
Foundation models for genome biology and drug discovery
Presented by Albi Celaj for Deep Genomics

The Deep Genomics team is growing!

deep genomics

Genome biology is complicated - RNA plays a central part and ML models are needed!

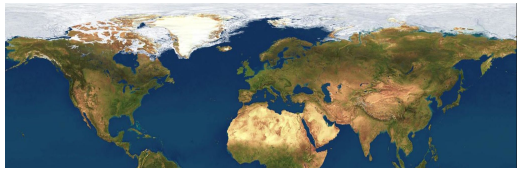
Deep Genomics Models

Algorithms: Multi-task learning, Pretraining, Fine tuning, Zero-shot learning

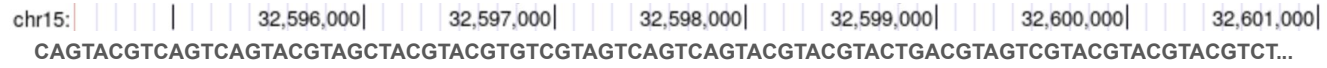
Architectures: Transformers, Structured state space models, Convolutional layers

Encoding genome biology through genome tracks: A treasure trove of experimental information for the human genome

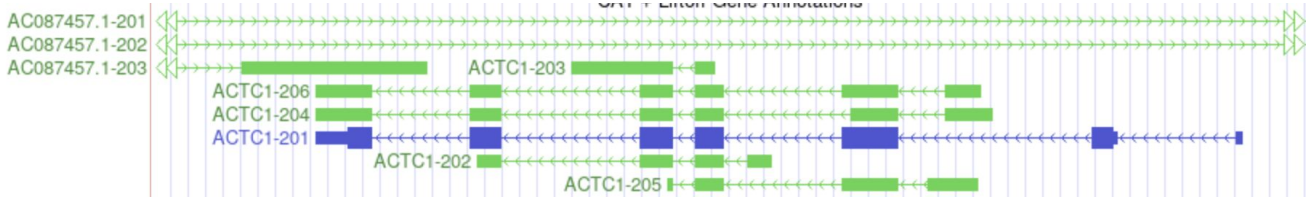
deep genomics



Raw data: DNA Sequence

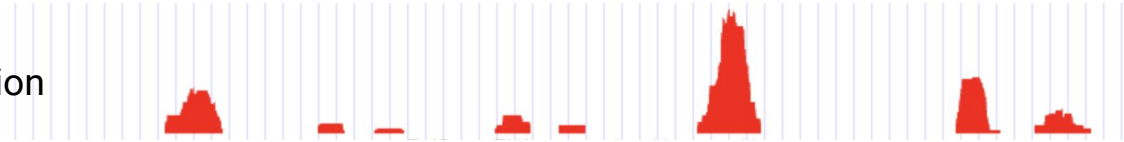


Human labels: Genes and Transcripts

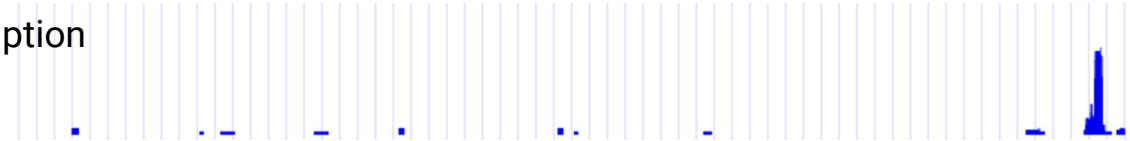


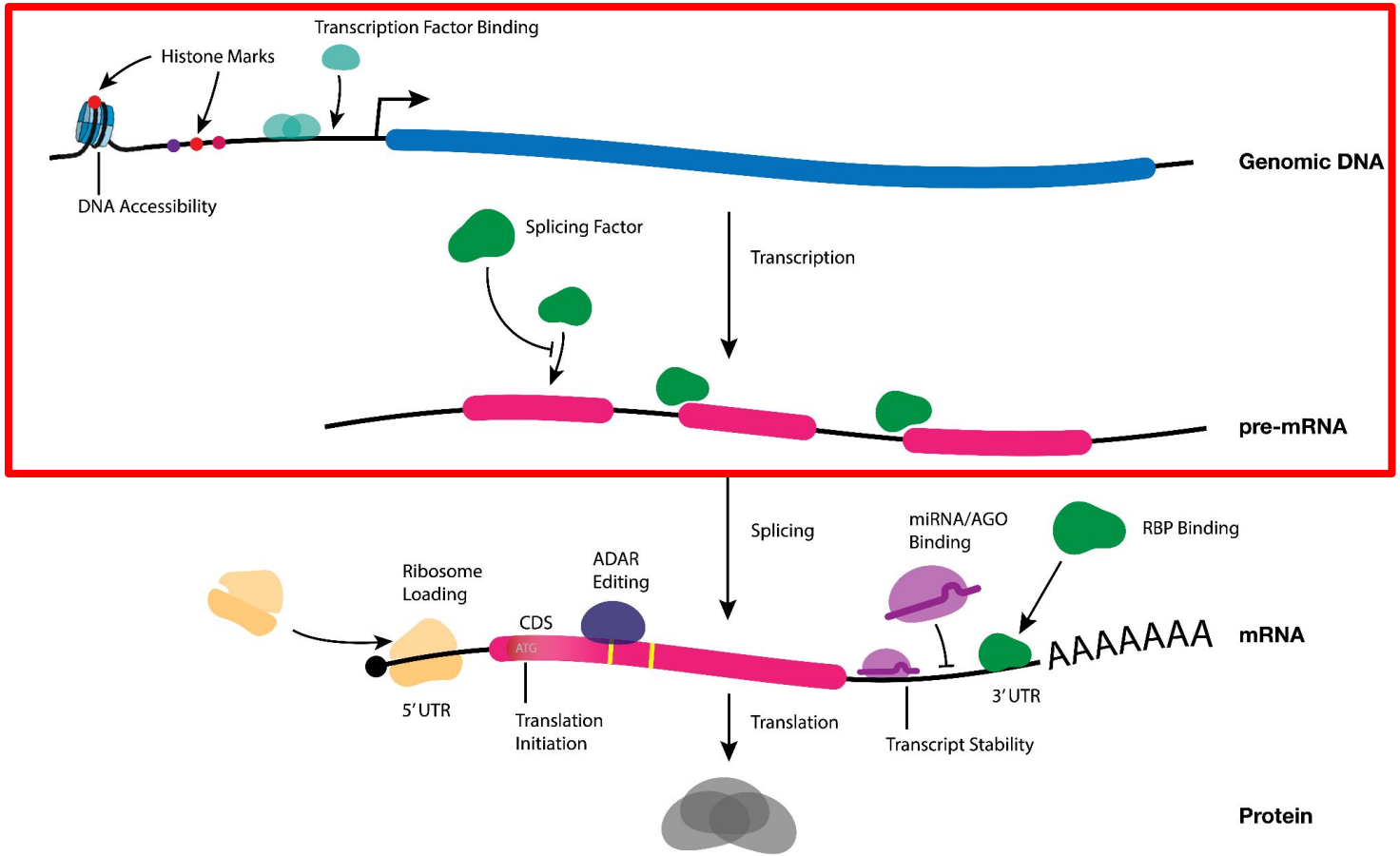
Experimental labels: Many experiments on the human genome!

RNA expression

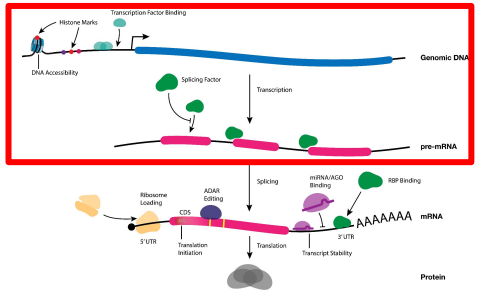


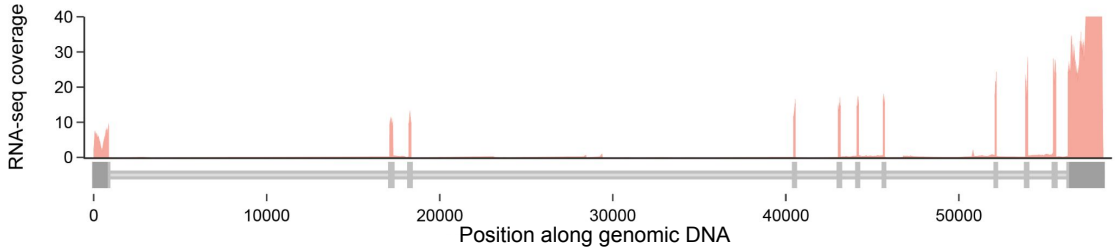
Transcription activity





Modeling some RNA genome tracks needs long-range information





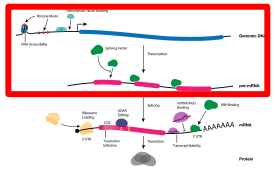
Does this look like a DNA sequence that starts a gene?

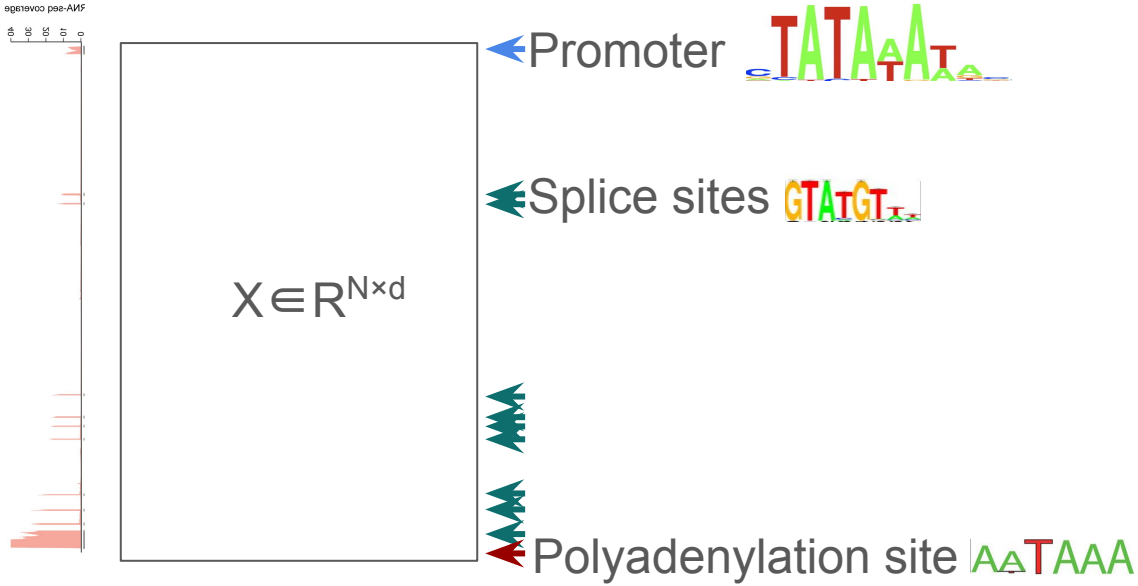
Does this look like a splice site?

Does this look like a DNA sequence that ends a gene (a polyadenylation site)?

Attention mechanism for DNA sequence

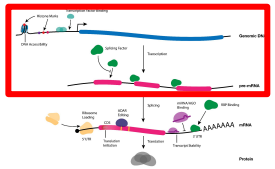
Convolutions can typically encode short range information about DNA





Embedding matrix X - N DNA basepairs by d hidden channels

Attention mechanism for DNA sequence



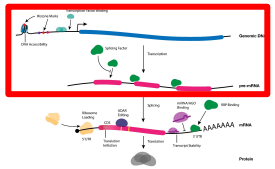
Step 1: Project DNA embeddings into Q, K, V (query, key, and value matrices)

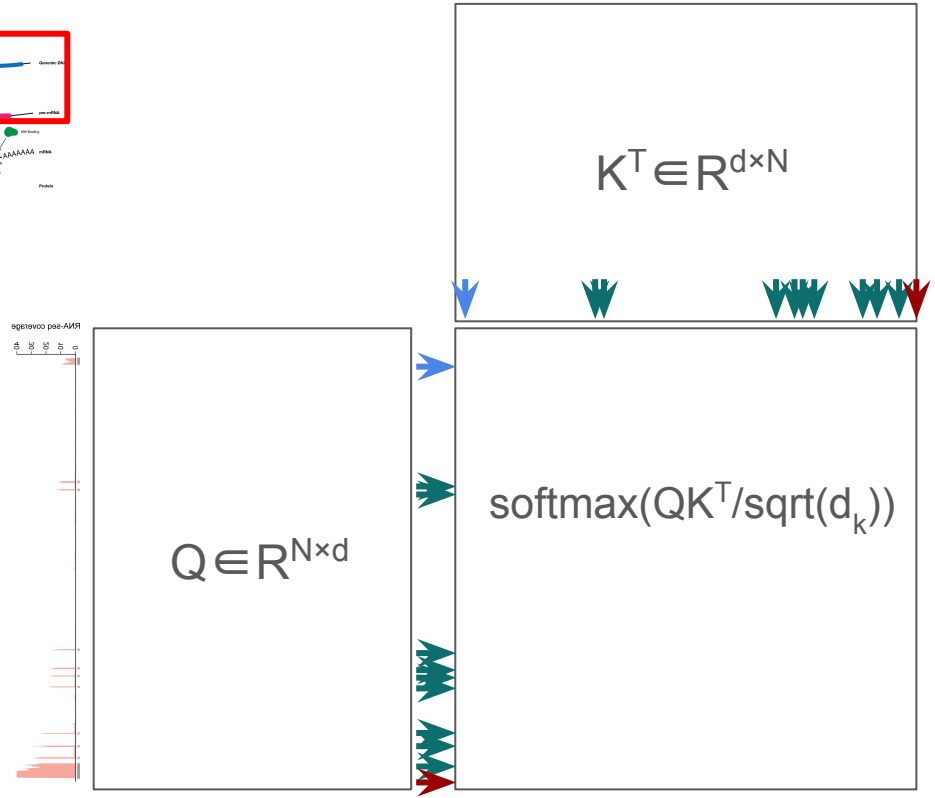
$X \in R^{N \times d}$ → Linear → $Q \in R^{N \times d}$ $K \in R^{N \times d}$ $V \in R^{N \times d}$

Embedding matrix X - N DNA basepairs by d hidden channels

Attention mechanism for DNA sequence

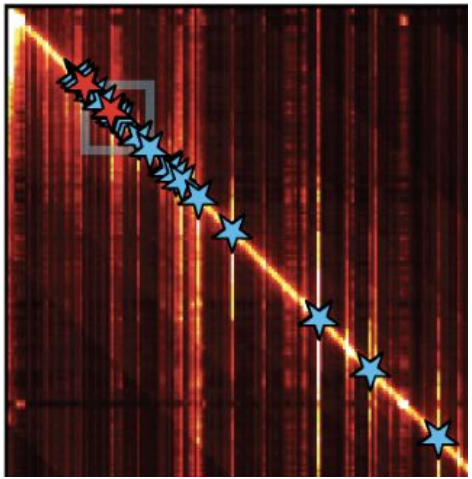
Step 2: Allow each token to attend to each other token





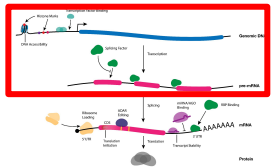
N DNA basepairs by d hidden channels

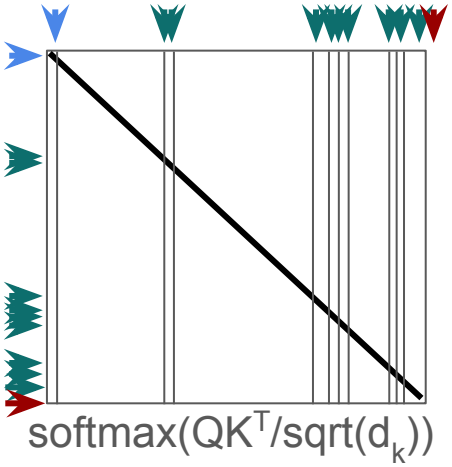
Linder et al 2023



Attention mechanism for DNA sequence

Step 3*: Transform the embeddings based on self-attention

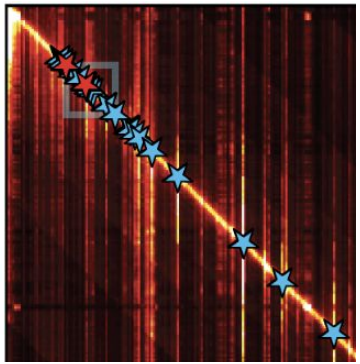




$X \in R^{N \times d}$

$X' \in R^{N \times d}$

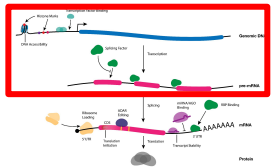
Linder et al 2023

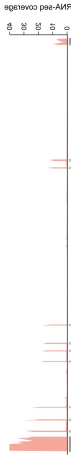


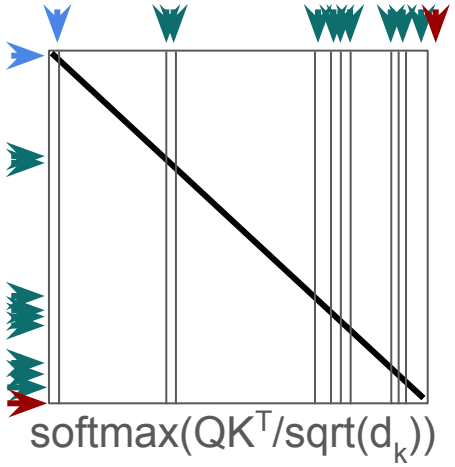
Embedding matrix X - N DNA basepairs by d hidden channels

Attention mechanism for DNA sequence

Step 3: Transform the embeddings based on self-attention



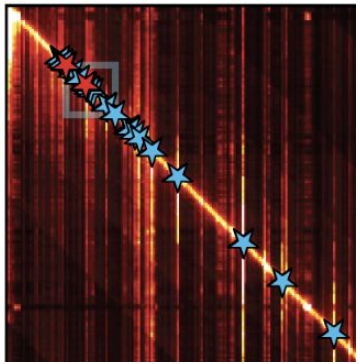




$V \in R^{N \times d}$

$X' \in R^{N \times d}$

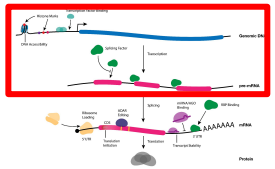
Linder et al 2023



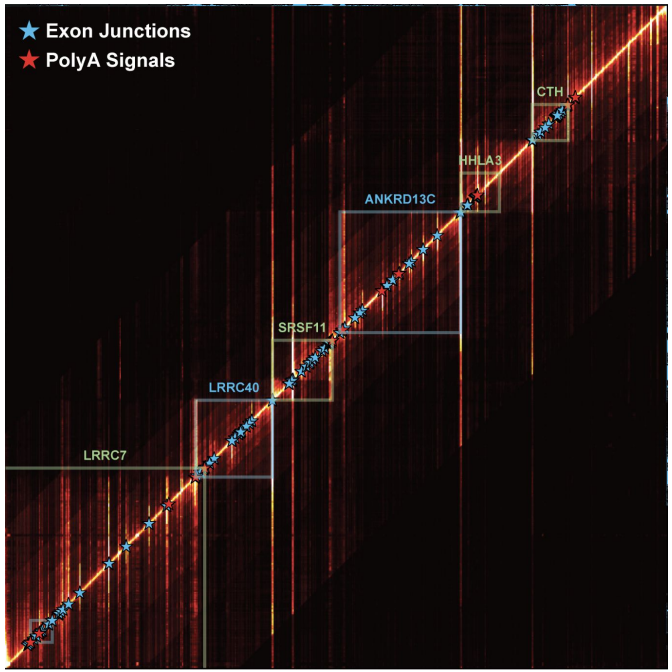
Embedding matrix X - N DNA basepairs by d hidden channels

Attention mechanism for DNA sequence

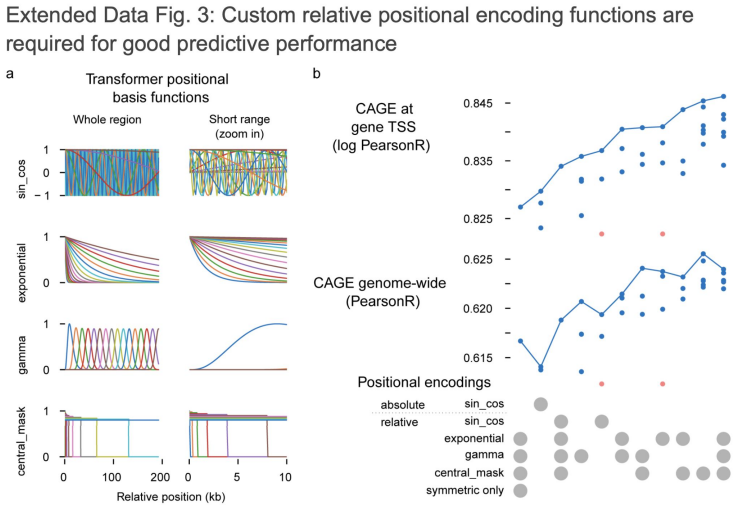
Caveat: Also need to encode position of elements



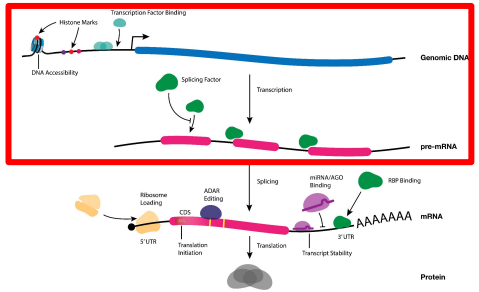
Linder et al 2023

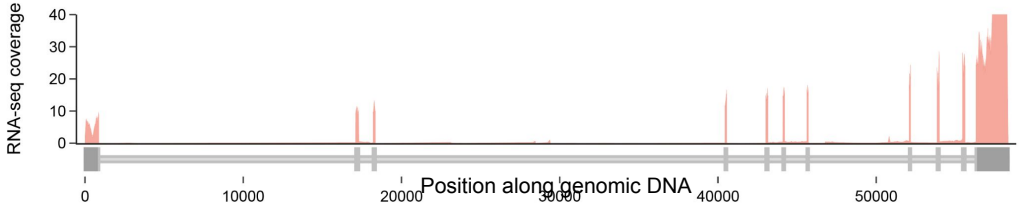


Avsec et al 2021



Modeling some RNA genome tracks needs long-range information





Does this look like a DNA sequence that starts a gene?

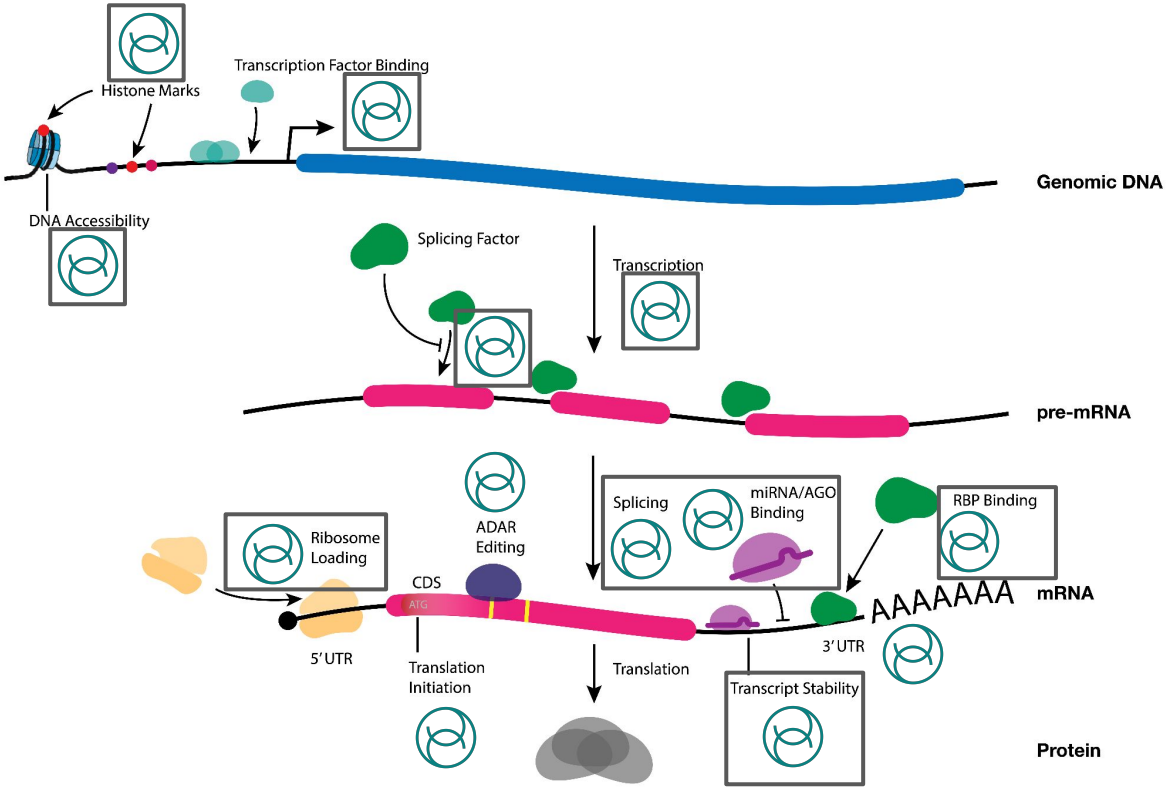
Does this look like a splice site?

Does this look like a DNA sequence that ends a gene (a polyA site)?

This is a gene start and the region allows making an RNA. The RNA looks stable

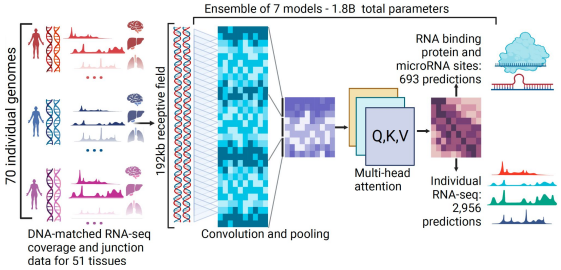
This is a splice site and we are within an RNA being transcribed. The RNA looks stable

Using a foundation model to model multiple RNA processes



BigRNA

Foundation model of RNA biology



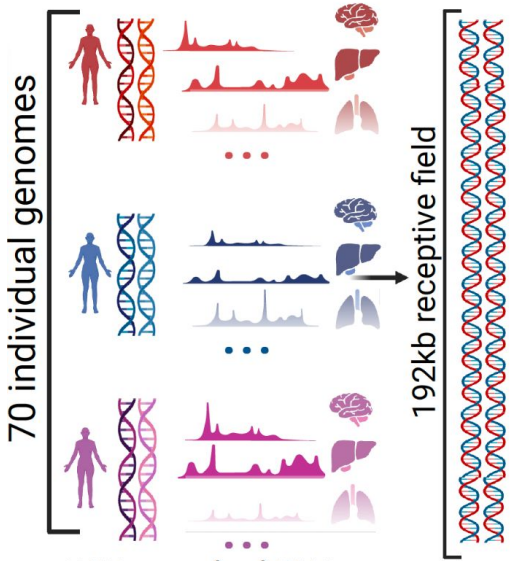

Alston Lo


Helen Zhu

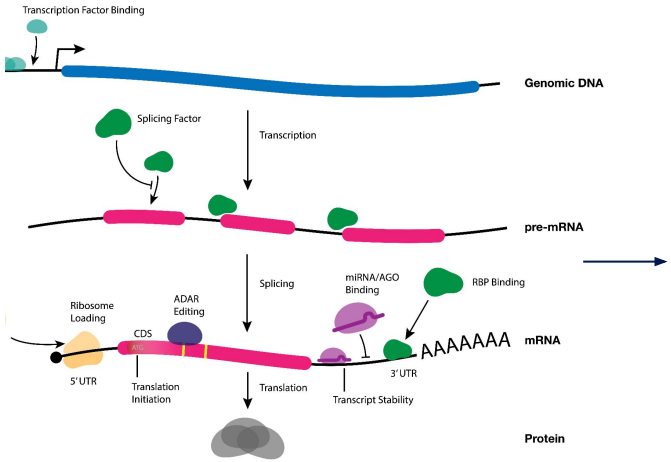

Vivian Chu

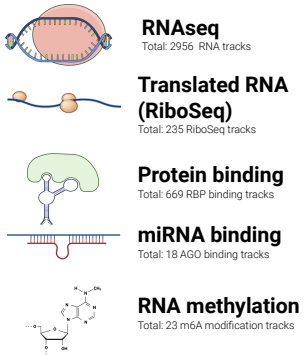

Nicole Zhang

BigRNA is an all-purpose “foundation model” for RNA genome biology



BigRNA predicts many genome tracks from DNA sequence





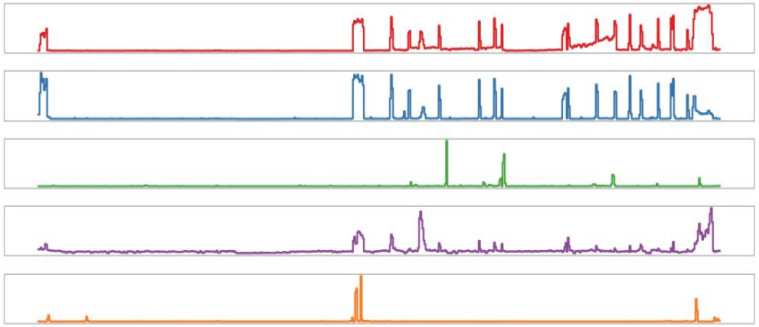
RNAseq
Total: 2956 RNA tracks

Translated RNA (RiboSeq)
Total: 235 RiboSeq tracks

Protein binding
Total: 669 RBP binding tracks

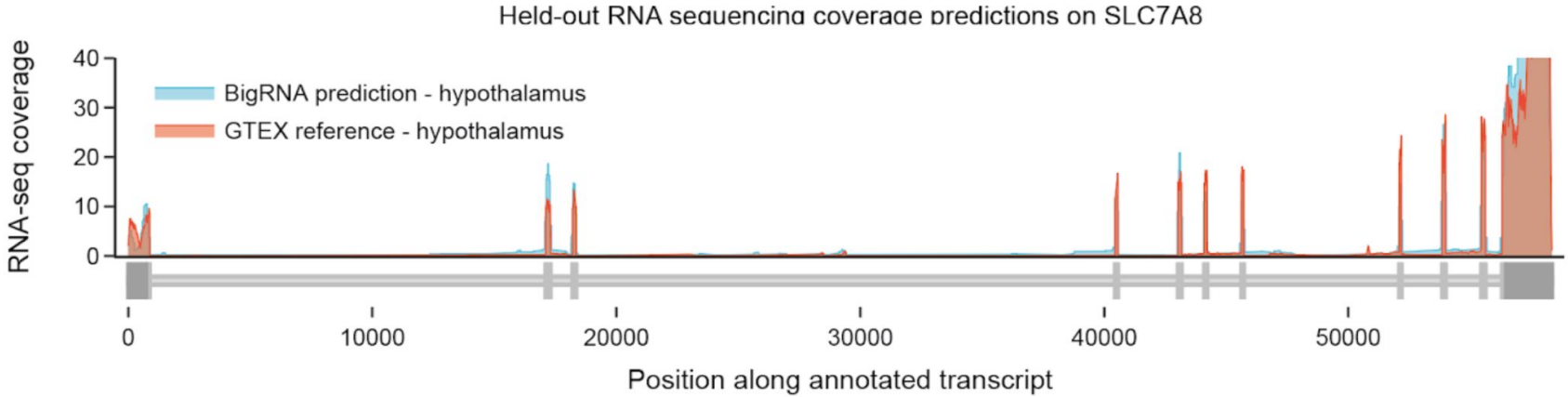
miRNA binding
Total: 18 AGO binding tracks

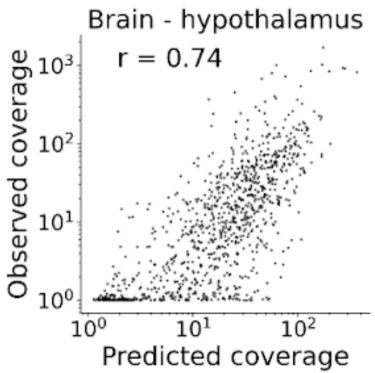
RNA methylation
Total: 23 m6A modification tracks

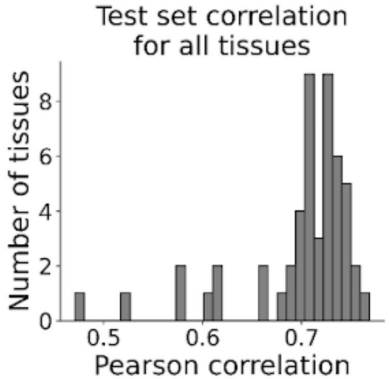


Total: 3901 labels x 3 billion DNA base pairs

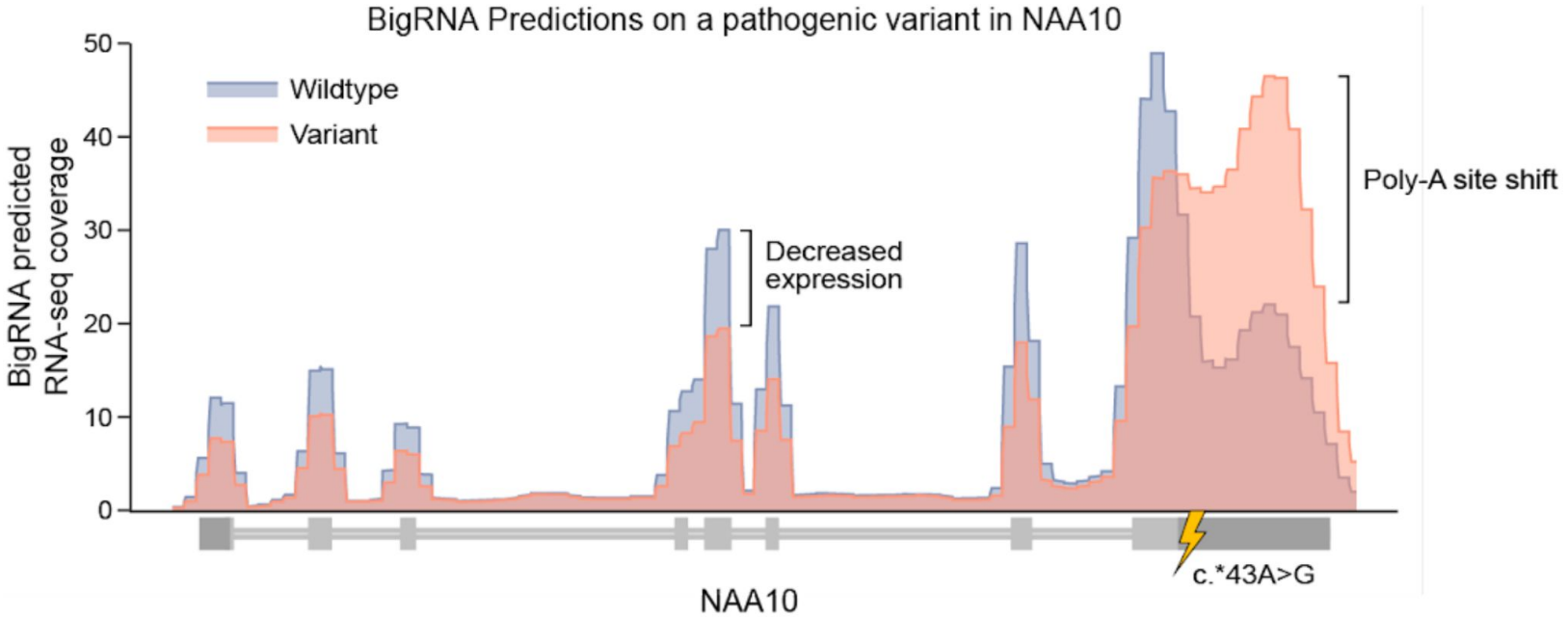
Predicting mRNA expression







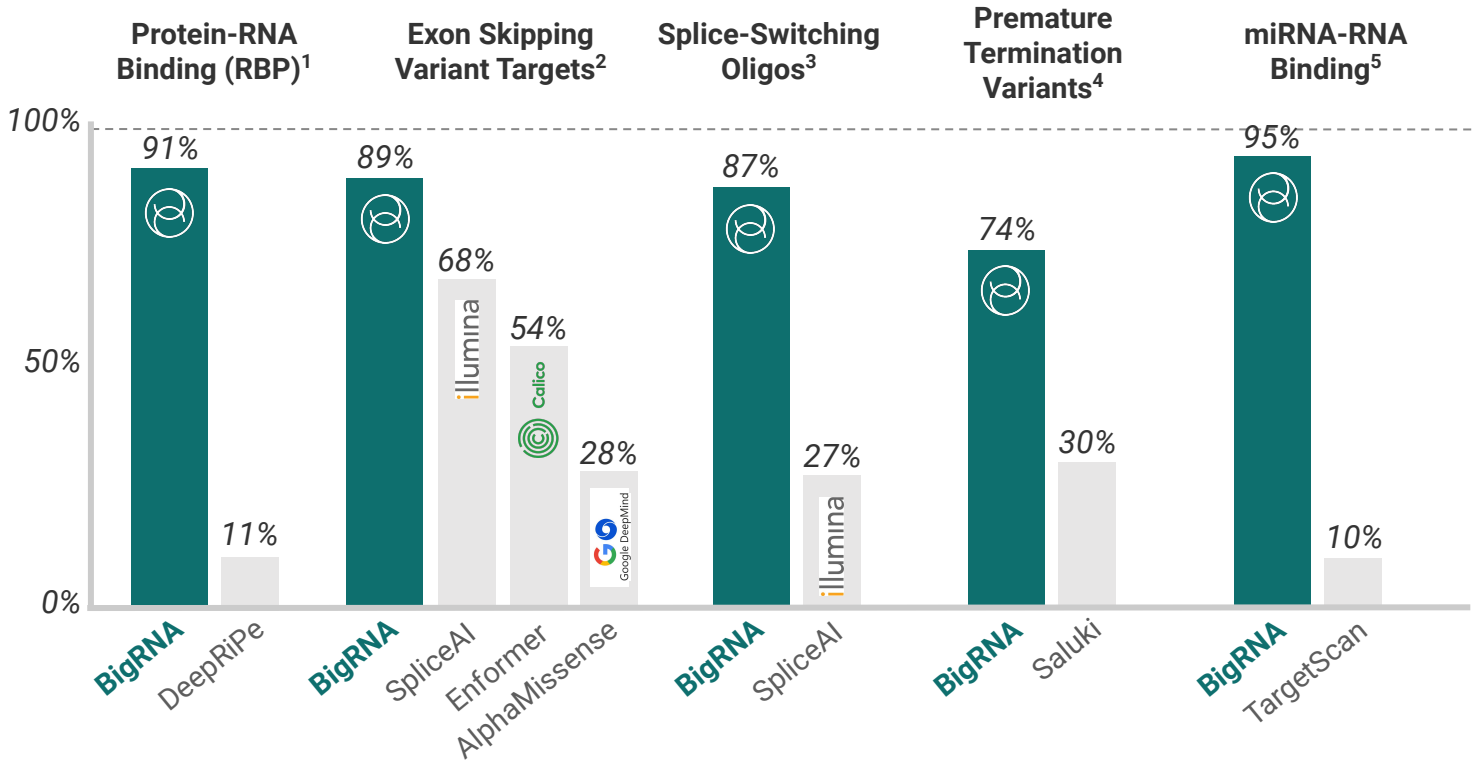
Predicting polyA-disrupting genetic variants



BigRNA outperforms state-of-the-art models across a range of tasks

Accuracy on Held Out Test Data (see notes)

1. Fraction of 141 proteins for which average precision for binding prediction >= 0.1.
2. True positive rate at FPR=0.15 for classifying MaPSy variants (341 positives, 754 negatives).
3. Fraction of 15 exons (in 12 genes) for which predicted exon inclusion of oligos correlate with measured values by >= 0.45.
4. Fraction of premature termination codon variants classified correctly at a 10% FPR vs common 5′ UTR SNVs.
5. Fraction of 19 cell models for which AUC >= 0.75 for identification of CLIP miRNA binding sites.



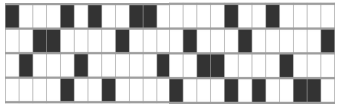
For more, see: https://www.biorxiv.org/content/10.1101/2023.09.20.558508v1

BigRNA acts as a zero-shot model of disease-causing mutations and RNA-blocking “oligo” drugs

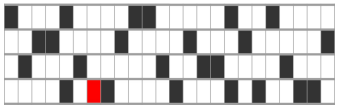
deep genomics

Model input

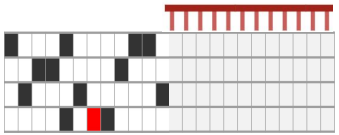
Healthy sequence



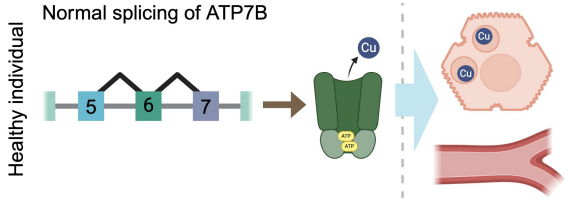
Disease sequence

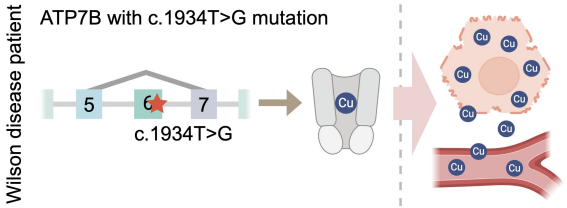


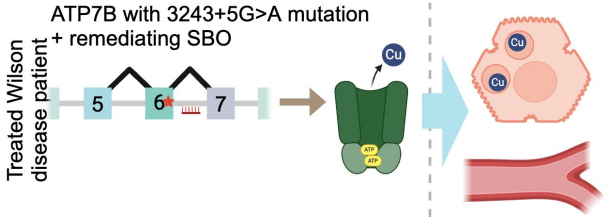
Disease sequence + RNA-blocking oligo



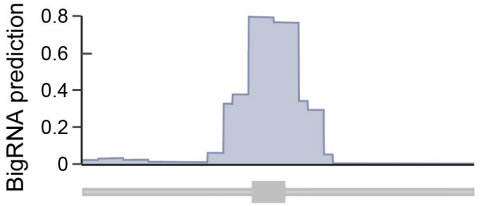
Expected outcome

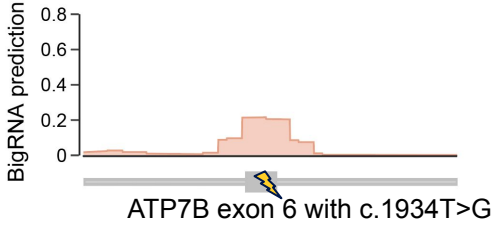


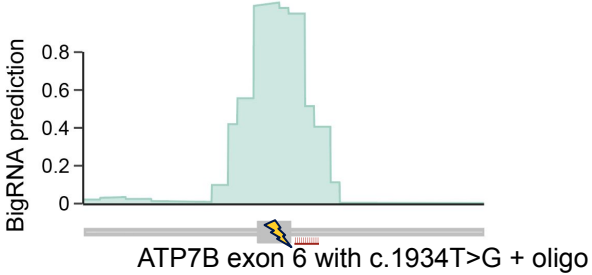




Model Output

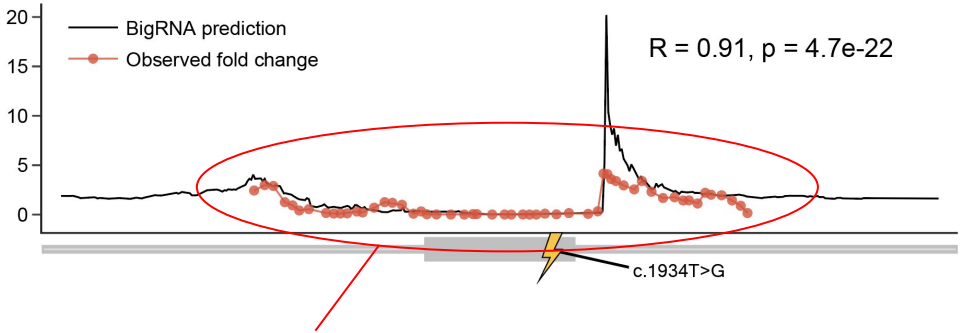






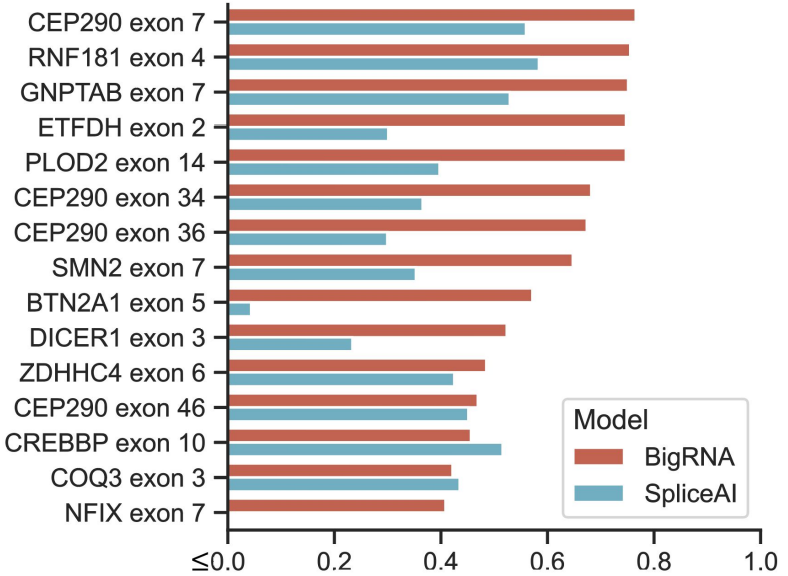
BigRNA predicts the top compound, and correlates remarkably well with multiple experiments

Predicted versus observed RNA-blocking drug effects on ATP7B exon 6



BigRNA prediction achieves strong correlation with experimental measurements

Predicted versus observed RNA-blocking drug effects in fifteen different experiments

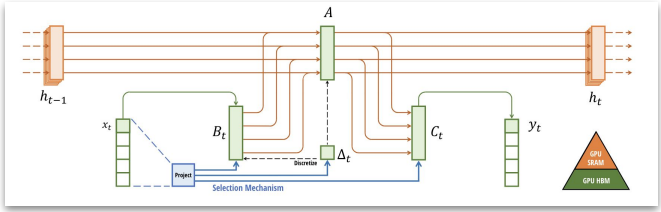


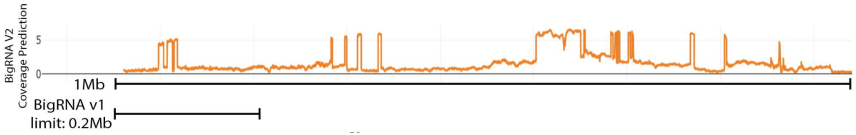
Correlation with experimentally-observed RNA-blocking drug effect

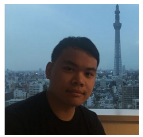
Latest developments: Newest version of BigRNA is high resolution and high context

From transformer ensemble to structured state space architecture (mamba)

- Flexible input size **allows predictions for all genes**
- **Enables scaling** via higher memory efficiency
- Increased accuracy for **variant predictions** and **therapeutics**



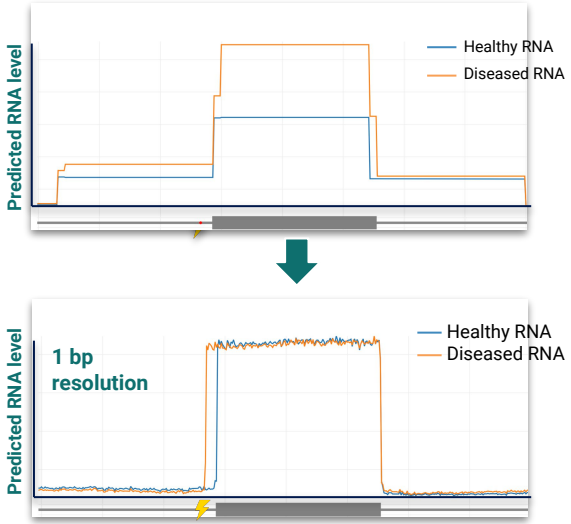




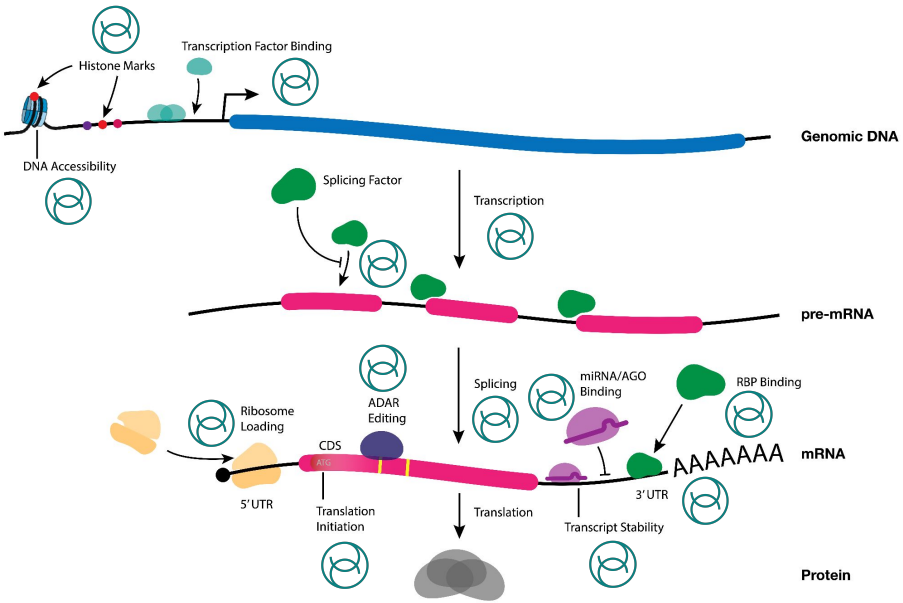
Alston Lo

Increase resolution over many other models to 1-bp

- Contrast Enformer (128 bp), Borzoi (32 bp)
- Detects short RNA **changes that are drivers of disease**



PAH exon 11




Sara Pour


Andrew Jung


Vivian Chu


Rory Gao


Alston Lo


Nicole Zhang

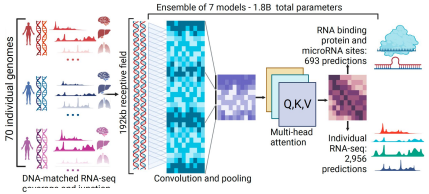

Junru Lin


Helen Zhu

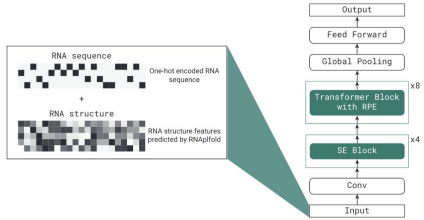
deep genomics

Deep Genomics Models

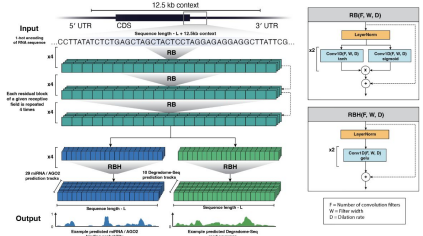
BigRNA
Foundation model of RNA biology



DeepADAR
Deep learning for RNA editing oligo design



REPRESS
Deep learning for repressive RNA elements



What we're working on

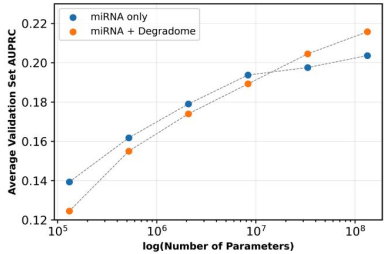
deep genomics

Fine-grained predictions of individual cells

Single-cell data, individual cell types and cell states

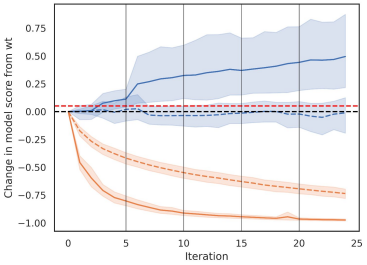
Improving accuracy and understanding its limits

Type and quantity of training data, e.g. multiple species and learning from human variation, improvements to model architecture and training objective



Beyond zero-shot learning: fine-tuning on drug screening data

Impressive zero-shot performance does not take advantage of internal data




Thank you!
Presented by Albi Celaj for Deep Genomics