A long-context RNA foundation model for predicting transcriptome architecture

Ali Saberi[1,2,*], Benedict Choi[3,4,5,6,*], Sean Wang[7], Mohsen Naghipourfar[8], Arsham Mikaeili Namini[2,9], Vijay Ramani[3,7,†], Amin Emad[1,10,11,†], Hamed S. Najafabadi[2,9,12,†], Hani Goodarzi[3,14,†]

[1] Department of Electrical and Computer Engineering, McGill University, Montreal, Canada

[2] Victor P. Dahdaleh Institute of Genomic Medicine, Montreal, QC, Canada

[3] Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, CA, USA

[4] Department of Urology, University of California, San Francisco, San Francisco, CA, USA

[5] Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, San Francisco, CA, USA

[6] Bakar Computational Health Sciences Institute, University of California, San Francisco, San Francisco, CA, USA

[7] Gladstone Institute for Data Science & Biotechnology, San Francisco, CA 94158

[8] Department of Bioengineering, University of California, Berkeley, CA USA

[9] Department of Human Genetics, McGill University, Montreal, QC, Canada

[10] Mila, Quebec AI Institute, Montreal, QC, Canada

[11] The Rosalind and Morris Goodman Cancer Institute, Montreal, QC, Canada

[12] McGill Centre for RNA Sciences, McGill University, Montreal, Canada

[14] Arc Institute, Palo Alto 94305, USA

*These authors contributed equally

[†]Correspondence: hani.goodarzi@ucsf.edu, hamed.najafabadi@mcgill.ca, amin.emad@mcgill.ca, vijay.ramani@gladstone.ucsf.edu

Abstract

Linking DNA sequence to genomic function remains one of the grand challenges in genetics and genomics. Here, we combine large-scale single-molecule transcriptome sequencing of diverse cancer cell lines with cutting-edge machine learning to build LoRNA^{SH}, an RNA foundation model that learns how the nucleotide sequence of unspliced pre-mRNA dictates transcriptome architecture—the relative abundances and molecular structures of mRNA isoforms. Owing to its use of the StripedHyena architecture, LoRNA^{SH} handles extremely long sequence inputs (~65 kilobase pairs), allowing for quantitative, zero-shot prediction of all aspects of transcriptome architecture, including isoform abundance, isoform structure, and the impact of DNA sequence variants on transcript structure and abundance. We anticipate that our public data release and proof-of-concept model will accelerate varying aspects of RNA biotechnology. More broadly, we envision the use of LoRNA^{SH} as a foundation for fine-tuning of any transcriptome-related downstream prediction task, including cell-type specific gene expression, splicing, and general RNA processing.

Introduction

Cell identity and behavior in higher eukaryotes is chiefly determined by transcriptome architecture—the quantitative levels (i.e., transcript abundances) and molecular structures (i.e., constitutively and alternatively spliced exons) of RNA isoforms. Transcriptome architecture is ultimately dictated by genomic sequence, as noncoding cis regulatory elements within and outside transcribed genes dictate the starts, ends, splicing decisions, and expression levels of RNA molecules. Transcriptome architecture is heavily perturbed in diseases such as cancer, raising the need for technologies capable of learning how the genomic sequence and other factors shape transcriptomes.

Recent advances in deep representation learning have significantly advanced computational biology, enabling models that capture complex biological processes such as protein-DNA binding, chromatin accessibility, RNA splicing, alternative polyadenylation of transcripts, and post-transcriptional regulation. Large-scale pre-trained (i.e. 'foundation') models have further enhanced these capabilities, offering versatile tools to model data across various domains; in genomics specifically, foundation models like the Nucleotide Transformer (NucT) and DNABERT have shown great potential in analyzing DNA sequences, excelling at tasks such as sequence classification and variant effect prediction (1,2). To date, however, a model for joint representation of the full-context genomic sequence and associated RNA molecular structures, at nucleotide resolution, is lacking.

To build an RNA foundation model that addresses this gap, two critical components are needed: a comprehensive knowledgebase of RNA molecular structures produced from different genomic sequences, to be used as training data, and a model architecture that can capture the length and complexity of these molecular structures. To date, short read RNA sequencing has been the primary source for inference of the transcriptome architecture. This type of data, however, does not capture the connectivity of transcription start sites, spliced exons, and 3' end sites, owing to the limited length of the reads (order of ~10^2 bp). Consequently, models trained on short read data fail to grasp the transcriptome architecture end-to-end. In this study, we first report a novel transcriptomic dataset, comprising high-quality long-read data from 26 cancer cell lines derived from nine distinct tissues (generated using PacBio IsoSeq), as well as many other additional samples across mouse and human. We use our dataset, which contains around 100 million full-length reads, to build a training dataset that contains more than 7 billion training tokens forming a strong foundation for building long-context models of transcriptome architecture directly from genomic sequence.

Two major architectures have been previously utilized to build RNA models. Convolutional neural networks, augmented by residual dilated convolution or attention layers, e.g., Borzoi (3) and SpliceAI (4), allow for long-context modeling of gene expression and/or RNA processing as prediction tasks. Apart from the reliance of these models on short-read data for training, their convolutional

blocks also limit the resolution at which long-range sequence interactions can be captured. Models based on the transformer architecture have recently emerged in recent years as alternatives, such as SpliceBERT (5), which leverages masked language modeling to learn RNA grammar solely from gene annotations. However, since attention mechanisms scale quadratically with sequence length, these models fail to handle full-length pre-mRNA. While Hyena operators, such HyenaDNA (6) and Evo (7), have offered improvements for processing long stretches of sequence, they are genome models at core and are not designed to model the transcriptome architecture.

In this study, we describe the long-read RNA model with StripedHyena (LoRNA^{SH}), the first RNA foundation model that learns and predicts transcriptome architecture solely from long-context sequences. We used causal language modeling and an expanded RNA token set to train LoRNA^{SH} on our large-scale full-length transcriptome architecture dataset. We demonstrate that LoRNA^{SH} is capable of predicting the entire transcriptomic architecture, that is the combined transcript start/end sites, position, length, and combinatorial usage of exons, and their abundances. LoRNA^{SH} generates informative representations (embeddings) and calculates meaningful likelihoods for transcript isoforms, addressing the limitations of DNA-focused models in representing RNA transcript diversity. We also report the ability of LoRNA^{SH} to surpass existing long-context genome models in three zero-shot prediction tasks: (1) absolute and relative isoform abundances for a given pre-mRNA sequence; (2) in-silico exon trapping frequencies, mimicking classic exon trapping assays in cells, and (3) nucleotide variant effects while controlling isoform structure. Finally, given that LoRNA^{SH} is a generative model, we show that synthetic transcripts generated by LoRNA^{SH} are similar to the real-world transcript architectures. Our results, taken together, establish LoRNA^{SH} as a frontier model that effectively learns the grammar underlying the complex transcriptome architecture.

Results

Building a map of cancer cell-line transcriptomes using long-read RNA sequencing (C2L2)

To generate a comprehensive snapshot of the transcriptomic architecture of the human genome, we generated a large IsoSeq (isoform-sequencing) dataset using the PacBio long-read RNA sequencing platform (8). For this, we profiled 26 cell lines representing cancers of nine distinct tissues, including skin, lung, breast, liver, kidney, pancreas, colon, prostate, and bone (**Figure 1A**). We generated 35 million long-reads with a median read length of 3 kilobases and a maximum read length of 16 kilobases (**Figure 1B**), corresponding to an average of 1.5 million reads per sample (**Figure 1C**). In parallel, we also performed cDNA-matched short-read RNA sequencing for all samples, with an average of 55 million reads per sample. This large dataset of matched long- and short-read data allows for systematic annotation and quantification of transcripts in each sample. The

raw reads were then preprocessed to enable transcriptome inference and isoform quantification (Methods).

As expected, we observed a high detection rate of polyA transcripts, including protein-coding genes and long non-coding RNAs (lncRNAs), in addition to expressed pseudogenes (**Figure 1D**). The majority of transcripts we identified from our compendium match annotated transcripts (**Figure 1E**). Nonetheless, we identified 1,307 novel genes alongside 23,691 known ones, and 10,139 novel transcripts in addition to the known 90,826 isoforms. Notably, most of these novel transcripts are composed of novel combinations of known annotated splice junctions, highlighting long-range interactions in RNA processing uncovered by IsoSeq (**Figure 1E**).

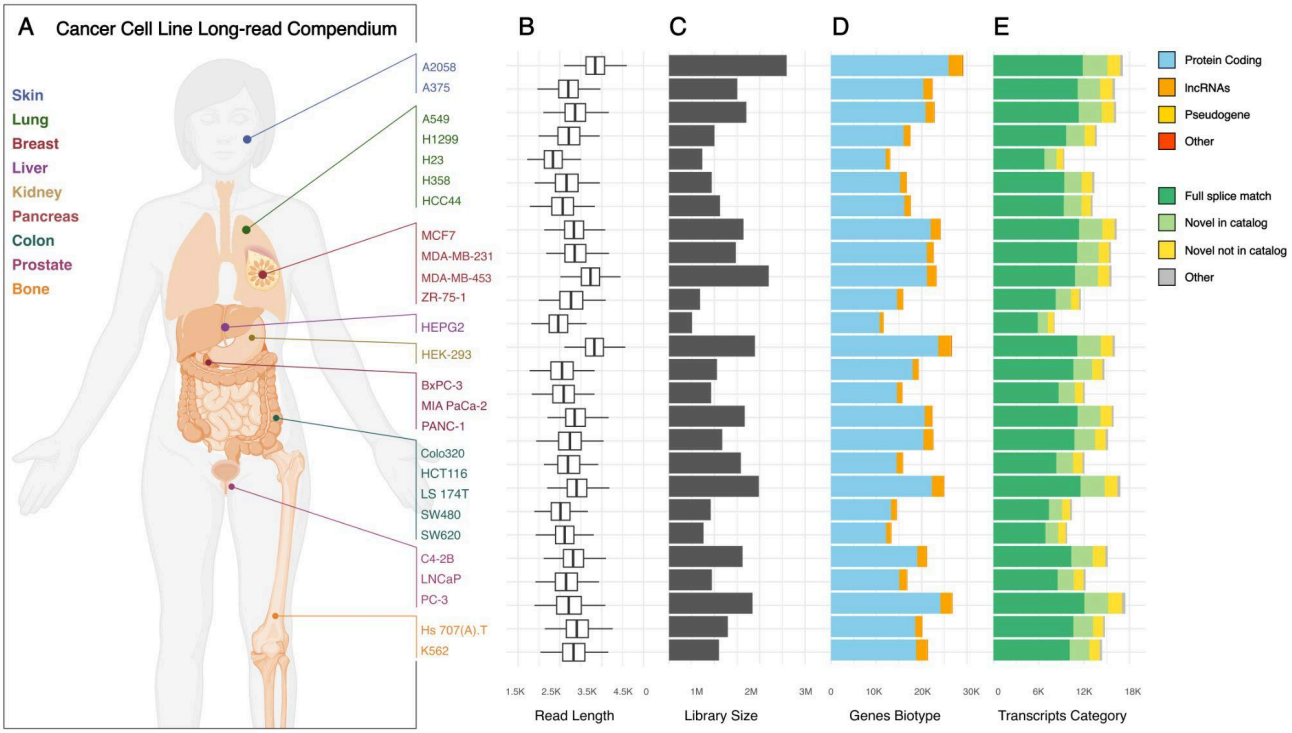

Figure 1. Overview of the Cancer Cell-Line Long-Read Compendium
(A) Visualization of the cancer cell lines sampled, mapped to their tissue of origin, with different colors representing distinct tissue types.
(B) Boxplot of read lengths across cell lines, with outliers removed for clarity. The mid-line within each box represents the median read length, while the edges of the box indicate the first (Q1) and third (Q3) quartiles, encompassing the interquartile range (IQR). The whiskers extend to 1.5 times the IQR, capturing the range of most data points, and any data points outside this range are considered outliers and have been removed for clarity.
(C) Barplot of library sizes for each cell line, indicating the total number of reads obtained.
(D) Stacked barplot of gene biotypes expressed (i.e., with at least one observed long read) in each cell line.
(E) Stacked barplot categorizing inferred transcripts by structure, including full splice match, novel in catalog (i.e., novel combination of annotated splice junctions), and novel not in catalog (i.e., transcripts with at least one unannotated splice junction, following the notation used in (9)).

Transcript quantification using our long-read data revealed highly reproducible transcriptomic profiles between biological replicates of each cell line (mean Pearson correlation 0.94). In contrast,

different cell lines showed markedly different transcriptomic profiles (mean Pearson correlation 0.60; **Figure 2A**). The divergence of cell lines is also supported by pairwise differential abundance analysis of cell lines , which revealed widespread differential expression of transcripts (**Figure 2B**). In total, we found 14,085 transcripts that were differentially expressed, FDR<0.05 (Benjamini–Hochberg False Discovery Rate), between at least two cell lines, with many transcripts showing significant differential abundance in multiple pairwise cell line comparisons (**Figure 2B**). In addition to absolute transcript abundance differences, we also observed widespread differential transcript usage (DTU; transcript usage is defined as the abundance of a transcript relative to the abundance of all isoforms of the same gene). Specifically, we identified 3,658 transcripts whose relative usage was significantly different across cell lines (beta-binomial likelihood ratio test FDR<0.05, and DTU>0.5 between at least two cell lines; **Figure 2C**). Together, these observations underline the reproducibility of cell line transcriptomic profiles obtained here and suggest highly divergent gene expression and isoform usage profiles across different cancer cell lines.

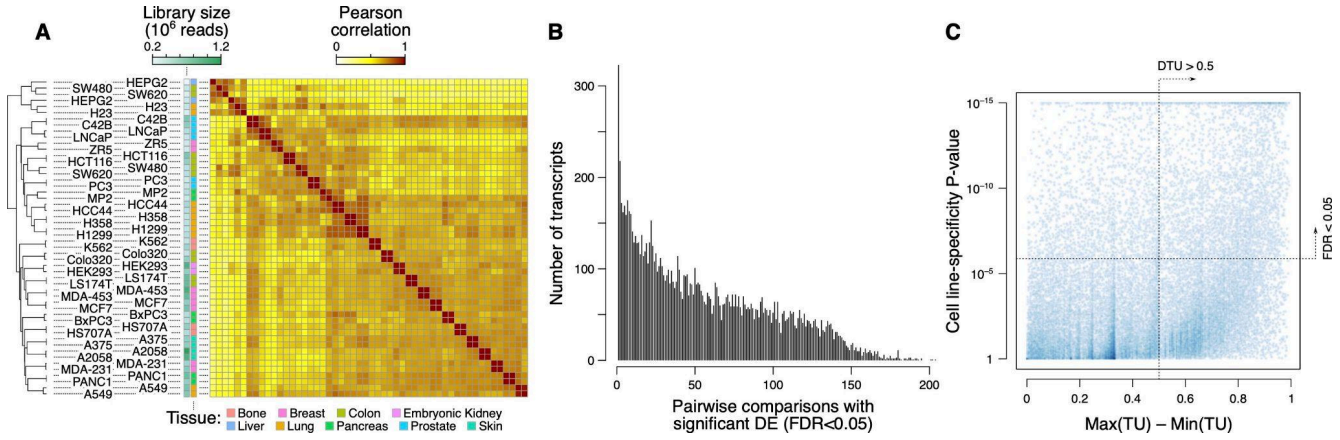

Figure 2. Overview of Transcriptome Diversity Across Cancer Cell Line Long-Read Compendium (C2L2) **(A)** Correlation heatmap showing the similarities and divergences across different cancer cell lines. Side bars represent the long-read library size and tissue of origin of each cell line. **(B)** Histogram of transcripts found to be differentially expressed between cell lines. The x-axis represents the number of pairwise cell line comparisons. **(C)** Summary of differential transcript usage (DTU) test. The x-axis represents the maximum transcript usage difference between any two samples. The y-axis shows the P-value of DTU in at least one cell line (likelihood ratio test based on beta-binomial regression). The upper-right quartile represents transcripts with a usage difference of at least 0.5 across the samples that are significant at FDR < 0.05.

Training a long-context RNA foundation model using causal language modeling

We designed LoRNA^{SH} as a transcriptomic model that is capable of handling the full length pre-mRNA sequences along with the molecular structure of resulting mature mRNAs (**Figure 3A**). The architecture of LoRNA^{SH} is based on the StripedHyena model, which combines multiple Hyena

layers—data-controlled convolutional operators—with multi-head attention layers enhanced by rotary position embeddings, allowing LoRNA^{SH} to handle sequences up to 64K nucleotides in length, a sufficiently long context to capture most pre-mRNAs. StripedHyena was previously used to build genome foundation models (7); however, we made a number of modifications to build a splice-aware long-context RNA model. A key step in training LoRNA^{SH} was our tokenization strategy, which allows the model to operate at single-nucleotide resolution, while capturing the overall RNA molecular structure by distinguishing between exonic and intronic regions. LoRNA^{SH} employs a specialized vocabulary of 16 effective tokens, where exonic regions are encoded in uppercase (A, C, U, G) and intronic regions in lowercase (a, c, u, g). To standardize the lengths of all isoforms of a gene, LoRNA^{SH} adds flanking upstream and downstream DNA sequences, making them identical in length and enabling more accurate and comparable likelihood assessments across different isoforms (**Figure 3B**).

LoRNA^{SH} was trained using causal language modeling (CLM) with a next-token-prediction (NTP) objective (Figure 3C), allowing it to learn the sequential dependencies inherent to RNA sequences by predicting each nucleotide based on its preceding context, which is crucial for tasks like isoform abundance prediction and variant impact analysis. To pre-train LoRNASH, as mentioned above, we compiled a comprehensive dataset by aggregating long-read RNA-seq data across a compendium of human cancer cell lines. This dataset includes 52 samples (2 biological replicates for 26 cell lines) from the C2L2 project. In addition, we incorporated 149 other human and mouse samples from various long-read RNA-seq projects, including those involving normal tissues, different cancer types, COVID-19 studies, control groups, and drug-treated conditions (unpublished data). Altogether, this dataset comprises 201 human samples, yielding 85 million FLNC reads and resulting in 277,481 unique transcripts, of which 191,273 have more than one exon and a maximum standardized length of 64K nucleotides. To further augment our training dataset and enhance LoRNASH's ability to capture sequence conservation across species, we also included 46 mouse samples, contributing 9.5 million FLNC reads and 153,133 unique transcripts, of which 106,453 have more than one exon and a maximum standardized length of 64K nucleotides. (**Table S1)**. Altogether, LoRNA^{SH} was trained on 297,726 unique transcripts, comprising over 7 billion training tokens (**Figure 3D)**, each tokenized at the single-nucleotide level using our expanded nucleotide token set and annotated with species-specific identifiers.

We first carried out a systematic analysis of scaling laws to gain insights into the optimal balance between model size, training data size, and computational resources. This analysis revealed that performance improves with increasing model size up to a certain point, after which further scaling yields diminishing returns (**Figure 3E**). Additionally, we benchmarked LoRNA^{SH}'s StripedHyena architecture against alternative architectures, including attention-only models and Hyena-only models. LoRNA^{SH} consistently outperformed these alternatives, demonstrating its ability to process long RNA

sequences while maintaining a strong contextual understanding, crucial for accurate splicing-aware RNA sequence modeling (**Figure 1F** and **Figure S1**). Based on this analysis, we selected a model consisting of 16 blocks with a width of 256 dimensions, totaling 6.5 million parameters. The architecture includes 13 Hyena layers interspersed with 3 rotary self-attention layers at regular intervals, providing an optimal balance of efficiency and performance (**Figure 3G**, **Methods**).

In addition to the expanded nucleotide token set and the special species tokens, we also included tokens for transcription start sites (TSS) and the polyA site (PAS). We used the PAS token embedding from LoRNA^{SH} as a representative of the entire transcript isoform sequence because it encapsulates the contextual information from all preceding tokens that build the totality of a transcript isoform. Interestingly, these embeddings revealed distinct clusters of transcripts, which may be driven by the presence of different regulatory elements along the transcript. As an example, we searched for sequence patterns that distinguish different clusters near the transcript 3' end site, and identified highly enriched sequence motifs matching the known preferences of different RNA-binding proteins and 3' processing signals (**Figure 3H**). This observation highlights the ability of LoRNA^{SH} in capturing regulatory information underlying RNA processing.

We also observed that isoform likelihoods are often sensitive to splicing signals, including but not limited to sequences at and around the splice-junctions. Specifically, to showcase LoRNA^{SH}'s understanding of splicing signals and regulatory elements, we analyzed an experimentally validated exon in the Grin1 gene in mice, where exon inclusion is critically dependent on two downstream Rbfox1 motifs (10). The model's predictions showed that the likelihoods for splice site signals and the Rbfox1 motifs were significantly higher in transcripts that included the exon compared to those that excluded it. This suggests that LoRNA^{SH} effectively recognizes not only the splice sites but also the regulatory elements, both proximal and distal, that are essential for splicing decisions (**Figure 3I**).

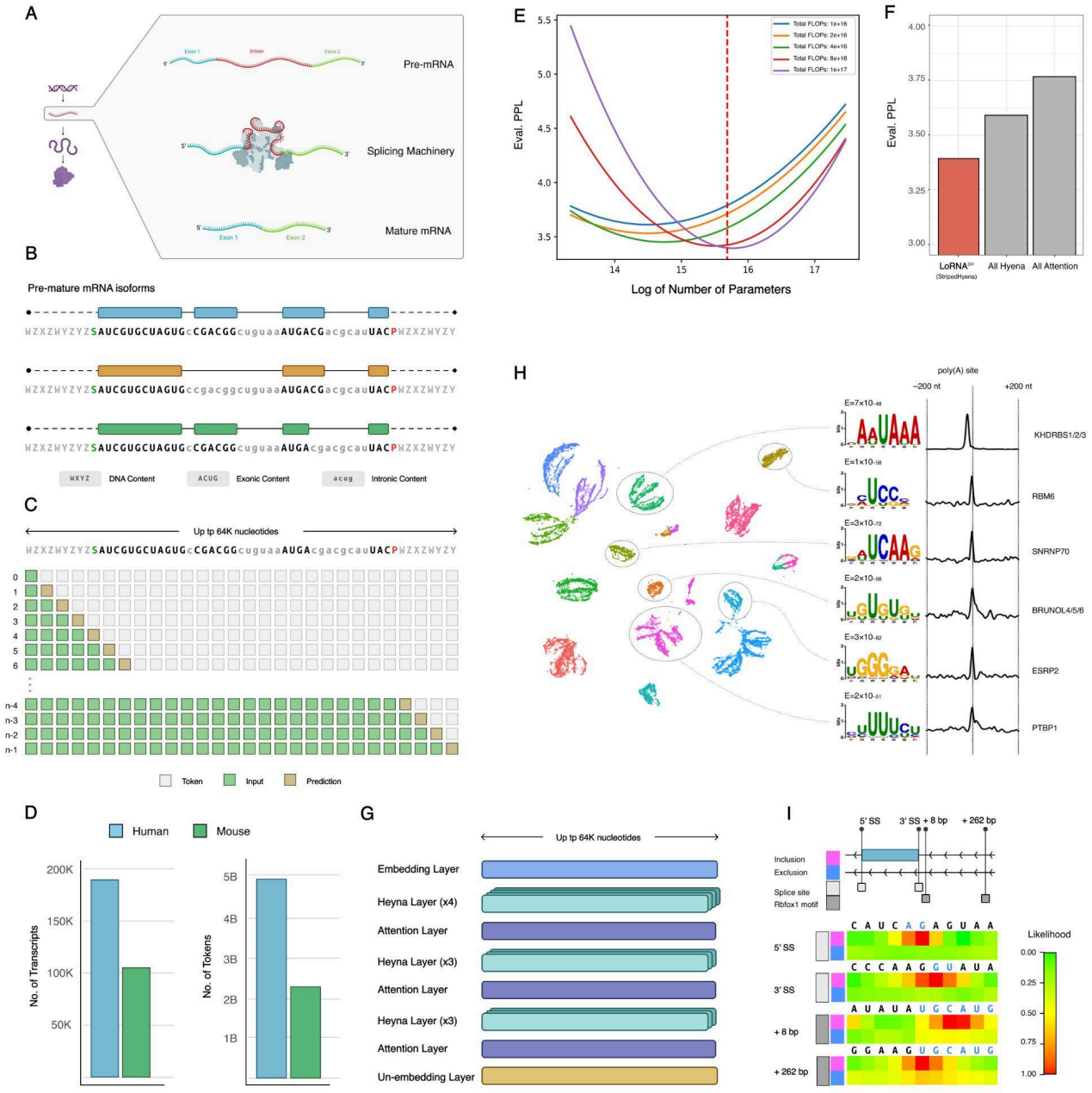

Figure 3. Detailed Representation of LoRNASH Model and Performance
(A) Diagram illustrating the transcription of pre-mRNA and the role of splicing machinery, which recognizes specific sequence features across the pre-mRNA to detect splice sites. The machinery then removes introns and joins exons to form mature mRNA.
(B) A toy example demonstratingLoRNA^{SH}'s tokenization approach given the genomic sequence and the exon-intron structure of different isoforms.
(C) Diagram of the causal language modeling used to train LoRNASH, integrating Hyena layers with attention mechanisms to handle long sequences.
(D) Summary of training data, displaying number of transcripts and tokens for human and mouse datasets.
(E) Bar plot comparing model performance for StripedHyena,Hyena-only, or attention-only architectures. PPL: Perplexity.
(F) Plot showing scaling laws, where evaluation perplexity decreases with increasing model size, measured in a

log number of parameters (ranging from 0.5 to 40 million). The x-axis represents the model size, while the y-axis shows evaluation loss. Different curves correspond to varying training budgets in FLOPs (Floating Point Operations, ranging from 1 x 10^16 to 1 x 10^17), highlighting the balance between model size and computational resources. Beyond an optimal point, further increases in model size lead to diminishing returns. **(G)** UMAP visualization of PAS embeddings. For a few example clusters, the top enriched motifs near the poly(A) sites are shown, along with their positional biases and example RBPs that may recognize those motifs. **(H)** Heatmaps showing average likelihoods for splice site signals and the Rbfox1 motif matches in the Grin1 gene, comparing transcripts with exon inclusion versus exclusion. In this experiment, each nucleotide in the sequence was systematically mutated into all possible alternatives, generating seven mutated sequences plus the original reference sequence—eight sequences in total. LoRNASH estimated the likelihood of each sequence, which was then normalized into a categorical probability distribution using the softmax function. The distribution highlights the probability of each sequence, where the only variation is a single nucleotide. A higher probability assigned to the sequence with the original reference nucleotide indicates that this nucleotide is critical for the model's predictions, demonstrating the model's sensitivity to specific nucleotide positions within RNA sequences.

LoRNA^{SH} Excels in Zero-Shot Prediction Across Diverse Transcriptomic Tasks

Zero-shot prediction is the ability of a model to make predictions on tasks it has not previously encountered during training. This capability is an indication of the model's ability to extract important, functional features of a sequence and understand their interactions, which in turn underlie the observations that are being predicted. To evaluate LoRNA^{SH}'s zero-shot prediction performance, we tested the model on three distinct downstream tasks: isoform absolute abundance or relative usage prediction, in-silico exon trapping prediction, and pathogenicity prediction for non-coding variants.

Absolute Isoform Abundance Prediction

Transcript abundance is determined by a combination of cis-regulatory features in the genomic sequence (which may act at the DNA or RNA level). We evaluated LoRNA^{SH}'s ability to predict the mean abundance across cancer cell lines using the polyA site (PAS) token likelihood, which encapsulates the model's understanding of the entire transcript. Our reasoning is that the isoforms with higher likelihood represent combinations of sequence and molecular structure that, based on LoRNA^{SH}'s internal model, are more consistent with expressed mRNAs; such isoforms are expected to be processed more frequently, which in turn results in higher mean isoform abundance. Indeed, we observed a significant correlation between LoRNA^{SH} likelihoods and absolute mRNA abundance (Pearson correlation 0.15, p-value = 8.572173e-237, **Figure 4A**). In contrast, two other state-of-the-art long-context genome models, HyenaDNA (6) and Nucleotide Transformer (1) showed poor correlation (Pearson correlations 0.03 and –0.02, respectively). We specifically benchmarked LoRNA^{SH} against these two models as, to our knowledge, they were the only models with a sufficiently long input context to enable whole isoform-level predictions. Interestingly, when the correlations are calculated within each gene, we see, on average, an even higher correlation between LoRNA^{SH} likelihoods and isoform abundances (mean Pearson correlation 0.19, **Figure 4A**), reflecting LoRNA^{SH}'s ability to predict relative transcript usage across isoforms of each gene. These results underscore LoRNA^{SH}'s

ability to model transcript-level features that govern isoform abundance. It is notable that this performance is achieved without any task-specific fine-tuning. For instance, training a simple downstream linear model on PAS token embeddings as representations of isoform sequences increased the Pearson correlation with absolute mRNA abundances to 0.38 (**Figure S3**), without even fine-tuning LoRNA^{SH} end-to-end on this task .

Trapped Exon Abundance Prediction

Using exon trapping, a recent genome-wide study identified approximately 1.25 million genomic fragments across the human genome with the ability to act as autonomous exons, suggesting that many genomic sequences, far more than the set of annotated exons within known genes, are able to get spliced if expressed as part of a pre-mRNA (11). This dataset offers a unique opportunity to step outside of known genes and annotated exons and evaluate the rules of splicing learned by models in an entirely held-out data. To evaluate LoRNA^{SH}'s performance in predicting the abundances of these trapped exons, we performed an in-silico exon trapping experiment by inserting each reported trapped exon along with 500 nucleotides upstream and downstream into the vector sequences used in the study. We then calculated the likelihood of the entire vector sequence, together with the RNA molecular structure expected from a successfully spliced trapped exon, using LoRNA^{SH}, and compared to the likelihood of the same sequence but with the RNA molecular structure expected from exon skipping. The resulting Δlog-likelihoods were indeed significantly correlated with experimentally observed trapped exon abundances (Pearson correlation 0.20, p-value = 2.143144e-28, Figure **S4**), much higher than Nucleotide Transformer, SpliceAI (4) (a model specifically trained to predict splice junctions from pre-mRNA sequence), and HyenaDNA, which had correlations of 0.14, 0.13, and 0.03, respectively (**Figure 4B**).

Pathogenicity Prediction for Non-coding Variants

Unlike coding genetic variation, which may lead to altered protein sequence and function, pathogenic non-coding variants impact human health through regulatory processes that govern gene expression. There are many non-coding variants that fall within intronic sequences or untranslated regions of the gene—if functional, these variants can influence RNA processing and splicing, which may lead to pathogenic phenotypes Accurate zero-shot prediction of non-coding variants can directly impact our understanding of specific human diseases, can help interpret the results of genome-wide association studies, and even guide clinical diagnostics and/or therapeutic development (12,13).

To examine LoRNA^{SH}'s ability to predict the pathogenicity of non-coding variants, we used a high-confidence subset of variants in the ClinVar dataset (2+ star-rated variants), and evaluated our ability to differentiate pathogenic/likely pathogenic variants from all other variants (VUS and benign)

based on the difference in likelihoods between wild-type and mutant sequences. LoRNA^{SH} achieved a strong area under the receiver operating characteristic curve (AUROC) of 0.80 and an area under the precision-recall curve (AUPRC) of 0.95 in this task. In comparison, SpliceAI scored an AUROC of 0.77 and an AUPRC of 0.64, while Nucleotide Transformer and HyenaDNA achieved AUROCs of 0.52 and 0.54, and AUPRCs of 0.86 and 0.83, respectively, **Figure 4C**). These results, obtained solely with next-token-prediction and without any task-specific fine-tuning, further highlight LoRNA^{SH}'s zero-shot performance abilities and its internal understanding of the gene regulatory logic of spliced transcripts.

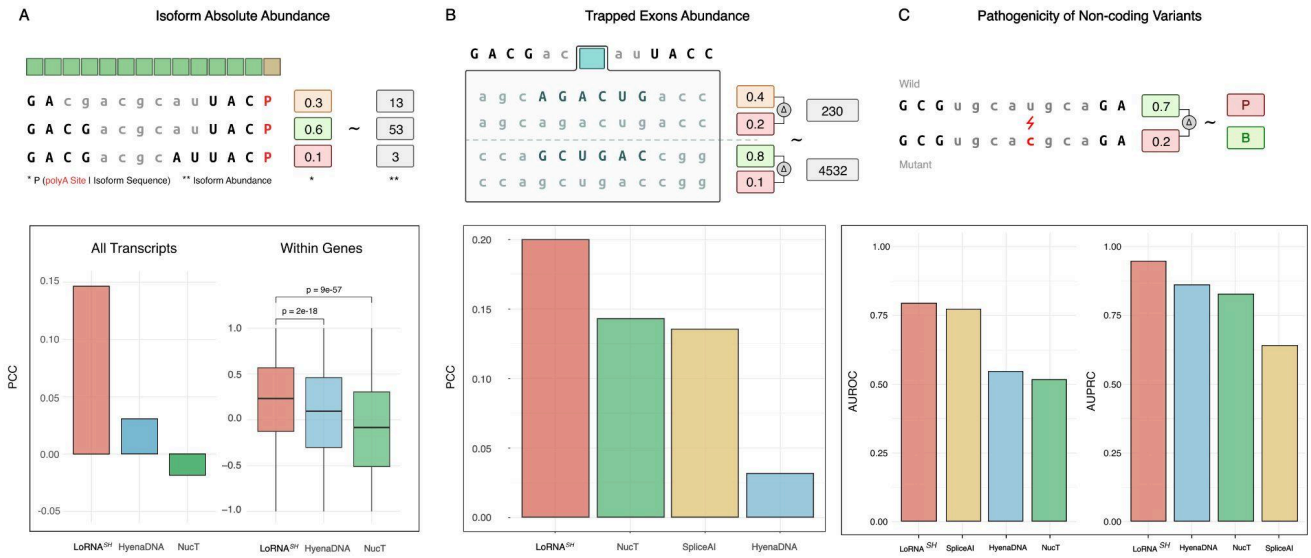

Figure 4. Model's Zero-Shot Prediction Performance on Downstream Tasks
(A) Top: Schematic of the experiment for predicting isoform absolute abundance based on the PAS token likelihood, illustrating how the model estimates the probability of isoform sequences. Bottom: The bar plots show the Pearson correlation of our model's PAS token likelihood vs absolute isoform abundances across all transcripts (all isoforms in the C2L2 dataset). The boxplots show the correlation of relative isoform usage versus likelihood within each gene.
(B) Top: Schematic of the experiment for predicting trapped exon absolute abundance, detailing how the model evaluates exon inclusion. Bottom: The correlation between the Δlog-likelihood and the absolute abundance of trapped exons.
(C) Top: Schematic of the experiment for predicting variant pathogenicity based on the difference in likelihoods between wild-type and mutant sequences. Bottom: The bar plots show the AUROC (left) and AUPRC (right).In each panel, LoRNA^{SH} is compared to other models on the same task.

Sequences generated using LoRNA^{SH} capture the architecture of human transcriptome

To assess LoRNA^{SH}'s capability in capturing the structural characteristics of transcript isoform sequences and producing synthetic exons and introns, we tasked the model with generating 1,000 transcripts. The model was prompted with 'HS', where the 'H' token directed it to simulate human transcripts, and the 'S' token instructed it to initiate from the transcription start site (TSS) and continue until reaching the 'P' token, which denotes the polyA site (PAS) (**Figure 5A**). For comparative purposes, we also sampled 1,000 transcript sequences from the annotated human transcriptome.

The analysis of exon and intron lengths between the generated and reference sequences revealed that LoRNA^{SH} recapitulates the overall structural features of human isoforms (**Figure 5B,C**). Additionally, the comparison of exon counts (**Figure 5D**) indicates that the model accurately reproduces exon-intron architecture, effectively switching between these regions in a manner consistent with naturally occurring sequences. Overall sequence length distributions (**Figure 5E**) also suggest that LoRNA^{SH} generates full-length transcripts resembling those found in the human transcriptome. Moreover, we observed that SpliceAI-based scoring of splice junctions could not distinguish synthetic and real transcripts (Wilcoxon rank-sum test, p-value = 0.06, **Figure 5F**), suggesting that LoRNA^{SH} is able to generate transcripts that, even based on other splicing models, are expected to be processed properly. It should be noted that these results were achieved without hyperparameter tuning for sequence generation.

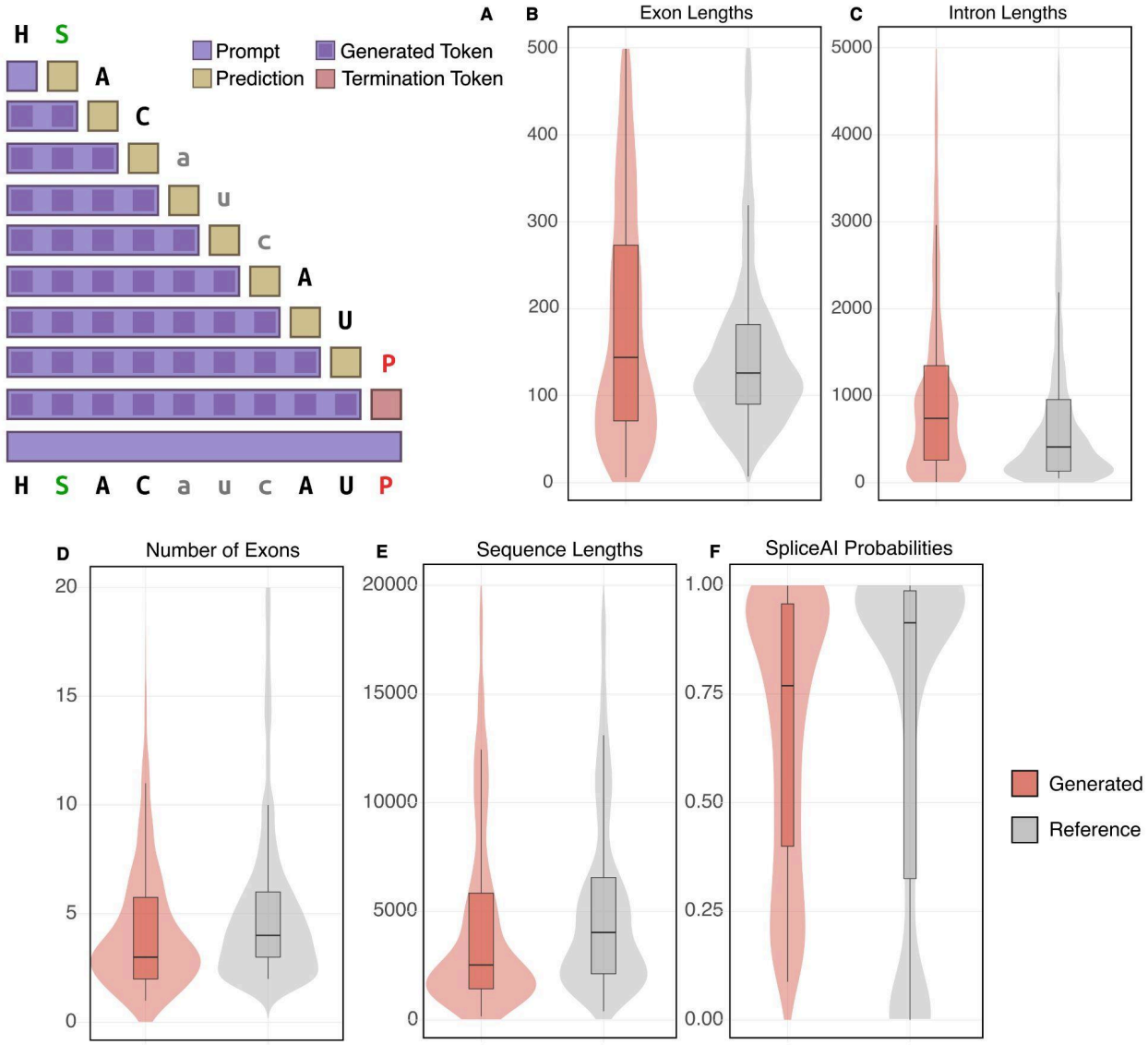

Figure 5. Comparison of Generated and Reference Transcript Sequences
(A) Schematic illustration of the procedure for transcript sequence generation using LoRNA^{SH}. The process begins with providing a prompt to the model, either H or M, to condition the model on the species (human or mouse, respectively). The first token (nucleotide) is generated based on the prompt. Subsequently, each next token is generated based on the previously generated tokens, continuing this sequence until the model generates the token P, indicating the transcript's 3' end site, at which point the generation stops. **(B-F)** Violin plots comparing key structural features between LoRNA^{SH}-generated transcripts (red) and real human transcripts (gray). **(B)** Exon lengths, **(C)** intron lengths, **(D)** number of exons, **(E)** total sequence lengths, and **(F)** SpliceAI-predicted splice site probabilities are shown.

Discussion

Our results present LoRNA^{SH} as a long-context RNA foundation model that accurately predicts transcriptome architecture directly from input genomic sequence. Building on a large number of prior efforts to model the transcriptome using deep learning (3–5,14), LoRNA^{SH} innovates in two key ways: first, via training on a high-quality long-read transcriptomic dataset spanning 26 different cancer cell lines (along with additional datasets); second, by incorporating StripedHyena operators that enable input scaling to >60,000 tokens. The resulting model excels at a variety of zero-shot tasks currently unaddressed by existing models, including the prediction of isoform abundance and structure, prediction of trapped exon inclusion in a given sequence of interest, and prediction of pathogenic non-coding variants. Future work building on our model will focus on validating model predictions, and further fine-tuning to extend these prediction tasks to relevant biotechnological applications, such as transcriptome engineering and variant effect prediction.

A key aspect of transcriptome architecture is the relative expression of all constituent isoforms. Here, LoRNA^{SH}'s ability to predict isoform abundance suggests it has learned key sequence determinants of isoform splicing and abundance. Though the current iteration of our model is agnostic to cell type, we anticipate the broad usage of LoRNA^{SH} as a foundation for future inquiries into the major drivers of cell-type specificity in transcriptome structure, including cis- and trans-acting regulatory factors governing alternative transcript start/end sites, exon usage, and isoform abundance.

Another distinguishing feature of LoRNA^{SH}, compared to other existing models relevant to RNA processing, is its generative capability. Leveraging this, we demonstrated that prompting our model with the species and transcription start site tokens generates mRNA sequences with expected length distributions of exons and introns, and splice junctions that closely resemble real transcripts. There are many possible future directions unlocked by this, including the design of transgenes with desirable splicing regulation patterns and expression levels.

Importantly, there are limitations to this current iteration of LoRNA^{SH}. First, our model has a maximum context length of 64k nucleotides – this length covers ~80% of transcripts, excluding longer transcriptional units. Second, our training, though performed on long-read RNA sequencing data, only covers datasets from human and mouse. As operators like StripedHyena become more efficient, we

anticipate both expanding context lengths beyond 64 kbp, and extending our long-read corpus of data across a broad range of evolutionary distinct species.

Predictive models bear much promise for biological research, but there are both molecular and computational challenges that must be addressed before they become a standard component of molecular biology research workflows. Here, we address two of these issues, by training a computational model with increased context awareness, including awareness of pre-mRNA and spliced RNA sequence/structure, using molecular biological data that captures the full context of transcriptomes. Our resulting model thus represents a powerful starting ground for new inquiries into the factors that shape human transcriptomes during development and in disease.

Methods

Data Collection

Cell culture

All cells were cultured in a 37 °C 5% CO_2 humidified incubator and in growth medium supplemented with antibiotics including penicillin (100 units/mL), streptomycin (100 µg/mL) and amphotericin B (1 µg/mL). The following cell lines were cultured in DMEM high-glucose medium (Gibco) supplemented with 10% fetal bovine serum (FBS), glucose (4.5 g/L), L-glutamine (4 mM) and sodium pyruvate (1 mM): MCF7 (ATCC HTB-22), MDA-MB-231 (ATCC HTB-26), MDA-MB-453 (ATCC HTB-131), LS174T (ATCC CL-188), A375 (CRL-1619), PANC-1 (ATCC CRL-1469), C4-2B (ATCC CRL-3315), Hs 707(A).T (ATCC CRL-7448), HEK-293 (ATCC CRL-1573) and A2058 (ATCC CRL-3601). The next cell lines were cultured in RPMI-1640 medium (Gibco) supplemented with 10% FBS, L-glutamine (2 mM) and HEPES (25 mM): BxPC-3 (ATCC CRL-1687), LNCaP (ATCC CRL-1740), ZR-75-1 (ATCC CRL-1500), H358 (ATCC CRL-5807), H1299 (ATCC CRL-5803), Colo320 (DZMC ACC-144), H23 (ATCC CRL-5800), HCC44 (DZMC ACC-534), MIA PaCa-2 (ATCC CRL-1420) and K562 (ATCC CCL-243). The cell lines including HCT116 (ATCC CCL-247), SW480 (ATCC CCL-228) and SW620 (ATCC CCL-227) were cultured in McCoy's 5A (modified) medium (Gibco) with 10% FBS, glucose (3 g/L) and L-glutamine (1.5 mM). The lines A549 (ATCC CCL-185) and PC-3 (CRL-1435) were cultured in Ham's F-12K medium (Gibco) supplemented with 10% FBS and L-glutamine (2 mM). The HEPG2 line (ATCC HB-8065) was cultured in Eagle's Minimum Essential Medium (ATCC) supplemented with 10% FBS, L-glutamine (2 mM) and sodium pyruvate (1 mM). All cell lines were routinely screened for mycoplasma with a PCR-based assay and tested negative. They were also routinely authenticated using STR profiling at UC Berkeley Sequencing Facility.

RNA isolation

Cells were collected and lysed in TRIzol reagent (ThermoFisher). The lysate was mixed with chloroform and phase-separated in a phaselock gel tube heavy (Quantabio) by centrifugation. The aqueous phase containing RNA was then mixed with one volume of 100% ethanol, and column-purified using Zymo QuickRNA Miniprep kit with in-column DNase treatment per the manufacturer's protocol. RNA integrity (RIN ≥ 9) was confirmed using the 4200 Tapestation System (Agilent) prior to library construction.

Short-read and long-read RNA-seq

Short-read RNA-seq libraries were prepared using the SMARTer Stranded Total RNA-seq kit v3 (TaKaRa Bio) following the manufacturer's instructions. The libraries were sequenced on a NovaSeq 6000 instrument at UCSF Center for Advanced Technology. Long-read RNA-seq libraries were constructed using the NEBNext Single Cell/Low Input cDNA Synthesis & Amplification Module, PacBio Iso-Seq Express Oligo Kit and SMRTbell express template prep kit 3.0 according to the manufacturer's protocol. The libraries were sequenced on a PacBio Sequel IIe at UCSF.

Data Preprocessing

We developed a comprehensive pipeline specifically designed for long-read RNA-seq data processing, incorporating several critical tools within Nextflow to ensure thorough and accurate data preparation. The workflow began with the IsoSeq pipeline, starting with pbccs, which was used to generate highly accurate HiFi reads using raw subreads through circular consensus sequencing (CCS). These high-quality reads, known for their length and precision, were then processed using lima to demultiplex the sequences based on barcodes and to remove cDNA primers. The isoseq-refine tool was subsequently applied for further refinement, including polyA tail trimming and concatemer removal, resulting in Full-Length Non-Chimeric (FLNC) reads that capture complete transcript structures with minimal errors. Following the generation of FLNC reads, the data were aligned to the reference genome using minimap2 (15), employing the splice preset optimized for long-read spliced alignment. This setting ensures that splicing events are accurately represented, with long deletions treated as introns, thereby reflecting true splicing patterns in the alignments. After alignment, samtools (16) was used to sort and index the BAM files. For transcript isoform quantification, we utilized bambu (17), a tool specifically designed for long-read RNA-seq data, operating in both discovery and quantification modes. In the discovery mode, bambu performed error correction on the reads and extended existing annotations to identify novel transcripts, effectively inferring the transcriptome by discovering previously unannotated isoforms. In the subsequent quantification mode, bambu accurately quantified transcript abundance by associating read classes

with specific transcripts, including both known and novel isoforms. By orchestrating these tools within Nextflow, we were able to automate and streamline the entire data processing workflow, ensuring consistency, reproducibility, and high-quality output across all steps.

Training Data

As detailed in the Results section, we curated a high-quality dataset from comprehensive long-read RNA-seq data. The dataset was filtered to include transcripts with more than one exon and a maximum standardized length of 65,536 nucleotides.

Species	Num. Transcripts	Num. Tokens
human	191,273	4,985,385,906
mouse	106,453	2,299,348,935
total	297,726	7,284,734,841

To rigorously assess LoRNA^{SH}'s generalization performance, the data was split into training, validation, and test sets based on chromosome assignment, with chromosome 10 reserved for validation and chromosome 8 designated as the test set.

Species	Split	Chromosome	Num. Transcripts	Num. Tokens	% of Tokens
human	train	1-7,9,11-22,X,Y,M	179,089	4,653,411,260	93.34
human	validation	10	6,169	177,889,685	3.57
human	test	8	6,015	154,084,961	3.09
mouse	train	1-7,9,11-19,X,Y	96,504	2,084,601,665	90.66
mouse	validation	10	4,510	93,748,826	4.08
mouse	test	8	5,439	120,998,444	5.26

Model Architecture

StripedHyena architecture

LoRNA^{SH} is based on StripedHyena (18), a state-of-the-art architecture for long-sequence language modeling (19) that combines Hyena (20) and rotary self-attention operators (21). The StripedHyena architecture can be described as a sequence of transformations applied to an input sequence $X \in \mathbb{R}^{L \times d}$ where L is the sequence length and d is the dimensionality of the model's hidden states. Each layer (block) in StripedHyena is either a Hyena layer or a rotary self-attention layer, interleaved to balance the strengths of convolutional and attention mechanisms. LoRNA^{SH} consists of 16 blocks with a model width of 256 dimensions, where each block features sequence

mixing and channel mixing layers to process information along the sequence and model width dimensions, respectively. Within the sequence mixing layers, LoRNA^{SH} utilizes 13 hyena layers, alternated with 3 rotary self-attention layers at regular intervals. The channel mixing layers in LoRNA^{SH} use gated linear units. Additionally, LoRNA^{SH} applies root mean square layer normalization to the inputs of each layer.

Hyena layers

The Hyena layers in LoRNA^{SH} are pivotal for efficiently capturing long-range dependencies in sequence data, offering a computationally efficient alternative to traditional attention mechanisms. The layers employ a blend of implicitly parametrized long convolutions and data-controlled gating functions, allowing them to process sequences with a computational complexity that scales sub-quadratically with the sequence length. While the self-attention mechanism captures global dependencies by computing pairwise interactions between all elements in a sequence—a powerful but computationally expensive approach with a complexity of $O(L^2)$—the Hyena mechanism achieves similar outcomes using a combination of implicit long convolutions and data-controlled gating mechanism. These convolutions can be efficiently computed using the fast Fourier transform (FFT), allowing the Hyena mechanism to model dependencies over varying scales without explicit pairwise comparisons. This results in a more scalable complexity of $O(NDL(\log L + D))$, where N is the order of the Hyena operator, D is the model width, and L is the sequence length. Additionally, while self-attention requires explicit positional encodings to maintain the order of tokens in a sequence, Hyena uses a combination of positional encoding and windowed convolutions to encode positional information implicitly, offering greater flexibility in handling sequences of varying lengths. Detailed information of the architecture is summarized in (20).

Self-Attention Layers

Self-attention mechanisms are integral to LoRNA^{SH}'s ability to model intricate dependencies between tokens in a sequence, regardless of their positions. Originally introduced in the Transformer architecture (21), self-attention computes the importance of each token in the sequence relative to others, enabling the model to focus on relevant parts of the sequence for a given task.

Given an input sequence $X \in \mathbb{R}^{L \times d}$, self-attention operates by computing three matrices: queries Q, keys K, and values V:

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V$$

where $W_Q, W_K, W_V \in \mathbb{R}^{d \times d_k}$ are learnable weight matrices, and d_k is the dimensionality of the queries and keys. The attention scores are computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

The softmax function ensures that the attention weights sum to one, allowing the model to distribute its focus across the sequence. The division by d_k is a scaling factor that stabilizes gradients during training.

In LoRNA^{SH}, the attention mechanism is enhanced with rotary position embeddings (RoPE) (22), which encode positional information by rotating the query and key vectors:

$$Q_i' = Q_i \cdot \cos(\theta_i) + K_i \cdot \sin(\theta_i), \quad K_i' = K_i \cdot \cos(\theta_i) - Q_i \cdot \sin(\theta_i)$$

where θ_i is a rotation angle that depends on the relative position of token i in the sequence. RoPE effectively captures relative positional relationships, allowing the model to distinguish between different token orders while maintaining position invariance.

Tokenization

Tokenization is a critical component of LoRNA^{SH}'s architecture, as it defines how RNA sequences are represented within the model. In LoRNA^{SH}, RNA sequences are tokenized at a single-nucleotide resolution, with an effective vocabulary of 16 tokens from a total of 32 characters. This precise tokenization strategy is essential for capturing the detailed information in RNA sequences, particularly for distinguishing between exonic and intronic regions. The tokenization process can be described mathematically as follows: Given an RNA sequence $S = (s_1, s_2, \ldots, s_L)$ of length L, each nucleotide s_i is mapped to a token t_i from the vocabulary V. The tokenization function $f : S \to T$ maps the sequence S to a sequence of tokens $T = (t_1, t_2, \ldots, t_L)$, where $t_i \in V$.

For exonic sequences, the nucleotides are encoded using uppercase tokens A, C, U, G, while intronic sequences use lowercase tokens a, c, u, g. Additionally, special tokens W, X, Y, Z are used to represent non-RNA (DNA) sequences flanking the transcripts, and tokens S and P mark the transcription start site (TSS) and polyA site (PAS), respectively. The tokens H and M are used to specify whether the sequence belongs to a human or mouse, respectively.

Training Procedure

LoRNA^{SH} was pre-trained using causal language modeling (CLM) with a next token prediction (NTP) objective. This task involves predicting the next token in a sequence based on its preceding context, a process critical for learning the sequential dependencies inherent in RNA sequences. The

training objective is to minimize the cross-entropy loss, which measures the divergence between the predicted token distribution and the actual distribution:

$$L_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\log p(y_i|x_i)$$

Here, L_{CE} is the cross-entropy loss, N is the number of tokens in the sequence, y_i is the true token at position i, and $p(y_i|x_i)$ is the predicted probability of token y_i given the input context x_i.

From a probabilistic standpoint, cross-entropy measures the difference between two probability distributions: the true distribution and the predicted distribution. In the context of LoRNA^{SH}, the true distribution is the actual sequence of tokens in the RNA sequence, while the predicted distribution is the model's output probabilities for each token. Cross-entropy quantifies how well the probability distribution predicted by the model matches the true distribution. A lower cross-entropy value indicates that the predicted probabilities are closer to the true labels, reflecting higher confidence and accuracy in the model's predictions. In statistical terms, minimizing the cross-entropy loss is equivalent to maximizing the likelihood of the observed data under the model, a principle known as Maximum Likelihood Estimation (MLE). The MLE approach seeks to find the model parameters θ that maximize the likelihood function:

$$\hat{\theta} = \arg\max_{\theta}\prod_{i=1}^{N} p(y_i|x_i;\theta)$$

Taking the negative log of the likelihood gives us the cross-entropy loss:

$$L_{\text{CE}}(\theta) = -\sum_{i=1}^{N}\log p(y_i|x_i;\theta)$$

Minimizing this loss function is equivalent to finding the parameter set θ that makes the observed data most probable under the model, thereby optimizing the model's predictions.

Perplexity, derived from cross-entropy, is a commonly used metric for evaluating language models. It is defined as:

$$\text{Perplexity} = 2^{L_{\text{CE}}}$$

Perplexity can be interpreted as the exponential of the average log-loss per token, measuring the model's uncertainty in predicting the next token.

Scaling Laws and Comparison with All-Hyena and All-Attention Architectures

In our study, we conducted a comprehensive scaling law analysis to compare the performance of various model architectures under a compute-optimized protocol. The primary objective was to determine the optimal balance between model size and dataset size to maximize the efficiency of computational resources. Our experiments explored model sizes ranging from 0.5 million to 40 million

parameters, maintaining a fixed context length of 65,536 tokens. To ensure the reliability of our results, we gradually reduced the learning rate and repeated the experiments until convergence was achieved. Model performance was evaluated using the perplexity metric, which measures the model's accuracy in predicting the next token in a sequence—a lower perplexity indicates better performance.

To carry out the scaling law analysis, we defined a series of compute budgets ranging from 1×10^{16} to 1×10^{17} FLOPs. For each budget, we calculated the FLOPs required to process a fixed input size, representing the "cost" of running the model. We then identified the compute-optimal allocation by selecting a range of model sizes and determining the number of training tokens needed to fully utilize the compute budget. To pinpoint the optimal compute allocation, we fitted a second-order polynomial to the relationship between log model size and perplexity. The minimum point of this function indicated the compute-optimal point, which signifies the most efficient allocation of model size and training tokens for a given compute budget. This methodical approach enabled us to predict performance trends as model size, dataset size, and compute resources scale, offering valuable insights into the relative efficiencies of different architectures. For further details, please refer to (7,23).

To further evaluate the performance of LoRNA^{SH}, after identifying the optimal architecture—comprising 13 Hyena layers interleaved with 3 attention layers (totaling 16 layers)—we compared it against two other configurations: an All-Hyena model and an All-Attention model. The All-Hyena model replaces the attention layers with Hyena layers, resulting in 16 Hyena layers similar to the HyenaDNA architecture (6). Conversely, the All-Attention model replaces the Hyena layers with attention layers, resulting in 16 attention layers akin to the GPT architecture (24). All models were trained using the same hyperparameters and were allowed to train until the validation loss reached a plateau.

Motif Discovery Across PAS Embeddings Clusters

LoRNA^{SH} is composed of 16 layers, with each layer progressively encapsulating more complex and abstract information from the input sequences. As the model deepens, it learns increasingly sophisticated concepts, patterns, and dependencies. To understand the type of information encoded within each layer, the embeddings from different layers can be analyzed to observe how the samples (transcripts) are distributed across these embeddings in the hidden space. Specifically, we focused on the penultimate layer's embeddings of the PAS token for each input sequence, using these embeddings as representations of the entire sequence.

To identify distinct patterns within these embeddings, we performed clustering and found that the PAS embeddings are best described by 18 distinct clusters, as indicated by the maximum silhouette value. We then visualized these clusters using UMAP, which provided a clear depiction of how the sequences are organized in the embedding space. To further investigate the biological relevance of

these clusters, we employed the MEME suite (25) to conduct a comprehensive motif analysis. This analysis revealed motifs that were significantly enriched in specific clusters. Additionally, we examined the localization profiles of these motif hits relative to the poly(A) site, identifying potential RNA-binding proteins (RBPs) that could recognize these motifs. This analysis provides insights into the regulatory elements captured by LoRNA^{SH}, highlighting its ability to uncover biologically meaningful patterns in transcript sequences.

Grin1 Splice Sites and Rbfox1 Motifs Likelihood Analysis

LoRNA^{SH} was trained using a next-token (nucleotide) prediction task, leveraging causal language modeling to develop a deep understanding of the biological language of pre-mRNA transcripts. This approach enables LoRNA^{SH} to become a splice-aware RNA foundation model, capable of generating sequence embeddings and estimating their likelihoods. To evaluate how effectively the model understands regulatory elements, a comprehensive mutation experiment was conducted to identify which input tokens (nucleotides) are critical for the model's likelihood estimation.

In this experiment, each token in the sequence was mutated into all possible alternatives, resulting in seven mutated sequences plus the original reference sequence—eight sequences in total. LoRNA^{SH} then estimated the likelihood of each sequence. These likelihoods were normalized into a categorical probability distribution using the softmax function, where each element of the distribution represents the probability of the corresponding sequence. Since the only difference between the sequences is one token, the resulting probability distribution indicates the importance of the reference nucleotide to the model. A higher probability for the sequence with the reference token suggests that this nucleotide is critical for the model's predictions.

This experiment was performed on all tokens across all transcripts of the Grin1 gene in the mouse genome. Grin1 was chosen because previous research (10) had investigated Rbfox1 motifs located downstream of one of the cassette exons in Grin1 transcript isoforms. The analysis revealed that LoRNA^{SH} assigns high likelihood scores to splice site signals across the gene. Furthermore, when examining the +8 and +262 tokens downstream of the exon reported in the referenced study, it was observed that the model recognizes the regulatory significance of Rbfox1 motifs, both in proximity to and at a distance from the cassette exon of interest. This demonstrates the model's capability to understand and predict key regulatory elements within transcript sequences.

Zero-shot Prediction Tasks

Isoform Absolute Abundance

To assess LoRNA^{SH}'s zero-shot prediction performance for isoform absolute abundance, we input the entire pre-mRNA sequence of each transcript isoform into LoRNA^{SH} and calculate the

likelihood of the PAS (transcription termination site) token, denoted as "P." The PAS token was chosen because its likelihood $\mathcal{P}(PAS \mid IsoformSequence)$ reflects the model's comprehension of the entire isoform sequence, embedding the sequence's overall information within the likelihood of the PAS token. For example, consider an isoform sequence represented as $HSACGTacgtACGTP$, where "H" indicates a human transcript, "S" marks the transcription start site (TSS), uppercase letters $ACGT$ denote exonic regions, lowercase letters $acgt$ denote intronic regions, and "P" signifies the PAS. The model computes the probability $\mathcal{P}(P \mid HSACGTacgtACGT)$, which serves as the likelihood for this isoform sequence.

Using the cancer cell line long-read compendium data that we generated, we inferred a comprehensive transcriptome and quantified the abundance of each isoform across various cell lines via bambu, as explained in the Data Processing part. We then averaged the isoform abundance across all cell lines to determine the absolute abundance of each isoform in the dataset.

Finally, we compared the log-likelihood of the isoforms with their log-absolute abundance and observed a promising correlation, indicating that LoRNA^{SH} effectively estimates isoform abundance through its likelihood predictions.

It is important to note that likelihoods for models like LoRNA^{SH} and HyenaDNA, which use causal language modeling, are calculated directly. However, for models trained using masked language modeling (e.g., BERT), the pseudo-likelihood method is applied. Pseudo-likelihood evaluates pretrained masked language models by masking each token in a sequence one by one and computing the log probability of the masked token given the remaining sequence. These log probabilities are then summed to yield the pseudo-log-likelihood (PLL) score for the entire sequence. For the Nucleotide Transformer, we used the pseudo-likelihood approach.

Trapped Exons Abundance Task

In exon trapping experiments, a vector sequence is used to contain a region into which an arbitrary sequence of nucleotides can be inserted. For this task, we utilized the exact vector sequence employed by (11) in their experiments. Their study reported approximately 1.25 million randomly fragmented genome sequences that were trapped as exons, with the abundance of these trapped exons measured and documented. This provided us with a comprehensive list of genomic regions and their corresponding abundance.

To estimate the likelihood of a sequence fragment being trapped as an exon, we generated two sequences for each fragment. In both sequences, the vector sequence remained identical; however, in the first sequence, the genomic fragment was included as an intron (represented in lowercase, e.g., acgt), and in the second sequence, it was included as an exon (represented in uppercase, e.g., ACGT). We then fed both sequences into LoRNA^{SH} and computed their log-likelihoods.

The log-likelihood of the entire sequence was calculated by summing the log-likelihood of each token (nucleotide) given the preceding token. For example, for the sequence $SACGTacgtACGTP$, the likelihood of the entire sequence $\mathcal{P}(SACGTacgtACGTP)$ based on the chain rule is computed as:

$$\mathcal{P}(S) \times \mathcal{P}(A \mid S) \times \mathcal{P}(C \mid AS) \times \cdots \times \mathcal{P}(P \mid SACGTacgtACGT)$$

Next, we calculated the delta-log-likelihood, which is the difference between the log-likelihood of the sequence with the exon versus the sequence with the intron. This delta-log-likelihood was used as an estimator for the likelihood of the exon being trapped. We then compared the trapped exon's log-absolute abundance with the delta-log-likelihood and measured the correlation between these values.

It is important to note that for the HyenaDNA and Nucleotide Transformer models, we could only estimate the likelihood (or pseudo-likelihood) of a single sequence, due to their identical tokenization for exonic and intronic regions.

Pathogenicity Prediction for Non-coding Variants

The ClinVar dataset (26) contains variants submitted by individuals, with each variant's significance described in terms of pathogenicity. Among the 2,743,456 variants in ClinVar, 2,490,892 are single nucleotide variants (SNVs). Of these, 382,645 SNVs have a review status of at least two stars, indicating that multiple submitters have provided criteria with no conflicts, and some variants may have been reviewed by an expert panel. Within this subset, 10,456 variants are located in intronic regions, and our evaluation focuses on these intronic variants. In ClinVar, the clinical significance of a variant is classified into five categories: pathogenic, likely pathogenic, variant of uncertain significance (VUS), likely benign, and benign. By definition, "likely pathogenic" indicates a probability of 95% or higher that the variant is pathogenic, while "likely benign" suggests a similar probability of being benign. For our analysis, we merged the "pathogenic" and "likely pathogenic" labels into a single "pathogenic" category and combined the "likely benign" and "benign" labels into a "benign" category. This resulted in three clinical significance labels for each variant: pathogenic, VUS, and benign. Among the 10,456 intronic variants in our dataset, 1,942 are classified as pathogenic, 1,133 as VUS, and 7,381 as benign. Given that VUS variants lack sufficient evidence to be categorized as either pathogenic or benign, distinguishing confirmed pathogenic variants from the rest is of practical importance. Therefore, we framed our task as a binary classification problem: identifying whether a variant is pathogenic or not, where "not" includes both VUS and benign variants. This classification task involved 1,942 pathogenic variants and 8,514 non-pathogenic variants.

For each variant in our evaluation set, we generated two transcript sequences: one with the reference allele and the other with the mutant allele. We input both sequences into LoRNA^{SH} to

estimate the log-likelihood of the entire sequences and then calculated the delta-log-likelihood, defined as the difference between the log-likelihoods of the reference and mutant transcript sequences. This delta-log-likelihood was used as a score to classify the clinical significance of the variants. We then calculated the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC) based on these scores. For the Nucleotide Transformer model, it was not feasible to input the entire transcript sequence. Instead, we input the 1,024 nucleotides flanking the mutant site and estimated the delta-log-pseudo-likelihood for this model. Additionally, for the SpliceAI model, we obtained SpliceAI scores for all the variants and used these scores for classification.

Transcript Sequence Generation

To generate transcript sequences, we utilized a causal language modeling approach with the LoRNA^{SH} model. Causal language models operate in an autoregressive manner, predicting the next token in a sequence based on the preceding context. This autoregressive nature allows these models to generate sequences in a coherent, step-by-step manner, maintaining both logical and semantic consistency throughout the sequence.

In our workflow, sequences were sampled from LoRNA^{SH} using standard techniques for autoregressive models, such as top-k sampling and temperature-based decoding. These methods help strike a balance between exploration and exploitation, allowing the model to produce diverse yet meaningful outputs. LoRNA^{SH} particularly benefits from the fast recurrent mode of its Hyena layers, which enhances computational efficiency by reducing both latency and memory requirements during sequence generation (27).

To guide the sequence generation process, we employed standard language model prompting techniques. This involved conditioning the model with a given prefix, which served as the initial context for generating the subsequent output. For species-conditional generation, we used a single token representing the desired species (H for human or M for mouse) as the prompt, enabling the model to generate transcript sequences specific to the chosen species.

While we have yet to extensively fine-tune and integrate advanced causal language modeling techniques and optimizations, our approach has already proven effective in generating coherent and biologically relevant transcript sequences. The generated sequences maintain key properties such as exon and intron lengths, the number of exons, and accurate splice site signals, showcasing the potential of LoRNA^{SH} in transcript sequence generation.

Bibliography

1. Dalla-Torre H, Gonzalez L, Mendoza Revilla J, Lopez Carranza N, Henryk Grywaczewski A, Oteri F, et al. The nucleotide transformer: building and evaluating robust foundation models for human genomics. BioRxiv. 2023 Jan 15;
2. 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;
3. Linder J, Srivastava D, Yuan H, Agarwal V, Kelley DR. Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. BioRxiv. 2023 Sep 1;
4. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, et al. Predicting Splicing from Primary Sequence with Deep Learning. Cell. 2019 Jan 24;176(3):535-548.e24.
5. Chen K, Zhou Y, Ding M, Wang Y, Ren Z, Yang Y. Self-supervised learning on millions of primary RNA sequences from 72 vertebrates improves sequence-based RNA splicing prediction. Brief Bioinformatics. 2024 Mar 27;25(3).
6. 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 Nov 14;
7. Nguyen E, Poli M, Durrant MG, Thomas AW, Kang B, Sullivan J, et al. Sequence modeling and design from molecular to genome scale with Evo. BioRxiv. 2024 Feb 27;
8. Au KF, Sebastiano V, Afshar PT, Durruthy JD, Lee L, Williams BA, et al. Characterization of the human ESC transcriptome by hybrid sequencing. Proc Natl Acad Sci USA. 2013 Dec 10;110(50):E4821-30.
9. Pardo-Palacios FJ, Arzalluz-Luque A, Kondratova L, Salguero P, Mestre-Tomás J, Amorín R, et al. SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms. Nat Methods. 2024 May;21(5):793–7.
10. Gehman LT, Stoilov P, Maguire J, Damianov A, Lin C-H, Shiue L, et al. The splicing regulator Rbfox1 (A2BP1) controls neuronal excitation in the mammalian brain. Nat Genet. 2011 May 29;43(7):706–11.
11. Stepankiw N, Yang AWH, Hughes TR. The human genome contains over a million autonomous exons. Genome Res. 2023 Dec 1;33(11):1865–78.
12. Gallagher MD, Chen-Plotkin AS. The Post-GWAS Era: From Association to Function. Am J Hum Genet. 2018 May 3;102(5):717–30.
13. Peña-Martínez EG, Rodríguez-Martínez JA. Decoding Non-coding Variants: Recent Approaches to Studying Their Role in Gene Regulation and Human Diseases. Front Biosci (Schol Ed). 2024 Mar 1;16(1):4.
14. Celaj A, Gao AJ, Lau TTY, Holgersen EM, Lo A, Lodaya V, et al. An RNA foundation model enables discovery of disease mechanisms and candidate therapeutics. BioRxiv. 2023 Sep 26;
15. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018 Sep 15;34(18):3094–100.
16. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078–9.

17. Chen Y, Sim A, Wan YK, Yeo K, Lee JJX, Ling MH, et al. Context-aware transcript quantification from long-read RNA-seq data with Bambu. Nat Methods. 2023 Aug;20(8):1187–95.

18. Poli M, Wang J, Massaroli S, Quesnelle J, Carlow R. StripedHyena: Moving Beyond Transformers with Hybrid Signal Processing Models.

19. Paving the way to efficient architectures: StripedHyena-7B, open source models offering a glimpse into a world beyond Transformers [Internet]. [cited 2024 Aug 22]. Available from: https://www.together.ai/blog/stripedhyena-7b

20. Poli M, Massaroli S, Nguyen E, Fu DY, Dao T, Baccus S, et al. Hyena Hierarchy: Towards Larger Convolutional Language Models. arXiv. 2023;

21. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv. 2017;

22. Su J, Lu Y, Pan S, Wen B, Liu Y. RoFormer: Enhanced Transformer with Rotary Position Embedding. arXiv. 2021;

23. Kaplan J, McCandlish S, Henighan T, Brown TB, Chess B, Child R, et al. Scaling laws for neural language models. arXiv. 2020;

24. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. arXiv. 2020;

25. Machanick P, Bailey TL. MEME-ChIP: motif analysis of large DNA datasets. Bioinformatics. 2011 Jun 15;27(12):1696–7.

26. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014 Jan;42(Database issue):D980-5.

27. [2310.18780] Laughing Hyena Distillery: Extracting Compact Recurrences From Convolutions [Internet]. [cited 2024 Aug 22]. Available from: https://arxiv.org/abs/2310.18780

Acknowledgements

AS acknowledges financial support from the Fonds de recherche du Québec – Nature et technologies (FRQNT) under the Doctoral Training Scholarships program.

This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC) [RGPIN-2019-04460] to AE and the Canada Foundation for Innovation (CFI) JELF [project 40781]. This research was enabled in part by support provided by Calcul Québec (calculquebec.ca) and the Digital Research Alliance of Canada (alliancecan.ca).

This work was supported by the Canadian Institutes of Health Research (CIHR) grant PJT-173317 to HSN, and resource allocations from the Digital Research Alliance of Canada. HSN holds a CIHR Canada Research Chair.

HG is an Arc Institute Core Investigator.

Supplementary material

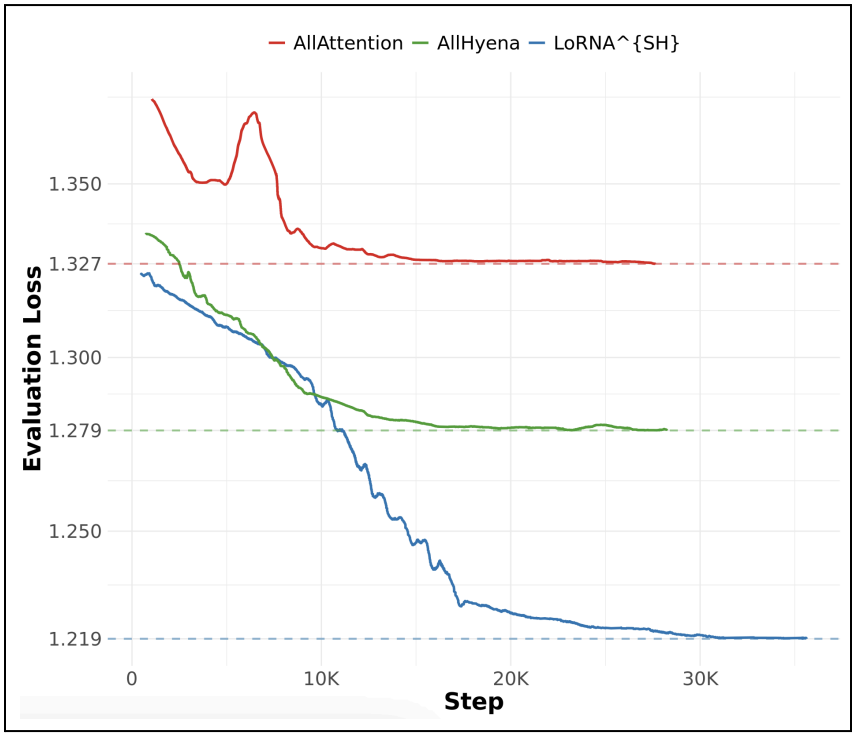

Figure S1. Evaluation Loss Curves for LoRNA^{SH}, All-Hyena, and All-Attention Architectures
This figure presents the evaluation loss curves for three different architectures: LoRNA^{SH} (blue), which utilizes the StripedHyena configuration with 13 Hyena layers and 3 interleaved attention layers; All-Hyena (green), consisting of 16 Hyena layers; and All-Attention (red), composed of 16 attention layers. The x-axis represents the training steps, while the y-axis shows the evaluation loss. As the training progresses, LoRNA^{SH} continues to decrease in loss, ultimately achieving the lowest evaluation loss. In contrast, the All-Hyena and All-Attention architectures reach a plateau at higher loss values, indicating that they have converged but with less optimal performance compared to the StripedHyena architecture in LoRNA^{SH}. The dashed horizontal lines mark the

points where each model's loss stabilizes, illustrating the superior convergence and lower loss of LoRNA^{SH} relative to the other two models.

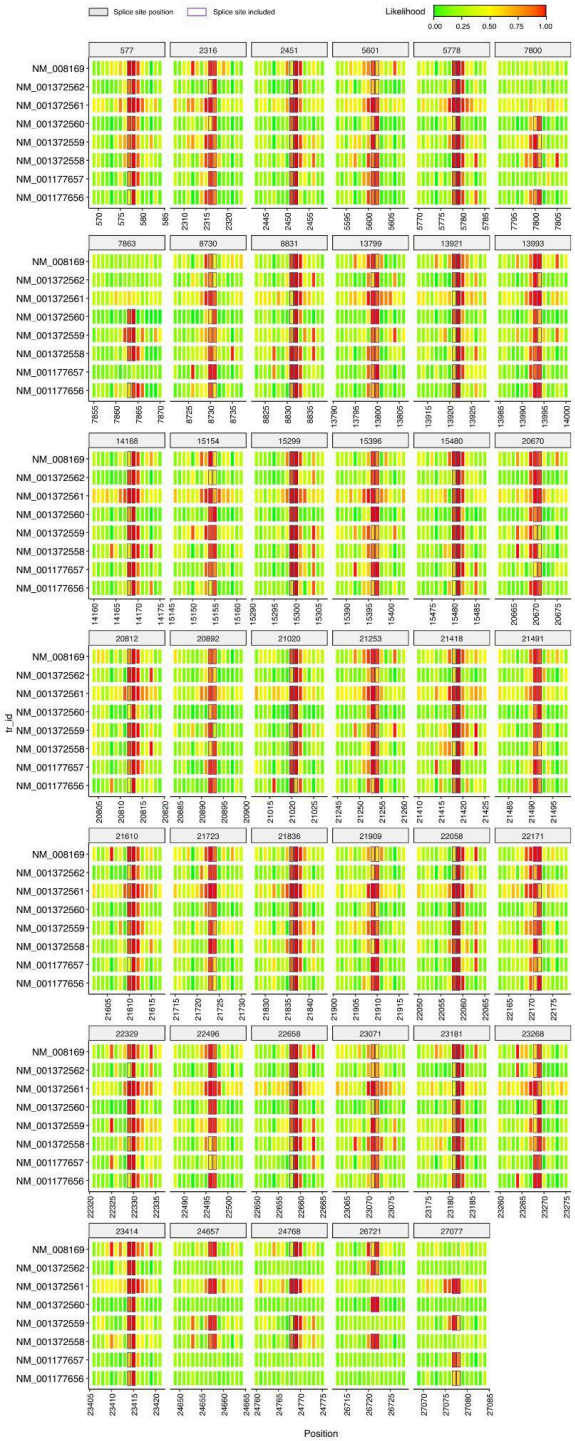

Figure S2. Likelihood of Nucleotide Positions in Grin1 Transcripts
This figure presents the likelihoods associated with each nucleotide position for all existing splice sites across eight transcripts of the *Grin1* gene in mouse. Each row corresponds to a specific transcript, and each column represents a nucleotide position within that transcript. The color scale reflects the likelihood of the entire sequence with the specific nucleotide at that position. Higher likelihoods (toward red) indicate that the nucleotide is critical for maintaining the sequence's overall likelihood, meaning that mutations at these positions would

significantly reduce the sequence's likelihood. Positions with borders represent actual splice sites in the corresponding transcript, emphasizing their importance in splice site recognition and sequence stability.

Species	Project	Sample Name	Replicate	Cell Line	Tissue	Disease	Long-read Library Size (FLNC)	Matched Short-read	Short-read Library Size
human	210222_AH_IsoSeq	MDA_MB_231_1	Rep1	MDA-MB-231	Breast	Adenocarcinoma	215,480	Yes	78,619,497
human	210222_AH_IsoSeq	MDA_MB_231_2	Rep2	MDA-MB-231	Breast	Adenocarcinoma	21,764	Yes	61,129,519
human	210222_AH_IsoSeq	MDA_MB_231_LM2_1	Rep1	MDA-MB-231	Breast	Adenocarcinoma	188,469	Yes	46,542,662
human	210222_AH_IsoSeq	MDA_MB_231_LM2_2	Rep2	MDA-MB-231	Breast	Adenocarcinoma	297,809	Yes	51,777,684
human	210222_AH_IsoSeq	SW480tr_1	Rep1	SW480	Colon	Adenocarcinoma	170,607	Yes	91,491,910
human	210222_AH_IsoSeq	SW480tr_2	Rep2	SW480	Colon	Adenocarcinoma	144,349	Yes	17,711,573
human	210222_AH_IsoSeq	SW480tr_LVM2_1	Rep1	SW480	Colon	Adenocarcinoma	144,479	Yes	85,970,880
human	210222_AH_IsoSeq	SW480tr_LVM2_2	Rep2	SW480	Colon	Adenocarcinoma	217,188	Yes	70,846,292
human	20220415_neurondiff_IsoSeq	hESC_Day0_1	Rep1	hESC	Embryo		485,819	Yes	82,999,812
human	20220415_neurondiff_IsoSeq	hESC_Day0_2	Rep2	hESC	Embryo		1,311,078	Yes	77,792,645
human	20220415_neurondiff_IsoSeq	SOX10_Day41_1	Rep1				478,909	Yes	87,092,630
human	20220415_neurondiff_IsoSeq	SOX10_Day41_2	Rep2				1,636,498	Yes	89,362,340
human	20220415_neurondiff_IsoSeq	SOX10_Day61_1	Rep1				2,108,768	Yes	77,316,444
human	20220415_neurondiff_IsoSeq	SOX10_Day61_2	Rep2				811,637	Yes	84,619,618

Table S1. Summary of Long-Read RNA-seq Data Across Human and Mouse Samples
This table provides a comprehensive overview of the long-read RNA-seq datasets generated from various human and mouse samples. Each row represents a sample, detailing the species, project name, sample name, replicate number, cell line, tissue type, and associated disease. For samples with matched short-read data, the short-read library size is also provided. Here, the first rows are shown and the full Table (248 rows) is provided as a supplemental table.

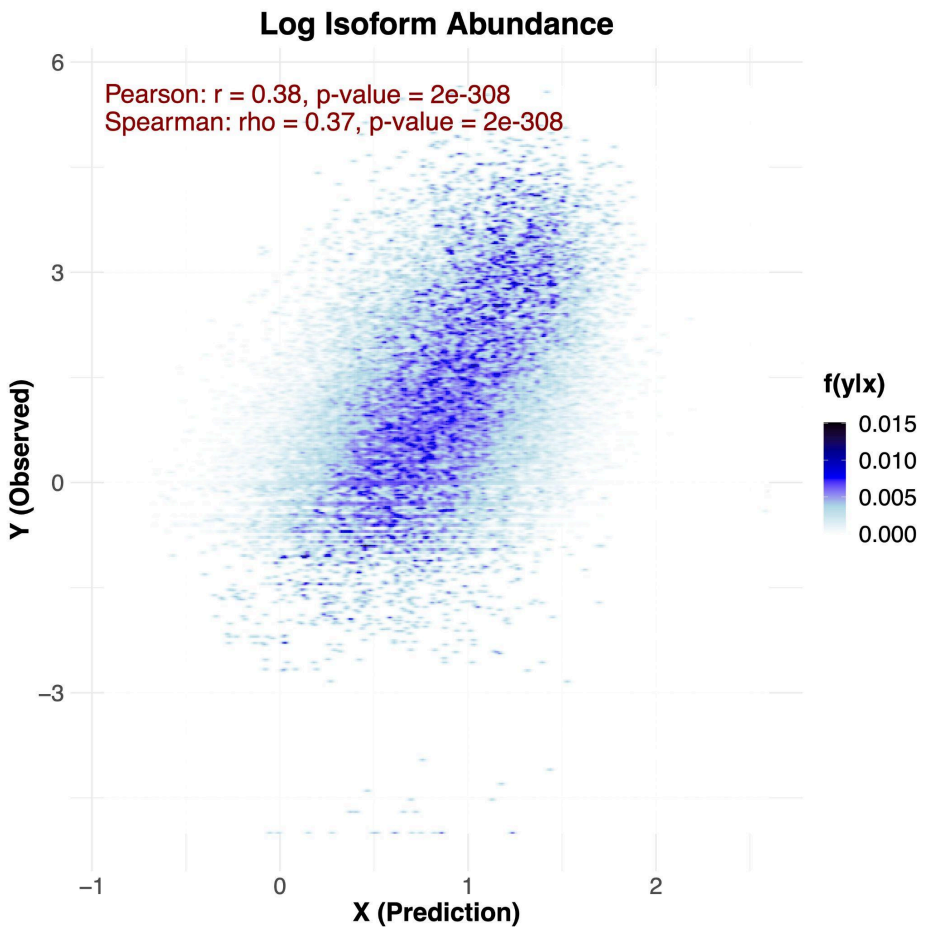

Figure S3. Correlation Between Predicted and Observed Expression Using PAS Embeddings
This figure demonstrates the predictive power of transcription termination site (PAS) embeddings for gene expression. By training a simple linear model on PAS embeddings to predict gene expression, we achieve a correlation of 0.38 in a leave-one-chromosome-out cross-validation. The x-axis represents the predicted

expression values, while the y-axis represents the observed (log10) expression values. The density of points highlights the relationship between predicted and observed expression, showing that PAS embeddings are highly informative for this task.

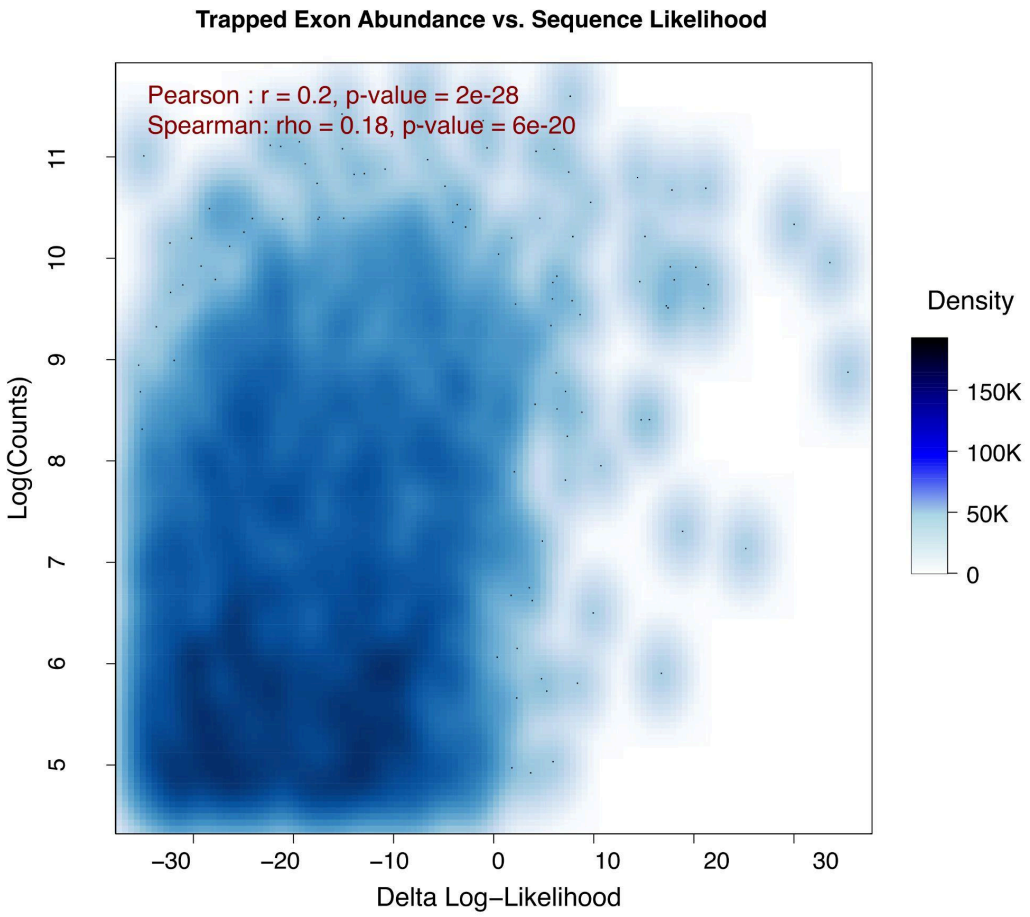

Figure S4. Correlation Between Trapped Exon Abundance and Sequence Likelihood
LoRNA^{SH} outperformed state-of-the-art models with a correlation of 0.20 between the delta log-likelihood of vector sequences when the trapped sequence is considered as an intron (lowercase tokenization) versus when it is considered as a trapped exon (uppercase tokenization) and the log of the absolute abundance counts. This smooth scatter plot visualizes the relationship between the change in log-likelihood (x-axis) and the log of exon abundance (y-axis), demonstrating LoRNA^{SH} 's ability to effectively predict trapped exon abundance based on sequence likelihood.