Determinants of RNA metabolism in the *Schizosaccharomyces pombe* genome

Philipp Eser[1,2*], Leonhard Wachutka[2*], Kerstin C. Maier[1], Carina Demel[1], Mariana Boroni[2], Srignanakshi Iyer[2], Patrick Cramer[1], and Julien Gagneur[2]

[1]Max-Planck-Institute for Biophysical Chemistry, Department of Molecular Biology, Am Faßberg 11, 37077 Göttingen, Germany.
[2]Gene Center Munich and Department of Biochemistry, Center for Integrated Protein Science CIPSM, Ludwig-Maximilians-Universität München, Feodor-Lynen-Straße 25, 81377 Munich, Germany.

[*]These authors contributed equally

Correspondence:
P.C. (patrick.cramer@mpibpc.mpg.de); J.G. (gagneur@genzentrum.lmu.de).

ABSTRACT

To decrypt the regulatory code of the genome, sequence elements must be defined that determine the kinetics of RNA metabolism and thus gene expression. Here we attempt such decryption in an eukaryotic model organism, the fission yeast *S. pombe*. We first derive an improved genome annotation that redefines borders of 36% of expressed mRNAs and adds 487 non-coding RNAs (ncRNAs). We then combine 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 identify functional sequence elements in DNA and RNA that control RNA metabolic rates, and quantify 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.

INTRODUCTION

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 messenger RNA (mRNA) and the cellular concentration of non-coding RNA (ncRNA). The rates of both RNA degradation and splicing contribute to the time required for reaching mature RNA steady-state levels following transcriptional responses (Jeffares et al., 2008; Rabani et al., 2014).

To estimate the kinetics of RNA metabolic events genome-wide, techniques including genomic run-on followed by RNA polymerase chromatin Immuno-precipitation (Pelechano et al., 2010), cytoplasmic sequestration of RNA polymerase (Geisberg et al, Cell, 2014), or metabolic RNA labeling (Miller et al., 2011; Rabani et al., 2011; Schulz et al., 2013; Zeisel et al., 2011) has been performed in various organisms and under different conditions. 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 (Marguerat et al., 2014; Rabani et al., 2014; Schwanhäusser et al., 2011). RNA degradation modulates and fine-tunes mRNA abundance, largely varies across conditions and between organisms, and can be dynamically changed to shape gene expression (Eser et al., 2014; Munchel et al., 2011; Pai et al., 2012; Sun et al., 2012).

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 (Windhager et al., 2012) and mouse (Rabani et al., 2011, 2014) indicate that the rates of splicing also vary within a wide range. However, how these rates are quantitatively encoded in the genome remains largely unknown.

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 (Käufer and Potashkin, 2000) and conserved consensus splice site (SS) sequences (Lerner et al., 1980; Roca and Krainer, 2009), heterochromatin silencing (Allshire et al., 1995), and RNA interference (Volpe et al., 2002). 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 (Dutrow et al., 2008; Rhind et al., 2011; Wilhelm et al., 2008), a map of polyadenylation sites (Mata, 2013; Schlackow et al., 2013), the 'translatome' as measured by ribosome profiling (Duncan and Mata, 2014), and an absolute quantification of protein and RNA (Marguerat et al., 2012).

Here 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. We provide an improved genome annotation and a quantitative description of RNA metabolism for an important eukaryotic model organism. The approach developed here enables quantitative, genome-wide studies of eukaryotic gene regulation, and provides a general route to help decrypting the regulatory code of the genome.

RESULTS

Strategy to describe RNA metabolism and regulatory elements

Our approach consists of three steps (Figure 1). First, we performed short and progressive metabolic labeling of RNA with 4-thiouracil coupled with strand-specific RNA-Seq (4tU-Seq, Materials and Methods). 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 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*.

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, at mean per-base read coverage of 385) from fission yeast grown in rich media (Materials and Methods). Genomic intervals of apparently uninterrupted transcription (Transcriptional Units, TUs, Figure 2A) were identified with a segmentation algorithm applied to the RNA-Seq read coverage signal (Materials and Methods). The three parameters of the algorithm, the minimum per base coverage, the minimum TU length, and the maximum gap within TUs were chosen to best match the existing genome annotation (Pombase version 2.22 (Wood et al., 2012), Figure S1A,B). TUs that did not show significant signal in the 4tU-Seq dataset were considered as artifacts and discarded (Materials and Methods).

The segmentation led to a total of 5,484 TUs (Figure 2B, Table S1), 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, Table S2), and no evidence for substantial alternative splicing at any given intron or circular RNAs was found (at least 10 supporting reads, Materials and Methods). These observations are in line with previous RNA-Seq studies of *S. pombe* showing that alternative splicing is prevalent but rare (Bitton et al., 2015; Rhind et al., 2011). A total of 402 ORFs (8%) in the existing annotation (Wood et al., 2012) were not recovered (Figure 2C, Materials and Methods), apparently because they were not expressed under the used growth condition (gene set enrichment analysis, Figure S1C).

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 (Materials and Methods, Figure 2B, C). 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 (Perocchi et al., 2007). 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 S1D). The remaining half non-recovered RNAs might be genuine ncRNAs that are not expressed in our growth condition. 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), consistent with a previously curated set of ORF transcript boundaries (Figure S1E, Lantermann et al., 2010). 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 S1F; for an example see Figure 2A). Our revised transcript 3'-ends were centered around experimentally mapped

polyadenylation (polyA) sites (Mata, 2013), whereas the previously annotated 3'-ends typically extended well beyond polyA sites (median difference = 3 nt versus 45 nt, Figure 2D). 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 aberrantly long UTRs.

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 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. These reads from unspliced RNA gradually ceased during the time course (Figure 3A,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 3C). 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 3C, D). We included in the model scaling factors that account for variations in sequencing depth, an overall increase of the labeled RNA fraction, cross-contamination of unlabeled RNA, and 4sU label incorporation efficiency (Materials and Methods). The model was fitted using maximum likelihood and assuming negative binomial distribution to cope with overdispersion of read counts (Robinson et al. 2010, Anders and Huber 2010).

Our method yields absolute splicing and degradation rates, but provides synthesis rates up to a scaling factor common to all TUs. Absolute synthesis rates were obtained by

scaling all values so that the median steady-state level of ORF-TUs matches the known median of 2.4 mRNAs per cell (Marguerat et al., 2012). To facilitate comparisons of the obtained RNA metabolic rates, we present the synthesis rate as the average time elapsed between the production of two transcripts in a single cell ('synthesis time'); the degradation rate as the time needed to degrade half of the mature RNAs ('half-life'); and the splicing rate as the time to process half of the precursor RNA junction ('splicing time') (Table S1, S2).

The synthesis times and half-lives inferred from distinct splice junctions of the same TU agreed well, demonstrating the robustness of our approach (Spearman rank correlation = 0.44 for synthesis time, $P < 2 \times 10^{-16}$ and Spearman rank correlation = 0.79 for half-life, $P < 2 \times 10^{-16}$, Figure 3E and S2A). 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, see below), 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 takes as input all reads overlapping the exon was used (exon model, Figure S2B). When applied to intron-containing genes, parameter estimates with the exon model were consistent with those obtained with the splice junction model (Figure S2C,D). Overall, synthesis and degradation rates correlated well with previous estimates from microarray data (Sun et al., 2012, Spearman rank correlation = 0.45, $P < 2 \times 10^{-16}$ for synthesis rate and Spearman rank correlation = 0.74, $P < 2 \times 10^{-16}$ for half-life, Figure S2G,H), strongly supporting our rate estimation procedure.

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 3F-H). These results are consistent with splicing of beta-globin introns within 20 to 30 sec as measured by *in vivo* single RNA imaging (Martin et al., 2013), and argue against earlier slower estimates for splicing times of 5 to 10 min (Singh and Padgett, 2009). Notably, ncTUs were synthesized at a significantly lower rate than ORF-TUs (median synthesis times of 23 min and 6.1 min, respectively, $P < 2 \times 10^{-16}$, Wilcoxon test), and were degraded slightly faster (median half-life of 12 min for ORF-TUs versus 7.4 min for ncTUs, $P < 2 \times 10^{-16}$, 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. Moreover, splicing time did differ significantly between the two transcript classes (median splicing time of 0.7 min for ORF-TUs versus 1.5 min for ncTUs, $P = 1.3 \times 10^{-4}$, Wilcoxon test). 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 3F-H). 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.

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 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, Materials and Methods). 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 (Materials and Methods). We found 12 motifs that significantly associated with RNA metabolism kinetics (Figure 4A). 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 4B).

Determinants of high expression

Motifs that were predictive of RNA synthesis times were only found in the promoter region, further validating our approach (Figure 4A). 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 (Witt et al., 1995 and Tanay et al, 2005), the 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 4B, 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 4C).

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 4B), 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 4F, Figure S3B), but also with increased half-life (50% and 31%) of the corresponding RNAs (Figure S3A, Figure S3C), likely because those RNAs are both highly synthesized and stable.

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 4C), and associated with a decrease in half-life by 30% per motif instance (Linear regression, Figure S3D). Four AC-rich motifs were found (CAACCA, AACCAC, ACCAAC, and

CCAACA) in 5' UTRs, preferentially located near the TU 5'end (Figure 4D) and were associated with an increased RNA half-life (Figure S3E-H). 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.

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 (Barreau et al., 2005; Shaw and Kamen, 1986) 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 4G). 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 4H, Figure S3I). These two motifs were found in a large number of ORF-TUs (466 and 514, 11% and 13% respectively, Figure 4B), and were co-occurring (FDR < 0.1, Figure 4B), 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 4E), 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 (Xie et al., 2005), indicating that their function is conserved from *S. pombe* to mammals.

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 (Materials and Methods). 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 5A, 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 (Figure 5A). 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 5B). Changes in positions flanking the motif have minor effects but may play functional roles (Figure 5A, B). Nucleotides associated with important effects tended to also be more frequent (Sequence logo, Figure 5A,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 (Figure S4).

New regulatory motifs predict effects of cis-regulatory variants

To further provide evidence for 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 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 (Clément-Ziza et al., 2014). 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 subpopulations: 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 subpopulations. Hence, differential gene expression between the two subpopulations 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 (Materials and Methods), 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 S5). 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 5C, $P = 2 \times 10^{-10}$, one-sided Wilcoxon test) consistent with the predicted 1.35-fold increased synthesis time (Figure 5A).

Two variants acting in an opposite fashion strongly supported 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 5B). Consistently, 7.A>G substitution occurring on the gene *SPCC794.06* led to a significantly increased expression level (Figure 5D, $P = 2 \times 10^{-4}$) whereas the (7.G>A) in the gene *mug65* led to a significantly decreased expression level (Figure 5E, $P = 10^{-4}$). 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, $P = 9 \times 10^{-4}$, Figure 5F), demonstrating the model predicted quantitatively the effects of single mutations and providing strong evidence for the functional role of these motifs.

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, Materials and Methods, Figure 6A). 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 (Smith et al., 2008; Staley and Guthrie, 1998)). 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 (Smith et al., 2008). 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 6A). Also, splicing time anti-correlated with the frequency of the core branch site sequence across the genome (Figure 6B). 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, $P = 5 \times 10^{-4}$, Figure 6C). Positions within the branch site region and downstream of the 5'SS are most commonly found as consensus and showed the largest effect on splicing kinetics (Figure 6C). The last nucleotide of the 5' exon is generally a guanine but did not influence splicing time (Figure 6C, 5'SS-1 position), indicating that other sources of selection influence this position.

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 = -0.16, $P < 2 \times 10^{-16}$, Figure S5A, B), 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 (Rabani et al., 2014). 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 6D) as in mouse cells (Rabani et al., 2014), independently of the relative position of the intron within the transcript (Fig S5C-E). 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 and Figure S5F). 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, $P < 2 \times 10^{-16}$, Figure 6E), in agreement with results in mouse (Rabani et al., 2014). 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 (Singh and Padgett, 2009), because synthesis rate is mostly determined by the rate of transcription initiation rather than elongation (Ehrensberger et al., 2013). Altogether, multivariate analysis (Materials and Methods) 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).

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 (Pelechano and Steinmetz, 2013; Xu et al., 2011). In our revised genome annotation, convergent TUs generally did not overlap (1022 out of 1616), typically leaving 75 bp of untranscibed sequence in between (Figure 7A). 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 (Figure S6, 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 7B). In fission yeast, antisense transcription could repress sense RNA synthesis, as in *S. cerevisiae* (Schulz et al., 2013), 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, $P = 9 \times 10^{-7}$), 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 7C). However, no difference regarding mRNA stability was observed (Figure 7D). 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.

DISCUSSION

Here, by combining metabolically labeled RNA profiling at high temporal resolution with computational kinetic modeling, we obtained *in vivo* RNA synthesis, splicing and degradation rates across an entire eukaryotic genome, providing insights into RNA metabolism and its sequence determinants. In addition, our systematic annotation of the transcribed genome of *S. pombe* redefines most ncRNAs and a large fraction of UTRs in mRNAs, in particular 5'UTRs and thus promoter regions. Hence, our data will be an important resource for the *S. pombe* community, and our approach will be of general use for systems biology studies of eukaryotic gene expression and its regulation.

So far, only Rabani and colleagues (2014), using mouse cells, have reported a computational tool to access RNA metabolism and 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. Multiple lines of evidence based on independent data, down to the contribution of individual base to splicing recapitulating nucleotide interactions between the precursor RNA and the spliceosome, support the high quality of our dataset.

We further introduced an approach to discover regulatory elements in the genome that combines *in vivo* quantification of RNA metabolic rates with robust regression on DNA sequence. 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 AC-rich sequences in promoters and in 5'UTRs. Our approach has several advantages. Conservation analysis is difficult in genomic regions that align poorly, as is often the case for regulatory regions, and give confounding results because selection can have regulatory and non-regulatory origins. By analyzing the relation between sequence and individual RNA metabolic rates, we uncouple the contribution of sequence elements to each step of RNA metabolism. Whereas standard motif enrichment analysis discriminates 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. Our method is general and can be applied to other organisms. Moreover it could be extended to study other layers of gene regulation such as ribosome recruitment, or translation.

Functional evidence of the discovered motifs was obtained by exploiting existing transcription profiles of genetically distinct strains. To this end, the analysis was restricted to genes harboring a single variant across the promoter and the whole gene body. Although one cannot exclude on every single gene that further independent mutations in linkage are causative of the observed expression changes, the agreement for each motif over multiple genes in direction and amplitude strongly indicate the functionality of the motifs. Transcription profiles across genetically distinct individuals are increasingly available and include recombinant panels of model organisms such as *S. cerivisiae*, fly, mouse, *A. thaliana*, and human. Hence, our approach could help interpreting the transcription profiling in human individuals. 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 (Montgomery and Dermitzakis, 2011).

MATERIALS AND METHODS

Strains

All experiments were done with the strain ED666 (BIONEER) (h+, *ade6*-M210, *ura4*-D18, *leu1*-32).

4tU labeling and RNA extraction

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 OD_{600} 0.1 and grown to OD_{600} 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 (Sun et al., 2012).

Sequencing

rRNA was depleted using the Ribo-Zero™ 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).

RNA-seq read mapping

Single- and paired-end RNA-seq reads were mapped to the reference genome (ASM294v2.26) with GSNAP (Wu and Nacu, 2010), allowing for novel splice site identification (Expanded view).

Transcriptional Units and their classification

Transcriptional units (TUs) were identified using a min-length max-gap algorithm on binarized RNA-Seq coverage track (Expanded view) resulting in 5,596 TUs. Of these, 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.

Read counts per exon, intron and splice junctions

Counts of reads aligning completely within exons or introns were obtained with the software HTseq-counts (Anders et al., 2014) 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 python 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.

Rate estimation, identification of sequence elements predictive for rates and linear regression, and analysis of recombinant strain panel

Detailed descriptions are found in the Expanded view.

AUTHOR CONTRIBUTIONS

Conceived and designed the experiments: PC. Performed the experiments: KM. Analyzed the data: PE LW MB JG CD SI. Wrote the paper: PE JG PC LW MB.

ACKNOWLEDGMENTS

We are thankful to Mario Halic for helpful discussions on genome annotation and to Mathieu Clément-Ziza for data sharing and analysis advices.

PE was supported by the Deutsches Konsortium für translationale Krebsforschung DKTK. PC was supported by the Deutsche Forschungsgemeinschaft, the Advanced Grant TRANSIT of the European Research Council, and the Volkswagen Foundation. JG was supported by the Bavarian Research Center for Molecular Biosystems and by the Bundesministerium für Bildung und Forschung through the Juniorverbund in der Systemmedizin “mitOmics” (FKZ 01ZX1405A). MB was supported by CNPq, Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brasil. CD was supported by a DFG Fellowship through the Graduate School of Quantitative Biosciences Munich (QBM).

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES

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 Dev. *9*, 218–233.

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

Barreau, C., Paillard, L., and Osborne, H.B. (2005). AU-rich elements and associated factors: Are there unifying principles? Nucleic Acids Res. *33*, 7138–7150.

Bitton, D.A., Atkinson, S.R., Rallis, C., Smith, G.C., Ellis, D.A., Chen, Y., Malecki, M., Codlin, S., Cotobal, C., Lemay, J.-F., et al. (2015). Widespread exon-skipping triggers degradation by nuclear RNA surveillance in fission yeast. Genome Res.

Clément-Ziza, M., Marsellach, F.X., Codlin, S., Papadakis, M.A., Reinhardt, S., Rodríguez-López, M., Martin, S., Marguerat, S., Schmidt, A., Lee, E., et al. (2014). Natural genetic variation impacts expression levels of coding, non-coding, and antisense transcripts in fission yeast. Mol. Syst. Biol. *10*, 764.

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

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.

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

Eser, P., Demel, C., Maier, K.C., Schwalb, B., Pirkl, N., Martin, D.E., Cramer, P., and Tresch, A. (2014). Periodic mRNA synthesis and degradation co-operate during cell cycle gene expression. Mol. Syst. Biol. *10*, 717.

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

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

Lantermann, A.B., Straub, T., Str?lfors, A., Yuan, G.-C., Ekwall, K., and Korber, P. (2010). Schizosaccharomyces pombe genome-wide nucleosome mapping reveals positioning mechanisms distinct from those of Saccharomyces cerevisiae. Nat Struct Mol Biol *17*, 251–257.

Lerner, M.R., Boyle, J.A., Mount, S.M., Wolin, S.L., and Steitz, J.A. (1980). Are snRNPs involved in splicing? Nature *283*, 220–224.

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

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 Biol. *11*, 12.

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 Rep. *4*, 1144–1155.

Mata, J. (2013). Genome-wide mapping of polyadenylation sites in fission yeast reveals widespread alternative polyadenylation. RNA Biol. *10*, 1407–1414.
Miller, C., Schwalb, B., Maier, K., Schulz, D., Dümcke, S., Zacher, B., Mayer, A., Sydow, J., Marcinowski, L., Dölken, L., et al. (2011). Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. Mol. Syst. Biol. *7*, 458.
Montgomery, S.B., and Dermitzakis, E.T. (2011). From expression QTLs to personalized transcriptomics. Nat. Rev. Genet. *12*, 277–282.
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.
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., et al. (2012). The contribution of RNA decay quantitative trait loci to inter-individual variation in steady-state gene expression levels. PLoS Genet. *8*, e1003000.
Pelechano, V., and Steinmetz, L.M. (2013). Gene regulation by antisense transcription. Nat Rev Genet *14*, 880–893.
Pelechano, V., Chávez, S., and Pérez-Ortín, J.E. (2010). A complete set of nascent transcription rates for yeast genes. PLoS One *5*, e15442.
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 Res. *35*, e128.
Rabani, M., Levin, J.Z., Fan, L., Adiconis, X., Raychowdhury, R., Garber, M., Gnirke, A., Nusbaum, C., Hacohen, N., Friedman, N., et al. (2011). Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. Nat. Biotechnol. *29*, 436–442.
Rabani, M., Raychowdhury, R., Jovanovic, M., Rooney, M., Stumpo, D.J., Pauli, A., Hacohen, N., Schier, A.F., Blackshear, P.J., Friedman, N., et al. (2014). High-Resolution Sequencing and Modeling Identifies Distinct Dynamic RNA Regulatory Strategies. Cell *159*, 1698–1710.
Rhind, N., Chen, Z., Yassour, M., Thompson, D.A., Haas, B.J., Habib, N., Wapinski, I., Roy, S., Lin, M.F., Heiman, D.I., et al. (2011). Comparative functional genomics of the fission yeasts. Science *332*, 930–936.
Roca, X., and Krainer, A.R. (2009). Recognition of atypical 5' splice sites by shifted base-pairing to U1 snRNA. Nat. Struct. Mol. Biol. *16*, 176–182.
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 *19*, 1617–1631.
Schulz, D., Schwalb, B., Kiesel, A., Baejen, C., Torkler, P., Gagneur, J., Soeding, J., and Cramer, P. (2013). Transcriptome surveillance by selective termination of noncoding RNA synthesis. Cell *155*, 1075–1087.
Schwanh?usser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. Nature *473*, 337–342.
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.

Singh, J., and Padgett, R.A. (2009). Rates of in situ transcription and splicing in large human genes. Nat. Struct. Mol. Biol. *16*, 1128–1133.
Smith, D.J., Query, C.C., and Konarska, M.M. (2008). "Nought May Endure but Mutability": Spliceosome Dynamics and the Regulation of Splicing. Mol. Cell *30*, 657–666.
Staley, J.P., and Guthrie, C. (1998). Mechanical devices of the spliceosome: motors, clocks, springs, and things. Cell *92*, 315–326.
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.
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 *297*, 1833–1837.
Wilhelm, B.T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C.J., Rogers, J., and Bähler, J. (2008). Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature *453*, 1239–1243.
Windhager, L., Bonfert, T., Burger, K., Ruzsics, Z., Krebs, S., Kaufmann, S., Malterer, G., L'hernault, A., Schilhabel, M., Schreiber, S., et al. (2012). Ultrashort and progressive 4sU-tagging reveals key characteristics of RNA processing at nucleotide resolution. Genome Res. *22*, 2031–2042.
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., et al. (2012). PomBase: a comprehensive online resource for fission yeast. Nucleic Acids Res. *40*, D695–D699.
Wu, T.D., and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics *26*, 873–881.
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.
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. Mol. Syst. Biol. *7*, 468.
Zeisel, A., Köstler, W.J., Molotski, N., Tsai, J.M., Krauthgamer, R., Jacob-Hirsch, J., Rechavi, G., Soen, Y., Jung, S., Yarden, Y., et al. (2011). Coupled pre-mRNA and mRNA dynamics unveil operational strategies underlying transcriptional responses to stimuli. Mol. Syst. Biol. *7*, 529.

FIGURE LEGENDS

Figure 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.

Figure 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 aberrantly long UTRs (e.g. 3'UTR of SPAC3G9.13c as TU.0899).

(B) Classification (Materials and Methods) 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).

(C) 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).

(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).

Figure 3. 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-state) along the UTRs (white boxes), the exons (dark boxes) and the introns (lines) of the TU encoding *cdc2*.
(B) Distribution of sequencing-depth normalized unspliced junction read counts (top panel) and normalized spliced junction read counts (lower panel) for the complete 4tU-Seq time series and the steady-state RNA-Seq samples.
(C) Schema of the junction first-order kinetics model. Each splice junction is modeled individually, assuming constant synthesis time, splicing time and half-life. Unspliced junction reads (blue) are specific to the precursor RNA and spliced junction reads (red) are specific to the mature RNA.
(D) Observed (circles) and fitted (lines) splice junction counts for the first intron of TU.0597 (*php3*). Unspliced (blue) and spliced (red) normalized counts (y-axis) are shown for all 4tU-Seq samples and the steady-state sample (x-axis).
(E) Half-life estimated from the first (x-axis) versus the second (y-axis) splice junction on TUs with two or more introns.
(F,G,H) Distribution of synthesis times (F), half-lives (G) and splicing times (H) for ORF-TUs (blue) and ncTUs (light blue). Median indicated as vertical line.

Figure 4. 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. All motif instances are provided in Table S3.
(C) 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).

(D), (E). Same as (C) for the 5'UTR motifs (D) and for the 3'UTR motifs with respect to polyA site (E). No positional preference was found when aligning 3'UTRs with respect to stop codon and 5'UTRs with respect to start codon.

F) Distributions of synthesis time among ORF-TUs that have zero, one or more than one occurrence of the motif CAGTCACA in their promoter sequence.

G) Distributions of half-lives of ORF-TUs that have zero, one, two or more than two occurrence(s) of the motif TATTTAT in their 3'UTR sequence.

H) Distributions of half-lives of ORF-TUs that have zero, one or more than one occurrence of the motif TTAATGA in their 3'UTR sequence.

Figure 5. 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). Coefficients for all motifs are available in Table S4.

(B) As in (A) for the 3' UTR motif TTAATGA.

(C,D,E) 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* (C), *SPCC794.06* (D), and *mug65* (E)

(F) 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.

Figure 6. 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. All coefficients are provided in Table S4.

(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)

Figure 7. 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.

Figure S1. Segmentation algorithm parameters and antisense artifacts in current genome annotation

(A) 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).

(B) 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.

(C) Top ten GO terms enriched (model-based gene set analysis, Bauer et al. 2011) among 402 non-recovered protein coding genes from Pombase.

(D) Sense mean coverage (x-axis) versus antisense mean coverage (y-axis) of 1011 non-recovered ncRNAs of the current annotation.
(E) Differences between 5' and 3' ends of ORF-TUs and Pombase v.2.22 (first and second box), difference between TUs 3'ends and polyA-sites mapped by Mata et al. (2013) (third box), and differences between 5' and 3' ends of ORFs defined by Lantermann et al. (2010) and Pombase v.2.22 (fourth and fifth box).
(F) 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.

Figure S2. Modeling RNA kinetics

(A) Estimate of synthesis time for ORF-TUs with two introns based on the first intron (x-axis) against estimate based on the second intron (y-axis).
(B) Exon-only model used to estimate synthesis times and half-lives of intronless TUs.
(C) Half-lives of intron-containing TUs estimated using the junction model (x-axis) versus the exon model (y-axis).
(D) As in (C) for synthesis times.
(E) Synthesis time of intron-containing TUs estimated using the exon model on the first exon (x-axis).
(F) As in (E) for half-lives..
(G) Comparison of synthesis times of ORF-TUs between this study (x-axis) against synthesis rates publised in (Sun et al. 2012).
(H) As in (G) for half-lives.

Figure S3.

(A) Boxplot showing the distribution of half-lives for all ORF-TUs, grouped by the number of occurences of the motif CAGTCACA in their promoter sequence. Number of instances per box in parentheses.
(B-I) As in (A) for all identified motifs and corresponding rates.

Figure S4.

(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.

Figure S5.

(A) Mature mRNA length (x-axis) of all ORF-TUs with the splicing time (y-axis) of the corresponding introns.
(B) Intron length (x-axis) versus splicing time (y-axis) for all introns.
(C-E) Intron position specific distributions of the deviation from the mean splicing time for ORF-TUs with two (C), three (D) or four (E) introns.
(F) Distribution of splicing time of individual introns predicted from their sequence at 5'SS, 3'SS and BS (y-axis) versus the total number of introns in the corresponding ORF-TU (x-axis). This plot is to be compared to Figure 6D.

Figure S6.

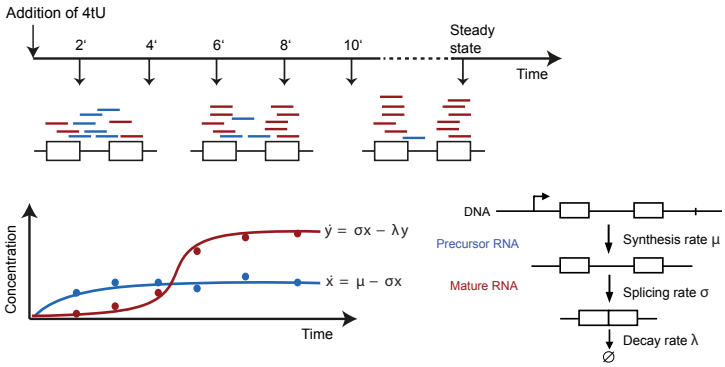
Observed and expected number of TU termination events in CDS, intron, 5'UTR and 3'UTR of antisense ORF-TUs. Expected counts are estimated by 999 times randomization of all overlapping sense-antisense TU-pairs. Dark grey bars show the mean and the range which contains 90% of expected counts ("error bars").

SUPPLEMENTAL TABLES

Supplemental table 1: Transcriptional units
Supplemental table 2: Splice junctions
Supplemental table 3: Motif sites
Supplemental table 4: Linear model coefficients
Supplemental table 5: Motifs affected by genetic variation

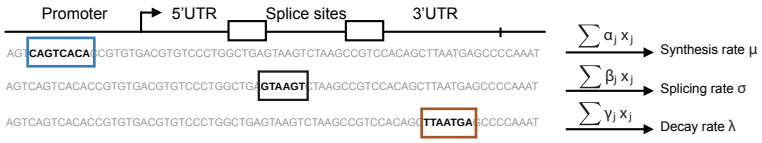
1. Genome-wide *in vivo* RNA metabolism kinetics

- 4tU RNA labeling and deep sequencing in fission yeast
- Mathematical modeling
- RNA metabolism rates for 5,484 genes and 4,958 introns



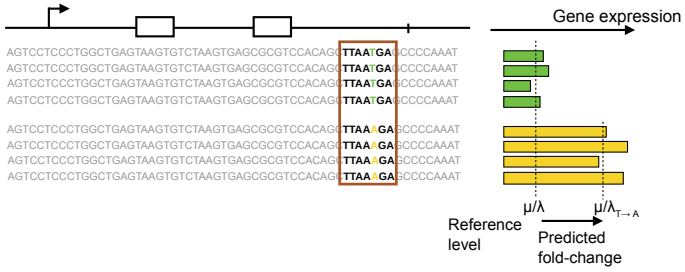
2. Predict RNA metabolism rates from DNA sequence

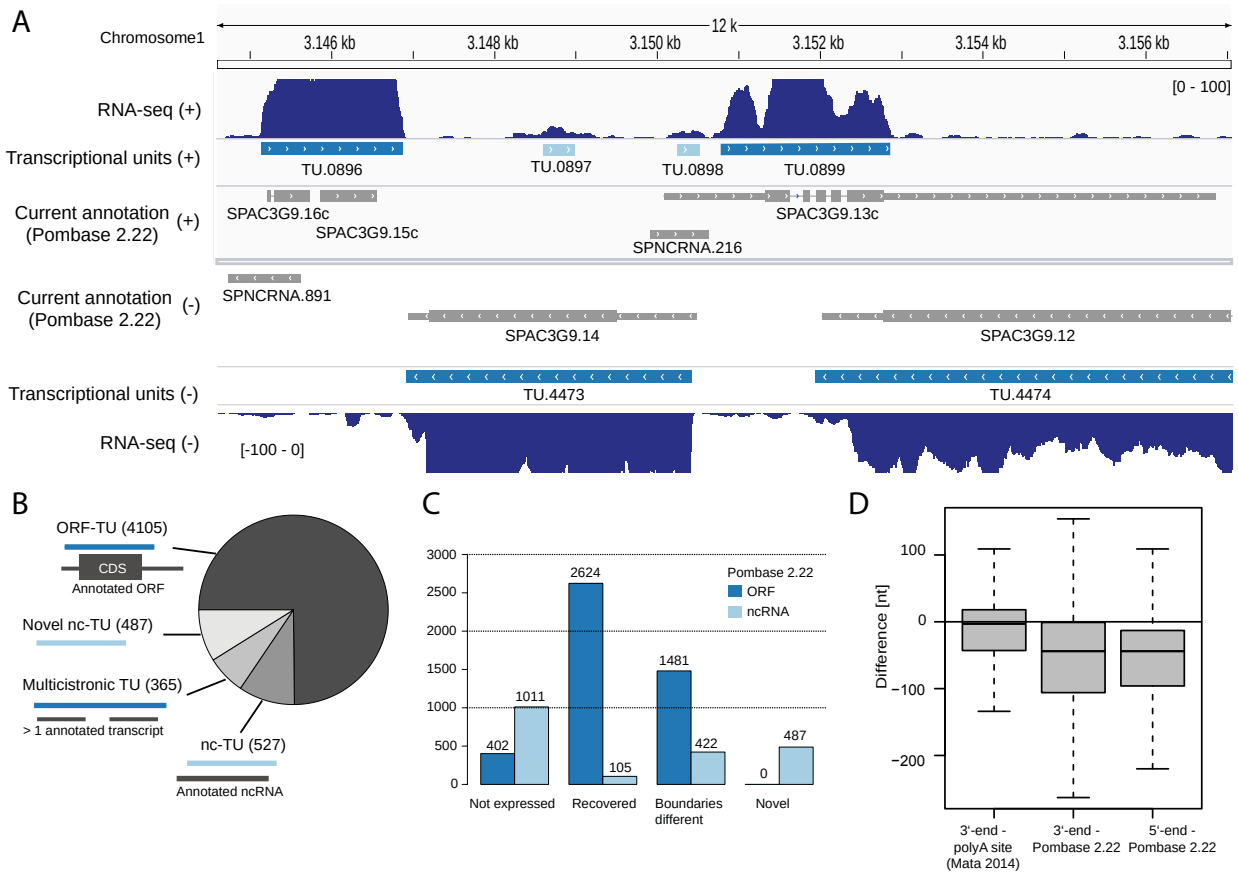
- De novo identification of predictive regulatory elements
- Quantification of single nucleotide effects on rates

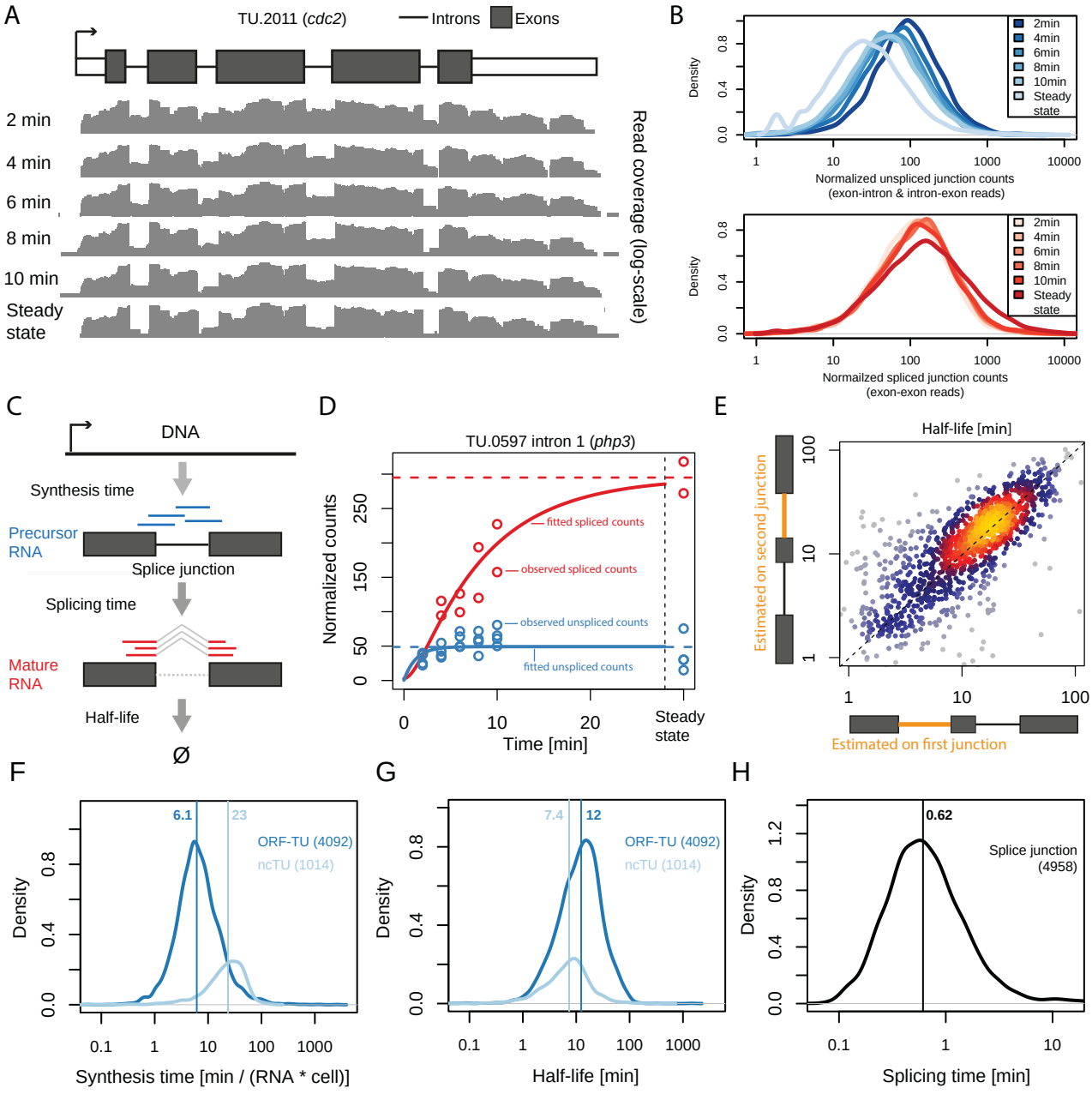


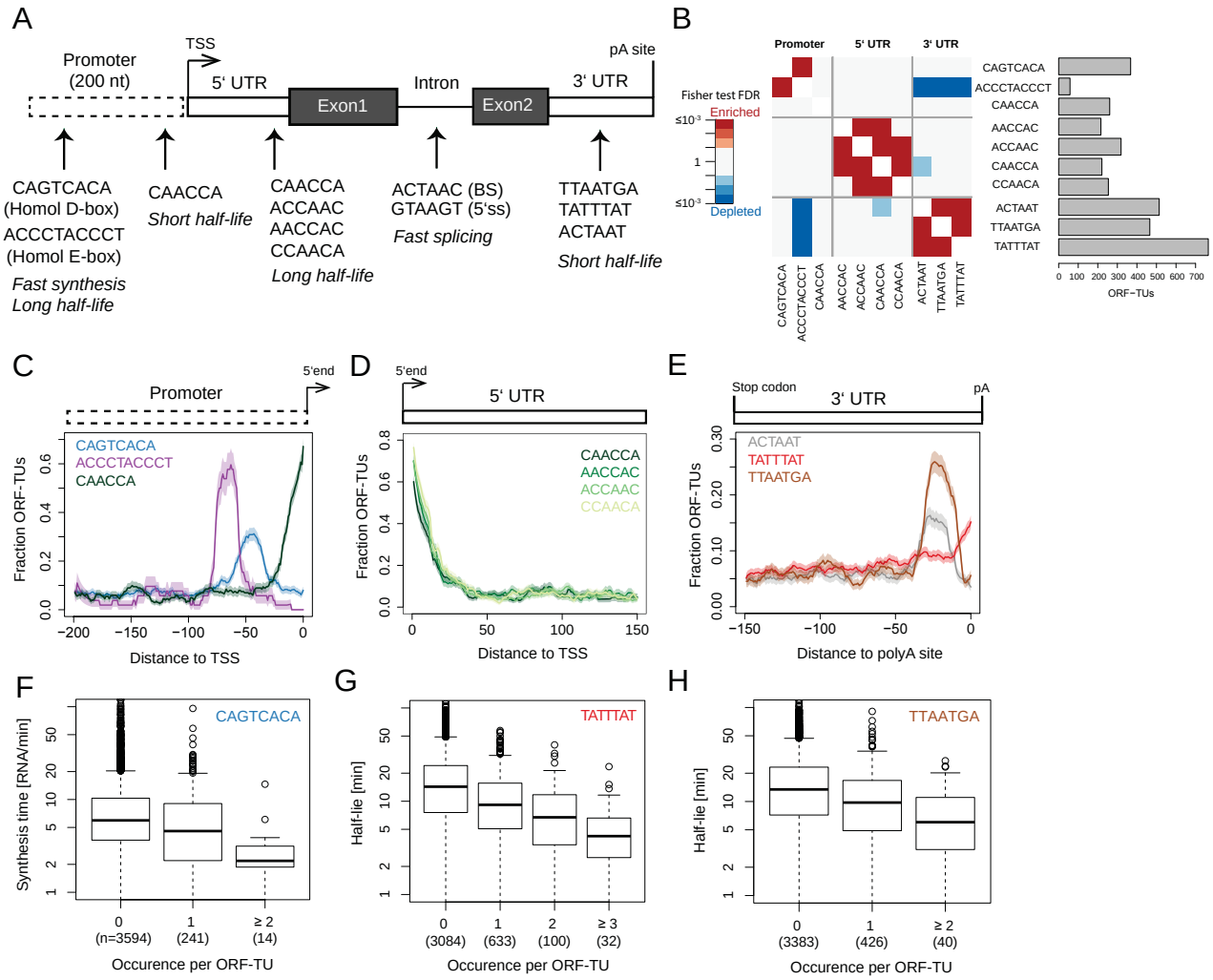
3. Validate regulatory elements using genetically distinct individuals

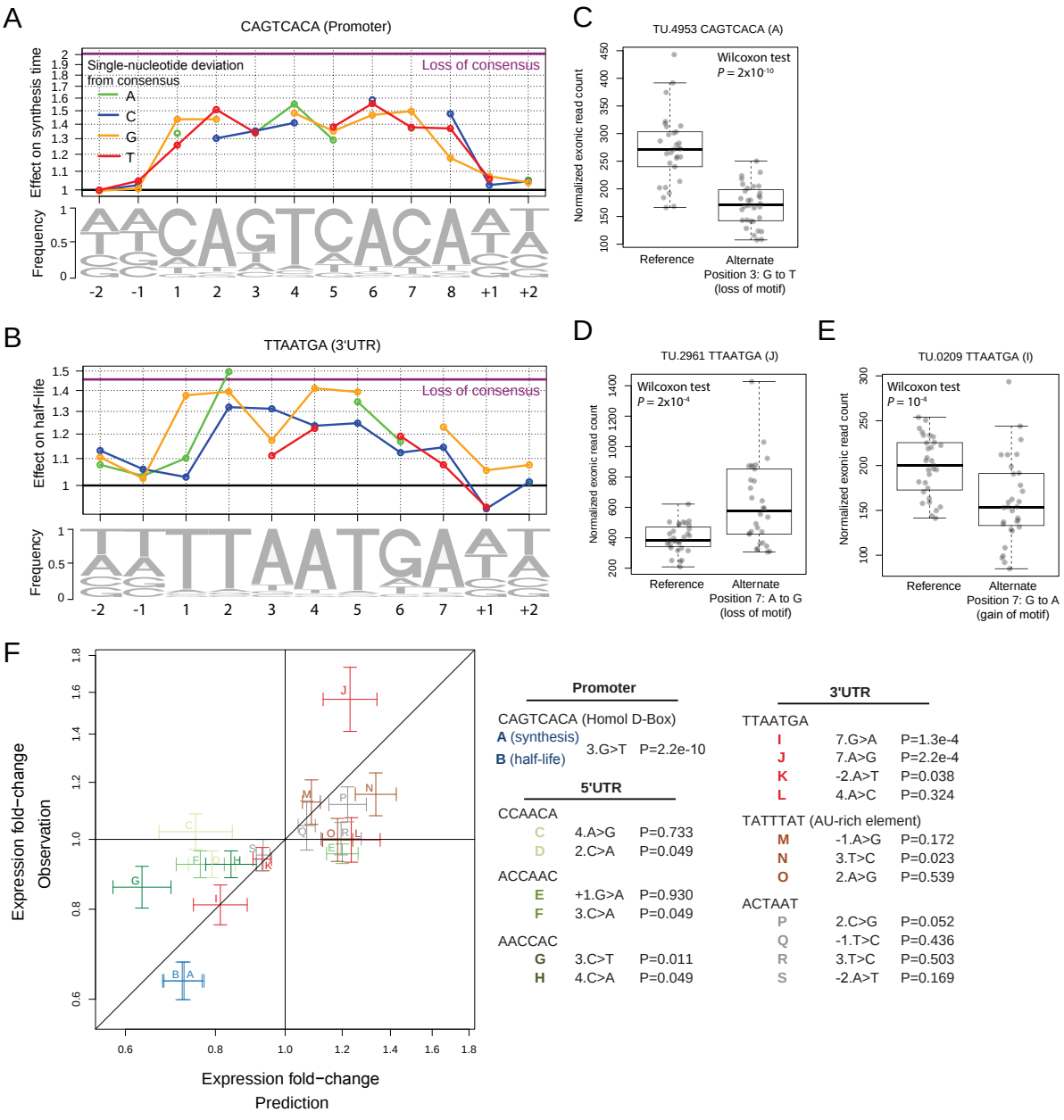
- Gene expression level across 44 strains of a recombinant strain panel (Clement-Ziza et al. 2015)
- Matched fold-changes with predictions demonstrate causality











Promoter

CAGTCACA (Homol D-Box)

A (synthesis)
B (half-life)

3.G>T	P=2.2e-10	

5'UTR

CCAACA

C	4.A>G	P=0.733
D	2.C>A	P=0.049

ACCAAC

E	+1.G>A	P=0.930
F	3.C>A	P=0.049

AACCAC

G	3.C>T	P=0.011
H	4.C>A	P=0.049

3'UTR

TTAATGA

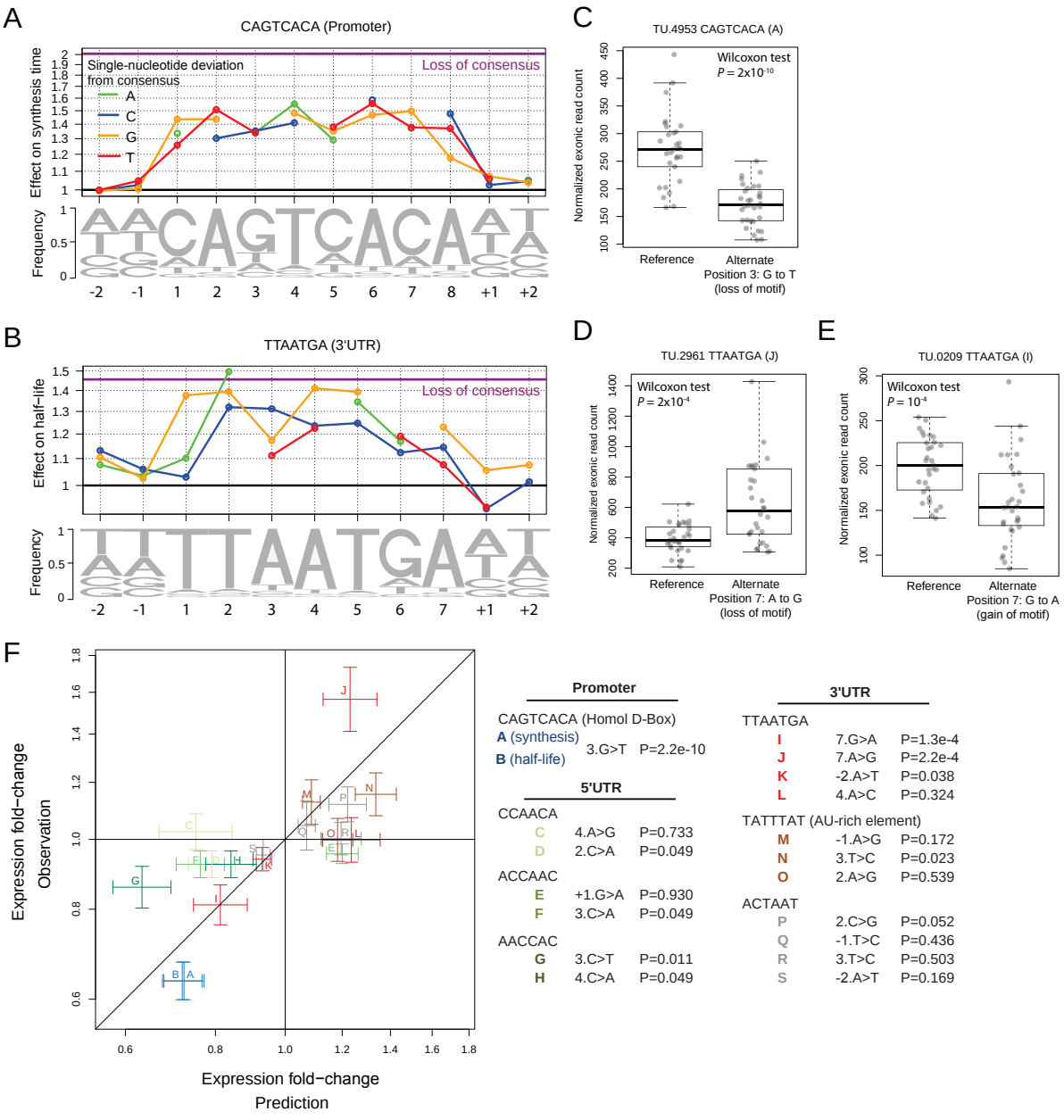
I	7.G>A	P=1.3e-4
J	7.A>G	P=2.2e-4
K	-2.A>T	P=0.038
L	4.A>C	P=0.324

TATTTAT (AU-rich element)

M	-1.A>G	P=0.172
N	3.T>C	P=0.023
O	2.A>G	P=0.539

ACTAAT

P	2.C>G	P=0.052
Q	-1.T>C	P=0.436
R	3.T>C	P=0.503
S	-2.A>T	P=0.169



Promoter

CAGTCACA (Homol D-Box)

A (synthesis)
B (half-life)

3.G>T	P=2.2e-10	

5'UTR

CCAACA

C	4.A>G	P=0.733
D	2.C>A	P=0.049

ACCAAC

E	+1.G>A	P=0.930
F	3.C>A	P=0.049

AACCAC

G	3.C>T	P=0.011
H	4.C>A	P=0.049

3'UTR

TTAATGA

I	7.G>A	P=1.3e-4
J	7.A>G	P=2.2e-4
K	-2.A>T	P=0.038
L	4.A>C	P=0.324

TATTTAT (AU-rich element)

M	-1.A>G	P=0.172
N	3.T>C	P=0.023
O	2.A>G	P=0.539

ACTAAT

P	2.C>G	P=0.052
Q	-1.T>C	P=0.436
R	3.T>C	P=0.503
S	-2.A>T	P=0.169