

Extremely fast and incredibly close: cotranscriptional splicing in budding yeast

EDWARD W.J. WALLACE^{1,2} and JEAN D. BEGGS²

¹School of Informatics, University of Edinburgh, EH8 9AB, United Kingdom

²Wellcome Trust Centre for Cell Biology, University of Edinburgh, EH9 3BF, United Kingdom

ABSTRACT

RNA splicing, an essential part of eukaryotic pre-messenger RNA processing, can be simultaneous with transcription by RNA polymerase II. Here, we compare and review independent next-generation sequencing methods that jointly quantify transcription and splicing in budding yeast. For many yeast transcripts, splicing is fast, taking place within seconds of intron transcription, while polymerase is within a few dozens of nucleotides of the 3' splice site. Ribosomal protein transcripts are spliced particularly fast and cotranscriptionally. However, some transcripts are spliced inefficiently or mainly post-transcriptionally. Intron-mediated regulation of some genes is likely to be cotranscriptional. We suggest that intermediates of the splicing reaction, missing from current data sets, may hold key information about splicing kinetics.

Keywords: CTD phosphorylation; introns; polymerase CTD; polymerase pausing; splicing kinetics

INTRODUCTION

RNA splicing is an essential process in the maturation of most transcripts produced by eukaryotic RNA polymerase II (Pol II). RNA molecules can be spliced while still being transcribed, as shown by pioneering electron micrography studies of *Drosophila* embryo transcripts (Beyer and Osheim 1988; Beyer et al. 1981). Since then, diverse experimental methods have provided extensive support for functional coupling between splicing and transcription (for review, see Alexander and Beggs 2010; Naftelberg et al. 2015; Alpert et al. 2016; Saldi et al. 2016). Evidently, transcription can affect splicing and vice versa, but how this is achieved and regulated is largely unknown. The speed of splicing varies from gene to gene, depending on the strength of splice sites as well as other factors, and expression of some genes is regulated through splicing. What are the transcriptome-wide patterns of splicing kinetics?

Recently, distinct next-generation sequencing approaches have measured the coupling of transcription to splicing in *Saccharomyces cerevisiae*: fast metabolic labeling with 4-thio-uracil (4tU-seq) (Barrass et al. 2015), nascent RNA-seq (Harlen et al. 2016), and single molecule intron tracking (SMIT) (Carrillo Oesterreich et al. 2016). These approaches

produce, for many genes, quantitative estimates of the speed of splicing, extent of cotranscriptional splicing, or polymerase position at splicing, respectively. Collectively, these data demonstrate that, in budding yeast, many introns are spliced out very soon after the intron is transcribed, a scenario that was controversial 10 years ago. Alternative next-generation sequencing methods measure polymerase position, without so far providing high-resolution measures of splicing: nascent elongating transcript sequencing (NET-seq) (Harlen et al. 2016), and crosslinking to modified Pol II and analysis of cDNA (mCRAC) (Milligan et al. 2016). Importantly, Harlen and colleagues and Milligan and colleagues also measure how the phosphorylation state of the carboxy-terminal domain (CTD) of a Pol II large subunit correlates with splicing factor recruitment or intron position genome-wide.

Here, we discuss the strengths and limitations of each approach, and the extent to which their results agree. One remarkable point of agreement is that ribosomal protein transcripts, the largest and most abundant class of spliced mRNAs, tend to be spliced faster and more cotranscriptionally.

METHODS OF MEASURING NASCENT RNA

4tU-seq (Barrass et al. 2015)

4tU-seq uses short-timescale metabolic labeling to measure the first minutes of RNA transcription and processing.

Abbreviations: 5'SS, 3'SS, 5' and 3' splice sites; BP, branch point; Pol II, RNA polymerase II; CTD, carboxy-terminal domain of Pol II; 4tU-seq, sequencing of 4-thio-Uracil labeled RNA; SMIT, single molecule intron tracking; nts, nucleotides; NET-seq, sequencing of the 3' ends of polymerase II-associated RNA; mCRAC, crosslinking to modified Pol II and analysis of cDNA; RP, ribosomal protein.

Corresponding author: Edward.Wallace@ed.ac.uk

Article is online at <http://www.rnajournal.org/cgi/doi/10.1261/rna.060830.117>. Freely available online through the RNA Open Access option.

© 2017 Wallace and Beggs This article, published in *RNA*, is available under a Creative Commons License (Attribution 4.0 International), as described at <http://creativecommons.org/licenses/by/4.0/>.

Nascent transcripts are labeled by incorporation of the uracil analog 4tU. After flash-freezing and cell lysis, the 4tU-labeled transcripts are biotinylated and then isolated based on their affinity for streptavidin beads, followed by randomly primed reverse transcription and sequencing. Using a statistical model to compare the spliced:unspliced ratio of transcripts labeled with 4tU for different times, the “relative speed of splicing” of each transcript is determined. The reported measure is the area under the curve (AUC) of the spliced:unspliced ratio, a time-weighted average of the fraction spliced within 5 min of synthesis. Note that the overall rate of conversion of pre-mRNA to spliced mRNA is determined in real time rather than relative to the movement of Pol II.

As 4tU-seq is essentially RNA-seq applied to the subset of RNA that is newly synthesized, a range of mature, validated tools are available to analyze the data, including building on the MISO algorithm for mRNA isoform quantification (Katz et al. 2010), to accommodate time-series labeling (DICE) (Huang and Sanguinetti 2016). These statistical tools provide robust gene-level quantification of splicing by combining strong information from the small proportion of reads that span a splice junction with individually weak information from the majority of nonjunction reads. However, 4tU-seq is uninformative about splicing intermediates (the products of the first catalytic step) or excised introns (the by-product of the second catalytic step): Indeed, at these short timescales the intron:exon ratio may detect variability in degradation rates of excised introns as well as variability in splicing rates (Box 1).

SMIT (Carrillo Oesterreich et al. 2016)

The SMIT approach cleverly exploits paired-end RNA sequencing to measure the splicing status of nascent transcripts that are assumed to be still associated with Pol II. Cells are harvested by centrifugation and washed in cold buffer, requiring the assumption that transcription and splicing are arrested to the same degree during sample preparation, then chromatin is isolated in a multistep procedure developed earlier by the same group (Carrillo Oesterreich et al. 2010), and mature poly(A)⁺ RNA is further depleted. Remaining chromatin-derived RNA is reverse-transcribed from a 3′ end-ligated oligonucleotide, then PCR amplified, using gene-specific primers upstream of target introns to ensure high read-depth for target transcripts. Most budding yeast intron-containing genes have short first exons, therefore 87 target genes were selected based on the suitability of first exons to prime PCR across the intron. To aid correct quantification despite dense sampling of target genes, random molecular barcodes were incorporated in the ligated oligonucleotide. By design, the 3′ end reads report polymerase position, while the 5′ end reads, because they start close to the introns, report splicing status. SMIT detects intron-containing transcripts prior to the first step of splicing and also detects spliced transcripts that have completed the second

step of splicing (exon joining, Box 1). SMIT cannot detect transcripts that have undergone only the first step of splicing, for which the 3′ end of nascent RNA is not contiguous with the 5′ primer site (Box 1). The fraction of transcripts that have undergone both steps is calculated at spatially binned Pol II positions for each tested gene, and the position is estimated at which 10%, 50%, and 90% of nascent transcripts are spliced comparably to “saturation,” i.e., fraction spliced for the most 3′ Pol II positions. Thus SMIT reports the “distribution of Pol II positions at splicing.” SMIT’s measurements of relative positions of polymerase and splicing are analogous to those made from the electron micrograph tracings of Beyer and Osheim (1988) and Beyer et al. (1981).

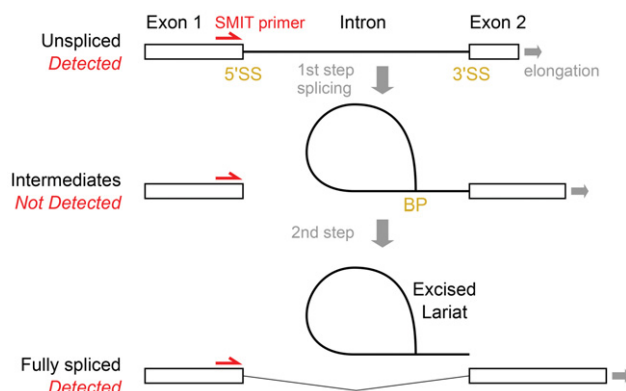
Importantly, Carrillo Oesterreich et al. (2010) validate the SMIT assay by comparison to other methods: Long-read sequencing confirms the onset of splicing soon after polymerase passes the 3′SS. Further, SMIT shows that splicing moves further 3′ from the 3′SS in a fast-elongating Pol II mutant, as predicted by kinetic competition of polymerization and splicing.

Nascent RNA-seq (Harlen et al. 2016)

For nascent RNA-seq, cells with affinity-tagged Pol II are flash-frozen and lysed, Pol II is immunoprecipitated, and the associated, nascent RNA transcripts are purified, fragmented, and sequenced. This approach allows estimation of the fraction of RNA spliced cotranscriptionally, but does not provide information about Pol II occupancy or splicing kinetics. Here we compare three estimates of the fraction spliced from the same data: from the fraction of intronic reads within 25 nt of the 3′SS (“3′SS int:ex”), from reads spanning the splice junction, or recalculated using all reads by DICE (Huang and Sanguinetti 2016). Only a small proportion of nascent RNA-seq reads span a splice junction or align within 25 nt of the 3′SS, thus for less abundant transcripts the fraction spliced cotranscriptionally is the ratio of two small numbers that are susceptible to count noise. More recent, noise-aware methods such as DICE circumvent this issue by pooling information from junction, intron, and exon reads.

NET-seq (Churchman and Weissman 2011; Harlen et al. 2016)

NET-seq quantifies Pol II occupancy genome-wide by sequencing the 3′ ends of all transcripts associated with Pol II (Churchman and Weissman 2011). Similar to nascent RNA-seq, RNA is extracted from immunoprecipitated Pol II, the crucial difference being that short reads (75 nt) are obtained from the 3′ ends of extracted RNA that has not been fragmented. As only a very small fraction of all nascent transcripts are spliced within these short distances, there is a very small number of spliced reads, so this approach is not suited to measuring splicing relative to the Pol II position.

BOX 1. Difficulties in detecting all stages of splicing

The pre-mRNA splicing reaction has two catalytic steps: In the first step, the 5'SS is cleaved and the lariat intron–exon intermediate species is formed by joining at the branch point (BP); and in the second, the 3' end of the upstream exon is joined to the downstream exon at the 3'SS, resulting in spliced mRNA and an excised lariat intron. Biased detection of these various RNA species complicates analysis in all methods discussed here.

4tU-seq quantifies splicing by the ratio of exonic to all reads for a given RNA, as well as incorporating the unique information from junction reads, implicitly assuming that excised introns are degraded quickly enough that most intronic reads are from unspliced pre-mRNA. Although excised introns are, in general, quickly degraded, some lariats might be degraded on a timescale comparable to that of 4tU-seq measurements. Indeed, some evolutionarily conserved noncoding RNAs, including snoRNAs, are processed from introns, and these may be degraded slowly (Hooks et al. 2016). Therefore, 4tU-seq may *underestimate* the rate of splicing for transcripts with stable intron-derived products.

SMIT quantifies splicing by sequencing PCR products that extend from a primer site upstream of the 5'SS to the 3' Pol II position. For splicing intermediates, these sites are not contiguous, and so are not detectable by SMIT. Therefore, SMIT may *overestimate* the extent of splicing of transcripts for which lariat intermediates represent a significant fraction of the total (slow second step), and overestimate the speed of splicing if spliceosome assembly and the first step of splicing occur before the 3'SS is transcribed (first step splicing is possible once the BP becomes available), such that distance from the 3'SS does not reflect all aspects of splicing.

Consider nascent transcripts whose 3' end is at position n . If the true count of pre-mRNA is P_n , splicing intermediates is I_n , and fully spliced mRNA is F_n , then the fraction that has completed splicing is $F_n/(P_n + I_n + F_n)$. However, SMIT estimates the fraction spliced as $F_n/(P_n + F_n)$, which is strictly greater; the difference depends on I_n .

This may explain a puzzle in the SMIT analysis, that some estimates of the position of onset of splicing conflict with structural models. The average position for onset (10% of transcripts spliced) of 26 nt after transcription of the 3' splice site is remarkably soon given that the distance between polymerase and spliceosome active sites is estimated as 24 nt (Carrillo Oesterreich et al. 2016). Perhaps this is not surprising: In a first-order kinetics approximation of splicing, the relative time or position of splicing would be an exponential distribution, with a mode at the first available point. However, beyond averages, SMIT's quantification algorithm reports that 24 of 87 measured genes have splicing onset <24 nt after the 3' SS, 9 are 50% spliced by then, and onset is not reported for a further 31 genes.

NET-seq detects Pol II-associated 3' ends of RNA species, including many reads exactly at the 3'SS which are most likely from excised introns attached to spliceosomes that are still associated with Pol II (Harlen et al. 2016). Interestingly, an earlier NET-seq data set observed an abundance of 3' end reads exactly at the 5'SS (Churchman and Weissman 2011), also consistent with a population of splicing intermediates that pull down with elongating Pol II.

Complementary methods quantify splicing intermediates by sequencing, even in wild-type cells (Gould et al. 2016; Qin et al. 2016). However, these rely on enrichment of lariats, so provide weak information on the kinetics of splicing. Methods that measure all steps of the cotranscriptional splicing reaction simultaneously are thus needed to fill in the gaps in our quantitative understanding of cotranscriptional splicing.

Detection of tRNA, which is produced by RNA Pol III, and accumulation of reads at the 3' ends of snoRNAs, suggests detectable levels of contamination from mature transcripts in NET-seq data. Similarly, 4tU-seq detects non-Pol II transcripts by design, and low levels of background RNA are isolated even without 4tU-labeling.

mCRAC (Milligan et al. 2016)

With CRAC (cross-linking and analysis of cDNA) (Granneman et al. 2009), short fragments of RNA are identi-

fied that are UV crosslinked to, and protected from nuclease digestion by, a tagged bait protein. The protocol is analogous to CLIP (Darnell 2010), but features a stringent denaturation step to reduce background. Modification CRAC (mCRAC) uses a two-step purification, first with tagged Pol II, then with antibodies specific to different phosphorylation states of Pol II's CTD, to obtain fragments of nascent transcripts associated with distinct states of the elongation machinery. Importantly, these data were used as input to a Bayesian model that suggests multiple distinct initiation and elongation states of Pol II. We address this data set only briefly here on

account of its relatively low depth of sequencing that does not permit an assessment of cotranscriptional splicing.

The data obtained by these approaches measure many aspects of transcription beyond those coupled to splicing: 4tU-seq, NET-seq, and mCRAC allow estimates of transcription rates as distinct from RNA abundance, including anti-sense and unstable transcripts, and a comparison would be instructive. However, we focus here on coupling of transcription and splicing.

RESULTS

Ribosomal protein genes tend to be spliced fast and cotranscriptionally

There is a great deal of agreement between results of these distinct methods (Fig. 1). As with most high-throughput data, between-study correlations are generally lower than within-study correlations. Using DICE to estimate the nascent RNA-seq fraction that is spliced generally correlates better with other studies; however, the nascent RNA-seq estimates that focus on the 3'SS reads correlate better with SMIT's saturation values, which also focus on the 3' end of the gene. SMIT's positions for near-complete (90%) splicing correlate better with alternative measurements than the positions of onset (10%) or median (50%) splicing.

Importantly, SMIT shows that the second step of splicing is complete on most transcripts when Pol II has moved less than 100 nt beyond the 3'SS, consistent with nascent RNA-seq's measure that the majority of splicing is cotran-

scriptional (Fig. 2A). Also, transcripts that are seen to be spliced fast by 4tU-seq are generally measured as spliced cotranscriptionally by SMIT (Fig. 2B) and nascent RNA-seq (Fig. 2C): in particular, ribosomal protein (RP) transcripts. Transcripts measured by SMIT to be spliced when Pol II is further from the 3'SS appear to be spliced more slowly by 4tU-seq analysis, including the well-characterized *ACT1*. Furthermore, *YRA1*, which is negatively autoregulated by an intron-dependent mechanism (Preker and Guthrie 2006; Dong et al. 2007), is spliced distally, slowly, and inefficiently (Fig. 2A,B). Transcripts that are mostly spliced at steady-state, but have low SMIT saturation values, are candidates for mainly post-transcriptional completion of splicing (Fig. 2D); none of these are RPs.

RP transcripts have other distinguishing patterns in these data: They have relatively fewer SMIT 3' end reads in the intron compared to exon 2, despite their longer introns (Fig. 2E; Supplemental Fig. S1). This suggests a large decrease in polymerase speed from intron to exon 2, consistent with NET-seq. Furthermore, RP genes have higher U1 occupancy as measured by CHIP-nexus (Harlen et al. 2016), so that the reported "high U1 occupancy genes" are essentially synonymous with the RP genes (Fig. 2F). Notably, the U1 occupancy differs only threefold between RP and non-RP genes, which is much less than the difference in mRNA abundance or, presumably, transcription rate. This is consistent with a low dynamic range or high background signal for U1 occupancy as measured by CHIP, but is also consistent with RP transcripts recruiting more U1 but for shorter times, due to their faster splicing.

These assays reveal differences in splicing between paralogous transcripts, for example, of the *RPS14A* and *RPS14B* genes. It has been shown that excess Rps14 protein can bind to a stem-loop structure in *RPS14B* pre-mRNA, inhibiting its splicing and leading to its rapid degradation (Fewell and Woolford 1999). *RPS14B* transcripts are spliced very slowly, inefficiently, and distally compared to other RPs (Fig. 2A–D); *RPS14A* was not measured by SMIT and behaves like most RPs in the other assays (Fig. 2C).

Some genes are spliced slowly, inefficiently, and/or post-transcriptionally

Curiously, some transcripts are apparently spliced slowly or less efficiently, yet close to the 3'SS: *SEC27*, *RFA2*, *NSP1*, *ARP2*, *DBP2*, *OM14*, *SAR1*, *RPL2A*, *RPS9A*, and *RPL30* (lower left quadrant in Fig. 2B). This apparent paradox could reflect the distinction between time and position of splicing: Pol II elongating more slowly or pausing near the 3'SS would allow slow splicing to occur cotranscriptionally. This could be an explanation for proximal splicing of *RPL30* and *RPL2A* transcripts, for which the SMIT saturation value is similar to the fraction spliced at steady state (Fig. 2D). Alternatively, the proximally spliced RNA may represent only a small fraction of the total, with most being spliced

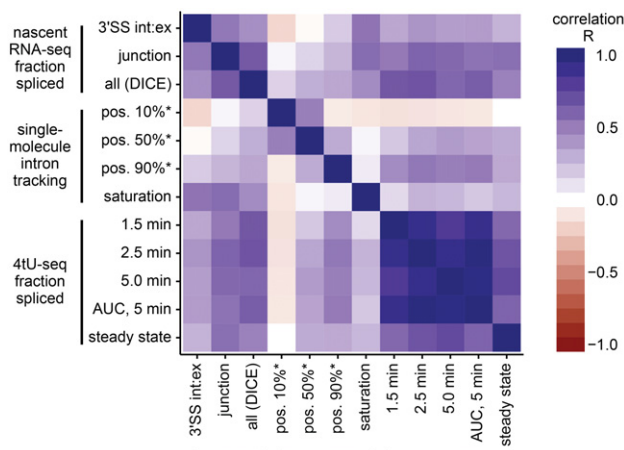


FIGURE 1. Estimates of cotranscriptional splicing, or splicing speed, mostly agree. Methods of measuring nascent RNA are described in the text; here we plot Pearson's correlation between genewise measurements. As expected, higher fraction cotranscriptionally spliced corresponds to smaller position of splicing, so for visual clarity, we reversed the sign of correlations for rows and columns containing position-based measurements (*).

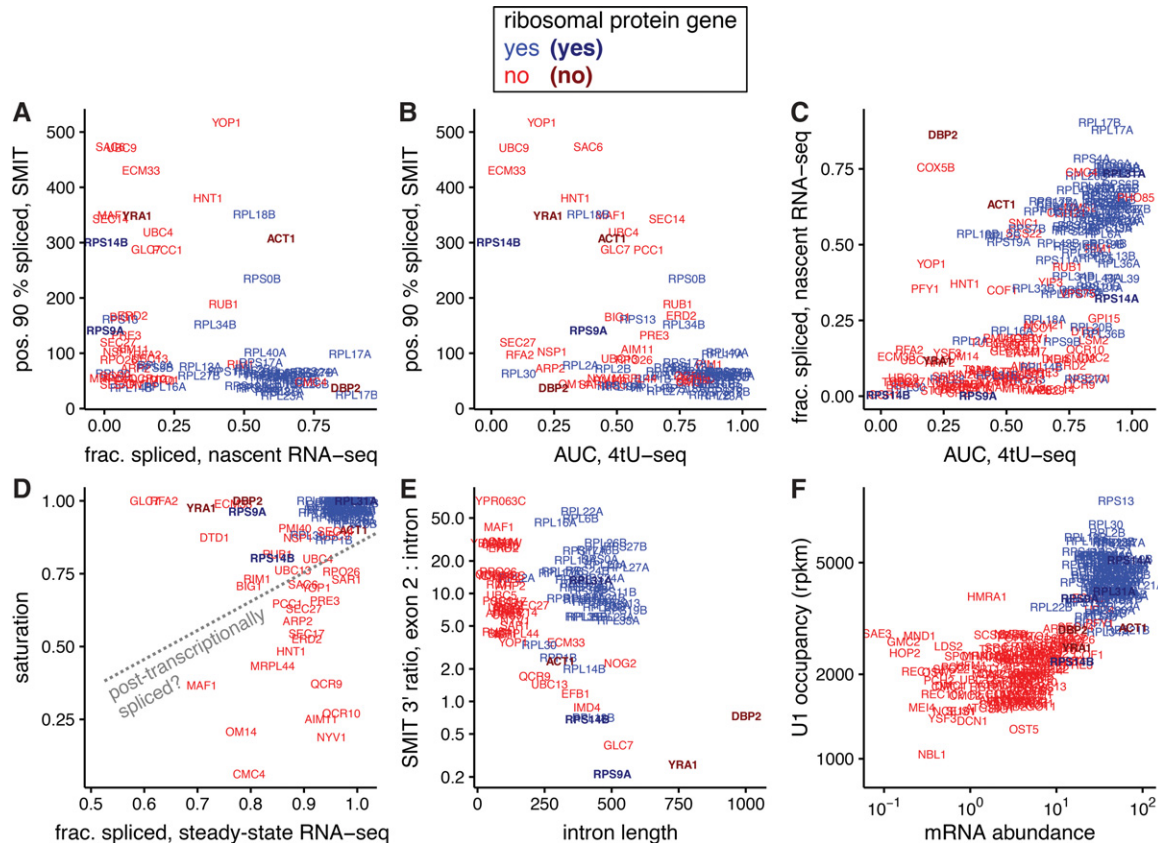


FIGURE 2. Intron-containing ribosomal protein transcripts (blue) tend to be spliced faster and more cotranscriptionally, compared to nonribosomal transcripts (red). Gene names are plotted with *center* indicating the position on each scale; select genes mentioned in text are shaded and bolded. (A) The distance in nucleotides of Pol II from the 3'SS when 90% of transcripts are spliced according to SMIT (pos. 90% spliced) (Carrillo Oesterreich et al. 2016) is plotted versus fraction spliced cotranscriptionally measured by nascent RNA-seq (Harlen et al. 2016). Note pos. 90% was not calculated for *RPL31A* in the SMIT analysis. (B) Pos. 90% spliced by SMIT versus AUC by 4tU-seq (a weighted average over the fraction spliced at 1.5, 2.5, and 5 min) (Barrass et al. 2015). (C) Fraction spliced according to nascent RNA-seq versus AUC by 4tU-seq. (D) Saturation (fraction spliced at most 3' Pol II positions) (Carrillo Oesterreich et al. 2016) plotted against fraction spliced in steady-state RNA-seq (Barrass et al. 2015). Transcripts of genes *below* the dotted line may be mostly post-transcriptionally spliced. (E) Ratio of SMIT 3' ends in introns and exon 2 (spliced or unspliced) (Carrillo Oesterreich et al. 2016) versus intron length. (F) U1 occupancy measured by ChIP-nexus (Harlen et al. 2016) versus mRNA abundance (Csárdi et al. 2015a).

post-transcriptionally. For example, *SEC27*, *ARP2*, and *OM14* are spliced inefficiently and proximally, but with low SMIT saturation values (Fig. 2D); in this respect the assays agree. The apparent discrepancy is caused by dividing the low fraction of proximally spliced reads by the still-low fraction of distally spliced reads that indicates mainly post-transcriptional splicing. This conclusion is supported by nascent RNA-seq also measuring a low fraction spliced for *SEC27*, *RFA2*, *NSP1*, *ARP2*, *OM14*, *SAR1*, *RPL2A*, *RPS9A*, and *RPL30* (lower left quadrant of Fig. 2A).

Artifacts in one or more of the assays could also explain the discrepancy: If an excised intron were degraded particularly slowly and sequenced efficiently, that would depress 4tU-seq estimates of splicing speed. Likewise, SMIT could overestimate the fraction of fully spliced transcripts near the 3'SS if splicing intermediates represent a large fraction of slowly spliced transcripts, as they are not detected in this assay (Box 1).

Intron-mediated inhibition may be cotranscriptional

The Rpl30 protein negatively regulates splicing of the *RPL30* transcript by binding cotranscriptionally to the intron (Macías et al. 2008). Concordantly, the *RPL30* intron is spliced very slowly, with only 24% spliced after 5 min as measured by 4tU-seq; however, SMIT reports that *RPL30* is 90% spliced when the polymerase has progressed only 70 nt beyond the 3'SS. As suggested above this could be caused by Pol II pausing, with Rpl30-intron binding causing Pol II to pause prior to 3'SS synthesis. The agreement of SMIT saturation with fraction spliced at steady state (Fig. 2D) suggests that, after the hypothesized pause is relieved and Pol II continues past the 3'SS, *RPL30* transcripts are spliced cotranscriptionally. Could other intron-regulated transcripts be regulated by similar cotranscriptional mechanisms?

The Dbp2 protein negatively regulates its own production by binding to its intron to inhibit splicing (Barta and Iggo

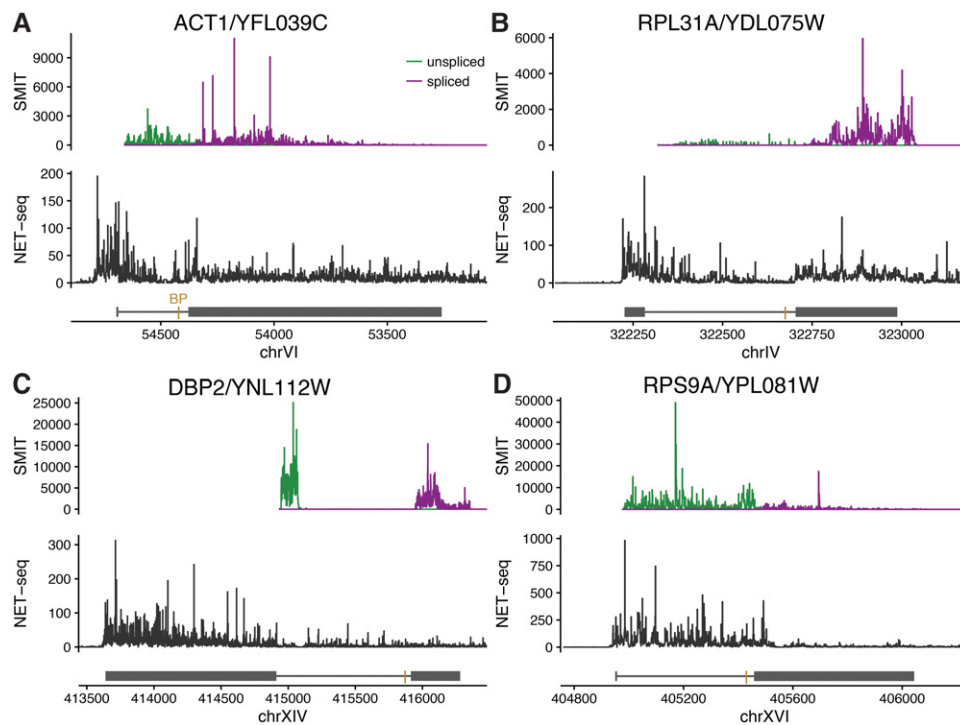


FIGURE 3. Comparison of SMIT and NET-seq profiles along individual genes, plotted in genomic co-ordinates. *Upper panels* show 3' end counts of unspliced (green) and spliced (purple) SMIT reads. *Lower panels* show 3' end counts of NET-seq reads. Line diagrams show exons represented by solid blocks, introns by lines, and the gold bar represents the branch point (BP); these selected genes are all on the plus strand (5' left, 3' right). Note different length and count scales for each gene. (A) *ACT1/YFL039C* is spliced with medium efficiency. (B) *RPL31A/YDL075W* is a fast-spliced ribosomal protein transcript: The low ratio of intron:exon 2 reads in SMIT, and the depletion of NET-seq reads toward the 3' end of the intron are typical of RP transcripts. *DBP2/YNL112W* (C) and *RPS9A/YPL081W* (D) have anomalous profiles.

1995); its transcript is slowly, yet apparently highly cotranscriptionally, spliced. To investigate at what stage in the transcript's lifetime this regulation might occur, we looked at the SMIT and NET-seq data in more detail (Fig. 3C). The unspliced SMIT reads decline abruptly ~160 nt into the intron, a position coinciding with a T-rich sequence that is presumably hard to align and also locally absent from NET-seq reads. Thus the apparently complete 3' SS-proximal splicing is due to the near-complete absence of unspliced reads near the 3' SS in the SMIT data. However, NET-seq reads continue along the intron and into exon 2, albeit at a low density compared to exon 1, consistent with incomplete passage of Pol II through the intron. Meanwhile, Dbp2 functions as a cotranscriptional RNA chaperone (Ma et al. 2013, 2016). We hypothesize that the intron-mediated negative autoregulation of *DBP2* splicing occurs cotranscriptionally and results in premature termination or cleavage of the nascent transcript.

Likewise, Rps9b protein binds to the *RPS9A* intron and 3' UTR in chromatin (suggesting that this happens cotranscriptionally) and represses splicing of the *RPS9A* intron (Petibon et al. 2016). Nevertheless, the *RPS9A* intron is highly cotranscriptionally spliced (according to SMIT) but with low efficiency (according to nascent RNA-seq). Again, both SMIT and NET-seq show a remarkably low density of reads in exon 2 compared to the intron (Fig. 3D). Petibon and col-

leagues further showed that repression of *RPS9A* splicing is promoter-independent, thus occurs after transcription initiation. We hypothesize that cotranscriptional repression of *RPS9A* splicing by Rps9b protein may result in Pol II pausing or premature termination.

These examples illustrate the rich information that can be obtained by comparing data sets from these different approaches, beyond a simple analysis of splicing efficiency, speed, or position. Indeed, comparing the SMIT profiles on individual genes highlights not only the distinction between RPs and other intron-containing genes, but also the 3' ends of the intronic snoRNAs embedded in *EFB1* and *IMD4* introns (Supplemental Fig. S1).

More generally, these vignettes emphasize that regulation of transcription is coupled to regulation of splicing.

Splicing affects transcription elongation and Pol II phosphorylation

Our group and others have observed that elongating Pol II can pause while the nascent pre-mRNA is being spliced. Pol II accumulates transiently, in a splicing-dependent manner, near the 3' splice site on reporter genes in budding yeast (Alexander et al. 2010), and certain splicing defects give rise to transcription defects at introns, suggesting a

transcriptional elongation checkpoint during cotranscriptional spliceosome assembly (Chathoth et al. 2014). Furthermore, Carrillo Oesterreich et al. (2010) reported an apparently distinct transcriptional pausing ~250 bp downstream from the 3'SS using tiling microarrays of a chromatin fraction. It remains unclear how widespread such pausing phenomena are: At which positions, in which genes, in which organisms, and in which conditions, does transcription wait for splicing, and for how long? Does pausing occur both before and after the first catalytic step of splicing?

Evidence from high-throughput data sets is so far mixed. Carrillo Oesterreich et al. (2016) sequenced 3' ends of chromatin-associated RNA in a single experiment, reporting no consistent Pol II accumulation close to the 3'SS or at positions of splicing onset or saturation, that might explain 3'SS proximal splicing. The first NET-seq data set (Churchman and Weissman 2011) did not disentangle RNA 3' ends that could be caused by Pol II pausing from an abundance of reads at the 3' ends of excised introns. On the other hand, Harlen et al. (2016), with higher read-depth, report an excess of NET-seq reads on yeast exons downstream from the 3'SS relative to upstream in the introns, and a peak immediately downstream from the 3'SS, that is interpreted as Pol II pausing. In agreement, Pol II was seen to accumulate after the 3'SS in a splicing-dependent manner in human cells with NET-seq (Nojima et al. 2015) and with chromatin-derived 3' end targeted RNA-seq (Mayer et al. 2015).

Splicing also leads to a change in state of the transcription elongation machinery, notably in the modifications of the Pol II CTD. CHIP-qPCR in yeast (Alexander et al. 2010; Chathoth et al. 2014) and CHIP-seq in humans (Nojima et al. 2015) showed a splicing-dependent phosphorylation of Ser5 in the CTD. Furthermore, CHIP-nexus (a modified, more precise, ChIP-seq) (Harlen et al. 2016) reported that Ser5 phosphorylation rises rapidly after the 3'SS in yeast, and Ser5 phosphorylated Pol II is enriched in 5'SS cleaved splicing intermediates in mammalian cells (Nojima et al. 2015). Applying a hidden Markov model to mCRAC data, Milligan et al. (2016) also detected systematic differences between overall modification states of Pol II on intron-containing versus intronless yeast transcripts, with changes of phosphorylation state coinciding with positions of splice sites.

Are these effects of splicing on Pol II functionally related? If specific modifications in Pol II are required for efficient transcription through the second exon, then conditions that alter these modifications could lead to transcriptional pausing. It has been suggested (Alexander et al. 2010; Chathoth et al. 2014) that Pol II pausing and phosphorylation could be evidence for splicing-dependent transcriptional checkpoints, implicating quality control of RNA processing in regulating transcription.

Clearly, more extensive and detailed analyses are needed to characterize RNA processing-dependent polymerase behavior genome-wide. If Pol II pausing happens between the first

and second catalytic steps of splicing, detecting the pause via sequencing requires splicing intermediates (specifically, intron–exon lariats) to be quantified simultaneously with the end products. If a subset of (modified) polymerase pauses transiently for splicing, then the signal is likely to be weak in the pool of total polymerase. If polymerase only pauses on a subset of genes, the signal within metagene analyses will be weak. If polymerase pausing is involved in a regulatory response or is modulated by environmental stimuli that affect RNA processing, pausing might be rare in some conditions, but common in others.

Which features of a transcript determine the cotranscriptionality or speed of splicing?

Intron-containing transcripts differ in the sequences that directly interact with the spliceosome (5'SS, 3'SS, BP) as well as their promoters, untranslated regions, introns, and coding regions, which can affect splicing by forming RNA secondary structure and/or recruiting regulatory proteins. The ideal sequence predictor would explain a substantial proportion of the genewise variance of all splicing measurements consistently, both for RP and non-RP transcripts. Scoring the 5'SS, 3'SS, or BP sequences against their consensus motifs falls far short of this ideal (Supplemental Fig. S2). In contrast, higher intron secondary structure (more negative ΔG per nucleotide) correlates with splicing being slower and less cotranscriptional (Supplemental Fig. S2), consistent with the multifactorial analysis in Barrass et al. (2015). However, differences between splicing metrics for RP and non-RP genes are more striking than the differences in these predictors.

Future directions

Collectively, the data sets examined here provide compelling evidence that RP transcripts are spliced both faster and more cotranscriptionally than the average transcript. RP genes produce the largest and most abundant class of spliced transcripts in yeast (Ares et al. 1999; Cherry et al. 2012), and RP gene expression and splicing are coherently regulated in response to a variety of environmental signals (Pleiss et al. 2007; Bergkessel et al. 2011). The majority of RP introns are required for growth in at least one condition (Parenteau et al. 2011), supporting a regulatory role. RP transcripts are also unusually stable in fast-growth conditions and unusually unstable after glucose starvation (Munchel et al. 2011). Which features of RP genes promote their efficient and cotranscriptional processing? Is it their unusually long (for yeast) introns or their near-consensus GUAUG 5'SS? Is it their stereotypical transcription factors or promoter architecture (Knight et al. 2014), or distinctive complement of interacting RNA-binding proteins (Hogan et al. 2008)? Could it be connected to reciprocal regulation of ribosomal protein and ribosomal RNA production (Lempiäinen and Shore 2009)? Are there distinct

spliceosomal factors that favor splicing of RP transcripts? Unraveling the distinctiveness of RP compared to non-RP gene expression is an important ongoing research focus.

The results discussed here were produced using innovative approaches; how robust will the conclusions seem after further methodological development? Protocols, especially those associated with next-generation sequencing, may need extended optimization to distinguish biological sequence-specific signals from noise and artifacts that arise in both sample preparation and data analysis. Ribosome profiling presents an instructive comparison: The initial analysis for yeast reported that ribosomes do not translate rare codons slowly (Ingolia et al. 2009), which was counterintuitive to conclusions from evolutionary biology and biochemistry (Hershberg and Petrov 2008). Active debate continued over several years as many groups adjusted experimental and analytical protocols, particularly regarding the use of ribosome-stalling drugs. Now, there is clear evidence that ribosome profiling does quantify slower translation at rare codons (Hussmann et al. 2015; Weinberg et al. 2016), in agreement with complementary next-generation sequencing approaches (Pelechano et al. 2016).

Nanopore sequencing of nascent RNA would circumvent the length restrictions of short-read Illumina sequencing; direct RNA sequencing promises to circumvent biases caused by reverse-transcription to cDNA (Garalde et al. 2016). Enrichment of spliced transcripts, analogous to SMIT, might be achievable by real-time selective sequencing on a nanopore sequencer (Loose et al. 2016).

The work discussed here has quantified splicing and transcription in the model yeast, *S. cerevisiae*, with unprecedented detail and scope: Multiple experimental approaches applied to the same system yield richer insights than any one viewed alone. Do these observations extend to other eukaryotes? For *Schizosaccharomyces pombe*, fast, cotranscriptional splicing was reported by long-read sequencing (Carrillo Oesterreich et al. 2016) and suggested by 4tU-seq (Eser et al. 2016). NET-seq in mammals supports cotranscriptional splicing, observing Pol II peaks (interpreted as transcriptional pauses) 3' to a subset of introns (Nojima et al. 2015) and within spliced exons (Mayer et al. 2015). However, very different efficiencies of cotranscriptional splicing have been reported for *Drosophila* and mouse (Khodor et al. 2012). A variety of gene features, including intron, exon and overall gene length, intron position, splice sites and other sequence elements, RNA structure and synthesis rate, contribute to differences in splicing kinetics (Khodor et al. 2012; Barrass et al. 2015; Eser et al. 2016). Explaining such differences is not simple and, clearly, much remains to be learned.

DATA DEPOSITION

Data were taken from cited publications, where possible with minimal processing, to reflect both the experimental and analysis

pipelines used. Nascent RNA-seq data by gene (used in Fig. 1) were a personal communication from K. Harlen, corresponding to Figure 6D of Harlen et al. (2016), and we recomputed the fraction spliced using DICE (Huang and Sanguinetti 2016), after aligning raw nascent RNA-seq reads from GEO (GSE68484; SRR2046809) to the *S. cerevisiae* genome using STAR (Dobin et al. 2013). SMIT data are from Carrillo Oesterreich et al. (2016), Supplemental Table S1, and 4tU-seq data from Barrass et al. (2015), Supplemental Table S7. SMIT intron:exon ratios in Figure 2E were calculated by aligning raw reads from GEO (GSE70908; pooled from smit_20genes and smit_extended data sets, excluding short RNA data) to the *S. cerevisiae* genome using STAR (Dobin et al. 2013); SMIT profiles in Figure 3 and Supplemental Figure S1 were from the same alignments. Intron lengths in Figure 2E were taken from Supplemental Table S8 of Barrass et al. (2015). In Figure 2F, mRNA abundances were taken from Csárdi et al. (2015b) (scer-mrna-protein-absolute-estimate.txt from data package) and U1 occupancy from Supplemental Table S4 of Harlen et al. (2016). NET-seq profiles in Figure 3 are from the bedgraph files (WT_NETseq) in GEO (GSE68484). Data were processed using the statistical language R (R core team) and plotted with ggplot2 (Wickham 2009). R scripts to create the figures are available upon request.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

We thank members of the Beggs laboratory, Guido Sanguinetti, Stirling Churchman, and Karla Neugebauer, for discussions and comments on the manuscript. We further thank Yuanhua Huang for help with using DICE. We thank two reviewers for their constructive comments. This work was funded by the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement no. 661179 to E.W., Wellcome Trust awards to J.B. (104648), and the Wellcome Trust Centre for Cell Biology core grant (092076).

NOTE ADDED IN PROOF

As this paper was in press, we were made aware that Schaughency et al. (2014) observed a transcription termination site within the DBP2 intron.

REFERENCES

- Alexander R, Beggs JD. 2010. Cross-talk in transcription, splicing and chromatin: who makes the first call? *Biochem Soc Trans* **38**: 1251–1256.
- Alexander RD, Innocente SA, Barrass JD, Beggs JD. 2010. Splicing-dependent RNA polymerase pausing in yeast. *Mol Cell* **40**: 582–593.
- Alpert T, Herzl L, Neugebauer KM. 2016. Perfect timing: splicing and transcription rates in living cells. *Wiley Interdiscip Rev RNA*. doi: 10.1002/wrna.1401.
- Ares M Jr, Grate L, Pauling MH. 1999. A handful of intron-containing genes produces the lion's share of yeast mRNA. *RNA* **5**: 1138–1139.
- Barrass JD, Reid JEA, Huang Y, Hector RD, Sanguinetti G, Beggs JD, Granneman S. 2015. Transcriptome-wide RNA processing kinetics revealed using extremely short 4tU labeling. *Genome Biol* **16**: 282.

- Barta I, Iggo R. 1995. Autoregulation of expression of the yeast Dbp2p "DEAD-box" protein is mediated by sequences in the conserved DBP2 intron. *EMBO J* **14**: 3800–3808.
- Bergkessel M, Whitworth GB, Guthrie C. 2011. Diverse environmental stresses elicit distinct responses at the level of pre-mRNA processing in yeast. *RNA* **17**: 1461–1478.
- Beyer AL, Osheim YN. 1988. Splice site selection, rate of splicing, and alternative splicing on nascent transcripts. *Genes Dev* **2**: 754–765.
- Beyer AL, Bouton AH, Miller OL. 1981. Correlation of hnRNP structure and nascent transcript cleavage. *Cell* **26**: 155–165.
- Carrillo Oesterreich F, Preibisch S, Neugebauer KM. 2010. Global analysis of nascent RNA reveals transcriptional pausing in terminal exons. *Mol Cell* **40**: 571–581.
- Carrillo Oesterreich F, Herzel L, Straube K, Hujer K, Howard J, Neugebauer KM. 2016. Splicing of nascent RNA coincides with intron exit from RNA polymerase II. *Cell* **165**: 372–381.
- Chathoth KT, Barrass JD, Webb S, Beggs JD. 2014. A splicing-dependent transcriptional checkpoint associated with prespliceosome formation. *Mol Cell* **53**: 779–790.
- Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, et al. 2012. *Saccharomyces* Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res* **40**: D700–D705.
- Churchman LS, Weissman JS. 2011. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* **469**: 368–373.
- Csárdi G, Franks A, Choi DS, Airoidi EM, Drummond DA. 2015a. Accounting for experimental noise reveals that mRNA levels, amplified by post-transcriptional processes, largely determine steady-state protein levels in yeast. *PLoS Genet* **11**: e1005206.
- Csárdi G, Franks A, Choi DS, Airoidi EM, Drummond DA. 2015b. Data from: Accounting for experimental noise reveals that mRNA levels, amplified by post-transcriptional processes, largely determine steady-state protein levels in yeast. *Dryad Digital Repository*. doi: 10.5061/dryad.d644f.
- Darnell RB. 2010. HITS-CLIP: panoramic views of protein-RNA regulation in living cells. *Wiley Interdiscip Rev RNA* **1**: 266–286.
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21.
- Dong S, Li C, Zenklusen D, Singer RH, Jacobson A, He F. 2007. YRA1 autoregulation requires nuclear export and cytoplasmic Edc3p-mediated degradation of its pre-mRNA. *Mol Cell* **25**: 559–573.
- Eser P, Wachutka L, Maier KC, Demel C, Boroni M, Iyer S, Cramer P, Gagneur J. 2016. Determinants of RNA metabolism in the *Schizosaccharomyces pombe* genome. *Mol Syst Biol* **12**: 857.
- Fewell SW, Woolford JL Jr. 1999. Ribosomal protein S14 of *Saccharomyces cerevisiae* regulates its expression by binding to RPS14B pre-mRNA and to 18S rRNA. *Mol Cell Biol* **19**: 826–834.
- Garalde DR, Snell EA, Jachimowicz D, Heron AJ. 2016. Highly parallel direct RNA sequencing on an array of nanopores. *bioRxiv* doi: 10.1101/068809.
- Gould GM, Paggi JM, Guo Y, Phizicky DV, Zinshteyn B, Wang ET, Gilbert WV, Gifford DK, Burge CB. 2016. Identification of new branch points and unconventional introns in *Saccharomyces cerevisiae*. *RNA* **22**: 1522–1534.
- Granneman S, Kudla G, Petfalski E, Tollervey D. 2009. Identification of protein binding sites on U3 snoRNA and pre-rRNA by UV cross-linking and high-throughput analysis of cDNAs. *Proc Natl Acad Sci* **106**: 9613–9618.
- Harlen KM, Trotta KL, Smith EE, Mosaheb MM, Fuchs SM, Churchman LS. 2016. Comprehensive RNA polymerase II interactomes reveal distinct and varied roles for each phospho-CTD residue. *Cell Rep* **15**: 2147–2158.
- Hershberg R, Petrov DA. 2008. Selection on codon bias. *Annu Rev Genet* **42**: 287–299.
- Hogan DJ, Riordan DP, Gerber AP, Herschlag D, Brown PO. 2008. Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. *PLoS Biol* **6**: e255.
- Hooks KB, Naseeb S, Parker S, Griffiths-Jones S, Delneri D. 2016. Novel intronic RNA structures contribute to maintenance of phenotype in *Saccharomyces cerevisiae*. *Genetics* **203**: 1469–1481.
- Huang Y, Sanguinetti G. 2016. Statistical modeling of isoform splicing dynamics from RNA-seq time series data. *Bioinformatics* **32**: 2965–2972.
- Hussmann JA, Patchett S, Johnson A, Sawyer S, Press WH. 2015. Understanding biases in ribosome profiling experiments reveals signatures of translation dynamics in yeast. *PLoS Genet* **11**: e1005732.
- Ingolia NT, Ghaemmghami S, Newman JRS, Weissman JS. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**: 218–223.
- Katz Y, Wang ET, Airoidi EM, Burge CB. 2010. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods* **7**: 1009–1015.
- Khodor YL, Menet JS, Tolan M, Rosbash M. 2012. Cotranscriptional splicing efficiency differs dramatically between *Drosophila* and mouse. *RNA* **18**: 2174–2186.
- Knight B, Kubik S, Ghosh B, Bruzzone MJ, Geertz M, Martin V, Déneraud N, Jacquet P, Ozkan B, Rougemont J, et al. 2014. Two distinct promoter architectures centered on dynamic nucleosomes control ribosomal protein gene transcription. *Genes Dev* **28**: 1695–1709.
- Lempiäinen H, Shore D. 2009. Growth control and ribosome biogenesis. *Curr Opin Cell Biol* **21**: 855–863.
- Loose M, Malla S, Stout M. 2016. Real-time selective sequencing using nanopore technology. *Nat Methods* **13**: 751–754.
- Ma WK, Cloutier SC, Tran EJ. 2013. The DEAD-box protein Dbp2 functions with the RNA-binding protein Yra1 to promote mRNP assembly. *J Mol Biol* **425**: 3824–3838.
- Ma WK, Paudel BP, Xing Z, Sabath IG, Rueda D, Tran EJ. 2016. Recruitment, duplex unwinding and protein-mediated inhibition of the DEAD-box RNA helicase Dbp2 at actively transcribed chromatin. *J Mol Biol* **428**: 1091–1106.
- Macías S, Bragulat M, Tardiff DF, Vilardell J. 2008. L30 binds the nascent RPL30 transcript to repress U2 snRNP recruitment. *Mol Cell* **30**: 732–742.
- Mayer A, di Iulio J, Maleri S, Eser U, Vierstra J, Reynolds A, Sandstrom R, Stamatoyannopoulos JA, Churchman LS. 2015. Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. *Cell* **161**: 541–554.
- Milligan L, Huynh-Thu VA, Delan-Forino C, Tuck A, Petfalski E, Lombaña R, Sanguinetti G, Kudla G, Tollervey D. 2016. Strand-specific, high-resolution mapping of modified RNA polymerase II. *Mol Syst Biol* **12**: 874.
- Munchel SE, Shultzaberger RK, Takizawa N, Weis K. 2011. Dynamic profiling of mRNA turnover reveals gene-specific and system-wide regulation of mRNA decay. *Mol Biol Cell* **22**: 2787–2795.
- Naftelberg S, Schor IE, Ast G, Kornbliht AR. 2015. Regulation of alternative splicing through coupling with transcription and chromatin structure. *Annu Rev Biochem* **84**: 165–198.
- Nojima T, Gomes T, Grosso ARF, Kimura H, Dye MJ, Dhir S, Carmo-Fonseca M, Proudfoot NJ. 2015. Mammalian NET-Seq reveals genome-wide nascent transcription coupled to RNA processing. *Cell* **161**: 526–540.
- Parenteau J, Durand M, Morin G, Gagnon J, Lucier J-F, Wellinger RJ, Chabot B, Abou Elela S. 2011. Introns within ribosomal protein genes regulate the production and function of yeast ribosomes. *Cell* **147**: 320–331.
- Pelechano V, Wei W, Steinmetz LM. 2016. Genome-wide quantification of 5'-phosphorylated mRNA degradation intermediates for analysis of ribosome dynamics. *Nat Protoc* **11**: 359–376.
- Petibon C, Parenteau J, Catala M, Elela SA. 2016. Introns regulate the production of ribosomal proteins by modulating splicing of duplicated ribosomal protein genes. *Nucleic Acids Res* **44**: 3878–3891.
- Pléiss JA, Whitworth GB, Bergkessel M, Guthrie C. 2007. Rapid, transcript-specific changes in splicing in response to environmental stress. *Mol Cell* **27**: 928–937.

- Preker PJ, Guthrie C. 2006. Autoregulation of the mRNA export factor Yra1p requires inefficient splicing of its pre-mRNA. *RNA* **12**: 994–1006.
- Qin D, Huang L, Wlodaver A, Andrade J, Staley JP. 2016. Sequencing of lariat termini in *S. cerevisiae* reveals 5' splice sites, branch points, and novel splicing events. *RNA* **22**: 237–253.
- Saldi T, Cortazar MA, Sheridan RM, Bentley DL. 2016. Coupling of RNA polymerase II transcription elongation with pre-mRNA splicing. *J Mol Biol* **428**: 2623–2635.
- Schaughency P, Merran J, Corden JL. 2014. Genome-wide mapping of yeast RNA polymerase II termination. *PLoS Genet* **10**: e1004632.
- Weinberg DE, Shah P, Eichhorn SW, Hussmann JA, Plotkin JB, Bartel DP. 2016. Improved ribosome-footprint and mRNA measurements provide insights into dynamics and regulation of yeast translation. *Cell Rep* **14**: 1787–1799.
- Wickham H. 2009. *ggplot2: elegant graphics for data analysis*. Springer-Verlag, New York.