Genome Biology

RESEARCH Open Access

Species-aware DNA language models capture regulatory elements and their evolution

Alexander Karollus[1,2†], Johannes Hingerl[1†], Dennis Gankin[1†], Martin Grosshauser[1], Kristian Klemon[1] and Julien Gagneur[1,2,3,4,5*]

†Alexander Karollus, Johannes Hingerl, and Dennis Gankin contributed equally to this work.

*Correspondence: gagneur@in.tum.de

[1] School of Computation, Information and Technology, Technical University of Munich, Garching, Germany
[2] Munich Center for Machine Learning, Munich, Germany
[3] Institute of Human Genetics, School of Medicine and Health, Technical University of Munich, Munich, Germany
[4] Computational Health Center, Helmholtz Center Munich, Neuherberg, Germany
[5] Munich Data Science Institute, Technical University of Munich, Garching, Germany

Abstract

Background: The rise of large-scale multi-species genome sequencing projects promises to shed new light on how genomes encode gene regulatory instructions. To this end, new algorithms are needed that can leverage conservation to capture regulatory elements while accounting for their evolution.

Results: Here, we introduce species-aware DNA language models, which we trained on more than 800 species spanning over 500 million years of evolution. Investigating their ability to predict masked nucleotides from context, we show that DNA language models distinguish transcription factor and RNA-binding protein motifs from background non-coding sequence. Owing to their flexibility, DNA language models capture conserved regulatory elements over much further evolutionary distances than sequence alignment would allow. Remarkably, DNA language models reconstruct motif instances bound in vivo better than unbound ones and account for the evolution of motif sequences and their positional constraints, showing that these models capture functional high-order sequence and evolutionary context. We further show that species-aware training yields improved sequence representations for endogenous and MPRA-based gene expression prediction, as well as motif discovery.

Conclusions: Collectively, these results demonstrate that species-aware DNA language models are a powerful, flexible, and scalable tool to integrate information from large compendia of highly diverged genomes.

Introduction

A typical eukaryotic genome contains large regions of non-coding DNA. These are not translated into proteins but contain regulatory elements which control gene expression in response to environmental cues. Finding these regulatory elements and elucidating how their combinations and arrangements determine gene expression is a major goal of genomics research and is of great utility for synthetic biology and personalized medicine.

In the last decade, significant progress has been made towards understanding the regulation of model species, particularly humans and mice, by leveraging the work of large consortia such as ENCODE [1] or FANTOM5 [2]. These groups have invested considerable resources to compile massive compendia of experiments which probe many steps of transcription control at high resolution and depth.

Nonetheless, estimates indicate that there are millions of eukaryotic species [3]. Many of these have agricultural, medical, or biotechnological relevance and even those lacking direct economic importance may still hold key insights about regulatory evolution. Accordingly, it would certainly be valuable if existing approaches were extended beyond model organisms. Nevertheless, generating an ENCODE for every species is not feasible at the current level of technology. What is more in reach, however, is to sequence the genomes of all species—and this is increasingly being done [4–8].

As the genomes of all organisms are evolutionarily related, it is possible to study regulatory elements through sequence comparison, without requiring additional experimental annotations [9]. Specifically, we expect that regulatory sequences and functional arrangements thereof are selected and thus generally more conserved than expected for neutral mutations [10]. The main method to determine whether particular nucleotides are conserved makes use of sequence alignment. Unfortunately, alignment is difficult for non-coding sequences, presumably because regulatory sequences evolve faster than coding sequences and because of a certain tolerance to the exact order, orientation, and spacing of regulatory elements in regulatory sequences [11, 12]. Thus, while alignments will certainly remain valuable, their ability to integrate information across large compendia of highly diverged species is limited.

Inferring regulatory elements from genomic sequences only, without requiring further gene expression-related experimental data, i.e., labels, is reminiscent of a typical challenge in another domain, natural language processing, where vast quantities of mostly unlabelled text are available. There, masked language modeling, a type of self-supervised representation learning, presents another way to generate insights from unlabelled data [13]. To train masked language models (LMs), parts of an input sequence are hidden (masked), and the model is tasked to reconstruct them. To do so, LMs learn an internal numerical representation of words and their context, capturing the syntax and semantics of natural language. In turn, these representations can be used as features for efficiently training supervised models for many downstream tasks. This approach has alleviated the scarcity of labeled data in many natural language processing predictive tasks such as translation or question answering.

In genomics, previous work has adopted masked language modeling as a method to build sequence representations for DNA [14], although initially this was focused on single species. Only recently, multi-species datasets have been used to train large genomic language models [15–17]. These models foremost serve as a basis to build predictors of molecular phenotypes. In principle, LMs should benefit from multi-species training through the ability to leverage evolution. While first results in this direction have been very encouraging [16], a recent analysis found that a model trained only on human sequences achieves better predictions for fruit fly enhancers compared to a multispecies model that includes the fruit fly genome [17]—an observation which is difficult to reconcile with the 700 million years of divergence between these species [18]. Performance on downstream tasks after fine-tuning, i.e., using the pre-trained model as initialization to then train the full model in a supervised fashion, is an indirect measure of the features learned by multispecies pretraining since the exact features needed to succeed at the given tasks are unknown, potentially learned during fine tuning and not investigated in these works.

Due to their focus on downstream tasks, previous works considered the actual task the language model is trained to perform—reconstructing masked nucleotides—as a means to an end and do not study it. The sole exception to this trend is a recent contribution [19], which showed that poorly reconstructed nucleotides were enriched for variants with low frequency in the *Arabidopsis thaliana* population—although the strongest variants driving this signal appear to be coding rather than regulatory. Moreover, while their model was trained on eight genomes, all evaluations were done on the *Arabidopsis thaliana* genome only, which the model was also trained on. Thus, while the authors suggest that the prediction of their language model can be considered a generalized conservation score, it is unclear whether their language model leverages between-species, rather than within-species, conservation of regulatory rules, and, more importantly, if their model accounts for changes in the code between species.

In this study, we aim to address these limitations by training masked LMs on a large number of highly evolutionarily diverged eukaryotes. Compared to previous approaches, we provide species information to our models to avoid the model having to infer the species context to account for the evolution of the regulatory code. We focus on non-coding regions and explicitly evaluate whether the models have learned meaningful species-specific and shared regulatory features during training across evolution and can transfer them to unseen species. Finally, we evaluate whether sequence representations provided by the models are predictive of important molecular phenotypes, such as RNA half-life or gene expression, and encode biologically meaningful motifs.

Results

Using language models as an alignment-free method to capture the conservation and evolution of regulatory elements

Sequence alignment is a well-known and highly effective method to study the evolution of biological sequences. It can be used to detect homologies, find conserved subsequences, and pinpoint sequence motifs. We first assessed whether sequence alignments could be a viable starting point to capture conserved regulatory sequence elements in 3′ regions across large evolutionary distances. To this end, we aligned annotated 3′ untranslated regions (UTRs) and, as control, coding sequences of *Saccharomyces cerevisiae* to the genomes of a variety of fungal species (Fig. 1A). Coding sequences could be successfully aligned even between highly diverged species. In contrast, the ability to align 3′ regions almost completely disappears beyond the *Saccharomyces* genus.

Nevertheless, many regulatory sequence elements of 3′ UTRs are conserved far beyond the genus boundary. The 3′ region of the cytochrome B assembly factor CBP3 illustrates this. Experiments have shown that the RNA-binding protein Puf3 binds 3′ UTR of this gene in *S. cerevisiae* and the far diverged (about 500 Mya [20]) *Neurospora crassa* [21]. Moreover, we found that the Puf3 consensus binding motif can be found 3′ to the stop of the CBP3 homolog in almost all yeasts. However, we observed that the motif appears to be highly mobile, complicating alignment (Fig. 1B). This example

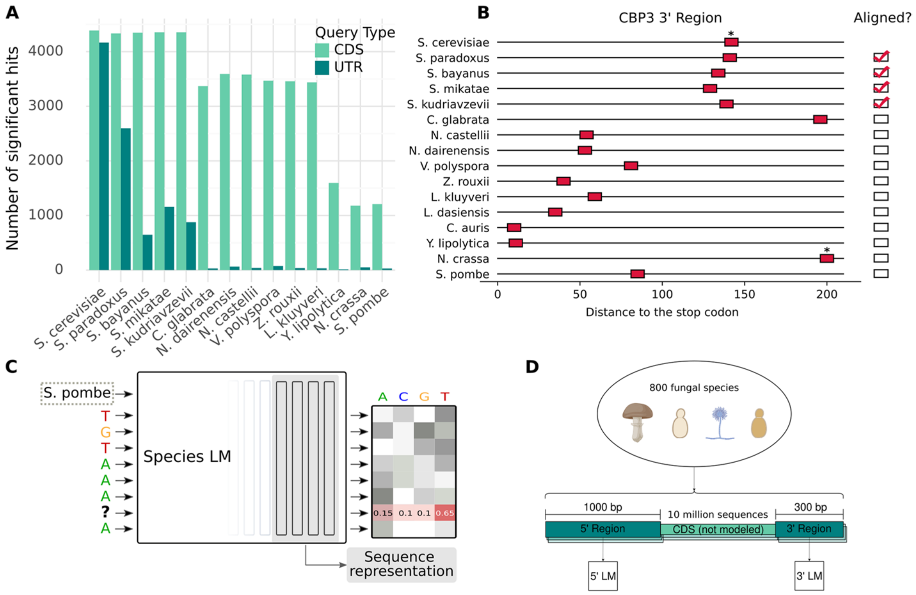

Fig. 1 Masked language modeling can serve as an alternative for alignments, which struggle to capture the conservation of regulatory elements over large evolutionary distances. **A** BLAST hits of *S. cerevisiae* CDS and 3′ UTRs in other fungal species. **B** Regions 3′ to the stop codon of CBP3 orthologues in different fungal species. Instances of Puf3-like motifs (TGTA*ATA) are indicated in red, and a star indicates experimental evidence of Puf3 binding. It appears that Puf3 binding is conserved whereas the location of the motif is not. **C** Masked language models are neural networks trained to reconstruct masked nucleotides from context. We illustrate this with the example of a Puf3 motif (TGTAAATA), where the second to last T has been hidden. Since this motif is highly conserved, the model may learn that, given this context, a T is most likely. For each masked nucleotide, the model returns a probability distribution over the letters A, C, G, and T. We can extract sequence representations from the model by pooling the hidden representations of the last four layers of the model. The architecture of the LM corresponds to DNABERT [14], with the modification that we make the model species-aware, by providing a token denoting the species where the sequence is originally from. **D** We train language models on hundreds of highly diverged fungi. In each genome, we locate the annotated coding sequences and we extract the non-coding sequences immediately before the start codon (5′ region) and immediately after the stop codon (3′ region). We train separate models for each region. Each model is trained on more than 10 million sequences

illustrates the need and potential for alignment-free approaches to leverage conservation across large evolutionary distances.

In principle, masked language modeling should be able to leverage evolutionary information without requiring alignment because their representation of sequence is more flexible and expressive, alleviating the rigid order constraint that sequence alignments subsume. Specifically, we expect that nucleotides with regulatory function should be more conserved, within and particularly across species, and therefore easier to reconstruct when masked than remaining background non-coding sequences.

Prior to applications of masked language models in genomics, methods addressed the issue of lack of alignments of non-coding sequences. Generally, these methods either make use of motif representations learned from experimental data in model organisms to investigate genomes of other species [22, 23] or are based on *k*-mer enrichments [24–28]. The language modeling approach is similar to the latter, as learning to predict masked nucleotides from context requires capturing subtle patterns of co-occurrence. However, in contrast to these approaches, language models implicitly model such patterns using flexible, nonlinear functions which enables them to learn informative representations of their input. Moreover, the attention mechanism guides the model towards the parts of the input relevant for a particular prediction. This, in theory, allows capturing high-order dependencies between nucleotides or motifs without requiring a tabulation of all possible dependencies or explicit assumptions regarding their shape and nature.

To explicitly test this, we trained masked LMs (Fig. 1C) on non-coding regions adjacent to genes extracted from a vast multispecies dataset, comprising 806 fungal species separated by more than 500 million of years of evolution. We trained distinct models for the 1000 nucleotides 5′ of the start codon (5′ region) and the 300 nucleotides 3′ of the stop codon (3′ region) of all annotated coding sequences (Fig. 1D). The 5′ region typically contains the 5′ UTR and the promoter of the gene [29]. The 3′ region typically contains the 3′ UTR [30]. Hence, we expect these regions to be enriched for non-coding sequences and to capture both transcriptional regulatory elements as well as post-transcriptional regulatory elements involved in mRNA stability and translation. Importantly, no annotation of UTR, promoters, nor transcription start site or polyA site was provided to the model. Models were trained with (species LM) and without species tokens in the input (agnostic LM). However, we provided no information about the phylogeny of the species, nor did we indicate to the models which sequences flank orthologous genes. The models were trained on all extracted sequences jointly. We focused on fungi, since many fungal genomes are available, they evolve quickly and their transcriptional control generally makes less use of extreme long-range interactions which are difficult to model with current approaches. We held out the entire *Saccharomyces* genus to test the generalization performance of our model in the well-studied species *Saccharomyces cerevisiae*.

Language models reconstruct known binding motifs in an unseen species

Binding motifs, which represent the sequence specificities of DNA and RNA-binding proteins (RBP), are considered to be the “atomic units of gene expression” [31]. Accordingly, to verify that LMs can capture aspects of the regulatory code in an alignment-free manner, we first needed to verify that they capture important known motifs.

To test this, we analyzed to what extent our 3′ and 5′ LMs could reconstruct nucleotides in the held-out species *S. cerevisiae*. We compared the reconstruction obtained by our LMs with a number of baselines, including approaches based on *k*-mer frequencies and alignment of species beyond the genus. We also computed the reconstruction achieved by aligning *S. cerevisiae* with other *Saccharomyces* species, which can be regarded as an estimate of the upper range of achievable reconstruction.

We found that all approaches—except the alignment of close species—perform similarly when we compute the reconstruction accuracy over all nucleotides (Fig. 2A and Additional file 1: Fig S1). This changes drastically when considering the reconstruction of known motifs. Instances matching the Puf3 consensus motif, for example, are reconstructed almost as well by the species-aware 3′ model as they are by the alignment of close species (Fig. 2A), strongly outperforming the alignment with far species and other baselines. We observe similar results for other RBP motifs, as well as the consensus motifs of a number of transcription factors (TF) in the 5′ region (Additional file 1: Fig S1). Remarkably, by applying Modisco clustering [32] on all nucleotides weighted by their information content as computed by our LMs, we were able to recover some of

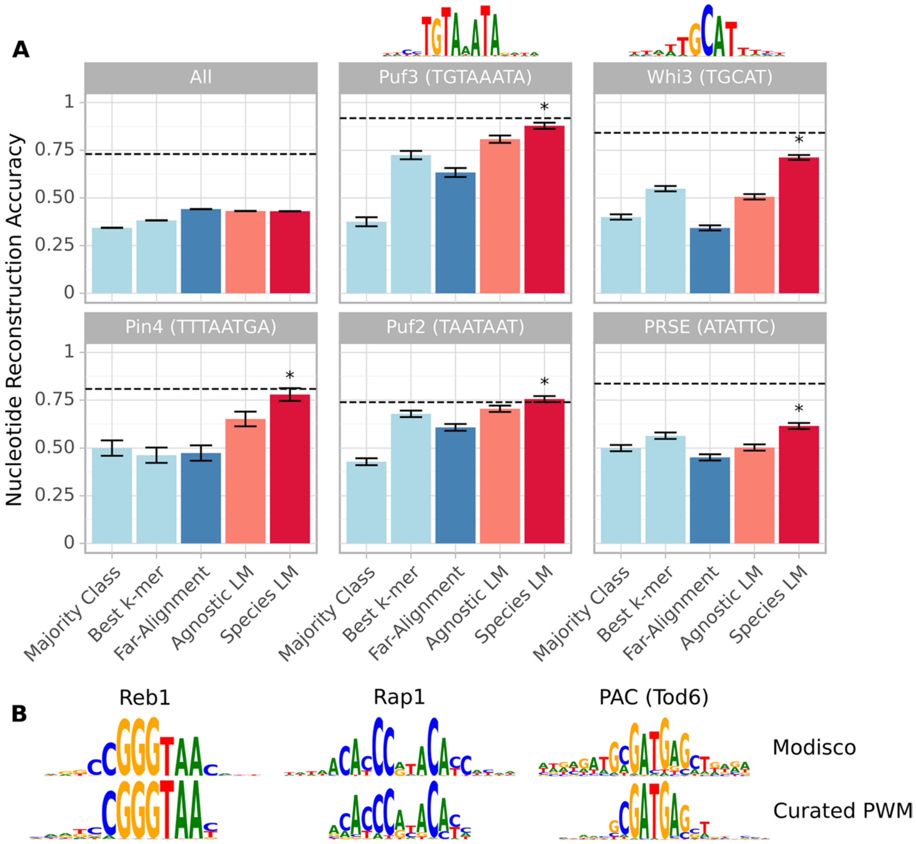

Fig. 2 Language models reconstruct likely regulatory sequences in a held-out species and recover known binding motifs. **A** Reconstruction accuracy for nucleotides within instances of RNA-binding protein consensus motifs and across all nucleotides in *S. cerevisiae* 3′ UTR sequences (those longer than 300 bp have been truncated). We compare the agnostic and species 3′ LM to a variety of baselines. The dashed line represents the accuracy achieved by the intra-genus alignment. Star indicates that the species LM significantly ($P < 0.05$, binomial test) outperforms the best baseline. For Puf3 and Whi3, Modisco clustering on the species LM reconstructions recovers the motif (depicted above the respective plots). **B** A sample of known transcription factor motifs recovered by applying Modisco clustering to the 5′ species LM reconstructions, (manually) matched to the respective high-confidence PWM from the YeTFaSCo [34] database

these motifs de novo (Fig. 2B, Additional file 1: Fig S2). This approach seems to recover a similar number of known motifs as STREME [33], a dedicated motif-finding tool (Additional file 1: Motif Discovery), although developing a general-purpose motif-finder based on the LM was out of the scope of this paper.

For many motifs we tested, the species-aware language models reconstructed slightly better than their already strong agnostic counterparts. In sum, our analysis clearly demonstrates that language models trained on a large compendium of highly diverged genomes are (1) able to learn conserved regulatory elements and (2) able to transfer this knowledge to unseen species.

Context-sensitive reconstruction of motifs is predictive of in vivo binding

It has been shown for many genomes that only a fraction of the instances of a particular motif is functional, i.e., binds the respective protein in vivo [35–37]. This is because, among other reasons, binding also depends on the context of the motif. Therefore, we next sought to evaluate whether our models can recognize these relationships.

An important piece of context for many motifs is their position with respect to certain genomic landmarks. One example of this is that RBP motifs can only be functional if they are located in a transcribed region. Another example is that in yeast TF binding sites tend to be located relatively close to the transcription start site (TSS) [38]. As a first test of whether our LMs are capable of locating these genomic landmarks—despite their location not having been indicated during training—we computed the actual and predicted nucleotide biases as a function of the distance to the TSS (imputed using CAGE data [39]) and the distance to the end of the 3′ UTR [30]. In both cases, we observed that the LMs track local changes in nucleotide biases (Fig. 3A, Additional file 1: Fig S3).

To explicitly test whether the 3′ LM locates and accounts for genomic landmarks when reconstructing motifs, we compared the reconstruction of instances of the Puf3 consensus motif located within annotated 3′ UTR regions to those that are located beyond the 3′ UTR yet still within 300 bp of the stop codon. We find that motif instances within the annotated 3′ UTR are reconstructed significantly better (Fig. 3B–C), consistent with the function of RBPs. In contrast, the phastCons [40] score, an alignment-based measure of conservation, appeared to be a poor predictor of whether a Puf3 site is within a 3′ UTR or not. We repeated this analysis for the other 3′ UTR motifs, finding similar results (Additional file 1: Fig S4A, B). For a TF, we expected this relationship to be reversed, and indeed the E-box motif was reconstructed better when found outside of 3′ UTR regions (Additional file 1: Fig S4C).

Having shown that the models did not simply learn a mere lexicon of over-represented motifs, but instead seemed to account for the context in which motif instances occur, we next asked whether the reconstruction fidelity could predict whether a motif is bound in vivo. To this end, we compared the reconstruction of Puf3 motif instances located in 3′ of genes that have been experimentally verified to bind Puf3p [41] to those without verified binding. Strikingly, we found that the reconstruction fidelity serves as a predictor of whether a gene containing a Puf3 consensus motif is bound by Puf3p—despite our models never having been exposed to binding data (Additional file 1: Fig S5).

We repeated a similar analysis for Tbf1 (Fig. 3D) and a variety of other transcription factors (Additional file 1: Fig S6, S7). Specifically, we evaluated the reconstruction of the consensus binding motifs of these TFs as a function of their distance to the closest upstream TSS. We observed that reconstruction improved when the motifs were located at a biologically plausible distance [38] from the TSS. Moreover, the reconstruction fidelity is highly predictive of in vivo binding to motif instances as measured by Chip-Exo [42], outperforming the phastCons score and, in some cases, expert-curated PWMs constructed using binding data (Fig. 3E, Additional file 1: Fig S6, S7) [34].

Distinct motifs have been established to exhibit associations with specific groups of co-regulated genes. An example in *S. cerevisiae* is the Rap1 motif, which is found primarily in the promoters of ribosomal protein genes [43, 44]. Accordingly, we find that instances of the Rap1 consensus motifs tend to be better reconstructed if they are found within 1 kb 5′ of a ribosomal protein gene (Additional file 1: Fig S8A). In other words, the reconstruction of the Rap1 motif serves as an indirect predictor of whether a gene belongs to the ribosomal protein module. We performed a similar analysis for

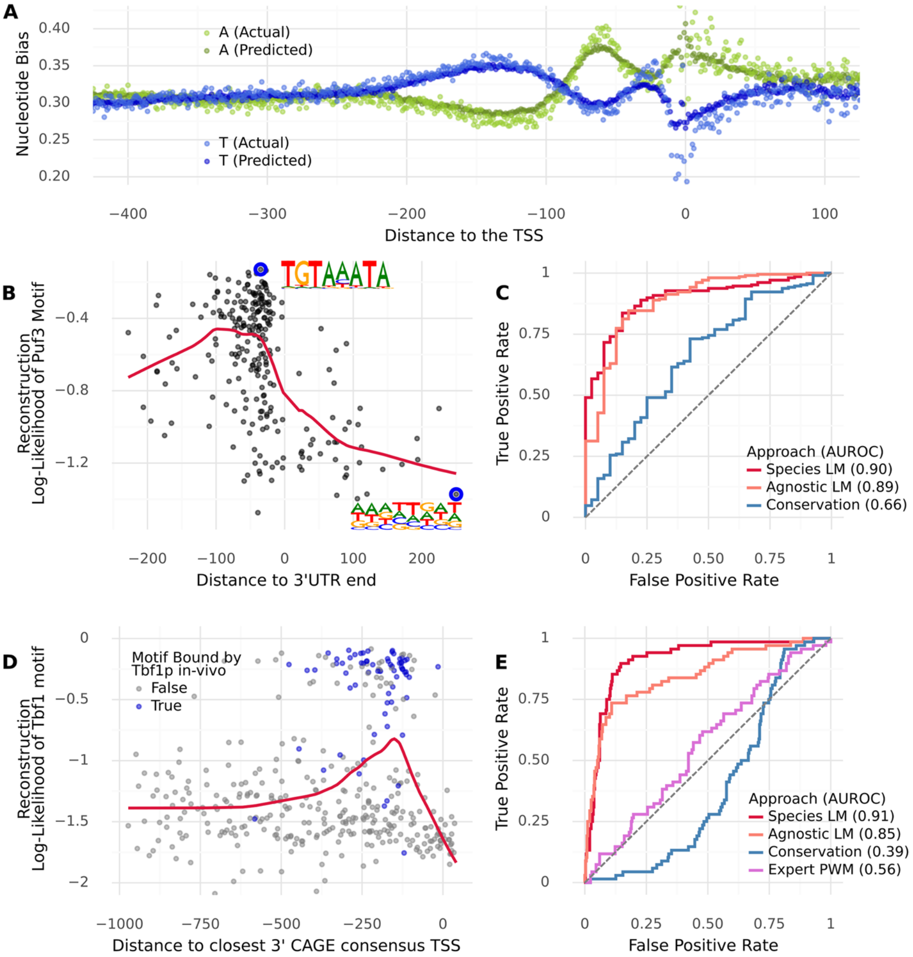

Fig. 3 Reconstruction of motifs depends on the context and predicts whether a motif instance will be bound in vivo. **A** Actual and predicted (by the 5′ species LM) nucleotide biases as a function of the distance to the TSS (imputed using CAGE data). The model keeps track of local variations in nucleotide biases. **B** Reconstruction fidelity (log-likelihood of the individual observed nucleotides, averaged per motif instance, according to the 3′ species LM) of instances of the Puf3 motif (TGTAAATA), as a function of the distance to the end of the annotated 3′ UTR. The predictions of the model for masked nucleotides are indicated for two instances of the motif (blue circles). Reconstruction fidelity is notably degraded beyond the 3′ UTR end ($P = 2.2 \times 10^{-15}$, Mann–Whitney *U*). **C** ROC curve evaluating to what extent the reconstruction fidelity of our 3′ LMs, as well as the phastCons conservation score, can serve as a predictor of whether a Puf3 motif instance is within or beyond the 3′ UTR boundary. The LMs greatly outperform the conservation score. **D** Reconstruction fidelity (log-likelihood of the observed nucleotides according to the 5′ species LM) of instances of the Tbf1 consensus motif (ARCCCTA), as a function of the distance to the closest 3′ TSS (imputed using CAGE data). Blue indicates that the motif instance was bound in vivo according to Chip-exo data. Motif instances that are around − 100 to − 250 nt to the TSS are better reconstructed than those further away or in the 5′ UTR ($P = 1.2 \times 10^{-11}$, Mann–Whitney *U*). **E** ROC curve evaluating to what extent the reconstruction fidelity of our 5′ LMs, as well as the phastCons conservation score and an expert-curated PWM, can serve as a predictor of whether a Tbf1 motif instance is bound in vivo. The LMs again greatly outperform the alternative methods

the RRPE motif, which is primarily found near genes involved in ribosome biogenesis, and obtained similar results (Additional file 1: Fig S8B).

Overall, this analysis demonstrates that LMs do not just learn a lexicon of conserved motifs, but additionally pick up on correlations between the motifs and their context which are predictive of whether motif instances are bound in vivo. Notably, this is achieved purely from genomic sequences without requiring any additional experimental data during training. This suggests that the attention mechanism provides an effective way to integrate motif interactions, although determining which exact interactions are learned is difficult to disentangle. Finally, the ability to outperform the phastCons conservation score shows the advantages of an alignment-free approach.

Language models account for changes in the regulatory code between species

Our previous analyses have focused on the held-out species *S. cerevisiae*. However, one of the main use cases we envision for genomic LMs is to serve as a method to explore understudied species. We thus analyzed how the LMs perform when evaluated across fungal species.

In Fig. 1B, we showed that the Puf3 motif, but not its location, is conserved in the 3′ regions of CBP3 homologs. We applied our 3′ model to these sequences and found that, in most species, the Puf3 motif tends to be well reconstructed compared to the background (Fig. 4A). We next applied Modisco clustering to the predictions of the model on the 3′ regions of all CBP3 homologs in our dataset. We recovered the Puf3 motif, as well as two versions of the Puf4 motif (Additional file 1: Fig S9A) [45]. This indicates that our method allows motif discovery not just across genes in one organism, but also for individual genes across organisms.

Over evolutionary timescales, certain motifs may either change drastically or fully disappear, particularly if the binding protein evolves or is lost. As an example, we considered Rap1p, a conserved protein known to control telomere length [46]. However, as noted previously, Rap1p additionally acts as a regulator of ribosomal protein expression in certain parts of the yeast lineage, a change that is associated with the acquisition of a transactivation domain by Rap1p [44]. To determine whether our models can reflect such changes, we analyzed the reconstruction of the *S. cerevisiae* Rap1 motif in 60 fungal species. Specifically, we computed for each species the difference in reconstruction of instances of the consensus motif and instances that correspond to shuffled versions thereof. This procedure controls for GC content and general differences in reconstruction fidelity between species. We found that in species close to *S. cerevisiae*, the motif instances are reconstructed significantly better than the shuffled instances (Fig. 4B). By contrast, in species where we cannot find a significant BLAST match to *S. cerevisiae* Rap1p we observed no such enrichment. An example of this is *Y. lipolytica*, a species known to not have a Rap1p homolog [47]. We performed a similar analysis for two other motifs, finding consistent results (Additional file 1: Fig S9B–C).

In addition to motif evolution, the proper context of motifs can change as well. A famous example is the positional constraint of the TATA box. In most eukaryotes including the *Schizosaccharomyces* yeasts, the TATA box is preferentially located about 30 bp 5′ from the TSS [29]. Budding yeasts, however, use a scanning mechanism to initiate transcription and therefore the position of the TATA box in these species is more flexible, but usually located between 50 and 120 bp 5′ from the TSS [29]. To verify whether our model correctly recapitulates these constraints, we located all instances of TATAWAWR in the respective species. We then assessed how well the model reconstructs these nucleotides as a function of the distance to the closest upstream TSS [39]. We found

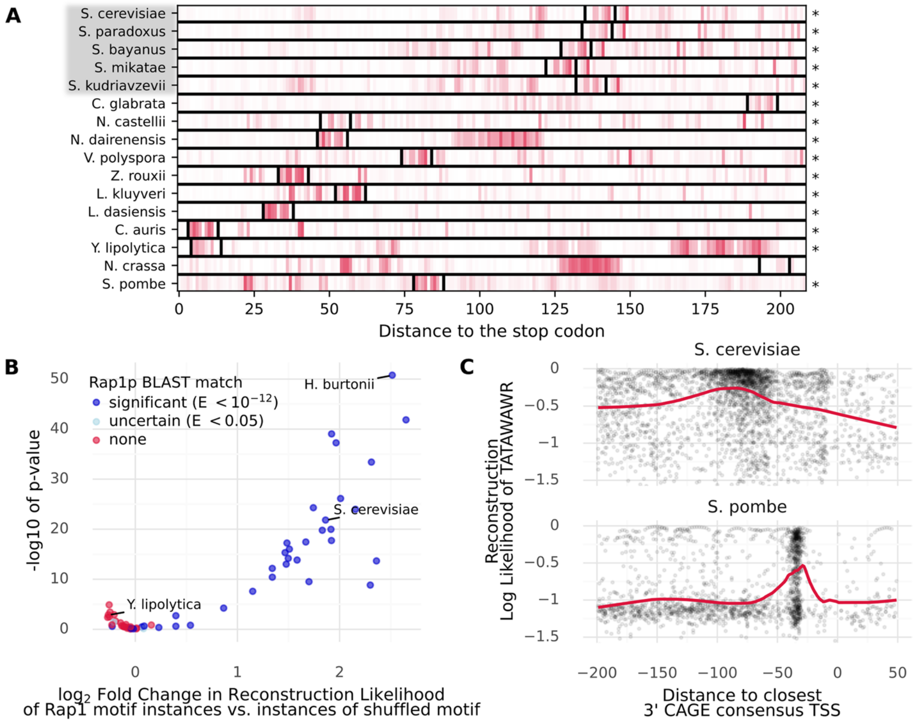

Fig. 4 LMs can trace the movement and disappearance of motifs across species and they account for the evolution of the transcription initiation mechanism. **A** We applied the 3′ species LM to the 3′ region of CBP3 homologs in a number of fungal species (compare Fig. 1A). Darker color indicates that the model assigns a higher probability to the correct nucleotide at that position. In most species, the Puf3 motif instance (delineated with black bars) is notably reconstructed better than the remaining nucleotides. Star indicates that this difference in reconstruction is significant ($P < 0.05$, Mann–Whitney *U*). Species with gray background were held out during LM training. **B** We computed the reconstruction fidelity (log-likelihood) achieved by the species 5′ LM for the *S. cerevisiae* consensus Rap1 motif (CAYCCRTACAY) instances and for instances matching shuffled versions of this motif in 60 fungal species. The difference in reconstruction between the true and shuffled motif instances, expressed as $\log_2$ fold change, is plotted against the $-\log_{10}$ *P*-value of this difference, computed using a Mann–Whitney *U* test. We observe that in species that have no BLAST match to *S. cerevisiae* Rap1p, the reconstruction fidelity of the *S. cerevisiae* Rap1 motif is generally not much better than that of shuffled versions thereof, indicating that the model correctly accounts for species context when reconstructing motifs. **C** Reconstruction fidelity (log-likelihood of the observed nucleotides according to the 5′ Species LM) of instances of the TATA-box (TATAWAWR), as a function of the distance to the closest 3′ TSS (imputed using CAGE data). Positive values indicate that the TATA-box instance is located in the 5′ UTR. We observe that in *S. pombe*, the TATA-box is best reconstructed when located ca. − 30 bp to the TSS. In *S. cerevisiae*, which uses a scanning mechanism to initiate transcription and therefore allows more flexible positioning of the TATA, the model reconstructs TATA well overall, but somewhat better when located 50 to 120 bp 5′ from the TSS

that in *S. pombe*, reconstruction fidelity of the TATA box notably peaks around 30 bp 5′ to the TSS, whereas instances located further or beyond the TSS are generally reconstructed poorly (Fig. 4C). In *S. cerevisiae*, on the other hand, TATA boxes were generally well reconstructed (likely also reflecting the AT-bias of the budding yeasts), but we observed a peak in the region 120 to 50 bp 5′ of the TSS. Thus, the model applies specific constraints when reconstructing motifs in a way that reflects the evolution of the initiation code.

One feature of the species LM is that it learns a representation for each species in the training dataset. To explore what information is contained in this representation, we first performed PCA analysis on the species representations of the 5′ species LM. We found that the first principal component, which explains about 10% of variance, is highly correlated with GC content of the respective species ($r = -0.86$, Additional file 1: Fig S10). Next, we found that species of the same taxonomic class tend to be closer together than those of different classes (Additional file 1: Fig S11, S12). This suggests that the representations encode features that at least partially correspond to the fungal taxonomy.

Next, we investigated whether changes in motif preferences are also captured by the species representation. To this end, we considered the previously characterized differential preferences of TF motifs in nucleosome-depleted regions between *S. cerevisiae* and *C. albicans* [48, 49]. Specifically, *S. cerevisiae* nucleosome depletion was associated with the Reb1 motif, whereas in *C. albicans* it was associated with the *k*-mer aCACGAC c ("Tsankov" motif here and elsewhere)—which to our knowledge is not known to have any role in *S. cerevisiae*. We compared how the reconstruction accuracy achieved by the 5′ species LM for these motifs is affected by swapping the species token. We find that with our *S. cerevisiae* proxy token (*K. africana*), the species LM reconstructed Reb1 instances better, both in *S. cerevisiae* and *C. albicans* 5′ regions (Additional file 1: Fig S13). Conversely, with the *C. albicans* token, the *C. albicans*-specific Tsankov motif is reconstructed better in both species.

In conclusion, we find considerable evidence that LMs can account for changes in the regulatory code and learn meaningful motif-context relationships in a species-specific manner.

Species-aware language model representations encode biologically meaningful features and are directly predictive of many molecular phenotypes

While our previous analyses demonstrated that the reconstructions of any LM are already very informative and could potentially be used to explore regulation in understudied species, we can also extract the learned sequence representations and leverage them to predict gene-expression-related traits in a supervised fashion. We note that this makes most sense for tasks that are data-constrained, such as gene-level measures of expression or half-life where there can only be as many data points as there are unique genes/transcripts. Accordingly, to test the predictive power of the representations themselves, we selected several gene-level omics assays, including RNA half-life measurements in *S. cerevisiae* and *S. pombe*, RNA-seq-based gene expression in *S. cerevisiae* and microarray measures of condition-specific gene expression for a number of yeast species [50–54]. We additionally included three reporter assays testing 3′ sequences [55, 56] and promoter sequences in isolation [57].

We then used our masked language models to generate sequence representations for the different sequences assayed in the respective experiments (Methods). We trained linear models in cross-validation using LM representations as input for different tasks. We used linear models specifically to ensure that the predictive power derives mostly from the LM and not from a fine-tuning procedure or a heavily engineered nonlinear fitting with many tunable hyperparameters. As a baseline for what can be achieved using "naive" sequence representations, we also trained linear models on *k*-mer counts. Furthermore, if available, we compared against state of the art.

We found that LM sequence representations outperform simple sequence representations based on *k*-mer counts across all tasks, a finding consistent with previous work

[16]. Remarkably, the species LM performed best on all tasks (Fig. 5A, Additional file 1: Table S1). On a task where most of the signal comes from the coding sequence, it performed on par (better, but not significantly) with expert hand-crafted features (Cheng et al. [50]) to predict mRNA half-life. Furthermore, simple ridge regressions trained on species-aware representations significantly outperformed hyperparameter-optimized deep neural networks (Zrimec et al. [53]) on gene expression prediction from non-coding sequence. This clearly shows that species LMs learn sequence representations rich in information without requiring labeled data.

One of the datasets we considered, the Shalem et al. 3′ MPRA [55], used a tiled mutagenesis design to measure the effect of individual subsequences of native 3′ sequences on expression. Here, the species LM performs extremely well, outperforming the agnostic model by 20 percentage points of explained variation (Fig. 5B, C). The results of this controlled experiment demonstrate that supervised models trained on species-LM sequence representations can capture causal determinants of expression found in 3′ UTR sequences.

While LM representations prove advantageous on species included in the dataset (*S. pombe*), we emphasize that they generalize to unseen species (*S. cerevisiae*). To ensure that LM performance is independent of dataset composition, and to ascertain that

A

	S. cerevisiae RNA half-life^{41}	*S. pombe* RNA half-life^{43}	*S. cerevisiae* MPRA^{40}	*S. cerevisiae* gene expression^{44}			Promoter MPRA^{47}	*C. glabrata* microarray^{45}
Region	3’	3’	3’	3’	5’	3’ + 5’	5’	3’ + 5’
Best k-mer	0.580	0.450	0.337	0.046	0.080	0.094	0.230	0.181
SOTA	0.586	0.422		0.230	0.414	0.492	0.440	
Agnostic LM	0.578	0.479	0.532	0.240	0.365	0.438	0.352	0.425
Species LM	**0.592**	**0.494***	**0.704***	**0.297***	**0.492***	**0.550***	**0.454***	**0.456***

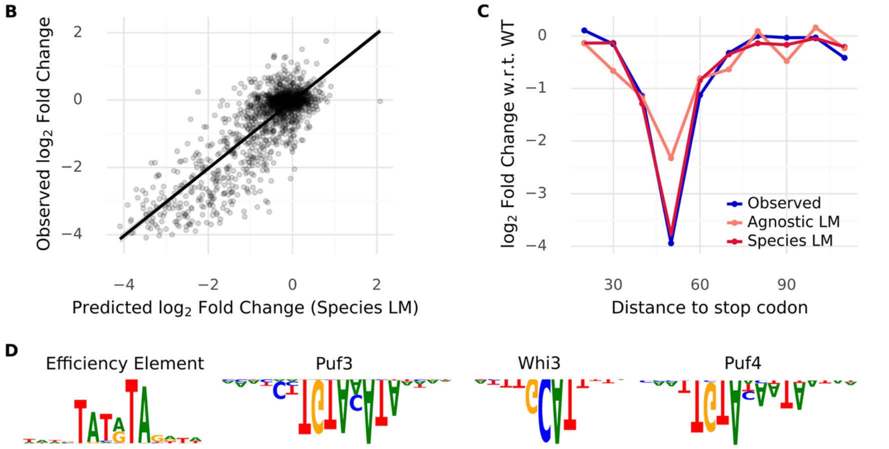

Fig. 5 Sequence representations of the species LM outperform other methods on a variety of downstream tasks. **A** Performance (R^2) of linear models trained on embeddings from language models compared to state-of-the-art models and *k*-mer count regressions, where the best k from {3, 4, 5} is shown. Star indicates that the Species LM significantly ($P < 0.05$) outperforms the second best. **B** Effect of mutation of 3′ sequences on expression. Observed $\log_2$ fold changes, as measured in Shalem et al. [55] are well predicted by the species LM representation. **C** Observed and predicted effects of mutation on expression as a function of distance to the stop codon for the YDR131C 3′ sequence. **D** Motifs recovered through in-silico mutagenesis followed by Modisco clustering on our linear model for the *S. cerevisiae* half-life task. Motifs with a negative effect on half-life are depicted upside down. We recover (2 of 4) motifs found by Cheng et al. [50]: the Puf3 motif and the Whi3 motif. Additionally, we find two motifs not found by this previous analysis, the Puf4 motif and the efficiency element, both of which have known effects on RNA stability

training on more species is beneficial, we repeated these analyses using different models pre-trained on different sets of species. We found that for all tasks, training on more species in a species-aware fashion led to better predictions (Additional file 1: Table S2).

However, in genomics, models are often used not primarily for prediction but rather as a method to discover motifs or mechanisms in large datasets. To verify that supervised learning models trained on LM sequence representations are suitable for motif discovery, we applied a standard model interpretation workflow (in-silico mutagenesis, followed by Modisco clustering, Methods) to the *S. cerevisiae* mRNA half-life data. In this way, we recovered two out of the four motifs originally found by Cheng et al. [50]: the Puf3 and the Whi3 motif—both of which our model correctly predicts as being destabilizing (Fig. 5D). We additionally found one more destabilizing element, namely the Puf4 motif [58] and one stabilizing element, the efficiency motif, which was shown to have large positive effects on expression in the Shalem mutagenesis reporter assay [55]—where we also recovered this motif using the same technique. In sum, applying standard interpretation techniques with minimal manual tuning to our embedding-based approach yielded biologically meaningful sequence features.

Discussion

In this study, we trained language models on the genomes of hundreds of fungal species, spanning more than 500 million years of evolution. We specifically directed our attention to non-coding regions, examining the ability of the models to acquire meaningful species-specific and shared regulatory attributes when trained on the genomes of many species. To our knowledge, we are the first to show that LMs are able to transfer these attributes to unseen species.

Through analysis of the masked nucleotide reconstructions provided by the models, we have demonstrated that they not only can preferentially reconstruct motifs, but they do so in a way that is sensitive to context. As a result, the reconstruction fidelity can serve as a predictor of whether a particular instance of a consensus motif will be bound in vivo. This suggests that these models could be used to discover candidate high-affinity regulatory elements in species where no binding data is available.

We have further illustrated that the models better reconstruct RBP binding sites if they are located within annotated 3′ UTRs and exhibit improved reconstruction of TATA-box instances if they are placed at a distance to the TSS that is appropriate for the given species. This is remarkable, as we indicated neither the TSS site nor the polyadenylation site during training. This suggests that the models can infer the location of these genomic landmarks. Consequently, LMs may prove useful for accurate genome annotation of understudied species.

Altogether, these analyses further indicate that the reconstruction of masked nucleotides, rather than just a means to an end, can be very informative by itself. We have focused on verifying that the models capture known features of the regulatory code, but similar techniques could potentially reveal novel associations between regulatory elements and their context.

Additionally, we have shown that providing species information to DNA LMs greatly improves their internal numerical representations of regulatory sequences. Strikingly, these representations, when used as input for simple linear models, achieve state-of-the-art predictive accuracy on a wide variety of tasks such as prediction of RNA abundance, condition-specific RNA expression, or mRNA half-life. We note that this approach requires neither retraining the whole model nor engineering a complex nonlinear downstream predictor. Despite its simplicity, we show that this procedure nevertheless recovers biologically meaningful motifs. Thus, LMs do not only learn predictive sequence representations but can also serve as a tool for biological discovery. As species awareness provides significant improvements at practically no cost (one additional token), we think that integrating a species representation will be useful for almost all DNA LMs. Perhaps, a similar strategy could be used to make the model aware of gene families or functions, by providing a token that indicates orthologues or contains some other representation of the gene. This could help the model to learn conserved gene-specific or pathway-specific motifs. However, requiring a gene token would hinder applications on synthetic assays, reporter genes, or genes with few homologs (e.g., many lncRNAs in humans).

This notwithstanding, the language modeling approach does have two important drawbacks. Firstly, it is computationally costly, particularly for longer sequence contexts. We initially experimented with state space models [59, 60], which scaled better to longer sequences. However, we ultimately adopted the standard DNABERT transformer architecture for simplicity, as our main goal was to explore the suitability of language models for multi-species modeling rather than a comprehensive evaluation of model architectures. Scaling language models to mammals with long-range regulatory interactions while maintaining single nucleotide resolution will require architectural improvements [61].

While we showed that LMs generalize across highly diverged species, our analysis focused on a single kingdom. Therfore, it falls short of demonstrating that training on the entire eukaryotic tree of life—from protists to blue whales—could further improve language models. It is conceivable that at a certain point regulatory mechanisms diverge so fundamentally that proper generalization is no longer possible. Additionally, massive differences in genome size may require careful dataset curation. On the other hand, we did not evaluate whether LMs would benefit from large collections of very similar species, such as the recently released 233 primate genomes [6]. Lastly, we also must note that most existing genome collections do not represent a random sample of the true phylogeny but tend to oversample species with medical or biotechnological relevance while undersampling non-western species and those that are hard to isolate [62]. Systematically studying how dataset composition influences what the model learns and how well it performs in different target species is an important avenue for future work.

Materials and methods

Genome data

We obtained 1,500 fungal genomes, comprising 806 different species, from the Ensembl fungi 53 database [63]. For each annotated protein-coding gene in each genome, we extracted 300 base pairs 3′ to the stop codon of genes and 1,000 bases 5′ to the start codon. While the actual transcribed 3′ untranslated regions (3′ UTR) vary in length, we expect that in most cases 300 bp will be sufficient to include the entire 3′ UTR [30]. Equally, in most species, 1 kb should be sufficient to cover the 5′ UTR and promoter—and scaling beyond this length becomes computationally infeasible for the modeling approach taken here. Overall our train set included in the order of 13 million sequences, meaning that the 3′ models are trained on 3.9 billion nucleotides, whereas the 5′ models were trained on 13 billion nucleotides.

As a test set, we used the widely studied species *Saccharomyces cerevisiae*. To prevent data leakage from closely related species, the train set excludes the entire *Saccharomyces* genus.

Sequence alignment

Sequence alignment of annotated *Saccharomyces cerevisiae* CDS and 3′ UTR sequences was performed using discontiguous megablast, which was specifically designed for cross-species alignment, version 2.13 with default parameters [64]. Sequence alignment for *S. cerevisiae* proteins was performed using tBlastn, with default parameters.

Masked language modeling

We performed masked language modeling on the fungi 3′ and 5′ regions. Specifically, we randomly masked nucleotides in each sequence and trained DNABERT models [14] to reconstruct these from context, so as to minimize cross-entropy.

For DNABERT, nucleotides were tokenized into overlapping six-mers before being passed to the model. In this context, we mask spans of overlapping 6-mers so that 80% of the nucleotides selected for masking were masked, 10% percent were randomly mutated, and the remaining 10% were left unchanged. This choice of masking strategy is based on the empirical investigations performed in the original BERT paper for natural language [13]. The model consists of 12 transformer encoder blocks and has around 90 M parameters. We employed Flash-attention [65] as a fast exact attention implementation. Models were trained for 200,000 steps using a batch size of 2048 using the Adam optimizer [66]. The learning rate was warmed up to $4x10^{-4}$ during the first 10,000 steps and then linearly decayed to 0 until training terminates. We increased the masking ratio after 100,000 steps from 15 to 20%. Note that these hyperparameters are the same as those used in the DNABERT paper.

To make the model species aware, the species label corresponding to each region was provided as an additional input token and prepended to the sequence. At the beginning of training the species token embedding was randomly initialized and learned during training. For purposes of comparison, we also trained a species-agnostic version of the model, which does not receive the species label.

As the test species was held out from the training set, the species LM could not be provided with the matching species token. To allow it to predict anyway, we provided the model with a proxy species token of a closely related species. For the 3′ species LM, we used *C. glabrata*, for the 5′ species LM we used *K. africana* for all analyses. An explanation of why these proxies were selected and a detailed investigation of the impact of the choice of proxy can be found in Supplementary Material—Species Token Choice.

We trained ablations of the 3′ models where we varied the dataset composition. The substantial computational cost of training the model prevented us from exploring more 5′ models.

Nucleotide reconstruction

We computed reconstruction predictions for each position in the 3′ and 5′ sequences of the test species *S. cerevisiae* using the respective species and agnostic LMs. For this, we masked each nucleotide individually by masking the span of six overlapping 6-mers that contains this nucleotide. We then averaged the prediction for this nucleotide over the overlapping 6-mers, to obtain one probability distribution per position.

To fit the *k*-mer-models, we tabulated the frequencies of nucleotides conditional on the identity of the (*k*-1)/2 flanking nucleotides, where k is an odd number in {7,…,13} across our training dataset. To reconstruct masked nucleotides, we extracted the (*k*-1)/2 flanking nucleotides on either side of this masked position and then calculated the probability of each nucleotide accordingly.

To reconstruct using alignment, we first downloaded the seven yeast alignment [40] and found the aligned position in the other species for each position in *S. cerevisiae*. We then computed the frequency over the nucleotides in the aligned positions to obtain a distribution. For the far-alignment we used the species *N. castellii* and *L. kluyveri*. For the intra-genus alignment, we used the *Saccharomyces* species.

To evaluate reconstruction for a particular motif, we used regex search to find positions in the 3′ or 5′ sequences matching the motif consensus. We allowed overlapping matches. When 5′ sequences overlap (which can occur, as they are 1 kb long), we do not double count matches, but instead keep only the match in the sequence where it is located closest to the upstream gene.

Many consensus motifs we used in our analyses have degenerate positions. When we computed metrics such as reconstruction accuracy or log-likelihood on these motifs, we only take into account the non-degenerate positions, as otherwise the metrics are inflated.

For all analyses where we use reconstruction fidelity to predict some kind of outcome, including whether the motif is in the 3′ UTR, whether it is bound and whether it is reconstructed better than a shuffled version, we always use the reconstruction log-likelihood (this is the log-likelihood of each nucleotide, averaged over nucleotides in the motif) as the metric. This is mainly because accuracy, while easier to interpret, is also more likely to produce ties.

3′ UTR and TSS annotations

We extracted 3′ UTR annotations for *S. cerevisiae* from Cheng et al. [50], who derived them from Pelechano et al. [30], and matched them to the 3′ sequences. Note that these annotations were only available for 4,388 genes.

To get TSS annotations, we gathered the consensus CAGE clusters for *S. cerevisiae* in YPD from YeastTSS [39]. We matched the locations of these consensus TSS sites with the 5′ sequences. We then matched motif instances to their closest 3′ TSS (or to the closest 5′ TSS site, if there was no 3′ TSS site between the motif and the start codon of the gene).

Binding data

To test whether the reconstruction fidelity is predictive of Puf3 binding, we collected all instances of the Puf3 consensus motifs 3′ (i.e., within 300 bp 3′ of the stop codon) of *S.*

cerevisiae genes. We considered as a positive set all instances of the Puf3 motif in genes with experimental evidence of Puf3p binding their 3′ UTR [41]. We resorted to this as the data does not indicate the exact binding site, just that the 3′ UTR was bound. The negative set comprised the remaining instances. We then computed the reconstruction log-likelihood for all motifs of interest. PhastCons conservation scores [40] for *S. cerevisiae* were downloaded from UCSC and extracted for regions of interest.

For transcription factors, we matched instances of the consensus motif with Chip-exo peaks [42]. Specifically, the data indicates the center of Chip-exo peaks. We extend this by 10 bp to either side and consider that an instance of the motif was bound if it overlapped the extended peak.

Downstream tasks

To evaluate the predictiveness of LM sequence representations, we considered the following tasks. Sun et al. [51] measured half-life for 4,388 *S. cerevisiae* mRNAs using nonperturbing metabolic RNA labeling. Cheng et al. [50] used this data to build a quantitative model to predict mRNA half-life from sequence, using handcrafted features from the coding sequence, 5′ and 3′ UTR—which to our knowledge is the state-of-the-art model of mRNA stability in yeast. Eser et al. [52] measured mRNA half-life in *S. pombe*. In the Shalem et al. [55] MPRA, the expression of a fixed reporter gene was measured when combined with different 3′ sequences. These include genomic 3′ sequences as well as mutated versions of these sequences where mutations were performed in such a way as to tile the sequence. Zrimec et al. [53] aggregated over 20,000 RNA-Seq experiments and built a convolutional neural network to predict the variation in mRNA abundance between genes from sequence. To our knowledge, this represents the state-of-the-art model for predicting endogenous gene expression in *S. cerevisiae*. Zrimec et al. also consider two additional tasks to evaluate the generalization of their model, which we adopt. Firstly, Keren et al. [57] measured, using a fluorescence reporter assay, the expression of a fixed reporter gene when combined with different endogenous *S. cerevisiae* promoter sequences. Yamanishi et al. [56] also used a fluorescence reporter assay to measure the impact of different *S. cerevisiae* terminator sequences. Finally, to have data from non-model species, we obtain microarray data measuring condition-specific gene expression in different stages of growth for a number of yeast species [54].

Generating sequence representations and downstream predictions using linear models

To generate LM sequence representations, we pass each sequence to the language models and extract the sequence representation of the last 4 layers for all tokens, which are then mean pooled to obtain a single embedding per sequence as in Devlin et al. [13] As our models expect fixed length sequences, we truncated longer input sequences. We did this in the following way: for the 3′ sequences, we truncate from the end, and for the 5′ sequences, we truncate from the start. If sequences were shorter than the input, for the 3′ model, we feed them as is. For the 5′ model, we pad them from the left using a fixed sequence. We do this because the 5′ model expects the 1000th sequence position to be immediately 5′ of the start codon.

For predicting half-life, we provided 5′ UTR and coding sequence (CDS) features, as described in Cheng et al. [50] for *S. cerevisiae* and Eser et al. [52] for *S.*

pombe, in addition to our (or competing methods') sequence representations of the $3'$ sequence to a ridge regression with the regularization parameter set by nested cross-validation. We performed 10-fold cross-validation to estimate generalization performance.

We proceeded similarly for predicting reporter expression in the experiment of Shalem et al. [55] but did not include any features besides the $3'$ sequence representation. However, as the reporter contains many almost-duplicate sequences that represent mutations of the same endogenous sequence, we performed grouped 10-fold cross-validation, where the groups consisted of said endogenous sequences. This prevented trivial overfitting.

Zrimec et al. [53] trained different convolutional neural networks using different parts of the regulatory sequence as input. To mimic this, we trained separate linear models for $3'$ and $5'$ as well as the combination thereof to predict gene expression. Here, we do not perform cross-validation but use the same train test split as in Zrimec et al. To predict expression driven by terminator and promoter sequences, Zrimec et al. do not train new models but instead directly apply their convolutional neural network trained on endogenous gene expression. To be comparable, we equally transfer linear model weights to these tasks.

For the microarray, we embed $3'$ and $5'$ sequences for each species and train separate ridge regressions for each species and condition combination. We again use 10-fold cross-validation to evaluate performance.

To assess statistical significance, we computed the residuals and for each task performed a paired Wilcoxon test to determine if the residuals of the species LM were significantly smaller than those of the next best-performing method. The only exception to this is the Zrimec et al. endogenous gene expression task. Here, we only had access to the aggregate performance measures (R^2) of their models. Thus, we used bootstrapping to compute 95% confidence intervals for the performance of the Species LM and determined its performance to be significantly better than Zrimec et al. if the confidence interval did not overlap their result.

Modisco clustering

For de novo motif discovery based on reconstruction probability only, we normalized the reconstruction probability p at each position: $p_{\text{normalized}} = p \cdot \log(p/p_{\text{mean}})$. We then passed this to tfmodisco-lite (https://github.com/jmschrei/tfmodisco-lite) to obtain motifs.

To perform motif discovery on downstream tasks, we calculated embeddings and predictions for all possible single nucleotide polymorphisms of the original input sequences (in silico mutagenesis). Since ridge models were trained using cross-validation, we always selected the model which had the non-mutagenized sequence in its test fold to predict mutation effects for this sequence. Having collected all variant effects, we removed their mean at each position (to also associate an attribution score to the reference nucleotide) and passed this to Modisco.

For every Modisco analysis, we set the sliding window size to 8, the flank size to 3, the target seqlet FDR to 0.05, and the number of Leiden runs to 3.

STREME

To discover motifs in S. cerevisiae 3′ and 5′ regions with STREME, we used the webinterface: https://meme-suite.org/meme/tools/streme. For 5′ regions, we used default parameters. For the 3′ regions, as here we mostly expect stranded RBP motifs, we used the RNA mode which uses only one strand. Moreover, we set the minimum width to five nucleotides, to allow STREME to discover motifs such as Whi3.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-024-03221-x.

Additional file 1. PDF containing supplementary figures and analyses cited in the main text.

Additional file 2. List of the specific datasets analyzed.

Additional file 3. Peer review history.

Acknowledgements
We thank all the scientists who isolated, sequenced, and assembled the genome data which builds the basis for our research. Figure 1D was created with BioRender.com.

Review history
The review history is available as Additional file 3.

Peer review information
Tim Sands was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions
JG and AK conceived the study. JG supervised the research. JH designed and trained the models used in the study. AK and JH designed the evaluations and analyzed the data. DG, MG, and KK explored prototype versions of the model and evaluations. AK, JH, and JG wrote the manuscript. All authors read and approved the final manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL. This study was funded by the German Bundesministerium für Bildung und Forschung (BMBF) through the Model Exchange for Regulatory Genomics project MERGE (031L0174A to JG). The funders had no role in the study design, data collection, analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials
The code is available on https://github.com/gagneurlab/SpeciesLM [67] under a MIT license.
All supporting data is available on Zenodo: https://doi.org/10.5281/zenodo.8247134 [68].
The species and agnostic LMs are available on figshare: https://doi.org/10.6084/m9.figshare.23732655 [69].
The models are additionally available on Huggingface: https://huggingface.co/gagneurlab/SpeciesLM [70].
A list of all datasets analyzed in the study can be found in Additional file 2.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 14 August 2023 Accepted: 20 March 2024
Published online: 02 April 2024

References

1. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.
2. Noguchi S, Arakawa T, Fukuda S, Furuno M, Hasegawa A, Hori F, et al. FANTOM5 CAGE profiles of human and mouse samples. Sci Data. 2017;4:170112.

3. Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B. How many species are there on Earth and in the ocean? PLOS Biol. 2011;9:e1001127.
4. Blaxter M, Archibald JM, Childers AK, Coddington JA, Crandall KA, Di Palma F, et al. Why sequence all eukaryotes? Proc Natl Acad Sci. 2022;119:e2115636118.
5. Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, et al. Towards complete and error-free genome assemblies of all vertebrate species. Nature. 2021;592:737–46.
6. Kuderna LFK, Gao H, Janiak MC, Kuhlwilm M, Orkin JD, Bataillon T, et al. A global catalog of whole-genome diversity from 233 primate species. Science. 2023;380:906–13.
7. Osmanski AB, Paulat NS, Korstian J, Grimshaw JR, Halsey M, Sullivan KAM, et al. Insights into mammalian TE diversity through the curation of 248 genome assemblies. Science. 2023;380:eabn1430.
8. Zhang G, Li C, Li Q, Li B, Larkin DM, Lee C, et al. Comparative genomics reveals insights into avian genome evolution and adaptation. Science. 2014;346:1311–20.
9. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature. 2003;423:241–54.
10. Kimura M. Evolutionary rate at the molecular level. Nature. 1968;217:624–6.
11. Weirauch MT, Hughes TR. Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. Trends Genet. 2010;26:66–74.
12. Hare EE, Peterson BK, Iyer VN, Meier R, Eisen MB. Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. PLOS Genet. 2008;4:e1000106.
13. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv; 2019. Available from: http://arxiv.org/abs/1810.04805. Cited 2023 Jan 18.
14. Ji Y, Zhou Z, Liu H, Davuluri RV. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. Bioinformatics. 2021;37:2112–20.
15. Zhou Z, Ji Y, Li W, Dutta P, Davuluri R, Liu H. DNABERT-2: efficient foundation model and benchmark for multi-species genome. arXiv; 2023. Available from: http://arxiv.org/abs/2306.15006. Cited 2023 Jul 22.
16. Dalla-Torre H, Gonzalez L, Revilla JM, Carranza NL, Grzywaczewski AH, Oteri F, et al. The nucleotide transformer: building and evaluating robust foundation models for human genomics. bioRxiv; 2023. p. 2023.01.11.523679. Available from: https://www.biorxiv.org/content/10.1101/2023.01.11.523679v1. Cited 2023 Jan 19.
17. Fishman V, Kuratov Y, Petrov M, Shmelev A, Shepelin D, Chekanov N, et al. GENA-LM: a family of open-source foundational models for long DNA sequences. bioRxiv; 2023. p. 2023.06.12.544594. Available from: https://www.biorxiv.org/content/10.1101/2023.06.12.544594v1. Cited 2023 Jul 22.
18. Hedges SB, Dudley J, Kumar S. TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics. 2006;22:2971–2.
19. Benegas G, Batra SS, Song YS. DNA language models are powerful zero-shot predictors of genome-wide variant effects. bioRxiv; 2023. p. 2022.08.22.504706. Available from: https://www.biorxiv.org/content/10.1101/2022.08.22.504706v2. Cited 2023 Jul 22.
20. Prieto M, Wedin M. Dating the diversification of the major lineages of Ascomycota (Fungi). PLoS One. 2013;8:e65576.
21. Wilinski D, Buter N, Klocko AD, Lapointe CP, Selker EU, Gasch AP, et al. Recurrent rewiring and emergence of RNA regulatory networks. Proc Natl Acad Sci. 2017;114:E2816–25.
22. Tanay A. Extensive low-affinity transcriptional interactions in the yeast genome. Genome Res. 2006;16:962–72.
23. Ward LD, Bussemaker HJ. Predicting functional transcription factor binding through alignment-free and affinity-based analysis of orthologous promoter sequences. Bioinformatics. 2008;24:i165–71.
24. Wolfertstetter F, Frech K, Herrmann G, Werner T. Identification of functional elements in unaligned nucleic acid sequences by a novel tuple search algorithm. Bioinformatics. 1996;12:71–80.
25. Elemento O, Tavazoie S. Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. Genome Biol. 2005;6:R18.
26. Bussemaker HJ, Li H, Siggia ED. Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. Proc Natl Acad Sci. 2000;97:10096–100.
27. Gordân R, Narlikar L, Hartemink AJ. Finding regulatory DNA motifs using alignment-free evolutionary conservation information. Nucleic Acids Res. 2010;38:e90.
28. Zielezinski A, Vinga S, Almeida J, Karlowski WM. Alignment-free sequence comparison: benefits, applications, and tools. Genome Biol. 2017;18:186.
29. Lu Z, Lin Z. The origin and evolution of a distinct mechanism of transcription initiation in yeasts. Genome Res. 2021;31:51-63.
30. Pelechano V, Wei W, Steinmetz LM. Extensive transcriptional heterogeneity revealed by isoform profiling. Nature. 2013;497:127–31.
31. Sahu B, Hartonen T, Pihlajamaa P, Wei B, Dave K, Zhu F, et al. Sequence determinants of human gene regulatory elements. Nat Genet. 2022;54:283–94.
32. Shrikumar A, Tian K, Avsec Ž, Shcherbina A, Banerjee A, Sharmin M, et al. Technical note on Transcription Factor Motif Discovery from Importance Scores (TF-MoDISco) version 0.5.6.5. arXiv; 2020. Available from: http://arxiv.org/abs/1811.00416. Cited 2022 Sep 25.
33. Bailey TL. STREME: accurate and versatile sequence motif discovery. Bioinformatics. 2021;37:2834–40.
34. de Boer CG, Hughes TR. YeTFaSCo: a database of evaluated yeast transcription factor sequence specificities. Nucleic Acids Res. 2012;40:D169–79.
35. Yang A, Zhu Z, Kapranov P, McKeon F, Church GM, Gingeras TR, et al. Relationships between p63 binding, DNA sequence, transcription activity, and biological function in human cells. Mol Cell. 2006;24:593–602.
36. Rossi MJ, Lai WKM, Pugh BF. Genome-wide determinants of sequence-specific DNA binding of general regulatory factors. Genome Res. 2018;28:497–508.
37. Gordân R, Shen N, Dror I, Zhou T, Horton J, Rohs R, et al. Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. Cell Rep. 2013;3:1093–104.

38. Erb I, van Nimwegen E. Transcription factor binding site positioning in yeast: proximal promoter motifs characterize TATA-less promoters. PLoS ONE. 2011;6:e24279.
39. McMillan J, Lu Z, Rodriguez JS, Ahn T-H, Lin Z. YeasTSS: an integrative web database of yeast transcription start sites. Database. 2019;2019:baz048.
40. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005;15:1034–50.
41. Lapointe CP, Stefely JA, Jochem A, Hutchins PD, Wilson GM, Kwiecien NW, et al. Multi-omics reveal specific targets of the RNA-binding protein Puf3p and its orchestration of mitochondrial biogenesis. Cell Syst. 2018;6:125–135.e6.
42. Rossi MJ, Kuntala PK, Lai WKM, Yamada N, Badjatia N, Mittal C, et al. A high-resolution protein architecture of the budding yeast genome. Nature. 2021;592:309–14.
43. Lieb JD, Liu X, Botstein D, Brown PO. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein–DNA association. Nat Genet. 2001;28:327–34.
44. Tanay A, Regev A, Shamir R. Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast. Proc Natl Acad Sci. 2005;102:7203–8.
45. Hogan GJ, Brown PO, Herschlag D. Evolutionary conservation and diversification of Puf RNA binding proteins and their mRNA targets. PLOS Biol. 2015;13:e1002307.
46. Li B, Oestreich S, de Lange T. Identification of human Rap1: implications for telomere evolution. Cell. 2000;101:471–83.
47. Kramara J, Willcox S, Gunisova S, Kinsky S, Nosek J, Griffith JD, et al. Tay1 protein, a novel telomere binding factor from Yarrowia lipolytica*. J Biol Chem. 2010;285:38078–92.
48. Tsankov AM, Thompson DA, Socha A, Regev A, Rando OJ. The role of nucleosome positioning in the evolution of gene regulation. PLOS Biol. 2010;8:e1000414.
49. Tsankov A, Yanagisawa Y, Rhind N, Regev A, Rando OJ. Evolutionary divergence of intrinsic and trans-regulated nucleosome positioning sequences reveals plastic rules for chromatin organization. Genome Res. 2011;21:1851–62.
50. Cheng J, Maier KC, Avsec Ž, Rus P, Gagneur J. Cis-regulatory elements explain most of the mRNA stability variation across genes in yeast. RNA. 2017;23:1648–59.
51. Sun M, Schwalb B, Pirkl N, Maier KC, Schenk A, Failmezger H, et al. Global analysis of eukaryotic mRNA degradation reveals Xrn1-dependent buffering of transcript levels. Mol Cell. 2013;52:52–62.
52. Eser P, Wachutka L, Maier KC, Demel C, Boroni M, Iyer S, et al. Determinants of RNA metabolism in the Schizosaccharomyces pombe genome. Mol Syst Biol. 2016;12:857.
53. Zrimec J, Börlin CS, Buric F, Muhammad AS, Chen R, Siewers V, et al. Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. Nat Commun. 2020;11:6141.
54. Thompson DA, Roy S, Chan M, Styczynsky MP, Pfiffner J, French C, et al. Evolutionary principles of modular gene regulation in yeasts. Tautz D, editor. Elife. 2013;2:e00603.
55. Shalem O, Sharon E, Lubliner S, Regev I, Lotan-Pompan M, Yakhini Z, et al. Systematic dissection of the sequence determinants of gene 3′ end mediated expression control. PLOS Genet. 2015;11:e1005147.
56. Yamanishi M, Ito Y, Kintaka R, Imamura C, Katahira S, Ikeuchi A, et al. A genome-wide activity assessment of terminator regions in Saccharomyces cerevisiae provides a "terminatome" toolbox. ACS Synth Biol. 2013;2:337–47.
57. Keren L, Zackay O, Lotan-Pompan M, Barenholz U, Dekel E, Sasson V, et al. Promoters maintain their relative activity levels under different growth conditions. Mol Syst Biol. 2013;9:701.
58. Fischer AD, Olivas WM. Multiple Puf proteins regulate the stability of ribosome biogenesis transcripts. RNA Biol. 2018;15:1228–43.
59. Gu A, Johnson I, Goel K, Saab K, Dao T, Rudra A, et al. Combining recurrent, convolutional, and continuous-time models with linear state-space layers. arXiv; 2021. Available from: http://arxiv.org/abs/2110.13985. Cited 2023 Jan 18.
60. Gupta A, Gu A, Berant J. Diagonal state spaces are as effective as structured state spaces. arXiv; 2022. Available from: http://arxiv.org/abs/2203.14343. Cited 2023 Jan 18.
61. Nguyen E, Poli M, Faizi M, Thomas A, Birch-Sykes C, Wornow M, et al. HyenaDNA: long-range genomic sequence modeling at single nucleotide resolution. arXiv; 2023. Available from: http://arxiv.org/abs/2306.15794. Cited 2023 Jul 22.
62. Marks RA, Hotaling S, Frandsen PB, VanBuren R. Representation and participation across 20 years of plant genome sequencing. Nat Plants. 2021;7:1571–8.
63. Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, et al. Ensembl 2022. Nucleic Acids Res. 2022;50:D988–95.
64. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.
65. Dao T, Fu DY, Ermon S, Rudra A, Ré C. FlashAttention: fast and memory-efficient exact attention with IO-awareness. arXiv; 2022. Available from: http://arxiv.org/abs/2205.14135. Cited 2023 Jul 22.
66. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv; 2017. Available from: http://arxiv.org/abs/1412.6980. Cited 2023 Jul 22.
67. Karollus A, Hingerl J, Gankin D, Grosshauser M, Klemon K, Gagneur J. gagneurlab/SpeciesLM. 2023. Available from: https://github.com/gagneurlab/SpeciesLM.
68. Karollus A, Hingerl J, Gankin D, Gagneur J. Supporting data for species-aware DNA language models. Zenodo; 2023. Available from: https://zenodo.org/records/8247134. Cited 2024 Mar 11.
69. Karollus A, Hingerl J, Gagneur J. Species and agnostic LM. figshare; 2023. Available from: https://figshare.com/articles/code/Species_and_Agnostic_LM/23732655/1. Cited 2024 Mar 11.
70. Karollus A, Hingerl J, Gagneur J. gagneurlab/SpeciesLM hugging face. Available from: https://huggingface.co/gagneurlab/SpeciesLM. Cited 2024 Mar 11.

Publisher's Note