Nucleic Acids Research, 2025, **53**, gkaf871
https://doi.org/10.1093/nar/gkaf871
Computational Biology

Investigating the performance of foundation models on human 3′UTR sequences

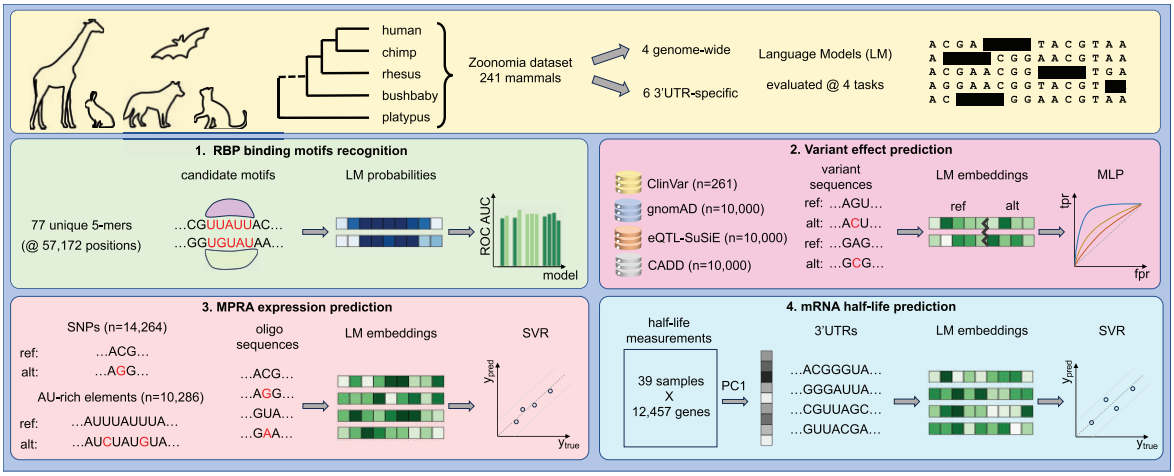
Sergey Vilov[1] **and Matthias Heinig**[1,2,3,*]

[1]Institute of Computational Biology, Computational Health Center, Helmholtz Zentrum München Deutsches Forschungszentrum für Gesundheit und Umwelt (GmbH), Neuherberg 85764, Germany
[2]Department of Computer Science, TUM School of Computation, Information and Technology, Technical University Munich, Garching 85748, Germany
[3]German Centre for Cardiovascular Research (DZHK), Partner Site Munich Heart Alliance, Berlin, 10785, Germany
[*]To whom correspondence should be addressed. Email: matthias.heinig@helmholtz-munich.de

Abstract

Foundation models, such as DNABERT and Nucleotide Transformer, have recently shaped a new direction in DNA research. Trained in an unsupervised manner on a vast quantity of genomic data, they can be used for a variety of downstream tasks, such as promoter prediction, DNA methylation prediction, gene network prediction, or functional variant prioritization. However, these models are often trained and evaluated on entire genomes, neglecting genome partitioning into different functional regions. In our study, we investigate the efficacy of various unsupervised approaches, including genome-wide and 3′ untranslated region (3′UTR)-specific foundation models on human 3′UTR regions. To this end, we train a set of popular transformer architectures on a 3′UTR-specific dataset comprising 3 783 714 3′UTR sequences (6.6B bp) of 241 Zoonomia species. Our evaluation includes downstream tasks specific for RNA biology, such as recognition of binding motifs of RNA-binding proteins, detection of functional genetic variants, prediction of expression levels in massively parallel reporter assays, and estimation of messenger RNA half-life. Remarkably, models specifically trained on 3′UTR sequences demonstrate superior performance when compared to established genome-wide foundation models in three out of four downstream tasks. Our results underscore the importance of considering genome partitioning into distinct functional regions when training and evaluating foundation models. In addition, the proposed set of 3′UTR-specific tasks can be used for benchmarking of future models.

Graphical abstract



Introduction

Foundation models have revolutionized various fields by harnessing self-supervised pre-training on vast datasets, enabling these models to capture intricate patterns and representations without requiring labeled data. After pre-training, these models can be utilized for specific downstream tasks either directly through zero-shot inference or by performing probing or fine-tuning on task-specific labeled datasets. This paradigm, initially established in natural language processing (NLP) with models like BERT [1] and GPT [2], has been successfully transferred to genomics, allowing the development of models with the ambitious goal of understanding the "DNA language."

Received: August 16, 2024. **Revised:** July 17, 2025. **Accepted:** August 1, 2025

In genomics, several large language models (LMs) for DNA sequences have emerged [3–5], leveraging transformer-based architectures to process and interpret genomic data. Among these, DNABERT [3], inspired by the success of BERT in NLP, focuses on capturing global contextual information across DNA sequences by learning to predict spans of randomly masked nucleotides in the sequence. However, trained solely on the human genome, DNABERT is not able to utilize evolutionary information, which has long been leveraged by alignment-based conservation models, such as PhyloP [6] or PhastCons [7]. Subsequent advancements, such as DNABERT2 [4], addressed this limitation by incorporating data from 135 species, enabling the model to capture evolutionary signals. Another major development in the field is the Nucleotide Transformer (NT) family [5], which includes multispecies models trained on sequences from a broad array of 850 species, spanning both vertebrates and non-vertebrates. State space attention models have been suggested as a more memory-efficient alternative [8]. All of these LMs have demonstrated promising results in such tasks as promoter prediction, enhancer detection, and prioritization of functional variants.

However, training genomic LMs may overlook the partitioning of the genome into distinct functional units, such as UTRs or coding regions. Training sequences are usually selected by chunking the genome without considering regional boundaries, and no explicit information about the functional context of these sequences is given to the model. Consequently, it can be hard for the model to infer region-specific properties of the sequences. While models trained on the whole genome should in principle be also applicable to region-specific tasks, this constraint could hinder their ability to capture essential dependencies and functional relationships inherent to distinct genomic regions.

The 3′ untranslated region (3′UTR) is a functionally crucial segment of the genome, playing a vital role in post-transcriptional regulation. It serves as a hub for interactions with RNA-binding proteins (RBPs) and microRNAs, influences messenger RNA (mRNA) stability, and regulates translation efficiency. The unique functional landscape of 3′UTRs raises the question of whether models trained specifically on these regions can better capture the functional dependencies required for 3′UTR-specific tasks.

We hypothesize that foundation models trained exclusively on 3′UTRs might outperform genome-wide models on these tasks by directly specializing on the unique features of this genomic region. To evaluate this hypothesis, we define a reusable benchmark of four biologically and clinically relevant tasks: recognition of RBP binding motifs, prioritization of functional variants, prediction of expression levels in massively parallel reporter assays (MPRA), and estimation of mRNA half-life.

We compare several 3′UTR-specific LMs against existing genome-wide models, including the DNABERT (89M parameters), DNABERT2 (117M parameters), NT (NT-MS-v2-100M, 98M parameters), and State Space (1M parameters) architectures, assessing their performance across the proposed tasks. Although the DNABERT and the multispecies NT models were originally designed with distinct goals—the former for human-specific tasks and the latter for general multispecies applications—comparing their performance on region-specific tasks like 3′UTRs allows us to assess the trade-offs between specialization and generalization in genomic modeling. In the context of region-specific modeling, it is also particularly compelling to investigate the potential of the lightweight State Space architecture, which has previously been applied to modeling of fungi genomes. We also emphasize that none of the considered models relies on specific biological knowledge or any other constraints other than the input sequences that would limit their application to the initial use case. By applying all these models to the proposed tasks, we aim to highlight the importance of region-specific training of genomic foundation models and establish a set of benchmarks for future studies focusing on 3′UTR biology.

Materials and methods

Multispecies data preparation

For multispecies training of LMs, we considered the 241 mammalian genomes of the Zoonomia project [9]. The Zoonomia dataset is particularly attractive, as it is supplied with a Phylop-241way model that provides an evolutionary-based conservation score computed on whole-genome alignment [9], derived using one of the most recent aligners named Progressive Cactus [10].

The 3′UTR plays a vital role in post-transcriptional regulation, and its proper annotation is crucial for training models that are aimed to capture region-specific features. However, due to the lack of comprehensive 3′UTR annotations across species, we developed an approximation method using highly conserved coding sequences as anchors.

Our 3′UTR-specific models are trained on RNA data that is prepared as follows. Sequence annotations for protein-coding human genes are extracted using BioMart (https://www.ensembl.org/biomart/martview/), using GRCh38/hg38 as a reference genome. To facilitate comparison between DNA- and RNA-based models, we only consider single-exon UTRs, resulting in 18 134 3′UTR sequences in total. For each gene, we consider the 3′UTR annotation of the Ensembl canonical transcript. Obtaining 3′UTR coordinates for non-human species is more challenging since our search in the public domain revealed a lack of 3′UTR annotations for most of the Zoonomia species. For example, Ensembl release 112 provides 3′UTR coordinates for only 57 species out of 241. One way to overcome this problem is to use the Zoonomia whole-genome Progressive Cactus alignment [9]. We then assume that the stop codons of protein-coding genes are well aligned due to the adjacent highly conserved coding sequences. The 5′end of a 3′UTR is located directly downstream of the stop codon. Hence, by knowing the 5′ end positions of 3′UTRs in the human genome, one can use the Zoonomia alignment to infer the corresponding positions in all other species. However, the positions of the 3′ends can not be determined equally easily without full length cDNA or RNA-seq data. As an estimation of 3′UTR coordinates in non-human genomes, we simply considered sequence segments downstream of the stop codon, with the length corresponding to the 3′UTR length of the respective human transcript (Supplementary Fig. S1). Most of the Zoonomia species are assembled at the scaffold level. When extracting the 3′UTR sequences, we assumed that the 3′UTR of a given transcript and its stop codon are located on the same scaffold. When evaluated on the 3′UTR annotations available in Ensembl release 112, our approach detects the position of 5′end of 3′UTRs with an accuracy of 93%, the median Jaccard index between our 3′UTR annotations and those from Ensembl being 70%. For both 5′ and 3′ ends, the

difference between the Ensembl 3′UTR coordinates and those detected by our approach has a median of 0 nt. The interquartile range is 0 and 648 nt for the 5′ and 3′ ends, correspondingly. While this approach does not capture all species-specific differences in 3′UTR length, it provides a consistent framework for tasks that primarily evaluate human-centric validation datasets. To train the RNA-based models, sequences mapped to the negative (reverse) strand of the human DNA, including human transcripts on the negative strand, were reverse complemented to match the human RNA orientation. In total, our Zoonomia-based 3′UTR-specific dataset comprised 3 783 714 3′UTR sequences with a total length of around 6.6B bp.

To train the whole-genome DNABERT2-ZOO model, we prepared a whole-genome dataset based on the Zoonomia sequences. To this end, we extracted all contigs for all Zoonomia species from the Zoonomia whole-genome alignment. We then split long sequences into chunks of 100 000 nt with an overlap of 50 nt and shuffled. To extract sequences from the Zoonomia alignment .hal file, we used HAL format API v.2.3.

Models

To evaluate whether region-specific training improves downstream task performance, we selected a set of representative models. This includes genome-wide models to serve as baselines and 3′UTR-specific models specifically trained to capture localized context. The selected models span a diverse set of architectural and training approaches. In particular, DNABERT was designed to assess performance on human-specific sequences, while the DNABERT2 and NT models offer insights into the benefits of incorporating multispecies data. Finally, the State Space models offer a linearly scalable implementation of the attention mechanism and might be more suitable for region-specific training due to the lower number of parameters. In total, 10 alignment-free and 4 alignment-based models were evaluated on human 3′UTR sequences, including 6 alignment-free models specifically trained on 3′UTRs. Among these is the DNABERT2-ZOO model, which is a modification of DNABERT2 that we trained on the whole-genome Zoonomia dataset to assess the impact of species selection on the performance of multispecies models.

Genome-wide transformer models:

DNABERT. DNABERT is the first BERT-like model for DNA analysis. Specifically, we used the version referred to as DNABERT-6, as it demonstrated the best performance in the original study [3]. This model employs an encoding scheme that splits the input sequence into overlapping 6-mer tokens and can accommodate sequences of up to 512 nt in length.

DNABERT2. This is a BERT-like model trained on whole-genome data from 135 species. The model uses byte pair encoding (BPE), which splits the input sequence into tokens of variable length. Sequences up to 1024 tokens in length (≈52 00 nt; the exact length in nt depends on the sequence content) can be used.

DNABERT2-ZOO. We retrained the original DNABERT2 model based on the whole Zoonomia dataset [9], which we describe in detail in the previous section.

NT-MS-v2-100M. The 100M-parameter multispecies NT model. Although the 2.5B multispecies NT model outperformed the other NT models across the downstream tasks [5], its training on the available GPU infrastructure would be challenging due to a substantial number of learning parameters. We therefore opted for the v2-100M version, which is faster to train. The model employs an encoding scheme that splits the input sequence into non-overlapping 6-mer tokens and can accommodate sequences of up to 1024 tokens (6144 nt) in length.

3′UTR-specific transformer models:

DNBT-3UTR-RNA, *DNBT2-3UTR-RNA*, and *NT-3UTR-RNA*. We retrained the previously proposed DNABERT, DNABERT2, and NT-MS-v2-100M models solely on 3′UTR sequences from the Zoonomia dataset [9]. We make the models RNA-specific by reversing the sequences on the negative strand while training, as described in the "Multispecies data preparation" section.

STSP-3UTR-RNA, STSP-3UTR-DNA, and STSP-3UTR-RNA-HS. Using the State Space model architecture previously used for language modeling on 3′UTR sequences in fungi [8], we train a 3′UTR-specific RNA-based multispecies model, a 3′UTR-specific DNA-based multispecies model, and a 3′UTR-specific RNA-based human-only model to assess the effect of including multiple species. These models employ an encoding scheme that treats each nucleotide as a single token and can accommodate sequences of up to 5000 nt in length.

The 3′UTR-specific LMs were trained on 2 NVIDIA A100 80G GPUs in parallel until the relative loss change for the last three epochs dropped below 1%. The DNABERT2-ZOO model was trained for ∼2 epochs on 10 GPUs in parallel. This corresponds to 270B tokens, which is comparable to the training set size of similar genome-wide LMs. For example, the NT-MS-v2-100M model was trained on 300B tokens. To train all models we used the AdamW optimizer [11]. To accelerate the training procedure, we used the mixed precision training and gradient accumulation techniques. The training parameters (Supplementary Table S1) were set as close as possible to the original model publication.

Many 3′UTRs exceed the maximum input length of transformer models. Chunking these regions while preserving overlapping context ensures that the models retain sequence dependencies and make accurate predictions. For each trained model, large sequences were split into chunks to match the model's field of view. We implemented a 50 nt overlap at both ends of each chunk to provide the model with context from the adjacent segments while minimizing the resulting number of redundant training tokens.

All the models are trained in a BERT-like fashion, by masking a random portion of the input sequence. For each input sequence, 15% of the tokens were randomly chosen for masking. Of these tokens, 80% were replaced with a MASK token, 10% were randomly mutated, and the remaining 10% were left unchanged. The selection of 15% of tokens is somewhat arbitrary but has gained popularity for training LMs since it was first introduced in the original BERT paper [1], where it was chosen with the rationale that the model struggles to learn effective representations when too much text is masked. While the random part helps prevent overfitting, the purpose of the unchanged part is to teach the model to rely not only on the neighboring positions when predicting a masked token k but also on the token k itself. This makes the training task closer to the inference task since no masking is used when generating embeddings. Since a large fraction of sequences in the Zoonomia dataset contained long N-chunks, we added an extra "NNNNNN" token when training the DNABERT2 and NT architectures.

All the models were trained using cross-entropy loss computed only on the masked positions in the sequence.

Alignment-based conservation models:

PhyloP-100way. This conservation model is based on whole genome alignment of 100 vertebrates [6].

PhyloP-241way. This conservation model is based on whole genome alignment of 241 mammalian genomes from the Zoonomia project [9].

CADD-1.7. Combined Annotation-Dependent Depletion model v. 1.7 [12]. This model is a logistic regression-based classifier that directly predicts deleteriousness or functionness of a genetic variant in the human genome by combining a wide range of genomic annotations derived from various sources, including conservation across species, gene and transcript models, protein LMs, etc. We categorize CADD-1.7 as an alignment-based model due to its reliance on PhyloP conservation scores.

Zoo-AL. As an alignment-based Zoonomia-specific allele-sensitive baseline model, we compute allele probabilities from the nucleotide frequencies across all columns of the 3′UTR Zoonomia alignments.

Predictions using zero-shot scores and probing

LMs are pre-trained in a self-supervised manner. They can then be utilized to make predictions for specific downstream tasks using zero-shot scores or through probing or fine-tuning. Zero-shot scores refer to the machine learning paradigm where a pre-trained model is evaluated on classes that were not used at training. They leverage the pre-trained model's understanding of sequence context without altering its parameters. These scores are typically derived directly from token predictions or from similarity metrics on embedding vectors. Similar to [5], we define probing as a prediction technique consisting of using embeddings generated by a LM as input features to a simpler task-specific model trained in a supervised manner. In this case, the parameters of the LM itself do not need to be adjusted. In contrast, fine-tuning involves updating the weights of all or only some layers of a pre-trained LM. Additionally, extra layers are usually added to the pre-trained LM to enable supervised training and prediction. In this study, we explore zero-shot scores and probing, as it is computationally much less expensive than fine-tuning, especially given the necessity of cross-validation (CV) due to the relatively small sizes of the task-specific datasets. For probing, we apply Ridge regression, multilayer perceptron (MLP) models and support vector regression (SVR) models, depending on the extent to which task-specific information is linearly encoded in the embeddings.

Recognition of RBP binding motifs: labeled data

As the first downstream task, we assess the ability of the models to recognize RBP binding motifs. The ground truth set of proxy-functional RBPs motifs was built based on the consensus of two experiments [13] that utilized distinct methodologies: RNA Bind-n-Seq (RBNS) and enhanced crosslinking and immunoprecipitation (eCLIP). The purpose of combining the eCLIP and RBNS datasets is to compensate for the possible mislabeling in eCLIP data by restricting the provided eCLIP peaks to the motifs associated with strong functional evidence in the RBNS experiment.

Specifically, the proxy-functional set was composed of motif positions resulting from the overlap of high-confidence eCLIP peaks with 5-mer binding sites detected in the RBNS experiment for the same RBPs. We define high-confidence eCLIP peaks as the peaks resulting from the irreproducible discovery rate analysis, which estimates positions of reproducible peaks based on experiments with two biological replicas [13]. Following the approach of [13], we additionally extend each such peak 50 nucleotides upstream of its 5′ end, as some RBP motifs can be found symmetrically around or only upstream of the peak start. Filtering for 3′UTR-specific sites resulted in 57 172 motif hits in the proxy-functional set, spanning 77 unique 5-mers for 20 RBPs. To construct the proxy-non-functional set, we considered 1 653 964 motif hits for the same 77 unique 5-mers not overlapping with any eCLIP peak for any biological replicate [13]. From these, we randomly sampled 57 172 motif positions to simplify calculations, ensuring the same number of instances and the same distribution of 5-mers as in the proxy-functional set. We assume that LM training should be robust to the intrinsic class imbalance, as functional and non-functional motifs are likely to appear in different sequencing contexts. Context-sensitive LMs should be able to recognize these distinctions, reducing the chance to confuse the two classes. Since our evaluation dataset undersamples the number of non-functional motifs, we chose receiver operating characteristic area under the curve (ROC AUC) as a metric robust to class imbalance, as described in the next section.

To compare model performance on motifs with different conservation, we examine two categories, with the proxy-functional motifs falling into the top 10% conservation (highly conserved motifs) and the bottom 10% conservation (weakly conserved motifs), as measured by PhyloP-241way. A given 5-mer can be associated with multiple RBPs. Assuming that the associated RBP defines the sequence context for a given 5-mer, we independently selected the top 10% of weakly conserved and 10% of highly conserved motifs for each group, defined by the tuple (RBP, 5-mer). In total, there are 102 such groups, each representing all positions of a specific 5-mer experimentally linked to a particular RBP within the eCLIP peaks. To create the set of proxy-functional motifs for the weakly (highly) conserved categories, we combined the 10% most weakly (highly) conserved motif positions from all (RBP, 5-mer) groups. We keep the same set of proxy-non-functional motifs for the highly conserved and weakly conserved categories.

Recognition of RBP binding motifs: predictions

To perform zero-shot predictions of binding sites, we computed for all proxy-functional and proxy-non-functional motifs an aggregate score Φ. For the alignment-based models, whose output can not be interpreted in terms of probability, we used the maximum of the model output across the motif positions: $\Phi = max(score(i)),\ i = 1..5$. For the CADD model, which provides scores for all possible single nucleotide variants (SNVs), the maximum SNV score across all motif positions was used as Φ. For the Zoo-AL model and LMs, which directly provide allele probability at the output, we used the average reference allele probability across the motif positions : $\Phi = 1/5 p_{ref}(i),\ i = 1..5$.

To derive per-nucleotide probabilities from LM predictions, we ran LM inference for all 18 134 3′UTR sequences in the human genome, masking the input tokens as described in the official model documentation or in the previous studies. In particular, for the DNABERT models, which use overlapping

tokenization, we employed the inference technique applied by [8]. This technique consists in masking a contiguous span of six 6-mers for each probed position and using the output for the fourth nucleotide of the third masked 6-mer as prediction. For the State Space models, the inference was performed by masking the nucleotide positions in the rolling masking fashion with stride = 50, as implemented in [8]. Finally, the predictions for the NT models were derived by consecutively masking each 6-mer. To predict the probability of a given nucleotide A, C, G, or T, we summed up all the model probabilities for all the tokens having a given nucleotide at a given position. For the DNABERT2-based models, Φ could not be computed, since the variable token length in the BPE encoding scheme impedes separation of contributions from individual positions along the sequence.

We consider a higher Φ as an indicator of a functional motif. Assuming that a more functional element is more conserved, it can be expected to occur repetitively within similar sequence contexts in the dataset. Such elements should then be associated with higher p_{ref} and higher Φ.

For each model, we then computed the ROC AUC metric, considering Φ as prediction and the motif proxy-functional status as the label. Note that the ROC AUC score is insensitive to class imbalance, since it is unaffected by multiplying the ratio between ground truth proxy-functional and proxy-non-functional labels by an arbitrary factor [14] and corresponds to the probability that a randomly chosen proxy-functional motif is ranked higher than a randomly chosen proxy-non-functional motif. It is, therefore, an appropriate metric for comparison between models. The resulting ROC AUC scores and their bootstrap-estimated 95% confidence intervals are reported in Supplementary Table S2. We considered two models equivalent when their 95% confidence intervals for ROC AUC overlap.

Prioritization of functional variants: labeled data

The second task is to predict the functional impact of genetic variants. We used four different sources to construct four sets of proxy-functional variants. As putative functional variants, we first considered SNVs with (likely) pathogenic annotations from ClinVar v. 2023.10.07 [15], dropping variants with *no_assertion* or *no_interpretation* annotations. However, as a high conservation score from PhyloP often serves as a criterion to include non-coding variants into ClinVar, the estimation of generalization performance of the PhyloP-100way model based on ClinVar variants is compromised. Hence, we also considered rare SNVs with allelic count of 1 from gnomAD v. 3.1.2 [16] as an alternative set of proxy-functional variants. As the third set, we utilized SNVs associated with gene expression, known as expression quantitative trait loci (eQTLs), extracted from the eQTL-SuSiE fine-mapping credible sets [17] with a stringent P-value threshold of $<10^{-25}$. Finally, we selected 3′UTR-specific "proxy-deleterious" variants used for CADD-1.7 training [12]. This is a set of simulated variants that do not naturally occur in primates and might thus be enriched for functional mutations. In total, we retained 261 ClinVar, 3 210 324 gnomAD, 11 301 eQTL, and 151 287 CADD SNVs.

As a matched set for proxy-functional variants from ClinVar, gnomAD, and eQTL data, we built a set of proxy-non-functional variants using 10 000 randomly chosen SNVs with gnomAD population allele frequency above 5%. These SNVs were selected to ensure no overlap with any proxy-functional variant. To ensure that the SNVs from the proxy-non-functional set do not exhibit eQTL effects, we also disregarded all gnomAD variants that overlapped eQTLs with P-value $< 10^{-12}$. As a matched set for the "proxy-deleterious" CADD variants, we selected the corresponding number of 3′UTR-specific "proxy-neutral" variants, also used in CADD training. These are variants that naturally occur in primates and are thus more likely to be benign. To reduce the computational burden, we used at most 10 000 randomly sampled proxy-functional or proxy-non-functional variants in each dataset.

Prioritization of functional variants: predictions

Predicting functional significance of genetic variants is a key step to understanding disease mechanisms. By leveraging embeddings and token probabilities, we aim to capture signals that distinguish functional and non-functional variants. To this end, we first calculated a set of *zero-shot* scores that could be used as an indicator of a variant's functional significance. Three of these scores were computed based on model predictions at the variant position: the logarithm of the reference allele probability ($\log(p_{\text{ref}})$), the logarithm of inverse alternative allele probability ($\log(p_{\text{alt}}^{-1})$), and the logarithmic ratio of these two probabilities ($\log(p_{\text{ref}}/p_{\text{alt}})$). The last two scores quantify the level of "surprise" the model experiences when encountering an alternative allele at the variant position. For more pathogenic variants, the alternative allele is expected to occur more rarely, leading to a higher score. To compute these scores, we used the same per-nucleotide probabilities that we derived for computing the motif scores in the first task.

For each variant we also computed the variant influence score (VIS), introduced by [18]. To do this, we first extracted a $W = 4096$ nt-long chunk ($W = 512$ nt for DNABERT-based models) of the reference sequence *refseq* centered at the variant's position. After that, the LM was employed to predict probabilities of all possible SNVs in this sequence. We then generated LM predictions for the alternative sequence *altseq* obtained by mutating the central (variant) nucleotide in the reference chunk to the alternative allele. In agreement with [18], the VIS for a variant at position i was computed as follows:

$$\text{VIS}_i = \frac{1}{W}\sum_{j\neq i}^{W} \max\left\{\left|\log_2\left(\frac{\text{odds}(n_j = k|altseq)}{\text{odds}(n_j = k|refseq)}\right)\right|\right\}_{k\in A,C,G,T}.$$

In simple terms, VIS quantifies to which extent the other loci in the sequence depend on the given variant. Following [18], when computing VIS, we retained masking only for the DNABERT-based models, since it was required to avoid a data leak due to the overlapping tokenization scheme used by the model. For the NT models, we used the same method for extracting nucleotide probabilities as in [18]. To mitigate the computational burden, we used at most 3000 variants per inference set and per label for VIS computation. For the 3′UTR-specific models, we did not include correlations with positions beyond the 3′UTR border.

For a given input sequence, a pre-trained LM can be used to extract a concise representation, called embedding. Technically, an embedding is usually the output of some intermediate layer of the model. Embeddings encode a compressed representation of the input sequence given the context learned

in training. Comparing embeddings for reference and alternative alleles provides a quantitative measure of how the model perceives changes in sequence functionness. While for non-functional variants, an alternative allele is likely to appear as often as the reference allele in a given context, for functional variants, it can be unexpected to appear in the same context. Therefore, for non-functional variants, the embeddings corresponding to the reference and alternative sequences can be expected to be more similar to each other than for functional variants. With the RNA-based models, we used RNA instead of DNA sequences, obtained by taking a reverse complement of the DNA sequence for the genes located on the negative strand.

For each tested LM, we extracted embeddings from the final layer, following the official model usage instructions or in the agreement with the original model publication. Based on these embeddings, we computed four additional zero-shot functionality scores using four similarity measures between embeddings for reference and alternative sequences (l1, l2, dot product, and cosine similarity). Two more scores were computed based on the loss on the alternative sequence and the difference between the alternative and reference losses.

In addition to the zero-shot scores, we also applied probing by using extracted embeddings to train supervised classifiers to predict functional variants. The embeddings were generated separately for the reference and alternative alleles by using a $W = 4096$ nt-long sequence window ($W = 512$ nt for DNABERT-based models) centered around each variant. The reference and alternative embeddings were then concatenated, resulting in a 2× longer vector used as an input to an MLP model. For the 3′UTR-specific models, we cut the window at the 3′UTR border.

The MLP was evaluated in a nested CV fashion, using the inner 5-fold CV for the hyperparameter search and the outer 10-fold CV for the estimation of model generalization performance. The MLP was trained for 300 epochs with learning rate of 1e-4 and batch size of 1024 using the AdamW optimizer [11]. The hyperparameter search for the number of layers, dropout probability, and weight decay was performed by running a Tree-structured Parzen estimator solver [19] 150 times. To reduce the computational burden, we retained the hyperparameters determined for the initial (first) outer CV split. The MLP was trained separately for each LM and each variant dataset, resulting in a total of 24 models.

For each functionality score and each variant dataset, we computed the ROC AUC measure, considering the functionality score as the predictor and the motif functional status as label. The zero-shot and MLP-based ROC AUC scores are reported alongside with their bootstrap-estimated 95% confidence intervals in Supplementary Tables S3 and S4, correspondingly. We considered two models equivalent when their 95% confidence intervals for ROC AUC overlap.

MPRA expression levels and mRNA half-life: labeled data

A model's ability to predict gene expression measured in MPRA experiments or mRNA half-life characterizes its capacity to capture the functional relevance of 3′UTRs, making MPRA and mRNA half-life prediction tasks crucial benchmarks for evaluation. Thus, our third and fourth tasks consist in predicting MPRA expression levels measured by [20] and [21], and to estimate mRNA half-life reported by [22].

We considered the mRNA steady-state measurements in six cell lines as prediction targets for the [20] data. In particular, we used the $\log_2$FoldChange_Ref_{cell_line} and $\log_2$FoldChange_Alt_{cell_line} values as prediction targets for the reference and alternative allele of each variant correspondingly. The length of oligo sequences used to generate embeddings was 101 nt.

For the [21] data, the prediction targets were the mRNA steady-state and mRNA stability measurements obtained in the original experiment for two distinct cell lines. In particular, we used the ratios_T0_GC_resid and ratios_T4T0_GC_resid values as measures of steady state and stability, correspondingly, as defined in the original study. The length of oligo sequences used to generate embeddings was 160 nt.

The sequence-level data in the original studies were reported in GRCh37/hg19 coordinates. To enable our analysis, we performed a liftover of variant and oligo coordinates to GRCh38/hg38, removing all oligos for which liftover failed and those who were not fully included in the hg38 UTRs.

In total, we generated embeddings for 14 264 oligo sequences from the [20] experiment, 10 286 oligo sequences from the [21] experiment on average per prediction target, and for 12 457 UTR sequences corresponding to genes for which mRNA half-life was reported by [22].

MPRA expression levels and mRNA half-life: predictions

The predictions were generated by probing of LM embeddings. As downstream regressors, we used Ridge regression and SVR models. As a baseline for each dataset, we also implemented feature encoding proposed in the corresponding original study (see the "Reimplementing models from previous studies" section).

When computing embeddings using the 3′UTR-specific models trained on RNA sequences, we used RNA sequences obtained by taking the reverse complement of the DNA sequence for the genes located on the negative strand. To generate embeddings for 3′UTRs longer than the LM field of view, we cut the sequences at the 3′ end, assuming that the relevant functional elements are located closer to the 5′ end of the 3′UTR.

All Ridge regression models were evaluated in a nested CV fashion, using the inner five-fold group-based CV for the hyperparameter search and the outer LeaveOneGroupOut CV for estimation of the generalization performance. The grouping criteria in the outer loop followed the conventions of the original study: groups for the MPRA data from [20] were composed of oligos derived from the same gene, groups for the MPRA data from [21] data matched chromosomes, and groups for the mRNA half-life data from [22] aligned with the folds defined in the original study.

In the MPRA task, expression levels were predicted independently for oligo sequences carrying the reference and alternative alleles, following the methodology in the original studies.

To fully leverage the potential of LM embeddings, we additionally trained SVR models on the same data. These models were evaluated using the same validation strategy as the Ridge regression models, following a nested CV approach. This involved an inner five-fold group-based CV for hyperparameter optimization and an outer LeaveOneGroupOut CV to estimate the model's generalization performance. To reduce the

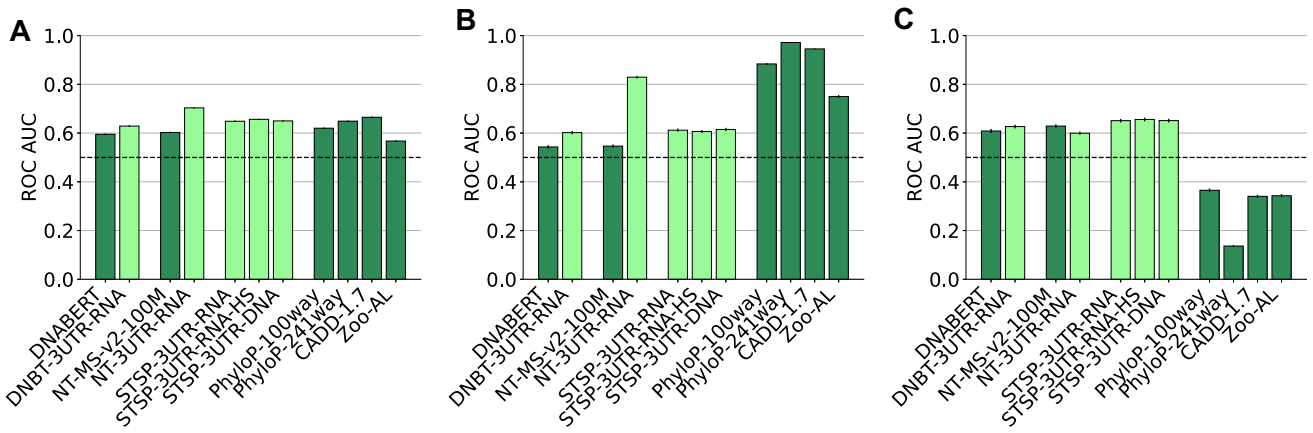


Figure 1. ROC AUC scores for RBP binding motif predictions for all motifs (**A**), proxy-functional motifs within the top 10% conservation (**B**), and proxy-functional motifs within the bottom 10% conservation (**C**), as assessed by PhyloP-241way. The error bars show the 95% confidence intervals. The color indicates genome-wide (dark green) or 3′UTR-specific models (light green).

computational burden, we retained the hyperparameters determined for the initial (first) outer CV split. The SVR hyperparameters (regularization strength C, RBF kernel coefficient γ, and epsilon-tube width ε) were searched on a logarithmic scale by running a Tree-structured Parzen estimator solver (Bergstra *et al*., 2011) 300 times.

The prediction quality was measured with the Pearson correlation coefficient r, computed between the predictions gathered from all test folds and the values measured in the experiment. We consider two models equivalent when their 95% confidence intervals for Pearson r overlapped. The 95% confidence interval was computed using Fisher transformation.

The alignment-based conservation models were excluded from the evaluation since they do not provide a way to generate sequence embeddings.

Reimplementing models from previous studies

Baseline models from previous studies provide a benchmark for evaluating the added value of LM embeddings. By reimplementing these models, we ensure a fair comparison between traditional feature-based methods and embedding-based approaches. Therefore, for the third and the fourth tasks, we trained additional regression models using sequence-based features proposed for the corresponding dataset in the original study.

In particular, for the MPRA data from [20], the input feature vector ($n = 58$) for each oligo sequence was constructed based on the following features: nucleotide percentage (across four bases, 4 features), dinucleotide percentages (16 features), exact dinucleotide counts (16 features), maximum homopolymer length (for each nucleotide A, T, C, G, and across all nucleotide types, five features), maximum dinucleotide length across all bases (16 features), and a measure of sequence uniformity (computed as follows: for i in range(1,len (seq)): if seq[i] == seq[i-1]: seq_uniformity = seq_uniformity + 1).

For the MPRA data from [21], the input feature vector ($n = 1024$) for each oligo sequence was constructed by counting the number of occurrences of each possible 5-mer in the input sequence since these features led to the best predictions in the original study.

For mRNA half-life prediction, the input feature vector ($n = 21\,844$) for each human 3′UTR sequence was constructed by counting all k-mers ($k = 1..7$) in the sequence.

Similarly to LM embeddings, the generated feature vectors were used as an input to Ridge and SVR models.

Results

Recognition of RBP binding motifs

By serving as binding regions for RBPs, RBP binding sites play a critical role in regulating gene expression through post-transcriptional processes such as mRNA stability, splicing, and translation. Accurate prediction of these binding sites is essential for unraveling complex biological networks and understanding disease mechanisms, especially in disorders associated with RNA dysregulation.

We therefore assessed the ability of each model to detect binding sites for 20 RBPs, represented by 77 unique 5-mers. This task might be particularly challenging for alignment-based models since regulatory elements might exhibit certain mobility [8]. In the context of multispecies alignment, the exact position of a given regulatory element may vary across different genomes and not necessarily match its position in the human sequence. The resulting model performance, assessed as ROC AUC, is illustrated on Fig. 1 and in Supplementary Table S2 for the whole dataset as well as for highly conserved (top 10%) and weakly conserved (bottom 10%) proxy-functional motifs separately.

The best performance across all motifs is delivered by the NT-3UTR-RNA model (AUC = 0.703). All 3′UTR-specific LMs outperform their genome-wide counterparts. To gain deeper insight into the mechanisms of alignment-free and alignment-based models, we assessed their ability to detect highly conserved and weakly conserved proxy-functional motifs separately (Fig. 1B and C). In the high conservation group, all LMs were outperformed by the alignment-based PhyloP and CADD models. In this group of motifs the alignment based models exploit the fact that the median PhyloP-241way conservation for the proxy-functional motifs (score $\approx$ 6.6) is about six times greater than for the proxy-non-functional (score $\approx$ 1.1). Among the LMs, NT-3UTR-RNA performs the best with AUC = 0.829. On the other hand, the State Space and DNABERT-based LMs perform worse than the alignment-based per-nucleotide probability model Zoo-AL (AUC = 0.731). These LMs might therefore struggle to use context information to predict p_{ref}. In-

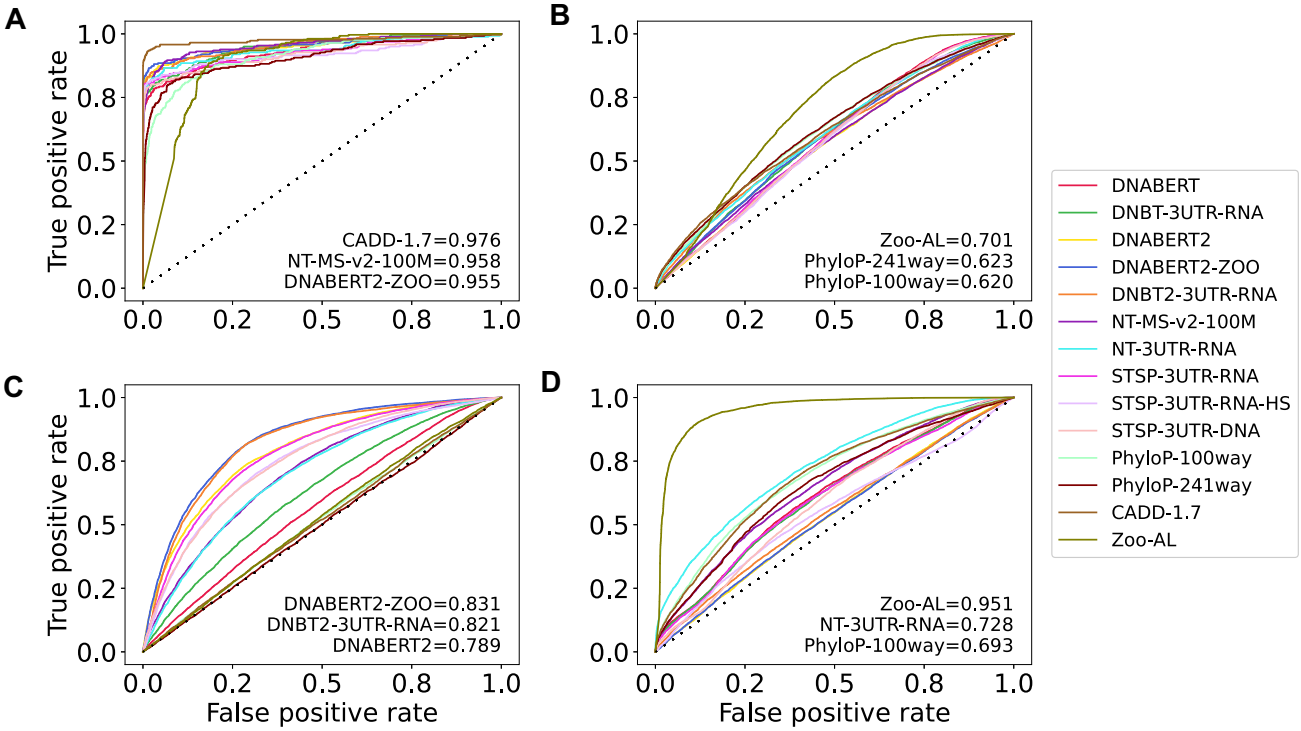


Figure 2. ROC curves for prediction of proxy-functional variants on (**A**) ClinVar, (**B**) gnomAD, (**C**) eQTL, and (**D**) CADD data using the best predictor for each model. The ROC AUC scores for the three best models are shown.

deed, when comparing the allele probabilities predicted by the LMs to the ground truth data from Zoo-AL (Supplementary Fig. S2), the NT-3UTR-RNA model achieves the highest accuracy, reconstructing the ground truth allele probability with a Pearson correlation of ∼0.54. In contrast, all other 3′UTR-specific multispecies LMs exhibit Pearson correlations around 0.4.

In contrast, all LMs outperformed the alignment-based models in the weak conservation group, with the State Space models performing the best (AUC = 0.656). In this group, the proxy-non-functional motifs are on average six times more conserved (score ≈ 1.1) compared to the proxy-functional motifs (score ≈ 0.2). Consequently, the ROC AUC score for alignment-based models falls below 0.5. In contrast, all LMs score above 0.5. Compared to the high conservation group, the NT-3UTR-RNA model degrades in performance (AUC = 0.600), whereas the other LMs perform similarly.

It is worth noting that the single-genome STSP-3UTR-RNA-HS model performs closely to the multispecies State Space models. This LM appears to leverage multiple occurrences of functional 5-mers appearing within similar sequence contexts across different 3′UTRs. One may also notice that some LMs even perform slightly better on the weakly conserved motifs (Supplementary Table S2). Since the same pattern is observed for the STSP-3UTR-RNA-HS and DNABERT models, which were not exposed to any multispecies data, we attribute this performance difference to possible mislabeling in the motifs dataset.

The NT-3UTR-RNA model appears to effectively use not only the shared functional role of the same 5-mer occuring in different 3′UTRs of the same species but also evolutionary conservation, as follows from its improved performance in the high conservation group. The possible reasons why the State Space and DNABERT-based models may not utilize evolutionary conservation so effectively include the smaller size of the former and the narrower field of view of the latter.

Prioritization of functional variants

Predicting functional significance of genetic variants is a key to understanding genotype-to-phenotype relationships. As the second task, we evaluated how well the analyzed models prioritize proxy-functional variants in 3′UTR sequences. To achieve this, we measured their performance on the ClinVar, gnomAD, eQTL, and CADD datasets, each modeling a distinct variant prediction scenario with a focus on identifying pathogenic or regulatory variants. We compared the performance of LMs with zero-shot scores (Supplementary Table S3) or probing (Supplementary Table S4) as well as alignment-based models (Supplementary Table S5). The results are summarized in Fig. 2.

First, we investigated which method performs best on each of the data sets. Alignment-based models perform the best on three out of four datasets. In particular, CADD-1.7 excels on the ClinVar data (AUC = 0.975), while Zoo-AL ranks first on the gnomAD (AUC = 0.701) and CADD (AUC = 0.951) datasets. In contrast, on the eQTL data the probing models using embeddings of the Zoonomia-based DNABERT2-ZOO and DNBT2-3UTR-RNA models achieve an exceptional score of AUC = 0.831 and 0.821 correspondingly, largely outperforming all other models. In this task, we notice no systematic improvement on 3′UTR-specific models compared to their whole genome counterparts.

Let us now have a closer look at the LM zero-shot, probing, and alignment-based predictions. The ROC AUC values for 10 zero-shot scores are shown in Supplementary Table S3. In this simple zero-shot scenario, most LMs already demonstrate some predictive capability, with the best ROC AUC

of 0.735 (NT-MS-v2-100M), 0.607 (NT-3UTR-RNA), 0.581 (STSP-3UTR-DNA), and 0.728 (NT-3UTR-RNA) on ClinVar, gnomAD, eQTL, and CADD, respectively. The NT-3UTR-RNA model achieves the best or equivalent performance in three out of four datasets, which correlates with its improved ability to predict the per-nucleotide allele probability (Supplementary Fig. S2).

Previous studies suggested [18] that the VIS, which considers dependencies between positions, could be a better predictor of functional variants than p_{ref}. We observe this trend on the ClinVar and eQTL datasets; however, on the gnomAD and CADD data, p_{ref} generally outperforms VIS. To better understand the factors influencing VIS, we plotted the difference between VIS-derived and p_{ref}-derived ROC AUC as a function of the observation window width W around the variant. As shown in Supplementary Fig. S3, the resulting curve varies across models. Notably, for some LMs, the AUC improvement due to VIS begins to decline at larger W for ClinVar variants. A possible explanation is that correlations in the immediate vicinity of the variant become masked by noise introduced by uncorrelated positions further away. In contrast, for some LMs, the AUC improvement increases with W on the eQTL data, suggesting that long-range interactions may play a key role in identification of eQTL variants.

To explore the LM performance in probing, we then trained a MLP classifier to predict proxy-functional variants based on LM embeddings. The resulting ROC AUC scores are summarized in Supplementary Table S4. Compared to zero-shot prediction, all LMs considerably improved in performance on the ClinVar and eQTL data. The AUC on Clinvar ranges from 0.920 to 0.958, with insignificant performance differences between all LMs. On the eQTL data, the DNABERT2-ZOO (AUC = 0.831) and DNBT2-3UTR-RNA (AUC = 0.821) models, which we specifically trained on the Zoonomia dataset, achieved equivalent top performance. It is worth mentioning that on this data, the DNABERT2-based models consistently outperform the DNABERT-based models despite their very similar architecture. It is, therefore, likely that the difference in performance results from the significantly larger field of view of the DNABERT2-based models that allows accounting for long-range interactions. This assumption is further supported by the fact that in additional experiments, we also observed that reducing the field of view strongly impacts the performance of the DNABERT2-based models on eQTL variants. In contrast, on the gnomAD and CADD data, the performance of most LMs shows only slight improvement or even declines compared to the zero-shot scores. One possible explanation for this is the reduced number of correlated positions for these variants, as suggested by VIS (Supplementary Table S3 and Supplementary Fig. S3). The lack of such correlations might cause the difference between the reference and alternative embeddings for proxy-functional and proxy-non-functional variants to remain on a similar scale.

It is also of note that similar to the first task, the single-genome DNABERT and STSP-3UTR-RNA-HS models performed comparably to some of their multispecies counterparts. Again, this can be possible due to similarities in sequence and regulatory function between different 3′UTRs of *Homo Sapiens*.

We then performed effect prediction using alignment-based models. The results are shown in Supplementary Table S5. These models outperform the LMs on ClinVar (AUC = 0.976, CADD-1.7), gnomAD (AUC = 0.701, Zoo-AL), and CADD (AUC = 0.951, Zoo-AL). These three datasets are designed in such a way that proxy-functional variants are expected to have a higher p_{ref} and lower p_{alt} compared to the proxy-non-functional. For the PhyloP models, conservation acts as a proxy for p_{ref} while Zoo-AL directly provides the nucleotide probability, leading to higher scores on two out of four datasets. In this regard, the inferior performance of LMs can be related to their poor ability to predict the alignment-based nucleotide probability (Supplementary Fig. S2). Interestingly, Zoo-AL outperforms CADD-1.7 (AUC = 0.690) on the CADD training dataset. A possible reason for this is the reliance of CADD-1.7 on the Zoonomia PhyloP scores, which also demonstrate suboptimal performance (Supplementary Table S5).

However, the performance of alignment-based models drops significantly on eQTL, where Zoo-AL achieves the highest AUC of only 0.525. The poor performance of alignment-based models on eQTL is likely due to poor evolutionary conservation of these variants. Indeed, 95% of proxy-functional eQTL variants have gnomAD allele frequency above 5%, which suggests that the alternative allele may also frequently occur in multiple species alignment. This makes the allele probability a poor predictor for these variants.

Prediction of MPRA expression levels and mRNA half-life

Regulatory sequence elements in 3′UTRs influence gene expression and mRNA stability. Predicting these properties provides an opportunity to evaluate whether LMs can effectively encode biologically relevant signals. To explore this potential, we designed two additional downstream tasks. Our third task consists in predicting reporter expression in two MPRA studies on 3′UTR variants. The fourth task consists in predicting mRNA half-life.

MPRA is a powerful experimental technique to explore the regulatory impact of variants. The alleles associated with the analyzed variant are first seeded into short oligonucleotide sequences, which are placed next to the reporter gene on a plasmid. By assessing reporter expression across different oligo sequences, the impact of the analyzed variant (alleles) on gene expression can be estimated.

Following the probing approach, we first applied Ridge regression to predict MPRA reporter expression from LM embeddings. MPRA data were measured in six human cell lines for several thousand variants associated with human disease and evolutionary selection [20]. The Pearson correlation coefficient r between expression predictions and the hold out data are shown in Table 1.

Across all cell lines, the region-specific STSP-3UTR-RNA model ($r = 0.30 - 0.50$) yields the best or equivalent performance across all the LMs and greatly outperforms the baseline model from [20]. All 3′UTR-specific LMs outperform their genome-wide counterparts.

Embeddings generated by distinct model architectures encode data differently. For some of them, the relevant information might not be encoded linearly with respect to the prediction target. In such cases, Ridge regression lacks the complexity required to reveal intricate patterns relevant to gene expression. Staying within the probing framework, we therefore trained a more sophisticated SVR regressor with a RBF kernel. SVR improves the prediction score for all the models by up to

Table 1. Pearson *r* correlation coefficient between Ridge-based predictions from sequence embeddings and ground truth MPRA expression from [20]

	HEK293FT	HMEC	HEPG2	GM12878	K562	SKNSH
DNABERT	0.28 ± 0.01	0.17 ± 0.02	0.27 ± 0.01	0.38 ± 0.01	0.23 ± 0.01	0.22 ± 0.01
DNBT-3UTR-RNA	0.38 ± 0.01	0.32 ± 0.01	0.35 ± 0.01	0.51 ± 0.01	0.35 ± 0.01	0.34 ± 0.01
DNABERT2	0.11 ± 0.02	0.10 ± 0.02	0.13 ± 0.02	0.31 ± 0.01	0.18 ± 0.01	0.15 ± 0.02
DNABERT2-ZOO	0.12 ± 0.02	0.08 ± 0.02	0.14 ± 0.02	0.31 ± 0.01	0.15 ± 0.02	0.11 ± 0.02
DNBT2-3UTR-RNA	0.29 ± 0.01	0.20 ± 0.01	0.28 ± 0.01	0.46 ± 0.01	0.29 ± 0.01	0.27 ± 0.01
NT-MS-v2-100M	0.14 ± 0.02	0.12 ± 0.02	0.16 ± 0.02	0.32 ± 0.01	0.20 ± 0.01	0.18 ± 0.02
NT-3UTR-RNA	0.29 ± 0.01	0.24 ± 0.01	0.28 ± 0.01	0.46 ± 0.01	0.30 ± 0.01	0.28 ± 0.01
STSP-3UTR-RNA	0.40 ± 0.01	0.30 ± 0.01	0.38 ± 0.01	0.50 ± 0.01	0.34 ± 0.01	0.32 ± 0.01
STSP-3UTR-RNA-HS	0.35 ± 0.02	0.22 ± 0.02	0.37 ± 0.02	0.45 ± 0.02	0.29 ± 0.02	0.26 ± 0.02
STSP-3UTR-DNA	0.33 ± 0.02	0.18 ± 0.03	0.34 ± 0.02	0.39 ± 0.02	0.23 ± 0.02	0.22 ± 0.02
Griesemer *et al*., 2021	0.24 ± 0.01	0.22 ± 0.01	0.24 ± 0.01	0.41 ± 0.01	0.27 ± 0.01	0.24 ± 0.01

The 95% confidence intervals are reported.

30% (Supplementary Table S6), with embeddings from STSP-3UTR-RNA leading to the best performance across all cell lines ($r = 0.39 - 0.56$). Interestingly, the human-only STSP-3UTR-RNA-HS model performs equivalently to STSP-3UTR-RNA on five out of six cell lines. This result emphasizes again the similarity in sequences and regulatory elements between different 3′UTR regions of *Homo Sapiens*. We also note that on this dataset, the RNA-based model STSP-3UTR-RNA outperforms the DNA-based model STSP-3UTR-DNA.

Again following the probing strategy, we evaluated the potential utility of LM embeddings in predicting mRNA steady-state and stability levels in Jurkat and BeasB2 cells measured in the second MPRA experiment [21]. The prediction results are shown in Supplementary Tables S7 and S8 for Ridge and SVR regressors correspondingly. For both regressors, the 3′UTR-specific models outperformed their genome-wide counterparts as well as the baseline 5-mer model from [21]. In Ridge regression, DNBT-3UTR-RNA showed the best results across all targets ($r = 0.32…0.54$), while in SVR regression it was outperformed by NT-3UTR-RNA for steady state prediction on Beas2B cells. Additionally, compared to [20], the difference between Ridge and SVR results is less pronounced. We observed that training on this dataset required stronger regularization (smaller SVR parameter *C*), which may level out the performance difference between simple (Ridge) and more complex (SVR) machine learning techniques. Again, the human-only STSP-3UTR-RNA-HS model performed equivalently to the multispecies STSP-3UTR-RNA model for all but one target.

To check if providing additional context can improve model performance, we additionally experimented with predictions based on a longer sequence context of 4096 nt, obtained by placing the original oligo sequences in their genomic context. This resulted in degraded performance for all the models. We suppose that the reason for this is that in MPRA experiments all oligo sequences are placed in a fixed context, independently of their actual position in the human genome.

Finally, we used embeddings generated by the LMs to predict mRNA half-life derived by [22]. In that study, the consensus mRNA half-life was obtained as the first principle component of a *sample x gene* matrix composed of half-life measurements from 39 human samples, spanning different cell types and measurement techniques. We note that the study of [22] did not reveal any cell type-specific differences between different samples, so the consensus half-life values are cell type-agnostic.

We first predicted the half-life from 3′UTR sequences alone, using various 3′UTR embeddings, including LM embeddings and *k*-mer ($k = 1..7$) embeddings used in the original study (Fig. 3A). In this task, the 3′UTR-specific LMs again outperformed their genome wide counterparts. The DNBT2-3UTR-RNA ($r = 0.43$) and NT-3UTR-RNA ($r = 0.42$) models demonstrated the best performance, while the *k*-mer embeddings performed the worst. Notably, DNBT-3UTR-RNA ($r = 0.41$) performed equivalently to the best models. Since the field of view of DNABERT-based models is limited to 512 nt, this suggests that the key elements defining mRNA half-life are primarily located within the first 512 nt from the 5′end of the 3′UTR. The results for most models improved when probed with SVR (Supplementary Table S9). The greatest improvement (∼80%) is achieved for *k*-mer embeddings. This clearly indicates that in contrast to LM embeddings, there is no linear relationship between the simple k-mer embeddings and mRNA half-life as the prediction target. In SVR regression, the DNBT2-3UTR-RNA and NT-3UTR-RNA models again showed the best results ($r = 0.45$).

Additional mRNA features might provide complementary information relevant for half-life prediction, which is not encoded in 3′UTR embeddings alone. The difference in performance of models using various 3′UTR embeddings might thus be less pronounced when these models are also equipped with additional mRNA features. To investigate this, we considered the BC3MS model from [22], which showed the best performance among all models utilizing manually crafted sequence-based features in the original study. This model relies on basic mRNA features (length and G/C content of 5′UTR, ORF, and 3′UTR; intron length; ORF exon junction density), mRNA codon counts, 3′UTR *k*-mers ($k = 1..7$), miRNA target repression, and SeqWeaver RBP binding prediction. To test our hypothesis, we predicted mRNA half-life using the BC3MS* model, which we constructed by replacing the original *k*-mer encodings for 3′UTR sequences in BC3MS with embeddings from the DNBT2-3UTR-RNA model. The DNBT2-3UTR-RNA embeddings were chosen as they led to the best results when predicting mRNA half-life from 3′UTR embeddings alone (Fig. 3A and Supplementary Table S9). Both embeddings led to equivalent performances (Fig. 3B), with Pearson *r* values of 0.67 ± 0.04 and 0.69 ± 0.03 for BC3MS and BC3MS*, respectively, considerably outperforming all the models based on 3′UTR embeddings only. This indicates that additional mRNA features indeed encode complementary information relevant for half-life prediction, thus reducing the

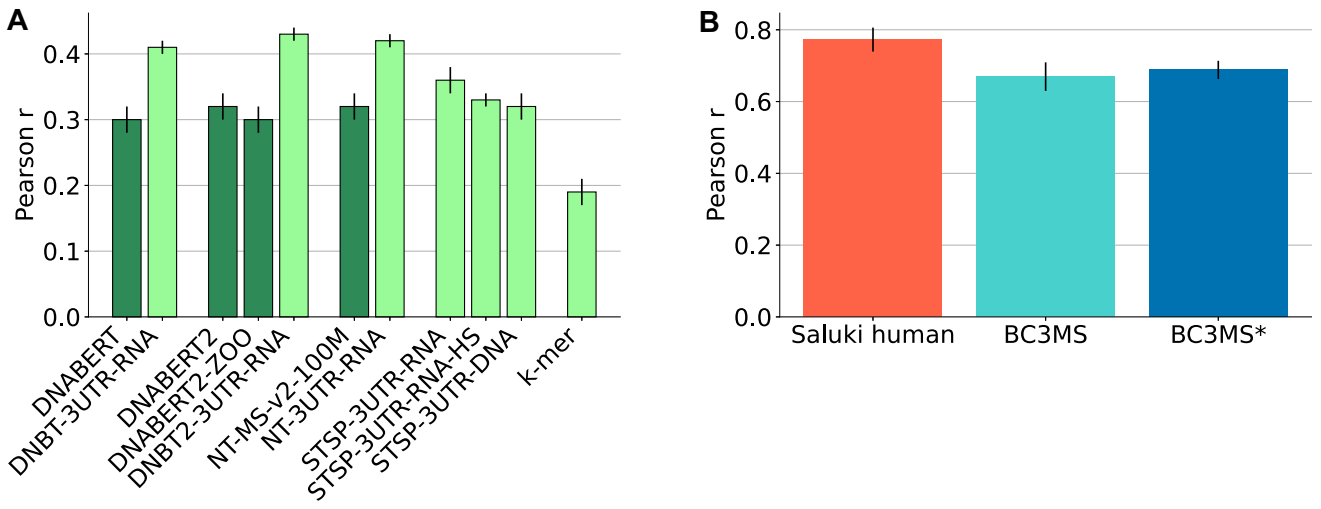


Figure 3. (**A**) Pearson *r* correlation coefficient between mRNA half-life prediction and ground truth data from [22], obtained by applying Ridge regression to different 3′UTR embeddings. (**B**) Pearson *r* correlation coefficient for mRNA half-life prediction with the BC3MS model based on different 3′UTR embeddings and the Saluki model. The performance of the BC3MS model and the BC3MS* model with DNBT2-3UTR-RNA embeddings is reported based on the SVR results. The performance of Saluki is reported as provided in the original study [22]. The error bars show the 95% confidence intervals.

impact of a particular approach used to generate 3′UTR embeddings. However, in comparison to the deep CNN-based Saluki model reaching a Pearson *r* of 0.77 ± 0.03 in [22], both BC3MS and BC3MS* exhibit inferior performance (Fig. 3B). We also observed that applying SVR produced equivalent quantitative results as Lasso regression used in the original study.

Discussion

In this work, we considered a range of 3′UTR-specific tasks to compare general-purpose genomic foundation models with LMs specifically trained on 3′UTR sequences and conservation-based models. In all tasks apart from variant effect prediction, LMs specifically trained on 3′UTR sequences outperformed their whole genome counterparts, with the optimal model being task-specific. This highlights the potential of region-specific training to capture unique regional dependencies that are overlooked by general-purpose models.

We first proceeded with identification of RBP binding motifs. In this task, the NT-3UTR-RNA model showed the best performance. When splitting motifs according to conservation, all LMs performed better than the alignment-based models on the weakly conserved group. This showcases a certain ability of LMs to leverage the specific sequencing context to predict the reference allele probability in an alignment-free manner.

We note, however, that all LMs evaluated in this task still struggle with accurate prediction of the nucleotide probability from the human-reference multispecies alignment. This suggests that sequence alignments still contain valuable information that is currently not exploited by the current strategies of self-supervised learning across many genomes. Developing new training strategies that leverage the information about orthologous sets of sequences across species could lead to LMs capable of more accurate prediction of nucleotide probability and improved prediction of functional motifs.

Another difficulty in this task is potential mislabeling of proxy-functional motifs. A given RBP is not equally likely to bind all of its 5-mers identified in the RBNS experiment. Instead, it exhibits varying degrees of affinity for each 5-mer, which we do not account for when constructing the proxy-functional set, potentially resulting in some non-functional motif hits labeled as proxy-functional. However, we believe that this way of constructing the dataset is acceptable as far as relative, rather than absolute, model performance is considered (e.g. for comparison between the models).

We then proceeded with variant effect prediction. We first observed that the LMs already exhibit certain predictive capacity by delivering meaningful zero-shot scores. Task-specific probing based on LM embeddings significantly improved variant classification on ClinVar and eQTL, suggesting non-linear relationships between LM embeddings and variant function. We note that the DNABERT2-ZOO and DNBT2-3UTR-RNA models specifically trained on the Zoonomia dataset achieved exceptional performance on 3′UTR-specific eQTLs. This indicates the strong predictive potential of Zoonomia-based models, and future studies should also consider evaluating the DNABERT2-ZOO model on other genomic regions. On the other hand, LMs with more accurate predictions of nucleotide probability might excel on the gnomAD and CADD datasets, as indicated by the superior performance of Zoo-AL on this data.

We would also like to comment on the superior performance of the alignment-based conservation models on the ClinVar and CADD datasets. In the case of ClinVar, conservation is often directly used to annotate causal non-coding variants, which might cause "label leakage." In the case of CADD, the labels of proxy-neutral variants are inferred as human lineage-derived sequence alterations in primate whole-genome alignments [12]. Therefore, the label does not directly depend on sequence conservation. However, conservation is an important predictor in the CADD model, which is in line with the good performance of conservation in predicting CADD labels observed in our analyses. Similar considerations might apply to a more subtle extent also to the gnomAD proxy-functional variants. This should be taken into account while choosing the best model for a particular application.

The superiority of region-specific LMs was corroborated when predicting MPRA expression measured in the experi-

ments of [20] and [21]. In this task, all region-specific models outperformed their genome-wide counterparts as well as the baseline models from the original studies.

The region-specific NT-3UTR-RNA and DNBT2-3UTR-RNA models also performed the best when predicting mRNA half-life based on 3′UTR embeddings only. In this task, they outperformed k-mer ($k = 1..7$) embeddings used to encode 3′UTR sequences in the original study [22]. However, using the 3′UTR embeddings alongside with other features in the BC3MS and BC3MS* models annihilated the performance difference, probably due to the additional features containing 3′UTR-specific information, such as G/C content and predicted binding score of human RBPs, as well as the features derived from other mRNA regions. Notably, encoding 3′UTRs using LM embeddings instead of k-mers did not bring the performance of the BC3MS-based models closer to that of the deep full-mRNA Saluki model [22]. This suggests that the superior performance of Saluki cannot be attributed to enhanced encoding of the 3′UTR alone, but rather to a more complex processing (compared to BC3MS) of the whole mRNA sequence.

Across all the tasks, we observed that the single-genome STSP-3UTR-RNA-HS model performed close or even equivalently to the State Space multispecies models. This may only be possible when regulatory elements with shared functional roles occur in a similar sequence context across different human 3′UTRs. The short training time of STSP-3UTR-RNA-HS (Supplementary Table S1) makes this model particularly attractive in resource-constrained settings.

We hypothesize that there could be a distinct reason why a particular LM excels in a given task. For example, it is apparently the improved ability of NT-3UTR-RNA to reconstruct the reference allele probability that makes it perform the best on motif recognition. On the other hand, the Zoonomia-based DNABERT2 models are likely to provide the exceptional results on eQTL variants due to their ability to capture long-range correlations more effectively than the other models. On the other hand, DNBT-3UTR-RNA and STSP-3UTR-RNA perform the best on MPRA effect prediction, where long-range interactions are not considered due to the small length of oligo sequences. The worse performance of the whole genome LMs in this task could be due to these models confusing short sequences coming from different genomic regions. In this regard, it was previously pointed out [5] that genome-wide NT models might encounter challenges in recognizing 3′UTR sequences. Finally, models with a greater field of view, such as DNBT-3UTR-RNA and NT-3UTR-RNA, excel on half-life predictions. Although the State Space models possess a comparable field of view, their small number of parameters might hinder their performance.

We observed that in the MPRA prediction on the [20] data, the DNA-based 3′UTR-specific model STSP-3UTR-DNA was outperformed by its RNA-based counterpart, STSP-3UTR-RNA. Notably, the evaluation strategy for this task involves training the LM on genomic alterations within a given subset of 3′UTRs, and then assessing its performance on alterations in a held-out set of 3′UTRs from a specific gene or chromosome. This approach assumes strong similarity between regulatory elements across different human 3′UTRs. Reversing sequences for genes on the negative strand (RNA-based models) may help the model recognize similar regulatory patterns across different regions, effectively augmenting the dataset. Additionally, this evaluation strategy suggests that MPRA predictions could be further enhanced by training on larger datasets, thereby increasing the chances that sequences in the training and test sets are more homogeneous (similar) in terms of sequences and functions.

When applying LMs to study fungi genomics, previous studies showed [8] that adding a species label to the model can lead to improved performance. To check this, we conducted additional tests by providing a species label to the STSP-3UTR-RNA input while training. The resulting species-aware model performed as well as STSP-3UTR-RNA across all our tasks. We speculate that the closer evolutionary distance between the Zoonomia species (up to 260 m.y.) compared to the fungi dataset from [8] (up to 500 m.y.) leads to the reduced divergence between the genomes, eliminating the impact of species-awareness.

To the best of our knowledge, there is no established strategy for selecting species for model training. To explore the impact of species selection, we trained the DNABERT2-ZOO model, which provided superior results on the variant effect prediction task compared to the original DNABERT2. While this clearly suggests that species choice can affect model performance, further research is needed to understand which characteristics of the species selection are the most important determinants of model performance. In addition to multispecies models, future studies could also explore 3′UTR-specific models trained on multiple human genomes (e.g. the 1000 Genomes dataset), since genome-wide versions of these models were shown to outperform multispecies LMs on some tasks [5].

Finally, as noted in the “Materials and methods” section, probing rather than fine-tuning was chosen in this study as the evaluation strategy due to the higher computational demands of the latter. Therefore, it remains an open question whether fine-tuned genome-wide models could outperform 3′UTR-specific LMs.

The proposed set of tasks provides a robust benchmark for evaluating genomic LMs, expanding their applicability to biologically and clinically relevant problems. By standardizing these evaluations, this benchmark lays the groundwork for future advancements in RNA-focused genomic modeling, fostering innovation and cross-comparison in the field.

Conclusion

We demonstrated that LMs trained on 3′UTR sequences from a large set of mammalian genomes outperform existing human single-genome and multispecies foundation models in three out of four 3′UTR-specific tasks. Our results highlight the importance of evaluating LMs in a region-specific manner. Additionally, the proposed set of 3′UTR-specific tasks can be used as a benchmark for future model development.

Acknowledgements

We would like to thank Julien Gagneur and Pedro Tomaz da Silva for an insightful discussion. We additionally thank Julien Gagneur for the provided computational resources that supported the start of this work.

Author contributions: S.V. and M.H. jointly conceived the study. S.V. implemented the analysis. S.V. and M.H. wrote the manuscript.

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

This work was supported by the German Ministry for Education and Research (BMBF) [031L0203A (VALE) to M.H.] within the computational life science program. M.H. is supported by the Chan Zuckerberg Initiative [2019-202666, 2021-237882]. Funding to pay the Open Access publication charges for this article was provided by Institute core funding. This study was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) via the IT Infrastructure for Computational Molecular Medicine (project #461264291).

Data availability

The pre-trained weights and the model implementation of the 6-mer DNABERT model were downloaded from the official GitHub repository: https://github.com/jerryji1993/DNABERT. The pre-trained weights and the model implementation of DNABERT2 were downloaded from the official HuggingFace page: https://huggingface.co/zhihan1996/DNABERT2-117M. The pre-trained weights and the model implementation of the NT-MS-v2-100M multispecies model were downloaded from the official HuggingFace page: https://huggingface.co/InstaDeepAI/nucleotide-transformer-v2-100M-multi-species. PhyloP-100way conservation scores were downloaded from the UCSC ftp server: http://hgdownload.soe.ucsc.edu/goldenPath/hg38/phyloP100way/. PhyloP241-way conservation scores (241-mammalian-2020v2.bigWig) and Zoonomia Progressive Cactus alignment (241-mammalian-2020v2.hal) were downloaded from the UC Santa Cruz Computational Genomics Lab web page: https://cglgenomics.ucsc.edu/november-2020-nature-mammalian-and-avian-alignments/. HAL format API was downloaded from the project's GitHub page: https://github.com/ComparativeGenomicsToolkit/hal. ClinVar data were downloaded from the NCBI ftp server: https://ftp.ncbi.nlm.nih.gov/pub/clinvar/. GnomAD data were derived from the official web page: https://gnomad.broadinstitute.org/. eQTL-SuSiE data were downloaded from the EBI ftp server: http://ftp.ebi.ac.uk/pub/databases/spot/eQTL/susie/. CADD training data were derived from the official CADD website: https://cadd.bihealth.org/. The implementation of the State Space architecture was adopted from https://github.com/DennisGankin/species-aware-DNA-LM. The raw model scores, model weights, and preprocessing data can be found at: https://doi.org/10.5281/zenodo.14993890. The scripts to process this data can be found at: https://github.com/heiniglab/investigating-foundation-models-3utr and https://doi.org/10.5281/zenodo.15286123.

References

1. Devlin J, Chang M-W, Lee K *et al.* Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies(NAACL-HLT 2019)*. Minneapolis, MN: Association for Computational Linguistics, 2019, 4171–86.
2. Brown TB, Mann B, Ryder N *et al.* Language models are few-shot learners. In: *Advances in Neural Information Processing Systems*. 2020, **33**, 1877–1901.
3. Ji Y, Zhou Z, Liu H *et al.* DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics* 2021;**37**:2112–20. https://doi.org/10.1093/bioinformatics/btab083
4. Zhou Z, Ji Y, Li W *et al.* DNABERT-2: efficient foundation model and benchmark for multi-species genomes. In: *Proceedings of the 12th International Conference on Learning Representations (ICLR 2024)* Vienna, Austria, 2024.
5. Dalla-Torre H, Gonzalez L, Mendoza-Revilla J *et al.* Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nat Methods* 2025; **22**:287–97. https://doi.org/10.1038/s41592-024-02523-z
6. Pollard KS, Hubisz MJ, Rosenbloom KR *et al.* Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* 2010;**20**:110–21. https://doi.org/10.1101/gr.097857.109
7. Siepel A, Bejerano G, Pedersen JS *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 2005;**15**:1034–50. https://doi.org/10.1101/gr.3715005
8. Karollus A, Hingerl J, Gankin D *et al.* Species-aware DNA language models capture regulatory elements and their evolution. *Genome Biol* 2024;**25**:83. https://doi.org/10.1186/s13059-024-03221-x
9. Zoonomia Consortium. A comparative genomics multitool for scientific discovery and conservation. *Nature* 2020;**587**:240–5. https://doi.org/10.1038/s41586-020-2876-6
10. Armstrong J, Hickey G, Diekhans M *et al.* Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* 2020;**587**:246–51. https://doi.org/10.1038/s41586-020-2871-y
11. Loshchilov I, Hutter F. Decoupled weight decay regularization. In: *Proceedings of the 7th International Conference on Learning Representations (ICLR 2019)*. New Orleans, LA, 2019.
12. Schubach M, Maass T, Nazaretyan L *et al.* CADD v1.7: using protein language models, regulatory CNNs and other nucleotide-level scores to improve genome-wide variant predictions. *Nucleic Acids Res* 2024;**52**:D1143–54. https://doi.org/10.1093/nar/gkad989
13. Van Nostrand EL, Freese P, Pratt GA *et al.* A large-scale binding and functional map of human RNA-binding proteins. *Nature* 2020;**583**:711–9. https://doi.org/10.1038/s41586-020-2077-3
14. Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett* 2006;**27**:861–74. https://doi.org/10.1016/j.patrec.2005.10.010
15. Landrum MJ, Lee JM, Benson M *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* 2018;**46**:D1062–7. https://doi.org/10.1093/nar/gkx1153
16. Chen S, Francioli LC, Goodrich JK *et al.* A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* 2024;**625**:330–9. https://doi.org/10.1038/s41586-023-06045-0
17. Kerimov N, Hayhurst JD, Peikova K *et al.* A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nat Genet* 2021;**53**:1290–9. https://doi.org/10.1038/s41588-021-00924-w
18. Tomaz Da Silva P, Karollus A, Hingerl J *et al.* Nucleotide dependency analysis of DNA language models reveals genomic functional elements. bioRxiv, https://doi.org/10.1101/2024.07.27.605418, 13 July 2025, preprint: not peer reviewed.
19. Bergstra JS, Bardenet R, Bengio Y *et al.* Algorithms for hyper-parameter optimization. In: Shawe-Taylor J, Zemel R, Bartlett P, Pereira F, Weinberger KQ (eds.), *Advances in Neural*

Information Processing Systems(NIPS 2011). Vol. **24**. Red Hook, NY: Curran Associates, Inc., 2011, 2546–54.
20. Griesemer D, Xue JR, Reilly SK *et al.* Genome-wide functional screen of 3′UTR variants uncovers causal variants for human disease and evolution. *Cell* 2021;**184**:5247–60. https://doi.org/10.1016/j.cell.2021.08.025
21. Siegel DA, Le Tonqueze O, Biton A *et al.* Massively parallel analysis of human 3′ UTRs reveals that AU-rich element length and registration predict mRNA destabilization. *G3 (Bethesda)* 2022;**12**:jkab404. https://doi.org/10.1093/g3journal/jkab404
22. Agarwal V, Kelley DR. The genetic and biochemical determinants of mRNA degradation rates in mammals. *Genome Biol* 2022;**23**:245. https://doi.org/10.1186/s13059-022-02811-x

Received: August 16, 2024. **Revised:** July 17, 2025. **Accepted:** August 1, 2025