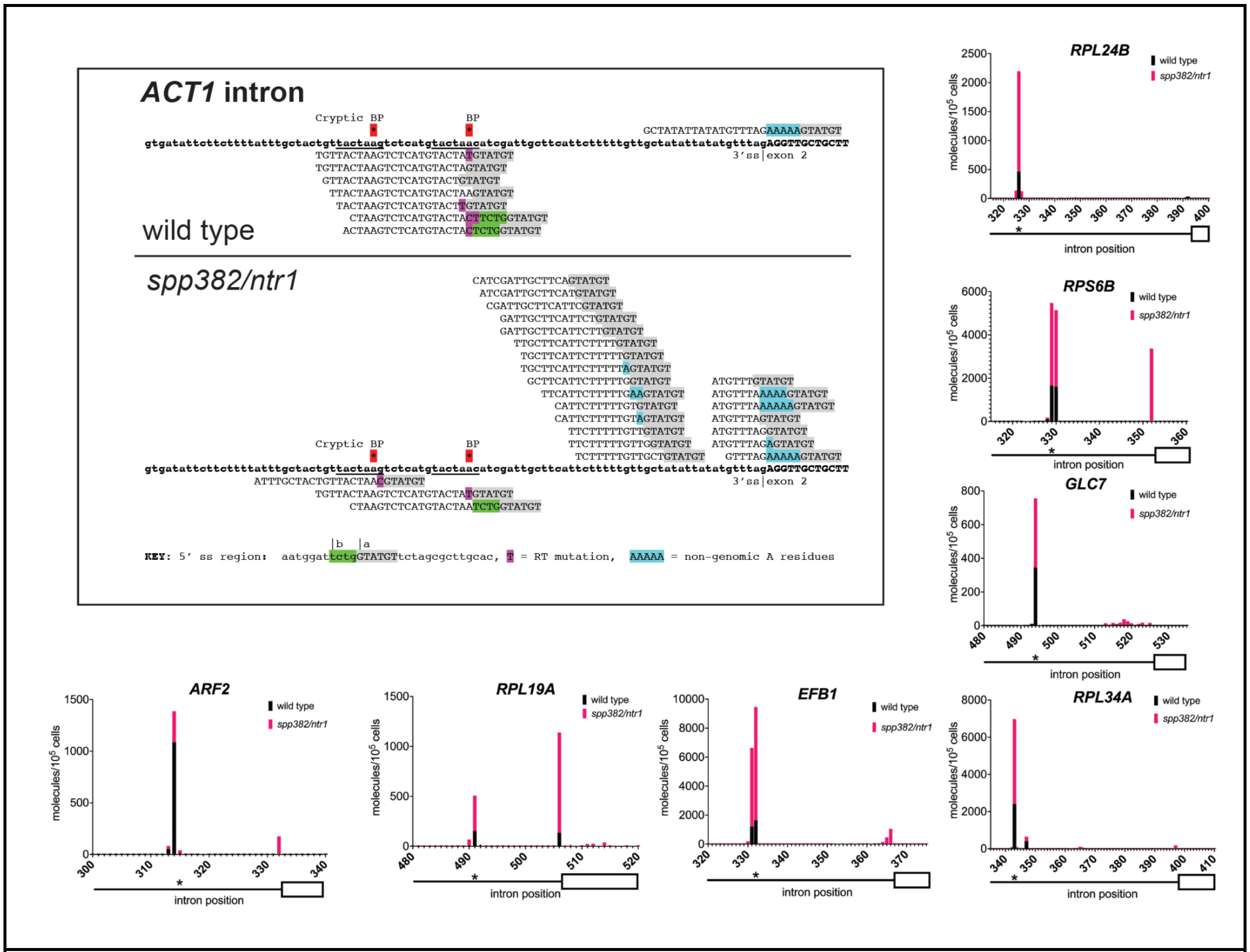
Supplemental Materials for:

Intron-lariat spliceosomes convert lariats to true circles: implications for intron transposition. Manuel Ares, Jr., Haller Igel, Sol Katzman, John P. Donohue

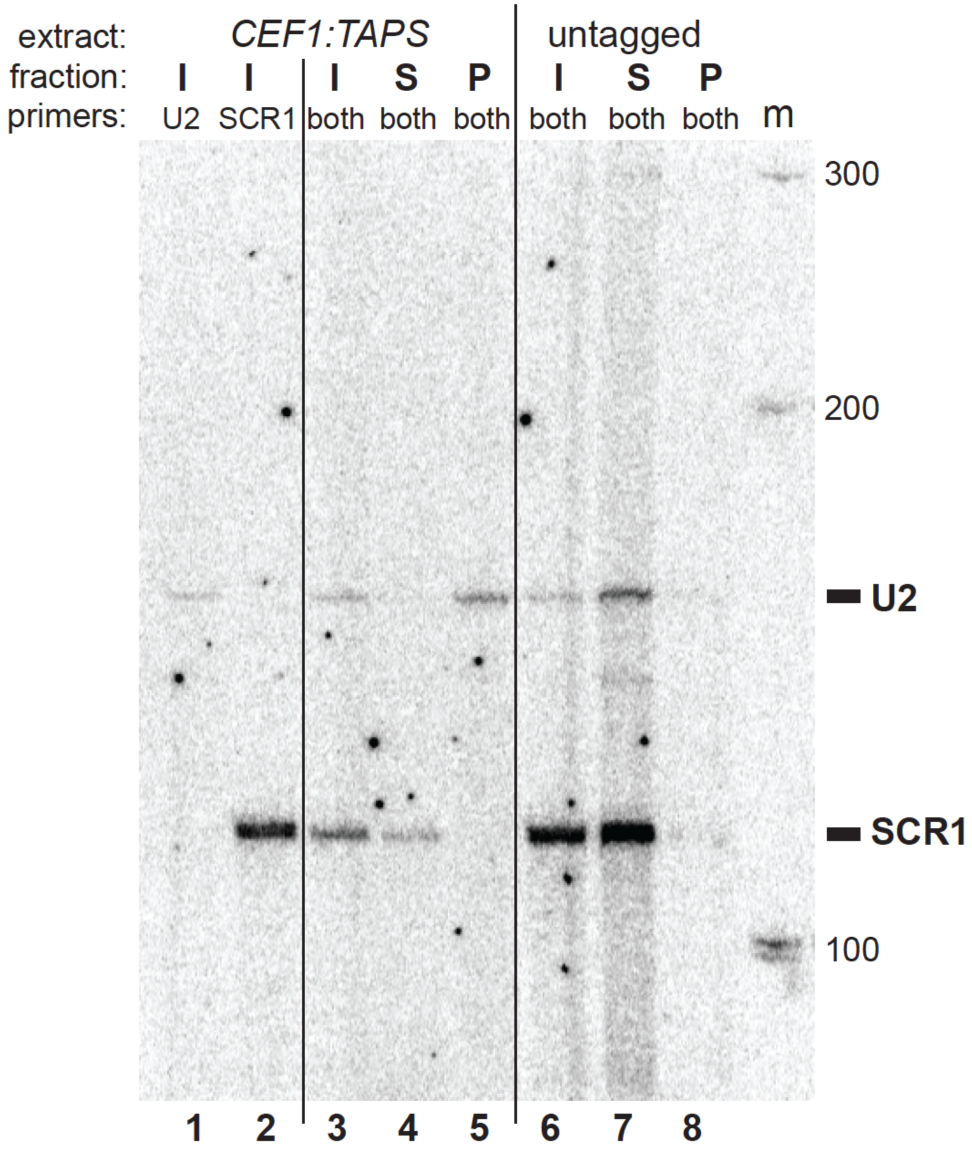
Supplemental Figures

Supplemental Figure S1



Supplemental Figure S1. Detection of intron circles for additional introns of *S. cerevisiae*. Upper left panel shows circular intron amplicon reads mapping to the *ACT1* intron from wild type (above the line) and the *spp382-1* mutant (below the line). The sequence of the intron is shown in bold. Circle reads are aligned above and lariat reads below the intron sequence with the 5'ss sequence highlighted in gray and non-genomic As highlighted in blue. A class of lariats derived from incorrect 5'ss selection are shown in green. RT errors at the branch are highlighted in purple, see key at bottom. The small graphs show distinct distributions of processed and unprocessed intron circles from different introns. Graphs show the junction locations (x-axis) and number of unique reads with junctions at that location (y-axis, calibrated to a spiked-in circular RNA) per 10^5 yeast cells, from wild type (black bars) or *spp382-1* (pink bars) cells. The 3' part of each intron (line) and its second exon (white box) are shown below the x-axis, with the asterisk indicating the position of the bp.

Supplemental Figure S2



Supplemental Figure S2. Partial purification of spliceosomes by IgG-sepharose chromatography of a CEF1-tagged extract. The *spp382-1* mutant was fitted with a TAPS tag at the C-terminus of its *CEF1* coding region for affinity purification of spliceosomes. Extracts from this strain and an untagged control were prepared (Input, I) and bound to IgG-sepharose. Unbound (Supernatant, S) and bound (Pellet, P) fractions were prepared. RNA from each fraction was used as a template for primer extension with a ^{32}P-labeled oligonucleotide complementary to U2 snRNA (lanes 1, 3-8) and SCR1 (lanes 2-8). Labels on the right indicate the expected migration of the cDNA from each primer, 152 nt U2 cDNA primed by oligo 23T, and a 113 nt SCR1 cDNA primed by oligo SCR1-110 (Table S7). The pellet fraction from the tagged extract (lane 5), but not the untagged extract (lane 8) is enriched in the spliceosomal U2 snRNA, free of the cytoplasmic RNA SCR1, which remains in the supernatant fraction of both extracts (lanes 4 and 7).

Supplemental Tables

Supplemental Table S1: Data for Fig1, Fig S1

Supplemental Table S2: Data for Fig 2C

Supplemental Table S3: Data for Fig 3B, C

Supplemental Table S4: Data for Fig 4

Supplemental Table S5: Data for Fig 5

Supplemental Table S6: Data for Fig 6

Supplemental Table S7: Yeast strains, Plasmids, Oligonucleotides

Supplemental Table S8: Mapping parameters

Supplemental Table S9: Mapping data for Trial 1

Supplemental Table S10: Mapping data for Trial 2

Supplemental Table S11: Mapping data for Trial 3

Supplemental Methods

Experimental Trials and Library Details

Trial 1 consisted of 12 libraries designed to test the amplicon method for several introns and to ask about two mutations in spliceosome disassembly proteins (*spp381-1* and *drn1Δ*) and the nuclear 3' processing machinery (*trf4Δ*, *trf5Δ*, and *rrp6Δ*) using total log phase RNA from wild type and the mutants. The first set of 6 libraries (1.1-1.6) created amplicons for 7 introns (*ACT1*, *EFB1*, *GLC7*, *RPL19A*, *RPL24B*, *RPL34A*, *RPP1B*) in wild type (1.1); *spp382-1* (1.2); *trf4Δ* (1.3); *trf5Δ* (1.4); *rrp6Δ* (1.5); *drn1Δ* (1.6). The second set of libraries (2.1-2.6) created amplicons for 3 introns: *ACT1*, *RPL17B*, *RPL24B* in wild type (2.1); *spp382-1* (2.2); *trf4Δ* (2.3); *trf5Δ* (2.4); *rrp6Δ* (2.5); *drn1Δ* (2.6). The libraries were quantified by qPCR using the Illumina i5 and i7 primers, pooled and sequenced by MA on a MiSeq in Doug Black's lab at UCLA.

Trial 2 consisted of 12 libraries designed to test the effect of Dbr1 and RNase R treatment and to replicate experiments with certain mutants, and to assess circles in stationary phase. This trial also included a linear control amplicon from the *TUB3* intron. Both sets of libraries in this trial employed primers for the introns in *ACT1*, *ARF2*, *ECM33*, *EFB1*, *GLC7*, *MPT5*, *RPL19A*, *RPL24B*, *RPL34A*, *RPP1B*, *RPS6B*, and *TUB3*. The first set of 6 libraries (3.1-3.6) used rRNA depleted RNA not treated with enzyme: wild type log phase RNA (3.1); *spp382-1* log phase (3.2); *trf4Δ* log phase (3.3); *rrp6Δ* log phase (3.4); wild type stationary phase (3.5); *spp382-1* stationary phase (3.6). The second set (3.7-3.12) are derived from rRNA depleted log phase RNA treated with enzymes: wild type, RNase R (3.7); *spp382-1*, RNase R (3.8); *dbr1Δ*, RNase R (3.9); wild type, Dbr1 (3.10); *spp382-1*, Dbr1 (3.11); *dbr1Δ*, Dbr1 (3.12). These libraries were sequenced on a NextSeq500 by the staff at the UCSC Paleogenomics Lab.

Trial 3 consisted of 16 libraries designed to quantify the numbers of circles per cell using a spiked in circle, and to extend the test of disassembly factor mutants in altering circle numbers as well as to assess the dependence of circle formation on the tRNA ligase Trl1. RNAs were spiked with an in vitro synthesized 812 nt circular RNA (cGFP) to achieve 1 cGFP molecule per 100 cell equivalents of RNA as follows. Cells were counted at harvest and yields agreed with expected total RNA amounts per cell. A yeast cell contains about 0.7 pg of RNA () so that 10 ug of total RNA arises from 1.43 x 10^7 cells. At 1 cGFP per 100 cells this requires 1.43 x 10^5 cGFP molecules. The molecular weight of an 812 nt circular RNA is 2.7 x 10^5 daltons. Our gel purified cGFP stock was 58.1 ng/ul or 215 nM, or about 1.3 x 10^{11} cGFP molecules/ul. We prepared two 1 ml aliquots of water with 8 ug of HEK293 cell total RNA as a carrier and transferred 1 ul of the cGFP stock (1.3 x 10^{11} molecules) to the first ml, and took 1 ul from that (1.3 x 10^8 molecules) and diluted it into the second ml of 8 ug/ml HEK cell total RNA to reach a final concentration of 1.3 x 10^5 cGFP molecules per ul. We then added 1.1 ul of this second dilution (1.43 x 10^5 cGFP molecules) to each 10 ug of total yeast RNA (from 1.43 x 10^7 cells) to be used for library creation. All samples were then treated with RNase R as described below. [Note that comparison of untreated vs RNase R treated libraries (Trial 2, Fig 3C) shows that RNase R reduces circular read counts and thus the experiments in Trial 3 may underestimate the true number of circles.] The libraries in this trial employed primers for the introns in *ACT1*, *ARF2*, *ECM33*, *EFB1*, *GLC7*, *MPT5*, *RPL19A*, *RPL24B*, *RPL34A*, *RPP1B*, *RPS6B*, *TUB3*, and the synthetic spiked in circle cGFP. The first set (4.1-4.8) consisted of two wild type replicates 1 and 2 (4.1, 4.2); two *spp382-1* replicates 1 and 2 (4.3, 4.4); two *rrp6Δ* replicates 1 and 2 (4.5, 4.6); and two *trf4Δ* replicates 2 and 3 (4.7, 4.8). The second set (4.9-4.14) consisted of pairs of strains, one wild type control and a mutant: *PRP43* wild type control (4.9); *prp43-Q423N* (4.10); U6 wild type control (4.11); U6*Δ*5 (4.12); *TRL1* wild type control (4.13); *trl1Δ* + *Δ*intron-tRNAs plasmid (4.14). The third set replaced Superscript IV with TIGRT in the first library step and compared wild type (4.15); with *dbr1Δ* (4.16). These libraries were sequenced on a MiSeq by the staff at the UCSC Paleogenomics Lab.

Raw data from all libraries is available in the Short Read Archive under the accession number PRJNA739208. Data in Fig. 1 and Supplemental Fig. S1 came from Trial 3 libraries 4.1-4.4, except for the *ACT1* data in Supplemental Fig. S1, which came from Trial 1 libraries 2.1 and 2.2. Data from Fig. 2C came from Trial 3 libraries 4.13 and 4.14. Data from Fig. 3B and C came from Trial 2 libraries 3.2, 3.8, and 3.11. Data from Fig. 4 are from Trial 3 libraries 4.1, 4.4, and 4.9-4.12. Fig. 5 data comes from Trial 3 libraries 4.1-4.8.

Detailed description of circular intron amplicon read processing

The following process was applied separately to each of the samples. Table S8 lists the parameters specific to each sample.

STEP1: UMI QUALITY FILTERING

The paired-end fastq files were first filtered for high Illumina quality value in all of the positions of the sequenced reads corresponding to the unique molecular index (UMI). This included 4 bases in read1 and 4 bases in read2. The minimum quality value (minQV) was chosen based on the distribution of quality values in those bases, which is dependent on the sequencing platform. The most common quality value was consistent in those bases for each sample and was above minQV. This step was taken to ensure that when duplicate reads with the same UMIs were removed in a later step, the UMI bases were accurate. This minimizes the possibility that a true molecular duplicate would appear to have a different UMI and thus be erroneously retained. All subsequent steps are performed using these minQV fastq files.

STEP2A: BOWTIE MAPPING FILTERING

For each intron and the circular cGFP spike-in, the paired-end fastq files were filtered to create a set of intron-specific fastq files using a procedure based on bowtie mapping as follows. A fasta file was created using the full length intronic sequences from the sacCer3 assembly (see Table S8). Each sequence corresponds to the rna-coding strand beginning at the 5' splice site (GT...) and ending at the 3' splice site (...AG). A bowtie 2 (Langmead and Salzberg 2012) target was created using this fasta file. For the purposes of mapping, the paired-end 76x76bp minQV fastq files were trimmed to paired-end 35x24bp by removing the first 4 UMI bases and the last 37 or 48 bases from read1 or read2, respectively. These trimmed fastq files were mapped to the bowtie2 target using the bowtie2 parameters "--end-to-end --sensitive --fr" (all others default values). The intent of this mapping is to filter the sequenced fragments to those having the expected primers and subsequent target bases in the introns (or cGFP spike-in). Because the primers are not in the standard orientation for bowtie mapping, and to eliminate spurious hybridization products, the bowtie mappings were further filtered to generate a set of read-IDs as follows. For the circular intron (and cGFP) targets, the bowtie SAM-format output was filtered with samtools (https://github.com/samtools/hts-specs) and the unix tool "awk", based on the expected strand of the read1/read2 primers, and the expected distance between them in the intron.

To extract the read1 mapping IDs for one intron (a "reference" in SAM nomenclature) this sequence of commands was used:

```
samtools view -S  -f 0x61 -F 0xf0e sample.sam |
awk '($3 == ITN && $7 == "=" && $9 < -100){print $1}' ITN=intronRef |
sort > intronRef.read1.okay.ids
```

To extract the read2 mapping IDs:

```
samtools view -S  -f 0x91 -F 0xf0e sample.sam |
awk '($3 == ITN && $7 == "=" && $9 >  100){print $1}' ITN=intronRef |
sort > intronRef.read2.okay.ids
```

The "-f" and "-F" options together specify: read1 or read2, its strand, the fact that both reads of the pair are mapped, and that this is NOT a "proper pair" by the standard bowtie2 determination. The third field ($3) of the SAM file is selected for the particular reference, and the seventh field ($7) specifies that the other read maps to the same reference. The sign and magnitude of the ninth field ($9) indicates the minimum required distance between the read locations in the intron as well as their relative positions in the intron. The first field ($1) is the desired read-ID. The 2 files of read-IDs were verified as identical, and either can be used in the next step. For the non-circular *TUB3* intron target, slightly different parameters were used because of the expected position and orientation of its primers.

For read1 of the linear *TUB3* intron target:

```
samtools view -S  -f 0x63 -F 0xf0c sample.sam |
awk '($3 == ITN && $7 == "=" && $9 >  100){print $1}' ITN=intronRef | \
sort > intronRef.read1.okay.ids
```

For read2 of the linear *TUB3* intron target:

```
samtools view -S  -f 0x93 -F 0xf0c - sample.sam | \
awk '($3 == ITN && $7 == "=" && $9 < -100){print $1}' ITN=intronRef | \
sort > intronRef.read2.okay.ids
```

Using the set of read-IDs for a given intron (whether circular or linear target), the original full-length paired-end fastq files were filtered to create a set of intron-specific fastq files. All subsequent steps are performed using these intron-specific fastq files.

STEP2B: EXPECTED SEQUENCE FILTERING

As indicated in Table S8, a specific range of bases from read2 in the fastq files (filtered as described in the previous step) were extracted. Only those fastq entries that had an exact match to the intron-specific expected sequence string were retained for the following analysis step. The range and expected string for this step were chosen based on the intronic bases adjacent to the 5' splice site of the intron (or the expected circularization point of the cGFP spike-in).

STEP3: SEQUENCE ANALYSIS INCLUDING DUPLICATE REMOVAL

As indicated in Table S8, a specific range of bases from read2 were analyzed for each intron. This range was chosen based on the expected position in read2 of the 5' splice site (5'ss) of the intron, relative to one of the primers used to extract the RNA fragment sequenced. In particular, the analysis bases were chosen to overlap the 5'ss position and include 6 expected intronic bases upstream and 14 unknown bases from the downstream part of the intron. The downstream bases should differ depending on whether the sequenced fragment was from a lariat, an RNA circle, or something else. The upstream bases might also differ from the expected string if the spliceosome used an incorrect 5'ss upstream of the expected 5'ss, but incorrect downstream 5'ss usage would be filtered out in the previous sequencing filtering step. To remove duplicates from the set of analysis strings, the following algorithm was used. First, the analysis string was extracted for all paired-end reads into a file "analysis.all". Similarly, the UMI bases from read1 and read2 were extracted into files "umird1.all", "umird2.all". To collapse duplicate strings having the same UMI, the following unix string of piped commands was executed:

```
paste analysis.all umird1.all umird2.all | sort | uniq | cut -f 1 | sort |
uniq -c | sort -k 1,1nr > analysis_dupremoved.sorted
```

The "paste" command associates each occurrence of a string with its UMI. The initial "sort | uniq" command removes all duplicates, leaving only one copy of a string-umi combination. The "cut" command strips off the UMI, leaving the dup-removed analysis strings. The next "sort | uniq -c" command counts occurrences of each distinct dup-removed string. The final "sort" command arranges these in decreasing order of counts. These counts and their sum are combined to generate the fraction and cumulative fraction for these distinct dup-removed strings, which can also be presented in reverse-complement form for easier downstream analysis. Each string represents a distinct 5'ss-3' intron end junction including the 6 bases of the 5'ss, and the counts of each represent their abundance in the sample.

STEP 4: MAPPING THE JUNCTION ENDS OF EACH UNIQUE CIRCLE

To identify the circle junction, we used BLAT (Kent 2002). The last 6 nt of the string were removed and each remaining sequence was used to query the intron sequences (except GLC7, which was done separately due to low information content segments of its intron). BLAT options used were:

```
 -minIdentity=50 -noTrimA -tileSize=6 -oneOff=1 -minScore=15  -out=pslx
```

From the pslx file output, the BLAT score (number of matching nucleotides), the position in the intron sequence of the end of the match (this is the nucleotide joined to the 5'ss), and the match sequence were reported along with the query, its number of counts and its rank. These values are reported in the Supplemental Tables S9-S11. Classification and counting of junctions as lariats or circles was done using the nucleotide positions specific for each intron as described in Supplemental Table S1.

References

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12**: 656–664.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359.