Supplemental Figures of

Predicting dynamic expression patterns in budding yeast with a fungal DNA language model

Kuan-Hao Chao[1, 2, *], Majed Mohamed Magzoub[3], Emily Stoops[3], Sean Hackett[3], Johannes Linder [3, *] and David R. Kelley[3, *]

[1]Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA

[2]Center for Computational Biology, Johns Hopkins University, Baltimore, MD 21218, USA

[3]Calico Life Sciences LLC, South San Francisco, CA 94080, USA

[*]*Corresponding authors:* kuanhao.chao@gmail.com, jlinder@calicolabs.com, drk@calicolabs.com

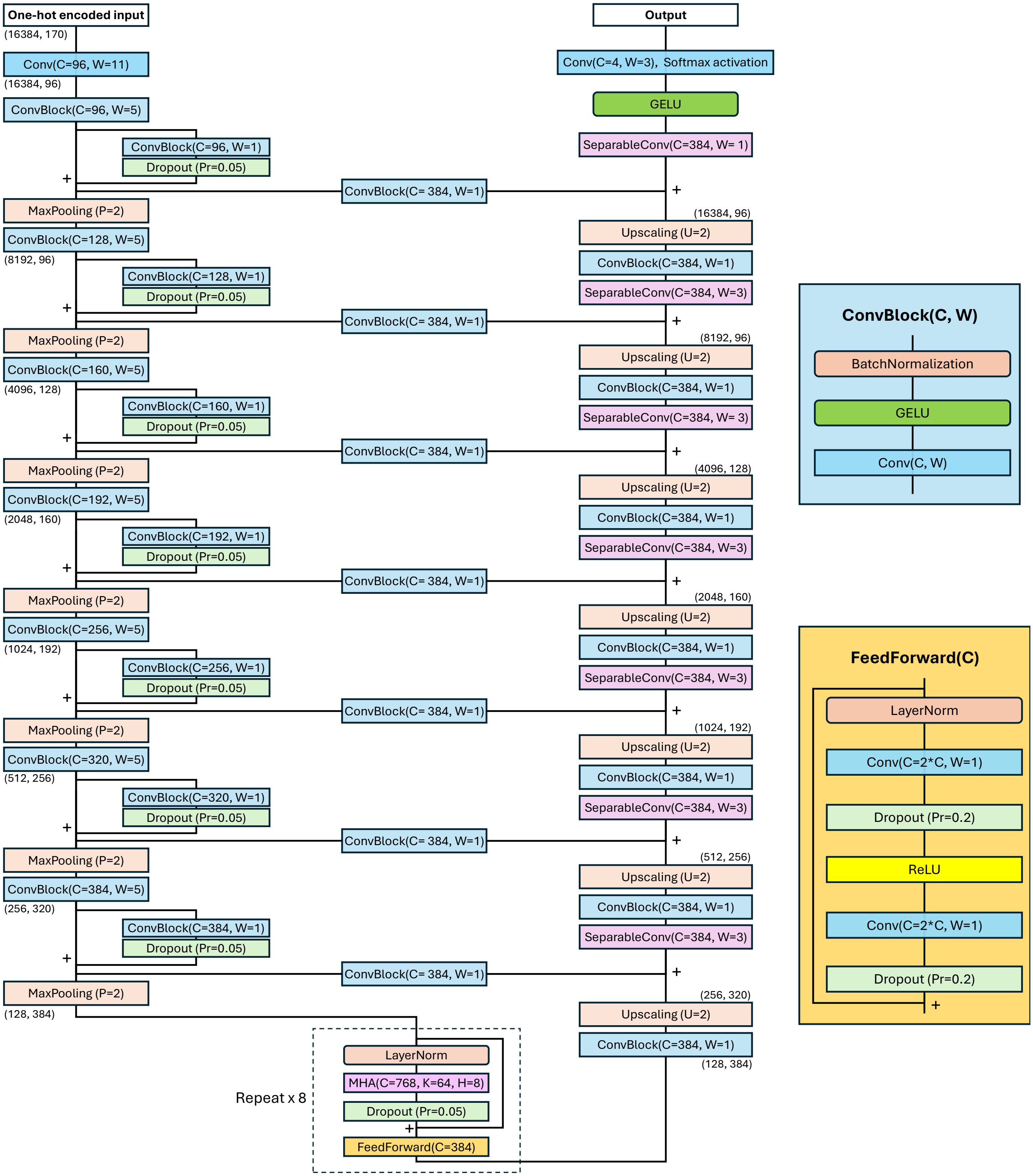


Figure S1. Architecture of the Shorkie LM model. The 13.7-million-parameter model predicts per-base nucleotide probabilities from genomic windows of 16,384 bp. Input sequences are augmented with stochastic shifts and passed through a 1D convolution layer (kernel size = 11, filters = 96), generating a 16,384 × 96 feature map. The encoder consists of seven residual down-sampling blocks, reducing the map to 128 × 384. This is followed by an eight-layer Transformer bottleneck that integrates long-range sequence context. The decoder mirrors the encoder with seven up-sampling stages, U-Net skip connections, and depthwise-separable convolutions, ultimately restoring the sequence to its original length of 16,384 positions. A final 1×1 convolutional layer and a softmax activation function across four channels produce per-position nucleotide probability distributions for A/C/G/T.

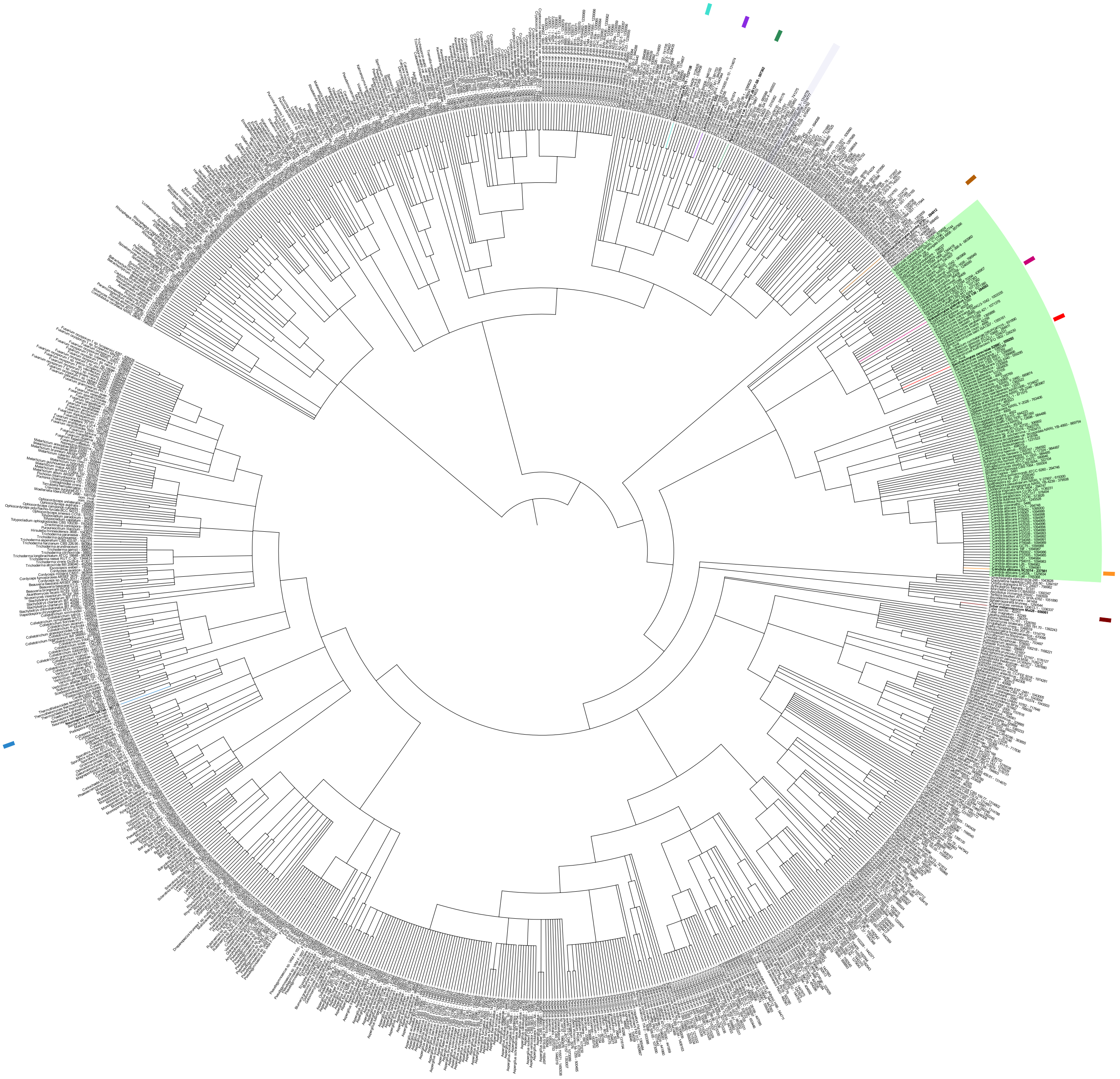


Figure S2. Phylogenetic tree of 1,341 fungal genomes. Species were assigned to NCBI Taxonomy IDs using the ETE Toolkit, and a minimum spanning tree of the target taxa was constructed with ete-ncbiquery. The tree, formatted in Newick, was visualized using iTOL v5 with a circular layout and annotated with color-coding for taxonomic groups. The order Saccharomycetales is highlighted in light green. Example annotated species include *Pleurotus ostreatus* (oyster mushroom) in cyan, *Lentinula edodes* (shiitake) in purple, *Agaricus bisporus* (white button mushroom) in green, *Tuber melanosporum* (black truffle) in brown, *Schizosaccharomyces pombe* (fission yeast) in orange, and *Saccharomyces cerevisiae* (brewer's yeast) in red.

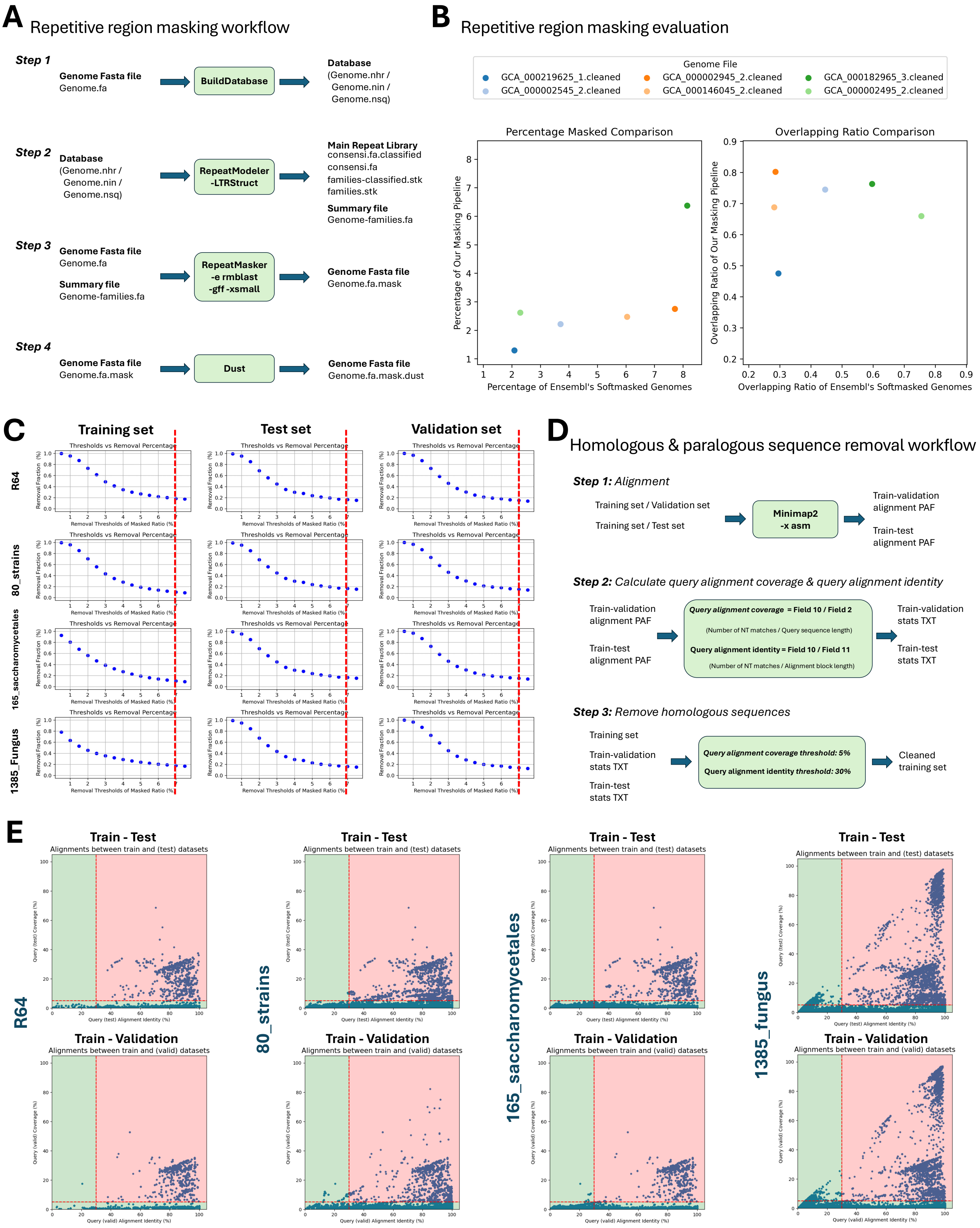

A
Repetitive region masking workflow
Step 1
Genome Fasta file
Genome.fa
BuildDatabase
Database
(Genome.nhr /
Genome.nin /
Genome.nsq)
Step 2
Database
(Genome.nhr /
Genome.nin /
Genome.nsq)
RepeatModeler
-LTRStruct
Main Repeat Library
consensi.fa.classified
consensi.fa
families-classified.stk
families.stk
Summary file
Genome-families.fa
Step 3
Genome Fasta file
Genome.fa
Summary file
Genome-families.fa
RepeatMasker
-e rmblast
-gff -xsmall
Genome Fasta file
Genome.fa.mask
Step 4
Genome Fasta file
Genome.fa.mask
Dust
Genome Fasta file
Genome.fa.mask.dust
B
Repetitive region masking evaluation
Genome File
● GCA_000219625_1.cleaned
● GCA_000002545_2.cleaned
● GCA_000002945_2.cleaned
● GCA_000146045_2.cleaned
● GCA_000182965_3.cleaned
● GCA_000002495_2.cleaned
Percentage Masked Comparison
Percentage of Our Masking Pipeline
Percentage of Ensembl's Softmasked Genomes
Overlapping Ratio Comparison
Overlapping Ratio of Our Masking Pipeline
Overlapping Ratio of Ensembl's Softmasked Genomes
C
Training set
Test set
Validation set
R64
80_strains
165_saccharomycetales
1385_Fungus
Thresholds vs Removal Percentage
Removal Fraction (%)
Removal Thresholds of Masked Ratio (%)
D
Homologous & paralogous sequence removal workflow
Step 1: Alignment
Training set / Validation set
Training set / Test set
Minimap2
-x asm
Train-validation
alignment PAF
Train-test
alignment PAF
Step 2: Calculate query alignment coverage & query alignment identity
Train-validation
alignment PAF
Train-test
alignment PAF
Query alignment coverage = Field 10 / Field 2
(Number of NT matches / Query sequence length)
Query alignment identity = Field 10 / Field 11
(Number of NT matches / Alignment block length)
Train-validation
stats TXT
Train-test
stats TXT
Step 3: Remove homologous sequences
Training set
Train-validation
stats TXT
Train-test
stats TXT
Query alignment coverage threshold: 5%
Query alignment identity threshold: 30%
Cleaned
training set
E
R64
80_strains
165_saccharomycetales
1385_fungus
Train - Test
Alignments between train and (test) datasets
Query (test) Coverage (%)
Query (test) Alignment Identity (%)
Train - Validation
Alignments between train and (valid) datasets
Query (valid) Coverage (%)
Query (valid) Alignment Identity (%)

Figure S3. Shorkie LM data preprocessing and homology filtering. **(A)** Two-tiered repeat masking approach: Each fungal genome from Ensembl was soft-masked using a custom library built with RepeatModeler v2.0 (incorporating RepeatScout, RECON, and LTR_retriever) merged with Dfam entries. RepeatMasker (RMBlast) was used with the "`-xsmall`" and "`-gff`" options to soft-mask interspersed repetitive elements, while residual low-complexity regions were masked using DUST from the MEME suite (with a score threshold of 20). **(B)** Comparison of the custom repeat masks against the original soft-masking from Ensembl in six representative genome assemblies. **(C)** Fully masked genomes were divided into 16,384 bp windows with a 4,096 bp stride; windows containing >7% masked content were excluded, resulting in the removal of approximately 20% of windows across all data splits. **(D)** Homology filtering to avoid data leakage: All training windows were aligned against the validation and test sets using Minimap2 ("`-x asm`"). For each alignment, coverage (percentage of query length covered) and identity (percentage of alignment matching the query) were computed. Any training window with ≥ 5% coverage and ≥ 30% identity to any validation or test window was discarded. **(E)** Scatter plots of coverage versus identity for all training-validation and training-test alignments, highlighting the thresholds used to define non-redundant training sequences.

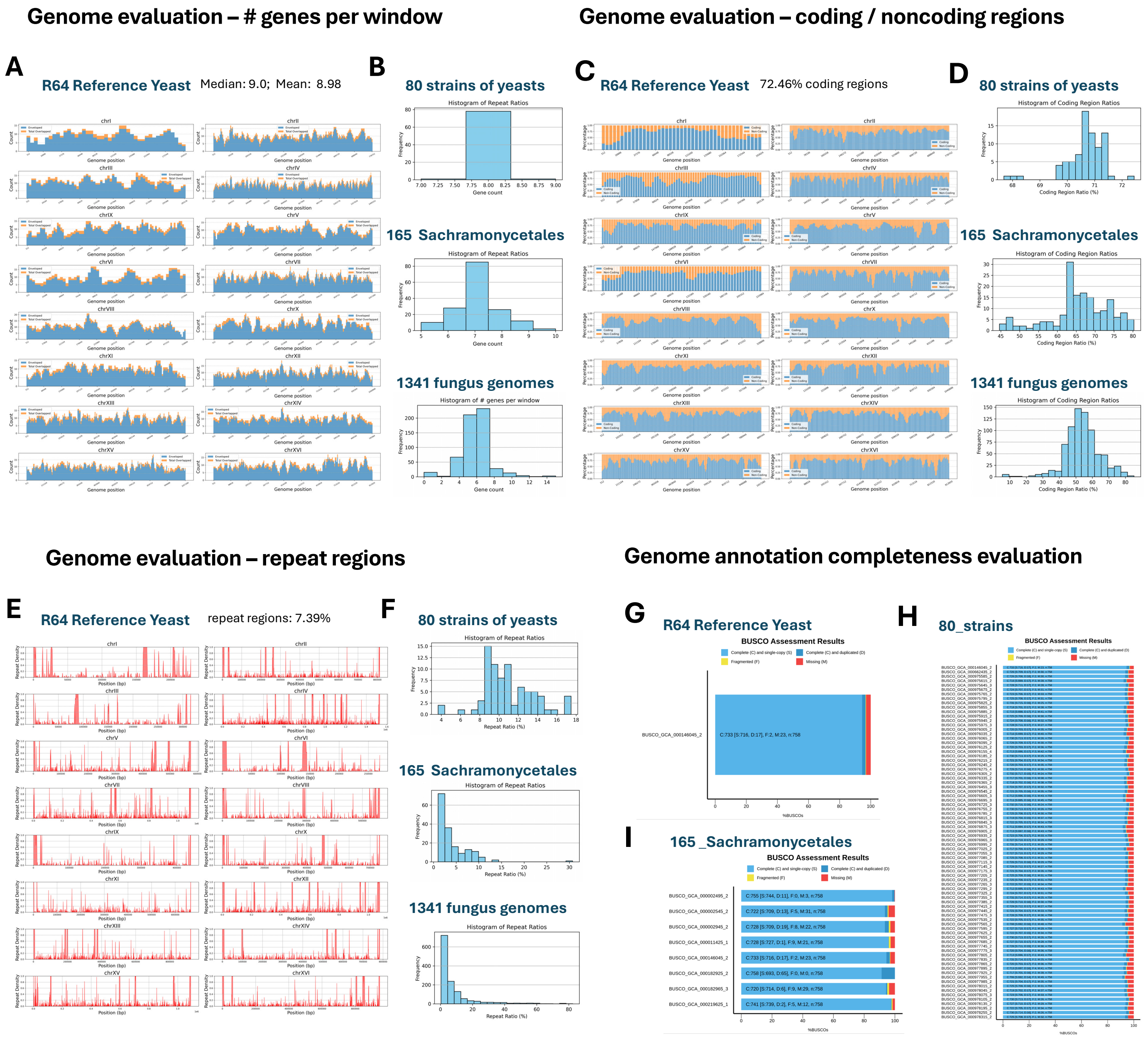
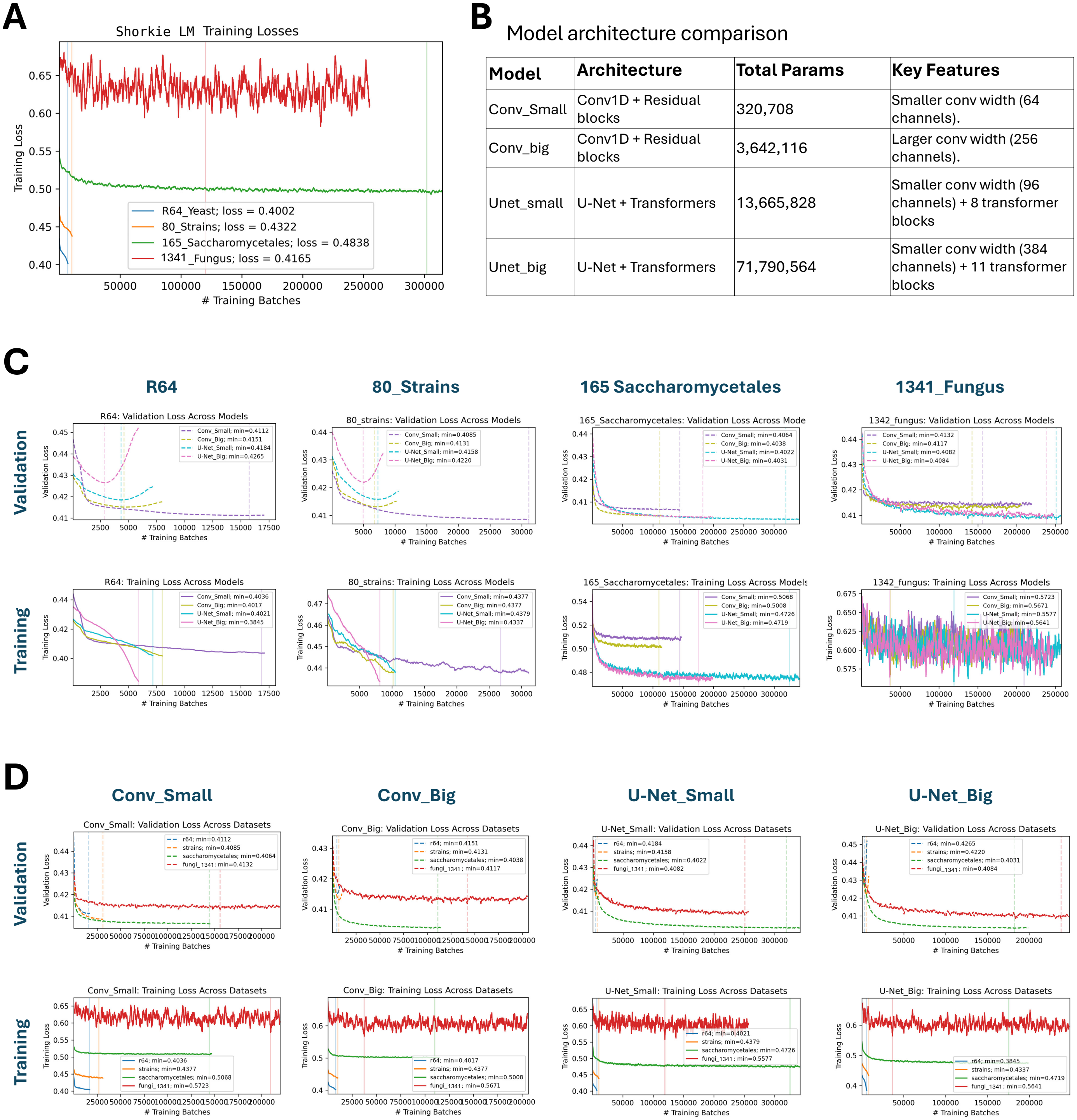


Figure S4. Window-based analysis of genomic features and BUSCO annotation completeness across fungal assemblies. Genomes were partitioned into 16,384 bp windows with a 4,096 bp stride, and the following analyses were performed: **(A)** gene count per window in the *S. cerevisiae* reference genome; **(B)** distribution of gene counts across 80 *S. cerevisiae* strains, 165 Saccharomycetales genomes, and 1,341 fungal assemblies; **(C)** ratio of coding-to-noncoding sequences per window in the *S. cerevisiae* reference genome; **(D)** distribution of coding-to-noncoding ratios across the three genome sets; **(E)** fraction of each window that is soft-masked as repetitive sequence in the *S. cerevisiae* reference genome; **(F)** distribution of repeat-masked fractions across the three genome sets; **(G–I)** BUSCO v5 (fungi_odb10) completeness scores: percentages of complete, fragmented, and missing orthologs for **(G)** the *S. cerevisiae* reference genome, **(H)** 80 *S. cerevisiae* strains, and **(I)** 165 Saccharomycetales genomes.



Model architecture comparison

Model	Architecture	Total Params	Key Features
Conv_Small	Conv1D + Residual blocks	320,708	Smaller conv width (64 channels).
Conv_big	Conv1D + Residual blocks	3,642,116	Larger conv width (256 channels).
Unet_small	U-Net + Transformers	13,665,828	Smaller conv width (96 channels) + 8 transformer blocks
Unet_big	U-Net + Transformers	71,790,564	Smaller conv width (384 channels) + 11 transformer blocks

Figure S5. Evaluation of Shorkie LM training across evolutionary divergence and model architectures. **(A)** Training and validation loss curves for models trained on four fungal genomic datasets, ordered by increasing evolutionary divergence: R64 (blue; least divergent), 80_strains (orange), 165_Saccharomycetales (green), and 1341_Fungus (red; most divergent). Solid lines represent training loss, while dashed lines represent validation loss across training iterations. **(B)** Overview of the four neural network architectures assessed: (1) Conv_Small (64-channel Conv1D with residual blocks), (2) Conv_Big (256-channel Conv1D with residual blocks), (3) U-Net_Small (U-Net backbone with eight transformer blocks and 96-to-384-channel convolutions), and (4) U-Net_Big (U-Net backbone with 11 transformer blocks, 384-to-768-channel convolutions, and larger convulational widths). **(C)** Final training and validation loss values for each architecture across the four datasets. The x-axis represents datasets (R64, 80_strains,

165_Saccharomycetales, 1341_Fungus), and colored lines denote different architectures. Solid lines indicate training loss, while dotted lines indicate validation loss. **(D)** Final training and validation loss values for each dataset across the four architectures. The x-axis represents architectures (Conv_Small, Conv_Big, U-Net_Small, U-Net_Big), with colored lines corresponding to datasets. Solid curves represent training loss, while dotted curves represent validation loss.

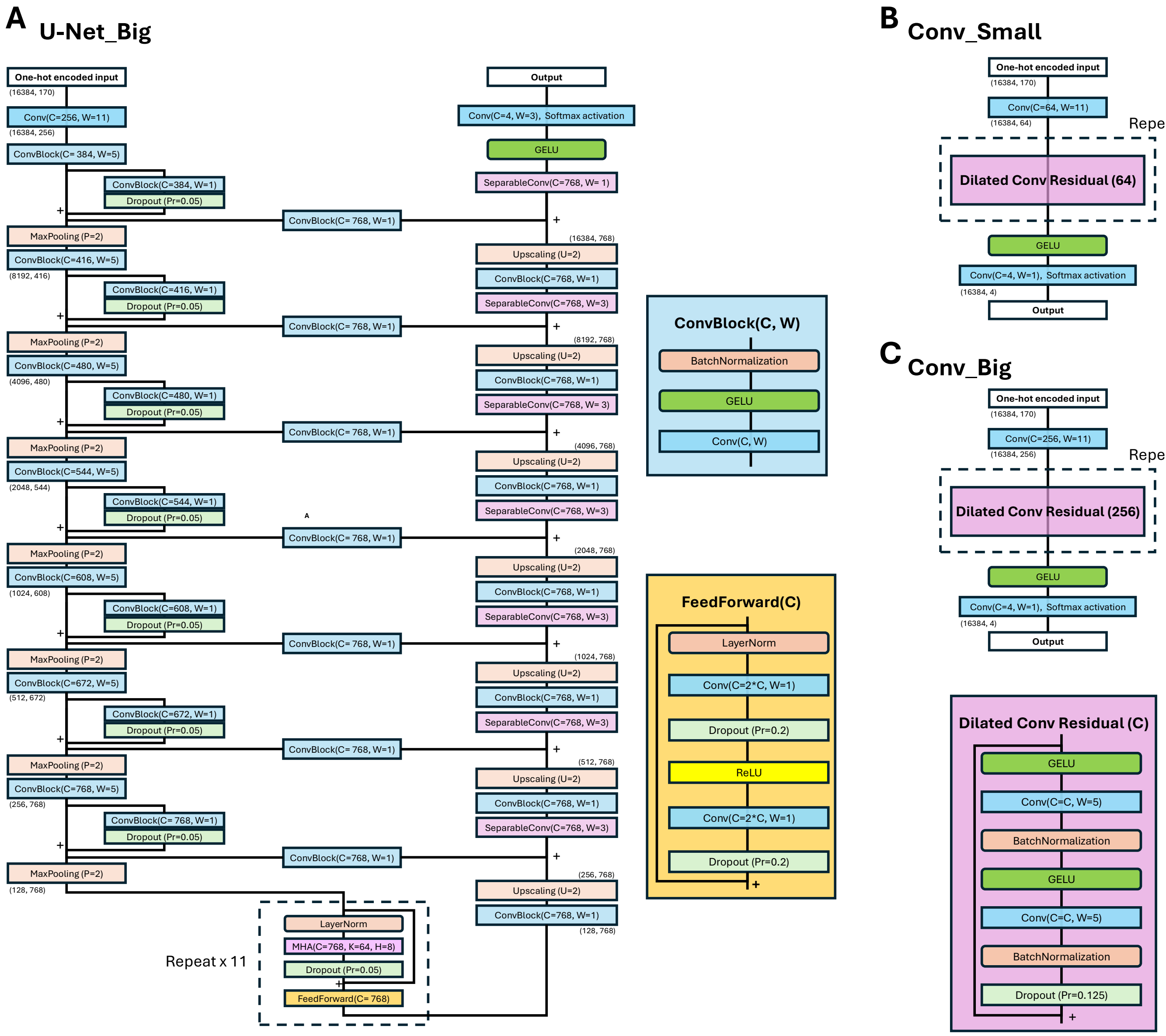


Figure S6. Schematic overviews of the three additional neural network architectures evaluated for Shorkie LM (U-Net_Small is described in Figure S1). **(A)** U-Net_Big: A U-Net–style encoder–decoder with 11 transformer blocks interleaved at the bottleneck, 384 convolutional channels per layer, skip connections between matching resolution stages, and a total of ~71.8 million parameters. **(B)** Conv_Small: A compact residual 1D convolutional network comprising four residual blocks with 64-channel convolutions, totaling ~320.7 thousand parameters. **(C)** Conv_Big: A deeper residual 1D convolutional network with four residual blocks using 256-channel convolutions, totaling ~3.64 million parameters. Each panel illustrates the flow of data from raw one-hot encoded sequence input through successive convolutional and, where applicable, transformer layers, to the final prediction head.

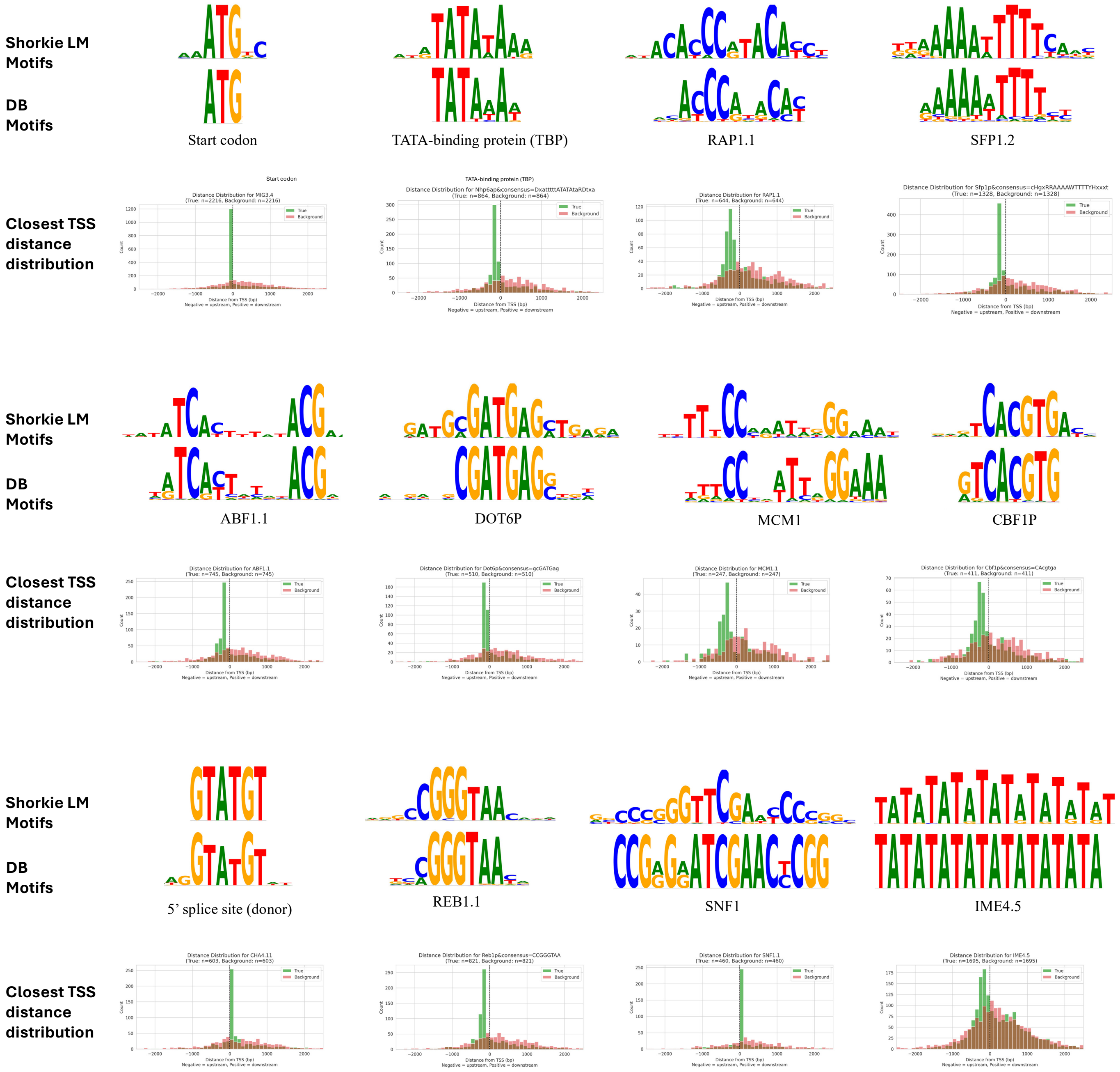


Figure S7. *De novo* motif discovery by Shorkie LM and positional enrichment relative to transcription start sites (TSSs). Using TF-MoDISco-Lite with importance scores derived from the *S. cerevisiae* reference genome, we recovered and clustered seqlets corresponding to the start codon, TATA-binding protein (TBP), Rap1.1, Sfp1.2, Abf1.1, Dot6p, Mcm1, Cbf1p, the 5′ splice-site donor motif, Reb1.1, Snf1 and Ime4.5. For each clustered motif, we plotted histograms of each motif's detected seqlet distances to the nearest TSS versus a shuffled background.

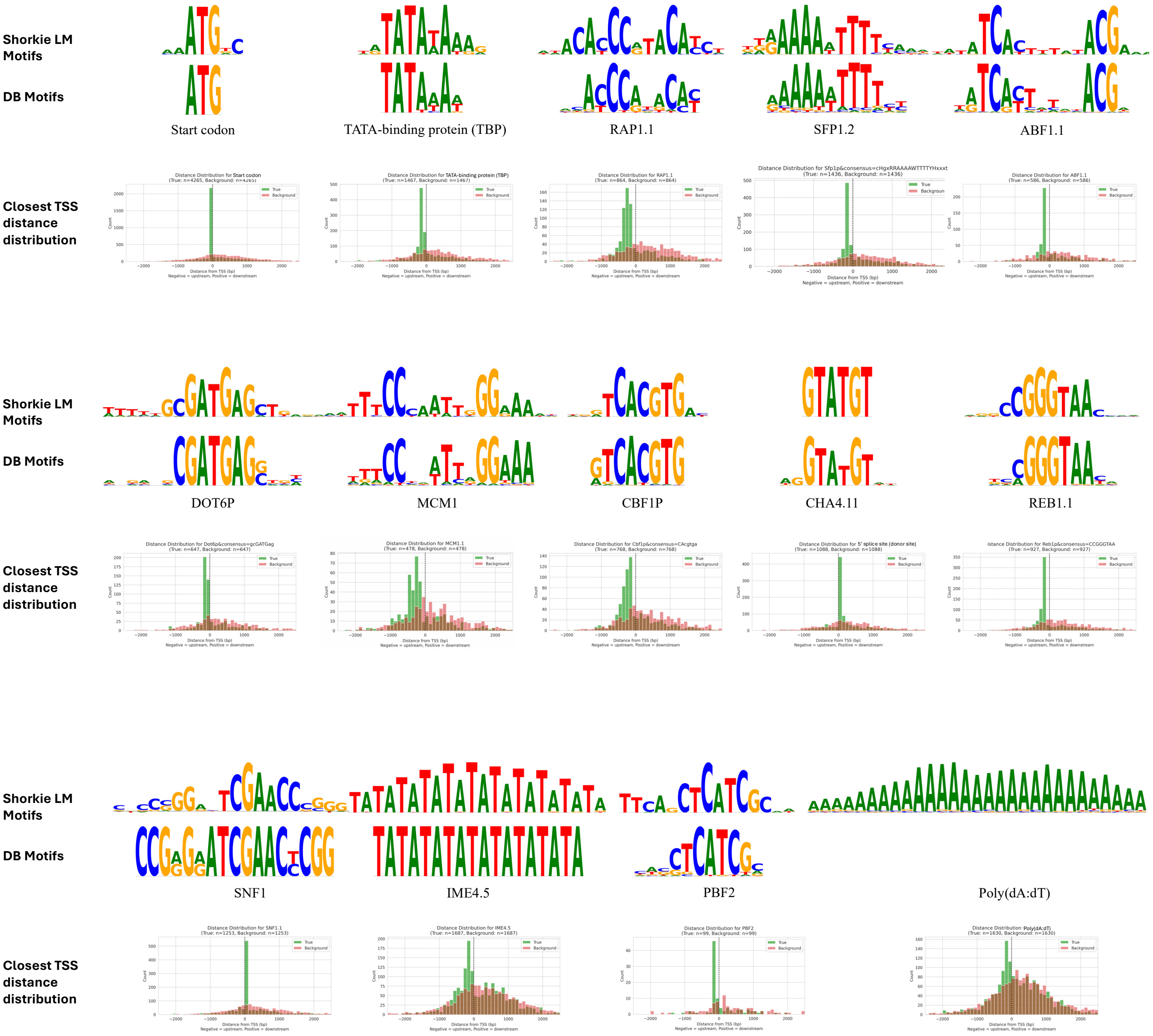


Figure S8. *De novo* motif discovery by Shorkie LM and positional enrichment relative to transcription start sites (TSSs). Using TF-MoDISco-Lite with importance scores derived from five Saccharomycetales genomes, we recovered and clustered seqlets corresponding to the start codon, TATA-binding protein, Rap1.1, Sfp1.2, Abf1.1, Dot6p, Mcm1, Cbf1p, the 5′ splice-site donor motif, Reb1.1, Snf1, Ime4.5, Pbf2, and Poly(dA:dT). For each clustered motif, we plotted histograms of each motif’s detected seqlet distances to the nearest TSS versus a shuffled background.

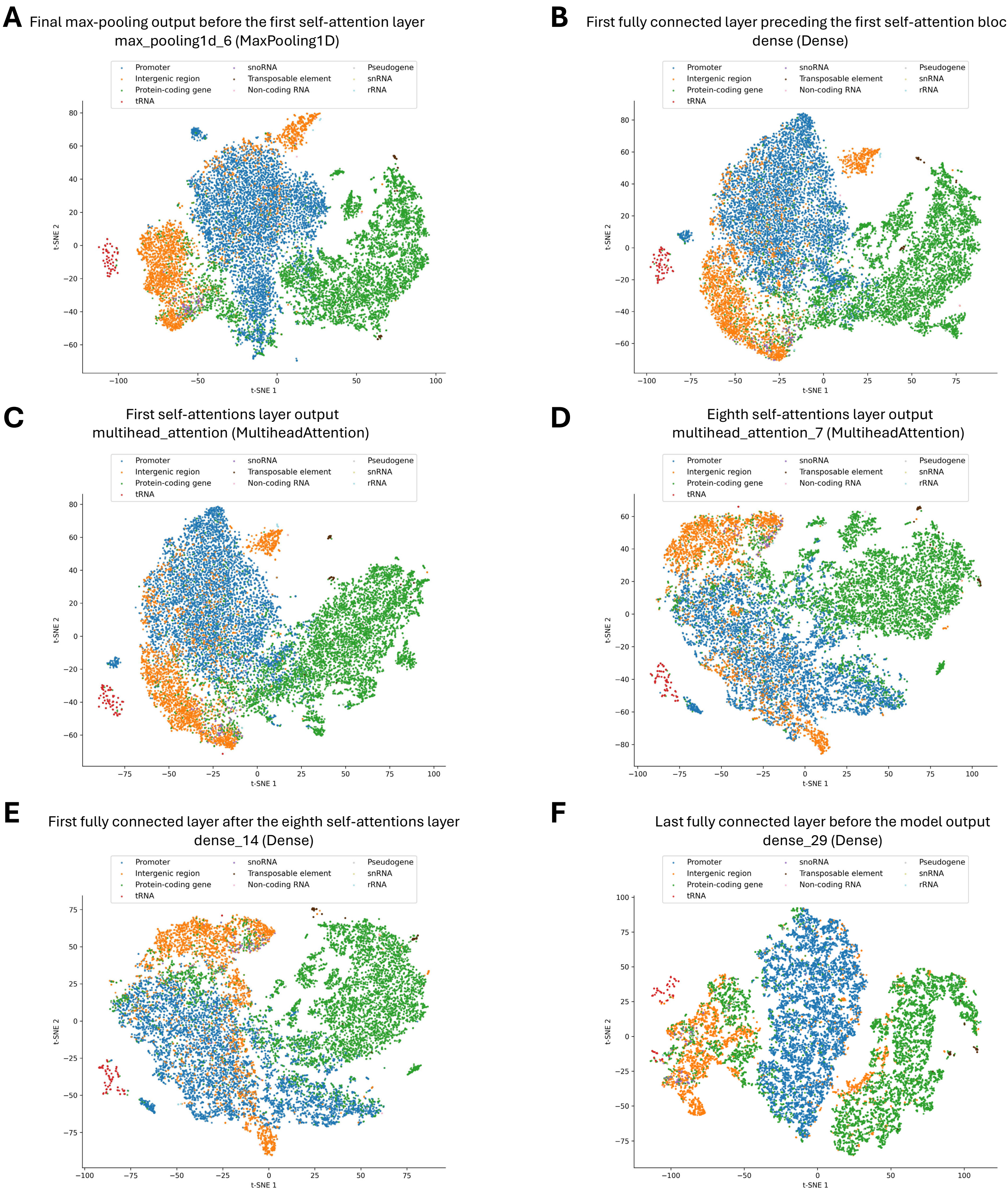


Figure S9. t-SNE projections of Shorkie LM embeddings at six network layers. Embedding vectors were extracted for the same set of genomic loci at six distinct layers in the Shorkie architecture and projected into two dimensions via t-SNE. Points are colored by features class. **(A)** Output of the first convolutional block before downsampling. **(B)** Output of the first max-pooling layer following the initial convolutional block. **(C)** Output of the first multi-head self-attention layer. **(D)** Output of the eighth multi-head self-attention layer. **(E)** Output of the first feed-forward (dense) layer following the eighth self-attention. **(F)** Output of the final feed-forward (dense) layer preceding the model's prediction.

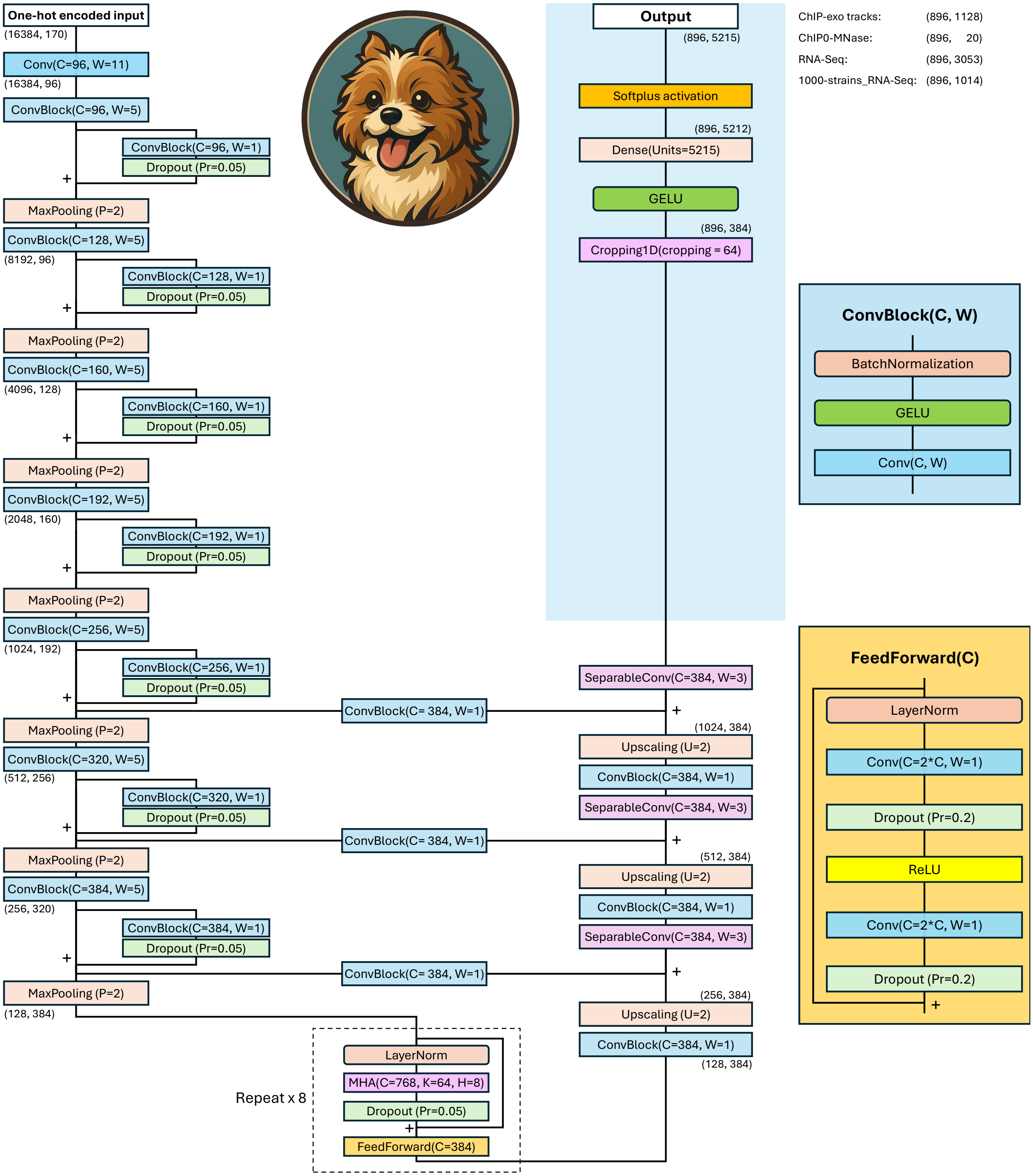


Figure S10. Shorkie model architecture. The Shorkie model, with 14.3 million parameters, processes one-hot-encoded 16,384 bp genomic windows and predicts 5,215 distinct genomic tracks at a resolution of 16 bp. The model architecture

begins with a 1D convolutional block (96 filters, kernel size = 11) followed by seven residual down-sampling stages, reducing the feature map from 16,384 × 96 to 128 × 384. A transformer bottleneck with eight layers at 128 × 384 integrates long-range dependencies. The decoder mirrors the encoder through three up-sampling stages and U-Net-style skip connections, reconstructing a 1,024 × 384 feature map. All encoder, transformer, and decoder layers are initialized with pre-trained weights from the Shorkie language model. In the final prediction head (highlighted in light blue), the feature map is cropped by 64 bins on each side, resulting in a 896 × 384 feature map, which is then passed through a fully connected layer with Softplus activation. The output is a 896 × 5,215 matrix representing track predictions: 1,128 ChIP-exo tracks, 20 ChIP-MNase tracks, 3,053 IDEA RNA-Seq tracks, and 1,014 1000-strain RNA-Seq tracks.

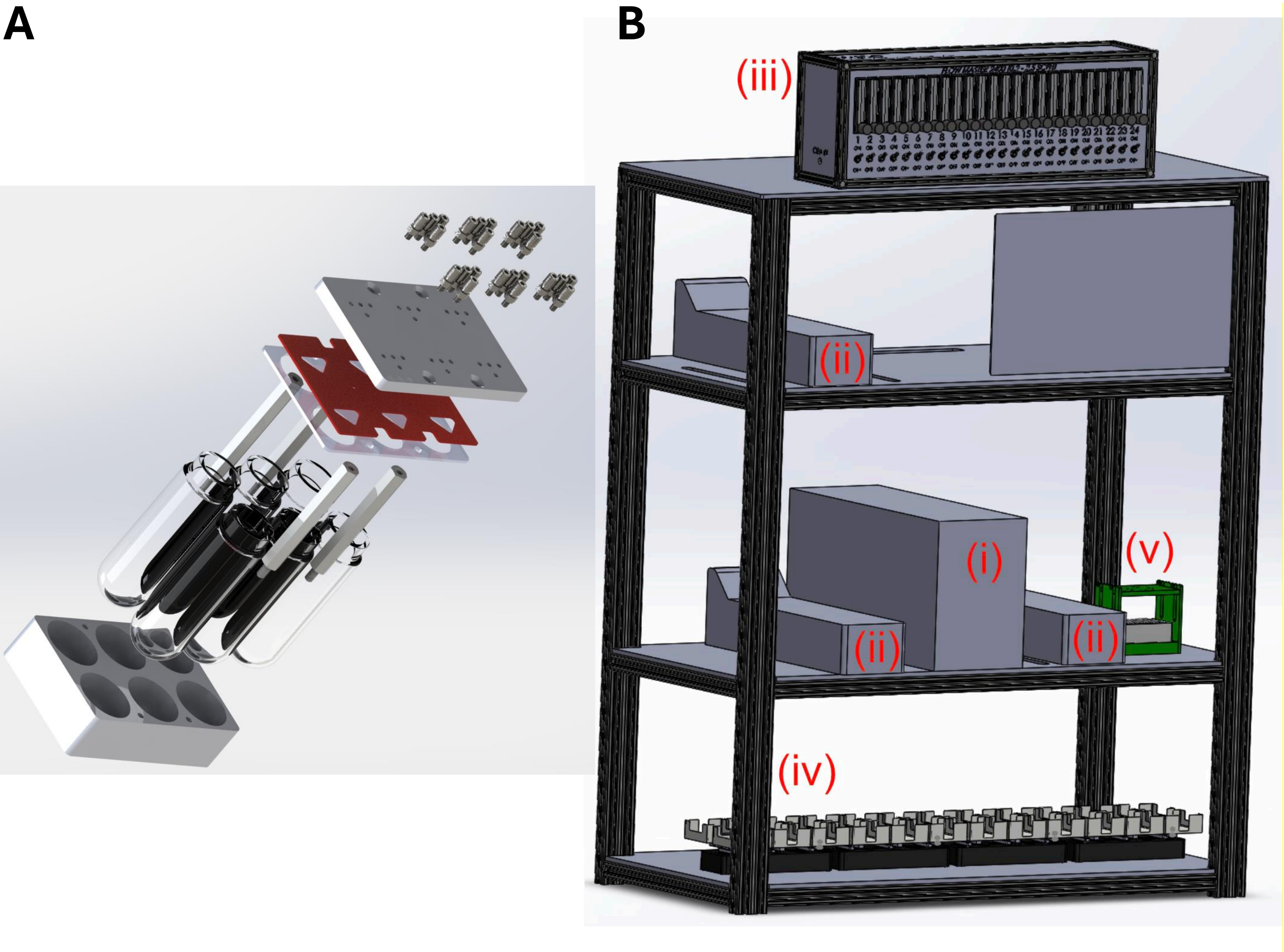


Figure S11. Overview of ministat array system. **(A)** Yeast were cultured in 100mL round bottom flasks housed in banks of 6-vessels. Each vessel supported connections for media, sampling, and an input for inoculating a bolus of cells or adding estradiol to activate transcriptional induction. **(B)** CAD mock up of major system components highlighted in red (i) 24-vessel manifold combined four 6-vessel manifolds into a complete ministat array (ii) three peristaltic pumps connected to each vessel via individual tubing for media, sampling, and cell/estradiol inputs (iii) air flow regulator with 24 lines whose flow rate was separately controlled (iv) balance system separately weighted the effluent of each ministat to calculate realized dilution rate for adding media to each vessel (v) sampling manifolds which housed plates for cells used for initial inoculation or estradiol for induction and a separate plate where final the samples for sequencing were collected.

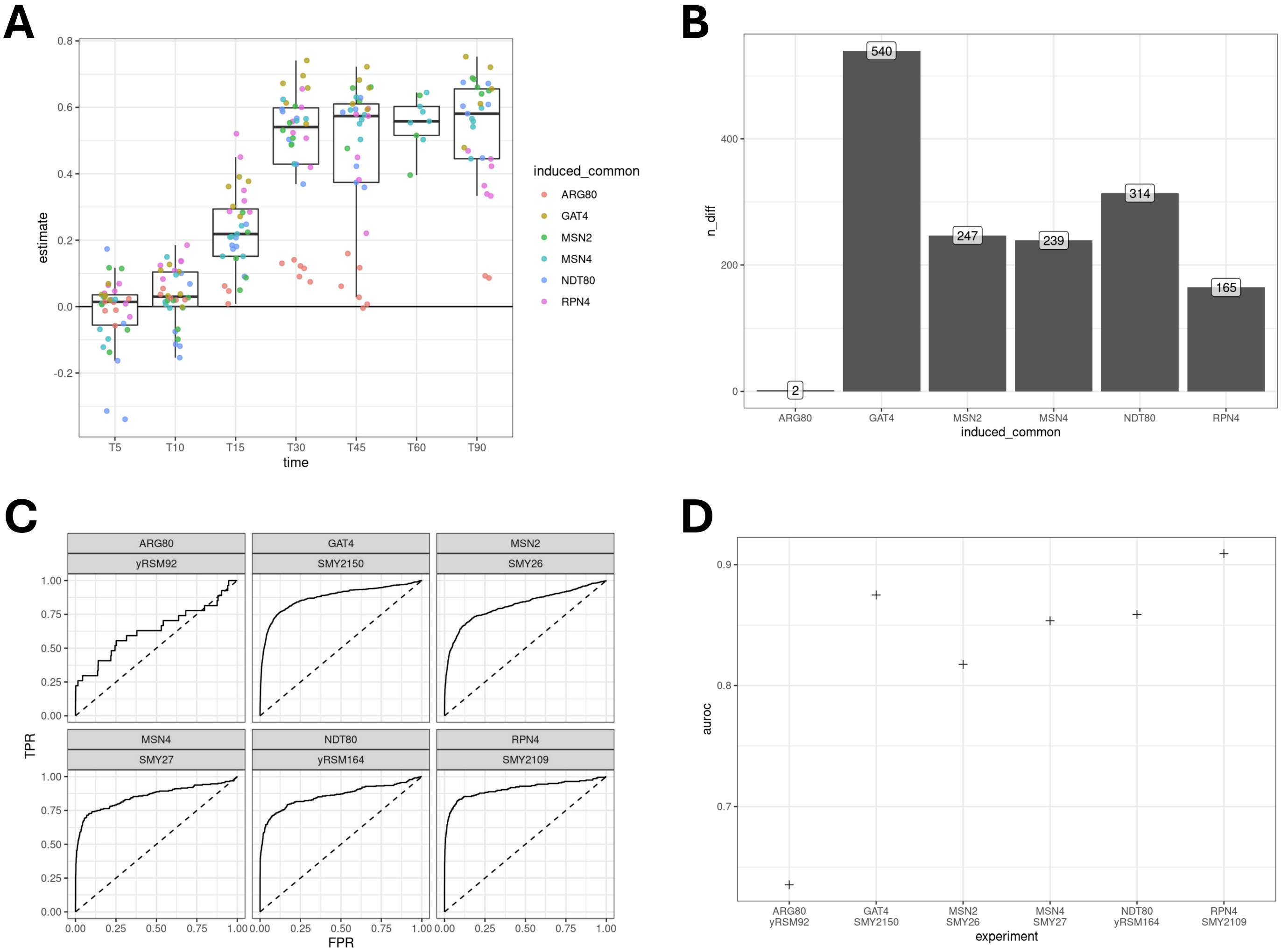


Figure S12. Ministat Array System Validation. **(A)** Box-and-whisker plots of Pearson correlations between $\log_2$ fold-changes measured by microarray versus RNA-seq for each replicate induction experiment; point colors correspond to the induced TF. **(B)** Bar plot of the number of differentially expressed genes for each TF perturbation. **(C)** ROC curves measuring accuracy in detecting differentially expressed genes for each TF perturbation using the new ministat array system. **(D)** AUROC values for each TF perturbation experiment.

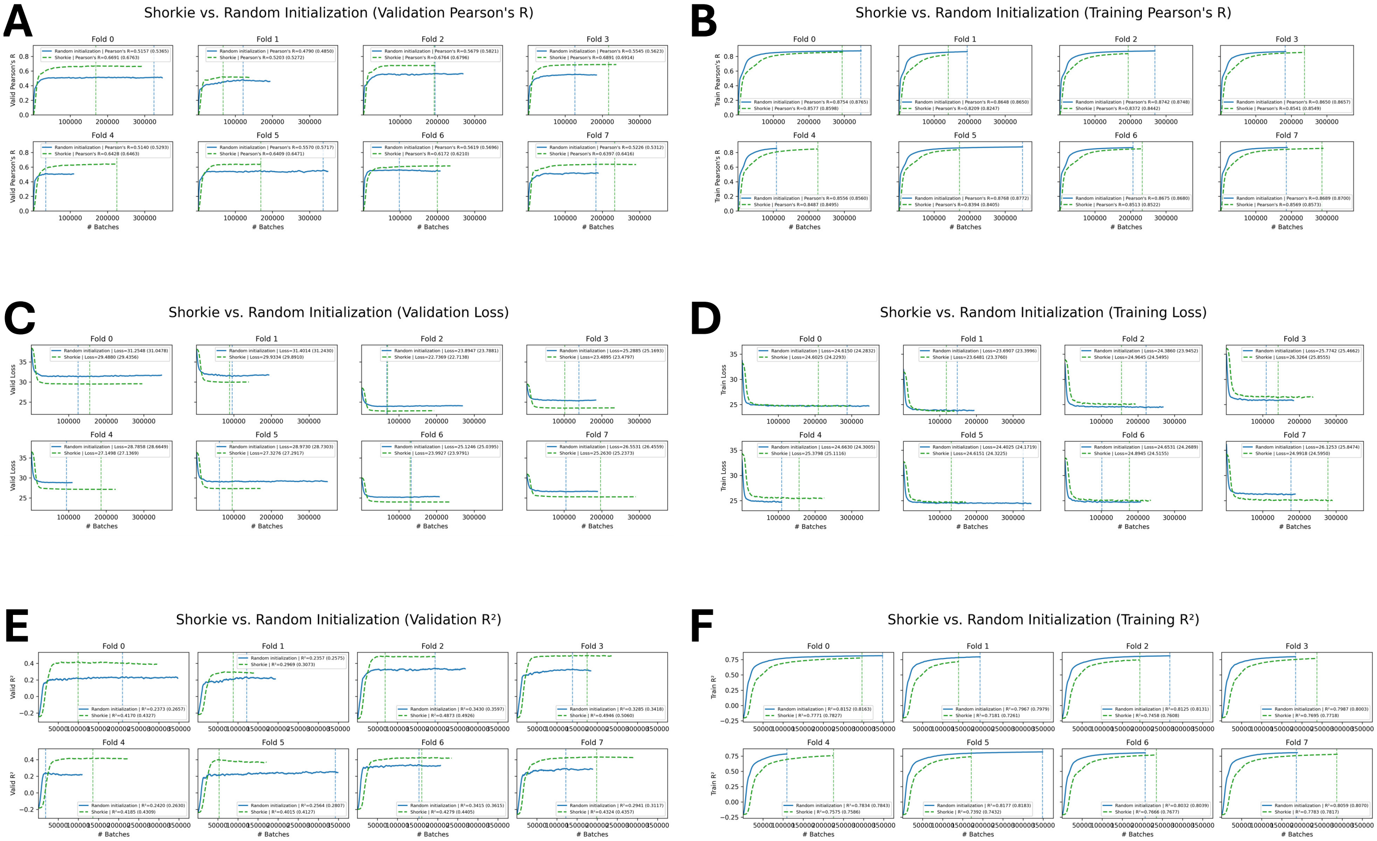


Figure S13. Training and validation performance of Shorkie and Shorkie_Random_Init. The figure shows six panels displaying the metric trajectories across training epochs for the Shorkie model (initialized with pretrained Shorkie LM weights; dashed lines) and the Shorkie_Random_Init model (initialized with random weights; solid lines). Panels A, C, and E represent validation metrics, while panels B, D, and F correspond to training metrics. **(A)** Validation Pearson's correlation coefficient (R); **(B)** training Pearson's correlation coefficient (R); **(C)** validation loss; **(D)** training loss; **(E)** validation R-squared (R²); and **(F)** training R-squared (R²).

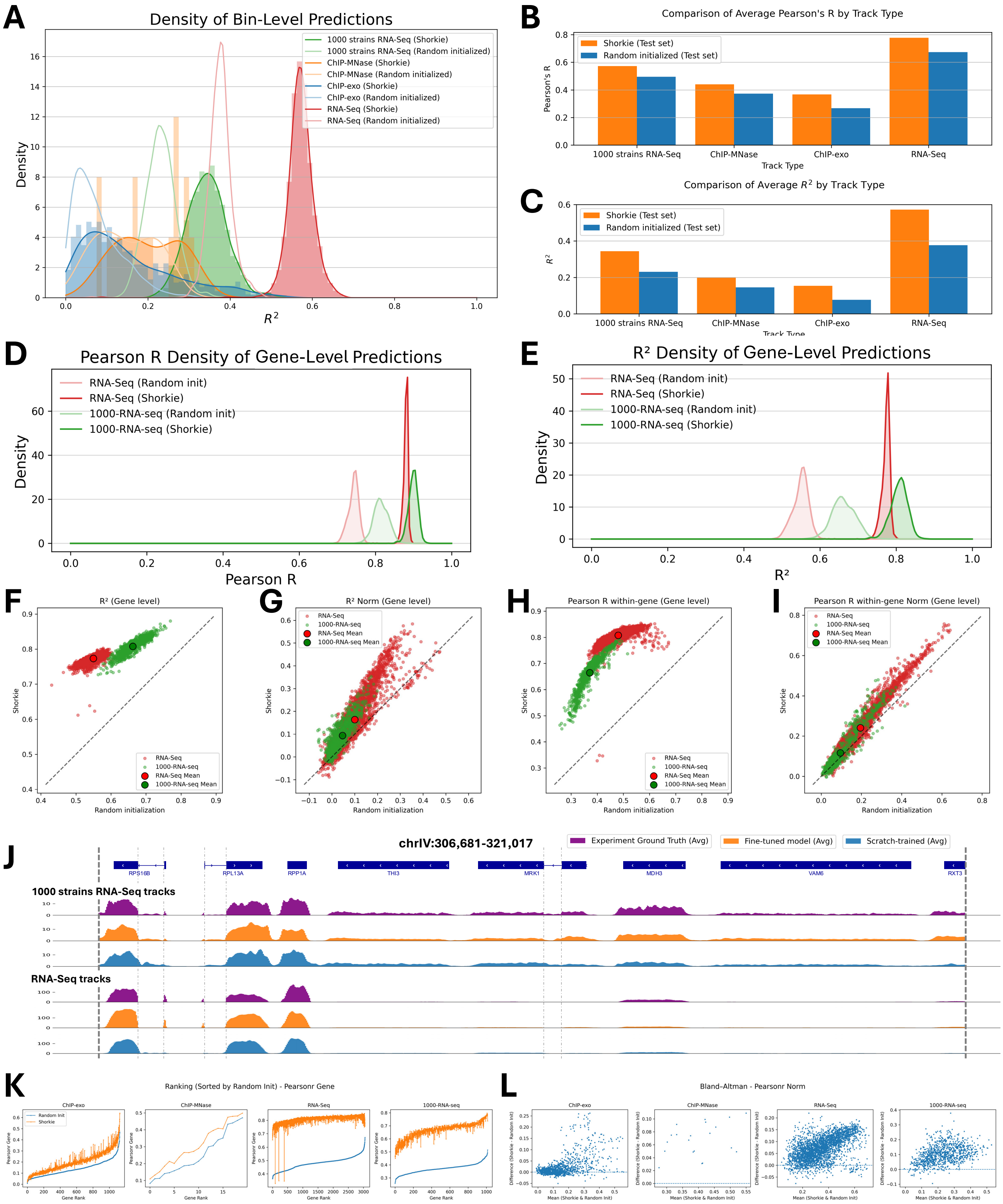


Figure S14. Performance overview of Shorkie's predictive capabilities at both bin- and gene-level resolutions on held-out test data. Panels A–C present the distribution and mean values of bin-level metrics across all track types: **(A)** shows the full distribution of R² scores, while **(B)** and **(C)** display the average Pearson's correlation coefficient (R) and R²,

respectively, comparing the pretrained Shorkie model (orange) with the randomly initialized baseline model, Shorkie_Random_Init (blue). Panels D–I depict gene-level metrics: **(D)** and **(E)** illustrate the distributions of Pearson's R and R^2 scores for the IDEA and 1000-strain RNA-Seq tracks, while the scatter plots in **(F)–(I)** compare gene-level R^2, normalized R^2, within-gene Pearson's R, and normalized within-gene Pearson's R, respectively, emphasizing the concordance between predicted and observed gene expression. **(J)** shows an RNA-Seq coverage snapshot over a 16 kb window (chrIV:306,681–321,017) spanning the two-exon genes RPS16B, RPL13A, and MRK1. **(K)** presents a rank-ordered plot of track-level R^2 scores for Shorkie_Random_Init, with Shorkie scores shown according to the same ranking. **(L)** shows a Bland-Altman plot comparing normalized Pearson's R between Shorkie and Shorkie_Random_Init, with the difference plotted against the mean.

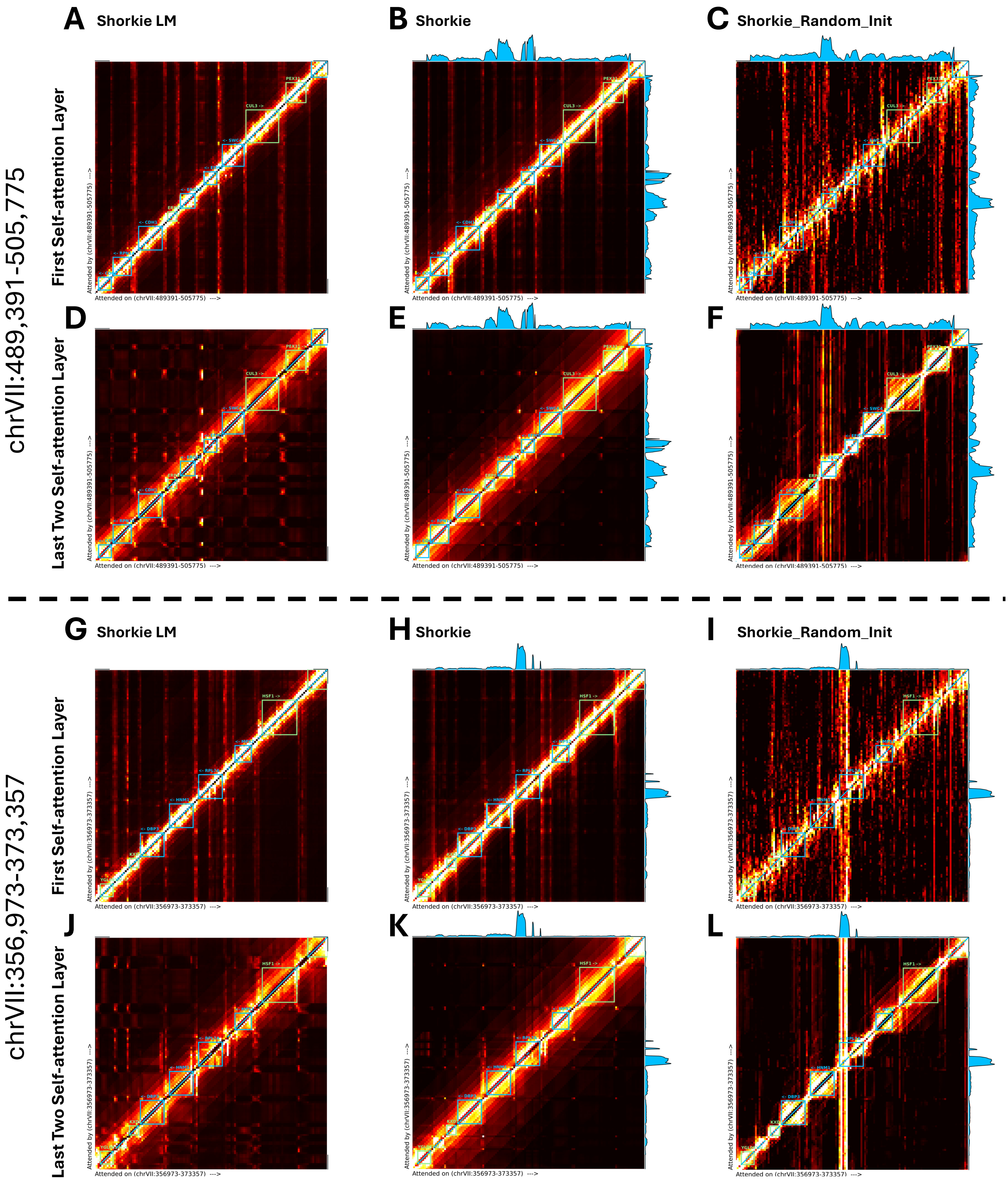


Figure S15. Self-attention reveals gene-structure encoding and refined regulatory focus. Self-attention scores, averaged across eight heads, are shown for (i) the first transformer block **(A–C, G–I)** and (ii) the last two self-attention blocks **(D–F, J–L)** for each model (Shorkie LM in column 1, Shorkie in column 2, and Shorkie_Random_Init in column 3, over 16,384 bp windows centered on the three-exon gene **EFM5**

(chrVII:489,391–505,775; top two rows) and the housekeeping gene **RPL7A** (chrVII:356,973–373,357; bottom two rows), using T_0 RNA-Seq tracks. All models focus attention on exonic segments within gene bodies, indicating robust gene-structure recognition. Shorkie LM and Shorkie additionally exhibit sharp peaks in flanking intergenic regions, whereas Shorkie_Random_Init shows a diffuse intergenic signal. Transfer learning further sharpens Shorkie's regulatory focus: last-layer attention **(D–F, J–L)** becomes tightly localized at specific intergenic loci, while first-layer attention **(A–C, G–I)** remains broadly distributed, reflecting progressive layer-wise specialization.

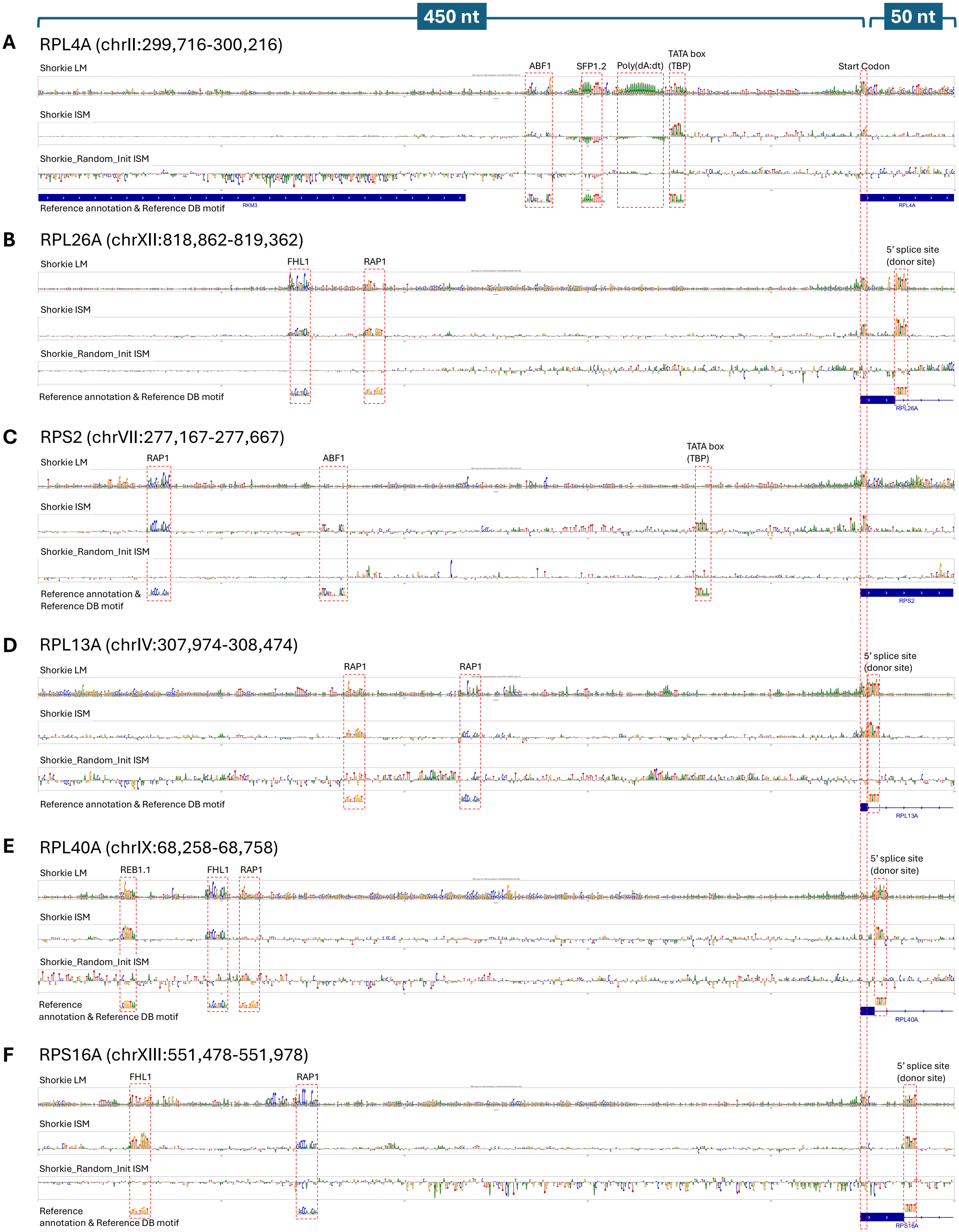


Figure S16. Motif analysis in promoter regions of six ribosomal protein genes. A 500-bp window spanning –450 to +50 bp around the TSS was extracted for **(A)** RPL4A (chrII:299,716–300,216), **(B)** RPL43B (chrXII:818,862–

819,362), **(C)** RPS9A (chrVII:277,167–277,667), **(D)** RPL13A (chrIV:307,974–308,474), **(E)** RPL40A (chrIX:68,258–68,758) and **(F)** RPS16A (chrXIII:551,478–551,978). For each promoter: Row 1 shows conservation-based DNA logos produced by Shorkie LM. Row 2 shows ISM maps from the fine-tuned Shorkie model, where each nucleotide was systematically substituted and the effect on predicted RNA-Seq signal recorded. Row 3 shows ISM maps from Shorkie_Random_Init trained from scratch. Row 4 overlays the reference gene annotation and curated TF-binding motifs from the yeast motif database.

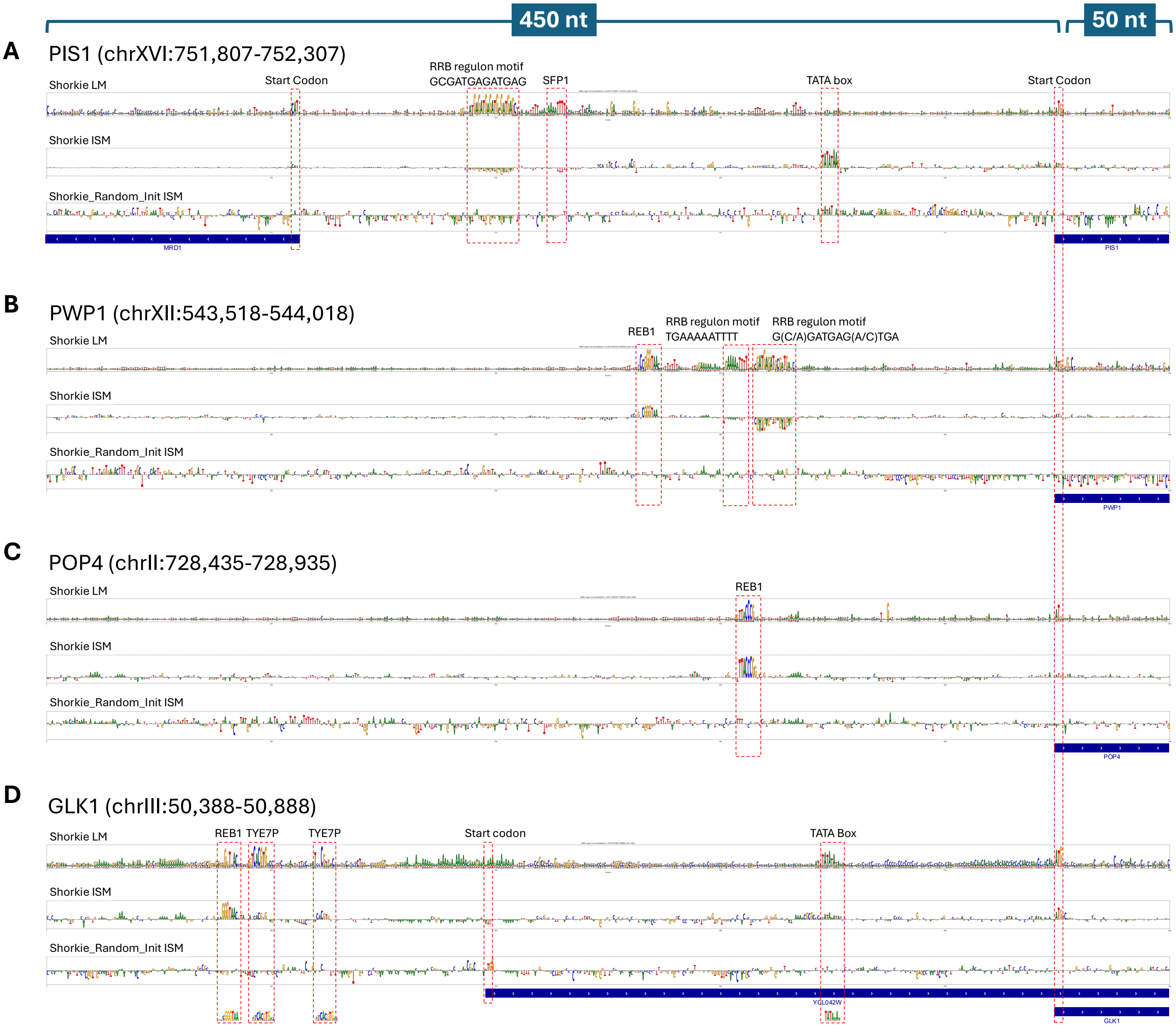


Figure S17. Motif analysis in promoter regions of four protein-coding genes. A 500-bp window spanning –450 to +50 bp around the TSS was extracted for **(A)** PIS1 (chrXVI:751,807–752,307), **(B)** PWP1 (chrXII:543,518–544,018), **(C)** POP4 (chrII:728,435–728,935), and **(D)** GLK1 (chrIII:50,388–50,888). For each promoter: Row 1 shows conservation-based DNA logos produced by Shorkie LM. Row 2 shows ISM maps from the fine-tuned Shorkie model, where each nucleotide was systematically substituted and the effect on predicted RNA-Seq signal recorded. Row 3 shows ISM maps from Shorkie_Random_Init trained from scratch. Row 4 overlays the reference gene annotation and curated TF-binding motifs from the yeast database, including the RRB regulon motifs (e.g. G(C/A)GATGAG(A/C)TGA and TGAAAAATTTT), REB1, TYE7, TATA box, and the annotated start codon.

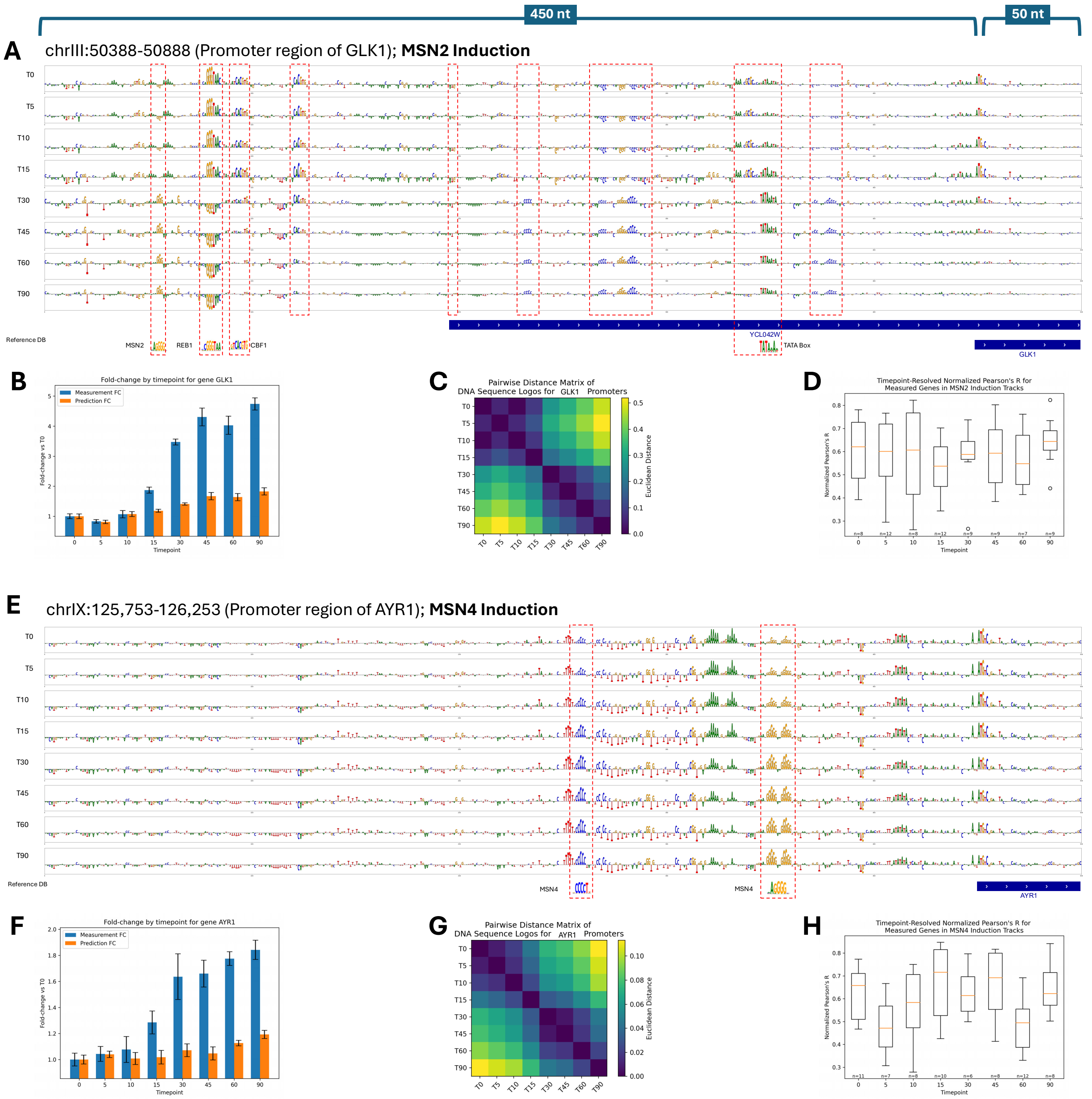


Figure S18. Time-course analysis of MSN2 and MSN4 induction in *Saccharomyces cerevisiae*. **(A–D)** MSN2 induction at the GLK1 promoter (chrIII:50,388–50,888): (A) Shorkie-generated ISM maps for the GLK1 promoter at eight timepoints following MSN2 induction (T0 = 0 min; T5 = 5 min; T10 = 10 min; T15 = 15 min; T30 = 30 min; T45 = 45 min; T60 = 60 min; T90 = 90 min). Rows correspond to successive timepoints, with the bottom row showing the reference sequence. Key motifs (TATA box, Cbf1, MSN2, Reb1) are annotated. **(B)** Experimental RNA-Seq fold change relative to T0 (blue) versus Shorkie predictions (orange) across the GLK1 locus. **(C)** Heatmap of pairwise Euclidean distances between the ISM maps for all MSN2 induction timepoint pairs. **(D)** Boxplot of normalized Pearson's R scores between experimental and predicted profiles across all yeast genes under MSN2 induction at each timepoint. **(E–H)** MSN4 induction at the AYR1 promoter (chrIX:125,753–126,253), with panels analogous to **A–D**: **(E)** ISM maps generated by Shorkie for the AYR1 promoter across the same eight timepoints. **(F)** Experimental versus

predicted fold change at different timepoints over the AYR1 locus. **(G)** Pairwise Euclidean distance heatmap of the AYR1 logos. **(H)** Boxplot of normalized Pearson's R scores for MSN4-induced samples at each timepoint.

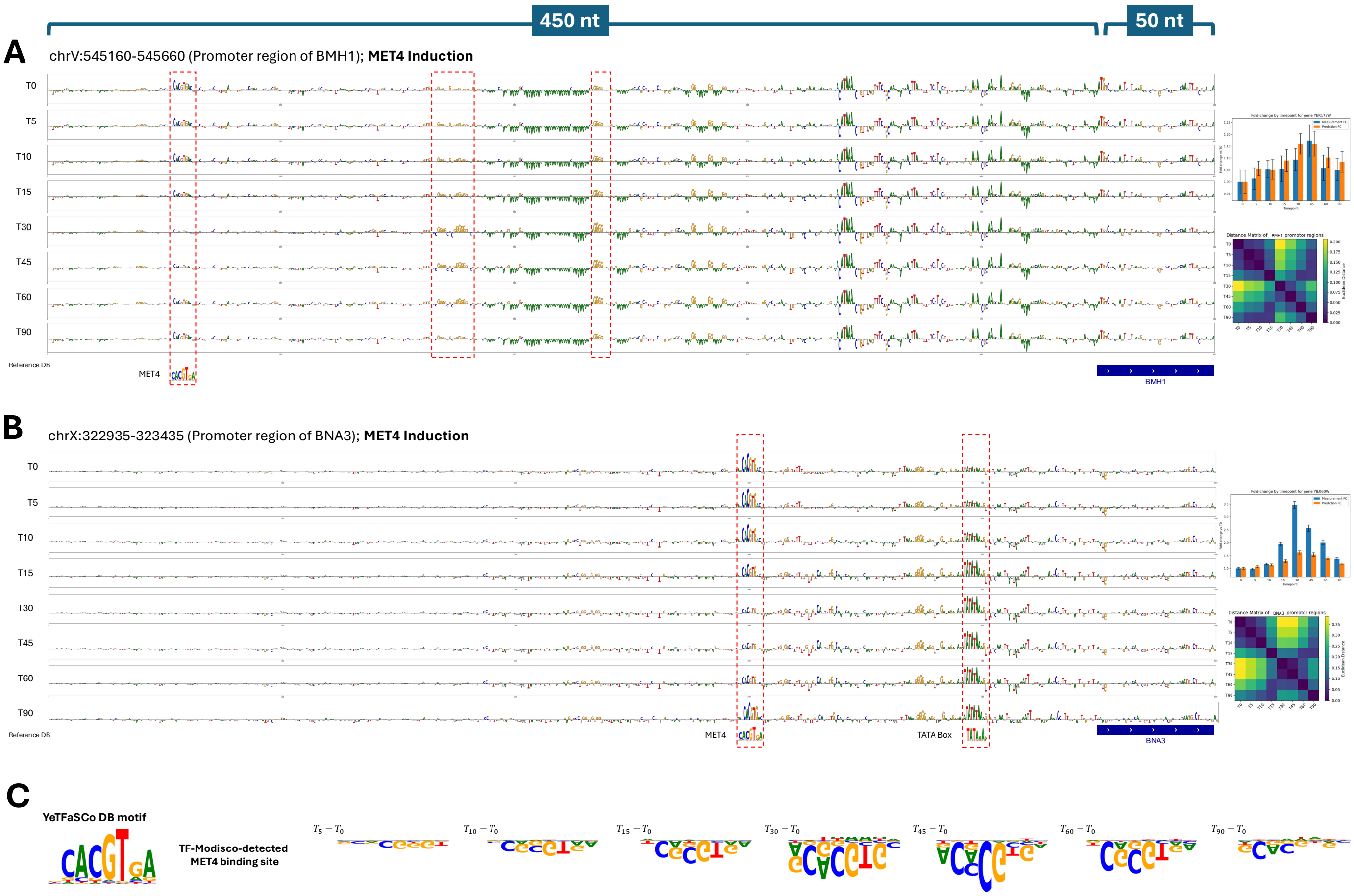


Figure S19. Time-course analysis of MET4 induction in *Saccharomyces cerevisiae*. Shorkie ISM was applied to 500 bp windows (–450 to +50 bp relative to the transcription start site, TSS) of the BMH1 (chrV:545,160–545,660) and BNA3 (chrX:322,935–323,435) promoters during MET4 induction. **(A)** ISM heatmaps for the BMH1 promoter at eight timepoints (T0 = 0 min, T5 = 5 min, T10 = 10 min, T15 = 15 min, T30 = 30 min, T45 = 45 min, T60 = 60 min, T90 = 90 min). Each row represents a logo at a single timepoint, with the bottom row showing the reference annotation and curated *Saccharomyces cerevisiae* transcription factor (TF)-binding motifs. Key regulatory elements, including the MET4 binding site and TATA box, are annotated. **(B)** ISM heatmaps for the BNA3 promoter, plotted as in **(A)**. **(C)** Consensus MET4 binding motif derived from TF-MoDISco, generated by clustering differential ISM scores across all timepoints relative to T0.

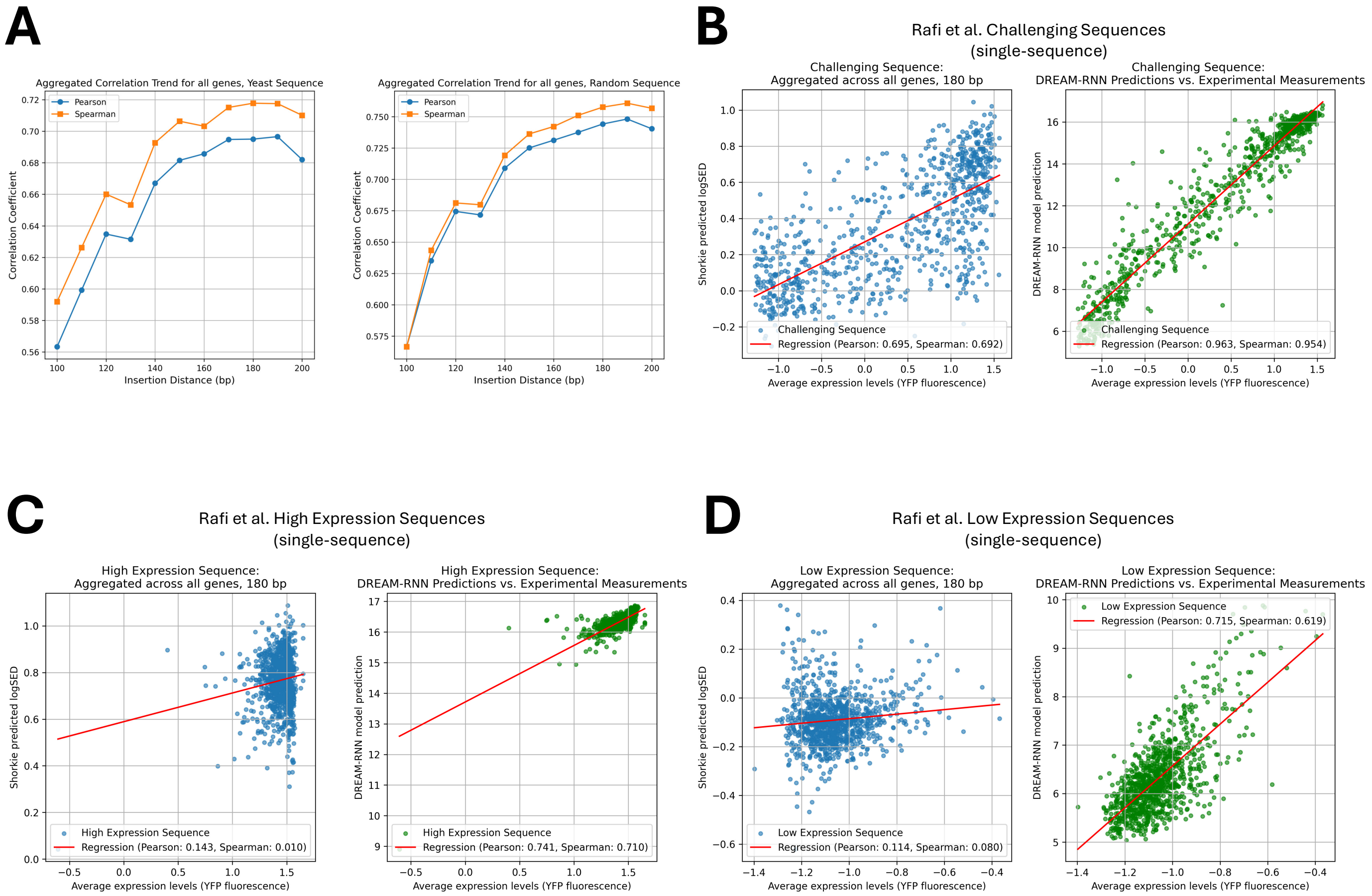


Figure S20. (A) Pearson's correlation coefficient (R) between Shorkie-predicted logSED scores at various positions upstream of the transcription start site (TSS) and mean YFP fluorescence levels measured by MPRA. Left: native yeast promoter sequences; right: randomized sequences. **(B–D)** Scatterplots of predicted logSED versus observed YFP expression for three distinct sequence subsets: **(B)** challenging constructs, **(C)** high-expression constructs, and **(D)** low-expression constructs. Each plot compares predictions from the Shorkie model (left) and the DREAM-RNN model (right).

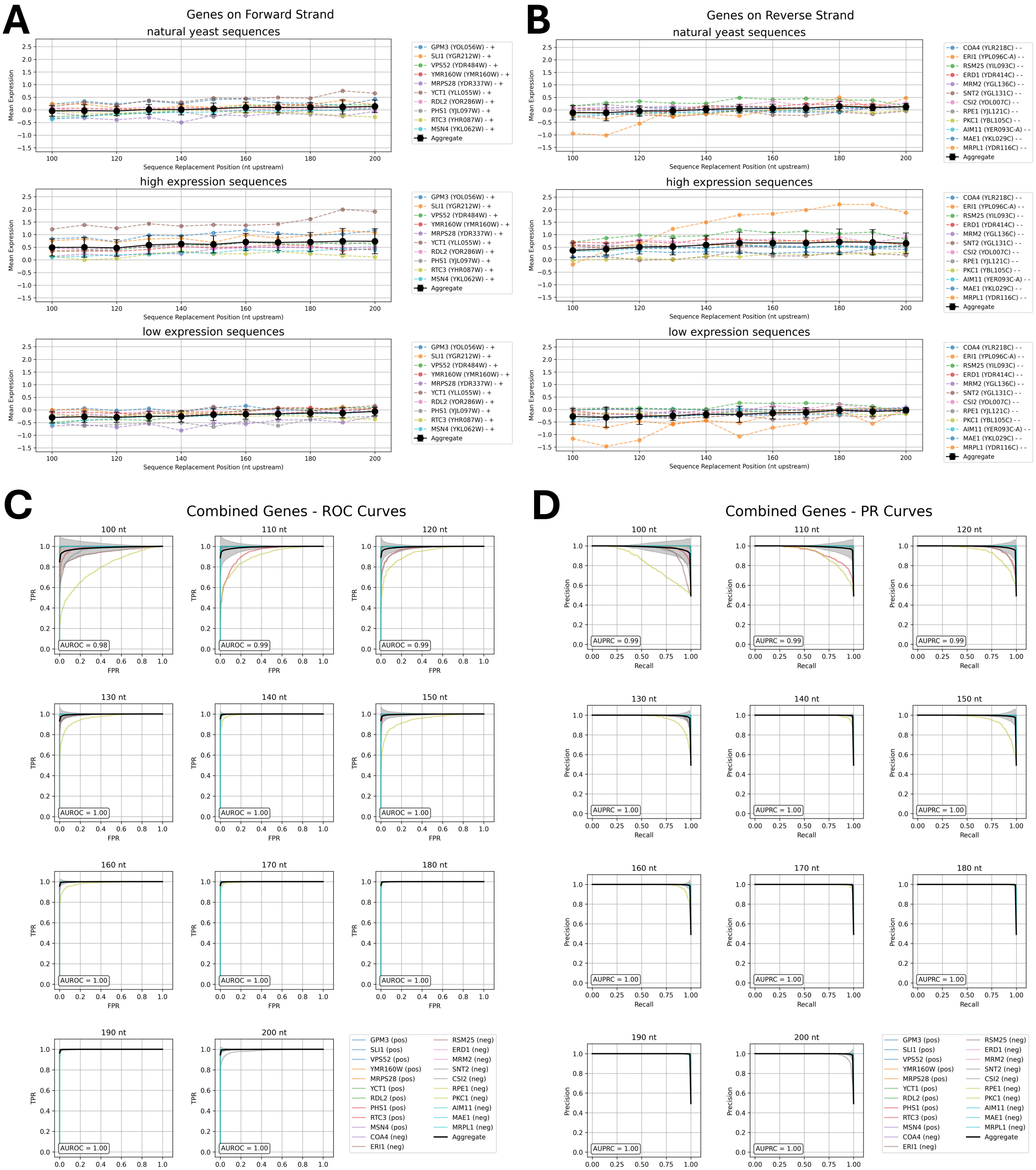


Figure S21. Shorkie prediction of cis-regulatory activity in a yeast MPRA assay. **(A–B)** Mean logSED scores predicted by Shorkie for native *S. cerevisiae* promoter sequences and for the highest- and lowest-expressing variants from the DREAM challenge, plotted at each nucleotide position relative to the transcription start site (TSS). **(A)** shows data for 10 genes on the forward strand, and **(B)** shows data for 12 genes on the reverse strand. Dashed colored lines represent the per-gene mean logSED across ~1,000 sequence variants at each position, while solid black lines indicate

the overall mean ± standard error across all 22 genes. **(C–D)** Performance of Shorkie in classifying high- versus low-expression variants from 100 to 200 bp upstream of the TSS, evaluated by **(C)** receiver operating characteristic (ROC) analysis and **(D)** precision–recall (PR) analysis. Gene-specific curves are shown in color, with the overall mean curves shown in black.

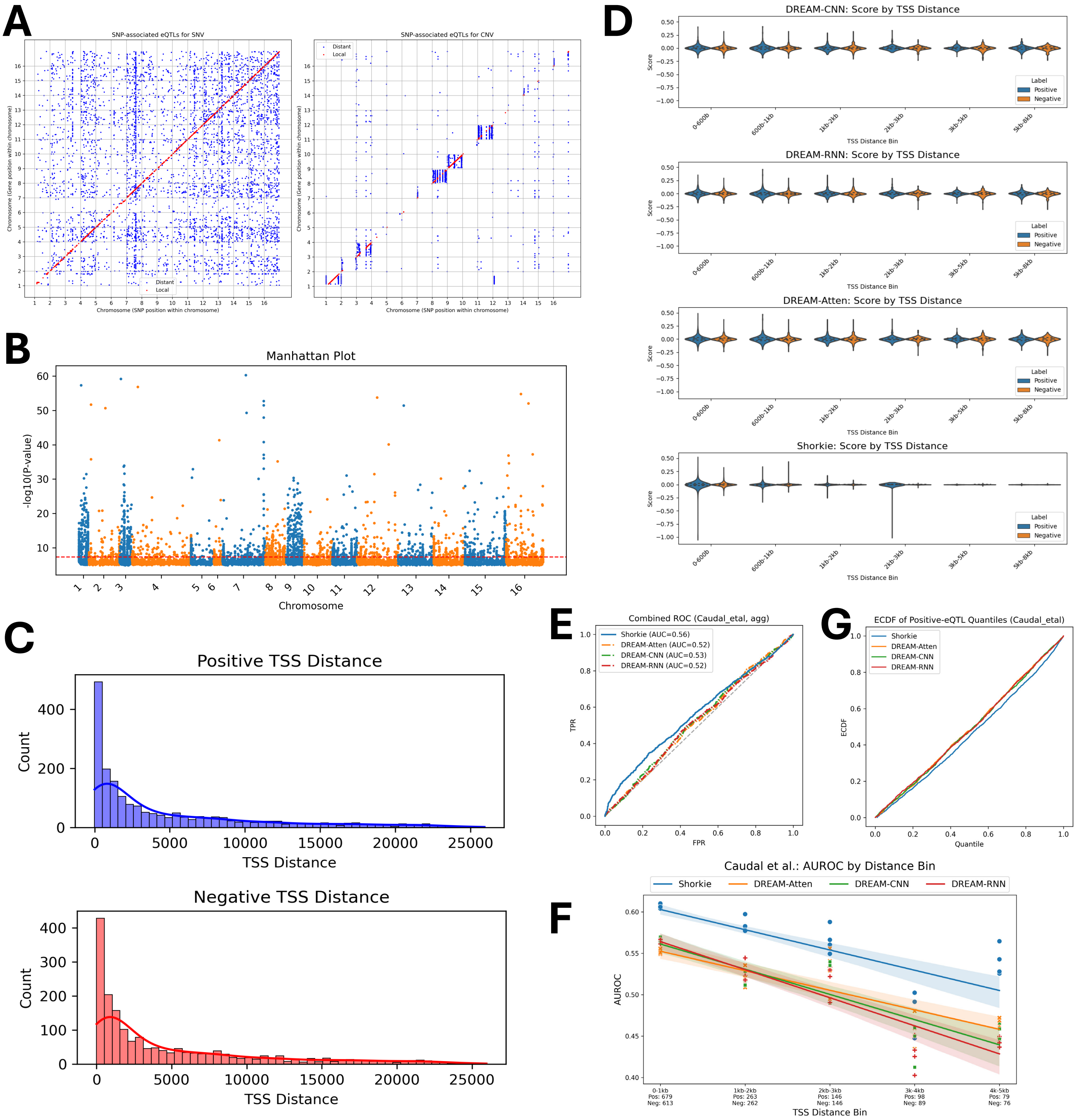


Figure S22. Predictions of variant effects on the Caudal et al. eQTL dataset by Shorkie and the DREAM Challenge models. **(A)** Genomic locations of SNP-associated eQTLs (left) and copy-number variant (CNV)-associated eQTLs (right), differentiated by local (red) and distant (blue) effects, adapted from Caudal et al. (2024). **(B)** Manhattan plot

displaying all SNP-associated eQTLs across the Saccharomyces cerevisiae genome. **(C)** Distribution of distances from eQTL variants (both positive and negative) to their target gene transcription start sites (TSS). **(D)** Score distributions for positive and negative eQTLs as predicted by Shorkie and the top three models from the DREAM Challenge (DREAM-CNN, DREAM-RNN, DREAM-Atten; Rafi et al., 2024). **(E)** Receiver operating characteristic (ROC) curves comparing the ability of each model to classify positive versus negative eQTLs. **(F)** Area under the ROC curve (AUROC) stratified by TSS distance bins, highlighting how classification performance varies with genomic distance. **(G)** Empirical cumulative distribution functions (ECDFs) of positive-eQTL quantile scores for Shorkie and the DREAM Challenge models, where a steeper rise at lower quantiles indicates stronger enrichment of high-effect predictions among true positives.

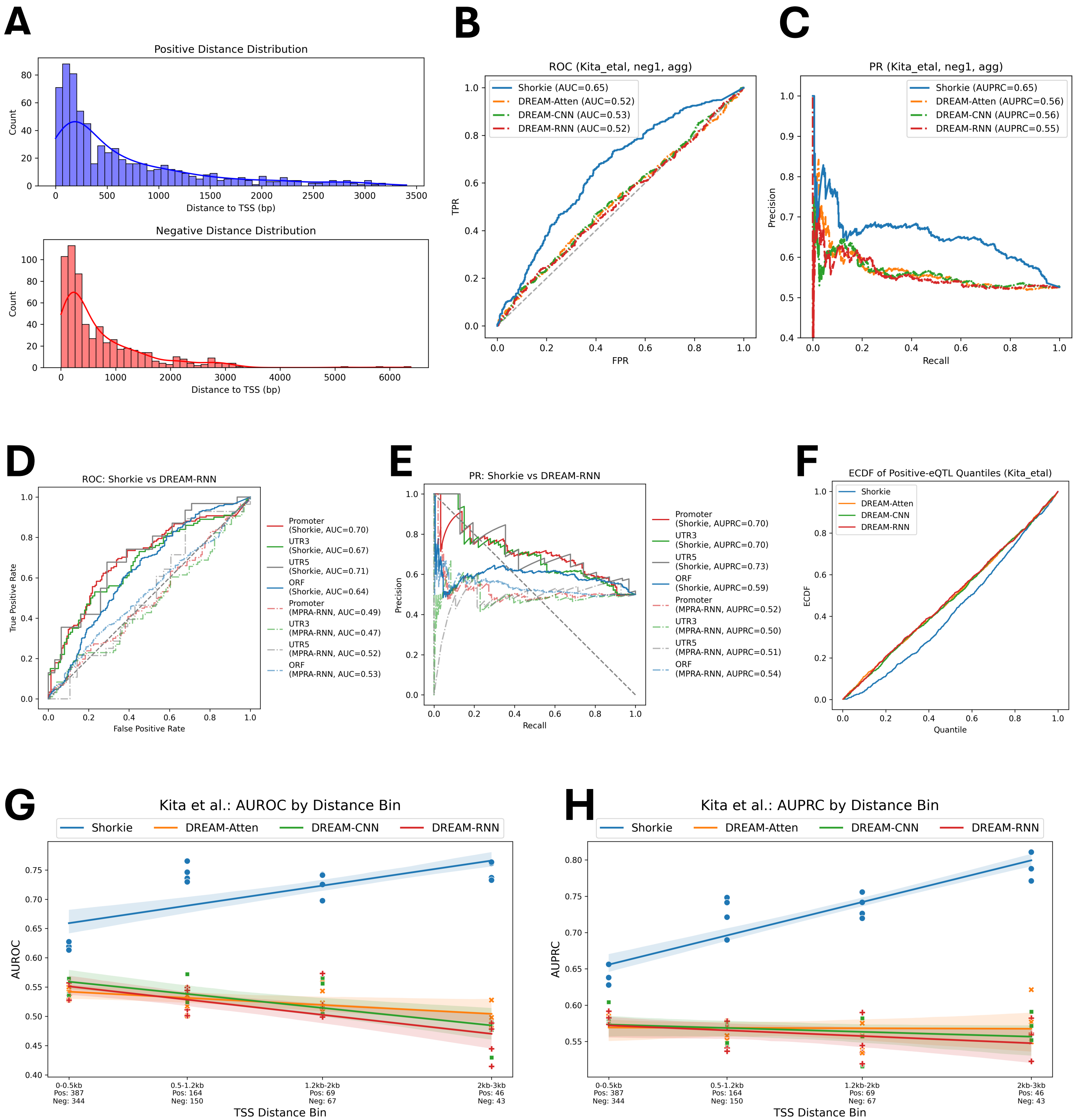


Figure S23. Evaluation of variant effect predictions on the Kita et al. eQTL dataset using Shorkie and DREAM Challenge models. **(A)** Distribution of distances from positive and negative eQTL variants to the transcription start site (TSS) of their target genes. **(B)** Receiver Operating Characteristic (ROC) curves and **(C)** Precision-Recall (PR) curves comparing the performance of Shorkie and DREAM Challenge models in distinguishing positive from negative eQTLs. **(D)** ROC and **(E)** PR curves for eQTL SNP locations: (1) promoter, (2) 3' UTR, (3) 5' UTR, and (4) open reading frame (ORF), comparing Shorkie with DREAM-RNN. **(F)** Empirical Cumulative Distribution Functions (ECDFs) of positive eQTL quantile scores for Shorkie and DREAM Challenge models. A steeper rise at lower quantiles indicates a higher enrichment of high-effect predictions among true positives. **(G)** Area under the

ROC curve (AUROC) and **(H)** Area under the PR curve (AUPRC) stratified by TSS-distance bins, illustrating the variation in model classification performance relative to genomic distance.

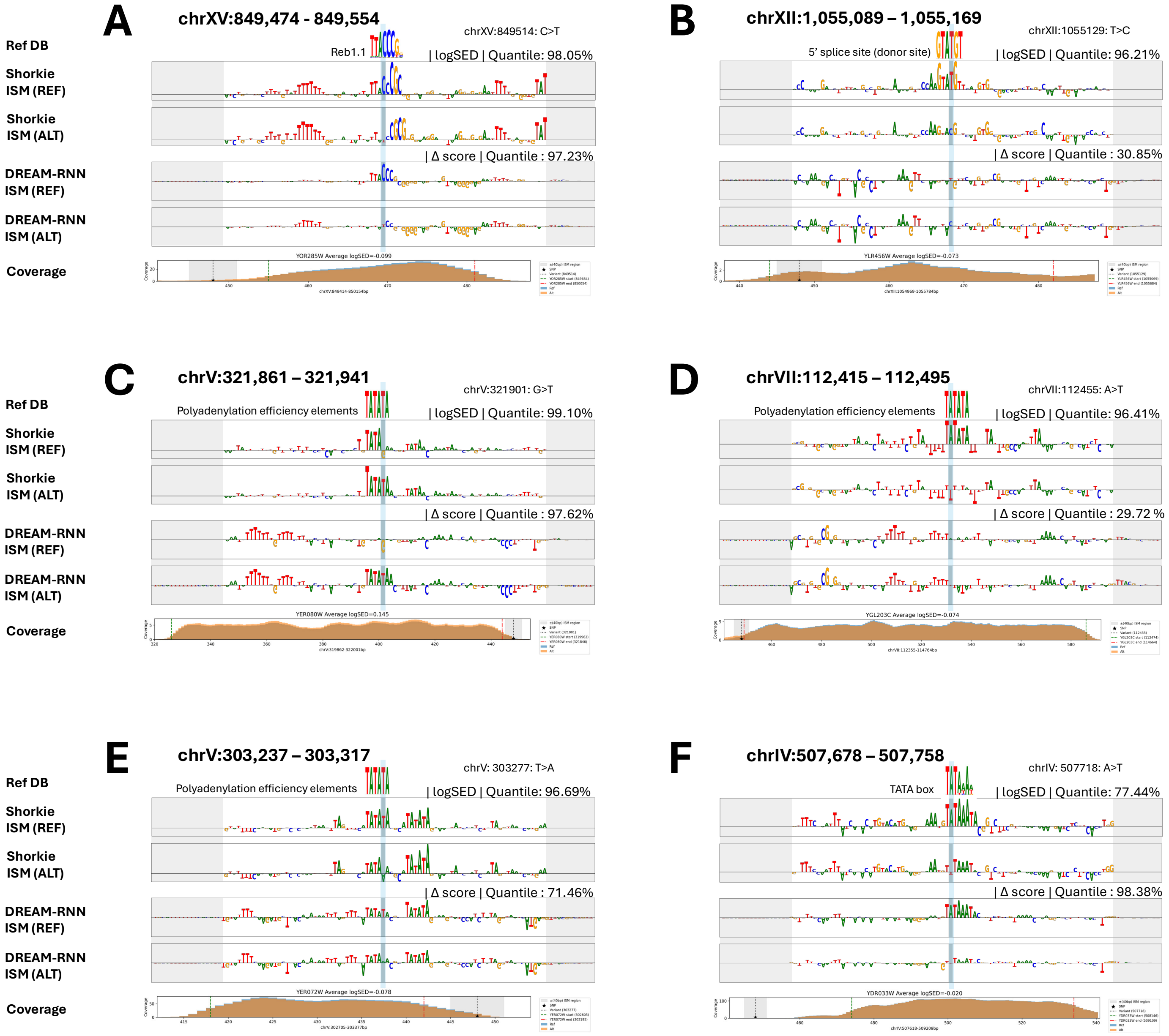


Figure S24. Shorkie *in silico* mutagenesis (ISM) analysis of predicted variant effects at eight eQTL loci. Shown are Shorkie ISM maps alongside DREAM-RNN Δ-score profiles for the reference (REF) and alternative (ALT) alleles, each plotted over an 80 bp window centered on the variant. Negative scores indicate a predicted decrease in expression for the ALT allele relative to REF, while positive scores indicate an increase. Below each panel, the exact genomic span (chromosome:start–end) and coverage (read depth) are provided.

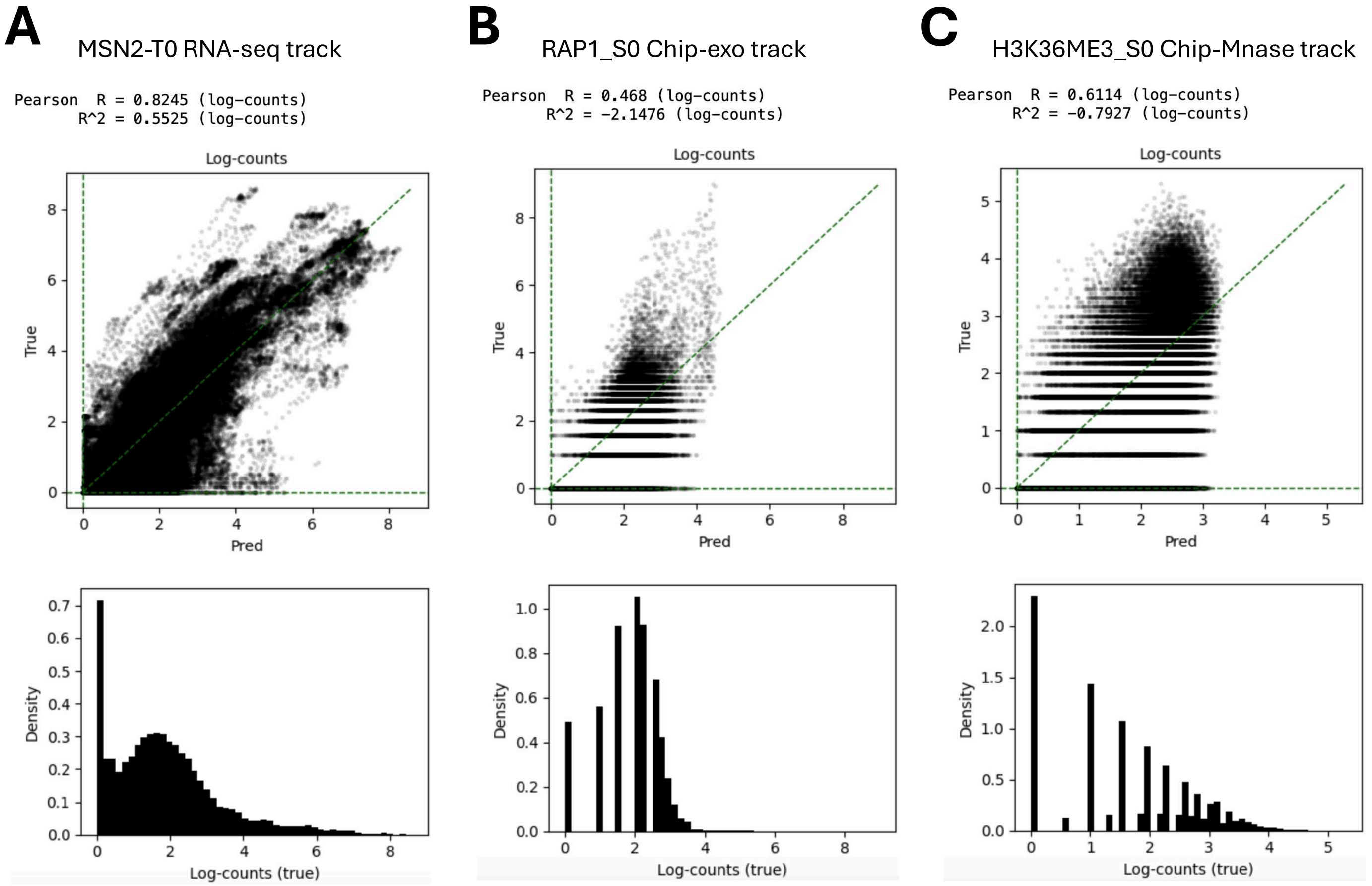


Figure S25. Bin-level predicted vs. measured signal across assays (fold-0 test set). The upper row shows scatterplots of model predictions (x-axis; log-counts) versus measurements (y-axis; log-counts) for representative tracks: **(A)** MSN2-T0 RNA-seq, **(B)** RAP1_S0 ChIP-exo, and **(C)** H3K36ME3_S0 ChIP–MNase. Each point is a genomic bin; the dashed diagonal marks the $y = x$ line (perfect agreement). Pearson's R and R^2 are computed on log-counts and reported in each panel. The bottom row shows histograms of the measured (true) log-counts for the same bins, illustrating assay-specific dynamic range. All examples use held-out regions from the fold-0 model.