



Native RNA sequencing in fission yeast reveals frequent alternative splicing isoforms

José Carlos Montañés, Marta Huertas, Simone G Moro, et al.

Genome Res. published online May 26, 2022

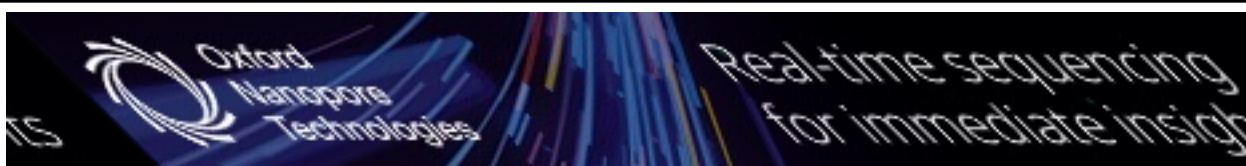
Access the most recent version at doi:[10.1101/gr.276516.121](https://doi.org/10.1101/gr.276516.121)

P<P Published online May 26, 2022 in advance of the print journal.

Accepted Manuscript Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.

Creative Commons License This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see <https://genome.cshlp.org/site/misc/terms.xhtml>). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).



To subscribe to *Genome Research* go to:
<https://genome.cshlp.org/subscriptions>

Published by Cold Spring Harbor Laboratory Press

Native RNA sequencing in fission yeast reveals frequent alternative splicing isoforms

José Carlos Montañés¹, Marta Huertas¹, Simone G. Moro¹, William R. Blevins^{1,#}, Mercè Carmona², José Ayté², Elena Hidalgo², M. Mar Albà^{1,3,*}.

¹Evolutionary Genomics Group, Research Programme on Biomedical Informatics, Hospital del Mar Medical Research Institute (IMIM) and Universitat Pompeu Fabra (UPF), Barcelona, Spain.

²Oxidative Stress and Cell Cycle Group, Universitat Pompeu Fabra (UPF), Barcelona, Spain.

³Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain

[#]Current address: CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona, Spain

^{*}Corresponding author: malba@imim.es

Running title: dRNA in fission yeast

Abstract

The unicellular yeast *Schizosaccharomyces pombe* (fission yeast) retains many of the splicing features observed in humans and is thus an excellent model to study the basic mechanisms of splicing. Nearly half the genes contain introns, but the impact of alternative splicing in gene regulation and proteome diversification remains largely unexplored. Here we leverage Oxford Nanopore Technologies native RNA sequencing (dRNA), as well as ribosome profiling data, to uncover the full range of polyadenylated transcripts and translated open reading frames. We identify 332 alternative isoforms affecting the coding sequences of 262 different genes, 97 of which occur at frequencies higher than 20%, indicating that functional alternative splicing in *S. pombe* is more prevalent than previously suspected. Intron retention events make about 80% of the cases; these events may be involved in the regulation of gene expression and, in some cases, generate novel protein isoforms, as supported by ribosome profiling data in 18 of the intron retention isoforms. One example is the *rpl22* gene, in which intron retention is associated with the translation of a protein of only 13 amino acids. We also find that lowly expressed transcripts tend to have longer poly(A) tails than highly expressed transcripts, highlighting an interdependence between poly(A) tail length and transcript expression level. Finally, we discover 214 novel transcripts that are not annotated, including 158 antisense transcripts, some of which also show translation evidence. The methodologies described in this work open new opportunities to study the regulation of splicing in a simple eukaryotic model.

Introduction

The unicellular eukaryote *S. pombe*, with around 7,000 genes, is an ideal model to study cellular processes that are conserved across eukaryotes (Wood et al. 2002; Kim et al. 2010). About 43% of the genes contain introns, often multiple ones. Thus, in contrast to other unicellular yeast species such as *S. cerevisiae*, which has a very limited number of introns, *S. pombe* can also be used to study the molecular basis of splicing. Previous studies using intron lariat sequencing, short read RNA-seq and Iso-seq have uncovered many low frequency alternative isoforms (Stepankiw et al. 2015; Bitton et al. 2015a; Kuang et al. 2017), suggesting that splicing fidelity in the species is relatively low.

Little is known about the impact of alternative splicing in generating functional isoforms and expanding the proteome of *S. pombe*. One of the few well-studied cases is *rem1*, encoding a cyclin required for meiosis. The expression of the Rem1 protein is regulated at the level of splicing, the retention of an intron ensures that no protein is produced before the start of meiosis (Malapeira et al. 2005; Moldón et al. 2008). At the same time, the intron retention isoform results in a 17 kDa protein with a role in recombination in the pre-meiotic S phase. Other possible examples of functional alternative splicing events are three exon skipping transcripts that have been reported to be conserved between *S. pombe* and humans (Awan et al. 2013). A complete catalog of alternative splicing isoforms occurring at high frequencies, together with the putative encoded proteins, is still missing.

Here we use native RNA sequencing (Garalde et al. 2018; Workman et al. 2019), in combination with ribosome profiling (Ingolia et al. 2009; Brar and Weissman 2015), to uncover the complete transcriptome and translome of *S. pombe*. Oxford Nanopore Technologies (ONT) direct RNA (dRNA) sequencing offers several important advantages over previous RNA-seq approaches: 1. It provides an unbiased snapshot of the native polyadenylated RNAs in the cell; 2. There is no need to assemble the transcripts using reads that are much shorter than the RNA molecule; 3. It is highly

quantitative, as each sequence corresponds to a single RNA molecule and; 4. It is very sensitive, millions of reads can be generated per experiment. dRNA sequencing has been successfully employed to discover new gene isoforms in *Homo sapiens* (Workman et al. 2019), *Arabidopsis thaliana* (Zhang et al. 2020) and *Caenorhabditis* (Li et al. 2020; Roach et al. 2020). Additionally, unlike Nanopore cDNA sequencing, dRNA provides information on the orientation of the transcript, which is essential to be able to detect new antisense transcripts. As we have recently shown in *S. cerevisiae*, antisense transcripts can originate rapidly during evolution, providing new functionalities (Blevins et al. 2021). No dRNA sequencing of *S. pombe* has yet been produced, limiting our knowledge on the complexity of the transcriptome of this model eukaryotic species.

Nanopore dRNA sequencing starts from the 3' end of the molecule, capturing the full-length poly(A) tail of each RNA. This enables the investigation of poly(A) tail variation among different transcripts and individual mRNA molecules (Workman et al. 2019). Poly(A) tail length is the result of polyadenylation and de-adenylation processes, and has been related to transcript stability and translatability (Dreyfus and Régnier 2002). Poly(A) tail shortening can initiate mRNA degradation in the cytoplasm (Parker and Song 2004). In humans, it has been shown that poly(A) polymerase activity can result in the decay of nuclear non-coding RNAs or mRNAs with retained introns (Bresson et al. 2015). By using dRNA it is possible to study both alternative splicing and alterations in the poly(A) length, obtaining new clues about the possible regulatory functions of poly(A) length.

Alternative splicing isoforms can encode proteins that are different from the canonical ones. These proteins remain poorly annotated because they are frequently short and partially overlap the annotated protein. A high throughput method to test for translation activity in putative open reading frames is ribosome profiling (Ribo-seq) (Ingolia et al. 2009). This technique is based on the sequencing of ribosome protected RNA fragments and has single nucleotide resolution. The three nucleotide periodicity of the reads has been used to discover novel translated ORFs in long non-coding RNAs and 5' untranslated regions (5'UTRs) (Ingolia et al. 2009; Duncan and Mata 2014; Ji et al.

2015; Ruiz-Orera and Albà 2019b) as well as in alternative transcript isoforms (Reixachs-Solé et al. 2019). Here we use Ribo-seq data to investigate the hallmarks of translation of alternative protein isoforms as well as to identify translated ncRNAs and novel transcripts. Our aim is to exploit Nanopore dRNA data in conjunction with Ribo-seq to uncover parts of the transcriptome and translome that might have remained hidden due to previous technical limitations.

Results

Native sequencing of polyA⁺ RNAs in *S. pombe*

We extracted total RNA from *S. pombe* cells growing at log-phase and subsequently performed polyA⁺ selection. Then we carried out direct RNA (dRNA) sequencing of the polyadenylated RNA using an ONT Gridion instrument (Garalde et al. 2018). We obtained a total of 7,297,642 dRNA sequencing reads from four sequencing runs. Each of these reads corresponds to a single native polyA⁺ RNA. The average read length was around 650 nucleotides (see Supplemental Table S1 for more details).

Nanopore reads are remarkably long compared to other short read sequencing technologies and they contain more errors, which need to be corrected (Amarasinghe et al. 2020). One commonly used approach to try to decrease the proportion of errors is to select reads that pass a certain quality score (typically $Q \geq 7$). However, we found that eliminating reads with $Q < 7$ had nearly no effect on the error rate (Figure 1A), and thus we did not apply this filter. Instead, we performed a correction based on Illumina reads with the program fmlrc (Wang et al. 2018), taking advantage of a previous *S. pombe* Illumina RNA-seq experiment performed in the same growth conditions as here (Blevins et al. 2019). The error rate decreased to about half its original values but some regions, such as the 3' end of transcripts, still remained largely uncorrected. For this reason we subsequently applied Transcript Clean, which uses the genome sequence as reference to correct ONT reads (Wyman and Mortazavi 2019). The final “clean” set had an average error rate of only 1.24%, which basically corresponded to short indels.

The reads were mapped to the PomBase gene annotations with minimap2 (Li 2018). The total number of mapped reads was 5,054,233. The longest mapped read was 13,899 nucleotides long. The mapped reads had an average length of 756 nucleotides and were significantly longer than the raw reads (Figure 1B). We could see expression of the vast majority of the protein-coding transcripts (97.8%), as well as of a very large percentage of non-coding RNAs or ncRNAs (87.8%) and smaller amounts of other RNA classes (Supplemental Fig S1). In general, ncRNAs were expressed at much lower levels than mRNAs (average 70 dRNA reads *versus* 1,130 dRNA reads).

We inspected the correlation between dRNA and Illumina derived transcript abundances. Whereas each dRNA read corresponds to one native molecule, Illumina sequencing involves cDNA synthesis and PCR amplification, and the number of mapped reads needs to be normalized by length. In addition, we found that 13.41% of the Illumina reads were multimapping, increasing the uncertainty in the transcript abundance estimates. There was a high positive correlation between the abundance estimates obtained with the two technologies (Spearman's $\rho = 0.849$, $p < 10^{-12}$, Figure 1C), after excluding transcripts with multimapping reads. Inclusion of the 1,800 transcripts with Illumina multimapping reads caused an overestimation of transcript expression levels for some of the transcripts (Supplemental Fig S2).

Nanopore mRNA sequencing starts from the 3' end of the transcript and proceeds towards the 5' end. Some of the mRNAs are sequenced to completion (full-length reads), whereas others are truncated at their 5' end. We estimated the number of full-length reads by mapping the reads to the gene annotations, and then comparing the length of each mapped read with the length of the corresponding annotated transcript. To be considered full length, the read had to be equal or longer than the annotated transcript or, in case it was shorter, the difference should be less than 50 nucleotides. This accounted for the fact that the first 10-15 nucleotides of the 5'UTR are systematically missed with Nanopore, and that the real 3' end might also show some variation with respect to the annotated transcript. We estimated that the total number of full-length reads was 1,013,789 (20.06%). Perhaps more importantly, the

vast majority of the transcripts with expression evidence had at least one full-length read (5,165 out of 6,453, 80.04%). As expected, the fraction of transcripts recovered as full-length reads decreased with transcript length, with the strongest effect being observed in transcripts longer than 3.45 kb (last decile in Figure 1D, Supplemental Fig S3). We observed that this subset of very long transcripts also tended to be expressed at lower levels than transcripts of intermediate length (Figure 1E, Supplemental Fig S4). Transcripts in the first decile (< 633 nucleotides) were expressed at even lower values but because of their short size, they were normally recovered as full length reads.

Poly(A) tail length depends on expression level but not transcript length

Extended poly(A) tail lengths have been previously associated with increased transcript's stability and translatability (Dreyfus and Régnier 2002). We used nanopolish to measure poly(A) tail length directly from the dRNA data. The average poly(A) length was around 50 nucleotides, similar to that observed in humans (Workman et al. 2019). Poly(A) tails tended to be slightly shorter in mRNAs (median = 48.9 nt) than ncRNAs (median = 51 nt)(Figure 1F). We found that poly(A) length and transcript abundance were negatively correlated (Fig. 1G, Spearman's rho = -0.376 and p-value < 10^{-5}). The median number of counts for the top 10% transcripts with the shortest poly(A) tail (length < 40) was 581.5, whereas the 10% of genes with the longest poly(A) tail (length > 58) had a median of 131 counts. Gene Ontology term enrichment analysis indicated that genes with the shortest poly(A) tail were significantly enriched in translation-related functions, whereas those with the longest poly(A) tail in sexual reproduction and meiosis-related functions (Supplemental Fig S5). Consistently, these two groups also showed clear differences in their expression levels, with the translation genes being expressed at very high levels and the meiosis genes at much lower levels (Figure 1G). No major differences in poly(A) length were observed in relation to transcript length (Figure 1H).

Identification of hundreds of alternative transcript isoforms

We used StringTie2 to identify possible transcript isoforms supported by the dRNA reads (Kovaka et al. 2019). This program has the advantage that it does not require that the reads are full-length, something which *a priori* cannot be determined for dRNA reads. StringTie2 yielded 5,799 transcripts, which showed a length distribution similar to that of annotated transcripts (Supplemental Table S2 and Supplemental Fig S6).

We identified a total of 332 alternative isoforms, in 262 different genes, which had an effect on the coding sequence. These events were novel, not annotated in PomBase. Because not all reads corresponded to full-length transcripts, a small proportion of the reads, 3.7%, mapped to different gene isoforms (7,271 multimapping reads out of 189,281). The formation of alternative isoforms decreased the relative amount of the reference protein, and in some cases could potentially lead to different protein products. We could distinguish between four types of events: intron retention (IR), intron inclusion (II), use of alternative splicing sites (AS) and exon skipping (ES) (Figure 2A). Intron retention events were denoted by dRNA sequences in which the intron was not spliced out. In the case of intron inclusion, a non-annotated intron was observed in a subset of the reads. Alternative splicing sites implied the use of different splice site donor or acceptor signals in a subset of the mRNA molecules. Finally, exon skipping was represented by sequences that lacked a complete exon. The most common event was intron retention, which represented about 80% of all events, followed by intron inclusion, in about 12% of the cases (Figure 2B).

In general, the number of alternative isoforms observed was one or at most two, although in some cases a larger number of isoforms could be observed (Figure 2C). The latter cases corresponded to genes of the killer meiotic drive system, a rapidly evolving family of parasitic and antidote genes (Eickbush et al. 2019). The maximum number of alternative isoforms recovered by StringTie2 was 9 in *wtf19*. These isoforms apparently originated from different types of exon/intron inclusion and exclusion events as well as by the use of alternative splice sites. The capacity of this gene to generate many alternative isoforms might be important for the ongoing arms race that characterizes the gene family.

In order to quantify the isoform expression levels we mapped the dRNA reads to the transcriptome and used only uniquely mapped reads. This allowed us to unambiguously distinguish between the reference and alternative isoform transcripts. As expected, alternative isoforms were in general found at lower frequencies than the annotated isoform (Figure 2D). Nevertheless, some non-annotated isoforms were found at very low frequencies, and there was a clear overlap between the expression levels of alternative isoforms and already annotated transcripts (Figure 2E). As many as 92 alternative isoforms had a frequency higher than 20%, and 172 cases higher than 10%. For example, retention of the first intron in *rpl22*, a gene encoding 60S ribosomal protein 22, showed a frequency of 30% (12,003 dRNA reads *versus* 28,910 for the reference mRNA). In *gdt2*, coding for a Golgi calcium ion transporter, the intron retention isoform was supported by 40% of the dRNA reads (565 *versus* 858). In the case of *elo1*, encoding an enzyme involved in fatty acid elongation, an isoform in which the third intron was included represented 42% of all transcripts (201 dRNA reads *versus* 276 for the reference mRNA). An extreme case was *etp1*, a gene involved in the adaptation to high concentrations of ethanol (Snowdon et al. 2009). In this case, the transcript containing the intron was the predominant one (85% of the dRNA reads, 182 out of 212). These examples were further validated by RT-PCR (Supplemental Fig S7 and Supplemental Table S7).

We observed a moderate but significant tendency for the first intron to be retained. In genes with two introns, and for which only one of the introns was retained, we found 64 cases in which the first intron was retained and 36 in which the second intron was retained (p-value =0.007 compared to 50/50, proportion test). We identified 9 exon skipping events, mostly affecting very small exons (Figure 2F). One example was *sir2*, encoding a histone deacetylase. An isoform in which the fourth exon was skipped represented 17.2% of the transcripts (37 dRNA reads *versus* 178 for the reference isoform).

Virtual translation of the sequences of the alternative isoforms indicated that, except in 10 cases, they resulted in proteins that were shorter than the annotated one (Figure

3A). We sought evidence of protein translation using previously published Ribo-seq data (Duncan and Mata 2017). We focused on intron retention isoforms, which are the easiest to analyze because we do not expect to have Ribo-seq reads mapping to the intron except if the intron is retained and translated. In 18 cases we found a minimum of 5 Ribo-seq reads supporting the alternative protein. One remarkable example was the translation of an ORF coding for a protein of only 13 amino acids (aa) in the intron retention isoform of *rpl22* (Figure 3B). The 13 aa protein was supported by 311 isoform-specific Ribo-seq reads. The number of Ribo-seq reads that map to a sequence can be used as a proxy of translation level, because each mapped Ribo-seq read potentially corresponds to a translating ribosome (Ingolia et al. 2009). Examination of the Ribo-seq coverage in the isoform-specific intronic region indicated that, although the 13 aa isoform was translated at levels estimated to be around 1/10 of the canonical 117 amino acid long protein, it was still among the top 10% most expressed proteins in the cell.

Another example of intron retention supported by ribosome profiling data was *uap2*, encoding the U2 snRNP-associated protein Uap2. About one fourth of the transcripts corresponded to retention of the first intron, resulting in a putative protein of 38 amino acids instead of the standard 367 aa protein (Figure 3C). Other genes displaying similar patterns were *mal3*, encoding a microtubule protein (des Georges et al. 2008), *rpb4*, encoding a RNA polymerase II complex subunit (Sakurai et al. 1999), and *not11*, encoding a CCR4-NOT complex subunit involved in the shortening of the poly(A) length and initiation of cytoplasmic mRNA decay (Ukleja et al. 2016). In *mal3*, intron retention was found at a frequency of 38% and resulted in a putative protein of 26 aa, in *rpb4* of 26.8% and a protein of 20 aa and, in *not11* of 60% and a protein of 55 aa. A very different case was *etp1*, the intron contained no stop codon in frame and, for this reason, the resulting protein was predicted to be 15 aa longer than the reference one (Figure 3D).

In other genes, different isoforms were generated by the use of alternative splice sites, by exon skipping or intron inclusion. One example was *pat10* (SPAC18B11.08c), a gene that encodes an endoplasmic reticulum protein which is part of a chaperone complex

involved in the biogenesis of proteins with multiple transmembrane domains (Chitwood and Hegde 2020). The reference transcript is composed of 5 exons and encodes a protein that is 95 amino acids long. The dRNA reads provided direct evidence of an alternative isoform arising from a downstream alternative splice site in intron 3 and skipping of exon 4. The alternative isoform had a frequency of 27% (161 dRNA reads alternative isoform vs 431 for the reference) and resulted in a putative protein of 74 amino acids (Supplemental Fig S8).

Intron retention is associated with extended Poly(A) tails

We next investigated poly(A) tail length with respect to alternative splicing events. First, for each type of event, we compared the poly(A) length of the dRNA reads in the alternative isoform and the reference isoform. Collectively, we could observe significant differences for intron retention events and alternative splicing site events (Figure 4A). In both cases, poly(A) tails tended to be longer in the alternative isoform. However, when we compared the differences in poly(A) length for each gene, taking the average value for all reads mapping to the same isoform, a consistent difference was only observed for intron retention events (Figure 4B).

In a recent study in *S. cerevisiae* the authors concluded that the poly(A) tail of newly transcribed transcripts is 50 adenosines long on average, and is shortened in the cytoplasm to 40 adenosines on average (Tudek et al. 2021). We thus considered the possibility that longer poly(A) tails could be indicative of not fully processed transcripts still retained in the nucleus. However, isoforms with translation evidence, and thus presumably located in the cytoplasm, also had longer poly(A) tails than the reference isoforms (Supplemental Fig S9). Thus, the data fit quite well the previously observed negative correlation between expression level and poly(A) length (Figure 4C). Another possible explanation was that only a fraction of the molecules was being translated, whereas the rest was retained in the nucleus, resulting in overall longer poly(A) tails. Because the latter possibility cannot be tested with the current data, the question remains open.

In general, intron retention isoforms tended to be less abundant than the reference isoform and also tended to have longer poly(A) tails, as shown in the examples in Figure 4D. When we examined cases in which the alternative and reference isoforms had relatively similar abundances, the results varied depending on the gene. In some cases, such as *not11*, the poly(A) tail length of the alternative and reference isoform was not significantly different. In other cases, including *mal3*, *rps13* and *vps38*, the differences were significant, although relatively small (Supplemental Fig S10). In contrast, *slm3* and *red1* showed very strong and significant differences in poly(A) tail length between the alternative and reference isoforms (approximately 80 *versus* 50 nt, respectively)(Supplemental Fig S11).

Discovery of new transcribed loci

The reconstruction of the transcriptome using the dRNA reads also resulted in the discovery of 214 completely novel transcripts, whose coordinates on the genome did not show any overlap to annotated genes on the same strand. Mapping the Illumina reads to these transcripts confirmed the expression of the majority of them (168 out of 214). We found that about three quarters of them, 158 (74%), overlapped other genes on the opposite orientation and were classified as antisense. The remaining 56 transcripts were located in regions with no other annotated features and were classified as intergenic.

The novel transcripts tended to be shorter than the annotated ones, especially the intergenic ones (Figure 5A and Supplemental Table S2). They also tended to be expressed at lower levels than annotated transcripts, although in this case there were no significant differences between antisense and intergenic transcripts. Since novel transcripts are lowly expressed, their detection might largely depend on the sequencing coverage. To explore this, we generated saturation curves by subsampling the number of original mapped dRNA reads. Whereas the number of known genes that could be detected reached a plateau at approximately 1.5 Million reads, the number of novel transcripts showed an approximately linear relationship with the number of sequencing reads (Supplemental Fig S12). We also investigated the fitness associated

with the novel transcripts by using data from a previous saturating transposon mutagenesis experiment (Grech et al. 2019). We found that the level of constraints in the set of novel transcripts showed no significant differences to that of annotated non-coding RNAs and was clearly weaker than in coding sequences (Supplemental Fig S13).

We next used the ribosome profiling data to investigate the translation patterns in the novel transcripts. We predicted putatively translated ORFs using RibORF (Ji et al. 2015). This program produces a score based on three nucleotide periodicity and homogeneity of the Ribo-seq reads along the ORF. In previous studies we established that a RibORF score > 0.7 was associated with significant translation activity (Blevins et al. 2021; Moro et al. 2021). As expected, the vast majority of the annotated coding sequences in mRNAs were classified correctly by the program (4,514 out of 4,560, 98.99%)(Figure 5B). In addition, 16% of the annotated ncRNAs also contained ORFs with evidence of translation (Supplemental Table S3 and Supplemental Fig S14). These findings are in line with the translation signatures observed in a large fraction of the long non-coding RNAs in other biological systems (Ruiz-Orera et al. 2014; Ji et al. 2015; Chen et al. 2020).

Among the newly discovered transcripts we identified 12 cases with evidence of translation, 8 antisense and 4 intergenic. One of the intergenic transcripts contained two putatively translated ORFs. The encoded proteins were small, with a median length of 44.5 aa for antisense transcripts and 25 aa for intergenic transcripts, compared to 393 aa for canonical ORFs (Figure 5C). The orientation of four of these novel transcripts, and the distance to the nearest transcription start site (< 400 bp), suggested divergent transcription from a bidirectional promoter (Supplemental Table S4). One of the newly identified proteins showed significant homology to several prokaryotic proteins as well as to an uncharacterized protein from the fungus *Macrophomina phaseolina* (Figure 5D). Given the sparse species distribution, it seems likely that this protein has originated by horizontal gene transfer, probably from a bacteria. The rest of genes did not have homology to any other annotated protein, neither to a set of novel translated ORFs recently discovered in *S. cerevisiae* (Blevins et

al. 2021). Therefore, these genes might have originated *de novo* in the *S. pombe* lineage.

Finally, we also used the dRNA sequencing data to annotate 5' and 3' untranslated regions (UTRs), focusing on those that were not yet annotated in PomBase (266 mRNA without a 5'UTR and 337 mRNAs without a 3'UTR). Nanopore data is expected to be very accurate for the 3'end but less so for the 5'end, as the first 10-15 nucleotides of the mRNA are normally not recovered. Using the dRNA based transcriptome we annotated 105 5'UTRs and 75 3'UTRs that were previously missing. The median size of these sequences was 217 and 256, respectively, which was comparable to the length of the annotated ones (median 168 for 5'UTR and 259 for 3'UTR, respectively). Thus, dRNA provides an effective way to annotate 3'UTRs and to some extent also 5'UTRs.

Discussion

Native or direct RNA sequencing (dRNA) provides unprecedented resolution to study the transcriptome. The technique has provided new insights into the features of the transcripts expressed in several eukaryotic species, including human, *C. elegans* and *Arabidopsis* (Workman et al. 2019; Li et al. 2020; Zhang et al. 2020; Roach et al. 2020). Here we applied dRNA to the fission yeast *S. pombe*, an intron-rich unicellular eukaryote that has become a very useful model to study splicing (Yan et al. 2015; Fair and Pleiss 2017). Our strategy was based on obtaining a very high coverage of the transcriptome to uncover alternative splice forms and lowly expressed transcripts. We obtained RNA sequences for 97% of the annotated mRNAs and 87% of the non-coding RNAs. Additionally, we characterized 332 non-annotated alternative isoforms and 214 completely new transcripts, about three quarters of which overlapped other genes in antisense orientation. The work presents a new view of the *S. pombe* transcriptome because a substantial number of the newly identified alternative splicing isoforms occur at high frequencies, and some are likely to translate alternative proteins, indicating that the transcriptome is more complex and functionally diverse than previously thought.

By using dRNA sequencing, it is possible to recover poly(A) tail length information from the sequencing reads. In eukaryotes, poly(A) tail lengthening is associated with increased mRNA stability and poly(A) tail shortening with mRNA degradation (Richter JD 2000; Dreyfus and Régnier 2002). Here we characterized poly(A) length in *S. pombe* and investigated if transcripts which showed diverse splicing patterns presented alterations in poly(A) length. For the complete *S. pombe* poly(A)⁺ transcriptome we found that the average poly(A) tail is around 50 nucleotides, very similar to humans (Workman et al. 2019) and *C. elegans* (Roach et al. 2020), highlighting the high evolutionary conservation of this trend. We also found that poly(A) length tends to be shorter in mRNAs encoding highly expressed proteins, such as translation related proteins, than in mRNAs that are expressed at low levels during exponential growth conditions, such as many meiosis-specific proteins. Similar results were recently observed using TAIL-Seq data in *C. elegans* (Lima et al. 2017). These results are unexpected given previous experimental evidence that poly(A) tail elongation promotes transcript stability and translatability (Preiss et al. 1998; Eichhorn et al. 2016), and point to yet poorly understood mechanisms controlling poly(A) tail dynamics in different kinds of transcripts.

Upon synthesis, transcripts are polyadenylated and later exported to the cytoplasm, where they eventually decay, a process that involves poly(A) tail deadenylation (Tudek et al. 2021). We found that the poly(A) tail of the alternative isoforms was generally longer than that of the reference transcripts, especially in the case of intron retention events. This could be explained by the negative relationship between expression level and poly(A) tail length, but it could also be that some intron retention isoforms were retained in the nucleus whereas others, including those for which we found translation evidence, are exported to the cytoplasm. In order to eliminate the influence of expression level, we examined the differences in poly(A) length for cases in which the reference and alternative isoform were expressed at similar levels. We found two different scenarios. In the first one, the alternative isoform had a similar poly(A) tail length than the reference isoform. One example was *not11*, with a median poly(A) length of 51.2 for the reference isoform and 50.1 for the intron retention isoform. In the case of *rps13*, poly(A) length was 34 for the reference isoform and 37.6 for the

intron retention isoform, in line with the high expression of this gene. In the second scenario, there was a very clear difference in poly(A) length between the two isoforms. This was the case of *slm3* and *red1*, with a median poly(A) tail of around 80 for the intron retention isoform and 50 for the fully spliced form. These results pointed to the existence of two classes of isoforms, the first class representing possibly functional alternative proteins and the second class transcripts retained in the nucleus. Nuclear retention of incompletely processed mRNAs might be an additional layer of gene expression control. For example, in mouse cells, *Gabbr1* RNA remains incompletely spliced on the chromatin in embryonic stem cells, being only fully processed and exported for translation upon neuronal differentiation (Yeom et al. 2021).

Nanopore native mRNA sequencing is a powerful technique to uncover the full set of transcripts generated by different combinations of exons and introns, which cannot be accurately solved by Illumina reads. At the same time, the cost and scalability it offers is comparable to that of Illumina sequencing. We generated around 7 Million dRNA reads in an organism with about 7,000 genes. This high coverage, and the lack of amplification biases, allowed us to perform a very precise estimation of the abundance of alternative splicing transcripts. We found that about one third of the events occurred at a very high frequency (> 20%), which suggests that many of the events are functional. Intron retention (IR) events were the most common ones, as also observed in other fungi and plants (Gonzalez-Hilarion et al. 2016; Ullah et al. 2018). IR often results in premature termination codons, which could potentially trigger nonsense mediated decay (NMD). However, studies in *Cryptococcus neoformans* have shown that IR is largely independent of NMD because mutants that do not express Upf proteins, which are the proteins that mediate NMD (Kervestin and Jacobson 2012), do not show IR upregulation (Gonzalez-Hilarion et al. 2016). By using ribosome profiling data we obtained evidence that some intron retention isoforms are likely to translate alternative proteins. Thus, in some cases, the same gene may be used to expressed multiple proteins. A previously described example is cardiolipin synthase, which is specifically produced by the intron IV retention isoform of SPAC22A12.08c mRNA (Virčíková et al. 2018). Here we found several possible examples of novel proteins generated by intron retention isoforms, which will need to be inspected in more detail.

In addition to intron retention, other not yet characterized proteins can be formed by intron inclusion, exon skipping or the use of alternative splice sites. Taken together, the results show that alternative splicing in fission yeast is likely to play a more important role in proteome diversification than previously anticipated.

ORFs encoding proteins smaller than 100 amino acids are difficult to annotate because they cannot be distinguished from randomly occurring ORFs using computational means. The emergence of ribosome profiling has changed this situation, because it enables the identification of ORFs with significant translation signatures regardless of the size of the ORF (Ingolia et al. 2009; Ruiz-Orera and Albà 2019b). The technique has revealed that the number of small proteins in the cells is probably much higher than previously suspected, including many micropeptides resulting from the translation of small ORFs in transcripts currently annotated as long non-coding RNAs (Ji et al. 2015; Calviello et al. 2016; Ruiz-Orera and Albà 2019a; Chen et al. 2020; Douka et al. 2021). When we examined the Ribo-seq data in *S. pombe* for the complete transcriptome we found very similar results to those previously described in mammals. The genome contains a relatively large number of annotated ncRNAs, 1,527, many of which are antisense to protein-coding genes. Here we identified 214 additional ones. In line with previous findings in *S. pombe* (Duncan and Mata 2014), a sizable fraction of these ncRNAs (16%) showed translation evidence. Some of the translated ORFs in ncRNAs might encode functional micropeptides, whereas others, especially for very lowly abundant ncRNAs, could represent pervasive non-functional translation activities.

In summary, deep native RNA sequencing using Nanopore has uncovered an unexpectedly large number of high frequency alternative splicing isoforms in *S. pombe*. Many of these isoforms could encode alternative, generally smaller, proteins, of yet unknown functions. We have also identified a group of intron retention RNAs that show abnormally long poly(A) tails and which could potentially be regulating gene expression. The work provides new resources and methodologies for researchers investigating differential splicing in the fission yeast model.

Material and Methods

S. pombe cultures

Schizosaccharomyces pombe (strain CBS5682) was grown in a rich medium at 30°C and harvested during log-phased growth (OD₆₀₀~0.5). The medium was identical to the one previously employed to perform RNA sequencing of the same *S. pombe* isolate using Illumina Technology (Blevins et al. 2019). The composition of the medium, defined by Tsankov and collaborators (Tsankov et al. 2010), can be found in Supplemental Tables S5 and S6.

RNA extraction

We extracted total RNA from *S. pombe* using the phenol chloroform extraction method (Castillo et al. 2003). Briefly, cells were grown to a final OD₆₀₀ of 0.5. Yeast cultures (25–50 ml) were then centrifuged at 1,500 rpm for 3 min and washed with H₂O, and cell pellets were immediately kept on ice. Each sample was then resuspended in 0.4 ml of AE buffer (50 mM sodium acetate, pH 5.3, 10 mM EDTA, pH 8.0). Sodium dodecyl sulfate was then added to a final concentration of 1%, and proteins and DNA were extracted by adding 0.6 ml of acidic phenol/chloroform (V/V), followed by incubation at 65°C for 5 min. The aqueous phase was separated by centrifugation at 14,000 rpm for 2 min at 4°C and washed with a volume of chloroform and separated by centrifugation at 14,000 rpm for 2 min at 4°C. RNA was precipitated from the aqueous phase with ethanol. RIN quality scores were in the range 9.6-10. We subsequently performed poly(A)⁺ RNA purification using NEBNext Poly(A) Magnetic Isolation Module and concentration with the Monarch RNA Cleanup Kit. The polyA⁺ purification steps were performed at the Genomics Core Facility of Universitat Pompeu Fabra.

Direct RNA sequencing

The poly(A)⁺ RNA was used for direct RNA (dRNA) sequencing in an Oxford Nanopore Technologies gridion X4. Direct RNA sequencing offers the advantage over cDNA

sequencing that strand orientation information is maintained. The protocol involves adaptor ligation and the molecules pass through a ionic current, adaptors and poly(A)⁺ tail first and then the rest of the molecule. The *S. pombe* samples were run in four flowcells. For each run we used about 600 ng of poly(A)⁺ RNA in 10 uL of volume. The direct RNA sequencing kit SQK-RNA002 was employed. The basecalling was performed on live mode (during the sequencing) through the Guppy v.4.0.11 integrated on minKNOW v.4.0.5, using the HAC model. Nanopore dRNA sequencing and basecalling was done by the Centro Nacional de Análisis Genómico (CNAG, Barcelona).

We pulled together the output of the four runs, obtaining a total of 7,297,641 reads. We discarded any reads smaller than 150 bases and longer than 15,000 bases (likely artefacts), and removed any possible adapters with Porechop (<https://github.com/rrwick/Porechop>).

Read mapping and correction

In order to decrease the error rate of the reads, and facilitate the subsequent *de novo* transcript assembly, we decided to correct the dRNA reads with Illumina RNA-seq reads from yeast grown in the same conditions, using the software fmlrc with default parameters (Wang et al. 2018). Subsequently, we used TranscriptClean to correct the remaining errors (Wyman and Mortazavi 2019). The set of Illumina reads comprised 22,389,887 strand-specific 50 bp reads (Blevins et al. 2019). To run fmlrc we first had to transform all uracil (U) bases in the dRNA reads to thymine (T) bases and, once the reads had been corrected, transform them back to Us. The reads were mapped to the genome using minimap2 (Li 2018) with the following options: minimap2 -t 6 -ax splice -uf -k14 --secondary=no -G 260. The Nanopore/Illumina hybrid reads showed an error rate of 7.2%, mainly because of poor coverage of the mRNA 3'ends by Illumina reads. Individual read error rate was calculated using the CIGAR values of the reads aligned to the reference genome. Average error rate was calculated using SAMtools stats (Li et al. 2009). Reads that had a mapping quality of less than 5 were eliminated. The final set of “clean” reads comprised 5,054,233 reads.

In order to estimate the number of reads that were full length we first obtained the length of the mapped reads with `bam_alignment_length.py` from the `wub` package. Reads with length equal or longer than the transcript in which they are aligned minus 50 nucleotides were estimated to be full length reads.

Transcript assembly

We used StringTie2 to obtain a *S. pombe* transcriptome directly from the set of dRNA clean reads, as previously described (Kovaka et al. 2019). The parameters were as follows: `-l - conservative -G -t -c 1.5 -f 0.05`. The program uses the reference genome and the mapped reads for the assembly and, optionally, a gene annotation file. We chose to use the gene annotations to guide the assembly because this option provided a direct mapping to already known genes, facilitating the assembly. The reference genome and gene annotation files were downloaded from the PomBase database on Feb 1 2021 (Lock et al. 2019). For the assembly we considered all mapped reads except those that mapped to multiple sites (those with a MAPQ score > 5 and which had alignments with the flags 0 and 16 in the SAM file). The number of reads used for the assembly was 5,054,233 reads (“clean” reads). We eliminated any assembled transcripts shorter than 150 nucleotides. The resulting annotation file was named “StringTie transcriptome”. It contained 5,799 different transcripts.

The identification of alternative isoforms was based on the Stringtie transcriptome. StringTie2 recovered all isoforms with an estimated frequency higher than 5% with respect to the most common isoform. Stringtie transcripts that corresponded to annotated genes but were different from the reference transcript were classified into one of the following transcript alternative isoform types: Intron retention, Intron inclusion, Exon skipping and Alternative splicing site. The classification was based on the number and genomic position of the exons represented in the dRNA reads. We focused on events affecting the coding sequence, as they were the ones most likely to have functional consequences. These selection steps resulted in 262 genes with alternative isoforms. The total number of events was 332 because some genes had more than one possible event.

The identification of novel transcribed loci was also based on the Stringtie transcriptome. Novel transcripts were defined as those not overlapping any annotated gene. We obtained 214 novel transcripts, 158 of which overlapped another gene in antisense orientation. Saturation curves were produced by subsampling the mapped reads with SAMtools -s (0.1 to 0.9 of the total number of reads) and running Stringtie2 again. The assembled transcripts were then compared to the set of novel or annotated transcripts using the intersect function from BEDTools (Quinlan et al. 2010).

Transcript expression quantification with Nanopore reads

The number of mapped dRNA transcripts was calculated first mapping the transcripts to the transcriptome with minimap2 (arguments were `-t 6 -ax map-ont -K 10G --for-only --no-long-join -r 10,10 --secondary=no`) and then using the script `bam_counts_reads.py` from nanoporetech package with the `-a` argument set to 5 to consider only uniquely mapped reads. To quantify the expression of already annotated transcripts we used the PomBase transcriptome. We mapped reads to 5,026 annotated mRNAs (97.8%), the average number of mapped reads per mRNA was 1,130 and the median 1,259. As each read corresponds to a native mRNA molecule, we consider the number of mapped reads to be equivalent to coverage. To quantify the expression of different isoforms and novel transcripts we used the Stringtie transcriptome. The data allowed to compare the relative abundance of different isoforms with high accuracy. Supplemental Table S8 contains information on gene expression data.

Transcript expression quantification with Illumina reads

Illumina reads from *S. pombe* grown in rich medium were obtained from a previous study (Blevins et al. 2019). The Illumina reads were 50 bp long and strand-specific. They were mapped to the reference transcriptome with HISAT2 (Kim et al. 2019) with `--rna-strandness "RF"` option. Reads that were not aligned in the expected direction (tag XS:A:-) were discarded from the resulting BAM file. The reads mapping to each

transcript were quantified with `bam_counts_reads.py` from Oxford Nanopore Technologies nanoporetech package (<https://github.com/nanoporetech/wub>). The counts were normalized to fragments per Kilobase per Million mapped reads (FPKM).

Poly(A) tail quantification

Poly(A) tail lengths were estimated at the read level using the nanopolish (v 0.13.3) `polya` script. As input to the command `nanopolish polya` we used the raw FAST5 and FASTQ files in addition to the corrected reads mapped to the genome of *S. pombe*. Finally, only those reads with the quality control provided by nanopolish with the tag “PASS” were considered. Supplemental Table S8 contains information on poly(A) length.

Gene Ontology enrichment

Gene Ontology (GO) term enrichment was performed using the web application AnGeLi (Bitton et al. 2015b). We identified overrepresented or underrepresented GO terms for genes in the top 10% or bottom 10% regarding average poly(A) tail length. As a pre-defined background we used all Genes. GO term information about individual genes was extracted from AnGeLi too.

Analysis of ribosome profiling data

To study the translatability of the transcripts we used previously published ribosome profiling data (Duncan and Mata 2017). The data was from untreated cells (ArrayExpress ERR1994961 and ERR1994962). We filtered out ribosomal RNAs and mapped the reads to the *S. pombe* genome with TopHat (v 2.1.1)(Kim et al. 2013) with default options. We used the script `offsetCorrect.pl` from RibORF to identify the P-site of each read (Ji et al. 2015). The number of mapped P-sites was used to evaluate the level of translation of the intronic regions in intron retention isoforms. Ribo-seq coverage was calculated in all the transcripts with CDS in PomBase using `htseq-count` (-m intersection-strict) with the genome and P-sites obtained before. Ribo-seq coverage

of the alternative isoform of SPAC11E3.15.1 was calculated using only the intronic region until the stop codon included in the intron. RibORF was also used to predict translated ORFs in novel transcripts and ncRNAs. We required at least 10 mapped reads and a ribORF score > 0.7. When 2 or more ORFs showed overlapped on the same transcript we kept the longest ORF. Supplemental Table S8 contains information on isoforms with translation evidence (≥ 5 mapped Ribo-Seq reads), novel non-annotated transcripts and UTRs. Supplemental Table S9 contains information on annotated lncRNAs with evidence of translation (RibORF score > 0.7).

Genomic DNA preparation

Genomic DNA was prepared from 10 ml of yeast cultures grown to saturation. Cells were pelleted at 1,500 rpm for 3 min, washed with H₂O, pellets were immediately frozen in liquid nitrogen. Samples were re-suspended in 0.2 ml of genomic DNA preparation buffer (10 mM Tris-HCl, pH 8.0, 100 mM NaCl, 2% Triton X-100, 1% SDS, 1 mM EDTA), 0.1 ml neutral phenol, and 0.1 ml chloroform. Glass beads were added and cells were lysed in Vortex Genie 2 (Scientific Industries). After removal of glass beads, homogenates were centrifuged at 20,000 g for 5 min (4°C), supernatants were collected, and 0.2 ml of chloroform was added. Following centrifugation, supernatants were collected, and DNA was precipitated with 1/10 volume of 3 M sodium acetate (pH 5.2) and 2.5 volumes of EtOH, followed by incubation 30 min at -80°C. Following centrifugation, pellets were washed with 1 ml EtOH (70%), air-dried and re-suspended in 40 μ l of TE-buffer (10 mM Tris-HCl, pH8.0, 1 mM EDTA) containing 1 μ l RNase A. RNA was digested for 30 min at 37°C and DNA was afterward stored at -20°C.

cDNA preparation

RNA (25 μ g) was treated with 0.5 μ l DNase I for 30 min at 37°C and then inactivated at 75°C for 10 min. Reverse transcriptase (RT) reactions were performed with 8 μ g of DNase I-digested RNA, following manufacturer's instructions (High Capacity Reverse Transcriptase kit, Thermo Fisher Scientific) (10 min at 25°C, 120 min at 37°C and 5 min at 85°C) in the presence or absence of RT.

Polymerase chain reactions (PCRs)

PCRs were performed in a total volume of 20 μ l using 0.5 μ l of the cDNA reactions with primers listed in Supplemental Table S7. 50 ng of genomic DNA were used as reference for the unspliced transcript. PCR products were separated on 2% agarose TBE gels. Digital images were acquired with Bio-Rad software.

Statistical tests and plots

The generation of plots and statistical tests was performed using the R package (R Core Team. R Foundation for Statistical Computing, Vienna 2020). Figures were made with ggplot2 (Wickham 2016).

Data access

The ONT dRNA raw sequences generated in this study have been submitted to the NCBI BioProject database (<https://www.ncbi.nlm.nih.gov/bioproject/>) under accession number PRJNA791394. Additional Supplemental Data can be found at Figshare <https://doi.org/10.6084/m9.figshare.19368146>. This comprises the set of clean reads used for transcript reconstruction and quantification, the dRNA-based Stringtie transcriptome (non-curated), the dRNA-based *S. pombe* transcriptome for genes with alternative isoforms (curated), 5'UTR and 3'UTR annotation files, as well as Supplemental Tables S8 and S9. Supplemental Table S8 contains information on gene expression values, alternative gene isoforms and transcript poly(A) tail length. Supplemental Table S9 contains information on non-coding RNAs containing ORFs with evidence of translation.

Competing interests statement

The authors declare no conflicts of interest.

Acknowledgements

This work benefited from preliminary Nanopore RNA-Seq data analyses performed by Bea Calvo and Audald Lloret-Villas, as well as discussions with Eduardo Eyra and Ivan de la Rubia. We acknowledge funding from Ministerio de Ciencia e Innovación (MCI), Agencia Estatal de Investigación (AEI) grant PGC2018–094091-B-I00, co-funded by Fondo Europeo de Desarrollo Regional (FEDER) and from Agència de Gestió d'Ajuts Universitaris i de Recerca, Generalitat de Catalunya, grant 2017SGR01020.

Author contributions

W.R.B., J.C.M., E.H., J.A. and M.M.A. contributed to the conceptualization of the study and design of experiments. W.R.B., J.C.M. and M.C. performed yeast growth cultures and RNA extraction. M.C. performed the RT-PCR experiments. J.C.M. developed most pipelines and performed the majority of analyses. M.H. and S.G.M. analyzed the data related to novel genes and the annotation of UTRs. J.C.M. and M.M.A. wrote the manuscript with input from all authors.

Figure legends

Figure 1. dRNA sequencing of *S. pombe*. A. Error rate distribution of raw and clean reads. Error rate is the percentage of aligned positions which contain a mismatch or indel. “Raw” refers to the original reads. “Raw P7” refers to the original reads with quality score $Q \geq 7$. “Clean” are the corrected reads used in this work. **B. Sequence length distribution of raw and clean reads.** Number of raw reads: 7,097,130. Number of clean reads: 5,054,233. Median value raw reads: 500. Median value clean reads: 620. Differences in the distribution are significant by a Wilcoxon test ($p\text{-value} < 2.2 \times 10^{-16}$). **C. Correlation transcript abundance ONT dRNA versus Illumina.** The reads were mapped to the PomBase transcriptome. In the case of dRNA reads we simply divided the number of reads by the number of Million reads (Reads per Million or RPM). For Illumina reads we calculated Transcripts per Million (TMP), normalizing by transcript

length as well as number of Million reads. We selected transcripts expressed in at least one of the two datasets; transcripts with multimapping Illumina reads were removed. Number of transcripts analyzed = 4,999. **D. Estimated number of transcripts with at least one full length read with respect to transcript length.** The data is shown for different transcript length deciles: [71.0 - 633.2), [633.2 - 923.4), [923.4 - 1,175.0), [1,175.0 - 1,395.0), [1,395.0 - 1,637.0), [1,637.0 - 1,911.2), [1,911.2 - 2,227.4), [2,227.4 - 2,695.0), [2,695.0 - 3,444.6), [3,444.6 - 15,022.0]. Number of transcripts: 6,453. **E. dRNA counts with respect to transcript length.** Bins are the same than in figure 1D. **F. Poly(A) tail distribution in mRNAs and ncRNAs.** Poly(A) tail is estimated as the mean of the poly(A) tail length of all the reads that map to each transcript. Differences are significant according to a Wilcoxon test ($p\text{-value} < 10^{-5}$). **G. Relationship between poly(A) tail length and transcript abundance.** For each transcript, the average poly(A) tail length of all the reads mapping to the transcript is taken. Only mRNAs are taken into account for this calculation ($n = 4,995$). Genes related with reproduction (GO:0000003) and translation (GO:0006412) are highlighted. Highly expressed transcripts tend to have shorter poly(A) tails. The correlation is significant (Spearman's $\rho = -0.376$; $p\text{-value} = 9.3 \times 10^{-168}$). **H. Distribution of poly(A) tail length with respect to transcript length.** Bins are the same as in Figure 1D. Poly(A) tail length is homogeneously distributed across different transcript length classes.

Figure 2. Identification of alternative isoforms using dRNA sequencing. **A. Alternative splicing isoform classes.** We built a transcriptome using the dRNA reads with Stringtie2. We identified 332 alternative isoforms in 263 different genes. The plot shows one example of each of the four main classes of alternative isoforms detected. Diagrams of the exons of the reference and the alternative isoform as shown, together with the dRNA coverage along the gene. **B. Number of different types of splicing isoforms.** Intron retention represents about 80% of the events. **C. Number of isoforms per gene.** In most cases only one alternative isoform was detected. The most extreme case corresponds to *wtf19*, with 2 annotated isoforms and 9 additional alternative isoforms detected here. **D. Relative abundance of reference and alternative isoforms for each gene.** Data is for the genes containing at least one alternative isoform. The abundance is computed using the number of mapped dRNA reads, the fraction is then

calculated over all isoforms containing mapped reads. Number of reference isoforms: 263 (two for *wtf19*). Number of new alternative isoforms detected here: 332. Median fraction reference isoforms: 0.873. Median fraction alternative isoforms: 0.105. **E. Abundance of reference and alternative isoforms.** Number of dRNA reads mapped to reference and alternative isoforms. Numbers of isoforms as in D. **F. Skipped exons tend to be smaller than the complete set of exons in the reference annotations.** Median length reference exons: 170; median length skipped exons 63; p-value = 0.00752 Wilcoxon test.

Figure 3. Proteome expansion by alternative splicing. A. Difference in size between the putative alternative protein and the reference protein. Data is shown for the four main classes of alternative splicing events. In the vast majority of cases the alternative protein would be smaller than the canonical protein. **B. Alternative 13aa protein isoform in the *rpl22* gene.** The diagram shows the putative coding sequence in the alternative intron retention isoform. A stop codon in frame in the intronic region results in the translation of a 13 aa protein. We obtained Ribo-seq support for the 13 aa alternative protein. We also estimated the Ribo-seq coverage of the Rpl22 canonical protein (SPAC11E3.15.1), and the 13 aa alternative isoform (STRG.1957.2), using isoform-specific coding sequences. The values are compared to those for the coding sequences of all transcripts with 5 or more Ribo-seq reads mapped to the P-site (n=5,669). The gray area covers 90% of cases. **C. Alternative 38 aa protein isoform in the *uap2* gene.** Intron retention in *uap2* generates a shorter coding sequence, encoding a putative 38 aa protein. **D. Alternative 361 aa protein isoform in the *etp1* gene.** Intron retention in *etp1* results in a protein that is 15 aa longer than the reference one.

Figure 4. Poly(A) length in alternative transcript isoforms. A Distribution of poly(A) length by isoform type. We computed poly(A) length for all dRNA reads with nanopore. Only reads with the label PASS were considered. n= number of reads with poly(A) length information; m=median poly(A) length. IR: Intron retention; II: Intron inclusion; AS: Alternative splicing site; ES: Exon skipping. Significant differences were identified for isoform retention and alternative splicing site events when compared to the corresponding reference isoforms (Wilcoxon test, *** p-value < 10⁻⁵). **B.**

Difference in the average poly(A) length between the alternative and reference isoforms. Longer poly(A) lengths were consistently observed for intron retention isoforms. **C. Negative correlation between average poly(A) length and expression level for reference and alternative transcript isoforms.** The data is only for genes in which we detected alternative isoforms. Reference refers to the annotated isoform. Spearman's $\rho = -0.41$, $p = 3.14 \times 10^{-23}$. **D. Examples poly(A) length differences between the reference and the intron retention isoforms.** We computed the poly(A) length for the dRNA reads that correspond to each of the isoforms. The first example corresponds to RNA polymerase II subunit 4 (*rpb4*, SPBC337.14), with about 27% of the reads corresponding to the intron retention isoform. The second example corresponds to a nucleolar RNA-binding protein also implicated in mRNA processing (*tam10*, SPBC14C8.19), with about 18% of the reads mapping to the intron retention isoform. In both cases poly(A) length showed a significant tendency to be longer in the intron retention isoform (Wilcoxon test, *** p -value $< 10^{-5}$, ns: non-significant).

Figure 5. New transcripts and peptides. A. Transcript length and gene expression for novel antisense and intergenic transcripts. Novel antisense transcripts tend to be longer than intergenic transcripts, and both classes of transcripts are shorter than already characterized ones (Wilcoxon test, p -value $< 10^{-5}$). Gene expression levels are lower in novel transcripts compared to not novel ones (Wilcoxon test, p -value $< 10^{-5}$). Novel antisense and intergenic transcripts show a similar expression level distribution. TPM: transcripts per Million, quantification provided by Stringtie. **B. Prediction of translated ORFs.** The plot shows the RibORF score *versus* the number of Ribo-seq mapped reads for different classes of transcripts. ORFs with at least 10 mapped reads and a RibORF score higher than 0.7 were selected as translated. The score is based on the Ribo-seq read three nucleotide periodicity and homogeneity. Nanopore novel indicates transcripts that did not overlap any annotated transcript, 12 of these transcripts contained ORFs with translation signatures. PomBase non-coding indicates annotated ncRNAs which also had translation signatures. **C. Novel translated ORFs are shorter than annotated ones.** Comparison of amino acid length for ORFs with evidence of translation in novel transcripts and annotated coding sequences. Differences are statistically significant (Wilcoxon test, p -value $< 10^{-5}$). **D. New protein**

identified in *S. pombe*. The protein labelled as STRG.2094.1 showed significant similarity to other bacteria proteins (BLASTP e-values between 10^{-2} and 10^{-7}) and, in eukaryotes, only to a protein from the fungus *Macrophomina phaseolina* (e-value $< 10^{-9}$). The blue box represents the homologous region; lines at the side represent additional protein sequence.

References

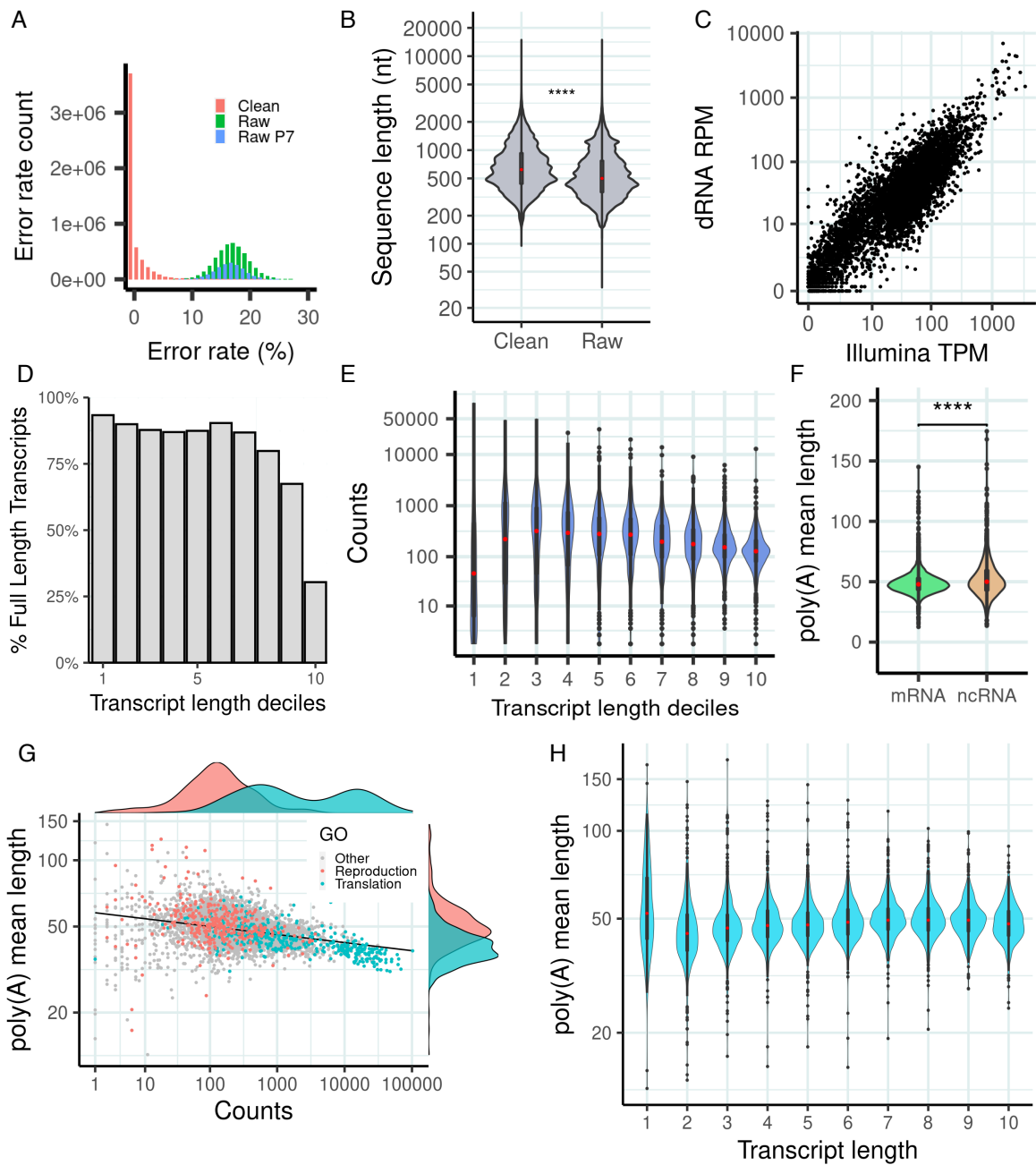
- Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. 2020. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol* **21**: 30.
- Awan AR, Manfredo A, Pleiss JA. 2013. Lariat sequencing in a unicellular yeast identifies regulated alternative splicing of exons that are evolutionarily conserved with humans. *Proc Natl Acad Sci U S A* **110**: 12762–7.
- Bitton DA, Atkinson SR, Rallis C, Smith GC, Ellis DA, Chen YYC, Malecki M, Codlin S, Lemay J-F, Cotobal C, et al. 2015a. Widespread exon skipping triggers degradation by nuclear RNA surveillance in fission yeast. *Genome Res* **25**: 884–96.
- Bitton DA, Schubert F, Dey S, Okoniewski M, Smith GC, Khadayate S, Pancaldi V, Wood V, Bähler J. 2015b. AnGeLi: A Tool for the Analysis of Gene Lists from Fission Yeast. *Front Genet* **6**: 330.
- Blevins WR, Carey LB, Albà MM. 2019. Transcriptomics data of 11 species of yeast identically grown in rich media and oxidative stress conditions. *BMC Res Notes* **12**: 250.
- Blevins WR, Ruiz-Orera J, Messeguer X, Blasco-Moreno B, Villanueva-Cañas JL, Espinar L, Díez J, Carey LB, Albà MM. 2021. Uncovering de novo gene birth in yeast using deep transcriptomics. *Nat Commun* **12**: 604.
- Brar GA, Weissman JS. 2015. Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nat Rev Mol Cell Biol* **16**: 651–664.
- Bresson SM, Hunter O V, Hunter AC, Conrad NK. 2015. Canonical Poly(A) Polymerase Activity Promotes the Decay of a Wide Variety of Mammalian Nuclear RNAs. *PLoS Genet* **11**: e1005610.
- Calviello L, Mukherjee N, Wyler E, Zauber H, Hirsekorn A, Selbach M, Landthaler M, Obermayer B, Ohler U. 2016. Detecting actively translated open reading frames in ribosome profiling data. *Nat Meth* **13**: 165–170.
- Castillo EA, Vivancos AP, Jones N, Ayte J, Hidalgo E. 2003. Schizosaccharomyces pombe cells lacking the Ran-binding protein Hba1 show a multidrug resistance phenotype due to

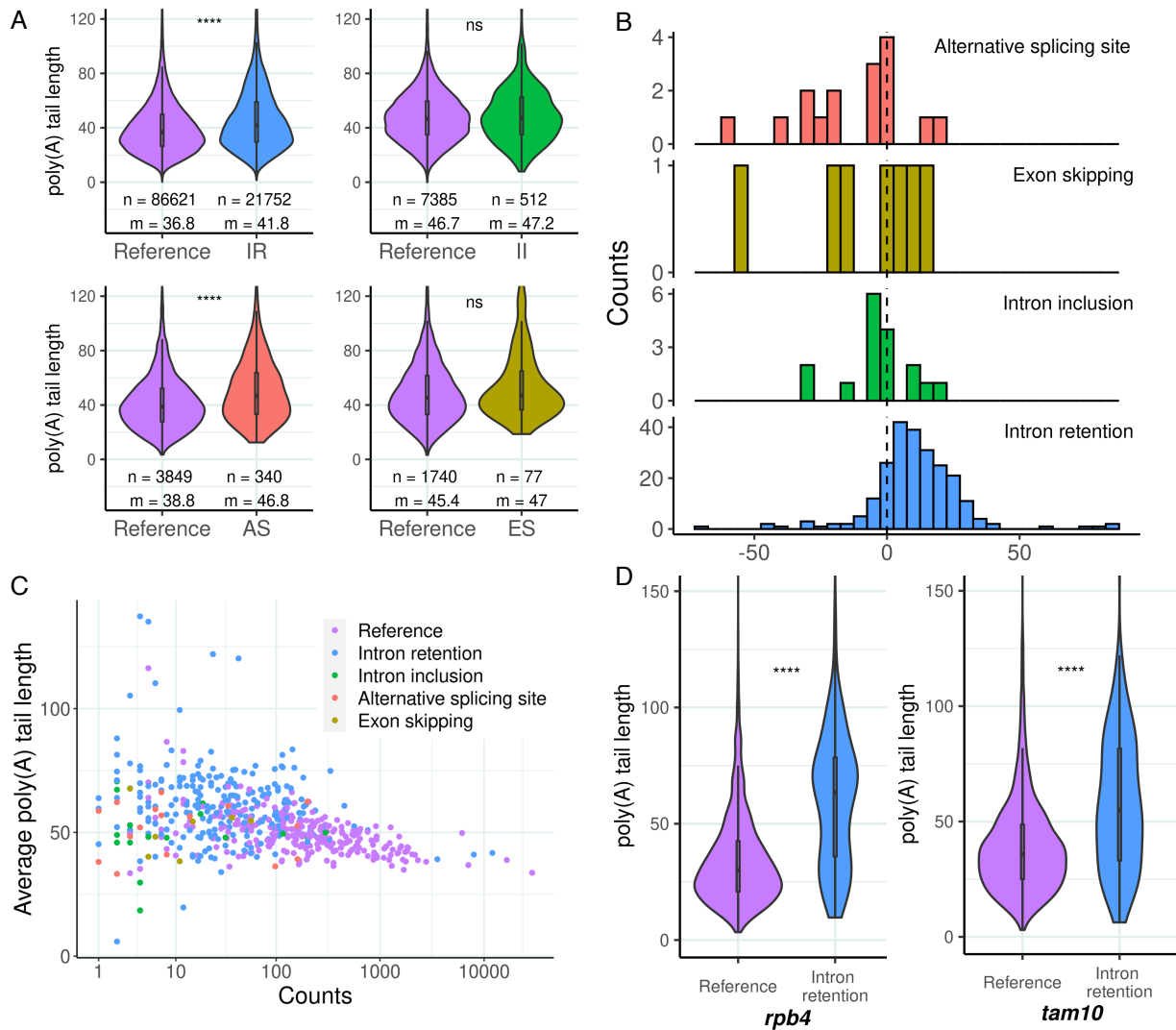
- constitutive nuclear accumulation of Pap1. *J Biol Chem* **278**: 40565–72.
- Chen J, Brunner A-D, Cogan JZ, Nuñez JK, Fields AP, Adamson B, Itzhak DN, Li JY, Mann M, Leonetti MD, et al. 2020. Pervasive functional translation of noncanonical human open reading frames. *Science (80-)* **367**: 1140–1146.
- Chitwood PJ, Hegde RS. 2020. An intramembrane chaperone complex facilitates membrane protein biogenesis. *Nature* **584**: 630–634.
- des Georges A, Katsuki M, Drummond DR, Osei M, Cross RA, Amos LA. 2008. Mal3, the *Schizosaccharomyces pombe* homolog of EB1, changes the microtubule lattice. *Nat Struct Mol Biol* **15**: 1102–8.
- Douka K, Birds I, Wang D, Kosteletos A, Clayton S, Byford A, Vasconcelos EJR, O’Connell MJ, Deuchars J, Whitehouse A, et al. 2021. Cytoplasmic long noncoding RNAs are differentially regulated and translated during human neuronal differentiation. *RNA* **27**: 1082–1101.
- Dreyfus M, Régnier P. 2002. The poly(A) tail of mRNAs: bodyguard in eukaryotes, scavenger in bacteria. *Cell* **111**: 611–3.
- Duncan CDS, Mata J. 2017. Effects of cycloheximide on the interpretation of ribosome profiling experiments in *Schizosaccharomyces pombe*. *Sci Rep* **7**: 10331.
- Duncan CDS, Mata J. 2014. The translational landscape of fission-yeast meiosis and sporulation. *Nat Struct Mol Biol* **21**: 641–647.
- Eichhorn SW, Subtelny AO, Kronja I, Kwasniewski JC, Orr-Weaver TL, Bartel DP. 2016. mRNA poly(A)-tail changes specified by deadenylation broadly reshape translation in *Drosophila* oocytes and early embryos. *Elife* **5**: e16955.
- Eickbush MT, Young JM, Zanders SE. 2019. Killer Meiotic Drive and Dynamic Evolution of the wtf Gene Family. *Mol Biol Evol* **36**: 1201–1214.
- Fair BJ, Pleiss JA. 2017. The power of fission: yeast as a tool for understanding complex splicing. *Curr Genet* **63**: 375–380.
- Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, Pantic N, Admassu T, James P, Warland A, et al. 2018. Highly parallel direct RNA sequencing on an array of nanopores. *Nat Methods* **15**: 201–206.
- Gonzalez-Hilarion S, Paulet D, Lee K-T, Hon C-C, Lechat P, Mogensen E, Moyrand F, Proux C, Barboux R, Bussotti G, et al. 2016. Intron retention-dependent gene regulation in *Cryptococcus neoformans*. *Sci Rep* **6**: 32252.
- Grech L, Jeffares DC, Sadée CY, Rodríguez-López M, Bitton DA, Hoti M, Biagosch C, Aravani D, Speekenbrink M, Illingworth CJR, et al. 2019. Fitness Landscape of the Fission Yeast Genome. *Mol Biol Evol* **36**: 1612–1623.

- Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**: 218–23.
- Ji Z, Song R, Regev A, Struhl K. 2015. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins ed. N. Sonenberg. *Elife* **4**: e08890.
- Kervestin S, Jacobson A. 2012. NMD: a multifaceted response to premature translational termination. *Nat Rev Mol Cell Biol* **13**: 700–712.
- Kim D-U, Hayles J, Kim D, Wood V, Park H-O, Won M, Yoo H-S, Duhig T, Nam M, Palmer G, et al. 2010. Analysis of a genome-wide set of gene deletions in the fission yeast *Schizosaccharomyces pombe*. *Nat Biotechnol* **28**: 617–623.
- Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusion. *Genome Biol* **14**: R36.
- Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**: 907–915.
- Kovaka S, Zimin A V., Pertea GM, Razaghi R, Salzberg SL, Pertea M. 2019. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol* **20**: 278.
- Kuang Z, Boeke JD, Canzar S. 2017. The dynamic landscape of fission yeast meiosis alternative-splice isoforms. *Genome Res* **27**: 145–156.
- Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, and 1000 Genome Project Data Processing Subgroup. 2009. The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* **25**: 2078-2079.
- Li R, Ren X, Ding Q, Bi Y, Xie D, Zhao Z. 2020. Direct full-length RNA sequencing reveals unexpected transcriptome complexity during *Caenorhabditis elegans* development. *Genome Res* **30**: 287–298.
- Lima SA, Chipman LB, Nicholson AL, Chen Y-H, Yee BA, Yeo GW, Collier J, Pasquinelli AE. 2017. Short poly(A) tails are a conserved feature of highly expressed genes. *Nat Struct Mol Biol* **24**: 1057–1063.
- Lock A, Rutherford K, Harris MA, Hayles J, Oliver SG, Bähler J, Wood V. 2019. PomBase 2018: user-driven reimplementations of the fission yeast database provides rapid and intuitive access to diverse, interconnected information. *Nucleic Acids Res* **47**: D821–D827.
- Malapeira J, Moldón A, Hidalgo E, Smith GR, Nurse P, Ayté J. 2005. A meiosis-specific cyclin regulated by splicing is required for proper progression through meiosis. *Mol Cell Biol* **25**: 6330–7.
- Moldón A, Malapeira J, Gabrielli N, Gogol M, Gómez-Escoda B, Ivanova T, Seidel C, Ayté J.

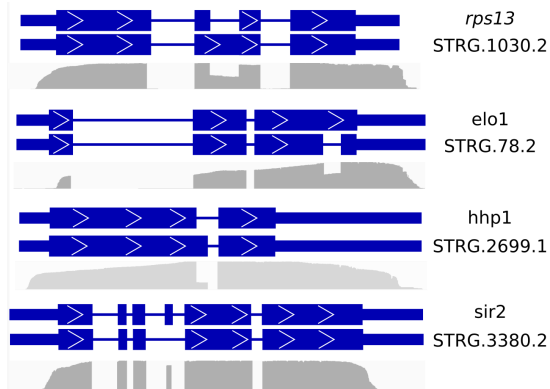
2008. Promoter-driven splicing regulation in fission yeast. *Nature* **455**: 997–1000.
- Moro SG, Hermans C, Ruiz-Orera J, Albà MM. 2021. Impact of uORFs in mediating regulation of translation in stress conditions. *BMC Mol cell Biol* **22**: 29.
- Parker R, Song H. 2004. The enzymes and control of eukaryotic mRNA turnover. *Nat Struct Mol Biol* **11**: 121–127.
- Preiss T, Muckenthaler M, Hentze MW. 1998. Poly(A)-tail-promoted translation in yeast: implications for translational control. *RNA* **4**: 1321–31.
- R Core Team. R Foundation for Statistical Computing, Vienna AU <https://www.R.org>. 2020. R: a language and environment for statistical computing.
- Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.
- Reixachs-Solé M, Ruiz-Orera J, Albà MM, Eyra E. 2019. Ribosome profiling at isoform level reveals an evolutionary conserved impact of differential splicing on the proteome. *Nat Commun* **11**: 1768.
- Richter JD. 2000. Translational Control of Gene Expression. In (eds. N. Sonenberg, J. Hershey, and M. MB), pp. 785–805, Cold Spring Harbor Laboratory Press, New York.
- Roach NP, Sadowski N, Alessi AF, Timp W, Taylor J, Kim JK. 2020. The full-length transcriptome of *C. elegans* using direct RNA sequencing. *Genome Res* **30**: 299–312.
- Ruiz-Orera J, Albà MM. 2019a. Conserved regions in long non-coding RNAs contain abundant translation and protein-RNA interaction signatures. *NAR genomics Bioinforma* **1**: e2.
- Ruiz-Orera J, Albà MM. 2019b. Translation of Small Open Reading Frames: Roles in Regulation and Evolutionary Innovation. *Trends Genet* **35**: 186–198.
- Ruiz-Orera J, Messegue X, Subirana JA, Alba MM. 2014. Long non-coding RNAs as a source of new peptides. *Elife* **3**: e03523.
- Sakurai H, Mitsuzawa H, Kimura M, Ishihama A. 1999. The Rpb4 subunit of fission yeast *Schizosaccharomyces pombe* RNA polymerase II is essential for cell viability and similar in structure to the corresponding subunits of higher eukaryotes. *Mol Cell Biol* **19**: 7511–8.
- Snowdon C, Schierholtz R, Poliszczuk P, Hughes S, van der Merwe G. 2009. ETP1/YHL010c is a novel gene needed for the adaptation of *Saccharomyces cerevisiae* to ethanol. *FEMS Yeast Res* **9**: 372–80.
- Stepankiw N, Raghavan M, Fogarty EA, Grimson A, Pleiss JA. 2015. Widespread alternative and aberrant splicing revealed by lariat sequencing. *Nucleic Acids Res* **43**: 8488–501.
- Subtelny AO, Eichhorn SW, Chen GR, Sive H, Bartel DP. 2014. Poly(A)-tail profiling reveals an embryonic switch in translational control. *Nature* **508**: 66–71.
- Tsankov AM, Thompson DA, Socha A, Regev A, Rando OJ. 2010. The role of nucleosome

- p>positioning in the evolution of gene regulation.
- PLoS Biol*
- 8**
- .
- Tudek A, Krawczyk PS, Mroczek S, Tomecki R, Turtola M, Matylla-Kulińska K, Jensen TH, Dziembowski A. 2021. Global view on the metabolism of RNA poly(A) tails in yeast *Saccharomyces cerevisiae*. *Nat Commun* **12**: 4951.
- Ukleja M, Cuellar J, Siwaszek A, Kasprzak JM, Czarnocki-Cieciura M, Bujnicki JM, Dziembowski A, Valpuesta JM. 2016. The architecture of the *Schizosaccharomyces pombe* CCR4-NOT complex. *Nat Commun* **7**: 10433.
- Ullah F, Hamilton M, Reddy ASN, Ben-Hur A. 2018. Exploring the relationship between intron retention and chromatin accessibility in plants. *BMC Genomics* **19**: 21.
- Virčíková V, Pokorná L, Tahotná D, Džugasová V, Balážová M, Griač P. 2018. *Schizosaccharomyces pombe* cardiolipin synthase is part of a mitochondrial fusion protein regulated by intron retention. *Biochim Biophys Acta Mol cell Biol lipids* **1863**: 1331–1344.
- Wang JR, Holt J, McMillan L, Jones CD. 2018. FMLRC: Hybrid long read error correction using an FM-index. *BMC Bioinformatics* **19**: 50.
- Wickham H. 2016. *ggplot2: Elegant graphics for Data Analysis*. Springer-Verlag New York Inc.
- Wood V, Gwilliam R, Rajandream M-A, Lyne M, Lyne R, Stewart A, Sgouros J, Peat N, Hayles J, Baker S, et al. 2002. The genome sequence of *Schizosaccharomyces pombe*. *Nature* **415**: 871–80.
- Workman RE, Tang AD, Tang PS, Jain M, Tyson JR, Razaghi R, Zuzarte PC, Gilpatrick T, Payne A, Quick J, et al. 2019. Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat Methods* **16**: 1297–1305.
- Wyman D, Mortazavi A. 2019. TranscriptClean: variant-aware correction of indels, mismatches and splice junctions in long-read transcripts ed. B. Berger. *Bioinformatics* **35**: 340–342.
- Yan C, Hang J, Wan R, Huang M, Wong CCL, Shi Y. 2015. Structure of a yeast spliceosome at 3.6-angstrom resolution. *Science* **349**: 1182–91.
- Yeom K-H, Pan Z, Lin C-H, Lim HY, Xiao W, Xing Y, Black DL. 2021. Tracking pre-mRNA maturation across subcellular compartments identifies developmental gene regulation through intron retention and nuclear anchoring. *Genome Res* **31**: 1106–1119.
- Zhang S, Li R, Zhang L, Chen S, Xie M, Yang L, Xia Y, Foyer CH, Zhao Z, Lam H-M. 2020. New insights into *Arabidopsis* transcriptome complexity revealed by direct sequencing of native RNAs. *Nucleic Acids Res* **48**: 7700–7711.

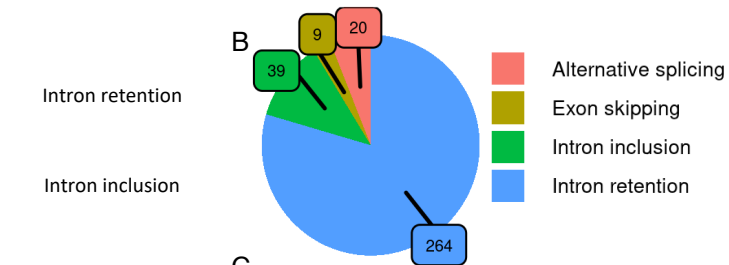




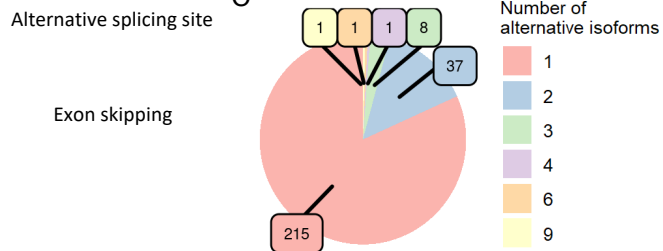
A



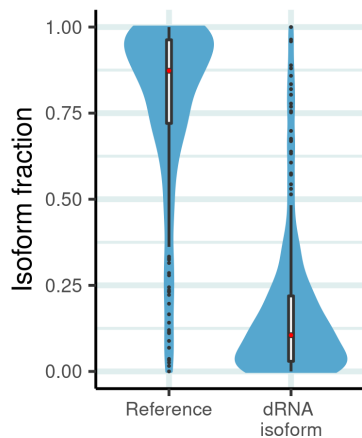
B



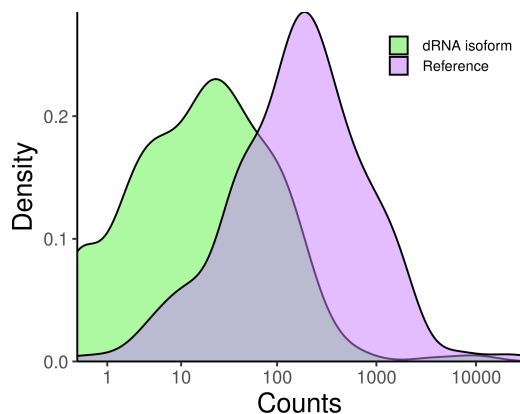
C



D



E



F

