



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

Kuan-Hao Chao

2025.08.25



khchao.com



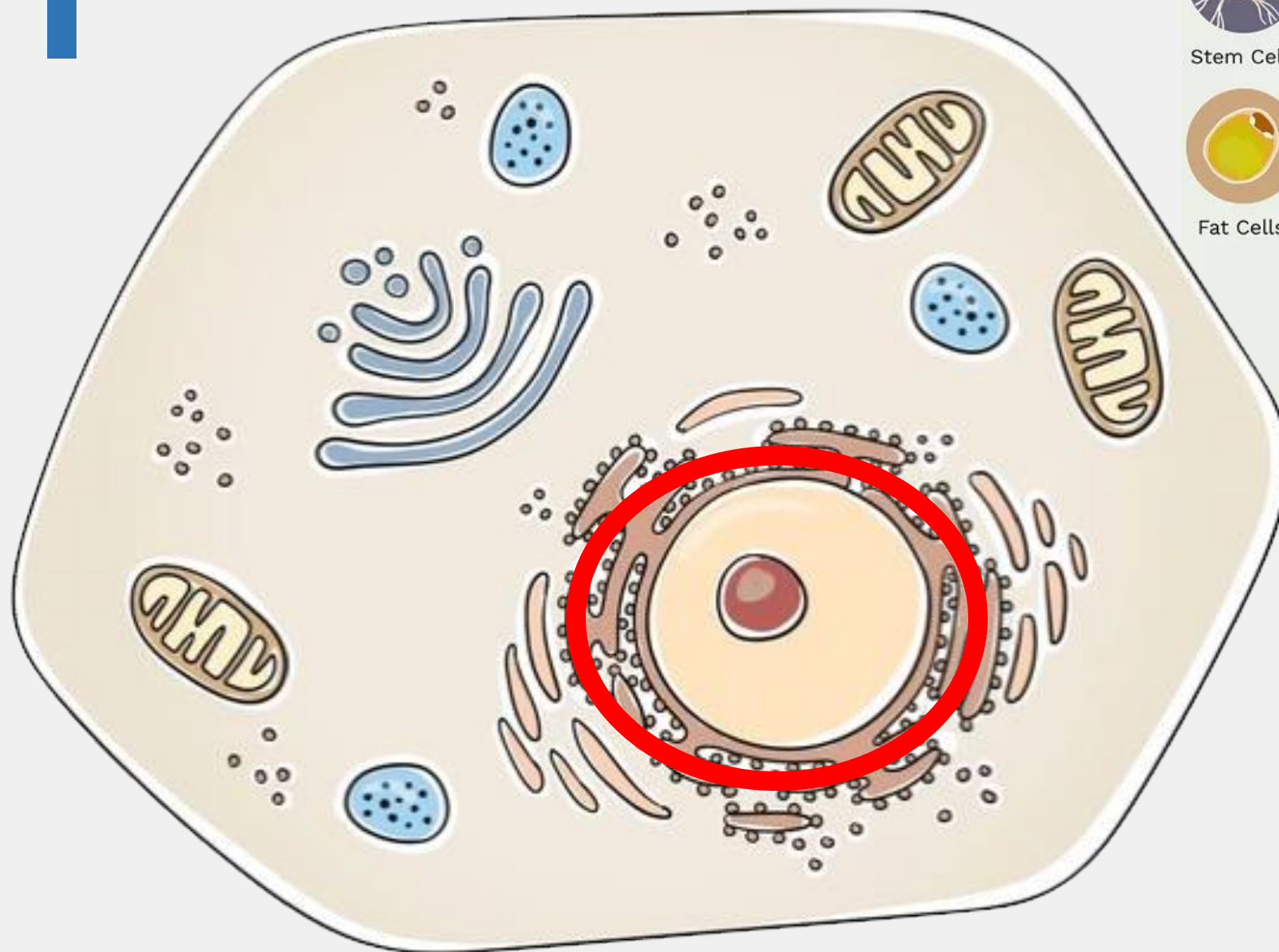
@kuanhaochao.bsky.social



@KuanHaoChao



Kuanhao-Chao



Stem Cells



Bone Cells



Blood Cells



Muscle Cells



Fat Cells



Skin Cells



Nerve Cells



Endothelial Cells



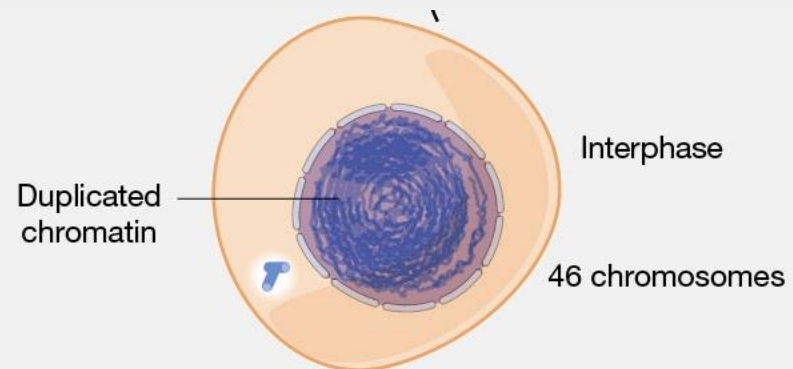
Sex Cells

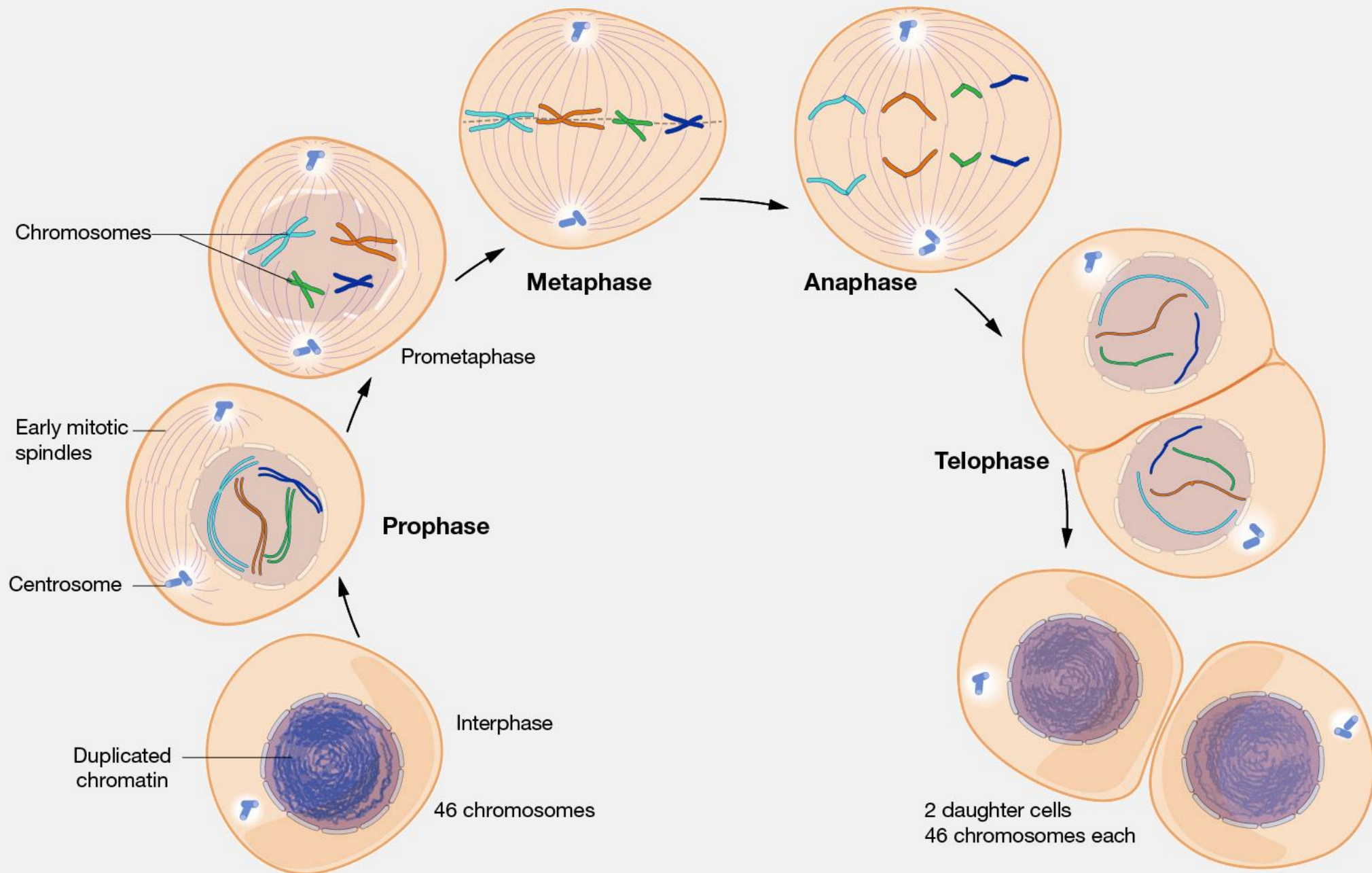


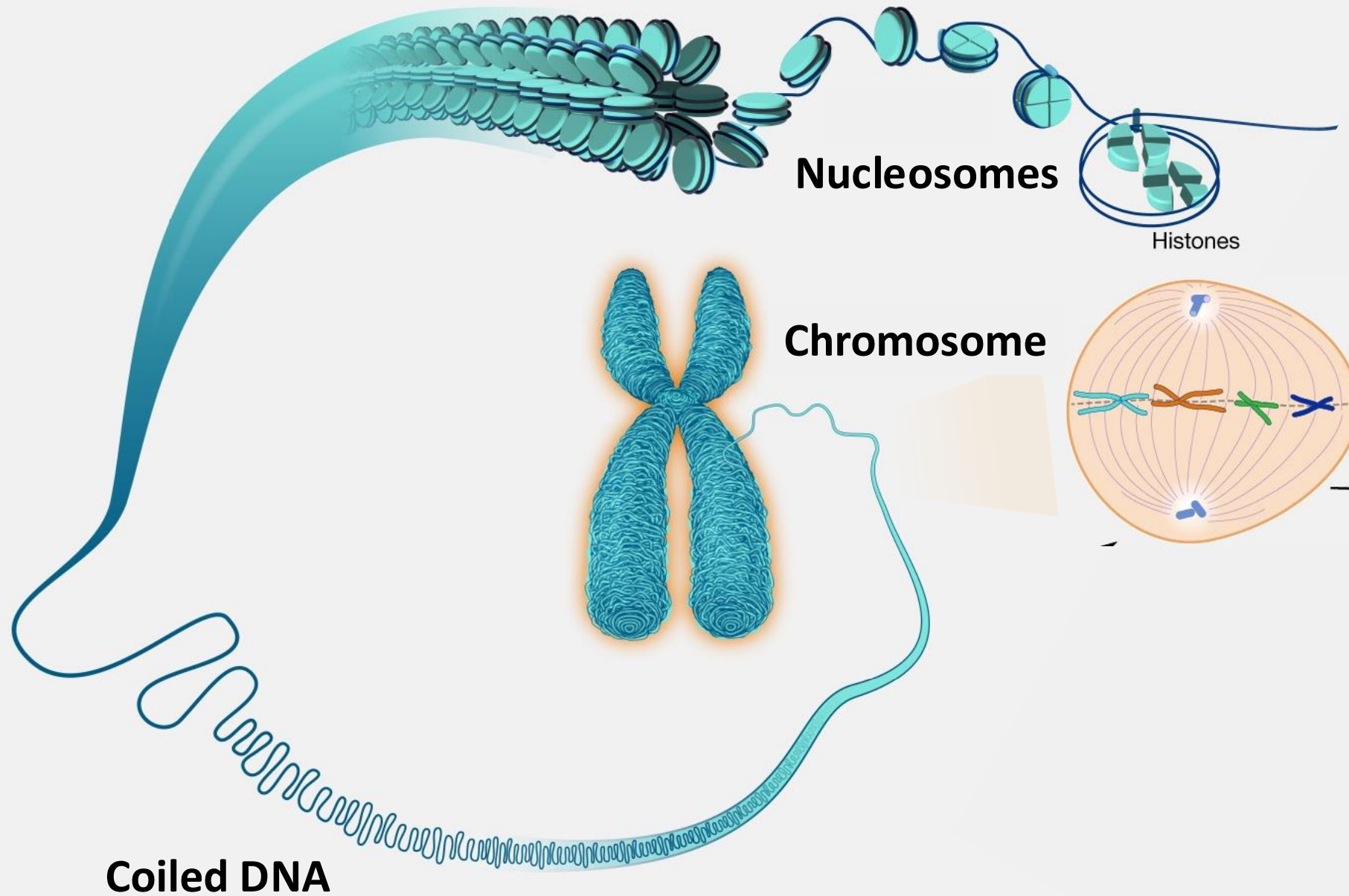
Pancreatic Cells

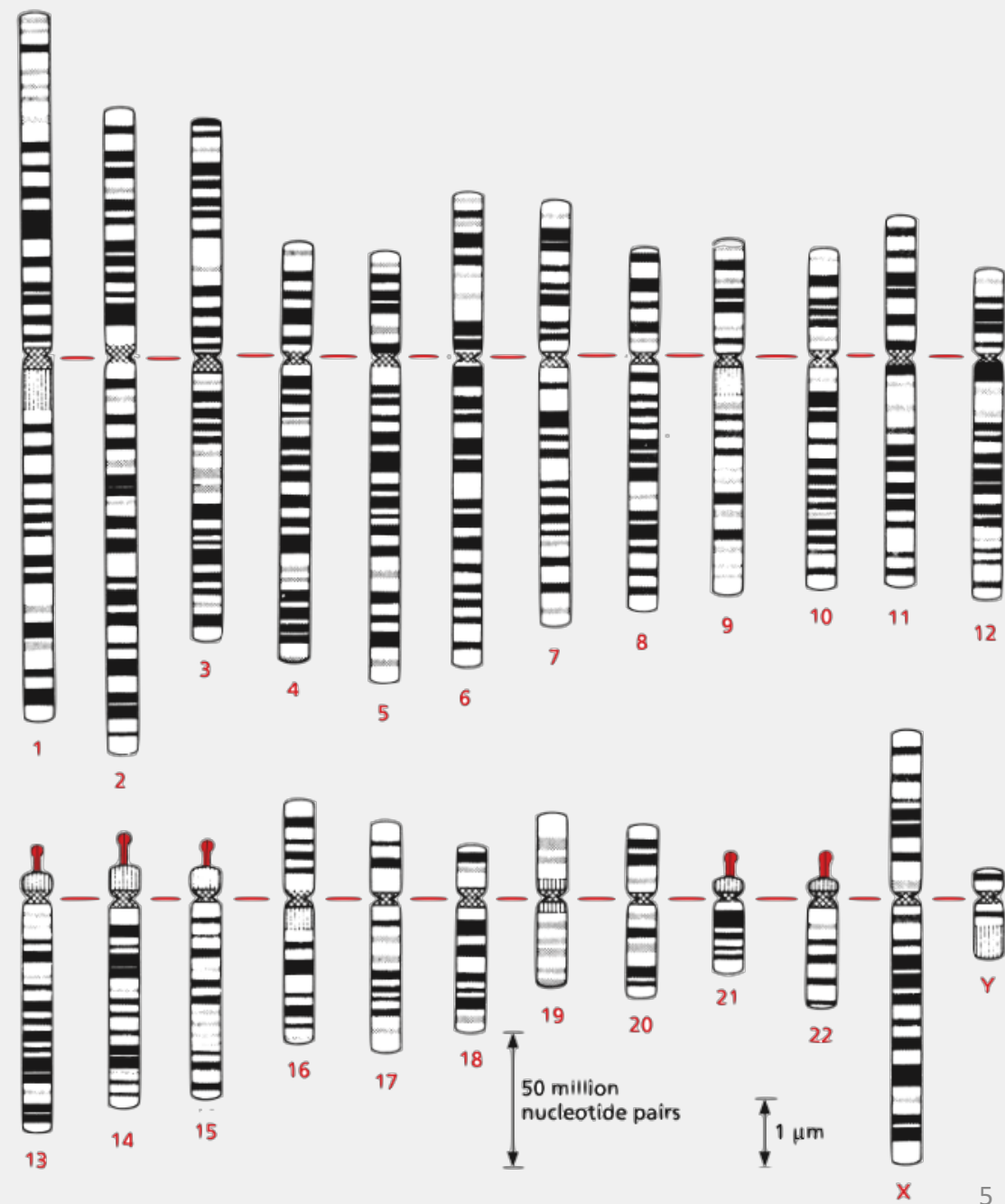


Cancer Cells









Introduction

Assembly & Indexing

Genome Annotation

Splice Site Prediction

RNA-Seq Prediction

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

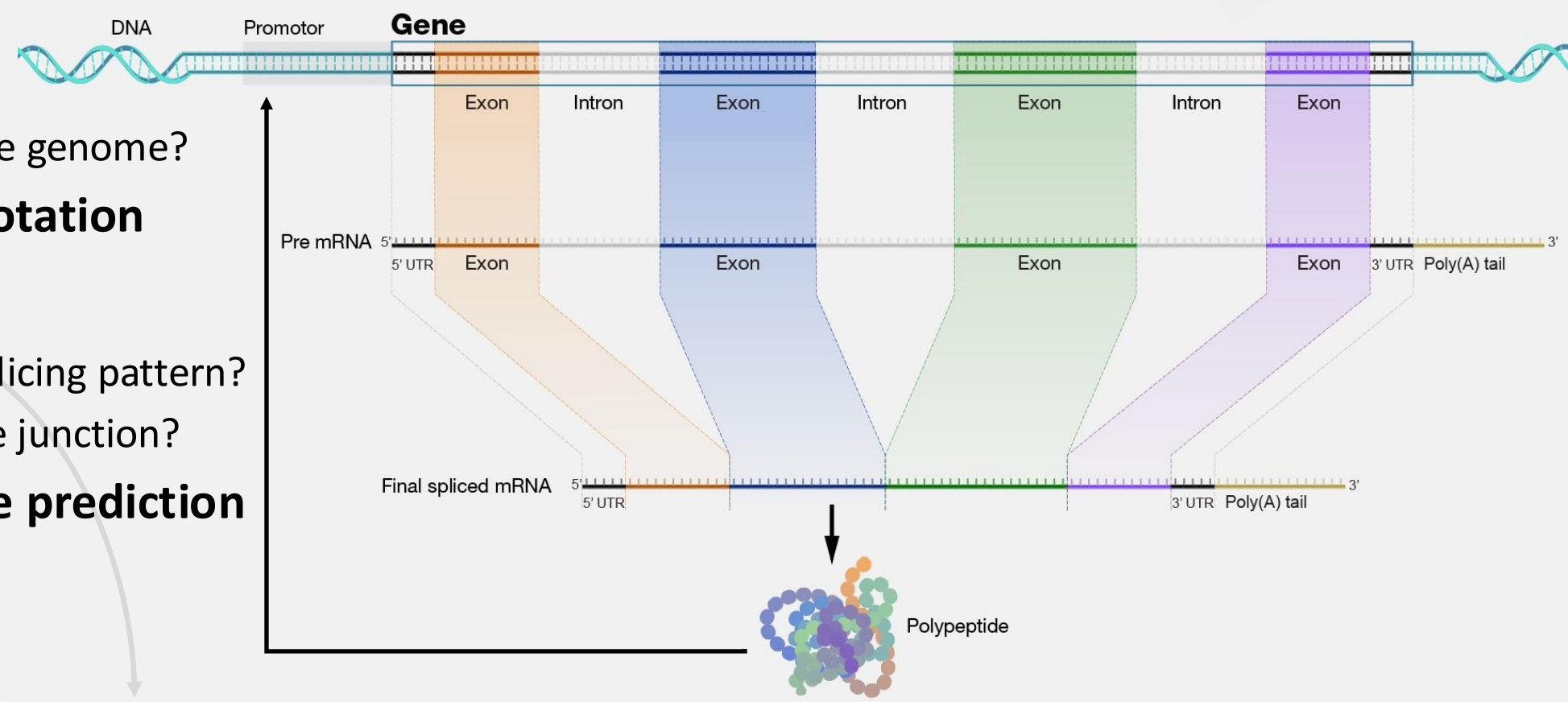


Where are the genes in the genome?

Part III: Genome Annotation

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

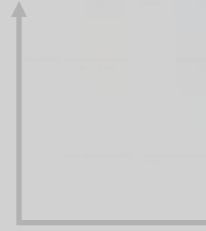
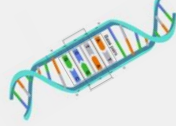
Part IV & V: Splice site prediction



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

Part VI: Shorkie. RNA-Seq coverage prediction

■ Part I & II



Steven Salzberg



Mihaela Pertea



Ben Langmead

Genome Assembly & Indexing

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



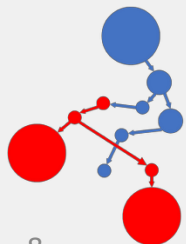
Chao, K. H., Zimin, A. V., Pertea, M., & Salzberg, S. L. (2023). The first gapless, reference-quality, fully annotated genome from a Southern Han Chinese individual. *G3: Genes, Genomes, Genetics*, 13(3), jkac321.



RECOMB-SEQ



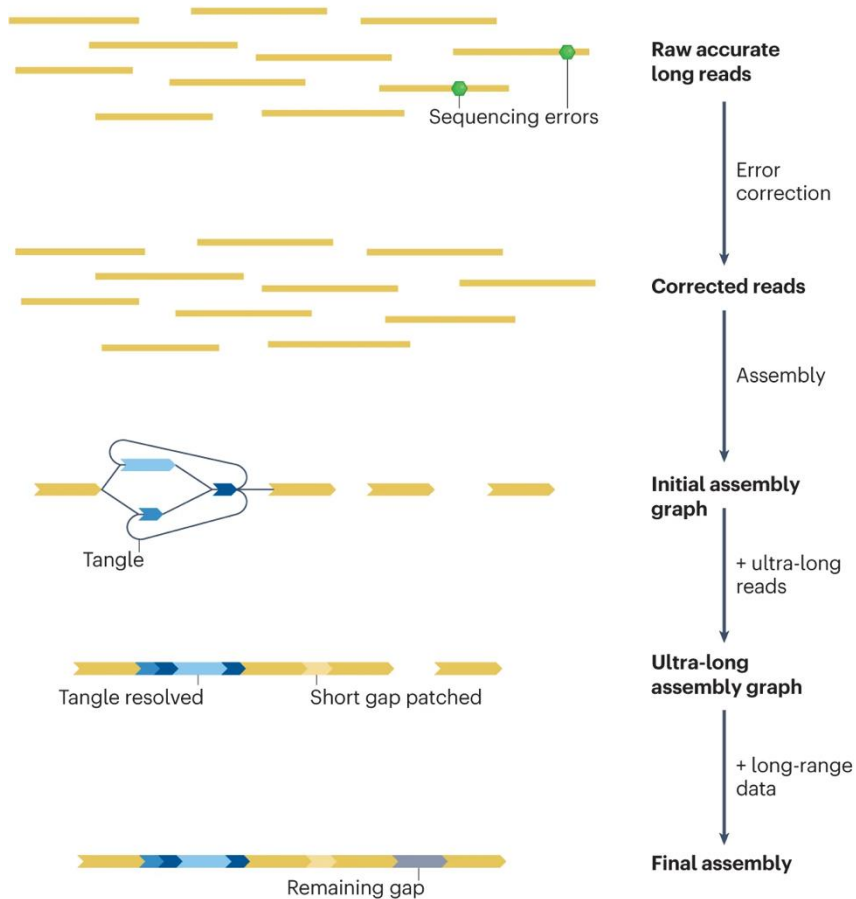
Chao, K. H., Chen, P. W., Seshia, S. A., & Langmead, B. (2023). WGT: Tools and algorithms for recognizing, visualizing, and generating Wheeler graphs. *iScience*, 26(8).



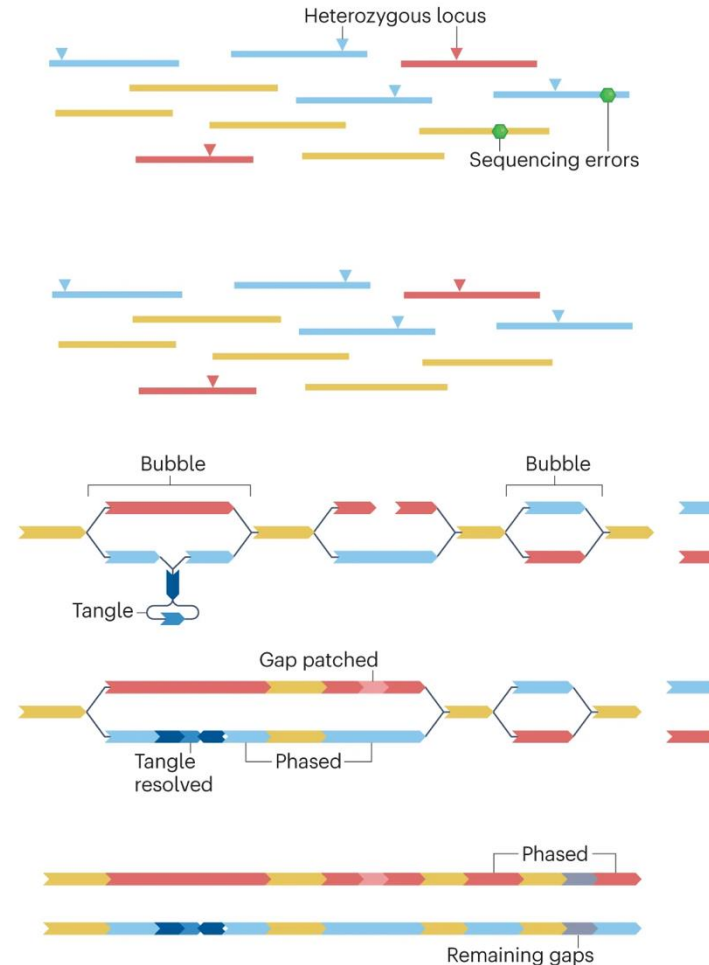
8

Genome Assembly

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

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

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.

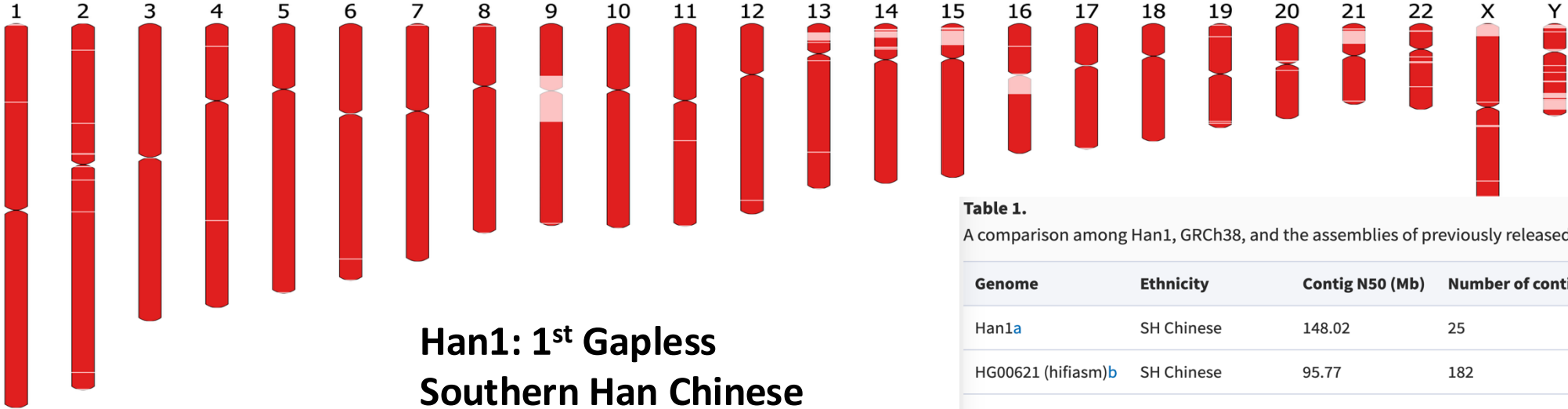
Genome Assembly: Han1



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.

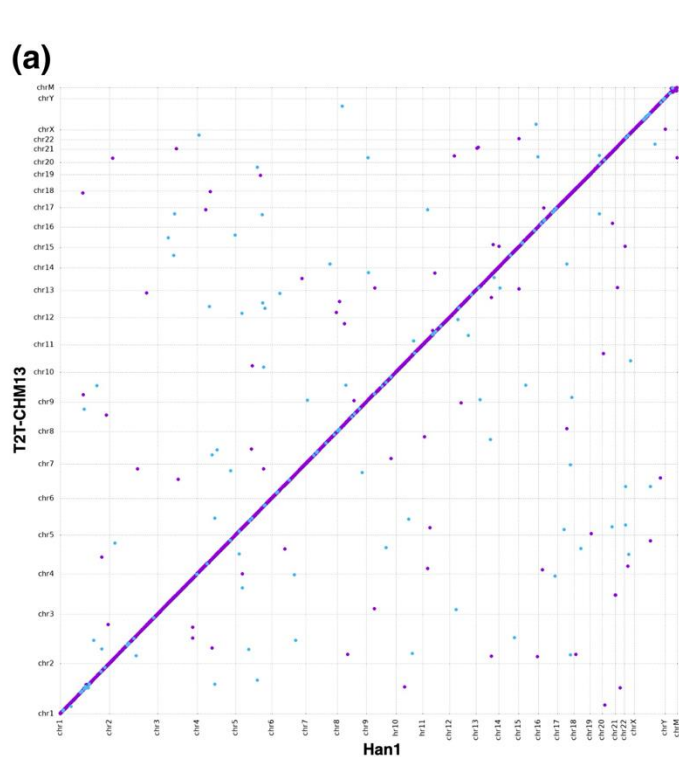
Genome	Ethnicity	Contig N50 (Mb)	Number of contigs	Number of gaps	Assembly size (Gb)
Han1a	SH Chinese	148.02	25	0	3.10
HG00621 (hifiasm)b	SH Chinese	95.77	182	157	3.11
T2T-CHM13v2.0c	Northern European	150.62	25	0	3.12
HJ-H1d	NH Chinese	28.15	1,330	427	3.07
HJ-H2d	NH Chinese	25.90	896	390	2.91
NH1d	NH 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.

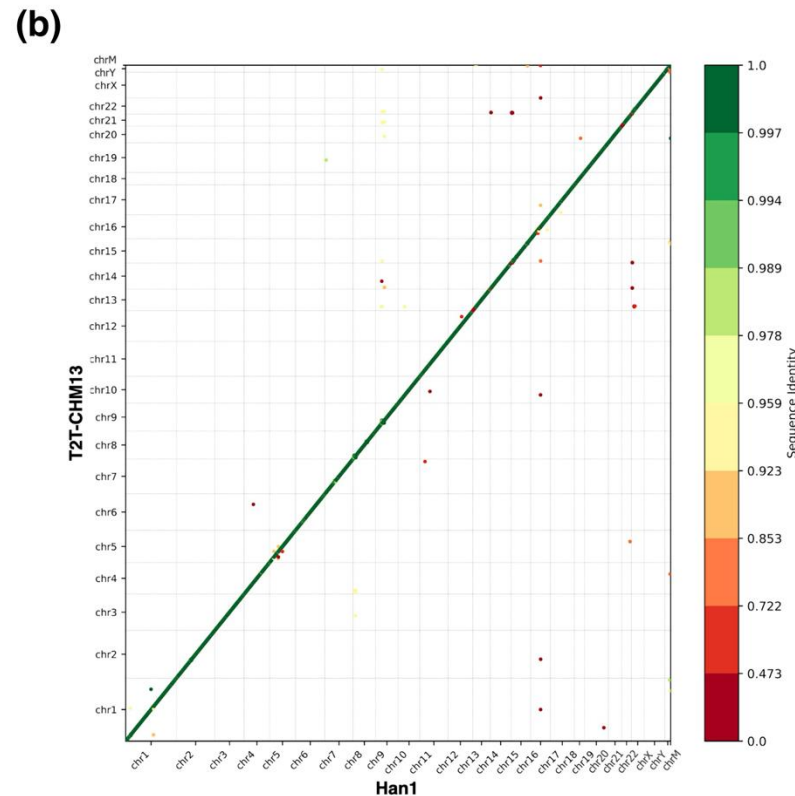
Assembler	Sequencing data	assembled sequence (bp)	Contig N50	Number of contigs	Quality value
Hifiasm v0.16.1-r375	PacBio HiFi	3,110,501,483	95,769,069	182	57.8a
Flye v2.5	ONT Ultralong	2,974,205,132	40,850,737	1,658	25.6a

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

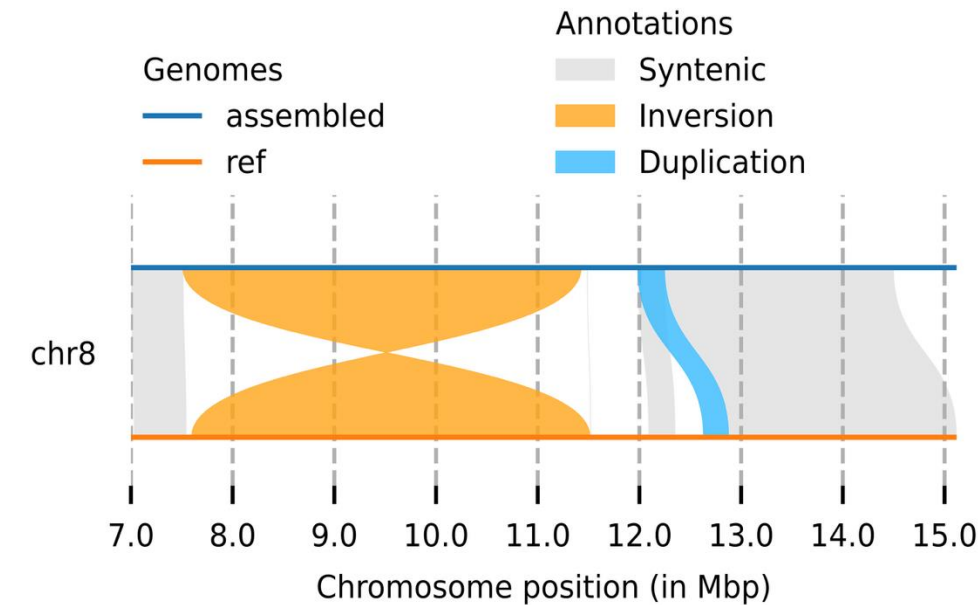
Mummer dot plot



Gene order plot



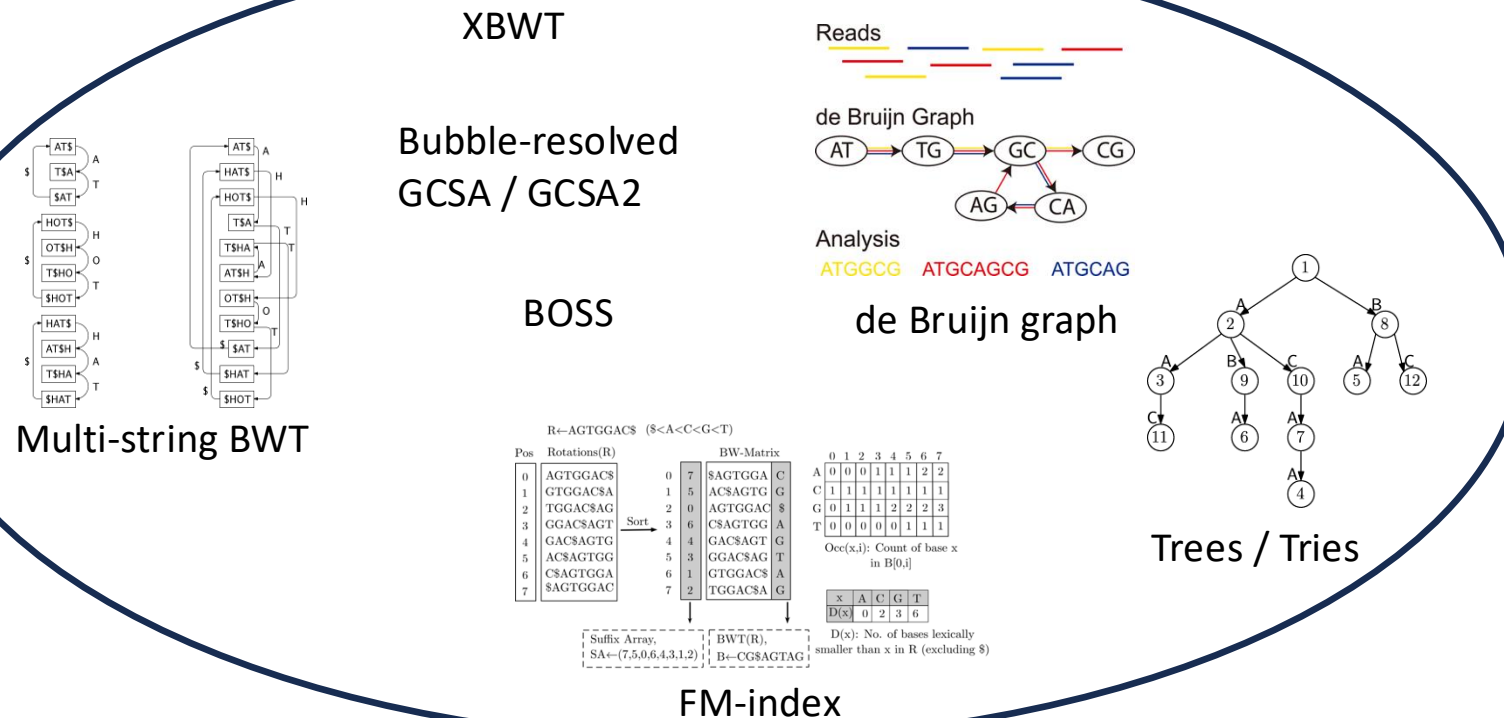
Inversion on β defensin gene cluster



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

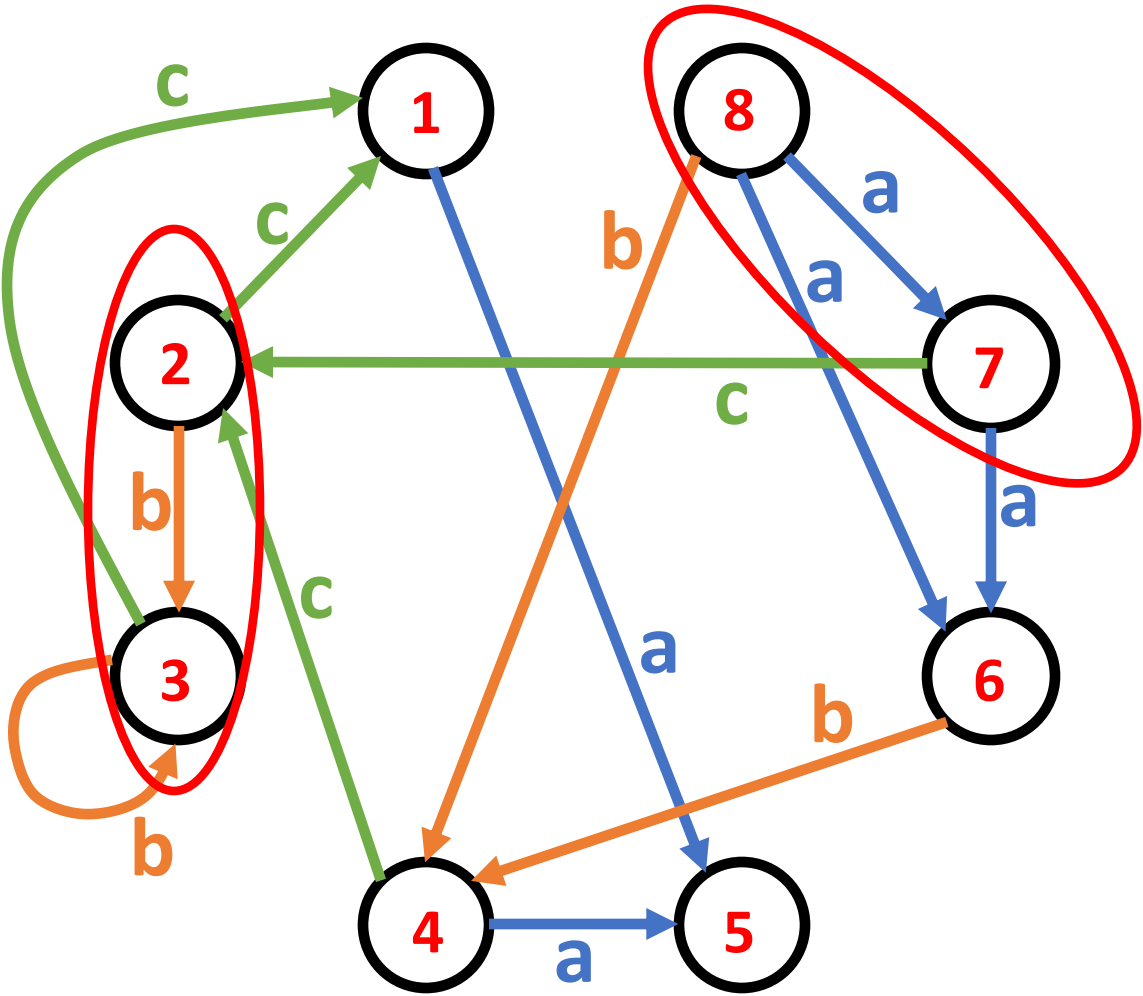
2020 Graph Burrows-Wheeler Transform (GBWT)

2021 vg Giraffe

Wheeler Graph Recognition

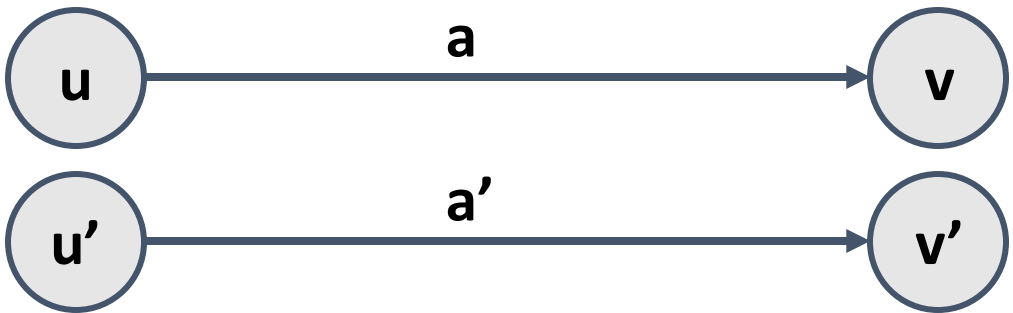


RECOMB-SEQ



Wheeler graph recognition problem

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





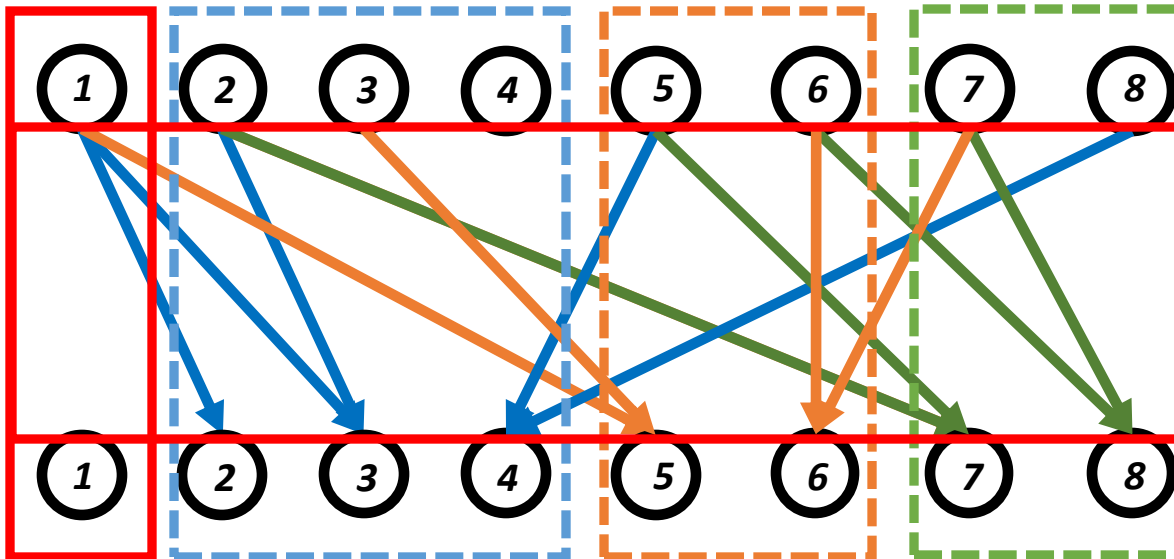
Satisfiability modulo theories (SMT)

- SMT solver = SAT solver + IDL theory solver**

1. Node with indegree 0 comes before every other nodes.

2. $a < a' \implies v < v'$

3. $(a = a') \wedge (u < u') \implies v \leq v'$



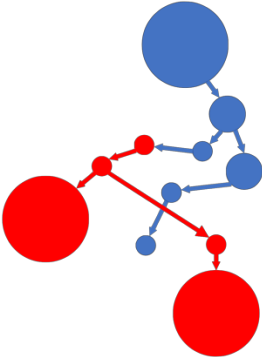
```
#####
## Adding Wheeler graph constraints
#####
# SMT solver initialization
s = SolverFor('QF_IDL')

# 1st WG rule: nodes without incoming edge should be in front
for u in no_incoming_edge:
    s.add(u <= len(no_incoming_edge))

# 2nd WG rule: a < a' => v < v'
for us, (lb, ub) in group_info:
    for u in us:
        s.add(And(u > lb, u <= ub))
        s.add(Distinct(*us))

# 3rd WG rule: (a = a') ^ (u < u') => v <= v'
for i in range(len(edges)):
    # for j in range(len(edges)):
    for j in range(i+1, len(edges)):
        if i != j:
            ui, vi, wi = edges[i]
            uj, vj, wj = edges[j]
            if wi == wj:
                s.add(Implies(ui < uj, vi <= vj))
                s.add(Implies(ui > uj, vi >= vj))
```

Wheeler Graph Recognition

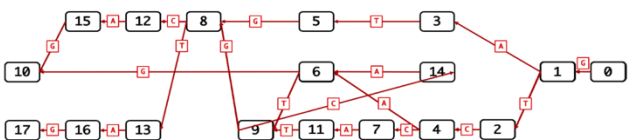
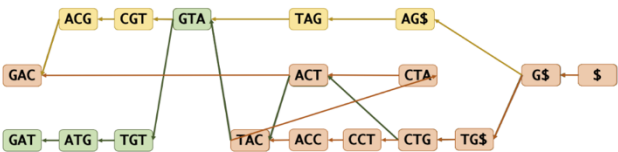


WGT
Wheeler Graph Toolkit



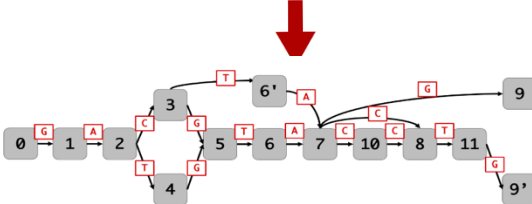
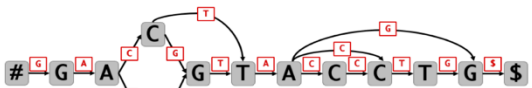
GACGTA-CTG
GACGTA--G
GATGTA-CTG
GAC-TACCTG

De Bruijn Graph



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

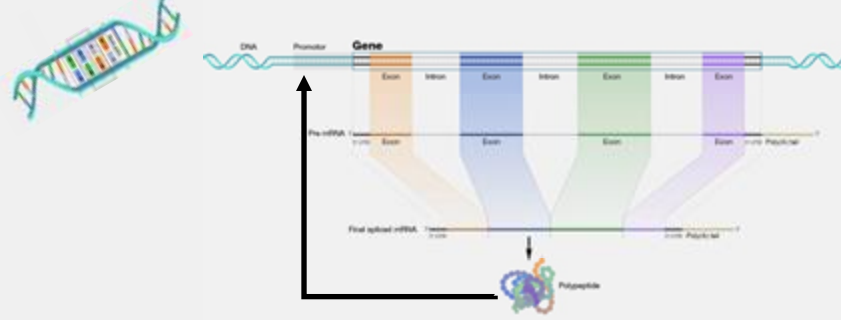


Original label	Wheeler order	Original label	Wheeler order
0	1	9	8
2	2	5	9
7	3	9'	10
3	4	4	11
10	5	6'	12
8	6	11	13
1	7	6	14

18 nodes, 21 edges
1 “distant gluing”

14 nodes, 16 edges
No “distant gluing”

■ Part III



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



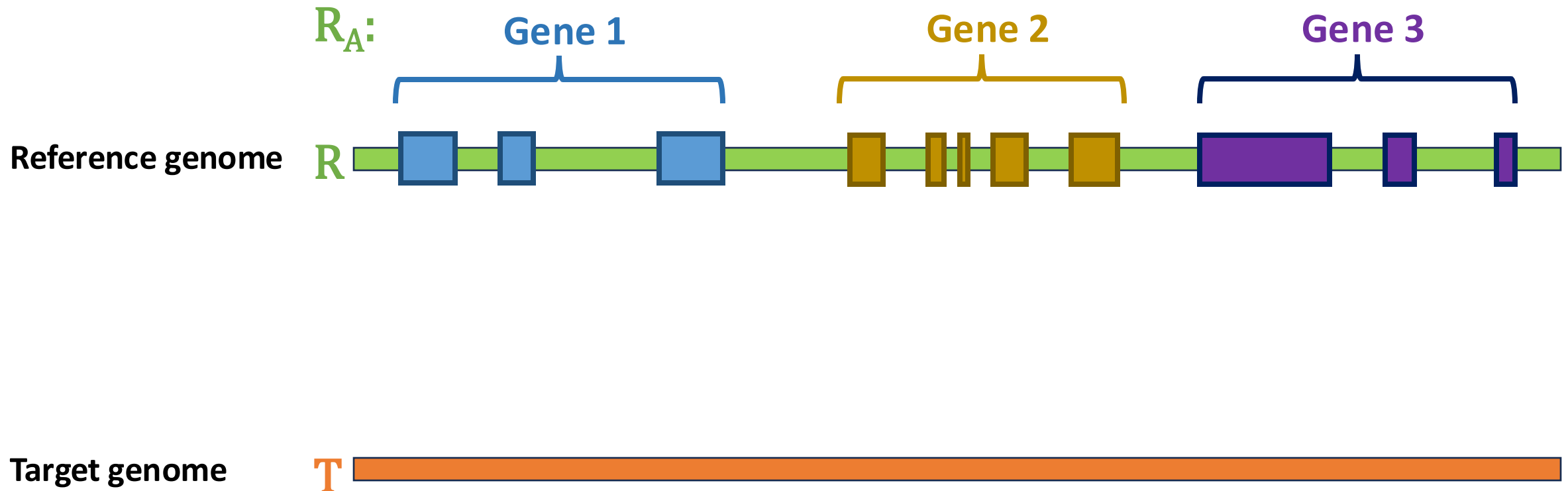
Genome (FASTA)

```
CAGCCCCCGGAGACCTtaaatacaggaagaaaaaggCAGGACAGAATTACAAGGTGCTGGCCCAGGGCGGGCAGCGGCCCT
GCCTCCTACCCCTTGCGCCTCATGACCAGCTTGTTGAAGAGATCCGACATCAAGTGCCACCTTGGCTCGTGGCTCTCACT
GCAACGGGAAAGCCACAGACTGGGGTGAAGAGTTCAAGTCACATGCGACCGGTgactccctgtccccacccccatgACACT
CCCCAGCCCTCCAAGGCCACTGTGTTTCCCAGTTAGCTCAGAGCCTCAGTCGATCCCTGACCCAGCACCGGGGCACTGATG
AGACAGCGGCTGTTTGAGGagccacctcccagccacctcggggccagggccaggggtgtGCAGCACCCTGTACAATGGGG
AAACTGGCCCAGAGAGGTGAGGCAGCTTGCTTGGGGTCACAGAGCAAGGCAAAAGCAGCGCTGGGTACAAGCTCAAACC
ATAGTGCCCAGGGCACTGCCGCTGCAGGCGCAGGCATCGCATCACACCAGTGTCTGCGTTTCACAGCAGGCATCATCAGTA
```

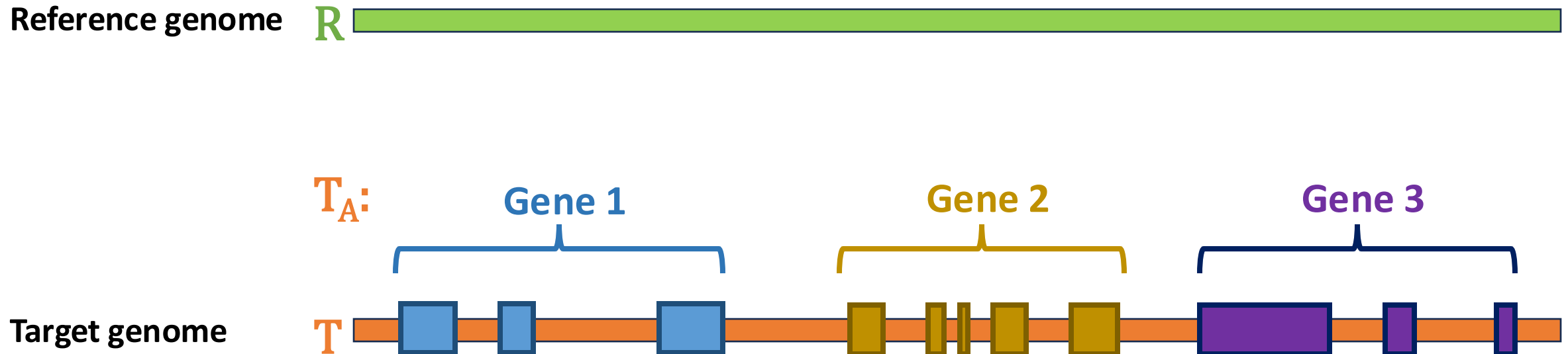
Annotation (GFF / GTF)

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;

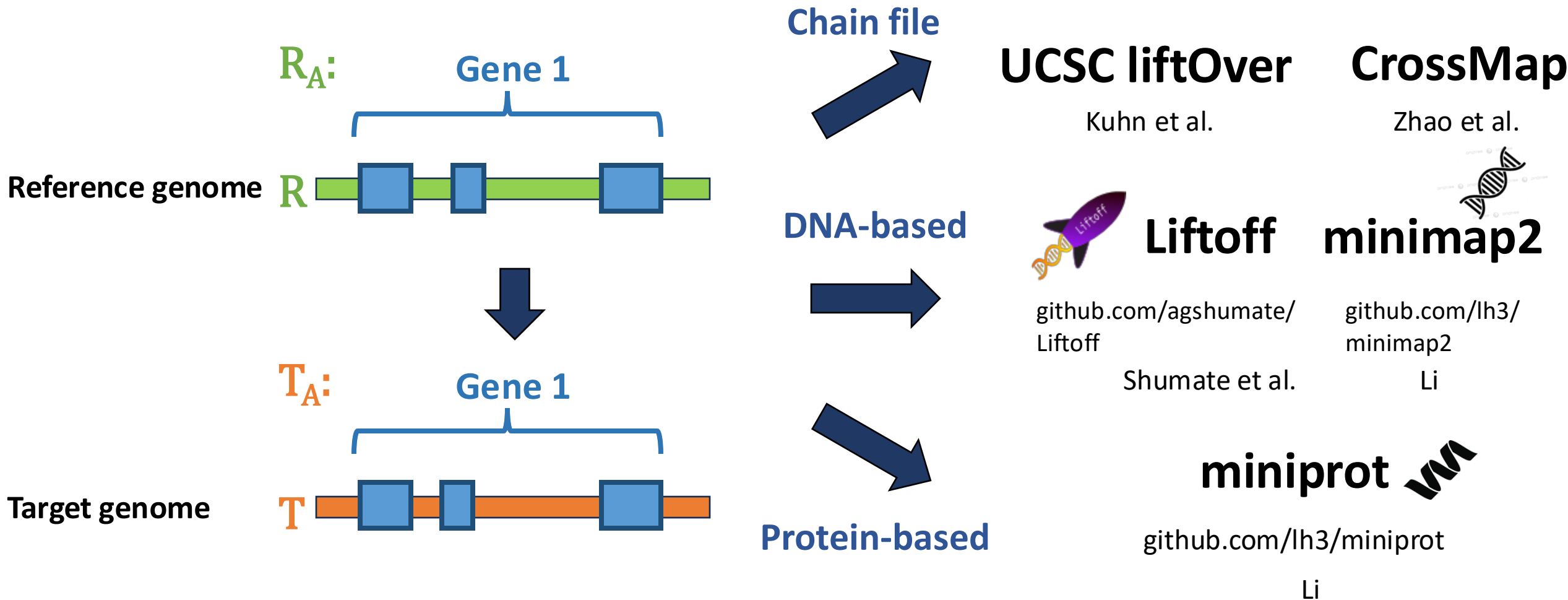
Lift-over Problem Definition:



Lift-over Problem Definition:



Lift-over problem, what methods are available?



Application: GRCh38 to T2T-CHM13 lift-over

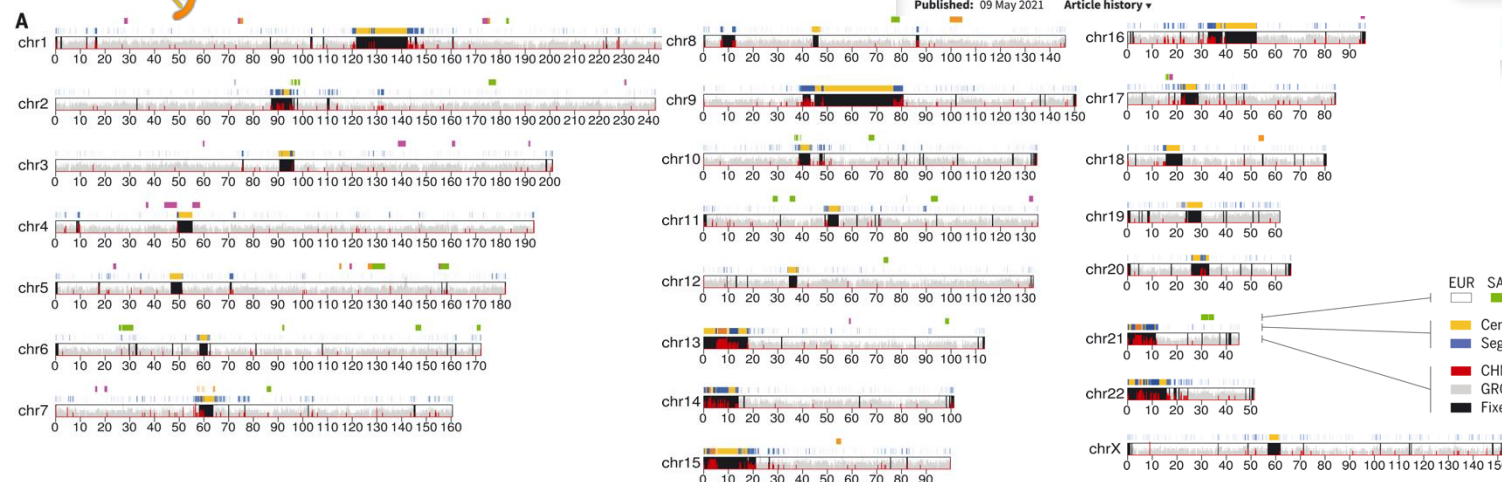
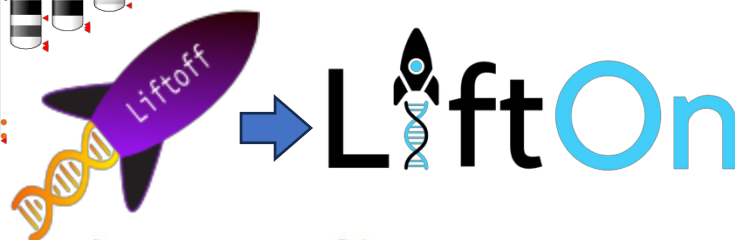
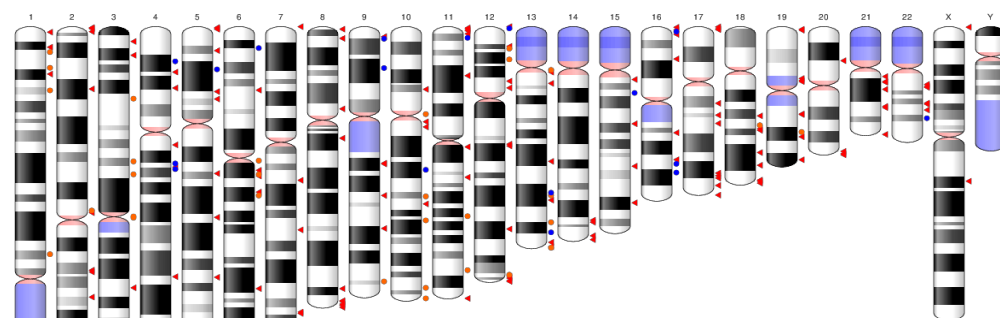
NCBI Build 34
NCBI Build 35
NCBI Build 36.1
GRCh37



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. AND YI-HONG 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

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, [...] AND ADAM M. PHILLIPPY +95 authors

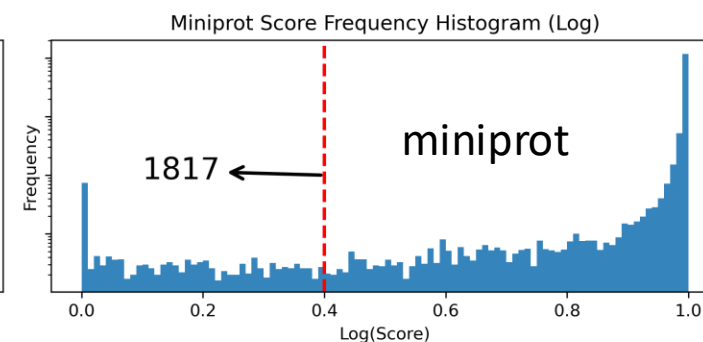
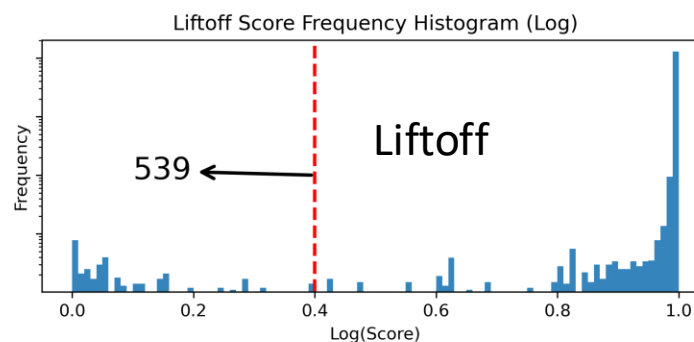
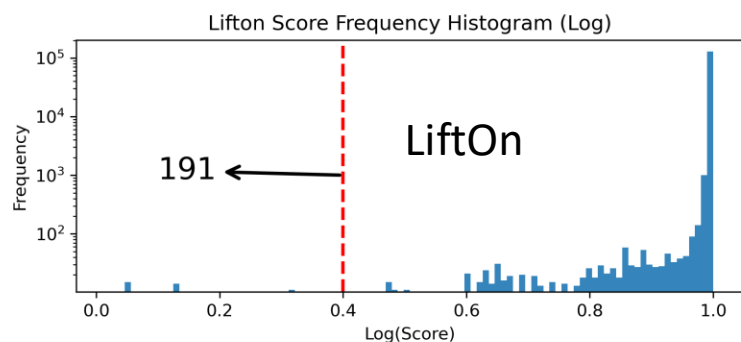
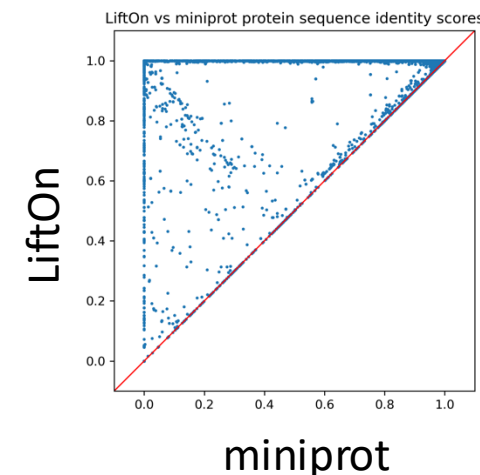
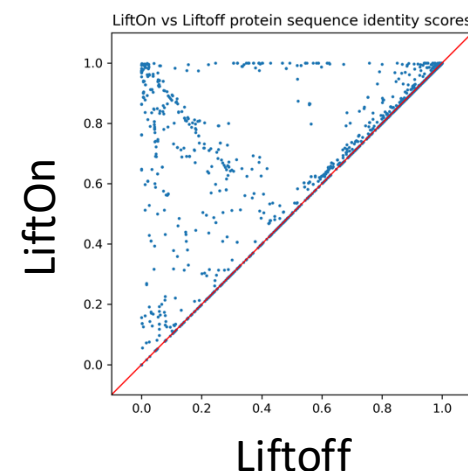
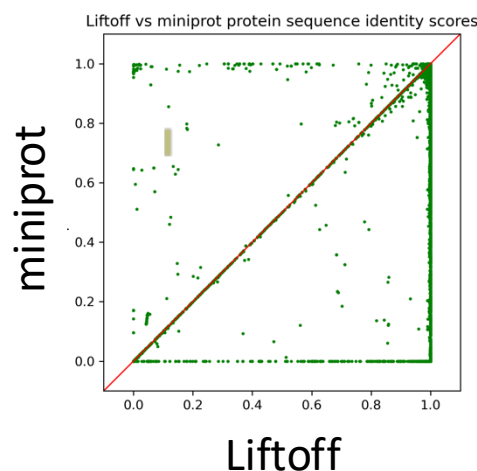
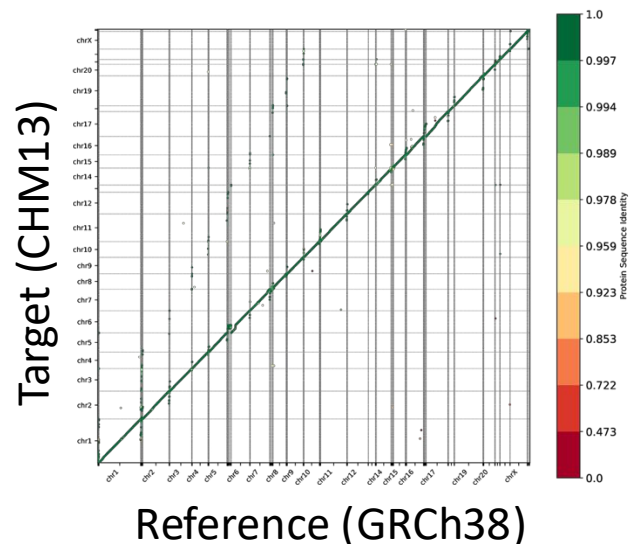
EUR SAS EAS AMR Ancestry
Centromeric satellites
Segmental duplications
CHM13 exclusive gene density
GRCh38 gene density
Fixed GRCh38 gaps and issues

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



Map RefSeq v220 from GRCh38 -> CHM13V2.0

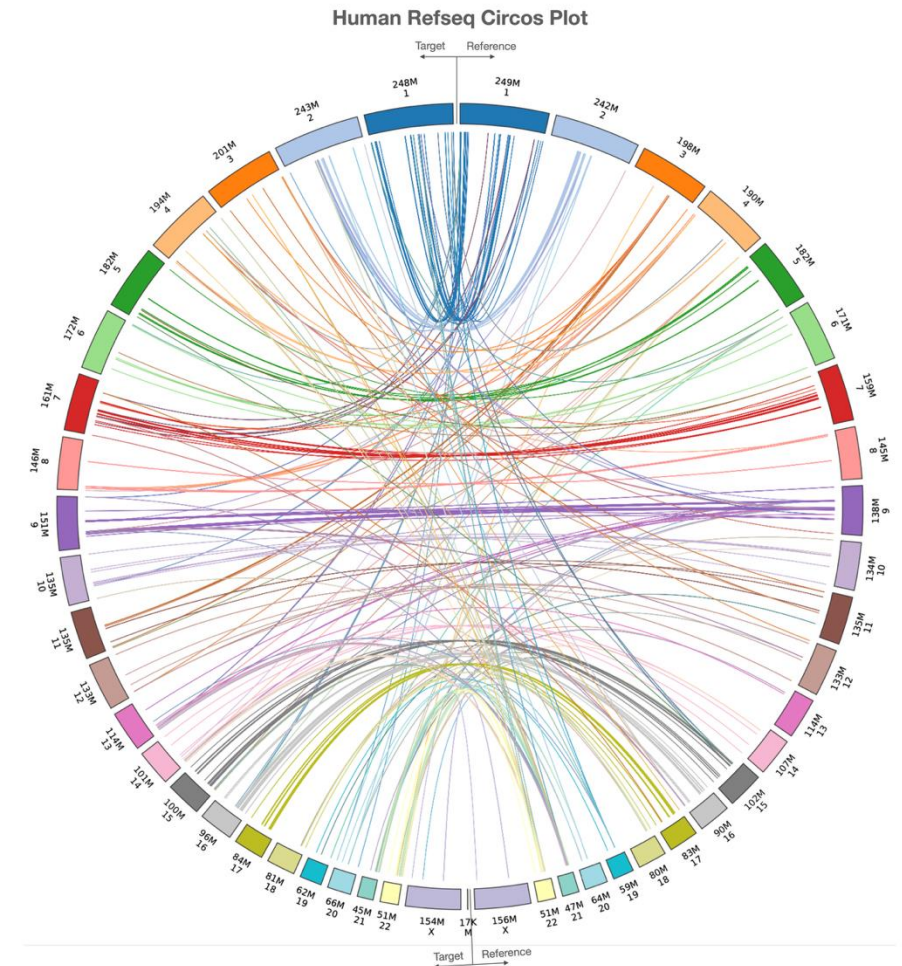
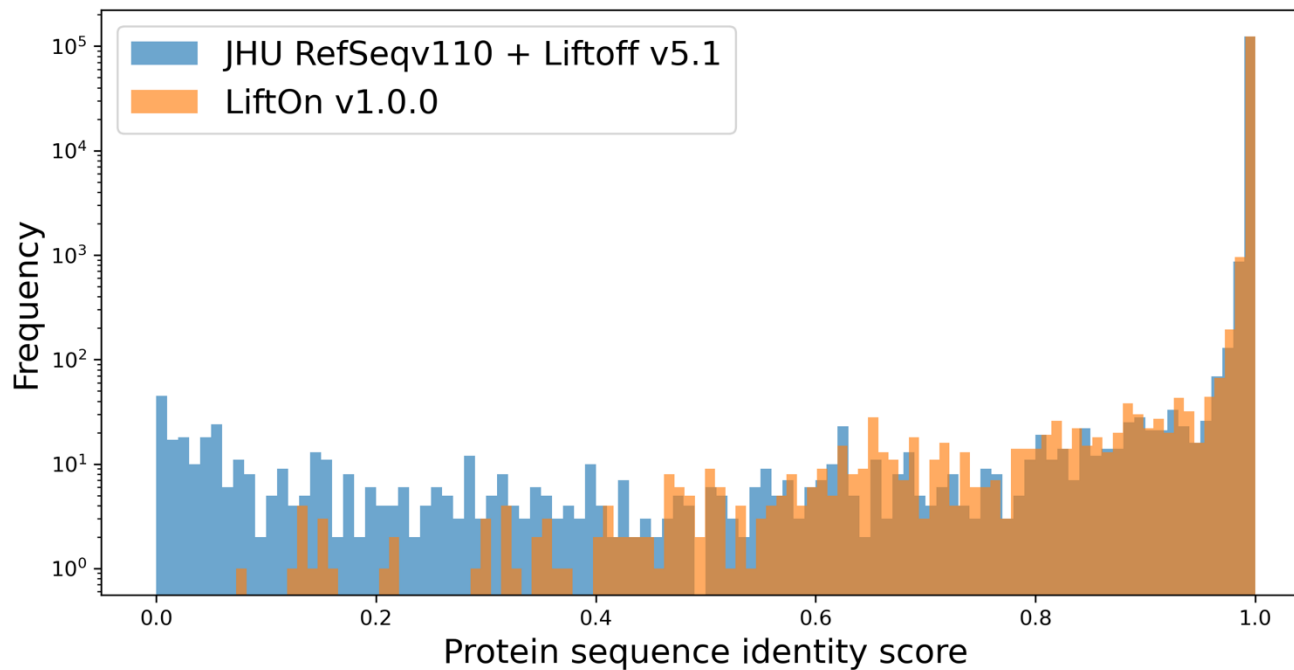
Compressed-gap protein sequence identity



Result 2: improve CHM13 protein annotations

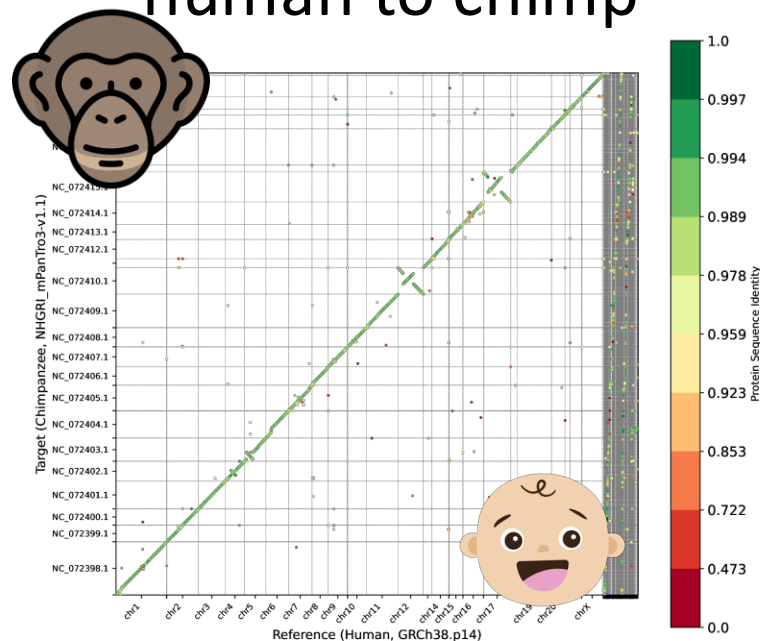


Protein sequence identity score frequency histogram



Result 3: improve distant species lift-over

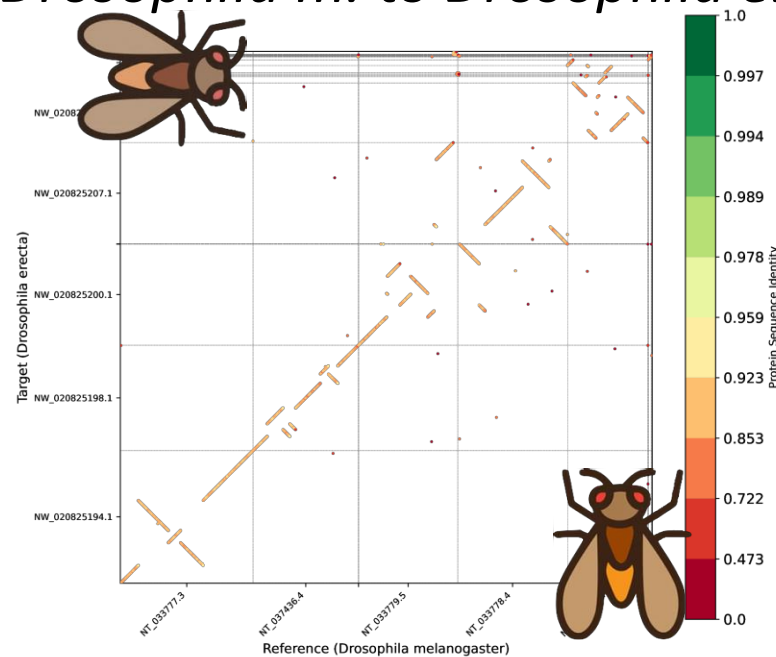
human to chimp



Mash: 0.013

Dashing2: 0.47

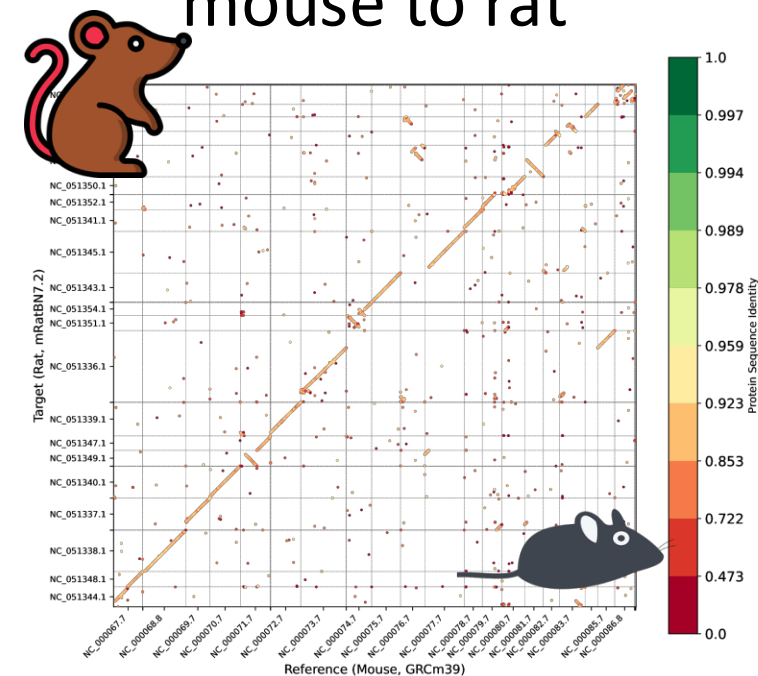
Drosophila m. to *Drosophila e.*



Mash: 0.077

Dashing2: 0.07

mouse to rat



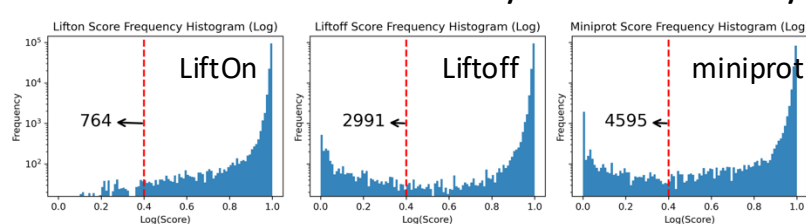
Mash: 0.12

Dashing2: 0.01

DNA + Protein

DNA-only

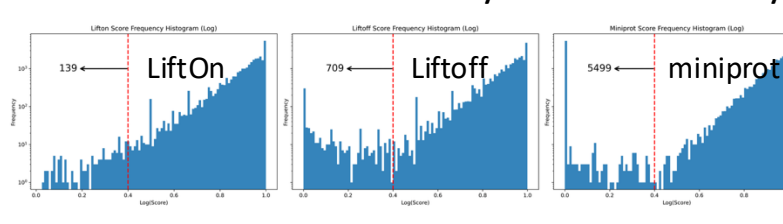
Protein-only



DNA + Protein

DNA-only

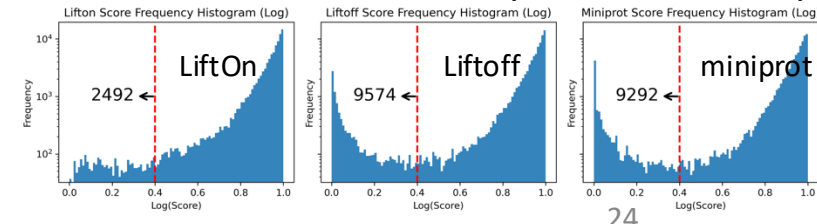
Protein-only



DNA + Protein

DNA-only

Protein-only



24

Introduction

Assembly & Indexing

Genome Annotation

Splice Site Prediction

RNA-Seq Prediction

L_{ft}On: Protein-maximization algorithm

A

Target genome +
Expected annotation



1. Liftoff annotation



2. miniprot annotation



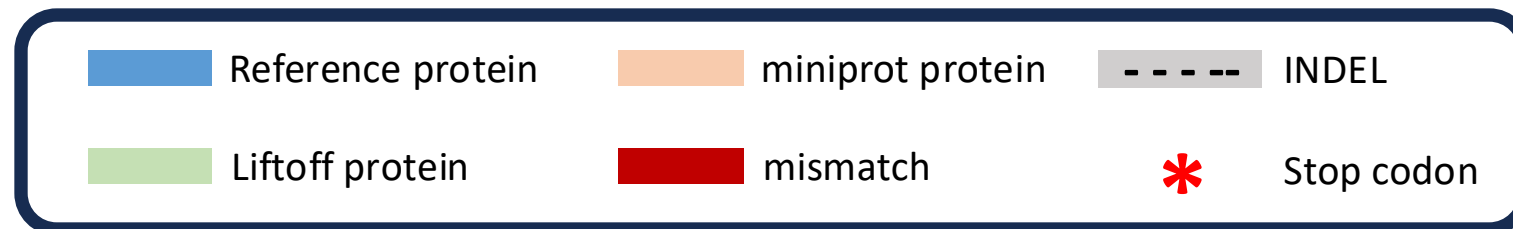
LiftOn: Protein-maximization algorithm

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

Liftoff protein alignment

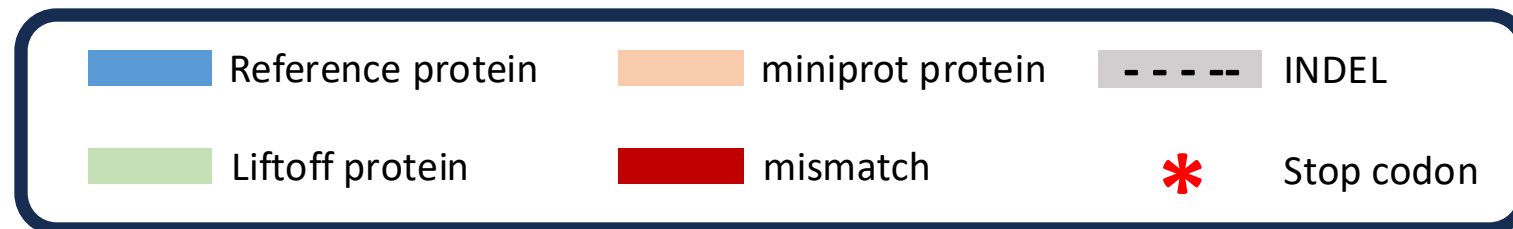
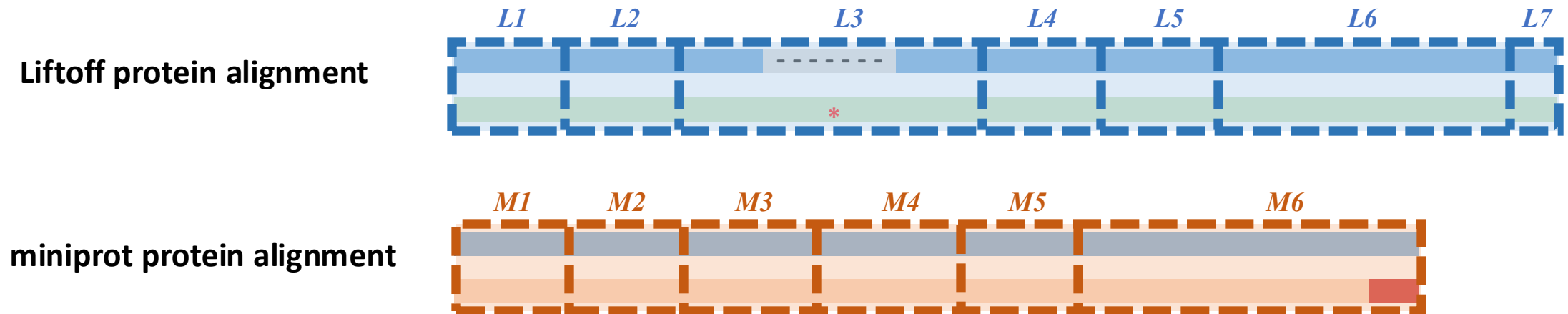


miniprot protein alignment



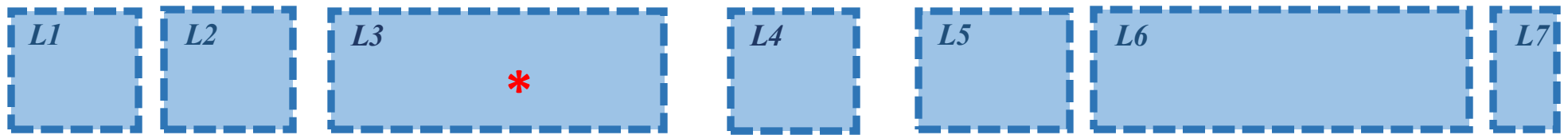
LiftOn: Protein-maximization algorithm

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”

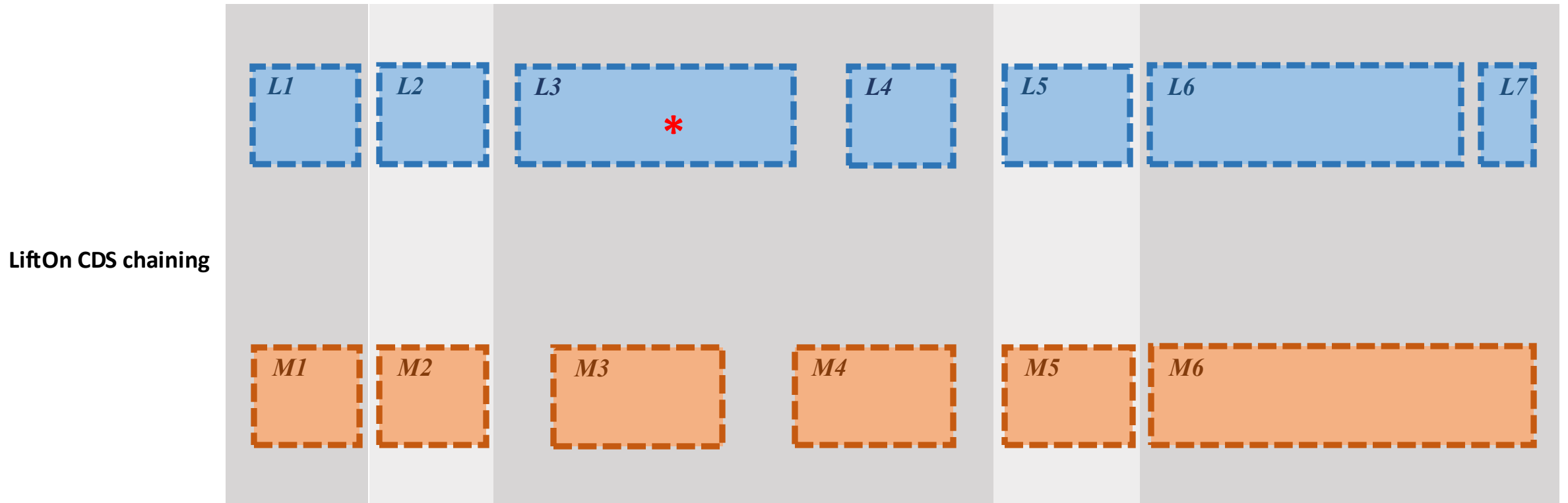


LiftOn CDS chaining



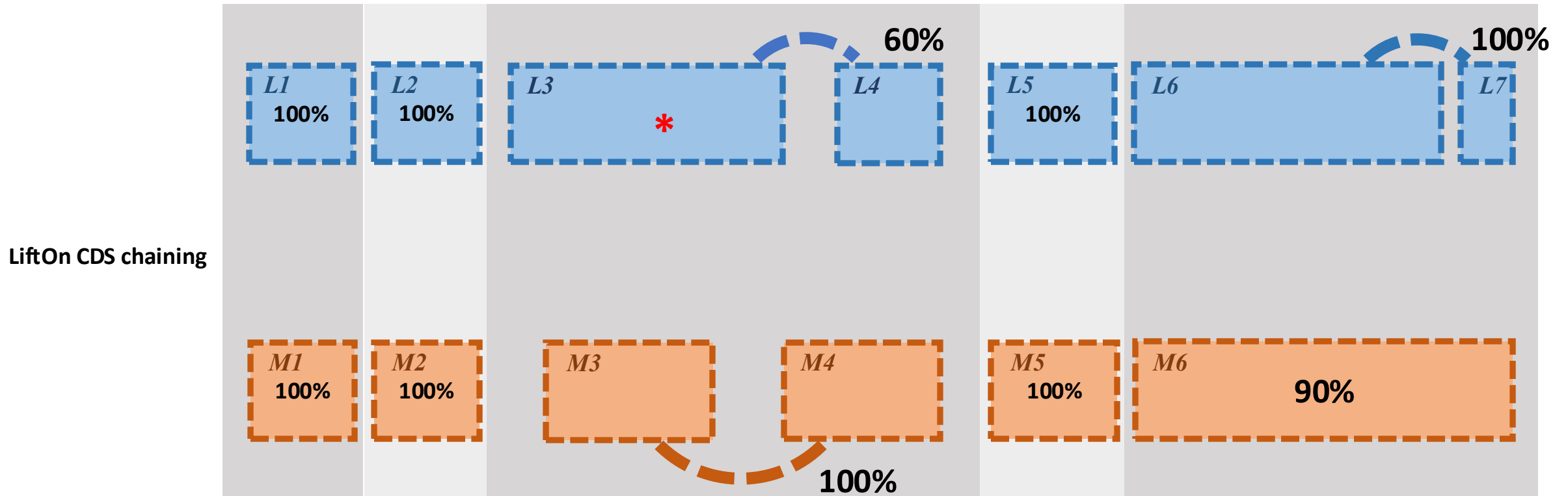
LiftOn: Protein-maximization algorithm

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



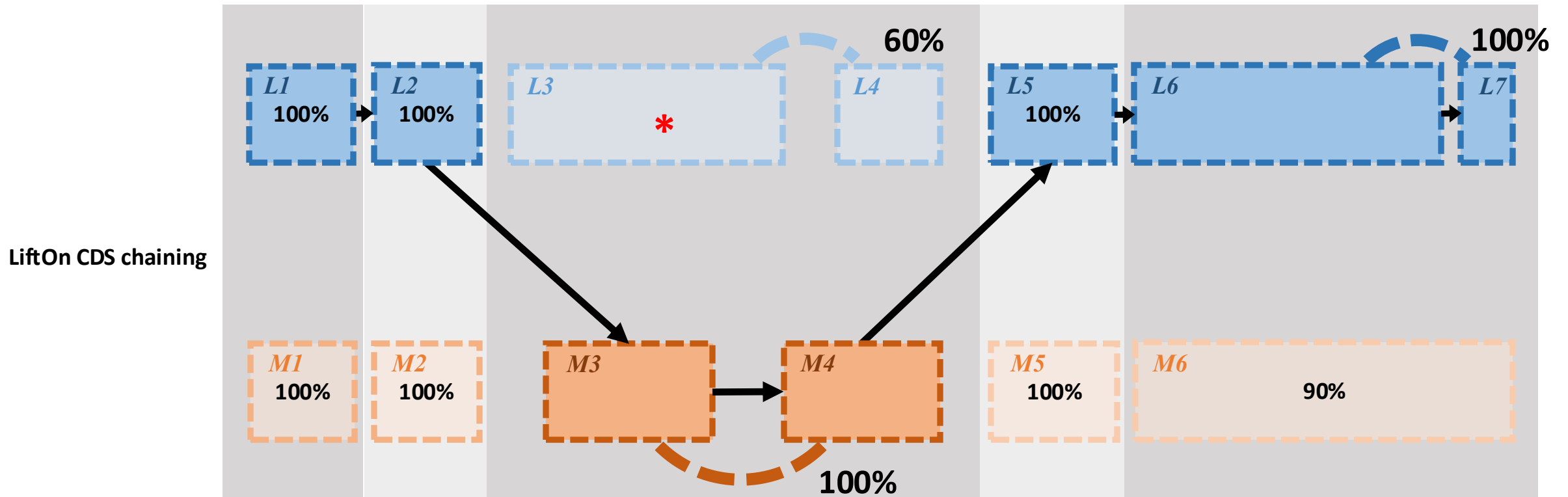
LiftOn: Protein-maximization algorithm

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

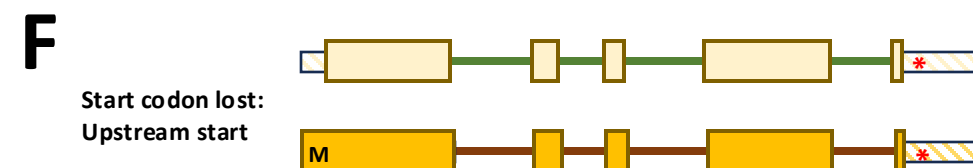
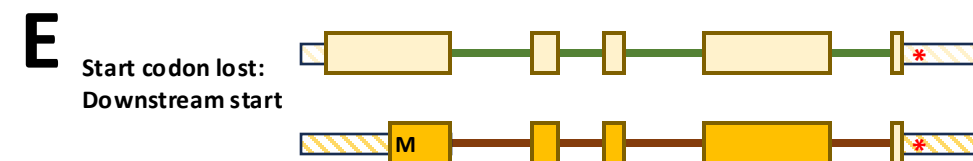
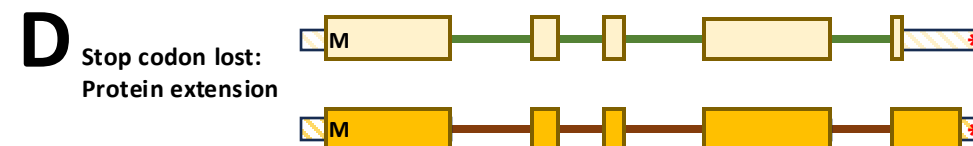
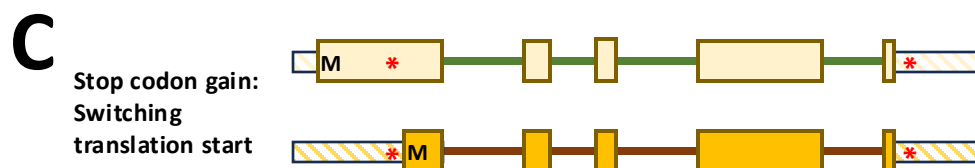
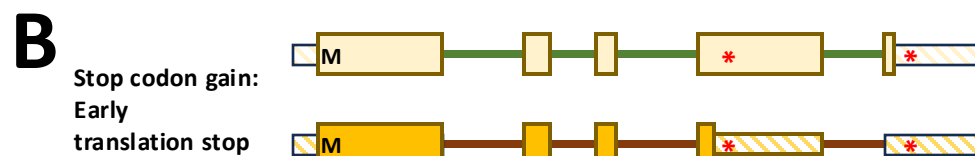
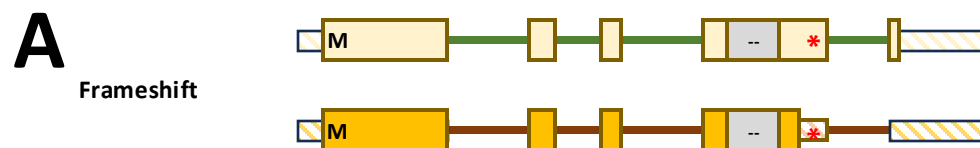
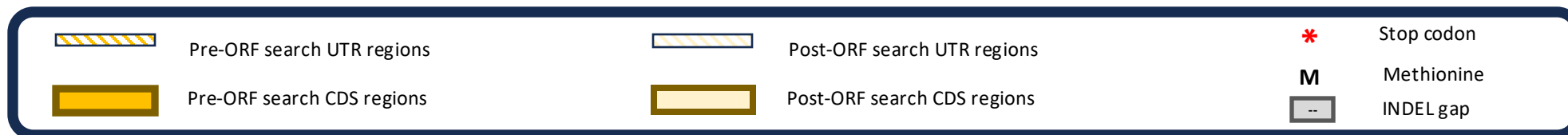


LiftOn: Protein-maximization algorithm

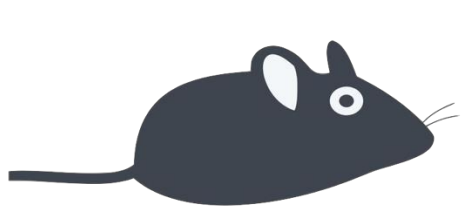
D Step 3: group CDSs by “accumulated AA in the reference protein”



LiftOn: Protein-maximization algorithm



LiftOn: more lift-over applications



House mouse
(*Mus musculus*)



Yeast
(*Saccharomyces cerevisiae*)



Thale cress
(*Arabidopsis thaliana*)



Honey bee
(*Apis mellifera*)

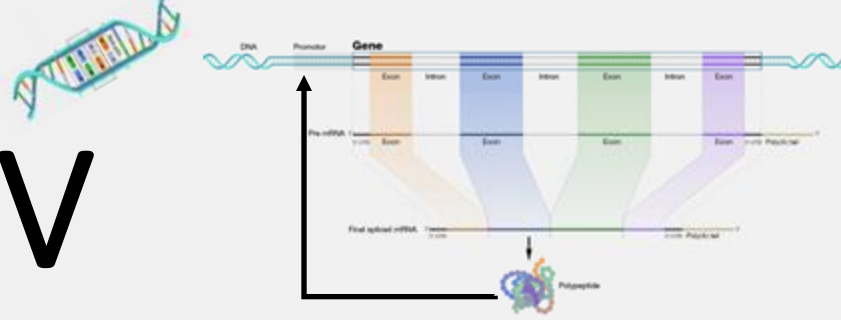


fruit fly
(*Drosophila melanogaster*)



Rice
(*Oryza sativa*)

Part IV & V



Mihaela Pertea



Steven Salzberg



Anqi Liu

Splice Site Prediction

- OpenSpliceAI
- Splam



Genome Biology

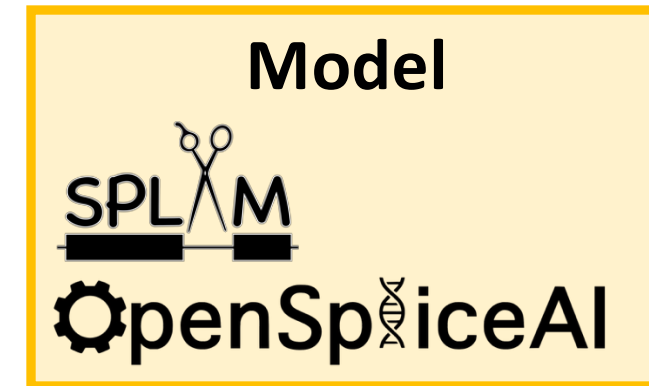
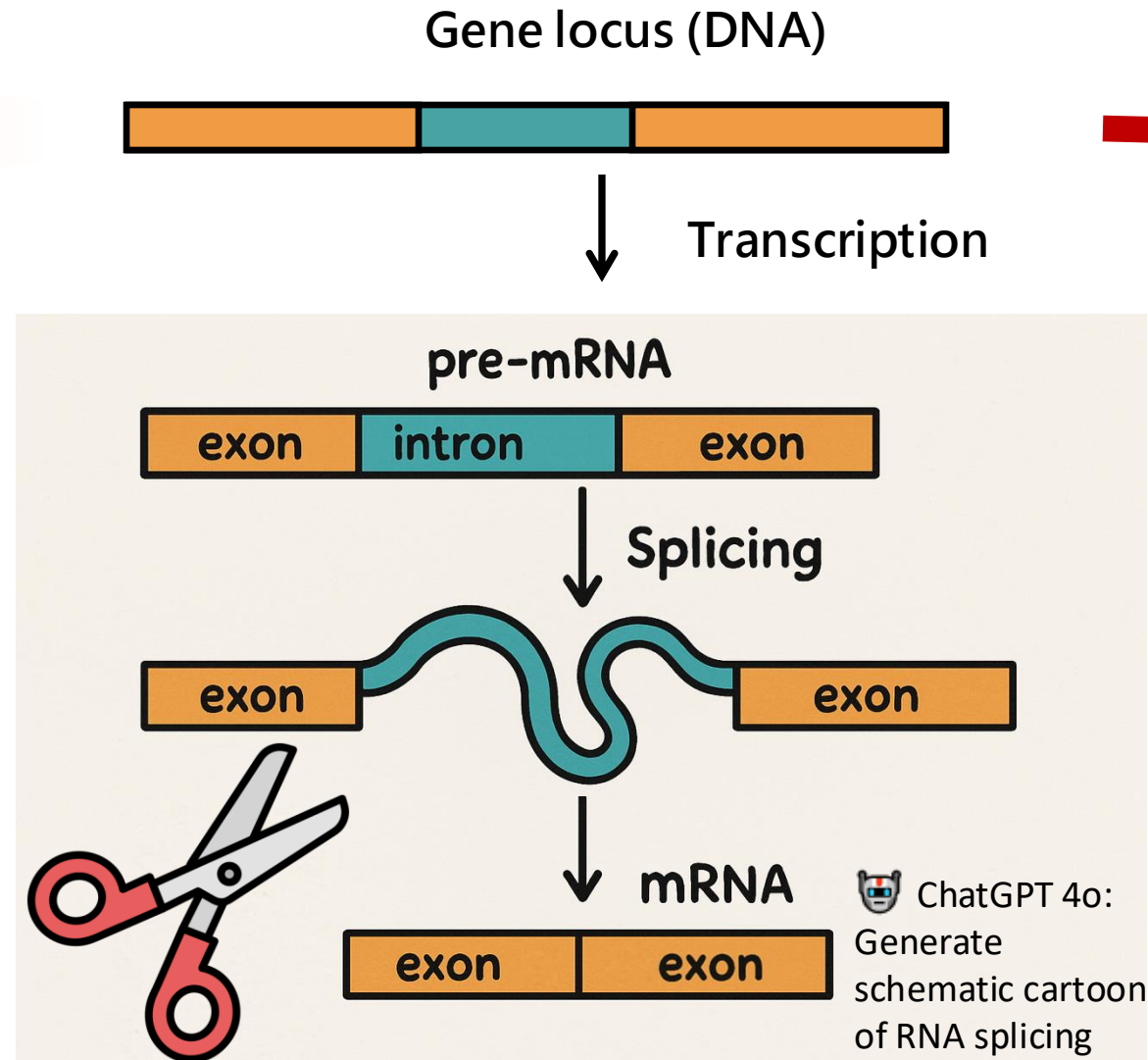
Chao, K. H., Mao, A., Salzberg, S. L., & Pertea, M. (2024). Splam: a deep-learning-based splice site predictor that improves spliced alignments. *Genome biology*, 25(1), 243.

OpenSpliceAI



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



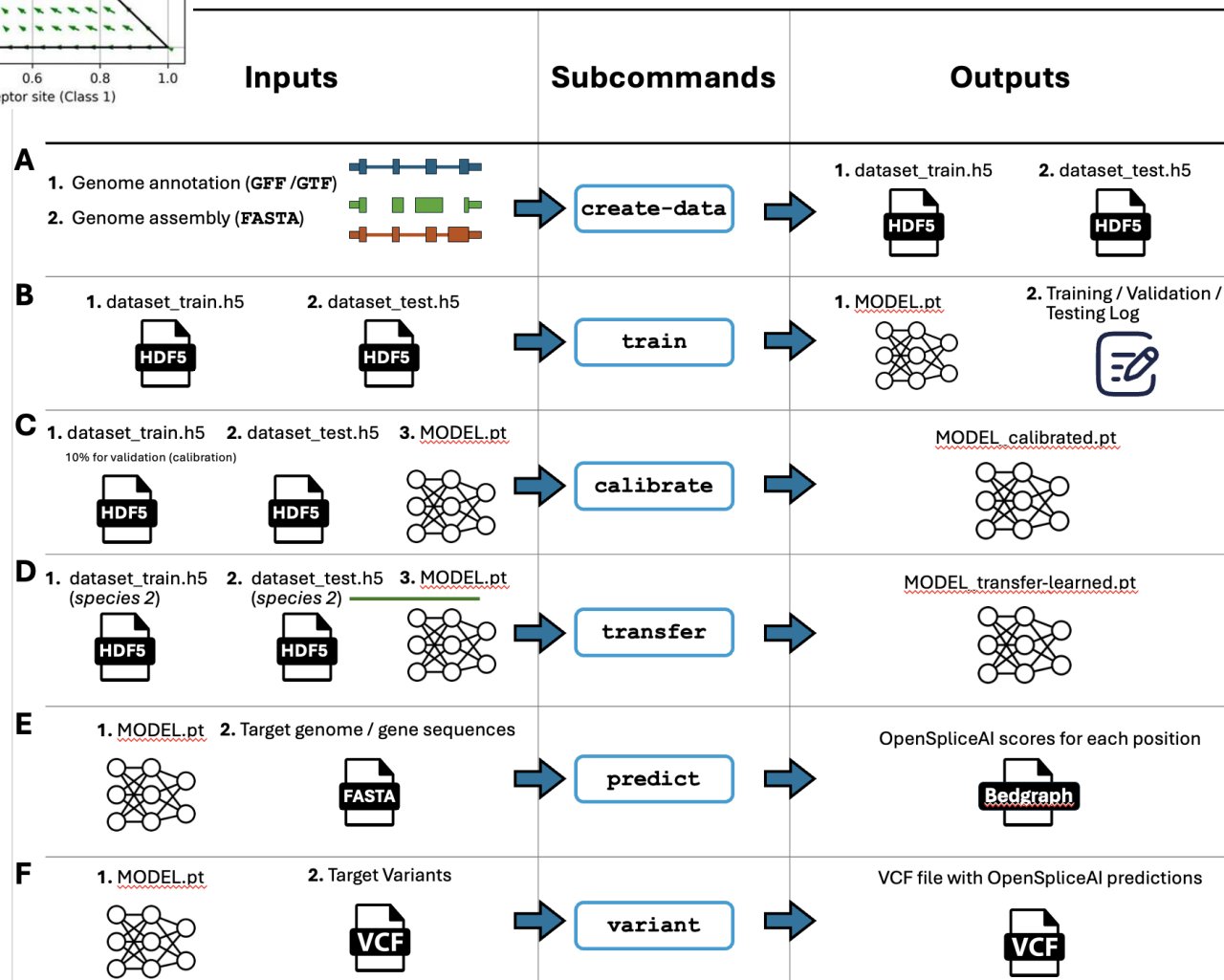
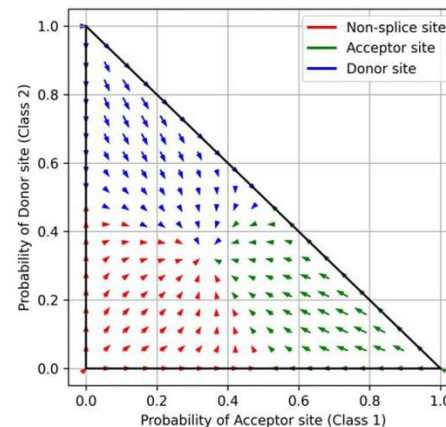
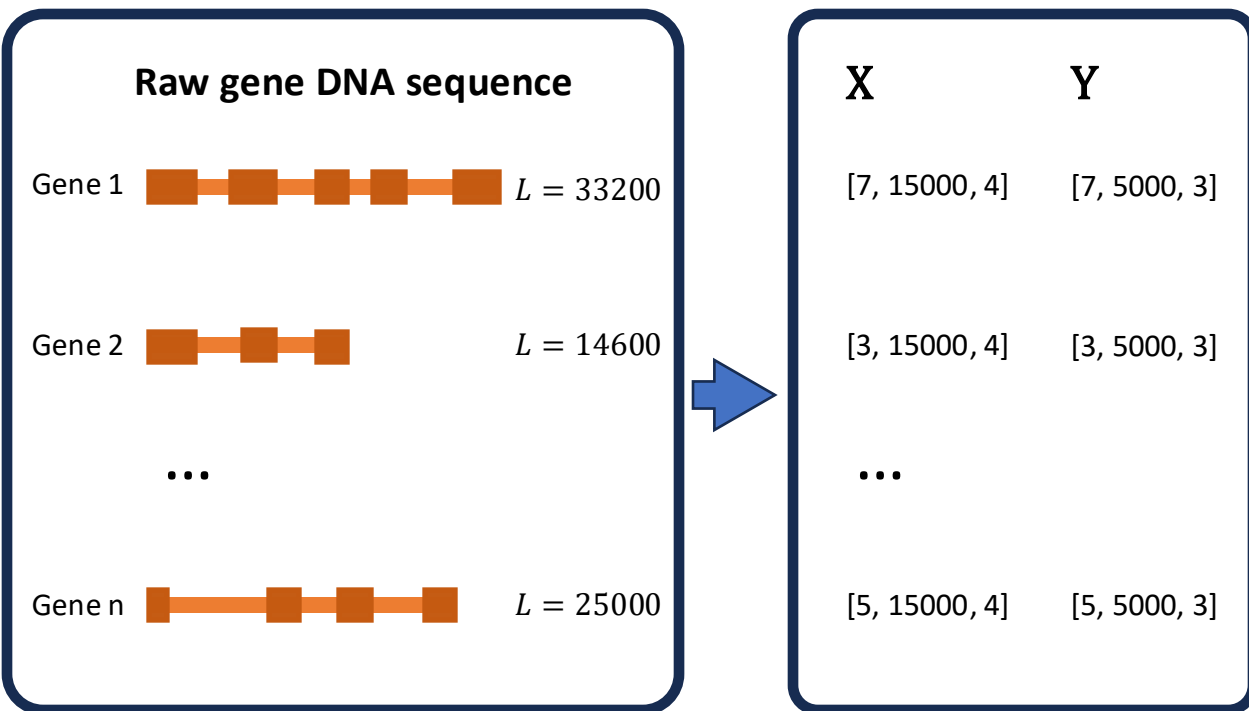
Where are the splice sites?

[illegible]

Neither: 0

OpenSpliceAI

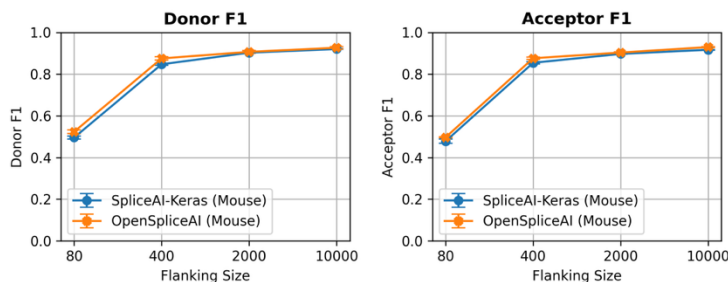
~20k protein-coding genes



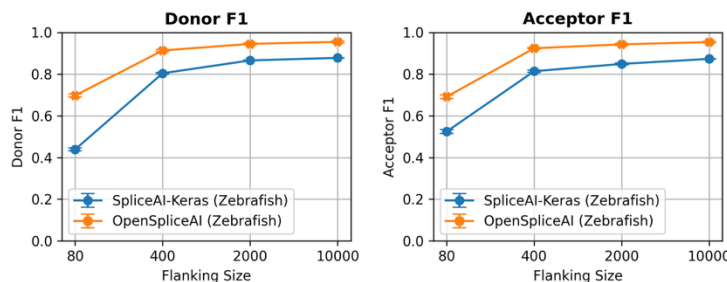
OpenSpliceAI: retrain on different species



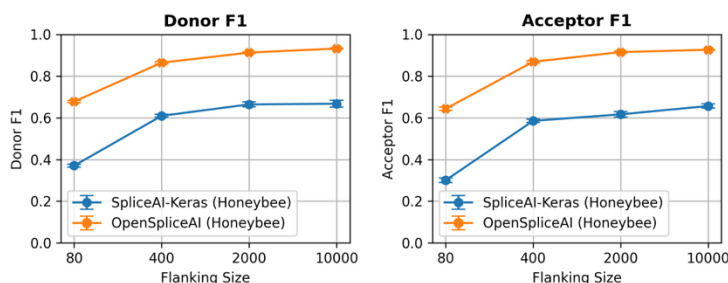
Splice site prediction metrics for Mouse



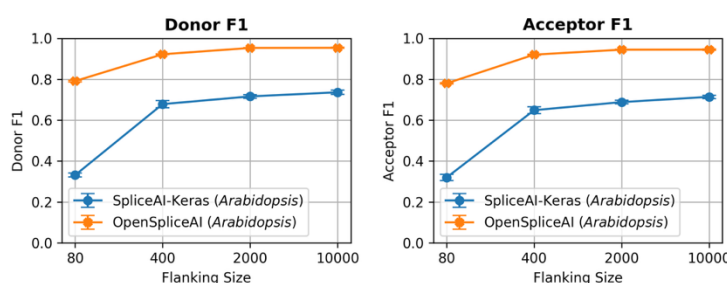
Splice site prediction metrics for Zebrafish



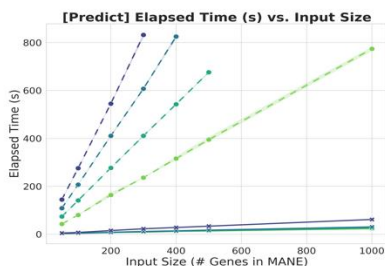
Splice site prediction metrics for Honeybee



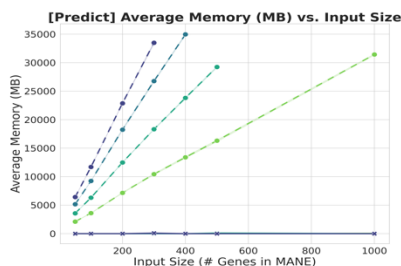
Splice site prediction metrics for *Arabidopsis*



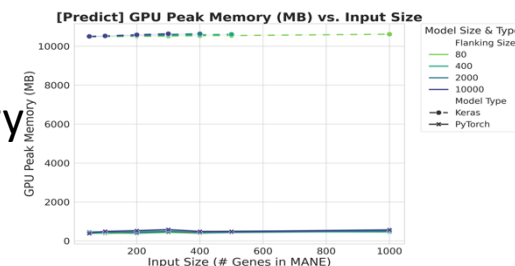
System time

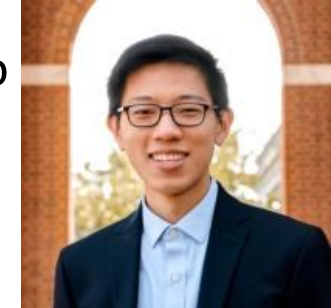


Memory

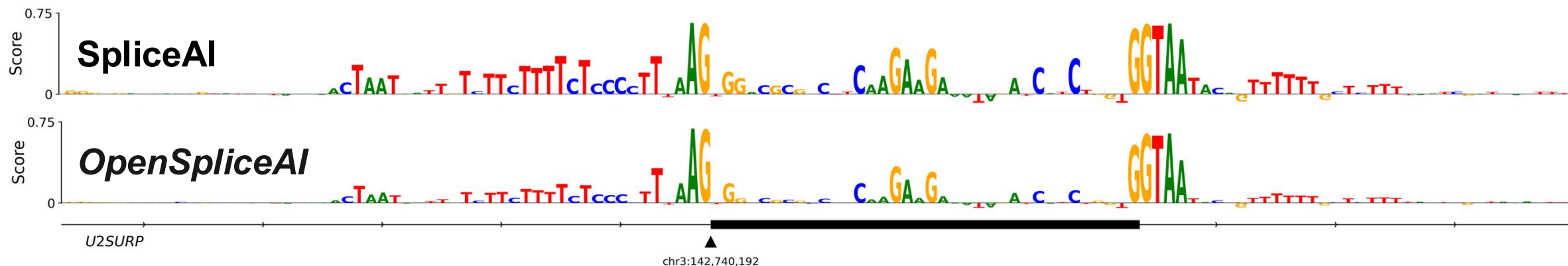


GPU Avg memory

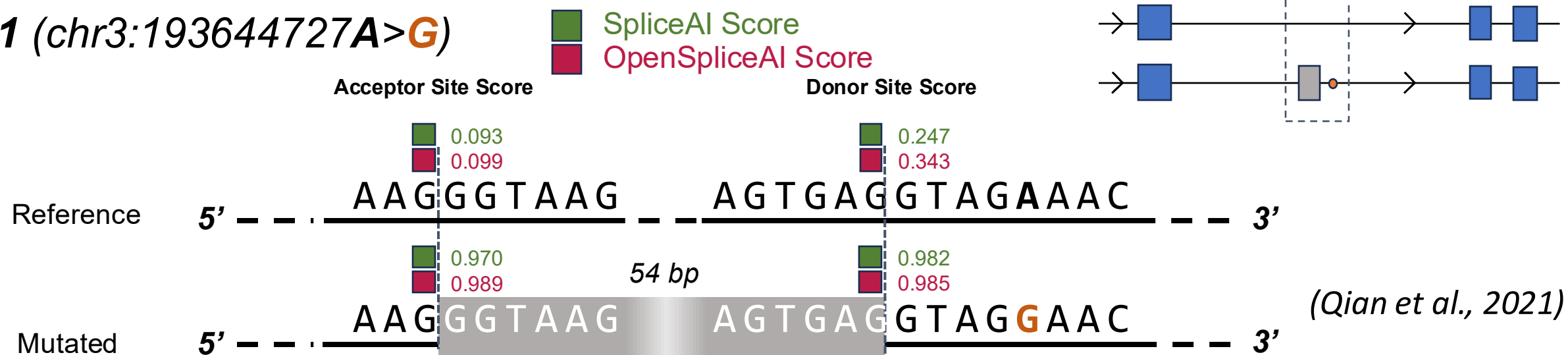




OpenSpliceAI: ISM & variant prediction



OPA1 (chr3:193644727A>G)

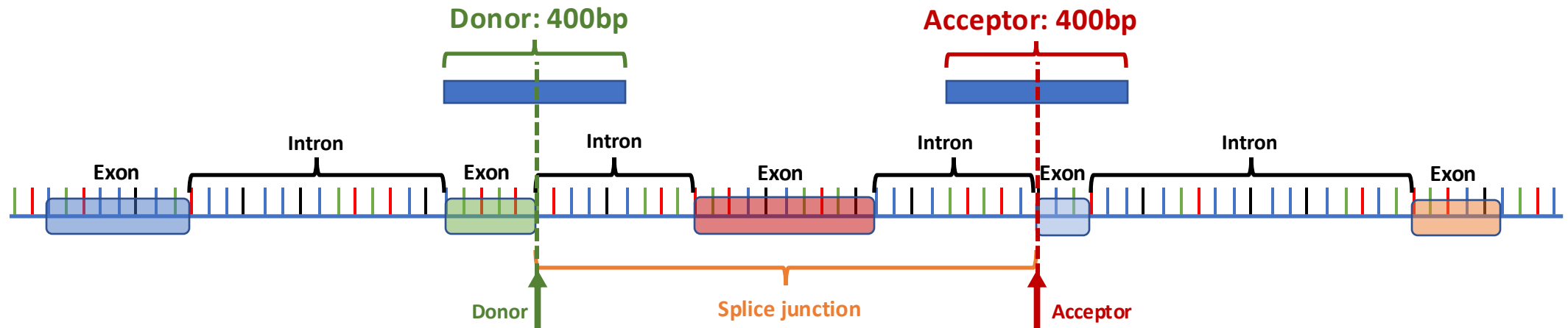




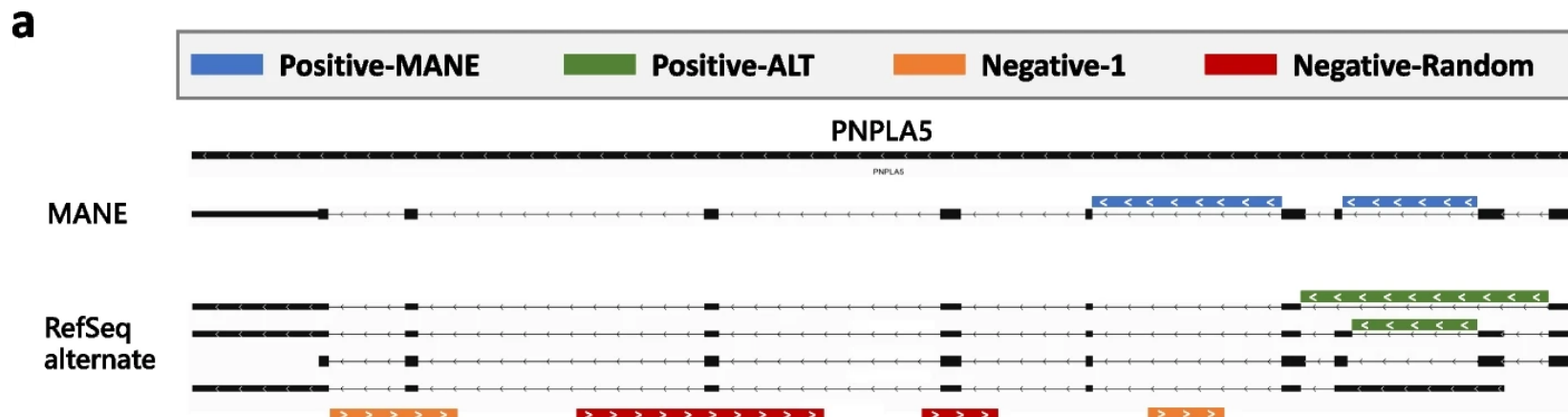
: predicting splice sites in a pair



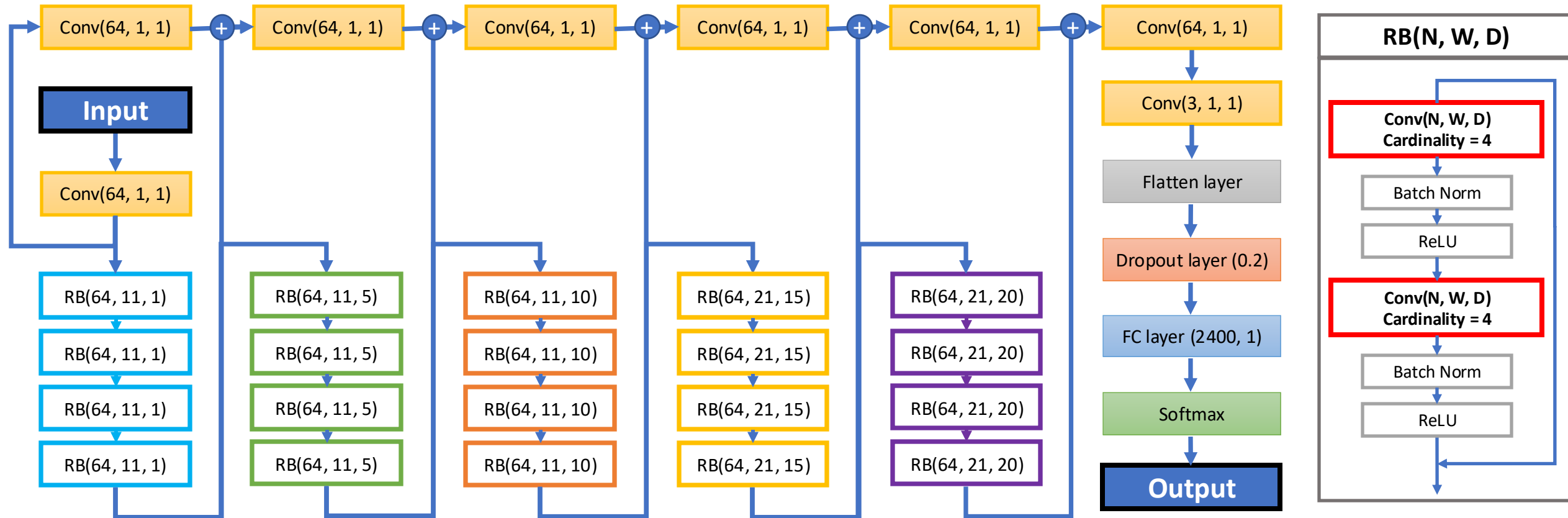
Chao, K. H., Mao, A., Salzberg, S. L., & Pertea, M. (2024). Splam: a deep-learning-based splice site predictor that improves spliced alignments. Genome biology, 25(1), 243.



Data

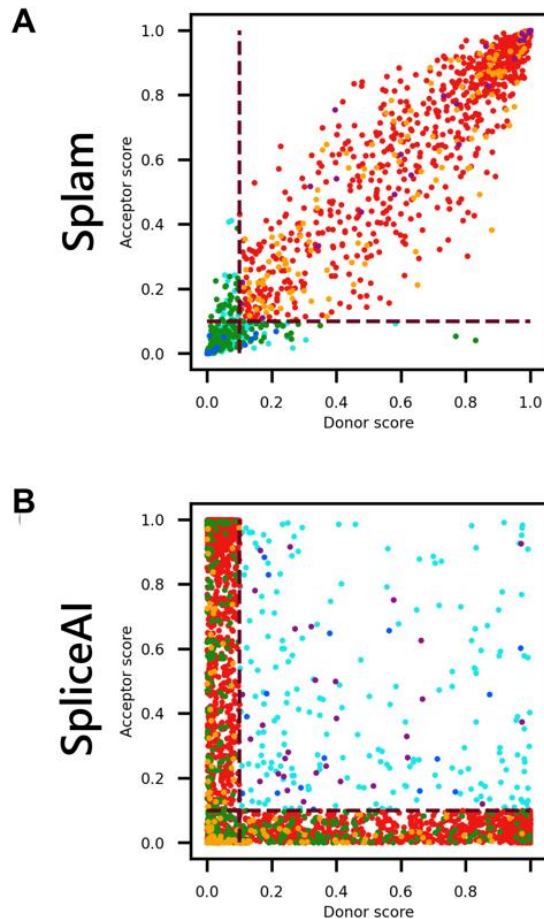


SPLM : Splam Model Architecture

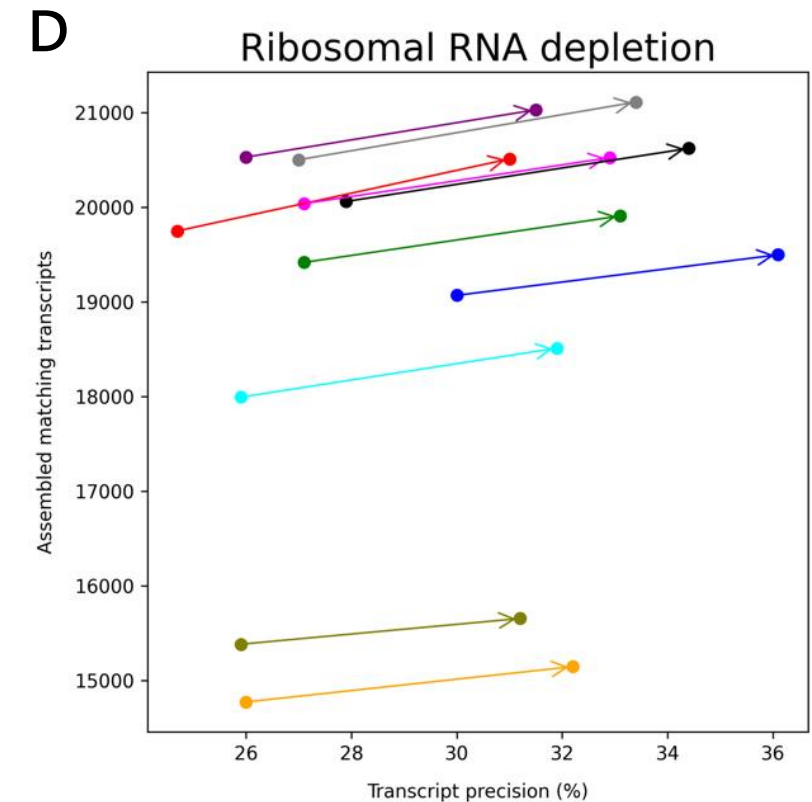
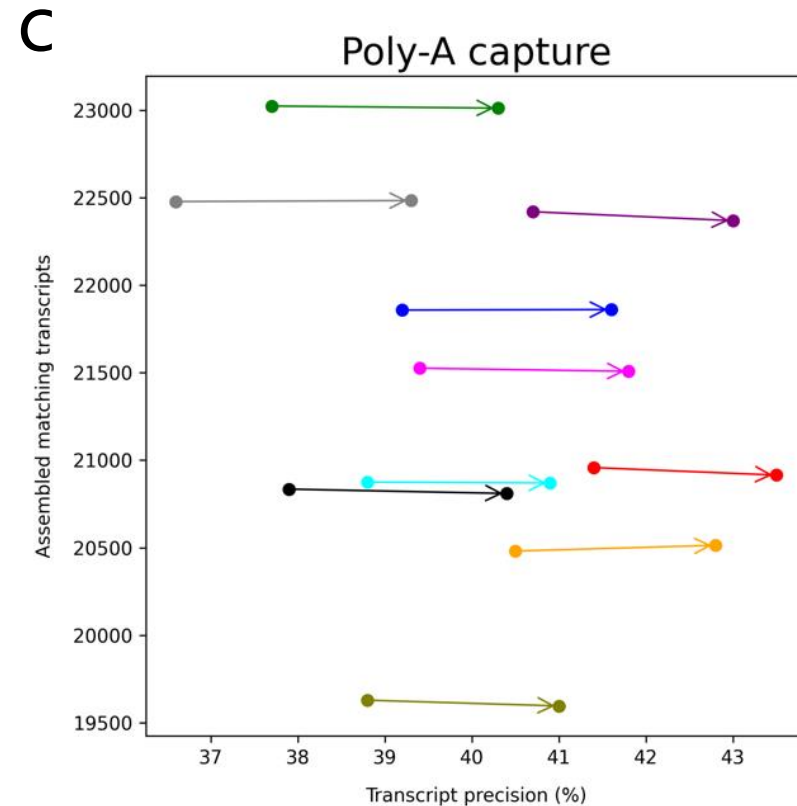


: improving transcriptome assembly

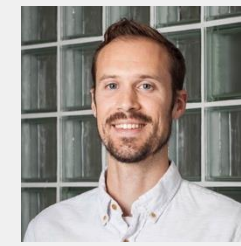
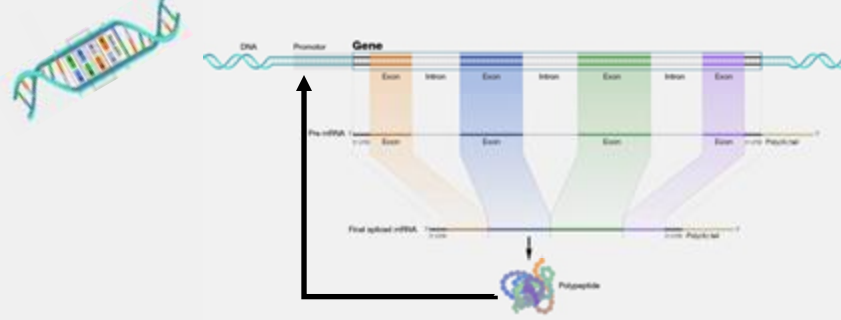
Score stability



Transcriptome assembly improvement



Part VI



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



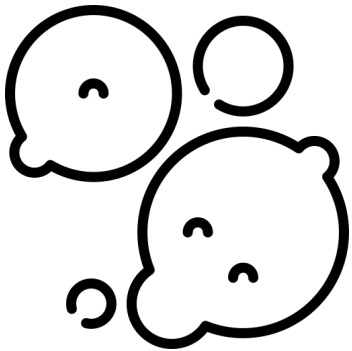
Chao, K. H., Magzoub, M., Stoops, E., Hackett, S., Linder J., * and Kelley, D. R., (manuscript in preparation). Predicting dynamic expression patterns in budding yeast with a fun-gal DNA language model

Manuscript in preparation.
Expected to be on bioRxiv soon

Questions we answer in this study:

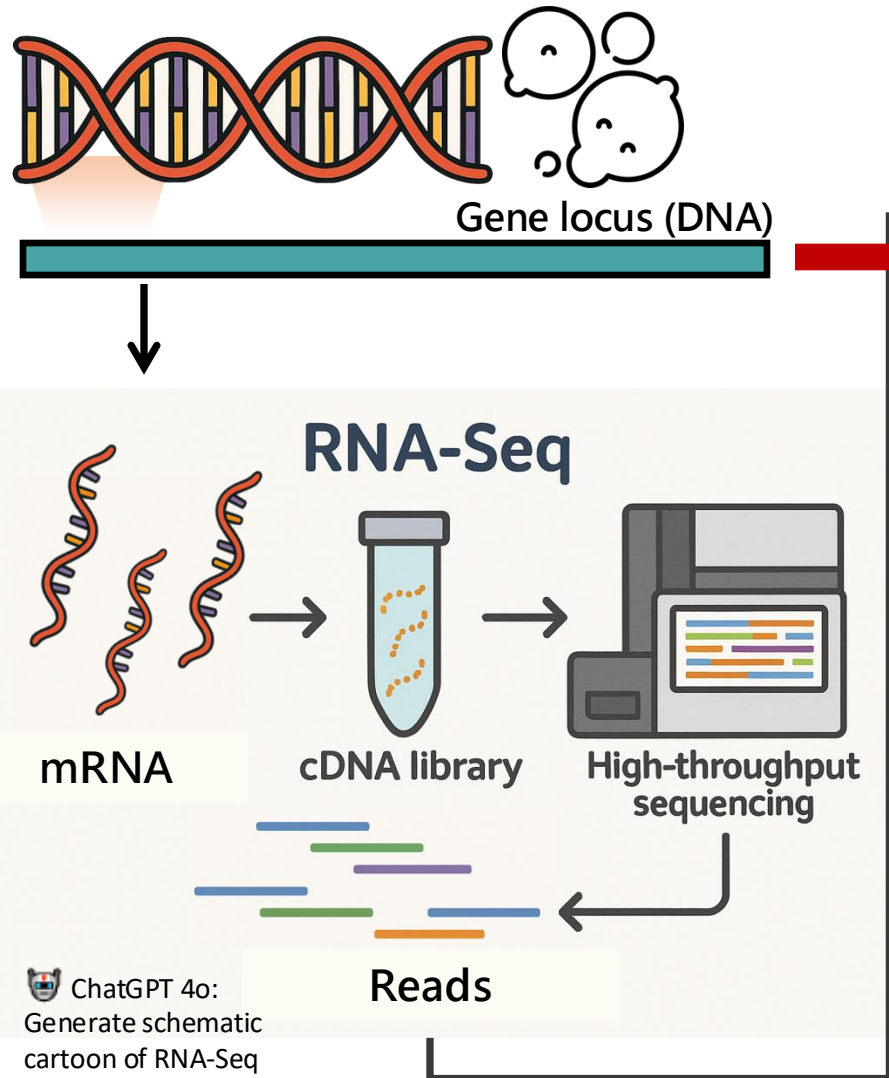
Main Goal

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

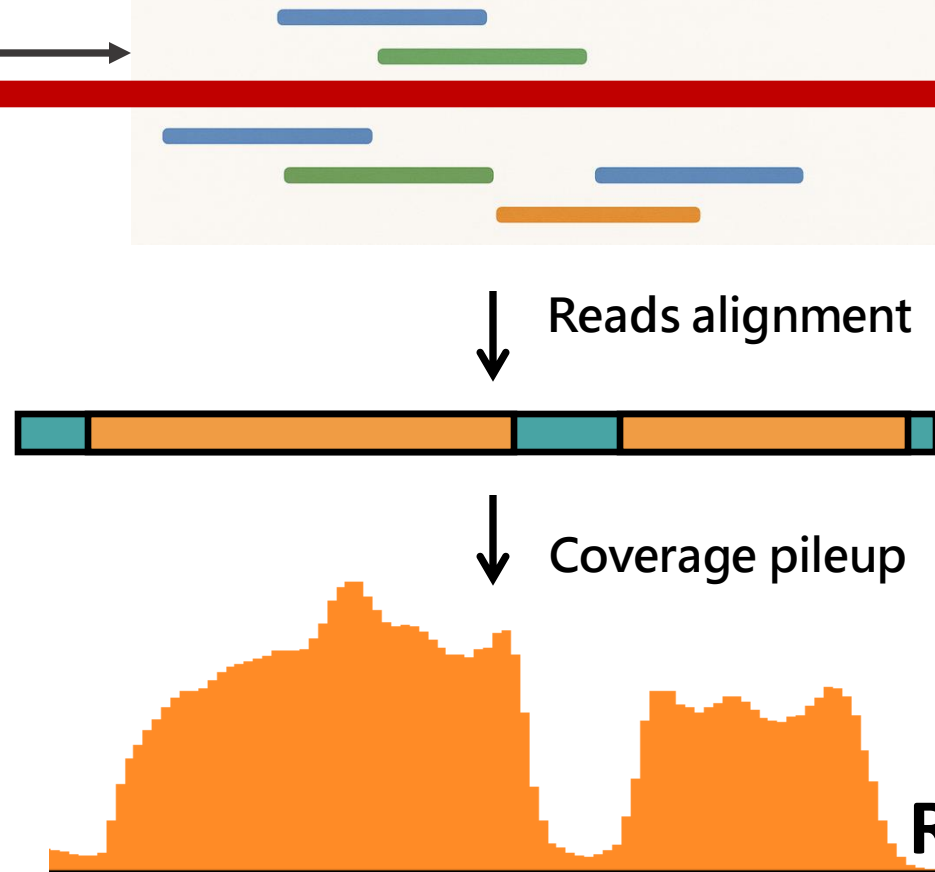


Yeast RNA-Seq Prediction

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



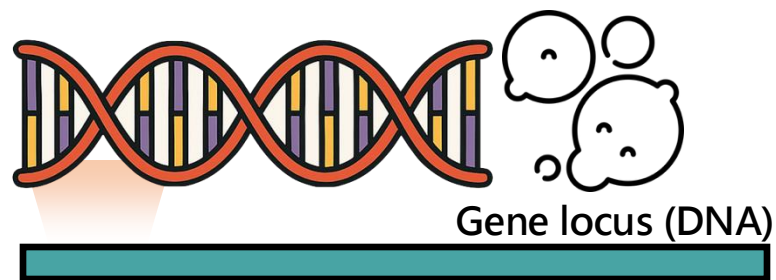
RNA-Seq Coverage



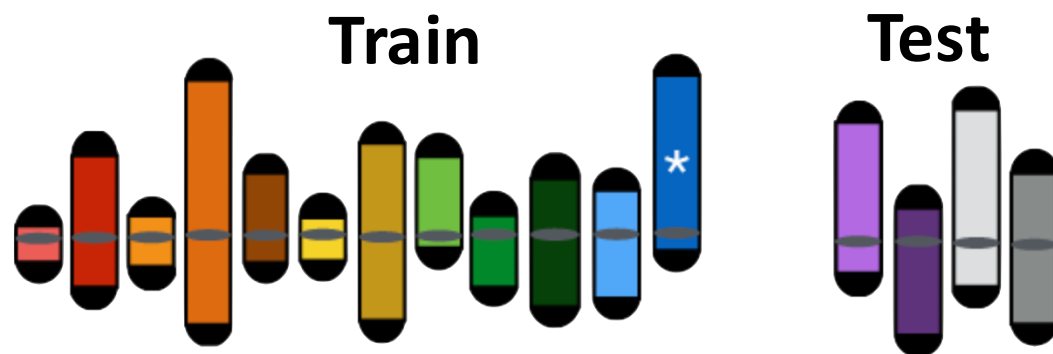
**Gene expression
model**

RNA-Seq expression?

Yeast RNA-Seq Prediction



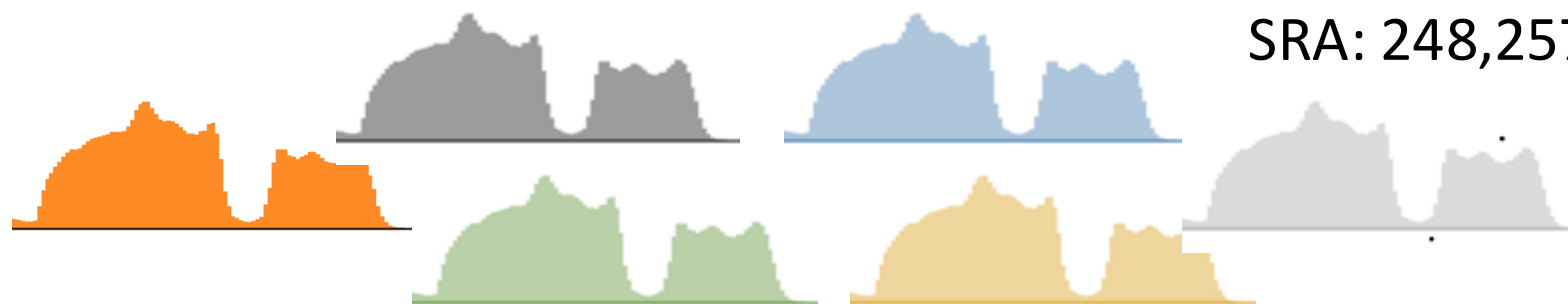
ChatGPT 4o:
Generate schematic cartoon
of RNA-Seq coverage



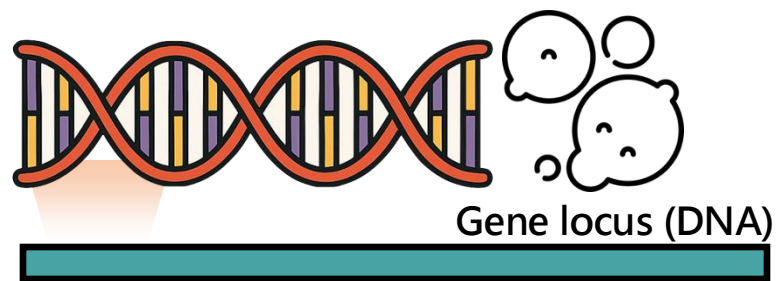
12 million base pairs

Genome size is too small, leading to overfitting.

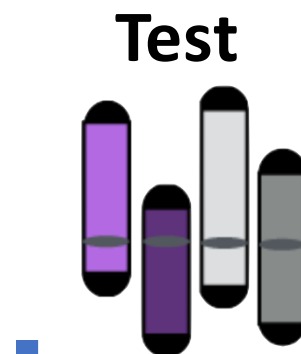
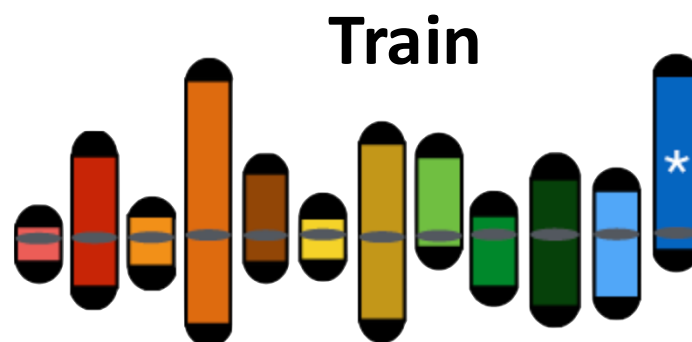
SRA: 248,257 search entries



Yeast RNA-Seq Prediction



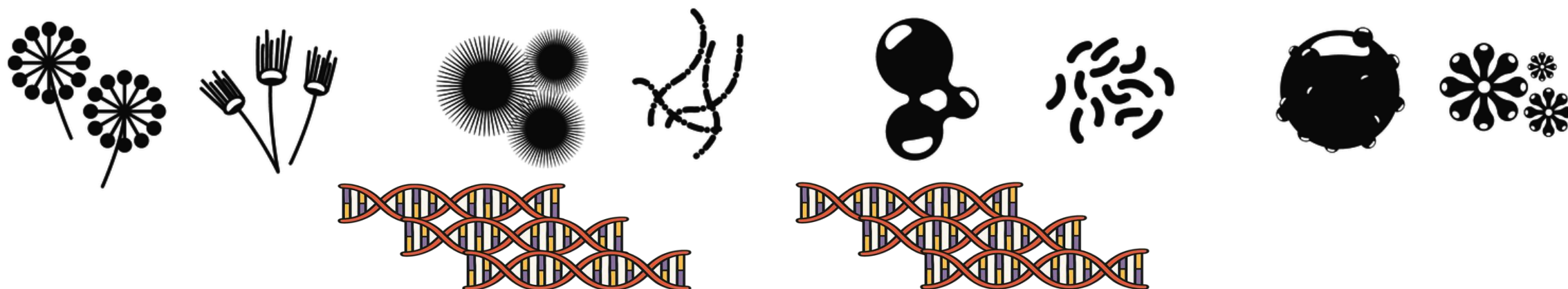
ChatGPT 4o:
Generate schematic cartoon
of RNA-Seq coverage



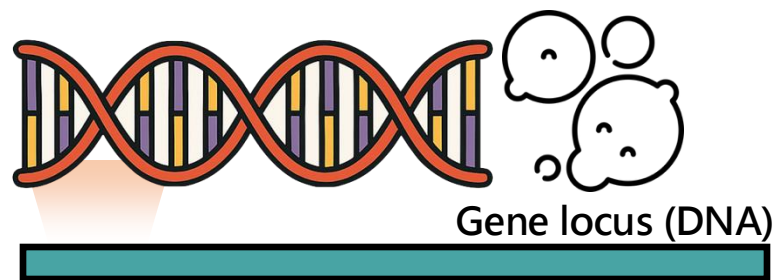
12 million base pairs

Fungal LM

Ensembl fungi: 1500 genomes

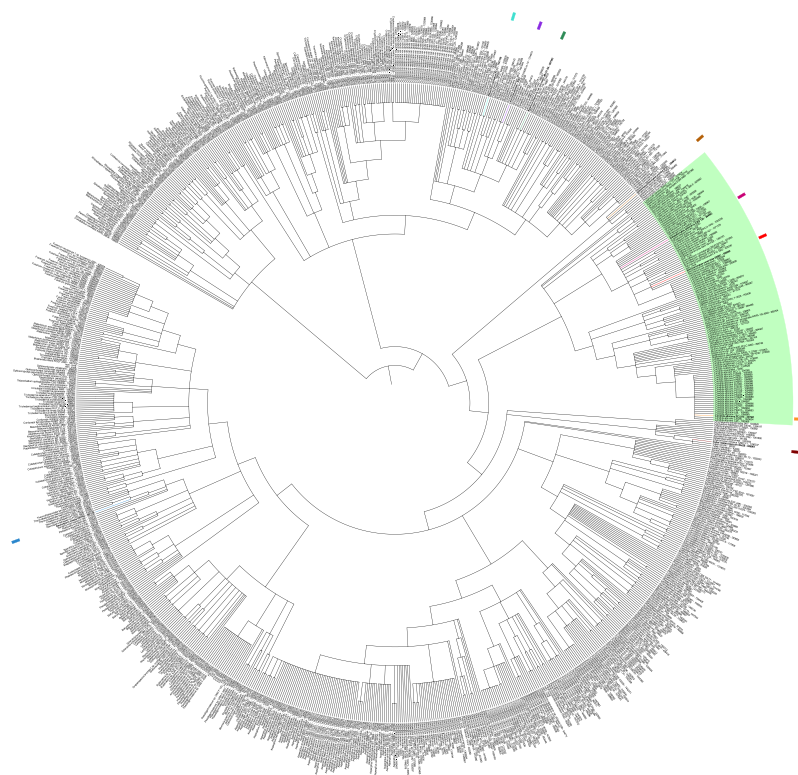
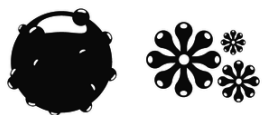
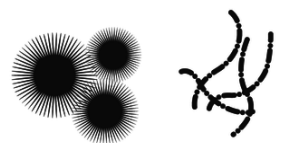


Yeast RNA-Seq Prediction



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

Fungal LM



Gene expression
model



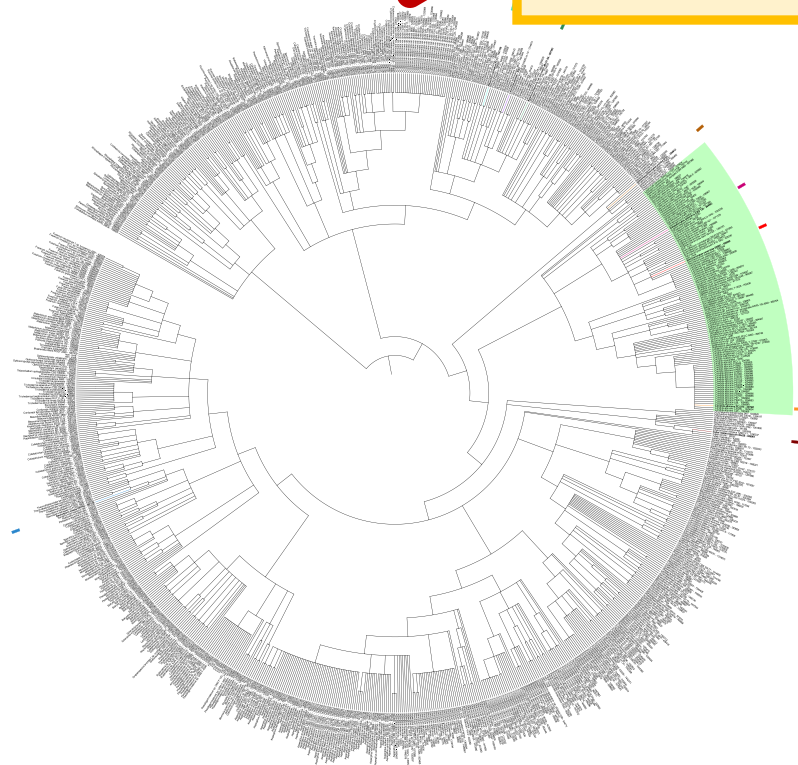
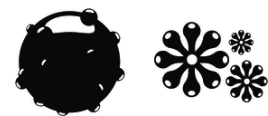
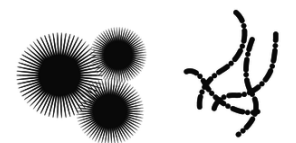
RNA-Seq expression?

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

Troyanskaya Lab Princeton

nature > nature

Brief Communication | Published: 24 August 2015

Predicting effects of noncoding RNA on gene expression with deep learning-based sequence models

Jian Zhou & Olga G Troyanskaya

DeepSEA
2015

CSH PRESS GENOME RESEARCH

HOME | ABOUT | ARCHIVE | SUBMIT | SUBSCRIBE | ADVERTISE | AUTHOR INFORMATION

Basset: learning the regulatory logic of 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 Primary Sequence 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 Efthymiou¹, Serafim Batzoglou¹, Stehfan 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 determinants 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 > articles

Analysis | Published: 27 July 2015

Predicting the sequence specificity of DNA- and RNA-binding proteins with deep learning

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

FUToronto

DeepBind
2015

CSH PRESS GENOME RESEARCH

HOME | ABOUT | ARCHIVE | SUBMIT | SUBSCRIBE | ADVERTISE | AUTHOR INFORMATION

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 primary sequence with Akita

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

Calico

Akita
2020

nature methods

Explore content About the journal Publish with us

nature > nature methods > articles > articles

Article | Published: 08 August 2022

scBasset: sequence-based model of gene expression using convolutional neural networks

Han Yuan & David R. Kelley

Calico

scBasset
2022

Bioinformatics

Gifford LabMIT

DNA-TF binding prediction with deep learning
2016

Article Navigation

JOURNAL ARTICLE

Convolutional neural networks for predicting protein binding

Haoyang Zeng, Matthew D. Edwards, Ge Liu, David K. Gifford

nature genetics

Explore content

nature >

Article | Published: 16 July 2018

Deep learning sequence-based prediction of variant effects on gene expression

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

Troyanskaya Lab Princeton

ExPecto
2018

ARTICLES

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

Effective gene expression prediction from DNA sequence by integrating multiple data types

Žiga Avsec^{1,2,3,4,5}, Vikram Agarwal^{1,2,4}, Daniel Viswanathan^{1,2,3,4,5}, Agnieszka Grabska-Barwinska¹, Kyle R. Taylor^{1,2,3,4,5} and David R. Kelley^{1,2,3,4,5}

DeepMind + Calico

Enformer
2021

nature genetics

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

Received: 28 August 2023

Accepted: 4 December 2024

Johannes Linder^{1,2,3,4,5}, Divyanshi Srivastava^{1,2,3,4,5}, Han Yuan^{1,2,3,4,5}, Vikram Agarwal^{1,2,3,4,5} & David R. Kelley^{1,2,3,4,5}

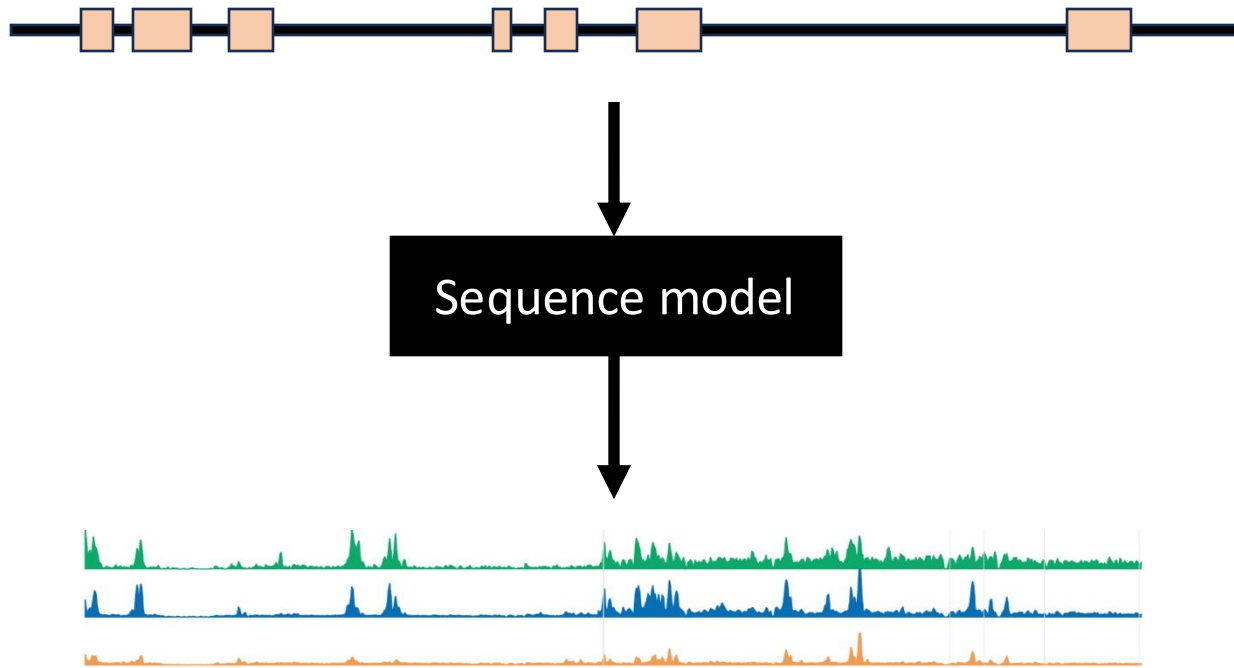
Calico

Borzo
2025

Supervised learning

Input: DNA sequences

Output: Genomics tracks



Protein Language model (PLM)

RESEARCH ARTICLE | BIOLOGICAL SCIENCES | 8

Facebook

Biological structure and function
from scaling up
million protein s

First LM attempt
PNAS 2020

Alexander Rives, Joshua Meier, Tom Serres, and Rob Fergus

Article | Open access | Published: 27 July 2022

Höcker Lab

**ProtGPT2 is a deep unsupervised
for protein design**

Noelia Ferruz, Steffen Schmitt

Nature Communications 13, Article

ProtGPT2
Nat Commun 2022

RESEARCH ARTICLE | STRUCTURE PREDICTION

f X

Meta

**Evolutionary-scale prediction
tein structure with a language**

ZEMING LIN, HALIL AKIN, L.-J. AND ALEXANDER RIVES

ESMFold
Science 2023

Transformer protein language

Facebook

unsup

Transformer protein LM
ICLR 2021

Roshan Rives

Article | Published: 03 October

Harvard + Columbia

**Single-sequence prot
using a language m**

Ratul Chowdhury, Nazim B.

Anant Kharkar, Koushik Roy

Zhang, George M. Church,

AminoBERT
Nat Biotechnol 2022

Primer | Published: 15 February 2024

Profluent

Designing protein

Jeffrey A. Ruffolo & Ali M.

Nature Biotechnology 42, 200–202 (2024)

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 Language**

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,

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 version)]

**ProLLaMA: A Protein Language
Multi-Task Protein Language**

Liuzhenghao Lv, Zongying Lin, Hao Li, Yuyang Liu, Ji

Peking Uni

ProLLaMA
arxiv, 2024

SYNTHESIS · Volume 12, Issue 6, P654-669.E3, June 16, 2021

Open Access

**Learning the protein
structure, and function**

Tristan Bepler, 1,2,3, Bor

MIT

Review
Cell Systems 2021

ARTICLE · Volume 14, Issue 11,

**ProGen2: Exploring the
language models**

Erik Nijkamp, Jeffrey A. Ruffolo

Ali Madani, 1,3,6

Salesforce + Profluent

ProGen2
Cell System 2023

~ 6000 citations

DNA Language model

DNABERT: pre-trained Bidirectional Encodings of Representations from Transformers for DNA-language in genomic context

Yanrong Ji, Zhihan Zhou, Han Liu

SBU

DNABERT
Bioinformatics 2021

DNABERT-2: Efficient Foundation Model Benchmark For Multi-Species

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik

SBU + NU

DNABERT-2
ICLR 2024

Sequence
molecular

Stanford + Arc Inst + TogetherAI

Eric Nguyen, Michael Poli, Matthew G. Durrant, Armin
Jeremy Sullivan, Madelena Y. Ng, Ashley Lewis, Aman P.
Stephen A. Baccus, Tina Hernandez-Boussard, Christo
doi: <https://doi.org/10.1101/2024.02.27.582234>

Evo
bioRxiv 2024

RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY |

UC Berkeley

DNA language models are powerful predictors of genome-wide variation

Gonzalo Benegas, Sanjit Singh Batra, and Yun S. Song

GPN
PNAS 2023

Article | [Open access](#) | Published: 23 July 2024

DNA language models capture genomic context in the human genome

Melissa Sanabria, Jonas Hirsch, Pierre

TUD

GROVER
Nat Mach Intell 2024

[Submitted on 5 Mar 2024]

Cornell + Princeton + CMU

Caduceus: Bi-Directional Range DNA Sequence Modeling

Yair Schiff, Chia-Hsiang Kao, Aaron Gokaslan, Tri

Caduceus
ICML 2024

The Nucleotide Transformer Foundation Models for Genomic Language

InstaDeep + Nvidia + TUM

Hugo Dalla
Adam Henr
Bernardo P
Marie Lope

Nucleotide Transformer
bioRxiv 2023

Article | [Open access](#) | Published: 03 April 2024

Genomic language model predicts regulation and function

Yunha Hwang, Andre L. Cornman, Elizabeth
R. Girguis

Harvard + MIT

Genomic LM
Nat Commun 2024

Cross-species
nucleotide r

Cornell + USDA-ARS + Simons

Jingjing Zhai, Aaron Gokaslan, Yair
Michelle C. Stitzer, M. Cinta Rom

doi: <https://doi.org/10.1101/2024.06>

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, Yosh
Baccus, Chris Ré

Stanford

HyenaDNA
NeurIPS 2023

Research | [Open access](#) | Published: 02 April 2024

Species-aware DNA language models capture regulatory elements across species

Alexander Karol
Julien Gagneur

TUM

Species-aware DNA LM
Genom Biol 2024

Genome Biology, 25, Article number: 83 (2024) | [Cite this article](#)

Nucleotide dependency analysis of DNA language models

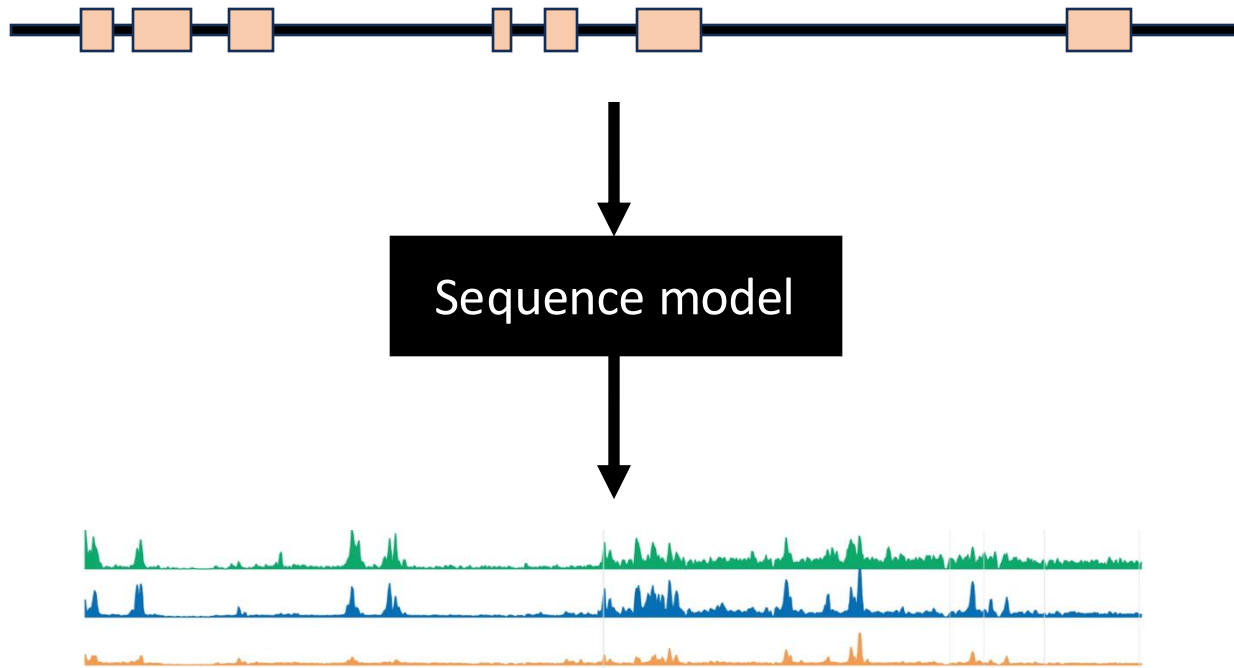
TUM

Nucleotide dependency
bioRxiv 2024

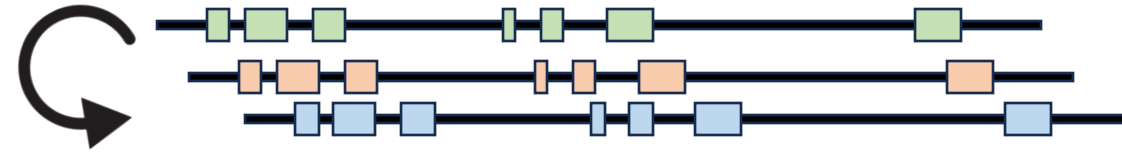
Supervised learning

Input: DNA sequences

Output: Genomics tracks



Stage 1 Self-supervised learning



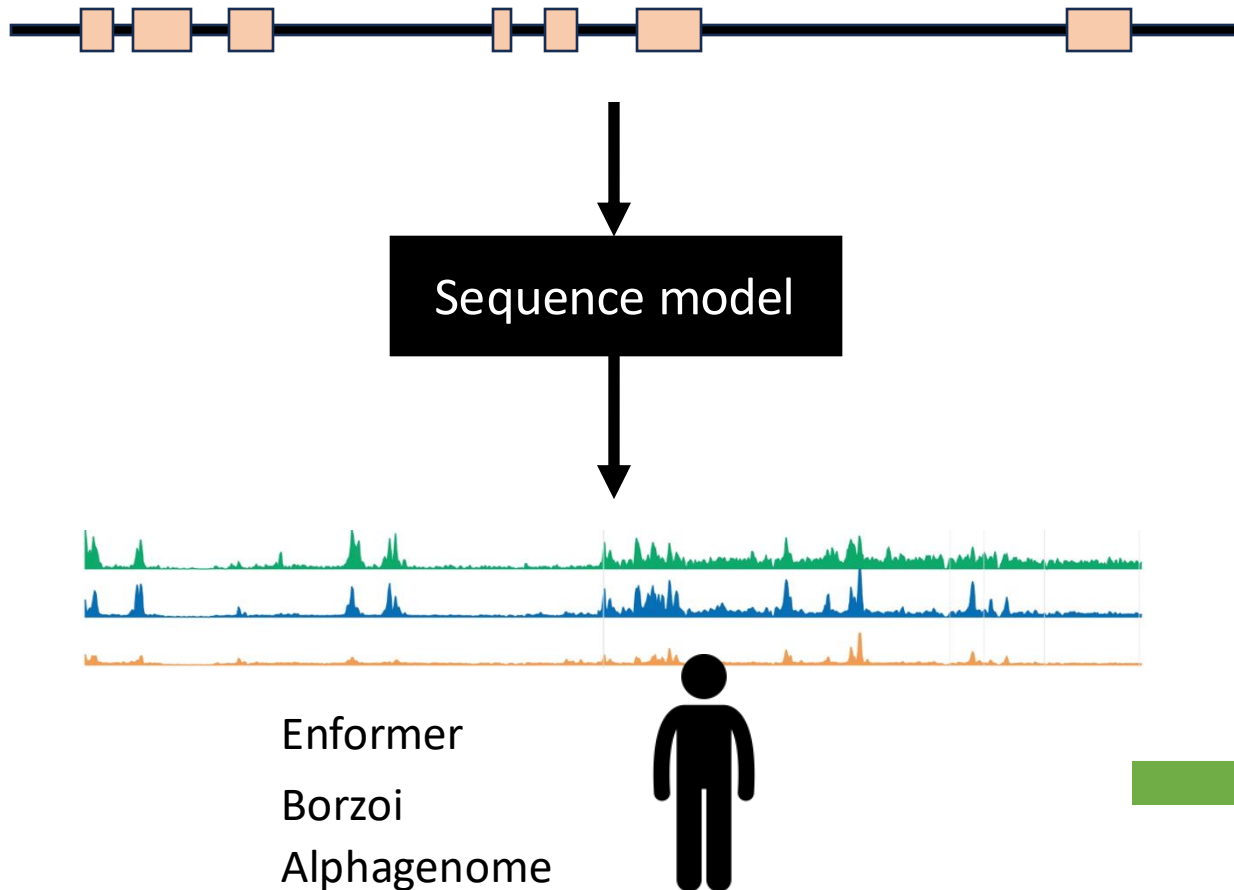
Language model (LM) /
Foundation model

Stage 2 Fine-tuning LM

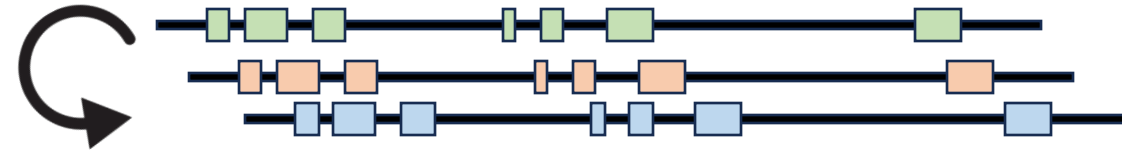


Enformer
Borzoï

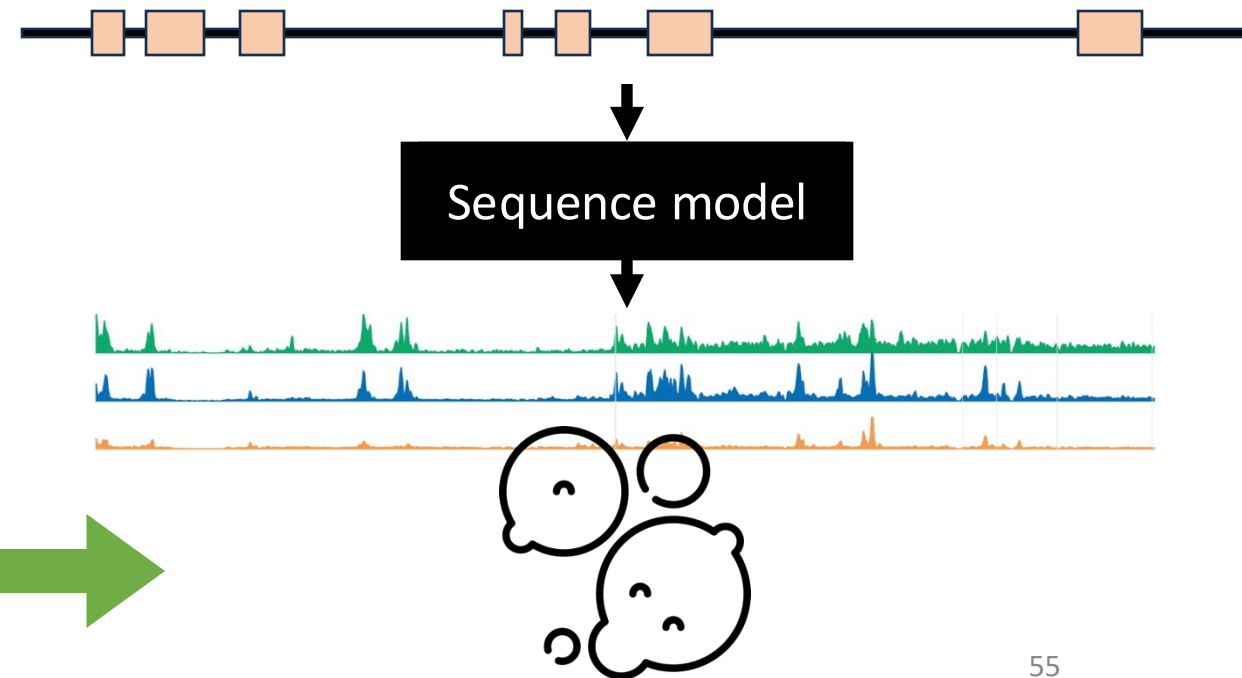
Supervised learning



Stage 1 Self-supervised learning



Stage 2 Fine-tuning LM



Transfer learning

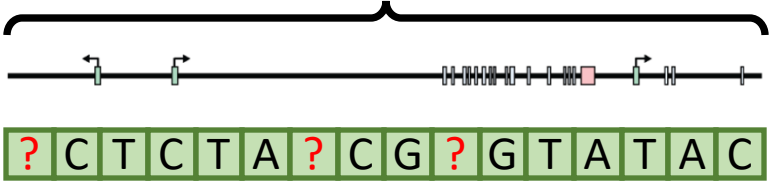
Interpretability

Applications



Shorkie

16384bp



Encoding: (4 + 1 + species_num)

Reverse complementary

16384 * 4 Masked language modeling loss

A

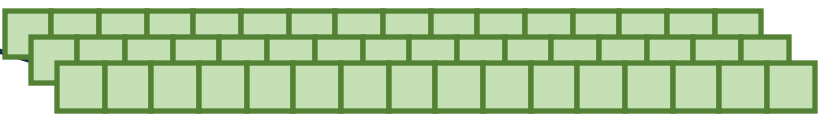
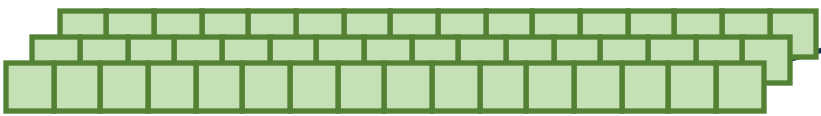
C

G

T

.8				.0		.0													
.1				.0		.7													
.1				.9		.1													
.0				.1		.2													

1bp res



1bp res

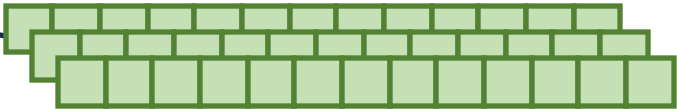
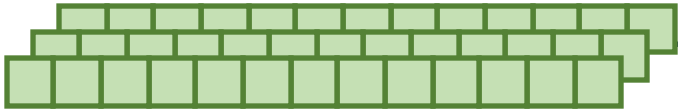
...

...

...

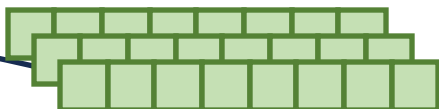
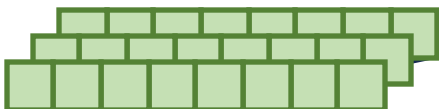
...

16bp res



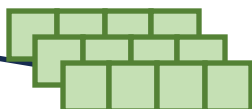
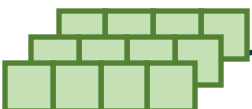
16bp res

32bp res



32bp res

64bp res



64bp res

128bp res



128bp res

Transformer
Blocks (8x)



Borzoi

Linder, J. et al. (2025). Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. Nature Genetics, 1-13.

RAP1



FHL1



LM predicts motifs in RP gene Promoter regions



Introduction

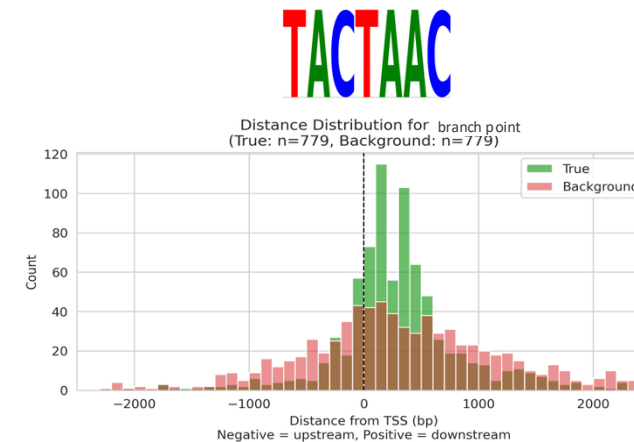
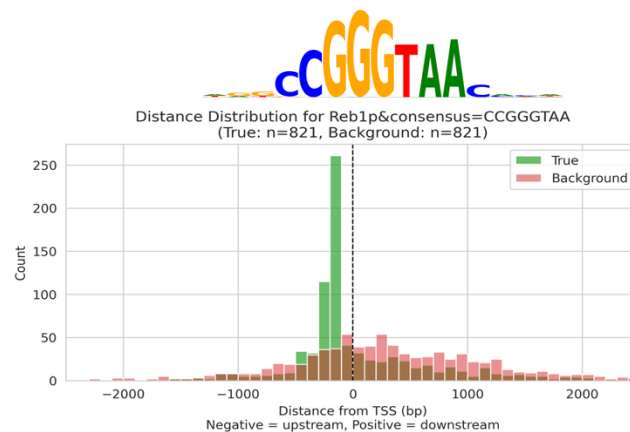
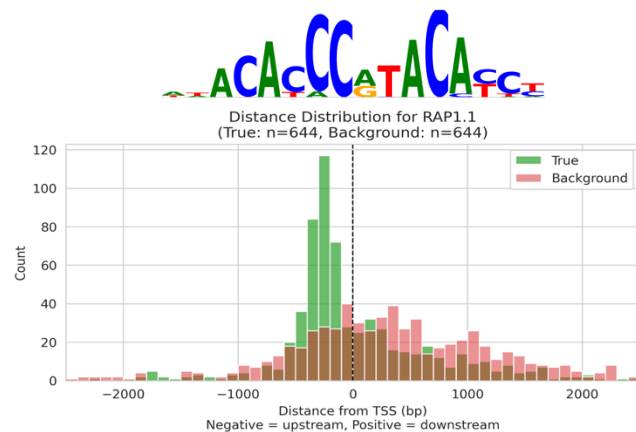
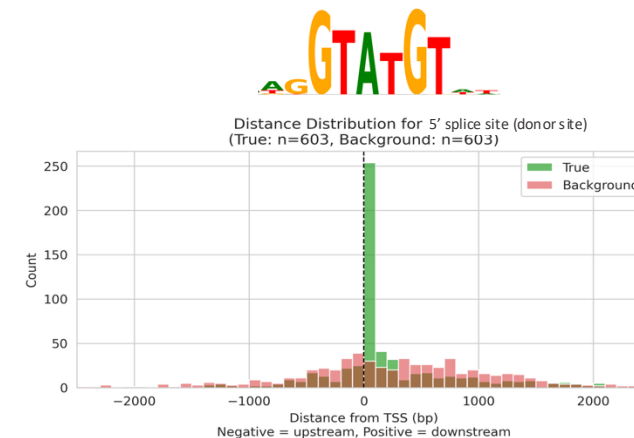
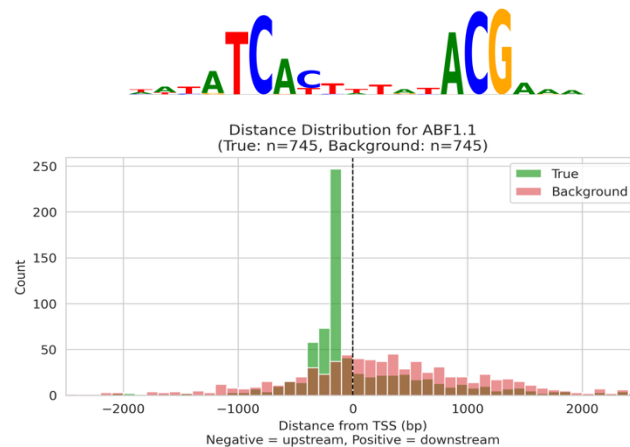
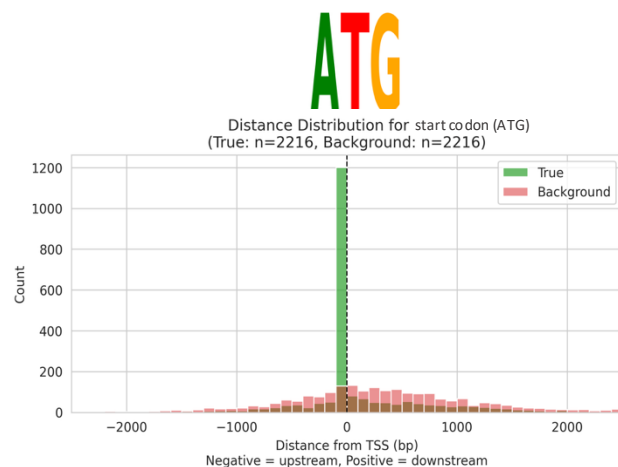
Fungal LM

Transfer learning

Interpretability

Applications

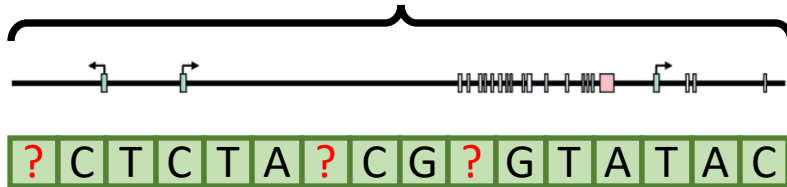
These motifs are enriched in promoter regions





Shorkie

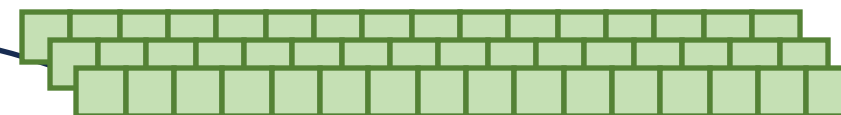
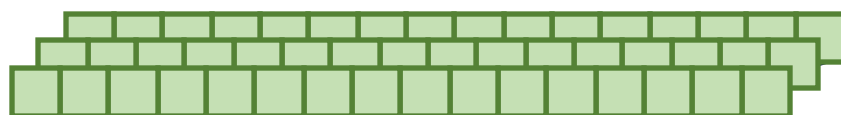
16384bp



16384 * 4

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

1bp res



1bp res

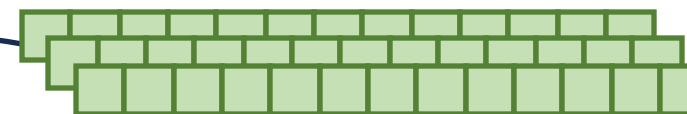
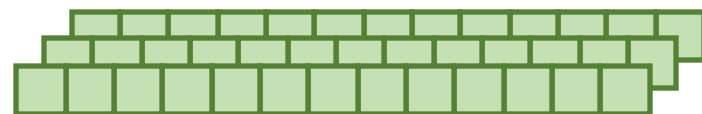
...

...

...

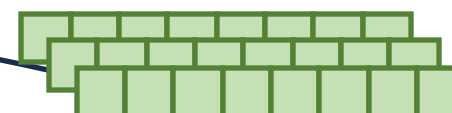
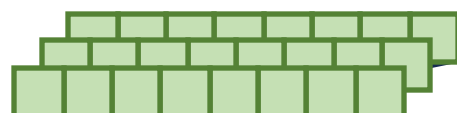
...

16bp res



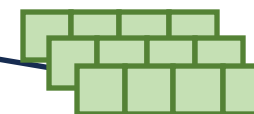
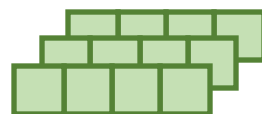
16bp res

32bp res



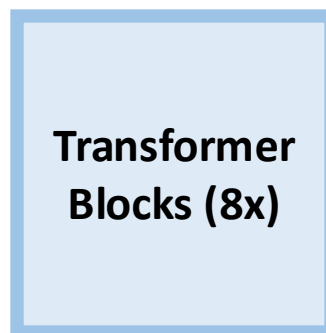
32bp res

64bp res



64bp res

128bp res

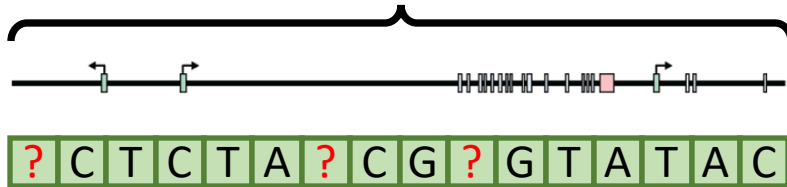


128bp res

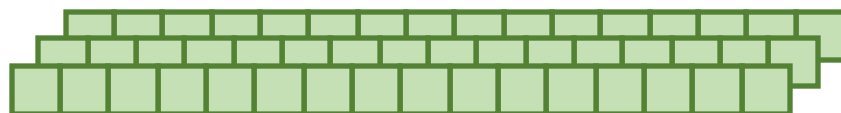


Shorkie

16384bp



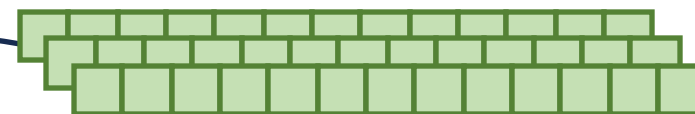
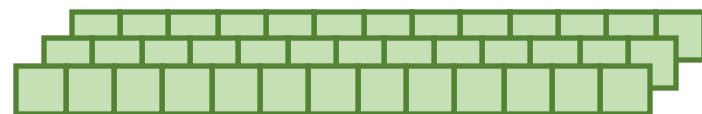
1bp res



...

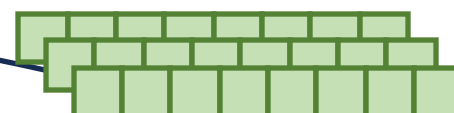
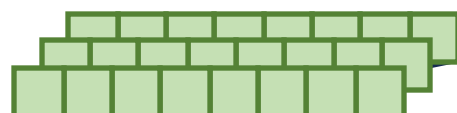
...

16bp res



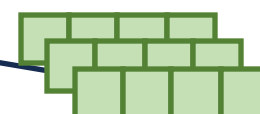
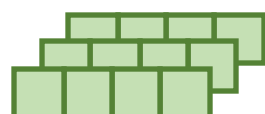
16bp res

32bp res



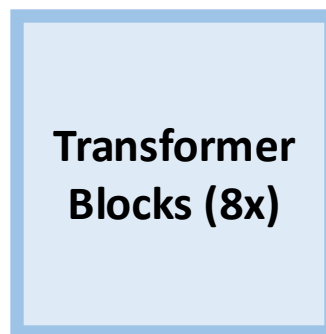
32bp res

64bp res

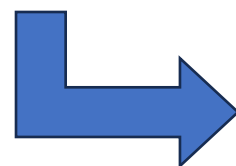


64bp res

128bp res



128bp res

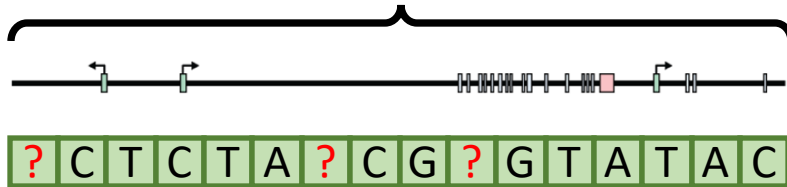


Transformer
Blocks (8x)

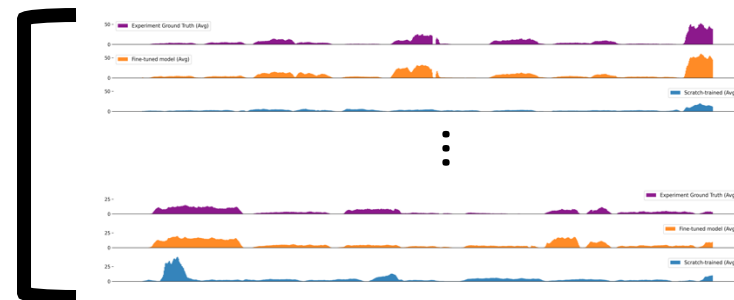


Shorkie

16384bp



Coverage Tacks (896 * 2488)



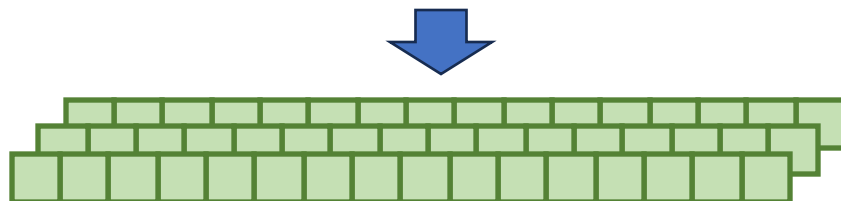
ChIP-exo (1128)

Histone marks (20)

RNA-Seq (3054)

1000-strains
RNA-Seq (1014)

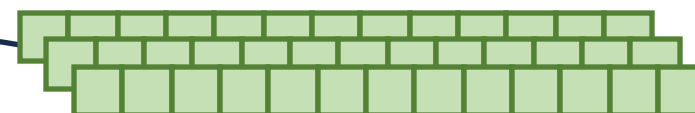
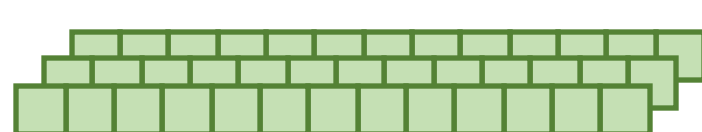
1bp res



...

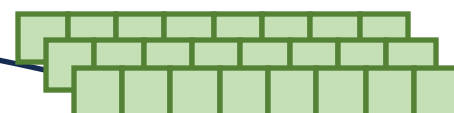
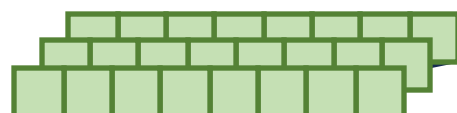
...

16bp res



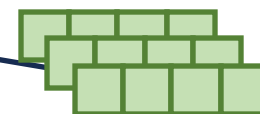
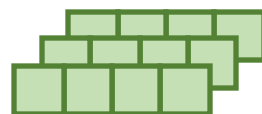
16bp res

32bp res



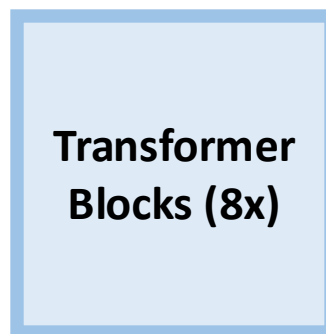
32bp res

64bp res



64bp res

128bp res

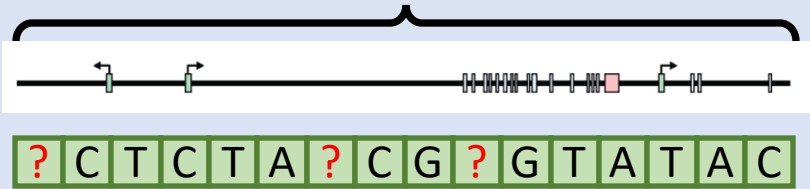


128bp res

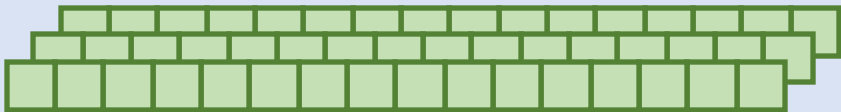


Shorkie

16384bp



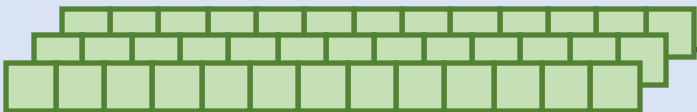
1bp res



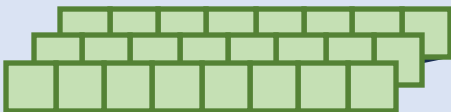
...

...

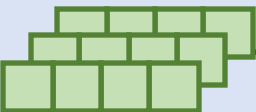
16bp res



32bp res



64bp res



Introduction

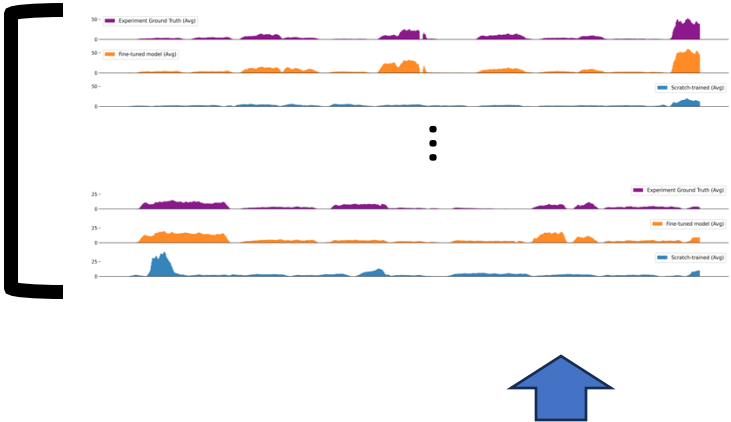
Fungal LM

Transfer Learning

Interpretability

Applications

Coverage Tacks (896 * 2488)



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

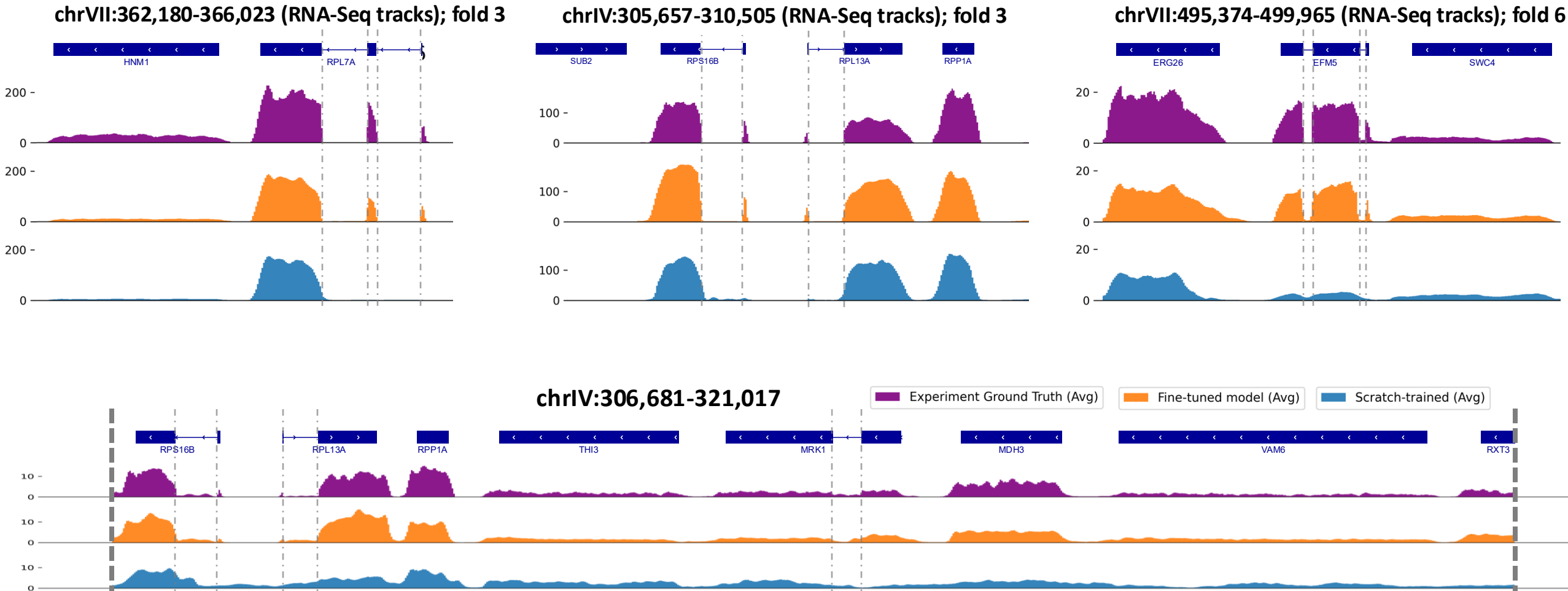
Experiment Ground Truth (Avg)

Fine-tuned model (Avg)

Scratch-trained (Avg)

Shorkie vs Random Initialization (Test set)

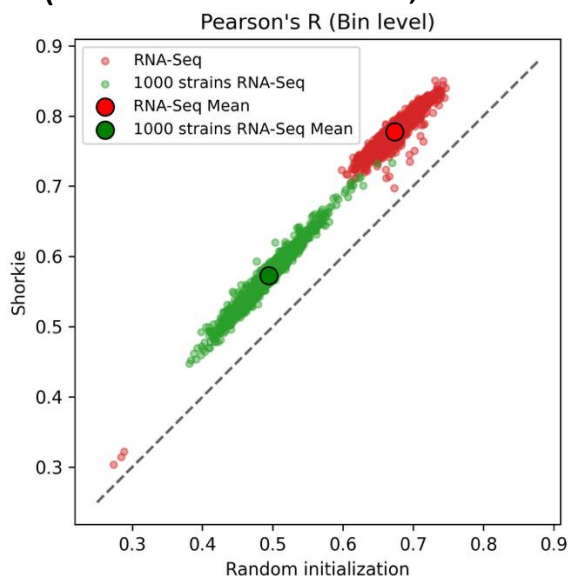
RNA-Seq tracks



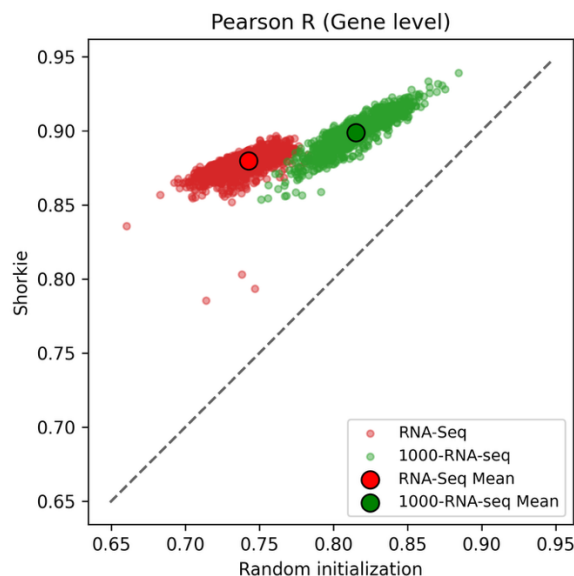
Shorkie vs Random Initialization (Test set)

Bin level

(each dot is a track; all bins)

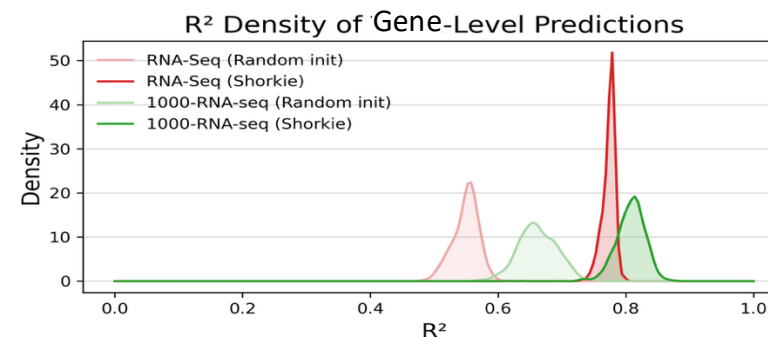
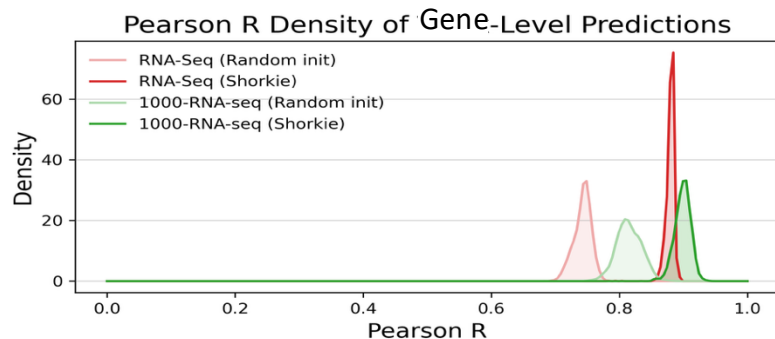
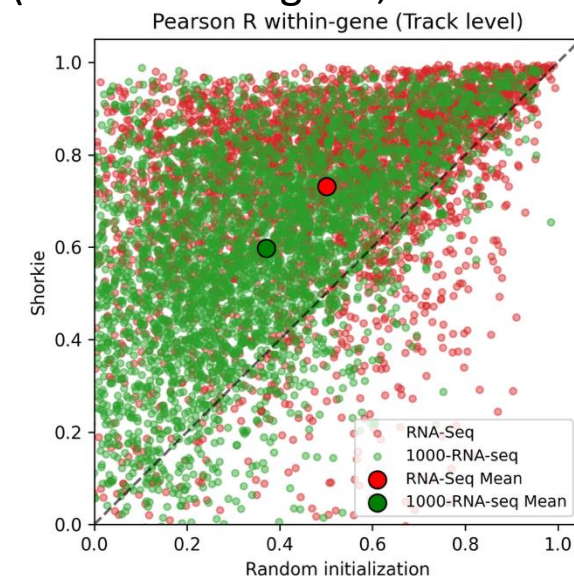


Gene level (each dot is a track; all genes)

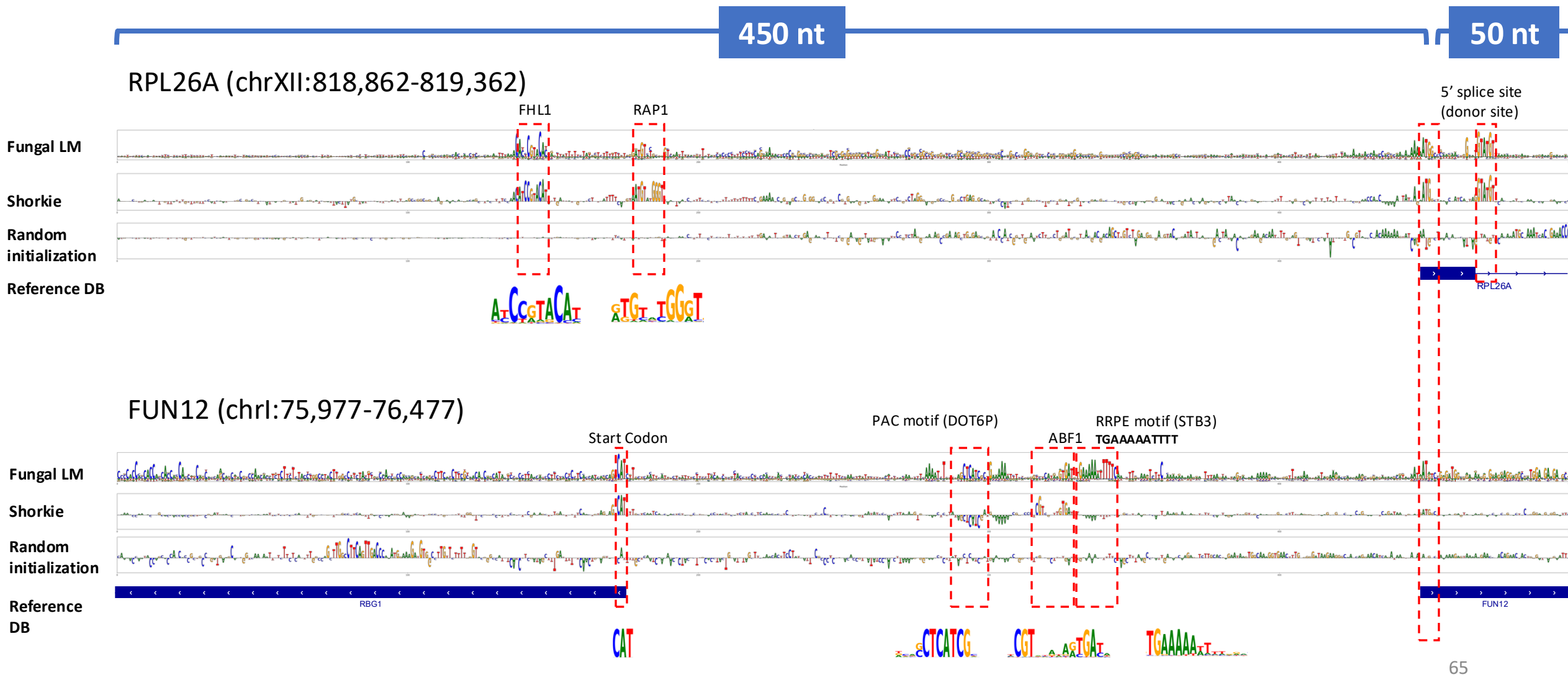


Track level

(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

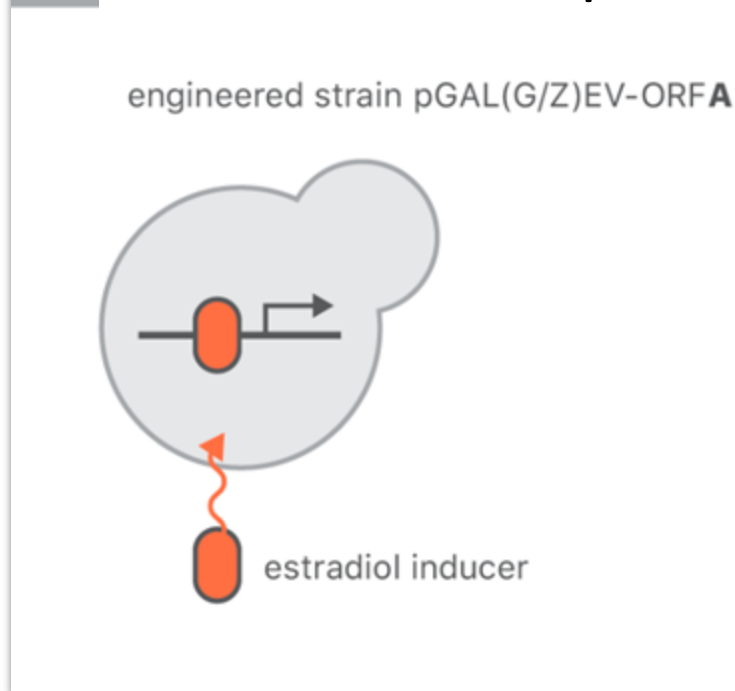
ACATAC

TTACCCG

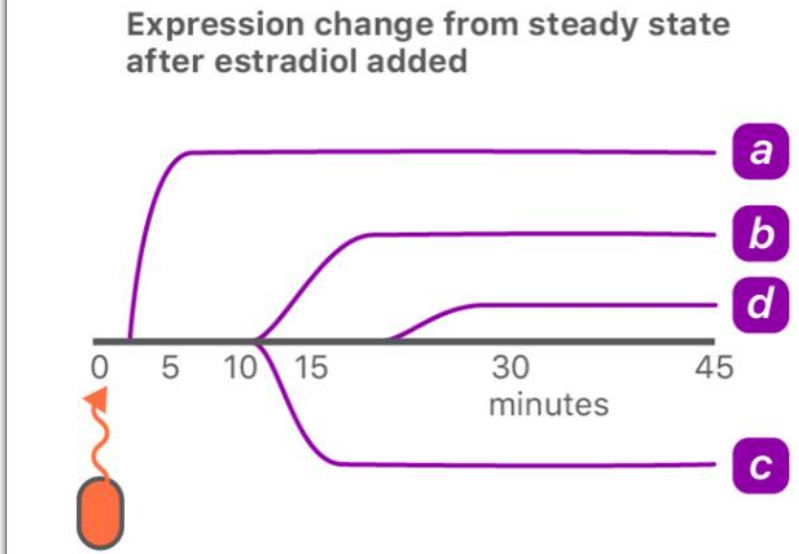
Time-course TF-induced RNA-Seq

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

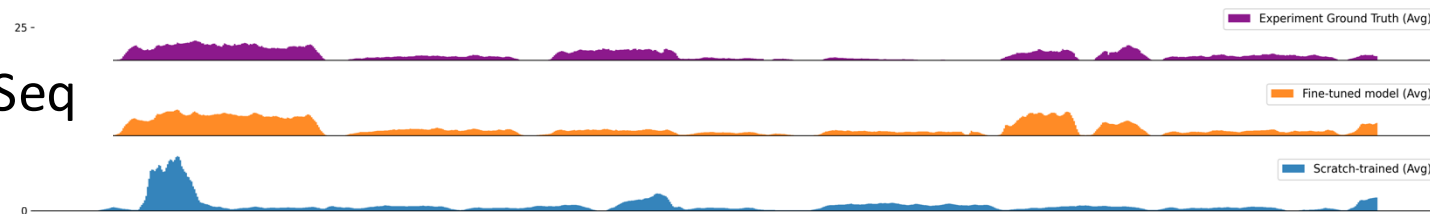
A TF induction experiments



B RNA-seq at different T

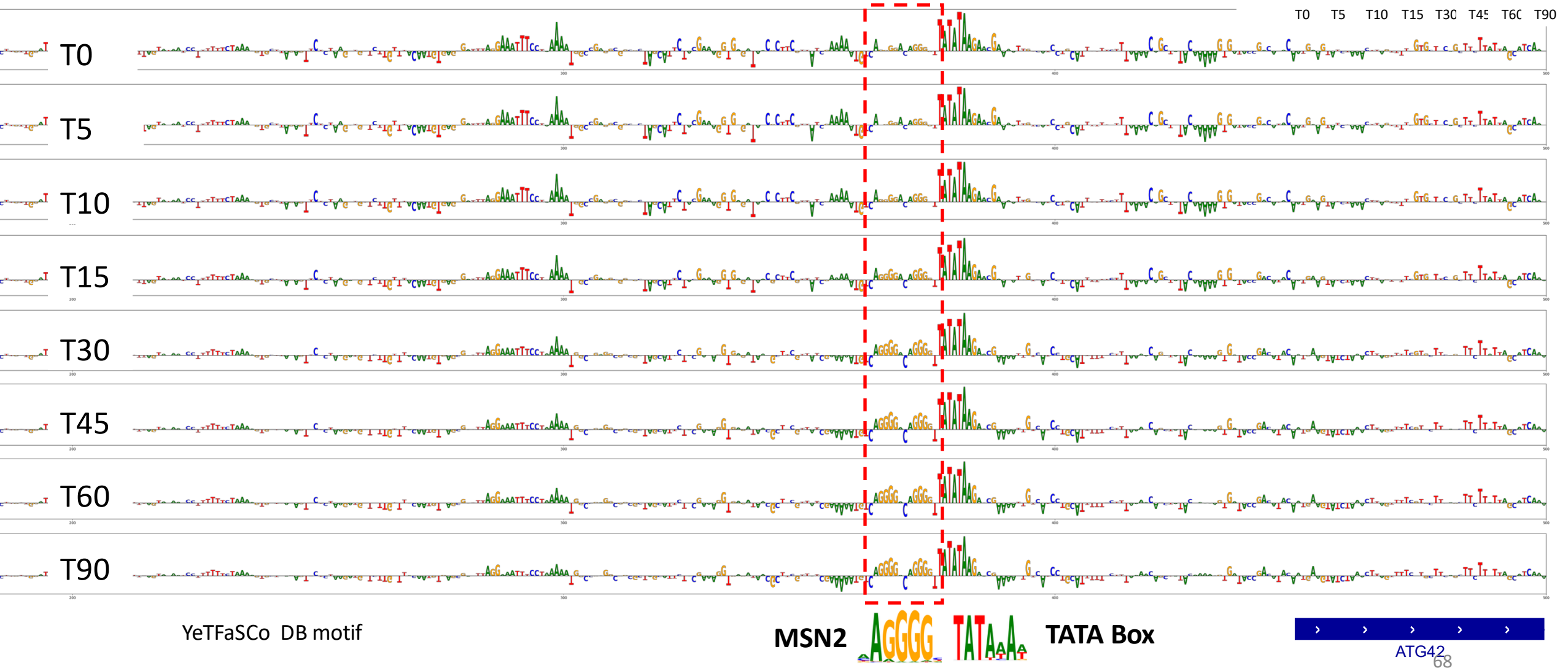
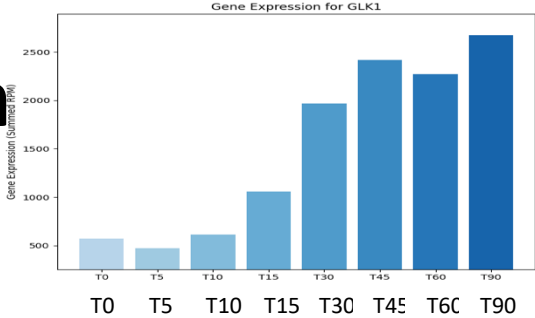


3054 RNA-Seq readouts



Shorkie MSN2 Induction temporal RNA-Seq prediction

chrIII:50388-50888 (Promoter region of GLK1)



YeTFaSCo DB motif

MSN2

TATA Box

ATG42₆₈

Introduction

Fungal LM

Transfer Learning

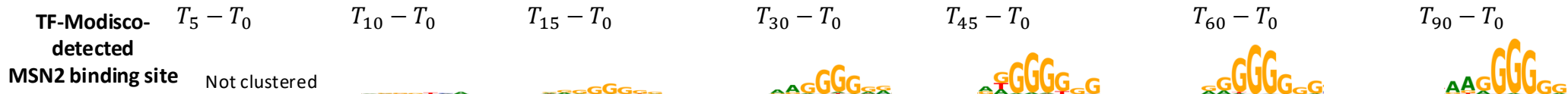
Interpretability

Applications

Average motif changes using TF-modisco

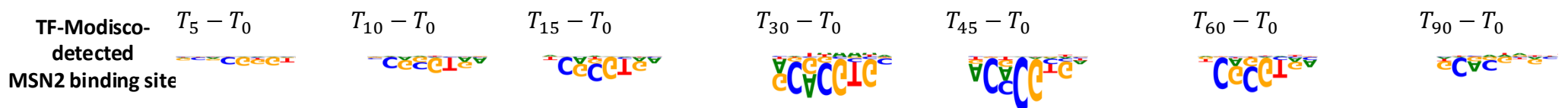
MSN2 Induction

YeTFaSCo
DB motif



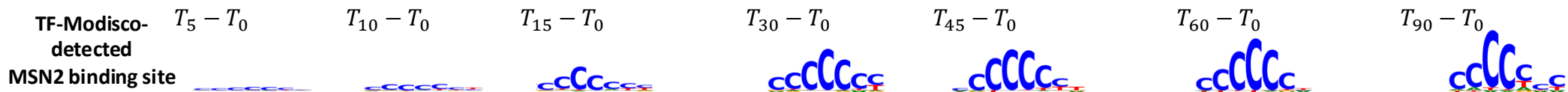
MET4 Induction

YeTFaSCo
DB motif



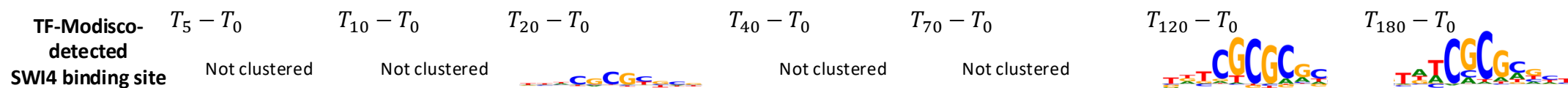
MSN4 Induction

YeTFaSCo
DB motif



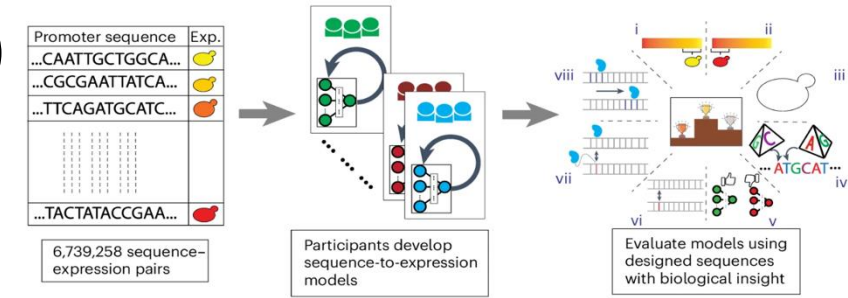
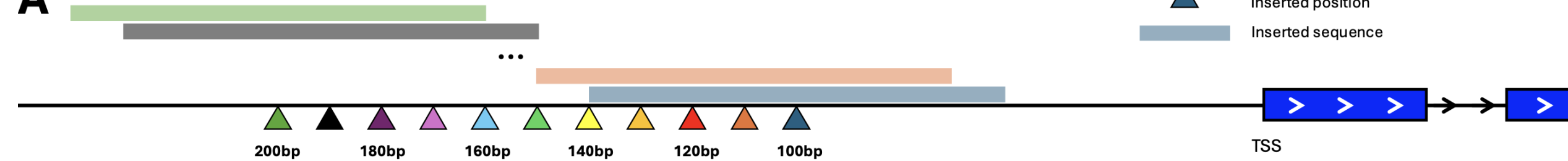
SWI4 Induction

YeTFaSCo
DB motif

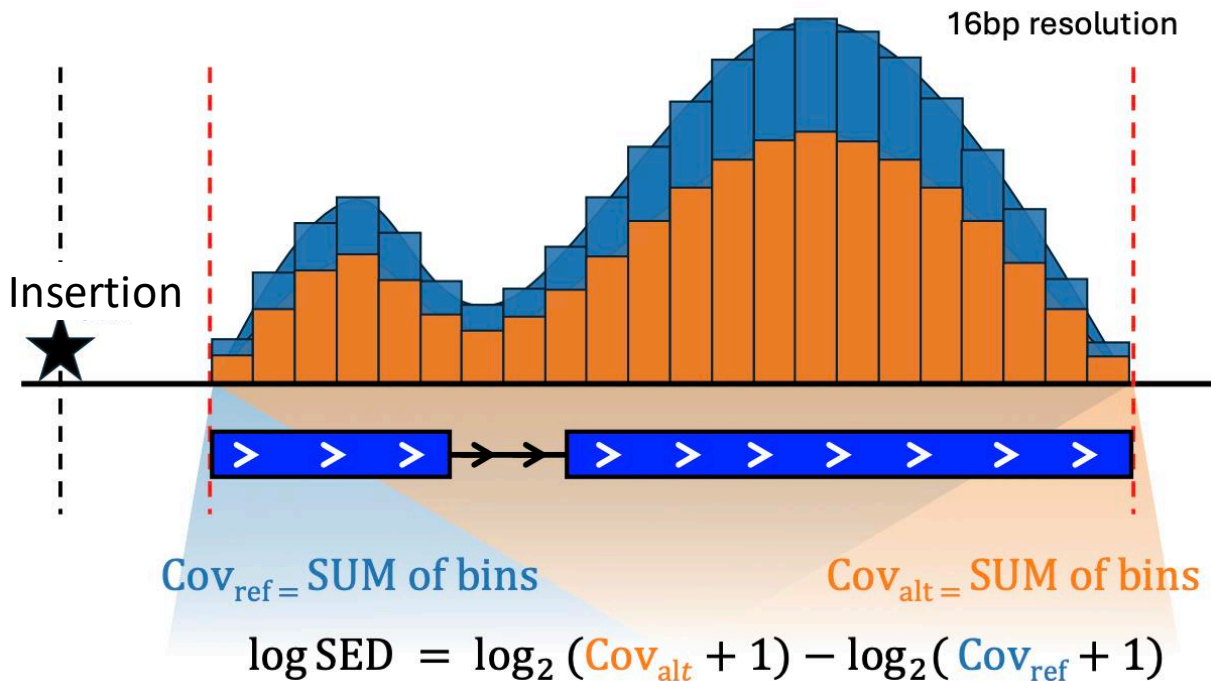


MPRA mutation effect prediction

A



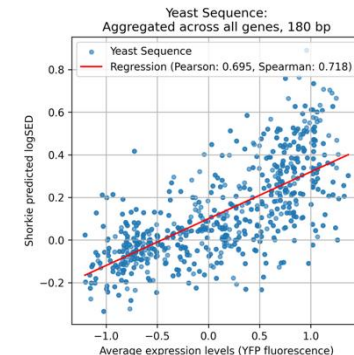
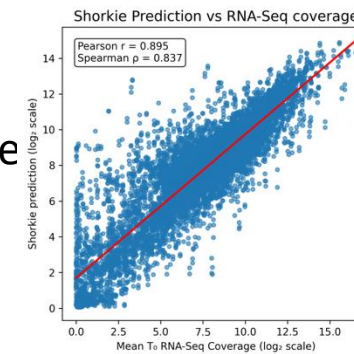
Inserted position
Inserted sequence



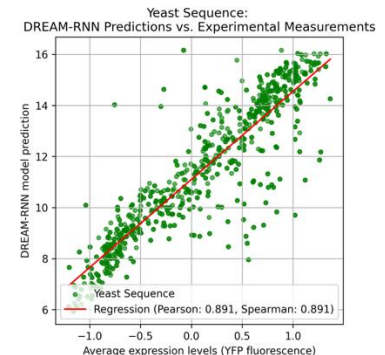
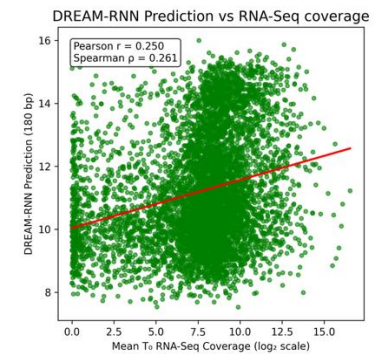
RNA-Seq Coverage Prediction

Rafi et al. Yeast Sequences (single-sequence)

Shorkie



DREAM-RNN



Introduction

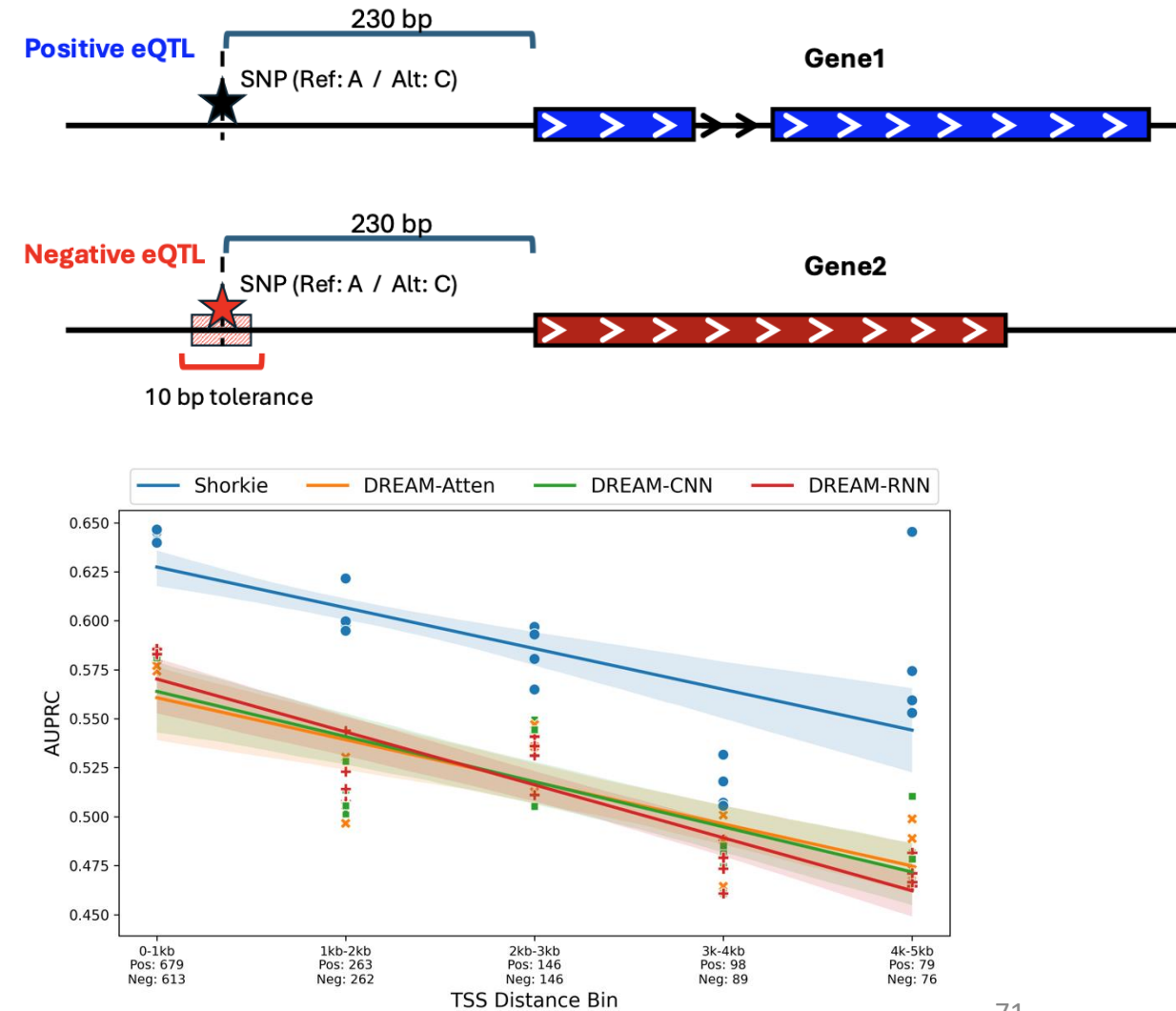
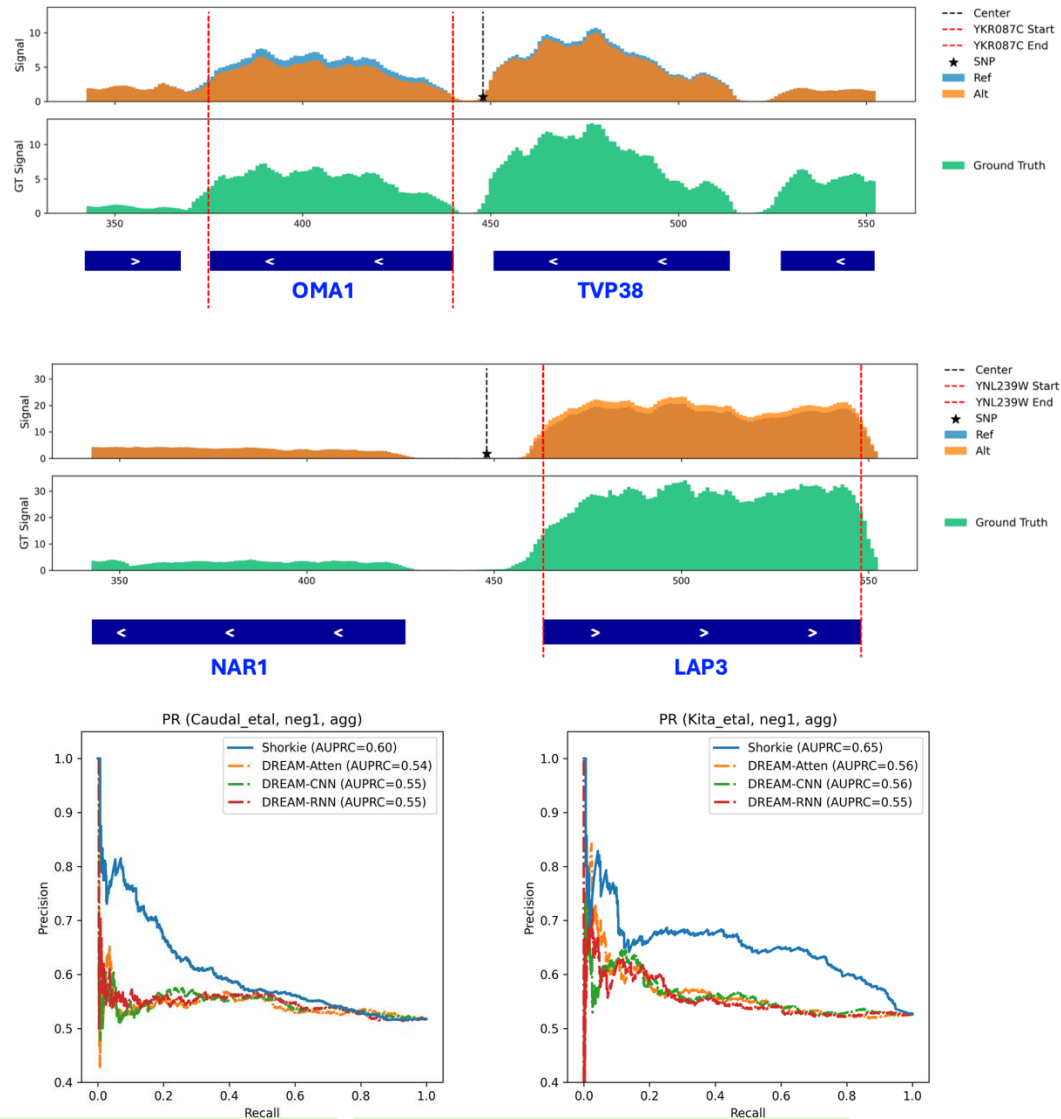
Fungal LM

Transfer Learning

Interpretability

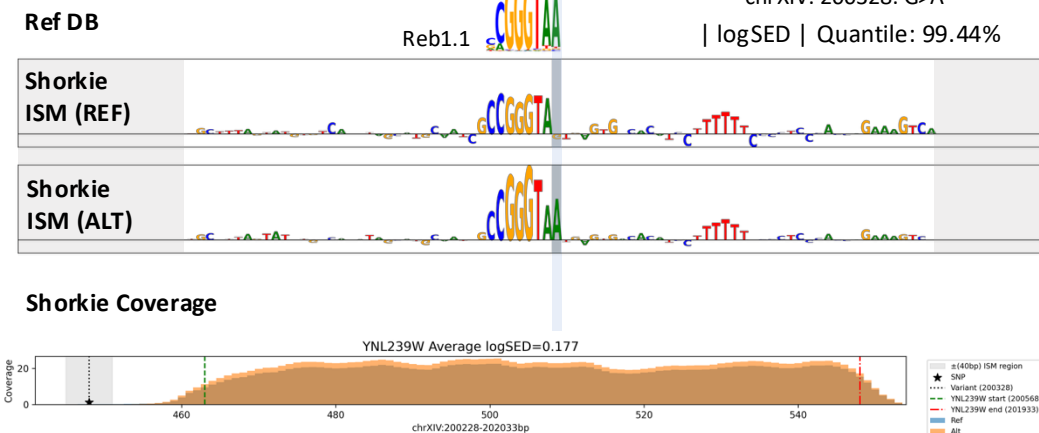
Applications

Variant effects on eQTLs and negatively selected eQTLs

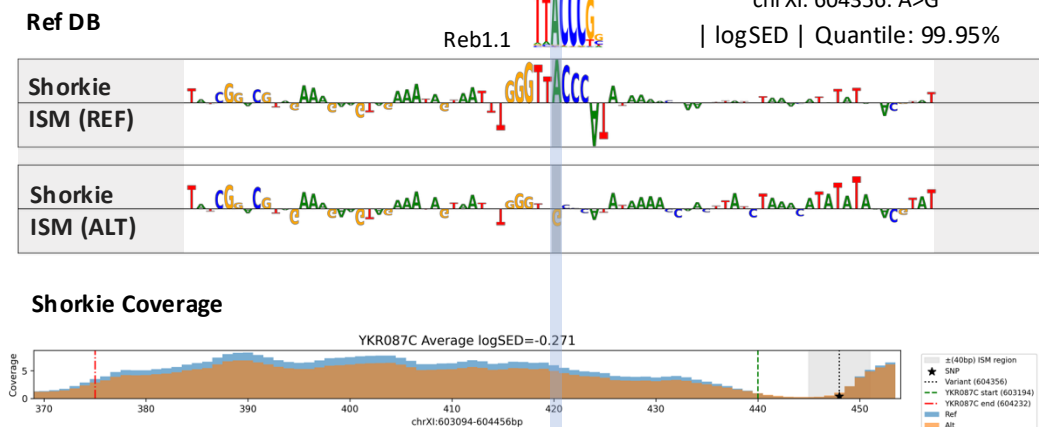


Variant effects on eQTLs and negatively selected eQTLs

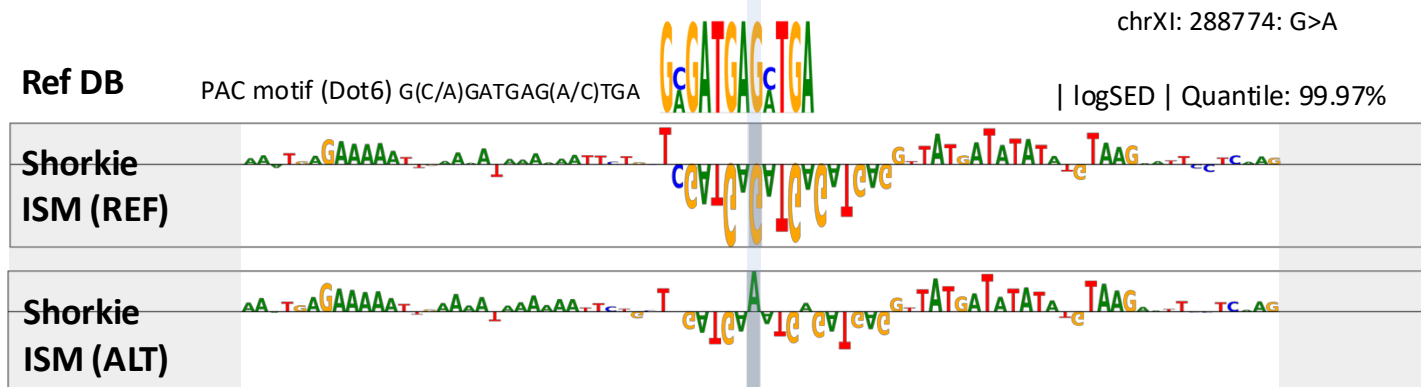
chrXIV:200,288 – 200,368



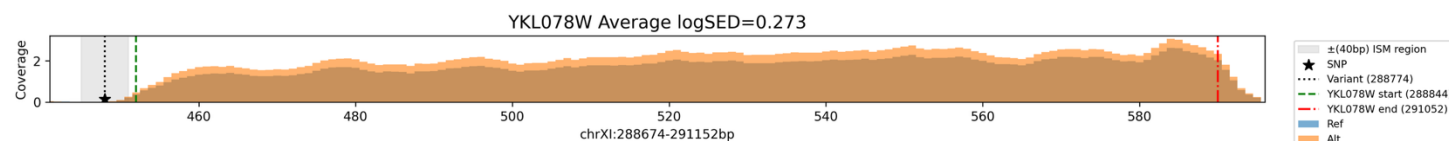
chrXI:604,316-604,396



chrXI:288,734 – 288,814



Shorkie Coverage



Introduction

Fungal LM

Transfer Learning

Interpretability

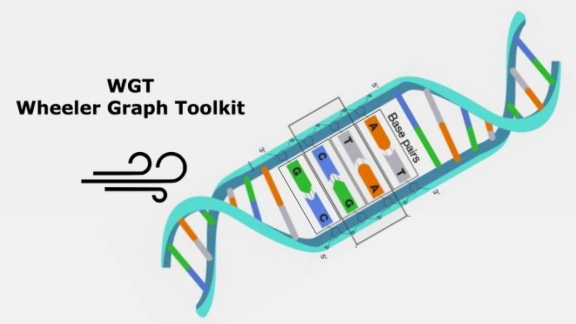
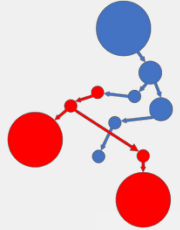
Applications

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

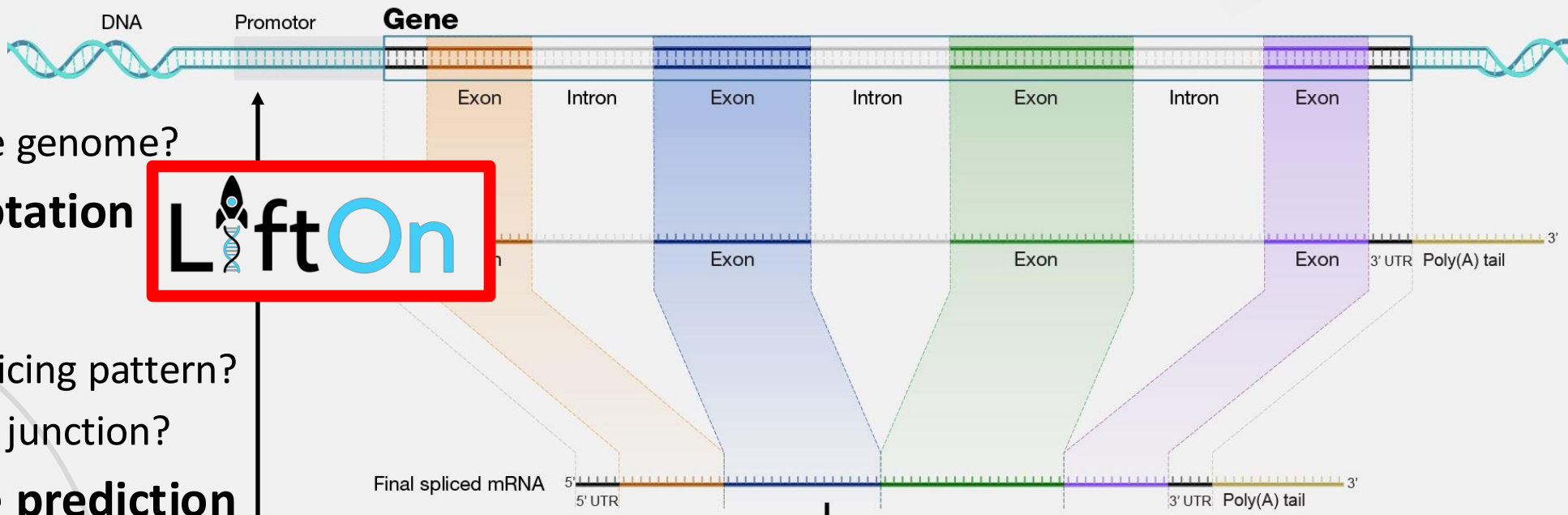
WGT



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

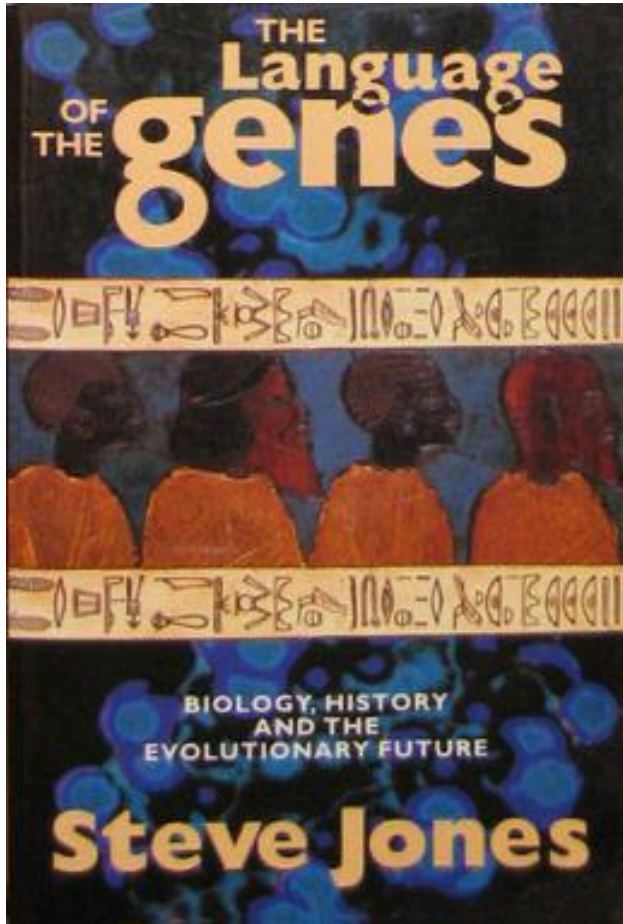
Can we predict gene expression by learning the regulatory grammar in the genome?

Part VI: Shorkie. RNA-Seq coverage prediction

Shorkie



Conclusions: Language of genomes



Steve Jones, 1993

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

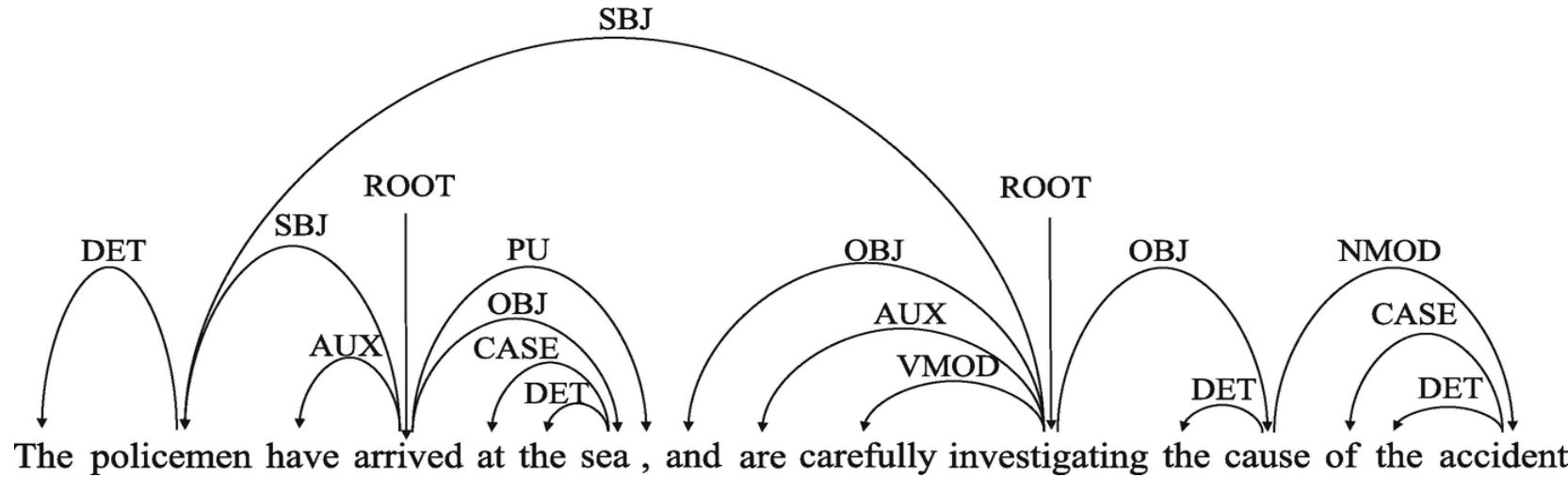
Zhou et al., 2023; Hwang et al., 2024; Sanabria et al., 2024; Nguyen et al., 2024; Theodoris, 2024; Dalla-Torre et al., 2025

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

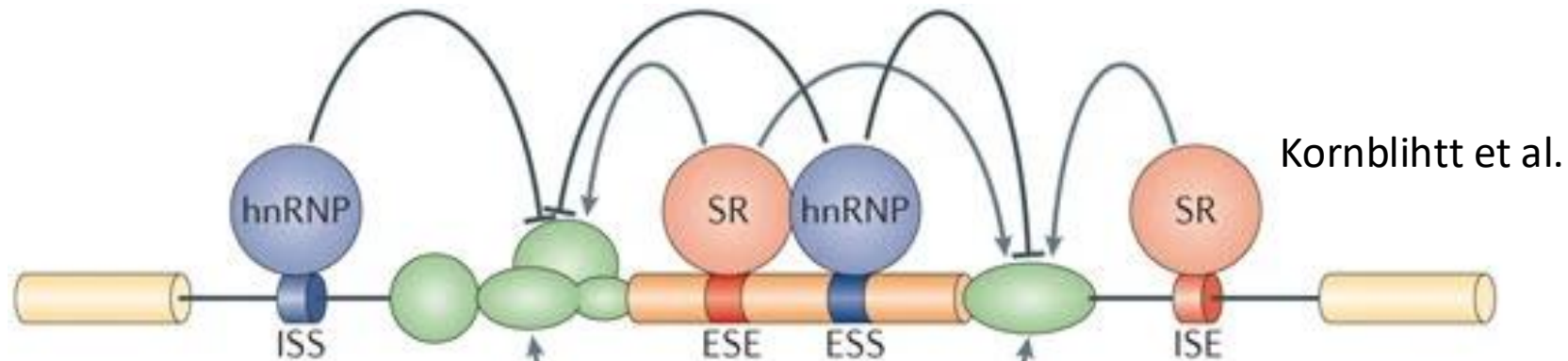
```
caccatgtttgccaggctggctctgaactcccgacccaggtgatccaggt  
ccctgtttgctcagtgaaatgaggggctgacctggcattccagcctgaggt  
ccttgactcctcaGTCAC TAGAGGATCTTGAAGACCAAAACGAAACGAA  
CGCCTGCTGCCCGCAGGCGTGTCTGTGCGTGTGGCGTGTGCAACGAACT  
TACATACCGCAAACCGTGTGTGAACCTGTCAACTCTCTGTCTTTCGAACT  
ACTTGAAGGACTGCTGTGTATTGTAAATAACAAACTATTGTCACTTGT  
TAACCACAAAACGCCATCGTCGTGAGGTAAGCTGTCTCTTACAAACTT  
ttatttgacaaatgctgaaaacttattctatctaatattacataaatatg  
ttctgttggttctctctctctctctctctctctctctctctctctctct  
tgagagaggttgagttcagggaagacagcctgccacacatccaggttarg  
gagggccacagtgaggacacagagtcacagacctgagccctccacccct  
gccccaaagaggttcaccagcaaggaactgtgtgtattttaaagcaggt  
gacatttcagctcttaaaaagaagcaggagatttcaaatgctagatatt  
aaccgagcctaggggtgacggggagcgaccaagctggcatcttctctct  
aattaggagataattctaagagttactaaaggatgatttattaaagaa  
gttatttgattgctctctatttggtccgggtgggaggtctgcccctct  
gcccctgactgtatttgagggtaccctgtggtccaaacgacccct  
cgctcccacgatccaccctccctggaggaccgagaggttcaggagat  
cccaaagcctgacatagtcagttaccctcaatctctatctctccactga  
gggaaggccaggaaagccctgggtgacaaggaacttagggcactctac  
caagtgtggccacacgcatgccctactccctctccctggccctgag  
cgtctctcatcggggacatatggcaaaaacttggccagagaggaatg  
jagaatggcatgaacccgggagttggagcttgagtgagccagatcgc  
jcatcaggggcagctgcacaggttgggactgcattactccagaggtg  
taacattgattaggtattaccacaggccaggtccatgctatgcaact  
taccctccacgtctgagcagtgagggtgcccgaagcagctgggct  
tcaccttagagaggaatcagcgggggaatcagagcagaaacgggtg
```

Conclusions: Language of genomes

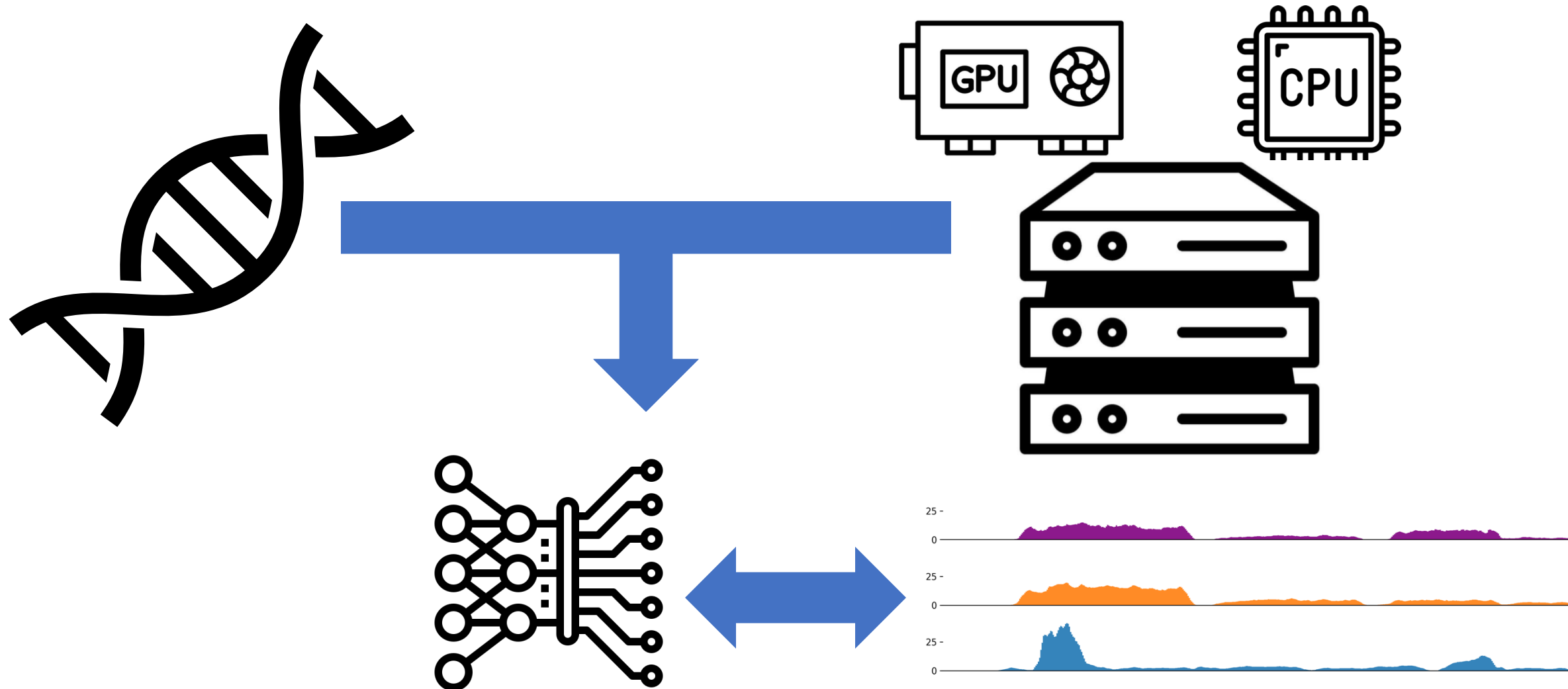
Natural Language Processing



DNA gene regulation



Conclusions: Language of genomes



Committee members



Steven Salzberg



Mihaela Pertea



Ben Langmead



David Kelley



Anqi Liu

Best labmates



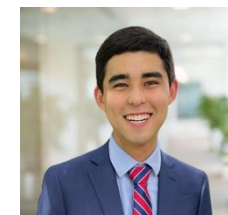
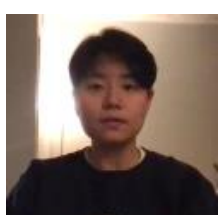
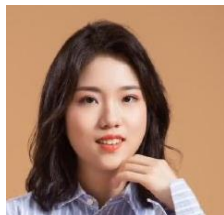
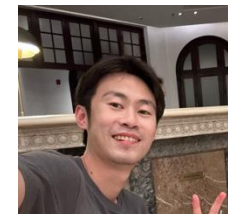
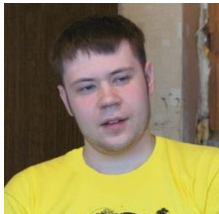
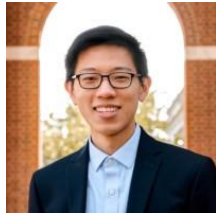
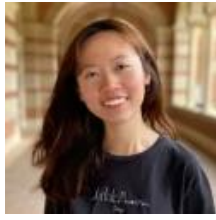
Salzberg Lab

Dr. Steven Salzberg



Perte Lab

Dr. Mihaela Perte



Hopkins Genomics + CS Friends





Calico Friends (2024 Summer)



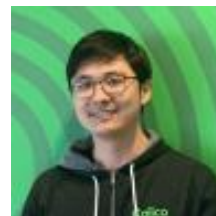
Dr. David Kelley



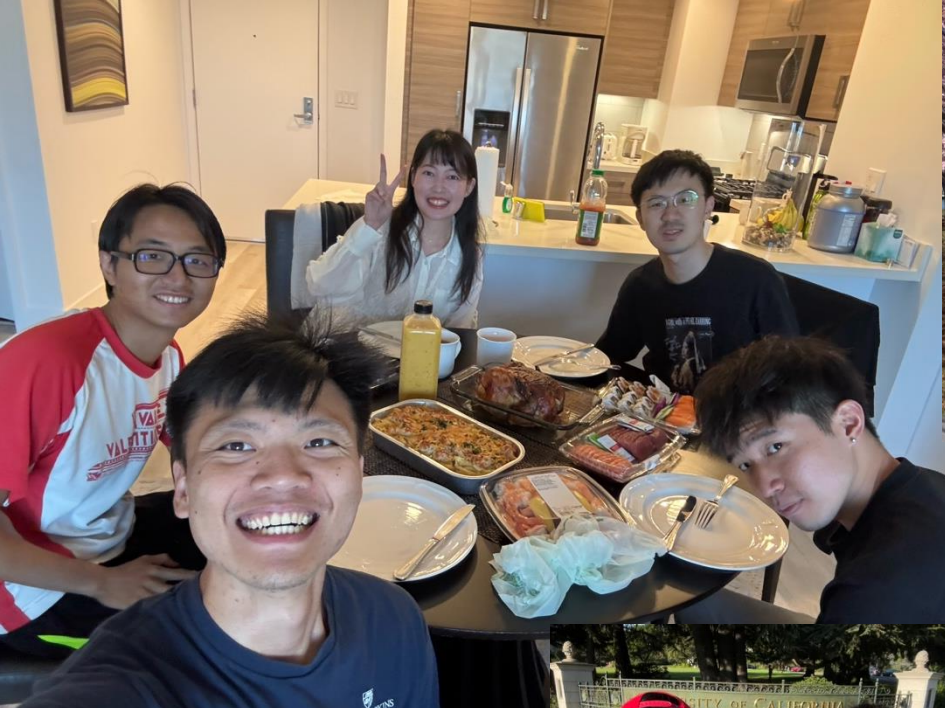
Dr. Johannes Linder



Majed Magzoub







Quarter-finals (Spring 2025)



JHU X UMD Sport Day



AC Malone (Sum 2023)



AC Malone Family



Playoff (Spring 2024)



TSA basketball family



Softball (Sum 2023)



10/22	Kuan-Hao Chao	Predicting RNA-
11/05	Mahler Revsine	HyenaDNA: Lon
11/19	Cristina Martin Linares	Machine-guided
12/03	Eduarda Vaz	Effective gene e
1/7	Gus Fridell	Applying
01/21	Stephen Hwang	Evoluti
02/11	Celine Hoh	A found
02/18	Ashton Omdahl	Small-c
03/04	Zoe Rudnick	Sequen
03/25	Bowen Chen	Fast an
04/08	Mahler Revsine	Person
04/22	Kuan-Hao Chao	Bench
05/06	Cristina Martin Linares	Design
05/27	Bohan Ni	Towar
06/03	Mao-Jan Lin	A scal
06/17	Gus Fridell	Diseas
07/01	Celine Hoh	EpiGeP
07/22	Eduarda Vaz	The lar
07/29	Jake Galvin	Param
08/12	Zoe Rudnick	Predict
08/26	Bowen Chen	Accura
		AlphaC

JHU Deep Learning in Genomics Study Group

Mahler Revsine



Model Ablation and Training Data Experiments

- EpiGePT vs EpiGePT-seq (no TF module)
 - Inclusion of TF module increases cross-cell-type DNase prediction 0.787 PCC → 0.74 PCC
- TF-only model without sequence module
 - Average 0.084 decrease in PCC for epigenomic signals
 - H3K4me1: PCC decreased from 0.408 to 0.329
 - Overall decrease: 0.074 across eight signals
- Eight separate models vs joint multi-task prediction
 - TF module with expression profiles as zeros, mean, median imputation
 - Larger number of missing TFs → greater performance decrease
 - Consistent across all imputation strategies
 - Varied number of cellular contexts
 - Performance increases with more training contexts
 - Even with a single training context, the incorporation of the TF module yields a 1.2% average improvement in PCC compared to the best sequence-based method

Models cannot predict individual variation

- Right: correlation between model outputs and gene expression **across individuals**
 - No correlation
- PrediXcan, a supervised linear model, is a lower bound for theoretical performance
 - Lots of signal is not being captured by deep learning models

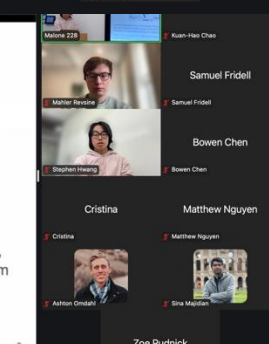
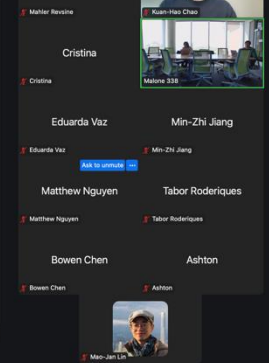
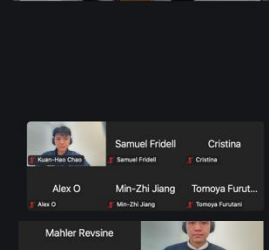
Atomic-resolution structure emerges in language models trained on protein sequences

Idea: Proteins on the scale of evolution capture biological structure and function. Evolution of a protein (mutations) is constrained by structural needs.

$$\mathcal{L}_{\text{MLM}} = - \sum_{i \in M} \log p(x_i | x_{\setminus M})$$

A randomly generated mask M that includes 15% of positions i in the sequence x , the model is tasked with predicting the identity of the amino acids in the mask from the surrounding context $x_{\setminus M}$, excluding the masked positions.

Trained over sequences in the UniRef protein sequence database: ~65 million unique sequences



Search & analysis

Genome-wide association study (GWAS) (all comparison)

Searching (MMseqs2)

alignment

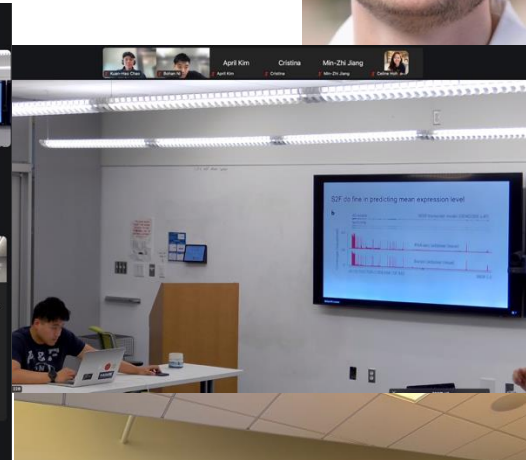
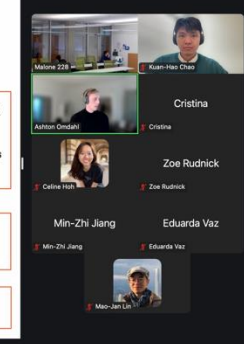
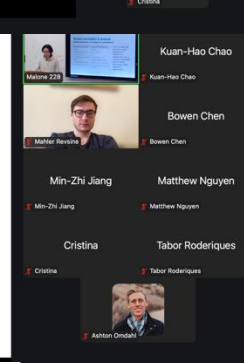
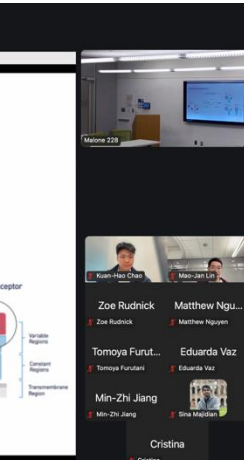
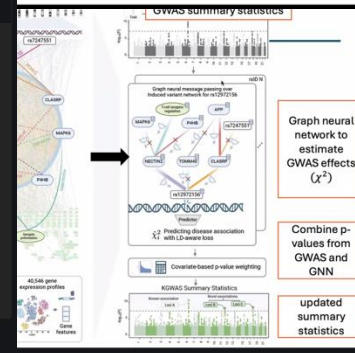
1,000-core cluster

of magnitude faster

for massive-scale datasets

scale sequence search &

- Can we describe the amino acid backbone of proteins as sequences over a structural alphabet and compare structures using sequence alignments?



Family



My research path

Next ...



Dr. David R Kelley



Dr. Johannes Linder



Majed Mohamed
Magzoub

Calico



Liu Yuh-san



Dr. Eric Y. Chuang



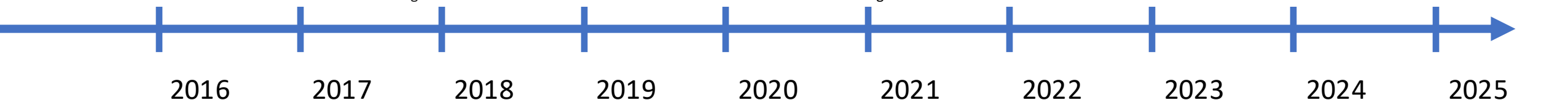
Dr. Tzu-Pin Lu



Dr. Robert Lanfear



Dr. Huai-Kuang Tsai



Taipei Municipal Chien
Kuo High School



國立臺灣大學

National Taiwan University

National
Taiwan
University



Australian National
University



Academia Sinica



Dr. Steven L Salzberg



Dr. Mihaela Pertea



Dr. Ben Langmead



Dr. Anqi Liu

■ Next: Foster City, CA



Dr. Kyle Farh



Dr Kishore Jaganathan

Senior Deep Learning Scientist

@ Illumina AI Lab

illumina®