Sidorenkov et al. 1998). TFIID, TFIIA, TFIIB, and the Mediator complex dissociate from the transcription machinery and stay attached to the promoter, allowing for a rapid reinitiation of Pol II (Section 1.3).

1.2 Transcription elongation

The transition from initiation to early elongation is accompanied by a conformational change of Pol II (Proudfoot et al. 2002). The Pol II C-terminal domain (CTD) consists of multiple YSPTSPS heptapeptide repeats (26 in yeast, 52 in mammals) (Corden et al. 1985), which serve as binding platform for RNA maturation factors that process RNA co-transcriptionally (Hirose et al. 2000; McCracken et al. 1997b). Except for proline (P), all amino acids of the CTD can be phosphorylated

and the CTD phosphoylation status changes dynamically during the transcription cycle ('CTD code') (Buratowski 2003; Hahn 2004; Nechaev et al. 2011; Proudfoot et al. 2002) and stimulates the pre-mRNA processing steps (Fong et al. 2001). During promoter binding, the CTD is mostly unphosphorylated and attracts factors such as Mediator that stabilize the PIC (Myers et al. 1998). Phosphorylation of the CTD at Serine-5 residues by the TFIIH subunit CDK7 destabilizes these interactions and favors promoter escape and therefore promotes the transition from initiation into early elongation (Liu et al. 2004).

Serine-5 phosphorylation (S5P) is also recognized by capping enzymes (Cho et al. 1997; McCracken et al. 1997a). When the nascent RNA reaches a length of 20-30 nt, a 7-methyl-guanosine cap is added to its 5' end (Rasmussen et al. 1993). Capping is important for RNA stability, nuclear export, and enhances translation (Proudfoot et al. 2002).

Transcription can be regulated during early elongation by promoter-proximal pausing (PPP) of Pol II (Core et al. 2008). The DRB-sensitivity inducing factor (DSIF) and the negative elongation factor (NELF) act as negative elongation factors on Pol II (Renner et al. 2001). The positive elongation factor P-TEFb phosphorylates DSIF and NELF, phosphorylated NELF is released, and DSIF acts as positive elongation factor (EF) on Pol II and transcription enters productive elongation (Marshall et al. 1995). The cyclin-dependent kinase CDK9, a subunit of P-TEFb, is responsible for increasing phosphorylatation of CTD Serine-2 (S2P) towards the 3' end of the gene (Marshall et al. 1995; Peterlin et al. 2006), while Serine-5 is gradually dephosphorylated by Ssu72 (Reyes-Reyes et al. 2007). The elevated levels of Serine-2 phosphorylation increase the affinity for the recruitment of the splicing complex, which performs co-transcriptional splicing on the nascent pre-mRNA (Ardehali et al. 2009).

1.3 Transcription termination and reinitiation

Hyper-phosphorylation of CTD at Serine-2 residues plays a key role in termination, as it recruits 3' end processing factors, such as the cleavage and polyadenylation specificity factor (CPSF) and the cleavage stimulatory factor (CstF), to the poly-

merase (Ahn et al. 2004; McCracken et al. 1997b). For eukaryotic protein-coding genes, 3' end processing and transcription termination are tightly coupled (Whitelaw et al. 1986). Once the polymerase transcribes the polyadenylation (pA) signal, a highly conserved AATAAA sequence followed by a G/T-rich downstream sequence element (DSE) (Colgan et al. 1997; Proudfoot et al. 2002; Zhao et al. 1999), CPSF and CstF recognize these sequence elements in the emerging transcript, promote pausing of Pol II, and induce cleavage of the nascent transcript directly after the pA site so that it gets released from Pol II (Gilmartin et al. 1989). Poly(A) polymerase (PAP) is recruited to the termination machinery and adds the poly(A) tail to the newly generated 3' end (Colgan et al. 1997; Moore et al. 1985). Polyadenylation is required for export to the cytoplasm (Huang et al. 1996), where translation takes place.

Transcription termination downstream the pA site can be explained by two different models. In the 'anti-termination' model, Pol II stays associated with the DNA until elongation factors (antiterminator factors) dissociate from the complex at the pA site (Logan et al. 1987; Proudfoot 2004). In the 'torpedo' model Pol II stays associated with the DNA and continues transcribing downstream the pA site (Connelly et al. 1988; Proudfoot 1989). The newly formed 5' end of the polymerase-associated RNA is uncapped and unprotected and is therefore attacked by the RNA 5' to 3' exonuclease Xrn2 for nucleolytic degradation. When Xrn2 catches up with Pol II, transcription terminates and Pol II is displaced from the DNA (Kim et al. 2004; Teixeira et al. 2004; West et al. 2004).

The released hypo-phosphorylated Pol II can enter a new round of transcription. Gene looping (Ansari et al. 2005) and the promoter-bound GTFs that form a 'reinitiation scaffold' (Yudkovsky et al. 2000) facilitate efficient recycling and reinitiation of Pol II on the same template (Dieci et al. 1996).

2 Regulation of transcription by enhancers

In metazoan cells, the synthesis of mRNAs from protein-coding genes during transcription is driven from promoters and activated by enhancers (Lenhard et al. 2012; Levine et al. 2014). Enhancers are regulatory units in the genome that contain

binding sites for sequence-specific transcription factors and can activate mRNA transcription over long distances (Banerji et al. 1981) (Section 2.1). Active enhancers adopt an open chromatin structure (Calo et al. 2013) and recruit co-activators such as Mediator (Fan et al. 2006). Mediator can apparently bridge between enhancers and promoters because it binds both transcriptional activators and the Pol II initiation complex at the promoter (Figure 2) (Liu et al. 2013; Malik et al. 2010). Promoter-enhancer interaction ('pairing') increases initiation complex stability and promotes Pol II escape from the promoter (Allen et al. 2015; DeMare et al. 2013; Splinter et al. 2006). Promoter-enhancer pairing requires DNA looping that is facilitated within insulated neighborhoods, which are genomic regions formed by looping of DNA between two CTCF-binding sites co-occupied by cohesin (Dowen et al. 2014; Hnisz et al. 2016a; Phillips-Cremins et al. 2013) (Section 2.2).

The genome-wide identification of enhancers is crucial for studying cellular regulation and differentiation, but remains technically challenging (Shlyueva et al. 2014). Enhancers may be distinguished from other genomic regions through a signature of histone modifications that can be mapped by chromatin immunoprecipitation (ChIP) (Heintzman et al. 2007; Schübeler 2007; Visel et al. 2009) or DNA accessibility assays (Shlyueva et al. 2014; Thurman et al. 2012; Xi et al. 2007). Regulatory active enhancers may be identified through their transcriptional activity, which is thought to be a good proxy for their function in promoter activation (Li et al. 2016; Melgar et al. 2011; Wu et al. 2014). Transcribed enhancers produce eRNAs (Djebali et al. 2012; Kim et al. 2010), which are difficult to detect because they are short-lived (Rabani et al. 2014; Schwalb et al. 2016), rapidly degraded by the exosome (Lubas et al. 2015), and generally not conserved over species (Andersson et al. 2014) (Section 2.3).

2.1 Enhancer characteristics and identification

Enhancers are *cis*-regulatory elements in the genome that were first discovered in the 1980s (Banerji et al. 1981; Maniatis et al. 1987; Orkin 1990). They can activate or alleviate the gene expression of nearby promoters (Banerji et al. 1981). This is achieved by binding of cell type-specific transcription factors to specific binding

sites in the enhancer sequence and delivering them to the transcription machinery at the promoter via DNA looping (Section 2.2). The function of enhancers is highly cell type and cell context-dependent. Different enhancer elements may regulate the same promoter under different conditions (Chan et al. 2010). Enhancers can act over long-range distances up to several megabases and independently of their orientation relative to their target genes (Banerji et al. 1981; Maniatis et al. 1987).

The transcription factor-binding nature of enhancers can be used to identify them via ChIP-seq. Especially the transcriptional coactivators CBP and p300 that interact with the transcription machinery and acetylate histone tails at enhancers to generate an open chromatin structure, have commonly been used to identify enhancers (Visel et al. 2009). ChIP-seq of characteristic histone modifications, such as acetylation of H3 lysine 27 (H3K27ac) by CBP (Tie et al. 2009) or distinct methylation patterns of H3K4, can also be used to identify enhancer elements. Enhancer sequences show high levels of H3K4 mono- and di-methylation (H3K4me1/2) and at the same time low levels of H3K4me3. As promoter sequences show the opposite pattern, the ratio H3K4me3/H3K4me1 is often used to distinguish enhancers from promoters (Heintzman et al. 2007). Additionally, the open chromatin structure at enhancers that allows binding of transcription factors (TFs) is related to DNase I hypersensitivity (Schaffner 2015), which can be identified by DNase-seq or ATAC-seq.

2.2 Chromatin looping and promoter-enhancer interactions

Chromatin is usually organized in compartments that are in close spatial proximity, called topologically associated domains (TADs) (Figure 2), which are separated from each other by the insulator protein CCCTC-binding factor (CTCF) and cohesin binding at TAD boundaries (Dixon et al. 2012; Gonzalez-Sandoval et al. 2016). The frequency of long-range DNA interactions is higher inside TADs than between TADs, therefore TADs are constraining enhancer-promoter interactions (Dixon et al. 2015).

Several methods have been developed to study chromatin interactions: Chromosome conformation capture (3C) (Dekker et al. 2002) and its derivatives 4C (circular 3C) (Simonis et al. 2006) and 5C (3C-carbon-copy) (Dostie et al. 2006) can detect

long-range interactions of specific loci or within confined genomic regions (~1 Mb) by ligation of cross-linked DNA fragment ends. Hi-C (Lieberman-Aiden et al. 2009) provides genome-wide long-range interaction at higher resolution (up to 1 kb). Chromatin interaction analysis using paired-end tag sequencing (ChIA-PET) utilizes the presence of cohesin at the bondaries of chromatin loops, as it identifies chromatin interaction sites that are bound by a specific factor (e.g. cohesin) via ChIP (Fullwood et al. 2009; Hnisz et al. 2016b).

These methods provided chromatin interaction maps for various species and cell types that show that TADs are highly conserved between cell types and across species and usually span about 1 MB (Dixon et al. 2012; Li et al. 2016; Schmitt et al. 2016). Cohesin and CTCF also form loop structures inside TADs, called insulated neighborhoods, that span ~190 kb and organize cell identity genes in clusters with their enhancers (Dowen et al. 2014; Hnisz et al. 2016a). The disruption of insulated neighborhoods has been linked to cancer, as the correct formation of insulating sites prevents activation of proto-oncogenes by enhancers (Hnisz et al. 2016b).

Chromatin interaction maps support the model in which enhancers and promoters are brought close to each other by looping (Figure 2).

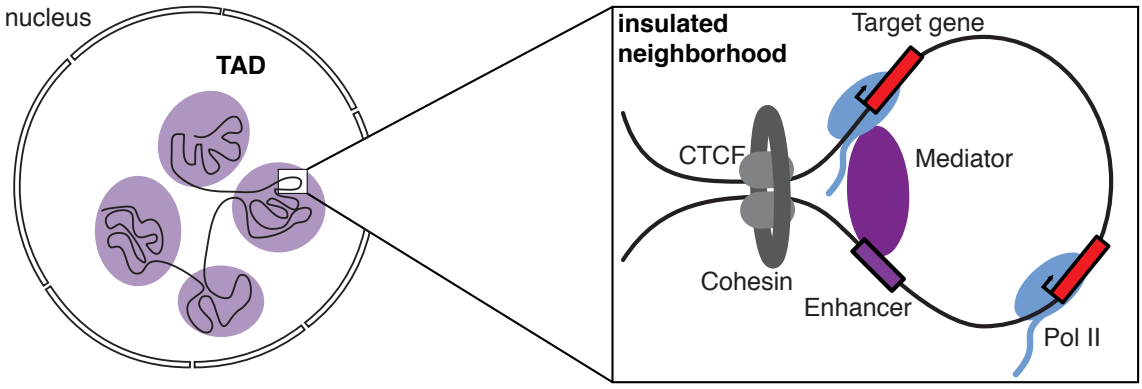


Figure 2: Chromatin in the nucleus is organized in TADs, which are composed of insulated neighborhoods. Chromatin loops, which are established by cohesin and Mediator, bring enhancers and target promoters close to each other. Adapted from (Hnisz et al. 2016a; Shlyueva et al. 2014; Gonzalez-Sandoval et al. 2016).

2.3 Enhancer transcription

The role of enhancer transcription and/or eRNAs remains unclear (Li et al. 2016). It is likely that the process of enhancer transcription has a functional role, maybe in re-

cruiting chromatin remodelers through their association with transcribing Pol II (Gribnau et al. 2000). Consistent with this model, enhancer transcription can precede target gene transcription (Arner et al. 2015; De Santa et al. 2010; Kaikkonen et al. 2013; Schaukowitch et al. 2014). It is also possible that eRNAs themselves have a function, because eRNA knockdown may impair target gene activation (Ilott et al. 2014; Li et al. 2013; Schaukowitch et al. 2014). eRNA knockdown may also have negative effects on promoter-enhancer pairing (Li et al. 2013), although some studies came to different conclusions (Hah et al. 2013; Schaukowitch et al. 2014).

Several studies have shown that the transcription of eRNAs can be activated or vary in a circadian manner (Fang et al. 2014; Kaikkonen et al. 2013; Step et al. 2014). Additionally, some eRNAs have a function in transcription activation: Via interaction with the Mediator complex and the androgen receptor complex at the enhancer, KLK3 eRNA facilitates chromosomal looping and enhances gene expression of *KLK3* and *KLK2* (Hsieh et al. 2014). In a single-gene study, it has been found out that the eRNA is transcribed, then it interacts with the NELF complex, which is subsequently released from Pol II at the target promoter *Arc*, and transcription of *Arc* mRNA takes place (Schaukowitch et al. 2014). These results would suggest that eRNAs need to be transcribed before their target genes in order to increase their gene expression levels. Dynamic changes in eRNA and mRNA transcription have been assessed via CAGE (cap analysis of gene expression) in 33 time courses of different biological stimuli and across different cell types and organisms. This study shows that in general eRNA expression can peak as early as 15 minutes after stimulation and is followed by an induction of mRNA expression levels (Arner et al. 2015).

3 T-cell activation

T-cell activation is a widely studied model system to investigate the cellular response to exogenous stimulation. Upon T-cell stimulation, the T-cell receptor (TCR) and the costimulatory receptor CD28 are activated, leading to a signaling cascade (Smith-Garvin et al. 2009). Phosphorylation of multiple factors at the plasma membrane leads to recruitment and activation of PLC-γ that cleaves

$PI(4,5)P_2$ in DAG and IP_3. First, DAG binds and activates PKCθ, which leads to activation and nuclear translocation of the transcription factors NF-κB and AP-1. Second, IP_3 diffuses away from the plasma membrane and activates calcium channel receptors on the endoplasmic reticulum (ER), increasing intracellular calcium ion concentration and leading to activation of calmodulin and calcineurin. Calcineurin activates NFAT that translocates to the nucleus and drives gene activation. T-cell stimulation via the T-cell receptor and CD28 can be mimicked by addition of PMA and ionomycin because phorbol esters activate PKC and calcium ionophores raise intracellular calcium levels (Weiss et al. 1987).

The T-cell response involves rapid changes in gene expression (Cheadle et al. 2005; Diehn et al. 2002; Feske et al. 2001; Marrack et al. 2000; Raghavan et al. 2002; Rogge et al. 2000). Responding genes were classified into immediate-early, early, and late response genes based on changes in RNA levels. Immediate-early response genes are transiently activated within the first hour after stimulation (Bahrami et al. 2016). There are ~40 immediate-early genes described, most of which code for transcription factors such as *FOS*, *FOSB*, *FRA1*, *JUNB*, *JUN*, *NFAT*, *NF-κB* and *EGR1* (Greenberg et al. 1984; Sheng et al. 1990). Several hours after stimulation, immediate-early factors activate early and late response genes, including cytokines, such as *IL-2, TGF-β* or *IFN-γ* (Crabtree 1989; Ellisen et al. 2001). Despite these studies, the immediate T-cell response and the primary events after T-cell stimulation remain incompletely understood.

4 Transcriptome profiling

4.1 Quantification of RNA abundance

The abundance of RNA in a cell can be measured on a large scale by microarrays or RNA sequencing (RNA-seq). These methods allow the genome-wide, high-resolution quantification of gene expression.

Microarrays contain thousands of oligonucleotide DNA probes on their surface that are complementary to specific genes of interest or genomic target loci. Microarrays allow the simultaneous measurement of gene expression for thousands

of transcripts, for which probes are present on the microarray. After converting the RNA probes to complementary DNA (cDNA), hybridization of flourescently labeled samples generates a optical signal, which is scanned and quantified (Shalon et al. 1996). One limitation of microarrays is the restriction to known transcript sequences. Next-generation sequencing techniques overcome this limitation by direct high-throughput sequencing of cDNA (RNA-seq). cDNA fragments with adaptors ligated to their ends can be sequenced from one (single-end sequencing) or both ends (paired-end sequencing). This generates million of reads of 30-400 bp in length (depending on the sequencing platform) at low cost in a very short time (Mortazavi et al. 2008; Wang et al. 2009b). RNA-seq is compatible with every species, but it requires that the reference genome is known, as the sequenced reads need to be mapped to the reference genome to determine their origin. This allows to quantify the gene expression by counting reads at each locus of interest in the genome. One advantage of RNA-seq is its power to detect low-abundance and novel transcripts. The single-base resolution of RNA-seq also provides the potential to detect novel 5' and 3' transcript boundaries and alternative splicing isoforms (Mortazavi et al. 2008; Wang et al. 2009b).

In recent years, many variations of RNA-seq have been established that focus on specific transcript properties. Some sequencing-based methods detect the usage of specific transcription start sites by sequencing 5' transcript ends (GRO-cap (Core et al. 2014)) or alternative poly(A)-sites. Several methods assess nascent RNA by nuclear run-on of isolated nuclei (GRO-seq (Core et al. 2008), PRO-seq (Kwak et al. 2013)), by immunoprecipitation of Pol II and sequencing of Pol II-associated transcripts (NET-seq (Churchman et al. 2011)), or by metabolic labeling (Miller et al. 2011; Paulsen et al. 2014) (Section 4.2).

4.2 Metabolic labeling to measure RNA synthesis

The levels of RNA in an eukaryotic cell are the result of regulated synthesis and degradation of RNA. Standard transcriptomics (e.g. RNA-seq) measure the total RNA abundance. Due to the long half-lives of mRNA, and thus high levels of mRNA in the cell, RNA-seq is not sensitive enough to observe fluctuations in transcript

synthesis or degradation. Additionally, the large amount of stable mRNA in the cell conceils ncRNAs that usually have a faster turnover (Schwalb et al. 2016).

Methods which measure the nascent RNA in a cell are sensitive to changes in the RNA synthesis and are not biased by stable transcripts. Several methods have been used to study nascent transcription that include arrest of Pol II *in vivo* and run-on *in vitro* (Core et al. 2008; Kwak et al. 2013; García-Martínez et al. 2004). Transcription can be arrested by sarkosyl, but this treatment inhibits cellular processes (Miller et al. 2011). Therefore, to measure RNA synthesis in a non-perturbing manner *in vivo*, nascent RNA can be marked by metabolic labeling (Cleary et al. 2005; Dölken et al. 2008; Friedel et al. 2009; Kenzelmann et al. 2007; Miller et al. 2009) and subsequently purified labeled transcripts can be quantified.

The additional advantage of unperturbing metabolic labeling is the inference of transcript degradation rates. Degradation rates can be directly measured after blocking transcription or heat shock perturbation (Grigull et al. 2004; Holstege et al. 1998; Lam et al. 2001; Wang et al. 2002), but both methods are perturbing the cellular system. Metabolic labeling followed by kinetic modeling leads to unperturbed RNA synthesis and degradation rates (Miller et al. 2011).

4.2.1 Dynamic Transcriptome Analysis (DTA)

Nascent RNA can be labeled by providing a labeling substrate such as 4-thiouracil (4tU) or 4-thiouridine (4sU). During transcription, 4sUTP is incorporated into the newly transcribed RNA instead of UTP. The thiol-labeled RNA ('labeled RNA') can be isolated from the total pool of RNA ('total RNA') by biotinylation and streptavidin-coated magnetic beads (Miller et al. 2011).

The nucleoside analog 4sU can be taken up by eukaryotic cells or by yeast cells expressing the nucleoside transporter human equilibriative nucleoside transporter (hENT1). Alternatively, 4tU can be used to label the budding yeast *Saccharomyces cerevisiae (S. cerevisiae)* or the fission yeast *Schizosaccharomyces pombe (S. pombe)* without the expression of an additional transporter. After the cellular uptake of 4tU, it gets efficiently converted to thiolated UTP (Eser et al. 2016).

4sU/4tU-labeled and total RNA fractions can then be quantified by microarray measurements (Eser et al. 2013; Miller et al. 2011; Sun et al. 2012) or deep sequencing

(4sU-seq/4tU-seq) (Eser et al. 2016; Schulz et al. 2013).

Comparative DTA (cDTA) is an extension of DTA, where *S. pombe* RNA is used as internal standard for normalization of different *S. cerevisiae* samples. Labeled *S. pombe* and *S. cerevisiae* cells are mixed in a defined ratio before cells lysis, total RNA purification, labeled RNA extraction, and hybridization on microarrays. cDTA allows to compare absolute changes between different samples (Sun et al. 2012).

4sU-seq/4tU-seq is more sensitive than RNA-seq in monitoring dynamic changes in RNA levels and allows to estimate synthesis and degradation rates by kinetic modeling (Sections 6 and 8.4).

4.2.2 Transient transcriptome sequencing

TT-seq (Schwalb et al. 2016) is based on 4sU-seq and aims at a uniform read distribution along long human transcripts, which is not given by 4sU-seq. Human protein-coding genes are on average 67 kb long. With a estimated elongation rate of ~4 kb/min (Ardehali et al. 2009; Darzacq et al. 2007; Singh et al. 2009) polymerase transcribes about 20 kb during the 5 minute labeling pulse. Therefore, during the short labeling pulse of 5 minutes, only a small 3' part of nascent RNA is labeled, while the 5' regions were already pre-existing before the addition of 4sU (Figure 3). Purification of labeled RNA fragments would result in the overrepresentation of 5' ends (5' bias) that were not synthesized during the labeling pulse. To overcome this bias, TT-seq was established. The RNA is fragmented via mild sonication, yielding fragments of about 1.5 kb, before labeled RNA is purified (Figure 3). The subsequent isolation of labeled fragments yields a transcript length-independent distribution of nascent RNAs (Schwalb et al. 2016).

Therefore, TT-seq provides a framework to estimate synthesis and degradation rates, observe transient RNAs, and observe fast changes in transcription kinetics in the human system.

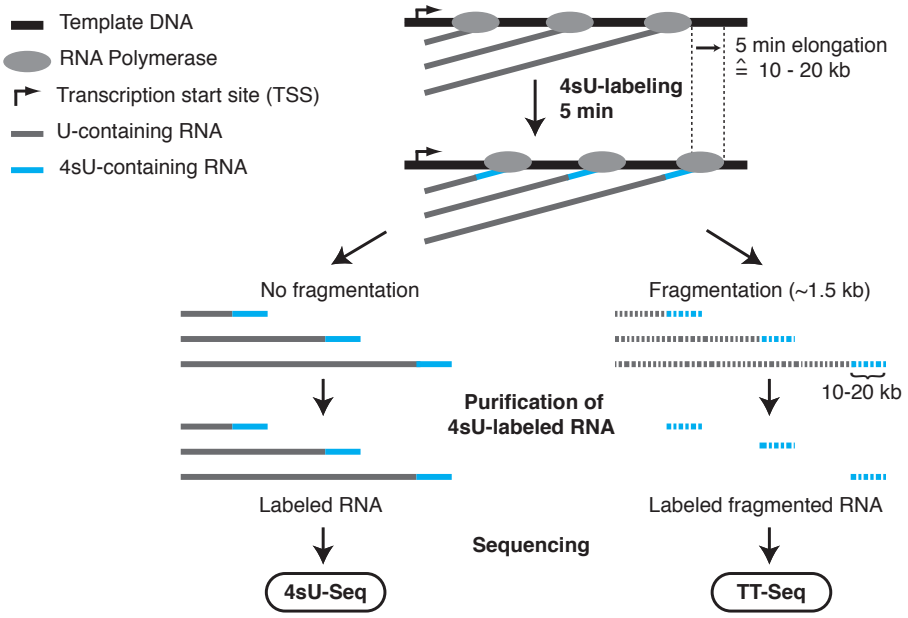


Figure 3: Schematic view of the 4sU-seq and TT-seq protocols. Taken from (Schwalb et al. 2016). Reprinted with permission from AAAS.

5 Conventional methods for modeling and normalization of RNA-seq data

For RNA-seq and its derivatives, the abundance of biological RNA fragments is not directly reflected in the observed read counts. The distribution of reads within a single sequenced sample allows to draw conculsions on relative ratios of gene expression levels between different genes, as the read counts are linearily related to the abundance of a transcript (Mortazavi et al. 2008). The absolute difference between different samples, however, cannot be directly estimated from sequenced read numbers. The number of sequenced reads/fragments (read counts) per transcript depends on the sequencing depth, i.e. how many reads have been sequenced for the sample, the length of the transcript, and its abundance in the cell. Hence, the quantitative assessment of differential gene expression across different samples requires normalization between samples that allows for comparisons between samples (Dillies et al. 2013).

So far, many different models have been proposed to infer levels of gene-expression and compare them quantitatively between different samples. A common normalization in quantitative transcriptomics is the calculation of reads per kilobase per million mapped reads (RPKM) for each transcript (Mortazavi et al. 2008). This

value accounts for different sequencing depths per sample and varying gene lengths, but is biased by highly-expressed transcripts (Bullard et al. 2010).

There are multiple tools that use statistical testing to infer differentially expressed genes while accounting for different sequencing depths between samples. In order to test if observed read counts between samples are significant different, these methods have to take into account the underlying distribution. Some methods assume a Poisson distribution, as it approximates a multinomial distribution which would reflect read counts that are independently sampled (Marioni et al. 2008; Wang et al. 2009a). The problem with the Poisson distribution is that its variance is equal to the mean. Overdispersion is a common feature in count data, where the variance is high for low count numbers. Therefore, the Poisson distribution is not suited to model count data. The negative binomial distribution (NB), which is determined by a mean μ and variance σ^2, has been proposed to model count data (Whitaker 1914). The *edgeR* package uses the NB read distribution to model read counts k_{ij} for a gene i in a sample j: $k_{ij} \sim NB(\mu_{ij}, \phi_j)$ with the dispersion parameter ϕ_j, while assuming that mean and variance are related: $\sigma^2_{ij} = \mu_{ij} + \phi_j \mu^2_{ij}$ (Robinson et al. 2007; Robinson et al. 2009). *edgeR* estimates the trimmed mean of M values (TMM), a weighed trimmed mean of log expression ratios in two different conditions. The TMM is inferred from gene-wide expression ratios under the assumption that the majority of genes is not differentially expressed (Robinson et al. 2010). The TMM is used to scale the total read counts per sample, obtaining the *effective library size*, and therefore normalizing for variations in sequencing depth. Another commonly used approach is *DESeq,* which addresses the library size normalization by linear scaling of a condition-dependent gene-specific value q_{ij} with the normalization factor s_j to obtain the expected count value for gene i in sample j, μ_{ij}: $\mu_{ij} = q_{ij} s_j$ (Anders et al. 2010). The variance is estimated as the sum of the shot noise (uncertainty in measuring a concentration by counting reads) and the sample-to-sample variation: $\sigma^2 = s_j \mu_{ij} + \phi s^2_j \mu^2_{ij}$ (Anders et al. 2012). *DESeq2* also includes shrinkage of fold changes for genes with low read counts, where the variance is usually higher compared to highly expressed genes. Variability between replicates is handled with a dispersion parameter (Love et al. 2014a; Love et al. 2014b).

All of the explained statistical methods are based on the hypothesis that the

majority of genes is not differentially expressed. However, this is not always true in experiments, as for example the knock-down or knock-out of a specific factor could lead to a global down-regulation of cellular transcription. This issue can only be assessed by the usage of internal standards in the experiment that are not influenced by environmental stimuli or genetic perturbations. In the cDTA method, RNA from a different organism is added to the samples of interest (Sun et al. 2012) (Section 4.2). It is also possible to use synthetic spike-ins to infer global normalization factors (Jiang et al. 2011; Lovén et al. 2012; Schwalb et al. 2016) (Section 8.1).

In the case of metabolic labeling with subsequent kinetic modeling the relative ratio of newly-synthesized labeled RNA and total cellular RNA has to be estimated in order to infer correct synthesis and degradation rates on an absolute scale. This normalization can be achieved by scaling each sample with the ratio of purified RNA quantities before and after labeled RNA purification (Rabani et al. 2011). In the *DRiLL* model two normalization factors are estimated that account for the relative abundance of 4sU-labeled RNA within the total RNA population and for a possible cross-contamination of unlabeled RNA in the labeled RNA fraction (Rabani et al. 2014). The *INSPEcT* framework jointly estimates normalization factors and calculates synthesis, degradation, and processing rates from 4sU-seq and total RNA-seq samples (de Pretis et al. 2015). In the case of multiple labeling time points, the increase of the labeling fraction over time and hence its convergence to steady-state levels can be used to estimate the normalization factor and the cross-contamination rate (Eser et al. 2016). All of these approaches estimate these normalization factors jointly across all genes. However, they are limited in the detection of global expression differences between samples.

6 Estimation of RNA metabolism kinetics

Standard transcriptomics combined with metabolic labeling can be used to model transcription kinetics. A set of differential equations describes the RNA synthesis, processing, and degradation processes. As it is reasonable to assume cytoplasmic RNA levels decay proportionally to their level, degradation is usually modeled by

first-order kinetics. Assuming that the cytoplasmic RNA amount T is degraded proportionally to its overall concentration with rate λ during time t, the following differential equation can be derived:

$$\frac{dT}{dt} = -\lambda T \tag{1}$$

The solution of this differential equation gives an estimate about the RNA concentration at time t dependent on the initial RNA concentration $T(0)$ assuming exponential decay:

$$T(t) = T(0) \cdot e^{-\lambda \cdot t} \tag{2}$$

The time $t_{1/2}$, after which half of the initial RNA amount $T(0)$ has been degraded, can be derived by setting $T(t) = \frac{1}{2}T(0)$. Then,

$$t_{1/2} = \frac{log(2)}{\lambda}. \tag{3}$$

During steady-state, mRNA synthesis and degradation are in equilibrium.

The *DTA* method has been designed to extract synthesis and degradation rates from 4sU-labeled and total RNA microarray measurements (Schwalb et al. 2012). The change of steady-state labeled and total RNA levels L and T during labeling time t can be described by

$$\frac{dL}{dt} = \mu - \lambda L \tag{4}$$

and

$$\frac{dT}{dt} = \mu - \lambda T \tag{5}$$

with constant synthesis rate μ and degradation rate λ during labeling time t.

DRiLL extends the model above by estimating also processing rates (Rabani et al. 2011; Rabani et al. 2014). This model implies that pre-mRNA P is synthesized and processed to mature RNA M in the nucleus before it gets exported to the cytoplasm, where degradation takes place. This can be modeled by relying on the assumption that pre-mRNA in the nucleus is not degraded and that the export to the cytoplasm happens immediately after maturation. Additionally, when the labeling time is short (<10 min), it can be assumed that 4sU-labeled RNA is mostly nuclear (Rabani et al.

2011). Then, the labeled RNA fraction corresponds to pre-mRNA levels:

$$L = P \tag{6}$$

and

$$T = P + M. \tag{7}$$

Pre-mRNA levels are reduced by RNA processing at rate ψ:

$$\frac{dP}{dt} = \mu - \psi P. \tag{8}$$

Mature RNA levels evolve from processed pre-mRNA levels and are degraded at rate λ:

$$\frac{dM}{dt} = \psi P - \lambda M. \tag{9}$$

The *INSPEcT* software additionally accounts for the fact that 4sU-labeled RNA can also contain mature RNA, so Equation (6) does not hold anymore. This is done by calculating exonic and intron RPKM values in 4sU-labeled and total RNA-seq samples that correspond to pre-mRNA and mRNA levels among 4sU-labeled and total RNA, respectively (de Pretis et al. 2015).

Eser et al. fit the same model as described in Equations (8) and (9) to a time series of multiple labeling time points, while accounting for increasing labeled RNA fractions with longer labeling time, sequencing depth, 4tU labeling efficiency, and cross-contamination of unlabeled RNAs in the labeled RNA fraction (Eser et al. 2016) (Section 19).

The development of TT-seq and its ability to measure nascent RNA unbiased by the length of the gene (Section 4.2.2) offers the possiblity to determine local synthesis and degradation rates at nucleotide resolution (Wachutka et al. 2016). The reads mapping to exonic or intronic bases, or to exon-intron, intron-exon, and exon-exon junctions are quantified in TT-seq and RNA-seq fractions and synthesis and degradation rates can be estimated. The synthesis rate of individual phosphodiester bonds corresponds to the transcription rate of exons, introns, or single junctions.

The degradation rate of phosphodiester bonds within exons or introns or bonds spanning exon-exon junctions correspond to the degradation rate of the exon, intron or junction, respectively. The degradation rate of bonds at exon-intron or intron-exon junctions gives the splicing rate at these junctions (Figure 4) (Wachutka et al. 2016).

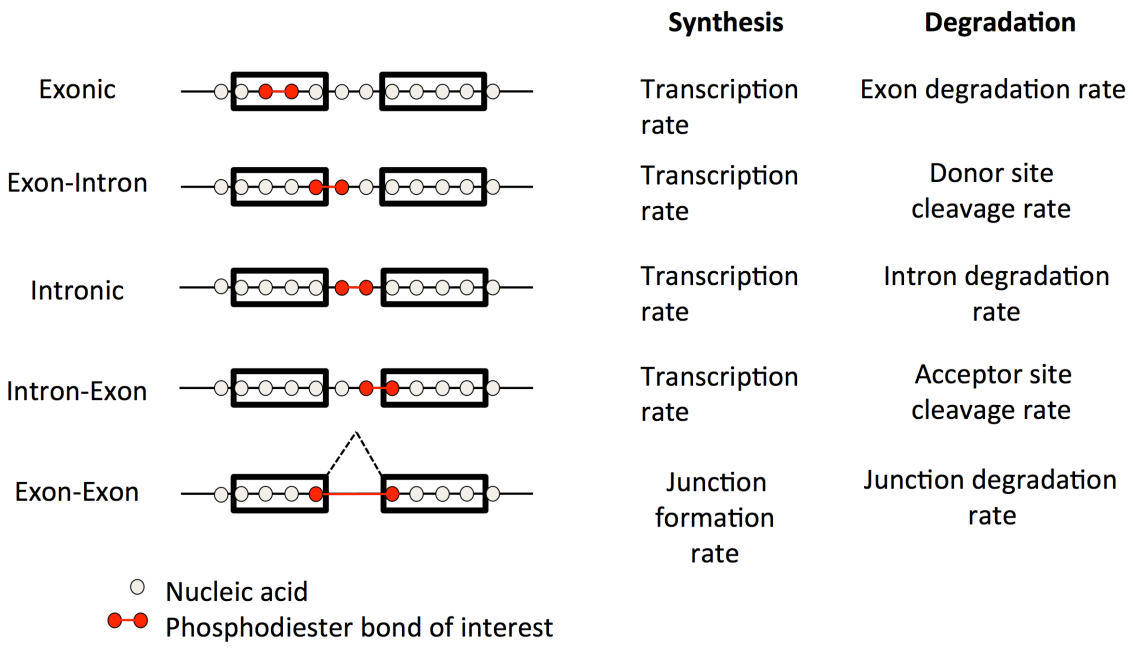


Figure 4: Synthesis and degradation rates at the level of individual phosphodiester bonds. The degradation rate at donor and acceptor splice sites corresponds to the splicing rate. Taken from (Wachutka et al. 2016).

7 Aims and scope of this thesis

Transcription and its regulation by enhancers have been intensively studied (Sections 1 and 2). Especially the human T-cell response is a widely used model system (Section 3). The development of high-resolution methods for transcriptome profiling such as RNA-seq (Section 4) provides the opportunity to measure gene expression on a genome-wide level and compare it across multiple experiments. However, different samples need to be normalized properly in order to compare them (Section 5). Most normalization methods are lacking a factor that accounts for global differences. Metabolic labeling of RNA measures RNA synthesis rate, but in order to infer RNA degradation rates correctly labeled and total RNA fractions need to be scaled relative to each other (Section 6).

In the first part of this thesis, I introduce a model to normalize TT-seq (or 4sU/4tU-seq) and RNA-seq data relative to each other, based on the use of artifical, *in vitro* labeled spike-ins. This normalization corrects for sequencing depth and the ratio of labeled to total RNA as well as for the cross-contamination of unlabeled RNA in the labeled fraction. The resulting normalization values are then used to model synthesis and degradation rates jointly across replicates of TT-seq and RNA-seq data.

In the second part of the thesis, I investigate the relationship between transcription from enhancers and promoters during the human T-cell response. To monitor immediate transcriptional changes after T-cell stimulation we use transient transcriptome sequencing (TT-seq) (Section 4.2.2). TT-seq was developed recently to detect short-lived RNAs such as eRNAs in human cells, and to estimate RNA synthesis and degradation rates (Schwalb et al. 2016) (Section 18). TT-seq involves short, 5 minute labeling of nascent RNA with 4-thiouridine (4sU). RNA is then fragmented, and the labeled RNA fragments are sequenced, providing a genome-wide view of RNA synthesis during the 5 minute labeling pulse. TT-seq is a sensitive method to detect eRNAs, because it has higher sensitivity than RNA-seq in detecting short-lived RNAs. It is more sensitive than standard 4sU labeling in detecting short RNAs because its fragmentation step confers a transcript length-independent sampling of the nascent transcriptome (Schwalb et al. 2016). Hence, TT-seq should

be ideally suited to map changes in eRNA and mRNA production during transcriptional activation, but this was not yet demonstrated. Studies, which investigate the timing of eRNA and mRNA activation, lack very early response systems, are not conducted genome-wide or are not performed *in vivo*. The T-cell activation system provides an optimal setup to study the very rapid gene expression response within minutes, that has not yet been studied applying the sensitivity of TT-seq to very early time points after activation. The temporal resolution of eRNA and mRNA expression changes are of special interest, as it is still unclear, if the eRNA transcript itself has a function on mRNA transcription regulation. eRNAs that are transcribed before their promoters could have a function on the gene expression, while co-transcribed eRNAs are more likely a transcriptional by-product without a function. Therefore, I investigate the timing of eRNA and mRNA production during early T-cell activation with TT-seq.

In the third part of this thesis, I present additional results that were obtained in collaborations applying 4tU-seq/TT-seq and the normalization method to different cell types and organisms.

Part II

Materials and Methods

In the following, I introduce mathematical and computational methods that were developed and/or applied to achieve the results presented in Section III. The first section presents the mathematical basis for the normalization based on spike-in read counts and estimation of synthesis and degradation rates from TT-seq and RNA-seq data (Section 8). The second part describes the analysis procedure of TT-seq data derived during T-cell activation (Section 9).

8 Normalization and modeling of TT-seq count data

This section presents the mathematical model to normalize TT-seq and RNA-seq data relative to each other using spike-ins. The correct ratio of labeled and total RNA in a cell can be used to infer synthesis and degradation rates. This model was implemented as an R package, for which a manuscript is under preparation. It has been first applied to human K562 data (Schwalb et al. 2016) (Section 18).

8.1 Spike-ins

Sequencing RNA involves the conversion of RNA to cDNA, fragmentation, and polymerase chain reaction (PCR) amplification of doule-stranded cDNA. Library preparation kits require specific volumes and concentrations of cDNA material as input. Therefore, varying RNA volumes due to biological variations between samples are concealed. The comparison of different RNA-seq samples regarding global effects on RNA expression has been a tedious task (Section 5). In order to account for global variations between samples from different experimental conditions, RNA from a related organism is often used as internal standard, such as *S. pombe* RNA for *S. cerevisiae* RNA measurements (Miller et al. 2012; Sun et al. 2012; Sun et al. 2013). In the case of metabolically labeled RNA samples, such as 4tU/4sU-seq or TT-seq, the labeling efficiency with 4sU is another factor that could be addressed during normalization. Additionally, the ratio of labeled RNA to total RNA in the

cells cannot directly be read out from the sequencing results, as same input volumes are used for the library preparation. Therefore, an internal control that accounts for labeling with 4sU is needed, which can be achieved by synthetic spike-ins from the External RNA Controls Consortium (ERCC) (External RNA Controls Consortium 2005). These spike-ins mimic eukaryotic RNA sequences and can be transcribed and labeled *in vitro* with 4sU (Frühauf 2015).

Katja Frühauf selected the ERCC spike-ins displayed in Table 1 based on the following criteria: similar length, similar uridine (U) content, and varying GC content between 30% and 50% (Frühauf 2015). Spike-ins were amplified via PCR with the forward primer including the T7 promoter sequence to facilitate *in vitro* transcription from the PCR product. Spike-ins were transcribed *in vitro* with either only UTP ('unlabeled spike-ins': Spike 5, Spike 9, Spike 12) or with 1:10 4sUTP:UTP ratio ('labeled spike-ins': Spike 2, Spike 4, Spike 8) (Schwalb et al. 2016). The spike-in sequences can be found in Appendix Section 24.1.

Spike-in	**ERCC-ID**	**length**	**number of Us**	**GC content [%]**	**4sU labeled**
Spike 2	ERCC-00043	1023	303	33	yes
Spike 12	ERCC-00170	1023	316	34	no
Spike 4	ERCC-00136	1033	268	42	yes
Spike 5	ERCC-00145	1042	266	44	no
Spike 8	ERCC-00092	1124	296	50	yes
Spike 9	ERCC-00002	1061	266	51	no

Table 1: Spike-ins used for normalization of TT-seq and RNA-seq data.

Margaux Michel prepared a stock of *in vitro* labeled and unlabeled spike-ins, which were mixed in equal amounts. This spike-in mix was used for the TT-seq experiments discussed in this thesis (Michel 2016). Detailed experimental methods can be found in Section 24.2. The same volume of spike-ins relative to the cell number was added to the cell lysate of every sample.

The artifical spike-ins are subjected to RNA isolation, fragmentation and labeled RNA purification, together with the sample of interest. In the total RNA fraction, all spike-ins should be present to a similar extend, and therefore have similar read counts after sequencing. The labeled RNA fraction should be depleted of unlabeled spike-ins due to labeled RNA purification.

8.2 Modeling count data obtained via sequencing 4sU-labeled and total RNA fractions

We propose a statistical model that complements the usage of artificial spike-ins in TT-seq and RNA-seq experiments. It describes observed read counts k_{ij} for a gene i in a sample j by gene-specific labeled and unlabeled RNA amounts, while accounting for various scaling factors.

In the following, the term *feature* describes either a gene or a spike-in, i.e. a region of the genome to which reads could be mapped to and counted. Let L_i be the effective length of feature i. The effective length L_i is the absolute length L_i^* of feature i minus the read length L_R.

$$L_i = L_i^* - L_R \tag{10}$$

The effective length represents the number of possible start positions of a read so that it maps with its entire length to the feature.

For every feature i, the number of expected reads in a sample j without labeled RNA purification ('total RNA', T) is dependent on the effective length L_i, the sequencing depth σ_j, the labeled and unlabeled RNA concentrations α_{ij} and β_{ij}, the RNA extraction efficiency γ_j and the number of cells N used for the experiment. Hence, the expected number of counts $E(k_{ij}^T)$ can be modeled as:

$$E(k_{ij}^T) = L_i \cdot \sigma_j \cdot (\alpha_{ij} + \beta_{ij}) \cdot \gamma_j \cdot N. \tag{11}$$

In a sample j with labeled RNA purification, the amount of labeled RNA α_{ij} is dependent on the purification efficiency δ_j for the labeled fraction. In theory, no unlabeled RNA fragments should appear in the labeled fraction, but due to some unspecific binding during labeled RNA purification we need to adjust for unlabeled RNA fragments by the cross-contamination rate ϵ_j. The cross-contamination rate is relative to the total RNA amount in the sample. The number of expected reads for feature i in a sample j with labeled RNA (L) purification can then be explained by

$$E\left(k_{ij}^L\right) = L_i \cdot \sigma_j \cdot (\delta_j \alpha_{ij} + \epsilon_j \beta_{ij}) \cdot \gamma_j \cdot N. \tag{12}$$

Note, Equation (11) can be derived from Equation (12) by setting $\delta_j = 1$ and $\epsilon_j = 1$. Hence, in a total RNA sample we get 100% of the labeled RNA (labeled RNA purification efficiency $\delta = 1$) and 100% of the unlabeled RNA (cross-contamination rate $\epsilon = 1$).

For labeled spike-ins i, the amount of unlabeled RNA $\beta_{ij} = 0$, since we assume every molecule is sufficiently labeled. The expected number of counts $E(k_{ij}^T)$ for a labeled spike-in i in a total RNA sample j can be formulated as:

$$E(k_{ij}^T) = L_i \cdot \sigma_j \cdot \alpha_i. \tag{13}$$

For a sample j with labeled RNA purification we get

$$E(k_{ij}^L) = L_i \cdot \delta_j \cdot \sigma_j \cdot \alpha_i. \tag{14}$$

Analogous to labeled spike-ins, for an unlabeled spike-in i, the amount of labeled RNA $\alpha_{ij} = 0$, since they were reverse transcribed in the absence of 4sUTP. Hence, the expected number of counts in total and labeled RNA samples for an unlabeled spike-in i can be formulated as:

$$E(k_{ij}^T) = L_i \cdot \sigma_j \cdot \beta_i \tag{15}$$

and

$$E(k_{ij}^L) = L_i \cdot \sigma_j \cdot \epsilon_j \cdot \beta_i. \tag{16}$$

From gene annotations and spike-in sequences, we can derive the effective length L_i for each feature i. For simplification, we set $N = 1$, because the number of cells is theoretically identical across all our samples (50 Mio, see Section 24.2).

The estimation of absolute values for the unknown parameters δ, ϵ, and σ is not possible, as they would always be relative to each other. Therefore, we set $\delta_j = 1$. Setting $\alpha_i = 1$ in Equation (14), we can get the sequencing depth σ_j for every sample j. From Equation (16) with $\beta_i = 1$ we can derive the cross-contamination rate ϵ_j for every sample j.

For each gene i, α_{ij}, β_{ij}, and γ_j are not identifiable from Equations (11 and 12).

This means we cannot control for RNA extraction efficiency and set $\gamma_j = 1$. Thus, Equations (11)-(16) can be reformulated as:

$$E(k_{ij}) = L_i \cdot \sigma_j \cdot (\alpha_i + \epsilon_j \beta_i) \tag{17}$$

with $\epsilon_j = 1$ for total RNA samples.

8.3 A model for normalization with spike-ins

The labeled and unlabeled spike-ins can be used to normalize TT-seq ('labeled RNA') and RNA-seq ('total RNA') samples and set them into relation to each other. Unlabeled spike-ins mimic unlabeled RNA fragments, while labeled spike-ins imitate newly synthesized, labeled RNA fragments. The spike-ins undergo the same steps of sample preparation as real RNA. All reads obtained by sequencing are mapped to a combined reference genome with additional 'chromosomes' containing the spike-in sequences (Appendix Section 24.1). The mapped reads on the spike-in sequences are then counted for each spike-in individually. Our proposed statistical model to normalize TT-seq and RNA-seq samples is based on these spike-in read counts.

In order to estimate the sample-specific parameters for sequencing depth σ_j and cross-contamination ϵ_j we perform a multiple regression analysis on spike-in read counts.

Equation (17) can be extended to also allow for some variability between spike-ins, e.g. sequence variations that lead to different polymerase chain reaction (PCR) amplification biases, by multiplication with a spike-in specific parameter ρ_i:

$$E(k_{ij}) = L_i \cdot \sigma_j \cdot (\alpha_i + \epsilon_j \beta_i) \cdot \rho_i. \tag{18}$$

We set $\alpha_i = 1$ and $\beta_i = 0$ for the labeled spike-ins, and $\alpha_i = 0$ and $\beta_i = 1$ for the unlabeled spike-ins. This assumption holds true because we assume that *in vitro* transcription in the presence of 4sU works efficiently and the spike-ins are labeled sufficiently for purification. We use the same volumes for all spike-ins in all samples (see Section 8.1), therefore, their relative ratio is 1. We fit a Generalized Linear Model (GLM) to estimate σ_j, ϵ_j, and ρ_i. As the effective length L_i for each spike-in

i is known, we use it as offset in the linear predictor:

$$log(E(k_{ij})) = log(L_i) + log(\sigma_j) + log(\epsilon_j) + log(\rho_i). \tag{19}$$

Using a GLM allows the response variables (i.e. the observed read counts) to follow other error distributions than a normal distribution. A negative binomial distribution is assumed to model the expected read counts. In count data, small numbers often have a large variation. This is the so-called overdispersion (Cameron et al. 1998), which is also true for sequencing read counts. Therefore, an appropriate distribution needs to be chosen to model read counts. While for the Poisson distribution variance and mean are equal, the negative binomial distribution is a generalization that allows for a larger variance, and it is commonly used to model sequencing read counts (Anders et al. 2010; Eser et al. 2016).

For the spike-in read counts, the GLM is fitted jointly for all samples by maximum likelihood (ML) estimation. Thereby, all spike-ins in one sample contribute to its sequencing depth value σ, the unlabeled spike-ins in a labeled sample to its cross-contamination estimate ϵ, and the read counts for each spike-in across the different samples contribute to the spike-in specific factor ρ (Table 2).

	4sU	Sample	sequencing depth	cross-contamination rate	spike-in specific factor
Spike 2	yes	TT-seq	σ_1		ρ_{Spike2}
Spike 4	yes		σ_1		ρ_{Spike4}
Spike 5	no		σ_1	ϵ_1	ρ_{Spike5}
Spike 8	yes		σ_1		ρ_{Spike8}
Spike 9	no		σ_1	ϵ_1	ρ_{Spike9}
Spike 12	no		σ_1	ϵ_1	$\rho_{Spike12}$
Spike 2	yes	RNA-seq	σ_2	1	ρ_{Spike2}
Spike 4	yes		σ_2	1	ρ_{Spike4}
Spike 5	no		σ_2	1	ρ_{Spike5}
Spike 8	yes		σ_2	1	ρ_{Spike8}
Spike 9	no		σ_2	1	ρ_{Spike9}
Spike 12	no		σ_2	1	$\rho_{Spike12}$

Table 2: Contribution of spike-ins to global scaling factors in different samples.

8.4 A model for estimating synthesis rates and half-lives during steady state

The method to estimate synthesis and degradation rates from 4sU-labeled and total RNA-seq data is based on previous studies (Miller et al. 2011; Schwalb 2012). Steady-state conditions are assumed, meaning the amount of RNA in the cell is constant and there is a dynamic equilibrium of a constant RNA synthesis and decay.

After applying the model from Section 8.3 to spike-in read counts, the values for sequencing depth σ_j and cross-contamination ϵ_j per sample j are fixed and used for the estimation of gene-specific parameters. Using Equations (11) and (12), the RNA amounts α_{ij} and β_{ij} can be estimated independently for every gene.

To account for the overdispersion of count data, a negative binomial function is assumed to model the read counts:

$$E(k_{ij}) \sim NB(\mu_{ij}, \phi_j). \tag{20}$$

The dispersion parameter ϕ_j is estimated per gene individually for labeled samples (4sU/4tU-seq, TT-seq) and total RNA samples, using the *DESeq2* implementation (Love et al. 2014b).

This model is fitted by maximum likelihood to transcript read counts to provide estimates of the labeled and unlabeled RNA amounts α_i and β_i for a pair of TT-seq and RNA-seq measurements. The total RNA amount υ_i for a gene i is the sum of labeled and unlabeled RNA amounts per cell:

$$\upsilon_i = \alpha_i + \beta_i. \tag{21}$$

Previous studies (Miller et al. 2011) have shown that labeled and total RNA amounts α_i and υ_i can be explained by the following equations:

$$\alpha_i = \frac{\mu_i}{\lambda_i} \cdot \left(1 - e^{-\lambda_i \cdot t}\right) \tag{22}$$

and

$$\upsilon_i = \frac{\mu_i}{\lambda_i}. \tag{23}$$

Therefore,

$$\alpha_i + \beta_i = \frac{\mu_i}{\lambda_i}. \tag{24}$$

The synthesis rate μ_i and the degradation rate λ_i can be calculated from Equations (22) and (24) assuming first-order kinetics:

$$\lambda_i = -\frac{1}{t} \cdot log\left(\frac{\beta_i}{\alpha_i + \beta_i}\right), \tag{25}$$

and

$$\mu_i = (a_i + \beta_i) \cdot \lambda_i. \tag{26}$$

Assuming exponential decay, gene-specific half-lives $t_{1/2,i}$ can be calculated from the estimated degradation rates (see Section 6):

$$t_{1/2,i} = \frac{log(2)}{\lambda_i}. \tag{27}$$

The estimated half-life times are on an absolute level (in minutes), but synthesis rates are on an arbitrary scale. Absolute synthesis rates can be obtained by scaling the values, so that total RNA levels match reported expression levels (Eser et al. 2016).

By applying this model individually to exon read counts and intron read counts, we can obtain local synthesis and degradation rates, which reflect synthesis rates and half-lives per nucleotide bond. Note that for data sets where the steady state assumption does not hold (e.g. activation time course), this model is not sufficient.

8.5 Implementation and availability

The model for normalization with spike-ins and subsequent estimation of synthesis and degradation rates was implemented in R (R Development Core Team 2011). Fitting this model to the observed spike-in read counts is done using the R function *glm.nb* from the MASS package (Venables et al. 2002) with the default 'log' link function. An open source R/Bioconductor package is under preparation in collaboration with Leonhard Wachutka and Julien Gagneur. The package includes additional functionalities such as counting reads for specific features and estimating splicing

rates thereby using the functionalities of the *SummarizedExperiment*-class (Morgan et al. 2016).

9 Analysis of a TT-seq data set obtained from an activation time course in human T-cells

This section describes computational procedures and methods that were carried out to analyze a set of TT-seq and RNA-seq data sets from an activation time course of human T-cells. Detailed experimental methods can be found in Section 24.2. The results are presented in Section III.

The methods described here have been published in:

TT-seq captures enhancer landscapes immediately after T-cell stimulation
M. Michel*, **C. Demel***, B. Zacher, B. Schwalb, S. Krebs, H. Blum, J. Gagneur, and P. Cramer
Molecular Systems Biology (2017)

For detailed author contributions see page ix.

9.1 Replicate measurements

We prepared TT-seq and total RNA-seq libraries for two biological replicates. For total RNA-seq, there were essentially no significant changes between time points, and the samples showed very high correlations (Spearman correlation ≥ 0.98) and can be seen as replicates (Figure A1). Replicate TT-seq libraries for time points 0 min and 10 min after T-cell activation were obtained and showed high correlation (Spearman correlation coefficient 0.97, Appendix Figure A2). Based on these results, it was clear that the data are highly reproducible and of high quality, making further replicate measurements obsolete. For all subsequent analyses, replicates were averaged after size factor normalization, where available. For transcriptome annotation (Section 9.5), all TT-seq samples were used, irrespective of their sequencing depth, as GenoSTAN (Zacher et al. 2017) places more weight on deeper sequenced samples.

9.2 Sequencing data processing

Paired-end 50 base reads with additional 6 base reads of unique barcodes were obtained for each of the samples. Reads were demultiplexed and 150-250 Mio read pairs per sample were mapped unambiguously with STAR (version 2.3.0) (Dobin et al. 2015) to the hg20/hg38 (GRCh38) genome assembly (Human Genome Reference Consortium). Samtools (Li et al. 2009) was used to quality-filter SAM files, whereby alignments with MAPQ smaller than 7 (*-q 7*) were skipped and only proper pairs (*-f99, -f147, -f83, -f163*) were selected. Further data processing was carried out using the R/Bioconductor environment (Gentleman et al. 2004; R Development Core Team 2011).

9.3 Data availability

The sequencing data sets have been deposited in the Gene Expression Omnibus (GEO) database under accession code GSE85201.

9.4 Antisense correction

For merged transcribed regions from GENCODE, we selected strand-specific genomic regions where no antisense annotation existed in GENCODE. This ensured to only take unique regions into account where antisense transcription should not be present. For all genomic positions in those regions, where the sense coverage exceeded 100 reads (i.e. highly expressed regions), we calculated the median ratio of antisense-to-sense coverage (including one pseudo-count). This value provides an estimate of the antisense bias (c) in every sample. We corrected the observed coverage/read counts for Watson and Crick strands, respectively, by solving the following formulas, which assume that the observed sense coverage is the sum of 'real' sense coverage and a small percentage (i.e. the antisense bias value c) of the 'real' antisense coverage:

$$Coverage^{sense}_{real} = \frac{Coverage^{sense}_{observed} + c \cdot Coverage^{antisense}_{observed}}{1 - c^2}. \tag{28}$$

For antisense correction of coverage profiles, the antisense coverage was averaged in a symmetrical 51 nt window around the position on the sense strand which should be normalized. For all further analyses (including transcriptome annotation, calculation of expression values, fold changes, and synthesis/degradation rates) antisense-corrected feature counts (rounded to the nearest integer) were used.

9.5 Transcription Unit (TU) annotation and classification

Genome-wide strand-specific coverage was calculated from fragment midpoints in consecutive 200 bp bins throughout the genome for all TT-seq samples. Binning reduced the number of uncovered positions within expressed transcripts and increased the sensitivity for detection of lowly synthesized transcripts. To overcome antisense bias due to highly expressed genes, an antisense correction was performed on each bin (as described in the previous paragraph). A pseudo-count was added to each bin to mask noisy signals. The R/Bioconductor package GenoSTAN (Zacher et al. 2017) was used to learn a two-state hidden Markov model with a PoissonLog-Normal emission distribution in order to segment the genome into 'transcribed' and 'untranscribed' states, which resulted in 139,507 transcribed regions.

Transcription units (TUs) that overlapped at least to 20% of their length with a protein-coding gene or a lincRNA annotated in GENCODE (gtf column 'transcript_type' either 'protein_coding' or 'lincRNA') and overlapped with an exon of the corresponding annotated feature, were classified as protein-coding/lincRNA, the rest was assumed to be ncRNAs. TUs mapping to exons of the same protein-coding gene/lincRNA were combined. In order to filter spurious predictions a minimal expression threshold for TUs was defined based on overlap with genes annotated in GENCODE. The threshold was optimized using the Jaccard index criterion, and resulted in 27,558 TUs with minimal 16.5 reads per kilobase (RPK) (Figure 6B). In order to overcome low expression or mappability issues, ncRNAs that are only 200 bp (1 bin) apart, were merged. Subsequently, TU start and end sites were refined to nucleotide precision by finding borders of abrupt coverage increase or decrease between two consecutive segments in the two 200 bp bins located around the initially assigned start and stop sites via fitting a piecewise constant curve to the coverage

profiles (whole fragments) for all TT-seq samples using the segmentation method from the R/Bioconductor package *tilingArray* (Huber et al. 2006). Overlapping transcripts (arising through overlaps with multiple annotated genes) were merged using the *reduce* function from the GenomicRanges package and assigned the corresponding protein-coding or lincRNAs GENCODE ID, if existing. 612 annotated transcripts (that were included to calculate DESeq size factors) that overlapped with multiple protein-coding genes by at least 75% of the GENCODE transcript length and 20% of our transcript were removed from further analyses, because they could not be clearly assigned to one gene. Protein-coding transcripts shorter than 5 kb and overlapping less than 10% with any GENCODE protein-coding gene were classified as 'ncRNA'. All ncRNAs with starting sites up to 1 kb downstream of a protein-coding gene on the sense strand were omitted in enhancer analysis or eRNA comparisons, as these reads might come from read-through transcription after the transcript sequence. This resulted in 22,141 non-ambiguously classified RNAs (8,878 protein-coding genes, 590 lincRNAs, and 12,673 ncRNAs), on which the rest of the analysis was focused. The class of eRNAs was comprised of 5,616 of our ncRNAs, where either the transcript or the region 1 kb upstream of the ncRNA overlapped with an enhancer annotated by GenoSTAN in at least one T-cell line (Zacher et al. 2017).

9.6 Estimation of RNA synthesis rates and half-lives

To overcome inconsistent coverage throughout a gene due to splicing and multiple isoforms, constitutive exons (Bullard et al. 2010) were determined for all our mRNA and lincRNA transcripts. Read counts for those constitutive exons and all other ncRNA classes across all TT-seq and RNA-seq samples were calculated using *HTSeq* (Anders et al. 2014). To estimate rates of RNA transcription and degradation we used the same approach as described in (Schwalb et al. 2016) and in Section 8. Briefly, we used a statistical model that describes read counts k_{ij} (in a TT-seq or RNA-seq sample) by the length of the feature (spike-in/transcript) i, L_i, and feature-specific labeled und unlabeled RNA amounts, α_i and β_i: $E(k_{ij}) = L_i\sigma_j(\alpha_{ij} + \epsilon_j\beta_{ij})$. We calculated the sequencing depths σ_j and cross-contamination ε_j rates per sample

j based on the spike-in read counts by setting $\alpha_{ij} = 1$ and $\beta_{ij} = 0$ for labeled spike-ins, and $\alpha_{ij} = 0$ and $\beta_{ij} = 1$ for unlabeled spike-ins. In a total RNA-seq sample, ε_j is fixed to 1, and in a TT-seq sample ε_j is close to 0, as we enrich for labeled RNA. Then, this model was fitted by maximum likelihood to transcript read counts to provide estimates of the labeled and unlabeled RNA amounts α_i and β_i for a pair of TT-seq and RNA-seq measurements. The synthesis rate μ_i and the degradation rate λ_i were calculated from α_i and β_i assuming first-order kinetics as in (Miller et al. 2011) in the following way: $\lambda_i = -\frac{1}{t} log(\beta_i/(\alpha_i + \beta_i))$ and $\mu_i = (a_i + \beta_i)\lambda_i$.

9.7 Differential gene expression

Gene expression fold changes upon T-cell stimulation for each time point were calculated using the R/Bioconductor implementation of *DESeq2* (Love et al. 2014b). The DESeq size factor was only estimated on our set of protein-coding genes. Differentially expressed genes were identified applying a fold change cutoff of 2 and an adjusted P-value cutoff of 0.05 comparing each time point to the 0 min measurements. For the absolute numbers of genes with changed synthesis, we checked if the TT-seq read count is significantly (adjusted P-value ≤ 0.05) changed at least 2-fold at any time point compared to time point 0 min.

9.8 Motif analysis

DNA motifs in the form of position weight matrices (PWMs) were downloaded from the JASPAR database via the R/Bioconductor package *JASPAR2016* (Tan 2015). Each PWM was screened against a positive and a negative set of sequences (e.g. 250 bp upstream sequences of eRNAs and remaining ncRNAs) with the *searchSeq* function in the *TFBSTools* package (Tan et al. 2016). We defined a cutoff to distinguish between motif occurrence and not-occurrence as 80% of the maximal score that the PWM could reach. The number of sequences in which the motif occurred was counted for the positive and negative set, and an odds ratio was calculated.

9.9 eRNA-mRNA pairing

We paired all eRNAs and mRNAs in all possible combinations, as long as both transcript TSSs are within the same insulated neighborhood, defined by ChIA-PET Anchor sites (Section 9.10) using the *findOverlaps* function from the *GenomicRanges* package (Lawrence et al. 2013). Pairs were removed, where the eRNA TSS fell into the region [TSS; TSS+1000] around the protein-coding gene's TSS.

9.10 External data processing

Experimentally validated enhancers were downloaded from the VISTA enhancer browser (`http://enhancer.lbl.gov/frnt_page_n.shtml`). ENCODE DNase-seq raw coverage files (for Figure 10D) and peak files (for Figure 9) for Jurkat cells were retrieved from

`https://genome.ucsc.edu/ENCODE/dataMatrix/encodeDataMatrixHuman.html`

and replicates were merged. Enhancer and DNaseI hypersensitivity sites (DHS) coordinates were converted to hg20 coordinates using the *liftover* function in the R/Bioconductor package *rtracklayer* (Lawrence et al. 2009). ChIA-PET interaction domains processed with the Mango pipeline were downloaded from a previous study (Hnisz et al. 2016b) and were selected for P-values < 0.2. ChIA-PET Anchor sites were converted to hg20 coordinates using the *liftover* function in the R/Bioconductor package *rtracklayer* (Lawrence et al. 2009) followed by a *reduce* with *min.gapwidth=60* which closes 90% of the gaps arising by liftover, in order to get continuous genomic regions.

Part III

Results and Discussion

All results presented in this section were obtained in collaboration with Margaux Michel and are published in:

TT-seq captures enhancer landscapes immediately after T-cell stimulation
M. Michel*, **C. Demel***, B. Zacher, B. Schwalb, S. Krebs, H. Blum, J. Gagneur, and P. Cramer
Molecular Systems Biology (2017)

The full article with supplementary materials and tables can be found at `http://msb.embopress.org/content/13/3/920`.

Contribution: *I carried out all bioinformatics analyses for this project.*

In this study, we use TT-seq to monitor the immediate T-cell response over the first 15 min after cell stimulation. We identify new immediate, direct target genes of the T-cell response, and show that activation of immediate enhancers and promoters, as defined by RNA production, occurs simultaneously. The results also establish TT-seq as a simple-to-use, very sensitive tool to investigate transcriptional responses at high temporal resolution, ideally suited to monitor rapid changes in enhancer landscapes and in transcriptional programs during cellular differentiation and reprogramming.

10 Monitoring the immediate T-cell response

We monitored immediate changes in RNA synthesis in Jurkat T-cells during the first 15 min after stimulation with ionomycin and PMA using both TT-seq and RNA-seq (Figure 5A, Sections 9.2 and 24.2, Appendix Figures A1 and A2). We selected time points before stimulation (0 min), and 5, 10, and 15 min after stimulation. The TT-seq data revealed strong up- and down-regulation of mRNA synthesis for immediately responding genes (Figures 5B and C). In TT-seq data, we also observed a high coverage of intronic regions and regions downstream of the polyadenylation site (PAS, annotated by GENCODE (Harrow et al. 2012)), demonstrating that

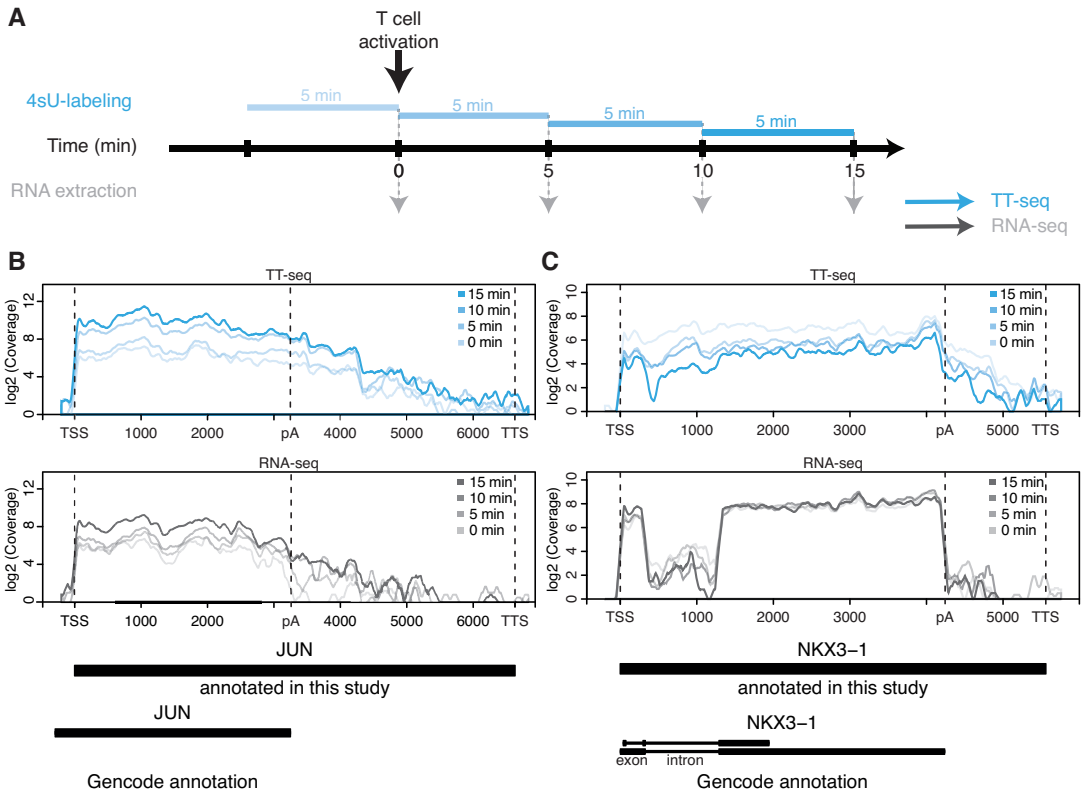


Figure 5: TT-seq analysis of immediate response to T-cell stimulation. **(A)** Experimental design. RNA in cells was labeled with 4-thiouracile (4sU) for consecutive 5 minute intervals. Total and 4sU-labeled RNA was extracted before T-cell stimulation and 5, 10, and 15 min after T-cell stimulation and subjected to deep-sequencing. **(B)** Exemplary genome browser view for an upregulated mRNA (*JUN*). Upper panel: TT-seq data for 0, 5, 10 and 15 min after stimulation; lower panel: total RNA-seq data for 0, 5, 10 and 15 min after stimulation. **(C)** Exemplary genome browser view for a downregulated mRNA (*NKX3-1*), analogous to (B).

TT-seq could trap short-lived RNA (Figures 5B and C).

We combined the TT-seq data to segment the genome into transcribed and non-transcribed regions using GenoSTAN (Zacher et al. 2017). Then, we automatically annotated a total of 22,141 transcribed regions ('transcripts') before and after T-cell stimulation (RPK cutoff = 16.5, Section 9.5, Figure 6). Comparison with the GENCODE annotation (Harrow et al. 2012) enabled us to classify our annotated transcripts into 8,878 mRNAs, 590 long non-coding RNAs (lincRNAs), by requiring at least 20% of the transcribed region to overlap with GENCODE annotated 'protein_coding' or 'lincRNA' (long, intervening noncoding RNA that can be found in evolutionarily conserved, intergenic regions) transcripts (Section 9.5). The 12,673 remaining transcripts we categorized as non-coding RNAs (ncRNAs) (Figure 6C). These RNAs may contain additional long non-coding RNAs that don't fall into GENCODE's 'lincRNA' definition (Section 9.5).

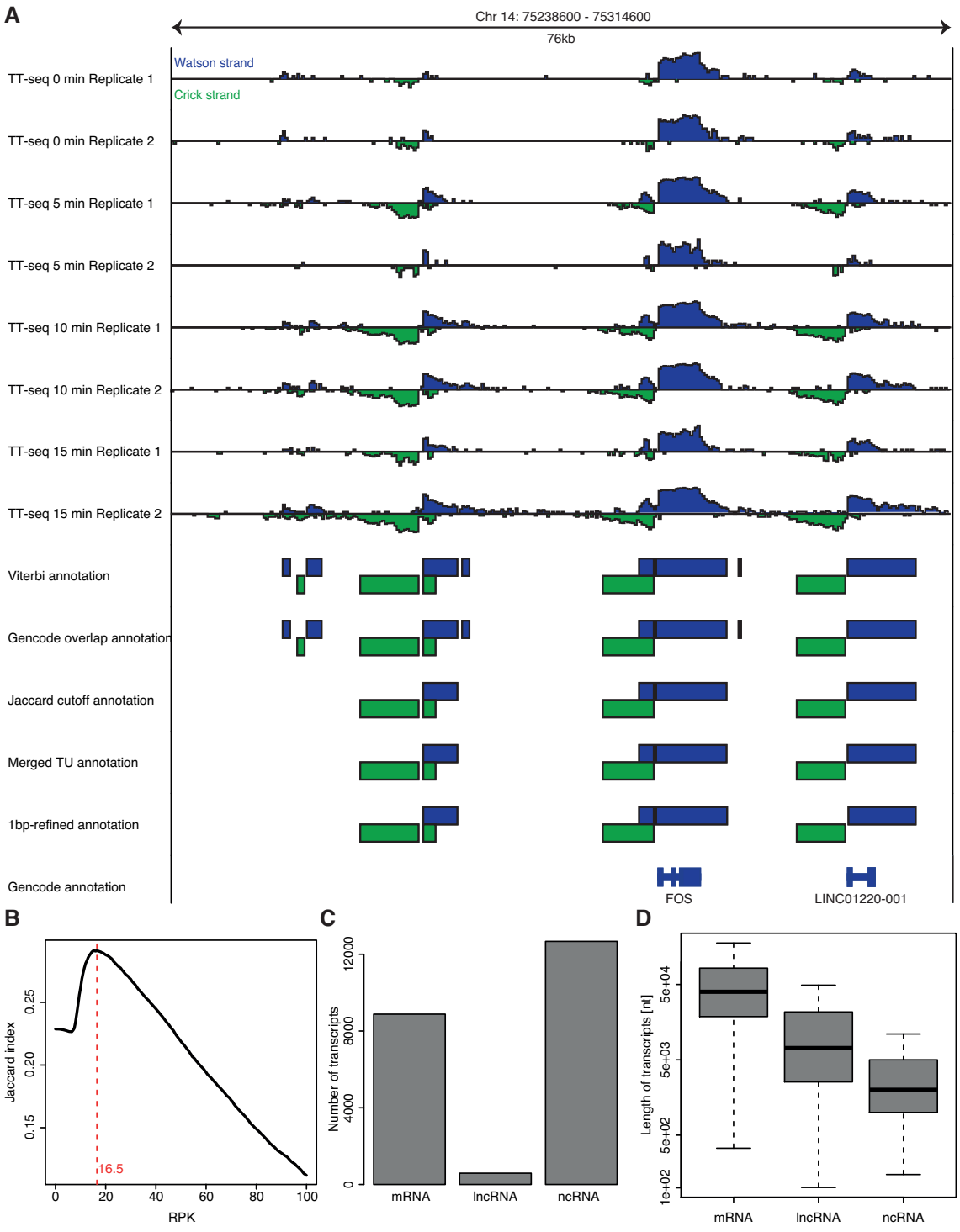


Figure 6: Annotation of transcripts. **(A)** Segmentation workflow. The Watson and Crick strands are in dark blue and green, respectively. The top 8 tracks show antisense-corrected TT-seq data tracks (log2 scale) that were used as input for GenoSTAN. The other tracks indicate the stepwise annotation of transcripts. From the GENCODE annotation, only full transcripts with transcript_support_level 1 are depicted. **(B)** Jaccard index (compared to GENCODE annotation) for different choices of thresholds (x-axis: Reads Per Kilobase (RPK)). The red line indicates the selected RPK value where the Jaccard index reaches the maximal value. **(C)** Number of transcripts per transcript class. **(D)** Distribution of transcript lengths per transcript class. Box limits are the first and third quartiles, the band inside the box is the median. The ends of the whiskers extend the box by 1.5 times the interquartile range.

The length distribution of RNAs in these classes (Figure 6D) resembled that in our previous study of human K562 cells (Schwalb et al. 2016). Steady-state RNA synthesis rate and half-life distributions also agreed with previous results (Figure 7, Section 9.6). Taken together, we obtained a transcriptome annotation for T-cells that included both stable transcripts present during steady-state growth and short-lived RNAs that are produced immediately after stimulation.

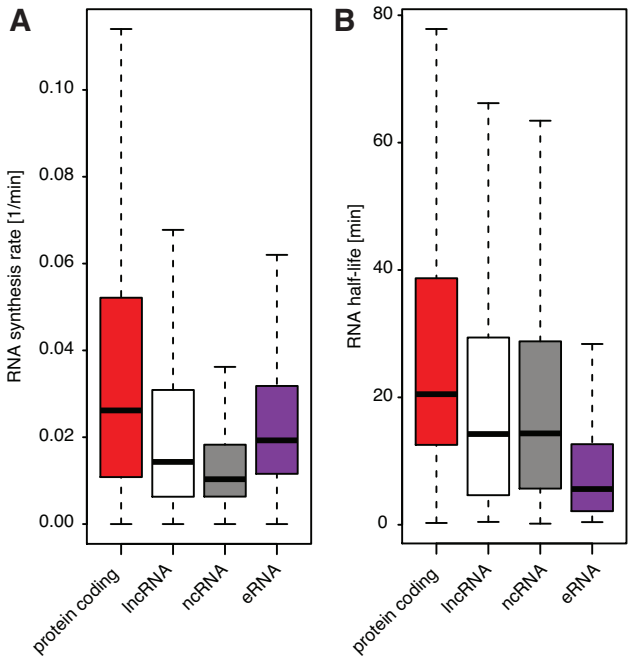


Figure 7: Half-life and synthesis rate distribution of transcript classes. **(A)** Distribution of synthesis rates for different transcript classes. **(B)** Distribution of half-lives for different transcript classes. Data information: Box limits are the first and third quartiles, the band inside the box is the median. The ends of the whiskers extend the box by 1.5 times the interquartile range.

11 TT-seq uncovers many immediate response genes

We next analyzed changes in transcript coverage upon T-cell stimulation after integrating reads over transcribed units at different time points. When we compared time points 5, 10, and 15 min with time point 0 min, RNA-seq data did not reveal any significant (FC $>$ 2, adjusted P-value $<$ 0.05, Methods Section 9.7) changes. In contrast, TT-seq uncovered hundreds of newly synthesized transcripts with significantly changed signals already after 5 min, and thousands of changed transcripts after 15 min following stimulation (FC $>$ 2, adjusted P-value $<$ 0.05, Section 9.7,

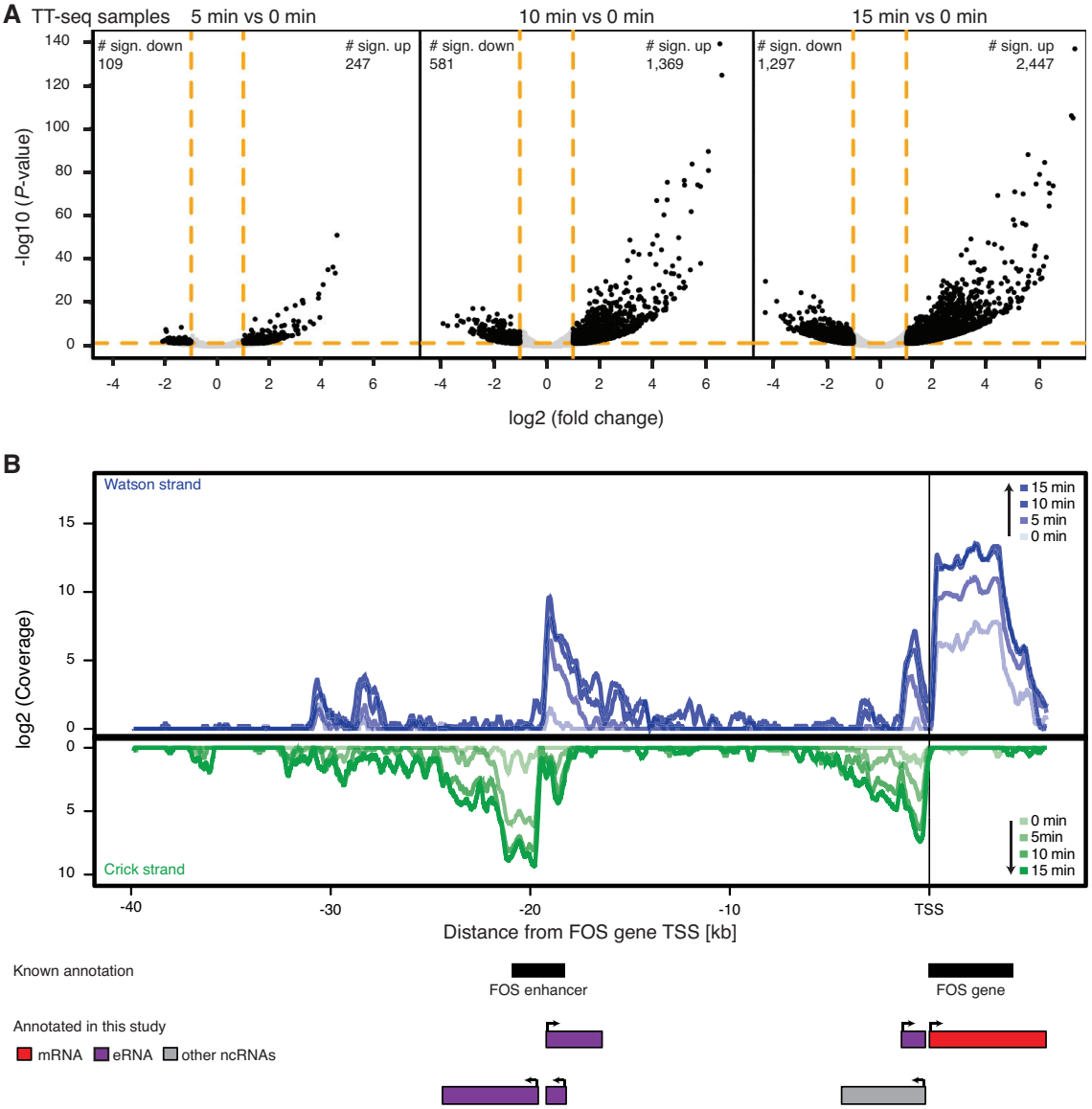


Figure 8: TT-seq captures transcriptional changes after T-cell stimulation. **(A)** TT-seq signal for statistically significant differentially expressed transcripts after 5, 10, and 15 minutes compared to time point 0 min, before cell stimulation. Significantly differentially expressed genes are indicated by black points. The numbers in the plots correspond to the numbers of significantly up/down-regulated transcripts at each time point. The vertical orange lines indicate the fold change cutoff of 2, and the horizontal orange line shows the *P*-value cutoff of 0.05. **(B)** Sense and antisense transcription mapped with TT-seq at the *FOS* gene and its annotated upstream enhancer. The rectangles with arrows indicate transcripts annotated in this study.

Figure 8A, Tables 3 and 4). These results show that transcription activity in T-cells changes immediately upon stimulation and that TT-seq captures transcriptional up- and down-regulation with great sensitivity long before changes in RNA levels are detected by RNA-seq.

Out of a total of 3,744 transcripts that showed significantly changed synthesis 15 min after stimulation, 638 were mRNAs, and 2,986 were ncRNAs, including

120 lincRNAs (Tables 3 and 4). Many up-regulated mRNAs encode known marker proteins of T-cell activation, such as *FOS*, *FOSB*, *JUN*, *JUNB* and *CD69*. Other up-regulated mRNAs stemmed from known immediate-early response genes, such as transcription factors *EGR1*, *EGR2*, *EGR3*, and *NR4A1*, and the stem cell identity factor *KLF4*. However, the majority of the up-regulated mRNAs that we detected had not been described in association with T-cell stimulation. Of the 638 differentially expressed mRNAs, only ~20% were known to be involved in T-cell activation (Cheadle et al. 2005; Diehn et al. 2002; Ellisen et al. 2001). Amongst the newly detected up-regulated genes were those that encode GPR50, KLF4, DUSP1, PPP1R15A, MASP2, and RGCC proteins that are involved in processes, such as MAPK signaling or other signaling pathways, the immune response, or the response to stimuli. Thus, the high sensitivity of TT-seq can uncover new target genes even in very well-studied systems.

	mRNAs	lincRNAs	ncRNAs (eRNAs)	Total
5 min	29	12	206 (135)	247
10 min	132	42	1,195 (594)	1,369
15 min	311	78	2,058 (897)	2,447

Table 3: Number of up-regulated transcripts per class and time point after activation. The number of ncRNAs includes the number of eRNAs, which is shown in parenthesis.

	mRNAs	lincRNAs	ncRNAs (eRNAs)	Total
5 min	9	1	99 (70)	109
10 min	42	16	523 (359)	581
15 min	327	42	928 (629)	1,297

Table 4: Number of down-regulated transcripts per class and time point after activation. The number of ncRNAs includes the number of eRNAs, which is shown in parenthesis.

12 Defining the dynamic landscape of transcribed enhancers

The vast majority of transcripts with significantly changed synthesis after stimulation were ncRNAs. When we investigated the TT-seq coverage at known enhancers,

we observed increasing RNA synthesis, showing that we could monitor eRNA production at transcribed enhancers such as the one at the *FOS* locus (Figure 8B). Within 15 min after stimulation, eRNA synthesis at this locus increased about 160-fold, whereas synthesis of *FOS* mRNA increased about 40-fold (Figure 8B). The TT-seq coverage profiles also immediately revealed bidirectional transcription at both the promoter and a known enhancer at the *FOS* locus. Thus, enhancer transcription is very well captured by TT-seq, encouraging us to fully describe the landscape of transcribed enhancers and its changes during T-cell stimulation.

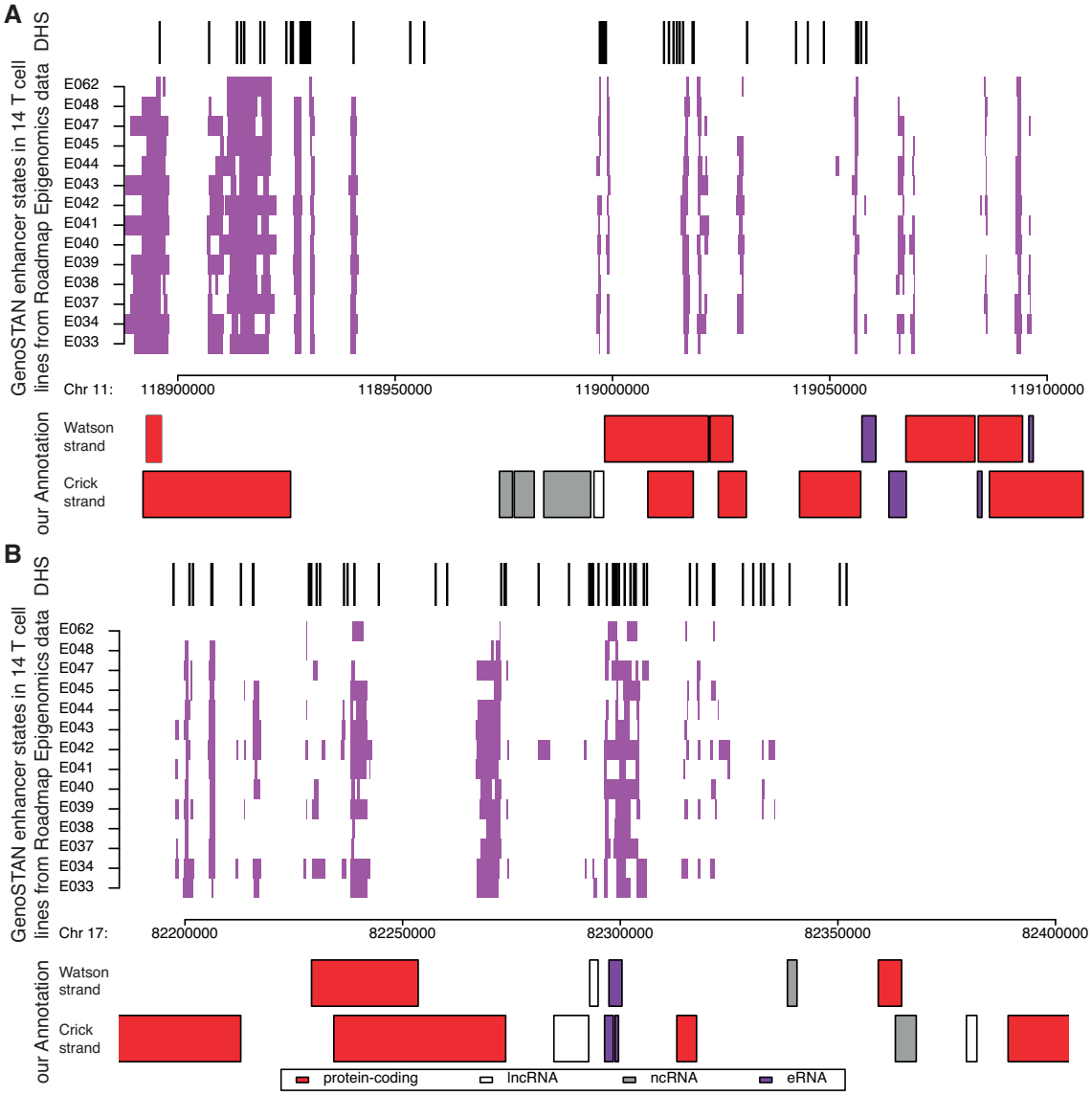


Figure 9: Example of eRNA identification using GenoSTAN. **(A)** Shown are the DNAse hypersensitivity signal (DHS) from ENCODE (top, black marks), the GenoSTAN enhancer states (violet colored) obtained from from 14 Roadmap Epigenomics T-cell lines (middle), and the obtained transcript annotation (bottom) for a region on chromosome 11. **(B)** As in A for a region on chromosome 17.

To select putative eRNAs from our annotated 12,673 ncRNAs, we used our recent GenoSTAN annotation of chromatin states in genomes of 14 T-cell lines, which is based on the integration of publicly available chromatin marks and DNA accessibility data (Zacher et al. 2017). We compared all ncRNAs with all enhancer states from T-cells (Zacher et al. 2017) (Figure 9). This resulted in 5,616 (44%) ncRNAs that overlapped with enhancer states either with their transcribed region or with the region 1,000 bp upstream, and were therefore classified as putative eRNAs (Figure 10A).

13 Immediate, nucleosome-depleted enhancers

Out of a total of 50,810 annotated T-cell enhancer states, 7,865 produced eRNAs in our cell line and under our conditions that we could detect. The obtained putative 5,616 eRNAs showed a similar length distribution as the remaining 7,057 ncRNAs (Figure 10B, Appendix Figure A3A), but had shorter half-lives (Figure 10C, Appendix Figure A3B), reflecting the known unstable nature of eRNAs. The sets of putative active eRNAs (applying the same cutoff as for the transcriptome annotation, RPK $\geq$ 16.5) comprises more than 5,000 actively transcribed eRNAs at each time point (Figure 10A). For a large fraction of eRNAs (29%) we observed significant changes in their synthesis during the time course compared to the initial time point (Section 9.7), showing that eRNA transcription is highly regulated.

Consistent with the chromatin state annotation of enhancers, the putative eRNAs were flanked by a region of high DNase hypersensitivity (The ENCODE Project Consortium 2012) immediately upstream, which was not the case for the remaining ncRNAs (Figure 10D). In addition, the region 250 bp upstream of the eRNA TSS was significantly enriched for binding sites of transcription factors that act during T-cell activation, namely *EGR1*, *EGR2*, *ERG3*, *JUNB*, *REL*, *FOSL2*, *FOS* (odds ratios 6.8, 4.3, 3.8, 1.6, 1.6, 1.5, and 1.3, respectively, Section 9.8), compared to other ncRNA upstream sequences (Figure 10E). This strongly indicates that our set of putative eRNAs represents transcripts originating from enhancers that are relevant for the T-cell response. Taken together, TT-seq can define the landscape of actively transcribed enhancers, and its changes during T-cell stimulation.

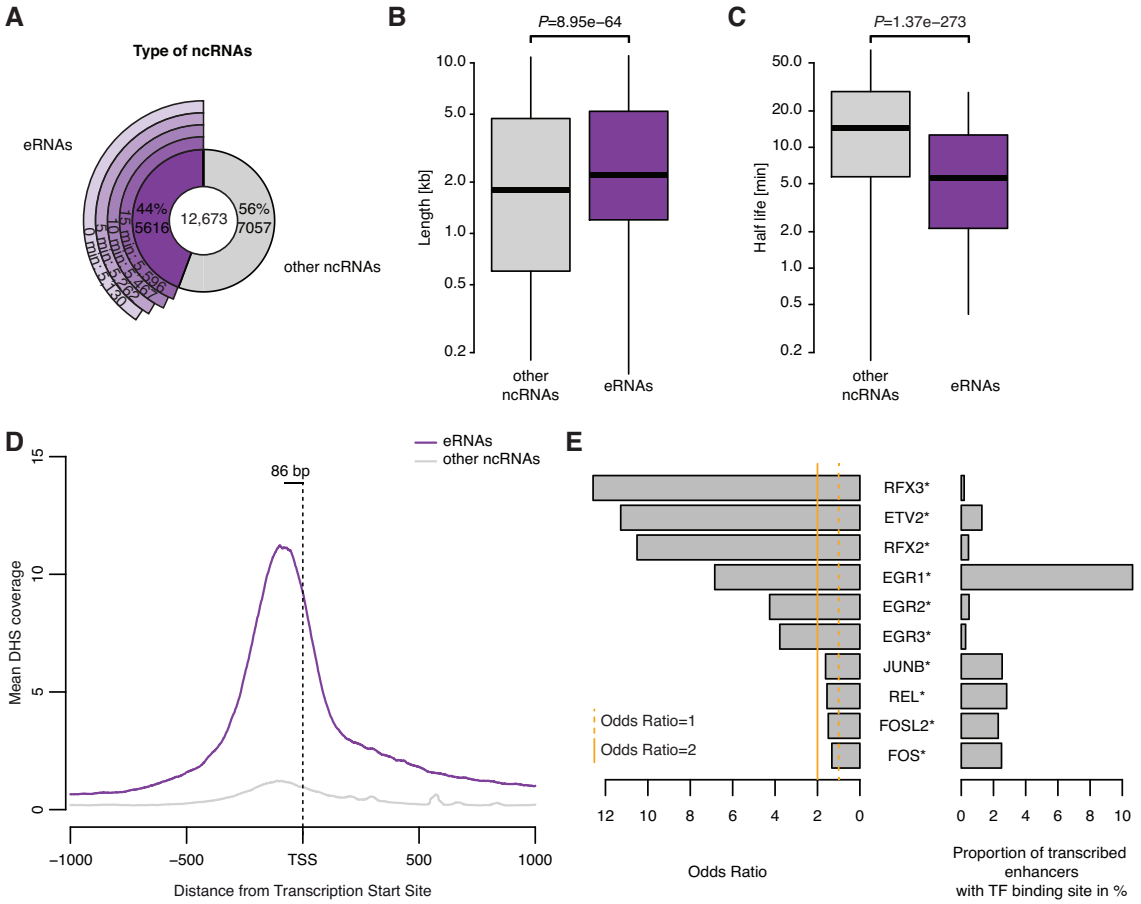


Figure 10: Characteristics of transcribed enhancers. **(A)** Distribution of identified eRNAs among ncRNAs annotated based on TT-seq signal. The outer circle segments show the number of actively transcribed eRNAs (RPK $\geq$ 16.5) at each time point. **(B)** Length distribution of eRNAs and other ncRNAs. The *P*-value was derived by two-sided Mann-Whitney *U* test. Box limits are the first and third quartiles, the band inside the box is the median. The ends of the whiskers extend the box by 1.5 times the interquartile range. **(C)** Half-life distribution of eRNAs and other ncRNAs. The *P*-value was derived by two-sided Mann-Whitney *U* test. Box limits are the first and third quartiles, the band inside the box is the median. The ends of the whiskers extend the box by 1.5 times the interquartile range. **(D)** Average DNase hypersensitivity sites (DHS) signal at the TSS of eRNAs and other ncRNAs. **(E)** Motif enrichment in the 250 bp upstream sequences of eRNA versus other ncRNAs. Displayed are only motifs of upregulated TFs with odds ratio > 1.2 and *P*-value < 0.05. The stars indicate TFs with statistical significant enrichment upon eRNAs after multiple testing correction (Benjamini-Hochberg method, FDR < 0.05).

14 Transcription from promoters and enhancers is correlated and distance-dependent

We next paired eRNAs and mRNAs that localized within insulated neighborhoods (Figure 11A) that were defined with a combination of cohesin ChIA-PET and CTCF ChIP-seq profiling performed in the Jurkat cell line (Hnisz et al. 2016b). After removing pairs of upstream divergent (1 kb upstream of sense TSS) and conver-

gent (1 kb downstream of sense TSS) transcripts and their bidirectional promoters, we obtained a total of 6,896 eRNA-mRNA pairs that represent putative enhancer-promoter pairs. These pairs contained 2,454 transcribed enhancers and 2,520 promoters. On average there were 1.3 transcribed enhancers and 1.5 promoters located within an insulated region (Figure 12A). The median transcribed enhancer-promoter distance within these pairs was 117 kb, with 52% of all paired transcribed enhancers residing within +/- 50 kb from their closest paired promoter (Figure 11B). There was no preference for eRNA orientation with respect to the mRNA orientation (Figure 12B). Due to the small size of the insulated neighborhoods and our conservative pairing, most transcribed enhancers (56%) and most promoters (72%) remained unpaired. The paired transcribed enhancers engaged on average with 2.8 promoters, whereas paired promoters engaged on average with 2.7 enhancers (Figure 11C).

We found that changes in RNA synthesis over time correlated very well between transcribed enhancers and their paired promoters (Figure 11D, Section 9.9). When we shuffled the transcribed enhancers and promoters and paired them randomly, irrespective of insulated neighborhoods, the correlation dropped (Figure 12C, P-value=9.99e-4). Moreover, the correlation was higher for transcribed enhancers located less than 10 kb from their paired promoter ('proximal enhancers') than for those located further apart ('distal enhancers') (Figure 11D). This indicates that enhancer transcription decreases with increasing distance from the activated target promoter, consistent with the observation that interacting enhancers tend to be close to their promoters (Dekker et al. 2013; He et al. 2014).

The distance between transcribed enhancer-promoter pairs is limited by the size of the insulated neighborhoods, but it is generally much shorter (P-value $<$ 2.2e-16, Figures 12D and E). Pairing within insulated neighborhoods leads to higher correlations than pairing every promoter with it's closest transcribed enhancer (Appendix Figure A4). There is no relationship between the distance and the correlation over time between closest transcribed enhancer-promoter pairs (Spearman correlation -0.03, Appendix Figure A5). When we splitted up the closest pairs dependent on their location within the same insulated neighborhood, the pairs within the same loop showed a higher correlation (P-value=0.00121, Appendix Figure A6).

These results indicate that the correlation in changes of RNA synthesis from transcribed enhancers and promoters depends on both genomic distance and location within insulated neighborhoods.

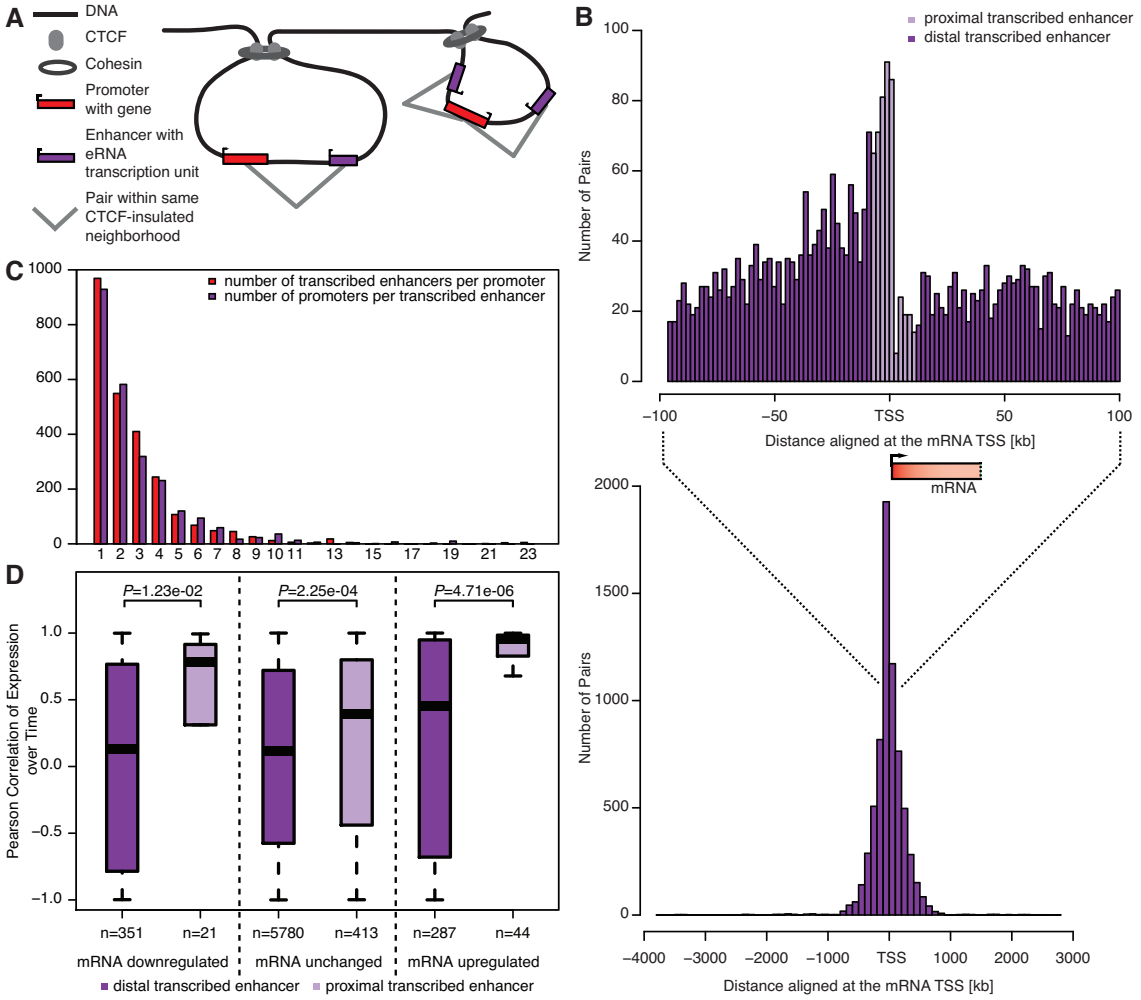


Figure 11: Pairings of transcribed enhancers with promoters. **(A)** Schematic of transcribed enhancer-promoter pairing based on CTCF-insulated neighborhoods. **(B)** Distance distribution between eRNA and mRNA TSS. The lower histogram depicts the full distance range in 100 kb steps. The upper histogram shows a zoom-in of the region [TSS-100 kb, TSS+100 kb] in 2 kb steps. The position of the paired mRNA is indicated together with its median length. **(C)** Number of transcribed enhancers per paired promoter and promoters per paired transcribed enhancer. **(D)** Correlation of TT-seq signal over time between proximal (left, dark violet) or distal (right, light violet) transcribed enhancers and promoters by change in promoter TT-seq signal (from left to right: downregulated, unchanged, upregulated promoters). The Pearson correlation coefficient was calculated between read counts across the time series (replicates averaged per time point) for each transcribed enhancer-promoter pair. The *P*-values were derived by two-sided Mann-Whitney *U* tests. Box limits are the first and third quartiles, the band inside the box is the median. The ends of the whiskers extend the box by 1.5 times the interquartile range.

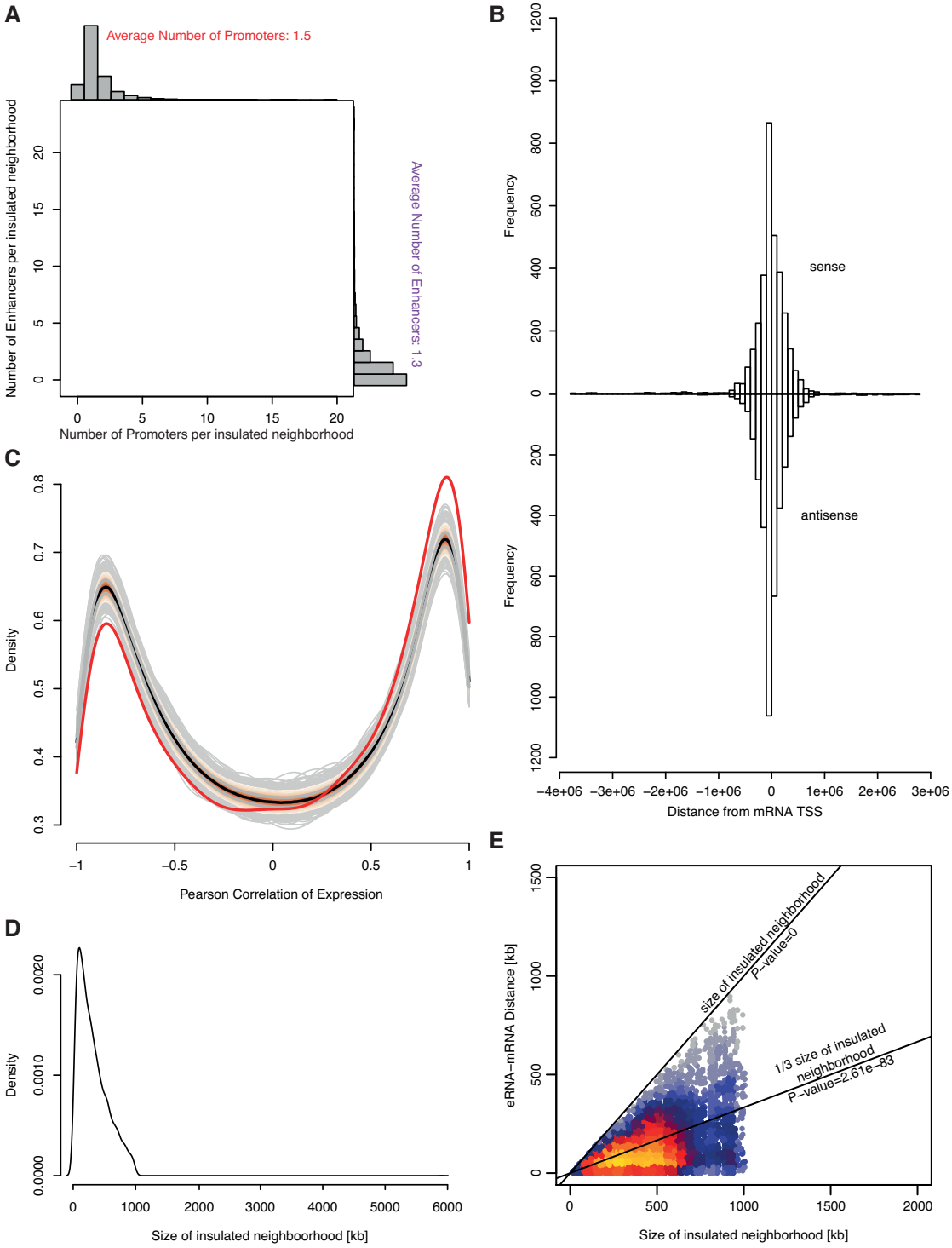


Figure 12: Details on transcribed enhancer-promoter pairs. **(A)** Number of transcribed enhancers and transcribed promoters per ChIA-PET defined insulated neighborhood. Count values were jittered for visualization purposes. **(B)** Location and orientation of transcribed enhancers with respect to their paired target gene TSS. Negative values indicate the transcribed enhancer location upstream of the promoter TSS, positive values downstream of the promoter TSS. Upper histogram: Distance distribution for transcribed enhancers on the same ('sense') strand as their target promoter. Lower histogram: Distance distribution for transcribed enhancers on the opposite, antisense strand. *Figure legend continued on next page.*

Figure 12: (C) Distribution of the Pearson correlation between observed TT-seq signal at transcribed enhancers and promoters for the enhancer-promoter pairs derived here (where both eRNA and mRNA change significantly between 0 min and 15 min timepoints; red line) and for 1,000 randomly shuffled enhancer-promoter pairs (grey lines). The colored profile indicates the quantiles (5-95% of the data) with the black median line and the grey 25% and 75% quantile lines. Observed correlations are enriched for positive correlations (right peak) and depleted for negative correlations (left peak). **(D)** Distribution of insulated neighborhood size by Cohesin-ChIA-PET and CTCF-ChIP-seq. The median size of an insulated neighborhood was 255 kb. **(E)** Distance of transcribed enhancer-promoter pairs versus the size of the corresponding insulated neighborhood. The lines indicate the size of the insulated neighborhood (by which the distance is bound by definition), and a third of the size of the insulated neighborhood (which is the expected distance between tow randomly drawn positions in a neighborhood). The P-value was derived by a one-sided Mann-Whitney U test. For visualization purposes, 127 points with insulated neighborhood size $>$ 2,000 kb are not shown.

15 Rapid up- and down-regulation via promoter-proximal elements

Our results raise the question how transcription can be activated from promoters without paired enhancers. It is known that promoters may contain proximal binding sites for transcriptional activators such as AP-1, a heterodimer of FOS and JUN proteins, that is induced upon T-cell stimulation. Indeed we found that upregulated but unpaired promoters were enriched for AP-1 binding sites (TGACTCA) in the promoter-proximal region 500-100 bp upstream of the TSS for mRNA transcription, compared to upregulated and paired mRNAs (odds ratio 2.24, P-value 0.031, Section 9.8). This shows that TT-seq can be used to disentangle promoter-based from enhancer-based activation of gene expression.

TT-seq also revealed a large number of downregulated genes upon T-cell stimulation, as captured by ceasing RNA synthesis. Such rapid downregulation cannot be observed by RNA-seq, due to the stability of most mRNAs. When we investigated sequence motifs for unpaired downregulated mRNAs compared to unpaired upregulated mRNAs, we did not find enriched motifs that are known to bind transcriptional repressors, but consistently found them to be depleted of binding sites for the transcriptional activator AP-1 in the region 500-100 bp upstream of the TSS (odds ratio 0.51, P-value 0.022, Section 9.8). Together these observations show that TT-seq is ideally suited to detect down-regulated genes and are consistent with the view that rapid gene regulation can be mediated by the promoter-proximal region.

16 Transcription from enhancers and promoters occurs simultaneously

We next investigated whether there are temporal differences in the onset of transcriptional changes between enhancers and promoters. In particular, we wished to find out whether enhancers were transcribed before their paired promoters. To this end, we selected pairs where the transcriptional change 15 min after stimulation for both the transcribed enhancer and the promoter in the TT-seq samples was at least two-fold increased ('up-regulated pairs') or two-fold decreased ('down-regulated pairs') and significant (FDR < 0.05). This selection ensures that both the promoters and the transcribed enhancer have been activated during the time course allowing to probe the relative timing of activation. The TT-seq data clearly showed that changes in RNA synthesis occurred simultaneously at paired transcribed enhancers and promoters, both for up- and down-regulated pairs, at the temporal resolution of our data and within a given variation (Figure 13). This shows that for an immediate transcription response the changes in RNA synthesis for enhancers and their paired promoters occur simultaneously, provided our current temporal resolution (Figures 13A and C).

In contrast, our RNA-seq data suggested that an increase in enhancer transcription preceded mRNA transcription for up-regulated pairs (Figures 13B and D). However, this does not mean that eRNA synthesis changes more rapidly than mRNA synthesis. Instead, the half-life of eRNAs is around two orders of magnitude shorter than that of mRNAs (Rabani et al. 2014; Schwalb et al. 2016), and this renders eRNA levels very sensitive to changes in their synthesis (Figure 7). Also, RNA-seq cannot detect changes in down-regulated pairs because mRNAs have long half-lives in the range of hours (Rabani et al. 2014; Schwalb et al. 2016), and therefore, a rapid shut-down in RNA synthesis does not change mRNA levels when monitored within minutes.

Taken together, TT-seq enables monitoring rapid changes in both eRNA and mRNA synthesis that cannot be detected by RNA-seq in an unbiased manner.

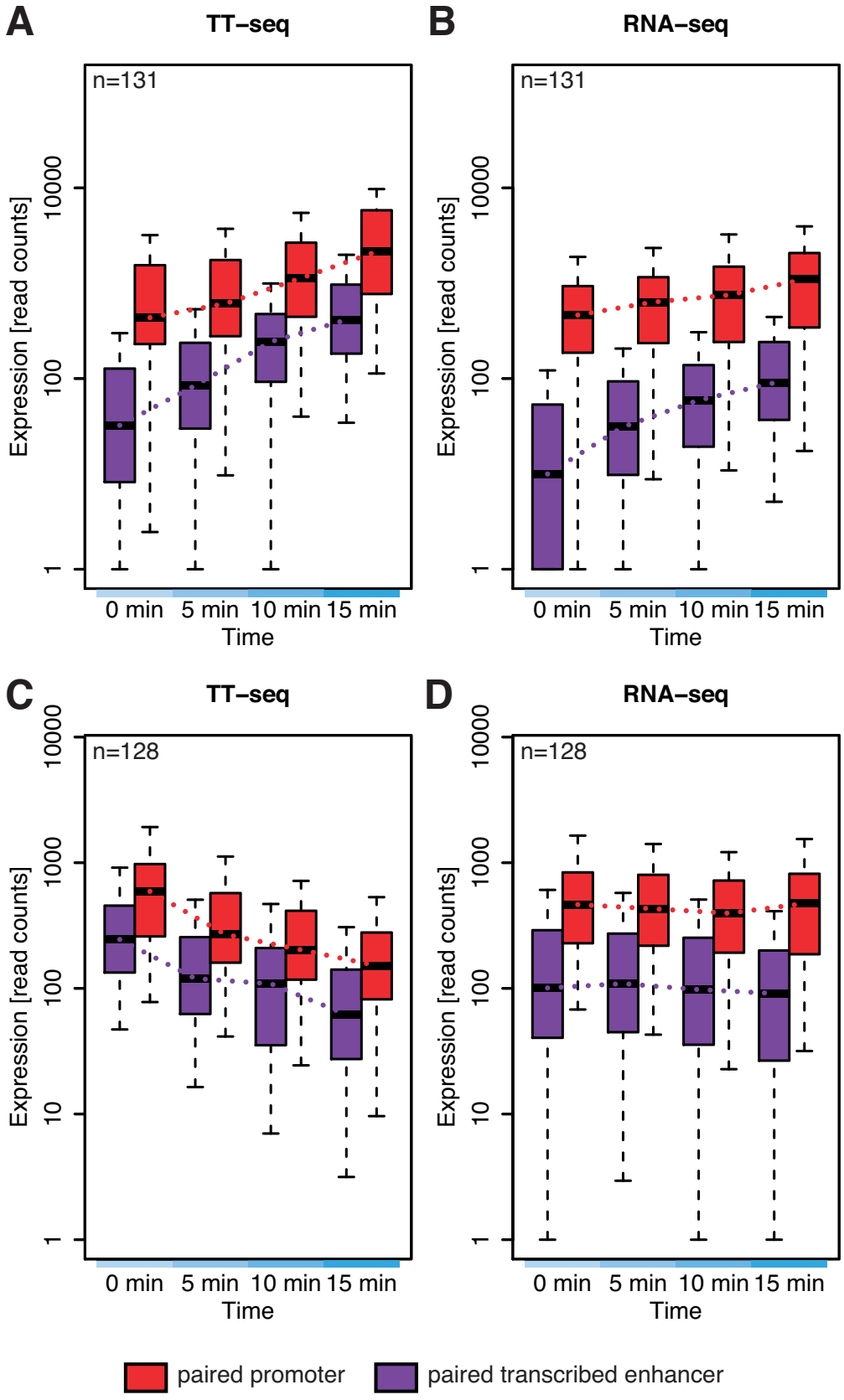


Figure 13: Temporal changes in enhancer and promoter transcription. *Figure legend continued on next page.*

Figure 13: Temporal changes in enhancer and promoter transcription. **(A)** Development of TT-seq signal over time after T-cell stimulation for paired promoters and enhancers (n=131) that are both significantly upregulated (FC $\geq$ 2, FDR $\leq$ 0.05) 15 min after stimulation (over the whole eRNA/ the first 2,200 bp of the mRNA). The y-axis shows the normalized read counts over the whole transcribed enhancer region (violet) and the first 2,200 bp (average length of eRNA) of the paired mRNA (red). The black line indicates the median. **(B)** As in panel A but using RNA-seq read counts. **(C)** TT-seq signal change as in panel A but for paired promoters and enhancers (n=128) that are both significantly downregulated (FC $\leq$ 1/2, FDR $\leq$ 0.05) 15 min after stimulation. **(D)** As in panel C but using RNA-seq read counts. Data information: Box limits are the first and third quartiles, the band inside the box is the median. The ends of the whiskers extend the box by 1.5 times the interquartile range.

17 Discussion

Here, we used TT-seq to monitor a very rapid transcription response in human T-cells, and show that TT-seq can globally detect very short-lived transcripts such as eRNAs in a highly dynamic system at high temporal resolution. We demonstrate that TT-seq is suitable for annotating potential eRNAs and quantifying transcriptional changes very early after stimulation and thus provides insights into gene regulation, activation, and enhancer identity. Our results have implications for understanding the T-cell response, the temporal sequence of enhancer and promoter transcription during gene activation, the nature of functional enhancer-promoter pairing, and the design of future studies of transcription regulation in human cells.

First, our results provide new insights into the immediate T-cell response. TT-seq enabled us to detect immediate changes in the synthesis of thousands of transcripts. These RNAs included most of the transcripts known to be altered during T-cell stimulation, confirming known studies. We also found, however, many new mRNAs and ncRNAs that show altered synthesis upon T-cell stimulation. Many of these have functions in signaling pathways (*PPP1R15A*, *KLF4*, *ARC*), the response to stimuli (*KLF4*, *GPR50*, *DUSP1*, *MASP2*), or have catalytic activities (*DUSP1*, *MASP2*). Our results of immediate transcriptional changes confirm very early single-locus radiolabeled nuclear run-on studies of T-cell activation (Greenberg et al. 1984). Our results thus extend and complement previous genome-wide studies of the T-cell response (Cheadle et al. 2005; Diehn et al. 2002; Ellisen et al. 2001), and help to more generally understand very early transcriptional responses.

Our work also identifies and characterizes eRNAs based on their synthesis, thereby mapping transcribed enhancers. We show that eRNA-producing enhancers can be

paired with their target promoters by taking advantage of previously published datasets on insulated neighborhoods (Hnisz et al. 2016b) and chromatin states (Zacher et al. 2017). This yields enhancer-promoter pairs with highly correlated temporal changes in RNA synthesis. These results are consistent with the idea that eRNA transcription is a very sensitive and good proxy for the activity of enhancers with respect to target gene activation (Hah et al. 2015). One limitation of the method, however, relates to the inability of TT-seq to detect intronic eRNAs in sense direction of mRNA transcription; however, only a small fraction of enhancers is missed this way.

The classification of non-coding RNAs remains challenging. The previous definitions of lincRNAs (long, stable, spliced, polyadenylated) and eRNAs (short, short-lived, transcribed from enhancer element) do not always allow for a clear distinction between them (Paralkar et al. 2016). Here we decided to exclude the GENCODE class of 'lincRNAs' from our eRNA set because these are evolutionary conserved and less likely to be cell type-specific enhancer transcripts. Due to the low number of transcripts overlapping GENCODE-annotated lincRNAs (n=590), we are not excluding many potential eRNAs, although some lincRNAs may stem from enhancers.

Our results also provide evidence that enhancer and promoter transcription can occur simultaneously during immediate gene activation. Our observations are derived from a single biological process with very fast response kinetics. Previous studies have observed that enhancer transcription precedes transcription from promoters, although in some cases, evidence for simultaneous transcription was also obtained (Arner et al. 2015; De Santa et al. 2010; Kaikkonen et al. 2013; Schaukowitch et al. 2014). These differences can to some extent be explained by the high sensitivity and temporal resolution of TT-seq, but may also reflect differences in the cellular responses monitored. Whereas we focused here on the immediate T-cell response that occurs within minutes, published work generally analyzed responses after hours, and these require changes in chromatin at enhancers (Kaikkonen et al. 2013). Changes in chromatin could lead to a time lag between enhancer and promoter transcription and likely do not occur during the immediate response we investigated here because immediate-early genes responding within minutes are poised for gene activation, and chromatin is in a pre-open state (Byun et al. 2009; Tullai et al.

2007). Similarly, enhancers are primed for activity and are DNase I hypersensitive and modified with H3K4me1 (Wang et al. 2015).

Most importantly, our results demonstrate that TT-seq is an easy-to-use tool that is ideally suited to monitor rapid changes in the genomic landscape of transcribed enhancers and gene transcription in a non-perturbing manner *in vivo*. In addition to its high sensitivity and high temporal resolution, TT-seq is uniquely suited to detect immediate down-regulation of genes, as it informs on drops in RNA synthesis when the mRNA product is long-lived and will give a signal in RNA-seq even at time points when transcription has been shut off for a long time already. In addition, TT-seq will map only those enhancers that produce eRNA at a certain time, providing apparently active enhancers rather than a list of all chromatin regions with enhancer signatures that may stem from past enhancer transcription events. TT-seq therefore facilitates the pairing of enhancers with putative target promoters. In the future, the application of TT-seq to other human cells, signaling and differentiation events, is expected to provide novel biological insights into fundamental changes in gene regulatory programs.

Part IV

Further Contributions

In this section, I present additional work from collaborations that has already been published. All of these studies are based on TT-seq and 4tU-seq experiments. They are briefly explained and summarized in the following.

18 TT-seq measures transcription rates for transient RNAs

The results presented in this section have been published in:

TT-seq maps the human transient transcriptome

B. Schwalb*, M. Michel*, B. Zacher*, K. Frühauf, **C. Demel**, A. Tresch, J. Gagneur, and P. Cramer

Science (2016), 352(6290), 1225-1228.

The full article with supplementary materials can be found at `http://science.sciencemag.org/content/352/6290/1225`.

Contribution: *I established the spike-in normalization method (presented in Section 8), which was first applied in this study.*

18.1 Introduction

Transcription of eukaryotic genomes produces protein-coding mRNAs and many non-coding RNAs (ncRNAs), including enhancer RNAs (eRNAs) that stem from regulatory elements (Andersson et al. 2014; Jensen et al. 2013). Most ncRNAs are rapidly degraded, difficult to detect, and generally impossible to map in their full length. Comprehensive and complete mapping of such transient RNAs, however, is required to understand how genomes are regulated and how RNA fate is controlled. We developed transient transcriptome sequencing (TT-seq), a protocol that uniformly maps the entire range of RNA-producing units and estimates rates of RNA synthesis and degradation.

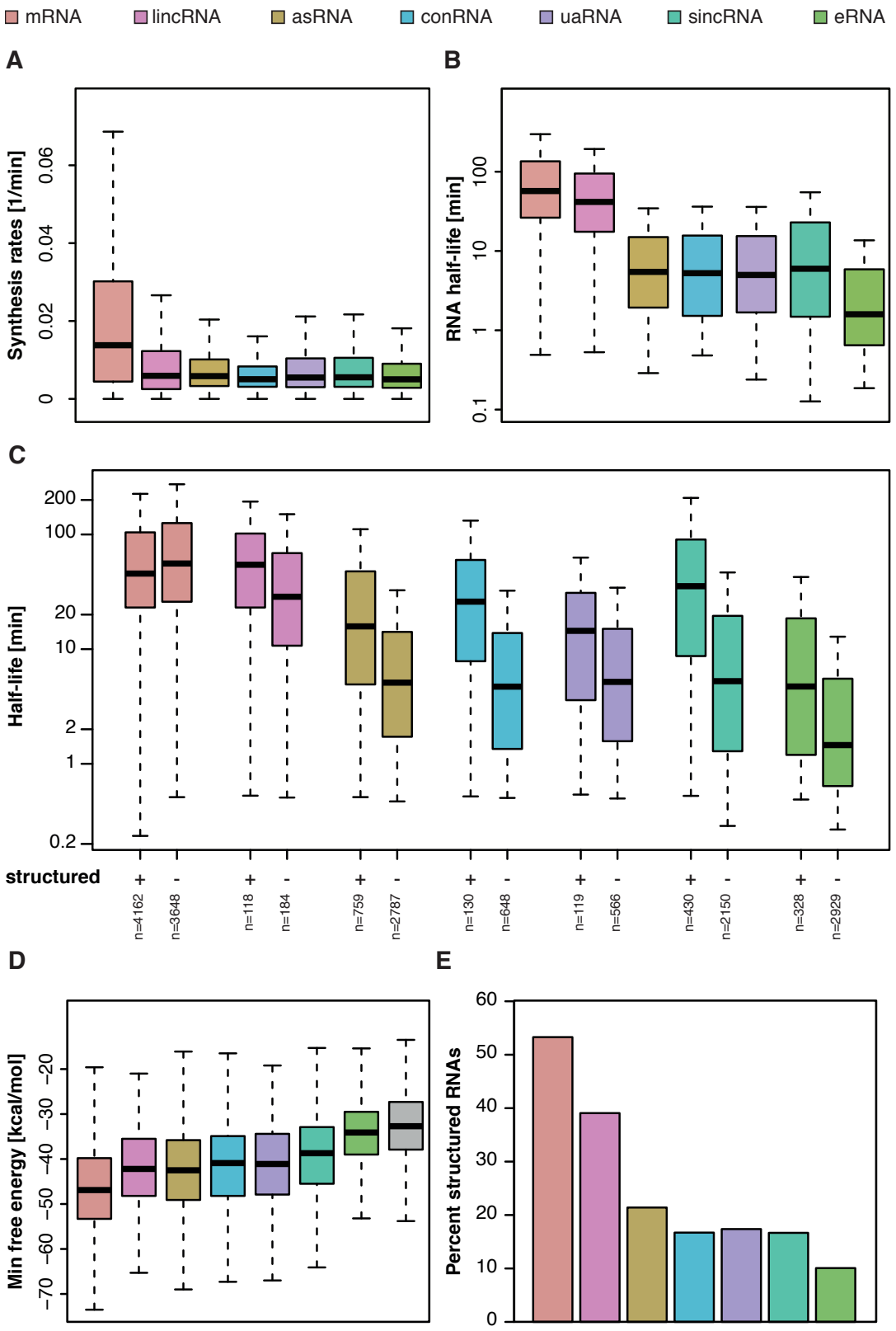


Figure 14: Estimated synthesis rates, half-lives, and predicted structure. Adapted from (Schwalb et al. 2016). Reprinted with permisson from AAAS. *Figure legend on continued on page.*

Figure 14: Estimated synthesis rates, half-lives, and predicted structure. **(A)** Distribution of synthesis rates per transcript class. Black bars represent the median, boxes represent upper and lower quartile, and whiskers represent 1.5 times the interquartile range. **(B)** Estimated RNA half-lives for different transcript classes. **(C)** Distribution of half-lives of different transcript classes depending on whether they are predicted to be structured or not (+, -; (Washietl et al. 2005)). **(D)** Distribution of the minimum free energy in the first 1000 nucleotides per transcript class. **(E)** Distribution of percentage of structured RNA in different transcript classes.

18.2 Results

Kinetic modeling of TT-seq and RNA-seq data (see also Section 8) enabled us to estimate rates of RNA synthesis and degradation (Figures 14A and B). We estimated rates of phosphodiester bond formation or breakage at each transcribed position and averaged these within TUs, thus obtaining estimates of relative transcription rates and RNA stabilities. We found that mRNAs and lncRNAs had the highest synthesis rates and longest half-lives. We determined a median mRNA half-life of ~50 min, compared to a previous estimate of ~139 min (Rabani et al. 2014). Other transcript classes had low synthesis rates and short half-lives, explaining why short ncRNAs are difficult to detect. eRNAs had half-lives of a few minutes, consistent with prior data (Rabani et al. 2014).

Short RNA half-lives correlated with a lack of secondary structure (Figure 14C). The folding energy of eRNAs was comparable to the genomic background level (Figure 14D), and only 10% of their sequence was predicted to be structured, compared with 52% in mRNAs (Figure 14E).

18.3 Summary

Application of TT-seq to human K562 cells recovers stable mRNAs and long intergenic non-coding RNAs, and additionally maps over 10,000 transient RNAs, including enhancer RNAs, antisense RNAs, and promoter-associated RNAs, both convergent and upstream antisense RNAs. TT-seq analysis shows that enhancer RNAs are short-lived and lack secondary structure.

19 Quantification of *Schizosaccharomyces pombe* RNA metabolism

The results presented in this section have been published in:

Determinants of RNA metabolism in the *Schizosaccharomyces pombe* genome
P. Eser*, L. Wachutka*, K.C. Maier, **C. Demel**, M. Boroni, S. Iyer, P. Cramer, and J. Gagneur
Molecular Systems Biology (2016), 12(2), 857.

The full article with supplementary materials can be found at `http://msb.embopress.org/content/12/2/857`.

Contribution: *I conducted initial analyses and provided count matrices for all relevant features, namely exons, introns, exon-intron junctions, and intron-exon junctions, that are the basis for fitting the kinetic model.*

19.1 Introduction

Gene expression can be regulated at each stage of RNA metabolism, during RNA synthesis, splicing, and degradation. The rates of both RNA degradation and splicing contribute to the time required for reaching mature RNA steady-state levels following transcriptional responses (Jeffares et al. 2008; Rabani et al. 2014).

The fission yeast *Schizosaccharomyces pombe (S. pombe)* is an attractive model organism to study eukaryotic RNA metabolism. *S. pombe* shares important gene expression mechanisms with higher eukaryotes, including splicing. To cover the typical range of synthesis, splicing, and degradation rates, cells in a steady-state culture were harvested after 2, 4, 6, 8, and 10 min following 4tU addition. Moreover, a matching total RNA-seq was performed after 10 min labeling to control for the slower doubling time in the presence of 4tU (285 min versus 180 min).

19.2 Results

The data contained many reads that stemmed from intronic sequences and reads comprising exon–intron junctions, showing that 4tU-seq captured short-lived precursor RNA transcripts. These reads from unspliced RNA gradually ceased during the time course (Figure 15A and B), indicating that the kinetics of RNA splicing

may be inferred from the data.

To globally estimate rates of RNA synthesis, splicing, and degradation, we used a first-order kinetic model with constant rates that describes the amount of labeled RNA as a function of time (Figure 15C). We modeled splicing of individual introns, where splicing refers to the overall process of removing the intron and joining the two flanking exons. The model was fit to every splice junction using the counts of spliced and unspliced junction reads (Figure 15C and D). We included in the model scaling factors that account for variations in sequencing depth, an overall increase of the labeled RNA fraction, cross-contamination of unlabeled RNA, and 4tU label incorporation efficiency. The model was fitted using maximum likelihood and assuming negative binomial distribution to cope with overdispersion of read counts (Anders et al. 2010; Robinson et al. 2010). Our method yields absolute splicing and degradation rates, but provides synthesis rates up to a scaling factor common to all TUs. Absolute synthesis rates were obtained by scaling all values so that the median steady-state level of ORF-TUs matches the known median of 2.4 mRNAs per cell (Marguerat et al. 2012). To facilitate comparisons of the obtained RNA metabolic rates, we present the synthesis rate as the average time elapsed between the production of two transcripts in a single cell ('synthesis time'), the degradation rate as the time needed to degrade half of the mature RNAs ('half-life'), and the splicing rate as the time to process half of the precursor RNA junction ('splicing time'). The synthesis times and half-lives inferred from distinct splice junctions of the same TU agreed well, demonstrating the robustness of our approach (Spearman rank correlation = 0.44 for synthesis time, $P < 2 \times 10^{-16}$ and Spearman rank correlation = 0.79 for half- life, $P < 2 \times 10^{-16}$, Figure 15E)

19.3 Summary

By combining metabolically labeled RNA profiling at high temporal resolution with computational kinetic modeling, we obtained *in vivo* RNA synthesis, splicing, and degradation rates across an entire eukaryotic genome, providing insights into RNA metabolism.

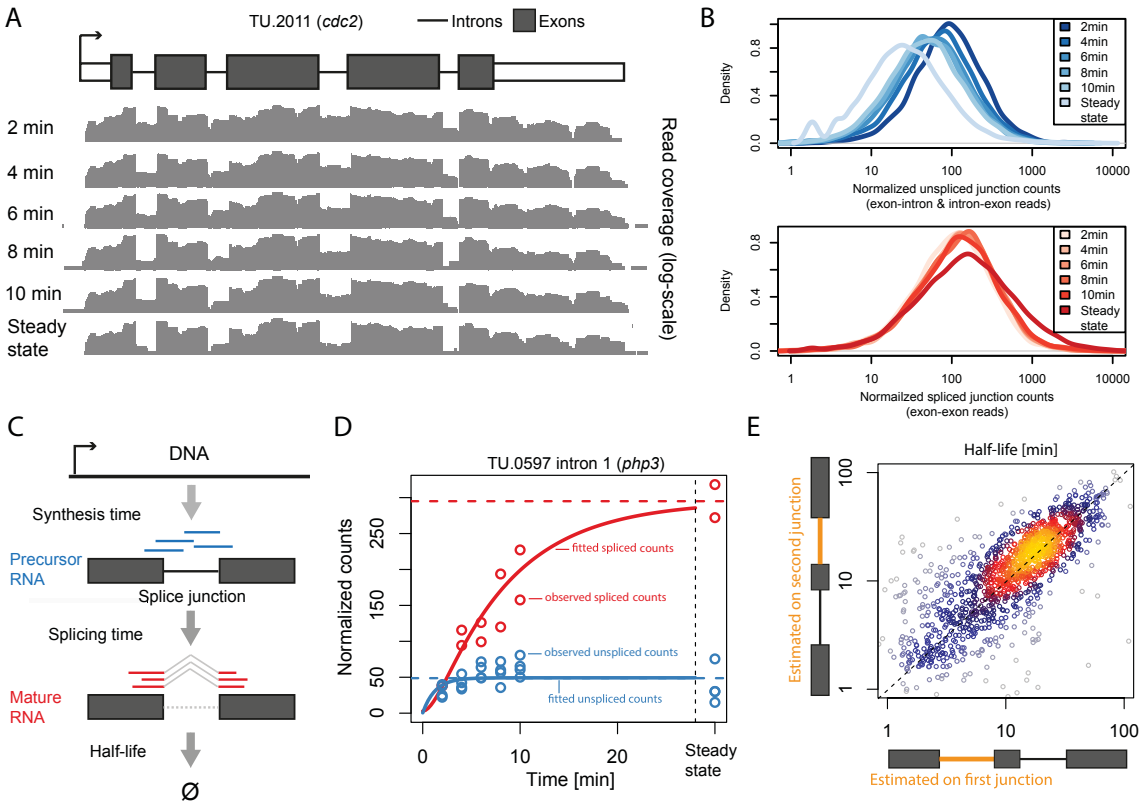


Figure 15: Estimating RNA processing rates using labeled RNA time series. **(A)** Per-base coverage (gray tracks) in a logarithmic scale of 4tU-seq samples at 2, 4, 6, 8, and 10 min. labeling and for one RNA-seq sample (i.e., steady state) along the UTRs (white boxes), the exons (dark boxes), and the introns (lines) of the TU encoding *cdc2*. **(B)** Distribution of sequencing depth normalized unspliced junction read counts (top panel) and normalized spliced junction read counts (lower panel) for the complete 4tU-seq time series and the steady-state RNA-seq samples. **(C)** Schema of the junction first-order kinetics model. Each splice junction is modeled individually, assuming constant synthesis time, splicing time and half-life. Unspliced junction reads (blue) are specific to the precursor RNA and spliced junction reads (red) are specific to the mature RNA. **(D)** Observed (circles) and fitted (lines) splice junction counts for the first intron of TU.0597 (*php3*). Unspliced (blue) and spliced (red) normalized counts (y-axis) are shown for all 4tU-seq samples and the steady-state sample (x-axis). **(E)** Half-life estimated from the first (x-axis) versus the second (y-axis) splice junction on TUs with two or more introns.

20 Spt5 is required for a normal rate of RNA synthesis

The results presented in this section have been published in:

Spt5 plays vital roles in the control of sense and antisense transcription elongation
A. Shetty*, S.P. Kallgren*, **C. Demel**, K.C. Maier, D. Spatt, B.H. Alver, P. Cramer, P.J. Park, and F. Winston
Molecular Cell (2017), 66, 1-12.

The full article with supplementary materials can be found at `http://www.cell.com/molecular-cell/fulltext/S1097-2765(17)30160-0`. *Relevant methods for the results presented here are given in Section* 26.

Contribution: *I analyzed all 4tU-seq data generated for this project.*

20.1 Introduction

In eukaryotes, Spt5 is an integral and essential part of the Pol II elongation complex (Mayer et al. 2010; Rahl et al. 2010). While Spt5 has been extensively studied, surprisingly little has been done to test its role as a positive transcription elongation factor genome-wide. To address the genome-wide role of Spt5 in transcription, we have comprehensively analyzed the effects of Spt5 depletion on transcription genome-wide in the model organism *Schizosaccharomyces pombe (S. pombe)*. We constructed a *S. pombe* strain that allows for efficient, auxin-inducible degradation of Spt5. Using this system, we were able to efficiently deplete Spt5 genome-wide, while maintaining cell viability. We then measured the level of Pol II across the *S. pombe* genome using ChIP-seq and NET-seq, comparing cells before and after Spt5 depletion. Our results showed that after Spt5 depletion, there was a globally reduced level of Pol II across transcribed regions. Importantly, the distribution of Pol II across genes also changed, with an accumulation over the first ~500 bp of genes, followed by a decreased level of Pol II downstream.

20.2 Results

To gain greater insight into the changes in elongation when Spt5 is depleted, we monitored cellular RNA synthesis before and after Spt5 depletion by metabolic labeling

with 4tU (Miller et al. 2011) and normalization with RNA spike-in probes (Section 8). The 4tU-seq results showed greatly decreased RNA synthesis rates genome-wide in the Spt5-depleted cells, consistent with a general elongation-stimulatory activity for Spt5 (Figure 16A). We observed a uniform decrease in RNA synthesis activity across transcription units (Figure 16B) and antisense of transcription units (Figure 16C), suggesting that the overall rate of transcription is decreased when Spt5 is depleted. Taken together with our ChIP-seq and NET-seq results, these findings suggest that Spt5 is required for a normal rate of transcription by Pol II in order to elongate past a site or barrier at a position within 500 bp from the transcription start site (TSS), possibly a nucleosome or an Spt5-dependent transcription checkpoint (Hartzog et al. 1998; Lidschreiber et al. 2013; Viladevall et al. 2009).

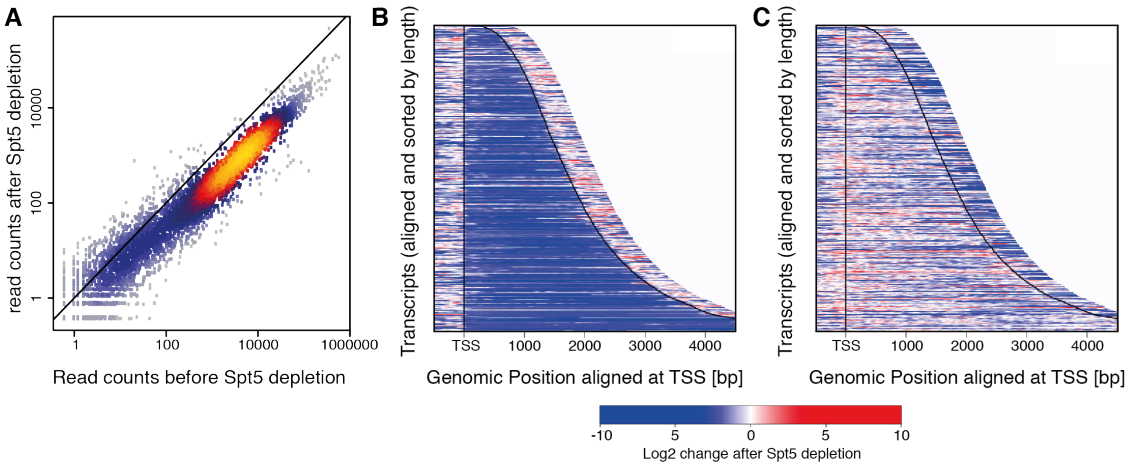


Figure 16: Spt5 is required for a normal rate of transcription genome-wide. **(A)** The scatterplot shows new RNA synthesis measured by 4tU-seq in Spt5-depleted versus non-depleted cells. Each point corresponds to the spike-in normalized signal from one transcript for merged replicate experiments. **(B)** The heatmap shows the log2 ratio of the spike-in normalized 4tU-seq signal in Spt5-depeleted versus non-depleted cells. Genes are sorted by length and aligned by their TSS. The TSS and pA-site are indicated by the solid black lines. **(C)** The heatmap depicts the 4tU-seq log2-fold change in signal obtained for the synthesis of antisense transcripts across transcribed regions in Spt5-depleted compared to non-depleted cells, analogous to (B).

20.3 Summary

Spt5 is an essential and conserved factor that functions in transcription and co-transcriptional processes. However, many aspects of the requirement for Spt5 in transcription are poorly understood. We have analyzed the consequences of Spt5

depletion in *Schizosaccharomyces pombe* using four genome-wide approaches. Our results demonstrate that Spt5 is crucial for a normal rate of RNA synthesis and distribution of Pol II over transcription units. In the absence of Spt5, Pol II localization changes dramatically, with reduced levels and a relative accumulation over the first ~500 bp, suggesting that Spt5 is required for transcription past a barrier. Spt5 depletion also results in widespread antisense transcription initiating within this barrier region. Deletions of this region alter the distribution of Pol II on the sense strand, suggesting that the barrier observed after Spt5 depletion is normally a site at which Spt5 stimulates elongation. Our results reveal a global requirement for Spt5 in transcription elongation.

Part V

Future Perspectives

The newly developed TT-seq protocol is a highly-sensitive method to investigate the transcription of transient RNAs. These technological and experimental advances pose new challenges to the field of computational biology, both in terms of mathematical models and analysis of this data.

To infer correct synthesis and degradation rates, the relative ratio of 4sU/4tU-labeled RNA to the total RNA in the sample has to be determined. The method presented in this thesis provides the theoretical basis to normalize 4sU-labeled and total RNA-seq samples relative to each other. This is accomplished with a GLM that is fit to read counts from artificial spike-ins, which can be labeled *in vitro*. Additionally, the use of spike-ins allows to observe global expression changes and to adjust for them.

The normalization scheme can be applied to genome-wide TT-seq and RNA-seq data sets and allows to infer synthesis and degradation rates for all transcription products at base-pair resolution.

The sensitivity of TT-seq facilitates the observation of newly synthesized transcription products. This led to the discovery of many unannotated transcripts (Michel et al. 2017; Schwalb et al. 2016). With the help of additional datasets, cell line-specific enhancer RNAs (eRNAs) can be identified and paired with their promoters (Michel et al. 2017). Knowing enhancers and their expression patterns is crucial to understand cellular regulatory mechanisms.

21 Extensions for the mathematical model

The model presented in Section 8 could be extended to also account for dynamic changes in the total RNA levels in the cell. In cases where the steady-state assumption does not hold true, synthesis and degradation rates are not constant and need to be estimated independently for time points in a dynamic time course experiment. One approximation could be made in the case of very long time courses. Then, it

can be assumed that for a few distinct time points in a long time course, synthesis and degradation rates within the short (5 minute) labeling time are constant. Therefore, the model presented in this thesis can be applied to estimate time-specific synthesis and degradation rates. Otherwise, time-specific synthesis and degradation rates need to be estimated. The model can also be extended to include the doubling time of cells, which affects the RNA levels in the cell.

22 Biological applications

The TT-seq method and its sensitivity to detect transient RNAs makes it an ideal tool to calculate genome-wide rates of transcription kinetics in an unperturbed manner. The presented normalization scheme allows the comparison of different biological samples on a global scale. Therefore, TT-seq and the presented mathematical method to calculate synthesis rates and half-lives can be used to address various biological questions. I will discuss some of them here in detail:

22.1 Human splicing rates

Splicing of introns contributes to the complexity of eukaryotic organisms. In human, ~95% of genes undergo alternative splicing (Pan et al. 2008). The short half-live of introns complicates measuring RNA splicing rates. *In vitro*, transcripts are spliced within 15-60 min after their synthesis (Das et al. 2006). Single-molecule assays have shown that splicing rates *in vivo* are much faster. Estimated splicing rates range from 30s to 10 min (Carrillo Oesterreich et al. 2011). They have been determined by various assays, such as direct visualization by electron microscopy (Beyer et al. 1988), reverse transcription with PCR amplification (RT-PCR) (Kessler et al. 1993), observing flourescence of single pre-mRNAs by microscopy (Martin et al. 2013), or transcription arrest followed by quantitative RT-PCR (Singh et al. 2009). Most of the approaches used for splicing rate estimation, however, are conducted *in vitro* or are limited to single, long genes. The application of TT-seq can result in genome-wide, *in vivo* splicing rates, independent of gene lengths. Short, progressive RNA labeling followed by TT-seq can be used to determine robust splicing rates for

individual junctions as in *S. pombe* (Eser et al. 2016). The estimation of degradation rates per phosphodiester bond allows to individually determine splicing rates for donor and acceptor sites. Varying rates for donor and acceptor splice sites can reveal alternative splicing events (Wachutka et al. 2016). The obtained splicing rates could be correlated to the position of the intron within the gene and with nearby sequence motifs (Eser et al. 2016).

It is also possible to quantify the effects of specific splicing inhibitors, such as pladienolide (Pla-B) (Kotake et al. 2007). The additional use of artifical spike-ins and application of the spike-in normalization presented in this thesis allows to compare the splicing effects of inhibitor-treated cells relative to wild-type cells.

22.2 Cellular differentiation

Cellular differentiation is a fundamental process in life. All cells in an organism have the same underlying genotype, but it can be expressed in various phenotypes. This differential gene expression is the result of highly special regulatory programs. As enhancers control tissue-specific gene expression (Visel et al. 2009), they could be crucial regulators of differentiation. Mapping eRNAs with TT-seq along a differentiation or trans-differentiation time course could shed light on the differential use of enhancer elements that govern tissue-specific gene expression. The usage of H3K4me1 ChIP-seq data and additional data to map open chromatin (DHS, ATAC-seq) could help to identify primed and active enhancer regions at different stages of differentiation. TT-seq could establish the link between active enhancer transcription and the spatio-temporal pattern of gene expression during cellular differentiation. The timing of events, especially the question if nucleosome rearrangement has to happen before histone modifications and transcription take place, is of particular interest.

22.3 Cancer transcriptomics

Enhancer RNAs (eRNAs) play an important role in the regulation of immediate early genes (IEGs), and therefore they also have an effect on disease and cancer development (Bahrami et al. 2016). TT-seq can reveal active enhancers in various

cancer cell lines by mapping transient RNAs. Additionally, the RNA synthesis and degradation rates of cancer cells can be quantified and compared to healthy control cells. Therefore, TT-seq can also provide novel and medically relevant insights into the RNA metabolism of different cancer cell lines.

23 Concluding remarks

TT-seq allows to measure RNA metabolism kinetics on a genome-wide scale *in vivo*. The sensitivity of TT-seq to map transient RNAs offers new opportunities to gain insights in dynamic changes in gene regulatory programs. The application of TT-seq, together with the proposed normalization strategy, can provide genome-wide rates of RNA metabolism and reveal global changes in gene expression. The ability of TT-seq to map only actively transcribed enhancers will provide novel insights into gene regulation in various biological questions, such as cellular differentiation, cancer development, or X chromosome inactivation.

Part VI

Appendix

24 Materials and Methods for Section III

24.1 Spike-in sequences

The following ERCC spike-ins were used to normalize the TT-seq samples (Section 8.1).

Spike-in 2

AATACCTTTACAAATGCTTTAACAAGAGGAAATTGTGTTTTTGCCAATTTAAGACCTAATTTAATAGTTAAACCATTAA-
CCTTAGTTGTTCCAAGGCATAATATAGAGAGTGAGATACAGGATGAGCTATTTCAGGGAGTTATTCAGTATGCAGTTG-
CCAAGGCAGTTGCTGATTTAGATTTAGATGAAGATTTAAAGGTTGTTGTCTCTGTTAATGTCCCAGAGGTTCCAATAAC-
CAATTTAAATAAAAGAAAACTCTTCCAATACTTCTATGCCTCAGCAAAGTTAGCTATAAACAGAGCTTTAAATGAATAT-
CCTTCAAAAGAGAAGGTAAAGAAAGAGAAATATAGAGCTTTGCATCCATTAGTTGGATTTAGGGATGTTAGATTGGAG-
TATCCTCCATATCTACAAATTGCTTTGGATGTCCCAACTATGGAGAATTTGGAATTTTTGTTACAAACAATTCCAAATA-
GCGACCACATCATCTTAGAGGCTGGAACACCACTAATTAAAAAGTTTGGTTTAGAGGTTATTGAAATAATGAGAGAAT-
ATTTTGATGGCTTTATTGTTGCTGATTTAAAAACCTTAGACACTGGAAGGGTTGAGGTAAGATTGGCATTTGAAGCAA-
CAGCTAATGCAGTGGCAATAAGTGGAGTAGCACCAAAATCAACAATAATTAAAGCTATCCACGAATGTCAAAAATGTG-
GTTTAATCAGCTATTTGGATATGATGAACGTCTCTGAACCTCAAAAATTATATGATTCATTAAAATTAAAGCCAGATGT-
TGTTATCTTGCATAGAGGGATTGATGAGGAGACATTTGGAATTAAAAAGGAATGGAAATTTAAGGAAAACTGCTTATT-
AGCAATTGCTGGAGGAGTTGGTGTGGAGAATGTTGAAGAGCTTTTAAAAGAATATCAAATATTAATCGTTGGTAGAGC-
AATTACAAAATCAAAAGACCCAGGAAGAGTAATTAGGATTTTATAAACAAGATGGGTTAAAAAAAAAAAAAAAAAAAA-
AAAA

Spike-in 4

TTTCGACGTTTTGAAGGAGGGTTTTAAGTAATGATCGAGATTGAAAAACCAAAAATCGAAACGGTTGAAATCAGCGAC-
GATGCCGAATTTGGTAAGTTTGTCGTAGAGCCACTTGAGCGTGGATATGGTACAACTCTGGGTAACTCCTTACGTCGT-
ATCCTCTTATCCTCACTCCCTGGTGCCGCTGTAACATCAATCCAGATAGATGGTGTACTGCACGAATTCTCGACAATTG-
AAGGCGTTGTGGAAGATGTTACAACGATTATCTTACACATTAAAAAGCTTGCATTGAAAATCTACTCTGATGAAGAGAA-
GACGCTAGAAATTGATGTACAGGGTGAAGGAACTGTAACGGCAGCTGATATTACACACGATAGTGATGTAGAGATCTT-
AAATCCTGATCTTCATATCGCGACTCTTGGTGAGAATGCGAGTTTCCGAGTTCGCCTTACTGCTCAAAGAGGACGTGGG-
TATACGCCTGCTGACGCAAACAAGAGAGGCGATCAGCCAATCGGCGTGATTCCGATCGATTCTATCTATACGCCAGTTT-
CCCGTGTATCTTATCAGGTAGAGAACACTCGTGTAGGCCAAGTTGCAAACTATGATAAACTTACACTTGATGTTTGGAC-
TGATGGAAGCACTGGACCGAAAGAAGCAATTGCGCTTGGTTCAAAGATTTTAACTGAACACCTTAATATATTCGCTGGT-
TTAACTGACGAAGCTCAACATGCTGAAATCATGGTTGAAGAAGAAGAAGATCAAAAAGAGAAAGTTCTTGAAATGACA-
ATTGAAGAATTGGATCTTTCTGTTCGTTCTTACAACTGCTTAAAGCGTGCGGGTATTAACACGGTTCAAGAGCTTGCGA-
ACAAGACGGAAGAAGATATGATGAAAGTTCGAAATCTAGGACGCAAATCACTTGAAGAAGTGAAAGCGAGACTAGAAG-
AACTTGGACTCGGACTTCGCAAAGACGATTGACTAGTTTCCCTTGTGAACTAGGATTTTCCCGGGTACAAAAAAAAAAA-
AAAAAAAAAAA

Spike-in 5

ACTGTCCTTTCATCCATAAGCGGAGAAAGAGGGAATGACATTGTTCTTACACGGCACAAGCAGACAAAATCAACATGGT-
CATTTAGAAATCGGAGGTGTGGATGCTCTCTATTTAGCGGAGAAATATGGTACACCTCTTTACGTATATGATGTGGCTT-
TAATACGTGAGCGTGCTAAAAGCTTTAAGCAGGCGTTTATTTCTGCAGGGCTGAAAGCACAGGTGGCATATGCGAGCA-
AAGCATTCTCATCAGTCGCAATGATTCAGCTCGCTGAGGAAGAGGGACTTTCTTTAGATGTCGTATCCGGAGGAGAGC-
TATATACGGCTGTTGCAGCAGGCTTTCCGGCAGAACGCATCCACTTTCATGGAAACAATAAGAGCAGGGAAGAACTGC-

GGATGGCGCTTGAGCACCGCATCGGCTGCATTGTGGTGGATAATTTCTATGAAATCGCGCTTCTTGAAGACCTATGTA-
AAGAAACGGGTCACTCCATCGATGTTCTTCTTCGGATCACGCCCGGAGTAGAAGCGCATACGCATGACTACATTACAAC-
GGGCCAGGAAGATTCAAAGTTTGGTTTCGATCTTCATAACGGACAAACTGAACGGGCCATTGAACAAGTATTACAATC-
GGAACACATTCAGCTGCTGGGTGTCCATTGCCATATCGGCTCGCAAATCTTTGATACGGCCGGTTTTGTGTTAGCAGCG-
GAAAAAATCTTCAAAAAACTAGACGAATGGAGAGATTCATATTCATTTGTATCCAAGGTGCTGAATCTTGGAGGAGGT-
TTCGGCATTCGTTATACGGAAGATGATGAACCGCTTCATGCCACTGAATACGTTGAAAAAATTATCGAAGCTGTGAAAG-
AAAATGCTTCCCGTTACGGTTTTGACATTCCGGAAATTTGGATCGAACCGGGCCGTTCTCTCGTGGGAGACGCAGGCA-
CAACTCTTTATACGGTTGGCTCTCAAAAAGAAGTGGATAAGCTGTACAATCGTTTCATCATTCGGCGTGCGAATTAAAA-
AAAAAAAAAAAAAAAAAAAAAA

Spike-in 8

AGATGTATATATGATGTCCTTGGACGGGGTGGCGCAGTATTACTGCAAGAGAGCGGACAGATTAGTGTGTTGGAGCCG-
ACACATCAAAGGTTCGTCCGGGGACCGATCTGCAGCCTACGGGACATTTATCCGTAAAAGCATGGCGCTGTTTCGTACT-
TATCGGAGGCCAGGTATCGTCGCGGCGAGTCTCCCCGACGACGGAGATGGGCGTTACTATCTGGGCCGTCTCGTACTC-
TGTTACTTGGCACAGATGCGAGCCCTCGTAATGTGCATCAGCTAAGGGCGATATTATAATGCGACGTTTGTACGGATT-
CGTTACTAACGTGTTGGACGCTAGTGGAATATGTGTCGTTGGTTAGCCTACCCATGGCTTTCGCGGCGACACATGCTTA-
GACTCTTTCAAAACTTCGGTGAAGTTCACTCAAGCCGCGGAGCGCCGTCGTAATTCACTAGGGATGGCGGTACCCGTG-
CCCGTCCGATTCGTAGCAACCTGCATCACGATTTTGTCTTCGGGCGACTTATCAGATACGGTAATGTAAATACCTGGCA-
TTTGGGCACTTCTTGCGTTTAAGCGGGAAAGATCGCGAGGGCCCGCTATTTGCGATACTTCCCATGTCGGTGCCGTCG-
CCTCTATGTACTCGGAGACGTTAATGCAGAGGCTAAGGACAATTTACCATGACTCGGTAATCCGTTCGTCAAGCAGGTA-
GCTCGAGTCTCCCCACGGACACGTAGTGGGTTTGTAACGATCGATACCGAGTCTTTTTGTCTAGTAGAACCAACCAACC-
ATTAAGGAGTTCACTAGCACATCTTTGCGACCCGATCGTCCGTGTGTCGCGTAATACTTTTGTTATGACGAGACATACG-
CTCAAGCCCTGGGTAGCTAGTCGCGGAGGCACGTTACCGCGCACAACCCCTATTCGTTTACATGTACATCGCATCTGAG-
GTAGTACACTTCCGGCGTACGTGAGTATTTGCGCGTAATAAGCGCGTGTTTAGCTGATCCCCTCTCGTATCGAGGTTAA-
GGCAGATTAGTGCCCAGTAATTGCGTTTTTTTGTCGTTGTCGCAGAACGCGATTTGCTCCGAAAGCTTTAAGCCGTGG-
AAAAAAAAAAAAAAAAAAAAAAA

Spike-in 9

TCCAGATTACTTCCATTTCCGCCCAAGCTGCTCACAGTATACGGGCGTCGGCATCCAGACCGTCGGCTGATCGTGGTTT-
TACTAGGCTAGACTAGCGTACGAGCACTATGGTCAGTAATTCCTGGAGGAATAGGTACCAAGAAAAAAACGAACCTTT-
GGGTTCCAGAGCTGTACGGTCGCACTGAACTCGGATAGGTCTCAGAAAAACGAAATATAGGCTTACGGTAGGTCCGAA-
TGGCACAAAGCTTGTTCCGTTAGCTGGCATAAGATTCCATGCCTAGATGTGATACACGTTTCTGGAAACTGCCTCGTCA-
TGCGACTGTTCCCCGGGGTCAGGGCCGCTGGTATTTGCTGTAAAGAGGGGCGTTGAGTCCGTCCGACTTCACTGCCCC-
CTTTCAGCCTTTTGGGTCCTGTATCCCAATTCTCAGAGGTCCCGCCGTACGCTGAGGACCACCTGAAACGGGCATCGTC-
GCTCTTCGTTGTTCGTCGACTTCTAGTGTGGAGACGAATTGCCAGAATTATTAACTGCGCAGTTAGGGCAGCGTCTGA-
GGAAGTTTGCTGCGGTTTCGCCTTGACCGCGGGAAGGAGACATAACGATAGCGACTCTGTCTCAGGGGATCTGCATAT-
GTTTGCAGCATACTTTAGGTGGGCCTTGGCTTCCTTCCGCAGTCAAAACCGCGCAATTATCCCCGTCCTGATTTACTGG-
ACTCGCAACGTGGGTCCATCAGTTGTCCGTATACCAAGACGTCTAAGGGCGGTGTACACCCTTTTGAGCAATGATTGCA-
CAACCTGCGATCACCTTATACAGAATTATCAATCAAGCTCCCCGAGGAGCGGACTTGTAAGGACCGCCGCTTTCGCTCG-
GGTCTGCGGGTTATAGCTTTTCAGTCTCGACGGGCTAGCACACATCTGGTTGACTAGGCGCATAGTCGCCATTCACAG-
ATTTGCTCGGCAATCAGTACTGGTAGGCGTTAGACCCCGTGACTCGTGGCTGAACGGCCGTACAACTCGACAGCCGGT-
GCTTGCGTTTTACCCTTAAAAAAAAAAAAAAAAAAAAAAA

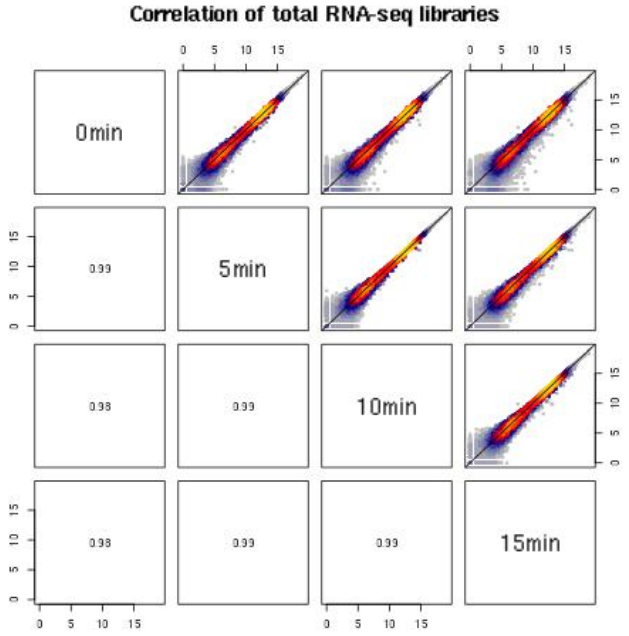
Spike-in 12

TATTGGTGGAGGGGCACAAGTTGCTGAAGTTGCGAGAGGGGCGATAAGTGAGGCAGACAGGCATAATATAAGAGGGG-
AGAGAATTAGCGTAGATACTCTTCCAATAGTTGGTGAAGAAAATTTATATGAGGCTGTTAAAGCTGTAGCAACTCTTCC-
ACGAGTAGGAATTTTAGTTTTAGCTGGCTCTTTAATGGGAGGGAAGATAACTGAAGCAGTTAAAGAATTAAAGGAAAA-
GACTGGCATTCCCGTGATAAGCTTAAAGATGTTTGGCTCTGTTCCTAAGGTTGCTGATTTGGTTGTTGGAGACCCATTG-
CAGGCAGGGGTTTTAGCTGTTATGGCTATTGCTGAAACAGCAAAATTTGATATAAATAAGGTTAAAGGTAGGGTGCTA-
TAAAGATAATTTAATAATTTTTGATGAAACCGAAGCGTTAGCTTTGGGTTATGAAACTCCATGATTTTCATTTAATTTT-
TTCCTATTAATTTTCTCCTAAAAAGTTTCTTTAACATAAATAAGGTTAAAGGGAGAGCTCTATGATTGTCTTCAAAAAT-
ACAAAGATTATTGATGTATATACTGGAGAGGTTGTTAAAGGAAATGTTGCAGTTGAGAGGGATAAAATATCCTTTGTG-
GATTTAAATGATGAAATTGATAAGATAATTGAAAAAATAAAGGAGGATGTTAAAGTTATTGACTTAAAAGGAAAATAT-
TTATCTCCAACATTTATAGATGGGCATATACATATAGAATCTTCCCATCTCATCCCATCAGAGTTTGAGAAATTTGTAT-
TAAAAAGCGGAGTTAGCAAAGTAGTTATAGACCCGCATGAAATAGCAAATATTGCTGGAAAAGAAGGAATTTTGTTTA-
TGTTGAATGATGCCAAAATTTTAGATGTCTATGTTATGCTTCCTTCCTGTGTTCCAGCTACAAACTTAGAAACAAGTGG-
AGCTGAGATTACAGCAGAGAATATTGAAGAACTCATTCTTTAGATAATGTCTTAGGTTAAAAAAAAAAAAAAAAAAAA-
AAA

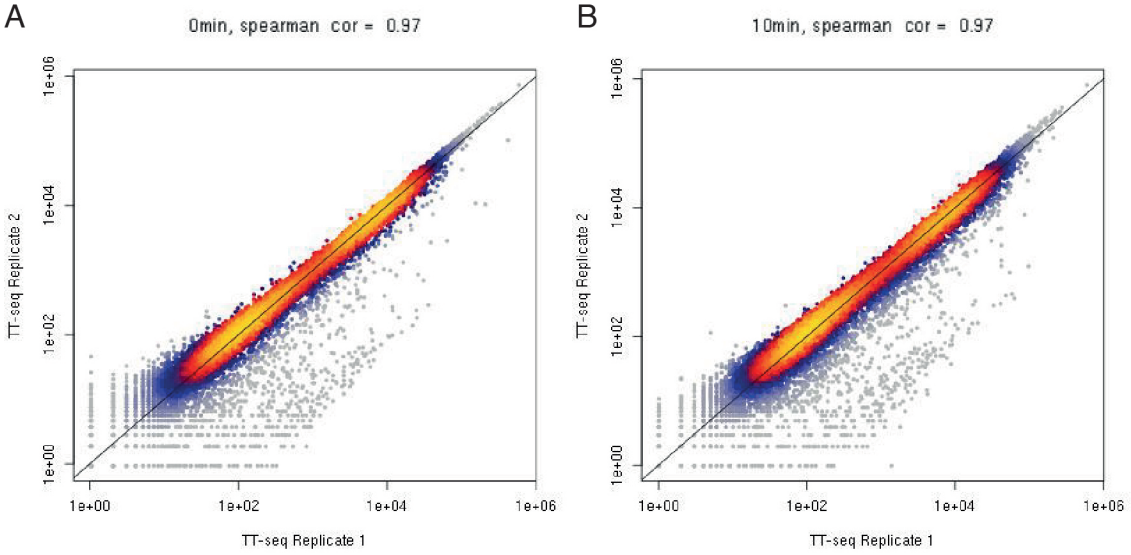
24.2 TT-seq protocol

Jurkat cells were acquired from DSMZ (Braunschweig, Germany). Cells were grown in RPMI 1640 medium (Gibco) supplemented with 10% heat-inactivated FBS (Gibco) and 1% Penicillin/Streptomycin (100x, PAA) at 37°C under 5% CO2. Cells were labeled in media for 5 min with 500 μM 4-thiouridine (4sU, Sigma-Aldrich) and activated with 50 mM PMA (Sigma-Aldrich) and 1 μM ionomycin (Sigma-Aldrich). Cells were harvested, spike-ins were added, and RNA was purified and fragmented as described (Schwalb et al. 2016). Fragmented RNA was subjected to purification of labeled RNA as described (Dölken et al. 2008). Labeled fragmented RNA (TT-seq) and total fragmented RNA (Total RNA-seq) were treated with 2 units of DNase Turbo (Life Technologies). Sequencing libraries were prepared with the Ovation Human Blood RNA-seq library kit (NuGEN) following the manufacturer's instructions. All samples were sequenced on an Illumina HiSeq 1500 sequencer.

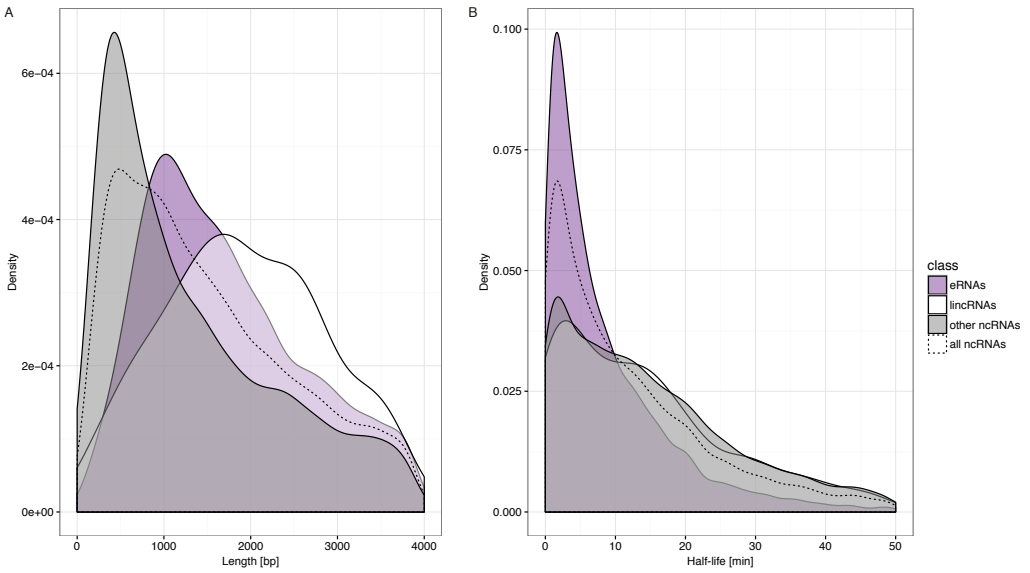
25 Appendix Figures for Section III



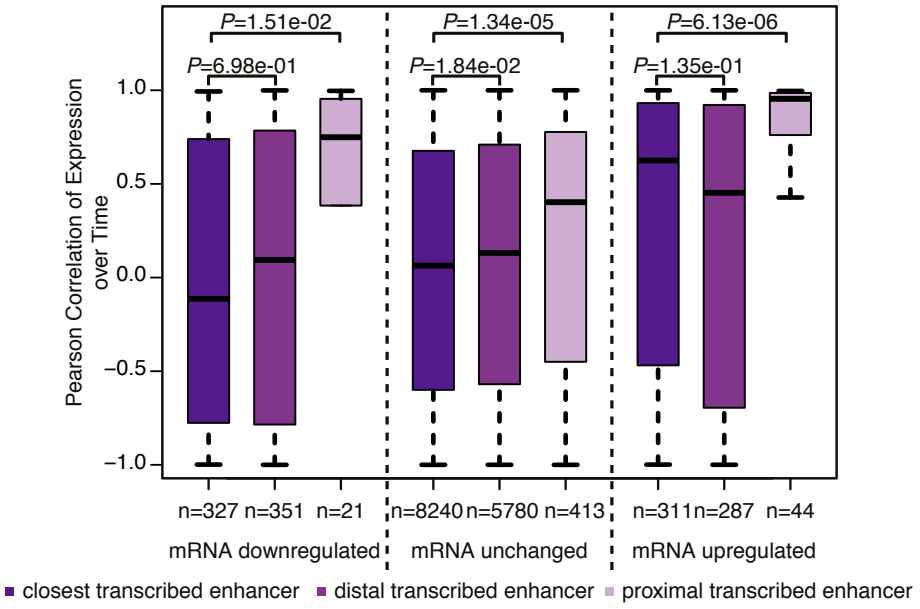
Appendix Figure A1: Correlation of read counts for total RNA-seq libraries. The single scatter plots show log2 read counts for individual total RNA-seq libraries. The lower triangle displays the Spearman correlation coefficients for any of the comparisons between two samples.



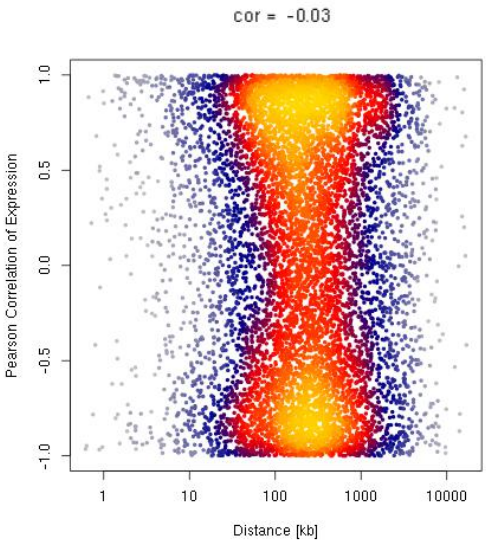
Appendix Figure A2: Correlation of TT-seq replicate measurements. **(A)** Correlation of read counts for TT-seq replicates before T-cell activation. **(B)** Correlation of read counts TT-seq replicates 10 min after T-cell activation.



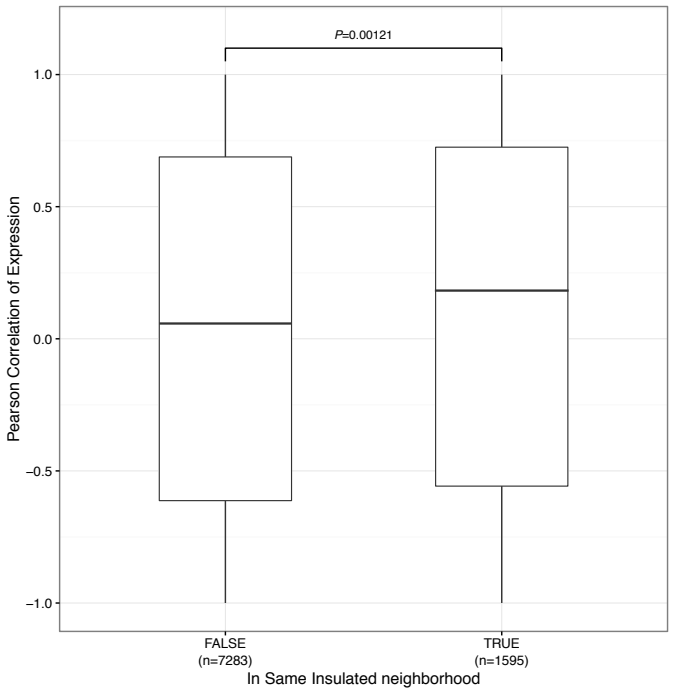
Appendix Figure A3: Characteristics of non-coding RNA classes. **(A)** Length distributions for different non-coding RNA classes (eRNAs, violet; lincRNAs, white; other ncRNAs; grey). The dashed line shows the distribution of all non-coding RNAs together. **(B)** Half-life distributions for different non-coding RNA classes.



Appendix Figure A4: Correlation of TT-seq signal over time. Correlation of TT-seq signal over time between closest (left boxes, dark violet), proximal (middle boxes, medium violet) or distal (right boxes, light violet) transcribed enhancers and promoters by change in promoter TT-seq signal (from left to right: downregulated, unchanged, upregulated promoters). Closest transcribed enhancers were taken for each mRNA irrespective of insulated neighborhood boundaries. Distal and proximal transcribed enhancers are located in the same insulated neighborhood as their respective promoters. The Pearson correlation coefficient was calculated between read counts across the time series (replicates averaged per time point) for each transcribed enhancer-promoter pair. The P-values were derived by two-sided Mann-Whitney U tests. Box limits are the first and third quartiles, the band inside the box is the median. The ends of the whiskers extend the box by 1.5 times the interquartile range.



Appendix Figure A5: Correlation vs Distance. Correlation of TT-seq signal over time between transcribed enhancers and promoters. Closest transcribed enhancers were taken for each mRNA irrespective of insulated neighborhood boundaries. The Pearson correlation coefficient was calculated between read counts across the time series (replicates averaged per time point) for each transcribed enhancer-promoter pair.



Appendix Figure A6: Correlation of TT-seq signal for closest eRNAs with their mRNAs dependent on location in same insulated neighborhood. The Pearson correlation coefficient was calculated between read counts across the time series (replicates averaged per time point) for each closest transcribed enhancer- promoter pair. The P-value was derived by a two-sided Mann-Whitney U test. Box limits are the first and third quartiles, the band inside the box is the median. The ends of the whiskers extend the box by 1.5 times the interquartile range.

26 Materials and Methods for Section 20

26.1 4tU-seq

The 4tU-seq experiments and analyses were performed as previously described (Eser et al. 2016; Miller et al. 2011). The Spt5 depletion strain was grown as described (Shetty et al. 2017) and cultures were labeled with 4tU for 10 min at 0 and 4.5 hr after the addition of thiamine and auxin. RNA spike-ins were added to cell pellets at the first step of RNA purification (Schwalb et al. 2016). The amount of spike-ins was adjusted to the cell number for each sample (120 ng of spike-in mix for 2.5 $\times 10^8$ cells for all samples). Sequencing libraries were prepared according to the manufacturer's recommendations using the Ovation Universal RNA-Seq System (NuGen). Libraries were sequenced on an Illumina HiSeq 2500 at LAFUGA, LMU Munich.

26.2 4tU-seq computational analysis

Paired-end 50 base reads with additional 6 base reads of barcodes were obtained in replicates for all samples. Reads were demultiplexed and mapped with STAR 2.3.0 (Dobin et al. 2015) to the concatenated S. pombe genome and external spike-in sequences with maximum one mismatch (–outFilterMismatchNmax 2) allowing for only one mapping position (–outFilterMultimapScoreRange 0). SAM files were filtered using Samtools (Li et al. 2009; Schwalb et al. 2016) for alignments with MAPQ of at least 7 (-q 7) and only proper pairs (-f99, -f149, -f84, -f163) were selected. This resulted in 44-95 million reads per sample. Further data processing was carried out using the R/Bioconductor environment (Gentleman et al. 2004; R Development Core Team 2011). Samples were normalized using external spike-ins as previously described (Schwalb et al. 2016) (Section 8).

26.3 Data availablility

The raw sequencing data reported in this paper have been deposited in the NCBI Gene Expression Omnibus under accession number GSE85182. Other data have been deposited to Mendeley Data and are available at `http://dx.doi.org/10.17632/v5jy3367rs.3`.

References

Ahn, S.H., Kim, M., and Buratowski, S. (2004). Phosphorylation of Serine 2 within the RNA Polymerase II C-Terminal Domain Couples Transcription and 3' End Processing. *Molecular Cell* 13(1): 67–76.

Allen, B.L. and Taatjes, D.J. (2015). The Mediator complex: a central integrator of transcription. *Nature Reviews Molecular Cell Biology* 16(3): 155–166.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology* 11(10): R106.

Anders, S. and Huber, W. (2012). Differential expression of RNA-Seq data at the gene level - the DESeq package. *Package Vignette*.

Anders, S., Pyl, P.T., and Huber, W. (2014). HTSeq - A Python framework to work with high-throughput sequencing data. *Bioinformatics.* 31(2): btu638.

Andersson, R. et al. (2014). An atlas of active enhancers across human cell types and tissues. *Nature* 507(7493): 455–461.

Ansari, A. and Hampsey, M. (2005). A role for the CPF 3'-end processing machinery in RNAP II-dependent gene looping. *Genes & Development* 19(24): 2969–2978.

Ardehali, M.B. and Lis, J.T. (2009). Tracking rates of transcription and splicing in vivo. *Nature Structural & Molecular Biology* 16(11): 1123–1124.

Arner, E. et al. (2015). Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science* 347(6225): 1010–1014.

Bahrami, S. and Drabløs, F. (2016). Gene regulation in the immediate-early response process. *Advances in Biological Regulation* 62(7491): 37–49.

Banerji, J., Rusconi, S., and Schaffner, W. (1981). Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell* 27(1): 299–308.

Basehoar, A.D., Zanton, S.J., and Pugh, B. (2004). Identification and Distinct Regulation of Yeast TATA Box-Containing Genes. *Cell* 116(5): 699–709.

Beyer, A.L. and Osheim, Y.N. (1988). Splice site selection, rate of splicing, and alternative splicing on nascent transcripts. *Genes & Development* 2(6): 754–765.

Bullard, J.H., Purdom, E., Hansen, K.D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 11: 94.

Buratowski, S. (1994). The Basics of Basal Transcription by RNA Polymerase II. *Cell* 77: 1–3.

Buratowski, S. (2003). The CTD code. *Nature structural biology* 10(9): 679–680.

Buratowski, S., Hahn, S., Guarente, L., and Sharp, P.A. (1989). Five Intermediate Complexes in Transcription Initiation by RNA Polymerase II. *Cell* 56: 549–561.

Byun, J.S. et al. (2009). Dynamic bookmarking of primary response genes by p300 and RNA polymerase II complexes. *Proceedings of the National Academy of Sciences of the United States of America* 106(46): 19286–19291.

Calo, E. and Wysocka, J. (2013). Modification of Enhancer Chromatin: What, How, and Why? *Molecular Cell* 49(5): 825–837.

Cameron, A.C. and Trivedi, P.K. (1998). *Regression Analysis of Count Data*. Cambridge University Press.

Carrillo Oesterreich, F., Bieberstein, N., and Neugebauer, K.M. (2011). Pause locally, splice globally. *Trends in Cell Biology* 21(6): 328–335.

Chan, Y.F. et al. (2010). Adaptive Evolution of Pelvic Reduction in Sticklebacks by Recurrent Deletion of a Pitx1 Enhancer. *Science* 327(5963): 302–305.

Cheadle, C. et al. (2005). Control of gene expression during T cell activation: alternate regulation of mRNA transcription and mRNA stability. *BMC genomics* 6: 75.

Cho, E.J., Takagi, T., Moore, C.R., and Buratowski, S. (1997). mRNA capping enzyme is recruited to the transcription complex by phosphorylation of the RNA polymerase II carboxy-terminal domain. *Genes & Development* 11(24): 3319–3326.

Churchman, L.S. and Weissman, J.S. (2011). Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* 469(7330): 368–373.

Cleary, M.D., Meiering, C.D., Jan, E., Guymon, R., and Boothroyd, J.C. (2005). Biosynthetic labeling of RNA with uracil phosphoribosyltransferase allows cell-specific microarray analysis of mRNA synthesis and decay. *Nature Biotechnology* 23(2): 232–237.

Colgan, D.F. and Manley, J.L. (1997). Mechanism and regulation of mRNA polyadenylation. *Genes & Development* 11(212): 2755–2766.

Connelly, S. and Manley, J.L. (1988). A functional mRNA polyadenylation signal is required for transcription termination by RNA polymerase II. *Genes & Development* 2(4): 440–452.

Corden, J.L., Cadena, D.L., Ahearn, J.M., and Dahmus, M.E. (1985). A unique structure at the carboxyl terminus of the largest subunit of eukaryotic RNA polymerase II. *Proceedings of the National Academy of Sciences of the United States of America* 82(23): 7934–7938.

Core, L.J., Waterfall, J.J., and Lis, J.T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* 322(5909): 1845–1848.

Core, L.J. et al. (2014). Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nature Genetics* 46(12): 1311–1320.

Crabtree, G.R. (1989). Contingent genetic regulatory events in T lymphocyte activation. *Science* 243(4889): 355–361.

Cramer, P. et al. (2008). Structure of Eukaryotic RNA Polymerases. *Annual Review of Biophysics* 37(1): 337–352.

Crick, F.H.C. (1970). Central Dogma of Molecular Biology. *Nature* 227: 561–563.

Darzacq, X. et al. (2007). In vivo dynamics of RNA polymerase II transcription. *Nature Structural & Molecular Biology* 14(9): 796–806.

Das, R. et al. (2006). Functional coupling of RNAP II transcription to spliceosome assembly. *Genes & Development* 20(9): 1100–1109.

De Santa, F. et al. (2010). A large fraction of extragenic RNA Pol II transcription sites overlap enhancers. *PLoS Biology* 8(5): e1000384.

De Pretis, S. et al. (2015). INSPEcT: a computational tool to infer mRNA synthesis, processing and degradation dynamics from RNA- and 4sU-seq time course experiments. *Bioinformatics* 31(17): 2829–2835.

Dekker, J., Marti-Renom, M.A., and Mirny, L.A. (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature Reviews Genetics* 14(6): 390–403.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing Chromosome Conformation. *Science* 295(5558): 1306–1311.

DeMare, L.E. et al. (2013). The genomic landscape of cohesin-associated chromatin interactions. *Genome Research* 23(8): 1224–1234.

Dieci, G. and Sentenac, A. (1996). Facilitated Recycling Pathway for RNA Polymerase III. *Cell* 84(2): 245–252.

Diehn, M. et al. (2002). Genomic expression programs and the integration of the CD28 costimulatory signal in T cell activation. *Proceedings of the National Academy of Sciences of the United States of America* 99(18): 11796–11801.

Dillies, M.A. et al. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics* 14(6): 671–683.

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

Dixon, J.R. et al. (2015). Chromatin architecture reorganization during stem cell differentiation. *Nature* 518(7539): 331–336.

Djebali, S. et al. (2012). Landscape of transcription in human cells. *Nature* 489(7414): 101–108.

Dobin, A. and Gingeras, T.R. (2015). Mapping RNA-seq Reads with STAR. *Current Protocols in Bioinformatics* 51: 11–14.

Dölken, L. et al. (2008). High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. *RNA* 14(9): 1959–1972.

Dostie, J. et al. (2006). Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome research* 16(10): 1299–1309.

Dowen, J.M. et al. (2014). Control of Cell Identity Genes Occurs in Insulated Neighborhoods in Mammalian Chromosomes. *Cell* 159(2): 374–387.

Ellisen, L.W. et al. (2001). Cascades of transcriptional induction during human lymphocyte activation. *European Journal of Cell Biology* 80(5): 321–328.

Eser, P. et al. (2013). Periodic mRNA synthesis and degradation co-operate during cell cycle gene expression. *Molecular systems biology* 10(1): 717.

Eser, P. et al. (2016). Determinants of RNA metabolism in the Schizosaccharomyces pombe genome. *Molecular Systems Biology* 12: 857.

External RNA Controls Consortium (2005). Proposed methods for testing and selecting the ERCC external RNA controls. *BMC genomics* 6: 150.

Fan, X., Chou, D.M., and Struhl, K. (2006). Activator-specific recruitment of Mediator in vivo. *Nature Structural & Molecular Biology* 13(2): 117–120.

Fang, B. et al. (2014). Circadian enhancers coordinate multiple phases of rhythmic gene transcription in vivo. *Cell* 159(5): 1140–1152.

Feske, S., Giltnane, J., Dolmetsch, R., Staudt, L.M., and Rao, A. (2001). Gene regulation mediated by calcium signals in T lymphocytes. *Nature Immunology* 2(4): 316–24.

Fong, N. and Bentley, D.L. (2001). Capping, splicing, and 3' processing are independently stimulated by RNA polymerase II: different functions for different segments of the CTD. *Genes & Development* 15(14): 1783–1795.

Friedel, C.C. and Dölken, L. (2009). Metabolic tagging and purification of nascent RNA: implications for transcriptomics. *Molecular BioSystems* 5(11): 1271–1278.

Frühauf, K. (2015). "Dissecting the regulation of gene expression during steroid hormone signaling in Drosophila by Dynamic Transcriptome Analysis (DTA)". PhD thesis. LMU München.

Fullwood, M.J. and Ruan, Y. (2009). ChIP-based methods for the identification of long-range chromatin interactions. *Journal of Cellular Biochemistry* 107(1): 30–39.

García-Martínez, J., Aranda, A., and Pérez-Ortín, J. (2004). Genomic Run-On Evaluates Transcription Rates for All Yeast Genes and Identifies Gene Regulatory Mechanisms. *Molecular Cell* 15(2): 303–313.

Gentleman, R. et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology* 5(10): R80.

Gilmartin, G.M. and Nevins, J.R. (1989). An ordered pathway of assembly of components required for polyadenylation site recognition and processing. *Genes & Development* 3(12 B): 2180–2190.

Gonzalez-Sandoval, A. and Gasser, S.M. (2016). On TADs and LADs : Spatial Control Over Gene Expression. *Trends in Genetics* 32(8): 485–495.

Greenberg, M.E. and Ziff, E.B. (1984). Stimulation of 3T3 cells induces transcription of the c-fos proto-oncogene. *Nature* 311(5985): 433–438.

Gribnau, J., Diderich, K., Pruzina, S., Calzolari, R., and Fraser, P. (2000). Intergenic Transcription and Developmental Remodeling of Chromatin Subdomains in the Human β-globin Locus. *Molecular Cell* 5(2): 377–386.

Grigull, J., Mnaimneh, S., Pootoolal, J., Robinson, M.D., and Hughes, T.R. (2004). Genome-Wide Analysis of mRNA Stability Using Transcription Inhibitors and Microarrays Reveals Posttranscriptional Control of Ribosome Biogenesis Factors. *Molecular and Cellular Biology* 24(12): 5534–5547.

Hah, N., Murakami, S., Nagari, A., Danko, C.G., and Lee Kraus, W. (2013). Enhancer transcripts mark active estrogen receptor binding sites. *Genome Research* 23(8): 1210–1223.

Hah, N. et al. (2015). Inflammation-sensitive super enhancers form domains of coordinately regulated enhancer RNAs. *Proceedings of the National Academy of Sciences of the United States of America* 112(3): E297–302.

Hahn, S. (2004). Structure and mechanism of the RNA Polymerase II transcription machinery. *Nature Structural & Molecular Biology* 11(5): 394–403.

Harrow, J. et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Research* 22(9): 1760–1774.

Hartzog, G.A., Wada, T., Handa, H., and Winston, F. (1998). Evidence that Spt4, Spt5, and Spt6 control transcription elongation by RNA polymerase II in Saccharomyces cerevisiae. *Genes & Development* 12(3): 357–69.

He, B., Chen, C., Teng, L., and Tan, K. (2014). Global view of enhancer-promoter interactome in human cells. *Proceedings of the National Academy of Sciences of the United States of America* 111(21): E2191–2199.

Heintzman, N.D. et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics* 39(3): 311–318.

Hirose, Y. and Manley, J.L. (2000). RNA polymerase II and the integration of nuclear events. *Genes & Development* 14: 1415–1429.

Hnisz, D., Day, D.S., and Young, R.A. (2016a). Insulated Neighborhoods: Structural and Functional Units of Mammalian Gene Control. *Cell* 167(5): 1188–1200.

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

Holstege, F.C.P. and Young, R.A. (1998). Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* 95(5): 717–728.

Hsieh, C.-L. et al. (2014). Enhancer RNAs participate in androgen receptor-driven looping that selectively enhances gene activation. *Proceedings of the National Academy of Sciences of the United States of America* 111(20): 7319–7324.

Huang, Y. and Carmichael, G.G. (1996). Role of polyadenylation in nucleocytoplasmic transport of mRNA. *Molecular and Cellular Biology* 16(4): 1534–1542.

Huber, W., Toedling, J., and Steinmetz, L.M. (2006). Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics* 22(16): 1963–1970.

Ilott, N.E. et al. (2014). Long non-coding RNAs and enhancer RNAs regulate the lipopolysaccharide-induced inflammatory response in human monocytes. *Nature Communications* 5: 3979.

Jacquier, A. (2009). The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. *Nature Reviews Genetics* 10(12): 833–844.

Jeffares, D.C., Penkett, C.J., and Bähler, J. (2008). Rapidly regulated genes are intron poor. *Trends in Genetics* 24(8): 375–378.

Jensen, T., Jacquier, A., and Libri, D. (2013). Dealing with pervasive transcription. *Molecular Cell* 52(4): 473–484.

Jiang, L. et al. (2011). Synthetic spike-in standards for RNA-seq experiments. *Genome Research* 21(9): 1543–1551.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science* 316(5830): 1497–1502.

Kaikkonen, M.U. et al. (2013). Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription. *Molecular Cell* 51(3): 310–325.

Kenzelmann, M. et al. (2007). Microarray analysis of newly synthesized RNA in cells and animals. *Proceedings of the National Academy of Sciences of the United States of America* 104(15): 6164–6169.

Kessler, O., Jiang, Y., and Chasin, L.A. (1993). Order of intron removal during splicing of endogenous adenine phosphoribosyltransferase and dihydrofolate reductase pre-mRNA. *Molecular and Cellular Biology* 13(10): 6211–6222.

Kim, M. et al. (2004). The yeast Rat1 exonuclease promotes transcription termination by RNA polymerase II. *Nature* 432(7016): 517–522.

Kim, T.-K. et al. (2010). Widespread transcription at neuronal activity-regulated enhancers. *Nature* 465(7295): 182–187.

Koleske, A.J. and Young, R.A. (1994). An RNA polymerase II holoenzyme responsive to activators. *Nature* 368: 466–469.

Kornberg, R.D. (2005). Mediator and the mechanism of transcriptional activation. *Trends in Biochemical Sciences* 30(5): 235–239.

Kotake, Y. et al. (2007). Splicing factor SF3b as a target of the antitumor natural product pladienolide. *Nature Chemical Biology* 3(9): 570–575.

Kugel, J.F. and Goodrich, J.A. (2002). Translocation after synthesis of a four-nucleotide RNA commits RNA polymerase II to promoter escape. *Molecular and Cellular Biology* 22(3): 762–773.

Kwak, H., Fuda, N.J., Core, L.J., and Lis, J.T. (2013). Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* 339(6122): 950–953.

Lam, L.T. et al. (2001). Genomic-scale measurement of mRNA turnover and the mechanisms of action of the anti-cancer drug flavopiridol. *Genome Biology* 2(10): research0041.1–research0041.11.

Lawrence, M., Gentleman, R., and Carey, V. (2009). rtracklayer: An R package for interfacing with genome browsers. *Bioinformatics* 25(14): 1841–1842.

Lawrence, M. et al. (2013). Software for Computing and Annotating Genomic Ranges. *PLoS Computational Biology* 9(8): e1003118.

Lenhard, B., Sandelin, A., and Carninci, P. (2012). Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nature Reviews Genetics* 13(4): 233–245.

Levine, M., Cattoglio, C., and Tjian, R. (2014). Looping back to leap forward: Transcription enters a new era. *Cell* 157(1): 13–25.

Li, B., Carey, M., and Workman, J.L. (2007). The Role of Chromatin during Transcription. *Cell* 128(4): 707–719.

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

Li, W., Notani, D., and Rosenfeld, M.G. (2016). Enhancers as non-coding RNA transcription units: recent insights and future perspectives. *Nature Reviews Genetics* 17(4): 207–223.

Li, W. et al. (2013). Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature* 498(7455): 516–520.

Lidschreiber, M., Leike, K., and Cramer, P. (2013). Cap completion and C-terminal repeat domain kinase recruitment underlie the initiation-elongation transition of RNA polymerase II. *Molecular and Cellular Biology* 33(19): 3805–3816.

Lieberman-Aiden, E. et al. (2009). of the Human Genome. *Science* 326: 289–294.

Liu, X., Bushnell, D.A., and Kornberg, R.D. (2013). RNA polymerase II transcription: structure and mechanism. *Biochimica et Biophysica Acta* 1829(1): 2–8.

Liu, Y. et al. (2004). Two cyclin-dependent kinases promote RNA polymerase II transcription and formation of the scaffold complex. *Molecular and Cellular Biology* 24(4): 1721–1735.

Logan, J., Falck-Pedersen, E., Darnell, J.E., and Shenk, T. (1987). 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. *Proceedings of the National Academy of Sciences of the United States of America* 84(23): 8306–8310.

Love, M.I., Anders, S., and Huber, W. (2014a). Differential analysis of count data - the DESeq2 package. *Package Vignette*.

Love, M.I., Huber, W., and Anders, S. (2014b). Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Genome Biology* 15(12): 550.

Lovén, J. et al. (2012). Revisiting Global Gene Expression Analysis. *Cell* 151(3): 476–482.

Lubas, M. et al. (2015). The human nuclear exosome targeting complex is loaded onto newly synthesized RNA to direct early ribonucleolysis. *Cell Reports* 10(2): 178–192.

Malik, S. and Roeder, R.G. (2005). Dynamic regulation of pol II transcription by the mammalian Mediator complex. *Trends in Biochemical Sciences* 30(5): 256–263.

Malik, S. and Roeder, R.G. (2010). The metazoan Mediator co-activator complex as an integrative hub for transcriptional regulation. *Nature Reviews Genetics* 11(11): 761–772.

Maniatis, T.O.M., Goodbourn, S., and Fischer, J.A. (1987). Regulation of Inducible and Tissue-Specific Gene Expression. *Science* 236: 1237–1245.

Margeat, E. et al. (2006). Direct observation of abortive initiation and promoter escape within single immobilized transcription complexes. *Biophysical Journal* 90(4): 1419–1431.

Marguerat, S. et al. (2012). Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells. *Cell* 151(3): 671–683.

Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., and Gilad, Y. (2008). RNA-seq : An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research* 18(9): 1509–1517.

Marrack, P. et al. (2000). Genomic-scale analysis of gene expression in resting and activated T cells. *Current Opinion in Immunology* 12(2): 206–209.

Marshall, N.F. and Price, D.H. (1995). Purification of P-TEFb, a transcription factor required for the transition into productive elongation. *The Journal of Biological Chemistry* 270(21): 12335–12338.

Martin, R.M., Rino, J., Carvalho, C., Kirchhausen, T., and Carmo-Fonseca, M. (2013). Live-Cell Visualization of Pre-mRNA Splicing with Single-Molecule Sensitivity. *Cell Reports* 4: 1144–1155.

Mayer, A. et al. (2010). Uniform transitions of the general RNA polymerase II transcription complex. *Nature Structural & Molecular Biology* 17(10): 1272–1278.

McCracken, S. et al. (1997a). 5'-Capping enzymes are targeted to pre-mRNA by binding to the phosphorylated carboxy-terminal domain of RNA polymerase II. *Genes & Development* 11(24): 3306–3318.

McCracken, S. et al. (1997b). The C-terminal domain of RNA polymerase II couples mRNA processing to transcription. *Nature* 385(6614): 357–361.

Melgar, M.F., Collins, F.S., and Sethupathy, P. (2011). Discovery of active enhancers through bidirectional expression of short transcripts. *Genome Biology* 12(11): R113.

Michel, M. (2016). "Transient transcriptome sequencing : development and applications in human cells". PhD thesis. LMU München.

Michel, M. et al. (2017). TT-seq captures enhancer landscapes immediately after T-cell stimulation. *Molecular Systems Biology* 13(3): 920.

Miller, C. et al. (2011). Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. *Molecular Systems Biology* 7(458): 458.

Miller, C. et al. (2012). Mediator phosphorylation prevents stress response transcription during non-stress conditions. *The Journal of Biological Chemistry* 287(53): 44017–44026.

Miller, M.R., Robinson, K.J., Cleary, M.D., and Doe, C.Q. (2009). TU-tagging: cell type-specific RNA isolation from intact complex tissues. *Nature Methods* 6(6): 439–441.

Moore, C.L. and Sharp, P.A. (1985). Accurate cleavage and polyadenylation of exogenous RNA substrate. *Cell* 41(3): 845–855.

Morgan, M., Obenchain, V., Hester, J., and Pagès, H. (2016). *SummarizedExperiment: SummarizedExperiment Container*.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* 5(7): 621–628.

Myers, L.C. et al. (1998). The Med proteins of yeast and their function through the RNA polymerase II carboxy-terminal domain. *Genes & Development* 12(1): 45–54.

Nagalakshmi, U. et al. (2008). The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science* 320(5881): 1344–1349.

Nechaev, S. and Adelman, K. (2011). Pol II waiting in the starting gates: Regulating the transition from transcription initiation into productive elongation. *Biochimica et Biophysica Acta* 1809(1): 34–45.

Orkin, S.H. (1990). Globin gene regulation and switching: Circa 1990. *Cell* 63(4): 665–672.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics* 40(12): 1413–1415.

Paralkar, V.R. et al. (2016). Unlinking an lncRNA from Its Associated cis Element. *Molecular Cell* 62(1): 104–110.

Parker, R. and Sheth, U. (2007). P Bodies and the Control of mRNA Translation and Degradation. *Molecular Cell* 25(5): 635–646.

Paulsen, M.T. et al. (2014). Use of Bru-Seq and BruChase-Seq for genome-wide assessment of the synthesis and stability of RNA. *Methods* 67(1): 45–54.

Peterlin, B.M. and Price, D.H. (2006). Controlling the elongation phase of transcription with P-TEFb. *Molecular Cell* 23(3): 297–305.

Phillips-Cremins, J.E. et al. (2013). Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* 153(6): 1281–1295.

Proudfoot, N.J. (1989). How RNA polymerase II terminates transcription in higher eukaryotes. *Trends in Biochemical Sciences* 14(3): 105–110.

Proudfoot, N. (2004). New perspectives on connecting messenger RNA 3' end formation to transcription. *Current Opinion in Cell Biology* 16(3): 272–278.

Proudfoot, N.J., Furger, A., and Dye, M.J. (2002). Integrating mRNA Processing with Transcription. *Cell* 108(4): 501–512.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*.

Rabani, M. et al. (2011). Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nature Biotechnology* 29(5): 436–442.

Rabani, M. et al. (2014). High-Resolution Sequencing and Modeling Identifies Distinct Dynamic RNA Regulatory Strategies. *Cell* 159(7): 1698–1710.

Raghavan, A. et al. (2002). Genome-wide analysis of mRNA decay in resting and activated primary human T lymphocytes. *Nucleic Acids Research* 30(24): 5529–5538.

Rahl, P.B. et al. (2010). c-Myc Regulates Transcriptional Pause Release. *Cell* 141(3): 432–445.

Rasmussen, E.B. and Lis, J.T. (1993). In vivo transcriptional pausing and cap formation on three Drosophila heat shock genes. *Proceedings of the National Academy of Sciences of the United States of America* 90(17): 7923–7927.

Ren, B. (2010). Transcription: Enhancers make non-coding RNA. *Nature* 465(7295): 173–174.

Renner, D.B., Yamaguchi, Y., Wada, T., Handa, H., and Price, D.H. (2001). A highly purified RNA polymerase II elongation control system. *The Journal of Biological Chemistry* 276(45): 42601–42609.

Reyes-Reyes, M. and Hampsey, M. (2007). Role for the Ssu72 C-terminal domain phosphatase in RNA polymerase II transcription elongation. *Molecular and Cellular Biology* 27(3): 926–936.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2009). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26(1): 139–140.

Robinson, M.D. and Smyth, G.K. (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* 23(21): 2881–2887.

Robinson, M. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* 11(3): R25.

Roeder, R.G. and Rutter, W.J. (1969). Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms. *Nature* 224(5216): 234–237.

Roeder, R.G. (1996). The role of general initiation factors in transcription by RNA polymerase II. *Trends in Biochemical Sciences* 21(9): 327–335.

Rogge, L. et al. (2000). Transcript imaging of the development of human T helper cells using oligonucleotide arrays. *Nature Genetics* 25(1): 96–101.

Schaffner, W. (2015). Enhancers, enhancers - From their discovery to today's universe of transcription enhancers. *Biological Chemistry* 396(4): 311–327.

Schaukowitch, K. et al. (2014). Enhancer RNA Facilitates NELF Release from Immediate Early Genes. *Molecular Cell* 56(1): 29–42.

Schmitt, A.D., Hu, M., and Ren, B. (2016). Genome-wide mapping and analysis of chromosome architecture. *Nature Reviews Molecular Cell Biology* 17(12): 743–755.

Schübeler, D. (2007). Enhancing genome annotation with chromatin. *Nature Genetics* 39(3): 284–285.

Schulz, D. et al. (2013). Transcriptome Surveillance by Selective Termination of Noncoding RNA Synthesis. *Cell* 1: 1–13.

Schwalb, B. (2012). "Dynamic transcriptome analysis (DTA): Kinetic modeling of synthesis and decay of mRNA transcripts upon perturbation in S.cerevisiae, S.pombe and D.melanogaster". PhD thesis.

Schwalb, B. et al. (2012). Measurement of genome-wide RNA synthesis and decay rates with Dynamic Transcriptome Analysis (DTA). *Bioinformatics* 28(6): 884–885.

Schwalb, B. et al. (2016). TT-seq maps the human transient transcriptome. *Science* 352(6290): 1225–1228.

Shalon, D., Smith, S.J., and Brown, P.O. (1996). A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Research* 6(7): 639–645.

Shandilya, J. and Roberts, S.G.E. (2012). The transcription cycle in eukaryotes: From productive initiation to RNA polymerase II recycling. *Biochimica et Biophysica Acta* 1819(5): 391–400.

Sheng, M. and Greenberg, M.E. (1990). The regulation and function of c-fos and other immediate early genes in the nervous system. *Neuron* 4(4): 477–485.

Shetty, A. et al. (2017). Spt5 Plays Vital Roles in the Control of Sense and Antisense Transcription Elongation Article Spt5 Plays Vital Roles in the Control of Sense and Antisense Transcription Elongation. *Molecular Cell* 66: 1–12.

Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics* 15(4): 272–286.

Sidorenkov, I., Komissarova, N., and Kashlev, M. (1998). Crucial role of the RNA:DNA hybrid in the processivity of transcription. *Molecular Cell* 2(1): 55–64.

Simonis, M. et al. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nature Genetics* 38(11): 1348–1354.

Singh, J. and Padgett, R.A. (2009). Rates of in situ transcription and splicing in large human genes. *Nature Structural & Molecular Biology* 16(11): 1128–1133.

Smith-Garvin, J.E., Koretzky, G.A., and Jordan, M.S. (2009). T Cell Activation. *Annual Review of Immunology* 27(1): 591–619.

Splinter, E. et al. (2006). CTCF mediates long-range chromatin looping and local histone modification in the β-globin locus. *Genes & Development* 20(17): 2349–2354.

Step, S.E. et al. (2014). Anti-diabetic rosiglitazone remodels the adipocyte transcriptome by redistributing transcription to PPARγ-driven enhancers. *Genes & Development* 28(9): 1018–1028.

Sun, M. et al. (2012). Comparative dynamic transcriptome analysis (cDTA) reveals mutual feedback between mRNA synthesis and degradation. *Genome Research* 22(7): 1350–1359.

Sun, M. et al. (2013). Global Analysis of Eukaryotic mRNA Degradation Reveals Xrn1-Dependent Buffering of Transcript Levels. *Molecular Cell* 52(1): 52–62.

Svejstrup, J.Q. (2004). The RNA polymerase II transcription cycle: cycling through chromatin. *Biochimica et Biophysica Acta* 1677: 64–73.

Tan, G. (2015). *JASPAR2016: Data package for JASPAR2016.*

Tan, G. and Lenhard, B. (2016). TFBSTools: an R/bioconductor package for transcription factor binding site analysis. *Bioinformatics* 32(10): 1555–1556.

Teixeira, A. et al. (2004). Autocatalytic RNA cleavage in the human β-globin pre-mRNA promotes transcription termination. *Nature* 432: 526–530.

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

Thurman, R.E. et al. (2012). The accessible chromatin landscape of the human genome. *Nature* 489(7414): 75–82.

Tie, F. et al. (2009). CBP-mediated acetylation of histone H3 lysine 27 antagonizes Drosophila Polycomb silencing. *Development* 136(18): 3131–3141.

Tullai, J.W. et al. (2007). Immediate-early and delayed primary response genes are distinct in function and genomic architecture. *Journal of Biological Chemistry* 282(33): 23981–23995.

Venables, B. and Ripley, B. (2002). *Modern Applied Statistics with S*. 4th Editio. Springer Verlag.

Viladevall, L. et al. (2009). TFIIH and P-TEFb Coordinate Transcription with Capping Enzyme Recruitment at Specific Genes in Fission Yeast. *Molecular Cell* 33(6): 738–751.

Visel, A. et al. (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* 457(7231): 854–8.

Wachutka, L. and Gagneur, J. (2016). Measures of RNA metabolism rates: Toward a definition at the level of single bonds. *Transcription* e1257972.

Wang, A. et al. (2015). Epigenetic priming of enhancers predicts developmental competence of hESC-derived endodermal lineage intermediates. *Cell Stem Cell* 16(4): 386–399.

Wang, D. et al. (2011). Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature* 474(7351): 390–4.

Wang, L., Feng, Z., Wang, X., Wang, X., and Zhang, X. (2009a). DEGseq: An R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics* 26(1): 136–138.

Wang, W., Carey, M., and Gralla, J. (1992). Polymerase II promoter activation: closed complex formation and ATP-driven start-site opening. *Science* 255(5043): 450–453.

Wang, Y. et al. (2002). Precision and functional specificity in mRNA decay. *Proceedings of the National Academy of Sciences of the United States of America* 99(9): 5860–5865.

Wang, Z., Gerstein, M., and Snyder, M. (2009b). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* 10(1): 57–63.

Washietl, S., Hofacker, I.L., Lukasser, M., Hüttenhofer, A., and Stadler, P.F. (2005). Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nature Biotechnology* 23(11): 1383–1390.

Weiss, A. and Imboden, J.B. (1987). Cell surface molecules and early events involved in human T lymphocyte activation. *Advances in Immunology* 41: 1–38.

West, S., Gromak, N., and Proudfoot, N.J. (2004). Human 5' to 3' exonuclease Xrn2 promotes transcription termination at co-transcriptional cleavage sites. *Nature* 432(7016): 522–525.

Whitaker, L. (1914). On the Poisson Law of Small Numbers. *Biometrika* 10(1): 36–71.

Whitelaw, E. and Proudfoot, N. (1986). Alpha-thalassaemia caused by a poly(A) site mutation reveals that transcriptional termination is linked to 3' end processing in the human alpha 2 globin gene. *The EMBO journal* 5(11): 2915–2922.

Wu, H. et al. (2014). Tissue-Specific RNA Expression Marks Distant-Acting Developmental Enhancers. *PLoS Genetics* 10(9): e1004610.

Xi, H. et al. (2007). Identification and Characterization of Cell Type-Specific and Ubiquitous Chromatin Regulatory Structures in the Human Genome. *PLoS Genetics* 3(8): e136.

Yudkovsky, N., Ranish, J.A., and Hahn, S. (2000). A transcription reinitiation intermediate that is stabilized by activator. *Nature* 408(6809): 225–229.

Zacher, B. et al. (2017). Accurate Promoter and Enhancer Identification in 127 ENCODE and Roadmap Epigenomics Cell Types and Tissues by GenoSTAN. *PLoS ONE* 12(1): 041020.

Zhao, J., Hyman, L., and Moore, C. (1999). Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. *Microbiology and Molecular Biology Reviews* 63(2): 405–445.

Abbreviations

3C	chromosome conformation capture
4sU	4-thiouridine
4tU	4-thiouracil
cDNA	complementary DNA
ChIA-PET	chromatin interaction analysis using paired-end tag sequencing
ChIP	chromatin immunoprecipitation
CPE	core promoter element
CPSF	cleavage and polyadenylation specificity factor
CstF	cleavage stimulatory factor
CTCF	CCCTC-binding factor
CTD	C-terminal domain
DHS	DNaseI hypersensitivity sites
DNA	deoxyribonucleic acid
DSE	downstream sequence element
DSIF	DRB-sensitivity inducing factor
EF	elongation factor
ERCC	External RNA Controls Consortium
eRNA	enhancer RNA
GLM	Generalized Linear Model
GTF	general transcription factor
lncRNA	long non-coding RNA
ML	maximum likelihood
mRNA	messenger RNA
ncRNA	non-coding RNA
NELF	negative elongation factor
pA	polyadenylation
PCR	polymerase chain reaction
PIC	pre-initiation complex
Pol II	RNA polymerase II
PWM	position weight matrix
RNA	ribonucleic acid
RNA-seq	RNA sequencing
RPK	reads per kilobase
RPKM	reads per kilobase per million mapped reads
S. cerevisiae	Saccharomyces cerevisiae
S. pombe	Schizosaccharomyces pombe
snoRNA	small nucleolar RNA

snRNA	small nuclear RNA
TAD	topologically associated domain
TF	transcription factor
TSS	transcription start site
TT-seq	transient transcriptome sequencing
TU	transcription unit
UAS	upstream activating sequence

List of Figures

List of Tables