RESEARCH SUMMARY

Shorkie: learning yeast regulatory code from related fungi

By Kuan-Hao Chao

Sep 19, 2025 · 13 min read · Updated Jul 1, 2026

Abstract

Shorkie asks whether related fungal genomes can provide useful pretraining signal for predicting dynamic expression in budding yeast. It follows the path from yeast-only supervised training to self-supervised fungal pretraining, showing that the best transfer came from the right evolutionary scope and then examining motifs, regulatory variants, time-dependent expression, and model limits.

Every cell carries essentially the same DNA, but cells do not use that DNA in the same way. A neuron, a skin cell, and a yeast cell responding to stress differ because different genes are turned on, at different strengths, at different times. That control program is written into regulatory DNA: transcription-factor binding sites, nucleosome-positioning sequence, promoter architecture, splice signals, and other patterns we still only partly understand.

If we could read that program directly from sequence, we could do more than predict whether a gene is active. We could test which bases influence a model's expression prediction, which variants may change it, and how regulation shifts when a cell enters a new state. That is the promise of sequence-to-function modeling. The hard part is separating learned regulatory biology from correlations and shortcuts in natural genomes.

Shorkie (Chao et al., 2025) was built for that test. It is a fungal DNA language model that is first pretrained on related genomes to learn general sequence patterns, then fine-tuned — adapted with labeled yeast data — to predict how budding yeast genes change expression over time. The central question was empirical: if yeast-only supervised training was limited by the amount of sequence available, could related fungal genomes provide a useful starting point? Shorkie asks whether we can learn regulatory grammar from those relatives, then transfer it back to Saccharomyces cerevisiae.

Why budding yeast?

Budding yeast (Saccharomyces cerevisiae) is one of the best systems we have for asking whether a model understands eukaryotic gene regulation. Its biology is deeply mapped: thousands of genes,

hundreds of transcription factors, decades of experiments on promoter logic, chromatin, stress response, metabolism, and aging. It is also compact enough to model end to end. The whole genome is only about 12 megabases.

That compactness is both the opportunity and the challenge. A supervised sequence-to-expression model learns from examples: one gene's sequence paired with its measured expression, over and over. Yeast has roughly 6,000 protein-coding genes, which is small by deep-learning standards. There is enough biological knowledge to evaluate a model carefully, but it was not obvious that yeast sequence alone would provide enough independent examples for the model we wanted to train.

That was the first thing we tested. Shorkie began as my summer-2024 internship in the Kelley Lab at Calico, where Borzoi (Linder et al., 2025) had shown how to predict RNA-seq coverage directly from DNA sequence. We trained the same supervised Shorkie architecture from random weights and pushed on the obvious knobs, including learning rate and model capacity. It fit the genes it had seen, but it did not generalize as strongly as we wanted to unseen genes. That result made pretraining the natural next experiment.

The bet: borrow power from evolution

The pretraining idea was to give the model a broader sequence context before asking it to predict yeast expression. If S. cerevisiae alone was a narrow training source for this model, its relatives might add useful signal. Across the fungal kingdom, evolution has run many related regulatory experiments. Functional sequence tends to be constrained; neutral sequence drifts. A model trained across many related genomes should see the repeated patterns often enough to learn which DNA words matter. That motivation follows the broader evidence that species-aware DNA language models can capture regulatory elements and their evolution (Karollus et al., 2024).

This is the same self-supervised idea behind language models, applied to DNA. Instead of predicting the next word in a sentence, the model sees masked stretches of genome and learns to reconstruct the hidden bases. It is not told which sites are regulatory. It has to infer useful patterns from sequence context itself.

The question was not only whether pretraining would help. It was also how far across the tree we should reach. More genomes provide more diversity, but distant organisms can carry regulatory rules that are less relevant to budding yeast. Too narrow a training set may not teach enough; too broad a training set may drown yeast's signal in unrelated biology. We suspected there would be a sweet spot.

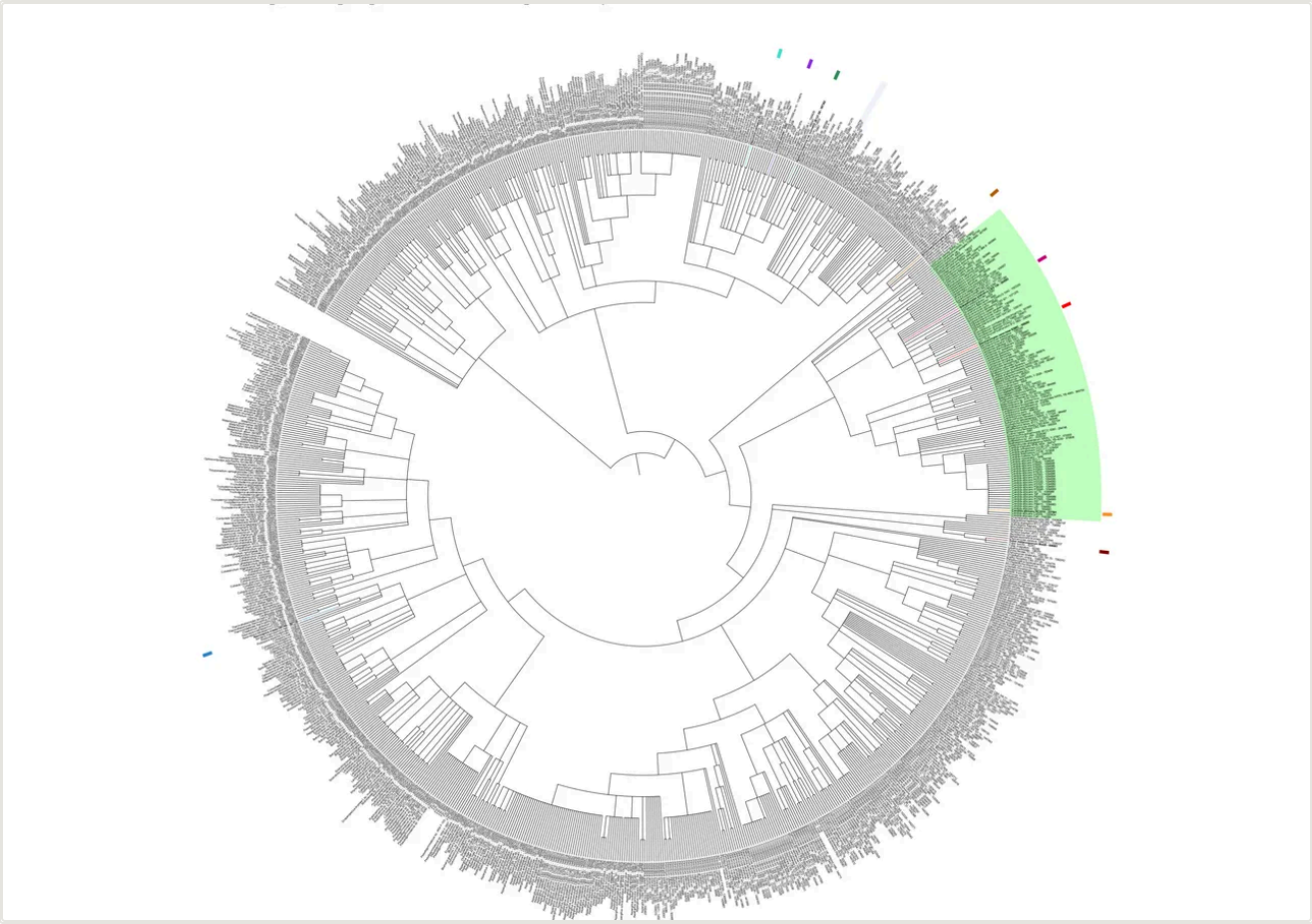


Figure 1. The tree of life we pretrain across — 1,341 fungal genomes spanning the kingdom, from oyster mushrooms, shiitake, and black truffles to the yeasts. The order Saccharomycetales around budding yeast is highlighted in green; the 165-genome slice of that order turned out to be the sweet spot for transferring regulatory grammar back to S. cerevisiae.

What Shorkie is

Shorkie has two stages.

First comes **pretraining**. We trained the same masked DNA language model on four alternative sequence corpora, each representing a different evolutionary scope:

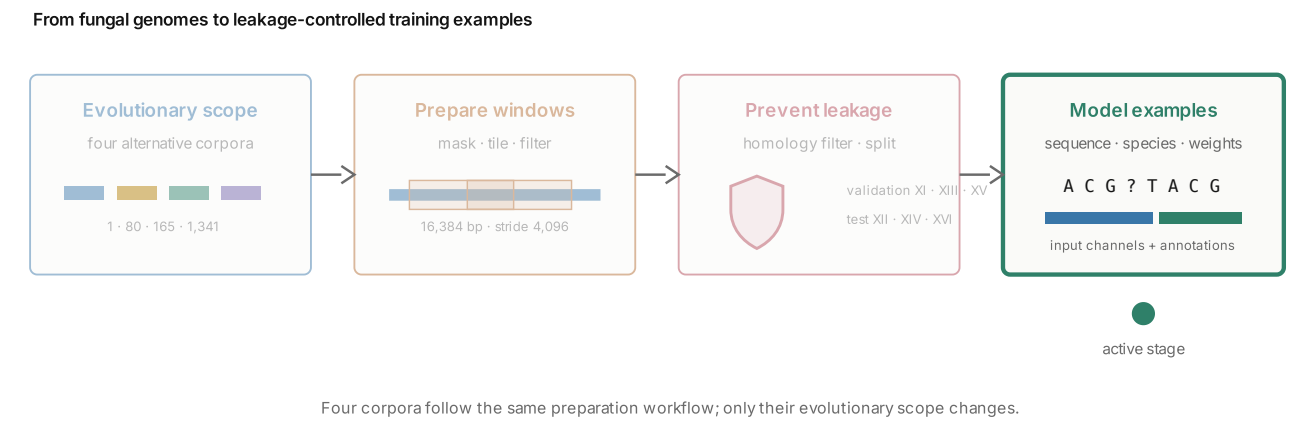
- one S. cerevisiae reference genome, R64, about 12 Mb;
- 80 S. cerevisiae strains, about 1 Gb;
- 165 genomes from the broader Saccharomycetales order, about 2 Gb;
- 1,341 genomes spanning the fungal kingdom, about 42 Gb.

This design turned the evolutionary-scope question into an experiment. If raw scale were all that mattered, the 1,341-genome kingdom model should win. If closeness to yeast were all that mattered, the

single-genome or strain models should win. Neither was quite right. Each corpus follows the same preparation workflow.

Prepare evolutionary training data

4 Create model-ready examples



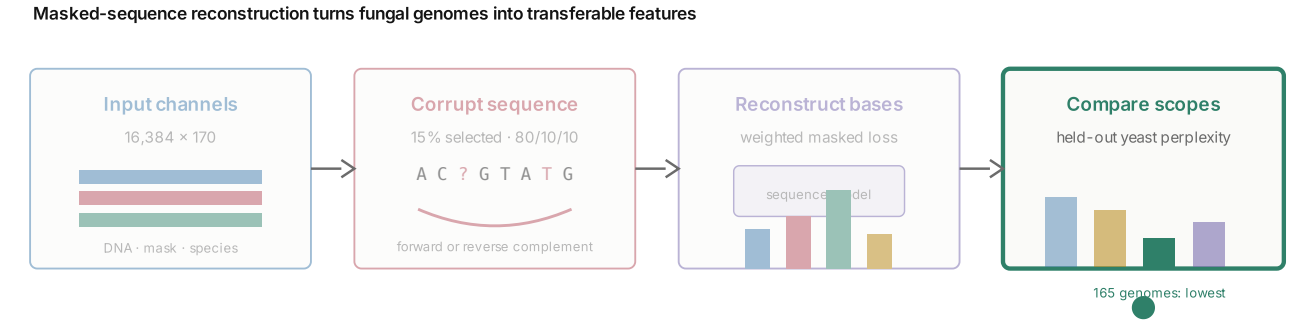
Four corpora follow the same preparation workflow; only their evolutionary scope changes.

For each scope, the pipeline repeat-masks the genomes, tiles them into overlapping 16,384 bp windows, removes repeat-heavy examples, and filters training sequence homologous to held-out yeast chromosomes. This last step prevents shared sequence from making reconstruction performance look better than it is.

The self-supervised objective then hides 15% of the bases in each training window and asks the model to reconstruct them from surrounding sequence and species context. The selected 165-genome model receives four DNA channels, one mask channel, and a one-hot identity for each species.

Learn fungal sequence context

4 Select the best evolutionary scope



Pretraining learns sequence context before Shorkie sees supervised yeast assays.

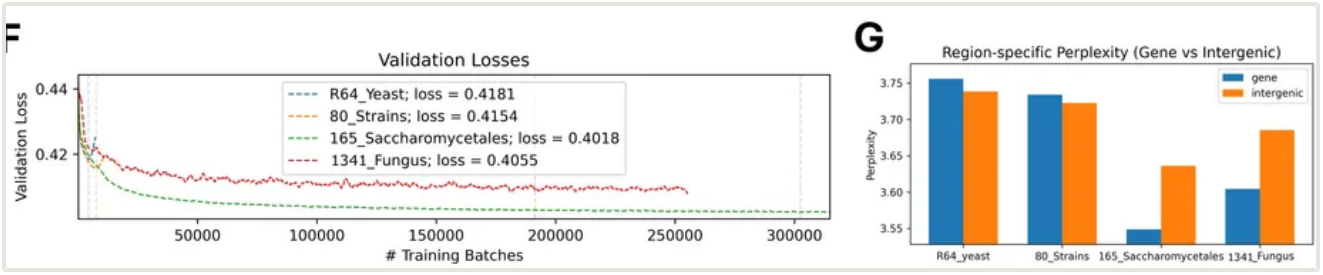


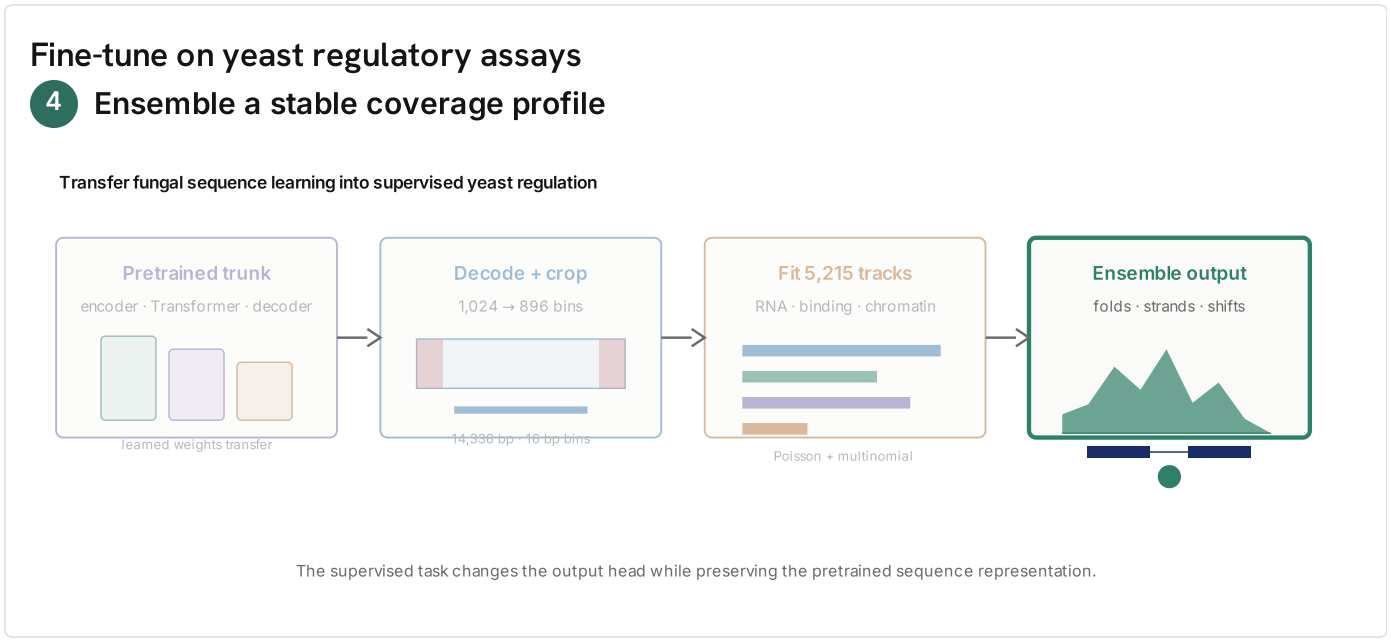
Figure 2. The language model's own report card, by evolutionary scope. Trained on each of the four corpora and scored on held-out S. cerevisiae, the masked LM reconstructs yeast best when pretrained on the 165-genome Saccharomycetales order — lowest validation loss (left) and lowest perplexity (right), ahead of the 1,341-genome kingdom set and the narrower strain and single-genome corpora. The language model itself already prefers the sweet spot.

The 165-genome Saccharomycetales model reconstructed held-out yeast sequence best. It had the lowest validation loss and the lowest perplexity, meaning the masked language model was least surprised by yeast DNA after training on that corpus. The result matters because it separates useful evolutionary diversity from sheer data volume. Fungal-kingdom pretraining helped, but it was not the best source for yeast. The closer single-genome and strain corpora helped less. The order-level set threaded the needle: broad enough to expose conserved regulatory grammar, close enough that the training signal still pointed toward yeast.

The kingdom model's noisier training is consistent with greater corpus heterogeneity and a harder optimization problem, but the experiments do not isolate one cause. Across four tested model capacities, however, the 165-genome order remained the best pretraining scope, so the observed ranking was not reversed by increasing capacity alone.

Second comes **fine-tuning**. At Calico, the team produced high-resolution RNA-seq time courses using miniaturized chemostats, or "ministats," to watch yeast genes respond minute by minute after transcription-factor induction. Shorkie fine-tunes the pretrained model on those measurements and other regulatory assays, unifying 5,215 experimental tracks: 3,053 induction RNA-seq timepoints, 1,014 strain RNA-seq profiles, 1,128 ChIP-exo transcription-factor binding datasets (Rhee and Pugh, 2011), and 20 histone-modification tracks.

Fine-tuning retains the pretrained sequence trunk but changes the task and output resolution. Three upsampling stages produce 16 bp bins; cropping the less-supported edges leaves 896 bins across the central 14,336 bp. A combined Poisson and multinomial objective trains the model to match both total signal and its distribution along the sequence.



Architecturally, Shorkie is compact. The 13.7-million-parameter language model compresses a 16,384 bp window through an initial convolution and seven residual downsampling stages to 128 positions with 384 channels. Eight Transformer layers (Vaswani et al., 2017), each with four attention heads in the released configuration, integrate long-range context. A U-Net decoder (Ronneberger et al., 2015) reconnects local sequence detail. Pretraining decodes to single-base A/C/G/T probabilities; fine-tuning stops at 16 bp resolution and predicts the regulatory tracks.

The evolutionary ranking carried into this downstream task. After identical fine-tuning, the models ranked **165-genome order > 1,341-genome kingdom > 80 strains > single genome > no pretraining**. The language model's own perplexity on held-out yeast sequence broadly tracked downstream fine-tuning performance, although not perfectly. That makes perplexity a useful screening signal before running the expensive supervised experiment.

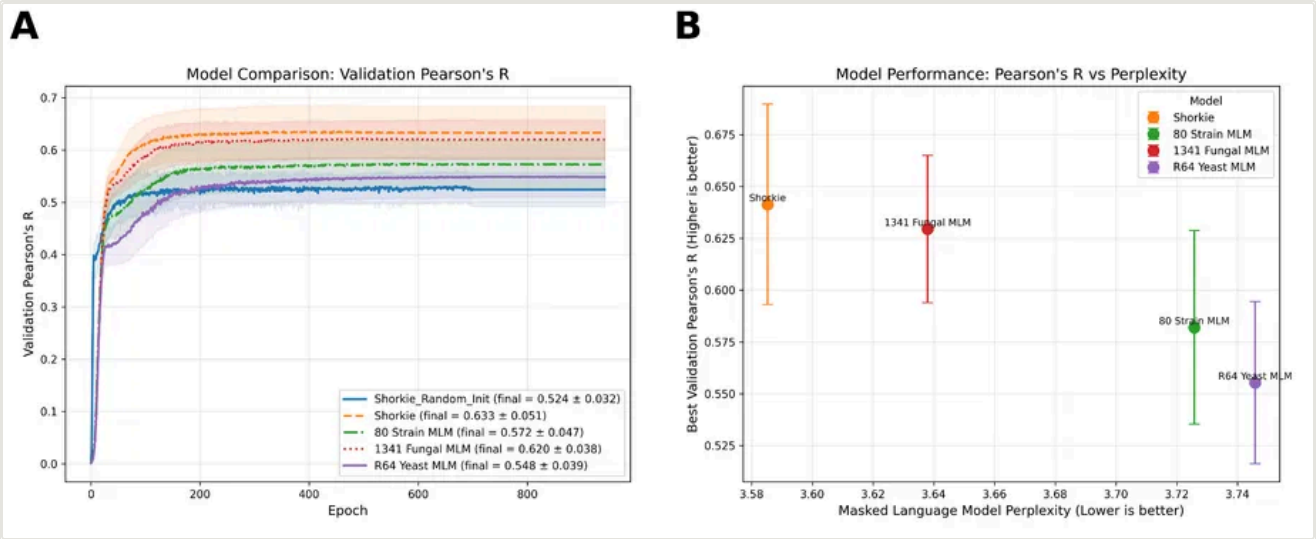


Figure 3. The pretraining choice carries through to expression prediction. After fine-tuning each pretrained model identically on the same yeast data, downstream expression accuracy (validation Pearson's r) lands in the same order — 165-genome order (0.63) > 1,341 kingdom (0.62) > 80 strains (0.57) > single genome (0.55) > no pretraining (0.52) (left). Lower held-out yeast perplexity broadly tracks better fine-tuning performance (right), making perplexity a useful but imperfect proxy for choosing pretraining data.

What it showed

The first result is the cleanest: pretraining transfers. On held-out induction RNA-seq tracks, Shorkie improves median bin-level Pearson's R (a correlation coefficient from −1 to 1, where higher means the predictions track the measurements more closely) from 0.67 to 0.78 over `Shorkie_Random_Init`. After aggregating over genes, mean gene-level Pearson's R rises from 0.74 to 0.88, and the pretrained model wins on 87.8% of genes. That 0.74 to 0.88 gene-level jump is the central result in one number. The model did not merely benefit from better tuning; we swept learning rates and tested smaller from-scratch networks, and the gap remained.

The second result is more important biologically: the models learned recognizable regulatory grammar. Using in silico mutagenesis (computationally substituting each base and measuring how much the prediction changes) and language-model sequence reconstruction, we can ask which bases the models treat as important. Without being given motif labels, Shorkie LM and the fine-tuned Shorkie model recover canonical transcription-factor motifs and regulatory sequence features: Reb1, Tye7, Cbf1, poly(dA:dT) tracts, TATA boxes, Rap1 sites in ribosomal-protein promoters, the PAC motif of ribosome biogenesis, Abf1 sites, 5′ splice donors, and branch-point signals.

That distinction matters. High predictive accuracy alone is not enough, because a model trained on natural genomes can exploit correlations that are not causal. Recovering known motifs is useful biological validation that the model has learned recognizable sequence structure, but it is not by itself proof that those features cause expression changes.

We also used two lenses that answer different questions. The pretrained language model reports what is predictable across fungal sequence. The fine-tuned model's in silico mutagenesis reports which substitutions change its expression prediction. Conservation and prediction sensitivity do not have to agree: a conserved site can be inactive in a condition, and an experimentally causal site need not be deeply conserved. Agreement between the two lenses is informative, but causal interpretation still requires perturbation data.

That is what happens at RPL26A. The language model and fine-tuned model both highlight the Fhl1 and Rap1 binding sites and the 5′ splice donor. The no-pretraining control, `Shorkie_Random_Init`, does not recover the same motif structure.

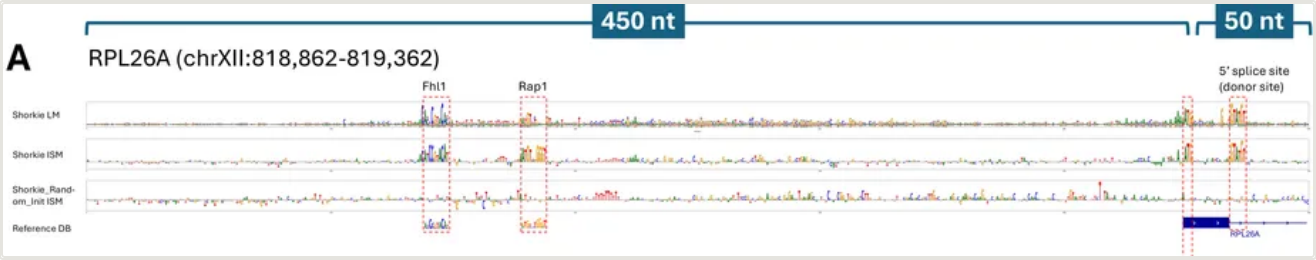


Figure 4. Reading the grammar back out, one gene at a time. At the ribosomal-protein gene RPL26A, two independent lenses agree: the pretrained language model's sequence-predictability signal (Shorkie LM) and in silico mutagenesis on the fine-tuned model (Shorkie ISM) both highlight the Fhl1 and Rap1 transcription-factor sites and the 5′ splice donor, matching the reference annotation at the bottom. The Shorkie_Random_Init track, an identical model without pretraining, does not recover the same motif-specific signal.

Variants, dynamics, and the hard caveat

A useful regulatory model should say something about variants. We tested Shorkie on two association-based yeast cis-eQTL datasets (variants statistically linked to differences in nearby gene expression) and an MPRA-validated benchmark of causal cis-regulatory variants (Renganaath et al., 2020). Across those three evaluations, Shorkie separated positive variants from matched controls better than the tested DREAM-challenge models, its randomly initialized counterpart, and the pretrained language model alone. The comparison shows that pretraining alone is not sufficient; fine-tuning on measured regulatory outputs adds task-specific signal.

The comparison has an important caveat. DREAM models were trained on synthetic promoter variants measured by massively parallel reporter assays (MPRAs). On synthetic MPRA constructs, the MPRA-trained model beats Shorkie, even though Shorkie never saw MPRA data. That is not a contradiction. Each model is strongest in the domain it was trained on: Shorkie on native genomic sequence, DREAM on designed reporter constructs.

This is part of a larger problem in regulatory genomics: natural genomes contain correlations that are not necessarily causal. MPRAs help break that structure by testing controlled sequence perturbations. In the Renganaath benchmark, Shorkie achieved the highest reported discrimination among the compared models (AUPRC 0.629 ± 0.013 and AUROC 0.618 ± 0.006 — standard precision-recall and ROC scores

for how well true positives are separated from controls). That result provides a stronger causal test than the association-based eQTL sets, while remaining specific to the variants and assay design evaluated in the paper.

The time-course data lets Shorkie address another piece of regulation: dynamics. Gene regulation is not static. During MSN2 and MSN4 induction, STRE motif signals sharpened over the first 90 minutes, matching the expected activation biology. Genome-wide, predicted and measured fold-changes after induction agreed with Pearson correlations of roughly 0.51-0.65 across 5-180 minutes, evaluating up to 218,806 gene-TF pairs per timepoint.

The bigger picture

For me, Shorkie is a clean demonstration of a principle I keep returning to in computational biology: when labeled data is scarce, borrow statistical strength from related data that still shares the underlying biology. Here, that means borrowing sequence from related fungi, then anchoring the model back to yeast with real expression, binding, chromatin, and time-course measurements.

The most useful lesson is not simply “more genomes help.” The lesson is that evolutionary scale matters. The best pretraining corpus was not the biggest one. It was the one close enough to preserve the relevant grammar and broad enough to give the model many examples of it. That sweet spot will differ by organism and phenotype, but Shorkie gives a way to look for it: train self-supervised models across candidate scopes, evaluate held-out target-organism perplexity, and fine-tune the best candidates.

There is still plenty Shorkie does not solve. Its supervised output is at 16 bp resolution rather than single-base resolution. ChIP-exo remains the most challenging assay, especially for extreme, narrow peaks, and RNA-seq time-course predictions show compressed dynamic range relative to the measurements. And the hard caveat remains: a model trained on natural sequence can still learn correlations that are not pure causes. More functionally validated variants are the honest next step.

But the throughline is clear. A small genome can be a powerful proving ground when we pair it with related genomes rather than ask it to supply all the training signal alone. A model trained across related genomes can learn a regulatory grammar that transfers back to the organism we care about. Once it does, we can interrogate that grammar directly, one nucleotide at a time.

Shorkie is open source — the code and trained models are on GitHub.

Read the preprint on bioRxiv, browse the code, or watch the talk. Shorkie was built with Majed Mohamed Magzoub, Emily Stoops, Sean R. Hackett, Johannes Linder, and David R. Kelley at Calico Life Sciences.

References

1. Chao, K.-H., Magzoub, M. M., Stoops, E., Hackett, S. R., Linder, J., and Kelley, D. R. Predicting dynamic expression patterns in budding yeast with a fungal DNA language model. bioRxiv (2025). doi:10.1101/2025.09.19.677475
2. Linder, J. et al. Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. Nature Genetics (2025). doi:10.1038/s41588-024-02053-6
3. Karollus, A. et al. Species-aware DNA language models capture regulatory elements and their evolution. Genome Biology (2024). doi:10.1186/s13059-024-03221-x
4. Rhee, H. S. and Pugh, B. F. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell (2011). doi:10.1016/j.cell.2011.05.042
5. Vaswani, A. et al. Attention is all you need. Advances in Neural Information Processing Systems (2017). https://arxiv.org/abs/1706.03762
6. Ronneberger, O., Fischer, P., and Brox, T. U-Net: convolutional networks for biomedical image segmentation. MICCAI (2015). doi:10.1007/978-3-319-24574-4_28
7. Renganaath, K. et al. Systematic identification of cis-regulatory variants that cause gene expression differences in a yeast cross. eLife (2020). doi:10.7554/eLife.62669