Title
Identification and Characterization of Systems with Defects in Transcription Termination

Author
Roth, Samuel

Publication Date
2021

UNIVERSITY OF CALIFORNIA SAN DIEGO

Identification and Characterization of Systems with Defects in Transcription Termination

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Bioinformatics and Systems Biology

by

Samuel J. Roth

Committee in charge:

Professor Christopher Benner, Chair
Professor Sven Heinz, Co-Chair
Professor Alon Goren
Professor Olivier Harismendy
Professor Gene Yeo
Professor Elina Zuniga

2021

The Dissertation of Samuel J. Roth is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

DEDICATION

To my parents, for putting up with me through my best and my worst.

EPIGRAPH

> *I can accept failure, everyone fails at something. But I can't accept not trying.*

- Michael Jordan

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

There are too many people who have influenced me in my path to possibly thank everyone, but I promise you if you have influenced my life I am grateful. My parents for being there for all of the good and bad moments. I have come to realize that I am a rare person in that I am both friends with my parents and hold them in the highest respect. Alan, my father, thank you for teaching me tireless work ethic and supporting me whenever I've made crazy choices you don't agree with. Sara, my mother, thank you for teaching me that my obsession with detail and critical eye were assets when used properly. Both of you provided the foundation for my scientific mind. There is a posthumous thank you to my grandfather Eddie. He was the first person to encourage my scientific curiosity and pursuits as a scientist. I am truly sad that you don't get to be here for these moments. Thank you to the rest of my family for putting up with my very particular interests in science, jiu jitsu, cooking, and all the odd obsessions I've acquired over the years.

Thank you to all my friends. You guys are the best. Thank you to Arjun Chandrasekhar for being there through the good, the bad, and the ugly. It's been a crazy journey. Thank you to Kevin Blansit for being my co-chef in cookouts and a genuine friend. Thank you to Townsend Montilla for somehow being a gym bro and a creative writing student. I still don't understand how you exist and I love it. Thank you to Max Kaplan. There are no words and there don't have to be. Thank you to Chris Maute for being the older brother I never had but always needed. Rest assured, I know you're always right. Thank you to Ross Bennie for late night conversations about business, wine, women, and all sorts of silliness. Thank you to my jiu jitsu friends for beating me up and letting me get the occasional leg lock. Thank you Andre Galvao, you are a black belt at jiu jitsu and at life. You have been so kind to me in so many ways. Thank you to my rescue pitbull Angus. Though you can't read, I hope you know that you taught me the meaning of unconditional love and kept my feet warm during thesis writing.

Thank you to all my professors at Wesleyan and UCSD. You have inspired me so much. Thank you to my thesis committee for providing guidance and insight into my research. Thank you to the Svenner lab for letting me be both the gym bro and the bioinformatics nerd. It was a good dual role. Thank you to Sascha Duttke for absurd conversations and an excellent partnership in research. Thank you to Camila De Arruda Saldanha for mentoring me in the wet lab. I would not have graduated without you. Thank you to my co-advisor Sven Heinz for somehow being convinced that I belonged in the wet lab and always providing enthusiastic insight at every turn. Your excitement is always contagious. Last but not least, thank you Chris Benner. You took a chance on me during the toughest year of my life and mentored me to be the kind of scientist I didn't think I could be. Thank you for being available and believing in me when I didn't believe in myself.

Chapter 2, in full, is a reprint of "ARTDeco: automatic readthrough transcription detection" as it appears in BMC Bioinformatics 2020. Samuel J. Roth, Sven Heinz, Christopher Benner. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in part, is material being prepared for submission contributed by Samuel J. Roth, Camila De Arruda Saldanha, Sven Heinz, Christopher Benner. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in part, is material being prepared for submission contributed by Samuel J. Roth, Sascha H.C. Duttke, Christopher Benner. The dissertation author was one of the primary investigators and authors of this paper.

VITA

2013	B.A. Biology, Computer Science, Wesleyan University, Middletown, CT
2013	M.A. Biology, Wesleyan University, Middletown, CT
2021	Ph.D., Bioinformatics and Systems Biology, University of California San Diego, La Jolla, CA, USA

PUBLICATIONS

Isaac Shamie, Sascha H. Duttke, Karen J. la Cour Karottki, Claudia Z. Han, Anders H. Hansen, Hooman Hefzi, Kai Xiong, Shangzhong Li, **Samuel J. Roth**, Jenhan Tao, Gyun Min Lee, Christopher K. Glass, Helene Faustrup Kildegaard, Christopher Benner, and Nathan E. Lewis. A Chinese hamster transcription start site atlas that enables targeted editing of CHO cells. Nucleic Acid Research Genomics and Bioinformatics, 3(3), 2021.

Sascha H. Duttke, **Samuel J. Roth**, and Christopher Benner. The human transcription factor motifs functionally defined by transcription initiation. In preparation, 2021.

Samuel J. Roth, Sven Heinz, and Christopher Benner. Artdeco: automatic readthrough transcription detection. BMC Bioinformatics, 21(214), 2020.

Brynne E. Lycette, Jacob W. Glickman, **Samuel J. Roth**, Abigail E. Cram, Tae Hee Kim, Danny Krizanc, and Michael P. Weir. N-terminal peptide detection with optimized peptide-spectrum matching and streamlined sequence libraries. Journal of Proteome Research, 15(9):2891—-2899, 2016.

Miin S. Lin, Justin J. Cherny, Claire T. Fournier, **Samuel J. Roth**, Danny Krizanc, and Michael P. Weir. Assessment of ms/ms search algorithms with parent-protein profiling. Journal of Proteome Research, 13(4):1823—-1832, 2014.

ABSTRACT OF THE DISSERTATION

Identification and Characterization of Systems with Defects in Transcription Termination

by

Samuel J. Roth

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California San Diego, 2021

Professor Christopher Benner, Chair
Professor Sven Heinz, Co-Chair

Transcription termination is a fundamental process in gene regulation. It is a critical step in mRNA maturation and it has been found that several cellular stresses can disrupt transcription termination. When termination is disrupted, transcription continues past the annotated 3’ end of genes (called readthrough transcription). This has many downstream effects such as novel elongated transcripts, changes in epigenetic state, and large alterations in 3D genome structure (Hennig et al. 2018; Heinz et al. 2018). Given the variety of both causes and consequences of this phenotype, it is critical to develop methods to both identify and characterize defects of transcription termination (DoTT). In my first chapter, I present a software package called Automatic Readthrough Transcription Detection (ARTDeco), which can quantify readthrough transcription in data generated by next generation sequencing (NGS) assays that measure transcription. We demonstrate ARTDeco's ability to discriminate between systems with DoTT and those with normal transcription termination. ARTDeco is able to quantify the degree of readthrough transcription in a system using three separate metrics. It is able to discriminate whether genes are transcribed due to gene activation (called primary induction genes) or due to readthrough transcription extending from the end of one gene through the body of its downstream gene (called read-in genes). We show that read-in genes represent analytical noise in the context of functional analyses. In addition, ARTDeco can identify downstream of gene (DoG) transcripts, which are intergenic transcripts originating from faulty termination. We show that ARTDeco can flexibly perform these functions across a variety of data types and organisms. In my second chapter, I deploy ARTDeco on NCBI's Gene Expression Omnibus (GEO) repository of NGS data to search for signs of DoTT in virally-infected samples. We find evidence that several viruses cause DoTT. Among these viruses, we identify a likely mechanism for readthrough transcription in Rift Valley Fever Virus (RVFV). We confirm that the RVFV's NSs protein causes DoTT by expressing it in THP-1 monocytes. Further, we compare the full range of transcriptional responses between NSs and the NS1 protein from influenza A virus (IAV). We find that both proteins cause global readthrough transcription and disrupt interferon signaling in distinct ways.

Finally, I develop a software package to address a different fundamental regulatory process: transcription initiation. Transcription initiation is known to occur as a result of multiple transcription factors (TFs) binding to a regulatory sequence and recruiting transcriptional machinery. Existing computational methods do not adequately capture the collaboration of the TFs from sequence alone. I developed the Dual HOMER method, which employs successive rounds of motif enrichment in order to infer cooperativity between TFs in transcription start site regions (TSRs). We show that Dual HOMER is able to recapitulate known interactions between TFs and lends novel insights into these interactions due to the properties of the transcriptional network it generates. In all, this thesis advances the understanding of two fundamental biological processes and outlines methods that lend biological insight to both.

CHAPTER 1: INTRODUCTION

One of the most important parts of transcription regulation is proper termination of transcripts. In normal conditions, polyadenylation-dependent transcription termination occurs when a transcribing RNA polymerase II (RNAPII) encounters a polyadenylation site (PAS), which triggers a change in conformation and recruitment of the cleavage and polyadenylation (CPA) machinery (Licatalosi et al. 2002). This leads to the release of the nascent pre-mRNA for further processing. RNAPII continues to transcribe RNAs that are degraded by the exonuclease XRN2. This continues until XRN2 catches up to the elongating RNAPII and causes its dissociation from the DNA (West, Gromak, and Proudfoot 2004; Kim et al. 2004)

Proper transcription termination is of critical importance for gene regulation. In addition to maintaining intact transcriptional units, defects of transcription termination (DoTT) can be deleterious to gene regulation and genomic integrity. One consequence of defective transcription termination is transcriptional interference (Greger and Proudfoot 1998; Shearwin, Callen, and Egan 2005). This is when transcription continues past the canonical 3’ end of a gene and disrupts transcription initiation at the promoter of a downstream gene on the same strand. If the genes are on different strands, two elongating RNAPII complexes on opposite strands can collide (Prescott and Proudfoot 2002; Hobson et al. 2012). In extreme cases, this can cause DNA damage and genome instability (Gaillard, García-Muse, and Aguilera 2015).

It has recently been observed that several cellular stresses such as heat shock, osmotic stress, oxidative stress, hypoxia, senescence, cancer, and infection by influenza A virus (IAV) and herpes simplex virus 1 (HSV-1) can cause DoTT (Bauer et al. 2018; Cardiello, Goodrich, and Kugel 2018; Grosso et al. 2015; Heinz et al. 2018; Hennig et al. 2018; Muniz et al. 2017; Rutkowski et al. 2015; Vilborg et al. 2015, 2017). While this is a stress-mediated phenotype, in IAV and HSV-1, the viral proteins NS1 and ICP27, respectively, are known to cause this phenotype by interacting with the CPA machinery (Nemeroff et al. 1998; X. Wang et al. 2020). In clear cell renal carcinoma, it is hypothesized that mutations to the methyltransferase SETD2 cause alterations in the epigenome that make the cells more susceptible to DoTT (Grosso et al. 2015). Despite this multitude of known causes of this phenotype, there is still much work to do to characterize the full scope of possible causes.

There are many phenotypic consequences for DoTT. Among these are alterations to the epigenome. Widespread transcription elongation in systems with DoTT leads to the opening of chromatin and reduced binding of transcription factors (TFs) (Hennig et al. 2018; Heinz et al. 2018). Transcripts that result from defective termination (also called readthrough transcription) do not go through normal mRNA maturation processes and, as a result, are not transcribed or translated (Rutkowski et al. 2015; Heinz et al. 2018). Given the scope of these phenotypic consequences, comprehensive characterization of DoTT is needed from the perspective of identifying mechanisms that induce DoTT as well as their downstream effects.

In addition to these epigenomic and transcriptomic consequences for DoTT, there are analytical consequences. In systems with widespread DoTT, many genes are transcribed due to readthrough transcription rather than promoter activation (we term these read-in genes) (Rutkowski et al. 2015; Heinz et al. 2018). Read-in genes represent analytical noise when examining patterns of gene expression using differential expression because these transcripts are retained in the nucleus and are never translated. Thus, they represent both non-functional transcripts as well as an artificial increase in transcription due to non-regulatory mechanisms. This can reduce discovery ability when investigating transcriptional networks in the context of systems with DoTT. Thus, it is important to be able to identify read-in genes in these systems.

Another fundamental issue in transcriptional regulation is understanding how pairs of TFs collaborate to establish transcription start site regions (TSRs). Regulatory regions of the genome are formed when TFs bind to their target sequences, open the chromatin in these regions, recruit transcriptional machinery, and initiate transcription (Heinz et al. 2010). Rather than operating one TF at a time, this is a collaborative process wherein multiple TFs work together to establish a regulatory region or TSR (Zhu, Shendure, and Church 2005; Heinz et al. 2010, 2015).

Several methods exist for inferring these cooperative interactions. However, these methods utilize statistical methods such as pointwise mutual information and the Fisher's exact test, which estimate the co-occurrence of TF motifs against a random background model (van Bömmel et al. 2018; Meckbach et al. 2015). This ensures that enrichments for co-occurrence are above random expectation. However, this does not ensure that the motifs are enriched relative to their genomic background and, therefore, biologically relevant. This is especially important in the context of cell-type and signal specificity.

In order to address issues in characterizing and identifying systems with DoTT, I developed a software tool called ARTDeco (automatic readthrough transcription detection). I demonstrated that ARTDeco is a useful tool for quantifying readthrough transcription in a system and can identify read-in genes. I then applied ARTDeco to discover readthrough in several viruses and test the mechanism of readthrough in Rift Valley Fever Virus by expressing the NSs protein in THP-1 monocytes. Finally, I developed a novel method for inferring TF motif co-occurrences using successive rounds of motif enrichment. I show that this method recapitulates known physical interactions between TFs and that networks generated using this method have a unique biological interpretation.

CHAPTER 2: ARTDECO: AUTOMATIC READTHROUGH TRANSCRIPTION DETECTION

2.1 Abstract

2.1.1 Background:

Mounting evidence suggests several diseases and biological processes target transcription termination to misregulate gene expression. Disruption of transcription termination leads to readthrough transcription past the 3′ end of genes, which can result in novel transcripts, changes in epigenetic states and altered 3D genome structure.

2.1.2 Results:

We developed Automatic Readthrough Transcription Detection (ARTDeco), a tool to detect and analyze multiple features of readthrough transcription from RNAseq and other next-generation sequencing (NGS) assays that profile transcriptional activity. ARTDeco robustly quantifies the global severity of readthrough phenotypes, and reliably identifies individual genes that fail to terminate (readthrough genes), are aberrantly transcribed due to upstream termination failure (read-in genes), and novel transcripts created as a result of readthrough (downstream of gene or DoG transcripts). We used ARTDeco to characterize readthrough transcription observed during influenza A virus (IAV) infection, validating its specificity and sensitivity by comparing its performance in samples infected with a mutant virus that fails to block transcription termination.

We verify ARTDeco's ability to detect readthrough as well as identify read-in genes from different experimental assays across multiple experimental systems with known defects in transcriptional termination, and show how these results can be leveraged to improve the interpretation of gene expression and downstream analysis. Applying ARTDeco to a gene expression data set from IAV- infected monocytes from different donors, we find strong evidence that read-in gene-associated expression quantitative trait loci (eQTLs) likely regulate genes upstream of read-in genes. This indicates that taking readthrough transcription into account is important for the interpretation of eQTLs in systems where transcription termination is blocked.

2.1.3 Conclusions:

ARTDeco aids researchers investigating readthrough transcription in a variety of systems and contexts.

Keywords: Readthrough transcription, Transcription termination, Transcriptomics, Gene expression, Next-generation sequencing analysis

2.2 Background

Transcription termination is a fundamental step in gene expression regulation. For most genes, transcription termination is triggered when RNA polymerase II (RNAPII) transcribes a polyadenylation site (PAS) that activates the cleavage and polyadenylation (CPA) complex associated with the C-terminal domain (CTD) of RNAPII (Licatalosi et al. 2002). There are two popular models for how CPA recruitment induces transcription termination. In the allosteric model, recruitment of CPA is accompanied by a conformational change in elongating RNAPII, causing dissociation from the DNA and release of the nascent pre-mRNA (H. Zhang, Rigo, and Martinson 2015). In the torpedo model, polyA-dependent cleavage of pre-mRNA by CPA leaves an uncapped nascent RNA emanating from elongating RNAPII. The exonuclease XRN2 degrades the unprotected nascent transcript until it catches up to transcribing RNAPII, causing its release from the DNA (Kim et al. 2004; West, Gromak, and Proudfoot 2004). Alternative transcription termination mechanisms have been described for histone genes, snRNAs, and transcripts generated by RNAPI and RNAPIII (Kawauchi et al. 2008; Nielsen, Yuzenkova, and Zenkin 2013; Richard and Manley 2009).

Recent studies have demonstrated that cellular stress can disrupt normal transcription termination, leading to aberrant transcription of intergenic regions downstream of canonical termination sites (termed readthrough transcription or downstream of gene [DoG] transcription) through an unknown mechanism (pictured in Figure 2.1A). These stresses include heat shock, osmotic stress, hypoxia, influenza A virus (IAV) infection, herpes simplex virus 1 (HSV-1) infection, senescence, and cancer (Bauer et al. 2018; Cardiello, Goodrich, and Kugel 2018; Grosso et al. 2015; Heinz et al. 2018; Hennig et al. 2018; Muniz et al. 2017; Rutkowski et al. 2015; Vilborg et al. 2015, 2017). In addition to exerting cellular stress, IAV expresses the viral non-structural protein 1 (NS1), which by itself can induce readthrough transcription, presumably by inactivating the poly(A) signal-recognition molecule cleavage and polyadenylation specificity factor (CPSF) 30 (Nemeroff et al. 1998). This causes inhibition of CPA activity at poly(A) signal-dependent genes and leads to widespread readthrough transcription (Bauer et al. 2018; Heinz et al. 2018; Nemeroff et al. 1998).

Analyzing gene expression data from samples exhibiting evidence for readthrough transcription poses several challenges: without proper termination, both splicing and polyadenylation of the pre-mRNA may be impaired (N. Zhao et al. 2018). Size-selected RNA-sequencing (North-seq) experiments indicate that readthrough/DoG RNAs are long (> 13.5 kb) and not exported from the nucleus (Heinz et al. 2018). Similarly, HSV-1 infection leads to decreased signal for readthrough transcripts in cytoplasmic RNA relative to both total and nuclear RNA (Hennig et al. 2018). Ribosome profiling in HSV-1-infected cells indicates that readthrough RNAs are not bound by ribosomes and thus not translated (Rutkowski et al. 2015). The observation that readthrough transcription impedes protein expression is important because RNA profiling methods are often used as proxies for gene expression in biomedical research. RNA-seq or microarray profiling in systems with readthrough transcription are therefore likely to provide incorrect estimates of protein levels.

Readthrough transcription can also impact the measurement of gene expression in genes located downstream of sites where transcription termination is inhibited. As aberrant transcription proceeds into downstream genes, RNA templated from these regions may be misinterpreted as evidence for expression of these downstream genes (e.g. FAP in Figure 2.1B) (Rutkowski et al. 2015; Heinz et al. 2018). Following Rutkowski et al., we will term these loci “read-in” genes. The regulation of read-in genes is easily misinterpreted because the RNAs produced at these loci are unlikely to be exported or translated, and their promoters and other regulatory elements do not regulate their transcript levels. Given that most functional analyses and systems-level studies rely on RNA levels as their primary approach to molecular profiling, this represents a potential source of error when analyzing systems with widespread readthrough transcription. Without correcting for read-in genes, these analyses suffer from the inclusion of aberrantly transcribed read-in genes when studying the molecular pathways and regulatory mechanisms underlying transcriptional responses.

In addition to generating non-canonical and novel transcripts, readthrough transcription can alter the epigenomic state of the genome (Cardiello, Goodrich, and Kugel 2018; Heinz et al. 2018; Hennig et al. 2018). In the case of heat shock, osmotic stress, and HSV-1 infection, it has been found that regions exhibiting transcriptional readthrough have increased chromatin accessibility (Hennig et al. 2018; Vilborg et al. 2017). Strikingly, in IAV infection, transcriptional readthrough causes dynamic changes in 3D genome structure. This phenomenon occurs as elongating RNAPII displaces cohesin, the ringlike complex that spatially constrains the strands of DNA at the base of chromatin loops (Heinz et al. 2018). In addition, IAV-induced readthrough can result in widespread changes in histone modifications and transcription factor (TF) binding site occupancy (Heinz et al. 2018).

Given the extensive impact that defects in transcription termination and readthrough transcription can have, computational tools are needed to identify and characterize their phenotypes from next-generation sequencing (NGS) profiling data. Although several studies have analyzed readthrough transcription, they have primarily used custom or ad hoc approaches (Hennig et al. 2018; Rutkowski et al. 2015; Vilborg et al. 2015, 2017; Wiesel, Sabath, and Shalgi 2018). Presently, there are two published methods designed to analyze readthrough transcription: DoGFinder, a tool that discovers and quantifies intergenic transcripts downstream of genes (DoG transcripts) (Hennig et al. 2018; Rutkowski et al. 2015; Vilborg et al. 2015, 2017; Wiesel, Sabath, and Shalgi 2018), and DogCatcher, a tool that discovers and quantifies DoGs, Antisense Downstream of Gene (ADoG), Previous of Gene Transcripts (PoGs), and Antisense Previous of Gene (APoG) transcripts in addition to being able to perform differential expression analysis on these transcripts (Melnick et al. 2019) Both tools provide a useful characterization of readthrough transcription and can aid in the discovery of systems exhibiting transcription termination defects. However, their functionality is limited to searching for aberrant transcripts in intergenic regions.

Here we present Automatic Readthrough Transcription Detection (ARTDeco), a framework for the quantification and characterization of readthrough transcription. ARTDeco expands on the functionality of existing approaches by implementing three separate strategies to quantify readthrough transcription by evaluating (1) the fraction of transcription starting upstream and continuing into a gene (‘read-in level’), (2) the fraction of transcription that continues past the end of genes (‘readthrough level’), and (3) detection of novel DoG transcripts created as a result of readthrough transcription (pictured schematically in Figure 2.1C). We assess the performance of ARTDeco on previously generated data for IAV infection and heat shock treatment. We also demonstrate how ARTDeco can be used to quantitatively assess readthrough transcription across large donor

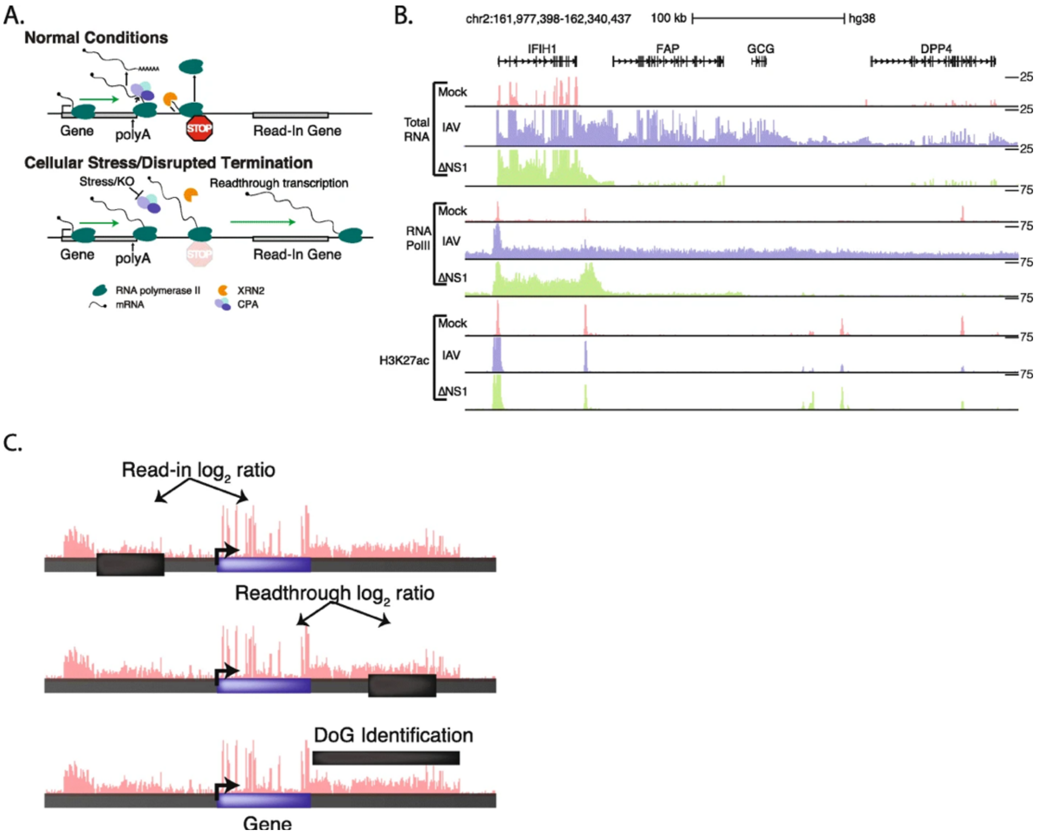


Figure 2.1. ARTDeco evaluates different aspects of readthrough transcription. **a** Schematic diagram of typical transcription termination (top) and readthrough transcription (bottom). **b** Total RNA-seq, RNA polymerase II ChIP-seq, and H3K27ac ChIP-seq data at *IFIH1* locus. Normalized read coverage ranges are indicated on the right and signals exceeding these levels may be clipped (e.g. RNA-seq coverage on the exons of *IFIH1*). *IFIH1* represents a primary induction gene while *FAP*, *GCG*, and *DPP4* represent read-in genes. **c** Schematic depicting the regions used to quantify read-in levels, readthrough levels, and DoG transcript discovery for each gene.

datasets and show that eQTLs for read-in genes likely control their upstream gene's transcription levels. We conclude that our tool is capable of quantifying key features of readthrough transcription to improve the analysis and interpretation of NGS experiments performed on samples with defects in transcription termination.

2.3 Implementation

ARTDeco is written in Python 3.6. It has the following software dependencies: BEDOPS (Neph et al. 2012), bx-python, DESeq2 (Love, Huber, and Anders 2014), HOMER (Heinz et al. 2010), NetworkX (Hagberg, A.A., Shult, D.A., Swart, P.J. 2008), NumPy (Oliphant 2006), Pandas (McKinney 2010), rpy2, RSeQC (L. Wang, Wang, and Li 2012), and Samtools (H. Li et al. 2009). Code is available at https://github.com/sjroth/ARTDeco.

2.3.1 ARTDeco analysis framework

ARTDeco requires aligned BAM files, a GTF file of gene annotations, and a chromosome sizes file. Optionally, a metadata file detailing the experimental design and a comparison file detailing the comparisons to be carried out during differential expression analysis can be supplied. The program will quantify expression at genic and intergenic regions (detailed below) and return summary statistics for readthrough transcription and DoG transcripts as well as read-in and readthrough ratios for each gene.

2.3.2 ARTDeco preprocessing

The input gene annotation (GTF file) is preprocessed into BED files representing the key genomic regions interrogated by ARTDeco. For each gene, all separate isoforms are condensed into a single region starting from the most upstream transcription start site [TSS] to most downstream transcription termination site [TTS]) to avoid misidentifying alternative isoforms as readthrough transcripts. Intergenic regions for detecting read-in and readthrough transcription relative to each gene are then selected as outlined schematically in Figure 2.1C and Supplementary Figure 2.1B. Genes were excluded from consideration if their annotation fell within another gene. Read-in quantification regions are placed a fixed distance (as defined by the user; 1 kb by default) upstream of the most upstream TSS for each gene to avoid variation in TSS location relative to annotations. Readthrough quantification regions are placed a fixed distance (as defined by the user; 10 kb by default) downstream of each gene to avoid detection of transcription that normally occurs in the region immediately 3′ of the poly(A) signal-dependent cleavage site. The default length of each read-in/readthrough detection region is set to 15 kb (can be user-defined). If another gene is present in the locus, the length of the read-in/readthrough regions are truncated such that they extend a maximum of one-third of the distance to the next gene to avoid detecting signal originating from the other gene. Thus, the length of the read-in and readthrough regions can be expressed as min (maxLength, 1/3*geneDist) where maxLength is the maximum length of a rea-din/readthrough region (15 kb by default) and geneDist is the distance to the upstream or downstream gene. The minimum length of both read-in and readthrough regions can be user-defined and is 100 bp by default. If genes are overlapping or too close in proximity, the readthrough/read-in region is removed and not reported for that gene. If one gene falls within the gene body of another gene (as is the case with many small RNAs), that gene is removed from consideration by ARTDeco. Inclusion of these genes leads to issues in interpretation and potential errors due to annotation rather than biological phenomena. Both read-in and readthrough regions are placed into BED files for downstream processing.

2.3.3 ARTDeco expression quantification

ARTDeco quantifies gene expression (both raw counts and FPKM) using HOMER's analyzeRepeats.pl and the user-supplied GTF file as well as expression at intergenic regions using HOMER's annotatePeaks.pl (Heinz et al. 2010). Expression is quantified across the whole gene body for each transcript in the GTF file and the most highly (maximum) expressed isoform (in FPKM) is stored for downstream processing of read-in and readthrough levels.

2.3.4 ARTDeco read-in and readthrough level quantification

For each gene, the expression in both raw counts and FPKM for both the maximum isoform of the gene and the intergenic region of interest are grouped together. Then, the log2 ratio of length-normalized counts is computed between the isoform and the read-in/readthrough region (outlined in Figure 2.1C and Supplementary Figure 2.1B). These ratios define the read-in and readthrough levels for each gene. ARTDeco then infers read-in genes based upon a user-defined threshold for read-in level (0 by default) as well as a user-defined expression threshold level (0.25 FPKM by default) to exclude genes with minimal expression. ARTDeco summarizes the basic statistics of read-in and readthrough levels for the most expressed genes (top 1000 by default).

2.3.5 ARTDeco gene expression deconvolution

ARTDeco can correct deconvolute the contribution of upstream readthrough transcription to total gene expression by using the upstream read-in expression. In order to do this, it subtracts the length-normalized raw expression in the read-in region from the length-normalized raw gene body expression. If the read-in region has higher expression than the gene body, the gene body expression is set to 0.

2.3.6 Combining read-in levels with differential expression information

Expression information can be combined with differential expression analysis as performed by DESeq2 (Love, Huber, and Anders 2014) to discriminate genes that are directly induced (termed “primary induction”) from those induced as a consequence of read-in transcription from upstream genes (termed “read-in”). This can be useful for enhancing the specificity of the analysis if the experimental condition is expected to impact transcription termination. DESeq2 is carried out on all transcripts in the GTF file as quantified by ARTDeco and this information is combined with read-in ratios for each gene. Genes are thresholded based upon log2 fold change (default is 2), adjusted p-value (BenjaminiHochberg correction as performed by DESeq2; default is 0.05), and expression in FPKM (default is 0.25) and categorized as a primary induction or read-in gene based upon read-in levels (default is 0).

2.3.7 ARTDeco DoG detection

ARTDeco uses a rolling window approach beginning at the TTS of each gene as defined by our condensed gene annotation. Over each window of the user-specified length (500 bp by default), transcription levels are quantified and the FPKM of the window must meet a user-specified threshold to be considered part of a DoG (0.15 FPKM by default). A DoG can be extended beyond a downstream gene’s TSS if that gene is labeled a read-in gene. After DoGs are discovered for each experiment, their expression is obtained (raw and FPKM). Then, they are combined into a single annotation by taking the union wherein the longest DoG annotation is kept for shared DoGs across experiments. The expression of the unified set of DoGs and their differential expression (if applicable) is also reported (raw and FPKM).

2.4 Results

ARTDeco processes NGS data (e.g., RNA-seq) to characterize the features of readthrough transcription genome-wide. This includes the identification of genes that exhibit transcription downstream of their 3′ ends (readthrough genes), genes that are transcribed as a result of readthrough transcription from upstream genes (read-in genes), as well as detection of novel DoG transcripts created as a result of readthrough transcription. The basic workflow of ARTDeco is detailed in Supplementary Figure 2.1A. ARTDeco can work with custom gene annotations and custom genomes. ARTDeco detects read-through events by comparing the levels of transcription in genic and intergenic regions for all genes, evaluating signal both upstream and downstream of genes to distinguish readthrough and read-in events. The intervals used to calculate intergenic transcription levels exclude regions immediately upstream of the transcription start site (TSS, > 1 kb) and downstream of the transcription termination site (TTS, > 10 kb) to avoid detection of RNA signal that arises from incorrect TSS assignment and post-poly(A) site cleavage transcripts that may accumulate during normal termination, respectively. Because closely spaced genes (< 10 kb distance between gene ends) limit the ability to infer intergenic expression levels, these genes are excluded from the analysis. The log2 transcript signal ratio of the read-in or readthrough regions versus the gene body expression can be used as a quantification of the degree of readthrough upstream (read-in level) or downstream (readthrough level) of a gene, respectively (Figure 2.1C, Supplementary Figure 2.1B). In studies where a specific experimental condition is suspected to induce transcription readthrough, ARTDeco can combine its analysis strategy with differential expression analysis to discriminate between genes that are likely regulated by primary induction (i.e. promoter activation) versus read-in genes among all induced genes. ARTDeco also detects unannotated DoG transcripts using a rolling window approach with a minimal FPKM threshold beginning at the TTS of each gene, similar to DoGFinder (Wiesel, Sabath, and Shalgi 2018).

2.4.1 Global quantification of read-through

To evaluate ARTDeco's ability to quantify transcriptional readthrough across multiple experiments, we analyzed previously generated transcriptomic and epigenetic data from monocyte-derived macrophages infected ex vivo with two strains of IAV as well as a mock infection condition (example of data in Figure 2.1B) (Heinz et al. 2018). The first influenza strain is the highly pathogenic IAV (subtype H5N1) virus (Influenza A/Vietnam/1203/2004 (H5N1) HAlo) used to model severe disease with an intact NS1 protein (called IAV here). The second strain has the same viral genetic background but is mutated to produce a truncated, non-functional NS1 protein (ΔNS1) (Heinz et al. 2018; Steel et al. 2009). These two strains induce a similar antiviral transcriptional response in the cell, but only IAV infection expresses an intact NS1 protein capable of inhibiting the CPA complex, leading to readthrough transcription. In effect, the ΔNS1 strain allows us to examine antiviral response activation without readthrough while the mock condition has neither antiviral response nor readthrough. This allows us to differentiate antiviral response transcription from readthrough transcription during IAV infection.

First, ARTDeco quantifies the global level of readthrough transcription in each sample, by calculating the genome-wide distributions of read-in and readthrough ratios for the top 1000 expressed genes (Figure 2.2A,B). We found that the distributions of both read-in and readthrough ratios were shifted to higher values in the IAV samples relative to both ΔNS1 or mock infection (Figure 2.2A,B).

Because transcription levels still decay after the cleavage site even when termination is inhibited, readthrough levels, which are measuring the signal produced by readthrough transcription at sites directly downstream of where termination is inhibited, often have a more pronounced signal than read-in levels, which are measured upstream of the next gene, 83,649 bp downstream of the TTS on average. Given that read-in transcription is likely mediated by readthrough transcription from adjacent genes, we quantified this relationship by comparing read-in levels for every expressed gene (> 0.25 FPKM) with the readthrough levels of their upstream gene, finding that these two values were significantly correlated (Figure 2.2C, $r = 0.55$; $p < 1e-151$). This result is quantitatively and qualitatively consistent with the hypothesized relationship between read-in levels and the readthrough levels of the upstream gene. In all, this confirms the ability of ARTDeco to use read-in and readthrough levels to quantify readthrough transcription.

Because read-in levels are defined as the log2 ratio of upstream readthrough transcription to genic transcription, they represent the relative contribution of readthrough to gene expression. Given this observation, we investigated whether read-in levels could potentially aid in deconvoluting the relative contributions of readthrough transcription and canonical gene activation to expression level. We examined all upregulated differentially expressed genes in the IAV condition relative to the mock condition and compared their expression

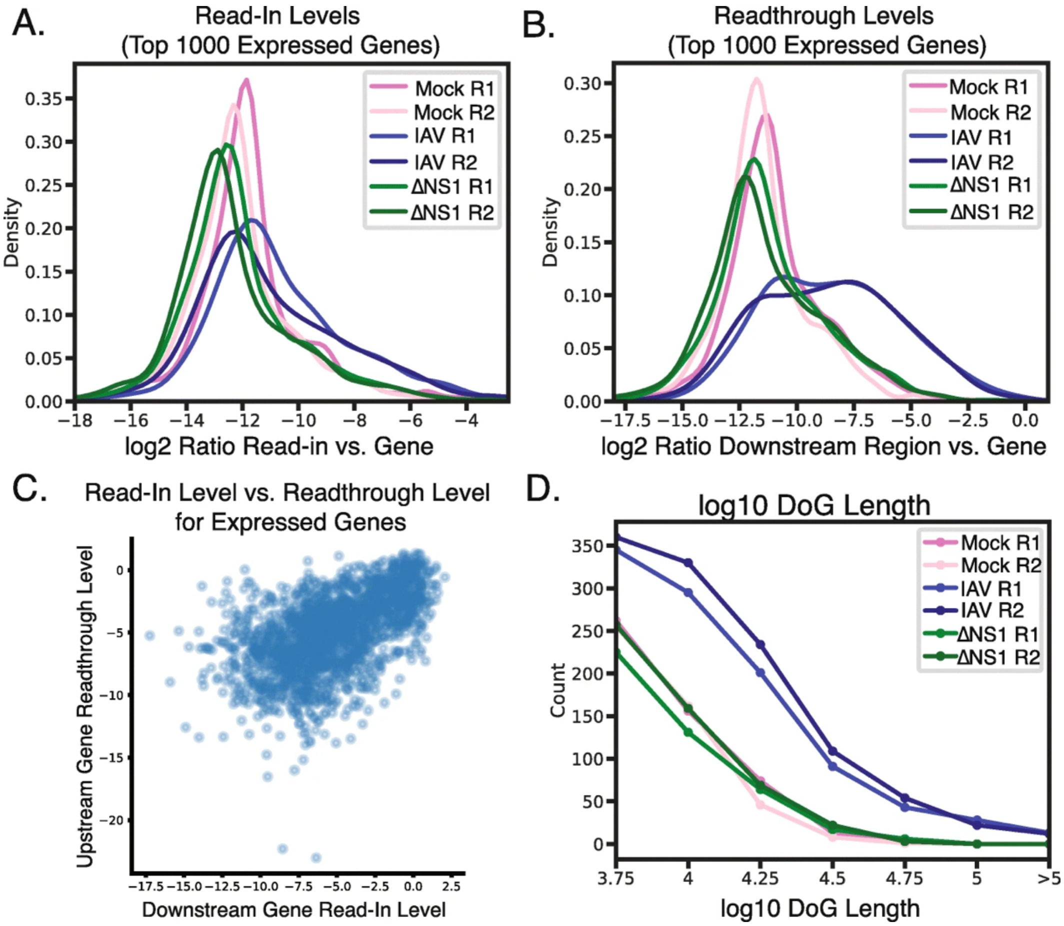


Figure 2.2. Quantification of readthrough phenotypes in IAV-infected monocyte derived macrophages. **a** Distribution of read-in levels (log2 ratio of reads in read-in region vs. gene body) for top 1000 expressed genes. **b** Distribution of readthrough levels (log2 ratio of reads in downstream region vs. gene body) for top 1000 expressed genes. **c** Downstream gene read-in level vs. upstream gene read-in level in the first replicate of the IAV condition. Both downstream and upstream genes were expressed at a level > 0.25 FPKM (r = 0.55; *p* < 1e-151). **d** Distribution of DoG lengths for DoGs discovered by ARTDeco using default settings (minimum length of 4 kb, window size of 500 bp, and minimum read density of 0.15 FPKM).

values between IAV and ΔNS1 conditions (Supplementary Figure 2.2A). We found that the expression levels between these two datasets was largely correlated (r = 0.72; p < 1e-87), however, many genes were expressed more highly in the IAV condition due to read-in transcription (Supplementary Figure 2.2A). We then corrected the expression values for both conditions by the estimated fraction of reads due to readthrough and compared their expression. We found that the correlation in gene expression was increased (r = 0.81; p < 1e-127) and that this increase was statistically significant (p < 0.001; Fisher's z transformation) (Supplementary Figure 2.2B). This suggests that the read-in level provides information about the relative contribution of readthrough transcription to gene expression and indicates that ARTDeco can estimate gene expression by removing contributing upstream readthrough.

Another method of quantifying readthrough transcription is the detection of DoG transcripts. Similar to the read-in and readthrough ratios, we performed DoG transcript discovery on mock-, IAV-, and ΔNS1-infected samples (Supplementary Table 2). We found more than twice as many DoGs in IAV-infected samples than the other conditions, consistent with the global increase in readthrough caused by NS1-mediated disruption of transcription termination. Additionally, DoGs found in the IAV condition were much longer than those in the ΔNS1 or mock conditions (almost twice as long on average), which were typically less than 10 kb in length (Figure 2.2D).

In order to compare ARTDeco's ability to detect DoG transcripts to existing methods, we independently used DoGFinder and Dogcatcher to identify DoGs in the IAV condition using default parameters (Supplemental Methods). Despite differences in how transcript detection is performed between the methods, all three methods exhibited comparable sensitivity and detected many of the same DoGs

(Supplementary Figure 2.3A,B). Notably, Dogcatcher found very few unique DoGs (Supplementary Figure 2.3A,B). This is likely because Dogcatcher screens DoGs similarly to DoGFinder (i.e., using a minimum coverage) while maintaining genic reads like ARTDeco. Differentially detected DoGs between the methods are largely explained by technical differences. DoGFinder and Dogcatcher screen DoGs based upon continuous coverage (presence or absence of reads spanning a portion of the screening window). In contrast, ARTDeco extends transcripts based upon a read density threshold measured in FPKM while keeping genic reads. This leads to DoGFinder-specific transcripts in regions with low signal but continuous coverage. Conversely, ARTDeco does not remove genic reads so some DoGs may represent mis-annotation of the TTS or inefficient transcription termination. These methodological differences are reflected by DoGFinder-specific transcripts with lower expression in FPKM (the criteria for ARTDeco) while ARTDecospecific transcripts have lower per-base coverage (the criteria for DoGFinder) (Supplementary Figure 2.3D,E).

In order to validate ARTDeco's ability to detect DoGs, we looked for independent evidence for transcription of DoGs by examining the levels of H3K36me3 and RNAPII phosphorylated on serine 2 of the CTD (RNAPII S2p) at DoG loci. Both H3K36me3 and RNAPII S2p are associated with transcription elongation, and should be enriched in readthrough regions relative to non-transcribed regions. Because ARTDeco and DoGFinder discovered the most distinct DoGs individually and Dogcatcher discovered very few unique DoGs (only 6; Supplementary Figure 2.3B), we chose to compare DoGs from ARTDeco and

DoGFinder. We found that DoGs shared between ARTDeco and DoGFinder had comparable occupancy of both signals while DoGs unique to DoGFinder had decreased signal (Supplementary Figure 2.3F,G). In summary, we find that ARTDeco has sensitivity comparable to DoGFinder and Dogcatcher and confirmed that the DoGs identified show evidence of transcription elongation.

2.4.2 Identification of read-in genes

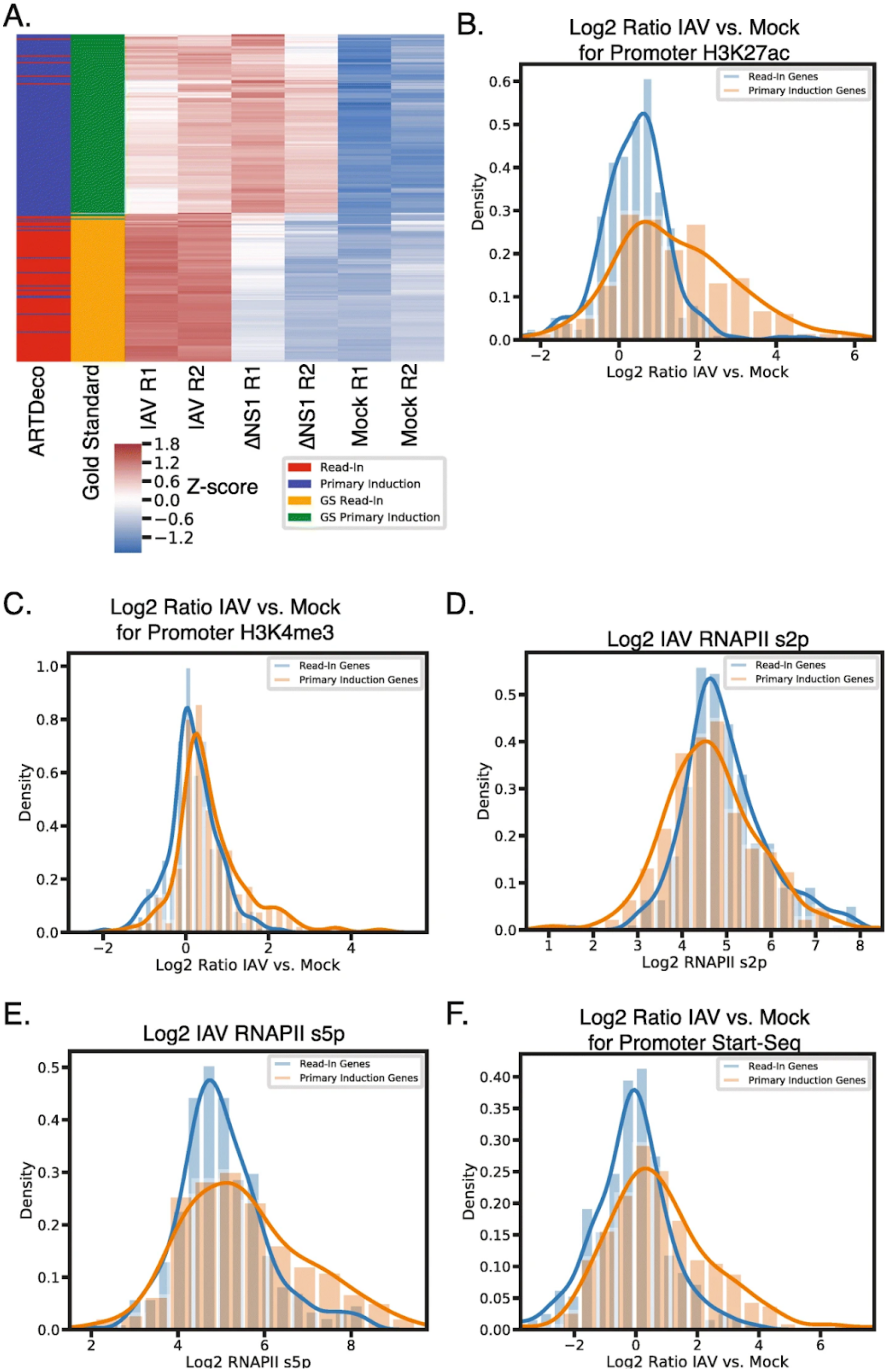
Because pre-mRNAs produced as a result of readthrough transcription are generally not exported from the nucleus and are unlikely to be translated (Heinz et al. 2018; Hennig et al. 2018; Vilborg et al. 2015), differential RNA levels in samples with readthrough transcription likely misrepresent gene expression levels of newly transcribed genes and may confound functional analyses. Furthermore, readthrough transcription can continue far past the 3′ end of transcribed genes leading to the increase of RNA signal at downstream “read-in” genes. This leads to the illusion that read-in genes are regulated by the biological process being studied. One of the novel functions of ARTDeco is to identify read-in genes to infer whether a given gene is “induced” by readthrough transcription (i.e. read-in) or if it is directly targeted for induction by the cell’s regulatory machinery (referred to here as ‘primary induction’ genes).

We sought to test the ability of ARTDeco to discriminate between primary induction and read-in genes among genes induced by IAV. In order to benchmark our method, we curated a gold standard set of primary induction and read-in genes based on differences in induction in the wild-type IAV and ΔNS1 viruses

(Supplemental Methods; Figure 2.3B). We considered gold standard primary induction genes to be upregulated in IAV relative to mock infection with clear signs of promoter activation in H3K27ac and RNAPII ChIP-seq data (Supplemental Methods; Supplemental Table 1; example Supplementary Figure 2.4A). Similarly, we considered gold standard read-in genes to be upregulated in IAV relative to both mock and ΔNS1 (log2 fold change > 2 and adjusted p-value < 0.05 according to DESeq2) with no signs of promoter activation (Supplemental Methods; Supplemental Table 1; example Supplementary Figure 2.4A). In total, there were 163 gold standard primary induction genes and 135 gold standard read-in genes (Supplemental Table 1).

ARTDeco was able to identify IAV primary induction and read-in genes with an F1 score (a measure of the accuracy of classification computed by taking the harmonic mean of the precision and recall; Supplemental Methods) of 0.95 relative to our gold standard. ARTDeco's performance when inferring read-in genes was robust to different parameters, but optimal when upregulated genes had a log2 fold change > 2, adjusted p-value < 0.05 and read-in level > − 2 (for all genes with expression > 0.25 FPKM; number of Gold Standard [GS] Primary Induction Genes = 163, number of GS Read-In Genes = 130, True Positives [TP] = 118, True Negative [TN] = 158, False Positive [FP] = 5, False Negative [FN] = 12) (Supplementary Figure 2.4C,D). We also found that ARTDeco was able to infer read-in genes on single experiments without differential expression information

Figure 2.3. ARTDeco successfully discriminates between genes that are directly induced by IAV infection (primary induction) and genes induced as a result of readthrough transcription (read-in). **a** Heatmap of *z*-normalized expression values and ARTDeco assignments for gold standard primary induction and read-in genes. Thresholds for assigning read-in genes were log2 fold change > 2, adjusted *p*-value < 0.05, and read-in level > − 2. Leftmost column is ARTDeco assignment (blue is primary induction and red is read-in). Next column is gold standard assignment (green is primary induction and gold is read-in). Remaining columns are z-normalized gene expression for IAV replicate 1, IAV replicate 2, ΔNS1 replicate 1, ΔNS1 replicate2, mock replicate 1, and mock replicate 2. **b** Distribution of log2 ratio of H3K27ac for IAV vs. Mock conditions at promoters for primary induction and read-in genes. (*p* < 1e-20; t-test) **c** Distribution of log2 ratio of H3K4me3 for IAV vs. Mock conditions at promoters for primary induction and read-in genes. (*p* < 1e-10; t-test) **d** Distribution of RNA PolII serine-2 phosphorylation (S2p) at promoters in the IAV condition for primary induction and read-in genes. (*p* < 0.001; t-test) **e** Distribution of RNA PolII serine-5 phosphorylation (S5p) in the IAV condition at promoters for primary induction and read-in genes. (*p* < 1e-5; t-test) **f** Distribution of log2 ratio of Start-seq signal for IAV vs. Mock at promoters for primary induction and read-in genes. (*p* < 1e-14; t-test).


A.
ARTDeco
Gold Standard
IAV R1
IAV R2
ΔNS1 R1
ΔNS1 R2
Mock R1
Mock R2
1.8
1.2
0.6
0.0
−0.6
−1.2
Z-score
Read-In
Primary Induction
GS Read-In
GS Primary Induction
B.
Log2 Ratio IAV vs. Mock for Promoter H3K27ac
Read-In Genes
Primary Induction Genes
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
−2
0
2
4
6
Log2 Ratio IAV vs. Mock
C.
Log2 Ratio IAV vs. Mock for Promoter H3K4me3
Read-In Genes
Primary Induction Genes
Density
0.0
0.2
0.4
0.6
0.8
1.0
−2
0
2
4
Log2 Ratio IAV vs. Mock
D.
Log2 IAV RNAPII s2p
Read-In Genes
Primary Induction Genes
Density
0.0
0.1
0.2
0.3
0.4
0.5
1
2
3
4
5
6
7
8
Log2 RNAPII s2p
E.
Log2 IAV RNAPII s5p
Read-In Genes
Primary Induction Genes
Density
0.0
0.1
0.2
0.3
0.4
0.5
2
4
6
8
Log2 RNAPII s5p
F.
Log2 Ratio IAV vs. Mock for Promoter Start-Seq
Read-In Genes
Primary Induction Genes
Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
−2
0
2
4
6
Log2 Ratio IAV vs. Mock

and thresholding only on read-in levels (Supplemental Methods; Supplementary Figure 2.5A, optimal performance using a read-in level > − 1). Performance was generally poorer when not including differential expression information due to an increase in false positives as reflected in the false discovery rate (FDR) (0.04 with differential expression vs. 0.44 without differential expression) (Supplementary Figure 2.5A; F1 = 0.67; GS Primary Induction Genes = 4188, GS Read-In Genes = 128, TP = 105, TN = 4106, FP = 82, FN = 23). One source of false positives were a result of ARTDeco detecting readthrough transcription in the read-in region despite no significant change in genic expression in IAV relative to either mock or ΔNS1 and signs of promoter activation in the downstream gene (ex. MON2 in Supplementary Figure 2.5B). The use of differential expression also helps filter the number of genes considered and, thus, limits potential exposure to errors due to incorrect gene annotations. Based upon this, we conclude that the addition of differential expression allows ARTDeco to improve specificity in experimental designs where readthrough transcription is expected to be regulated in a specific condition.

After using the above parameters (log2 fold change > 2, adjusted p-value < 0.05, and read-in level > − 2) to infer read-in genes with differential expression information, we sought independent validation of our inference. We clustered gene expression profiles for all gold standard genes and found that gene assignments showed expected expression patterns (i.e., true positives [read-in genes] were expressed exclusively in IAV while true negatives [primary induction genes] were expressed in both IAV and ΔNS1 but not in mock) (Figure 2.3B). Because read-in genes are transcribed as a result of upstream expression rather than transcription initiation, we hypothesized that promoters of read-in genes would show decreased signs of promoter activation and transcription initiation relative to primary induction genes. As expected, promoters of primary induction genes were enriched for both H3K27ac and H3K4me3 (epigenomic signals associated with promoter activation) in IAV relative to mock while the promoters of read-in genes were not (Figure 2.3B,C). Similarly, we examined the phosphorylation state of RNAPII at promoters. Primary induction genes showed higher RNAPII serine-5 phosphorylation (S5p) (a mark of transcription initiation) occupancy at promoters while read-in genes showed higher RNAPII serine-2 phosphorylation (S2p) (a mark of transcription elongation) occupancy (Figure 2.3D,E). These data are consistent with the hypothesis that the promoters of primary induction genes are activated by IAV while the promoters of read-in genes are not.

In order to assess whether the promoters of primary induction genes showed more evidence of transcription initiation than those of read-in genes, we also examined Startseq data at promoters in both IAV- and mock-infected THP-1 cells (a human monocytic cell line) (Heinz et al. 2018). Start-seq captures newly initiating short RNAs that approximate rates of transcription initiation at TSSs (Scruggs et al. 2015). We observed increased signals of transcription initiation at promoters of primary induction genes as compared to read-in genes despite differences in cell type (Figure 2.3F). This further strengthens the conclusion that primary induction genes represent a stimulus-specific response while read-in genes are expressed due to upstream readthrough transcription rather than promoter activation. In all, these data show that ARTDeco is able to discriminate between primary induction and read-in genes in a set of differentially expressed genes.

2.4.3 Functional analysis of primary induction and read-in genes

Read-in genes represent over half (301/545) of all upregulated genes despite not being directly activated by IAV infection (Figure 2.4A). Given these read-in genes are not directly targeted for activation by the host transcriptional machinery and likely not expressed as proteins, it is possible that these genes represent biological noise and could dilute the results of functional analyses. With this in mind, we assessed the impact of read-in genes on common functional analyses such as gene ontology (GO) enrichment (Ashburner et al. 2000). Assessing GO enrichment separately on primary induction and read-in genes, we found that primary induction genes were strongly enriched for GO terms consistent with viral defense and immune response. In contrast, read-in genes showed minimal evidence for GO term enrichment, consistent with the hypothesis that read-in genes represent transcriptional noise. (Figure 2.4C). We also compared these enrichments with the GO enrichment for all upregulated genes, finding that inclusion of read-in genes did not identify additional enriched GO terms and diluted the fraction of regulated genes in each of the enriched terms relative to just analyzing the primary induction genes (Figure 2.4C). Given that GO is incomplete and has known biases such as method of investigation, curation practices, and authorship, it is possible that read-in genes are not properly functionally annotated

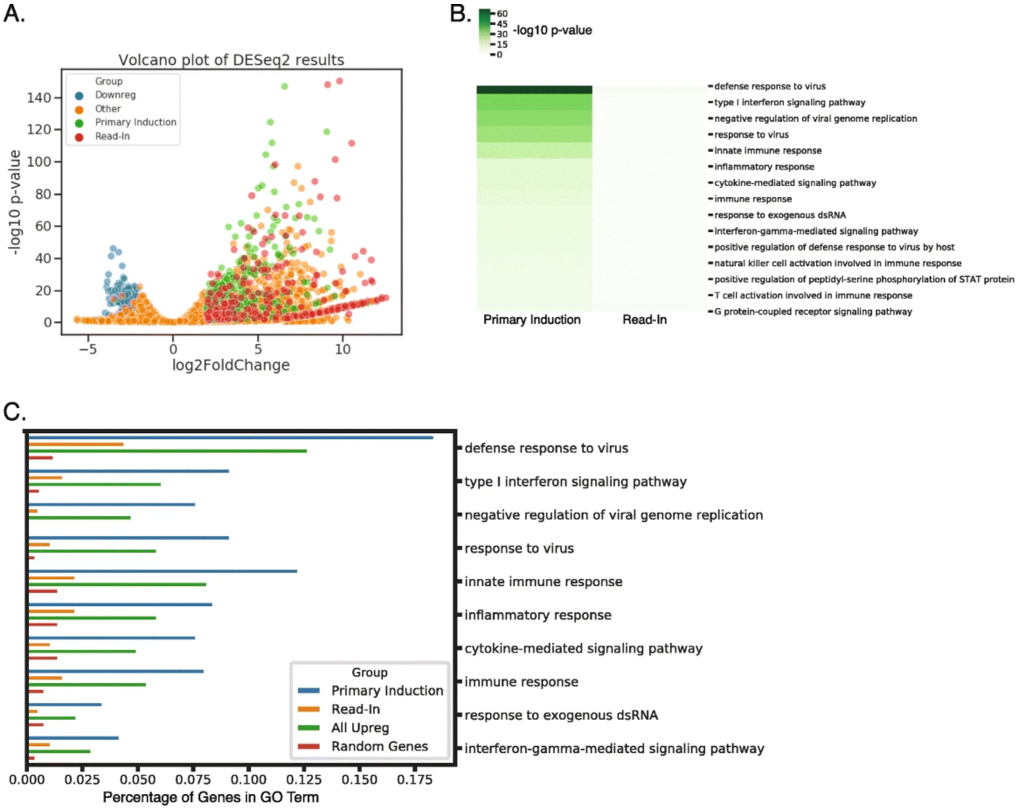
(Altenhoff et al. 2012; Thomas et al. 2012). With this in mind, we analyzed the TF binding motifs in the promoters of primary induction and read-in genes, reasoning that promoter sequences directly activated by the infection should be enriched for binding motifs for TFs activated during viral infection. We performed motif-finding using HOMER and found that promoters of primary induction genes were enriched for interferon-stimulated response elements (ISRE) while promoters of read-in genes lacked significant enrichment for TF binding motifs (Figure 2.4D). Together, our findings suggest that read-in genes are not directly activated as part of the immune response to infection and therefore should be excluded from functional or regulatory element analysis when attempting to infer regulatory mechanisms or functional responses in systems with readthrough transcription.

2.4.4 Extension of ARTDeco to other experimental systems and NGS data types

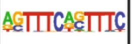
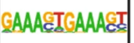
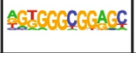
In order to validate ARTDeco on non-IAV datasets, we reanalyzed data from heat shock-treatment of NIH 3T3 cells (Vilborg et al. 2017), another stimulus known to induce transcriptional readthrough (Figure 2.5A). Similar to IAV data, we observed that all global signals of readthrough were elevated (i.e., distribution of read-in/readthrough level, DoG length, and DoG expression) (Figure 2.5B-D). Next, we assigned primary induction and read-in genes for the heat shock data. Similar to IAV, for primary induction genes we found significant GO term and TF motif enrichment that was consistent with a heat shock response while no significant enrichment was found for read-in genes (Figure 2.5E,F). These results demonstrate that ARTDeco can successfully identify transcriptional readthrough and define primary and read-in gene sets in additional datasets, using the optimized default parameters determined in IAV infection.

In order to demonstrate the flexibility and general applicability of ARTDeco to different experimental data types, we applied it to two methods that assess transcription by measuring RNAPII engagement: RNAPII ChIP-seq and mNET-seq. RNAPII ChIP-seq directly measures DNA binding of the RNAPII complex, while mNET-seq measures nascent transcripts that are associated with the RNAPII complex (Nojima et al. 2015). First, we applied ARTDeco to RNAPII ChIP-seq data from IAV, ΔNS1-, and mock-infected cells (Supplementary Figure 2.6A). Consistent with previous analyses, the distribution of readthrough levels reflects a defect in termination present in IAV infected samples but not the other two conditions, similar to the results generated using total RNA-seq, despite the different data type (Figure 2.2B, Supplementary Figure 2.6A). Additionally, we found that total RNAseq data was robust to different downstream readthrough distances while RNAPII ChIP-seq was not (Figure 2.2A, Supplementary Figure 2.6A-C). Distributions of readthrough levels with a 5 kb distance were more similar between conditions and readthrough was therefore harder to detect on a global level compared to analysis using a 10 kb distance (Supplementary Figure 2.6A,B). Thus, ARTDeco's default parameter of a 10 kb downstream readthrough distance is flexible with respect to data type. Next, we applied ARTDeco to a published data set that used mNET-seq to profile transcription in response to influenza infection (IAV H1N1 WSN/33, IAV H1N1 Puerto Rico/8/34, IAV H3N2 Udorn/72, IAV H3N2

Figure 2.4. Read-in genes mainly contribute noise to downstream functional analysis of differentially regulated genes. **a** Volcano plot of DESeq2 results for the maximum expressed isoform for each gene from IAV-infected macrophages vs. Mock-infected controls. Genes were considered up- or down-regulated if they had |log2 fold change| > 2 and adjusted p-value < 0.05 as well as FPKM > 0.25. Genes were considered primary induction if they were upregulated by this standard and had read-in levels < -2. Read-in genes were similarly upregulated but had read-in levels > -2. **b** Heat map of GO enrichment (−log10 p-value) for top 15 GO terms in primary induction and read-in genes. **c** Bar chart of proportion of genes from each gene list in a given GO term for top 10 GO terms for primary induction genes, read-in genes, all upregulated genes, and random 500 genes with expression > 0.25 FPKM. **d** HOMER motif enrichments (q-value) for top 3 known motifs in primary induction genes expression).



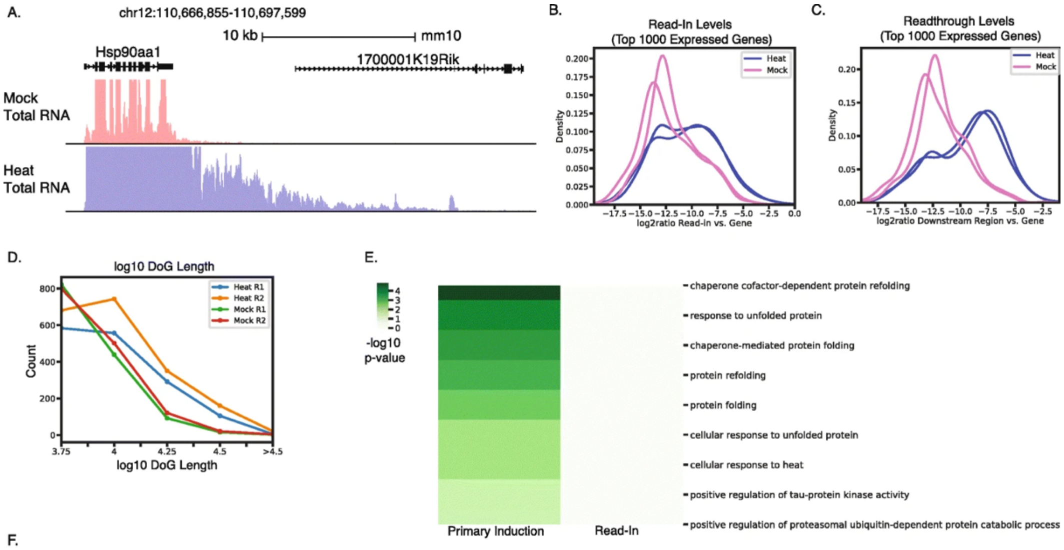
D.

Name	Motif	Primary Induction % of Target Sequence	Primary Induction % of Background Sequence	Primary Induction q-value	Read-In % of Target Sequence	Read-In % of Background Sequence	Read-In q-value
ISRE		33.66%	5.55%	0.000	7.46%	5.54%	1
IRF1		24.75%	2.87%	0.000	4.48%	2.80%	1
Sp5		61.39%	41.72%	0.003	29.85%	33.88%	1

Udorn/72: NS1Δ99, and Influenza B virus [IBV] Florida/04/2006) as well as an siRNA construct for the CPSF complex, salt shock treatment using KCl, and inducible expression of wild-type and mutant NS1 proteins (Bauer et al. 2018).Consistent with their reported results, we found that cells infected with influenza virus, subjected to KCl treatment, or deficient in the CPSF complex had higher readthrough levels relative to cells in the mock condition, reflecting decreased transcription termination efficiency. Interestingly, we confirmed the presence of readthrough transcription in IAV H3N2, which contains a deletion in the NS1 protein (Supplementary Figure 2.6D). This is consistent with the hypothesis of (Bauer et al. 2018) that cellular stress may drive part of the readthrough phenotype in A549 and HEK293 cells. In summary, we show that ARTDeco is compatible with multiple NGS data types with different characteristics.

2.4.5 Reinterpretation of eQTLs identified in data with readthrough transcription

To demonstrate how ARTDeco can improve the analysis of large-scale datasets that exhibit signs of readthrough transcription, we used ARTDeco to reanalyze RNA-seq profiles from primary human monocytes derived from 200 individual donors. Within the original study, monocytes from each donor were genotyped and infected with IAV (H1N1 strain A/USSR/90/1977) or stimulated with individual donors. Within the original study, monocytes from each donor were genotyped and infected with IAV (H1N1 strain A/USSR/90/1977) or stimulated with lipopolysaccharide (LPS), Pam3CSK4, or R848 in vitro to elicit innate immune responses with the goal of mapping expression quantitative trait loci (eQTLs) (Quach et al. 2016). We assessed the presence of readthrough transcription in these datasets by quantifying the median readthrough level of the top 1000 expressed genes as a summary statistic for samples from each donor in each condition. This analysis revealed that IAV-infected samples showed significantly greater median readthrough ratios relative to the other stimuli profiled, consistent with the expected inhibition of transcription termination in samples infected with IAV (Figure 2.6A).



F.

Name	Motif	Primary Induction % of Target Sequence	Primary Induction % of Background Sequence	Primary Induction q-value	Read-In % of Target Sequence	Read-In % of Background Sequence	Read-In q-value
HRE		14.81%	3.16%	0.004	8.31%	2.82%	1

Figure 2.5. ARTDeco analysis of readthrough transcription induced by heat shock in NIH 3T3 cells. **a** Total RNA-seq levels at the *Hsp90aa1* locus in mouse fibroblasts for heat shock and mock conditions from (Vilborg et al. 2017). *Hsp90aa1* represents the primary induction genes while *1700001K19Rik* is a read-in gene defined by ARTDeco. **b** Distribution of read-in levels for top 1000 expressed genes following heat shock. **c** Distribution of readthrough levels for top 1000 expressed genes following heat shock. **d** Distribution of DoG lengths in both mock and heat shock conditions. **e** Heat map of GO term enrichment (−log10 p-value) for top 15 enriched GO terms for primary induction and read-in genes. **f** HOMER motif enrichment for primary induction and read-in genes.

While analyzing IAV samples, we observed that some samples generally had higher levels of readthrough transcription than others, prompting us to consider whether ARTDeco could be used to quantitatively assess differences in global readthrough across samples. For example, samples from donors of European origin (EUB) had significantly higher median readthrough ratios than samples from donors of African ancestry (AFB) (Figure 2.6B, $p < 1e-8$, t-test), suggesting readthrough ratios may offer a quantitative estimate of the degree to which transcription termination is impacted by infection. In order to corroborate these observations, we compared the median readthrough level from each sample to the expression of viral NS1 RNA in each sample, finding the values to be highly correlated (Figure 2.6C, $r^2 = 0.53$, $p < 1e-33$). NS1 mRNA levels are likely correlated with other aspects of infection, including the efficiency of viral entry, viral replication rates, and antiviral host responses, and it was noted in the original study that AFB samples showed higher expression of immune response genes such as chemokines and cytokines and thus were likely more resistant to infection (Quach et al. 2016). However, given the fact that NS1 is both necessary and sufficient to inhibit transcription termination (Bauer et al. 2018; Heinz et al. 2018), the correlation between readthrough transcription levels and NS1 expression is consistent with the molecular functions of the viral protein.

In view of the widespread evidence for readthrough transcription in the IAV-infected samples, we hypothesized that eQTLs that map to genes aberrantly transcribed by readthrough transcription (i.e. read-in genes) may be regulating transcription in upstream regions rather than directly controlling transcription activation of the eQTL-associated read-in gene (Figure 2.6D). Using our list of inferred primary induction and read-in genes, we reexamined eQTLs (as inferred in the original analysis) defined in IAV-infected conditions. We hypothesized that eQTLs mapping to read-in genes would also map to upstream genes that serve as the source of readthrough transcription, while eQTLs mapping to primary induction genes would be more likely to map near or within the gene itself. We found that 9/32 (28%) of eQTLs mapping to ARTDeco-defined read-in genes also mapped to their upstream genes, while none of the eQTLs mapping to primary induction genes also mapped to their upstream genes (Figure 2.6E, $p < 1e-3$, Fisher's Exact Test, Supplementary Table 3). For example, in the case of the read-in gene SCN1B, the SNP rs2651133 was also assigned as eQTL to its upstream gene, GRAMD1A, in the IAV condition (Figure 2.6F). This SNP falls near a promoter-distal enhancer upstream of GRAMD1A, where it likely influences regulatory mechanisms such as TF binding or promoter-enhancer interactions to modulate the activity of GRAMD1A. Since the promoter of SCN1B lacks epigenetic evidence for activation after IAV infection (Figure 2.6F, bottom), it is likely that the same eQTL affects the expression of SCN1B by directly modulating the expression of GRAMD1A, which then leads to readthrough transcription into the SCN1B locus. These findings underscore the need to be careful when interpreting the functions of eQTLs in the presence of readthrough transcription.

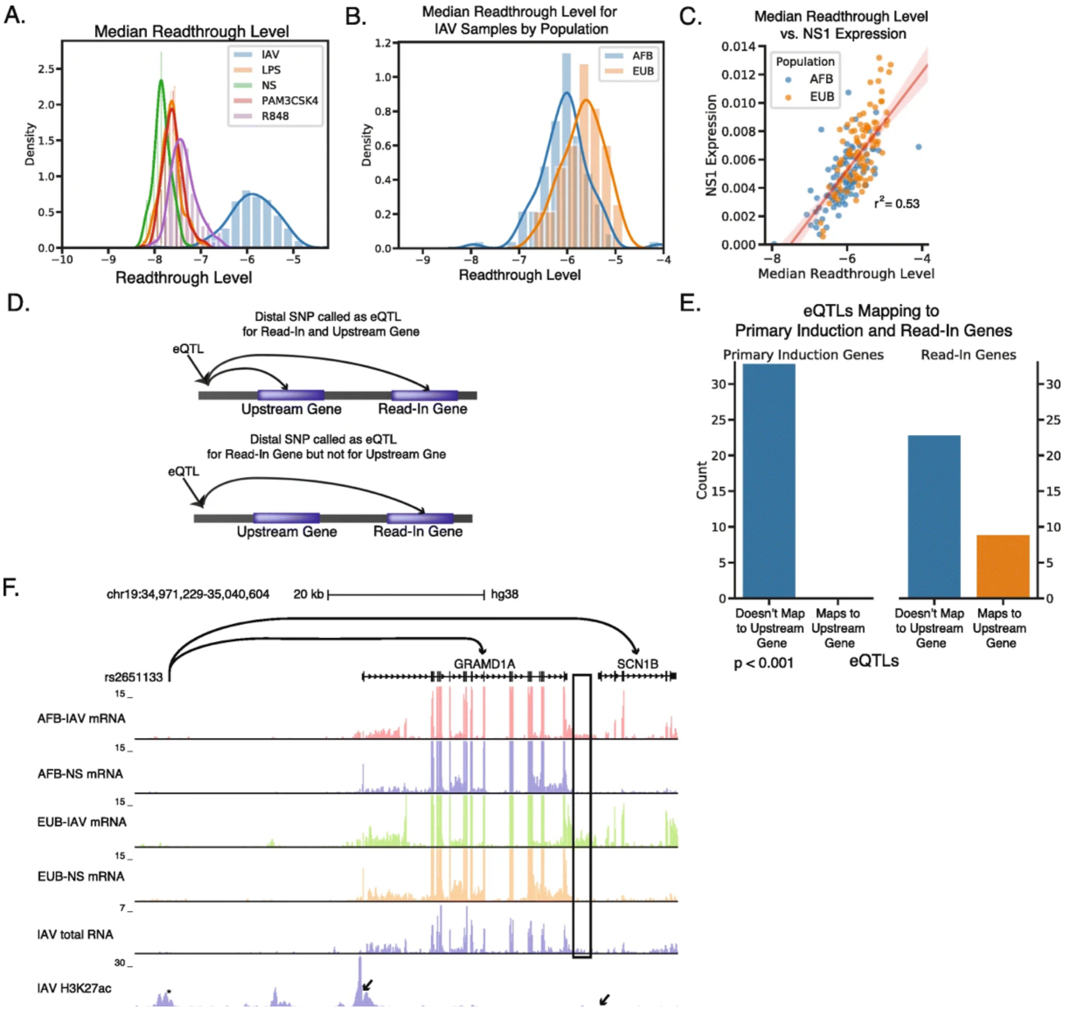
2.5 Discussion

Here we present ARTDeco, a framework for comprehensively characterizing and quantifying readthrough transcription from NGS data. ARTDeco globally quantifies the degree of readthrough transcription using read-in levels, readthrough levels, and detection of DoG transcripts. We demonstrate that the medians of the read-in and readthrough level distributions for the top-expressed genes represent useful summary statistics for characterizing the degree of readthrough in a given sample. These measures represent a novel advance in the detection of readthrough transcription. ARTDeco expands upon existing methods for DoG transcript discovery by allowing the discovered transcripts to extend into annotated gene bodies to avoid arbitrary truncation (Melnick et al. 2019; Wiesel, Sabath, and Shalgi 2018). This allows for a more precise quantification of readthrough as well as more representative transcripts from large regions of transcriptional readthrough that extend through multiple genes (Figure 2.1B). ARTDeco's approach is robust to multiple data types including RNA-seq, mNET-seq, and RNAPII ChIPseq (Figures 2.2, 2.6, Supplementary Figure 2.6) making it a versatile tool for the characterization and detection of transcriptional readthrough. Additionally, it requires less preprocessing and has a nearly 2-fold faster runtime than DoGFinder and a nearly 5-fold faster runtime than Dogcatcher (Supplemental Methods; Table 2.1). ARTDeco's flexibility and performance in addition to its novel measures of readthrough transcription represent a significant advance in analytical tools for studying defects in transcription termination.

In addition to global quantification of readthrough transcription, ARTDeco provides per-gene quantification. This provides an opportunity to study readthrough at the level of single genes in the context of both downstream readthrough and upstream read-in. The quantification of read-in levels can also enable the deconvolution of gene expression in systems with transcriptional readthrough. Additionally, each method of readthrough quantification enables us to pinpoint loci of interest in order to study the effects of readthrough on the epigenome and genome structure. Many of the mechanisms of how these changes occur are still unclear. For example, change in genome 3D structure due to transcriptional readthrough has been noted in both IAV infection and heat shock (Cardiello, Goodrich, and Kugel 2018; Heinz et al. 2018). Using readthrough levels and DoG transcripts, we may be able to better characterize the specific loci that are affected. This would lend great insight into how the mechanism of transcription induces these changes in genome 3D structure and epigenetic regulation.

An open question is what determines the level of readthrough. Work in HSV-1 infection suggests that sequence context at the TTS is a more important determinant of readthrough than expression level (Hennig et al. 2018). ARTDeco's quantification of readthrough levels could potentially lend insight to this and hint at potential mechanisms. Additionally, it has been posited that readthrough has an effect on the expression of downstream genes via mechanisms such as transcriptional interference (Greger and Proudfoot 1998; Shearwin, Callen, and Egan 2005). It remains unclear to what degree this impacts transcriptional regulation and gene expression writ large. Quantification of read-in level allows us to more directly measure these effects by elucidating the relationship between upstream readthrough transcription and gene expression.

Figure 2.6. ARTDeco analysis of donor monocytes infected with IAV reveals that eQTLs mapping to read-in genes also frequently map to upstream genes. **a** Distribution of median readthrough levels for top 1000 expressed genes for all samples from Quach et al. (2016). Grouped by treatment condition. **b** Distribution of median readthrough levels for top 1000 expressed genes for IAV samples from Quach et al. (2016). Grouped by population of origin. **c** Scatter plot comparing median readthrough level of top 1000 expressed genes with proportion of reads mapping to IAV NS1 gene (r2 = 0.53, $p < 1e\text{-}33$). **d** Schematic of two eQTL assignments that are difficult to interpret when readthrough transcription is present. On the top, a SNP is assigned as an eQTL for both the upstream gene and the read-in gene. On the bottom, a SNP located in the upstream gene is assigned as an eQTL for the read-in gene only. The first case represents eQTLs that may modulate the expression of the read-in gene by changing the expression of the upstream gene. **e** Bar chart showing the number of eQTLs mapped by Quach et al. (2016) to genes assigned as read-in and primary induction genes. eQTLs are classified as either mapping to the upstream gene as outlined in 6B or not mapping to the upstream gene. Enrichment was computed using Fisher exact test ($p < 0.001$). **f** Example of an eQTL (rs2661133) mapped by Quach et al. (2016) that maps to both a read-in gene (*SCN1B*) and the upstream gene (*GRAMD1A*) in IAV-infected samples. Genome browser tracks corresponding to mRNA from an African Belgian (AFB) and a European Belgian (EUB) from IAV-infected and non-stimulated (NS) conditions as well as total RNA and H3K27ac for IAV infection from Heinz et al. (2018). The readthrough region is outlined in the black box.


A.
Median Readthrough Level
IAV
LPS
NS
PAM3CSK4
R848
Density
Readthrough Level
B.
Median Readthrough Level for
IAV Samples by Population
AFB
EUB
Density
Readthrough Level
C.
Median Readthrough Level
vs. NS1 Expression
Population
AFB
EUB
NS1 Expression
r² = 0.53
Median Readthrough Level
D.
Distal SNP called as eQTL
for Read-In and Upstream Gene
eQTL
Upstream Gene
Read-In Gene
Distal SNP called as eQTL
for Read-In Gene but not for Upstream Gne
eQTL
Upstream Gene
Read-In Gene
E.
eQTLs Mapping to
Primary Induction and Read-In Genes
Primary Induction Genes
Read-In Genes
Count
Doesn't Map to Upstream Gene
Maps to Upstream Gene
Doesn't Map to Upstream Gene
Maps to Upstream Gene
p < 0.001
eQTLs
F.
chr19:34,971,229-35,040,604
20 kb
hg38
rs2651133
GRAMD1A
SCN1B
AFB-IAV mRNA
AFB-NS mRNA
EUB-IAV mRNA
EUB-NS mRNA
IAV total RNA
IAV H3K27ac

A novel function of ARTDeco is the identification of read-in genes. To our knowledge, it is the first software tool that is designed to characterize this phenomenon. This is important as many functional analyses rely on gene expression levels to make inferences (e.g., differential expression, co-expression, etc.) and read-in genes represent a potential source of noise when employing these techniques. We demonstrated the ability to confidently identify read-in genes from NGS profiling data, and showed that these genes likely represent noise in functional analysis when analyzing differentially regulated genes in two different conditions (IAV and heat shock). Our analyses underscore the advantage of treating these genes as noise rather than a potential false signal in the data.

We showed that in a population study of transcriptional responses to IAV infection that a significant proportion of eQTLs mapping to read-in genes also mapped to genes upstream (Figure 2.6C,D). In these cases, readthrough transcription is the probable mechanism by which the eQTL influences expression for variants mapped to read-in genes. Given the known difficulty of both mapping and interpreting the functional impact of these SNPs, it is important to correct for transcriptional readthrough when studying gene expression variation in populations in the context of systems with disrupted transcription termination. Our findings suggest that readthrough transcription analysis should be routinely incorporated into population-scale analyses of systems that may contain readthrough in order to better interpret eQTLs.

Table 2.1. Run time comparison for DoGFinder, Dogcatcher, and ARTDeco

Task	Number of Runs	Average Run Time (s)
ARTDeco Full	10	1095.76
ARTDeco DoG Mode	10	982.83
Dogcatcher Preprocessing	10	1307.25
Dogcatcher (no differential expression)	10	4085.60
Dogcatcher (with differential expression)	10	4593.81
DoGFinder Preprocessing	10	982.71
DoGFinder	10	1065.85

2.6 Conclusions

Readthrough transcription is an emergent phenotype that has been characterized in several systems including IAV infection, HSV-1 infection, heat shock, salt stress, senescence and renal carcinoma (Bauer et al. 2018; Cardiello, Goodrich, and Kugel 2018; Grosso et al. 2015; Heinz et al. 2018; Hennig et al. 2018; Muniz et al. 2017; Rutkowski et al. 2015; Vilborg et al. 2015, 2017). Given its relative novelty, it is likely that more stresses cause defects in transcription termination, and this phenotype may be more common than previously thought. The use of median readthrough level for top expressed genes as a summary statistic greatly aids discovery of these stresses. Further, ARTDeco can be used to analyze systems where components of the transcription termination machinery are knocked out in order to further analyze mechanisms of termination. In all, ARTDeco will aid future researchers by providing a systematic characterization of readthrough transcription.

2.7 Availability and requirements

Project name: ARTDeco.

Project home page: https://github.com/sjroth/ARTDeco

Operating system(s): Platform independent.

Programming language: Python.

Other requirements: Python 3.6, BEDOPS 2.4 or higher, bx-python 0.8 or higher, DESeq2 1.2 or higher, HOMER 4.9 or higher, NetworkX 2.2 or higher, NumPy 1.16 or higher, Pandas 0.24 or higher, rpy2 2.9, RSeQC 3.0 or higher, and Samtools 1.9 or higher.

License: MIT License.

Any restrictions to use by non-academics: No restrictions.

2.8 Availability of data and materials

Data from Heinz et al. (2018) was obtained from GEO accession GSE103477 (available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE103477). Data from Vilborg et al. (2017) was obtained from GEO accession GSE98906 (available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE98906). Data from Bauer et al. (2018) was obtained from NCBI SRA SRP132032 (available at https://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP132032). Data from Quach et al. (2016) was obtained from the EGA accession EGAS00001001895 (available at https://www.ebi.ac.uk/ega/studies/ EGAS00001001895).

2.9 Acknowledgements

We would like to thank Alon Goren for guidance and helpful comments. We would also like to thank Max Chang for troubleshooting ARTDeco implementations.

Chapter 2, in full, is a reprint of “ARTDeco: automatic readthrough transcription detection” as it appears in BMC Bioinformatics 2020. Samuel J. Roth, Sven Heinz, Christopher Benner. The dissertation author was the primary investigator and author of this paper.

2.10 Supplementary Methods

2.10.1 NGS data processing

Data from Heinz et al. (2018) and Vilborg et al. (2017) were obtained from GEO accessions GSE103477 and GSE98906, respectively. Data from Bauer et al. (2018) was obtained from NCBI SRA SRP132032. Data from Quach et al. (2016) was obtained from the EGA accession EGAS00001001895. Reads from these data (and all data types therein) were trimmed using Cutadapt v2.4 (Martin 2011). All RNA-seq data was aligned to reference genome using STAR v. 2.7.0d (Dobin et al. 2013). RNA-seq data was either aligned to a combined genome of hg38 and Influenza A/Vietnam/1203/2004 (H5N1) HAlo, the mm10 genome, or a combined genome of hg38 and Influenza A/A/USSR/90/1977 (H1N1) for data from Heinz et al. (2018), Vilborg et al. (2017), and Quach et al. (2016), respectively. After alignment, RNA-seq data was processed by ARTDeco (detailed below). mNETseq data from Bauer et al. (2018) was aligned to the hg38 genome and aligned files were further processed using mNET_snr (Nojima et al. 2016) prior to ARTDeco processing. ChIP-seq and Start-seq data was from Heinz et al. (2018) was aligned to the hg38 genome using Bowtie2 v. 2.3.5 (Langmead and Salzberg 2012). Tag directories and peaks were called using HOMER v. 4.10 with the exception of RNAPII data which was processed by ARTDeco (detailed below) (Heinz et al. 2010). All NGS data were visualized on the UCSC genome browser using HOMER makeMultiWigHub.pl (Heinz et al. 2010; Kent 2002).

2.10.2 ARTDeco data processing

All RNA-seq data, RNAPII ChIPseq data from Heinz et al. (2018), and mNETseq data from Bauer et al. (2018) were run through the standard ARTDeco preprocessing and quantification (outlined above). GTF files from GENCODE (Frankish et al. 2019) were used as input (hg38 v28 for data mapped to hg38 [or a combined genome containing hg38 and viral genomes] and mm10 vM17 for data mapped to mm10). Only gene types in the categories protein_coding, lincRNA, bidirectional_promoter_lncRNA, and processed_transcript as defined by GENCODE were considered for read-in gene analysis. Chromosome sizes files were generated using Samtools (H. Li et al. 2009). For total RNA-seq data from Heinz et al. (2018) and Vilborg et al. (2017), read-in genes were called with expression of >0.25 FPKM and read-in ratios of > -1 when not using differential expression information. When using differential expression information, genes were considered upregulated if they had log2 fold change > , p-value < 0.05, and FPKM > 0.25. These were assigned as read-in genes if read-in levels were > -2 for Heinz et al. (2018) and > -1 for Vilborg et al. (2017) and if they fell into the above-mentioned gene categories. Thresholds were determined based upon benchmarking Heinz et al. (2018) data (described below). DoGs were called using default parameters for all datasets.

2.10.3 Deconvolution of Gene Expression using Read-In Expression

We took all upregulated genes as called by DESeq2 in the IAV condition relative to the mock condition and compared raw and corrected gene expression values in the IAV and ΔNS1 conditions. Similar to above, only gene types in the categories protein_coding, lincRNA, bidirectional_promoter_lncRNA, and processed_transcript as defined by GENCODE were considered.

2.10.4 Benchmarking Read-In Gene Inference

We curated a set of gold standard read-in and promoter-activated/primary induction genes using differential expression output from total RNA-seq from Heinz et al. (2018). As a positive control for identifying primary induction genes, we used data from samples infected with an IAV virus that expresses a truncated NS1 protein (ΔNS1) that does not cause readthrough transcription. We expected that genes considered upregulated in both IAV and ΔNS1 samples represent primary induction genes while genes upregulated in IAV samples but not ΔNS1 samples represent read-in genes. Differential expression analysis was carried out using DESeq2 (Love, Huber, and Anders 2014) as performed in the ARTDeco pipeline. Gold standard read-in genes were defined as true positives while gold standard promoter-activated/primary induction genes were defined as true negatives for performance evaluation. A gold-standard read-in gene was defined as being upregulated in IAV relative to ΔNS1 (log2 fold change > 2 and $p < 0.05$), expression in IAV > 0.25 FPKM, and expression in ΔNS1 < 0.5 FPKM while having no promoter-proximal H3K27ac or RNAPII ChIP-seq peaks. Promoter-activated genes (true negative when not using differential expression to infer read-in genes) were defined as having expression in IAV > 0.25 FPKM and having both H3K27ac and RNAPII ChIP-seq peaks near/on the promoter. Primary induction genes (true negative when using differential expression to infer read-in genes) were defined as upregulated in both IAV and ΔNS1 relative to the mock condition (log2 fold change > 2 and $p < 0.05$) with expression in both IAV and ΔNS1 above 0.25 FPKM and having promoter-proximal H3K27ac and RNAPII ChIP-seq peaks.

We computed various measures of performance such as false positive rate (FPR), false negative rate (FNR), false discovery rate (FDR) and F1 score. These were calculated as follows:

$$PR = \frac{FP}{FP + TN}$$

$$FNR = \frac{FN}{FN + TP}$$

$$FDR = \frac{FP}{FP + TP}$$

$$F1 = \frac{2 * TP}{2 * TP + FP + FN}$$

where true positives (TP) are correctly assigned read-in genes, false positives (FP) are incorrectly assigned read-in genes, true negatives (TN) are correctly assigned primary induction genes, and false negatives (FN) are incorrectly assigned primary induction genes. We then varied parameters such as log2 fold change, p-value, and read-in level to test the ability of ARTDeco to infer read-in genes both with and without differential expression information included.

2.10.5 Functional analysis of read-in genes

Read-in genes were inferred using differential expression as described above for both the Heinz et al. (2018) and Vilborg et al. (2017) data. Gene ontology (GO) enrichment was performed using GOATOOLS (Klopfenstein et al. 2018) on read-in and primary induction genes. Additionally, motif enrichment of promoters was performed using HOMER (Heinz et al. 2010).

2.10.6 DoGFinder and Dogcatcher DoG Comparison

DoGFinder (Wiesel, Sabath, and Shalgi 2018) and Dogcatcher (Melnick et al. 2019) were run in “window mode” with a window of 500 bp and coverage of 0.6 using both IAV replicates. DoGs discovered using DoGFinder in each replicate were combined using the Union_DoGs_annotation function. DoGs discovered using Dogcatcher were combined using the 2.5_Dogcatcher_filter.py script. The characteristics of DoGs discovered by DoGFinder (i.e., identity, length, epigenomic signatures of transcription elongation) were compared to the set of combined ARTDeco DoGs for both IAV replicates in order to assess similarities and differences in transcript detection. Random DoG regions were generated using bedtools to shuffle genomic locations of DoGs discovered by ARTDeco in IAV replicates (Quinlan and Hall 2010). Per base coverage of DoGs was computed using bedtools coverage (Quinlan and Hall 2010).

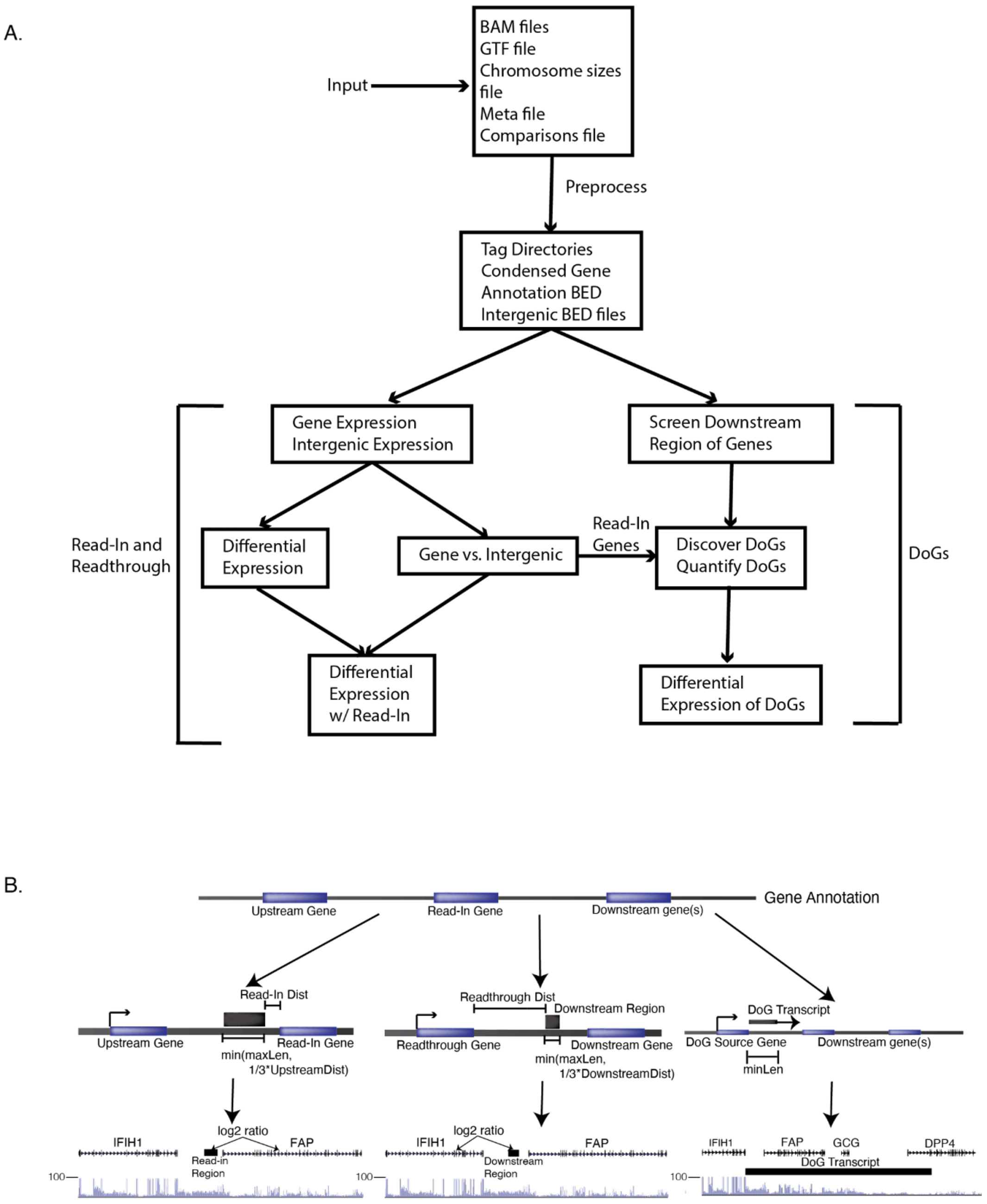
2.10.7 DoGFinder and Dogcatcher Runtime Comparison

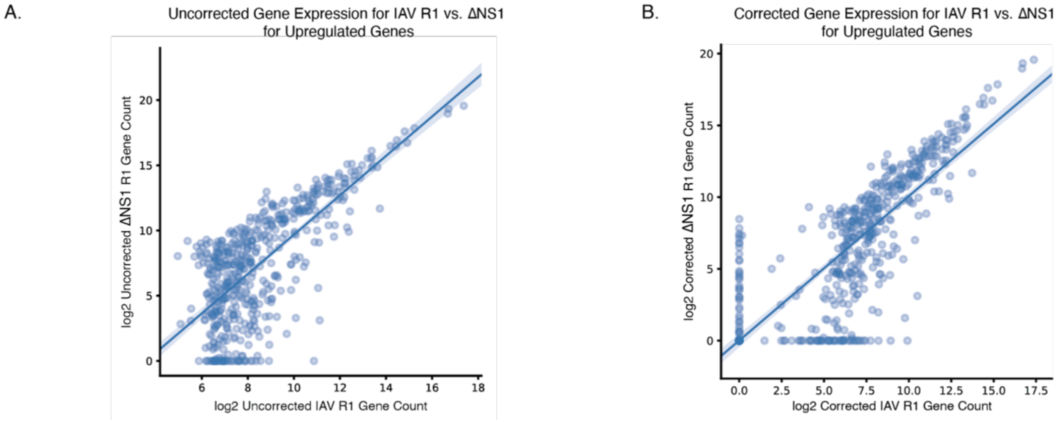
DoGFinder, Dogcatcher and ARTDeco were each run 10 times on mock, IAV, and ΔNS1 in order to assess runtime. All runs were performed on 50 Intel Xeon E5-2697 v3 @ 2.60GHz CPUs. DoGFinder was run in two stages. First, preprocessing was performed on these BAM files in order to ensure proper formatting for DoGFinder and Dogcatcher. A Snakemake workflow (Köster and Rahmann 2018) that combined custom scripts and Samtools (H. Li et al. 2009) was implemented for both DoGFinder and Dogcatcher to convert the BAM files to SAM files, switch strand orientation, sort the SAM files, and index the resulting BAM files. BAM files were converted into bedGraphs using bedtools (Quinlan and

Hall 2010) for Dogcatcher preprocessing in addition to the above steps. Then, DoGFinder was performed as detailed above with the addition of generating expression data for each set of DoGs discovered for each experiment as well as all DoGs (as discovered by Union_DoGs_annotation) using the Get_DoGs_rpkm function. ARTDeco was run as described above in both full mode (i.e., using all functions including read-in gene inference and differential expression) and in DoG discovery mode. Dogcatcher was run with and without differential expression (i.e., including or excluding the following scripts 3.0_Create_R_subread_DESeq2_script.py, 4.0_Dogcatcher_Rsubread_DESeq2.py, and 5.0_filter_sig_DESeq2.py.

2.11 Supplementary Figures

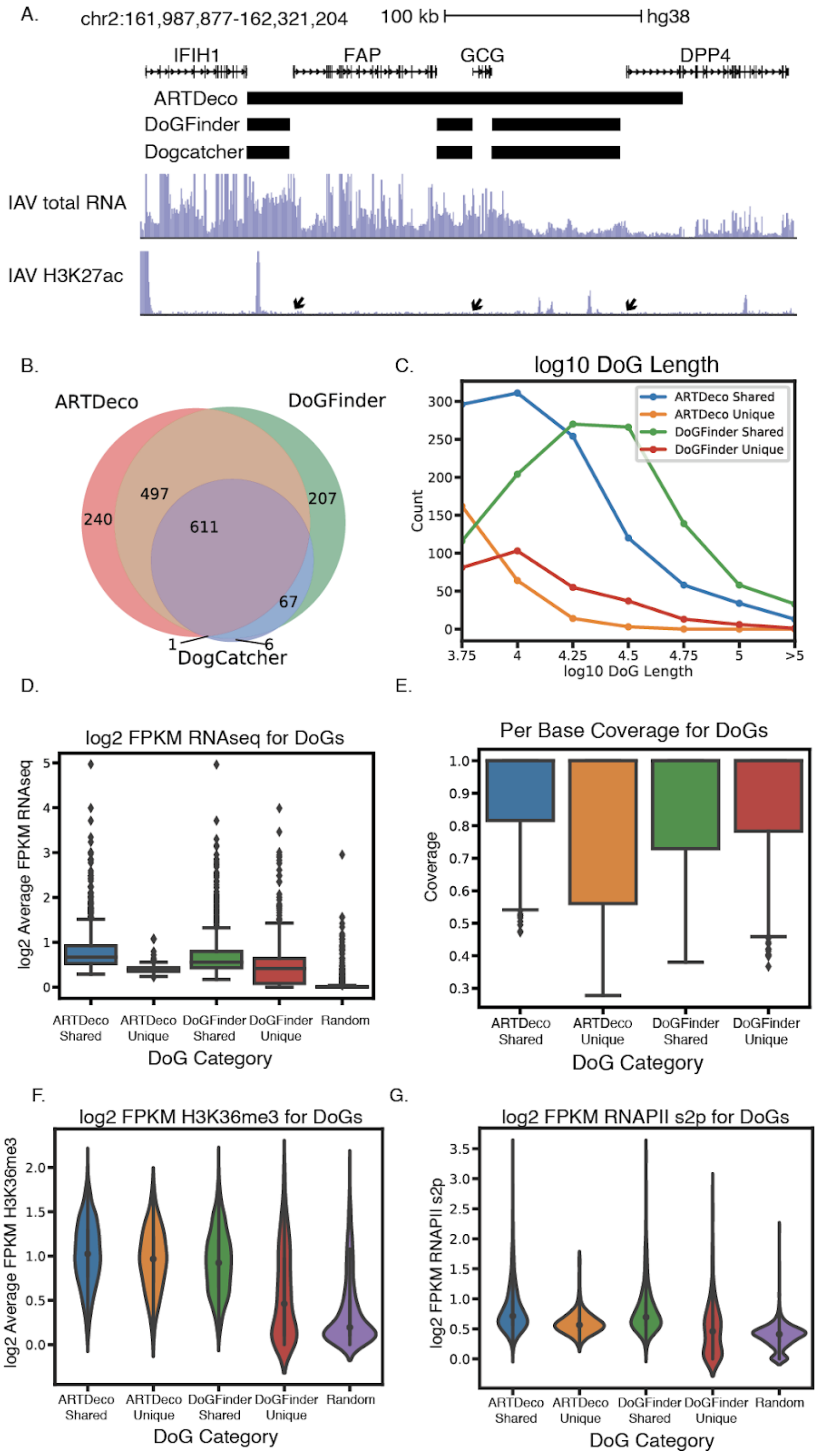
Supplementary Figure 2.1 Basic outline of ARTDeco data processing (a) Basic flowchart of ARTDeco functions. Program inputs are BAM files, a GTF file, and a chromosome sizes file as well as optional inputs for differential expression modes comprised of a meta file and a comparisons file. Data files are preprocessed into HOMER tag directories, a condensed gene annotation BED, and intergenic (read-in and downstream) BED files. From here, ARTDeco can compute read-in and readthrough statistics (left branch) or detect DoGs. Read-in levels for genes are used for DoG transcript discovery (details in Methods). (b) Schematic depicting the regions used to quantify read-in levels, readthrough levels, and DoG transcript discovery for each gene (maxlen is 15 kb by default). Examples of each region and total RNA-seq levels during IAV infection are depicted for the IFIH1 locus.


A.
Input
BAM files
GTF file
Chromosome sizes file
Meta file
Comparisons file
Preprocess
Tag Directories
Condensed Gene Annotation BED
Intergenic BED files
Gene Expression
Intergenic Expression
Screen Downstream Region of Genes
Read-In and Readthrough
Differential Expression
Gene vs. Intergenic
Read-In Genes
Discover DoGs
Quantify DoGs
DoGs
Differential Expression w/ Read-In
Differential Expression of DoGs
B.
Upstream Gene
Read-In Gene
Downstream gene(s)
Gene Annotation
Read-In Dist
Upstream Gene
Read-In Gene
min(maxLen, 1/3*UpstreamDist)
Readthrough Dist
Downstream Region
Readthrough Gene
Downstream Gene
min(maxLen, 1/3*DownstreamDist)
DoG Transcript
DoG Source Gene
Downstream gene(s)
minLen
log2 ratio
IFIH1
FAP
Read-in Region
100
log2 ratio
IFIH1
FAP
Downstream Region
100
IFIH1
FAP
GCG
DPP4
DoG Transcript
100

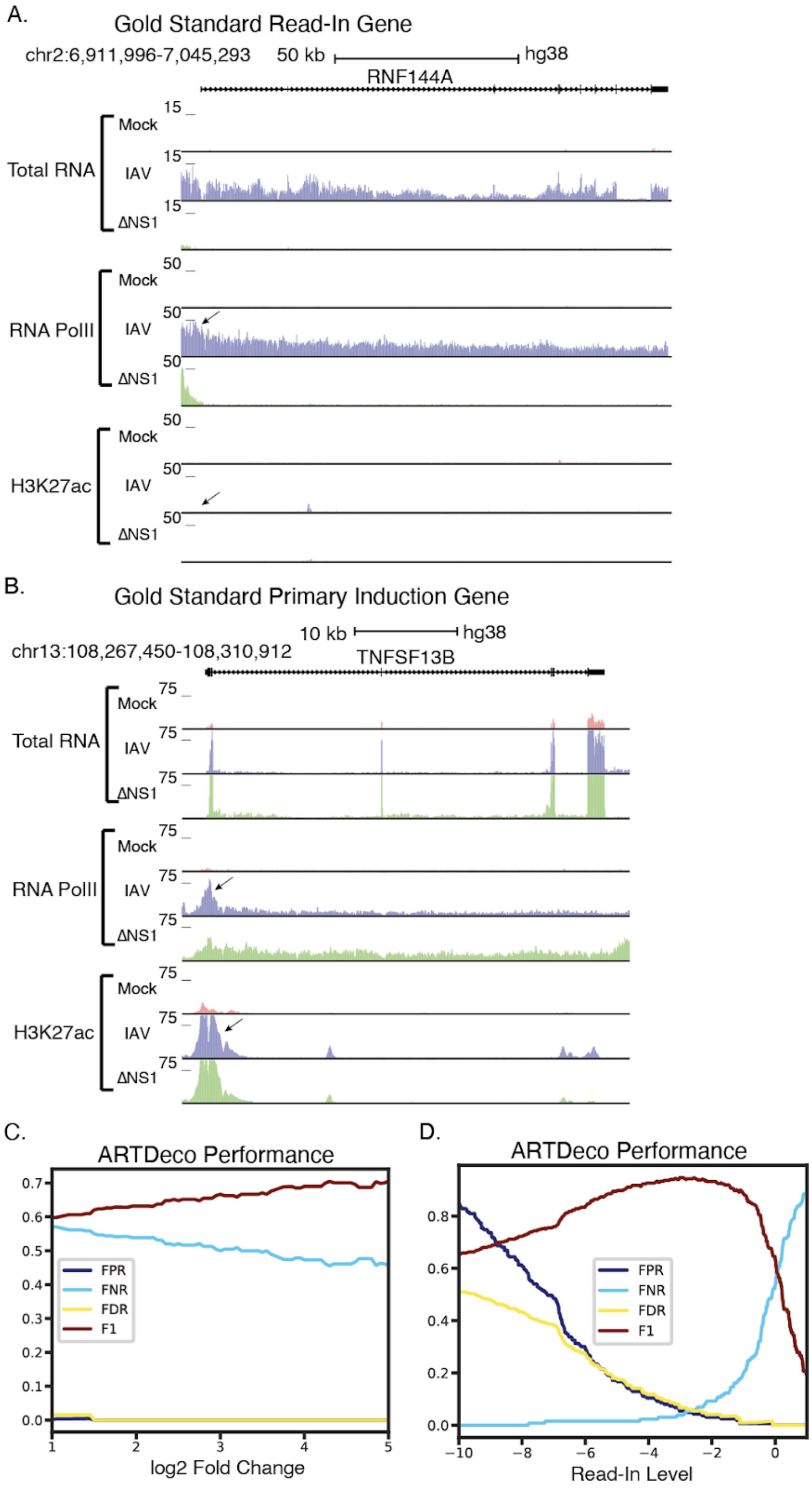


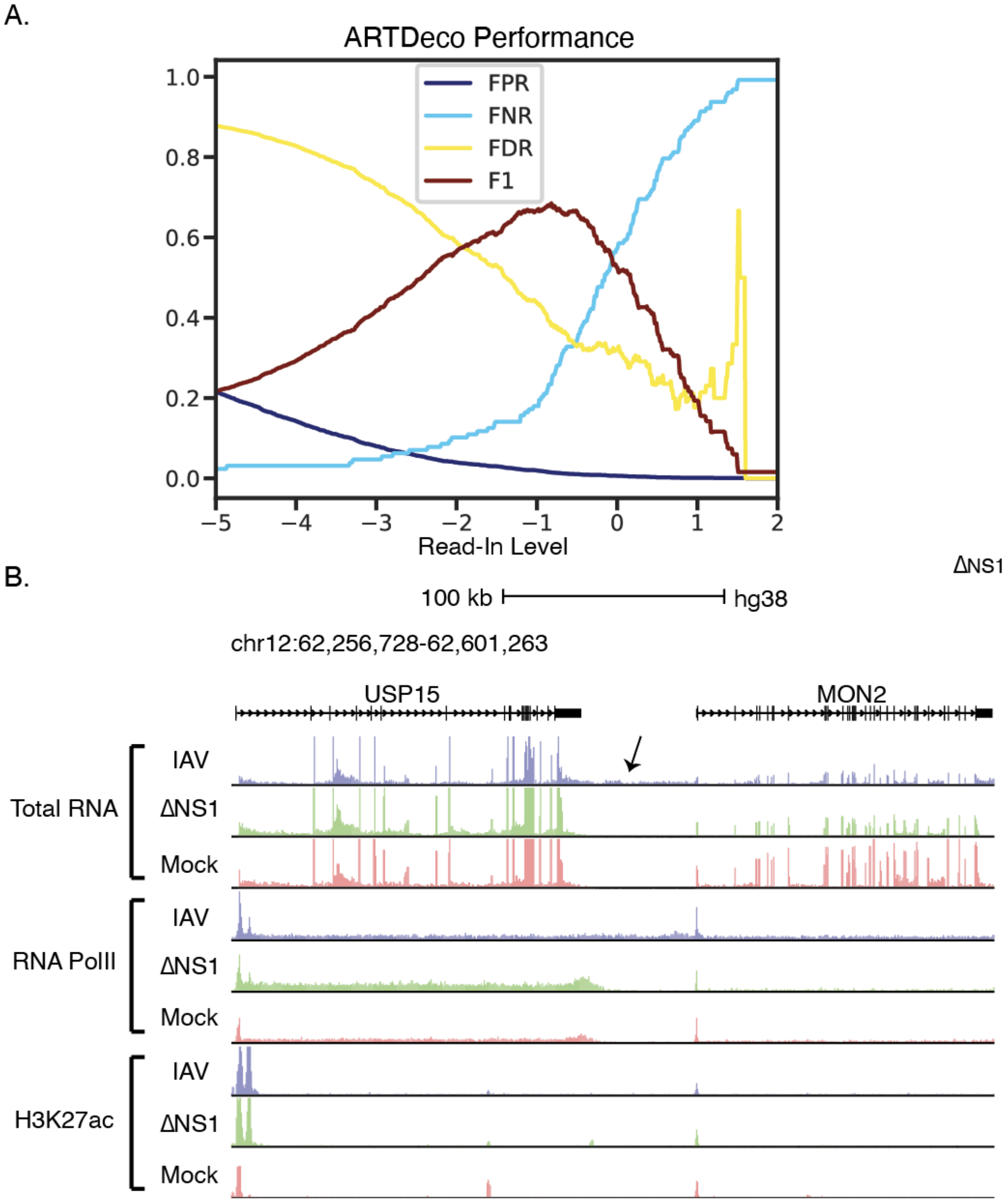
Supplementary Figure 2.2 Deconvolution of gene expression for upregulated genes in IAV relative to mock. (a) Uncorrected expression for IAV replicate 1 and ΔNS1 replicate 1. ($r = 0.72$; $p < 1e\text{-}77$) (b) Corrected expression for IAV replicate 1 and ΔNS1 replicate 1. ($r = 0.81$; $p < 1e\text{-}127$).

Supplementary Figure 2.3 Assessment of Downstream of Gene (DoG) transcripts. (a) Total RNAseq and H3K27ac ChIPseq at the IFIH1 locus and DoGs identified by ARTDeco and DoGFinder. (b) Venn diagram of all DoGs called by ARTDeco and DoGFinder using both IAV replicates using default coverage parameters and a sliding window of 500 bp. (c) Distribution of DoG lengths for DoGs called by ARTDeco and DoGFinder. (d) Distribution of RNA-seq FPKM values for DoGs identified by ARTDeco and DoGFinder. (e) Distribution of RNA-seq read coverage for DoGs identified by ARTDeco and DoGFinder. (f) Log2 FPKM H3K36me3 occupancy for DoGs assigned by ARTDeco and DoGFinder as well as random regions. (g) Log2 FPKM RNAPII s2p occupancy for DoGs assigned by ARTDeco and DoGFinder as well as random regions.

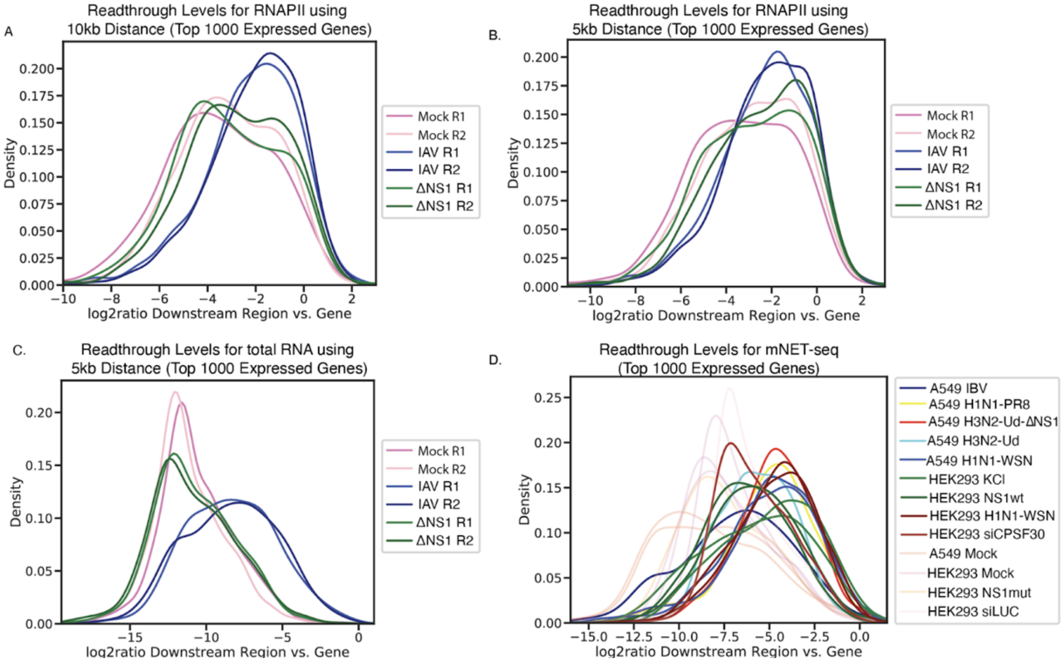

A.
chr2:161,987,877-162,321,204
100 kb
hg38
IFIH1
FAP
GCG
DPP4
ARTDeco
DoGFinder
Dogcatcher
IAV total RNA
IAV H3K27ac
B.
ARTDeco
DoGFinder
240
497
611
207
67
1
6
DogCatcher
C.
log10 DoG Length
ARTDeco Shared
ARTDeco Unique
DoGFinder Shared
DoGFinder Unique
Count
300
250
200
150
100
50
0
3.75
4
4.25
4.5
4.75
5
>5
log10 DoG Length
D.
log2 FPKM RNAseq for DoGs
log2 Average FPKM RNAseq
5
4
3
2
1
0
ARTDeco Shared
ARTDeco Unique
DoGFinder Shared
DoGFinder Unique
Random
DoG Category
E.
Per Base Coverage for DoGs
Coverage
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
ARTDeco Shared
ARTDeco Unique
DoGFinder Shared
DoGFinder Unique
DoG Category
F.
log2 FPKM H3K36me3 for DoGs
log2 Average FPKM H3K36me3
2.0
1.5
1.0
0.5
0.0
ARTDeco Shared
ARTDeco Unique
DoGFinder Shared
DoGFinder Unique
Random
DoG Category
G.
log2 FPKM RNAPII s2p for DoGs
log2 FPKM RNAPII s2p
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
ARTDeco Shared
ARTDeco Unique
DoGFinder Shared
DoGFinder Unique
Random
DoG Category

Supplementary Figure 2.4 Examples of primary induction and read-in genes from IAV-infected macrophages. (a) Example of a gold standard true positive (read-in) gene (RNF144A). Gene expression is upregulated in IAV relative to ΔNS1 and mock with low (> 0.5 FPKM) expression in ΔNS1. Additionally, there are no RNA PolII and H3K27ac ChIP-seq peaks (as called by HOMER) at the promoter regions. (b) Example of gold standard true negative (primary induction) gene (TNFSF13B). Gene expression is upregulated in IAV and ΔNS1 relative to mock. Additionally, there are both RNA PolII and H3K27ac peaks (as called by HOMER) at the promoter region indicating transcription initiation. (c) Benchmarking of ARTDeco performance for inference of read-in genes using false positive rate (FPR), false negative rate (FNR), false discovery rate (FDR), and F1 score while varying DESeq2 log2 fold change. Values for adjusted p-value, FPKM, and read-in level are 0.05, 0.25 and 0, respectively. (d) Benchmarking for ARTDeco performance for inference of read-in genes using FPR, FNR, FDR, and F1 score while varying read-in level. Values for log2 fold change, adjusted p-value, and FPKM are 2, 0.05, and 0.25, respectively.


A.
Gold Standard Read-In Gene
chr2:6,911,996-7,045,293
50 kb
hg38
RNF144A
Total RNA
Mock
IAV
ΔNS1
15
15
15
RNA PolII
Mock
IAV
ΔNS1
50
50
50
H3K27ac
Mock
IAV
ΔNS1
50
50
50
B.
Gold Standard Primary Induction Gene
chr13:108,267,450-108,310,912
10 kb
hg38
TNFSF13B
Total RNA
Mock
IAV
ΔNS1
75
75
75
RNA PolII
Mock
IAV
ΔNS1
75
75
75
H3K27ac
Mock
IAV
ΔNS1
75
75
75
C.
ARTDeco Performance
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
FPR
FNR
FDR
F1
1
2
3
4
5
log2 Fold Change
D.
ARTDeco Performance
0.8
0.6
0.4
0.2
0.0
FPR
FNR
FDR
F1
−10
−8
−6
−4
−2
0
Read-In Level



Supplementary Figure 2.5 Evaluation of read-in gene identification without using a control condition. (a) Benchmarking for ARTDeco performance for inference of read-in genes without differential expression while varying read-in level. Gene expression is > 0.25 FPKM. (b) Example of a gene (MON2) that was marked as a read-in gene despite being initiated. There is substantial readthrough originating from the upstream gene USP15.



Supplementary Figure 2.6 Analysis of RNAPII ChIP-seq and mNet-seq data using ARTDeco. (a) Distribution of readthrough levels for IAV, ΔNS1, and mock for top 1000 expressed genes based on ARTDeco's analysis of RNAPII ChIP-seq data (instead of RNA-seq data) using the default 10 kb downstream readthrough distance. (b) Distribution of readthrough levels for IAV, ΔNS1, and mock for top 1000 expressed genes based on ARTDeco's analysis of RNAPII ChIP-seq data using a 5 kb downstream readthrough distance. (c) Distribution of readthrough levels for IAV, ΔNS1, and mock for top 1000 expressed genes based on ARTDeco's analysis of total RNA-seq data using a 5 kb downstream readthrough distance. (d) Distribution of readthrough levels for mNET-seq data from Bauer et al. (2018) for top 1000 expressed genes. Cell types are denoted in legend as A549 and HEK293. Treatment conditions are as follows: IAV H1N1 WSN/33, IAV H1N1 Puerto Rico/8/34, IAV H3N2 Udorn/72, IAV H3N2 Udorn/72: NS1Δ99, Influenza B virus [IBV] Florida/04/2006, KCl, wildtype and mutant NS1 proteins, siLUC, and siCPSF. Conditions where readthrough was observed in the original analysis conducted by Bauer et al. (2018) have distribution curves with higher opacity.

CHAPTER 3: ANALYSIS OF READTHROUGH TRANSCRIPTION IN NGS DATASETS ENABLES IDENTIFICATION OF VIRALLY-INDUCED DEFECTS OF TRANSCRIPTION TERMINATION

3.1 Abstract

Host-pathogen interactions are an essential part of disease progression during viral infections. Two viruses (influenza A virus [IAV] and herpes simplex virus 1 [HSV-1]) are known to induce defects of transcription termination (DoTT). This causes the expression of novel transcripts, epigenomic remodeling, and changes in 3D chromatin structure. We hypothesized that this phenotype was not limited to these viruses and searched publicly available data for virally-infected systems with DoTT. We find evidence of significant DoTT for the first time in three additional viruses: Rift Valley Fever virus (RVFV), Zika virus (ZIKV), and Sindbis virus (SINV). We then investigated the cause of this phenotype in RVFV by expressing in vitro-transcribed (IVT) mRNA for the viral NSs protein in THP-1 monocytes. Expression of NSs revealed evidence of DoTT as well as a downregulation of host immune response genes. In all, we establish a pipeline for the discovery and validation of both the presence and mechanism DoTT in a novel system. This will enable future researchers to identify more viruses that induce DoTT and further characterize the causes and consequences of this phenotype.

3.2 Introduction

One key aspect of understanding disease progression and cytopathogenesis during viral infection is the characterization of host-pathogen interactions in the context of host gene expression. Host cells express genes encoding interferons (IFNs), pro-inflammatory cytokines, and chemokines in response to detection of acute viral infection (Rai et al. 2021; Chen et al. 2018; Iwasaki and Pillai 2014). Several viruses have evolved strategies to evade this host IFN response including shutdown of host transcription and/or translation (García-Sastre 2017; Lyles 2000; Rai et al. 2021). Understanding and characterizing these strategies can lead to the discovery of druggable viral targets.

One understudied phenotype has been virally-induced DoTT. In normal conditions (i.e., without DoTT), RNA polymerase II (RNAPII) elongates along the gene body and encounters a polyadenylation site (PAS), which causes a conformational change and recruitment of the cleavage and polyadenylation (CPA) machinery (Licatalosi et al. 2002). The CPA complex associates with the C-terminal domain (CTD) of RNAPII and the nascent pre-mRNA is released for further mRNA processing and maturation (H. Zhang, Rigo, and Martinson 2015). In systems with DoTT, RNAPII is not properly released from the DNA and transcription extends beyond the annotated transcription termination site (TTS) (called readthrough transcription).

DoTT has a variety of effects on the host cell. Among these are changes in 3D chromatin structure, opening of chromatin, and increase in transcription factor binding (Hennig et al. 2018; Heinz et al. 2018). Additionally, mRNAs that undergo readthrough transcription are not processed properly and are not exported out of the nucleus and translated (Rutkowski et al. 2015; Heinz et al. 2018). In traditional gene expression analyses using RNA-seq, these mRNAs represent analytical noise. The degree to which DoTT and readthrough transcription contribute to viral pathogenesis is poorly understood. In order to address these concerns and further characterize DoTT, our lab recently developed a software tool called ARTDeco that can identify and quantify readthrough transcription in systems with DoTT (Roth, Heinz, and Benner 2020).

Two viruses, IAV and HSV-1 are known to induce DoTT via similar mechanisms (N. Zhao et al. 2018; Rutkowski et al. 2015; Heinz et al. 2018; Bauer et al. 2018; Hennig et al. 2018; X. Wang et al. 2020). In IAV, the NS1 protein binds to the CPSF30 subunit of the CPSF and prevents recognition of the PAS (Nemeroff et al. 1998). In HSV-1, the ICP27 protein physically associates with many different subunits of the CPSF (CPSF160, CPSF100, CPSF73, CPSF30, Wdr33, and Fip1) in order to disrupt host transcription termination as well as aid in viral 3' mRNA processing (X. Wang et al. 2020). To date these are the only two viruses with validated mechanisms for inducing DoTT. Despite this shared phenotype, little else is similar between the two viruses. However, notably, both viruses engage in host transcription shutoff in order to evade host immune defenses (He et al. 2020; Khaperskyy and McCormick 2015; Bercovich-Kinori et al. 2016)

RVFV is a single-stranded, ambisense RNA virus known to infect animals (cattle, goats, and sheep), insects (mosquitoes), and humans (Davies and Martin 2006; Petrova et al. 2020). In livestock, the contraction of the virus can be fatal and is known to cause spontaneous abortion in pregnant animals (Glyn Davies, Martin, and Food and Agriculture Organization of the United Nations 2003). In humans, it can cause a range of symptoms ranging from mild flu-like symptoms to liver failure, retinal hemorrhage, and CNS dysfunctions typical of a hemorrhagic fever (Petrova et al. 2020; Glyn Davies, Martin, and Food and Agriculture Organization of the United Nations 2003). Similar to IAV and HSV-1, RVFV causes host transcription shutoff (Billecocq et al. 2004). Currently, only one genome-wide transcriptomic study of cells infected with RVFV exists and the effects of infection on host transcription have not been fully investigated (de la Fuente et al. 2018).

Here we use ARTDeco to screen public datasets in NCBI's Gene Expression Omnibus (GEO) in order to identify viruses that cause DoTT. We find several viruses that show evidence of DoTT. Of these, we proposed and validated a mechanism of DoTT for RVFV by expressing IVT mRNA of its NSs protein into THP-1 monocytes and then performing total RNA-seq. We characterize both patterns of readthrough transcription as well as host IFN shutoff and compare these phenotypes to those induced by expression of IAV NS1 protein. We find that global patterns of readthrough transcription are similar between both proteins indicating that they both inhibit transcription termination similarly. We also examined their effect on gene expression and found that they inhibit distinct sets of immune response and IFN-stimulated genes (ISGs).

3.3 Methods

3.3.1 Computational screening for DoTT, data processing, and bioinformatic analysis

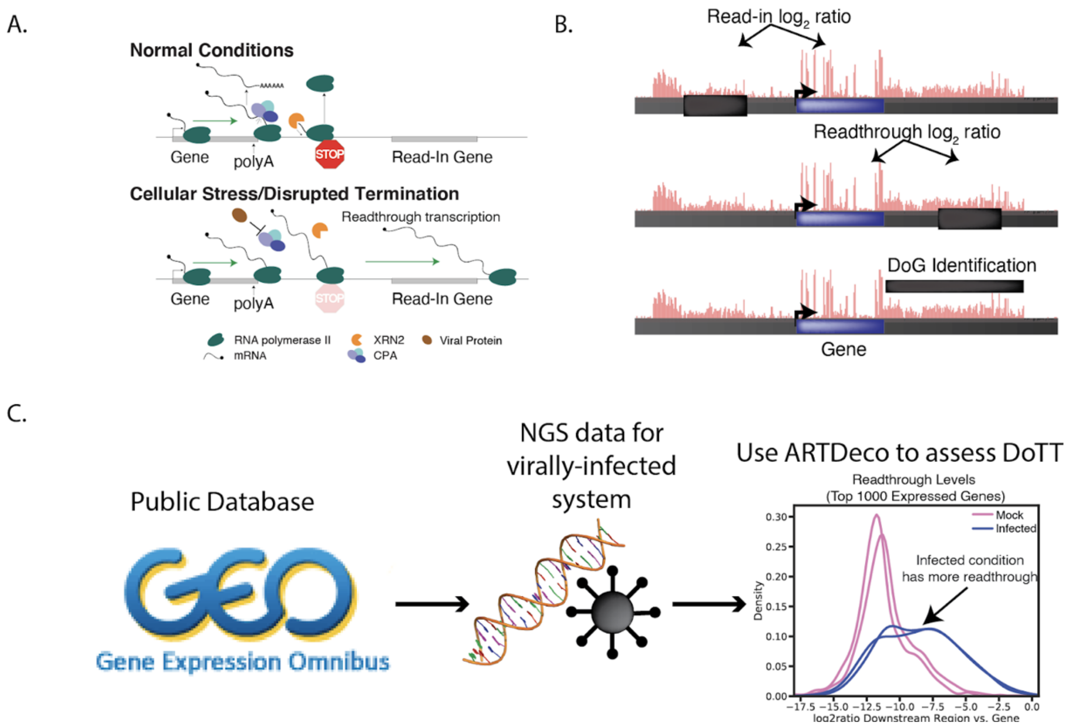


Figure 3.1. ARTDeco can measure and identify readthrough transcription. **A.** Diagram of normal transcription termination and how viral proteins induce DoTT by interfering with cleavage and polyadenylation (CPA) machinery. **B.** Schematic of the three measures of readthrough transcription as defined by ARTDeco. The read-in and readthrough levels are the log2 ratio of reads mapping to an upstream “read-in” or downstream “readthrough” regions (respectively) to reads mapping to the gene body. Downstream of gene (DoG) transcripts are regions of continuous transcript coverage downstream of the annotated TTS. **C.** Proposed pipeline for utilizing publicly available NGS data to discover systems with readthrough transcription in the context of viral infection.

Data for cells infected with Rift Valley Fever Virus, Zika virus, Sindbis virus, and Ebola virus were downloaded from NCBI’s GEO database (accessions in table). Raw fastq files were trimmed for adapters using Cutadapt v2.4. Then, it was aligned using STAR v. 2.7.0d to either the hg38 reference genome or a concatenation of the hg38 and the virus genome in samples with viral infection (references in table). Aligned files were then processed using ARTDeco with default parameters in readthrough mode using the GENCODE v31 gene annotation (Frankish et al. 2019). Only genes in the categories protein_coding, lincRNA, bidirectional_promoter_lncRNA, and processed_transcript as defined by GENCODE were considered for analysis of global patterns of readthrough (i.e., readthrough levels of top 1000 genes). Datasets were considered to have DoTT if infected samples showed higher readthrough levels than mock samples. This same processing pipeline was used for data generated in total RNA-seq experiments (outlined below).

Clustering of both readthrough levels and gene expression values was performed using the Python package Seaborn and its clustermap function (Waskom 2021). All NGS data were visualized on the UCSC genome browser using HOMER makeMultiWigHub.pl. GO analysis on gene sets was performed using Metascape. Motif analysis of gene promoters was performed using HOMER findMotifs.pl.

3.3.2 Generation of IVT mRNAs

We generated IVT mRNAs for the NS1 and NSs viral proteins from IAV and RVFV, respectively, as well as EGFP as a control. NS1 IVT mRNA was the same as that used in Heinz et al. (2018) and was generously provided by Sven Heinz. Similarly, EGFP sequence and primers were the same as used in Heinz et al. (2018). NSs sequence was codon optimized with a C-terminal HA tag using the IDT Codon Optimizer. NSs IVT mRNA was generated using a T7-promoter containing DNA template which was amplified using NS1_T7_f/NS1_T7_r primers. PCR products were phenol:chloroform extracted and resulting template was transcribed using mMESSAGE mMACHINE™ T7 Transcription Kit (Invitrogen). IVT mRNA was extracted using MEGAClear Transcriptional Clean Up Kit (Ambion) and treated with Calf Intestinal Alkaline Phosphatase (NEB). Resulting mRNA was then purified using phenol:chloroform:isoamyl alcohol (Invitrogen).

3.3.3 Electroporation

We introduced viral mRNAs (and EGFP) into cells via electroporation. We used 3M THP-1 cells per electroporation. Cells were pelleted by centrifugation (8 minutes at 1200 rpm at room temperature) and washed with 10 mL OPTIMEM twice. Cells were electroporated in 200 μL OPTIMEM in a 4 mm cuvette using a Gene Pulser Xcell (Bio-Rad) with a rectangular pulse of 400 V and 5 ms duration to deliver 3 μg of NSs and EGFP IVT mRNA and 9 μg of NS1 IVT mRNA. Cells were transferred to a pre-warmed 6 well plate with 3 mL of media (RPMI-1640 containing 10% low-endotoxin FBS (Peak Serum Inc.), 1x (100 μM) non-essential amino acids (Life Technologies), 1x MEM Sodium Pyruvate Solution (1 mM) (Life Technologies), 1x GlutaMax I (Life Technologies), 50 μM β-mercaptoethanol (Life Technologies) and allowed to incubate for 6 hours at 37℃, 5% CO2, 100% humidity. For NS1, NSs, and one EGFP sample, media was treated with 500 units of IFN-β per mL of media with another EGFP sample left untreated as a control. After incubation, 0.5M cells were reserved for Western blot verification of protein expression using anti-NS1 and anti-HA antibodies for NS1 and NSs proteins, respectively.

3.3.4 RNA-sequencing and library preparation

After electroporation, cells were lysed using Trizol LS (Thermo). RNA was isolated using Direct-zol RNA Miniprep kit (Zymo Research) according to manufacturer's instructions. Library preparation and sequencing were performed by UCSD's IGM Genomics center.

Table 3.1 Reagents

Reagent	**Source**	**Identifier**
Human IFN-β	R&D Systems	Cat# 11415-1
Calf Intestinal Alkaline Phosphatase	NEB	M0290S
UltraPure Phenol:Chloroform:Isoamyl Alcohol	Invitrogen	Cat# 15593031
Direct-zol RNA MiniPrep	Zymo Research	R2050
MEGAClear Transcription Clean-up Kit	Ambion	AM1908
Human: Cell line THP-1	ATCC	TIB-202
mMESSAGE mMACHINE™ T7 Transcription Kit	Invitrogen	AM1344
Q5® Hot Start High-Fidelity 2X Master Mix	NEBNext	M0494S

Table 3.2 GEO Accessions

Data set	**Study**	**GEO Accession**
Rift Valley Fever Virus	de la Fuente et al. (2018)	GSE102481
Zika Virus	Carlin et al. (2018)	GSE118305
Sindbis Virus	Garcia-Moreno et al. (2019)	GSE125182
Ebola Virus	Smith et al. (2017)	GSE100839

Table 3.3 Viral Genome Accession

Virus	GenBank Accession(s)
Rift Valley Fever Virus MP-12 Strain	DQ380154.1, DQ375404.1, DQ380208.1
Rift Valley Fever Virus ZH-548 Strain	DQ380206.1, DQ375407.1, DQ380151.1
Zika Virus	KU955593.1
Sindbis Virus	NC_001547.1
Ebola Virus	AF086833.2

3.4 Results

3.4.1 Screening of publicly available data reveals evidence for more virally-induced DoTT

Our lab previously developed a software tool for identifying and characterizing systems with DoTT called ARTDeco (Roth, Heinz, and Benner 2020). In brief, ARTDeco measures the amount of readthrough transcription in a given system using three different metrics (read-in level, readthrough level, and downstream of gene [DoG] transcripts). Our previous results indicated that the distribution of readthrough levels among the top 1000 expressed genes is a good measure of DoTT in a system as well as being robust to data type (Roth, Heinz, and Benner 2020). Given these observations, we deployed ARTDeco as a discovery tool for identifying systems with readthrough transcription (schematically outlined in Figure 3.1).

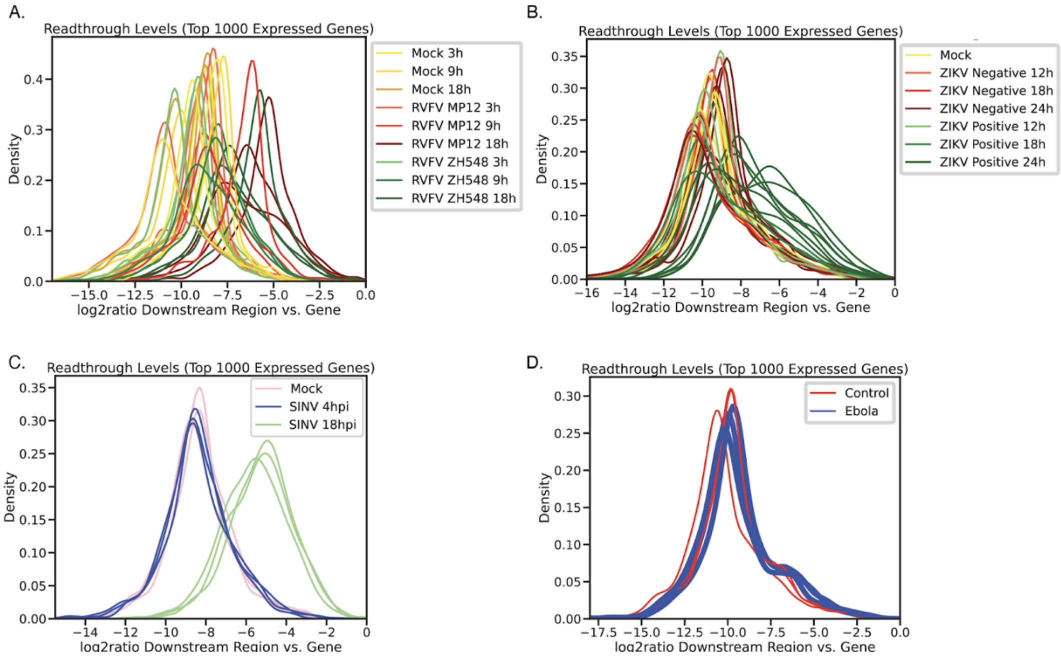


Figure 3.2. Distribution of readthrough levels in top 1000 expressed genes in publicly available datasets. Distribution of the readthrough levels of the top 1000 expressed genes in RVFV (**A)**, ZIKV (**B**), and SINV (**C**) indicate presence of readthrough transcription while it does not for Ebola virus (**D)**.

Two highly dissimilar viruses (IAV and HSV-1) are known to cause DoTT (N. Zhao et al. 2018; Rutkowski et al. 2015; Heinz et al. 2018; X. Wang et al. 2020; Bauer et al. 2018; Hennig et al. 2018). Thus, we hypothesized that other viruses also cause this phenotype. We focused our search on viruses that induce host transcription shutoff as this is one of the only commonalities between IAV and HSV-1 infections but also screened other viruses for completeness (Khaperskyy and McCormick 2015; He et al. 2020). With this in mind, we utilized ARTDeco to screen virally-infected systems with data in the NCBI GEO database to see if host DoTT was induced (Barrett et al. 2013) (Figure 3.1C).

We found evidence of DoTT in total RNA-seq experiments for three separate viruses: RVFV, Zika virus (ZIKV), and Sindbis virus (SINV) (de la Fuente et al. 2018; Carlin et al. 2018; Garcia-Moreno et al. 2019). These are genetically diverse viruses with RVFV being a phlebovirus while ZIKV and SINV belong to the flavivirus and alphavirus genuses, respectively (Ikegami 2012; Musso and Gubler 2016; Adouchief et al. 2016). Importantly, all three of these viruses induce host transcription shutoff in infected cells (Billecocq et al. 2004; Akhrymuk, Kulemzin, and Frolova 2012; Fros and Pijlman 2016; Carlin et al. 2018). All viruses showed more readthrough at later timepoints in infection conditions relative to the mock condition (i.e., no infection) (Figures 3.2A-C,3.3). ZIKV samples were sorted by infection status (i.e., infected cells vs. bystander cells) and only infected cells at 24 hours post virus treatment showed signs of DoTT (Carlin et al. 2018) (Figure 3.2B,3.3B). This is consistent with readthrough transcription resulting from viral infection rather than a technical artifact. In addition to viruses that show evidence of readthrough transcription, we found that many viruses (e.g., ARPE-19 cells infected with Ebola virus from Smith et al. (2017) in Figure 3.2D) showed no evidence of this phenotype. This indicates that while readthrough transcription is more common in viruses than previously appreciated, it is not a universal phenotype. In all, we show that ARTDeco can be deployed as a tool for discovery of DoTT and identify three viruses with the readthrough transcription phenotype.

3.4.2 Expression of RVFV NSs protein results in readthrough transcription

After identifying DoTT in these viruses, we then sought to discover the mechanism for DoTT. We chose to focus on RVFV as we were able to identify its NSs protein as a strong candidate for the cause of this phenotype. We chose the NSs protein because it has been implicated in host transcription and translation shutoff processes during RVFV infection and is localized to the nucleus (Billecocq et al. 2004; Ly and Ikegami 2016). Additionally, it is known to interfere with mRNA export (similar to the NS1 protein in IAV) (Copeland, Van Deusen, and Schmaljohn 2015). Because of these phenotypes, we hypothesized that NSs was the cause of readthrough transcription in RVFV. In order to test our hypothesis, we expressed NSs in THP-1 monocytes via electroporation of IVT mRNAs (Figure 3.4A). This approach is similar to both previous work in our lab where we expressed NS1 to examine the effects of readthrough transcription on the transcriptome and epigenome of THP-1 cells and the Pfizer and Moderna mRNA vaccines which lead to expression of the SARS-CoV-2 spike protein (Heinz et al. 2018; Sahin et al. 2020; Jackson et al. 2020). Briefly, we generated NS IVT mRNA, electroporated them into cells treated with IFN-β, and subsequently performed total RNA-seq. In addition to treating with NSs, we also electroporated in IVT mRNAs for NS1 and EGFP to compare NSs to a known cause of readthrough (NS1) as well as a negative control (EGFP) (Figure 3.4A). Additionally, we electroporated EGFP into THP-1 cells that were untreated in order to control for effects of IFN-β treatment.

We observed readthrough transcription in both NS1 and NSs samples though no readthrough in EGFP treated samples. NS1-treated cells showed more readthrough than NSs-treated cells indicating that it is likely a more potent inhibitor of transcription termination, though this may be due to differences in protein

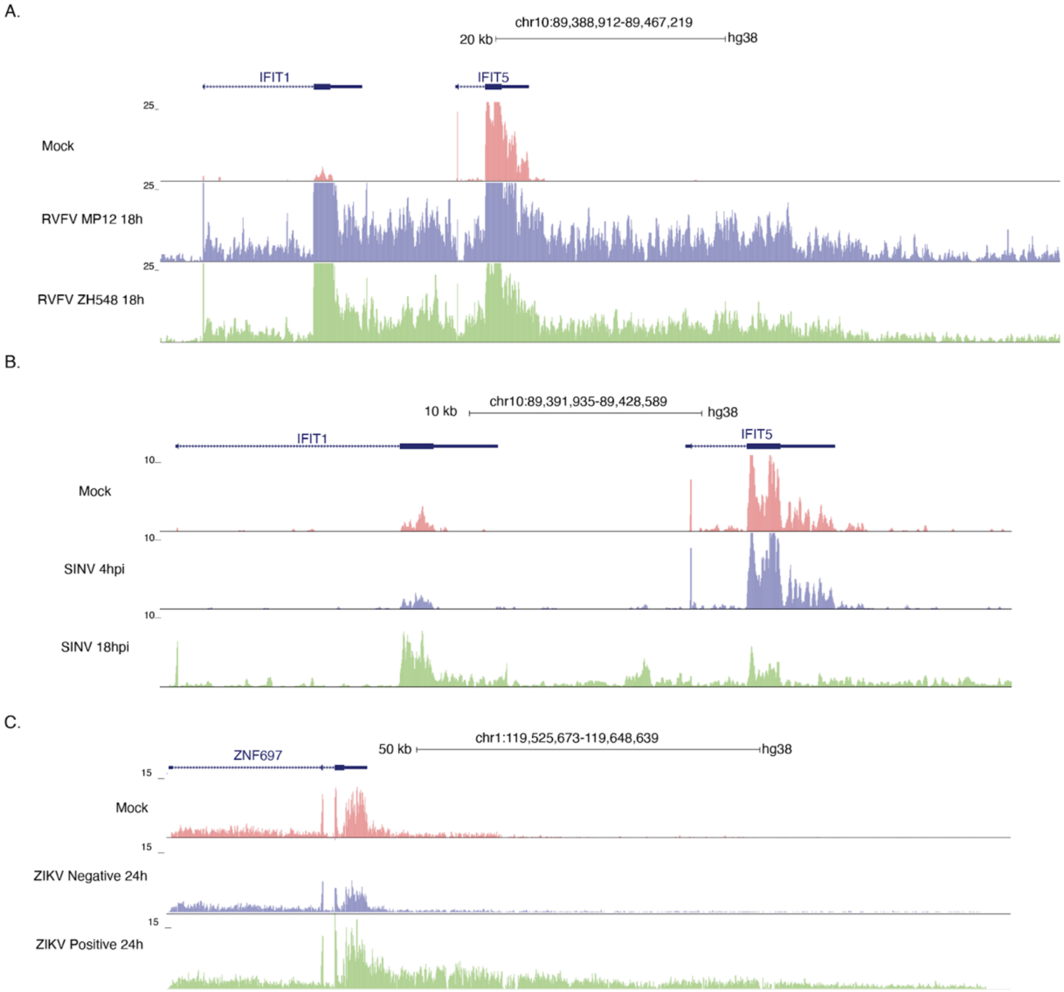


Figure 3.3**.** UCSC Genome Browser shots for viruses with DoTT. UCSC Genome Browser shots of loci exhibiting readthrough transcription in RVFV (**A)**, ZIKV (**B**), and SINV (**C**). ZIKV cells were sorted by infection status prior to sequencing so ZIKV positive and ZIKV negative correspond to infected and bystander cells, respectively.

Expression (Figure 3.4B,C).

Given that we had confirmed our hypothesis that NSs is a cause of DoTT in RVFV infection, we sought to examine patterns of readthrough. While NS1 inhibits PAS recognition and causes global DoTT for polyadenylated genes, it is unknown whether RVFV infection induces DoTT on a genome-wide scale or targets specific loci (N. Zhao et al. 2018). We clustered the readthrough levels of the top 1000 expressed genes from all samples (Figure 3.4C). Much like NS1, NSs appears to induce readthrough transcription on a genome-wide level but at a lower magnitude than NS1 (Figure 3.4D).

Next, we wanted to assess which genes experienced substantial readthrough transcription. It is currently unknown whether viruses target particular groups of genes for disabling termination or if certain genes are more susceptible to termination defects. We extracted two major clusters that showed distinct patterns of readthrough (Figure 3.4D). The first cluster showed significant readthrough for NS1 and the second cluster showed readthrough for both NS1 and NSs. We used Metascape to examine functional enrichment among genes in these clusters (Zhou et al. 2019). Interestingly, we found that both clusters had highly significant enrichment for immune defense genes (Figure 3.4E). The reason for this is unknown. It is possible that readthrough transcription is targeting immune response programs with a yet undiscovered mechanism.

3.4.3 Expression of NSs and NS1 leads to distinct patterns of downregulation of immune response and ISGs

In addition to examining DoTT caused by NS1 and NSs, we wanted to assess the effects of these proteins on host gene expression. Both of these proteins have been reported to interact with transcriptional machinery that affects transcription initiation in addition to their role in disrupting transcription termination and downstream mRNA processing (Nogales et al. 2018; Le May et al. 2004, [a] 2008). With this in mind, we sought to examine their effects on transcription apart from inducing DoTT. We first examined genes that were upregulated in IFN-β-treated and untreated cells that were electroporated with EGFP mRNA. We saw relatively few upregulated genes (considered upregulated with log2 fold change > 1 and unadjusted p-value < 0.05) (Figure 3.5A). These few genes were primarily known ISGs and immune response loci, indicating some activation of IFN pathways. Overall, IFN activation appeared modest.

We then looked at genes that were downregulated in NS1- and NSs-treated cells relative to EGFP-treated cells (with IFN-β). Interestingly, we found more downregulated genes in these samples than genes upregulated as a result of IFNβ-stimulation (Figure 3.5A). Using Metascape, we examined functional enrichments among these genes (Zhou et al. 2019) (Figure 3.5C). Despite largely distinct gene sets, both NS1- and NSs-downregulated genes showed significant enrichment for immune response pathways (Figure 3.5C). We clustered gene expression profiles to verify that these differences were due to distinct patterns of host immune gene suppression rather than thresholding effects. Indeed, we confirm that these genes have distinct patterns of expression. This suggests that NS1 and NSs shutdown host ISG expression in unique ways (Figure 3.5B).

NSs shutdown of host gene expression is hypothesized to be due to inhibition of TFIIH and, expression of IFN-β-specifically, recruitment of the SAP30 protein (Le May et al. 2004; Kainulainen et al. 2014; Kalveram, Lihoradova, and

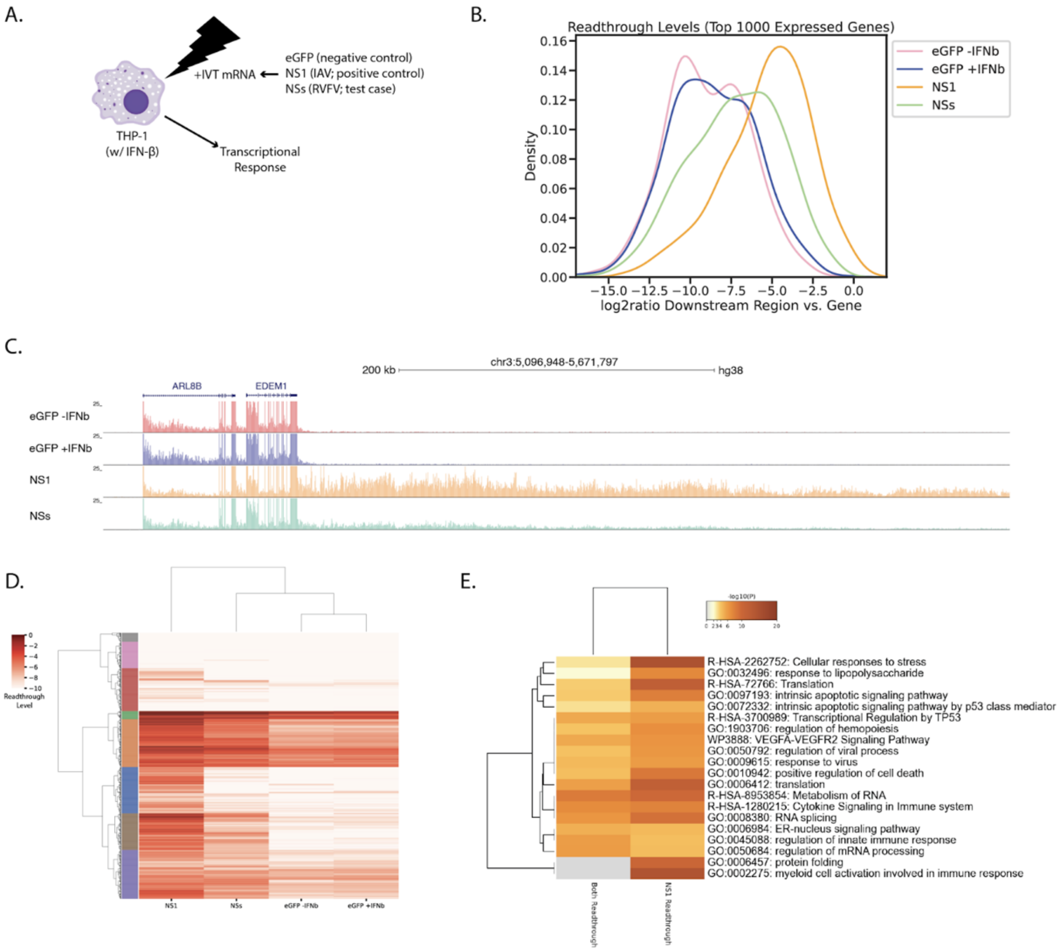


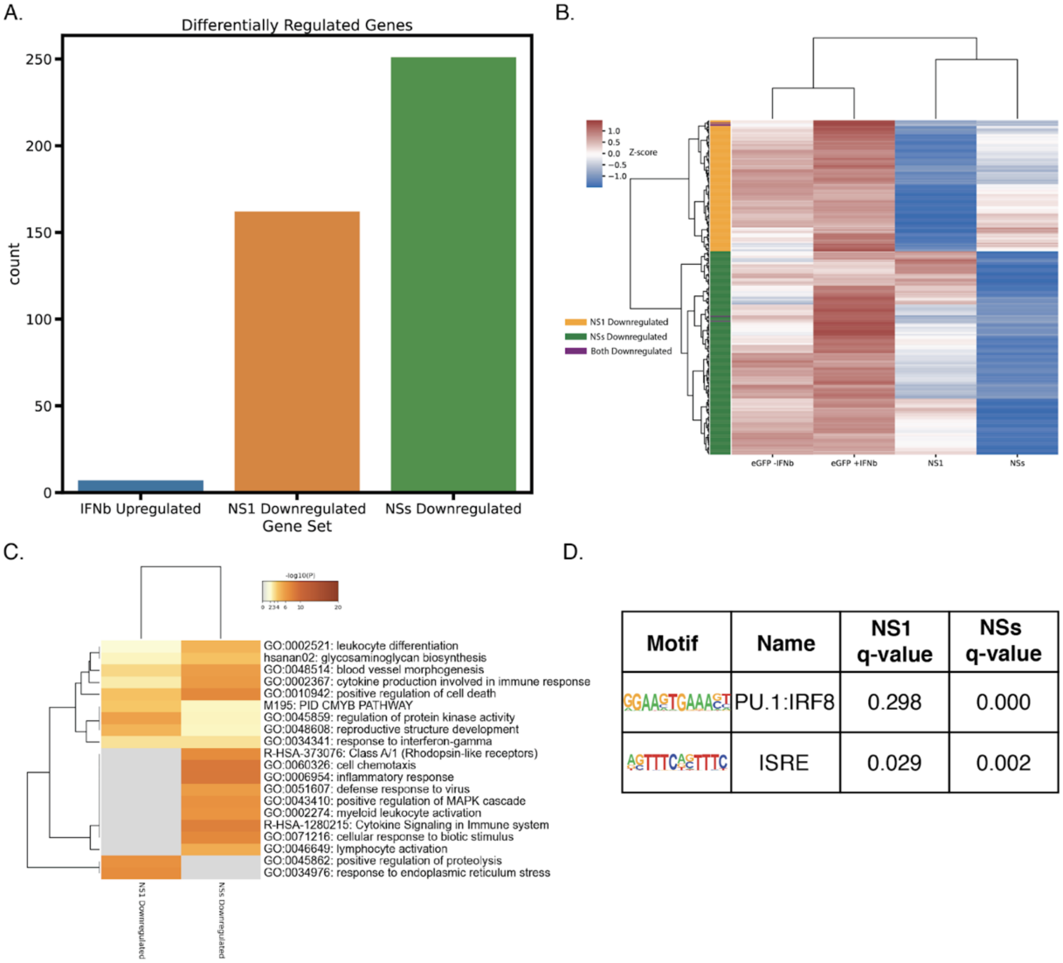
Figure 3.4**.** Cells expressing NSs show evidence of readthrough transcription. **A.** Schematic of experimental system for testing if viral proteins induce DoTT. Viral (or EGFP) IVT mRNAs are electroporated into THP-1 cells treated with IFN-β which are then subjected to total RNA-seq. **B.** Distribution of readthrough levels for the top 1000 expressed genes for cells treated with NS1, NSs, and EGFP IVT mRNAs. **C.** Example of locus showing readthrough transcription for cells treated with NS1 and NSs. **D.** Clustering of readthrough levels for the top 1000 expressed genes in each experiment. Clusters are in the leftmost column. **E.** Metascape (Zhou et al. 2019) clustering of functional enrichment for genes with evidence of readthrough in NS1-treated cells (Blue cluster in D) and readthrough in both NS1- and NSs-treated cells (bottom two clusters).

Ikegami 2011; Terasaki, Ramirez, and Makino 2016; Le May et al. 2008). We wanted to examine whether NS1 and NSs promoter elements have sequence motifs that are reflective of shutdown of ISGs. As expected, we found that promoters of genes that are downregulated by both proteins contain interferon-stimulated response elements (ISRE), consistent with the hypothesis that these proteins shut down host ISGs (Figure 3.5D). Surprisingly, we found no enrichment for YY1 motifs in promoters of genes downregulated by NSs (Figure 3.5D). NSs's interaction with SAP30 is thought to suppress YY1-mediated transcription initiation at the promoter region of IFN-β (Le May et al. 2008). It is possible that this is not a global mechanism of transcription inhibition, but analyses of transcription initiation patterns are necessary to investigate this relationship.

3.5 Discussion

Previous analyses of virally-induced DoTT and readthrough transcription have focused on one virus or one dataset at a time. Here we outline a pipeline for discovery of systems with this phenotype utilizing ARTDeco. Briefly, researchers can access public databases such as GEO, dbGaP, TCGA, ENCODE, etc., download transcriptomic data, and process it using ARTDeco to assess whether DoTT are present. Given the ever increasing amount of data in these databases, this represents a high throughput way to search for and characterize systems with readthrough transcription.

We were able to discover DoTT in three viruses (RVFV, ZIKV, and SINV) not previously known to display this phenotype (Figure 3.2A-C). While these were all total RNA-seq datasets, there was remarkable heterogeneity among them. While both RVFV and SINV datasets used whole cell populations, the ZIKV dataset utilized cell sorting to isolate infected and bystander cells (Carlin et al. 2018).



Motif	Name	NS1 q-value	NSs q-value
GGAAGTGAAAGT	PU.1:IRF8	0.298	0.000
AGTTTCAGTTTC	ISRE	0.029	0.002

Figure 3.5**.** Assessment of differentially regulated genes in response to treatment with IVT mRNAs. **A.** Counts of genes upregulated in EGFP-treated cells with IFN-β treatment (relative to without IFN-β treatment), genes downregulated in NS1-treated cells relative to EGFP-treated cells, and genes downregulated in NSs-treated cells relative to EGFP-treated cells. **B.** Clustering of gene expression of genes downregulated in both NS1- and NSs-treated cells relative to EGFP-treated cells. **C.** Metascape (Zhou et al., 2019) clustering of functional enrichments for genes downregulated in NS1- and NSs-treated cells relative to EGFP-treated cells. **D.** Motif enrichments at promoters for genes downregulated in NS1- and NSs-treated cells relative to EGFP-treated cells.

Additionally, RVFV and SINV datasets were infected at different multiplicities of infection (MOI), 5 and 10, respectively (de la Fuente et al. 2018; Garcia-Moreno et al. 2019). Finally, all three datasets came from different cell types: human small airway epithelial cells (HSAECs) for RVFV, macrophages for ZIKV, and HEK293 cells for SINV (de la Fuente et al. 2018; Carlin et al. 2018; Garcia-Moreno et al. 2019).

In addition to systems where we were able to discover DoTT, we also showed an example of a virally-infected system that did not show evidence of this phenotype (Ebola virus) (Figure 3.2D). These cells were infected at an MOI of 5 so it is unlikely that this observation is due to an insufficient infectivity (Smith et al. 2017). The effects of the Ebola virus on host transcription are largely unknown and poorly characterized, however, it is unlikely the induction of readthrough transcription is a consequence of infection (Speranza and Connor 2017). This demonstrates ARTDeco's flexibility and ability to identify readthrough transcription in a multitude of systems as well as ability to discern systems with DoTT from systems without DoTT. Given the remarkable diversity of viruses and viral datasets, this is critically important when investigating this phenotype.

With the exception of RVFV (discussed below), it is unknown how these viruses cause DoTT. Based upon literature review, we hypothesize the mechanisms of DoTT in ZIKV and SINV. In ZIKV, the NS5 protein seems like a likely candidate for causing readthrough transcription. It localizes to the nucleus and is known to interact with gene regulation and RNA processing machinery

(Davidson 2009; Shah et al. 2018). Further, it is directly implicated in host IFN shutdown (Davidson 2009; Shah et al. 2018; Z. Zhao et al. 2021). Similarly, in SINV, the nsP2 protein is known to localize to the nucleus and inhibit host immune response (Fros and Pijlman 2016; Akhrymuk, Kulemzin, and Frolova 2012). Interestingly, nsP2 is thought to inhibit host immune responses by disabling host transcription entirely by degrading the RBP1 subunit of RNAPII (Akhrymuk, Kulemzin, and Frolova 2012). Similarly, Carlin et al. (2018) noted that RNAPII levels were lower in ZIKV-infected cells compared to bystander cells. How can RNAPII be degraded and simultaneously show signs of DoTT? As observed in Carlin et al. (2018), while there may be a global decrease in RNAPII occupancy on a genome-wide level, there may still be engaged RNAPII on the genome producing transcripts. In fact, this is the case for HSV-1 where ICP27 is known to both induce DoTT and degradation of RNAPII (X. Wang et al. 2020; Dai-Ju et al. 2006). Further investigation is required to confirm these hypotheses.

Of the viruses that showed evidence of DoTT in public data, we were able to confirm the mechanism in RVFV. By utilizing insights from the available literature on the virus, we hypothesized that the NSs protein caused the phenotype. We confirmed this hypothesis by expressing NSs in THP-1 cells. To date, this is the third viral protein that has been identified to cause DoTT with the other two being NS1 in IAV and ICP27 in HSV-1 (Heinz et al. 2018; X. Wang et al. 2020). It is notable that we were able to generate and test our hypothesis that NSs causes readthrough transcription without culturing the virus in our facilities. Instead, we used ARTDeco to screen publicly available data, examined literature surrounding the biology of the virus, and electroporation of viral mRNAs into cells in order to validate the cause of DoTT. This represents a novel discovery approach that is similar to how SARS-CoV-2 mRNA vaccines were generated (Sahin et al. 2020; Jackson et al. 2020).

Our screening platform allowed us to examine the effects of expressing viral proteins on both transcription termination as well as general effects on the transcriptome. We found that NS1 induced more readthrough transcription than NSs did (Figure 3.3B-D). Both proteins caused DoTT on a genome-wide level suggesting that NSs disrupts transcription termination through a general mechanism rather than a locus-specific one (Figure 3.3B-D).

We also analyzed the effects of these proteins on host gene expression. We found that, consistent with prior observations, both NS1 and NSs inhibit ISGs and innate immune response genes (Le May et al. 2004; Kainulainen et al. 2014; Terasaki, Ramirez, and Makino 2016; Le May et al. 2008). This is an interesting observation as there was an enrichment for immune response in genes that showed significant readthrough (Figure 3.4E). Given prior observations indicate that transcripts with considerable readthrough are not exported and translated like normally processed mRNAs, this suggests that these viruses mount a dual-pronged attack against host immune response wherein they disable both transcriptional activation and transcription termination of host ISGs (Heinz et al. 2018; Rutkowski et al. 2015).

Despite the fact that both viruses disrupted transcriptional activation of ISGs and innate immune response genes, genes downregulated by either NS1 or NSs were largely distinct (Figure 3.5B). This suggests that NS1 and NSs inhibit ISG activation in distinct ways. More experiments are needed to flesh out the mechanisms responsible for this difference. NSs has been hypothesized to suppress promoter activation of ISGs via binding to SAP30, but the extent of this mechanism is unknown (Le May et al. 2008). To date, this aspect of NSs has not been interrogated on a genome-wide level and we have the ability to extend our experimental system to incorporate other assays that can address these questions.

DoTT may be an understudied hallmark of certain viruses. While the characterization of this phenotype adds to basic knowledge of viral mechanisms, it also represents a potential target for therapeutic interventions. It is unknown how cytopathic DoTT is to host cells. If DoTT is part of the host transcription shutoff as we hypothesize, perhaps targeting the protein responsible would represent an effective treatment. Additionally, if viruses can be engineered to lack DoTT as a phenotype, they may make good vaccines. This has already been accomplished with IAV so it may be possible to extend it to other viruses (Steel et al. 2009). In all, studying DoTT and readthrough transcription provides both mechanistic insights into virology as well as a potentially druggable phenotype.

3.6 Acknowledgements

I would like to thank all of the wet lab personnel in the Svenner lab for helpful tips in generating mRNAs and collecting samples, especially Sascha Duttke and Carlos Guzman. Extreme thanks are due to Camila De Arruda Saldanha for mentoring me through the entire process. This would not have been possible without your guidance.

Chapter 3, in part, is material being prepared for submission contributed by Samuel J. Roth, Camila De Arruda Saldanha, Sven Heinz, Christopher Benner. The dissertation author was the primary investigator and author of this paper.

CHAPTER 4: PREDICTING FUNCTIONAL INTERACTIONS BETWEEN TRANSCRIPTION FACTORS USING CONSECUTIVE ROUNDS OF MOTIF ENRICHMENT

4.0 Disclosure

This chapter diverges from the previous theme of systems with defects of transcription termination (DoTT) and readthrough transcription and instead focuses on developing a method for analyzing transcription start site regions (TSRs). My work in the Benner Lab has focused on many aspects of transcriptional regulation including transcription initiation. While a major portion of the work has focused on DoTT, I have also taken an interest in the grammar of transcription initiation and worked on various projects that address transcriptional regulatory grammar writ large. This chapter is reflective of this aspect of my work. In it, I develop a novel method for computing motif co-occurrences that, in addition to recapitulating known transcription factor (TF) interactions, offers unique interpretational insights into these TF interactions. Thus, this chapter is representative of a significant portion of my work.

4.1 Abstract

Gene expression is regulated by the combinatorial action of transcription factors (TFs). This combinatorial action can be inferred by examining the co-occurrence of TF binding motifs at regulatory regions. Here we develop a novel method of identifying motif co-occurrences that employs consecutive rounds of motif enrichment for target sequences against background genomic sequences.

We deploy this method on transcription start site data and find that our method, called the Dual HOMER method, is able to recapitulate known interactions between TFs better than an approach that does not incorporate motif enrichment. Finally, we show that the transcriptional network generated by the Dual HOMER network is directional and lends interpretable insight into the grammar of TF cooperativity.

4.2 Introduction

In the human body, all cells have approximately the same genome, however, there are many different cell types with distinct morphology and function. These differences are determined by differences in gene expression as influenced by cis-regulatory architecture (i.e., promoters, enhancers, and other regulatory sequences) (Maston, Evans, and Green 2006). The cis-regulatory architecture is determined by the binding of transcription factors (TFs) to regulatory sequences. TFs establish cell-type specific regulatory regions by binding sequence motifs (hereafter referred to as motifs), opening up compacted chromatin, and allowing recruitment of regulatory machinery (Heinz et al. 2015). Once this machinery has been recruited, many regulatory regions (i.e., active promoters and active enhancers) are transcribed into mRNAs (in the case of promoters at active genes) or enhancer RNAs (eRNAs) at enhancers. While there is extensive expression of eRNAs across the genome, the precise function and guiding grammar of these sites remains largely unknown (Djebali et al. 2012; Andersson et al. 2014; W. Li, Notani, and Rosenfeld 2016).

Rather than acting one at a time, TFs act collaboratively (Heinz et al. 2010, 2015; Zhu, Shendure, and Church 2005). There have been several approaches to studying cooperative binding of TFs both experimentally and computationally. Experimentally, the most common method for investigating whether two TFs are co-bound is chromatin immunoprecipitation with deep sequencing (ChIP-seq) and similar methods. For example, the ENCODE project was able to investigate patterns of TF binding for a panel of TFs on a number of cell lines (Moore et al. 2020). However, this approach is highly limited because there are as many 1500 TFs encoded in the human genome and the most expansive collection of TF binding data (ENCODE) only had antibodies for 653 (the most mapped in any single cell line was 171) (Moore et al. 2020; Partridge et al. 2020; Lambert et al. 2018). Methods like sequential ChIP-seq (also known as re-ChIP) can profile whether two TFs collaborate by using successive rounds of antibody purification for both TFs to find sequences bound by both factors (Furey 2012). This approach can be highly laborious and suffers from the same limitations outlined above for conventional ChIP-seq. The final experimental approach is to perform a knockout on a given TF and examine how binding patterns of another factor (or factors) are affected. This approach is limited by the ability to knockout a given factor as well as the identification/immunoprecipitation of the hypothesized collaborating factor or design an applicable reporter construct.

Another approach is to computationally predict binding partners from ChIP-seq data for a single TF (Whitington et al. 2011; Levitsky et al. 2019). In these approaches, the top scoring motif among binding sites is inferred as the target motif for the TF in question and collaborative binding is inferred based upon the frequency, and other attributes such as orientation and spacing, of other motifs. This is more high-throughput than relying solely on experimental data, but it suffers from similar weaknesses to only requiring experimental evidence (i.e., the existence of the antibody for the TF in question). Further, this is limited to one target motif at a time and therefore limits the number of possible combinations of collaborative TFs.

Another approach to inferring TF collaboration is to utilize computational predictions of TF-binding events from open chromatin data as well as promoter annotations (van Bömmel et al. 2018; Meckbach et al. 2015; Vandenbon et al. 2012; Myšičková and Vingron 2012; Hu and Gallo 2010; Jankowski, Prabhakar, and Tiuryn 2014) (Compared in Table 4.1). Despite this diversity of methods, there are notable flaws. Of methods that mine sequence characteristics to predict co-occurring motifs, none of them incorporate expected motif frequencies. Instead, they use statistical measures such as Fisher's exact test (coTRaCTE) or pointwise mutual information (PC-TraFF) (van Bömmel et al. 2018; Meckbach et al. 2015). While this may allow for detection of co-occurrences above random expectation, it does not ensure that the motifs are enriched relative to the background frequencies of these motifs in the genome and, therefore, that the co-enrichment of two motifs is biologically significant.

In order to overcome existing methodological shortcomings in existing methods for TF co-occurrence, we developed a novel method that we term the dual HOMER approach to motif co-dependencies. HOMER performs motif finding by using an empirical estimation of motif frequencies in background sequences matched for GC content. It then scores the enrichment of motifs in a set of target sequences relative to these background sequences using the hypergeometric test (Heinz et al. 2010). This approach has the advantage of recovering enriched motifs in target sequences relative to the non-random background sequences (unlike the approaches cited above). This is useful in the context of the genome where there are known sequence biases.

Table 4.1. Summary of major motif co-occurrence methods.

Method	**Reference**	**Regions Profiled**	**Background Frequency Model**	**Directional Network**
Dual HOMER	This study	All	Random Genomic	Yes
PC-TraFF	Meckbach et al. (2015)	Promoters	Statistical	No
Frequency Ratio	Vandenbon et al. (2012)	Promoters	Statistical	No
TACO	Jankowski, Prabhakar, and Tiuryn (2014)	Cell-type DNase Hypersensivity sites	All DNase Hypersensivity sites	None Inferred
coTRaCTE	van Bömmel et al. (2018)	All	Statistical	No

Our approach utilizes HOMER by applying it in two rounds. In the first round, we screen target sequences for enriched motifs relative to a random genomic background. Subsequently, we take those enriched motifs and screen target sequences containing those motifs against random genomic background sequences containing those motifs (i.e., for a given motif, both the target and background sequences on the second round of HOMER contain that motif). This allows for a straightforward and interpretable co-enrichment calculation. It also avoids the problems outlined above by screening for enriched motifs in the first round of HOMER application.

We apply our dual HOMER method to transcription start site region (TSR) data generated using capped short RNA-seq (csRNA-seq) on a panel of 13 commonly used cell lines. We find that we are able to recapitulate known enrichments for cooperative TFs on both a global and cell-type specific level. We verified the accuracy of these co-enrichments using the STRING protein-protein interaction (PPI) network (Szklarczyk et al. 2019). Further, using the novel interpretation of TF cooperation implied by our method, we are able to dissect the nature of common TF cooperation. In all, we present a novel method for motif co-occurrence calculation that provides insight into the biology of TF cooperation in the context of TSRs.

4.3 Methods

4.3.1 Overview of Dual HOMER method

In brief, the Dual HOMER method works by executing two successive runs of the HOMER motif-finding algorithm for known motifs (illustrated schematically in Figure 4.2A) (Heinz et al. 2010). On the first run, for a given peak set, HOMER motif enrichments are run for a user-supplied motif file (or the HOMER library by default). The user can specify the number of background sequences in order to maintain balance between the target and background sequences on the second iteration. Results for this run are deposited in a user-specified directory. After this first run, Dual HOMER extracts enriched sequences as filtered by q-value (Benjamini-Hochberg corrected p-value; 0.01 by default).

Dual HOMER extracts all target and background sequences and finds all instances of all motifs using the “tab2fastq.pl” and “homer2 find” commands (Heinz et al. 2010). Then, for each enriched motif, the HOMER motif-finding algorithm is applied again for target and background sequences containing the motif in question. Each set of results is then deposited in a subdirectory that is named after the enriched motif.

4.3.2 Extraction of co-enrichments

Co-enrichments can be extracted using two main techniques. The first is to combine enrichment results into a matrix using HOMER’s “combineGO.pl” command (Heinz et al. 2010). This combines results from the enrichment outputs of each motif that is enriched on the initial HOMER run. In this matrix, the columns are the enriched motifs from the first run while the rows are the enrichments for all motifs from the second run.

The other technique is to construct a directed co-enrichment graph (here using Networkx (Hagberg, A.A., Shult, D.A., Swart, P.J. 2008). Each motif enriched in the first run represents a node and a directed edge is drawn from that node to every motif enriched in the second run of HOMER (q-value < 0.05 by default). This graph can be compressed into an undirected graph by requiring reciprocal directed edges. Edges in the undirected graph correspond to mutual dependency (and, therefore, cooperativity) between motifs. Networks were visualized using ipycytoscape. Communities were determined using Louvain clustering (Blondel et al. 2008).

4.3.3 Assessment of Fisher's exact test as a method of co-occurrence

We generated co-occurrence networks using the Fisher's exact test in a method similar to coTRaCTE (van Bömmel et al. 2018). TSRs were annotated for what motifs were present using the HOMER "annotatePeaks.pl" script (Heinz et al. 2010) The contingency table to determine co-occurrence between two motifs was set up as outlined in Table 1. Similar to above, each node represents a motif and each edge represents a significant co-occurrence (p-value < 0.05, odds ratio > 2). Networks were visualized using ipycytoscape. Communities were determined using Louvain clustering (Blondel et al. 2008).

Table 4.2. Contingency table for Fisher's exact test assessment of motif co-occurrence.

	TF A	
TF B	Both Occur	TF B Only
	TF A Only	Neither

4.3.4 Motif Library Generation

Due to substantial collinearity in the HOMER motif library, we eliminated redundant motifs from the default HOMER library using the compareMotifs.pl script using a similarity threshold of 0.6 (i.e., all motifs that were more similar were reduced into a single motif) (Heinz et al. 2010). This motif library was used throughout.

4.3.5 Identification of TSRs

csRNA-seq was performed using the protocol from Duttke et al. (2019) on the following cell lines: A549, GM12878, H9, hCMEC/D3, HCT116, HEK293T, HepG2, K562, MCF7, MDA-MB-231, OvCar8, THP-1, and U2OS (manuscript in preparation). Fastq files for both csRNA and input files were aligned using STAR aligner and processed into HOMER tag directories (Dobin et al. 2013; Heinz et al. 2010). TSRs were called using HOMER's findcsRNATSS.pl script and subsequently merged using the mergePeaks script (merge performed for replicates of the same cell line as well as for all TSRs) (Heinz et al. 2010).

4.3.6 Assessment of TF co-occurrences using PPI

In order to verify the accuracy of Dual HOMER co-occurrences, the interaction between TFs in the STRING PPI network was used (Szklarczyk et al. 2019). Motifs from our library (discussed above) were converted into their corresponding gene names which were then mapped onto STRING PPI protein IDs using Ensembl BioMart (Howe et al. 2021). The PPI was rendered into a graph using NetworkX for ease of interaction assessment (Hagberg, A.A., Shult, D.A., Swart, P.J. 2008).

We performed bootstrap analyses by randomly sampling pairs of TFs to create a random network 1,000,000 times to assess the performance of the network of all TSRs and 1000 times to assess the performance of the network of THP-1 cells. This gave us an empirical p-value to assess enrichments for physical interactions in a given network.

4.4 Results

4.4.1 Characterizing the motif composition of TSRs across 13 cell lines

In order to investigate how TF binding sites (TFBS) affect transcription initiation on a genome-wide level, we profiled transcription start site regions (TSRs) across 13 common cell lines (listed in Methods) using csRNA-seq. csRNA-seq is a method developed by our lab wherein short initiating RNAs are isolated and then subjected to high throughput sequencing (Figure 4.1A). Previous work in our lab has indicated that it is able to capture nascent transcription initiation at nucleotide resolution (Duttke et al. 2019). We found that cell lines had an average of 39,155 TSRs, which totaled to 171,726 unique TSRs across all of the cell lines (Figure 4.1B).

We then performed motif enrichment using HOMER on different groupings of TSRs (Heinz et al. 2010). First, we compared TSRs unique to each cell line to ubiquitous TSRs. We found that motifs that were enriched in individual cell lines corresponded to known lineage-determining TFs (LDTFs). For example, K562 TSRs showed a significant enrichment for the GATA2 motif, an LDTF for hematopoiesis and a hallmark of that cell line (Linnemann et al. 2011; Fujiwara et al. 2009). Similarly, HepG2 TSRs showed enrichment for HNF1b, HNF4a, and FOXA1 motifs, LDTFs for hepatocytes (Lau et al. 2018). In contrast, ubiquitous TSRs were enriched for general transcriptional activators such as Sp1, YY1, NFY, and NRF1 (Kaczynski, Cook, and Urrutia 2003; Gordon et al. 2005; Mantovani 1999; Y. Zhang and Xiang 2016). This same pattern emerges if we group TSRs by the number of cell lines in which they are present. Ubiquitous factors such as Sp1

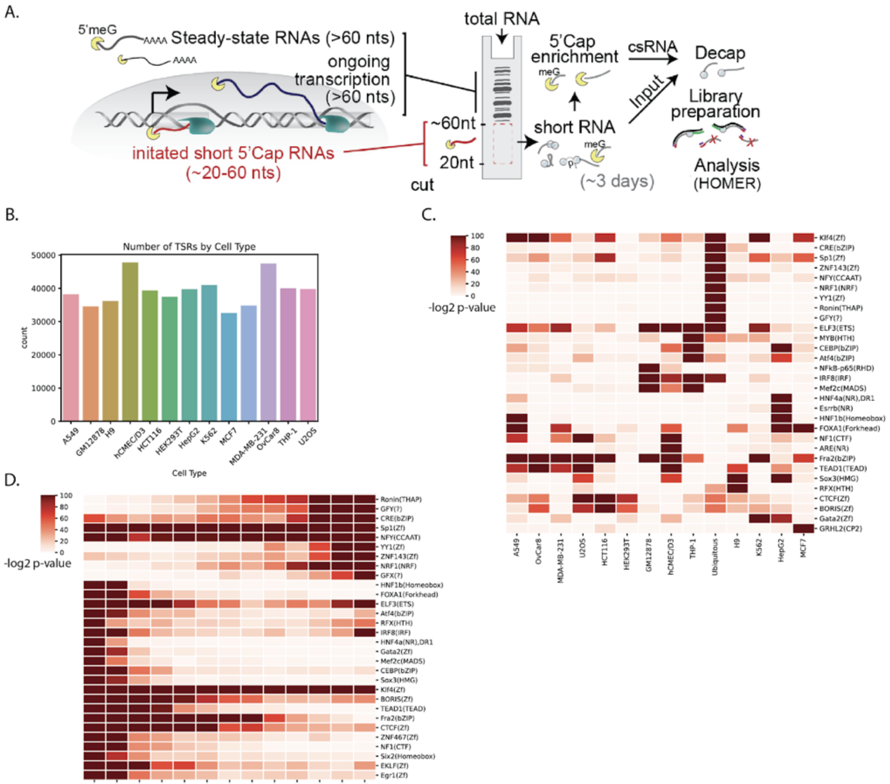


Figure 4.1. Initial assessment of motif enrichment in TSRs across 13 common cell lines. **(A)** Schematic diagram of csRNA-seq adapted from Duttke et al. (2018). Total RNA is size-selected and enriched for capped short RNAs then subjected to deep sequencing. Downstream analysis is performed by HOMER. **(B)** Counts of TSRs identified in each cell line. **(C)** Heatmap of HOMER motif enrichments for the top 30 most variant motifs for TSRs unique to each cell line as well as ubiquitous TSRs. **(D)** Heatmap of HOMER motif enrichments for the top 30 most variant motifs for TSRs grouped by how many cell lines in which they appear.

and NFY are found across a variety of cell lines while more specific factors such as HNF4a only occur in one or two. Interestingly, several factors such as NRF1 and YY1 are only enriched in ubiquitous peaks. This suggests that these factors regulate almost exclusively ubiquitously expressed genes and do not serve as general transcriptional activators for cell type specific expression.

4.4.2 Dual HOMER approach is a novel way of quantifying motif co-enrichments that recapitulates known interactions between TFs

Although normal motif enrichment analysis as performed above provided insight into the grammar of transcription initiation, it has been noted that TFs act collaboratively at regulatory regions (Heinz et al. 2010, 2015; Zhu, Shendure, and Church 2005). With this in mind, we examined motif co-occurrences at TSRs to examine TF collaboration in the context of transcription initiation. Given limitations in existing methods for computing motif co-occurrences, we developed our own method. Our method (covered in more detail in the Methods section) consists of consecutive applications of the HOMER motif finding algorithm. The first run of HOMER is on a user-provided set of TSRs. From this first run, all enriched motifs are collected as well as all sequences (target and background) that contain those motifs are collected. Then, for each enriched motif, HOMER motif finding is applied to target and background sequences containing that motif. Motifs that are enriched in this second round of HOMER are considered to be co-enriched. We call this the Dual HOMER method.

This co-enrichment strategy has several advantages to existing methods. Firstly, the strength of co-enrichment (as quantified in the second round of HOMER motif finding) has a directional component. From our example in our schematic

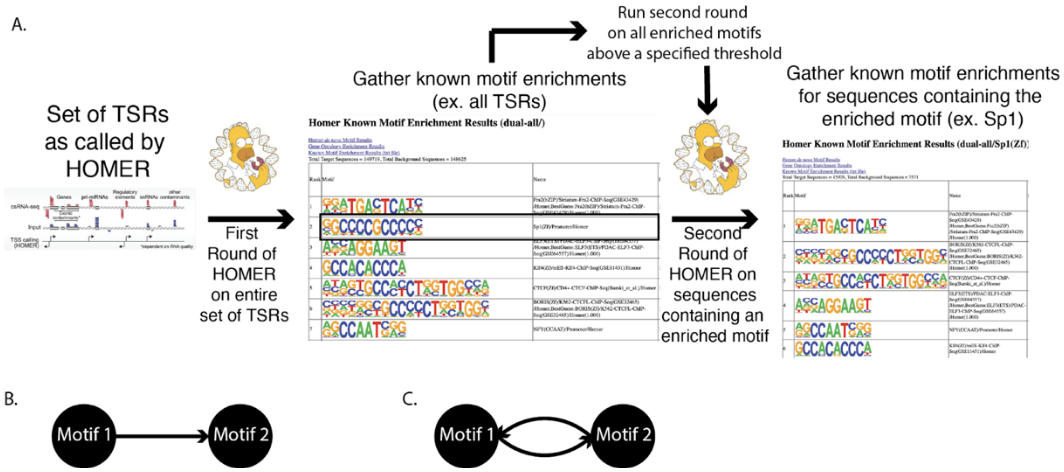


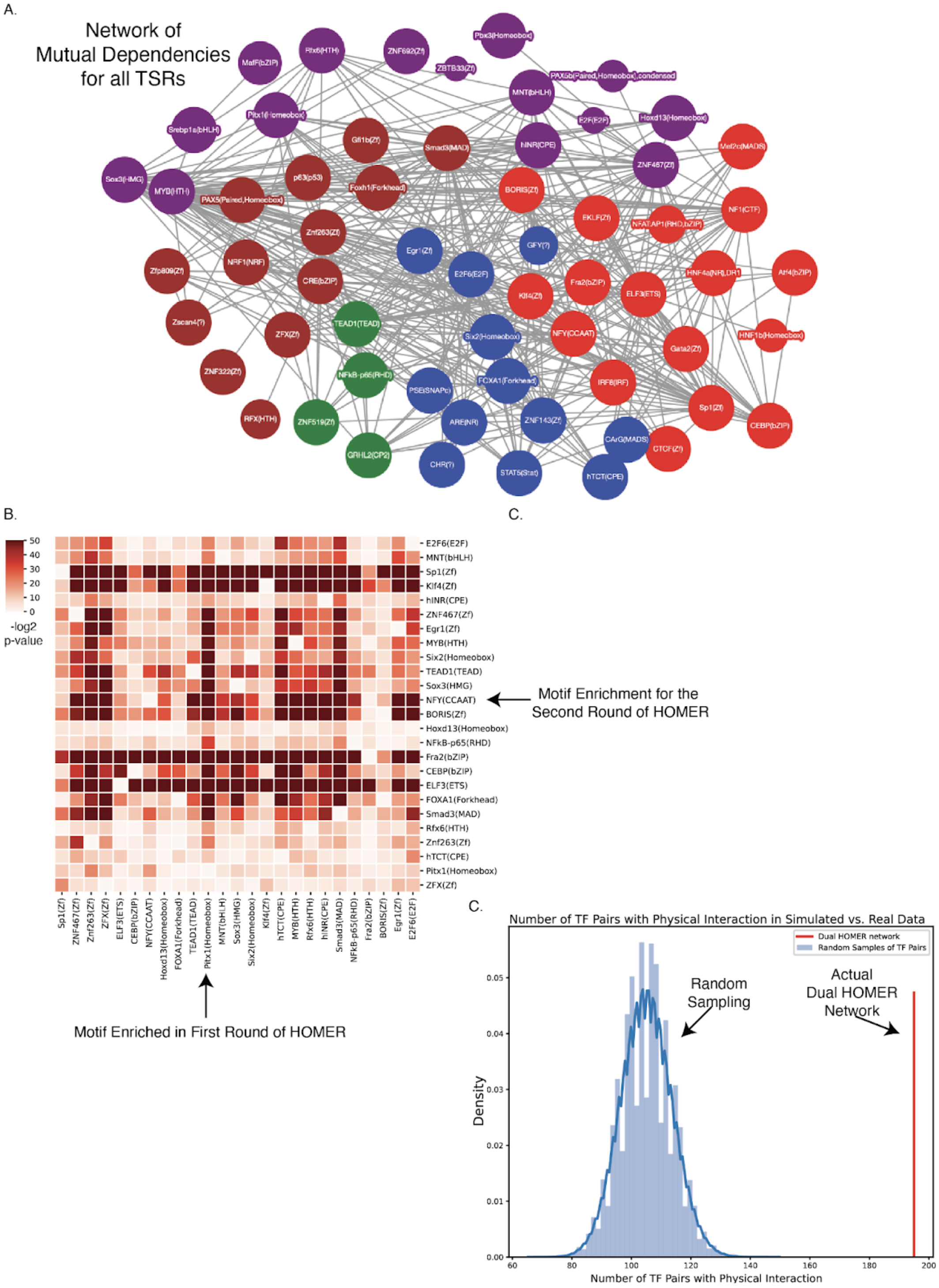
Figure 4.2. Outline of the Dual HOMER method**.** **(A)** Schematic of the Dual HOMER method. A user-provided set of peaks is subjected to an initial round of HOMER motif finding. Then, each enriched motif is subjected to another round of HOMER motif finding where target and background sequences possess the motif (Sp1 pictured here). **(B)** and **(C)** are diagrams of two different co-enrichment outcomes for enriched motifs. In **(B)**, Motif 2 is found to be enriched in peaks with Motif 1. This suggests that Motif 1 is dependent on Motif 2 for the given peaks. In **(C)**, both Motif 1 and Motif 2 are co-enriched in sequences containing either motif. This suggests that both motifs are dependent on each other and thus collaborative.

(Figure 4.2A), Sp1 is co-enriched with Fra2. These co-enrichments imply that the first motif (Sp1) is dependent on the second (Fra2) because the second motif is enriched in TSRs containing the first motif. This implies that TSRs with the Sp1 motif require the Fra2 motif for activation. Likewise, the enrichment for Fra2 in peaks containing Sp1 indicates the dependency of Fra2 on Sp1. If both motifs are dependent on each other, we call this a mutual dependency and reflective of a cooperative relationship. Figures 4.2B and 4.2C detail the nature of these relationships. To our knowledge, Dual HOMER is the only method to incorporate this notion of dependency.

We applied the Dual HOMER to the set of all TSRs as well as individual cell lines. We chose two different representations of co-enrichment. The first is a network of mutual co-dependencies (Figure 4.3A). Given the reasoning outlined above and in Figure 4.2C, we believe this best captures TF cooperativity in the context of motifs. We observed that many of the motifs that were enriched in the set of ubiquitous TSRs (i.e., various ETS factors, Sp1, NFY, and Fra2) segregated to the same community (red) while more cell-type specific motifs segregated to the other communities (Figure 4.3A).

The other representation we chose was a heatmap of co-enrichments as measured by the -log2 p-value of the second round of HOMER motif finding for each enriched motif (Figure 4.3B). This provides a quantitative insight into the dependency of motifs. Each column corresponds to an enriched motif in the first round of HOMER while each row represents motif enrichment in the second round of HOMER. In this framework, high enrichments along a column means that the motif has many dependencies (ex. Smad3) and high enrichments across a row means that the motif is depended upon (ex. Sp1). This provided different insights than the graph representation. General transcriptional activators such as Sp1, Fra2, and Klf4 have relatively modest dependencies (i.e., weaker enrichments in its column) but are a dependency for many factors (i.e., stronger enrichments in its row) (Figure 4.3B). By contrast, Smad3 had many dependencies but was rarely

Figure 4.3. Dual HOMER recapitulates known enrichments and lends insight into co-enrichments. (**A**) Network of mutual dependencies for the set of all TSRs. Motifs are nodes (scaled to number of occurrences) while mutual dependencies are edges. (**B**) Heatmap of co-enrichments. Columns are enriched motifs from the first round of HOMER motif finding and rows are HOMER motif enrichments in the second round (i.e., in TSRs containing the motifs in the column). (**C**) Bootstrap analysis of physical interactions between TFs in the Dual HOMER network. Random pairs of TFs were randomly sampled to create networks 1000 times and then assessed for the number of physical interactions. The Dual HOMER network was significantly enriched for physically-interacting TFs (p-value < 1e-6).


A.
Network of
Mutual Dependencies
for all TSRs
MafF(bZIP)
Rfx6(HTH)
ZNF692(Zf)
ZBTB33(Zf)
Pbx3(Homeobox)
PAX5b(Paired,Homeobox),condensed
MNT(bHLH)
E2F(E2F)
Hoxd13(Homeobox)
Srebp1a(bHLH)
Pitx1(Homeobox)
Gfi1b(Zf)
Smad3(MAD)
hINR(CPE)
Mef2c(MADS)
ZNF467(Zf)
Sox3(HMG)
MYB(HTH)
PAX5(Paired,Homeobox)
p63(p53)
Foxh1(Forkhead)
BORIS(Zf)
EKLF(Zf)
NFAT:AP1(RHD,bZIP)
NF1(CTF)
Znf263(Zf)
Egr1(Zf)
GFY(?)
Zfp809(Zf)
NRF1(NRF)
CRE(bZIP)
E2F6(E2F)
Klf4(Zf)
Fra2(bZIP)
ELF3(ETS)
HNF4a(NR),DR1
Atf4(bZIP)
Zscan4(?)
ZFX(Zf)
TEAD1(TEAD)
Six2(Homeobox)
NFY(CCAAT)
HNF1b(Homeobox)
Gata2(Zf)
ZNF322(Zf)
NFkB-p65(RHD)
PSE(SNAPc)
FOXA1(Forkhead)
IRF8(IRF)
RFX(HTH)
ZNF519(Zf)
ARE(NR)
ZNF143(Zf)
Sp1(Zf)
CEBP(bZIP)
CArG(MADS)
CTCF(Zf)
GRHL2(CP2)
CHR(?)
STAT5(Stat)
hTCT(CPE)
B.
C.
50
40
30
20
10
0
-log2
p-value
E2F6(E2F)
MNT(bHLH)
Sp1(Zf)
Klf4(Zf)
hINR(CPE)
ZNF467(Zf)
Egr1(Zf)
MYB(HTH)
Six2(Homeobox)
TEAD1(TEAD)
Sox3(HMG)
NFY(CCAAT)
BORIS(Zf)
Hoxd13(Homeobox)
NFkB-p65(RHD)
Fra2(bZIP)
CEBP(bZIP)
ELF3(ETS)
FOXA1(Forkhead)
Smad3(MAD)
Rfx6(HTH)
Znf263(Zf)
hTCT(CPE)
Pitx1(Homeobox)
ZFX(Zf)
Sp1(Zf)
ZNF467(Zf)
Znf263(Zf)
ZFX(Zf)
ELF3(ETS)
CEBP(bZIP)
NFY(CCAAT)
Hoxd13(Homeobox)
FOXA1(Forkhead)
TEAD1(TEAD)
Pitx1(Homeobox)
MNT(bHLH)
Sox3(HMG)
Six2(Homeobox)
Klf4(Zf)
hTCT(CPE)
MYB(HTH)
Rfx6(HTH)
hINR(CPE)
Smad3(MAD)
NFkB-p65(RHD)
Fra2(bZIP)
BORIS(Zf)
Egr1(Zf)
E2F6(E2F)
Motif Enrichment for the
Second Round of HOMER
Motif Enriched in First Round of HOMER
C.
Number of TF Pairs with Physical Interaction in Simulated vs. Real Data
Dual HOMER network
Random Samples of TF Pairs
Random
Sampling
Actual
Dual HOMER
Network
Density
0.05
0.04
0.03
0.02
0.01
0.00
60
80
100
120
140
160
180
200
Number of TF Pairs with Physical Interaction

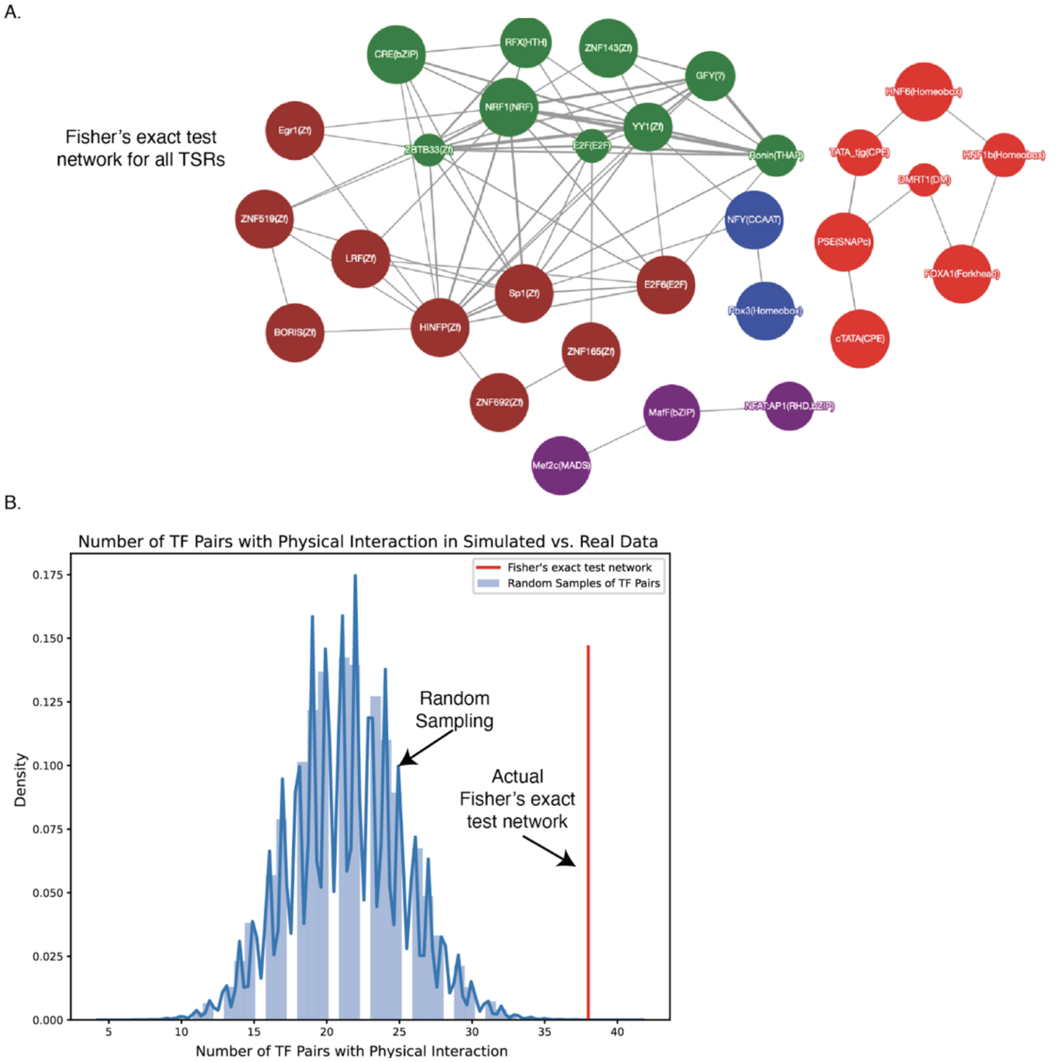


Figure 4.4. Assessment of the Fisher's exact test co-occurrence network for all TSRs. (**A**) Fisher's exact test network for all TSRs. Motifs are nodes (scaled to occurrences) while edges represent enriched co-occurrences. (**B**) Bootstrap analysis of physical interactions between TFs in the Fisher's exact test network. Random pairs of TFs were randomly sampled to create networks 1000 times and then assessed for the number of physical interactions. The Fisher's exact test network was significantly enriched for physically-interacting TFs (p-value = 2.9e-5).

depended on (stronger column enrichments than row enrichments) (Figure 4.3B). This is consistent with Smad3's known role as a signal-dependent TF (SDTF) in the TGF-β signaling cascade rather than a stand-alone transcriptional activator

(Massagué and Chen 2000). In general, motifs tended to either be depended on or have many dependencies) (Figure 4.3B).

Given that both of our representations visually recapitulated expected observations (Figure 34.A,B), we sought to validate our network of mutual dependencies. In order to do this, we utilized the STRING PPI of physically interacting proteins (Szklarczyk et al. 2019) to assess if mutually dependent motifs represented physically interacting TFs (i.e., edges in our network correspond to physical interaction in the PPI). We performed a bootstrap analysis wherein we randomly sampled pairs of TFs and compared the frequency of direct interactions in the PPI to the frequency of direct interactions in our network. We found that our network was highly enriched for direct interactions (Figure 4.3C). This validates the prediction of cooperativity in our network.

4.4.3 Dual HOMER outperforms the Fisher's exact test as a metric of co-occurrence

Although we were able to validate the Dual HOMER method, we sought to compare it to another common approach: the Fisher's exact test. This statistical test is commonly used to assess frequency of co-occurrence. The Fisher's exact test can be used to assess motif co-occurrence (e.g., coTRaCTE) as well as mutual exclusivity in cancer mutations (van Bömmel et al. 2018; Babur et al. 2015; Leiserson et al. 2015). We implemented this test on our data and compared results to the Dual HOMER method.

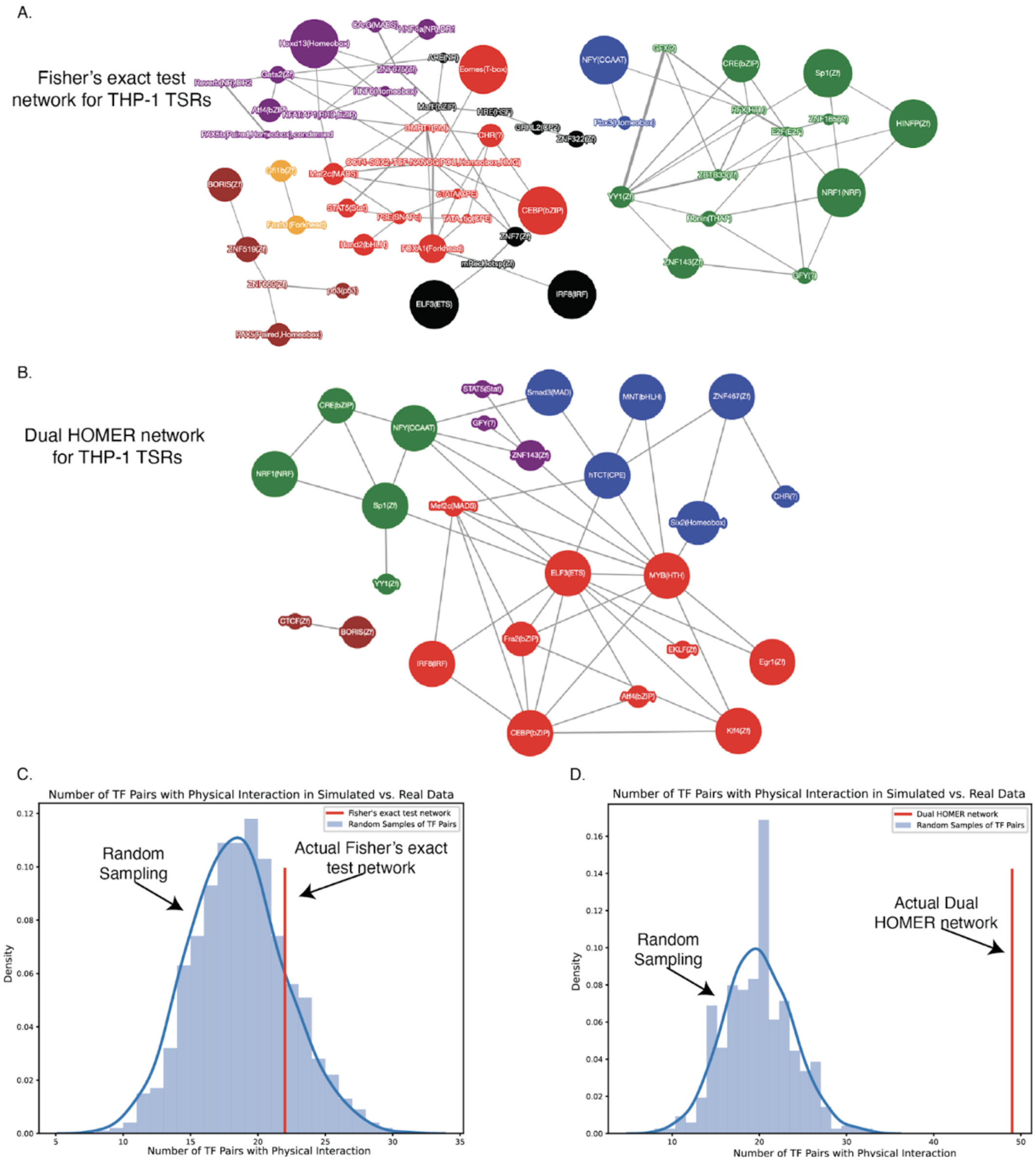
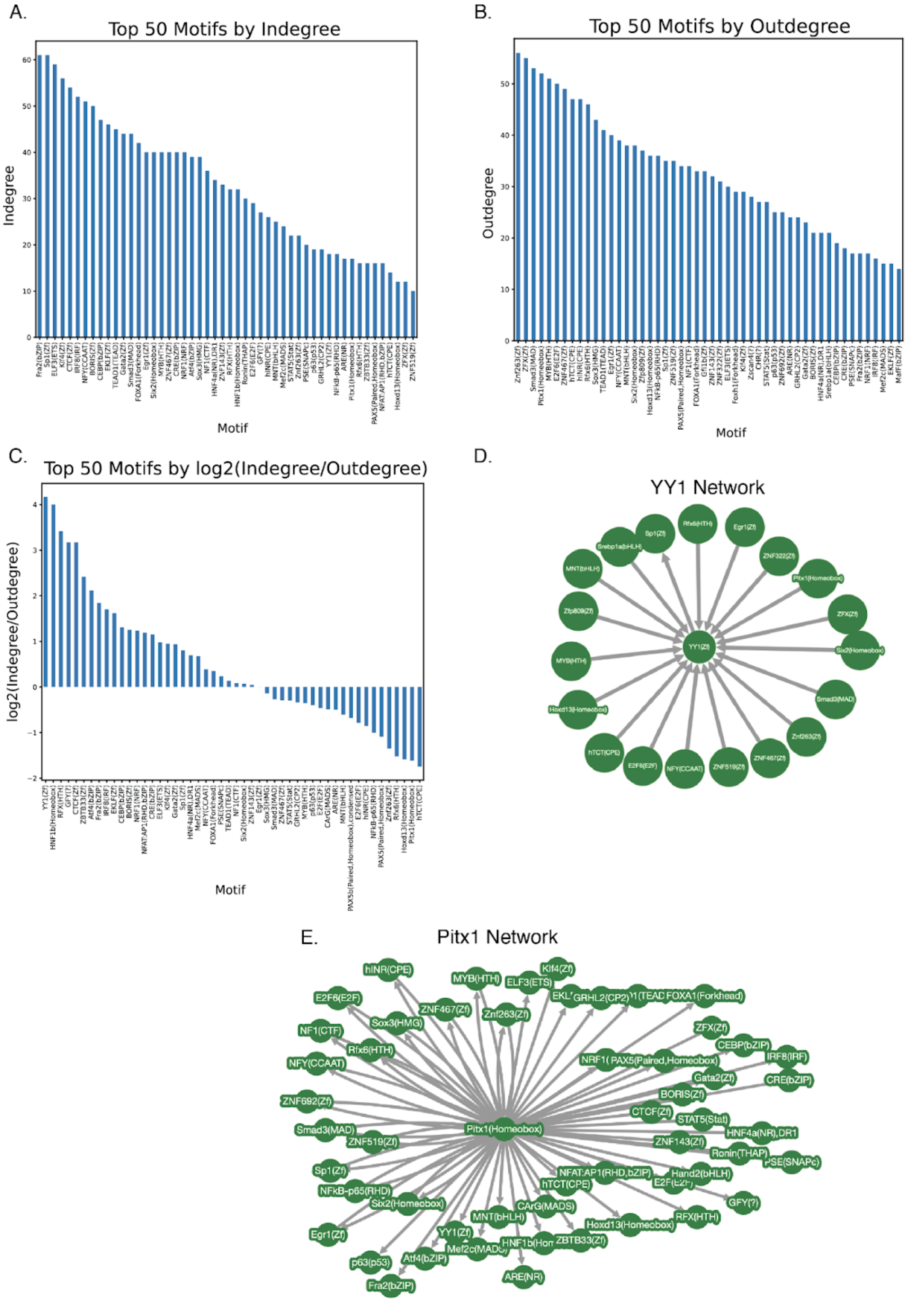


Figure 4.5. Comparison between Fisher's exact test and Dual HOMER THP-1 TSR networks. Fisher's exact test (**A)** and Dual HOMER (**B**) networks for THP-1 TSRs. Nodes are scaled to motif occurrences. Bootstrap analysis of physical interactions between TFs in the Fisher's exact test (**C**) and Dual HOMER (**D)** networks. Random pairs of TFs were randomly sampled to create networks 1000 times and then assessed for the number of physical interactions. The Fisher's exact test network was not significantly enriched for physically-interacting TFs (p-value = 0.19). The Dual HOMER network was significantly enriched (p-value < 0.001).

The Fisher's exact test network formed from examining all TSRs looked sparse but had some of the major motifs such as Sp1, NRF1, and YY1 (Figure 4.4A). However it was missing some notable motifs such as Fra2 and many of the ETS factors (Figure 4.4A). We checked the network to see if physical interactions in the STRING PPI were present, similar to the analysis of the Dual HOMER network. We found that the Fisher's exact test network was enriched for physical interactions (Figure 4.4B). This enrichment was about an order of magnitude lower than the enrichment for physical interactions in the Dual HOMER method (p-value = 2.9e-5 vs p-value < 1e-6) (Figures 4.3C,4.4B).

To get a better picture of the differences between the two networks, we compared them using TSRs identified in THP-1 cells (Figure 4.5). The Fisher's exact test network was considerably larger than the Dual HOMER network (Figure 4.5A,B). Many of the enrichments in the Fisher's exact network were for motifs such as HNF4a and HNF6 that are not enriched in THP-1 cells under a single HOMER run and are known to be liver-specific LDTFs (J. Li, Ning, and Duncan 2000; Nagaki and Moriwaki 2008; Samadani and Costa 1996; Hayhurst et al. 2001). In contrast, the Dual HOMER approach did not have these unexpected enrichments and showed expected enrichments in Egr1, Fra2 and Myb, which are known to be involved in monocyte differentiation and did not appear in the Fisher's exact test network (Figure 5.5B) (Matsui et al. 1990; Friedman 2007; Valledor et al. 1998; Krishnaraju, Hoffman, and Liebermann 2001).

Figure 4.6. Exploration of graph properties of the Dual HOMER network**.** (**A**) **I**ndegree of motifs for the top 50 motifs sorted by indegree. (**B**) Outdegree of motifs for the top 50 motifs sorted by outdegree. (**C**) log2(indegree/outdegree) of motifs for the top 50 motifs sorted by log2(indegree/outdegree). (**D**) Dual HOMER network for YY1. (**E**) Dual HOMER network for Pitx1.


A.
Top 50 Motifs by Indegree
Indegree
0
10
20
30
40
50
60
Fra2(bZIP)
Sp1(Zf)
ELF3(ETS)
Klf4(Zf)
CTCF(Zf)
IRF8(IRF)
NFY(CCAAT)
BORIS(Zf)
CEBP(bZIP)
EKLF(Zf)
TEAD1(TEAD)
Gata2(Zf)
Smad3(MAD)
FOXA1(Forkhead
Egr1(Zf)
Six2(Homeobox
MYB(HTH)
ZNF467(Zf)
CRE(bZIP)
NRF1(NRF)
Atf4(bZIP)
Sox3(HMG)
NF1(CTF)
HNF4a(NR),DR1
ZNF143(Zf)
RFX(HTH)
HNF1b(Homeobox
Ronin(THAP)
E2F6(E2F)
GFY(?)
hINR(CPE)
MNT(bHLH)
Mef2c(MADS)
STAT5(Stat)
Znf263(Zf)
PSE(SNAPc)
p63(p53)
GRHL2(CP2)
YY1(Zf)
NFkB-p65(RHD)
ARE(NR)
Pitx1(Homeobox
Rfx6(HTH)
ZBTB33(Zf)
PAX5(Paired,Homeobox
NFAT:AP1(RHD,bZIP)
hTCT(CPE)
Hoxd13(Homeobox
ZFX(Zf)
ZNF519(Zf)
Motif
B.
Top 50 Motifs by Outdegree
Outdegree
0
10
20
30
40
50
Znf263(Zf)
ZFX(Zf)
Smad3(MAD)
Pitx1(Homeobox
MYB(HTH)
E2F6(E2F)
ZNF467(Zf)
hTCT(CPE)
hINR(CPE)
Rfx6(HTH)
Sox3(HMG)
TEAD1(TEAD)
Egr1(Zf)
NFY(CCAAT)
MNT(bHLH)
Six2(Homeobox
Zfp809(Zf)
Hoxd13(Homeobox
NFkB-p65(RHD)
Sp1(Zf)
ZNF519(Zf)
PAX5(Paired,Homeobox
NF1(CTF)
FOXA1(Forkhead
Gfi1b(Zf)
ZNF143(Zf)
ZNF322(Zf)
ELF3(ETS)
Foxh1(Forkhead
Klf4(Zf)
Zscan4(?)
CHR(?)
STAT5(Stat)
p63(p53)
ZNF692(Zf)
ARE(NR)
GRHL2(CP2)
Gata2(Zf)
BORIS(Zf)
HNF4a(NR),DR1
Srebp1a(bHLH)
CEBP(bZIP)
CRE(bZIP)
PSE(SNAPc)
Fra2(bZIP)
NRF1(NRF)
IRF8(IRF)
Mef2c(MADS)
EKLF(Zf)
MafF(bZIP)
Motif
C.
Top 50 Motifs by log2(Indegree/Outdegree)
log2(Indegree/Outdegree)
−2
−1
0
1
2
3
4
YY1(Zf)
HNF1b(Homeobox
RFX(HTH)
GFY(?)
CTCF(Zf)
ZBTB33(Zf)
Atf4(bZIP)
Fra2(bZIP)
IRF8(IRF)
EKLF(Zf)
CEBP(bZIP)
BORIS(Zf)
NRF1(NRF)
NFAT:AP1(RHD,bZIP)
CRE(bZIP)
ELF3(ETS)
Klf4(Zf)
Gata2(Zf)
Sp1(Zf)
HNF4a(NR),DR1
Mef2c(MADS)
NFY(CCAAT)
FOXA1(Forkhead
PSE(SNAPc)
TEAD1(TEAD)
NF1(CTF)
Six2(Homeobox
ZNF143(Zf)
Egr1(Zf)
Sox3(HMG)
Smad3(MAD)
ZNF467(Zf)
STAT5(Stat)
GRHL2(CP2)
MYB(HTH)
p63(p53)
E2F(E2F)
CArG(MADS)
ARE(NR)
MNT(bHLH)
PAX5b(Paired,Homeobox),condensed
E2F6(E2F)
hINR(CPE)
NFkB-p65(RHD)
PAX5(Paired,Homeobox
Znf263(Zf)
Rfx6(HTH)
Hoxd13(Homeobox
Pitx1(Homeobox
hTCT(CPE)
Motif
D.
YY1 Network
YY1(Zf)
Sp1(Zf)
Rfx6(HTH)
Egr1(Zf)
Srebp1a(bHLH)
ZNF322(Zf)
MNT(bHLH)
Pitx1(Homeobox)
Zfp809(Zf)
ZFX(Zf)
MYB(HTH)
Six2(Homeobox)
Smad3(MAD)
Hoxd13(Homeobox)
Znf263(Zf)
hTCT(CPE)
ZNF467(Zf)
E2F6(E2F)
NFY(CCAAT)
ZNF519(Zf)
E.
Pitx1 Network
Pitx1(Homeobox)
hINR(CPE)
Klf4(Zf)
MYB(HTH)
ELF3(ETS)
E2F6(E2F)
EKL
GRHL2(CP2)
TEAD
FOXA1(Forkhead)
ZNF467(Zf)
Znf263(Zf)
Sox3(HMG)
ZFX(Zf)
NF1(CTF)
CEBP(bZIP)
Rfx6(HTH)
IRF8(IRF)
NRF1
PAX5(Paired,Homeobox)
NFY(CCAAT)
Gata2(Zf)
CRE(bZIP)
BORIS(Zf)
ZNF692(Zf)
CTCF(Zf)
STAT5(Stat)
Smad3(MAD)
HNF4a(NR),DR1
ZNF519(Zf)
ZNF143(Zf)
Ronin(THAP)
PSE(SNAPc)
Sp1(Zf)
NFAT:AP1(RHD,bZIP)
Hand2(bHLH)
NFkB-p65(RHD)
hTCT(CPE)
E2F(E2F)
Six2(Homeobox)
CArG(MADS)
GFY(?)
MNT(bHLH)
Hoxd13(Homeobox)
RFX(HTH)
Egr1(Zf)
YY1(Zf)
HNF1b(Hom
ZBTB33(Zf)
Mef2c(MADS
Atf4(bZIP)
p63(p53)
ARE(NR)
Fra2(bZIP)

Finally, we compared the Fisher's exact test and Dual HOMER networks using the STRING PPI. We found that the Dual HOMER network had significant enrichment for physical interactions while the Fisher's exact test network did not (p-value < 0.001 vs. p-value = 0.19, respectively) (Figure 4.5C,D). This confirms that the unexpected enrichments likely represent false positive interactions. In all, the Dual HOMER method outperforms the Fisher's exact test in identifying cooperative TF pairs.

4.4.4 Examination of graph properties of Dual HOMER network lends insight into differences between general and cell-type specific transcriptional activators

Given that our approach recapitulated known patterns of TF cooperativity, we sought to investigate the properties of our motif network in order to better understand the grammar of transcription initiation. As noted earlier, our network can be rendered as a directed graph where, for a given motif, outbound edges indicate dependencies that motif has while inbound edges represent motifs that depend on the target motif (Figure 4.2B,C). This directional aspect of interactions is not captured by any existing method for co-occurrence inference and lends for unique interpretational power (Table 4.1).

With this in mind, we investigated the graph properties of motifs in the Dual HOMER network. We plotted the indegree, outdegree, and log2(indegree/outdegree) of motifs in our network (Figure 4.6A,B,C). We found that general transcriptional activators had high indegree while tissue-specific TFs had high outdegree (Figure 4.6A,B). In general, we found more general TFs had a higher log2(indegree/outdegree) value than TFs (Figure 4.6C). This trend makes sense as general transcriptional activators are depended upon by many factors for activation while more tissue-specific TFs likely retain their specificity by not initiating transcription without the proper epigenomic context and coactivators.

We wanted to further examine the relationship of indegree and outdegree with some examples. We chose YY1 and Pitx1 as examples of TFs that have a high and low log2(indegree/outdegree) (4.17 vs. -1.80), respectively (Figure 4.6C-E). YY1 is noted as a general transcriptional activator so it's higher indegree is expected as it is depended upon frequently and in many different contexts (Gordon et al. 2005; Verheul et al. 2020). Interestingly, YY1 had only one dependency, Sp1 (Figure 4.6D). This is notable as Sp1 has been found to physically interact with YY1 and amplify transcription initiation where this interaction takes place (Seto, Lewis, and Shenk 1993; Lee, Galvin, and Shi 1993). Manual examination of Sp1's co-enrichments found that YY1's enrichment in peaks with Sp1 was q-value = 0.011 when the cutoff was q-value < 0.01 so this may be a mutual dependency.

On the opposite end of the spectrum, Pitx1 had many dependencies and was rarely depended upon (Figure 4.6C,E). When Pitx1 was depended upon, it was only in mutual dependencies (Figure 4.6E). Pitx1 is multifaceted homeobox TF that is involved with many developmental and disease processes. It is known to be critical in several developmental processes, most notably the pituitary gland

and hindlimb area (Tran and Kioussi 2021). It is a tumor suppressor gene in the context of breast cancer due to its ability to repress ERα (Stender et al. 2011). This repressor function may explain Pitx1's dependence on CTCF as both are known to repress TERT expression, although collaboration in the context of this phenotype has not been characterized to our knowledge (Qi et al. 2011; Stender et al. 2011).

4.5 Discussion

Here we outline a novel method for detecting motif co-occurrences and inferring TF collaboration. The Dual HOMER method detects co-occurrences by utilizing successive rounds of HOMER motif enrichment. In the first round of motif enrichment, Dual HOMER identifies enriched motifs in target sequences against a random genomic background. In subsequent rounds, Dual HOMER iterates through these motifs and identifies which motifs are co-enriched in target and background sequences containing the original motif (schematically outlined in Figure 4.2A). We demonstrate that this method effectively recapitulates known physical interactions between TFs and outperforms the Fisher's exact test as a measurement of co-occurrence (Figures 4.3C,4.5). Further, Dual HOMER's output of co-enrichment of motifs has a directional interpretation, which lends to unique biological insights (Figure 4.2B,C).

The novelty of the Dual HOMER method is two-fold. First, it employs motif enrichment analysis against background sequences in order to infer co-occurrences. One other method employs this strategy (MCOT), however, it limits

its co-enrichment strategy to one primary motif at a time (presumably, the factor of interest in a ChIP-seq experiment) and background sequences are the results of permutations of target sequences rather than random genomic background (as in Dual HOMER's approach) (Levitsky et al. 2019). By leveraging motif enrichment in target sequences relative to genomic background, the Dual HOMER method is able to detect important motifs in the target sequence while ruling out motifs that occur by chance relative to background frequencies. An example of this is in the comparison between the Dual HOMER and Fisher's exact test networks in TSRs occurring in the THP-1 cell line (Figure 4.5A,B). The Fisher's exact test network had enrichment for TFs that are LDTFs for other cell types (e.g., HNF4a and HNF6, both LDTFs for the hepatocyte lineage) and was missing critical factors to the monocyte lineage that were present in the Dual HOMER network (e.g., Egr1, Fra2 and Myb) (J. Li, Ning, and Duncan 2000; Nagaki and Moriwaki 2008; Samadani and Costa 1996; Hayhurst et al. 2001; Matsui et al. 1990; Friedman 2007; Valledor et al. 1998; Krishnaraju, Hoffman, and Liebermann 2001). In this scenario, Dual HOMER is both more sensitive to critical LDTFs and does not include as many false discoveries.

The other major novelty of the Dual HOMER method is the directionality of the network. In our network, directed edges can be drawn from one motif to another if the second motif is enriched in HOMER enrichment for sequences containing the first (schematically outlined in Figure 4.2B,C). This aspect of the network is biologically realistic as some TFs are general transcriptional activators (e.g., Sp1) while others are cell-type and/or stimulus specific (HNF factors). Further, our notion of mutually dependent TFs is consistent with collaborative TF binding from a conceptual standpoint (Heinz et al. 2015, 2010; Zhu, Shendure, and Church 2005). We were able to use the properties of the directed graph to characterize TFs. We found that TFs that had many dependencies tended to be cell-type specific and signal-dependent TFs (ex. Pitx1) while TFs that were frequently depended upon tended to be general transcriptional activators (ex. YY1) (Figure 4.6). Given this property of the Dual HOMER network, it allows for more precise biological characterization of TF networks.

Interestingly, Dual HOMER markedly outperformed a Fisher's exact test network in the context of a cell-type specific network (i.e., all TSRs in the THP-1 cell line; discussed above) (Figure 4.5). This is in contrast to the performance of the two methods on the network of all TSRs where Dual HOMER only slightly outperformed the Fisher's exact test (Figures 4.3,4.4). This difference can be explained by Dual HOMER filtering for enriched motifs. In the context of the global set of TSRs, filtering for enriched motifs is not as critical since there is a relatively large search space (171,726 total TSRs). However, in a single cell line (e.g., THP-1), there are fewer TSRs (40,011), which likely skews expected motif frequency estimations when not factoring in genomic background. Further, intuitively, in a cell-type specific context, there is a prior expectation of enrichment for LDTFs since these determine cell identity (Heinz et al. 2015). This represents a distinct advantage for the Dual HOMER method and may indicate that it is most useful when applied to cell-type specific data.

Dual HOMER's ability to recover critical information about cell-type specific TF cooperativity could yield benefits for understanding disease states. As an example, the development of cancer can be linked to the rewiring of regulatory networks (Islam et al. 2021; Melton et al. 2015) It has also recently been observed that variants for heritable traits do not restrict themselves to expected causal pathways but instead exert regulatory influence over the entire genome (termed the omnigenic model of heritable disease) (Boyle, Li, and Pritchard 2017). Under this model, most heritable variants affect gene regulatory networks, which then affect the aforementioned causal pathways. In both of these cases, Dual HOMER could be deployed to compare disease states with healthy states in order to understand how transcriptional regulatory networks differ between these two states. This could lend insight into mechanisms of disease and possible drug targets. In all, the Dual HOMER method represents a step forward in detecting motif co-occurrences that recapitulates known TF cooperativity and lends unique interpretability.

4.6 Acknowledgements

Thanks to Sascha Duttke for generating the data and for helpful analysis discussions.

Chapter 4, in part, is material being prepared for submission contributed by Samuel J. Roth, Sascha H.C. Duttke, Christopher Benner. The dissertation author was one of the primary investigators and authors of this paper.

CHAPTER 5 CONCLUSION

Both transcription termination and transcription initiation are fundamental processes in gene regulation. Here I investigated key aspects of both of these processes. In the context of transcription termination, I developed ARTDeco, a software package aimed at characterizing and quantifying readthrough transcription. Defects of transcription termination (DoTT) are a relatively novel phenotype. I was able to quantify readthrough transcription using three different metrics: read-in level, readthrough level, and downstream of gene (DoG) transcript discovery (Figure 2.1C). Further, we show that these measures can discriminate between systems with readthrough transcription and those without it (Figure 2.2A,B,D). This represents a major advance in the study of systems with DoTT. Previous packages aimed at quantifying readthrough could only perform the function of DoG discovery (Melnick et al. 2019; Wiesel, Sabath, and Shalgi 2018). ARTDeco not only outperformed these packages in terms of runtime, DoGs discovered by ARTDeco showed more signs of transcription (Table 2.1, Supplementary Figure 2.3).

Another major advancement that ARTDeco provides is the identification of read-in genes. Read-in genes are genes that are transcribed due to readthrough transcription rather than promoter activation. I show that ARTDeco can infer whether a gene is a read-in or primary induction gene using the read-in level as a metric (Figure 2.3). Further, read-in genes were shown to represent functional noise (Figure 2.4). This is critical for analysis of gene expression in systems with

DoTT. Given that many analysis utilize differential expression to curate gene sets, it is critical to know whether there is a polluting signal such as read-in genes. The read-in level can also be used to denoise gene expression estimations. Because the read-in level represents the relative contribution of upstream readthrough to gene expression, it can be used to remove upstream readthrough as a source of noise (Supplementary Figure 2.2).

The final application of ARTDeco in Chapter 2 involved deploying it on a population-level study of gene expression response to influenza A virus (IAV). I showed that ARTDeco was able to successfully quantify readthrough in this data set and that patterns of readthrough matched expectations based upon the original study (Figure 2.6A-C) (Quach et al. 2016). More importantly, it was demonstrated that variants that modulated the expression of read-in genes were enriched for affecting the upstream gene (Figure 2.6D-F). This is significant as it shows that readthrough can pollute eQTL enrichments in systems with readthrough.

In Chapter 3, I extend ARTDeco's functionality into a discovery pipeline for systems with DoTT (Figure 3.1). ARTDeco discovered readthrough in three viruses: Rift Valley Fever virus (RVFV), Zika virus (ZIKV), and Sindbis virus (SINV) (Figure 3.2A-C). This more than doubles the known viruses that induce DoTT. I examined the literature on these viruses in order to hypothesize mechanisms of readthrough for each of these viruses. I hypothesized that the NSs protein caused DoTT in RVFV due to its role in host transcription shutoff (Le May et al. 2008; Ly and Ikegami 2016; Billecocq et al. 2004; Bouloy et al. 2001). Host transcription shutoff is among the only phenotypes shared between the two viruses previously known to cause DoTT, IAV and herpes simplex virus 1 (HSV-1) (Nogales et al. 2018; Khaperskyy and McCormick 2015; He et al. 2020).

I was able to demonstrate that NSs induces readthrough transcription by expressing the NSs protein in THP-1 monocytes and subsequently performing total RNA-seq (Figure 3.4A-C). This adds a crucial validation step to the existing discovery pipeline. I compared the readthrough induced by NS1 (the causative protein in IAV) to the readthrough induced by NSs. Both proteins induced readthrough on a global level with NS1 inducing slightly more readthrough. This suggests that NSs interferes with transcription termination machinery in a manner similar to NS1. In the future, it will be informative to express the ICP27 protein (the cause of DoTT in HSV-1) and compare patterns of readthrough between all of the proteins. It is possible that differences between these proteins cause distinct patterns of readthrough.

I also examined gene expression in THP-1 cells expressing the NSs and NS1 proteins and found that the host immune response was downregulated in distinct ways (Figure 3.5). This difference had not been noted elsewhere in the literature. Further, this suggests that both of these proteins mount a dual-pronged attack on host transcription wherein both gene activation and transcription termination are attacked. Future experiments are needed to validate and characterize this phenomenon.

In Chapter 4, I introduce a new method of detecting motif co-occurrences. We applied this method to csRNA-seq data for 13 commonly used cell lines in order to profile transcription factor (TF) collaboration at transcription start site regions (TSRs). This method provides several unique aspects. First, no existing method of motif co-occurrences incorporates motif enrichment of target sequences against genomic background sequences. In addition to this, the Dual HOMER method offers a novel presentation of a transcriptional network. To date, it is the only motif co-occurrence technique that returns a directional network (pictured in Figure 4.2B,C). This directionality arises from the nature of the method. Dual HOMER runs successive rounds of motif enrichment (schematically outlined in Figure 4.2A). The first round is on all TSRs. Subsequent rounds are on a subset of TSRs that contain an enriched motif from the first round as well as background sequences containing that motif. This second round of enrichment establishes the dependency of the motif being queried (i.e., in the sequences) on this round of enriched motifs (schematically discussed in Figure 4.2B,C).

This dependency relationship lends to direct interpretability of the transcriptional network generated. I noted that general transcriptional activators tend to be depended upon more often while cell-type specific and signal-dependent TFs tend to have more dependencies (Figure 4.6). When examining individual TFs, this gave unique insights. For example, YY1's only dependency was to Sp1 (Figure 4.6D). This is in line with literature that these two factors physically interact and that Sp1 specifically increases YY1's ability to activate transcription (Seto, Lewis, and Shenk 1993; Lee, Galvin, and Shi 1993). The directional aspect of Dual HOMER's network allows for these unique insights and may shed light on TF interactions at a mechanistic level.

I validated the performance of this network using the STRING protein-protein interaction network (PPI) (Szklarczyk et al. 2019). I found that mutual dependencies in the Dual HOMER network for both all TSRs and TSRs specific to the THP-1 cell line were enriched for physical interactions between TFs (Figures 4.3C,4.4D). Further, the Dual HOMER network outperformed another common statistical measure of co-occurrence (the Fisher's exact test) (Figures 4.3C,4.4,4.5). This was especially the case for TSRs occurring in the THP-1 cell line where the Fisher's exact test network performed exceptionally poorly in comparison to the Dual HOMER network (Figure 4.5). This is likely because in the context of fewer TSRs motif enrichment becomes more important for ruling out false co-occurrences. This makes sense in the context of the THP-1 cell line TSRs where the Fisher's exact test network found co-occurrences between motifs whose corresponding TFs are not expressed in that lineage (Figure 4.5A).

Finally, it is important to note the potential translational impact of these findings. ARTDeco's ability to denoise gene expression data in the context of DoTT will be immensely useful in characterizing disease states that cause this phenotype. By quantifying and understanding how DoTT manifests, this can help identify biomarkers as well as druggable targets. Further, it will aid in discovering the full scope of stresses that induce readthrough transcription. I outlined a pipeline that uses ARTDeco to discover these stresses in the context of viruses. My work in this thesis demonstrates that, with the aid of ARTDeco, it is possible to identify and confirm the cause readthrough in viruses. Given that DoTT is likely part of cytopathogenesis, the proteins responsible for this phenotype represent potential druggable targets. Finally, Dual HOMER could potentially be utilized to investigate how transcriptional networks are rewired in disease states. This holds most potential for cancer wherein it is well known that somatic mutations substantially rewire transcriptional networks. Perhaps Dual HOMER's network can lend insight into what dependencies are present and this may help identify druggable genes or pathways.

REFERENCES

Adouchief, Samuel, Teemu Smura, Jussi Sane, Olli Vapalahti, and Satu Kurkela. 2016. “Sindbis Virus as a Human Pathogen-Epidemiology, Clinical Picture and Pathogenesis.” *Reviews in Medical Virology* 26 (4): 221–41.

Akhrymuk, Ivan, Sergey V. Kulemzin, and Elena I. Frolova. 2012. “Evasion of the Innate Immune Response: The Old World Alphavirus nsP2 Protein Induces Rapid Degradation of Rpb1, a Catalytic Subunit of RNA Polymerase II.” *Journal of Virology* 86 (13): 7180–91.

Altenhoff, Adrian M., Romain A. Studer, Marc Robinson-Rechavi, and Christophe Dessimoz. 2012. “Resolving the Ortholog Conjecture: Orthologs Tend to Be Weakly, but Significantly, More Similar in Function than Paralogs.” *PLoS Computational Biology* 8 (5): e1002514.

Andersson, Robin, Claudia Gebhard, Irene Miguel-Escalada, Ilka Hoof, Jette Bornholdt, Mette Boyd, Yun Chen, Xiaobei Zhao, Christian Schmidl2, Takahiro Suzuki, Evgenia Ntini, Erik Arner, Eivind Valen, Kang Li, Lucia Schwarzfischer, Dagmar Glatz, Johanna Raithel, Berit Lilje, Nicolas Rapin, Frederik Otzen Bagger, Mette Jørgensen, Peter Refsing Andersen, Nicolas Bertin, Owen Rackham, A. Maxwell Burroughs, J. Kenneth Baillie, Yuri Ishizu, Yuri Shimizu, Erina Furuhata, Shiori Maeda, Yutaka Negishi, Christopher J. Mungall, Terrence F. Meehan, Timo Lassmann, Masayoshi Itoh, Hideya Kawaji, Naoto Kondo6,14, Jun Kawai6,14, Andreas Lennartsson15, Carsten O. Daub, Peter Heutink, David A. Hume, Torben Heick Jensen, Harukazu Suzuki, Yoshihide Hayashizaki, Ferenc Mueller4 , The FANTOM Consortium, Alistair R. R. Forrest, Piero Carninci, Michael Rehli, and Albin Sandelin. 2014. “An Atlas of Active Enhancers across Human Cell Types and Tissues.” *Nature* 507 (7493): 455–61.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. 2000. “Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium.” *Nature Genetics* 25 (1): 25–29.

Babur, Özgün, Mithat Gönen, Bülent Arman Aksoy, Nikolaus Schultz, Giovanni Ciriello, Chris Sander, and Emek Demir. 2015. “Systematic Identification of Cancer Driving Signaling Pathways Based on Mutual Exclusivity of Genomic Alterations.” *Genome Biology* 16 (February): 45.

Barrett, Tanya, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, Kimberly A. Marshall, Katherine H. Phillippy, Patti M. Sherman, Michelle Holko, Andrey Yefanov, Hyeseung Lee, Naigong Zhang, Cynthia L. Robertson, Nadezhda Serova, Sean Davis, and Alexandra Soboleva. 2013. “NCBI GEO: Archive for Functional Genomics Data Sets--Update.” *Nucleic Acids Research* 41 (Database issue): D991–95.

Bauer, David L. V., Michael Tellier, Mónica Martínez-Alonso, Takayuki Nojima, Nick J. Proudfoot, Shona Murphy, and Ervin Fodor. 2018. “Influenza Virus Mounts a Two-Pronged Attack on Host RNA Polymerase II Transcription.” *Cell Reports*. https://doi.org/10.1016/j.celrep.2018.04.047.

Bercovich-Kinori, Adi, Julie Tai, Idit Anna Gelbart, Alina Shitrit, Shani Ben-Moshe, Yaron Drori, Shalev Itzkovitz, Michal Mandelboim, and Noam Stern-Ginossar. 2016. “A Systematic View on Influenza Induced Host Shutoff.” *eLife* 5 (August). https://doi.org/10.7554/eLife.18311.

Billecocq, Agnès, Martin Spiegel, Pierre Vialat, Alain Kohl, Friedemann Weber, Michèle Bouloy, and Otto Haller. 2004. “NSs Protein of Rift Valley Fever Virus Blocks Interferon Production by Inhibiting Host Gene Transcription.” *Journal of Virology* 78 (18): 9798–9806.

Blondel, Vincent D., Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. “Fast Unfolding of Communities in Large Networks.” *Journal of Statistical Mechanics: Theory and Experiment*. https://doi.org/10.1088/1742-5468/2008/10/p10008.

Bömmel, Alena van, Michael I. Love, Ho-Ryun Chung, and Martin Vingron. 2018. “coTRaCTE Predicts Co-Occurring Transcription Factors within Cell-Type Specific Enhancers.” *PLoS Computational Biology* 14 (8): e1006372.

Bouloy, M., C. Janzen, P. Vialat, H. Khun, J. Pavlovic, M. Huerre, and O. Haller. 2001. “Genetic Evidence for an Interferon-Antagonistic Function of Rift Valley Fever Virus Nonstructural Protein NSs.” *Journal of Virology* 75 (3): 1371–77.

Boyle, Evan A., Yang I. Li, and Jonathan K. Pritchard. 2017. “An Expanded View of Complex Traits: From Polygenic to Omnigenic.” *Cell* 169 (7): 1177–86.

Cardiello, Joseph F., James A. Goodrich, and Jennifer F. Kugel. 2018. “Heat Shock Causes a Reversible Increase in RNA Polymerase II Occupancy Downstream of mRNA Genes, Consistent with a Global Loss in Transcriptional Termination.” *Molecular and Cellular Biology* 38 (18). https://doi.org/10.1128/MCB.00181-18.

Carlin, Aaron F., Edward A. Vizcarra, Emilie Branche, Karla M. Viramontes, Lester Suarez-Amaran, Klaus Ley, Sven Heinz, Christopher Benner, Sujan Shresta, and Christopher K. Glass. 2018. “Deconvolution of pro- and Antiviral

Genomic Responses in Zika Virus-Infected and Bystander Macrophages." *Proceedings of the National Academy of Sciences of the United States of America* 115 (39): E9172–81.

Chen, Xiaoyong, Shasha Liu, Mohsan Ullah Goraya, Mohamed Maarouf, Shile Huang, and Ji-Long Chen. 2018. "Host Immune Response to Influenza A Virus Infection." *Frontiers in Immunology*. https://doi.org/10.3389/fimmu.2018.00320.

Copeland, Anna Maria, Nicole M. Van Deusen, and Connie S. Schmaljohn. 2015. "Rift Valley Fever Virus NSS Gene Expression Correlates with a Defect in Nuclear mRNA Export." *Virology* 486 (December): 88–93.

Dai-Ju, Jenny Q., Ling Li, Lisa A. Johnson, and Rozanne M. Sandri-Goldin. 2006. "ICP27 Interacts with the C-Terminal Domain of RNA Polymerase II and Facilitates Its Recruitment to Herpes Simplex Virus 1 Transcription Sites, Where It Undergoes Proteasomal Degradation during Infection." *Journal of Virology* 80 (7): 3567–81.

Davidson, Andrew D. 2009. "Chapter 2. New Insights into Flavivirus Nonstructural Protein 5." *Advances in Virus Research* 74: 41–101.

Davies, F. Glyn, and Vincent Martin. 2006. "Recognizing Rift Valley Fever." *Veterinaria Italiana* 42 (1): 31–53.

Djebali, Sarah, Carrie A. Davis, Angelika Merkel, Alex Dobin, Timo Lassmann, Ali Mortazavi, Andrea Tanzer, Julien Lagarde, Wei Lin, Felix Schlesinger, Chenghai Xue, Georgi K. Marinov, Jainab Khatun, Brian A. Williams, Chris Zaleski, Joel Rozowsky, Maik Roeder, Felix Kokocinski, Rehab F. Abdelhamid, Tyler Alioto, Igor Antoshechkin, Michael T. Baer, Nadav S. Bar, Philippe Batut, Kimberly Bell, Ian Bell, Sudipto Chakrabortty, Xian Chen, Jacqueline Chrast, Joao Curado, Thomas Derrien, Jorg Drenkow, Erica Dumais, Jacqueline Dumais, Radha Duttagupta, Emilie Falconnet, Meagan Fastuca, Kata Fejes-Toth, Pedro Ferreira, Sylvain Foissac, Melissa J. Fullwood, Hui Gao, David Gonzalez, Assaf Gordon, Harsha Gunawardena, Cedric Howald, Sonali Jha, Rory Johnson, Philipp Kapranov, Brandon King, Colin Kingswood, Oscar J. Luo, Eddie Park, Kimberly Persaud, Jonathan B. Preall, Paolo Ribeca, Brian Risk, Daniel Robyr, Michael Sammeth, Lorian Schaffer, Lei-Hoon See, Atif Shahab, Jorgen Skancke, Ana Maria Suzuki, Hazuki Takahashi, Hagen Tilgner1, Diane Trout, Nathalie Walters, Huaien Wang, John Wrobel, Yanbao Yu, Xiaoan Ruan, Yoshihide Hayashizaki, Jennifer Harrow, Mark Gerstein, Tim Hubbard, Alexandre Reymond, Stylianos E. Antonarakis, Gregory Hannon, Morgan C. Giddings, Yijun Ruan, Barbara Wold, Piero Carninci, Roderic Guigo and Thomas R. Gingeras. 2012. "Landscape of Transcription in Human Cells." *Nature* 489 (7414): 101–8.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. “STAR: Ultrafast Universal RNA-Seq Aligner.” *Bioinformatics* 29 (1): 15–21.

Duttke, Sascha H., Max W. Chang, Sven Heinz, and Christopher Benner. 2019. “Identification and Dynamic Quantification of Regulatory Elements Using Total RNA.” *Genome Research* 29 (11): 1836–46.

Frankish, Adam, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M. Mudge, Cristina Sisu, James Wright, Joel Armstrong, If Barnes, Andrew Berry, Alexandra Bignell, Silvia Carbonell Sala, Jacqueline Chrast, Fiona Cunningham, Tomas Di Domenico, Sarah Donaldson, Ian T. Fiddes, Carlos Garcıa Giron, Jose Manuel Gonzalez, Tiago Grego, Matthew Hardy, Thibaut Hourlier , Toby Hunt, Osagie G. Izuogu, Julien Lagarde, Fergal J. Martin, Laura Martınez, Shamika Mohanan, Paul Muir, Fabio C.P. Navarro, Anne Parker, Baikang Pei, Fernando Pozo, Magali Ruffier, Bianca M. Schmitt, Eloise Stapleton, Marie-Marthe Suner , Irina Sycheva, Barbara Uszczynska-Ratajczak, Jinuri Xu, Andrew Yates, Daniel Zerbino, Yan Zhang, Bronwen Aken, Jyoti S. Choudhary, Mark Gerstein, Roderic Guigo, Tim J.P. Hubbard, Manolis Kellis, Benedict Paten, Alexandre Reymond, Michael L. Tress and Paul Flicek. 2019. “GENCODE Reference Annotation for the Human and Mouse Genomes.” *Nucleic Acids Research* 47 (D1): D766–73.

Friedman, A. D. 2007. “Transcriptional Control of Granulocyte and Monocyte Development.” *Oncogene* 26 (47): 6816–28.

Fros, Jelke J., and Gorben P. Pijlman. 2016. “Alphavirus Infection: Host Cell Shut-Off and Inhibition of Antiviral Responses.” *Viruses* 8 (6). https://doi.org/10.3390/v8060166.

Fuente, Cynthia de la, Chelsea Pinkham, Deemah Dabbagh, Brett Beitzel, Aura Garrison, Gustavo Palacios, Kimberley Alex Hodge, Emanuel F. Petricoin, Connie Schmaljohn, Catherine E. Campbell, Aarthi Narayanan, Kylene Kehn-Hall. 2018. “Phosphoproteomic Analysis Reveals Smad Protein Family Activation Following Rift Valley Fever Virus Infection.” *PloS One* 13 (2): e0191983.

Fujiwara, Tohru, Henriette O’Geen, Sunduz Keles, Kimberly Blahnik, Amelia K. Linnemann, Yoon-A Kang, Kyunghee Choi, Peggy J. Farnham, and Emery H. Bresnick. 2009. “Discovering Hematopoietic Mechanisms through Genome-Wide Analysis of GATA Factor Chromatin Occupancy.” *Molecular Cell* 36 (4): 667–81.

Furey, Terrence S. 2012. “ChIP-Seq and beyond: New and Improved Methodologies to Detect and Characterize Protein-DNA Interactions.” *Nature Reviews. Genetics* 13 (12): 840–52.

Gaillard, Hélène, Tatiana García-Muse, and Andrés Aguilera. 2015. “Replication Stress and Cancer.” *Nature Reviews Cancer*. https://doi.org/10.1038/nrc3916.

Garcia-Moreno, Manuel, Marko Noerenberg, Shuai Ni, Aino I. Järvelin, Esther González-Almela, Caroline E. Lenz, Marcel Bach-Pages, et al. 2019. “System-Wide Profiling of RNA-Binding Proteins Uncovers Key Regulators of Virus Infection.” *Molecular Cell* 74 (1): 196–211.e11.

García-Sastre, Adolfo. 2017. “Ten Strategies of Interferon Evasion by Viruses.” *Cell Host & Microbe* 22 (2): 176–84.

Glyn Davies, F., Vincent Martin, and Food and Agriculture Organization of the United Nations. 2003. *Recognizing Rift Valley Fever*. Food & Agriculture Org.

Gordon, S., G. Akopyan, H. Garban, and B. Bonavida. 2005. “Transcription Factor YY1: Structure, Function, and Therapeutic Implications in Cancer Biology.” *Oncogene* 25 (8): 1125–42.

Greger, I. H., and N. J. Proudfoot. 1998. “Poly(A) Signals Control Both Transcriptional Termination and Initiation between the Tandem GAL10 and GAL7 Genes of Saccharomyces Cerevisiae.” *The EMBO Journal* 17 (16): 4771–79.

Grosso, Ana R., Ana P. Leite, Sílvia Carvalho, Mafalda R. Matos, Filipa B. Martins, Alexandra C. Vítor, Joana M. P. Desterro, Maria Carmo-Fonseca, and Sérgio F. de Almeida. 2015. “Pervasive Transcription Read-through Promotes Aberrant Expression of Oncogenes and RNA Chimeras in Renal Carcinoma.” *eLife* 4 (November). https://doi.org/10.7554/eLife.09214.

Hagberg, A.A., Shult, D.A., Swart, P.J. 2008. “Exploring Network Structure, Dynamics, and Function Using Networkx.” In *Proceedings of the 7th Python in Science Conference*, edited by Varoquaux, G., Vaught, T., Millman, J., 11–15.

Hayhurst, Graham P., Ying-Hue Lee, Gilles Lambert, Jerrold M. Ward, and Frank J. Gonzalez. 2001. “Hepatocyte Nuclear Factor 4α (Nuclear Receptor 2A1) Is Essential for Maintenance of Hepatic Gene Expression and Lipid Homeostasis.” *Molecular and Cellular Biology*. https://doi.org/10.1128/mcb.21.4.1393-1403.2001.

Heinz, Sven, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C. Lin, Peter Laslo, Jason X. Cheng, Cornelis Murre, Harinder Singh, and Christopher K. Glass. 2010. “Simple Combinations of Lineage-Determining Transcription Factors Prime Cis-Regulatory Elements Required for Macrophage and B Cell Identities.” *Molecular Cell*. https://doi.org/10.1016/j.molcel.2010.05.004.

Heinz, Sven, Casey E. Romanoski, Christopher Benner, and Christopher K. Glass. 2015. “The Selection and Function of Cell Type-Specific Enhancers.” *Nature Reviews. Molecular Cell Biology* 16 (3): 144–54.

Heinz, Sven, Lorane Texari, Michael G. B. Hayes, Matthew Urbanowski, Max W. Chang, Ninvita Givarkes, Alexander Rialdi, Kris M. White, Randy A. Albrecht, Lars Pache, Ivan Marazzi, Adolfo Garcı´a-Sastre, Megan L. Shaw, and Christopher Benner. 2018. “Transcription Elongation Can Affect Genome 3D Structure.” *Cell* 174 (6): 1522–36.e22.

Hennig, Thomas, Marco Michalski, Andrzej J. Rutkowski, Lara Djakovic, Adam W. Whisnant, Marie-Sophie Friedl, Bhaskar Anand Jha, Marisa A. P. Baptista, Anne L’Hernault, Florian Erhard, Lars Dolken, and Caroline C. Friedel. 2018. “HSV-1-Induced Disruption of Transcription Termination Resembles a Cellular Stress Response but Selectively Increases Chromatin Accessibility Downstream of Genes.” *PLoS Pathogens* 14 (3): e1006954.

He, Tianqiong, Mingshu Wang, Anchun Cheng, Qiao Yang, Ying Wu, Renyong Jia, Mafeng Liu, Dekang Zhu, Shun Chen, Shaqiu Zhang, Xin-Xin Zhao, Juan Huang, Di Sun, Sai Mao, Xuming Ou, Yin Wang, Zhiwen Xu, Zhengli Chen, Lin Zhu, Qihui Luo, Yunya Liu, Yanling Yu, Ling Zhang, Bin Tian, Leichang Pan, Mujeeb Ur Rehman, and Xiaoyue Chen. 2020. “Host Shutoff Activity of VHS and SOX-like Proteins: Role in Viral Survival and Immune Evasion.” *Virology Journal* 17 (1): 68.

Hobson, David J., Wu Wei, Lars M. Steinmetz, and Jesper Q. Svejstrup. 2012. “RNA Polymerase II Collision Interrupts Convergent Transcription.” *Molecular Cell* 48 (3): 365–74.

Howe, Kevin L., Premanand Achuthan, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M. Ridwan Amode, Irina M. Armean, Andrey G. Azov, Ruth Bennett, Jyothish Bhai, Konstantinos Billis, Sanjay Boddu, Mehrnaz Charkhchi, Carla Cummins, Luca Da Rin Fioretto, Claire Davidson, Kamalkumar Dodiya, Bilal El Houdaigui, Reham Fatima, Astrid Gall, Carlos Garcia Giron, Tiago Grego, Cristina Guijarro-Clarke, Leanne Haggerty, Anmol Hemrom, Thibaut Hourlier, Osagie G Izuogu, Thomas Juettemann, Vinay Kaikala, Mike Kay, Ilias Lavidas, Tuan Le, Diana Lemos, Jose Gonzalez Martinez, Jose Carlos Marugan, Thomas Maurel, Aoife C. McMahon, Shamika Mohanan, Benjamin Moore, Matthieu Muffato, Denye N. Oheh, Dimitrios Paraschas, Anne Parker, Andrew Parton, Irina Prosovetskaia, Manoj P. Sakthivel, Ahamed I. Abdul Salam, Bianca M. Schmitt, Helen Schuilenburg, Dan Sheppard, Emily Steed, Michal Szpak, Marek Szuba, Kieron Taylor, Anja Thormann, Glen Threadgold, Brandon Walts, Andrea Winterbottom, Marc Chakiachvili, Ameya Chaubal, Nishadi De Silva, Bethany Flint, Adam Frankish , Sarah E. Hunt, Garth R. Ilsley, Nick Langridge, Jane E. Loveland , Fergal J. Martin, Jonathan M. Mudge, Joanella Morales, Emily Perry, Magali Ruffier, John Tate, David Thybert, Stephen J. Trevanion, Fiona Cunningham, Andrew D. Yates , Daniel R. Zerbino and Paul Flicek. 2021. “Ensembl 2021.” *Nucleic Acids Research* 49 (D1): D884–91.

Hu, Zihua, and Steven M. Gallo. 2010. “Identification of Interacting Transcription Factors Regulating Tissue Gene Expression in Human.” *BMC Genomics* 11 (January): 49.

Ikegami, Tetsuro. 2012. “Molecular Biology and Genetic Diversity of Rift Valley Fever Virus.” *Antiviral Research* 95 (3): 293–310.

Islam, Zeyaul, Ameena Mohamed Ali, Adviti Naik, Mohamed Eldaw, Julie Decock, and Prasanna R. Kolatkar. 2021. “Transcription Factors: The Fulcrum Between Cell Development and Carcinogenesis.” *Frontiers in Oncology* 11 (June): 681377.

Iwasaki, Akiko, and Padmini S. Pillai. 2014. “Innate Immunity to Influenza Virus Infection.” *Nature Reviews Immunology*. https://doi.org/10.1038/nri3665.

Jackson, Lisa A., Evan J. Anderson, Nadine G. Rouphael, Paul C. Roberts, Mamodikoe Makhene, Rhea N. Coler, Michele P. McCullough, et al. 2020. “An mRNA Vaccine against SARS-CoV-2 - Preliminary Report.” *The New England Journal of Medicine* 383 (20): 1920–31.

Jankowski, Aleksander, Shyam Prabhakar, and Jerzy Tiuryn. 2014. “TACO: A General-Purpose Tool for Predicting Cell-Type-Specific Transcription Factor Dimers.” *BMC Genomics* 15 (March): 208.

Kaczynski, Joanna, Tiffany Cook, and Raul Urrutia. 2003. “Sp1- and Krüppel-like Transcription Factors.” *Genome Biology* 4 (2): 206.

Kainulainen, Markus, Matthias Habjan, Philipp Hubel, Laura Busch, Simone Lau, Jacques Colinge, Giulio Superti-Furga, Andreas Pichlmair, and Friedemann Weber. 2014. “Virulence Factor NSs of Rift Valley Fever Virus Recruits the F-Box Protein FBXO3 To Degrade Subunit p62 of General Transcription Factor TFIIH.” *Journal of Virology*. https://doi.org/10.1128/jvi.02914-13.

Kalveram, B., O. Lihoradova, and T. Ikegami. 2011. “NSs Protein of Rift Valley Fever Virus Promotes Posttranslational Downregulation of the TFIIH Subunit p62.” *Journal of Virology*. https://doi.org/10.1128/jvi.02255-10.

Kawauchi, Junya, Hannah Mischo, Priscilla Braglia, Ana Rondon, and Nick J. Proudfoot. 2008. “Budding Yeast RNA Polymerases I and II Employ Parallel Mechanisms of Transcriptional Termination.” *Genes & Development* 22 (8): 1082–92.

Kent, W. J. 2002. “The Human Genome Browser at UCSC.” *Genome Research*. https://doi.org/10.1101/gr.229102.

Khaperskyy, Denys A., and Craig McCormick. 2015. “Timing Is Everything: Coordinated Control of Host Shutoff by Influenza A Virus NS1 and PA-X Proteins.” *Journal of Virology* 89 (13): 6528–31.

Kim, Minkyu, Nevan J. Krogan, Lidia Vasiljeva, Oliver J. Rando, Eduard Nedea, Jack F. Greenblatt, and Stephen Buratowski. 2004. “The Yeast Rat1 Exonuclease Promotes Transcription Termination by RNA Polymerase II.” *Nature* 432 (7016): 517–22.

Klopfenstein, D. V., Liangsheng Zhang, Brent S. Pedersen, Fidel Ramírez, Alex Warwick Vesztrocy, Aurélien Naldi, Christopher J. Mungall, Jefrey M.Yunes, Olga Botvinnik, Mark Weigel, Will Dampier, Christophe Dessimoz, Patrick Flick, and Haibao Tang. 2018. “GOATOOLS: A Python Library for Gene Ontology Analyses.” *Scientific Reports* 8 (1): 10872.

Köster, Johannes, and Sven Rahmann. 2018. “Snakemake-a Scalable Bioinformatics Workflow Engine.” *Bioinformatics* 34 (20): 3600.

Krishnaraju, Kandasamy, Barbara Hoffman, and Dan A. Liebermann. 2001. “Early Growth Response Gene 1 Stimulates Development of Hematopoietic Progenitor Cells along the Macrophage Lineage at the Expense of the Granulocyte and Erythroid Lineages.” *Blood*. https://doi.org/10.1182/blood.v97.5.1298.

Lambert, Samuel A., Arttu Jolma, Laura F. Campitelli, Pratyush K. Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R. Hughes, and Matthew T. Weirauch. 2018. “The Human Transcription Factors.” *Cell*. https://doi.org/10.1016/j.cell.2018.01.029.

Langmead, Ben, and Steven L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4): 357–59.

Lau, Hwee Hui, Natasha Hui Jin Ng, Larry Sai Weng Loo, Joanita Binte Jasmen, and Adrian Kee Keong Teo. 2018. “The Molecular Functions of Hepatocyte Nuclear Factors – In and beyond the Liver.” *Journal of Hepatology*. https://doi.org/10.1016/j.jhep.2017.11.026.

Lee, J. S., K. M. Galvin, and Y. Shi. 1993. “Evidence for Physical Interaction between the Zinc-Finger Transcription Factors YY1 and Sp1.” *Proceedings of the National Academy of Sciences of the United States of America* 90 (13): 6145–49.

Leiserson, Mark D. M., Hsin-Ta Wu, Fabio Vandin, and Benjamin J. Raphael. 2015. “CoMEt: A Statistical Approach to Identify Combinations of

Mutually Exclusive Alterations in Cancer." *Genome Biology*. https://doi.org/10.1186/s13059-015-0700-7.

Le May, Nicolas, Sandy Dubaele, Luca Proietti De Santis, Agnès Billecocq, Michèle Bouloy, and Jean-Marc Egly. 2004. "TFIIH Transcription Factor, a Target for the Rift Valley Hemorrhagic Fever Virus." *Cell* 116 (4): 541–50.

Le May, Nicolas, Zeyni Mansuroglu, Psylvia Léger, Thibaut Josse, Guillaume Blot, Agnès Billecocq, Ramon Flick, Yves Jacob, Eliette Bonnefoy, and Michèle Bouloy. 2008. "A SAP30 Complex Inhibits IFN-Beta Expression in Rift Valley Fever Virus Infected Cells." *PLoS Pathogens* 4 (1): e13.

Levitsky, Victor, Elena Zemlyanskaya, Dmitry Oshchepkov, Olga Podkolodnaya, Elena Ignatieva, Ivo Grosse, Victoria Mironova, and Tatyana Merkulova. 2019. "A Single ChIP-Seq Dataset Is Sufficient for Comprehensive Analysis of Motifs Co-Occurrence with MCOT Package." *Nucleic Acids Research* 47 (21): e139.

Licatalosi, Donny D., Gabrielle Geiger, Michelle Minet, Stephanie Schroeder, Kate Cilli, J. Bryan McNeil, and David L. Bentley. 2002. "Functional Interaction of Yeast Pre-mRNA 3′ End Processing Factors with RNA Polymerase II." *Molecular Cell*. https://doi.org/10.1016/s1097-2765(02)00518-x.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25 (16): 2078–79.

Li, J., G. Ning, and S. A. Duncan. 2000. "Mammalian Hepatocyte Differentiation Requires the Transcription Factor HNF-4alpha." *Genes & Development* 14 (4): 464–74.

Linnemann, Amelia K., Henriette O'Geen, Sunduz Keles, Peggy J. Farnham, and Emery H. Bresnick. 2011. "Genetic Framework for GATA Factor Function in Vascular Biology." *Proceedings of the National Academy of Sciences of the United States of America* 108 (33): 13641–46.

Li, Wenbo, Dimple Notani, and Michael G. Rosenfeld. 2016. "Enhancers as Non-Coding RNA Transcription Units: Recent Insights and Future Perspectives." *Nature Reviews. Genetics* 17 (4): 207–23.

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

Ly, Hoai J., and Tetsuro Ikegami. 2016. “Rift Valley Fever Virus NSs Protein Functions and the Similarity to Other Bunyavirus NSs Proteins.” *Virology Journal* 13 (July): 118.

Lyles, D. S. 2000. “Cytopathogenesis and Inhibition of Host Gene Expression by RNA Viruses.” *Microbiology and Molecular Biology Reviews: MMBR* 64 (4): 709–24.

Mantovani, Roberto. 1999. “The Molecular Biology of the CCAAT-Binding Factor NF-Y.” *Gene*. https://doi.org/10.1016/s0378-1119(99)00368-6.

Martin, Marcel. 2011. “Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads.” *EMBnet.journal*. https://doi.org/10.14806/ej.17.1.200.

Massagué, Joan, and Ye-Guang Chen. 2000. “Controlling TGF-β Signaling.” *Genes & Development* 14 (6): 627–44.

Maston, Glenn A., Sara K. Evans, and Michael R. Green. 2006. “Transcriptional Regulatory Elements in the Human Genome.” *Annual Review of Genomics and Human Genetics*. https://doi.org/10.1146/annurev.genom.7.080505.115623.

Matsui, M., M. Tokuhara, Y. Konuma, N. Nomura, and R. Ishizaki. 1990. “Isolation of Human Fos-Related Genes and Their Expression during Monocyte-Macrophage Differentiation.” *Oncogene* 5 (3): 249–55.

McKinney, Wes. 2010. “Data Structures for Statistical Computing in Python.” *Proceedings of the 9th Python in Science Conference*. https://doi.org/10.25080/majora-92bf1922-00a.

Meckbach, Cornelia, Rebecca Tacke, Xu Hua, Stephan Waack, Edgar Wingender, and Mehmet Gültas. 2015. “PC-TraFF: Identification of Potentially Collaborating Transcription Factors Using Pointwise Mutual Information.” *BMC Bioinformatics* 16 (December): 400.

Melnick, Marko, Patrick Gonzales, Joseph Cabral, Mary A. Allen, Robin D. Dowell, and Christopher D. Link. 2019. “Heat Shock in C. Elegans Induces Downstream of Gene Transcription and Accumulation of Double-Stranded RNA.” *PLOS ONE*. https://doi.org/10.1371/journal.pone.0206715.

Melton, Collin, Jason A. Reuter, Damek V. Spacek, and Michael Snyder. 2015. “Recurrent Somatic Mutations in Regulatory Regions of Human Cancer Genomes.” *Nature Genetics*. https://doi.org/10.1038/ng.3332.

Moore, Jill E., Michael J. Purcaro, Henry E. Pratt, Charles B. Epstein, Noam Shoresh, Jessika Adrian, Trupti Kawli, Carrie A. Davis, Alexander Dobin, Rajinder Kaul, Jessica Halow, Eric L. Van Nostrand, Peter Freese, David U. Gorkin, Yin Shen, Yupeng He, Mark Mackiewicz, Florencia Pauli-Behn, Brian A. Williams, Ali Mortazavi, Cheryl A. Keller, Xiao-Ou Zhang, Shaimae I. Elhajjajy, Jack Huey, Diane E. Dickel, Valentina Snetkova, Xintao Wei, Xiaofeng Wang, Juan Carlos Rivera-Mulia, Joel Rozowsky, Jing Zhang, Surya B. Chhetri, Jialing Zhang, Alec Victorsen, Kevin P. White, Axel Visel, Gene W. Yeo, Christopher B. Burge, Eric Lécuyer, David M. Gilbert, Job Dekker, John Rinn, Eric M. Mendenhall, Joseph R. Ecker, Manolis Kellis, Robert J. Klein, William S. Noble, Anshul Kundaje, Roderic Guigó, Peggy J. Farnham, J. Michael Cherry, Richard M. Myers, Bing Ren, Brenton R. Graveley, Mark B. Gerstein, Len A. Pennacchio, Michael P. Snyder, Bradley E. Bernstein, Barbara Wold, Ross C. Hardison, Thomas R. Gingeras, John A. Stamatoyannopoulos, and Zhiping Weng . 2020. “Expanded Encyclopaedias of DNA Elements in the Human and Mouse Genomes.” *Nature* 583 (7818): 699–710.

Muniz, Lisa, Maharshi Krishna Deb, Marion Aguirrebengoa, Sandra Lazorthes, Didier Trouche, and Estelle Nicolas. 2017. “Control of Gene Expression in Senescence through Transcriptional Read-Through of Convergent Protein-Coding Genes.” *Cell Reports* 21 (9): 2433–46.

Musso, Didier, and Duane J. Gubler. 2016. “Zika Virus.” *Clinical Microbiology Reviews* 29 (3): 487–524.

Myšičková, Alena, and Martin Vingron. 2012. “Detection of Interacting Transcription Factors in Human Tissues Using Predicted DNA Binding Affinity.” *BMC Genomics* 13 Suppl 1 (January): S2.

Nagaki, Masahito, and Hisataka Moriwaki. 2008. “Transcription Factor HNF and Hepatocyte Differentiation.” *Hepatology Research*. https://doi.org/10.1111/j.1872-034x.2008.00367.x.

Nemeroff, M. E., S. M. Barabino, Y. Li, W. Keller, and R. M. Krug. 1998. “Influenza Virus NS1 Protein Interacts with the Cellular 30 kDa Subunit of CPSF and Inhibits 3’end Formation of Cellular Pre-mRNAs.” *Molecular Cell* 1 (7): 991–1000.

Neph, Shane, M. Scott Kuehn, Alex P. Reynolds, Eric Haugen, Robert E. Thurman, Audra K. Johnson, Eric Rynes, et al. 2012. “BEDOPS: High-Performance Genomic Feature Operations.” *Bioinformatics* 28 (14): 1919–20.

Nielsen, Soren, Yulia Yuzenkova, and Nikolay Zenkin. 2013. “Mechanism of Eukaryotic RNA Polymerase III Transcription Termination.” *Science* 340 (6140): 1577–80.

Nogales, Aitor, Luis Martinez-Sobrido, David J. Topham, and Marta L. DeDiego. 2018. “Modulation of Innate Immune Responses by the Influenza A NS1 and PA-X Proteins.” *Viruses* 10 (12). https://doi.org/10.3390/v10120708.

Nojima, Takayuki, Tomás Gomes, Maria Carmo-Fonseca, and Nicholas J. Proudfoot. 2016. “Mammalian NET-Seq Analysis Defines Nascent RNA Profiles and Associated RNA Processing Genome-Wide.” *Nature Protocols* 11 (3): 413–28.

Nojima, Takayuki, Tomás Gomes, Ana Rita Fialho Grosso, Hiroshi Kimura, Michael J. Dye, Somdutta Dhir, Maria Carmo-Fonseca, and Nicholas J. Proudfoot. 2015. “Mammalian NET-Seq Reveals Genome-Wide Nascent Transcription Coupled to RNA Processing.” *Cell* 161 (3): 526–40.

Oliphant, Travis E. 2006. *A Guide to NumPy*.

Partridge, E. Christopher, Surya B. Chhetri, Jeremy W. Prokop, Ryne C. Ramaker, Camden S. Jansen, Say-Tar Goh, Mark Mackiewicz, Kimberly M. Newberry, Laurel A. Brandsmeier, Sarah K. Meadows, C. Luke Messer, Andrew A. Hardigan, Candice J. Coppola, Emma C. Dean, Shan Jiang, Daniel Savic, Ali Mortazavi, Barbara J. Wold, Richard M. Myers, and Eric M. Mendenhall. 2020. “Occupancy Maps of 208 Chromatin-Associated Proteins in One Human Cell Type.” *Nature* 583 (7818): 720–28.

Petrova, Velislava, Paul Kristiansen, Gunnstein Norheim, and Solomon A. Yimer. 2020. “Rift Valley Fever: Diagnostic Challenges and Investment Needs for Vaccine Development.” *BMJ Global Health* 5 (8). https://doi.org/10.1136/bmjgh-2020-002694.

Prescott, E. M., and N. J. Proudfoot. 2002. “Transcriptional Collision between Convergent Genes in Budding Yeast.” *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.132270899.

Qi, Dong-Lai, Takahito Ohhira, Chikako Fujisaki, Toshiaki Inoue, Tsutomu Ohta, Mitsuhiko Osaki, Eriko Ohshiro, Tomomi Seko, Shinsuke Aoki, Mitsuo Oshimura, and Hiroyuki Kugoh. 2011. “Identification of PITX1 as a TERT Suppressor Gene Located on Human Chromosome 5.” *Molecular and Cellular Biology* 31 (8): 1624–36.

Quach, Hélène, Maxime Rotival, Julien Pothlichet, Yong-Hwee Eddie Loh, Michael Dannemann, Nora Zidane, Guillaume Laval, Etienne Patin, Christine Harmant,Marie Lopez, Matthieu Deschamps, Nadia Naffakh, Darragh Duffy, Anja Coen, Geert Leroux-Roels, Frederic Clement, Anne Boland, Jean-Francois Deleuze, Janet Kelso, Matthew L. Albert, and Lluis Quintana-Murci. 2016. “Genetic

Adaptation and Neandertal Admixture Shaped the Immune System of Human Populations." *Cell* 167 (3): 643–56.e17.

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

Rai, Kul Raj, Prasha Shrestha, Bincai Yang, Yuhai Chen, Shasha Liu, Mohamed Maarouf, and Ji-Long Chen. 2021. "Acute Infection of Viral Pathogens and Their Innate Immune Escape." *Frontiers in Microbiology* 12 (June): 672026.

Richard, P., and J. L. Manley. 2009. "Transcription Termination by Nuclear RNA Polymerases." *Genes & Development*. https://doi.org/10.1101/gad.1792809.

Roth, Samuel J., Sven Heinz, and Christopher Benner. 2020. "ARTDeco: Automatic Readthrough Transcription Detection." *BMC Bioinformatics* 21 (1): 214.

Rutkowski, Andrzej J., Florian Erhard, Anne L'Hernault, Thomas Bonfert, Markus Schilhabel, Colin Crump, Philip Rosenstiel, Stacey Efstathiou, Ralf Zimmer, Caroline C. Friedel, and Lars Dolken. 2015. "Widespread Disruption of Host Transcription Termination in HSV-1 Infection." *Nature Communications* 6 (May): 7126.

Sahin, Ugur, Alexander Muik, Evelyna Derhovanessian, Isabel Vogler, Lena M. Kranz, Mathias Vormehr, Alina Baum, Kristen Pascal, Jasmin Quandt, Daniel Maurus, Sebastian Brachtendorf, Verena Lörks, Julian Sikorski1 , Rolf Hilker1 , Dirk Becker1 , Ann-Kathrin Eller1 , Jan Grützner, Carsten Boesler, Corinna Rosenbaum, Marie-Cristine Kühnle, Ulrich Luxemburger, Alexandra Kemmer-Brück, David Langer, Martin Bexon, Stefanie Bolte, Katalin Karikó, Tania Palanche, Boris Fischer, Armin Schultz, Pei-Yong Shi, Camila Fontes-Garfias, John L. Perez, Kena A. Swanson, Jakob Loschko, Ingrid L. Scully, Mark Cutler, Warren Kalina, Christos A. Kyratsous , David Cooper, Philip R. Dormitzer , Kathrin U. Jansen, and Özlem Türeci. 2020. "COVID-19 Vaccine BNT162b1 Elicits Human Antibody and T1 T Cell Responses." *Nature* 586 (7830): 594–99.

Samadani, U., and R. H. Costa. 1996. "The Transcriptional Activator Hepatocyte Nuclear Factor 6 Regulates Liver Gene Expression." *Molecular and Cellular Biology* 16 (11): 6273–84.

Scruggs, Benjamin S., Daniel A. Gilchrist, Sergei Nechaev, Ginger W. Muse, Adam Burkholder, David C. Fargo, and Karen Adelman. 2015. "Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin." *Molecular Cell*. https://doi.org/10.1016/j.molcel.2015.04.006.

Seto, E., B. Lewis, and T. Shenk. 1993. "Interaction between Transcription Factors Sp1 and YY1." *Nature* 365 (6445): 462–64.

Shah, Priya S., Nichole Link, Gwendolyn M. Jang, Phillip P. Sharp, Tongtong Zhu, Danielle L. Swaney, Jeffrey R. Johnson, John Von Dollen, Holly R. Ramage, Laura Satkamp, Billy Newton, Ruth Huttenhain, Marine J. Petit, Tierney Baum, Amanda Everitt, Orly Laufman, Michel Tassetto, Michael Shales, Erica Stevenson, Gabriel N. Iglesias, Leila Shokat, Shashank Tripathi, Vinod Balasubramaniam, Laurence G. Webb, Sebastian Aguirre, A. Jeremy Willsey, Adolfo Garcia-Sastre, Katherine S. Pollard, Sara Cherry, Andrea V. Gamarnik, Ivan Marazzi, Jack Taunton, Ana Fernandez-Sesma, Hugo J. Bellen, Raul Andino, and Nevan J. Krogan. 2018. “Comparative Flavivirus-Host Protein Interaction Mapping Reveals Mechanisms of Dengue and Zika Virus Pathogenesis.” *Cell* 175 (7): 1931–45.e18.

Shearwin, K., B. Callen, and J. Egan. 2005. “Transcriptional Interference – a Crash Course.” *Trends in Genetics*. https://doi.org/10.1016/j.tig.2005.04.009.

Smith, Justine R., Shawn Todd, Liam M. Ashander, Theodosia Charitou, Yuefang Ma, Steven Yeh, Ian Crozier, Michael Z. Michael, Binoy Appukuttan, Keryn A. Williams, David J. Lynn, and Glenn A. Marsh. 2017. “Retinal Pigment Epithelial Cells Are a Potential Reservoir for Ebola Virus in the Human Eye.” *Translational Vision Science & Technology* 6 (4): 12.

Speranza, Emily, and John H. Connor. 2017. “Host Transcriptional Response to Ebola Virus Infection.” *Vaccines* 5 (3). https://doi.org/10.3390/vaccines5030030.

Steel, John, Anice C. Lowen, Lindomar Pena, Matthew Angel, Alicia Solórzano, Randy Albrecht, Daniel R. Perez, Adolfo García-Sastre, and Peter Palese. 2009. “Live Attenuated Influenza Viruses Containing NS1 Truncations as Vaccine Candidates against H5N1 Highly Pathogenic Avian Influenza.” *Journal of Virology* 83 (4): 1742–53.

Stender, Joshua D., Fabio Stossi, Cory C. Funk, Tze Howe Charn, Daniel H. Barnett, and Benita S. Katzenellenbogen. 2011. “The Estrogen-Regulated Transcription Factor PITX1 Coordinates Gene-Specific Regulation by Estrogen Receptor-Alpha in Breast Cancer Cells.” *Molecular Endocrinology* 25 (10): 1699.

Szklarczyk, Damian, Annika L. Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T. Doncheva, John H. Morris, Peer Bork, Lars J. Jensen, and Christian von Mering. 2019. “STRING v11: Protein-Protein Association Networks with Increased Coverage, Supporting Functional Discovery in Genome-Wide Experimental Datasets.” *Nucleic Acids Research* 47 (D1): D607–13.

Terasaki, Kaori, Sydney I. Ramirez, and Shinji Makino. 2016. “Mechanistic Insight into the Host Transcription Inhibition Function of Rift Valley Fever Virus NSs and Its Importance in Virulence." *PLoS Neglected Tropical Diseases* 10 (10): e0005047.

Thomas, Paul D., Valerie Wood, Christopher J. Mungall, Suzanna E. Lewis, Judith A. Blake, and on behalf of the Gene Ontology Consortium. 2012. "On the Use of Gene Ontology Annotations to Assess Functional Similarity among Orthologs and Paralogs: A Short Report." *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1002386.

Tran, Thai Q., and Chrissa Kioussi. 2021. "Pitx Genes in Development and Disease." *Cellular and Molecular Life Sciences: CMLS* 78 (11): 4921–38.

Valledor, Annabel F., Francesc E. Borràs, Martin Cullell-Young, and Antonio Celada. 1998. "Transcription Factors That Regulate Monocyte/macrophage Differentiation." *Journal of Leukocyte Biology*. https://doi.org/10.1002/jlb.63.4.405.

Vandenbon, Alexis, Yutaro Kumagai, Shizuo Akira, and Daron M. Standley. 2012. "A Novel Unbiased Measure for Motif Co-Occurrence Predicts Combinatorial Regulation of Transcription." *BMC Genomics* 13 Suppl 7 (December): S11.

Verheul, Thijs C. J., Levi van Hijfte, Elena Perenthaler, and Tahsin Stefan Barakat. 2020. "The Why of YY1: Mechanisms of Transcriptional Regulation by Yin Yang 1." *Frontiers in Cell and Developmental Biology* 8 (September): 592164.

Vilborg, Anna, Maria C. Passarelli, Therese A. Yario, Kazimierz T. Tycowski, and Joan A. Steitz. 2015. "Widespread Inducible Transcription Downstream of Human Genes." *Molecular Cell* 59 (3): 449–61.

Vilborg, Anna, Niv Sabath, Yuval Wiesel, Jenny Nathans, Flonia Levy-Adam, Therese A. Yario, Joan A. Steitz, and Reut Shalgi. 2017. "Comparative Analysis Reveals Genomic Features of Stress-Induced Transcriptional Readthrough." *Proceedings of the National Academy of Sciences of the United States of America* 114 (40): E8362–71.

Wang, Liguo, Shengqin Wang, and Wei Li. 2012. "RSeQC: Quality Control of RNA-Seq Experiments." *Bioinformatics* 28 (16): 2184–85.

Wang, Xiuye, Thomas Hennig, Adam W. Whisnant, Florian Erhard, Bhupesh K. Prusty, Caroline C. Friedel, Elmira Forouzmand, William Hu, Luke Erber, Yue Chen, Rozanne M. Sandri-Goldin, Lars Dölken, and Yongsheng Shi. 2020. "Herpes Simplex Virus Blocks Host Transcription Termination via the Bimodal Activities of ICP27." *Nature Communications* 11 (1): 293.

Waskom, Michael. 2021. "Seaborn: Statistical Data Visualization." *Journal of Open Source Software*. https://doi.org/10.21105/joss.03021.

West, Steven, Natalia Gromak, and Nick J. Proudfoot. 2004. “Human 5’ --> 3' Exonuclease Xrn2 Promotes Transcription Termination at Co-Transcriptional Cleavage Sites.” *Nature* 432 (7016): 522–25.

Whitington, Tom, Martin C. Frith, James Johnson, and Timothy L. Bailey. 2011. “Inferring Transcription Factor Complexes from ChIP-Seq Data.” *Nucleic Acids Research* 39 (15): e98.

Wiesel, Yuval, Niv Sabath, and Reut Shalgi. 2018. “DoGFinder: A Software for the Discovery and Quantification of Readthrough Transcripts from RNA-Seq.” *BMC Genomics* 19 (1): 597.

Zhang, Huimin, Frank Rigo, and Harold G. Martinson. 2015. “Poly(A) Signal-Dependent Transcription Termination Occurs through a Conformational Change Mechanism That Does Not Require Cleavage at the Poly(A) Site.” *Molecular Cell* 59 (3): 437–48.

Zhang, Yiguo, and Yuancai Xiang. 2016. “Molecular and Cellular Basis for the Unique Functioning of Nrf1, an Indispensable Transcription Factor for Maintaining Cell Homoeostasis and Organ Integrity.” *Biochemical Journal*. https://doi.org/10.1042/bj20151182.

Zhao, Nan, Vittorio Sebastiano, Natasha Moshkina, Nacho Mena, Judd Hultquist, David Jimenez-Morales, Yixuan Ma, Alex Rialdi, Randy Albrecht, Romain Fenouil, Maria Teresa Sánchez-Aparicio, Juan Ayllon, Sweta Ravisankar, Bahareh Haddad, Jessica Sook Yuin Ho, Diana Low, Jian Jin, Vyacheslav Yurchenko, Rab K. Prinjha, Alexander Tarakhovsky, Massimo Squatrito, Dalila Pinto, Kimaada Allette, Minji Byun, Melissa Laird Smith, Robert Sebra, Ernesto Guccione, Terrence Tumpey, Nevan Krogan, Benjamin Greenbaum, Harm van Bakel, Adolfo García-Sastre, and Ivan Marazzi. 2018. “Influenza Virus Infection Causes Global RNAPII Termination Defects.” *Nature Structural & Molecular Biology* 25 (9): 885–93.

Zhao, Zikai, Mengying Tao, Wei Han, Zijing Fan, Muhammad Imran, Shengbo Cao, and Jing Ye. 2021. “Nuclear Localization of Zika Virus NS5 Contributes to Suppression of Type I Interferon Production and Response.” *The Journal of General Virology* 102 (3). https://doi.org/10.1099/jgv.0.001376.

Zhou, Yingyao, Bin Zhou, Lars Pache, Max Chang, Alireza Hadj Khodabakhshi, Olga Tanaseichuk, Christopher Benner, and Sumit K. Chanda. 2019. “Metascape Provides a Biologist-Oriented Resource for the Analysis of Systems-Level Datasets.” *Nature Communications*. https://doi.org/10.1038/s41467-019-09234-6.

Zhu, Zhou, Jay Shendure, and George M. Church. 2005. “Discovering Functional Transcription-Factor Combinations in the Human Cell Cycle.” *Genome Research* 15 (6): 848–55.