8-2016

Computational Analyses of mRNA Ribosome Loading in Arabidopsis Thaliana

Joseph Benjamin Ernest
University of Tennessee, Knoxville, jernest1@utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Joseph Benjamin Ernest entitled "Computational Analyses of mRNA Ribosome Loading in Arabidopsis Thaliana." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Life Sciences.

Albrecht von Arnim, Major Professor

We have read this dissertation and recommend its acceptance:

Mariano Labrador, Michael Gilchrist, Michael Langston

Accepted for the Council:
Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Computational Analyses of mRNA Ribosome Loading in Arabidopsis Thaliana

A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Joseph Benjamin Ernest
August 2016

Acknowledgements

My completion of this dissertation and development as a scientist have been possible only through the help of many people. Nobuyo Maeda from the University of North Carolina and John Parks and Lew Nelson from Wake Forest University School of Medicine helped me in my decision to pursue a PhD in the biological sciences. Many others have taught, counseled, and encouraged me along my journey through the Graduate School of Genome Science and Technology (GST) at The University of Tennessee.

Fortunately for me, Tamah Friedman left no child behind in her computational biology course. Arnold Saxton patiently trained and encouraged me in biostatistics. Brynn Voy, my initial research advisor at UT, demonstrated sincere interest in my overall development and well-being, helped me to discover what my niche was and was not, and encouraged me through stressful times.

Then there's Albrecht. As my research advisor, he challenged, encouraged, taught, and guided me the whole way. His dedication to and love of science are truly infectious and paralleled only by his sincere desire for his students and those around him to succeed.

My other committee members—Mike Langston, Mike Gilchrist, and Mariano Labrador—helped me in unique ways. Mike Langston helped me to "see the light" (i.e, my computer screen) when I first considered trying computational biology as a full-time job, and his "foreman", Charles Phillips, helped me to learn the tools of the trade as a computational biologist. Mike Gilchrist gave me just the right amount of "tough love" to help me to do my best and strive to be better. And Mariano Labrador, who was on my committee the whole time, helped me to build confidence in myself.

This dissertation literally would not exist without the high-quality experimental work of two gifted post-doctoral fellows, Anamika Missra and Ju Guan. Anamika "teed the ball up" for me by performing all of the experiments and generating the data that made chapter 2 and the corresponding publication possible, and Ju generated the excellent data used in chapter 3 and contributed valuable biological insight into the computational approaches.

Many lab members in the Voy, Campagna, von Arnim, and Nebenfuehr labs helped me reach this point through their encouragement, friendliness, and dedication to science.

Outside of science I have made some amazing friends in Knoxville who have helped me to take often needed breaks from thinking about science and to think instead about beer, the outdoors in East Tennessee, and Tennessee football.

Lastly, Mom, Dad—thank you, and I love you.

Abstract

Translation of mRNA into protein is a critical step in gene expression, but the principles guiding its regulation at the genome level are not completely understood. Translation can be quantified at a genome scale by measuring the ribosome loading of mRNA—the extent to which mRNA is associated with ribosomes. In this dissertation, I present investigations into how genome-wide ribosome loading is controlled in Arabidopsis thaliana. In chapter 1, I give an overview of regulation of ribosome loading and translation. In chapter 2, I present research demonstrating for the first time that genome-wide ribosome loading in plants is partially controlled by the circadian clock. In chapter 3, I present a study of a computational model that describes how various biochemical steps control ribosome loading. And in chapter 4, I conclude by briefly summarizing the dissertation as a whole and discussing future perspectives.

Table of contents

List of tables

List of figures

Chapter 1

Introduction

All cellular organisms make proteins to carry out the biological processes necessary for life. Cells make proteins according to the central dogma of molecular biology, which states that DNA is used to make RNA, and RNA is used to make proteins (Crick 1958, Crick 1970). Cells adjust the type and abundance of proteins they make in order to facilitate growth and development and to maintain homeostasis in a changing intracellular and extracellular environment. Since each type of protein is encoded by a specific DNA sequence in the form of a gene, understanding how genes are used to make proteins is central to understanding how cells function. The basic principles of this process of gene expression are clear. DNA from each gene is transcribed into specific mRNAs by enzymes called RNA polymerases, and mRNAs are translated into specific proteins by cellular machines called ribosomes.

The biochemical steps involved in gene expression have been characterized in much detail (Moore 2005). Many fundamental concepts of mRNA expression were established using early hybridization approaches, such as northern blotting, *in situ* hybridization, RNAse protection assays, and reporter gene expression assays, which target one RNA species at a time. However, much remains to be understood about how gene expression is regulated. Stemming from advances in technologies for studying RNA in the 1980s and 1990s, including the polymerase chain reaction (PCR) and microarrays, much of the research focus on gene expression regulation at the genome scale has been at the level of transcription.

Indeed, mRNA levels are important determinants of protein levels. But, once developments in high-throughput proteomics enabled comparisons of mRNA and protein levels at a genome scale, it became clear that mRNA levels are not perfect predictors of protein levels (de Sousa Abreu, Penalva et al. 2009), and the importance of post-transcriptional regulation in gene expression became increasingly recognized. Protein synthesis is energetically very costly

for a cell, so translation is tightly regulated in order for a cell to manage its resources efficiently (Jackson, Hellen et al. 2010, Vogel and Marcotte 2012).

Advances in experimental and computational approaches now allow detailed exploration into post-transcriptional regulation of mRNA. For instance, a pool of mRNA from a biological sample can be fractionated based on the number of ribosomes associated with each mRNA, and each fraction can be quantified in order to indirectly estimate the translation efficiency of each mRNA species. Isotope labeling can be used to directly quantify mRNA synthesis and turnover rates (Schwanhausser, Busse et al. 2011). RNA sequencing (RNA-seq), combined with techniques for isolating ribosomes and proteins, can be used to track the positions of ribosomes and proteins along an mRNA (Ingolia, Ghaemmaghami et al. 2009). The drug harringtonine stalls ribosomes near the start codon of mRNA but allows others to continue translation. Cycloheximide causes stalling of ribosomes at any location along mRNA. Therefore, treatment of cultured cells with harringtonine, followed by cycloheximide at various time points and sequencing of ribosome-bound mRNA fragments, can be used to monitor the elongation speed of ribosomes along mRNA at the genome scale (Ingolia, Lareau et al. 2011). Combined with advances in computer hardware and software for deciphering meaning from data generated from high-throughput experiments, the field of study devoted to post-transcriptional regulation of mRNA is expanding our understanding of the central dogma.

The research presented in this dissertation focuses on control of ribosome loading—the number of ribosomes associated with an mRNA molecule in a cell. While proteomics can be used to measure absolute protein synthesis rates in a high-throughput manner, it cannot reveal how transcription and post-transcriptional factors work in concert to give rise to a particular rate of protein synthesis. Measuring ribosome loading provides a closer look at post-transcriptional

processes that act upon mRNA and which, combined with mRNA abundance, control the protein synthesis rate. In this introduction, I first review the process of protein translation in eukaryotes. Second, I review high-throughput approaches for quantifying translation, focusing on RNA-based techniques. Third, I review the post-transcriptional processes that control ribosome loading and translation. Fourth, I provide context for the research presented in chapter 2, which is a research article describing our characterization of diel and circadian regulation of ribosome loading in Arabidopsis. And last, I give a brief overview of the remaining chapters of this dissertation. The research presented throughout involves control of ribosome loading in plants, so when relevant, information that is unique to plants is highlighted.

Protein translation: a review

Pre-mRNA processing

Transcription in eukaryotes occurs in the nucleus and produces a precursor mRNA (pre-mRNA) which is processed into mRNA, complexed with proteins as a messenger ribonucleoprotein complex (mRNP), and exported into the cytoplasm. Pre-mRNA processing involves three main steps, all of which are important regulators of ribosome loading and translation.

5' capping

The first step in pre-mRNA processing is 5' capping, in which 7-methylguanosine (the "5' cap") is added to the 5' end. Capping is important because the 5' cap is recognized by the eukaryotic initiation factor 4E (eIF4E), which facilitates initiation by recruiting the 43S pre-

initiation complex (PIC; composition described in "Initiation" section below) to the 5' end of the mRNA through a relay of interacting proteins. eIF4E interacts with eIF4G, which interacts with the eIF3 complex, which binds to the 43S small ribosomal subunit (43S). eIF4E, eIF4G, and the RNA helicase eIF4A are components of the eIF4F complex, which unwinds the mRNA in order to allow the PIC to bind near the 5' cap and scan toward the start codon. Once the PIC recognizes the start codon, the 60S large ribosomal subunit (60S) joins it in order to form the full ribosome and begin elongation.

Polyadenylation

The second step in pre-mRNA processing is polyadenylation, during which multiple adenine bases ("poly(A) tail") are added to the 3' end. Polyadenylation is important because the 3' poly(A) tail is recognized by the poly(A)-binding protein (PABP) which also interacts with eIF4G at the 5' cap in order to form the "closed loop" configuration. The closed loop enhances translation in two ways—first, by helping to recycle ribosomes after they complete translation, and second, by keeping eIF4F associated with the mRNA even if it loses contact with the 5' end (Wells, Hillner et al. , Gray, Coller et al. 2000, Jackson, Hellen et al. 2010). Polyadenylation involves cleavage of the pre-mRNA at the poly(A) cleavage site by a protein complex called cleavage and polyadenylation specificity factor (CPSF), followed by polyadenylation by polyadenylate polymerase (Mandel, Kaneko et al. 2006, Hunt, Xu et al. 2008). Many mRNAs have multiple poly(A) cleavage sites, so selection among multiple poly(A) sites, termed "alternative polyadenylation" (APA), determines the length and nucleotide content of the 3' poly(A) tail (Wu, Liu et al. 2011, Xing and Li 2011).

Splicing

The third and final step in pre-mRNA processing is splicing, during which introns are removed and exons are stitched together to form a mature mRNA. Splicing is an important regulator of ribosome loading and translation because it determines the nucleotide sequence of the mature mRNA, which influences many aspects of ribosome loading and translation as discussed in detail in the section "Post-transcriptional control of ribosome loading and translation" in this chapter, as well as the actual protein product. Following processing and export into the cytoplasm, a mature mRNA is ready to be translated into a protein via initiation, elongation, and termination, before eventually being destroyed.

Initiation

Initiation involves three main steps, 1) recruitment of the PIC to an mRNA, 2) scanning of the 5' UTR by the PIC, and 3) recognition of and commitment to a start codon.

The PIC is composed of 1) the 40S, 2) a ternary complex composed of eIF2 bound to GTP and a charged methionyl initiator tRNA (eIF2-GTP-Met-tRNA$_i^{Met}$), and 3) initiation factors eIF3, eIF1, and eIF1A. The eIF4F complex, consisting of subunits eIF4E, eIF4A, and eIF4G, associates with the 5' cap via eIF4E. eIF4G is a scaffold protein that binds eIF4E, eIF4A, PABP, and eIF3. eIF4G induces a structural conformation of eIF4E which stabilizes its association with the cap. Helicases, including eIF4A, unwind the secondary structure of the 5' UTR, which promotes efficient binding of the PIC. eIF4H and eIF4B bind to both the mRNA and eIF4A and enhance the helicase activity of eIF4A by preventing re-annealing of unwound mRNA and promoting movement of the PIC along the 5' UTR (Marintchev, Edmonds et al.

2009, Sun, Atas et al. 2012). The PIC attaches to the large assembly of proteins at the 5' UTR via eIF3.

Following attachment of the PIC to the mRNA near the 5' cap, it scans the 5' UTR in search of a start codon via an alternating sequence of steps consisting of unwinding mRNA secondary structure and movement along the unwound mRNA. Scanning requires that the PIC adopts a specific structural conformation that is induced by eIF1 and eIF1A (Pestova and Kolupaeva 2002, Passmore, Schmeing et al. 2007). The unwinding of mRNA secondary structure is mediated by some of the same key players as in the initial attachment of the PIC. eIF4A, eIF4G, and eIF4B, as well as ATP, are required for scanning, with eIF4A and ATP required in amounts that are proportional to the degree of mRNA secondary structure (Jackson 1991, Svitkin, Pause et al. 2001), demonstrating the importance of unwinding for the scanning process. Many key details of the scanning process are still poorly understood (Pestova and Kolupaeva 2002, Jackson, Hellen et al. 2010).

Third and finally, the PIC recognizes and commits to a start codon. The start codon chosen is usually an AUG in a favorable sequence context—GCC(A/G)CCAUGGG, also termed the Kozak consensus sequence named after its discoverer (Kozak 1991). eIF1 and eIF1A, in addition to inducing structural changes in the PIC necessary for scanning (Pestova and Kolupaeva 2002, Passmore, Schmeing et al. 2007), help to ensure that the PIC recognizes the correct start codon and help to disrupt assembly of ribosomal complexes on incorrect codons (Pestova, Borukhov et al. 1998, Pestova and Kolupaeva 2002, Pisarev, Kolupaeva et al. 2006). The PIC recognizes start codons via codon-anticodon base pairing, and eIF1 and eIF1A stabilize a "closed" conformation of the PIC when it encounters a start codon with a strong Kozak consensus sequence (Unbehaun, Borukhov et al. 2004, Pisarev, Kolupaeva et al. 2006). Once the

PIC stops at the correct start codon and eIF1 and eIF1A help to lock the PIC in place, eIF5 binds to the β-subunit of eIF2 (eIF2β) of the PIC and induces eIF2γ to hydrolyze eIF2-GTP into eIF2-GDP, which leads to displacement of eIF2-GDP, eIF1, eIF1A, and eIF3 and joining of the 60S to form the complete 80S ribosome (Shin, Kim et al. 2011). When initiation ends, Met-tRNA$_i^{Met}$ occupies the P site of the ribosome, the A site is vacant, and elongation is ready to begin.

Elongation

Elongation is a repeating process which begins with a peptidyl-tRNA occupying the P site of the ribosome (Kapp and Lorsch 2004, Chen, Tsai et al. 2012, Lareau, Hite et al. 2014). A tRNA carrying the next amino acid to be added to the growing peptide becomes associated with a ternary complex containing GTP and elongation factor 1A (aa-tRNA-GTP-eEF1A). While any such ternary complex can bind in the A site, conformational changes and hydrolysis of GTP by eEF1A ensure that only the correct tRNA is able to remain bound and continue to the next step of elongation during which a peptide bond will be formed (Rodnina and Wintermeyer 2001). In this next step, the ribosome carries out its peptidyl transferase reaction, forming a peptide bond between the peptidyl-tRNA in the P site and the amino acid associated with the tRNA in the A site. eEF2 then facilitates movement of the tRNA formerly bound to the peptide into the E site and the peptidyl-tRNA—now one amino acid longer—into the P site, leaving the A site once again vacant. This cycle continues until the ribosome encounters a stop codon (Wintermeyer, Savelsbergh et al. 2001).

Termination

Termination results in hydrolysis of the bond between the peptide chain and the tRNA occupying the P site of the ribosome in a reaction catalyzed by the peptidyl transferase center of the ribosome (Caskey, Beaudet et al. 1971, Rodnina 2013, Yusupova and Yusupov 2014). This peptidyl transferase activity is induced by release factors, which detect the presence of a stop codon in the A site of the ribosome (Kapp and Lorsch 2004).

Approaches for quantifying translation at a genome scale

In order to address the question of how a cell regulates translation, translation must be quantified. Translation is a multi-faceted process, so the choice of measurement technique depends on the question being asked and the aspect of translation being investigated. Below, I give a brief overview of mass spectrometry-based proteomics and its limitations before discussing RNA-based techniques for quantifying translation.

Mass spectrometry-based proteomics

Mass spectrometry has made impressive gains in its ability to measure genome-wide protein abundance, and it is beginning to be applied for measuring rates of protein synthesis and turnover (Schwanhausser, Busse et al. 2011). To measure genome-wide rates of protein synthesis using mass spectrometry, researchers can monitor the incorporation of a detectable label into proteins over time. For instance, in a technique called “stable isotope labeling by amino acids in cell culture”, or SILAC, cultured cells are treated with media containing a heavy stable isotope version of an essential amino acid, which can be distinguished from the normal

light version by mass spectrometry because of its slightly different mass. The increase in the abundance of the heavy isotope versions of proteins over time can be used to determine the protein synthesis rates at a genome scale (Ong, Blagoev et al. 2002). To compare rates between treatments or conditions, an extension of SILAC called pulsed SILAC (pSILAC) can be used. In pSILAC, two isotope versions can be used, with one treatment group being treated with a medium-heavy version, and the other treatment group being treated with a heavy version. Following this pulse with medium-heavy or heavy isotope amino acids, protein samples are prepared from both groups, mixed together (to avoid variance between instrument runs), and analyzed by mass spectrometry. The relative abundance of medium-heavy to heavy versions of proteins allows for the estimation of the relative protein synthesis rates between the two treatments or conditions (Schwanhäusser, Gossen et al. 2009).

Despite significant advances, however, mass spectrometry-based proteomics still has its limitations. Proteome coverage, or the percent of proteins that can be quantified for a given organism, is only beginning to approach transcriptome coverage. Even so, to obtain the same coverage as a routine RNA-seq experiment requires much more time and effort in sample preparation and instrument time, with one group spending 12 days of instrument analysis in order to quantify 10,300 proteins, while in the same experiment, routine RNA-seq quantified over 12,000 mRNA transcripts (Nagaraj, Wisniewski et al. 2011). Even as proteomics begins to approach transcriptomics in terms of genome coverage, it is still difficult to quantify the dynamic cellular processes that give rise to a particular protein expression pattern. For these reasons, while mass spectrometry-based proteomics is an important and growing analytical approach to studying translation, techniques that quantify translation by targeting RNA are reasonable and powerful alternatives.

RNA-based approaches

cDNA microarrays and RNA-seq are the platforms of choice for investigating cellular events at the level of mRNA at the genome scale, including synthesis, turnover, molecular interactions, and cellular localization, all of which influence translation. Approaches to quantifying genome-wide translation from mRNA measurements can be divided into two classes—polysome profiling and ribosome footprinting. Neither polysome profiling nor ribosome footprinting by themselves can distinguish mRNA with actively translating ribosomes from mRNA with stalled ribosomes (Ingolia, Lareau et al. 2011), and ribosomes bound to mRNA are not necessarily engaged in translation (Guttman, Russell et al. 2013), so two assumptions must often be made in order to use ribosome loading as an indication of the protein synthesis rate. The first is that all ribosomes on an mRNA are actively engaged in translation, and the second is that all ribosomes carry out elongation at the same rate. If these assumptions hold, then the number of ribosomes bound to an mRNA is proportional to the protein synthesis rate.

Polysome profiling

Polysome profiling is an approach for quantifying how actively translated each mRNA species is based on the degree of association of each mRNA species with ribosomes. Successive rounds of initiation on the same mRNA produce an mRNA with multiple ribosomes attached to it in a complex called a poly-ribosome, or polysome. Experimentally, polysomes of different sizes (i.e., number of ribosomes) can be separated using centrifugation, yielding multiple polysome fractions, with each fraction corresponding to mRNA with a specific number (or range of numbers) of ribosomes (Johannes, Carter et al. 1999, Zong, Schummer et al. 1999, Arava,

Wang et al. 2003). Alternatively, polysomes of all sizes can be isolated in one fraction using immunoprecipitation with antibodies that recognize ribosomal proteins (Zanetti, Chang et al. 2005, Branco-Price, Kaiser et al. 2008). The abundance of each mRNA species in each fraction can be quantified by microarray or RNA-seq. The mRNA abundances in the various fractions and the total mRNA sample can be used to calculate a value which reflects the translation activity of each mRNA species which is independent of mRNA abundance. These techniques yield measurements commonly referred to as translation state (TL) (Zong, Schummer et al. 1999), ribosome occupancy (Tiruneh, Kim et al. 2013), or ribosome loading (Kawaguchi, Girke et al. 2004).

Our lab has used an approach to polysome profiling which aims to estimate the number of protein molecules being synthesized per mRNA per unit time (Missra, Ernest et al. 2015). In our approach, mRNA is fractionated based on polysome size into three fractions. The non-polysomal (NP) fraction corresponds to poorly translated mRNAs having zero or one ribosome. The small polysome (SP) fraction corresponds to moderately translated mRNAs having a medium number of ribosomes. And the large polysome (LP) fraction corresponds to highly translated mRNAs having many ribosomes. The expected number of ribosomes bound to mRNA in each fraction has been estimated at zero, two, and seven, for NP, SP, and LP, respectively (unpublished data from the von Arnim lab). TL for each gene is calculated as $\frac{2 \cdot SP + 7 \cdot LP}{NP + SP + LP}$, where NP, SP, and LP are the abundance of mRNA in those fractions. Using this approach, TL is an estimate of the average number of ribosomes per mRNA and serves as an approximation of the number of protein molecules being synthesized per mRNA per unit time.

Ribosome footprinting

A second method for quantifying translation at the genome scale based on mRNA measurements is referred to as ribosome footprinting or ribosome profiling (Ingolia 2014). This approach involves enzymatically digesting mRNA that is not protected by ribosomes and using RNA-seq to count the number of ribosome-protected fragments, which reflects the number of ribosomes bound to each type of mRNA and their positions (Ingolia, Ghaemmaghami et al. 2009, Ingolia, Brar et al. 2012). This approach has certain advantages over polysome profiling. One advantage is avoiding the polysome fractionation process, which is time- and labor-intensive and cannot always distinguish large polysomes that differ by only one or two ribosomes. A second advantage is the ability to track positions of ribosomes along mRNAs. On the other hand, without separating mRNA based on ribosome loading, some important information is lost. While the number of ribosome-protected mRNA fragments per gene reflects the extent of ribosome loading, it cannot be used to directly quantify the average number or distribution of numbers of ribosomes per mRNA.

Post-transcriptional control of ribosome loading and translation

In this section, I discuss how ribosome loading and translation of mRNA are controlled by key post-transcriptional processes, including translation initiation, elongation, marking for degradation, and degradation, and how each process is regulated. The rates at which each of these processes operate determine which of them is rate-limiting for the overall process of translation. The structural and chemical properties of an mRNA molecule, which are determined by its sequence, can influence each of the biochemical steps in translation, so these properties of

an mRNA molecule have the potential to determine the rate-limiting step in translation and therefore, in a sense, to establish the rate of translation. However, in addition to the mRNA sequence, which is generally unchangeable, processes that are dynamically regulated also have the potential to modulate each of the biochemical steps. For instance, initiation may always be the rate-limiting step for one gene, so this gene may preferentially modulate initiation in order to alter its translation. A different gene, however, could have elongation as its rate-limiting step, in which case it may target elongation in order to regulate its overall translation rate.

Therefore, as discussed below, the translation rate for a gene is controlled by a combination of fixed, mRNA-specific properties, as well as dynamic processes that respond to stimuli from inside the cell and from its external environment. This phenomenon is evident from several observations. First, there can be substantial variation in ribosome density in different regions of an mRNA. For instance, in one of the first ribosome footprinting studies it was found that ribosome density was three-fold higher on average in the first 30-40 codons compared to the rest of the mRNA across all genes (Ingolia, Ghaemmaghami et al. 2009). Second, there can be substantial variation in ribosome loading of the same mRNA under different environmental conditions, as many studies, including ours, have shown (Kawaguchi, Girke et al. 2004, Brengues, Teixeira et al. 2005, Juntawong, Girke et al. 2014, Missra, Ernest et al. 2015). And lastly, there is substantial variation in ribosome loading among mRNAs encoded by different genes. For instance, in the same original yeast experiment, there was over 100-fold variation in the ratio of ribosome-protected mRNA fragments to total mRNA fragments across genes, and in plants, there was nearly 100 thousand-fold variation in the number of ribosome-protected mRNA fragments per gene (in ribosome-protected fragments per kilobase per million reads, or rpKM) (Juntawong, Girke et al. 2014).

Translation regulation via initiation

Initiation is the primary rate-limiting step in translation

While the question of how each gene utilizes various biochemical processes in modulating ribosome loading has not been thoroughly addressed at the genome scale, it has been explored. It is widely held that initiation is the primary rate-limiting step at which translation is regulated for most genes. One major theoretical basis for this assumption is the following. It is assumed that the total amount of mRNA in a cell far exceeds the number of ribosomes and that ribosomes that terminate translation on an mRNA dissociate and do not undergo re-initiation. Once a ribosome terminates translation of an mRNA from one gene, it will almost certainly be "captured" by an mRNA from a different gene. Therefore, increasing the elongation or termination rate for one mRNA species would not be expected to influence the number of ribosomes that translate an mRNA species per unit time (Andersson and Kurland 1990, Bulmer 1991). However, it is also theorized that if ribosomes are so abundant in a cell that a ribosome can bind to an mRNA immediately after the one preceding it moves past the initiation site, then the elongation rate can become rate-limiting (Bulmer 1991). Similarly, an early simulation model suggested that elongation could become rate-limiting if it slows drastically, such as in response to amino acid starvation (Harley, Pollard et al. 1981). Nevertheless, most theoretical and empirical evidence suggests that initiation is typically rate-limiting. For instance, ribosome occupancy has been shown to be well below the maximum packing density along mRNA in several organisms, including *E. coli* (Ingraham, Maaløe et al. 1983) and yeast (Arava, Wang et al. 2003). Additionally, inserting rare, presumably slowly translated codons into an mRNA typically has no effect on the protein synthesis rate (Robinson, Lilley et al. 1984).

Computational models describing the dynamics of protein translation have been developed which are also consistent with initiation being the primary rate-limiting step in translation. For example, Plotkin's group developed a model in which each ribosome and tRNA in a cell were considered to be either freely diffusing or bound to a specific mRNA molecule at a specific codon position. Rates of translation initiation and elongation depend on values of model parameters, such as abundances of ribosomes, tRNA molecules, and mRNA molecules, as well as cell volume and other parameters, many of which were previously determined experimentally by other researchers. The model suggested that depletion of free ribosomes, which would slow initiation rates for many mRNAs, slowed protein synthesis (Shah, Ding et al. 2013). As discussed above, initiation, like other biochemical steps involved in translation, can be regulated by fixed, intrinsic properties of mRNA, as well as dynamically regulated processes.

Regulation of initiation by fixed mRNA properties

Fixed properties of mRNA that influence initiation include mRNA secondary structure, the start codon sequence context, upstream open reading frames (uORFs), mRNA length, and indirectly, the elongation rate, as will be explained here. First, mRNA secondary structure influences initiation because tightly folded mRNA in the 5' UTR can hinder recruitment of the PIC and its scanning of the 5' UTR. Plotkin's group in this case took a more experimental approach to studying translation by generating 154 different mRNA sequences that all encoded the same green fluorescent protein (GFP) but used different combinations of codons (Kudla, Murray et al. 2009). When expressed in *E. coli*, these sequences yielded 250-fold variation in protein abundance, with as much as 59% of this variation explained by the predicted folding energy of the first 40 nucleotides. Consistently, adding an mRNA tag with weak predicted

structure to the 5' end of these mRNAs led to an increase in protein abundance. These findings in prokaryotes and similar findings in both prokaryotes and eukaryotes (Wen, Lancaster et al. 2008, Cannarozzi, Schraudolph et al. 2010, Fredrick and Ibba 2010, Gu, Zhou et al. 2010, Tuller, Carmi et al. 2010, Tuller, Waldman et al. 2010) suggest that weaker mRNA folding near the start codon improves PIC recruitment and thus the efficiency of initiation.

A second example of a fixed property of mRNA influencing its translation at the initiation step is the start codon sequence context, or Kozak context. The closer the start codon sequence context adheres to the Kozak consensus sequence, the more efficiently it will be recognized by the PIC (Kozak 1986). In support of this, it was shown that "anti-Kozak" sequences, or sequences that differ strongly from the Kozak consensus sequence, in the 30 nucleotides upstream and downstream of the start codon, help to ensure that initiation occurs efficiently at the main start codon (Zur and Tuller 2013). Tuller et al. speculated in their attempt to predict ribosome loading using various sequence features, that considering the Kozak context would improve their model's performance (Tuller, Veksler-Lublinsky et al. 2011).

A third example of fixed mRNA properties influencing initiation is the presence of start codons in the 5' untranslated region (UTR) of an mRNA, known as upstream open reading frames (uORFs). uORFs are present in 31% of genes in Arabidopsis (Kim, Cai et al. 2007) and 50% of genes in humans and mice (Calvo, Pagliarini et al. 2009). They typically repress translation, likely by promoting their own initiation and termination, which is followed by ribosome dissociation, all of which hinder initiation at the main ORF (Kozak 1991, Morris and Geballe 2000, Calvo, Pagliarini et al. 2009). However, a number of uORFs have been identified which instead promote initiation and translation of the main ORF under specific conditions (see

subsection “Regulation of initiation by dynamically regulated processes” immediately following this one for examples).

Fourth, ribosome density of mRNA—the number of ribosomes per unit length—is negatively correlated with the length of the coding sequence, although the reasons are unclear (Arava, Wang et al. 2003). Fifth and lastly, the elongation rate, which as discussed below is largely dictated by an mRNA’s sequence, can potentially affect the initiation rate because it affects the rate at which ribosomes are cleared from the initiation site (Chu, Kazana et al. 2013).

Regulation of initiation by dynamically regulated processes

Since initiation is the primary step at which translation is regulated, its control by dynamically regulated processes, including stimuli from a cell’s internal and external environments, plays a critical role in gene regulation. Generally, factors that up-regulate initiation in a cell do so in a global manner, affecting many mRNAs. This is because initiation occurs via interactions between the eIF4F complex and the 5’ cap, which most mRNAs have. On the other hand, factors that down-regulate initiation do so in either a global or gene-specific manner. During stress, phosphorylation of eIF2α causes it to bind tightly to and inhibit the guanine exchange factor (GEF) eIF2B. Active eIF2B helps to convert eIF2-GDP into eIF2-GTP which becomes part of the ternary complex of the PIC, so phosphorylation of eIF2α represses initiation by inhibiting PIC assembly. Thus, repression of initiation via phosphorylation of eIF2α is a global event.

A second mechanism for global repression of initiation is inhibition of the cap-recognition process (Marcotrigiano, Gingras et al. 1999, Mathews, Sonenberg et al. 2007). The eIF4E-binding proteins (4E-BPs) bind to eIF4E and inhibit its interaction with eIF4G, thereby

hindering the ability of the eIF4F complex as a whole to bind to the 5' cap of mRNA and recruit the PIC. Phosphorylation of the 4E-BPs, for instance, by mTOR in response to nutrients and growth factors, inhibits their binding to eIF4E and thus allows efficient initiation of translation for many mRNAs (Holz, Ballif et al. 2005).

While these cellular events repress translation of many mRNAs, some mRNAs are able to escape global repression of initiation in response to stress. In yeast, for example, phosphorylated eIF2α (eIF2α-P) promotes translation of *GCN4* mRNA by blocking the inhibitory effects of uORFs on reinitiation (Hinnebusch, Dever et al. 2007). In mammalian cells, a similar mechanism promotes reinitiation of *ATF4* mRNA in response to stress while most mRNAs are translationally repressed (Vattem and Wek 2004, Hinnebusch, Dever et al. 2007).

In contrast to nearly global suppression of initiation in response to certain conditions, suppression of initiation can be mRNA-specific. The classic example of mRNA-specific suppression of initiation is microRNAs which, as discussed in the section "Regulation of mRNA degradation by dynamically regulated processes", recognize specific mRNAs by direct base pairing with the 3' UTR of their targets. While the predominant mode of action of microRNAs in plants is to promote degradation of their target mRNAs, they also appear capable of inhibiting initiation, although the underlying mechanisms and the extent of this phenomenon in plants are controversial (Rogers and Chen 2013).

Translation regulation via elongation

While initiation is the primary step at which translation is regulated for most genes in most organisms, several lines of evidence point to an important role for elongation in translation regulation, both by properties of mRNA and by dynamically regulated processes. The fixed

properties primarily include the codons used by an mRNA and their arrangement, in addition to the three-dimensional structure of an mRNA. In this section, I first explain how different codons can be translated at different rates, before discussing the broader topic of how fixed mRNA properties and dynamic processes influence the elongation rate.

How are different codons translated at different rates?

Over several decades it has been established that synonymous codons, which are different codons encoding the same amino acid, are translated at different rates. It was proposed very early that the concentrations of each tRNA species play an important role in regulation of protein synthesis (Ames and Hartman 1963). Once DNA sequence data for several organisms accumulated, it was quickly realized that synonymous codons were not used randomly within a particular genome and that each organism's genome used a consistent coding strategy across nearly all genes, a concept referred to as the "genome hypothesis" (Grantham 1980, Grantham, Gautier et al. 1980, Grantham, Gautier et al. 1981, Ikemura 1985). Not long after this realization, it became clear that an organism's codon strategy is related to its population of isoaccepting tRNAs (Post, Strycharz et al. 1979, Ikemura, Osawa et al. 1980, Post and Nomura 1980, Ikemura 1981, Bennetzen and Hall 1982, Ikemura 1982), which are tRNAs that are charged with the same amino acid but differ in the codon that they recognize. It was also realized that the extent of codon usage bias (CUB) is highly correlated with the level of protein expression across genes (Grantham, Gautier et al. 1981, Ikemura 1981, Bennetzen and Hall 1982, Ikemura 1982, Ikemura and Ozeki 1983). Specifically, these studies have shown that highly expressed genes tend to have extreme CUB, while poorly and moderately expressed genes tend to have low to moderate CUB. The usage of any given isoaccepting tRNA had also been

shown to depend heavily on its cellular abundance (Ikemura and Ozeki 1977, Ikemura, Osawa et al. 1980, Ikemura 1981, Ikemura 1982, Ikemura and Ozeki 1983, Ikemura 1985).

Finally, it was directly determined in *E. coli* that the rate-limiting step of elongation was tRNA selection, which nicely explained how the rate of elongation was variable over the length of an mRNA (Varenne, Buc et al. 1984). This finding was repeatedly confirmed and it was shown that the translation rate for a given codon specifically depended on the competition over recognition of the codon between the cognate amino acyl-tRNA—the correct one—and non- and near-cognate amino acyl-tRNAs (Akashi 2003, Fluitt, Pienaar et al. 2007, Man and Pilpel 2007). Thus, variation in abundances of different tRNAs, (Ikemura and Ozeki 1977, Dong, Nilsson et al. 1996, Kanaya, Yamada et al. 1999, Duret 2000), determined primarily by their gene copy numbers (Duret 2000, Reis, Savva et al. 2004, Ran and Higgs 2010, Iben and Maraia 2014), is responsible for the variation in translation rate for synonymous codons. Consistently, it was shown in mammalian cells that genome-wide rates of elongation were not constant across mRNAs, and that there were over 1,500 loci at which elongation paused (Ingolia, Lareau et al. 2011).

Regulation of elongation by fixed mRNA properties

Like initiation, the sequence and structure of an mRNA molecule influence its elongation rate. It has been known for decades that significant variation exists in codon usage and guanine-cytosine (GC) content across organisms as well as across genes within the same organism (Sueoka 1961). Perhaps unsurprisingly, the GC content of genomes and genes is highly related to codon usage, that is, genomes and genes with high GC content tend to have high representation of codons and amino acids that also utilize relatively more G and C, and vice-

versa for adenosine (A) and thymine (T). Indeed, models can very accurately predict codon usage for all amino acids in genomes and genes on the basis of GC content, in (Knight, Freeland et al. 2001). Despite this expected trend, the codons used by an mRNA molecule as well as their arrangement, for reasons discussed above, are important for elongation speed.

As it became clear that protein expression level, tRNA gene copy number, and codon usage were so closely related, Ikemura devised four rules that determine which codon is "optimal" for each amino acid (Ikemura 1985). First, the most abundant isoaccepting tRNA is preferred. Second, A-terminated codons are preferred over G-terminated codons. Third, uracil (U)- and C-terminated codons are preferred over A-terminated codons. And fourth, for (A/U)-(A/U)-(C/U) codons, C is preferred in the third position over U.

Sharp and Li proposed the codon adaptation index (CAI, Equations 1.1-1.4) as a measure of CUB for a given gene. CAI indicates how adapted an mRNA sequence is for optimal translation, based on its codon usage, compared to a reference set of highly expressed genes (Sharp and Li 1987). Calculation of CAI involves four steps.

1) The relative synonymous codon usage (RSCU) is determined for each codon from a reference set of highly expressed genes (Equation 1.1). RSCU is the relative occurrence of a codon in the reference set compared to all of the synonymous codons that encode the same amino acid.

$$RSCU_{ij} = \frac{x_{ij}}{\frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}}. \qquad (1.1)$$

x_{ij} is the number of occurrences of the j^{th} codon for the i^{th} amino acid in the reference set, and n_i is the number of synonymous codons encoding the i^{th} amino acid.

2) Second, the observed CAI (CAI_{obs}) is calculated, which is the geometric mean of the RSCU values for all codons of an mRNA sequence.

$$CAI_{obs} = \left(\prod_{k=1}^{L} RSCU_k \right)^{\frac{1}{L}}. \quad (1.2)$$

L is the number of codons in the mRNA sequence and k is the codon number or position in the gene sequence.

3) Third, the maximum CAI (CAI_{max}) is calculated, which is the maximum possible CAI for the given amino acid sequence, that is, if all codons were optimal.

$$CAI_{max} = \left(\prod_{k=1}^{L} RSCU_{kmax} \right)^{\frac{1}{L}}. \quad (1.3)$$

For each codon, $RSCU_{kmax}$ is the maximum RSCU of all of the synonymous codons coding for the same amino acid.

4) Lastly, CAI is calculated as the ratio of CAI_{obs} to CAI_{max}:

$$CAI = \frac{CAI_{obs}}{CAI_{max}}. \quad (1.4)$$

CAI, therefore, is the observed CAI for the gene relative to the maximum possible CAI for a gene with the same amino acid composition. Savva's group introduced the tRNA adaptation index (tAI), which was inspired by the CAI but instead quantifies how adapted a codon is to the tRNA usage in a genome (dos Reis, Wernisch et al. 2003, dos Reis, Savva et al. 2004). An optimal codon, according to the tAI, is one that uses the tRNA that has the highest gene copy number among all isoaccepting tRNAs and that has perfect codon-anticodon pairing.

One advantage of the tAI over the CAI is not having to select a reference set of highly expressed genes, as a given highly expressed gene may or may not use codons that are optimized for translation efficiency.

The CAI and more recently, the tAI, have been used in a number of studies as indicators of translation efficiency in local regions of mRNA. von der Haar's group used simulations and experimental approaches to show that movement of a ribosome past "slow" codons near the start codon of an mRNA could be rate-limiting for initiation (Chu, Kazana et al. 2013) and thus the overall process of translation for some genes. As mentioned above, the elongation rate has been shown by several groups to be variable across an mRNA molecule. Pilpel's group identified a universally conserved "ramp" of translation efficiency, based on the tAI (Cannarozzi, Schraudolph et al. 2010, Tuller, Carmi et al. 2010). In this ramp, the first 30-50 codons at the start of mRNAs had low translation efficiency, while the last approximately 50 codons had the highest translation efficiency of any location along the mRNA (Cannarozzi, Schraudolph et al. 2010, Tuller, Carmi et al. 2010). It was hypothesized that this slowing down of ribosomes early in the elongation process and speeding up farther downstream helps to prevent collisions and jamming among ribosomes along most of the transcript.

A number of groups have sought explanations for CUB. Shah and Gilchrist showed in yeast that the CUB in the mRNA sequences of highly expressed genes could be almost completely explained by variation in elongation times among synonymous codons, while for poorly expressed genes this extent of over-representation could be explained by biased mutation rates (Shah and Gilchrist 2011). This supports the hypothesis that CUB helped to direct evolution because it has a real effect on translation. Other studies have identified correlations, both positive and negative, between the length of mRNA and the extent of CUB in different

organisms (Moriyama and Powell 1997, Moriyama and Powell 1998, Coghlan and Wolfe 2000), which exist as the result of different selective pressures. As these authors discuss, in one organism, strong CUB may facilitate efficient synthesis of long proteins which are energetically expensive to make and potentially more prone to missense errors. In another organism, weak CUB may favor the production of many smaller proteins if they are often sufficient to do the job of larger proteins.

Interestingly, selection for pairs of codons occurring next to each other in *E. coli* appears to be quite different from eukaryotes. Heck's group took advantage of the fact that transcription and translation occur together in prokaryotes in order to empirically compare elongation rates among multiple pairs of codons in *E. coli* (Irwin, Heck et al. 1995). They created a gene expression construct in which if a ribosome proceeded quickly along the mRNA behind RNA polymerase, it would run into a stop codon quickly and dissociate from the mRNA, allowing an RNA stem-loop structure to form. However, if the ribosome proceeded slowly and therefore remained on the mRNA for a longer period of time, it would obstruct the formation of the stem-loop structure. The stem-loop structure was upstream of the structural genes of the *lac* operon, so a relatively fast elongation rate could be detected by increased β-galactosidase expression. Using this experimental approach, Heck's group demonstrated that highly utilized pairs of codons, occurring next to each other in mRNA sequences more often than expected by chance, were translated more slowly than pairs that were not over-represented. This was perhaps due to differences in compatibilities of pairs of tRNAs next to each other on a translating ribosome. Aside from differences between yeast and *E. coli*, it is unclear why over-represented single codons tend to be those that are translated faster, while over-represented neighboring pairs of codons tend to be those that are translated more slowly. Nevertheless, these observations suggest

that specific mRNA sequences have been selected not only to control the structure and function of their protein products, but also to influence the rates of peptide elongation.

One additional way in which the mRNA sequence can affect the elongation rate is in determining the charges of amino acids along the polypeptide emerging from the ribosome exit tunnel (Tenson and Ehrenberg 2002). The electrostatic potential of the inside of the ribosome exit tunnel is negative, so stretches of positively charged amino acids in the nascent polypeptide chain may slow down ribosome movement (Voss, Gerstein et al. 2006, Lu, Kobertz et al. 2007, Lu and Deutsch 2008). This hypothesis was supported by modeling work by Tuller et al. (Tuller, Veksler-Lublinsky et al. 2011). They showed using modeling in yeast and *E. coli* that negatively charged amino acids predicted to be in the ribosome exit tunnel, in addition to higher tAI and weaker mRNA folding energy, were correlated with faster elongation rates along specific segments of mRNAs (Tuller, Veksler-Lublinsky et al. 2011). Further, specific amino acid sequence motifs in the N-terminus of a polypeptide as it emerges from the ribosome exit tunnel have been shown to cause stalling of elongation near its C-terminus (Nakatogawa and Ito 2002). Using ribosome footprinting combined with a pulse-chase approach, Weissman's group identified over 1,500 pauses in elongation in the coding sequences of genes and 420 pauses at stop codons (Ingolia, Lareau et al. 2011). Taken together, these observations point to an important role of fixed properties of mRNA, especially sequence and structure, in determining the elongation rate for an mRNA molecule.

Regulation of elongation by dynamically regulated processes

In addition to fixed properties of mRNA, elongation can be regulated by dynamic processes. Darnell's group sequenced polysome-bound mRNAs that were also associated with

the FMRP protein in mouse brain cells (Darnell, Van Driesche et al. 2011). They demonstrated that the FMRP protein stalled ribosomes, through an unknown mechanism, and thereby slowed elongation, which they viewed as a sort of "translational brake" which protects against fragile X syndrome and autism-related disorders (Darnell, Van Driesche et al. 2011). Thus, the activity of the FMRP protein can modulate elongation.

Stress responses are a second example of elongation regulation by dynamic processes. A large body of evidence since the 1960s suggests that protein folding can begin co-translationally, that is, as the polypeptide is being synthesized (Zipser and Perrin 1963, Kiho and Rich 1964, Fedorov and Baldwin 1997, Komar 2009, Han, David et al. 2012, Wells, Bergendahl et al. 2015). Qian's group demonstrated that a protein folding inhibitor caused pausing of ribosome movement along mRNAs in human cells, which mimicked the effects of disrupting molecular chaperones by mutations and chemical inhibitors (Liu, Han et al. 2013). This suggested that molecular chaperones play a role in regulating elongation. In that study and others (Oh, Becker et al. 2011), the site at which ribosomes tend to pause along an mRNA normally (Han, David et al. 2012) and in response to stress (Oh, Becker et al. 2011) is 30-60 codons from the start codon. Since the ribosome exit tunnel is approximately 40 amino acids long, these observations are consistent with the idea that molecular chaperones function in stress response by sensing a stress signal and relaying it to the translation machinery through their effects on folding of nascent polypeptides that are just emerging from the ribosome exit tunnel. This may then lead to stalling of elongation and, consequently, as discussed above, initiation. Thus, environmental stresses, such as heat shock (Shalgi, Hurt et al. 2013) and oxidative stress (Gerashchenko, Lobanov et al. 2012), which are known to impact folding, have the potential to impact protein synthesis at multiple levels.

Translation regulation via mRNA degradation

While translation regulation at the initiation and elongation steps regulate the rate of translation from a given amount of mRNA, mRNA degradation impacts the abundance of mRNA (Wilusz, Wormington et al. 2001) and thus indirectly regulates translation. In eukaryotic cells, mRNAs that are targeted for destruction, as well as proteins involved in degradation can aggregate in foci called processing bodies (P-bodies) (Bashkirov, Scherthan et al. 1997, Sheth and Parker 2003, Cougot, Babajko et al. 2004, Teixeira, Sheth et al. 2005). Because of the central role of mRNA degradation in basic cell processes and its potent effects in living systems, mRNA degradation has been harnessed in the form of investigational tools and techniques and has even shown potential as a therapeutic target in humans (Zimmermann, Lee et al. 2006, Frank-Kamenetsky, Grefhorst et al. 2008). Like the other steps in translation, the influence of mRNA degradation on translation can be through fixed properties of an mRNA molecule or dynamically regulated processes.

Regulation of mRNA degradation by fixed mRNA properties

The half-life of mRNA varies widely across genes in a number of model systems, including mouse fibroblasts (Friedel, Dölken et al. 2009, Schwanhausser, Busse et al. 2011), human cell lines (Yang, van Nimwegen et al. 2003, Friedel, Dölken et al. 2009), *Bacillus subtilis* (Hambraeus, Wachenfeldt et al. 2003), yeast (Miller, Schwalb et al. 2011), and Arabidopsis (Narsai, Howell et al. 2007). Perhaps the most obvious and widely studied property of mRNA that affects its half-life is its sequence. An mRNA's sequence per se could theoretically influence its half-life. While mRNAs in a cell are generally complexed with proteins which protect them from degradative enzymes, the sequence of any region of an mRNA that is not

protected may influence the mRNA's half-life. In this sense, genome-wide variation in sequences of unprotected regions of mRNA may influence mRNA half-lives due to variation in their susceptibility to attacks by enzymes, as well as variation in their susceptibility to degradation due to oxidation, heat, acidity, and other factors. Consistently, mRNA half-lives are affected by oxygen levels (Klug 1991) and temperature (Goldenberg, Azar et al. 1996) in bacteria.

The more physiologically relevant way in which the sequence of an mRNA affects its half-life is through dictating the mRNA's interactions with proteins and cellular processes that affect its stability. Narsai et al.'s analysis in Arabidopsis mentioned previously (Narsai, Howell et al. 2007) revealed three mRNA sequence characteristics that were correlated with half-lives. First, genes with one or more introns have longer half-lives than genes with no introns, irrespective of mRNA length, number of introns, or nucleotide composition. Second, specific sequence motifs are enriched in the 3' UTRs of mRNAs with short half-lives, while other motifs are enriched in mRNAs with long half-lives. Third, mRNAs that are targets of microRNAs tend to have shorter half-lives than mRNAs that are not. This and other studies in plants (Abler and Green 1996) as well as mammals (Ross 1995), including humans (Shaw and Kamen 1986, Yang, van Nimwegen et al. 2003), have also revealed that AU-rich sequence motifs are enriched in the 3' UTRs of relatively unstable mRNAs. Zubiaga et al. found an extremely potent AU-rich motif, UUAUUUAUU, enriched in the 3' UTRs of unstable mRNAs in mouse fibroblasts, which accelerated deadenylation and degradation of mRNA (Zubiaga, Belasco et al. 1995). While various patterns and trends regarding the relationship between the sequence of mRNA and its decay rate have been identified, a thorough understanding is still lacking.

Regulation of mRNA degradation by dynamically regulated processes

While the half-life of mRNA is under the influence of fixed properties, it can also be modulated by dynamic processes. Like other steps in gene expression, mRNA stability is sensitive to a wide variety of stimuli from a cell's internal and external environment. For example, circadian clock associated 1 (*CCA1*) mRNA is more rapidly degraded in response to light, and this behavior is likely dictated by a sequence in its coding region (Yakir, Hilman et al. 2007). In yeast, in the presence of carbon sources that do not require mitochondrial function for their metabolism, the Puf3p protein promotes degradation of mRNA targets involved in mitochondria-mediated metabolism, while in the presence of carbon sources that do require mitochondrial function, this decay activity is suppressed (Miller, Russo et al. 2014). Numerous other examples of mRNA degradation being controlled by environmental cues in yeast (García-Martínez, Delgado-Ramos et al. 2015, Russo and Olivas 2015, Braun, Dombek et al. 2016) plants (Banerjee, Lin et al. 2009), and mammals (Müllner and Kühn 1988) have been reported. But how do diverse signals from a cell's internal and external environment lead to the targeting of specific mRNAs to these very generic destruction pathways? One major way is through signaling pathways that lead to RNA-binding proteins recognizing and binding specific mRNAs, usually via sequences within the 3' UTR. These bound proteins then recruit other proteins to the mRNA which facilitate deadenylation, decapping, and exonucleolytic digestion (Shim and Karin 2002), as in the Puf family of RNA-binding proteins (Olivas and Parker 2000, Ulbricht and Olivas 2008, Wang, Opperman et al. 2009, Miller and Olivas 2011).

A second important way for cells to target specific mRNAs for degradation is through RNA interference (RNAi), which is mediated by small interfering RNAs (siRNAs), as well as microRNAs (Bartel 2004, Filipowicz 2005, Bartel 2009). Plants (Rhoades, Reinhart et al. 2002,

Jones-Rhoades and Bartel 2004, Jones-Rhoades, Bartel et al. 2006) and animals (Selbach, Schwanhausser et al. 2008, Friedman, Farh et al. 2009) have hundreds of siRNAs and microRNAs, with many targeting large numbers of mRNAs, so RNAi and microRNAs have the potential to dramatically affect protein production for a wide variety of genes. Since mRNA degradation mediated by RNAi and microRNAs is a consequence of mRNA sequence, it can be considered a fixed property. However, since abundance of siRNAs and microRNAs is subject to regulation of their transcription and processing (Reinhart, Weinstein et al. 2002, Baskerville and Bartel 2005), in another sense, it is a dynamically regulated process.

In RNAi, ~20-bp siRNAs with high sequence complementarity to their target mRNAs are synthesized from double-stranded RNAs which are derived from DNA or RNA. Then, they are incorporated into RNA-induced silencing complexes (RISCs) which destroy the target mRNAs with the help of the Argonaute family of proteins (Liu, Carmell et al. 2004, Jones-Rhoades, Bartel et al. 2006). MicroRNAs are ~20-bp RNAs derived from genome-encoded RNA precursors. It was supposed that the predominant mode of action of microRNAs in decreasing protein levels was via translation repression in animals (Carrington and Ambros 2003, Pillai, Bhattacharyya et al. 2005, Baek, Villén et al. 2008) and mRNA degradation in plants (Bartel 2004, Baumberger and Baulcombe 2005), but more recent analyses suggest that mRNA degradation is the predominant mode of action even in animals (Guo, Ingolia et al. 2010). Nevertheless, there are interesting differences between animals and plants, with microRNAs tending to hybridize imperfectly with their target mRNAs in the 3' UTR in animals, but nearly perfectly with coding regions in plants (Carrington and Ambros 2003). In plants, microRNAs facilitate destruction of mRNA in at least two ways, including promotion of exonucleolytic digestion via the deadenylation-dependent degradation pathway (Wu, Fan et al. 2006, Rogers

and Chen 2013), as well as by slicing the mRNA somewhere in the middle via endonuclease activity by Argonaute proteins (Mi, Cai et al. 2008, Montgomery, Howell et al. 2008, Takeda, Iwasaki et al. 2008, Ji, Liu et al. 2011, Maunoury and Vaucheret 2011, Zhu, Hu et al. 2011). Whether facilitated by RNA-binding proteins, siRNA, or microRNAs, targeted mRNA destruction commonly occurs in P-bodies, although P-bodies are not required. Evidence from yeast suggests that P-bodies form as a result of widespread mRNA degradation in a cell as RNA-binding proteins from different mRNPs interact and aggregate into larger structures (Teixeira, Sheth et al. 2005, Teixeira and Parker 2007) (Eulalio, Behm-Ansmant et al. 2007).

Translation regulation via marking mRNA for degradation

In order for a cell to destroy mRNA, it must first identify which mRNAs to destroy and decide how to destroy them, a process that I refer to here as "marking". In this section, I focus on the predominant means of marking—the deadenylation-dependent pathway. As with the other major biochemical processes controlling ribosome loading of mRNA, I discuss inherent, fixed properties of mRNA that influence marking, in addition to dynamically regulated processes.

Deadenylation-dependent mRNA degradation

The predominant pathway in eukaryotes for marking mRNA for degradation is the deadenylation-dependent pathway, which consists of three steps, 1) deadenylation of the 3' poly(A) tail by deadenylases, 2) removal of the 5' cap by decapping enzymes, and 3) digestion from the ends of the mRNA by 5'→3' and 3'→5' exonucleases (Moore 2005, Goldstrohm and Wickens 2008, Moore and Proudfoot 2009, Chen and Shyu 2011). The poly(A) tail and PABP

protect the 3' end of mRNA from degradation, so their removal exposes the tail and allows 3'→5' exonucleolytic digestion (Wilusz, Gao et al. 2001, Mangus, Evans et al. 2003). However, the more important consequence of deadenylation is that the mRNA becomes a substrate for decapping enzymes, which remove the 5' cap and expose the mRNA to 5'→3' exonucleolytic digestion and reduce or abolish initiation (Wilson and Treisman 1988, Shyu, Belasco et al. 1991, Decker and Parker 1993, Hsu and Stevens 1993, Beelman, Stevens et al. 1996). Removal of the poly(A) tail disrupts the association between the 3' end and PABP, which may disrupt the closed loop and expose the 5' end to decapping enzymes. In the third and final step of deadenylation-dependent degradation, exonucleases digest the mRNA from the unprotected 5' end. As mentioned above, some marking pathways are closely tied to initiation.

mRNAs that are translationally repressed by loss of association with ribosomes and are associated with decapping machinery can assemble into P-bodies (Franks and Lykke-Andersen 2008, Chen and Shyu 2013). It appears that inhibition of initiation, not elongation, is the trigger for P-body formation, as the following three lines of evidence suggest. First, Teixeira, et al. reported that in yeast, conditions that cause widespread repression of translation initiation, including stress and mutations, caused an increase in the number and size of P-bodies, while inhibiting elongation by trapping ribosomes on mRNA caused disassembly of P-bodies (Teixeira, Sheth et al. 2005). Second, this group and others found proteins involved in mRNA degradation, mRNA surveillance, translation repression, and RNAi in P-bodies, but not ribosomal proteins or other proteins involved in initiation (Andrei, Ingelfinger et al. 2005, Teixeira, Sheth et al. 2005). Third, the average size of P-bodies in a cell was found to be correlated with the concentration of ribosome-free mRNA (Franks and Lykke-Andersen 2008). mRNAs in P-bodies can either be destroyed or returned to the translationally active pool

(Brengues, Teixeira et al. 2005). While it is clear that P-bodies are important sites of mRNA processing, the nature and extent of their cellular functions remain controversial, as questions remain about how common it is for mRNA in P-bodies to escape destruction (Arribere, Doudna et al. 2011) and how active P-bodies are in dictating the fate of mRNAs residing in them (Franks and Lykke-Andersen 2008).

Regulation of marking by fixed mRNA properties

Like the other biochemical processes that control ribosome loading of mRNA, marking can be influenced by fixed properties of mRNA as well as dynamically regulated processes. One example of a fixed property of mRNA influencing marking is the effect of the pre-mRNA sequence on polyadenylation. The sequence context around the poly(A) cleavage site at the 3' end of an mRNA is an important determinant of the poly(A) tail length (Tian, Hu et al. 2005, Sandberg, Neilson et al. 2008, Shen, Ji et al. 2008), as APA sites vary in their frequency of use (Proudfoot, Furger et al. 2002, Xing and Li 2011). Also, the position of the APA site that is used dictates how much of the pre-mRNA is included in the mature 3' UTR, in other words, whether various sequence elements (e.g., microRNA recognition sites) will be included. These properties influence marking and degradation of the mRNA. Further, alternative polyadenylation (APA) dictates the length and nucleotide sequence of the mature mRNA (Wu, Liu et al. 2011, Xing and Li 2011). For genes subject to APA, the sequence context at each poly(A) site in a pre-mRNA, as well as alternative splicing (Wachter, Tunc-Ozdemir et al. 2007), influence which site will be used and, as a result, also influence marking and degradation.

Regulation of marking mRNA for degradation by dynamically regulated properties

While the sequence of an mRNA—a fixed property—can influence its marking by dictating the length and position of the poly(A) tail, the predominant source of influence on marking is from dynamically regulated processes. As discussed above, dephosphorylation of the 4E-BPs causes inhibition of eIF4E. This suppresses recruitment of ribosomes to the 5' end, increases access of decapping enzymes (e.g., DCP1) to the 5' cap, and disrupts the closed loop facilitated by the eIF4F complex (Franks and Lykke-Andersen 2008, Arribas-Layton, Wu et al. 2013). These events, which often respond to stress, act on mRNAs with a 5' cap, and thus contribute to marking of mRNA for degradation in a global fashion. On the other hand, a cell triggers deadenylation of specific mRNAs through pathways that result in binding of proteins with those mRNAs, which they recognize by conserved sequence elements. Those RNA-binding proteins then recruit deadenylases. One example of this is the PUF family of RNA-binding proteins, which regulate cell proliferation, development, and signaling among neurons in various animal systems (Crittenden, Bernstein et al. 2002, Wickens, Bernstein et al. 2002, Menon, Sanyal et al. 2004). In yeast, PUF proteins directly interact with POP2, a CAF1-family protein and component of the CCR4-CAF1-NOT complex. In response to upstream cues, PUF proteins bind to specific mRNAs by recognizing a specific short sequence element and recruit the CCR4-CAF1-NOT deadenylase complex to the mRNA. As a result of PUF proteins' interaction with POP2, the yeast decapping enzymes DCP1 and DHH1 are subsequently recruited to the mRNA (Goldstrohm, Hook et al. 2006).

Regulation of translation by the circadian clock and other factors

In this section, I provide context for chapter 2. First, I summarize what has been learned from genome-wide polysome profiling in Arabidopsis. Second, I give a basic overview of the Arabidopsis circadian clock. And third, I summarize what was previously known about diel and circadian regulation of translation, which provides rationale for the hypothesis tested in chapter 2, that the circadian clock influences genome-wide ribosome loading in Arabidopsis.

Overview of polysome profiling experiments in Arabidopsis

Progress has been made in understanding how ribosome loading and translation are regulated by many molecular events. However, key questions are still being asked about how the translation machinery senses cues from within the cell and its external environment and responds to them in order to grow and develop and to maintain homeostasis. A number of groups have measured genome-wide translation states (TLs) by polysome profiling in order to investigate how translation is influenced by various environmental conditions and genetic backgrounds. From these experiments in Arabidopsis, which are summarized in Table 1.1, several trends have become clear.

Table 1.1: Public genome-wide polysome profiling data sets from Arabidopsis

Research group	Experimental backgrounds	Growth conditions	References
Whitham lab, Iowa State University, Ames, IA	- Turnip mosaic virus - Control - Genotypes: Col-O RPL18 overexpression	- Age: 4 weeks - Soil/media: LC soil - Day length: 12h - Temperature: 22°C	(Moeller, Moscou et al. 2012)
Wu lab, Academica Simica, Taipei, Taiwan	- Dark - 0.5h light - 4h light - Genotypes: Col-O	- Age: 4 days - Soil/media: Half-strength MS media - Day length: NA - Temperature: 22°C	(Liu, Wu et al. 2012)
Bailey-Serres lab, University of California, Riverside, CA	- 12h hypoxia - Control - Genotypes: Ler	- Age: 7 days - Soil/media: MS media - Day length: 16h - Temperature: 20°C	(Branco-Price, Kawaguchi et al. 2005)
Bailey-Serres lab, University of California, Riverside, CA	- 2h, 9h hypoxia - 2h, 9h control - 9h hypoxia/1h recovery - Genotypes: Col-O RPL18 overexpression	- Age: 7 days - Soil/media: MS media - Day length: 16h - Temperature: 23°C	(Branco-Price, Kaiser et al. 2008)
Bailey-Serres lab, University of California, Riverside, CA	3 genotypes (Col-O background) - RPL18 overexpression - CSP1 overexpression - CSP1 knockdown 2 conditions: cold, control	- Age: 10 days - Soil/media: MS media - Day length: 16h - Temperature: 23°C	(Juntawong, Sorenson et al. 2013)
Bailey-Serres lab, University of California, Riverside, CA	- 1h light - 1h dark - 1h dark/10min re-illumination - Genotypes: Col-O	- Age: 14 days - Soil/media: MS media - Day length: 16h - Temperature: 23°C	(Juntawong and Bailey-Serres 2012)
Von Arnim lab, University of Tennessee, Knoxville, TN	- Wild type (Col-O) - eIF3h mutant (Ws)	- Age: 10 days - Soil/media: MS media - Day length: Continuous light - Temperature: 22°C	(Kim, Cai et al. 2007)
Von Arnim lab, University of Tennessee, Knoxville, TN	- Wild type (Col-O) - Rpl24b mutant (Ws)	- Age: 10 days - Soil/media: MS media - Day length: 16h - Temperature: 22°C	(Tiruneh, Kim et al. 2013)
Von Arnim lab, University of Tennessee, Knoxville, TN	2 genotypes (Col-O background) - Wild type - CCA1 overexpression 4 time points: 6am, 12pm, 6pm, 12am	- Age: 10 days - Soil/media: MS media - Day length: 16h - Temperature: 22°C	(Missra, Ernest et al. 2015)
Castellano lab, INIA-UPM, Madrid, Spain	- Heat - Control - Genotypes: Col-O	- Age: 7 days - Soil/media: MS media - Day length: 16h - Temperature: 22°C	(Yángüez, Castro-Sanz et al. 2013)

First, TL can be a more responsive state of gene expression than mRNA transcript abundance. The Bailey-Serres group investigated the genome-wide responses to hypoxia at the transcriptional and translational levels (Branco-Price, Kaiser et al. 2008). They measured mRNA levels and TLs by microarray in plants maintained in normal conditions for 2 or 9 hours, hypoxia for 2 or 9 hours, and hypoxia for 9 hours followed by a 1-hour recovery. I obtained their published microarray data and created a heatmap which shows the genome-wide transcript and TL profiles for each experimental group, shown in Figure 1.1. The genes and samples were clustered using hierarchical clustering based on the Pearson coefficient. Based on the clustering pattern and visual inspection, the transcriptional profiles from the 2-hour hypoxia group most closely resemble the two control groups, suggesting that 2 hours of hypoxia was not enough time for the plants to have a dramatic genome-wide response at the transcriptional level. However, the 9-hour hypoxia group clusters separately from the control groups, suggesting that 9 hours of hypoxia was long enough. The 9-hour-hypoxia/1-hour-recovery plants most closely resemble the 9-hour hypoxia plants, suggesting that 1 hour of recovery is not long enough for the plants to return to normal in terms of their genome-wide transcriptional profiles. Looking at the translational level gives a different view. Both hypoxia groups cluster completely separately from the control groups, suggesting that 2 hours of hypoxia is enough time for the plants to have a dramatic response in their genome-wide TL profiles. Importantly, though, the 9-hour-hypoxia/1-hour-recovery plants more closely resemble the control plants than the hypoxia-treated plants, suggesting that 1 hour of recovery was enough time for the plants to return back to normal in terms of their genome-wide TL profiles.

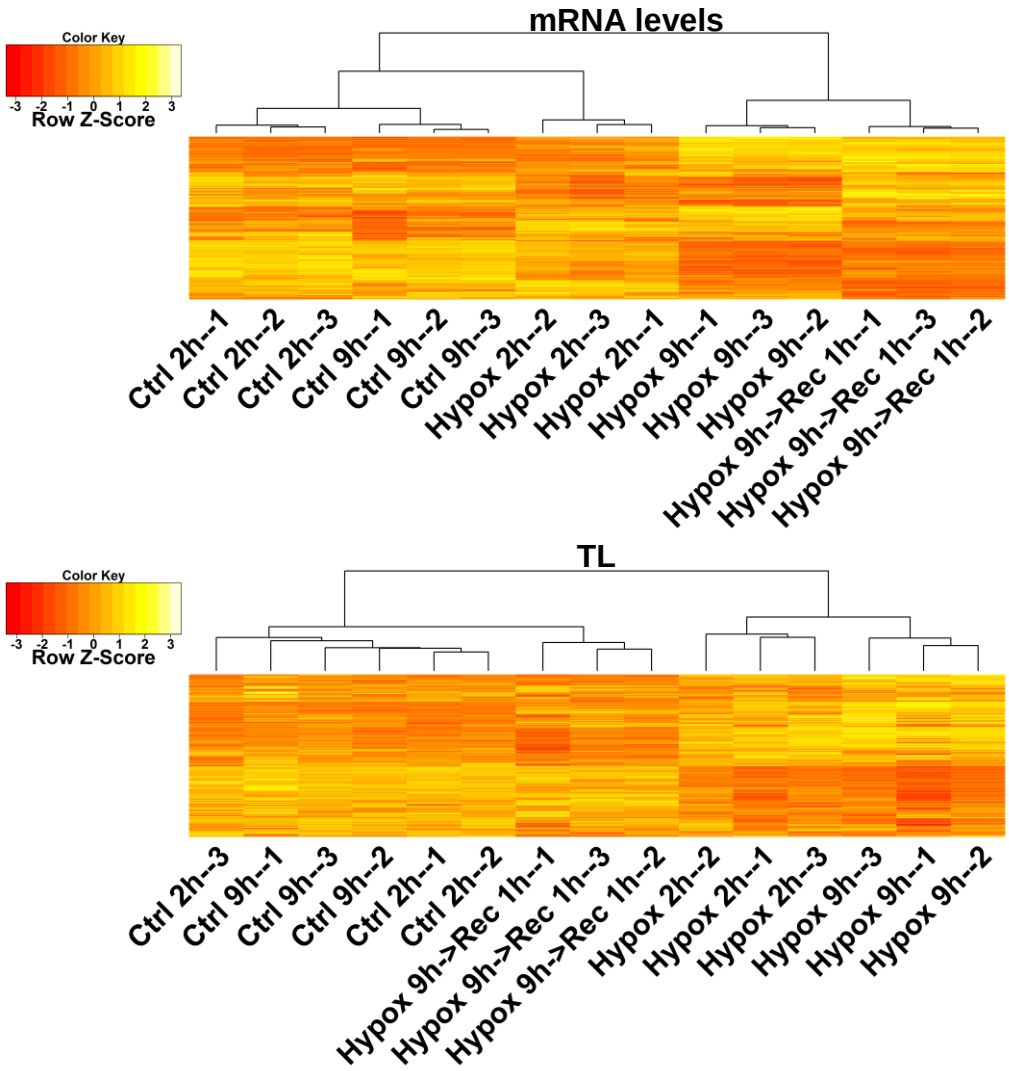

Figure 1.1: Transcriptional and TL profiles in response to hypoxia.
The data are from the Bailey-Serres hypoxia data set (Branco-Price, Kaiser et al. 2008). The experimental groups include control conditions for 2 (Ctrl 2h) or 9 hours (Ctrl 9h), hypoxia for 2 (Hypox 2h) or 9 hours (Hypox 9h), and 9 hours of hypoxia followed by a 1-hour recovery (Hypox 9h->Rec 1h). The numbers to the right of the dashes indicate biological replicates from the same experimental group. Both the genes and the samples were clustered using hierarchical clustering based on the Pearson coefficient. The colors reflect the Z-scores, which are the number of standard deviations each value is above or below the mean of its row.

Second, like the processes of initiation and degradation, TL regulation can be a global event, affecting most mRNA species, or mRNA-specific, targeting a few specific ones. I performed analysis of variance (ANOVA) on the same 2008 Branco-Price TL data, treating each experimental condition as a different treatment group and TL as the response variable. From this I obtained p-values and retained the 2000 genes with the lowest p-values, which represent genes whose TLs were the most strongly influenced by the experimental conditions. From these 2000 genes, I created a heatmap as before, which is shown in Figure 1.2. From this it is clear that many genes respond in a co-regulated way, either up- or down-regulated at the TL level in response to hypoxia. However, there are smaller groups of genes that do not follow the major trends. For instance, a few genes have different TLs between the 2-hour control and 9-hour control groups, perhaps due to minor differences in growth conditions at those two different time points. Additionally, while many genes behave the same between the controls and 9-hour-

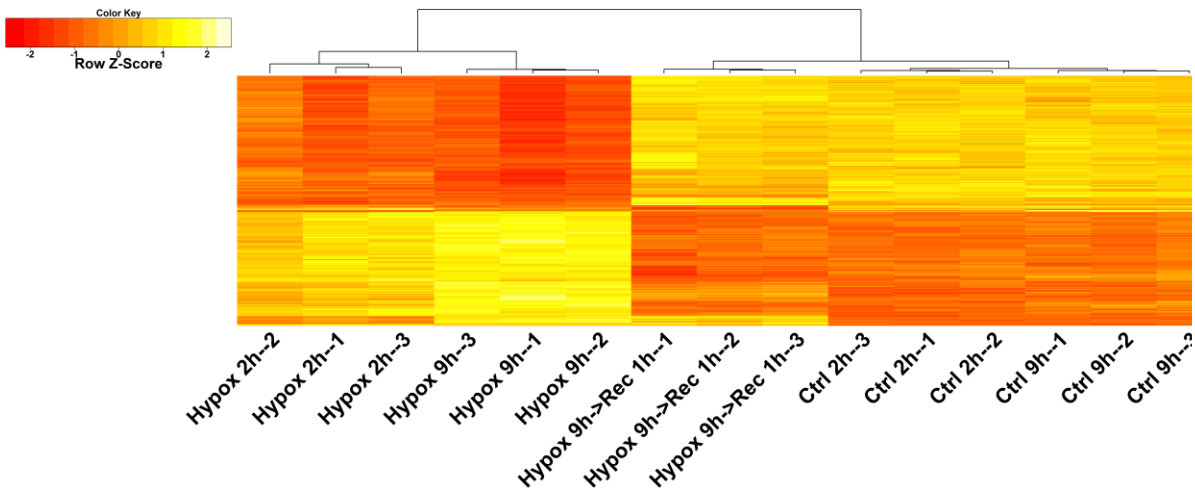

Figure 1.2: TL profiles for hypoxia-sensitive genes.
ANOVA p-values were obtained reflecting the effect of the experimental conditions on TL, and 2000 genes with the lowest p-values are shown here. The colors reflect Z-scores (see Figure 1.1).

hypoxia/1-hour-recovery groups, small groups of genes remained up- or down-regulated compared to the controls after 9 hours of hypoxia.

Third, TL is influenced by a wide variety of environmental conditions and genetic backgrounds, including temperature, oxygen levels, water, light, virus infection, genotype, and ecotype. Close analysis of each of the data sets by the original authors uncovered genes that respond at the TL level to each experimental condition, as well as groups of genes that share similar TL behavior and biological functions. Figure 1.3 shows a heatmap that I made from all of the TL data sets in Table 1.1, except the 2005 Branco-Price data and the CCA1-ox genotype from the von Arnim data, totaling 9 data sets, 29 experimental backgrounds (treating controls or wild type from different experiments as different experimental backgrounds), and 72 individual samples. As the authors of each study have found, there are many genes that respond at the TL level to the experimental treatments and conditions in each study. However, there are also clear signatures in TL profiles that are unique to each data set and distinct from other data sets, independent of the given experimental treatments in each study. This could reflect differences in growth conditions used in different laboratories, plant age, time of day, or other factors which have real biological effects on ribosome loading, as well as different protocols used to prepare the RNA samples, which do not truly affect ribosome loading. Further analysis of these meta-data, for instance, modeling the effects of variables, such as ecotype, age, soil composition, sucrose concentration, and light intensity, on TL might help to explain these signatures in TL profiles that are unique to each data set.

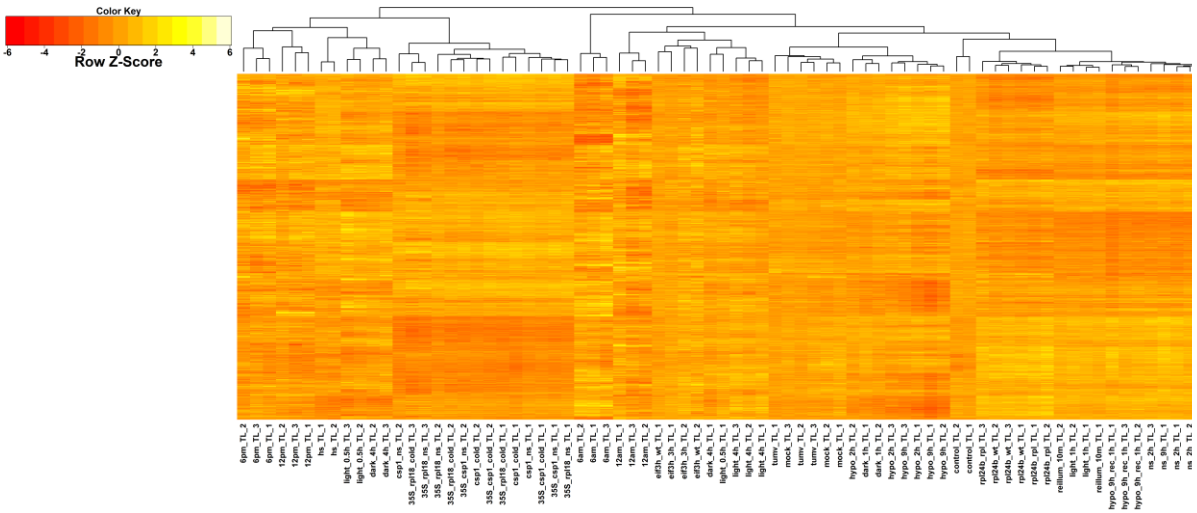

Figure 1.3: TL profiles for all available data from Arabidopsis.
The heatmap was made using the public data sets in Table 1.1, with hierarchical clustering used to cluster the genes and samples based on the Pearson coefficient. The colors reflect Z-scores (see Figure 1.1).

The circadian clock in Arabidopsis

The circadian clock is a physiological system in many organisms that helps them to carry out certain activities at the appropriate times (Hsu and Harmer 2014). In a broad sense, the circadian clock is composed of three main parts, the central oscillator, input pathways, and output pathways. The central oscillator is the core of the clock, as it sustains its cyclical nature and can operate without external cues. Input pathways help to set the clock and establish its timing. Output pathways connect the clock timing with the physiological processes that are to occur at specific times.

In plants, the central oscillator is largely set by the light-dark cycle through input pathways. It consists of three sets of genes that repress each other in a cycle over the course of the day. They do this through their protein products, which either directly repress transcription

of their target genes or promote transcription of target genes which then lead to repression of other genes in the central oscillator. One set of genes, the "morning genes", are expressed highly in the morning and repress the evening genes. The day genes are expressed highly during the day and repress the morning genes, and the evening genes are expressed highly during the evening and repress the day genes. The genes in the central oscillator not only regulate each other, but they also regulate genes outside of the central oscillator through output pathways, enabling plants to carry out physiological processes at the appropriate times of day.

Input pathways help to entrain the clock based on environmental cues. The light-dark cycle is one widely studied environmental cue that is known to affect the expression of central oscillator genes at multiple levels, including transcription, mRNA stability, and translation. For instance, the ZEITLUPE (ZTL) protein, encoded by the *ZTL* gene, responds to light during the day by binding with increased affinity to GIGANTEA (GI). However, as light intensity decreases at dusk, ZTL loses its affinity for GI and targets TIMING OF CAB EXPRESSION 1 (TOC1) and PSEUDO-RESPONSE REGULATOR 5 (PRR5) proteins for degradation via the ubiquitin-proteasome pathway (Más, Kim et al. 2003). Phytochrome proteins sense red and far-red light, and cryptochrome proteins sense blue light, and both types of proteins help to set the speed of the clock (Strasser, Sánchez-Lamas et al. 2010, Hu, Franklin et al. 2013). Along with the light-dark cycle, temperature is also a widely studied factor that affects the clock via input pathways. Like the light-dark cycle, temperature affects the clock at multiple levels of gene expression. For instance, temperature affects the alternative splicing of several clock genes, including *CCA1*, *LHY*, *PRR9*, *PRR7*, *PRR5*, *PRR3*, and *TOC1*, with splicing of *CCA1* into its full-length vs. truncated variants implicated in cold resistance (James, Syed et al. 2012).

Output pathways connect the clock with the physiological systems that it regulates, particularly growth, metabolism, and response to abiotic and biotic stresses. The hypocotyl, for example, exhibits a daily rhythm in growth that is influenced by the clock as well as light, sucrose, and hormonal signaling. Many genes involved in photosynthesis and carbohydrate metabolism are regulated by the clock (Michael, Mockler et al. 2008). Photosynthesis results in a daily accumulation of starch during the day, with 95% of it consumed by the plant by the following dawn (Graf, Schlereth et al. 2010) at a rate that the plant determines by taking stock of the amount of starch accumulated and the anticipated length of the dark period (Pal, Liput et al. 2013, Scialdone, Mugford et al. 2013). Lastly, the clock helps plants to respond to abiotic and biotic stresses. A classic example of circadian regulation of responses to abiotic stresses is cold resistance (Eriksson and Webb 2011). C-REPEAT BINDING FACTOR 1 (CBF1), CBF2, and CBF3 are clock-regulated transcription factors that promote expression of genes involved in cold resistance, which lead to increased levels of cryoprotectant molecules. Expression of these *CBF* genes peaks at midday under the direction of the clock, and it has been hypothesized that this behavior allows plants to "interpret" cold temperatures during the day as a sign of seasonal changes, but not cold temperatures at night, which do not signify seasonal changes (Robertson, Skeffington et al. 2009).

The Kay group identified a number of additional examples of physiological systems influenced by the clock via output pathways in their study of genome-wide transcriptional regulation by the clock in Arabidopsis (Harmer, Hogenesch et al. 2000). Their analysis indicated that 23 genes that had circadian cycles of transcription and reached their peak transcript levels before dawn encoded enzymes in the phenylpropanoid biosynthetic pathway (Harmer, Hogenesch et al. 2000). This suggested that the circadian clock helps plants to produce pigments

early in the day that protect them from ultraviolet light, which was consistent with previous reports (Li, Ou-Lee et al. 1993, Landry, Chapple et al. 1995). Another clock-regulated output pathway suggested by the Kay group's analysis was cold resistance. Changes in the fatty acid composition of cotton seedlings was previously found to have a circadian pattern that correlated with cold resistance (Rikin, Dillwith et al. 1993). The Kay group found circadian patterns in the expression of lipid desaturase enzymes, which peaked just before dusk, as well as transcription factors and other genes involved in chilling resistance (Harmer, Hogenesch et al. 2000). A number of genes involved in other physiological processes in plants that are especially important at specific times of the day, such as the light-harvesting reactions of photosynthesis, starch metabolism, and sucrose metabolism, were also transcribed in a circadian fashion over the course of the day.

Diel and circadian regulation of translation

Translation has been known to play a role in the circadian clock for decades. In the 1970s, Jacklet found that applying anisomycin, a drug that inhibits protein synthesis by interfering with the ribosome, to the eye of the sea slug (*Aplysia californica*) shifted the phase of its circadian pattern of action potentials (Jacklet 1977). This provided one of the first key pieces of evidence that translation of new proteins played a role in the circadian clock in any organism. In the 1990s, however, using Northern blotting and other pre-polymerase chain reaction (PCR) techniques, researchers uncovered a system of transcriptional regulation that was critical in clock function (Aronson, Johnson et al. 1994). With technological innovations such as PCR and microarrays, which greatly accelerated the study of transcription, transcriptional regulation became the focus of research on regulation of the circadian clock (Harmer, Hogenesch et al.

2000). Thus, the role of translation regulation, particularly at the genome scale, in the circadian clock has received much less attention.

While the role of translation in the circadian clock has not been thoroughly characterized in any organism to date, it has been shown that translation in plants can fluctuate in response to daily external cues. The Stitt group monitored changes in ribosome loading of total cellular mRNA over a light-dark cycle in Arabidopsis (Pal, Liput et al. 2013). They found that the percent of total mRNA associated with large polysomes increased in response to light exposure and decreased after the lights were turned off again, while the percent of mRNA not associated with ribosomes did the opposite. This suggested that ribosome loading of mRNA in Arabidopsis is sensitive to fluctuations in light which occur over the course of a day.

Given the range of physiological processes that are regulated by the circadian clock in plants, the range of physiological processes controlled in part at the level of translation, the established link between the clock and translation, and the reports of post-transcriptional gene regulation by the circadian clock, it is not difficult to imagine that translation might play an important role in the circadian clock in plants. In chapter 2, I present a research article from the von Arnim lab in which genome-wide changes in transcript levels and ribosome loading of mRNA under the direction of the circadian clock were characterized in Arabidopsis. From the experimental work by a former post-doctoral fellow, Dr. Anamika Missra, and my data analysis, we made several interesting and novel findings. First, we demonstrated for the first time that translation is partially controlled by the circadian clock in plants. Second, translation for ribosomal proteins peaked late at night, suggesting that the clock may help plants to efficiently utilize “spare capacity” for translation at a time of day when many other proteins are not needed. Third, the clock can actually suppress daily translation cycles and uncouple translation from

transcription, adding sophistication to gene regulation. And lastly, the clock helps many genes to adjust transcription in anticipation of daily changes in light. While this work suggests that clock output pathways influence translation, it does not address the mechanism. Given the range of physiological processes that are regulated by clock output pathways, some of which are closely tied with translation, such as sucrose metabolism, it is unclear whether translation is regulated directly by clock output pathways, or indirectly.

Overview of dissertation

In this first chapter, I have reviewed the process of translation and the principles of its regulation and provided context for chapter 2. Chapter 2 is a research article describing the characterization of translational control by the circadian clock in Arabidopsis (Missra, Ernest et al. 2015). In chapter 3, I present a collaborative project with Dr. Michael Gilchrist at The University of Tennessee that involved testing and improving a computational model of translation that he developed with biological insight from Dr. von Arnim. The model describes how ribosome loading of mRNA is controlled by the rates of several biochemical steps in the translation process which were described in this chapter.

Chapter 2

The circadian clock modulates global daily cycles of mRNA ribosome loading

The content of this chapter is published in The Plant Cell in September 2015 (Missra, Ernest et al. 2015). I contributed to this article as a co-first author by designing and performing the bulk of the computational analyses. Raw data are available from the Gene Expression Omnibus (GEO) repository at http://www.ncbi.nlm.nih.gov/geo/ with accession number GSE61899. Supplemental datasets and figures not included in this chapter are available online from The Plant Cell at http://www.plantcell.org/content/27/9/2582.

Abstract

Circadian control of gene expression is well-characterized at the transcriptional level, but little is known about diel or circadian control of translation. Genome-wide translation state profiling of mRNAs in *Arabidopsis thaliana* seedlings grown in long day was performed to estimate ribosome loading per mRNA. The experiments revealed extensive translational regulation of key biological processes. Notably, translation of mRNAs for ribosomal proteins and mitochondrial respiration peaked at night. Central clock mRNAs are among those subject to fluctuations in ribosome loading. There was no consistent phase relationship between peak translation states and peak transcript levels. The overlay of distinct transcriptional and translational cycles can be expected to alter the waveform of the protein synthesis rate. Plants that constitutively overexpress the clock gene *CCA1* showed phase shifts in peak translation, with a six hour delay from midnight to dawn or from noon to evening being particularly common. Moreover, cycles of ribosome loading that were detected under continuous light in the wild type collapsed in the CCA1 overexpressor. Finally, at the transcript level, the CCA1-ox strain adopted a global pattern of transcript abundance that was broadly correlated with the light-

dark environment. Altogether, these data demonstrate that gene-specific, diel cycles of ribosome loading are controlled in part by the circadian clock.

Introduction

To adjust to a changing environment over the course of a day, plant cells regulate gene expression in a diel context. In *Arabidopsis*, for example, about one third of all genes are transcribed under the direction of the circadian clock (Covington, Maloof et al. 2008). Since most of the energy required for gene expression is spent on translation, it is plausible that translation itself may also be dielly regulated in order for plants to respond to the environment in an energy-efficient manner. Indeed, *Arabidopsis* plants undergo global cycles of ribosome loading (Pal, Liput et al. 2013) but not ribosome abundance (Piques, Schulze et al. 2009) over the course of the light-dark cycle.

The translation state (TL) of an mRNA is often estimated from its ribosome loading. The more ribosomes bind to an mRNA, the more efficiently it is translated (Mathews et al., 2007)(Mathews, Sonenberg et al. 2007). mRNA-ribosome complexes (polysomes) can be fractionated according to the number of ribosomes, and the proportion of the mRNA pool that resides in the different fractions can be used to determine TL. As defined here, TL is independent of the mRNA transcript level. That is, an mRNA's TL value will be the same between two different samples, as long as the average number of ribosomes per mRNA molecule is the same, even if the total amount of the mRNA differs between the two samples. TL for many Arabidopsis mRNAs is sensitive to a variety of environmental conditions, including hypoxia, heavy metal, drought, sugar, virus, heat, and light exposure, as well as various genetic backgrounds (Kawaguchi, Girke et al. 2004, Nicolai, Roncato et al. 2006, Kim, Cai et al. 2007,

Branco-Price, Kaiser et al. 2008, Juntawong and Bailey-Serres 2012, Liu, Wu et al. 2012, Moeller, Moscou et al. 2012, Tiruneh, Kim et al. 2013, Yanguez, Castro-Sanz et al. 2013). Extensive co-regulation of TL has been reported, for example, for ribosomal protein mRNAs (Kawaguchi, Girke et al. 2004, Kim, Cai et al. 2007, Juntawong and Bailey-Serres 2012, Tiruneh, Kim et al. 2013), suggesting that the transcriptome is organized into regulons of translational control.

The central oscillator of the circadian clock in *Arabidopsis* is based on a group of interlocked transcriptional feedback loops. At the core of the oscillator are three groups of genes that regulate gene expression in a cyclical fashion throughout the day by mutual transcriptional repression (Nagel and Kay 2012, Pokhilko, Fernandez et al. 2012). Transcripts for evening genes (e.g. *TOC1*, *LUX*, *ELF3*, and *ELF4*) peak late in the day and repress day genes (e.g. *PRR5*, *PRR7*, and *PRR9*). Day genes repress morning genes, and morning genes (e.g. *CCA1* and *LHY*) repress evening genes. Constitutive overexpression of the morning gene, *CCA1*, disrupts normal clock function (Wang and Tobin 1998, Green, Tingay et al. 2002). Under continuous light, the clock of the CCA1-overexpressor strain (CCA1-ox) is disordered and arrhythmic, as many central clock genes and clock output mRNAs are continuously expressed, while the partner of CCA1, LHY, is continuously repressed (Wang and Tobin 1998, Matsushika, Makino et al. 2002). In contrast, under light-dark cycle conditions, mRNAs for several central clock genes and clock outputs continue to cycle in the CCA1-ox strain (Matsushika, Makino et al. 2002). Clock genes such as *LHY* and *CCR2*/*GRP7* still respond to light in CCA1-ox but typically do not anticipate the dark-to-light transition, in keeping with the defect in the clock (Green, Tingay et al. 2002).

Early studies established that translation of new proteins plays a fundamental role in the operation of the circadian clock (Jacklet 1977, Nakashima, Perlman et al. 1981). However, subsequent investigations identified a robust mechanism of transcriptional control at the core of several circadian clocks (Hardin, Hall et al. 1992, Aronson, Johnson et al. 1994, Sehgal, Rothenfluh-Hilfiker et al. 1995, Schaffer, Ramsay et al. 1998, Wang and Tobin 1998, Strayer, Oyama et al. 2000). Thus, the role of translational control in circadian clock function has not been studied in detail at the genome level. In recent years, post-transcriptional control of diel and circadian gene expression has attracted significant attention. In particular, many clock mRNAs are alternatively spliced (Staiger, Zecca et al. 2003, Staiger and Green 2011, Filichkin and Mockler 2012, Park, Seo et al. 2012) and this must be regulated for proper clock function (Sanchez, Petrillo et al. 2010, Jones, Williams et al. 2012). Alternative splicing has also been implicated in temperature compensation of the clock (James, Syed et al. 2012, James, Syed et al. 2012, Seo, Park et al. 2012, Kwon, Park et al. 2014). By comparison, control of diel gene expression at the translational level has received comparatively little attention (Kim, Song et al. 2003).

Here, we have characterized translational control over the course of the diel light-dark cycle by measuring the ribosome loading of mRNAs in ten-day-old *Arabidopsis* seedlings grown in a long day. Approximately one in seven mRNAs are subject to robust diel cycles of ribosome loading. These cycles are partially controlled by the circadian clock, given that the translation cycles are substantially remodeled in the CCA1-ox strain. Diel and circadian translational control are particularly common among mRNAs for ribosome biogenesis, the inner mitochondrial membrane, and the photosynthetic apparatus. In summary, we provide the first genome-wide characterization of circadian control of gene-specific translation in a plant.

Results

Polysome loading over a diel cycle

We monitored polysome loading in 10-day-old wild-type (WT) *Arabidopsis* seedlings over a 16-hour light, 8-hour dark cycle at 6am (Zeitgeber time ZT0), 12pm (ZT6), 6pm (ZT12), 12am (ZT18), and again at 6am (ZT24). RNA was fractionated into non-polysomal (NP), small polysomal (SP) and large polysomal (LP) fractions using sucrose density centrifugation. We quantified polysome loading as the fraction of RNA found in SP and LP fractions relative to the total, which also includes NP RNA: (SP+LP)/(NP+SP+LP). In WT, polysome loading began at its lowest level at dawn, 6am (ZT0), peaked during the day, remained elevated through 12am (ZT18), and returned to low levels again the next dawn (Figure 2.1A, B). These data, obtained in seedlings grown in long day on artificial medium with 1% sucrose, follow a similar pattern as those from vegetative rosettes grown in a 12-hour light-dark cycle on soil (Pal et al., 2013), although the drop in translation towards dawn was less pronounced in our experiments. In the CCA1-ox strain, which has a disrupted circadian clock due to constitutive overexpression of *CCA1*, the pattern was similar to WT, but less dramatic (Figure 2.1A, C). Invoking the well-supported notion that ribosome loading reflects the rate of translation initiation (Mathews et al., 2007), these data suggested that diel control of translation may depend on a functional clock.

In order to obtain gene-specific ribosome loading data over the diel cycle, the mRNAs in the NP, SP, LP, and total (TX) RNA fractions were quantified by microarray hybridization at 6am (ZT0), 12pm (ZT6), 6pm (ZT12), and 12am (ZT18). A translation state (TL) was calculated for each mRNA: TL=(2xSP+7xLP)/(NP+SP+LP). SP and LP fractions were weighted by 2 and 7, respectively, because mRNA molecules are estimated to be bound by two and seven

ribosomes in these fractions, on average (Supplemental Figure 1). TL values were calculated for those 12,342 nuclear-encoded genes that were reliably detected in the SP and LP fractions at all four time points in all three replicates (Supplemental Dataset 1). ΔTL is defined as the TL at the peak minus the TL at the trough. Genes with varying TL across the diel cycle were first identified by significance analysis of microarrays, or SAM (Tusher, Tibshirani et al. 2001). Using SAM, 1825 mRNAs (15% of 12,342) varied in their TL value across time points at a collective false discovery rate (FDR) of 10% (Figure 2.2A and C**)**. In lieu of a gene-wise FDR, which SAM does not provide, we calculated an empirical permutation-based p-value, which confirmed that, besides a large fraction of strong translation cycles, many of the moderate translation cycles ($0.3 < \Delta TL < 0.7$) were statistically significant (Figure 2.2E). The TL of the majority of mRNAs peaked at noon (ZT6) or midnight (ZT18); peaks at dawn (ZT0) or in the evening (ZT12) were less common. More frequently than not, peak and trough were offset by 12 hours (Figure 2.2A).

Translation states of six representative genes were also analyzed by quantitative real time (reverse transcriptase) PCR (qRT-PCR; Supplemental Figure 2) as an independent technique, by calculating TL from the levels of transcripts present in the NP, SP, and LP fractions. Evidently, qRT-PCR and microarray results showed similar trends. The qRT-PCR data also confirmed that the changes in TL reflect diel cycles rather than monotonic trends over developmental time.

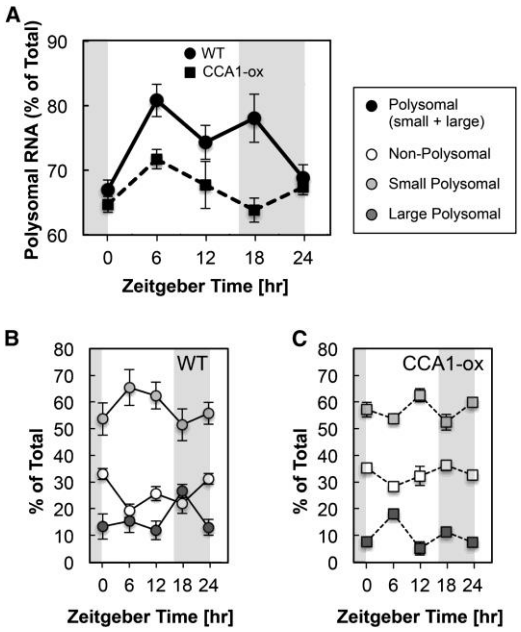

Figure 2.1: Polysome loading over a diel cycle.
Ten-day-old wild-type and CCA1-ox seedlings were grown in a 16-h-light/8-h-dark cycle. Tissue extracts from five time points harvested every 6 h were individually subjected to polysome density gradient centrifugation. The fraction of total RNA recovered in polysomal fractions (small and large polysomes) is plotted in **(A)** as a percentage of total RNA (nonpolysomes and small and large polysomes shown in **[B]** and **[C]**) at each ZT time. Error bars show standard deviations from three biological replicates. The difference in polysome loading between the wild type and CCA1-ox was significant by unpaired two-tailed t test for ZT6 ($P = 0.0058$) and ZT18 ($P = 0.0041$). The elevated polysome loading compared with ZT0 was significant in the wild type for ZT6, ZT12, and ZT18 and in CCA1-ox for ZT6.

Figure 2.2: Diel changes in ribosome loading of Arabidopsis mRNAs.
TL is an estimate of ribosome numbers per mRNA, and ΔTL is the difference between the highest and lowest TL for a gene, averaged over replicate samples. (A), (C), (E), (G), and (I) are the wild type. (B), (D), (F), (H), and (J) are CCA1-ox.
(A) and **(B)** mRNAs were filtered for a translational cycle using SAM. The mRNAs were first sorted into four predefined clusters according to the time of peak TL. Each cluster was then subdivided according to the time of the TL trough. Data were displayed using the heatmap.2 function from the gplots package in R. The number of mRNAs per cluster is given on the right.
(C) and **(D)** Distribution of all ΔTL values (black trace). ΔTL values were binned in increments of 0.1. The subset of cycling genes that were selected by ANOVA (AOV $P < 0.05$) (red trace) or SAM (10% FDR) (blue trace) is also illustrated. **(E)** and **(F)** Relationship between ΔTL and SAM P value. We used the test statistic computed by SAM and its permutation-based null distribution to compute an empirical P value for each prefiltered gene. Here, we plotted the negative log of this P value versus the ΔTL for each gene. The 2437 genes with $P < 0.05$ lie to the right of the broken line.
(G) and **(H)** Venn diagram showing the overlap among the four classes of differentially translated genes.
(I) and **(J)** Relationship between raw ANOVA P values from comparing average TL across time points and the corresponding adjusted P values using the Benjamini-Hochberg method. For the wild type and CCA1-ox, a raw P value of 0.05 (red vertical line) corresponded to a FDR of 0.25 and 0.15, respectively. Fifty-four and 1254 genes passed the FDR < 0.05 threshold, respectively (red horizontal line).

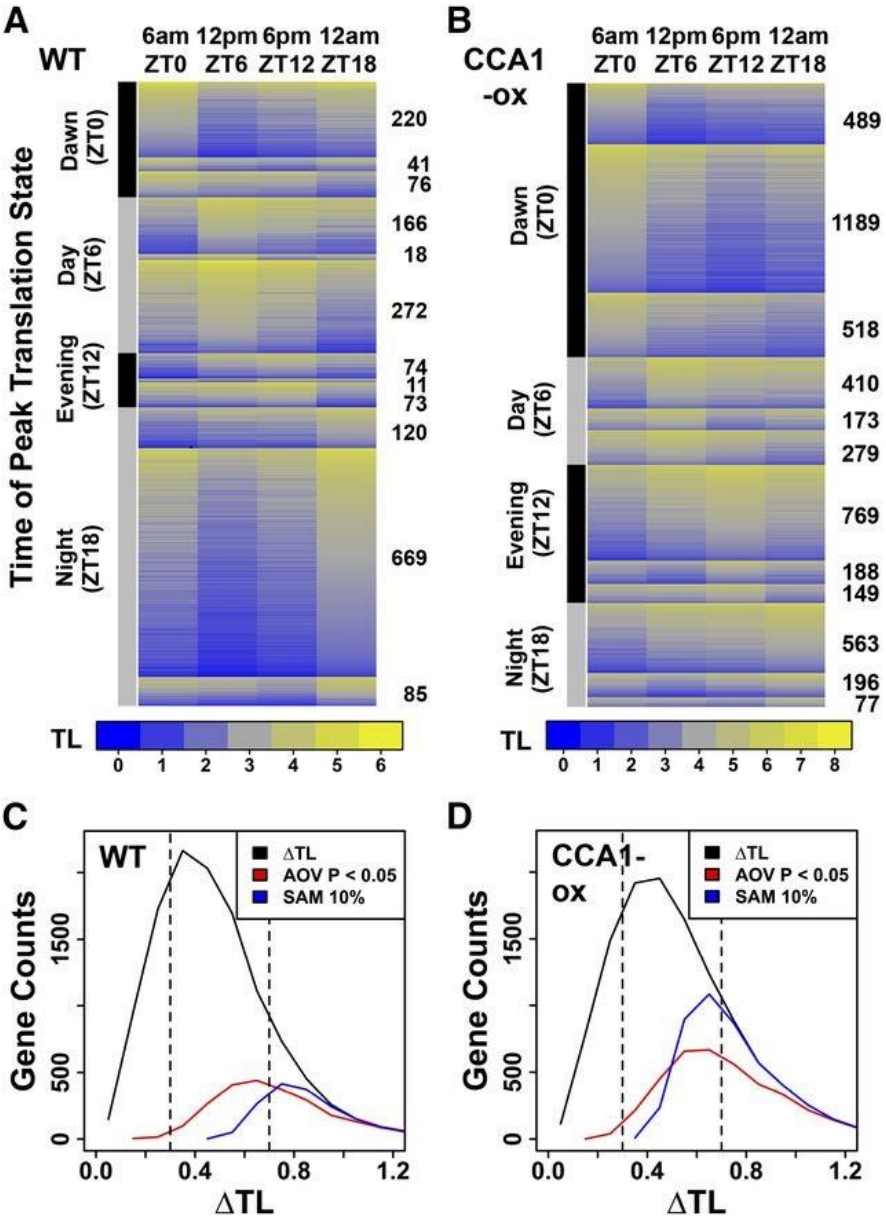

Figure 2.2 continued.

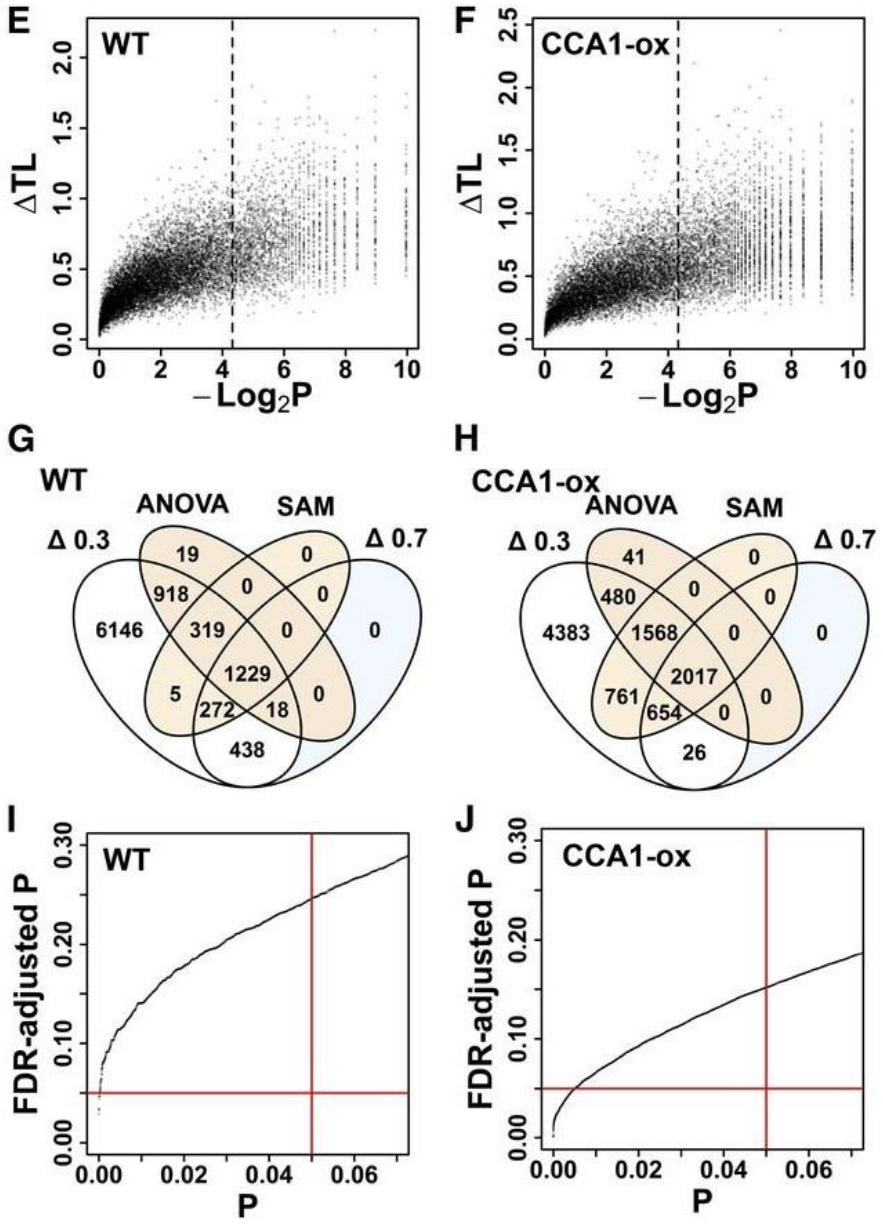

Figure 2.2 continued.

Because the SAM filter will miss valid cycling genes because of the arbitrary FDR cutoff of 10%, and to capture the majority of true positives while balancing the risk of false positives, we applied two other filters to the raw data. One-way analysis of variance (ANOVA), to identify significant variation in TL across time points, yielded 2,503 genes (20%) with an uncorrected p-value below 0.05 and an FDR of below 0.25 (Figure 2.2C, G, and I), indicating that about 2,000 are true positives. As an alternative to ANOVA, we simply applied a moderately stringent threshold (ΔTL>0.7) or a lenient threshold (ΔTL >0.3) to the data (Figure 2.2G**,** also see Supplemental Dataset 1). Based on SAM, ANOVA, and the ΔTL>0.3 threshold, about 4,000 mRNAs were translationally invariant. Taken together, about 2,000 mRNAs have statistically robust translational changes over the diel cycle, about 4,000 mRNAs are invariant, and the remaining 6,300 may have statistically marginal changes in their translation. Evidently, a large fraction of the seedling transcriptome is affected.

We modeled diel variation in TL and TX as sine waves with a 24 hour period (see Methods). Figure 2.3A shows four genes whose large R^2 values document a good fit to a sine model and whose translation states peak around ZT0, ZT6, ZT12 and ZT18. Figure 2.3B shows distributions of R^2 values for TX and TL in WT and CCA1-ox (*see below*). We used an R^2 value of 0.6 as an arbitrary threshold to identify genes with sinusoidal expression, and we identified a peak time as the time at which the sine function reached its maximum. The distribution of peak times is shown in Figure 2.3C for WT TX and TL, as well as for CCA1-ox. While the sine model makes an additional assumption, it has merit because it uses information that we did not previously take into account, namely the wave form of the data. The sine-modeled phase of our wild-type transcript cycles (Figure 2.3C) matched those of eleven published experiments (Mockler et al., 2007) with R^2 values of up to 0.91, confirming the accuracy of our transcript

data and the value of the sine modeling approach. In general, the patterns were consistent with our original analysis but provided additional information. At the TL level, in WT there is a strong preference for genes to peak at ZT19, with a secondary preference around ZT7 to ZT10. It appeared from our original analysis that many mRNAs peaked at ZT18 (12am), while others peaked at ZT0 (6am). The new perspective, made possible by the sine modeling approach, suggests that these two groups may actually behave as one larger group, whose TL peaks are centered about ZT19.

Comparison of transcript levels and translation state over a diel cycle

In the wild type, the majority of genes had phase offsets between the peak in transcript abundance (TX) and the TL peak (Figure 2.4A). After calculating an odds ratio (Figure 2.4B), coincidence between the TX and TL peak was slightly overrepresented compared to the three other phase relationships. Similar results were obtained when genes were classified by their trough times. In summary, phase shifts between TX and TL cycles are common and gene-specific.

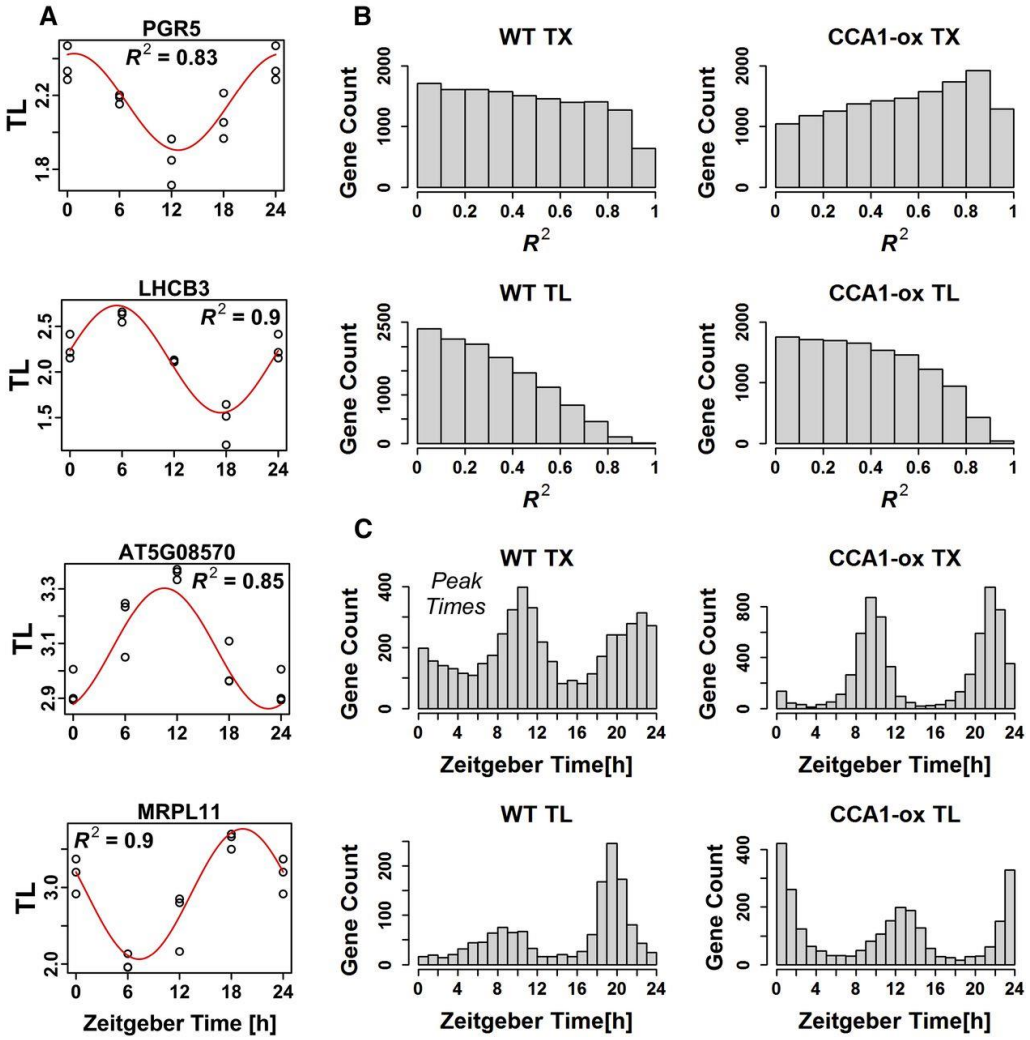

Figure 2.3: Diel cycles of transcript levels and translation states modeled as sine waves. **(A)** Examples of the fit of representative translation state data to sine waves. The R^2 indicates the fraction of the variation in TL that is explained by the sine model. **(B)** Distributions of R^2 values for TX and TL in the wild type and CCA1-ox. Genes in CCA1-ox have a greater tendency to behave like sine waves in terms of their TX and TL. **(C)** Genes with an R^2 >0.6 were selected and were binned according to their estimated peak TX and peak TL under the assumption of a sine model. For wild-type TL, the median confidence interval for peak TL was estimated by bootstrapping to be ±1.8 h (5th to 95th percentile range 1.0 to 2.6 h). In (A), PGR5 is photosystem I protein PROTON GRADIENT REGULATION5. LHCB3 is light-harvesting chlorophyll a/b binding protein 3 of photosystem II. At5g08570 is annotated as a pyruvate kinase family protein. MRPL11 is protein 11 of the mitochondrial ribosomal large subunit.

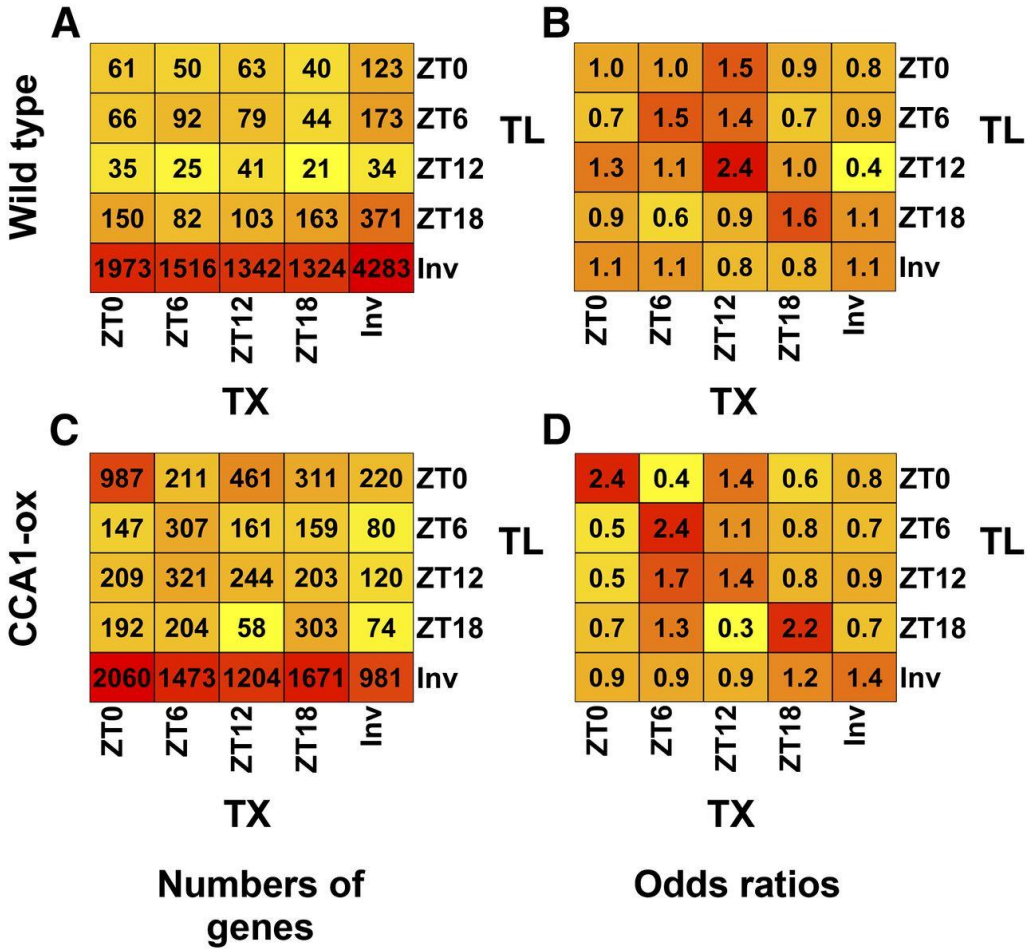

Figure 2.4: Relationship between diel cycles of mRNA levels and TL in the wild type and CCA1-ox.

Genes were classified as cycling at the TX or TL level, as identified by SAM with a 10% FDR. **(A)** and **(C)** Gene counts. Genes sharing the same phase relationship between peak TX and peak TL were binned together in the indicated cells and counted. Each cell contains the number of genes that peaked at the indicated times based on TX and TL. Invariant genes (Inv) were those that passed our prefiltering step but whose TX or TL did not vary significantly across time points. Coloring indicates the (log2) values in each cell, with yellow indicating lower values, orange indicating medium values, and red indicating higher values. A heat map of genes clustered based on TX cycles is shown in Supplemental Figure 7. (A) is the wild type, and (C) is CCA1-ox. **(B)** and **(D)** Odds ratios for the data shown on the left. Odds ratios were calculated by (1) calculating the odds of a transcript with TX peak at time x having a TL peak at time y, (2) calculating the odds of a transcript not peaking at time x having its TL peak at time y, and (3) dividing the odds from step 1 by the odds from step 2. (B) is the wild type, and (D) is CCA1-ox.

Gene ontology enrichment in co-regulated genes over a diel cycle

To identify functional processes that are under diel regulation at the translational level, we searched for enriched functional terms among genes with translation peaks at each time point using gene ontology analysis (Figure 2.5, Table 2.1). At dawn (ZT0), photosystem proteins were enriched, as well as biosynthesis of sulfur compounds such as glucosinolates and sulfur amino acids. At noon (ZT6), cell growth and division processes such as microtubules were slightly enriched, and these terms were strongly depleted in the evening (ZT12). Of note, hypocotyl growth peaks around noon (Nozue, Covington et al. 2007). In WT, there was no enrichment of any terms in the evening. The most striking enrichment was observed at night (ZT18). Ribosomal protein translation was highly enriched at this time, together with RNA methylation, nucleolar proteins and small nuclear ribonucleoproteins. The coordinate diel translation of ribosomal protein mRNAs is displayed in Figure 2.6. Evidently, the ribosome loading of the majority of these mRNAs was high at midnight (ZT18) to dawn (ZT0) and low around noon (ZT6) and evening (ZT12). The small minority of mRNAs that bucked the trend generally represented a less expressed paralog within their gene family; some of them had either very low or very high ribosome loading (not shown). Taken together, many ribosomal proteins belong to one regulon of translational control whose diel ribosome loading peaks at night. Besides the ribosomal proteins, several other small functional categories were preferentially translated at night, in particular mitochondrial proteins, the prefoldin complex, a protein that aids in co- or post-translational protein folding, V-type proton-ATPase, and the DNA directed RNA polymerase IV and V complexes (Figure 2.5, Table 2.1).

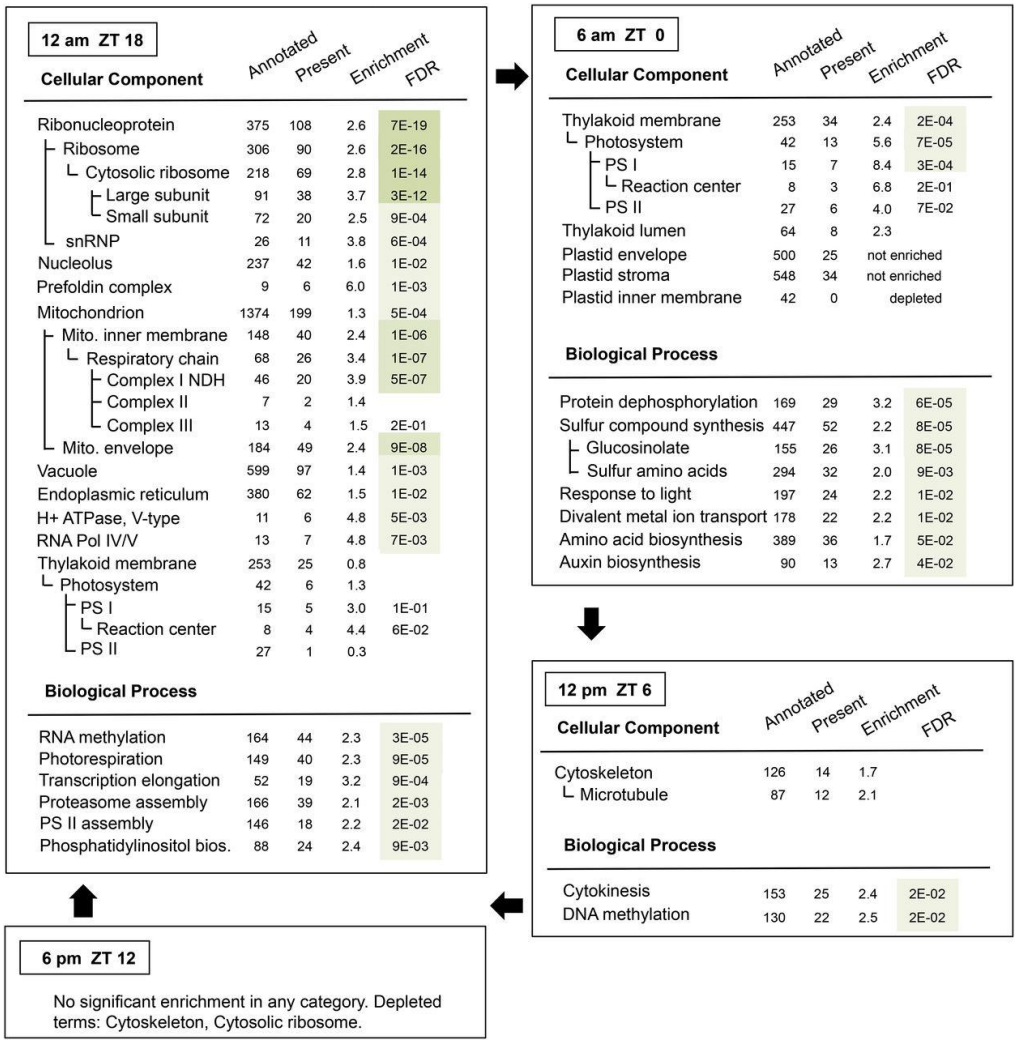

Figure 2.5: Functional enrichment among groups of mRNAs with common peaks in their ribosome loading cycle (wild type).

Genes identified as cycling by either SAM at 10% FDR, ΔTL > 0.7, or ANOVA were searched for enrichment of functional gene annotations. The topGO R package was used with default settings and with all 12,342 reliably expressed genes as a background set to identify enriched biological processes (BP) and cellular components (CC). Significantly enriched functional categories are presented separately in a nested fashion. The table presents the number of genes within the background set that are annotated with the given term, the number whose TL peaks at the given time, the enrichment factor, and its FDR-corrected P value. FDR values below 1E-10 are shaded dark green, those below 1E-05 are medium green, and those below 5E-02 are light green. FDRs above 0.05 are not listed. E, to the power of 10; bios, biosynthesis; PS, photosystem; NDH, NADH dehydrogenase; Pol, polymerase; snRNP, small nucleolar ribonucleoprotein particle.

Table 2.1: Changes in peak translation upon disruption of the circadian clock in CCA1-ox. Cohorts of mRNAs with similar diel ribosome loading cycles were searched for enriched functional categories using Gene Ontology (GO). An X indicates that the term is enriched among the mRNAs whose translation peaks at the given time; a dash indicates absence of enrichment. Detailed data are in Figure 2.5 (wild type) and Supplemental Figure 10 (CCA1-ox).

Table 2.1 continued.

	Peak in Wild Type				Peak in CCA1-ox			
	6 am	12 pm	6 pm	12 am	6 am	12 pm	6 pm	12 am
GO Enrichment Term ZT:	0	6	12	18	0	6	12	18
Translation								
Cytosolic ribosome				X	X			
RNA methylation				X	X			
Prefoldin complex				X	X			
Signal recognition particle						X		
tRNA metabolism						X		
tRNA aminoacylation						X	X	
rRNA processing						X		
mRNA catabolism								X
Mitochondrion								
Inner membrane				X	X			
Envelope				X	X			
Photorespiration				X	X			
Plastid								
Thylakoid	X				X	X		
Photosystem I	X			X	X			
Photosystem II	X				X			
Photosystem II assembly				X		X		
Stroma						X		
Plastid nucleoid						X		
Plastid envelope						X		
Protein targeting to plastid						X		
Carbohydrate synthesis							X	
Carbohydrate catabolism					X			
Carotenoid synthesis						X	X	
Sulfur compound synthesis	X					X		
Amino acid synthesis					X	X		
Nucleus								
Nuclear chromatin						X	X	
Chromatin silencing								X
DNA methylation		X						
Histone lysine methylation							X	
RNA Pol IV/V				X				
Cell division								
Cytokinesis or cell cycle		X					X	X
Microtubule		X					X	
Other								
V-type H+ ATPase				X	X			
Proteasome				X		X		
Phosphatidylinositol synthesis				X	X			
ATP binding (kinase, helicase)							X	X
Protein phosphorylation								X
Protein dephosphorylation	X							
Divalent metal ion transport	X				X			

A more detailed view of ribosome loading dynamics in central energy metabolism is displayed in Supplemental Figure 4. Of those mRNAs that had significant cycles, photosystem I mRNAs were well coordinated (night/dawn peak), as were most of the translocators in the plastid envelope (night peak) and a majority of the light harvesting proteins (dawn/noon peak). Photosystem II proteins yielded less information and Calvin cycle proteins had essentially no TL cycles. Overall, different functional groups of chloroplast proteins have different patterns of ribosome loading. Among mitochondrial proteins with translation cycles, most of which function in oxidative phosphorylation in the inner membrane, most peaked at night/dawn (pattern I), while a smaller subset peaked during the day (pattern II). In glycolysis and associated enzymes, ribosome loading tended to peak during the day. Of note, most glycolytic enzymes did not have translation cycles. Of those three that did, phosphofructokinase and pyruvate kinase both catalyze energetically downhill reactions, and are considered to be highly regulated. Regulation of ribosome loading may add another layer of regulation to these enzymes. Also, of the multiple paralogous genes that encode several of these enzymes, not all undergo changes in ribosome loading, and the phase of the ribosome loading cycle can differ between paralogs (e.g. invertase, phosphoglycerate mutase). This finding suggests that translational regulation is one more way for duplicated genes to evolve new and distinct patterns of regulation.

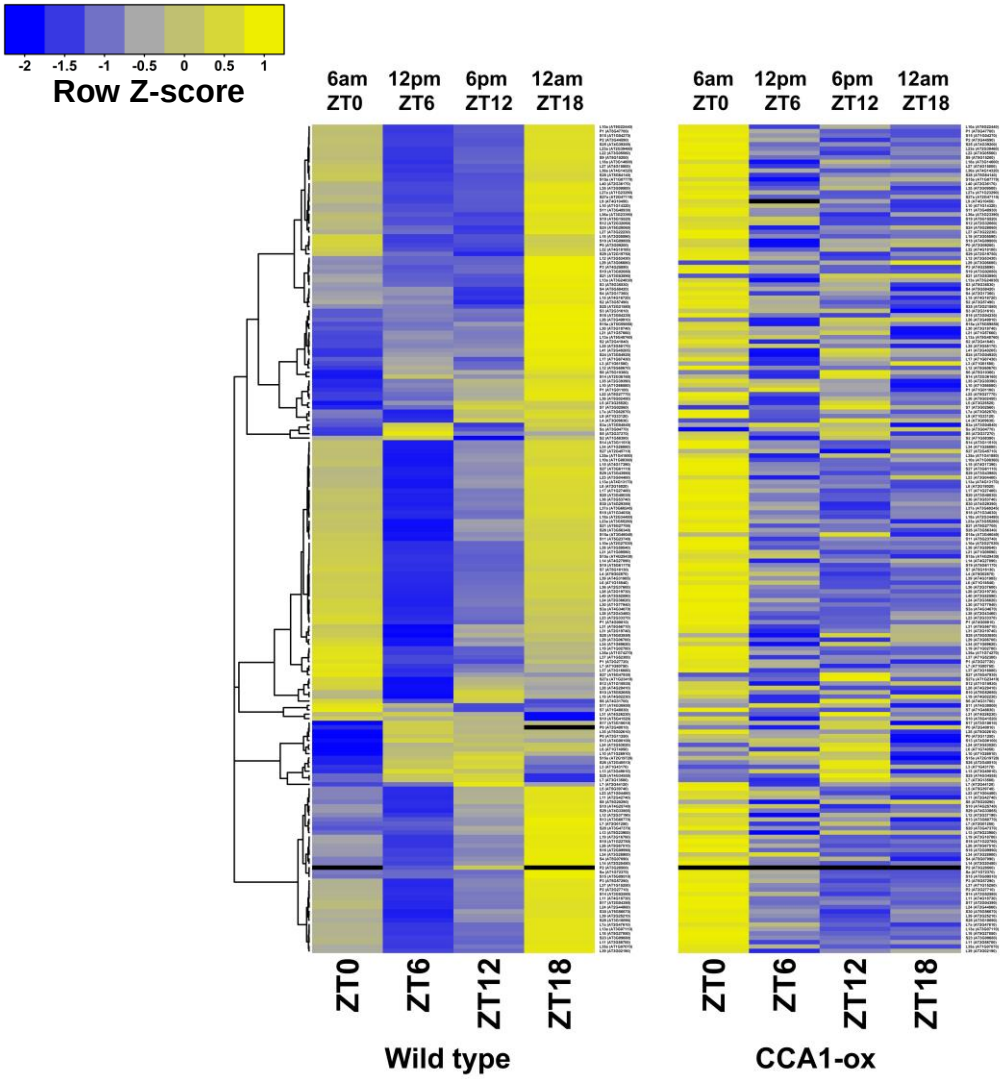

Figure 2.6: Heatmap of translation states for 189 cytosolic ribosomal protein mRNAs. mRNAs (Barakat et al., 2001; Browning and Bailey-Serres, 2015) were matched to probeset IDs (228 out of the 249 genes). Of these, 189 genes passed a criterion for sufficient and reliable gene expression for at least two time points in wild type. Wild-type ribosome loading data were averaged from three replicates and then clustered hierarchically using their Pearson correlation coefficient, resulting in the dendrogram on the left. The equivalent data for CCA1-ox were not clustered and are displayed in the same gene order as for wild type. For display, the ribosome loading of each gene was z-score transformed, where z=+1 (yellow) indicates a translation state one standard deviation above the mean translation state of the gene. Positive and negative z-scores are yellow and blue respectively. The light period lasts from ZT0 to ZT16. Data points with insufficient data, such as low mRNA level, are masked in black. Protein names are readable when viewed on a screen with sufficient zoom.

Two other functional groups of mRNAs with noteworthy translation dynamics are shown in Supplemental Figure 5 and 6. Many redox-related enzymes had translation cycles, but the timing of peak TL differed among individual mRNAs (Supplemental Figure 5A). These enzymes also had a tendency for relatively high absolute ribosome loading (Supplemental Figure 5B). Proteins that function in protein turnover also revealed biases in absolute translation states, with proteases and proteasome subunits scoring high and E3 ligases scoring low, on average (Supplemental Figure 6**)**. Taken together, these data suggest that besides the ribosomal protein mRNAs, several other functional groups of mRNAs also form translational regulons.

Global transcript profile in the clock-deficient CCA1-ox strain

To distinguish whether the diel cycles of translation are driven by the circadian clock or by diel light-dark changes, we then prepared to compare WT and a strain overexpressing the central oscillator gene, *CCA1* (CCA1-ox). As a first step we determined transcript cycles of clock genes because these had not been described at a global level (Figure 2.7). Compared to WT, the majority of transcript profiles in CCA1-ox correlated with light and darkness indicating that, as expected, the plants' ability to anticipate the lights-on signal was severely curtailed. In contrast, WT had a variety of diel TX patterns, including many genes that appear to anticipate light changes. Assuming that transcript cycles conform to a sine model, in CCA1-ox the distribution of TX peak times was relatively narrow, restricted around ZT9 (3pm) and ZT21 (3am) (Figure 2.3C). In WT, many more mRNAs peaked between ZT12-ZT16 (evening) and between ZT0-ZT6 (morning).

The broad correlation between light phase and transcript phase is reminiscent of the hypocotyl growth in CCA1-ox (Nozue, Covington et al. 2007), which, having escaped from

control by the clock, is also strongly driven by light and darkness. As expected (Wang and Tobin 1998, Green, Tingay et al. 2002, Matsushika, Makino et al. 2002), cycling of key clock-regulated transcripts was muted in CCA1-ox (Figure 2.8A**)**, *e.g.* for *ELF4, PRR5, PRR7, PRR9*. Additionally, *LHY* was transcriptionally repressed at all time points. Although diel cycling of LHY mRNA was clearly disrupted in CCA1-ox, in our hands, LHY retained a small peak at the end of night at ZT0. Therefore we cannot rule out that the CCA1-ox plants may retain weak residual clock activity in long day.

In CCA1-ox, the following broad functional annotation patterns were observed (Supplemental Figure 8**)**. At dawn (ZT0), the cohort of peak transcripts was enriched for RNA-biology processes and protein localization. At noon (ZT6) and in the evening (ZT12), the majority of enriched processes were chloroplast- and photosynthesis-associated. Remarkably, of 309 light-reaction mRNAs, 229 peaked during the light period, either at ZT6 or ZT12, a stronger enrichment than in the wild type (note red values for chloroplast (ZT6) and PS light reaction (ZT12) in Supplemental Figure 8, column WT). Similar trends were seen for many functional groups responsible for carbohydrate metabolism. In the evening (ZT12), categories such as apoplast, glucose catabolism, and glucosinolates became prominent as well. Finally, at night (ZT18), ion transport and defense responses became predominant. Notably, the majority of these patterns in CCA1-ox were similar, but more accentuated, as compared to wild type. However, for other functional terms, the transcript patterns in CCA1-ox were muted (*e.g.* defense response, cell wall at ZT18; green values indicate stronger enrichment in WT). The term cytokinesis/cell division was biased towards noon (ZT6) in WT, yet only weakly enriched, at ZT12, in CCA1-ox.

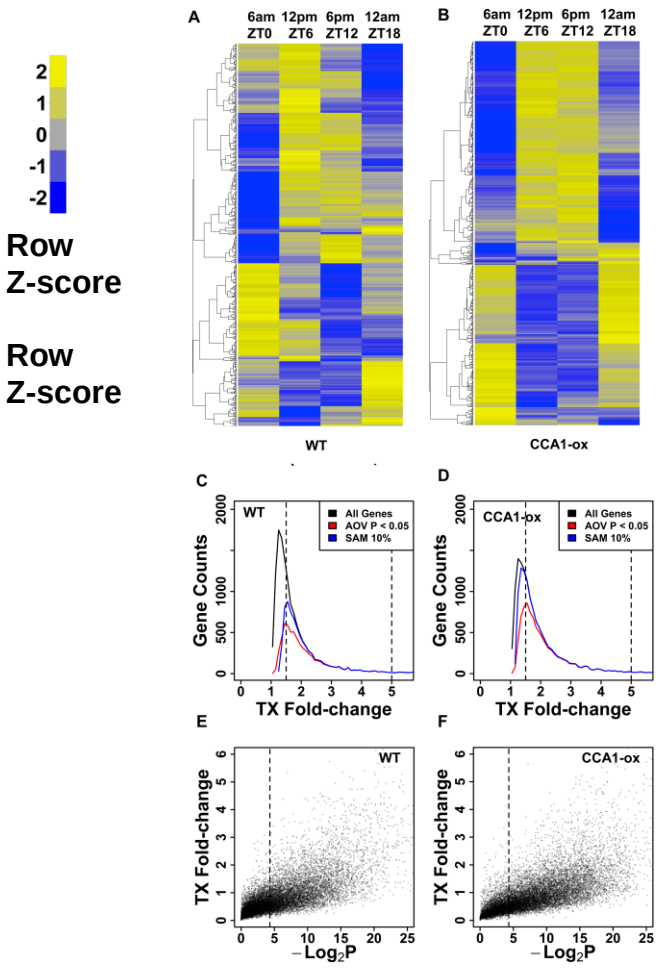

Figure 2.7: Diel cycles of transcript levels in wild type and CCA1-ox.
Heat map of total mRNA levels obtained under long day conditions at 6am (ZT0), 12pm (ZT6), 6pm (ZT12), and 12am (ZT18). Data are averages from three biological replicates. Only mRNAs with a fold-change higher than 5-fold are included. **(A)** WT, 615 genes, **(B)** CCA1-ox. Transcript levels are displayed after z-score transformation (yellow=high). The clustering tree on the left reveals how, in CCA1-ox, most mRNAs fall into two large clusters that correlate with light (12pm and 6pm) and darkness (6am and 12am). In contrast, in wild type, transcript levels fall into four major clusters that were less prone to reflect the light environment. Panels **(C)** and **(D)** show a line histogram of transcript fold changes for the entire dataset (13,625 genes) similar to Figure 2.2C and D. Panels **(E)** and **(F)** show the relation between ANOVA p-value (AOV) and fold-change in transcript level for 14,218 genes (volcano plots). The stippled line represents p=0.05.

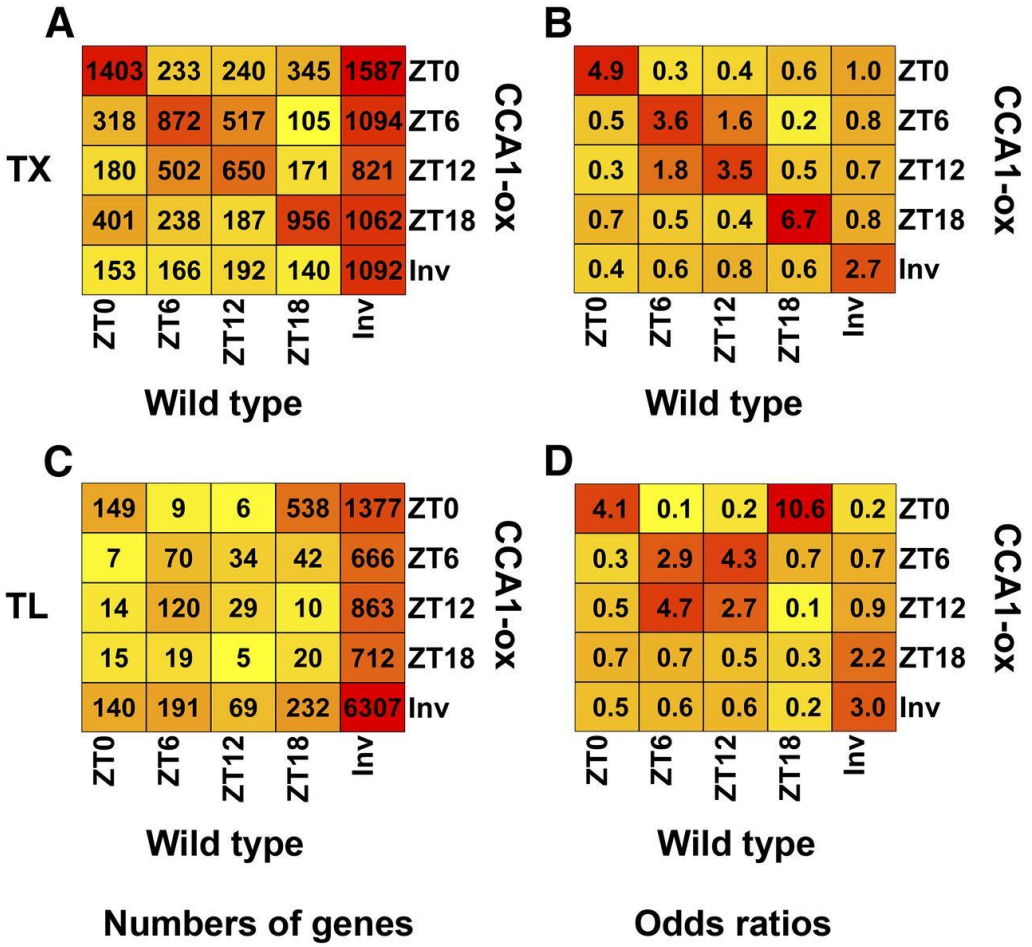

Figure 2.8: Phase diagram comparing the expression cycles between the wild type and CCA1-ox.

Genes with significant TX cycles ([A] and [B]) or translation cycles ([C] and [D]) were identified using SAM with the 10% FDR cutoff. All other genes were called invariant (Inv). The genes were binned according to the peak in the wild type and the peak in CCA1-ox. **(A)** and **(C)** Numbers of genes with a given phase relationship are indicated as a heat map. **(B)** and **(D)** Odds ratios for the data shown on the left. For details, see legend to Figure 2.4.

In summary, the defect of the clock in the CCA1-ox strain disturbs the coordinate transcription of certain functional classes of mRNAs, but also opens the door to a more tightly coordinated transcription for other classes of mRNAs, especially photosynthesis-related mRNAs. In wild-type plants the functional clock uncouples large numbers of transcripts from the tight control exerted by the overt light environment.

Translational cycling in plants with a disrupted circadian clock

We initially hypothesized that translational regulation by the diel cycle would be dampened in CCA1-ox plants. However, polysome microarray analysis of the CCA1-ox strain revealed robust translation cycles (Figure 2.2B, 2D). The clusters with dawn and evening peaks were enlarged at the expense of the day and night peaks. Thanks in part to a lower variation between replicate experiments, the CCA1-ox TL data yielded a larger number of genes (5,521 versus 2,780 in WT) that were scored statistically significant in their translation cycle by either the SAM or ANOVA methods (Figure 2.2D, F, H, J). In CCA1-ox, only 2,533 genes were not identified as cycling by any method and were therefore classified as surely invariant.

In CCA1-ox, TX and TL cycles were more highly coordinated than in WT (Figure 2.4C). This was particularly striking at dawn (6am, ZT0), as evident from the high odds-ratios along the diagonal in Figure 2.4D. Taken together with the previous data from WT plants (Figure 2.4A and B), this observation suggests that the clock does not just regulate translation. The fully functional clock in the WT may also work to separate transcriptional control from translational control, making them more independent of each other. Whilst in the clock-defective CCA1-ox strain, transcription and translation cycles may be coming under an alternative form of joint control, possibly light-dark transitions and the associated shift in the cellular energy balance.

Diel cycles of translation are disturbed by malfunction of the clock

Next, we examined how the clock defect in the CCA1-ox strain affected the phase of the diel translation cycles (Figure 2.8C and D). Among the mRNAs with robust translation cycles in WT, about one-third lost their cycle in CCA1-ox (Figure 2.8C). Of those that maintained a cycle in CCA1-ox, the minority maintained their peak at the same time as WT, while the majority shifted their translation peak to a different time. A 6-hour delay was most common. In particular, the vast majority of mRNAs with a WT TL peak at 12am (ZT18) peaked six hours later in CCA1-ox. These mRNAs preferentially encode ribosomal proteins and mitochondrial proteins (Table 2.1). Finally, we draw attention to more than 3,000 mRNAs that had no robust cycle in WT, yet started to cycle in CCA1-ox, thus revealing translation cycles that may be suppressed in WT by the fully functional clock (Figure 2.8C). Overall, the shift in TL in the CCA1-ox strain was also evident when TL was modeled as a sine wave (Figure 2.3C). Taken together, these data indicate that a functional clock is important for maintaining translation cycles for certain mRNAs and suppressing the cycles of others.

For comparison, at the transcript level, it was more common for genes to peak at the same time in WT and CCA1-ox (Figure 2.8A and B). Only a small fraction lost their cycle. However, like at the translation level, many mRNAs that had no consistent transcript cycle in WT did cycle in CCA1-ox.

The shifts in translation cycles in the CCA1-ox strain stood out in starker contrast after gene ontology analysis (Figure 2.5 for WT, Supplemental Figure 9 for CCA1-ox, and Table 2.1 for a summary). The translation phase of the following major processes was delayed by six hours: cytosolic ribosome (Supplemental Figure 3) and RNA methylation, mitochondrial inner membrane and envelope, photorespiration, sulfur compound synthesis, and cytokinesis. For some

of the smaller categories, such as the prefoldin complex, V-type proton ATPase, and microtubules, the 6-hour delay was very striking, given that a sizable fraction of genes in each group was translationally regulated in a coordinated fashion. Second, certain functional classes maintained a peak in TL at a given time or showed only a slight shift, for example photosystem proteins and metal ion transport. Third, several functional categories only revealed coordinate translation in CCA1-ox, after disruption of the clock, but not in WT. For example, mRNAs for tRNA metabolism, specifically aminoacyl-tRNA synthetases, typically did not cycle in WT but did cycle in CCA1-ox with a peak at ZT6 or ZT12. The opposite pattern, loss of coordinate translation, also occurred occasionally (e.g. RNA Polymerase IV/V).

It should be understood that not all mRNAs within one larger category follow the same dynamic. For example, in WT the mRNAs for amino acid biosynthesis had gene-specific translation peaks at each time-point (not shown), but appeared to coalesce into a broader peak with enrichment at ZT0 and ZT6 in CCA1-ox. These data again underscore coregulation of both large and small, functionally related, groups of mRNAs and suggest that, besides the well-known ribosomal protein mRNAs, many other mRNAs are targets of translational control.

The changes in ribosome loading over the diel cycle differ dramatically from changes previously described to occur in response to shorter, unexpected, dark or light treatments (Supplemental Figure 10). Using k-means clustering, it is evident that three different shifts from light to darkness (*e.g.* ZT12 evening to ZT18 night) did not resemble the response that occurs when light-grown seedlings are exposed to one hour of darkness in midday at ZT8 (L➔1hD, Juntawong et al., 2012). Instead, it resembled more closely the response of dark-exposed seedlings to reillumination (Juntawong et al., 2012) and the response of dark-grown seedlings to four hours of white light, which stimulates their deetiolation (Liu et al., 2012). Vice versa, the

diel response to a dark-to-light shift (ZT0 to ZT6) resembled the response to a one hour exposure to darkness. Of the six clusters defined by k-means, clusters 1, 2, 3, and 4 all supported this anti-correlated pattern. In fact, in previous studies, ribosomal proteins were among the most highly repressed mRNAs in response to darkness, and induced by light, whereas here, ribosomal proteins (cluster 1 and 2) were translationally stimulated during the night (Table 2.1). Evidently, many other mRNAs follow a similar pattern, including mitochondrial (Cluster 1) and cytoskeletal proteins (Cluster 3), and polysaccharide synthesis (Cluster 4). Only clusters 5 and 6 (enriched for chloroplast, Golgi, cell wall, glucose catabolism) followed the more intuitive expectation of a correlated behavior. These data suggest that mRNAs belonging to the night cluster under diel conditions also tend to be translationally inhibited when darkness is experienced by the plant as a stress condition.

Translational control of circadian clock mRNAs

Next, we addressed whether clock mRNAs were subject to translational control. Because the levels of many central clock mRNAs drop below the reliably detectable limit at one time point, and were therefore filtered out in our previous analysis, we re-filtered the raw array data such that genes must be present in all three replicates of SP and LP fractions for at least two time points, but not all four time points (14,397 genes). In CCA1-ox the transcript cycles of many clock-associated genes became muted, and the peaks and troughs often coincided with darkness (ZT0, ZT18) or light (ZT6, ZT12, see Figure 2.9A), not unlike the transcriptome as a whole (Supplemental Figure 7).

At the translation level, clock-associated transcripts (Figure 2.9B**)** fell into two broad groups in WT, one with peaks at dawn or noon (ZT0 or ZT6, e.g. GRP8, COP1, PHYD), and a

second with peaks in the night (ZT18, e.g. LUX, UVR8). Upon disruption of the clock in CCA1-ox both groups changed dramatically. The late night group tended to be delayed to dawn or noon (e.g. LUX, UVR8), while the dawn-noon group was delayed until noon (GRP8, COP1) or evening (e.g. photoreceptors including PHYE, PHYD, CRY2, PHOT2) or beyond (Figure 6B), roughly consistent with the pattern across the entire translatome (Figure 2.8C, D).

Figure 2.10 displays the same cycles of transcript levels and translation states in the context of a clock wiring diagram (Pokhilko, Fernandez et al. 2012). For WT, several evening genes showed a 6-hour phase delay between peak transcript level and peak translation (e.g. *TOC1*, *LUX*). Among genes with a transcript peak around noon, *PRR5* and *GI* also had an offset in their translation peak. In contrast, among genes with a dawn transcript peak, coincidence of TX and TL was more common (e.g. *LHY*, *CCA1*, also see *CRY1*, *HYH* in Figure 2.9). While inferences about the behavior of single genes are sensitive to noise, in the aggregate it seems clear that translation states are a novel way for plants to fine-tune the expression of clock-regulatory mRNAs. These data suggest that the clock relies not only on transcriptional control but also on translational control at the level of ribosome loading for proper functionality.

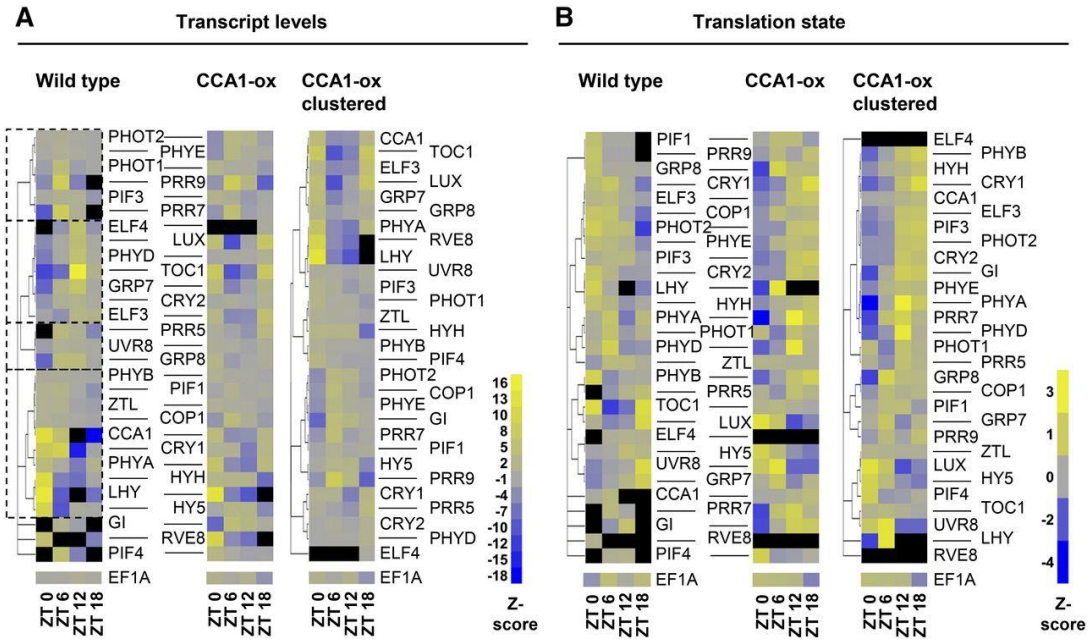

Figure 2.9: Diel cycles of translation states and mRNA transcript levels for clock-associated genes.

Diel cycles of translation states and mRNA transcript levels for clock-associated genes. Previously described clock-associated genes were hand-selected, focusing on the central oscillator, the light input pathways, and selected outputs. EF1A was included for comparison as a weakly cycling mRNA. For better comparability between genes, the signals are row scaled. In detail, for each gene, the mean signal was scaled to 0 (gray), and an average SD was calculated. The signal value for each time point was row-scaled so as to indicate the number of standard deviations that separate each value from the mean for that gene (unitless Z-score, yellow = high). Black indicates that the gene did not pass our prefilter for that time point. The original transcript abundances and translation states are displayed in Supplemental Figure 8.

(A) Transcript abundance. Left: the wild type. Genes were clustered according to their original mRNA expression values using hierarchical clustering based on Pearson coefficients, with replicates averaged. Four major clusters are boxed to aid in interpretation. Middle: CCA1-ox. The genes are ordered according to the wild-type clustering tree. Gene names are shown between the left and middle panels and apply to both panels. Right: CCA1-ox data were reclustered on their own; gene names are shown on the right.

(B) Translation patterns of clock-associated genes are displayed as in (A). TL states were calculated as described in Methods. Note the large cluster of mRNAs whose TL peak shifted from morning/noon (ZT0/ZT6) in the wild type to evening/night (ZT12/ZT18) in CCA1-ox.

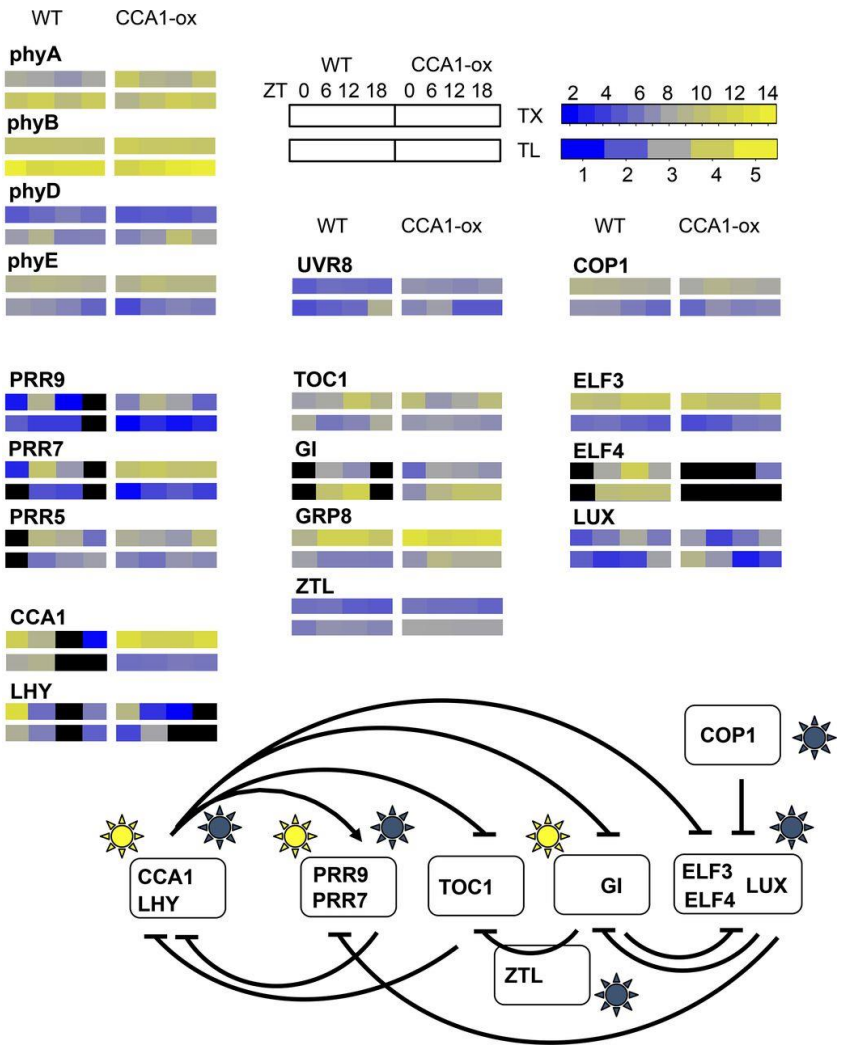

Figure 2.10: Translation cycles affecting the Arabidopsis circadian clock.
A wiring diagram for the Arabidopsis clock (Pokhilko et al., 2012) was amended to illustrate cycles of TX levels and TLs. Arrows and T-bars indicate stimulation and repression, respectively. Yellow suns indicate light-regulated transcription, and blue suns indicate light-dependent posttranscriptional effects. TX levels and TLs are displayed in the form of heat maps. The upper row of each set of rectangles indicates the TX and the lower row indicates the TL. The left set indicates the wild type and the right set CCA1-ox. Data are from Supplemental Figure 8; n = 3. For ELF4, TL data are omitted because of poor expression data. Black symbolizes mRNA expression below threshold.

Given that the amplitude of TL cycles is generally small, we wanted to simulate how the waveform of the protein synthesis rate might be affected by phase offsets between a TX cycle and a TL cycle (Figure 2.11). Panel (A) assumes that TX and TL both follow plain sinusoidal patterns. The rate of protein synthesis was calculated by multiplying TX by TL. A 9-hour offset between TX and TL causes an asymmetric peak in the synthesis rate, and a 12-hour offset can produce a plateau-like pattern. It is very challenging to measure protein synthesis rates empirically (Schwanhausser et al., 2011), especially with enough precision to distinguish relatively small differences, and especially in plants, where the key technique, stable isotope labeling with amino acids, is just barely becoming more common (Lewandowska et al., 2013). This is why a simulation is useful. Data from selected Arabidopsis mRNAs are shown to exemplify how TL may modulate the synthesis rate (Figure 2.11B-I). The photoreceptors PHOT2 and UVR8 (AT5G63860, UV-B receptor) predict that the translation dynamic amplifies the TX dynamic or extends the protein synthesis into a plateau-like pattern, respectively. For data as those from LP1 (AT2G38540, lipid transfer protein 1) and RCE1 (AT4G36800, Rub1 conjugating enzyme) the offset between TX and TL cycles causes a delay or advance, respectively, in the protein synthesis rate by several hours, which is not expected from the TX level alone. For MRPL11 (AT4G35490, mitochondrial ribosomal protein L11) and RPL P1 (ribosomal protein P1) the protein synthesis rate is driven primarily by the TL pattern, yet phase-shifted slightly by the weaker TX cycle. Finally, PFK1 and PFK3 are isoforms of phosphofructokinase (At4g29220 and At4g26270, Supplemental Figure 4) that differ in their phase offsets between TX and TL. The PFK data demonstrate that genes related by an ancient gene duplication could evolve different patterns of translational regulation. Taken together, it is plausible that differences in the waveform of protein synthesis rates may help to fine-tune gene

function. For example, high translation of the TOC1 and LUX mRNAs at night (Figure 2.10) may allow these proteins to continue to repress transcription of morning genes, CCA1 and LHY, and of day genes such as GI and PRR9, respectively. In conclusion, relatively shallow cycles of ribosome loading that affect a large number of mRNAs may contribute in a non-linear fashion to the functioning of complex cellular networks.

Translation cycles in constant light

Finally, we examined the role of the circadian clock in the cycle of translation by measuring TLs in continuous light. CCA1-ox plants grown in a light-dark cycle still exhibit circadian rhythmicity in transcriptional behavior, but CCA1-ox plants grown in constant light do not (Green et al., 2002). We examined polysome loading during and after a shift from long day to constant light conditions (LD>LL) in both WT and CCA1-ox. On a global level (Figure 2.12A), the fraction of wild-type RNA recovered in polysomes (PL) rose during the day, as expected (Figure 2.1), dropped during the subjective night despite continuous illumination, then rose and fell again during Day 2. This pattern was clearly evident in the individual dynamics of NP and LP. In CCA1-ox, in contrast, PL stayed high during the first subjective night, and then fell during Day 2. These data further indicate that global polysome loading is controlled by the circadian clock.

Figure 2.11: Simulation of protein synthesis rates from mRNA transcript level data and TL data. mRNA levels were multiplied with the corresponding translation states in order to simulate the protein synthesis rate from the given mRNA. The results were extrapolated to 42 h to better visualize the cycling behavior. The mRNA levels and simulated protein synthesis rates are mean-centered (mean = 100) and are displayed on the left y axis, while the TL is displayed as is on the right y axis. **(A)** A mathematical simulation of TX level (stippled trace), TL in three possible phases (pale traces), and the three calculated protein synthesis rates (dark traces). **(B-I)** Actual TX and TL data and the protein synthesis rates calculated from them. The difference in the shape of the transcript level trace and the protein synthesis trace is accentuated with dark shading. ZT, zeitgeber time.

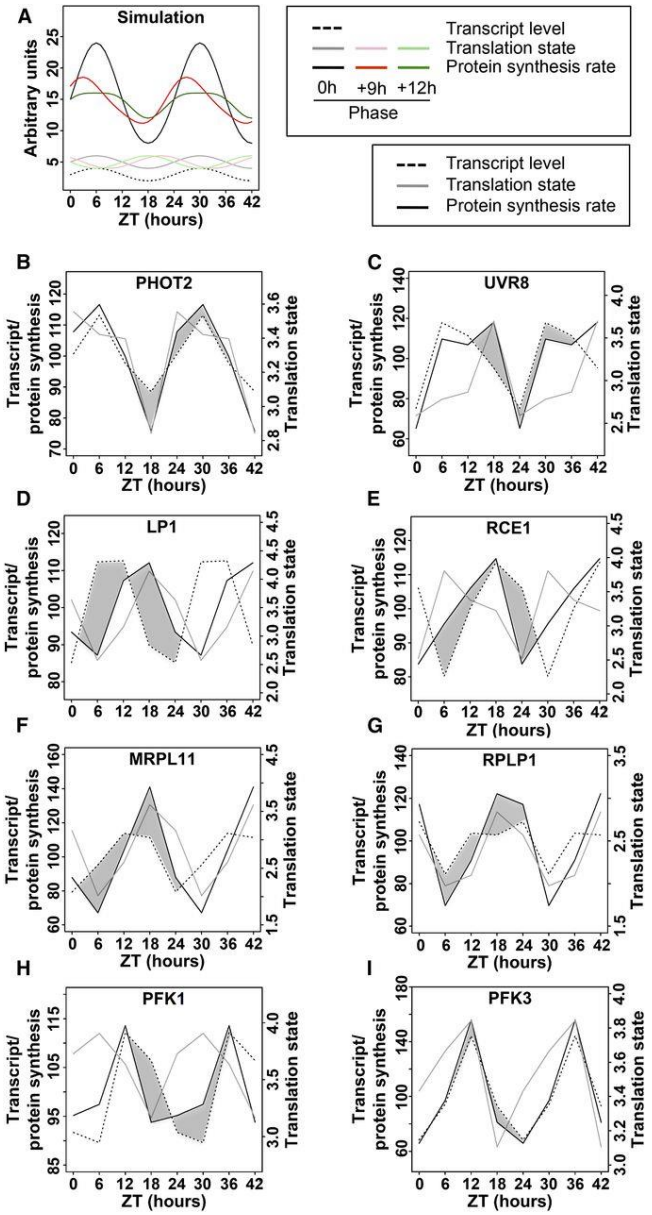

Figure 2.11 continued.

We then analyzed the translation states of specific mRNAs using qRT-PCR (Figure 2.12B). The mRNAs for two chlorophyll binding proteins (CAB4 and LHCA1), and one ribosomal protein (RPL26B) experienced dynamic fluctuations in ribosome loading in the wild type, wheras in CCA1-ox these fluctuations were suppressed. Two central clock mRNAs, PRR3 and possibly CCA1, also displayed cyclical changes in wild type but not in CCA1-ox. In contrast, the EF1A mRNA maintained an even ribosome loading throughout. Although there was considerable variation between replicates, as illustrated in detail for LHCA1 in the wild type, the general trend for an amplified dynamic on Day 2 of the time course was consistently observed. It is noteworthy that the dynamics in wild type often did not conform to a strict 24-hour period (see trough-to-trough period for LHCA1, PRR3, and RPL26B). Together, these observations suggest that wild type has a trend for cyclical ribosome loading of various mRNAs, even under continuous light, in keeping with the global dynamic (Figure 2.12A). Moreover, a functional clock appears to be required for these ribosome loading cycles. However, the erratic patterns seen under continuous light suggest that regular light-dark changes assist the clock in organizing the ribosome loading of various mRNAs.

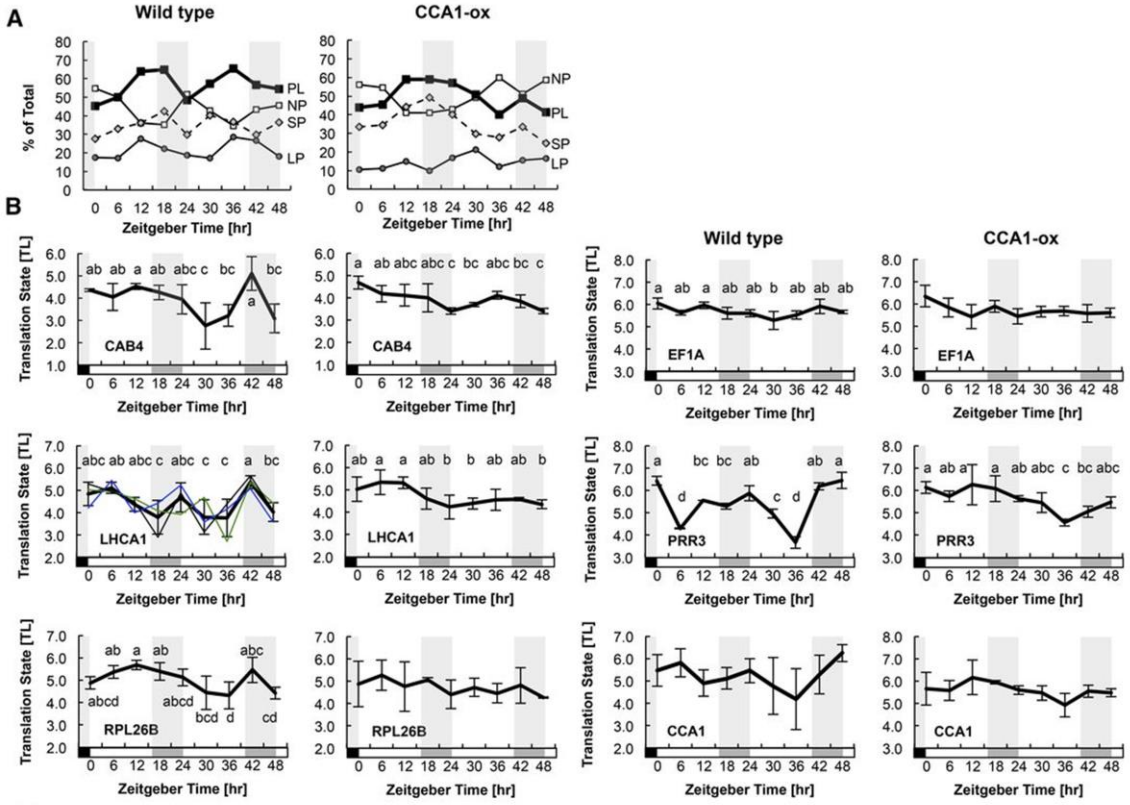

Figure 2.12: TLs under continuous light conditions.
Plants were grown for 10 d under long-day conditions (LD) and shifted to continuous light (LL) at ZT0. Plant samples were harvested at 6-h intervals for 48 h and fractionated by sucrose density gradient into NP, SP, and LP fractions. The left column contains data from the wild type and the right column is from CCA1-ox. The subjective night is indicated by gray shading. **(A)** The percentage of RNA measured by UV absorption that was detected in NP, SP, and LP portions of the gradient. The trace labeled PL (polysomes) is the sum of SP and LP. **(B)** The abundances of individual mRNAs for selected genes were quantified by qRT-PCR. Translation states were calculated. The traces from the three biological replicates were centered on their average TL. The error bars indicate the range of the data from n = 3 biological replicates (2 for PRR3). For LHCA1 as a representative mRNA, the traces of the three individual replicates are also shown with thin lines. Letters above each data point indicate which data points are significantly different by ANOVA with Tukey post-hoc test ($\alpha = 0.05$). Data points that share the same letter are not different from each other. Data sets that lack lettering had few or no significantly different pairs.

Discussion

In this study we document diel and circadian regulation of ribosome loading of mRNA in *Arabidopsis* seedlings. These fluctuations are extensive, affecting about 15% of transcripts in WT seedlings, and affect different classes of genes differently. The specific diel patterns of ribosome loading that are observed in wild-type plants require a functional circadian clock. Although diel effects on translation have been documented earlier, here we provide information across the entire transcriptome, gene by gene. Clock-controlled diel translation must be integrated with other internal and external cues that are known to affect translation.

Diel regulation of ribosome loading is extensive

Of approximately 12,000 mRNAs that were reliably detected in Arabidopsis seedlings, about one fifth (2,503) passed ANOVA with a standard significance threshold ($p<0.05$), far more than the 600 expected to pass this threshold by chance alone if data were randomly distributed. More than 7% (890) of genes were scored as dielly fluctuating by SAM at an FDR of 0.05. These are respectable yields of statistically validated genes given that our measure of TL is calculated from three independent data points, i.e. NP, SP, and LP mRNA, that each come with their own margin of error. To optimize biological inference, we applied different filters to our raw translation data, depending on the question at hand. For example, to search for functional enrichment using gene ontology, we started with a lenient filter which selected genes through any of three different measures of significance. This approach yielded substantial insight into translationally co-regulated mRNAs by uncovering enrichment of numerous functional categories, which were evaluated rigorously for statistical significance by guarding against

multiple comparisons (Figure 2.5). Other analyses, such as comparisons between TX levels and TLs, as well as phase shift diagrams between WT and CCA1-ox, were conducted with a more rigorously preselected set of genes (SAM with FDR<10%), again while performing downstream statistical tests (Figure 2.4, Figure 2.8).

Here, the diel fluctuations in ribosome loading of *Arabidopsis* mRNAs were revealed in seedling shoots growing on defined medium with 1% sucrose and entrained by a 16-hour day (Figure 2.1). When Pal and coworkers performed global analysis of ribosome loading over the diel cycle in vegetative rosettes in a 12-hour day (Piques, Schulze et al. 2009, Pal, Liput et al. 2013), their global drop in ribosome loading was deeper than ours, suggesting that diel translation under natural conditions, i.e. with fluctuating levels of photoassimilate, may be more pronounced than seen here.

Diel phase affects translation in more than one way

We found peaks in ribosome loading for mRNAs at all four time points that we surveyed, dawn, noon, evening, and night. Together with the large cohort of translationally flat genes, there appear to be five groups of mRNAs. Of these five, the groups peaking at night, at dawn, and at noon had distinct functional enrichment, underscoring that they are regulated differently. The 'night' cohort had the most striking bias, being enriched for mRNAs for cytosolic ribosomal proteins, mitochondrial proteins, and several smaller protein assemblies. Significant functional biases were also evident in the dawn (ZT0) and noon (ZT6) cohorts. Of the photosynthetic apparatus, photosystem I mRNAs were translationally stimulated at night and dawn, photosystem II was biased more toward dawn, and most light harvesting proteins were stimulated well into the light period, possibly shedding light on translational control of assembly

of the photosynthetic apparatus. No positive enrichment was detected among mRNAs whose ribosome loading is flat, nor in the smaller evening cohort (peak at ZT12), although the evening group was strongly depleted for terms that peak at noon and at night, i.e. cytoskeleton and ribosome.

Ribosomal protein mRNAs are the best defined translational regulon in plants, being coregulated under essentially all experimental conditions that have been examined. Their coregulation here further validates the quality of the data. These mRNAs are preferentially repressed in their translation after abiotic stress such as heat (Yanguez, Castro-Sanz et al. 2013), drought (Kawaguchi, Girke et al. 2004), and hypoxia (Branco-Price, Kaiser et al. 2008). They are translationally stimulated by light (Liu, Wu et al. 2012) and in mutants defective in the translation apparatus (Kim, Cai et al. 2007, Tiruneh, Kim et al. 2013). These mRNAs are translationally repressed by one hour of unanticipated darkness in the middle of the day (Juntawong and Bailey-Serres 2012), while they are translationally stimulated after two hours of anticipated darkness at night, in our data. These findings suggest that ribosomal protein mRNAs are regulated by darkness in a circadian context.

Diel translation is a function of the circadian clock

Three pieces of evidence indicate that the diel regulation of ribosome loading is an output of the circadian clock. First, the clock-compromised CCA1-ox strain has a very distinct translational profile compared to WT. Translation cycles continued to be very evident in CCA1-ox, but the dawn and evening peaks were much more pronounced while the night peak and noon peak were relatively smaller. For several specific functional categories, the translation cycles in CCA1-ox appear to be delayed by about six hours, for example ribosomal proteins, indicating

that the WT clock works to advance the time of peak translation. One caveat when using the CCA1-ox strain is that CCA1 may have additional functions that are entirely independent of the clock. For example, by exposing the cell to CCA1 protein in the evening, when its level is usually low, CCA1 might interact spuriously with partners that have no relation to the clock and that CCA1 does not normally encounter. For this reason, experiments were also performed under continuous light conditions. Second, cycles of diel translation were also evident in WT under free-running continuous light (LL) conditions (Figure 2.12), suggesting that they are governed by the circadian clock. Third, these fluctuations of translation under continuous light were broadly disrupted in the CCA1-ox strain (Figure 2.12). Given that CCA1-ox is substantially clock-deficient under continuous light (Wang and Tobin 1998), we conclude that cycles of diel translation are driven by the circadian clock.

The clock's output pathways may well affect translation indirectly, because none of the core oscillator components are known translational regulators. Most direct outputs of the clock are transcriptional, and these may well mediate the translational effects seen here. For example, in CCA1-ox plants many RNA-biology transcripts peak around ZT0, whereas in the wild type fewer of them do (Supplemental Figure 8). This result adds credence to the hypothesis that the clock may regulate translation via the primary layer of clock output genes. Should one consider whether CCA1-dependent changes in translation may be the far-downstream consequence of physiological or developmental alterations in the CCA1-ox strain? CCA1-ox does have a clearly elongated hypocotyl (Wang and Tobin, 1998). In addition, our wild-type seedlings expressed the floral inducer FT, while CCA1-ox plants did not. Overall, however, CCA1-ox is morphologically quite normal and physiologically vigorous at the time of our experiments. Therefore, we doubt

that the translational changes in CCA1-ox can be attributed solely to mechanisms that are several degrees of separation away from the clock itself.

We collected microarray data of the global transcriptome in CCA1-ox over the diel cycle. Many transcripts assume a cycle that mirrors the light and dark conditions, in keeping with the clock defect. However, the *LHY* transcript, while expectedly repressed, rose slightly at ZT0 (6am), an effect we interpret as a residual anticipation of dawn, and thus residual clock activity (Figure 2.9). Even though CCA1-ox is not completely clock-deficient in the light-dark cycle as previously noted (Green, Tingay et al. 2002, Matsushika, Makino et al. 2002), the strain was broadly affected in its translation, indicating that CCA1 cycling is critical for dielly regulated translation.

The clock-deficient CCA1-ox strain revealed translational cycling of several hundred new mRNAs, which did not cycle in WT (Figure 2.8). With the caveat that lack of statistical significance does not prove absence of a cycle in WT, this finding suggests that the clock helps to suppress fluctuations in ribosome loading driven by diel light-dark cycle conditions. Likewise, de-repression of cycles in the CCA1-ox strain was also evident at the transcript level.

Global clock control of ribosome loading

While alternative splicing has been implicated repeatedly as a clock output and as a regulator of circadian clock function, clock control of translation or ribosome loading has rarely been examined. Aside from the global cycles of ribosome loading (Piques, Schulze et al. 2009, Pal, Liput et al. 2013) in *Arabidopsis*, translation of the LHY mRNA was shown to be stimulated by light, a phenomenon thought to sharpen the peak of LHY protein abundance in the morning (Kim, Song et al. 2003). Our results also showed higher ribosome loading of LHY in the

morning, when LHY mRNA peaks, than at the end of the day, when the mRNA is low. This pattern of translational regulation may keep LHY protein levels from rising too early in the night.

Translational control of a clock output has been described in the dinoflagellate Gonyaulax and in the green alga Chlamydomonas (Morse, Milos et al. 1989, Mittag, Lee et al. 1994). More recently, translational regulation of ribosomal protein mRNAs in the mouse was shown to peak at night (Jouffe, Cretenet et al. 2013), which is at first glance similar to the situation in Arabidopsis. However, mice are nocturnal animals, feeding at night and living off their fat deposits during the day. In Arabidopsis, in contrast, energy is harvested and thus more abundant during the day, while the plant lives off its starch deposits at night. Thus, considering the overall energy infrastructure of the mouse and Arabidopsis, the resemblance in the nocturnal peaks of ribosomal protein translation is probably coincidental.

Synthesis of ribosomal proteins occupies a substantial fraction of translational capacity especially in growing cells (Warner 1999, Piques, Schulze et al. 2009). The question arises why the plant would translate ribosomal proteins preferentially at night rather than during the day when energy is more directly available, which would circumvent losses during starch deposition and conversion. The reason is unknown, but may be associated with the following. Translational capacity is finite. Ribosomes lying idle in the cell may be a sign of poor stewardship of growth-limiting resources, especially nitrogen. During the light period, the cell may utilize most of its translational capacity for bulk maintenance, including maintenance of the photosynthetic apparatus, leaving little capacity for ribosomal protein translation. Thus, the finding that ribosomal protein translation preferentially occurs at night may suggest that the cell has spare capacity at that time.

Integration of clock-controlled diel translation with other signals

Most signals known to regulate translation are exogenous environmental cues, including light, darkness, drought, and temperature. In contrast, only a few endogenous cues are known to regulate translation, e.g. sucrose and amino acids (Nicolai, Roncato et al. 2006, Lageix, Lanet et al. 2008, Zhang, Wang et al. 2008, Roy and von Arnim 2013). The clock is another endogenous mechanism now known to affect translation. Clock-controlled and diel ribosome loading must be integrated with other signals that affect ribosome loading simultaneously. These signals can be expected to play a relatively minor role under constant light conditions on medium containing sucrose, but will affect what happens under light-dark cycle conditions. For example, polysome loading rises rapidly upon lights-on, peaking after 1-4 hours and then slowly declining towards the end of the day. Conversely, a transient drop in polysome loading occurs around 15-30 minutes after lights-off (Pal, Liput et al. 2013). How the effects of light-dark transitions, day length, photosynthate, and other signals are integrated with signals from the clock will be an important area of future research.

Methods

Plant material and polysome gradient fractionation

Arabidopsis ecotype Columbia was grown on agar plates containing full strength Murashige and Skoog salts, pH 5.7, and 1% sucrose for ten days in a 16-hour light (~80 μmole/m^2/sec) and 8-hour dark cycle at 22˚C. The CCA1-ox strain overexpresses the CCA1 protein under the control of the cauliflower mosaic virus 35S promoter (35S:CCA1, (Wang and Tobin 1998) and was grown likewise. Three biological replicates were collected. RNA extraction

and microarray hybridization closely followed an earlier procedure (Kim, Cai et al. 2007), with a few modifications described earlier (Missra et al., 2014). After sucrose-gradient fractionation we generated three fractions of mRNAs: the non-polysomal fraction (NP), the small polysomes (SP, 1-3 ribosomes per mRNA), and the large polysomes (LP, 4 and more ribosomes per mRNA) (Supplemental Figure 1). We measured the RNA abundance after fractionation in order to reveal the global shift in ribosome loading over the course of the day. Samples for total transcripts were also collected alongside. Following manufacturer's protocols, LP, SP, and NP and total RNA fractions were converted to cDNA and hybridized to GeneChip® Arabidopsis ATH1 Genome Arrays, which contain 22,746 probe sets representing approximately 24,000 genes.

If a given experimental treatment causes a global reduction in polysome loading, the global shift becomes masked during the standardized experimental procedure. The global shift is measured from RNA abundance data after fractionation (Kawaguchi, Girke et al. 2004). A small global shift was detected with a peak at ZT6 (noon) and trough at ZT0 (end of night) (Figure 2.1). However, no global adjustment of our array data was performed. Therefore, a gene that displays a TL cycle identical to the global shift will be regarded as non-cycling. The TL cycles described in Figure 2.2 are cycles beyond the global cycle.

Microarray data analysis

Statistical and bioinformatic analyses were carried out using R version 3.1.1 (R Core Team, 2014) and Bioconductor version 2.14. Raw signal intensities for each probe set were extracted from CEL files, the Affymetrix (Santa Clara, CA) proprietary data format, using the *affy* package version 1.44.0 (Gautier et al., 2004) and normalized using the *gcrma* package version 2.38 (Wu et al.), all with default settings. Normalization using the *rma* method from the

affy package yielded similar results on this dataset. Hybridization signals were classified as Present (P), Marginal (M), or Absent (A), using the *mas5calls* function from the *affy* package. qRT-PCR was performed on a Bio-Rad iQ5 instrument. Primer sequences for qRT-PCR are listed in Supplemental Table 1.

Calculation of translation states

Translation states (TLs) were calculated for every time point for genes with P calls in all four SP and LP samples, while A calls were permitted in the NP samples. Genes with a variance below 0.001 in hybridization signals for any time point were discarded as artifactual, resulting in 12,342 genes in WT. Signals from the *gcrma* output were unlogged and TL was calculated according to the following formula:

TL = (0 x NP + 2 x SP + 7 x LP) / (NP + SP + LP)

This calculation is based on the estimate that mRNAs in the NP, SP, and LP fractions are bound by an average of, respectively, zero, two, and seven ribosomes. If all ribosomes were equally active in translation, then TL would indicate a translation rate of proteins produced per mRNA per unit time. We cannot rule out that the average number of ribosomes per mRNA varies with time, for example, it may be lower than seven in the LP at ZT0, when the global polysome loading is lowest. Because we were not able to settle on more precise estimates for each timepoint, the given values of two and seven were applied to all samples equally. The translation profiles may be skewed slightly as a result. The abundance of total mRNAs (TX) was displayed on a $\log_2$ scale, as usual.

Identification of genes with diel fluctuation of their translation state

For each gene, the difference between the peak TL and trough TL is the ΔTL value. Differentially translated mRNAs were identified by one of four criteria, ΔTL >0.3, ΔTL >0.7, ANOVA with p<0.05, and SAM (Tusher, 2001) with a collective FDR <0.10. The lenient cutoff of 0.3 was only used to identify those genes that clearly lack a translation cycle. For SAM (Tusher, et al., 2001) we ran the R function "samr" with response type "Multiclass" to identify genes whose TL varied significantly across time points. We used 1000 permutations to compute a test statistic (T) for each gene, as well as a null distribution of 1000 Ts for each gene obtained from random shuffling of the class labels. We then computed an empirical p-value for each gene as the number of randomized Ts that were greater than the original, divided by 1000. As an additional check, we permutated the timepoint labels of our WT TL data and then used SAM to search for false-positive TL cycles. Such permutations yielded an average of only 46 translationally cycling genes (range: 0 to 312), as compared to 1825 from the original data. For ANOVA with a raw p-value <0.05 we also calculated the FDR per gene using the Benjamini-Hochberg method.

Modeling diel cycles as sine waves

We modeled diel variation in TX and TL as sine waves using a linear model approach to precisely estimate the phase and amplitude, as described elsewhere in more detail (Stolwijk et al., 1999). A sine wave can be described by the function $y(t) = A \times sin\left(\frac{2\pi t}{T} - \varphi\right)$, where t is time, A is the amplitude, T is the period, and φ is the phase. This sine function can be transformed into a linear regression formula, $f(t) = \beta_1 \times sin\left(\frac{2\pi t}{T}\right) + \beta_2 \times cos\left(\frac{2\pi t}{T}\right)$, and fitting

a linear model with this formula yields the coefficients β_1 and β_2. The amplitude can then be calculated as $\sqrt{\beta_1^2 + \beta_2^2}$, and the phase as $\arctan\left(\frac{\beta_1}{\beta_2}\right)$. For each gene, we first subtracted the mean TX or TL from the data series and then followed this approach to estimate the phase and amplitude with the "lm" function from the standard R "stats" package, assuming a period of 24 hours. The Pearson coefficient (R^2) between empirical data and predicted data indicates the percent of the variation in TX or TL for a gene that is explained by the sine wave. This approach is convenient due to its simplicity and computational speed, and it has the advantage that the phase can be any value over a continuous range from 0 to 2π (ZT0 to ZT24), so the peak and trough of TX or TL can occur between the times at which data were collected. A confidence interval was calculated for each TL peak time using bootstrapping.

Clustering and higher level analyses

mRNAs with diel fluctuations in TL were clustered using R according to the time of peak TL and, secondarily, the time of trough TL. Hierarchical clustering was performed using the Pearson coefficient as the similarity metric. This and all overlap plots were made using the heatmap.2 function from the *gplots* package in R package version 2.14.2 (Warnes et al., 2014). Where indicated, the data from individual genes were mean-centered and row-scaled by their Z-score (distance from the mean as multiples of their standard deviation) to better display trends in translation over time. When individual expression values in a time series did not pass our data quality filter the expression value was replaced with NA. Enrichment of functional annotations was determined using the *topGO* R package version 2.18 (Alexa et al., 2010) with all 12,342 expressed genes as the reference set. For gene ontology analysis, genes were preselected using

SAM, or ANOVA, or a ΔTL value above 0.7. Functional terms had to be significantly enriched with an FDR of 0.05 or less in either WT or CCA1-ox to be considered for presentation. Terms that were substantially overlapping with other terms were omitted.

Accession numbers

The original microarray hybridization data, metadata, and extracted and normalized data are accessible in NCBI-GEO under superseries GSE61899: Polysome profiling in wild type and CIRCADIAN CLOCK ASSOCIATED1-overexpressing (CCA1-ox) Arabidopsis thaliana over a 24-hour diel cycle. Four datasets are grouped together under GSE61899: GSE61895, transcript levels in WT; GSE61896, transcript levels in CCA1-ox; GSE61897, polysome WT; GSE61898, polysome CCA1-ox. A list of AGI numbers is provided in Supplemental Table 2.

Chapter 3

Computational modeling of mRNA ribosome loading

Abstract

A computational model of translation was previously developed that describes how five steps in gene expression—transcription, initiation, elongation, marking for degradation, and degradation—control ribosome loading of mRNA. Empirical data were previously collected with the goal of fitting the model to the data in order to quantify the five rates for many genes across the Arabidopsis genome. Despite significant progress, various contributions made toward this goal have remained largely isolated due to the inter-disciplinary nature of this collaborative research. In this chapter, I have synthesized the past contributions, including explaining the background and procedures involved in terms of biological and computational aspects. I have expanded the usability of the computer code for implementing the model in terms of user-friendliness, flexibility, and performance. And I have explored the model's behavior through simulation studies. A major challenge is that the model output cannot be directly matched to the empirical data, and work continues in order to improve the current procedure. This chapter serves as a progress report on the work done by people in the Gilchrist and von Arnim labs in developing and testing the model.

Preface

Cellular organisms change their physiology in order to grow and develop and to adapt to a changing environment. A key aspect of this capacity for physiological change at the cellular level is the ability to alter the abundance of proteins. A cell controls protein synthesis by controlling the abundance of each mRNA species—through transcription and degradation—and the rates of their translation—through initiation and elongation. Regulation of these processes

and their effects on ribosome loading and translation were reviewed in chapter 1. To summarize, ribosome loading and translation are influenced by a variety of post-transcriptional factors, which can be classified as either fixed properties of each mRNA species or dynamically regulated processes which respond to a cell's internal or external environment. Physiologically, the most important net effect of these transcriptional and post-transcriptional factors is the resulting protein abundance. However, measuring protein abundance alone is limited in that it does not provide information about how various factors combine to yield a particular protein abundance, and despite advances in instrumentation and analytics, proteomics measurements are still typically limited to a few thousand different proteins.

The von Arnim lab and others have established that ribosome loading in Arabidopsis is regulated in a genome-wide and gene-specific manner by numerous environmental conditions. However, the extent to which each gene utilizes transcription and various post-transcriptional processes in modulating ribosome loading is not well understood. The Gilchrist lab developed a computational model that describes how a particular pattern of ribosome loading, that is, the amount of mRNA with different numbers of ribosomes bound, results from the combined effects of five processes which will be referred to as "translation parameters"—rates of transcription, initiation, elongation, marking, and degradation. The von Arnim lab generated genome-wide mRNA measurements from nine polysome fractions as well as un-fractionated mRNA. The two research groups share the goal of using the model and the empirical data to quantify the five parameters for many genes across the Arabidopsis genome. Ultimately the model may be used to describe how rates of the translation parameters change in response to changing environmental conditions.

Quantifying the parameters for an individual gene requires an approach for determining whether a particular set of parameter values are consistent with the empirical data for the gene. This in turn requires predicting the empirical data based on the parameter values. Mathematical optimization is used to search the parameter space for a solution that matches the empirical data. In the research presented, the task of accurately predicting the empirical data is especially challenging because the model output does not have the same format as the data that were collected. After reviewing the experimental procedures, implementing and testing the computational procedures, and attempting to fit the model to the empirical data, it is clear that the current procedure for converting the model output into the same format as the empirical data is flawed. Largely for this reason, efforts to fit the model to the empirical data have been unsuccessful, and further work is needed to improve the procedure. Thus, while the ultimate goal of this research is to apply the model to the empirical data, this chapter will focus on describing the context of the research, the computational procedures and their rationale, model implementation improvement, insight from simulation studies, and areas for improvement.

A number of people have been involved at various stages of this project. Drs. von Arnim and Gilchrist conceived the conceptual basis for the model and how it could be applied to the empirical data. Dr. Gilchrist formalized the model mathematically and, with assistance from his student, Nate Pollesch, first implemented it in computer code. Dr. von Arnim oversaw the data collection, including generating plant samples and collecting the microarray data, which was primarily done by a post-doctoral fellow, Dr. Ju Guan. And Joe Carboni began an implementation of the model in R. A major challenge has been to bring together the contributions made by the various people in order to progress toward the goals of the research, which requires synthesizing concepts and methodologies from mathematics, computer science,

and biology. Moving the research forward requires a comprehensive and centralized documentation of the computational and biological rationale and procedures, including data collection and model testing, which was previously lacking.

In addition to synthesizing the work already done, moving the research forward would also require improving the usability and efficiency of the implementation of the model. The model had been implemented in Wolfram Mathematica, which is a suitable language and computing platform. However, for practical considerations, a flexible and robust R version that expanded on the initial implementation would offer several benefits. First, R is open source, while the Mathematica version was limited in its parallel computing capacity due to its commercial license. Second, R's wide use across scientific disciplines and the availability of R software for computational biology made it an ideal computing platform for continuing this collaborative project. Third, efficient exploration and application of the model would require additional computational procedures, for instance, better automation, utilization of high-performance computing, and visualization, in addition to flexibility in choices of various algorithms, inputs, and statistical methods.

This chapter is organized as follows. In the introduction, I provide context for the research by giving overviews of existing translation models and mathematical optimization, and I describe the model. In the methods, I describe collection and processing of the empirical data and the procedures used to implement and improve the model. In the results, I describe simulation studies that should aid future work, and I report results from fitting the model to the empirical data. Lastly, in the discussion, I summarize the insights gained from this research and areas that need improvement.

Introduction

Existing translation models

Computational models are useful for studying biological processes when they are more practical than direct experimentation. Such is the case for some of the dynamic aspects of translation. Current high-throughput techniques for studying translation, including polysome profiling, ribosome footprinting, and proteomics, even paired with isotope-labeling, do not have single-molecule resolution, so molecular events cannot be connected to specific mRNA molecules. Thus, dynamic events such as recruitment of a pre-initiation complex (PIC) to a 5' untranslated region (5' UTR), scanning of a PIC along a 5' UTR, movement of a ribosome along an mRNA, and interactions between mRNA and bound and unbound ribosomes, are difficult to monitor directly, so instead their net effects are often quantified in a population of molecules. Further, the question of how multiple processes combine in a gene-specific manner to yield a protein synthesis rate is difficult to approach through direct experimentation. Because of these issues, a number of computational models have been developed in order to investigate some of the dynamic aspects of translation which are often intractable (von der Haar 2012). These are described below and summarized in Table 3.1.

Stochastic translation models

A number of translation models are stochastic, incorporating randomness. In stochastic models, a system is represented as various types of reactions or steps, such as recruitment of a PIC to mRNA or translocation of a ribosome from one codon to the next. Each reaction is associated with a probability of it occurring in some time interval, with the probability depending on other probabilistic factors, such as whether a neighboring codon is occupied by a ribosome

and the aminoacyl-tRNA (aa-tRNA) concentration. Thus, model predictions depend on each of the probabilities and can fluctuate, so repeated simulations may be required before predictions converge on a stable expected outcome. Stochastic models can be powerful and accurate portrayals of cellular processes because many cellular processes are stochastic in nature, being dependent on stochastic processes like diffusion.

Agent-based models

Agent-based models (ABMs) are a class of stochastic models that simulate a system of components or "agents" and their interactions based on rules. The purpose is often to uncover emergent properties of the system that would be difficult to decipher by analyzing the individual components or rules separately (An 2001, Bankes 2002, Bonabeau 2002, Grimm, Revilla et al. 2005, Macal and North 2010). Chu et al. developed an ABM that simulates the behavior of all mRNAs and ribosomes in a system (Chu, Zabet et al. 2012, Chu, Thompson et al. 2014). In the system, mRNAs compete with each other for ribosomes and aa-tRNAs according to first-order kinetics, and an elongating ribosome progresses from one codon to the next if the correct aa-tRNA has bound to the next codon. Ribosomes may collide with each other along an mRNA, causing "traffic jams" which slow translation. This group applied their ABM to yeast, adding support to the hypothesis that the elongation speed for different codons can significantly affect the overall protein synthesis rate, but with two caveats. First, when there are a large number of ribosomes translating an mRNA, the translation rate depends on the arrangement of fast and slow codons, which can potentially cause traffic jams. Second, the elongation speed does not affect the overall protein synthesis rate if the initiation rate is slow.

Table 3.1: Examples of translation models

Class	Sub-type	Notes	Applications	References
Stochastic	Agent-based, TASEP	Represent translation as particles moving along 1-D lattice	Describe parallel bio-polymerization in a polysome	(MacDonald, Gibbs et al. 1968)
			Describe interactions among mRNAs, tRNAs, and ribosomes	(Shah, Ding et al. 2013)
			Relate codon usage bias to translation efficiency	(Gilchrist and Wagner 2006)
			Predict ribosome density from codon adaptation, amino acid charge, and mRNA structure	(Tuller, Veksler-Lublinsky et al. 2011)
			Determine conditions in which elongation affects translation rate	(Chu, Thompson et al. 2014)
Deterministic	ODEs	Limited scale due to number of equations	Describe dependence of initiation on 5' UTR and poly(A) tail	(Bi and Goss 2000)
			Describe linear and non-linear protein buildup	(Gerst and Levine 1965)
	Ribosome flow		Determine codon order to optimize translation	(Zarai, Margaliot et al. 2014)

The totally asymmetric simple exclusion process

The totally asymmetric simple exclusion process (TASEP) is a stochastic modeling approach that describes movement of particles along a one-dimensional lattice of discrete positions (Spitzer 1970). Each position can either be empty or occupied by a particle, and a particle at one position blocks movement of the particle at the position behind it. Particles advance along the lattice according to probability. Particle reservoirs are at each end of the lattice, so a particle at the last position on the lattice can jump into the reservoir and a particle in the reservoir can jump onto the first position if it is empty. The goal of TASEP is often to determine what set of probabilities for the various events lead to a particular outcome (Blythe and Evans 2007). The applicability to translation, being the movement of ribosomes (the particles) along codon positions of an mRNA (the lattice), is obvious. Some of the first reported models of translation used concepts that were later incorporated into TASEP (MacDonald, Gibbs et al. 1968, MacDonald and Gibbs 1969), and TASEP has continued to be used for this general purpose (Shaw, Zia et al. 2003, Zia, Dong et al. 2011).

Concepts of TASEP have been applied in different types of translation models, including kinetic models which have been used to explore mechanisms that dictate protein synthesis rates (Vassart, Dumont et al. 1971, Heinrich and Rapoport 1980, Chou and Lakatos 2004, Gilchrist and Wagner 2006, Mitarai, Sneppen et al. 2008, Zia, Dong et al. 2011, Chu and von der Haar 2012, Chu, Zabet et al. 2012). Shah et al. developed a TASEP model, more specifically a Markov chain model, that tracks all mRNAs, ribosomes, and tRNAs, which they applied to yeast (Shah, Ding et al. 2013). In their model, ribosomes and tRNAs can either be diffusing freely in a cell, or they can be engaged in elongation at a specific codon along an mRNA. This model improved upon previous TASEP models in that it assumed a limited supply of ribosomes.

Parameterizing the model with empirically-derived measurements, they predicted which genes were rate-limited by initiation and which were rate-limited by elongation. Additionally, they were able to recapitulate the "ramp" feature of both types of mRNAs in which ribosome density decreases in the 5'→3' direction (Ingolia, Ghaemmaghami et al. 2009, Reid and Nicchitta 2012). The reason was due to an increased initiation rate rather than slow codons near the start codon, as reported elsewhere (Tuller, Carmi et al. 2010, Reuveni, Meilijson et al. 2011, Tuller, Veksler-Lublinsky et al. 2011). In addition, they predicted parameter regimes under which certain behaviors would be observed, for example, for which combinations of mRNA concentrations and codon adaptation index (CAI) scores would codon usage bias influence the protein synthesis rate.

Deterministic models

A number of translation models are deterministic, not involving any randomness.

Ordinary differential equations

Ordinary differential equations (ODEs) have been used in translation models, but for different purposes than ABMs and TASEP (von der Haar 2012). To describe all possible elongation reactions for an mRNA with l codons and a maximum of n ribosomes attached would require up to $(l-1)^n$ ODEs when all possible combinations of ribosome occupancies are considered. Further, ODEs applied in this way would not account for ribosomes hindering each other's movement, as some stochastic models have done by incorporating probabilities of ribosomes advancing along an mRNA that are conditional on there being no ribosomes blocking them. Thus, translation models using ODEs to describe the detailed kinetics of elongation have been limited in scale (Gerst and Levine 1965, Garrick 1967, Singh 1996, Heyd and Drew 2003,

Zouridis and Hatzimanikatis 2007, Zhang, Fedyunin et al. 2010). Nevertheless, ODEs and stochastic differential equations (SDEs) have been used to compute numerical approximations describing certain aspects of translation (Bi and Goss 2000, Berthelot, Muldoon et al. 2004, Dimelow and Wilkinson 2009, De Silva, Krishnan et al. 2010, You, Coghill et al. 2010). Dr. Gilchrist's model presented in this chapter uses a system of ODEs to describe how ribosome loading depends on transcription, initiation, termination, marking, and degradation.

Ribosome flow model

The ribosome flow model (RFM) describes the flow of ribosomes along an mRNA in terms of its parameters, an initiation rate and either a constant (Reuveni, Meilijson et al. 2011) or codon-specific (Poker, Zarai et al. 2014) elongation rate. Because the steady-state translation rate is a concave function of the initiation rate and one or more elongation rates, a global maximum translation rate under a given set of constraints can be computed and used to inform the design of efficiently translated mRNAs in synthetic biology (Reuveni, Meilijson et al. 2011, Poker, Zarai et al. 2014, Zarai, Margaliot et al. 2014).

Review of mathematical optimization

Mathematical optimization is the process of determining the best solution to a mathematical problem. In our case, the mathematical problem is an objective function, which quantifies how closely the model solution for a given set of parameter values matches the observed data (empirical or simulated). Our objective function is nonlinear, so determining the optimal parameter values requires nonlinear optimization. The "search space" or "parameter space" is the set of all possible parameter values from which the algorithm can choose. A "point" or "location" in the search space is one particular set of parameter values. Optimization

algorithms work by computing the value of the objective function repeatedly with different combinations of parameter values. At each iteration, the algorithm must decide whether it has reached an optimal solution or to keep trying. After each iteration, it decides which parameter values to change and by how much, which are simplifications of concepts known as the search direction and step distance, respectively. As discussed below, algorithms differ in the strategies they use to search the parameter space. A number of optimization algorithms are provided in the built-in "stats" package ("optim" function) and the "optimx" package ("optimx" function) in R. Here, I review these algorithms briefly since they have been reviewed elsewhere (Nocedal and Wright 1999, Bonnans, Gilbert et al. 2006, Venter 2010, Marthaler 2013, Nash 2014). Specific references to methods available in optim or optimx are indicated with bold font, and these methods are summarized in Table 3.2.

Nonlinear optimization algorithms can be classified according to several types of criteria. First, they can be designed for convex or nonconvex optimization problems.

Convex vs. nonconvex optimization

In a convex problem, the values of the objective function form a convex set in which there is a globally optimal solution (Nash 2014). Graphically, a convex function forms a bowl shape in two or three dimensions. In a nonconvex problem, such as the one in our study, the objective function is not convex, so multiple local optima are possible. Methods for solving convex and nonconvex problems may use gradient or non-gradient methods.

Gradient optimization algorithms

Conceptually similar to a derivative, a gradient is the rate of change in the objective function at a given point in the parameter space (Nash 2014). Gradient methods use the gradient in deciding how to vary the search direction and step distance, and their two classes, line search and trust region methods, differ in how they use gradient information.

Line search optimization algorithms

Line search methods compute the search direction at each iteration and then choose a step distance from a sample of trial steps (Nocedal and Wright 1999). Line search methods differ in how they choose the search direction. Steepest descent (or gradient descent) methods choose the direction as the one along which the objective function decreases most rapidly. Conjugate gradient (CG) methods improved upon steepest descent's efficiency by using information from previous search directions. An early and popular CG method is that of Fletcher-Reeves ("**CG**") (Fletcher and Reeves 1964), with an update added in the **Rcgmin** method in optimx. The spectral projected gradient (**SPG**) method speeds convergence for convex optimization by projecting the search direction into the search space without requiring that each iteration decreases the objective function (Birgin, Martínez et al. 2001).

Newton optimization algorithms

Newton methods are a sub-type of line search methods. At each iteration, Newton methods first formulate a quadratic function, containing the parameters of the objective function, that approximates its behavior in the local search space. Second, they compute the so-called Hessian matrix containing the second derivatives of all parameters elementwise with respect to all other parameters, as well as the determinant of the Hessian matrix, which is known as the Hessian and describes the local curvature of the objective function in terms of multiple (possibly

many) parameters. Newton methods, such as the "**nlm**" and "**nlminb**" methods in optimx, compute the Hessian in each iteration, which can hinder performance on large-scale problems with many parameters (Nash 2014). However, nlminb performed best out of all algorithms tested in terms of identifying the true parameter values in the simulation study presented below, perhaps because of the small number of parameters in the objective function.

Modified Newton optimization algorithms

To improve the efficiency of the Newton method, modified Newton methods have been developed that combine elements of the Newton method with other line search methods. One popular variation, the quasi-Newton or variable metric method, approximates the Hessian once and then updates it in each iteration with new information, which speeds convergence (Nash 2014). Quasi-Newton variants, such as symmetric-rank-one (SR1) and **BFGS**, differ in the formulas they use to update the Hessian approximation (Nocedal and Wright 1999). L-BFGS is a limited-memory quasi-Newton method that uses the BFGS formula but uses information from fewer past iterations to update the Hessian approximation (Byrd, Lu et al. 1995). **Rvmmin** and **L-BFGS-B** use the BFGS and L-BFGS strategies, respectively, but also facilitate constraints on the parameters being searched (Nash and Varadhan 2011).

Trust region optimization algorithms

In contrast to line search methods, which first choose the descent direction and then the step distance, trust region methods first choose a step distance, known as a trust region, and then choose the search direction. **Newuoa**, **bobyqa**, **ucminf**, and Levenberg–Marquardt are examples of nonlinear optimization algorithms that use trust region strategies (Nielsen and Mortensen 2009, Bates, Mullen et al. 2014).

Derivative-free optimization algorithms

Derivative-free methods, such as **Nelder-Mead**, **newuoa**, and **bobyqa**, avoid computing derivatives and so do not use the gradient (Nash and Varadhan 2011, Nash 2014). For example, Nelder-Mead computes the objective function at vertices of a simplex, which is an extension of a polygon to additional dimensions. In each iteration, the simplex is stretched, shrunken, and/or rotated so that it moves toward the lowest possible value of the objective function.

Stochastic optimization algorithms

Finally, optimization algorithms can be stochastic. Simulated annealing ("**SANN**") is a stochastic algorithm that is typically applied to discrete data (Nash 2014). It continues to vary the parameter values by small amounts as long as the objective function decreases. However, if the objective function does not decrease, it uses a stochastic process to decide whether to move to a new search location. The probability of moving is based on a "temperature" parameter and decreases with each iteration. It repeats this process many times per iteration, keeping track of the parameter values that produced the lowest value of the objective function. Given enough time, simulated annealing is guaranteed to find the global optimum.

The model

Basic model overview

The model uses the five translation parameters to predict the steady-state amounts of an mRNA species having different numbers of ribosomes attached. It describes a system containing two states, "unmarked" and "marked", with each state containing a number of "ribosome classes" (Figure 3.1). Based on the transcription rate (λ), mRNA enters the system in unmarked class 0 (m_0), in which there are no ribosomes associated with the mRNA. i_{max} denotes the

highest ribosome class for a particular mRNA species, which is its maximum ribosome occupancy. From m_0 through $m_{i_{max}-1}$, mRNA moves up in class according to the initiation rate (κ). mRNA moves down in class at the rate at which ribosomes complete elongation and termination, which is referred to here as simply the elongation rate (τ). From any unmarked class m_i, mRNA enters the corresponding marked class m_i^* according to the marking rate (μ). From any class in the marked state, initiation cannot occur, but ribosomes that are already attached can terminate, so mRNA still moves down in class according to the elongation rate. From marked class 0 (m_0^*), mRNA exits the system according to the degradation rate (δ).

Table 3.2: Survey of nonlinear optimization algorithms available in optim and optimx in R

Algorithm	Optimx input code	Search type	Derivative type	Allows constraints	Comments	References
Nelder-Mead	Nelder-Mead	Local, deterministic	Derivative-free	No	Simplex method; popular; default in R	(Nelder and Mead 1965)
Broyden-Fletcher-Goldfarb-Shanno	BFGS	Local, deterministic	2nd	No	Quasi-Newton; updates Hessian with BFGS formula	(Fletcher 1970, Nash 1979)
Conjugate gradient	CG	Local, deterministic	1st	No	Poor convergence; ideal for large problems; improved in Rcgmin	(Fletcher and Reeves 1964)
Limited-memory BFGS with box constraints	L-BFGS-B	Local, deterministic	1st	Yes	Default for constrained problem; most flexible; ideal for large problems	(Byrd, Lu et al. 1995)
UNCMIN	nlm	Local, deterministic	2nd	No	Good convergence and efficiency	(Schnabel, Koonatz et al. 1985, Dennis and Schnabel 1996)
Nonlinear minimization subject to box constraints	nlminb	Local, deterministic	2nd	Yes	Gradient function recommended; little documentation	(Fox, Hall et al. 1978, Gay 1983, Gay 1984, Dongarra and Grosse 1987)
Spectral projected gradient	spg	Local, deterministic	1st	Yes	Efficient on large-scale problems	(Birgin, Jos et al. 2001, Varadhan and Gilbert 2009)
General purpose unconstrained optimization	ucminf	Local, deterministic	1st	No	Blend of line search (BFGS) and trust region; similar to Rvmmin	(Nielsen 2000)
New unconstrained optimization algorithm	newuoa	Local, deterministic	Derivative-free	No	Model-based	(Bates, Mullen et al. 2014)
Bound optimization by quadratic approximation	bobyqa	Local, deterministic	Derivative-free	Yes	Newuoa with constraints	(Bates, Mullen et al. 2014)
Nelder-Mead optimization algorithm for derivative-free optimization	nmkb	Local, deterministic	Derivative-free	Yes	Nelder-Mead with constraints	(Kelley 1999)
Hooke-Jeeves derivative-free minimization	hjkb	Local, deterministic	Derivative-free	Yes	Reliable; inefficient	(Kelley 1999)
Conjugate gradient (with Dai/Yuan update)	Rcgmin	Local, deterministic	1st	Yes	Reaches global optimum if conditions are met	(Nash 1979, Dai and Yuan 2001)
Variable metric nonlinear function minimization	Rvmmin	Local, deterministic	1st	Yes	Constrained version of BFGS	(Nash 2014)
Simulated annealing	SANN (in optim)	Global, stochastic	Derivative-free	No	Approximates global optimum with enough iterations	(Claude, xe et al. 1992)

i_{max} for each gene is not known with certainty, but a number of studies have suggested a theoretical upper limit of 1/30th of the coding sequence length in nucleotides. For instance, studies involving ribosome-protected mRNA have indicated that in a number of species, including Arabidopsis, one ribosome protects approximately 30 nucleotides of mRNA (Steitz 1969, Ingolia, Ghaemmaghami et al. 2009, Oh, Becker et al. 2011, Bazzini, Lee et al. 2012, Juntawong, Girke et al. 2014). Consistently, empirical measurements in yeast based on polysome profiling have suggested a maximum ribosome density of 3.3 ribosomes per 100 nucleotides (Arava, Wang et al. 2003). Nevertheless, it is not known whether these findings apply equally to mRNAs of different lengths, or how i_{max} influences ribosome loading for individual mRNA species.

Model details

The model is a system of coupled ordinary differential equations (ODEs) which relate the rates of change in mRNA abundance in all unmarked and marked ribosome classes to the rates of the translation parameters and the steady-state mRNA abundances in each class. Flux through a particular class is the difference between the flux into it and out of it. Table 3.3 contains definitions of symbols and variables used. Equations 3.1 and 3.2 contain the ODEs for the unmarked and marked states, respectively. The model predicts steady-state levels of mRNA in each ribosome class, so the equations are set to zero. Only one rate is used per parameter for a given gene, which assumes that the rates are independent of the number of ribosomes attached to the mRNA. The model details, including equations and symbols used, have been reproduced from Dr. Gilchrist and Nate Pollesch largely in the original format.

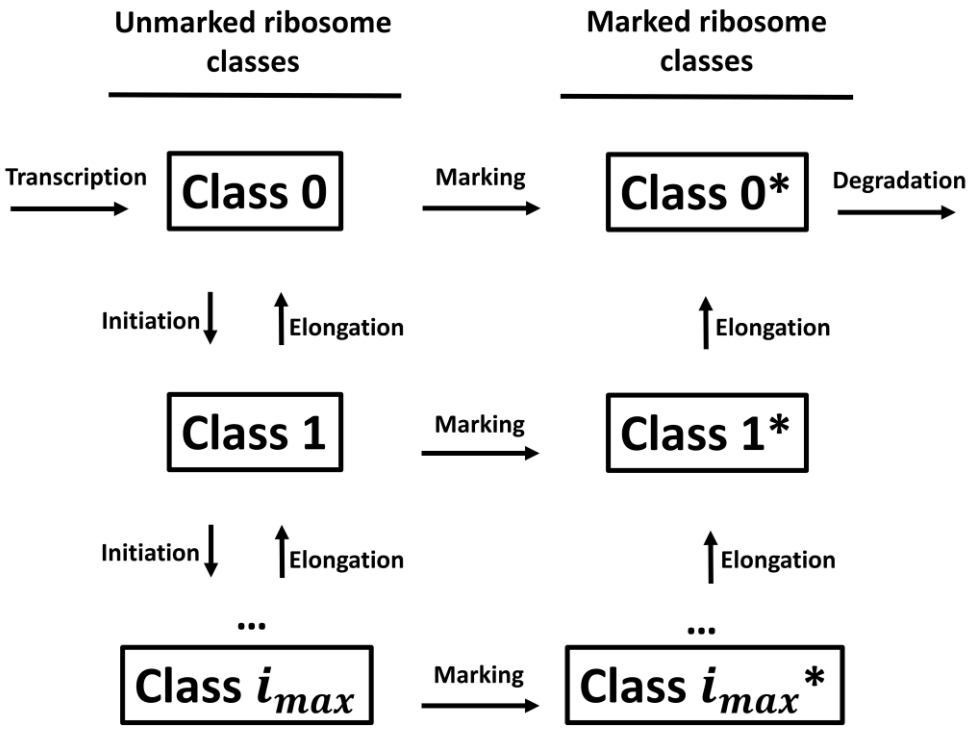

Figure 3.1: Model schematic.
The model describes the flux of an mRNA species among unmarked and marked ribosome classes. mRNA enters unmarked class 0 (m_0) according to the transcription rate (λ), where it is not associated with any ribosomes. i_{max} denotes the highest class that a particular mRNA species can occupy, which is the maximum number of ribosomes that can be associated with the mRNA. mRNA in class m_0 through $m_{i_{max}-1}$ moves up in class according to the initiation rate (κ). mRNA in all unmarked classes moves into the corresponding marked class according to the marking rate (μ). mRNA in all classes except m_0 and m_0^* moves down in class according to the elongation rate (τ). mRNA in the marked class 0 (m_0^*) exits the system according to the degradation rate.

Table 3.3: Variables and symbols used in the model

Symbol	Description
State variables	
$\boldsymbol{m_i}$	mRNA abundance in the unmarked i^{th} ribosome class
$\boldsymbol{m_i^*}$	mRNA abundance in the marked i^{th} ribosome class
Parameter definitions	
$\boldsymbol{i_{max}}$	Maximum number of ribosomes able to bind an mRNA
$\boldsymbol{\kappa}$	Translation initiation rate
$\boldsymbol{\tau}$	Translation elongation rate
$\boldsymbol{\mu}$	Marking rate
$\boldsymbol{\lambda}$	Production (transcription) rate for mRNA into the unmarked class 0 (m_o)
$\boldsymbol{\delta}$	Removal (degradation) rate for mRNA from the marked class 0 (m_0^*)

$$\frac{dm_0}{dt} = \lambda + \tau \cdot m_1 - \kappa \cdot m_0 - \mu \cdot m_0$$

$$\begin{aligned}
\frac{dm_1}{dt} &= \kappa \cdot m_0 + \tau \cdot m_2 - \tau \cdot m_1 - \kappa \cdot m_1 - \mu \cdot m_1 \\
&\vdots \\
\frac{dm_i}{dt} &= \kappa \cdot m_{i-1} + \tau \cdot m_{i+1} - \tau \cdot m_i - \kappa \cdot m_i - \mu \cdot m_i \\
&\vdots \\
\frac{dm_{i_{max}}}{dt} &= \kappa \cdot m_{i_{max}-1} - \tau \cdot m_{i_{max}} - \mu \cdot m_{i_{max}}
\end{aligned} \qquad (3.1)$$

$$\frac{dm_0^*}{dt} = \mu \cdot m_0 + \tau \cdot m_1^* - \delta \cdot m_0^*$$

$$\begin{aligned}
\frac{dm_1^*}{dt} &= \mu \cdot m_1 + \tau \cdot m_2^* - \tau \cdot m_1^* \\
&\vdots \\
\frac{dm_i^*}{dt} &= \mu \cdot m_i + \tau \cdot m_{i+1}^* - \tau \cdot m_i^* \\
&\vdots \\
\frac{dm_{i_{max}}^*}{dt} &= \mu \cdot m_{i_{max}} - \tau \cdot m_{i_{max}}^* - \mu \cdot m_{i_{max}}
\end{aligned} \qquad (3.2)$$

Solving the model

Solutions to the model are computed using a matrix algebra approach. The unmarked system can be represented as a matrix equation describing the rates of change in the abundance of mRNA in each unmarked ribosome class,

$$\begin{pmatrix} m_0' \\ m_1' \\ \vdots \\ m_i' \\ \vdots \\ m_{i_{max}}' \end{pmatrix} = \begin{pmatrix} -\kappa-\mu & \tau & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ \kappa & -\kappa-\mu-\tau & \tau & 0 & \dots & \dots & \dots & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & \kappa & -\kappa-\mu-\tau & \tau & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & \dots & \dots & \dots & \dots & 0 & \kappa & -\mu-\tau \end{pmatrix} \begin{pmatrix} m_0 \\ m_1 \\ \vdots \\ m_i \\ \vdots \\ m_{i_{max}} \end{pmatrix} + \begin{pmatrix} \lambda \\ 0 \\ 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \qquad (3.3)$$

Equation 3.3 can also be represented as

$$\vec{m}' = A \cdot \vec{m} + \vec{b}. \qquad (3.4)$$

At equilibrium, the rates of change in mRNA abundance in each unmarked ribosome class (left side of equation 3.3) are zero, so the mRNA abundances in the unmarked classes are described by

$$\vec{m} = -A^{-1} \cdot \vec{b}, \qquad (3.5)$$

which is equal to

$$\vec{m} = -\frac{1}{Det[A]} Adj[A] \cdot \vec{b}. \qquad (3.6)$$

$Adj[A]$ from equation 3.6 can be simplified and the equation restructured, yielding

$$\vec{m} = -\frac{1}{Det[A]} \cdot \lambda \begin{vmatrix} Det[A_{i_{max}-1}] \\ -s_1 Det[A_{i_{max}-2}] \\ -s_1 s_2 Det[A_{i_{max}-3}] \\ \vdots \\ (-1)^i \left(\prod_{j=1}^{i} s_j\right) Det[A_{i_{max}-(i+1)}] \\ \vdots \\ (-1)^{i_{max}-2} \left(\prod_{j=1}^{i_{max}-2} s_j\right) Det[A_1] \\ (-1)^{i_{max}-1} \left(\prod_{j=1}^{i_{max}-1} s_j\right) Det[A_0] \end{vmatrix}. \qquad (3.7)$$

In equation 3.7, $Det[A_o] = 1$, A_i is the $i \times i$ lower right sub-matrix, and s_i is the sub-diagonal entry in the i^{th} row in the full A matrix, or $\frac{i_{max}-(i-1)}{i_{max}}\kappa$.

The "Solve" command in Wolfram Mathematica and various manipulations were used to find an equation for $Det[A]$:

$$Det[A] = (-1)^{i_{max}+1}\mu\prod_{i=1}^{i_{max}}\left(\frac{i}{i_{max}}\kappa + \mu + i\tau\right). \qquad (3.8)$$

For finding the determinants of the submatrices in equation 3.7, one can take advantage of the fact that the large matrix in equation 3.3, or A in equations 3.4-3.7, is tri-diagonal, taking the general form,

$$\begin{pmatrix} d_0 & p_0 & 0 & 0 & 0 & 0 \\ s_1 & d_1 & p_1 & 0 & 0 & 0 \\ 0 & s_2 & d_2 & p_2 & 0 & 0 \\ 0 & 0 & \ddots & \ddots & \ddots & 0 \\ 0 & 0 & 0 & s_{n-1} & d_{n-1} & p_{n-1} \\ 0 & 0 & 0 & 0 & s_n & d_n \end{pmatrix}, \qquad (3.9)$$

where n is the size of the tri-diagonal matrix, which in our case is the number of ribosome classes (i_{max}+1).

The determinant of a sub-matrix A_i ($Det[A_i]$) of a general tri-diagonal matrix with n rows, such as that in equation 3.9, can be solved by a recurrence relation,

$$a(n,i) = d_{n-1}\cdot a(n,i-1) - p_{n-i}\cdot s_{n-i+1}\cdot a(n,i-2), \qquad (3.10)$$

with the following equalities,

$$\begin{aligned} d_i &= -\left(\frac{i_{max} - i}{i_{max}}\kappa + \mu + i\tau\right) \\ p_i &= (i+1)\tau \\ s_i &= \left(\frac{i_{max}-(i-1)}{i_{max}}\right)\kappa, \end{aligned} \qquad (3.11)$$

and with initial conditions,

$$\begin{aligned} a(n,1) &= d_n \\ a(n,2) &= d_{n-1} \cdot a(n,1) - p_{n-1}s_n. \end{aligned} \qquad (3.12)$$

The marked class abundances are then calculated as

$$m_i^{*eq} = \frac{\mu}{\tau}\frac{\sum_{j=i}^{n} m_i^{eq}}{\gamma(i)}, \text{ where } \gamma(i) = \begin{cases} \frac{\delta}{\tau} & i = 0 \\ i & i > 0 \end{cases}, \qquad (3.13)$$

where m_i^{*eq} is the mRNA abundance in the i^{th} marked class at equilibrium, and m_i^{eq} is the mRNA abundance in the i^{th} unmarked class at equilibrium.

Non-dimensionalization

The model solutions are not unique. Multiple sets of rates that have the same relative values will return the same ribosome class mRNA abundances. Therefore, when performing searches for the parameter values, one parameter can be kept constant. In this study, the termination rate is kept constant, so the model parameters can be interpreted as their rates relative to the termination rate, which naturally non-dimensionalizes the system. This is beneficial because it speeds up the parameter searches and increases the identifiability of the four parameters that are estimated, although I demonstrate in the simulation study that keeping additional parameters constant may be required.

Objective function and maximum likelihood estimation

The maximum likelihood estimation (MLE) approach is used to fit the model. The objective function minimized is the negative logarithm of a likelihood function (negative log-likelihood, or NLL). A likelihood function describes the likelihood of observing a specific value from a distribution based on specific parameters. The goal is to find parameter values that maximize the likelihood of observing the empirical data, or equivalently, minimize the negative likelihood. It is assumed that the measured mRNA abundance in each polysome fraction is log-normally distributed. The probability density function of the normal distribution gives the likelihood (lik) of a value x being observed from a normal distribution with mean μ and standard deviation σ:

$$Lik(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \qquad (3.14)$$

Out of convenience and convention, we minimize the NLL, which is the negative logarithm of equation 3.14, simplified as

$$-LLik(x) = \ln\left(\sigma\sqrt{2\pi}\right) + \frac{(x-\mu)^2}{2\sigma^2}. \qquad (3.15)$$

Since the first term in equation 3.15 is a constant, the second term only is minimized as our objective function. The denominator of the second term could even be omitted since σ, the assumed standard deviation among replicate measurements, is being kept constant in our procedures. The NLL for all of the data for one gene, empirical or simulated, is obtained by summing the results over all observations:

$$-LLik(G) = \sum_{i=1}^{n} \sum_{j=1}^{k} \frac{\left(g_{ij} - x_j\right)^2}{2\sigma^2}, \qquad (3.16)$$

where k is the number of polysome fractions, n is the number of replicates per fraction, G is the full set of values for a gene with k fractions and n replicates per fraction, x_j is the model

prediction for the j^{th} fraction, and g_{ij} is the empirical observation for the i^{th} replicate in the j^{th} fraction.

Methods

Empirical data

Microarray data

Genome-wide mRNA abundances were measured by microarray from polysome fractions collected along a sucrose density gradient from wild type Arabidopsis. The plant growth conditions and fractionation process, as well as cDNA preparation and microarray hybridization, were similar to those described in the Methods section of Chapter 2. Briefly, from one plant, three replicate sucrose gradients were run, and twelve polysome fractions were collected from the same positions along each gradient. For each replicate, fractions 1 through 4 were pooled and called the non-polysomal (NP) fraction, and fractions 5 through 12 (F5-F12) were kept separate. The NP fractions, F5-F12, and un-fractionated total mRNA samples for three replicates were analyzed by microarray, yielding 30 microarray samples altogether.

Microarray data processing and normalization

The microarray data were processed and normalized using the "ProcessRawData" function described in the procedures below.

Relationship between mRNA length and ribosome loading

mRNA length is correlated with ribosome loading in yeast (Arava, Wang et al. 2003). I examined the correlation between the coding sequence (CDS) length, obtained from the Ensembl BioMart database (Kasprzyk 2011), on the $\log_2$ scale and the percent of each mRNA species in

each fraction across the genome in Arabidopsis. I found modest positive and negative correlations (Figure 3.2), with fraction 7 having the highest negative correlation (R^2 = 0.21) and fraction 12 having the highest positive correlation (R^2 = 0.26). There was virtually no correlation for the NP fraction (R^2 = 0.007), which contains mRNA associated with 0 or 1 ribosome. The observation that the correlations get increasingly positive in higher polysome fractions is consistent with previous studies and with the idea that mRNA length is related to ribosome loading. This relationship could exist because of the influences of at least two factors. The first is a longer elongation time for longer mRNAs due to the longer distance a ribosome must travel. If ribosomes take longer to translate and dissociate from longer mRNAs than shorter ones, on average, then longer mRNAs would, on average, be associated with more ribosomes at any point in time. A second and closely related possible explanation is differences in codon adaptation between longer and shorter mRNAs. Studies in yeast have shown that shorter mRNAs tend to be more highly expressed and have stronger codon usage bias (CUB) than longer mRNAs (Coghlan and Wolfe 2000), so in addition to traveling a shorter distance, ribosomes may also carry out elongation at a faster rate on shorter mRNAs compared to longer ones. Despite the relationship between mRNA length and ribosome loading, the lack of correlation between mRNA length and the percent of each mRNA species in the NP fraction implies that factors other than length control ribosome loading of mRNA.

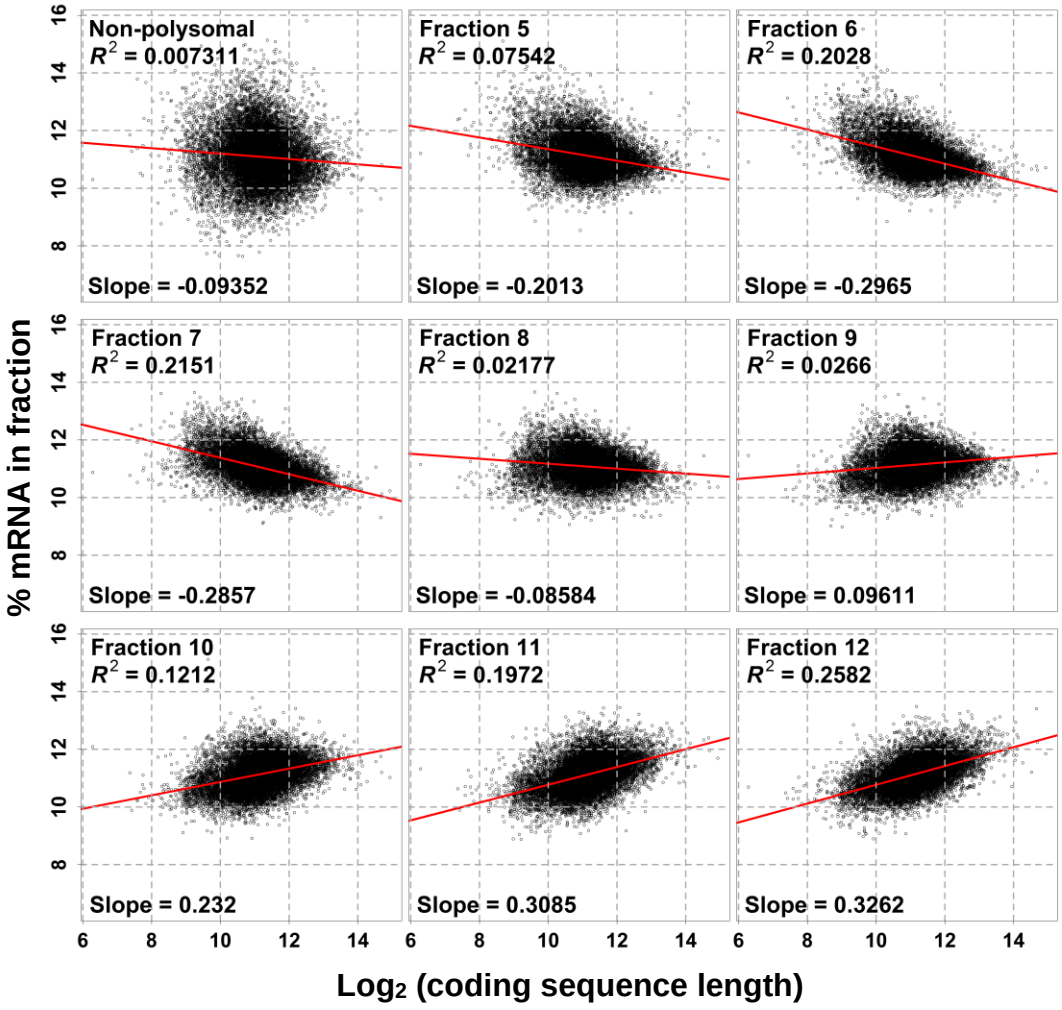

Figure 3.2: Correlations between gene length and the percent of mRNA in each polysome fraction. The red lines are regression lines fit to the data points which have the Pearson R^2 values and slopes indicated.

UV absorbance data

Larger polysomes migrate faster through a density gradient than smaller polysomes. UV absorbance is proportional to RNA concentration, so the resulting UV absorbance spectrum of the fractionated sample reflects the relative concentration of total RNA along the gradient, with migration distance along the horizontal axis and relative RNA concentration along the vertical axis. A representative example is shown in Figure 3.3.

Connecting the model with empirical data

The microarray data consist of genome-wide mRNA abundances in nine polysome fractions, while the model predicts mRNA abundances in a number of ribosome classes (unmarked and marked combined). Since the polysome fractions do not correspond exactly to ribosome classes, each fraction is a mixture of mRNA from multiple classes. Therefore, to predict the polysome fraction abundances from a set of parameter values, a procedure is needed for converting ribosome class abundances into fraction abundances. This involves, first, estimating how much mRNA from each ribosome class migrates into each polysome fraction, and second, correcting the predicted fraction abundances to account for variation in RNA yield across fractions.

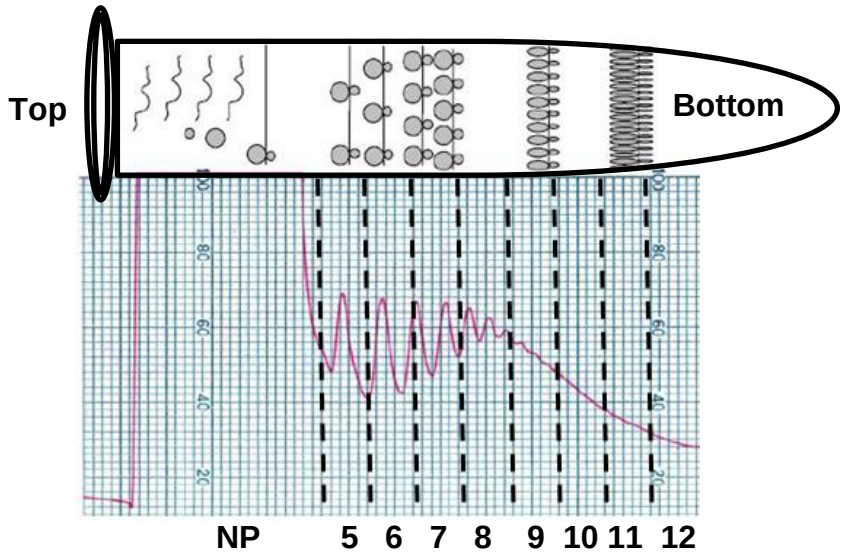

Figure 3.3: UV absorbance profile from the sucrose gradient.
The large peak on the left corresponds to mRNA with 0 or 1 ribosome. To the immediate right, moving left to right, the peaks correspond to discrete ribosome classes with specific numbers of ribosomes per mRNA molecule. Polysomes corresponding to 2 ribosomes up to approximately 8 are visually resolvable. The diagram of the test tube with polysomes of various sizes above the trace is meant to convey the idea that the larger polysomes migrate farther through the gradient. The approximate bounds of the fractions collected from the gradient are indicated as NP (non-polysomal) and 5-12 (polysome fractions 5-12).

Predicting fraction abundances from class abundances

To convert ribosome class abundances into fraction abundances, we must predict how far along the sucrose gradient a population of polysomes from a given class will migrate. We assume that the migration distances for a population of polysomes of a given class are normally distributed. Nate Pollesch used the coordinates from a UV absorbance profile to derive a function that predicts the mean migration distance for polysomes based on their class. The mean migration distances for the first eight ribosome classes were estimated by importing the UV profile image into Wolfram Mathematica and using the "get-coordinates" tool to determine the coordinates of each peak center. The first peak corresponds to RNA associated with at most one complete ribosome, and peaks 2 through 8 correspond to RNA associated with the same number of ribosomes as their rank (i.e., peak 4 corresponds to RNA from ribosome class 4, having four ribosomes per polysome). Ribosome class numbers and the center positions of their corresponding peaks were modeled using several types of functions in Mathematica using the "FindFit" tool, and the best-fitting type, based on least squares model fitting, was obtained:

$$\mu_{fit}(i) = \frac{a + b \cdot i^c}{d + i^c}. \qquad (3.17)$$

$\mu_{fit}(i)$ is the mean migration distance for polysomes, including all associated RNA, corresponding to class i, and a, b, c, and d are coefficients that are obtained from fitting this model to data. The equation from the model solution, referred to as the "migration distance function", was used to predict the mean migration distance for a population of polysomes from a given ribosome class i (see Figure 3.4 for the model fit):

$$\mu_{fit}(i) = \frac{404.49 + 3040.38 \cdot i^{0.912}}{6.56 + i^{0.912}}. \qquad (3.18)$$

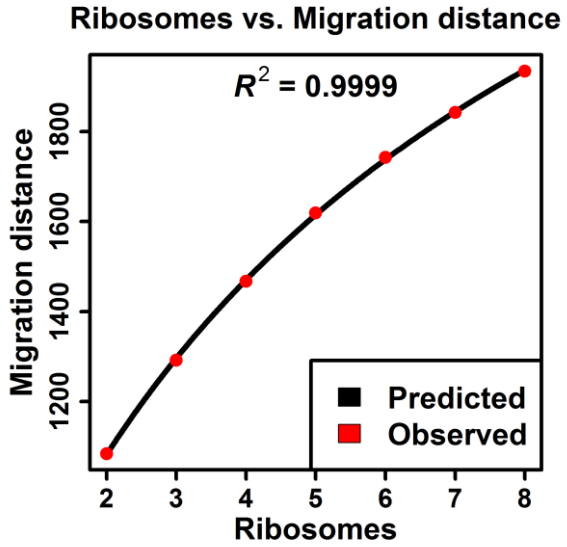

Figure 3.4: Model fit of the migration distance function.
The migration distances predicted by the model solution in equation 3.18 are plotted against the corresponding empirical data, with the function in black and the empirical data in red.

To estimate the migration distance standard deviation (σMD) for a given peak, it was assumed that each peak was normally distributed. A quadratic function was fit to x and y coordinates for each peak, and the second derivative was determined. The second derivative was set equal to the equation for the second derivative of the normal distribution function, where the only unknown value was σMD, and σMD was determined algebraically. For simplicity, σMD for each class was averaged to obtain a universal σMD. Since the sucrose gradients were split evenly into twelve fractions, each fraction corresponds to a lower and upper bound of migration distance. Assuming normality, the probability of an mRNA from a given ribosome class i migrating to a given fraction j can be computed by applying the cumulative distribution function (CDF) of the normal distribution:

$$\Pr\left[a_j \leq X_i \leq b_j\right] = \Phi\left(b_j\right) - \Phi(a_j) \qquad (3.19)$$

where

$$X_i \sim N(\mu_{fit}(i), \sigma MD). \quad (3.20)$$

Here, X_i is the theoretical distribution of migration distances for mRNA from a given ribosome class i, a_j is the lower bound of fraction j, b is the upper bound of fraction j, $\mu_{fit}(i)$ is the predicted mean migration distance for mRNA in class i, σMD is the standard deviation of the migration distance, and Φ is the CDF for the normal distribution. In R, $\Phi(x)$ is computed using the built-in "pnorm" function, where the argument "x" is the appropriate fraction bound, "mean" is the mean migration distance for a given ribosome class, and "sd" is σMD. This procedure is applied for each combination of ribosome class and polysome fraction, yielding a matrix of fraction location probabilities, referred to as the "FLP matrix", where the rows correspond to ribosome classes and the columns correspond to polysome fractions. The dot product of the FLP matrix and ribosome class abundances is computed, yielding predicted fraction abundances.

Computing correction factors

The predicted fraction abundances ("model predictions") are the theoretical abundances of mRNA in each fraction for a gene with specific parameter values. Our goal is to find parameter values for each gene for which the model predictions match the empirical data. A key challenge, however, is the fact that the empirical fraction abundances were subjected to two adjustments from the experimental procedure, creating a discrepancy between the model predictions and the empirical data. The first adjustment was that the RNA concentrations were adjusted to correct for variation in RNA yield across fractions by known "RNA dilution factors". The concentrations were measured by UV absorbance, which measures total RNA, including mRNA and ribosomal RNA (rRNA). After the RNA adjustment, the total RNA concentrations were equivalent across samples, but because each fraction has a different size distribution of

polysomes, the (mRNA)/(total RNA) ratios varied across fractions. From these RNA samples, cDNA was synthesized by reverse transcription. Because of the variation in the (mRNA)/(total RNA) ratio, the cDNA yield is expected to vary across fractions because the fractions have different amounts of mRNA, and mRNA, not rRNA, gives rise to cDNA. The cDNA concentrations were adjusted to the same concentrations by unknown "cDNA dilution factors" before being applied to the microarrays. The cDNA dilution factors are scaling parameters that are also expected to create a discrepancy between the model predictions and the empirical data. The cDNA dilution factors are unknown because this step and the subsequent microarray hybridization were performed according to a standardized protocol that did not allow a record to be kept. The RNA and cDNA dilution factors effectively adjust the amount of mRNA in each fraction twice before the mRNA abundances are measured.

The following is a more formal description of how the RNA and cDNA dilution factors affect the expected microarray signals.

The total amount of UV absorbance in a given fraction (T_f) is proportional to the total amount of RNA in the fraction, which is the sum of the amounts of mRNA (R_m) and rRNA (R_r),

$$T_f = R_m + R_r. \qquad (3.20)$$

R_m is given by

$$R_m = \sum_g \sum_i n_g \cdot m_{ig} \cdot P_f(i), \qquad (3.21)$$

where g is a given gene, i is a given ribosome class (i.e., number of ribosomes per polysome), n_g is the number of nucleotides per mRNA from gene g, m_{ig} is the number of mRNA molecules in

ribosome class i from gene g, and $P_f(i)$ is the probability of mRNA from ribosome class i migrating to the fraction. Similarly, R_r is given by

$$R_r = \sum_g \sum_i i \cdot n_r \cdot m_{ig} \cdot P_f(i), \qquad (3.22)$$

where n_r is the number of nucleotides of rRNA per ribosome. During the sample preparation, after the RNA from the fraction was diluted by the known RNA dilution factor a_f, the resulting RNA yield (R_t) would be given by

$$R_t = T_f \cdot a_f. \qquad (3.23)$$

cDNA was synthesized by reverse transcription, and the total amount of cDNA (C_f) in the fraction is proportional to the total amount of mRNA as represented above, multiplied by a_f and a constant K_1 which represents the PCR amplification efficiency,

$$C_f = \left(\sum_g \sum_i n_g \cdot m_{ig} \cdot P_f(i) \right) \cdot a_f \cdot K_1. \qquad (3.24)$$

The cDNA sample was then diluted by an unknown factor b_f, and resulting cDNA abundance (C_f^x) would be given by

$$C_f^x = C_f \cdot b_f, \qquad (3.25)$$

which can also be expressed as

$$C_f^x = R_m \cdot a_f \cdot K_1 \cdot b_f. \qquad (3.26)$$

The expected microarray signal for gene g from the fraction is equal to the amount of cDNA corresponding to the gene and fraction (C_{gf}^{x}), multiplied by a constant K_2 for that gene which represents the hybridization efficiency of the cDNA to the microarray chip. This expected microarray signal (x_{gf}) is given by

$$x_{gf} = K_2 \cdot C_{gf}^{x}. \quad (3.27)$$

R_{mg} is the abundance of mRNA in the fraction corresponding to gene g,

$$R_{mg} = \sum_{i} n_g \cdot m_{ig} \cdot P_f(i). \quad (3.28)$$

Therefore, the expected microarray signal for gene g from the fraction is given by

$$x_{gf} = R_{mg} \cdot a_f \cdot b_f \cdot K_1 \cdot K_2. \quad (3.29)$$

R_{mg} is predicted based on the model output and the fraction location probabilities, a_F is the known RNA dilution factor, and K_1 and K_2 can be assumed to be constant. Therefore, the only unknown quantity remaining is b_F, the unknown cDNA dilution factor. The greater the mRNA abundance in a given polysome fraction, the more cDNA would be synthesized by reverse transcription and the more the resulting cDNA sample would need to be diluted in order to be the same as the others. Therefore, b_F is expected to be proportional to the abundance of mRNA across all genes. The mRNA abundance in a given ribosome class across all genes can be determined from the total RNA abundance, since the (mRNA)/(total RNA) ratio is given by the ribosome class number (i.e., number of ribosomes per polysome).

To match the model predictions to the empirical data, we need to adjust the model predictions in the same way as the polysome fraction samples were adjusted. Correcting for the

RNA dilution factor is simple—we just divide the model predictions by the known RNA dilution factors. The second correction is more complicated, as it requires estimating b_f, or equivalently, the (mRNA)/(total RNA) ratio in each polysome fraction. Based on reviewing the computational and experimental procedures and fitting the model to the empirical data, it appears that this step is a main obstacle in being able to fit the model to the data, as I will explain subsequently. Nevertheless, I will describe the approach that has been used and discuss in general terms one alternative approach that may improve upon the current procedure.

Calculating unknown cDNA dilution factors in polysome fractions

Here, I describe the method we have used for estimating the unknown cDNA dilution factors (b_F) in each polysome fraction (Method 1) and discuss in general terms one alternative method (Method 2) that may improve upon the current procedure.

Method 1

1) The rRNA/mRNA ratio in each ribosome class (not fraction) i is estimated as

$$\left(\frac{rRNA}{mRNA}\right)_i = \frac{n_r \cdot i}{n_g} \qquad (3.30)$$

where n_r is the total rRNA length in nucleotides in a ribosome (~5400, from Ju Guan) and n_g is the mean mRNA length in nucleotides (1278 in Arabidopsis) (Wortman, Haas et al. 2003). This is because in class i each polysome consists of i ribosomes and 1 mRNA.

2) The contribution of rRNA from class i to fraction j, relative to mRNA, is calculated as

$$Contribution_{ij} = i \cdot FLP_{ij} \cdot \left(\frac{rRNA}{mRNA}\right)_i, \qquad (3.31)$$

where FLP_{ij} is the probability of an mRNA or polysome in class i migrating to fraction j, based on probabilities described above (Equations 3.18 and 3.19), and $\left(\frac{rRNA}{mRNA}\right)_i$ is the rRNA/mRNA ratio in polysomes in class i estimated in step 1 (Equation 3.30).

3) For a given polysome fraction, the contributions of rRNA (Equation 3.31) from each class are summed and b_f is estimated as the inverse of this sum, which considers the mRNA abundance in the fraction and resulting cDNA yield to be proportional to this estimate of the mRNA/rRNA ratio. This approach also makes the assumption that the total amount of RNA in each ribosome class is the same, which is almost certainly not the case.

Method 2

The following is a very general description of one alternative method for estimating b_F that is currently being developed. The distributions of migration distances for RNA in higher ribosome classes appear to be highly overlapping. Because of this overlap, RNA from multiple ribosome classes can contribute to the total UV absorbance at a given migration distance. Therefore, the areas of the peaks from ribosome class 9 and above are less than what is predicted by their corresponding peak height and σMD. To estimate the relative abundances of RNA corresponding to each ribosome class more accurately, an approach should be used that finds $\mu_{fit}(i)$ (Equation 3.18) and σMD for each class that are consistent with the total UV absorbance profile. In other words, at a given migration distance, the total UV absorbance should equal the sum of the absorbance resulting from overlapping peaks at that position.

1) Coordinates are selected along the UV absorbance profile, and the predicted mean migration distance for each class i is assumed to be $\mu_{fit}(i)$ (Equation 3.18).

2) Normal distribution functions are fit to the coordinates, using optimization to find the σMD that yields a UV absorbance profile that best matches the empirical data, based on minimizing an objective function. The "mixtools" R package (Benaglia, Chauveau et al. 2009) provides methods for implementing this procedure.
3) The total RNA abundance corresponding to each ribosome class across all genes would be estimated by determining the area under each curve. The percent of total RNA in a given ribosome class i that is mRNA (P_{mi}) can then be estimated as $P_{mi} = \frac{n_g}{n_g + n_r \cdot i}$. The relative abundance of mRNA from each ribosome class can then be estimated as the product of P_{mi} and the peak area corresponding to each class.
4) The portion of a given peak in each polysome fraction can be determined using integration to find the area under the curve that lies within the fraction bounds, and these percentages can be considered as probabilities that a polysome or mRNA from ribosome class i migrates to the fraction. From these probabilities, a fraction location probability matrix can be created and combined with the estimated mRNA abundances in each ribosome class to estimate the abundance of mRNA in each polysome fraction, which would be considered an estimate of the unknown cDNA factors (b_f).

Reversing the correction procedure

Because Method 1 yielded model predictions that consistently did not match the empirical data (see results section), I wondered whether any model solutions existed that would match the empirical data using Method 1. Therefore, I developed a procedure for reversing the conversion procedure to see what model solutions would be needed in order for Method 1 to yield predictions that matched the empirical data. In the original conversion procedure, as described, the fraction abundances are computed as a dot product where the FLP matrix is

multiplied by the vector of ribosome class abundances element-wise and the values in each column of the resulting matrix are summed. This can be represented by the following system of equations:

$$\begin{aligned} a_{01}x_0 + a_{11}x_1 + \cdots + a_{i_{max}1}x_{i_{max}} &= b_1 \\ a_{02}x_0 + a_{12}x_1 + \cdots + a_{i_{max}2}x_{i_{max}} &= b_2 \\ &\vdots \\ a_{09}x_0 + a_{19}x_1 + \cdots + a_{i_{max}9}x_{i_{max}} &= b_9. \end{aligned} \qquad (3.32)$$

In Equation 3.32, a_{ij} is the probability in the FLP matrix of a polysome in class i migrating to fraction j, x_i is the predicted mRNA abundance in ribosome class i, and b_j is the predicted mRNA abundance in fraction j. The goal of reversing the procedure is to determine the vector of ribosome class values x_0 through $x_{i_{max}}$. Equation 3.32 can be expressed in a matrix algebra form as

$$A \cdot X = B. \qquad (3.33)$$

In Equation 3.33, A is now a transpose of the FLP matrix (i.e., its rows and columns have been swapped), containing the values of a in Equation 3.32, which are treated as coefficients. As before, X is the vector of unknown ribosome class abundances, and B is the vector of known fraction abundances. Singular value decomposition (SVD) is a technique for solving such a system of equations to determine X. I used SVD to determine the model solutions in reverse for several real genes, and found values that would be impossible for the model to produce, including negative values. This provided clear evidence that Method 1 is not appropriate. This procedure should be useful in the future for testing whether a given procedure can accurately convert the model output into predicted microarray data.

Model implementation

Dr. Gilchrist and Nate Pollesch first implemented the model in Wolfram Mathematica. Joe Carboni had worked toward an implementation in R, but it needed substantial work before it could be fully usable. I sought to improve upon these previous efforts in three main aspects. First, I sought to implement a fully usable version in R. R is very widely used across scientific disciplines, including the computational and life sciences, so we chose to continue working on the model in R in order to support continued collaboration. Second, we wanted to utilize the parallel processing functionality in R provided by various packages, for example, "doParallel" and "doMPI". Mathematica has site licenses that limit the number of CPUs that can be used by a program in parallel, while R, being an open-source programming language and computing platform, has no such limitations. Third, we wanted to create an efficient, high-throughput way of running parameter searches. I created command-line scripts written in R that can be submitted as batch jobs on the Newton high-performance computing (HPC) system at the University of Tennessee. It uses the doMPI package to divide the parameter searches among a chosen number of CPUs, using message passing interface (MPI) to communicate among CPUs on different compute nodes.

Procedures

I created a number of procedures for processing data, exploring the model parameter space, fitting the model to data, visualizing results, and other tasks. I describe some of the most important ones here, roughly in the order in which they would typically be used. These are summarized in Table 3.4.

Table 3.4: Procedures for implementing the model

Procedure	Description
ProcessRawData	Process and normalize microarray data
SelectAOV	Find genes with large range in mRNA abundance and low noise
CalcFLPMatrix	Compute FLP matrix
CalcCorrectionFactors	Compute correction factors
CalcUnmarkedClass	Compute unmarked class abundances from parameter values
CalcMarkedClass	Compute marked class abundances
ModelSolveI	Compute total class abundances
AddExp, SubtractExp, CrossProdLogVectors, CrossProdLogMatrix, SumExp, SubtractLogVectors	Perform various mathematical operations on log-transformed data
CreateSignal	Compute predicted fraction abundances
NLLikFunction	Objective function; compute NLL from data and parameter values
MLELocalMultiSearch	Perform MLE parameter search
CreateBootstrapData	Generate bootstrapped data by simulation
MLEBootstrapSearch	Parameter search on bootstrapped data
PlotModelFit	Create plot with model predictions and data
CreateProfileImage	Draw UV absorbance profile based on parameter values
PlotModelSolveI	Interactive visualization tool

Processing microarray data

I created an R function called "ProcessRawData" which processes the raw microarray data from CEL files from a specified directory using the "affy" R package (Gautier, Cope et al. 2004) from Bioconductor (Gentleman, Carey et al. 2004). It normalizes the data using the "rma" method by default, but normalization can be omitted if desired. It uses the mas5calls function to classify each microarray signal as "present" ("P"), "marginal" ("M"), or "absent" ("A"). It keeps genes in the data set if all three total mRNA samples as well as all three replicates for at least one polysome fraction had "P" calls. It removes control probe sets as well as probe sets corresponding to chloroplast and mitochondrial genes. It returns the normalized and processed data and optionally exports it to a file.

Selecting a subset of genes for testing

I created an R function called "SelectAOV" which is used to select genes with desirable properties for testing the model. A good test set should have low noise in microarray signals among replicate measurements from the same polysome fraction and high variance across fractions, as this would represent genes that were measured with good precision that also had robust, identifiable polysome profiles. A good test set should also include genes with a range of profiles, having peak mRNA abundances in different fractions, as this would test the ability of the model and model-fitting process to distinguish among different patterns of ribosome loading. The function performs analysis of variance (ANOVA) on each gene, treating sets of replicates from each polysome fraction as groups, yielding a p-value for each gene. Lower p-values have higher ratios of between-fraction/within-fraction variance. Each gene is classified based on the polysome fraction in which its mRNA abundance is highest, yielding nine classes of genes.

Within each class, the genes are sorted based on their ANOVA p-values. The function returns microarray data for a specified number of genes in each class with the lowest p-values.

Computing fraction location probabilities

I created a function called "CalcFLPMatrix" which computes the "fraction location probabilities", stored in a matrix (the FLP matrix), as described above. The function calculates the FLP matrix based on a given i_{max}, migration distance standard deviation (σMD), and migration distance function, which by default is Equation 3.18.

Calculating correction factors

I created a function called "CalcCorrectionFactors" which calculates so-called "correction factors" used to adjust the predicted fraction abundances as discussed above. The function accepts a number of input arguments. Inputs include i_{max}, σMD, and the choice of adjustment methods. These methods include Method 1 described above and two modifications of Method 1 (not described). Additional arguments include mean mRNA length (default = 1278 nucleotides) and total rRNA length (default = 5400 nucleotides). Lastly, the function gives a choice of whether or not to divide the predicted fraction abundances by the dilution factors, with the default being to do so.

Solving the model

I created an R function called "ModelSolveI" which computes the model solutions, which are ribosome class mRNA abundances, based on i_{max} and the five translation parameters. It first uses another function called "CalcUnmarkedClass" to compute the unmarked class abundances. Second, it passes the unmarked class abundances into the function "CalcMarkedClass", which computes the marked class abundances. The unmarked and marked class abundances are added together to yield the model predictions.

Avoiding excessively large values by using the log scale

To compute the unmarked class abundances, CalcUnmarkedClass uses the recurrence relation in equation 3.10, which can generate very large numbers that R codes as "Inf" (infinity), which is undefined. To avoid these large numbers, I modified the calculations so that they use $\log_2$-transformed values. I wrote several helper functions that carry out the appropriate mathematical operations on the log scale. For instance, the "AddExp" function takes two log-transformed values and returns a value equal to that obtained by adding them together on the original scale and then log-transforming them. It uses the following identity:

$$\log_z(x+y) = \log_z x + \log_z\left(1+\frac{y}{x}\right). \qquad (3.34)$$

AddExp uses the right side of the equation below to combine two log-transformed values, here $\log_z x$ and $\log_z y$, so that the result is the same as if the values had been back-transformed with base z, added, and then log-transformed again:

$$\log_z(x+y) = \log_z x + \log_z(1+z^{\log_z y - \log_z x}). \qquad (3.35)$$

Other helper functions that I created use either the same identity or other related ones in order to add, subtract, multiply, divide, and perform matrix operations correctly (dot products, cross products) using values on the log scale, thus avoiding problems due to very large values in R.

Predicting fraction abundances from parameters

I created an R function called "CreateSignal" which computes model predictions, the polysome fraction abundances, from a set of parameter values and other inputs. Many other functions and procedures rely on this function. CreateSignal carries out three steps.

1) A model solution is computed using ModelSolveI.
2) The model solutions are converted into polysome fraction abundances by computing the dot product of the FLP matrix and the vector of model solutions.

3) The predicted fraction abundances are adjusted using correction factors as discussed above. A background signal is then added, which by default is 3.0, the approximate mean of all expression signals in the empirical data classified as "A" (absent) by the mas5calls function.

All of the optional procedures discussed thus far are available to CreateSignal. Replicate data can be generated, in which case values are sampled randomly from a normal distribution with a specified mean and standard deviation.

Generating data by simulation

I created an R function called "CreateBootstrapData" which creates a "bootstrapped" data set based on given parameters. It calls CreateSignal a specified number of times to generate one or more data sets with a specified number of replicates, adding noise by sampling from a normal distribution.

Calculating the negative log-likelihood

I created an R function called "NLLikFunction" which computes the negative log-likelihood (NLL). It uses CreateSignal to compute model predictions, then compares the model predictions to empirical or bootstrapped data using Equation 3.16. NLLikFunction can accept all input arguments that CreateSignal accepts. NLLikFunction is the objective function that is minimized in the optimization procedure.

Performing an MLE parameter search

The procedure that is most central to this project is estimating the model parameters. This is done using nonlinear optimization to search for parameter values that best fit the data based on the NLL. I created an R function called "MLELocalMultiSearch" which carries out

this task. MLELocalMultiSearch uses either the "optimx" function from the "optimx" R package (Nash and Varadhan 2011) or the "optim" function from the built-in "stats" package, depending on which optimization method is chosen. Simulated annealing (option "SANN") is only available in optim. See Table 3.2 for a survey of the available algorithms.

Summary of inputs that control the MLE parameter search

Here, I summarize the most important inputs to MLELocalMultiSearch.

1) First is a list of gene IDs or corresponding row numbers from a matrix containing the data on mRNA abundances per polysome fraction (can be simulated or empirical data).
2) Second are one or more sets of starting values for the five translation parameters plus the migration distance standard deviation (σMD). The five translation parameters, in addition to σMD, the background microarray signal (βNoise), and i_{max} can be optimized, but typically, the termination rate (τ), σMD, and i_{max} are kept constant. i_{max} can be specified in two ways. First, one or more i_{max} values can be provided as an input argument, in which case each value will be used for each gene in combination with all other starting values. i_{max} values specified in this way can either be kept constant or optimized as a parameter in each parameter search. The second way i_{max} can be specified is by providing as a separate input argument one or more values representing the number of nucleotides a ribosome occupies on an mRNA. The total coding sequence (CDS) length for the gene is then divided by this number and rounded up to an integer to yield a gene-specific i_{max}. i_{max} specified in this way is kept constant. Like the first way of specifying i_{max}, each value will be used for each gene and in combination with each set of starting values.

3) Third are the parameters that should be kept constant. MLELocalMultiSearch handles fixed parameters by defining an objective function that executes NLLikFunction with the fixed parameters passed as constant inputs to it, but with the parameters to be optimized passed as variables.
4) One useful additional input specifies whether to optimize the parameters on the log scale. If the parameters are optimized on their original scale, the function sets constraints so that parameter search avoids negative values. Doing this imposes limitations due to the fact that, as discussed, some search algorithms cannot accommodate constraints. However, optimizing the parameters on the log scale avoids this problem. If the log scale is used, then for example, instead of optimizing the value x, we optimize e^x.

Other input arguments can be provided, including those already described which are used by the functions called by MLELocalMultiSearch, with most of them being passed into CreateSignal.

Parallelizing the MLE parameter search

MLELocalMultiSearch divides the parameter searches among available processing cores. If run on a local computer (shared memory system) with multiple cores on the same compute node, the user must load the "doParallel" package (Analytics and Weston 2014) or a similar package which provides "%dopar%" and "foreach" methods. If run on a distributed memory system with cores distributed across multiple compute nodes, such as the Newton HPC system, the "doMPI" package (Weston 2013) is required, which relies on the "Rmpi" package to communicate among processes using the Open MPI library (Graham, Woodall et al.).

I implemented the MLE parameter search as a command-line script called "mleSearch.R", which runs on the Newton HPC system. A typical job file to run the script with multiple cores on Newton is shown below.

```
#$ -N MLE
#$ -q short*
#$ -pe openmpi* 12
#$ -l cores_per_node=12,proc_vendor=Intel
mpirun Rscript mleSearch.R MetaInfo.R >& output.txt
```

"-N MLE" designates "MLE" as the name of the job. "-q short*" requests a short queue on Newton which allows jobs that complete in 2 hours. "-pe openmpi* 12" instructs the queue system to reserve 12 compute nodes. "-l cores_per_node=12,proc_vendor=Intel" instructs the queue system to reserve only compute nodes with 12 cores per node (12 nodes x 12 cores/node = 144 cores) and only those that use Intel processors. The latter instruction is necessary because the "Rmpi" library (Yu 2002), which doMPI needs, must be compiled and run on the same type of processor. Since Newton contains some compute nodes with Intel processors and some with AMD, I compiled Rmpi using both processors so that the code can be run on compute clusters with both kinds of processors. "mpirun" instructs the Newton system to run the command with the requested number of processors. "Rscript" is the R interpreter that executes stand-alone R scripts. "mleSearch.R" is the main R program that runs the parameter search. "MetaInfo.R" is the name of a file containing inputs to the program assigned to variables. This is where the user specifies the data file, genes, algorithm, starting values, output file name, and other inputs. The inputs should be one per line, and set to variables that can be interpreted by R. For instance, the input specifying which genes to use could be "geneIDs = 1:20" or "geneIDs =

c(“10000_at”,”10001_at”)”. “>& output.txt” specifies that all output should be written to the designated file.

Parametric bootstrapping

I created an R function called “MLEBootstrapSearch” which performs parametric bootstrapping. Bootstrapping is a general approach for assessing the precision of parameter estimates by simulating the process of repeating an experiment and calculating measures of precision from the repeated “experiments” (DiCiccio and Efron 1996). There are four main steps in parametric bootstrapping.

1) Estimate the parameters.
2) Generate many random data sets from a distribution based on the parameter estimates.
3) Estimate the parameters from the random data sets.
4) Calculate measures of precision, such as the standard error (SE) and confidence intervals (CIs), from the repeated parameter estimates.

Aside from assessing the precision of parameter estimates, parametric bootstrapping enables us to explore how the choice of starting values, noisy data, and various model assumptions affect our ability to identify the true parameter values.

MLEBootstrapSearch reads a file created by CreateBootstrapData which typically contains many data sets stochastically generated from the same model parameter values with noise added by sampling from a normal distribution with a mean of 0 and a chosen standard deviation. In this context, I use the term “data set” to refer to one specific sampling of values from a distribution of data based on a set of parameters, with one data set having the same format as a gene from the empirical data. “Data series” is a collection of data sets generated under the

same conditions, that is, using the same parameter values and model assumptions, but with noise added. The user supplies one or more sets of values which are used as starting values along with the true parameters used to generate the simulated data. MLELocalMultiSearch is used to perform parameter searches for all bootstrapped data sets and all sets of starting values. The results are exported into a CSV file in a similar format as those from MLELocalMultiSearch, and a number of plots are created in a PDF file. The function can be used to run the full procedure described, which can take a long time for many bootstrapped data sets, or to simply read in a previous analysis and change the appearance of the plots. I created an R script called "bsSearch.R" which carries out this analysis on Newton in a similar manner as "mleSearch.R".

Visualizing model fits

I created an R function called "PlotModelFit" which draws a plot showing the data used to fit the model as well as model predictions based on a set of parameter values. It is used for visualizing how well a set of parameter values fit a data set. It accepts all input arguments that can control ModelSolveI and CreateSignal, including the five translation parameters, i_{max}, σMD, βNoise, and others. Examples can be seen in the results in Figure 3.9.

Visualizing UV absorbance profiles predicted by model parameters

I created an R function called "CreateProfileImage" which draws a plot showing the UV absorbance profile that would be expected for one mRNA species based on a set of parameter values. The UV absorbance profile would be expected to change based on the ribosome loading of mRNA, since higher ribosome loading causes mRNA to migrate farther through a sucrose density gradient. An example is shown in Figure 3.5. This plotting routine carries out four steps.

1) ModelSolveI is used to predict the ribosome class mRNA abundances from the parameter values.
2) For each ribosome class i, the mean migration distance $\mu_{fit}(i)$ is calculated from Equation 3.18.
3) For each class i, for a series of positions from 0 to the maximum migration distance (for instance, 1000 positions), the probability densities are computed based on a normal distribution with a mean of $\mu_{fit}(i)$ and standard deviation of σMD. These probabilities correspond to the percent of mRNA from class i that migrate to each position across the gradient.
4) The expected UV absorbance profile for a single class i is obtained by multiplying these probabilities by the predicted mRNA abundance for class i.
5) Each class-specific profile and the total profile, which is the sum of the class-specific profiles, are plotted.

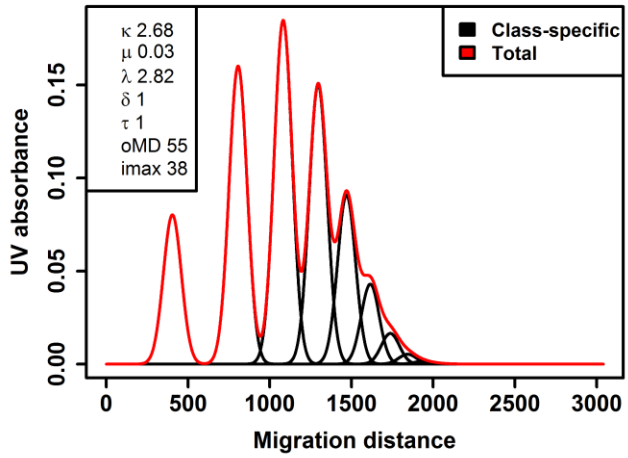

Figure 3.5: UV absorbance profile.
The x-axis is the sucrose gradient migration distance, and the y-axis is the relative UV absorbance. For the parameter values shown in the box, the UV profile corresponding to each ribosome class (here, class 0 to 38) are plotted individually ("Class-specific", in black). The total profile, which is the sum across classes, is also plotted ("Total", in red). Note that the UV profile is for mRNA only; it does not contain the additional contribution from the ribosomal RNAs.

Interactive visualization tool

In order to guide our intuition for how the parameters and different procedures affect the model predictions, I implemented the model as an interactive application called "PlotModelSolveI". PlotModelSolveI uses the "shiny" R package (Chang, Cheng et al. 2015), which provides methods for taking input through interactive controls such as slider bars, buttons, check boxes, and text input and displaying the output. A screenshot is shown in Figure 3.6. PlotModelSolveI accepts nearly all inputs that have been discussed, including the translation parameters, a constant i_{max} used in calculating the correction factors, and a gene-specific i_{max} used in computing the model solutions and converting them into polysome fraction abundances, correction factor method, and whether or not to include the dilution factors (first RNA

adjustment discussed above) in computing the correction factors. PlotModelSolveI displays three plots. The first is a plot of the predicted unmarked, marked, and total mRNA abundances in each ribosome class, the second is a plot of the predicted mRNA abundances in each polysome fraction, and the third is a plot of the predicted UV absorbance profile.

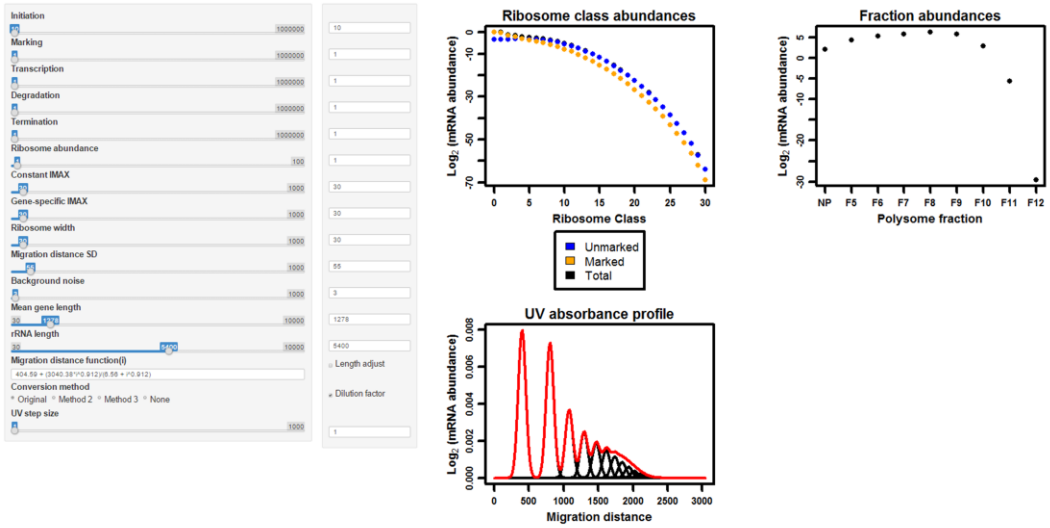

Figure 3.6: PlotModelSolveI interactive visualization tool

Optimization of computational efficiency

I profiled the R code to identify bottlenecks in the processing time for performing the parameter search. NLLikFunction is the objective function that is minimized in the parameter search and is therefore called many times, so I aimed to optimize this function. The main functions called by NLLikFunction are depicted in Table 3.5. I used the "Rprof" function from the built-in "utils" package in R to determine what percent of the overall processing time of NLLikFunction were spent on each function. Rprof works by checking at specific time intervals (by default, every 20 milliseconds) which function is currently being executed. I executed

NLLikFunction 1000 times and used Rprof to calculate the time and percent of the total time spent on each function called by it.

As shown in Table 3.5, the bulk of the processing time for NLLikFunction, 94.34%, was spent executing ModelSolveI. ModelSolveI spent most of its time executing CalcUnmarkedClass (82.66% of the total for NLLikFunction), and CalcUnmarkedClass spent most of its time executing SubDet (58.94% of the total for NLLikFunction). SubDet computes the matrix sub-determinants corresponding to each ribosome class, which are used to solve the model, using the recurrence relation in Equation 3.10, and it must be called for every ribosome class in the system. SubDet had already been sped up using memoization, that is, storing the result each time it is called with new inputs and returning the pre-computed results when it is called with the same inputs again, which avoids repeating the same calculation. Memoization helps speed up recurrence functions because they repeat the same calculation multiple times. Memoization was done using the "memoise" function from the "memoise" R package (Wickham, Hester et al. 2016).

I did not see a clear way to further optimize SubDet in the R code. However, code written in the C++ language can be pre-compiled into an R function using the Rcpp R package (Eddelbuettel, François et al. 2011), which can provide a speed-up. Therefore, I implemented SubDet (memoized) and several other R functions in C++. I used the "microbenchmark" function from the "microbenchmark" R package to compare the processing times of NLLikFunction written in R and C++. As shown in Table 3.5, the C++ version of SubDet was approximately 37 times faster than the R version. Ultimately, the optimized version of NLLikFunction was about three times faster than the original version. This typically resulted in approximately a 3-fold speed-up in the parameter search for any number of genes and sets of

starting values. A single parameter search using the optimized NLLikFunction on one CPU can take anywhere from a few seconds to over 30 seconds, depending on the starting values, the search algorithm, and i_{max}. Efforts will continue to further improve the efficiency of the code.

Results

A simulation study

The ultimate goal of this research is to estimate the translation parameters for many genes by fitting the model to the empirical data. Estimating the parameters poses a number of challenges, however. The goal of this simulation study was to better understand two main challenges in the parameter estimation process. First, because the parameter search is a local minimization of the objective function, there is no guarantee that a given set of parameter estimates are the best ones. I demonstrate that with poor starting values, some search algorithms can yield sub-optimal model fits, likely because they become trapped in local minima where the objective function is minimized in the local parameter space while better ones exist elsewhere. The "nlminb" algorithm (a Newton method), however, significantly outperforms the others in finding parameter estimates that fit the data best, regardless of the starting values.

Table 3.5: Improvement in the efficiency of the objective function used in the parameter search

Function	Time before Optimization (% total time)	Time after Optimization (% total time)	Benchmark test [median(lq - uq)]	
			Before	After
NLLikFunction	10.96 secs (100%)	4.82 secs (100%)	15.1 ms (14.6-15.9)	5.95 ms (5.79-6.33)
CreateSignal	10.88 secs (99.27%)	4.72 secs (97.93%)	16.4 ms (14.6-16.0)	5.9 ms (5.7-6.1)
ModelSolveI	10.34 secs (94.34%)	4.36 secs (90.46%)	14.27 ms (13.8-15.0)	5.2 ms (5.1-5.4)
CalcUnmarkedClass (eqs. 3.10-.12)	9.06 secs (82.66%)	3.44 secs (71.37%)	12.3 ms (12.1-13.0)	3.8 ms (3.6-4.1)
SubDet (eq. 3.10)	6.46 secs (58.94%)	0.16 secs (3.32%)	9.5 ms (9.2-10.0)	0.26 ms (0.24-0.28)
CalcMarkedClass (eq. 3.13)	1.28 secs (11.68%)	0.90 secs (18.67%)	1.53 ms (1.47-1.63)	1.35 ms (1.28-1.43)

The arrangement of the functions illustrates the order in which they are called by NLLikFunction. For example, ModelSolveI calls the function CalcUnmarkedClass, which calls SubDet. Once CalcUnmarkedClass completes, ModelSolveI calls CalcMarkedClass. NLLikFunction was executed 1000 times, so the timings are estimates of the amount of time spent on a particular function when NLLikFunction was run 1000 times. The percentages should not add up to 100% because each function is called within another function, and they are the percentages of the total time of NLLikFunction spent on that function. Microbenchmark was used to compare the processing times before and after optimizing the code. The functions were run 100 times and the distributions of processing times compared. The values in the "Benchmark test" column are the median processing time, with the lower and upper quartiles in parentheses. secs = seconds; ms = milliseconds. Functions that correspond to specific equations in this dissertation are listed with the equation number in parentheses.

The second challenge is noise in the data. With more noise in the data, multiple sets of parameter values can fit the same data equally well. When this is the case, good parameter estimates that are close to the true parameter values cannot be distinguished from bad ones based on how well they fit the data, making the parameters unidentifiable. Fortunately, certain ribosome loading patterns only result from parameter values that are in specific proportions to each other, in which case the relationships among parameters are constrained by the data. In this case, keeping additional model parameters constant limits the range of parameter estimates that fit the data, making the parameters more identifiable. More noise in the data can also make parameter estimates less reliable, reflected in wider confidence intervals (CIs) of parameter estimates. Used throughout this simulation study, parametric bootstrapping helps in assessing the reliability of parameter estimates, given a certain amount of noise in the data. Here, I first briefly describe the generation of data for this simulation study. I then describe these observations in more detail, all the while discussing their implications for the overall goals of this research.

Generation of data by simulation

Here, "data set" refers to one sampling of data from a distribution based on a set of parameter values. One data set has the same format as the microarray data for one gene, consisting of three replicate values from nine polysome fractions. "Data series" refers to a collection of 100 replicate data sets generated from the same parameter values, with noise sampled from a log-normal distribution with a specific standard deviation (SD). I used PlotModelSolveI to identify three sets of parameter values that yield low, medium, and high ribosome loading profiles (Table 3.6), and I refer to these sets of parameter values and the data

as “low RL”, “medium RL”, and “high RL”, respectively. The low RL data results from a high transcription rate combined with relatively low values for the other parameters. The medium RL data results from the initiation rate being moderately higher than the marking rate. The high RL data results from a high initiation rate combined with low values for the other parameters. I used CreateBootstrapData to generate twelve data series, including the three ribosome loading profiles and four levels of noise. Ribosome class abundances, UV absorbance profiles, and polysome fraction abundances predicted by the three sets of parameter values are shown in Figure 3.7. $\boldsymbol{\theta}$ (“theta”) refers to the true parameter values used to generate a data set or series, and can refer to one or any number of the five translation parameters.

Table 3.6: Parameter values used to generate data with a low, medium, or high ribosome loading profile

Parameter	Low ribosome loading	Medium ribosome loading	High ribosome loading
Transcription ($\boldsymbol{\lambda}$)	1000	10	1
Initiation ($\boldsymbol{\kappa}$)	1	10	1000
Marking ($\boldsymbol{\mu}$)	1	1	1
Degradation ($\boldsymbol{\delta}$)	1	10	1
Elongation ($\boldsymbol{\tau}$)	1	1	1

Comparison of algorithms for local parameter searches

I used MLEBootstrapSearch to run parameter searches for the three data profiles listed in Table 3.6 with no noise added to see how well different search algorithms could identify the true parameter values (θ). I used 11 of the 15 algorithms listed in Table 3.2. The exceptions were “CG” , “hjkb”, “Rcgmin”, and “Rvmmin” due to various technical issues. I used 8 sets of starting values chosen arbitrarily and ranging from 10,000-fold below to 10,000-fold above θ. I

compared the algorithms in terms of how many good model fits they returned out of 8, as well as the total time required for MLEBootstrapSearch to complete the 8 searches. These and all subsequent analyses were carried out on a local Windows 8 computer with 2 physical cores (4 logical cores), specifically 1.8GHz Intel Core i7 (4^{th} gen) processors. 4 searches were run simultaneously. I considered a model fit "good" if all 4 parameter estimates were within 1% of their respective θ. Table 3.7 gives the numbers of good model fits out of 8 and the run times.

Just under half of the model fits were nearly perfect, yielding parameter estimates nearly identical to θ along with low negative log-likelihoods (NLLs), so good parameter estimates were distinguishable from bad ones. The nlminb algorithm was the clear winner, yielding parameter estimates that were most often virtually identical to θ along with the lowest NLLs. For the low RL data, all 4 parameter estimates in all 8 model fits were within 1e-7 of θ and NLLs were approximately 1e-14. For the medium and high RL data, 7 of 8 searches yielded equally good parameter estimates and NLLs. Newuoa performed nearly as well in terms of model fits, but was 3-4 times slower. Other algorithms performed nearly as well for parameter searches begun near θ but poorly for searches begun farther away. Nlminb was used for all subsequent parameter searches.

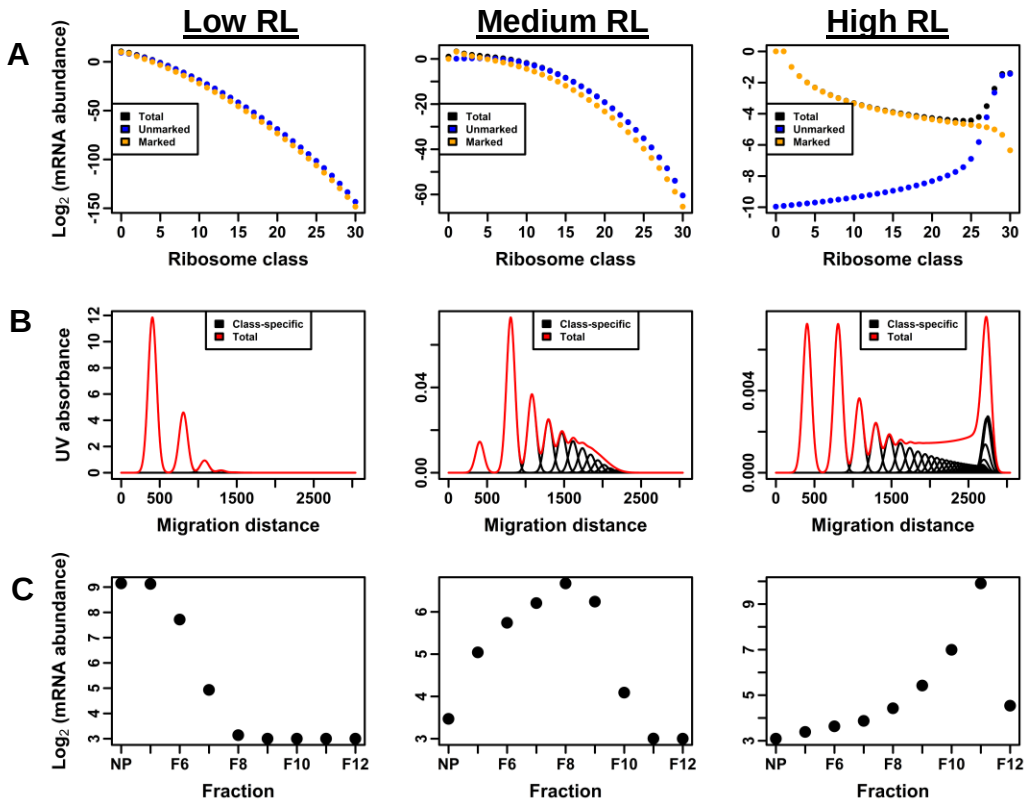

Figure 3.7: Ribosome loading profiles generated for the simulation study. The low ("Low RL"), medium ("Medium RL"), and high ("High RL") ribosome loading profiles used in this simulation study are visualized using various plots. **(A)** Ribosome class abundances, **(B)** UV absorbance profiles, and (**C**) polysome fraction abundances. See PlotClassAbundances and PlotModelSolveI procedures for details about (A) and CreateProfileImage for (B). Fraction abundances in (C) were produced from the PlotFractionAbundances function and include the RNA and cDNA yield corrections and background signal described in the text.

Table 3.7: Performance of 11 search algorithms

	Low ribosome loading		**Medium ribosome loading**		**High ribosome loading**	
Algorithm	No. within 1%	Time (s)	No. within 1%	Time (s)	No. within 1%	Time (s)
Nelder-Mead	1	13	1	21	1	16
BFGS	3	31	3	17	4	27
L-BFGS-B	5	18	4	31	6	29
nlm	3	16	4	12	3	13
nlminb	8	13	7	20	7	18
spg	4	338	0	1050	3	284
ucminf	3	9	6	21	4	141
newuoa	8	66	6	72	6	47
bobyqa	6	57	5	119	6	45
nmkb	2	11	2	21	2	17
SANN	0	211	0	486	0	465

Parameter searches were run for the low, medium, and high ribosome loading profiles generated from the parameter values in Table 3.6 with no noise. 11 search algorithms were used with 8 sets of starting values ranging from 10,000-fold below to 10,000-fold above the true parameter values (θ). The numbers of model fits out of 8, for which all 4 parameter estimates were within 1% of θ, are given in the column "No. within 1%". The times in seconds for MLEBootstrapSearch to complete the 8 searches, running 4 in parallel, are given in the column "Time (s)".

Noise makes the model parameters unidentifiable

As demonstrated, parameter searches using the nlminb algorithm consistently identified the true parameter values (θ) for the three data profiles with no noise, and good parameter estimates could be distinguished from bad ones because they fit the data much better. The bad model fits resulting from poor starting values were obvious, yielding abnormally large negative log-likelihoods (NLLs) and parameter estimates that were far from θ. In that sense, the parameters were highly identifiable. I demonstrate here that noise in the data does not necessarily cause higher occurrences of bad model fits (i.e., bad NLLs). However, noise can make parameters less identifiable because bad parameter estimates can fit the data as well as good ones.

As mentioned above, I generated data from the same three sets of parameter values with varying amounts of noise. I used CreateBootstrapData to generate 12 data series consisting of 100 data sets each. Noise was sampled from a log-normal distribution with a standard deviation (SD) of 0.01, 0.1, 0.33, and 0.66, with 0.33 being the average over the polysome fractions across all genes in the empirical data. I ran the parameter searches for these data using 9 sets of starting values, including the same as before (ranging from 10,000-fold below to 10,000-fold above θ, the true parameter values for each data profile) in addition to θ. 900 parameter searches were run on each data series.

In this simulation study, the true parameter values (θ) that produced the data were known. With knowledge of θ, the parameter estimates could be compared in terms of how close they were to θ. Typically, for a given data set, most of the model fits resulting from different starting values had low NLLs that were equal to each other to many significant digits, along with parameter estimates that were very similar, though to fewer significant digits. This majority set of model fits could be considered to have "good" parameter estimates and NLLs. Other model fits, however, had parameter estimates that were farther from θ, so these estimates could be considered "bad". In some cases, the bad parameter estimates yielded NLLs that were higher than and thus distinguishable from the group of nearly identical, low NLLs. For that reason, in these cases, the parameters would be considered identifiable. In other cases, the bad parameter estimates yielded NLLs that were indistinguishable from those of the good estimates. In those cases, the parameters would be considered less identifiable. This would be a problem for the empirical data, for which we do not know θ *a priori*, because we must rely on the NLLs to decide which parameter estimates are best.

Because noise was added to the data, the set of very similar, good parameter estimates fit the data better than exactly θ. Therefore, the medians of the parameter estimates for a given data set, rather than exactly θ, reflected the best-fitting parameter values for the data set. To distinguish the majority set of very similar, good parameter estimates from the bad ones, I considered estimates that were within 1% of their median as good, while the others were considered bad. To distinguish the set of nearly identical, lower NLLs from the larger ones, I used a stricter criterion, where an NLL was considered bad if it was larger than the minimum NLL after rounding to six significant digits. This criterion was chosen because in this study it distinguished the majority set of nearly identical, low NLLs for a given data set from the minority set of NLLs that were quite similar to them, but nevertheless larger.

For each data series, I determined how many model fits had a good NLL out of those with at least one bad parameter estimate, and these results are given as fractions in Table 3.8. More model fits with bad parameter estimates (denominator) indicate that the parameter estimates were relatively more sensitive to the choice of starting values, and a higher number of those with good NLLs indicates that the parameters were less identifiable, since bad parameter estimates fit the data as well as good ones. The degradation rate (δ) estimates exhibited more variability than the other parameter estimates without significantly affecting the model fits, for reasons that are discussed below. δ was thus excluded from the criteria for determining whether a model fit had good or bad parameter estimates.

Table 3.8: Decreasing parameter identifiability with increasing noise

Data	SD = 0.01	SD = 0.1	SD = 0.33	SD = 0.66
Low RL	0/3	13/66	321/417	472/596
Medium RL	0/149	0/144	0/137	0/138
High RL	0/98	0/101	60/192	139/301

Parameter searches were carried out for 12 data series with the indicated ribosome loading (RL) profiles and levels of noise, based on the standard deviations (SDs) of the distributions from which the noise was sampled. There were 900 model fits for each data series, resulting from parameter searches using 9 sets of starting values for 100 data sets. The table gives the numbers of model fits that had good negative log-likelihoods (numerators) out of those that had bad parameter estimates (denominators). "Good" and "bad" are defined in the text.

For the low and high RL data, more noise was associated with lower parameter identifiability. The low RL data with SD = 0.01 only yielded 3 model fits where one or more parameter estimates were considered bad, based on being more than 1% different from the median of 9 estimates for the same data set (Table 3.8). None of the bad parameter estimates yielded a good NLL. As the noise in the low RL data increased, more model fits yielded bad parameter estimates, and a higher percentage of them yielded good NLLs. The high RL data also yielded more model fits with bad parameter estimates as the noise in the data was increased, and while none of the model fits with bad parameter estimates yielded good NLLs for the data with SD = 0.01 or 0.1, these numbers did increase for the data with SD = 0.33 and 0.66. The medium RL data yielded a similar number of model fits with bad parameter estimates for data with all levels of noise, but out of these, none yielded good NLLs. In summary, more noise was associated with decreased parameter identifiability, being defined here as for the low and high RL data, but not the medium RL data.

Noise makes parameters estimates less reliable

More noise in the data can also decrease the reliability of parameter estimates in that they exhibit more variation when the data are re-sampled, or bootstrapped. Figure 3.8 shows the distributions of negative log-likelihoods (NLLs) and parameter estimates resulting from searches

begun at the true parameter values (θ) for the low, medium, and high ribosome loading (RL) data with different amounts of noise. Consistent with the previous observations, increasing noise most dramatically affected the parameter estimates for the low RL data, with more noise associated with wider confidence intervals (CIs) for all of the parameter estimates, especially data with SD = 0.33 and 0.66. More noise also yielded wider CIs for the high RL data, especially the initiation (κ) and degradation (δ) rate estimates for data with SD = 0.33 and 0.66. As before, parameter estimates for the medium RL data were the least impacted by increasing levels of noise.

One might reasonably speculate that parameter estimates that are closer to θ fit the data better than those that are farther away, but that is not the case. Figure 3.9 shows an example where two very different sets of parameter estimates fit the same data nearly equally well. The example is one data set from the low RL data with SD = 0.33.

Figure 3.8: Distributions of negative log-likelihoods and parameter estimates for data with different amounts of noise.
Parameter searches were run for the low (**A**), medium (**B**), and high (**C**) ribosome loading (RL) data generated from the parameter values in Table 3.6, with different amounts of noise, as indicated by their standard deviations (SDs) on the right. The distributions of parameter estimates are on the natural log scale. The red lines correspond to the true parameter values (θ) used to generate the data, and the black dashed lines are at the bounds of the 95% confidence intervals. The 5 columns of plots correspond to the 5 labels at the top. NLL = negative log-likelihood.

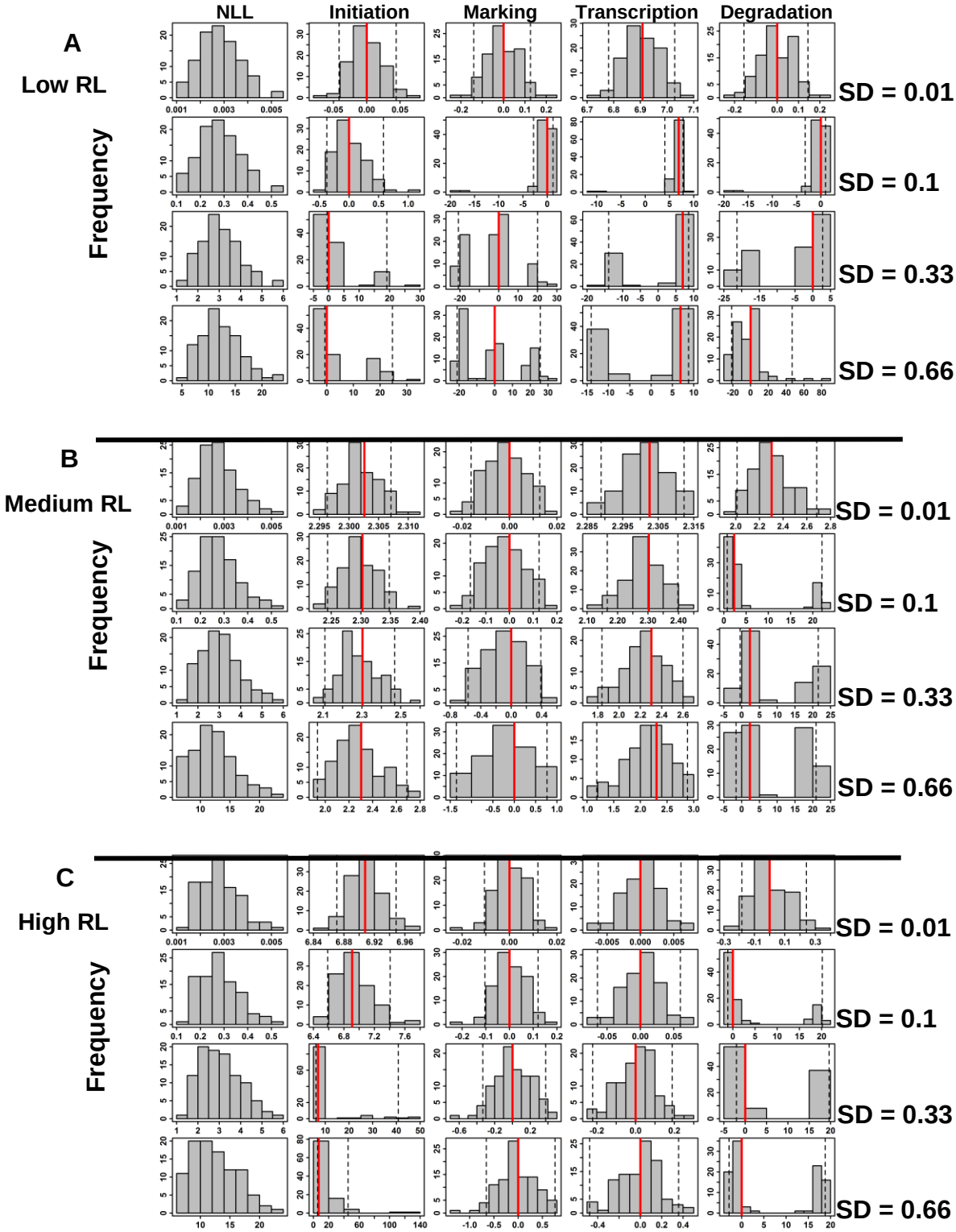

Figure 3.8 continued.

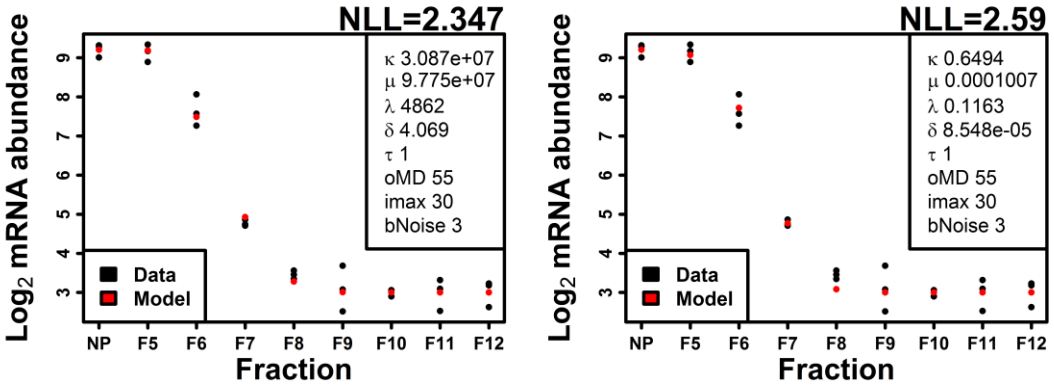

Figure 3.9: Example of different sets of parameter values that fit the same data well. The original data ("Data" in black) and model predictions ("Model" in red) are shown for one data set from the low ribosome loading data with SD = 0.33. The negative log-likelihoods (NLLs) are indicated above the plots. The parameter values are given in the right legend. See PlotModelFit procedure and Table 3.3 for details and symbols used.

Keeping additional parameters constant can improve identifiability

As I demonstrated above, noise in the data can make parameters less identifiable, where multiple sets of parameter values fit the same data equally well. I investigated how multiple sets of parameter values can fit the same data, and found that even when individual parameters are poorly constrained by the data, relationships among them can still be highly constrained. In other words, to yield certain ribosome loading profiles, the parameters can be adjusted in a variety of ways as long as they are kept in the same proportions to each other. In this section, I first describe the constraints in the relationships among the parameters in the data used in this simulation study. I then show how these constraints enable some parameters to be more precisely estimated, despite noise, by keeping others constant. I use the data with the highest level of noise (SD = 0.66) to illustrate these concepts, since those data suffered the most from a lack of parameter identifiability.

Constraints in relationships among model parameters

In the low ribosome loading (RL) data, the ribosome class mRNA abundances (Figure 3.7A) and resulting polysome fraction abundances (Figure 3.7C) exist in a certain equilibrium resulting from a high transcription rate (λ) and lower initiation (κ) and marking (μ) rates that are equal to each other. The same equilibrium results if transcription is slower but initiation is sufficiently faster than marking. The slower transcription rate would cause less mRNA to enter the system, but the higher initiation rate, relative to marking, would cause a higher percent of the mRNA to move up in ribosome class instead of entering the marked classes where mRNA would only move down in class. This relationship among the three parameters is evident in the nearly perfect negative correlation (R^2 = 0.9987) between $\log(\lambda)$ estimates and $\log(\kappa) - \log(\mu)$ across the 100 data sets, which is depicted in Figure 3.10A.

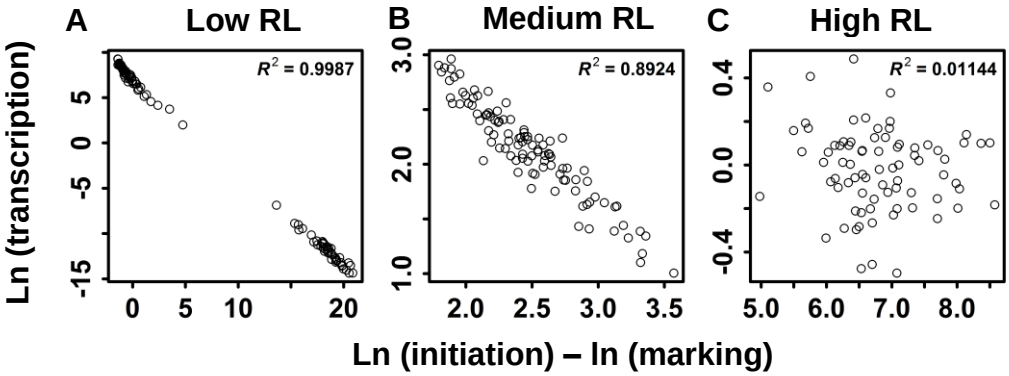

Figure 3.10: Relationships among estimated rates of transcription, initiation, and marking.
For the low (**A**), medium (**B**), and high (**C**) ribosome loading (RL) data, the differences between the ln (initiation) and ln (marking) rate estimates are plotted along the x-axis, and ln (transcription) is plotted along the y-axis. The R^2 values are given in the upper right corner of each plot.

As the initiation rate gets increasingly larger than the marking rate, the transcription rate decreases, yielding nearly the same model predictions as θ. The medium RL data show a similar relationship, though not quite as strong (R^2 = 0.89), and the range of parameter values is much

smaller than in the low RL data. This correlation is not present in the high RL data (R^2 = 0.01), however. The high RL data was based on a high initiation rate (1000) and lower rates for the other parameters. A wide range of initiation rate estimates fit these data well, due to a combination of factors. First, the majority of mRNA in each ribosome class from 21 to 30 is predicted to migrate into fraction 11, based on the migration distance function (Equation 3.18) and the fraction location probabilities (Equation 3.19). Second, with the high initiation rate, 96% of the unmarked mRNA is predicted to be in class 21 or higher (Figure 3.7A, "High RL"). Therefore, most of the unmarked mRNA is found in fraction 11 (Figure 3.7C, "High RL"). With nearly all of the unmarked mRNA in fraction 11, increasing the initiation rate has little effect on the distribution of mRNA among polysome fractions. It could be argued that a higher i_{max} may be more appropriate when the initiation rate is very high, as this would allow more mRNA to populate fraction 12, which cannot possibly have more mRNA than fraction 11 with an i_{max} = 30.

The distributions of degradation rate (δ) estimates were especially wide for all but the lowest noise level (Figure 3.8). In the low RL data, where mRNA was abundant in the NP fraction (Figure 3.7C, "Low RL"), changing degradation alone would affect marked class 0 (m_0^*) and therefore the NP fraction. The model system can compensate for changing the degradation rate by changing the transcription and marking rates proportionately in the same direction, which maintains the same amount of mRNA in m_0^*. Figure 3.11 shows plots of δ vs. λ and μ estimates for the 100 data sets in the three data series with SD = 0.66. The relationships among δ, λ, and μ appear constrained for 56 low RL data sets in the lower left region of Figure 3.11A, where δ estimates are highly positively correlated with λ (R^2 = 0.995) and μ (R^2 = 0.988). The other 44

low RL data sets had higher δ estimates which, for unknown reasons, were not correlated with the λ or μ estimates.

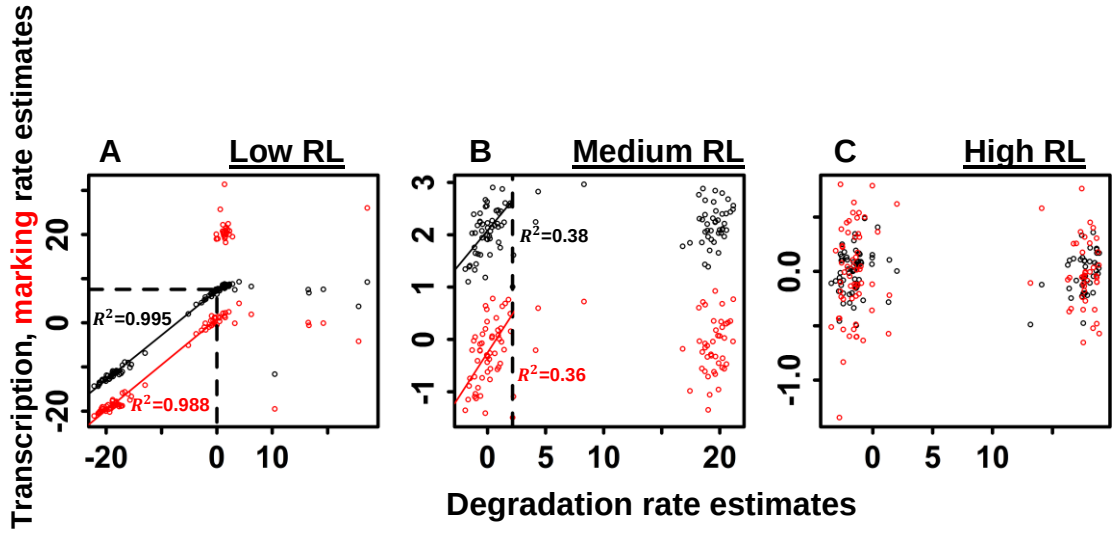

Figure 3.11: Relationships among estimated rates of transcription, marking, and degradation.
Estimates of the degradation rate (δ) are plotted along the x-axis against estimates of transcription (λ, black) and marking (μ, red) rates along the y-axis. Both axes are on the natural log scale. The parameter estimates are those resulting from parameter searches for the 100 data sets from the low, medium, and high ribosome loading (RL) data with SD = 0.66. In (**A**), the box defines the range of δ estimates that were highly correlated with λ and μ, and the regression lines and R^2 values correspond to the points in the box. In (**B**), the vertical dashed line is at ln(10) (the true δ value), and the R^2 values correspond to the data to the left of the line. In (**C**), the parameter estimates are poorly correlated so no regression lines are drawn.

In the medium RL data, the NP fraction has a low abundance of mRNA (Figure 3.7C, "Medium RL"). Increasing degradation alone has little effect on the system, so any value higher than θ (10) can fit the data well. Accordingly, there is no correlation between the degradation rate estimates above 10 and the other parameters (R^2 = 0.0013, 0.013, and 0.0022 for correlations between δ and λ, κ, and μ, respectively; Figure 3.11B shows plots of δ vs. λ and μ). Decreasing degradation alone, however, would increase accumulation of mRNA in m_0^*, but the system can compensate by decreasing the other three parameter values in concert. Therefore, the

degradation rate estimates below 10 are moderately positively correlated with the others (R^2 = 0.38, 0.30, and 0.36 for correlations with λ, κ, and μ, respectively). The reason why these correlations are weaker than those in the low RL data may be because there are more ways to compensate for changes in the degradation rate than in the low RL data, so changes in the degradation rate are moderately correlated with all three of the others rather than strongly with one or two. In the low RL data, for the 56 data sets that have low degradation rate estimates, the system depends almost entirely on modulating transcription and marking, and not initiation, for compensating for changes in degradation, while in the medium RL data, the system can modulate transcription, initiation, and marking in a variety of ways in order to compensate for changes in degradation and yield the same model predictions.

In the high RL data, the NP fraction also has very little mRNA (Figure 3.7B, "High RL"). Therefore, increasing the degradation rate alone has little effect on the model predictions. Decreasing the degradation rate alone 100-fold would increase accumulation of mRNA in m_0^*. However, the parameter searches did not return any degradation rate estimates for the high RL data that were this low, likely because there were no values for the other parameters that fit the data well when combined with such low degradation rate values.

In summary, noise in the data can make parameters less identifiable, but relationships among the parameters can still be highly constrained by some data.

Keeping some parameters constant can improve identifiability of others

As I have demonstrated, noise can make parameters less identifiable, where different starting values yield different parameter estimates that fit the same data equally well. Nevertheless, constraints in the relationships among parameters from certain data, described

above, make it possible to increase the identifiability of those that are estimated by keeping others constant.

I ran parameter searches on the low, medium, and high RL data with the highest level of noise (SD = 0.66), estimating two or three parameters, having already carried out searches where four were estimated. As before, I used 9 sets of starting values for each data series, ranging from 10,000-fold below to 10,000-fold above the true parameter values (θ), in addition to θ, for a total of 900 model fits for each data series. I used the same criteria as before for judging the NLLs and parameter estimates as "good" and "bad". I only considered κ and μ in judging whether the parameter estimates from a model fit were good or bad, since they were estimated in every search. With these two criteria, I determined how many model fits for each data series had a good NLL out of those with either a bad κ or μ estimate. As described above, more model fits with bad parameter estimates (denominator) indicate that the parameter estimates were relatively more sensitive to the choice of starting values, and a higher percentage of those with good NLLs indicates that the parameters were less identifiable, since bad parameter estimates fit the data as well as good ones. These proportions are given in Table 3.9 for parameter searches where two, three, and four parameters were estimated.

Table 3.9: Parameter identifiability with different numbers of parameters estimated

Data	$\boldsymbol{\kappa, \mu}$	$\boldsymbol{\kappa, \mu, \lambda}$	$\boldsymbol{\kappa, \mu, \delta}$	$\boldsymbol{\lambda, \kappa, \mu, \delta}$
Low RL	0/1	82/94	8/34	472/596
Medium RL	0/175	0/70	0/148	0/138
High RL	126/306	160/319	118/321	139/301

Parameter searches were carried out for low, medium, and high RL data series generated from the parameter values in Table 3.6, all with a standard deviation (SD) of 0.66. There were 900 model fits for each data series, resulting from parameter searches using 9 sets of starting values for 100 data sets. The table gives the numbers of model fits that had good negative log-likelihoods (numerators) out of those that had bad parameter estimates (denominators). "Good" and "bad" are defined in the main text. κ=initiation; μ=marking; λ=transcription; δ=degradation.

For the low RL data, when only κ and μ were estimated, only 1 out of 900 model fits had a bad parameter estimate, but its NLL was also bad, making κ and μ identifiable. When λ was added to the parameters being estimated, 94 model fits had bad parameter estimates, and 82 of those had good NLLs. Thus, good κ and μ estimates became less distinguishable from bad ones and therefore less identifiable. Estimating δ instead of λ improved identifiability, as 34 model fits had bad parameter estimates, with only 8 of them yielding good NLLs. Estimating four parameters yielded the most model fits with bad parameter estimates, with the largest percentage of them having good NLLs.

Because the relationships between λ and κ and μ were highly constrained, keeping λ constant limited the range of values that κ and μ could take to fit the data well. Keeping δ constant yielded fewer model fits with bad parameter estimates compared to when four parameters were estimated, but most of them yielded good NLLs. This is likely because the relationships between δ and λ and μ were not constrained by the low RL data. Although their estimates were highly correlated for 56 of 100 data sets (Figure 3.11A), there were many other values that δ could take that fit the data well that were not dependent on κ or μ. Therefore, wide ranges of κ, μ, and λ fit the data well, so parameter searches that begun far away from θ yielded bad parameter estimates that still fit the data well based on their NLLs. In addition, since the data have noise, exactly θ would not be expected to fit the data as well as another combination of values, so fixing parameters at θ could potentially lead to worse model fits, although this possibility has not been systematically explored.

For the medium RL data, even with four parameters estimated, none of the 138 model fits with bad parameter estimates yielded good NLLs. Thus, the parameters were already

identifiable in this case because good parameter estimates could be distinguished from bad ones, so keeping additional parameters constant provided little benefit.

For the high RL data, when four parameters were estimated, 139 model fits yielded good NLLs out of 301 that had bad parameter estimates. Keeping additional parameters constant did not improve identifiability for these data because the relationships among them were not constrained by the data. For example, even when only κ and μ were estimated for the high RL data, 126 model fits yielded good NLLs out of the 306 that had bad parameter estimates.

Insights from the simulation study

The goals of the simulation study were to better understand challenges in estimating the model parameters in order to guide future work with the empirical data. One challenge is that the parameter search is a local minimization of the objective function, which can yield sub-optimal model fits when poor starting values are used. Comparing 11 search algorithms indicated that the nlminb algorithm performed nearly perfectly in identifying the true parameter values (θ) for data with no noise, even with starting values that varied 100 million-fold. Nlminb was several times faster than the newuoa algorithm, the only one that gave comparable performance in terms of accuracy. The nlminb algorithm generally avoids getting trapped in local minima if better model fits exist elsewhere in the parameter space.

The second challenge in estimating the parameters is noise in the data. I demonstrated that with a low amount of noise in the data, the parameter estimates do not depend on the choice of starting values, as most of the estimates resulting from different starting values for a given data set were nearly identical. However, with more noise, searches using different sets of starting values can yield widely varying sets of parameter estimates that fit the same data equally well. This problem particularly affected the low and high ribosome loading (RL) data, while not

affecting the medium RL data as dramatically. When good and bad parameter estimates fit the data equally well, the parameters are unidentifiable. More noise can also make parameter estimates less reliable, where widely different parameter estimates are obtained when the data are re-sampled. As with identifiability, noise particularly affected the reliability of the parameter estimates for the low and high RL data, with more noise associated with wider confidence intervals (CIs) of each parameter estimate across replicate data sets.

Although noise can decrease the identifiability of some parameters and the reliability of their estimates, the relationships among parameters can be highly constrained by some data. I explained how the low and medium RL data could result from many combinations of transcription, initiation, and marking rates, as long as their ratios stay the same. This relationship was evident in the strong correlations among the transcription, initiation, and marking rate estimates in the low and medium RL data. The high RL data did not exhibit such constraints, however. This was because the distribution of mRNA abundances was shifted toward the upper polysome fractions as much as possible, so any combination of initiation and marking rates would yield the same model predictions as long as the initiation rate was sufficiently higher than the marking rate.

Lastly, I demonstrated that constraints in the relationships among parameters make it possible to improve identifiability by keeping additional parameters constant. The low RL data suffered from poor parameter identifiability, but the transcription, initiation, and marking rates had to be in the correct proportions to each other in order to fit the data well. Keeping one of those parameters constant, as I demonstrated with the transcription rate, limits the values of the other parameters that can fit the data well. As a result, parameter estimates resulting from different starting values become more consistent, and poor parameter estimates are more easily

distinguished from good ones because they do not fit the data as well, based on their NLLs. In this way, keeping additional parameters constant can make those that are estimated more identifiable. The medium RL data did not suffer from poor parameter identifiability, so keeping additional parameters constant did not improve the parameter estimation. The high RL data, as discussed, did suffer from poor parameter identifiability, but it had no constraints in the relationships among the parameters. Thus, keeping additional parameters constant did not improve the parameter search for these data either.

If keeping additional model parameters constant improves the identifiability of those that are estimated, future success in estimating the model parameters for empirical data may require empirical knowledge of certain parameter values. Narsai et al. measured genome-wide decay rates of mRNAs in Arabidopsis (Narsai, Howell et al. 2007). Conceivably, these decay rates could be combined with our empirical measurements of mRNA abundance in order to infer transcription rates, and the transcription rate, possibly in addition to the degradation rate, could be kept constant. One limitation of this approach might be that microarray expression signals cannot typically be compared across different genes due to variation in hybridization properties, while RNA-seq measurements are probably more proportional to the true number of mRNA molecules per gene in a biological sample. A number of groups have compared mRNA measurements between microarrays and RNA-seq in Arabidopsis (Giorgi, Del Fabbro et al. 2013). These data could be used to identify genes for which microarray measurements are good predictors of their absolute mRNA abundance, and rates of transcription and degradation that are at least proportional to their real rates for these genes could be estimated.

Fitting the model to empirical data

The ultimate goal of the research in this chapter is to accurately estimate the translation parameters for many genes across the Arabidopsis genome. As mentioned in the abstract and methods sections, efforts to fit the model to the empirical data have been unsuccessful to date. Possible sources of problems and suggestions for future work are considered in the discussion section below. Here, I briefly describe typical results from fitting the model to the empirical data in order to illustrate the need for future improvement.

I have focused on a set of 27 genes for testing the model fitting procedure on the empirical data. I selected the genes using the SelectAOV procedure, which first categorizes genes according to the polysome fraction where their mRNA abundance is highest, and then sorts the genes in each category according to a p-value. The p-value is from a one-way analysis of variance (ANOVA) where the replicate measurements in each fraction are considered as treatment groups, so lower p-values reflect higher ratios of between-fraction/within-fraction variance. The reasoning behind this approach is that these genes should represent a range of ribosome loading states and should have low noise and distinguishable polysome profiles. I selected 27 genes using this approach, with 3 genes having their peak mRNA abundance in each polysome fraction. The ANOVA p-values for these 27 genes ranged from 1.5e-13 to 5.9e-9, while the median 50% of all 14,199 genes was 0.0012 to 0.36. The polysome profiles for these genes are shown as a heatmap in Figure 3.12.

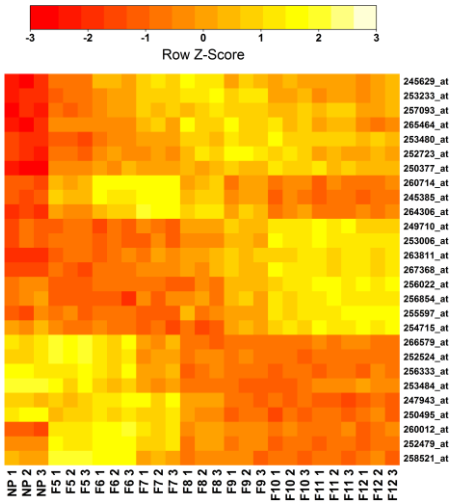

Figure 3.12: Heatmap of empirical mRNA levels for 27 genes.
The rows (genes, indicated by their Affymetrix probe ID) were arranged by hierarchical clustering, using the Pearson coefficient as the similarity metric. For coloring, the data in each row were scaled so that the colors indicate their Z-scores, which are the number of standard deviations below or above the mean of the row. The column labels indicate the fraction and replicate number (e.g., "NP 1" and "F10 2" are non-polysomal replicate 1 and fraction 10 replicate 2, respectively).

I ran parameter searches on these 27 genes in the same way as in the simulation study, using the nlminb search algorithm and 24 sets of starting values ranging from 0.0001 to 1e7, totaling 648 parameter searches. The lowest negative log-likelihoods (NLLs) for the 27 genes ranged from 51.5 to 226.2. Model fits for two genes with the lowest NLLs are shown in Figure 3.13. Clearly, the best-fitting parameter estimates do not fit the data well. The model predictions are generally on a similar scale as the empirical data, but their shapes are different. For both genes, mRNA is relatively abundant in the higher fractions, including fraction 12, but the abundances predicted by the parameter estimates have their peak in fraction 10 and are lower in fraction 11, with their lowest level in fraction 12.

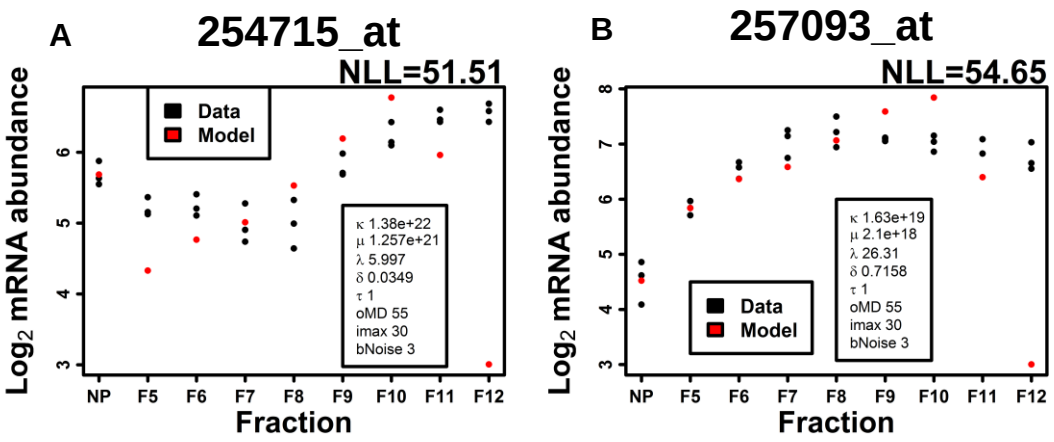

Figure 3.13: Plots of model fits for two genes from the empirical data.
Plots of model fits are shown for two genes from the empirical data with the lowest negative log-likelihoods (NLL) from parameter searches. The original data ("Data") are shown in black, and the model predictions ("Model") based on the parameter estimates are shown in red. The parameter estimates and constant parameter values are indicated in the second legend. The Affymetrix probe IDs and NLLs are indicated above the plots. See the PlotModelFit procedure and Table 3.3 for more details and symbols used.

There were several similarities between the results of parameter searches in the simulation study and for the empirical data. First, as in the simulation study, the individual parameters were not generally constrained by the data, as the parameter estimates resulting from different starting values varied widely. For instance, the relative standard deviations (RSDs) of the initiation rate estimates over the 24 model fits for these two genes were 2.7e111% and 8.2e98% of their medians. Second, while the parameter estimates varied widely, they were highly correlated, with the initiation and marking rate estimates being correlated in this case, as Figure 3.14 shows. Third, the NLLs were highly consistent, despite the widely varying parameter estimates. The RSDs of the NLLs for each gene ranged from 7.3e-8% to 1.5% of their medians. These observations suggest that the problem with fitting the model to the data is not because the parameter search gets trapped in local minima. Rather, there is typically one best

NLL for each gene, and the search finds it, although many combinations of parameter estimates yield the same NLL as long as they are in the correct proportions to each other.

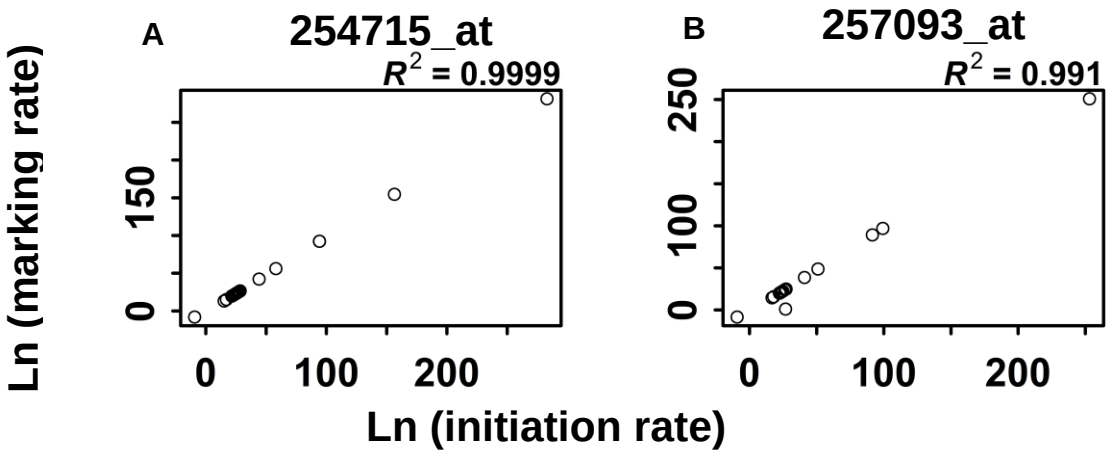

Figure 3.14: Relationship between estimated rates of initiation and marking in empirical data.
Parameter searches were run on 27 genes from the empirical data, using 24 sets of starting values. For the two genes with the lowest negative log-likelihoods (NLLs), the initiation rate estimates are plotted against the marking rate estimates for the 24 model fits. The Affymetrix probe IDs are indicated above the plots, along with the Pearson R^2 values.

Discussion

This chapter is more of a progress report than a report of biological discoveries. A great deal of work has gone into developing the model, implementing and testing it, and collecting data by a number of people. However, their contributions have until now remained largely separated, making it difficult to understand the status of the research and the best way to move the research forward. In this chapter, I have synthesized these contributions, including explaining the context of the research in terms of the biological and computational aspects and

documenting the experimental and computational procedures involved. In addition to bringing together prior contributions, I have made contributions to the research. I have expanded the usability of the computer code for implementing the model in terms of user-friendliness, flexibility, and performance. I have explored the process of parameter estimation through simulation studies. And from fitting the model to the empirical data, I have identified aspects that need improvement.

From the simulation studies, I first demonstrated that search algorithms can get trapped in local minima when very poor starting values are used. However, the nlminb algorithm significantly outperforms the others in terms of identifying the true parameter values (θ), and it is competitive in terms of speed.

Second, I described the effect of noise in the data on the identifiability of the model parameters and the reliability of their estimates. As I demonstrated for data with no noise, even a good search algorithm can occasionally return poor model fits and parameter estimates. So, to tell whether a parameter was identifiable in a given context, I did not focus on what percent of the parameter searches gave parameter estimates that were close to θ. Instead, I assessed whether good parameter estimates could be distinguished from bad ones based on how well they fit the data. Typically, most model fits for the same data resulting from different starting values had nearly identical negative log-likelihoods (NLLs) and parameter estimates, while fewer of them had a higher NLL, a different parameter estimate, or both. If a particular parameter estimate is different from the rest of them, but it fits the data just as well with the same NLL, then these good and bad parameter estimates cannot be distinguished, which makes the parameters unidentifiable.

For some data, higher levels of noise decreased parameter identifiability and the reliability of the parameter estimates, but the relationships among the parameters were still highly constrained. I showed that keeping one or two additional parameters constant can improve identifiability of other ones in cases where the relationships among parameters are constrained by the data. For the low RL data, the parameters were unidentifiable when all four were estimated, but the relationships among the rates of transcription, initiation, and marking were highly constrained by the data. Therefore, keeping the transcription rate constant greatly improved the identifiability of the initiation and marking rates. The medium RL data did not suffer from poor parameter identifiability, so keeping additional parameters constant had little impact. The high RL data did suffer from poor parameter identifiability, but the relationships among the parameters were not constrained, so keeping additional parameters constant did not significantly improve identifiability in this case. With this knowledge, future work in estimating the parameters for the empirical data may, in principle, utilize information about these parameters obtained experimentally.

Finally, I gave examples of typical results from estimating the parameters for the empirical data to demonstrate that these efforts have been largely unsuccessful to date. In principle, there are three possible areas in which problems could occur in estimating the parameters for the empirical data. First, the model itself could potentially be a poor representation of the regulation of mRNA ribosome loading. Second, the procedure for predicting polysome fraction mRNA abundances based on the model output, which are ribosome class mRNA abundances, may not be accurate. And third, the parameter search could fail to find the parameter values that best fit the data and instead become trapped in local minima due to poor starting values.

While no model in biology is perfect, there is no evidence yet that the model itself is the main source of the problem. Additionally, the simulation study indicated that the parameter searches nearly always find parameter estimates that fit the data as well as any others, despite poor starting values and noise in the data. In other words, searches using different starting values usually reach nearly identical NLLs. The main challenge in the parameter search, rather, is to unambiguously identify the true parameter values when multiple sets of parameter estimates fit the same data well. This leaves the procedure for predicting the empirical data based on the model output as a likely source of problems.

A singular value decomposition approach was used to "undo" the conversion procedure, which showed that infeasible model output would be required to match the empirical data based on the current procedure. Further, from a biological standpoint, the conversion procedure makes one important assumption that is not supported by empirical evidence or biological intuition. The conversion procedure first predicts how much mRNA migrates from each ribosome class to each polysome fraction, based on probabilities that are supported by empirical data (i.e., the UV absorbance profile). Second, the conversion procedure attempts to adjust these predicted fraction abundances to account for the effect of varying amounts of ribosomal RNA (rRNA) across polysome fractions. The procedure attempts to estimate how much rRNA is present in each polysome fraction by first estimating how much rRNA from each ribosome class across all genes migrates to each polysome fraction. The problematic assumption is that the mRNA in a cell from all genes is equally distributed among ribosome classes. The UV absorbance profile, however, suggests that there are different amounts of mRNA in each ribosome class, and that there is less mRNA in the higher ribosome classes. Thus, with the problematic assumption, the

conversion procedure likely incorrectly estimates the amount of rRNA in each fraction and therefore also the expected microarray expression signals.

Future success in estimating the translation parameters for the empirical data will likely require an improved procedure for predicting the empirical data based on the model output. In the methods section, I briefly described an alternative method for estimating how much total RNA is in each ribosome class based on the UV absorbance profile. In this procedure, image analysis software would be used to estimate the areas of the absorbance peaks that are well-defined, which correspond to ribosome classes from 2 up to 8. A deconvolution approach would be used to estimate the areas of the higher peaks that are not well-defined. The relative areas of these peaks would be expected to be proportional to the total RNA abundances, which could then be used to easily approximate how much rRNA is in each ribosome class, based on knowing how much rRNA corresponds to each ribosome. This method could possibly serve as an alternative to the current statistical approach for estimating the fraction location probabilities. These probabilities give the proportions of mRNA from each ribosome class that are expected to migrate to each polysome fraction. They are computed by assuming that the center of each peak represents the mean migration distance for the corresponding mRNA, and the widths of the peaks are used to estimate the migration distance standard deviation. In an alternative method, the proportion of a given absorbance peak that is located in each polysome fraction could be obtained easily using image analysis software. These proportions could then be used as fraction location probabilities. While there are no obvious problems with the current procedure, it would be worthwhile to compare the original statistical approach with a more empirical approach.

In conclusion, I have brought together the work of a number of people in the Gilchrist and von Arnim labs in developing a computational model of ribosome loading. The model will

hopefully add to our current understanding of gene expression regulation by describing how various biochemical steps combine to yield a particular state of protein expression. The model is unique in its field because it is deterministic and based on ordinary differential equations, unlike the many stochastic models of translation. Though efforts to fit the model to the empirical data have not been successful to date, important areas for improvement have been identified and work continues to address them.

Chapter 4

Conclusions and future perspectives

In this dissertation, I have presented investigations into translation regulation in Arabidopsis thaliana using computational approaches. As described in chapter 1, translation can be quantified at the genome scale by a variety of techniques, most of which target either proteins or RNA. In chapters 2 and 3, measurements of genome-wide translation were based on ribosome loading, which is the extent to which mRNA is associated with ribosomes. In chapter 2, mRNA samples were fractionated based on ribosome loading into three fractions which were quantified by microarrays and used to compute a translation state (TL) for each gene, yielding approximations of protein synthesis rates. While some information is lost by consolidating fractionated mRNA from a density gradient into a few fractions, this procedure made it feasible to carry out the large-scale experiment in chapter 2 involving multiple genotypes, polysome fractions, time points, and replicates. In chapter 3, mRNA abundances from nine polysome fractions were quantified by microarrays, giving a higher-resolution view of ribosome loading and providing data for parameterizing the computational model described.

The work in chapter 2 yielded new insights into the regulation of gene expression at the transcriptional and translational levels by the circadian clock. That study did investigate some biological mechanisms. For instance, by comparing diel cycles of genome-wide transcript levels and TL between wild-type plants and those with a dysfunctional circadian clock—due to constitutive overexpression of the *CCA1* gene—it revealed that the clock not only regulates a large proportion of the Arabidopsis transcriptome, as was already known. The clock also regulates the degree of ribosome loading of mRNAs from diverse functional categories. Nevertheless, it did not address the specific mechanisms underlying the diel cycles in transcript levels and TL or the differences in these cycles between genotypes. By contrast, the work in chapter 3 explored the mechanisms behind control of ribosome loading itself. Diel TL

cycles, as well as differences in cycles between genotypes described in chapter 2, could be explained by changes in the abundance of mRNA in non-polysomal, small polysome, or large polysome fractions, or some combination of the three. Indeed, the simulation studies in chapter 3 were consistent with the notion that a given pattern of ribosome loading can result from widely varying combinations of rates of biochemical translation reactions. Chapters 2 and 3, therefore, complement each other in the following sense. While chapter 2 was largely descriptive and focused on identifying genes whose ribosome loading was influenced by the circadian clock, chapter 3 was more mechanistic and focused on the fundamental question of how ribosome loading is regulated.

Chapters 2 and 3 leave a number of questions unanswered and highlight important topics for further study. Chapter 2 revealed that many genes exhibit diel cycles of ribosome loading in Arabidopsis and that many of these diel cycles are partially controlled by the circadian clock. It also validated the ribosomal proteins as a "translational regulon" since their TL cycles were highly coordinated. However, one question left open is, what cellular pathways connect the clock—specifically the cycling in *CCA1* expression—with ribosome loading and translation? Second, the study adds the circadian clock to the large and growing list of both environmental and endogenous cues that regulate ribosome loading. But how does the endogenous circadian clock interact with these external factors in regulating transcription and translation? Third, chapter 2 showed that global ribosome loading increased during the day and decreased at night under the influence of the clock, as was the case for many specific mRNAs. mRNAs encoding ribosomal proteins did the opposite, with ribosome loading peaking at night and dropping to the lowest levels near midday. This begs the question of how and why these mRNAs escape the global ribosome loading pattern and behave in their own way.

The computational model presented in chapter 3 and many procedures for implementing it are still in active development. Once the model fits the empirical data better, new challenges will likely be encountered. As the simulation studies showed, the model parameters are largely unidentifiable since many combinations of parameter values can yield the same ribosome loading pattern. Success in estimating the model parameters for a large number of genes will likely require obtaining some parameter values from empirical measurements in the literature. Once the model parameters can be estimated with good precision, the next goal will be to address the question of how the biochemical rates change in response to environmental stimuli, such as stresses. Finally, future studies of the regulation of ribosome loading and translation will likely benefit from techniques that are becoming established for studying translation at the genome level, including RNA sequencing and mass spectrometry-based proteomics.

List of references

1. Abler, M. L. and P. J. Green (1996). "Control of mRNA stability in higher plants." Plant molecular biology **32**(1-2): 63-78.
2. Akashi, H. (2003). "Translational selection and yeast proteome evolution." Genetics **164**(4): 1291-1303.
3. Ames, B. N. and P. E. Hartman (1963). "The Histidine Operon." Cold Spring Harbor Symposia on Quantitative Biology **28**: 349-356.
4. An, G. (2001). "Agent-based computer simulation and sirs: building a bridge between basic science and clinical trials." Shock **16**(4): 266-273.
5. Analytics, R. and S. Weston (2014). "doParallel: Foreach parallel adaptor for the parallel package." R package version **1**(8).
6. Andersson, S. G. and C. G. Kurland (1990). "Codon preferences in free-living microorganisms." Microbiological Reviews **54**(2): 198-210.
7. Andrei, M., D. Ingelfinger, R. Heintzmann, T. Achsel, R. Rivera-Pomar and R. Luehrmann (2005). "A role for eIF4E and eIF4E-transporter in targeting mRNPs to mammalian processing bodies." RNA **11**(5): 717-727.
8. Arava, Y., Y. Wang, J. D. Storey, C. L. Liu, P. O. Brown and D. Herschlag (2003). "Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae." Proceedings of the National Academy of Sciences **100**(7): 3889-3894.
9. Aronson, B. D., K. A. Johnson, J. J. Loros and J. C. Dunlap (1994). "Negative feedback defining a circadian clock: autoregulation of the clock gene frequency." Science **263**(5153): 1578-1584.
10. Arribas-Layton, M., D. Wu, J. Lykke-Andersen and H. Song (2013). "Structural and functional control of the eukaryotic mRNA decapping machinery." Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms **1829**(6–7): 580-589.
11. Arribere, Joshua A., Jennifer A. Doudna and Wendy V. Gilbert (2011). "Reconsidering Movement of Eukaryotic mRNAs between Polysomes and P Bodies." Molecular Cell **44**(5): 745-758.
12. Baek, D., J. Villén, C. Shin, F. D. Camargo, S. P. Gygi and D. P. Bartel (2008). "The impact of microRNAs on protein output." Nature **455**(7209): 64-71.
13. Banerjee, A. K., T. Lin and D. J. Hannapel (2009). "Untranslated Regions of a Mobile Transcript Mediate RNA Metabolism." Plant Physiology **151**(4): 1831-1843.
14. Bankes, S. C. (2002). "Agent-based modeling: A revolution?" Proceedings of the National Academy of Sciences **99**(suppl 3): 7199-7200.
15. Bartel, D. P. (2004). "MicroRNAs: Genomics, Biogenesis, Mechanism, and Function." Cell **116**(2): 281-297.
16. Bartel, D. P. (2009). "MicroRNAs: Target Recognition and Regulatory Functions." Cell **136**(2): 215-233.
17. Bashkirov, V. I., H. Scherthan, J. A. Solinger, J.-M. Buerstedde and W.-D. Heyer (1997). "A Mouse Cytoplasmic Exoribonuclease (mXRN1p) with Preference for G4 Tetraplex Substrates." The Journal of Cell Biology **136**(4): 761-773.
18. Baskerville, S. and D. P. Bartel (2005). "Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes." RNA **11**(3): 241-247.
19. Bates, D., K. M. Mullen, J. C. Nash and R. Varadhan (2014). "minqa: Derivative-free optimization algorithms by quadratic approximation. R package version 1.2.4.".

20. Baumberger, N. and D. C. Baulcombe (2005). "Arabidopsis ARGONAUTE1 is an RNA Slicer that selectively recruits microRNAs and short interfering RNAs." Proceedings of the National Academy of Sciences of the United States of America **102**(33): 11928-11933.
21. Bazzini, A. A., M. T. Lee and A. J. Giraldez (2012). "Ribosome Profiling Shows That miR-430 Reduces Translation Before Causing mRNA Decay in Zebrafish." Science **336**(6078): 233-237.
22. Beelman, C. A., A. Stevens, G. Caponigro, T. E. LaGrandeur, L. Hatfield, D. M. Fortner and R. Parker (1996). "An essential component of the decapping enzyme required for normal rates of mRNA turnover."
23. Benaglia, T., D. Chauveau, D. Hunter and D. Young (2009). "mixtools: An R package for analyzing finite mixture models." Journal of Statistical Software **32**(6): 1-29.
24. Bennetzen, J. L. and B. Hall (1982). "Codon selection in yeast." Journal of Biological Chemistry **257**(6): 3026-3031.
25. Berthelot, K., M. Muldoon, L. Rajkowitsch, J. Hughes and J. E. McCarthy (2004). "Dynamics and processivity of 40S ribosome scanning on mRNA in yeast." Molecular microbiology **51**(4): 987-1001.
26. Bi, X. and D. J. Goss (2000). "Kinetic proofreading scanning models for eukaryotic translational initiation: the cap and poly (A) tail dependency of translation." Journal of theoretical biology **207**(2): 145-157.
27. Birgin, E. G., Jos, #233, M. Mart, #237, nez and M. Raydan (2001). "Algorithm 813: SPG—Software for Convex-Constrained Optimization." ACM Trans. Math. Softw. **27**(3): 340-349.
28. Birgin, E. G., J. M. Martínez and M. Raydan (2001). "Algorithm 813: SPG—software for convex-constrained optimization." ACM Transactions on Mathematical Software (TOMS) **27**(3): 340-349.
29. Blythe, R. A. and M. R. Evans (2007). "Nonequilibrium steady states of matrix-product form: a solver's guide." Journal of Physics A: Mathematical and Theoretical **40**(46): R333.
30. Bonabeau, E. (2002). "Agent-based modeling: Methods and techniques for simulating human systems." Proceedings of the National Academy of Sciences **99**(suppl 3): 7280-7287.
31. Bonnans, J.-F., J. C. Gilbert, C. Lemaréchal and C. A. Sagastizábal (2006). Numerical optimization: theoretical and practical aspects, Springer Science & Business Media.
32. Branco-Price, C., K. A. Kaiser, C. J. Jang, C. K. Larive and J. Bailey-Serres (2008). "Selective mRNA translation coordinates energetic and metabolic adjustments to cellular oxygen deprivation and reoxygenation in Arabidopsis thaliana." Plant J **56**(5): 743-755.
33. Branco-Price, C., K. A. Kaiser, C. J. H. Jang, C. K. Larive and J. Bailey-Serres (2008). "Selective mRNA translation coordinates energetic and metabolic adjustments to cellular oxygen deprivation and reoxygenation in Arabidopsis thaliana." The Plant Journal **56**(5): 743-755.
34. Branco-Price, C., R. Kawaguchi, R. B. Ferreira and J. Bailey-Serres (2005). "Genome-wide analysis of transcript abundance and translation in Arabidopsis seedlings subjected to oxygen deprivation." Annals of botany **96**(4): 647-660.

35. Braun, K. A., K. M. Dombek and E. T. Young (2016). "Snf1-Dependent Transcription Confers Glucose-Induced Decay upon the mRNA Product." Molecular and Cellular Biology **36**(4): 628-644.
36. Brengues, M., D. Teixeira and R. Parker (2005). "Movement of Eukaryotic mRNAs Between Polysomes and Cytoplasmic Processing Bodies." Science **310**(5747): 486-489.
37. Bulmer, M. (1991). "The selection-mutation-drift theory of synonymous codon usage." Genetics **129**(3): 897.
38. Byrd, R. H., P. Lu, J. Nocedal and C. Zhu (1995). "A Limited Memory Algorithm for Bound Constrained Optimization." SIAM Journal on Scientific Computing **16**(5): 1190-1208.
39. Calvo, S. E., D. J. Pagliarini and V. K. Mootha (2009). "Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans." Proceedings of the National Academy of Sciences **106**(18): 7507-7512.
40. Cannarozzi, G., N. N. Schraudolph, M. Faty, P. von Rohr, M. T. Friberg, A. C. Roth, P. Gonnet, G. Gonnet and Y. Barral (2010). "A Role for Codon Order in Translation Dynamics." Cell **141**(2): 355-367.
41. Carrington, J. C. and V. Ambros (2003). "Role of MicroRNAs in Plant and Animal Development." Science **301**(5631): 336-338.
42. Caskey, C. T., A. L. Beaudet, E. M. Scolnick and M. Rosman (1971). "Hydrolysis of fMet-tRNA by Peptidyl Transferase." Proceedings of the National Academy of Sciences **68**(12): 3163-3167.
43. Chang, W., J. Cheng, J. Allaire, Y. Xie and J. McPherson (2015). "Shiny: web application framework for R." R package version 0.11 **1**.
44. Chen, C.-Y. A. and A.-B. Shyu (2011). "Mechanisms of deadenylation-dependent decay." Wiley Interdisciplinary Reviews: RNA **2**(2): 167-183.
45. Chen, C.-Y. A. and A.-B. Shyu (2013). Deadenylation and P-Bodies. Ten Years of Progress in GW/P Body Research. L. E. K. Chan and J. M. Fritzler. New York, NY, Springer New York**:** 183-195.
46. Chen, J., A. Tsai, S. E. O'Leary, A. Petrov and J. D. Puglisi (2012). "Unraveling the dynamics of ribosome translocation." Current Opinion in Structural Biology **22**(6): 804-814.
47. Chou, T. and G. Lakatos (2004). "Clustered Bottlenecks in mRNA Translation and Protein Synthesis." Physical Review Letters **93**(19): 198101.
48. Chu, D., E. Kazana, N. Bellanger, T. Singh, M. F. Tuite and T. von der Haar (2013). "Translation elongation can control translation initiation on eukaryotic mRNAs." The EMBO Journal **33**(1): 21-34.
49. Chu, D., J. Thompson and T. von der Haar (2014). "Charting the dynamics of translation." Biosystems **119**: 1-9.
50. Chu, D. and T. von der Haar (2012). "The architecture of eukaryotic translation." Nucleic Acids Research **40**(20): 10098-10106.
51. Chu, D., N. Zabet and T. von der Haar (2012). "A novel and versatile computational tool to model translation." Bioinformatics **28**(2): 292-293.
52. Claude, J. P. B., xe and lisle (1992). "Convergence Theorems for a Class of Simulated Annealing Algorithms on R^d." Journal of Applied Probability **29**(4): 885-895.

53. Coghlan, A. and K. H. Wolfe (2000). "Relationship of codon bias to mRNA concentration and protein length in Saccharomyces cerevisiae." Yeast **16**(12): 1131-1145.
54. Cougot, N., S. Babajko and B. Séraphin (2004). "Cytoplasmic foci are sites of mRNA decay in human cells." The Journal of Cell Biology **165**(1): 31-40.
55. Covington, M. F., J. N. Maloof, M. Straume, S. A. Kay and S. L. Harmer (2008). "Global transcriptome analysis reveals circadian regulation of key pathways in plant growth and development." Genome Biol **9**(8): R130.
56. Crick, F. (1970). "Central dogma of molecular biology." Nature **227**(5258): 561-563.
57. Crick, F. H. (1958). On protein synthesis. Symposia of the Society for Experimental Biology.
58. Crittenden, S. L., D. S. Bernstein, J. L. Bachorik, B. E. Thompson, M. Gallegos, A. G. Petcherski, G. Moulder, R. Barstead, M. Wickens and J. Kimble (2002). "A conserved RNA-binding protein controls germline stem cells in Caenorhabditis elegans." Nature **417**(6889): 660-663.
59. Dai, Y. H. and Y. Yuan (2001). "An Efficient Hybrid Conjugate Gradient Method for Unconstrained Optimization." Annals of Operations Research **103**(1-4): 33-47.
60. Darnell, Jennifer C., Sarah J. Van Driesche, C. Zhang, Ka Ying S. Hung, A. Mele, Claire E. Fraser, Elizabeth F. Stone, C. Chen, John J. Fak, Sung W. Chi, Donny D. Licatalosi, Joel D. Richter and Robert B. Darnell (2011). "FMRP Stalls Ribosomal Translocation on mRNAs Linked to Synaptic Function and Autism." Cell **146**(2): 247-261.
61. De Silva, E., J. Krishnan, R. Betney and I. Stansfield (2010). "A mathematical modelling framework for elucidating the role of feedback control in translation termination." Journal of theoretical biology **264**(3): 808-821.
62. de Sousa Abreu, R., L. O. Penalva, E. M. Marcotte and C. Vogel (2009). "Global signatures of protein and mRNA expression levels." Molecular BioSystems **5**(12): 1512-1526.
63. Decker, C. J. and R. Parker (1993). "A turnover pathway for both stable and unstable mRNAs in yeast: evidence for a requirement for deadenylation." Genes & Development **7**(8): 1632-1643.
64. Dennis, J. and R. Schnabel (1996). Numerical Methods for Unconstrained Optimization and Nonlinear Equations, Society for Industrial and Applied Mathematics.
65. DiCiccio, T. J. and B. Efron (1996). "Bootstrap confidence intervals." Statistical science: 189-212.
66. Dimelow, R. J. and S. J. Wilkinson (2009). "Control of translation initiation: a model-based analysis from limited experimental data." Journal of The Royal Society Interface **6**(30): 51-61.
67. Dong, H., L. Nilsson and C. G. Kurland (1996). "Co-variation of tRNA Abundance and Codon Usage in Escherichia coli at Different Growth Rates." Journal of Molecular Biology **260**(5): 649-663.
68. Dongarra, J. J. and E. Grosse (1987). "Distribution of mathematical software via electronic mail." Commun. ACM **30**(5): 403-407.
69. dos Reis, M., R. Savva and L. Wernisch (2004). "Solving the riddle of codon usage preferences: a test for translational selection." Nucleic acids research **32**(17): 5036-5044.

70. dos Reis, M., L. Wernisch and R. Savva (2003). "Unexpected correlations between gene expression and codon usage bias from microarray data for the whole Escherichia coli K-12 genome." Nucleic acids research **31**(23): 6976-6985.
71. Duret, L. (2000). "tRNA gene number and codon usage in the C. elegans genome are co-adapted for optimal translation of highly expressed genes." Trends in Genetics **16**(7): 287-289.
72. Eddelbuettel, D., R. François, J. Allaire, J. Chambers, D. Bates and K. Ushey (2011). "Rcpp: Seamless R and C++ integration." Journal of Statistical Software **40**(8): 1-18.
73. Eriksson, M. E. and A. A. R. Webb (2011). "Plant cell responses to cold are all about timing." Current opinion in plant biology **14**(6): 731-737.
74. Eulalio, A., I. Behm-Ansmant, D. Schweizer and E. Izaurralde (2007). "P-Body Formation Is a Consequence, Not the Cause, of RNA-Mediated Gene Silencing." Molecular and Cellular Biology **27**(11): 3970-3981.
75. Fedorov, A. N. and T. O. Baldwin (1997). "Cotranslational Protein Folding." Journal of Biological Chemistry **272**(52): 32715-32718.
76. Filichkin, S. A. and T. C. Mockler (2012). "Unproductive alternative splicing and nonsense mRNAs: a widespread phenomenon among plant circadian clock genes." Biol Direct **7**: 20.
77. Filipowicz, W. (2005). "RNAi: The Nuts and Bolts of the RISC Machine." Cell **122**(1): 17-20.
78. Fletcher, R. (1970). "A new approach to variable metric algorithms." The Computer Journal **13**(3): 317-322.
79. Fletcher, R. and C. M. Reeves (1964). "Function minimization by conjugate gradients." The Computer Journal **7**(2): 149-154.
80. Fluitt, A., E. Pienaar and H. Viljoen (2007). "Ribosome kinetics and aa-tRNA competition determine rate and fidelity of peptide synthesis." Computational Biology and Chemistry **31**(5–6): 335-346.
81. Fox, P. A., A. P. Hall and N. L. Schryer (1978). "The PORT Mathematical Subroutine Library." ACM Trans. Math. Softw. **4**(2): 104-126.
82. Frank-Kamenetsky, M., A. Grefhorst, N. N. Anderson, T. S. Racie, B. Bramlage, A. Akinc, D. Butler, K. Charisse, R. Dorkin, Y. Fan, C. Gamba-Vitalo, P. Hadwiger, M. Jayaraman, M. John, K. N. Jayaprakash, M. Maier, L. Nechev, K. G. Rajeev, T. Read, I. Röhl, J. Soutschek, P. Tan, J. Wong, G. Wang, T. Zimmermann, A. de Fougerolles, H.-P. Vornlocher, R. Langer, D. G. Anderson, M. Manoharan, V. Koteliansky, J. D. Horton and K. Fitzgerald (2008). "Therapeutic RNAi targeting PCSK9 acutely lowers plasma cholesterol in rodents and LDL cholesterol in nonhuman primates." Proceedings of the National Academy of Sciences **105**(33): 11915-11920.
83. Franks, T. M. and J. Lykke-Andersen (2008). "The Control of mRNA Decapping and P-Body Formation." Molecular Cell **32**(5): 605-615.
84. Fredrick, K. and M. Ibba (2010). "How the Sequence of a Gene Can Tune Its Translation." Cell **141**(2): 227-229.
85. Friedel, C. C., L. Dölken, Z. Ruzsics, U. H. Koszinowski and R. Zimmer (2009). "Conserved principles of mammalian transcriptional regulation revealed by RNA half-life." Nucleic Acids Research **37**(17): e115.

86. Friedman, R. C., K. K.-H. Farh, C. B. Burge and D. P. Bartel (2009). "Most mammalian mRNAs are conserved targets of microRNAs." Genome Research **19**(1): 92-105.
87. García-Martínez, J., L. Delgado-Ramos, G. Ayala, V. Pelechano, D. A. Medina, F. Carrasco, R. González, E. Andrés-León, L. Steinmetz, J. Warringer, S. Chávez and J. E. Pérez-Ortín (2015). "The cellular growth rate controls overall mRNA turnover, and modulates either transcription or degradation rates of particular gene regulons." Nucleic Acids Research.
88. Garrick, M. D. (1967). "The kinetics of the translation of messenger RNA into protein." Journal of Theoretical Biology **17**(1): 19-30.
89. Gautier, L., L. Cope, B. M. Bolstad and R. A. Irizarry (2004). "affy—analysis of Affymetrix GeneChip data at the probe level." Bioinformatics **20**(3): 307-315.
90. Gay, D. (1984). A trust-region approach to linearly constrained optimization. Numerical Analysis. D. Griffiths, Springer Berlin Heidelberg. **1066:** 72-105.
91. Gay, D. M. (1983). "Algorithm 611: Subroutines for Unconstrained Minimization Using a Model/Trust-Region Approach." ACM Trans. Math. Softw. **9**(4): 503-524.
92. Gentleman, R. C., V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge and J. Gentry (2004). "Bioconductor: open software development for computational biology and bioinformatics." Genome biology **5**(10): R80.
93. Gerashchenko, M. V., A. V. Lobanov and V. N. Gladyshev (2012). "Genome-wide ribosome profiling reveals complex translational regulation in response to oxidative stress." Proceedings of the National Academy of Sciences **109**(43): 17394-17399.
94. Gerst, I. and S. N. Levine (1965). "Kinetics of protein synthesis by polyribosomes." Journal of Theoretical Biology **9**(1): 16-36.
95. Gilchrist, M. A. and A. Wagner (2006). "A model of protein translation including codon bias, nonsense errors, and ribosome recycling." Journal of Theoretical Biology **239**(4): 417-434.
96. Giorgi, F. M., C. Del Fabbro and F. Licausi (2013). "Comparative study of RNA-seq-and microarray-derived coexpression networks in Arabidopsis thaliana." Bioinformatics **29**(6): 717-724.
97. Goldenberg, D., I. Azar and A. B. Oppenheim (1996). "Differential mRNA stability of the cspA gene in the cold-shock response of Escherichia coli." Molecular Microbiology **19**(2): 241-248.
98. Goldstrohm, A. C., B. A. Hook, D. J. Seay and M. Wickens (2006). "PUF proteins bind Pop2p to regulate messenger RNAs." Nature structural & molecular biology **13**(6): 533-539.
99. Goldstrohm, A. C. and M. Wickens (2008). "Multifunctional deadenylase complexes diversify mRNA control." Nat Rev Mol Cell Biol **9**(4): 337-344.
100. Graf, A., A. Schlereth, M. Stitt and A. M. Smith (2010). "Circadian control of carbohydrate availability for growth in Arabidopsis plants at night." Proceedings of the National Academy of Sciences **107**(20): 9458-9463.
101. Graham, R. L., T. S. Woodall and J. M. Squyres Open MPI: A flexible high performance MPI, Springer.
102. Grantham, R. (1980). "Working of the genetic code." Trends in Biochemical Sciences **5**(12): 327-331.

103. Grantham, R., C. Gautier, M. Gouy, M. Jacobzone and R. Mercier (1981). "Codon catalog usage is a genome strategy modulated for gene expressivity." Nucleic Acids Research **9**(1): 213.
104. Grantham, R., C. Gautier, M. Gouy, R. Mercier and A. Pavé (1980). "Codon catalog usage and the genome hypothesis." Nucleic Acids Research **8**(1): 197.
105. Gray, N. K., J. M. Coller, K. S. Dickson and M. Wickens (2000). "Multiple portions of poly (A)-binding protein stimulate translation in vivo." The EMBO Journal **19**(17): 4723-4733.
106. Green, R. M., S. Tingay, Z. Y. Wang and E. M. Tobin (2002). "Circadian rhythms confer a higher level of fitness to Arabidopsis plants." Plant Physiol **129**: 576-584.
107. Grimm, V., E. Revilla, U. Berger, F. Jeltsch, W. M. Mooij, S. F. Railsback, H.-H. Thulke, J. Weiner, T. Wiegand and D. L. DeAngelis (2005). "Pattern-oriented modeling of agent-based complex systems: lessons from ecology." science **310**(5750): 987-991.
108. Gu, W., T. Zhou and C. O. Wilke (2010). "A Universal Trend of Reduced mRNA Stability near the Translation-Initiation Site in Prokaryotes and Eukaryotes." PLoS Comput Biol **6**(2): e1000664.
109. Guo, H., N. T. Ingolia, J. S. Weissman and D. P. Bartel (2010). "Mammalian microRNAs predominantly act to decrease target mRNA levels." Nature **466**(7308): 835-840.
110. Guttman, M., P. Russell, Nicholas T. Ingolia, Jonathan S. Weissman and Eric S. Lander (2013). "Ribosome Profiling Provides Evidence that Large Noncoding RNAs Do Not Encode Proteins." Cell **154**(1): 240-251.
111. Hambraeus, G., C. Wachenfeldt and L. Hederstedt (2003). "Genome-wide survey of mRNA half-lives in Bacillus subtilis identifies extremely stable mRNAs." Molecular Genetics and Genomics **269**(5): 706-714.
112. Han, Y., A. David, B. Liu, J. G. Magadán, J. R. Bennink, J. W. Yewdell and S.-B. Qian (2012). "Monitoring cotranslational protein folding in mammalian cells at codon resolution." Proceedings of the National Academy of Sciences **109**(31): 12467-12472.
113. Hardin, P. E., J. C. Hall and M. Rosbash (1992). "Circadian oscillations in period gene mRNA levels are transcriptionally regulated." Proc Natl Acad Sci U S A **89**(24): 11711-11715.
114. Harley, C. B., J. W. Pollard, C. P. Stanners and S. Goldstein (1981). "Model for messenger RNA translation during amino acid starvation applied to the calculation of protein synthetic error rates." Journal of Biological Chemistry **256**(21): 10786-10794.
115. Harmer, S. L., J. B. Hogenesch, M. Straume, H.-S. Chang, B. Han, T. Zhu, X. Wang, J. A. Kreps and S. A. Kay (2000). "Orchestrated Transcription of Key Pathways in Arabidopsis by the Circadian Clock." Science **290**(5499): 2110-2113.
116. Heinrich, R. and T. A. Rapoport (1980). "Mathematical modelling of translation of mRNA in eucaryotes; steady states, time-dependent processes and application to reticulocytest." Journal of theoretical biology **86**(2): 279-313.
117. Heyd, A. and D. A. Drew (2003). "A mathematical model for elongation of a peptide chain." Bulletin of mathematical biology **65**(6): 1095-1109.
118. Hinnebusch, A. G., T. E. Dever and K. Asano (2007). "Mechanism of Translation Initiation in the Yeast Saccharomyces cerevisiae." Cold Spring Harbor Monograph Archive **48**: 225-268.

119. Holz, M. K., B. A. Ballif, S. P. Gygi and J. Blenis (2005). "mTOR and S6K1 mediate assembly of the translation preinitiation complex through dynamic protein interchange and ordered phosphorylation events." Cell **123**(4): 569-580.
120. Hsu, C. L. and A. Stevens (1993). "Yeast cells lacking 5'-->3' exoribonuclease 1 contain mRNA species that are poly(A) deficient and partially lack the 5' cap structure." Molecular and Cellular Biology **13**(8): 4826-4835.
121. Hsu, P. Y. and S. L. Harmer (2014). "Wheels within wheels: the plant circadian system." Trends in plant science **19**(4): 240-249.
122. Hu, W., K. A. Franklin, R. A. Sharrock, M. A. Jones, S. L. Harmer and J. C. Lagarias (2013). "Unanticipated regulatory roles for Arabidopsis phytochromes revealed by null mutant analysis." Proceedings of the National Academy of Sciences **110**(4): 1542-1547.
123. Hunt, A. G., R. Xu, B. Addepalli, S. Rao, K. P. Forbes, L. R. Meeks, D. Xing, M. Mo, H. Zhao, A. Bandyopadhyay, L. Dampanaboina, A. Marion, C. Von Lanken and Q. Q. Li (2008). "Arabidopsis mRNA polyadenylation machinery: comprehensive analysis of protein-protein interactions and gene expression profiling." BMC Genomics **9**: 220-220.
124. Iben, J. R. and R. J. Maraia (2014). "tRNA gene copy number variation in humans." Gene **536**(2): 376-384.
125. Ikemura, T. (1981). "Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system." Journal of molecular biology **151**(3): 389-409.
126. Ikemura, T. (1982). "Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes: differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs." Journal of molecular biology **158**(4): 573-597.
127. Ikemura, T. (1985). "Codon usage and tRNA content in unicellular and multicellular organisms." Molecular Biology and Evolution **2**(1): 13-34.
128. Ikemura, T., S. Osawa, H. Ozeki, H. Uchida and T. Yura (1980). "The frequency of codon usage in E. coli genes: correlation with abundance of cognate tRNA." Genetics and evolution of RNA polymerase, tRNA, and ribosomes. University of Tokyo Press, Tokyo, and Elsevier/North Holland, Amsterdam: 519-534.
129. Ikemura, T. and H. Ozeki (1977). "Gross map location of Escherichia coli transfer RNA genes." Journal of molecular biology **117**(2): 419-446.
130. Ikemura, T. and H. Ozeki (1983). Codon usage and transfer RNA contents: organism-specific codon-choice patterns in reference to the isoacceptor contents. Cold Spring Harbor symposia on quantitative biology, Cold Spring Harbor Laboratory Press.
131. Ingolia, N. T. (2014). "Ribosome profiling: new views of translation, from single codons to genome scale." Nat Rev Genet **15**(3): 205-213.
132. Ingolia, N. T., G. A. Brar, S. Rouskin, A. M. McGeachy and J. S. Weissman (2012). "The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments." Nat. Protocols **7**(8): 1534-1550.
133. Ingolia, N. T., S. Ghaemmaghami, J. R. S. Newman and J. S. Weissman (2009). "Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling." Science **324**(5924): 218-223.

134. Ingolia, Nicholas T., Liana F. Lareau and Jonathan S. Weissman (2011). "Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes." Cell **147**(4): 789-802.
135. Ingraham, J. L., O. Maaløe and F. C. Neidhardt (1983). Growth of the bacterial cell, Sinauer Associates.
136. Irwin, B., J. D. Heck and G. W. Hatfield (1995). "Codon Pair Utilization Biases Influence Translational Elongation Step Times." Journal of Biological Chemistry **270**(39): 22801-22806.
137. Jacklet, J. (1977). "Neuronal circadian rhythm: phase shifting by a protein synthesis inhibitor." Science **198**(4312): 69-71.
138. Jacklet, J. W. (1977). "Neuronal circadian rhythm: phase shifting by a protein synthesis inhibitor." Science **198**(4312): 69-71.
139. Jackson, R. J. (1991). "The ATP requirement for initiation of eukaryotic translation varies according to the mRNA species." European Journal of Biochemistry **200**(2): 285-294.
140. Jackson, R. J., C. U. T. Hellen and T. V. Pestova (2010). "The mechanism of eukaryotic translation initiation and principles of its regulation." Nat Rev Mol Cell Biol **11**(2): 113-127.
141. James, A. B., N. H. Syed, S. Bordage, J. Marshall, G. A. Nimmo, G. I. Jenkins, P. Herzyk, J. W. Brown and H. G. Nimmo (2012). "Alternative splicing mediates responses of the Arabidopsis circadian clock to temperature changes." Plant Cell **24**(3): 961-981.
142. James, A. B., N. H. Syed, S. Bordage, J. Marshall, G. A. Nimmo, G. I. Jenkins, P. Herzyk, J. W. S. Brown and H. G. Nimmo (2012). "Alternative splicing mediates responses of the Arabidopsis circadian clock to temperature changes." The Plant Cell **24**(3): 961-981.
143. James, A. B., N. H. Syed, J. W. Brown and H. G. Nimmo (2012). "Thermoplasticity in the plant circadian clock: how plants tell the time-perature." Plant Signal Behav **7**(10): 1219-1223.
144. Ji, L., X. Liu, J. Yan, W. Wang, R. E. Yumul, Y. J. Kim, T. T. Dinh, J. Liu, X. Cui and B. Zheng (2011). "ARGONAUTE10 and ARGONAUTE1 regulate the termination of floral stem cells through two microRNAs in Arabidopsis." PLoS Genet **7**(3): e1001358.
145. Johannes, G., M. S. Carter, M. B. Eisen, P. O. Brown and P. Sarnow (1999). "Identification of eukaryotic mRNAs that are translated at reduced cap binding complex eIF4F concentrations using a cDNA microarray." Proceedings of the National Academy of Sciences **96**(23): 13118-13123.
146. Jones-Rhoades, M. W. and D. P. Bartel (2004). "Computational Identification of Plant MicroRNAs and Their Targets, Including a Stress-Induced miRNA." Molecular Cell **14**(6): 787-799.
147. Jones-Rhoades, M. W., D. P. Bartel and B. Bartel (2006). "MicroRNAs AND THEIR REGULATORY ROLES IN PLANTS." Annual Review of Plant Biology **57**(1): 19-53.
148. Jones, M. A., B. A. Williams, J. McNicol, C. G. Simpson, J. W. Brown and S. L. Harmer (2012). "Mutation of Arabidopsis spliceosomal timekeeper locus1 causes circadian clock defects." Plant Cell **24**(10): 4066-4082.
149. Jouffe, C., G. Cretenet, L. Symul, E. Martin, F. Atger, F. Naef and F. Gachon (2013). "The circadian clock coordinates ribosome biogenesis." PLoS Biol **11**(1): e1001455.

150. Juntawong, P. and J. Bailey-Serres (2012). "Dynamic Light Regulation of Translation Status in Arabidopsis thaliana." Frontiers in plant science **3**: 66.
151. Juntawong, P. and J. Bailey-Serres (2012). "Dynamic light regulation of translation status in Arabidopsis thaliana." Frontiers Plant Sci **3**: 66.
152. Juntawong, P., T. Girke, J. Bazin and J. Bailey-Serres (2014). "Translational dynamics revealed by genome-wide profiling of ribosome footprints in Arabidopsis." Proceedings of the National Academy of Sciences **111**(1): E203-E212.
153. Juntawong, P., R. Sorenson and J. Bailey-Serres (2013). "Cold shock protein 1 chaperones mRNAs during translation in Arabidopsis thaliana." The Plant Journal **74**(6): 1016-1028.
154. Kanaya, S., Y. Yamada, Y. Kudo and T. Ikemura (1999). "Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis." Gene **238**(1): 143-155.
155. Kapp, L. D. and J. R. Lorsch (2004). "The molecular mechanics of eukaryotic translation." Annual review of biochemistry **73**(1): 657-704.
156. Kasprzyk, A. (2011). "BioMart: driving a paradigm change in biological data management." Database **2011**: bar049.
157. Kawaguchi, R., T. Girke, E. A. Bray and J. Bailey-Serres (2004). "Differential mRNA translation contributes to gene regulation under non-stress and dehydration stress conditions in Arabidopsis thaliana." Plant J **38**(5): 823-839.
158. Kawaguchi, R., T. Girke, E. A. Bray and J. Bailey-Serres (2004). "Differential mRNA translation contributes to gene regulation under non-stress and dehydration stress conditions in Arabidopsis thaliana." The Plant Journal **38**(5): 823-839.
159. Kelley, C. T. (1999). Iterative methods for optimization, Siam.
160. Kiho, Y. and A. Rich (1964). "Induced enzyme formed on bacterial polyribosomes." Proceedings of the National Academy of Sciences of the United States of America **51**(1): 111.
161. Kim, B.-H., X. Cai, J. N. Vaughn and A. G. von Arnim (2007). "On the functions of the h subunit of eukaryotic initiation factor 3 in late stages of translation initiation." Genome Biology **8**(4): 1-20.
162. Kim, B. H., X. Cai, J. N. Vaughn and A. G. von Arnim (2007). "On the functions of the h subunit of eukaryotic initiation factor 3 in late stages of translation initiation." Genome Biol **8**(4): R60.
163. Kim, J. Y., H. R. Song, B. L. Taylor and I. A. Carre (2003). "Light-regulated translation mediates gated induction of the Arabidopsis clock protein LHY." EMBO J **22**(4): 935-944.
164. Klug, G. (1991). "Endonucleolytic degradation of puf mRNA in Rhodobacter capsulatus is influenced by oxygen." Proceedings of the National Academy of Sciences **88**(5): 1765-1769.
165. Knight, R., S. Freeland and L. Landweber (2001). "A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes." Genome Biology **2**(4): research0010.0011 - research0010.0013.
166. Komar, A. A. (2009). "A pause for thought along the co-translational folding pathway." Trends in Biochemical Sciences **34**(1): 16-24.

167. Kozak, M. (1986). "Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes." Cell **44**(2): 283-292.
168. Kozak, M. (1991). "Structural features in eukaryotic mRNAs that modulate the initiation of translation." Journal of Biological Chemistry **266**(30): 19867-19870.
169. Kudla, G., A. W. Murray, D. Tollervey and J. B. Plotkin (2009). "Coding-Sequence Determinants of Gene Expression in Escherichia coli." Science **324**(5924): 255-258.
170. Kwon, Y. J., M. J. Park, S. G. Kim, I. T. Baldwin and C. M. Park (2014). "Alternative splicing and nonsense-mediated decay of circadian clock genes under environmental stress conditions in Arabidopsis." BMC Plant Biol **14**: 136.
171. Lageix, S., E. Lanet, M. N. Pouch-Pelissier, M. C. Espagnol, C. Robaglia, J. M. Deragon and T. Pelissier (2008). "Arabidopsis eIF2alpha kinase GCN2 is essential for growth in stress conditions and is activated by wounding." BMC Plant Biol **8**: 134.
172. Landry, L. G., C. C. S. Chapple and R. L. Last (1995). "Arabidopsis mutants lacking phenolic sunscreens exhibit enhanced ultraviolet-B injury and oxidative damage." Plant physiology **109**(4): 1159-1166.
173. Lareau, L. F., D. H. Hite, G. J. Hogan and P. O. Brown (2014). "Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments." eLife **3**: e01257.
174. Li, J., T.-M. Ou-Lee, R. Raba, R. G. Amundson and R. L. Last (1993). "Arabidopsis flavonoid mutants are hypersensitive to UV-B irradiation." The Plant Cell **5**(2): 171-179.
175. Liu, B., Y. Han and S.-B. Qian (2013). "Cotranslational Response to Proteotoxic Stress by Elongation Pausing of Ribosomes." Molecular Cell **49**(3): 453-463.
176. Liu, J., M. A. Carmell, F. V. Rivas, C. G. Marsden, J. M. Thomson, J.-J. Song, S. M. Hammond, L. Joshua-Tor and G. J. Hannon (2004). "Argonaute2 Is the Catalytic Engine of Mammalian RNAi." Science **305**(5689): 1437-1441.
177. Liu, M. J., S. H. Wu and H. M. Chen (2012). "Widespread translational control contributes to the regulation of Arabidopsis photomorphogenesis." Mol Syst Biol **8**: 566.
178. Liu, M. J., S. H. Wu, H. M. Chen and S. H. Wu (2012). Widespread translational control contributes to the regulation of Arabidopsis photomorphogenesis.
179. Lu, J. and C. Deutsch (2008). "Electrostatics in the Ribosomal Tunnel Modulate Chain Elongation Rates." Journal of Molecular Biology **384**(1): 73-86.
180. Lu, J., W. R. Kobertz and C. Deutsch (2007). "Mapping the Electrostatic Potential within the Ribosomal Exit Tunnel." Journal of Molecular Biology **371**(5): 1378-1391.
181. Macal, C. M. and M. J. North (2010). "Tutorial on agent-based modelling and simulation." Journal of simulation **4**(3): 151-162.
182. MacDonald, C. T. and J. H. Gibbs (1969). "Concerning the kinetics of polypeptide synthesis on polyribosomes." Biopolymers **7**(5): 707-725.
183. MacDonald, C. T., J. H. Gibbs and A. C. Pipkin (1968). "Kinetics of biopolymerization on nucleic acid templates." Biopolymers **6**(1): 1-25.
184. Man, O. and Y. Pilpel (2007). "Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species." Nature genetics **39**(3): 415-421.
185. Mandel, C. R., S. Kaneko, H. Zhang, D. Gebauer, V. Vethantham, J. L. Manley and L. Tong (2006). "Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end-processing endonuclease." Nature **444**(7121): 953-956.

186. Mangus, D. A., M. C. Evans and A. Jacobson (2003). "Poly (A)-binding proteins: multifunctional scaffolds for the post-transcriptional control of gene expression." Genome Biol **4**(7): 223.
187. Marcotrigiano, J., A.-C. Gingras, N. Sonenberg and S. K. Burley (1999). "Cap-dependent translation initiation in eukaryotes is regulated by a molecular mimic of eIF4G." Molecular cell **3**(6): 707-716.
188. Marintchev, A., K. A. Edmonds, B. Marintcheva, E. Hendrickson, M. Oberer, C. Suzuki, B. Herdy, N. Sonenberg and G. Wagner (2009). "Topology and Regulation of the Human eIF4A/4G/4H Helicase Complex in Translation Initiation." Cell **136**(3): 447-460.
189. Marthaler, D. (2013). An Overview of Mathematical Methods for Numerical Optimization. Numerical Methods for Metamaterial Design. K. Diest, Springer Netherlands. **127:** 31-53.
190. Más, P., W.-Y. Kim, D. E. Somers and S. A. Kay (2003). "Targeted degradation of TOC1 by ZTL modulates circadian function in Arabidopsis thaliana." Nature **426**(6966): 567-570.
191. Mathews, M., N. Sonenberg and J. Hershey (2007). Translational control in biology and medicine. Cold Spring Harbor, NY, Cold Spring Harbor Laboratory Press.
192. Matsushika, A., S. Makino, M. Kojima, T. Yamashino and T. Mizuno (2002). "The APRR1/TOC1 quintet implicated in circadian rhythms of Arabidopsis thaliana: II. Characterization with CCA1-overexpressing plants." Plant Cell Physiol **43**(1): 118-122.
193. Maunoury, N. and H. Vaucheret (2011). "AGO1 and AGO2 act redundantly in miR408-mediated Plantacyanin regulation." PLoS One **6**(12): e28729.
194. Menon, K. P., S. Sanyal, Y. Habara, R. Sanchez, R. P. Wharton, M. Ramaswami and K. Zinn (2004). "The translational repressor Pumilio regulates presynaptic morphology and controls postsynaptic accumulation of translation factor elF-4E." Neuron **44**(4): 663-676.
195. Mi, S., T. Cai, Y. Hu, Y. Chen, E. Hodges, F. Ni, L. Wu, S. Li, H. Zhou and C. Long (2008). "Sorting of small RNAs into Arabidopsis argonaute complexes is directed by the 5′ terminal nucleotide." Cell **133**(1): 116-127.
196. Michael, T. P., T. C. Mockler, G. Breton, C. McEntee, A. Byer, J. D. Trout, S. P. Hazen, R. Shen, H. D. Priest and C. M. Sullivan (2008). "Network discovery pipeline elucidates conserved time-of-day–specific cis-regulatory modules." PLoS Genet **4**(2): e14.
197. Miller, C., B. Schwalb, K. Maier, D. Schulz, S. Dümcke, B. Zacher, A. Mayer, J. Sydow, L. Marcinowski, L. Dölken, D. E. Martin, A. Tresch and P. Cramer (2011). "Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast." Molecular Systems Biology **7**(1).
198. Miller, M. A. and W. M. Olivas (2011). "Roles of Puf proteins in mRNA degradation and translation." Wiley Interdisciplinary Reviews: RNA **2**(4): 471-492.
199. Miller, M. A., J. Russo, A. D. Fischer, F. A. Lopez Leban and W. M. Olivas (2014). "Carbon source-dependent alteration of Puf3p activity mediates rapid changes in the stabilities of mRNAs involved in mitochondrial function." Nucleic Acids Research **42**(6): 3954-3970.
200. Missra, A., B. Ernest, T. Lohoff, Q. Jia, J. Satterlee, K. Ke and A. G. von Arnim (2015). "The Circadian Clock Modulates Global Daily Cycles of mRNA Ribosome Loading." The Plant Cell **27**(9): 2582-2599.

201. Mitarai, N., K. Sneppen and S. Pedersen (2008). "Ribosome Collisions and Translation Efficiency: Optimization by Codon Usage and mRNA Destabilization." Journal of Molecular Biology **382**(1): 236-245.
202. Mittag, M., D. H. Lee and J. W. Hastings (1994). "Circadian expression of the luciferin-binding protein correlates with the binding of a protein to the 3' untranslated region of its mRNA." Proc Natl Acad Sci U S A **91**(12): 5257-5261.
203. Moeller, J. R., M. J. Moscou, T. Bancroft, R. W. Skadsen, R. P. Wise and S. A. Whitham (2012). "Differential accumulation of host mRNAs on polyribosomes during obligate pathogen-plant interactions." Mol Biosyst **8**(8): 2153-2165.
204. Moeller, J. R., M. J. Moscou, T. Bancroft, R. W. Skadsen, R. P. Wise and S. A. Whitham (2012). "Differential accumulation of host mRNAs on polyribosomes during obligate pathogen-plant interactions." Molecular BioSystems **8**(8): 2153-2165.
205. Montgomery, T. A., M. D. Howell, J. T. Cuperus, D. Li, J. E. Hansen, A. L. Alexander, E. J. Chapman, N. Fahlgren, E. Allen and J. C. Carrington (2008). "Specificity of ARGONAUTE7-miR390 interaction and dual functionality in TAS3 trans-acting siRNA formation." Cell **133**(1): 128-141.
206. Moore, M. J. (2005). "From Birth to Death: The Complex Lives of Eukaryotic mRNAs." Science **309**(5740): 1514-1518.
207. Moore, M. J. and N. J. Proudfoot (2009). "Pre-mRNA Processing Reaches Back toTranscription and Ahead to Translation." Cell **136**(4): 688-700.
208. Moriyama, E. N. and J. R. Powell (1997). "Codon usage bias and tRNA abundance in Drosophila." Journal of molecular evolution **45**(5): 514-523.
209. Moriyama, E. N. and J. R. Powell (1998). "Gene length and codon usage bias in Drosophila melanogaster, Saccharomyces cerevisiae and Escherichia coli." Nucleic Acids Research **26**(13): 3188-3193.
210. Morris, D. R. and A. P. Geballe (2000). "Upstream Open Reading Frames as Regulators of mRNA Translation." Molecular and Cellular Biology **20**(23): 8635-8642.
211. Morse, D., P. M. Milos, E. Roux and J. W. Hastings (1989). "Circadian regulation of bioluminescence in Gonyaulax involves translational control." Proc Natl Acad Sci U S A **86**(1): 172-176.
212. Müllner, E. W. and L. C. Kühn (1988). "A stem-loop in the 3′ untranslated region mediates iron-dependent regulation of transferrin receptor mRNA stability in the cytoplasm." Cell **53**(5): 815-825.
213. Nagaraj, N., J. R. Wisniewski, T. Geiger, J. Cox, M. Kircher, J. Kelso, S. Pääbo and M. Mann (2011). "Deep proteome and transcriptome mapping of a human cancer cell line." Molecular systems biology **7**(1).
214. Nagel, D. H. and S. A. Kay (2012). "Complexity in the wiring and regulation of plant circadian networks." Curr Biol **22**: R648-657.
215. Nakashima, H., J. Perlman and J. F. Feldman (1981). "Cycloheximide-induced phase shifting of circadian clock of Neurospora." Am J Physiol **241**(1): R31-35.
216. Nakatogawa, H. and K. Ito (2002). "The Ribosomal Exit Tunnel Functions as a Discriminating Gate." Cell **108**(5): 629-636.
217. Narsai, R., K. A. Howell, A. H. Millar, N. O'Toole, I. Small and J. Whelan (2007). "Genome-Wide Analysis of mRNA Decay Rates and Their Determinants in Arabidopsis thaliana." The Plant Cell **19**(11): 3418-3436.

218. Nash, J. C. (1979). Compact numerical methods for computers: linear algebra and function minimisation, Bristol: A. Hilger.
219. Nash, J. C. (2014). Nonlinear Parameter Optimization Using R Tools, Wiley.
220. Nash, J. C. and R. Varadhan (2011). "Unifying Optimization Algorithms to Aid Software System Users: optimx for R." 2011 **43**(9): 14.
221. Nelder, J. A. and R. Mead (1965). "A Simplex Method for Function Minimization." The Computer Journal **7**(4): 308-313.
222. Nicolai, M., M. A. Roncato, A. S. Canoy, D. Rouquie, X. Sarda, G. Freyssinet and C. Robaglia (2006). "Large-scale analysis of mRNA translation states during sucrose starvation in arabidopsis cells identifies cell proliferation and chromatin structure as targets of translational control." Plant Physiol **141**(2): 663-673.
223. Nielsen, H. B. (2000). UCMINF - an Algorithm for Unconstrained, Nonlinear Optimization, Informatics and Mathematical Modelling (IMM), Technical University of Denmark.
224. Nielsen, H. B. and S. B. Mortensen (2009). "ucminf: General-purpose unconstrained non-linear optimization." URL http://CRAN. R-project. org/package= ucminf, R package version **1**.
225. Nocedal, J. and S. J. Wright (1999). Numerical Optimization, New York: Springer.
226. Nozue, K., M. F. Covington, P. D. Duek, S. Lorrain, C. Fankhauser, S. L. Harmer and J. N. Maloof (2007). "Rhythmic growth explained by coincidence between internal and external cues." Nature **448**(7151): 358-361.
227. Oh, E., Annemarie H. Becker, A. Sandikci, D. Huber, R. Chaba, F. Gloge, Robert J. Nichols, A. Typas, Carol A. Gross, G. Kramer, Jonathan S. Weissman and B. Bukau (2011). "Selective Ribosome Profiling Reveals the Cotranslational Chaperone Action of Trigger Factor In Vivo." Cell **147**(6): 1295-1308.
228. Olivas, W. and R. Parker (2000). "The Puf3 protein is a transcript-specific regulator of mRNA degradation in yeast." The EMBO Journal **19**(23): 6602-6611.
229. Ong, S.-E., B. Blagoev, I. Kratchmarova, D. B. Kristensen, H. Steen, A. Pandey and M. Mann (2002). "Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics." Molecular & cellular proteomics **1**(5): 376-386.
230. Pal, S. K., M. Liput, M. Piques, H. Ishihara, T. Obata, M. C. Martins, R. Sulpice, J. T. van Dongen, A. R. Fernie, U. P. Yadav, J. E. Lunn, B. Usadel and M. Stitt (2013). "Diurnal changes of polysome loading track sucrose content in the rosette of wild-type arabidopsis and the starchless pgm mutant." Plant Physiol **162**(3): 1246-1265.
231. Pal, S. K., M. Liput, M. Piques, H. Ishihara, T. Obata, M. C. M. Martins, R. Sulpice, J. T. van Dongen, A. R. Fernie, U. P. Yadav, J. E. Lunn, B. Usadel and M. Stitt (2013). "Diurnal Changes of Polysome Loading Track Sucrose Content in the Rosette of Wild-Type Arabidopsis and the Starchless pgm Mutant." Plant Physiology **162**(3): 1246-1265.
232. Park, M. J., P. J. Seo and C. M. Park (2012). "CCA1 alternative splicing as a way of linking the circadian clock to temperature response in Arabidopsis." Plant Signal Behav **7**(9): 1194-1196.
233. Passmore, L. A., T. M. Schmeing, D. Maag, D. J. Applefield, M. G. Acker, M. A. Algire, J. R. Lorsch and V. Ramakrishnan (2007). "The eukaryotic translation initiation factors

eIF1 and eIF1A induce an open conformation of the 40S ribosome." Molecular cell **26**(1): 41-50.

234. Pestova, T. V., S. I. Borukhov and C. U. T. Hellen (1998). "Eukaryotic ribosomes require initiation factors 1 and 1A to locate initiation codons." Nature **394**(6696): 854-859.
235. Pestova, T. V. and V. G. Kolupaeva (2002). "The roles of individual eukaryotic translation initiation factors in ribosomal scanning and initiation codon selection." Genes & development **16**(22): 2906-2922.
236. Pillai, R. S., S. N. Bhattacharyya, C. G. Artus, T. Zoller, N. Cougot, E. Basyuk, E. Bertrand and W. Filipowicz (2005). "Inhibition of Translational Initiation by Let-7 MicroRNA in Human Cells." Science **309**(5740): 1573-1576.
237. Piques, M., W. X. Schulze, M. Hohne, B. Usadel, Y. Gibon, J. Rohwer and M. Stitt (2009). "Ribosome and transcript copy numbers, polysome occupancy and enzyme dynamics in Arabidopsis." Mol Syst Biol **5**: 314.
238. Pisarev, A. V., V. G. Kolupaeva, V. P. Pisareva, W. C. Merrick, C. U. T. Hellen and T. V. Pestova (2006). "Specific functional interactions of nucleotides at key -3 and +4 positions flanking the initiation codon with components of the mammalian 48S translation initiation complex." Genes & Development **20**(5): 624-636.
239. Poker, G., Y. Zarai, M. Margaliot and T. Tuller (2014). "Maximizing protein translation rate in the non-homogeneous ribosome flow model: a convex optimization approach." Journal of The Royal Society Interface **11**(100).
240. Pokhilko, A., A. P. Fernandez, K. D. Edwards, M. M. Southern, K. J. Halliday and A. J. Millar (2012). "The clock gene circuit in Arabidopsis includes a repressilator with additional feedback loops." Mol Syst Biol **8**: 574.
241. Post, L. E. and M. Nomura (1980). "DNA sequences from the str operon of Escherichia coli." Journal of Biological Chemistry **255**(10): 4660-4666.
242. Post, L. E., G. D. Strycharz, M. Nomura, H. Lewis and P. P. Dennis (1979). "Nucleotide sequence of the ribosomal protein gene cluster adjacent to the gene for RNA polymerase subunit beta in Escherichia coli." Proceedings of the National Academy of Sciences **76**(4): 1697-1701.
243. Proudfoot, N. J., A. Furger and M. J. Dye (2002). "Integrating mRNA processing with transcription." Cell **108**(4): 501-512.
244. Ran, W. and P. G. Higgs (2010). "The Influence of Anticodon–Codon Interactions and Modified Bases on Codon Usage Bias in Bacteria." Molecular Biology and Evolution **27**(9): 2129-2140.
245. Reid, D. W. and C. V. Nicchitta (2012). "Primary role for endoplasmic reticulum-bound ribosomes in cellular translation identified by ribosome profiling." Journal of Biological Chemistry **287**(8): 5518-5527.
246. Reinhart, B. J., E. G. Weinstein, M. W. Rhoades, B. Bartel and D. P. Bartel (2002). "MicroRNAs in plants." Genes & Development **16**(13): 1616-1626.
247. Reis, M. d., R. Savva and L. Wernisch (2004). "Solving the riddle of codon usage preferences: a test for translational selection." Nucleic Acids Research **32**(17): 5036-5044.
248. Reuveni, S., I. Meilijson, M. Kupiec, E. Ruppin and T. Tuller (2011). "Genome-Scale Analysis of Translation Elongation with a Ribosome Flow Model." PLoS Comput Biol **7**(9): e1002127.

249. Rhoades, M. W., B. J. Reinhart, L. P. Lim, C. B. Burge, B. Bartel and D. P. Bartel (2002). "Prediction of Plant MicroRNA Targets." Cell **110**(4): 513-520.
250. Rikin, A., J. W. Dillwith and D. K. Bergman (1993). "Correlation between the circadian rhythm of resistance to extreme temperatures and changes in fatty acid composition in cotton seedlings." Plant physiology **101**(1): 31-36.
251. Robertson, F. C., A. W. Skeffington, M. J. Gardner and A. A. R. Webb (2009). "Interactions between circadian and hormonal signalling in plants." Plant molecular biology **69**(4): 419-427.
252. Robinson, M., R. Lilley, S. Little, J. S. Emtage, G. Yarranton, P. Stephens, A. Millican, M. Eaton and G. Humphreys (1984). "Codon usage can affect efficiency of translation of genes in Escherichia coli." Nucleic acids research **12**(17): 6663-6671.
253. Rodnina, M. V. (2013). "The ribosome as a versatile catalyst: reactions at the peptidyl transferase center." Current Opinion in Structural Biology **23**(4): 595-602.
254. Rodnina, M. V. and W. Wintermeyer (2001). "Fidelity of aminoacyl-tRNA selection on the ribosome: Kinetic and structural mechanisms." Annual Review of Biochemistry **70**: 415-435.
255. Rogers, K. and X. Chen (2013). "Biogenesis, Turnover, and Mode of Action of Plant MicroRNAs." The Plant Cell **25**(7): 2383-2399.
256. Ross, J. (1995). "mRNA stability in mammalian cells." Microbiological Reviews **59**(3): 423-450.
257. Roy, B. and A. G. von Arnim (2013). "Translational regulation of cytoplasmic mRNAs." The Arabidopsis Book **11**: e0165.
258. Russo, J. and W. M. Olivas (2015). "Conditional regulation of Puf1p, Puf4p, and Puf5p activity alters YHB1 mRNA stability for a rapid response to toxic nitric oxide stress in yeast." Molecular Biology of the Cell **26**(6): 1015-1029.
259. Sanchez, S. E., E. Petrillo, E. J. Beckwith, X. Zhang, M. L. Rugnone, C. E. Hernando, J. C. Cuevas, M. A. Godoy Herz, A. Depetris-Chauvin, C. G. Simpson, J. W. Brown, P. D. Cerdan, J. O. Borevitz, P. Mas, M. F. Ceriani, A. R. Kornblihtt and M. J. Yanovsky (2010). "A methyl transferase links the circadian clock to the regulation of alternative splicing." Nature **468**(7320): 112-116.
260. Sandberg, R., J. R. Neilson, A. Sarma, P. A. Sharp and C. B. Burge (2008). "Proliferating cells express mRNAs with shortened 3'untranslated regions and fewer microRNA target sites." Science **320**(5883): 1643-1647.
261. Schaffer, R., N. Ramsay, A. Samach, S. Corden, J. Putterill, I. A. Carre and G. Coupland (1998). "The late elongated hypocotyl mutation of Arabidopsis disrupts circadian rhythms and the photoperiodic control of flowering." Cell **93**(7): 1219-1229.
262. Schnabel, R. B., J. E. Koonatz and B. E. Weiss (1985). "A modular system of algorithms for unconstrained minimization." ACM Trans. Math. Softw. **11**(4): 419-440.
263. Schwanhausser, B., D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen and M. Selbach (2011). "Global quantification of mammalian gene expression control." Nature **473**(7347): 337-342.
264. Schwanhäusser, B., M. Gossen, G. Dittmar and M. Selbach (2009). "Global analysis of cellular protein translation by pulsed SILAC." Proteomics **9**(1): 205-209.

265. Scialdone, A., S. T. Mugford, D. Feike, A. Skeffington, P. Borrill, A. Graf, A. M. Smith and M. Howard (2013). "Arabidopsis plants perform arithmetic division to prevent starvation at night." Elife **2**: e00669.
266. Sehgal, A., A. Rothenfluh-Hilfiker, M. Hunter-Ensor, Y. Chen, M. P. Myers and M. W. Young (1995). "Rhythmic expression of timeless: a basis for promoting circadian cycles in period gene autoregulation." Science **270**(5237): 808-810.
267. Selbach, M., B. Schwanhausser, N. Thierfelder, Z. Fang, R. Khanin and N. Rajewsky (2008). "Widespread changes in protein synthesis induced by microRNAs." Nature **455**(7209): 58-63.
268. Seo, P. J., M. J. Park, M. H. Lim, S. G. Kim, M. Lee, I. T. Baldwin and C. M. Park (2012). "A self-regulatory circuit of CIRCADIAN CLOCK-ASSOCIATED1 underlies the circadian clock regulation of temperature responses in Arabidopsis." Plant Cell **24**(6): 2427-2442.
269. Shah, P., Y. Ding, M. Niemczyk, G. Kudla and Joshua B. Plotkin (2013). "Rate-Limiting Steps in Yeast Protein Translation." Cell **153**(7): 1589-1601.
270. Shah, P. and M. A. Gilchrist (2011). "Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift." Proceedings of the National Academy of Sciences **108**(25): 10231-10236.
271. Shalgi, R., Jessica A. Hurt, I. Krykbaeva, M. Taipale, S. Lindquist and Christopher B. Burge (2013). "Widespread Regulation of Translation by Elongation Pausing in Heat Shock." Molecular Cell **49**(3): 439-452.
272. Sharp, P. M. and W.-H. Li (1987). "The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications." Nucleic Acids Research **15**(3): 1281-1295.
273. Shaw, G. and R. Kamen (1986). "A conserved AU sequence from the 3′ untranslated region of GM-CSF mRNA mediates selective mRNA degradation." Cell **46**(5): 659-667.
274. Shaw, L. B., R. K. P. Zia and K. H. Lee (2003). "Totally asymmetric exclusion process with extended objects: A model for protein synthesis." Physical Review E **68**(2): 021910.
275. Shen, Y., G. Ji, B. J. Haas, X. Wu, J. Zheng, G. J. Reese and Q. Q. Li (2008). "Genome level analysis of rice mRNA 3′-end processing signals and alternative polyadenylation." Nucleic acids research **36**(9): 3150-3161.
276. Sheth, U. and R. Parker (2003). "Decapping and Decay of Messenger RNA Occur in Cytoplasmic Processing Bodies." Science **300**(5620): 805-808.
277. Shim, J. and M. Karin (2002). "The control of mRNA stability in response to extracellular stimuli." Molecules and cells **14**(3): 323-331.
278. Shin, B.-S., J.-R. Kim, S. E. Walker, J. Dong, J. R. Lorsch and T. E. Dever (2011). "Initiation factor eIF2γ promotes eIF2–GTP–Met-tRNAiMet ternary complex binding to the 40S ribosome." Nat Struct Mol Biol **18**(11): 1227-1234.
279. Shyu, A. B., J. G. Belasco and M. E. Greenberg (1991). "Two distinct destabilizing elements in the c-fos message trigger deadenylation as a first step in rapid mRNA decay." Genes & Development **5**(2): 221-231.
280. Singh, U. (1996). "Polyribosome dynamics: Size-distribution as a function of attachment, translocation and release of ribosomes." Journal of theoretical biology **179**(2): 147-159.
281. Spitzer, F. (1970). "Interaction of Markov processes." Advances in Mathematics **5**(2): 246-290.

282. Staiger, D. and R. Green (2011). "RNA-based regulation in the plant circadian clock." Trends Plant Sci **16**(10): 517-523.
283. Staiger, D., L. Zecca, D. A. Wieczorek Kirk, K. Apel and L. Eckstein (2003). "The circadian clock regulated RNA-binding protein AtGRP7 autoregulates its expression by influencing alternative splicing of its own pre-mRNA." Plant J **33**(2): 361-371.
284. Steitz, J. A. (1969). "Polypeptide Chain Initiation: Nucleotide Sequences of the Three Ribosomal Binding Sites in Bacteriophage R17 RNA." Nature **224**(5223): 957-964.
285. Strasser, B., M. Sánchez-Lamas, M. J. Yanovsky, J. J. Casal and P. D. Cerdán (2010). "Arabidopsis thaliana life without phytochromes." Proceedings of the National Academy of Sciences **107**(10): 4776-4781.
286. Strayer, C., T. Oyama, T. F. Schultz, R. Raman, D. E. Somers, P. Mas, S. Panda, J. A. Kreps and S. A. Kay (2000). "Cloning of the Arabidopsis clock gene TOC1, an autoregulatory response regulator homolog." Science **289**(5480): 768-771.
287. Sueoka, N. (1961). "Compositional Correlation between Deoxyribonucleic Acid and Protein." Cold Spring Harbor Symposia on Quantitative Biology **26**: 35-43.
288. Sun, Y., E. Atas, L. Lindqvist, N. Sonenberg, J. Pelletier and A. Meller (2012). "The eukaryotic initiation factor eIF4H facilitates loop-binding, repetitive RNA unwinding by the eIF4A DEAD-box helicase." Nucleic Acids Research.
289. Svitkin, Y. V., A. Pause, A. Haghighat, S. Pyronnet, G. Witherell, G. J. Belsham and N. Sonenberg (2001). "The requirement for eukaryotic initiation factor 4A (elF4A) in translation is in direct proportion to the degree of mRNA 5'secondary structure." Rna **7**(3): 382-394.
290. Takeda, A., S. Iwasaki, T. Watanabe, M. Utsumi and Y. Watanabe (2008). "The mechanism selecting the guide strand from small RNA duplexes is different among argonaute proteins." Plant and cell physiology **49**(4): 493-500.
291. Teixeira, D. and R. Parker (2007). "Analysis of P-Body Assembly in Saccharomyces cerevisiae." Molecular Biology of the Cell **18**(6): 2274-2287.
292. Teixeira, D., U. Sheth, M. Valencia-Sanchez, M. Brengues and R. Parker (2005). "Processing bodies require RNA for assembly and contain nontranslating mRNAs." RNA (New York, NY) **11**(4): 371.
293. Tenson, T. and M. Ehrenberg (2002). "Regulatory Nascent Peptides in the Ribosomal Tunnel." Cell **108**(5): 591-594.
294. Tian, B., J. Hu, H. Zhang and C. S. Lutz (2005). "A large-scale analysis of mRNA polyadenylation of human and mouse genes." Nucleic acids research **33**(1): 201-212.
295. Tiruneh, B., B.-H. Kim, D. Gallie, B. Roy and A. von Arnim (2013). "The global translation profile in a ribosomal protein mutant resembles that of an eIF3 mutant." BMC Biology **11**(1): 123.
296. Tiruneh, B. S., B.-H. Kim, D. R. Gallie, B. Roy and A. G. Von Arnim (2013). "The global translation profile in a ribosomal protein mutant resembles that of an eIF3 mutant." BMC biology **11**(1): 123.
297. Tiruneh, B. S., B. H. Kim, D. R. Gallie, B. Roy and A. G. von Arnim (2013). "The global translation profile in a ribosomal protein mutant resembles that of an eIF3 mutant." BMC Biol **11**(1): 123.

298. Tuller, T., A. Carmi, K. Vestsigian, S. Navon, Y. Dorfan, J. Zaborske, T. Pan, O. Dahan, I. Furman and Y. Pilpel (2010). "An Evolutionarily Conserved Mechanism for Controlling the Efficiency of Protein Translation." Cell **141**(2): 344-354.
299. Tuller, T., I. Veksler-Lublinsky, N. Gazit, M. Kupiec, E. Ruppin and M. Ziv-Ukelson (2011). "Composite effects of gene determinants on the translation speed and density of ribosomes." Genome Biology **12**(11): R110.
300. Tuller, T., Y. Y. Waldman, M. Kupiec and E. Ruppin (2010). "Translation efficiency is determined by both codon bias and folding energy." Proceedings of the National Academy of Sciences **107**(8): 3645-3650.
301. Tusher, V. G., R. Tibshirani and G. Chu (2001). "Significance analysis of microarrays applied to the ionizing radiation response." Proc Natl Acad Sci U S A **98**(9): 5116-5121.
302. Ulbricht, R. J. and W. M. Olivas (2008). "Puf1p acts in combination with other yeast Puf proteins to control mRNA stability." RNA **14**(2): 246-262.
303. Unbehaun, A., S. I. Borukhov, C. U. T. Hellen and T. V. Pestova (2004). "Release of initiation factors from 48S complexes during ribosomal subunit joining and the link between establishment of codon-anticodon base-pairing and hydrolysis of eIF2-bound GTP." Genes & Development **18**(24): 3078-3093.
304. Varadhan, R. and P. Gilbert (2009). "BB: An R Package for Solving a Large System of Nonlinear Equations and for Optimizing a High-Dimensional Nonlinear Objective Function." 2009 **32**(4): 26.
305. Varenne, S., J. Buc, R. Lloubes and C. Lazdunski (1984). "Translation is a non-uniform process: Effect of tRNA availability on the rate of elongation of nascent polypeptide chains." Journal of Molecular Biology **180**(3): 549-576.
306. Vassart, G., J. E. Dumont and F. R. L. Cantraine (1971). "Translational control of protein synthesis: A simulation study." Biochimica et Biophysica Acta (BBA) - Nucleic Acids and Protein Synthesis **247**(3): 471-485.
307. Vattem, K. M. and R. C. Wek (2004). "Reinitiation involving upstream ORFs regulates ATF4 mRNA translation in mammalian cells." Proceedings of the National Academy of Sciences of the United States of America **101**(31): 11269-11274.
308. Venter, G. (2010). Review of Optimization Techniques. Encyclopedia of Aerospace Engineering, John Wiley & Sons, Ltd.
309. Vogel, C. and E. M. Marcotte (2012). "Insights into the regulation of protein abundance from proteomic and transcriptomic analyses." Nat Rev Genet **13**(4): 227-232.
310. von der Haar, T. (2012). "Mathematical and computational modelling of ribosomal movement and protein synthesis: an overview." Computational and structural biotechnology journal **1**(1): 1-7.
311. Voss, N. R., M. Gerstein, T. A. Steitz and P. B. Moore (2006). "The Geometry of the Ribosomal Polypeptide Exit Tunnel." Journal of Molecular Biology **360**(4): 893-906.
312. Wachter, A., M. Tunc-Ozdemir, B. C. Grove, P. J. Green, D. K. Shintani and R. R. Breaker (2007). "Riboswitch control of gene expression in plants by splicing and alternative 3′ end processing of mRNAs." The Plant Cell **19**(11): 3437-3450.
313. Wang, Y., L. Opperman, M. Wickens and T. M. T. Hall (2009). "Structural basis for specific recognition of multiple mRNA targets by a PUF regulatory protein." Proceedings of the National Academy of Sciences **106**(48): 20186-20191.

314. Wang, Z. Y. and E. M. Tobin (1998). "Constitutive expression of the CIRCADIAN CLOCK ASSOCIATED 1 (CCA1) gene disrupts circadian rhythms and suppresses its own expression." Cell **93**(7): 1207-1217.
315. Warner, J. R. (1999). "The economics of ribosome biosynthesis in yeast." Trends Biochem Sci **24**(11): 437-440.
316. Wells, Jonathan N., L. T. Bergendahl and Joseph A. Marsh (2015). "Co-translational assembly of protein complexes." Biochemical Society Transactions **43**(6): 1221-1226.
317. Wells, S. E., P. E. Hillner, R. D. Vale and A. B. Sachs "Circularization of mRNA by Eukaryotic Translation Initiation Factors." Molecular Cell **2**(1): 135-140.
318. Wen, J.-D., L. Lancaster, C. Hodges, A.-C. Zeri, S. H. Yoshimura, H. F. Noller, C. Bustamante and I. Tinoco (2008). "Following translation by single ribosomes one codon at a time." Nature **452**(7187): 598-603.
319. Weston, S. (2013). "doMPI: Foreach parallel adaptor for the Rmpi package." R package version 0.2: 16.
320. Wickens, M., D. S. Bernstein, J. Kimble and R. Parker (2002). "A PUF family portrait: 3 ' UTR regulation as a way of life." Trends in Genetics **18**(3): 150-157.
321. Wickham, H., J. Hester and K. Müller (2016). "memoise: Memoisation of Functions. R package version 1.0.0.".
322. Wilson, T. and R. Treisman (1988). "Removal of poly (A) and consequent degradation of c-fos mRNA facilitated by 3′ AU-rich sequences."
323. Wilusz, C., M. Gao, C. Jones, J. Wilusz and S. Peltz (2001). "Poly (A)-binding proteins regulate both mRNA deadenylation and decapping in yeast cytoplasmic extracts." RNA **7**(10): 1416.
324. Wilusz, C. J., M. Wormington and S. W. Peltz (2001). "The cap-to-tail guide to mRNA turnover." Nat Rev Mol Cell Biol **2**(4): 237-246.
325. Wintermeyer, W., A. Savelsbergh, Y. P. Semenkov, V. I. Katunin and M. V. Rodnina (2001). "Mechanism of Elongation Factor G Function in tRNA Translocation on the Ribosome." Cold Spring Harbor Symposia on Quantitative Biology **66**: 449-458.
326. Wortman, J. R., B. J. Haas, L. I. Hannick, R. K. Smith, R. Maiti, C. M. Ronning, A. P. Chan, C. Yu, M. Ayele, C. A. Whitelaw, O. R. White and C. D. Town (2003). "Annotation of the Arabidopsis Genome." Plant Physiology **132**(2): 461-468.
327. Wu, L., J. Fan and J. G. Belasco (2006). "MicroRNAs direct rapid deadenylation of mRNA." Proceedings of the National Academy of Sciences of the United States of America **103**(11): 4034-4039.
328. Wu, X., M. Liu, B. Downie, C. Liang, G. Ji, Q. Q. Li and A. G. Hunt (2011). "Genome-wide landscape of polyadenylation in Arabidopsis provides evidence for extensive alternative polyadenylation." Proceedings of the National Academy of Sciences **108**(30): 12533-12538.
329. Xing, D. and Q. Q. Li (2011). "Alternative polyadenylation and gene expression regulation in plants." Wiley Interdisciplinary Reviews: RNA **2**(3): 445-458.
330. Yakir, E., D. Hilman, M. Hassidim and R. M. Green (2007). "CIRCADIAN CLOCK ASSOCIATED1 Transcript Stability and the Entrainment of the Circadian Clock in Arabidopsis." Plant Physiology **145**(3): 925-932.

331. Yang, E., E. van Nimwegen, M. Zavolan, N. Rajewsky, M. Schroeder, M. Magnasco and J. E. Darnell (2003). "Decay Rates of Human mRNAs: Correlation With Functional Characteristics and Sequence Attributes." Genome Research **13**(8): 1863-1872.
332. Yanguez, E., A. B. Castro-Sanz, N. Fernandez-Bautista, J. C. Oliveros and M. M. Castellano (2013). "Analysis of genome-wide changes in the translatome of Arabidopsis seedlings subjected to heat stress." PLoS One **8**(8): e71425.
333. Yángüez, E., A. B. Castro-Sanz, N. Fernández-Bautista, J. C. Oliveros and M. M. Castellano (2013). "Analysis of Genome-Wide Changes in the Translatome of Arabidopsis Seedlings Subjected to Heat Stress." PLoS ONE **8**(8): e71425.
334. You, T., G. M. Coghill and A. J. Brown (2010). "A quantitative model for mRNA translation in Saccharomyces cerevisiae." Yeast **27**(10): 785-800.
335. Yu, H. (2002). "Rmpi: parallel statistical computing in R." R News **2**(2): 10-14.
336. Yusupova, G. and M. Yusupov (2014). "High-resolution structure of the eukaryotic 80S ribosome." Annual review of biochemistry **83**: 467-486.
337. Zanetti, M. E., F. Chang, F. Gong, D. W. Galbraith and J. Bailey-Serres (2005). "Immunopurification of polyribosomal complexes of Arabidopsis for global analysis of gene expression." Plant Physiology **138**(2): 624-635.
338. Zarai, Y., M. Margaliot and T. Tuller (2014). "Maximizing protein translation rate in the ribosome flow model: the homogeneous case." IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) **11**(6): 1184-1195.
339. Zhang, G., I. Fedyunin, O. Miekley, A. Valleriani, A. Moura and Z. Ignatova (2010). "Global and local depletion of ternary complex limits translational elongation." Nucleic acids research **38**(14): 4778-4787.
340. Zhang, Y., Y. Wang, K. Kanyuka, M. A. Parry, S. J. Powers and N. G. Halford (2008). "GCN2-dependent phosphorylation of eukaryotic translation initiation factor-2alpha in Arabidopsis." J Exp Bot **59**(11): 3131-3141.
341. Zhu, H., F. Hu, R. Wang, X. Zhou, S.-H. Sze, L. W. Liou, A. Barefoot, M. Dickman and X. Zhang (2011). "Arabidopsis Argonaute10 specifically sequesters miR166/165 to regulate shoot apical meristem development." Cell **145**(2): 242-256.
342. Zia, R. K. P., J. J. Dong and B. Schmittmann (2011). "Modeling Translation in Protein Synthesis with TASEP: A Tutorial and Recent Developments." Journal of Statistical Physics **144**(2): 405-428.
343. Zimmermann, T. S., A. C. H. Lee, A. Akinc, B. Bramlage, D. Bumcrot, M. N. Fedoruk, J. Harborth, J. A. Heyes, L. B. Jeffs, M. John, A. D. Judge, K. Lam, K. McClintock, L. V. Nechev, L. R. Palmer, T. Racie, I. Röhl, S. Seiffert, S. Shanmugam, V. Sood, J. Soutschek, I. Toudjarska, A. J. Wheat, E. Yaworski, W. Zedalis, V. Koteliansky, M. Manoharan, H.-P. Vornlocher and I. MacLachlan (2006). "RNAi-mediated gene silencing in non-human primates." Nature **441**(7089): 111-114.
344. Zipser, D. and D. Perrin (1963). "Complementation on Ribosomes." Cold Spring Harbor Symposia on Quantitative Biology **28**: 533-537.
345. Zong, Q., M. Schummer, L. Hood and D. R. Morris (1999). "Messenger RNA translation state: The second dimension of high-throughput expression screening." Proceedings of the National Academy of Sciences **96**(19): 10632-10636.

346. Zouridis, H. and V. Hatzimanikatis (2007). "A Model for Protein Translation: Polysome Self-Organization Leads to Maximum Protein Synthesis Rates." Biophysical Journal **92**(3): 717-730.
347. Zubiaga, A. M., J. G. Belasco and M. E. Greenberg (1995). "The nonamer UUAUUUAUU is the key AU-rich sequence motif that mediates mRNA degradation." Molecular and Cellular Biology **15**(4): 2219-2230.
348. Zur, H. and T. Tuller (2013). "New Universal Rules of Eukaryotic Translation Initiation Fidelity." PLoS Comput Biol **9**(7): e1003136.

Vita

Ben Ernest was born in Winston-Salem, North Carolina on February 18th, 1985 to Mac and Sheila Ernest. He attended college at Wake Forest University where he earned a Bachelor of Science in Health and Exercise Science in 2007. While deciding on a career path, he took post-baccalaureate courses in biological sciences and stastistics at the University of North Carolina at Greensboro, worked in John Parks' lab at Wake Forest University School of Medicine, and taught tennis throughout Winston-Salem. Initially interested in genetics and biochemistry applied to nutrition and metabolism, he began graduate work in The University of Tennessee's Graduate School of Genome Science and Technology (GST), working for Brynn Voy. He later joined Albrecht von Arnim's lab for the remainder of his graduate work where he focused on applying computational approaches to study translation regulation in plants. He received a PhD in Life Sciences through the GST program in 2016 and will begin working as a post-doctoral fellow in August 2016 in Dan Camerini's lab at the National Institutes of Health in Bethesda, Maryland.