Supplementary Methods

Meiosis time courses

Meiotic time courses were performed as described by Berchowitz *et al* (Berchowitz et al. 2013). Briefly, SK1 yeast cells were grown to saturation in YPD, diluted in BYTA medium to an OD_{600} = 0.25, and grown for 20 hours to reach a G1 phase state. These cells were washed once with water, then resuspended in sporulation medium to an OD_{600} = 1.9. Cells were allowed to proceed through meiosis for 6 hours at 30°C. During this time, cells accumulated at a meiotic prophase arrest due to the *NDT80* transcription factor being under an inducible promoter (Benjamin et al. 2003). After harvesting samples for the 6-hour time point, cells were released from the prophase arrest by the induction of *NDT80* with 1uM β-estradiol, allowing synchronous progression through the meiotic divisions (Carlile and Amon 2008). RNA and immunofluorescence samples were harvested in parallel at the indicated time points. For the RNA samples, 2mL of cells were pelleted, flash frozen in liquid nitrogen, and stored at -80°C until further processing. Tubulin immunofluorescence was performed as described by Berchowitz *et al*. We also observe in the *dbr1Δ* strain, introns that are abundant in the rRNA subtraction are depleted in the poly(A) selection, explaining discrepancies between these samples in the clustering since reads inside the intron specifically identify the retained intron isoform instead of the spliced isoform.

RNA isolation

RNA isolation for Lariat-seq and Branch-seq was performed as follows. Yeast were grown to OD_{600} 0.94-0.98 and were collected by centrifugation at 7000 RPM for 5 min at 4°C. Media was poured off and yeast were washed twice in water and frozen at -80°C. Cells were thawed and transferred to tubes containing 2.8mm ceramic beads and 1mL Trizol (Life Technologies) was added to 1/10 cell pellet. An Omni Bead Ruptor was used to lyse the cells, twice for 20 seconds on ½ max speed and once for 10 seconds on max speed. Samples were incubated at room temp for 5 min, 1/5 volume of chloroform was added and

mixed, samples incubated at room temp 2-3 min and were spun at max speed for 15 min at 4°C. The upper aqueous layer was transferred to a new tube and precipitated with ½ volume isopropanol. After 5min on ice, samples were spun at max speed, 4°C for 25 min. The RNA pellet was washed with 70% ethanol before storage at -80°C.

RNA isolation for RNA-seq was performed as follows. Overnight yeast cultures were grown in 5mL YPD media and were diluted in the morning into 50mL YPD and grown to log phase (OD_{600} 0.5 to 1), spun down, and the pellets were frozen in liquid nitrogen. RNA was isolated as in (Clarkson et al. 2010). Pellets were resuspended in 1mL Acid Phenol and an equal volume of AES buffer (50mM NaAcetate pH 5.2, 10mM EDTA, 1% SDS) was added. In 2mL Eppendorf tubes, samples were incubated at 65°C for 10 min with vortexing every minute. Samples were incubated on ice for 5 min and then transferred to a phaselock tube and one volume chloroform was added. After spinning, the top aqueous layer was transferred to a fresh phaselock tube and one volume of phenol:chloroform:isoamyl alcohol (25:24:1) was added, tubes were spun, one volume of chloroform was added, tubes were spun, and the aqueous layer was transferred to a fresh tube to be precipitated with 50uL 3M NaOAc (pH 5.5) and 550uL isopropanol. Samples were spun at max speed for 25 minutes at 4°C. The pellet was washed twice with 70% ethanol and resuspended in water.

Isolation of in vitro-spliced *Drosophila melanogaster* lariat RNA

Radio labeled *FTZ* lariat RNA was used to assess the fidelity of each step of the Branch-seq protocol during protocol development. The *FTZ* lariat RNA was included as a positive control spike-in during all subsequent Branch-seq experiments to ensure successful debranching. *FTZ* lariat RNA was generated using Hela nuclear extracts for in-vitro splicing. Hela nuclear extracts were a kind gift from the Reed Lab (Folco et al. 2012). Coupled in vitro transcription and splicing were performed similar to Folco and Reed (Folco and Reed 2014) except without addition of α-amanitin to obtain as many lariats as possible. Reactions were digested with RNase R (Epicenter) at 37°C for 1 hour to obtain radio labeled *FTZ* lariats.

Debranching enzyme purification

S. cer. DBR1 cDNA was generated from WT S288C yeast and cloned into the pET151 expression vector from Invitrogen. Protein was expressed in Rosetta 2(DE3)pLysS competent cells grown in YT media at 37°C until they were induced with IPTG and grown at 18°C. Bacteria were lysed using Native Lysis Buffer (Qiagen). Protein was purified with a Ni-NTA column (Qiagen) and subsequently over an S200 column (Buffer: 125 mM KCL, 20mM HEPES pH 7.3, 1mM DTT, 10% glycerol). Protein was concentrated (final 50% glycerol) and flash frozen. Protein was tested for RNase activity and debranching activity on linear RNA and an in vitro spliced lariat, respectively.

Reverse transcription

Reverse transcription was performed using primer /5Phos/AGATCGGAAGAGCGTCGTGTAGGGAAAG/iSp18/CACTCA/iSp18/GTGACTGGAGTTCCTTGGCACCCGAGAATTCCA/TTTTTTTTTTTTTTTTTTTTVN (designed in collaboration with Yarden Katz (Katz et al. 2014)) incubated with SuperScriptIII RT (Invitrogen) for 30 min at 48°C. Subsequently 2.1 uL of 1M NaOH was added and samples were incubated at 98°C for 15 min. The RT primer is a modified version if the ribosome footprint profiling RT primer where the 5' end of the RNA gets sequenced first and paired end, barcoded sequencing is possible (Ingolia et al. 2009).

The samples were then run on a 6% TBE-urea gels (Invitrogen) for 93 min at 200V to remove excess RT primer. Gels were stained with SYBR gold and gel slices were excised where product was observed to run above the RT primer for the top, middle, and bottom lariat samples. Gel slices were shredded and DNA was eluted in 400uL PAGE elution buffer overnight (see 2D gels Methods). Gel was removed before precipitation using a Nanosep column.

Circularization

Circligase (Epicentre) was used to circularize the gel isolated RT products for 1 hour at 60°C and the enzyme was inactivated by heating at 80°C for 10 minutes.

PCR

Phusion high-fidelity polymerase (NEB) was used to amplify the circularized products. Illumina PCR primer 1.0
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
was paired with Illumnia barcode primers (RPI#s)
(RPI1) CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA
(RPI2) CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA
(RPI3) CAAGCAGAAGACGGCATACGAGATGCCTAAGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA
(RPI4) CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA
Samples were removed after 6, 8, 10 and 12 PCR cycles and run on an 8% TBE gel (Invitrogen) for 40 min at 200V. PCR products were gel isolated by shredding the gel through a hole poked with a needle in the bottom of a 0.5 mL Eppendorf tube and eluted in 400uL PAGE elution puffer (see above) at 65°C , shaking at 1400RPM for one hour. Gel was removed with a Nanosep column and precipitated with isopropanol.

http://epigenome.usc.edu/docs/resources/core_protocols/Illumina%20Sequence%20Information%20for%20Customers%20DEC2008.pdf

Sequencing

For Branch-seq, one Illumina MiSeq flow cell was sequenced at the MIT Bio Micro Center (November 2011). 5' end reads were 50 bases and 3' end reads were 250 bases. 3' end reads were sequenced with custom sequencing primer GTGACTGGAGTTCCTTGGCACCCGAGAATTCCATTTTTTTTTTTTTTTTTTT to avoid sequencing the untemplated As added by the poly(A) tailing reaction. The 3' end sequencing primer was gel purified prior to use in sequencing (primer design might need to be changed for sequencing on other Illumina machines).

Lariat-seq library prep

Reverse transcription was performed on 2D gel isolated lariat RNA using 1ul Random hexamer Primers (3ug/ul) (Invitrogen) and SuperScript III reverse transcriptase (Invitrogen). RNA and primer mix was heated at 70°C for 10 minutes and then put on ice.

12 uL of Mix A (mix A: 4uL 5x 1st strand buffer, 2uL 100mM DTT, 1uL dNTPs (10mM), 4uL Actinomycin D [1mg/1mL], 1uL SuperaseIn (20U/ul)) was added to the RNA and primer. Then 1 uL of SSIII was added and the RT program was run: 25^{0}C 10 minutes, 42^{0}C 50 minutes, 70^{0}C 15 minutes, 4^{0}C hold. Sample volume was brought up to 200uL with water and then samples were phenol choloroform extracted and ethanol precipitated. Second strand synthesis was performed with DNA pol I and dUTP to make strand specific libraries. Next the samples underwent SPRI-TE (end repair, adenlyation, adapter ligation, gel purification #1). Subsequently uracil digestion was performed with USER, samples underwent PCR and gel purification before sequencing (1/30 of a HiSeq2000 lane).

RNA-seq library prep

RNA was isolated using the hot acid phenol method (see RNA isolation above) to ensure isolation of high quality RNA. All 6 samples, 2 WT, 2 *dbr1Δ*, 2 *upf1Δ,* had RQN (quality) values of 8.8 or higher as measured on the Advanced Analytical machine. Strand specific libraries were prepared by the MIT Bio Micro Center using the TruSeq™ RNA Sample Prep Kit v2 (RS-122-2101 kit) through cDNA after which LM-PCR was preformed using the Beckman Coulter SPRIte system with a 200-400bp size cutoff. Samples were barcoded and all sequenced in one HiSeq2000 lane, 60 X 60 bp.

Ribosomal RNA subtraction for the rapamycin treatment, *dbr1 Δ*, and the meiosis time course samples was performed using the Illumina Ribo-Zero Gold rRNA Removal Kit (Yeast) followed by standard strand-specific library preparation. All 20 samples were sequenced in one NextSeq lane, 75 x 75 bp.

Branch-seq read mapping

Reads were trimmed to 30 by 30 nt and mapped with Bowtie1 (Langmead et al. 2009) (bowtie-1.0.0) using the following parameters: **bowtie -S -m 1 -1 end1reads.fastq -2 end2reads.fastq.** Branch-seq reads for each gel slice were mapped to the genome and then combined using **samtools merge** (samtools-0.1.7a) (Li et al. 2009). Reads were initially mapped to SacCer2 (S288C_reference_genome_R61-1-1_20080605) and subsequently to SacCer3 (S288C_reference_genome_R64-1-1_20110203) downloaded from SGD. Peak

calling was performed using the SacCer2 genome and peak calls were converted to SacCer3 coordinates using liftOver tool (http://genome.ucsc.edu/) for some analyses. Peaks were called using the combined reads from the top, middle, and bottom sections of the arc. For Figure 1d if there were multiple peaks within 3 nt of the annotated BP, the annotated BP was only counted once.

winBP peak calling

A sliding window approach adapted from Arribere and Gilbert (Juneau et al. 2009; Arribere and Gilbert 2013) was used with some modifications in the winBP peak caller. A 200nt region was taken starting at the 5' end of each chromosome. Average read coverage per nucleotide, α, for this region was calculated using only BP end (second end) reads and was required to be at least 0.1. A sliding window of 5 nt (196 of these windows/200nt region) within each 200 nt region was used to reduce spurious calls in regions with uneven coverage. If coverage in the 5 nt sliding window was at least 12α a peak was called. At least 1nt was required between reported peaks. Peak calling was performed for each strand, always in the 5' to 3' direction. The 200nt regions were shifted 100 nt down the chromosome, and the steps outlined above were repeated until reaching the end of the chromosome. winBP recovered 58% (153/260) (Table 1) of annotated BPs in expressed genes. GEM-BP peak calling is described in Supplemental Methods.

GEM-BP peak calling

To discover BP events from the data, we extended the ChIP-seq and ChIP-exo peak caller GEM[4] that calls events with high spatial resolution. Unlike other peak callers, GEM does not assume any specific distribution of reads, and therefore is flexible to adapt to a new data type by learning a data-specific empirical spatial read distribution. We used a +/- 10bp window around the confident set of annotated BPs to learn the empirical read distribution (Fig. 1c) and used it for peak calling by GEM. To avoid including noisy reads from the non-BP strand, we modified GEM to perform single-strand peak calling and used only the 3' end (BP end) reads as input. As part of the integrated event finding and motif

discovery process, GEM discovered the consensus BP motif TACTAAC, some variants that are similar to the consensus motifs, and a poly A motif that represents technical artifacts resulting from anchored oligo(dT) RT step of the protocol. To distinguish events associated with different motifs, we modified GEM to use multiple position weight matrix (PWM) motifs as the positional priors for event discovery. If a base position is matched by multiple motifs, GEM chooses the PWM model that has a more significant p-value to set the positional prior. For each called event, GEM computes an event shape score that quantifies the similarity of the event read distribution to the empirical read distribution. The event shape score is defined as the Pearson correlation of read count values across the +/-10bp bases between the called event and the empirical read distribution. The new functionalities of the GEM software, which we called GEM-BP, were implemented in version 2.6. The following parameters were used to analyzed the Branch-seq data: --k 7 --a 2 --q 1 --bp --pp_pwm --not_update_model --nrf --nf.

We then post-processed the GEM-BP event calls to discover BP events using a Random Forest classifier (Breiman 2001) in the MATLAB software (MathWorks 2012). The features for the Random Forest include GEM-BP event read count, event shape score, and the binary motif categorical variables. We used the GEM-BP calls that overlap with the annotated BPs as the positive training set, and those that overlap with the tRNA genes as the negative training set. The trained Random Forest classifier was then applied to all of the GEM-BP event calls to make the final BP event calls.

In total, GEM-BP discovered 546 BPs (Table 1), including 75% of expressed BPs (196/260) (Table 1) within 3 nt of their annotated locations (Fig. 1d). Of 546 GEM-BP predicted BPs, 47 (8.6%) had more than one mismatch from the BP consensus motif TACTAAC, compared to 74 (21.5%) of the 344 peaks identified by the winBP approach. These numbers indicate that the GEM-BP predictions are more biased toward consensus BP, presumably because of its use of motif information and training on annotated BP, which match the consensus very closely, information which is not used by the winBP approach. Thus, we used the union of predictions made by both peak callers for subsequent analyses.

Typical 5'SS filter for putative novel BPs

GEM-BP and winBP together called numerous unannotated BPs in the yeast genome; the union of their peak calls yielded 430 putative novel BP peaks in all (Table 1). To define a high confidence subset of putative novel BPs, the paired-end sequencing information from Branch-seq was used as a built-in quality control for BP identification. Branch-seq data contains strand-specific read pairs connecting the BPs and 5'SS. Authentic putative novel BP resulting from splicing should be associated with a plausible 5'SS motif at the start of the associated 5' end reads, while any artefactual putative novel BP peaks would not be expected to have such a motif (or only at the background frequency of this motif in the genome).

For each BP, we took all BP end reads (3' end) within 5nt of the BP peak, accounting for strand. We obtained the paired 5'SS read for each BP read in this set and noted the location of the 5'SS read start. We calculated the mode position from all 5'SS read starts for that BP and looked at the 6mer motif at that position and one position 3'. We considered 6mers that matched the yeast 5'SS consensus GTATGT perfectly or with at most one mismatch as 'typical 5'SS motifs', and all others as 'atypical 5'SS motifs'.

Almost all (97%) annotated yeast introns in nuclear genes have typical 5'SS motifs by this definition (Table S3). Of the Branch-seq 3' end peaks that were associated with annotated BP, 76% (149/196) and 90% (138/153) had 5' end peaks at the annotated 5'SS for GEM-BP and winBP, respectively (Table S1). This result indicates that our approach can reliably and comprehensively map both the BP and 5'SS of introns, as intended.

After applying the typical 5'SS filter to the 430 putative novel BP, 268 cnBP remained. This subset of 268 should be treated as highly confident and was used for all downstream analyses. We estimate the FDR for the set of 268 BP is 1.1%, which is the genomic background frequency of 6mers matching typical 5'SS motifs in the yeast nuclear genome (Table S3, see bellow). Note, if the GEM-BP and winBP peaks were very close together, only the GEM-BP peak was counted.

Nicked lariats are expected to be extremely unstable in vivo. However, the presence of such species in the 2D gel arc used in Branch-seq could result in read pairs that have 3' ends located at arbitrary positions in an intron paired with authentic 5'SS, potentially resulting in artefactual BP identification. The putative BP sequence identified from such read pairs should be sampled more or less randomly from intron locations, and therefore

would show no bias for proximity to a BP sequence motif, allowing estimation of the frequency of such artifacts from the frequency of poor matches to the BP consensus motif. In our analysis, we used a window +/- 15 bp from the peak of Branch-seq reads to identify BP motifs (see "Novel and annotated BP motifs"). Noting that for 247/268 cnBP the best match to the yeast consensus TACTAAC was within 2 nt of the peak of Branch-seq reads (a window of 5 start positions), and most matched well to the consensus, we calculated the probability P3mm+ (= 45%) that the best matching 7mer to the yeast consensus within 15 bp of an arbitrary position in an annotated yeast intron was NOT within 2 bp of the center, AND was a poor match to the consensus (3 mismatches or more from TACTAAC). Thus, we expect 45% of artefactual cases to have these features. From the observed number of cnBPs that have these features (= 5), one can estimate the number of artefactual BPs within the cnBP set as 5/0.45 = ~11, or about 4% of the total. One might expect that the peak of Branch-seq reads should correspond very closely to the BP, as it did for annotated BPs (Fig. 2E), and we observed that the proportion of cnBPs with 3+ mismatches was much lower (7/247 = ~3%) when the best BP motif was within 2 bp of the Branch-seq peak, than when it was not (5/21 = 24%). These considerations motivate that cases where the best BP motif is within 2 bp of the peak be designated as "high confidence novel BPs" (hcnBPs) and cases where it is located further away – which likely have a much higher frequency of artifacts – be designated "low confidence novel BPs" (lcnBPs). Columns in Table S2 distinguish these two categories.

Another potential concern is that proteins bound to RNA may prevent exonucleolytic trimming causing identification of erroneous BP peaks downstream of annotated BPs. This is likely uncommon as 1) expanding the x-axis on Fig. 1C does not reveal strong peaks downstream of annotated BPs and 2) there are only 4 cases where a cnBP is located 3' of the annotated BP where the annotated and cnBP share the same 5'SS.

As a note, the overlap between the GEM-BP and winBP cnBP was only 80 BPs (Table 1), further suggesting that the two methods have different strengths and weaknesses in their ability to call novel BPs and there is benefit to using both methods.

One random position was selected in each of the 298 nuclear encoded intron containing genes in the SacCer3 genome annotation. The 6mer motif beginning at this location was score for number of mismatches from "GTATGT." This was done 10 times to

obtain 2980 simulated 5'SS in introns. 10 motifs had 0 mismatches and 24 motifs had 1 mismatch for an estimated FDR of 1.1% ((10+24)/2980) (Table S3).

Lariat tails are largely absent in vivo

Lariat tails appear to be efficiently digested *in vivo,* as previously reported, evidenced by a dearth of Lariat-seq reads in the long lariat tail of *UBC13* (Fig. S1B). With Branch-seq we are able to see RT priming preferences based on the nucleotides left down stream of the BP nucleotide after digestion of the lariat tail. It appears 2 nt are generally left after the BP, resulting in RT priming peaks that begin at the +1 or +2 position relative to the BP (Fig S1C) depending on the genomic sequence at those positions (Fig. S1D-E). The peak at -2 relative to the BP is likely to miss-priming of RT (Fig. S1D). See Fig. S1 legend for more information. To our knowledge, this is the first report of the precise number of nucleotides downstream of the BP nucleotide left undigested in lariat tails from RNA isolated from *dbr1Δ* yeast.

Comparison of cnBP to Qin et al. novel BPs

All 268 cnBP were compared to the 41 novel BP identified by Qin et al. Qin et al coordinates were liftedover from SacCer3 to SacCer2 for this comparison. Overall, 22 of the 41 novel BP reported by Qin et al were located within 7nt of cnBP (Table S2). Of the remaining 19 Qin BP, only 9 which had a good BP motif had no support from Branch-seq data and 7 of those 9 came from lariat loops > 100nt in length. We observe that Branch-seq is better at recovering lariat loops <100nt in length (Fig S3), explaining some of these 9 BPs. The remaining 10 Qin et al BPs are explained by a combination of difference reference BP annotations used, 5'SS with more than one mismatch from "GTATGT", and sequencing depth at those BP positions.

Mapping lariat junction reads

Lariat junction reads were identified and aligned in four main steps:

1. Reads were attempted to be aligned to the *S.cer.* genome using the Bowtie (version 1.0.0) read aligner and those aligning with fewer than 4 mismatches were omitted from further analysis.

2. Each unalignable read was split into two fragments such that each fragment was at least 12 bases long and the hexamer beginning the second fragment had maximum probability of being sampled from the *S.cer.* 5'ss position weight matrix. Reads for which this maximum probability was less than 0.01 were omitted from further analysis. The fragments will be referred to by their position at the 3' or 5' end of the original read moving forwards.

3. The fragment pairs were mapped to the *S.cer.* genome using the bowtie read aligner allowing no mismatches. The fragments were required to map in an inverted order (3' fragment upstream of 5' fragment). The final base of each 5' fragment, the putative BP nucleotide, was omitted from this alignment due to the prevalence of mismatches at this position.

4. For all fragment pairs with a valid alignment, the final base of each 5' fragment was re-added. The aligned position of the 3' end of the 5' fragment was called as a BP and the aligned position of the 5' end of the 3' fragment was called as the corresponding 5'ss.

Skipping across lariat 5'SS-BP junctions

We found that reverse transcriptase often introduces short insertions and deletions when crossing a lariat junction. This results in the 3' end of 5' fragment of lariat junction reads not always ending directly at the BP. The frequency of these events was determined by comparing the BP location called by each lariat junction read to a known BP location as annotated by Meyer et al. within 25 nts if one exists. Figure S2D reports the distribution when allowing no mismatches as used elsewhere in this paper. This criterion precluded observing insertion events as they were found to always have the sequence UACUACU at the 3' end of the 5' fragment, resulting in mismatches in the last two positions when aligned to the BP consensus motif.

BP calling from lariat junction reads

In order to make precise BP calls from the lariat junction reads, a probabilistic model based on the observed skipping rates in introns with annotated BP and a self-learned BP motif position weight matrix (PWM) was used.

Reads were separated into clusters based on proximity of their downstream ends. The i^{th}cluster of reads is denoted by R_i. The distribution $P(B_i = x \mid R_i)$, where B_i is a RV indicating the location of the BP generating R_i, was computed using the proportion $P(B_i = x \mid R_i) \propto P(R_i \mid B_i = x) * P(B_i = x)$. Assuming a uniform prior and that reads are independent given a BP, we rewrite this proportion as $P(B_i = x \mid R_i) \propto \prod_{r \in R_i} P(r \mid B_i = x)$. Note that $P(r \mid B_i = x)$ is simply the probability of observing a deletion of the size in read r given $B_i = x$.

An EM framework was used to learn a BP motif PWM, which was then used to improve precision. Beginning with an unbiased motif, the following protocol was repeated until the motif did not change between iterations:

1. Calculate $P(B_i = x \mid R_i, M)$, where M is the current motif, by multiplying $P(B_i = x \mid R_i)$ and the probability that the motif implied by $B_i = x$ would be sampled from $M$and then normalizing by the sum across each cluster.
2. Refine M based on the updated distribution. For each nucleotide in all positions in M, start with a pseudocount of 1. For all possible x, in all clusters i, add $P(B_i = x \mid R_i, M)$ to the count for the nucleotide in the respective position, for each position in the motif. Normalize by dividing all counts by the number of clusters plus 4.

Mapping RNA-seq reads for entropy calculations

60 X 60 bp reads (WT, upf1 null, and dbr1 null samples) were initially mapped with TopHat2 (Kim et al. 2013) (tophat-2.0.0.Linux_x86_64) giving TopHat no annotations and allowing it to discover novel splice junctions using the following parameters: **tophat -i 20 -I 10000 -a 10 --segment-length 15 --bowtie1 SacCer3 end1.fastq end2.fastq** Each barcoded sample was mapped on its own and additionally all samples were mapped together to find as many novel splice junctions as possible. A custom Bowite index was created for all splice junctions found by Tophat by concatenating the 50nt of sequence

immediately before and after the junction to ensure the reads had at least a 10nt overhang on each side of the junction. Bowtie1 was run with this custom index (genome + novel splice junctions) on each end of each sequencing library separately because parried end reads would be able to map to this custom index with many 100nt fragments. Bowtie was run as follows: **bowtie -S -m 1 -SacCer3_custom_index one_end_reads.fastq outfile.sam**. Bowtie read mapping to the custom splice index was used to calculate entropy of each splice junction (Graveley et al. 2011) using the formula below, as in Graveley et al., using the positions around the junction where read starts may fall.

pi = reads at offset i / total reads to junction window
Entopy = - sumi(pi * log(pi) / log2)

The entropy cutoff of 2 bits corresponds to uniform coverage of at least 4 distinct read start positions around each splice junction, or more variable coverage of a larger number of positions (Fig. S4A).

***RPL30* AT-AC isoforms**

These isoforms insert a stop codon early in the message, generating an upstream open reading frame (uORF). These isoforms might therefore be translated under specific conditions via uORF-medated translational regulation (Hinnebusch 1993), potentially producing a truncated protein comprising the C-terminal half of full length RPL30. RPL30 is known to regulate splicing and translation of transcripts from the *RPL30* locus by binding to RNA secondary structure at the 5' end of the pre-mRNA or mRNA.

Sequence conservation

PhastCons scores were downloaded from the UCSC genome browser (phastCons7way) for the novel BP and novel splice site analyses. For the novel splice site plots, the entire region surrounding the splice site in the figure had to fall into the region of question (i.e., intron or CDS). “Intergenic” refers to any region completely outside of a CDS or intron. For the BP conservation plot, only the location of the BP was considered for classifying the BP by location.

Protein length analysis

For all novel splice junctions with entropy at least 2 that overlap an annotated gene, the protein sequence of the resultant transcript was constructed. The length of each novel protein sequence was compared to the length of the annotated protein from the same gene and reported in Figure 4C. When constructing the novel protein sequences, the following assumptions were followed:

1. In cases where a gene has multiple novel splice junctions, only one is considered at a time (i.e. if there are 3 novel splice junctions in one gene, three protein sequences are created).
2. All annotated introns are spliced out, except if they overlap the novel splice junction being considered at the time.
3. If a novel splice junction removes the annotated translation start site, the next available AUG is used.

MISO analysis of splicing

In order to produce Figure 6A, retained intron annotations were created from all splice junctions with entropy >=2. Retained introns were splice junctions detected in the WT, upf1 null, or dbr1 null samples that did not overlap any other splice junctions detected, annotated or novel. To build the RI MISO annotations 200nt flanking the intron was used as exonic sequence. MISO (misopy/0.4.6) was run. For Waern et al data (downloaded from http://downloads.yeastgenome.org/published_datasets/Waern_2013_PMID_23390610/fastq/), --read-length = 76. For Brar et al. data (GEO accession number GSE34082), only reads of length 28-30 nt were used and --read-length was set to 29. Only footprints are shown for Brar et al. data because the total RNA libraries had few reads that fell into the 28-30 nt range. Prior to mapping Brar et al. data, poly(A) adaptor sequences were trimmed off of the reads using Cutadapt. Brar et al. and Waern et al. reads were mapped to the genome, defined splice junctions (UCSC, sacCer3), and novel splice junctions with entropy ≥ 2 in the WT, upf1 null, and dbr1 null RNA-seq (see above) using Tophat2. Summary tables from MISO output were generated for events with x=1, y=0, n=20, psi confidence = 0.5 (see

"Using the read class counts" https://miso.readthedocs.org/en/fastmiso/). These were considered "confident" psi values (see bellow).

In order to produce Figure 6B, annotations for all splice junctions detected in any sample with entropy ≥ 2 and all annotated introns were created. Splicing events that overlapped each other were grouped into one "gene" agnostic to known gene boundaries. The 200 nt flanking the furthest upstream 5'ss and the furthest downstream 3'ss for a given "gene" were used as exonic sequence. Miso (misopy/0.5.3) was run for all samples of both polyA and riboZero data. Events with a PSI confidence interval >= .5 or less than 20 reads uniquely assignable to either the given event isoform or the retained intron isoform were filtered out.

Clustering of PSI values

If an event had confident PSI values in at least half of the conditions, the missing psi values were replaced with the mean PSI from the confident samples. Clustering was done with heatmap.2 in R (Warnes et al. 2015).

Cufflinks (RNA-seq FPKMs)

Cufflinks(Trapnell et al. 2012) (version 2.2.1) was used to calculate FPKMs for the RNA-seq data using the command **cuffdiff -o . --library-type fr-firststrand -u -N -b SacCer3.fsa saccharomyces_cerevisiae_R64-1-1_20110208.gff wt1.bam,wt1.bam dbr1-1.bam,dbr1-2.bam upf1-1.bam,upf1-2.bam**

Branch-seq CPM calculations

Branch-seq CPMs were calculated using the formula CPM = F/((L)(M/1,000,000)) Where M is the total number of mapped reads. F is the number of strand-specific BP (3' end) reads within the L nucleotides centered on the BP peak. L=11 nt.

Genes with multiple BPs

5'SS-BP pairs from annotated introns with computationally predicted BPs (282)(Meyer et al. 2011) and all 268 cnBPs with typical 5'SS 5'SS-BP were considered in this analysis for a total of 550 5'SS-BP pairs. Any overlapping 5'SS-BP pairs on the same

strand were grouped into one "intron island." For islands that contain 2 or more BPs, it was required that there was a BP motif with 2 or fewer mismatches from "TACTAAC" within 3nt of the BP peak the keep the peak for downstream analyses. This yielded 11 intron islands that use 2 BPs and one intron island that uses 3 BPs. For the genes that use 2 BPs the distance from the 5'SS to the BP is the distance for each BP to its paired 5'SS. BP1 is the more 5'SS BP in the intron island. Sequence logos made with WebLogo (Crooks et al. 2004).

Novel and annotated BP motifs

Sequence 15nt up and downstream of the BP peaks were submitted to MEME (Bailey et al. 2009) (Version 4.10.0) to generate sequence logos. Only BP detected by Branch-seq are in the logos in Figure 2.

Human BP motif was generated using sequences 10 nt up and downstream of the BP nt from Mercer et al's (Mercer et al. 2015) annotated BPs. 1000 sequences were submitted to MEME (maximum MEME accepts) to generate the motif.

Conditions in Figure 6A

Conditions in (A), from table 1 of (Waern and Snyder 2013) include Exponential growth: YPD medium, Salt: 1M NaCl for 45 min, DNA damage: 1mM MMS for 1 hr, Alpha factor: 2.5mM for 45 min and add additional 50 uL to 25 mL yeast for additional 30 min, Sorbitol: 1M for 45 min, Oxidative stress: 0.4M H2O2 for 45 min, Heat shock: 37 deg C for 1 hr, Stationary phase: 18 day at 30 deg C, SC glycerol media: 4% glycerol instead of glucose, High calcium = 10mM calcium chloride medium, Low nitrogen: 20% of normal amount of Yeast Nitrogen Base in YPAD, Calcoflour: 0.1% for 1 hr, Hydroxyurea: 0.075M for 1 hr, Grape juice: filtered Walgreen's brand grape juice, Benomyl: 5 ug/mL for 1 hr, Congo red: 30 ug/mL for 1 hr.

***LSM2* qPCR primer sequences**

Actin primers:

ScerACT1_junct_F: ATGGATTCTGAGGTTGCTGCT

ScerACT1_mRNA_Rev: GGAGTCTTTTTGACCCATACCGA

LSM2 constitutive exon:

LSM2 qPCR Exon 2F constitutive: TAAAAAACGACATTGAAATAAAAGGTACA

LSM qPCR Exon 2R constitutive: TTCATCTGTGCATGATATGTTGTCTA

LSM2 novel 3'SS (PTC isoform):

LSM2 qPCR new 3'ss junction F: GTGGTCGTAGAGTCAAGTACTAAC

LSM qPCR Exon 2R constitutive: TTCATCTGTGCATGATATGTTGTCTA

LSM2 annotated 3'SS isoform:

LSM2 qPCR canonical (normal) 3'ss junction F: GTGGTCGTAGAGTTAAAAAACGAC

LSM qPCR Exon 2R constitutive: TTCATCTGTGCATGATATGTTGTCTA

RNA14 (NMD negative control):

GG10_for: ATGTCCAGCTCTACGACTCCTGAT

GG11_rev: GCGTATGACTCTTGAGTTTCCAAA (From Joshua Arribere(Arribere and Gilbert 2013))

TCA17 (NMD positive control):

GG8_for:GCCTTGCTTCGTATCATTGATAGA

GG9_rev:CATCATCAGCTCCACTTAGGCTTT (From Joshua Arribere(Arribere and Gilbert 2013))

RPL30 primer sequences

RT: SuperScript II protocol (Invitrogen)

GG13_YGL030W_rev: AAGCCAACTTTTGGTTGATAGA

PCR: Phusion (NEB)

GG14:YGL030W_5'end_for: agaccggagtgtttaagaacct

GG15:YGL030W_rev_ATACjunc: TAACTGGGGCctgttgaaat

SED1 primers

For Figure S4B:

RT: Random hexamers (Invitorgen), following SuperScript II protocol (Invitrogen).

PCR: Phusion (NEB)

GG17:SED1_for: TACATCTTTGCCACCAAGCA

GG18:SED1_rev: TTTGGTGGTAGTGCCCTTAGA

For Figure S5E - SED1 apparent RT artifact

Colony PCR was performed to put a T7 primer onto the start of the SED1 sequence. PCR product was gel extracted and used as a temple for T7 in vitro transcription (Epicentre AmpliScribe™ T7-Flash™ Transcription Kit), DNA was digested, and RNA product was cleaned via phenol chloroform extraction. RNA was gel extracted using UV shadowing visualization. RT and PCR were performed as in Figure S5B.
Scer_SED1_colony_Forward: TAATACGACTCACTATAGGGgacaagcaaaataaaatacgttcg
Scer_SED1_colony_Reverse: ttaaactacccctattgcttttaga

Plotting

Additional plots in this paper were made with ggplot2(Wickham 2009), IGV (Robinson et al. 2011), matplotlib, Pictogram, WebLobo, and MEME.

Legends to Supplemental Figures

Figure S1. Additional details pertaining to Branch-seq protocol.
(A) Left: 2D gel used to isolate lariats from top, middle, and bottom sections of arc. Right: Top and bottom splices excised. D1: 6% TBE-urea. D2: 20% TBE-urea. (B) Read coverage (green) in *UBC13* intron from Larita-seq. Depletion of reads between BP and 3'SS indicates lariat tails are digested when lariats accumulate in *dbr1*Δ yeast (Chapman and Boeke 1991). (C) Additional examples like inset in figure 1B of read start plots for BPs in 4 individual introns. The majority of reads are located at +1 or +2 position on an intron by intron basis. (D) Hypothesis for predominant +1 vs +2 read start position in individual introns. RNA sequence in black, question marks are unknown nucleotides after the BP. BP A in red. The RT primer, green, may prime at different locations, and produce sequencing products (blue arrow), starting at different positions relative to the BP nucleotide. +1 sequencing is expected if nucleotide after TACTAAC is an A because of anchored oligo(dT) priming step in RT. Similarly, +2 position is expected if nucleotide after TACTAAC is C, G, or T. Sequencing at -2 is due to mis-priming of anchored oligo(dT) primer over the terminal C of the BP motif. (E) Genomic sequence immediately downstream of annoated BPs (boxed) with maximum peak from (C) at +1, left, and +2, right, confirms hypothesis in (D). (F) Branch-seq reads in the *EFM5* intron are shifted 5 nt from the annotated BP location (blue underline) corresponding to a AACTAAC BP (red underline).

Figure S2. Further characterization of novel BPs.
(A) Left: Novel BPs (blue) are not conserved compared to annotated BPs (red). Right: novel BPs from blue line in left plot broken down by genomic location. (B) 5'SS motif of 162 putative novel BP with atypical 5'SS. (C) Novel BP overlapping *YDL138W* ORF (plus strand) comes from the minus strand, potentially from a longer form of the annotated CUT/SUT on the minus strand. Novel BP is confirmed by one Branch-seq read pair and several Lariat-seq junction reads. (D) RT sometimes skips over the BP nucleotide in Lariat-seq junction reads (see methods).

Figure S3. Characteristics of lariats captured by Branch-seq.
(A) Comparison of expression levels of lariats recovered in Branch-seq (combined top, middle, and bottom slices of arc) to expression of their parent mRNA in poly(A) selected RNA-seq. Only annotated BPs are plotted. (B) Same as (A) but regression calculated for different lariat sizes, suggested that Branch-seq read counts are semi-quantitative for lariat loops smaller than 100 nt. (C) Expression level of annotated and novel BPs recovered by Branch-seq. (D) Lariat loop lengths recovered by Branch-seq and Lariat-seq LJ reads.

Figure S4. Novel introns confirmed by entropy resemble annotated introns but preferentially come from short transcripts.
(A) Entropy of annotated (green) and novel (pink) splice junctions, separated by splice site motif AT/AC, GC/AG, GT/AG. A cutoff of entropy of 2 was used to define novel splice junctions(Graveley et al. 2011). (B) 5'SS and 3'SS motifs for annotated (top) and novel (bottom) splice sites. (C) Gene lengths (TSS to poly(A) site) (Pelechano et al. 2013) for genes containing novel BPs identified in Branch-seq and genes containing novel introns with entropy ≥ 2 identified in RNA-seq data.

Figure S5. Experimental testing of AT-AC splice site introns.
RT-PCR on total RNA to verify (A) *RPL30* and (B) *SED1* AT-AC splice sites. *SED1* AT-AC splice site intron is located inside a long repeat (C) highlighted in green and (D) shown in a dot plot. (E) RT-PCR on in-vitro transcribed full length *SED1* RNA. The presence of a product here of the expected spliced size suggests the presence of some sort of RT artifact.

Figure S6. Conservation of novel intron splice sites from isoforms that show splicing patterns similar to annotated introns.

Arrows above each splice site indicate sequence direction. UCSC browser snapshots are shown for splice sites located outside of coding sequences.

Figure S7. Translation of YNL194-YNL195C fusion transcript changes throughout meiosis time course.

Sashimi plots depict reads in exons and reads spanning splice junction (numbered arcs) with PSI value shown to the right with confidence bounds (tie fighter plot). Plots are ordered by progression through meiosis time course from Brar et al. (Brar et al. 2012) for (A) ribosome footprint profiling data.

Table S1: Branch-seq BP peaks paired 5'SS motifs.

Table S2: SacCer2 coordinates of GEM-BP and winBP peaks.

Table S3: GTATGT motif frequency at 5'SS and generally in introns.

Table S4: Branch-seq CPMs.

Table S5: Saccer3 coordinates of lariat junction reads.

Table S6: Novel splice junctions with entropy ≥ 2bits.

Table S7: Figure 6A PSI values.

Table S8: Figure 6A event annotations.

Table S9: Figure 6B PSI values.

Supplemental references

Arribere JA, Gilbert WV. 2013. Roles for transcript leaders in translation and mRNA decay revealed by transcript leader sequencing. *Genome research* **23**: 977–987.

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**: W202–8.

Benjamin KR, Zhang C, Shokat KM, Herskowitz I. 2003. Control of landmark events in meiosis by the CDK Cdc28 and the meiosis-specific kinase Ime2. *Genes Dev* **17**: 1524–1539.

Berchowitz LE, Gajadhar AS, van Werven FJ, De Rosa AA, Samoylova ML, Brar GA, Xu Y, Xiao C, Futcher B, Weissman JS, et al. 2013. A developmentally regulated translational control pathway establishes the meiotic chromosome segregation pattern. *Genes Dev* **27**: 2147–2163.

Brar GA, Yassour M, Friedman N, Regev A, Ingolia NT, Weissman JS. 2012. High-resolution view of the yeast meiotic program revealed by ribosome profiling. *Science* **335**: 552–557.

Carlile TM, Amon A. 2008. Meiosis I is established through division-specific translational control of a cyclin. *Cell* **133**: 280–291.

Chapman KB, Boeke JD. 1991. Isolation and characterization of the gene encoding yeast debranching enzyme. *Cell* **65**: 483–492.

Clarkson BK, Gilbert WV, Doudna JA. 2010. Functional overlap between eIF4G isoforms in Saccharomyces cerevisiae. *PLoS ONE* **5**: e9114. http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=20161741&retmode=ref&cmd=prlinks.

Crooks GE, Hon G, Chandonia J-M, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome research* **14**: 1188–1190.

Folco EG, Lei H, Hsu JL, Reed R. 2012. Small-scale nuclear extracts for functional assays of gene-expression machineries. *J Vis Exp*.

Folco EG, Reed R. 2014. In vitro systems for coupling RNAP II transcription to splicing and polyadenylation. *Methods in molecular biology (Clifton, NJ)* **1126**: 169–177.

http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=24549664&retmode=ref&cmd=prlinks.

Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, van Baren MJ, Boley N, Booth BW, et al. 2011. The developmental transcriptome of Drosophila melanogaster. *Nature* **471**: 473–479.

Hinnebusch AG. 1993. Gene-specific translational control of the yeast GCN4 gene by phosphorylation of eukaryotic initiation factor 2. *Mol Microbiol* **10**: 215–223.

Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. 2009. Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science* **324**: 218–223.

Katz Y, Li F, Lambert NJ, Sokol ES, Tam W-L, Cheng AW, Airoldi EM, Lengner CJ, Gupta PB, Yu Z, et al. 2014. Musashi proteins are post-transcriptional regulators of the epithelial-luminal cell state. *eLife* **3**: e03915. http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=25380226&retmode=ref&cmd=prlinks.

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 fusions. *Genome Biol* **14**: R36.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **10**: R25. http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=19261174&retmode=ref&cmd=prlinks.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25**: 2078–2079. http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=19505943&retmode=ref&cmd=prlinks.

Mercer TR, Clark MB, Andersen SB, Brunck ME, Haerty W, Crawford J, Taft RJ, Nielsen LK, Dinger ME, Mattick JS. 2015. Genome-wide discovery of human splicing branchpoints. *Genome research*.

Meyer M, Plass M, Pérez-Valle J, Eyras E, Vilardell J. 2011. Deciphering 3'ss selection in the yeast genome reveals an RNA thermosensor that mediates alternative splicing. *Mol Cell*

43: 1033–1039.

Pelechano V, Wei W, Steinmetz LM. 2013. Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature* **497**: 127–131.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29**: 24–26.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**: 562–578.

Waern K, Snyder M. 2013. Extensive transcript diversity and novel upstream open reading frame regulation in yeast. *G3 (Bethesda, Md)* **3**: 343–352. http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=23390610&retmode=ref&cmd=prlinks.

Warnes GR, Ben Bolker, Bonebakker L, Gentleman R, Liaw WHA, Lumley T, Maechler M, Magnusson A, Moeller S, Schwartz M, et al. 2015. gplots: Various R Programming Tools for Plotting Data. R package version 2.16.0. *CRANR-projectorg*.

Wickham H. 2009. ggplot2: elegant graphics for data analysis.

Figure S1

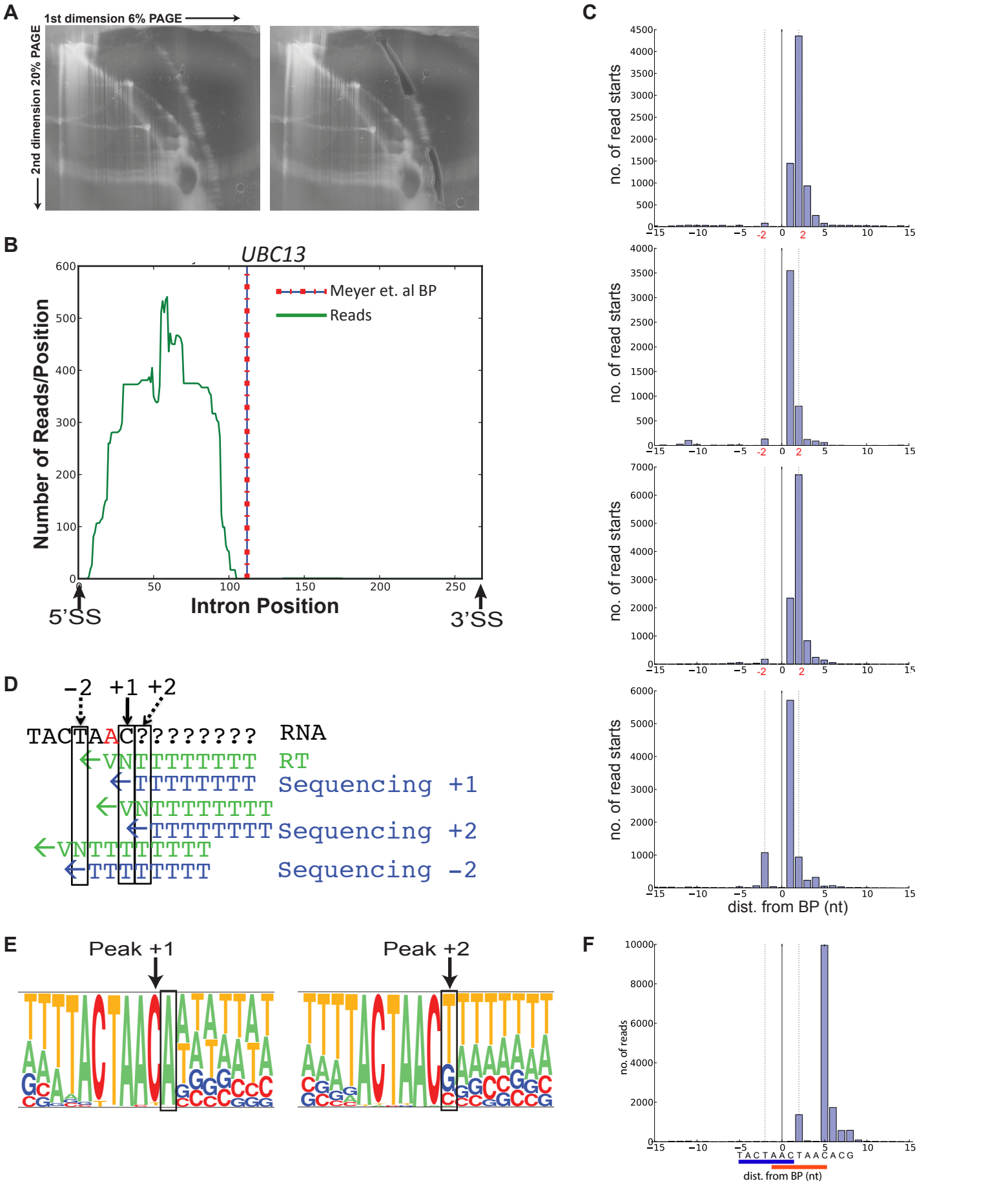
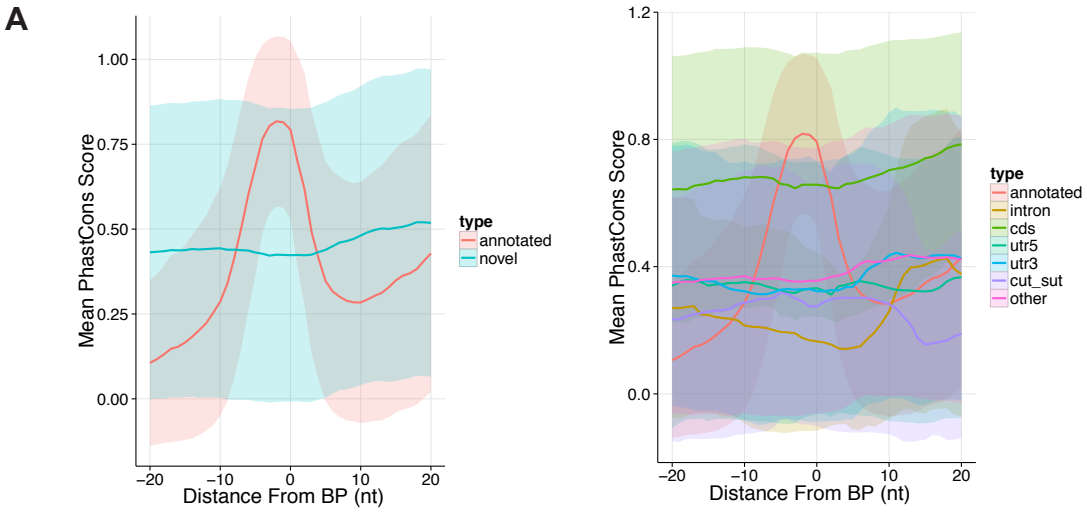
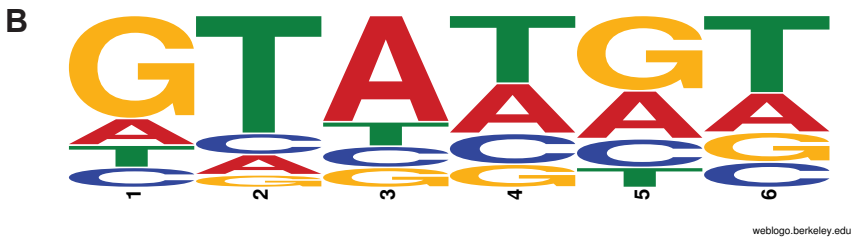
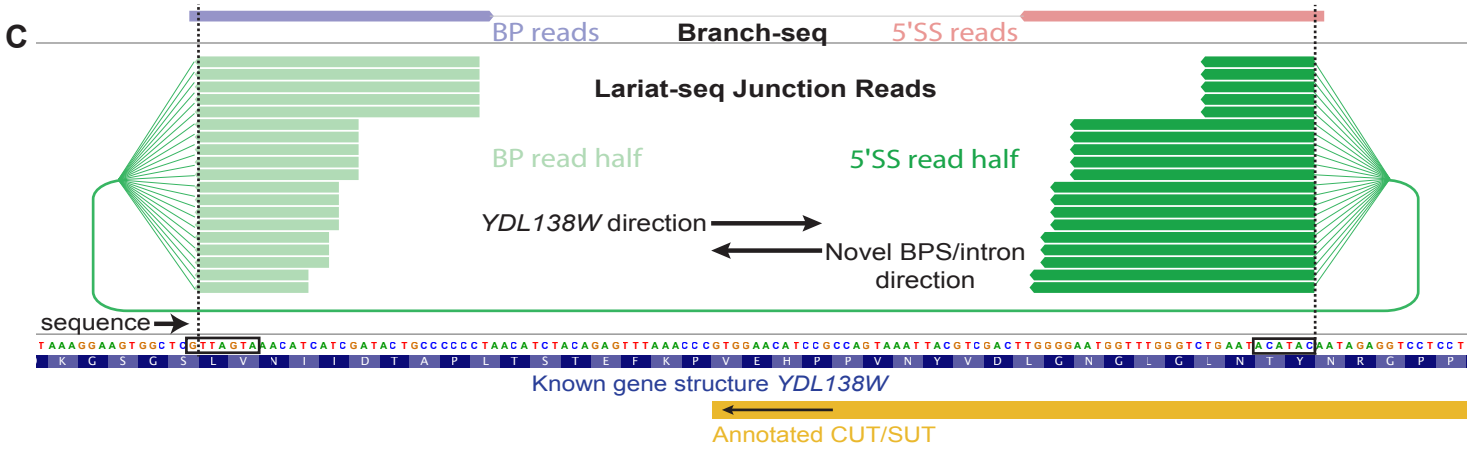

A
1st dimension 6% PAGE
2nd dimension 20% PAGE
B
UBC13
Meyer et. al BP
Reads
Number of Reads/Position
Intron Position
5'SS
3'SS
C
no. of read starts
dist. from BP (nt)
D
−2 +1 +2
TACTAAC???????? RNA
VNTTTTTTTT RT
TTTTTTT Sequencing +1
VNTTTTTTTT
TTTTTTTT Sequencing +2
VNTTTTTTTT
TTTTTTTT Sequencing -2
E
Peak +1
Peak +2
F
no. of reads
TACTAACTAACACG
dist. from BP (nt)

Figure S2


A
Mean PhastCons Score
1.00
0.75
0.50
0.25
0.00
−20
−10
0
10
20
Distance From BP (nt)
type
annotated
novel
Mean PhastCons Score
1.2
0.8
0.4
0.0
Distance From BP (nt)
type
annotated
intron
cds
utr5
utr3
cut_sut
other


B
weblogo.berkeley.edu


C
BP reads
Branch-seq
5'SS reads
Lariat-seq Junction Reads
BP read half
5'SS read half
YDL138W direction
Novel BPS/intron direction
sequence
Known gene structure YDL138W
Annotated CUT/SUT

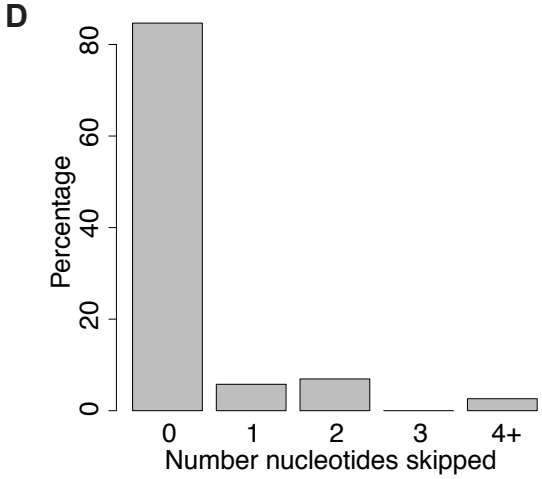

D
Percentage
0
20
40
60
80
0
1
2
3
4+
Number nucleotides skipped

Figure S3

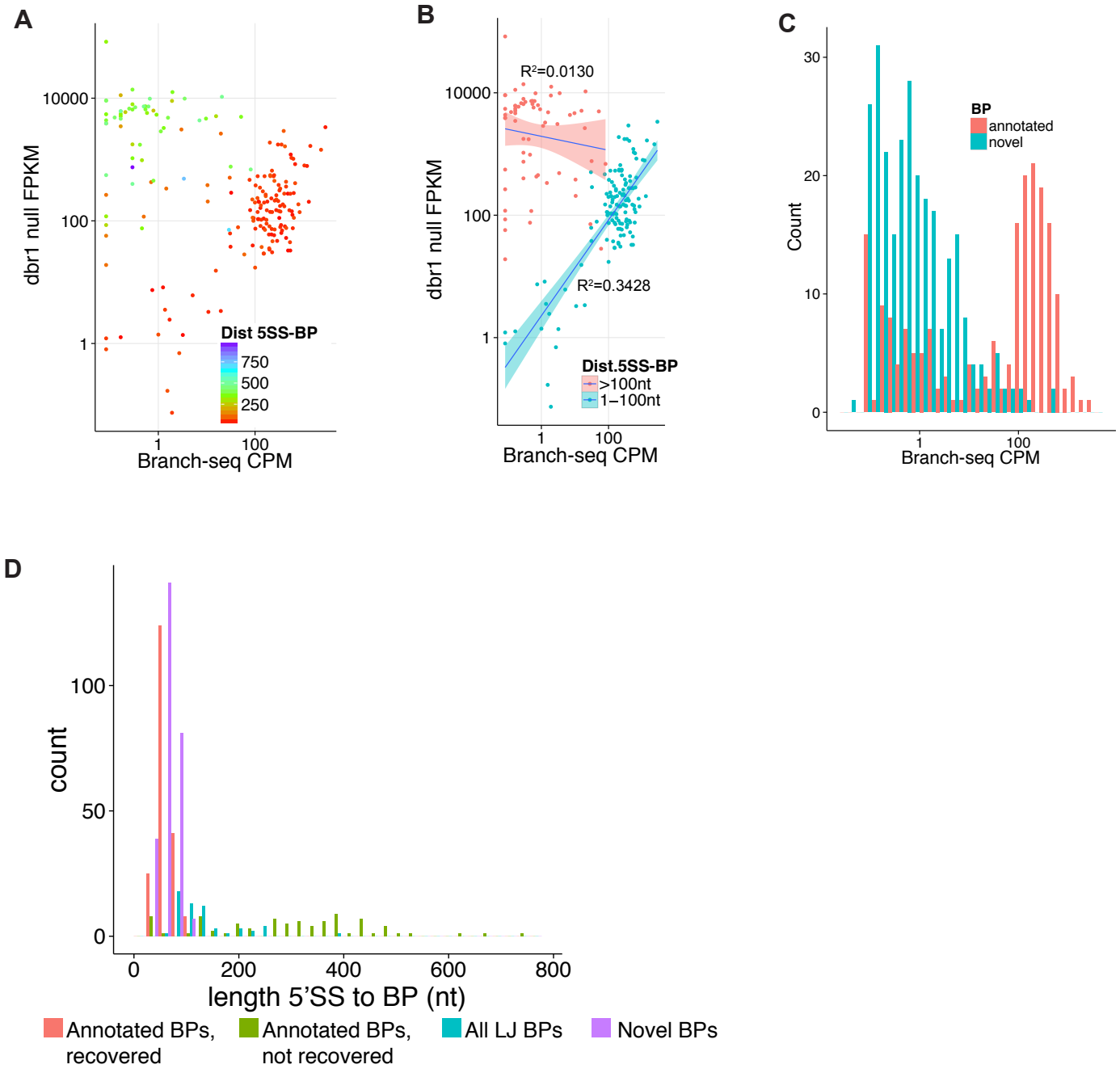

A
dbr1 null FPKM
10000
100
1
Dist 5SS-BP
750
500
250
1
100
Branch-seq CPM
B
R²=0.0130
R²=0.3428
Dist.5SS-BP
>100nt
1–100nt
C
Count
30
20
10
0
BP
annotated
novel
D
count
100
50
0
0
200
400
600
800
length 5'SS to BP (nt)
Annotated BPs, recovered
Annotated BPs, not recovered
All LJ BPs
Novel BPs

Figure S4

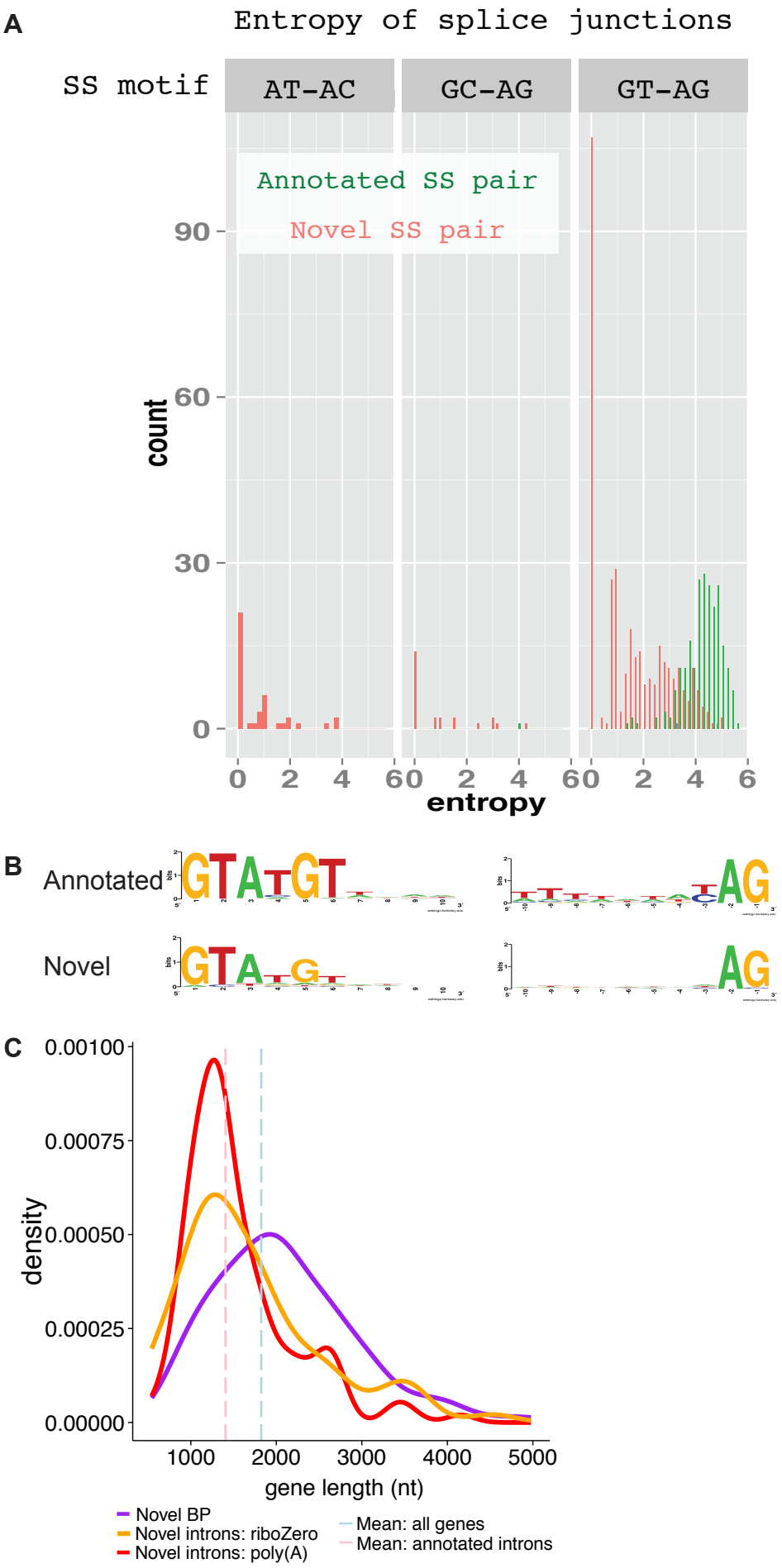
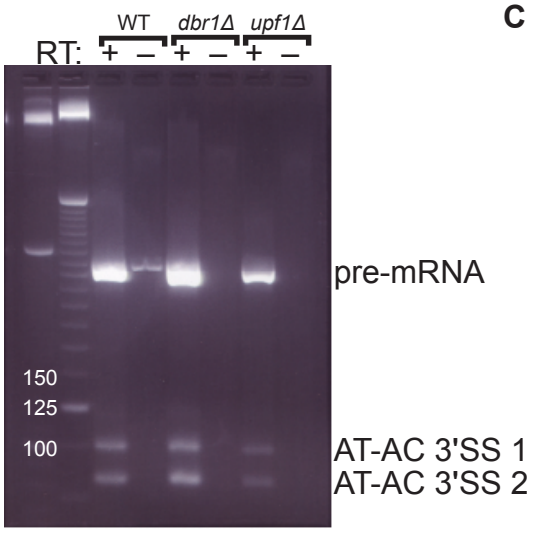

A
Entropy of splice junctions
SS motif
AT-AC
GC-AG
GT-AG
Annotated SS pair
Novel SS pair
count
90
60
30
0
0 2 4 60 2 4 60 2 4 6
entropy
B
Annotated
Novel
C
density
0.00100
0.00075
0.00050
0.00025
0.00000
1000 2000 3000 4000 5000
gene length (nt)
Novel BP
Novel introns: riboZero
Novel introns: poly(A)
Mean: all genes
Mean: annotated introns

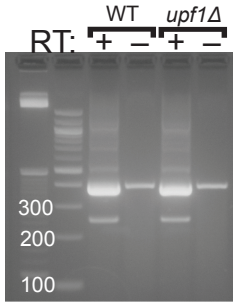
Figure S5

A



RPL30 PCR Expected Sizes
pre-mRNA = 297nt
AT-AC = 83nt

B

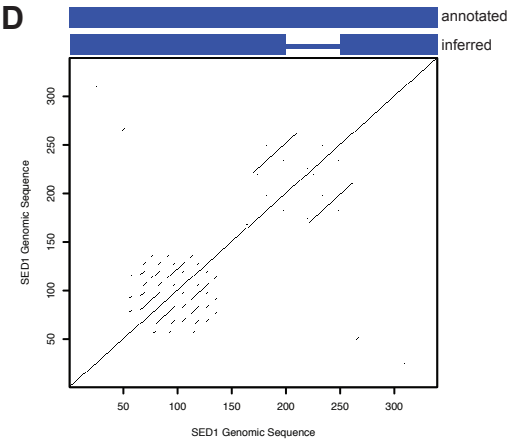


SED1 PCR Expected Sizes
No intron = 409nt
predicited (ent) AT-AC = 301nt
actual AT-AC = 256nt

C

```
CLUSTAL 2.1 multiple sequence alignment

SED1_Genomic      ATGAAATTATCAACTGTCCTATTATCTGCCGGTTTAGCCTCGACTACTTTGGCCCAATTT 60
PCR-band-05       ------------------------------------------------------------

SED1_Genomic      TCCAACAGTACATCTGCTTCTTCCACCGATGTCACTTCCTCCTCTTCCATCTCCACTTCC 120
PCR-band-05       ------------------------------------------------------------

SED1_Genomic      TCTGGCTCAGTAACTATCACATCTTCTGAAGCTCCAGAATCCGACAACGGTACCAGCACA 180
PCR-band-05       ------------------------------------------------------------

SED1_Genomic      GCTGCACCAACTGAAACCTCAACAGAGGCTCCAACCACTGCTATCCCAACTAACGGTACC 240
PCR-band-05       ------------------------------------------------------------

SED1_Genomic      TCTACTGAAGCTCCAACCACTGCTATCCCAACTAACGGTACCTCTACTGAAGCTCCAACT 300
PCR-band-05       ------------------------------------------------------------

SED1_Genomic      GATACTACTACTGAAGCTCCAACCACCGCTCTTCCAACTAACGGTACTTCTACTGAAGCT 360
PCR-band-05       ------------------------------------------------------------

SED1_Genomic      CCAACTGATACTACTACTGAAGCTCCAACCACCGGTCTTCCAACCAACGGTACCACTTCA 420
PCR-band-05       ------------------------------------------------------------

SED1_Genomic      GCTTTCCCACCAACTACATCTTTGCCACCAAGCAACACTACCACCACTCCTCCTTACAAC 480
PCR-band-05       ------------------------------------------------------------

SED1_Genomic      CCATCTACTGACTACACCACTGACTACACTGTAGTCACTGAATATACTACTTACTGTCCA 540
PCR-band-05       -CATCTACTGACTACACCACTGACTACACTGTAGTCACTGAATATACTACTTACTGTCCA 59
                   ************************************************************

SED1_Genomic      GAACCAACCACTTTCACCACAAACGGTAAGACTTACACCGTCACTGAACCAACCACATTG 600
PCR-band-05       GAACCAACCACTTTCACCACAAACGGTAAGACTTACACCGTCACTGAACCAACCAC---- 115
                  ************************************************************

SED1_Genomic      ACTATCACTGACTGTCCATGCACCATTGAAAAGCCAACAACCACATCAACCACCGAATAC 660
PCR-band-05       ------------------------------------------------------------

SED1_Genomic      ACTGTAGTCACTGAGTACACTACTTACTGTCCAGAACCAACCACTTTCACCACAAACGGT 720
PCR-band-05       ------------------------------------------------------------

SED1_Genomic      AAGACTTACACCGTCACTGAACCAACCACTTTGACTATCACTGACTGTCCATGTACTATT 780
PCR-band-05       ----------------------------TTTGACTATCACTGACTGTCCATGTACTATT 146
                                              *******************************

SED1_Genomic      GAAAAGAGCGAAGCCCCTGAGTCTTCTGTCCCAGTTACCGAATCTAAGGGCACTACCACC 840
PCR-band-05       GAAAAGAGCGAAGCCCCTGAGTCTTCTGTCCCAGTTACCGAATCTAAGGGCACTACCACC 206
                  ************************************************************

SED1_Genomic      AAAGAAACAGGTGTTACTACCAAACAAACCACAGCCAACCCAAGTCTAACCGTCTCCACA 900
PCR-band-05       AAA--------------------------------------------------------- 209
                  ***

SED1_Genomic      GTCGTCCCAGTTTCATCCTCTGCTTCTTCTCATTCCGTTGTCATCAACAGTAACGGTGCT 960
PCR-band-05       ------------------------------------------------------------

SED1_Genomic      AACGTCGTCGTTCCAGGTGCTTTAGGTTTGGCTGGTGTTGCTATGTTATTCTTATAA 1017
PCR-band-05       ---------------------------------------------------------
```

D



E

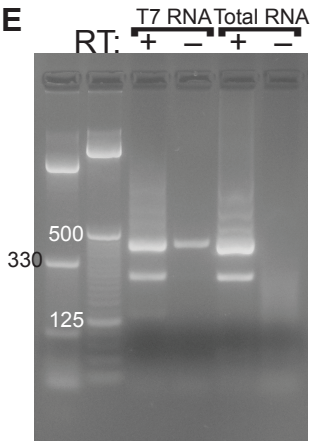


Figure S6

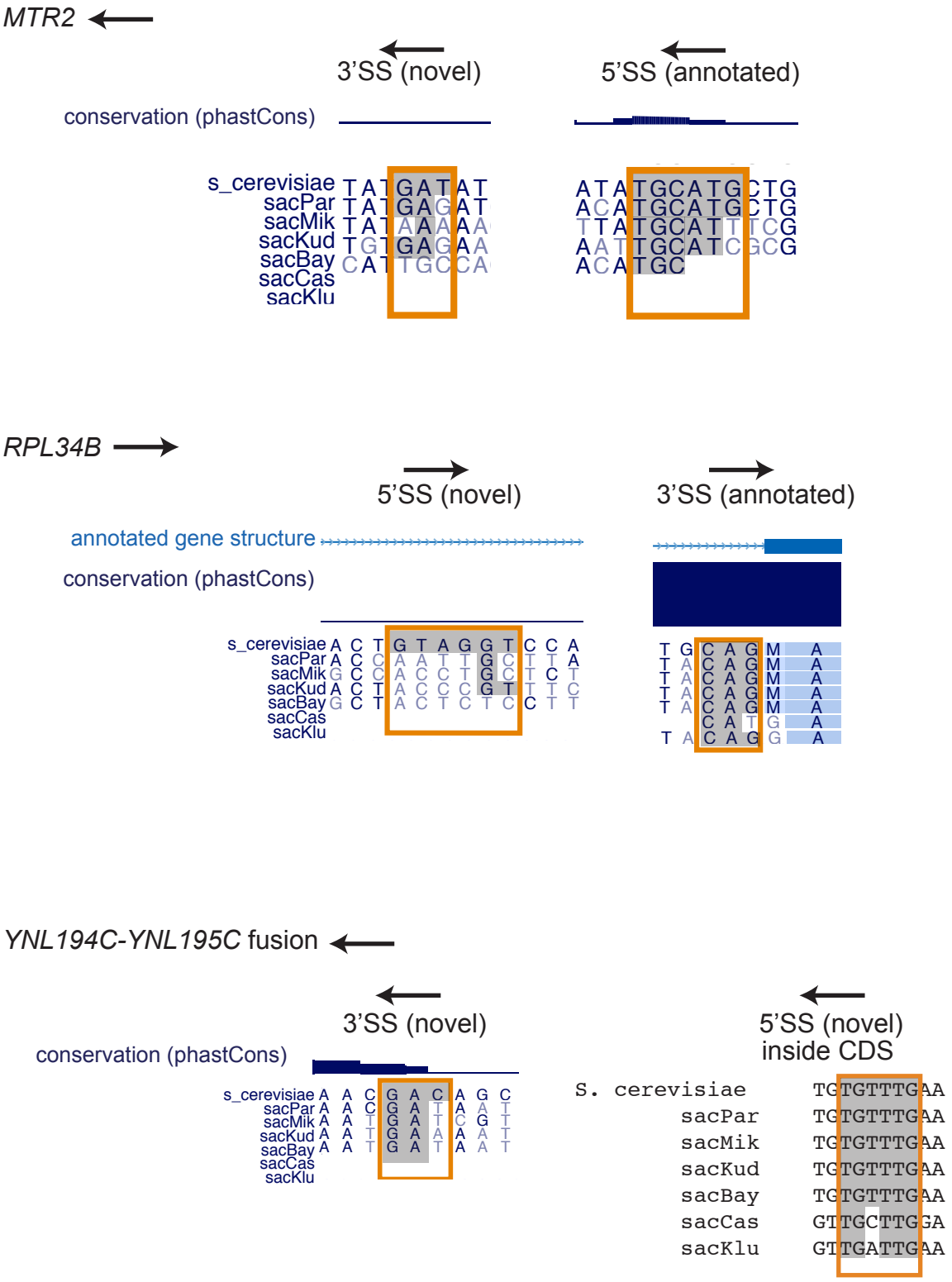
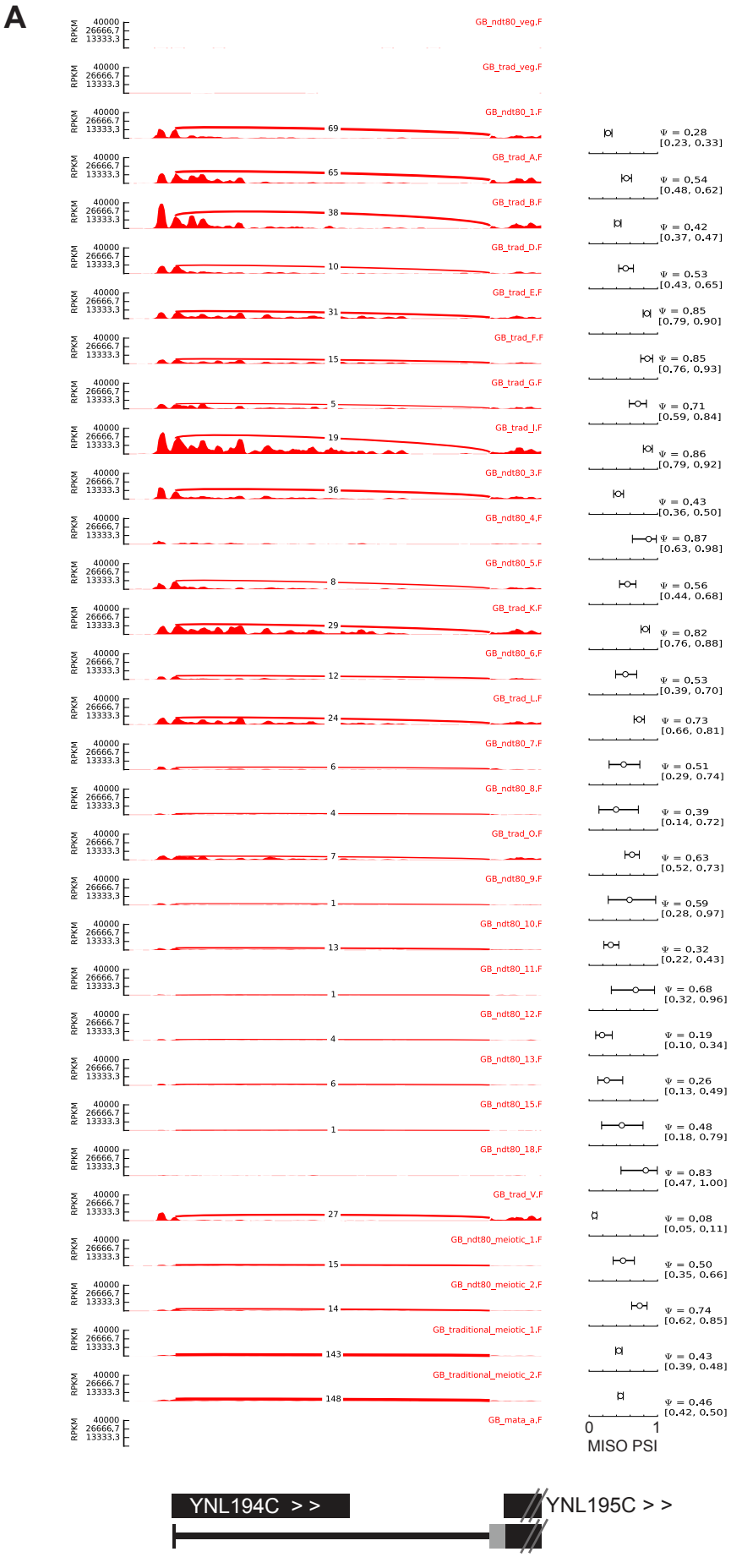

MTR2
3'SS (novel)
5'SS (annotated)
conservation (phastCons)
s_cerevisiae
sacPar
sacMik
sacKud
sacBay
sacCas
sacKlu
RPL34B
5'SS (novel)
3'SS (annotated)
annotated gene structure
conservation (phastCons)
YNL194C-YNL195C fusion
3'SS (novel)
5'SS (novel)
inside CDS
conservation (phastCons)
S. cerevisiae
sacPar
sacMik
sacKud
sacBay
sacCas
sacKlu
TGTGTTTGAA
TGTGTTTGAA
TGTGTTTGAA
TGTGTTTGAA
TGTGTTTGAA
GTTGCTTGGA
GTTGATTGAA

Figure S7


A
RPKM
40000
26666.7
13333.3
GB_ndt80_veg.F
GB_trad_veg.F
GB_ndt80_1.F
69
GB_trad_A.F
65
GB_trad_B.F
38
GB_trad_D.F
10
GB_trad_E.F
31
GB_trad_F.F
15
GB_trad_G.F
5
GB_trad_I.F
19
GB_ndt80_3.F
36
GB_ndt80_4.F
GB_ndt80_5.F
8
GB_trad_K.F
29
GB_ndt80_6.F
12
GB_trad_L.F
24
GB_ndt80_7.F
6
GB_ndt80_8.F
4
GB_trad_O.F
7
GB_ndt80_9.F
1
GB_ndt80_10.F
13
GB_ndt80_11.F
1
GB_ndt80_12.F
4
GB_ndt80_13.F
6
GB_ndt80_15.F
1
GB_ndt80_18.F
GB_trad_V.F
27
GB_ndt80_meiotic_1.F
15
GB_ndt80_meiotic_2.F
14
GB_traditional_meiotic_1.F
143
GB_traditional_meiotic_2.F
148
GB_mata_a.F
Ψ = 0.28 [0.23, 0.33]
Ψ = 0.54 [0.48, 0.62]
Ψ = 0.42 [0.37, 0.47]
Ψ = 0.53 [0.43, 0.65]
Ψ = 0.85 [0.79, 0.90]
Ψ = 0.85 [0.76, 0.93]
Ψ = 0.71 [0.59, 0.84]
Ψ = 0.86 [0.79, 0.92]
Ψ = 0.43 [0.36, 0.50]
Ψ = 0.87 [0.63, 0.98]
Ψ = 0.56 [0.44, 0.68]
Ψ = 0.82 [0.76, 0.88]
Ψ = 0.53 [0.39, 0.70]
Ψ = 0.73 [0.66, 0.81]
Ψ = 0.51 [0.29, 0.74]
Ψ = 0.39 [0.14, 0.72]
Ψ = 0.63 [0.52, 0.73]
Ψ = 0.59 [0.28, 0.97]
Ψ = 0.32 [0.22, 0.43]
Ψ = 0.68 [0.32, 0.96]
Ψ = 0.19 [0.10, 0.34]
Ψ = 0.26 [0.13, 0.49]
Ψ = 0.48 [0.18, 0.79]
Ψ = 0.83 [0.47, 1.00]
Ψ = 0.08 [0.05, 0.11]
Ψ = 0.50 [0.35, 0.66]
Ψ = 0.74 [0.62, 0.85]
Ψ = 0.43 [0.39, 0.48]
Ψ = 0.46 [0.42, 0.50]
0
1
MISO PSI
YNL194C > >
YNL195C > >