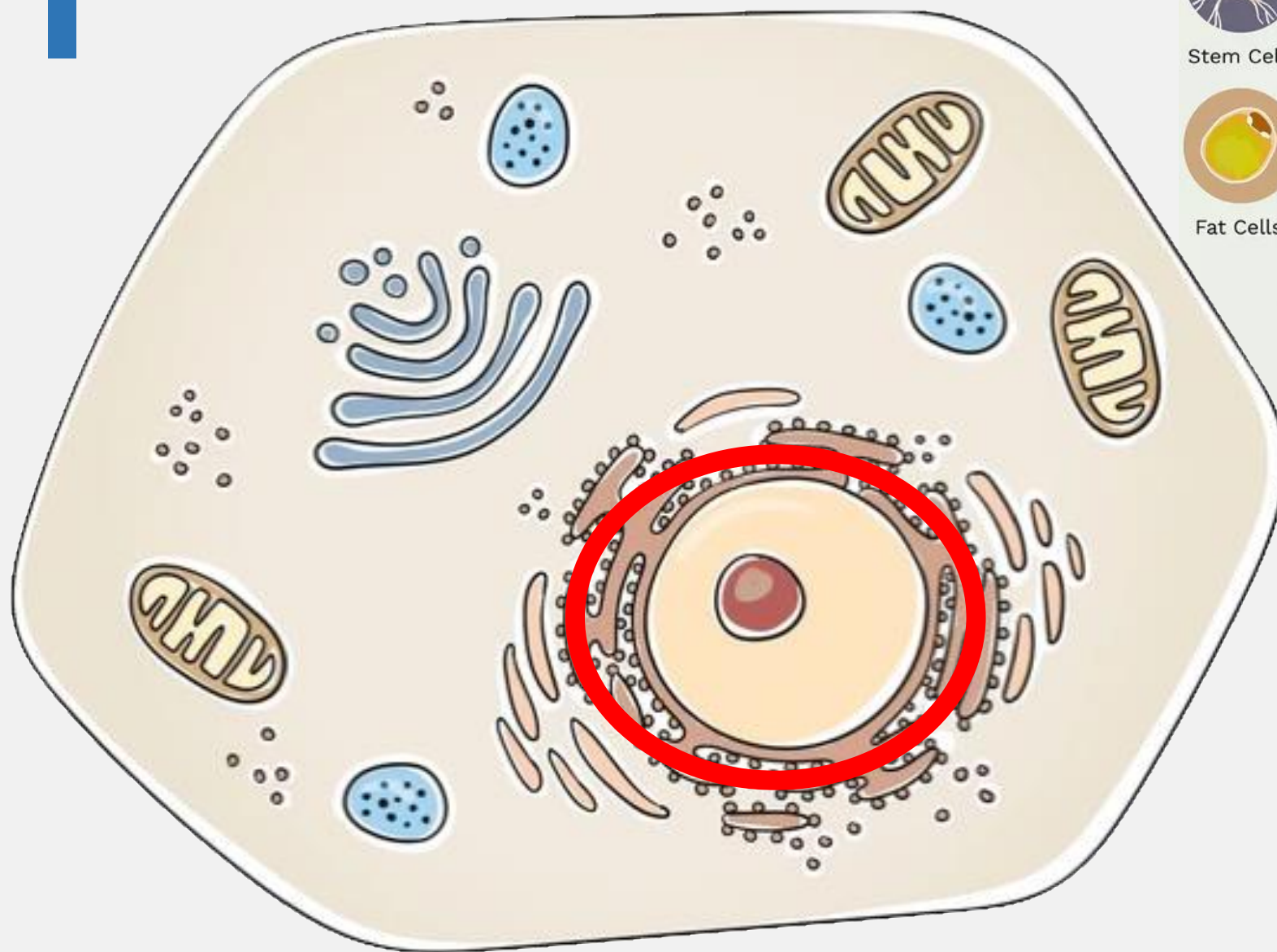




Decoding the Language of Genomes: Bridging Sequences and Function through Deep Learning

Kuan-Hao Chao

2026.07.21



Stem Cells



Bone Cells



Blood Cells



Muscle Cells



Fat Cells



Skin Cells



Nerve Cells



Endothelial Cells



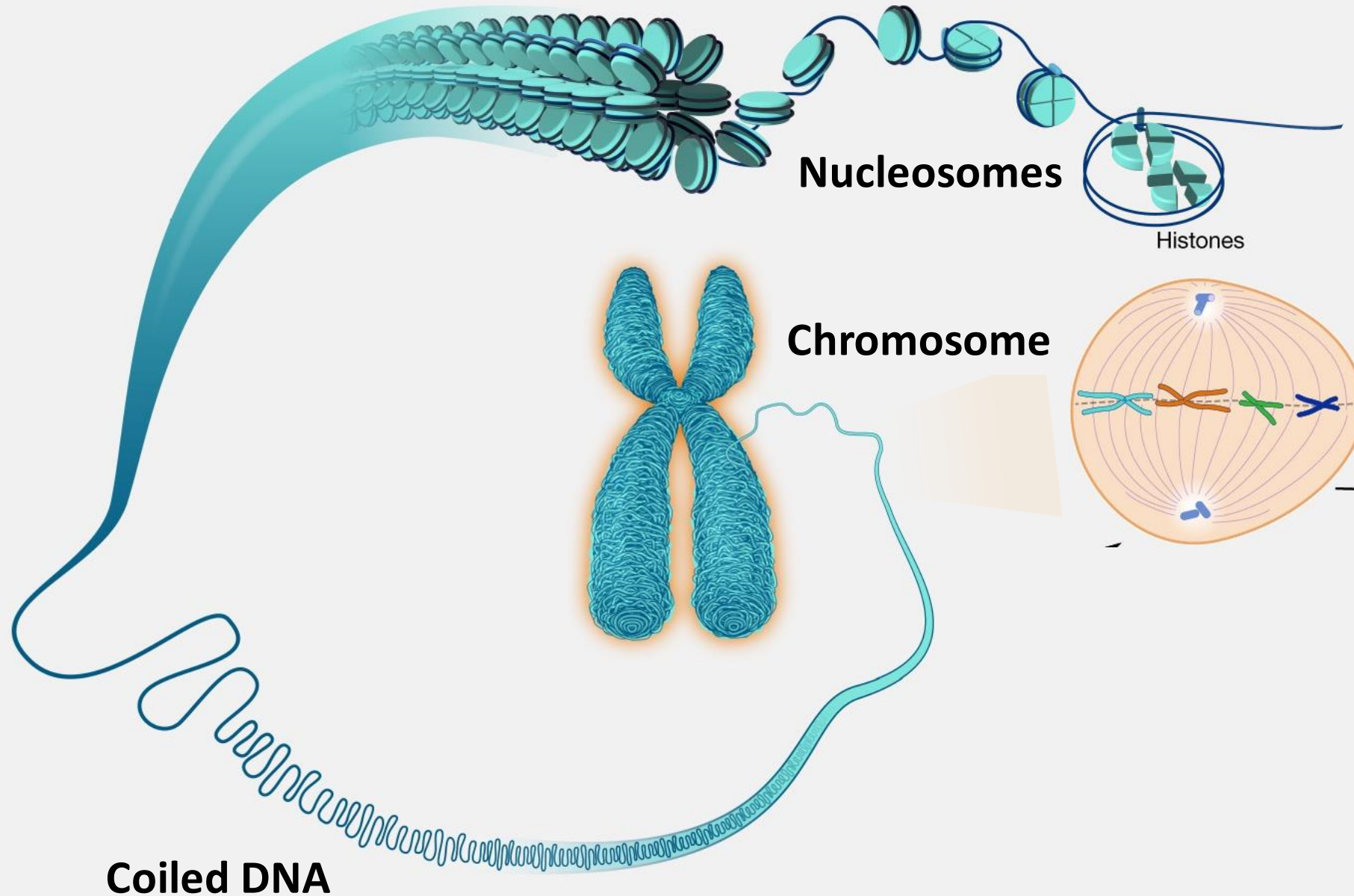
Sex Cells

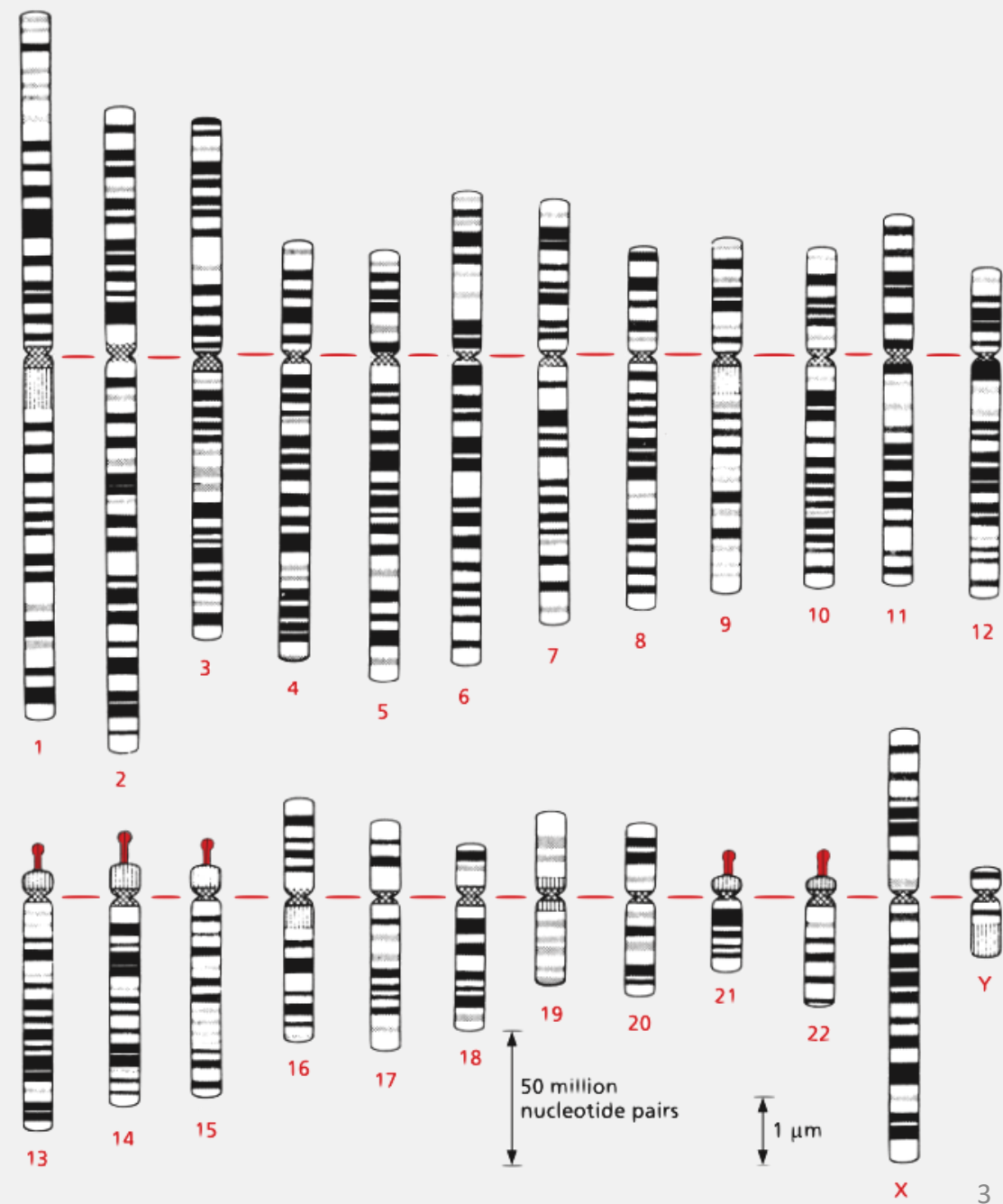


Pancreatic Cells



Cancer Cells





Introduction

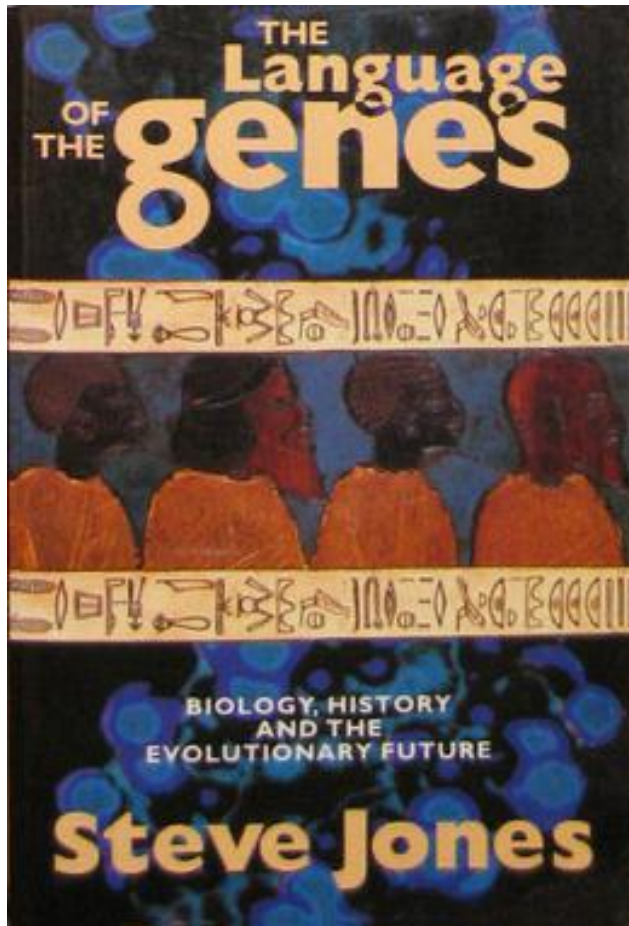
Assembly & Indexing

Genome Annotation

Splice Site Prediction

RNA-Seq Prediction

Conclusions: Language of genomes



Steve Jones, 1993

Codons as “words” and genes as “sentences”.

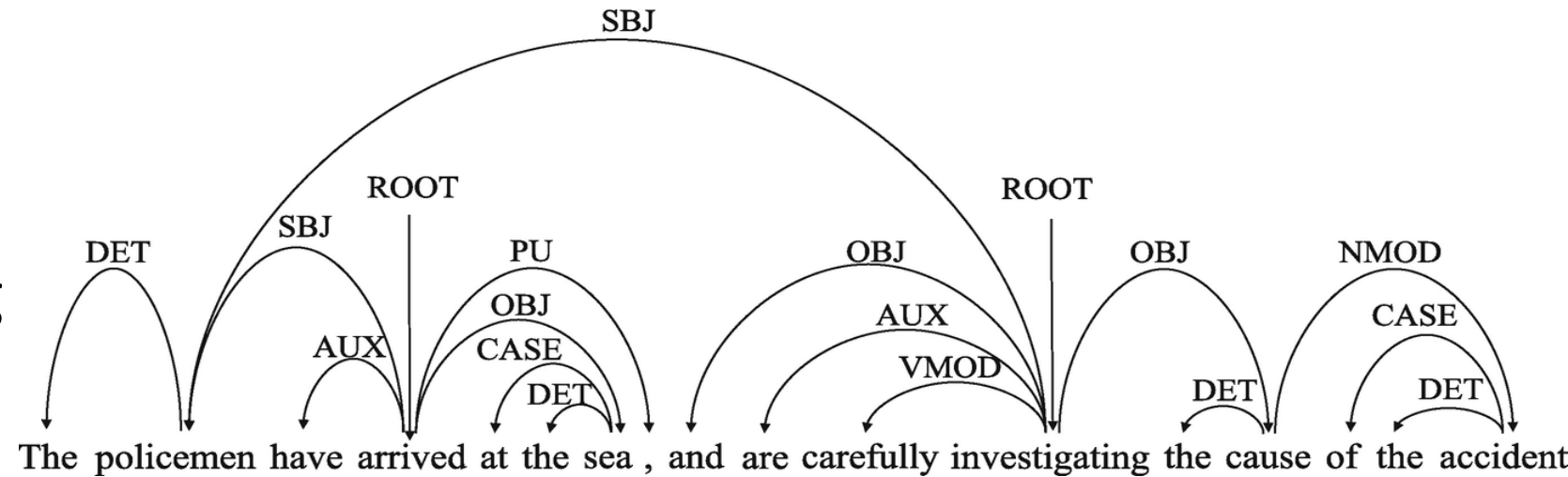
Zhou et al., 2023; Hwang et al., 2024; Sanabria et al., 2024; Nguyen et al., 2024; Theodoris, 2024; Dalla-Torre et al., 2025

Recurring k-mer motifs as discrete tokens,
modeling long-range dependencies as grammatical rules.

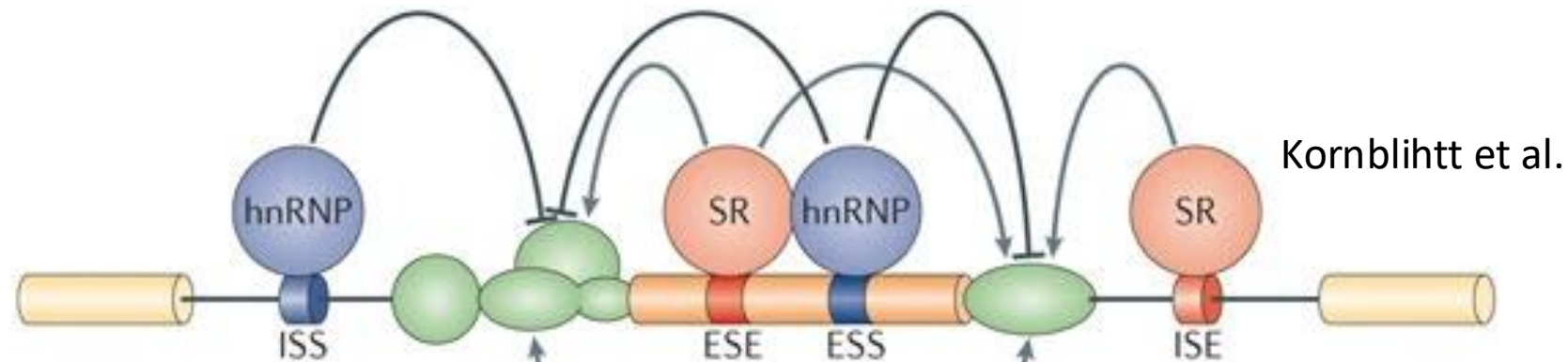
```
accatgttggccaggctggctctgaactccgagccaggtgatccaggt  
ccctgttggctcagtgaaatgaggggctgacctggcattccagcctgaggtg  
tctctgactcctcaGTCAC TAGAGGATCTTGAAGACCAAAACGAAACGAA  
CGCCTGCTGCCCGCAGGCGTGTCTGTGCGTGTGGCGTGTGCAACGAACT  
TACATACCGCAAACCGTGTGTGAACCTGTCAACTCTCTGCTCTTGGAACT  
ACTTGAAGGACTGCTGTGTATTGTAAATAACAAACTATTGTGACTGTGHT  
TAACCACAAAACGCCATCGTCGTGAGGTAAGCTGTCTCTTACAACCTCT  
ttatttgcacaaatgctgaaaacttattctatctaaattattacataaatat  
ttctgttgggtctctctctctctctctctctctctctctctctctctctct  
tgagagaggttgagttcagggaagacagcctgccacacatccaaggttarg  
ggaggccacagtgaggacacagagtcacagcctgagccctccacccctct  
gccccaaagaggttcaccagcaaggaactgtgtgtattttaaagtcaggt  
gacatttccagctcttaaaaagaagcaaggagatttcaaatgctagatct  
aacccagcctaggggtgacggggagcgacccaagctggcatcttctctct  
aataggagataattctaagagttactaaaggatgatttatttaagagat  
gttatttgattgctctctatttgggtccgggtgggaggtctgcccctct  
gcccctgactgtaatttgagggtaccctgtggtccaaacgacccctct  
cgctcccacgatccaccctccctggaggaccgagaggtctcaggagat  
cccaaagcctgacatagtcagtagcctcaatctctatctctccactga  
gggaaggccaggaaagccctgggtgacaaggaacttagggcactctaccc  
caagtgtggccacacgcatgccctactccctctccctggccctgagag  
cgctctctcatgcgggacataggcaaaacttggccagagaggaatgact  
jagaatggcatgaacccgggagttggagcttgagtgagccgagatcgc  
jcatccaggccagctgcacaggttgggactgcattactccagaggtg  
taacattgattaggcatttaccacaggccaggtccatgctatgcaact  
taccactccacgtctgagcagtgagggtgcccgaagcagctgggctct  
tcacctagagaggaatcagcgggggaatcagagcagaaacccgctg  
tcacctagagaggaatcagcgggggaatcagagcagaaacccgctg
```

Conclusions: Language of genomes

Natural Language Processing



DNA gene regulation



How do we assemble the complete 3-billion-nucleotide genome?
How can we efficiently represent multiple genomes for fast pattern matching?

Part I: Genome Assembly & Indexing

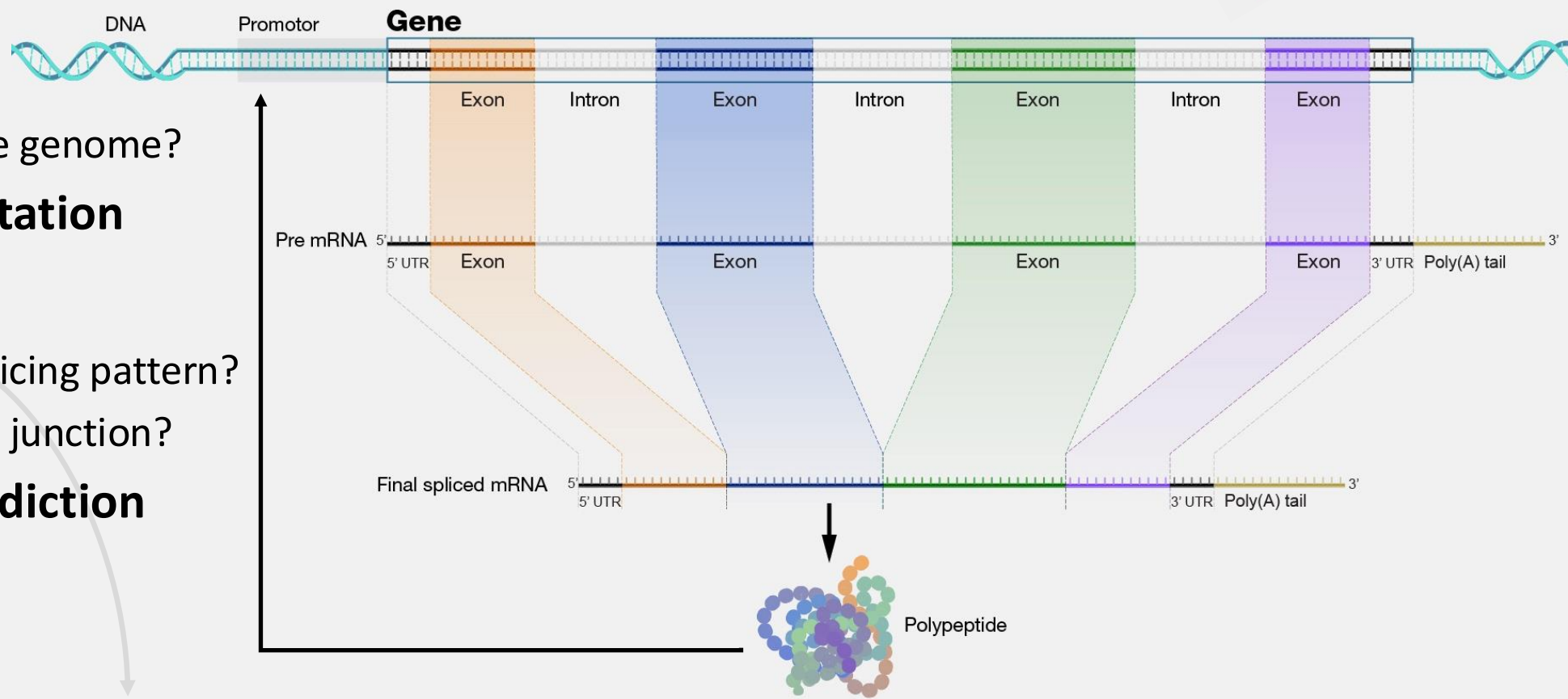


Where are the genes in the genome?

Part II: Genome Annotation

What are the canonical splicing pattern?
Alternative splicing? Splice junction?

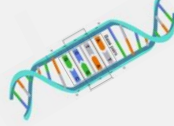
Part III: Splice site prediction



Can we predict gene expression by learning the regulatory grammar in the genome?

Part IV: Shorkie. RNA-Seq coverage prediction

Part I



Steven Salzberg



Mihaela Pertea



Ben Langmead

Genome Assembly & Indexing

- Han1: first gapless Southern Han Chinese genome
- WGT: Wheeler graph recognition algorithm



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.



RECOMB-SEQ



Chao, K. H., Chen, P. W., Seshia, S. A., & Langmead, B. (2023). WGT: Tools and algorithms for recognizing, visualizing, and generating Wheeler graphs. *Iscience*, 26(8).

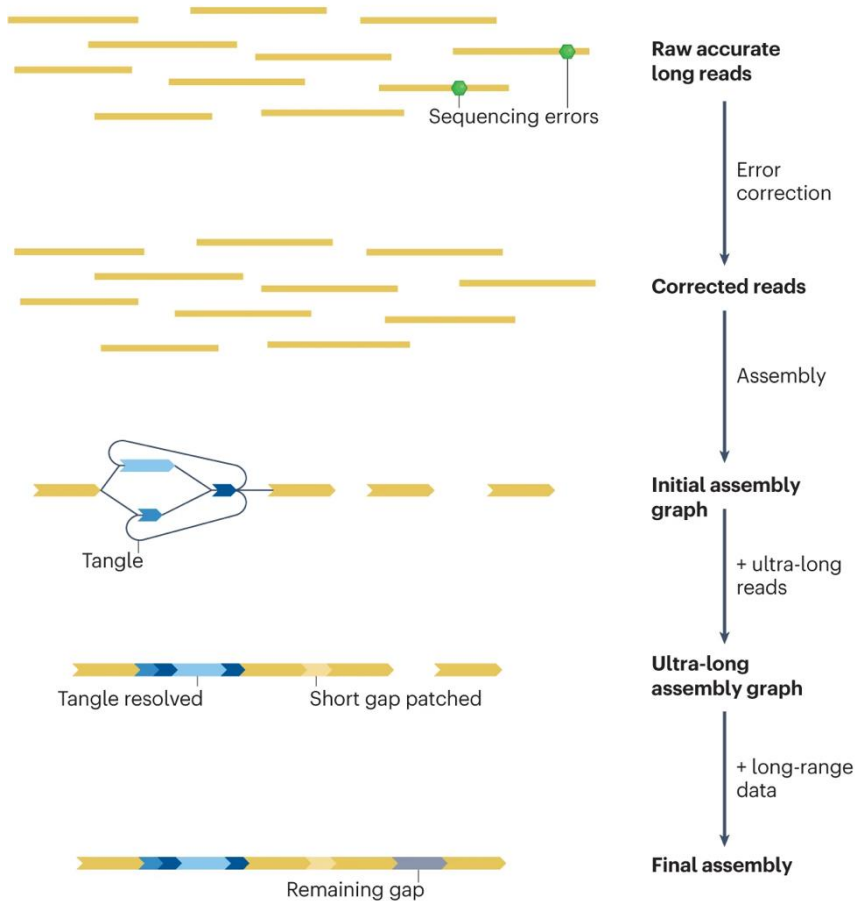


8

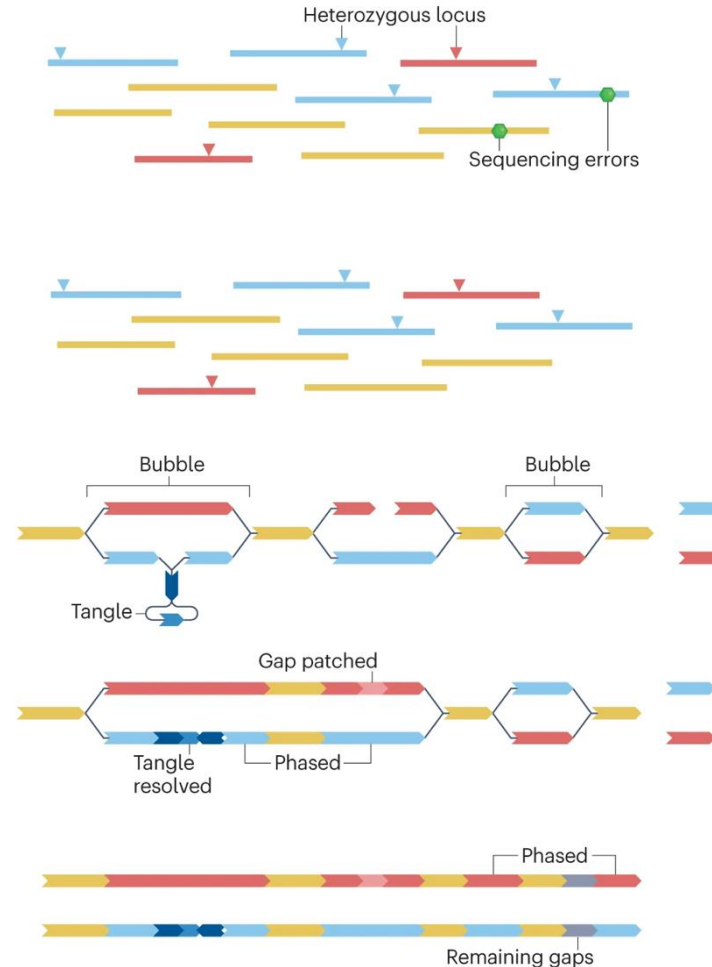
Genome Assembly

Li, H., Durbin, R. Genome assembly in the telomere-to-telomere era. Nat Rev Genet 25, 658–670 (2024).

a Homozygous genome



b Heterozygous genome



1972 Bacteriophage MS2 coat gene

1976 Bacteriophage MS2 ~3.5 kb

1977 Sanger sequencing

1990-2000: BAC-by-BAC shot gun strategy

2003 near-complete human genome

2000s second-generation sequencing
short-read sequencing
(Illumina)

2010s third-generation sequencing
long-read sequencing
(PacBio; ONT)

2019 Telomere-to-Telomere (T2T) era

2022 First complete human genome

2026 First complete diploid human
genome

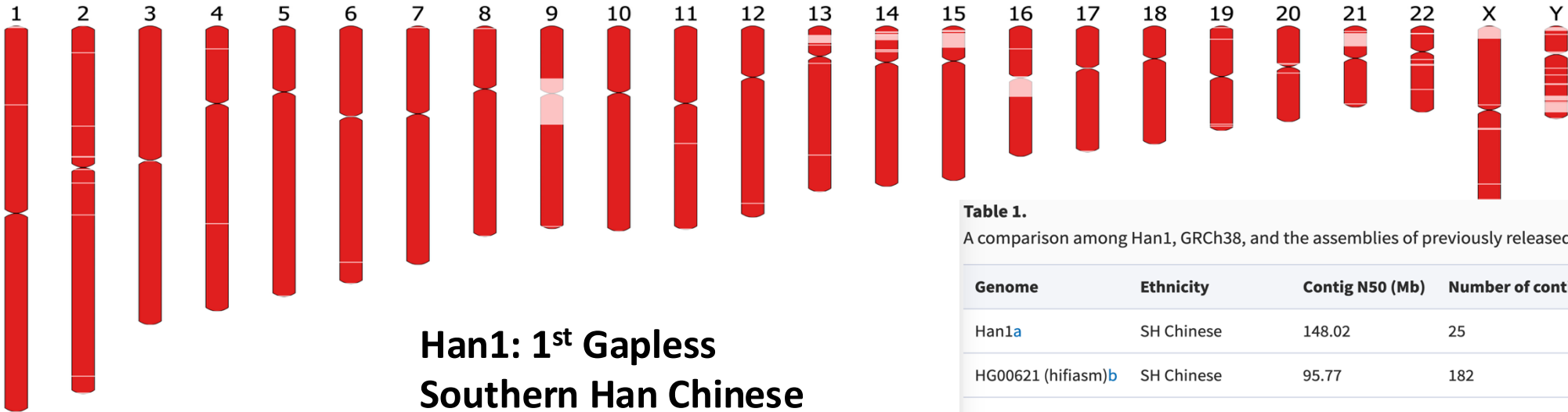
Genome Assembly: Han1



GenBank assembly:
GCA_024586135.1



Genes | Genomes | Genetics



Han1: 1st Gapless
Southern Han Chinese

Table 1.
A comparison among Han1, GRCh38, and the assemblies of previously released Chinese genomes.

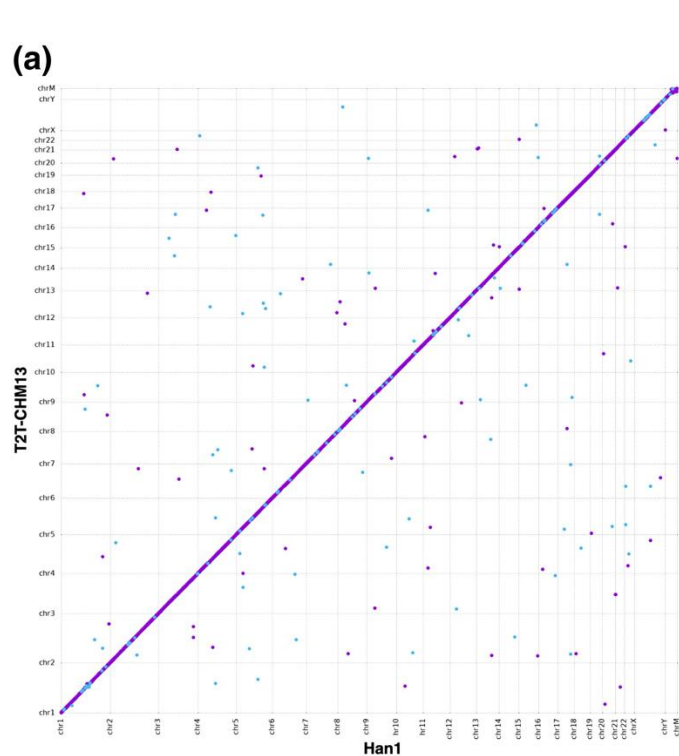
Genome	Ethnicity	Contig N50 (Mb)	Number of contigs	Number of gaps	Assembly size (Gb)
Han1a	SH Chinese	148.02	25	0	3.10
HG00621 (hifiasm)b	SH Chinese	95.77	182	157	3.11
T2T-CHM13v2.0c	Northern European	150.62	25	0	3.12
HJ-H1d	NH Chinese	28.15	1,330	427	3.07
HJ-H2d	NH Chinese	25.90	896	390	2.91
NH1d	NH Chinese	3.60	11,019	8,484	2.89
HX1d	SH Chinese	8.33	5,843	4,025	2.93
YH2.0e	SH Chinese	0.02	361,157	235,514	2.91
TJ1.p0f	Tujia	13.67	1,430	907	2.87
TJ1.p1f	Tujia	13.70	1,426	873	2.87
ZF1g	Tibetan	23.62	1,384	1360	2.85
GRCh38.p14h	Mixed	57.88	994	804	3.10

Table 3.
Statistics for the preliminary Han1 assemblies.

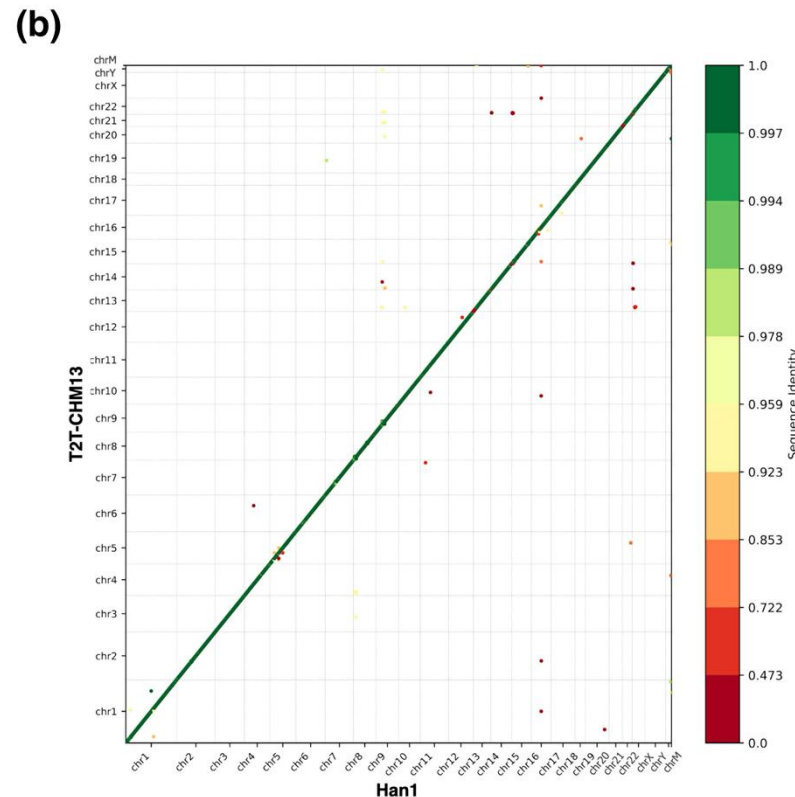
Assembler	Sequencing data	assembled sequence (bp)	Contig N50	Number of contigs	Quality value
Hifiasm v0.16.1-r375	PacBio HiFi	3,110,501,483	95,769,069	182	57.8a
Flye v2.5	ONT Ultralong	2,974,205,132	40,850,737	1,658	25.6a

GRCh38, T2T-CHM13, and other assemblies of Chinese genome

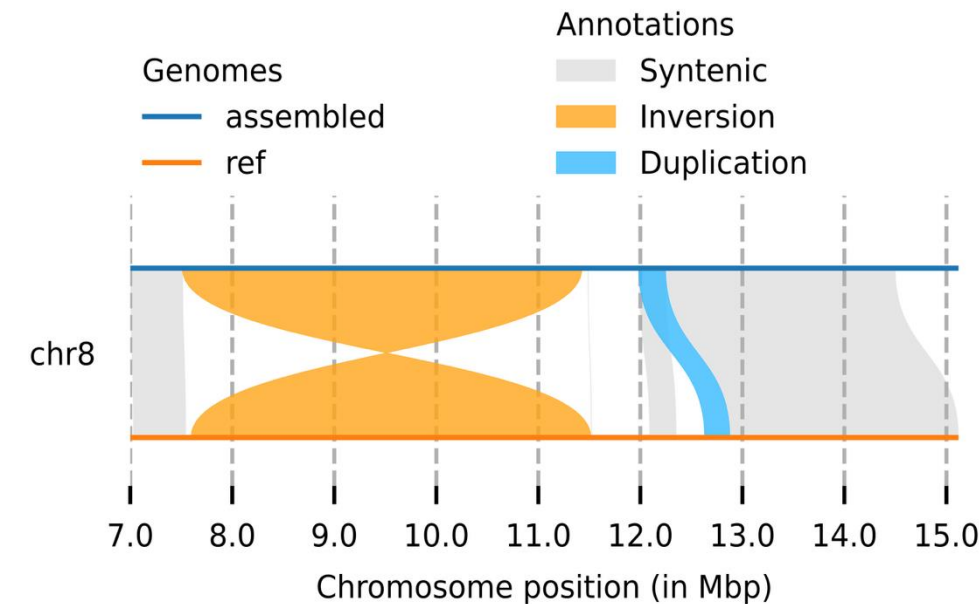
Mummer dot plot



Gene order plot



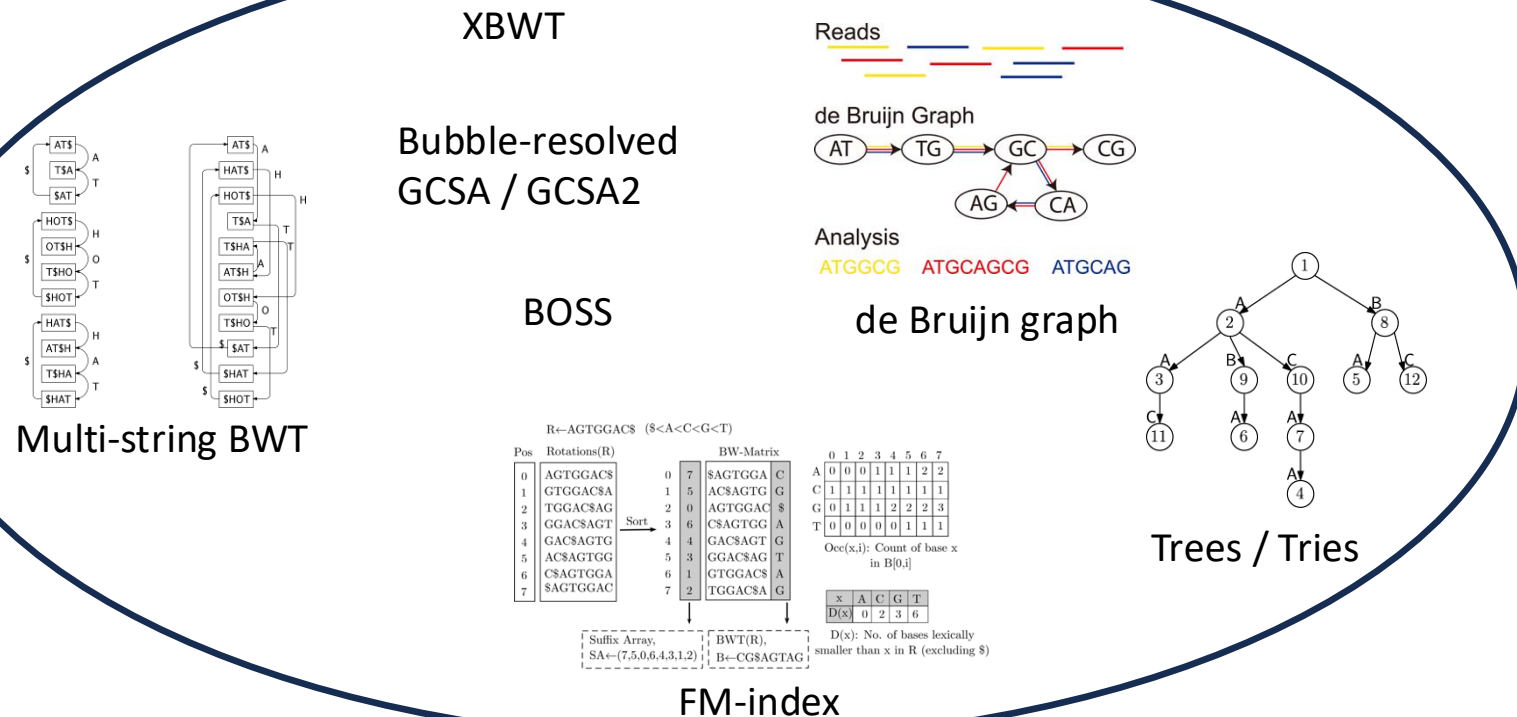
Inversion on β defensin gene cluster



Wheeler Graph Recognition

- Wheeler graphs are the basis of many pangenome and other sequence analysis tools

Wheeler graph



1994 Burrows-Wheeler Transform (BWT)

2000 FM-index

2009 Bowtie

2012 colored de Bruijn graphs

2012 BOSS data structure

2014 GCSA

2017 GCSA2

2017 VARI

2017 Wheeler graph

2018 Vg toolkit

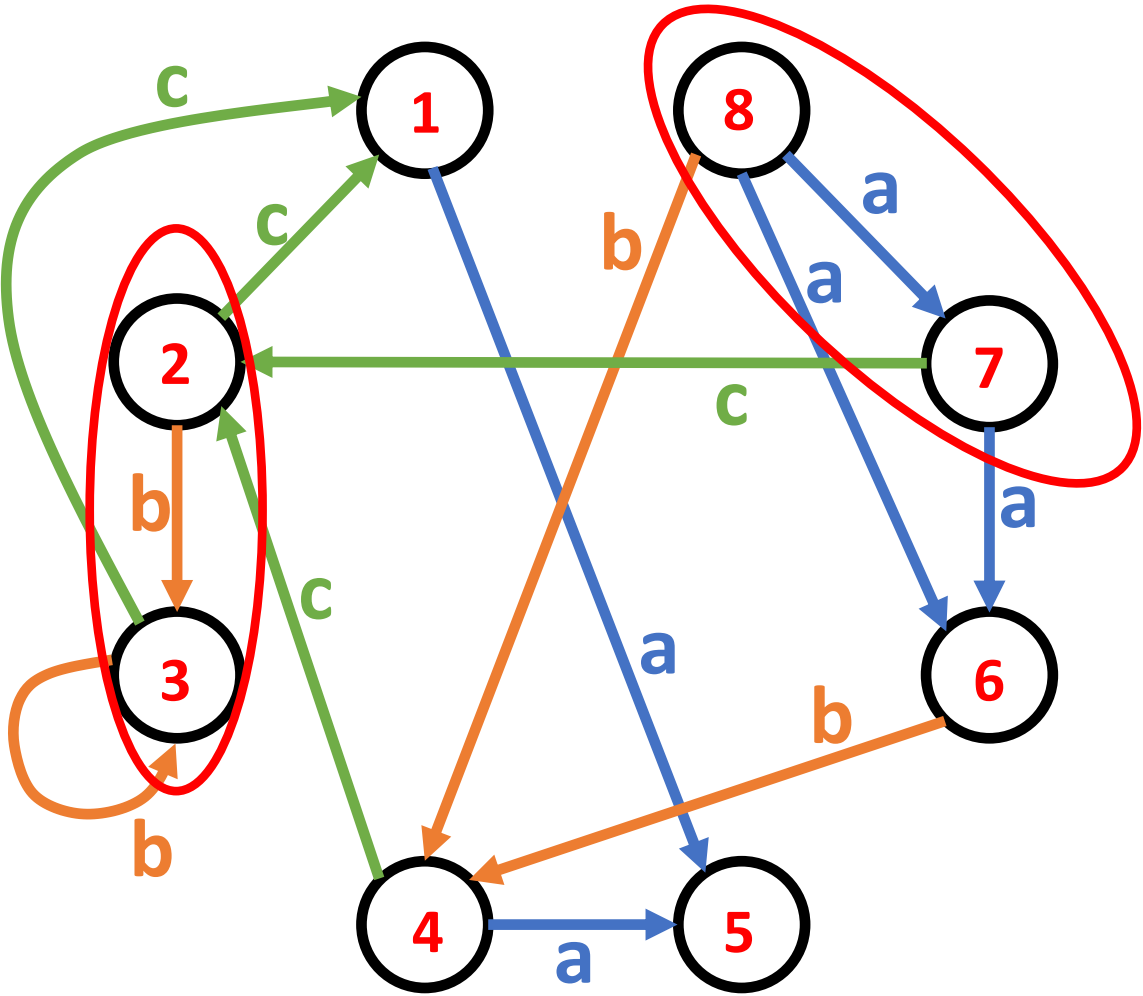
2020 Graph Burrows-Wheeler Transform (GBWT)

2021 vg Giraffe

Wheeler Graph Recognition

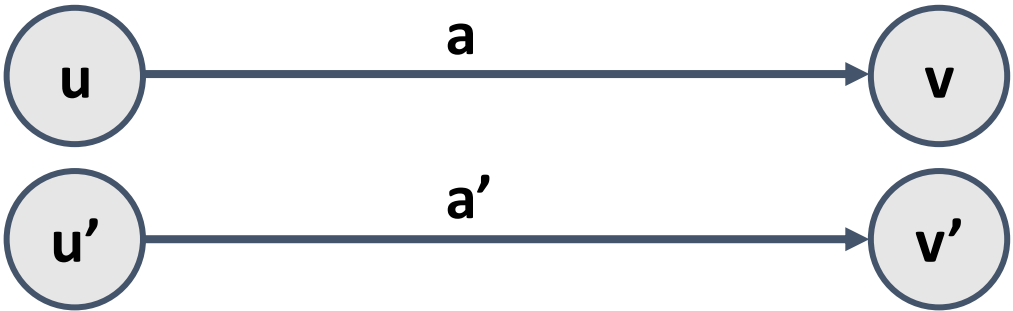


RECOMB-SEQ

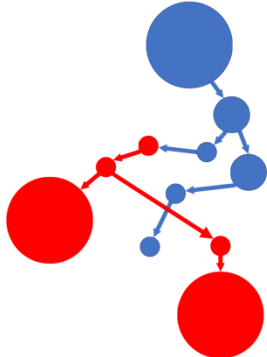


Wheeler graph recognition problem

1. Node with indegree 0 comes before every other nodes.
2. $a < a' \implies v < v'$
3. $(a = a') \wedge (u < u') \implies v \leq v'$



Wheeler Graph Recognition

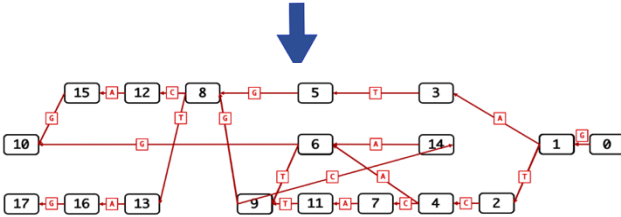
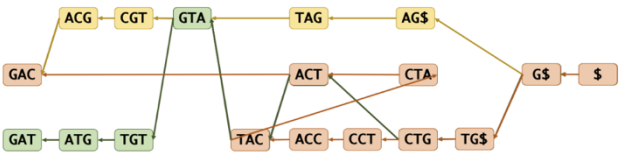


WGT
Wheeler Graph Toolkit



GACGTA-CTG
GACGTA--G
GATGTA-CTG
GAC-TACCTG

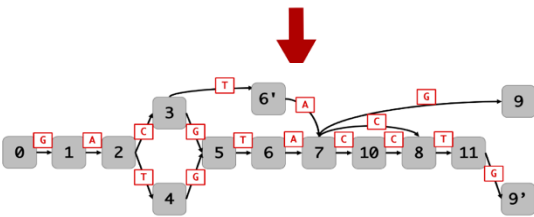
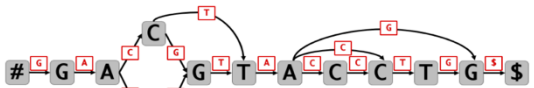
De Bruijn Graph



18 nodes, 21 edges
1 “distant gluing”

Original label	Wheeler order	Original label	Wheeler order
0	1	4	10
11	2	1	11
15	3	10	12
6	4	17	13
3	5	8	14
16	6	9	15
7	7	5	16
12	8	2	17
14	9	13	18

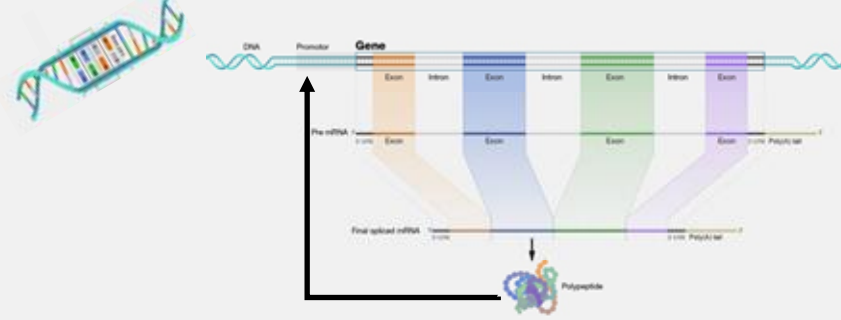
GCSA (Sirén et al., 2014)
Reverse deterministic graph



14 nodes, 16 edges
No “distant gluing”

Original label	Wheeler order	Original label	Wheeler order
0	1	9	8
2	2	5	9
7	3	9'	10
3	4	4	11
10	5	6'	12
8	6	11	13
1	7	6	14

Part II



Steven Salzberg



Mihaela Pertea

Genome Annotation

- LiftOn: genome annotation lift-over
- Application: CHES (GRCh38 - > CHM13; GRCh38 - > Han1)



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 Annotation



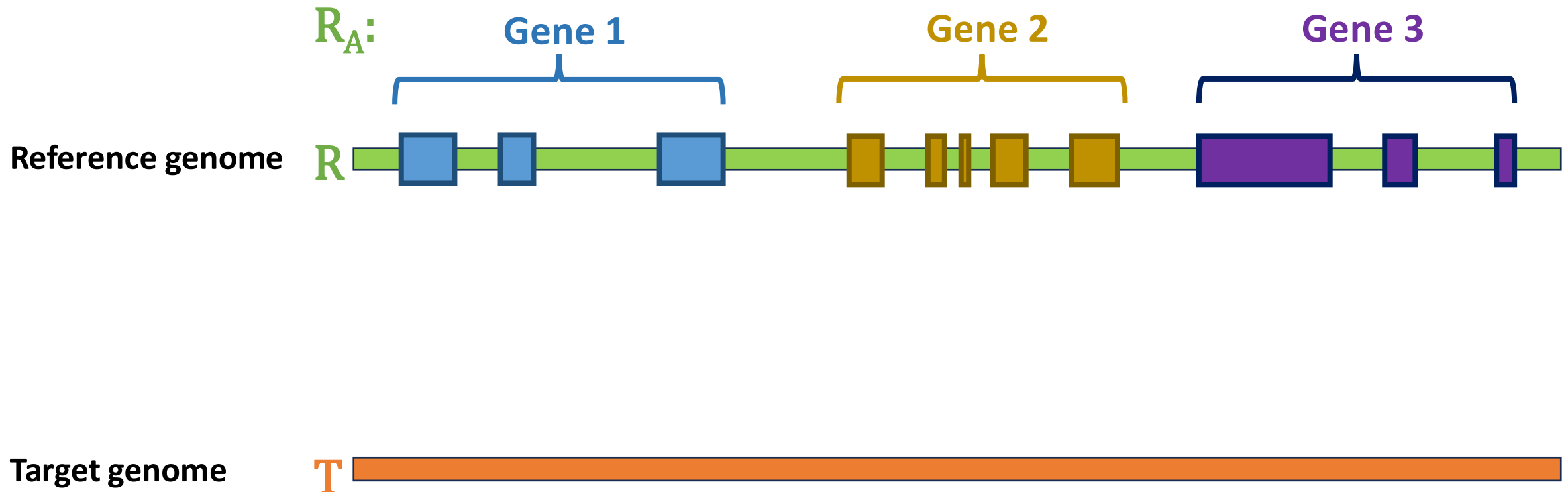
Genome
(FASTA)

```
CAGCCCCCGGAGACCTtaaatacaggaagaaaaaggCAGGACAGAATTACAAGGTGCTGGCCCAGGGCGGGCAGCGGCCCT
GCCTCCTACCCTTGCGCCTCATGACCAGCTTGTTGAAGAGATCCGACATCAAGTGCCACCTTGGCTCGTGGCTCTCACT
GCAACGGGAAAGCCACAGACTGGGGTGAAGAGTTCAAGTCACATGCGACCGGTgactccctgtccccacccccatgACACT
CCCCAGCCCTCCAAGGCCACTGTGTTTCCCAGTTAGCTCAGAGCCTCAGTCGATCCCTGACCCAGCACCGGGCACTGATG
AGACAGCGGCTGTTTGAGGagccacctcccagccacctcggggccagggccagggtgtGCAGCACCACTGTACAATGGGG
AAACTGGCCCAGAGAGGTGAGGCAGCTTGCTTGGGGTCACAGAGCAAGGCAAAAGCAGCGCTGGGTACAAGCTCAAAACC
ATAGTGCCCAAGGGCACTGCCGCTGCAGGCGCAGGCATCGCATCACACCAGTGTCTGCGTTTCACAGCAGGCATCATCAGTA
```

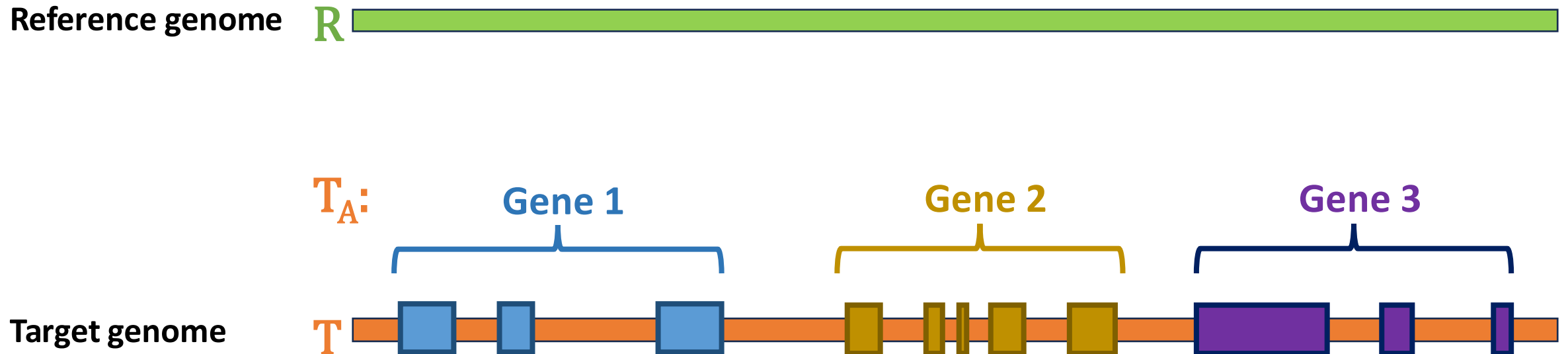
Annotation
(GFF / GTF)

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;

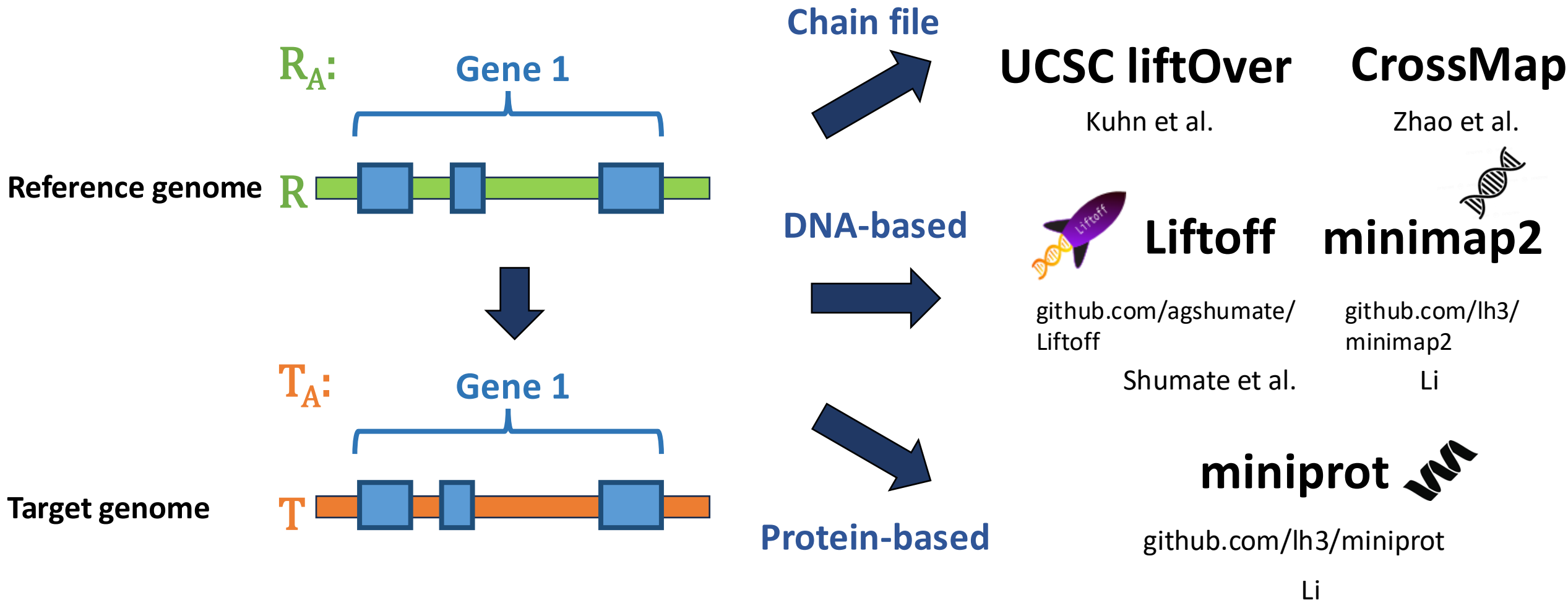
Lift-over Problem Definition:



Lift-over Problem Definition:



Lift-over problem, what methods are available?



Application: GRCh38 to T2T-CHM13 lift-over

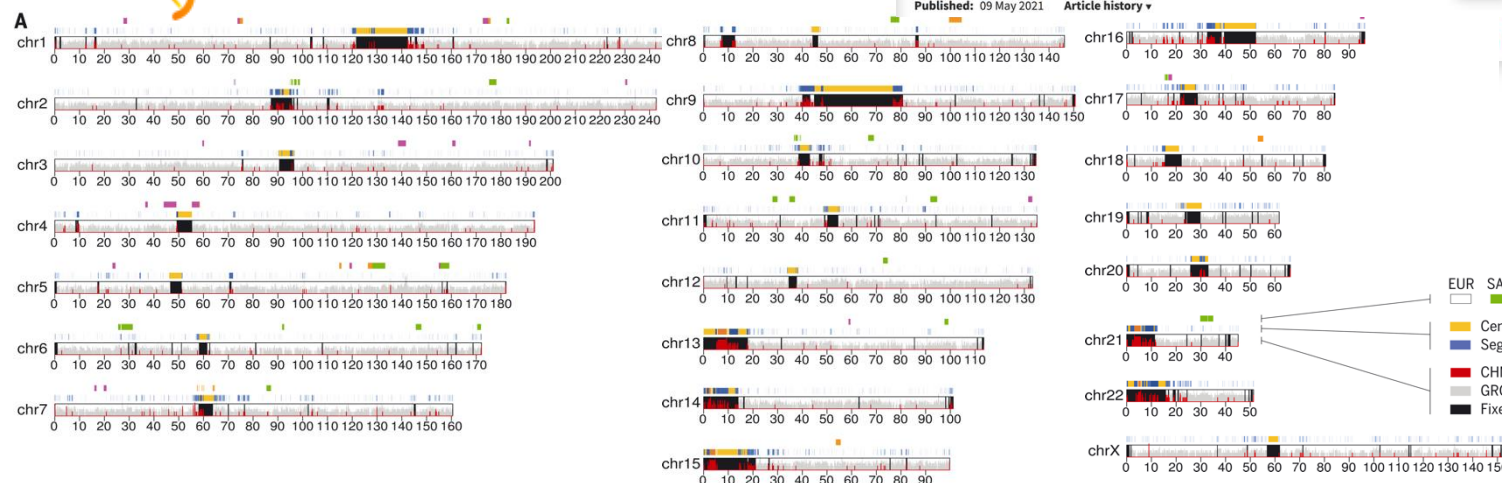
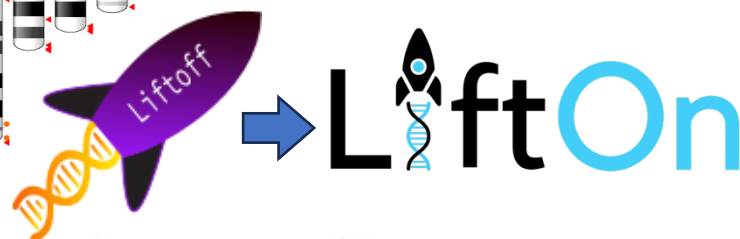
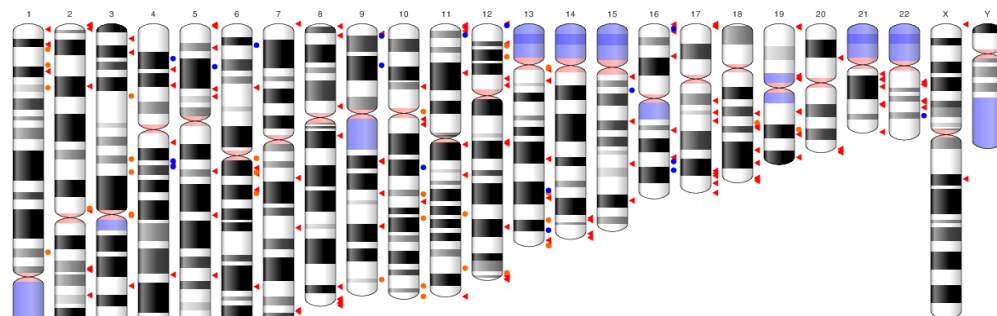
NCBI Build 34
NCBI Build 35
NCBI Build 36.1
GRCh37



GRCh38
(Dec 2013)



T2T-CHM13
(Jan 2022)



nature

Explore content About the journal Publish with us

nature > article

Original Article | Published: 01 February 2001

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium

Science

Current Issue First release papers

HOME > SCIENCE > VOL. 291, NO. 5507 > THE SEQUENCE OF THE HUMAN GENOME

SPECIAL REVIEWS

The Sequence of the Human Genome

J. CRAIG VENTER, MARK D. ADAMS, EUGENE W. MYERS, PETER W. LI, L. I. AND YAOHONG ZHU +269 authors Authors Info & Affiliations

Bioinformatics



Article Navigation

JOURNAL ARTICLE

Liftoff: accurate mapping of gene annotations

Alaina Shumate, Steven L. Salzberg

Bioinformatics, Volume 37, Issue 12, June 2021, Pages 1639–1643,
<https://doi.org/10.1093/bioinformatics/btaa1016>

Published: 09 May 2021

Article history

- 10 Mbp of rDNAs
- 182 Mbp of entirely novel sequence
- 1956 novel genes including 99 protein-coding

Science

Current Issue First release p

HOME > SCIENCE > VOL. 376, NO. 6588 > THE COMPLETE SEQUENCE OF A HUMAN GENOME

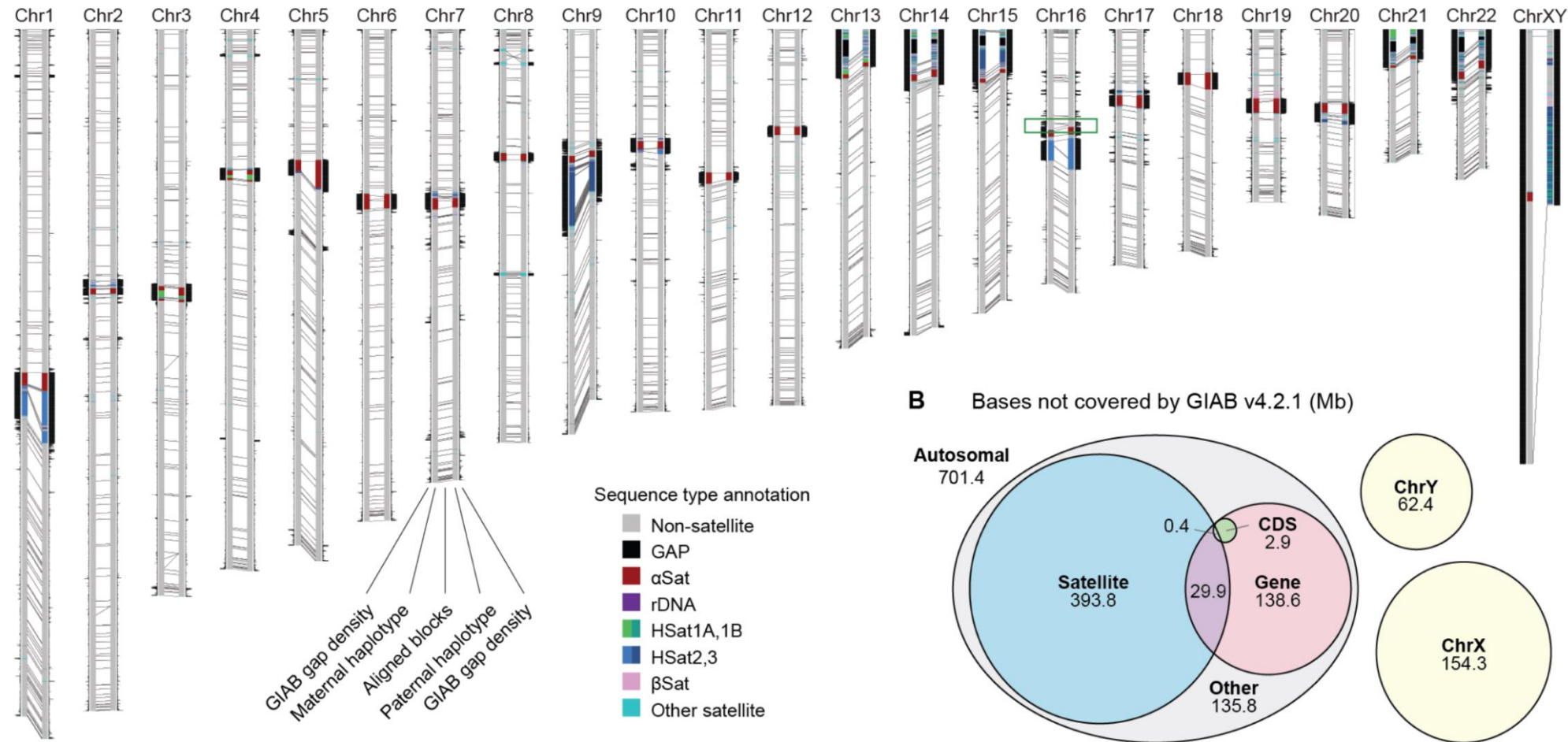
SPECIAL ISSUE RESEARCH ARTICLE | HUMAN GENOMICS

The complete sequence of a human genome

SERGEY NURK, SERGEY KOREN, ARANG RHIE, MIKKO RAUTAIENEN, L. I. AND ADAM M. PHILLIPPY +95 authors

EUR SAS EAS AMR Ancestry
Centromeric satellites
Segmental duplications
CHM13 exclusive gene density
GRCh38 gene density
Fixed GRCh38 gaps and issues

Application: GRCh38 to T2T-HG002 diploid genome lift-over



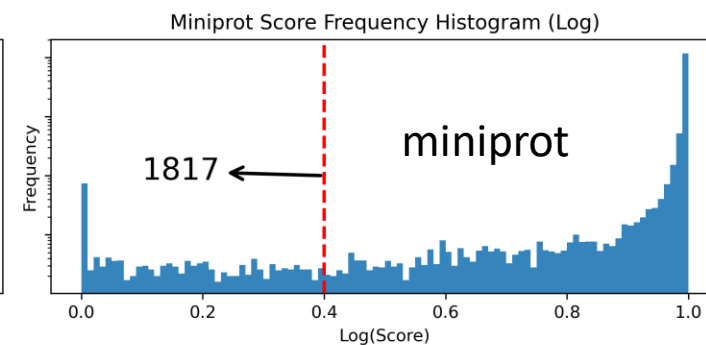
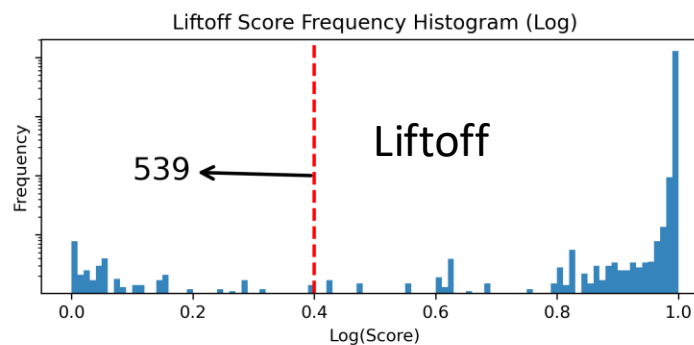
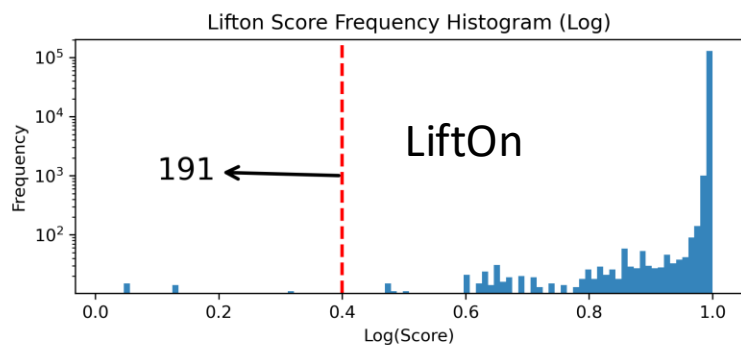
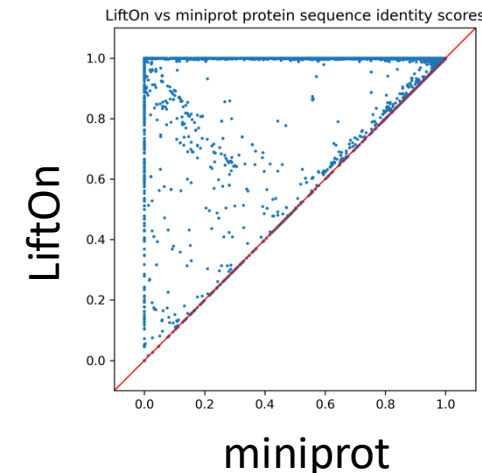
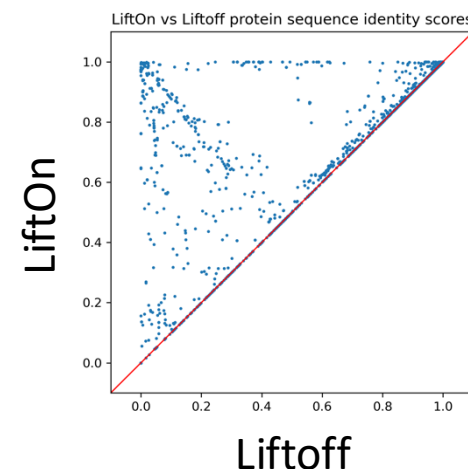
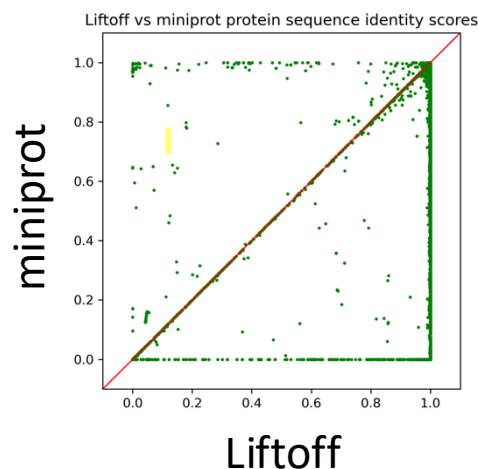
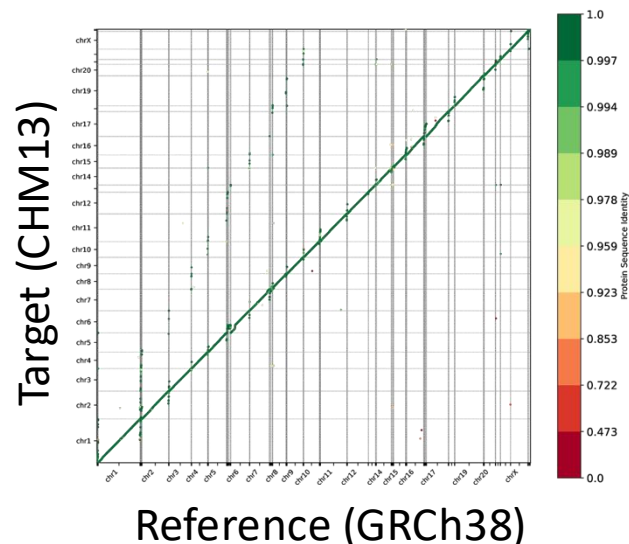
Hansen, N. F., Dwarshuis, N., Ji, H. J., Rhie, A., Loucks, H., Logsdon, G. A., ... & Phillippy, A. M. (2025). A complete diploid human genome benchmark for personalized genomics. *Biorxiv, Cell in press*

Result 1: improves DNA & protein-based lift-over



Map RefSeq v220 from GRCh38 -> CHM13V2.0

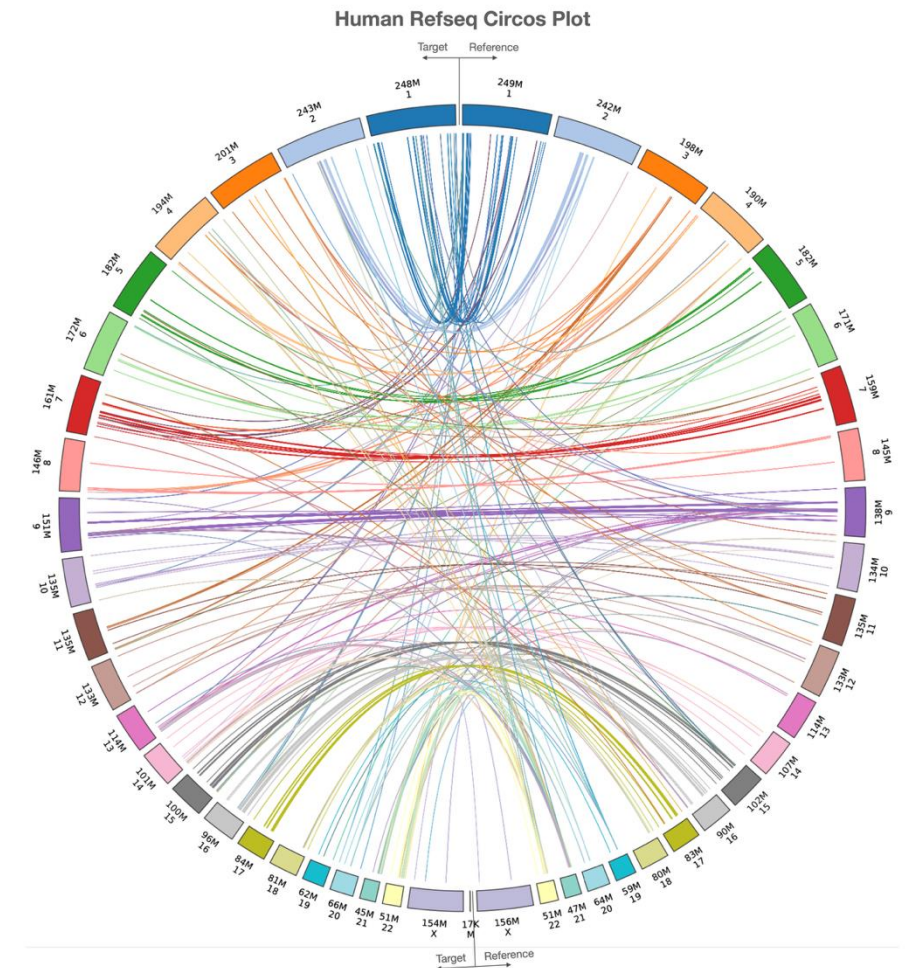
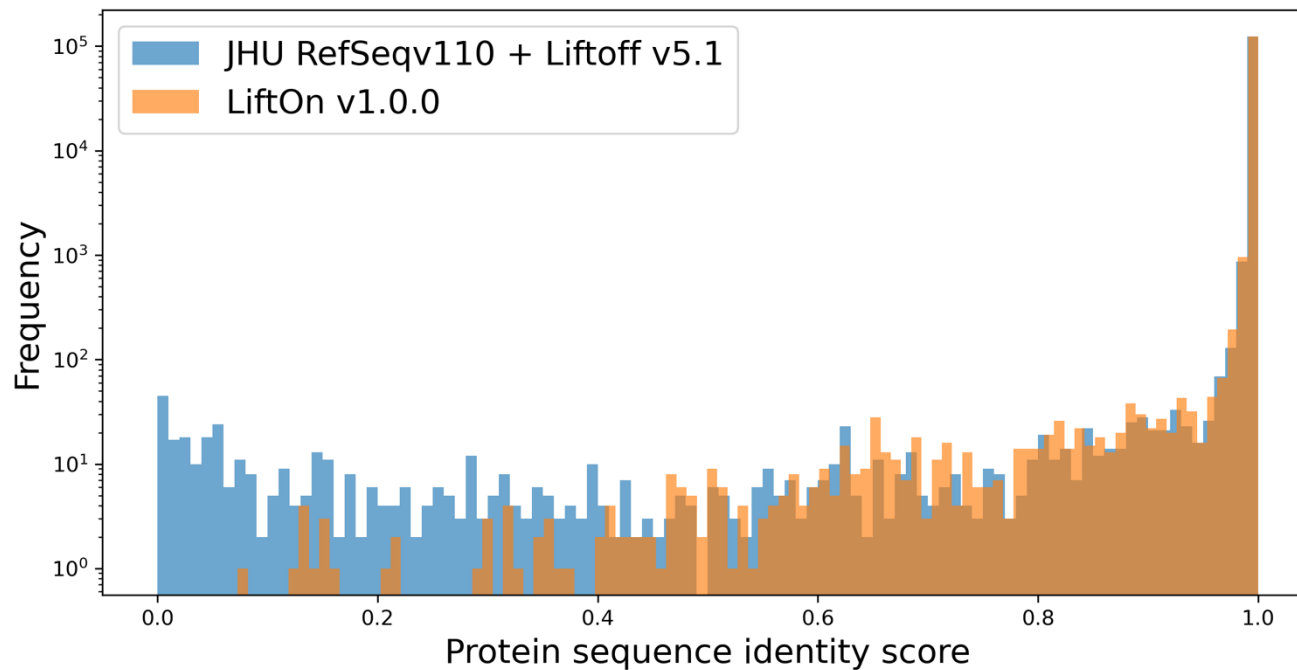
Compressed-gap protein sequence identity



Result 2: improve CHM13 protein annotations

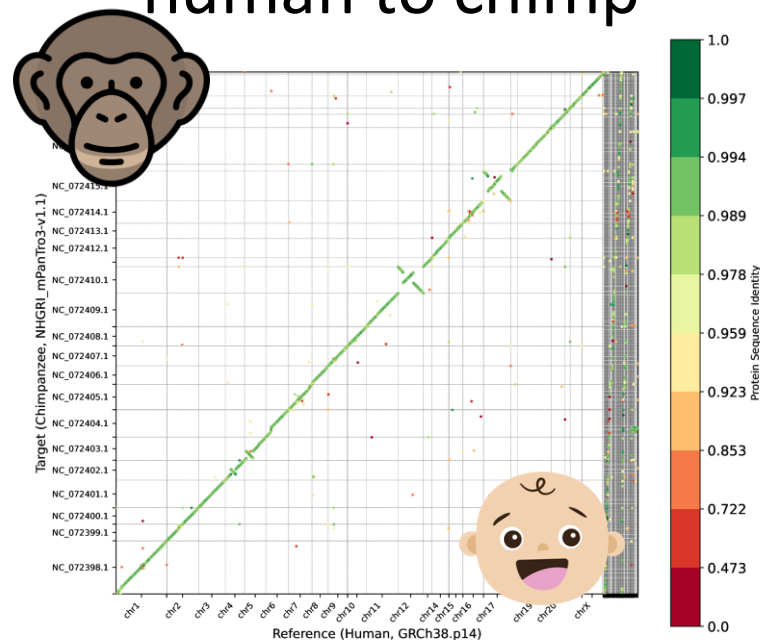


Protein sequence identity score frequency histogram



Result 3: improve distant species lift-over

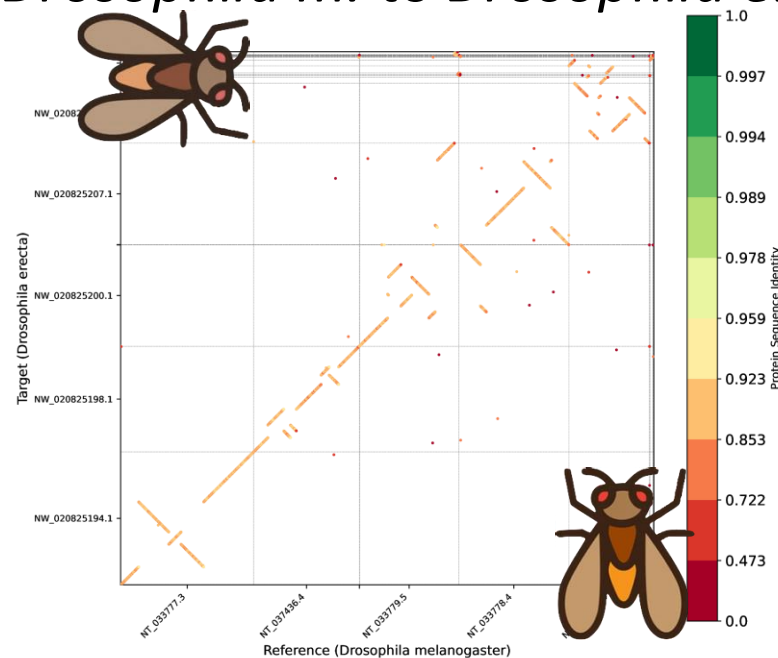
human to chimp



Mash: 0.013

Dashing2: 0.47

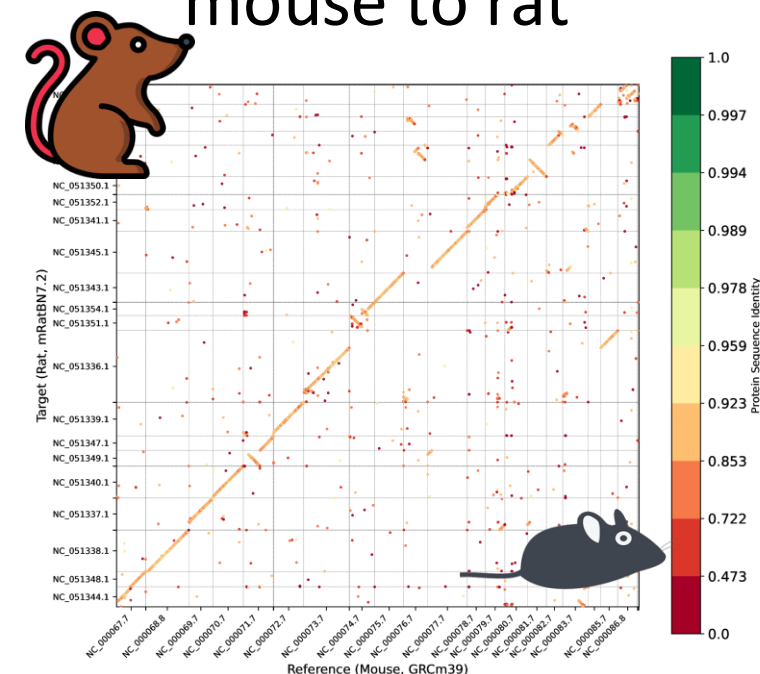
Drosophila m. to *Drosophila e.*



Mash: 0.077

Dashing2: 0.07

mouse to rat



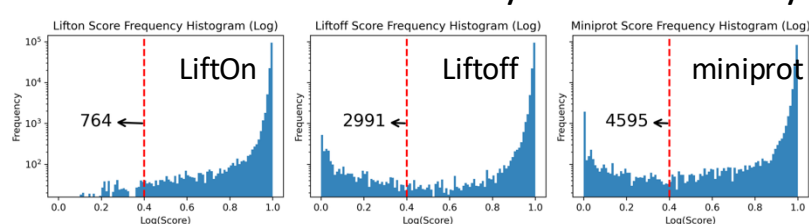
Mash: 0.12

Dashing2: 0.01

DNA + Protein

DNA-only

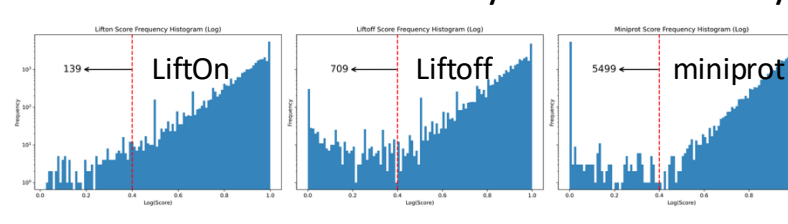
Protein-only



DNA + Protein

DNA-only

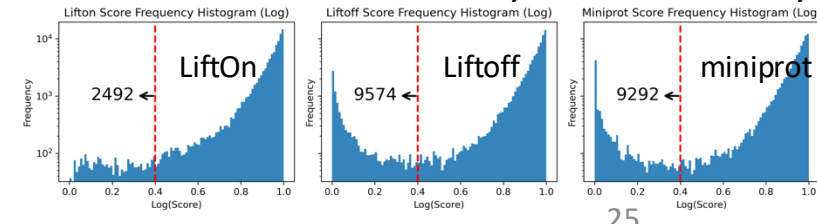
Protein-only



DNA + Protein

DNA-only

Protein-only



25

Introduction

Assembly & Indexing

Genome Annotation

Splice Site Prediction

RNA-Seq Prediction

LiftOn: Protein-maximization algorithm

A

Target genome +
Expected annotation



1. Liftoff annotation



2. miniprot annotation



LifftOn: Protein-maximization algorithm

B Step 1: Align Liftoff & miniprot proteins to reference protein

Liftoff protein alignment

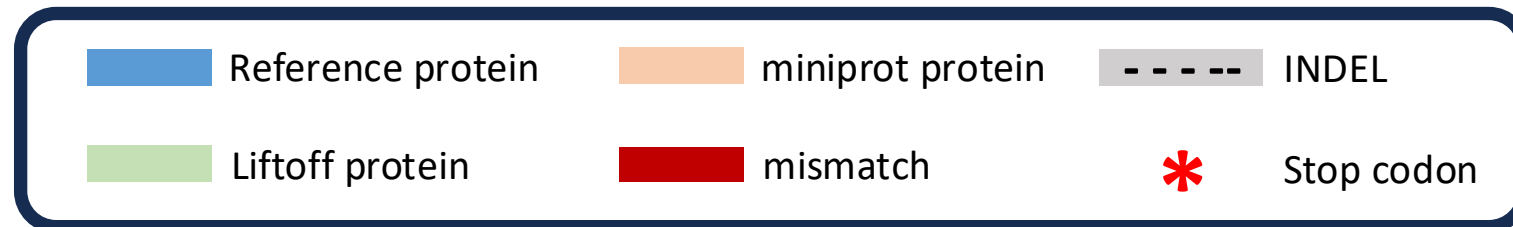
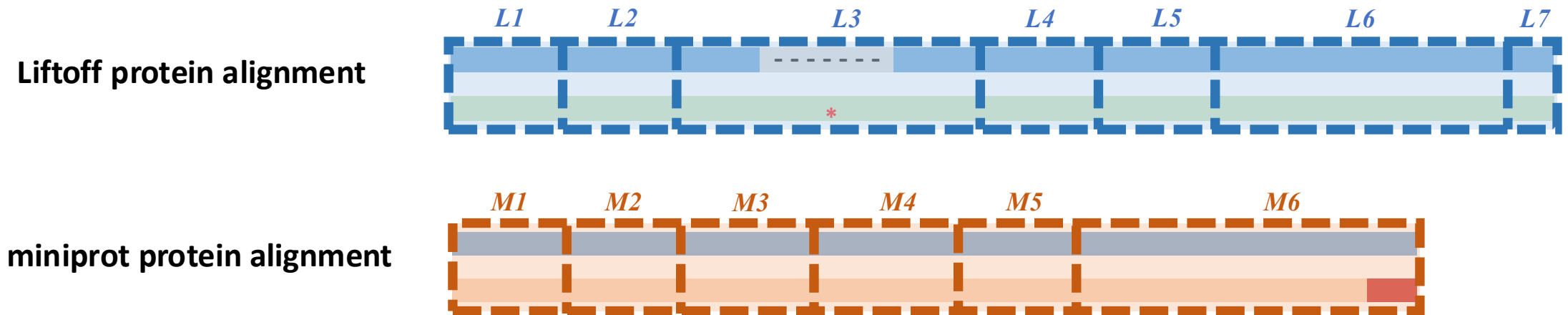


miniprot protein alignment



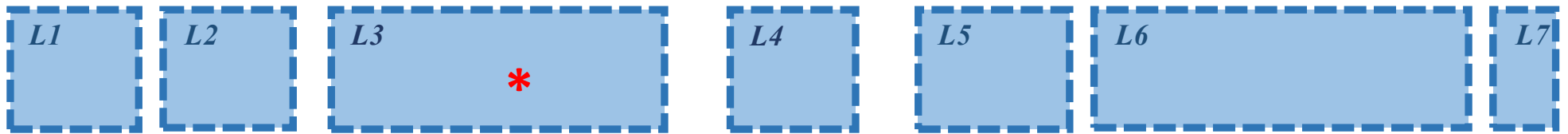
LiftoOn: Protein-maximization algorithm

C Step 2: Mapped CDS boundaries onto Liftoff & miniprot protein alignments



LiftOn: Protein-maximization algorithm

D Step 3: group CDSs by “accumulated AA in the reference protein”

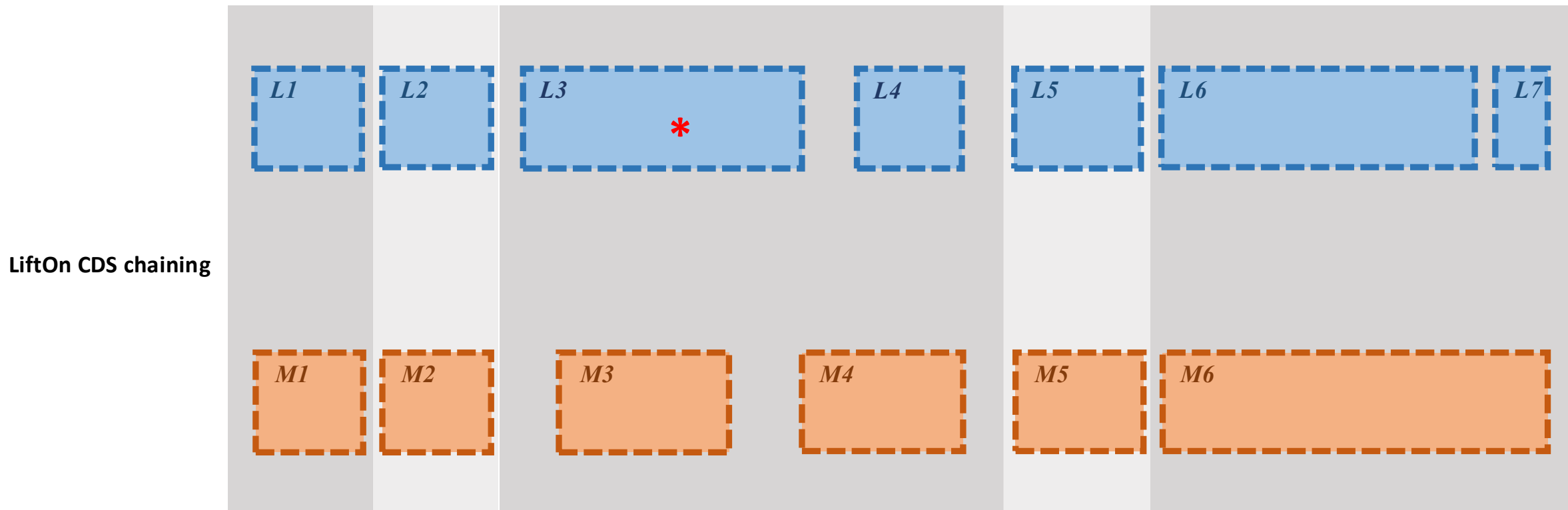


LiftOn CDS chaining



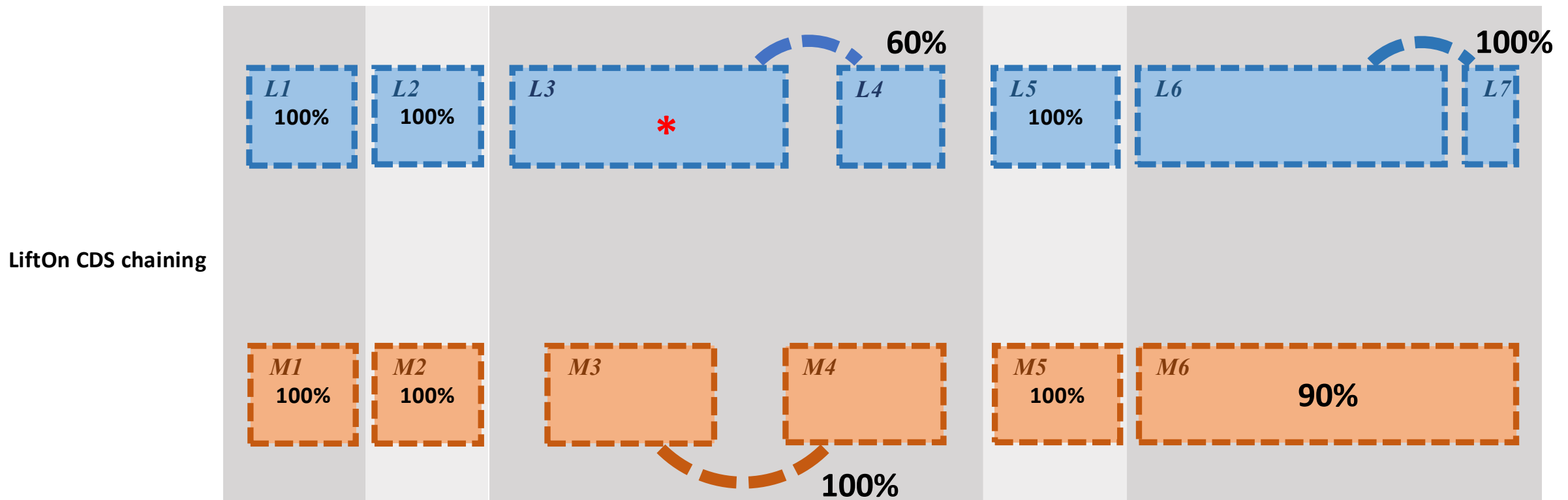
LiftOn: Protein-maximization algorithm

D Step 3: group CDSs by “accumulated AA in the reference protein”



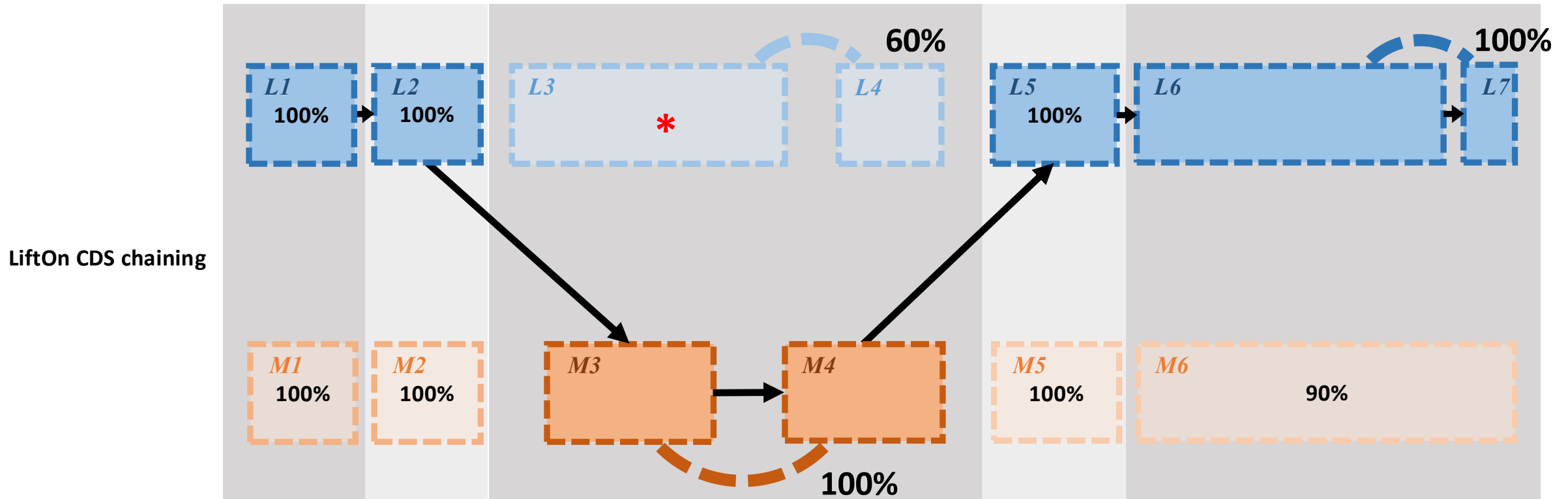
LiftOn: Protein-maximization algorithm

D Step 3: group CDSs by “accumulated AA in the reference protein”

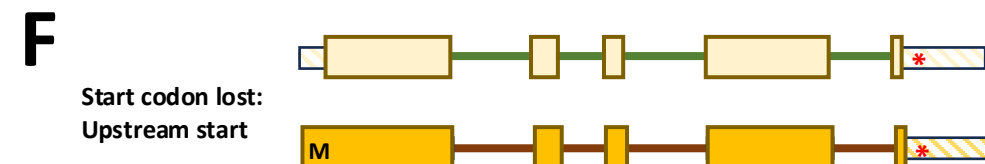
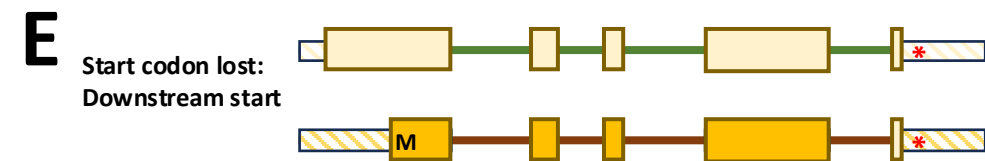
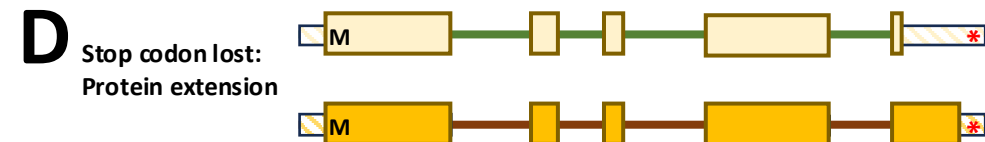
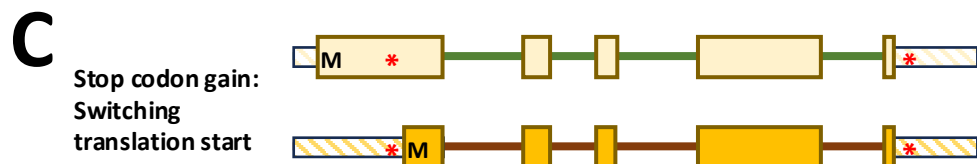
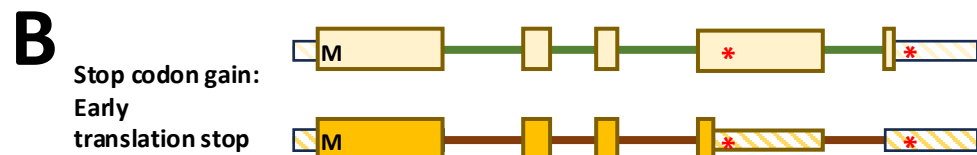
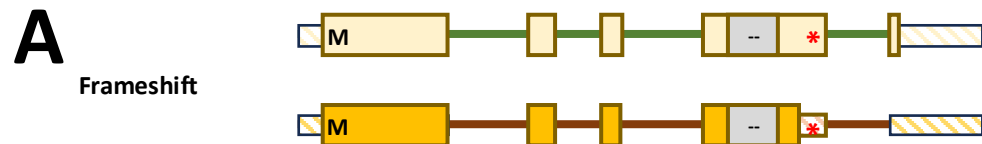
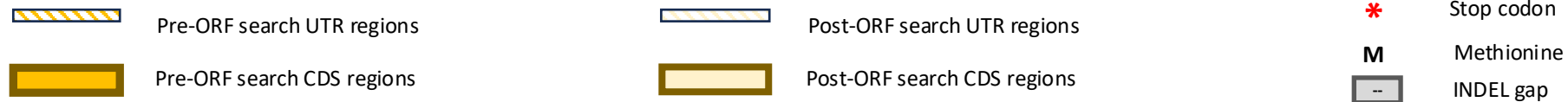


LiftOn: Protein-maximization algorithm

D Step 3: group CDSs by “accumulated AA in the reference protein”



LiftOn: Protein-maximization algorithm



LiftOn: more lift-over applications



House mouse
(*Mus musculus*)



Yeast
(*Saccharomyces cerevisiae*)



Thale cress
(*Arabidopsis thaliana*)



Honey bee
(*Apis mellifera*)

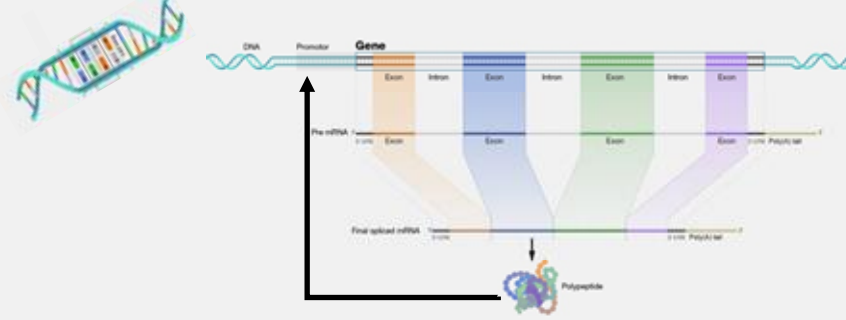


fruit fly
(*Drosophila melanogaster*)



Rice
(*Oryza sativa*)

Part III



Mihaela Pertea



Steven Salzberg



Anqi Liu

Splice Site Prediction

- OpenSpliceAI
- Splam



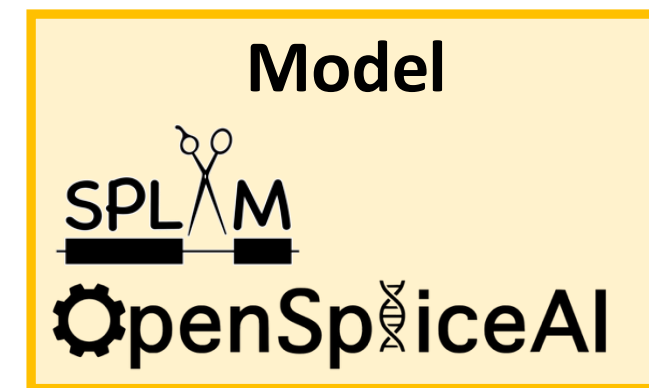
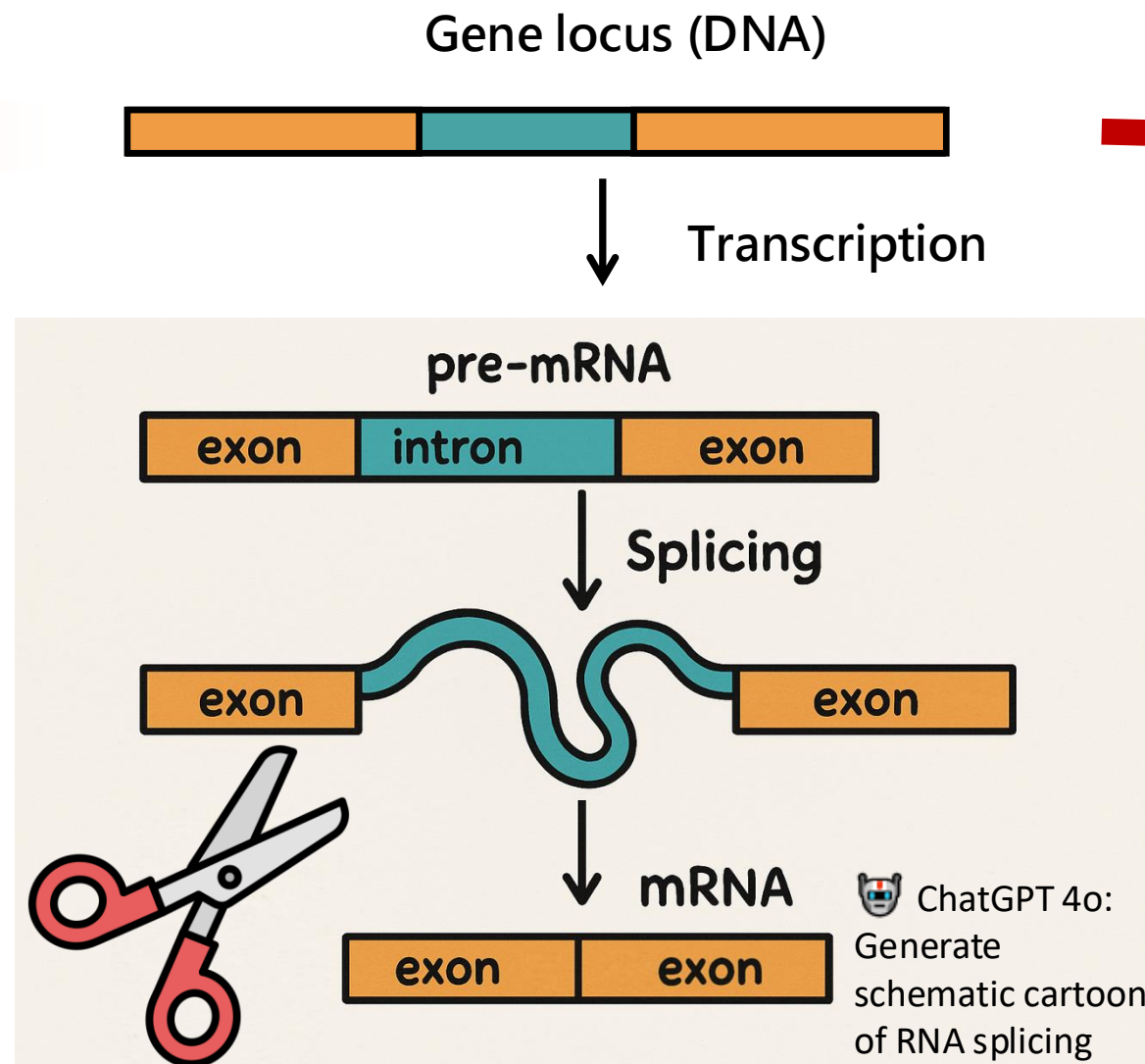
Genome Biology

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.



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. *eLife*, 2025-03.

Splice Site Prediction



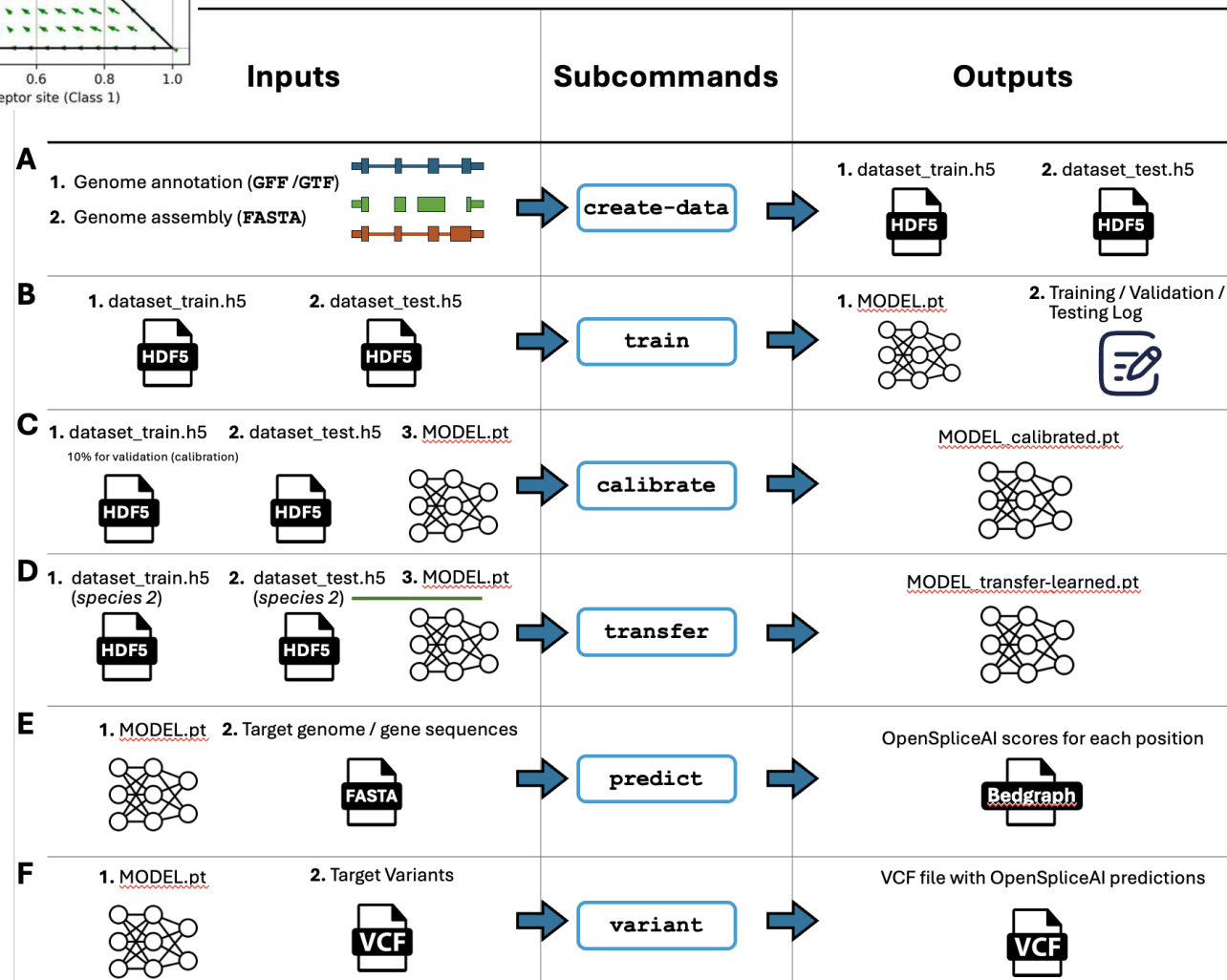
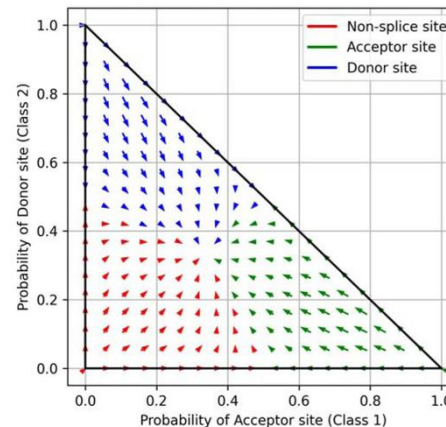
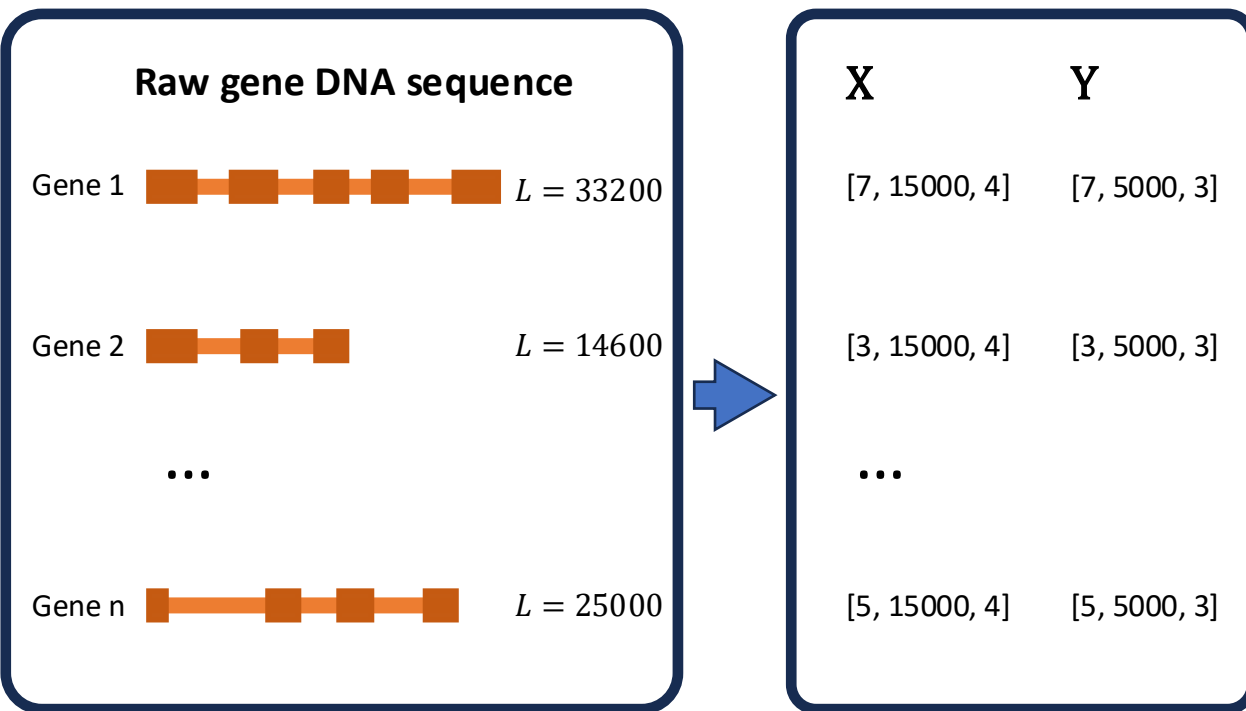
Where are the splice sites?

[illegible]

Neither: 0

OpenSpliceAI

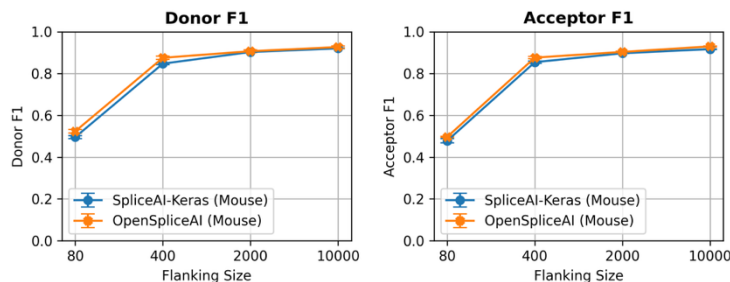
~20k protein-coding genes



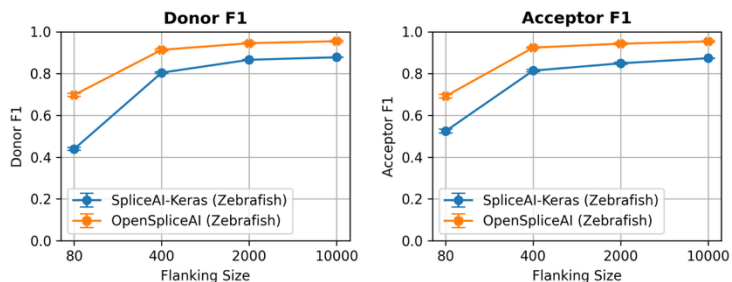
OpenSpliceAI: retrain on different species



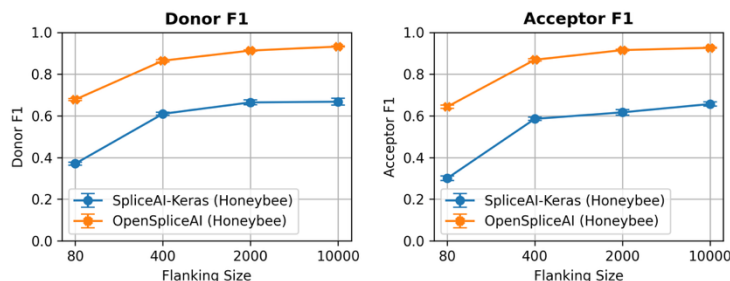
Splice site prediction metrics for Mouse



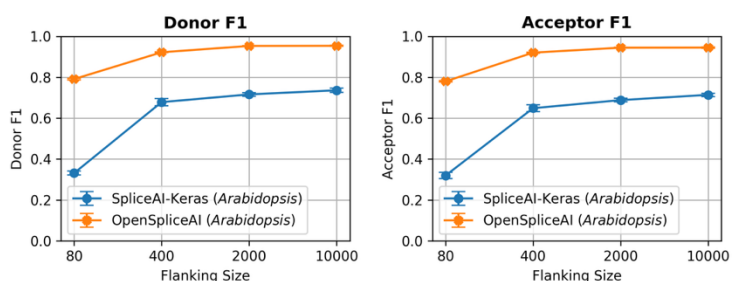
Splice site prediction metrics for Zebrafish



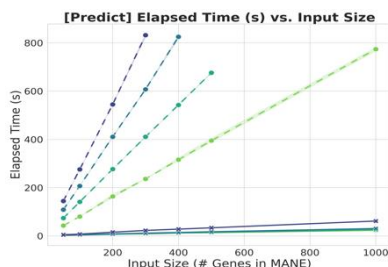
Splice site prediction metrics for Honeybee



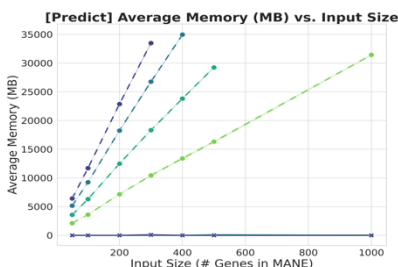
Splice site prediction metrics for Arabidopsis



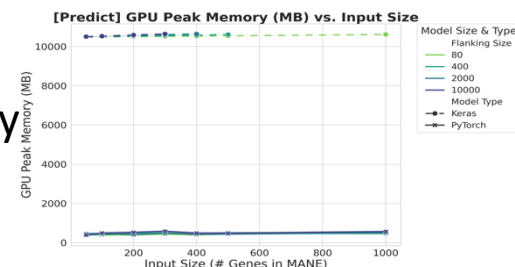
System time



Memory

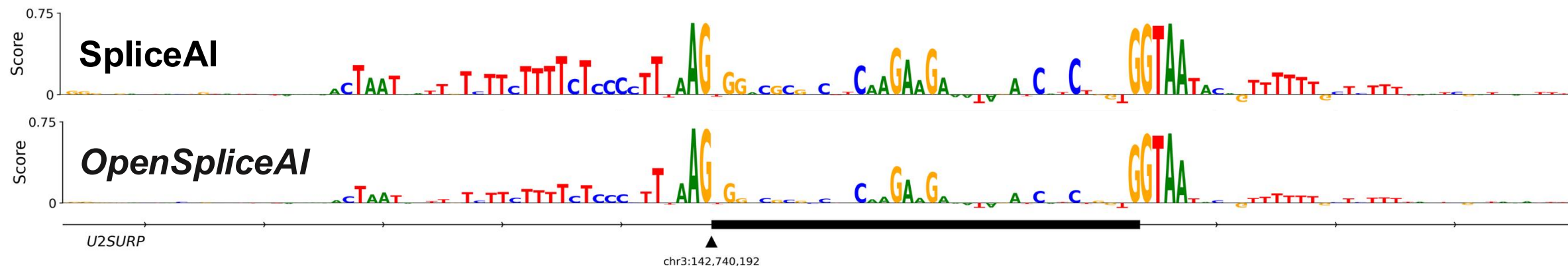


GPU Avg memory

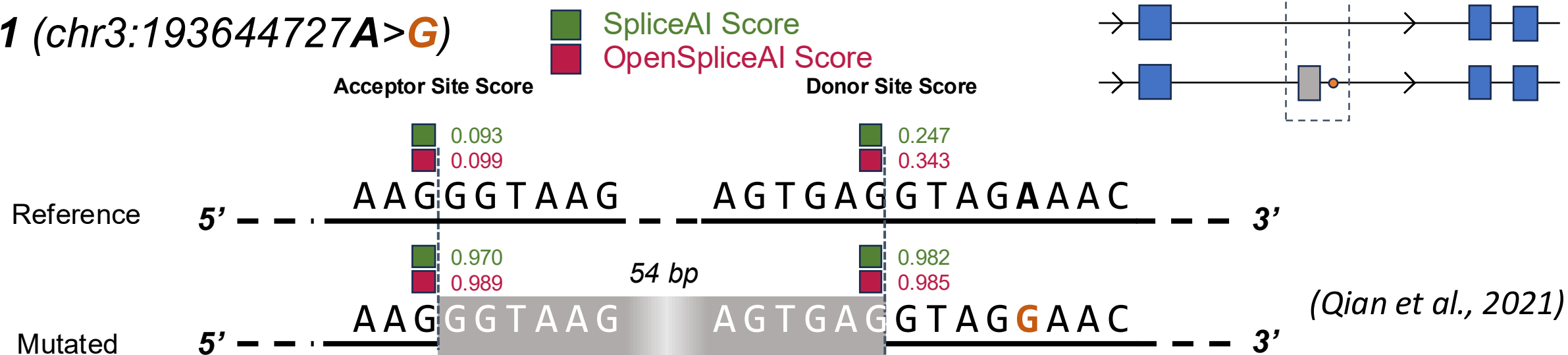




OpenSpliceAI: ISM & variant prediction



OPA1 (chr3:193644727A>G)

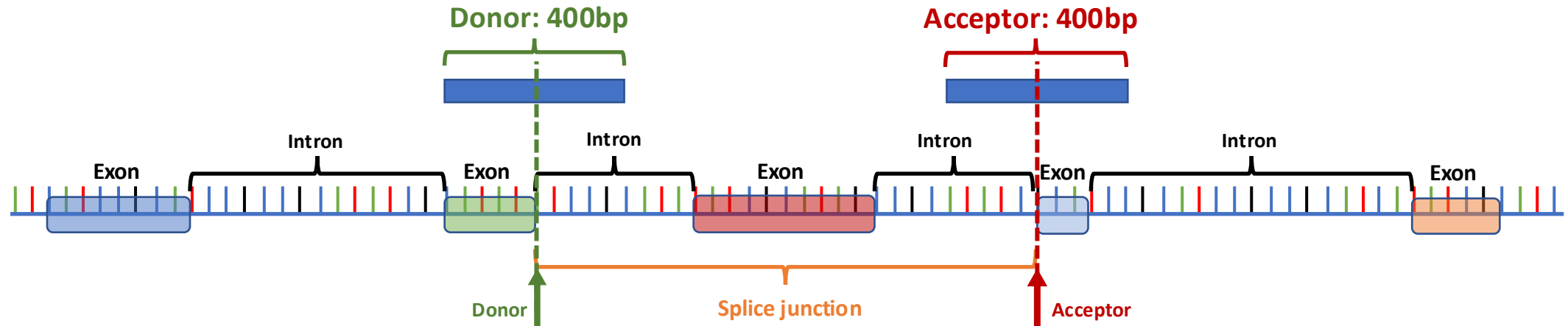




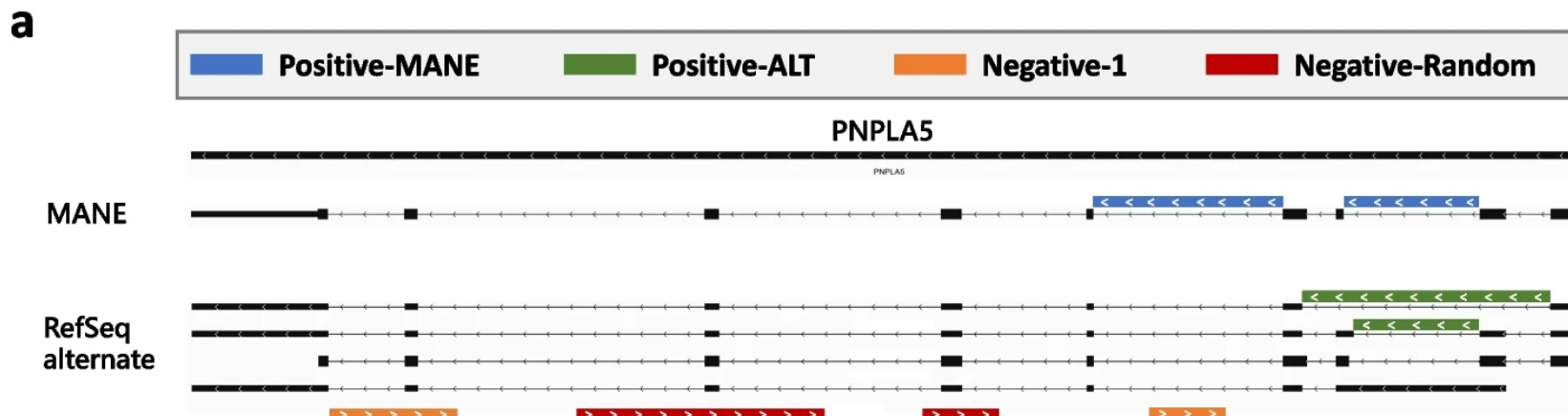
SPLM: predicting splice sites in a pair



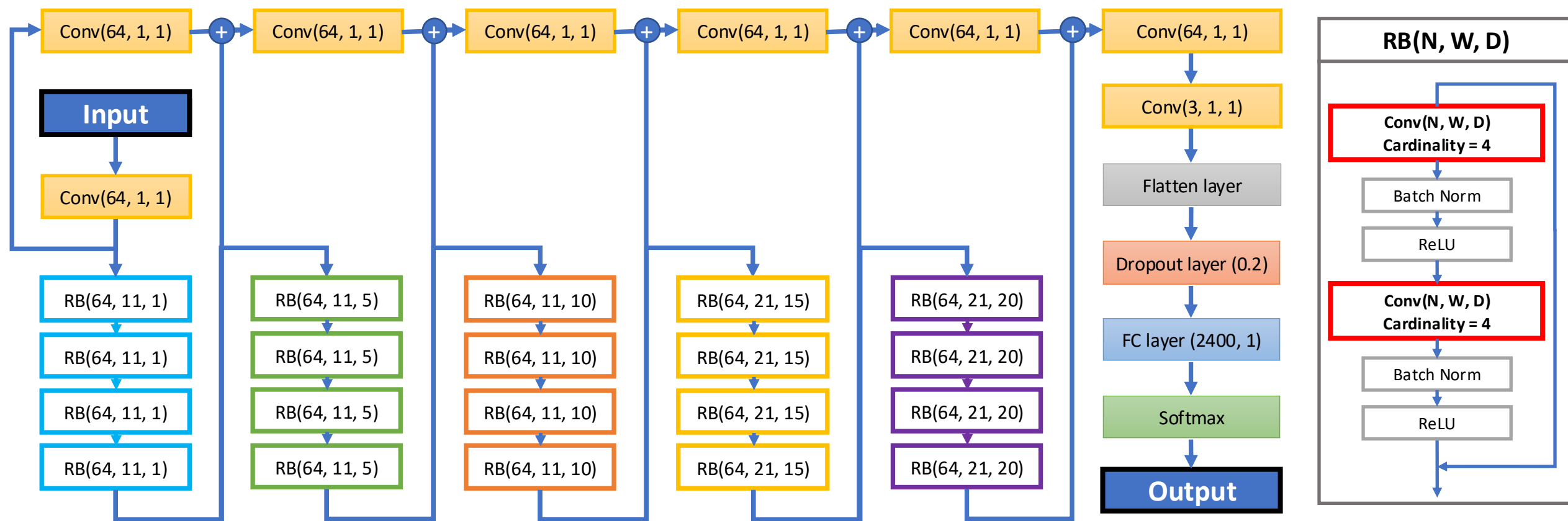
Chao, K. H., Mao, A., Salzberg, S. L., & Pertea, M. (2024). Splam: a deep-learning-based splice site predictor that improves spliced alignments. Genome biology, 25(1), 243.



Data

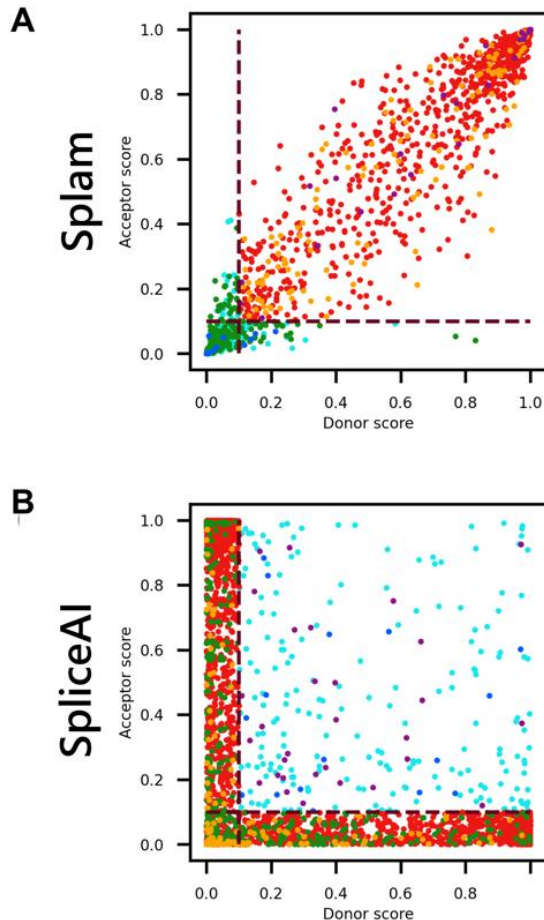


SPLAM : Splam Model Architecture

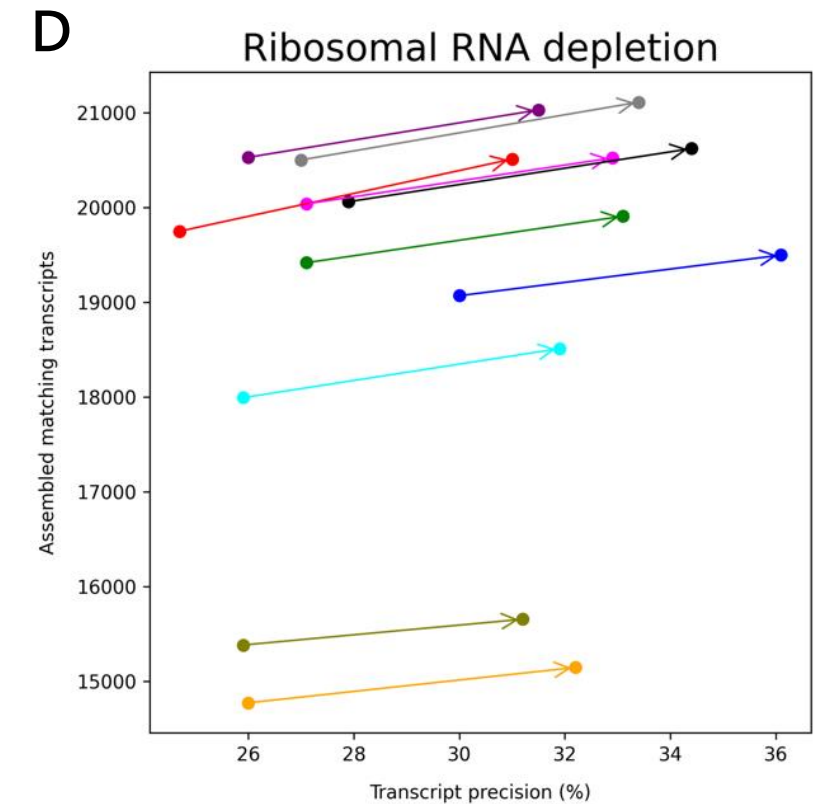
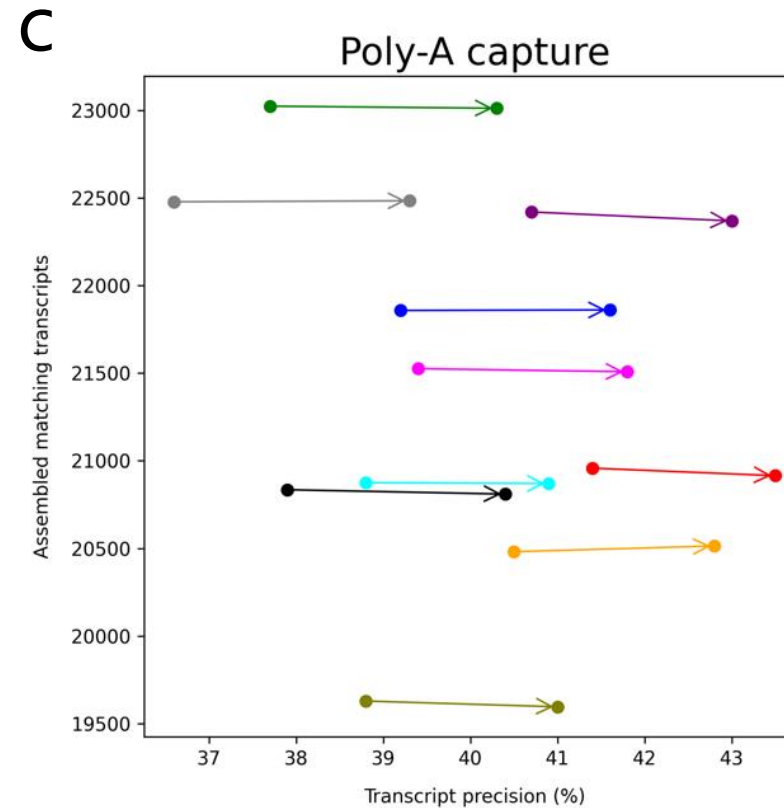


: improving transcriptome assembly

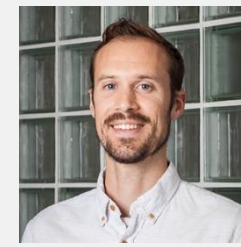
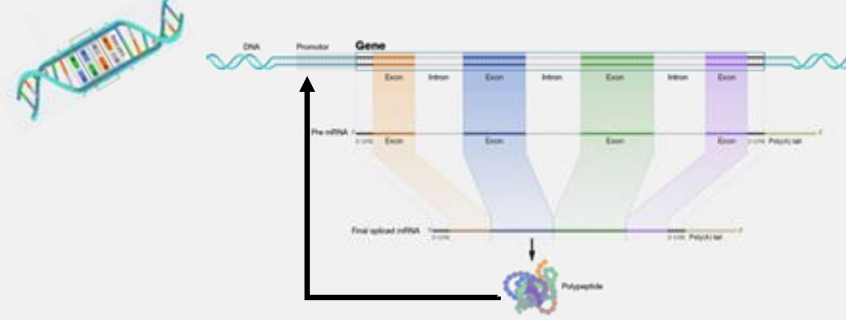
Score stability



Transcriptome assembly improvement



Part IV



David Kelley



Johannes Linder



Majed Mohamed Magzoub

Yeast RNA-Seq Prediction

- Fungal language model (LM)
- **Shorkie**: models to predict yeast RNA-Seq coverages
- Model interpretability



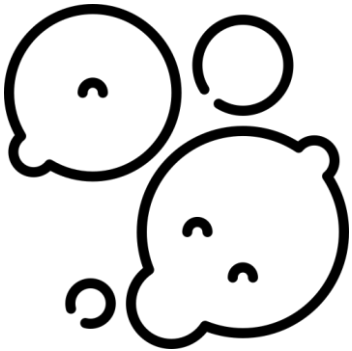
Chao, K. H., Magzoub, M., Stoops, E., Hackett, S., Linder J., * and Kelley, D. R., (manuscript in preparation). Predicting dynamic expression patterns in budding yeast with a fun-gal DNA language model

Manuscript in preparation.
Expected to be on bioRxiv soon

Questions we answer in this study:

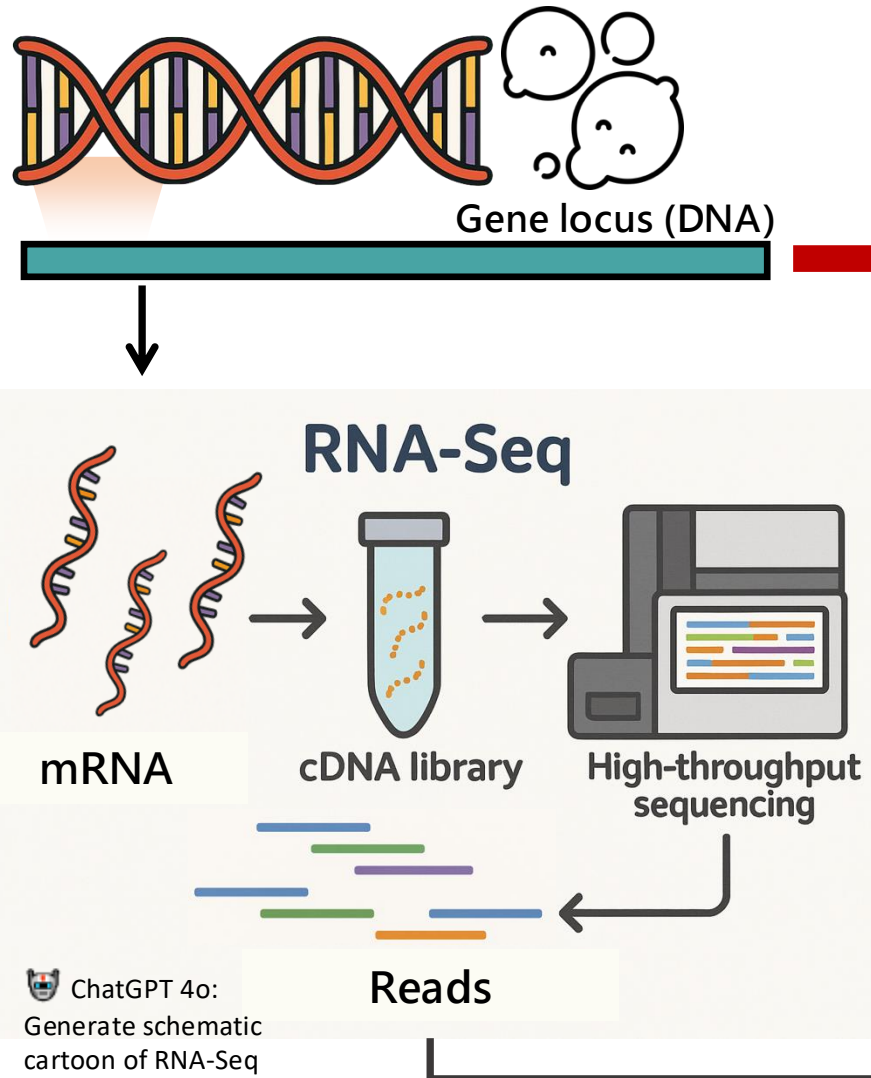
Main Goal

- Predict yeast gene expression (RNA-Seq coverage) directly from DNA sequences.

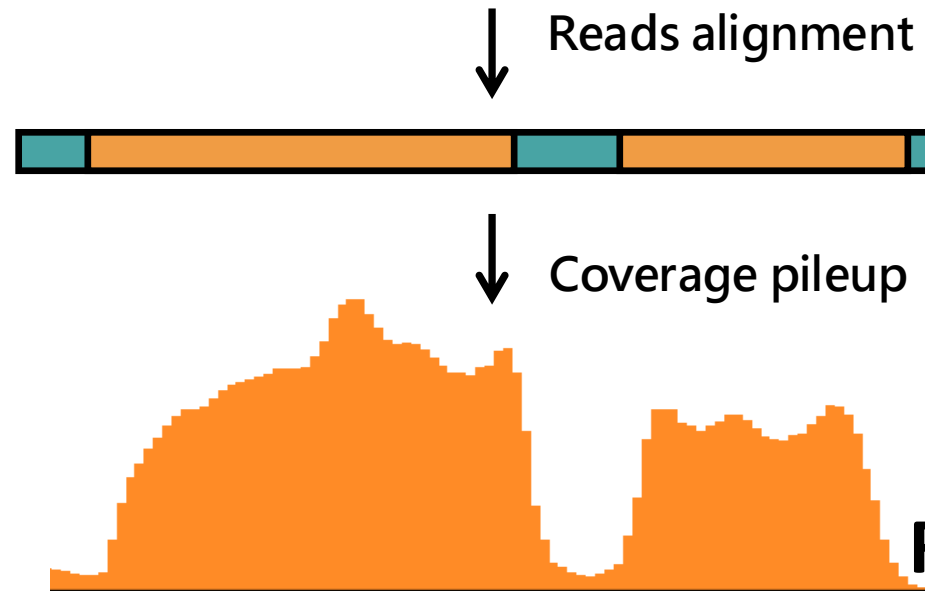


Yeast RNA-Seq Prediction

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



RNA-Seq Coverage



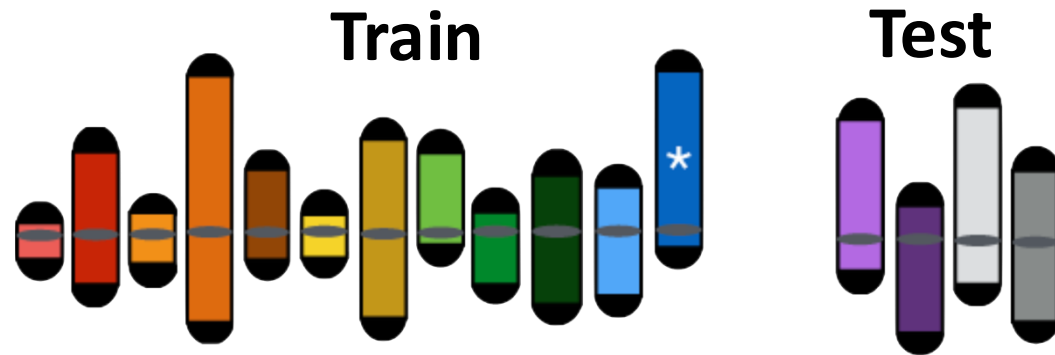
**Gene expression
model**

RNA-Seq expression?

Yeast RNA-Seq Prediction



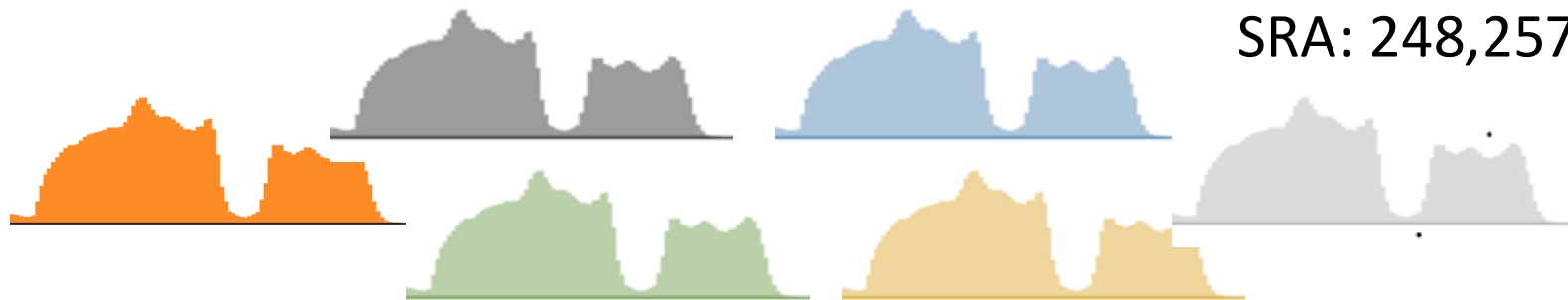
ChatGPT 4o:
Generate schematic cartoon
of RNA-Seq coverage



12 million base pairs

Genome size is too small, leading to overfitting.

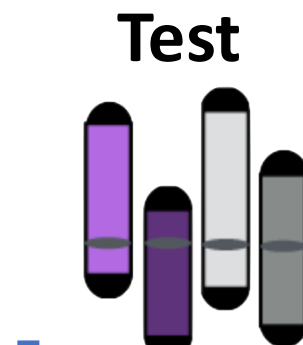
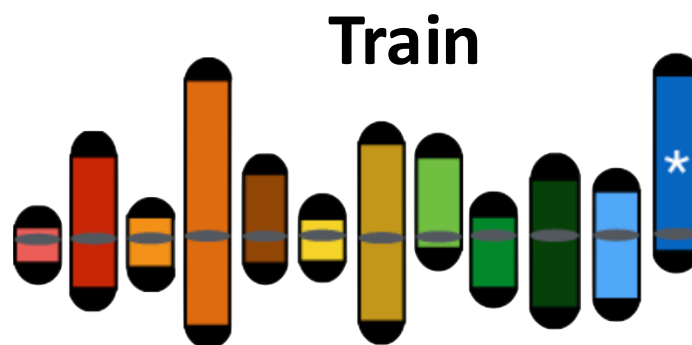
SRA: 248,257 search entries



Yeast RNA-Seq Prediction



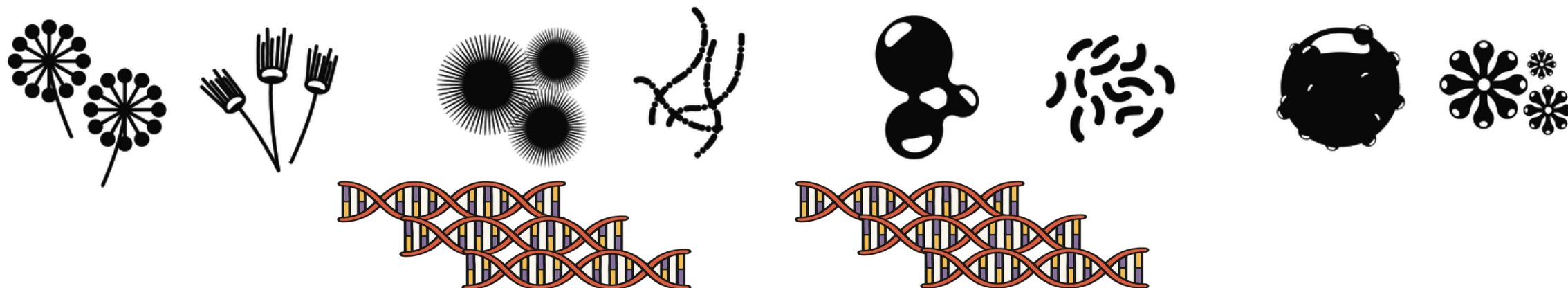
ChatGPT 4o:
Generate schematic cartoon
of RNA-Seq coverage



12 million base pairs

Fungal LM

Ensembl fungi: 1500 genomes



Yeast RNA-Seq Prediction



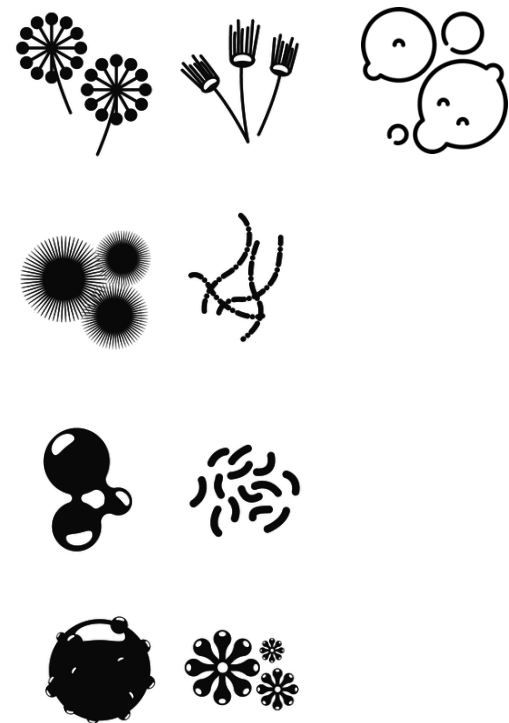
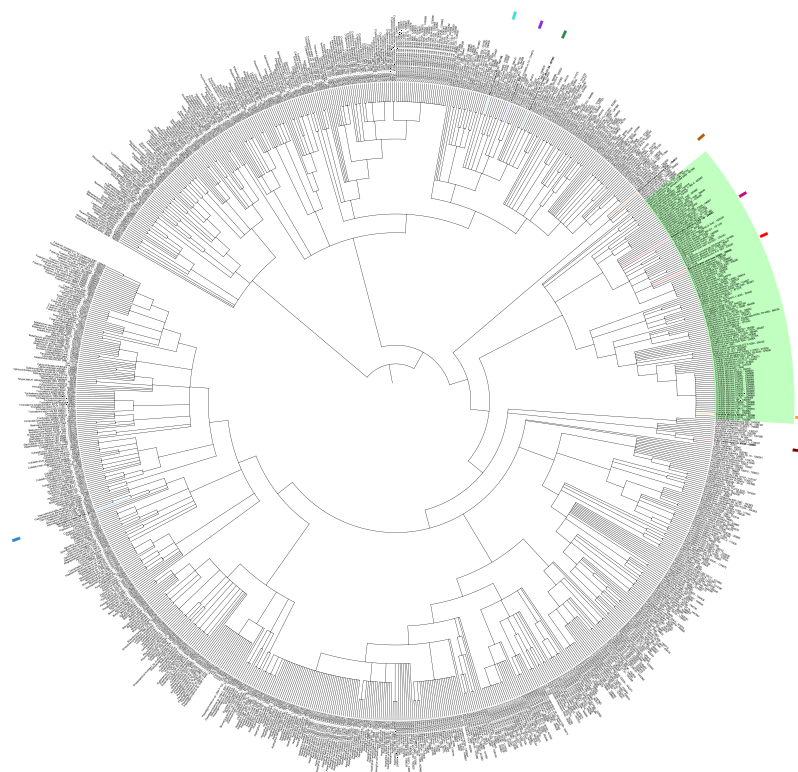
ChatGPT 4o:
Generate schematic cartoon
of RNA-Seq coverage

Fungal LM

Gene expression
model



RNA-Seq expression?

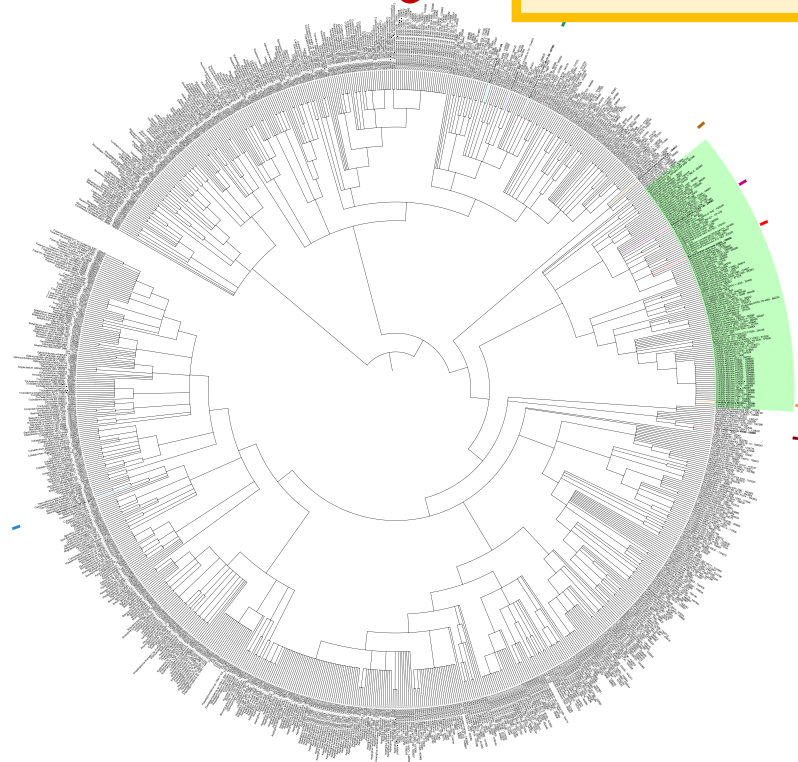
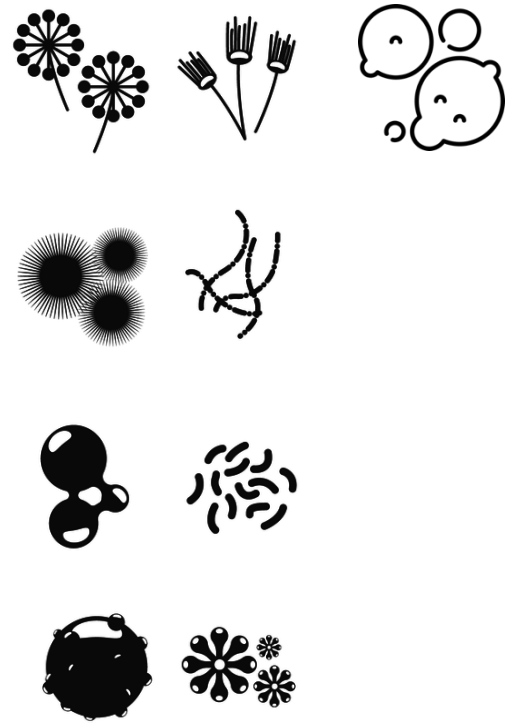


Yeast RNA-Seq Prediction



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

Fungal LM



Model
Fungal Model



RNA-Seq expression?

Deep learning-based DNA sequence model

nature methods View all journals Search

Explore content

nature > nature

Brief Communication | Published: 24 August 2015

Predicting effects of noncoding DNA- and RNA-binding protein learning-based sequence models

Jian Zhou & Olga G Troyanskaya

Troyanskaya Lab Princeton

DeepSEA
2015

CSH PRESS GENOME RESEARCH

HOME | ABOUT | ARCHIVE | SUBMIT | SUBSCRIBE | ADVERTISE | AUTH

Institution: MILTON S EISENHOWER LIBRARY Sign

Basset: learning the regulatory activity across the accessible genome with deep convolutional neural networks

David R. Kelley¹, Jasper Snoek² and John L. Rinn¹

Calico

Basset
2016

Cell

Volume 176, Issue 3, 24 January 2019, Pages 535-548.e24

Article
Predicting Splicing from Protein Binding with Deep Learning

Kishore Jaganathan^{1,6}, Sofia Kyriazopoulou Pangalotopoulos^{1,6}, Siavash Fazel Darbandi², David Knowles³, Yang Li^{1,3}, Jack Wenwu Cui¹, Grace B. Schwartz², Eric D. Chow¹, Efsthios Serafim Batzoglou¹, Stepan J. Sanders², Kyle Kai-How Farh^{1,7}

Illumina

SpliceAI
2019

Agarwal and Kelley Genome Biology (2022) 23:245
https://doi.org/10.1186/s13059-022-02811-x

RESEARCH

The genetic and biochemical detection of mRNA degradation rates in mammals

Vikram Agarwal^{1,2*} and David R. Kelley^{1*}

Calico

Saluki
2022

nature biotechnology

Explore content About the journal Publish with

nature > nature biotechnology > analyses > article

Analysis | Published: 27 July 2015

Predicting the sequence-specific DNA- and RNA-binding protein learning

Babak Alipanahi, Andrew Delong, Matthew T Weirauch & Brendan J Frey

FUToronto

DeepBind
2015

CSH PRESS GENOME RESEARCH

HOME | ABOUT | ARCHIVE | SUBMIT | SUBSCRIBE | ADVERTISE | AUTH

Institution: MILTON S EISENHOWER LIBRARY Sign

Sequential regulatory activity across chromosomes with convolutional neural networks

David R. Kelley¹, Yakir A. Reshef², Maxwell Bileschi³, David Cory Y. McLean³ and Jasper Snoek³

Calico

Basenji
2018

nature methods

Predicting 3D genome folding from Hi-C with Akita

Geoff Fudenberg^{1,2,3,4,5}, David R. Kelley^{2,5,6} and Katherine S. Pollard⁶

Calico

Akita
2020

nature methods

Explore content About the journal Publish with us

nature > nature methods > articles > article

Article | Published: 08 August 2022

scBasset: sequence-based model of chromatin accessibility using convolutional neural networks

Han Yuan & David R. Kelley

Calico

scBasset
2022

Bioinformatics

Gifford Lab MIT

DNA-TF binding
2016

nature genetics

Explore content

nature >

Article | Published: 16 July 2018

Deep learning sequence-based prediction of variant effects on expression and protein binding

Jian Zhou, Chandra L. Theesfeld, Kevin Yao, Kathleen M. Gifford

Troyanskaya Lab Princeton

ExPecto
2018

ARTICLES

https://doi.org/10.1038/s41592-021-01252-x

OPEN
Effective gene expression prediction from DNA sequence by integrating chromatin accessibility

Žiga Avsec^{1,2,3,4}, Vikram Agarwal^{1,4}, Daniel V. Agnieszka Grabska-Barwinska¹, Kyle R. Taylor¹ and David R. Kelley^{2,3,4}

DeepMind + Calico

Enformer
2021

nature genetics

Article

Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation

Received: 28 August 2023
Accepted: 4 December 2024
Johannes Linder^{1,2,3,4}, Divyanshi Srivastava¹, Han Yuan¹, Vikram Agarwal^{1,2,3,4} & David R. Kelley^{1,2,3,4}

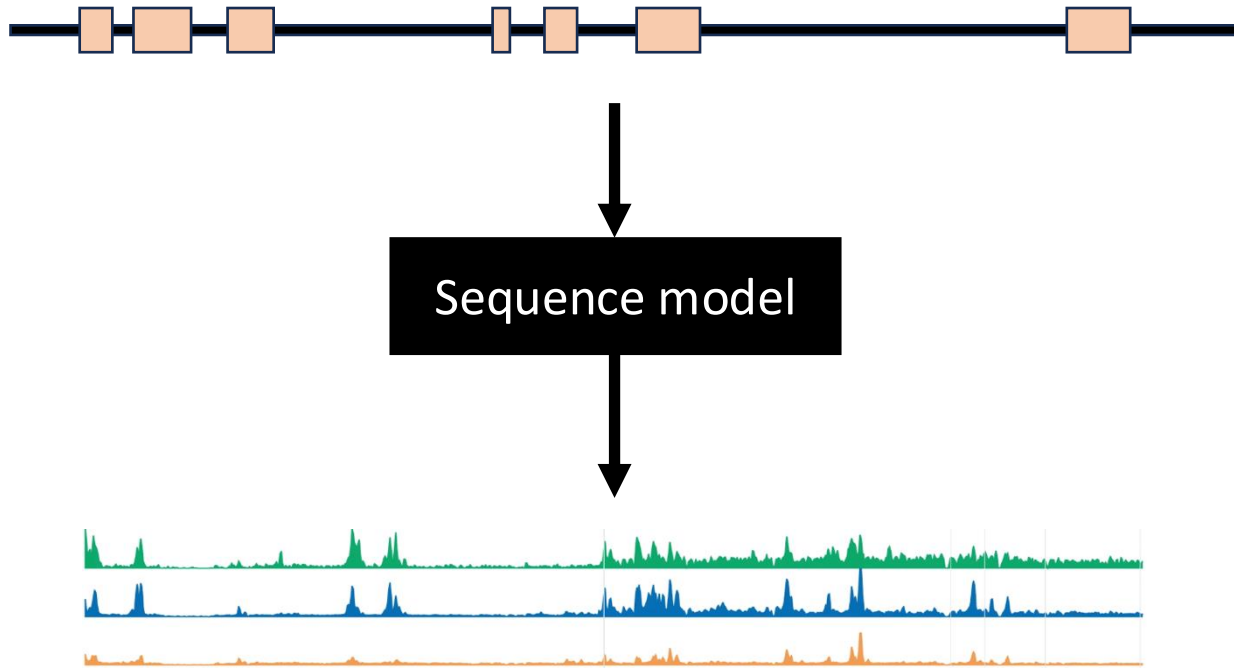
Calico

Borzoï
2025

Supervised learning

Input: DNA sequences

Output: Genomics tracks



Protein Language model (PLM)

RESEARCH ARTICLE | BIOLOGICAL SCIENCES |

Facebook

Biological structure and function
from scaling up
million protein s

First LM attempt
PNAS 2020

Alexander Rives, Joshua Meier, Tom Sercu, +7, and Rob Fergus Authors Info & Affiliations

Article | Open access | Published: 27 July 2022

Höcker Lab

ProtGPT2 is a deep unsupervised
for protein design

Noelia Ferruz, Steffen Schmi

Nature Communications 13, Art

ProtGPT2
Nat Commun 2022

RESEARCH ARTICLE | STRUCTURE PREDICTION

Meta

Evolutionary-scale prediction
tein structure with a langua

ZEMING LIN, HALIL AKIN, [...] AND ALEXANDER RIVES +12 au

ESMFold
Science 2023

Transformer protein language

Facebook

unsup

Roshan Ra
Rives

Transformer protein LM
ICLR 2021

Article | Published: 03 October

Harvard + Columbia

Single-sequence prot
using a language r

Ratul Chowdhury, Nazim B

Anant Kharkar, Koushik Ro

Zhang, George M. Church,

AminoBERT
Nat Biotechnol 2022

Primer | Published: 15 February 2024

Profluent

Designing protei

Jeffrey A. Ruffolo & Ali M

Nature Biotechnology 42,

PLM
Nat Biotechnol 2024

arXiv:2006.15222 (cs)

[Submitted on 26 Jun 2020 (v1), last revised 28 Mar 2021]

BERTology Meets Biology:
Attention in Protein Langu

Jesse Vig, Ali Madani, Lav R. Varshney, Caiming Xiong, R

Salesforce + UIUC

BERTology
ICLR 2021

Article | Published: 26 January 2023

Salesforce

Large language models generate f
protein sequences across diverse f

Ali Madani, Ben Krause, B

Mohr, James M. Holton, Jos

Sun, Richard Socher, James

ProGen
Nat Biotechnol 2023

arXiv:2402.16445 (cs)

[Submitted on 26 Feb 2024 (v1), last revised 16 Jul 2024 (this ver

ProLLaMA: A Protein Lan
Multi-Task Protein Lang

Liuzhenghao Lv, Zongying Lin, Hao Li, Yuyang Liu, J
Yonghong Tian

Peking Uni

ProLLaMA
arxiv, 2024

SYNTHESIS · Volume 12, Issue 6, P654-669.E3, June 16, 2

Open Access

Learning the protein
structure, and functi

Tristan Bepler, 1,2,3, Bor

MIT

Review
Cell Systems 2021

ARTICLE · Volume 14, Issue 11,

ProGen2: Exploring th
language models

Erik Nijkamp, 1,5 · Jeffrey A. Ruffo

· Ali Madani, 1,3,6

Salesforce + Profluent

ProGen2
Cell System 2023

~ 6000 citations

DNA Language model

DNABERT: pre-trained Bidirectional Encodings of Representations from DNA language in genomic context
Yanrong Ji, Zhihan Zhou, Han Liu et al.

DNABERT
Bioinformatics 2021

DNABERT-2: Efficient Foundation Model for Genomic Data
Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Joshi et al.

DNABERT-2
ICLR 2024

Sequence representation from DNA language models
Eric Nguyen, Michael Poli, Matthew G. Durrant, Armin Thomas, Jeremy Sullivan, Madelena Y. Ng, Ashley Lewis, Aman Patel, Stephen A. Baccus, Tina Hernandez-Boussard, Christopher J. M. Brown et al.

Evo
bioRxiv 2024

RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY | 8
DNA language models are powerful predictors of genome-wide variation
Gonzalo Benegas, Sanjit Singh Batra, and Yun S. Song et al.

GPN
PNAS 2023

Article | Open access | Published: 23 July 2024
DNA language model captures genomic context in the human genome
Melissa Sanabria, Jonas Hirsch, Pierre Schumacher et al.

GROVER
Nat Mach Intell 2024

[Submitted on 5 Mar 2024]
Caduceus: Bi-Directional Range DNA Sequence Modeling
Yair Schiff, Chia-Hsiang Kao, Aaron Gokaslan, Tri Dao et al.

Caduceus
ICML 2024

The Nucleotide Transformer Foundation Models for Genomic Data
Hugo Dalla Bernardini, Adam Henry, Bernardo P. L. F. de Souza, Marie Lopez et al.

Nucleotide Transformer
bioRxiv 2023

Article | Open access | Published: 03 April 2024
Genomic language model predicts regulatory elements
Yunha Hwang, Andre L. Cornman, Elizabeth R. Girguis et al.

Genomic LM
Nat Commun 2024

Cross-species nucleotide representation
Jingjing Zhai, Aaron Gokaslan, Yair Schiff, Michelle C. Stitzer, M. Cinta Romera et al.

PlantCaduceus
bioRxiv 2024

arXiv:2306.15794 (cs)
[Submitted on 27 Jun 2023 (v1), last revised 14 Nov 2023 (this version, v2)]
HyenaDNA: Long-Range Context Modeling at Single Nucleotide Resolution
Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Aman Patel, Clayton Rabideau, Stefano Massaroli, Yoshinori Baccus, Chris Ré et al.

HyenaDNA
NeurIPS 2023

Research | Open access | Published: 02 April 2024
Species-aware DNA language models capture regulatory elements across species
Alexander Karol, Julien Gagneur et al.

Species-aware DNA LM
Genom Biol 2024

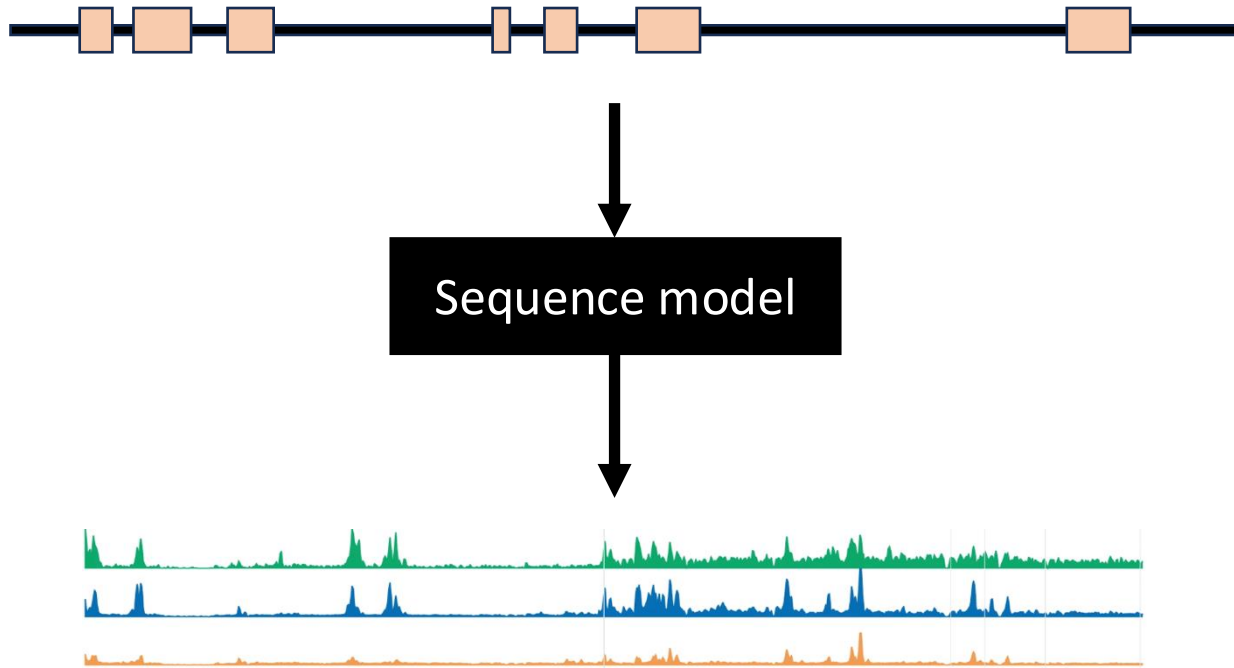
Nucleotide dependency analysis of DNA language models
Yair Schiff, Aaron Gokaslan, Yair Schiff, Michelle C. Stitzer, M. Cinta Romera et al.

Nucleotide dependency
bioRxiv 2024

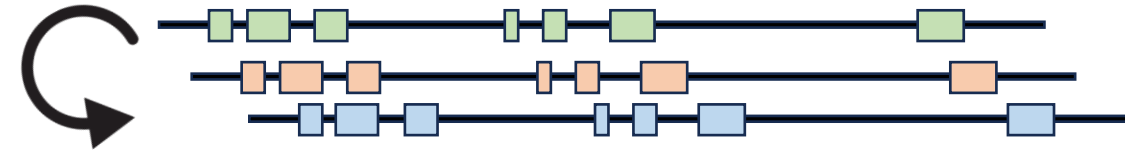
Supervised learning

Input: DNA sequences

Output: Genomics tracks



Stage 1 Self-supervised learning



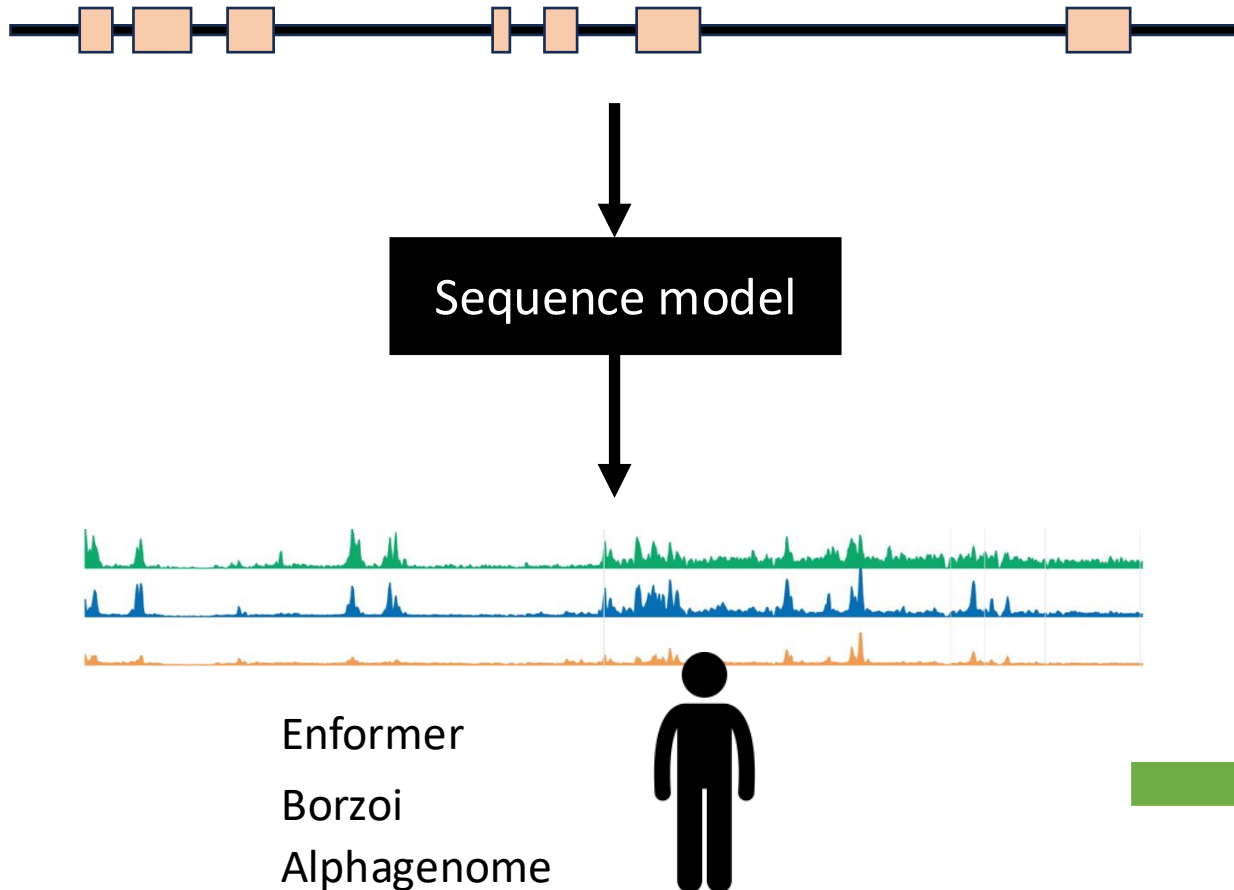
Language model (LM) /
Foundation model

Stage 2 Fine-tuning LM



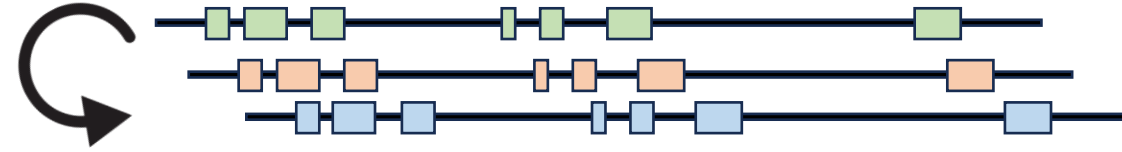
Enformer
Borzoï

Supervised learning



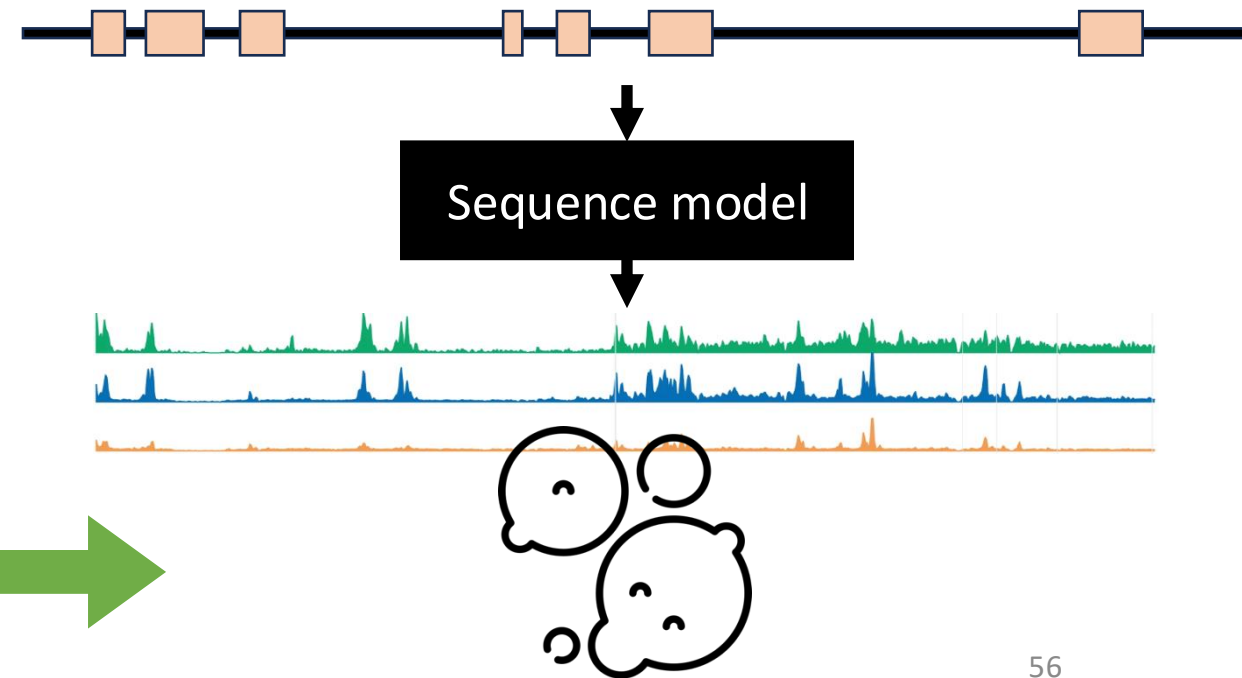
Stage 1

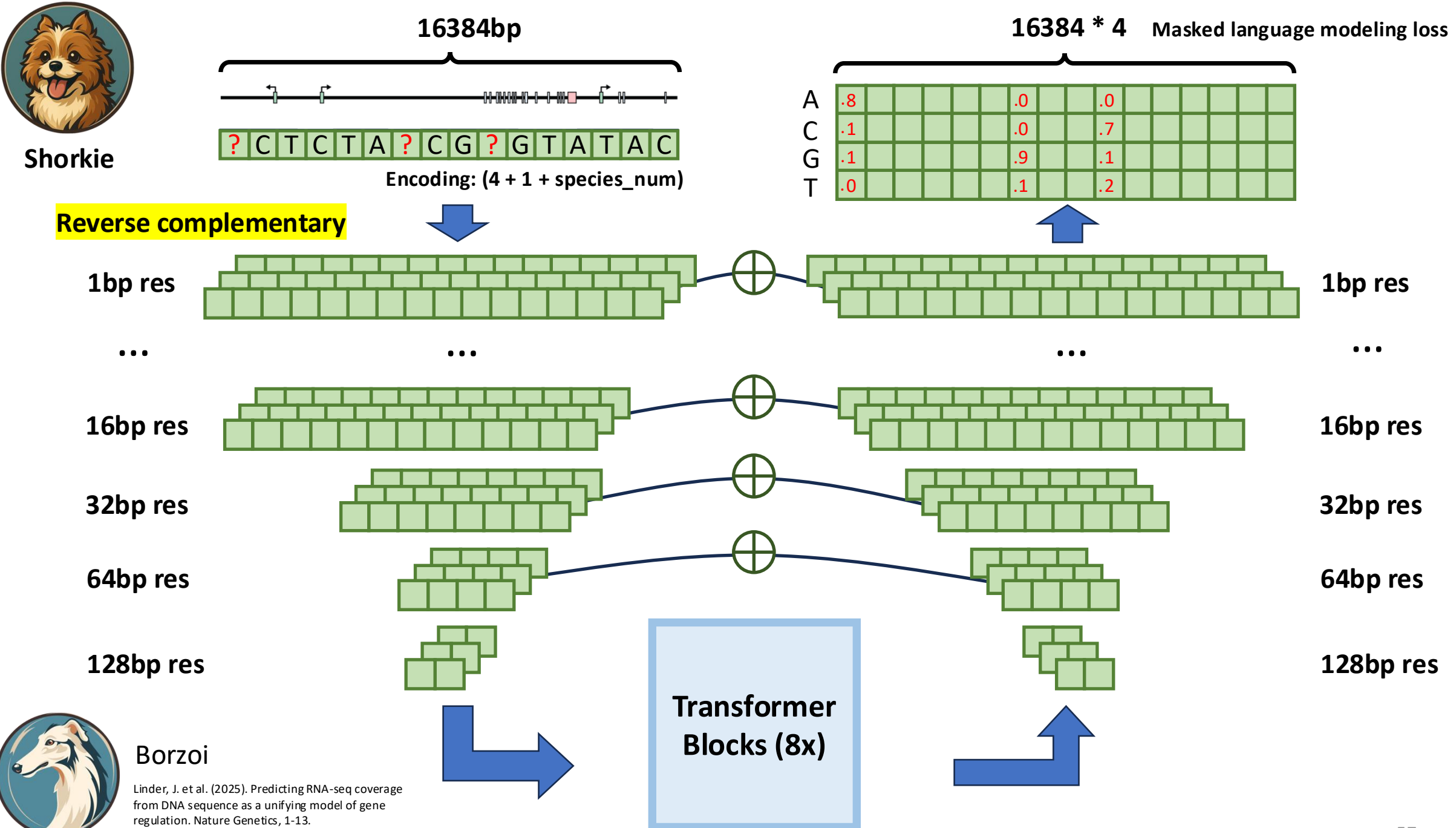
Self-supervised learning



Stage 2

Fine-tuning LM





RAP1



FHL1



LM predicts motifs in RP gene Promoter regions

450 nt

50 nt



Introduction

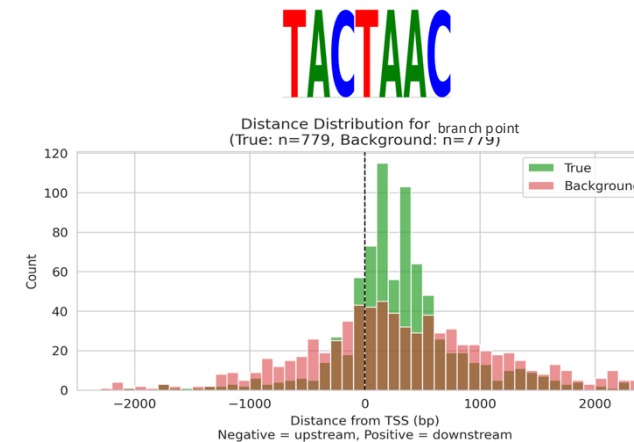
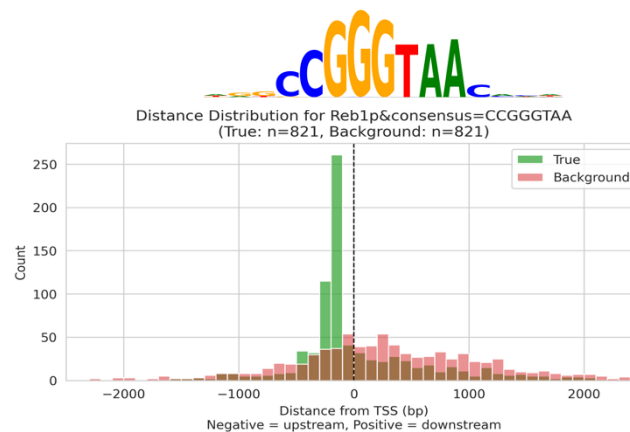
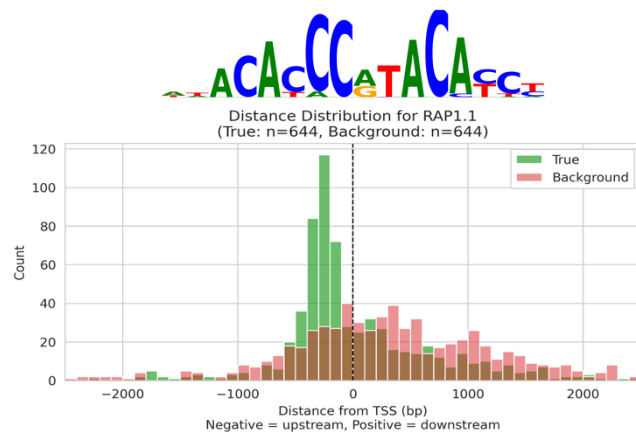
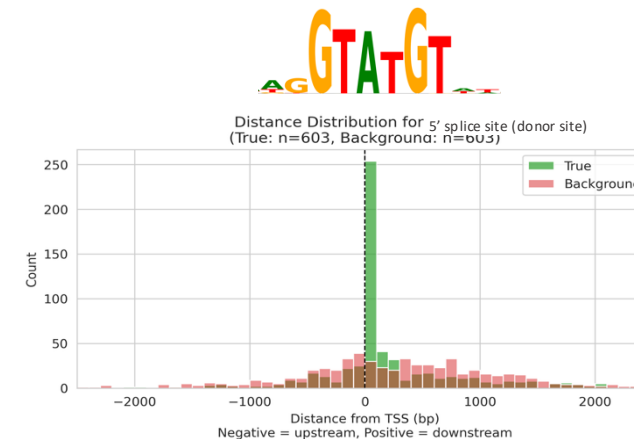
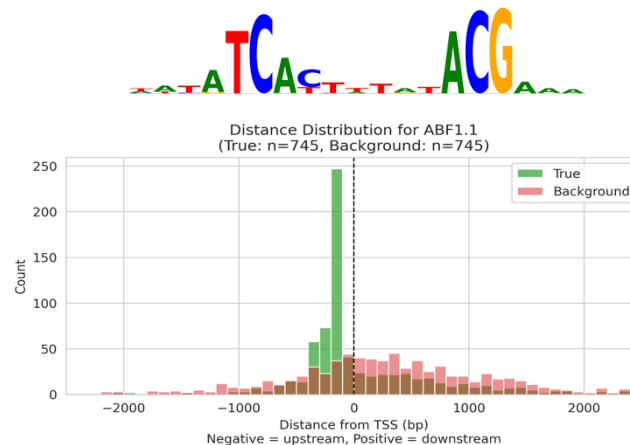
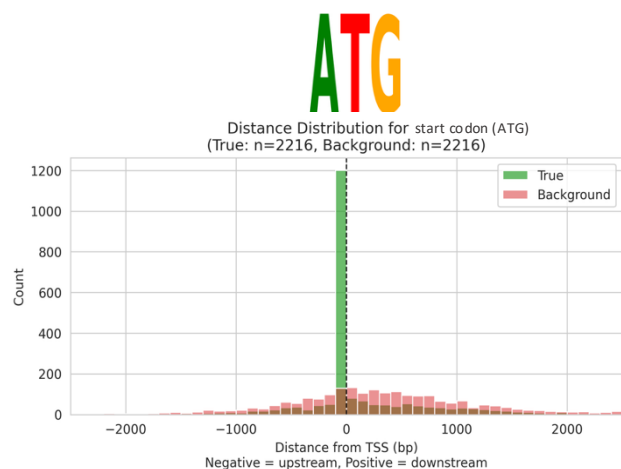
Fungal LM

Transfer learning

Interpretability

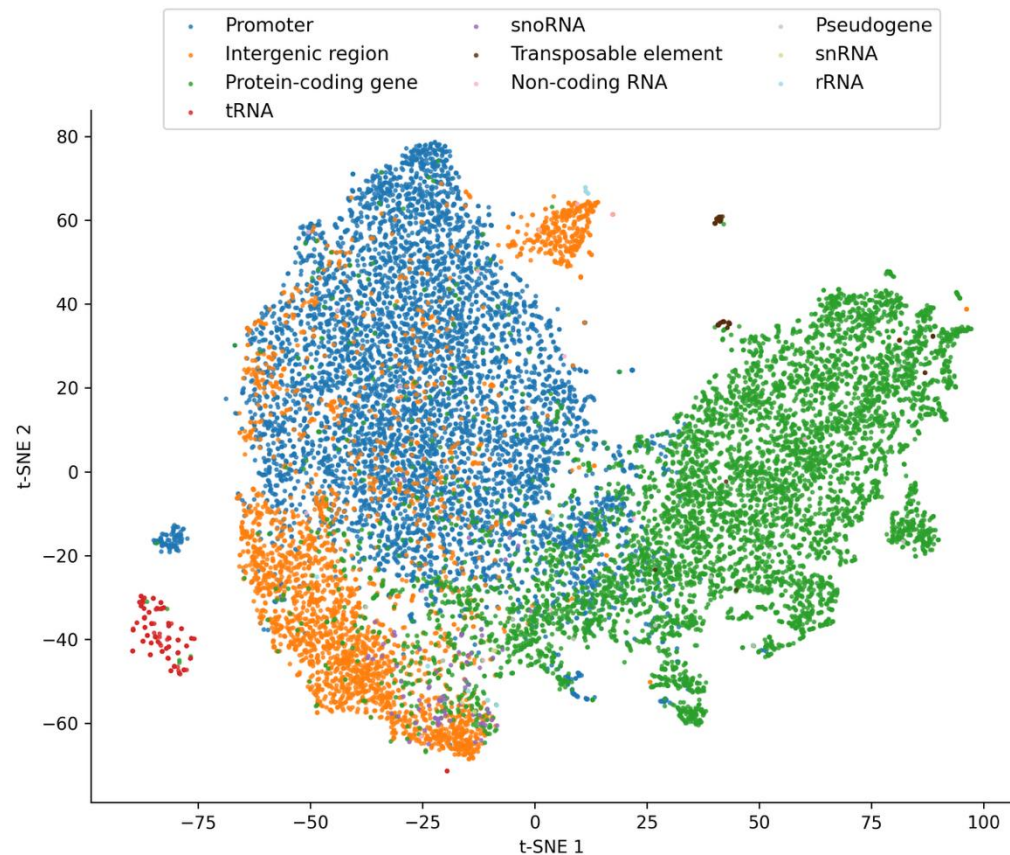
Applications

These motifs are enriched in promoter regions

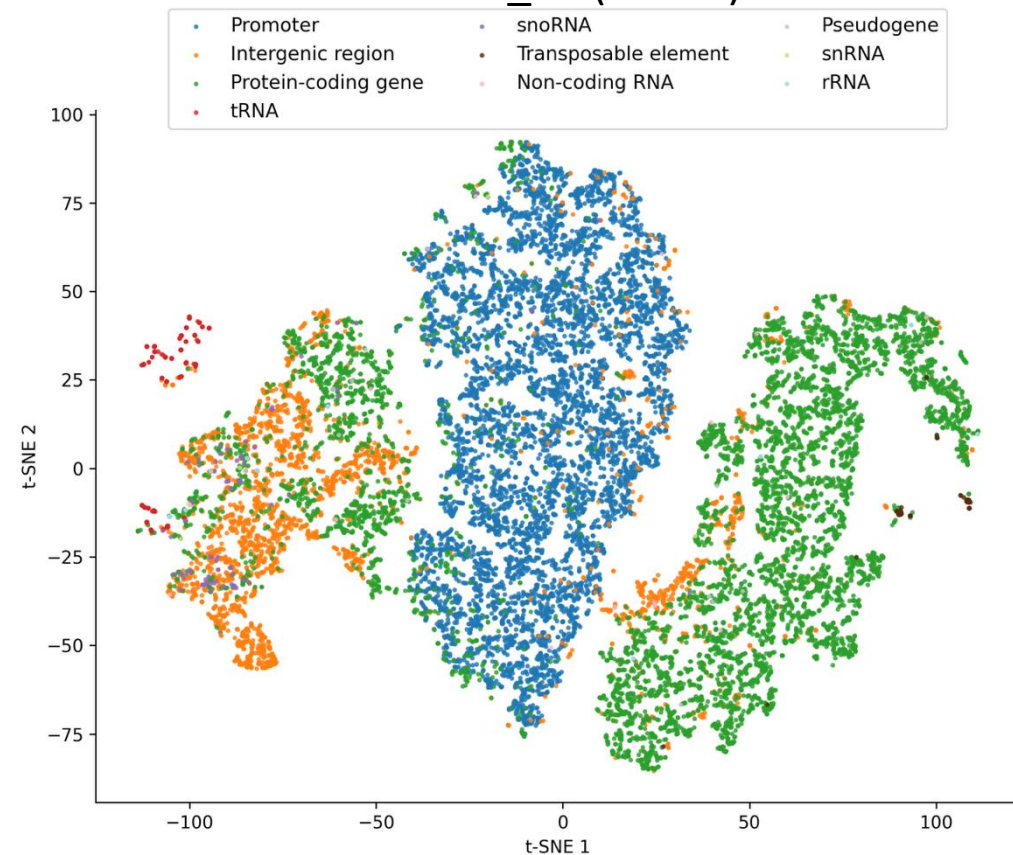


t-SNE Projection of Fungal LM Encodings for Genomic Regions

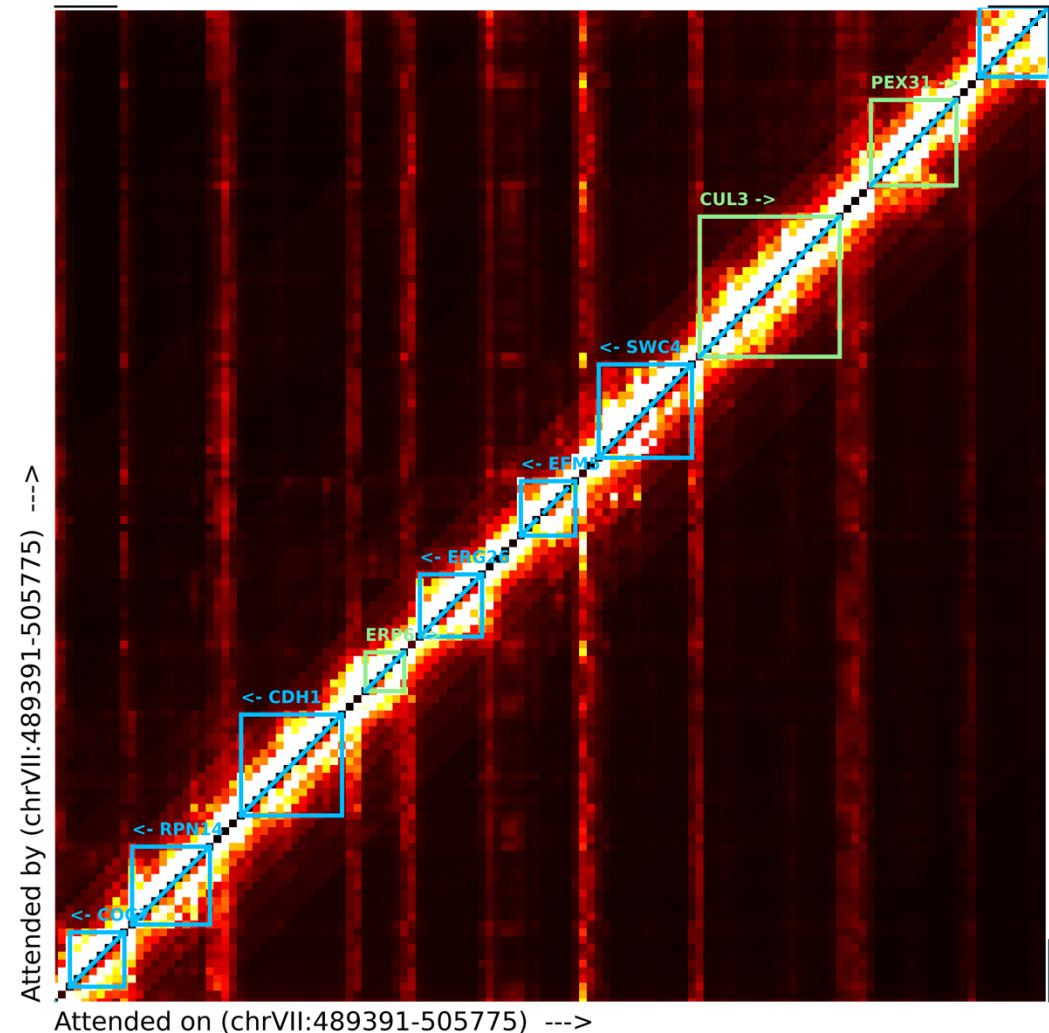
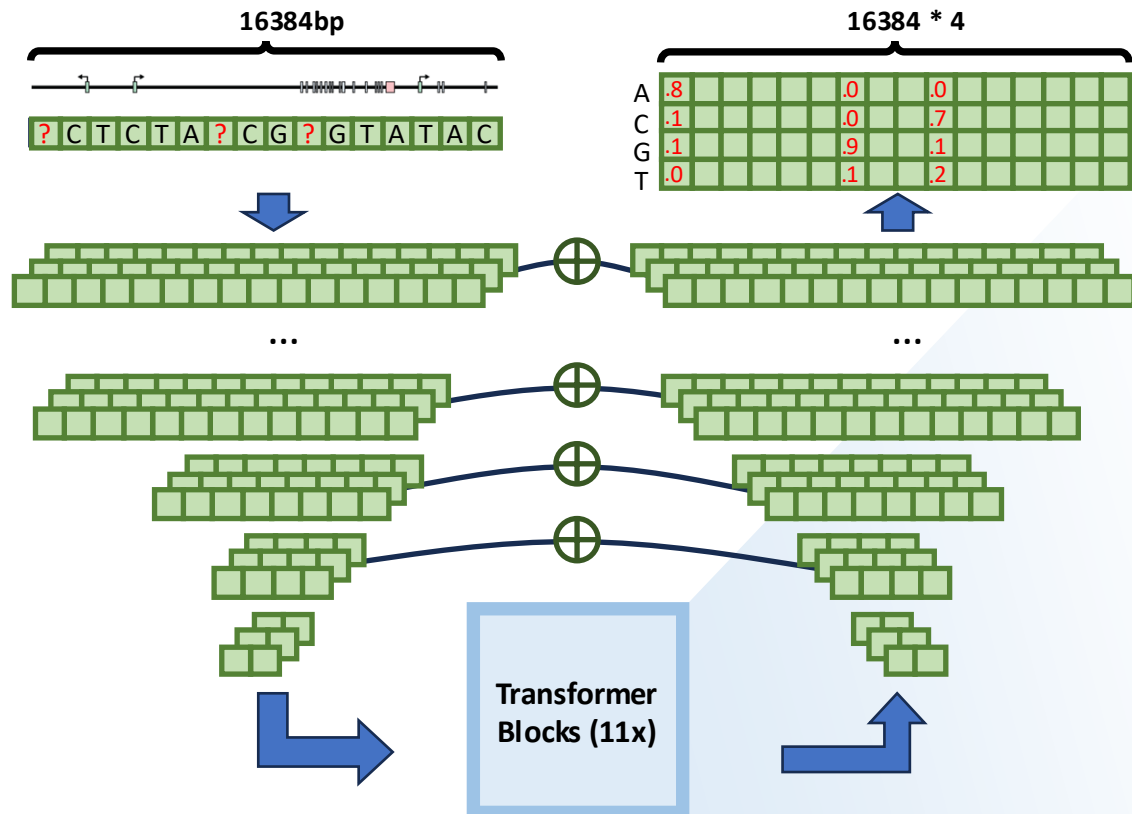
First self-attentions layer output
multihead_attention (MultiheadAttention)



Last fully connected layer before the model output
dense_29 (Dense)



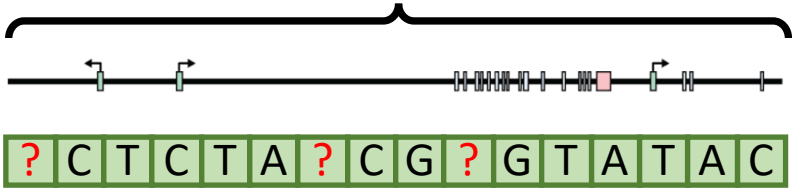
Self-attention maps





Shorkie

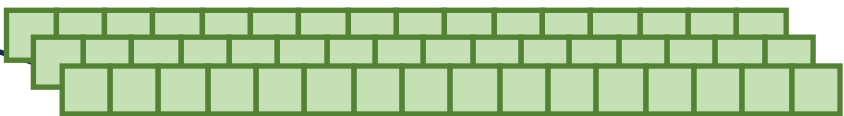
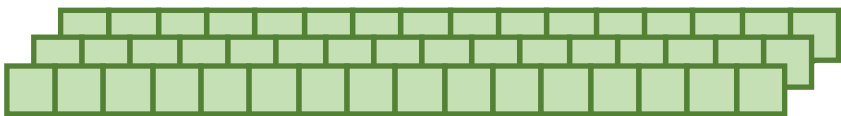
16384bp



16384 * 4

A	.8					.0			.0					
C	.1					.0			.7					
G	.1					.9			.1					
T	.0					.1			.2					

1bp res



1bp res

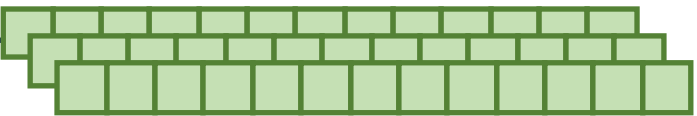
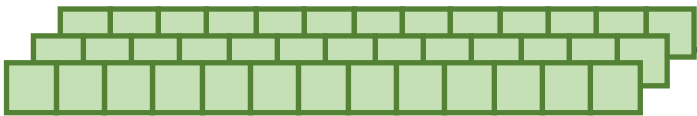
...

...

...

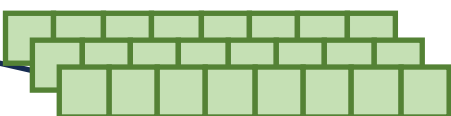
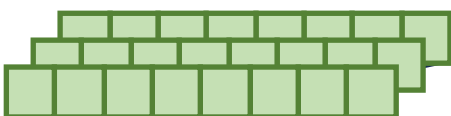
...

16bp res



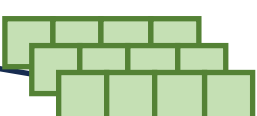
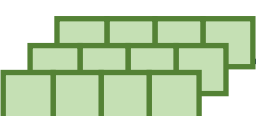
16bp res

32bp res



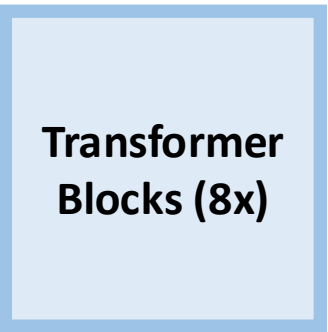
32bp res

64bp res



64bp res

128bp res

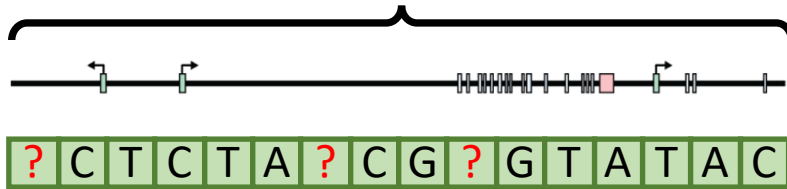


128bp res

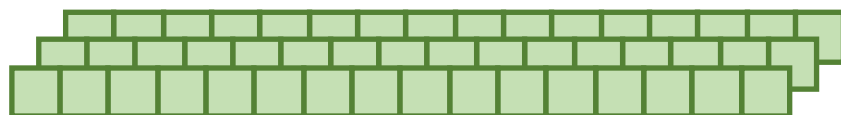


Shorkie

16384bp



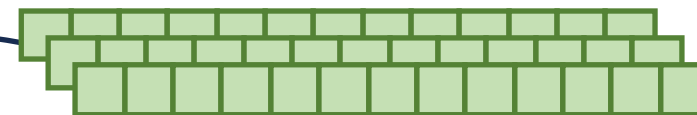
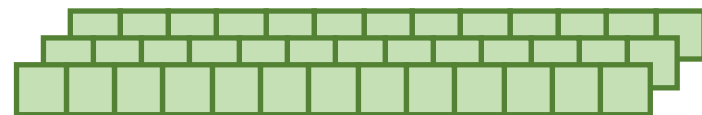
1bp res



...

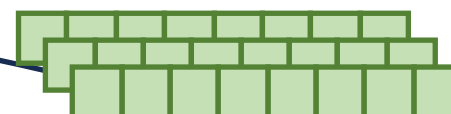
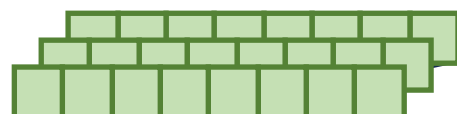
...

16bp res



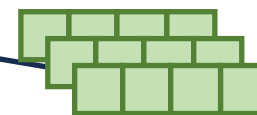
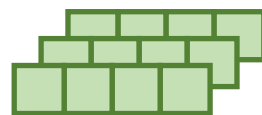
16bp res

32bp res



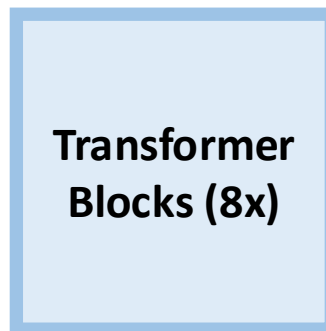
32bp res

64bp res

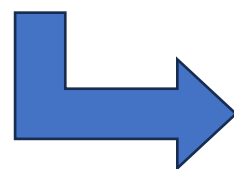


64bp res

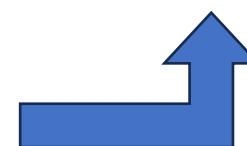
128bp res



128bp res

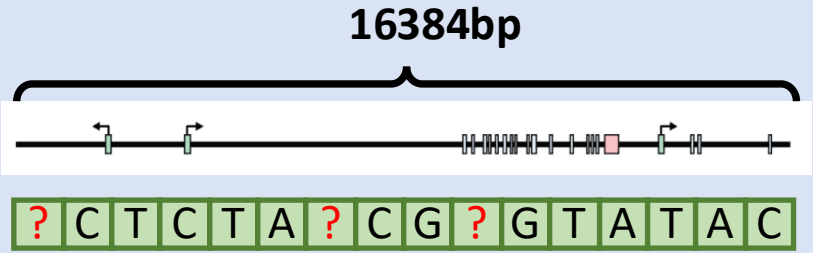


Transformer
Blocks (8x)

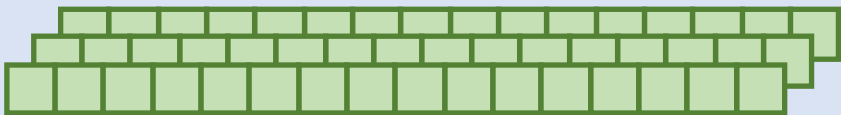




Shorkie



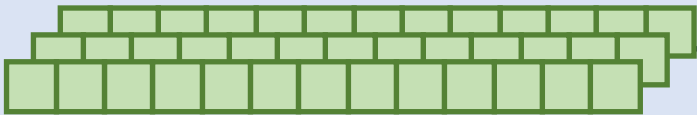
1bp res



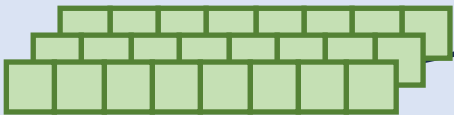
...

...

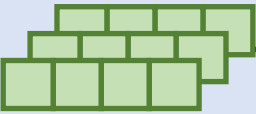
16bp res



32bp res

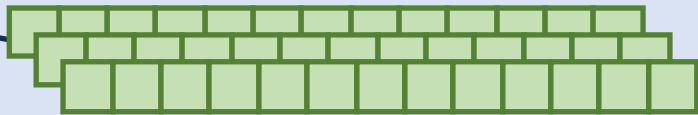


64bp res

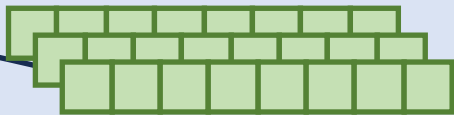


Transformer
Blocks (8x)

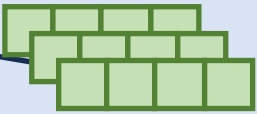
16bp res



32bp res



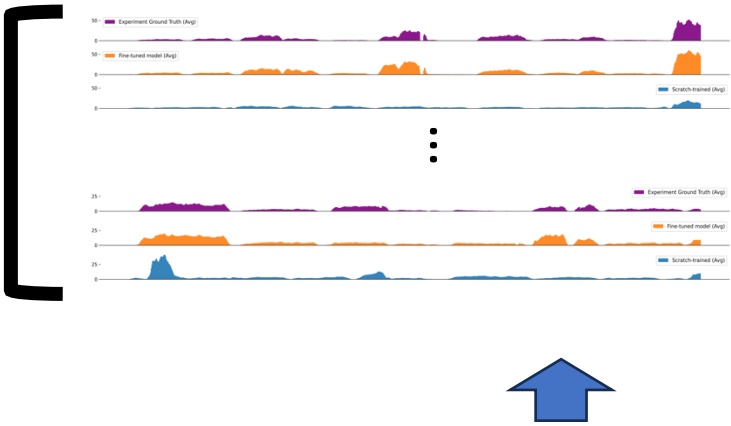
64bp res



128bp res



Coverage Tacks (896 * 2488)



ChIP-exo	(1128)
Histone marks	(20)
RNA-Seq	(3054)
1000-strains RNA-Seq	(1014)

Experiment Ground Truth (Avg)

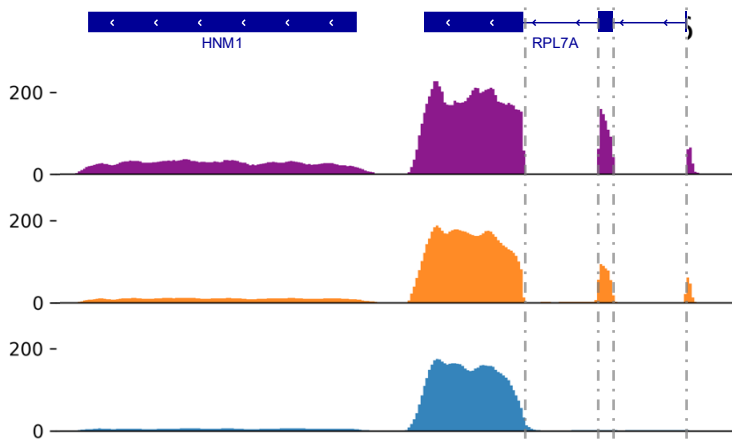
Fine-tuned model (Avg)

Scratch-trained (Avg)

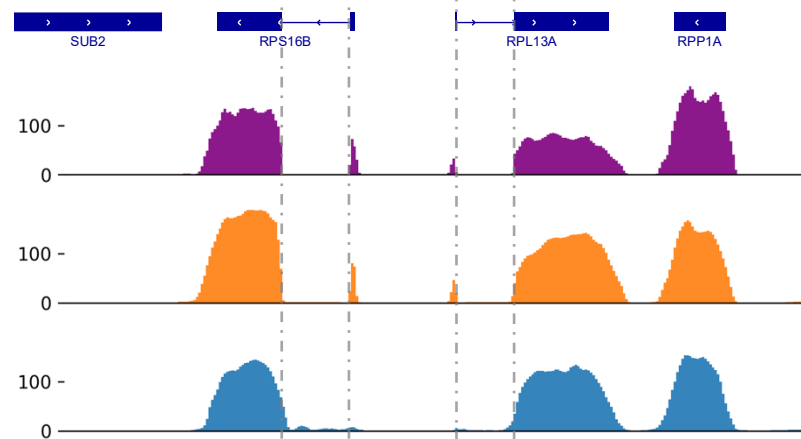
Shorkie vs Random Initialization (Test set)

RNA-Seq tracks

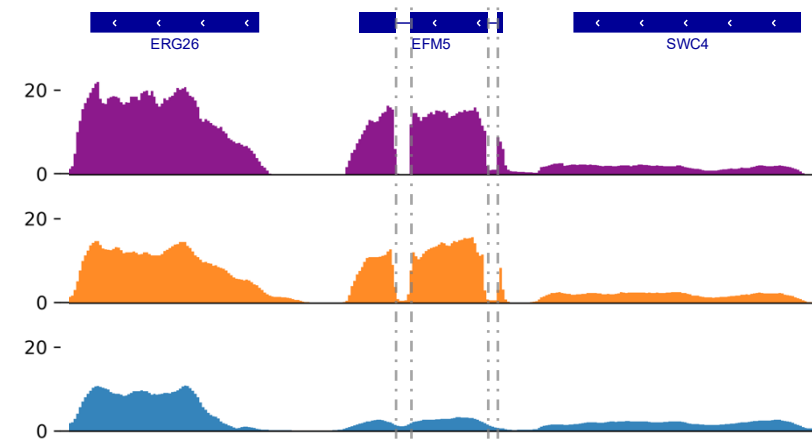
chrVII:362,180-366,023 (RNA-Seq tracks); fold 3



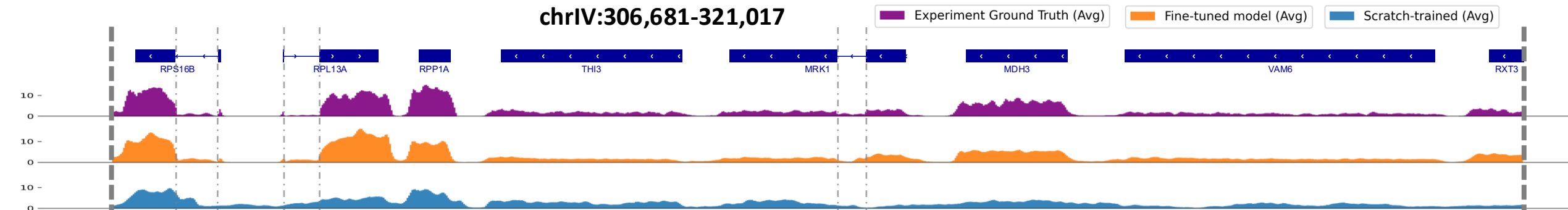
chrIV:305,657-310,505 (RNA-Seq tracks); fold 3



chrVII:495,374-499,965 (RNA-Seq tracks); fold 6



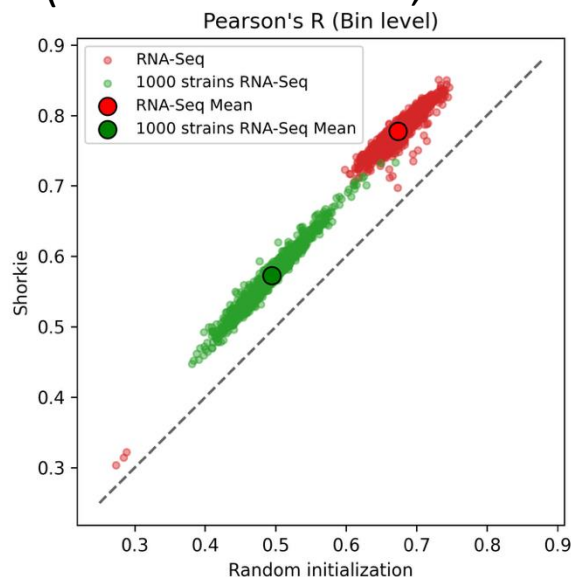
chrIV:306,681-321,017



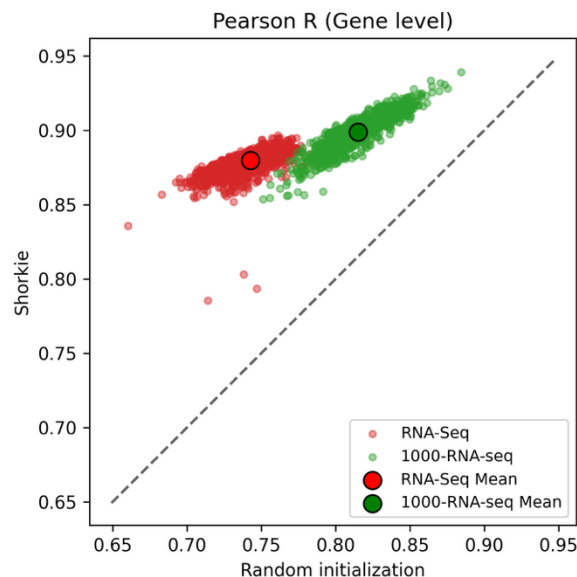
Shorkie vs Random Initialization (Test set)

Bin level

(each dot is a track; all bins)

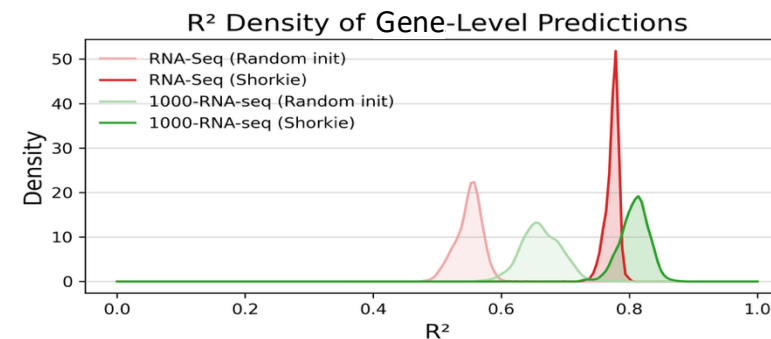
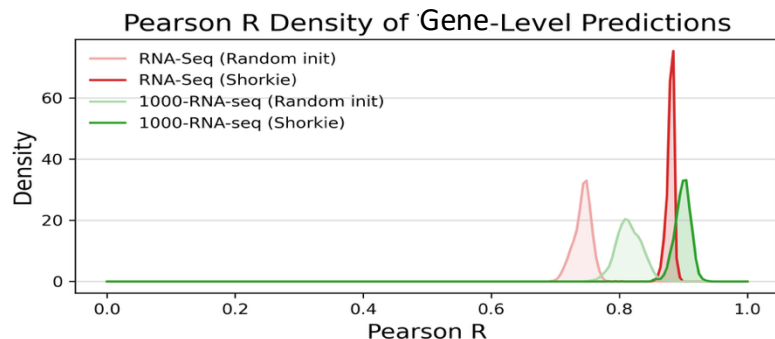
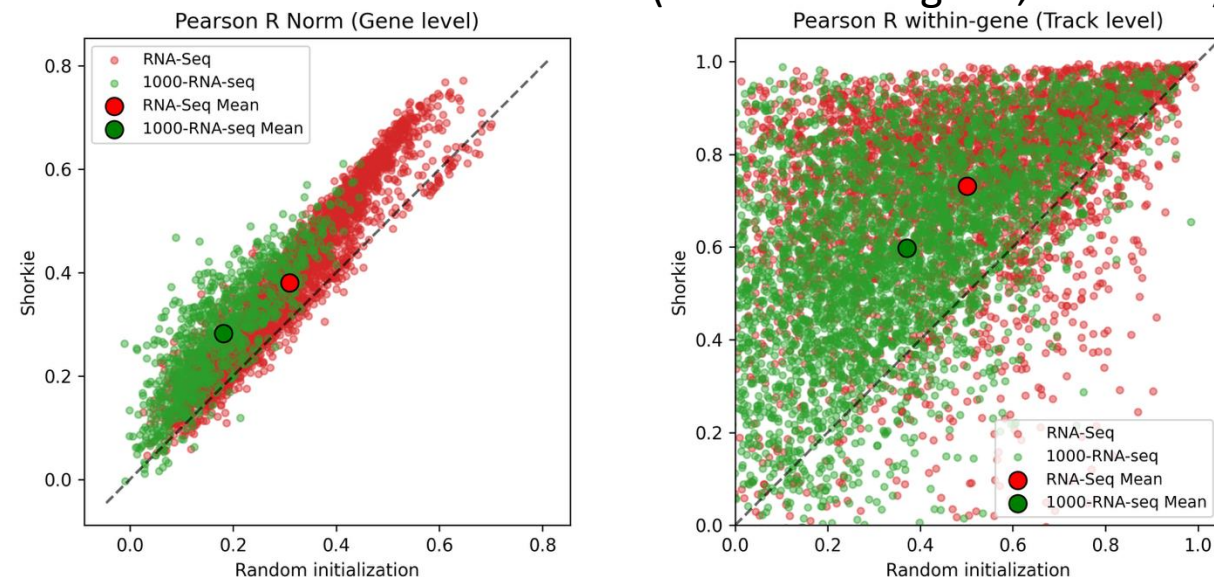


Gene level (each dot is a track; all genes)

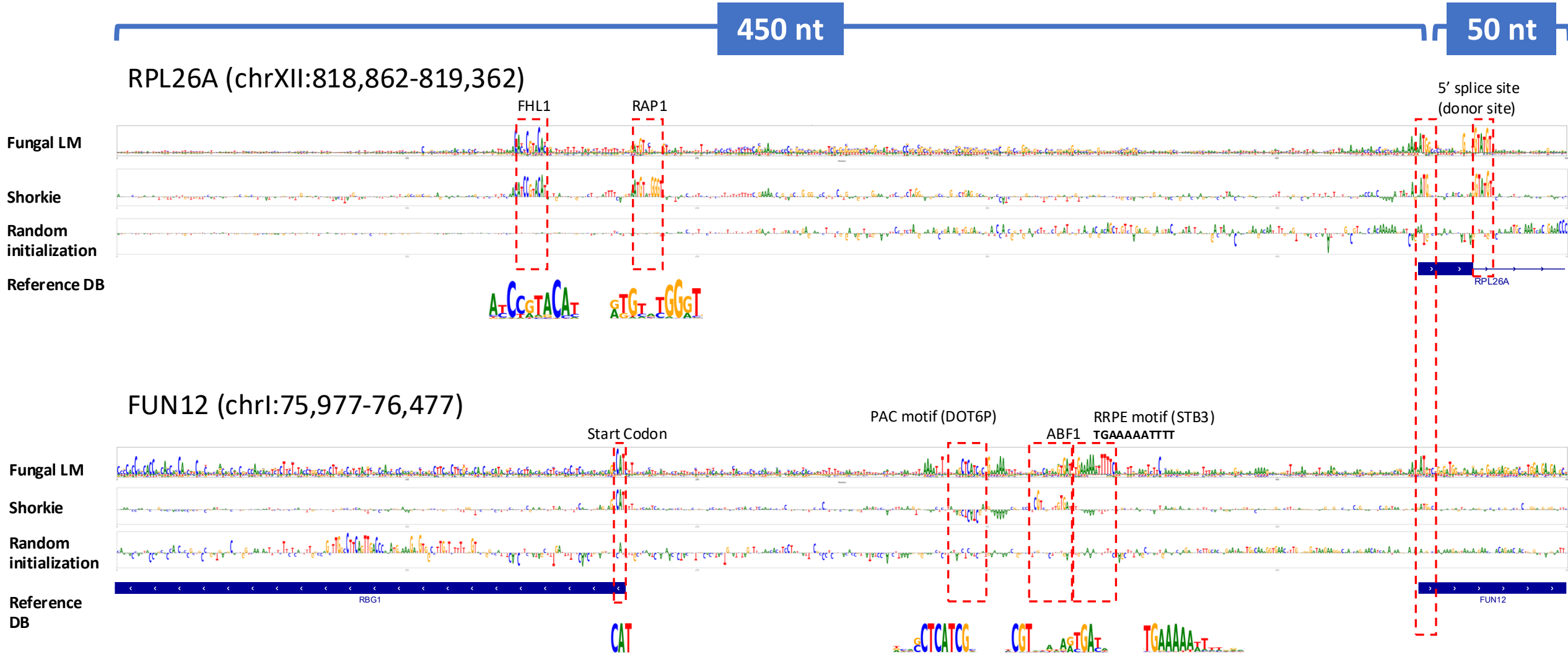


Track level

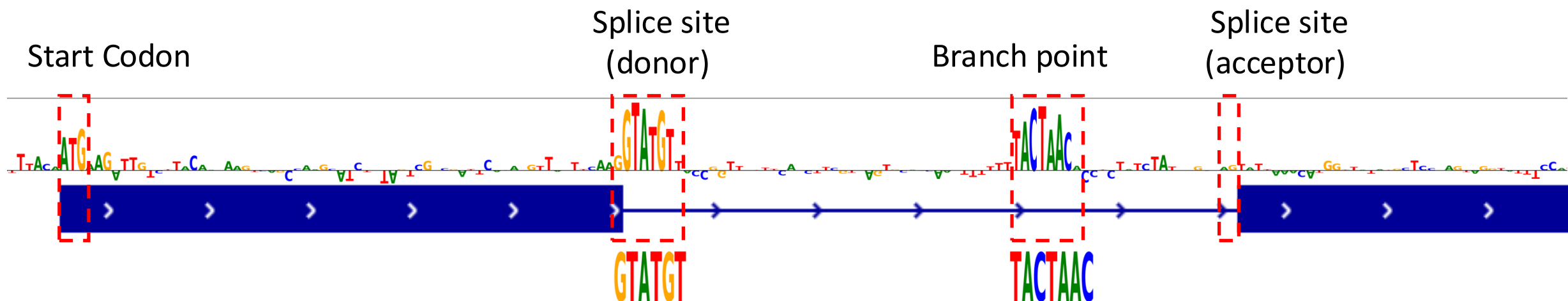
(each dot is a gene; all tracks)



Shorkie ISM maps: Ribosomal genes & RRB genes



Shorkie ISM maps: Splicing motifs



MAD1 (chrVII:346,669-347,169)

Branch point

5' splice site (donor site)

Start codon

REB1

Fungal LM

Shorkie

Random initialization

Reference DB

(reverse strand)

(forward strand)

GTTAGTA

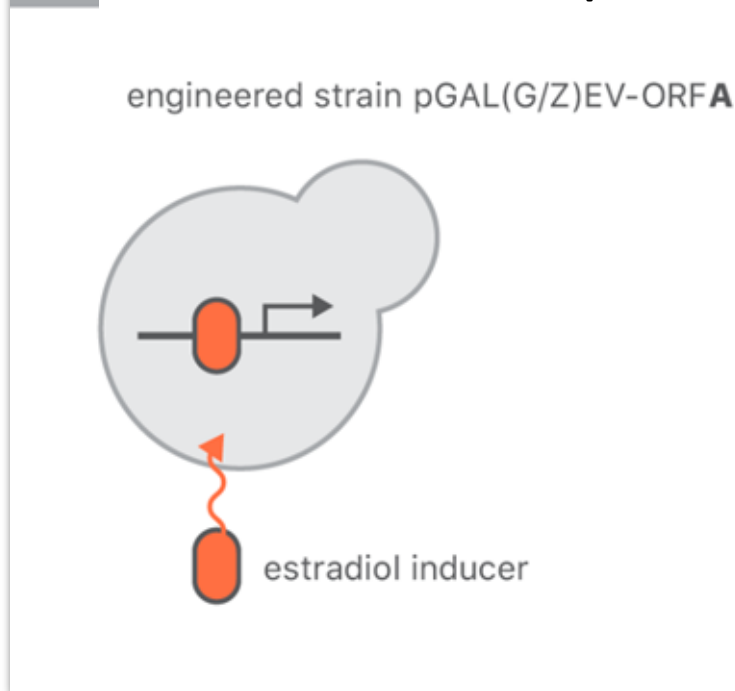
ACATAC

TTACCCG

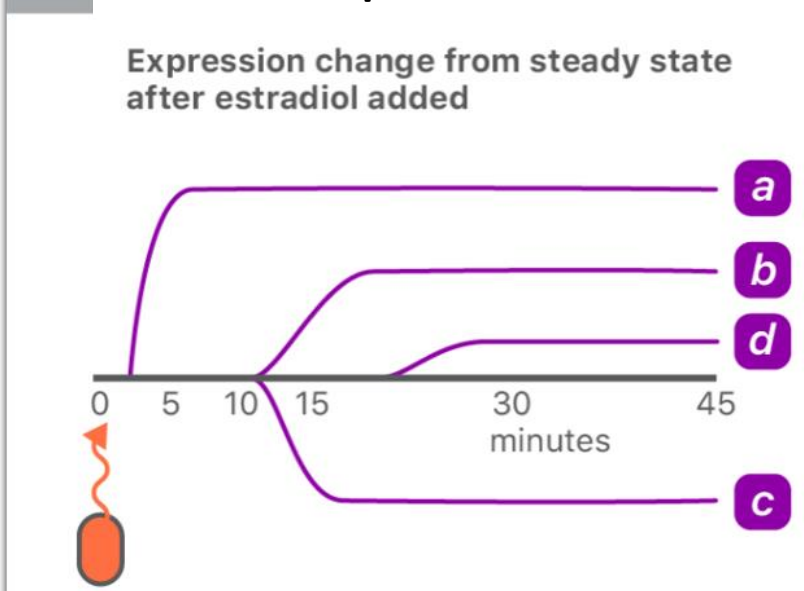
Time-course TF-induced RNA-Seq

- Genome-scale transcription factor **perturbation** (**1340** experiments; **3054** RNA-Seq readouts)
- Aggregating dynamics across many **time-courses**

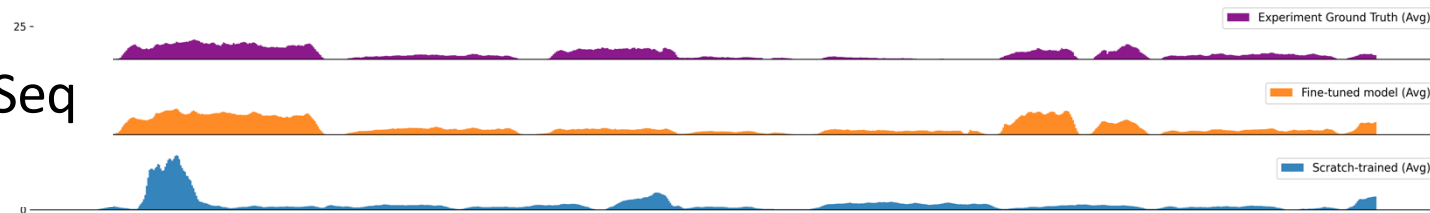
A TF induction experiments



B RNA-seq at different T

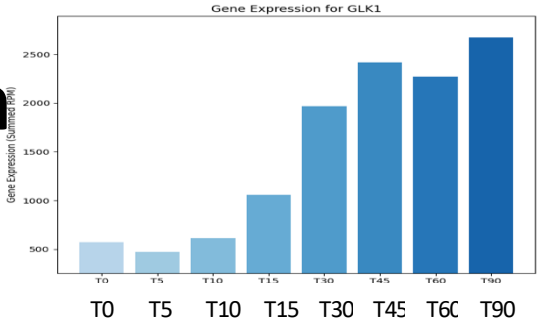


3054 RNA-Seq readouts



Shorkie MSN2 Induction temporal RNA-Seq prediction

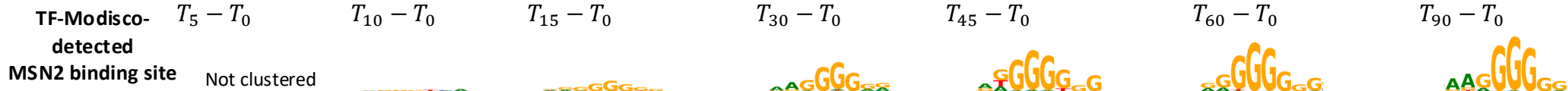
chrIII:50388-50888 (Promoter region of GLK1)



Average motif changes using TF-modisco

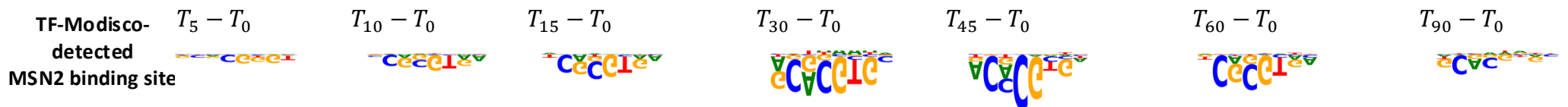
MSN2 Induction

YeTFaSCo
DB motif



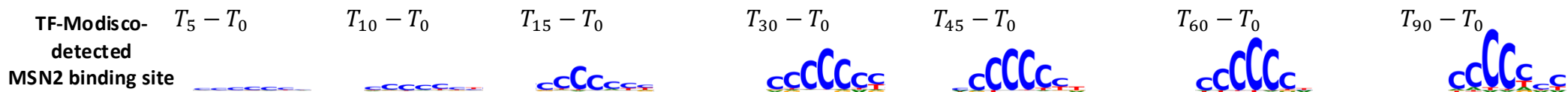
MET4 Induction

YeTFaSCo
DB motif



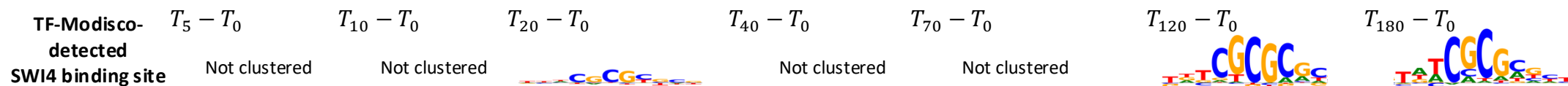
MSN4 Induction

YeTFaSCo
DB motif



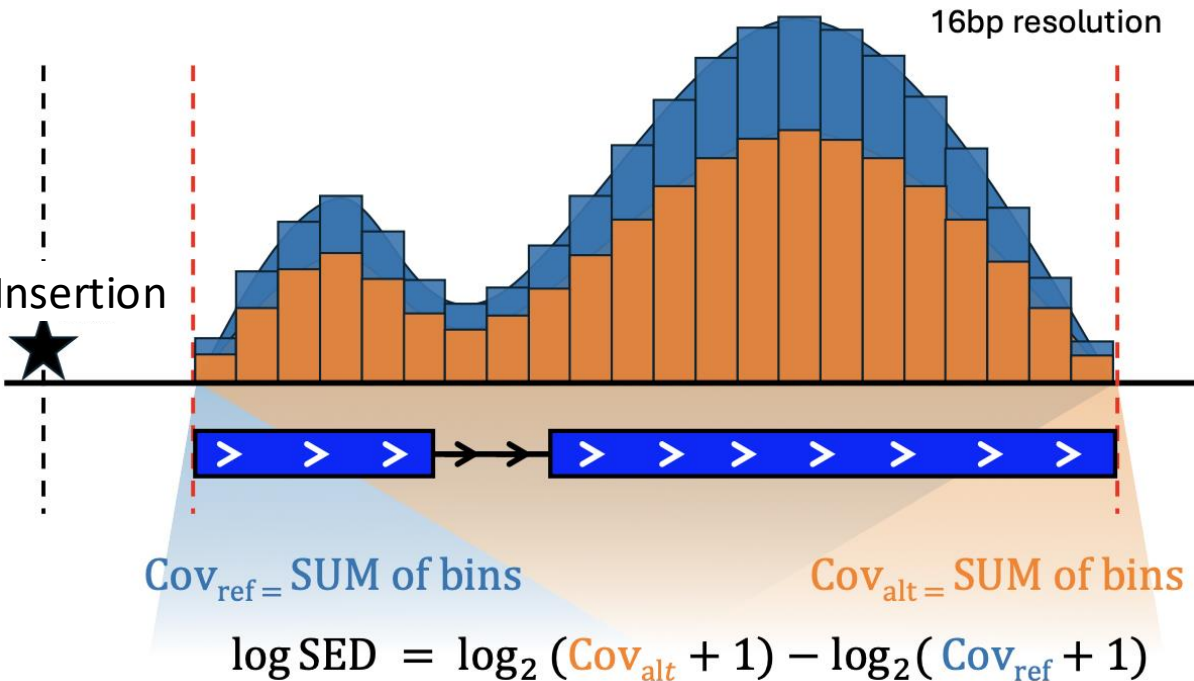
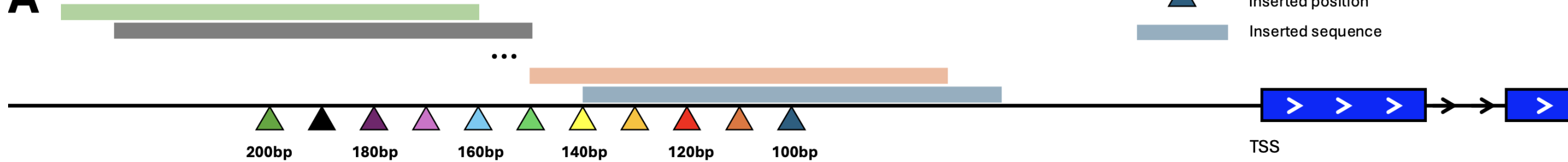
SWI4 Induction

YeTFaSCo
DB motif



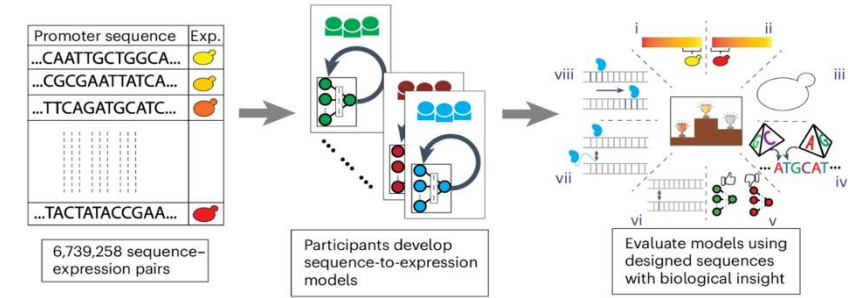
MPRA mutation effect prediction

A

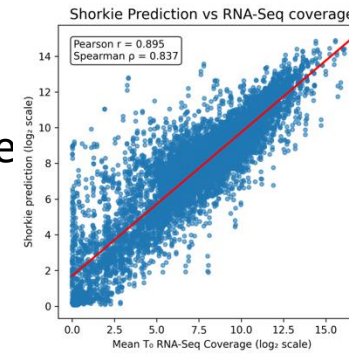


RNA-Seq Coverage Prediction

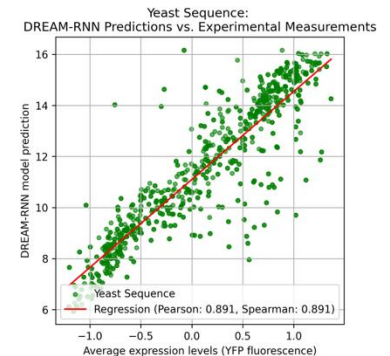
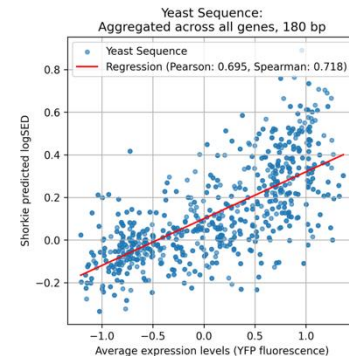
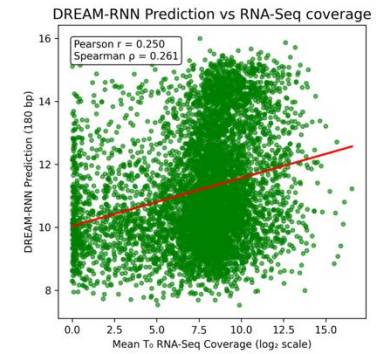
Rafi et al. Yeast Sequences (single-sequence)



Shorkie



DREAM-RNN



Introduction

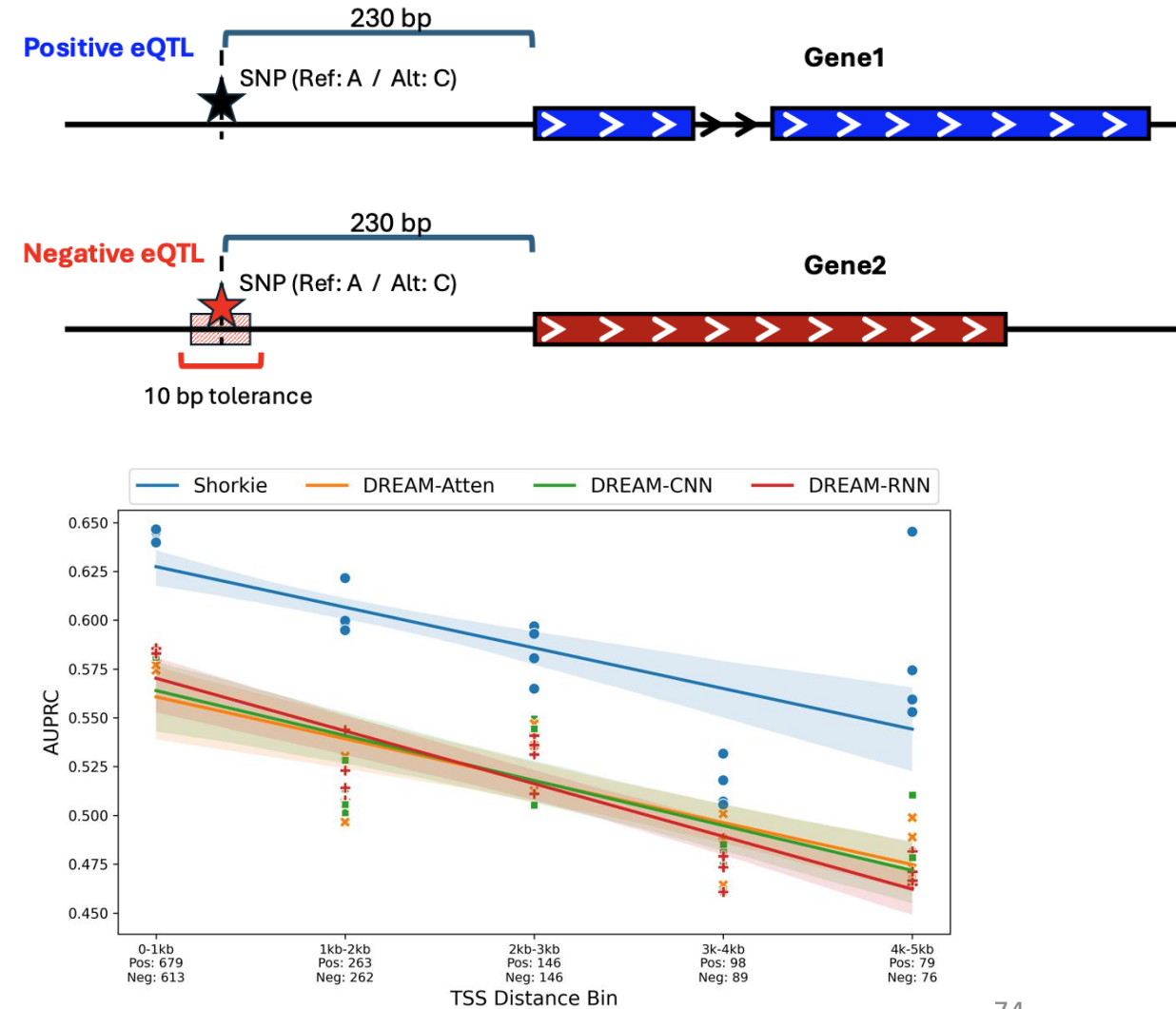
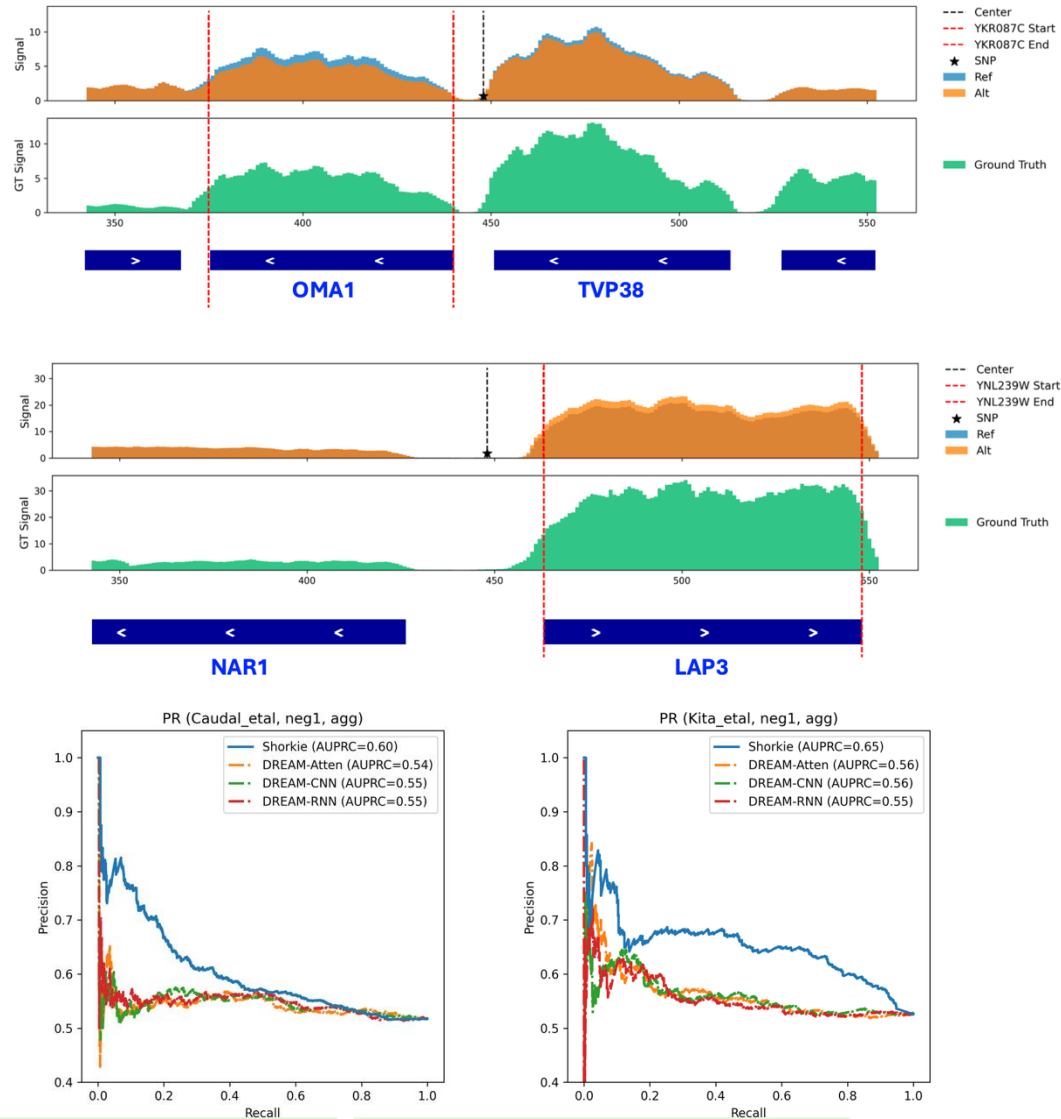
Fungal LM

Transfer Learning

Interpretability

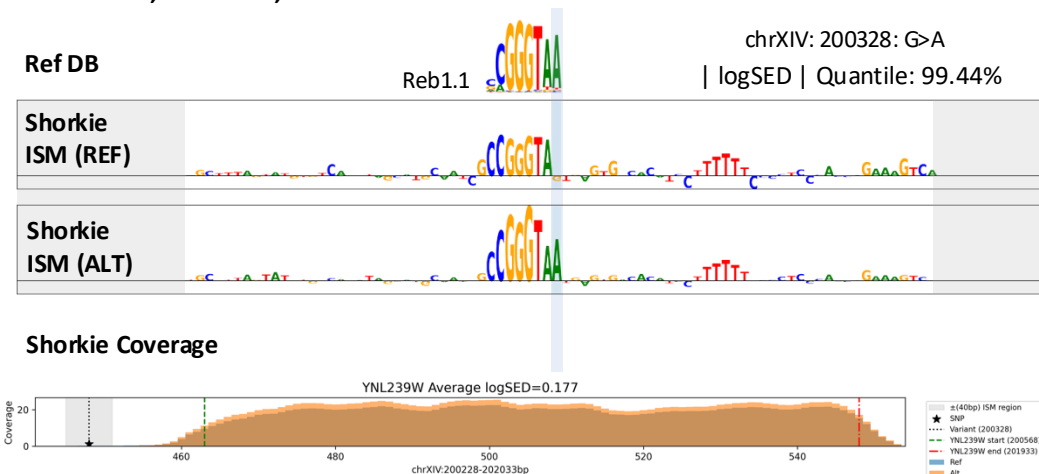
Applications

Variant effects on eQTLs and negatively selected eQTLs

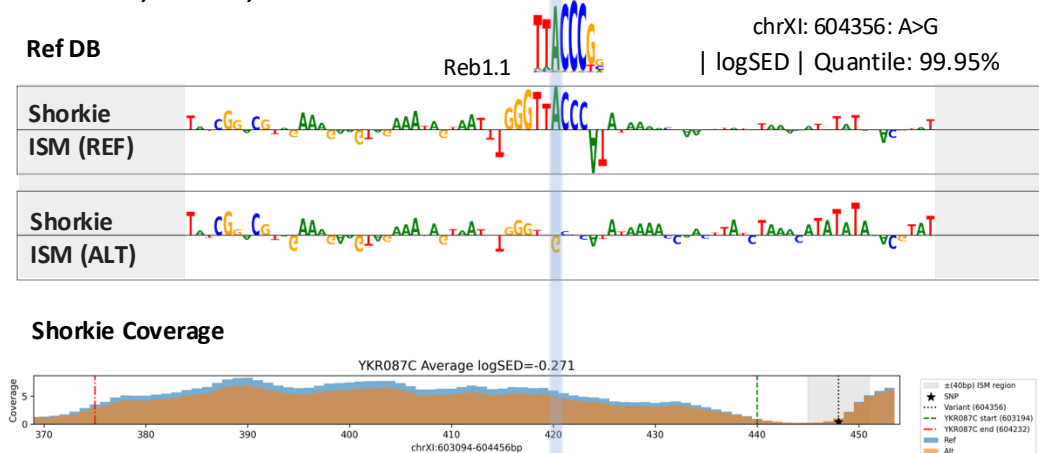


Variant effects on eQTLs and negatively selected eQTLs

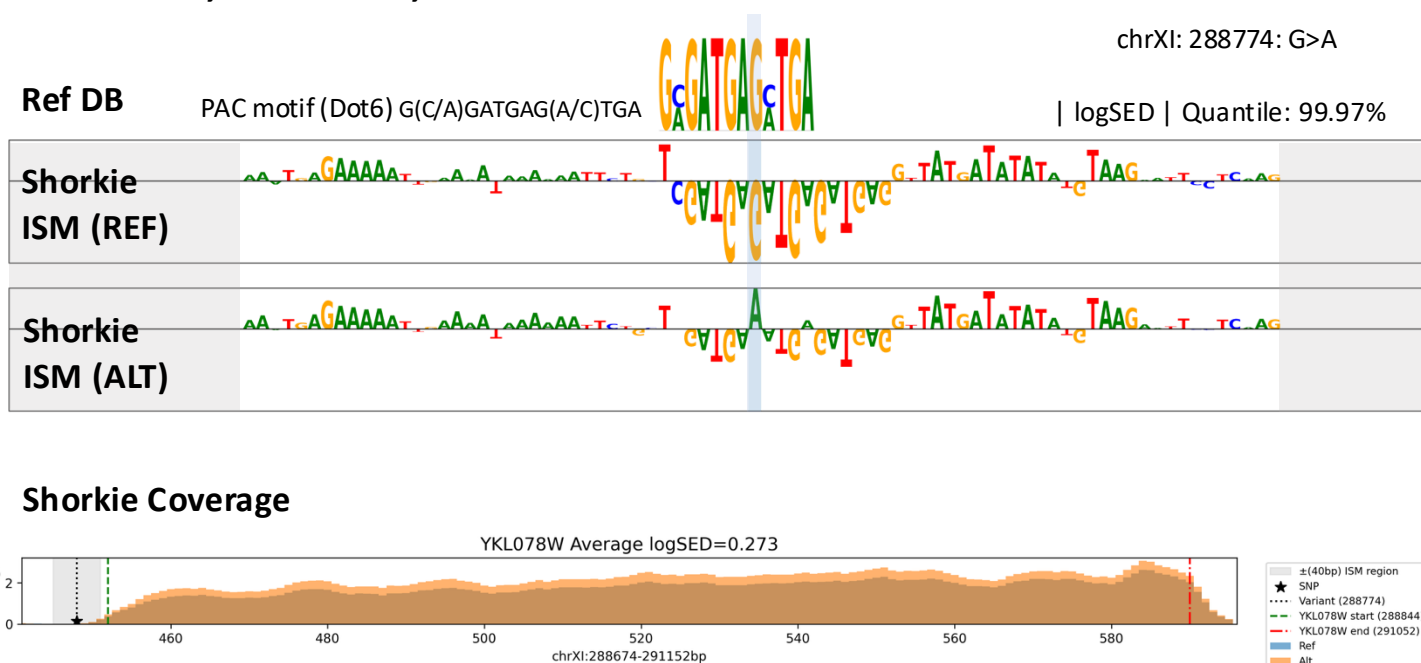
chrXIV:200,288 – 200,368



chrXI:604,316-604,396



chrXI:288,734 – 288,814



Introduction

Fungal LM

Transfer Learning

Interpretability

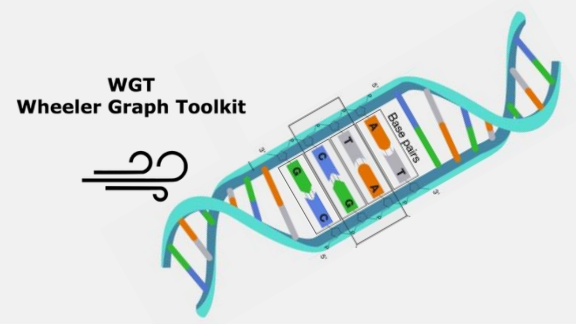
Applications

How do we assemble the complete 3-billion-nucleotide genome?
How can we efficiently represent multiple genomes for fast pattern matching?

Part I & II: Genome Assembly & Indexing

Han1

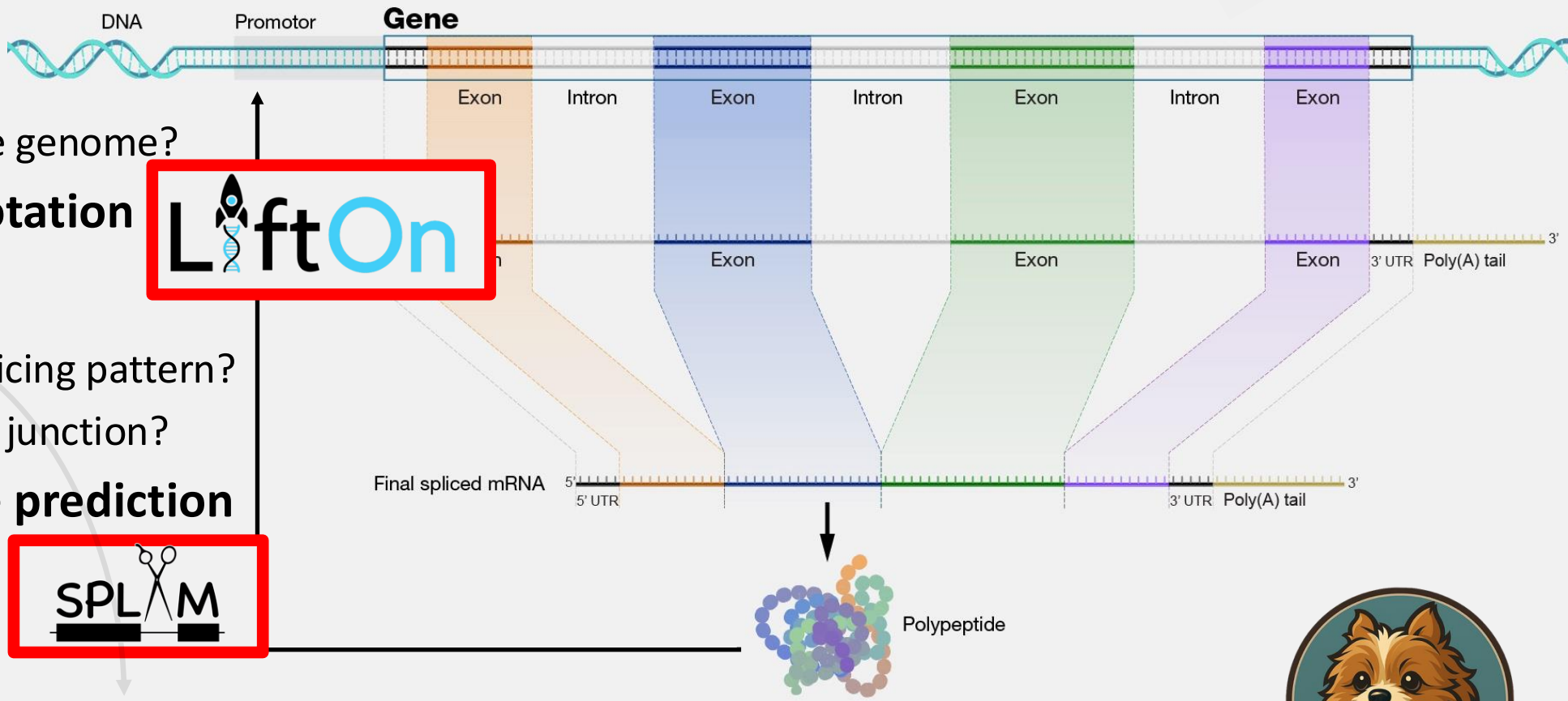
WGT



Where are the genes in the genome?

Part III: Genome Annotation

LiftOn



What are the canonical splicing pattern?
Alternative splicing? Splice junction?

Part IV & V: Splice site prediction

OpenSpiceAI

SPLM

Can we predict gene expression by learning the regulatory grammar in the genome?

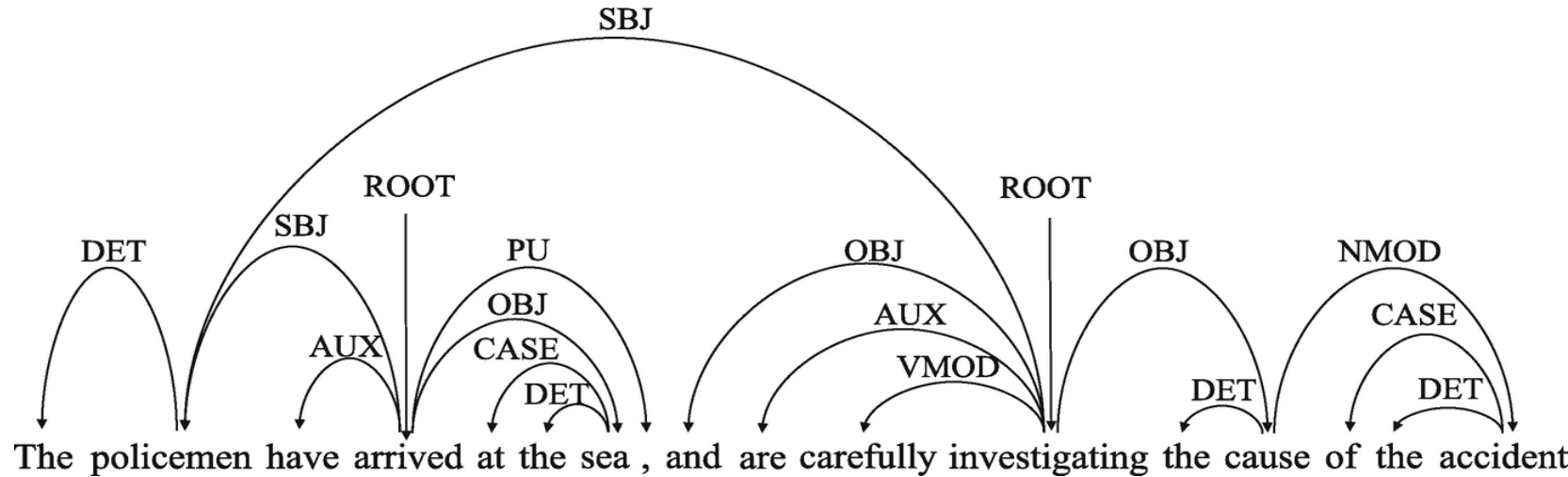
Part VI: Shorkie. RNA-Seq coverage prediction

Shorkie



Conclusions: Language of genomes

Natural Language Processing



DNA gene regulation

