JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING
Department of Computer Science

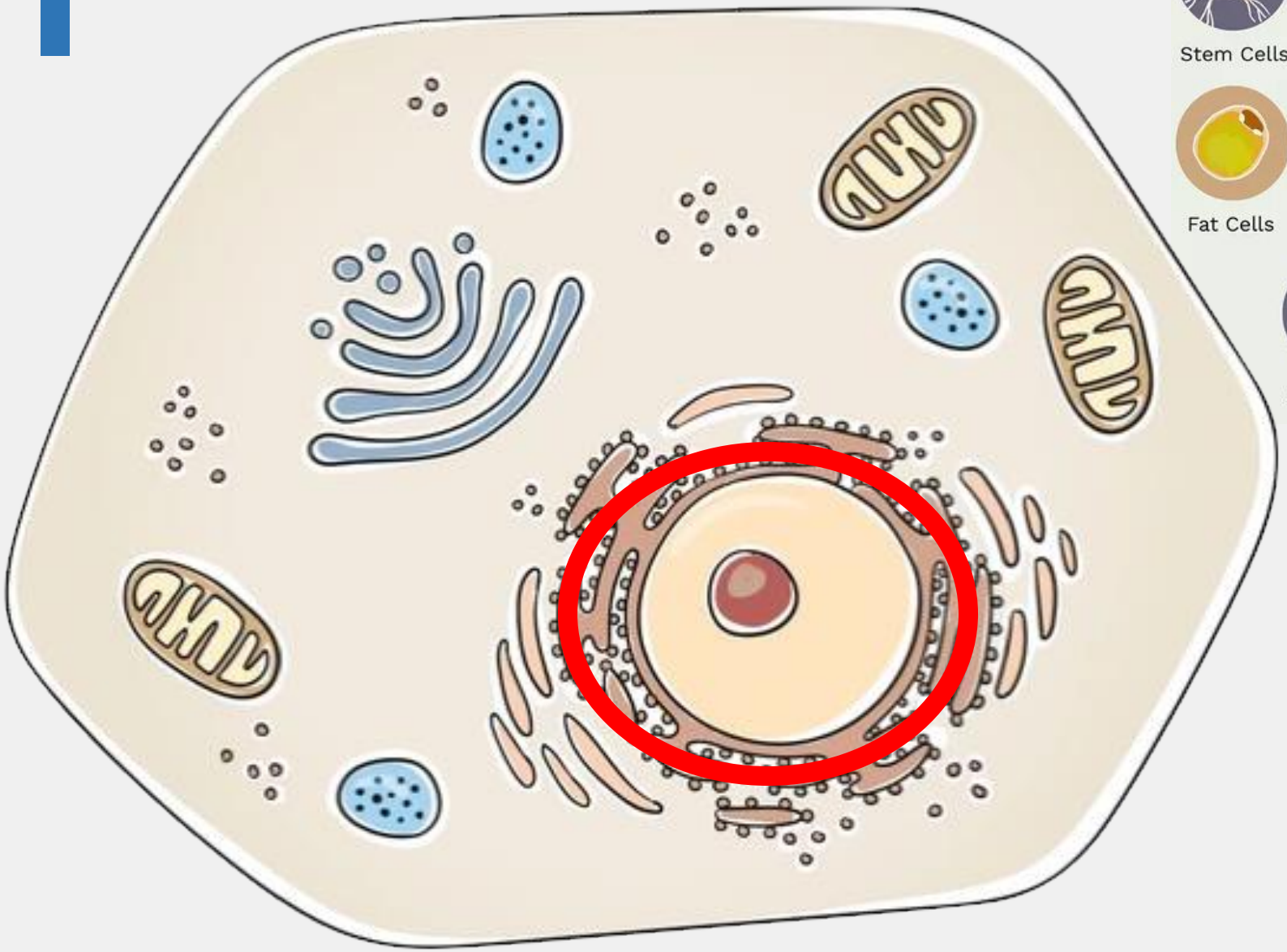
JOHNS HOPKINS UNIVERSITY
CENTER FOR COMPUTATIONAL BIOLOGY
CCB

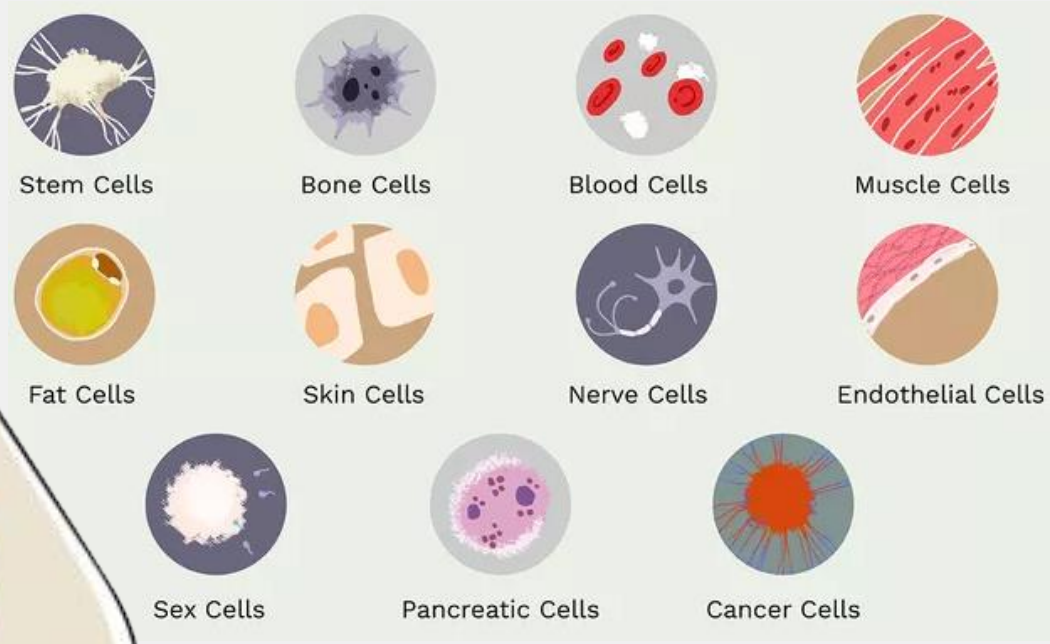
Decoding the Language of Genomes: Bridging Sequences and Function through Deep Learning

Kuan-Hao Chao

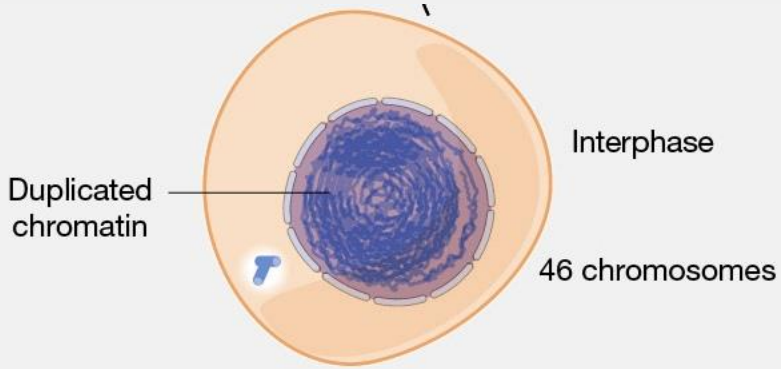
2025.08.25

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

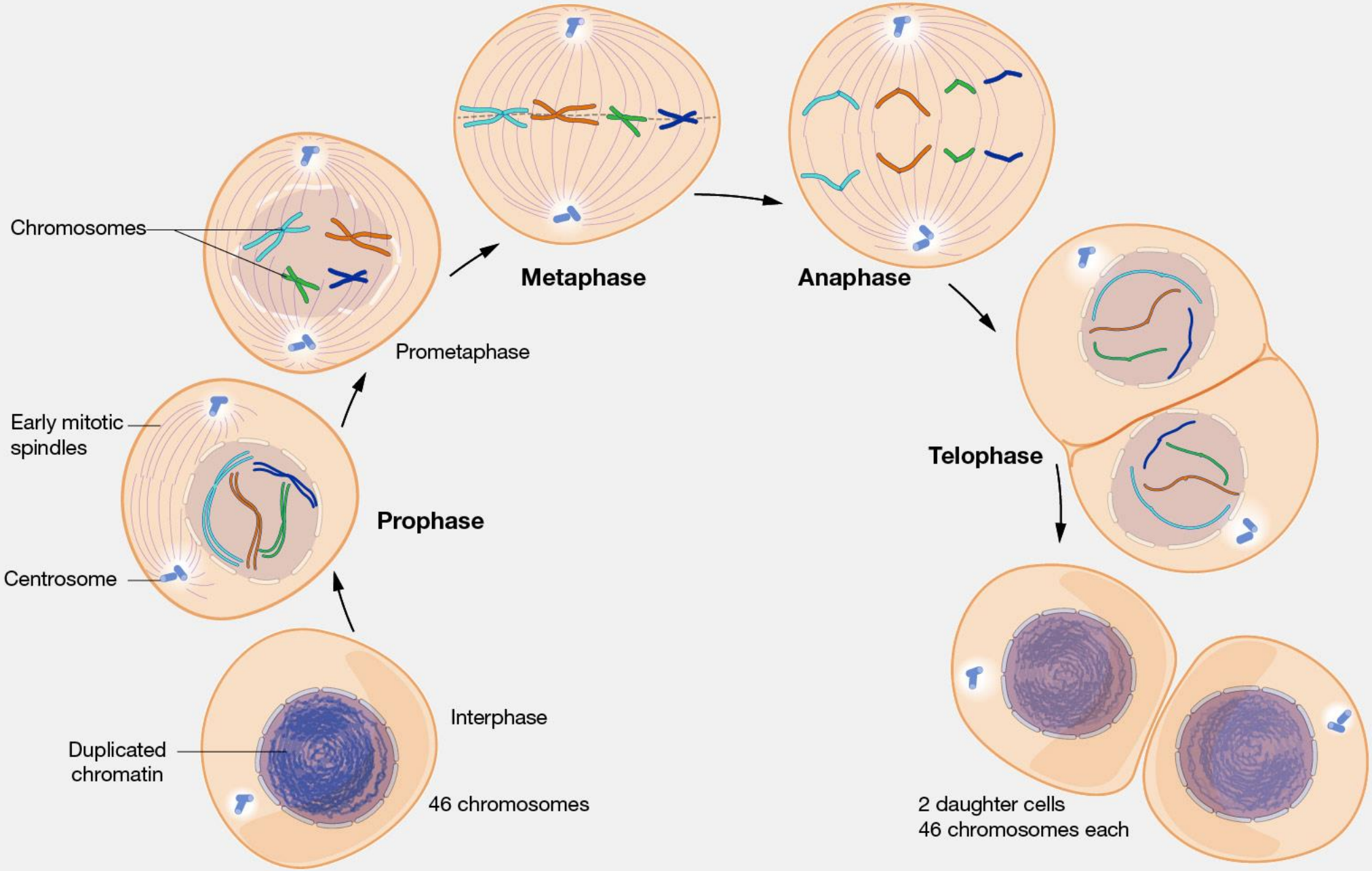



Stem Cells
Bone Cells
Blood Cells
Muscle Cells
Fat Cells
Skin Cells
Nerve Cells
Endothelial Cells
Sex Cells
Pancreatic Cells
Cancer Cells

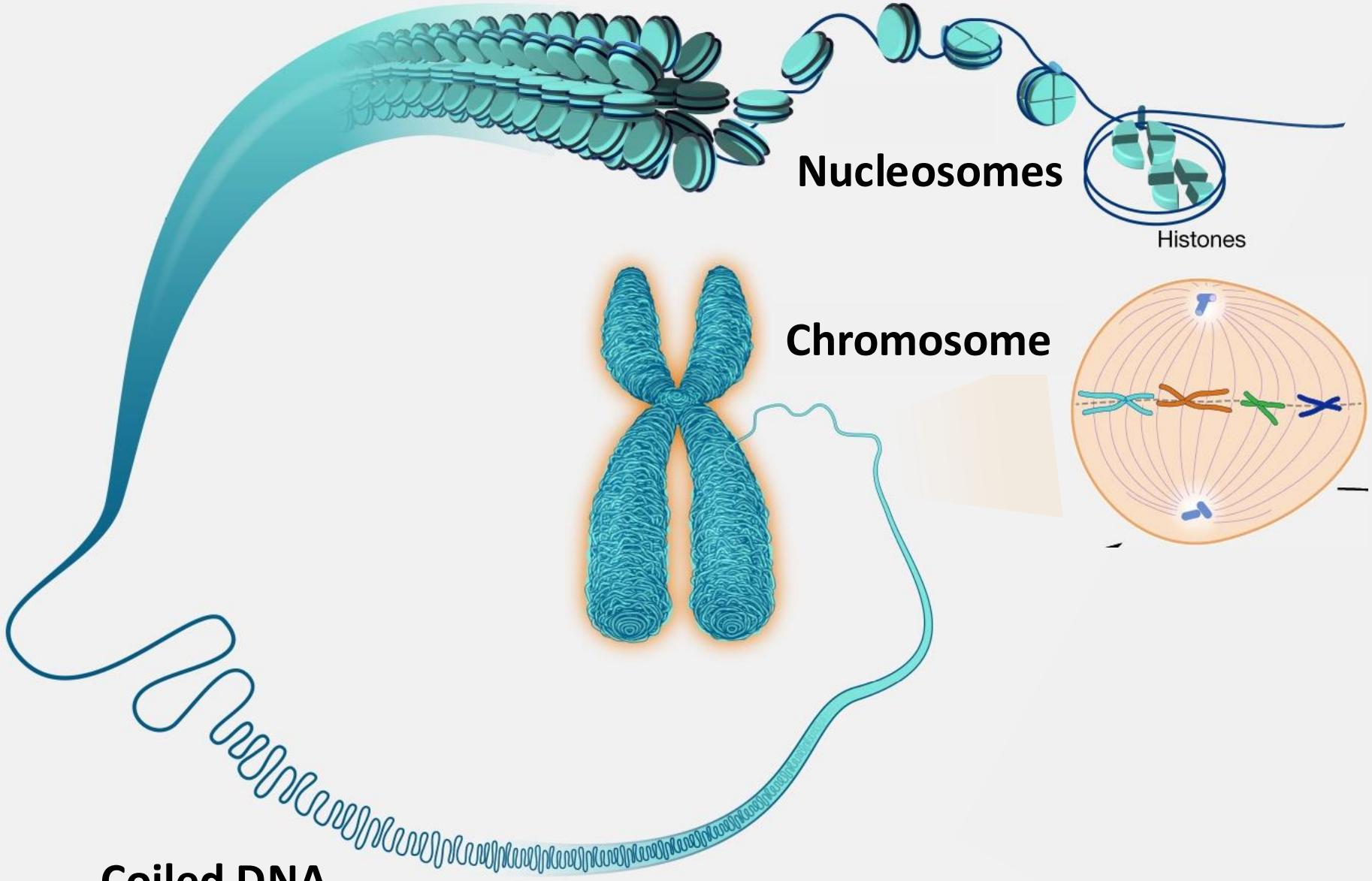
https://www.genome.gov/genetics-glossary/Mitosis



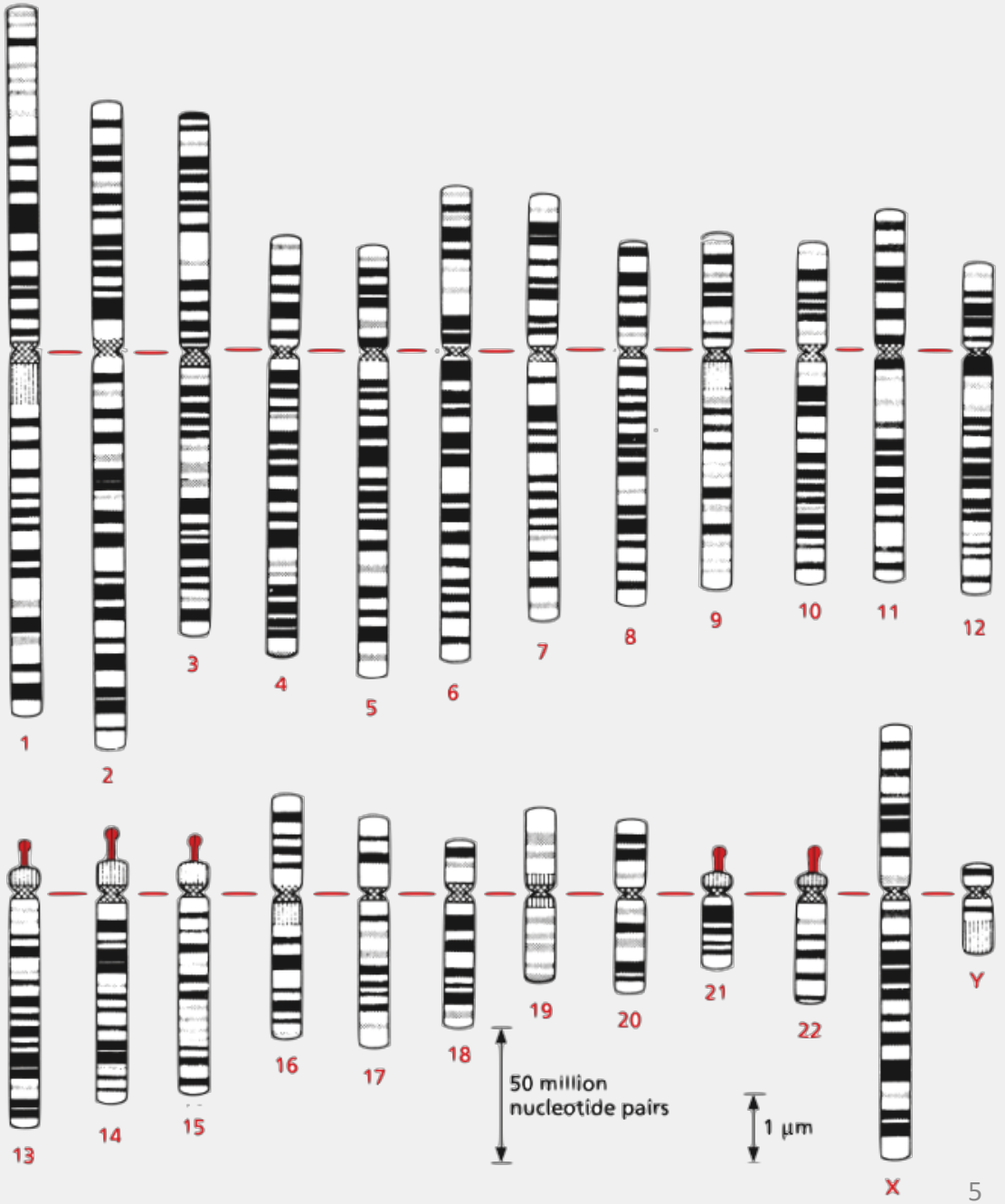
https://www.genome.gov/genetics-glossary/Mitosis

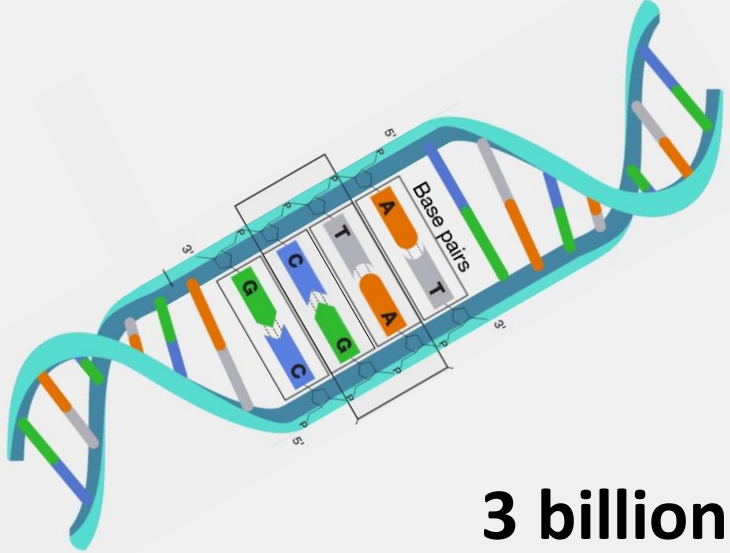


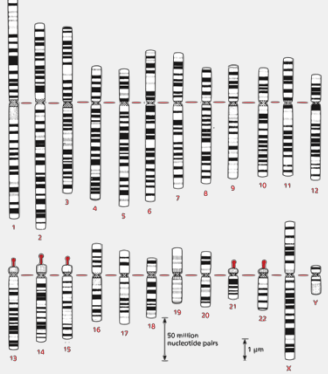
https://www.genome.gov/genetics-glossary/Chromosome




Deoxyribonucleic acid (DNA)
Base pairs
A
T
C
G
5'
3'

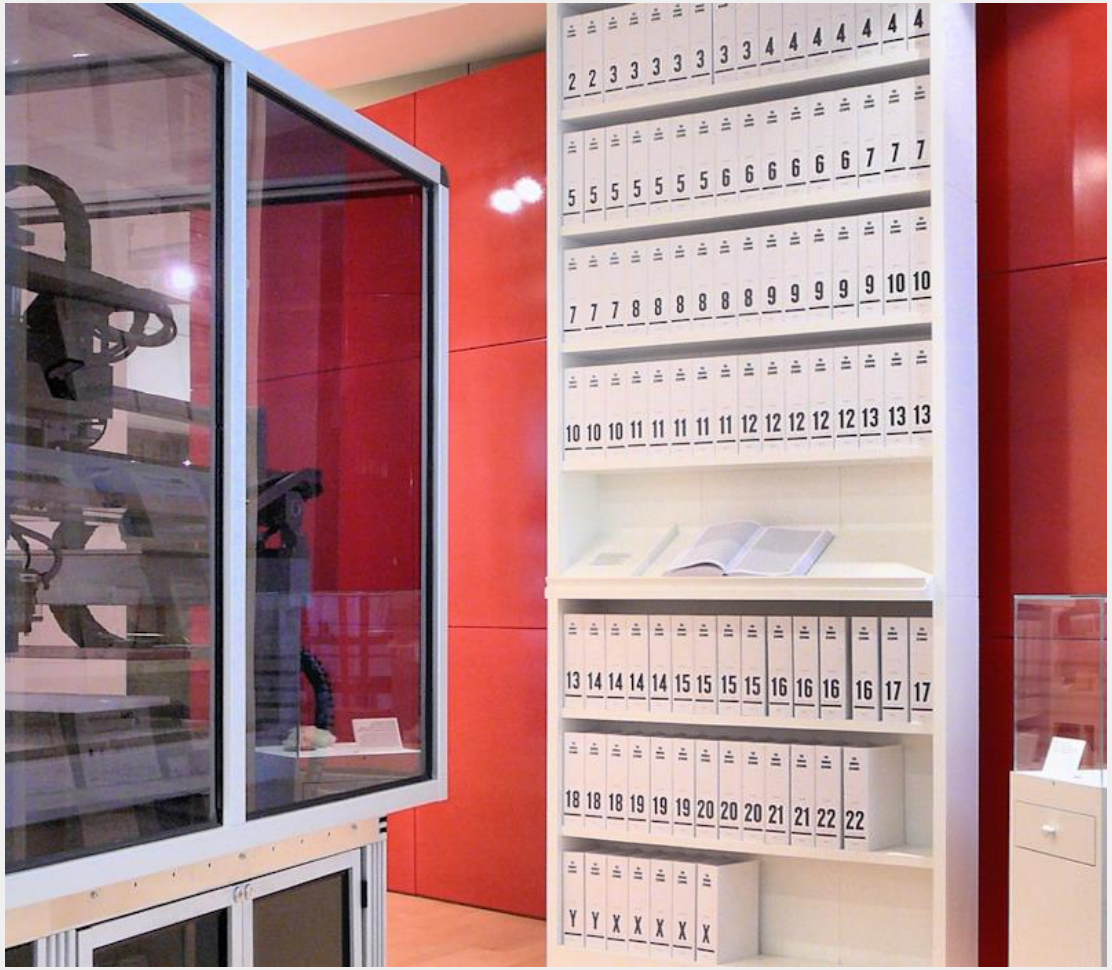

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
50 million
nucleotide pairs
1 µm





3 billions of ACGT!

Wellcome genome bookcase



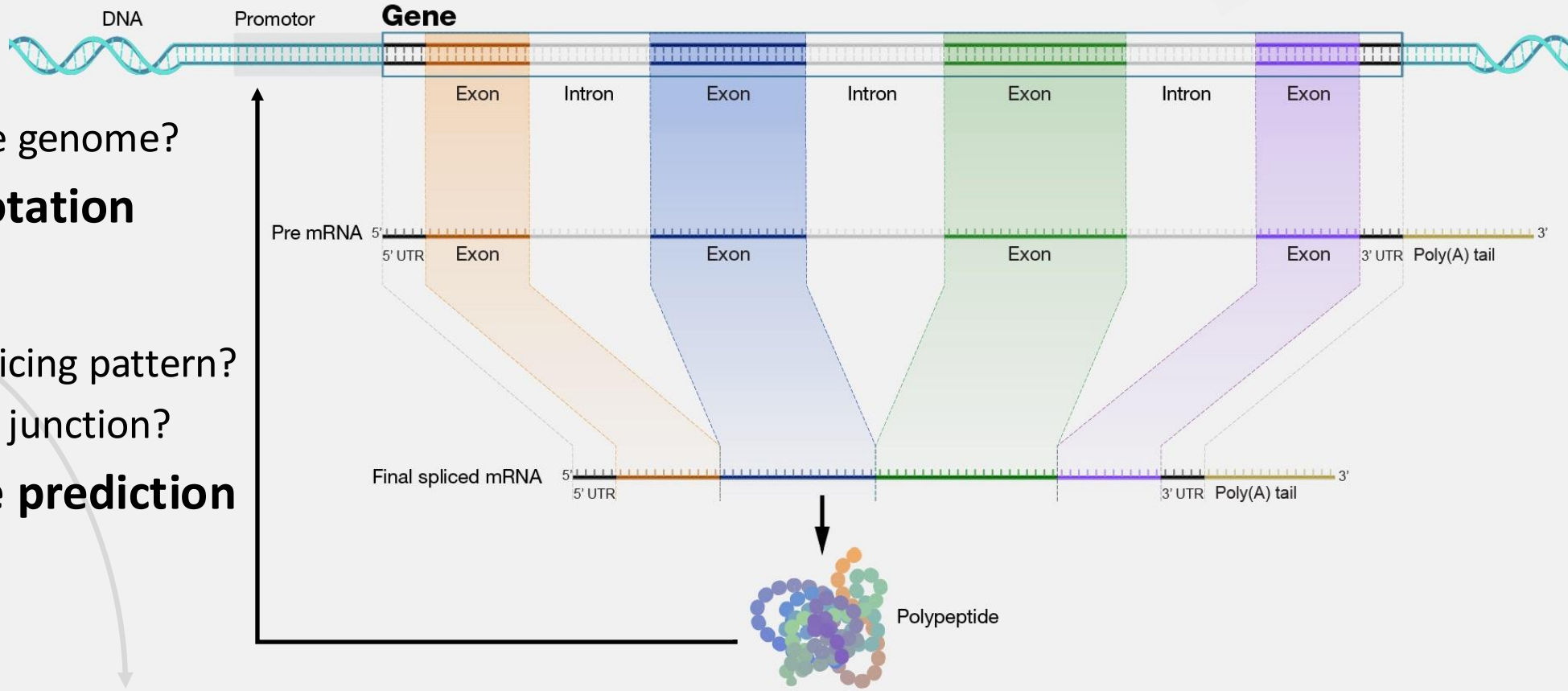
How do we assemble the complete 3-billion-nucleotide genome?

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

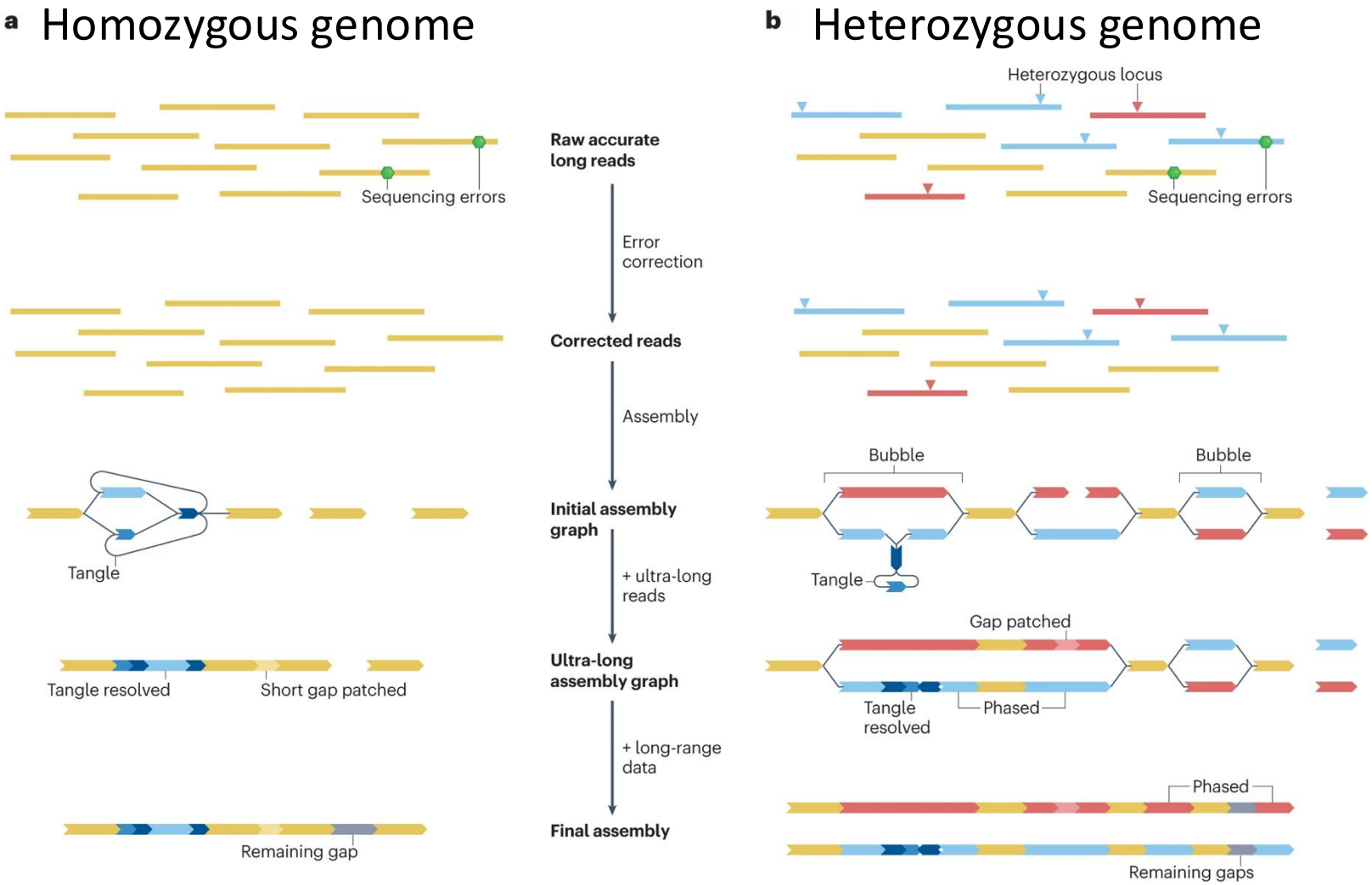
Genome Assembly & Indexing

- 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.

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).

Genome Assembly



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

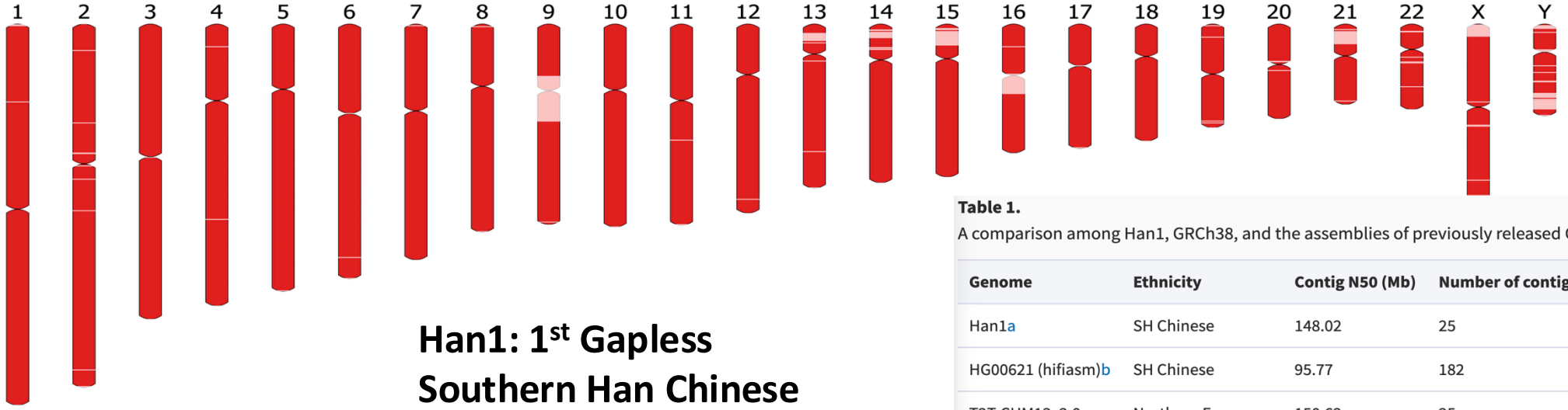
Li, H., Durbin, R. Genome assembly in the telomere-to-telomere era. Nat Rev Genet 25, 658–670 (2024).

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

NCBI GenBank assembly: GCA_024586135.1

G3 Genes | Genomes | Genetics



Han1: 1st Gapless Southern Han Chinese

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

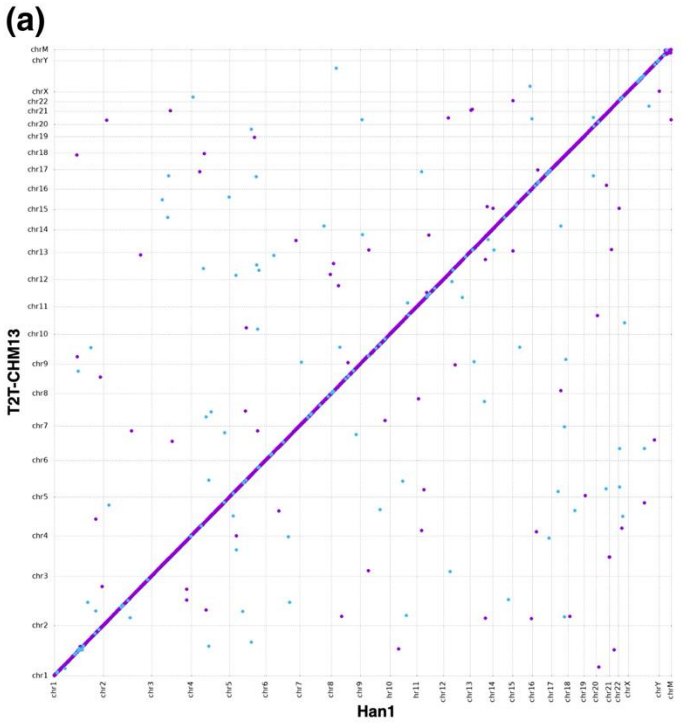
Genome	Ethnicity	Contig N50 (Mb)	Number of contigs	Number of gaps	Assembly size (Gb)
Han1a	SH Chinese	148.02	25	0	3.10
HG00621 (hifiasm)b	SH Chinese	95.77	182	157	3.11
T2T-CHM13v2.0c	Northern European	150.62	25	0	3.12
HJ-H1d	NH Chinese	28.15	1,330	427	3.07
HJ-H2d	NH Chinese	25.90	896	390	2.91
NH1d	NH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

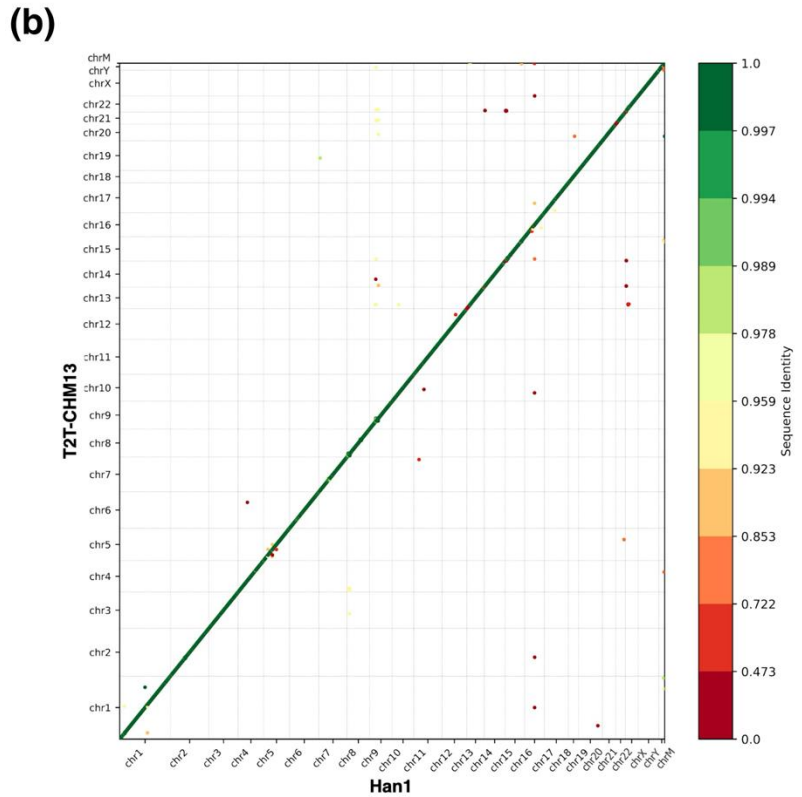
Assembler	Sequencing data	assembled sequence (bp)	Contig N50	Number of contigs	Quality value
Hifiasm v0.16.1-r375	PacBio HiFi	3,110,501,483	95,769,069	182	57.8a
Flye v2.5	ONT Ultralong	2,974,205,132	40,850,737	1,658	25.6a

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

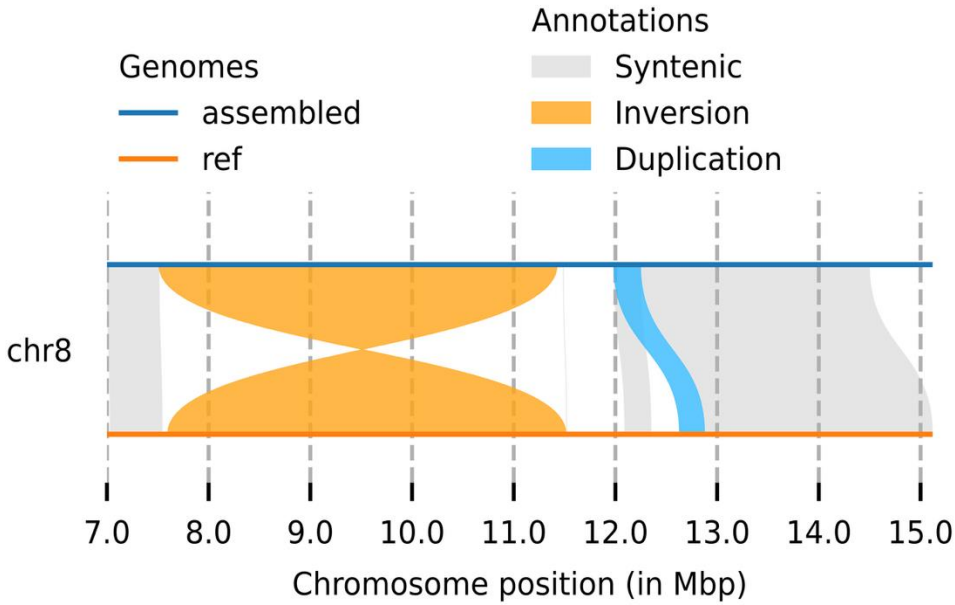
Mummer dot plot



Gene order plot

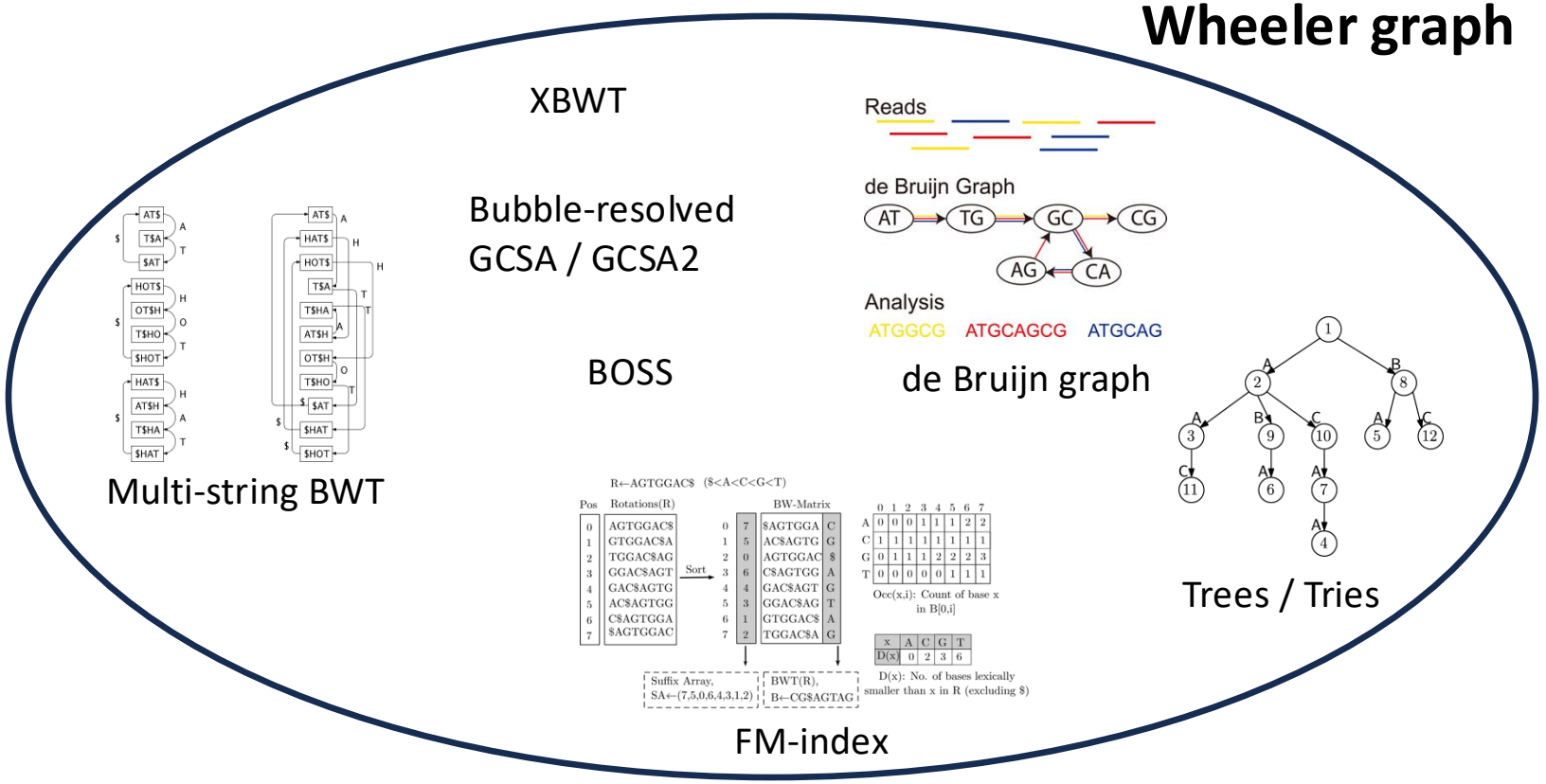


Inversion on β defensin gene cluster



Wheeler Graph Recognition

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



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)

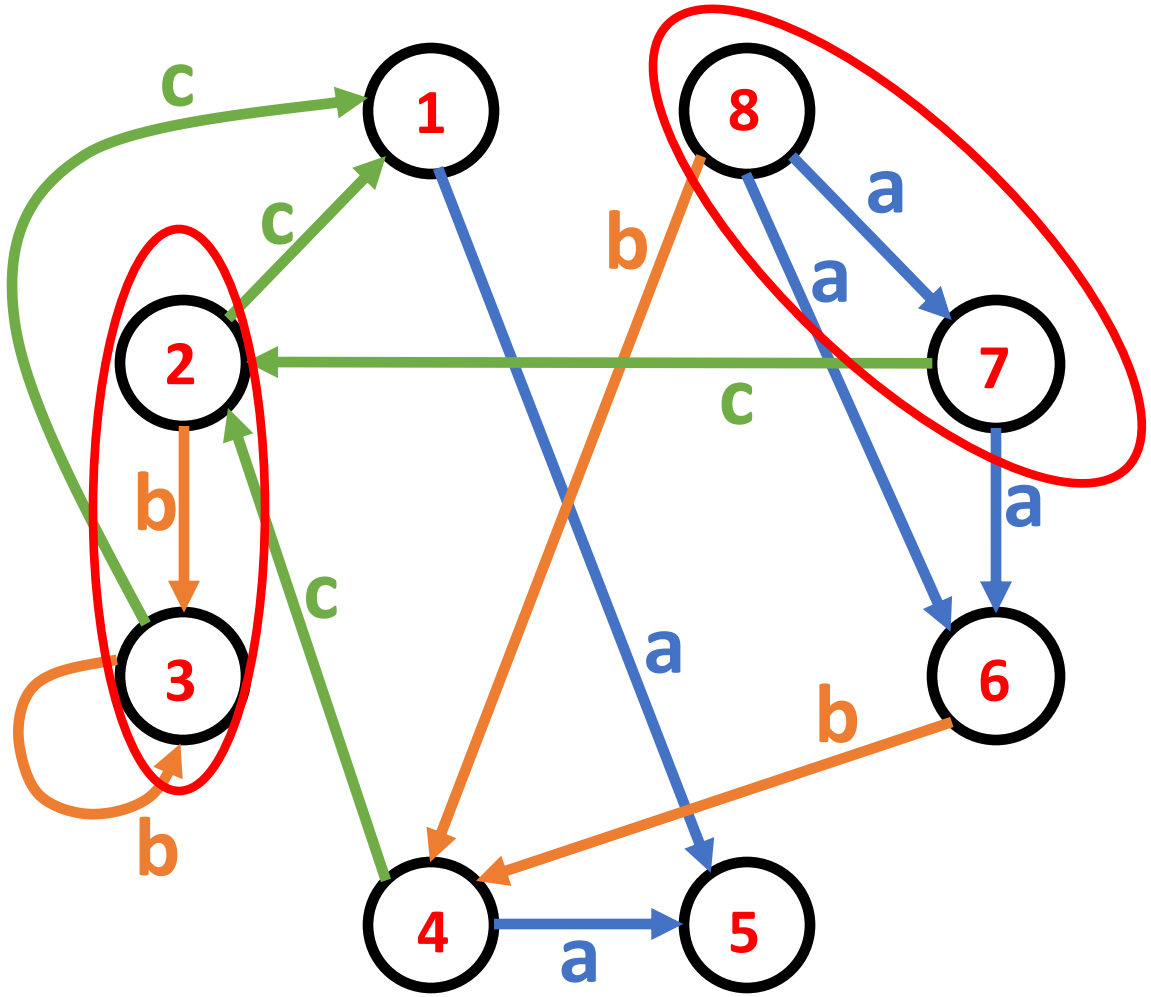
2021 vg Giraffe

Gagie et al. Theoretical computer science (2017)

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).

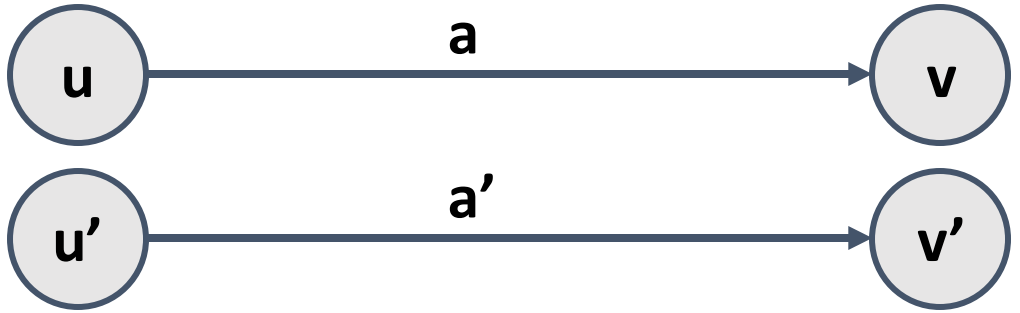
Wheeler Graph Recognition

RECOMB-SEQ iScience



Wheeler graph recognition problem

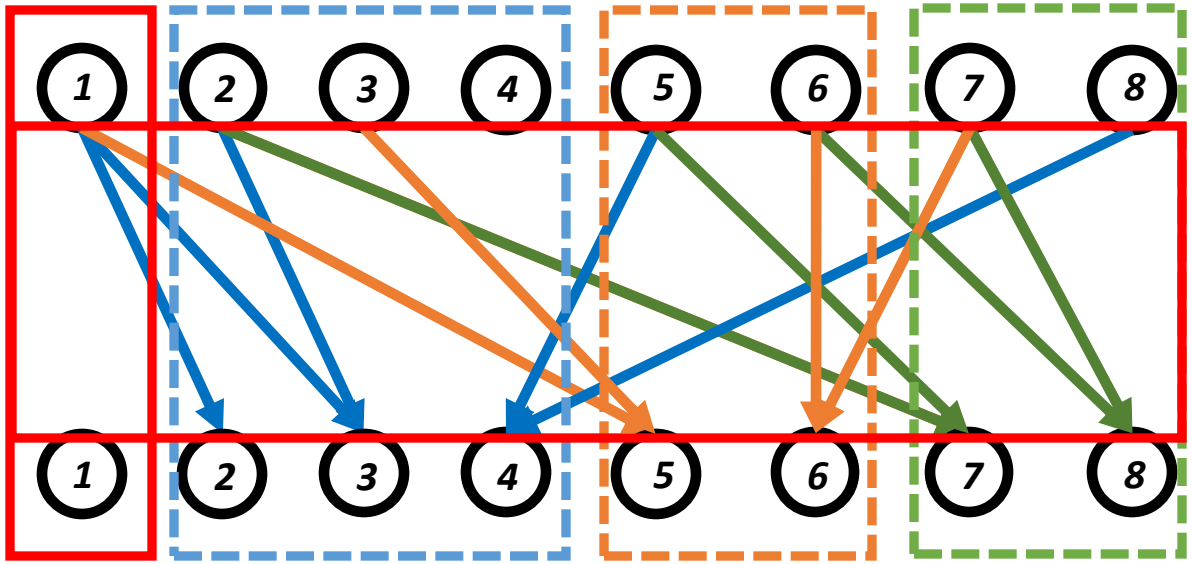
1. Node with indegree 0 comes before every other nodes.
2. $$a \prec a' \implies v < v'$$
3. $$(a = a') \wedge (u < u') \implies v \leq v'$$



Satisfiability modulo theories (SMT)

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

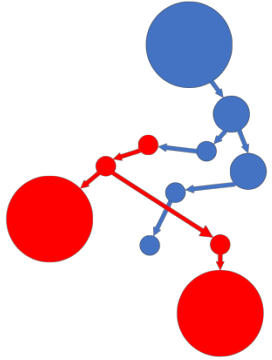
1.	Node with indegree 0 comes before every other nodes.
2.	$a \prec a' \implies v < v'$
3.	$(a = a') \wedge (u < u') \implies v \leq v'$




```
##############################
## 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))
```

Gagie et al. Theoretical computer science (2017)

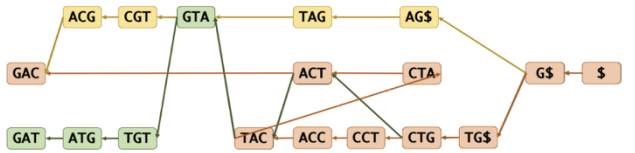
Wheeler Graph Recognition

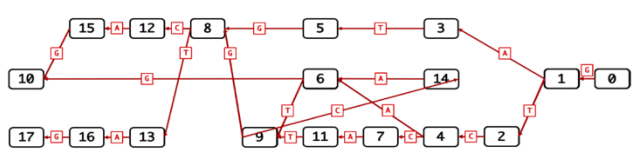


WGT
Wheeler Graph Toolkit

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

De Bruijn Graph



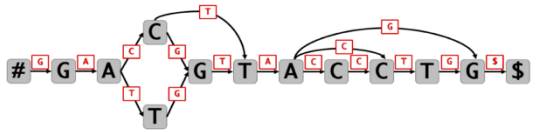


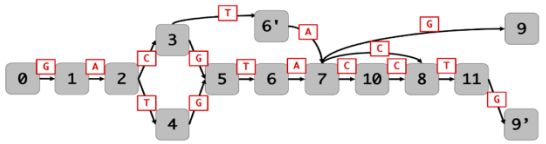
Original label	Wheeler order	Original label	Wheeler order
0	1	4	10
11	2	1	11
15	3	10	12
6	4	17	13
3	5	8	14
16	6	9	15
7	7	5	16
12	8	2	17
14	9	13	18

18 nodes, 21 edges
1 “distant gluing”

GCSA (Sirén et al., 2014)

Reverse deterministic graph

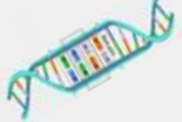


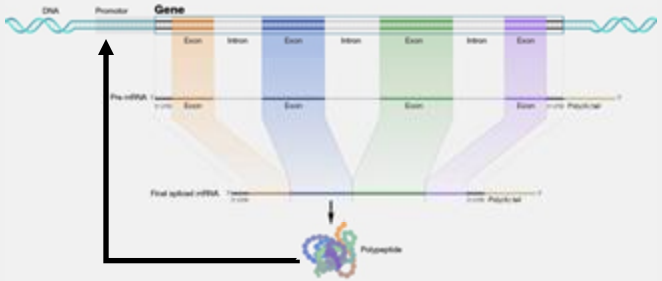


Original label	Wheeler order	Original label	Wheeler order
0	1	9	8
2	2	5	9
7	3	9'	10
3	4	4	11
10	5	6'	12
8	6	11	13
1	7	6	14

14 nodes, 16 edges
No “distant gluing”

Part III





Steven Salzberg

Mihaela Pertea

Genome Annotation

- LiftOn: genome annotation lift-over
- Application: CHESS (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 RESEARCH

LiftOn

Genome Annotation

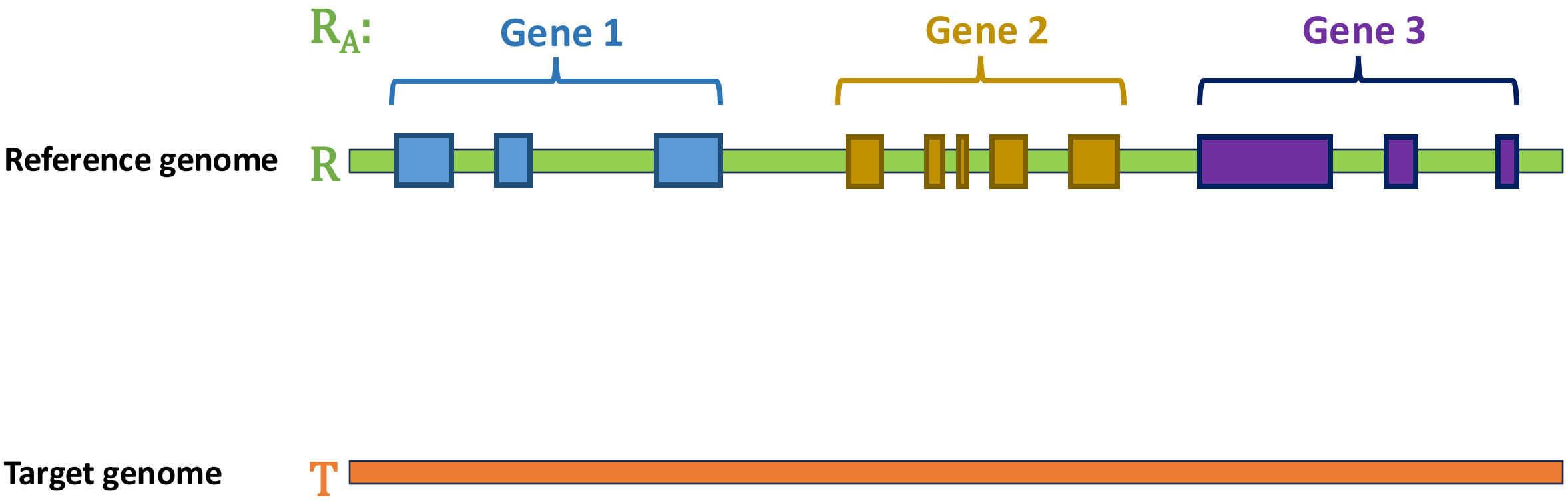
Genome (FASTA)

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

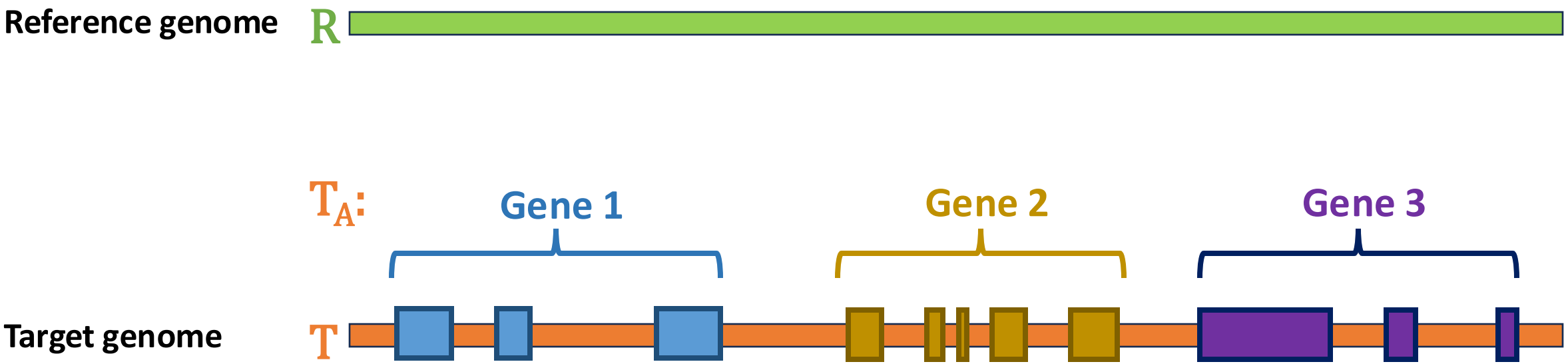
Annotation (GFF / GTF)

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

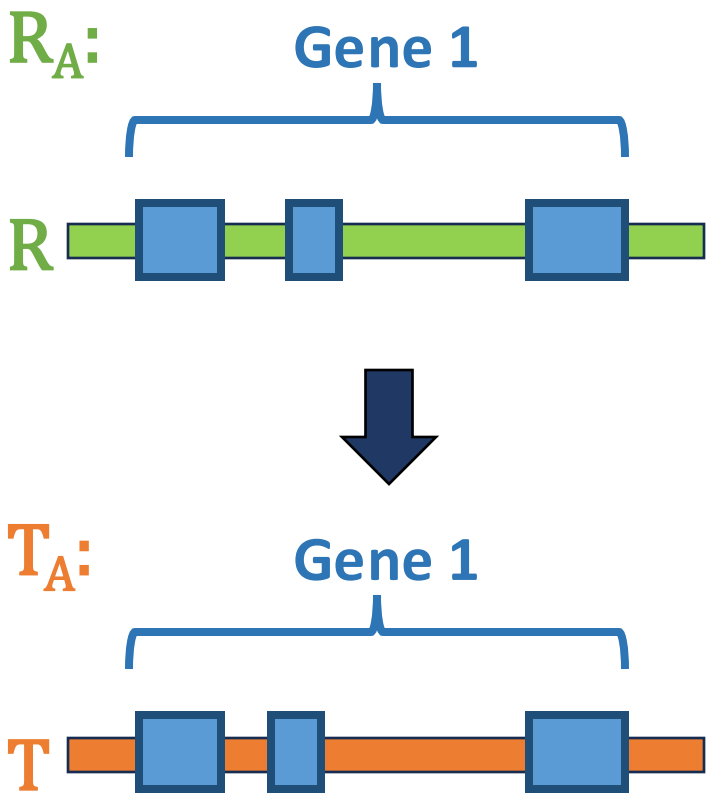
Lift-over Problem Definition:



Lift-over Problem Definition:



Lift-over problem, what methods are available?



Chain file

UCSC liftOver

Kuhn et al.

CrossMap

Zhao et al.

DNA-based

Liftoff

github.com/agshumate/Liftoff

Shumate et al.

minimap2

github.com/lh3/minimap2

Li

Protein-based

miniprot

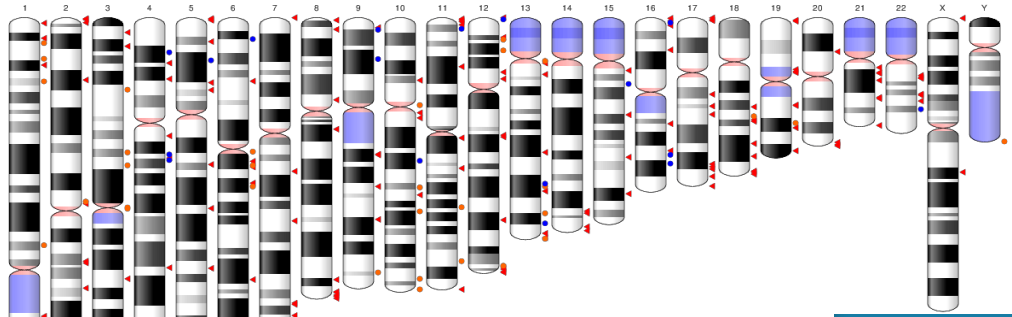
github.com/lh3/miniprot

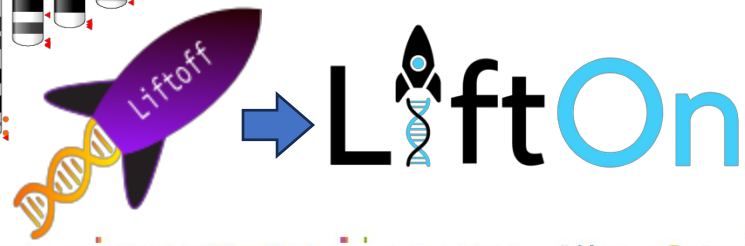
Li

Application: GRCh38 to T2T-CHM13 lift-over

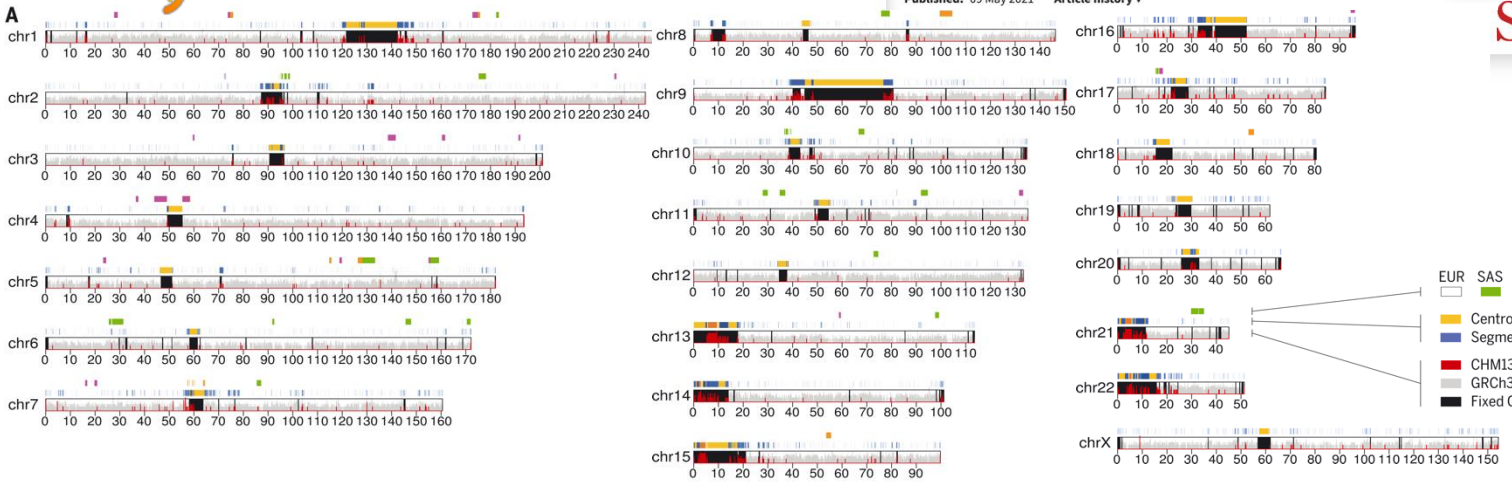
NCBI Build 34
NCBI Build 35
NCBI Build 36.1
GRCh37

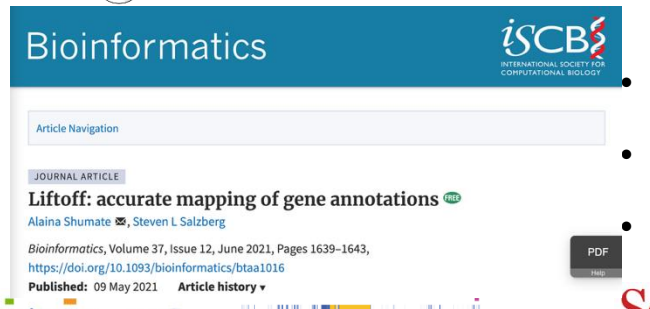
GRCh38
(Dec 2013)



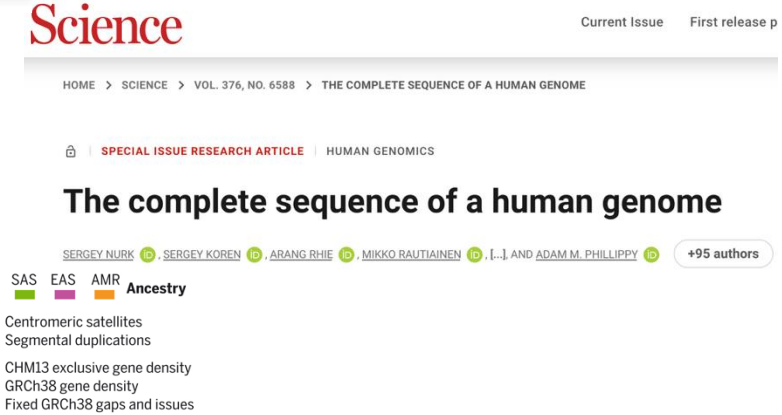


T2T-CHM13
(Jan 2022)



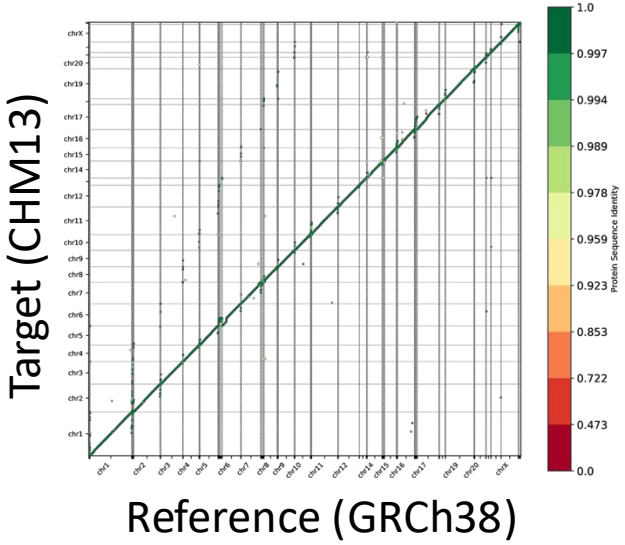


- 10 Mbp of rDNAs
- 182 Mbp of entirely novel sequence
- 1956 novel genes including 99 protein-coding

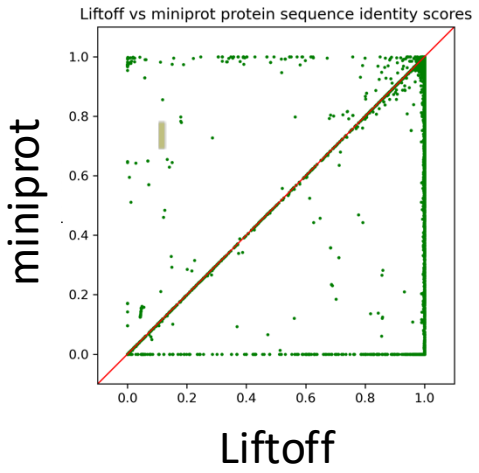


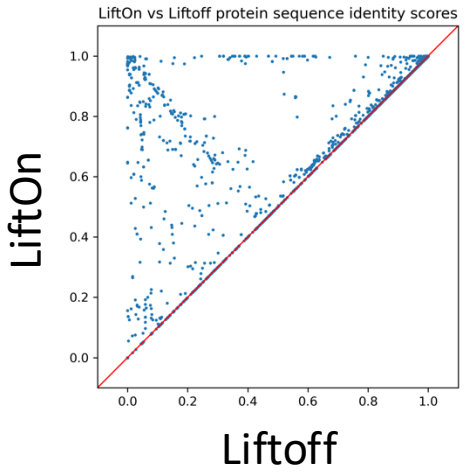
Result 1: improves DNA & protein-based lift-over

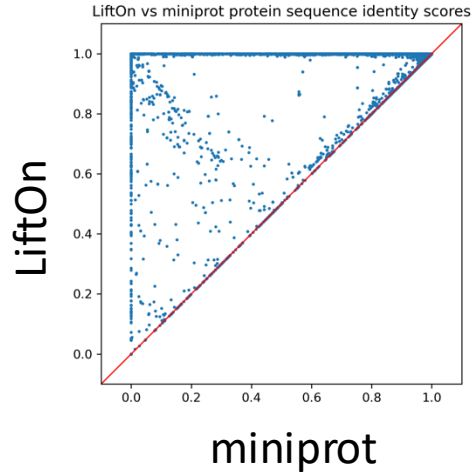
Map RefSeq v220 from GRCh38 -> CHM13V2.0

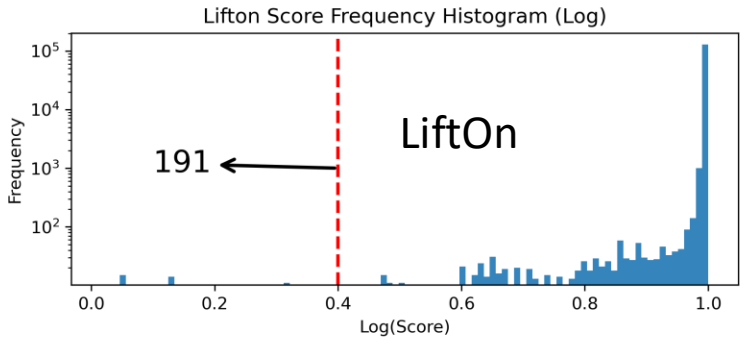


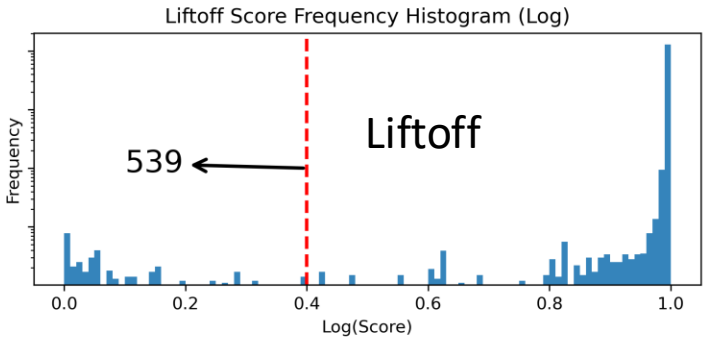
Compressed-gap protein sequence identity

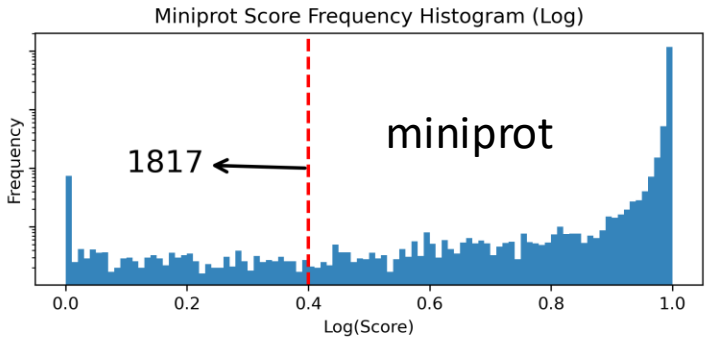




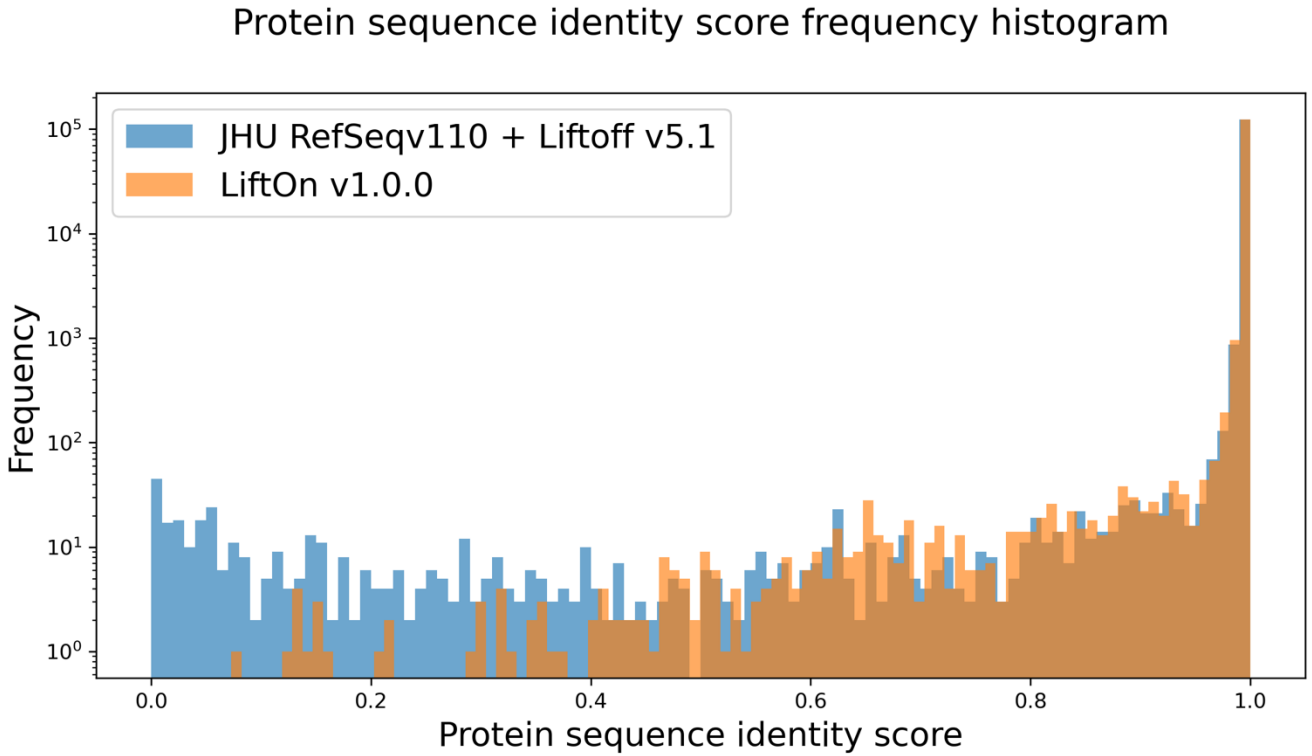


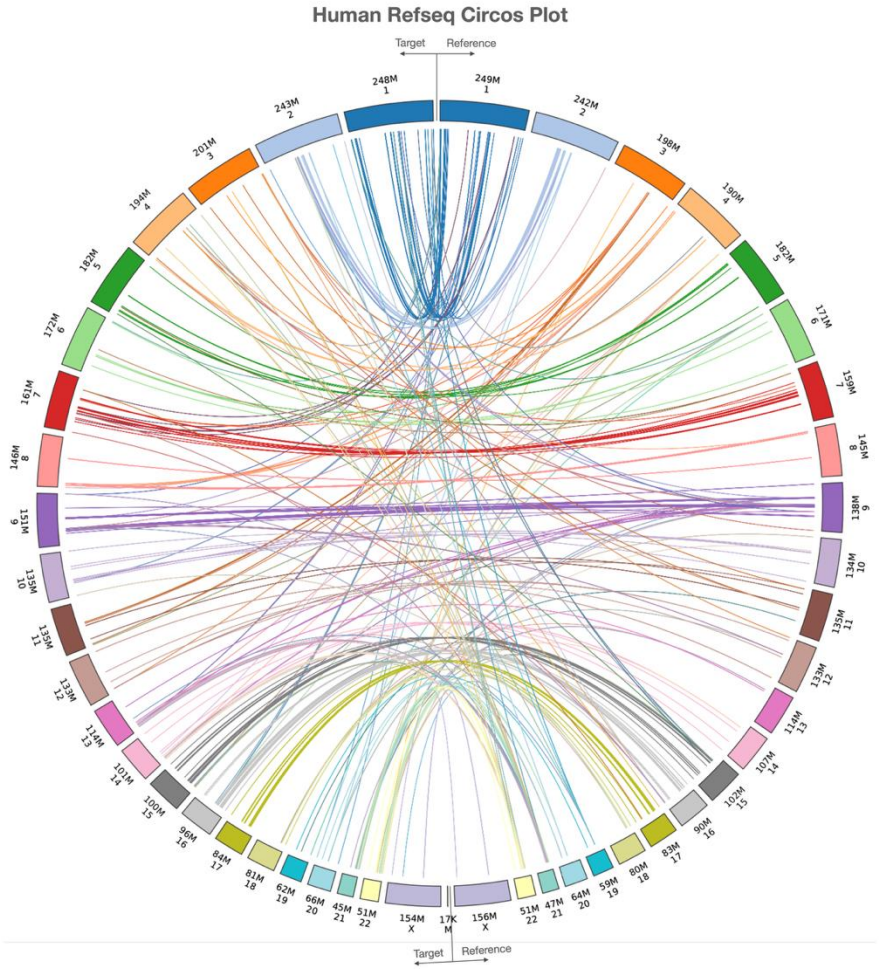






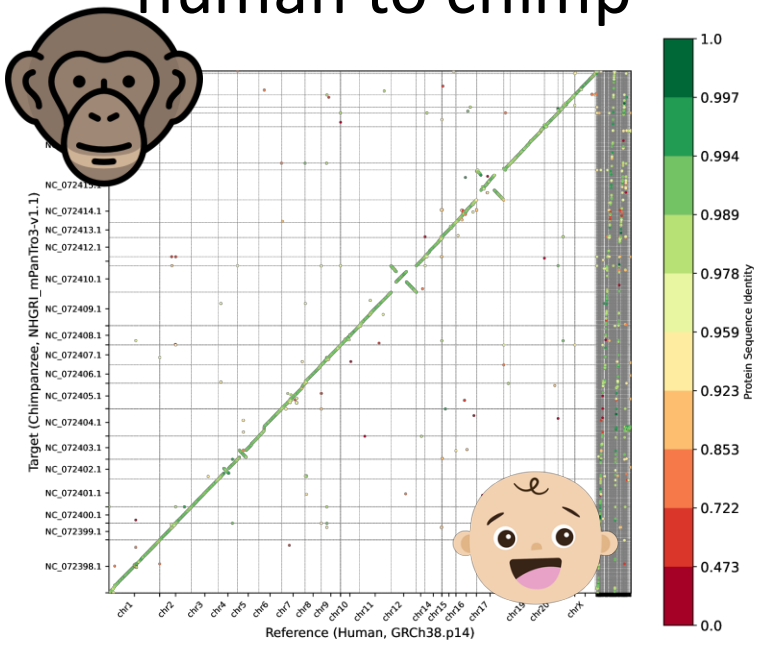
Result 2: improve CHM13 protein annotations





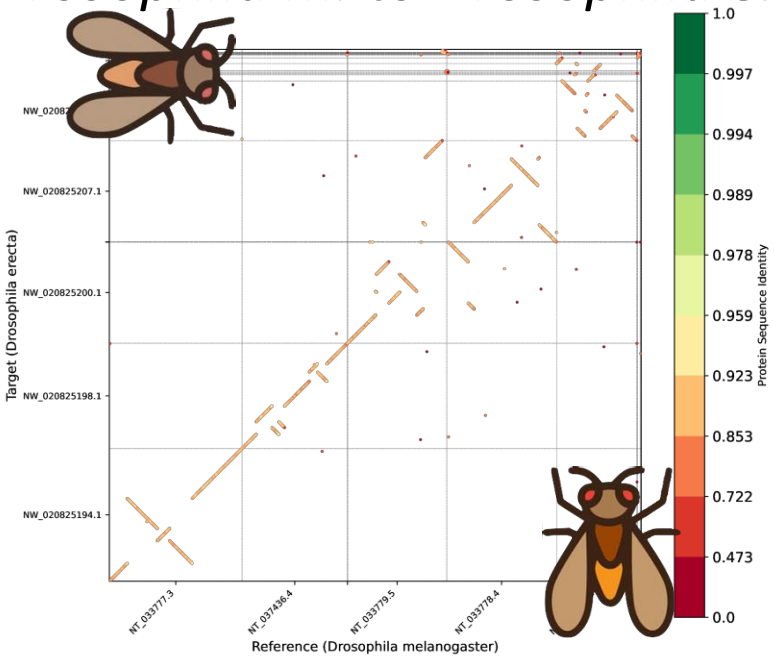
Result 3: improve distant species lift-over

human to chimp



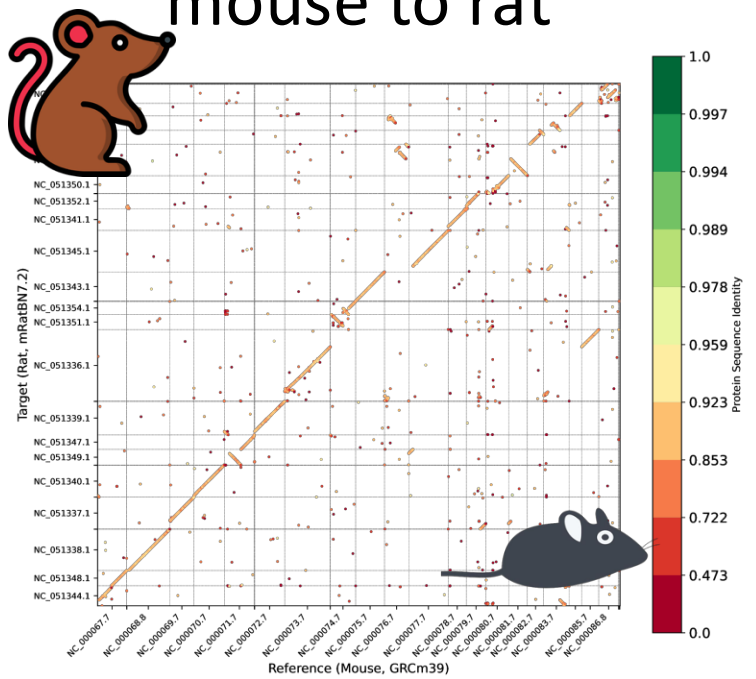
Mash: 0.013 Dashing2: 0.47

Drosophila m. to *Drosophila e.*



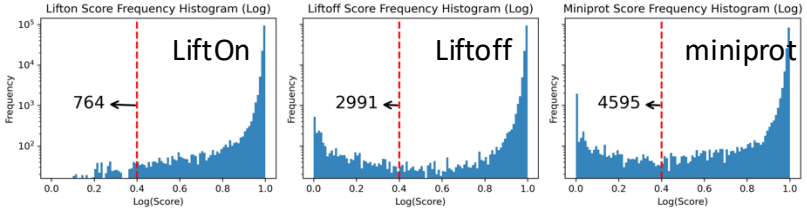
Mash: 0.077 Dashing2: 0.07

mouse to rat

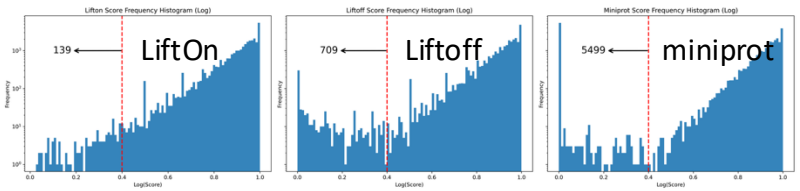


Mash: 0.12 Dashing2: 0.01

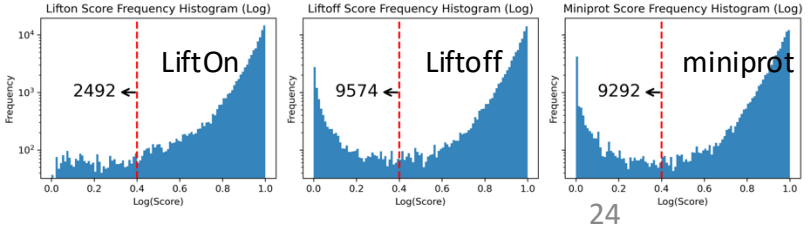
DNA + Protein DNA-only Protein-only



DNA + Protein DNA-only Protein-only

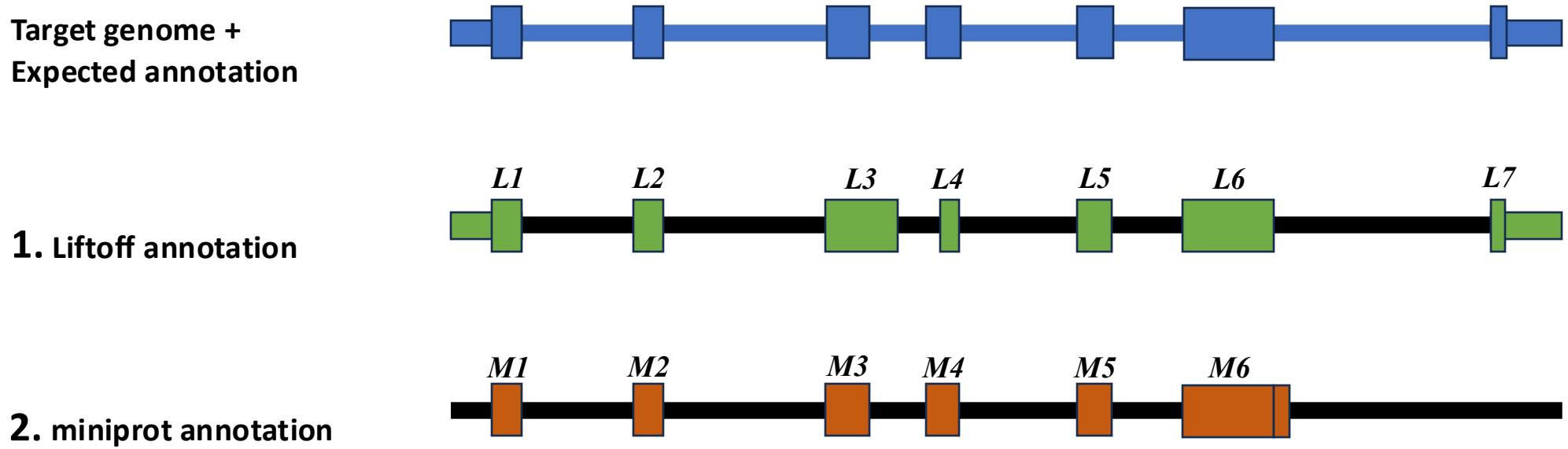


DNA + Protein DNA-only Protein-only



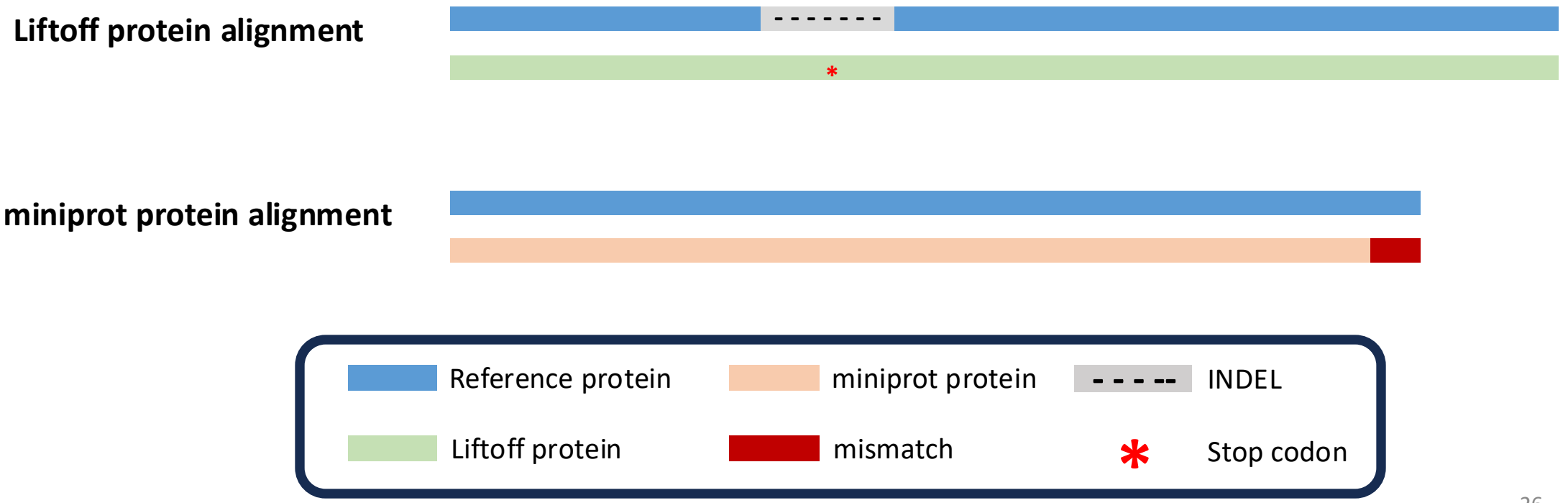
LiftOn: Protein-maximization algorithm

A



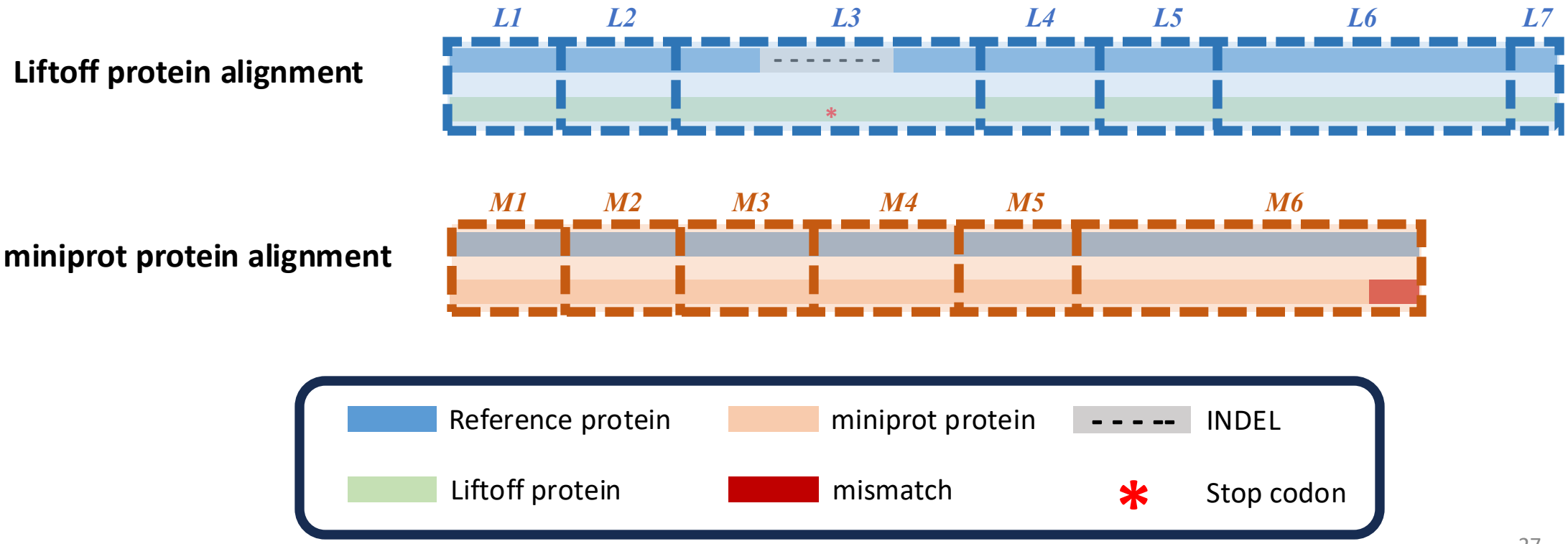
LiftOn: Protein-maximization algorithm

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



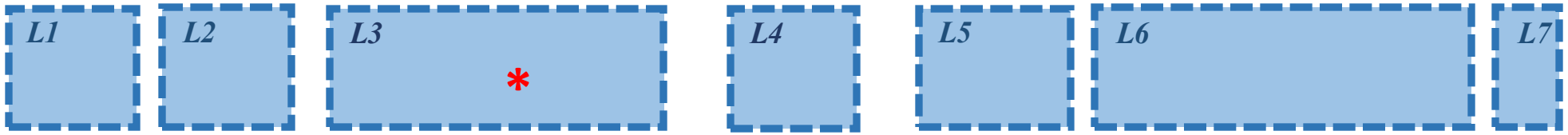
LiftOn: Protein-maximization algorithm

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



LiftOn: Protein-maximization algorithm

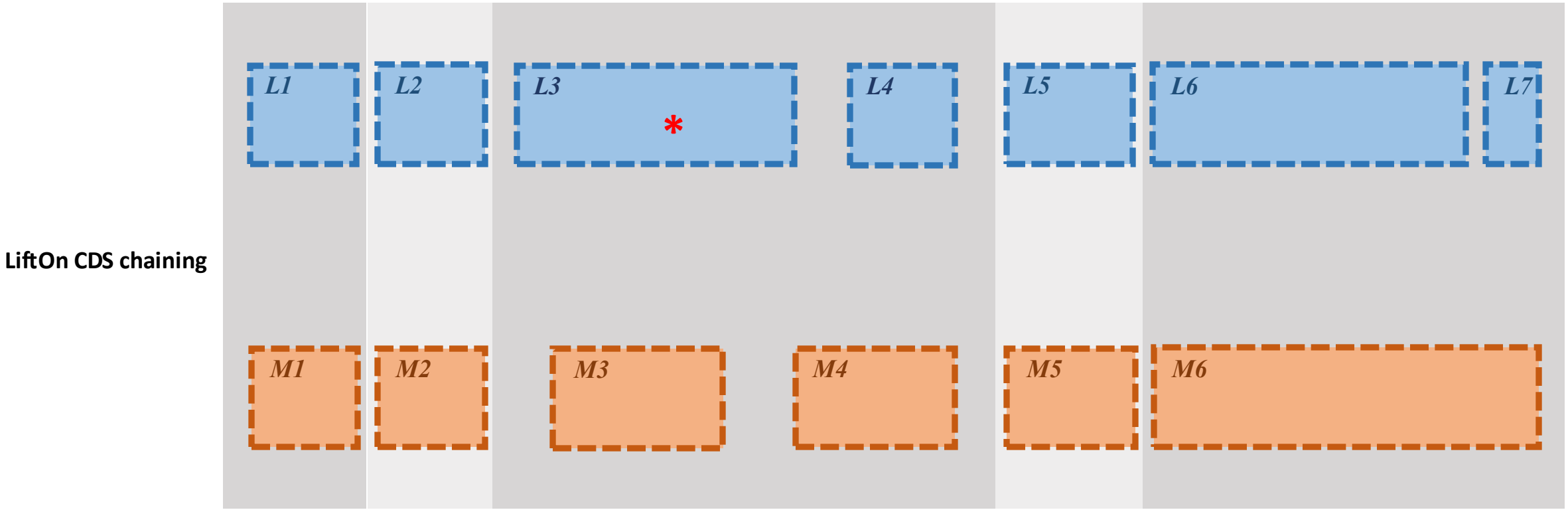
D Step 3: group CDSs by “accumulated AA in the reference protein”



LiftOn CDS chaining

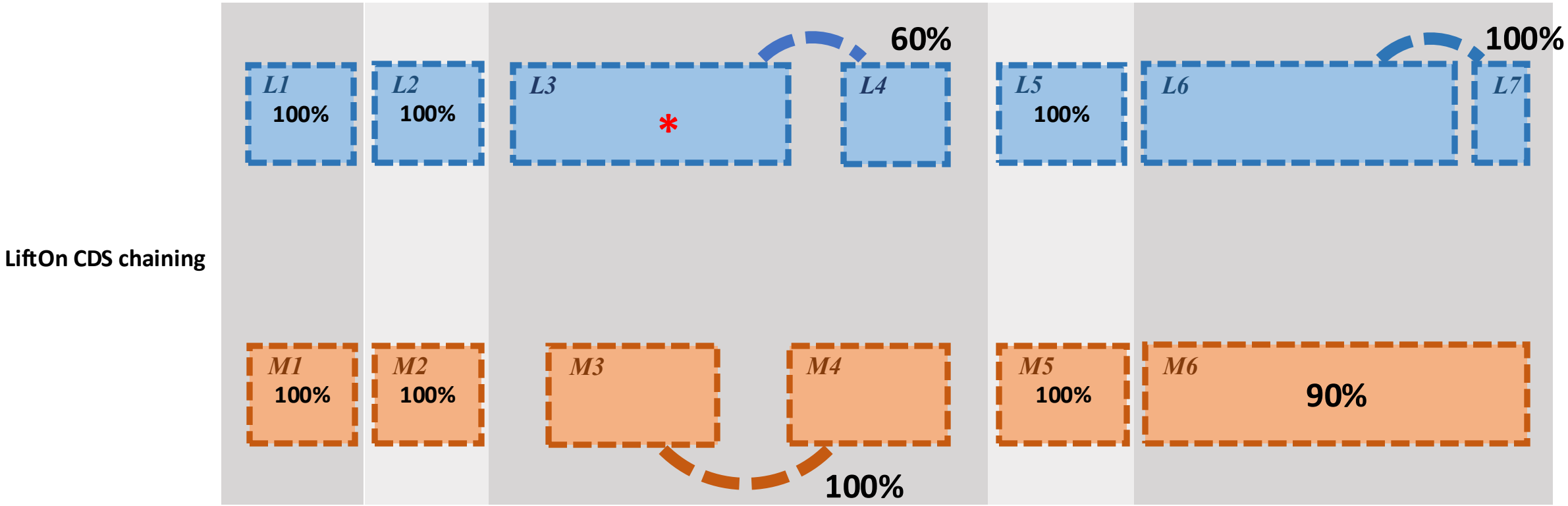
LiftOn: Protein-maximization algorithm

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



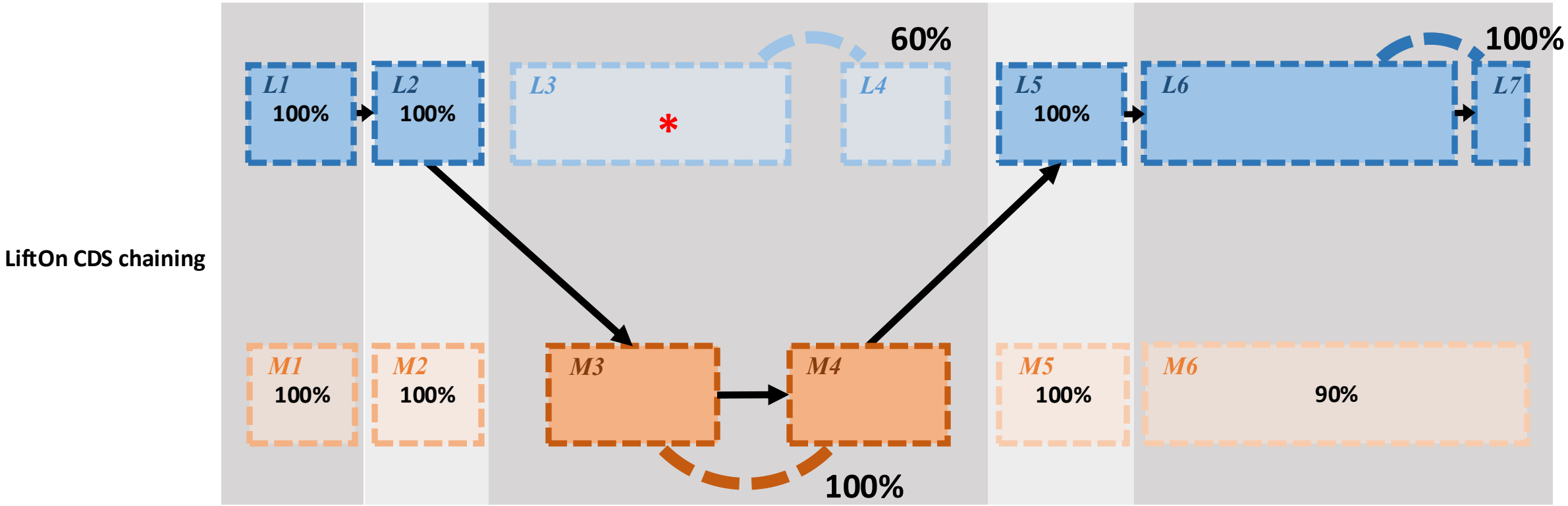
LiftOn: Protein-maximization algorithm

D **Step 3: group CDSs by "accumulated AA in the reference protein"**

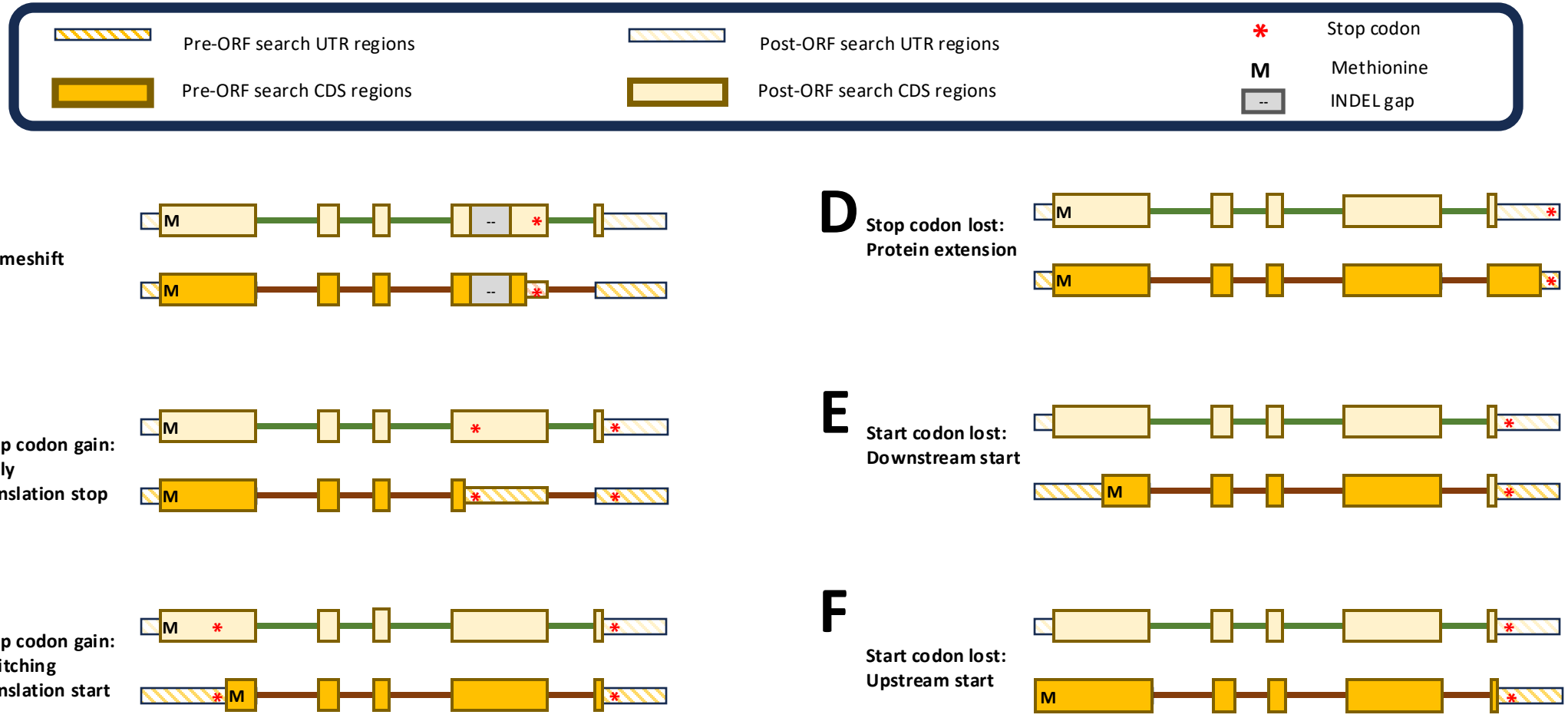


LiftOn: Protein-maximization algorithm

D **Step 3: group CDSs by "accumulated AA in the reference protein"**



LiftOn: Protein-maximization algorithm



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.

LiftOn: more lift-over applications

GENOME RESEARCH

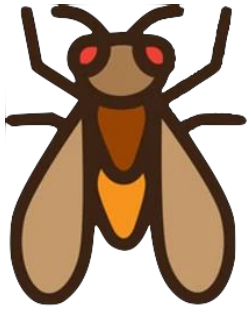
LiftOn

House mouse
(*Mus musculus*)

Yeast
(*Saccharomyces cerevisiae*)

Thale cress
(*Arabidopsis thaliana*)

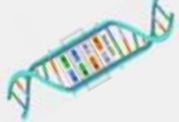
Honey bee
(*Apis mellifera*)

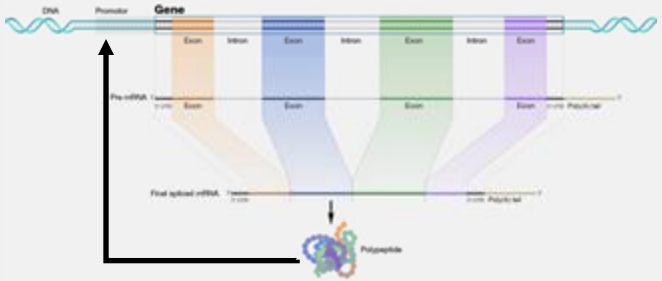


fruit fly
(*Drosophila melanogaster*)

Rice
(*Oryza sativa*)

Part IV & V






Mihaela Pertea


Steven Salzberg


Anqi Liu

Splice Site Prediction

- OpenSpliceAI
- Splam

SPLAM

Genome Biology

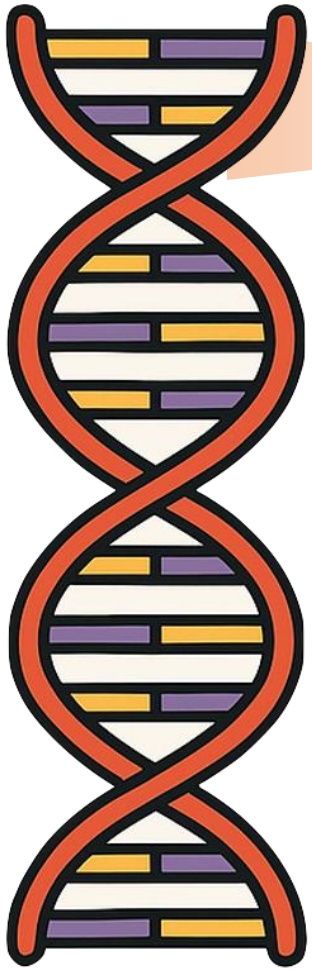
Chao, K. H., Mao, A., Salzberg, S. L., & Pertea, M. (2024). Splam: a deep-learning-based splice site predictor that improves spliced alignments. Genome biology, 25(1), 243.

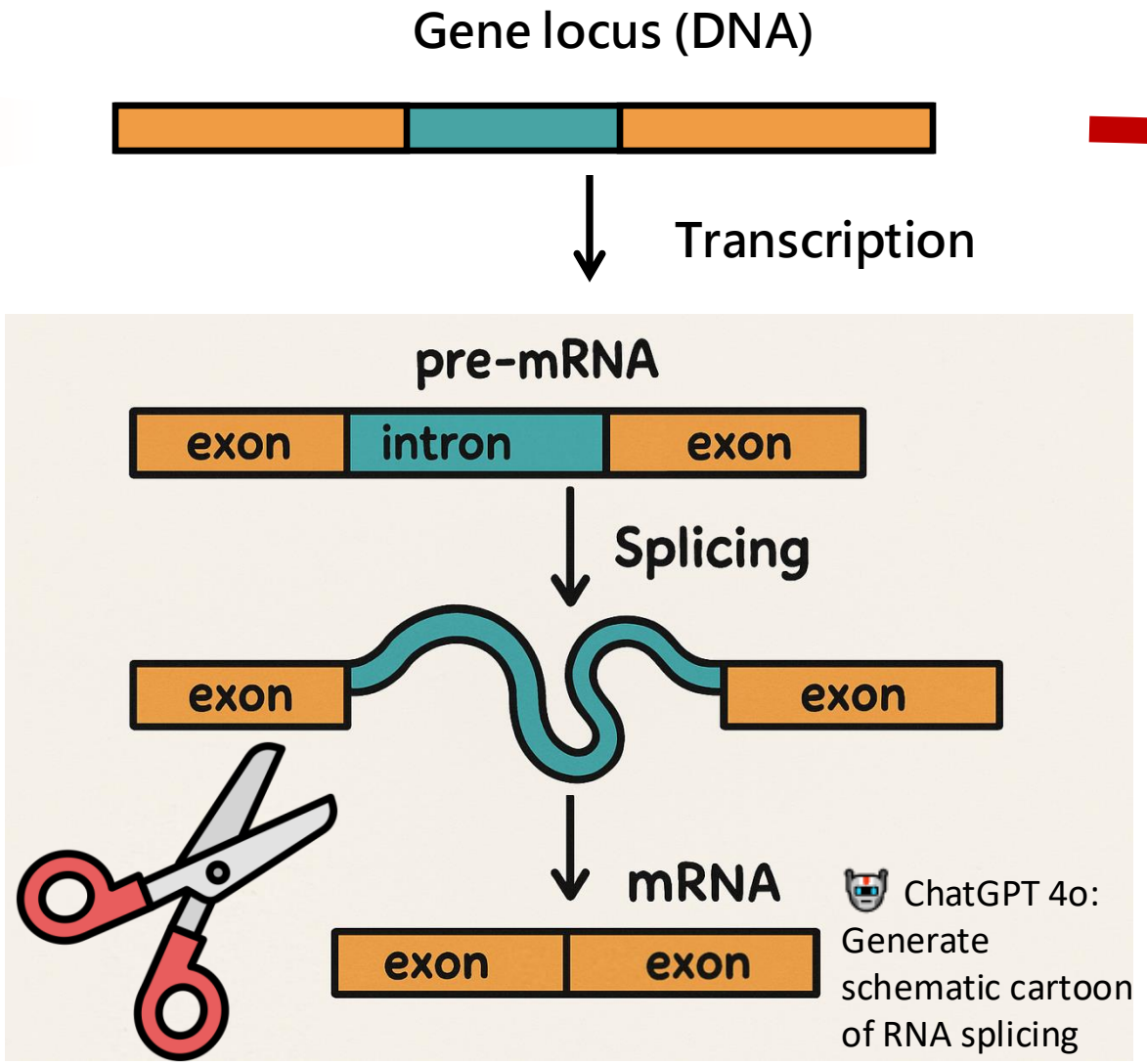
OpenSpliceAI

eLife

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





Model

SPLAM

OpenSpliceAI

Where are the splice sites?

OpenSpliceAI: Splice site prediction

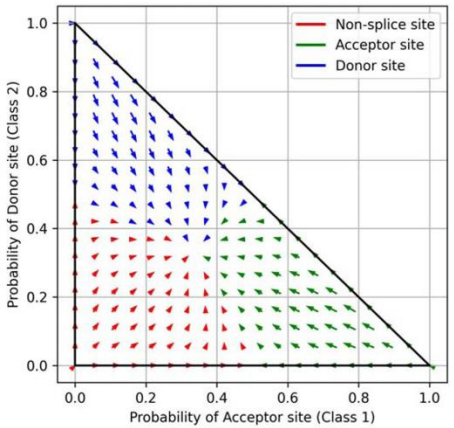
Chao, K. H., Mao, A., Liu, A., Salzberg, S. L., & Pertea, M. (2025). OpenSpliceAI: An efficient, modular implementation of SpliceAI enabling easy retraining on non-human species. eLife

X AGACTCAGCCCCCGGAGACTTAGTTAGAGGAAGAAAAAGGTAGGACAGAAGAAAAAGGCAGGACATACAAGGTGCTGGCCCAGGGCGG

Y 00000000000000000000002000000010000000020000000000010000000000000000000000000000000000000

Donor: 2 Acceptor: 1 Neither: 0

OpenSpliceAI



~20k protein-coding genes

Raw gene DNA sequence

Gene 1 $L = 33200$

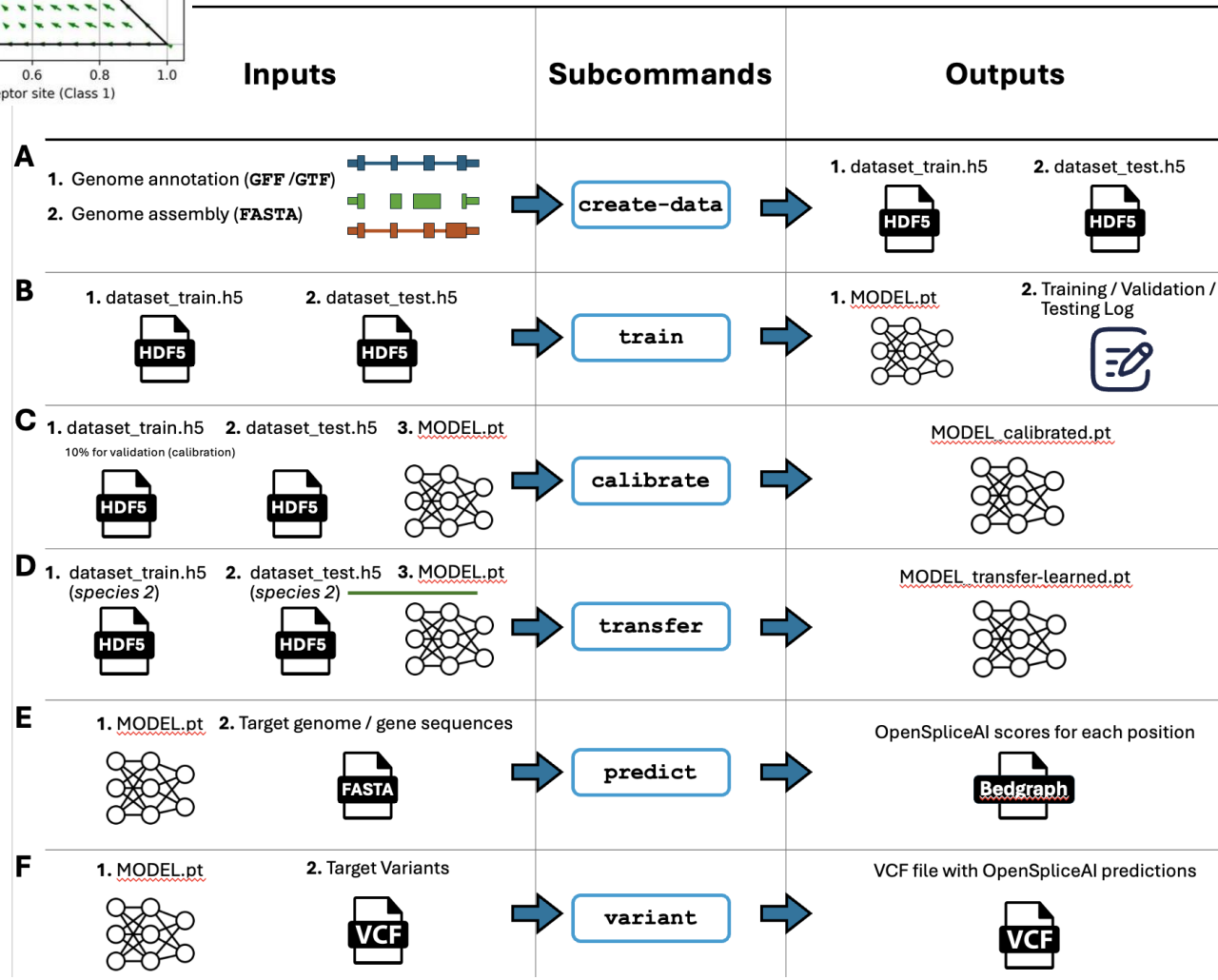
Gene 2 $L = 14600$

...

Gene n $L = 25000$

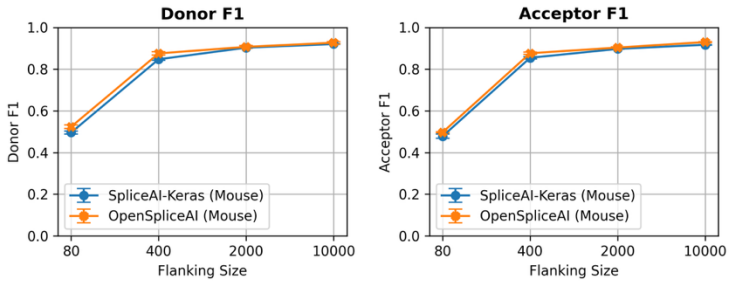
X	Y
[7, 15000, 4]	[7, 5000, 3]
[3, 15000, 4]	[3, 5000, 3]
...	
[5, 15000, 4]	[5, 5000, 3]

$W = 5000$ $F = 10{,}000$

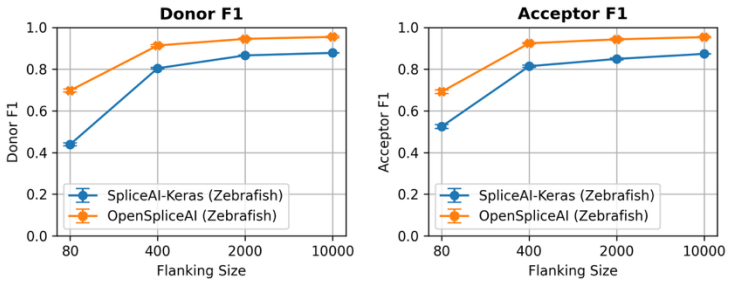


OpenSpliceAI: retrain on different species

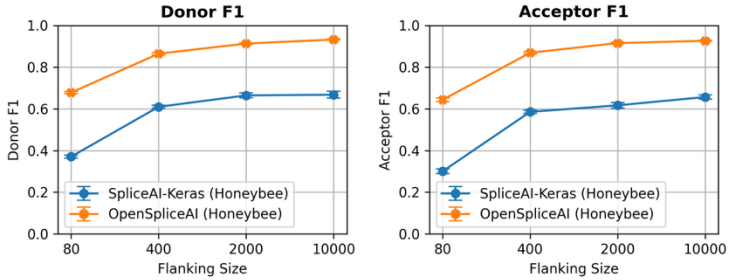
Splice site prediction metrics for Mouse



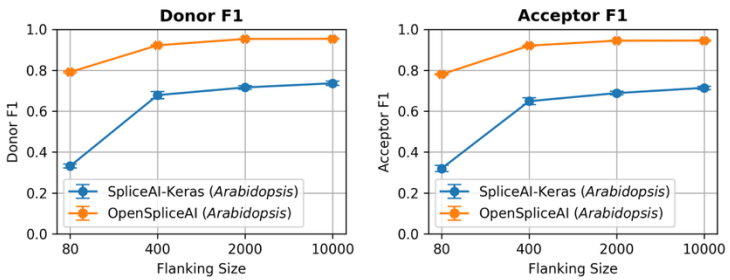
Splice site prediction metrics for Zebrafish



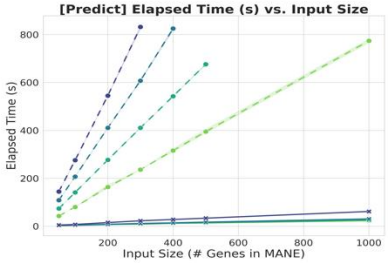
Splice site prediction metrics for Honeybee



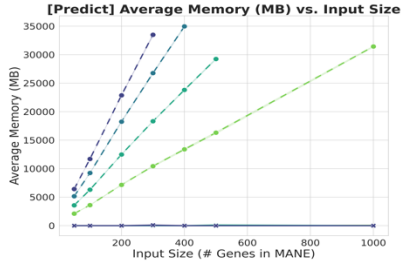
Splice site prediction metrics for *Arabidopsis*



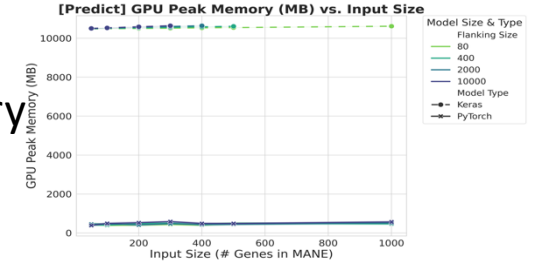
System time



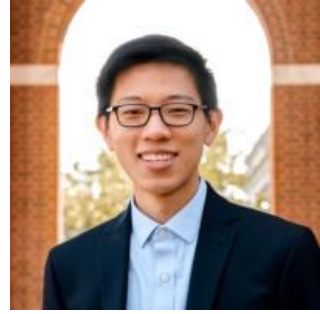
Memory



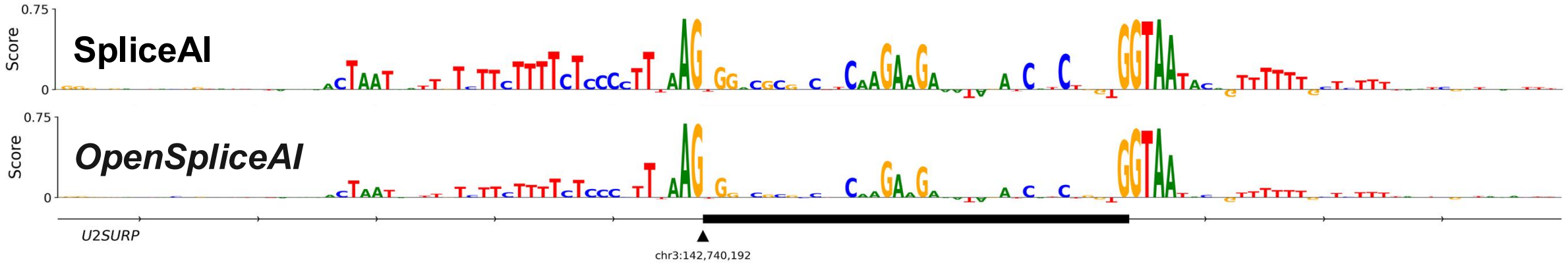
GPU Avg memory



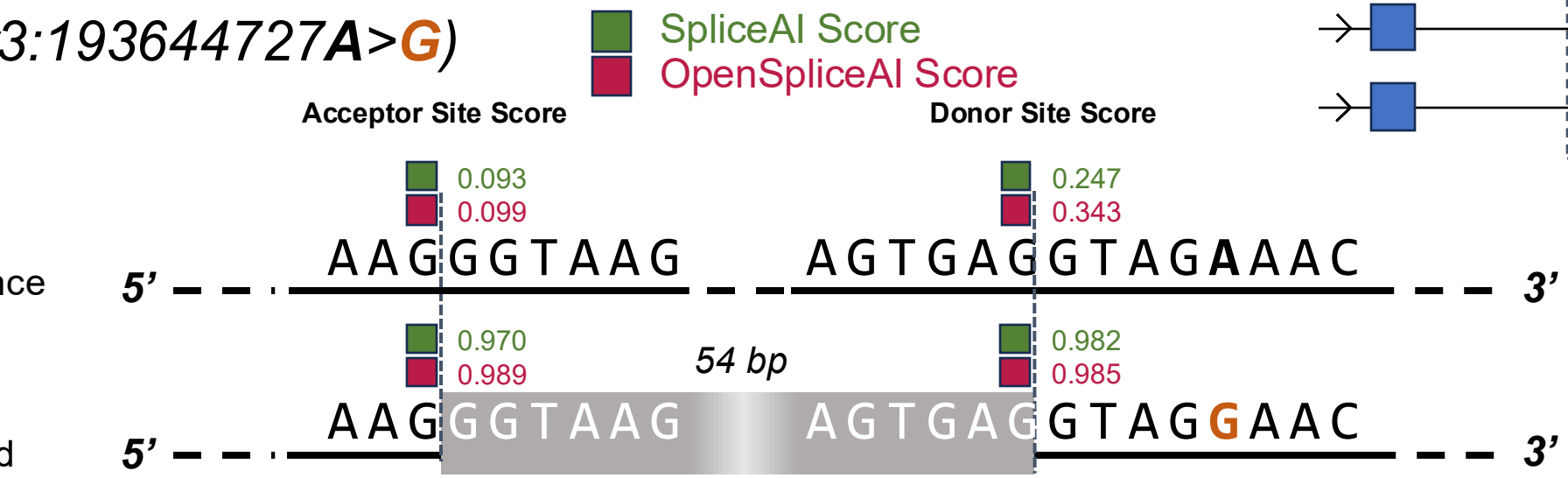
Alan Mao

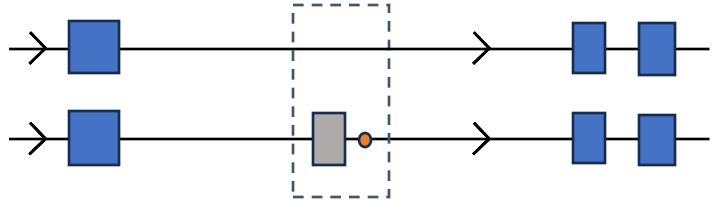


OpenSpliceAI: ISM & variant prediction



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



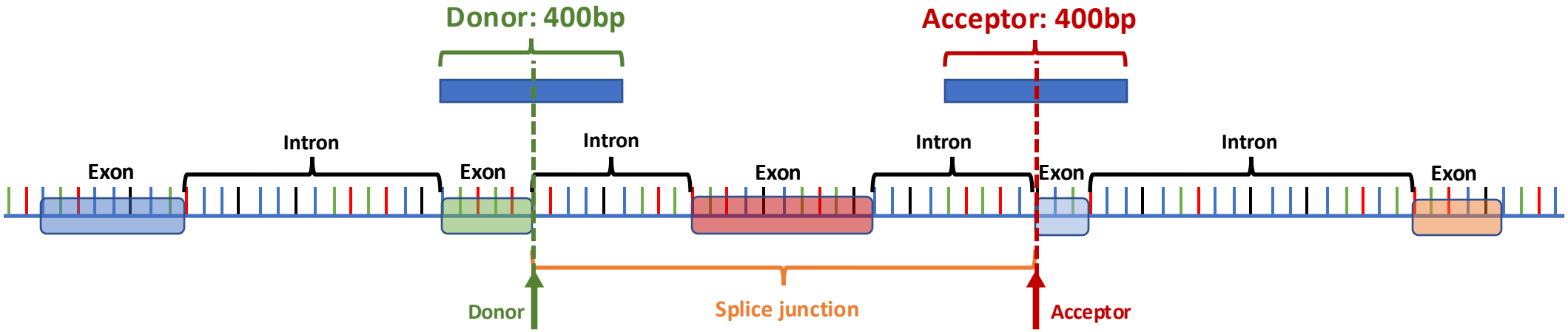


(Qian et al., 2021)

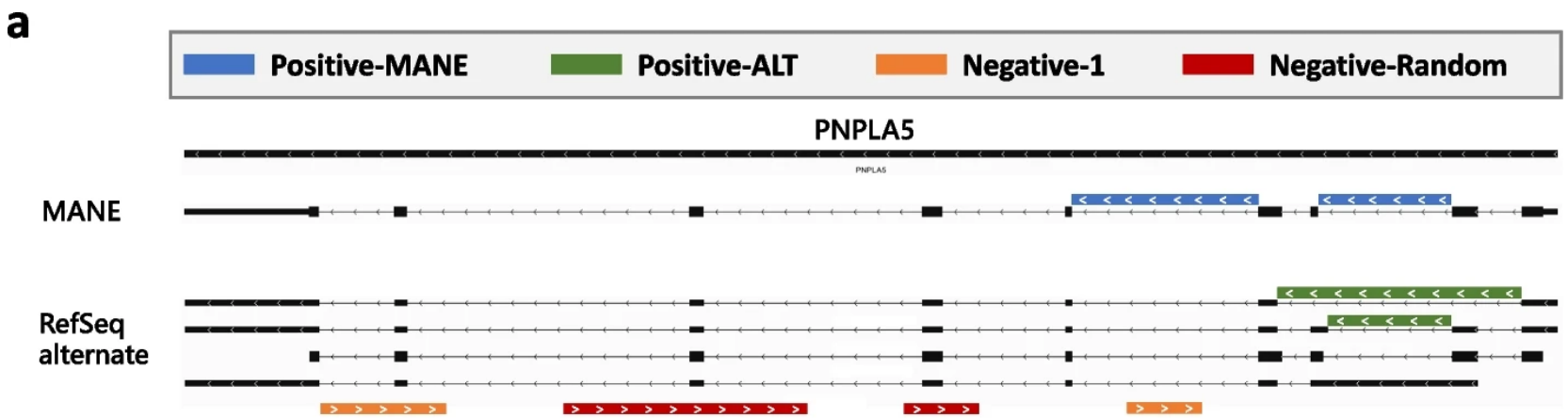
SPLAM: predicting splice sites in a pair

Genome Biology

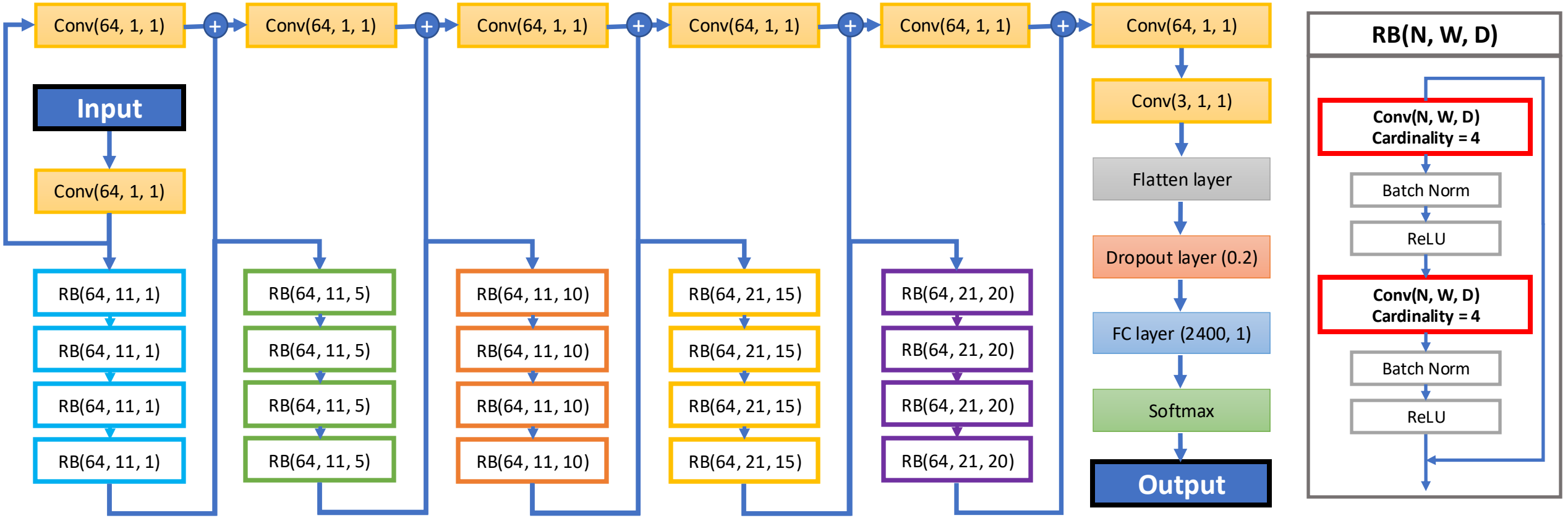
Chao, K. H., Mao, A., Salzberg, S. L., & Pertea, M. (2024). Splam: a deep-learning-based splice site predictor that improves spliced alignments. Genome biology, 25(1), 243.



Data

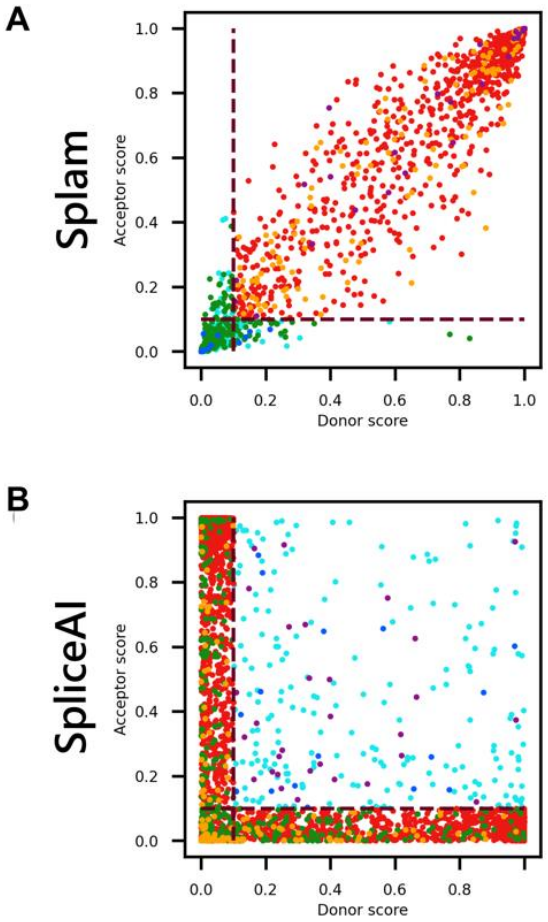


SPLAM : Splam Model Architecture

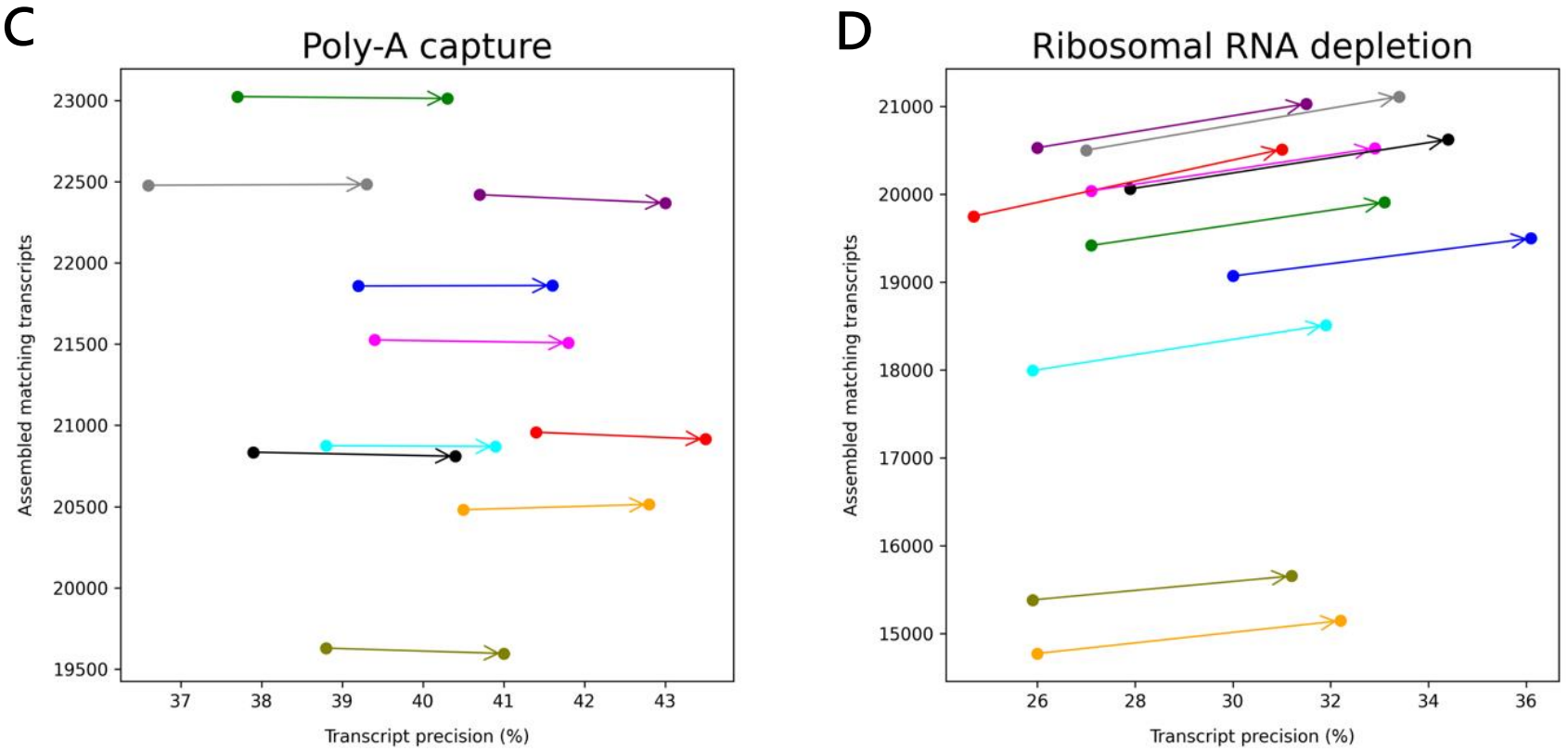


SPLAM: improving transcriptome assembly

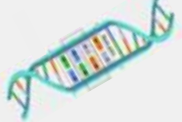
Score stability

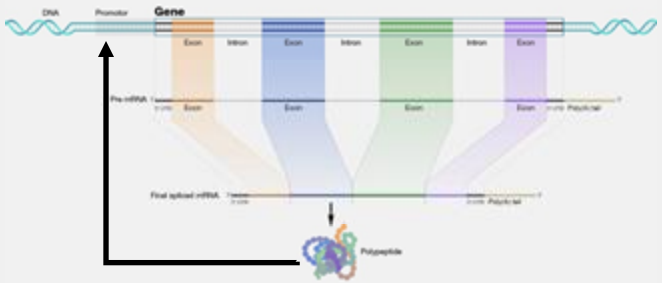


Transcriptome assembly improvement



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

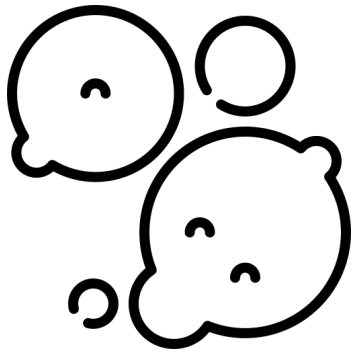
Chao, K. H., Magzoub, M., Stoops, E., Hackett, S., Linder J., * and Kelley, D. R., (manuscript in preparation). Predicting dynamic expression patterns in budding yeast with a fun-gal DNA language model

Manuscript in preparation.
Expected to be on bioRxiv soon

Questions we answer in this study:

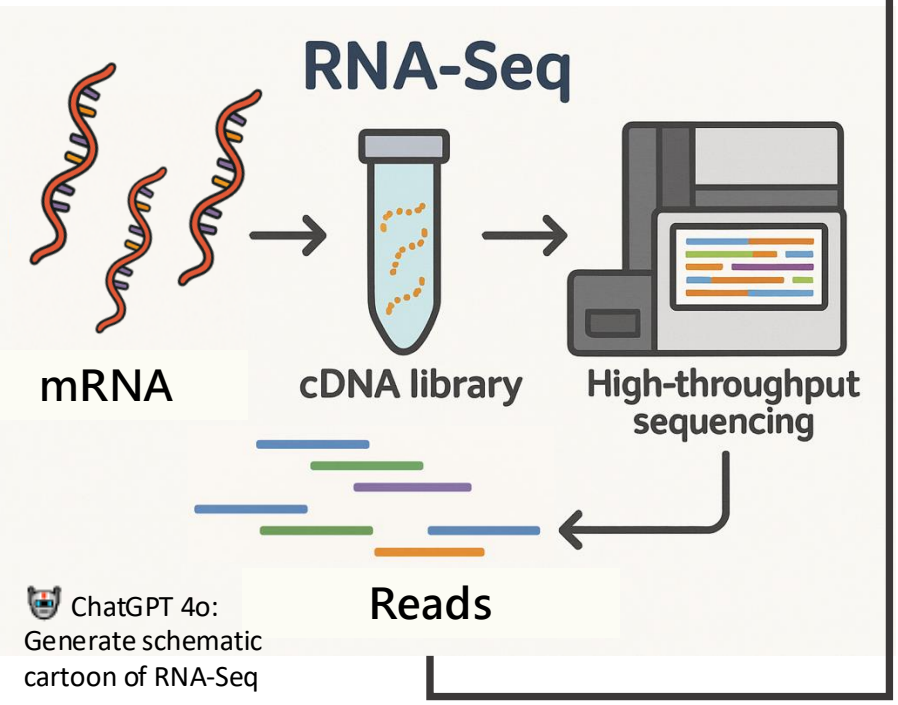
Main Goal

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



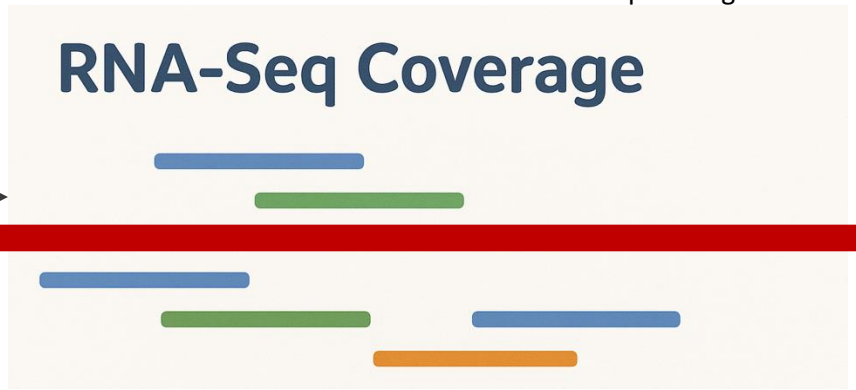
Yeast RNA-Seq Prediction

Gene locus (DNA)



ChatGPT 4o:
Generate schematic cartoon of RNA-Seq

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



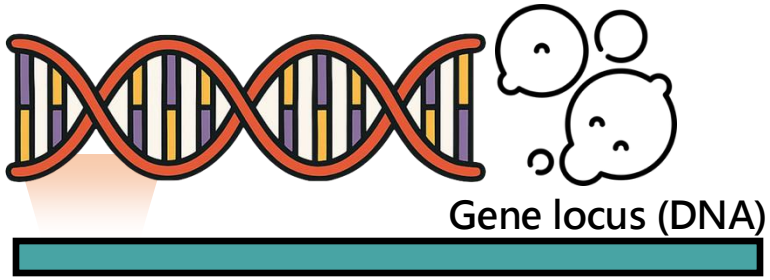
Reads alignment

Coverage pileup

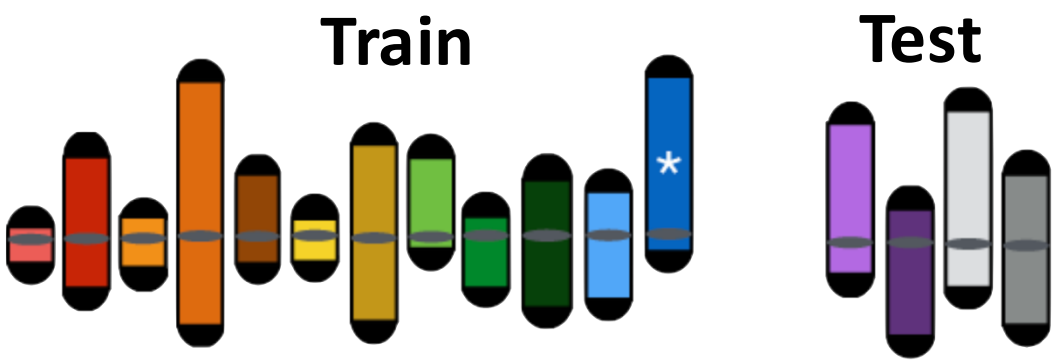
Gene expression model

RNA-Seq expression?

Yeast RNA-Seq Prediction

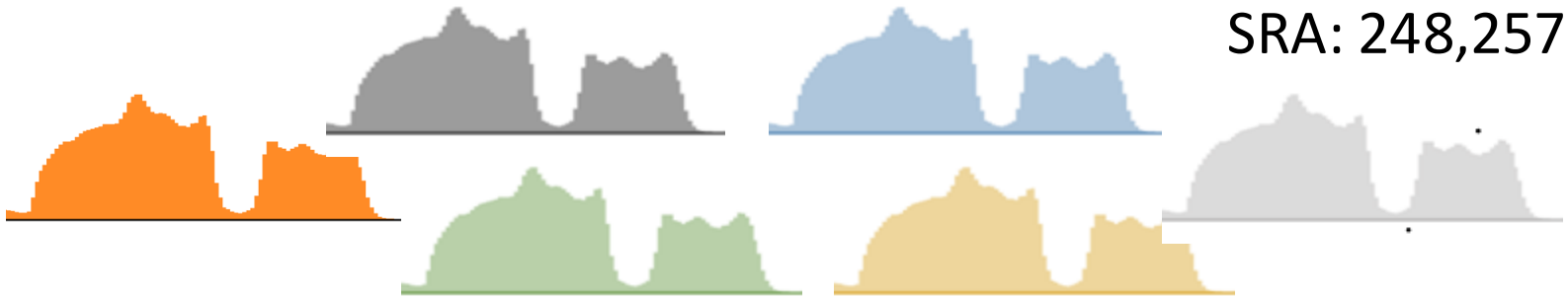


ChatGPT 4o:
Generate schematic cartoon
of RNA-Seq coverage



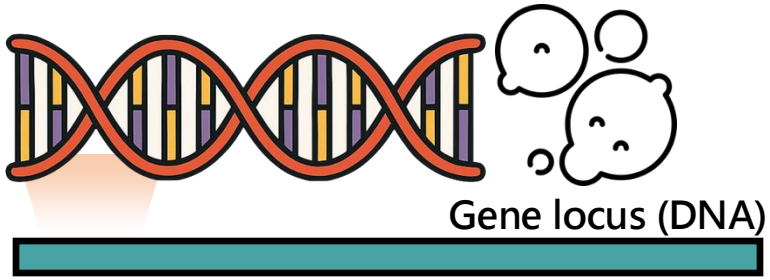
12 million base pairs

Genome size is too small, leading to overfitting.

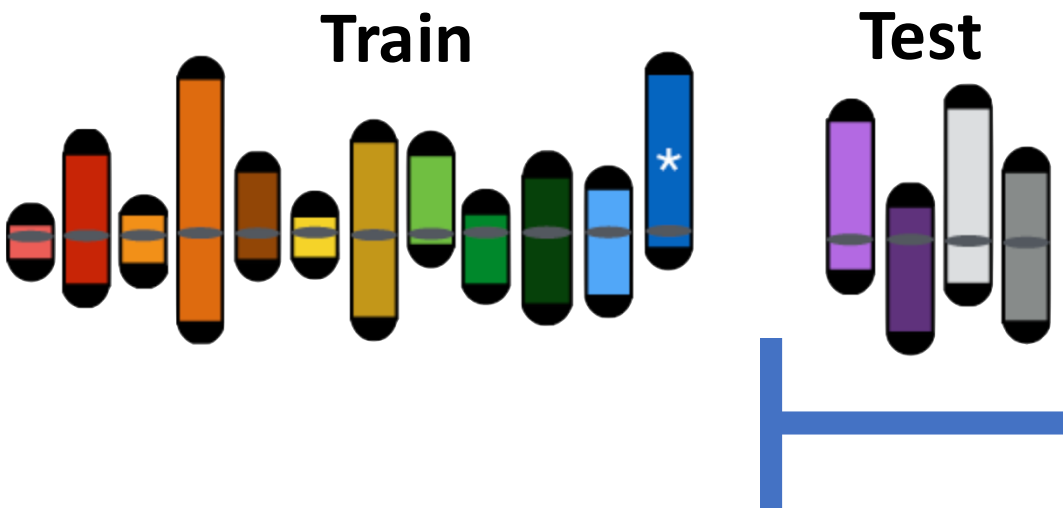


SRA: 248,257 search entries

Yeast RNA-Seq Prediction



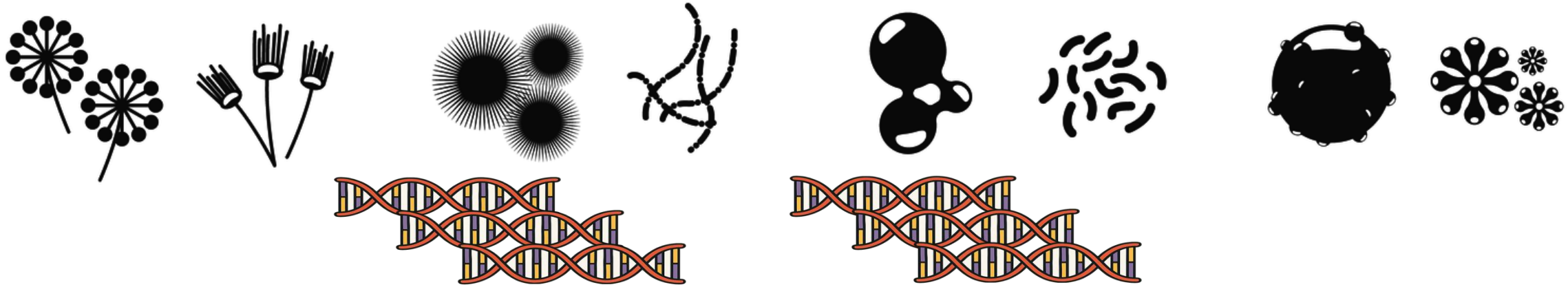
ChatGPT 4o:
Generate schematic cartoon
of RNA-Seq coverage



12 million base pairs

Fungal LM

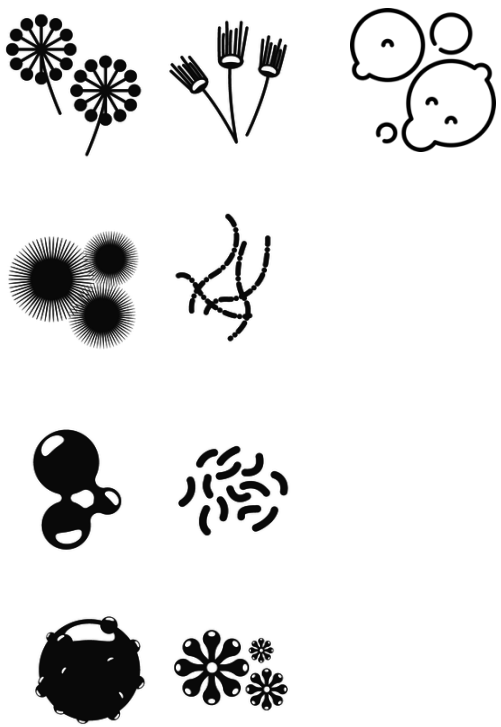
Ensembl fungi: 1500 genomes

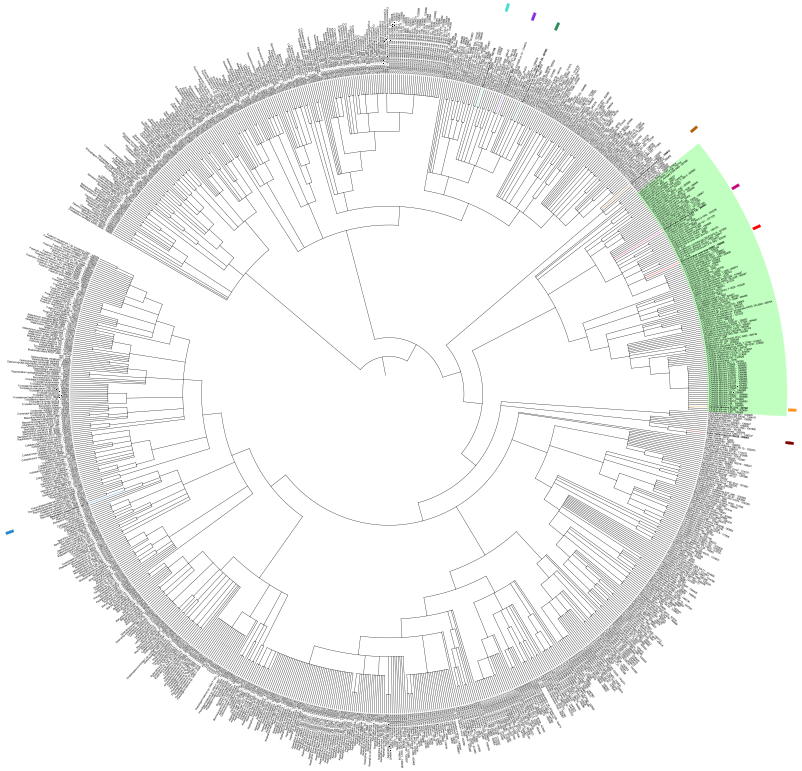


Yeast RNA-Seq Prediction

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

Gene locus (DNA)





Fungal LM

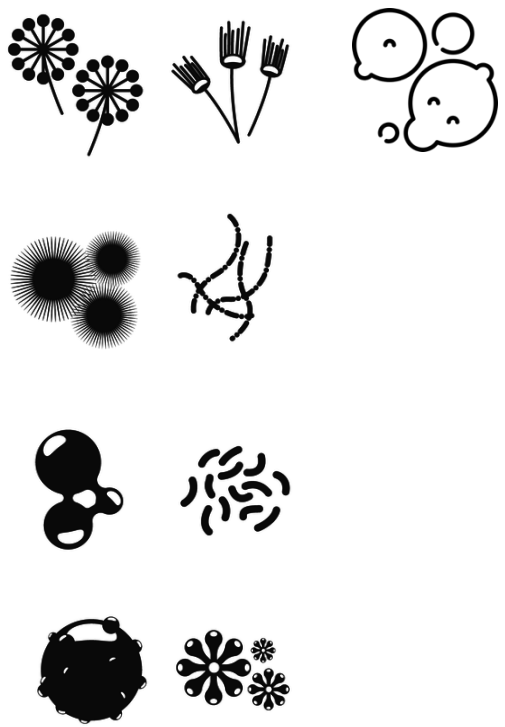
Gene expression model

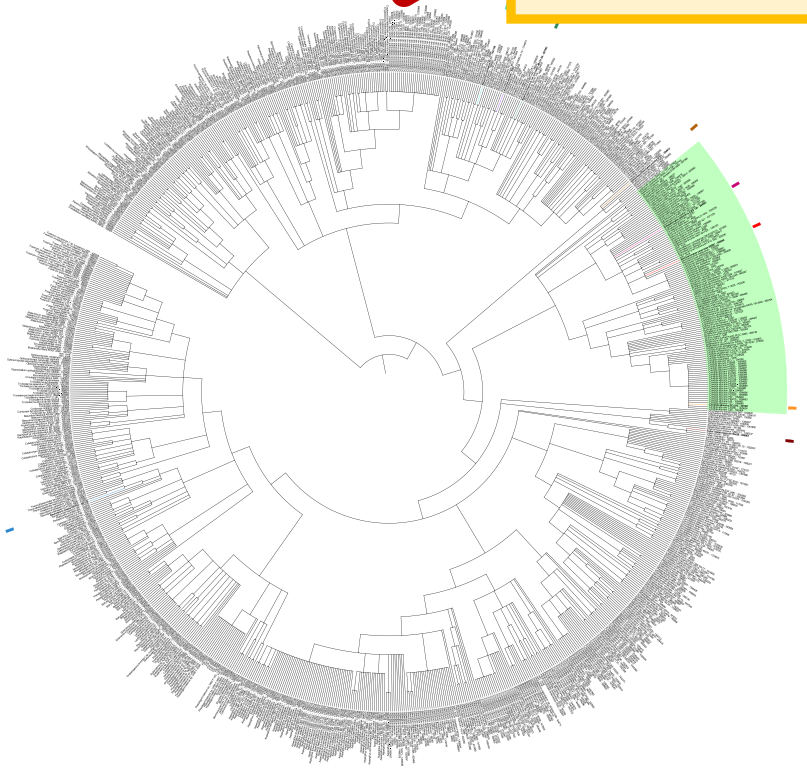
RNA-Seq expression?

Yeast RNA-Seq Prediction

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

Gene locus (DNA)



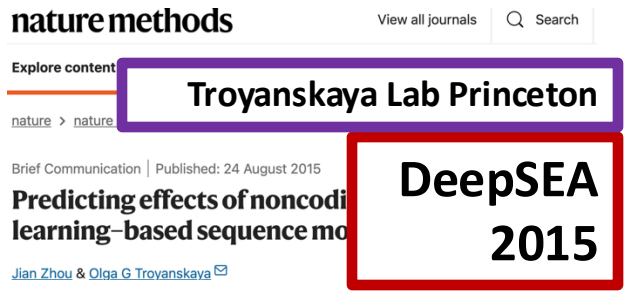


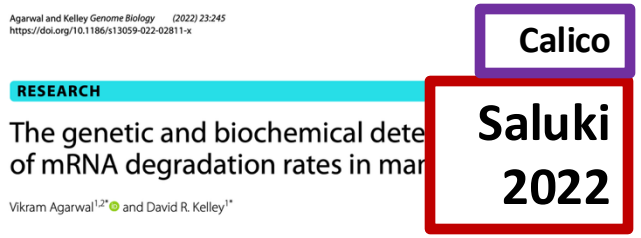
Fungal LM

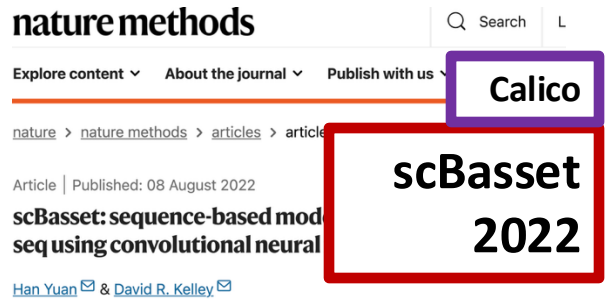
Model
Fungal Model

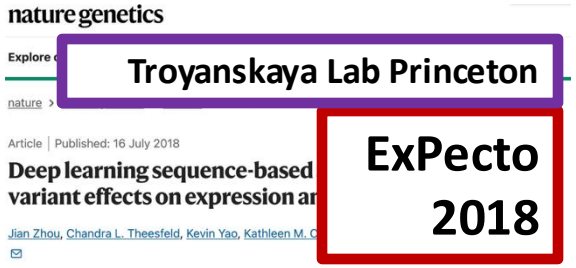
RNA-Seq expression?

Deep learning-based DNA sequence model





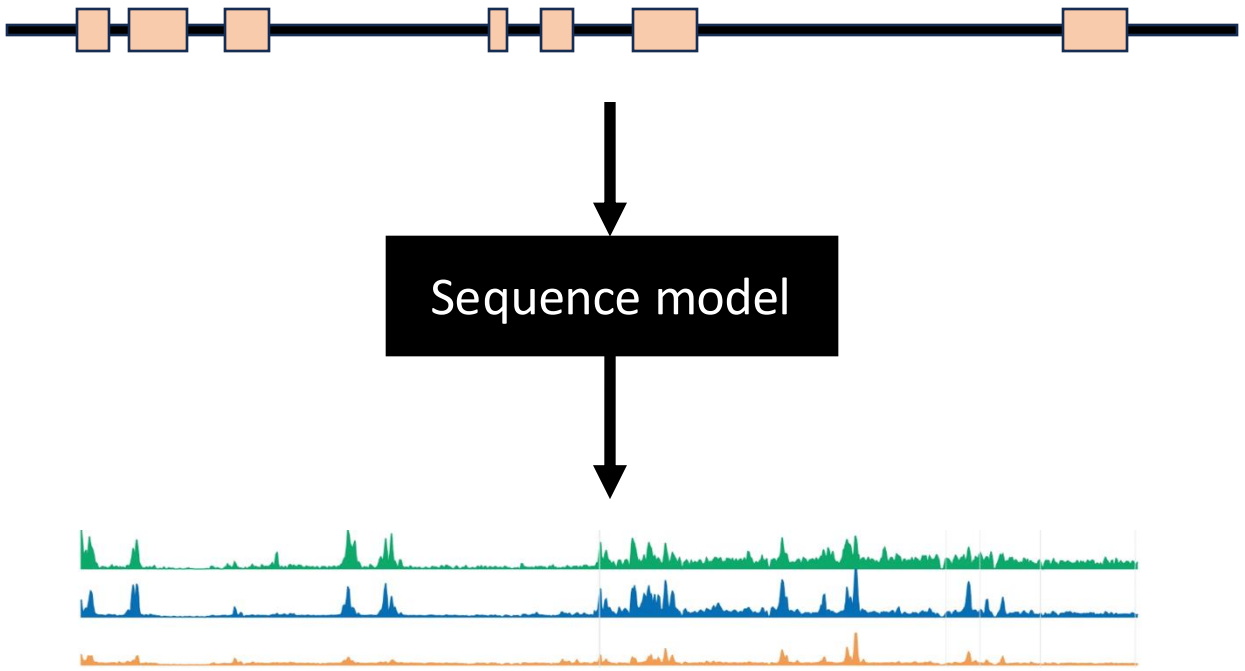




Supervised learning

Input: DNA sequences

Output: Genomics tracks



Protein Language model (PLM)

RESEARCH ARTICLE | BIOLOGICAL SCIENCES

Biological structure and functi
from scaling un
million protein s

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

Facebook

First LM attempt
PNAS 2020

Transformer protein language
unsup

Roshan Ra
Rives

Facebook

Transformer protein LM
ICLR 2021

arXiv:2006.15222 (cs)

[Submitted on 26 Jun 2020 (v1), last revised 28 Mar 202

BERTology Meets Biology:
Attention in Protein Langu

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

Saleforce + UIUC

BERTology
ICLR 2021

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 | Open access | Published: 27 July 2022

ProtGPT2 is a deep unsupervised l
for protein design

Noelia Ferruz, Steffen Schmid

Nature Communications 13, Arti

Höcker Lab

ProtGPT2
Nat Commun 2022

Article | Published: 03 October

Single-sequence prot
using a language m

Ratul Chowdhury, Nazim Bo
Anant Kharkar, Koushik Roy
Zhang, George M. Church,

Harvard + Columbia

AminoBERT
Nat Biotechnol 2022

Article | Published: 26 January 2023

Large language models generate fu
protein sequences across diverse f

Ali Madani, Ben Krause, E
Mohr, James M. Holton, Jos
Sun, Richard Socher, James

Saleforce

ProGen
Nat Biotechnol 2023

ARTICLE · Volume 14, Issue 11,

ProGen2: Exploring th
language models

Erik Nijkamp [1,5] · Jeffrey A. Ruffol
· Ali Madani [1,3,6]

Saleforce + Profluent

ProGen2
Cell System 2023

RESEARCH ARTICLE | STRUCTURE PREDICTION

Evolutionary-scale predictio
tein structure with a langua

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

Meta

ESMFold
Science 2023

Primer | Published: 15 February 2024

Designing protei

Jeffrey A. Ruffolo & Ali M

Nature Biotechnology 42

Profluent

PLM
Nat Biotechnol 2024

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

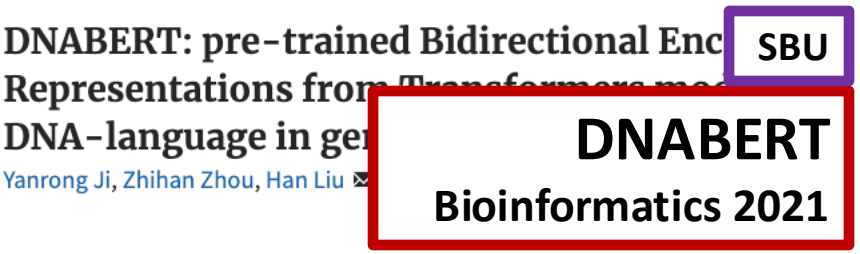
Liuzhenghao Lv, Zongying Lin, Hao Li, Yuyang Liu, Ji
Yonghong Tian

Peking Uni

ProLLaMA
arxiv, 2024

~ 6000 citations

DNA Language model


DNABERT: pre-trained Bidirectional Enc
Representations from
DNA-language in ge
Yanrong Ji, Zhihan Zhou, Han Liu

SBU

DNABERT
Bioinformatics 2021


DNABERT-2: Efficient Foundation M
Benchmark For Multi-Specie
Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik

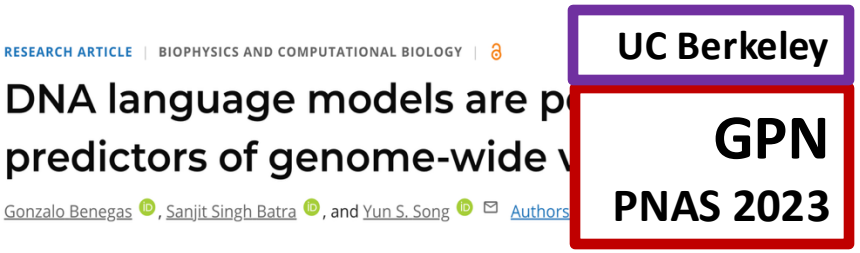
SBU + NU

DNABERT-2
ICLR 2024


Sequenc
molecul
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

Stanford + Arc Inst + TogetherAI

Evo
bioRxiv 2024


RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY
DNA language models are p
predictors of genome-wide v
Gonzalo Benegas, Sanjit Singh Batra, and Yun S. Song

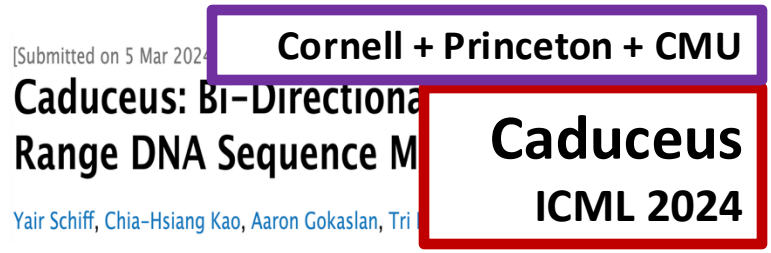
UC Berkeley

GPN
PNAS 2023


Article | Open access | Published: 23 July 2024
DNA language model
context in the human
Melissa Sanabria, Jonas Hirsch, Pierre

TUD

GROVER
Nat Mach Intell 2024


[Submitted on 5 Mar 2024
Caduceus: Bi-Directiona
Range DNA Sequence M
Yair Schiff, Chia-Hsiang Kao, Aaron Gokaslan, Tri

Cornell + Princeton + CMU

Caduceus
ICML 2024


The Nucleotide Transf
Foundation Models for
Hugo Dalla
Adam Henn
Bernardo P
Marie Lope

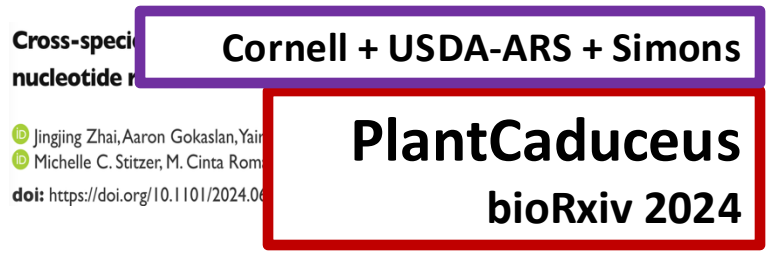
InstaDeep + Nvidia + TUM

Nucleotide Transformer
bioRxiv 2023


Article | Open access | Published: 03 April 2024
Genomic language model predi
regulation and functior
Yunha Hwang, Andre L. Cornman, Elizabe
R. Girguis

Harvard + MIT

Genomic LM
Nat Commun 2024


Cross-speci
nucleotide
Jingjing Zhai, Aaron Gokaslan, Yair
Michelle C. Stitzer, M. Cinta Rom
doi: https://doi.org/10.1101/2024.06

Cornell + USDA-ARS + Simons

PlantCaduceus
bioRxiv 2024


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

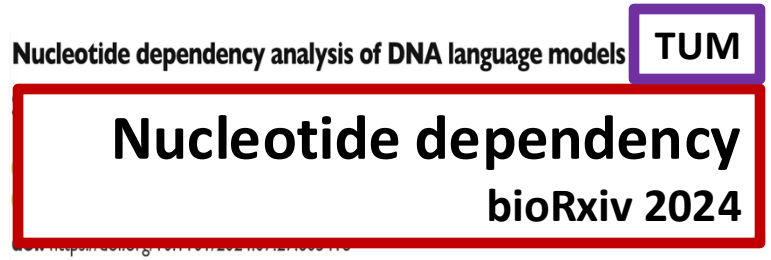
Stanford

HyenaDNA
NeurIPS 2023


Research | Open access | Published: 02 April 2024
Species-aware DNA language models capture regulatory
elements a
Alexander Karoll
Julien Gagneur
Genome Biology 25, Article number: 83 (2024) | Cite this article

TUM

Species-aware DNA LM
Genom Biol 2024


Nucleotide dependency analysis of DNA language models

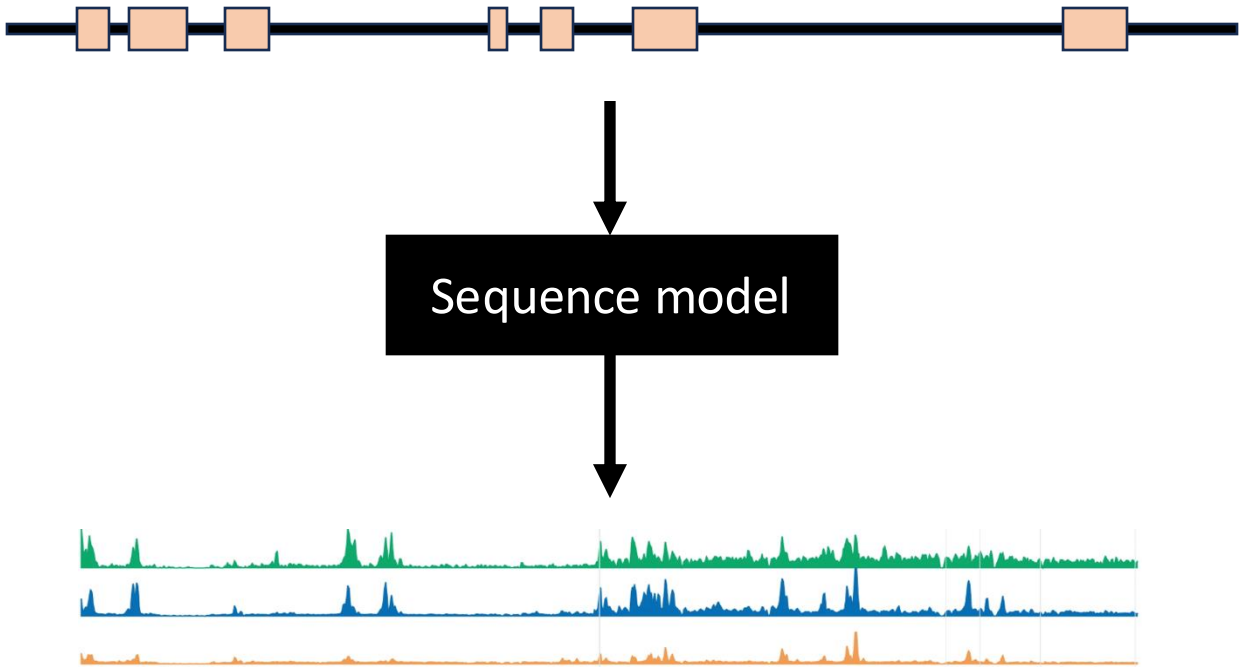
TUM

Nucleotide dependency
bioRxiv 2024

Supervised learning

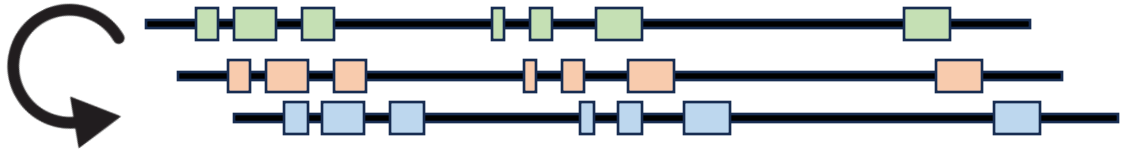
Input: DNA sequences

Output: Genomics tracks



Stage 1

Self-supervised learning



Language model (LM) / Foundation model

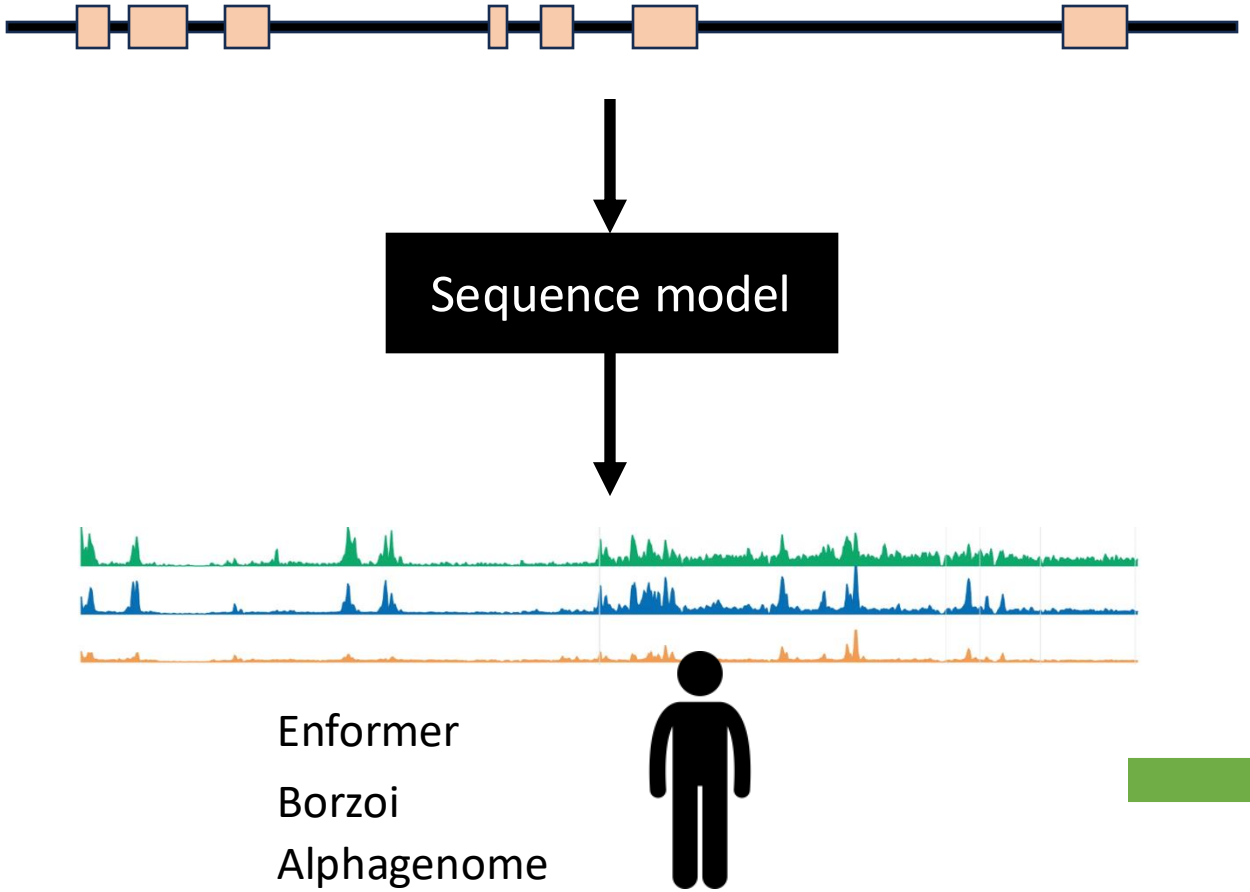
Stage 2

Fine-tuning LM

Enformer

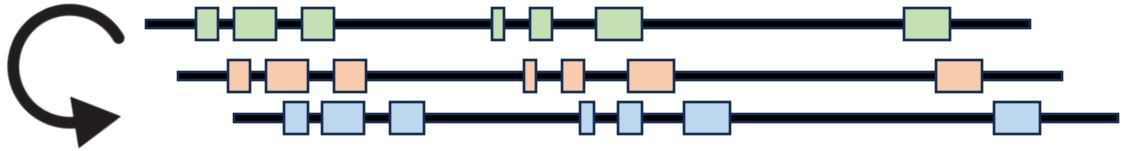
Borzoi

Supervised learning



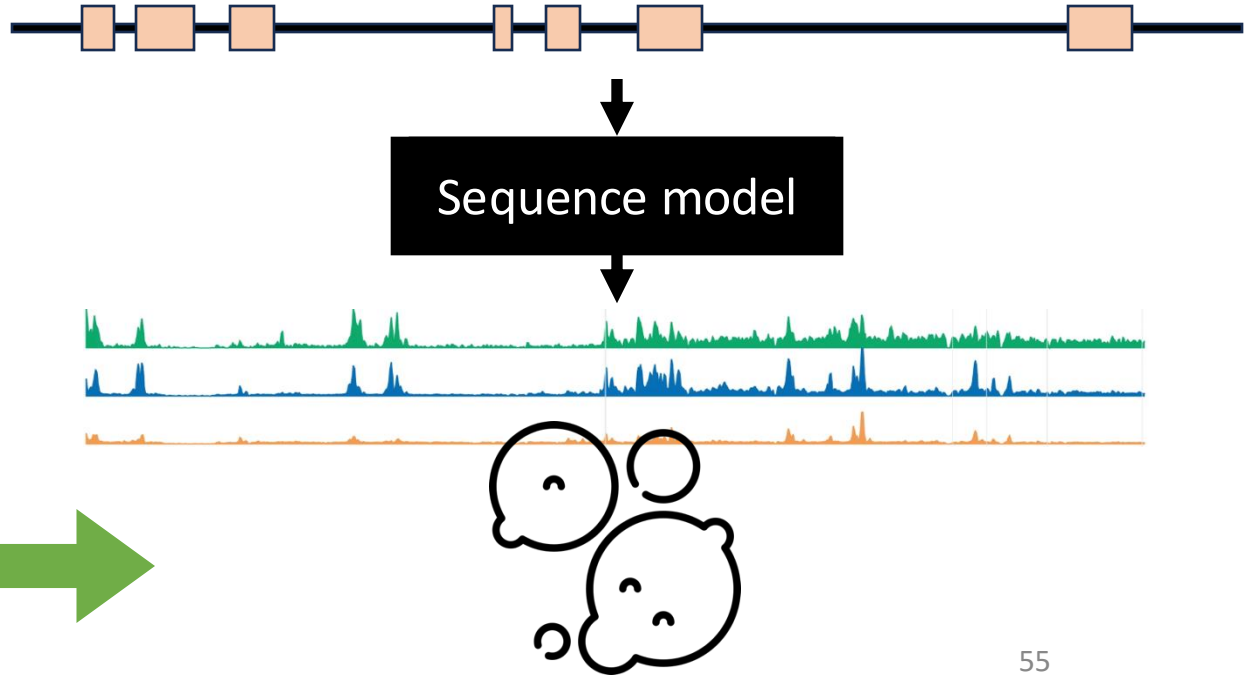
Stage 1

Self-supervised learning



Stage 2

Fine-tuning LM



Shorkie

16384bp

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

Encoding: (4 + 1 + species_num)

16384 * 4 Masked language modeling loss

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

Reverse complementary

1bp res ... ... 1bp res

16bp res 16bp res

32bp res 32bp res

64bp res 64bp res

128bp res 128bp res

Transformer Blocks (8x)

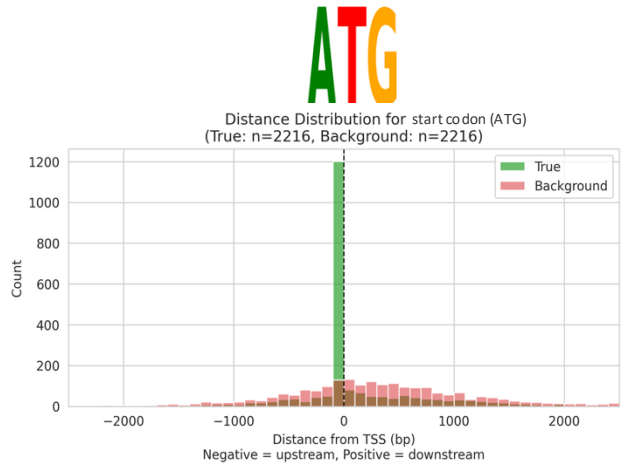
Borzoi

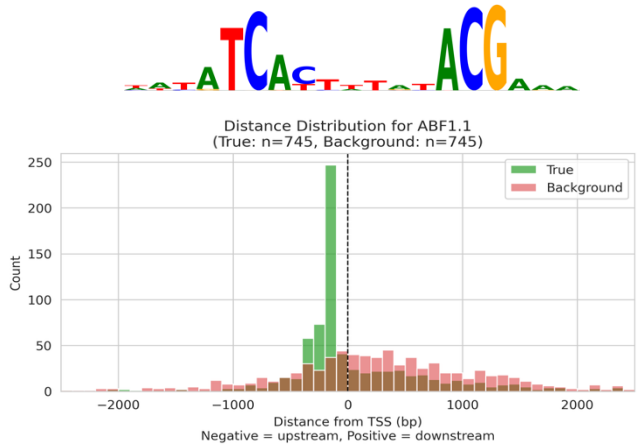
Linder, J. et al. (2025). Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. Nature Genetics, 1-13.

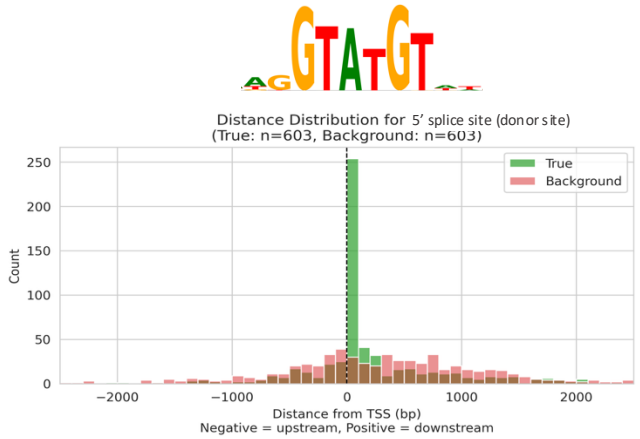
RAP1  FHL1 

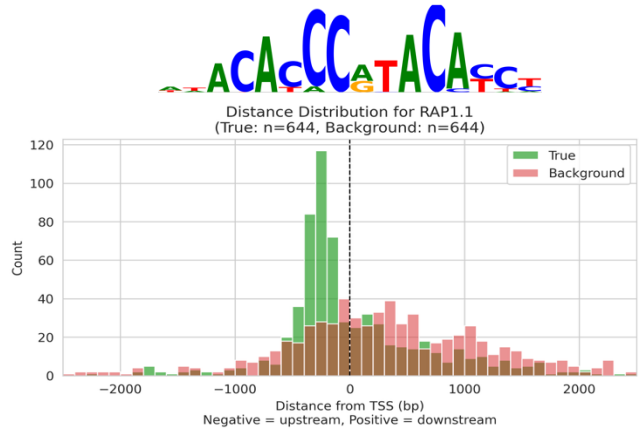
LM predicts motifs in RP gene Promoter regions

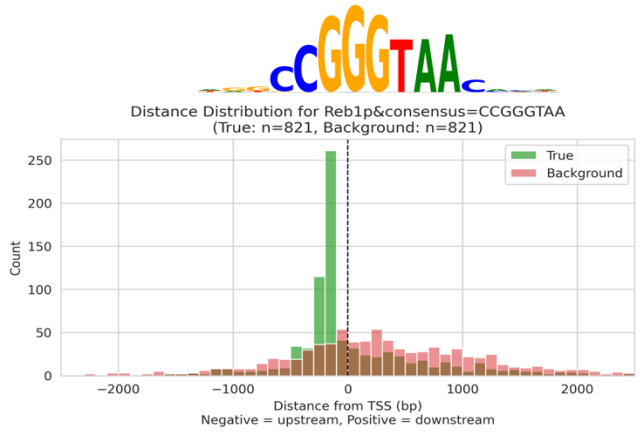
These motifs are enriched in promoter regions

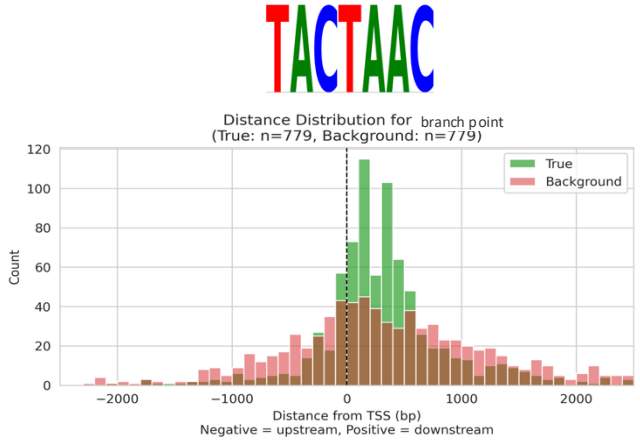


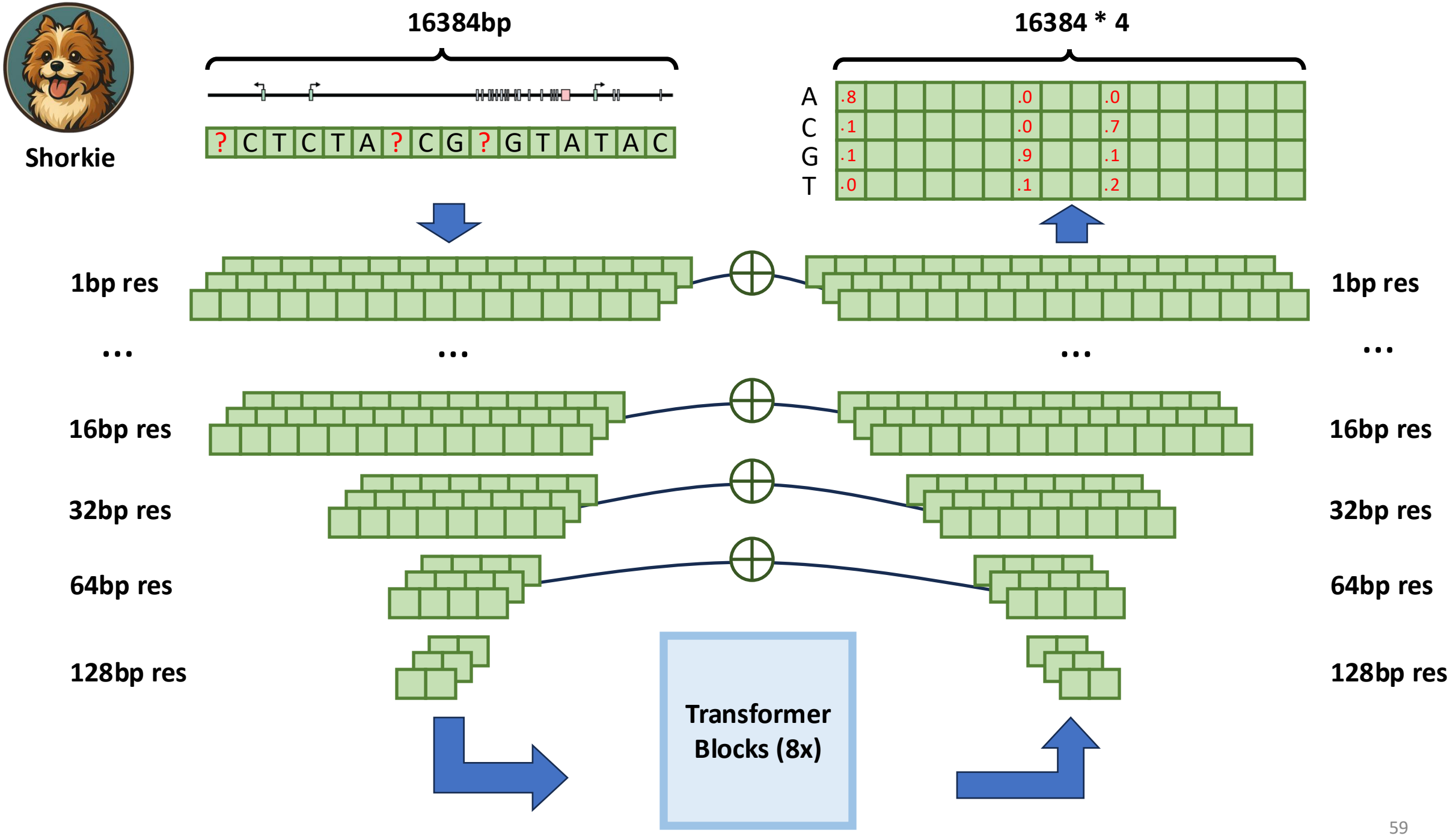




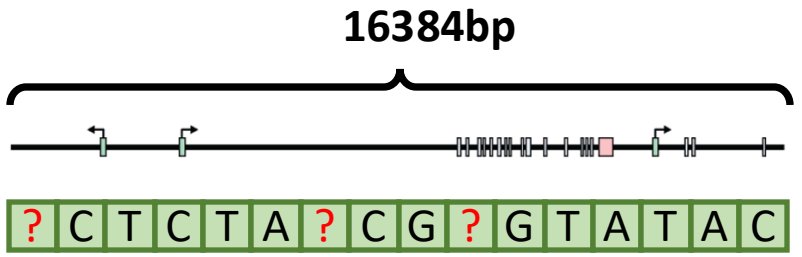


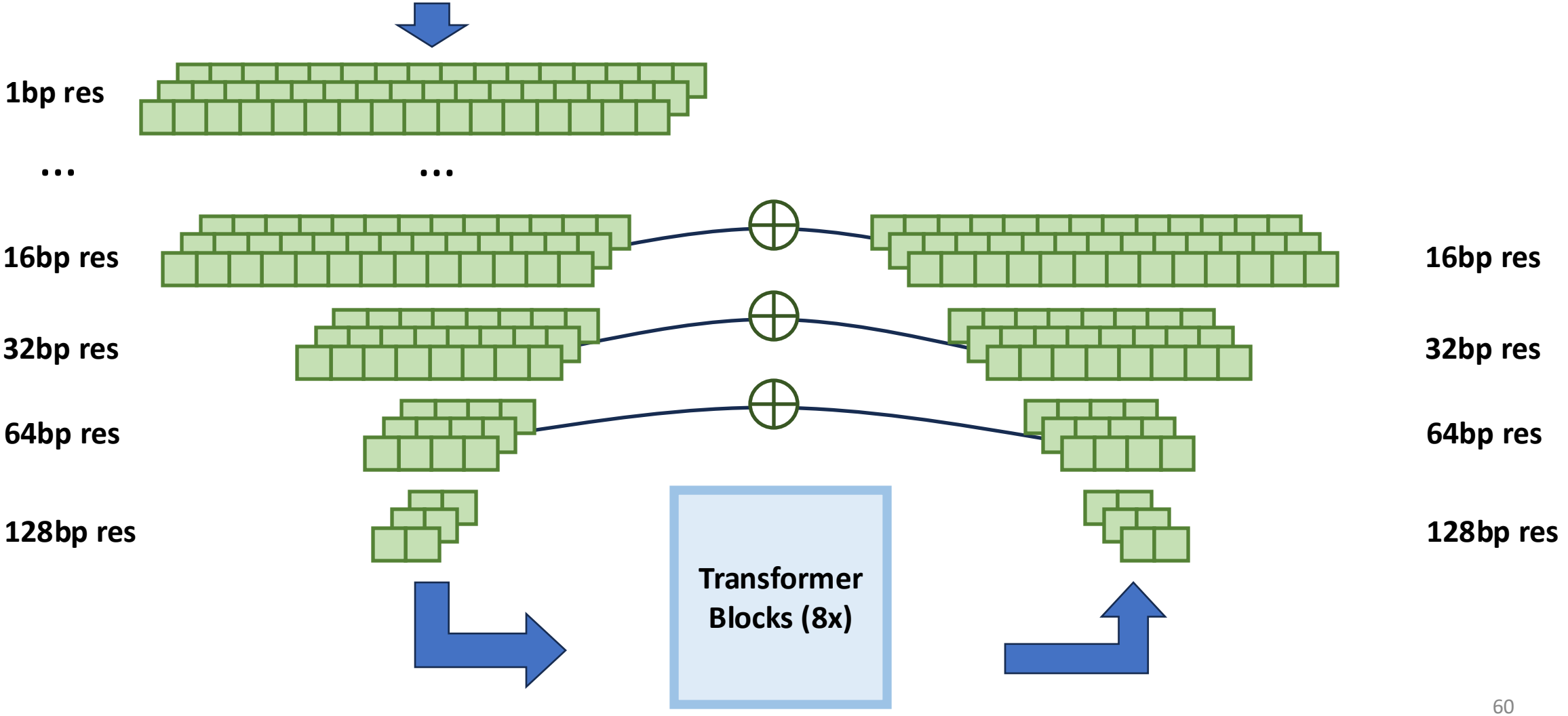





Shorkie
16384bp
? C T C T A ? C G ? G T A T A C
16384 * 4
A .8 .0 .0
C .1 .0 .7
G .1 .9 .1
T .0 .1 .2
1bp res
16bp res
32bp res
64bp res
128bp res
Transformer Blocks (8x)

Shorkie


16384bp
? C T C T A ? C G ? G T A T A C

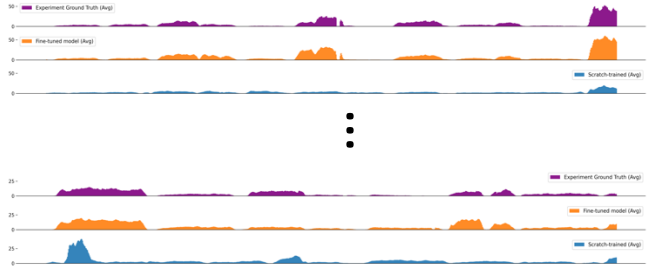

1bp res
...
...
16bp res
32bp res
64bp res
128bp res
Transformer Blocks (8x)
16bp res
32bp res
64bp res
128bp res

Shorkie

16384bp

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

Coverage Tacks (896 * 2488)



CHiP-exo (1128)

Histone marks (20)

RNA-Seq (3054)

1000-strains RNA-Seq (1014)

1bp res

...

...

16bp res

32bp res

64bp res

128bp res

Transformer Blocks (8x)

16bp res

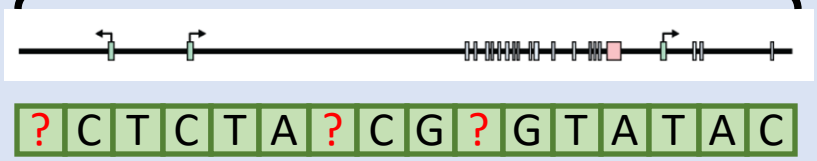
32bp res

64bp res

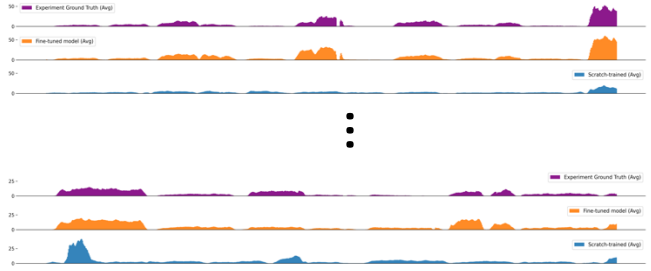
128bp res

Shorkie

16384bp



Coverage Tacks (896 * 2488)



CHiP-exo (1128)

Histone marks (20)

RNA-Seq (3054)

1000-strains RNA-Seq (1014)

1bp res

...

...

16bp res

32bp res

64bp res

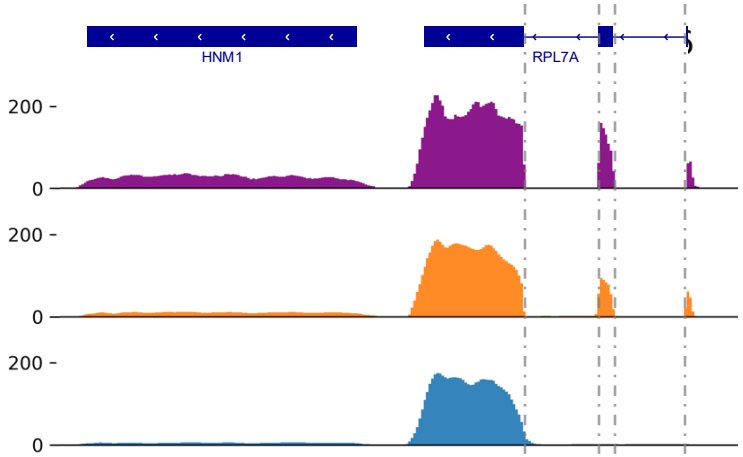
128bp res

Transformer Blocks (8x)

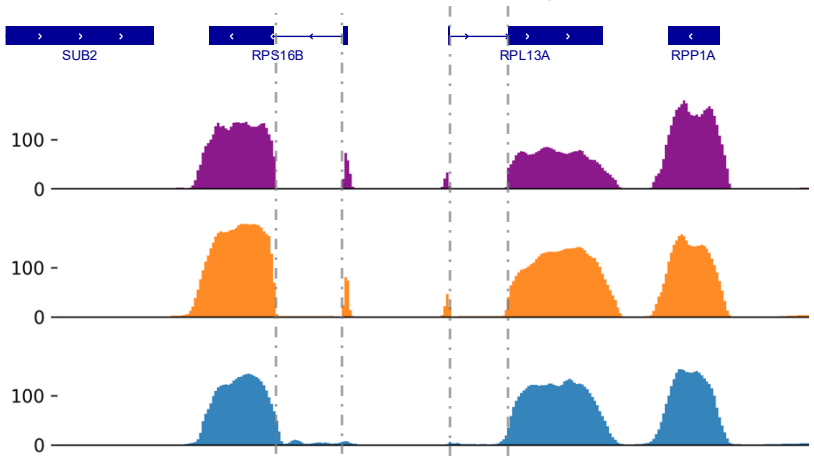
Shorkie vs Random Initialization (Test set)

RNA-Seq tracks

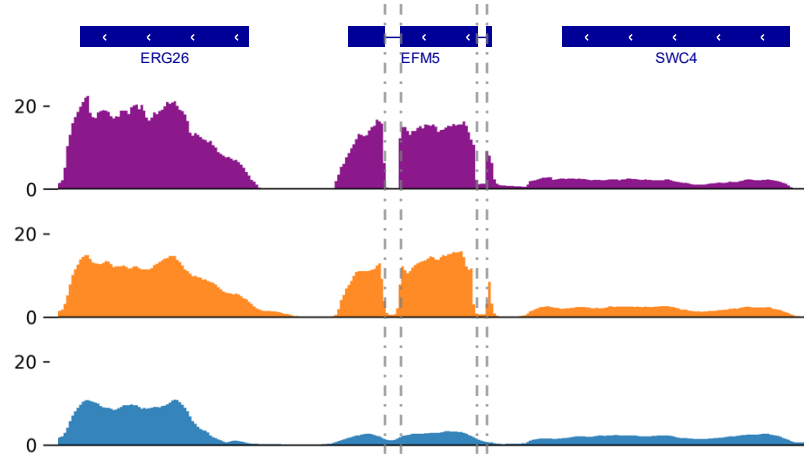
chrVII:362,180-366,023 (RNA-Seq tracks); fold 3



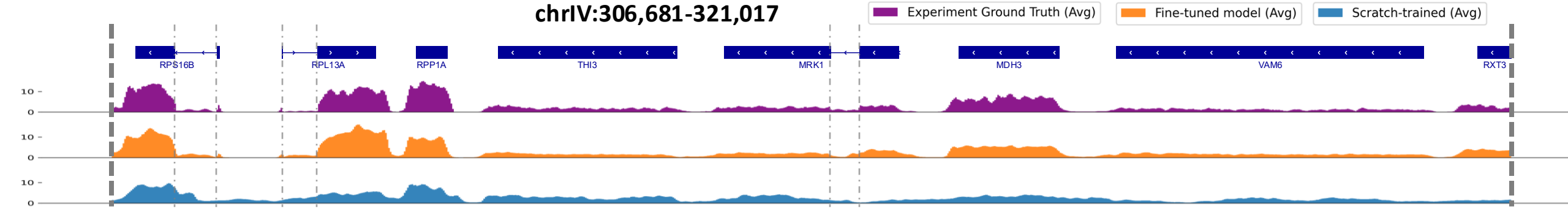
chrIV:305,657-310,505 (RNA-Seq tracks); fold 3



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

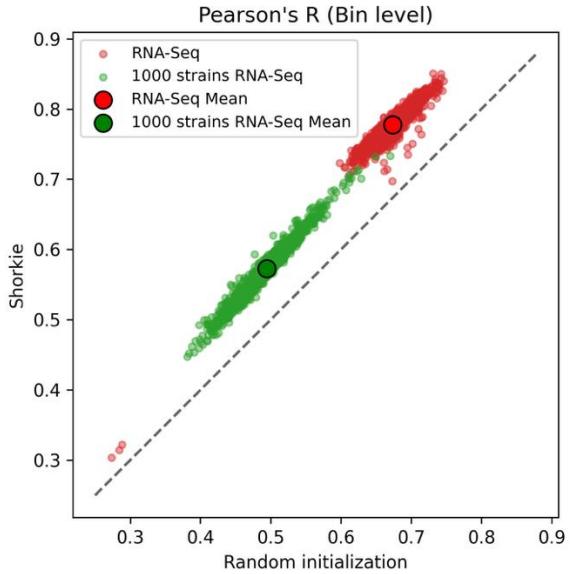


chrIV:306,681-321,017

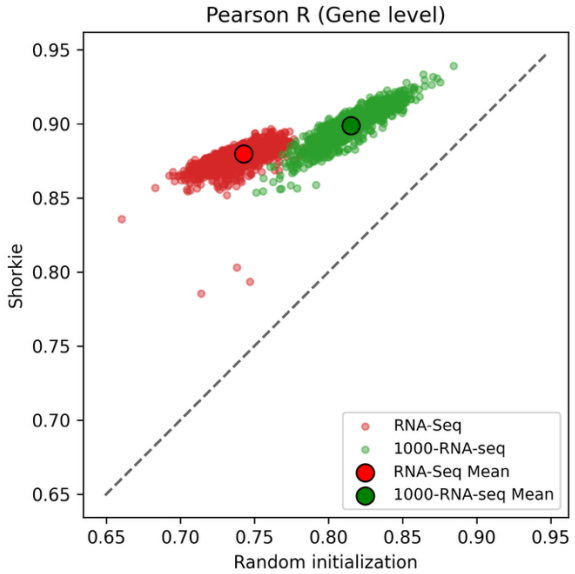


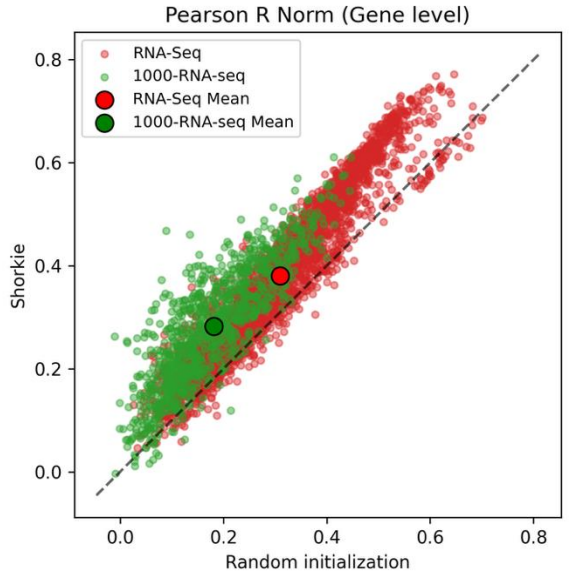
Shorkie vs Random Initialization (Test set)

Bin level
(each dot is a track; all bins)

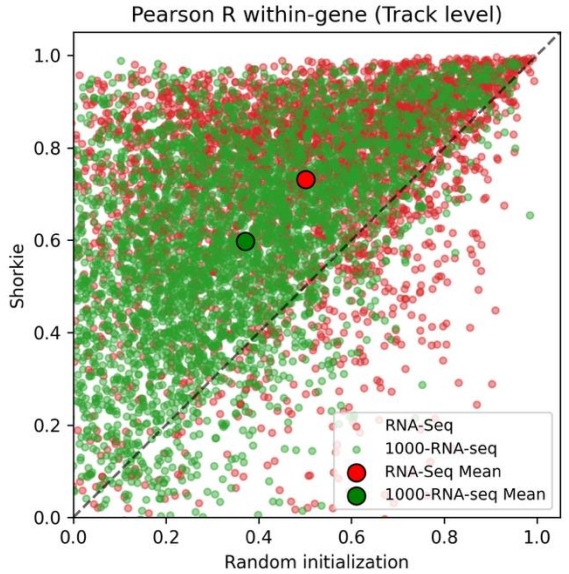


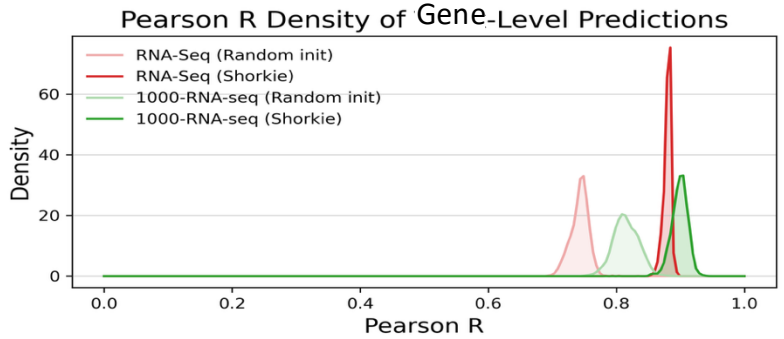
Gene level (each dot is a track; all genes)

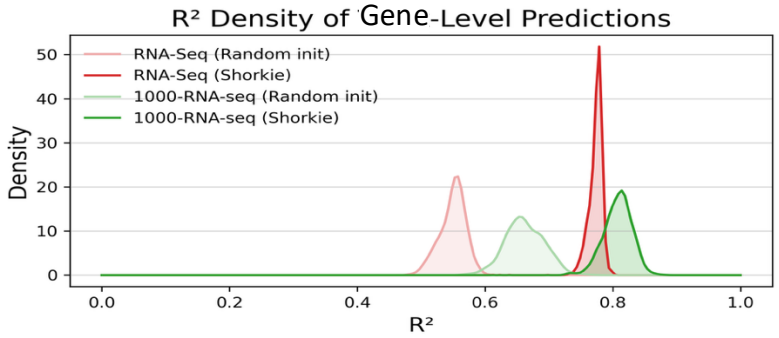




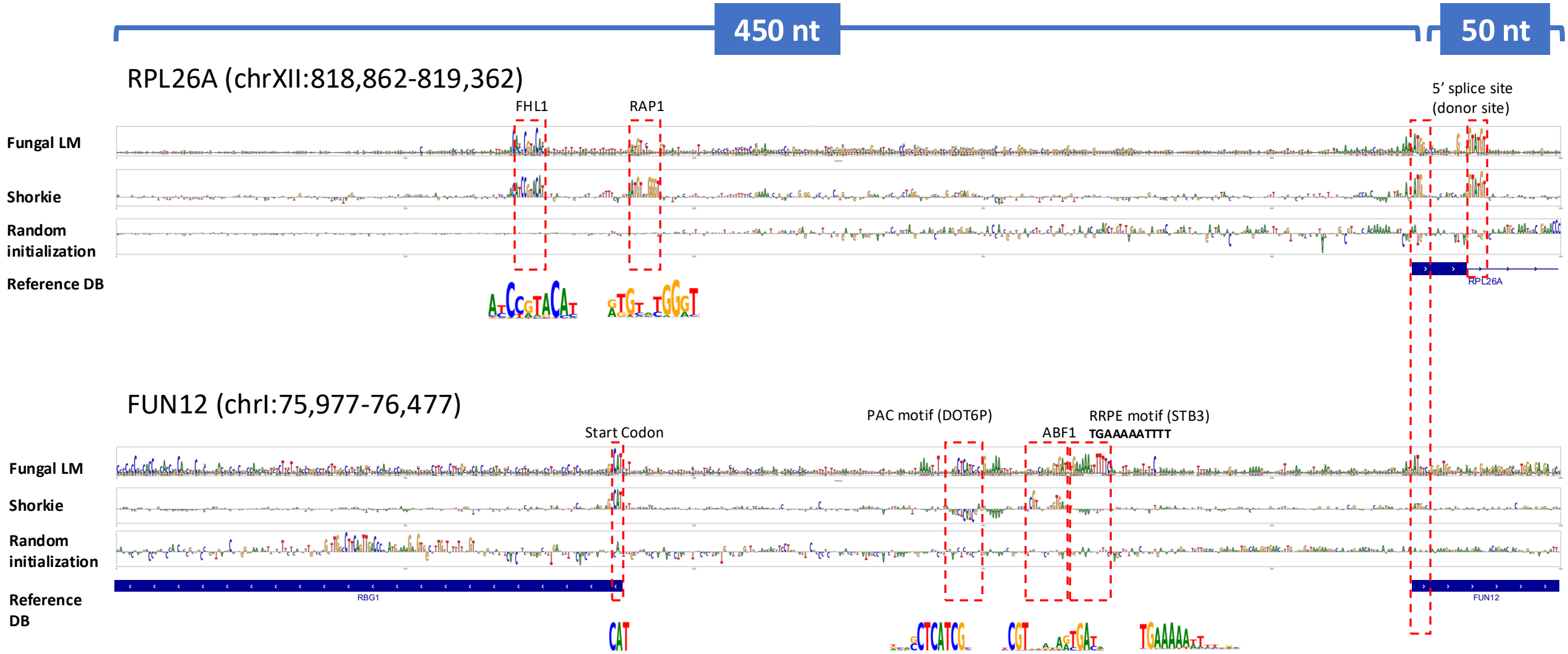
Track level
(each dot is a gene; all tracks)



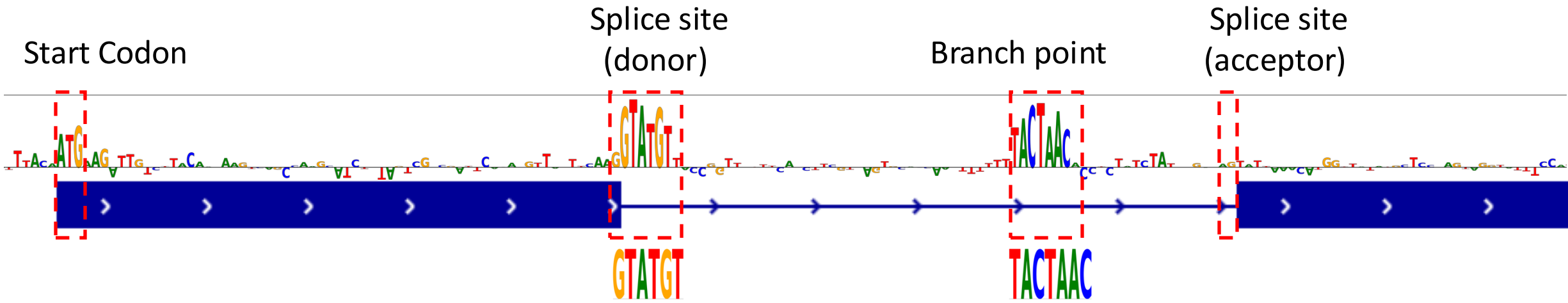


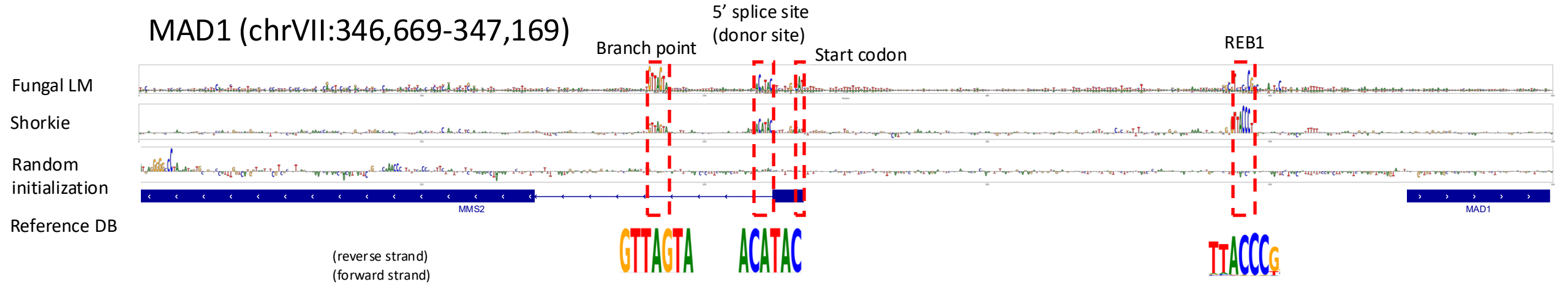


Shorkie ISM maps: Ribosomal genes & RRB genes



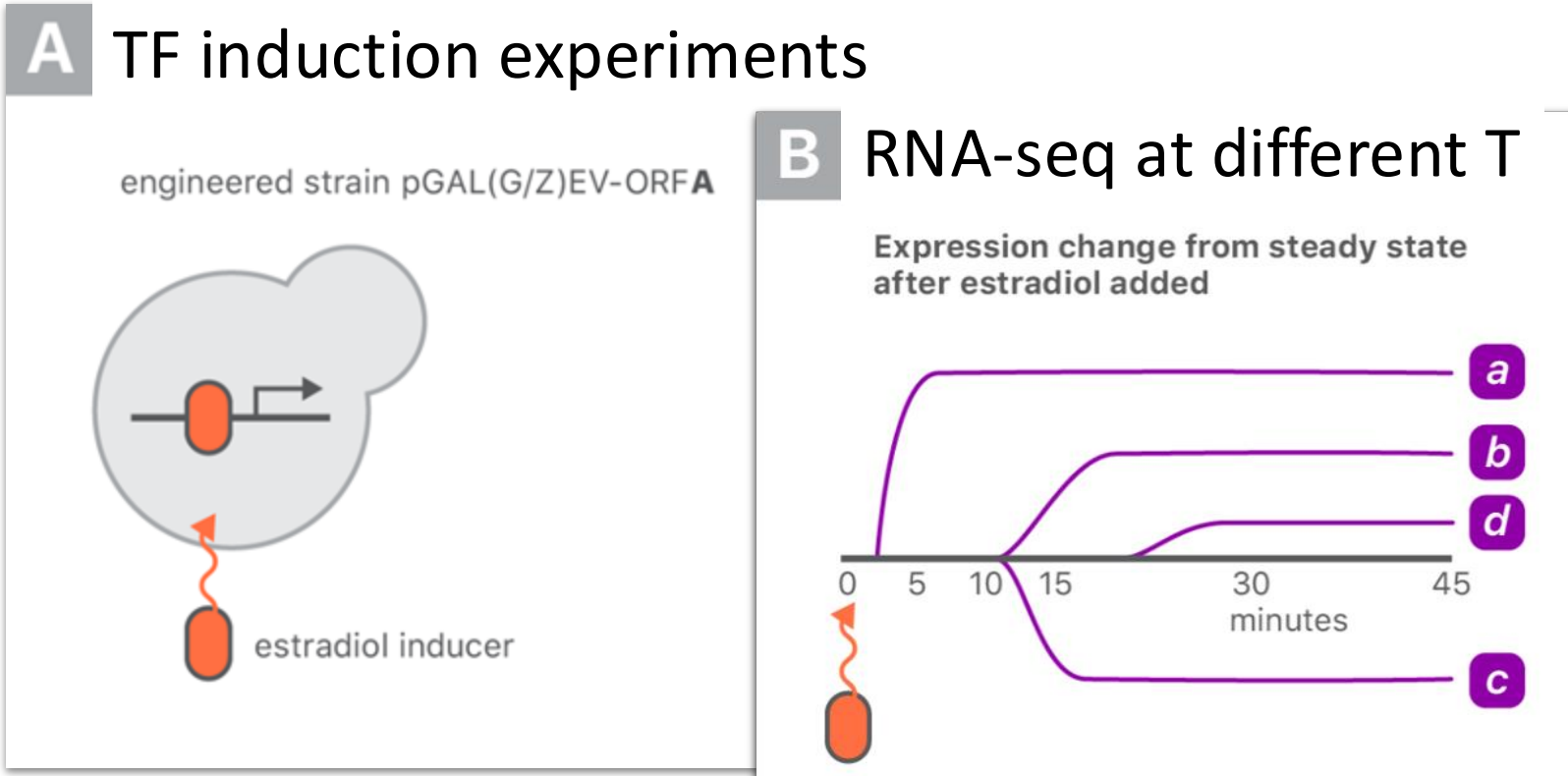
Shorkie ISM maps: Splicing motifs



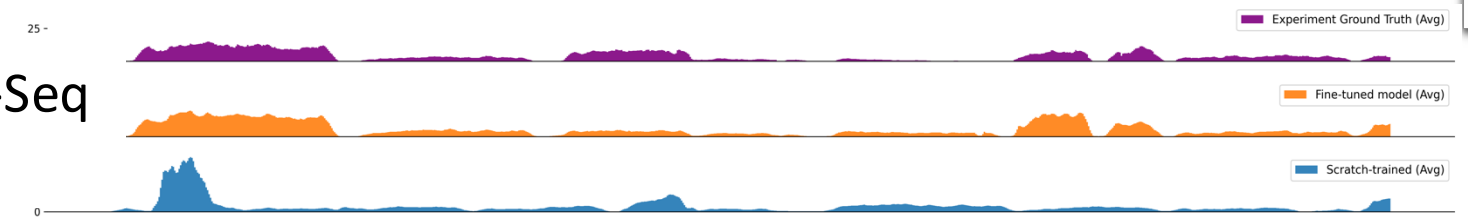


Time-course TF-inducted RNA-Seq

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

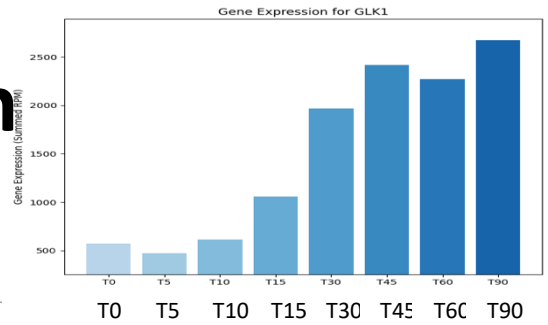


3054 RNA-Seq readouts



Shorkie MSN2 Induction temporal RNA-Seq prediction

chrIII:50388-50888 (Promoter region of GLK1)



T0 T5 T10 T15 T30 T45 T60 T90

T0

T5

T10

T15

T30

T45

T60

T90

YeTFaSCo DB motif

MSN2 **TATA Box**

ATG42

Average motif changes using TF-modisco

MSN2 Induction

YeTFaSCo DB motif

TF-Modisco-detected MSN2 binding site

$T_5 - T_0$	$T_{10} - T_0$	$T_{15} - T_0$	$T_{30} - T_0$	$T_{45} - T_0$	$T_{60} - T_0$	$T_{90} - T_0$
Not clustered						

MET4 Induction

YeTFaSCo DB motif

TF-Modisco-detected MSN2 binding site

$T_5 - T_0$	$T_{10} - T_0$	$T_{15} - T_0$	$T_{30} - T_0$	$T_{45} - T_0$	$T_{60} - T_0$	$T_{90} - T_0$

MSN4 Induction

YeTFaSCo DB motif

TF-Modisco-detected MSN2 binding site

$T_5 - T_0$	$T_{10} - T_0$	$T_{15} - T_0$	$T_{30} - T_0$	$T_{45} - T_0$	$T_{60} - T_0$	$T_{90} - T_0$

SWI4 Induction

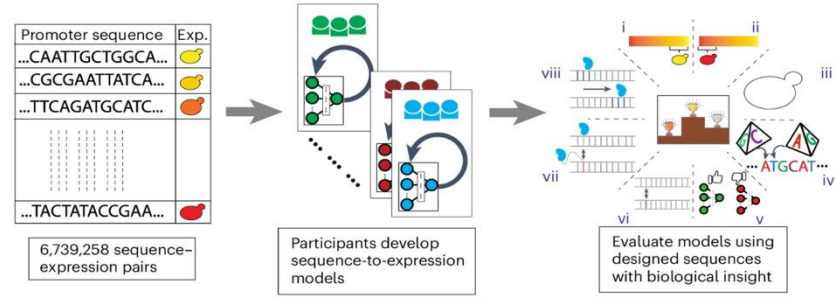
YeTFaSCo DB motif

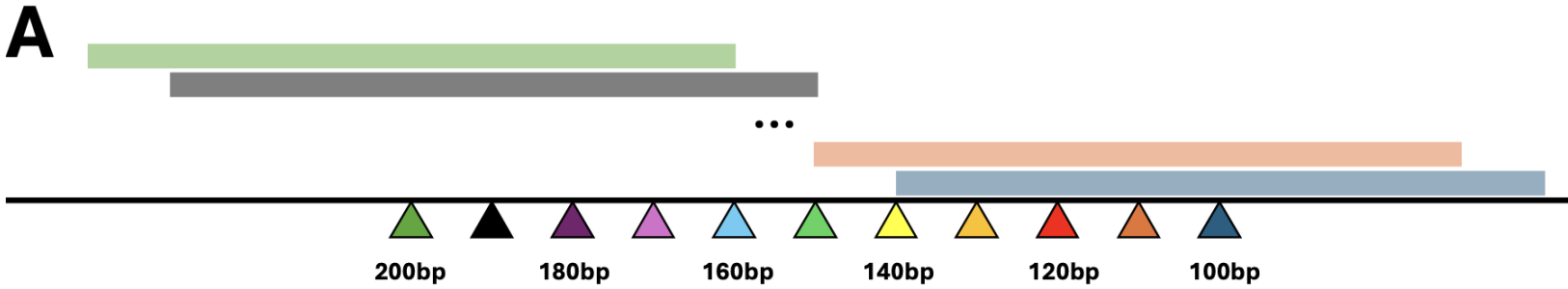
TF-Modisco-detected SWI4 binding site

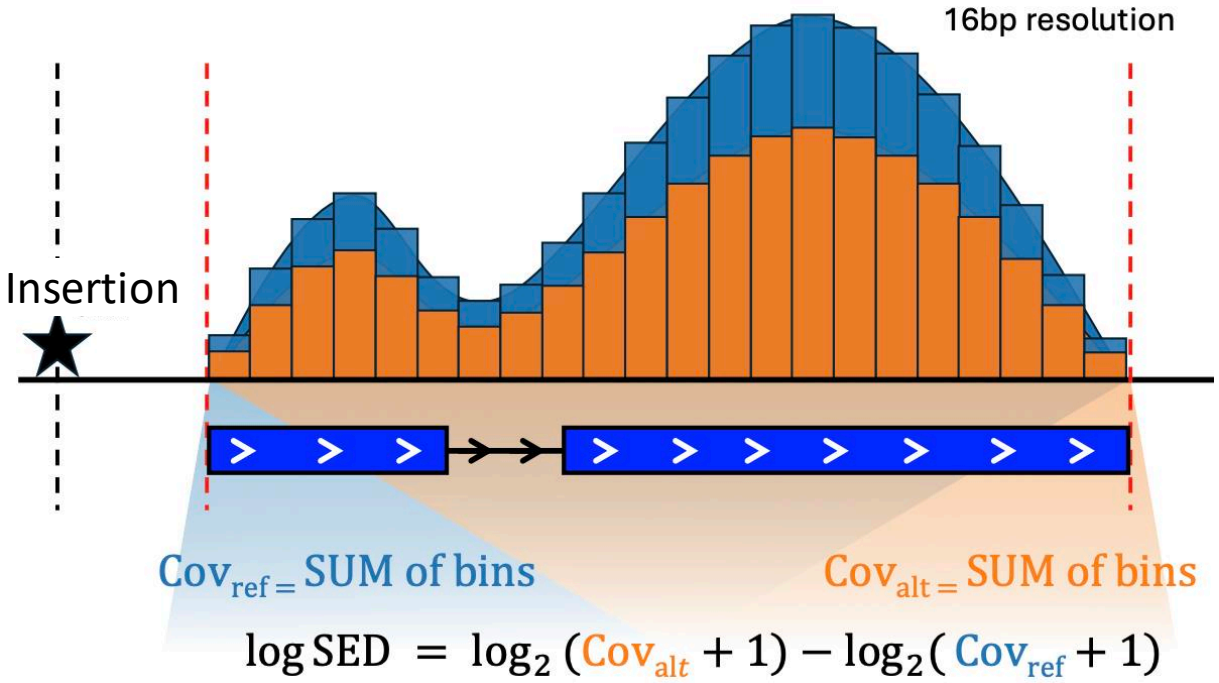
$T_5 - T_0$	$T_{10} - T_0$	$T_{20} - T_0$	$T_{40} - T_0$	$T_{70} - T_0$	$T_{120} - T_0$	$T_{180} - T_0$
Not clustered	Not clustered		Not clustered	Not clustered		

Rafi et al., Nature Biotech (2024)

MPRA mutation effect prediction



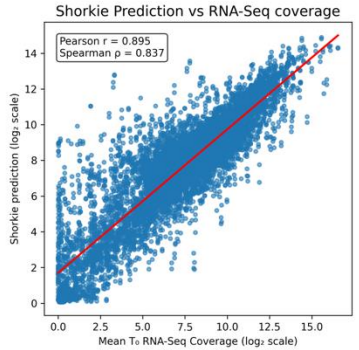


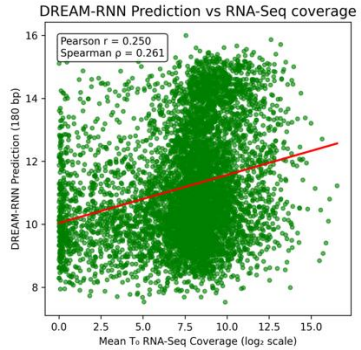


Shorkie

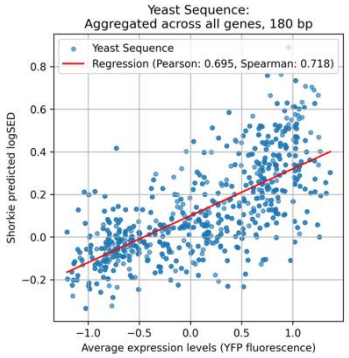
DREAM-RNN

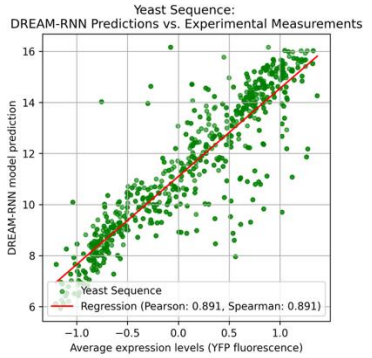
RNA-Seq Coverage Prediction





Rafi et al. Yeast Sequences (single-sequence)

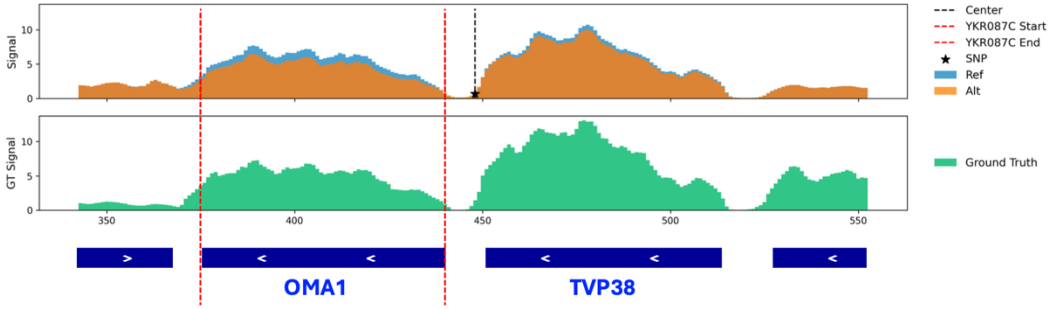


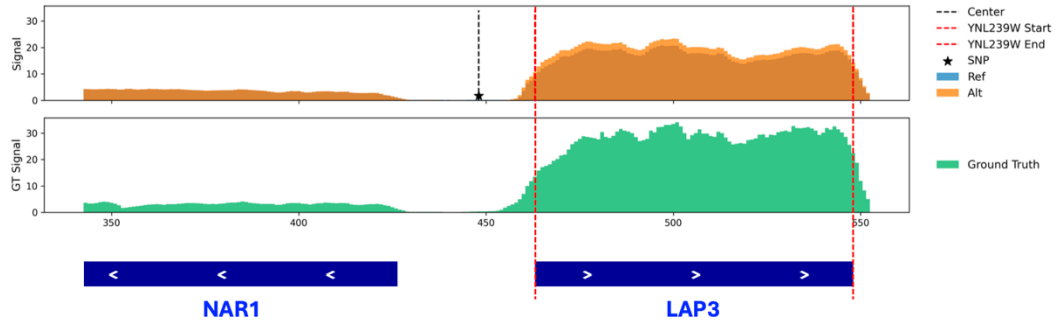


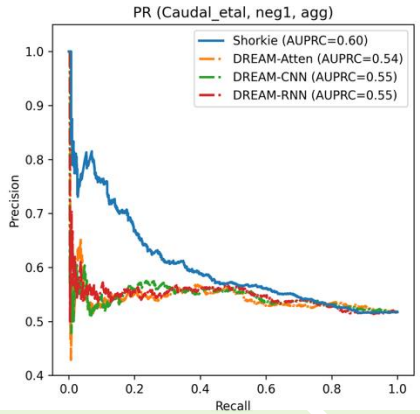
Caudal et al., Nature genetics (2024)

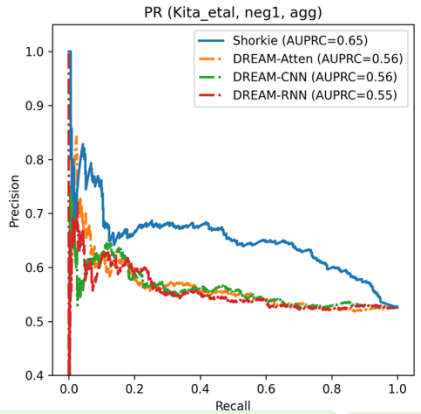
Kita et al., Nature genetics (2024)

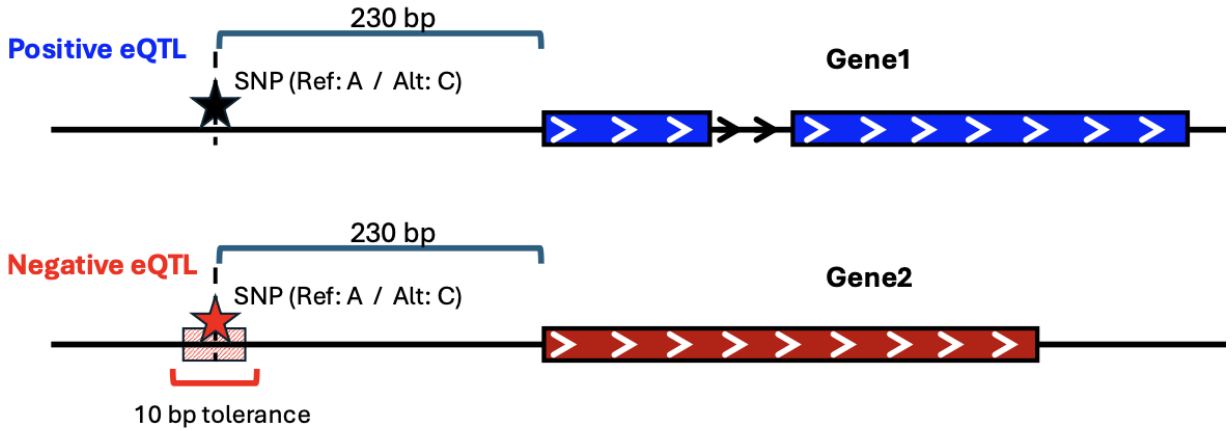
Variant effects on eQTLs and negatively selected eQTLs

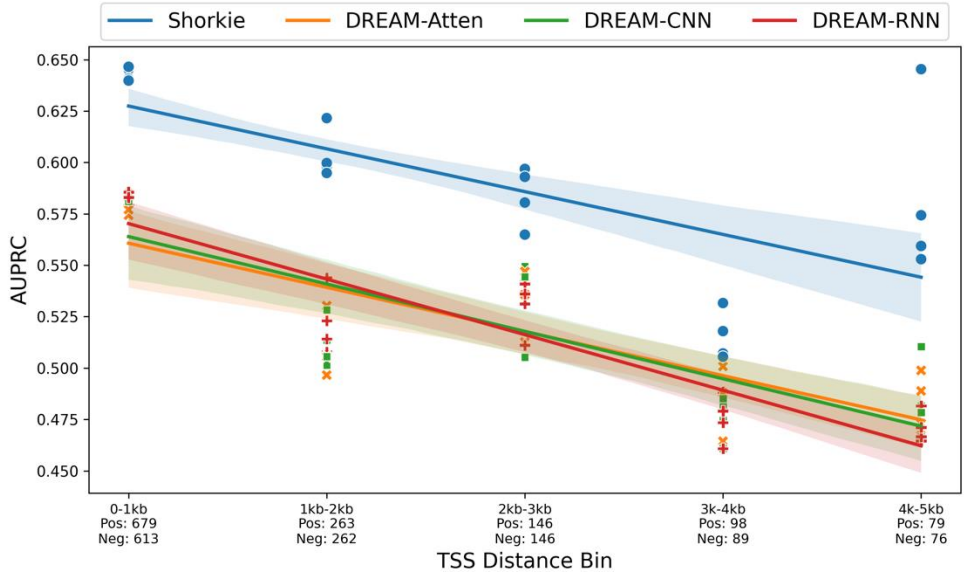






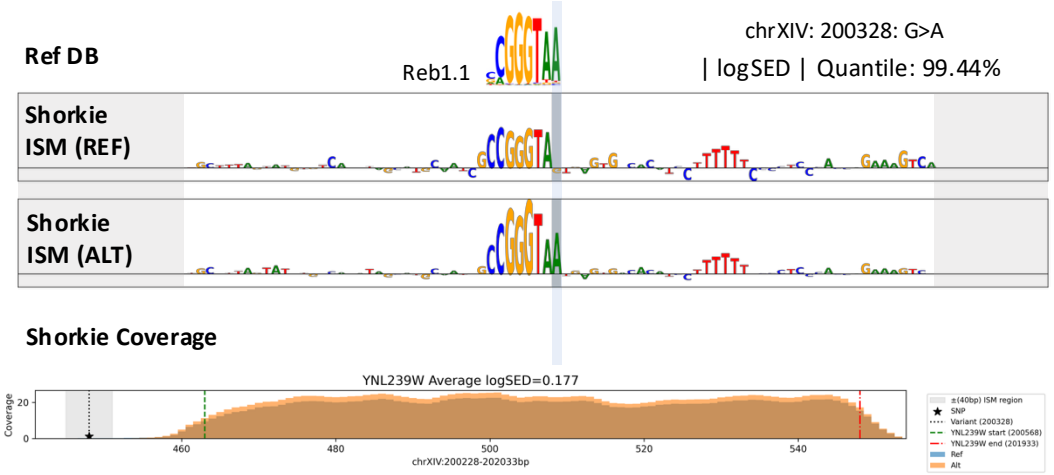




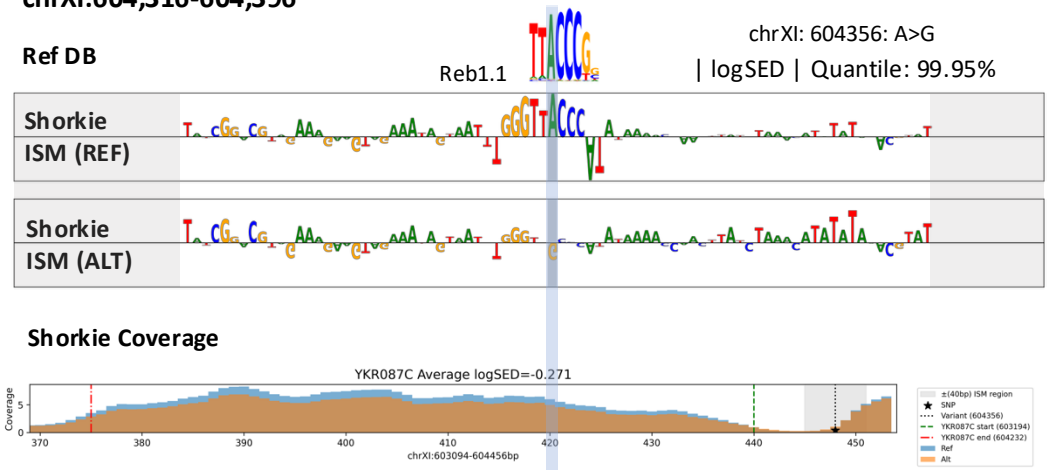


Variant effects on eQTLs and negatively selected eQTLs

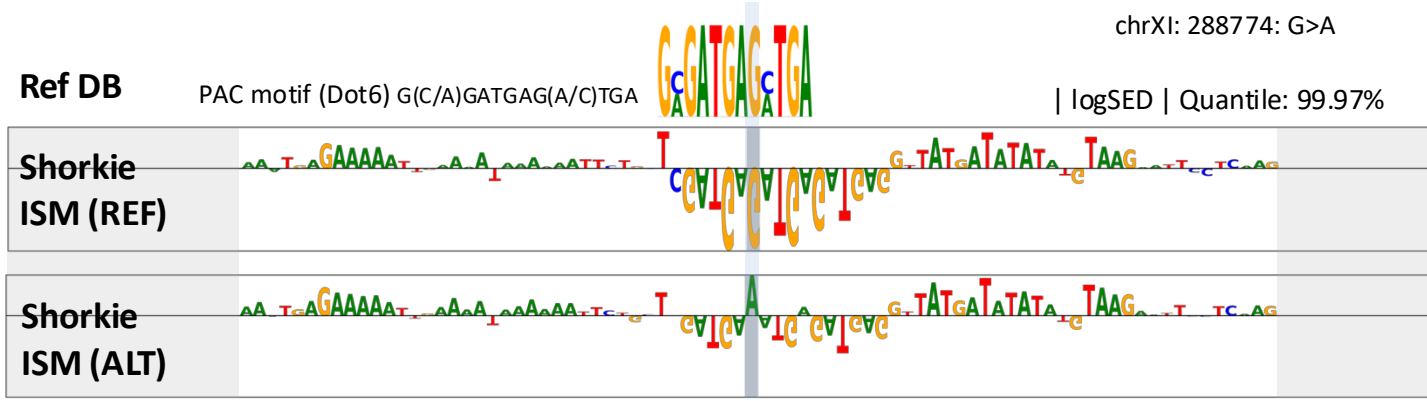
chrXIV:200,288 – 200,368



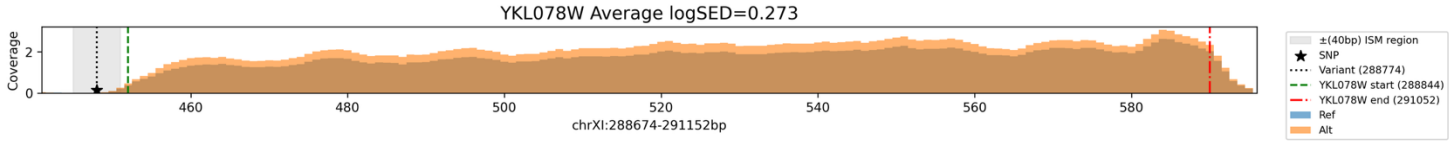
chrXI:604,316-604,396



chrXI:288,734 – 288,814



Shorkie Coverage



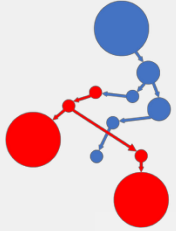
How do we assemble the complete 3-billion-nucleotide genome?

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

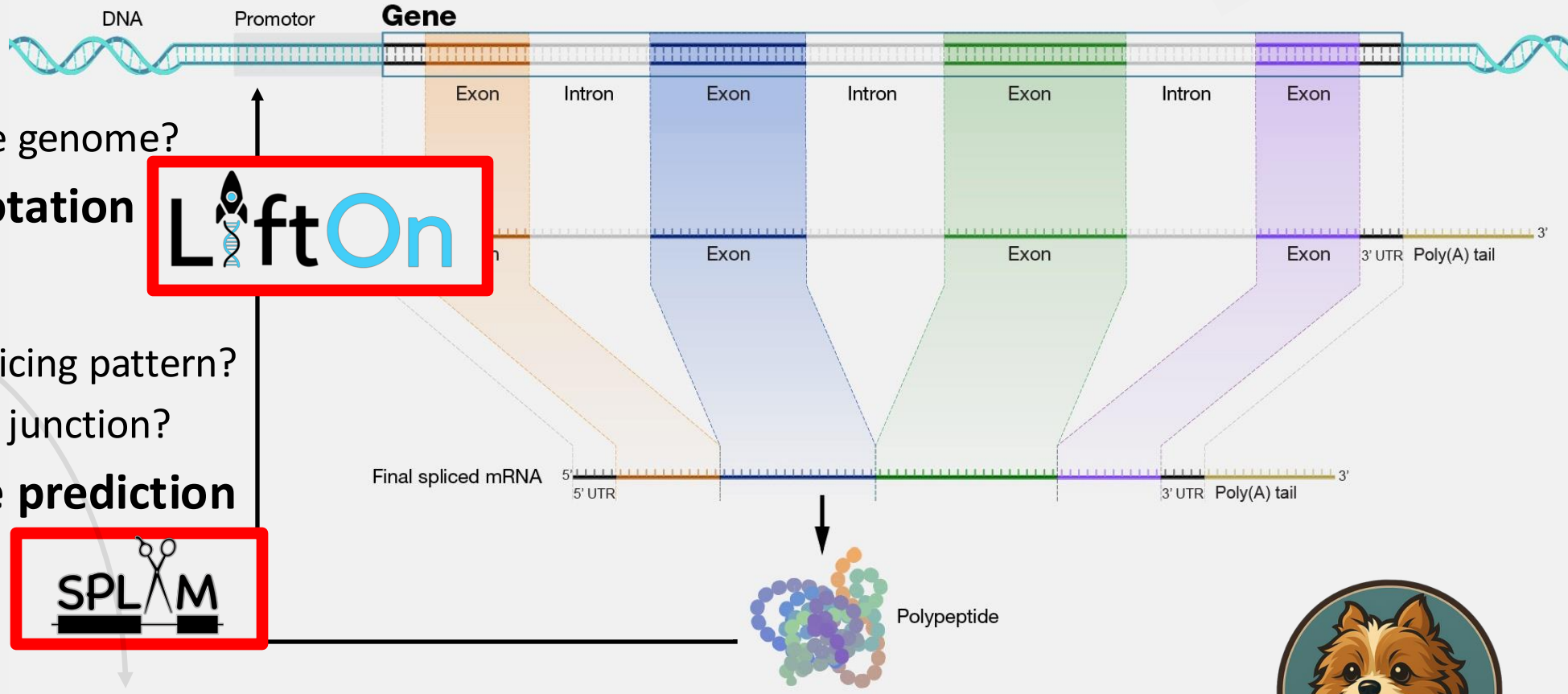
What are the canonical splicing pattern?

Alternative splicing? Splice junction?

Part IV & V: Splice site prediction

OpenSpliceAI

SPLAM

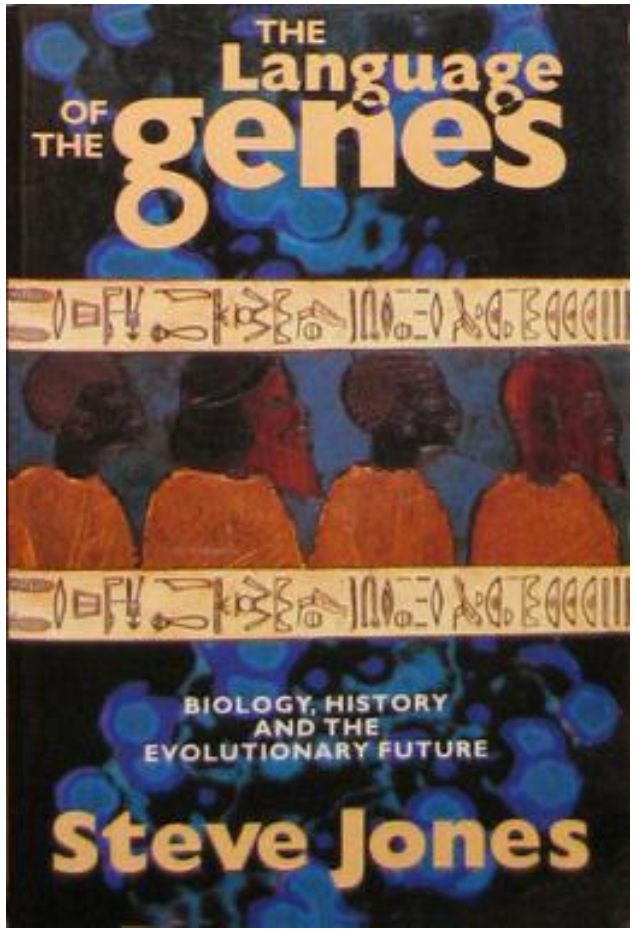


Can we predict gene expression by learning the regulatory grammar in the genome?

Part VI: Shorkie. RNA-Seq coverage prediction

Shorkie

Conclusions: Language of genomes



Steve Jones, 1993

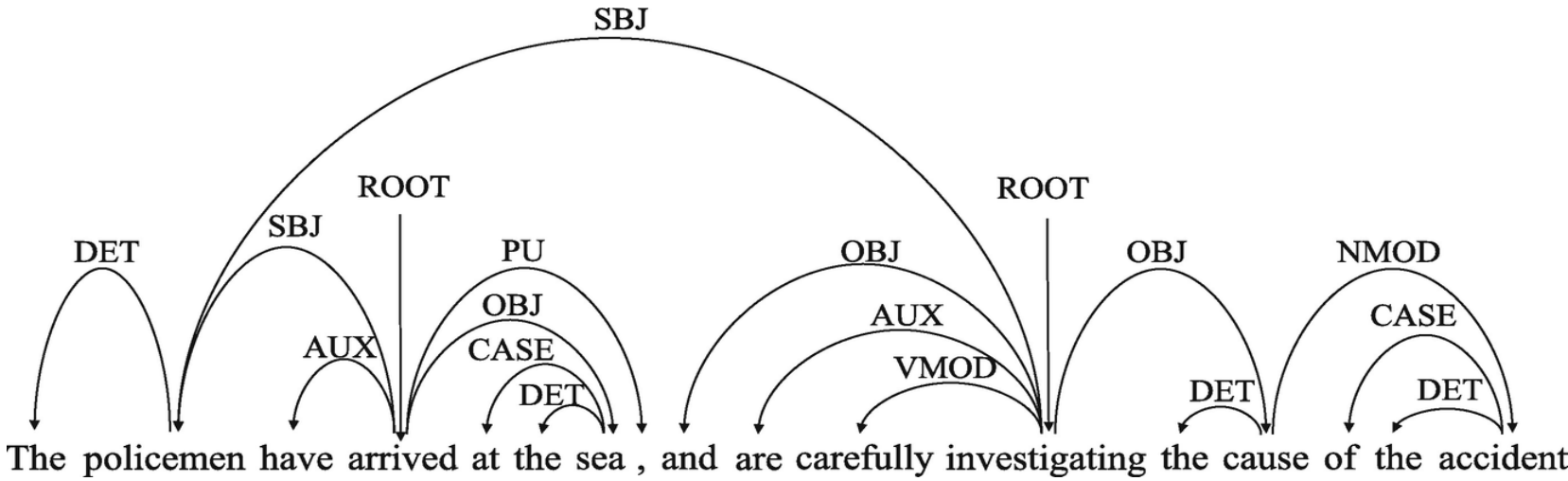
Codons as “words” and genes as “sentences”.

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

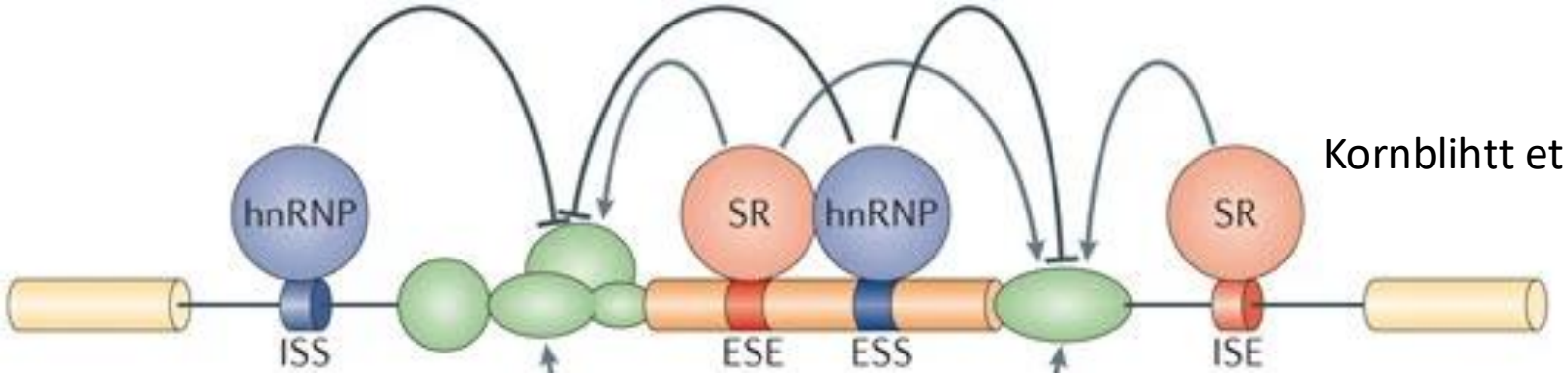
Recurring k-mer motifs as discrete tokens,
modeling long-range dependencies as grammatical rules.

Conclusions: Language of genomes

Natural Language Processing

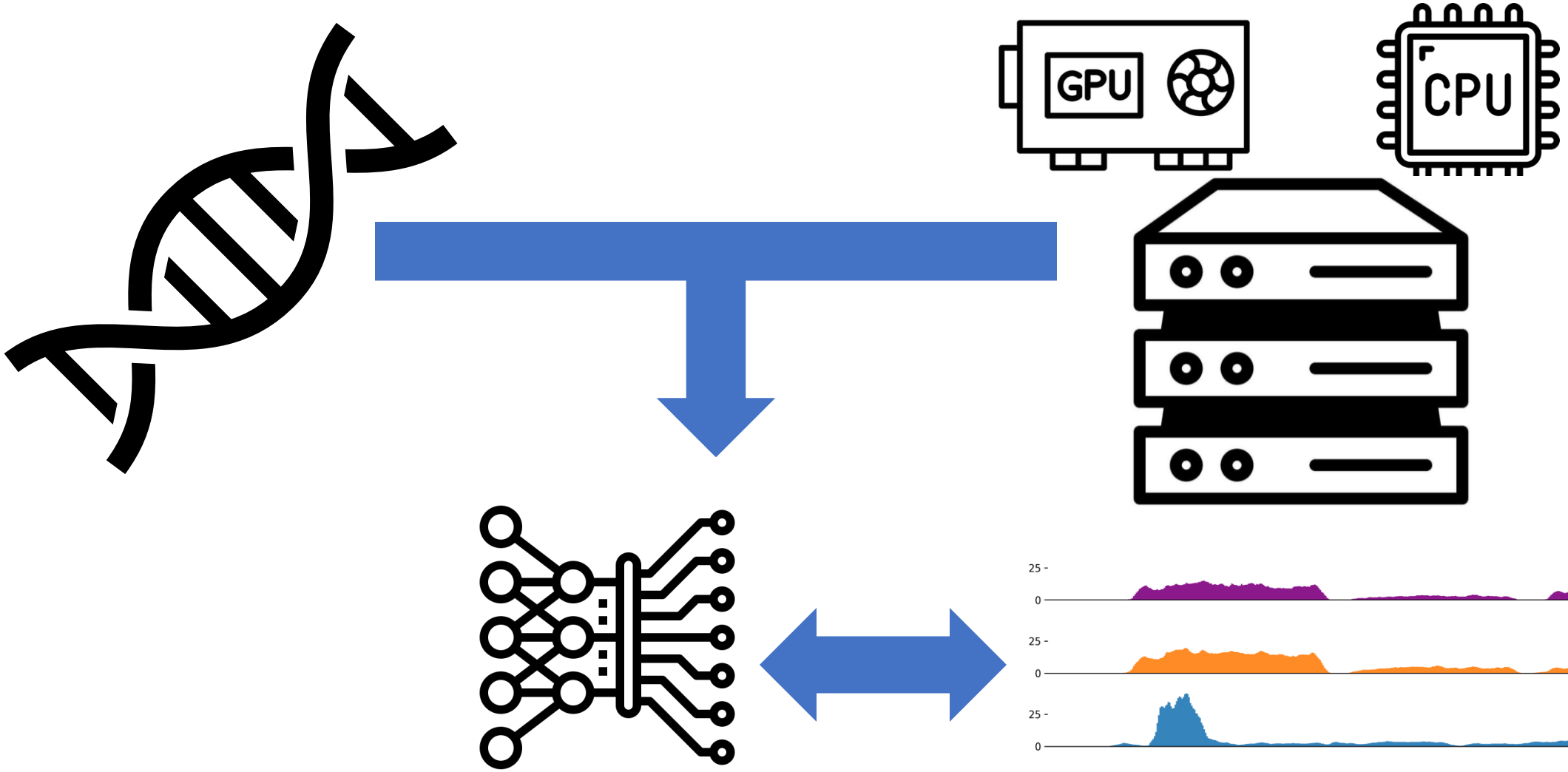


DNA gene regulation



Kornblihtt et al.

Conclusions: Language of genomes



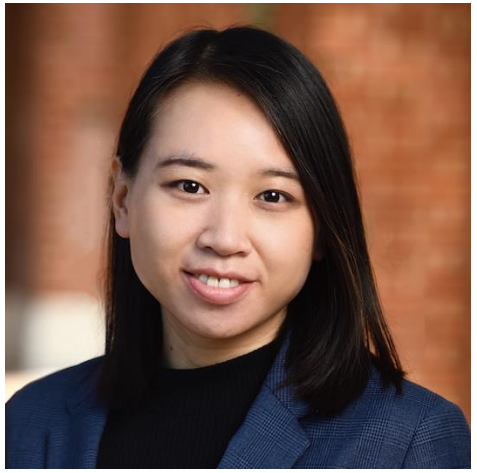
Committee members

Steven Salzberg

Mihaela Pertea

Ben Langmead

David Kelley



Anqi Liu

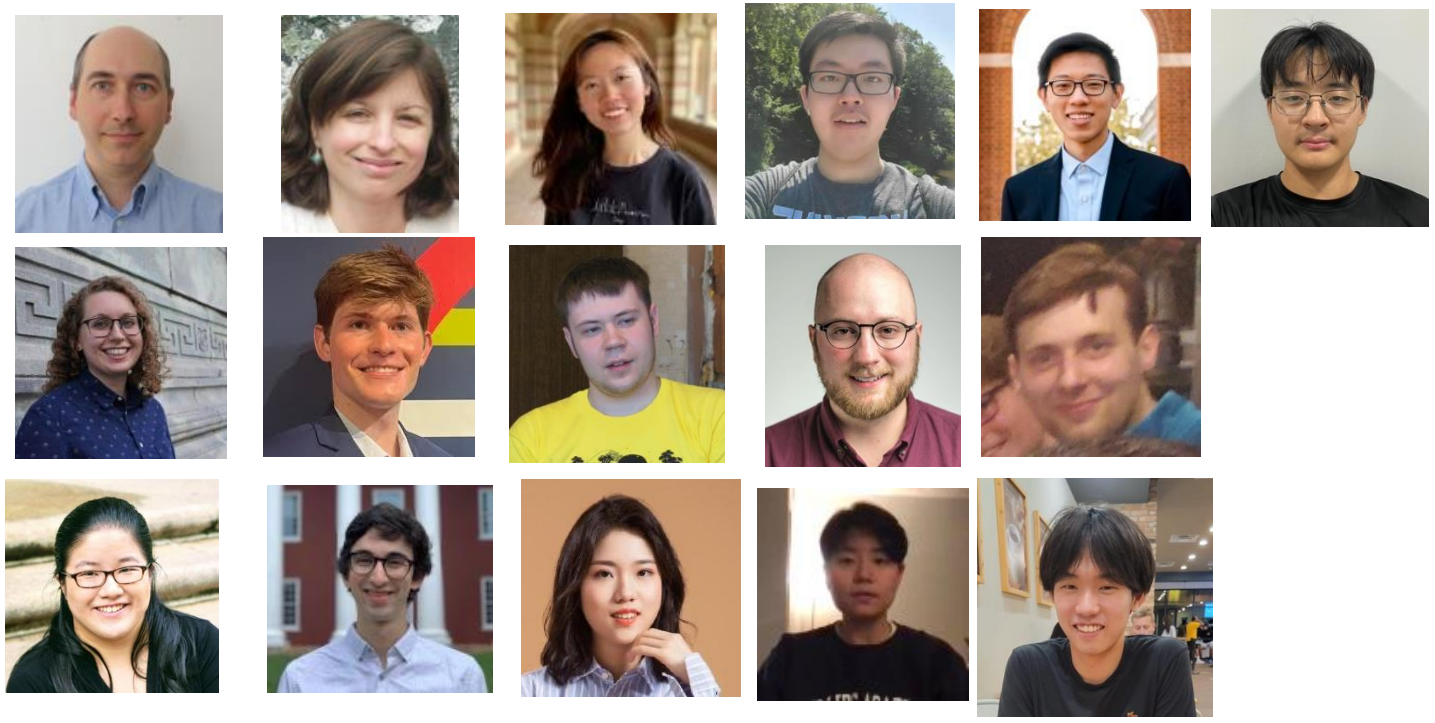
Best labmates

Salzberg Lab

Dr. Steven Salzberg

Pertea Lab

Dr. Mihaela Pertea



Hopkins Genomics + CS Friends

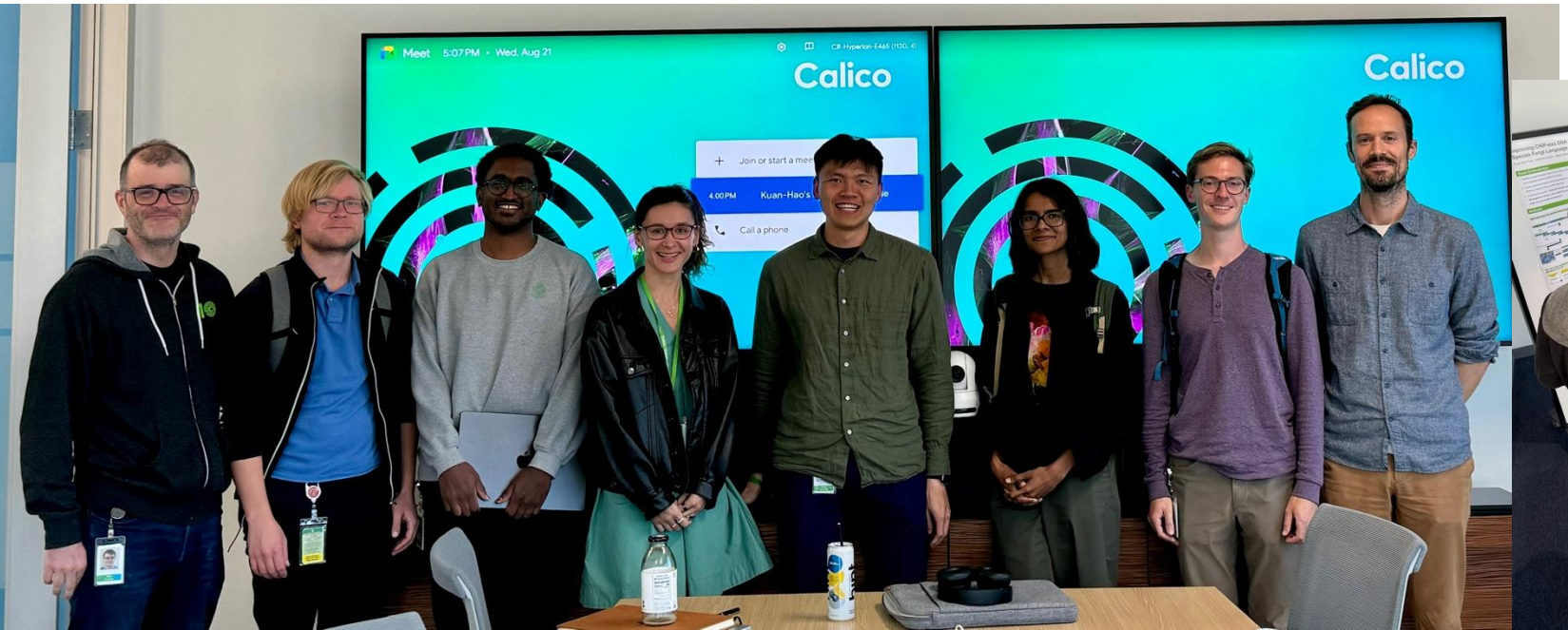

bioRxiv
Tangerine

Calico Friends (2024 Summer)

Dr. David Kelley

Dr. Johannes Linder

Majed Magzoub



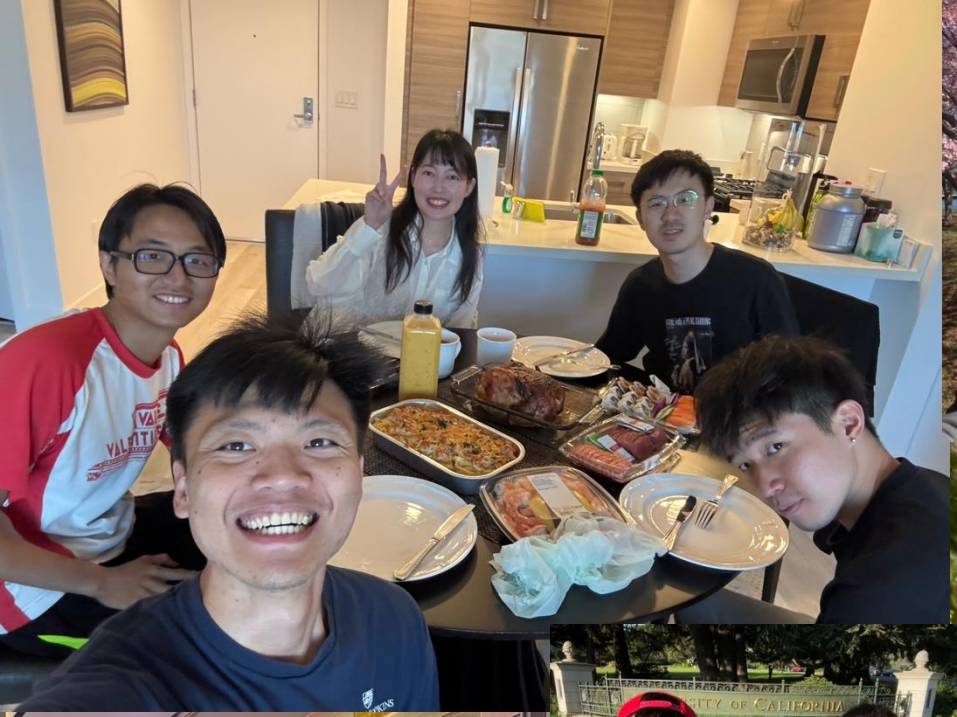









Brotherly Love
21
NTUEE
HARKER
CROSS COUNTY

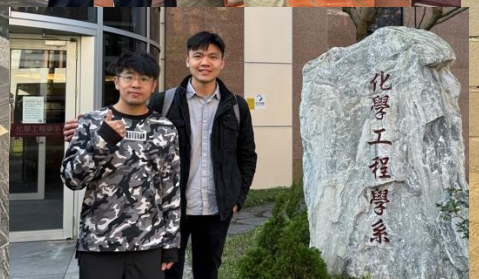












化學工程學系

GenoMalone Family

Quarter-finals (Spring 2025)

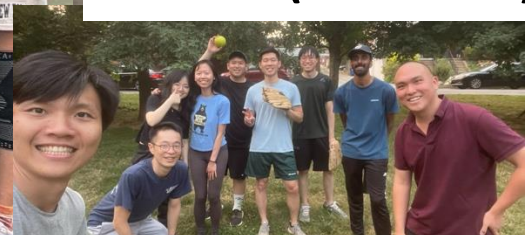
Playoff (Spring 2024)

TSA basketball family

JHU X UMD Sport Day

AC Malone (Sum 2023)

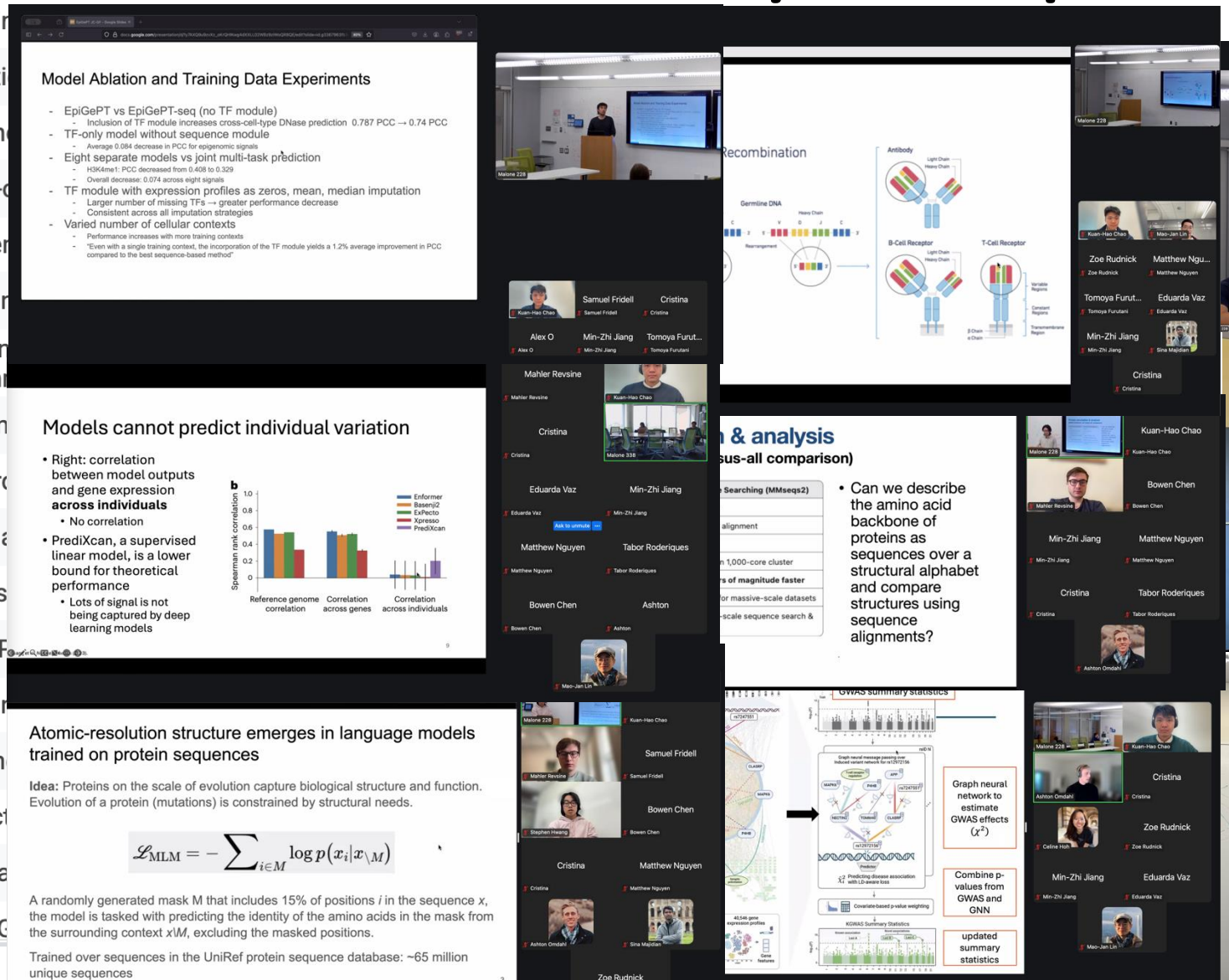
Softball (Sum 2023)



JHU Deep Learning in Genomics Study Group

Mahler Revsine

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	Applyin
01/21	Stephen Hwang	Evoluti
02/11	Celine Hoh	A foun
02/18	Ashton Omdahl	Small-c
03/04	Zoe Rudnick	Sequer
03/25	Bowen Chen	Fast ar
04/08	Mahler Revsine	Person Benchr
04/22	Kuan-Hao Chao	Design
05/06	Cristina Martin Linares	Towarc
05/27	Bohan Ni	A scala
06/03	Mao-Jan Lin	Diseas
06/17	Gus Fridell	EpiGeP
07/01	Celine Hoh	The lar
07/22	Eduarda Vaz	Param
07/29	Jake Galvin	Predict
08/12	Zoe Rudnick	Accura
08/26	Bowen Chen	AlphaG



Family

My research path

Next …


Dr. David R Kelley


Dr. Johannes Linder


Majed Mohamed Magzoub

Calico

Calico


Liu Yuh-san


Dr. Eric Y. Chuang


Dr. Tzu-Pin Lu


Dr. Robert Lanfear


Dr. Huai-Kuang Tsai

2016 2017 2018 2019 2020 2021 2022 2023 2024 2025

Taipei Municipal Chien Kuo High School

National Taiwan University

ANU

AS

Taiwan Army

Johns Hopkins University

Taipei Municipal Chien Kuo High School

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®