JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING
Department of Computer Science

2025 INAUGURAL
High-Performance Computing Symposium

JOHNS HOPKINS UNIVERSITY
CENTER FOR COMPUTATIONAL BIOLOGY
CCB

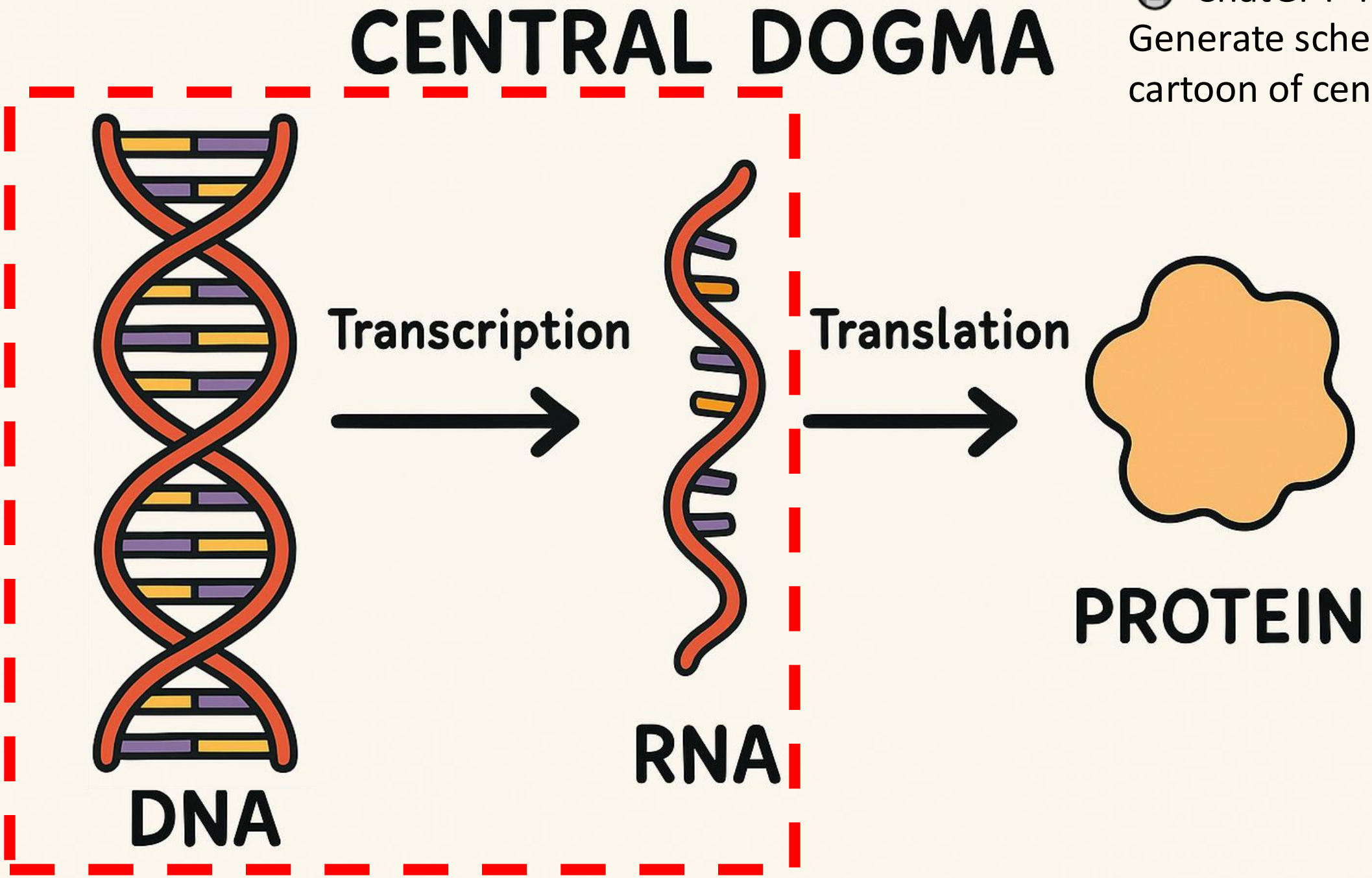
Teaching machines to learn biology:
splice site prediction and gene expression prediction

Kuan-Hao Chao

2025.04.02

khchao.com
@kuanhaochao.bsky.social
@KuanHaoChao
Kuanhao-Chao

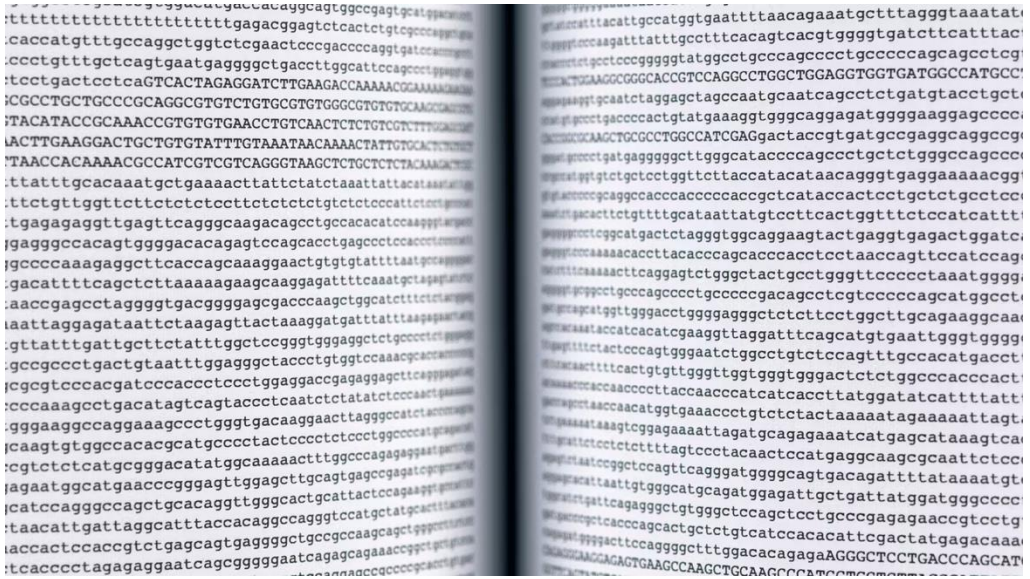
ChatGPT 4o:
Generate schematic cartoon of central dogma

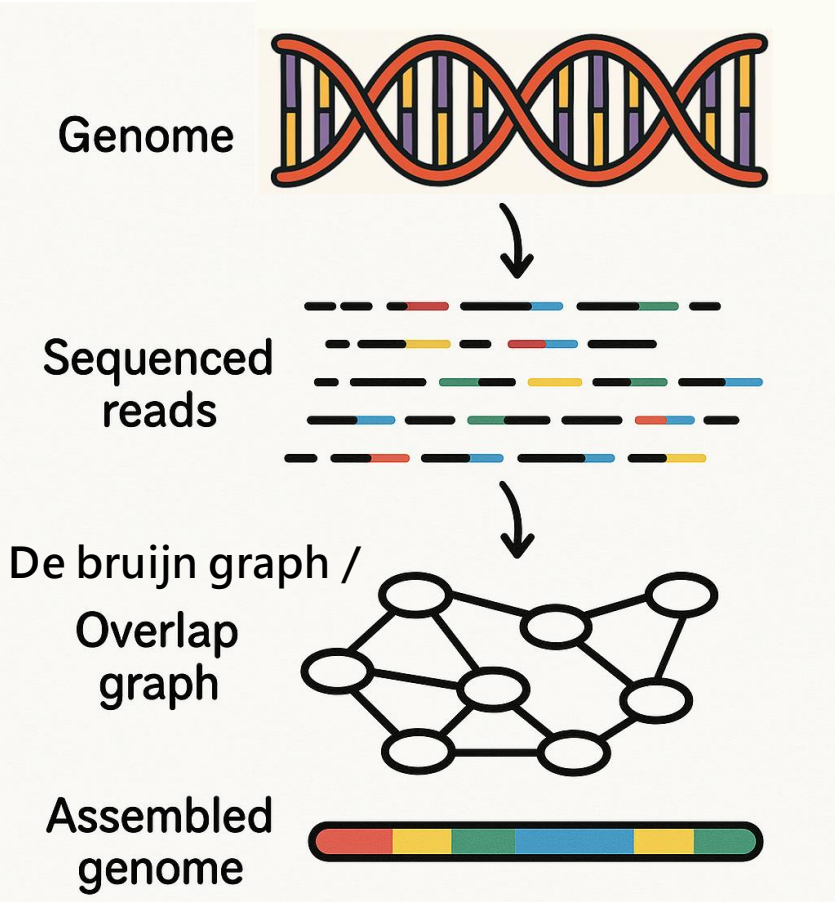


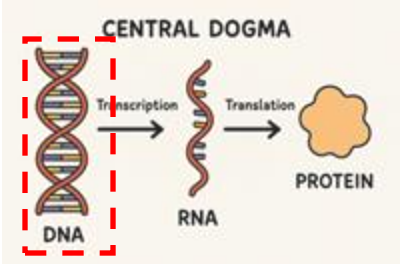
Genome Assembly

Wellcome genome bookcase

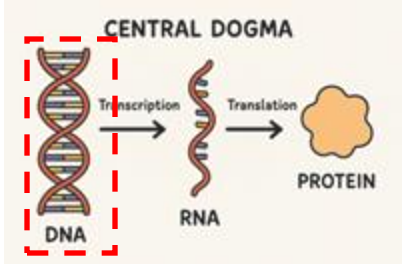
3 billions of ACGT!

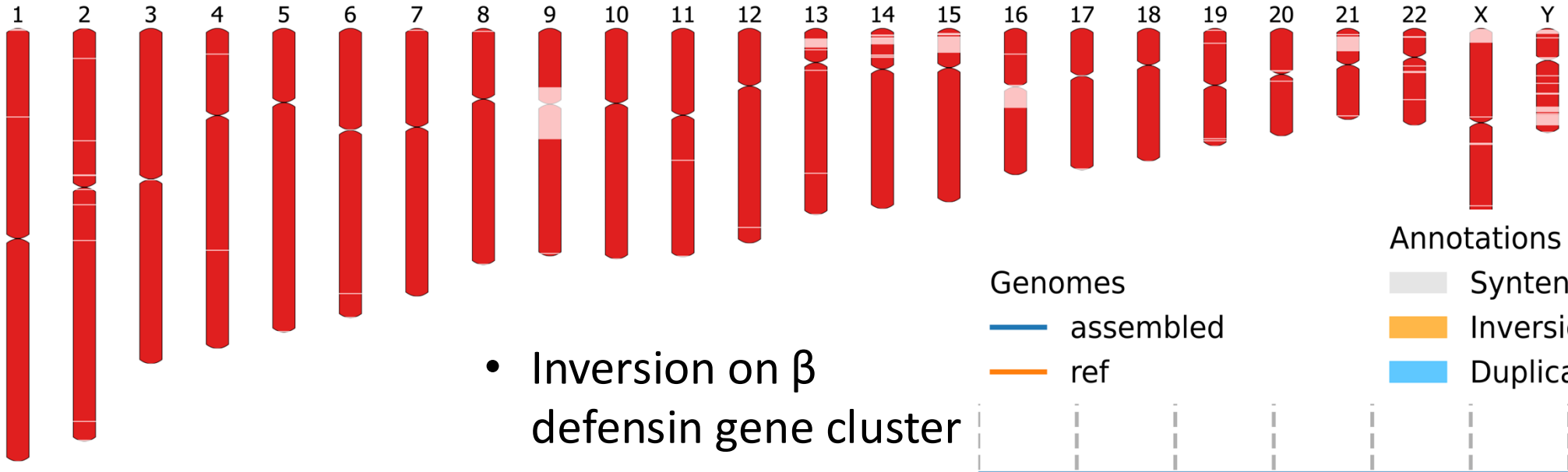




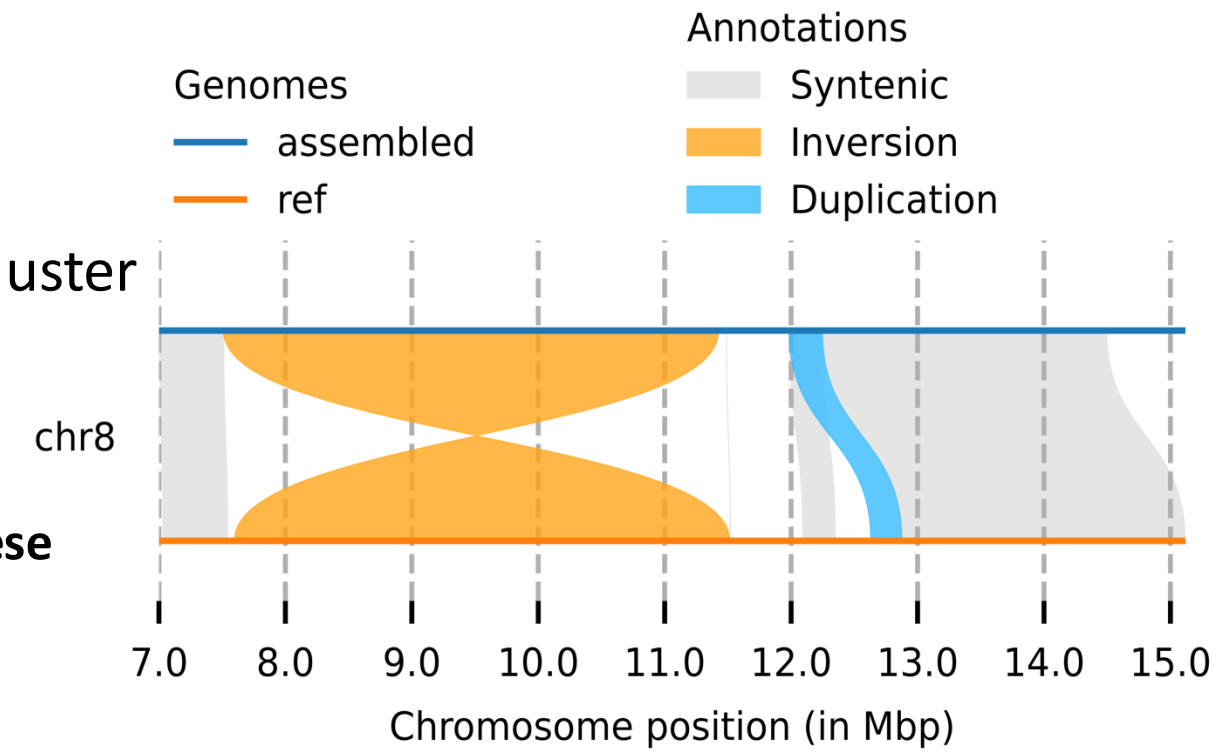


Genome Assembly: Han1





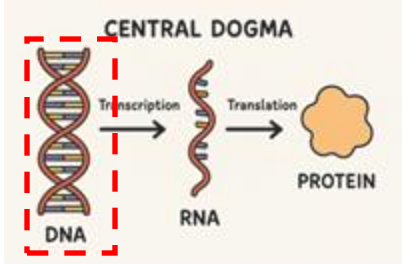
- Inversion on β defensin gene cluster



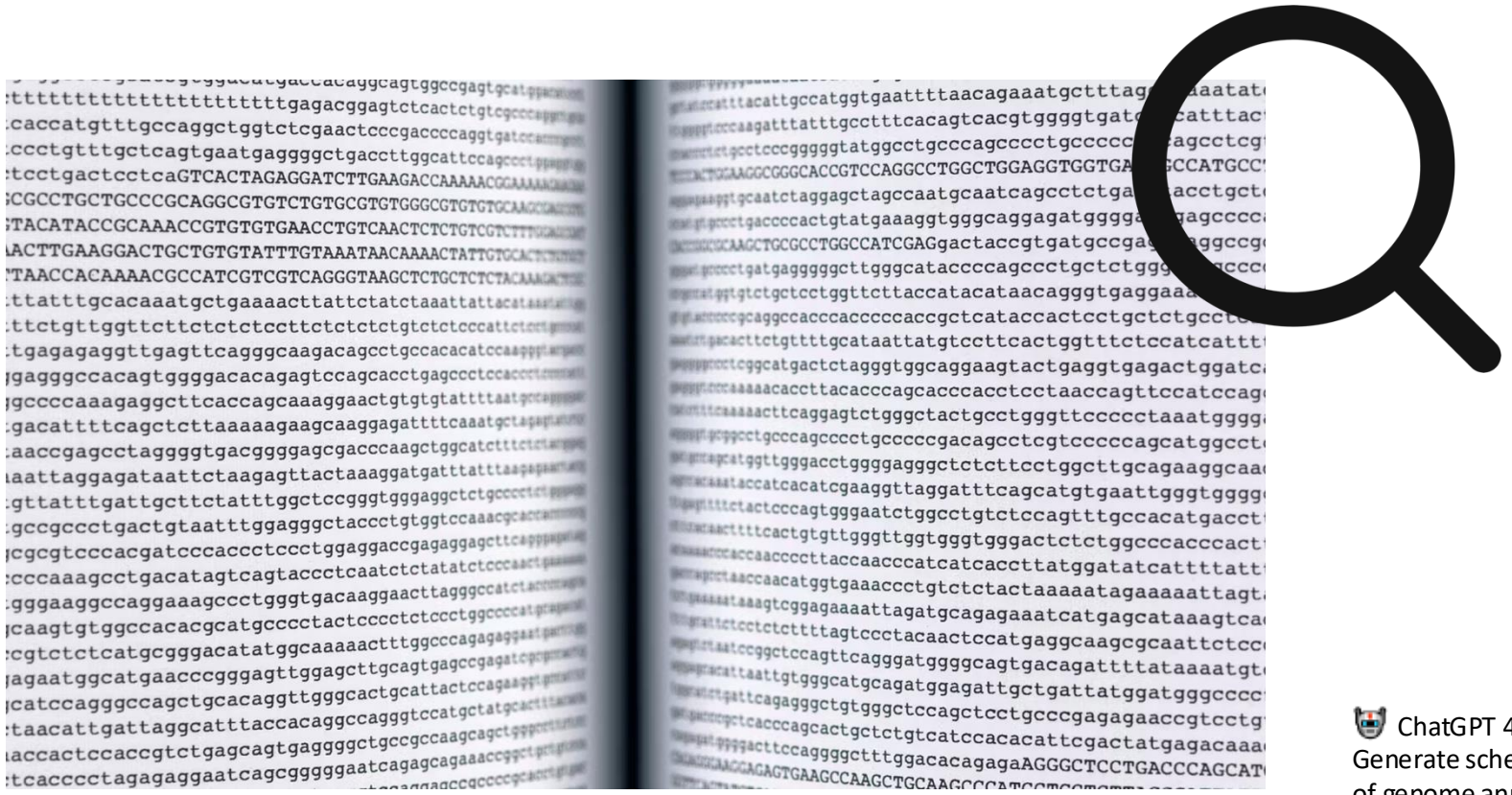
Han1: 1st Gapless Southern Han Chinese

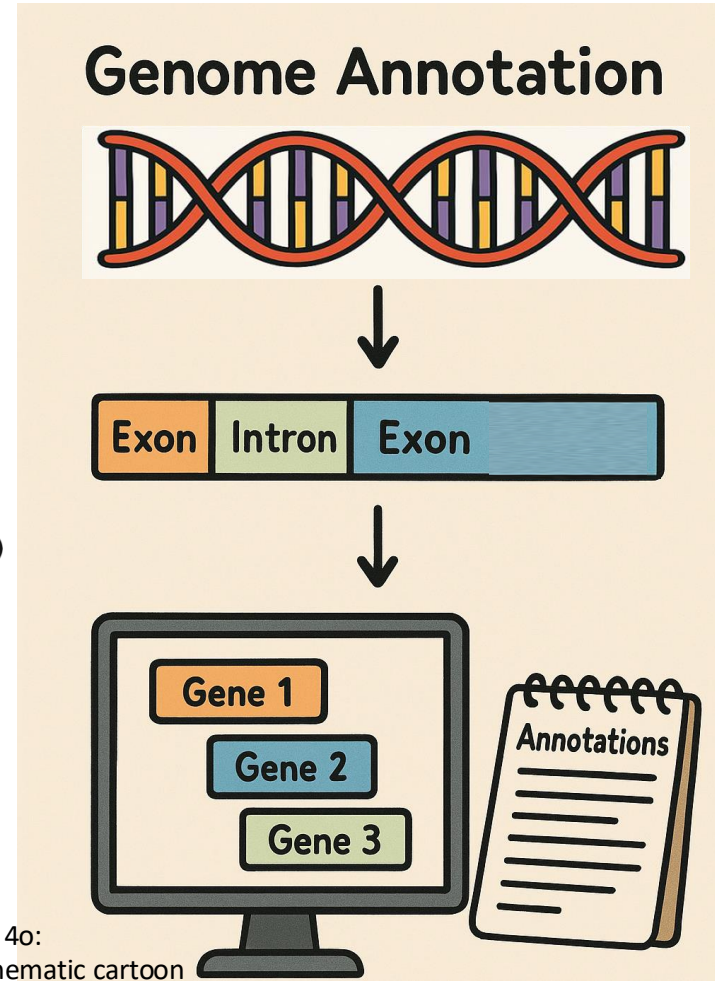
Chao, K. H., Zimin, A. V., Pertea, M., & Salzberg, S. L. (2023). The first gapless, reference-quality, fully annotated genome from a Southern Han Chinese individual. G3: Genes, Genomes, Genetics, 13(3), jkac321.

Genome Annotation



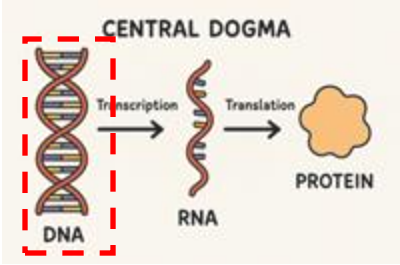
Where are the meaningful regions (Genes)?





ChatGPT 4o:
Generate schematic cartoon of genome annotation

Genome Annotation



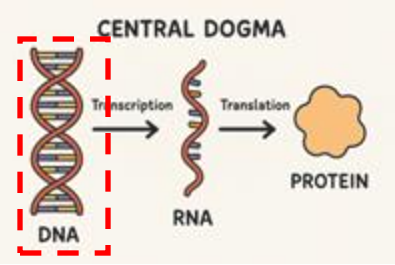
Genome (FASTA)

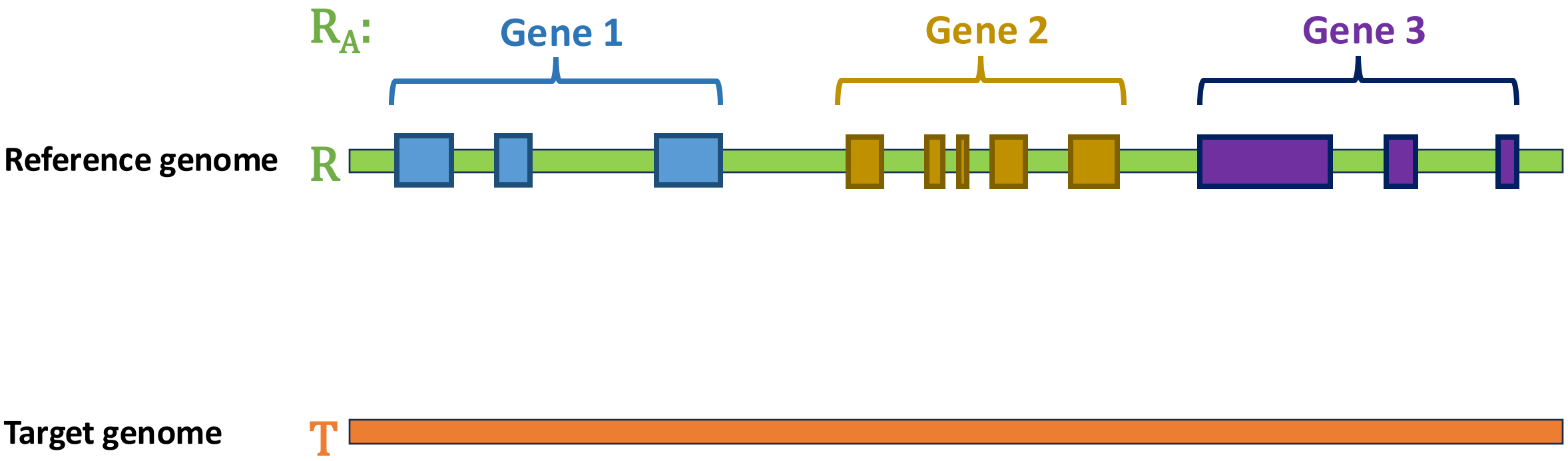
```
CAGCCCCCGGAGACTtaaatacaggaagaaaaaggCAGGACAGAATTACAAGGTGCTGGCCCAGGGCGGGCAGCGGCCCT
GCCTCCTACCCTTGCGCCTCATGACCAGCTTGTTGAAGAGATCCGACATCAAGTGCCCACCTTGGCTCGTGGCTCTCACT
GCAACGGGAAAGCCACAGACTGGGGTGAAGAGTTCAGTCACATGCGACCGGTgactccctgtccccacccccatgACACT
CCCCAGCCCTCCAAGGCCACTGTGTTTCCCAGTTAGCTCAGAGCCTCAGTCGATCCCTGACCCAGCACCGGGCACTGATG
AGACAGCGGCTGTTTGAGGagccacctcccagccacctcggggccagggccagggtgtGCAGCACCACTGTACAATGGGG
AAACTGGCCCAGAGAGGTGAGGCAGCTTGCCTGGGGTCACAGAGCAAGGCAAAAGCAGCGCTGGGTACAAGCTCAAAACC
ATAGTGCCCAGGGCACTGCCGCTGCAGGCGCAGGCATCGCATCACACCAGTGTCTGCGTTCACAGCAGGCATCATCAGTA
```

Annotation (GFF / GTF)

```
chr1  BestRefSeq  gene  450740 451678 .  -  .  ID=gene-OR4F29;
chr1  BestRefSeq  mRNA  450740 451678 .  -  .  ID=rna-NM_001005221.2;Parent=gene-OR4F29;
chr1  BestRefSeq  exon  450740 451678 .  -  .  ID=exon-NM_001005221.2-1;Parent=rna-NM_001005221.2;
chr1  BestRefSeq  exon  452658 453675 .  -  .  ID=exon-NM_001005221.2-2;Parent=rna-NM_001005221.2;
chr1  BestRefSeq  exon  454672 459678 .  -  .  ID=exon-NM_001005221.2-3;Parent=rna-NM_001005221.2;
chr1  BestRefSeq  CDS  450740 451678 .  -  0  ID=cds-NP_001005221.2-1;Parent=rna-NM_001005221.2;
chr1  BestRefSeq  CDS  452658 453675 .  -  0  ID=cds-NP_001005221.2-2;Parent=rna-NM_001005221.2;
```

Genome Annotation: Lift-over Problem



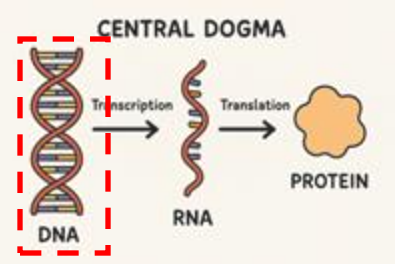


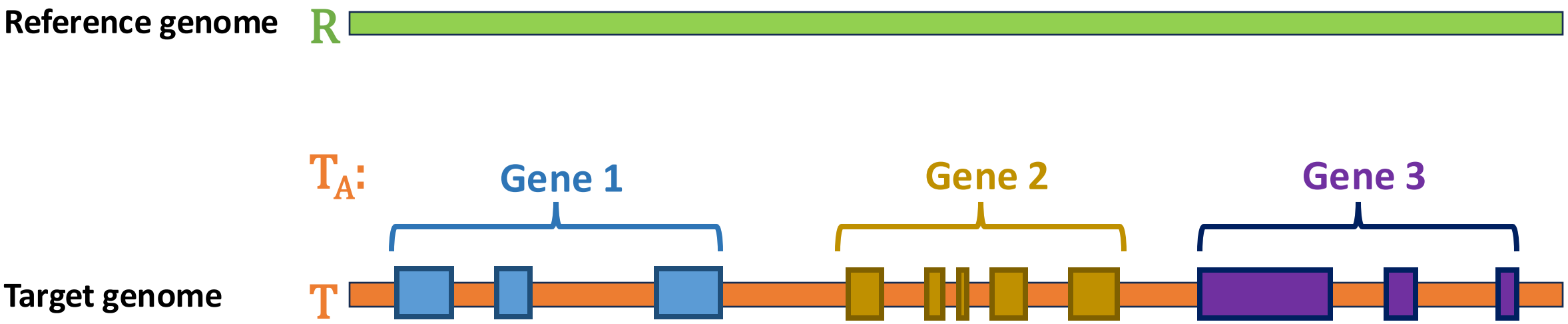
Chao, K. H., Heinz, J. M., Hoh, C., Mao, A., Shumate, A., Pertea, M., & Salzberg, S. L. (2025). Combining DNA and protein alignments to improve genome annotation with LiftOn. Genome Research, 35(2), 311-325.

GENOME RESEARCH

LiftOn

Genome Annotation: Lift-over Problem

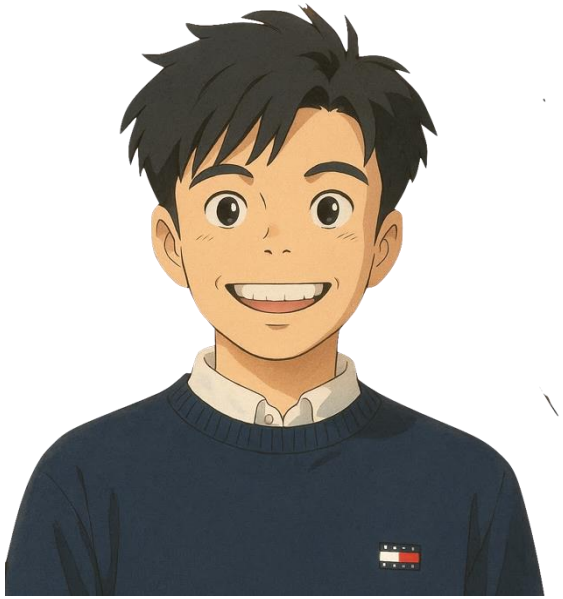




Chao, K. H., Heinz, J. M., Hoh, C., Mao, A., Shumate, A., Pertea, M., & Salzberg, S. L. (2025). Combining DNA and protein alignments to improve genome annotation with LiftOn. Genome Research, 35(2), 311-325.

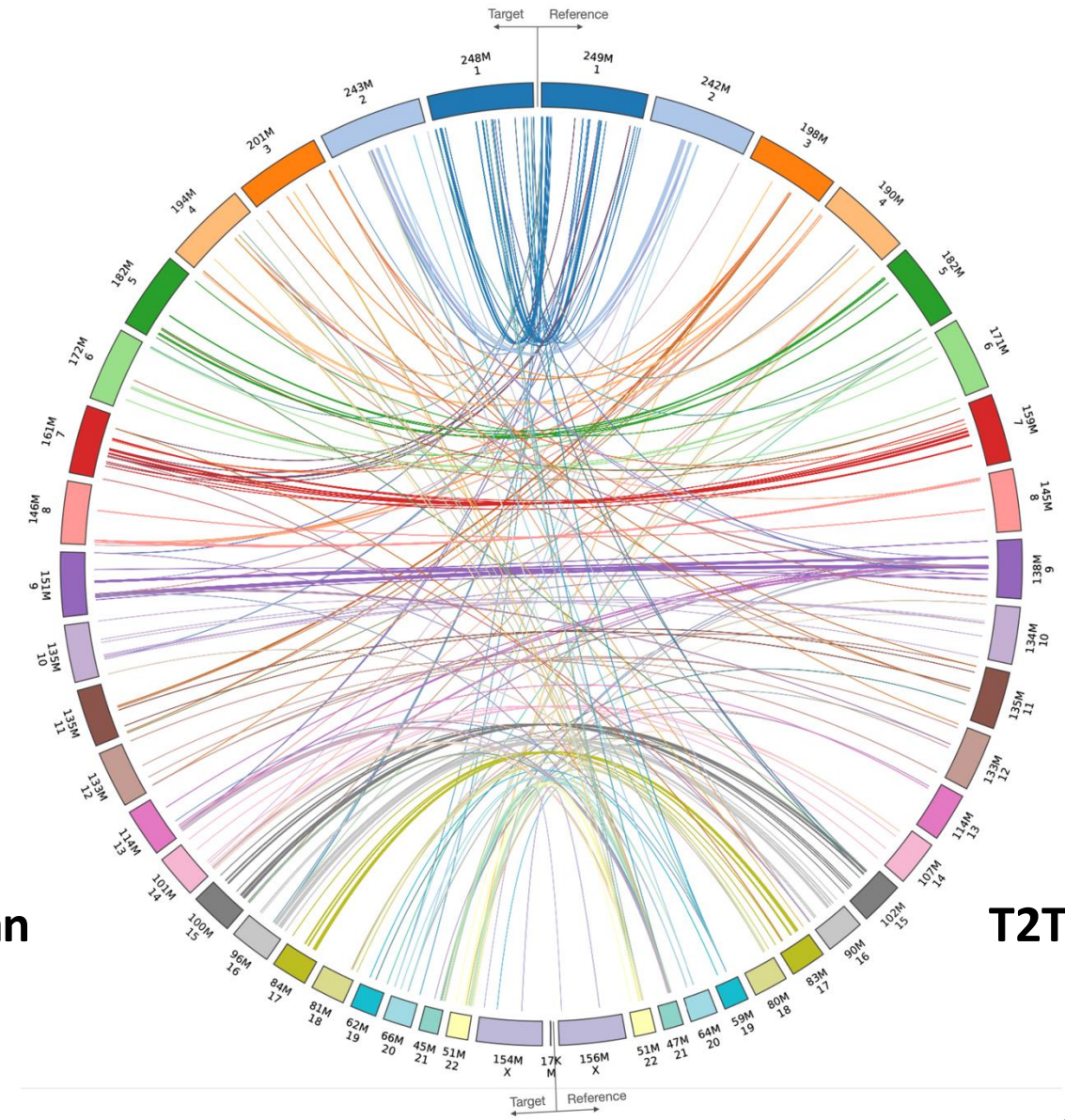
GENOME RESEARCH

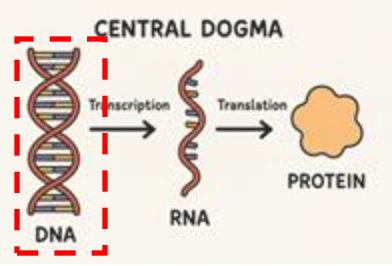
LiftOn



Han1 – Southern Chinese Han Individual

ChatGPT 4o:
Generate my profile photo in Ghibli style.





T2T-CHM13 – Northern European Individual

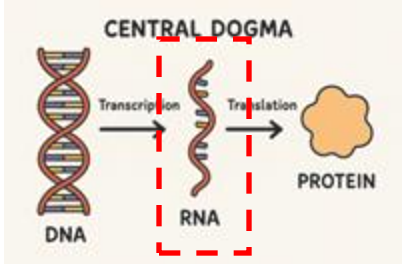
ChatGPT 4o:
Generate a European origin individual icon in Ghibli style.

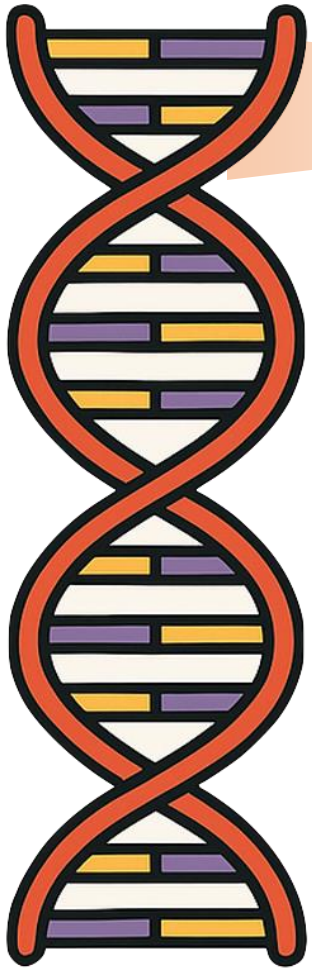
Chao, K. H., Heinz, J. M., Hoh, C., Mao, A., Shumate, A., Pertea, M., & Salzberg, S. L. (2025). Combining DNA and protein alignments to improve genome annotation with LiftOn. Genome Research, 35(2), 311-325.

GENOME RESEARCH

LiftOn

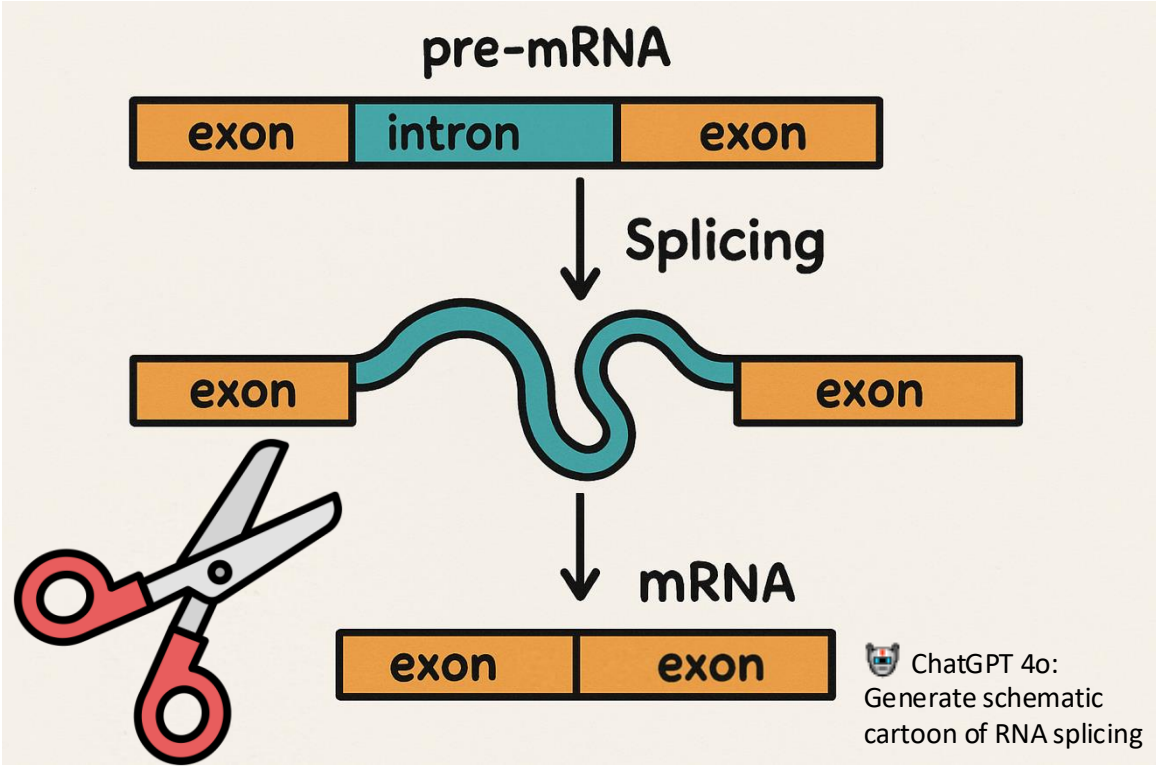
Splice Site Prediction





Gene locus (DNA)

Transcription



ChatGPT 4o: Generate schematic cartoon of RNA splicing

Model

SPLAM

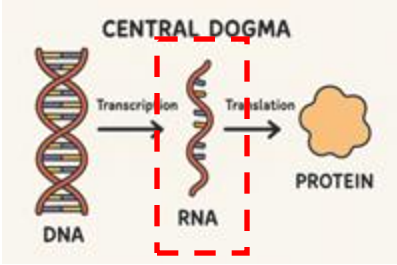
OpenSpliceAI

Where are the splice sites?

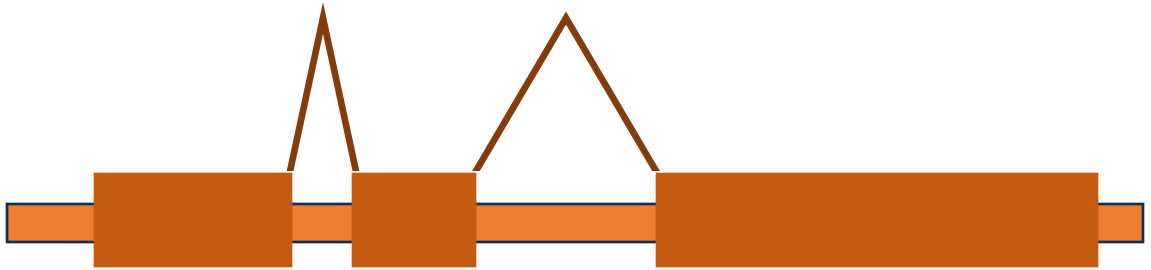
Splice Site Prediction

SPLAM

OpenSpliceAI



Pre-mRNA



Genome Biology

Chao, K. H., Mao, A., Salzberg, S. L., & Pertea, M. (2024). Splam: a deep-learning-based splice site predictor that improves spliced alignments. Genome biology, 25(1), 243.

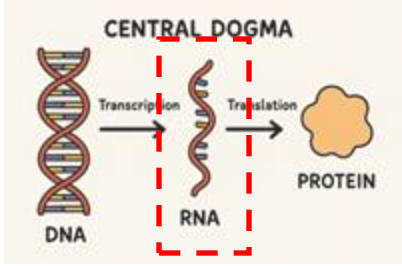
HUB

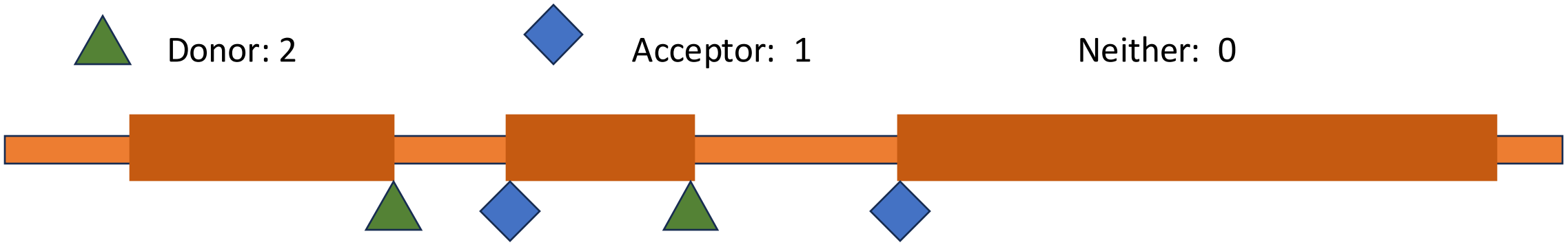
New AI tool pinpoints gene splicing with unmatched precision

bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

Chao, K. H., Mao, A., Liu, A., Salzberg, S. L., & Pertea, M. (2025). OpenSpliceAI: An efficient, modular implementation of SpliceAI enabling easy retraining on non-human species. *bioRxiv*, 2025-03.

Splice Site Prediction





X AGACTCAGCCCCCGGAGACTTAGTTAGAGGAAGAAAAAGGTAGGACAGAAGAAAAAGGCAGGACATACAAGGTGCTGGCCCAGGGCGG

Y 000000000000000000000002000000001000000002000000000001000000000000000000000000000000000

Genome Biology

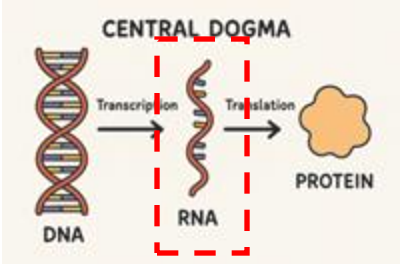
Chao, K. H., Mao, A., Salzberg, S. L., & Pertea, M. (2024). Splam: a deep-learning-based splice site predictor that improves spliced alignments. Genome biology, 25(1), 243.

HUB

New AI tool pinpoints gene splicing with unmatched precision

bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

Chao, K. H., Mao, A., Liu, A., Salzberg, S. L., & Pertea, M. (2025). OpenSpliceAI: An efficient, modular implementation of SpliceAI enabling easy retraining on non-human species. *bioRxiv*, 2025-03.



Splice Site Prediction: data processing

~20k protein-coding genes

Raw gene DNA sequence

Gene 1 — $L = 33200$

Gene 2 — $L = 14600$

...

Gene n — $L = 25000$

X	Y
[7, 15000, 4]	[7, 5000, 3]
[3, 15000, 4]	[3, 5000, 3]
...	
[5, 15000, 4]	[5, 5000, 3]

Genome Biology

Chao, K. H., Mao, A., Salzberg, S. L., & Pertea, M. (2024). Splam: a deep-learning-based splice site predictor that improves spliced alignments. Genome biology, 25(1), 243.

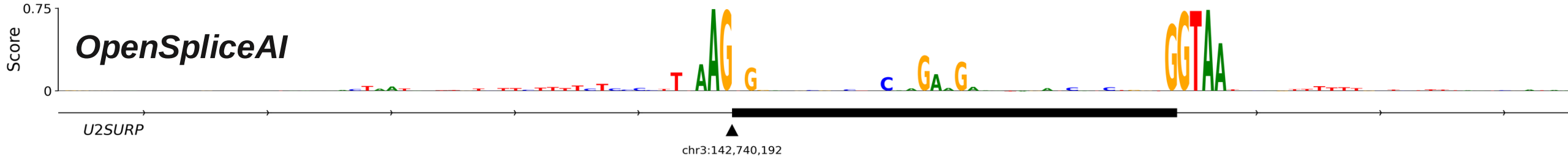
HUB

New AI tool pinpoints gene splicing with unmatched precision

bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

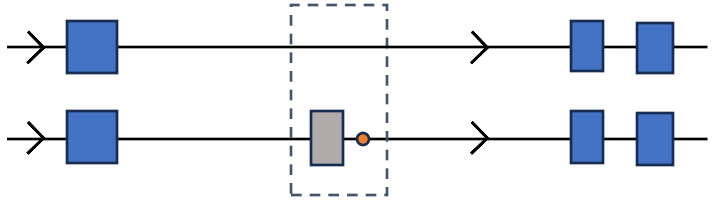
Chao, K. H., Mao, A., Liu, A., Salzberg, S. L., & Pertea, M. (2025). OpenSpliceAI: An efficient, modular implementation of SpliceAI enabling easy retraining on non-human species. *bioRxiv*, 2025-03.

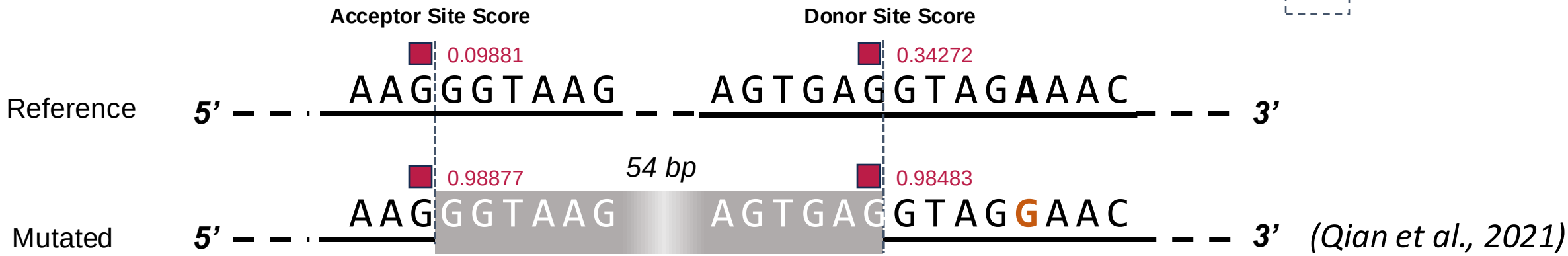
Splice Site Prediction: what did model learn?



OPA1 *(chr3:193644727**A**>**G**)*

OpenSpliceAI Score

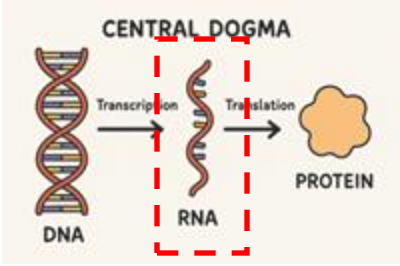




(Qian et al., 2021)

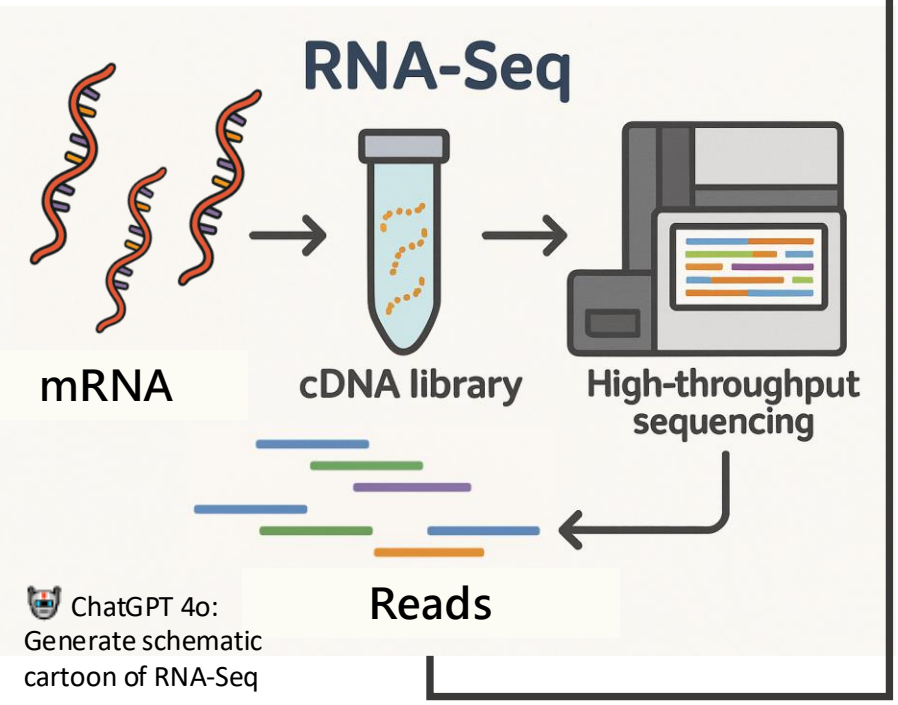
Calico

RNA-Seq Prediction



ChatGPT 4o:
Generate schematic cartoon of RNA-Seq coverage

Gene locus (DNA)



ChatGPT 4o:
Generate schematic cartoon of RNA-Seq

Reads alignment

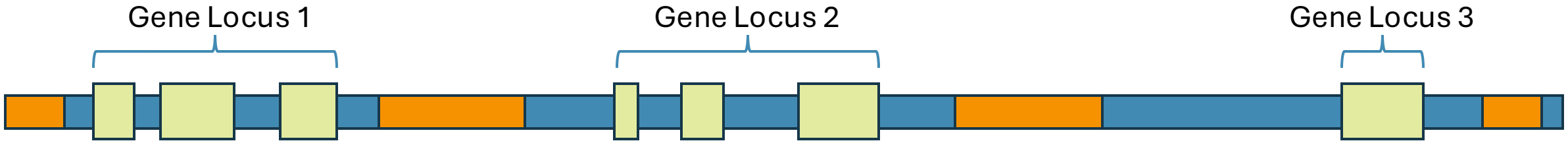
Coverage pileup

Model
Fungal Model

RNA-Seq expression?

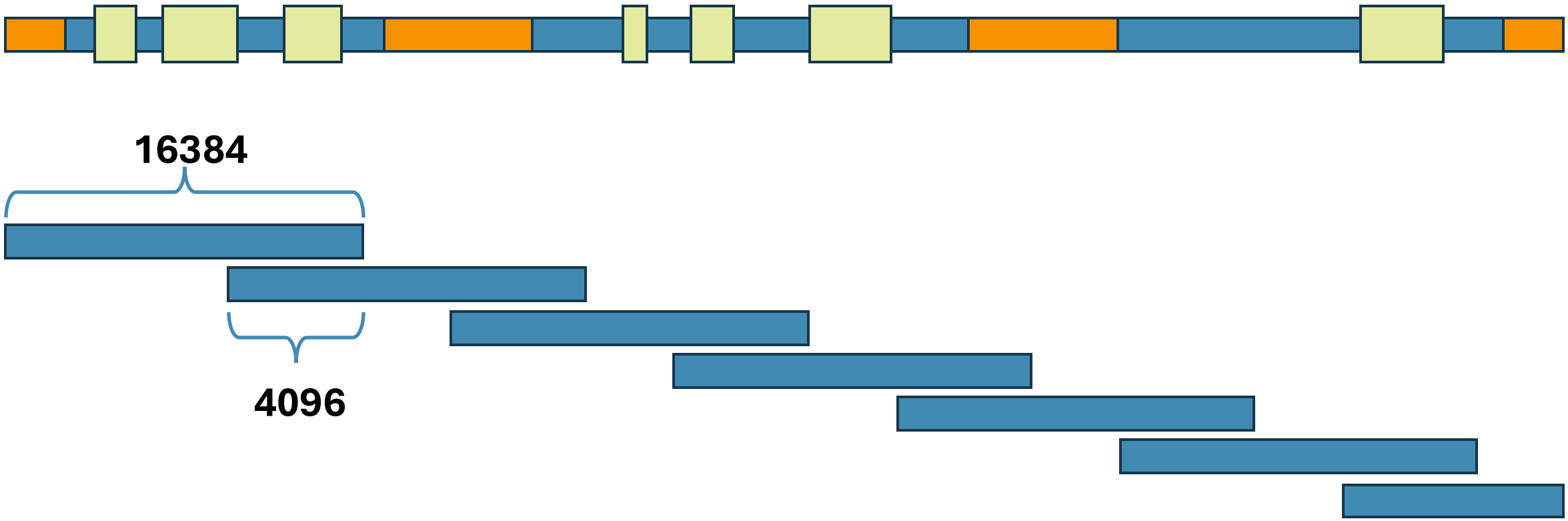
Calico

Processing 1K fungal genomes



Calico

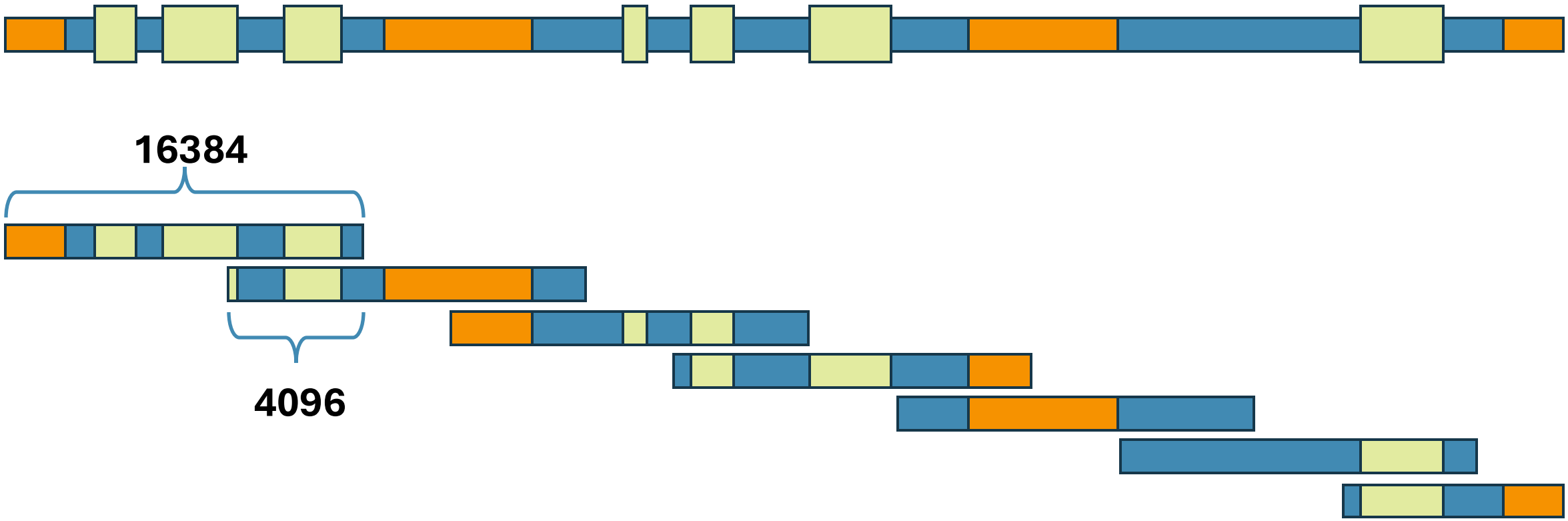
Processing 1K fungal genomes



~ 7 genes per window

Calico

Processing 1K fungal genomes



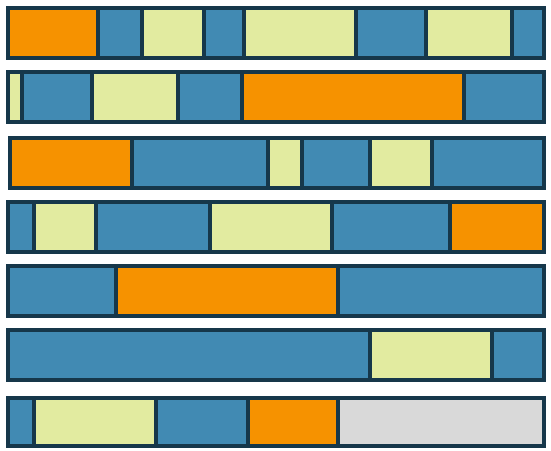
~ 7 genes per window

Calico

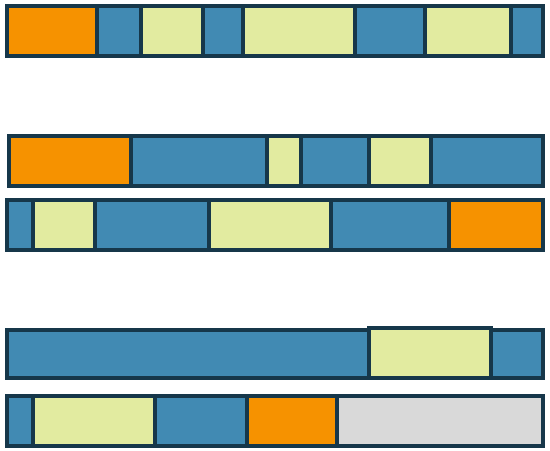
Processing 1K fungal genomes

Repeat regions

Coding regions



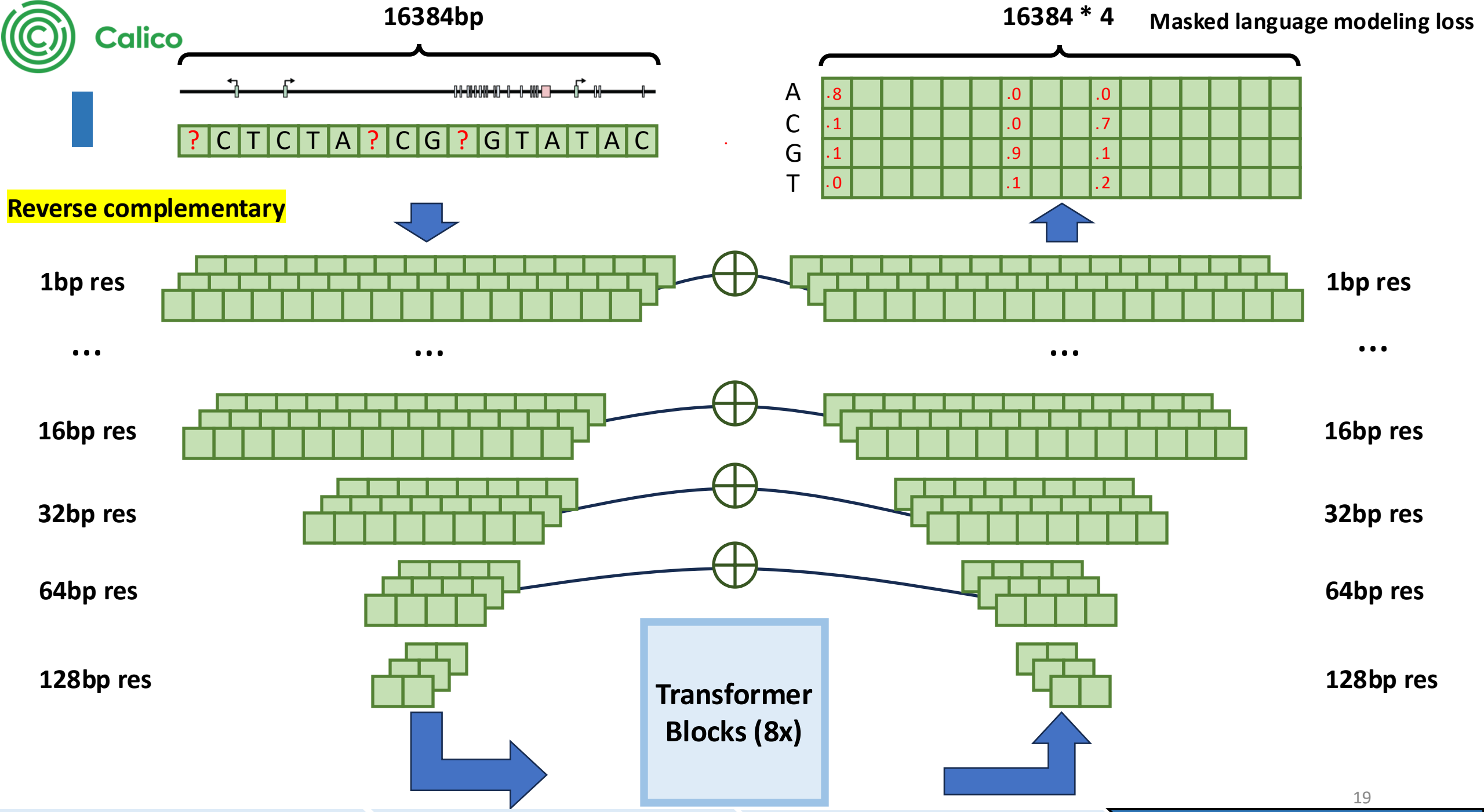
7% repeat threshold



Training

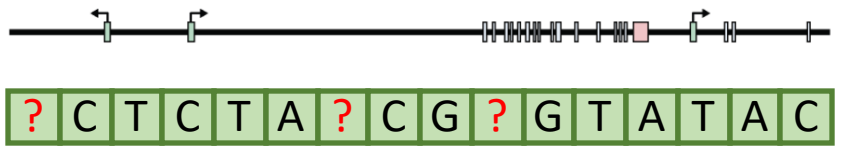
Validation
(chrXI, chrXIII, chrXV)

Testing
(chrXII, chrXIV, chrXVI)


16384bp
? C T C T A ? C G ? G T A T A C
16384 * 4
Masked language modeling loss
A .8 .0 .0
C .1 .0 .7
G .1 .9 .1
T .0 .1 .2
Reverse complementary
1bp res
16bp res
32bp res
64bp res
128bp res
Transformer Blocks (8x)

Calico

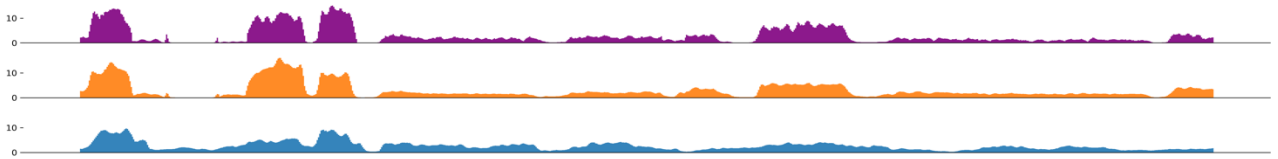
16384bp



?	C	T
C	T	A
?	C	G
?	G	T
A	T	A
C		

[CHiP-exo (1128)
Histone marks (20)
RNA-Seq (1340)]

Coverage Tacks



Reverse complementary

1bp res

...

...

16bp res

32bp res

64bp res

128bp res

Transformer Blocks (8x)

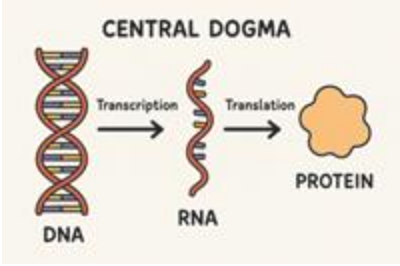
16bp res

32bp res

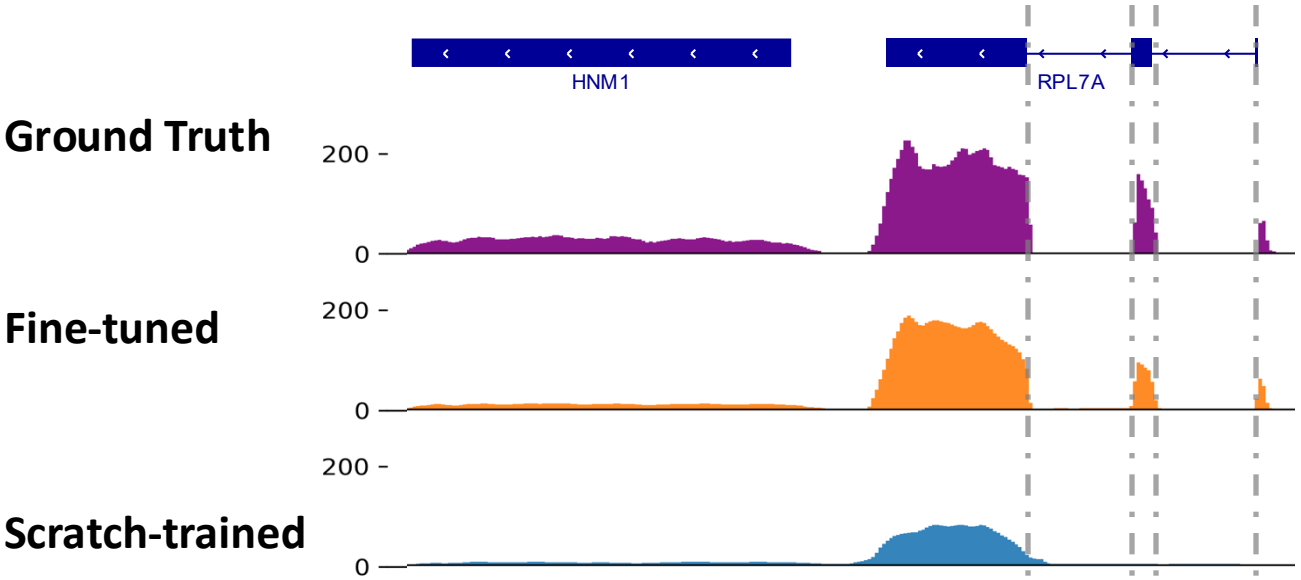
64bp res

128bp res

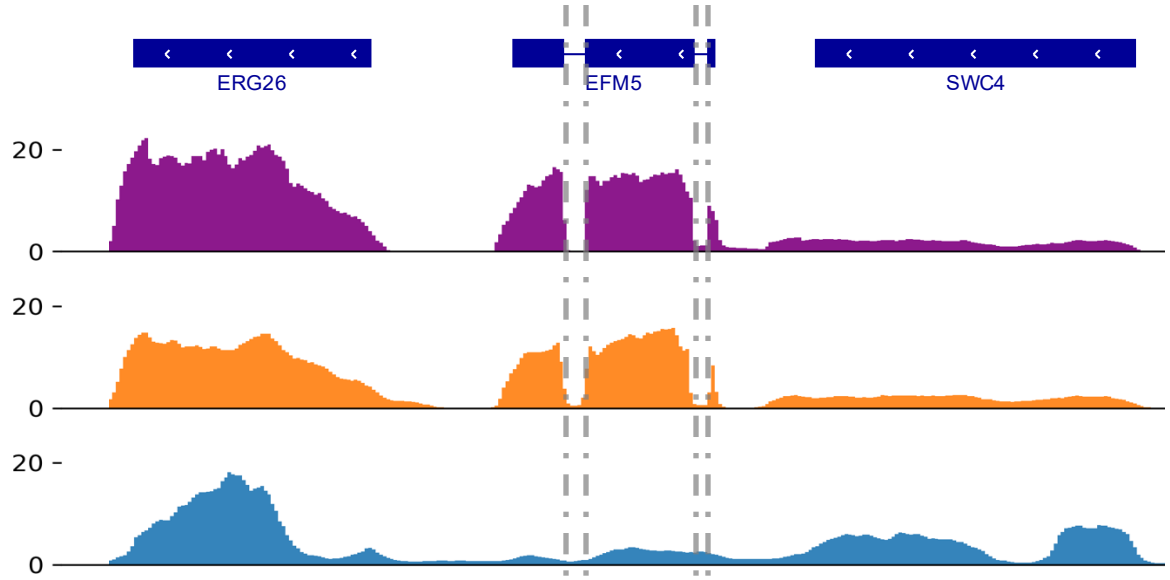
RNA-Seq Coverage Prediction Examples



chrVII:362,180-366,023 (RNA-Seq tracks)

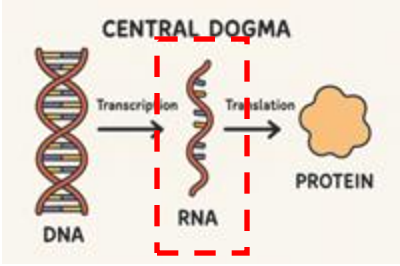


chrVII:495,374-499,965 (RNA-Seq tracks)



Calico

Learned Regulatory Elements



Fine-tuned motif Reconstruction from ISM PWM

Motif DB

RAP1

FHL1

SFP1.1

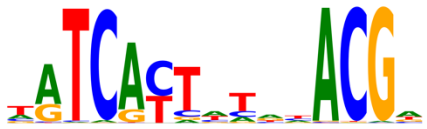
TATA Box

REB1

Fine-tuned motif Reconstruction from ISM PWM

Motif DB

ABF1

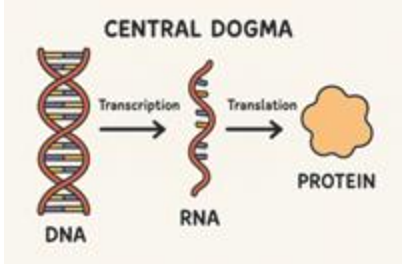


TBF1.1

CBF1

UME6.2

PBF1



Learned Splicing Motifs

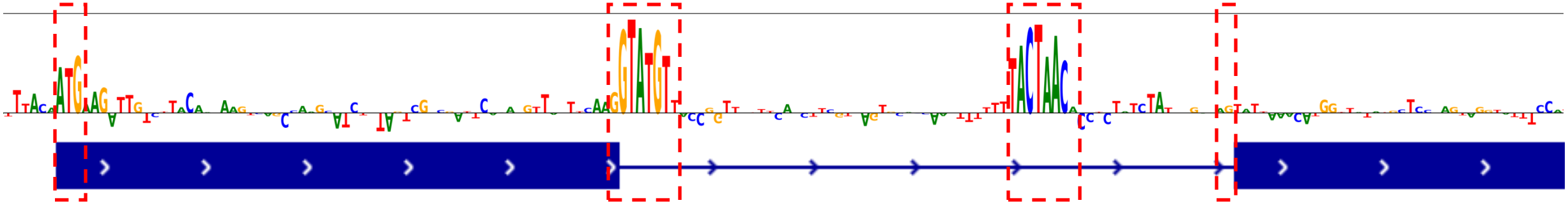
Yeast splicing motifs
(Schirman et. al.)

Start Codon

Splice site
(donor)

Branch point

Splice site
(acceptor)



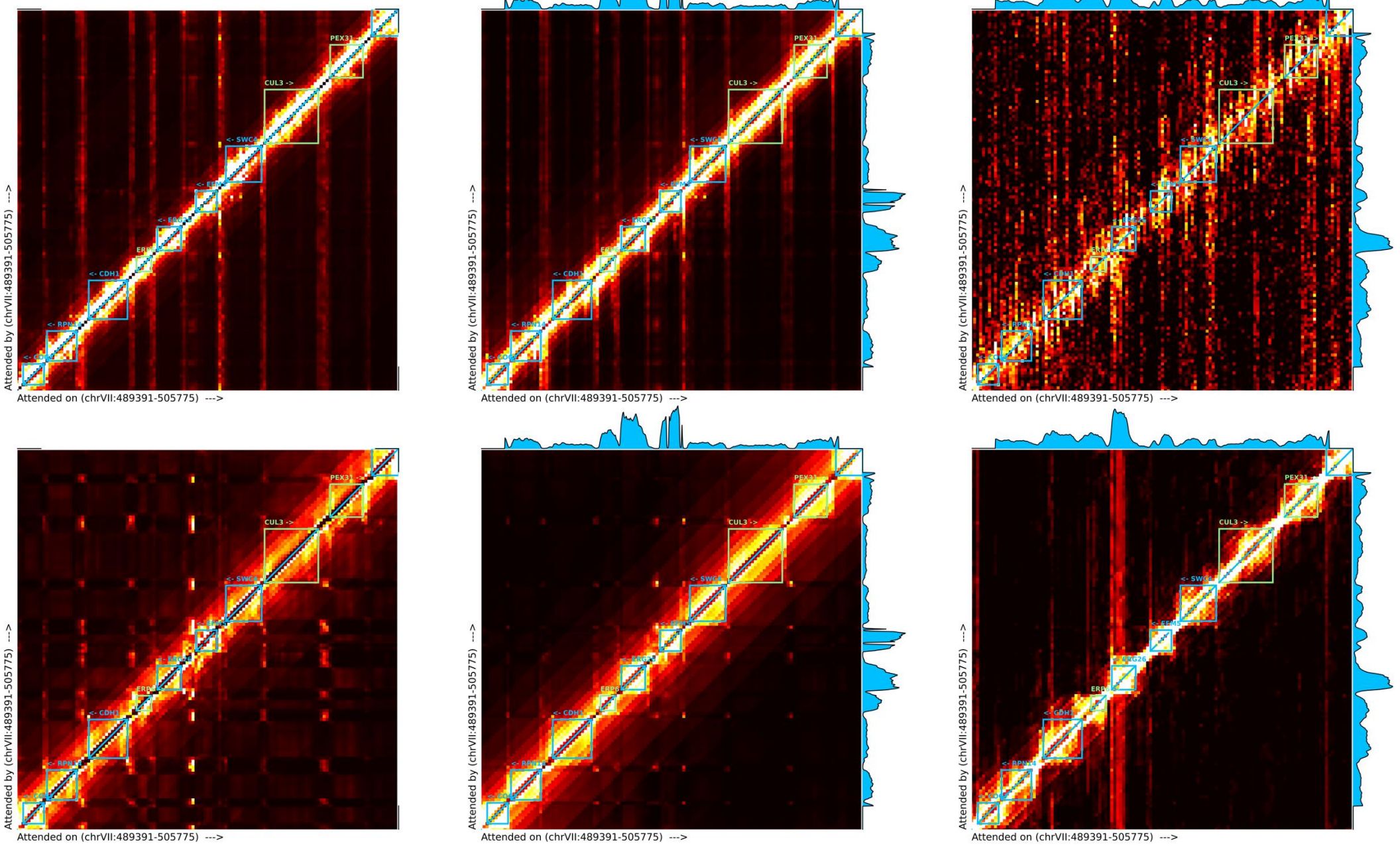
Language model

Fine-tuned RNA-Seq model

Scratch-trained RNA-Seq model

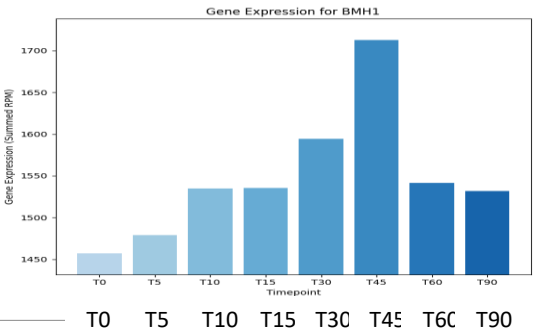
First Self-attention Layer

Last Two Self-attention Layer



MET4 Induction

chrX:322935-323435 (Promoter region of BNA3)



T0 T5 T10 T15 T30 T45 T60 T90

T0

T5

T10

T15

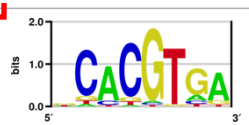
T30

T45

T60

T90

YeTFaSCo DB motif



MET4

“If you think of mathematics as the perfect description language for physics, then AI might be the perfect one for biology.”

Demis Hassabis, CEO of DeepMind, 2022

Acknowledgement