Models of Dynamic RNA Regulation in Mammalian Cells

by

Michal Rabani

B.Sc., Hebrew University of Jerusalem (2005)
M.Sc., Weizmann Institute of Science (2008)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2013

© Massachusetts Institute of Technology 2013. All rights reserved.

Author ........................................
Department of Electrical Engineering and Computer Science
Aug 8, 2013

Certified by ..........................................................,
Aviv Regev
Associate Professor of Biology
Thesis Supervisor

Accepted by ...............................................
Professor Leslie A. Kolodziejski
Chairman, Department Committee on Graduate Theses
Models of Dynamic RNA Regulation in Mammalian Cells

by

Michal Rabani

Submitted to the Department of Electrical Engineering and Computer Science on Aug 8, 2013, in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Abstract

Complex molecular circuits, consisting of multiple intertwined feedback loops and non-linear interactions, are a hallmark of every living cell, and a model of a dynamic complex network. Here, I systematically study the dynamic changes in the cellular circuits that control RNA levels in mammalian cells, focusing on the model response of immune dendritic cells to pathogens, through an integration of comprehensive computational models and innovative empirical approaches. I establish a computational framework to follow the dynamics of processes for RNA birth (production, by transcription), maturation (processing), and death (degradation), and their integration in the dynamic RNA life cycle. I study the kinetics of a gene’s RNA population with a model of its production and degradation, and generalize the system as an ensemble of genes. I further model genes as composite particles and study the regulation and kinetics of altering their internal structure. To allow robust statistical inference from these models, I develop innovative laboratory assays and collect extensive experimental data on the system. I directly measure RNA production rates by coupling short RNA metabolic labeling with advanced RNA quantification. I leverage recent improvements in RNA quantification by next-generation sequencing technology, to significantly increase the resolution of metabolic labeling in both time and gene-structure. Finally, I collect perturbation data, by monitoring RNA levels when specific elements of the network are disabled. In this way, I formulated several general principles of RNA regulation and its temporal evolution in mammalian cells. I find that temporal changes in production provide a dominant input in computing RNA levels by the cell over time. Yet, dynamic degradation changes contribute to shaping expression peaks, and dynamic processing changes allow a fast accumulation of mature transcripts. Static degradation and processing rates vary between genes and between individual splicing junctions, consistently with their function and expression dynamics. This study is broadly applicable to many normal as well as diseases misregulated cellular networks, and is also relevant for a more general analysis of complex systems dynamics.
Acknowledgments

None of this work would been possible without the help and support of many people. First and foremost, I would like to express my deepest gratitude to my research advisor Aviv Regev. Thank you Aviv, for supporting me in every step I took through my graduate school years. I was incredibly fortunate to have an advisor that put her students' satisfaction, professionally and personally, at the first priority. You are a true role model as a woman and a scientist. Two of my collaborators, Ido Amit and Nir Friedman, deserve many many thanks for advising me without the ‘official’ hat of a research advisor. Thank you Ido, for mentoring me through the this work. It is only thanks to you that I can add to my skills set aslo experimental work. Your sharp thinking and deep understanding of the biological system truly inspired me. Thank you Nir, for welcoming me into your lab. Your ability to ask the right questions and move forward with elegant solutions is an inspiration, which I hope to follow up with.

I would also like to acknowledge everyone in the Regev lab for both scientific and life advice, and most importantly to Tal, my officemate and my very close friend. Thank you Tal, I could not survive through these years without your day to day support and encouragement. Many thanks to all the members of the Friedman lab and the Amit lab that were always ready to offer their help and advice.

Finally, I would like to thank my family for their loving support through this journey. To my parents that encouraged, supported and helped from across the ocean. To my grandmother that always accepted me with so much love. To my brother and sister that are always so proud of me. And, most importantly, to my wonderful husband, that steppd into my life in the middle of this journey, and I can never imagine it without him. Thank you Yotam, none of this would be possible without you by my side (near or far) and your unlimited support. And finally, to my daughter and the newest member of my family. Thank you Ayalula for being so pure and cute. I have learned so much from you already.
# Contents

1 Introduction ............................................. 13
   1.1 Motivation ............................................. 13
   1.2 Regulation of cellular RNA levels ....................... 14
      1.2.1 The central dogma of molecular biology ............. 14
      1.2.2 The dynamic RNA life cycle ......................... 15
      1.2.3 RNA production through transcription ................. 16
      1.2.4 RNA maturation by processing ....................... 18
      1.2.5 RNA degradation .................................... 19
   1.3 Experimental methods for studying dynamic RNA regulation ............. 20
      1.3.1 Methods for assessing reaction rates within the RNA life cycle, and their limitations ......................... 20
      1.3.2 RNA quantification assays .......................... 22
      1.3.3 Targeted inactivation of genes in vivo ............... 23
   1.4 Mathematical formalisms in dynamic RNA regulation .................. 24
      1.4.1 Models of regulatory interactions that control RNA levels .... 24
      1.4.2 Dynamic time-series models of cellular responses .......... 25
      1.4.3 Inferring RNA abundance from sequencing reads .......... 27
   1.5 Mammalian immune dendritic cells ....................... 27

2 A model of RNA production and degradation dynamics ................. 29
   2.1 Introduction ............................................. 29
   2.2 A first order dynamic model of RNA production and degradation .... 30
      2.2.1 A model of dynamic RNA regulation .................... 30
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.2.2 Measurement dynamics</td>
<td>31</td>
</tr>
<tr>
<td>2.2.3 Estimating constant production rates</td>
<td>32</td>
</tr>
<tr>
<td>2.3 Inferring dynamic rates from measurements</td>
<td>33</td>
</tr>
<tr>
<td>2.3.1 Noise model</td>
<td>33</td>
</tr>
<tr>
<td>2.3.2 Maximal likelihood optimization by gradient descent</td>
<td>34</td>
</tr>
<tr>
<td>2.3.3 Initializing the optimum search</td>
<td>36</td>
</tr>
<tr>
<td>2.3.4 A Gaussian mixture prior on the parameter space</td>
<td>36</td>
</tr>
<tr>
<td>2.4 Hypothesis testing</td>
<td>37</td>
</tr>
<tr>
<td>2.4.1 Likelihood ratio test to select between competing hypotheses for degradation</td>
<td>37</td>
</tr>
<tr>
<td>2.4.2 Goodness of fit test</td>
<td>38</td>
</tr>
<tr>
<td>2.5 Summary</td>
<td>38</td>
</tr>
</tbody>
</table>

3 Principles of RNA production and degradation dynamics in mammalian cells

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Introduction</td>
<td>41</td>
</tr>
<tr>
<td>3.2 Direct estimation of RNA production by short metabolic labeling of RNA with 4sU</td>
<td>42</td>
</tr>
<tr>
<td>3.3 RNA transcription and degradation dynamics of a signature gene set</td>
<td>45</td>
</tr>
<tr>
<td>3.3.1 A high-resolution temporal response of signature genes</td>
<td>45</td>
</tr>
<tr>
<td>3.3.2 Dynamic changes in RNA expression levels usually lag behind transcription rate changes by 15-30 min</td>
<td>45</td>
</tr>
<tr>
<td>3.3.3 Changes in transcription rates shape temporal RNA profiles of most genes</td>
<td>47</td>
</tr>
<tr>
<td>3.3.4 Temporal changes in degradation are important for shaping 'peaked' responses</td>
<td>49</td>
</tr>
<tr>
<td>3.4 Genome-wide RNA transcription and degradation dynamics</td>
<td>50</td>
</tr>
<tr>
<td>3.4.1 The genome-wide response to LPS stimulation by high throughput RNA sequencing</td>
<td>50</td>
</tr>
</tbody>
</table>
3.4.2 Consistent genome-wide and small-scale measurements of newly transcribed RNA

3.4.3 Constant degradation rates are a genome-wide phenomenon and contribute to shaping temporal RNA levels

3.5 Summary

3.6 Methods

4 Models of RNA processing dynamics

4.1 Introduction

4.2 A second order model of the dynamic RNA life-cycle: production, processing and degradation

4.2.1 A single step junction processing model

4.2.2 Multi-step transcript processing models

4.3 Inferring precursor and mature RNA abundance from RNA sequencing

4.3.1 A generalized RPKM estimates an average precursor abundance

4.3.2 A binomial model estimates junction-specific precursor and mature RNA levels

4.3.3 Generalizing the binomial model to predict levels of a transcript's splicing intermediates

4.3.4 A case studies of two typical splice junctions

4.3.5 Maximum likelihood optimization of the binomial model

4.4 Inferring dynamic RNA processing rates

4.4.1 Direct inference of constant processing rates from RNA-4sU data

4.4.2 Maximum likelihood optimization by gradient descent

4.4.3 Maximum likelihood optimization from sequencing counts by derivative free methods

4.5 Identifying transcript editing positions from RNA-Total and RNA-4sU sequencing

4.6 Summary
5 The genome-wide dynamic regulation of RNA splicing and editing in mammalian cells

5.1 Introduction ........................................ 85

5.2 RNA maturation at whole transcript resolution .......... 86

5.2.1 Estimating a transcript's average maturation rate with the single-step processing model .......... 88

5.2.2 Differences in transcript maturation rates between genes contribute to regulation of RNA levels .......... 88

5.3 RNA splicing dynamics at junction resolution .......... 89

5.3.1 Measuring junction-specific precursor and mature transcript abundance from RNA sequencing .......... 90

5.3.2 The junction-specific temporal response of mouse dendritic cells to LPS stimulus .......... 91

5.3.3 Dynamic changes in a junction's mature RNA levels are preceded by highly correlated changes in precursor and newly produced RNA levels .......... 94

5.3.4 Differences in processing rates between junctions contribute to shaping dynamic RNA levels .......... 97

5.3.5 A junction's exon and intron sequence composition is associated with differences in processing rate .......... 101

5.3.6 Spatial coordination of junction processing within a transcript .......... 102

5.4 RNA editing sites in mouse dendritic cells .......... 104

5.4.1 Few transcripts are edited in mouse DCs during the LPS response .......... 104

5.4.2 Predicted editing events are located at non-coding introns and 3'UTRs, suggesting that they contribute to post-transcriptional regulation of expression .......... 105

5.5 Summary .......... 106

5.6 Methods .......... 108

6 Molecular mechanisms of RNA regulation .......... 111
6.1 Introduction .................................................. 111
6.2 A molecular model for the regulation of RNA degradation ........ 112
  6.2.1 The kinetics of an RNA degradation protein regulator ...... 112
  6.2.2 A model of RNA dynamics in a perturbational study .......... 115
  6.2.3 Likelihood ratio test to select between competing hypotheses for a degradation regulator ........................................ 116
  6.2.4 Inferring dynamic rates from measurements .................. 116
6.3 The RNA binding protein Zfp36 regulates RNA stability in mouse DCs 117
  6.3.1 The RNA binding protein Zfp36 ................................ 117
  6.3.2 Knock out of Zfp36 affects the degradation rates of candidate target transcripts .................................................. 119
6.4 A knock-down screen suggests candidate RNA splicing regulators 121
  6.4.1 A knock-down screen of 24 RNA binding proteins .......... 123
  6.4.2 Perturbation data uncovers regulatory interactions that determine RNA processing rates ........................................ 123
6.5 Summary ......................................................... 124
6.6 Methods .......................................................... 126

7 Conclusion ......................................................... 129
  7.1 Contributions .................................................. 129
  7.2 Future directions .............................................. 131
  7.3 Concluding remarks ............................................. 133
Chapter 1

Introduction

1.1 Motivation

Complex systems, consisting of a large number of interacting parts, exist in many computational and real-life domains. Many complex networks change dynamically over time in their composition (number or frequency of system elements) and their connectivity (interactions between elements). As these are generally too complex to allow informal reasoning, mathematical and computational formalisms are used to create models that describe and analyze such systems: clarify intuitions, predict system behavior, promote their theoretical understanding, and develop methods to manage and manipulate large scale data, as appropriate.

The molecular circuits that control RNA levels within living cells and their dynamic evolution, provide a compelling model of such a phenomena. In every living cell, complex molecular circuits process signals from the environment into an appropriate response (section 1.2.1). The regulation of cellular RNA levels, both in time and amplitude, is a main component of these circuits, as cells compute an accurate quantitative level of each type of RNA at each time point (section 1.2.2). To control RNA levels, cells tightly regulate their 'birth' (production by transcription, section 1.2.3), maturation (processing, section 1.2.4) and 'death' (degradation, section 1.2.5). Each of these steps is controlled by complex circuits, consisting of multiple intertwined feed-
back loops that connect DNA, RNA and proteins in non-linear interactions, which together encompass the dynamic life cycle of RNA.

Although the regulation of cellular RNA levels has been the focus of substantial research, it remains poorly understood, in part due to its complex control by non-linear interactions. While many studies monitor cellular RNA levels, monitoring the individual contribution of each separate regulatory process within the RNA life cycle has been technically challenging (section 1.3.1). Moreover, although cellular responses to environmental changes unfold over time, in the absence of precise RNA quantification methods (section 1.3.2), most studies rely only on steady-state measurements rather than on time-series experiments. There is also a big gap in existing theoretical and mathematical formalisms to study how cells implement dynamic transitions at different stages of the RNA life cycle (section 1.4.1), and how they compute precise temporal RNA levels (section 1.4.2).

Therefore, we still lack a quantitative genome-scale understanding of the individual components of RNA regulation, and of their temporal integration to shape cellular regulatory programs. Likewise, the molecular mechanisms that implement and control these processes are still largely unknown. Gaining such understanding is crucial for deciphering the principals that govern cellular regulatory programs and their mis-regulation in many diseases. Moreover, theoretical tools to study how complex systems implement dynamic transitions should be broadly applicable across many computational and real-life domains, ranging from other biological processes like neural networks, evolution and ecosystems to communication and social networks.

1.2 Regulation of cellular RNA levels

1.2.1 The central dogma of molecular biology

Cells are the basic building blocks of all living organisms. All cells encode a primary heritable blueprints of their development and response to external stimuli, commonly referred to as their genetic material (or genome), and made of deoxyribonucleic acid
(DNA) [1]. In eukaryotic cells, which include all mammalian cells, the DNA is contained within a membrane enclosed internal structure referred to as the nucleus (figure 1-1a), while the area outside the nucleus is the cell's 'cytoplasm'. The DNA is a polymer of four bases (also called nucleotides): adenine, cytosine, guanine and thymine, which is commonly represented as an array of four characters (A,C,G and T respectively). DNA is structured as a double-helix with complementary base-pairing (A to T, G to C) allowing replication [2].

Large areas of the genome are transcribed (or expressed, see 1.2.3) into ribonucleic acid (RNA), and referred to as "genes". Although both RNA and DNA are nucleic acids, RNA is chemically different than DNA, commonly existing as a single strand, and containing the nucleotide uracil (U) instead of T.

One of the primary types of RNAs encoded by the genome is messenger RNAs (mRNAs). These RNAs are exported from the nucleus where they are made, into the cytoplasm where they are translated using a 3-letter code into proteins (figure 1-1a). Each of the 64 3-base sequences (or codons) specifies one of 20 amino acids, which are the building blocks of proteins. Finally, proteins form the structural basis and the molecular machinery that catalyze most of the chemical reactions within the cell, including the transcription of RNA and replication of DNA. This is the central dogma of molecular biology.

1.2.2 The dynamic RNA life cycle

While every cell in a multi-cellular organism contains essentially the same DNA, cells themselves have dramatically different morphologies, behavior, and functions. For example, the structure and function of a neuron in the brain is completely distinct than that of a skin cell. This diversity is largely driven through the specific regulation of the repertoire of RNAs that are expressed in a cell, and thus cells invest significant efforts in controlling the precise level of each type of RNA at each time, through a tight regulation of its life cycle.

The life of an mRNA begins with its production (section 1.2.3), and processing (section 1.2.4) in the nucleus. Once completed, the mRNA is exported from the nucleus
into the cytoplasm where it will be translated. While in the nucleus the mature mRNA molecule is linear, but once exported into the cytoplasm, certain protein complexes bind both its ends making it effectively circular. Most commonly, while in the cytoplasm, an mRNA is associated with ribosomes and translated into protein. The circular structure of the molecule allows for an easy recycling of ribosomes and continuous production of the protein. However, recent evidence suggest that an mRNA can also remain sequestered in cytoplasmic granules under certain conditions [3] such that it is not available for translation. Ultimately, an mRNA is degraded (section 1.2.5) into its building blocks.

But mRNA is not a mere messenger of the DNA blueprint for making proteins. Every mRNA molecule includes two main untranslated regions (UTRs) that are located before the coding region (5'UTR) or after it (3'UTR) and serve as regulatory sequences that independently or by interacting with the hundreds of RNA binding proteins (RBPs) in the cell determine the transcript's location, rate of translation and finally degradation.

While mRNAs are a key element of the central dogma, most RNA transcripts in the cell (95%-99%) do not code for proteins at all, and are made to use as structural, regulatory or catalytic molecules. For example, the production of proteins by the ribosome is catalyzed by an RNA based (ribosomal RNA, rRNA) rather than a protein based machinery, and micro-RNAs (miRNAs) are short regulatory RNA molecules that bind mRNAs and change their degradation and translation rates. The three major regulatory events of production, processing and degradation apply to practically all known RNA transcripts, although the specific details of each depend more on the specific type of molecule.

1.2.3 RNA production through transcription

Not all parts of the DNA are equally active. The repertoire of RNAs (and consequently also proteins) that are generated from the cell’s DNA template through the
**Figure 1-1:** The central dogma of molecular biology. (a) Areas of the DNA are transcribed within the cell’s nucleus into pre-mRNA, which is later processed in several steps to produce a mature and active mRNA. While in the nucleus the mature mRNA molecule is linear, but once exported into the cytoplasm, certain protein complexes bind both its ends making it effectively circular. While in the cytoplasm mRNA can be translated into proteins, sequestered in granules and finally degraded into its building blocks. (b) The production of RNA through transcription and its processing are coupled to each other in the nucleus. After transcription is initiated, a 5’ cap is added to the growing RNA chain at its 5’ end. The cleavage of the 3’ end and its polyadenylation lead to the termination of transcription. (c) Degradation of mRNA commonly start by deadenylation at the 3’ end and degradation from 3’ to 5’ by the exosome, but can also be initiated by decapping the 5’ end and 5’ to 3’ degradation or by endonucleolitic cleavage at the middle of the molecule which exposes 2 ends for both 3’ to 5’ and 5’ to 3’ degradation.

The process of transcription determines a cell’s function and survival. Thus RNA transcription is highly regulated, as cells transcribe a precise level of RNA from each gene at each time. Indeed, while some genes produce many copies of RNA, others are completely inactive. This different behavior is determined through precise regulatory mechanisms, including the structure of the segment of DNA that encodes the gene, as well as the binding of specialized proteins, known as transcription factors (TFs), which bind to the region of DNA near the start of the gene (the promoter of the gene) and enhance or suppress how frequently that gene’s DNA is copied to RNA.
Eukaryotic transcription is a complex process, involving several steps and enzymatic reactions [4] (figure 1-1b). Both RNA and DNA are nucleic acids, which use base pairs of nucleotides as a complementary language that allows to convert back and forth from DNA to RNA. During transcription, the enzyme RNA polymerase reads a DNA sequence and produces a complementary, antiparallel RNA strand. First, specific protein factors initiate transcription by enabling binding of RNA polymerase to promoter DNA sequences. In the following elongation step, the complementary DNA nucleotides break apart while RNA polymerase adds matching RNA nucleotides that pair with one of the DNA strands, and form a newly synthesized RNA strand. The new RNA strand is directional, its start is referred to as the 5’ end while its end is called the 3’ end. Transcription termination involves cleavage of the new RNA transcript which is coupled to template-independent addition of A-s at its new 3’ end (polyadenylation).

1.2.4 RNA maturation by processing

The conversion of an initial precursor into a functional mature transcript is key to the cellular regulation of RNA levels, as most transcripts go through several precise processing events, which separate a transcript’s production from its activity [5, 6]. These processing events commonly involve the removal of segments from the middle (splicing) or ends (trimming) of a transcript, chemically modifying certain bases within it (editing, capping) or adding bases to the end of the transcript (e.g., polyadenylation). As with many other levels of regulation, specific protein factors bind near processing sites (splice junctions, edited positions and others) by recognizing a sequence or structure signal, and control their conversion (e.g., inclusion or exclusion of exons, modification of edited nucleotides, selecting a polyadenylation site). All these decisions change the transcripts’ structure and eventually affect downstream events such as the RNA half life (e.g., through nonsense mediated decay) or its translation efficiency.

Eukaryotic mRNAs go through several standard processing steps [4] (figure 1-1b) before they become mature transcripts that can serve as templates for protein trans-
lation. A 5' cap structure is added to the transcript shortly after the 5' end emerges from the RNA polymerase in a process known as **capping**. In most mRNAs specific segments (called 'introns') are removed from the initial transcript through a process called **splicing**. Splicing is usually performed by an RNA-protein complex named the spliceosome, but some RNA molecules are also capable of catalyzing their own splicing. Sometimes identical pre-mRNA messages can be alternatively spliced in several different ways, allowing a single gene to encode multiple alternative transcripts and subsequently proteins. Termination of transcription involves the cleavage of the transcript by endonucleases associated with RNA polymerase, and adding a polymer of adenyl (i.e., A of variable length) to the RNA chain. Finally, in certain messages specific bases are **edited** by enzymes which convert them into a different base. The most well known example is the conversion of Adenine to Inosine (compatible to Guanosine) by the ADAR enzyme [7].

Not only mRNAs are processed. Many transcripts are produced as precursors and later converted into their mature form. For example, 22 nucleotides long miRNAs are transcribed as long pri-miRNAs (several hundred bases or more) that are processed through a double cleavage to generate the small active molecule. Ribosomal RNAs are also processed through a series of editing and cleaving events, which produce the different mature components of the ribosome.

### 1.2.5 RNA degradation

The lifespan of RNA molecules in living cells varies significantly, ranging from few minutes to several days, depending on the specific cell type and the specific RNA [8]. The regulation of a transcript’s stability significantly affects its abundance, and, most importantly, also determines how upstream regulatory changes (such as production and processing) integrate over time and space. For example, a highly stable transcript will continue to accumulate even though its production stops, while a quick change in transcription levels must be accompanied by fast degradation to affect the effective transcript amount. RNA stability and decay is regulated through a combination of proteins and RNA sequence and structure elements. Notably, since RNA
DNA) can be catalytic, some sequence elements are active by themselves and not only through interaction with proteins (e.g. ribozymes).

RNA is degraded by a combination of ribonuclease enzymes, including endonucleases, 3'-exonucleases, and 5'-exonucleases that act in several RNA degradation pathways. The "canonical" RNA degradation in eukaryotic cells (figure 1-1c) begins by shortening of the poly(A) tail of the mRNA by specialized exonucleases [4]. These are commonly targeted to specific mRNAs by a combination of regulatory sequences on the RNA and RNA-binding proteins. Poly(A) tail removal disrupts the circular structure of the message and destabilize the 5' cap binding complex. The mRNA is then subject to degradation by either the exosome complex, degrading it from the 3' to the 5' end or the de-capping complex, degrading it from the 5' to the 3' end. In this way, once degradation is initiated, messages are destroyed quickly.

Over the years, several other pathways for RNA degradation were suggested [8], which mostly vary in the way degradation is initiated. For example, endonucleases can cleave transcripts at specific positions, exposing a 3' and 5' ends which are than quickly degraded by exonucleases. RNA degradation is also sometimes initiated from the 5' end of the transcript, which is de-capped and further degradation follows.

1.3 Experimental methods for studying dynamic RNA regulation

1.3.1 Methods for assessing reaction rates within the RNA life cycle, and their limitations

Cellular RNA levels are the result of integrating several highly regulated processes for their production, processing and degradation. Therefore, in order to decipher the cellular response, the RNA life cycle has to be dissected and its components studied both individually and in integration. Consequently, over the years, several methods were presented, which dissect independent processes within the RNA life cycle and measure their rate.
The rate of RNA production can be measured through nuclear run-on assays [9, 10, 11, 12]. In this assay, transcription is halted in vivo and then reinitiated in isolated nuclei under conditions where new transcription is not initiated, and that allow labeling of the nascent RNA chains (e.g., by radiolabeled nucleotides), thereby enabling them to be distinguished from bulk RNA. In other assays, RNA polymerase is immunoprecipitated, and all RNAs associated with it are quantified [13].

Approaches to measure RNA degradation rates typically rely on measuring the level of existing RNAs' at several time points following blocking cellular RNA production (transcription) [14, 15, 16, 17]. Transcription is blocked either using drugs (e.g. actinomycin D) [14, 15] or by conditional RNA-polymerase mutants (e.g. inactive in higher temperatures) [16, 17].

Measuring the rate of RNA splicing is challenging. Earlier works transiently block transcription, either by tetracycline-regulated promoters [18] or reversible polymerase inhibitors (e.g., dichlorobenzimidazole ribofuranoside) [19], and quantify the level of intron inclusion either in specific splice junctions or genome wide by RNA sequencing (see 1.3.2). Recent works isolate chromatin associated RNA, and quantify intron inclusion levels within this population of recently transcribed RNA, on a genome-scale [20]. Many works predict positions with high editing frequency in transcripts by comparing frequency of edited nucleotide in measured RNA sequences to a reference genome [21, 22, 23, 24].

However, all these methods provide only a partial and biased perspective due to several systemic shortcoming. Nuclear run-on assays for measuring transcription rates are conducted ex-vivo in isolated nuclei [9, 10, 11], and estimating degradation rates by transcriptional inhibition severely affects normal cells growth and survival [25], which limits their relevance in vivo. Fractionation-based methods including polymerase [13] and chromatin [20] associated RNA extraction are limited by obtaining clean fractions, due to non-specifically bound RNAs, re-association of proteinRNA complexes in vitro and additional co-precipitating RNA binding proteins. Significant error rates in RNA sequencing methods puts in doubt any high editing frequency result [26, 27, 28]. Finally, their technical complexity makes their adaptation to dynamic settings more
1.3.2 RNA quantification assays

There is an enormous technical challenge in quantitatively measuring levels of each of tens of thousands of transcripts in an unbiased and systematic way. Advanced RNA quantification assays, systematically follow the quantitative dynamics of many transcripts simultaneously, unbiasedly and with high precision.

Quantitative real-time polymerase chain reaction (qRT-PCR) allows to simultaneously quantify several dozen RNA molecule. Using a non-specific fluorescent dye allows to follow in real time after the progression of DNA amplification with a sequence-specific probes, and the increase in fluorescence at every step directly correlates to the increase in starting material. Unlike other methods, it is a very quick and flexible approach that provides results within a couple of hours.

The Nanostring nCounter technology [29] accurately measures even small quantities of RNA. In this technique, a pre-designed set of probes is hybridized to an RNA sample. Each probe has a unique color-code and is uniquely matching a single gene in the sample. Following hybridization, the instrument counts the number of molecules from each color from images of the RNA sample. This approach allows to accurately quantify even small RNA quantities without using any amplification steps.

Massively parallel sequencing generates genome-scale data [30]. This approach determines the precise sequence of nucleotides within short (commonly 50-100 bases long) segments of RNA molecules. Repeating the process many times (\( \sim 10^8 \) times), each time randomly selecting a different starting point on a different RNA transcript produce some evidence for almost all transcribed bases. The quantity of reads that match to the sequence of a certain gene directly correlates to the relative number of transcripts of the gene.
1.3.3 Targeted inactivation of genes in vivo

A gene **knockout** (KO) is a genetic technique in which one of an organism's genes is made inoperative ("knocked out" of the organism), and is not able to produce any active protein [31]. Knockout is accomplished through a combination of techniques. A plasmid, a bacterial artificial chromosome or other DNA construct is generated in the lab which contains an aberrant copy of the gene in interest or parts of it. Individual cells are genetically transfected with this DNA construct. The construct is engineered to recombine with the target gene, which is accomplished by incorporating sequences from the gene itself into the construct. Recombination then occurs in the region of that sequence within the gene, resulting in the insertion of a foreign sequence to disrupt the gene. With its sequence interrupted, the altered gene in most cases will be translated into a nonfunctional protein, if it is translated at all. When the goal is to create a transgenic animal that has the altered gene, embryonic stem cells are transformed and then inserted into early embryos. The resulting animals with the genetic change in their germline cells can then often pass the gene knockout to future generations.

A gene **knockdown** (KD) is a genetic technique by which the expression of one or more of an organism’s genes are reduced [32]. The reduction can occur either through genetic modification (similar to knockout) or by treatment with short DNA or RNA oligonucleotides that have a sequence complementary to either gene or an mRNA transcript. The binding of this oligonucleotide to the active gene or its transcripts causes decreased expression either through the blocking of transcription (in the case of gene-binding), the degradation of the mRNA transcript (e.g. by small interfering RNA (siRNA) or RNase-H dependent antisense), or through the blocking of either mRNA translation, pre-mRNA splicing sites, or nuclease cleavage sites used for maturation of other functional RNAs.
1.4 Mathematical formalisms in dynamic RNA regulation

1.4.1 Models of regulatory interactions that control RNA levels

The cellular regulatory interactions that control RNA levels in cells are crucial for understanding how cells survive and respond to their environment. As these are generally too complex to allow informal reasoning, substantial computational efforts were invested in developing quantitative models that couple an appropriate mathematical formulation with the correct algorithmic solutions.

Indeed, regulatory interactions within cells can be approached at two levels of details, including (1) interactions between regulatory events that integrate to generate a final RNA level and (2) molecular interactions between regulators and sequences that yield RNA levels. Dozens of currently available approaches attempt to predict the molecular mechanisms that control temporal cellular responses [33, 34], while many fewer aim to build models that integrate several regulatory steps into a comprehensive computational framework that follows them dynamically and explains how they compute cellular RNA levels.

Most of these can be broadly classified into three categories: (1) co-expression network, (2) differential equation based and (3) dynamic bayesian networks.

Co-expression network are coarse-scale, simplistic models that rely directly on pairwise or low-order conditional association measures, such as correlation (partial or time-lagged) or (conditional) mutual information, for inferring the connectivities between genes [35, 36, 37, 38]. To increase their biological relevance, these methods commonly also integrate information on the genes’ sequence. These methods have the advantage of low computational complexity, and can scale up to very large networks of thousands of genes, but they do not model the network dynamics, and hence cannot perform prediction.
Differential equation based approaches are well established methods which have long been used for modeling biochemical phenomena, including gene regulation [39, 40] including at the level of RNA degradation [41]. This approach accurately models the detailed dynamics of biochemical systems in continuous time, but are also much more computationally intensive, and so far are only applicable to relatively small networks of a handful of genes. Moreover, these models typically require a previous knowledge of the components of the system and the interactions between them.

Dynamic bayesian networks models are based on solid principles of probability and statistics [42, 43]. These models accurately and compactly represent the joint distribution of a set of variables, using probability and graph theories. The graph structure is used to predict of the gene network behavior in unknown conditions, albeit not as detailed a level as differential equations based approaches.

The success of these tools is tightly coupled to the ability to measure circuit input, output, and wiring on a genomic scale. In the absence of some of this information, one inevitably faces major computational challenges in separating direct and indirect effects, inferring causality and working with hidden variables and unobserved data. As a result, most large-scale models do not incorporate detailed regulatory functions and either address sequence inputs or protein regulator inputs but not both, and attempt to learn the regulatory mechanisms by generalizing circuit models across regulators and expression levels of multiple genes. When available, temporal data can provide more insights into mechanism and help distinguishing correlation from causality.

Yet despite the many dozen studies, still most computational efforts focus on the regulatory circuits that take protein regulators and DNA sequences as input and yield RNA levels as output, while mostly ignoring the individual steps in RNA life-cycle, which are much more rarely measured directly.

### 1.4.2 Dynamic time-series models of cellular responses

Cellular responses unfold over time in reaction to environmental changes (e.g., differentiation, disease), yet much of the analysis of these responses relied on steady-state
measurements rather than time-series experiments or ignored their temporal aspects [34]. Models primarily focused on identifying elements that share common responses across experiments, and associating them with various cellular processes based on their response profiles. Some models did leverage the power of dynamic data, and attempt to model the dynamics of expression time courses.

The vast majority of dynamic models investigate RNA expression levels, which is the most easily obtained molecular quantification of cellular responses. Model of dynamics expression time-courses attempt to extract biologically meaningful timing aspects of the individual responses, and compare them across different conditions [44]. Indeed, temporal data can help filter noise in measurements, impute missing values and study transient changes. Several approaches have focused on capturing the dynamics of cell cycle time courses, yet these methods are tailored to the sinusoidal patterns in the cell cycle, and do not generalize to other types of time series. Other strategies showed how splines can be used to encode continuous gene expression profiles, and successfully align similar expression profiles that exhibit different temporal properties. Some methods have defined shape-based similarity metrics for expression time courses, for the purpose of gene clustering, but without attempting to extract or evaluate specific timing properties. Other approaches use a probabilistic or regression-based time series model to capture the temporal dynamics of gene expression data [45]. These methods use generic function representation, capable of capturing a broad family of response profiles, and avoid over-fitting by estimating the model parameters using clusters of genes, possibly obscuring finer-grained signal. Parametric modeling of the temporal response capture the typical temporal pattern, and explicitly represents biologically meaningful temporal properties [46], and can also be used to to group genes based on temporal properties [47].

Few models specifically investigate individual steps within the RNA life cycle, mostly due to the rarity of such measurements. A common approach equates production rates with RNA abundance measurements (expression levels), this making a major simplifying assumption (either implicitly or explicitly) that RNA levels are controlled by affecting production alone, while the molecule's subsequent processing, localiza-
tion and degradation rates are quick and constant [33, 34]. When quantification is available, models of individual steps within the RNA life cycle usually take the form of low-order conditional association measures (such as correlation) which divide genes into typical behaviors and do not include temporal relations in their models [14, 11]. Degradation changes over time are modeled either by a single change [11, 15, 16] or as a continuous shift [9, 14] over time.

1.4.3 Inferring RNA abundance from sequencing reads

High-throughput RNA sequencing (see section 1.3.2) offers the ability to discover new transcripts and measure their expression in a single assay. Several efficient and statistically principled algorithms [48] are available to analyze the enormous volume of raw sequencing reads that RNA sequencing experiments produce. RNA-seq analysis tools generally fall into three categories: (1) read alignment (2) transcript assembly and genome annotation (either through genome independent or genome guided methods) and (3) transcript and gene quantification.

Although several models have been applied to study the processing of RNA precursors [5], most genome-scale works focused on estimating the frequency and regulation of alternative processing from RNA sequences, either in a single condition [49], by comparing them across different cell types and organisms [50, 51], or dynamically over time [52, 53].

1.5 Mammalian immune dendritic cells

An organism’s immune system is a set of structures and processes that attempt to protects against harmful foreign living agents (e.g., bacteria, virus, fungi), termed pathogens. An immune system can detect a wide variety of foreign agents, distinguish them from the organism’s own healthy tissue and act to destroy them.

Dendritic cells (DCs) are antigen-presenting cells of the mammalian immune system [54, 55]. DCs are present in an immature state in tissues that are in contact with the external environment, (e.g. skin, inner lining of the nose, lungs, stomach and
intestines and the blood), and constantly sample the surrounding environment for pathogens through chemical signatures which are unique to subsets of pathogens. Upon detection, DCs become activated into mature dendritic cells and consequently grow the branched projections after which they are named. Upon maturation, DCs phagocytose the pathogens, degrade their proteins into small pieces, and present those fragments (antigens) at their cell surface. They migrate to the lymph node, where they present their surface antigens to other cells of the immune system (T and B cells), thus acting as messengers between the innate and the adaptive immunity, that initiate and shape the organism’s adaptive immune response.

The response of DCs to pathogen stimulation provides a compelling model of a temporal regulatory program in mammalian cells [56, 38]. Upon stimulation with pathogen components, DCs activate a major regulatory program, which unfolds over 24 h and involves the activation of ~1,700 genes and repression of ~2,000 genes [38], some peaking as early as 30 min, whereas others peak after 6 h or more. A recent study [38] identified over a hundred transcription factors, and at least a dozen RNA binding proteins in controlling this response, suggesting a regulatory control in several stages of the RNA life cycle.
Chapter 2

A model of RNA production and degradation dynamics

2.1 Introduction

Deriving general principles of RNA dynamics from genome-scale data requires quantitative models that couple an appropriate mathematical formulation with the correct algorithmic solutions. Although many studies attempt to predict the molecular mechanisms that control temporal cellular responses [33, 34], only very few [41] aim to integrate several RNA regulatory steps (i.e., production and degradation) into a comprehensive computational framework that follows them dynamically and explains how they compute cellular RNA levels (see also section 1.4.1). Approaches from both categories can be broadly classified into three categories: simplistic co-expression network that rely on pairwise associations, dynamic bayesian networks and detailed ordinary differential equation (ODEs) based models.

The success of these tools is tightly coupled to the ability to measure the cellular circuit’s input, output, and wiring on a genomic scale. While RNA expression levels can be easily measured, assessing the contribution of individual regulatory steps has been technically challenging. In the absence of some of this information, most works focus on transcriptional mechanisms and tacitly assume that degradation rates (per
gene) are constant over time [45, 42, 35, 37, 43]. Indeed, most large-scale models do not incorporate detailed regulatory functions, and attempt to learn the regulatory mechanisms by generalizing models across multiple genes.

Here, I develop a computational approach which combines ODE models with a parametric model of dynamic rate time evolution [57] to infer dynamic rate functions from temporal measurements of standard cellular RNA abundance (*RNA-Total*) and of newly transcribed RNA (obtained as metabolically labeled RNA, *RNA-4sU*, see section 3.2). The model estimates dynamic and gene specific profiles of RNA production and degradation rates, and allow us to test alternative models in order to discrimimates between temporally constant and dynamic degradation.

### 2.2 A first order dynamic model of RNA production and degradation

#### 2.2.1 A model of dynamic RNA regulation

A first order dynamic equation describes the gene-specific dynamic RNA levels under several simplifying assumptions: (1) assume that RNA levels integrate production and degradation rates alone, and ignore other events within the RNA life cycle (including its processing), and (2) assume that degradation acts equally on all RNA molecules of a specific gene, regardless of their sequence, the proteins bound to them or their location in the cell.

For a gene X, define the following rates:

\[
\alpha(T) = \text{production (by transcription) rate of gene } X \text{ at time } T \left( \frac{RNA}{min \cdot cell} \right)
\]

\[
\beta(T) = \text{degradation rate of gene } X \text{ at time } T \left( \frac{1}{min \cdot cell} \right)
\]

A first order dynamic model of gene-specific cellular RNA levels is therefore (figure
Figure 2-1: A model of dynamic RNA regulation. (a) The 'constant degradation' and 'varying degradation' models. A first-degree dynamical model (formula, right) models the expression level of a gene (grey curve) as a function of transcription (black) and degradation (green) rates. Parameters include an 'impulse' model for transcription (black curve), and either a constant function for degradation ('constant degradation' model, solid green line), or an 'impulse' model ('varying degradation' model, dashed green line). I fit them to the data (left, RNA-Total, blue, and RNA-4sU, red) by optimizing the likelihood function (separately per gene). (b) I compare the model's fit (black and grey curves) to the data (red and blue curves, respectively) and calculate the error.

2-1a):

\[
\frac{dX}{dt} = \alpha(t) - \beta(t)X(t)
\]

which is solved by integration:

\[
X(T) = \int_{-\infty}^{T} \alpha(t) - \beta(t)X(t)dt
\]

2.2.2 Measurement dynamics

For a gene X, let \(X(T)\) be its \(RNA-Total\) levels at time \(T\) after induction (\(RNA_{\text{cell}}\)), which is the overall RNA abundance within the cell of the gene X at the measurement time \(T\) (figure 2-1b). \(RNA-Total\) integrates RNA levels over the entire lifetime of the cell, whereas I am interested in the dynamics during the specific response (which
starts at time \( T = 0 \). Therefore, I use the initial condition:

\[ X(T = 0) = X_0 \]

and get the following dynamics:

\[ X(T) = X_0 + \int_0^T \alpha(t) - \beta(t)X(t)dt \]

While \( RNA-Total \) measures the overall RNA abundance, \( RNA-4sU \) measures only RNA transcripts that are metabolically labeled with 4-Thiouridine (4sU) during a short labeling pulse (see section 3.2). Let \( X^*(T; t_L, d_L) \) be the gene \( X \)'s \( RNA-4sU \) levels at time \( T \) after induction (figure 2-1b). This quantity represents only RNA molecules which were actively produced during a pre-defined labeling time starting at \( t_L \) minutes after induction and lasting for \( d_L \) minutes during which 4sU is present in the medium of responding cells (i.e., between \( t_L \) and \( t_L + d_L \) minutes after stimulation).

Therefore, \( RNA-4sU \) only locally integrates RNA levels over the labeling period, and at all times \( T \leq t_L \) no labeling occurs giving the initial condition:

\[ X^*(t_L; t_L, d_L) = 0 \]

and the following dynamics:

\[ X^*(T; t_L, d_L) = \begin{cases} \int_{t_L}^{T} \alpha(t) - \beta(t)X^*(t)dt & T > t_L \\ 0 & T \leq t_L \end{cases} \]

### 2.2.3 Estimating constant production rates

I make two simplifying assumptions based on current biological knowledge, which allow me to directly estimate a constant production rate from each \( RNA-4sU \) measurement, representing the average production rate during the relevant labeling period. These two assumptions are (1) the production rate \( \alpha \) is constant during the (very short) labeling period (between \( t_L \) and \( t_L + d_L \)), and (2) RNA is produced in
the nucleus while degradation is (mostly) a cytoplasmic process. After short labeling pulse \((d_L \leq 10\) minutes), most RNA-4sU is nuclear and therefore not degraded \((\beta_{t_L} = 0, \text{ see section 3.2})\). Based on these assumptions, I parametrize the model with a constant production rate \(\alpha\) and \(\beta(T) = 0\), simplifying the dynamic equation to:

\[
\frac{dX}{dt} = \alpha
\]

and allowing an exact solution using the initial condition \(X^*(t_L) = 0\):

\[
X^*(T; t_L, d_L) = \alpha T - \alpha t_L = \alpha(T - t_L)
\]

which gives the estimator:

\[
\alpha_{t_L} = \frac{X^*(T; t_L, d_L)}{d_L}
\]

### 2.3 Inferring dynamic rates from measurements

#### 2.3.1 Noise model

I assume an additive and independent Gaussian Noise \((N)\). For a measured RNA level \(x\), and a model prediction \(X(\theta)\) I can write -

\[
N \sim N(0, \sigma)
\]

\[
x = X(\theta) + N
\]

and consequently get that:

\[
x \sim N(X(\theta), \sigma)
\]

\[
\frac{x - X(\theta)}{\sigma} \sim N(0, 1)
\]

I independently estimate \(\sigma_{Total}\) for RNA-Total data, and \(\sigma_{4sU}\) for RNA-4sU data from experimental repeats.
2.3.2 Maximal likelihood optimization by gradient descent

I use a gradient descent based optimization to find the maximal likelihood parameters of the model \((\theta = [\alpha, \beta, X_0])\), which are the rate predictions.

I define the following annotation for measured data \((D)\) and model predictions \((\Theta)\):

\[
\begin{align*}
D & = \{x_t\}_{t=1}^m = \text{temporal data, including all measurements (RNA-4sU, RNA-Total)} \\
\Theta & = \{X(t; \theta)\}_{t=1}^m = \text{model predictions based on a specific choice of parameters (\theta)}
\end{align*}
\]

The likelihood term is therefore:

\[
L(D; \theta) = L(\{x_t\}; \{X(t; \theta)\}, \sigma_{Total}, \sigma_{4sU}) = \prod_{t=1}^m p(x_t | X(t; \theta), \sigma_{Total}, \sigma_{4sU})
\]

\[
= \prod_{t=1}^m N\left(\frac{x_t - X(t; \theta)}{\sigma}\right)
\]

For optimization, I consider the log likelihood ratio:

\[
LLR(D; \theta) = \log \frac{L(D; \theta)}{L(\Theta; \theta)} = \log L(D; \theta) - \log L(\Theta; \theta)
\]

\[
= \sum_{t=1}^m \log p(x_t | X(t; \theta), \sigma_{Total}, \sigma_{4sU}) - \\
- \log p(X(t; \theta) | X(t; \theta), \sigma_{Total}, \sigma_{4sU})
\]

\[
= \sum_{t=1}^m \log N\left(\frac{x_t - X(t; \theta)}{\sigma}\right) - \log N(0)
\]

\[
N(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}
\]

\[
\log N(0) - \log N(x) = \log \frac{1}{\sqrt{2\pi}} e^0 - \log \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} = \log e^0 - \log e^{-\frac{x^2}{2}} = \frac{x^2}{2}
\]

\[
-LLR(D; \theta) = \sum_{t=1}^m \log N(0) - \log N\left(\frac{x_t - X(t; \theta)}{\sigma}\right) = \sum_{t=1}^m \left(\frac{x_t - X(t; \theta)}{\sigma}\right)^2
\]

And optimize the following error function:

\[
E(\theta; D) = -LLR(D; \theta) = \sum_{t=1}^m \left(\frac{x_t - X(t; \theta)}{\sqrt{2\sigma}}\right)^2
\]
\[ \theta_{\text{opt}} = \arg \min_\theta (E(\theta; D)) \]

Therefore the error function takes the form of sum of squares, and can be optimized (to find a local optimum \( \theta_{\text{opt}} \)) by non-linear least squares curve fitting.

The derivative with respect to \( \theta \) is:

\[
\frac{d}{d\theta} E(\theta; D) = \sum_{t=1}^{m} \frac{d}{d\theta} \left( \frac{x_t - X(t; \theta)}{\sigma} \right)^2
\]

\[
= 2 \cdot \sum_{t=1}^{m} \left( \frac{x_t - X(t; \theta)}{\sigma} \right) \frac{d}{d\theta} \left( \frac{x_t - X(t; \theta)}{\sigma} \right)
\]

\[
= -\sum_{t=1}^{m} \frac{x_t - X(t; \theta)}{\sigma^2} \left( \frac{d}{d\theta} X(t; \theta) \right)
\]

and the gradient of the error function, is given by:

\[
\nabla(E) = -\left( \sum_{t=1}^{m} \frac{x_t - X(t; \theta)}{\sigma^2} \left( \frac{d}{d\theta_1} X(t; \theta) \right) \right)
\]

\[
= -\sum_{t=1}^{m} \frac{x_t - X(t; \theta)}{\sigma^2} \left( \frac{d}{d\theta_1} X(t; \theta) \right)
\]

\[
= -\sum_{t=1}^{m} \frac{x_t - X(t; \theta)}{\sigma^2} \nabla(X(t; \theta))
\]

The gradient \( \nabla(X(t; \theta)) \) (of \( X(t; \theta) \) with respect to the parameters \( \theta \)) is estimated from the dynamic equation by:

\[
\frac{d}{d\theta} \frac{d}{dt} X(t; \theta) = \frac{d}{d\theta} \alpha(t; \theta) - X(t; \theta) \frac{d}{d\theta} \beta(t; \theta) - \beta(t; \theta) \frac{d}{d\theta} X(t; \theta)
\]

\[
\frac{d}{dt} \frac{d}{d\theta} X(t; \theta) = \frac{d}{d\theta} \alpha(t; \theta) - X(t; \theta) \frac{d}{d\theta} \beta(t; \theta) - \beta(t; \theta) \frac{d}{d\theta} X(t; \theta)
\]

\[
\frac{d}{dt} \nabla(X) = \nabla(\alpha) - X(t; \theta) \nabla(\beta) - \beta(t; \theta) \nabla(X)
\]
For RNA-4sU, using the initial condition $X(T = t_L) = 0$, I get that:

\[ \nabla(X(t_L)) = \frac{d}{d\theta} X(0; \theta) = 0 \]

\[ \nabla(X(T)) = \int_{t_L}^{T} \nabla(\alpha(t)) - X(t; \theta) \nabla(\beta(t)) - \beta(t; \theta) \nabla(X(t)) dt \]

and for RNA-Total, using the initial condition $X(T = 0) = X_0$, I get that:

\[ \nabla(X(0)) = \frac{d}{d\theta} X(0; \theta) = 0 \]

\[ \nabla(X(T)) = \int_{0}^{T} \nabla(\alpha(t)) - X(t; \theta) \nabla(\beta(t)) - \beta(t; \theta) \nabla(X(t)) dt \]

### 2.3.3 Initializing the optimum search

As with every gradient descent method, I need to start the search from some initial guess for the model parameters. This guess has a significant impact on the optimization, since gradient descent will only find a local maximum.

To initialize the production function ($\alpha(t)$) I estimate constant production rates directly from RNA-4sU data (see section 2.2.3), and fit an impulse model by multiple random initializations. To initialize the degradation function, I first estimate degradation rates $\beta(t)$ directly from the dynamic equation by $\beta(t) = \frac{1}{X} \left( \alpha(t) - \frac{dX}{dt} \right)$, and fit either a constant function or an impulse model to these values (using multiple random initializations for fitting the impulse model). The parameters of the production and degradation functions ($\alpha(t)$ and $\beta(t)$) are the initialization point for gradient descent.

### 2.3.4 A Gaussian mixture prior on the parameter space

To assist optimization, I use a a Gaussian mixture prior on the parameter space.

I estimate a Gaussian from optimal parameters of each of the 8 expression clusters in the signature data (see section 3.3), resulting in a mixture of 8 Gaussian models for each parameter.

I adapt the likelihood function where the second term is derived as before, and the
first term is the prior:

\[
\log L(D; \Theta) = \log p(\Theta) \cdot p(D|\Theta) = \log p(\Theta) + \log p(D|\Theta)
\]

\[
P(\Theta) = p(\{\theta_j\}) = \frac{1}{8} \sum_{i=1}^{8} P(\{\theta_j\}|\mu_i, \sigma_i)
\]

\[
P(\{\theta_j\}|\mu_i, \sigma_i) = \sum_{j=1}^{n} N(\theta_j|\mu_{ij}, \sigma_{ij})
\]

### 2.4 Hypothesis testing

#### 2.4.1 Likelihood ratio test to select between competing hypotheses for degradation

I use a single parameter constant function \((\text{Constant}(T; \theta) = \theta)\) to describe a static response, and an 'impulse model' with 6 parameters \([46, 57]\) to describes a dynamic response to stimulation:

\[
\theta = [h_0, h_1, h_2, t_1, t_2, \lambda]
\]

\[
\text{Impulse}(T; \theta) = \frac{1}{h_1} \cdot \left( h_0 + \frac{h_1 - h_0}{1 + e^{-\lambda(T-t_1)}} \right) \cdot \left( h_2 + \frac{h_1 - h_2}{1 + e^{\lambda(T-t_2)}} \right)
\]

I therefore fit two alternative models to the data, independently per gene:

<table>
<thead>
<tr>
<th>Model</th>
<th>Production</th>
<th>Degradation</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>(\alpha(T) = \text{Impulse}(T; \theta_\alpha))</td>
<td>(\beta(T) = \beta)</td>
<td>6 + 1 = 7</td>
</tr>
<tr>
<td>Dynamic</td>
<td>(\alpha(T) = \text{Impulse}(T; \theta_\alpha))</td>
<td>(\beta(T) = \text{Impulse}(T; \theta_\beta))</td>
<td>6 + 6 = 12</td>
</tr>
</tbody>
</table>

The simpler 'constant degradation' model assumes that each gene has a temporally constant degradation rate, which can vary between genes. This simple model is often implicitly assumed in computational models of gene expression \([45, 42, 35, 37, 43, 40]\).

The more complex 'varying degradation' model assumes that the degradation rate of a gene changes over time, and represents such changes with an 'impulse' model.

I compare these two nested alternative models (\(\Theta_{\text{null}}\) is contained in \(\Theta_{\text{alt}}\)) by the likelihood ratio test. Using Wilkes theorem, when the sample size \((of\ D)\) is big
enough, I know how a likelihood ratio term ($E$) is distributed:

$$E(\Theta_{Null}, \Theta_{Alt}) = \log \frac{L(D; \Theta_{null})}{L(D; \Theta_{alt})} - 2 \cdot E \sim \chi^2(|\Theta_{alt}| - |\Theta_{Null}|)$$

Therefore, I can decide on a significance level where to reject the null hypothesis.

In this case, the null model is the 'constant degradation' model, which is nested in the 'dynamic degradation' model. I reject the null hypothesis in favor of the alternative dynamic model with significance level of 1% by comparing $E$ to $\chi^2$ distribution.

### 2.4.2 Goodness of fit test

I test the 'goodness of fit' test to find out how well the model fits the set of observations.

I compare the model’s predictions to the measurements (per gene) by assuming an additive and independent Gaussian noise model $N(x, \sigma)$. I estimate the variance parameter $\sigma$ either from experimental repeats (if available) or from genome-wide measurements to be the variance in expression values measured genome-wide.

I reject the tested model (usually 'constant degradation' model) with significance level of 1% by comparing it to $\chi^2$ distribution, but do not suggest an alternative model.

### 2.5 Summary

In this work I develop a computational approach which combines ODE models with a parametric model of the dynamic rate evolution [57] to model RNA production and degradation dynamics during biological responses to stimulation. I use temporal measurements of standard cellular RNA abundance ($RNA\text{-}Total$) and metabolically labeled RNA ($RNA\text{-}4sU$) to infer gene specific dynamic rate functions of RNA production and degradation. This model also discriminates between temporally constant and dynamic degradation, and identifies the specific temporal patterns of the changes. This provides a new tool to study how complex interactions between production and
degradation dynamics implement the dynamic transitions in RNA expression, and is broadly applicable across many computational and real-life domains.
Chapter 3

Principles of RNA production and degradation dynamics in mammalian cells

3.1 Introduction

In addition to transcriptional regulation, changes in RNA degradation can also significantly affect differential gene expression, and particularly in mammalian cells, where RNA half-lives are typically longer [15, 16].

Two key questions on the roles of transcription and degradation in regulating RNA levels arise: (1) Which of the two processes contributes most to shaping changes in RNA levels over time? and (2) Do such changes primarily result from variation of constant rates between genes or from variability of the rates for each gene over time? Answering these questions is hampered by the shortcomings of the indirect methods used for determining transcription and degradation rates, which may limit their relevance in vivo (see section 1.3.1).

The extent to which RNA stability contributes to dynamic changes in RNA levels is still unclear and debated. Most works focus on transcriptional mechanisms [42, 37, 43], tacitly assuming that degradation rates (per gene) are constant over time [40].
However, recent studies suggested that changes in a gene's mRNA level following stimulation are strongly affected by corresponding changes in its RNA degradation rate [40, 36], which may determine up to half of the temporal changes in RNA levels in mammalian cells [9]. As most previous studies concentrated on transcriptional changes, dynamic changes in degradation rates were rarely studied [14]. Here I combine metabolic labeling of RNA at high temporal resolution with advanced RNA quantification and computational modeling to estimate RNA transcription and degradation rates during the response of mouse dendritic cells to lipopolysaccharide, and discovered key principles of temporal RNA regulation in mammalian cells.

3.2 Direct estimation of RNA production by short metabolic labeling of RNA with 4sU

Metabolic labeling of RNA with 4-thiouridine (4sU), a naturally occurring modified Uridine, allows to distinguish recently-transcribed RNA from the overall RNA population, with minimal interference to normal cell growth [58, 59, 60, 61, 62, 63]. The modified base is incorporated into the growing RNA chain in place of Uridine, marking it, and serving as an attachment point for a biotin tag for easy separation of newly transcribed RNA from the total RNA population (figure 3-1a). Previous studies with 4sU labeling suffered from low resolution and lacked a systematic dynamic analysis. In these studies, labeled RNA was hybridized to standard microarrays, requiring relatively large quantities of RNA and hence lengthier 4sU labeling times (1-2h). Thus, most existing studies focused on variation between genes during steady state conditions [59, 62, 63], and a single 4 time points microarray study [60], though promising, lacked a systematic dynamic analysis.

Here, I used short metabolic labeling with 4sU to directly estimate RNA transcription rates in DCs. I added 4sU to DCs for a pre-defined labeling time (figure 3-1a), such that RNA molecules that were actively transcribed during that time are labeled by
Figure 3-1: Metabolic labeling of RNA with 4-thiouridine (4sU). (a) After stimulation, I add 4sU (red squares) to growing cells for a pre-defined time, collect the cells and extract total RNA (blue). Using biotin capture with streptavidin magnetic beads, I purify labeled RNA (red) from the total RNA extract. (b) Transcription with and without 4sU. When 4sU is present, it is incorporated into the growing RNA chain in place of uridine. (c) Purification using streptavidin magnetic beads. Total RNA extract is biotinylated by covalently linking biotin (orange) to 4sU, followed by binding to Streptavidin coated magnetic beads (light blue). Biotinylated (4sU labeled) RNA is magnetically isolated, whereas unlabeled RNA is washed out. Finally, cleaving the biotin-4sU disulfide bond releases the labeled RNA from the beads. (d) After 10 minutes labeling, most 4sU purified RNA is nuclear. Shown is the fraction of nuclear 4sU-RNA (out of nuclear and cytoplasmic 4sU-RNA) from DCs collected at 3h post-LPS stimulation, after different metabolic labeling times (10 and 45 minutes, light and dark blue, respectively). Expression levels in each fraction were quantified using qRT-PCR for 3 induced genes (ifit2, irfl and stat1) and two controls (28S, 18S). The 28S measurement is used for normalization, and thus not shown. The low nuclear levels of 18S might be due to contamination with unlabeled rRNA, which is more significant for rRNA because of its high abundance in the cell (98% of cellular RNA).
4sU (figure 3-1b). I isolated the entire cellular RNA population (RNA-Total), used an in vitro reducing chemical reaction to specifically and covalently link biotin to 4sU residues, and separated the 4sU labeled RNA (RNA-4sU) using biotin capture with streptavidin magnetic beads (figure 3-1c). Finally, I quantified the RNA levels of key genes of the LPS response in both populations, and showed that 4sU metabolic labeling specifically labels newly transcribed RNA, is reproducible, is consistently measured by qRT-PCR and nCounter, and has no significant effect on cellular function or transcriptional response of primary DCs.

Total cellular RNA levels (RNA-Total) globally integrate the effects of RNA transcription and degradation over the entire lifetime of the cell, whereas newly-transcribed RNA (RNA-4sU) contains only RNA that was actively transcribed during the labeling pulse, and hence represents a 'local integration' of average transcription and degradation. When labeling time is sufficiently short, the labeled RNA is still in the nucleus, and is subjected to little, if any, degradation (with the notable exception of aberrant transcripts), thus reflecting the average transcription rate. I extracted RNA-4sU separately from nuclear and cytoplasmic fractions, and quantified each fraction. Indeed, after 10 min labeling RNA-4sU is predominantly nuclear (more than 70% of mRNA-4sU is present in the nuclear fraction) and much less after longer labeling times (only 50% after 45 min, figure 3-1d). I therefore chose a labeling time of 10 minutes as an appropriate 'short' duration. Moreover, promoter binding by RNA polymerase II (Pol-II) peaks at or before RNA-4sU measurements following short labeling (figure 3-1e), supporting short-4sU metabolic labeling as a direct measurement of RNA transcription rates.
3.3 RNA transcription and degradation dynamics of a signature gene set

3.3.1 A high-resolution temporal response of signature genes

I used short metabolic labeling (10 min) followed by nCounter measurements [29] to assess transcription rates and RNA levels of 254 representative signature genes along a high-resolution time course, during the response of DCs to LPS (figure 3-2a). I selected the 254 transcripts based on a previous study [38], as representative of global mRNA profiles in this response, including key regulators, cytokines and other effectors, whose expression changes in this system. I measured RNA-Total and RNA-4sU at 15 min intervals over the first 3 hours post-LPS stimulation (spanning most changes in mRNA abundance in this response [38]).

3.3.2 Dynamic changes in RNA expression levels usually lag behind transcription rate changes by 15-30 min

I used k-means clustering of RNA-Total and RNA-4sU expression data (standardized and log2 transformed), using multiple executions (with random initialization) and finally selecting the result that minimizes the distances between genes and their cluster’s centroid. I iteratively increased the number of clusters as long as none of the clusters had less than 2% of the genes. I found eight coherent groups with distinct temporal patterns that cluster based on their transcription rates and expression profiles (figure 3-2c, section 3.6), distinguishing subtle temporal differences. For example, both group III (e.g., Egr1, Zfp36) and group IV (e.g., Cxcl1, Tnf) genes peak early in the response, but with a 30-minute difference in their peak times. Likewise, the expression of groups VI (e.g., Il12b, Il6,
**Figure 3-2:** RNA transcription and degradation dynamics of a signature gene set. (a) Measuring transcription rates with short metabolic labeling. I used short metabolic labeling (10 min, red lines), and measured the expression of RNA-Total (blue) and RNA-4sU (red) for 254 signature genes at 13 time points in 15 min intervals (rows) over the first 3 hours post-LPS stimulation. (b) Changes in RNA-4sU levels follow changes in pol-II binding and precede changes in total RNA levels. Shown are example time course profiles for selected genes for RNA-4sU expression (nCounter, red), RNA-Total expression (nCounter, blue) and pol-II binding at the promoter (ChIP, dashed gray). (c) Distinct temporal clusters of newly transcribed and total RNA. Shown are clusters of expression profiles (nCounter) for 254 signature genes (rows) based on RNA-Total (left) and RNA-4sU (right) measurements across 13 time points (columns). Cluster I includes the control genes. Cluster numbers (I-VIII) are noted on right; representative member genes are noted on left. Purple: high relative expression; white: mean expression; pink: low relative expression. (d) Peak transcription precedes peak expression by 15-30 minutes. Shown are average profiles (Y axis) for RNA-4sU (red) and RNA-Total (blue) for each cluster at each time point (X axis), ordered by cluster numbers (cluster I topmost; cluster VIII bottommost). The size of each cluster is indicated in brackets. Pearson correlation coefficient $\rho$ of the best time-lag correlation between transcription and expression is indicated on right, with the optimal time lag in square brackets.
Nfkbl) and VII (e.g., Ifit2, Il12a) genes constantly increases, but with different onset and saturation times.

I calculated a gene’s $k$-th time-shifted Pearson correlation between a gene’s expression and transcription profiles, then averaged over clusters, and looked for the shift $k$ with the maximal average correlation.

Changes in RNA-Total levels temporally correlate with corresponding changes in transcription rates ($\text{mean } r^2 = 0.73 \pm 0.3 \text{ SD}$), but on average lag behind them by 15-30 min (figure 3-2d, section 3.6). There is about twice as long a delay for down-regulated than for up-regulated genes, possibly because degradation rates are generally slower than transcription rates, delaying their effect on RNA levels. Overall, using the optimal time shift, variation in transcription rates explains 64% of the variation in RNA levels.

### 3.3.3 Changes in transcription rates shape temporal RNA profiles of most genes

I developed a computational approach to estimate the temporal profiles of transcription and degradation rates from measurements of RNA-Total and RNA-4sU. I compared two alternative models for degradation. The simpler ‘constant degradation’ model assumes that each gene has a temporally constant degradation rate, which can vary between genes, while the more complex ‘varying degradation’ model assumes that the degradation rate of a gene changes over time, and represents such changes with an ‘impulse’ model (see chapter 2 for more details).

I estimated the percent of variance in a set of measurements that is explained by the model’s predictions by the standard coefficient of determination ($r^2$) of the regression model, and corrected to account for the non-linearity of the model.

Most signature genes (83%, 210/254) retain the ‘constant degradation’ hypothesis (figure 3-3a), suggesting that changes in degradation rates contribute minimally to
Figure 3-3: The temporal profiles of transcription and degradation rates. (a) The 'constant degradation' model fits the majority of genes well. Shown is the distribution of the log likelihood ratios between the 'constant degradation' and 'varying degradation' models. Dashed line indicates the threshold for rejecting constant degradation ($p < 0.01$). (b) Genes that reject the 'constant degradation' model (red) show a bigger shift between rates predicted based on Actinomycin D data before LPS treatment (Y-axis) and at 2.5 hours after LPS stimulation (X-axis) compared with genes that retain the 'constant degradation' model (blue). Most commonly, degradation rates for these genes increase at 2.5 hours. Only genes with r2 of 0.8 or higher for predicted degradation rate are considered (100/254 genes). (c) The percent of genes per cluster (numbered as in figure 3-2) that reject the constant degradation model. (d) Varying degradation profiles estimated for the 44 genes that reject the 'constant degradation' model. Right: estimated degradation rates (relative rate; purple: high; pink: low) for the 44 genes (rows), clustered into 3 groups (A-C), across 12 time points (columns; excluding t=0 which is highly sensitive to noise due to low RNA levels). Asterisk: known regulators of RNA degradation. Left: mean degradation rate profile per cluster (bracket: number of genes in cluster). (e) Genes with peaked responses reject the 'constant degradation' model. Shown are two example genes (top: Cxcl1, bottom: Zfp36).
shaping their RNA levels dynamics during the first 3 hours of the response. On average, the 'constant degradation' model explains 78% (19% SD) of a gene's RNA levels variation, while the 'varying degradation' model, despite its substantial complexity, extends this by less than 10% (86% 14% SD). While the predicted constant RNA half-lives range from 10 to 70 min, the predicted varying half-lives span a much wider range (1-200 min). Measuring degradation rates by standard Actinomycin D treatment (see 3.6), although severely impacting cell growth and survival [25], supports the predicted fit to these models (figure 3-3b).

These results contradict recent works [9, 14], which proposed that changes in RNA degradation rates significantly affect temporal expression patterns in mammalian sensory pathways. Instead, they support an earlier view [45, 42, 35, 37, 43, 40] of the 'constant degradation hypothesis' that during such responses RNA levels are predominantly affected by changes in transcription rates.

3.3.4 Temporal changes in degradation are important for shaping 'peaked' responses

In a minority of genes (44 genes, 17%) changes in degradation rate significantly affect changes in RNA levels (figure 3-3a). The predicted degradation rate profiles are partitioned into 3 groups (A, B, C, figure 3-3d) and have only one or two prominent rate changes (figure 3-3d,e), suggesting a single underlying regulatory event.

These changes in degradation may uniquely contribute to shaping sharp 'peaks'. First, the genes that reject the 'constant degradation' model are concentrated in three clusters (III-V, figure 3-3c) containing many immediate-early genes (e.g., Fos, Jun, Egr1, Zfp36) with a sharper 'peak' in their expression profile. Second, I defined the 'maximal peak height' in the temporal profile \(X(t)\) by \(\max_t\{2 \cdot X(t) - X(t - 1)\} - \)
\( X(t+1) \), and ranked genes according to the 'peakiness' of their expression profiles (genes with highest 'maximal peak height' have the sharpest peaks). I find that the maximal peak height and the error of the fit to the 'constant degradation' model are highly correlated (Pearson \( r = 0.61, p < 10^{-26} \)). Finally, the fit error concentrates at the peak, and changes in the 'varying degradation' rate overlap the peak (figure 3-3e).

### 3.4 Genome-wide RNA transcription and degradation dynamics

#### 3.4.1 The genome-wide response to LPS stimulation by high throughput RNA sequencing

To generalize these findings to genome-scale, I adapted the 4sU metabolic labeling protocol for massively parallel sequencing (4sU-Seq, figure 3-4a). First, to extract labeled RNA in sufficient quantities for sequencing, I extended the labeling time to 45 minutes. As labeling time increases, RNA-4sU no longer directly approximates transcription rates, but a local integration of the average transcription and degradation rates. Second, due to the lower proportion of rRNA and other stable RNAs in newly transcribed RNA (1.8-4 fold reduction), I generated 4sU-Seq libraries without prior poly-A selection (see also 5).

I applied 4sU-Seq to mouse DCs in response to LPS stimulation, with an expanded temporal scope and reduced temporal resolution (figure 3-4a). I estimated expression by Reads Per Kilobase exon model per Million mapped reads (RPKM), and found that these match the nCounter data.
Figure 3-4: The genome-wide RNA transcription dynamics and its degradation profile. (a) Experiment overview. I isolated RNA-4sU (after 45 min of 4sU labeling, red) and polyA+ RNA-Total (blue) at 1h intervals (rows) over the first 6 hours of the response of DCs to LPS stimulation, and used massively parallel sequencing to measure RNA levels. (b) Assignment of the genome-wide sequencing data to the 8 distinct temporal clusters of newly-transcribed and total RNA identified based on the signature set data. Shown are expression profiles (RPKM) for the 10,106 expressed genes (rows) based on RNA-Total (left) and RNA-4sU (right) measurements across 11 time points (columns). Clusters are numbered on right and separated by solid horizontal black lines. Horizontal dashed lines: further sub-clustering of the data. Purple: high relative expression; white: mean expression; pink: low relative expression. (c) Distribution of predicted constant mRNA half-lives for the 9,448 genes expressed during the first 6 hours of the response to LPS stimulation that do not reject the 'constant degradation' model. Dashed lines distinguish 10 deciles (A-J, 10% increments, 35 transcripts with >200min half-life are included in the last decile).

3.4.2 Consistent genome-wide and small-scale measurements of newly transcribed RNA

I assigned each RNA-Seq gene to the nCounter cluster (out of clusters II-VIII) with the maximal Pearson correlation between the gene’s RNA-Seq expression profile and
the cluster's centroid (as calculated by k-means at 0, 1, 2 and 3 hours). If the correlation p-values for all clusters was non significant ($p > 0.01$), I assigned the gene to cluster I (the control genes cluster). I divided each cluster into sub-groups by applying a k-means clustering separately per cluster, and required a minimal cluster size of at least 3% of the genes (globally).

The genome-wide analysis shows similar patterns as the signature set. Of the 9,838 expressed Refseq [64] genes, the majority (83%) fit the patterns of the signature-set clusters (clusters II-VIII) based on their 0-3h expression profiles (figure 3-4b), with subgroups of genes (in clusters II, III, VI) that differ in their behavior during later time points (4-6h). The minority of genes (17%, cluster I) that did not fit the signature patterns included genes repressed (Ia) or induced (Ib) after 3h.

Each of the eight clusters is enriched with distinct functional annotations, consistent with their temporal pattern. For example, Cluster I, showing minimal changes in transcription rate or expression (during 0-3h), is enriched for glycolysis and ribosomal proteins, whereas Clusters II and III, with transient or sustained down-regulation of transcription rates, are enriched for cell cycle, mitochondrial, or oxidative phosphorylation genes. The transiently early induced Cluster V (1h peak) is enriched for inflammatory regulators (e.g. Tnf, Nfkbiz, II1a, II1b) and differentiation factors. Conversely, in cluster VI, which is enriched for their targets (inflammatory and immune signaling genes and differentiation factors' targets), the early induction is sustained.

### 3.4.3 Constant degradation rates are a genome-wide phenomenon and contribute to shaping temporal RNA levels

To account for the reduced temporal resolution and the longer labeling time in the genomic data, I adapted the modeling approach, and used only the simpler 'constant degradation' model (see chapter 2 for more details).

For the vast majority of genes, dynamic changes in degradation rates contribute little to changes in expression during the first 6 hours of the response. Most expressed
genes (94%, 9,274/9,838) show only a minimal discrepancy between the measured data and the expected values under the constant degradation model (goodness of fit test). For the remaining 6% (564/9,838), the constant degradation model is rejected, suggesting that at least one of the model’s assumptions is inaccurate (such as varying degradation or other post-transcriptional events). This group is enriched for inflammatory and immune signaling genes, and for targets of NFkB signaling, suggesting that these processes are dynamically regulated both transcriptionally and post-transcriptionally. Notably, unlike the high-resolution signature data, I cannot fit the 'variable degradation model' to 4sU-Seq data, nor can I use a likelihood ratio test (Methods) to determine that the degradation model per se is at fault.

Large differences in constant degradation rates between genes are associated with distinct functional and temporal patterns, and suggest that variation in degradation rates between genes contributes to shaping temporal RNA levels. For the 93% ‘constant degradation’ genes, the estimated 'constant' mRNA half-lives range from 3 to 200 min (figure 3-4c), match the nCounter predictions, and are correlated albeit shorter on average to other available estimates in this system and others.

I divided the 9,448 genes with non-significant discrepancy from the 'constant degradation' model into 10 deciles (~ 900 genes per group) based on their predicted mRNA half-life (A-J, figure 3-4c). The long half-life genes (groups I-J) are enriched for proteasome, ribosomal proteins, oxidative-phosphorylation and glycolysis genes: all are stable mRNAs that constitute a major fraction of total cellular mRNA. These genes are mostly members of cluster I (e.g. Rpl18a, Cox8a) with a stable, minimally changing expression pattern in the first 3h of the response. Short half-life genes (groups A-B) are enriched for transcription regulators (e.g., Klf7, Dmtf1), and for targets of many known miRNAs and TFs (e.g., Foxo1, Hif1a, p53). Furthermore, some sharp 'peaked' expression genes from clusters III-V (e.g., Il10, Btg2) have constant high degradation rates (rather than varying rates), an alternative means to the same end.
3.5 Summary

In this work I use metabolic labeling coupled with advanced RNA quantification assays and computational modeling to study RNA regulation in the response of mouse DCs to LPS. Leveraging the Nanostring nCounter technology for accurate multiplex measurement of RNA [29] and massively parallel sequencing [30], I significantly reduce metabolic labeling time to directly measure RNA transcription rates at high temporal resolution for a selected set of signature genes, and at a lower temporal resolution on a genome-scale. A new computational approach allows to decompose RNA levels into the separate contributions of RNA production and degradation, and estimate changes in degradation rates between genes and over time.

I discover key principles of temporal RNA regulation in mammalian cells. I find that changes in transcription rate highly correlate with changes in RNA level, preceding them by 15-30 min, with about twice as long a delay in down-regulated than up-regulated genes. In contrast to recent works [9, 14], I find that dynamic changes in degradation rates have minimal effect on most RNA profiles, but that they do play a unique role in genes with sharp 'peaked' responses. Genome-wide analysis shows substantial variation in degradation rates between genes, rather than over time, consistent with their regulatory and functional differences.

The RNA degradation rates that I estimated rank similarly (Spearman $\rho = 0.56$), but are consistently shorter in absolute values than other RNA half life estimates measured previously in fibroblasts by 1 hour of metabolic labeling [60], and are also highly correlated but consistently shorter than the half-lives I measured with Actinomycin D in DCs. While a proportional decrease in rates may result from difference in cell type, it is more likely to result from the inherent limitations in all existing methods to measure RNA stability. Specifically, the ability to accurately and reliably estimate the fraction of newly produced RNA-4sU out of the overall RNA population is highly limited with current labeling methods, and leads to a global shift in predicted half lives, just as we observe. However, this has no effect on the RNA stability rankings, on which we base our analysis, which are indeed more comparable between meth-
ods and reliable. Notably, some attempts to overcome this limitation of metabolic labeling use a known amount of cells from a different specie in the samples [65], a solution that is easily available in yeast but hard to implement in our case. Indeed, this discrepancy in absolute degradation rates highlights the inherent limitations in all techniques, and the need for direct and reliable methods to measure degradation rates.

Overall, this method provides a new and effective tool to simultaneously study several key cellular regulatory processes and model their interactions with each other, generating a complementary view to any RNA expression analysis, and deepening our understanding of the RNA life cycle.

3.6 Methods

RNA sample collection from mouse DCs. All animal protocols were reviewed and approved by the MIT / Whitehead Institute / Broad Institute Committee on Animal Care (CAC protocol 0609-058-12). Details on the DCs isolation and treatment are presented in full in [38]. For metabolic labeling of RNA, 4-thiouridine (Sigma) was added to a 150 μM final concentration at the appropriate times prior to RNA extraction. Total RNA was extracted with the miRNeasy kit’s procedure (Qiagen), and sample quality was tested on a 2100 Bioanalyzer (Agilent). For the high-resolution (signature) analysis, I extracted RNA starting at 0h and until 3h after LPS stimulation in 15 minutes intervals. I generated replicated samples for the 0, 30, 45, 60 and 75 minutes samples. For the lower resolution analysis (4sU-Seq), I extracted RNA starting 0h and until 6h after LPS stimulation in 60 minutes intervals.

I used 20 μg total RNA for the biotinylation reaction. 4sU-labeled RNA was biotinylated using EZ-Link Biotin-HPDP (Pierce), dissolved in dimethylformamide (DMF) at a concentration of 1 mg/mL, and stored at -80C. Biotinylation was done in labeling buffer (10 mM Tris pH 7.4, 1 mM EDTA) and 0.2 mg/mL Biotin-HPDP for 2h at room temperature. Unbound Biotin-HPDP was removed by chloroform/isoamylalcohol (24:1) extraction using MaXtract (high density) tubes (Qiagen). RNA was precipitated at 20,000g for 20 min with
a 1:10 volume of 5M NaCl and an equal volume of isopropanol. The pellet was washed with
an equal volume of 75% ethanol and precipitated again at 20,000g for 10 min. The pellet
was re-suspended in 100 μL RNase-free water. Biotinylated RNA was captured using Dyn-
abeads MyOneTM Streptavidin T1 beads (Invitrogen). Biotinylated RNA was incubated
with 100 μL Dynabeads with rotation for 15 min at room temperature. Beads were magnetics-
ically fixed and washed with 1x Dynabeads washing buffer. Flow-through was collected for
unlabeled preexisting RNA recovery. RNA-4sU was eluted with 100 μL of freshly prepared
100 mM dithiothreitol (DTT). RNA was recovered from eluates and washing fractions with
RNeasy MinElute Spin columns (Qiagen).

**Cell fractionation.** Nuclear and cytoplasmic fractions were separated using NE-PER
nuclear and cytoplasmic extraction (Thermo Scientific). RNA from each compartment
was extracted following the miRNeasy kit’s procedure (Qiagen), and sample quality was
tested on a 2100 Bioanalyzer (Agilent). On average, I obtained 5 times more cytoplasmic
than nuclear RNA. I used 20 μg total RNA from the cytoplasmic fraction, and 4 μg total
RNA from the nuclear fraction for the biotinylation reaction to reflect the original ratio.
Purification of labeled RNA was done as described above.

**qPCR measurement.** RNA was reverse transcribed with the Sensiscript RT kit (Qia-
gen). Real time quantitative PCR reactions were performed on the LightCycler 480 system
(Roche) with FastStart Universal SYBR Green Master Mix (Roche), and every reaction
was run in duplicate. The 28S-rRNA levels were used as an endogenous control for normal-
ization.

**RNA-polymerase II ChIP.** Dendritic cells were fixed with 1% formaldehyde, quenched
with glycerol, and washed with ice-cold PBS. Pelleted cells were re-suspended in SDS ly-
sis buffer. Samples were sonicated with a Branson 250 Sonifier, centrifuged at 13,000g
for 10 min, and diluted 10-fold in ChIP dilution buffer. After removing a control aliquot
(whole-cell extract), sample was incubated at 4C overnight with antibodies against the
CTD domain of Pol-II (Covance MMS-128P). Complexes were precipitated with protein
G-Dynabeads. Beads were washed sequentially with low-salt immune complex wash buffer,
high-salt immune complex wash buffer, LiCl immune complex wash buffer, and TE. Im-
munoprecipitated chromatin was eluted in elution buffer, incubated at 65C for 8 hr, and
treated with RNase A, proteinase K and purified with a MinElute Kit (Qiagen). Quantification of relative binding was done by hybridization to an nCounter codeset covering the promoter regions of the indicated genes.

**nCounter measurements and data processing.** Details on the nCounter system are presented in full in [29]. I hybridized 50-100 ng of RNA for 16 hours with the code set and loaded into the nCounter Prep Station followed by quantification using the nCounter Digital Analyzer. Code sets were designed and constructed to detect the 254 signature genes, and to capture multiple splice isoforms. Each probe set matches 100 bases long exonic sequence of the target genes, and therefore detects both pre-mRNA and mature mRNA. I normalized the nCounter data in two steps: (1) with internal positive spiked-in controls provided by the nCounter system, and (2) with a set of 8 control genes. I used the ratio of RNA quantity in each sample before and after RNA-4sU purification (0.022 0.015 SD), to normalize RNA-total and RNA-4sU samples relative to each other.

**Measuring degradation rates in Actinomycin D treated cells.** I added actD (3 μg/ml) directly to untreated cells ('0h experiment') and collected cells at 0.5, 1, 2, 4 and 6 hours after actD addition. I added actD (3 μg/ml) directly to LPS-treated cell cultures for 2.5h ('2.5h experiment') without removal of LPS, and collected cells at 2, 3 and 4h after actD addition. RNA levels for the 'signature set' genes were quantified using nCounter, and normalized only with the internal nCounter provided controls.

I used the standard first-degree model:

\[ \frac{dX}{dt} = \alpha(t) - \beta(t)X(t) \]

I assumed that after actD treatment: (1) transcription is blocked, i.e. \( \alpha(t) = 0 \) and (2) a constant decay rate; and therefore the model becomes:

\[ \frac{dX}{dt} = \beta X(t) \]

I compared the actD data to the model predictions with two complementary methods. First, I compared RNA levels from actD experiment and model predictions: I used the 'constant degradation' model with \( \alpha(t) = 0 \) and with the \( \beta(t) \) predicted by the model at
actD treatment time (0h or 2.5h), and the RNA-total levels measured at actD treatment
time, to predict expected RNA levels following actD treatment. I compared these with actD
data by least square error (normalized by each gene’s average expression level). Second, I
compared the degradation rates ($\beta$) predicted by the models with those calculated directly
from actD decay profiles. Based on the assumptions (above), the analytic solution is a
first-order exponential decay: $I$ fit a first-order linear model to the log-decay profile to
estimate the degradation rate ($\beta$). Only genes with a good fit ($R^2 > 0.8$) were used for
further analysis (100/254 genes).

I estimated mRNA half-life directly from the degradation rate ($\beta$), per time point, by:

\[
X(t_{1/2}) = X_0 \cdot e^{-\beta t_{1/2}} = \frac{X_0}{2},
\]

\[
t_{1/2}(T) = \frac{\log(2)}{\beta(T)}
\]

Therefore, for varying degradation rates, the half-life at time $T$ represents the expected
half-life if the degradation rate was fixed to its value at that time point.

**Preparation and processing of RNA-Seq and 4sU-Seq libraries.** I prepared the
4sU-Seq libraries using the 'control (nonstrand-specific)' protocol as described in [30], with
the following modifications: after DNase treatment, I used the entire RNA sample (without
polyA-RNA isolation) and fragmented RNA by incubation in 1X RNA fragmentation buffer
(Affymetrix) at 80 degrees for 4 minutes. I prepared the RNA-A+-Seq libraries using the
'dUTP second strand (strand specific)' protocol as described in [30], except I fragmented
RNA as for the 4sU samples. I used the Illumina Genome Analyzer (GAII). I sequenced two
lanes for each RNA-A+-Seq sample, corresponding to 45 million paired-end reads/sample
(90 million single reads, 76 bases long) on average; and a single lane for RNA-4sU sample,
corresponding to 33 million paird-end reads/sample (68 million single reads, 76 bases long)
on average.

**Read Alignment.** I aligned all reads to the mouse reference genome (NCBI 37, MM9)
using the TopHat aligner [49] with default parameters. To estimate rRNA levels, I mapped
all reads (with the same procedure) to mouse rRNA sequences from GenBank [66]. About
80% of RNA-Seq (RNA-total) reads were uniquely mapped to the genome, and 85% of these
mapped in pairs (average insert size of 644 bases 2886 SD), while only 0.2% of the reads
mapped to rRNA. About 60% of 4sU-Seq (RNA-4sU) reads were uniquely mapped to the genome, and 65% of these mapped in pairs (average insert size of 699 bases \( \pm 5746 \) SD), while 30% of the reads mapped to rRNA. Thus, in 4sU-Seq libraries, rRNA accounts for the high fraction of reads that were not aligned (or not uniquely aligned) to the genome.

**Transcriptome reconstruction.** I extracted mouse RefSeq genes (exons and introns), microRNA and tRNA annotations from the UCSC genome browser [67]. I extracted rRNA sequences from GenBank [66] and other non-coding RNA sequences from Rfam [68]. I generated non-coding RNA annotations by aligning their sequences to the mouse genome using BLAST [69].

**Quantification of transcript abundance from RNA-Seq data.** I estimated the expression of a transcript X in both RNA-A+-Seq and RNA-4sU-Seq by standard 'Reads Per Kilobase exon model per Million mapped reads' (RPKM), as previously described in [70], but defined it over exons alone. I normalized RPKM estimates using: (1) a selected set of control genes as previously described [71], and (2) adjusting the mean and standard deviation of each sample’s \( \log_2(RPKM) \) values. I used the ratio between RNA quantity in each sample before and after RNA-4sU purification (0.046 \( \pm 0.019 \) SD), to normalize RNA-total and RNA-4sU relative to each other. I took genes with \( \log_2(RPKM) \geq 2 \) for at least one time point (10,106 genes, which are 40% of the Refseq [64] genes), and excluded transcripts with significant antisense expression (to avoid biases arising from strand-specificity), leaving us with 9,838 genes.

**Functional Enrichments.** I calculated the enrichment of a query set of genes X for an annotation A using hypergeometric p-value between two groups [72], and a 1% False Discovery Rate (FDR) significance threshold, across all annotations I tested. Functional annotations of the mouse genome were taken from the Molecular Signature Database (MSigDB) [73], c2 (curated gene sets), c3 (motif gene sets) and c5 (GO gene sets).
Chapter 4

Models of RNA processing dynamics

4.1 Introduction

Genome-scale data on the RNA sequence provides a unique opportunity to study the processing of a precursor RNA transcript into a functional mature RNA, and specifically the removal of intronic segments (splicing) and the alteration of specific bases on the transcript post-transcriptionally (editing). However, the enormous volume of raw sequencing data that RNA sequencing experiments produce necessitate statistically principled formulations and efficient algorithms to convert it into biologically meaningful quantities. Coupling the appropriate measurement resolution with the correct algorithmic solutions allows to study specific events within the multi step processing of a transcript at a uniquely high resolution and quantify each step independently from others. Indeed, measuring RNA splicing at per-junction resolution rather than averaging it over an entire transcript, allows to obtain a much more biologically accurate picture, and to also study the integration of these individual events into a complete processing procedure.

In order to quantify and model transcript processing we must measure transcripts at high sequence resolution. Traditional approaches to study RNA splicing used EST
data [74] and exon arrays [75], and therefore were inherently limited by the sequencing depth (EST) or prior annotations (exon arrays). Measurement quality was greatly enhanced by the introduction of next generation sequencing, which provided new and exciting opportunities to model transcript splicing [5]. However, the limited ability to isolate and measure unstable precursors and processing intermediates bounds most analysis to study the fully processed forms. Most genome-scale works focused on estimating the frequency and structure of mature and alternative transcripts, either in a single condition [49], by comparing them across different cell types and organisms [50, 51], or dynamically over time [52, 53]. Recently, more elaborate models of RNA splicing used enrichment of precursors within distinct cellular fractions of nuclear or chromatin associated RNAs [20], yet still averaged splicing rates across an entire transcript. Therefore, to date, junction specific RNA splicing rates were rarely measured and analyzed on a genome-scale, and never dynamically.

Many works aimed to predict positions in transcripts which are altered post transcriptionally (edited) by comparing transcripts sequences to a reference genome [21, 22, 23, 24]. These are significantly limited by the high error frequency of sequencing methods [26, 27, 28] and by single nucleotide variants within a non-controlled population (e.g., in humans). There are only a few well established editing events, and probably many more which are unknown, yet their rate of editing and its dynamic evolution were never studied.

Here I propose a computational approach to (1) recover the splicing-junction specific precursor and mature RNA levels and (2) predict edited positions and their editing levels from deep sequencing of ribosomal depleted (see chapter 5) total and metabolically labeled RNA (see section 3.2). These estimates are then used within a dynamic ODE model to infer dynamic rate functions per processing event. This model estimates dynamic and event specific profiles of RNA production, processing and degradation rates, and discriminates between temporally constant and dynamic rates.
4.2 A second order model of the dynamic RNA life-cycle: production, processing and degradation

The processing of a transcript is a complex process that commonly involve multiple steps, some occur co-transcriptionally [76] and others only after production is completed. I therefore take a gradual approach in modeling RNA dynamics, starting from a simple single-step processing and following with a multi-step processing model that considers either a single or multiple junctions.

4.2.1 A single step junction processing model

I consider the processing of a specific junction $J$ within a transcript (figure 4-1a,b), which might be the only processing step for some transcripts with a single splicing junction.

I develop a dynamic system model under the assumptions that (1) degradation acts \textit{only} on the mature transcript (after processing), (2) degradation acts equally on all mature RNAs (regardless of their location, sequence or structure) and (3) a precursor is spliced into a mature transcript in a single step.

I define the following rates:

\begin{align*}
\alpha(T) &= \text{junction’s transcription (production) rate at time } T \ (RNA/min \cdot cell) \\
\beta(T) &= \text{junction’s degradation rate at time } T \ (1/min \cdot cell) \\
\gamma(T) &= \text{junction’s processing rate at time } T \ (1/min \cdot cell)
\end{align*}

Then, by the assumptions above, I describe the junction’s time evolution (including...
Figure 4-1: Models of the dynamic RNA life-cycle. (a) An example transcript with 2 splice junctions: between exons A and B, and between exons B and C. The 4 different splicing intermediates are denoted below. (b) Independent junction processing model. Includes a separate model for each of the two junctions, with specific parameters for production, processing and degradation. (c) Integrated whole transcript processing models. A single step model (top) assumes the entire transcript (both junctions) are processed in a single step, therefore only two RNA species exist: the precursor and the mature RNA. A multi step processing model (bottom) includes separate processing steps for each junction. There are 3 options to describe the order: combinatorial (top) considers all orders of splicing between the junctions, 5' to 3' processing (middle) assumes the 5'-most junction is processed first, and 3' to 5' processing (bottom) assumes the 3'-most junction is processed first. (d) A co-transcriptional combinatorial processing model allows for processing of a junction to occur once both exons flanking the junction are produced. Therefore splicing can occur before the entire transcript is produced, and this assumption significantly increase the number of states in the model.

its production, degradation and processing) with a second order model:

\[
\frac{dP}{dt} = \alpha(t) - \sum_i \gamma_i(t)P \\
\frac{dM_i}{dt} = \gamma_i(t)P - \beta(t)M_i
\]

where \( P \) is the precursor level and \( M_i \) are the mature RNA levels (including alternative transcripts). Assuming also that \( X = M_1 + ... + M_m + P \) (no additional states except
precursor and mature transcripts), I converge to the following dynamics:

\[
\frac{dX}{dt} = \alpha(t) - \beta(t) \sum_i M_i
\]

### 4.2.2 Multi-step transcript processing models

When there are two or more junctions within a single transcript, modeling each junction independently does not allow to share parameters, such as the production and degradation rates, between junctions. Therefore, there is an advantage in modeling the entire processing of a transcript within a single model that considers multiple processing steps and estimates multiple processing intermediates. To do that I remove assumption (3) in the previous model (section 4.2.1) but add a new assumption: (3) splicing occurs following the complete transcription of the gene.

For a gene $G$ with $m$ splice-junctions, I first define:

- $J^k = \text{the set of all } \binom{m}{k} \text{ splice junction groups of size } k$
- $J = \cup_{i=0}^{m} J^i = \text{the set of all splice junction groups } \left( \sum_{i=0}^{n} \binom{n}{k} = 2^m \right)$
- $Q^S = \text{the precursor with the junctions in } S \text{ unspliced } (\forall S \in J)$

I consider several alternative hypotheses about the order of processing:

1. First, consider the case that splicing occurs as a **single step** (as with the junction specific model in section 4.2.1) (figure 4-1c). Therefore, the system description converges to:

   \[
   \frac{dQ^{\{1,\ldots,m\}}}{dt} = \alpha(t) - \gamma(t)Q^{\{1,\ldots,m\}}
   \]

   \[
   \frac{dQ^o}{dt} = \gamma(t)Q^{\{1,\ldots,m\}} - \beta(t)Q^o
   \]

2. Splicing of junctions occurs in spatial order **from 5' end to 3' end** of a
transcript (figure 4-1c), allowing the following system's description:

\[
\frac{dQ^{(1,\ldots,m)}}{dt} = \alpha(t) - \left( \sum_{i=1}^{m} \gamma_i(t) \right) Q^{(1,\ldots,m)}
\]

\[2 \leq k \leq m - 1 \quad \frac{dQ^{(k,\ldots,m)}}{dt} = \gamma_{k-1}(t)Q^{(k-1,\ldots,m)} - \gamma_k(t)Q^{(k,\ldots,m)}
\]

\[\frac{dQ^S}{dt} = \gamma_m(t)Q^S - \beta(t)Q^S
\]

This description includes \( m \) states \( \{Q^{(k,\ldots,m)}\} \), and is parametrized by the functions \( \alpha(t), \beta(t), \gamma_1(t), \ldots, \gamma_m(t) \).

3. Splicing of junctions occurs in spatial order from 3' end to 5' end of a transcript (figure 4-1c), resulting in the following dynamic model:

\[
\frac{dQ^{(1,\ldots,m)}}{dt} = \alpha(t) - \left( \sum_{i=1}^{m} \gamma_i(t) \right) Q^{(1,\ldots,m)}
\]

\[2 \leq k \leq m - 1 \quad \frac{dQ^{(1,\ldots,k)}}{dt} = \gamma_{k-1}(t)Q^{(1,\ldots,k)} - \gamma_k(t)Q^{(1,\ldots,k)}
\]

\[\frac{dQ^S}{dt} = \gamma_m(t)Q^S - \beta(t)Q^S
\]

This description includes \( m \) states \( \{Q^{(1,\ldots,k)}\} \), and is parametrized by the functions \( \alpha(t), \beta(t), \gamma_1(t), \ldots, \gamma_m(t) \).

4. Splicing of junctions can occur in any order (combinatorial) (figure 4-1c), giving us the following dynamic model:

\[\forall 1 \leq k \leq m - 1; \forall S \in J^k \quad \frac{dQ^S}{dt} = \sum_{i \in S} \gamma_i(t)Q^{(S,i)} - \left( \sum_{i \in S} \gamma_i(t) \right) Q^S
\]

\[\frac{dQ^S}{dt} = \sum_{i=1}^{m} \gamma_i(t)Q^{(i)} - \beta(t)Q^S
\]

This description includes \( 2^m \) states \( \{Q^S\}_{S \in J} \), and is parametrized by the functions \( \alpha(t), \beta(t), \gamma_1(t), \ldots, \gamma_m(t) \).
Finally, note that the assumption that splicing occurs following the complete transcription of the gene is inaccurate by some evidence on co-transcriptional splicing [76]. Relieving this assumption will significantly increase the number of states in the model, as it will include also partially produced transcripts (figure 4-1d), and is therefore not sufficiently supported by the available data (which includes only short sequencing reads) and computational resources.

4.3 Inferring precursor and mature RNA abundance from RNA sequencing

In order to infer dynamic models from RNA sequencing data, we must develop the appropriate statistical formalism that converts simple counts of reads in an RNA sequencing library into biologically meaningful estimates of transcript levels, including RNA precursor, mature and splicing intermediates. Here I develop approaches to quantify precursor and mature RNA levels (1) as a transcript average (2) independently per splice junction and (3) per-junction estimates within a larger transcript context.

4.3.1 A generalized RPKM estimates an average precursor abundance

A standard measure of transcript expression from RNA-Seq data is the "Reads per Kilobase of exon Model" (RPKM):

\[
\begin{align*}
R_{\text{exon}}(X) &= \text{number of reads mapped to the exons of transcript } X \\
L_{\text{exon}}(X) &= \text{the length (in kilo-bases) of the exons of transcript } X \\
R\text{PKM}(X) &= \frac{10^9 \cdot R_{\text{exon}}(X)}{L_{\text{exon}}(X) \cdot \sum_x R_{\text{exon}}(x)}
\end{align*}
\]
Figure 4-2: Inferring precursor and mature RNA abundance from RNA sequencing. (a) Using RNA-4sU sequencing data to study RNA processing. Sequencing reads in the RNA-4sU libraries originate from either precursor mRNA (P) or mature mRNA (M). While mRNA reads map only to exons, the precursor mRNA reads map to both exons and introns. A 'generalized RPKM' estimates newly transcribed pre-mRNA expression by the RPKM of a gene's introns alone, and overall newly transcribed RNA expression (pre-mRNA + mRNA) by the RPKM of a gene's exons. (b) Junction-specific precursor and mature RNA levels estimation from ribosomal depleted RNA sequencing data. The alignment of sequencing reads (colored rectangles) to the genome (gray) around a splicing junction includes 6 different types: reads aligned entirely within either the donor or the acceptor exon (green), reads aligned entirely within the intron (orange), reads alignment starts at the donor exon and continues into the intron (dark blue), reads alignment starts at the intron and continues into the acceptor exon (light blue), and finally reads alignment starts at the donor exon, skips the intron and continues at the acceptor exon (red).
Figure 4-2: Using the counts of reads from the 6 different alignment types within a binomial model allows to estimate two separate values for the level of precursor (P) and of mature RNA (M). (c) The alignment of sequencing reads (colored rectangles) to the genome (gray) at a splicing junction with an alternatively skipped exon. Alignments are of 12 different types: aligned entirely within either the donor, acceptor or skipped exon (green), aligned entirely within one of the two constitutive introns (orange), alignment starts at an intron and continues into an exon or vice versa (blue), an alignment that starts at one exon, skips the intron and continues to the following exon (light red) and an alignment that skips both introns and the alternative exon (dark red). The basic binomial model for a single isoform is extended for transcripts with alternative splicing. The extended model estimates the expression levels of a single precursor (P) and several alternative mature isoforms (M1, M2).

As reads mapped to exons can arise from either the mature or the precursor of a transcript, this estimates the overall RNA levels of a gene X (precursor, mature and any partially processed RNA).

To estimate the average abundance of a transcript’s precursor, I use intronic reads, which predominantly arise from unprocessed precursors (figure 4-2a). I therefore generalize the RPKM quantification separately on either introns or exons of the transcript, to approximate the average precursor abundance:

\[ RPKM_E(X) = \frac{10^6 \cdot R_{\text{exon}}(X)}{L_{\text{exon}}(X) \cdot \sum_x R_{\text{intron}}(x) + R_{\text{exon}}(x)} \sim \text{Precursor} + \text{Mature} \]

\[ RPKM_I(X) = \frac{10^6 \cdot R_{\text{intron}}(X)}{L_{\text{intron}}(X) \cdot \sum_x R_{\text{intron}}(x) + R_{\text{exon}}(x)} \sim \text{Precursor} \]

As with standard RPKM, generalized RPKM is a relative measure, estimating the relative abundance of precursor and mature RNA as well as that of different transcripts. Notably, although in theory intronic reads can also arise from excised introns and lead to over-estimated precursor abundance. However, previous studies [77, 78, 79, 80] suggest that excised introns are degraded very fast, and I also find an equivalent read coverage of introns and intron-exon junctions in the sequencing data, suggesting that this is unlikely and making the original assumption valid.
4.3.2 A binomial model estimates junction-specific precursor and mature RNA levels

Not all introns of a transcript are spliced simultaneously, hence a model that predicts precursor and mature RNA levels specifically per junction of a transcript is much more biologically accurate. In such a model, for each junction I estimate how many transcripts either have the unspliced junction (precursor) or the spliced junction (mature).

I develop a binomial model to estimate from RNA sequencing reads the \( \geq 2 \) RNA species that arise at a single splice junction: \( P \), the precursor RNA, which is the initial transcript (before processing occurred), and \( M_1, \ldots, M_m \) (\( m \geq 1 \)), the mature RNAs, which are the processed result, and can be several (due to alternative splicing).

I start with defining the following constants, which are characteristic of the sequencing data:

\[
\begin{align*}
N & = \text{the number of reads in the entire sequencing library} \\
r & = \text{the read length (in bases) for the sequencing library} \\
w & = \text{margin (in bases) for mapping to junctions (to minimize mapping errors)}
\end{align*}
\]

I term as \textit{splicing features} all the different genomic segments that define a splicing event, including the constitutive donor and acceptor exons, and one or more alternative introns and exons, as well as the junctions between them (figure 4-2b,c). Each of the splicing features defines a unique type of reads (e.g., reads that map to the donor exon, reads that map to the exon-exon junction etc).

Let \( F \) be the set of all splicing features for a given junction. I define the following matrices:

\[
R = \begin{pmatrix}
R_1 \\
\vdots \\
R_{|F|}
\end{pmatrix}
\]
\[ L = \begin{pmatrix} L_1 \\ \vdots \\ L_{|F|} \end{pmatrix} \]

\[ L_i = \begin{cases} |i| - r & i \text{ is an exon or an intron} \\ r - w & i \text{ is a junction} \end{cases} \]

\[ E = \begin{pmatrix} E_{1,P} & \ldots & E_{|F|,P} \\ E_{1,M_1} & \ldots & E_{|F|,M_1} \\ \vdots & \ddots & \vdots \\ E_{1,M_m} & \ldots & E_{|F|,M_m} \end{pmatrix} \]

\[ E_{i,X} = \begin{cases} 1 & i \in X \\ 0 & \text{otherwise} \end{cases} \]

where \( R_i \) is the number of reads that map to feature \( i \), \( L_i \) is the effective length (in bases) of feature \( i \) and \( E_i \) is a vector that assigns each feature to the RNA species that contain it (e.g., an intron is assigned only to the precursor \( P \)).

I make the basic assumption that the probability of seeing a read that starts at each base is uniform across the transcriptome. Therefore, the number of expected reads for each feature \( i \in F \) follows a binomial distribution with parameter \( p_i \) as the probability of 'success', i.e., that a read falls within the feature's region (and therefore counts towards \( R_i \)):

\[ R_i \sim \text{Binomial}(N, p_i) \]

\[ p(R_i|N, p_i) = \binom{N}{R_i} \cdot p_i^{R_i} \cdot (1 - p_i)^{N-R_i} \]

This probability \( p_i \) is directly related to the frequency \( \psi_P, \psi_{M_1}, \ldots, \psi_{M_m} \) of each RNA specie (i.e., precursor and mature RNAs):

\[ |\text{transcriptome}| = \sum_{x \in \text{transcriptome}} P_x |P_x| + \sum_j M_j^x |M_j^x| \]

\[ \psi_P = \frac{P_x}{|\text{transcriptome}|} \]

71
\[
\psi_{M_j} = \frac{M_j^2}{\text{transcriptome}} \\
\bar{\psi} = [\psi_P, \psi_{M_1}, ..., \psi_{M_m}] \\
\Rightarrow p_i = \bar{\psi} \cdot L_i \bar{E}_i
\]

and these frequencies \(\psi_P, \psi_{M_1}, ..., \psi_{M_m}\) estimate the expression of \(P, M_1, ..., M_m\) (up to a global normalization constant).

Mapping of RNA sequencing reads to the genome, and specifically to splicing junctions is not perfect, and can introduce biases in the data. For example, if only the last few bases of a read match a splicing junction, either skipping the intron or continuing into it, most mapping algorithms will not be able to accurately assign these last bases, leading to a bias in measuring junctions.

### 4.3.3 Generalizing the binomial model to predict levels of a transcript’s splicing intermediates

The above formalism can be easily generalized to predict the levels of each splicing intermediate that is considered by a multi step processing model \((Q^{S_0}...Q^{S_n})\), which will have \(n = 2^m\) for the combinatorial model, \(n = m + 1\) for the processive models or \(n = 1\) for the single step model, see also section 4.2.2).

For this generalized case, consider a gene \(G\) with \(m\) constitutive splice-junctions that is measured by \(5m + 1\) splicing features \((m + 1\) exons and \(m\) introns, exon-exon, intron-exon and exon-intron junctions). Let \(F\) be the set of all \(5m + 1\) features of the gene, so I generalize \(E\) to define whether or not each feature is part of the intermediate RNA species:

\[
E = \begin{pmatrix}
e_{1,Q^{S_0}} & \cdots & e_{5m+1,Q^{S_0}} \\
\vdots & \ddots & \vdots \\
e_{1,Q^{S_n}} & \cdots & e_{5m+1,Q^{S_n}}
\end{pmatrix}
\]

\[
e_{i,X} = \begin{cases} 
1 & i \in X \\
0 & \text{otherwise}
\end{cases}
\]
leaving both the expression for the frequency of all RNA species \( \tilde{\psi} = [Q_{S_0}, ..., Q_{S_n}] \)
and for the error function unchanged.

### 4.3.4 A case studies of two typical splice junctions

I consider two case studies of typical junctions, which indeed cover the majority of likely events.

**Case study: a constitutively expressed junction.** Consider a constitutive splicing event between two exons: \( E_{\text{donor}} \) and \( E_{\text{acceptor}} \), where a precursor \( P \) is converted into a mature junction \( M \) (figure 4-2b). This event is measured by 6 'features':

<table>
<thead>
<tr>
<th>Feature id</th>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( E_d )</td>
<td>donor exon</td>
</tr>
<tr>
<td>2</td>
<td>( E_a )</td>
<td>acceptor exon</td>
</tr>
<tr>
<td>3</td>
<td>( I )</td>
<td>intron</td>
</tr>
<tr>
<td>4</td>
<td>( EE )</td>
<td>exon-exon junction</td>
</tr>
<tr>
<td>5</td>
<td>( IE_d )</td>
<td>intron-exon junction at the donor site</td>
</tr>
<tr>
<td>6</td>
<td>( IE_a )</td>
<td>intron-exon junction at the acceptor site</td>
</tr>
</tbody>
</table>

I therefore define the following length matrices:

\[
L = \begin{pmatrix}
L_1 \\
\vdots \\
L_6
\end{pmatrix} = \begin{pmatrix}
|E_d| - r \\
|E_a| - r \\
|I| - r \\
r - 2w \\
r - 2w \\
r - 2w
\end{pmatrix}
\]

\[
E = \begin{pmatrix}
E_{1,P} & \ldots & E_{6,P} \\
E_{1,M} & \ldots & E_{6,M}
\end{pmatrix} = \begin{pmatrix}
1 & 1 & 1 & 0 & 1 & 1 \\
1 & 1 & 0 & 1 & 0 & 0
\end{pmatrix}
\]

**Case study: a skipped exon.** Consider a junction with 3 exons: \( E_{\text{donor}} \), \( E_{\text{acceptor}} \) and \( E_{\text{alternative}} \), where the alternative exon is either excluded \( (M_1) \) or included \( (M_2) \)
from the mature transcript (figure 4-2c). This event is measured by 9 types of read 'features':

<table>
<thead>
<tr>
<th>Feature id</th>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$E_d$</td>
<td>donor exon</td>
</tr>
<tr>
<td>2</td>
<td>$E_{a1}$</td>
<td>acceptor exon 1</td>
</tr>
<tr>
<td>3</td>
<td>$E_{a2}$</td>
<td>acceptor exon 2</td>
</tr>
<tr>
<td>4</td>
<td>$I$</td>
<td>intron</td>
</tr>
<tr>
<td>5</td>
<td>$EE_1$</td>
<td>exon-exon junction 1</td>
</tr>
<tr>
<td>6</td>
<td>$EE_2$</td>
<td>exon-exon junction 2</td>
</tr>
<tr>
<td>7</td>
<td>$IE_d$</td>
<td>intron-exon junction at the donor site</td>
</tr>
<tr>
<td>8</td>
<td>$IE_{a1}$</td>
<td>intron-exon junction at acceptor site 1</td>
</tr>
<tr>
<td>9</td>
<td>$IE_{a2}$</td>
<td>intron-exon junction at acceptor site 2</td>
</tr>
</tbody>
</table>

I therefore define the following read and length matrices:

$$L = \begin{pmatrix} L_1 \\ \vdots \\ L_9 \end{pmatrix} = \begin{pmatrix} |E_d| - r \\ |E_{a1}| - r \\ |E_{a2}| - r \\ |I| - r \\ r - 2w \\ r - 2w \\ r - 2w \\ r - 2w \end{pmatrix}$$

$$E = \begin{pmatrix} E_{1,P} & \ldots & E_{9,P} \\ E_{1,M_1} & \ldots & E_{9,M_1} \\ E_{1,M_2} & \ldots & E_{9,M_2} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

74
4.3.5 Maximum likelihood optimization of the binomial model

To infer the expression of \( P, M_1, ..., M_m \) I look for the \( \psi_P, \psi_{M_1}, ..., \psi_{M_m} \) values that optimize the likelihood of the data by the binomial model:

\[
LL(\psi; R, L, N) = \log \prod_i p(R_i|N, L, \psi) = \log \prod_i \left( \frac{N}{R_i} \right)^{R_i} (1 - \psi \cdot L_i \bar{E}_i)^{N-R_i} = \sum_i \log \left( \frac{N}{R_i} \right) + R_i \log(\psi \cdot L_i \bar{E}_i) + (N - R_i) \log(1 - \psi \cdot L_i \bar{E}_i) \]

\[
\hat{\psi} = \arg\max_\psi LL(\psi; R, L, N) = \arg\max_\psi \sum_i \log p(R_i|N, L, \psi)
\]

Giving rise to the following error function:

\[
\text{Err}(\psi|R, L, N) = - \sum_i \left( \log \left( \frac{N}{R_i} \right) + R_i \log(\psi \cdot L_i \bar{E}_i) + (N - R_i) \log(1 - \psi \cdot L_i \bar{E}_i) \right)
\]

The gradient of the error function is therefore:

\[
\nabla(\text{Err}) = -N \begin{pmatrix} 
\sum_i L_i E_{i,P} \frac{R_i}{N - \psi \cdot L_i \bar{E}_i} \\
\sum_i L_i E_{i,M_1} \frac{R_i}{N - \psi \cdot L_i \bar{E}_i} \\
\vdots \\
\sum_i L_i E_{i,M_m} \frac{R_i}{N - \psi \cdot L_i \bar{E}_i} 
\end{pmatrix}
\]

Which should be equal to zero at the optimum (\( \forall X \in \{ P, M_1, ..., M_m \} \)):

\[
\sum_i L_{i,X} \frac{R_i}{N - \psi \cdot L_i \bar{E}_i} \frac{\psi \cdot L_i \bar{E}_i}{(1 - \psi \cdot L_i \bar{E}_i)(\psi \cdot L_i \bar{E}_i)} = 0
\]

Unfortunately, this equation does not have an exact solution, and therefore I must use approximate algorithms to optimize the error function. Using derivative-free methods, such as the 'Neadler-Mead simplex algorithm' [81] (matlab implementation), I find a
Figure 4-3: Inferring dynamic RNA processing rates. (a) An overview of the inference from the single junction processing model. The model includes a production (black), processing (orange) and degradation (green) functions. By integrating production and processing, the level of precursor ($P$; blue) is estimated, and by integrating that with the degradation rate the level of mature RNA ($M$; red) is estimated. (b) I fit the model parameters to the data (left; RNA-4sU and RNA-4sU) by optimizing the likelihood function (separately per gene).

local optimum of this error function, and thus infer the expression of $P, M_1, ..., M_m$.

4.4 Inferring dynamic RNA processing rates

Here, I develop a framework which combines 2nd order ODE models with a parametric model of dynamic rate time evolution [57] to infer dynamic rate functions from temporal measurements of standard cellular RNA abundance (RNA-Total) and of newly transcribed RNA (obtained as metabolically labeled RNA, RNA-4sU, see section 3.2). These models estimates dynamic profiles of RNA production, processing and degradation rates either at a transcript resolution, or at a higher per-junction resolution.
4.4.1 Direct inference of constant processing rates from RNA-4sU data

Let $P^*(T; t_L, d_L), M^*(T; t_L, d_L)$ be the gene $X$'s precursor and mature RNA levels within an RNA-4sU population (newly produced) at time $T$ after induction. This quantity represents only RNA molecules which were actively produced during a pre-defined labeling time starting at $t_L$ minutes after induction and lasting for $d_L$ minutes during which 4sU is present in the medium of responding cells (i.e., between $t_L$ and $t_L + d_L$ minutes after stimulation). Therefore, RNA-4sU only locally integrates RNA levels over the labeling period, and at all times $T \leq t_L$ no labeling occurs (meaning $X^*(T; t_L, d_L) = 0$), which gives the initial condition:

$$P^*(t_L; t_L, d_L) = M^*_i(t_L; t_L, d_L) = 0$$

and the following dynamics:

$$P^*(T; t_L, d_L) = \begin{cases} \int_{t_L}^{T} \alpha(t) - \sum_i \gamma_i(t)P^*(t)dt & T > t_L \\ 0 & T \leq t_L \end{cases}$$

$$M^*_i(T; t_L, d_L) = \begin{cases} \int_{t_L}^{T} \gamma_i(t)P^*(t) - \beta(t)M^*_i(t)dt & T > t_L \\ 0 & T \leq t_L \end{cases}$$

I make several simplifying assumptions, which allows to directly estimate a constant processing rate using $P^*$ and $M^*_i$ values: (1) the 'single step processing' model (2) no alternative splicing (3) The rates $\alpha, \gamma$ are constant during the (very short) labeling period (between $t_L$ and $t_L + d_L$), and (4) labeled RNA is not degraded ($\beta = 0$) during the first $d_L$ min (it takes more than $d_L$ min for newly-transcribed RNA to reach the cytoplasm).

Under the single step processing model I get the following system description:

$$\frac{dP}{dt} = \alpha - \gamma P$$

$$\frac{dM}{dt} = \gamma P$$
which is solved by:

\[ P^*(T) = \frac{\alpha}{\gamma}(1 - e^{-\gamma(T-t_L)}) \]
\[ M^*(T) = \alpha(T-t_L) - \frac{\alpha}{\gamma}(1 - e^{-\gamma(T-t_L)}) \]

Now I define:

\[ f(x) = \frac{1 - e^{-x}}{x} = y \Rightarrow f^{-1}(y) = \frac{1}{y} - W_0 \left( \frac{e^{1/y}}{y} \right) \]

Lambert's W:

\[ W_0 \left( \frac{e^{1/y}}{y} \right) \cdot \frac{W_0 \left( \frac{e^{1/y}}{y} \right)}{y} = \frac{e^{1/y}}{y} \]

and get the estimators:

\[ M^*(T) = \alpha(T-t_L) - \frac{\alpha}{\gamma}(1 - e^{-\gamma(T-t_L)}) = \alpha_{t_L}(T - t_L) - P^*(T) \]
\[ \Rightarrow \alpha_{t_L} = \frac{P^*(T) + M^*(T)}{T - t_L} \]
\[ \frac{M^*(T)}{P^*(T)} = \frac{\alpha(T-t_L) - \frac{\alpha}{\gamma}(1 - e^{-\gamma(T-t_L)})}{\frac{\alpha}{\gamma}(1 - e^{-\gamma(T-t_L)})} = \frac{\gamma(T - t_L)}{1 - e^{-\gamma(T-t_L)}} - 1 \]
\[ = \frac{\gamma d_L}{1 - e^{-\gamma d_L}} - 1 \]
\[ \frac{P^*(T) + M^*(T)}{P^*(T)} = 1 + \frac{M^*(T)}{P^*(T)} = 1 + \frac{\gamma d_L}{1 - e^{-\gamma d_L}} - 1 = \frac{\gamma d_L}{1 - e^{-\gamma d_L}} \]
\[ \frac{P^*(T)}{P^*(T) + M^*(T)} = \frac{1 - e^{-\gamma d_L}}{\gamma d_L} = f(\gamma d_L) \]
\[ \Rightarrow \gamma_{t_L} = \frac{1}{d_L} f^{-1} \left( \frac{P^*(T)}{P^*(T) + M^*(T)} \right) \]

4.4.2 Maximum likelihood optimization by gradient descent

I use a gradient descent based optimization method to find the maximum likelihood parameters of the model \((\theta = [\alpha, \beta, \gamma, X_0])\), just as I did in section 2.3.2. This approach is based on comparing the model predictions for each of the model states (corresponding to distinct RNA species of the transcript) termed \(x\) to a direct estimation (from sequencing counts) of their levels, termed \(X\) (figure 4-3). This direct
estimation is available through the frameworks I developed earlier (section 4.3) for estimating RNA abundance (either the generalized RPKM or the binomial model) of different processing intermediates and alternative mature transcripts.

I optimize the sum of squares error function (see section 2.3.2), where the data \( x \) and the model predictions \( X \) include both precursor \( (P) \) and mature \( (M) \) abundance:

\[
Err(\theta; D) = \sum_{t=1}^{m} \left( \frac{x_t - X(t; \theta)}{\sqrt{2\sigma}} \right)^2
\]

\[
\theta_{opt} = \arg\min_{\theta}(Err(\theta; D))
\]

\[
\nabla(Err) = -\sum_{t=1}^{m} \frac{x_t - X(t; \theta)}{\sigma^2} \nabla(X(t; \theta))
\]

The gradient \( \nabla(X(t; \theta)) \) (of \( X(t; \theta) \) with respect to the parameters \( \theta \)) is directly calculated from the model's dynamic equations, and is used to find a local optimum \( \theta_{opt} \) by non-linear least squares curve fitting.

For example, for the junction specific single step processing model, the gradient is estimated by:

\[
\frac{d}{dt} \frac{d}{d\theta} P(t; \theta) = \frac{d}{d\theta} \alpha(t; \theta) - P(t; \theta) \frac{d}{d\theta} \gamma(t; \theta) - \gamma(t; \theta) \frac{d}{d\theta} P(t; \theta)
\]

\[
\frac{d}{dt} \nabla (P) = \nabla(\alpha) - P(t; \theta) \nabla (\gamma) - \gamma(t; \theta) \nabla (P)
\]

\[
\frac{d}{dt} \frac{d}{d\theta} M(t; \theta) = P(t; \theta) \frac{d}{d\theta} \gamma(t; \theta) + \gamma(t; \theta) \frac{d}{d\theta} P(t; \theta) - M(t; \theta) \frac{d}{d\theta} \beta(t; \theta)
\]

\[
-\beta(t; \theta) \frac{d}{d\theta} M(t; \theta)
\]

\[
\frac{d}{dt} \nabla (M) = P(t; \theta) \nabla (\gamma) + \gamma(t; \theta) \nabla (P) - M(t; \theta) \nabla (\beta) - \beta(t; \theta) \nabla (M)
\]

4.4.3 Maximum likelihood optimization from sequencing counts by derivative free methods

An alternative optimization strategy is to incorporate the binomial model into the dynamic model, and optimize both simultaneously by comparing to sequencing counts.

In this approach, instead of comparing the model predictions \( x \) to their estimators \( X \), I plug them into the binomial model (i.e., \( \psi = x \)), and predict read counts, which
are directly compared to the counts I have in the data.

Formally, let $F$ be the set of all splicing features, let $t = 1...n$ be the sample times, let $\theta$ be the parameters of the dynamic model and let $Q^1, ..., Q^k$ be its states. I define the following matrices:

$$O^\theta = [\bar{o}(t; \theta)] = \begin{pmatrix} Q^1(1; \theta) & \cdots & Q^k(1; \theta) \\ \vdots & \ddots & \vdots \\ Q^1(n; \theta) & \cdots & Q^k(n; \theta) \end{pmatrix}$$

$$R = \begin{pmatrix} R_1(1) & \cdots & R_1(|F|) \\ \vdots & \ddots & \vdots \\ R_n(1) & \cdots & R_n(|F|) \end{pmatrix}$$

$$N = \begin{pmatrix} N_1 \\ \vdots \\ N_n \end{pmatrix}$$

where $O_{ij}$ is the predicted expression levels (by $\theta$) per model state and time, $R_{ij}$ is the read counts per splicing feature and time and $N_t$ is the library sizes per time.

The binomial model parameters $p$ are given as a function of the expression levels $\psi$ by $p = \psi \cdot (L \cdot E)$. I simply substitute the parameters $\psi$ of the binomial distribution with the expression levels predicted by the model ($O^\theta$) such that that $p = O^\theta \cdot (L \cdot E)$.

More explicitly, the probability for a feature $i$ at time $t$ is

$$p_i(t) = \bar{o}(t; \theta) \cdot L_i \tilde{E}_i = \sum_{j=1}^{k} Q^k(t; \theta) L_i E_{i,M_j}$$

which defines the following probability of the number of reads $R_i(t)$ as a function of the parameters $\theta$:

$$\log p(R_i(t)|N_t, p_i(t)) = \log \left( \frac{N}{R_i(t)} \right) + R_i(t) \log(\bar{o}(t; \theta) \cdot L_i \tilde{E}_i)$$

$$+ (N - R_i(t)) \log(1 - \bar{o}(t; \theta) \cdot L_i \tilde{E}_i)$$
\[ \log p(R|\theta, N, L) = \sum_{t=1}^{n} \sum_{i=1}^{|F|} \log p(R_i(t)|N_t, p_i(t)) \]

Let \( \hat{\psi} \) be the optimal \( \psi \) by fitting the binomial model directly to reads, then:

\[ \hat{\psi}_t = \hat{\psi}_t \cdot L_t \hat{E}_t \]

\[ \log p(R_i(t)|N_t, \hat{\psi}_t) = \log \left( \frac{N}{R_i(t)} \right) + R_i(t) \log \hat{\psi}_t + (N_t - R_i(t)) \log(1 - \hat{\psi}_t) \]

\[ \log p(R|\hat{\psi}, N, L) = \sum_{t=1}^{n} \sum_{i=1}^{|F|} \log p(R_i(t)|N_t, \hat{\psi}_t) \]

which allows to define the likelihood of the model parameters \( \theta \) by:

\[ LL(\theta; R) = \log p(R|\theta, N, L) \]

\[ LLR(\theta; R) = \log p(R|\hat{\psi}, N, L) - \log p(R|\theta, N, L) \]

\[ = \sum_{t=1}^{n} \sum_{i=1}^{|F|} \log p(R_i(t)|N_t, \hat{\psi}_t) - \log p(R_i(t)|N_t, p_i(t)) \]

\[ = \sum_{t=1}^{n} \sum_{i=1}^{|F|} R_i(t)(\log \hat{\psi}_t - \log p_i(t)) + (N - R_i(t))(\log(1 - \hat{\psi}_t)) \]

\[ - \log(1 - p_i(t)) \]

To estimate the model parameters \( \theta \), optimize the likelihood:

\[ \theta_{opt} = \arg\max_{\theta} LL(\theta; R) = \arg\min_{\theta} LL(\theta; R) = \arg\min_{\theta} LLR(\theta; R) \]

and the error function is:

\[ Err(\theta; R) = LLR(\theta; R) \]

\[ = \sum_{t=1}^{n} \sum_{i=1}^{|F|} R_i(t)(\log \hat{\psi}_t - \log p_i(t)) + (N - R_i(t))(\log(1 - \hat{\psi}_t)) \]

\[ - \log(1 - p_i(t)) \]

I can use derivative-free methods, such as the 'Neadler-Mead simplex algorithm' [81]
Figure 4-4: Identifying transcript editing positions from RNA-Total and RNA-4sU sequencing. I identify candidate edited positions where the original nucleotide (red) is at higher frequency in the newly-produced (RNA-4sU, dark gray, top) population, while the distribution shifts to the edited nucleotide (blue) in RNA-Total (light gray, bottom). I count single nucleotide distribution in newly-produced (RNA-4sU) and total-RNA, and look for positions with different distributions by maximal likelihood test, with 5% FDR correction (top). I require no significant difference (rank-sum test) between RNA-4sU (dark gray) and RNA-Total (light gray) in the distribution of 3 quality control measurements: base qualities, nucleotide’s position on the read and the read’s strand, as such differences might indicate on a sequencing bias (bottom). Only positions that pass both tests are considered as editing sites.

(matlab implementation) to find a local optimum of this error function.

4.5 Identifying transcript editing positions from RNA-Total and RNA-4sU sequencing

Since RNA-4sU is enriched for recently produced transcripts that are still not fully processed, it will include many fewer edited transcripts than RNA-Total. I therefore expect a different distribution of nucleotides in edited positions between the two RNA populations: the edited nucleotide will have a higher frequency in RNA-Total data than in RNA-4sU data. I thus predict candidate editing sites by comparing nucleotide distributions between RNA-Total and RNA-4sU, and selecting positions with a significantly different distribution between the two.
Formally, for each position $p$ in the genome, I denote by $E_p(x|L)$ the frequency of the nucleotide $x$ within sequencing reads aligned to position $p$ in the library $L$. If $p$ is edited from $x$ to $y$, then $E_p(y|4sU) < E_p(y|Total)$ and $E_p(x|4sU) > E_p(x|Total)$, resulting in a different nucleotide distribution at position $p$ between RNA-Total and RNA-4sU sequencing libraries.

Within sequencing reads that align to position $p$, the probability of observing each nucleotide follows a multinomial distribution. I infer the multinomial model at position $p$ with a maximum likelihood estimator from the RNA-Total sequencing data:

$$D(p) \sim \text{multinomial}(f_A, f_C, f_G, f_T)$$

$$f_x = \frac{\text{count}(X)}{\#A + \#C + \#G + \#T + \#N}$$

and predict the probability of the RNA-4sU sequencing data gives this model:

$$\text{prob}([\#A, \#C, \#G, \#T, \#N]; M(p)) = f_A^{\#A} \cdot f_C^{\#C} \cdot f_G^{\#G} \cdot f_T^{\#T}$$

I use information of the reference genome sequence as a prior on the distribution, as a Dirichlet distribution with the 10 pseudo-counts for the reference nucleotide, and a single pseudo-count for each of the other nucleotides. I select as candidate editing sites all positions where $\text{prob} < 0.01$, with 5% FDR correction. Notably, this approach works well with a fully homozygous reference, as is the case here using an homozygous laboratory mice strain, but will require some adaptation to work on less uniform data (e.g., human samples).

I use 3 quality control measurements to filter candidate sites where the difference in distribution is likely to result from sequencing biases.

1. **Read position.** require a uniform distribution of the position within the sequencing read that generated each observed nucleotide (using a Rank-Sum test).

2. **Quality score.** require a uniform distribution of the sequencing quality score of the sequencing read that generated each observed nucleotide (using a Rank-Sum test).
3. **Read strand.** require a uniform distribution of the strand to which the sequencing read that generated each observed nucleotide is mapped to (by a Fisher exact test).

I only consider positions for which all 3 tests have a non-significant pvalue ($\text{prob} \geq 0.01$).

### 4.6 Summary

In this work I developed a new modeling approach which provides a unique look into (1) the relative frequencies of precursor and mature RNA species within the overall population of RNAs present in cells and (2) the dynamic response at each of a transcript’s junctions. This provides a new tool to study how complex interactions between RNA production, processing and degradation dynamics implement the dynamic transitions in RNA expression.

The models I describe here make several assumptions that are likely to be biologically inaccurate in some cases. First, all dynamic models I develop here assume that splicing occurs following the complete transcription of the gene despite some evidence for co-transcriptional splicing [76]. Second, in theory intronic reads can arise not just from unspliced junctions but also from excised introns and lead to over-estimated precursor abundance. However, previous studies [77, 78, 79, 80] suggest that excised introns are degraded very fast, and make this case highly unlikely. Finally, the approach for predicting editing sites works well with a fully homozygous reference, but will require some adaptation for less uniform data (e.g., human samples). Although not accurate in all cases, these assumptions are likely a good approximation in most, and moreover, relieving these assumptions is not sufficiently supported by the currently available data.

Overall the methods I develop here provide a new tool to study dynamic transitions within complex systems, and is broadly applicable across many computational and real-life domains.
Chapter 5

The genome-wide dynamic regulation of RNA splicing and editing in mammalian cells

5.1 Introduction

The conversion of an initial precursor into a functional mature transcript is key to the cellular regulation of RNA levels, as most transcripts go through several precise processing events, including splicing and editing, which separate a transcript’s production from its activity [5, 6].

However, significant technical and computational challenges still limit the availability of genome-wide data on these processes. Next-generation sequencing of total cellular RNA (RNA-seq) allows to study the outcome of RNA splicing at whole transcriptome level [49, 50, 51], and to discover new alternative isoforms of transcripts. However, the kinetics of RNA splicing and processing can hardly be resolved with these techniques. Endogenous rates of RNA splicing reactions in mammalian systems have proven more difficult to measure, and although the kinetics of excision of individual introns was described [19], larger scale analyses are limited by low levels and high turn-over rate of unprocessed transcripts [20]. Several recent works aimed to predict positions in
transcripts which are edited by comparing RNA sequencing data to a reference genome [21, 24, 22, 23]. These attempts are strongly biased by the high error frequency of sequencing methods, which questions their relevance in vivo [26, 27, 28]. Therefore, little is known about the contribution of alterations in RNA processing to gene expression: the rate at which individual splicing and editing events occur, their dynamic regulation and how they are temporally integrated into the observed transcript levels.

To answer these questions, I combine deep sequencing of ribosomal depleted RNA with RNA metabolic labeling at high resolution in both time and sequence space and computational modeling (chapter 4), and study RNA splicing and editing dynamics during the model response of mouse dendritic cells to pathogens. I leverage that unique view to formulate key principles of temporal RNA regulation in mammalian cells.

### 5.2 RNA maturation at whole transcript resolution

Metabolic labeling of RNA by 4-thiouridine (4sU) provides a unique view into the unstable transcriptome. Selectively labeling only the newly produced RNA during a short labeling pulse (see section 3.2) enriches the population with unstable transcripts (figure 5-1a). This bias has two important implications: (1) significant reduction in ribosomal RNA levels (by at least 40%) which allows to sequence whole transcriptome, without prior polyA selection, and (2) enrichment of processing intermediates and precursors of RNAs within the sequenced population. Therefore, the genome wide RNA-4sU sequencing data (see section 3.4.1) although in low temporal resolution, still provides enough information to start monitoring RNA maturation at the transcript level.
Figure 5-1: Monitoring RNA maturation at whole transcript resolution. (a) 4sU-RNA sequencing captures a broader representation of transcripts compared to polyA+ RNA-Seq. Shown is the fraction of reads in RNA-4sU sequencing libraries (left) and polyA+ RNA-Seq libraries (right), across several annotation categories. Only reads that are mapped to a unique location in the genome or to rRNA are considered. (b) Expression profiles (generalized RPKM) of both exons (representing both mature (M) and un-processed (P) transcripts; left) and introns (representing only un-processed (P) transcripts; right) for the 2,122 expressed genes (rows) across 6 time points (columns). Genes are partitioned to 8 expression clusters (I-VIII, right). Purple: high relative expression; white: mean expression; pink: low relative expression. (c) Distribution of predicted constant processing rates for 3,011 genes with exonic and intronic expression during the first 6 hours of the response to LPS stimulation. Dashed lines distinguish 5 quintiles (a-e, 20% increments), and transcripts with >30min half-life are added to the last bin. Pre-mRNA half-lives for illustrative genes are denoted in each bin. (d) Transcripts with low or high pre-mRNA half-lives are enriched in functional categories, clusters, exon structures or transcript lengths.
Figure 5-1: Shown are the enrichments (P-value, hypergeometric test, grey), of the overlap between the genes in each of the half-life bins in (c) (A-E, columns) and each tested category (rows). Only categories with at least one significant enrichment are shown.

5.2.1 Estimating a transcript’s average maturation rate with the single-step processing model

I leverage this unique view of unstable transcripts to monitor mRNA maturation. I use low-resolution RNA-4sU sequencing data and conservatively analyze only the 2,122 genes for which both exons and introns are expressed above a minimal value, excluding genes with expressed anti-sense transcripts or significant alternative isoforms biases (see section 5.6). I estimate an average level of transcript’s precursor and mature levels in RNA-4sU sequencing data by the generalized RPKM approach (see section 4.3.1), which distinguishes intronic and exonic reads. I predict average transcript processing rates by fitting these estimates with a single step transcript processing model (see section 4.2.2), parametrized with a constant processing and a constant degradation function, and a dynamic transcription function (an 'impulse' model). While this single step model is over simplistic in some cases, it is a good approximation for many transcripts, and discovers some basic principles of the response.

5.2.2 Differences in transcript maturation rates between genes contribute to regulation of RNA levels

Generalized RPKM estimates show that expressed introns and exons profiles correlate well (Spearman r=0.81), and have similar patterns in all clusters (figure 5-1b). Most expressed genes (95%, 2,014/2,122) show only a minimal discrepancy between the measurements and the expected values under the constant processing and degradation model ('goodness of fit' test). Of these, 8% (161/2,014) are genes that rejected the simpler constant degradation model (section 2.4.1), suggesting that for these genes, regulation of pre-mRNA maturation is important for shaping final mRNA levels. For
the remaining 5% (108/2,122), one or more of the model's assumptions is inaccurate (e.g., changes in RNA processing rate or in RNA stability), but I cannot determine which one. Predicted constant pre-mRNA half-lives (for 95% of genes) are shorter than the mRNA half-life (21 ± 13% SD of mRNA half-life), and range between 1-30 min (figure 5-1c). Pre-mRNA half-lives significantly correlate with mRNA half-lives (Spearman r=0.46) possibly due to the relationship between the 'constant degradation' and 'constant degradation and processing' models, which makes the processing and degradation rates estimates somewhat dependent.

Differences in constant pre-mRNA half-lives among genes are consistent with exon structure and temporal expression pattern, and suggest that variation in RNA maturation rates between genes serves as a regulatory mechanism (figure 5-1d). The slow processed RNAs are enriched for constantly expressed genes (clusters I), with 1-10 exons, and with short transcripts (1-16 kb unprocessed length). Conversely, fast processed RNAs (group A) are enriched for genes with sharp 'peaked' expression (cluster III), with few (1-4) exons, and surprisingly, with long transcripts (> 60kb unprocessed length) and many exons (> 17). Indeed, this long transcript bias might be a result of the single step processing assumption in the model, which is better fitted for short transcripts.

5.3 RNA splicing dynamics at junction resolution

In most genes, the formation of the mature transcript from its precursor requires several splicing events of individual junctions. Therefore, there is a significant advantage in monitoring distinct splicing events within a transcript and studying the dynamics of RNA splicing at a junction resolution.
5.3.1 Measuring junction-specific precursor and mature transcript abundance from RNA sequencing

The ability to monitor the dynamics of RNA splicing at a junction resolution depends (1) on the ability to measure the response with high resolution in both sequence and time, and (2) on measuring several distinct and complementary RNA populations.

I therefore generated a new high temporal resolution time series data, by following RNA levels during the response of mouse DCs’ to LPS stimulation and taking an RNA sample every 15 minutes for the first 3 hours of the response (13 samples in total). High coverage sequencing of strand specific and ribosomal depleted RNA (figure 5-2a) provides the necessary sequencing depth on exons, introns, and splicing junctions for accurate modeling.

I measure a broad representation of the dynamic transcriptome through deep sequencing of three complementary RNA populations. First, RNA-Total measures all transcripts that exist within the cell at a specific time, regardless of their transcription time (figure 5-2a). Second, I isolate recently made RNA-4sU using metabolic labeling (see section 3.2 and figure 5-2a), which measures only transcripts that were actively transcribed (at least in part) during the short labeling time. RNA-4sU is therefore enriched for short-lived transcripts, including precursors of quickly processed junctions and processing intermediates, and therefore provide necessary information on processing kinetics. Finally, I also measure standard RNA-polyA+ that is enriched for the fully processed and mature transcripts. Indeed, while exons have higher read counts in RNA-Total and RNA-polyA+ data, introns have higher coverage in RNA-4sU data.

I consider each splicing junction as an independent processing event, and use a binomial model (see section 4.3.2) to estimate, per junction, the precursor and mature RNA levels. I further enhance this analysis beyond constitutive exons and distinguish alternative mature isoforms. The predicted mature RNA levels are indeed enriched in RNA-total and RNA-polyA+ data, while precursor levels are enriched in RNA-4sU data (figure 5-2b).
Two examples highlight the utility of this approach. The Zfp36 gene has a single isoform with one splice junction (figure 5-2c). The binomial model predictions show the expected peak in mature transcript levels at 60 min. post stimulation, while precursor levels peak about 30 min. earlier. The RNA-4sU sequencing data includes two notable differences: (1) higher precursor levels and (2) a peak in mature RNA that is 15 min. earlier than in total-RNA. The Mbnl1 gene requires 8 splicing events to express two alternative isoforms in DCs (figure 5-2d). The binomial model predicts a similar dynamics for all splicing events, with lower expression of the alternative transcript, which fits sequencing counts. Notably, while the mature RNA shows a temporally constant dynamic, the precursor levels drop significantly 60 min. post stimulation.

5.3.2 The junction-specific temporal response of mouse dendritic cells to LPS stimulus

With this dataset of genome-scale and high-resolution measurements I study the dynamic response of mouse DCs at a per-junction resolution. I annotate the DCs’ transcriptome using standard polyA+ sequencing data (see section 5.6), including all expressed transcripts and splicing junctions. Notably, this annotation technique do not account for (1) transcripts with polyA- mature isoforms (e.g. histone genes) and (2) the precursor’s full length. I identify 11,444 expressed genes with 115,341 splicing junctions. Most junctions (96%) express a single isoform, and 4% have up to 6 alternative isoforms, corresponding to 30% of genes with alternative full-length transcripts. Alternative isoforms are mainly of 7 standard types (figure 5-3a), and are in line with previous measurements of alternative splicing in a collection of tissues [50]. I selected a subset of significantly expressed junctions with mapped reads above a minimal value (see section 5.6), resulting in a set of 30,181 junctions (26.1%) from
Figure 5-2: Junction specific precursor and mature transcript abundance from RNA sequencing. (a) After stimulation, I add 4sU (red squares) to growing cells for a pre-defined time, collect the cells and extract total RNA (blue). I use ribosomal depletion methods (Ribo-Zero) to deplete ribosomal RNA from RNA samples. Using biotin capture with streptavidin magnetic beads, I purify labeled RNA (red) from the ribosomal RNA depleted RNA extract.
4,996 genes (43.7%). These are enriched for immune genes and transcription factors that are important for the specific LPS response, as well as for RNA processing factors. Only 498 of significantly expressed junctions (1.7%) have annotated alternative isoforms, corresponding to 443 of expressed genes (8.9%) with alternative transcripts (figure 5-3a).

I applied the binomial model to the dynamic sequencing data of the 30,181 significantly expressed junctions based on these annotations, and predicted precursor and mature RNA levels at single splice junction resolution and in 15 minutes temporal intervals.

Indeed, the temporal expression of separate junctions on a single transcript is highly correlated, with a higher similarity between the mature junctions, and if the junctions are adjacent (figure 5-3b). Alternative mature isoforms, however, are comparable to separate transcripts in their expression levels similarity (mature RNA only), as these are indeed separate transcripts.
Figure 5-3: Expressed junction during the LPS response. (a) In 4% of the expressed junctions there are evidence for alternative splicing. Shown are the 7 most common types of alternative splicing evident in the data (left). Middle column states the number of junctions out of all expressed junctions that match each type of alternative splicing, and the right column states the number out of 'significantly expressed' junctions. (b) Distribution of the pairwise distances between the temporal expression vectors of either precursor (dashed lines) or mature (solid lines) RNA levels of pairs of junctions. Junction pairs are either consecutive on the same transcript (red), further apart on the same transcript (green), two isoforms of the same junction (blue) and all other pairs of junctions from different transcripts (purple).

5.3.3 Dynamic changes in a junction’s mature RNA levels are preceded by highly correlated changes in precursor and newly produced RNA levels

Predictions from the binomial model provides a unique perspective into the genome-wide dynamics of responding DCs at single splicing junction resolution, and provide the necessary data to answer several important questions, including how is the tem-
poral expression of a single junction regulated? does the processing of a junction contributes to shaping its mature RNA expression? and how prominent it is relative to its production and degradation rates? I therefore set to study the junction specific dynamic regulation, and answer some of these questions.

Of the 30,181 significantly expressed junctions, most (86%) are differentially expressed (figure 5-4a), and are either up (41%) or down (45%) regulated, while 14%, although expressed at significant levels, are not differentially regulated (less than 2-fold change in expression) during the first 3 hours of the LPS response. Indeed, up-regulated junctions are enriched with immune genes including inflammatory, anti viral and anti bacterial genes, for transcription and signaling molecules and for developmental and differentiation processes (including DCs maturation factors). On the other hand, down regulated junctions are enriched with housekeeping genes and basal cellular functions (e.g., oxidative phosphorylation) and for energy metabolism. Surprisingly, non differentially expressed junctions are enriched with spliceosome components.

Based on the mature RNA expression levels in RNA-Total (which are parallel to standard polyA+ measurements) I identify several distinct temporal behaviors, both within the up- and within the down-regulated genes (figure 5-4a). These dynamic expression patterns match functionality. For example, cluster 5 is enriched with inflammatory and bacterial response genes (including cytokines, chemokines and Tnf signaling genes) as well as for apoptosis and cell development genes, while clusters 7 and 10 are enriched with viral and inflammatory response genes (including Nfkb targets), and cluster 9 is enriched with interferon responsive genes. Conversely, cluster 23 is enriched with basal cellular functions (including housekeeping genes, and oxidative phosphorylation genes).

I next asked how the precursor dynamics relates to that of mature RNA and of new RNA (RNA-4sU). In principle, based on simulation, the temporal shift between precursor and mature RNA levels is determined both by the processing and by the degradation rates (figure 5-4b), and allows to estimate their contribution to the observed dynamics. A high turn-over rate (of either precursor or mature) minimize the regulatory effect on the final RNA levels, while a lower rate allows integration in
A high degradation rate (short mature RNA half life) results in a minimal shift between precursor and mature RNA profiles, while a high processing rate (short precursor half life) leads to a minimal shift between RNA-4sU and precursor RNA-Total profiles (figure 5-4b). I therefore predict a temporal shift between time courses using time lagged correlation, and compare precursor and mature RNA as well as Total and 4sU-tagged RNA of each junction.

I find that dynamic changes in a junction’s mature RNA levels are preceded by highly correlated changes in precursor levels. Precursor levels precede by a typical time-gap of 15-30 minutes (figure 5-4c), corresponding to slower degradation rates and longer mature half lives (in the order of tens of minutes). Although RNA-4sU and RNA-Total are also highly correlated (figure 5-4d), the time-gap between them is significantly shorter (mostly no shift), indicative of very fast processing rate and short precursor half life. These results reinstate RNA production as a major regulator of the dynamic response, and find, on a global view, that RNA degradation has a bigger regulatory role than that of RNA processing.

Finally, I note several limitations of the current analysis. First, a global view disregards local effects that in many cases represent significant effects. Identifying groups of genes with a significantly longer time gap and lower correlation between precursor, mature and new RNA, elucidate these effects. Moreover, using Pearson correlation as a measure of similarity between time courses, as I do here, might be less appropriate. Indeed, Pearson correlation is not taking into account the temporal nature of the data in comparing two time courses. Therefore, a distance measure that is tailored specifically for time courses, such as [57, 46], or methods that identify temporal differential expression between time courses [82] can be better suited to answer the specific questions I am interested in. Finally, even though the correlation is more prominent between precursors, the mature RNA is also mostly not shifted between RNA-4sU and RNA-Total, contradicting the predicted longer mature half lives, and suggesting some level of contamination with RNA-Total within the measured RNA-4sU population, which I am working on controlling for computationally.

Indeed, to reach more rigorous and systematic conclusions about dynamic regulation
at the junction resolution, I turn to dynamic modeling.

5.3.4 Differences in processing rates between junctions contribute to shaping dynamic RNA levels

I use a single step processing model (see section 4.2.1) to model the temporal kinetics of a junction, and unfold the relative contributions of its production, processing and degradation to the observed dynamics. Indeed this model is well suited for looking at single junctions processing.

I test several hypotheses on the dynamics of RNA processing and degradation, allowing either one to be constant or dynamic (an impulse model). To select between these non-nested models, I use a 'likelihood ratio test' to compare the dynamic and constant processing rate models using either a constant or a dynamic degradation model as an underlying assumption. And similarly, I select between a constant and dynamic degradation rate with either processing model.

For the vast majority of junctions (93%), a constant processing rate is sufficient to explain the data with either degradation function (constant or dynamic) and is consistent between the two degradation functions (Pearson $r = 0.98$ between predicted processing rates with either constant and dynamic degradation). Only a small minority of junctions (3%) reject the 'constant processing' hypothesis with either degradation function (figure 5-5a). But, this minority group is enriched for immune genes and transcriptional regulators that are major players in the LPS response, suggesting that dynamic regulation of processing rates uniquely contributes to the observed immune response. Processing rates correspond to precursor half-lives that range from less than a minute to over 30 minutes. Constant rates are mostly distributed within the 1-30 minutes range (figure 5-5b), while extreme values (both slow and fast) are more abundant within the dynamic rates.

A similar analysis of degradation rates identifies 84% of junctions which retain a con-
**Figure 5-4:** The junction-specific temporal response of mouse DCs to LPS stimuli. (a) Clusters of mature RNA-Total expression for all 30,181 'significantly expressed' junctions (rows) measurements across 13 time points (columns),
Figure 5-4: including data for RNA-Total (Precursor and Mature; left) and RNA-4sU (Precursor and Mature; right). Cluster numbers (1-24) are noted on right; up- (red; left) and down- (green; left) regulated junctions are divided into 23 clusters, and cluster 24 includes the non-differentially expressed junctions (gray; left). Red: high relative expression; white: mean expression; blue: low relative expression. (b) Simulated data showing expected RNA-4sU (dashed) and RNA-Total (solid) measurements of precursor (blue) and mature (red) RNA with the pre-defined production function (left, green) and the stated (top) precursor and mature RNA half lives (corresponding to processing and degradation rates respectively). (c) Time shifted correlations between precursor and mature RNA. Top: the median correlation (y-axis) across all genes using the specified time-shift (x-axis) between precursor and mature for either RNA-Total (black) or RNA-4sU (bottom). Bottom: distribution of the optimal time-shift between precursor and mature (x-axis) per gene (y-axis, fraction of genes) for either RNA-Total (black) or RNA-4sU (gray). (d) Time shifted correlations between RNA-4sU and RNA-Total. Top: the median correlation (y-axis) across all genes using the specified time-shift (x-axis) between RNA-Total and RNA-4sU for either precursor (blue) or mature (red). Bottom: distribution of the optimal time-shift between RNA-Total and RNA-4sU (x-axis) per gene (y-axis, fraction of genes) for either precursor (blue) or mature (red).

stant degradation assumption, while 12% of junctions reject it in favor of a dynamic degradation rate (figure 5-5c). Degradation rates correspond to mature RNA half lives that range between single minutes to few hours (figure 5-5d), and are indeed slower than predicted processing rates, as suggested by the time shift analysis. Indeed, it is not surprising that with a 15 minute time resolution data I detect more dynamic changes in the slower degradation rates than in processing rates. Finally, I note that I did not detect any global relationship between a junction’s precursor and mature RNA half lives (figure 5-5e).

Variation in processing and degradation rates are associated with junctions’ expression pattern (figure 5-6a). Non differentially expressed genes (cluster 24) are enriched for slow degradation rates and fast processing rates, suggesting these junctions are processed quickly and maintain the mature product for a longer time. Clusters 6-10 of junctions that are stably expressed for several hours are enriched for slow processing and slow degradation, while clusters 4-5 of transiently induced junctions are enriched for dynamically changing degradation rates and, in cluster 5, also dynamically changing processing rates.
Figure 5-5: Dynamic changes in a junction’s precursor and mature RNA levels. (a) Competing hypotheses on processing rate dynamics. I compared constant and dynamic processing either under a constant degradation assumption (rows) or a dynamic degradation assumption (columns) using maximum likelihood test. Numbers in cells represent overlaps in junctions between the two tests. (b) Distribution of predicted precursor RNA half-lives for the 28,633 junctions with predicted constant processing (black) or for the 929 junctions with predicted dynamic processing (red), evaluated every 10 minutes over 3 hours. (c) Competing hypotheses on degradation rate dynamics. I compared constant and dynamic degradation either under a constant processing assumption (rows)
5.3.5 A junction's exon and intron sequence composition is associated with differences in processing rate

To gain insights into the mechanisms that control processing rates, I next searched for over-represented sequences (1-4 nucleotides long) across all junctions in the dataset. I compared the distribution of sequence frequencies within junctions of each quantile to the overall genomic distribution, separately for exons (compared to exon background) or introns (compared to intron background) using a Kolmogorov-Smirnov (KS) p-value. Over or under represented sequences can relate to specific protein binding sites, but given the low sequence specificity of RNA binding proteins, these are more likely to represent general sequence properties that are recognized by splicing factors with a "fuzzy" sequence preference.

I find a distinct set of over-expressed sequences uniquely enriched either within junctions processed at a mid-distribution rate (quintiles 2-6), or within junctions with extreme constant rates (quintiles 1 and 9-10) and dynamically regulated rates (which are indeed more extreme) (figure 5-6b). The composition of over-represented intronic sequences is bimodal: AU-rich sequences are enriched in mid-distribution junctions, while GC-rich sequences are enriched in junctions processed at extreme (very fast or very slow) or dynamic rates (figure 5-6c). Indeed, intronic sequences all over the genome are AU-rich compared to other genomic regions (including exons) [83], and the GC-richness of extremely processed junctions is relative to that background. The composition of exonic over-represented sequences, on the other hand, changes gradually with processing rates: it shifts from GC-rich in fast-processed and dynamic junctions, through CU- then AU-rich mid distribution and finally AG-rich in slowly processed junctions.
processed junctions. To further validate these enriched sequences, it will be important to also test their conservation across species.

Notably, even though I compare here sequence composition between introns, it is possible that the extreme quantiles contain more junctions with annotation errors (wrongly classifying exons as introns) thus predicting low processing rates and observing GC enrichment for these junctions. Although annotation errors are hard to avoid completely, using available de-novo transcript reconstruction algorithms could help to minimize such errors.

5.3.6 Spatial coordination of junction processing within a transcript

Another interesting question is how the processing of different junctions on a single transcript relate to each other? and how these local rates integrate and shape the expression of the entire transcript? are all junctions on one transcript processed at the same rate? or is it that downstream junctions are processed faster than upstream junctions or vice versa?

To get insights into these questions, I associated the different processing quantiles with distinct junction’s structural characteristics. Indeed, dynamically processed and slowly processed junctions (quantiles 1-2) are more commonly alternatively spliced, while faster processed junctions (quantiles 4-10) mostly express a single alternative form. Extreme-rate (quantiles 1 and 9-10) and dynamically processed junctions are enriched with short introns that are located near the 5’ end of transcripts with overall fewer junctions (figure 5-6d). Mid-range junctions (quantiles 4-8) on the other hand, are enriched with long introns that are located towards the 3’ end of longer transcripts with many junctions.

Of the 4,996 transcripts in the 'significantly expressed' set, 1,534 have more than 80% of their junctions within this set, including 1,425 transcripts (29%) with more
**Figure 5-6:** Enrichments of differential splicing rates between genes. I divided the junctions with a predicted constant processing rate into 10 deciles based on their predicted precursor half life. Other 3 groups are junctions with a dynamic processing rate, and a constant or dynamic processing rate that depends on the degradation rate function (constant/dynamic and dynamic/constant). I search for specific enrichments within each of the groups. (a) Grayscale matrices represent enrichments per decile for each of the 24 clusters as in figure 5-4a). Red and blue matrices represent a one sided KS test for distribution of processing and degradation rate within each cluster (blue = enriched with long precursor half life, red = enriched with short precursor half life). (b) Over represented 4-mers either in the intron of junctions (vs. other introns in the genome) or the exons of junctions (vs. other exons in the genome). Shown is the number of overlapping 4-mers between each decile, and the dynamic junctions (red: high overlap, white: no overlap). (c) Nucleotide frequencies (x-axis) within enriched 4-mers in exons (left) and introns (right) for each of the 10 deciles and junctions with predicted dynamic rate (y-axis).
than one splicing junction. For each gene, I use a linear regression model to fit the junctions’ positions and their processing rates. Indeed, the first junction of genes with a negative slope (processing is faster towards the 5’ end) commonly fall within the fast processing quantiles, while the first junction of genes with a positive slope (processing is faster towards the 3’ end) fall into the slow processing quantiles. Within this set of 1,425 transcripts, most have a minimal slope (absolute value < 0.5), but in 6% of the transcripts I find a more prominent negative slope and in about twice as many (12%) a more prominent positive slope, suggesting that an increase in degradation rate from 5’ to 3’ of a transcript is more common than a decrease.

5.4 RNA editing sites in mouse dendritic cells

I take advantage of 4sU metabolic labeling to look at another type of transcript processing through RNA editing [7], by comparing sequencing errors between RNA-4sU and RNA-total sequences (see section 4.5).

5.4.1 Few transcripts are edited in mouse DCs during the LPS response

I identify 70 positions in which the sequencing data support editing (figure 5-7a), significantly less than estimations in human [24, 22, 23], but supported by other works that predict low RNA editing levels in mouse in the absence of the primate specific Alu repeats [21]. Only 53 of these fall within the boundaries of 32 known transcripts, and indeed 11 transcripts contain within them more than one and up to 6 predicted editing site. The gene Arhgapi2 has the highest number of 6 editing sites, all located within its introns.

Most changes I identify are editing by deamination: either A to I(G) changes (54%),
which are catalyzed by the enzyme Adenine deaminase (ADAR), or C to U(T) changes (9%) that are catalyzed by Cytidine deaminase. Notably, the other two frequently detected changes are their reverse complement (i.e., T to C and G to A) and are likely be a result of strand specificity biases in the sequencing data (figure 5-7a). The other 9 editing sites are more surprising changes. I also identify a significant amount of changes (340 genomic positions) in the other direction, namely the RNA-Total base rather than the RNA-4sU base fits the reference genome. Indeed these are almost exclusively C to T mutations (99%), a known error of sequencing RNA that contains 4sU.

5.4.2 Predicted editing events are located at non-coding introns and 3’UTRs, suggesting that they contribute to post-transcriptional regulation of expression

Surprisingly, predicted editing sites do not fall within coding regions of genes (figure 5-7b), but are most commonly found in 3’UTR (39%) and introns (37%), with a few in inter-genic regions (24%), which probably contain non annotated transcripts. Moreover, in all 11 transcripts that contain more than one editing site, all sites fall within the same type of non-coding region (all intronic in 7 transcripts, and all 3’UTR in 4 transcripts). The predicted editing sites are also enriched for forming stem-loop structures with an upstream sequence ($p < 5.7e^{-12}$), but not with a downstream sequence, which support editing by the ADAR enzyme that recognizes its targets through a dsRNA binding site.

I compare these predictions to recent work that validated a set of 289 editing sites in the mouse brain [23]. Of the 289 genes, only 50% (144) are in transcripts that are also expressed in mouse DCs. Indeed, these are found primarily in introns (67%) and 3’UTRs (29%) and only 5% (7 sites) are in coding regions. Since this set of editing sites was validated in a distinct tissue (brain), it provides no support for these sites being edited at all in mouse DCs, and indeed, only one site was overlapping with these predictions.
Figure 5-7: A spatial coordination of junction processing within a transcript. (a) The nucleotide changes at 70 genomic positions, which are predicted as editing sites (top, light gray). Y-axis labels indicate the genomic base, the RNA-4sU base and the RNA-Total base, in this order. X-axis represents fraction out of all sites, and numbers are the actual counts of each type of change. The nucleotide content of 340 genomic positions with a significant amount of 4sU induced mutations (bottom, dark gray). These differ from RNA-Editing in that the RNA-Total base rather than the 4sU-RNA base fits the reference genome. (b) Location of predicted editing sites - either at 3’UTR (yellow), exon (blue), intron (green) or intergenic (red) regions, for sites that change in RNA-Total (70 editing sites, left), sites that change in RNA-4sU (340 sites, middle) or validated sites based on Danecek et. al.

5.5 Summary

In this work I combine deep sequencing of several RNA populations, including polyA+, total and recently-made RNA with computational modeling to study the regulation of RNA splicing and editing in the response of mouse DCs to Lipopolysaccharide (LPS). Metabolic labeling of RNA with 4sU allows to isolate recently made RNA, which is enriched for short-lived transcripts including RNA precursors and processing intermediates, and leverage that unique view to study the genome-wide kinetics and dynamics at which transcripts mature and individual splicing and editing events occur. I measure RNA maturation at whole transcript resolution from low resolution sequencing data, and estimate RNA splicing dynamics at per junction resolution from ribosomal depleted high resolution (in both time and sequence space) sequencing.

A new computational scheme estimates precursor and mature RNA levels from se-
quencing counts both per-transcript and per-junction, and combines those with a
dynamic model to predict transcripts' overall maturation rates, and splice junction
specific processing rates on a genome scale. Local differences in base composition
between recently produced and overall RNA allows to identify editing events.
I discover several key principles of temporal RNA regulation in mammalian cells.
Changes in mature expression are preceded by correlated changes in precursor and
newly produced RNA, supporting a prominent role for RNA production, and sug-
gest that RNA decay rates have an overall stronger contribution to shaping RNA
dynamics than that of RNA processing rates. While most junctions (95%) are spliced
at a temporally constant rate, a dynamically changing splicing rate is important in
specific cases. Differences in RNA processing rates between genes and between indi-
vidual splicing junctions are coordinated with expression and function, and highlight
the contribution of RNA processing for regulating changes in expression between
genes. Coordination of processing rates with junctions' structure provide mechanistic
insights into splicing regulation. Predicted editing events throughout the expressed
transcriptome are mostly located at non-coding 3’UTRs, and most likely contribute
to post-transcriptional regulation of expression.
This analysis highlights some promising directions for further investigation. First,
using similarity measurements oriented towards temporal data can further highlight
differences between mature, precursor and newly produced RNA, and identify more
subtle effects of RNA processing and degradation control. Improving the existing
annotations by employing de-novo and ab-initio transcript reconstruction methods
rather than relying on existing Refseq annotations can remove many of the current
annotation biases, and identify more prominent roles for alternative splicing. Studying
the conservation of enriched sequences between species can provide further support
for the effects I find in mouse. Finally, applying the multi-step processing models
to the data will help to differentiate between competing hypothesis on the order in
which separate junctions on one transcript are spliced, and provide insights into a
highly debated question.
Overall, this work provides new and effective tools to study RNA maturation including
splicing and editing at both transcript and per junction resolution, and model its
interactions with other key cellular regulatory processes, generating a complementary
view to any RNA expression analysis, and deepening our understanding of the RNA
life cycle.

5.6 Methods

RNA sample collection from mouse DCs. All animal protocols were reviewed and
approved by the MIT / Whitehead Institute / Broad Institute Committee on Animal Care
(CAC protocol 0609-058-12). DCs isolation and treatment, 4sU labeling and RNA isolation
were done as described in section 3.6, with the following modifications. For the high-
resolution analysis, I extracted RNA starting at 0h and until 3h after LPS stimulation in
15 minutes intervals. 4-thiouridine (Sigma) was added to a 500 $\mu$M final concentration 10
minutes before RNA collection. For the lower resolution analysis, I extracted RNA starting
0h and until 6h after LPS stimulation in 60 minutes intervals, and 4-thiouridine (Sigma)
was added to a 150 $\mu$M final concentration 10 minutes before RNA collection. Total RNA
was extracted with the miRNeasy kit’s procedure (Qiagen), and sample quality was tested
on a 2100 Bioanalyzer (Agilent). I used 10 $\mu$g total RNA and removed rRNA with the
RiboZero kit (Epicenter). I saved an aliquot of 100 ng ribosomal depleted total RNA, and
followed with 4sU purification for the reminder of the sample.

Preparation and processing of RNA-Seq libraries. The small scale sequencing data
was generated as described in 3.6. For the large scale sequencing data, I prepared the
RNA-Seq libraries using the 'dUTP second strand (strand specific)' protocol as described
in [30], with the following modifications: after DNase treatment, I used the entire RNA
sample (without polyA-RNA isolation) and fragmented RNA by incubation in 1X RNA
fragmentation buffer (Affymetrix) at 80 degrees for 4 minutes. I sequenced all 26 samples
using Illumina sequencing, and generated $1.3 \cdot 10^9$ paired-end reads ($1.3 \cdot 10^9$ million single
reads, 101 bases long), with 80 – 200 million reads per sample.

Read Alignment. I aligned all reads to the mouse reference genome (NCBI 37, MM9)
using the TopHat aligner [49] with default parameters. To estimate rRNA levels, I mapped
all reads (with the same procedure) to mouse rRNA sequences from GenBank [66]. About 36% of RNA-total reads were mapped to rRNA, and only 34% uniquely mapped to the genome. In RNA-4sU data, only 7% of the reads were mapped to rRNA, and 60% were uniquely mapped to the genome. Indeed, the difference in rRNA levels accounts for the difference in uniquely mapped reads.

**Selecting expressed precursors from low-resolution 4sU-Seq data.** I focused on genes with significantly expressed introns and exons (RPKM $> 2$) for at least one time point (3,011/10,106; 30% of expressed genes). Approximately 27% (2,737/10,106) of the Refseq genes express only one major transcript, and over 75% express less than 5 (between 1-4) transcripts. I also find evidence for antisense transcription in 3% (268/10,106) of the RefSeq genes. I therefore excluded transcripts with antisense transcription or significant alternative isoforms biases, and is left with 2,122 genes.

**Transcriptome reconstruction.** I used Trinity de novo transcript assembly [84] and Cufflinks ab-initio transcript reconstruction [85] softwares to reconstruct mature transcripts from polyA+ sequencing data (low resolution). I intersected the predictions with known mouse gene annotations from the UCSC genome browser [67] (Refseq genes, UCSC genes and microRNA and tRNA) to select the set of expressed transcripts and isoforms.

**Selecting significantly expressed junctions.** I select a set of significantly expressed junctions by considering the overall reads that mapped to a junction (over time) and selected only junctions with counts above cutoff. I summed normalized read counts (RPKM) for each junction at all times in either RNA-Total or RNA-4sU data, and required that in at least one of the two this sum is more than 5 times the median junction's normalized count.

**Functional Enrichments.** I calculated the enrichment of a query set of genes X for an annotation A using hypergeometric p-value between two groups [72], and a 1% False Discovery Rate (FDR) significance threshold, across all annotations I tested. Functional annotations of the mouse genome were taken from the Molecular Signature Database (MSigDB) [73], c2 (curated gene sets), c3 (motif gene sets) and c5 (GO gene sets).
Chapter 6

Molecular mechanisms of RNA regulation

6.1 Introduction

The ability to measure transcription, processing and degradation during a complex temporal response provides a unique framework to start deciphering the underlying molecular mechanisms that control variations in these rates between genes or time points (for one gene). While as a first approximation, I considered each expressed gene independently and ignored the interactions between their products, at the next step I can add to the system's description the implementation of the regulatory circuit, through molecular interactions between cellular components (binding motifs and active regulators). As components' activity and their interactions change in response to both external signals and internal interactions, they give rise to the observed complex dynamic behavior.

Here I collected perturbation data, by monitoring RNA levels when certain components in the network are disabled, either by knock-out when available, or knock-down with short hairpin RNA (shRNA). Using this data I infer the network structure and the temporal regulators' activity levels, which provide insights on the molecular mechanisms that lie behind the observed dynamic changes in RNA regulation.
6.2 A molecular model for the regulation of RNA degradation

6.2.1 The kinetics of an RNA degradation protein regulator

The kinetic model in figure 6-1a describes the activity of a protein regulator $R$ (commonly an RNA binding protein) which is a degradation activator of gene $X$. This
model includes both a factor-mediated degradation from the bound state \((XR)\) and a basal degradation from the unbound stat \((X)\), which happen at a different rate. The change in RNA concentration over time is defined by:

\[
\frac{d[X_T]}{dt} = \alpha(t) - k_1(t)[X] - k_2(t)[XR]
\]

where \(\alpha\) is the production rate, \(k_1\) is the basal degradation rate and \(k_2\) is the maximal degradation rate achieved by the factor \(R\).

In order to study this dynamic reaction, I start by making the assumption that the RNA exists in one of two states - free or bound by the regulator, therefore \([X] + [XR] = [X_T]\). Using this assumption, I get that:

\[
\frac{d[X_T]}{dt} = \alpha(t) - k_1(t)[X] - k_2(t)[XR] = \alpha(t) - k_1(t)([X_T] - [XR]) - k_2(t)[XR] = \alpha(t) - k_1(t)[X_T] - (k_2(t) - k_1(t))[XR]
\]

**Michaelis-Menten kinetics assumption.** A Michaelis-Menten kinetics means that the concentration of RNA bound by the regulator is at quasi steady state, and therefore \(\frac{d[X_R]}{dt} = 0\), resulting in the following estimation of occupied sites:

\[
\frac{d[X_R]}{dt} = k_b[X][R] - k_d[XR] - k_2[XR] = 0
\]

\[
k_b[X][R] = (k_d + k_2)[XR]
\]

\[
\frac{k_d + k_2}{k_b} = \frac{[X][R]}{[XR]} = \frac{[R]([X_T] - [XR])}{[XR]} = \frac{[X_T][R]}{[XR]} - [R]
\]

\[
[XR] = [X_T] \cdot \frac{[R]}{k_d + k_2} + [R]
\]

**Hill coefficient kinetics assumption.** An alternative to the Michaelis-Menten kinetics, is to estimate the fraction of occupied sites by the Hill coefficient \(n\):

\[
[XR] = [X_T] \cdot \frac{[R]^n}{\left(\frac{k_d + k_2}{k_b}\right)^n + [R]^n}
\]
The Hill coefficient \( n \) is related to cooperativity: \( n = 1 \) no cooperativity, \( n < 1 \) negative cooperativity, \( n > 1 \) positive cooperativity.

Finally, taking the previous equation with either one of the assumptions, gives:

\[
\frac{d[X^T]}{dt} = \alpha - k_1 \cdot [X^T] - (k_2 - k_1) \cdot \frac{[R]^n}{\left(\frac{k_d+k_2}{k_b}\right)^n + [R]^n} \cdot [X^T]
\]

\[
= \alpha - \left(k_1 + (k_2 - k_1) \cdot \frac{[R]^n}{\left(\frac{k_d+k_2}{k_b}\right)^n + [R]^n}\right) \cdot [X^T]
\]

Which converges to a Michaelis-Menten kinetics when \( n = 1 \).

By defining: \( \beta_1 = k_1, \beta_2 = k_2 - k_1 \) and \( K = \frac{k_d+k_2}{k_b} \), the degradation term becomes:

\[
\beta(t) = k_1 + (k_2 - k_1) \cdot \frac{[R]^n}{\left(\frac{k_d+k_2}{k_b}\right)^n + [R]^n}
\]

\[
= \beta_1 + \beta_2 \cdot \frac{[R]^n}{K^n + [R]^n}
\]

I parameterize the model with 9 transcript-specific and 6 global parameters (figure 6-1b):

\[
\alpha(t) = \text{Impulse}(t; \theta_\alpha)
\]

\[
\beta(t) = \beta_1 + \beta_2 \cdot \frac{[R]^n}{K^n + [R]^n}
\]

\[
R(t) = \text{Impulse}(t; \theta_R)
\]

I use an 'impulse model' with 6 parameters [46, 57] to describe the dynamics of both a transcript specific production function (\( \alpha \)) and a global concentration of free regulator ([\( R \)]) that is not bound to any transcript. The degradation dynamics (\( \beta \)) follows from the factor \( R \) kinetics and 3 transcript specific parameters \( \beta_1, \beta_2, K \). The Michaelis-Menten constant \( K \) is the concentration of the regulator \( R \) that is required for 1/2-maximal degradation of a transcript, and relates to the binding motif activity. The parameters \( \beta_1, \beta_2 \) describe the basal degradation rate with and without the regulator.
If \( \beta_2 \to 0, k_b \to 0 \) or \( k_d \to \infty \) than the degradation function converges to \( \beta(t) = \beta_1 \), which means that the transcript is not regulated by \( R \).

6.2.2 A model of RNA dynamics in a perturbational study

For a gene \( X \), let \( X(T) \) be its RNA-Total levels at time \( T \) after induction, and let \( X^*(T; t_L, d_L) \) be the gene \( X \)'s RNA-4sU levels at time \( T \) with a short \( d_L \) labeling pulse. As before (section 2.2.2), the dynamics of these two measurements is:

\[
X(T) = X_0 + \int_0^T \alpha(t) - \beta(t)X(t)dt \\
= \int_0^T \alpha(t) - \left( \beta_1 + \beta_2 \cdot \frac{R(t)}{K + R(t)} \right) X(t)dt
\]

\[
X^*(T; t_L, d_L) = \begin{cases} 
\int_{t_L}^T \alpha(t)dt & T > t_L \\
0 & T \leq t_L 
\end{cases}
\]

To these I now add \( X_R(T) \) and \( X_R^*(T; t_L, d_L) \), the RNA-Total and RNA-4sU of the gene \( X \) when the regulator \( R \) is disabled. Under these conditions, \( \beta_2 = 0, K = 0 \) as the regulator is inactive, and therefore:

\[
X_R(T) = X_0 + \int_0^T \alpha(t) - \beta_1X(t)dt
\]

\[
X_R^*(T; t_L, d_L) = \begin{cases} 
\int_{t_L}^T \alpha(t)dt & T > t_L \\
0 & T \leq t_L 
\end{cases}
\]

Notably, this derivation assumes that the regulator's activity is completely blocked, which is likely the case when using a knock-out mice. However, other perturbation techniques, such as gene knock-down with shRNAs, only reduce the regulator activity rather than block it completely, and therefore with such data this assumption will need to be relaxed accordingly.
6.2.3 Likelihood ratio test to select between competing hypotheses for a degradation regulator

I fit two alternative models to the data:

<table>
<thead>
<tr>
<th>Model</th>
<th>Production</th>
<th>Degradation</th>
<th>Gene-specific parameters</th>
<th>Global parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_{NaN}$</td>
<td>$\alpha(t) = \text{Impulse}(t; \theta_\alpha)$</td>
<td>$\beta(t) = \beta_1$</td>
<td>$6 + 1 = 7$</td>
<td>0</td>
</tr>
<tr>
<td>$\theta_R$</td>
<td>$\alpha(t) = \text{Impulse}(t; \theta_\alpha)$</td>
<td>$\beta(t) = \beta_1 + \beta_2 \cdot \frac{R(t)}{K+R(t)}$</td>
<td>$6 + 3 = 9$</td>
<td>6</td>
</tr>
</tbody>
</table>

The simpler '$\theta_{NaN}$' model assumes a constant degradation rate, and no activity of the regulator $R$, while the more complex '$\theta_R$' model assumes that the regulator $R$ is active, and can control the degradation of some genes, allowing it to change dynamically over time. I compare these two nested alternative models ($\Theta_{NaN}$ is nested in $\Theta_R$) by the likelihood ratio test (see section 2.4.1), separately per gene, and identify genes where the regulator $R$ contributes to their dynamic expression.

6.2.4 Inferring dynamic rates from measurements

Available data contains 4 types of temporal measurements: RNA-Total and RNA-4sU levels of a gene, either with an active regulator or with a disabled regulator. I can easily fit the simpler '$\theta_{NaN}$' model using maximal likelihood optimization by gradient descent (as was done in section 2.3.2).

However, fitting the more complex '$\theta_R$' model requires certain adaptations to allow fitting both gene-specific and global parameters. I therefore use the following iterative optimization approach:

foreach gene $x \in X$:

Initialize gene specific parameters $\theta_x$ from perturbation data ($R$ is fully disabled)

loop until convergence:
Estimate global parameters $\theta_R$ | gene specific parameters $\theta_x$
oreach gene $x \in X$:
Estimate gene specific parameters $\theta_x$ | global parameters $\theta_R$

6.3 The RNA binding protein Zfp36 regulates RNA stability in mouse DCs

6.3.1 The RNA binding protein Zfp36

The RNA binding protein Zfp36 (TTP) is a known regulator of RNA degradation, and one of the best-characterized AU-rich element (ARE) binding protein. After binding to AREs, Zfp36 initiates and facilitates the assembly of the mRNA degradation machinery on the mRNA thereby increasing their degradation rate (i.e., de-stabilizing the transcript) and leading to their elimination. In addition to regulating its own stability (auto-regulation), Zfp36 regulates several known targets, including important players in the LPS response (e.g., Tnf, Socs3, Cxcl1, Cxcl2, Nfkbia, Il1a and Il1b. The core Zfp36 binding site is represented by an UAUUUAU ARE heptamers, which is located on the 3’UTR of its targets [86]. As the number of such heptamers increases, the regulatory effect is stronger. There is evidence that ARE-mediated mRNA stability is increased by cooperativity and combinatorial effects of ARE-binding proteins, as well as cooperation of ARE-binding proteins and microRNAs.

Several lines of evidence suggest that Zfp36 may play an important role in regulating changes in degradation rates in this system. First, Zfp36 is known to auto-regulate its own stability, consistent with the increased degradation rate, which the model predicted for it (section 3.3.4). Second, Zfp36’s most well-established target, Tnf [87], also shows a similar increase in degradation rate. Third, the same signaling pathways that activate the transcriptional machinery are known to also induce Zfp36 and other proteins that regulate RNA degradation [88, 86]. Therefore, I decided to
Figure 6-2: The RNA binding protein Zfp36 regulates RNA stability in mouse DCs. (a) Distribution of Zfp36 binding heptamer within 267 representative genes. Shown is the number of genes (y-axis) per number of AREs (x-axis) in 3'UTR (red), 5'UTR (green), CDS (blue) or introns (black, dashed). (b) Change in predicted constant degradation rate between WT and ΔZfp36 cells. Shown is the degradation rate predicted by the constant degradation model either in WT cells (x-axis, log scale) or ΔZfp36 cells (y-axis, log scale). Green dots represent genes with predicted 'strong' Zfp36 binding motif, red represent known Zfp36 targets and black represent genes that fit both. (c-d) Predicted temporal changes in degradation rate (y-axis) either in WT cells (solid) or in ΔZfp36 cells (dashed) for the genes Cxcl2 (c) and Tnf (d). (e) Ratio between dynamic degradation rates predicted for a gene in WT and ΔZfp36 cells. Shown is the average across groups of genes with increasing numbers of binding heptamers (0 to 4, grayscale) and for 'strong' Zfp36 targets with 5 or more heptamers (red).

conduct additional studies to test the hypothesis that Zfp36 regulates RNA stability in this system.
6.3.2 Knock out of Zfp36 affects the degradation rates of candidate target transcripts

I measured the response to LPS stimulation is DCs derived from wild-type mice (WT) and mice with a knock out of the Zfp36 gene (ΔZfp36), resulting in an inactive protein. I collected RNA-total and RNA-4sU samples every 15 minutes for the first 3 hours of the response, and quantified with the nCounter system for a set of 267 signature genes.

I searched for the Zfp36 binding heptamer within the sequence of each of the 267 signature genes (figure 6-2a). Notably, only 10% of all heptamers occur within the CDS or 5’UTR of the genes, while most occur at the 3’UTR as expected. In 27 genes, I find multiple occurrences of the binding heptamer (5 or more), and consider those as candidate Zfp36 targets with a 'strong' binding site.

I fitted both a constant and a dynamic (impulse model) degradation models (see section 2.4.1) to either WT or ΔZfp36 data, thereby predicting the degradation rate of each mRNA at each time during the response, either for WT or ΔZfp36 cells. By comparing the constant degradation rates predicted by the model on the WT and the ΔZfp36 data, I find a slight overall increase of degradation rates in ΔZfp36 cells. However, for the group of candidate targets (based on binding site) and known targets, degradation rate are decreased (figure 6-2b), suggesting these mRNAs are stabilized in the absence of Zfp36, and supporting a role for Zfp36 in regulating their stability. As currently available UTR annotations are inaccurate in many cases, it might account for some genes with a significant effect on degradation rate in ΔZfp36 cells but without a strong Zfp36 binding site.

I next asked whether inactivation of Zfp36 alters dynamic changes in degradation rates. Indeed, of the 43 genes with predicted dynamic rate in WT cells, 28 genes retain the constant degradation hypothesis in ΔZfp36 cells. For example, the gene Cxcl2 (figure 6-2c) has a WT degradation rate that increases about 1 hour after stimulation, while the ΔZfp36 cells retain the constant degradation hypothesis (no change). For another 10 genes, degradation is dynamic in both WT and ΔZfp36
Figure 6-3: Predictions from the Zfp36 molecular model. (a) Measured Zfp36 protein levels in WT cells by mass spectrometry (black) for two replicates (dashed lines) and their average (solid line) and the levels of the regulator protein that are predicted by the model (red). The RNA levels of Zfp36 in WT (solid) and ΔZfp36 (dashed) cells are indicated at the bottom panel (blue). (b) The distribution of Km values predicted by the model for targets with either none, a single, 2-4 or more than 4 Zfp36 binding heptamers on their 3'UTR (x-axis).

cells, but the change in degradation rate predicted in WT is reduced in ΔZfp36 cells at the last 1-2 hours of the response. For example, the Tnf mRNA (figure 6-2d) is predicted to have an increase in degradation rate about 1 hour after stimulation in WT cells, but in the ΔZfp36 cells there is a much weaker change in the degradation rate that occurs 30 minutes later (possibly a result of a weaker effect of one of the other members of the TTP family). Comparing the dynamically changing rates, we find a much stronger effect in the group of candidate Zfp36 targets with 5 or more binding heptamers (figure 6-2e), where the WT rate is much higher than the ΔZfp36 rate after 1 hour of induction.

Finally, to rigorously study the changes between WT and ΔZfp36 cells, I analyzed...
the RNA levels that I measured in either the WT or ΔZfp36 cells together through the molecular regulatory model (section 6.2), by using Zfp36 as an activator of RNA degradation. This model predicts 26 genes as targets for Zfp36 regulation, including 8 out of the 12 known targets ($p < 4.7e^{-8}$). The activity function of Zfp36 protein, as predicted by the model, fits its measured protein level (figure 6-3a). The model also predicts a hill coefficient of $n = 9.89$, supporting cooperativity in Zfp36 binding, as suggested by the multiple binding heptamers. Moreover, the RNA concentration at which degradation reaches half of its maximal rate, as predicted by the model parameter $K_m$, decreases as the number of AU-rich heptamers found in the gene’s 3’UTR is higher, and most notably in the group of candidate targets with 5 or more binding heptamers (figure 6-3b), supporting its role in Zfp36 binding.

### 6.4 A knock-down screen suggests candidate RNA splicing regulators

While regulation by Zfp36 accounts for many of the dynamic degradation rate changes during the LPS response, others are still not well explained. Moreover, several other effects are not accounted for, including differences in the basal degradation and processing rates between genes and splicing junctions, as well as dynamic shifts in processing rates. I therefore used a knock-down screen to uncover additional molecular regulators. Indeed, using short hairpin RNAs (shRNAs) to reduce the expression levels of specific genes has several limitations. First, shRNA knock-downs have a high variability in ability to effectively reduce RNA levels even in repeated use of the same shRNA. Second, reducing the levels of the mRNA does not guarantee a similar reduction in protein levels. Finally, the virus which delivers the shRNA into the cells affect their response (and particularly for immune cells) even without an effective shRNA (e.g., using a mock shRNA). However, using an assay which is quick and easy to implement has significant advantages for a screening approach, as I take here. Finally I point out that this is still an ongoing project, which is far from reaching
**Figure 6-4:** A knock-down screen of RNA binding proteins. (a) DCs were infected with high titer lentiviruses encoding shRNA that target one of 25 RNA binding proteins. Infected cells were selected by adding puromycin, and stimulated with LPS. RNA samples were collected before stimulation (time 0), and after 45 and 135 minutes. RNA samples were quantified using the nCounter system. For a selected group of 6 RNA binding proteins, RNA was treated to deplete ribosomal RNA (by Ribo-Zero), and quantified by high throughput sequencing. (b) A list of all 25 RNA binding proteins in the screen, and their predicted functionality, as well as 4 control shRNAs used. Lines marked in red were included in sequencing analysis. (c) Knock down efficiency (x-axis) for all samples (including replicates for some genes, y-axis), at time 0 (red), 45 (green) and 135 (blue) minutes. The 60% knock down efficiency threshold is indicated by a dashed line.

final conclusions.
6.4.1 A knock-down screen of 24 RNA binding proteins

I selected 24 candidate RNA binding proteins (RBPs) that are highly expressed during the first 3 hours of the LPS response in mouse DCs, and generated a lentiviral shRNA library to systematically perturb each of these RBPs in DCs (figure 6-4a,b). I collected RNA samples from affected DCs before LPS stimulation, and 45 and 135 min. post stimulation, and used the nCounter system to measure RNA-Total levels for a set of 267 representative LPS genes, including 30 genes with both intron specific and exon specific probes. I validated a > 60% knock-down efficiency in 19 out of 24 shRNAs (figure 6-4c).

I measured the effect of shRNA over mock infection for 30 genes with an nCounter probe in both the exon and the intron. Of the 24 KDs, I find only two with a > 5 fold increase in intron levels of any of these 30 genes. In the KD of the splicing factor Ybxl there is an increase in intron levels of the anti bacterial response gene Acpp (figure 6-5a), while in the KD of the RNA endonuclease RNaseK there is an increase in intron levels of 5 anti viral genes (figure 6-5b).

Based on this signature response, I selected a set of 6 RBPs, including three splicing factors (Rbm39, Snrpd3, Ybxl), one editing enzyme (Adar) and two regulators of RNA degradation (RNaseK, Zfp36), and used massive parallel sequencing of ribosomal-depleted RNA to measure their genome-wide effect.

6.4.2 Perturbation data uncovers regulatory interactions that determine RNA processing rates

I applied the binomial model (section 4.3.2) to estimate junction specific precursor and mature RNA levels in each of the 6 KDs. In all samples, the majority of junctions show minimal expression difference between a mock shRNA (control) and the KD, both at the precursor and at the mature RNA level. However, for each of the 6 KDs, I identify a set of few hundred targets that are > 10 fold increased or decreased in the KD compared with a mock infection (figure 6-4c,d). I identify many more effects on the mature RNA rather than on the precursor of junctions. Notably, the KD of the
RNA editing enzyme *ADAR* leads to an increase in the levels of a significantly bigger group of junctions than other KDs (figure 6-4c), supporting a role for regulation by RNA editing. Indeed, when predicting editing sites in this response, we only predict a much smaller set of 53 transcripts with A to I editing (see section 5.4), but indirect effects can account for the reminder of the effect. Sequencing data provides the opportunity to directly check if indeed the levels of editing in any of these sites is reduced as a result of *ADAR* inactivation.

The current analysis can be further developed in several directions. First, there is a benefit in testing if the effect of each KD is junction specific or common to several junctions on a single transcript. While a repeated effect across several junctions can provide some support to the observed changes, a junction specific effect can suggest regulation at the splicing level. Another interesting question is whether the effect on specific junctions is limited to either precursor or mature RNA, or is it joint to both. An effect that is limited to the precursor suggests regulation at the splicing level, while an effect only at the mature RNA level suggests an RNA degradation effect, and a combined effect on both precursor and mature RNA predicts a regulation at the production level, which is likely to be an indirect effect since all KDs are RNA binding proteins. The KD effect can also be limited to just one of the time points, or can repeat in all three measurements. Finally, by looking for similar expression patterns within the group of targets of each KD, I can get insights into the molecular network that regulates the effect.

### 6.5 Summary

Deciphering the molecular mechanisms that control the dynamic RNA life cycle is a highly difficult task, as the molecular interactions between cellular components (binding motifs and active regulators) consist of multiple intertwined feedback loops that connect DNA, RNA and protein in non linear interactions. The ability to reach
mechanistic insights is tightly coupled to the ability to accurately measure the system's behavior at the molecular level, including direct interactions between molecular components and the effect of inactivating specific regulators.

Here I collected perturbation data, by monitoring RNA levels when specific proteins in the network are disabled. By following the LPS response in high resolution after knock out of the RNA binding protein Zfp36, I can infer its temporal regulator’s activity levels and the network structure that suggest Zfp36 as a candidate regulator of several dynamic changes in RNA stability during the LPS response in DCs. I measure genome-wide RNA levels in knock-downs of six candidate RNA binding proteins, which are involved in either degradation, splicing or editing of RNA, and predict the targets of each protein.
This analysis is part of an ongoing project, and highlights some promising directions for further investigation. First, using the molecular regulatory model with a Hill function on the Zfp36 data can help to identify cooperativity in the regulator activity, as suggested by the repeated binding sites observed in many of its targets. Second, an in-depth analysis of the genome-wide targets predicted for each of the six knock-downs can help in deciphering their effect, and the regulatory level (production, splicing, editing or degradation of RNA) at which it occurs. Finally, sequencing data provides the opportunity to directly check the effect of ADAR knock down on editing levels in predicted A to I editing sites, and further support these predictions. Overall, these results demonstrate the importance of combining computational approaches with measuring perturbation data in order to decipher the molecular mechanisms that control RNA levels.

6.6 Methods

RNA sample collection from mouse DCs. All animal protocols were reviewed and approved by the MIT / Whitehead Institute / Broad Institute Committee on Animal Care (CAC protocol 0609-058-12). DCs isolation and treatment, 4sU labeling and RNA isolation were done as described in section 3.6, with the following modifications. For the high-resolution analysis, I extracted RNA starting at 0h and until 3h after LPS stimulation in 15 minutes intervals. 4-thiouridine (Sigma) was added to a 500 μM final concentration 10 minutes before RNA collection. For the low resolution KD analysis, I extracted RNA at 0m, 45m and 135m after LPS stimulation, and 4-thiouridine (Sigma) was added to a 150 μM final concentration 30 minutes before RNA collection. Total RNA was extracted with the miRNeasy kit’s procedure (Qiagen), and sample quality was tested on a 2100 Bioanalyzer (Agilent). Due to low RNA yield, I did not follow with RNA-4sU purification.

shRNA knock down experiments. High titer lentiviruses encoding shRNA-targeting genes of interest were obtained from The RNAi Consortium (TRC; Broad Institute, Cambridge, MA, USA). Bone marrow cells were plated in complete RPMI on non-tissue culture treated plates. At day 2 cells were infected with shRNA-encoding lentiviruses supplemented
with polybrene (8 µg/mL). At day 4, infected cells were selected by adding puromycin (5 µg/mL) to the culture. Infected cells were used for analysis 3 days after initiating puromycin selection.

**qRT-PCR measurement.** RNA was reverse transcribed with the Sensiscript RT kit (Qiagen). Real time quantitative PCR reactions were performed on the LightCycler 480 system (Roche) with FastStart Universal SYBR Green Master Mix (Roche), and every reaction was run in duplicate.

**nCounter measurements and data processing.** nCounter sample preparation, capture, analysis and normalization were done as described in section 3.6.

**Preparation and processing of RNA-Seq libraries.** I prepared the KD sequencing libraries using the 'nonstrand-specific' protocol as described in [30], with the following modifications: after DNase treatment, I used the entire RNA sample (without polyA-RNA isolation) and fragmented RNA by incubation in 1X RNA fragmentation buffer (Affymetrix) at 80 degrees for 4 minutes. I sequenced 21 KDs samples using Illumina sequencer, generating $1.06 \cdot 10^9$ pair-end reads ($2.1 \cdot 10^9$ single reads, 101 bases long).

**Read Alignment.** I aligned all reads to the mouse reference genome (NCBI 37, MM9) using the TopHat aligner [49] with default parameters. To estimate rRNA levels, I mapped all reads (with the same procedure) to mouse rRNA sequences from GenBank [66]. Sequencing data contained about 20% rRNA reads (globally for all samples), yet only 20% of reads were uniquely aligned to the genome.

**Transcriptome reconstruction.** I used the same set of transcripts as was defined in section 5.6.
Chapter 7

Conclusion

7.1 Contributions

The ability to continuously respond to changing input signals from the environment is a hallmark of all living systems. A main component of this response is achieved through complex molecular circuits, consisting of multiple intertwined feedback loops and non-linear interactions, that compute a quantitative level of each type of RNA at each time. Living cells control RNA levels by tightly regulating the processes for their 'birth' (production, by transcription), maturation (processing and transport), compartmentalization (in granules) and 'death' (degradation), which together encompass the dynamic RNA life cycle. However, their immense complexity limited available research that describe, analyze and predict their behavior.

In this work I integrate comprehensive computational models with innovative empirical approaches in order to systematically study the dynamic changes in the cellular circuits that control RNA levels in mammalian cells, focusing on the response of immune dendritic cells to pathogens.

I developed an innovative laboratory assay to directly measure RNA transcription rates at high temporal resolution through short metabolic labeling of RNA. I significantly reduce metabolic labeling times by leveraging advanced RNA quantification assays, and combine metabolic labeling with quantification by the nCounter technol-
ogy for accurate multiplex measurement of RNA and by massively parallel sequencing. I collect extensive experimental data on the system, measuring in high resolution in both time and sequence space the dynamic response of mouse immune dendritic cells to LPS stimulation. Through deep sequencing of complementary RNA populations I measure a broad representation of the dynamic transcriptome. Metabolically labeled RNA is enriched with unstable transcripts, including precursors of lowly expressed or quickly processed junctions, while polyA+ RNA is enriched for the fully processed species, and ribosome depleted total RNA correctly represents their relative abundance. Finally, I collect perturbation data, by monitoring RNA levels when specific elements of the network are disabled.

I developed a new and effective quantitative approach to simultaneously model several key regulatory steps in RNA regulation, and their interactions with each other. I model genes as composite particles and study the regulation and kinetics of altering their internal structure. Generalized RPKM measures a transcript’s overall maturation rate, while a binomial model decomposes RNA levels into the relative abundance of precursor and mature transcripts (including splicing isoforms) at the splicing junction level. A new computational approach identifies RNA editing sites by comparing the nucleotide distributions across measured RNA populations. I study the kinetics of a gene’s RNA population and generalize the system as an ensemble of RNA genes. Dynamic models elucidate the separate contributions of RNA production, processing and degradation, estimate changes in these rates between genes and over time, and quantitatively measure how their integration shapes overall cellular RNA levels. Molecular models of RNA degradation help to identify the targets of a protein regulator from perturbation data.

In this way, I formulate several general principles of RNA regulation and its temporal evolution in mammalian cells. I find that changes in transcription rate highly correlate with changes in RNA level and precede them by 15-30 min. Dynamic changes in degradation rates have minimal effect on most RNA profiles, but they do play a unique role in shaping the expression of genes with sharp 'peaked' responses. Genome-wide analysis shows substantial variation in constant degradation rates be-

130
between genes, consistent with their regulatory and functional differences. I find that changes in mature RNA levels are preceded by correlated changes in precursor and newly produced RNA. While most junctions (95%) are spliced at a temporally constant rate, a dynamically changing splicing rate is important in specific cases. Marked differences in RNA processing rates between genes and between individual splicing junctions are correlated with genes' function and with junctions' structure and provide mechanistic insights into splicing regulation. Predicted editing events throughout the expressed transcriptome are mostly located at non-coding 3'UTRs, and most likely contribute to post-transcriptional regulation of expression. Perturbation of Zfp36, a key regulator of RNA degradation elucidate its role in controlling dynamic changes in the stability of specific transcripts.

7.2 Future directions

The experimental tools and quantitative framework are broadly applicable and open the way to many further research directions. This work highlights the need for an unbiased method to directly measure degradation rates on a genome scale. While existing direct methods are significantly biased, and particularly in dynamic settings, the approach proposed here avoids such biases, but only indirectly infers degradation rates. In comparing different estimates I find predictions from this work to be highly correlated albeit consistently shorter than other estimates in this system and others. While a proportional decrease in rates may result from difference in cell type or from the inherent limitations in all existing methods (labeling based and others), the RNA stability rankings, on which I base this analysis, seems more comparable between methods and reliable. One promising direction in settling these discrepancies is chasing 4sU with Uridine [59], which allows to estimate RNA degradation directly and with minimal biases, but may be challenging for highly induced genes. Correlating RNA production and degradation rates to proteomics and ribosome occupancy studies can connect the mRNA life cycle with that of its protein products,
as well as add more layers of regulation at the translation and sequestration of RNA in the cytoplasm.

It will be important to generalize these findings to other systems and responses, and test whether the difference between the conclusions of this work and those achieved in other immune cells [9, 14] stem from technology or biological factors. In some systems and responses, regulation of RNA levels post-transcriptionally might be a dominant factor, due to inherent limitations in regulating production rates.

Sequencing of 4sU labeled RNA provides opportunities to study many additional post-transcriptional regulatory processes, including poly-adenylation and possibly also 5’ capping. This broad transcriptome high resolution sequencing data allows to improve coding and non-coding RNA annotations, including microRNAs and anti-sense transcription, provide evidence on still debated RNA classes, such as enhancer associated RNAs, and to identify new classes of short-lived functional transcripts.

Monitoring RNA production, processing and degradation at the single-cell level will allow to distinguish between changes that arise from intrinsic properties of the molecular system and those that are affected by cell-to-cell variation. It will allow to understand the population’s behavior and learn how individual cells regulate different aspects of the RNA life cycle.

Expanding the perturbation studies, and combining them with measurements of direct interaction of RNA binding proteins and regulatory RNAs with their targets will provide more data to analyze the physical implementation of the regulatory circuit at the molecular level, and study how it is integrated with other molecular regulatory networks inside the cell, such as the transcriptional regulatory network, or the protein interaction network. In this way I can add to the system’s description also the ‘implementation’ of the regulatory circuit, through molecular interactions between cellular components (binding motifs and active regulators).

Finally, looking at a wide variety of stimulations, including those that change in time, can further elucidate new behaviors of the regulatory circuits.
7.3 Concluding remarks

It has been known for decades that RNA levels in cells are regulated by an intricate interplay between separate processes governing the dynamic RNA life-cycle. Cells invest tremendous resources to regulate each of these processes individually, and it is estimated that over 30% of the genes in the mammalian genome regulate the RNA life cycle. Despite its central cellular role, the dynamics of the RNA life cycle remain surprisingly poorly understood. Even though regulatory circuits have been studied and modeled extensively, most of the work focused only on transcriptional regulation and overall RNA levels, and ignored the many other steps that control the observed RNA levels.

Indeed, maybe most importantly, the experimental tools and quantitative framework developed in this work has important implications for many areas, and are relevant for a more general analysis of complex systems dynamics. Such a quantitative study of RNA regulation as I propose is a complementary view to any RNA expression analysis, as a tool to provide a separate view of each RNA regulatory event. Such a unique tool for studying the cellular regulatory programs is also crucial for understanding their mis-regulated in many diseases, including cancer, autoimmune disorders, and diabetes. For instance, key proteins in the RNA life cycle are disrupted in up to 50% of cancers, including aberrations in RNA production (e.g. over-production of $cMyc$ and $p53$), splicing (e.g. the splicing factor $SF2/ASF$) and stability (e.g. mutations in the RBP $Zfp36$), and contribute to tumorigenesis. But, while I used these tools to study how cells implement dynamic transitions, this fundamental description is likely to be broadly-applicable across many complex systems, including biological processes like evolution, as well as computer or cellular networks.
Bibliography


