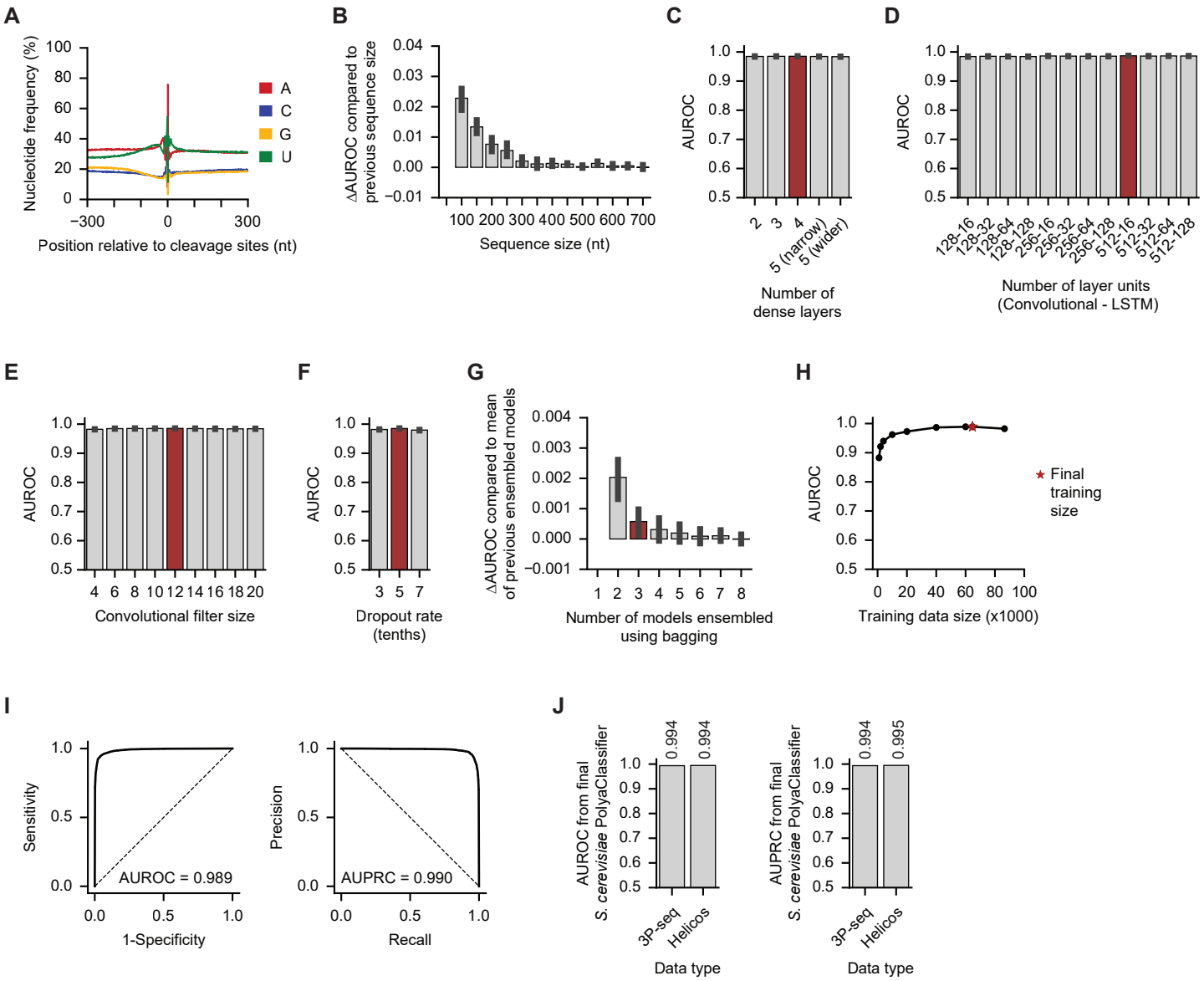
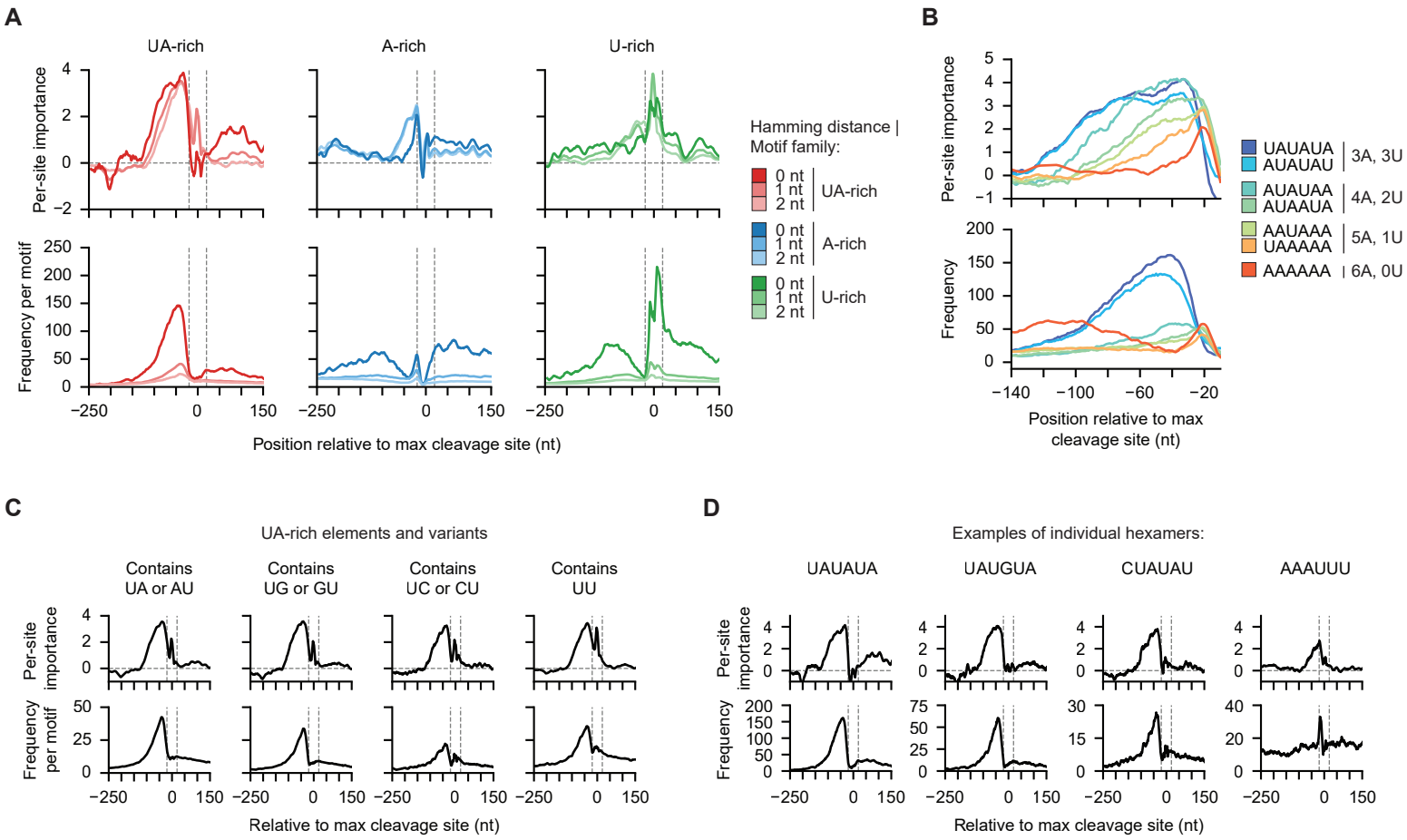
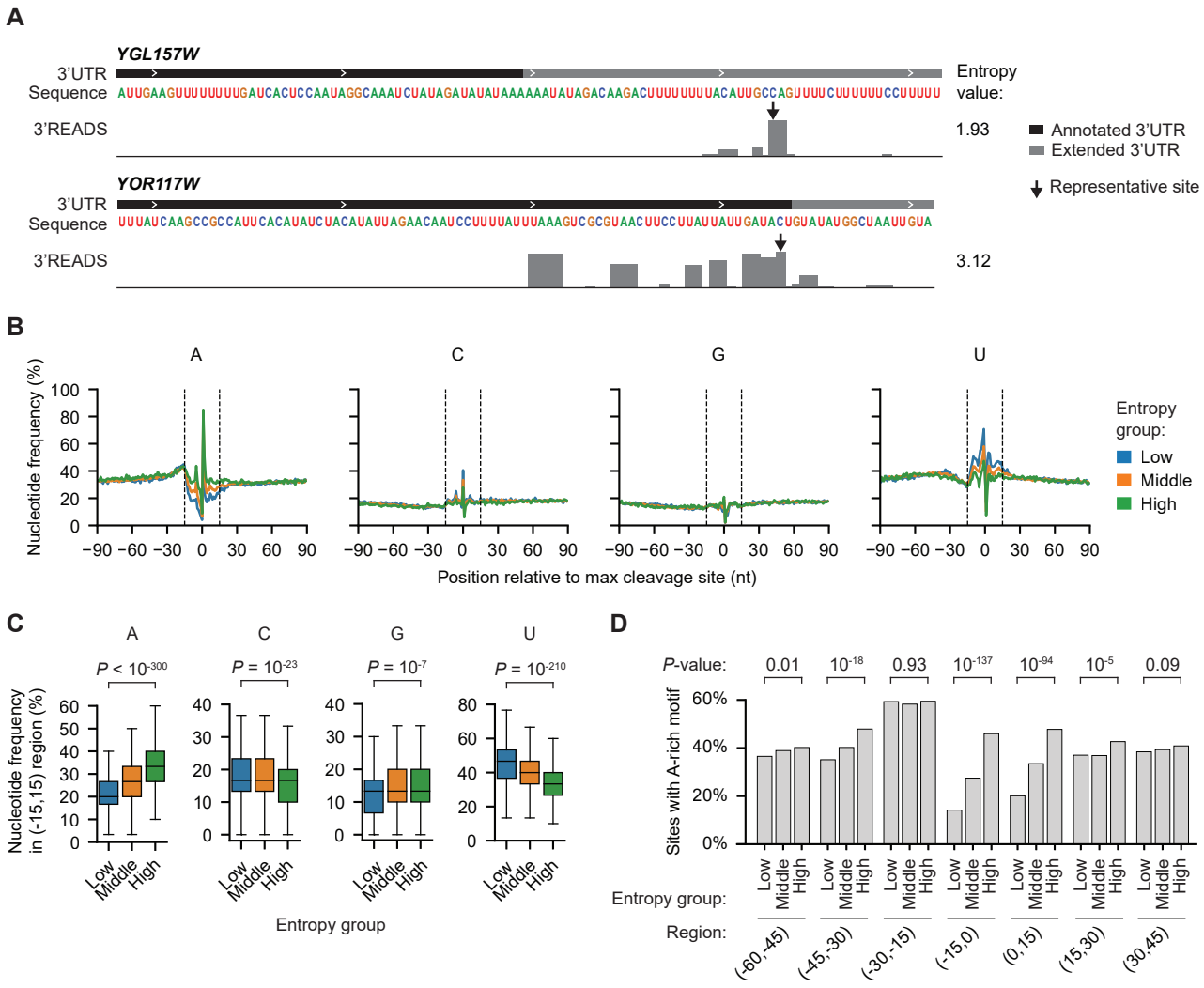


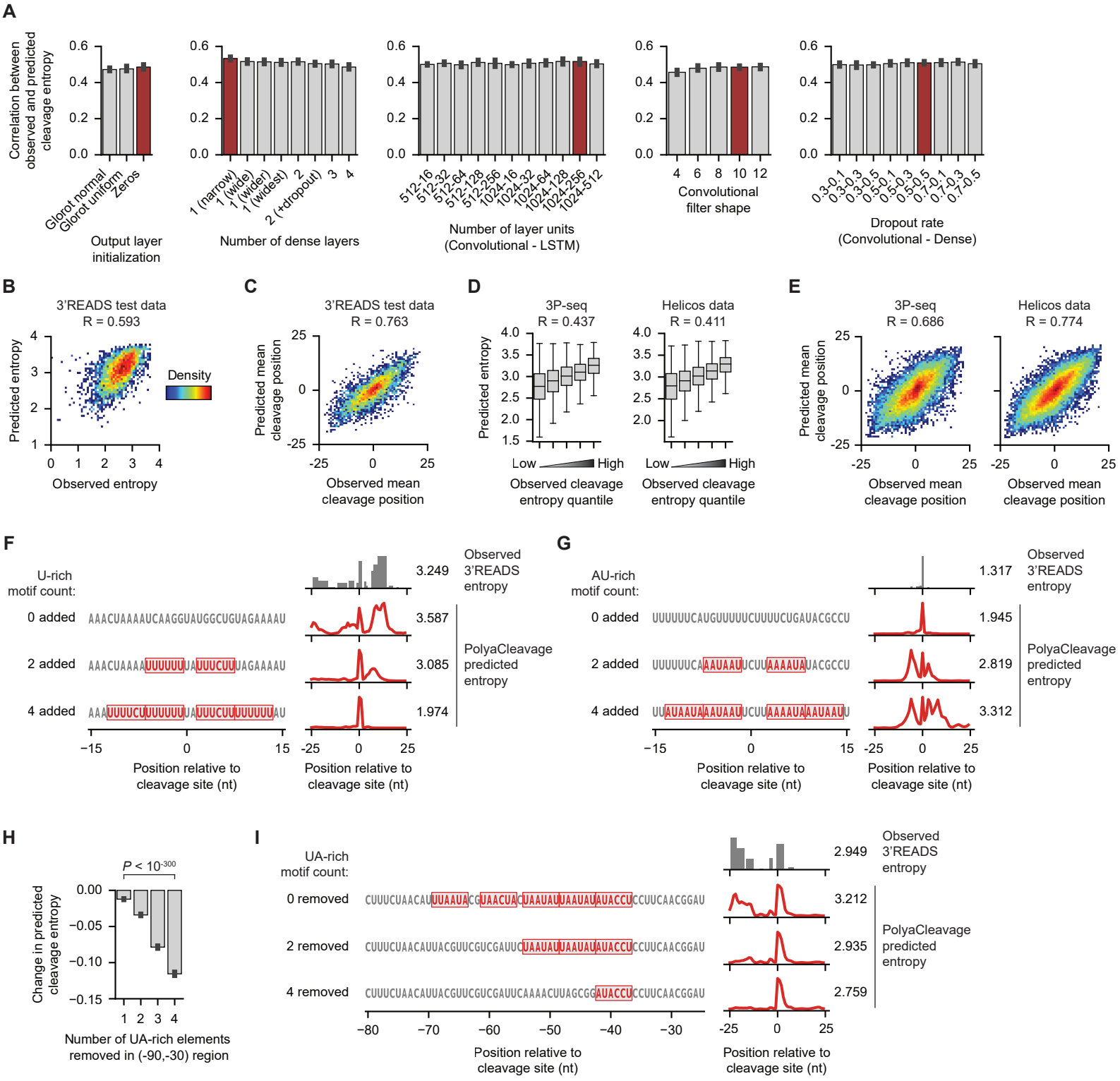
Supplemental Figure S1. *S. cerevisiae* PolyaClassifier model development and optimization.
(A) The distribution of nucleotide frequency for an extended region surrounding cleavage sites in *S. cerevisiae*.
(B) The change in AUROC as the input sequence size is expanded by 50 nt each step. The ΔAUROC is calculated relative to the previous sequence size. Data is shown as the mean and 95% confidence interval (error bar) of 10 models built based on different negative control sets.
(C) AUROC values as the model architecture was built based on different dense layers. Data is shown as the mean and 95% confidence interval (error bar) of 10 models trained on different negative control samples. The optimal value is highlighted in dark red.
(D) Similar to (C), except showing the AUROC values as the numbers of 1D-convolutional and bidirectional LSTM units are varied.
(E) Similar to (C), except showing the AUROC values with different 1D-convolutional filter sizes.
(F) Similar to (C), except showing the AUROC values with different dropout rates.
(G) The change in AUROC values as more models are combined using bagging to create an ensembled model. Data is shown as the mean and 95% confidence interval (error bar) of 10 unique model combinations. The optimal number of models to bag (3) is highlighted in dark red, after which the improvement in AUROC is minimal.
(H) AUROC values as the number of input training sequences is varied. The number of sequences used in our final model is marked in red.
(I) The receiver operating characteristic curve (left) and the precision-recall curve (right) measuring the performance of our final *S. cerevisiae* PolyaClassifier bagged model on the testing. The areas under the curve are shown.
(J) The AUROC and AUPRC values measuring the performance of *S. cerevisiae* PolyaClassifier model on cleavage sites defined by alternative 3'-end sequencing technologies including 3P-seq and Helicos sequencing data.



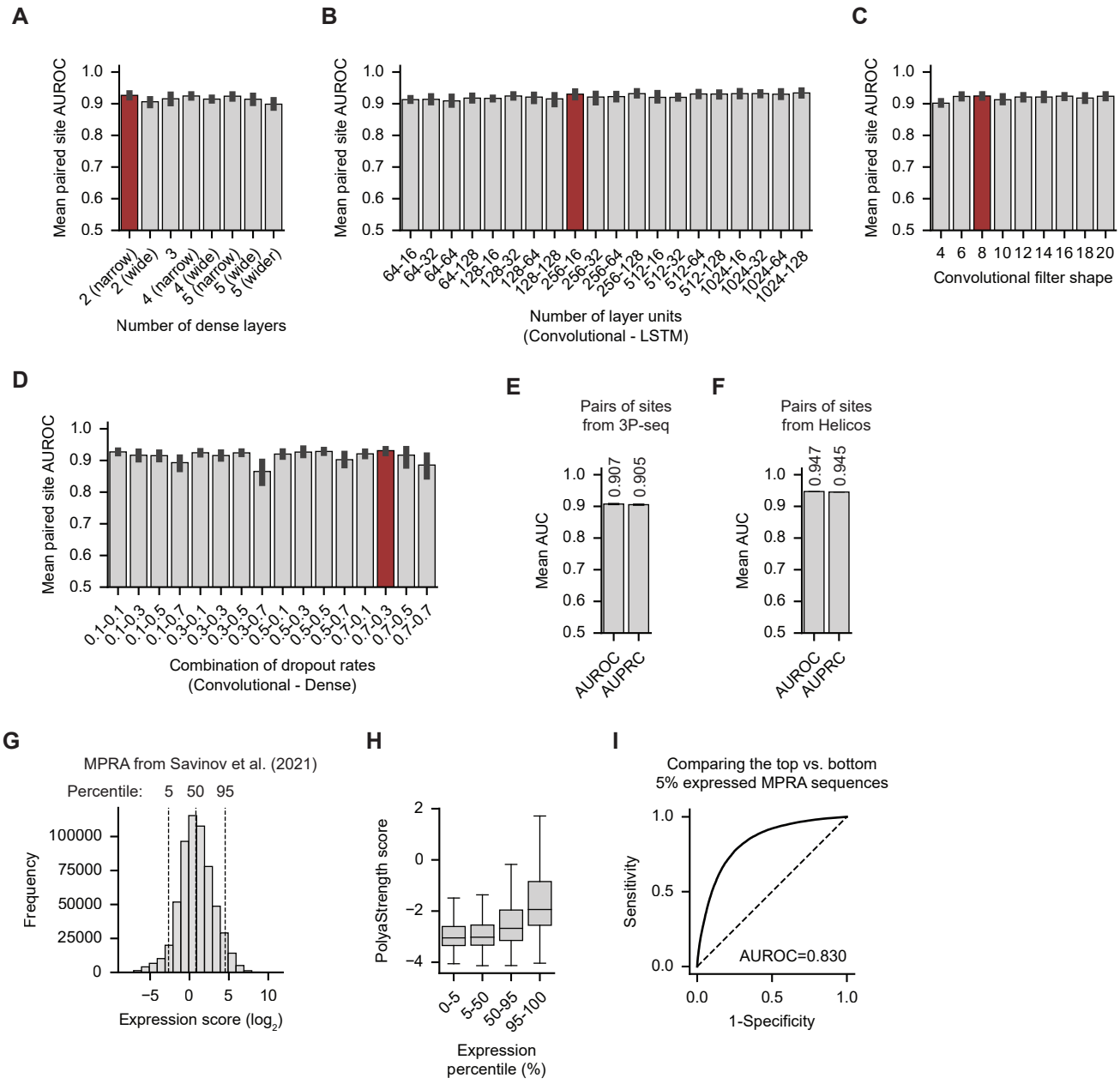
Supplemental Figure S2. Examine motif importance to poly(A) site identification by motif family.
(A) The per-site importance (top) and frequency (bottom) profiles for the *S. cerevisiae* motif families. In each family, the importance profile is grouped by the Hamming distance to the representative motif (0, 1, or 2 nt). The dashed vertical lines mark -20 and 20 nt around the cleavage site.
(B) Examples of per-site importance (top) and frequency (bottom) for individual motifs grouped by the number of A and Us.
(C) The per-site importance (top) and frequency (bottom) profiles for the UA-rich motifs. Motifs are grouped by whether they contain nucleotide variants. The dashed vertical lines mark -20 and 20 nt around the cleavage site.
(D) The per-site importance (top) and frequency (bottom) of example UA-rich motifs. The dashed vertical lines mark -20 and 20 nt around the cleavage site.



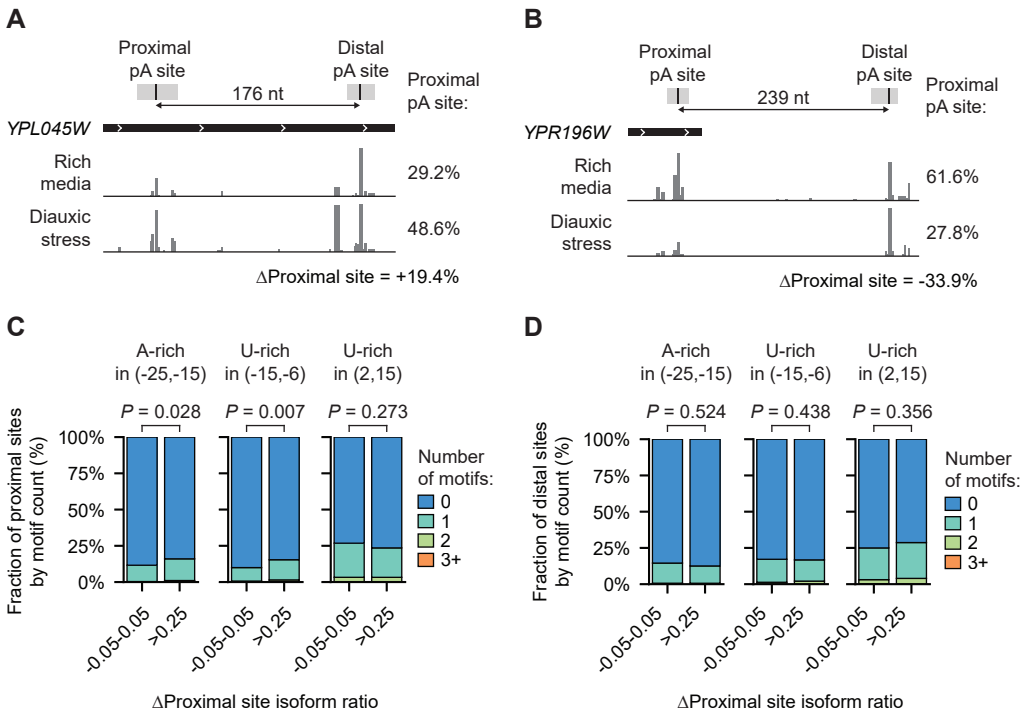
Supplemental Figure S3. Motif analyses of poly(A) sites from entropy groups.
(A) Examples of the PASS read distribution surrounding a low entropy site (top) and high entropy site (bottom). The observed entropy values are shown.
(B) The nucleotide frequency surrounding the maximum cleavage site grouped by cleavage entropy. The vertical dashed lines mark -15 nt and 15 nt.
(C) The quantification of nucleotide density in the (-15,15) nt region grouped by cleavage entropy. The *P*-values from the Wilcoxon rank-sum test comparing low vs. high entropy sites are shown.
(D) The fraction of sites in each cleavage entropy group that contain an A-rich motif in the noted regions. The *P*-value from the two proportions hypothesis test is shown.


A
Correlation between observed and predicted cleavage entropy
Glorot normal
Glorot uniform
Zeros
Output layer initialization
1 (narrow)
1 (wide)
1 (wider)
1 (widest)
2
2 (+dropout)
3
4
Number of dense layers
512-16
512-32
512-64
512-128
512-256
1024-16
1024-32
1024-64
1024-128
1024-256
1024-512
Number of layer units (Convolutional - LSTM)
Convolutional filter shape
Dropout rate (Convolutional - Dense)
B
3'READS test data
R = 0.593
Density
Predicted entropy
Observed entropy
C
3'READS test data
R = 0.763
Predicted mean cleavage position
Observed mean cleavage position
D
3P-seq
R = 0.437
Helicos data
R = 0.411
Predicted entropy
Low
High
Observed cleavage entropy quantile
E
3P-seq
R = 0.686
Helicos data
R = 0.774
Predicted mean cleavage position
Observed mean cleavage position
F
U-rich motif count:
0 added
2 added
4 added
Observed 3'READS entropy
PolyaCleavage predicted entropy
Position relative to cleavage site (nt)
G
AU-rich motif count:
0 added
2 added
4 added
Observed 3'READS entropy
PolyaCleavage predicted entropy
Position relative to cleavage site (nt)
H
P < 10-300
Change in predicted cleavage entropy
Number of UA-rich elements removed in (-90,-30) region
I
UA-rich motif count:
0 removed
2 removed
4 removed
Observed 3'READS entropy
PolyaCleavage predicted entropy
Position relative to cleavage site (nt)

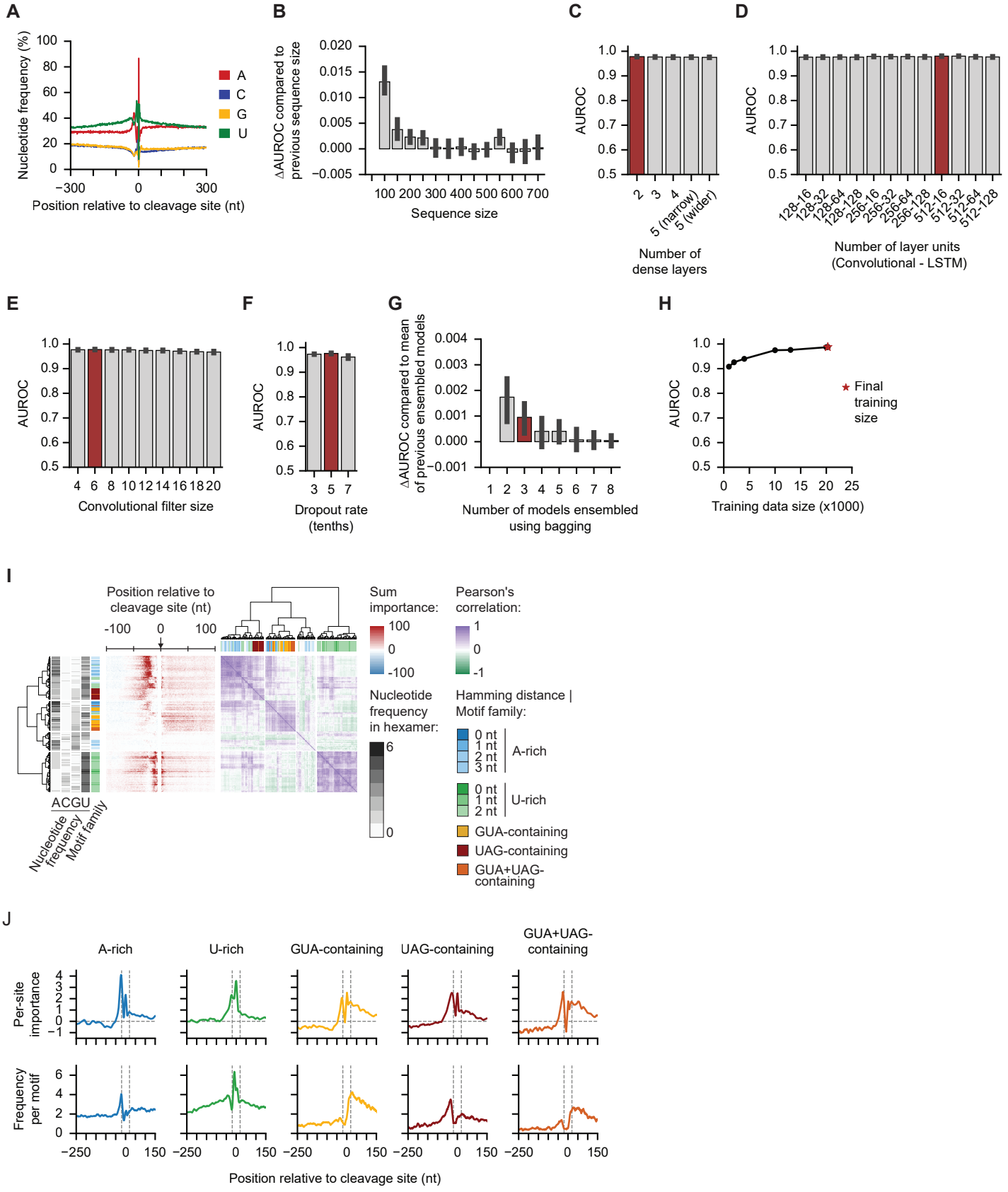
Supplemental Figure S4. Modeling cleavage heterogeneity using the PolyaCleavage model.
(A) The correlation between observed and predicted cleavage entropies in the holdout testing set (N = 4046 sites) as model parameters are varied one at a time. The Y-axis represents the Pearson's correlation coefficient. Data is shown as the mean and 95% confidence interval (error bar) using 5-fold cross-validation. The optimal value for each parameter is highlighted in dark red.
(B) 2D heatmaps showing the correlation between the observed and predicted entropy scores measuring the cleavage heterogeneities in the 3'READS holdout testing set.
(C) 2D heatmaps showing the correlation between the observed and predicted mean cleavage position in the 3'READS holdout testing set.
(D) The relationship between observed and predicted cleavage entropy for 3P-seq and Helicos sequencing-defined poly(A) sites. The sites were split into 5 evenly-sized groups. The Pearson's correlation coefficient is noted.
(E) 2D heatmaps showing the correlation between the observed and predicted mean cleavage position in poly(A) sites defined by 3P-seq and Helicos sequencing. The Pearson's correlation coefficient is noted.
(F) An example showing the effect of adding U-rich motifs in the (-15,+15) nt region around the cleavage site in the *YMR080C* gene. The added U-rich motifs are highlighted by red boxes. The observed PASS reads surrounding the site are shown in gray (top right) and the PolyaCleavage predictions for the original and altered sequences are shown in red (right).
(G) Similar to (F), except showing the effect of adding AU-rich motifs in the (-15,+15) nt region around the cleavage site in the *YNL146W* gene. The added AU-rich motifs are highlighted by red boxes. The observed PASS reads surrounding the site are shown in gray (top right) and the PolyaCleavage predictions for the original and altered sequences are shown in red (right).
(H) The changes in predicted cleavage entropy as upstream UA-rich elements are sequentially disrupted in the (-90,-30) region. High entropy sites with 5 upstream UA-rich motifs in this region were included (N = 305 actual cleavage sites, 122,000 altered sequences). The data is shown as the mean and the 95% confidence interval (error bar). The *P*-value from the Wilcoxon signed-rank paired test comparing sequences with 1 motif removed vs. 4 motifs removed is shown.
(I) An example showing the effect of disrupting upstream UA-rich motifs in the *YLR180W* gene. The removed UA-rich motifs are highlighted by red boxes. The observed 3'READS surrounding the site are shown in gray (top right) and the PolyaCleavage predictions for the original and altered sequences are shown in red (right).



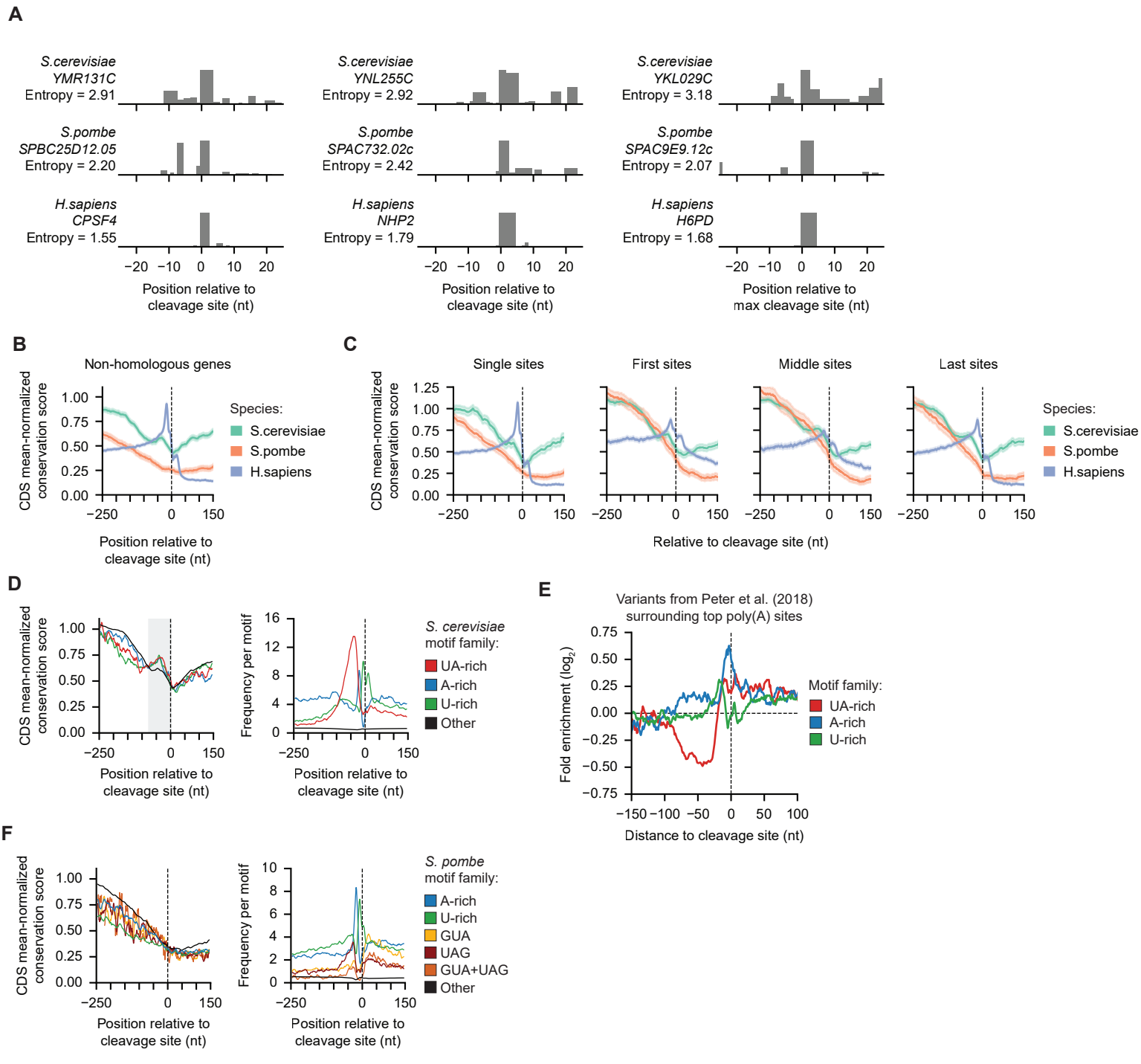
Supplemental Figure S5. The evaluation of PolyaStrength model performance.
(A-D) The results from the PolyaStrength model parameter grid search, which varied the number of dense layers (A), the number of convolutional and LSTM units (B), the convolutional filter shape (C), and dropout rates (D). The data is shown as the mean paired-site AUROC and 95% confidence interval from 5-fold cross-validation. The best parameter value is highlighted in dark red. The paired-site AUROC is calculated using the PolyaStrength score to separate highly vs. lowly expressed 3'UTR APA sites.
(E-F) The performance of PolyaStrength scores separate highly vs. lowly expressed 3'UTR APA sites (expression level difference of the paired sites >8-fold) defined by 3P-seq (E) and Helicos sequencing (F). The mean and standard deviation (error bar) AUROC and AUPRC values from 10 sets of paired sites are shown.
(G) The distribution of expression scores for MPRA sequences. We split the sequences into 4 groups based on the $\log_2$-expression score.
(H) The boxplot showing the predicted PolyaStrength scores of the MPRA sequences grouped in (G).
(I) The ROC curve showing the classification performance of PolyaStrength scores to distinguish highly vs. lowly-expressed MPRA sequences (top 5% vs. bottom 5% expressed sequences; N = 29,502 and 29,529, respectively).



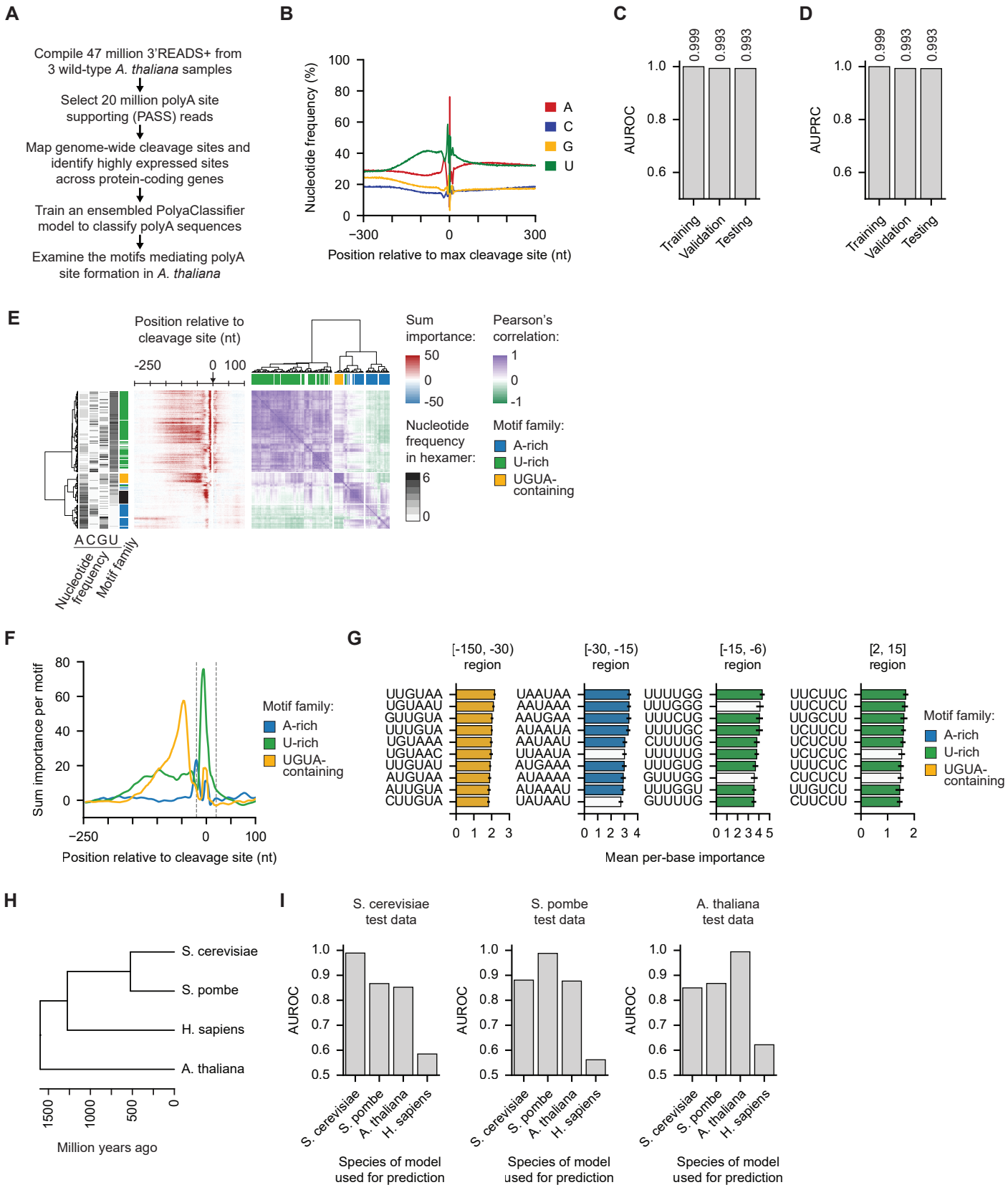
Supplemental Figure S6. APA regulation under the diauxic stress and motif configuration analysis.
(A) An example gene *YPL045W* shows moderately increased proximal poly(A) site isoforms under diauxic stress. The proximal and distal poly(A) site clusters are shown above in gray with the maximum cleavages site in each cluster marked with a black line. The distance between maximum cleavage sites is noted. The isoform expression of the proximal poly(A) site in each condition is shown.
(B) Similar to (A), showing the *YPR196W* gene which demonstrates decreased proximal poly(A) site isoforms under diauxic stress and a distal shift.
(C-D) Stacked bar chart showing the number of A-rich and U-rich elements present in the specified regions surrounding proximal (C) and distal (D) sites grouped by Δproximal site isoform ratio. The *P*-values from the Chi-squared tests are shown.


A
Nucleotide frequency (%)
A
C
G
U
Position relative to cleavage site (nt)
B
ΔAUROC compared to previous sequence size
Sequence size
C
AUROC
Number of dense layers
D
AUROC
Number of layer units (Convolutional - LSTM)
E
AUROC
Convolutional filter size
F
AUROC
Dropout rate (tenths)
G
ΔAUROC compared to mean of previous ensembled models
Number of models ensembled using bagging
H
AUROC
Final training size
Training data size (x1000)
I
Position relative to cleavage site (nt)
Sum importance:
Pearson's correlation:
Nucleotide frequency in hexamer:
Hamming distance | Motif family:
A-rich
U-rich
GUA-containing
UAG-containing
GUA+UAG-containing
ACGU
Nucleotide frequency
Motif family
J
A-rich
U-rich
GUA-containing
UAG-containing
GUA+UAG-containing
Per-site importance
Frequency per motif
Position relative to cleavage site (nt)

Supplemental Figure S7. Development of the *S. pombe* PolyaClassifier model.
(A) The distribution of nucleotide frequency for an extended region surrounding cleavage sites in *S. pombe*.
(B) The change in AUROC values as the input sequence size is expanded by 50 nt each step. The ΔAUROC is calculated relative to the previous sequence size. Data is shown as the mean and 95% confidence interval (error bar) of 10 models trained based on different sets of negative controls.
(C) The AUROC values as the model is trained based on different dense layers. Data is shown as the mean and 95% confidence interval (error bar) of 10 models. The optimal value is highlighted in dark red.
(D) Similar to (C), except showing the change in model performance as the numbers of 1D-convolutional and bidirectional LSTM units are varied.
(E) Similar to (C), except showing the change in model performance as the sizes of the 1D-convolutional filters are varied.
(F) Similar to (C), except showing the change in model performance as the dropout rate is varied.
(G) The change in AUROC values as more models are combined using bagging to create an ensembled model. Data is shown as the mean and 95% confidence interval (error bar) of 10 unique model combinations. The optimal number of models to bag (3) is highlighted in dark red, after which the improvement in AUROC is minimal.
(H) The AUROC values as the size of the input training dataset is varied. The input size of our final model is marked in dark red.
(I) Heatmaps showing the sum classification importance (left) and Pearson's correlation between classification importance profiles (right) for cis-regulatory elements significantly contributing to polyA site definition in *S. pombe* (N = 230). Each row represents a hexamer. The A-rich and U-rich motif families are defined by the Hamming distance between the motif of interest and the archetypical motifs AAAAAA or UUUUUU, respectively. Families for motifs containing GUA, UAG, or both are also shown.
(J) The per-site importance (top) and frequency (bottom) profiles for A-rich, U-rich, GUA-containing, UAG containing, and GUA+UAG-containing motifs that significantly contribute to poly(A) site definition in *S. pombe*.


A
S.cerevisiae YMR131C Entropy = 2.91
S.cerevisiae YNL255C Entropy = 2.92
S.cerevisiae YKL029C Entropy = 3.18
S.pombe SPBC25D12.05 Entropy = 2.20
S.pombe SPAC732.02c Entropy = 2.42
S.pombe SPAC9E9.12c Entropy = 2.07
H.sapiens CPSF4 Entropy = 1.55
H.sapiens NHP2 Entropy = 1.79
H.sapiens H6PD Entropy = 1.68
−20 −10 0 10 20
Position relative to cleavage site (nt)
Position relative to max cleavage site (nt)
B
Non-homologous genes
CDS mean-normalized conservation score
1.00 0.75 0.50 0.25 0.00
−250 0 150
Position relative to cleavage site (nt)
Species:
S.cerevisiae
S.pombe
H.sapiens
C
Single sites
First sites
Middle sites
Last sites
1.25 1.00 0.75 0.50 0.25 0.00
Relative to cleavage site (nt)
D
Frequency per motif
16 12 8 4 0
S. cerevisiae motif family:
UA-rich
A-rich
U-rich
Other
E
Variants from Peter et al. (2018) surrounding top poly(A) sites
Fold enrichment (log2)
0.75 0.50 0.25 0.00 −0.25 −0.50 −0.75
−150 −100 −50 0 50 100
Distance to cleavage site (nt)
Motif family:
UA-rich
A-rich
U-rich
F
10 8 6 4 2 0
S. pombe motif family:
A-rich
U-rich
GUA
UAG
GUA+UAG
Other

Supplemental Figure S8. Analyses of yeast poly(A) site cleavage heterogeneity and site conservation.
(A) Additional examples of the PASS read distribution surrounding the top poly(A) site in homologous genes across *S. cerevisiae*, *S. pombe*, and *H. sapiens*. The observed cleavage entropy values are shown.
(B) A meta plot showing the conservation score for the (-250,150) nt region surrounding the top expressed poly(A) sites in all non-homologous genes across the three species (N = 2225 *S. cerevisiae* sites, 1260 *S. pombe* sites, and 12325 *H. sapiens* sites). The conservation score was normalized to the mean score in coding regions. The data is shown as the mean with the shaded region indicating the 95% confidence interval at each position.
(C) Similar to (B), except for well-expressed poly(A) sites grouped by their relative position in the gene.
(D) A metaplot showing the conservation score (left) and per-motif frequency (right) for motifs within (-250,150) nt surrounding the top expressed poly(A) site in *S. cerevisiae*. Significant motifs from the PolyaClassifier analysis were included. Motifs are grouped by family (N = 40 UA-rich, 31 A-rich, and 51 U-rich). “Other” motifs were those not assigned to a family and not significant according to PolyaClassifier. Locations with missing conservation scores were excluded. The (-75,0) nt region upstream of the cleavage site is highlighted and the Wilcoxon rank-sum test $P < 10^{-33}$ comparing UA-rich, A-rich, and U-rich elements vs. other motifs.
(E) The relative enrichment of genetic variants affecting UA-rich, A-rich, and U-rich motifs surrounding the top cleavage sites in *S. cerevisiae*. The Y-axis represents the $\log_2$(enrichment ratio) value.
(F) A metaplot showing the conservation score (left) and per-motif frequency (right) for motifs within (-250,150) nt surrounding the top expressed poly(A) site in *S. pombe*. Significant motifs from the PolyaClassifier analysis were included (N = 63 A-rich, 87 U-rich, 23 GUA-containing, 19 UAG-containing, and 11 GUA+UAG-containing motifs). “Other” motifs were those not assigned to a family and not significant according to PolyaClassifier.


A
Compile 47 million 3'READS+ from 3 wild-type A. thaliana samples
Select 20 million polyA site supporting (PASS) reads
Map genome-wide cleavage sites and identify highly expressed sites across protein-coding genes
Train an ensembled PolyaClassifier model to classify polyA sequences
Examine the motifs mediating polyA site formation in A. thaliana
B
Nucleotide frequency (%)
Position relative to max cleavage site (nt)
A
C
G
U
C
AUROC
0.999
0.993
0.993
Training
Validation
Testing
D
AUPRC
E
Position relative to cleavage site (nt)
Sum importance:
Pearson's correlation:
Nucleotide frequency in hexamer:
Motif family:
A-rich
U-rich
UGUA-containing
Nucleotide frequency
Motif family
F
Sum importance per motif
Position relative to cleavage site (nt)
G
[-150, -30) region
[-30, -15) region
[-15, -6) region
[2, 15] region
Mean per-base importance
H
S. cerevisiae
S. pombe
H. sapiens
A. thaliana
Million years ago
I
S. cerevisiae test data
S. pombe test data
A. thaliana test data
Species of model used for prediction

Supplemental Figure S9. The deep learning modeling of *A. thaliana* poly(A) sites.
(A) Overview of the data processing workflow to train the *A. thaliana* PolyaClassifier model.
(B) Distribution of nucleotide frequency surrounding the representative cleavage sites in *A. thaliana* (N = 47,618).
(C) The AUROC values showing the constituent model performance for the training, validation, and testing set.
(D) The AUPRC values showing the constituent model performance for the training, validation, and testing set.
(E) Heatmaps showing the sum classification importance (left) and Pearson's correlation between classification importance profiles (right) for *cis*-regulatory elements significantly contributing to poly(A) site definition in *A. thaliana* (N = 255). Each row represents a hexamer.
(F) The per-motif sum classification importance profiles centered at the maximum cleavage site (N = 66 A-rich, 133 U-rich, and 17 UGUA-containing motifs).
(G) Bar plots showing the per-site importance of the top 10 motifs in each region surrounding the maximum cleavage site. Bars are colored by the family to which the motif belongs. Data is presented as the mean and the 95% confidence interval (error bar).
(H) Phylogenetic tree showing the evolutionary relationship between humans, yeast, and *A. thaliana*.
(I) Barplots showing the AUROC values for each species' holdout test set using predictions from species-specific PolyaClassifier models. For humans, the PolyaID model we built before was used.