JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING
Department of Computer Science

JOHNS HOPKINS UNIVERSITY
CENTER FOR COMPUTATIONAL BIOLOGY
CCB

Improving CHiP-exo DNA-binding and gene expression predictions with a multi-species fungi language model

Kuan-Hao Chao

2024.08.21

khchao.com

@KuanHaoChao

Kuanhao-Chao

Improving ChIP-exo DNA Binding and Gene Expression Predictions with a Multi-Species Fungi Language Model
Calico

Improving ChIP-exo DNA Binding and Gene Expression Predictions with a Multi-Species Fungi Language Model
Calico

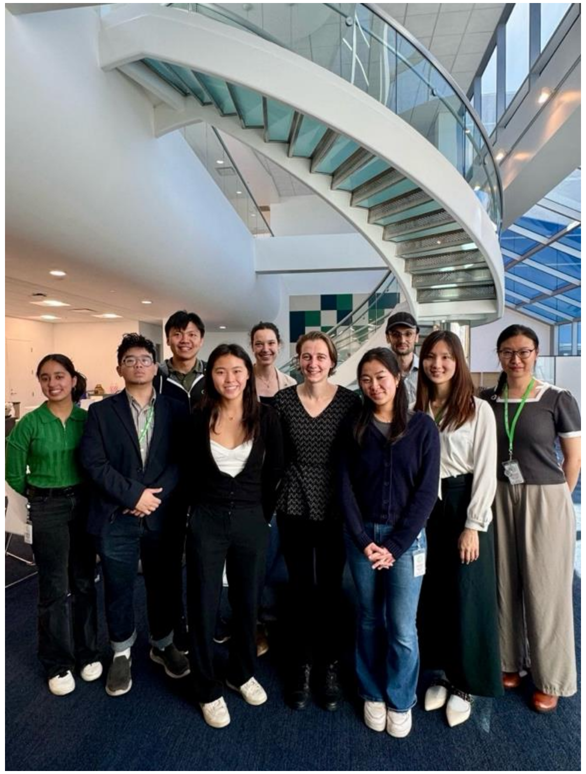
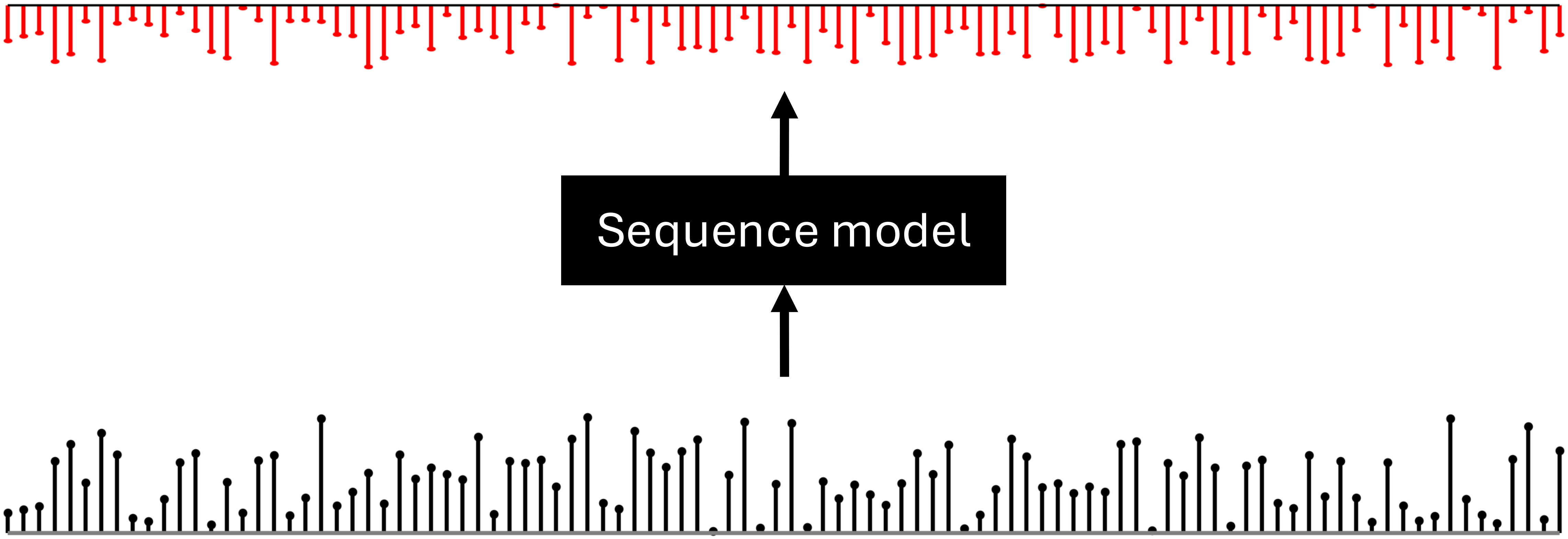


Photo with you!

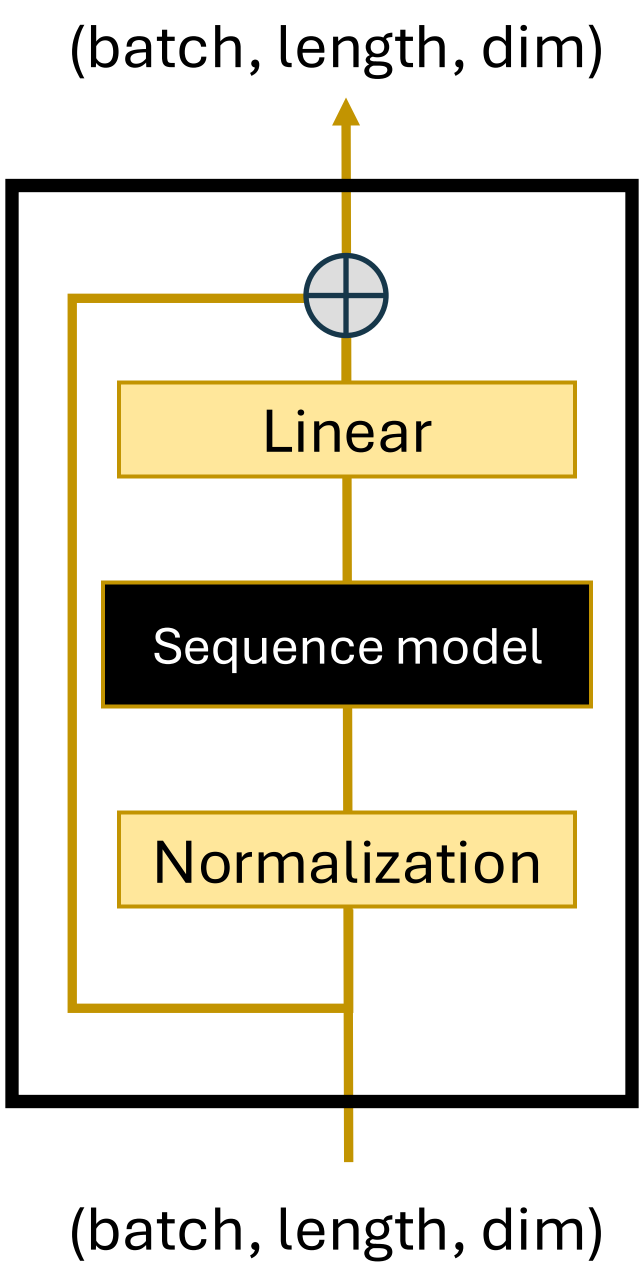
Reference: https://www.youtube.com/watch?v=luCBXCErkCs&t=197s



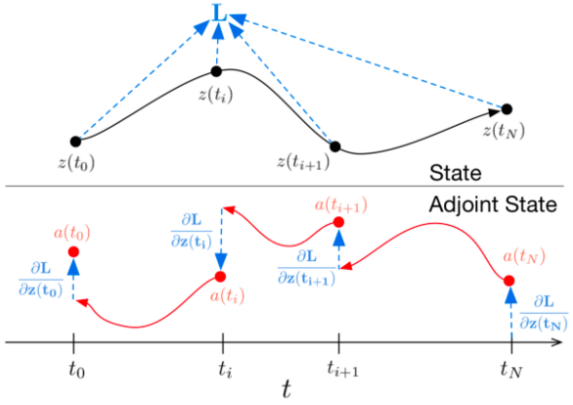
Sequence model

Sequence models map a sequence to a sequence

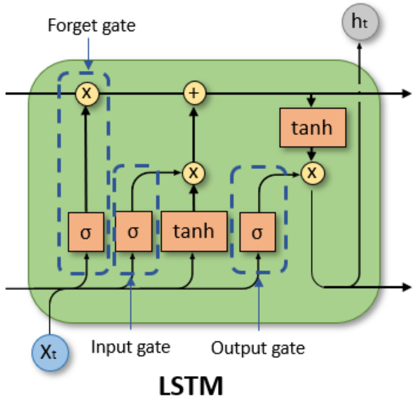
Reference: https://www.youtube.com/watch?v=luCBXCErkCs&t=197s



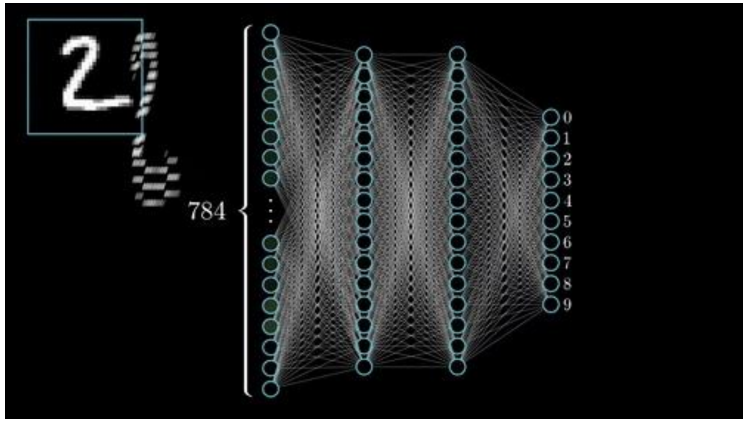
Neural ODEs


L
$z(t_i)$
$z(t_0)$
$z(t_{i+1})$
$z(t_N)$
State
Adjoint State
$a(t_0)$
$\frac{\partial L}{\partial z(t_0)}$
$\frac{\partial L}{\partial z(t_i)}$
$a(t_i)$
$a(t_{i+1})$
$\frac{\partial L}{\partial z(t_{i+1})}$
$a(t_N)$
$\frac{\partial L}{\partial z(t_N)}$
t_0
t_i
t_{i+1}
t_N
t

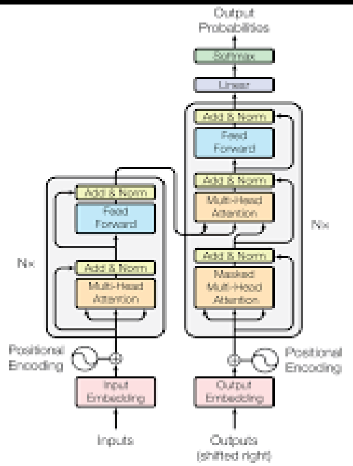
RNN

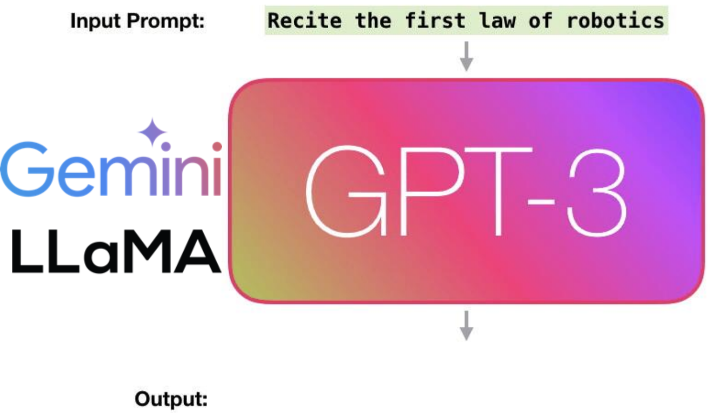


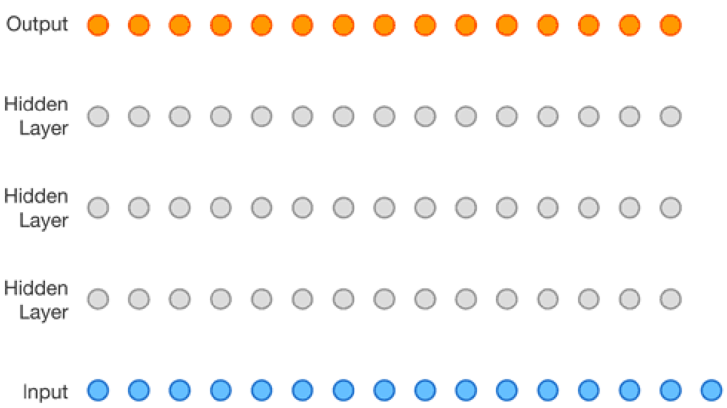
CNNs



Transformers

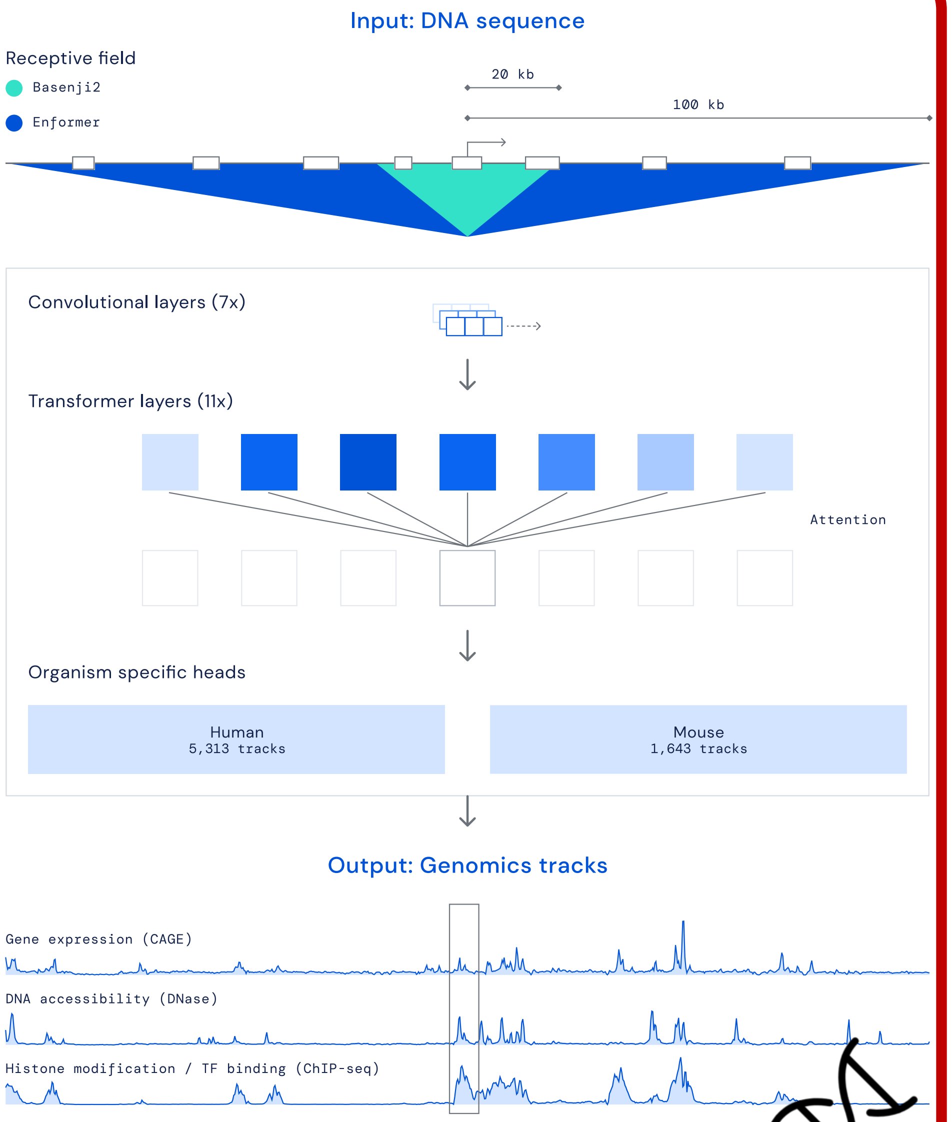






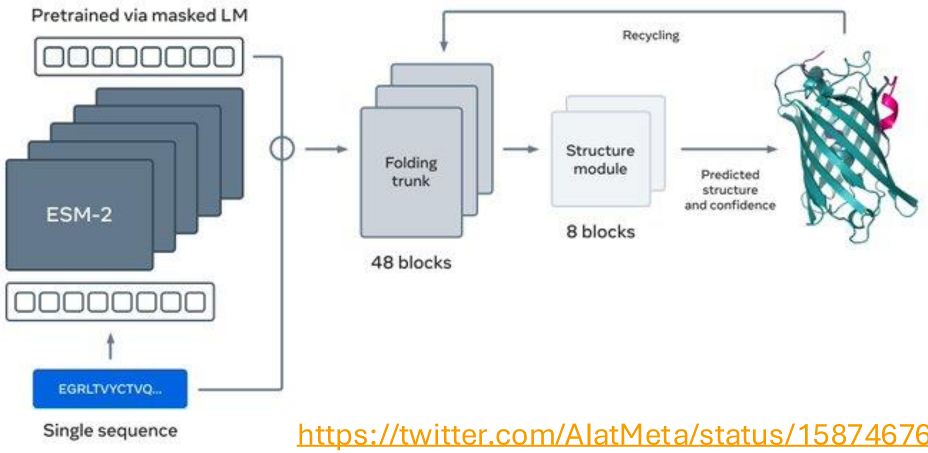
https://jalammar.github.io/how-gpt3-works-visualizations-animations/

https://deepmind.google/discover/blog/wavenet-a-generative-model-for-raw-audio/



Enformer

https://deepmind.google/discover/blog/predicting-gene-expression-with-a

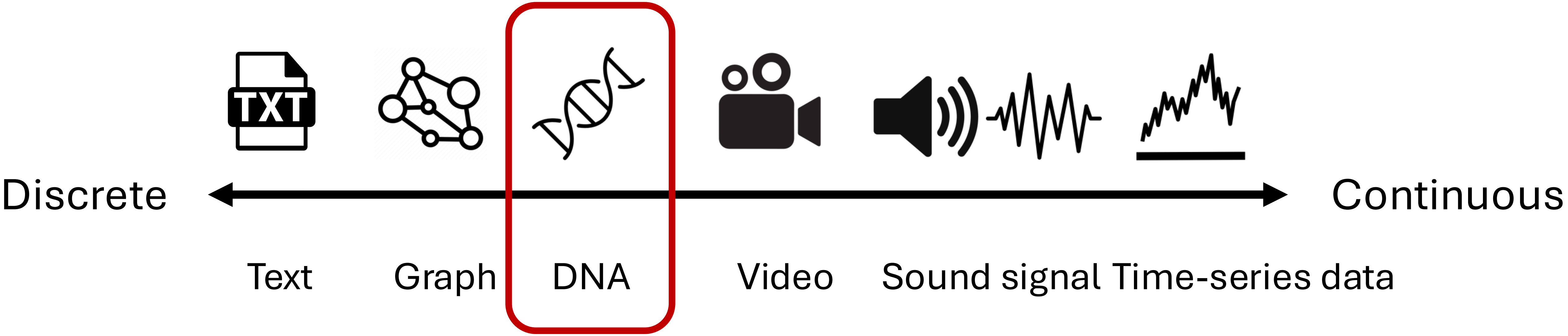


ESMFold

https://twitter.com/AIatMeta/status/1587467600413351937/photo/1

Reference: https://www.youtube.com/watch?v=luCBXCErkCs&t=197s

Spectrum of Sequential Data



Discrete

Continuous

Text

Graph

DNA

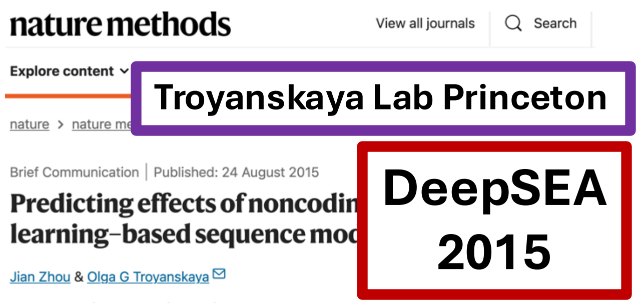
Video

Sound signal

Time-series data

DeepMind Calico illumina® Troyanskaya Laboratory

Why Deep learning sequence models to DNA？



nature methods View all journals Search

Explore content

nature > nature me

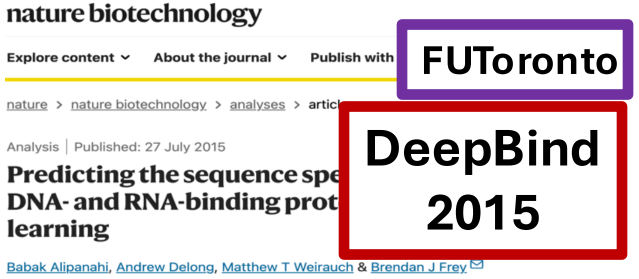
Brief Communication | Published: 24 August 2015

Predicting effects of noncodin learning–based sequence moc

Jian Zhou & Olga G Troyanskaya

Troyanskaya Lab Princeton

DeepSEA 2015



nature biotechnology

Explore content About the journal Publish with

nature > nature biotechnology > analyses > artic

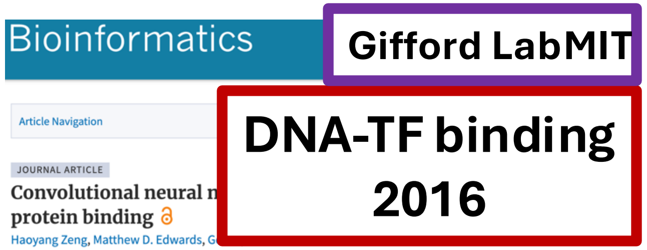
Analysis | Published: 27 July 2015

Predicting the sequence spe DNA- and RNA-binding prot learning

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

FUToronto

DeepBind 2015



Bioinformatics

Article Navigation

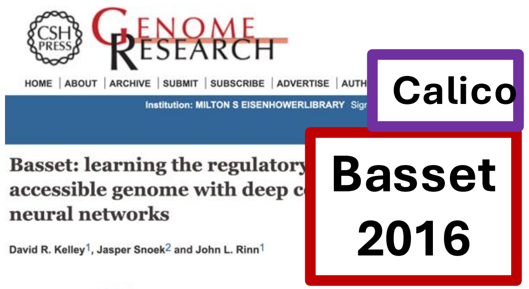
JOURNAL ARTICLE

Convolutional neural n protein binding

Haoyang Zeng, Matthew D. Edwards, G

Gifford LabMIT

DNA-TF binding 2016



CSH PRESS GENOME RESEARCH

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

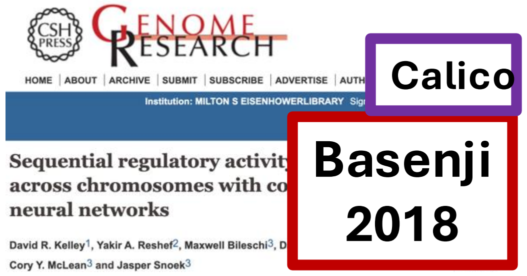
Institution: MILTON S EISENHOWERLIBRARY Sign

Basset: learning the regulatory accessible genome with deep c neural networks

David R. Kelley[1], Jasper Snoek[2] and John L. Rinn[1]

Calico

Basset 2016



CSH PRESS GENOME RESEARCH

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

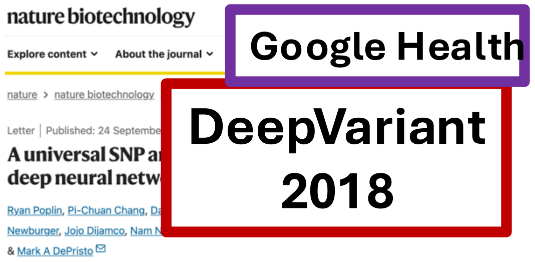
Institution: MILTON S EISENHOWERLIBRARY Sign

Sequential regulatory activit across chromosomes with co neural networks

David R. Kelley[1], Yakir A. Reshef[2], Maxwell Bileschi[3], D Cory Y. McLean[3] and Jasper Snoek[3]

Calico

Basenji 2018



nature biotechnology

Explore content About the journal

nature > nature biotechnology

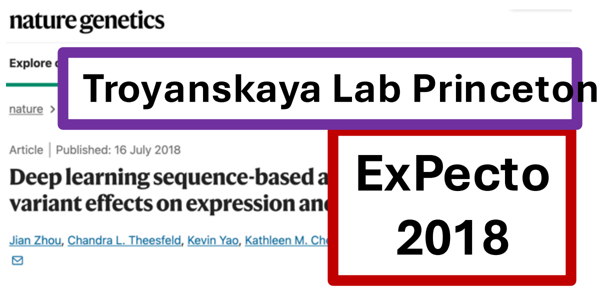
Letter | Published: 24 Septembe

A universal SNP ar deep neural netw

Ryan Poplin, Pi-Chuan Chang, Da Newburger, Jojo Dijamco, Nam N & Mark A DePristo

Google Health

DeepVariant 2018



nature genetics

Explore c

nature >

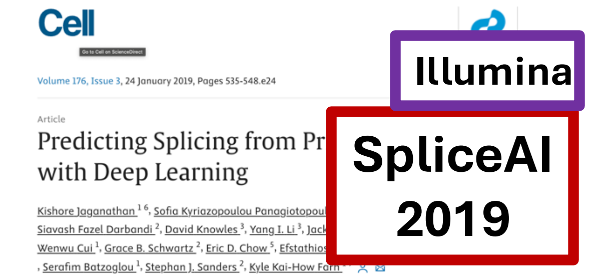
Article | Published: 16 July 2018

Deep learning sequence-based a variant effects on expression anc

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

Troyanskaya Lab Princeton

ExPecto 2018



Cell

Go to Cell on ScienceDirect

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

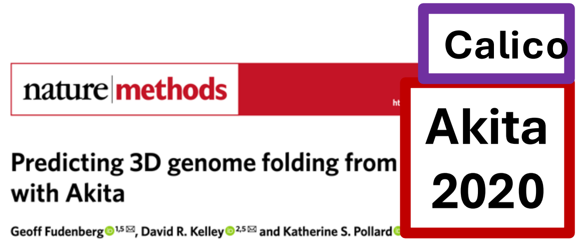
Article

Predicting Splicing from Pr with Deep Learning

Kishore Jaganathan[1 6], Sofia Kyriazopoulou Panagiotopou Siavash Fazel Darbandi[2], David Knowles[3], Yang I. Li[3], Jack Wenwu Cui[1], Grace B. Schwartz[2], Eric D. Chow[5], Efstathios , Serafim Batzoglou[1], Stephan J. Sanders[2], Kyle Kai-How Farh

Illumina

SpliceAI 2019



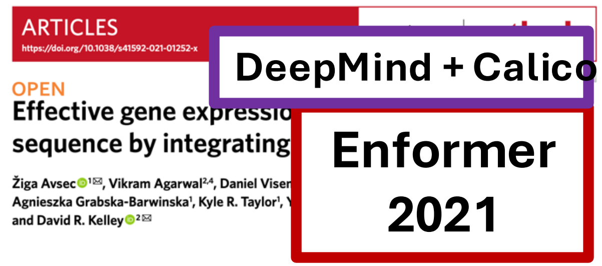
nature methods

Predicting 3D genome folding from with Akita

Geoff Fudenberg[1,5], David R. Kelley[2,5] and Katherine S. Pollard

Calico

Akita 2020



ARTICLES

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

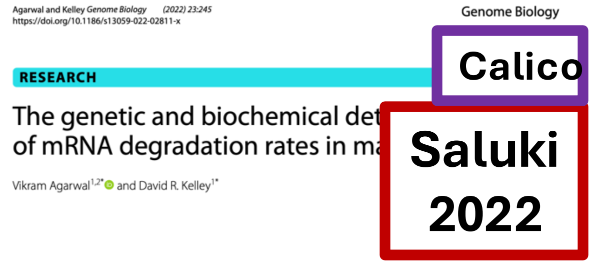
OPEN

Effective gene expressio sequence by integrating

Žiga Avsec[1], Vikram Agarwal[2,4], Daniel Viser Agnieszka Grabska-Barwinska[1], Kyle R. Taylor[1], Y and David R. Kelley[2]

DeepMind + Calico

Enformer 2021



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

Genome Biology

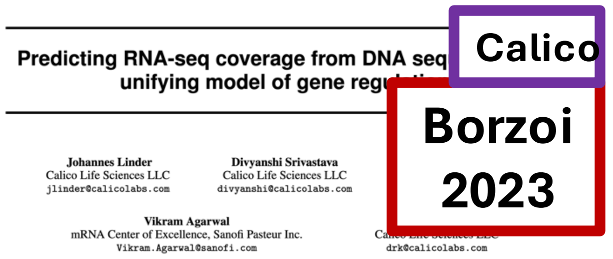
RESEARCH

The genetic and biochemical det of mRNA degradation rates in ma

Vikram Agarwal[1,2*] and David R. Kelley[1*]

Calico

Saluki 2022



Predicting RNA-seq coverage from DNA seq unifying model of gene regu

Johannes Linder
Calico Life Sciences LLC
jlinder@calicolabs.com

Divyanshi Srivastava
Calico Life Sciences LLC
divyanshi@calicolabs.com

Vikram Agarwal
mRNA Center of Excellence, Sanofi Pasteur Inc.
Vikram.Agarwal@sanofi.com

drk@calicolabs.com

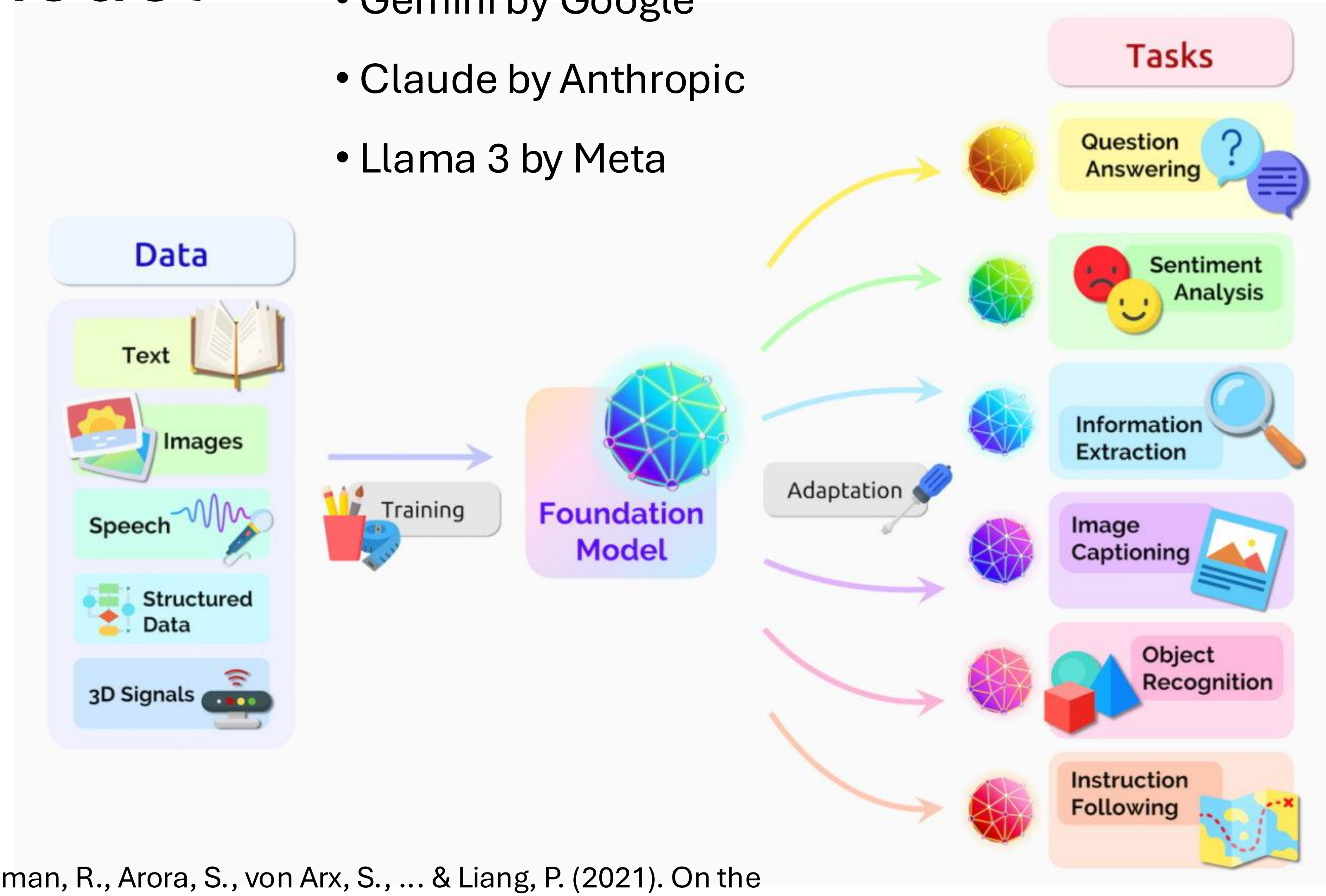
Calico

Borzoi 2023

Foundation model

- Stanford researchers called transformers “foundation models” in an August 2021 paper because they see them driving a paradigm shift in AI.

- GPT-3, GPT-4 by OpenAI
- Gemini by Google
- Claude by Anthropic
- Llama 3 by Meta



Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., … & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

Foundation model

- **Versatility**: wide range of downstream tasks
- **Transfer learning**: learn general representation of data. Task-specific is limited
- **Efficiency**: computational efficiency of fine-tuning models
- **Generalization**: “zero-shot” or “few-shot”
- **Emergent abilities**:
  - basic arithmetic
  - simple programming tasks
  - summarization, translation, or question-answering.

Goals

- Building an interpretable fungi LLM to help Calico construct gene regulatory networks (GRN) in the future.
- Predicting ChIP-exo, histone marks, and RNA-Seq
- Does fine-tuning a pretrained LM outperform training a new model from scratch under the exact model architecture?

Why yeast?

- Simple Eukaryotic Model
- Rapid Growth and Easy Culturing
- Genetic Manipulability
- Well-Characterized Genome
- Conserved Regulatory Mechanisms

Part I

Fungi Language Model

- Q: To what evolutionary distance should we include in our LM?
- Q: What is the quality of the annotation? Coding vs non-coding regions
- Q: How repetitive are the genomes?

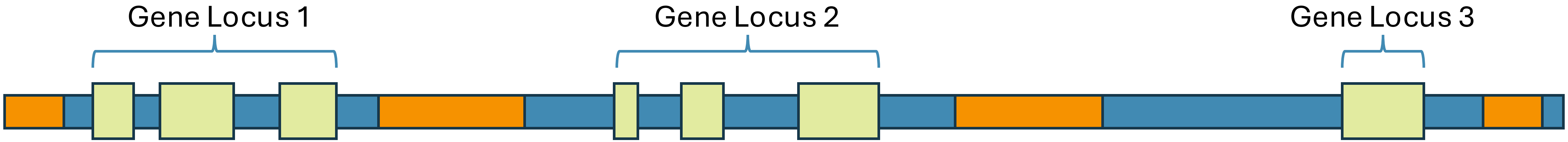
Why building a Fungi Language Model?

- Yeast genome is small. 12Mbps.
- Thousands of fungal genomes with high quality. No supervised measurements
- Language model pre-training on all available genomes followed by transfer learning to the smaller yeast genome.

Data preprocessing

🟧 Repeat regions

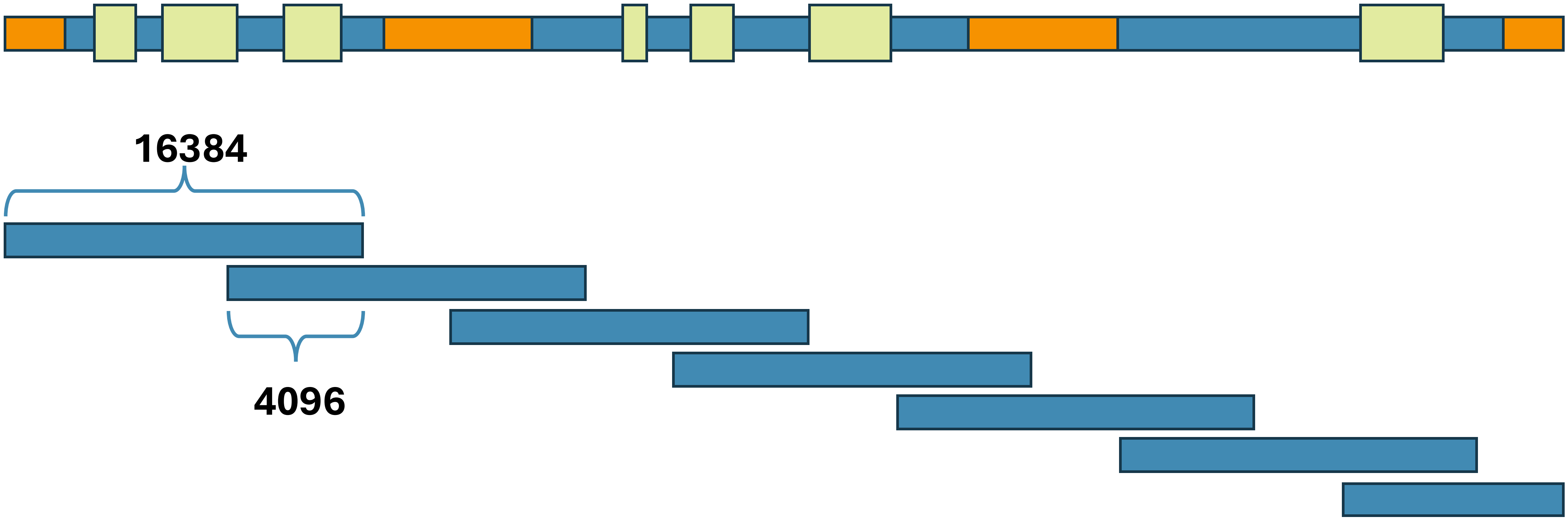
🟨 Coding regions



Data preprocessing

🟧 Repeat regions

🟨 Coding regions

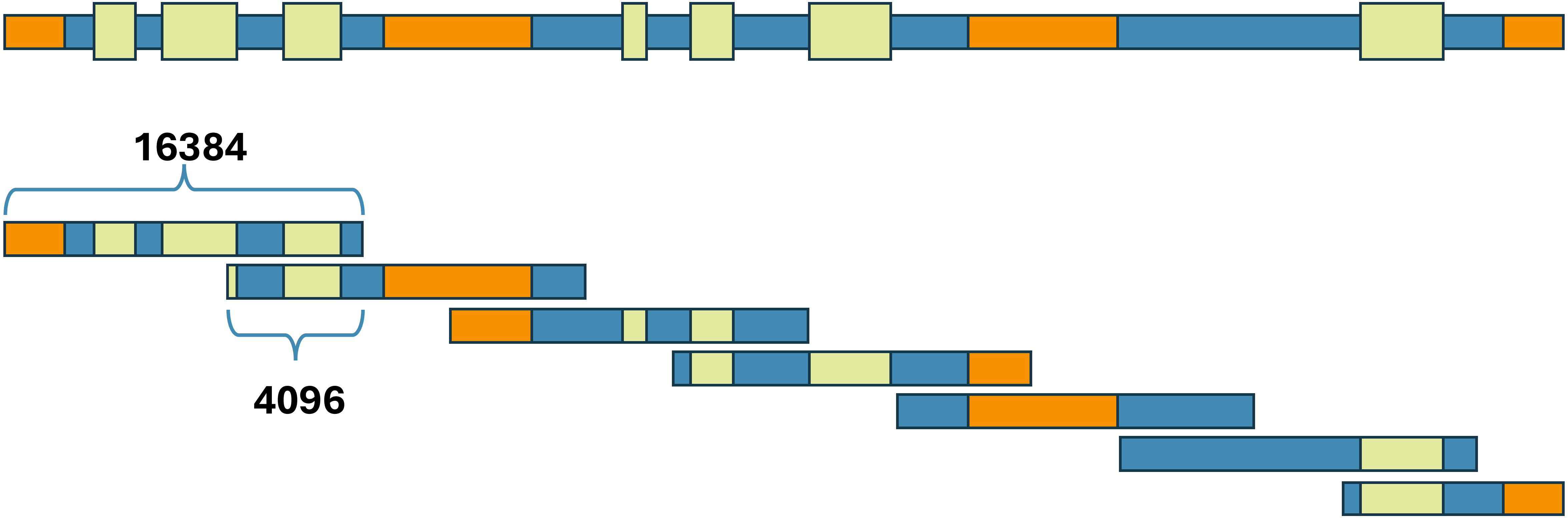


~ 7 genes per window

Data preprocessing

🟧 Repeat regions

🟨 Coding regions

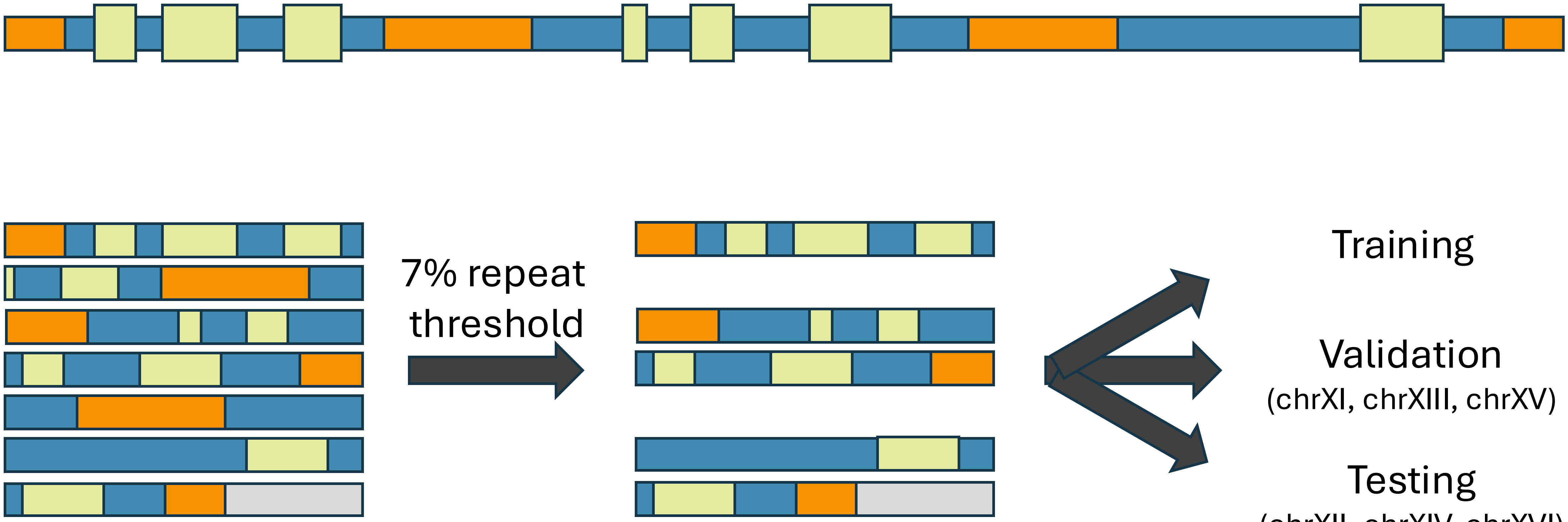


~ 7 genes per window

Data preprocessing

🟧 Repeat regions

🟨 Coding regions



7% repeat threshold

Training

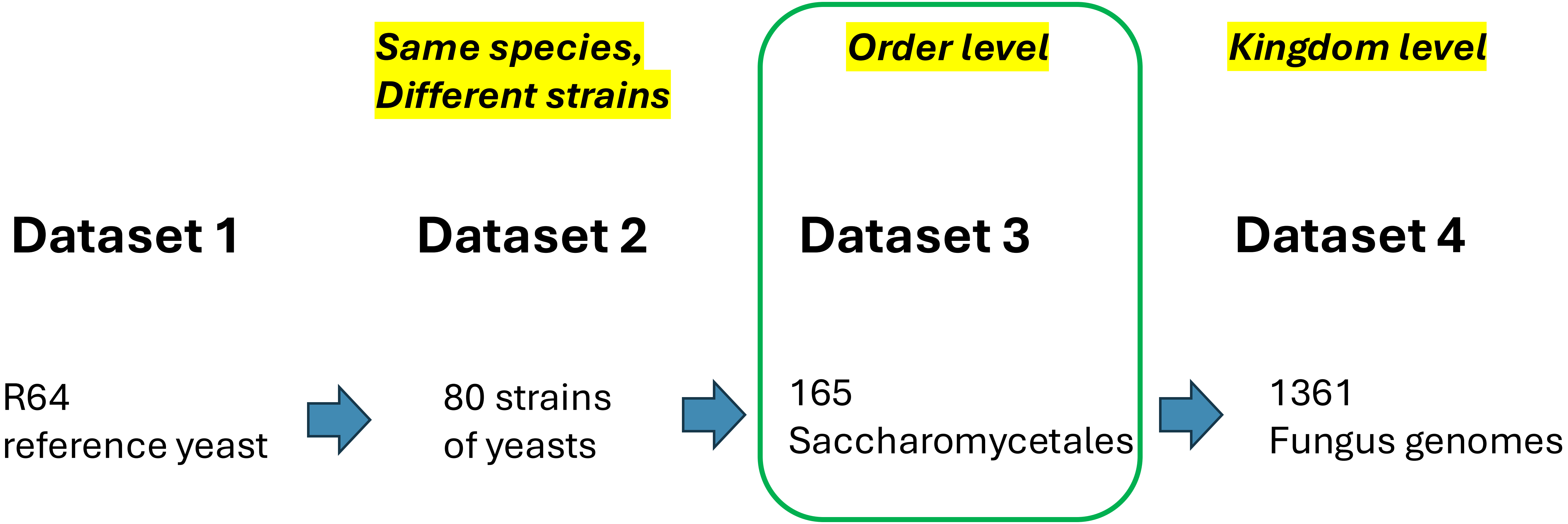
Validation
(chrXI, chrXIII, chrXV)

Testing
(chrXII, chrXIV, chrXVI)

Q1: To what evolutionary distance should we include in our LM?

Selected Genomes for LM

Fungi diverged from other life around 1.5 billion years ago



Same species, Different strains

Order level

Kingdom level

Dataset 1

R64 reference yeast

Dataset 2

80 strains of yeasts

Dataset 3

165 Saccharomycetales

Dataset 4

1361 Fungus genomes

Q1: Diversity of strains?

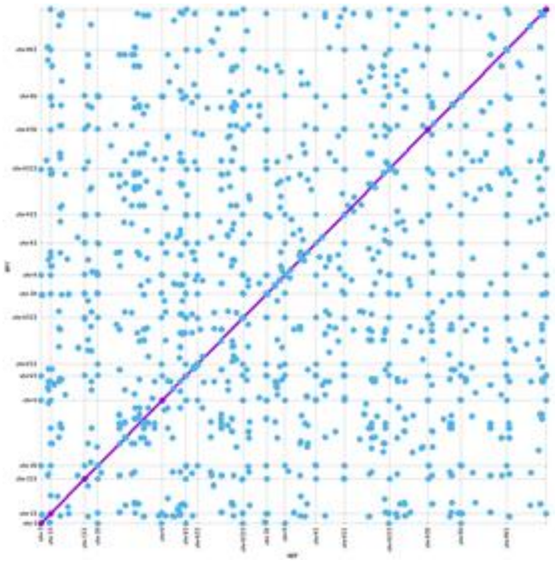
Q2: Diversity of species?

Q3: Even more diverse?

Genome distance evaluation

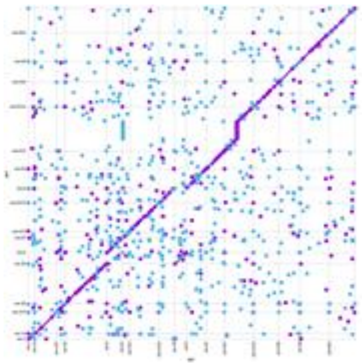
R64 Reference Yeast

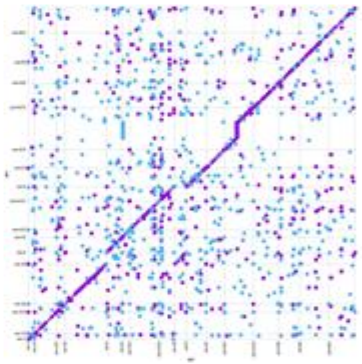
Mummerplot



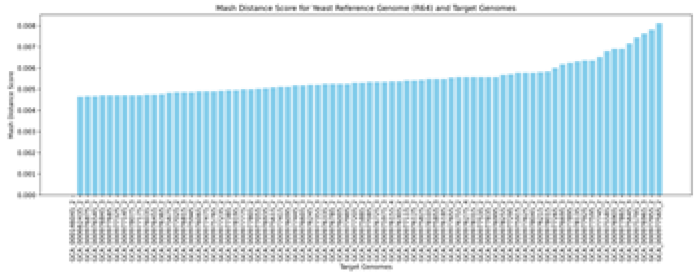
80 strains of yeasts

Mummerplot

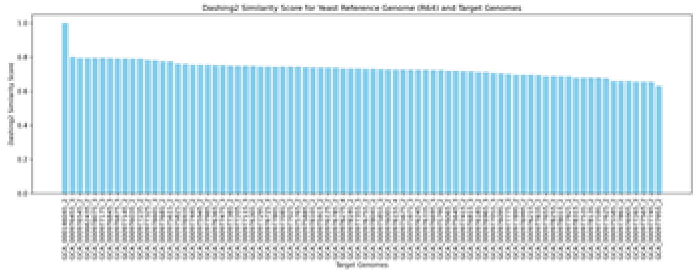




Mash distance

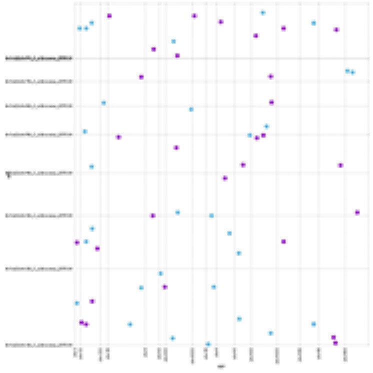


Dashing2 similarity

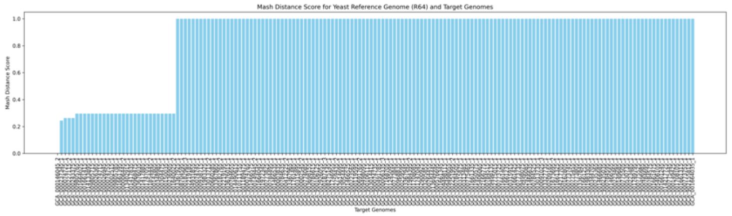


165 Saccharomycetales

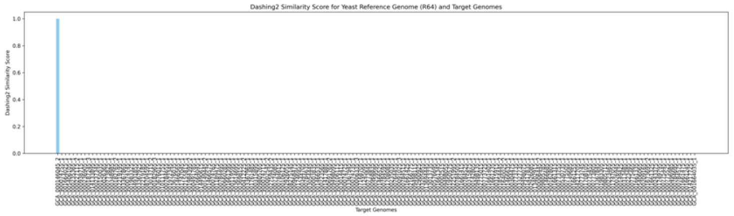
Mummerplot

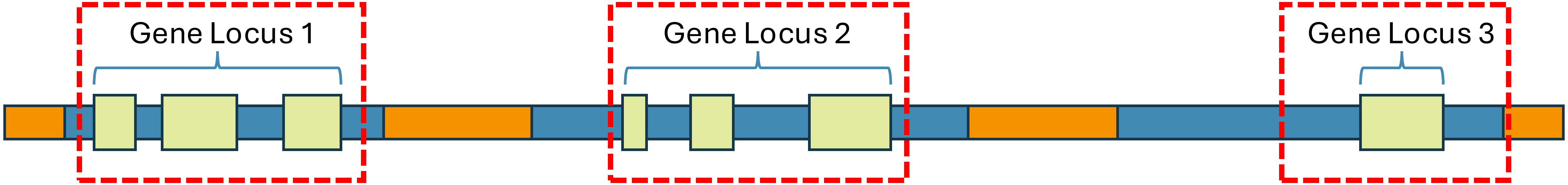


Mash distance



Dashing2 similarity

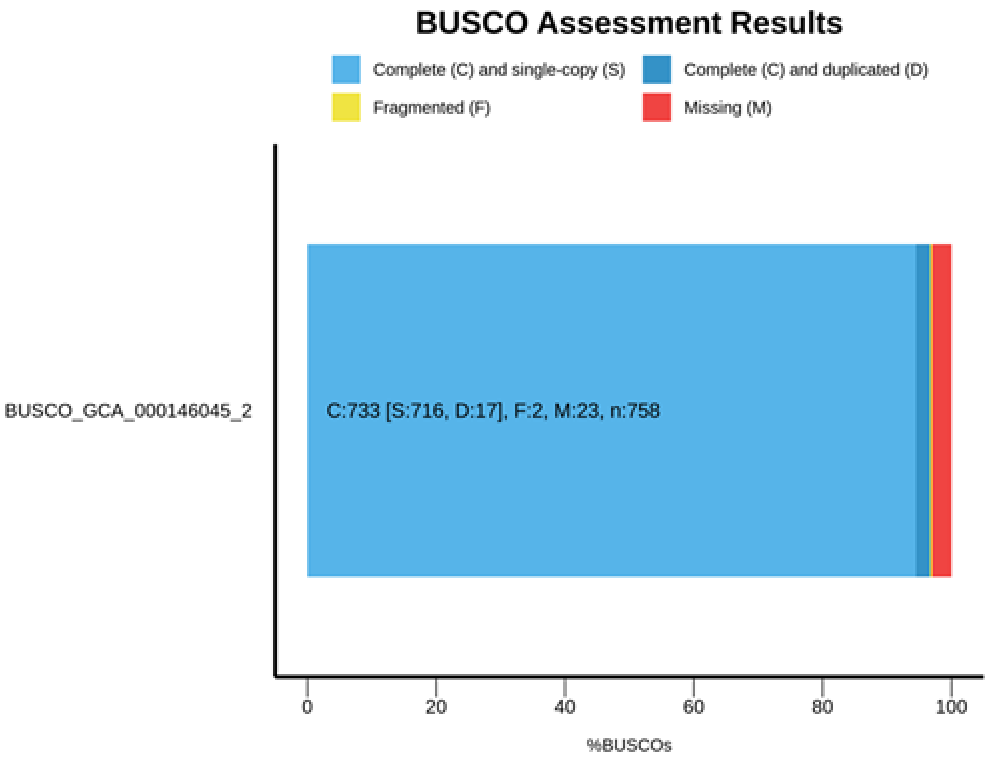




Q2: What is the quality of the annotation? Coding vs non-coding regions?

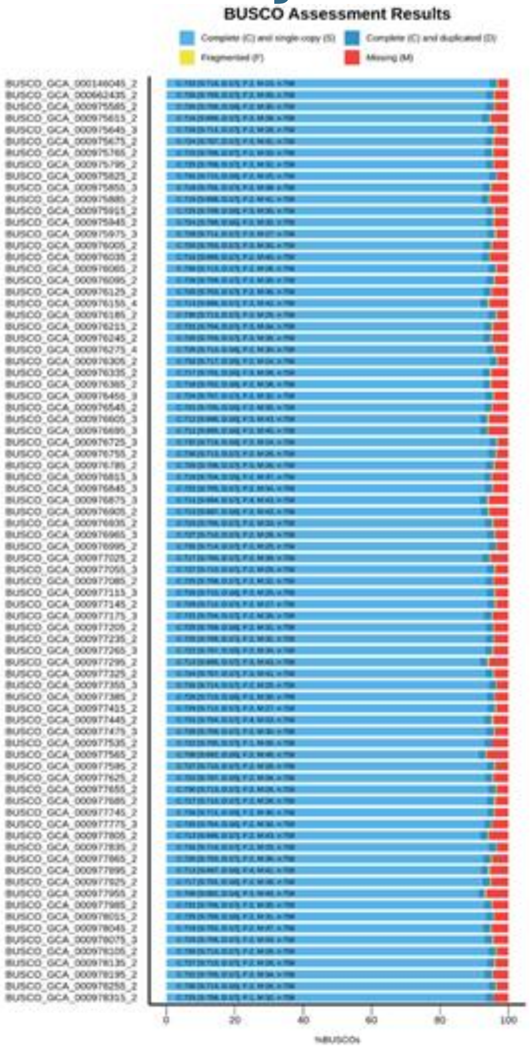
Genome annotation completeness evaluation

R64 Reference Yeast

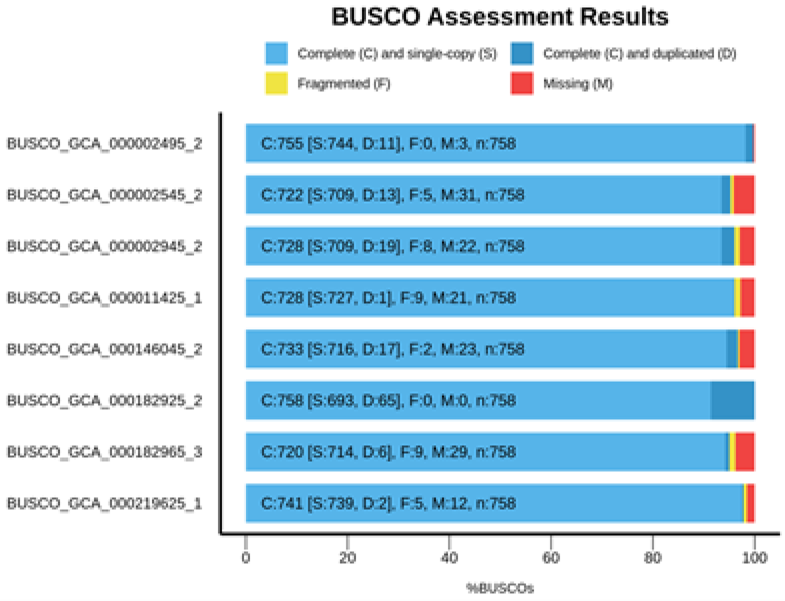


Conclusion:
~95% completeness

80 strains of yeasts



165 Sachramonycetales



Mosé Manni, Matthew R Berkeley, Mathieu Seppey, Felipe A Simão, Evgeny M Zdobnov, ***BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes.*** *Molecular Biology and Evolution, Volume 38, Issue 10, October 2021, Pages 4647–4654*

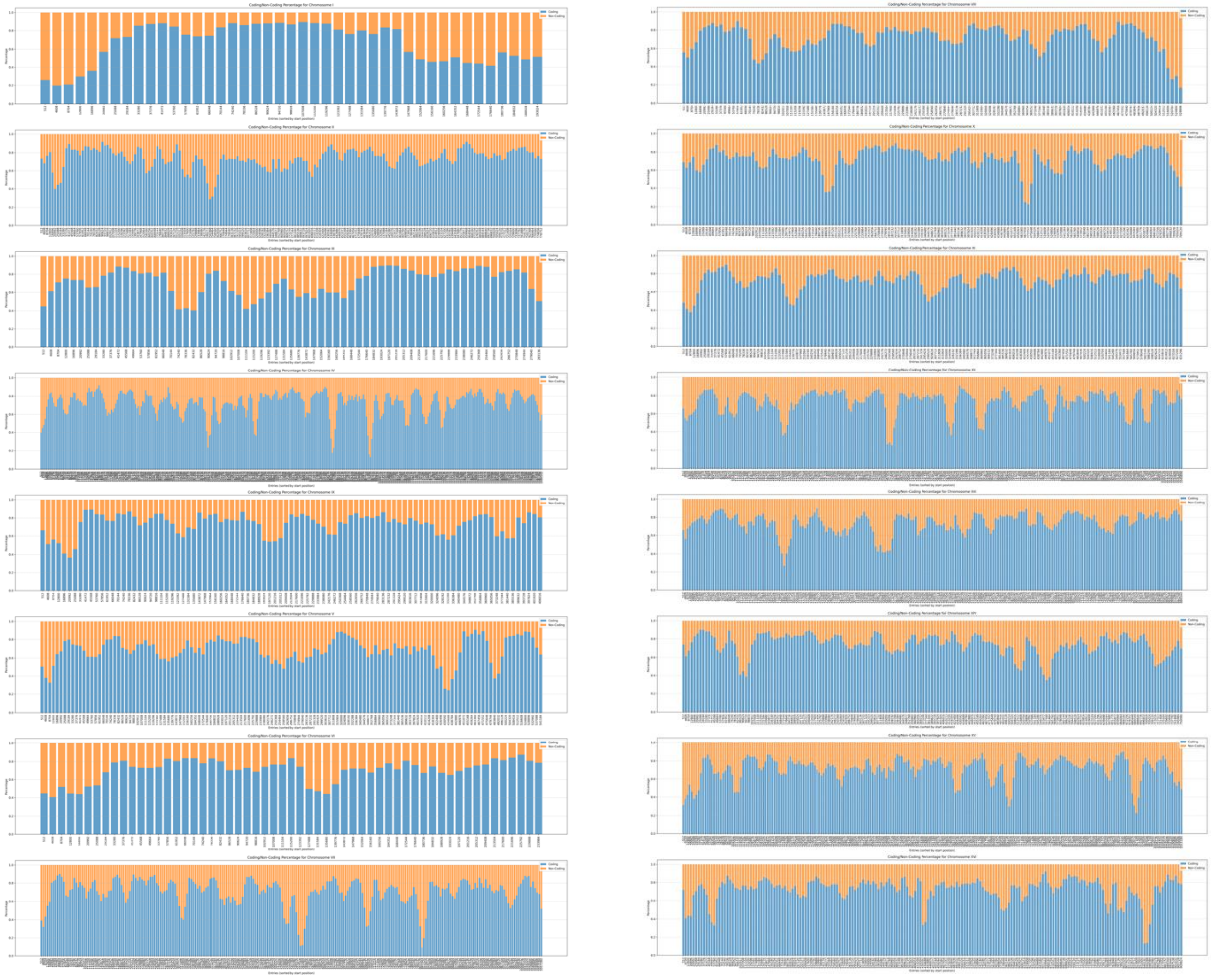
The following protocol covers the various BUSCO running modes and workflows, BUSCO setup, guidelines to interpret the results, and additional analyses, e.g., for building phylogenomic trees and visualizing syntenies using BUSCO results:

Manni, M., Berkeley, M. R., Seppey, M., & Zdobnov, E. M. (2021). ***BUSCO: Assessing genomic data quality and beyond.*** *Current Protocols, 1, e323. doi: 10.1002/cpz1.323*

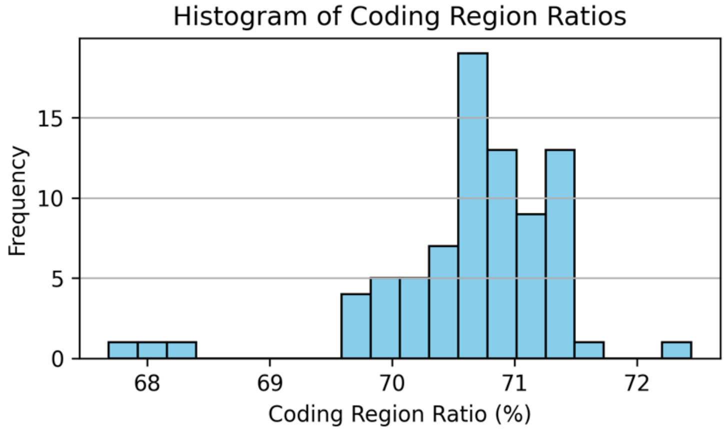
Genome evaluation – coding / noncoding regions

R64 Reference Yeast

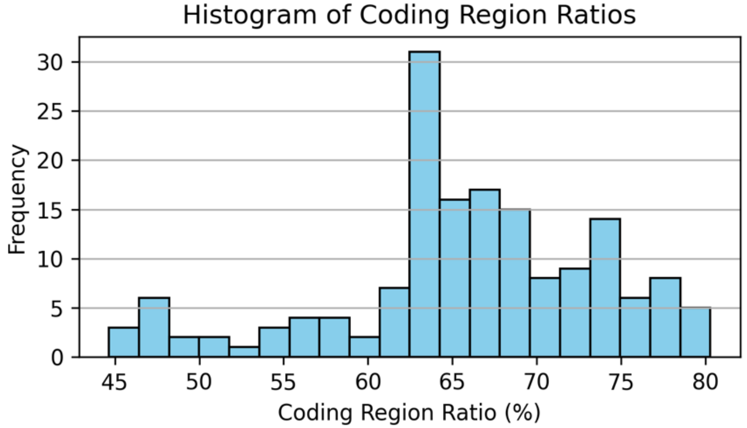
72.46% coding regions



80 strains of yeasts



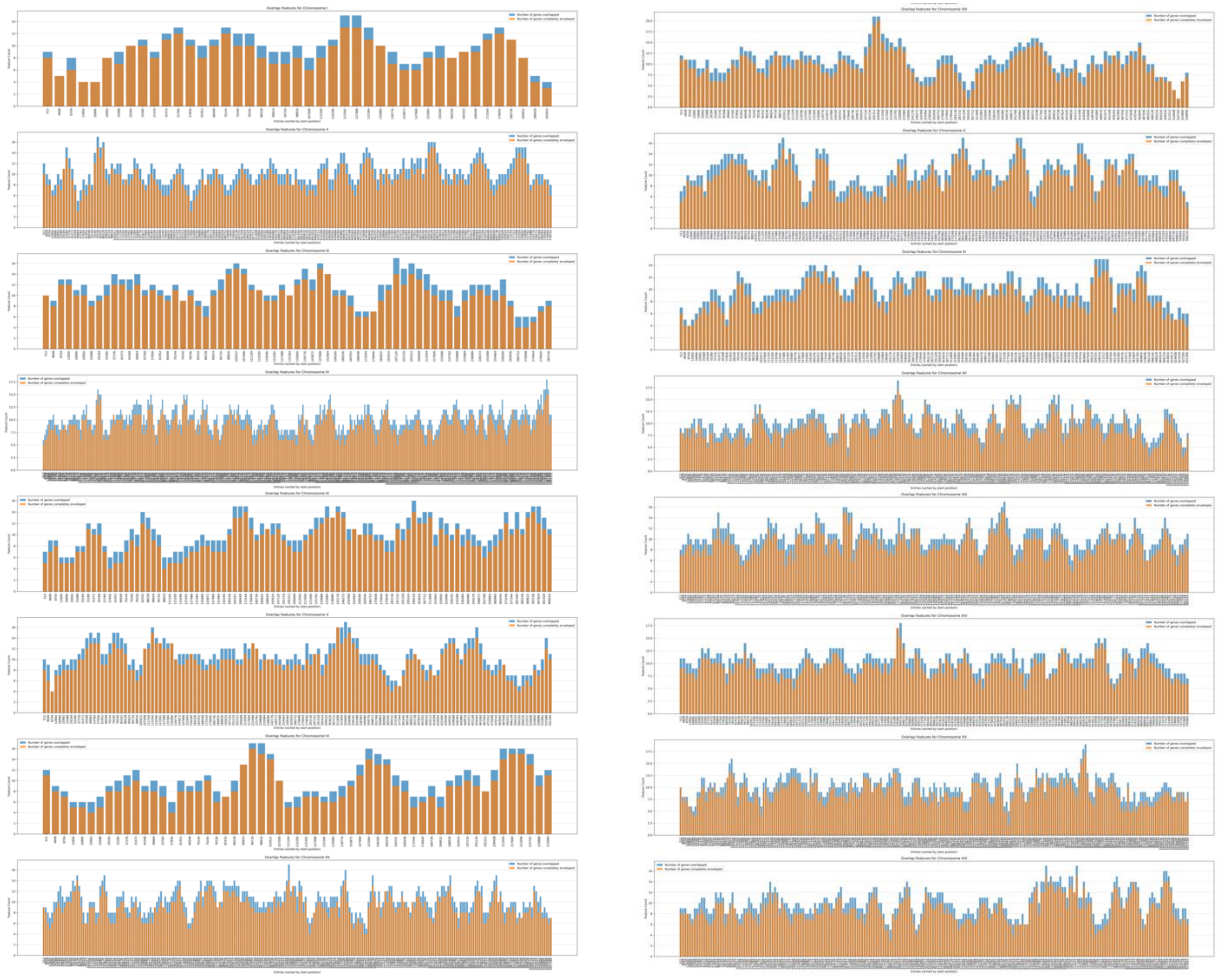
165 Sachramonycetales



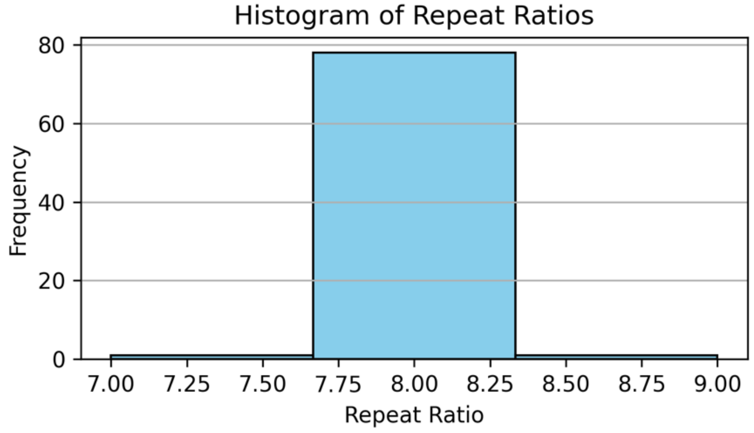
Genome evaluation – # genes per window

R64 Reference Yeast

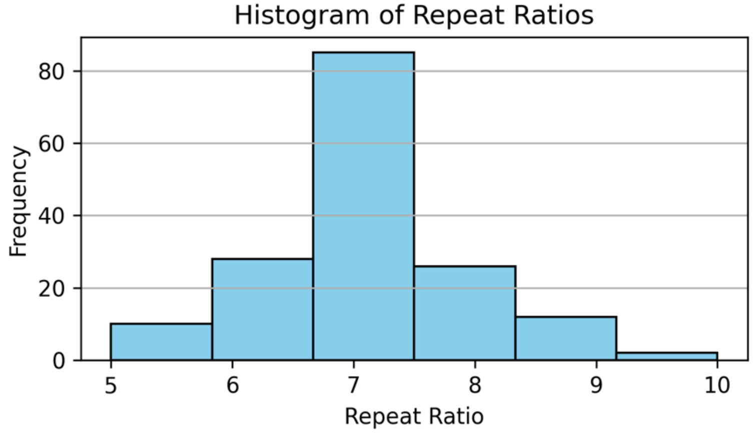
Median: 9.0; Mean: 8.98

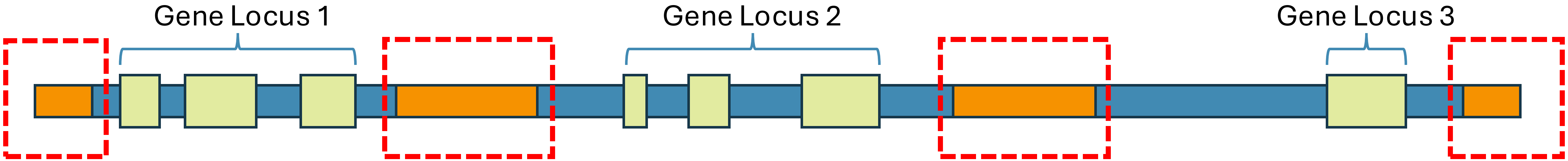


80 strains of yeasts

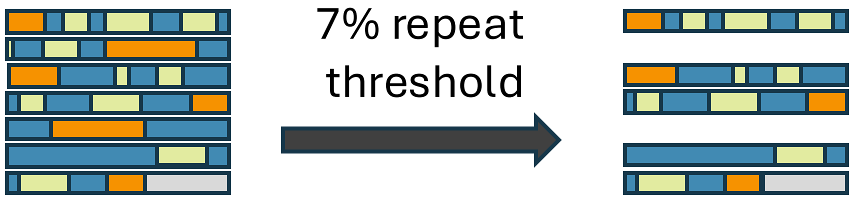


165 Sachramonycetales



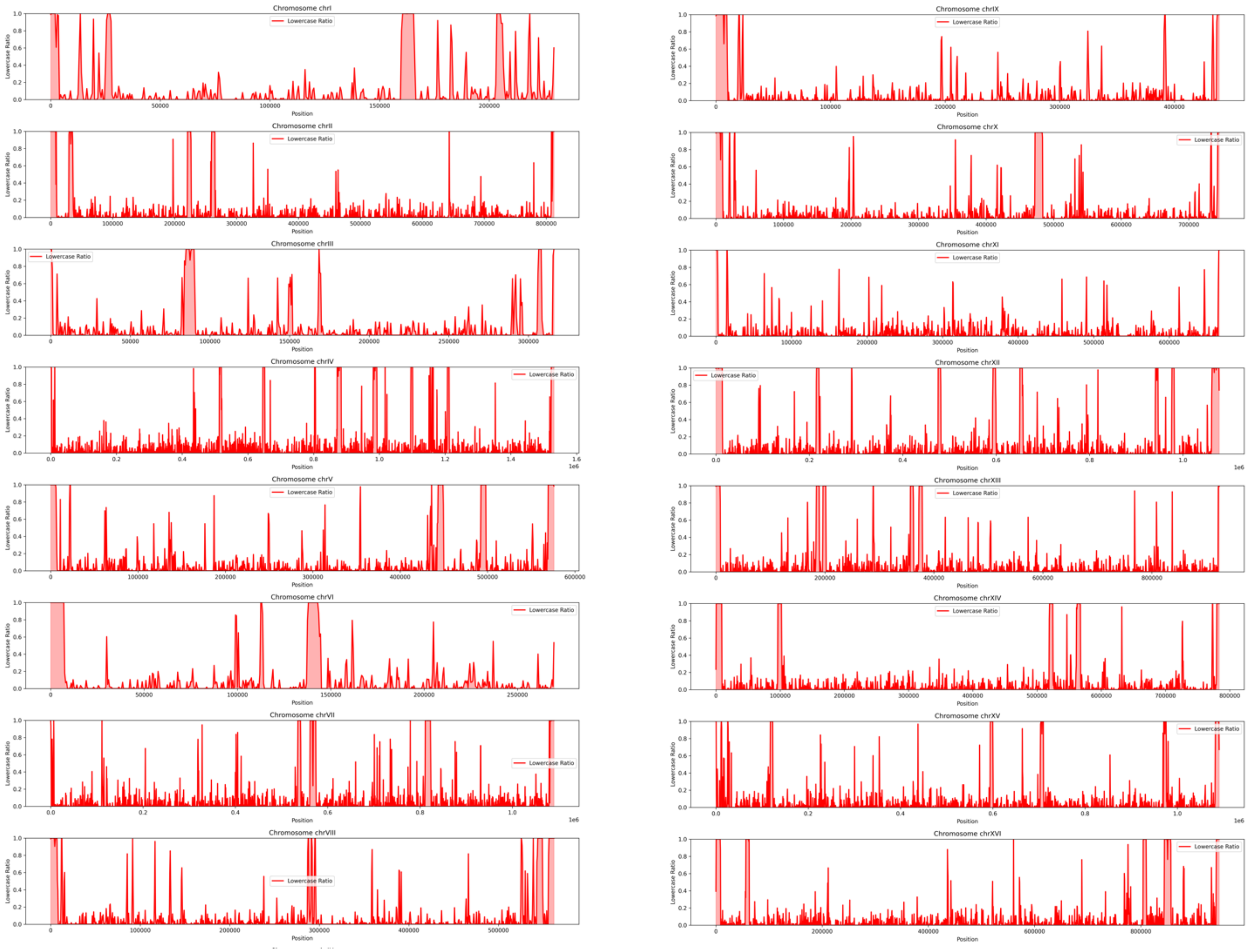


Q3: How repetitive are the genomes?

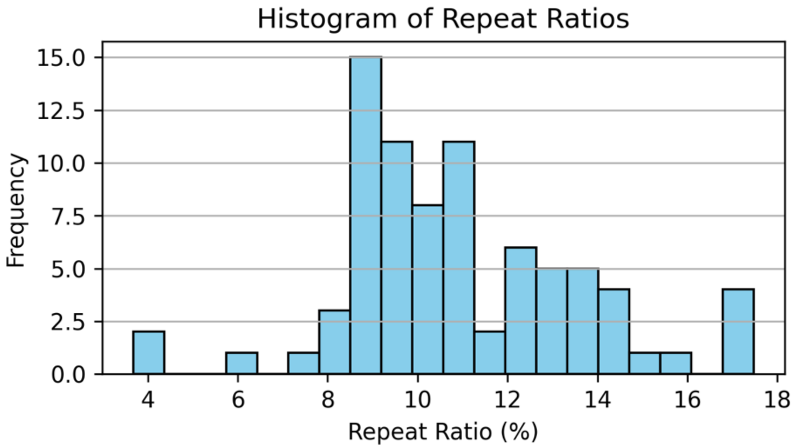


Genome evaluation – repeat regions

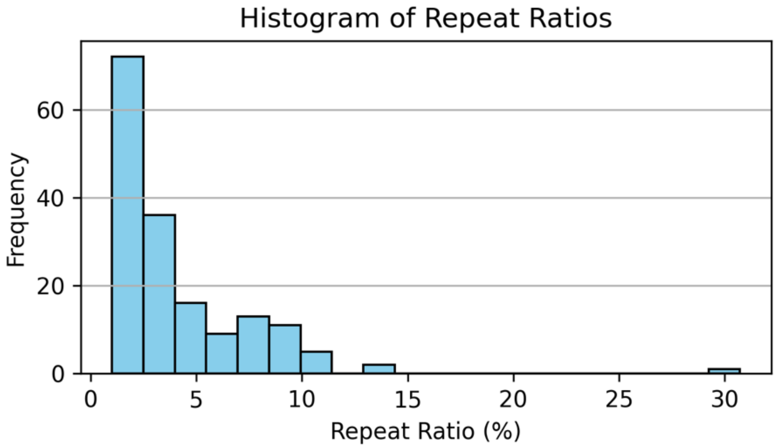
R64 Reference Yeast *7.39% repeat regions*



80 strains of yeasts



165 Sachramonvcetales



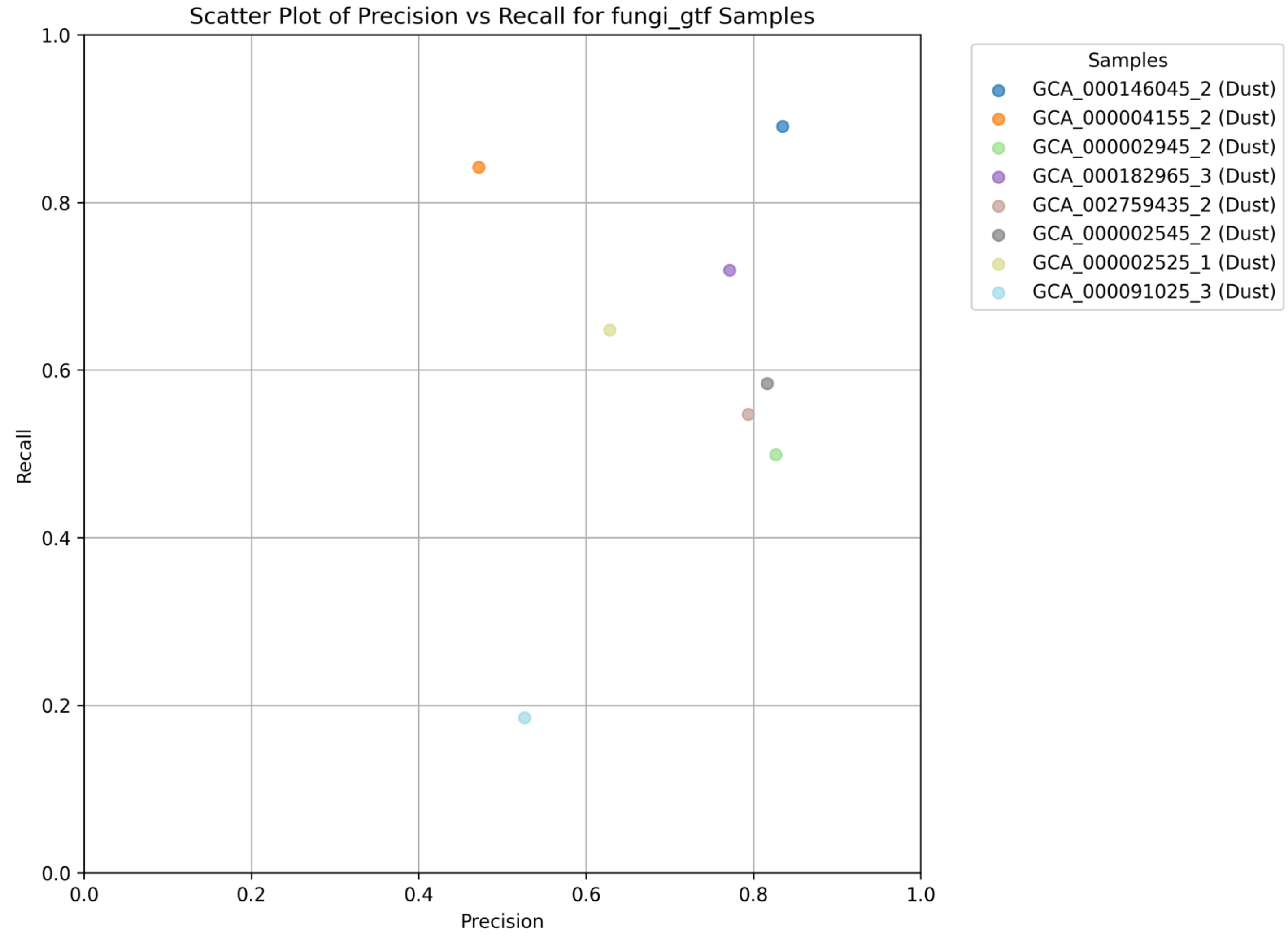
Repeats Detection

- **RepeatModeler**: Identifies de novo transposable element (TE) families.
  - BuildDatabase
  - RepeatModeler

- **RepeatMasker**: Screens DNA sequences for interspersed repeats and low complexity DNA sequences using Dfam (or RepBase, **30K**💰) database.

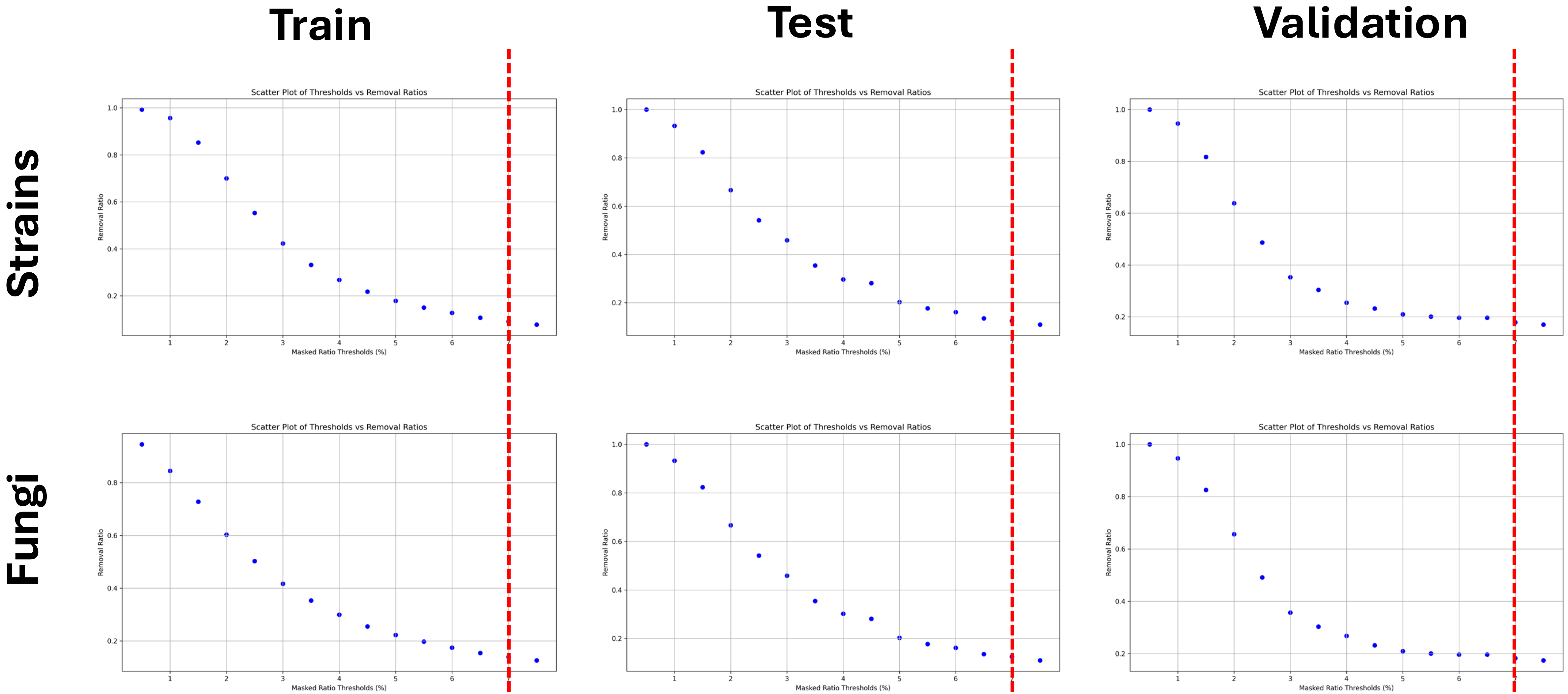
- **Dust**: Masks low-compexity regions

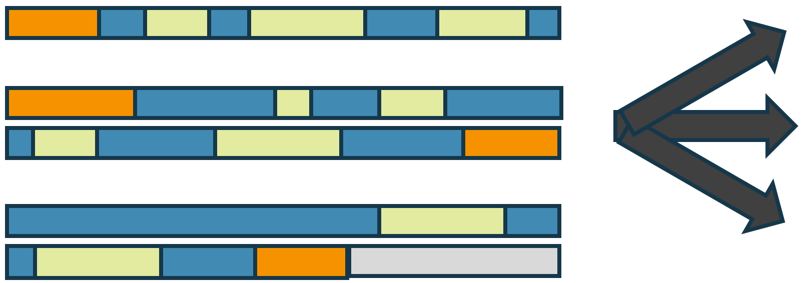
Repeats masking evaluation



a 7% threshold removes ~10% of the sequences.

Data cleaning – repeats removal



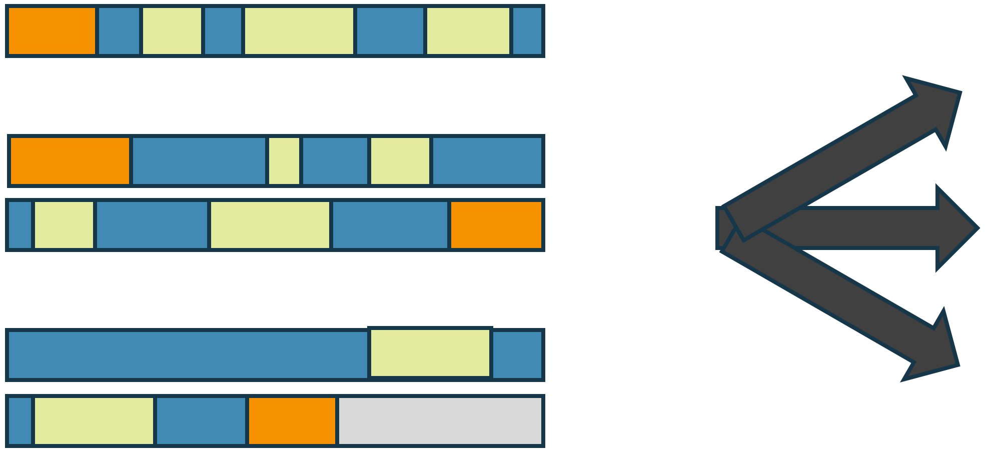


Training

Validation (chrXI, chrXIII, chrXV)

Testing (chrXII, chrXIV, chrXVI)

Q4: How many homologous sequences are there between training and testing?



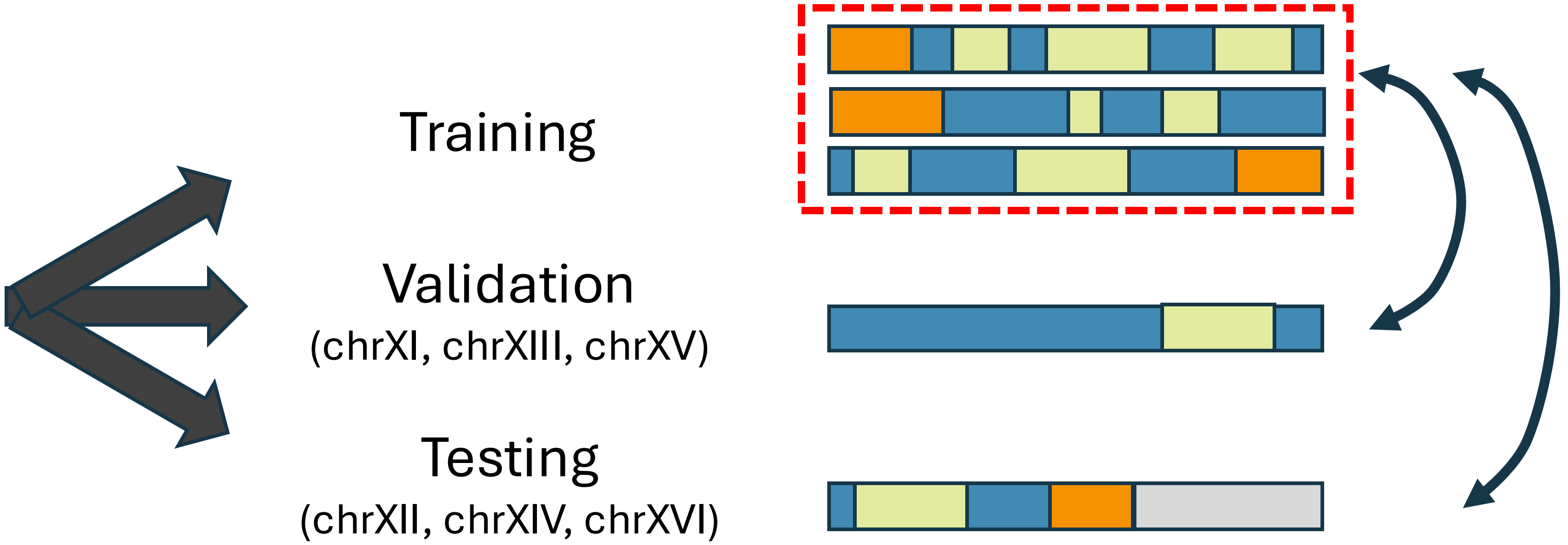
Training

Validation

(chrXI, chrXIII, chrXV)

Testing

(chrXII, chrXIV, chrXVI)



Training

Validation
(chrXI, chrXIII, chrXV)

Testing
(chrXII, chrXIV, chrXVI)

Detect homologous sequence using **DNA sequence aligner**

Homology sequence removal

- Nucmer:
  - minimum length of maximal exact matches (MEMs) (20) MEMs shorter than this length will be ignored.
  - A cluster is a group of MEMs that are close to each other and are used to build the alignment (65) Smaller clusters will be ignored

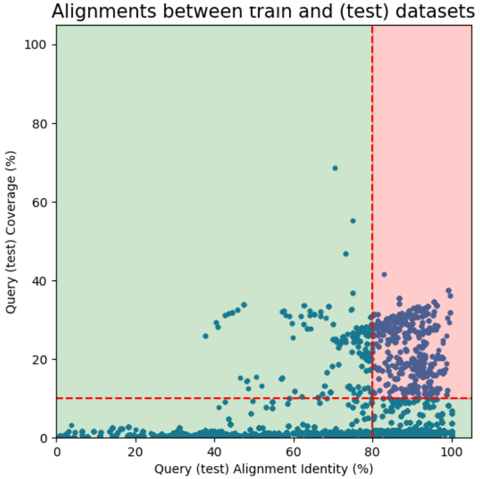
- Minimap2: minimap2 -x asm20
  - - asm5/asm10/asm20: - asm-to-ref mapping, for ~0.1/1/**<u>5%</u>** sequence divergence

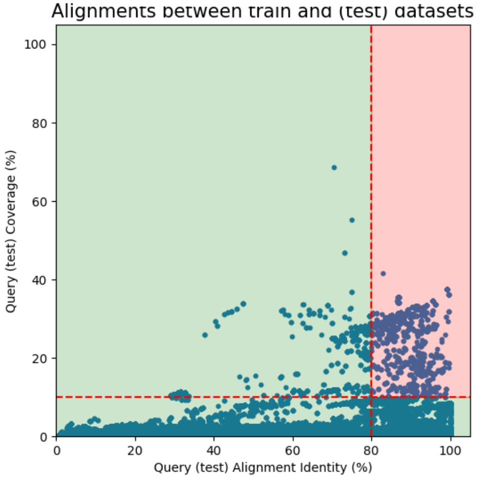
test_chrXI_chrXIII_chrXV / valid_chrXII_chrXIV_chrXVI / train remaining chromosomes

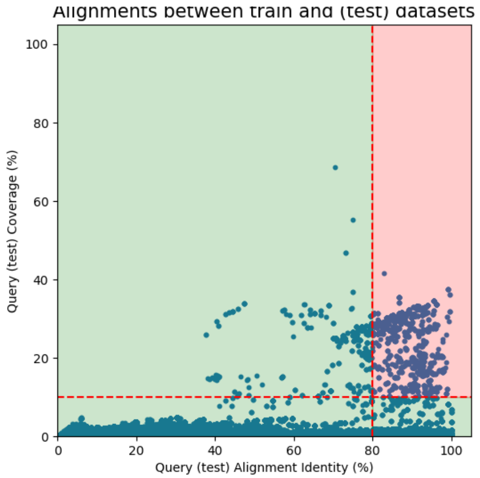
Homology sequence removal evaluation

r64 | **80 strains** | **165 Saccharomycetales**

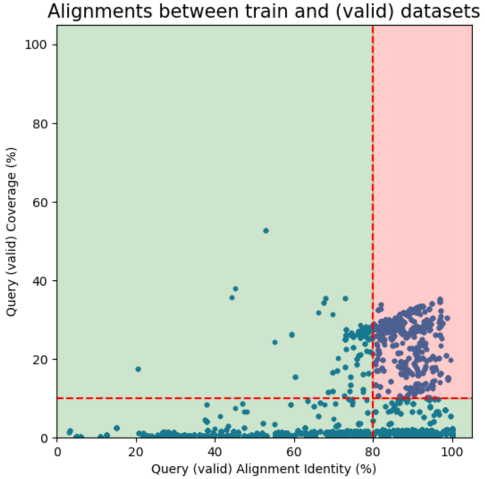
Train - Test

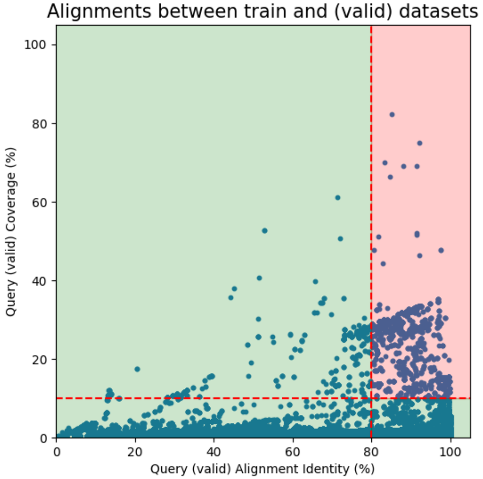

Alignments between train and (test) datasets

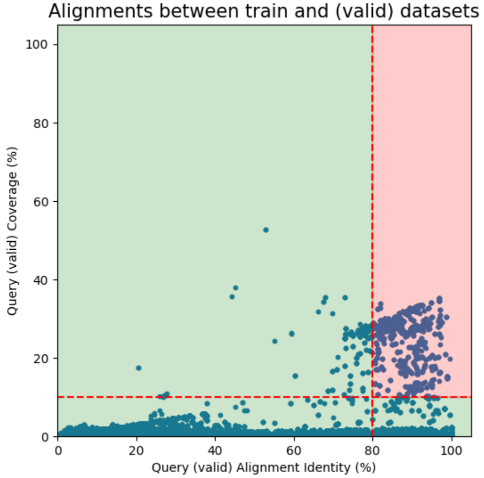

Alignments between train and (test) datasets


Alignments between train and (test) datasets

Train - Validation


Alignments between train and (valid) datasets


Alignments between train and (valid) datasets

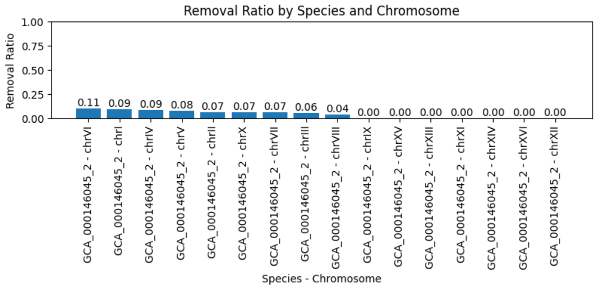

Alignments between train and (valid) datasets

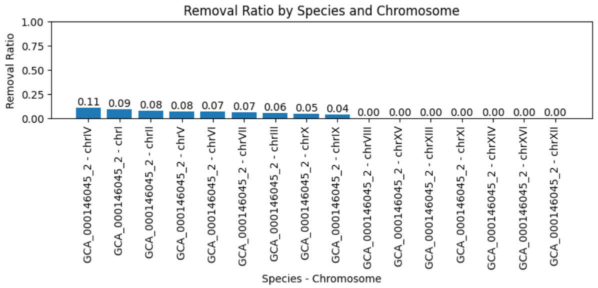
Homology sequence removal evaluation (minimap2)

Train - Test

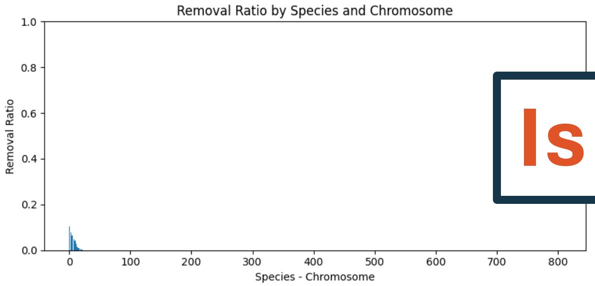
Train - Validation

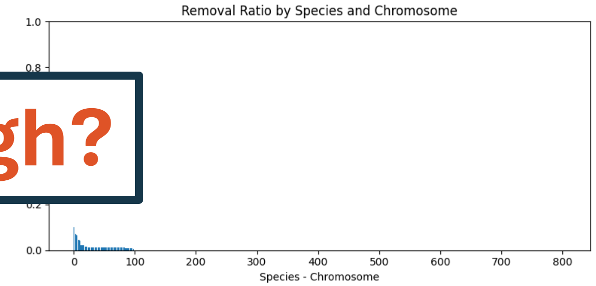
r64


Removal Ratio by Species and Chromosome


Removal Ratio by Species and Chromosome

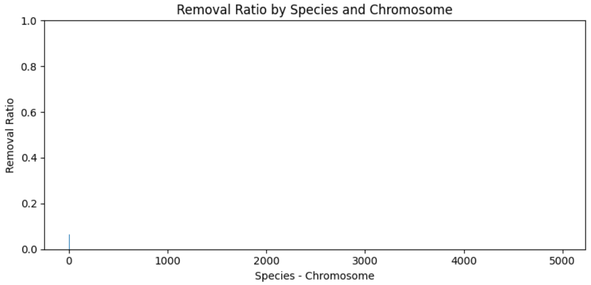
Strains

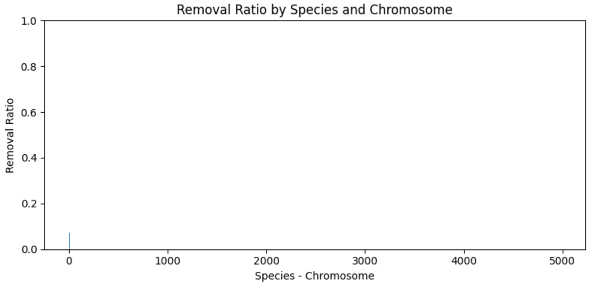

Removal Ratio by Species and Chromosome


Removal Ratio by Species and Chromosome

Is it good enough?

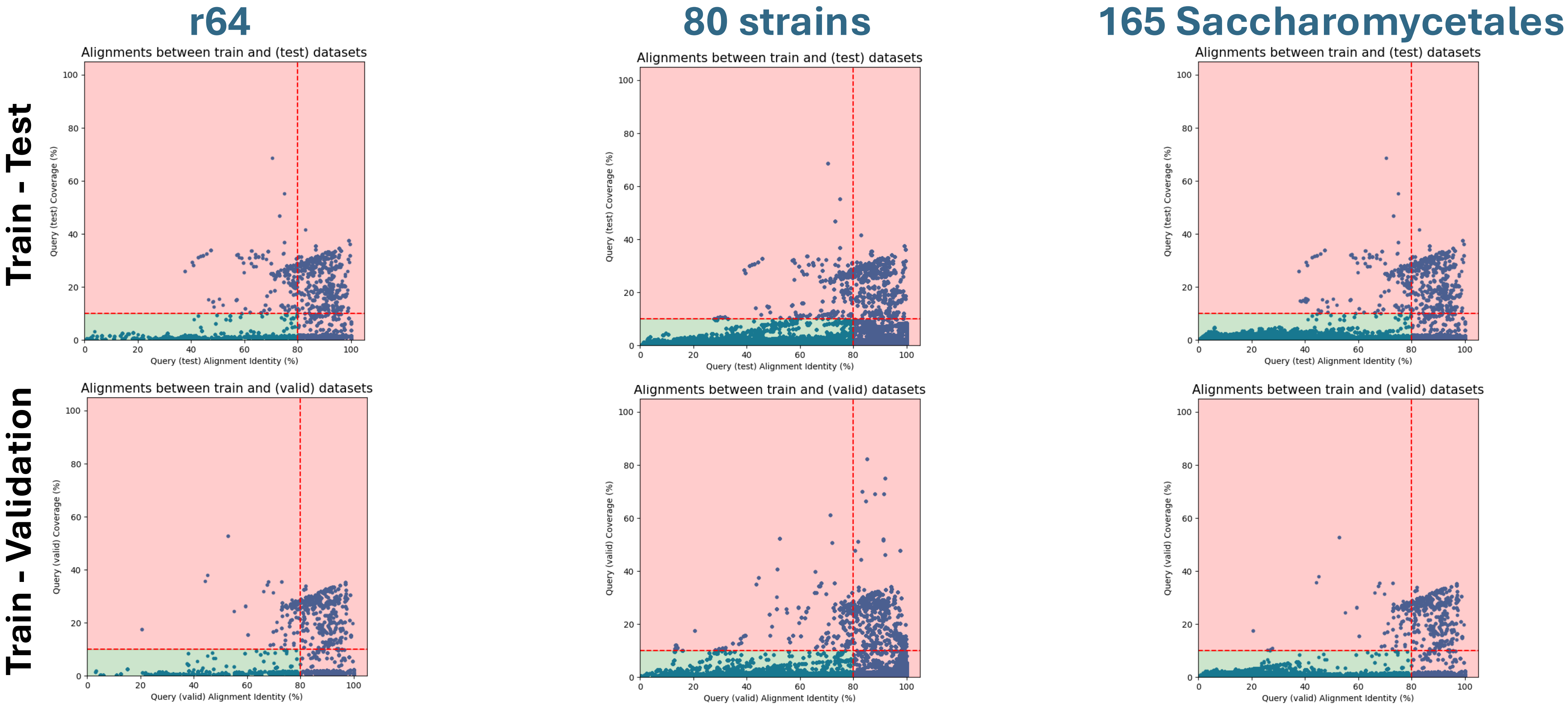
Saccharomycetales


Removal Ratio by Species and Chromosome


Removal Ratio by Species and Chromosome

test_chrXI_chrXIII_chrXV / valid_chrXII_chrXIV_chrXVI / train remaining chromosomes

Homology sequence removal evaluation


r64

80 strains

165 Saccharomycetales

Train - Test

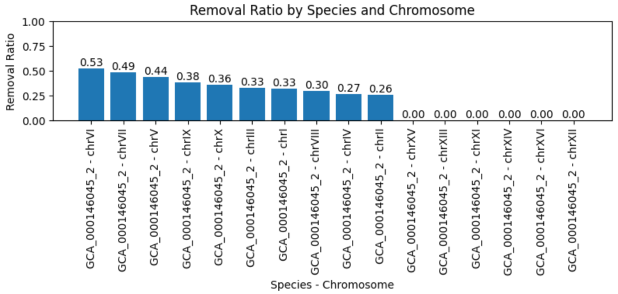
Train - Validation

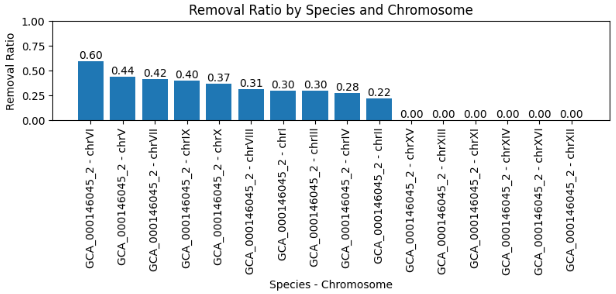
Homology sequence removal evaluation (minimap2)

Train - Test

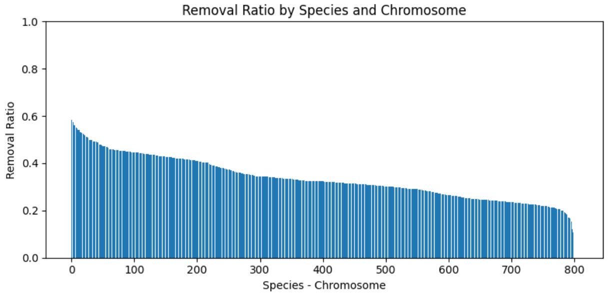
Train - Validation

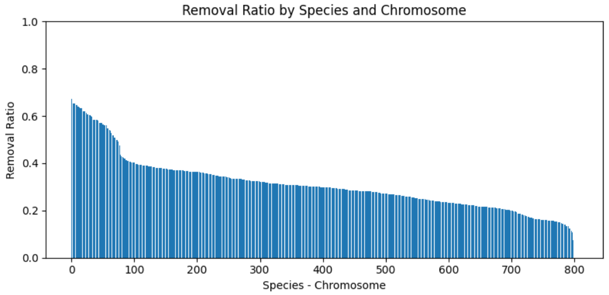
r64


Removal Ratio by Species and Chromosome

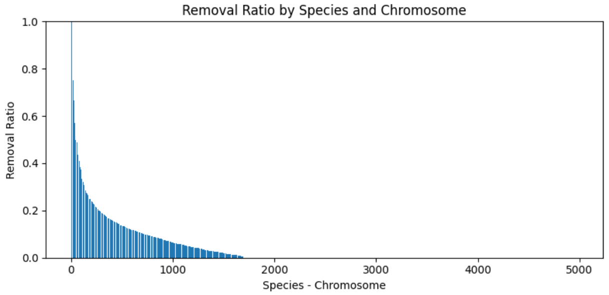

Removal Ratio by Species and Chromosome

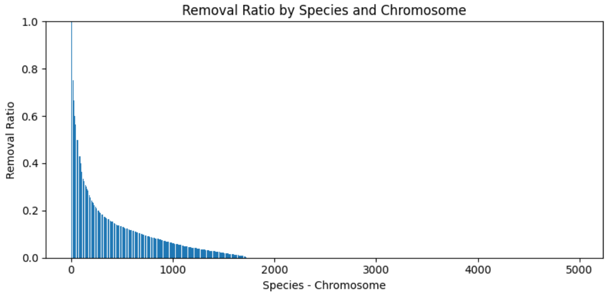
Strains


Removal Ratio by Species and Chromosome


Removal Ratio by Species and Chromosome

Saccharomycetales


Removal Ratio by Species and Chromosome


Removal Ratio by Species and Chromosome

Final sequence for training / testing / validation

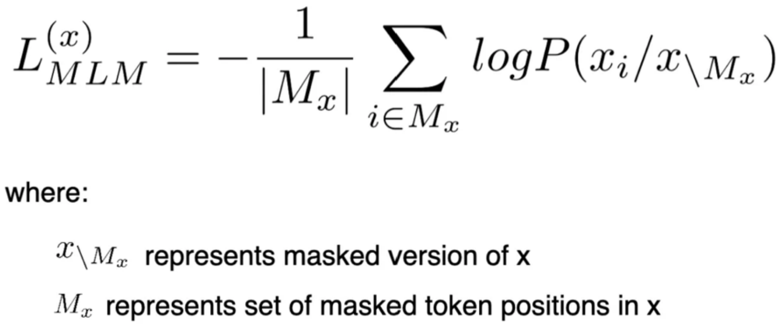
		r64	80 strains	165 Saccharomycetales
Before cleaning	Train	1440	108960	404608
	Test	608	608	608
	Validation	576	576	576
		-597 (-41.4%)	-40447 (-37.1%)	-65442 (-16.2%)
After cleaning	Train	843	68513	339166
	Test	507	507	507
	Validation	488	488	488

Fungi Language Model Architecture

Different model architecture we've tried

• Dilated convolutional neural network (small)	Total params: 320,708 (1.22 MB)
• Dilated convolutional neural network (large)	Total params: 3,642,116 (13.89 MB)
• Transformer-based unet (small)	Total params: 13,665,828 (52.13 MB)
• Transformer-based unet (large)	Total params: 71,790,564 (273.86 MB)

Masked language modeling loss

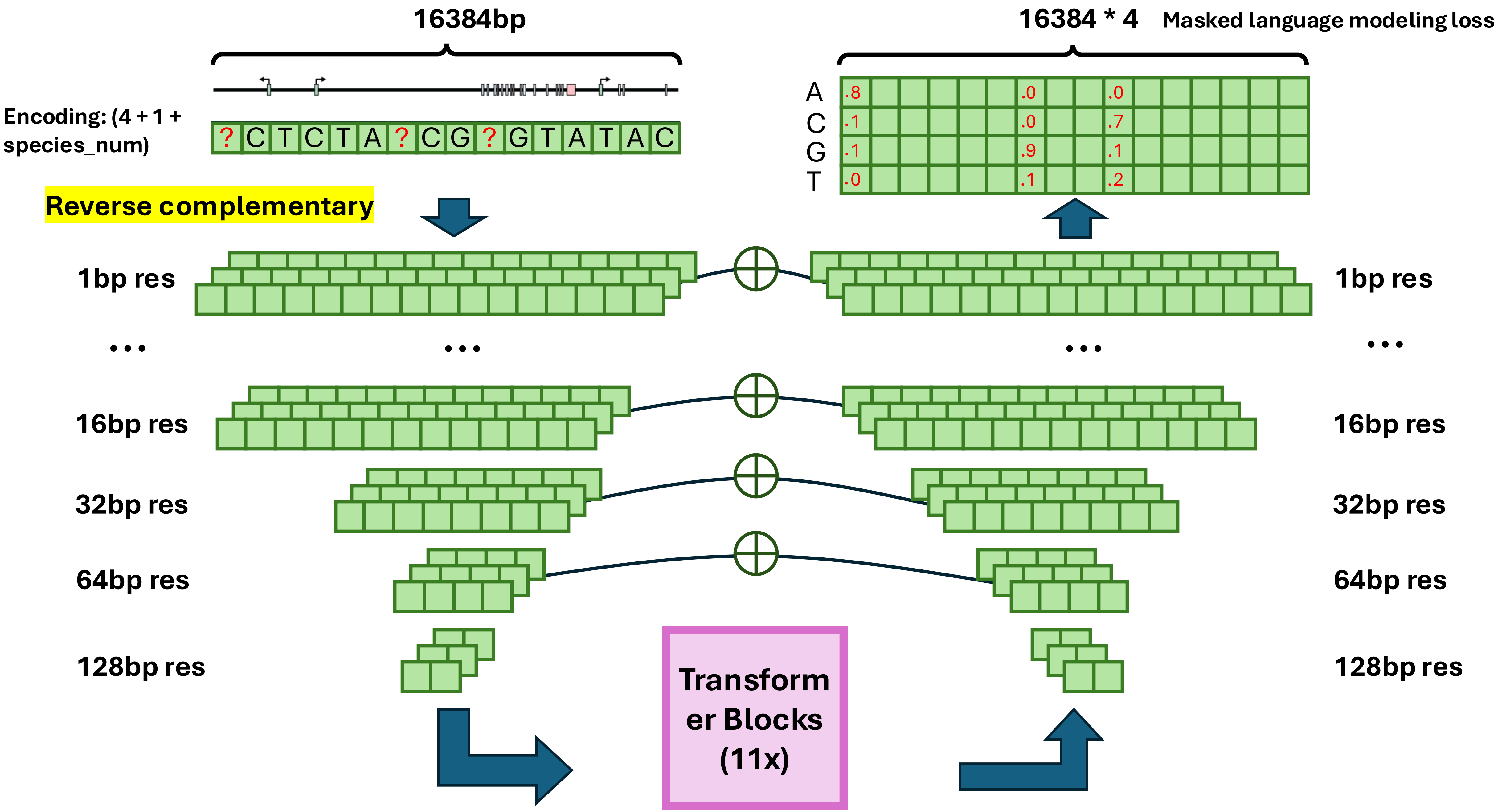


$$L_{MLM}^{(x)} = -\frac{1}{|M_x|}\sum_{i \in M_x} logP(x_i/x_{\backslash M_x})$$

where:

$x_{\backslash M_x}$ represents masked version of x

M_x represents set of masked token positions in x

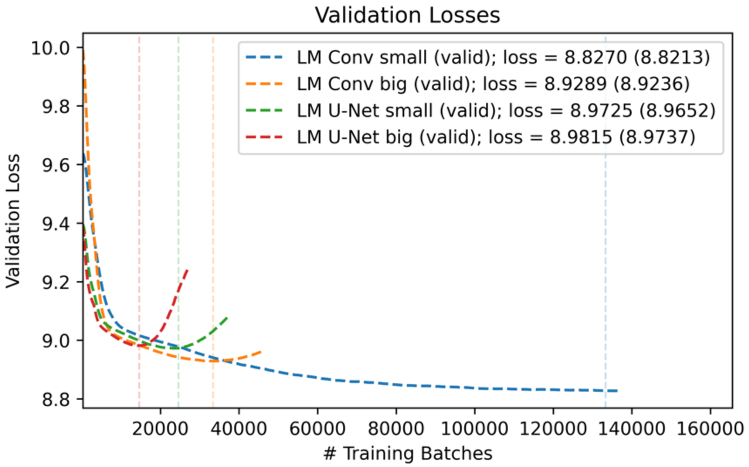

16384bp
Encoding: (4 + 1 + species_num)
? C T C T A ? C G ? G T A T A C
Reverse complementary
16384 * 4
Masked language modeling loss
A .8 .0 .0
C .1 .0 .7
G .1 .9 .1
T .0 .1 .2
1bp res
…
16bp res
32bp res
64bp res
128bp res
Transformer Blocks (11x)
1bp res
…
16bp res
32bp res
64bp res
128bp res

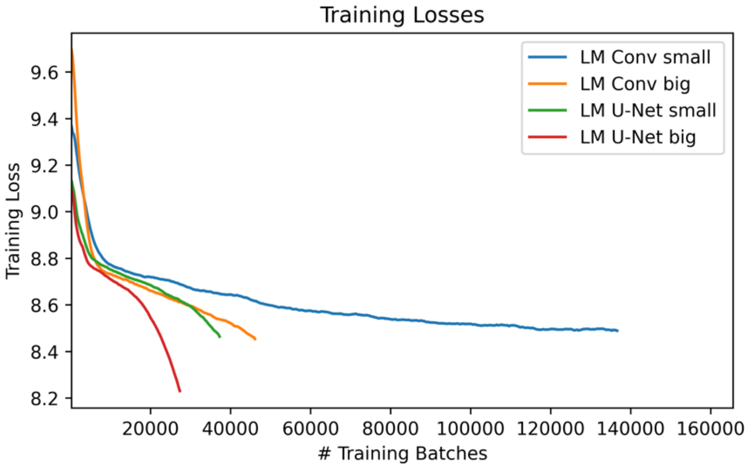
Self-supervised Fungi LM

Language Model Results

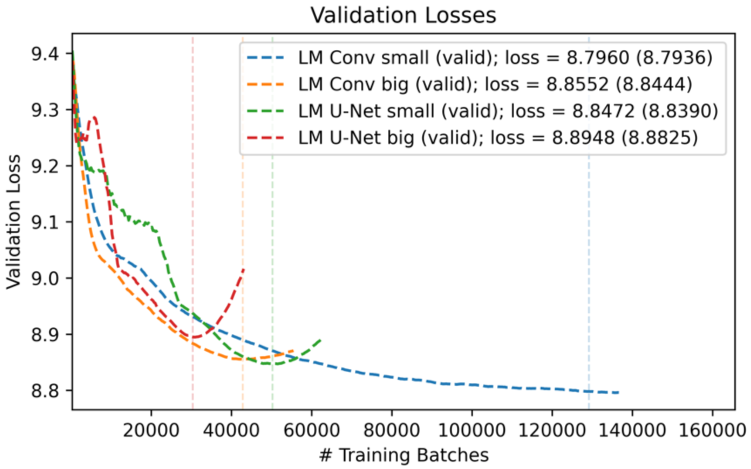
Model comparison

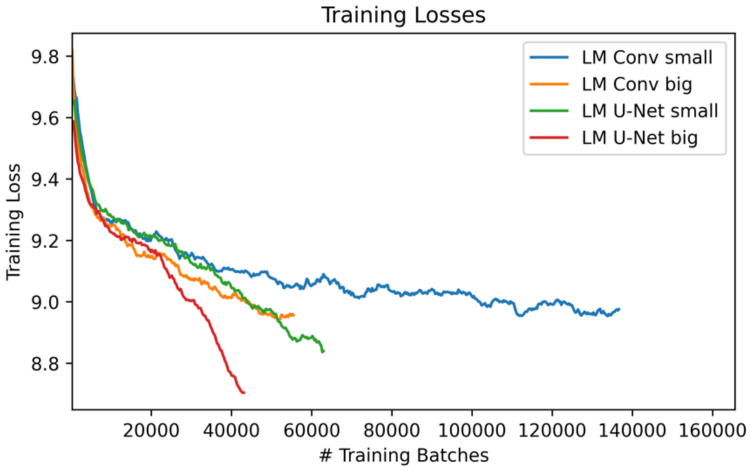
r64



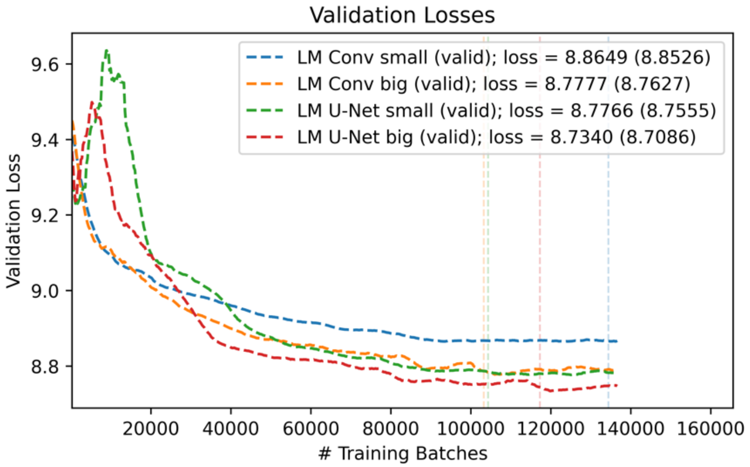


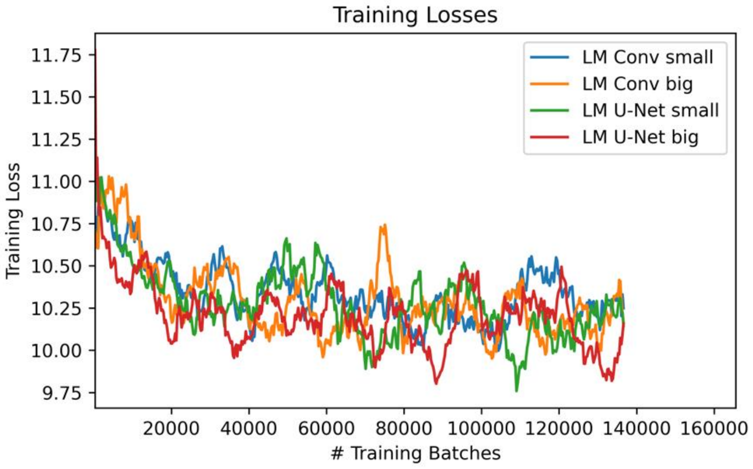
80 strains



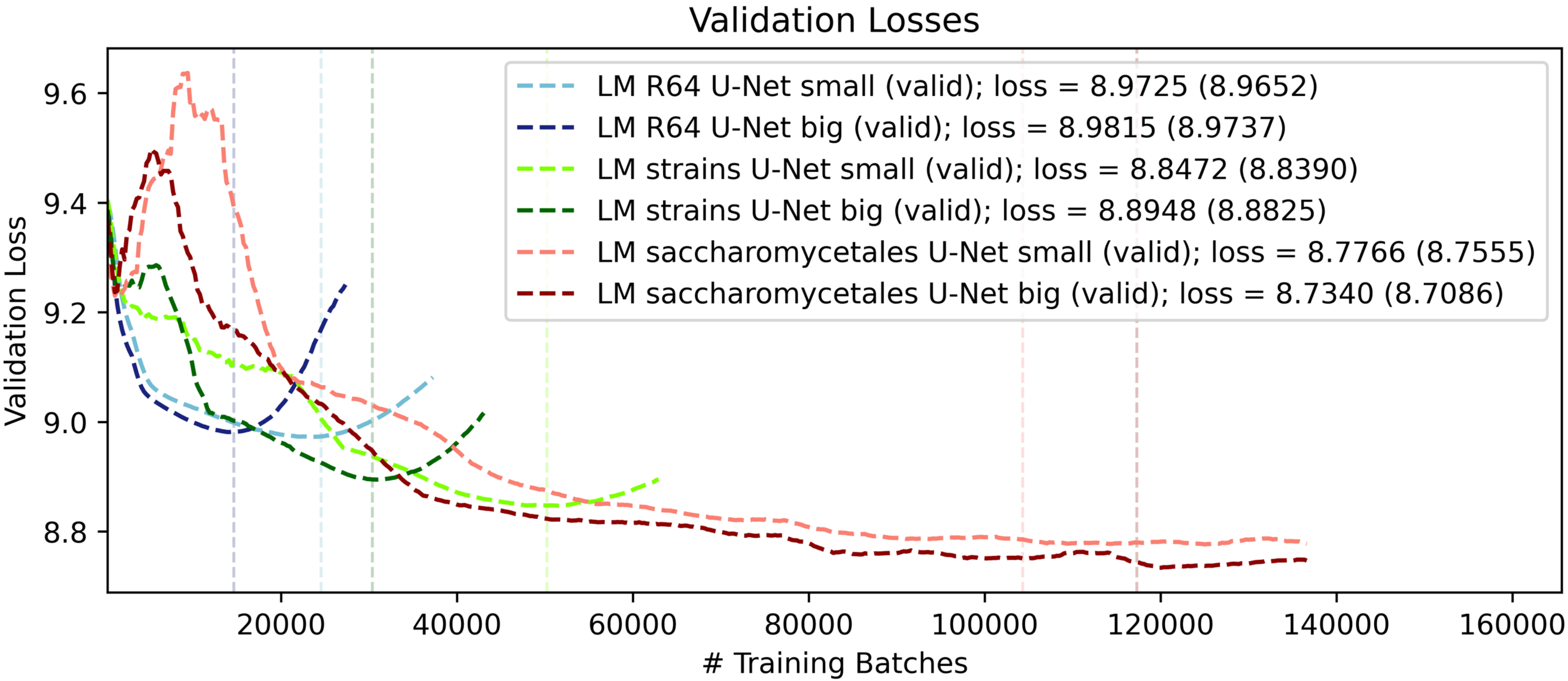


165 Saccharomycetales

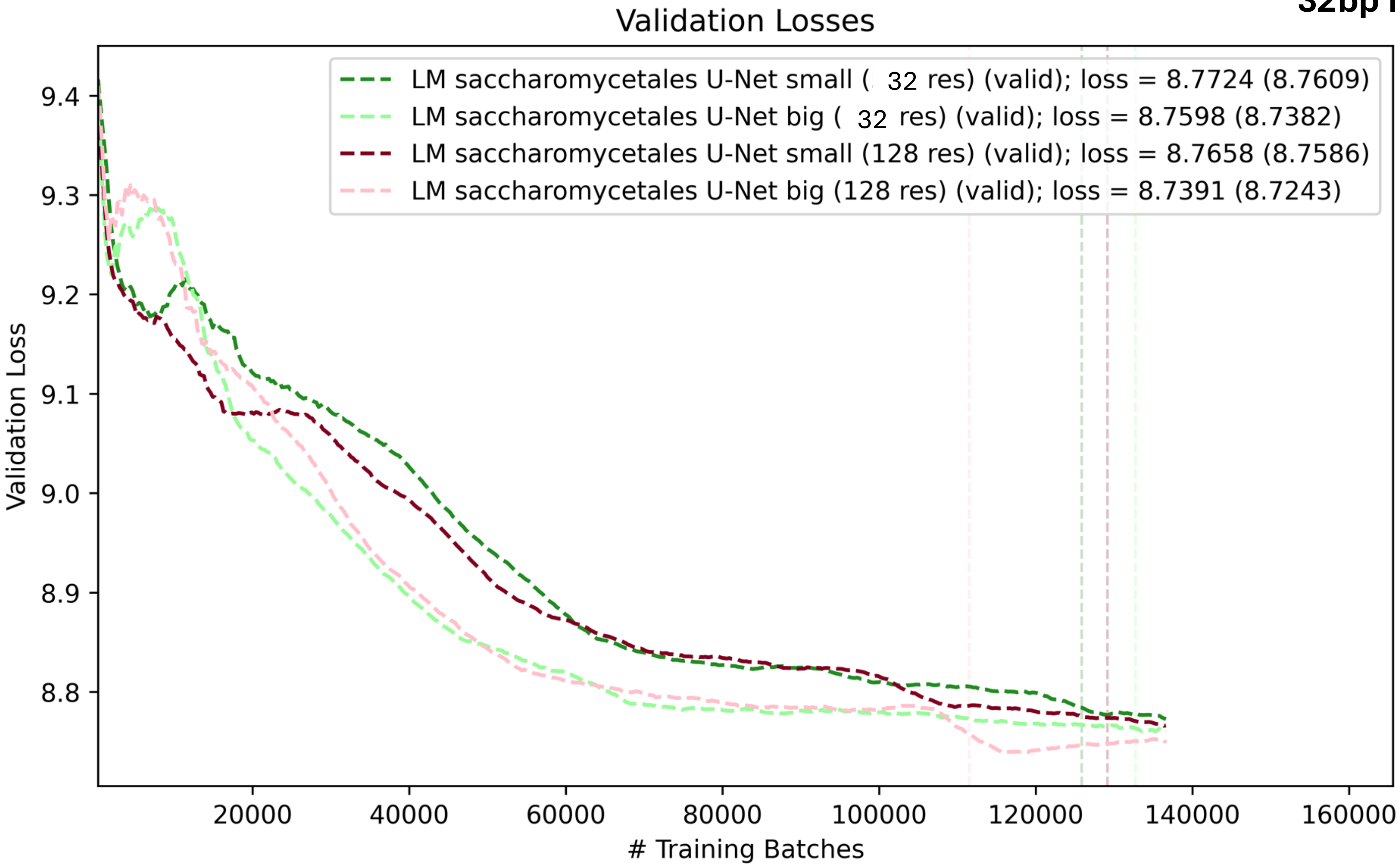


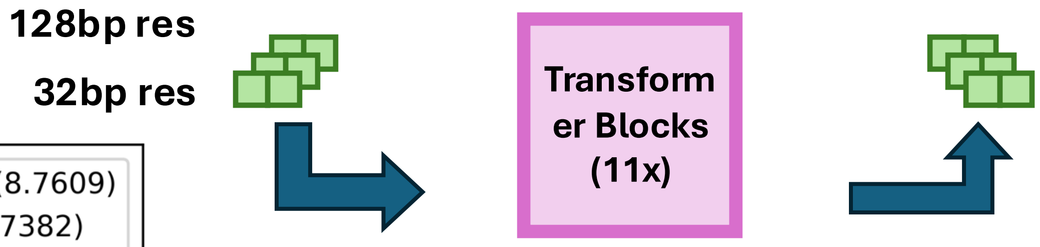


Dataset comparison



Different resolutions of input to transformer blocks





Both resolutions reach the similar loss

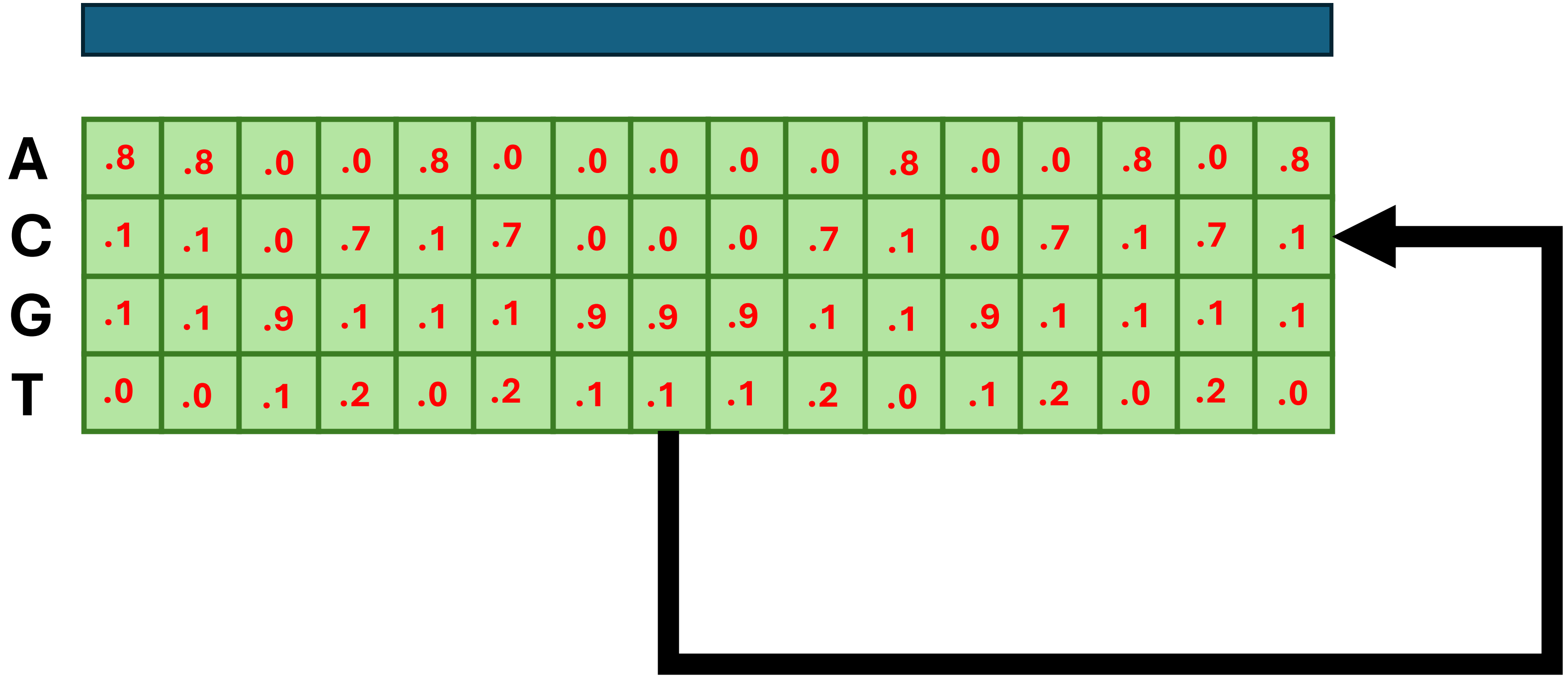
Fungi LM Language Model

Motif inference

Constructing PWM from Fungi LM

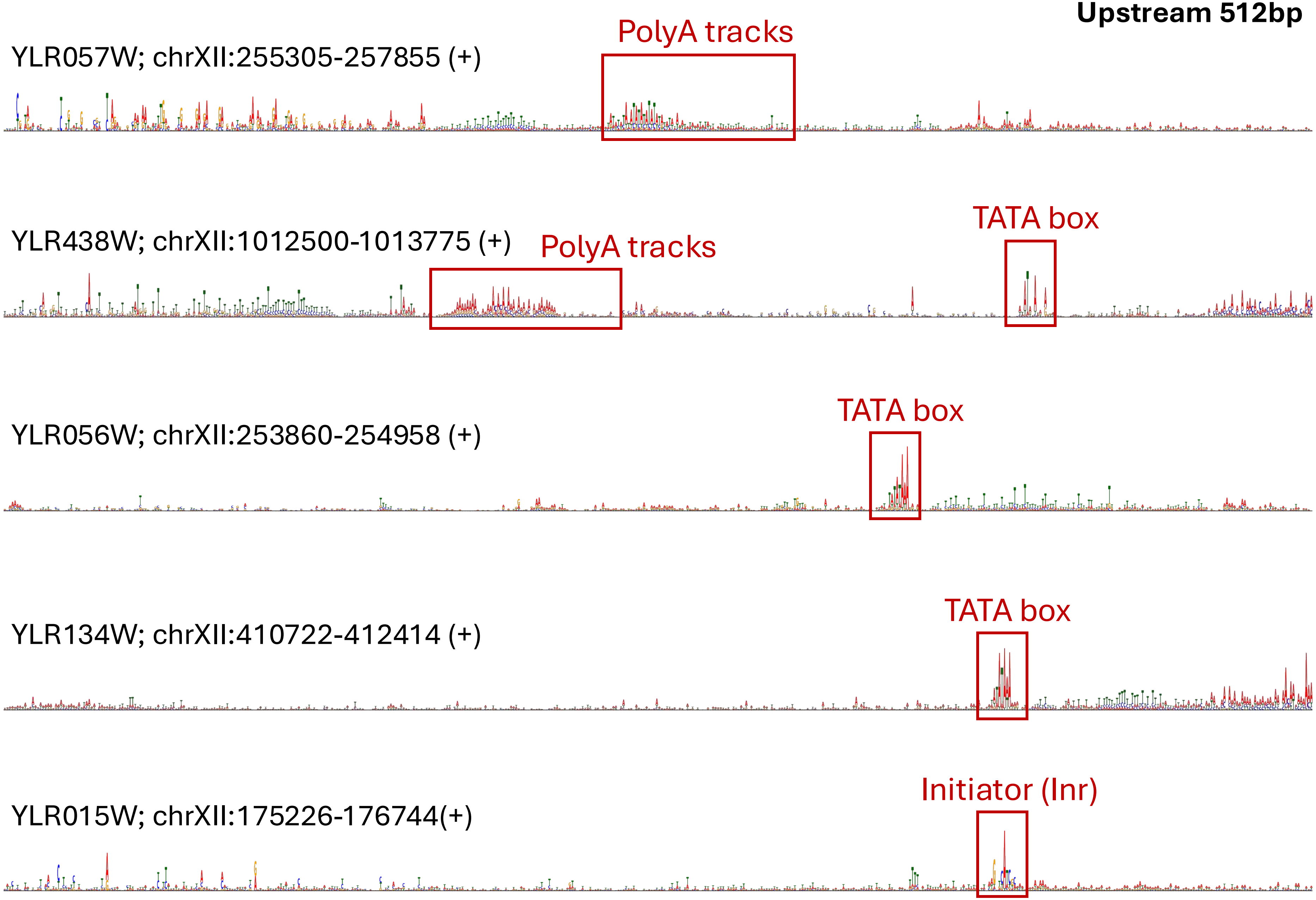
Predicting 15 % masked regions for each iteration

Testing
(chrXII, chrXIV, chrXVI)



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

Upstream 512bp



YLR057W; chrXII:255305-257855 (+)

YLR438W; chrXII:1012500-1013775 (+)

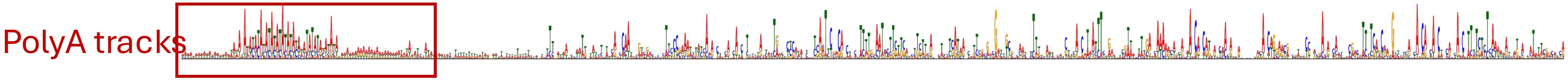
YLR056W; chrXII:253860-254958 (+)

YLR134W; chrXII:410722-412414 (+)

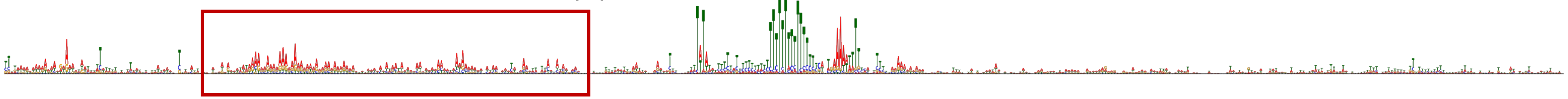
YLR015W; chrXII:175226-176744(+)

Downstream 512bp

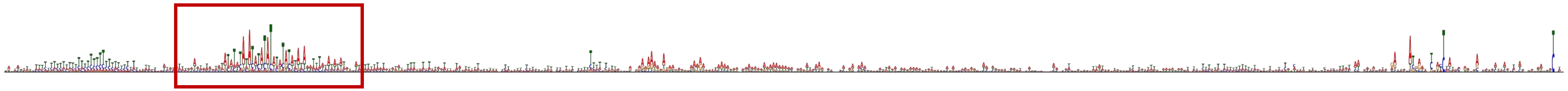
YLR057W; chrXII:255305-257855 (+)



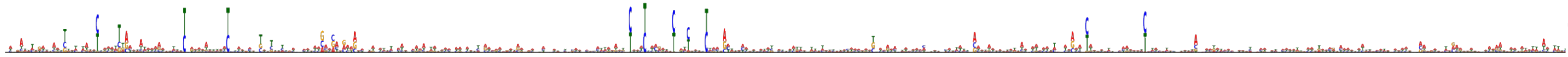
YLR438W; chrXII:1012500-1013775 (+)



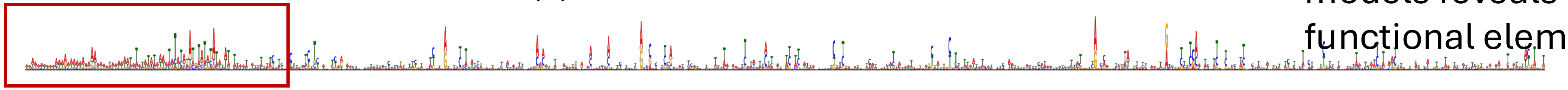
YLR056W; chrXII:253860-254958 (+)



YLR134W; chrXII:410722-412414 (+)



YLR015W; chrXII:175226-176744(+)



Tomaz da Silva et al., (2024). Nucleotide dependency analysis of DNA language models reveals genomic functional elements. bioRxiv

Fungi LM: Summary

1. Fungi language model: The **Saccharomycetales order** is a good evolutionary distance, offering good species diversity.

2. Orthologous gene annotations are **95%** complete.

3. Coding regions make up **50% - 75%** of the genome (**72.46%** in r64). Down-weighting is important!

4. A window size of 16,384 captures approximately **5-10 genes** (**9** in r64).

5. Repetitive regions account for **~2% - 15%** of the genome (**7.39%** in r64). Down-weighting is important!

6. Homologous sequence removal between train-test/validation is crucial (**40% / 60% / 16%**)

7. Transformer-based U-Net architecture overfits in r64 but generalizes best in Saccharomycetales.

8. Self-supervised learning is able to capture cis-regulatory motifs (preliminary results)

Part II

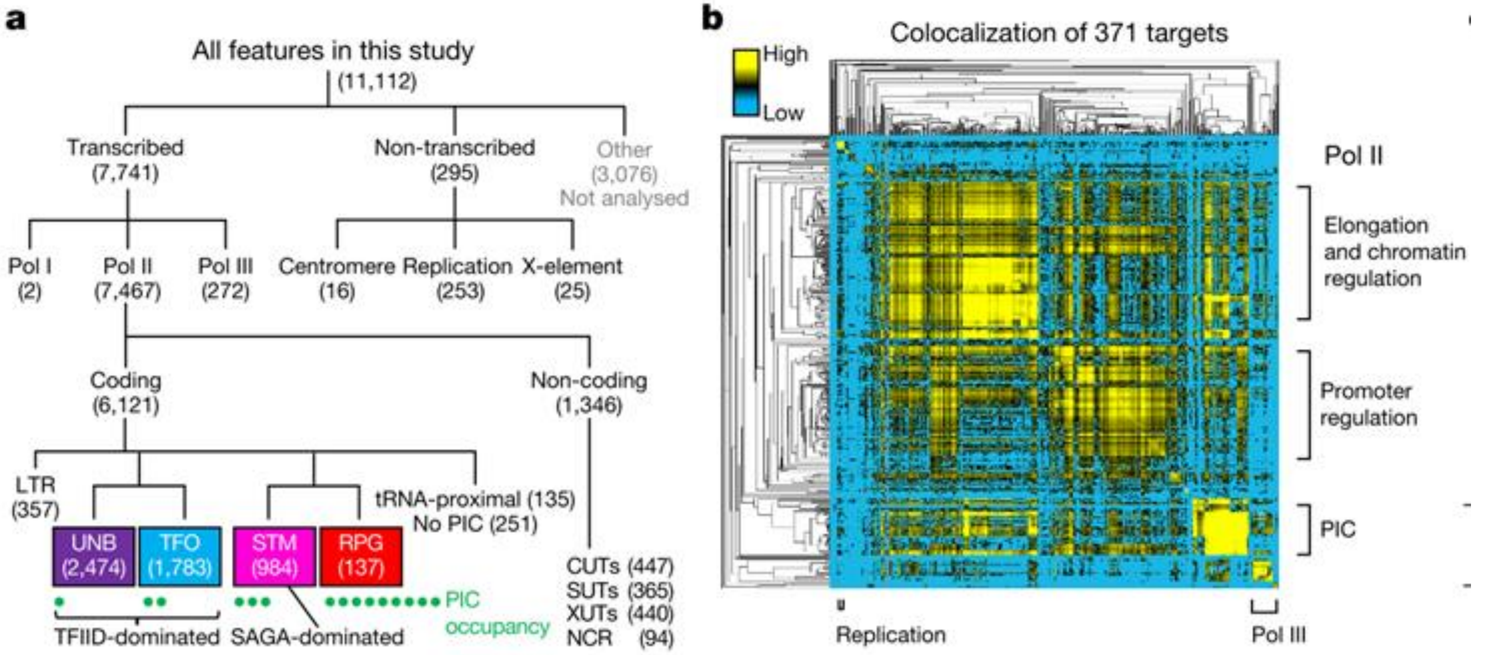
Supervised ChiP-exo, histone marks, RNA-Seq prediction

Linder, J., Srivastava, D., Yuan, H., Agarwal, V., & Kelley, D. R. (2023). Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. Biorxiv, 2023-08.

Label data introduction & preprocessing

ChiP-exo + Histone Marks

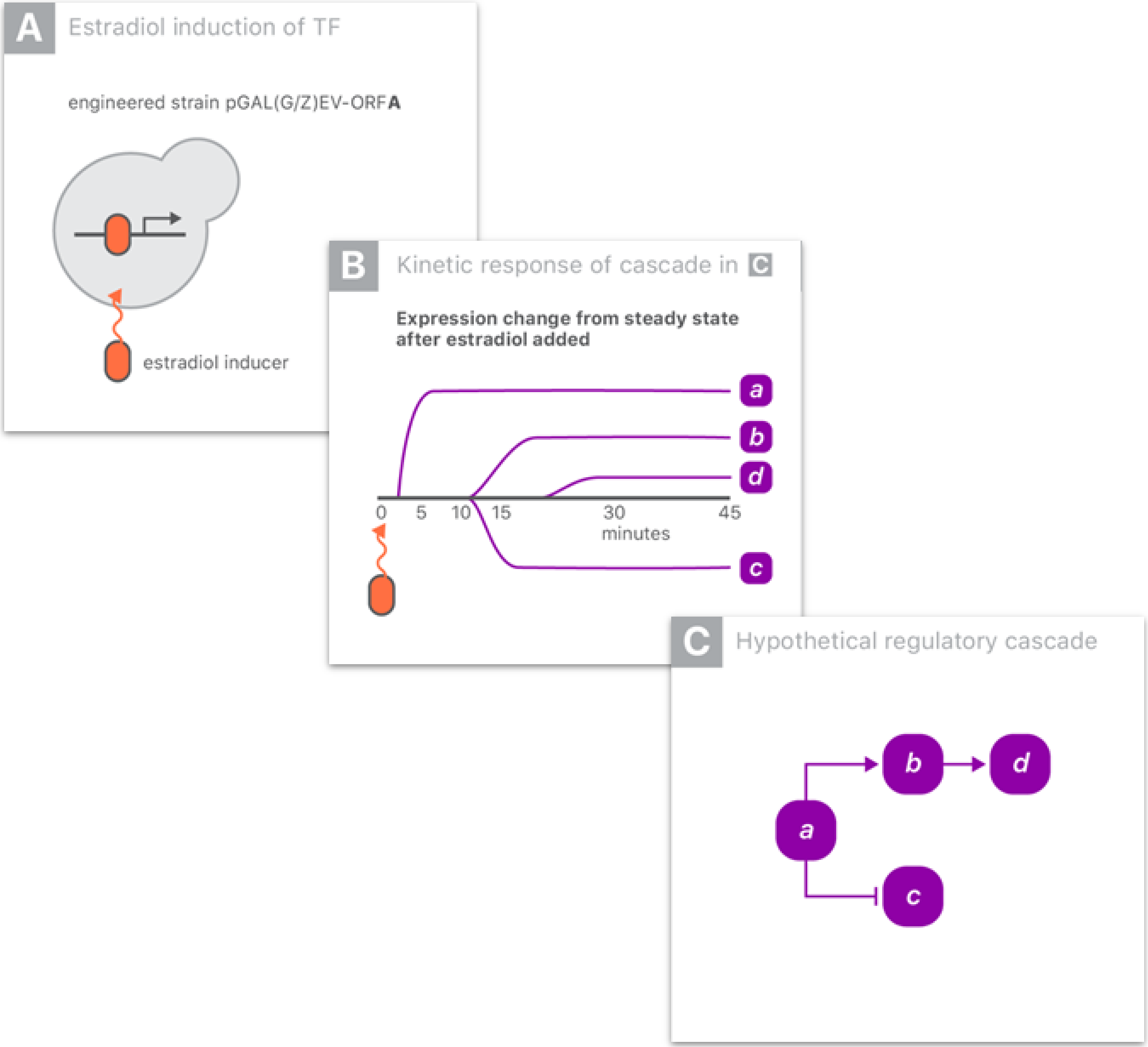
- ChIP-exo provides high res view of DNA binding
- Dataset includes 800 ChIP-exo experiments:
- Epigenetic regulators, DNA replication, centromeres, subtelomeres, transposons, RNA polymerase I/II/III
- 161 matched TF ChIP-exo from IDEA 1.0
- Histone Mods MNase-ChIP-seq



Rossi, Matthew et al. Nature. 2021.

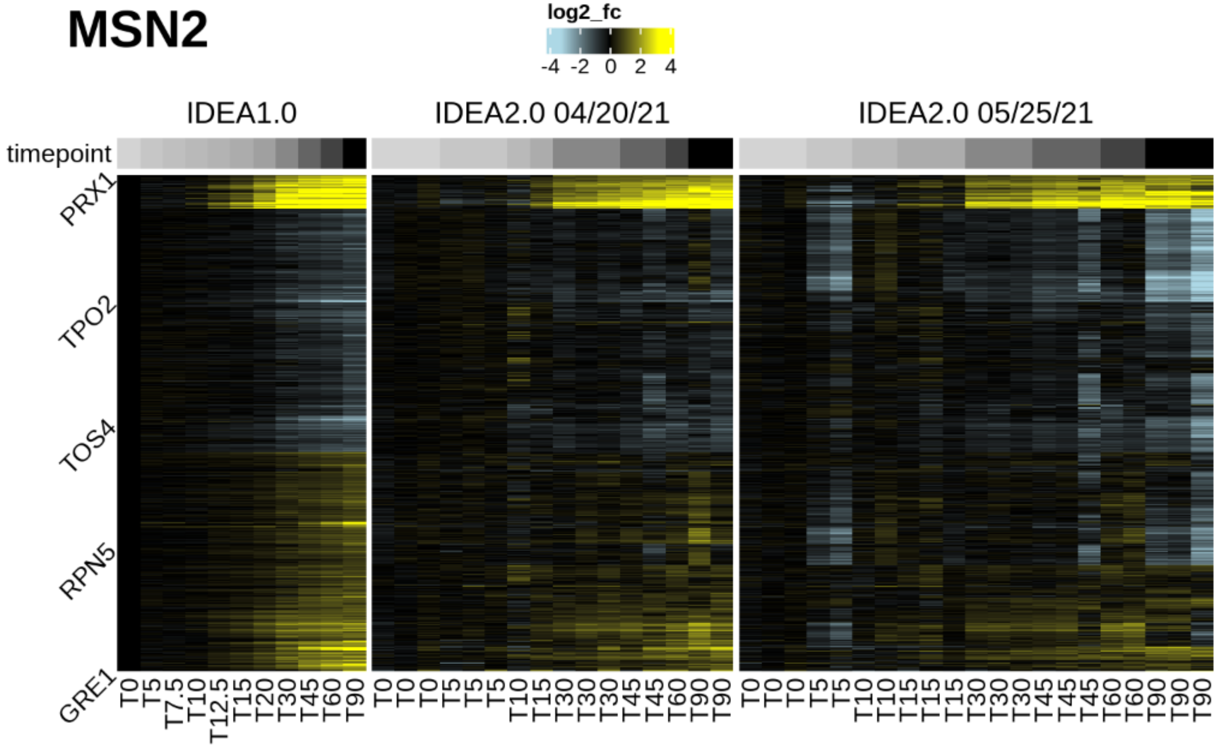
RNA-Seq

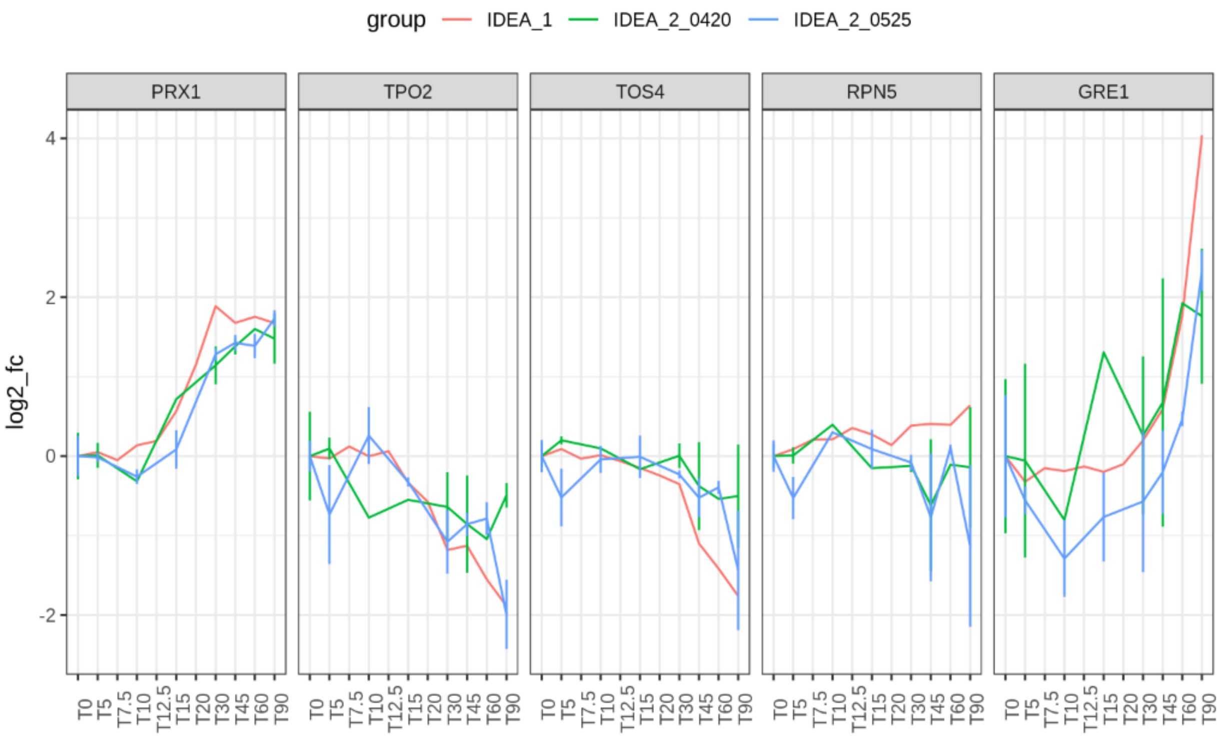
- Genome-scale perturbation dynamics propagate signals across regulatory networks
- Measuring dynamics allows events to be ordered
- Aggregating dynamics across many time-courses enables disambiguation of cause > effect relationships


A
Estradiol induction of TF
engineered strain pGAL(G/Z)EV-ORFA
estradiol inducer
B
Kinetic response of cascade in C
Expression change from steady state after estradiol added
a
b
d
c
0
5
10
15
30
45
minutes
C
Hypothetical regulatory cascade
a
b
d
c

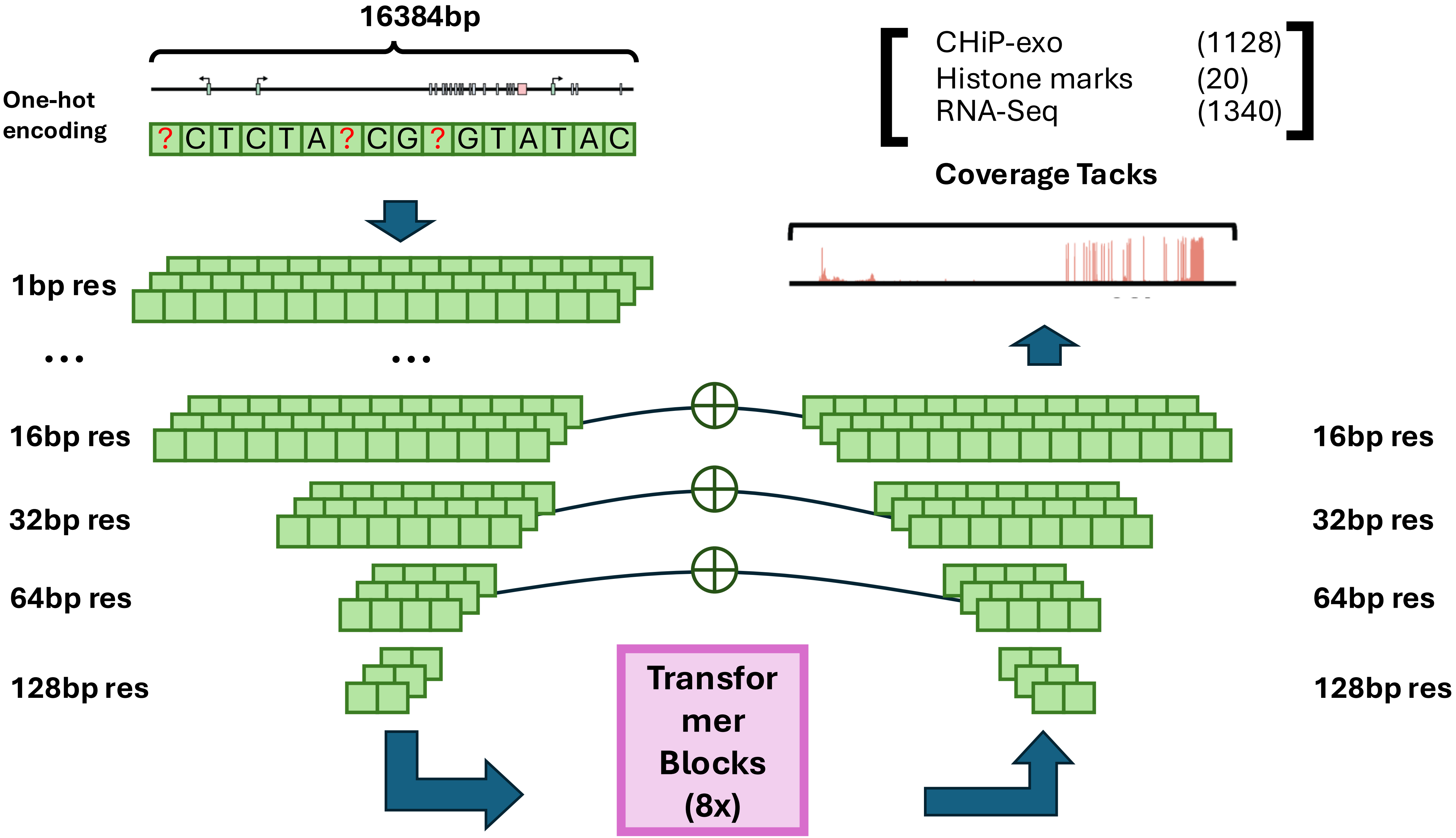
RNA-Seq

- IDEA (the Induction Dynamics gene Expression Atlas)





Supervised model architecture



16384bp

One-hot encoding

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

Coverage Tacks

1bp res

...

16bp res

32bp res

64bp res

128bp res

Transformer Blocks (8x)

Basenji Model Training

- Divide genome into 8 folds.
- Train 8 models with distinct validation and test folds.

Kristy May for Pets Adviser

Fold0: 743 seq, 1406020 nt (0.1244)

chrXIV:	0-628758
chrX:	0-436307
chrXI:	440246-666816
chrIII:	0-114385

Fold1: 736 seq, 1433427 nt (0.1268)

chrXI:	0-440129
chrV:	0-151987
chrV:	152104-576874
chrXIII:	0-268031
chrVI:	0-148510

Fold2: 806 seq, 1521492 nt (0.1346)

chrII:	238323-813184
chrVII:	0-496920
chrIV:	0-449711

Fold3: 755 seq, 1408276 nt (0.1246)

chrXVI:	0-555957
chrIV:	449821-990877
chrVI:	48627-270161
chrVIII:	0-105586
chrIX:	355745-439888

Fold4: 732 seq, 1444997 nt (0.1278)

chrIV:	990877-1531933
chrXII:	614562-1078177
chrII:	0-238207
chrIII:	114501-316620

Fold5: 742 seq, 1284157 nt (0.1136)

chrVII:	497038-1090940
chrX:	436425-745751
chrI:	0-151465
chrI:	151582-230218
chrXII:	0-150828

Fold6: 785 seq, 1446481 nt (0.1280)

chrXIII:	268149-924431
chrXII:	150947-614562
chrXV:	0-326584

Fold7: 733 seq, 1360020 nt (0.1203)

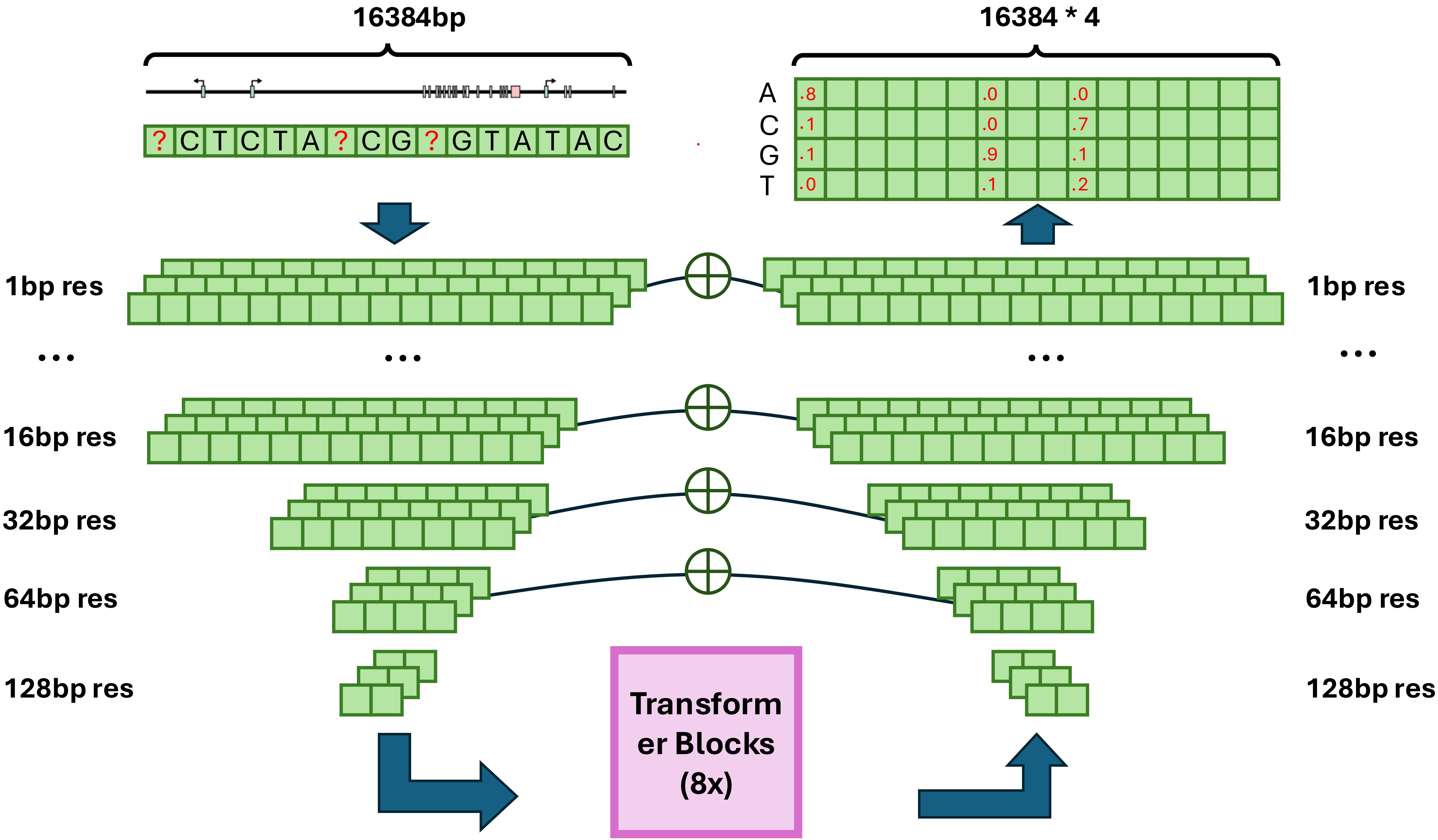
chrVIII:	105703-562643
chrXVI:	556073-948066
chrIX:	0-355629
chrXIV:	628875-784333

Part III

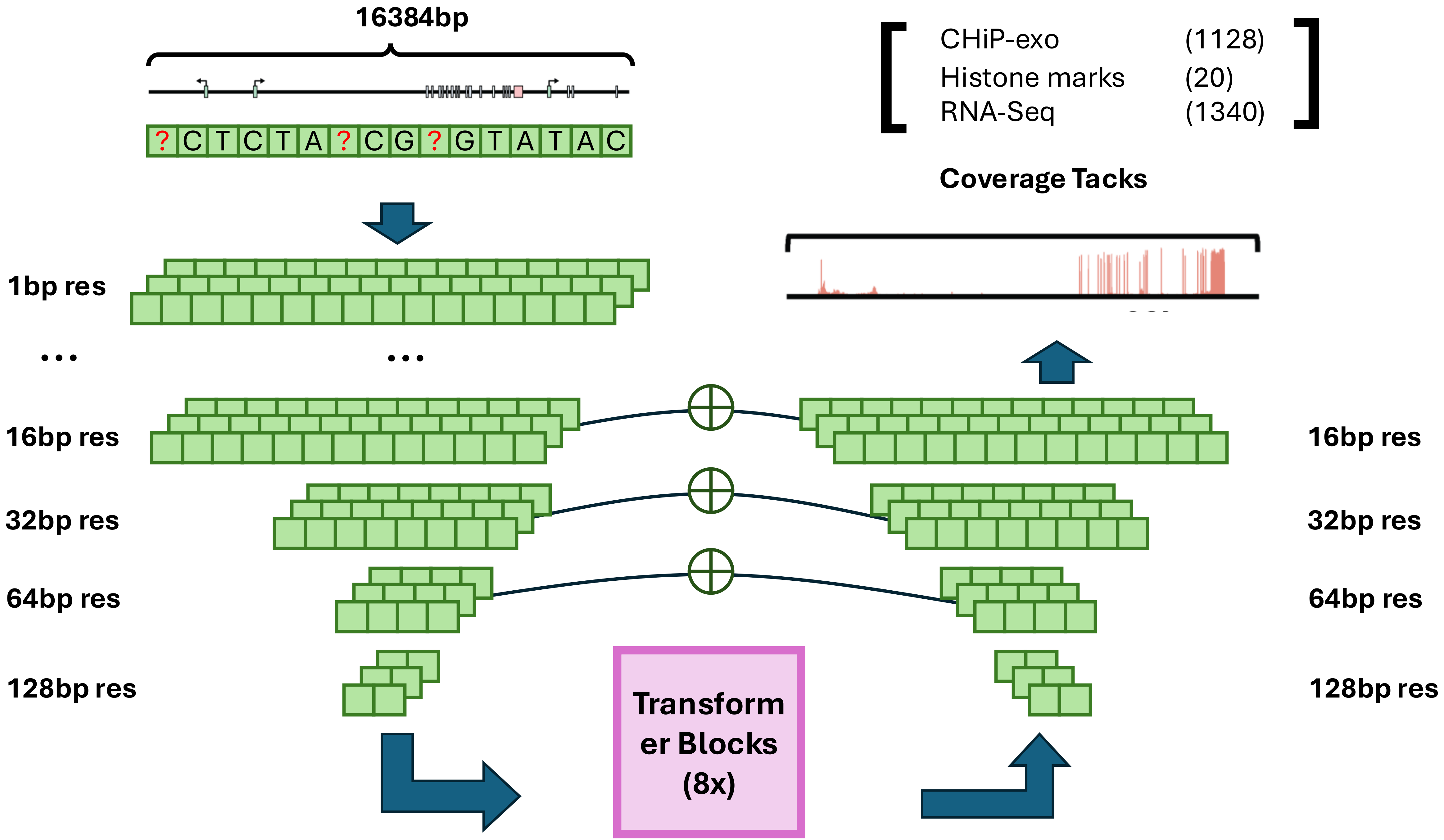
Fine-tuning Fungi Language Model

Q: Does fine-tuning a pretrained LM outperform training a new model from scratch under the exact model architecture?

Supervised Fungi model VS Fine-tuning Language Model



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



16384bp

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

CHiP-exo (1128)

Histone marks (20)

RNA-Seq (1340)

Coverage Tacks

1bp res

16bp res

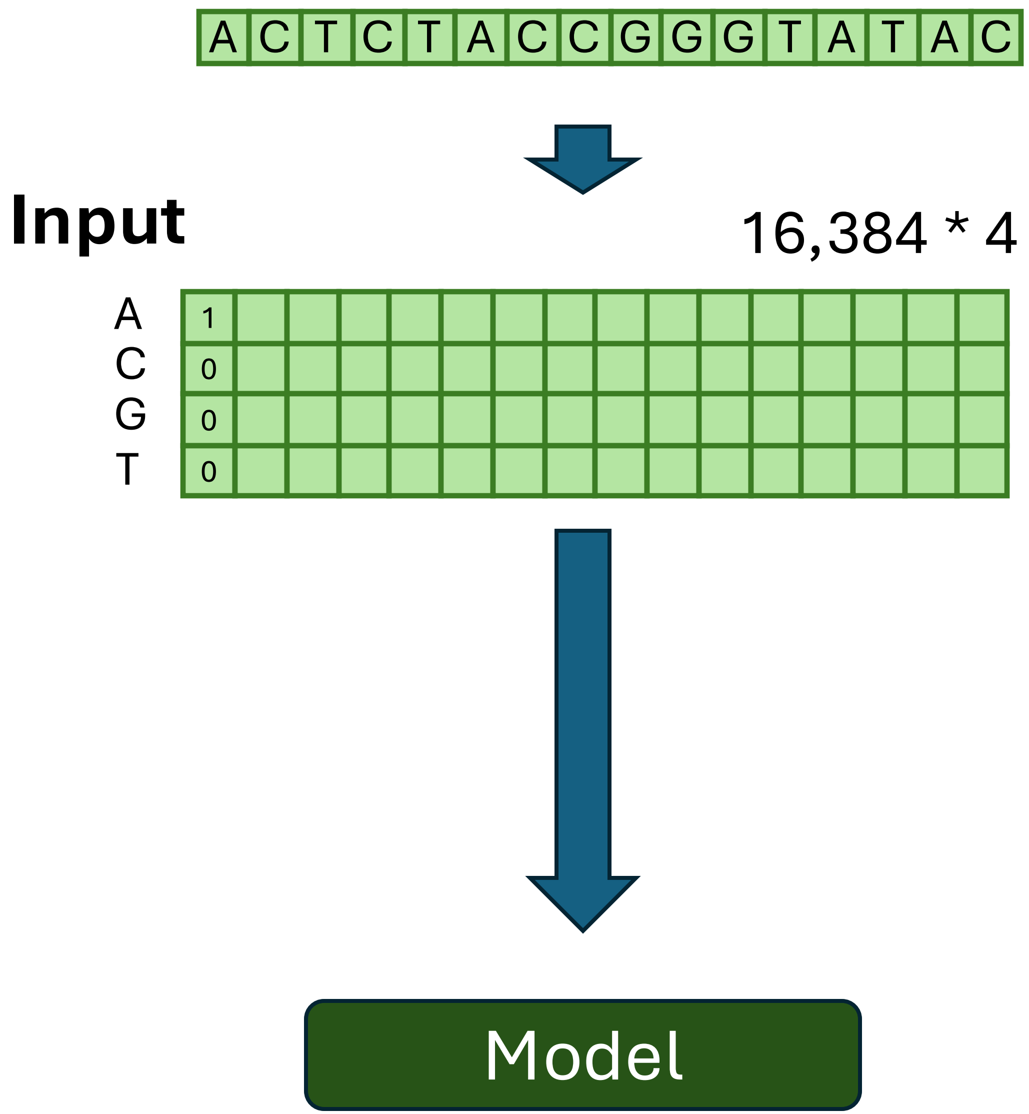
32bp res

64bp res

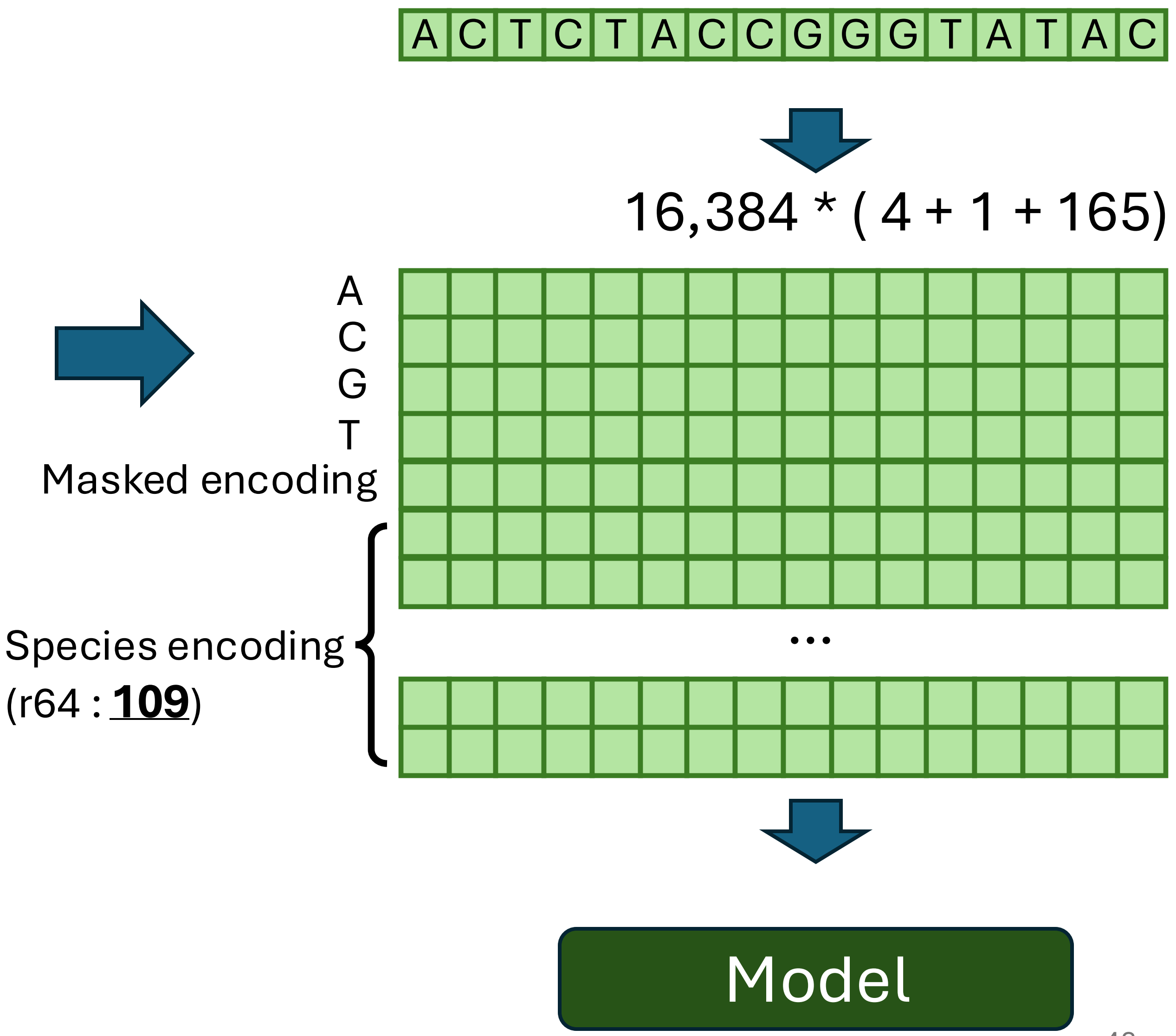
128bp res

Transformer Blocks (8x)

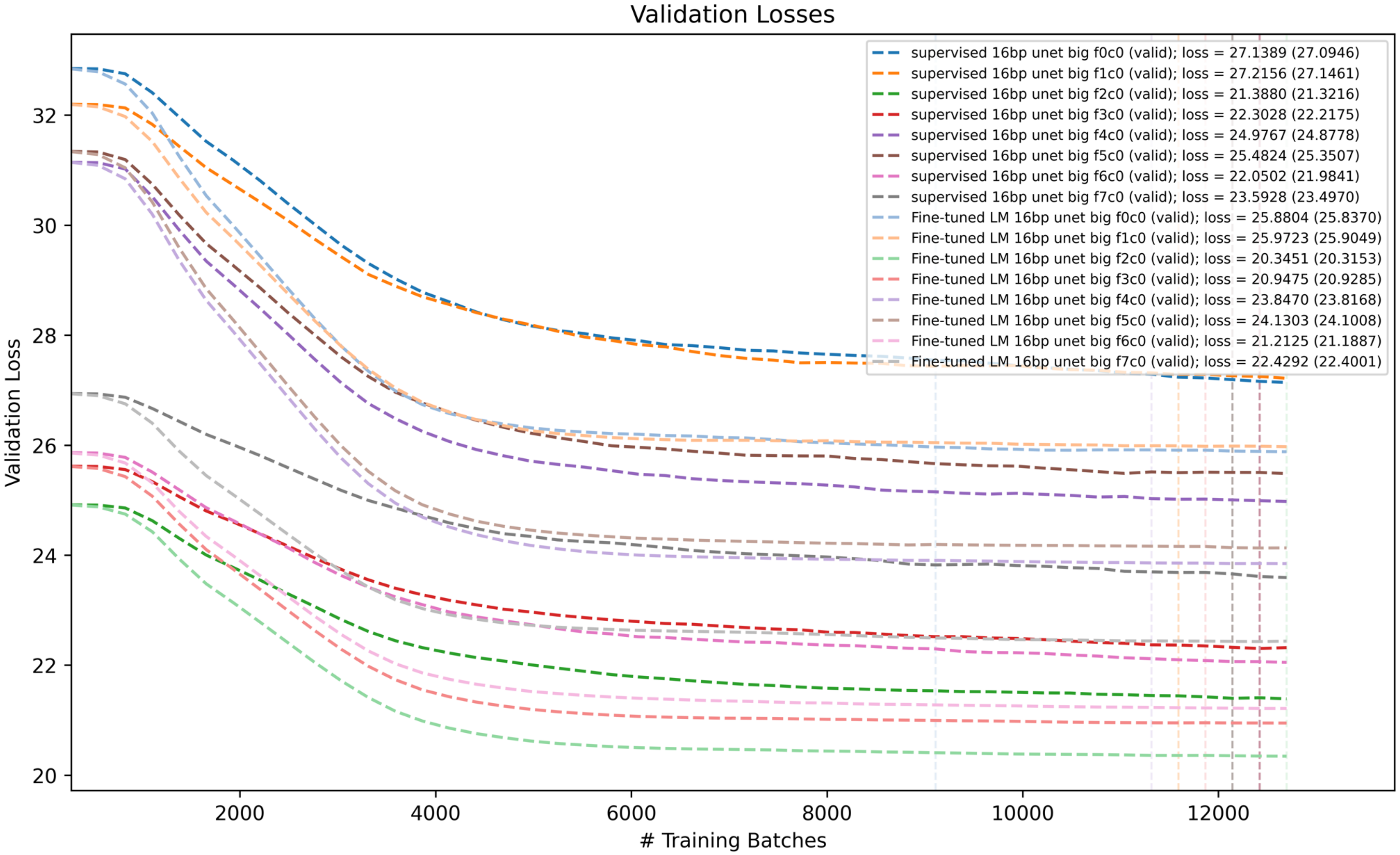
Supervised



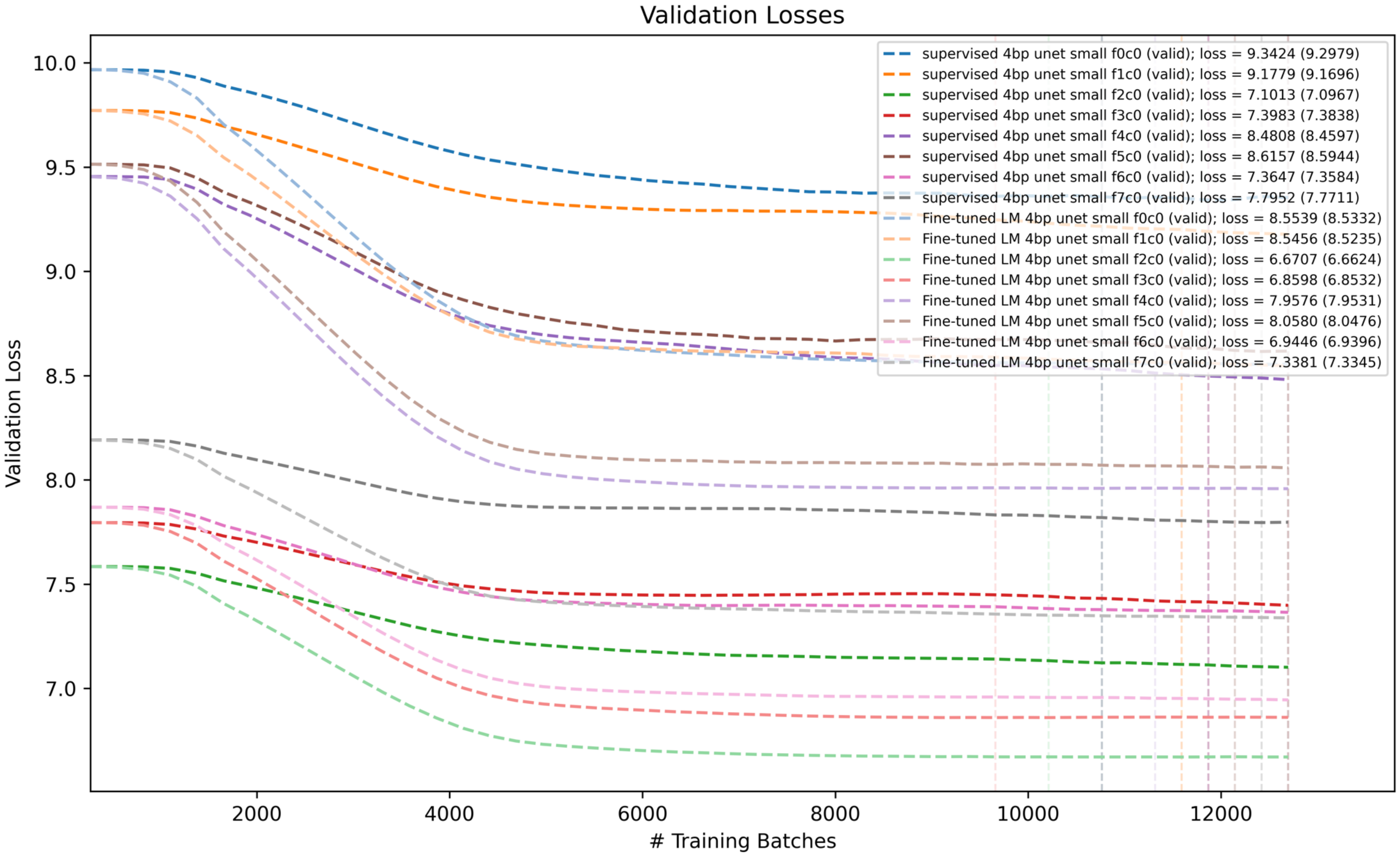
Fine-tuning LM

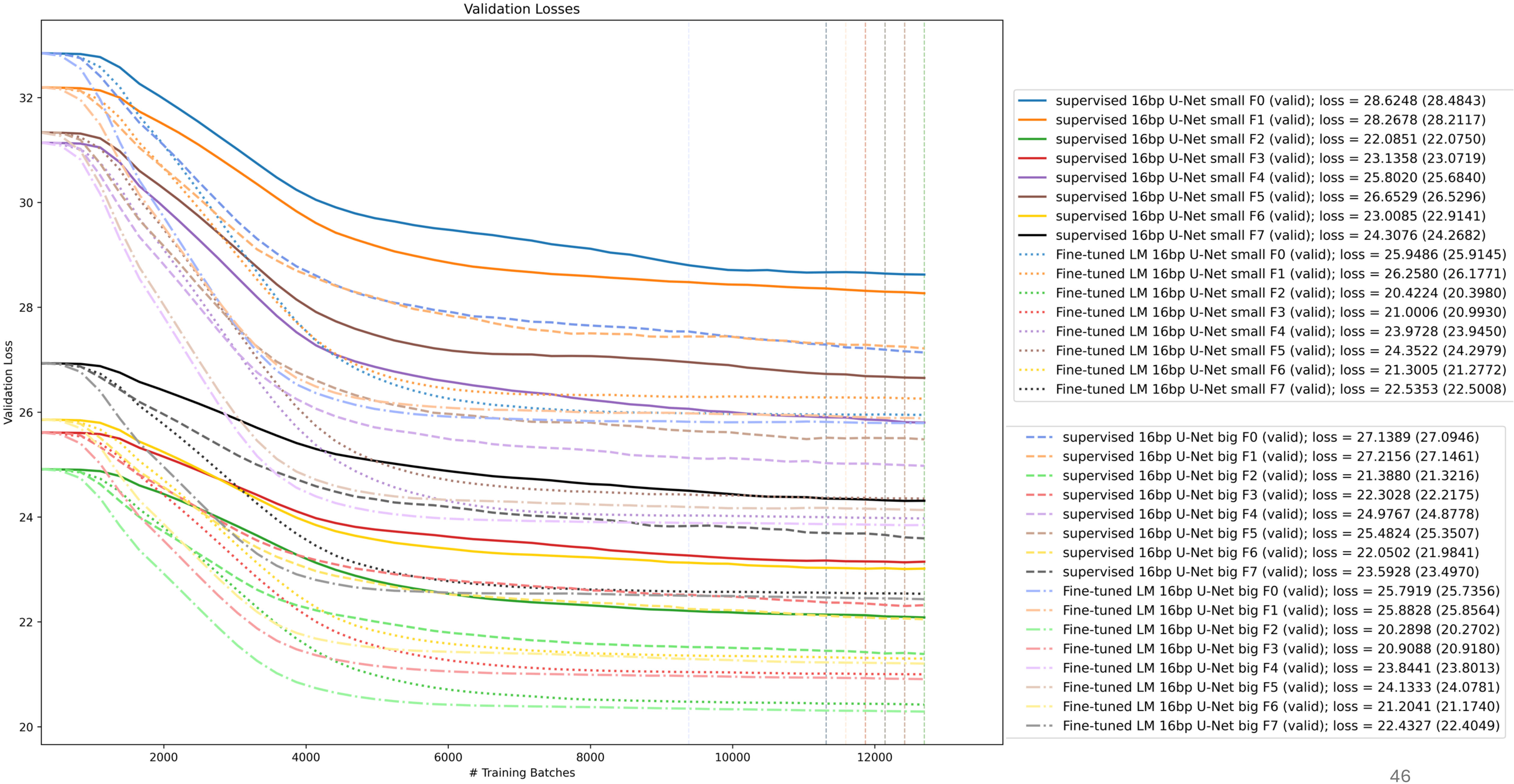


Fine-tuning vs Training from Scratch (16 bp resolution)



Fine-tuning vs Training from Scratch (4 bp resolution)



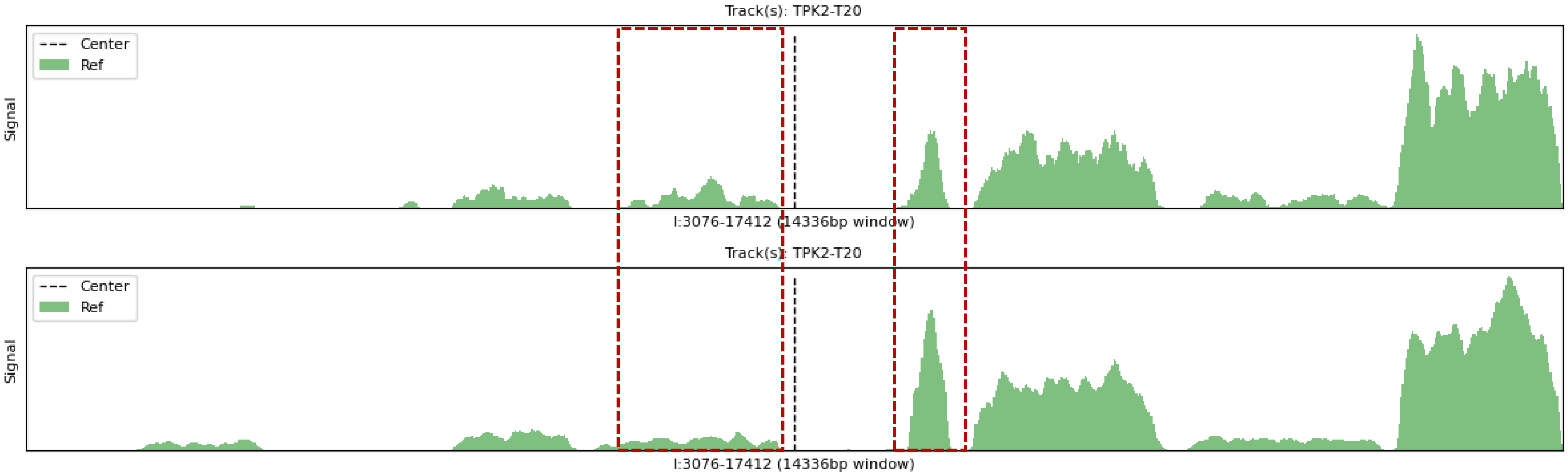

Validation Losses
Validation Loss
# Training Batches
20
22
24
26
28
30
32
2000
4000
6000
8000
10000
12000
supervised 16bp U-Net small F0 (valid); loss = 28.6248 (28.4843)
supervised 16bp U-Net small F1 (valid); loss = 28.2678 (28.2117)
supervised 16bp U-Net small F2 (valid); loss = 22.0851 (22.0750)
supervised 16bp U-Net small F3 (valid); loss = 23.1358 (23.0719)
supervised 16bp U-Net small F4 (valid); loss = 25.8020 (25.6840)
supervised 16bp U-Net small F5 (valid); loss = 26.6529 (26.5296)
supervised 16bp U-Net small F6 (valid); loss = 23.0085 (22.9141)
supervised 16bp U-Net small F7 (valid); loss = 24.3076 (24.2682)
Fine-tuned LM 16bp U-Net small F0 (valid); loss = 25.9486 (25.9145)
Fine-tuned LM 16bp U-Net small F1 (valid); loss = 26.2580 (26.1771)
Fine-tuned LM 16bp U-Net small F2 (valid); loss = 20.4224 (20.3980)
Fine-tuned LM 16bp U-Net small F3 (valid); loss = 21.0006 (20.9930)
Fine-tuned LM 16bp U-Net small F4 (valid); loss = 23.9728 (23.9450)
Fine-tuned LM 16bp U-Net small F5 (valid); loss = 24.3522 (24.2979)
Fine-tuned LM 16bp U-Net small F6 (valid); loss = 21.3005 (21.2772)
Fine-tuned LM 16bp U-Net small F7 (valid); loss = 22.5353 (22.5008)
supervised 16bp U-Net big F0 (valid); loss = 27.1389 (27.0946)
supervised 16bp U-Net big F1 (valid); loss = 27.2156 (27.1461)
supervised 16bp U-Net big F2 (valid); loss = 21.3880 (21.3216)
supervised 16bp U-Net big F3 (valid); loss = 22.3028 (22.2175)
supervised 16bp U-Net big F4 (valid); loss = 24.9767 (24.8778)
supervised 16bp U-Net big F5 (valid); loss = 25.4824 (25.3507)
supervised 16bp U-Net big F6 (valid); loss = 22.0502 (21.9841)
supervised 16bp U-Net big F7 (valid); loss = 23.5928 (23.4970)
Fine-tuned LM 16bp U-Net big F0 (valid); loss = 25.7919 (25.7356)
Fine-tuned LM 16bp U-Net big F1 (valid); loss = 25.8828 (25.8564)
Fine-tuned LM 16bp U-Net big F2 (valid); loss = 20.2898 (20.2702)
Fine-tuned LM 16bp U-Net big F3 (valid); loss = 20.9088 (20.9180)
Fine-tuned LM 16bp U-Net big F4 (valid); loss = 23.8441 (23.8013)
Fine-tuned LM 16bp U-Net big F5 (valid); loss = 24.1333 (24.0781)
Fine-tuned LM 16bp U-Net big F6 (valid); loss = 21.2041 (21.1740)
Fine-tuned LM 16bp U-Net big F7 (valid); loss = 22.4327 (22.4049)

RNA-Seq track visualization

YAL067C (chrI: 7235 – 9016) **YAL065C (chrI: 11565 – 11951)**

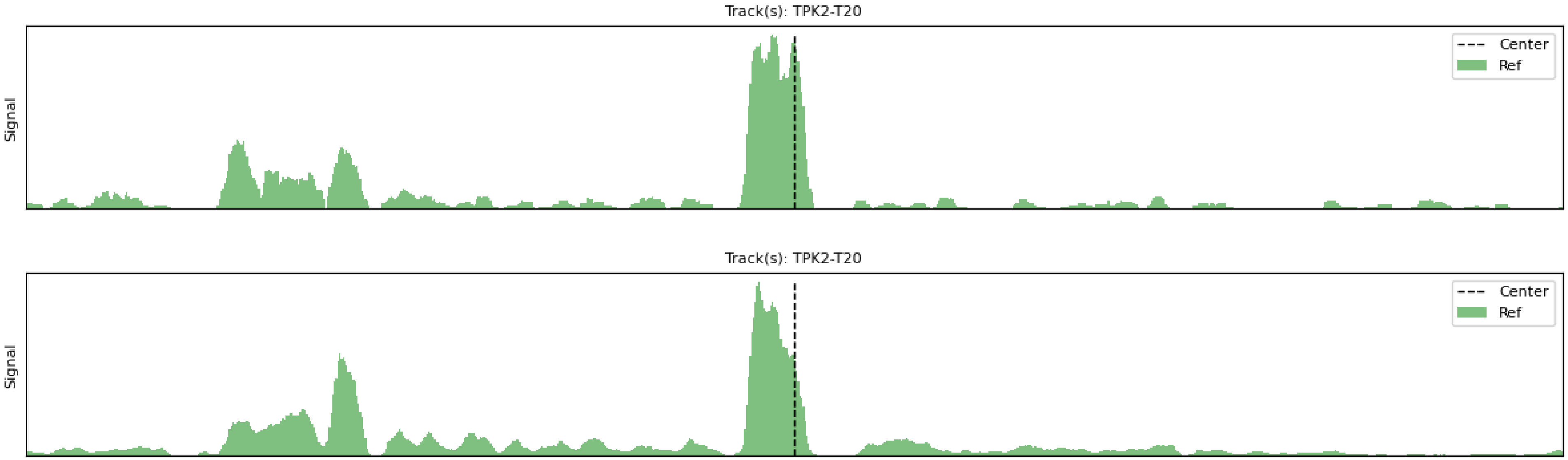
Label

Prediction



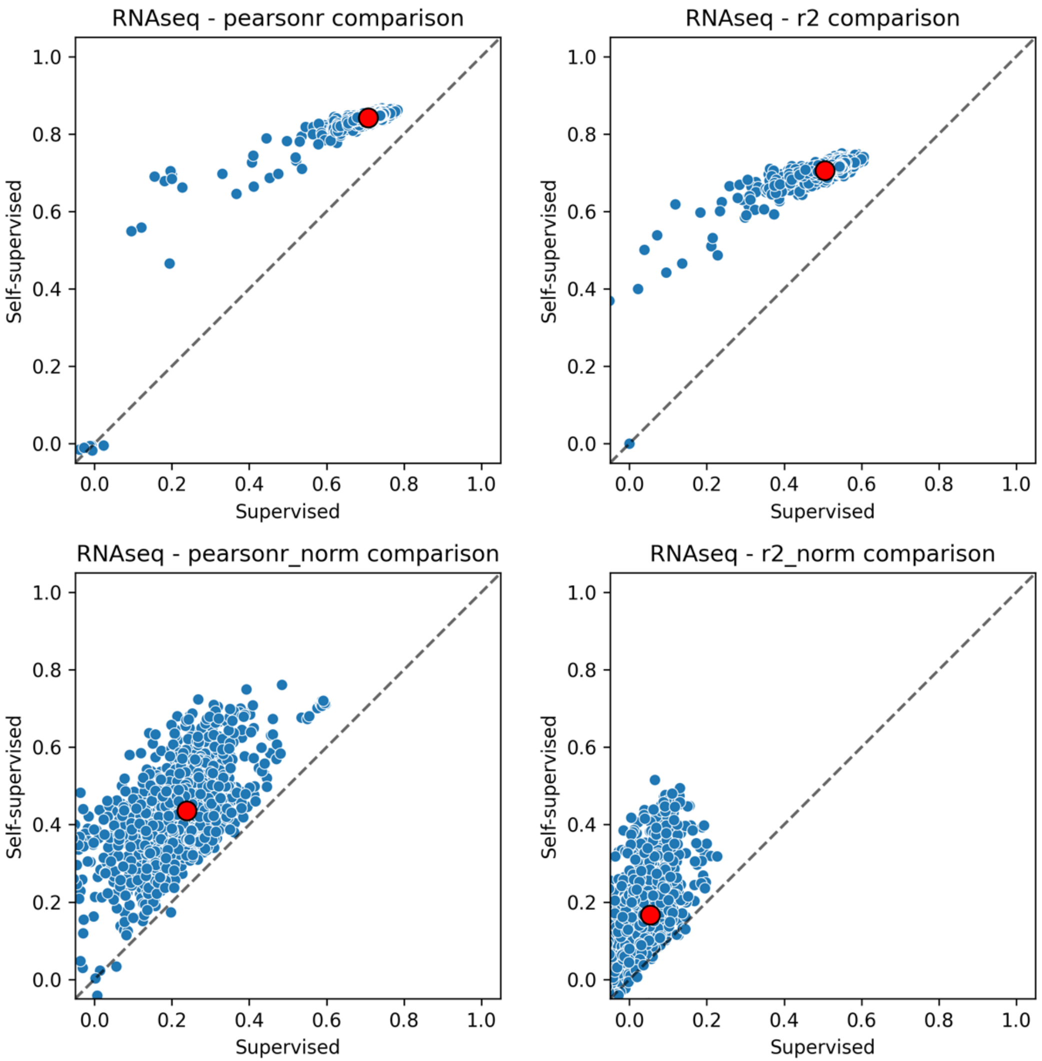
Label

Prediction

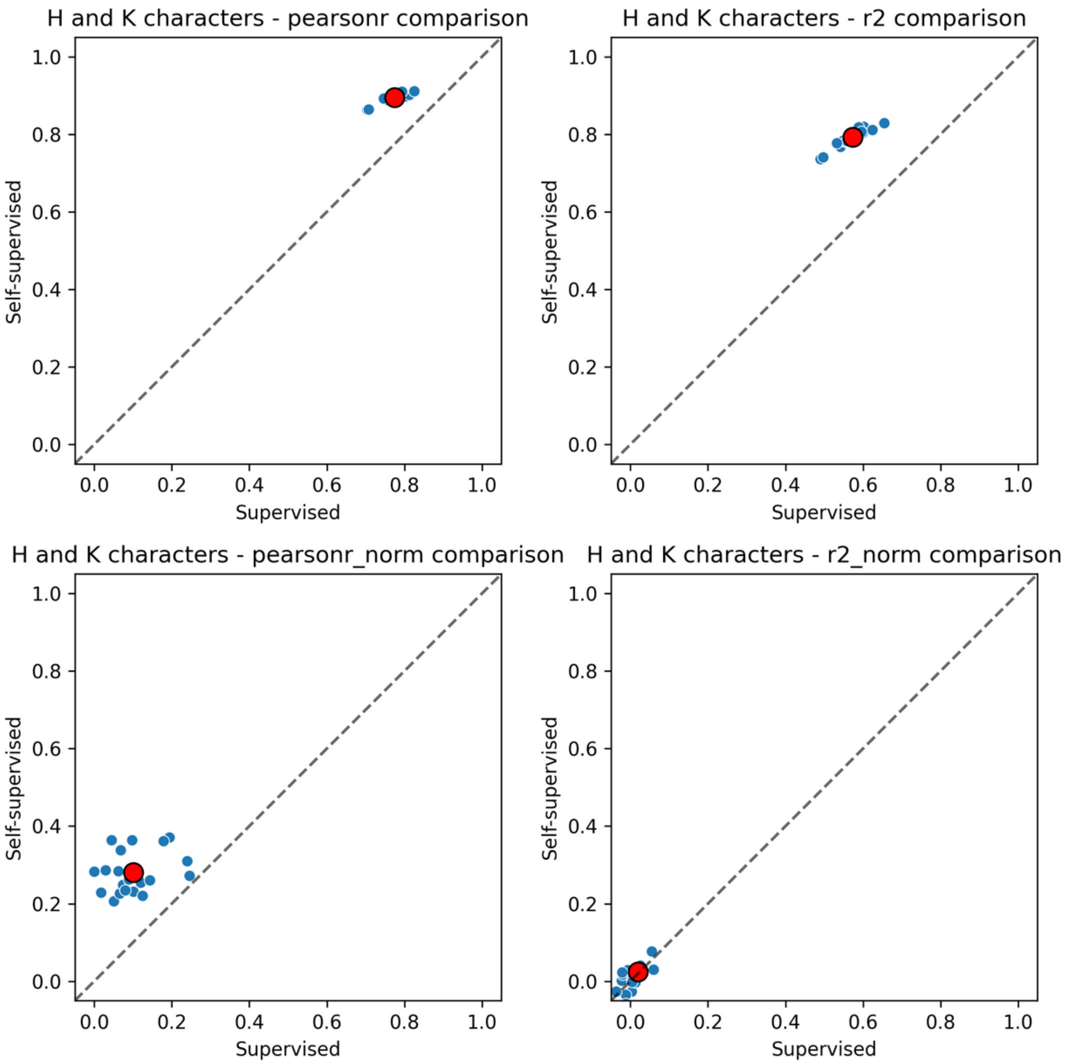


Track level prediction evaluation

RNA-Seq

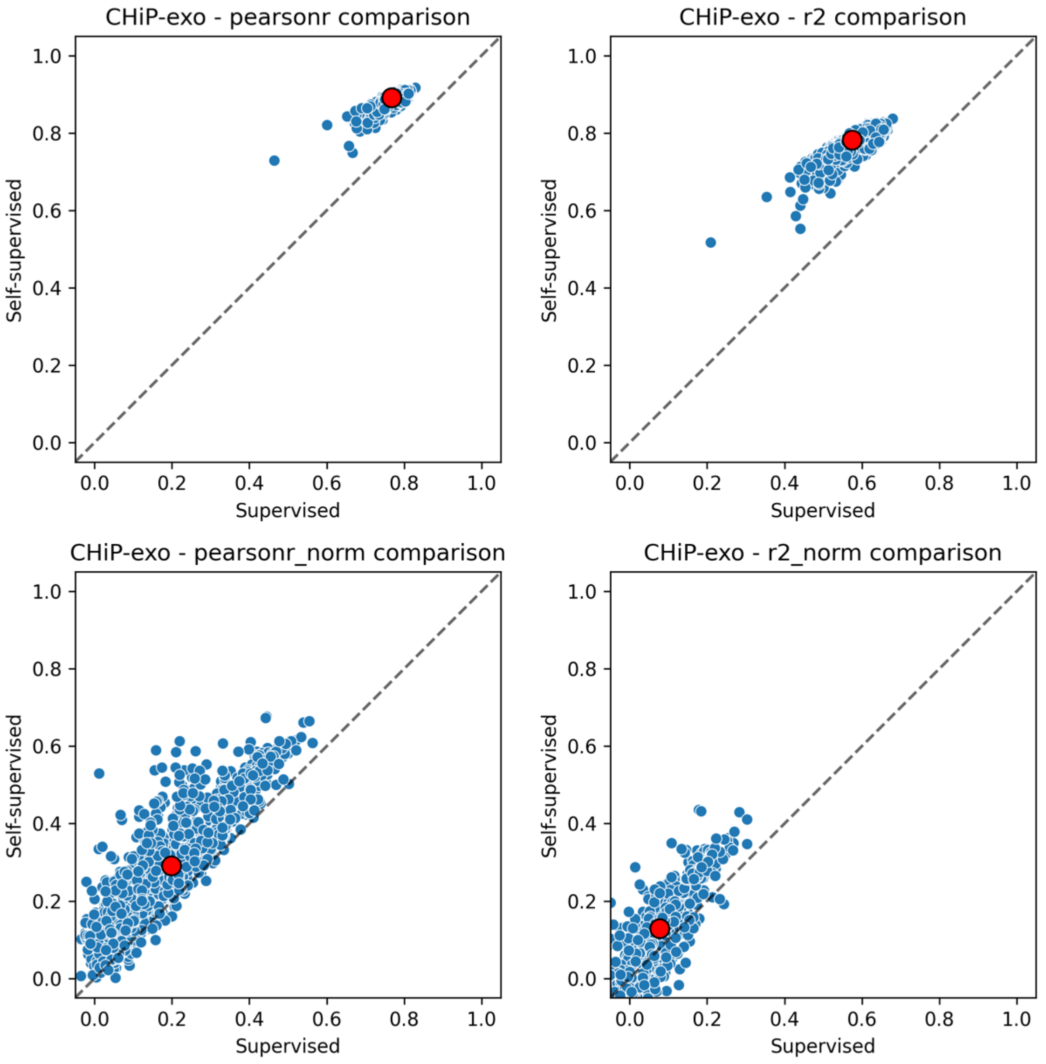


Histone Marks

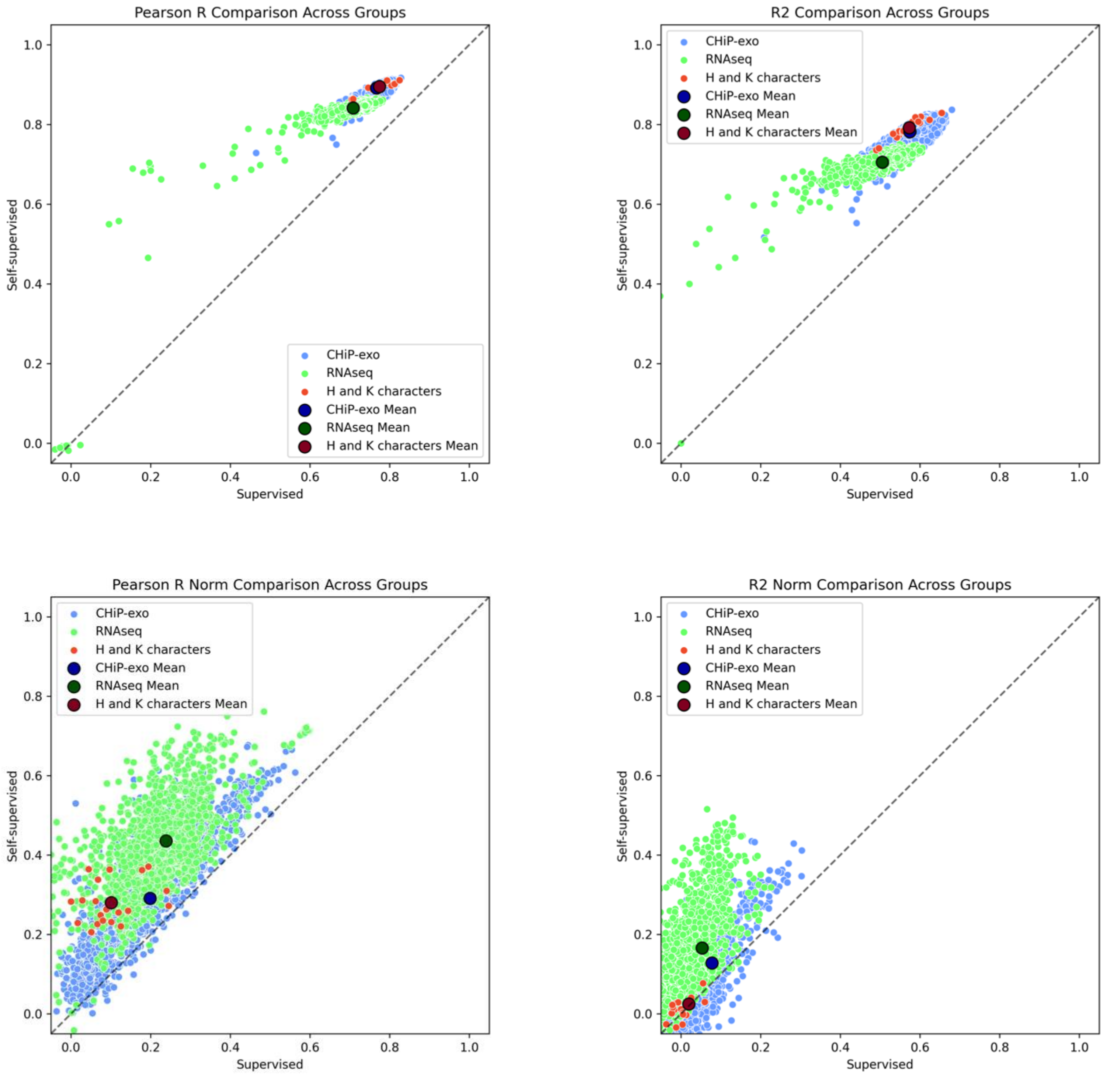


Average results across 8 folds. Each dot is a track.

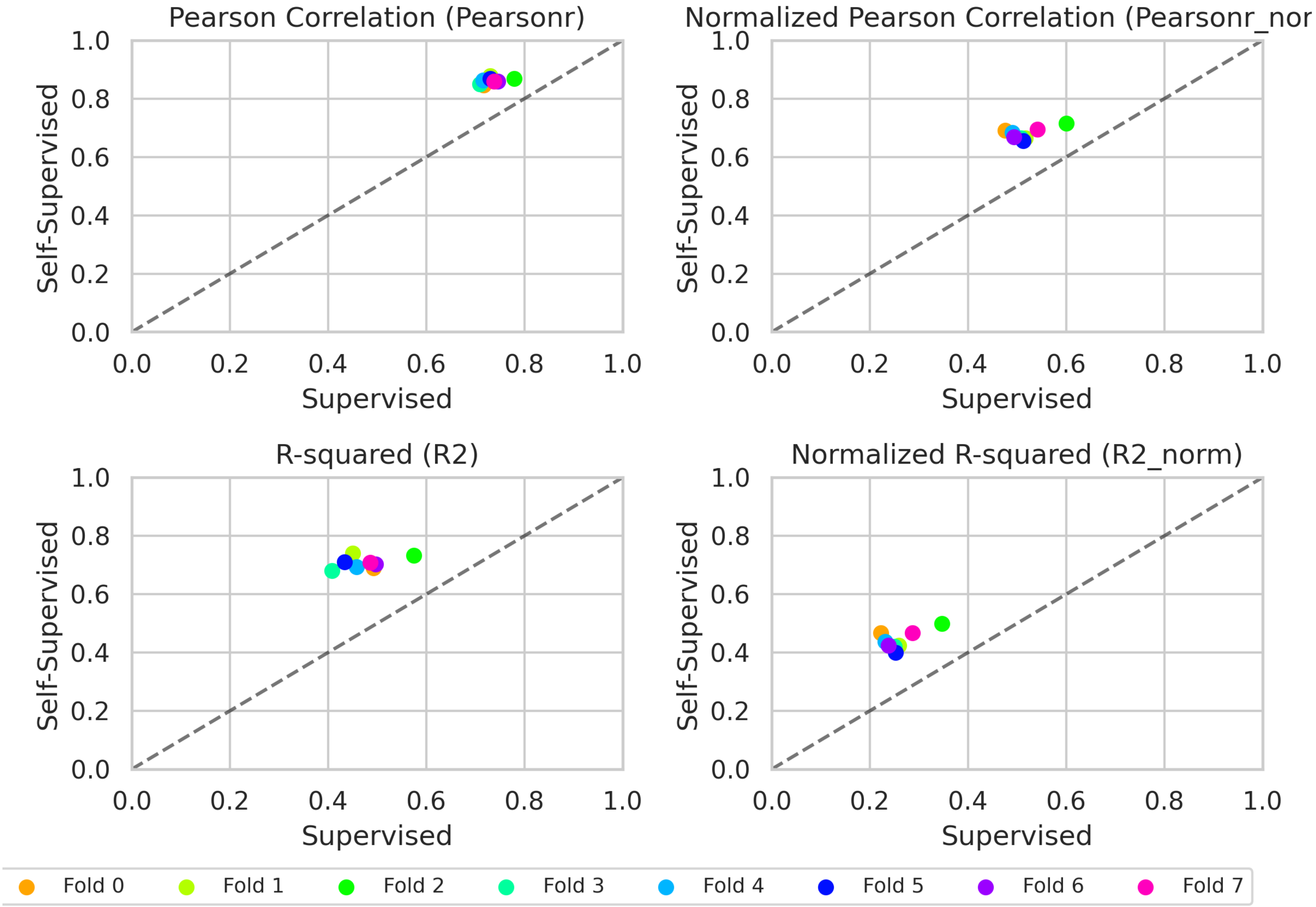
CHiP-exo



All (RNA-Seq + Histone Marks + CHiP-exo)



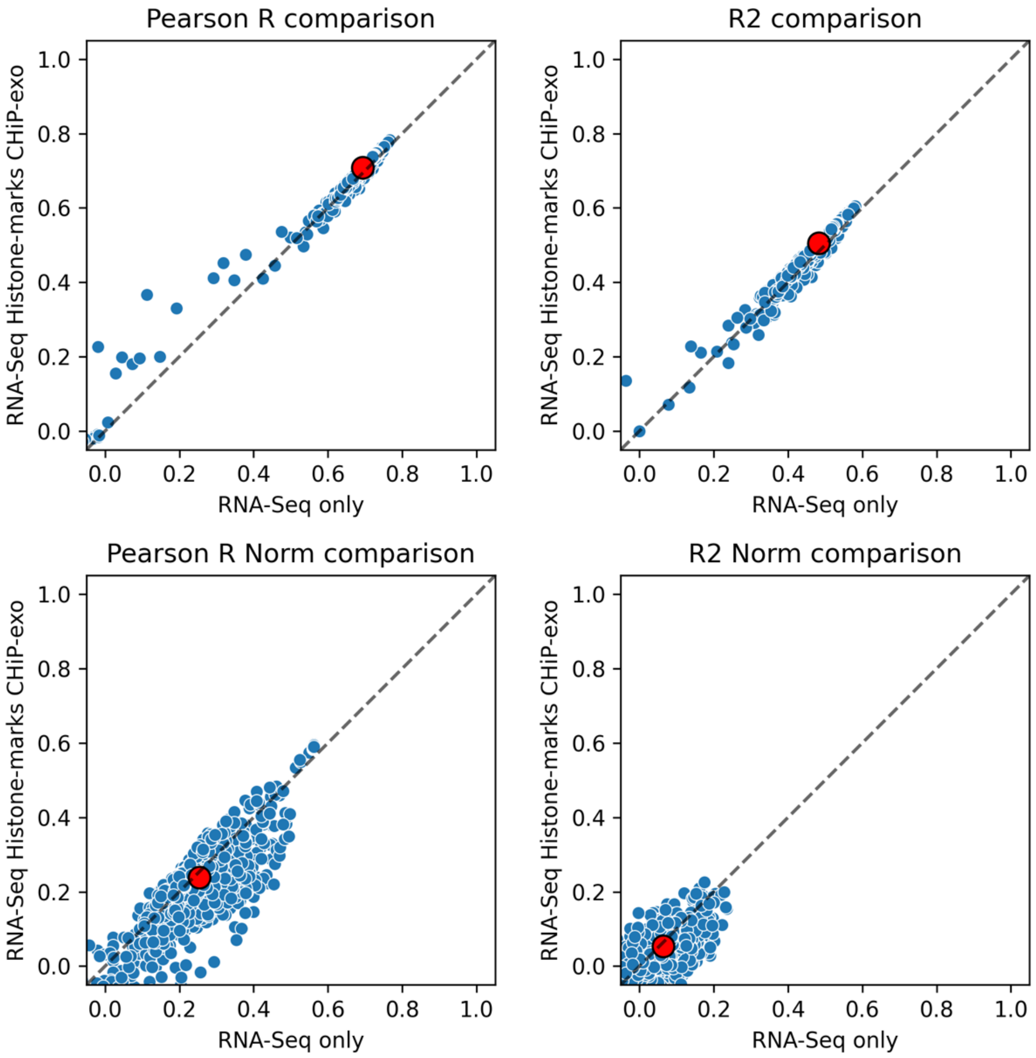
Average results across 8 folds. Each dot is a track.



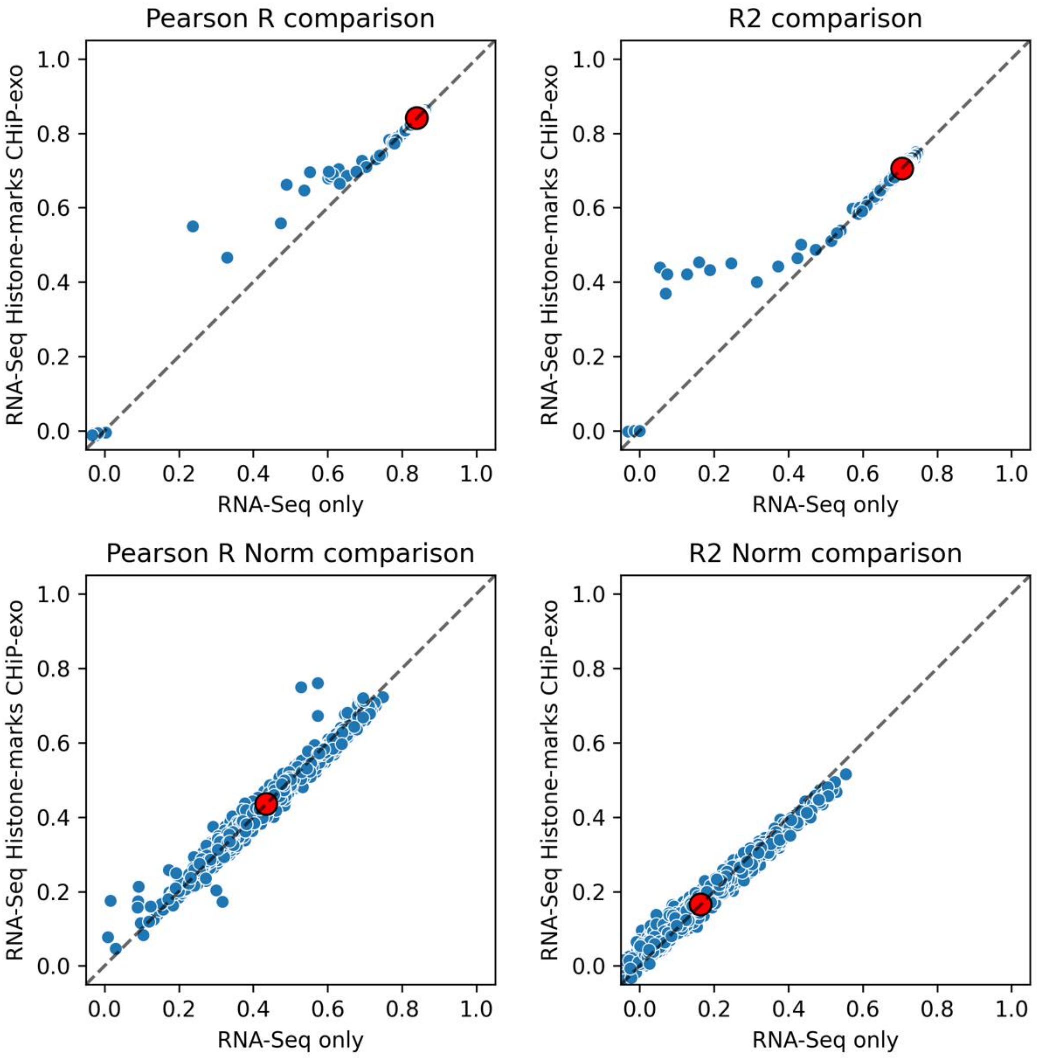
Average results across tracks.
Each dot is a fold

RNA-Seq tracks alone VS
RNA-Seq + Histone Marks +
CHiP-exo tracks

Supervised trained models



Self-supervised trained models



Average results across 8 folds. Each dot is a track.

Project Conclusion

1. Built the first fungi language model. The Saccharomycetales order is a good evolutionary distance, offering good species diversity. Processing 1361 fungus genomes.
2. Under the exact model architecture, pretrained LM weights & fine-tuning can outperform training a model from scratch.
   - Loss / gene level Pearson R / gene level R^2

khchao.com @KuanHaoChao Kuanhao-Cha

Acknowledgement

Johannes Linder Majed Mohamed Magzoub David Kelley Sean Hackett

Kelley Lab & Calico Computing Team **Great mentors, collaborators and good friends!**