Quantitative genome-wide studies of RNA metabolism in yeast

Philipp Eser

München 2016

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

Quantitative genome-wide studies of RNA metabolism in yeast

Philipp Eser
aus München

2016

Erklärung:

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

Eidesstattliche Versicherung:

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

München,

_____

Philipp Eser

Dissertation eingereicht am 29.03.2016

1. Gutachter: Prof. Dr. Patrick Cramer

2. Gutachter: PD. Dr. Dietmar Martin

Mündliche Prüfung am 02.05.2016

Acknowledgments

Above all I want to thank my parents Sylvia and Frank who always provided me with lots of love, support and trust. I also want to thank my wife Claudia and my son Moritz who give meaning to my life and who are the sweetest motivation for proceeding in my professional career. Also I want to thank my older brother Stefan for lots of advice and discussions concerning research and important decisions in life.
I am truly grateful to Patrick Cramer for giving me the chance to do my PhD in his lab, for giving me two nice research projects and for being such a great mentor and supervisor. I was always motivated by his enthusiasm and positive spirit and appreciated the encouraging and supportive discussions. Likewise, I want to thank Julien and Achim who both made a tremendous effort in supervising data analysis and writing the publications. I am truly grateful for their ideas and help with mathematical and statistical methods and implementations. I want to especially thank Kerstin who generated all the high quality data that formed the basis for my work and all biological findings. I am also very grateful that I could work together and share an co-authorship with Leonhard. It was always great to work with him and I will miss having interesting on topic and off topic discussions. Likewise, I want to thank Carina for her great contribution to the publications and ongoing analyses. I also want to thank Michael for introducing me to the lab, sharing his office with me and being a great friend. There are many more people within the Gene Center that I want to thank, especially all current and former members of the Cramer, Gagneur and Tresch groups who created an incredible warm and scientific atmosphere.

Abstract

Gene expression and its regulation are fundamental processes in every living cell and organism. RNA molecules hereby play a central role by translating the genetic information into proteins, by regulating gene activity and by forming structural components. The kinetics of RNA metabolism differ widely between genes and conditions and play an important role for cellular processes, but how this is achieved remains poorly understood. Here, we used a novel experimental protocol that allows profiling of newly transcribed RNAs in conjunction with an advanced computational modeling pipeline to explore the kinetics of RNA metabolism and the underlying genetic determinants.

In the first study, we investigated cell cycle regulated gene expression and the contributions of synthesis and degradation to mRNA levels in *S.cerevisiae.* During the cell cycle, the levels of hundreds of mRNAs change in a periodic manner, but how this is carried out by alterations in the rates of mRNA synthesis and degradation has not been studied systematically. We were able to derive mRNA synthesis and degradation rates every 5 minutes during the cell cycle, and thus provide for the first time a high-resolution time series of RNA metabolism during the cell cycle. A novel statistical model identified 479 genes that show periodic changes in mRNA synthesis and generally also periodic changes in their mRNA degradation rates. Peaks of mRNA degradation follow peaks of mRNA synthesis, resulting in sharp and high peaks of mRNA levels at defined times during the cell cycle. Whereas the timing of mRNA synthesis is set by upstream DNA motifs and their associated transcription factors (TFs), the synthesis rate of a periodically expressed gene is apparently set by its core promoter.

In the second study, we developed metabolic labeling with RNA-Seq (4tU-Seq) and novel computational methods to gain further insights into the kinetics of RNA metabolism and its regulation. To decrypt the regulatory code of the genome, sequence elements must be defined that determine RNA turnover and thus gene expression. Here we attempt such decryption in an eukaryotic model organism, the fission yeast *S. pombe.* We first derived an improved genome annotation that redefines borders of 36% of expressed mRNAs and adds 487 non-coding RNAs (ncRNAs). We then combined RNA labeling *in-vivo* with mathematical modeling to obtain rates of RNA synthesis and degradation for 5,484 expressed RNAs and splicing rates for 4,958 introns. We identified functional sequence elements in DNA and RNA that control RNA metabolic rates, and quantified the contributions of individual nucleotides to RNA synthesis, splicing, and degradation. Our

approach reveals distinct kinetics of mRNA and ncRNA metabolism, separates antisense regulation by transcription interference from RNA interference, and provides a general tool for studying the regulatory code of genomes.

Publications

2016 **Determinants of RNA metabolism in the Schizosaccharomyces pombe genome**
Philipp Eser*, Leonhard Wachutka*, Kerstin C. Maier, Carina Demel, Mariana Boroni, Srignanakshi Iyer, Patrick Cramer, and Julien Gagneur
Molecular Systems Biology (2016) 12:857
*contributed equally

2015 **Quantitative genome-wide RNA kinetics in fission yeast**
Philipp Eser
Conference presentation, *Statistical Methods for Post Genomic Data,* Munich 2015.

2014 **Periodic mRNA synthesis and degradation co-operate during cell cycle gene expression**
Philipp Eser, Carina Demel, Kerstin C Maier, Björn Schwalb, Nicole Pirkl, Dietmar E Martin, Patrick Cramer and Achim Tresch
Molecular Systems Biology (2014) 10:717.

2014 **MoPS: Model-based Periodicity Screening.**
Eser P and Tresch A. Bioconductor R package (2014)
`https://www.bioconductor.org/packages/release/bioc/html/MoPS.html`

Contents

Part I.

Introduction

1. Gene expression and RNA life

1.1. Genome, transcriptome and proteome

The genetic information that is encoded in DNA contains the blueprint for all known proteins in unicellular and multicellular organisms. Proteins act as macromolecular machines that determine the cellular structure and carry out biochemical functions. The structure and function of proteins in turn is dictated by the nucleotide sequence of genes [1]. Genes are first transcribed into RNA molecules and then translated into amino acid sequences which are then folded to yield functional proteins. (Figure 1.1, [2]). The proteome differs widely between organisms and cell types and can be dynamically changed e.g. in different phases of the cell cycle or to cope with changing conditions like starvation or osmotic stress [3]. Protein levels and thus the composition of the proteome within a cell at any

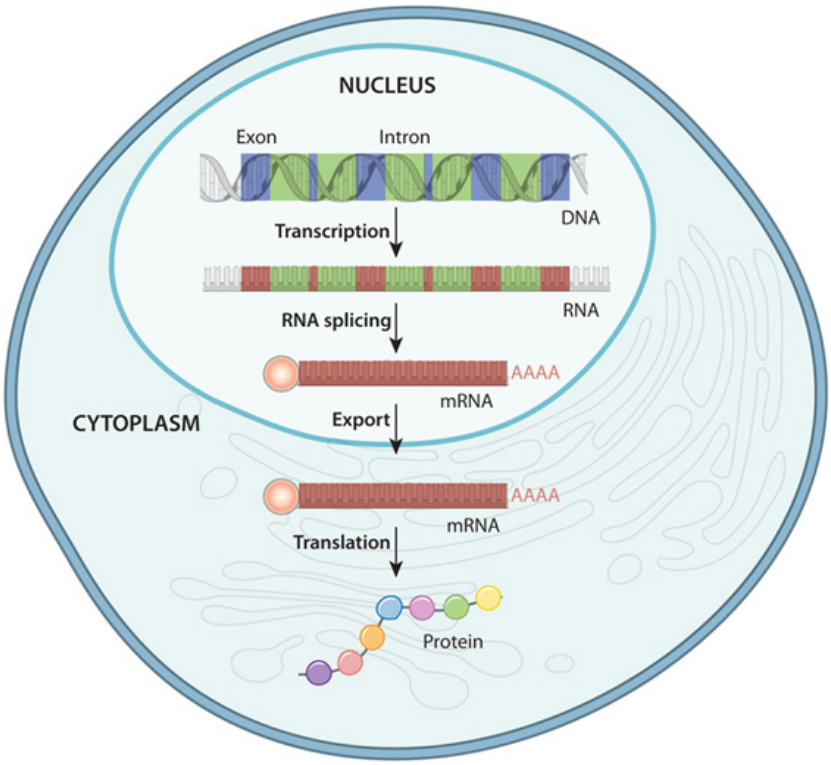

Figure 1.1.: Flow of information from DNA to protein. Taken from Scitable - Gene expression *(Nature Education*, 2010).

given time depend on the balance of protein production and degradation. The metabolism of proteins is heavily regulated at different levels including the rate of transcription, post-transcriptional processing and translation (reviewed in [4]). Regulation of gene expression and the amount of RNA molecules available to the translational machinery therefore take central roles for the biochemical integrity of cells. [5, 6].

1.2. RNA synthesis

How and when genes are switched on and off is determined by complex networks of protein-DNA interactions and the organization of chromatin [2]. Open chromatin allows the binding of proteins that recruit the core transcription machinery and co-factors like Mediator and SAGA at promoters of genes [7, 8]. Transcription initiation is further regulated by transcription factor binding sites upstream and within promoters that act as activators or repressors [9]. In the initial step of RNA synthesis, RNA polymerase escapes from promoter proximal regions and makes the transition to productive elongation of the RNA chain [10, 11]. After the synthesis of the complete transcript, polymerase is released from the DNA template and terminates transcription (Figure 1.2, reviewed in [12]). To orchestrate the transcription cycle, a diverse set of general transcription factors interacts with components of RNA polymerase at each stage of the cycle. Kinases and phosphatases act on the flexible C-terminal repeat domain (CTD) of polymerase to regulate transcription [13]. The phosphorylation state of CTD residues serves as a master regulator for the transitions from initiation to elongation and termination [14, 15]. Furthermore it couples transcription and RNA processing including efficient splicing and 3' processing [16, 17, 18]. Recent findings revealed a connection between transcription termination and RNA degradation, especially for the fate of non-coding RNAs [19, 20].
Transcription predominantly initiates at the promoters of protein-coding genes and produces sense transcripts. However, recent studies show that polymerases often transcribe in both directions resulting in bi-directional promoters and in the production of non-coding RNAs [21, 22]. These divergent and antisense non-coding transcripts as well as the transcriptional activity itself exhibit diverse functions in the regulation of gene expression. This includes the control of chromatin states by serving as scaffold for the chromatin-modifying machinery, gene silencing by transcriptional interference and small RNA mediated degradation of sense transcripts [23, 24, 25].

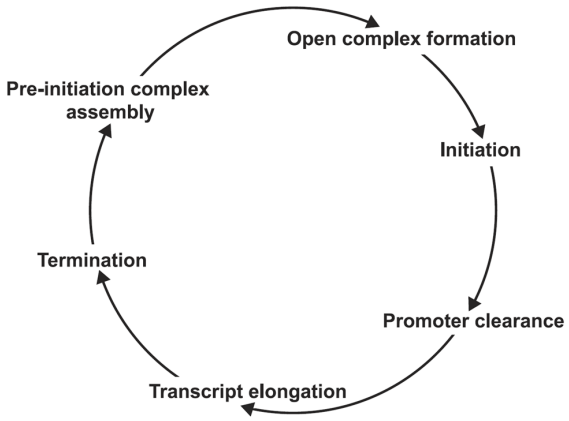

Figure 1.2.: The transcription cycle of RNA polymerases. Taken from [12].

1.3. RNA splicing

In eucaryotes, protein-coding genes contain intronic sequences that are not part of the mature mRNA. In the process of splicing, introns in the nascent pre-mRNA are excised and exons are joined together. For the majority of genes, this is performed by a highly conserved and flexible ribonucleoprotein complex, the spliceosome [26]. In a two-step biochemical reaction involving two transesterification reactions, the RNA components of the spliceosome interact with conserved sequence elements in the intronic RNA sequence to cleave the intron and ligate the adjacent exons (Figure 1.3, [27, 28]). These sequence elements comprise the core splicing signals that are required for spliceosome assembly: the 5' splice site, the 3' splice site and the branchpoint sequence [29, 30]. Outside of the core splice site motifs, the bulk of the information required for splicing is thought to be contained in exonic and intronic cis-regulatory elements that function by recruitment of sequence-specific RNA-binding protein factors that either activate or repress the use of adjacent splice sites [31]. This allows cells to make use of alternative splicing to expand the repertoire of mRNAs and thus the proteome (reviewed in [32]).
Splicing can either occur co-transcriptionally at nascent RNAs that are still bound to chromatin or post-transcriptionally at full length transcripts [33, 34]. Emerging evidence suggests that transcription and splicing are physically and functionally coupled [35]. In yeast, it has been shown that polymerases pause at the 3' end of introns in order to allow the splicing reaction to occur [36]. In human, antisense transcription was found to regulate alternative splicing of genes [37].

1.4. RNA degradation

Cells need to dynamically adjust their protein levels during proliferation, cell division or to cope with changed environmental conditions. This is achieved by the degradation of

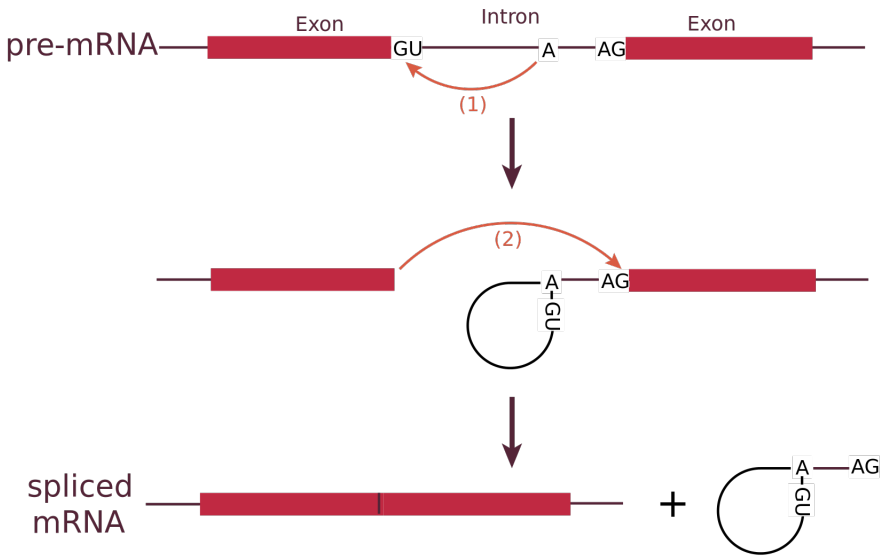

Figure 1.3.: Splicing reactions to remove introns from pre-mRNAs.

proteins and in addition, the reduction of mRNA levels by active degradation. Various tightly regulated RNA degradation pathways exist that further accomplish the renewal of mRNA pools, degrade aberrant or misfolded RNAs and control the life-time of non-coding and regulatory RNAs [38]. There are three major classes of RNA-degrading enzymes (RNases): endonucleases that cut RNA internally, 5' exonucleases that hydrolyze RNA from the 5' end, and 3' exonucleases that degrade RNA from the 3' end [39]. During transcription, the nascent RNA is first capped at the 5'end and then polyadenylated at the 3' end which prevents cytoplasmic degradation [40]. The initial step in mRNA degradation involves the removal of the poly-A tail. Deadenylated mRNAs can then either be 'decapped', a process that removes the 5' cap and thus allows the degradation by the Xrn1 exoribonuclease [41], or destroyed by the exosome that exhibits 3'-5' nuclease activity [42].

2. Transcriptome analysis

2.1. Profiling the transcriptome

Large-scale RNA quantification platforms allow the simultaneous measurement of transcripts expressed in cells. Two major techniques exist: hybridization-based (Microarrays) and sequencing-based (RNA-Seq). Microarrays contain thousands of different DNA sequence probes on their surface that are designed to be complementary to target gene sequences. Fluorescently labeled target sequences that bind to a probe sequence generate an optical signal [43]. The strength of this signal depends on the amount of target sample binding to the probe and hence can be used to estimate relative transcript abundances [44]. One major disadvantage of Mircroarrays is the limitation to the measurement of known transcripts. Sequencing based methods do not rely on specific probe-target matching and thus allow the identifcation and quantification of all expressed RNAs genome-wide [45, 46, 47]. Furthermore, RNA-Seq provides a higher dynamic range than Microarrays, meaning a higher resolution to detect lowly expressed transcripts as well as avoiding saturation effects with highly expressed transcripts [46, 48]. Strand-specific RNA-Seq allows the exact identification of transcript boundaries on each strand which led to the discovery of pervasive antisense transcription (reviewed in [49]). Studies using RNA-Seq revealed new classes of non-coding transcripts including functional long non-coding RNAs that exhibit cell-type and developmental time-point restricted expression patterns in mammalian genomes [50, 51].

Several experimental techniques that make use of RNA-Seq have been developed to investigate different aspects of transcription and genome regulation. Cap-analysis gene expression (CAGE) allows to isolate and sequence the initial bases at the 5' end of capped RNAs thereby permitting to map transcription initiation sites genome-wide [52]. Other methods have been developed to selectively sequence the 3' ends of untranslated regions of transcripts to study alternative poly(A) site usage [53, 54]. RNA-Seq is also used to systematically study protein-RNA interactions by sequencing and mapping of RNAs bound to specific RNA binding proteins [55].

Recently, new experimental methods have been developed that use custom RNA-Seq protocols together with mathematical modeling to quantify the contributions of RNA synthesis and degradation to cellular RNA levels (see section 2.2).

2.2. RNA metabolism kinetics

Gene expression can be regulated at each stage of RNA metabolism, during RNA synthesis, splicing, and degradation. The ratio between the rates of RNA synthesis and degradation determines steady-state levels of mature RNA, thereby controlling the amount of mRNA and the cellular concentration of ncRNAs. The rates of both RNA degradation and splicing contribute to the time required for reaching mature RNA steady-state levels following transcriptional responses [56, 57].

Several methods to estimate RNA turnover rates genome-wide have been presented including genomic run-on followed by RNA polymerase chromatin Immuno-precipitation [58], cytoplasmic sequestration of RNA polymerase [59], and metabolic RNA labeling [60, 61, 19, 62]. Metabolic RNA labeling allows the isolation of newly synthesized (labeled) transcripts with minimal perturbation. After separation of labeled from pre-existing RNAs, transcript abundances in each fraction can be quantified and synthesis and decay rates estimated [60, 63].

Quantifying the individual contributions of synthesis and degradation led to an improved understanding of how these processes are coordinated and how they control mRNA levels. The rates of RNA synthesis show large variation across genes and are the major determinants of constitutive and temporally or conditionally changing mRNA levels [64, 57, 5]. RNA degradation modulates and fine-tunes mRNA abundance, largely varies across conditions and between organisms, and can be dynamically changed to shape gene expression [65, 66, 67, 68]. In contrast to synthesis and degradation rates, accurate genome-wide kinetic parameters of splicing are still lacking, likely because sequencing depth is more limiting to get measurements of short-lived precursor RNAs. Nonetheless, recent studies in human [69] and mouse [61, 57] indicate that the rates of splicing also vary within a wide range.

3. Cell-cycle regulated gene expression

3.1. Periodic gene expression

The eukaryotic cell cycle is controlled by periodic gene expression. Gene expression changes during the cell cycle have been studied extensively in the budding yeast *S. cerevisiae* and in the fission yeast *Schizosaccharomyces pombe* (reviewed in [70]). These studies have revealed transcriptional regulatory proteins that drive cell cycle progression, their DNA-binding motifs, and their target genes [71, 72]. Parts of the regulatory networks that drive periodic gene expression could be reverse engineered [73, 74]. Cyclin-dependent kinases (CDKs) are pacemakers of the cell-cycle oscillator [75], although the sequential expression of TFs is sufficient to produce periodic expression for many cell cycle genes in the absence of mitotic cyclins [76]. A model suggesting the coupling of a TF network to CDK activity for robust oscillations in the cell cycle has been proposed [77].
The basis for these discoveries was laid by measurements of gene expression along the cell cycle, followed by identification and quantification of cell cycle regulated genes [78, 79, 80]. Different studies have identified diverse sets of 300-1500 genes that are periodically expressed [81, 82] (for a comprehensive overview of the results of different studies see the Cell Cycle database, [83]). The variation in the total number and the overlap of reported cell cycle genes arises from variation in experimental conditions like synchronization, strain, technological platform, and the type of computational analyses [81].

3.2. Computational identification of periodically expressed genes

There are two principal approaches to the computational identification of periodically expressed genes from time series measurements, non-parametric (model-free) approaches [80, 84] and parametric (model-based) methods [85, 86, 87]. Non-parametric methods do not assume a specific shape of a periodic time course, nor do they make particular assumptions on the distribution of the measurement errors. As such, they are inherently robust. However, they merely provide a measure for ranking genes according to their 'periodicity' without extracting information on the actual shape of the gene's time course.

Parametric methods explicitly infer the 'true' expression time course of a gene as a basis for a periodicity test. A proper modeling of the time course will not only increase the sensitivity of periodicity detection, it will provide valuable additional information for the grouping of periodically expressed genes. On the other hand, parametric models involve the risk of over-fitting, leading to a low specificity in the periodicity test. A careful choice of an appropriate model for periodic gene expression with a sparse parameter set is therefore essential [88]. A successful screening method needs to account for measurement noise and outliers, and ideally provides a smoothed, error-corrected estimate of the expression time course. Additionally, it has to account for the loss of synchronization of cells along the time course, which is caused by variability in progression through the cell cycle [81].

3.3. Cell cycle regulated mRNA synthesis and degradation

The regulation of mRNA levels not only involves changes in mRNA synthesis but also changes in mRNA degradation. Periodically expressed genes are enriched among genes that are subject to cytoplasmatic capping which might also contribute to controlling mRNA stability in the cell cycle [89]. Recently, long non-coding (lnc) RNAs have been found to modulate cell cycle transcription and post-transcriptional events by associating to the mRNA of cyclin-dependent kinases, thereby affecting their stability [90]. mRNA degradation is known to determine cellular mRNA equilibrium levels [65], and time-variable mRNA degradation can help in establishing a timely and precise adaption of mRNA levels [60, 61, 91]. Single-cell, single-molecule studies identified the mitotic genes CLB2 and SWI5 for which the process of periodic mRNA synthesis is corroborated by time-delayed periodic fluctuations in the degradation of their transcripts [92, 93]. Periodically expressed transcripts often encode proteins that are needed at a specific time of the cell cycle [94, 95]. Therefore any mechanism that sharpens the temporal profile of a periodically expressed mRNA is potentially beneficial. Despite these efforts, major questions concerning cell cycle gene expression remain. First, how do mRNA synthesis rates for periodically expressed genes change during the cell cycle? Second, what are the mechanistic determinants for the timing and magnitude of these synthesis rate changes? Third, do mRNA degradation rates also change during the cell cycle, and if so, how do these changes contribute to the observed changes in mRNA levels, i.e. transcript abundance?

4. Aims and scope of this thesis

All cellular biochemical processes in living organisms depend on the regulated expression of the genome. Genome-wide expression profiling by quantification of transcript abundance has thus become important standard in molecular biology research. Due to experimental limitations, gene expression studies have focused for decades solely on the quantification and comparison of steady-state RNA levels (total RNA). But changes in cellular transcript abundance originate from changes in RNA metabolism, specifically synthesis, processing and degradation rates, which total RNA measurements cannot resolve. To overcome this limitation, new experimental and computational methods are needed that allow the uncoupling of those processes.

The aim of this thesis was the development of computational methods and visualizations for the estimation and analysis of RNA metabolism from high resolution tanscriptomic datasets. The work covered here, builds on an established protocol to estimate RNA synthesis and degradation rates by metabolic labeling of nucleotides *in-vivo* and subsequent quantification of total and labeled RNAs with Microarrays (comparative Dynamic Transcriptome Analysis (cDTA), [67]).

In our first study, we applied cDTA to synchronized cells from *S.cerevisiae* allowing for the first time to monitor cell cycle regulated RNA metabolism in an eucaryote model organism. Software and mathematical models were developed to identify genes that are periodically expressed with high confidence and to estimate genome-wide mRNA synthesis and degradation rates. Cell cycle specific transcriptional regulators were identified and their contributions to target gene expression timing and level estimated. Finally, based on a novel dynamic model of regulated RNA degradation, we found evidence for a general destabilization mechanism that achieves high and sharp expression peaks of cell cycle genes.

In the second study, we aimed to improve the cDTA protocol by using RNA-Seq to quantify total and labeled RNA (4tU-Seq). In conjunction with the development of an advanced computational analysis pipeline, we make use of this high-resolution data for RNA turnover analyses. First, we established normalization and processing procedures for 4tU-Seq data and developed a computational approach to estimate robust synthesis, degradation and splicing rates. Next, we applied 4tU-Seq to wild type *S.pombe* cells in a time series with very short labeling times of only two minutes. We further revised the current *S.pombe* annotation resulting in the alteration of many transcript boundaries and the identification

of novel ncRNAs. Finally, the combination of the improved annotation and precise RNA turnover rate estimates enabled us to identify functional sequence elements in DNA and RNA that control RNA metabolic rates, and even quantify the contributions of individual nucleotides.

Part II.

Periodic mRNA synthesis and degradation cooperate during cell cycle gene expression

This work has been published in

Philipp Eser, Carina Demel, Kerstin C. Maier, Björn Schwalb, Nicole Pirkl, Dietmar E. Martin, Patrick Cramer and Achim Tresch. Periodic mRNA synthesis and degradation co-operate during cell cycle gene expression. *Molecular Systems Biology* (2014) 10: 717

5. Methods

5.1. cDTA of the yeast cell cycle

A BAR1 deletion strain was generated from WT strain BY4741 by replacing the BAR1 open reading frame from its start- to stop- codon with a KanMX module. BAR1 is a protease that cleaves and inactivates alpha factor and so recovers cells from alpha factor-induced cell cycle arrest. The Δbar1 strain was inoculated from a fresh overnight culture at OD600 0.1. At OD600 0.4 alpha factor was added at a final concentration of 600 ng/mL for 2 hours leading to cell cycle arrest in G1. Synchronization was followed visually by counting the number of budding cells under the microscope. Cells were centrifuged for 2 min at 1600 x g at 30°C and washed once with 3 x the original culture volume prewarmed YPD. Cells were then re suspended in the original culture volume with prewarmed YPD and released from alpha factor induced arrest. 41 consecutive samples were labeled for 5 min with 4-thiouracil every 5 min for 200 min. Labeling and sample processing was performed as described [67]. In particular, *S.pombe* mRNA spike-ins were used as an internal standard to estimate absolute abundance of *S.cerevisiae* mRNA levels in total and labeled data. Total RNA purification, separation of labeled RNA as well as sample hybridization and microarray scanning were carried out as previously described [67]. The quantification of labeled and total mRNA time courses was performed in two independent biological replicates.

5.2. Model-based screening for periodic fluctuations in time series (MoPS)

5.2.1. Description of the overall strategy

The MoPS algorithm is designed to recognize periodic behavior in a observation time series $g = (g(t_1), g(t_2), ..., g(t_K))$, having in mind the application to gene expression time series in our cell cycle data. We will use a likelihood ratio statistic to decide whether a time series displays periodic fluctuations or not. To that end, we will define a family of test functions $\mathcal{F}$, which consists of functions that we believe to exhaustively represent

time courses of periodically expressed genes. On the other hand, we will define a set of non-periodic test functions, $\overline{\mathcal{F}}$, that we believe to represent all typical time courses of genes that are not periodic, e.g. constant genes, or genes that show temporal drift (monotonically increasing/decreasing genes). Given a time course measurement g, and a continuous function f, let $L(f;\, g)$ denote the likelihood of f, given the observations on g. We determine the maximum likelihood fit $f_g \in \mathcal{F}$ respectively $\overline{f}_g \in \overline{\mathcal{F}}$ for the likelihood function specified in Section 5.2.2. Our test statistic, termed periodicity score, becomes

$$log \frac{L(f_g;\, g)}{L(\overline{f}_g;\, g)} \tag{5.1}$$

The larger the periodicity score, the more likely g shows periodic fluctuations.

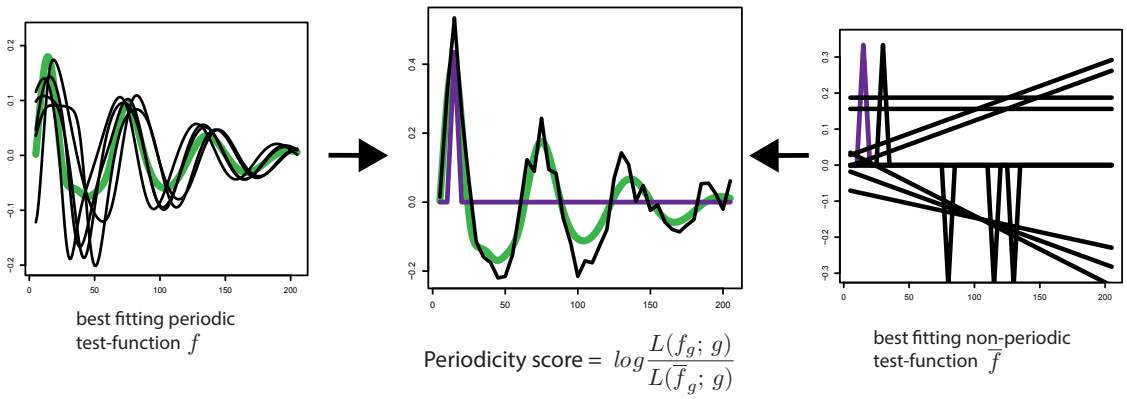

Figure 5.1.: MoPS periodic and non-periodic test-functions. Illustration of the statistical test used in MoPS to determine periodicity in time series data.

5.2.2. Preprocessing and Error Model

The raw total and labeled mRNA level measurements were corrected for 4-thiouridine labeling bias as described in [60]. The cDTA protocol uses spike-in control RNAs of *S.pombe* as an internal standard to normalize total mRNA arrays (resp. labeled mRNA arrays) between time points. We multiplicatively rescaled all total measurements such that the sum of all total gene expression levels at time zero equals $6 \cdot 10^4$, a recent estimate of the number of transcripts per *S.cerevisiae* cell [96]. The true ratio between the (mean) labeled expression measurements and the (mean) total expression measurements of all genes cannot be obtained from our measurements. This normalization factor was derived from the mean transcript half life in *S.cerevisiae* in wild type conditions, and it was chosen as in [67].

Erratic deviations in lowly abundant genes (whose measurements have a high coefficient of variation) might cause good periodic fits and hence false periodic gene calls if one

assume constant errors (constant variance of measurements) across the whole range of gene expression. We account for this by using a heteroscedastic error model. Let $g(t_k, i)$ denote the (normalized) measurement of gene g, $g \in G$, at time t_k, $k = 1, .., K$, in replicate $i \in I$. Let $g = (g(t_k, i);\ k = 1, ..., K,\ i \in I)$. The likelihood function $L(f;\ g)$ measures the goodness-of-fit by which a continuous function f approximates g at the measurement time points $t_1, ..., t_K$. Our likelihood function itself is standard, we assume independence of observations, and, as usual for gene expression measurements, Gaussian errors on the logarithmic values of g,

$$L(f;\ g) = \prod_{i \in I} \prod_{k=1}^{K} \frac{1}{\sqrt{2\pi\sigma_{g,t_k}}} \exp\left(-\frac{(\log g(t_k, i) - \log f(t_k))^2}{2\sigma^2_{g,t_k}} \right) \tag{5.2}$$

For each gene $g \in G$, we measured each time course (labeled or total RNA) in two replicates, namely $g = (g(t_k, i))$. Denote by Θ_g the full parameter set (specified in Section 5.6.2) which characterizes the approximation functions for g. Our target function is the negative log likelihood $l(\Theta_g;\ g)$,

$$l(\Theta_g;\ g) = \sum_{i \in I} \sum_{k=1}^{K} \frac{(\log\ g(t_k, i) - \log\ \hat{g}(t_k; \Theta_g))^2}{2 \cdot \sigma^2_{g,t_k}} \tag{5.3}$$

where $\hat{g}(t_k; \Theta_g)$ is the approximation function for g. Our loss function combines the idea of measuring similarity by correlation with the automatic penalization of genes whose seemingly periodic variation is in the range of their measurement error. Note that in Equation (5.3), σ^2_{g,t_k} is used to describe the variance for the total mRNA levels. These quantities still need to be defined. In our application, given merely 2 replicate measurements per gene and time point, we face the challenge that the number of observations is not sufficient to estimate the variances σ^2_{g,t_k} meaningfully from the 2 replicates alone. Therefore, we use a maximum-a-posteriori approach to regularize the gene-wise empirical variance by an estimate of the overall, intensity-dependent variance of a microarray. For the estimation of σ^2_{g,t_k}, we let $\log \bar{g}(t_k)$ be the mean of the replicates $\log g(t_k, i)$, $i \in I$. We assume that the replicate measurements $\log g(t_k, i)$ are i.i.d. samples from a Gaussian distribution,

$$\log g(t_k, i) \sim \mathcal{N}(\log g(t_k, i);\ \log \bar{g}(t_k), \sigma^2_{g,t_{k,}}) \ ,\ i \in I$$

For each time point, we calculate a global, intensity-dependent estimate of the variance by fitting a loess curve $m_{t_k}(.)$ to the point set $(\bar{g}(t_k), var(g(t_k, i); i \in I))$, $g \in G$. Here, $var(g(t_k, i); i \in I)$ denotes the empirical variance.

We assume a Gamma prior on σ^2_{g,t_k} , given by

$$\sigma_{g,t_k} \sim \Gamma\left(\sigma_{g,t_k},\, k = k(\bar{g}(t_k)), \theta = \theta(\bar{g}(t_k))\right) \propto \left(\sigma_{g,t_k}\right)^{k-1} \exp(-\frac{\sigma_{g,t_k}}{\theta}) \tag{5.4}$$

(where $\gamma(k)$ is the Gamma function). The shape parameter k and the scale parameter θ are chosen such that the expectation value of $\Gamma(\sigma;\, k, \theta)$ equals $m(\log \bar{g}(t_k))$, and its variance equals a parameter ν which is set to the mean of the squared residuals of the loess fit. This is achieved by letting

$$k = \frac{m(\log \bar{g}(t_k))^2}{\nu} \quad , \quad \theta = \frac{\nu}{m(\log \bar{g}(t_k))} \tag{5.5}$$

The regularized standard deviation is taken as the maximum a posteriori estimate

$$\sigma_{g,t_k}^{reg} = \arg\max_{\sigma_{g,t_k}} \left[\prod_{i \in I} \mathcal{N}(\log g(t_k, i);\, \log \bar{g}(t_k), \sigma^2_{g,t_{k,}}) \cdot \Gamma\left(\sigma_{g,t_k},\, k = k(\bar{g}(t_k)), \theta = \theta(\bar{g}(t_k))\right) \right]. \tag{5.6}$$

To safely guard against biases in the low intensity range, we additionally assume a minimum level for σ_{g,t_k}^{reg}, given by the 25% quantile of the respective residuals distribution.

5.2.3. Definition of periodic and non-periodic test functions

Definition of periodic test functions. Commonly, a gene g is called periodically expressed with period $\lambda' \in (0, \infty)$ and phase $\varphi \in [0, 2\pi]$ if its expression (in one cell) can be approximated, up to linear rescaling, by a cosine function

$$f(t) = cos(2\pi \cdot \frac{t}{\lambda'} - \varphi) \tag{5.7}$$

The phase φ describes the time at which g assumes its maximum expression divided by the cell cycle length; φ will therefore be also called the relative peak time. Accordingly, we call $\frac{\varphi}{2\pi} \cdot \lambda'$ the (absolute) peak time of g.

We wish to be less restrictive with respect to the shape of the periodic function. We use a slightly more general definition of a periodic gene. Let $\langle x \rangle$ the remainder x modulo 2π, i.e. the smallest non-negative number such that $x = \langle x \rangle + 2\pi z$ for some integer z. Let $\psi : [0, 2\pi] \to [0, 2\pi]$ be a monotonically increasing bijection of the unit interval. We consider a gene periodically expressed with period λ', phase φ and shape ψ if its expression

can be approximated, up to linear rescaling, by a function $f = f(t;\, \lambda', \varphi, \psi)$,

$$f(t;\, \lambda', \varphi, \psi) = cos(\psi \left\langle 2\pi \cdot \frac{t}{\lambda'} \right\rangle - \varphi) \tag{5.8}$$

Note that fixing ψ to the identity function yields the original notion of a periodic gene (Equation 5.7).

We are measuring the population average of a large number of cells. Not all cells proliferate at exactly the same speed. We assume that the cell cycle period length in the sample is not constant for individual cells in the sample, it is distributed according to a random variable $\lambda' = \lambda'(\lambda, \sigma)$ with mean period length λ and a standard deviation of σ. The measured expression of a periodic gene within our sample population will therefore resemble, up to linear rescaling, a function

$$\gamma(t;\, \lambda, \varphi, \psi, \sigma) \;\; = \;\; \int f(t;\, \lambda', \varphi, \psi)\, d\lambda'(\lambda, \sigma) \tag{5.9}$$

We finally arrive at the definition of the family of periodic test functions, given by

$$\begin{aligned} \mathcal{F} \;\; = \;\; \{\; & a \cdot \gamma(t;\, \varphi, \psi, \lambda, \sigma^2) + b \;| \\ & \varphi \in [0, 2\pi),\;\; \psi : [0, 2\pi] \to [0, 2\pi] \text{ a monotonically increasing bijection}, \\ & \lambda \in (0, \infty),\;\; \sigma^2 \in (0, \infty),\;\; a, b \in \mathbb{R}\} \end{aligned} \tag{5.10}$$

Choice of period length distribution. We tested three different classes of period length distributions for λ'. We scaled the parameters of the respective distributions such that they all have an expectation value of λ and a variance of σ^2. First, we chose a Gaussian distribution that has been cropped to the interval $[20, 200]$,

$$\lambda' \sim U_{[20,200]} * \mathcal{N}(\text{mean} = \lambda,\;\; \text{variance} = \sigma^2)\;, \tag{5.11}$$

The cropping was necessary to avoid negative cell cycle times. Secondly, we chose a log-normal distribution

$$\lambda' \sim \mathcal{LN}(\text{logmean} = ln\lambda - \tau/2,\;\; \text{logsigma} = \sqrt{\tau})\;, \tag{5.12}$$

with $\tau = ln(\sigma^2/\lambda^2 + 1)$ and thirdly, we selected a Gamma distribution

$$\lambda' \sim \text{Gamma}(\text{shape} = \frac{\lambda^2}{\sigma^2},\ \text{scale} = \frac{\sigma^2}{\lambda}) \tag{5.13}$$

It turns out that the the mean and standard deviation of the cell cycle length distribution λ' are enough to determine the dampening of the test function $\gamma(t;\, \lambda, \varphi, \psi, \sigma)$ up to irrelevant fluctuations. No matter which of the above distribution classes we chose, the results were almost identical, so we decided to use the log-normal distribution henceforth.

Definition of non-periodic test functions. Our goal is to discriminate periodic genes from non-periodic genes. To avoid false positive periodicity calls, the complementary set of non-periodic test functions should exhaustively cover time courses that a non-periodic gene can assume. Most often, a non-periodic gene has constant expression over time. Alternatively, due to continuous changes in the experimental conditions, non-periodic genes may show a constant drift, i.e., they are monotonically increasing or decreasing. There might also be genes that have one extraordinarily high / low peak at exactly one time point (in particular at $t = 0$). This might be due to a failure of the measurement, or due to synchronization at the beginning of the time course. We therefore define a family of non-periodic prototype test functions, consisting of the constant null function τ^0, a linearly increasing function τ^+, a linearly decreasing function τ^-, and the delta functions $\delta_k^+(t) = \begin{cases} 1 & \text{if } t = t_k \\ 0 & \text{else} \end{cases}$ and $\delta_k^-(t) = \begin{cases} -1 & \text{if } t = t_k \\ 0 & \text{else} \end{cases}$, $k = 1, ..., K$. We define the family $\overline{\mathcal{F}}$ of non-periodic functions as the set of all affine-linear transforms of the prototype test functions.

5.2.4. Parametrization of and screening for periodic genes

Maximum likelihood estimation in $\mathcal{F}$. The infinite family of periodic test functions $\mathcal{F}$ is parameterized by the tuple $(a, b, \lambda, \varphi, \psi, \sigma^2)$ (Equation (5.10)). Given a time series g, our task is to find the maximum likelihood estimate $f_g = argmin_{\gamma \in \mathcal{F}} l(g, \gamma)$. To that end, we construct a finite set of "prototype" functions $\mathcal{G}$, whose affine hull $\bar{\mathcal{G}} = \{a\gamma + b \mid \gamma \in \mathcal{G},\ a, b \in \mathbb{R}\}$ is assumed to lie sufficiently dense in $\mathcal{F}$. An approximation of the maximum likelihood estimate in $\mathcal{F}$ is then given by

$$\begin{aligned} f_g &= argmin_{\gamma \in \mathcal{F}} l(\gamma;\, g) \ \approx\ argmin_{\gamma \in \bar{\mathcal{G}}} l(\gamma;\, g) \\ &= argmin_{\gamma \in \mathcal{G}} \Big[argmin_{(a,b)} l(a\gamma + b;\, g) \Big]\, l_2 = \arg\max \end{aligned} \tag{5.14}$$

The minimization problem for (a, b), given γ, can be solved analytically by a weighted

linear regression, using the error model in (Equation (5.3)):

$$g(t_k) \sim \gamma(t_k) \quad , \text{ with weights } \sigma_{g,k}^{-2} \ , \ k = 1, ..., K \tag{5.15}$$

The slope of the regression line determines a, and the intercept determines b. The minimization over $\gamma \in \mathcal{G}$ is done by exhaustive search. The set $\mathcal{G}$ is defined as $\mathcal{G} = \{\psi(t;\, \lambda, \varphi, \psi, \sigma^2) \mid \lambda, \varphi, \psi, \sigma^2$ taken independently from a representative grid$\}$. The grid $\mathcal{G}_\lambda$ for λ runs from 30min to 90min in steps of 2.5min. The grid $\mathcal{G}_{\sigma^2}$ for the variance σ^2 runs from 1min^2 to 15min^2 by steps of 1min^2, and the grid $\mathcal{G}_\varphi$ for the peak time φ runs from 0 to $2\pi \cdot \frac{39}{4}$ in 40 equidistant steps. ψ runs through a representative set of piece-wise linear, monotonically increasing functions that are parameterized by a vector $y = (y_1, ..., y_{r-1})$ in the following way: Let $r = 4$, and let $t_j = j/r$, $j = 0, ..., r$. Define $\psi(t;\, y)$ as the piece-wise linear function which linearly interpolates the points (t_j, y_j), $j = 0, ..., r$ (set $(t_0, y_0) = (0, 0)$ and $(t_r, y_r) = (1, 1)$). Formally,

$$\psi(t;\, y) = r \cdot [(t - t_{j-1}) \cdot y_{j-1} + (t_j - t) \cdot y_j] \quad \text{if } t \in [t_{j-1}, t_j]$$

The values $y_1, ..., y_{r-1}$ are chosen from a finite grid

$$\mathcal{G}_\psi = \{(y_1, ..., y_{r-1}) \mid y_j \in \{\frac{0}{d}, \frac{1}{d}, ..., \frac{d}{d}\};\, y_1 \le y_2 \le ... \le y_{r-1}\} \ ,$$

for a given grid density d (we chose $d = 5$). In this way, a function $f \in \mathcal{F}$ is completely characterized by the tuple

$$(\varphi, y = (y_1, ..., y_{r-1}), \lambda, \sigma^2, a, b) \in \mathcal{G} = \mathcal{G}_\varphi \times \mathcal{G}_\psi \times \mathcal{G}_\lambda \times \mathcal{G}_{\sigma^2} \times \mathbb{R} \times \mathbb{R}$$

With our choice of $r = 4$, these are in total 7 parameters.

Note however, that the parameters are redundant: Assume that r is an even number. Let $\psi = \psi(.;\, (0, y, 1))$ be parameterized by a vector $y \in \mathcal{G}_\psi$ as described above. Define $y' = (y'_1, ..., y'_{r-1})$ by $y'_j = \begin{cases} y_{j+r/2} - y_{r/2} & \text{if } j = 1, ..., \frac{r}{2} - 1 \\ y_{j-r/s} - y_{r/2} + 1 & \text{if } j = \frac{r}{2}, ..., r - 1 \end{cases}$ (check that $y' \in \mathcal{G}_\psi$). By elementary calculations, it can be shown that

$$f(t;\, \lambda, \varphi, \psi(.;\, (0, y, 1))) = -f(t + \frac{\lambda}{2};\, \lambda, \varphi + \pi, \psi(.;\, (0, y', 1))$$

I.e., a phase shift by $\frac{\lambda}{2}$ can be described by a re-parametrization of the shape parameters, and by switching the sign. In other words, the parameter tuples $(\varphi, y, \lambda, \sigma^2, a, b)$ and $(\varphi + \pi, y', \lambda, \sigma^2, -a, b)$ describe identical test functions. Therefore, we only need to screen for φ values between 0 and π. In order to assign the correct peak time afterward, we simply need to check the sign of a in the linear regression. If $a \ge 0$, we keep the parameter

set $(\varphi, y, \lambda, \sigma^2, a, b)$. If a is negative, the corresponding set $(\varphi + \pi, y', \lambda, \sigma^2, -a, b)$ is the one with the correct peak time.

Maximum likelihood estimation in $\bar{\mathcal{F}}$. Since $\bar{\mathcal{F}}$ is the affine hull of a finite set of prototype functions, we proceed as in (Equation (5.14)). The maximum likelihood estimate in $\bar{\mathcal{F}}$ can be calculated exactly as

$$\begin{aligned} \bar{f}_g &= argmin_{\gamma \in \bar{\mathcal{F}}} l(\gamma;\, g) T \\ &= argmin_{\gamma \in \text{prototypes}} \left[argmin_{(a,b)} l(a\gamma + b;\, g) \right] \end{aligned} \tag{5.16}$$

Initial screen for periodic genes. The set $\tilde{P}$ of periodic genes is defined as the set of genes g for which the (log) likelihood ratio statistic $log \frac{L(f_g; g)}{L(\bar{f}_g; g)}$ exceeds some threshold value t_{min}. Genes that are not in $\tilde{P}$ are considered non-periodic. Assuming a fraction of 1/10 of periodic genes among all genes, t_{min} is determined by requiring that our criterion for periodicity have a false discovery rate of $\alpha = 0.05$). In this way, we identify genes that are periodically expressed with high confidence and can estimate the mean cell-cycle length and variation for each time series.

Refined screening to determine gene-specific parameters.

We estimate two gene-independent parameters, the mean cell cycle length λ and the variance σ^2 of the cell cycle length distribution in the initial screen from high-confidence periodic genes. For fixed λ, σ^2, let $h_{g,\lambda,\sigma^2} = \hat{a}_g \cdot \gamma(t;\, \lambda, \hat{\varphi}_g, \hat{\psi}_g, \sigma^2) + \hat{b}_g$ be the maximum likelihood approximation of g under the constraint that $\lambda_g = \lambda$ and $\sigma_g^2 = \sigma^2$, i.e.,

$$(\hat{a}_g, \hat{b}_g, \hat{\varphi}_g, \hat{\psi}_g) = argmin_{(a_g, b_g, \varphi_g, \psi_g)} l(a_g \cdot \gamma(t;\, \lambda, \varphi_g, \psi_g, \sigma^2) + b_g;\, g) \tag{5.17}$$

To determine the most likely global parameters $\hat{\lambda}$, $\hat{\sigma}^2$, we solve

$$(\hat{\lambda}, \hat{\sigma}^2) = argmin_{(\lambda, \sigma^2)} \sum_{g \in \tilde{P}} l(h_{g,\lambda,\sigma^2};\, g) \tag{5.18}$$

Note that the sum in Equation (5.18) is taken only over the initially defined periodic genes, because these are the candidates that are informative for the estimation of the global cell cycle parameters.

A gene g is called periodic, if

$$log \frac{L(h_{g,\hat{\lambda},\hat{\sigma}^2};\, g)}{L(\bar{f}_g;\, g)} > t_{min} \tag{5.19}$$

The set of all periodic genes is denoted by P. Genes that are defined as significant periodic

in the refined screen are implicitly also significant periodic in the initial screen ($\tilde{P} \subseteq P$), since $L(h_g;\, g) \geq L(h_{g,\hat{\lambda},\hat{\sigma}^2};\, g)$.

Accounting for replicate experiments. Our cell cycle experiment was done in two biological replicate time series, we do not only have one, but two time series, $(g^{(r)}(t_k))_{k=1,\ldots,K}$, $r \in R = \{1, 2\}$, for each gene g . We noticed that there are slight differences in the cell cycle length, thus we estimate two global parameter sets $\lambda^{(r)}$, $(\sigma^{(r)})^2$, $r = 1, 2$. The gene-specific parameters $a_g^{(r)}$, $b_g^{(r)}$ are estimated separately for each experiment, because it is not unlikely that there are slight differences in the magnitude of regulation due to slightly changed environmental conditions. The parameters $\hat{\varphi}_g$, $\hat{\psi}_g$ that determine the shape of the test function however are assumed to be common to all replicates. Finally, our screening procedure for periodic genes can be stated:

Definition of the Periodicity score and screen for periodic genes.

- Input: Expression time series measurements $g^{(r)} = (g^{(r)}(t_k))_{k=1,\ldots,K}$, $g \in G$, $r \in R$.
- For $g \in G$, $r \in R$, estimate the maximum likelihood fit of $g^{(r)}$ in $\mathcal{F}$ resp. $\overline{\mathcal{F}}$,

$$h_g^{(r)} = argmin_{\gamma \in \overline{\mathcal{G}}} l(\gamma;\, g^{(r)}) \in \mathcal{F} \quad , \quad \bar{f}_g^{(r)} = argmin_{\gamma \in \overline{\mathcal{F}}} l(\gamma,\, g^{(r)}) \in \overline{\mathcal{F}}$$

- Initial screen: determine a threshold t_{min}, and find all periodic genes $\tilde{P}^{(r)}$ in replicate $r \in R$,

$$\tilde{P}^{(r)} = \{g \in G \mid log \frac{L(h_g^{(r)};\, g^{(r)})}{L(\bar{f}_g^{(r)};\, g^{(r)})} > t_{min}\}$$

- For each replicate $r \in R$, calculate the global parameters $\hat{\lambda}^{(r)}$, $(\hat{\sigma}^{(r)})^2$ by

$$(\hat{\lambda}^{(r)}, (\hat{\sigma}^{(r)})^2) = argmin_{(\lambda to identify, \sigma^2)} \sum_{g \in \tilde{P}^{(r)}} l(h_{g,\lambda,\sigma^2}^{(r)},\, g^{(r)}) \;, \tag{5.20}$$

  where $h_{g,\lambda,\sigma^2}^{(r)}$ is the maximum likelihood approximation of $g^{(r)}$ in $\overline{\mathcal{G}}$ under the constraints $\lambda_g^{(r)} = \lambda$, $(\sigma^{(r)})^2 = (\sigma^{(r)})^2$ (see Equation (5.17)).

- Refined screening: for each gene $g \in G$, calculate $\hat{\varphi}_g$, $\hat{\psi}_g$, $a_g^{(r)}$, $b_g^{(r)}$, $r \in R$ by

$$(\hat{\varphi}_g, \hat{\psi}_g, a_g^{(r)}, b_g^{(r)};\, r \in R) = argmin_{(\varphi_g, \psi_g, a_g^{(r)}, b_g^{(r)};\, r \in R)} \sum_{r \in R} l(a_g^{(r)} \cdot \gamma(t;\, \hat{\lambda}^{(r)}, \varphi_g, \psi_g, \hat{\sigma}^2) + b_g^{(r)};\, g^{(r)}) \tag{5.21}$$

  Let $f_g^{(r)} = \hat{a}_g^{(r)} \cdot \gamma(t;\, \hat{\lambda}^{(r)}, \hat{\varphi}_g, \hat{\psi}_g, (\hat{\sigma}^{(r)})^2) + \hat{b}_g^{(r)}$.

- Define the periodicity score $T(g)$ as

$$T(g) = \sum_{r \in R} log \frac{L(f_g^{(r)};\, g^{(r)})}{L(\bar{f}_g^{(r)};\, g^{(r)})} \tag{5.22}$$

- Define the set P of periodic genes, $P = \{g \in G \mid T(g) > |R| \cdot t_{min}\}$.
- Output: The set P of periodic genes, and a set of parameters $\{\hat{\lambda}^{(r)}, (\hat{\sigma}^{(r)})^2, \hat{\varphi}_g, \hat{\psi}_g, a_g^{(r)}, b_g^{(r)};\, g \in G,\, r \in R\}$.

5.3. Significance of MoPS periodicity scores

MoPS computes a periodicity score for each gene and thus allows ranking of all genes according to their likelihood ratio to be periodically expressed respectively constantly expressed. However, there is no obvious way to assign significance to this score. We want to make use of existing knowledge derived from published studies about periodically expressed genes. To do this, we define a positive set and a negative set. The positive set comprises the top 200 periodic genes from Cyclebase [83] and the negative set consists of genes that have never been classified as cell-cycle regulated in any cell-cycle expression study considered [80, 82, 71]. The empirical distribution f of all MoPS scores is fitted by a mixture of the empirical distributions $f_+$and f_-scores of the positive respectively the negative set,

$$f \approx \mu \cdot f_+ + (1 - \mu) \cdot f_- \quad ,$$

where the mixture coefficient $\mu \in [0, 1]$ estimates the fraction of periodic genes among all genes. Fitting of μ was done be minimization of the Kolmogoroff-Smirnov statistic. μ, f_+ and f_- were then used to calculate the false discovery rate $FDR(c)$ as a function of the cutoff value c by

$$FDR(c) = \frac{(1 - \mu) \cdot \int_c^\infty f_-(t)dt}{\int_c^\infty f(t)dt}$$

5.4. Estimation of global and gene-specific parameters

In an initial screen we fit periodic test-functions that represent different combinations of cell-cycle length (λ), cell-cycle length variation in the population (σ) and phase (φ) to each expression profile. Using a strict periodicity score cutoff (FDR $< 5\%$, scores with best fitting λ, σ for each gene) this results in a set of periodic genes for each dataset with associated loss for all examined λ, σ combinations. The globally best fitting λ and σ are

then estimated by minimizing the overall loss for each combination over all genes. The distribution of estimated gene-specific λ and σ agree well within each dataset and between datasets. λ values range from 55 to 65 minutes and σ are mostly estimated to be in the range of 4 to 8 minutes.

A second screening is then performed using the dataset-specific global parameters λ and σ together with a refined set of periodic test-functions which are constructed from a exhaustive combination of the gene-specific parameters. Expression time courses of all genes are fitted to those periodic test-functions, separately for each dataset. This refines the initial screening by estimating gene-specific characteristic parameters. The derived characteristic expression time courses, the timing of peak expression and periodicity score are highly correlated between replicates. Since the periodicity scores of labeled and total datasets are highly similar for genes with positive scores, we averaged the scores and estimated one cutoff to obtain one set of cell-cycle regulated genes. Controlling the false discovery rate at 20%, we derive a cutoff of 0.78, which results in 479 significantly cell-cycle regulated genes (see Figure 5.2 for examples). For each gene, the best fitting 1 min resolution characteristic time course and its peak timing are averaged in total and labeled replicate time series.

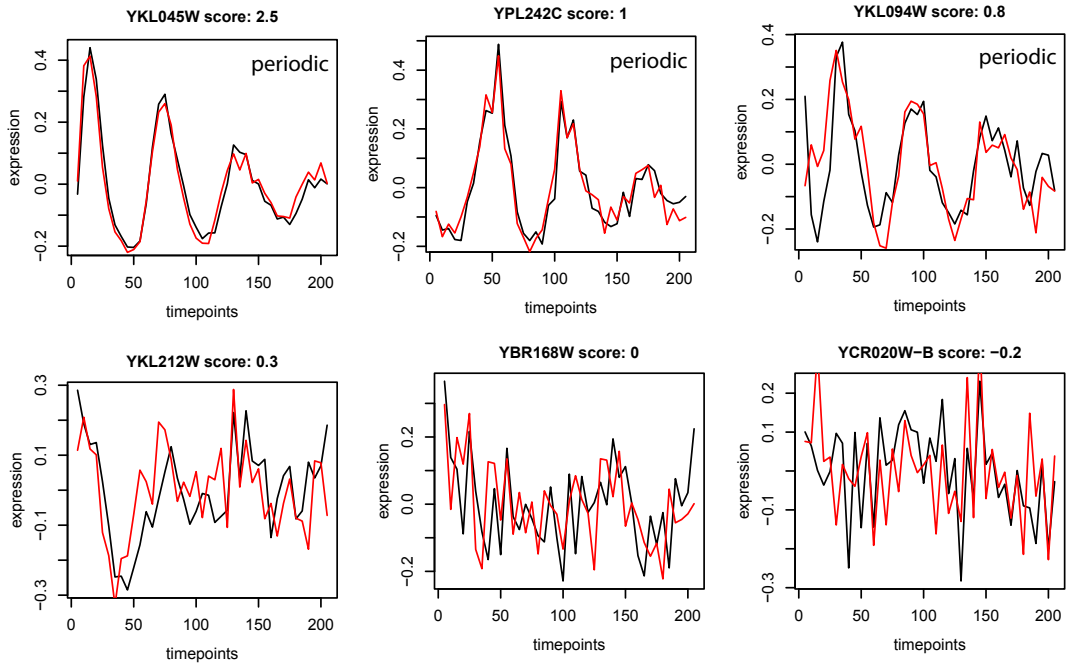

Figure 5.2.: Examples of six genes with various periodicity scores estimated with MoPS. Shown are total (black) and labeled (red) expression time courses (replicates averaged). All genes with a score above 0.78 are deemed periodic and considered for downstream analyses.

5.5. Motif search, association of TFs to periodic transcripts

Genes were grouped with k-means clustering (k = 10) according to their modeled 1 min resolution labeled expression time courses. Sequences 500 bases upstream of the respective transcription start site (SGD project, www.yeastgenome.org/download-data/sequence, Genome Release 64-1-1) were used as input for XXmotif [97] for each cluster. XXmotif was used with standard parameters, medium threshold for merging of similar motifs and set to report motifs that can occur multiple times per sequence. Motifs with an E-value higher than one were discarded. The positional weight matrices (PWMs) derived from ChIP-chip data [98] and the software TOMTOM (with standard parameters and Pearson correlation as comparison function) were used to assign the XXmotif found motifs to significantly similar, known TF-associated motifs (E-value < 1). Subsets of the 479 periodic genes are formed by using ChIP-chip derived associations (p-value < 0.01) of TFs and their targets [98]: genes that are regulated by a common set of cell cycle transcription factors (of 32 TFs identified in our TF screen).

5.6. Dynamic RNA turnover model and screen for periodic fluctuations in RNA degradation

5.6.1. A model for mRNA synthesis and degradation

Let $T(t)$ respectively $L(t)$ denote the time-dependent total respectively labeled mRNA amount of a certain transcript at time t. We assume that the mRNA population of a gene is synthesized with a time-dependent synthesis rate $\mu = \mu(t)$. We further assume that mRNA decays exponentially at a time-dependent rate $\delta = \delta(t)$. The amount of degraded mRNA molecules during the time interval dt can be expressed as $\delta(t)T(t)$. The synthesis rate function $\mu(t)$, the decay rate function $\delta(t)$ and the initial mRNA level $T(0)$ determine the total expression $T(t)$ and its labeled expression $L(t)$ by the differential equation

$$\frac{dT(t)}{dt} = \mu(t) - \delta(t)T(t). \tag{5.23}$$

Note that in [60], [67], we needed to account for an increase in the cell number with time. Here, we only follow the first cell cycle of the experiment, because the fluctuations in later cell cycles are attenuated too much to be informative for degradation estimation. Without loss, we may therefore assume a constant cell number, which simplifies our calculations considerably. Furthermore, we do not model cell growth, since we follow a synchronized

population of cells for one cell cycle, therefore the growth rate $\alpha = 0$. Equation (5.23) can be solved efficiently for arbitrary, sufficiently smooth functions μ and δ using a numerical ODE solver. Assuming piece-wise linear functions for μ and piece-wise constant functions for δ, it is even possible to derive the analytical solution to Equation (5.23).

We start labeling at time point t_0 and set

$$\Theta_g(t, t_0) := \int_{t_0}^{t} \delta(\xi) d\xi \tag{5.24}$$

The slope of the piece-wise linear function for μ changes at time points m_i with $i = 0, ..., k$. The piece-wise constant degradation rate δ changes at d_i with $i = 0, ..., n$. We set $H := \{h_i \mid i = 0, ..., k+n\} = \{m_i \mid i = 0, ..., k\} \bigcup \{d_i \mid i = 0, ..., n\}$with $\quad h_i \leq h_{i+1}$ for all i. On each interval $[a, b]$ ($a = h_i$,$b = h_{i+1}$) we can calculate

$$\phi(a, b) = \int_a^b \left[\mu^a + \frac{\mu^b - \mu^a}{m^b - m^a} \cdot (\xi - m^a)\right] e^{\alpha\xi + \Theta(\xi, 0)} d\xi \,.$$

Equation (5.23) can then be solved as

$$T(t) - T(t_j) = e^{-\Theta(t, t_j)} \left[T(t_j) + N \sum_{i|t>h_i} \phi(h_{i-1}, h_i) + \phi(h_{max}, t)\right] \tag{5.25}$$

with $h_{max} = max(h_i > t)$. The total amount of mRNA can therefore be derived using $t_j = 0$ and $T(0) = T_0$. The amount of labeled mRNA at time point t_j is obtained from Equation (5.25) by $L(t_j) = T(t_j + t_{lab}) - T(t_j)$, where t_{lab} is the length of the labeling interval.

5.6.2. Model specification

Given total and labeled time courses $T(t_k, i)$ and $L(t_k, i)$ of a gene in replicates $i \in I$, our main purpose is testing for the existence of periodic changes in mRNA degradation. We compare a model with constant decay rate, $\delta(t) = \delta$, with a model for regulated decay, in which $\delta(t)$ is a cosine function with average decay level δ_m, peak time φ and amplitude a. The synthesis rate $\mu(t)$ is modeled as a piece-wise linear function with 10 min intervals between interpolation points $(5 \text{ min}, \mu_0), (15 \text{ min}, \mu_1), ..., (65 \text{ min}, \mu_6)$. Additionally, we need to rescale the measured labeled mRNA fractions $L(t)$ by an unknown factor c in order to match the true fraction of newly synthesized mRNA given the amount of measured

total mRNA. This parameter reflects the true ratio between the (mean) labeled expression measurements and the (mean) total expression measurements of all genes at time 0. Given a complete parameter set Θ for one of the models, the synthesis and degradation rates are then converted into predictions for the labeled and total mRNA time courses, $\hat{T}(t_k;\Theta)$ and $\hat{L}(t_k;\Theta)$. Our target $l(\Theta)$ function measures the goodness-of-fit for both time courses, where goodness-of-fit is given by Equation (5.3). Hence,

$$\begin{aligned}\ell(\Theta) &= \sum_{i\in I}\ell(\Theta;T(t_k,i)) + \sum_{i\in I}\ell(\Theta;L(t_k,i)) \\ &= \sum_{i\in I}\sum_{k=1}^{K}\frac{(log\,T(t_k,i) - log\,\hat{T}(t_k;\Theta))^2}{2\cdot\sigma^2_{T,t_k,i}} + \sum_{i\in I}\sum_{k=1}^{K}\frac{(log\,L(t_k) - \frac{1}{c}\cdot log\,\hat{L}(t_k;\Theta))^2}{2\cdot\sigma^2_{L,t_k,i}}\end{aligned} \tag{5.26}$$

where $\sigma^2_{T,t_k,i}$ and $\sigma^2_{L,t_k,i}$ are the regularized replicate-, gene- and time-specific standard deviations obtained in Section 5.2.2.

Thus, the full model M_1 assuming constant decay for one gene is parameterized by

$$\Theta_{M_1} = \{c,\, \delta,\, \mu_0, ...,\, \mu_k\} \tag{5.27}$$

and the competing model M_2 using a sigmoidal function for the decay is parameterized by

$$\Theta_{M_2} = \{c,\, \delta_m,\, \varphi,\, a,\, \mu_0, ...,\, \mu_k\} \tag{5.28}$$

Both models are fitted using standard Metropolis-Hastings MCMC (we use Gaussian proposal functions truncated to the positive real values).

5.6.3. Detection of genes with variable degradation rate

Applying both the constant and the regulated decay model to a gene profile, this results in a score for the constant model Θ_{M_1} and a score for the regulated Model Θ_{M_2} (compare Equation (5.26)). For each gene profile we compare the fit of the two models by calculating a *Variable Degradation Score,* VDS, which is given by the log-likelihood ratio between the two models:

$$VDS = \ell(\Theta_{M_1}) - \ell(\Theta_{M_2}) \tag{5.29}$$

Since constant degradation is a special case of variable degradation with a max/min ratio of 1, the constant degradation model never scores better than the variable degradation

model. Consequently, the Variable Degradation Score is never negative. It is zero when both models fit equally well, and it is higher the more the variable degradation model is required to explain the data. By simulations we determined the sensitivity and specificity of different Variable Degradation Score cutoffs. We concluded that a Variable Degradation Score cutoff of 0.3 ensures sufficiently high sensitivity and specificity for genes with a degradation rate amplitude (max/min ratio) of at least 1.5.

6. Results and Discussion

6.1. cDTA monitors mRNA synthesis and degradation during the cell cycle

To measure mRNA synthesis rates over the yeast cell cycle, we synchronized cells using alpha factor as described [82]. For consistency with prior studies, we generated and used a bar1 deletion strain of yeast (Methods 5.1). After release of cells in G1 phase we used cDTA [60, 67] to measure the amount of newly synthesized and total RNA at 41 time points separated by 5 minutes, covering 200 min, corresponding to three cell cycle periods. At each time point newly synthesized RNA was labeled with 4-thiouracil for 5 minutes (Figure 6.1A). Using *S. pombe* as an internal standard, we normalized the labeled and total mRNA fractions across the time series to get absolute expression estimates (Methods 5.1). The entire time series experiment was performed in two biological replicates. Because labeled mRNA levels correlate well with mRNA synthesis rates, these data represent the first genome-wide estimation of mRNA synthesis rates in synchronized cells at different time points in the cell cycle.

We did extensive checks to verify the quality of our data set. Correlations (within labeled respectively total samples) were consistently above 0.93 (an example is shown in Figure 6.1B). Strikingly, periodic expression already shows in the samples correlation structure. Samples taken at similar time points in the cell cycle have a higher correlation than samples taken at more distant time points in the cell cycle. This leads to a characteristic tri-band diagonal correlation structure, corresponding to the three cell cycles that we monitored. A principal component plot automatically places consecutive samples in a 'cell cycle clock', a clock-wise spiral, demonstrating that most variation in the data (>74%) is due to periodic expression fluctuations (Figure 6.1C).

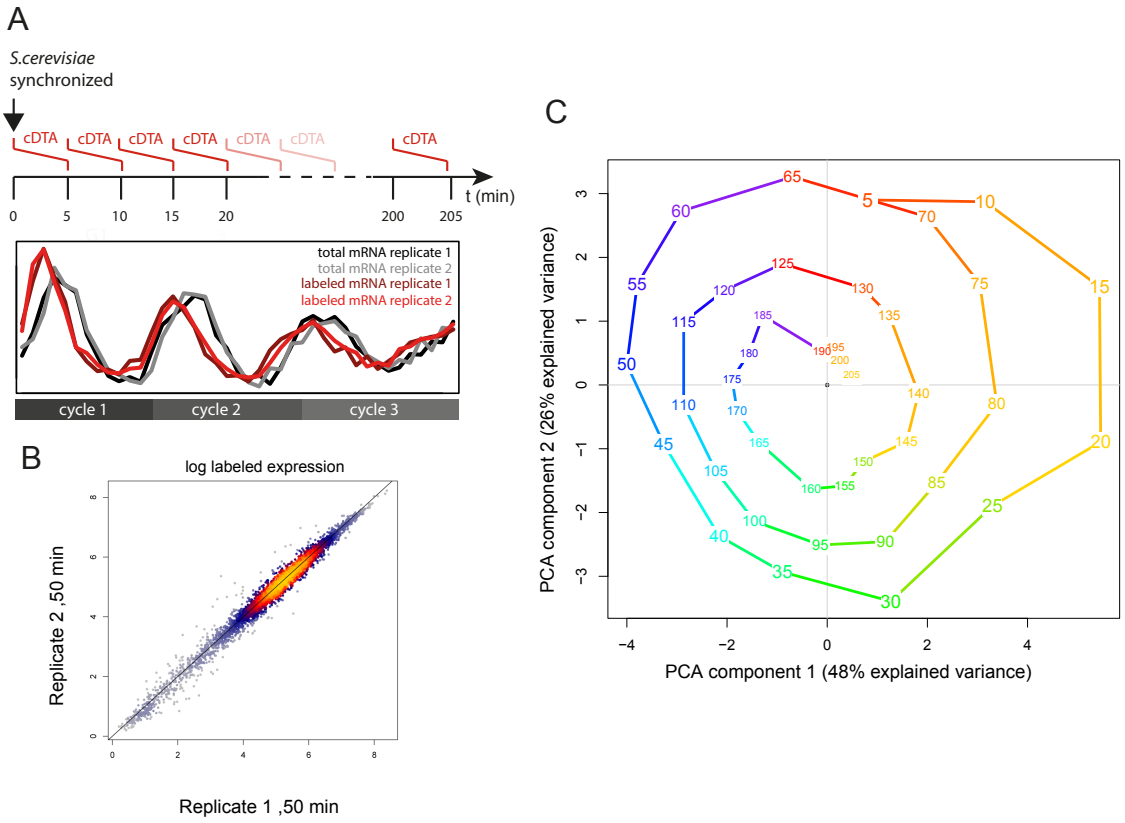

Figure 6.1.: cDTA cell cycle time course experiment and quality. (A) mRNAs were labeled with 4-thiouracil every 5 minutes until t=200min. After 5 min labeling time the respective sample was stopped and further processed according to the cDTA protocol. The experiment was performed in two replicates. For each mRNA, we obtained two time series of total mRNA levels (black and grey lines), and two time series of labeled mRNA levels (dark red and light red lines). (B) As a representative example, the scatter plot shows a comparison of the log labeled expression levels for all genes 50 minutes after synchronization in the two independent time series. (C) The yeast 'cell cycle clock'. Each point corresponds to the microarray measurements of one time point. Its coordinates are the projection of the corresponding expression vectors onto the two first principal components in a principal component analysis. Color coding is according to time in the cell cycle.

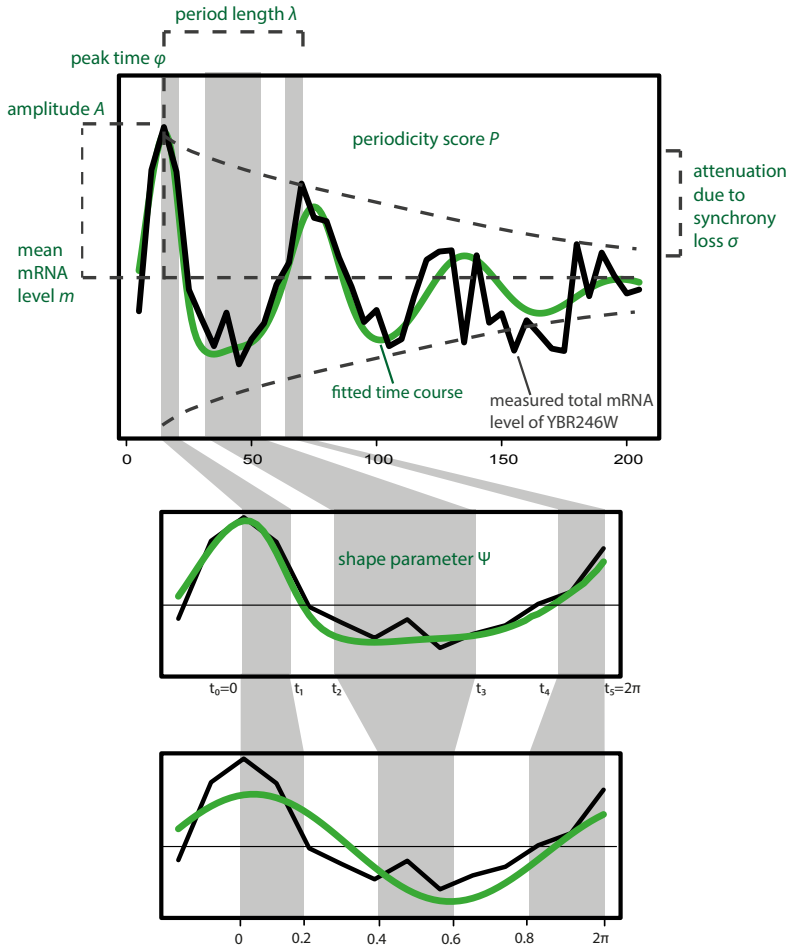

Figure 6.2.: Parametric modeling of periodic time courses by MoPS. Top panel: the peak time ρ denotes the time of maximum mRNA level relative to the time of transcription release. The period length λ is the shortest interval after which the expression pattern repeats, m and A denote the mean mRNA level and the amplitude, respectively. The decrease of the amplitude along several cell cycles is due to synchrony loss, σ. Bottom panel: The cosine wave, our basic model of periodic expression, is adapted to the time series by the shape parameter ψ, which is a monotonic transformation of the 'clock' that ticks along the interval $[0, 2\pi]$.

6.2. Model-based periodicity screening (MoPS)

6.2.1. Overview

To study cell-cycle regulated RNA metabolism, we first developed a new parametric screening method for the detection of periodic fluctuations in time (Model-based Periodicity Screen, MoPS, Figure 6.2, Methods 5.2.4). MoPS is available as an R/Bioconductor package at www.bioconductor.org and can be widely applied to various kinds of datasets. Here, we used MoPS to examine each labeled and total mRNA expression time course for periodicity and thus cell-cycle regulation. MoPS calculates a likelihood ratio statistic that compares the best fit of a periodic expression curve to that of a non-periodic curve (Methods 5.2). Periodic expression is modeled by a dampened, deformed cosine wave using 6 parameters (Figure 6.2).The cell cycle length λ (min) corresponds to the time difference between the first expression peak at peak time when the maximum mRNA level is observed and the next expression peak. The periodically changing mRNA level is described by its mean m and its amplitude A. The decrease of the mRNA level amplitude with time due to progressive loss of synchronization between cells is described as the 'synchrony loss' σ. We explain this effect by variation in cell cycle length of individual cells in our synchronized population. The parameter σ describes the dispersion of the cell cycle length distribution. The deformation of the cosine wave is described by a 'shape' parameter ψ, a bijective transformation of the interval $[0, 2\pi]$. Since the mean cell cycle length and the loss of synchronization are a characteristic of the cell population, these two parameters are common to all examined transcripts. This reduces the number of fitted parameters for individual gene expression profiles to four, which makes MoPS extremely robust, despite its flexibility that ensures excellent fits.

6.2.2. Simulation study to evaluate MoPS

Prior to the analysis of the cDTA time series, we conducted a simulation study to assess the discriminatory power and the reconstruction accuracy of MoPS. We randomly generated a set of 200 periodic functions f with period length $\lambda = 12$ min, no attenuation ($\sigma = 0$), shape and peak time chosen at random, mean 0, and amplitude 1. Each function was used to generate a periodic time course $y_j = f(t_j) + \epsilon_j$, $j = 1, ..., 37$, for $t_j = \frac{\lambda}{6} \cdot j$, and Gaussian noise $\epsilon_j \sim \mathcal{N}(0, \sigma_j^2)$. We simulated heteroscedastic noise by choosing the variance σ_j^2 dependent on the expression level, $\sigma_j^2 = \sigma_0^2 + \frac{\sigma_0^2}{2} \cdot f(t_j) \in [\frac{\sigma_0^2}{2}, \frac{3\sigma_0^2}{2}]$. We varied the global signal-to-noise ratio σ_0 between -1 and 1.5. The data set was complemented by 200 time courses containing only noise of the same variance. We combined rigid or flexible curve modeling (using only sine waves as basic functions, or the more flexible curves) with a constant (homoscedastic) or a variable (heteroscedastic) error model, resulting in

4 screening strategies. For each signal to noise ratio, all 400 time courses were ranked according to their periodicity score, and discriminatory power of the periodicity score was measured by the area under receiver operating curve (AUC, Figure 6.3A). It turns out that a heteroscedastic error model is always beneficial if the magnitude of the errors is modeled correctly. Remarkably, the discriminatory power is almost constant if one includes the shape of the curve as a parameter, which indicates the robustness of our model against overfitting. We additionally quantified the quality of the fit by calculating the average residual sum of squares (RSS, Figure 6.3B) for all periodic time courses. For moderate or low noise levels (log2 signal/noise › 0.5), flexible curve fitting yields a more accurate estimate of the true curve, while at the same time extracting more information from the data.
Taken together, the characterization of the space of periodic functions by a sparse, interpretable parameter set makes the algorithm flexible enough to yield a concise fit of periodic functions while at the same time being robust against overfitting. MoPS also accounts for heteroscedastic noise in the measurements, which substantially improves the quality of fit (examples in Figure 6.3C).

6.3. MoPS applied to cDTA time series

6.3.1. Identification and characterization of periodically expressed genes

To identify a reliable set of genes that are periodically expressed, we applied MoPS separately to total and labeled mRNA from both replicate cDTA time series. The cell cycle length λ and the synchronization loss σ were estimated for each gene. Among genes that have a positive periodicity score, the distribution of obtained cell cycle lengths λ sharply peaks at a median of 62.5 min, and the distribution of the synchrony losses σ has a median of 7 min (Figure 6.4A). The cell cycle length estimate of 62.5 min agrees well with that of 65min in [82] who used the same strain and the same synchronization method. In a second step, we fixed the parameters λ=62.5 min and σ=7 min, and recalculated all other parameters, namely the phase of expression, the characteristic shape of its time course and the periodicity score for all genes. The obtained values were in excellent agreement between replicates, and also in good agreement between labeled and total mRNA (Figure 6.4B,C). Genes were then ranked according to their periodicity score (for a representative selection of genes and their periodicity scores see Figure 5.2).

MoPS computes a periodicity score for each gene and thus allows ranking of all genes according to their likelihood ratio to be periodically expressed resp. constantly expressed. However, there is no obvious way to assign significance to this score. We used existing

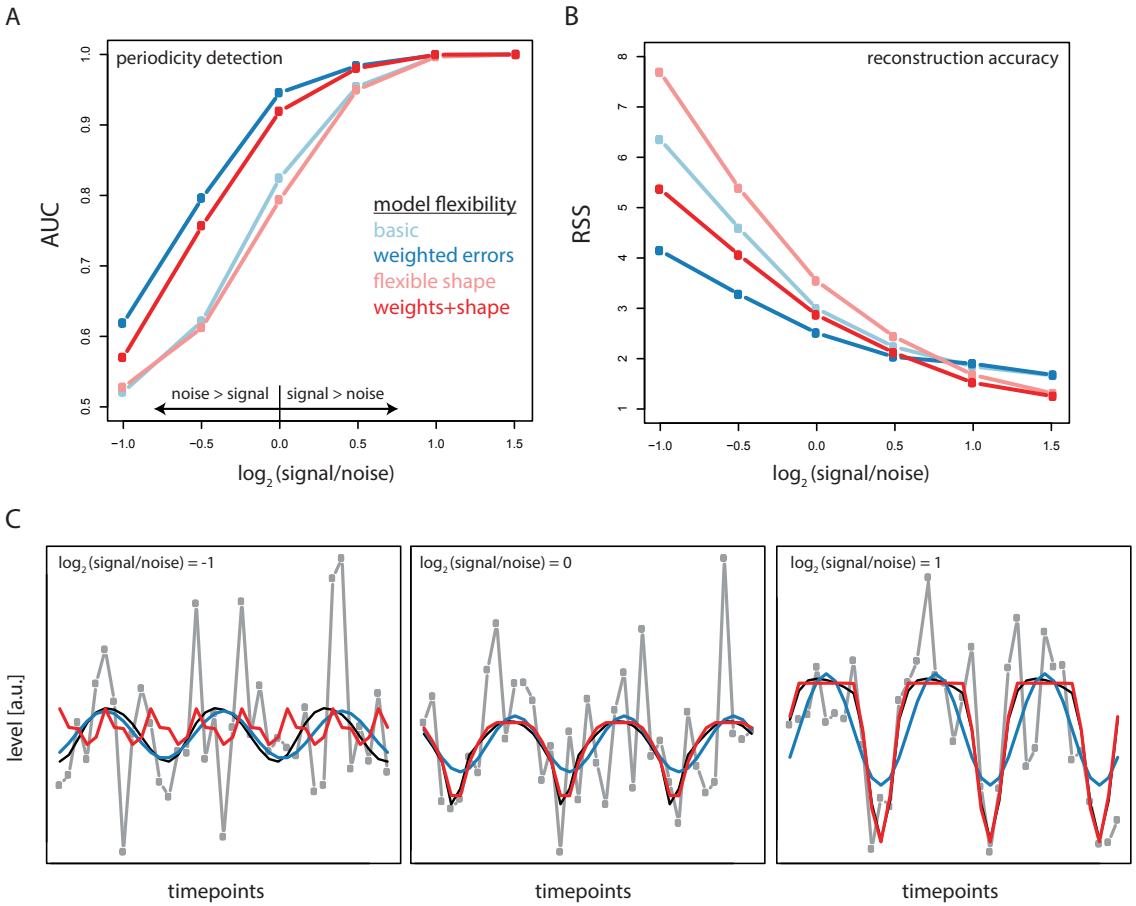

Figure 6.3.: Top panel: Detection specificity and sensitivity for screening with fixed sine shape (blue) and variable curve shape (red), with constant error model (faint colors) and variable error model (saturated colors). A) Each square represents the Area under curve value (AUC, y-axis) for one method at a certain signal-to-noise level (x-axis). B) Each square represents the average quadratic distance (RSS, y-axis) of the predictions to the true values at $t_1, ..., t_{37}$, for all periodic time courses. Bottom panel (C): Examples of the true signal (black curve), the measurements (grey squares connected by lines), for different noise levels, $\log_2$(signal-to noise) = -1, 0, 1. Colored lines show curve fits derived from the measurements using sine waves (blue) or flexible curves (red).

knowledge derived from published studies about periodically expressed genes to define a positive set and a negative set. The positive set comprises the top 200 periodic genes from Cyclebase [83] and the negative set consists of genes that have never been classified as cell cycle regulated in any cell cycle expression study. The empirical distribution of all MoPS scores is fitted by a mixture of the empirical distributions of the MoPS scores of the positive respectively the negative set (see Methods 5.3). A cut-off value was chosen to control the false discovery rate at a 20% level (Figure 6.4F). The power of our screening method was increased by combining the periodicity scores obtained from the total and labeled mRNA from both replicates into one sum.

6.3.2. Validation of periodically expressed genes

6.3.2.1. Comparison with other cell-cycle expression studies

MoPS identified a total of 479 periodic genes with high confidence (Figure 6.4). In the literature, different periodicity screening methods yield between 300 and 1500 genes that are considered cell cycle regulated in yeast [99, 80, 100]. The agreement between the datasets/methods is moderate [81], but nevertheless highly significant ($p < 10^{-10}$ in all pair wise Fisher tests). This shows that the detection of periodic genes strongly depends on the method, the experimental conditions, and the stringency cut-off that has been applied. Our set of significantly periodic genes can be compared to other studies by visualizing the overlap in identified periodic genes (Figure 6.5). Three studies are chosen for comparison: Spellman et al. [80] as the pioneering cell-cycle Microarray study; Granovskaia et al. [82] as the most recent study; Cyclebase [83] as a meta-study that combines several studies. Only 246 genes are found to be cell-cycle regulated by all studies, while there are 523 genes that are only identified in one study.

6.3.2.2. Benchmark on identification of bona-fide cell-cycle genes

We validated our periodicity screening using a framework proposed by de Lichtenberg et al. [81]. They developed their own periodicity screening method and applied it to 6 different cell cycle expression data sets. The resulting 6 ranked lists plus a combined reference list of periodically expressed genes are accessible from the Cyclebase repository (www.cyclebase.org, [83]). We compared our ranking of periodic genes to these 7 lists using the benchmark scheme as in [81]. The ranked lists were retrieved from Cyclebase. These lists correspond to different cell-cycle microarray datasets that have been normalized in the same manner and are ranked according to periodicity by the method of Lichtenberg et al. [81]. Additionally, it contains a ranked list that was derived by combining all datasets.

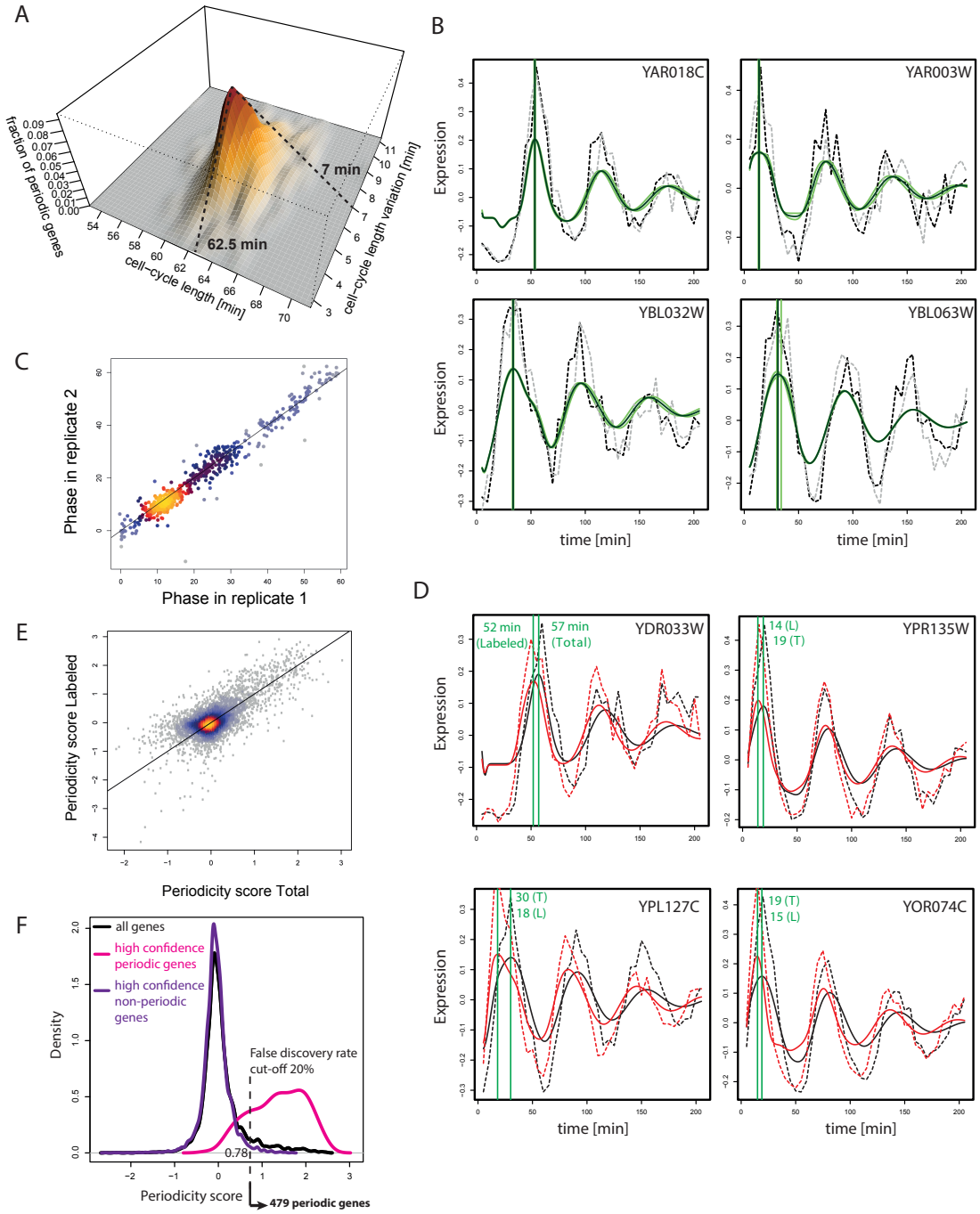

Figure 6.4.: (A) Identification of the global parameters cell cycle length (λ, x-axis) and synchrony loss (σ, y-axis). Each gene yields an estimate (λ,σ). The 3D surface plot shows their joint distribution with the medians λ=62.5min and σ=7min. (B) Replicate expression measurements (black and grey dotted lines) of selected genes together with the fitted characteristic time courses and timing of peak expression (light green). The estimates are averaged for further analyses (dark green) (C) Estimated phases of genes that are significant periodic in labeled data (Pearson correlation 0.97) (D) Total (black) and Labeled (red) time courses (dotted lines) are shown together with the MoPS fitted characteristic time course (solid lines). Peak times as estimated from MoPS are shown as green lines. (E) Comparison of individual periodicity scores for total and labeled mRNA. (F) The distribution of periodicity scores (black distribution) is approximated as a mixture of the periodicity score distributions of a set of bona fide periodic (red distribution) and non-periodic genes (blue distribution). Based on this fit, the 20% false discovery rate cutoff is calculated as 0.78.

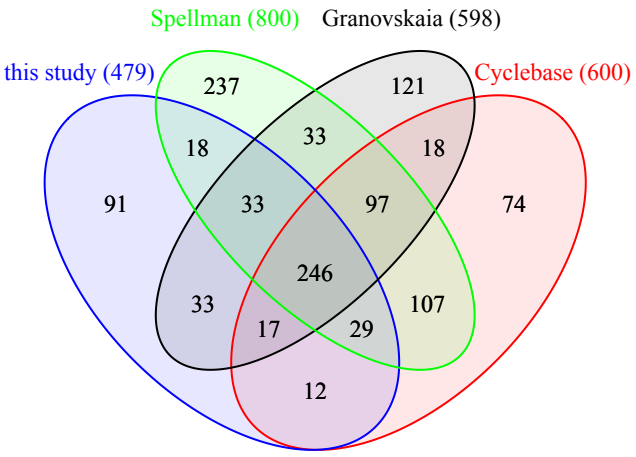

Figure 6.5.: Overlap of MoPS cDTA periodic genes with results from other cell-cycle expression studies.

Three different benchmark sets were used as a gold standard to assess the quality of a gene list by a receiver operating characteristic (ROC) analysis:

- Set B1 - A total of 113 genes previously identified as periodically expressed in small-scale experiments.
- Set B2 – 352 genes whose promoters were bound (P-value below 0.01) by at least one of nine known cell cycle transcription factors in two independent Chromatin IP studies.
- Set B3 – 518 genes annotated in MIPS [101] as 'cell cycle and DNA processing'.

A comparison of our ranked list with the other lists was performed as proposed in [81]. In all 7 cases and for all 3 benchmark sets, the de Lichtenberg method has been proven to perform better or at least as good as competing methods [102]. Our ranking, when included in the ROC analysis, performs comparably to the Lichtenberg method in all 3 benchmark scenarios (Figure 6.6). Out of the top 200 periodic genes from the combined ranking, we find 152 to be significantly periodic with our approach applied to our dataset. Visual inspection shows that the 48 genes that we did not classify as periodic in our dataset, indeed exhibit predominantly periodic profiles in labeled and total but show low correlation between replicates or show deranged profiles in the first 30 minutes after synchronization.

6.3.2.3. Robustness of peak time assignment

We follow the validation approach as described by Guo et al [87]. to estimate the robustness of peak time assignment to experimental noise. We added varying amounts of Gaussian noise to the measured time course of a gene and extract the peak timing of expression. We then compare the perturbed estimates with the original peak times. As in [87], we

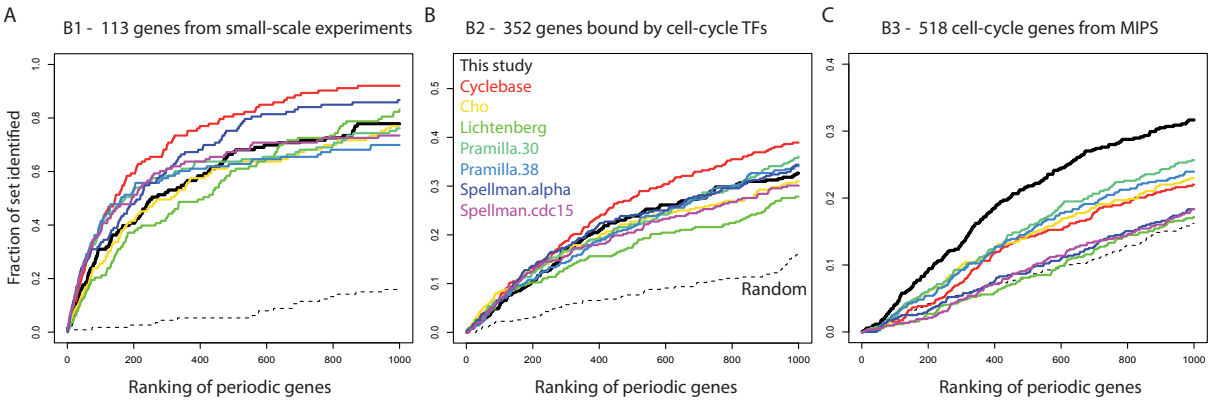

Figure 6.6.: Benchmark of MoPS. Shown are ROC-like curves, one for each ranked list of bona-fide cell-cycle genes. Three different benchmark sets of cell-cycle regulated genes are used as a gold standard for validation. The number of top n genes of a ranked list (x-axis) is plotted versus the fraction of the benchmark set which is contained in the top n genes, respectively. (A) The gold standard set B1 consists of 113 genes identified in small-scale experiments. (B) Benchmark set B2 consists of 352 genes identified in two independent Chromatin IP (ChIP) studies. These genes were found to be bound by known cell-cycle associated transcription factors. (C) Benchmark set B3 consists of 518 genes annotated in MIPS as 'cell cycle' or 'DNA processing'.

select the top 100 genes ranked by our periodicity score for benchmarking. For varying levels of noise, we generate 10 perturbed time courses for each gene, estimate the peak time with MoPS and compute the unsigned timing differences to the original estimates. The level of noise that is added at every time point is taken from a normal distribution (mean =0, sd = noise.level * error). The error is estimated from the calculated variation in our experimental replicate time series (see Section 5.2.2). We use four different levels of noise: 0.5 (more precise than actual measurements), 1, 1.2 and 1.5. The median peak time deviation was in the range of 1-2.5 min, confirming the accuracy of our estimates (Figure 6.7).

6.4. Three expression waves during the cell cycle

We sorted all periodically expressed genes by their synthesis peak time (Figure 6.8). Among the periodically expressed genes were many prominent cell cycle genes [79] including all six genes of the minichromosome maintenance family (MCM2-7), cyclins, and histone genes. These genes were used to assign cell cycle phases G1, S, G2, and M to measurement time points in our data. Periodically expressed genes appeared to be grouped in three expression waves, in agreement with previous observations [103]. A first wave shows peak synthesis in G1 phase, a second during S phase, and a third at the onset of M phase (Figure 6.8).

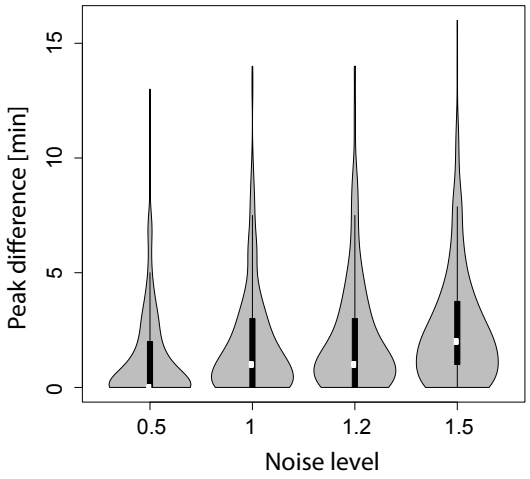

Figure 6.7.: Violin plot showing peak time variation in simulated perturbed expression measurements. Each violin shows the distribution of unsigned differences between peak timing estimated from original and perturbed time courses (labeled data).

6.5. Recovery of cell cycle transcription factors

Our 479 periodically expressed genes also contained 8 transcription factors (TFs) that potentially regulate the cell cycle. Since there is no consensus set of TFs that regulate the cell cycle we systematically screened for transcriptional regulators of periodic genes (Figure 6.9A). We used only the labeled mRNA data in this screen, because these represent transcriptional regulation better than total mRNA profiles. The extracted shapes of the periodic genes were grouped into 10 clusters by Euclidean distance average linkage k-means clustering. For each of the clusters, we performed an XXmotif search [97] for DNA sequence motifs in a region 500 bp upstream of the experimentally defined transcription start site. In total, 50 motifs with E-value smaller than 1 were recovered. Each motif was then matched to known DNA-binding protein motifs with TOMTOM (Methods 5.5). We obtained a total of 50 DNA motifs that were associated with a total of 32 DNA-binding transcription factors. The top motif identified from a G1 cluster perfectly matched the known Mlu1 cell cycle box (MCB) motif (Figure 6.9B). The MCB motif is enriched in promoters of genes required for DNA synthesis. TOMTOM identified two TFs that were significantly associated with the MCB motif, MBP1 and SWI6, which form the MBF heterodimer in which MBP1 acts as a sequence-specific, DNA-binding trans-activator. MBF regulates expression during the G1/S transition [104]. The second best motif that was found to be enriched in M phase was matched by multiple TFs (MCM1, NDD1, YOX1, FKH2, DIG1, ASH1, and FKH1), reflecting a complex interaction network of activators and repressors. The repressor Yox1 and the activator Fkh2-Ndd1 compete for binding to Mcm1, although they associate at opposite sides of the dimeric Mcm1 transcription factor. This competition determines the expression of late mitotic genes in yeast [105].

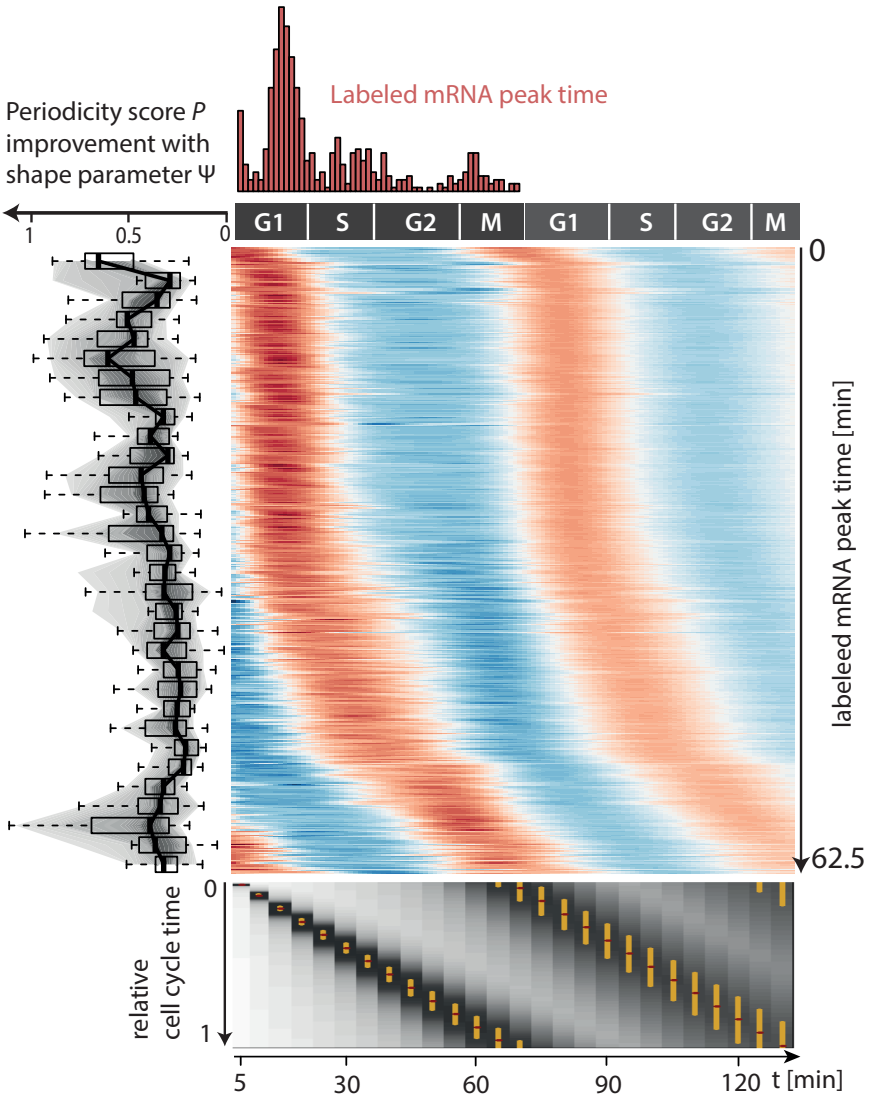

Figure 6.8.: Fitted time courses of the 479 periodic genes. Each row corresponds to one gene. Time is measured in terms of cell cycle phases G1,S,G2,M (x-axis). Genes were sorted according to their peak time, starting with genes peaking in G1 phase. High (low) expression is encoded in red (blue), where expression is taken relative to the gene's mean expression. The histogram on top shows the distribution of the peak times along the cell cycle. The snake plot to the left shows the improvement of the MoPS fit over a fit with a sine wave. Each box in the plot summarizes 15 consecutive genes. The bottom plot shows how the synchronization of cells decreases with time. Each measurement time point is represented by one column, which is a grey scale-coded representation of the individual cell cycle time distribution across the cell cycle. The golden bar marks the central 50% interval of the respective distribution. The dark red dot within the gold bar marks the modes of these distributions.

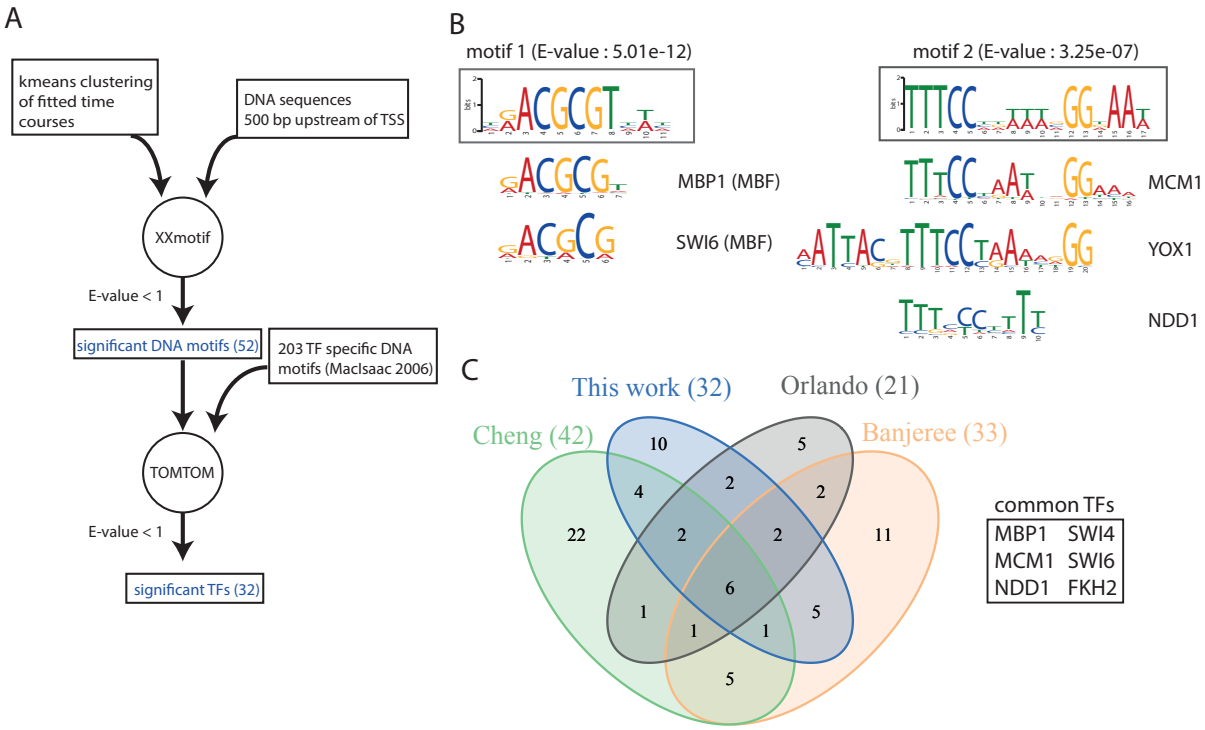

Figure 6.9.: Identification of cell-cycle related DNA motifs and transcription factors (A) Workflow, consisting of a motif discovery step using XXmotif and a TF detection step using TOMTOM. The input to XXmotif are the 500 bp upstream sequences of sets of co-regulated genes. The resulting list of significantly enriched motifs is processed by TOMTOM to find TFs with matching binding sites. (B) Sequence logos of the two top motifs, their associated TFs (MBF for motif 1, MCM1, YOX1, NDD1 for motif 2), together with their E-value. (C) Venn diagram showing the overlap of various integrative bioinformatics methods for the prediction of cell-cycle related TFs.

The obtained set of 32 predicted cell cycle TFs partially overlaps with TFs in other studies that integrate expression data with motif-discovery tools [106, 107, 76, 108, 109] (Figure 6.9C). It is evident that the association of TFs with cell cycle regulation is only clear for a core set of a few TFs. Wu and Li performed a benchmark on TFs annotated as known cell cycle regulators [110] with the Jaccard index as a measure of agreement. Their method scores best (Jaccard index 0.293), whereas our set of TFs led to a Jaccard index of 0.275, which is higher than 0.245, the score of the second best TF set [108] in their study.

6.6. TFs govern the expression timing of periodic genes

We investigated the influence of cell cycle regulating TFs on the mRNA synthesis of their target genes. Using ChIP-chip derived TF-target gene associations [98] of our 32 cell cycle regulators to our 479 periodic genes, we compare the total mRNA time course of a TF to the labeled time course of its targets. 8 of our 32 cell-cycle TFs are periodically expressed themselves. Their time course of total mRNA levels corresponds to their regulatory role in

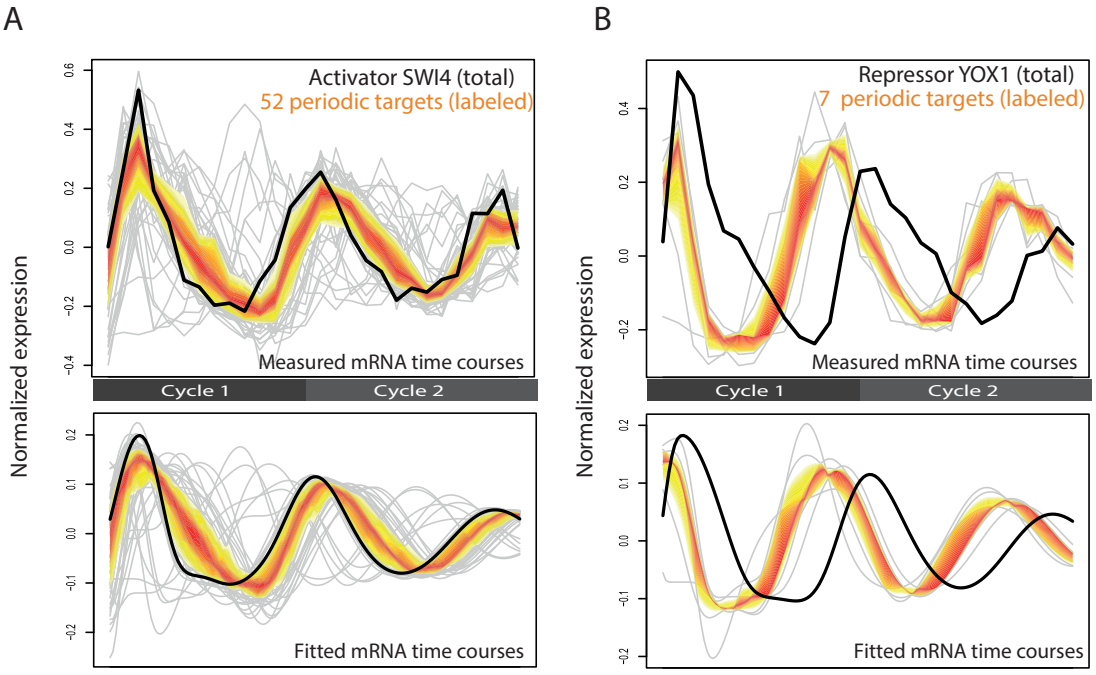

Figure 6.10.: Cell cycle regulators and their target genes. Two examples of identified cell-cycle regulating TFs (A: SWI4, B: YOX1) and their periodically expressed target genes. The top panel shows the measured, mean centered and re-scaled time courses of the TF (total mRNA, black line) and its periodically expressed targets (labeled mRNA, grey lines). The yellow-orange band marks the expression range of the central 50% of the targets at each time point. The bottom panel shows the corresponding fitted time courses as derived from our screening procedure.

cell cycle-associated transcription activation or repression. The expression of an activating TF is expected to precede the synthesis of its target genes. This is in accordance with our observations. SWI4 is a known activator working together with SWI6 to activate G1-specific transcription of targets. Indeed, the level of SWI4 mRNA peaks shortly before the synthesis peaks of its periodic target genes (Figure 6.10A). In contrast, the expression of a repressive TF should be preceded by the synthesis peak of its target genes. Indeed, the transcriptional repressor YOX1 that regulates genes expressed in M/G1 phase [111] shows high expression after peak synthesis of its target genes, and low mRNA levels when the synthesis rate of its targets is high (Figure 6.10B).

The periodically expressed gene FKH2 is described as having a dual role as activating and repressing TF [112]. Its targets peak either at the onset of M phase, shortly after the FKH2 peak, or at late G1 phase, shortly before the FKH2 peak. The first group is consistent with an activating role of FKH2, the second group seems to be repressed by FKH2 (Figure 6.11). Targets of non-periodically expressed TFs show also coherent timing, the most compelling example being the TF MBP1 and genes exclusively targeted by MPB1. The same effect was observed for all target gene sets with identical motif composition in their upstream region. Thus, in many instances the expression levels of regulatory TFs could explain the synthesis rates of their target genes.

6.7. Quantification of absolute mRNA abundance

The mean expression and amplitude is estimated for all 479 periodic genes by fitting their MoPS estimated characteristic time course to absolute mRNA concentrations. The minimization problem is solved with linear regression (see section 5.2.4). This extends the MoPS estimated time courses by adding information about the absolute mRNA levels (Figure 6.12A,B). Mean expression levels of non-periodic genes are determined by using the mean expression in the time course of the first cell cycle. We observe a very high correlation of absolute mean mRNA levels in replicates of total and labeled datasets (Figure 6.12C,D). The distribution of the mean expression of the periodic genes is comparable to that of all genes, with the exception of the left tail of weakly expressed genes. This is not surprising, because periodic genes fluctuate in their expression, which necessarily leads to a certain minimum mean expression level.

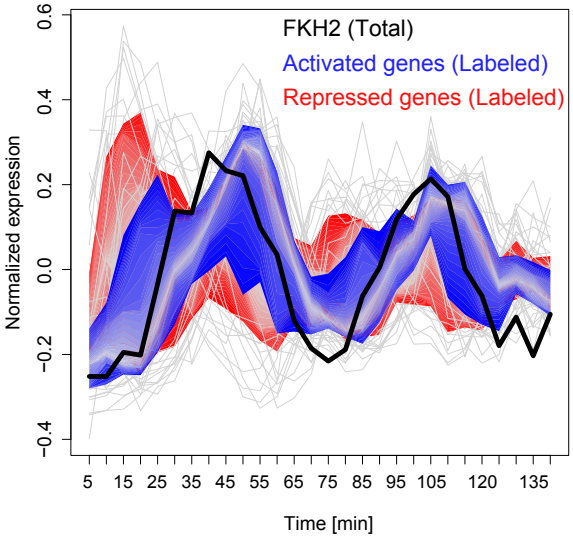

Figure 6.11.: Labeled expression time courses of 47 targets of the cell-cycle transcription factor FKH2. FKH2 total mRNA peaks at the beginning of M-phase (black line). One group of periodic targets show labeled expression peaks approx. 10 minutes after FKH2 peak expression (blue), a second group comprises genes that show labeled peak expression when FKH2 levels are low (red).

6.8. The core promoter governs the synthesis rates of periodic genes

The genes exclusively regulated by MBP1, though agreeing well in their timing, showed a remarkable diversity in their synthesis mean and amplitude (Figure 6.13A). The distribution of their mean synthesis rates resembles that of all periodic genes. This could also be observed with other sets of target genes which are regulated by common cell cycle TF(s). This suggested that TFs determine the timing but not the magnitude of the transcription rate of their target genes. We therefore checked whether the synthesis rate is rather set by the target gene core promoter sequence. We analyzed the deviation of the TATA box sequence from the TATA consensus. Genes were partitioned into 3 groups, genes with a perfect TATA box (0 mismatches to the TATA consensus motif), and TATA-less genes showing 1 or 2 mismatches compared to the TATA box consensus at the

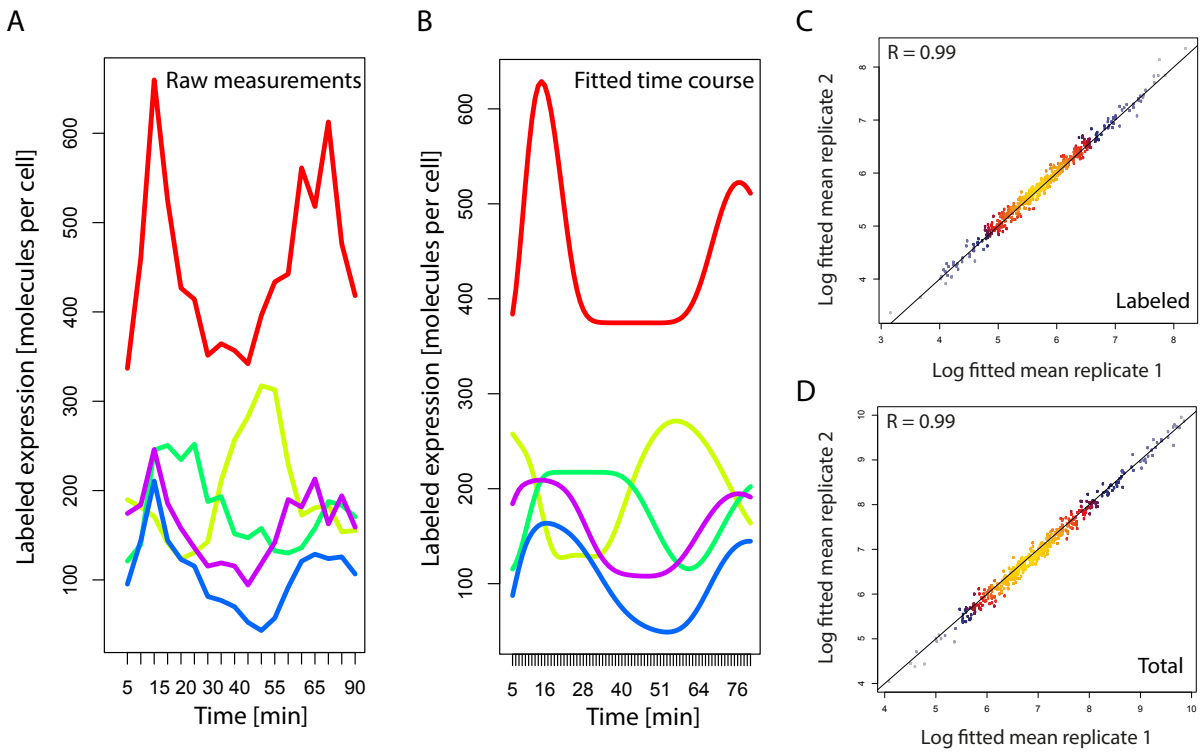

Figure 6.12.: (A,B) Fitted absolute expression time courses and corresponding raw measurements in labeled data for five periodic genes. (C,D) Correlation of estimated absolute mean expression of 479 periodic genes between replicates in labeled and total data.

experimentally defined location where the transcription pre-initiation complex is formed [113]. We excluded genes with more than 2 mismatches from the analysis, since only three of these genes were periodically expressed. For non-periodic genes the distribution of mean synthesis rates peaked at similar values for all TATA groups, with perfect TATA-containing genes peaking only slightly higher than TATA-less genes (p-value $<$ 1.7e-5, Wilcoxon test) (Figure 6.13B). For periodic genes, however, the perfect TATA box group showed a substantially higher mean synthesis rate than the imperfect TATA box groups (p-value $<$ 10e-10, Wilcoxon test). Although the differences are significant for non-periodic and periodic genes, the effect is 3-fold stronger for periodic genes. Indeed, periodic genes with very high levels in total and labeled mRNA were almost exclusively found in the perfect TATA box group. Gene Ontology analysis [114] of the 80 periodic genes with a consensus TATA box (using all 479 periodic genes as background) showed enrichment for processes of cell cycle progression, with the most significant process being the regulation of CDK activity by cyclins (CLN1, CLN2, CLB1, CLB6, PCL7, and PCL2). Further enriched Gene Ontology categories include DNA replication (POL12, POL30), chromosome organization during meiosis (MCD1, SGO1, GNA1), and chromatin assembly and histone formation (HTB1, HTA1, HHT2) (Figure 6.13B). To corroborate these findings, we compared the occupancy levels of the general transcription initiation factor TFIIB at core promoters

[113] to the mean labeled mRNA levels of periodic and non-periodic genes. We observed a high correlation of TFIIB occupancy with the expression mean (Figure 6.13C) and amplitude. The highest correlation was found for periodic TATA-box containing genes. Whereas other initiation factors behave like TFIIB, the initiation factor TFIID occupancy correlated only weakly with expression levels of periodic genes, regardless of the core promoter sequence. This is in line with the proposed role of TFIID in the transcription of constitutively expressed genes [115]. To conclude, the mRNA synthesis of cell-cycle regulated genes is governed by the sequence of the core promoter rather than the binding of upstream TFs, which however control the timing of expression.

6.9. Degradation rates of periodic mRNAs are not constant

Assuming that all copies of a transcript in an mRNA population share the same hazard of being degraded, the time course of an mRNA population is described by the differential equation

$$dT/dt = \mu(t) - \delta(t) \cdot T \tag{6.1}$$

where T is the mRNA level, $\mu(t)$ is the time-dependent synthesis rate and $\delta(t)$ is the time-dependent degradation rate for that population. Given $\mu(t)$ and $\delta(t)$, Equation 6.1 can predict the time course of total and labeled mRNA levels. Note that Equation 6.1 leaves one degree of freedom, the boundary condition on T. By setting T(0) to the total RNA level at time 0, the resulting solution T(t) to Equation 6.1 is the time course of the total RNA. By letting T(tj)=0, the solution T(t), for t>tj, is the amount of labeled RNA obtained after a (t- tj) min labeling pulse starting at time tj. For a description of the numerical and analytical solutions to Equation 6.1 see Methods 5.6.1. We used Equation 6.1 to simulate how a peak in mRNA synthesis translates into total mRNA in different degradation rate scenarios (Figure 6.14A). In particular, we computed the peak time delay between synthesis rate peak and total mRNA peak. A constant, low degradation rate leads to a broad peak in total RNA with a large peak time delay. A constant, high degradation rate reduces this time delay substantially, yet at the expense of a reduced total mRNA level. A variable degradation rate with a peak following the synthesis rate peak however results in a shorter peak time delay while maintaining a high total mRNA peak. The simulation shows that appropriate changes in the mRNA degradation rate minimize the peak time delay while still achieving a quantitatively high total mRNA

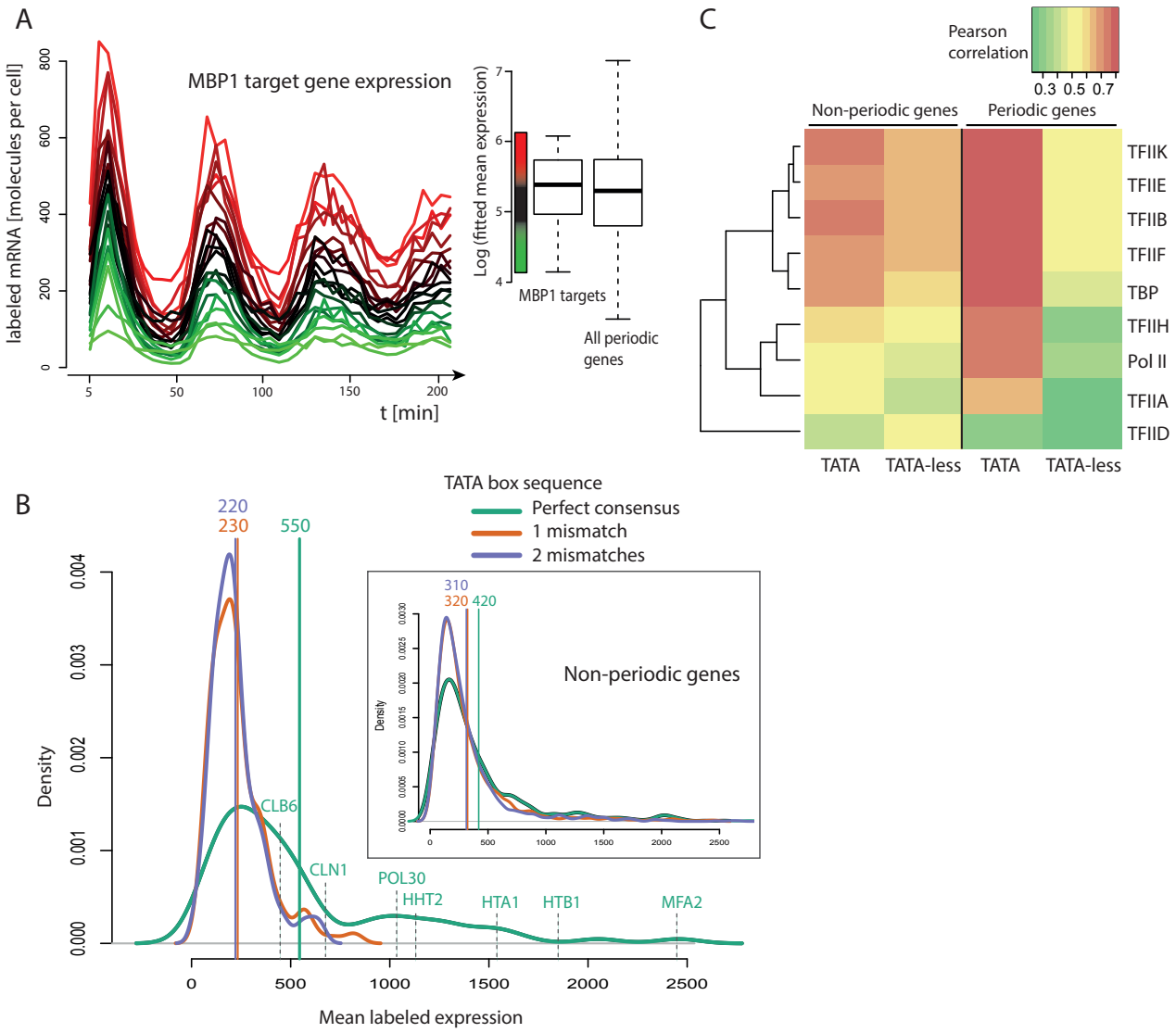

Figure 6.13.: Promoter and enhancer structure determine expression strength and timing (A) Time courses (absolute labeled mRNA measurements) of 22 periodically expressed genes that are exclusively annotated as MBP1 targets. Colors correspond to mean expression levels extracted from our fitting procedure. Box plots on the right show the expression ranges of the 22 MBP1 targets (left) and all 479 periodic genes (right) in logarithmic scale. (B) Densities of mean labeled mRNA expression of periodic genes respectively non-periodic genes (inset), stratified for the number of mismatches to the TATA consensus motif at core promoter (green: 0 mismatches, orange: 1 mismatch, purple: 2 mismatches). The green, orange and purple vertical lines indicate the mean (rounded to nearest tenth) of the respective distributions. Selected periodic genes that have a consensus TATA-box and are associated with enriched cell-cycle processes (7 out of 80, see text) are marked. (C) Correlations between the mean labeled expression of non-periodic and periodic genes grouped into TATA-containing (0 mismatches) and TATA-less with core promoter occupancies of general TFs involved in pre-initiation complex formation. The heatmap colors range from green (moderate correlation) to yellow (good correlation) and red (high correlation).

response. We therefore compared the peak time in labeled and total mRNA for each periodic gene (Figure 6.14B). This revealed a median time delay of 2 min between the total RNA with respect to the labeled RNA peak (mean 2.8, 1st quantile 0, 3rd quantile 4.0 min). According to our dynamic model, the expected time delay on the basis of a median transcript half-life of 11.5 min however is 8 min. Assuming constant degradation rates, the observed short peak time delays could only be explained by very short half-lives in the range of 1-2 min. This is far below any estimate in the literature [60, 116, 65]. For example, the ten cyclins which are found as periodically expressed in our data have a mean peak shift of 0.8 min. The observed short delays between synthesis and total mRNA peaks in periodic transcripts are therefore incompatible with the assumption of constant degradation rates.

6.10. Periodic changes in mRNA degradation shape expression peaks

To investigate the potential role of mRNA degradation rate changes quantitatively, we extended the DTA method such that it allows for the estimation of changes in mRNA synthesis and degradation rates. We exploit the fact that equation 6.1 translates a synthesis time course $\mu(t)$ and a degradation time course $\delta(t)$ into predictions of total and labeled mRNA (Figure 6.14A). This can be used to reverse engineer $\mu(t)$ and $\delta(t)$ from a pair of observed labeled and total mRNA time courses (Figure 6.15A). We model the synthesis rate as a piece-wise linear function, whereas the degradation rate $\delta(t)$ is modeled as sine function (Methods 5.6.2). Note that we did not use the smoothed synthesis rate estimate of MoPS, because MoPS aimed at the detection of periodic expression, and did not take into account changes in mRNA degradation. Moreover, we wanted to exclude any model bias and avoid findings due to slightly biased model assumptions. The measurement error that determines the quality of fit was as in MoPS. The parameters were then fitted to the measured cDTA data by Markov Chain Monte Carlo (Methods 5.6.2). This enabled us for the first time to decompose cell cycle dependent mRNA expression into the processes of mRNA synthesis and degradation.

We further developed a score quantifying the strength of periodic mRNA degradation. It is based on the comparison of two models for the explanation of the labeled and total mRNA time series of a gene. One model assumes a constant mRNA degradation rate, $\delta(t) = const$, and the other assumes a sinusoidal degradation rate, $\delta(t) = a * cos(t - \varphi) + const$. The log likelihood ratio of the respective best fits, termed 'variable degradation score', was used to rank genes according to their fluctuations in mRNA degradation (Methods 5.6.3). The variable degradation score was averaged over both replicate time series. Periodic transcripts had a mean variable degradation score of 0.64 (+-0.47 std.dev), as opposed to

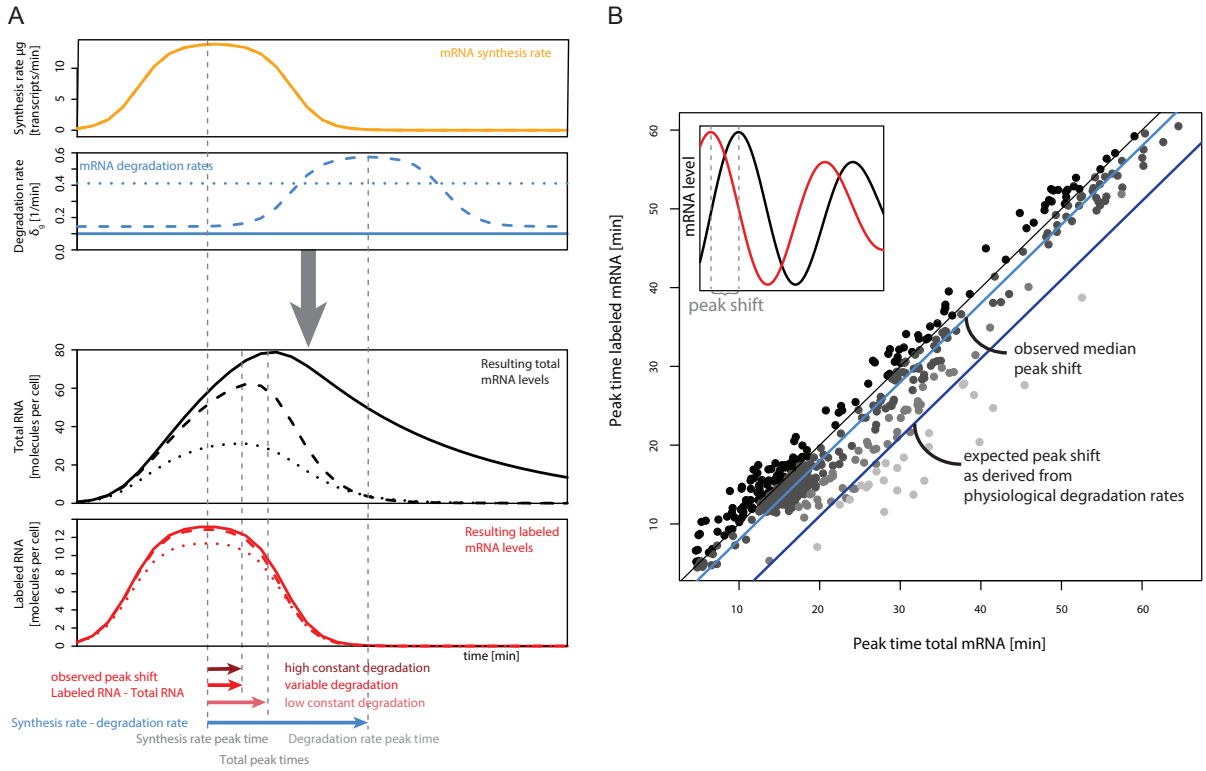

Figure 6.14.: (A) Calculation of labeled and total mRNA time courses as functions of mRNA synthesis and decay. A peak in RNA synthesis rate (top, orange line) translates into different time courses of total RNA concentration (3rd panel, black lines) and labeled RNA concentration (bottom, red lines) according to different time courses of its degradation rate (2nd panel, blue lines). Shown are three realistic degradation scenarios: The solid respectively dotted blue line corresponds to a constant low respectively high degradation rate, the dashed blue line shows a scenario in which degradation peaks a while after the synthesis rate peak. A low (high) constant degradation rate leads to a long (short) peak shift between total mRNA and RNA synthesis rate. A variable, peaked degradation leads to a short peak shift, while at the same time achieving almost the same amplitude of total RNA variation. (B) Scatter plot of labeled vs. total RNA peak time for 479 periodically expressed genes. The distance of a point to the main diagonal measures the peak shift for the corresponding gene (see inset, illustrating the peak shift between a total (black) and labeled (red) mRNA time course). Solid light blue line: observed median peak time delay = 2 min, corresponding to a constant mRNA degradation rate of at least 0.4 (half-life of 1.7 min). Solid dark blue line: expected median peak shift = 8 min corresponding to the average physiological degradation rate $\delta = 0.06$ in *S.cerevisiae* (half-life of 11.5 min).

non-periodic transcripts (mean 0.40, +-0.44 std.dev). Conversely, genes with a variable degradation score above 0.3 comprised 74.7% of all periodic transcripts. Additionally, the variable degradation score was positively correlated with the periodicity score of periodic transcripts (Figure 6.15B, Spearman correlation = 0.2, p<10e-10). This indicates that periodic variation in mRNA degradation is a common feature of periodic transcripts. We chose a conservative score cutoff of 0.3 to call genes with variable degradation. The 479 periodic genes were highly and significantly enriched for genes with variable degradation (odds ratio 3.3, p-value < 10e-10 in a Fisher test). Changes in degradation rates might be confined to a single cell cycle phase or might be gene-specific. We grouped the 358 periodic genes with variable degradation according to the cell cycle phase in which their transcription peaks and examined the distributions of their degradation peaks (Figure 6.15C). It turns out that there is no specific cell cycle phase where the degradation of all transcripts is maximal. Instead, there appears to be a preferential time delay between synthesis peak and degradation peak of 21 min on average (Figure 6.15D).

To examine the influence of the time delay between synthesis and degradation peak time on the amplitude of total mRNA expression, we conducted a simulation. Therefore, the same cosine-shaped synthesis rate was used while shifting the cosine-shaped degradation rate from 0 to 2π(= cell cycle length) (Figure 6.16A). The corresponding total mRNA levels were computed according to Equation (5.25). Due to the periodicity property of our model, the degradation rate peak shift by 2π corresponds to the results where the degradation rate is not shifted, and thus results for the shift by 2π are not shown here. The amplitudes of the total level vary with the shift of the degradation rate from the synthesis rate. The maximal amplitude in the total level is reached when the degradation rate is shifted by $\pi = 30$ min. Figure 6.16B shows the resulting amplitudes and peak time delays between the total RNA level and the synthesis rate for the simulated shifts in degradation rate peak time. For our set of periodic genes with variable degradation the observed peak time delay between synthesis and degradation rate corresponds to 21 min. It turns out that a time delay of 20-30 min strikes an optimum balance between total RNA peak height and peak shift. Thus, a quantitatively high and sharp expression response can be achieved at a much lower degradation rate than for constant RNA degradation. We conclude that periodic changes in mRNA degradation rates are a common, functionally relevant property of periodically expressed genes. Periodic changes in degradation efficiently achieve a sharp peaking of mRNA expression at defined time points during the cell cycle.

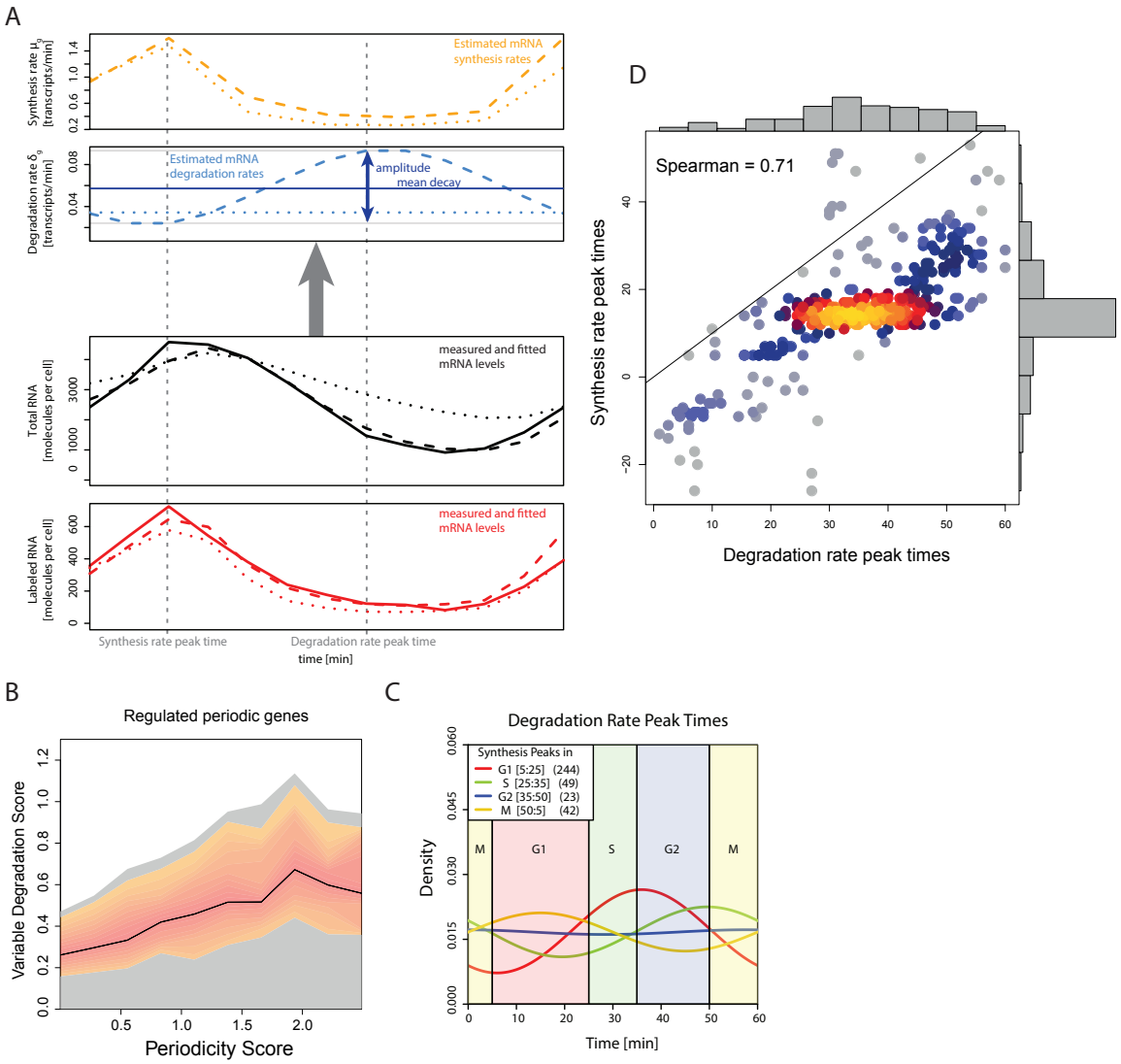

Figure 6.15.: (A) Fitting of the dynamic mRNA turnover model to experimental data. The two lower panels show the measured labeled RNA (solid red line) and total mRNA (solid black line) time courses for YNL312W (RFA2). The time courses were fitted using either a variable degradation scenario (red / black dashed lines) or a constant degradation scenario (red / black dotted lines). In both scenarios, the synthesis rates are estimated by a piece-wise linear function (first panel, dashed / dotted orange lines) and the degradation rates are estimated by a constant respectively a sine function (second panel, dashed / dotted blue lines). (B) Correlation between the periodicity score and the variable degradation score for 2584 genes with periodicity score > 0. Quartiles of the regulated-degradation score distribution (y-axis) are shown as a function of the periodicity score (x-axis). The interquartile range (25%-75% quantile is shown in orange-red, the extreme regions (0%-25% and 75%-100% quantile) are shown in grey, the central black line is the median line (C) Distribution of degradation rate peak times in the cell cycle. The genes are grouped according to the peak time of their synthesis rate (G1:red, S:green, G2:blue, M:yellow). The numbers in brackets correspond to the numbers of genes in each group. Generally the degradation peak is shifted by approximately 20 minutes relative to the synthesis peak. (D) Correlation of degradation rate peak times (x-axis) and synthesis rate peak times (y-axis) for periodically expressed genes with regulated degradation. For genes where the synthesis peaks at the end of a cell cycle and the degradation rate peaks at the beginning of the subsequent cell cycle, the synthesis rates were shifted by -60 minutes.

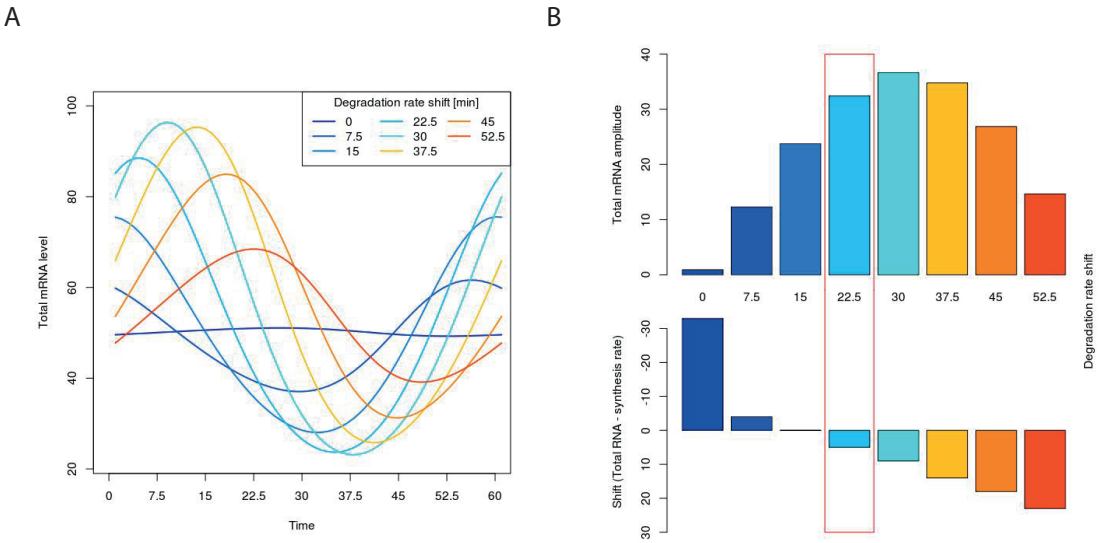

Figure 6.16.: (A) Total mRNA time courses retrieved using degradation rates that only differ in their degradation peak shift relative to the synthesis rate peak. The resulting amplitudes of the total mRNA levels vary. (B) Shifting the cosine shaped degradation rate relative to the synthesis rate results in different amplitudes and peak times of the total mRNA level. Top: Derived amplitudes for the total mRNA expression level. Bottom: Shift between the peak times of the total mRNA levels and the synthesis rate. The red rectangle indicates the bars which correspond to the degradation rate shift which is most similar to the one we observed for our set of regulated periodic genes (21 min).

Part III.

Determinants of RNA metabolism in the Schizosaccharomyces pombe genome

This work has been published in

Philipp Eser*, Leonhard Wachutka*, Kerstin C Maier, Carina Demel, Mariana Boroni, Srignanakshi Iyer, Patrick Cramer, Julien Gagneur. Determinants of RNA metabolism in the Schizosaccharomyces pombe genome. *Molecular Systems Biology* (2016) 12: 857

*these authors contributed equally

7. Methods

7.1. 4tU labeling, RNA extraction and sequencing

All experiments were done with the strain ED666 (BIONEER) (h+, ade6-M210, ura4-D18, leu1-32). A fresh plate (YES) was inoculated from glycerol stock. An over-night culture was inoculated (YES medium) from a single colony and grown at 30 °C. In the morning a 120 mL culture (YEA medium) was started at OD600 0.1 and grown to OD600 of 0.8 at 32 °C in a water bath at 150 rpm. 4-thiouracil was added to 110 mL of culture at 5 mM final concentration. 20 mL samples were taken out after 2, 4, 6, 8, and 10 minutes. Each sample was centrifuged immediately at 32 °C, at 3,500 rpm for 1 min. The supernatant was discarded and the pellet was frozen in liquid nitrogen. All experiments were performed in two independent biological replicates. Total RNA was extracted and samples were DNase digested with Turbo DNase (Ambion). Labeled RNA was purified as published [67]. ribosomal RNA was depleted using the Ribo-ZeroTM Gold Kit (Yeast, Epicentre) according to the manufacturer's recommendation with 1.5 µg labeled RNA and with 2.5 µg total RNA as input. Sequencing libraries for the time series samples were prepared according to the manufacturer's recommendations using the ScriptSeq™ v2 RNA-Seq Library Preparation Kit (Epicentre). Libraries were sequenced on Genome Analyzer IIx (Illumina). See Table 7.1 for an overview of the generated RNA-Seq data.

Table 7.1.: RNA-Seq datasets used to annotate transcribed regions and to study RNA metabolism in *S.pombe*.

Sample	Description	Application	Replicate	RNA fraction	Run mode	# raw reads	# filtered mapped reads	strandedness [%]
A	Steady-state	annotation	1	Total	paired-end	113292612	42904610	97,6
B	Steady-state	annotation	2	Total	paired-end	189347452	56629032	98,0
C	4tU - 2 min	modeling	1	Labeled	single-end	19102425	14977187	96,0
D	4tU - 2 min	modeling	2	Labeled	single-end	20123933	16047260	93,9
E	4tU - 4 min	modeling	1	Labeled	single-end	19715587	15967411	95,1
F	4tU - 4 min	modeling	2	Labeled	single-end	20311245	16261056	94,9
G	4tU - 6 min	modeling	1	Labeled	single-end	5986881	4917973	95,6
H	4tU - 6 min	modeling	2	Labeled	single-end	23557057	18833825	93,7
I	4tU - 8 min	modeling	1	Labeled	single-end	13207441	10895573	92,5
J	4tU - 8 min	modeling	2	Labeled	single-end	18534117	14905967	96,0
K	4tU - 10 min	modeling	1	Labeled	single-end	23565650	19340203	93,3
L	4tU - 10 min	modeling	2	Labeled	single-end	17075745	13611139	95,5
M	Steady-state	modeling	1	Total	single-end	18330704	11973808	95,5
N	Steady-state	modeling	2	Total	single-end	19597405	12056019	94,8

7.2. RNA-Seq read mapping

Single- and paired-end RNA-Seq reads were mapped to the reference *S.pombe* genome (ASM294v2.26) with the splice-aware aligner GSNAP [117], excluding split reads and read pairs longer than 5000 nt, and allowing up to 7 % mismatches. To detect known and novel splice sites, a splice site definition file compiled from the current annotation (Pombase V2.22) was supplied to GSNAP and the probabilistic model to identify splice junctions de novo (*flag –novelsplicing*) was used. After mapping, aligned paired-end reads were further filtered based on SAM flags contained in the alignment files to keep only pairs with proper pairing and orientation (*-f 99, -f 147*). Finally, PCR duplicates were removed with samtools [118] rmdup (standard parameters). Splice sites were identified by using the CIGAR string of all mapped paired-end total RNA reads (replicates added). All sites supported by 10 or more spliced reads were considered for downstream analyzes. As only 48 introns were found with alternative splice sites (with identical start but different end coordinate, or vice-versa), alternative splicing was later on not considered. For each of those 48 splice sites, the splice junction with the highest read support was kept.

7.3. Mapping of Transcriptional Units

To map transcriptional units (TUs) we applied a segmentation algorithm to the paired-end Total RNA-Seq data separately for each strand. The per base coverage was extracted by considering the full fragments of a read-pair, e.g. from the start coordinate of the first read in the pair to the end coordinate of the second read in the pair. The cumulative read coverage vectors over the two biological replicate datasets was considered. The algorithm takes as input a coverage vector and three segmentation parameters: coverage cutoff, minimal length (min-length) and maximal gap (max-gap).

First, all positions in the genome that exceeded the coverage cutoff were marked. Second, non-marked positions that were located between two marked positions that have a separation less than max-gap are also marked. Third, regions with a consecutive number of marked positions greater than min-length were reported. We used the current *S. pombe* annotation (Pombase V2.22) to estimate suitable values for the three segmentation parameters in our data: first, a coverage cutoff was estimated by an approach similar to [119]. The distribution of the per base coverage between current annotations was modeled as a bi-modal distribution consisting in few non-annotated transcribed regions and a majority of non-transcribed (background) regions. The background-region distribution was modeled as a Gaussian distribution. The mode m of the background-region distribution was set to the median of the whole distribution. The variance of the background-region distribution was estimated as the variance of the distribution that is obtained by mirroring the part of

the mixture with values lower than the mode m about the axis $y = m$. The cutoff (10.26) distinguishing background from transcribed regions was then set to a one-sided nominal p-value of 0.01 for the fitted Gaussian.
Second, min-length and max-gap are estimated simultaneously in an exhaustive search over all combinations of min-length and max-gap values between 10 and 1,000: for each combination, a segmentation of the genome was performed and scored based on the overlap with transcripts of the current annotation. We used the Jaccard index as a similarity measure, which is defined as the size of the intersection divided by the size of the union of the two sets. The Jaccard index reached maximal values for min-length and max-gap in the range between 50 and 500. Since there was no single optimal combination and the *S. pombe* transcriptome is very dense, we chose rather small parameters with min-length 200 and max-gap 80. With these parameters (coverage cutoff = 10.26, min-length = 200, max-gap = 80), the segmentation resulted in 7,062 TUs. To further improve this map of TUs, we only kept TUs that showed significant read coverage (average per base coverage < 20 in the two-minute labeled 4tU-Seq samples, normalized for sequencing depth using annotated ORFs read counts and following [120]. This filter resulted in a high-confidence set of 5,596 TUs.

These TUs are then classified according to the overlap with known transcripts: 112 partially overlapped ORFs and were discarded for further analysis. The remaining final set of 5,484 TUs were classified into four disjoint classes: i) ORF-TUs entirely contain one ORF only and not more than 70% of any annotated ncRNA ii) nc-TUs do not contain entire ORFs, overlap at least 70% of an annotated ncRNA and not more than 70% of any other annotated ncRNA iii) Novel nc-TUs do not overlap by more than 70% any annotated ncRNA and do not overlap any ORF iv) multicistronic TUs contain multiple ORFs entirely or overlap 70% of two or more annotated transcripts.

7.4. Read counts per exon, intron and splice junctions

Counts of reads aligning completely within exons or introns were obtained with the software HTseq-counts [121] with settings *–stranded=yes* and *-m intersection-strict*. To count reads that map to splice sites we used HTseq with one different parameter (*-m union*) to allow counting of reads that spanned the junctions. For each intron, we defined the 5'SS as the 2 nt region that contains the last position of the upstream exon and the first position of the intron. Accordingly, we defined the 3'SS as the 2 nt region that contains the last position of the intron and the first position of the downstream exon. To distinguish spliced and unspliced junction mapping reads, a custom script checked the cigar string of each alignment for occurrences of skipped reference bases ("N"). Alignments containing "N" and overlapped with a splice site, were counted as spliced junction reads.

7.5. Estimation of RNA metabolism rates from 4tU-Seq data

7.5.1. Overview

We used a probabilistic model that relates read counts of some kind (exonic reads, spliced and unspliced junction reads) to a set of model parameters $\mathbf{\Theta}$ which includes the RNA metabolism rates and technical nuisance parameters. With casual notations, we modeled the probability of observing read counts k of one kind in one sample as $p(k|\mathbf{\Theta}) = \mathrm{NB}(k|\mathrm{mean}(\mathbf{\Theta}), \mathrm{dispersion})$, where NB() is the negative binomial distribution. Subsections 7.5.2 - 7.5.5 model the RNA species concentrations in the sequenced samples and subsection 7.5.6. models the expected number of reads sequenced given these concentrations. This gives mean($\mathbf{\Theta}$). Subsection 7.5.7 describes the parameter estimation procedure.

7.5.2. Junction Model

For a given junction, let [**precursor RNA**] be the cellular concentration of the unspliced RNA and [**mature RNA**] the cellular concentration of the spliced RNA. With synthesis rate μ, splicing rate σ and degradation rate λ the following ODEs describe the dynamic of the system assuming first-order kinetics:

$$\begin{aligned}\frac{d[\textbf{precursor RNA}]}{dt} &= \mu - \sigma[\textbf{precursor RNA}] \\ \frac{d[\textbf{mature RNA}]}{dt} &= \sigma[\textbf{precursor RNA}] - \lambda[\textbf{mature RNA}]\end{aligned}$$

with following initial conditions:

$$\begin{aligned}[\textbf{precursor RNA}]_{\textbf{labeled}}|_{t=0} &= 0 \\ [\textbf{mature RNA}]_{\textbf{labeled}}|_{t=0} &= 0 \\ [\textbf{precursor RNA}]_{\textbf{unlabeled}}|_{t=0} &= \frac{\mu}{\sigma} \\ [\textbf{mature RNA}]_{\textbf{unlabeled}}|_{t=0} &= \frac{\mu}{\lambda}\end{aligned}$$

Under the assumption that introduction of labeled Uracils in the media ($t = 0$), the solutions are:

$$
\begin{aligned}
[\textbf{precursor RNA}]_{\textbf{labeled}}(t) &= \frac{\mu}{\sigma}(1 - e^{-\sigma t}) \\
[\textbf{mature RNA}]_{\textbf{labeled}}(t) &= \frac{\mu}{\lambda(\lambda - \sigma)}(\lambda(1 - e^{-\sigma t}) - \sigma(1 - e^{-\lambda t})) \\
[\textbf{precursor RNA}]_{\textbf{unlabeled}}(t) &= \frac{\mu}{\sigma}e^{-\sigma t} \\
[\textbf{mature RNA}]_{\textbf{unlabeled}}(t) &= \frac{\mu}{\lambda(\lambda - \sigma)}(\lambda e^{-\sigma t} - \sigma e^{-\lambda t})
\end{aligned}
$$

7.5.3. Exon model

For single-exon TUs, there is no processing. Following the same rate notations, we obtain the same kinetics as for the precursor RNA:

$$
\begin{aligned}
[\textbf{exon RNA}]_{\textbf{labeled}}(t) &= \frac{\mu}{\sigma}(1 - e^{-\sigma t}) \\
[\textbf{exon RNA}]_{\textbf{unlabeled}}(t) &= \frac{\mu}{\sigma}e^{-\sigma t}
\end{aligned}
$$

7.5.4. Uracil Bias

Not all uracils available to the transcription machinery are labeled, leading to a labeling bias against transcripts with a small number of Us [60]. Following Miller et al. [60], the probability $p(\text{4tUI})$ that one transcript incorporates at least one 4tU was modeled as:

$$p(\text{4tUI}) = 1 - (1 - p(\text{4tU replaces U}))^{\text{Number of U in transcript}}$$

This correction was difficult to apply to the junction model because of all possible RNA variants (isoforms, precursor and mature RNAs) overlapping the junction. However, we found that a U-bias correction would have negligible effects for intron-containing TUs because even their mature RNAs were generally containing many Us (short TUs were almost all single-exon). Hence, for typical values of $p(\text{4tU replaces U})$, $p(\text{4tUI})$ was very close to 1 for intron-containing TUs. In the following, U-bias correction was only applied to the exon model, which became:

$$[\textbf{exon RNA}]_{\textbf{labeled}}(t) = p(4\text{tUI})\frac{\mu}{\sigma}(1 - e^{-\sigma t})$$
$$[\textbf{exon RNA}]_{\textbf{unlabeled}}(t) = \frac{\mu}{\sigma}e^{-\sigma t} + (1 - p(4\text{tUI}))\frac{\mu}{\sigma}(1 - e^{-\sigma t})$$

7.5.5. Cross-contamination

What we measure is the purified and the not purified (so-called total) fractions of RNA. Measurements are sensitive to small amount of cross-contamination of unlabeled RNAs in the purified fraction, because unlabeled RNAs can represent the vast majority of RNAs especially at early time points. Thus, we introduced a cross-contamination factor χ that we assumed to be common to all RNA species for simplicity. Up to sample-specific factors common to all RNA species, the concentration of purified and not purified RNA relates to the RNA cellular concentrations as:

$$[\textbf{purified RNA}] = (1 - \chi)[\textbf{labeled RNA}] + \chi[\textbf{unlabeled RNA}]$$
$$[\textbf{not purified RNA}] = [\textbf{labeled RNA}] + [\textbf{unlabeled RNA}]$$

7.5.6. Expected number of reads given RNA species concentrations

7.5.6.1. Expected number of reads

Let $x_{i,j}$ be the concentration of feature i in sample j (e.g. [**purified precursor RNA**] is the concentration of the feature 'unspliced read' in labeled samples). The expected counts $k_{i,j}$ of feature i in sample j was modeled as:

$$E(k_{i,j}) = F_j N_i x_{i,j}$$

where F_j is sample-specific scaling factor (see below) and N_i is the effective length of feature i (see below).

7.5.6.2. Controlling for overall amount of labeled RNA and sequencing depth

The RNA sequencing protocol requires a constant amount of starting material and yields approximately the same number of reads per sample. Hence, the overall increase of labeled RNA over time is not reflected in the total amount of reads obtained. Therefore,

normalization of the samples relative to each other had to be performed using sample-specific factors F_j. This normalization factor also allows controlling for variations in sequencing depth.

7.5.6.3. Controlling for TU length

The exon model is based on all reads overlapping the exon, and therefore depends on the exon length, yet not in a simple proportional fashion. Indeed, purified transcripts are sonicated into fragments of a typical length, in our case about 200bp (mean fragment length, the actual number is not essential, it is only used in the derivation step). For asymptotically long transcripts, the expected number of fragments per transcript is:

$$N_i \approx \frac{\text{length of transcript } i}{\text{mean fragment length}}, \text{for } i \text{ long intronless TU}$$

However, this approximation fails for short transcripts. Indeed, sonication of short transcripts (about less than 2 times the mean fragment length) leads to a large fraction of very short fragments that are selected against during library preparation and do not get sequenced. Hence, to model the relation between fragment length and expected number of sequenced fragments for the whole range of transcript lengths, we empirically used a linear approximation that includes an offset L_{off} (see estimation below). This led us to an effective length such that:

$$N_i = \frac{\text{length of transcript } i + L_{\text{off}}}{\text{mean fragment length}}, \text{for } i \text{ intronless TU}$$

In contrast, the junction model relies on spliced and unspliced reads that overlap junctions. Reads that overlap a junction satisfy two criteria: i) they originate from fragments that overlap the junction and ii) the reads themselves overlap the junction. Junctions are typically further away from transcript ends compared to the fragment length. We therefore assumed that the expected number of possible fragments satisfying criterion i) is the same for all junctions genome-wide. Criterion ii) implies that the effective length for the junction model is proportional to read length. Matching junction model and exon model estimates to the same scale was achieved by setting the effective length of the junction model to read length (78bp in our case) over mean fragment length:

$$N_i = \frac{\text{read length}}{\text{mean fragment length}}, \text{for } i \text{ spliced or unspliced junction}$$

7.5.7. Parameter estimation

In the following we develop a method for estimating all parameters based on the observed count data by maximizing the likelihood.

Assuming negative binomial distribution of RNA-Seq read counts [120], the log likelihood reads as:

$$ll = \sum_{i,j} \log(\mathrm{NB}(k_{i,j}|E_{i,j}(\Theta), \alpha) \tag{7.1}$$

where Θ is the set of parameters $\{\mu_i, \sigma_i, \lambda_i, F_j, \chi, L_{\mathrm{off}}, p(4tUreplacesU)\}$ for all junctions or exons i and for all samples j, and where α is the dispersion parameter of the negative binomial. We assumed that the dispersion parameter is uniform over all samples and features, which we believe is a reasonable assumption.

7.5.7.1. Estimation of the dispersion parameter

Due to the complexity of the model and the large number of parameters it is not practically feasible to directly optimize the log likelihood. In a first step the mean $E_{i,j}$ of each data point (a data point is given by the number of reads belonging to one transcriptional feature e.g. exonic reads, junction reads at one time point) between the two replicates was computed. In a second step, the dispersion α was fitted by maximum likelihood letting the $E_{i,j}$ fixed:

$$\alpha = \mathrm{argmax} \sum_{i,j} \log(\mathrm{NB}(k_{i,j}|E_{i,j}, \alpha)) \tag{7.2}$$

Then the actual model was fitted using this value of α as fixed parameter. The expected counts obtained by this model were used again with (2) to get an improved estimate for α. Two rounds of iterations showed that α is a stable parameter and does not differ much from the first order guess (about 10% change). Forced changes of α by factor of 10 and 0.1 showed that the actual model parameters Θ are quite robust against variation of α, since the estimated rates did not change significantly (relative changes 10^{-4}). Hence, we did not increase the number of iterations.

7.5.7.2. Overall estimation procedure

After extensive testing and numerical simulations, we found that the best results were obtained using the following procedure.

Transcripts with a length less than 120 base pairs were excluded from the analysis because of insufficient coverage, as the read length itself comprises 80 base pairs. We used the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm using the R function `optim()` with analytical gradient. We actually maximized the logarithm of the log likelihood, which turned out to give more reproducible results.

1. Start with $L_{\text{off}} = 0$,$p(4tUreplacesU) = 1$, $\chi = 0$, and, and keep them as fixed parameters in the BFGS optimization procedure.
2. Select the 350 "best" intronless genes (in terms of coverage by visual inspection in IGV) and apply BFGS with the exon model (because we have the least number of assumptions here). The fitting is repeated 100 times using different start parameters, which allows us to estimate the robustness of the model. In this way we were able to extract the set of normalizing constants F_j for each sample (median of the fits). Because they are relative quantities, we deliberately set $F_1 = 1$.
3. Run the exon model and the junction model using the F_j as fixed input parameters. Each fit is done with 100 different initial values. Define the median of the calculated parameters as estimate.
4. Repeat step 2 and 3 using different values for the cross-contamination ranging from 0 to 5 % (independent experiments with spike-ins motivated this range), different L_{off} and $p(4tUreplacesU)$. These two parameters were set according to criteria described below.

Criteria for step 4 are as follows. Under the assumption that all junctions within one gene should have the same synthesis rate we chose the level of cross-contamination with the best correlation of the synthesis rates between the first and second junction of genes with 2 or more introns. When setting $p(4tUreplacesU)$ for the exon model so that these correlations match each other, we observed no significant correlations between synthesis rate and gene length. This result was in agreement with the junction model, for which no correlation between synthesis rate and gene length was found either. Moreover, the best value for $p(4tUreplacesU)$ was 1%, which is strictly within the expected range and close to the value of 0.5% estimated by [60] when profiling *S. cerevisiae*. Nonetheless, one should keep in mind that the lack of correlation between synthesis rate and gene length for intronless genes in our data is due to a modeling assumption and not a result of our investigations. The results were used to improve our estimate of the dispersion parameter and steps 1 to 4 were repeated to improve the model parameters even further.

7.5.8. Rescaling of synthesis rate

The two quantities [purified RNA], [not purified RNA] are both linear in μ. Hence, the synthesis rate can only be estimated up to a global constant. We therefore arbitrarily set

$F_1 = 1$ for the fitting. Absolute synthesis rates were then obtained by scaling all values so that the median steady-state expression level of ORF-TUs matches the one reported by a genome-wide absolute quantification study (median of 2.4 coding mRNAs per cell, Marguerat et al. [64]).

7.6. Identification of sequence elements predictive for rates and linear regression

The goal of this procedure was to identify sequence elements predictive for a given rate of interest (synthesis, splicing, or degradation), in a given gene region of interest (promoter for all TUs and 5'UTR, coding sequence, introns, 3'UTR for ORF-TUs) and to estimate coefficients for each nucleotide at each position of these sequence elements. The procedure consisted of two consecutive stages 'seed finding' and 'seed extension and regression'.

The output of the 'seed finding' stage are initial sequence elements that associate with the rate. To this end, a linear mixed model was considered to assess the effect of each possible 6-mer in turn, while controlling for random effects over all 6-mers. We followed here an idea proposed by Liyang et al. [122] to estimate the activity of microRNAs. Formally, the effect of the j-th 6-mer on the rate was modeled according to:

$$\begin{aligned} y &= W\alpha + x_j\beta_j + u + \epsilon \\ p(u) &= N(u|0, \lambda\tau^{-1}K) \\ p(\epsilon) &= N(\epsilon|0, \tau^{-1}I) \end{aligned}$$

, where y is a n-vector of rates over all n TUs (respectively splice sites), W is an optional $n \times c$ matrix of covariates, α is the corresponding vector of coefficients, x_j is the n-vector of the number of instances of the j-th 6-mer in the region over all TUs (respectively splice sites), β_j is its coefficient, u is a n-vector of random effects, and ϵ is the n-vector of errors. For all rates, we considered as covariate the unit vector in order to model an intercept. We also considered as covariate the length of the 3'UTR in the case of the degradation rate, which we had found to be significantly associated with degradation. For other rates, no further covariate was used. The covariance matrix K was set to $X^T X$ where X, whose columns are the x_j, is the matrix of 6-mers counts. The covariance on the random effects allows controlling for the effects of all other 6-mers. The model was fitted using the GEMMA software [123]. All 6-mers significantly associated with the rate (FDR <0.1, likelihood-ratio test with Benjamini-Hochberg correction for multiple testing) were retained. If both a 6-mer and its reverse complement were found significant, the two were considered as a single unstranded 6-mer, and the other ones as stranded 6-mers. Significant 6-mers overlapping by all but one or two base (eg. TTAATG and TAATGA)

and sharing more than half of their genome-wide instances reciprocally were recursively assembled into longer k-mers (in this example TTAATGA). This procedure led to stranded and unstranded k-mers that we coined 'seed'.

The goal of the 'seed extension and regression' stage is to extend seeds to cover neighbor nucleotides significantly and to estimate the effect of each nucleotide. This is achieved with the following iterative procedure:
1. Initialization: The 'sites' are initialized by all elements in the region of interest matching the seed up to one mismatch (two mismaches for the long HOMOL-box motifs) together with 2 nucleotides 5' and 2 nucleotides 3' of it. For the unstranded motifs (two homol boxes) we also considered the reverse complements of the motifs as match.

2. Linear regression: We denote by n_i the number of sites for the i-th TU and L the length of a site. The 'consensus' sequence is defined as the sequence of the position-wise most frequent nucleotides over all sites. The following linear model is fitted by maximum likelihood:

$$y_i = \beta_0 + \sum_{j=1}^{n_i} \beta_{\text{cons}} + \sum_{k=1}^{L} \beta_{k,w_{i,j,k}} + \epsilon_i$$

where β_0 is the intercept, i.e. the average level in the absence of any site, β_{cons} is the effect of one consensus site, $w_{i,j,k}$ is the k-th nucleotide of the j-th site of the i-th TU, and $\beta_{k,A}$, $\beta_{k,C}$, $\beta_{k,G}$, $\beta_{k,T}$ are the effects of each nucleotide at position k relative to the nucleotide of the consensus site at the same position. By definition $\beta_{k,w}$ is constrained to be 0 if w equals the k-th nucleotide of the consensus sequence. The errors ϵ_i are assumed to be independently and identically normally distributed. Reverse complemented motifs enter in their canonical form.

3. Extension: For each site considered in step 2, its overall effect $\beta_{\text{cons}} + \sum_{k=1}^{L} \beta_{k,w_{i,j,k}}$ is tested to be significantly different from 0 ($P < 0.05$). To compute the p-value we evaluate the multivariate t-statistic (using glht of the multcomp package in R). A position weight matrix (PWM) is constructed based on all significant sites extended by 2 nucleotides 5' of the 5'-most significant position and 2 nucleotide 3' of the 3'-most significant position. To construct the PWM, the genomic nucleotide distribution is taken as background distribution. The sequences significantly matching the PWM (P> 0.80, multinomial model with a Dirichlet conjugate prior) are considered as the new sites. Step 2 and 3 are repeated until sites do not get extended in length. This is decided by visual inspection of the obtained PWM (the extended bases equal the background distribution). It turned out that the extension stage was only necessary and useful for the two well conserved Homol boxes.

The final motif sequence we report is the consensus sequence of the motifs. We searched again in the regions of interest for them (allowing 1 mismatch for all but the two HOMOL-Boxes (2 mismatches)). We finally applied the same linear model as in point 2. on these found sites to obtain the final coefficients.

7.7. Validation of sequence model using an eQTL dataset

We compared fold-change associated with local genetic variants in Clément-Ziza et al. [124] with the predicted effects from the sequence-to-rate model described in section 7.6.

7.7.1. Read counts

This study profiled steady-state RNA levels and not the newly synthesized RNAs. Hence, the coverage on introns was too poor to perform accurate quantification of the precursor RNAs. We thus focused on the quantification of steady-state levels of mature RNAs of our TUs. To this end, RNA-Seq data from recombinant *S. pombe* strains libraries [124] were downloaded from ArrayExpress (E-MTAB-2640). Genetic variants and strain genotypes were obtained as supplementary Datasets from the manuscript. RNA-Seq reads from each strain were mapped separately to the reference genome using STAR (version 2.4.0i) Dobin et al. [125] with default options. We considered for further analysis $k_{i,j}$, the number of reads overlapping at least one exon for TU i in sample j.

7.7.2. Fold change associated with local genetic variants

The read counts $k_{i,j}$ defined above were modeled according to the following generalized linear model:

$$\begin{aligned} k_{i,j} &\sim \mathrm{NB}(\mu_{i,j}, \alpha_i) \\ \mu_{i,j} &= s_j \times q_{i,j} \\ \log_2(q_{i,j}) &= \beta_i^0 + \beta_i^{local} g_{i,j} + \sum_{b,batch} \beta_b^{batch} x_{j,b}^{batch} \end{aligned}$$

where NB is the negative binomial distribution, α_i is a gene-specific dispersion parameter; s_j is the size factor of sample j; $g_{i,j}$ is the genotype (0 for the reference allele, 1 for the alternative allele) at the variant of interest for gene i in sample j; $x_{j,b}^{batch}$ is 1 if sample j is from batch b and 0 otherwise. The model was implemented with the R/Bioconductor package DESeq2 [120], which provides robust estimation of the size factors, of the dispersion parameters and the fold changes. The log-fold change of interest, β_i^{local}, together with its standard error, was then considered for further analysis. Effect of batches, reported

in the original study, were dominating the signal and important to control for. We also investigated controlling for hotspots (8 eQTL hotspots were reported in the original study) but this led to an increased variance for little bias reduction.

7.8. Multivariate analysis of splicing time

We performed linear regression of log splicing time of each junction against i) all the nucleotides of 5'SS, the BS and the 3'SS region, ii) TU log-synthesis time, iii) the TU length and iv) the number of introns in the TU. First regression was done against each covariate individually. Then, a joint model was build incrementally including each covariate in this order. Fraction of explained variance for both procedures are reported in table 7.2.

Table 7.2.: Fraction of explained variance in splicing rates.

Covariate	Individual	Incremental
Sequence	0.50	0.50
log-synthesis time	0.45	0.42
TU length	0.09	0.07
Number of introns	0.034	0.0024

Except for the number of introns, all covariates contributed approximately equally in the individual and in the incremental model, showing that they are independently predicting synthesis time. In contrast, the number of introns did not added explained variance, likely because the predicted splicing time already correlated with number of introns. To determine the important nucleotides of the 5'SS, the BS and the 3'SS region, we used cross-validation. We started with all nucleotides +/-10 of the 5'SS, the BS and the 3'SS and decreased the window sizes systematically in several reduce steps. First we increased the starting position of the 5'SS -10 to -9,-8,-7, ... until cross-validation showed a loss in predictive power. Then we decreased the 5'SS +10 position in a similar manner. We continued analogously with the BS and 3'SS. We also used several different orders of removing the nucleotides (e.g. starting with 3'SS), to assert that we get similar results which are not biased by the order we apply the reduce step. We used 10-fold cross-validation. We trained the model on 9 parts and validated on the 10th part. Hereby we received a set of 10 models. To report the accuracy of our estimates we use the standard deviations of the coefficients reported by each model, the reported coefficients are the median of all 10 models.

8. Results and Discussion

8.1. Yeast as a model organism to study eukaryotic mRNA metabolism

The fission yeast *Schizosaccharomyces pombe* (*S. pombe*) is an attractive model organism to study eukaryotic RNA metabolism. *S. pombe* shares important gene expression mechanisms with higher eukaryotes that are not prominent or even absent in the budding yeast *S. cerevisiae*. These include splicing, which occurs for ~50% of the genes and is achieved with conserved spliceosomal components [126] and conserved consensus splice site (SS) sequences [127, 128], heterochromatin silencing [129], and RNA interference [130]. Because of its relevance for studying eukaryotic gene expression, S. pombe has been extensively characterized by genomic studies, and this led to an annotation of transcribed loci that includes ncRNAs [131, 132, 133], a map of polyadenylation sites [54, 134], the 'translatome' as measured by ribosome profiling [135], and an absolute quantification of protein and RNA [136].

8.2. Strategy to study RNA metabolism and regulatory elements in *S.pombe*

We used the fission yeast *S. pombe* as a model system to quantify RNA metabolism genome-wide, to identify genomic regulatory elements at single-nucleotide resolution, and to quantify the contribution of these elements to the kinetics underlying RNA metabolism. Our approach consists of three steps (Figure 8.1). First, we performed short and progressive metabolic labeling of RNA with 4-thiouracil coupled with strand-specific RNA-Seq (4tU-Seq, section 7.1). With the use of advanced computational modeling, we obtained accurate estimates of RNA synthesis and degradation rates for 5,484 transcribed loci and splicing rates for 4,958 splice sites. Second, a novel statistical modeling procedure quantifies the contribution of each single nucleotide in predicting RNA metabolic rates and thereby identifies sequence features that contribute to RNA metabolism rates. We then supported a causal role of these features by comparing RNA expression fold-changes between strains

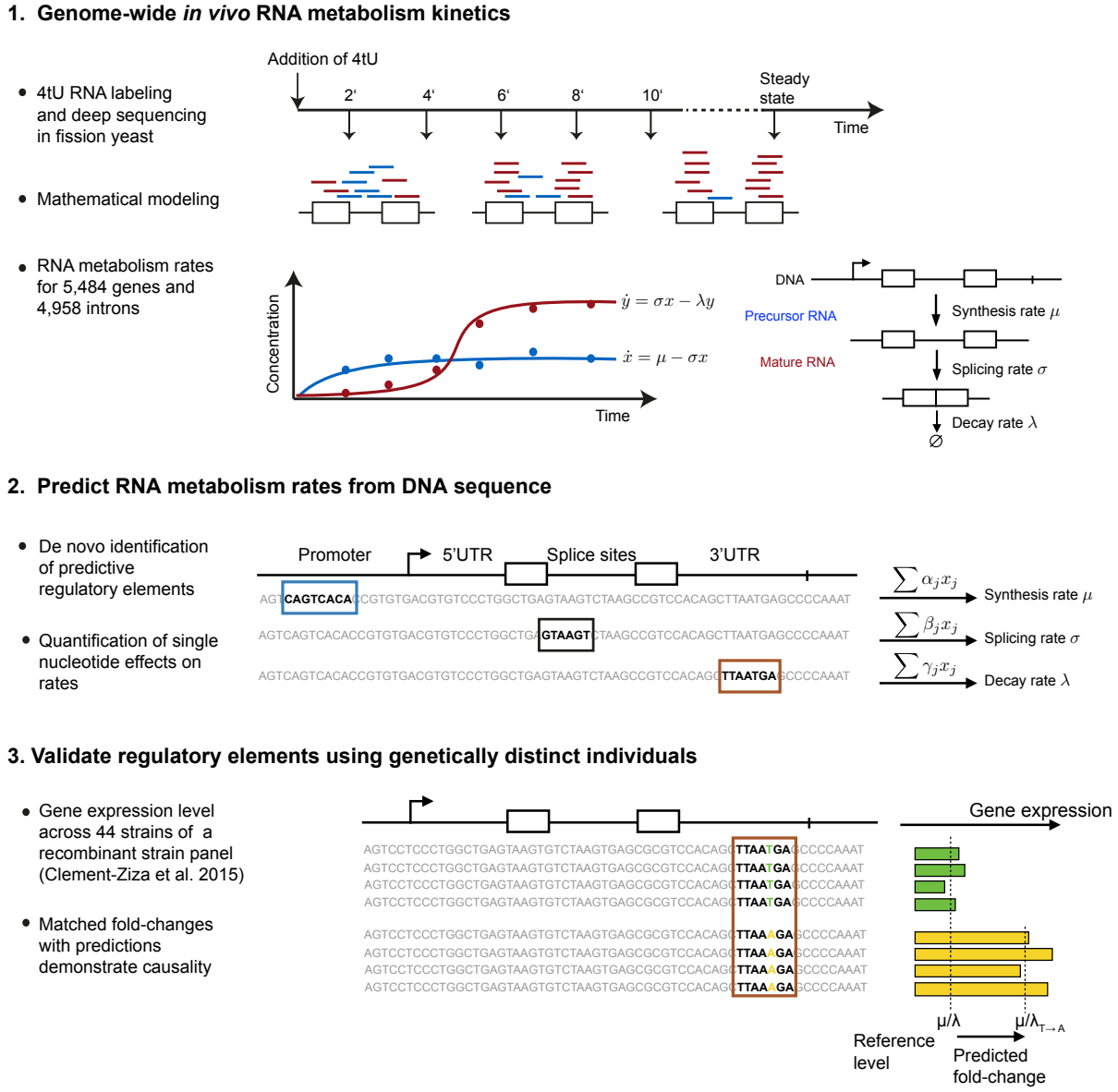

Figure 8.1.: Overview of the approach. Our approach for identifying regulatory elements that quantitatively determine RNA metabolism rates consists of three steps. In step 1 (top), genome-wide estimate of in vivo synthesis, splicing and degradation rates are obtained from the analysis of 4tU RNA labeling time series. In step 2 (middle), sequence motifs (colored boxes) that are predictive for each rate are identified. The method provides for each motif and each nucleotide in a motif an estimate of its quantitative contribution to the rate. In step 3 (bottom), the elements identified in step 2, which might be predictive by mere correlation, are tested for causality. To this end, ratio of average expression levels in a population harboring the reference allele versus a population harboring a single nucleotide variant are compared to model-predicted fold-change.

differing by a single nucleotide at these sites with the corresponding fold-changes predicted by the model. Our approach relies on an accurate annotation of the genome. In particular, accurate transcript boundaries are important for quantifying RNA metabolism. We therefore first set out to precisely define the transcriptional units in *S. pombe*.

8.3. Mapping transcriptional units in S. pombe

To map transcribed regions in the *S. pombe* genome, we carried out strand-specific, paired-end deep sequencing of total RNA (RNA-Seq) from fission yeast grown in rich media (section 7.1). Genomic intervals of apparently uninterrupted transcription (Transcriptional Units, TUs, Figure 8.2A) were identified with a segmentation algorithm applied to the RNA-Seq read coverage signal (section 7.3). The three parameters of the algorithm, the minimum per base coverage, the minimum TU length and the maximum gap between TUs, were chosen to best match the existing genome annotation (Pombase version 2.22 [137], Figure 8.2B,C). TUs that did not show significant signal in the 4tU-Seq dataset were considered as artifacts and discarded (section 7.3). The segmentation led to a total of 5,484 TUs (Figure 8.2D), of which 4,105 were containing a complete, annotated open reading frame (ORF-TU), 1,014 were non-coding TUs (ncTU), and the remaining 365 TUs contained two or more annotated adjacent transcripts, and thus may be multicistronic RNAs. Only a small number of novel splice sites were identified (148 out of 4,958), and no evidence for alternative splicing or circular RNAs was found, in line with previous RNA-Seq studies of *S. pombe* [132]. A total of 402 ORFs (8%) in the existing annotation were not recovered.

8.4. Significantly revised *S. pombe* genome annotation

The resulting annotation of ncTUs in *S. pombe* differed largely from the current one. We identified 487 novel ncTUs, changed the boundaries by more than 200 nt of 422 (27%) previously annotated ncRNAs and could not recover 1011 (66%) of the previously annotated ncRNAs (Figure 8.3A). A large fraction of the latter apparently represent spurious antisense RNAs that are often generated with conventional protocols, but their generation was suppressed here with the use of actinomycin D [138]. Indeed, 49% of those non-recovered ncRNAs were located antisense to highly expressed ORF-TUs and showed on average 66-fold higher antisense than sense coverage (Figure 8.3B). Thus, we redefined the location and boundaries of most ncRNAs in *S. pombe*, leaving only 105 of the currently annotated ncRNAs unchanged.

We also redefined boundaries for 1,481 coding transcripts that differed from the existing annotation by at least 200 nt. Untranslated regions (UTRs) of ORF-TUs were generally much shorter than previously annotated (mean difference 91 nt). This difference apparently also stemmed from spurious antisense RNAs in previous datasets because 68% of the 376 3'UTRs that were at least 250 nt shorter in our annotation showed higher antisense than sense coverage (Figure 8.3C). Our revised transcript 3'-ends were centered around experimentally mapped polyadenylation (polyA) sites [54], whereas the previously annotated

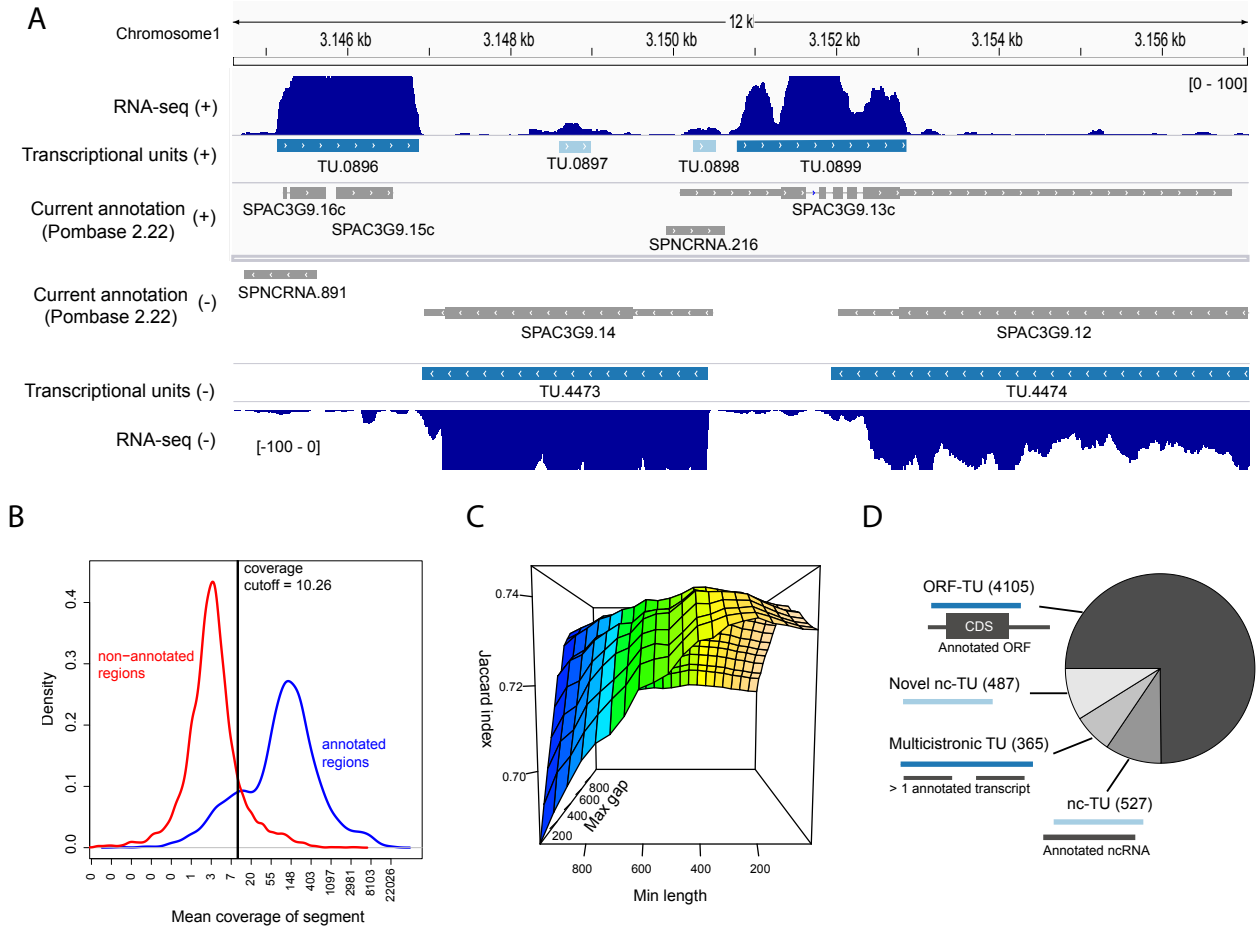

Figure 8.2.: Improved annotation of transcribed loci. (A) Example of transcribed loci annotation on a 12 kb region of chromosome 1. Data is displayed symmetrically in horizontal tracks for the plus strand (upper half) and the minus strand (lower half). From most external to most central tracks: RNA-Seq per base read coverage (marine blue), identified ORF-TUs (dark blue) and identified ncTUs (light blue), currently annotated ORFs and ncRNAs (gray, Pombase v2.22). On the current annotation tracks, UTRs are marked as thin rectangles and introns as lines. Typical changes that this study provides to the current annotation include the merging of adjacent transcripts into a single TU (e.g. SPAC3G9.16c and SPAC3G9.15c into TU.0896), the identification of novel ncRNAs (e.g. TU.0897), not recovered ncRNAs (SPNCRNA.891) and correction of aberrant long UTRs (e.g. 3'UTR of SPAC3G9.13c as TU.0899). (B) Distribution of mean RNA-Seq read coverage per segments for currently annotated (blue) and not currently annotated regions (red) and mean coverage cutoff for the segmentation algorithm to call a region expressed (vertical line). (C) Jaccard index (z-axis) when computing per base overlap between automatic segmentation and current annotation versus min-length and max-gap parameters of the segmentation algorithm. (D) Classification of the 5,484 TUs into ORF-containing (ORF-TUs), nc-TUs overlapping 70% of an annotated ncRNA (nc-TU), TUs overlapping more than one annotated TU (multicistronic TUs), and novel non-coding (Novel nc-TU).

3'-ends typically extended well beyond polyA sites (median difference = 3 nt versus 45 nt, Figure 8.3D). Thus our map of TUs provides a significantly revised annotation of the S. pombe genome that removes false positive ncRNAs from the current annotation and shortens aberrant long UTRs.

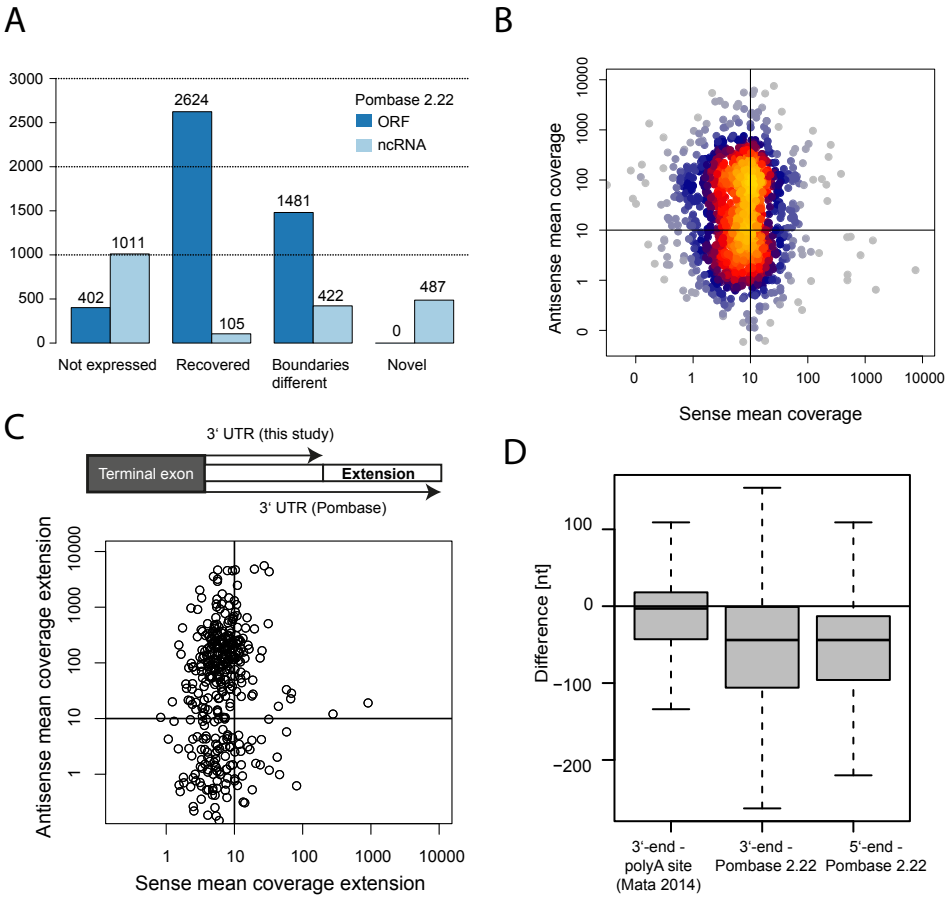

Figure 8.3.: Comparison of our annotation with Pombase. (A) From left to right: number of currently annotated transcripts that could not be recovered, are fully recovered, differ by more than 200 nt, and novel TUs for ORFs (dark blue) and ncRNAs (light blue). (B) Sense mean coverage (x-axis) versus antisense mean coverage (y-axis) of 1011 non-recovered ncRNAs of the current annotation. (C) Mean sense coverage (x-axis) and antisense coverage (y-axis) of Pombase 3'UTR regions that extend TU defined 3'UTRs by 250nt or more. Per base coverage is extracted from total RNA-Seq data used in this study. The mass of the data in upper left quadrant indicate that long Pombase UTRs mostly arise from antisense artifacts in former studies. (D) Differences between 3' ends of ORF-TUs and polyA-sites mapped by Mata et al. (2013) (left), between 3' ends of ORF-TUs and the corresponding currently annotated 3'UTR end (middle), and between 5' ends of ORF-TUs and the corresponding currently annotated 5'UTR end (right).

8.5. Quantification of S. pombe RNA metabolism

To quantify the kinetics of RNA synthesis, splicing, and degradation genome-wide, we sequenced newly synthesized RNA after metabolic RNA labeling with 4-thiouracil (4tU-

Seq) and used the obtained data for kinetic modeling (Figure 8.1, step 1). In cells, the nucleobase 4tU gets efficiently converted to thiolated UTP and incorporated during transcription into newly synthesized RNAs, which can then be isolated and sequenced. To cover the typical range of synthesis, splicing, and degradation rates, cells in a steady-state culture were harvested after 2, 4, 6, 8, and 10 minutes following 4tU addition. The data contained many reads that stemmed from intronic sequences and reads comprising exon-intron junctions, showing that 4tU-Seq captured short-lived precursor RNA transcripts (section 7.4). These reads from unspliced RNA gradually ceased during the time course (Figure 8.4A,B), indicating that the kinetics of RNA splicing may be inferred from the data.

To globally estimate rates of RNA synthesis, splicing, and degradation, we used a first-order kinetic model with constant rates that describes the amount of labeled RNA as a function of time (Figure 8.4C). We modeled splicing of individual introns, where splicing refers to the overall process of removing the intron and joining the two flanking exons. The model was fit to every splice junction using the counts of spliced and unspliced junction reads (Figure 8.4C, D). We included in the model scaling factors that account for variations in sequencing depth, an overall increase of the labeled RNA fraction, and cross-contamination of unlabeled RNA (section 7.5). The model was fitted using maximum likelihood and assuming negative binomial distribution to cope with overdispersion of read counts [139, 140]. Our method yields absolute splicing and degradation rates, but provides synthesis rates up to one factor common to all TUs. Absolute synthesis rates were obtained by scaling all values so that the median steady-state level of ORF-TUs matches the known median of 2.4 mRNAs per cell [136]. To facilitate comparisons of the obtained RNA metabolic rates, we present the synthesis rate as the average time to synthesize one transcript in a single cell ('synthesis time'); the degradation rate as the time needed to degrade half of the mature RNAs ('half-life'); and the splicing rate as the time to process half of the precursor RNA junction ('splicing time').

The synthesis times and half-lives inferred from distinct splice junctions of the same TU agreed well, demonstrating the robustness of our approach (Spearman rank correlation = 0.44 for synthesis time, $P < 2$ x 10-16 and Spearman rank correlation = 0.79 for half-life, $P < 2$ x 10-16, Figure 8.4E). Based on this comparison, we estimated the accuracy to be typically 46% for synthesis times and 31% for half-lives (mean coefficient of variation). Estimation of the accuracy based on comparing the estimates obtained from the two time series replicates indicate that the accuracy of the estimates of splicing times is between the accuracy for half-lives and synthesis times. The variations in the rate estimates were much smaller than the dynamic range of the rates (about 50-fold each), allowing us to interpret rate differences. Supported by the good agreement of rates across junctions, we took the mean synthesis times and half-lives as estimates for the entire TU.

In order to estimate synthesis and degradation rates of intronless genes, a kinetic model that

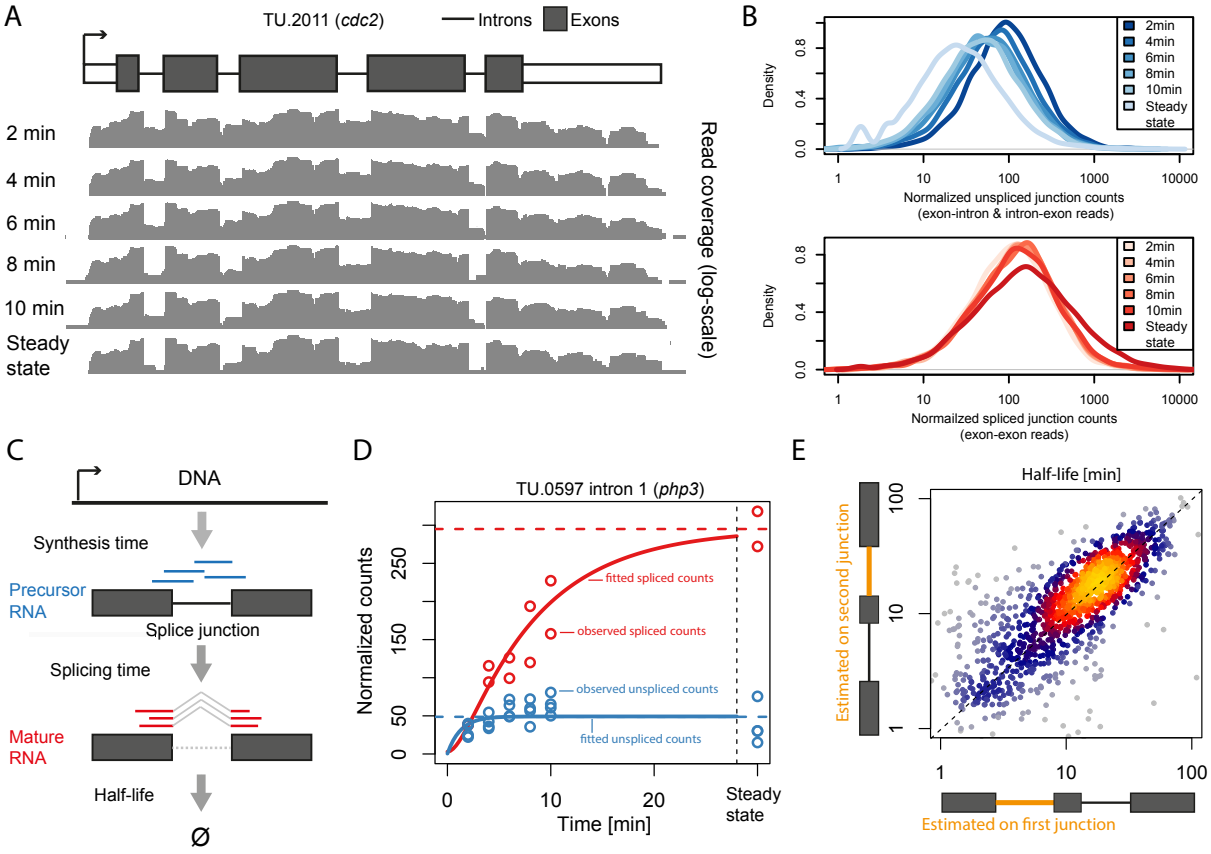

Figure 8.4.: Estimating RNA processing rates using labeled RNA time series. (A) Per base coverage (grey tracks) in a logarithmic scale of 4tU-Seq samples at 2, 4, 6, 8, and 10 min. labeling and for one RNA-Seq sample (i.e. steady-sate) along the UTRs (white boxes), the exons (dark boxes) and the introns (lines) of the TU encoding cdc2. (B) Distribution of sequencing-depth normalized unspliced junction read counts (top panel) and normalized spliced junction read counts (lower panel) for the complete 4tU-Seq time series and the steady-state RNA-Seq samples. (C) Schema of the junction first-order kinetics model. Each splice junction is modeled individually, assuming constant synthesis time, splicing time and half-life. Unspliced junction reads (blue) are specific to the precursor RNA and spliced junction reads (red) are specific to the mature RNA. (D) Observed (circles) and fitted (lines) splice junction counts for the first intron of TU.0597 (php3). Unspliced (blue) and spliced (red) normalized counts (y-axis) are shown for all 4tU-Seq samples and the steady-state sample (x-axis). (E) Half-life estimated from the first (x-axis) versus the second (y-axis) splice junction on TUs with two or more introns.

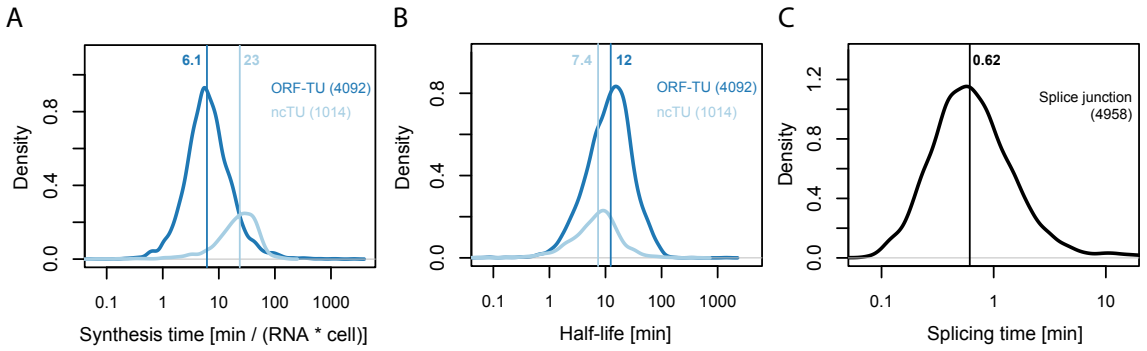

Figure 8.5.: Distribution of synthesis times (A), half-lives (B) for ORF-TUs (blue) and ncTUs (light blue) and splicing times for junctions (C). Median indicated as vertical line.

takes as input all reads overlapping the exon was used. When applied to intron-containing genes, parameter estimates with the exon model were consistent with those obtained with the splice junction model. Overall, synthesis and degradation rates correlated well with previous estimates from microarray data ([67], Spearman rank correlation = 0.45, P < 2 x 10-16 for synthesis rate and Spearman rank correlation = 0.74, P < 2 x 10-16 for half-life), strongly supporting our rate estimation procedure.

8.6. Distinct kinetics of mRNA and ncRNA metabolism

Overall, RNA synthesis and degradation occurred on similar time scales (median synthesis time of 7.4 min compared to a median half-life of 11 min) and about an order of magnitude slower than splicing (median splicing time 37 sec Figure 8.5). These results are consistent with splicing of beta-globin introns within 20 to 30 sec as measured by in vivo single RNA imaging [141], and argue against earlier slower estimates for splicing times of 5 to 10 min [142]. Notably, ncTUs were synthesized at a significantly lower rate than ORF-TUs (median synthesis times of 23 min and 6.1 min, respectively, P < 2 x 10-16, Wilcoxon test), and were degraded slightly faster (median half-life of 12 min for ORF-TUs versus 7.4 min for ncTUs, P < 2 x 10-16, Wilcoxon test). Thus, the differences in steady-state levels of mRNAs and ncRNAs are achieved both by longer synthesis times and shorter half-lives for ncRNAs, although the differences in synthesis times dominate. Transcription is known to be the major determinant of gene expression. However, among genes expressed above background level as investigated here, the dynamic ranges across the bulk of all TUs (95% equi-tailed interval) showed similar amplitudes for all three rates (53-fold for synthesis, 47-fold for half-life, and 33-fold for splicing time, Figure 8.5). Hence, there are large and comparable variations between genes at the level of RNA synthesis, degradation, and splicing. In the following, we first analyze the determinants for RNA synthesis and degradation, and then discuss the determinants for splicing rates.

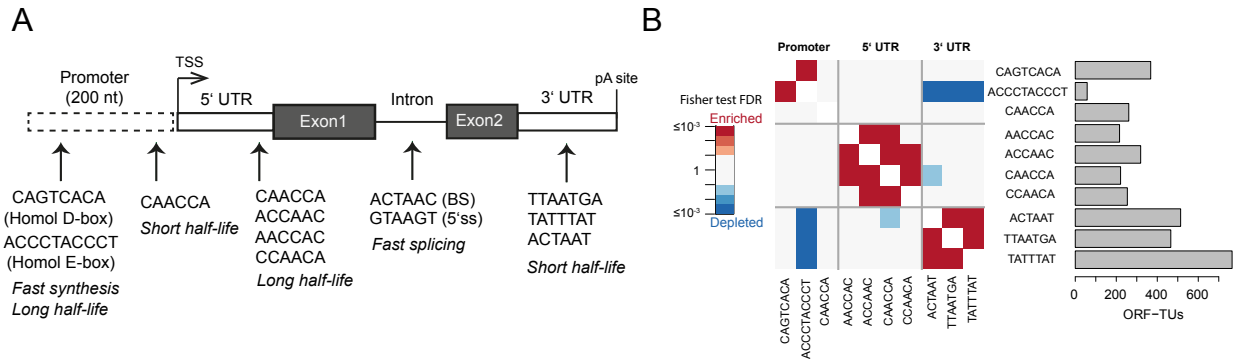

Figure 8.6.: Sequence motifs associated with in vivo degradation and synthesis rates (A) The 12 motifs found in promoter, 5'UTR, intron and 3'UTR sequences of ORF-TUs are shown, together with their qualitative effects on RNA metabolism rates. No motif was found in coding sequences. (B) Number of ORF-TUs with at least one occurrence (horizontal bar) and significant (FDR < 0.1) co-occurrence enrichment (red) and depletion (blue) for all motif pairs. Significance was assessed using Fisher test within ORF-TUs with a mapped polyA site (Mata et al.), followed by Benjamini-Hochberg multiple testing correction.

8.7. Sequence motifs associated with RNA metabolism

We systematically searched for motifs in ORF-TU sequences that could influence RNA synthesis, splicing, and degradation rates (Figure 8.1, step 2). First, 6-mer motifs were identified, whose frequency in a given gene region (promoter, 5'UTR, coding sequence, intron, 3'UTR) significantly correlated with either rate while controlling for other 6-mer occurrences (multivariate linear mixed model, section 7.6). Next, overlapping motifs associating with the same rate in the same direction were iteratively merged and extended to include further nucleotides that significantly associated with the rate. We found 12 motifs that significantly associated with RNA metabolism kinetics (Figure 8.6A). Motifs found within TUs were strand-specific, consistent with their function as part of RNA, whereas motifs found in the promoter region (except one, CAACCA), occurred in both orientations, suggesting that they function in double-stranded DNA. These observations strongly supported the functional relevance of the discovered motifs. The number of ORF-TUs per motif ranged from 58 (ACCCTACCCT) to 765 (TATTTAT) with motifs in the 3' UTR being the most abundant (Figure 8.6B).

8.8. Determinants of high expression

Motifs that were predictive of RNA synthesis times were only found in the promoter region, further validating our approach (Figure 8.6A). We identified de novo the Homol D-box (CAGTCACA), a fission yeast core promoter element, and the Homol E-box (ACCCTACCCT), providing positive controls. In agreement with literature [143, 144], the

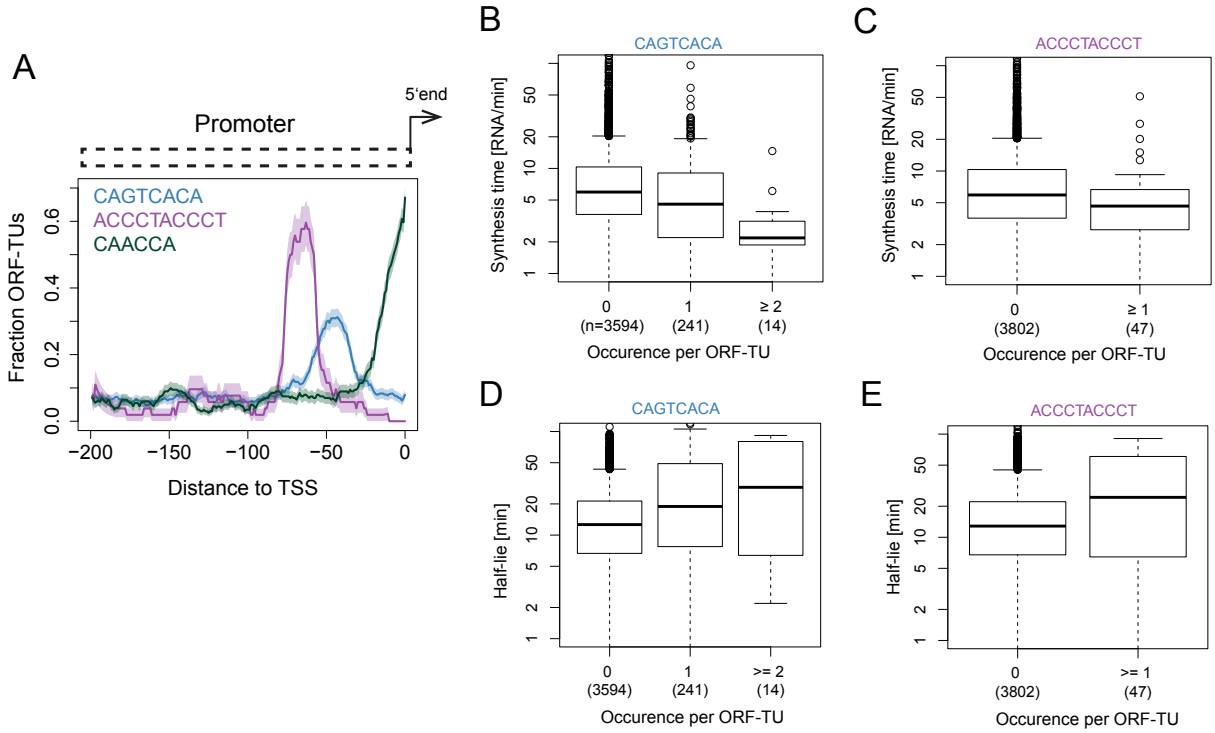

Figure 8.7.: Location and effect on rates of motifs in promoters. (A) Fraction of ORF-TUs containing the motif (y-axis) within a 20 bp window centered at a position (x-axis) upstream of the TSS for the Homol D-box (blue), the Homol E-box (purple) and the CAACCA motif (dark green). (B) Distributions of synthesis time among ORF-TUs that have zero, one or more than one occurrence of the motif CAGTCACA in their promoter sequence. (C) As in (B) for the motif ACCCTACCCT. (D) As in (B) for half-life. (E) As in (C) for half-life.

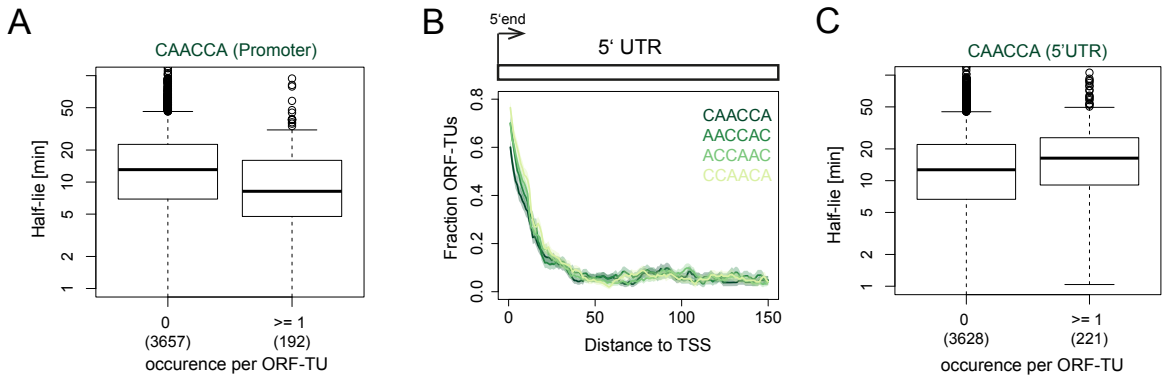

Figure 8.8.: Effect on half-life and location of the AC-rich motif. (A) Distributions of half-lives of ORF-TUs that have zero and one or more occurrence(s) of the motif CAACCA in their Promoter sequence. (B) Fraction of ORF-TUs containing the motif (y-axis) within a 20 bp window centered at a position (x-axis) upstream of the TSS for the AC-rich motifs. (C) Distributions of half-lives of ORF-TUs that have zero and one or more occurrence(s) of the motif CAACCA in their 5'UTR sequence.

Homol D-box and the Homol E-box motifs were enriched in ribosomal protein genes (32% and 41% of all ORF-TUs with these motifs), frequently co-occurred in promoters (Figure 8.6B, Fisher test, False Discovery Rate < 0.1) and showed strong localization preference at a distance of around 45 bp (Homol D-box) and 65 bp (Homol E-box) upstream of the TU 5'end (Figure 8.7A). The 3'UTRs of ORF-TUs with a Homol E-box were significantly depleted for all three motifs that we found to be associated with mRNA instability (FDR <0.1, Figure 8.6B), indicating that the high levels of expression of these genes are achieved by a combination of efficient promoter activity and RNA-stabilizing 3'UTRs. Both motifs associated with decreased synthesis time by 28% (Homol D-box) and 32% (Homol E-box) per motif instance (Linear regression, Figure 8.7B,C), but also with increased half-life (50% and 31%) of the corresponding RNAs (Figure 8.7D,E), likely because those RNAs are both highly synthesized and stable.

8.9. Determinants of RNA half-life

Motifs that were predictive of RNA half-lives were found in the promoter and in UTRs. A novel AC-rich promoter motif (CCAACA) is located near the TU 5'end (Figure 8.7A), and associated with a decrease in half-life by 30% per motif instance (Linear regression, Figure 8.8A). Four AC-rich motifs were found (CAACCA, AACCAC, ACCAAC, and CCAACA) in 5' UTRs, preferentially located near the TU 5'end (Figure 8.8B) and were associated with an increased RNA half-life (only one example shown in Figure 8.8C). Thus, for the AC-rich motif CCAACA the associated effect with half-life is the opposite, depending on whether the motif is located upstream or downstream of the TU 5'end.

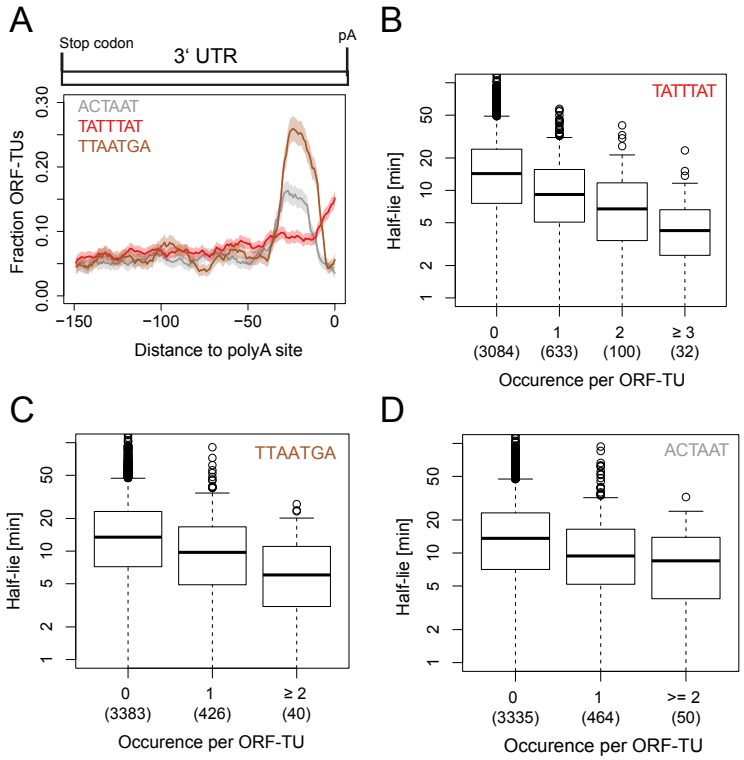

Figure 8.9.: Motifs in the 3'UTR of ORF-TUs. (A) Fraction of ORF-TUs containing the motif (y-axis) within a 20 bp window centered at a position (x-axis) upstream of the polyA site. (B) Distributions of half-lives of ORF-TUs that have zero, one, two or more than two occurrence(s) of the motif TATTTAT. (C) and (D) As in (B) for the motif TTAATGA and ACTAAT.

Three motifs were detected in 3' UTRs of ORF-TUs that all were associated with decreased RNA half-lives. One of these (TATTTAT) corresponds to the known AU-rich element (ARE) that destabilizes RNAs [145, 146] and that was found in 19% of the ORF-TUs and for which we estimated a half-life decrease per motif instance of 33% (Figure 8.9B). The second motif (TTAATGA) and the third motif (ACTAAT) are novel and associated with a reduction in transcript half-lives by similar extents (30% and 27%, Figure 8.9C,D). These two motifs were found in a large number of ORF-TUs (466 and 514, 11% and 13% respectively), and were co-occurring (FDR < 0.1, Figure 8.6B), yet not overlapping with each other. These findings suggest that TTAATGA and ACTAAT are widespread RNA elements that determine important RNA stability regulatory pathways. In contrast to the AU-rich element, the two novel 3'UTR motifs were sharply peaking 28 bp (ACTAAT) and 25 bp (TTAATGA) upstream of the polyA site (Figure 8.9A), indicating that they could implicate similar mechanisms, that are distinct from the AU-rich element pathway, and that are related to RNA polyadenylation or involve interactions with the polyA tail. Two of our motifs, the AC-rich element in the promoter region and the ACTAAT in 3'UTRs are enriched in the same regions of human, mouse, rat, and dog genes [147], indicating that their function is conserved from *S. pombe* to mammals.

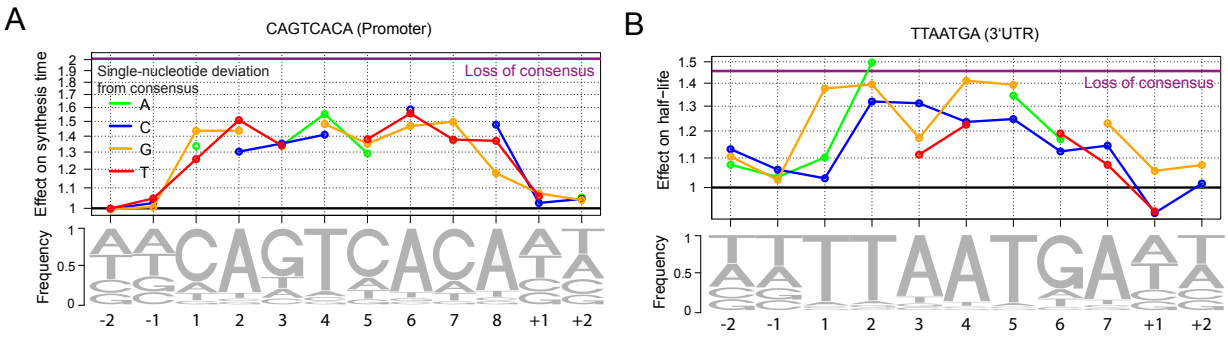

Figure 8.10.: Single-base substitution effects on RNA synthesis and half-life. (A) Nucleotide frequency within motif instances (lower track) and prediction of the relative effect on synthesis time (upper track) for single nucleotide substitution in the Homol D-box consensus motif and of complete loss of the consensus motif (purple line). (B) As in (A) for the 3' UTR motif TTAATGA.

8.10. Effects of single nucleotides on RNA kinetics

We next asked whether deviations from the consensus sequence of the discovered motifs can predict changes in synthesis time and half-life. We considered a linear model that included the effect of changes at each base position and the number of motifs present in each gene or RNA and fitted across all genes allowing for mismatches (Methods 7.6). Generally, deviations from the consensus sequence associate with decreased effects of the motif on synthesis time or half-life. These changes often neutralize the effect of the motif. For instance, loss of the consensus Homol D-box apparently increased synthesis time two-fold (Figure 8.10A, purple line). A single-nucleotide deviation from the consensus Homol D-box motif by a C at the 6th position associated with a 1.6-fold increased synthesis time. Similarly, a T to G substitution at the 5th position of the TTAATGA motif was predicted to lead to a 1.4-fold increased half-life, similar to the loss of the complete consensus motif (Figure 8.10B). Changes in positions flanking the motif have minor effects but may play functional roles. Nucleotides associated with important effects tended to also be more frequent (Sequence logo, Figure 8.10A,B) indicating that there is evolutionary pressure on these positions and further indicating that these motifs are functional. Similar results were obtained for all motifs (see Appendix).

8.11. New regulatory motifs are functional

Although the above analysis strongly indicated that the identified motifs were functional, the evidence remained correlative. In order to test the functional role of these new motifs, we asked whether genetic variants affecting these sequence elements resulted in a perturbed expression level in a direction and extent that match the predictions (Figure 8.1, step

3). We analyzed expression data of an independent study that profiled steady-state RNA levels of a library of 44 different recombinant strains obtained from a cross between the standard laboratory strain 968, also profiled here, and a South African isolate Y0036 [124]. In recombinant panels, the alleles of a reference and of an alternate parental strain are randomly shuffled by meiosis recombination within the population. For a variant of interest, recombinant strains group in two sub populations: about one half carries the reference allele and the other half the alternate allele. Variants that are not in linkage with the one of interest, for lying on another chromosome or far away on the same chromosome, are approximately equally inherited within the two sub populations. Hence, differential gene expression between the two sub populations reflects local regulatory variants, such as promoter and RNA motifs, while controlling for distant, trans-acting regulatory variants. To evaluate the effects due to perturbations of the motifs, we restricted the analysis to ORF-TUs with a variant that we predicted to significantly affect the rate (Methods 7.7), and harboring no further variant within the promoter region and the whole TU. These variants affected 20 motifs and were all single nucleotide variants (Table 8.1).

Table 8.1.: Comparison of predicted and observed effects on gene expression when cis-regulating motifs are mutated. Description of columns: **1)** ID **2)** strand **3)** p-value of motif **4)** region **5)** motif **6)** associated rate **7)** sequence **8)** mutated sequence **9)** log10 observed fold-change **10)** standard error of observed fold-change **11)** log10 fold change predicted **12)** p-value of predicted effect **13)** standard error of predicted fold-change.

1	2	3	4	5	6	7	8	9	10	11	12	13
TU.0209	+	2,06E-03	UTR3	TTAATGA	half-life	TTTTAATGGTT	TTTTAATGATT	-0,091	0,028	-0,090	1,60E-02	0,037
TU.0236	+	7,19E-06	UTR5	CCAACA	half-life	CACCACCACC	CACAACCACC	-0,035	0,019	-0,102	1,76E-03	0,033
TU.0236	+	2,09E-03	UTR5	ACCAAC	half-life	CCACCACCAC	CCACAACCAC	-0,035	0,019	-0,118	4,21E-04	0,033
TU.0236	+	4,42E-03	UTR5	AACCAC	half-life	ACCACCACCA	ACCACAACCA	-0,035	0,019	-0,075	3,25E-02	0,035
TU.0369	+	9,32E-04	UTR5	AACCAC	half-life	TAAACGACTT	TAAATGACTT	-0,066	0,029	-0,199	8,22E-07	0,040
TU.0488	+	6,80E-05	UTR5	ACCAAC	half-life	GTATCAACGT	GTATCAACAT	-0,020	0,014	0,079	3,02E-04	0,022
TU.0737	+	2,52E-03	UTR3	ACTAAT	half-life	AAATTAATCA	TAATTAATCA	-0,022	0,022	-0,031	1,31E-03	0,010
TU.0796	+	1,47E-02	UTR3	ACTAAT	half-life	TTACCAATTA	TCACCAATTA	0,019	0,034	0,030	1,21E-02	0,012
TU.2921	+	1,21E-02	UTR3	ACTAAT	half-life	GAACAAATAG	GAAGAAATAG	0,049	0,024	0,087	8,76E-04	0,026
TU.2961	+	5,39E-05	UTR3	TTAATGA	half-life	GCTTAATGACC	GCTTAATGGCC	0,194	0,045	0,090	1,60E-02	0,037
TU.3073	+	1,77E-03	UTR3	TTAATGA	half-life	ATTTAATAAAT	TTTTAATAAAT	-0,027	0,016	-0,032	1,26E-02	0,013
TU.3189	+	1,71E-03	UTR3	ACTAAT	half-life	TTACTATTTG	TTACTATCTG	NA	NA	0,094	5,18E-03	0,034
TU.4953	-	4,38E-04	Promoter	CAGTCACA	synthesis	GAGAGTCACATC	GAGATTCACATC	-0,196	0,026	-0,141	4,71E-07	0,028
TU.5006	-	2,41E-03	UTR3	TATTTAT	half-life	TATAATTATGA	TGTAATTATGA	0,052	0,031	0,036	3,31E-03	0,012
TU.5217	-	9,81E-12	UTR3	TATTTAT	half-life	ATTATTTATAG	ATTGTTTATAG	-0,001	0,030	0,073	5,84E-04	0,021
TU.5869	-	3,41E-02	UTR5	CCAACA	half-life	TTCCAATATA	TTCCAGTATA	0,011	0,025	-0,124	1,46E-02	0,051
TU.6215	-	1,72E-02	UTR3	ACTAAT	half-life	TTAGTAATTA	TTAGCAATTA	0,001	0,023	0,078	5,61E-03	0,028
TU.6266	-	4,26E-02	UTR3	TATTTAT	half-life	AGTGTTTATGA	AGTGCTTATGA	0,063	0,030	0,126	1,01E-05	0,028
TU.6563	-	4,22E-04	UTR3	TTAATGA	half-life	TATTAATAAAA	TATTACTAAAA	-0,001	0,031	0,092	2,17E-02	0,040
TU.4953	-	4,38E-04	Promoter	CAGTCACA	half-life	GAGAGTCACATC	GAGATTCACATC	-0,196	0,026	-0,143	2,01E-07	0,027

A positive control was provided by the alternate allele of the gene rctf1, which differed from the reference allele by a single nucleotide, a G-to-T substitution at the third position of a Homol D-box motif in its promoter. Recombinant strains harboring the alternate allele showed significantly lower steady-state expression levels (Figure 8.11A, P = 2 x 10-10, one-sided Wilcoxon test) consistent with the predicted 1.35-fold increased synthesis time (Figure 8.10A). Two variants acting in an opposite fashion demonstrated the functional role of the 3'UTR motif TTAATGA. The linear model predicted a 1.23-fold increased

half-life for a A to G substitution at the 7th position (7.A>G, Figure 8.10B). Consistently, 7.A>G substitution occurring on the gene SPCC794.06 led to a significantly increased expression level (Figure 8.11B, P = 2 x 10-4) whereas the (7.G>A) in the gene mug65 led to a significantly decreased expression level (Figure 8.11C, P = 10-4). Among the novel motifs, the TTAATGA could be validated (3 out of 4 genes with a significant change in expression in the predicted direction $P < 0.05$) as well as the AACCAC motif (2 out of 2 genes with a significant change in expression in the predicted direction $P < 0.05$). The other motifs generally did not yield significant changes, possibly because the predicted and the observed effects were of small amplitude. Over all 20 variants, the observed and predicted fold-changes did not only agree in direction but also in amplitude (Pearson correlation, P = 9 x 10-4, Figure 8.11D, Table 8.1), demonstrating that most motifs were functional and that the model predicted quantitatively the effects of single mutations.

8.12. Intron sequences determining splicing kinetics

Sequence motifs predictive of splicing times were found only in introns, and here only in the donor region downstream of the 5'-splice site (5'SS) and at the branch site (BS). We complemented this set with the 3'-splice site (3'SS) and extended motifs in each direction as far as significant single nucleotide effects were found (Linear regression and cross-validation, Methods 7.6, Figure 8.12A). Significant effects were found up to six nucleotides downstream of the 5'SS. These bases are those pairing with the spliceosome component U6 small nuclear RNA during the first catalytic step of splicing (reviewed in [148, 149]). We also found significant effects up to seven nucleotides 5' of the branch point adenosine and one nucleotide 3' of it, entailing all but one of the seven nucleotides pairing with the U2 small nuclear RNA [148]. These two regions showed the strongest effects, with typically 1.1- to 1.5-fold decreased splicing time compared to consensus, showing that exact base-pairing with U6 and U2, although not required for splicing, is a determinant for its kinetics. Significant but weaker effects (less than 1.1-fold) extending up to 8 nucleotides 3' and 5' of the 3'SS were also found. Deviations from the consensus sequence invariably associated with increased splicing time (Figure 8.12A). Also, splicing time anti-correlated with the frequency of the core branch site sequence across the genome (Figure 8.12B).
These observations indicate that there is selective pressure on all introns for rapid splicing in *S. pombe*. We then asked whether the selective strength at these positions always reflected their quantitative contribution to the rate of splicing. Overall, the mean effect of a deviation from the consensus significantly correlated with how little variable the base was across all introns genome-wide (Kullback-Leibler information, Spearman rank correlation = 0.61, P = 5 x 10-4, Figure 8.12C). Positions within the branch site region and downstream of the 5'SS are most commonly found as consensus and showed the largest

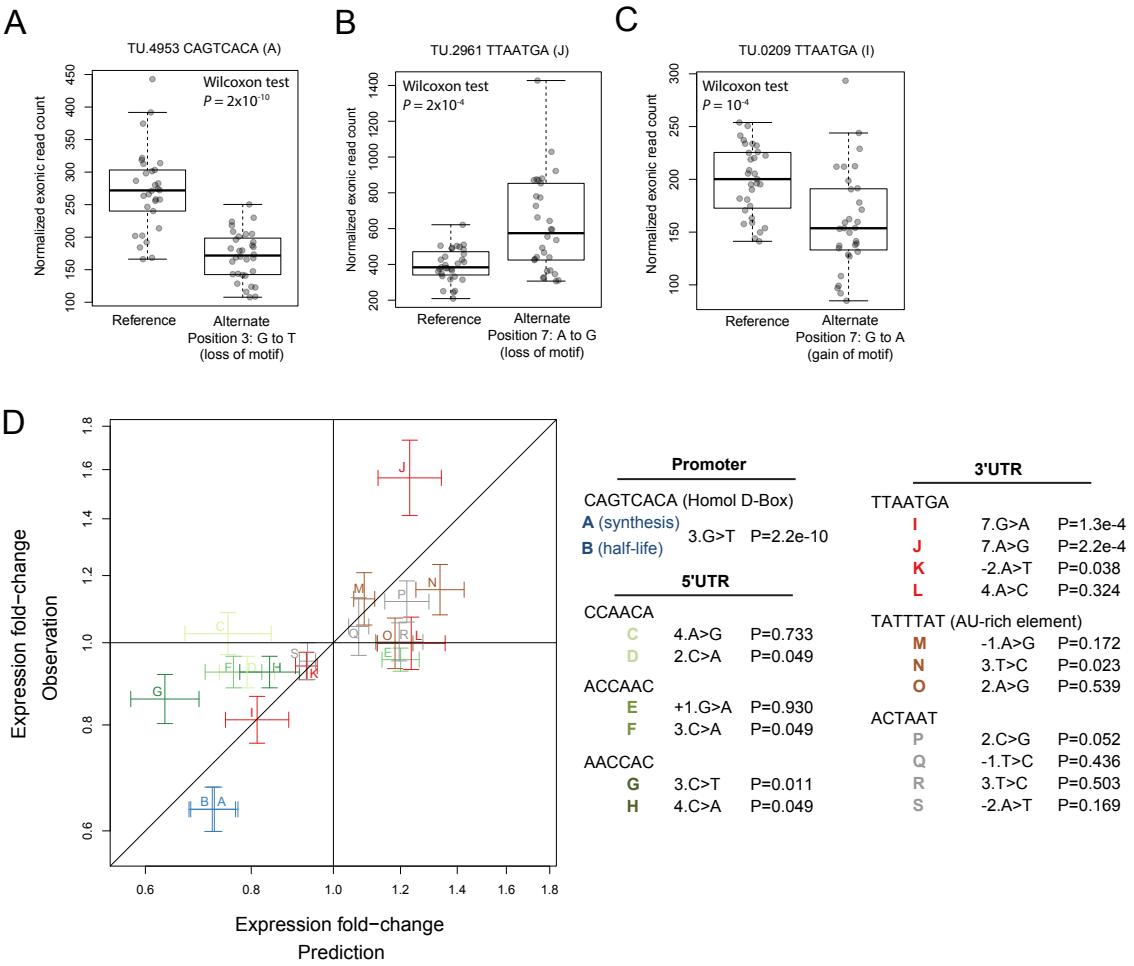

Figure 8.11.: Validation of motif SNP effects on expression. (A,B,C) Boxplot and individual data point of exonic read counts normalized for sequencing depth and batch effects (y-axis) for strains grouped by genotype (x-axis) for the gene rctf1 (A), SPCC794.06 (B), and mug65 (C) (D) Validation of motifs using expression data of a recombinant strain library (Clément-Ziza et al., 2015). Fold-change in steady-state expression level due to a single nucleotide variant as predicted from our models (x-axis) against average expression fold-change between strains harboring the variant and strains harboring the reference allele (y-axis). Estimated standard errors for the prediction and the observed are represented by the vertical and horizontal segments. The overall Spearman rank correlation is 0.76 (P=0.006). In legend: SNP code and one-sided Wilcoxon test P-value.

effect on splicing kinetics. The last nucleotide of the 5' exon is generally a guanine but did not influence splicing time (5'SS-1 position), indicating that other sources of selection influence this position.

8.13. Splicing kinetics also depends on RNA synthesis

Splicing time did not strongly correlate with intron length (Spearman rank correlation = 0.03, P = 0.05) and correlated negatively with TU length (Spearman rank correlation

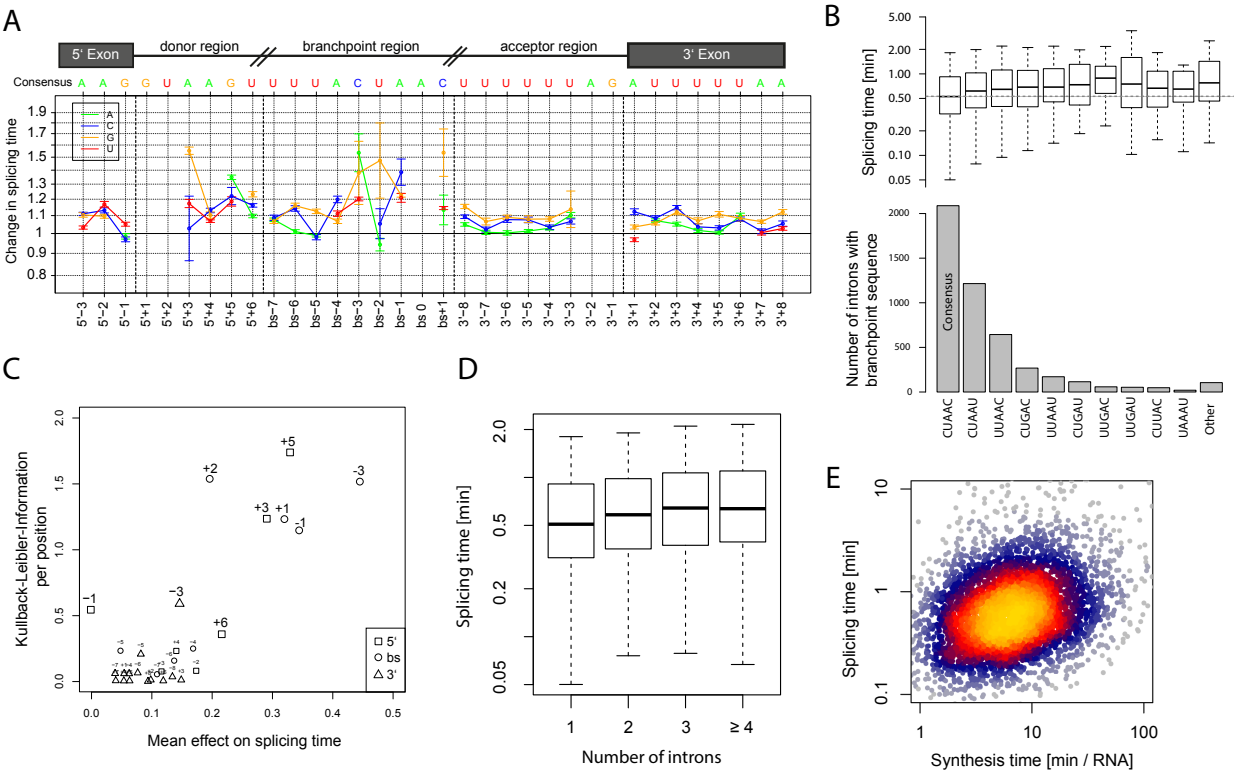

Figure 8.12.: Determinants of in vivo splicing rates. (A) Prediction of the relative effect on splicing time (y-axis) for single nucleotide substitution compared to consensus sequence around the 5'splice site, the branch site and the 3'splice site (cartoon top panel). Effects at invariant positions (5'SS: GU, BS: A and 3'SS: AG) cannot be computed. (B) Occurrence (bottom panel) and distribution of half-splicing times (top panel) per BS motif (x-axis) sorted by frequency. The median splicing time of introns with consensus sequence is indicated with a dashed line. (C) Information content (y-axis) versus mean effect on splicing time (x-axis) for each position (relative numbers) of the 5'SS (squares), BS (circles) and 3'SS (triangles). Positions with information content > 0.3 are highlighted. (D) Distribution of splicing times (y-axis) versus number of introns in the TU (x-axis). (E) Splicing time (y-axis) versus synthesis time (x-axis)

= -0.16, P $<$ 2 x 10-16), showing that short transcripts are spliced more slowly. This is in contrast to observations in mouse, where short transcripts and short introns are more rapidly spliced than longer ones [57]. This apparent discrepancy might be due to the fact that *S. pombe* neither contains very long genes nor very long introns. Splicing time increased with the number of introns (Figure 8.12D) as in mouse cells [57], independently of the relative position of the intron within the transcript. However, this correlation could be explained by the fact that genes with few introns also have efficient splice site and branch site sequences (multivariate analysis, section 7.8). Thus it is not the number of introns per se that affects splicing, rather, genes that give rise to rapidly processed RNAs evolved to have few introns and efficient splicing RNA elements. Splicing time correlated positively with synthesis time (Spearman rank correlation = 0.28, P $<$ 2 x 10-16, Figure 8.12E), in agreement with results in mouse. This may be due to co-evolution of synthesis and splicing, or because highly transcribed loci are more readily accessible

to the splicing machinery. This finding is not in contradiction to the understanding that fast RNA polymerase elongation inhibits splicing [142], because synthesis rate is mostly determined by the rate of transcription initiation rather than elongation [150]. Altogether, multivariate analysis (Methods 7.8) indicated that sequence elements, synthesis time, and TU length independently enhance splicing, where sequence is the major contributor (50% of the explained variance), followed by synthesis rates (42% of the explained variance).

8.14. Antisense transcription affects mRNA synthesis, not stability

Repression by antisense transcription is increasingly being recognized as an important mode of regulation of gene expression, but its mechanisms remain poorly understood [151, 152]. In our revised genome annotation, convergent TUs generally did not overlap (1022 out of 1616), typically leaving 75 bp of untranscribed sequence in between (Figure 8.13A). Among overlapping convergent pairs, TU 3'-ends were enriched within introns (P = 0.001) and depleted within exons (P = 0.001) of the opposite strand (1,000 random permutations of TU pairs), likely because coding sequence is highly restrained and may impair encoding of polyadenylation and termination signals for the opposite strand. Although transcripts are generally not antisense to each other, we found 520 ncTUs antisense of ORF-TUs (one example in Figure 8.13B).

In fission yeast, antisense transcription could repress sense RNA synthesis, as in *S. cerevisiae* [19], or affect RNA stability by RNA interference, because fission yeast, unlike budding yeast, contains the RNAi machinery. ORF-TUs with antisense ncTUs overlapping at least 40% exhibited significantly increased synthesis times (Wilcoxon test, P = 9 x 10-7), consistent with repression of mRNA synthesis by antisense transcription. This effect was higher when the antisense ncTU covered a larger area of the ORF-TU (Figure 8.13C). However, no difference regarding mRNA stability was observed (Figure 8.13D). Taking together, these results indicate that expression levels of those ORF-TUs were mainly regulated by means of mutually exclusive transcription rather than by RNA interference, which would be predicted to affect transcript stability.

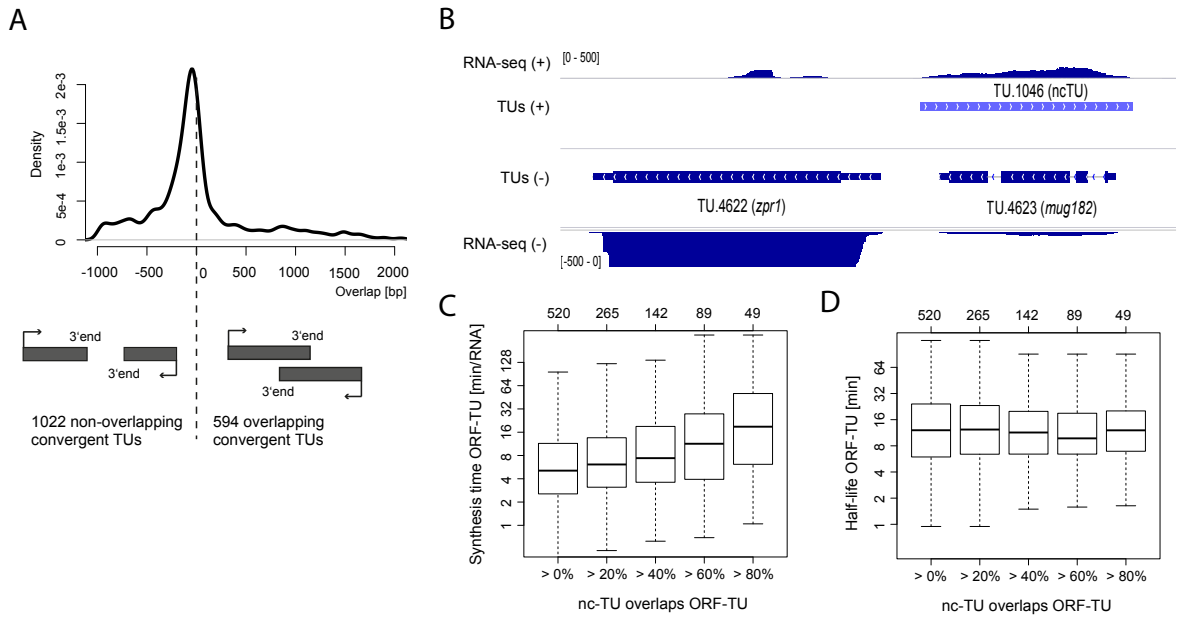

Figure 8.13.: Antisense transcription represses ORF-TUs synthesis. (A) Distribution of the overlap of the 1,616 convergent TUs separated by less than 1,000 bp. Most convergent TUs did not overlap. (B) Example of an ORF-TU (mug182) that is covered completely by a ncTU (TU.1046) on the opposite strand. The RNA-Seq read coverage (steady-state expression) of mug182 is considerably lower than of the adjacent gene zpr1. (C) Distribution of synthesis times of ORF-TUs grouped by the fraction of overlap by antisense ncTUs. (D) As in (C) for half-lives.

Part IV.

Conclusions and Outlook

9. Conclusions and Outlook

The work presented in this thesis elucidates novel molecular mechanisms that regulate *in-vivo* RNA metabolism. By combining metabolically labeled RNA profiling with computational modeling we could obtain genome-wide estimates for RNA synthesis, splicing and degradation in two different experimental setups. For both studies, we made use of a novel experimental protocol that allows the purification and quantification of newly synthesized RNAs with minimal perturbation. Advanced mathematical modeling was used to translate the raw time series measurements of relative RNA abundances into comparable and meaningful estimates of RNA turnover kinetics. Robust statistical methods and procedures were developed and applied in a biology driven way to investigate cellular processes and genetic sequences associated with and regulating RNA metabolism in cells.

9.1. Periodic mRNA synthesis and degradation cooperate during cell cycle gene expression

In the first study (Part II), we conducted the first systematic investigation of mRNA synthesis and degradation rates during the cell cycle, using as an eukaryotic model system the yeast *S. cerevisiae*. The labeling protocol was applied to monitor mRNA synthesis and degradation of synchronized cells along three cell cycles. We developed and released the software package MoPS, a general-purpose, model-based screening algorithm for the identification of periodic changes in time course measurements. By integrating total and labeled mRNA replicate time series, MoPS identified a reliable set of 479 genes with periodic expression during the cell cycle. In contrast to other periodicity screening methods, MoPS is particularly robust and extracts meaningful parameters from a periodic time course. These parameters, like expression peak time, peak height, and the shape of the expression time course laid a solid basis for an in-depth analysis of the underlying biological phenomena.

We found that labeled and total mRNA time courses are highly similar for most genes. This indicates that transcription is the key determinant of cell-cycle phase specific mRNA expression. By clustering of the fitted gene-specific parameters of labeled expression, we identified groups of co-transcribed genes. We were able to retrieve known regulatory

DNA motifs and identify transcription factors that determine cell cycle phase-specific transcription, confirming and extending previous work. By comparing gene expression levels of TFs with synthesis rates of their target genes, we consistently observed that total mRNA levels of activating (repressive) cell cycle TFs peak when transcription of its targets is maximal (minimal). Further investigation of co-regulated gene clusters revealed that the timing and the magnitude of periodic expression have different causes. Genes that have common binding sites for cell cycle TFs show coherent timing of expression, but differ in their mRNA synthesis rates. Striking examples are genes exclusively regulated by MBP1, a transcription factor that has a well-studied role in regulating expression of late G1 genes [153, 154]. Although these genes have very similar temporal profiles, they exhibit large differences in their synthesis rates and total mRNA levels. These differences are related to the composition of the core promoter TATA sequence, and correlate with the binding of general transcription factors. Periodic genes that drive cell cycle progression or regulate fundamental processes like chromatin organization in S-phase are found to be highly induced and tend to have a consensus TATA box.

The most intriguing finding from our results is however that most periodically expressed genes show periodic changes in the degradation rates of their mRNAs. We realized that total mRNA levels peak on average only 2 min after labeled mRNA, which indicates the peak of mRNA synthesis activity. This short time delay could not be explained when constant degradation rates were assumed. Computational modeling of degradation kinetics of periodically transcribed genes indicated that the stability of mRNAs decreases shortly after transcription ceases. This highlights the importance of post-transcriptional control on the regulation of genes involved in cell cycle-associated processes. Varying mRNA degradation rates during the cell cycle were previously observed [93]. In this study, two mitotic periodic genes SWI5 and CLB2 show a decrease in mRNA stability after peak expression to prevent carryover of mRNAs into the next cycle. Our results extend these findings to the majority of all periodic transcripts.

It is an open question how these changes are achieved, but due to the generality of the phenomenon we suggest that increased transcript degradation following a peak of mRNA synthesis is a passive phenomenon [155]. Other studies propose destabilizing, specific RNA-binding factors. Since the two hypotheses are not mutually exclusive, we expect a combination of both mechanisms. Whereas the molecular mechanisms underlying this phenomenon remain to be uncovered, our study revealed that periodic changes in mRNA synthesis and temporally delayed changes in degradation are common events that achieve concise and strong mRNA expression changes during the cell cycle.

Taken together, we obtained for the first time, RNA synthesis and degradation rates in synchronized cells and found evidence for a global mechanism that destabilizes or actively degrades cell-cycle regulated mRNAs. Furthermore, the excellent reproducibility of the measurements and the high temporal resolution at which mRNA synthesis rates and total

mRNA expression were determined will make our data an ideal resource for more advanced reverse engineering approaches of cell cycle related gene expression networks.

9.2. Determinants of RNA metabolism in the *Schizosaccharomyces pombe* genome

In the second study (Part III), we performed metabolically labeled RNA profiling with RNA-Seq (4tU-Seq) at high temporal resolution in wild-type fission yeast cells.
We estimated genome-wide RNA synthesis, splicing and degradation rates and systematically related them to regulatory sequence elements. We could identify known and novel DNA and RNA motifs that significantly affect RNA metabolism kinetics. We were able to estimate the effect of single bases on turnover rates and could validate the identified motifs with independent data from genetically distinct *S.pombe* strains [124].
The basis for this study was led by the transition from Microarrays to RNA-Seq for quantification of labeled and total RNA, because it permits the generation of strand-specific data with a higher dynamic range and single-nucleotide resolution [46]. We developed advanced computational kinetic modeling and statistical procedures that make use of the sequencing data in multiple ways. First, we use the paired-end information of the strand-specific reads to get an accurate map of transcript boundaries across the *S.pombe* genome. Second, it allowed us to profile not only the abundance of mRNAs but the complete transcriptome including short-lived non-coding RNAs. Third, the relative occurrences of split and unsplit reads across splice junctions enabled us to investigate splicing kinetics.
With this data, we first systematically annotated the transcribed genome of *S.pombe,* thereby redefining most ncRNAs and a large fraction of UTRs in mRNAs, in particular 5'UTRs and thus promoter regions. Advanced kinetic modeling using the 4tU-Seq time series data then allowed us to accurately estimate RNA metabolism rates for all transcripts of the new annotation (Figure 8.1, step 1).
We further introduced an approach to discover regulatory elements in the genome that combines the metabolic rates with robust regression on DNA sequence (Figure 8.1, step 2). Without using further information than simple gene architecture (promoter, UTRs, exons and introns), this approach recovered known regulatory motifs de novo, such as core promoter elements and the 3'UTR AU-rich element, but also provided two novel 3'UTR motifs, and functional AC-rich sequences in promoters and in 5'UTRs. An important advantage of this approach is the ability to determine the contributions of individual sequence elements to each step of RNA metabolism. Whereas standard motif enrichment analysis discriminates only between two classes of data (e.g. highly versus lowly expressed), we used quantitative regression and therefore could exploit the full range of the data without applying any cutoff. Regression furthermore has the benefit to provide quantitative

predictions regarding genetic perturbations that could be directly compared to expression fold-changes for functional validations.
Validation of the discovered motifs was achieved by exploiting existing transcription profiles of genetically distinct strains from an independent study [124]. We asked whether genetic variants affecting the motifs resulted in a perturbed expression level in a direction and extent that match the predictions (Figure 8.1, step 3). Overall, the observed and predicted fold-changes did not only agree in direction but also in amplitude, demonstrating that most motifs were functional and that the model predicted quantitatively the effects of single mutations. The beauty of this validation approach lies in the fact that we can exploit natural occurring genetic variation and do not need to introduce any modifications or perturbations into the cells. This is important, since it has been shown that even small changes can lead to aberrant behavior of molecular processes in cells [67].
Due to the high temporal resolution of the 4tU-Seq dataset, we were further able to investigate the kinetics of splicing. We revealed that splicing in *S.pombe* takes in average less than one minute and that splicing rates profoundly depend on conserved nucleotides around the branch point and the donor and acceptor sites. So far, only Rabani and colleagues [57], using mouse cells, have reported a computational tool to estimate genome-wide splicing rates. That study had used mammalian cells, resulting in a limitation in sequencing depth that restricted many parts of the analysis to the 10% most expressed splice junctions. Due to the higher sequencing depth, our analysis in *S. pombe* could be global. Another advantage of fission yeast is the absence of alternative splicing, which simplifies the analysis and makes rate estimation very robust.

The approach presented here is a pioneering approach to profile RNA metabolism at high temporal resolution. Further research can build on this and investigate differences in rates between different conditions or mutants. With the increased resolution compared to common total RNA-Seq, one could investigate changes in RNA metabolism upon external stimuli like heat stress or intracellular perturbations like blocking of certain pathways or the knockdown of general transcription or degradation factors. Mutants of the C-terminal repeat domain of RNA polymerase II are a particular interesting target since its role in the coupling of transcription and splicing is only poorly understood.
The presented method to decrypt the regulatory code of RNA metabolism is general and could be applied to other higher organisms. In the future, the application of our model may help to understand the consequences of regulatory variation in the human genome, with important implications for understanding gene regulation and interpreting the many disease-risk variants that fall outside of protein-coding regions [156].

Appendix

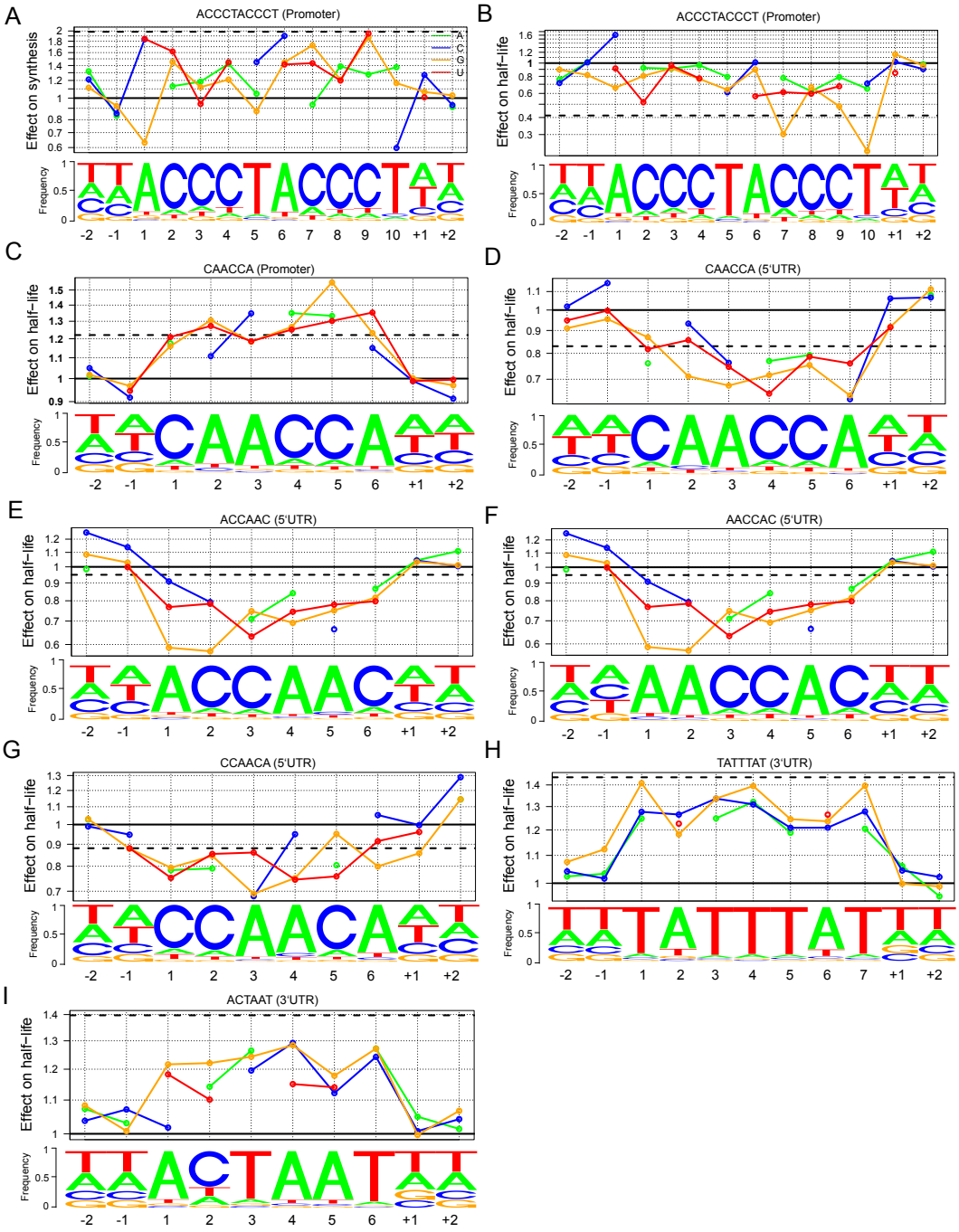

Figure 9.1.: Effects of single nucleotides on RNA kinetics. (A) Nucleotide frequency within motif instances (lower track) and prediction of the relative effect on synthesis time (upper track) for single nucleotide substitution in the Homol E-box consensus motif and of complete loss of the consensus motif (purple line). (B-I) As in (A) for all identified motifs and corresponding rates.

List of Figures

List of Tables

Bibliography

[1] Crick, F. (1970) Central dogma of molecular biology. *Nature 227*, 561–3.

[2] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. Molecular Biology of the Cell. 2002; `http://www.ncbi.nlm.nih.gov/books/NBK21054/`.

[3] Selevsek, N., Chang, C.-Y., Gillet, L. C., Navarro, P., Bernhardt, O. M., Reiter, L., Cheng, L.-Y., Vitek, O., and Aebersold, R. (2015) Reproducible and consistent quantification of the Saccharomyces cerevisiae proteome by SWATH-MS. *Molecular & Cellular Proteomics* 1–16.

[4] Vogel, C., and Marcotte, E. M. (2012) Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Reviews Genetics 13*, 227–232.

[5] Schwanhäusser, B. et al. (2011) Global quantification of mammalian gene expression control. *Nature 473*, 337–342.

[6] Koussounadis, A., Langdon, S. P., Um, I. H., Harrison, D. J., and Smith, V. A. (2015) Relationship between differentially expressed mRNA and mRNA-protein correlations in a xenograft model system. *Scientific reports 5*, 10775.

[7] Roeder, R. G. (1996) The role of general initiation factors in transcription by RNA polymerase II. *Trends in biochemical sciences 21*, 327–35.

[8] Plaschka, C., Larivière, L., Wenzeck, L., Seizl, M., Hemann, M., Tegunov, D., Petrotchenko, E. V., Borchers, C. H., Baumeister, W., Herzog, F., Villa, E., and Cramer, P. (2015) Architecture of the RNA polymerase II-Mediator core initiation complex. *Nature 518*, 376–80.

[9] Arnone, M. I., and Davidson, E. H. (1997) The hardwiring of development: organization and function of genomic regulatory systems. *Development (Cambridge, England) 124*, 1851–64.

[10] Adelman, K., Marr, M. T., Werner, J., Saunders, A., Ni, Z., Andrulis, E. D., and Lis, J. T. (2005) Efficient release from promoter-proximal stall sites requires transcript cleavage factor TFIIS. *Molecular cell 17*, 103–12.

[11] Nechaev, S., Fargo, D. C., dos Santos, G., Liu, L., Gao, Y., and Adelman, K. (2010) Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. *Science (New York, N.Y.) 327*, 335–8.

[12] Svejstrup, J. Q. The RNA polymerase II transcription cycle: Cycling through chromatin. 2004.

[13] Mayer, A., Lidschreiber, M., Siebert, M., Leike, K., Söding, J., and Cramer, P. (2010) Uniform transitions of the general RNA polymerase II transcription complex. *Nature structural & molecular biology 17*, 1272–8.

[14] Mayer, A., Heidemann, M., Lidschreiber, M., Schreieck, A., Sun, M., Hintermair, C., Kremmer, E., Eick, D., and Cramer, P. (2012) CTD Tyrosine Phosphorylation Impairs Termination Factor Recruitment to RNA Polymerase II. *Science 336*, 1723–1725.

[15] Heidemann, M., Hintermair, C., Voß, K., and Eick, D. Dynamic phosphorylation patterns of RNA polymerase II CTD during transcription. 2013.

[16] Hsin, J.-P., and Manley, J. L. (2012) The RNA polymerase II CTD coordinates transcription and RNA processing. *Genes & development 26*, 2119–37.

[17] Meinhart, A., and Cramer, P. (2004) Recognition of RNA polymerase II carboxy-terminal domain by 3'-RNA-processing factors. *Nature 430*, 223–226.

[18] Bird, G., Zorio, D. A. R., and Bentley, D. L. (2004) RNA polymerase II carboxy-terminal domain phosphorylation is required for cotranscriptional pre-mRNA splicing and 3'-end formation. *Molecular and cellular biology 24*, 8963–9.

[19] Schulz, D., Schwalb, B., Kiesel, A., Baejen, C., Torkler, P., Gagneur, J., Soeding, J., and Cramer, P. (2013) XTranscriptome surveillance by selective termination of noncoding RNA synthesis. *Cell 155*, 1075–1087.

[20] Tudek, A., Porrua, O., Kabzinski, T., Lidschreiber, M., Kubicek, K., Fortova, A., Lacroute, F., Vanacova, S., Cramer, P., Stefl, R., and Libri, D. (2014) Molecular basis for coordinating transcription termination with noncoding RNA degradation. *Molecular cell 55*, 467–81.

[21] Xu, Z., Wei, W., Gagneur, J., Perocchi, F., Clauder-Münster, S., Camblong, J., Guffanti, E., Stutz, F., Huber, W., and Steinmetz, L. M. (2009) Bidirectional promoters generate pervasive transcription in yeast. *Nature 457*, 1033–7.

[22] Wei, W., Pelechano, V., Järvelin, A. I., and Steinmetz, L. M. (2011) Functional consequences of bidirectional promoters. *Trends in genetics : TIG 27*, 267–76.

[23] Latos, P. A., Pauler, F. M., Koerner, M. V., Senergin, H. B., Hudson, Q. J., Stocsits, R. R., Allhoff, W., Stricker, S. H., Klement, R. M., Warczok, K. E.,

Aumayr, K., Pasierbek, P., and Barlow, D. P. (2012) Airn transcriptional overlap, but not its lncRNA products, induces imprinted Igf2r silencing. *Science 338*, 1469–1472.

[24] Pelechano, V., and Steinmetz, L. M. (2013) Gene regulation by antisense transcription. *Nat Rev Genet 14*, 880–893.

[25] Moazed, D. (2009) Small RNAs in transcriptional gene silencing and genome defence. *Nature 457*, 413–420.

[26] Will, C. L., and Lührmann, R. (2011) Spliceosome structure and function. *Cold Spring Harbor Perspectives in Biology 3*, 1–2.

[27] Fica, S. M., Tuttle, N., Novak, T., Li, N.-S., Lu, J., Koodathingal, P., Dai, Q., Staley, J. P., and Piccirilli, J. A. (2013) RNA catalyses nuclear pre-mRNA splicing. *Nature 503*, 229–34.

[28] Wahl, M. C., Will, C. L., and Lührmann, R. The Spliceosome: Design Principles of a Dynamic RNP Machine. 2009.

[29] Moore, M. J., and Sharp, P. A. (1993) Evidence for two active sites in the spliceosome provided by stereochemistry of pre-mRNA splicing. *Nature 365*, 364–8.

[30] Nilsen, T. W., and Graveley, B. R. (2010) Expansion of the eukaryotic proteome by alternative splicing. *Nature 463*, 457–63.

[31] Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P., and Burge, C. B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature 456*, 470–476.

[32] Kornblihtt, A. R., Schor, I. E., All?, M., Dujardin, G., Petrillo, E., and Mu?oz, M. J. (2013) Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nat Rev Mol Cell Biol 14*, 153–165.

[33] Pandya-Jones, A., and Black, D. L. (2009) Co-transcriptional splicing of constitutive and alternative exons. *RNA 15*, 1896–1908.

[34] Tardiff, D. F., Lacadie, S. A., and Rosbash, M. (2006) A genome-wide analysis indicates that yeast pre-mRNA splicing is predominantly posttranscriptional. *Mol Cell 24*, 917–929.

[35] Montes, M., Becerra, S., S?nchez-?lvarez, M., and Su??, C. (2012) Functional coupling of transcription and splicing. *Gene 501*, 104–117.

[36] Alexander, R. D., Innocente, S. A., Barrass, J. D., and Beggs, J. D. (2010) Splicing-dependent RNA polymerase pausing in yeast. *Mol Cell 40*, 582–593.

[37] Morrissy, A. S., Griffith, M., and Marra, M. A. (2011) Extensive relationship between antisense transcription and alternative splicing in the human genome. *Genome Res 21*, 1203–1212.

[38] Houseley, J., and Tollervey, D. (2009) The Many Pathways of RNA Degradation. *Cell 136*, 763–776.

[39] Li de la Sierra-Gallay, I., Zig, L., Jamalli, A., and Putzer, H. (2008) Structural insights into the dual activity of RNase J. *Nature structural & molecular biology 15*, 206–12.

[40] Shatkin, a., and Manley, J. (2000) The ends of the affair: capping and polyadenylation. *Nature structural biology 7*, 1–5.

[41] Franks, T. M., and Lykke-Andersen, J. The Control of mRNA Decapping and P-Body Formation. 2008.

[42] Houseley, J., LaCava, J., and Tollervey, D. (2006) RNA-quality control by the exosome. *Nature reviews. Molecular cell biology 7*, 529–39.

[43] Shalon, D., Smith, S. J., and Brown, P. O. (1996) A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome research 6*, 639–45.

[44] Leung, Y. F., and Cavalieri, D. Fundamentals of cDNA microarray data analysis. 2003.

[45] Wang, Z., Gerstein, M., and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet 10*, 57–63.

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

[47] Wilhelm, B. T., Marguerat, S., Goodhead, I., and B?hler, J. (2010) Defining transcribed regions using RNA-seq. *Nat Protoc 5*, 255–266.

[48] Ozsolak, F., and Milos, P. M. (2011) RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet 12*, 87–98.

[49] Faghihi, M. A., and Wahlestedt, C. (2009) Regulatory roles of natural antisense transcripts. *Nature reviews. Molecular cell biology 10*, 637–643.

[50] Atkinson, S. R., Marguerat, S., and B?hler, J. (2012) Exploring long non-coding RNAs through sequencing. *Semin Cell Dev Biol 23*, 200–205.

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

[52] Takahashi, H., Lassmann, T., Murata, M., and Carninci, P. (2012) 5' end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. TL - 7. *Nature protocols 7 VN - re*, 542–561.

[53] Wilkening, S., Pelechano, V., Järvelin, A. I., Tekkedil, M. M., Anders, S., Benes, V., and Steinmetz, L. M. (2013) An efficient method for genome-wide polyadenylation site mapping and RNA quantification. *Nucleic acids research 41*, e65.

[54] Mata, J. (2013) Genome-wide mapping of polyadenylation sites in fission yeast reveals widespread alternative polyadenylation. *RNA Biol 10*, 1407–1414.

[55] Baejen, C., Torkler, P., Gressel, S., Essig, K., Söding, J., and Cramer, P. (2014) Transcriptome maps of mRNP biogenesis factors define pre-mRNA recognition. *Molecular cell 55*, 745–57.

[56] Jeffares, D. C., Penkett, C. J., and B?hler, J. (2008) Rapidly regulated genes are intron poor. *Trends Genet 24*, 375–378.

[57] Rabani, M., Raychowdhury, R., Jovanovic, M., Rooney, M., Stumpo, D. J., Pauli, A., Hacohen, N., Schier, A. F., Blackshear, P. J., Friedman, N., Amit, I., and Regev, A. (2014) High-resolution sequencing and modeling identifies distinct dynamic RNA regulatory strategies. *Cell 159*, 1698–1710.

[58] Pelechano, V., Chávez, S., and Pérez-Ortín, J. E. (2010) A complete set of nascent transcription rates for yeast genes. *PloS one 5*, e15442.

[59] Geisberg, J. V., Moqtaderi, Z., Fan, X., Ozsolak, F., and Struhl, K. (2014) Global analysis of mRNA isoform half-lives reveals stabilizing and destabilizing elements in yeast. *Cell 156*, 812–824.

[60] Miller, C., Schwalb, B., Maier, K., Schulz, D., D?mcke, S., Zacher, B., Mayer, A., Sydow, J., Marcinowski, L., D?lken, L., Martin, D. E., Tresch, A., and Cramer, P. (2011) Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. *Mol Syst Biol 7*, 458.

[61] Rabani, M., Levin, J. Z., Fan, L., Adiconis, X., Raychowdhury, R., Garber, M., Gnirke, A., Nusbaum, C., Hacohen, N., Friedman, N., Amit, I., and Regev, A. (2011) Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nat Biotechnol 29*, 436–442.

[62] Zeisel, A., Köstler, W. J., Molotski, N., Tsai, J. M., Krauthgamer, R., Jacob-Hirsch, J., Rechavi, G., Soen, Y., Jung, S., Yarden, Y., and Domany, E. (2011) Coupled pre-mRNA and mRNA dynamics unveil operational strategies underlying transcriptional responses to stimuli. *Molecular Systems Biology 7*, 529.

[63] Dölken, L., Ruzsics, Z., Rädle, B., Friedel, C. C., Zimmer, R., Mages, J., Hoffmann, R., Dickinson, P., Forster, T., Ghazal, P., and Koszinowski, U. H. (2008) High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. *RNA (New York, N.Y.) 14*, 1959–72.

[64] Marguerat, S., Lawler, K., Brazma, A., and Bähler, J. (2014) Contributions of transcription and mRNA decay to gene expression dynamics of fission yeast in response to oxidative stress. *RNA Biology 11*, 12.

[65] Munchel, S. E., Shultzaberger, R. K., Takizawa, N., and Weis, K. (2011) Dynamic profiling of mRNA turnover reveals gene-specific and system-wide regulation of mRNA decay. *Mol Biol Cell 22*, 2787–2795.

[66] Pai, A. A., Cain, C. E., Mizrahi-Man, O., De Leon, S., Lewellen, N., Veyrieras, J.-B., Degner, J. F., Gaffney, D. J., Pickrell, J. K., Stephens, M., Pritchard, J. K., and Gilad, Y. (2012) The contribution of RNA decay quantitative trait loci to inter-individual variation in steady-state gene expression levels. *PLoS Genet 8*, e1003000.

[67] Sun, M., Schwalb, B., Schulz, D., Pirkl, N., Etzold, S., Larivi?re, L., Maier, K. C., Seizl, M., Tresch, A., and Cramer, P. (2012) Comparative dynamic transcriptome analysis (cDTA) reveals mutual feedback between mRNA synthesis and degradation. *Genome Res 22*, 1350–1359.

[68] Shalem, O., Dahan, O., Levo, M., Martinez, M. R., Furman, I., Segal, E., and Pilpel, Y. (2008) Transient transcriptional responses to stress are generated by opposing effects of mRNA production and degradation. *Mol Syst Biol 4*, 223.

[69] Windhager, L. et al. (2012) Ultrashort and progressive 4sU-tagging reveals key characteristics of RNA processing at nucleotide resolution. *Genome Research 22*, 2031–2042.

[70] Wittenberg, C., and Reed, S. I. (2005) Cell cycle-dependent transcription in yeast: promoters, transcription factors, and transcriptomes. *Oncogene 24*, 2746–2755.

[71] Pramila, T., Wu, W., Miles, S., Noble, W. S., and Breeden, L. L. (2006) The Forkhead transcription factor Hcm1 regulates chromosome segregation genes and fills the S-phase gap in the transcriptional circuitry of the cell cycle. *Genes Dev 20*, 2266–2278.

[72] Simon, I., Barnett, J., Hannett, N., Harbison, C. T., Rinaldi, N. J., Volkert, T. L., Wyrick, J. J., Zeitlinger, J., Gifford, D. K., Jaakkola, T. S., and Young, R. A. (2001) Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell 106*, 697–708.

[73] Hu, Z., Killion, P. J., and Iyer, V. R. (2007) Genetic reconstruction of a functional transcriptional regulatory network. *Nat Genet 39*, 683–687.

[74] Wu, W.-S., Li, W.-H., and Chen, B.-S. (2006) Computational reconstruction of transcriptional regulatory modules of the yeast cell cycle. *BMC Bioinformatics 7*, 421.

[75] Lu, Y., and Cross, F. R. (2010) Periodic cyclin-Cdk activity entrains an autonomous Cdc14 release oscillator. *Cell 141*, 268–279.

[76] Orlando, D. A., Lin, C. Y., Bernard, A., Wang, J. Y., Socolar, J. E. S., Iversen, E. S., Hartemink, A. J., and Haase, S. B. (2008) Global control of cell-cycle transcription by coupled CDK and network oscillators. *Nature 453*, 944–947.

[77] Kovacs, L. A. S., Mayhew, M. B., Orlando, D. A., Jin, Y., Li, Q., Huang, C., Reed, S. I., Mukherjee, S., and Haase, S. B. (2012) Cyclin-dependent kinases are regulators and effectors of oscillations driven by a transcription factor network. *Mol Cell 45*, 669–679.

[78] Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J., and Davis, R. W. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell 2*, 65–73.

[79] Rustici, G., Mata, J., Kivinen, K., Li?, P., Penkett, C. J., Burns, G., Hayles, J., Brazma, A., Nurse, P., and B?hler, J. (2004) Periodic gene expression program of the fission yeast cell cycle. *Nat Genet 36*, 809–817.

[80] Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell 9*, 3273–3297.

[81] de Lichtenberg, U., Jensen, L. J., Fausb?ll, A., Jensen, T. S., Bork, P., and Brunak, S. (2005) Comparison of computational methods for the identification of cell cycle-regulated genes. *Bioinformatics 21*, 1164–1171.

[82] Granovskaia, M. V., Jensen, L. J., Ritchie, M. E., Toedling, J., Ning, Y., Bork, P., Huber, W., and Steinmetz, L. M. (2010) High-resolution transcription atlas of the mitotic cell cycle in budding yeast. *Genome Biol 11*, R24.

[83] Gauthier, N. P., Larsen, M. E., Wernersson, R., de Lichtenberg, U., Jensen, L. J., Brunak, S., and Jensen, T. S. (2008) Cyclebase.org–a comprehensive multi-organism online database of cell-cycle experiments. *Nucleic Acids Res 36*, D854–D859.

[84] Wichert, S., Fokianos, K., and Strimmer, K. (2004) Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics 20*, 5–20.

[85] Johansson, D., Lindgren, P., and Berglund, A. (2003) A multivariate approach applied to microarray data for identification of genes with cell cycle-coupled transcription. *Bioinformatics 19*, 467–473.

[86] Lu, X., Zhang, W., Qin, Z. S., Kwast, K. E., and Liu, J. S. (2004) Statistical

resynchronization and Bayesian detection of periodically expressed genes. *Nucleic Acids Res 32*, 447–455.

[87] Guo, X., Bernard, A., Orlando, D. a., Haase, S. B., and Hartemink, A. J. (2013) Branching process deconvolution algorithm reveals a detailed cell-cycle transcription program. *Proc Natl Acad Sci U S A 110*, E968–77.

[88] Bar-Joseph, Z., Gitter, A., and Simon, I. (2012) Studying and modelling dynamic biological processes using time-series gene expression data. *Nat Rev Genet 13*, 552–564.

[89] Mukherjee, C., Patil, D. P., Kennedy, B. A., Bakthavachalu, B., Bundschuh, R., and Schoenberg, D. R. (2012) Identification of cytoplasmic capping targets reveals a role for cap homeostasis in translation and mRNA stability. *Cell Rep 2*, 674–684.

[90] Kitagawa, M., Kitagawa, K., Kotake, Y., Niida, H., and Ohhata, T. (2013) Cell cycle regulation by long non-coding RNAs. *Cell Mol Life Sci*

[91] Romero-Santacreu, L., Moreno, J., Pérez-Ortín, J. E., and Alepuz, P. (2009) Specific and global regulation of mRNA stability during osmotic stress in Saccharomyces cerevisiae. *RNA (New York, N.Y.) 15*, 1110–1120.

[92] Gill, T., Cai, T., Aulds, J., Wierzbicki, S., and Schmitt, M. E. (2004) RNase MRP cleaves the CLB2 mRNA to promote cell cycle progression: novel method of mRNA degradation. *Mol Cell Biol 24*, 945–953.

[93] Trcek, T., Larson, D. R., Mold?n, A., Query, C. C., and Singer, R. H. (2011) Single-molecule mRNA decay measurements reveal promoter- regulated mRNA stability in yeast. *Cell 147*, 1484–1497.

[94] Jensen, L. J., Jensen, T. S., de Lichtenberg, U., Brunak, S., and Bork, P. (2006) Co-evolution of transcriptional and post-translational cell-cycle regulation. *Nature 443*, 594–597.

[95] Yu, H. (2007) Cdc20: a WD40 activator for a cell cycle degradation machine. *Mol Cell 27*, 3–16.

[96] Trcek, T., Chao, J. a., Larson, D. R., Park, H. Y., Zenklusen, D., Shenoy, S. M., and Singer, R. H. (2012) Single-mRNA counting using fluorescent in situ hybridization in budding yeast. *Nature protocols 7*, 408–19.

[97] Hartmann, H., Guthöhrlein, E. W., Siebert, M., Luehr, S., and Söding, J. (2013) P-value-based regulatory motif discovery using positional weight matrices. *Genome Research 23*, 181–194.

[98] MacIsaac, K. D., Wang, T., Gordon, D. B., Gifford, D. K., Stormo, G. D., and Fraenkel, E. (2006) An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics 7*, 113.

[99] de Lichtenberg, U., Wernersson, R., Jensen, T. S., Nielsen, H. B., Fausb?ll, A., Schmidt, P., Hansen, F. B., Knudsen, S., and Brunak, S. (2005) New weakly expressed cell cycle-regulated genes in yeast. *Yeast 22*, 1191–1201.

[100] Peng, X., Karuturi, R. K. M., Miller, L. D., Lin, K., Jia, Y., Kondu, P., Wang, L., Wong, L.-S., Liu, E. T., Balasubramanian, M. K., and Liu, J. (2005) Identification of cell cycle-regulated genes in fission yeast. *Mol Biol Cell 16*, 1026–1042.

[101] Mewes, H. W., Heumann, K., Kaps, A., Mayer, K., Pfeiffer, F., Stocker, S., and Frishman, D. MIPS: A database for genomes and protein sequences. 1999.

[102] Marguerat, S., Jensen, T. S., de Lichtenberg, U., Wilhelm, B. T., Jensen, L. J., and B?hler, J. (2006) The more the merrier: comparative analysis of microarray studies on cell cycle-regulated genes in fission yeast. *Yeast 23*, 261–277.

[103] Rowicka, M., Kudlicki, A., Tu, B. P., and Otwinowski, Z. (2007) High-resolution timing of cell cycle-regulated gene expression. *Proc Natl Acad Sci U S A 104*, 16892–16897.

[104] Koch, C., Moll, T., Neuberg, M., Ahorn, H., and Nasmyth, K. (1993) A role for the transcription factors Mbp1 and Swi4 in progression from G1 to S phase. *Science (New York, N.Y.) 261*, 1551–7.

[105] Darieva, Z., Clancy, A., Bulmer, R., Williams, E., Pic-Taylor, A., Morgan, B. A., and Sharrocks, A. D. (2010) A Competitive Transcription Factor Binding Mechanism Determines the Timing of Late Cell Cycle-Dependent Gene Expression. *Molecular Cell 38*, 29–40.

[106] Banerjee, N., and Zhang, M. Q. (2003) Identifying cooperativity among transcription factors controlling the cell cycle in yeast. *Nucleic Acids Res 31*, 7024–7031.

[107] Cheng, C., and Li, L. M. (2008) Systematic identification of cell cycle regulated transcription factors from microarray time series data. *BMC Genomics 9*, 116.

[108] Tsai, H.-K., Lu, H. H.-S., and Li, W.-H. (2005) Statistical methods for identifying yeast cell cycle transcription factors. *Proc Natl Acad Sci U S A 102*, 13532–13537.

[109] Wu, W.-S., and Li, W.-H. (2008) Systematic identification of yeast cell cycle transcription factors using multiple data sources. *BMC Bioinformatics 9*, 522.

[110] Mewes, H. W., Ruepp, A., Theis, F., Rattei, T., Walter, M., Frishman, D., Suhre, K., Spannagl, M., Mayer, K. F. X., St??mpflen, V., and Antonov, A. (2011) MIPS: Curated databases and comprehensive secondary data resources in 2010. *Nucleic Acids Research 39*.

[111] Ubersax, J. a., Woodbury, E. L., Quang, P. N., Paraz, M., Blethrow, J. D., Shah, K., Shokat, K. M., and Morgan, D. O. (2003) Targets of the cyclin-dependent kinase Cdk1. *Nature 425*, 859–864.

[112] Sherriff, J. A., Kent, N. A., and Mellor, J. (2007) The Isw2 chromatin-remodeling ATPase cooperates with the Fkh2 transcription factor to repress transcription of the B-type cyclin gene CLB2. *Molecular and cellular biology 27*, 2848–60.

[113] Rhee, H. S., and Pugh, B. F. (2012) Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature 483*, 295–301.

[114] Bauer, S., Robinson, P. N., and Gagneur, J. (2011) Model-based gene set analysis for bioconductor. *Bioinformatics 27*, 1882–1883.

[115] Huisinga, K. L., and Pugh, B. F. (2004) A genome-wide housekeeping role for TFIID and a highly regulated stress-related role for SAGA in Saccharomyces cerevisiae. *Molecular Cell 13*, 573–585.

[116] Wang, Y., Liu, C. L., Storey, J. D., Tibshirani, R. J., Herschlag, D., and Brown, P. O. (2002) Precision and functional specificity in mRNA decay. *Proc Natl Acad Sci U S A 99*, 5860–5865.

[117] Wu, T. D., and Nacu, S. (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics 26*, 873–881.

[118] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics 25*, 2078–2079.

[119] David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C. J., Bofkin, L., Jones, T., Davis, R. W., and Steinmetz, L. M. (2006) A high-resolution map of transcription in the yeast genome. TL - 103. *Proceedings of the National Academy of Sciences of the United States of America 103 VN* -, 5320–5325.

[120] Love, M. I., Huber, W., and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology 15*, 550.

[121] Anders, S., Pyl, P. T., and Huber, W. *HTSeq A Python framework to work with high-throughput sequencing data*; 2014; p 002824.

[122] Diao, L., Marcais, A., Norton, S., and Chen, K. C. (2014) MixMir: microRNA motif discovery from gene expression data using mixed linear models. *Nucleic acids research 1027*, 1–13.

[123] Zhou, X., and Stephens, M. (2012) Genome-wide efficient mixed-model analysis for association studies. *Nature genetics 44*, 821–4.

[124] Clément-Ziza, M., Marsellach, F. X., Codlin, S., Papadakis, M. A., Reinhardt, S., Rodríguez-López, M., Martin, S., Marguerat, S., Schmidt, A., Lee, E., Workman, C. T., Bähler, J., and Beyer, A. (2014) Natural genetic variation impacts expression levels of coding, non-coding, and antisense transcripts in fission yeast. *Molecular systems biology 10*, 764.

[125] Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England) 29*, 15–21.

[126] Käufer, N. F., and Potashkin, J. (2000) Analysis of the splicing machinery in fission yeast: a comparison with budding yeast and mammals. *Nucleic acids research 28*, 3003–3010.

[127] Roca, X., and Krainer, A. R. (2009) Recognition of atypical 5' splice sites by shifted base-pairing to U1 snRNA. *Nature structural & molecular biology 16*, 176–182.

[128] Lerner, M. R., Boyle, J. A., Mount, S. M., Wolin, S. L., and Steitz, J. A. (1980) Are snRNPs involved in splicing? *Nature 283*, 220–4.

[129] Allshire, R. C., Nimmo, E. R., Ekwall, K., Javerzat, J. P., and Cranston, G. (1995) Mutations derepressing silent centromeric domains in fission yeast disrupt chromosome segregation. *Genes & Development 9*, 218–233.

[130] Volpe, T. A., Kidner, C., Hall, I. M., Teng, G., Grewal, S. I. S., and Martienssen, R. A. (2002) Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science (New York, N.Y.) 297*, 1833–7.

[131] Dutrow, N., Nix, D. A., Holt, D., Milash, B., Dalley, B., Westbroek, E., Parnell, T. J., and Cairns, B. R. (2008) Dynamic transcriptome of Schizosaccharomyces pombe shown by RNA-DNA hybrid mapping. *Nat Genet 40*, 977–986.

[132] Rhind, N. et al. (2011) Comparative functional genomics of the fission yeasts. *Science 332*, 930–936.

[133] Wilhelm, B. T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C. J., Rogers, J., and B?hler, J. (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature 453*, 1239–1243.

[134] Schlackow, M., Marguerat, S., Proudfoot, N. J., Bähler, J., Erban, R., and Gullerova, M. (2013) Genome-wide analysis of poly(A) site selection in Schizosaccharomyces pombe. *RNA (New York, N.Y.) 19*, 1617–31.

[135] Duncan, C. D. S., and Mata, J. (2014) The translational landscape of fission-yeast meiosis and sporulation. *Nat Struct Mol Biol 21*, 641–647.

[136] Marguerat, S., Schmidt, A., Codlin, S., Chen, W., Aebersold, R., and B?hler, J. (2012) Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells. *Cell 151*, 671–683.

[137] Wood, V., Harris, M. A., McDowall, M. D., Rutherford, K., Vaughan, B. W., Staines, D. M., Aslett, M., Lock, A., Bähler, J., Kersey, P. J., and Oliver, S. G. (2012) PomBase: a comprehensive online resource for fission yeast. *Nucleic acids research 40*, D695–9.

[138] Perocchi, F., Xu, Z., Clauder-Münster, S., and Steinmetz, L. M. (2007) Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin D. *Nucleic acids research 35*, e128.

[139] Robinson, M., Mccarthy, D., and Smyth, G. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England) 26*, 139.

[140] Anders, S., and Huber, W. (2010) Differential expression analysis for sequence count data. *Genome biology 11*, R106.

[141] Martin, R. M., Rino, J., Carvalho, C., Kirchhausen, T., and Carmo-Fonseca, M. (2013) Live-cell visualization of pre-mRNA splicing with single-molecule sensitivity. *Cell reports 4*, 1144–55.

[142] Singh, J., and Padgett, R. A. (2009) Rates of in situ transcription and splicing in large human genes. *Nature structural & molecular biology 16*, 1128–1133.

[143] Witt, I., Kwart, M., Gross, T., and Käufer, N. F. (1995) The tandem repeat AGGGTAGGGT is, in the fission yeast, a proximal activation sequence and activates basal transcription mediated by the sequence TGTGACTG. *Nucleic acids research 23*, 4296–302.

[144] Tanay, A., Regev, A., and Shamir, R. (2005) Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast. *Proceedings of the National Academy of Sciences of the United States of America 102*, 7203–8.

[145] Barreau, C., Paillard, L., and Osborne, H. B. AU-rich elements and associated factors: Are there unifying principles? 2005.

[146] Shaw, G., and Kamen, R. (1986) A conserved AU sequence from the 3' untranslated region of GM-CSF mRNA mediates selective mRNA degradation. *Cell 46*, 659–667.

[147] Xie, X., Lu, J., Kulbokas, E. J., Golub, T. R., Mootha, V., Lindblad-Toh, K., Lander, E. S., and Kellis, M. (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature 434*, 338–345.

[148] Smith, D. J., Query, C. C., and Konarska, M. M. "Nought May Endure but Mutability": Spliceosome Dynamics and the Regulation of Splicing. 2008.

[149] Staley, J. P., and Guthrie, C. Mechanical devices of the spliceosome: Motors, clocks, springs, and things. 1998.

[150] Ehrensberger, A., Kelly, G., and Svejstrup, J. (2013) Mechanistic Interpretation of Promoter-Proximal Peaks and RNAPII Density Maps. *Cell 154*, 713–715.

[151] Pelechano, V., Wei, W., and Steinmetz, L. M. (2013) Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature 497*, 127–131.

[152] Xu, Z., Wei, W., Gagneur, J., Clauder-Münster, S., Smolik, M., Huber, W., and Steinmetz, L. M. (2011) Antisense expression increases gene expression variability and locus interdependency. *Molecular systems biology 7*, 468.

[153] Harris, M. R., Lee, D., Farmer, S., Lowndes, N. F., and de Bruin, R. A. M. (2013) Binding Specificity of the G1/S Transcriptional Regulators in Budding Yeast. *PLoS ONE 8*.

[154] Breeden, L. (1996) Start-specific transcription in yeast. *Current topics in microbiology and immunology 208*, 95–127.

[155] Deneke, C., Lipowsky, R., and Valleriani, A. (2013) Complex Degradation Processes Lead to Non-Exponential Decay Patterns and Age-Dependent Decay Rates of Messenger RNA. *PLoS ONE 8*.

[156] Montgomery, S. B., and Dermitzakis, E. T. (2011) From expression QTLs to personalized transcriptomics. *Nature Reviews Genetics 12*, 277–282.