Supplemental Material for

Delineating yeast cleavage and polyadenylation signals using deep learning

Emily Kunce Stroup and Zhe Ji

This document contains the following information:

Supplemental methods;

Supplemental figure legends.

Supplemental table legends.

SUPPLEMENTAL METHODS

The optimization of the PolyaClassifer model.

Parameter grid search and model selection

To identify the optimal PolyaClassifier architecture, we employed a grid search protocol where each parameter of the model was varied one at a time as other parameters were held constant at a baseline value. For each parameter value tested, we trained 10 models, each using a different set of sampled negative controls to provide a measure of performance variability and to control for the effect of individual negative samples. Each set of negative sequences contained the same number of positive sequences. 70% of them were sampled from the random genome sequences without the PASS reads in the middle 50-nt regions. The remaining 30% were from shuffled transcript sequences keeping the single nucleotide compositions the same.

We began by exploring the impact of the input poly(A) sequence size, which we varied from 50 to 1000 nt in 50 nt steps. For each step, we trained 10 models using the same positive sequences and 10 randomly sampled negatives as described above. Based on the averaged AUROC values from the 10 models, we examined the impact of the sequence length parameter on model performance. In our final model, we selected 500 nt (+/- 250 nt centered around the cleavage site) as the input size because improvement in AUROC values was minimal for longer sequences.

Using a similar approach, we tested other model architecture parameters, including the number of dense layers, the number of convolutional and LSTM units, the shape of convolutional filters, and the dropout rate. The optimal parameter value was chosen to maximize the mean AUROC for the validation datasets for the 10 models trained. If multiple parameter values produced models within 1 standard error of the maximum AUROC, the most parsimonious parameter value was chosen. The final model configuration for each species was trained using the combination of the best parameters identified in the grid search and the final models are described in detail in Supplemental Table S2.

Developing an ensembled PolyaClassifier model using bagging

When working with severely imbalanced data, a common approach is to create an ensembled model using bagging (Khoshgoftaar et al. 2011). Multiple models are trained on resamples of the data and then the predictions are averaged to create a final "bagged" model prediction. This approach can improve model accuracy and robustness by leveraging the diversity of the available majority-class data, in this case, the large number of possible negative sequences (Galar et al. 2012). To investigate if bagging could improve PolyaClassifier performance, we sampled the positive poly(A) sequences and negative controls with replacement to create 20 independent "resamples" of the full dataset. We then used each resampled dataset to train a PolyaClassifier model using the species-specific configuration (Supplemental Table S2).

To choose the best number of models to bag together, we compared the improvement in AUROC as one more model was incorporated into the ensemble. For each number of models in the ensemble, we randomly selected 10 random groups of the specific number of models to ensemble and then calculated the change in AUROC over the mean of the previous number. For example, we randomly selected 10 groups of 3 models, calculated the concatenated AUROC for each, and then subtracted the mean AUROC from 10 random groups of 2 models. Using this approach, we found that bagging 3 models was the optimal number. When more than 3 models were combined, the 95% confidence interval of the mean delta AUROC crossed zero, indicating no consistent improvement in model performance.

To further examine the impact of negative control sampling on model performance, we randomly sampled another 10 random negative sets and positive sequences not included in the model training, and used our 3-bagged PolyaClassifer model in *S. cerevisiae* to calculate the AUROC values. The standard deviation was 0.0005, indicating the robustness of our model.

Examining the impact of positive site selections

The sequencing depth and the number of 3'READS available are variable for different species. Therefore, we chose appropriate read count thresholds to select high-quality cleavage sites for model training and downstream analysis. We required that high-quality cleavage sites show PASS reads ⩾2% of the most expressed site in the same gene and be supported by ⩾10 PASS reads in *S. cerevisiae* and *A. thaliana*, and by ⩾5 PASS reads in *S. pombe*. Further, we selected the most highly-expressed cleavage sites in a top-down fashion with the requirement that sites were ⩾5 nt apart. As a result, we included different numbers of positive poly(A) sequences in the model training across species.

To examine whether enough positive poly(A) sequences were included for building the PolyaClassifier model, we sampled the training data to include 1000, 2000, 4000, 10000, 20000, 40000, 60000, 100000, or all positive poly(A) sequences. We used these datasets to train a 3-bagged PolyaClassifier model for each training set size and calculated the AUROC values on the holdout test set. The AUROC values of our final model were all in the saturation phase, indicating that a sufficient number of positive sequences were used.

Cross-species PolyaClassifier performance

To evaluate the species-specificity of poly(A) signals, we made predictions for the testing set data using the *S. cerevisiae*, *S. pombe*, and *A. thaliana* PolyaClassifier models, as well as the human PolyaID model we previously developed (Stroup and Ji 2023). We then calculated the AUROC values for the model performances when each model was applied to predict the poly(A) sites across species.

Examining the performances of our deep learning models using 3P-seq and Helicos sequencing data

3P-seq data

We downloaded 3P-seq data for one wild-type S. cerevisiae sample SRR1049516 from GSE53310 (Subtelny et al. 2014). The data were processed using the steps as 3'READS. We required that high-quality cleavage sites were supported by ⩾10 3P-seq PASS reads and ⩾2% reads of the maximum expressed site in a gene. Then we applied an iterative, top-down sampling approach to select representative cleavage sites with a minimum distance ⩾5 nt from adjacent selected sites. We identified 15,808 representative cleavage sites that were used for model evaluation. To evaluate the performance of our bagged *S. cerevisiae* PolyaClassifier model on 3P-seq data, we combined the representative cleavage sites with an equal number of randomly sampled negative controls. We then made predictions for these sequences and calculated the AUROC and AUPRC values.

To evaluate the performance of our PolyaStrength model, we used the same approach as described for 3'READS to define poly(A) site clusters, and selected APA site pairs in 3'UTRs showing >8-fold expression differences. We used the PolyaStrength predicted scores to classify the highly vs. lowly expressed sites within a pair. AUROC and AUPRC values were calculated to evaluate the classification performance.

Helicos sequencing data

We downloaded normalized read count tracks of the sample GSM1959710 from GSE75587 (Roy et al. 2016). We selected cleavage sites with expression level ⩾1 RPM. After applying the iterative selection procedure, we found 21,006 representative cleavage sites ⩾5 nt from adjacent selected sites. We used similar downstream analysis steps as described for 3P-seq data to evaluate the performances of the PolyaClassifer and PolyaStrengh models.

The analyses of massively parallel reporter assays (MPRA)

A study used the massively parallel reporter assay that combined the HIS3 coding region with a 3'UTR derived from the *CYC1* gene in *S. cerevisiae*. They introduced hundreds of thousands of random 50 nt sequences upstream of a fixed cleavage site (Savinov et al. 2021). They determined relative HIS3 protein expression levels from the fitness of transformants using a growth selection. Based on their results, the polyadenylation activity is one major regulator of higher HIS3 expression, and RNA stability can also contribute to the regulation. We predicted the PolyaStrength score of each randomized sequence, padding with additional *HIS3* CDS sequence upstream and genomic sequence downstream to reach the 500 nt input length. We then partitioned the MPRA sequences into 4 groups based on the measured expression and calculated the AUROC by comparing the PolyaStrength scores classifying the highest (top 5%, 29,502 sequences) vs. lowest (bottom 5%, 29,529 sequences) expressed groups.

PolyaCleavage parameter search and model selection

For the PolyaCleavage model development, we started from the same pool of representative cleavage sites used in the PolyaClassifier model. The sites were split into 80% training, 10% validation, and 10% testing sets at the gene level. The training set was further split into 5 groups for 5-fold cross-validation. We used a parameter grid search approach, similar to that used for PolyaClassifier. We calculated the mean correlation between observed and predicted cleavage entropy and the correlation between observed and predicted mean cleavage position. The optimal value for each parameter was chosen to maximize these two performance measures. If multiple parameter values produced models within 1 standard error of the maximum correlation, the most parsimonious parameter value was chosen. The final PolyaCleavage model was trained using the combination of the best parameters identified by the grid search and is described in detail in Supplemental Table S2.

Examining genomic parameters regulating cleavage heterogeneity using the PolyaCleavage model.

We found that both the nucleotide composition around the cleavage site and the presence of upstream UA-rich motifs contribute to increased cleavage heterogeneity. To verify these findings using our PolyaCleavage model, we altered the poly(A) site compositions by adding or removing the *cis*-elements. For the low entropy sites, we randomly added non-overlapping AU-rich motifs into the (-15,+15) nt region around the cleavage site. The AU-rich elements included those identified through our motif enrichment analysis shown in Figure 3B and were most significant: AUAAUA, UAAUAA, AAUAAU, AAAAAU, or AAAAUA. We then measured the predicted changes to cleavage entropy using the PolyaCleavage model. We required that sites included in this analysis were in the bottom 20% low entropy group, and containing ⩾1 U-rich motif in both the (-15,0) and (0,15) regions immediately surrounding the cleavage site (N = 222 sites).

Inversely, for high entropy sites, we randomly added non-overlapping U-rich motifs to the -15 to +15 nt region around cleavage sites. In this analysis, we introduced the top U-rich motifs that were significantly enriched in low entropy sites from our motif analysis: UUUUUU, UUCUUU, UCUUUU, UUUCUU, and UUUUCU. We required that sites included in this analysis were in the top 20% high entropy sites, and containing ⩾1 UA-rich motif in both the (-15,0) and (0,15) regions immediately surrounding the cleavage site (N = 182 sites).

For each experiment, we randomly placed up to 4 AU-rich or U-rich motifs into the cleavage region one at a time without crossing the cleavage site. We repeated this random placement 100 times for each input site. We highlighted representative examples by showing the original sequence and PASS read distribution followed by the modified sequence and predicted cleavage vector after each sequential motif addition.

We also modified the number of upstream UA-rich elements to confirm the influence of these motifs on the cleavage heterogeneity. We sequentially disrupted the upstream efficiency elements of consensus high entropy sites with five existing UA-rich motifs in the (-90,-30) region. We sequentially replaced each UA-rich element with randomly sampled nucleotides until only one UA-rich motif was left and repeated this disruption 100 times. The randomly sampled nucleotides were chosen from a distribution matching the nucleotide distribution in the (-90,-30) region of consensus low entropy sites without creating new UA-rich motifs. We then measured the predicted change in cleavage entropy after modifying the sequence using the PolyaCleavage model. We highlighted a representative example by showing the original sequence and PASS read distribution followed by the modified sequence and predicted cleavage vector after each sequential motif removal.

PolyaStrength parameter search and model selection

To develop the PolyaStrength model, we split the set of 3'UTR APA sites into 70% training, 10% validation, and 20% testing sets at the gene level. The training set was further split into 5 groups to use for 5-fold cross-validation. We used a parameter grid search approach, similar to that used for PolyaClassifier described above. We calculated the mean paired site AUROC and the mean correlation between observed and predicted usage scores across all 5 cross-validation folds, as described above. The optimal value for each parameter was chosen to maximize these two performance measures. The final PolyaStrength model is described in detail in Supplemental Table S2.

Identifying motifs contributing to poly(A) site strength

We used the similar hexamer disruption and motif analysis approaches described for the PolyaClassifer model to identify motifs determining the poly(A) strength. We used 9,725 clustered poly(A) sites with ⩾100 reads and ⩾5% of the maximum expressed site in a gene. Next, we characterized the *cis*-regulatory elements that significantly contributed to poly(A) site strength using the two-step filtering procedure described above for the PolyaClassifier model, except we used the 99.9th percentile in each region during the second filtering step.

Studying the APA regulation under diauxic stress

To study the APA regulation under diauxic stress, we selected the top two expressed poly(A) site clusters in the 3'UTR or extended 3'UTR region of each coding gene by pooling the 3'READS measured under rich media and diauxic stress culture conditions. We required that the genes and poly(A) sites included in this analysis be well-expressed, with the gene supported by ⩾50 total PASS reads in the 3'UTR and the two poly(A) sites supported by ⩾10 PASS reads and located ⩾ 50 nt apart. We then compared the usage of the poly(A) site clusters under rich media and diauxic stress conditions by comparing the fraction of 3'READS supporting each poly(A) site. We selected genes showing significant APA under the stress condition comparing proximal vs. distal sites using the cutoff Benjamini-Hochberg correction *P*-value <0.05 (Fisher's exact test). For these genes, we calculated the proximal poly(A) site isoform abundance ratio as the ratio between the PASS read number supporting proximal sites vs. the sum reads of both proximal and distal sites. We next grouped poly(A) sites based on their differential usage levels between the diauxic stress vs. rich media conditions. We characterized these sites using the PolyaStrength scores, and the distances between the two sites.

We found that proximal poly(A) sites showing increased isoform ratios under diauxic stress tend to be weaker according to PolyaStrength and are paired with stronger distal sites. We examined the motif configurations around these sites to confirm the PolyaStrength predictions. We tabulated the frequency of UA-rich motifs in the (-90,-25) nt region, A-rich motifs in the (-25,-15) region, and U-rich motifs in the (-15,-6) and (2,15) nt regions. Motifs with a Hamming distance <= 2 nt compared with the archetypical motifs UAUAUA/AUAUAU, AAAAAA, and UUUUUU that were significant according to the PolyaStrength model were included in this analysis. The frequencies of the motifs in the indicated regions were compared using the Chi-squared test between the high Δratio group (>0.25) and the low Δratio group (-0.05 – 0.05).

We also confirmed the observed differences in motif frequencies using an unbiased motif enrichment analysis like that described during the cleavage heterogeneity analysis. We quantified the enrichment of 5-mers in the (-90,-25) nt region upstream of proximal and distal sites, comparing the high Δratio group (>0.25) and the low Δratio group (-0.05 – 0.05).

Examining the context of UAG- and GUA-elements in *S. pombe*.

To quantify the importance of the nucleotides surrounding UAG and GUA elements in *S. pombe* and to determine their most significant context, we padded "UAG" and "GUA" with all combinations of 3 nucleotides on each end to create 4096 possible 9-mers with UAG or GUA in the center. We then applied our systematic disruption approach to quantify the importance of each padded 9-mer. We examined UAG-containing 9mers where the UAG was found (-80,-30) nt upstream of the cleavage site. We also examined GUA-containing 9mers where the GUA was found (15,60) nt downstream of the cleavage site. For each position around the central UAG or GUA element, we grouped the motifs containing each nucleotide and calculated the mean importance score. These scores were combined and normalized to sum to 1 to create a position-probability matrix (PPM). We calculated the log-likelihood of the PPM assuming that all nucleotides are equally important and plotted the position-weight matrix for the nucleotides with positive values.

Studying poly(A) site sequence conservation across species

To examine the sequence conservation surrounding poly(A) sites across yeast and humans, we used the phastCons conservation tracks (Siepel et al. 2005). For humans, we used the 100-way alignment from UCSC. For *S. cerevisiae*, we used the 7-way alignment comparing *S. cerevisiae* to other *Saccharomyces* family members. For *S. pombe*, we used the 4-way alignment from the Fungal Genome Initiative comparing *S. pombe* to other *Schizosaccharomyces* family members (Rhind et al. 2011). We quantified the mean conservation score within coding regions and used this as a normalization factor. We performed the analyses for homologous genes across the three species defined by the PomBase database (Harris et al. 2022). We selected the top expressed poly(A) site from each homologous protein-coding gene (N = 3296 for *S. cerevisiae*, 2720 for *S. pombe*, and 3777 for *H. sapiens*). We calculated the averaged conservation scores and the 95% confidence interval values at each position normalized to the mean coding region conservation.

We also performed similar analyses for poly(A) sites from non-homologous genes (N = 2225 for *S. cerevisiae*, 1260 for *S. pombe*, and 12,325 for *H. sapiens*), and well-expressed sites grouped based on their relative genomic locations.

We further investigated the conservation of motif families surrounding top poly(A) sites in homologous protein-coding genes. We mapped the location of each hexamer surrounding these poly(A) sites and calculated the CDS-normalized mean conservation score. We grouped motifs that were significant according to PolyaClassifier into families and used not-significant motifs assigned to no motif family as the "other" group for reference. For *S. cerevisiae*, this included 40 UA-rich, 31 A-rich, and 51 U-rich, and 3535 other motifs. For *S. pombe*, we included 63 A-rich, 87 U-rich, 23 GUA-containing, 19 UAG-containing, 11 GUA+UAG-containing, and 2792 other motifs. We then plotted the distribution of CDS-normalized mean conservation scores for each motif family surrounding conserved top poly(A) sites.

Examining genetic variants impacting poly(A) motifs in *S. cerevisiae*

We analyzed genetic variants identified in 1,011 *S. cerevisiae* isolates (Peter et al. 2018). For significant poly(A) motifs defined by our deep learning models in *S. cerevisiae*, we examined their overlap with SNPs from the variant dataset. We grouped the motifs and variants by the distance to the cleavage sites and calculated the enrichment of overlapping variants relative to the background region +/- 1 kb around the poly(A) sites. At each position x around the cleavage sites, the enrichment statistic E for motif family f was calculated as:

$$E_{f,x} = \log_2 \left[\left(\frac{O_{f,x}}{N_{f,x}} \right) / \left(\frac{V_f}{M_f} \right) \right]$$

Where $O_{f,x}$ is the number of variants overlapping motifs from family f at position x, $N_{f,x}$ is the motif coverage for family f at position x, V_f is the number of variants overlapping motifs from family f in the wider background region, and M_f is the motif coverage for family f in the background region. The motif coverage calculates the number of motifs from family f that cross position x.

The analyses of *A. thaliana* poly(A) sites

We downloaded the 3'READS data for *A. thaliana* from the GEO database (Supplemental Table S1) (Guillermina Kubaczka et al. 2024). For read mapping, we used the Ensembl reference genome assembly TAIR10 and the corresponding gene annotation from release 58. The data analysis steps were the same as for *S. cerevisiae*. To build the PolyaClassifer model, the 500-nt sequences surrounding 47,618 representative cleavage sites were used as positives, and a similar grid search approach for *S. cerevisiae* was used to find the best model training parameters. The hexamer disruption approach was taken to identify the poly(A) motifs. The significant motifs we identified in *A. thaliana* can be grouped into four families: (1) the A-rich family which contained motifs with at least 4 As, (2) the U-rich family with motifs containing at least 4 Us and no UGUA, and (3) the UGUA-containing family. We then calculated the mean per-site and sum importance profiles for these families of motifs.

References

Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F. 2012. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* **42**: 463-484.

Guillermina Kubaczka M, Godoy Herz MA, Chen W-C, Zheng D, Petrillo E, Tian B, Kornblihtt AR. 2024. Light regulates widespread plant alternative polyadenylation through the chloroplast. *bioRxiv* doi:10.1101/2024.05.08.593009: 2024.2005.2008.593009.

Harris MA, Rutherford KM, Hayles J, Lock A, Bahler J, Oliver SG, Mata J, Wood V. 2022. Fission stories: using PomBase to understand Schizosaccharomyces pombe biology. *Genetics* **220**.

Khoshgoftaar TM, Van Hulse J, Napolitano A. 2011. Comparing Boosting and Bagging Techniques With Noisy and Imbalanced Data. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* **41**: 552-568.

Peter J, De Chiara M, Friedrich A, Yue JX, Pflieger D, Bergstrom A, Sigwalt A, Barre B, Freel K, Llored A et al. 2018. Genome evolution across 1,011 Saccharomyces cerevisiae isolates. *Nature* **556**: 339-344.

Rhind N, Chen Z, Yassour M, Thompson DA, Haas BJ, Habib N, Wapinski I, Roy S, Lin MF, Heiman DI et al. 2011. Comparative functional genomics of the fission yeasts. *Science* **332**: 930-936.

Roy K, Gabunilas J, Gillespie A, Ngo D, Chanfreau GF. 2016. Common genomic elements promote transcriptional and DNA replication roadblocks. *Genome Res* **26**: 1363-1375.

Savinov A, Brandsen BM, Angell BE, Cuperus JT, Fields S. 2021. Effects of sequence motifs in the yeast 3' untranslated region determined from massively parallel assays of random sequences. *Genome Biol* **22**: 293.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**: 1034-1050.

Stroup EK, Ji Z. 2023. Deep learning of human polyadenylation sites at nucleotide resolution reveals molecular determinants of site usage and relevance in disease. *Nat Commun* **14**: 7378.

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.

SUPPLEMENTAL FIGURES

SUPPLEMENTAL TABLES

Supplemental Table S1. The 3’-end sequencing datasets analyzed in this study.

Supplemental Table S2. The detailed architectures of deep learning models developed in this study.

Supplemental Table S3. The motifs showing significant importance scores from the PolyaClassifer model in *S. cerevisiae*.

Supplemental Table S4. The poly(A) sites identified in *S. cerevisiae*.

Supplemental Table S5. The motifs showing significant importance scores from the PolyaStrength model in *S. cerevisiae*.

Supplemental Table S6. The poly(A) sites identified *in S. Pombe*.

Supplemental Table S7. The motifs showing significant importance scores from the PolyaClassifer model in *S. Pombe*.

Supplemental Table S8. The poly(A) sites identified *in A. thaliana*.

Supplemental Table S9. The motifs showing significant importance scores from the PolyaClassifer model in *A. thaliana*.