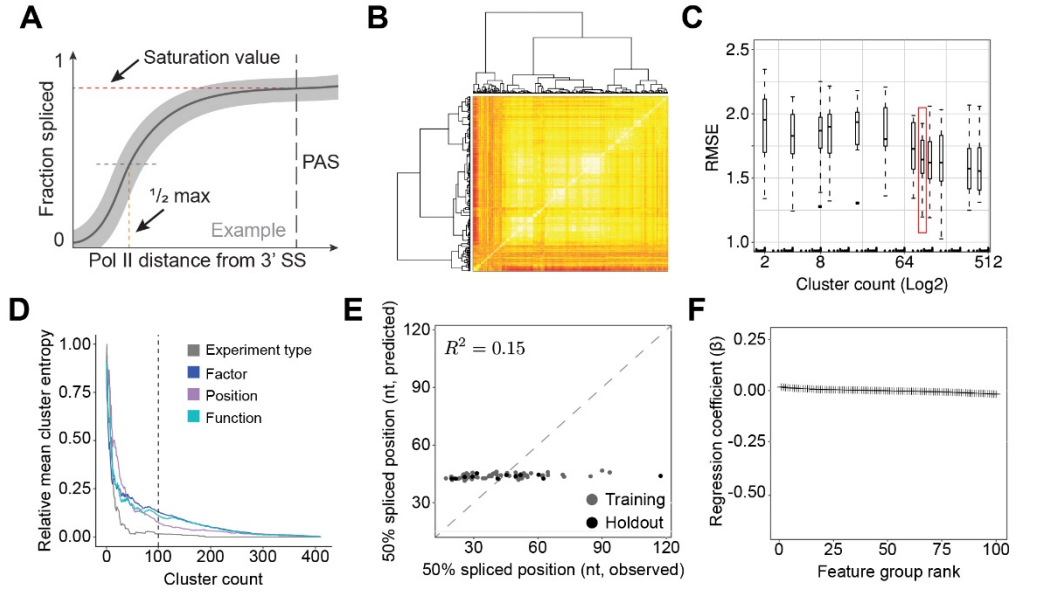
Cell Reports, Volume *33*

Supplemental Information

Widespread Transcriptional Readthrough Caused by Nab2 Depletion Leads to Chimeric Transcripts with Retained Introns

Tara Alpert, Korinna Straube, Fernando Carrillo Oesterreich, and Karla M. Neugebauer

SUPPLEMENT



G

Protocol variation	Cycle number for: First PCR	Second PCR	Pooled primers	Forward amplification
1	15	15	No	No
2	15	15	Yes	No
3	20	5	No	Yes
4	25	5	No	Yes
5	10	5	Yes	Yes
6	20	5	Yes	Yes
7	22	5	Yes	Yes
8	25	5	Yes	Yes
9	30	5	Yes	Yes

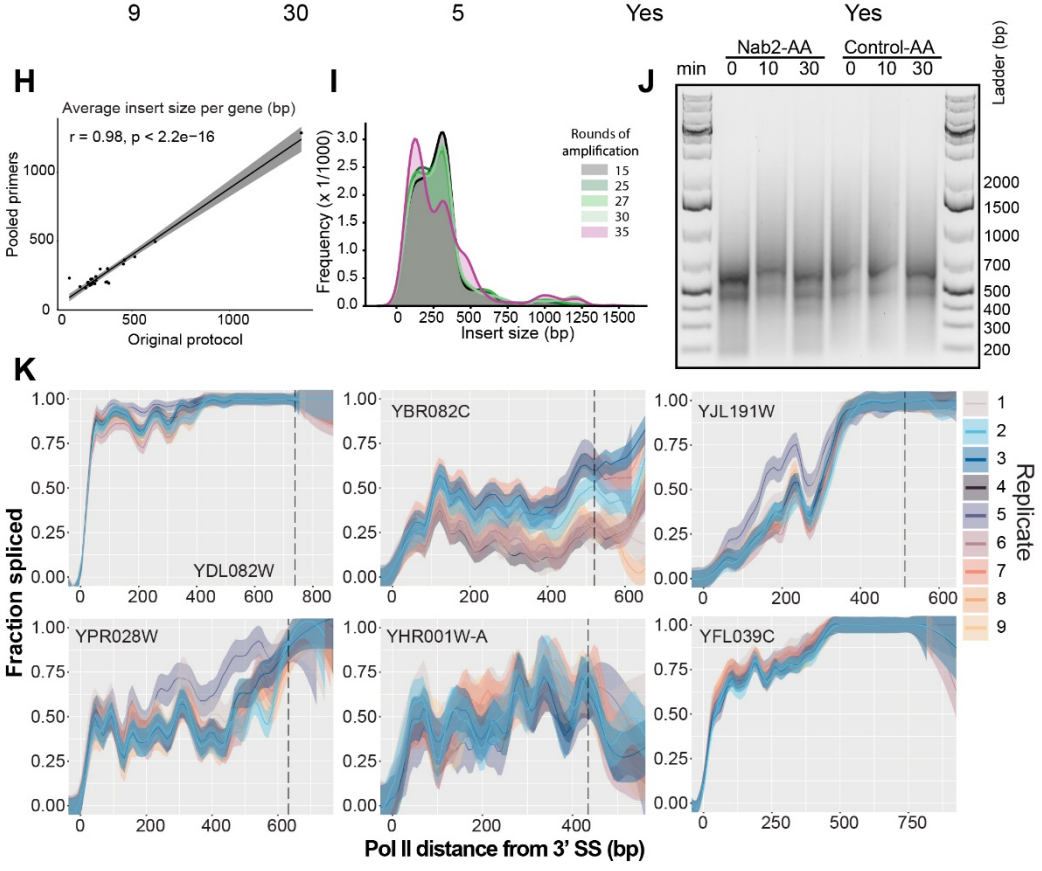


Figure S1. Related to Figure 1. Machine learning and optimizing SMIT for reproducibility and multiplexing.

A. Example of a splicing profile with two parameters indicated, saturation value and ½ max. Saturation value is calculated as mean fraction spliced of the last four 30 bp bins before the PAS (or last available bins if data does not extend to PAS). The Pol II distance from 3′ SS corresponding to half of the saturation value is the ½ max value.

B. Hierarchical clustering of input features (i.e. gene characteristics). Pair-wise correlation coefficients between features are represented in a heat map. Pearson correlation coefficients are color-coded, ranging from -1 (red) to +1 (white). Rows and columns are ordered by hierarchical clustering using squared Pearson correlation coefficient as a similarity measure. Clustering is represented by dendrograms. Dimensionality of the data is reduced by representing the data by feature-groups (i.e. clusters), generated by cutting the dendrogram at different heights.

C. Prediction performance of lasso regression as a function of cluster-count used to represent the data. Root mean squared error (RMSE) plotted against number of clusters (Log2 space). RMSEs of 3-times repeated 5-fold cross validation are represented as box plots. RMSE values for data represented by 100 clusters is highlighted (red box).

D. Separation of experiment type (grey), factor identity (blue), gene-position (purple) and biological function (teal) as a function of cluster-count used to represent the data. Separation is measured by averaging Shannon-entropy for each cluster and normalized to the Shannon-entropy observed for data-represented by only one cluster (root-node). Values for data represented by 100 clusters are highlighted (black line).

E. Lasso regression model was trained to predict the ½ max parameter and results of observed ½ max values are plotted against predicted.

F. Regression coefficients are plotted for all feature groups input into the model.

G. Table describing the SMIT protocol variations (rows) that were tested. PCR cycle number was varied to test amplification bias. Gene-specific forward primers were either pooled into a single PCR reaction or each primer was input into a separate PCR reaction. Finally, a forward amplification step was tested which uses only the forward primers and synthesizes only the first strand of DNA. This step amplifies the sample in a linear fashion rather than exponential growth of traditional PCR. 50 rounds of forward amplification PCR were optionally included prior to the first SMIT PCR.

H. Average insert size per gene for the original protocol (variation 1) is plotted against the values for the comparable protocol with pooled primers (variation 2). Linear regression modeled (black line) with a 95% confidence interval (grey ribbon).

I. Frequency of insert sizes are shown for protocol variations 5-9 which differ only in first PCR cycle number.

J. Agarose gel shows final amplified SMIT library for Nab2-AA and Control-AA samples treated with rapamycin for 0-, 10-, or 30-minutes.

K. Extensive replication was performed for a subset of genes to determine the reproducibility of SMIT. Vertical dashed line indicates position of the PAS. Data points are modeled using a Loess smoothing method and a 95% confidence interval.

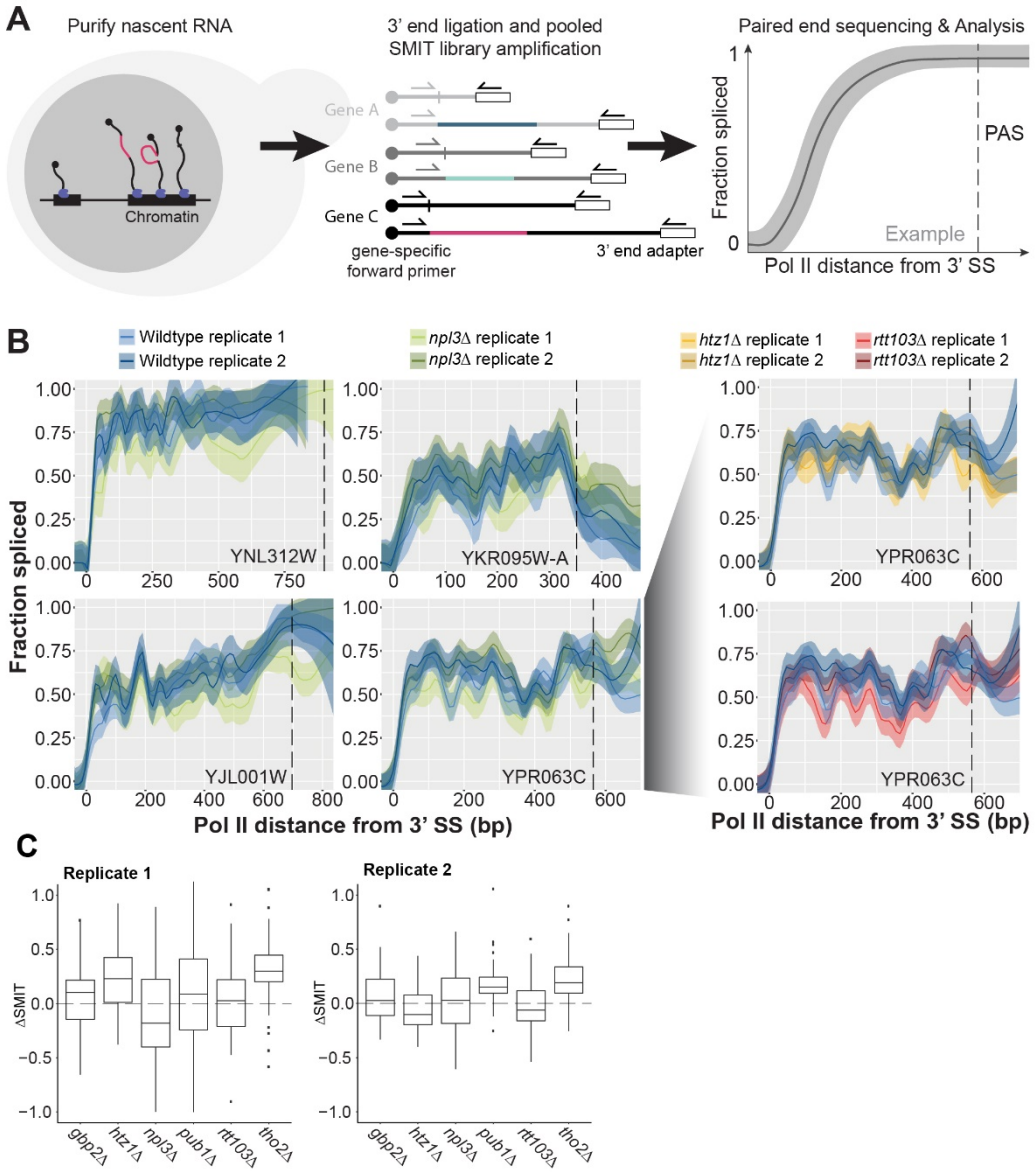


Figure S2. Related to Figure 1. Testing machine learning predictions by multiplexing SMIT.

A. Schematic of SMIT experiment shows Pol II associated nascent RNA purified from chromatin of budding yeast (left), SMIT adapter ligation and library amplification (center) and model data (blue) with saturation value and ½ max parameters indicated (right).

B. We sourced deletion strains from the Genome Deletion Project (Giaever et al., 2002; Winzeler et al., 1999), however this collection has been shown to harbor frequent secondary mutations (Teng et al., 2013). To ensure our samples didn't harbor undetectable compensatory mutations, we amplified out the deletion cassette and retransformed into a stable background strain. Wildtype (blue) and *npl3Δ* (green) splicing profiles are compared for four example genes (left). Additional splicing profiles for YPR063C from the *htz1Δ* (yellow) and *rtt103Δ* (red) samples are shown as well (right). Data points are modeled using a Loess smoothing method and a 95% confidence interval.

C. For each gene, the difference between the deletion strain and the wildtype SMIT values was calculated as the Euclidean distance of fraction spliced for the first 300 nt of each second exon (binned by 60 nt to minimize sequencing noise). The distribution of these ΔSMIT values are plotted for each strain where the middle bar represents the median and the edges of the box represent the first and third quartiles.

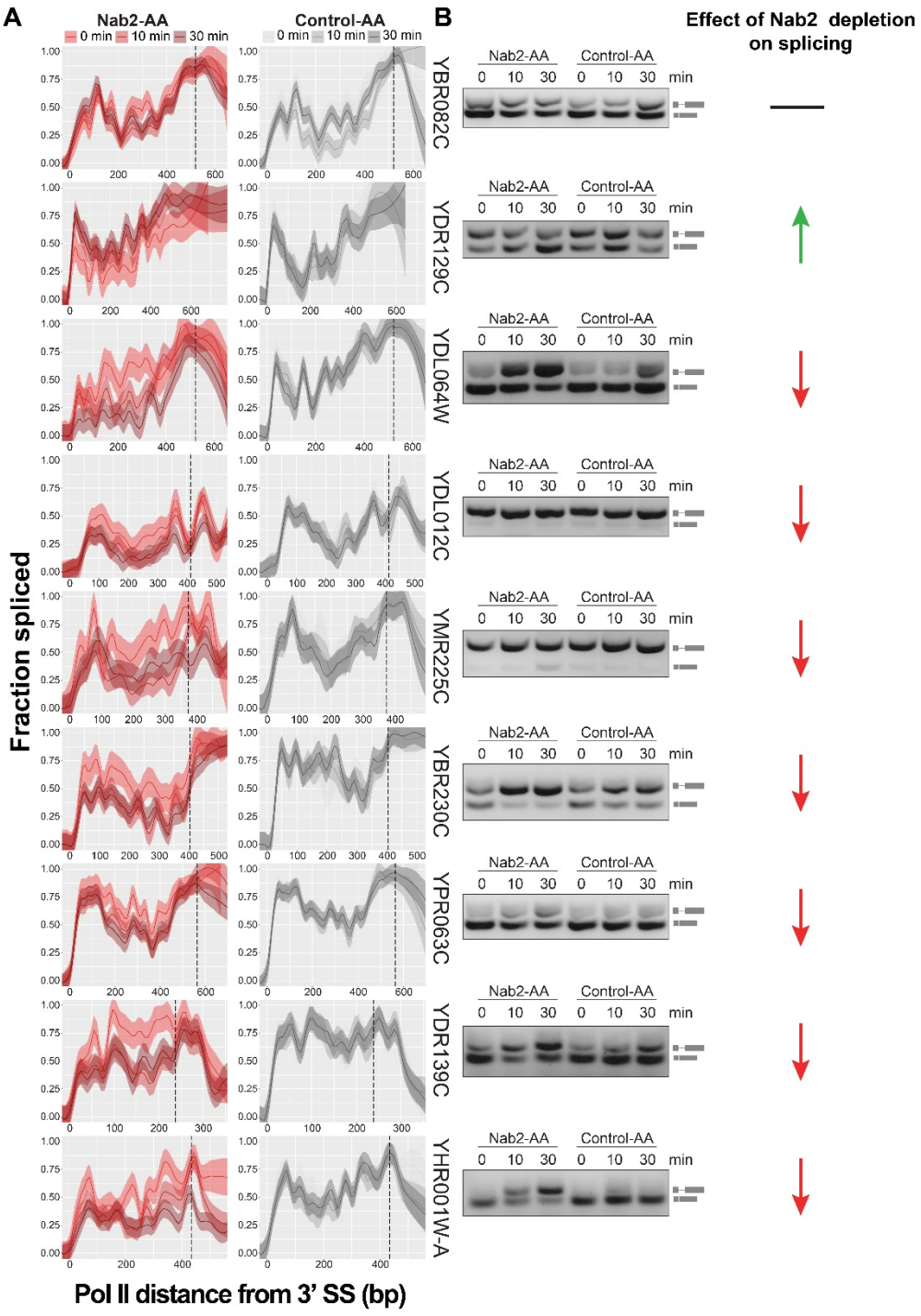


Figure S3. Related to Figure 2. Gene-specific changes in co-transcriptional splicing upon Nab2 depletion as detected by SMIT and validated by RT-PCR.

A. Splicing profiles are shown for a selection of genes during Nab2 depletion (red) and in the control (grey). Data points are modeled using a Loess smoothing method and a 95% confidence interval.

B. To validate the effect Nab2 has on each splicing profile, RT-PCR was performed on nascent RNA from Nab2-AA and Control-AA samples. RNA was reverse transcribed using random hexamers and intron-spanning primers then amplified both spliced (bottom) and unspliced (top) product which is visualized on 1% agarose (right). Arrows on the right indicate the effect of Nab2 depletion on splicing as shown in both splicing profiles and RT-PCR. Horizontal black line indicates no change, green arrows indicate an increase in splicing, red arrows indicate a decrease in splicing.

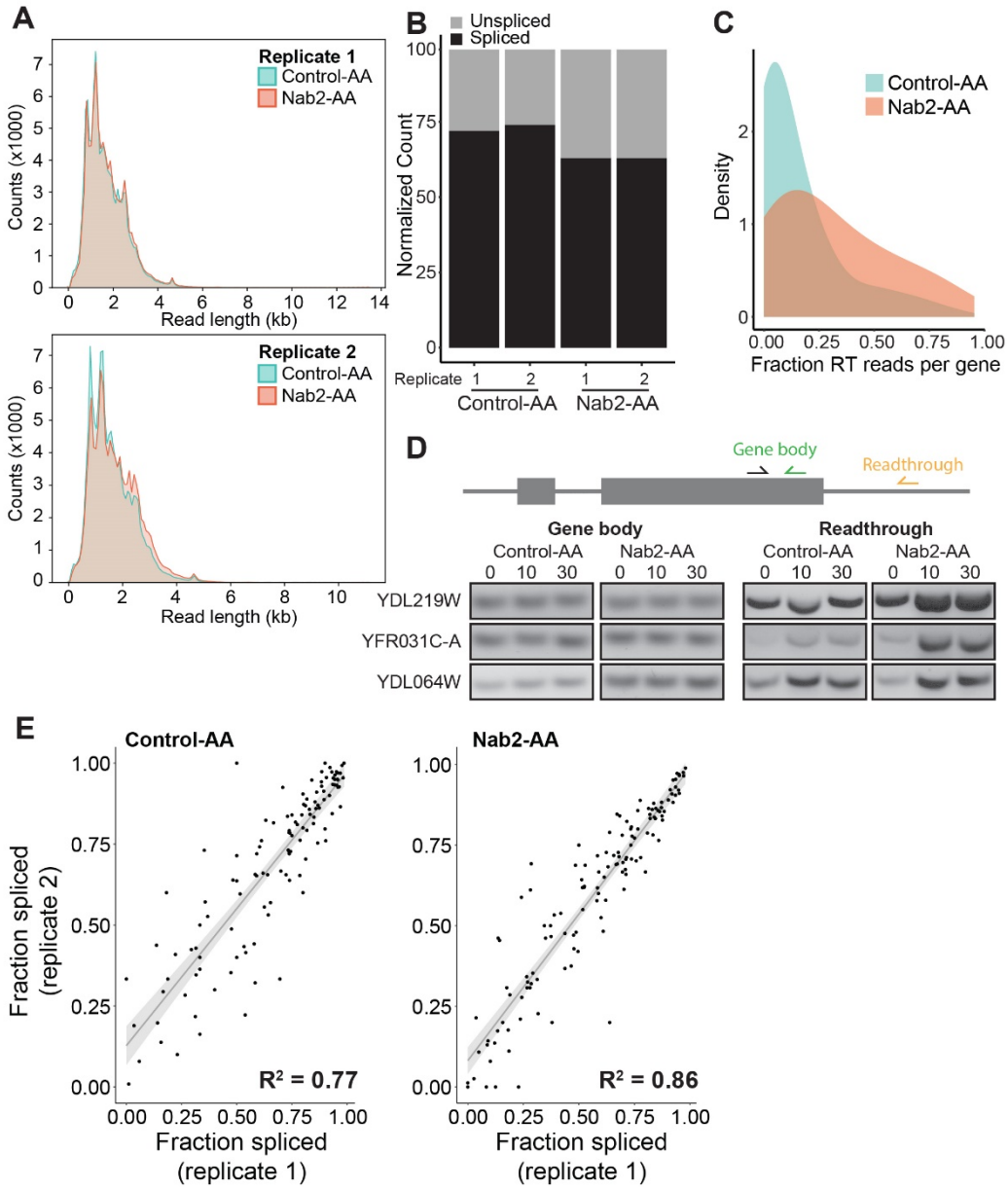


Figure S4. Related to Figure 3. Long read sequencing of nascent RNA upon Nab2 depletion.

A. Read length distribution for long read sequencing datasets for both Control-AA (teal) and Nab2-AA (orange).

B. Count of spliced (black) and unspliced (grey) reads for each replicate are normalized to reach 100.

C. The fraction of reads per gene which readthrough the polyA site of that gene are plotted as a distribution for both control (teal) and Nab2-AA (orange).

D. RT-PCR was performed to validate the readthrough phenotype of Nab2-AA observed in the sequencing data. Nascent RNA was reverse transcribed with random hexamers and then amplified with a common forward primer (black) in the gene body and a reverse primer either in the gene body (green) or in the region downstream of the PAS (yellow). PCR products are visualized on agarose gels for gene body (left) and downstream readthrough (right) for 0-, 10-, and 30-minute time points of rapamycin treatment in Control-AA and Nab2-AA cells.

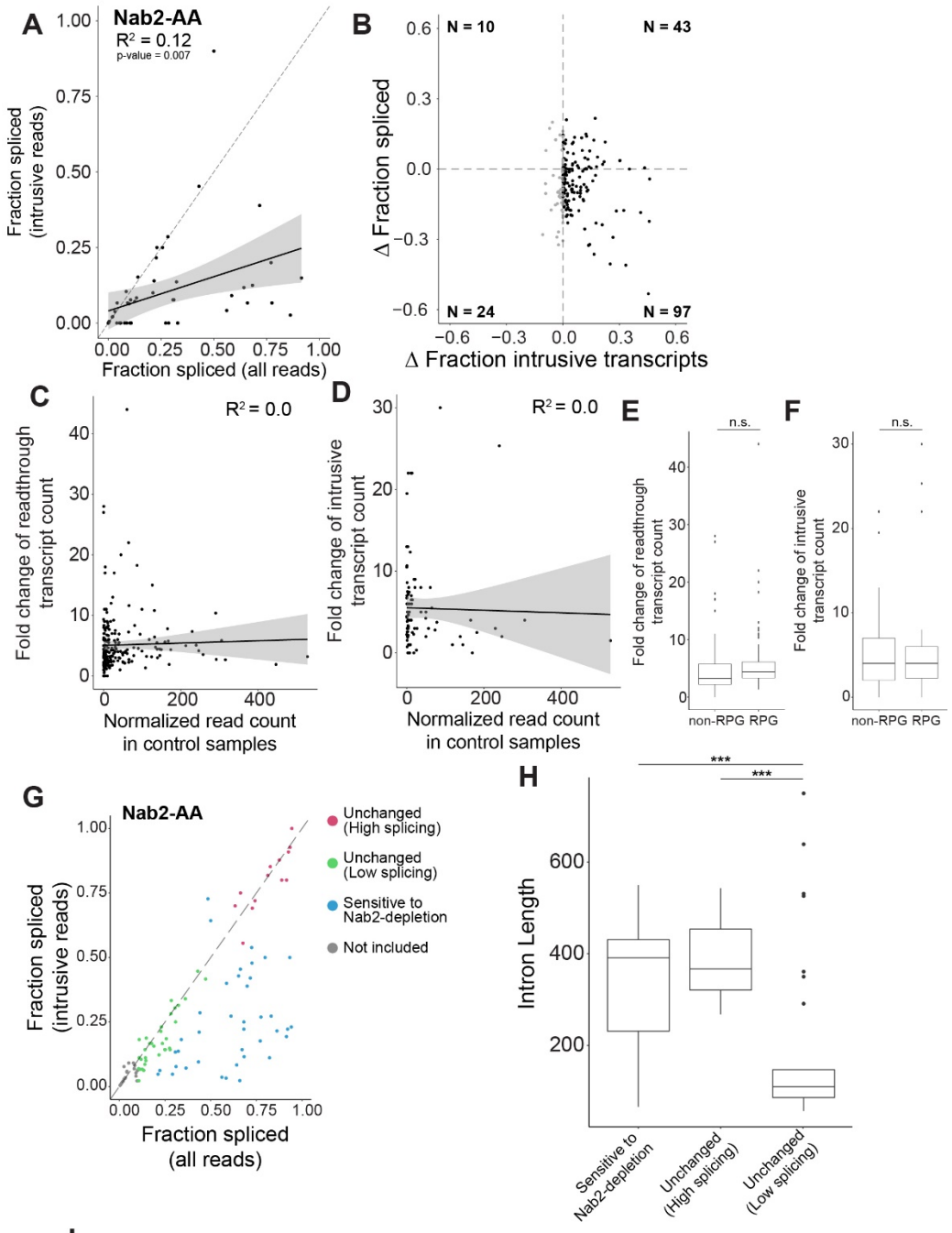
E. Fraction spliced are calculated for reads which start within 50 bp of the TSS (excluding intrusive transcripts) and values are plotted for each replicate. Adjusted R^2 values are shown for linear regression models (grey) and the 95% confidence interval.


YPL080C
YPL081W
YPL079W
YPL078C
YPL077C
Control-AA
1000 bp


YPL080C
YPL081W
YPL079W
YPL078C
YPL077C
Nab2-AA
1000 bp

Figure S5. Related to Figure 4. Nab2 depletion induces pervasive readthrough transcription.

Representative unfiltered long read data are shown for a segment of the genome (annotation above in grey) including the intronless genes. Reads are too numerous to show in their entirety, so a representative subset was chosen for display here using the default organization of Integrative Genomics Viewer (Robinson et al., 2011). Reads from two biological replicates are combined for Nab2-AA (orange) and Control-AA (teal).



I

	5`SS		
Category	**Consensus**	**Non-consensus**	**Percent non-consensus**
Unchanged (High splicing)	13	1	7%
Unchanged (Low splicing)	24	4	14%
Sensitive to Nab2-depletion	30	8	21%

Figure S6. Related to Figure 4. Intrusive transcripts generated by failed cleavage events are unspliced.

A. This plot uses the same analysis as in Figure 4D; however, a more stringent definition of intrusive transcripts was used where reads must overlap with the upstream gene. This stringent filtering exacerbates the relationship shown in Figure 4. Each datapoint is an individual gene with at least 10 reads intrusive reads. The plot for Control-AA is not shown because very few genes (< 5) met this criterion given the low levels of readthrough in wild type conditions.

B. The change in fraction of intrusive transcripts (Nab2-AA - Control-AA) is plotted against the change in fraction spliced (Nab2-AA - Control-AA). Number of points in each quadrant are shown with positive delta values for intrusive transcripts shown in black and negative values shown in grey.

C. Read count in the Control-AA sample was calculated to represent gene expression in these datasets (intrusive reads were removed from this value). The fold change (Nab2-AA / Control-AA) of readthrough reads is then plotted according to our expression values. The adjusted R^2 of the linear regression fit is displayed along with the 95% confidence interval.

D. The same expression values calculated for D where used for comparison against the fold change of intrusive reads.

E. The distribution of fold change values for readthrough reads is shown for ribosomal protein genes (RPGs) and non-RPGs. The difference between the two is not significant using the Mann-Whitney U test (n.s.).

F. The distribution of fold change values for intrusive reads is shown for RPGs and non-RPGs. The difference between the two is not significant using the Mann-Whitney U test (n.s.).

G. The right panel from Figure 4D is reproduced here with data points from the Nab2-AA nanopore dataset colored according to their designation into the categories listed: Unchanged (High splicing) (pink), Unchanged (Low splicing) (green), or Sensitive to Nab2-depletion (blue). A small number of poorly spliced genes were excluded from analysis because of their low read count. Dashed line represents y = x. Genes were categorized based on their distance from the y = x axis and whether their fraction spliced (all reads) was above or below 0.5.

H. Intron length was compared for the genes in each category defined in H. Genes with shorter introns are shown to splice less efficiently (Carrillo Oesterreich et al., 2010), so the results here were expected. Significance (Mann-Whitney U test) as follows: $p \leq 0.05$ (*), $p \leq 0.01$ (**), $p \leq 0.001$ (***), $p \leq 0.0001$ (****).

I. The 5′SS for each gene in I was identified as being either the consensus sequence (‘GTATGT’) or a variant of this sequence (non-consensus). The percent non-consensus value is displayed alongside the 5′SS counts for each category defined in I.