

# Deep Learning + Genomics Study Group

Kuan-Hao Chao / Mahler Revsine

2024.10.22

Why are we starting this study  
group?

# Deep learning-based DNA sequence model

nature methods

[View all journals](#) | [Search](#)

Explore content

Troyanskaya Lab Princeton

DeepSEA  
2015

Brief Communication | Published: 24 August 2015  
Predicting effects of noncoding  
learning-based sequence mo

[Jian Zhou & Olga G Troyanskaya](#)

GENOME  
RESEARCH

[HOME](#) | [ABOUT](#) | [ARCHIVE](#) | [SUBMIT](#) | [SUBSCRIBE](#) | [ADVERTISE](#) | [AUTHOR](#)

Institution: MILTON S EISENHOWER LIBRARY [Sign In](#)

Calico

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

David R. Kelley<sup>1</sup>, Jasper Snoek<sup>2</sup> and John L. Rinn<sup>1</sup>

Basset  
2016

Cell

[Volume 176, Issue 3, 24 January 2019, Pages 535-548.e24](#)

Article

Predicting Splicing from Pri  
with Deep Learning

Kishore Jaganathan<sup>1,6</sup>, Sofia Kyriazopoulou Pangiotopoul<sup>1</sup>,  
Siavash Fazel Darbandi<sup>2</sup>, David Knowles<sup>3</sup>, Yang Li<sup>1</sup>, Jack  
Wenwu Cui<sup>1</sup>, Grace B. Schwartz<sup>2</sup>, Eric D. Chow<sup>3</sup>, Efstratios  
Serafim Batzoglou<sup>1</sup>, Stephan J. Sanders<sup>2</sup>, Kyle Kai-How Forh<sup>1,7</sup>

Illumina

SpliceAI  
2019

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

RESEARCH

The genetic and biochemical deter  
of mRNA degradation rates in mar

Vikram Agarwal<sup>1,2\*</sup> and David R. Kelley<sup>1\*</sup>

Calico

Saluki  
2022

nature biotechnology

Explore content | About the journal | Publish with us

FUToronto

DeepBind  
2015

Analysis | Published: 27 July 2015  
Predicting the sequence spe  
DNA- and RNA-binding pro  
learning

[Babak Alipanahi, Andrew Delong, Matthew T Weirauch & Brendan J Frey](#)

GENOME  
RESEARCH

[HOME](#) | [ABOUT](#) | [ARCHIVE](#) | [SUBMIT](#) | [SUBSCRIBE](#) | [ADVERTISE](#) | [AUTHOR](#)

Institution: MILTON S EISENHOWER LIBRARY [Sign In](#)

Calico

Sequential regulatory activity  
across chromosomes with co  
neural networks

David R. Kelley<sup>1</sup>, Yakir A. Reshef<sup>2</sup>, Maxwell Bileschi<sup>3</sup>, Da  
Cory Y. McLean<sup>3</sup> and Jasper Snoek<sup>3</sup>

Basenji  
2018

nature methods

Explore content | About the journal | Publish with us

Calico

Calico

Akita  
2020

Predicting 3D genome folding from  
with Akita

Geoff Fudenberg<sup>1,5</sup>, David R. Kelley<sup>2,5</sup> and Katherine S. Pollard<sup>3</sup>

[Han Yuan](#) & David R. Kelley

scBasset  
2022

Bioinformatics

Gifford Lab MIT

DNA-TF binding  
2016

Article Navigation  
JOURNAL ARTICLE  
Convolutional neural  
protein binding

[Haoyang Zeng, Matthew D. Edwards, Ge Liu, David K. Gifford](#)

nature genetics

Explore content

nature >

Troyanskaya Lab Princeton

Article | Published: 16 July 2018

Deep learning sequence-based  
variant effects on expression ar

[Jian Zhou, Chandra L. Theesfeld, Kevin Yao, Kathleen M. C](#)

ExPecto  
2018

ARTICLES

<https://doi.org/10.1038/ng.41592-021-01252-x>

OPEN

Effective gene express  
sequence by integratin

Ziga Avsec<sup>1,2\*</sup>, Vikram Agarwal<sup>2,4</sup>, Daniel Vis  
Agnieszka Grabska-Barwinska<sup>1</sup>, Kyle R. Taylor  
and David R. Kelley<sup>2,3</sup>

DeepMind + Calico

Enformer  
2021

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

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

Calico Life Sciences LLC  
drk@calicolabs.com

Calico

Borzoi  
2023



# DNA Language model

DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model DNA-language in genomic contexts  
Yanrong Ji, Zihuan Zhou, Han Liu et al. SBU Bioinformatics 2021

RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY | 8 DNA language models are powerful predictors of genome-wide variation  
Gonzalo Benegas, Sanjit Singh Batra, Yun S. Song et al. UC Berkeley GPN PNAS 2023

The Nucleotide Transformer: Foundation Models for InstaDeep + Nvidia + TUM Nucleotide Transformer bioRxiv 2023

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

DNABERT-2: Efficient Foundation Model Benchmark For Multi-Species Sequencing of Molecules  
Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dandekar, SBU + NYU ICLR 2024

Article | Open access | Published: 23 July 2024 DNA language models predict gene context in the human genome  
Melissa Sanabria, Jonas Hirsch, Pierre Lefebvre et al. TUD GROVER Nat Mach Intell 2024

Article | Open access | Published: 03 April 2024 Genomic language model predicts gene regulation and function  
Yunha Hwang, Andre L. Cornman, Elizabeth J. Hartwell, R. Girguis et al. Harvard + MIT Genomic LM Nat Commun 2024

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

Sequencing of molecules  
Eric Nguyen, Michael Poli, Matthew G. Durrant, Armin Thomas, Jeremy Sullivan, Madelena Y. Ng, Ashley Lewis, Aman Patel, Stephen A. Baccus, Tina Hernandez-Boussard, Christopher J. Burtt et al. Stanford + Arc Inst + TogetherAI Evo bioRxiv 2024

[Submitted on 5 Mar 2024] Caduceus: Bi-Directional Range DNA Sequence Model  
Yair Schiff, Chia-Hsiang Kao, Aaron Gokaslan, Tripti Agarwal et al. Cornell + Princeton + CMU Caduceus ICML 2024

Cross-species analysis of nucleotide dependency  
Jingjing Zhai, Aaron Gokaslan, Yair Schiff, Michelle C. Stitzer, M. Cinta Romo et al. Cornell + USDA-ARS + Simons PlantCaduceus bioRxiv 2024

Nucleotide dependency analysis of DNA language models  
TUM Nucleotide dependency bioRxiv 2024

- **Why We're Starting This Group:** We've noticed that Hopkins doesn't have a dedicated communication channel for this area. We're excited to create a collaborative environment where we can share and explore interesting papers together.
- **Date and Time:** ~~Every or~~ Every other Tuesday, starting from October 22<sup>nd</sup>, 12:00 pm - 1:00 pm. We'll release the schedule before the next meeting (November 5<sup>th</sup> , 12:00 pm - 1:00 pm).
- **Location:** In-person in Room 228 at Malone, or online (link next slide)

# Zoom link

- Topic: Deep learning + Genomics Reading Group

Time: Oct 22, 2024 12:00 PM Eastern Time (US and Canada)  
Join Zoom Meeting

[https://JHUBlueJays.zoom.us/j/96753976668?pwd=gt20QNIDkcJebPE  
TgP3A4bCQbd2uaY.1](https://JHUBlueJays.zoom.us/j/96753976668?pwd=gt20QNIDkcJebPETgP3A4bCQbd2uaY.1)

- Meeting ID: 967 5397 6668

Passcode: 083354

- **Meeting Format:** Each meeting, a member will present a paper. It's not required, but we hope that every attendee can present at least one paper per semester.
- **Slack Channel:** We've set up a Slack channel under the JHU Genomics Collective: [#deep-learning-reading-group](#)
- **Reading List:** We've curated a selection of papers on deep learning and language models for DNA and proteins. We're definitely happy if you want to suggest additional papers to include!
  - <https://docs.google.com/spreadsheets/d/1Ufsju9EwDYyEU82OnT0W5ZkXmrbx2iaPOTTBdesW79k/edit?usp=sharing>.

# Links

- **Schedule for Fall 2024:**
  - [https://docs.google.com/spreadsheets/d/1mRFnzRyX5ThY69ine4oGkJJjmoyl4N0d\\_7tl2tC1Ts/edit?usp=sharing](https://docs.google.com/spreadsheets/d/1mRFnzRyX5ThY69ine4oGkJJjmoyl4N0d_7tl2tC1Ts/edit?usp=sharing)
- **Presenter registration form:**
  - [https://docs.google.com/forms/d/e/1FAIpQLSfgNDPOMdSNxO3OT2qdUhoPVETKvxtOxF7T-3dDLJlyilaMsg/viewform?usp=sf\\_link](https://docs.google.com/forms/d/e/1FAIpQLSfgNDPOMdSNxO3OT2qdUhoPVETKvxtOxF7T-3dDLJlyilaMsg/viewform?usp=sf_link)

Reference: [https://youtu.be/4J\\_NL5S3eYc?si=Tmc\\_TvCEMDf\\_Xleq](https://youtu.be/4J_NL5S3eYc?si=Tmc_TvCEMDf_Xleq)

<https://doi.org/10.1101/2023.08.30.555582>

<https://github.com/calico/borzoi>



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

Kuan-Hao Chao

2024.10.22

# GWAS highlights MANY genomic loci

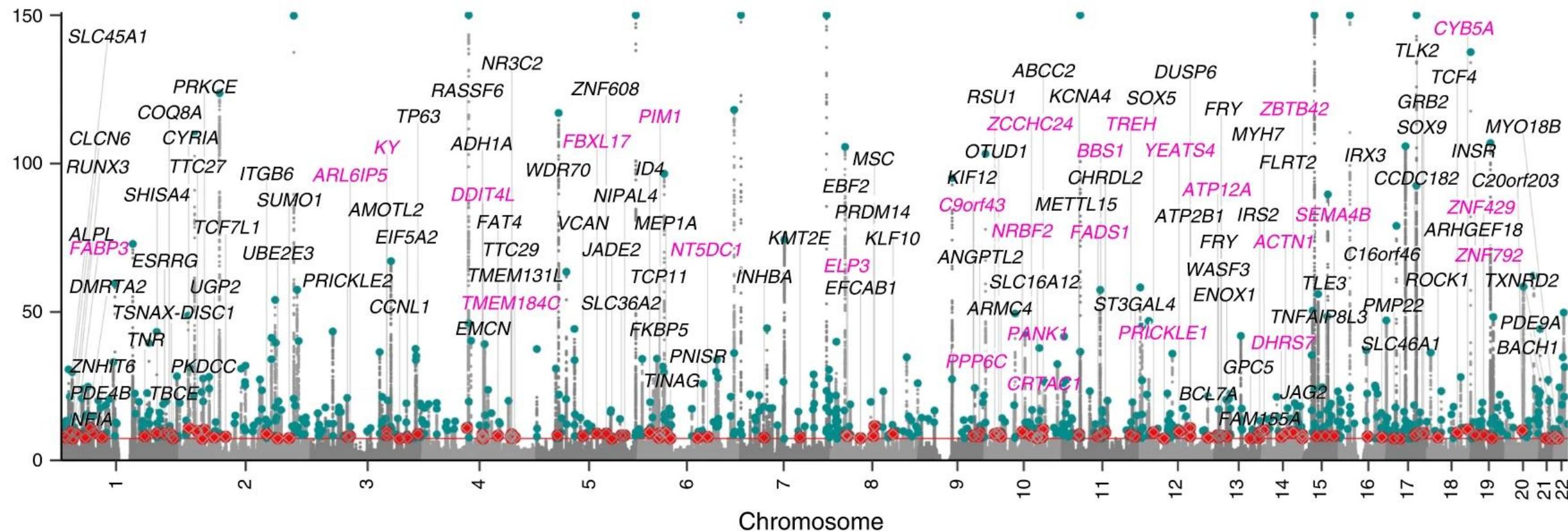
a

eGFRcrea GWAS ( $n = 1,508,659$  individuals)

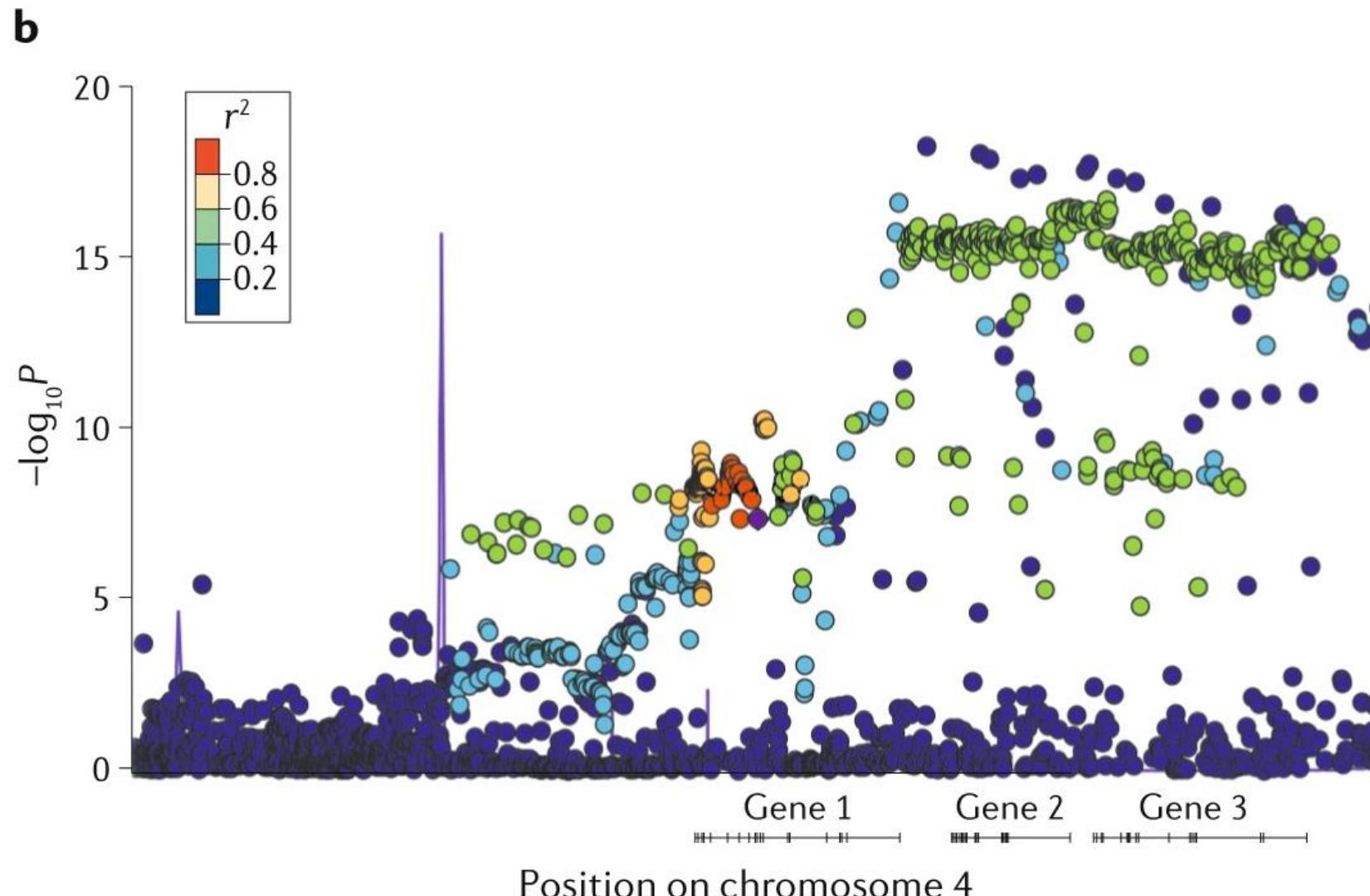
Independent loci ( $n = 878$ )

● Known ( $n = 752$ )

◆ New ( $n = 126$ )



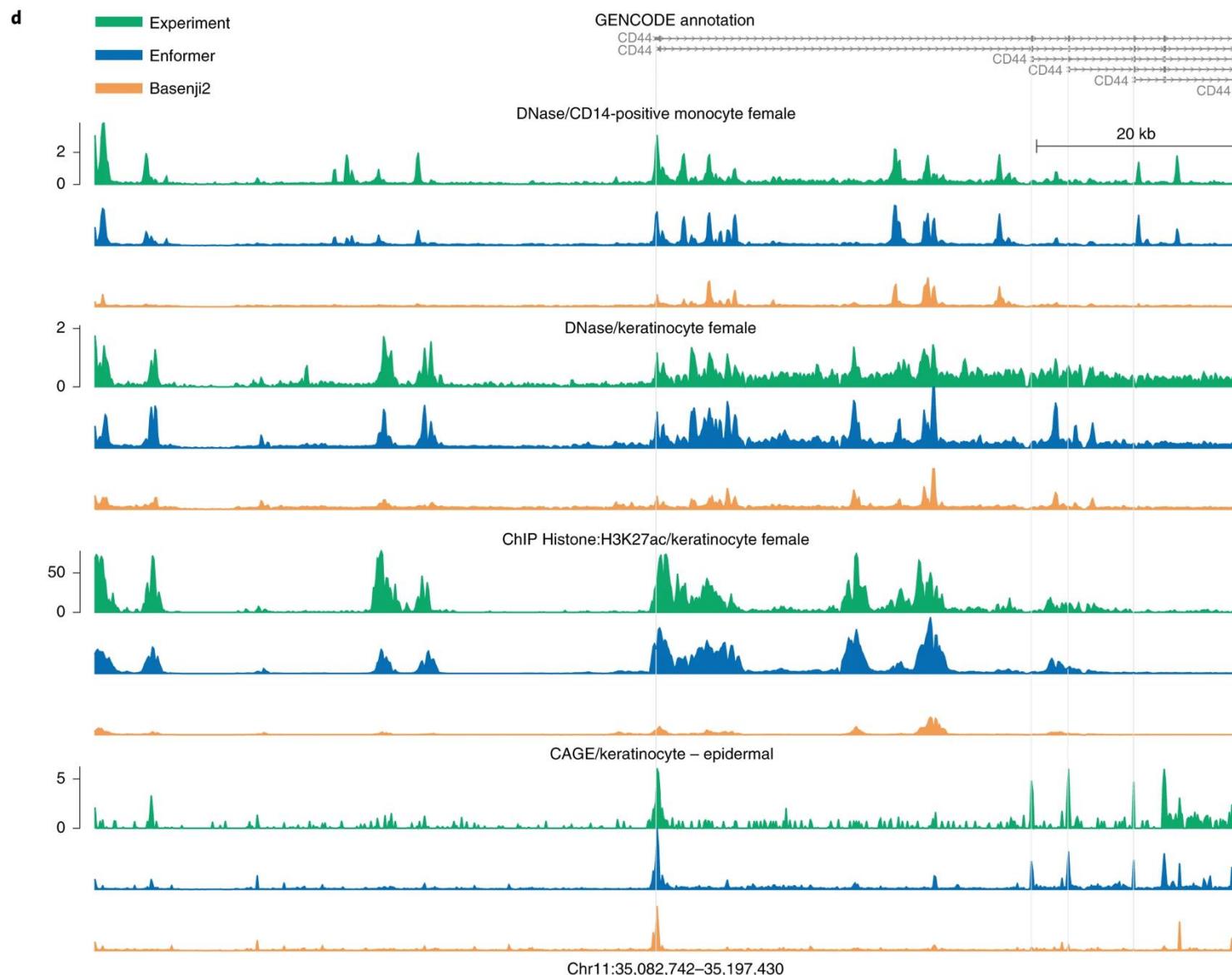
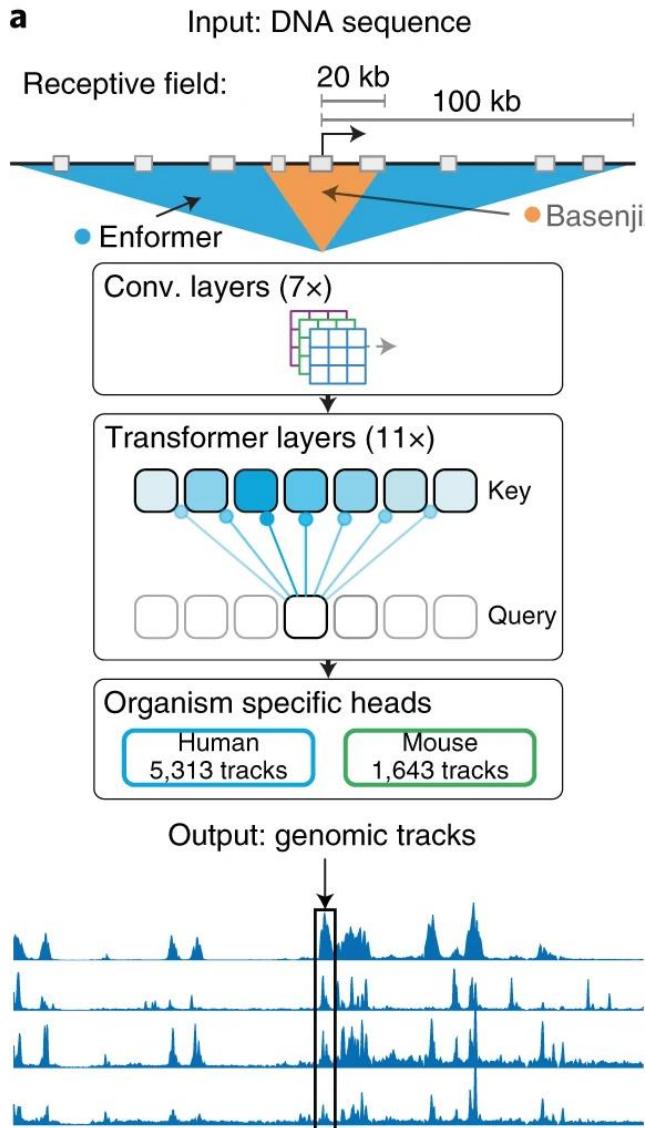
# But associations are typically ambiguous



# Objectives of Borzoi and Beyond

- Learning regulatory grammars from DNA sequences
- Understand Different expression profiles under different cell types and situations
- Precisely identify causal variants within associated loci
- Determine how these variants affect phenotypes
  - Which genes are affected?
  - What does it affect the gene expression?

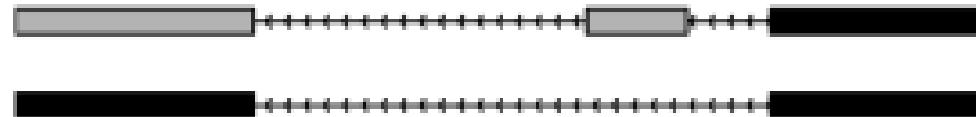
# ML problem – predict assays from sequences



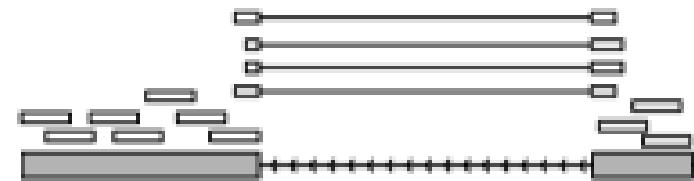
# Why aren't you predicting RNA-Seq?

A

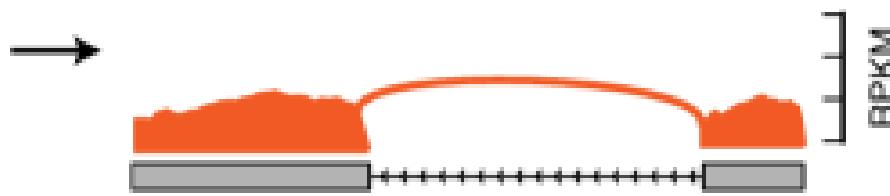
Annotation



Read alignments



Sashimi plot



# RNA-Seq potential

- **Why Revisit RNA-Seq Prediction:**

- Enormous amounts of data, across many more species.
- Abundant RNA-Seq data across species, unlike ATAC-Seq or ChIP-Seq.
- Opportunity to model species lacking other data types.
- Single-cell genomics: sc RNA-Seq prediction is challenging compared to sc ATAC-Seq (scBasset).

- **Advantages of Integrating RNA-Seq:**

- Models could capture multi-layered regulatory processes (e.g., transcription and splicing).
- Post-transcriptional assays readout with RNA-Seq

# New model for RNA-Seq: Borzoi

Enformer (131,072)

x4

524 kilobases (hg38 / mm10)  $2^{19}$

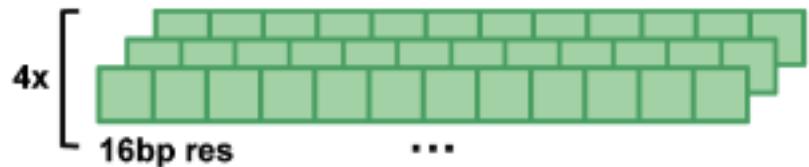


[ChIP  
ATAC/DNase  
CAGE  
RNA] = 7,611  
(human)  
= 2,608  
(mouse)

Coverage Track(s)

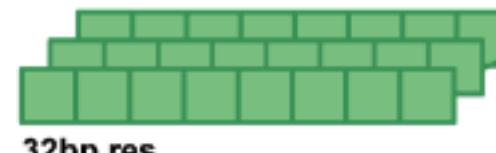
Finer resolution

32bp res

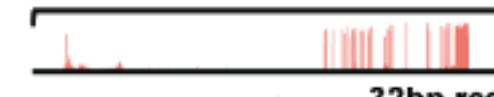


16bp res

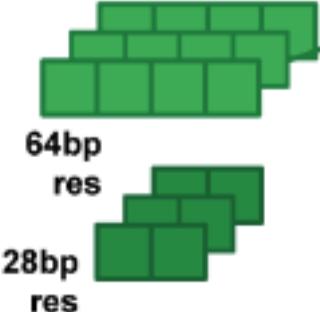
...



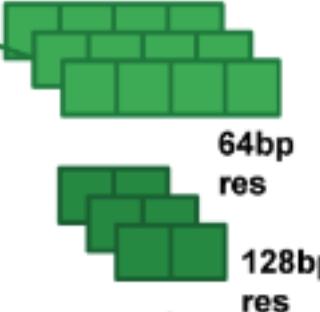
32bp res



32bp res



64bp res



128bp res

64bp res

128bp res

Enformer (128bp)

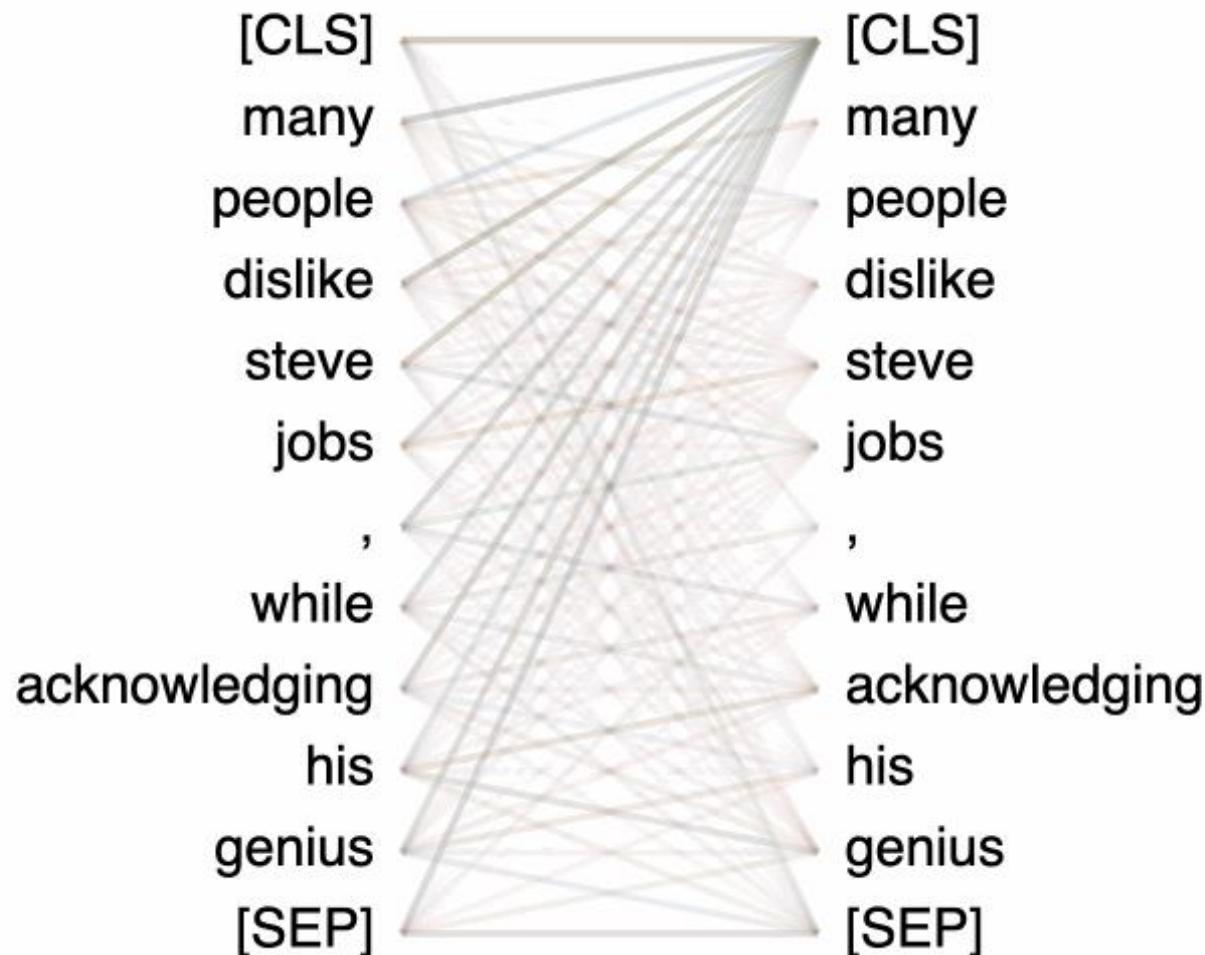


Transformer  
Blocks (8x)

128-bp resolution because it roughly represents a well-studied length of regulatory elements

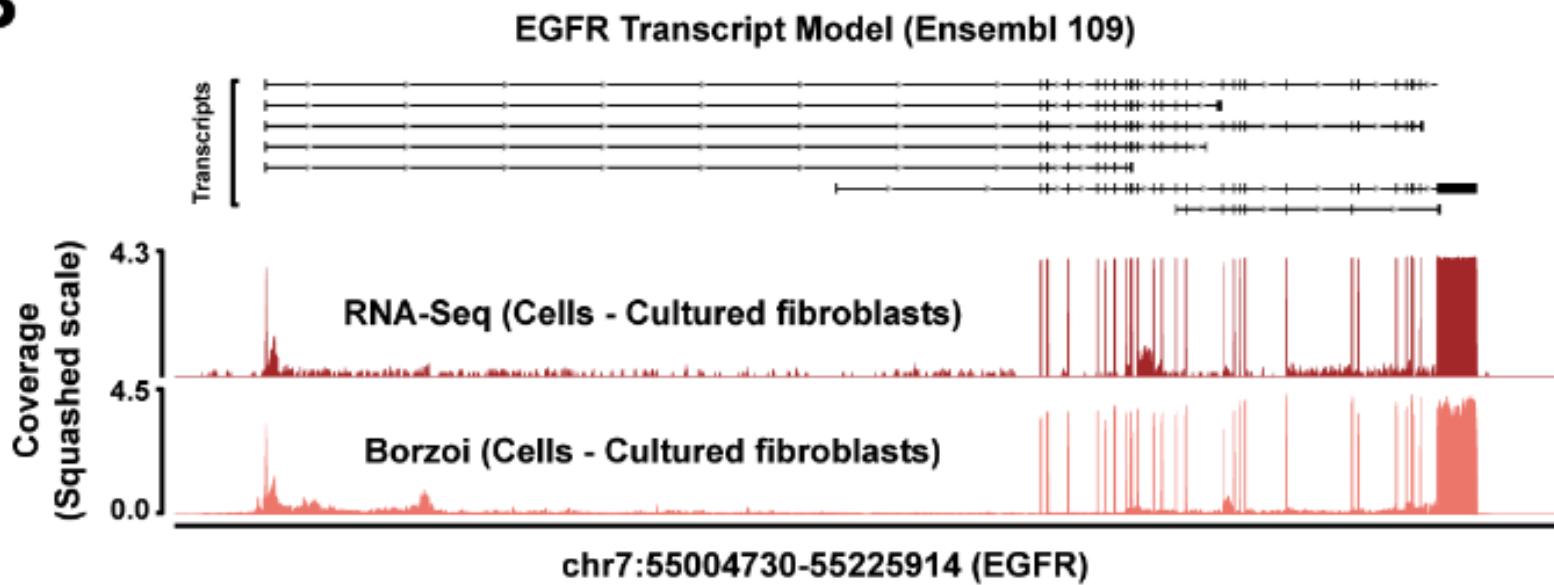
# Self-attention

Layer: 0 ▾ Attention: All ▾

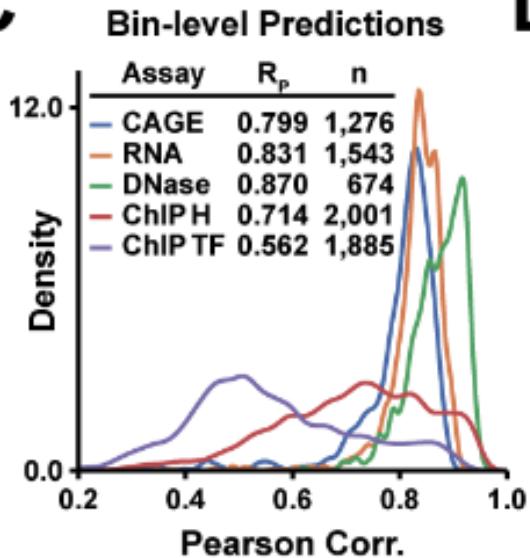


# Borzoi predicts RNA-Seq well

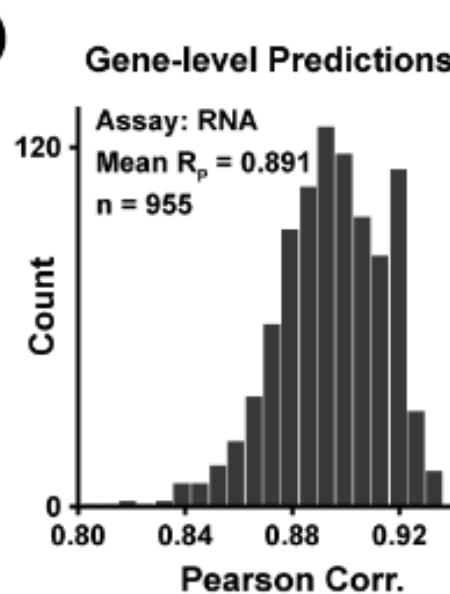
**B**



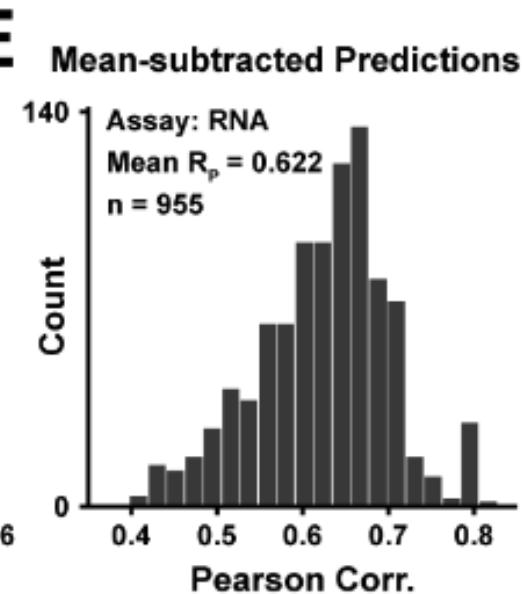
**C**



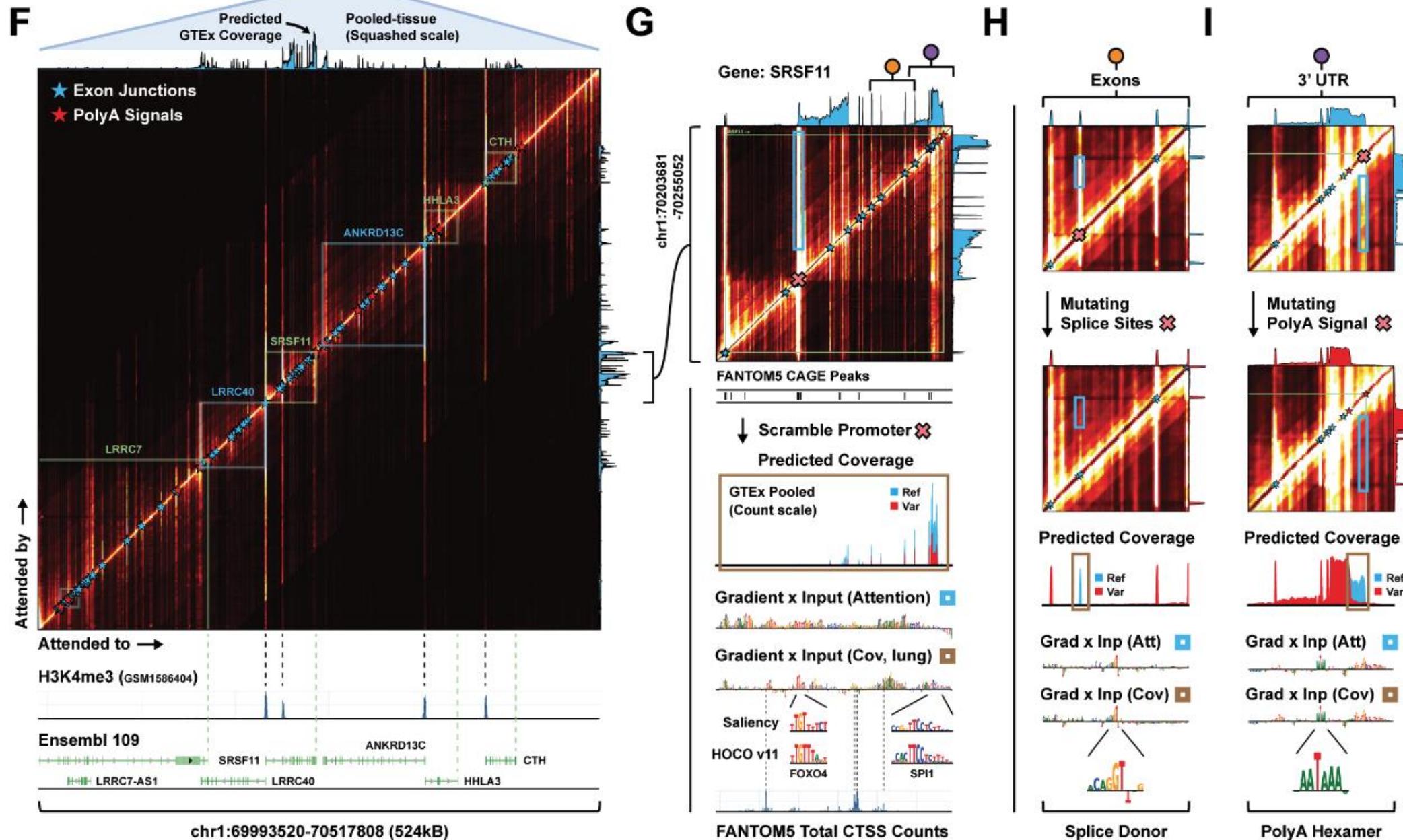
**D**



**E**

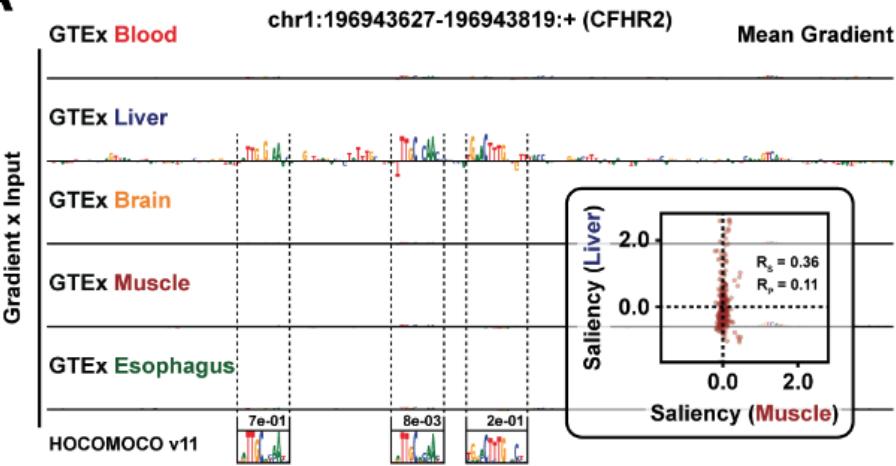


# Borzoi's attention map interpretation

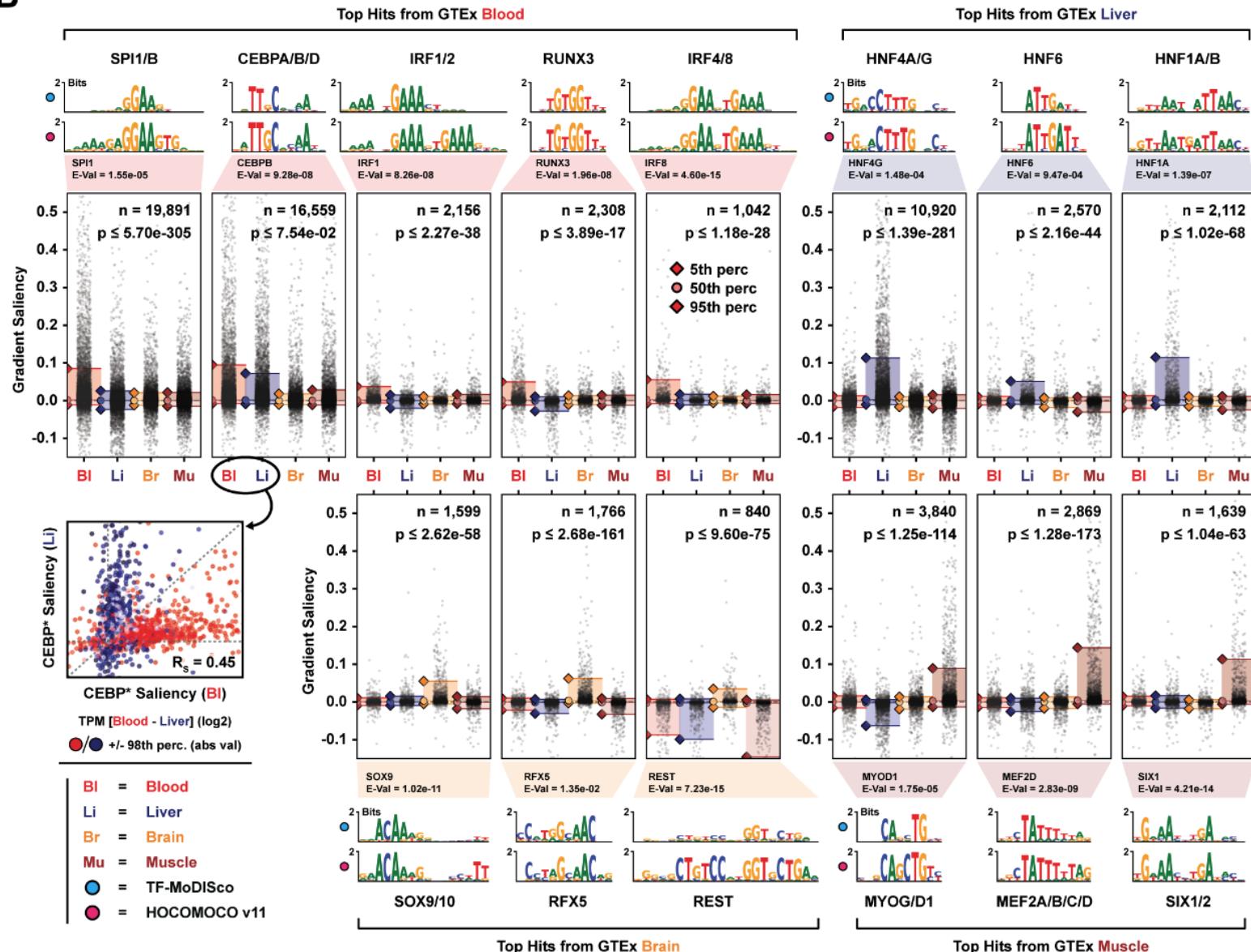


# TF motif discovery from Borzoi's gradient x input

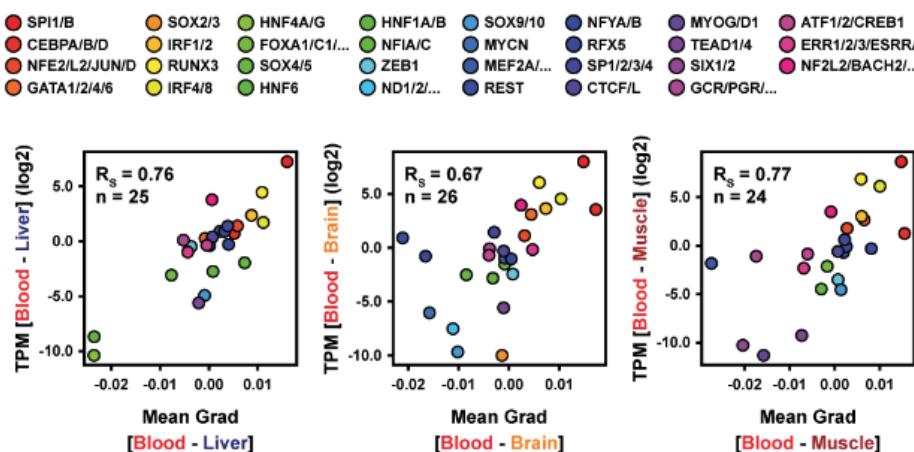
**A**



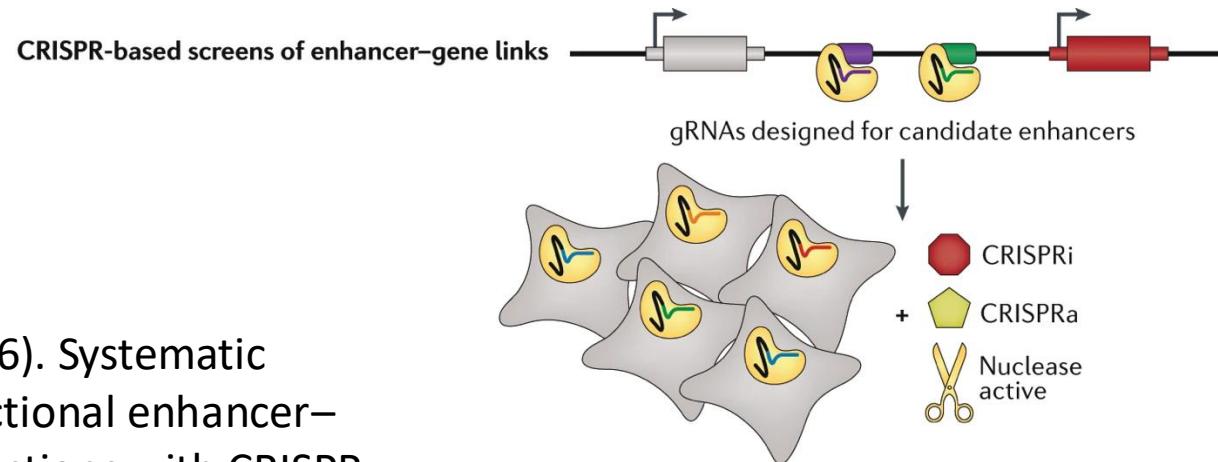
**B**



**C**

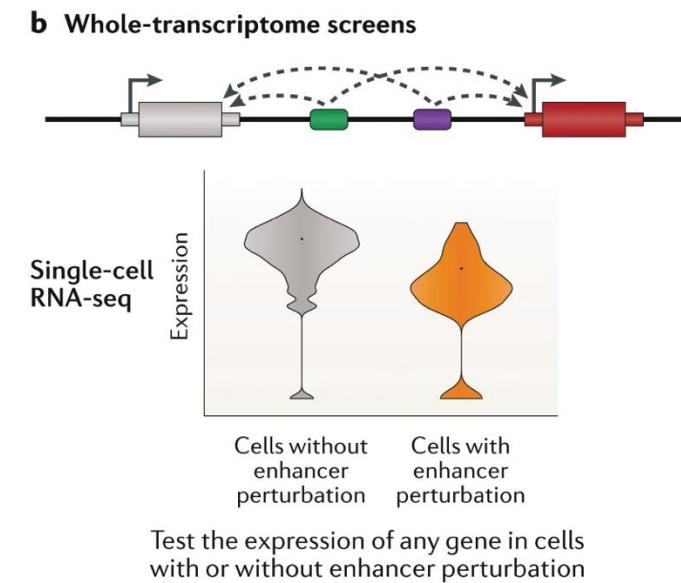
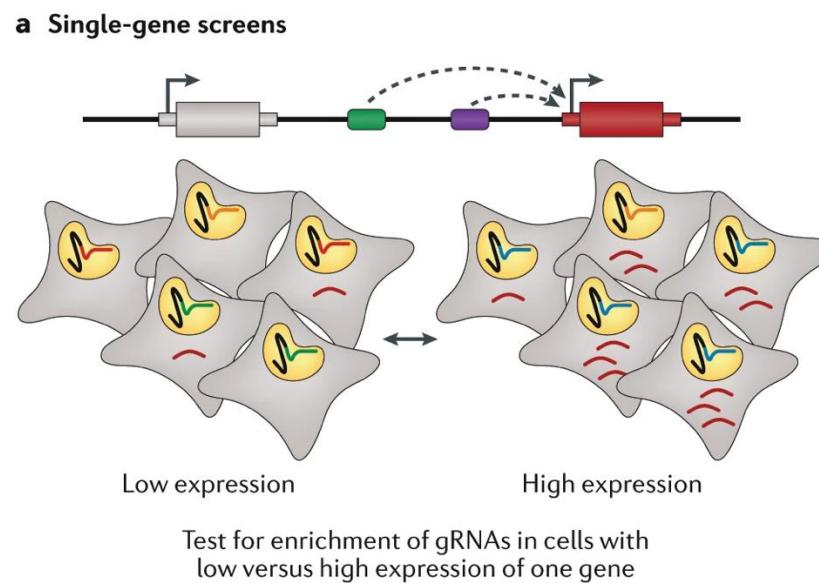


# Benchmarking long range usage w/CRISPR screens



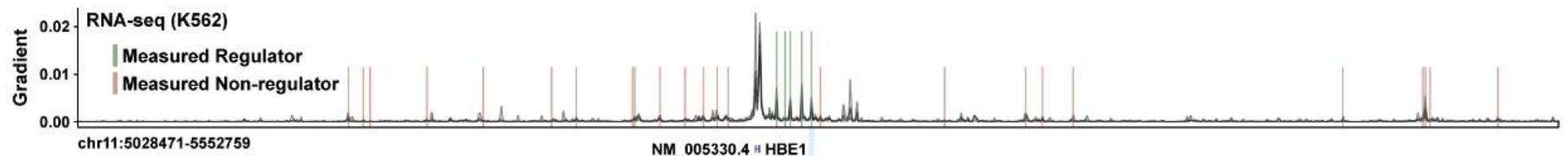
Fulco et al. (2016). Systematic mapping of functional enhancer–promoter connections with CRISPR interference. *Science*

Gasperini et al. (2019). A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell*

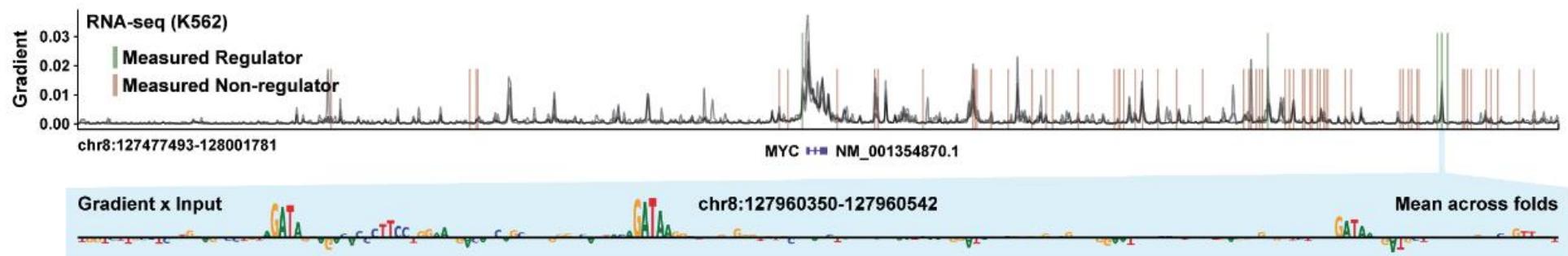


# CRISPR enhancer screens

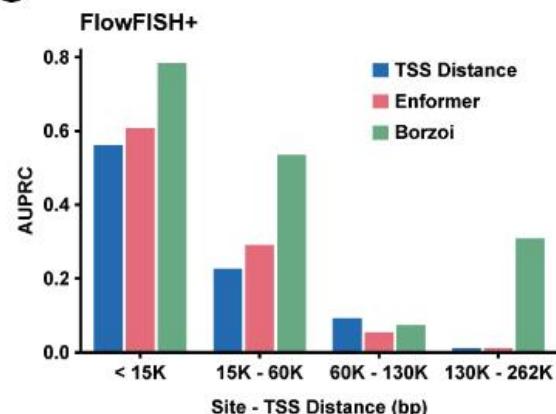
A



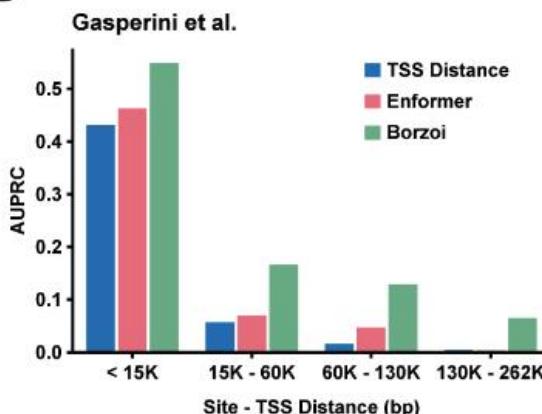
B



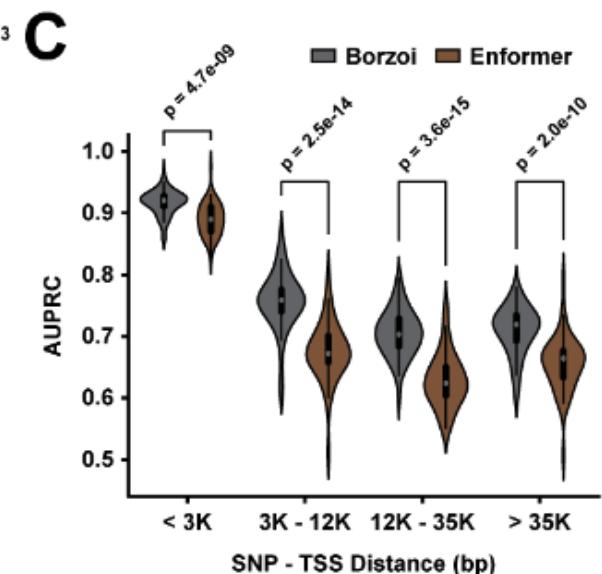
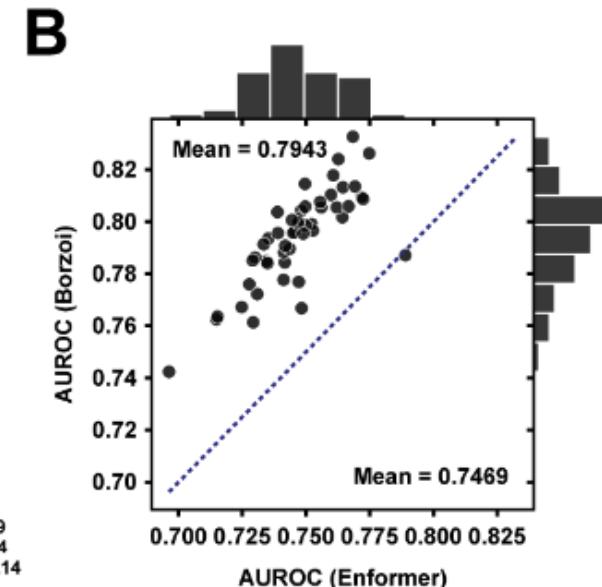
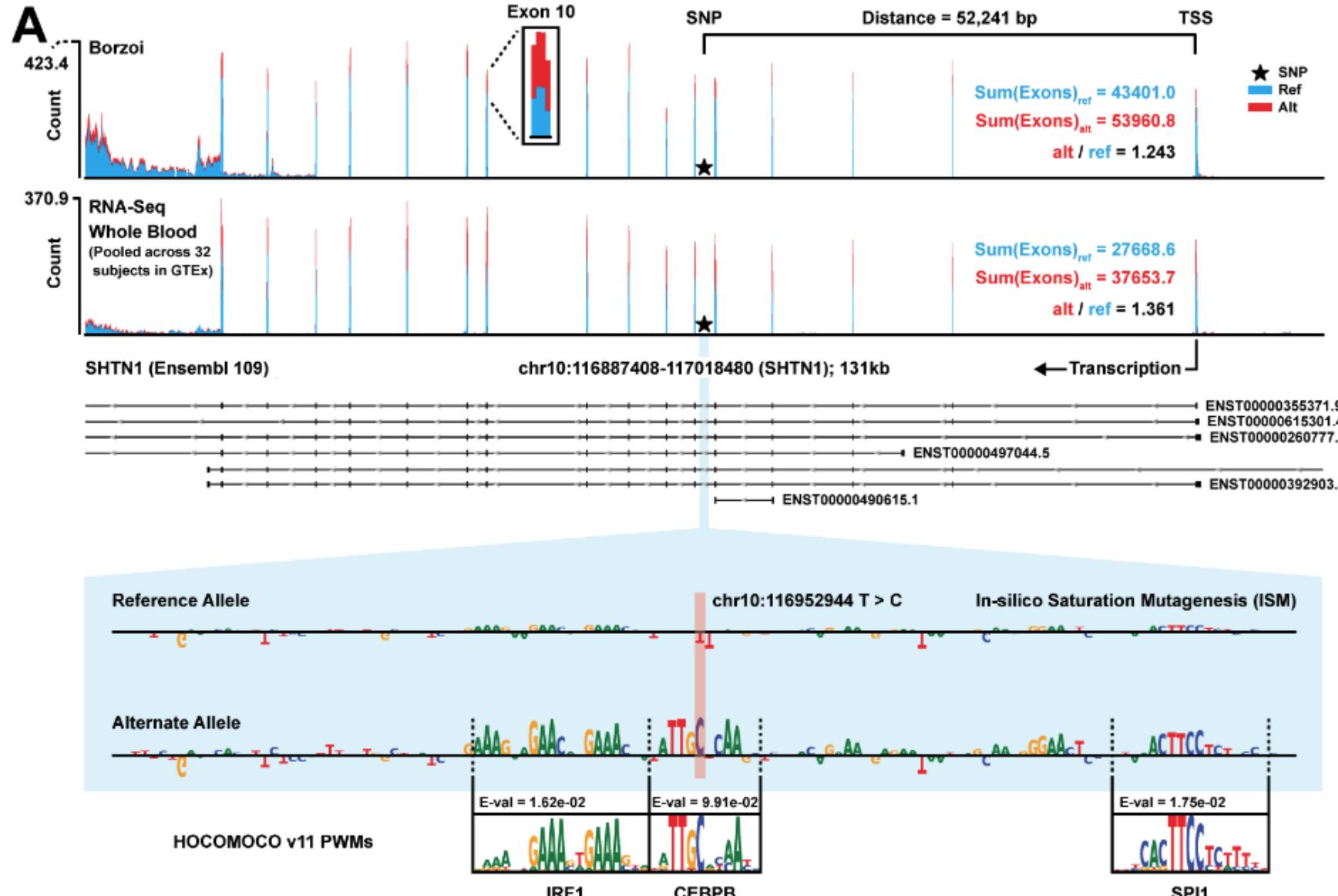
C



D



# eQTL – Noncoding variant interpretation



# Improved prediction on eQTL coefficients & Limitation

nature genetics

Explore content ▾ About the journal ▾ Publish with us ▾

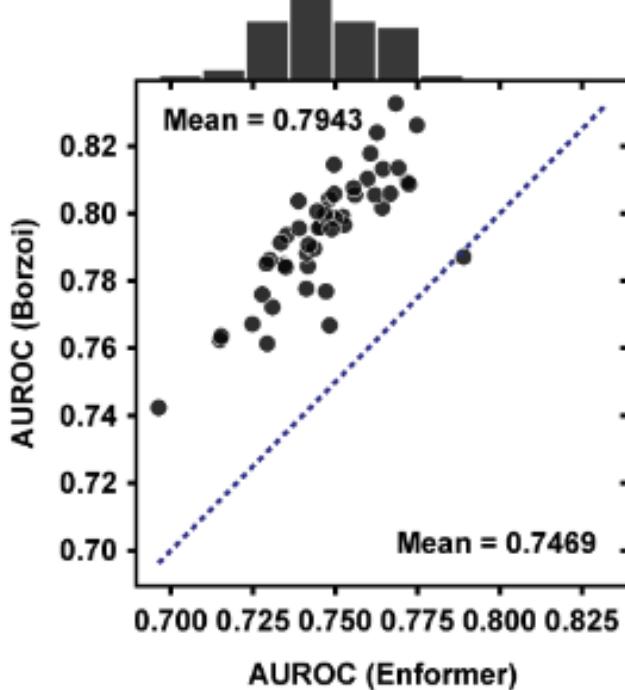
nature > nature genetics > letters > article

Letter | Published: 30 November 2023

## Benchmarking of deep neural networks for predicting personal gene expression from DNA sequence highlights shortcomings

Alexander Sasse, Bernard Ng, Anna E. Spiro, Shinya Tasaki, David A. Bennett, Christopher Gaiteri, Philip L. De Jager, Maria Chikina & Sara Mostafavi

B



nature genetics

Explore content ▾ About the journal ▾ Publish with us ▾

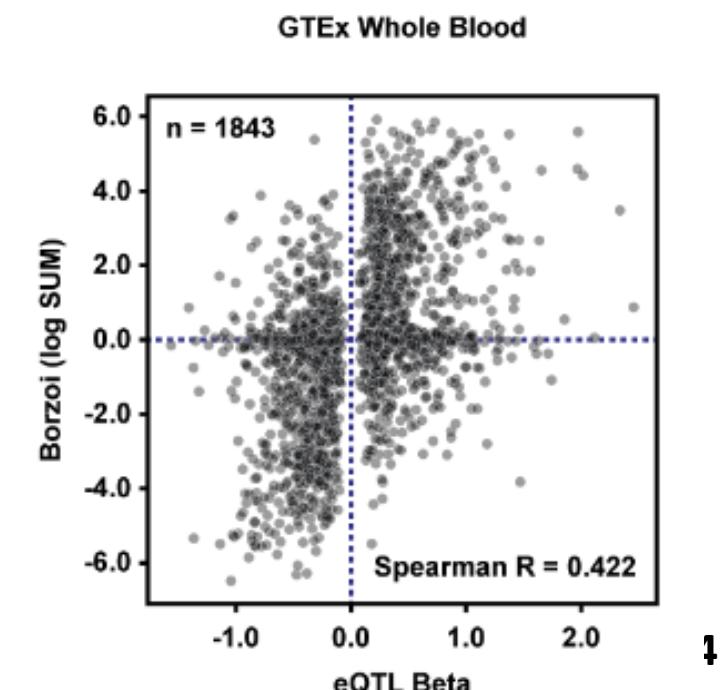
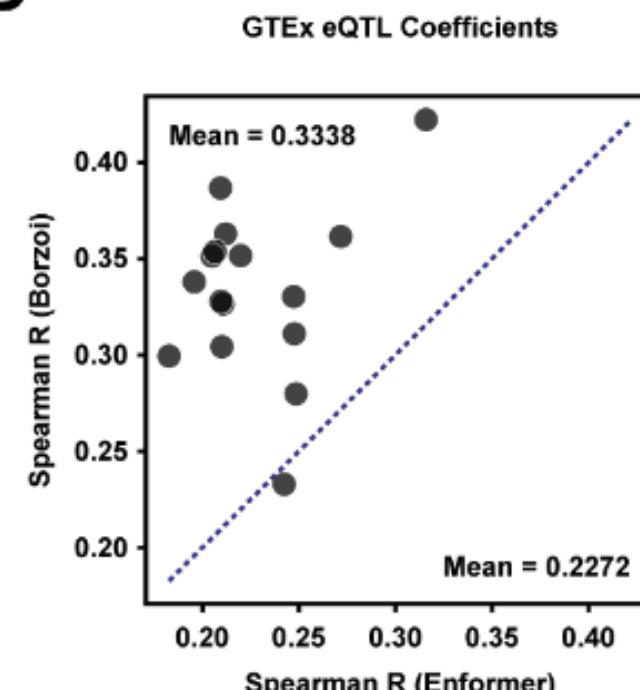
nature > nature genetics > brief communications > article

Brief Communication | Open access | Published: 30 November 2023

## Personal transcriptome variation is poorly explained by current genomic deep learning models

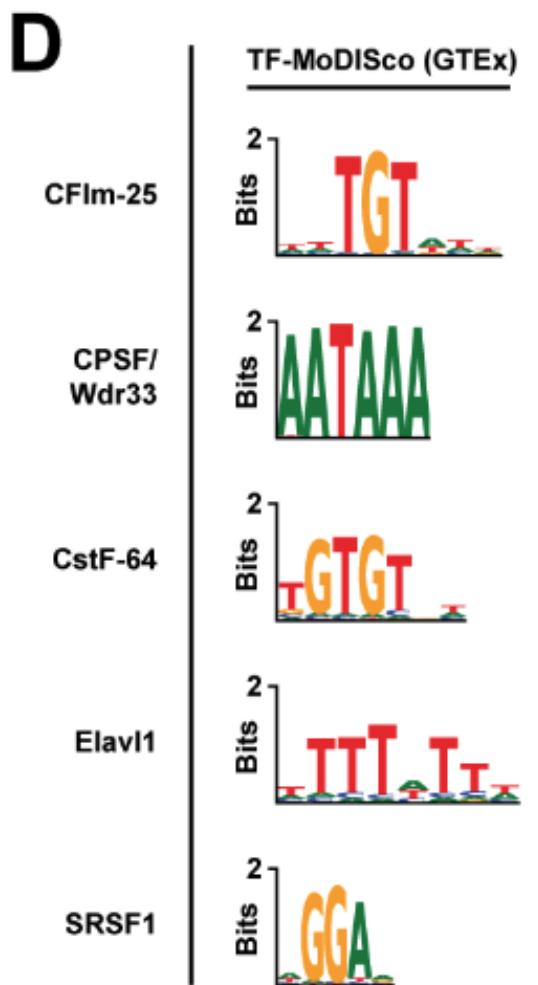
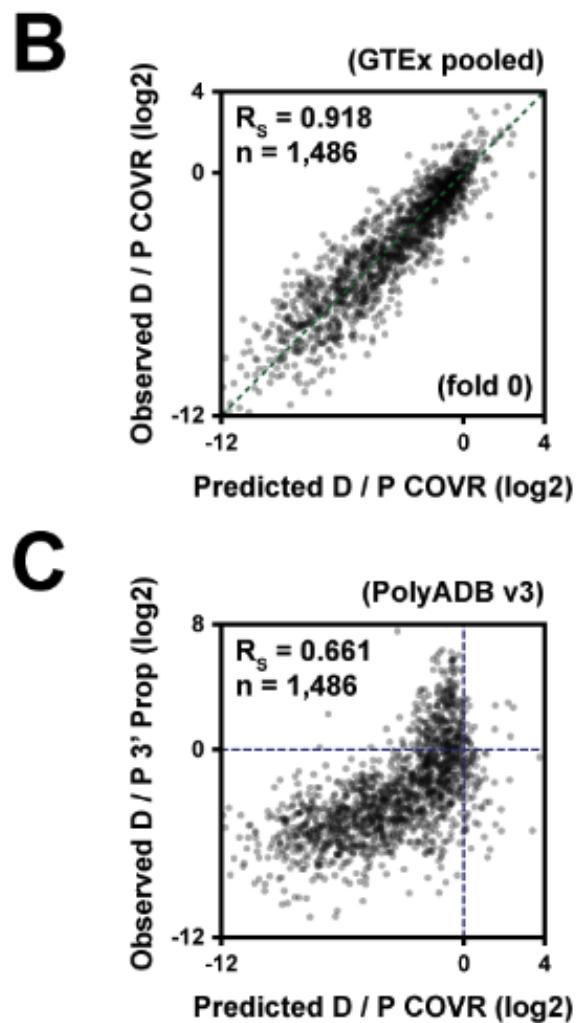
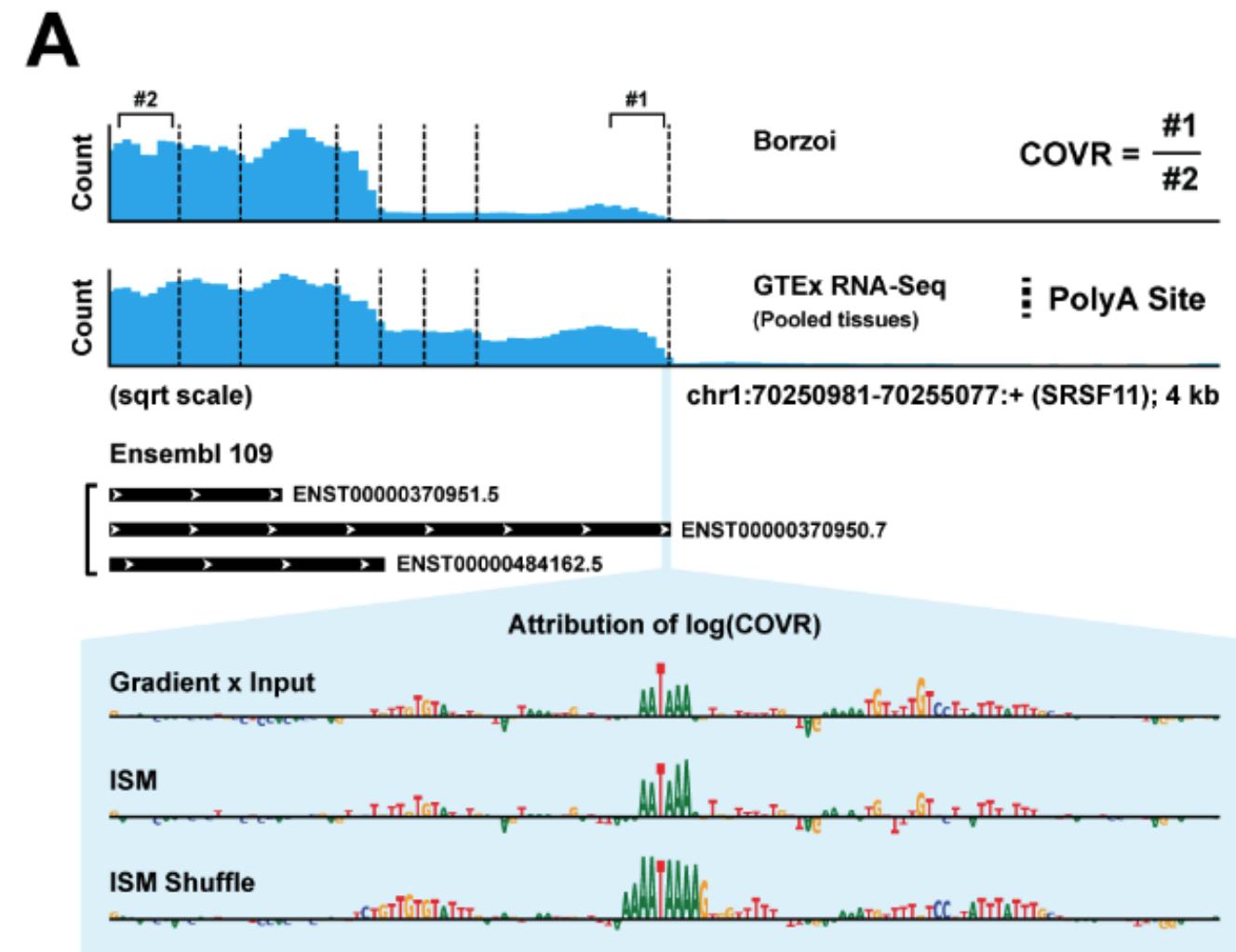
Connie Huang, Richard W. Shuai, Parth Baokar, Ryan Chung, Ruchir Rastogi, Pooja Kathail & Nilah M. Ioannidis

D



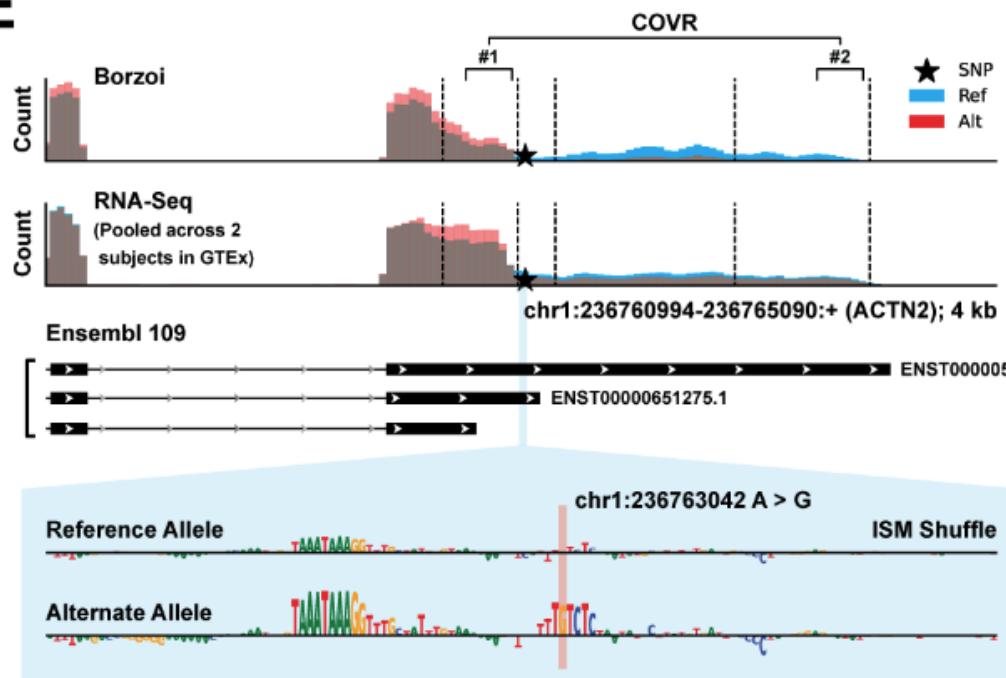
↓

# Polyadenylation



# Polyadenylation QTL

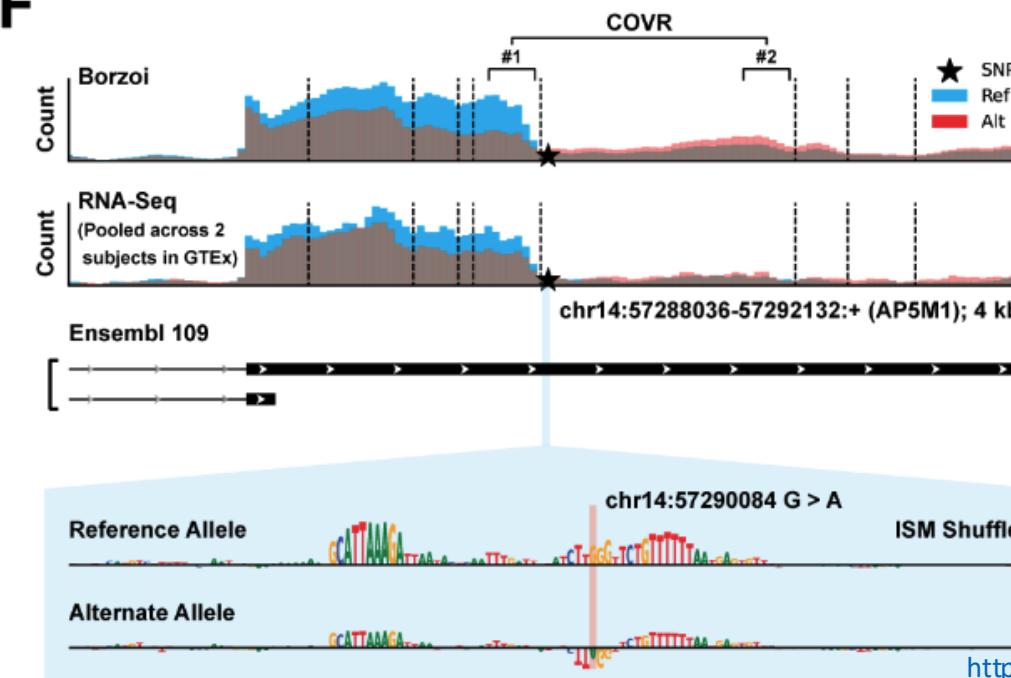
E



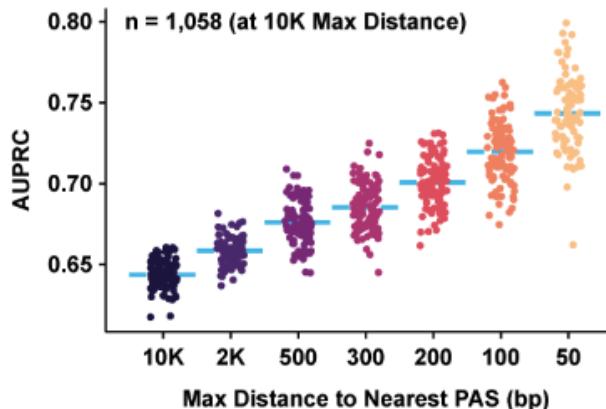
## eQTL Catalog

Kerimov, et al. (2021). A compendium of uniformly processed human gene expression and splicing quantitative trait loci. Nature genetics

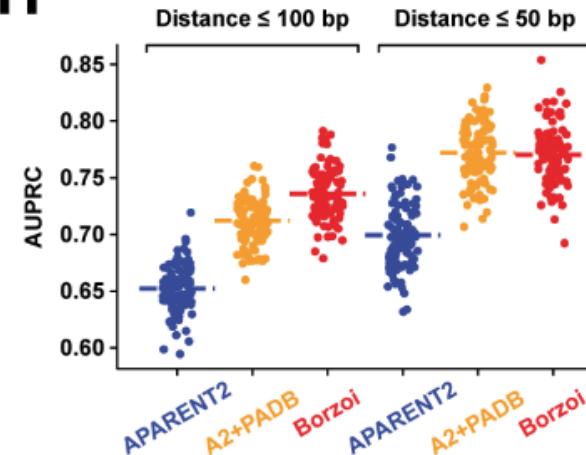
F



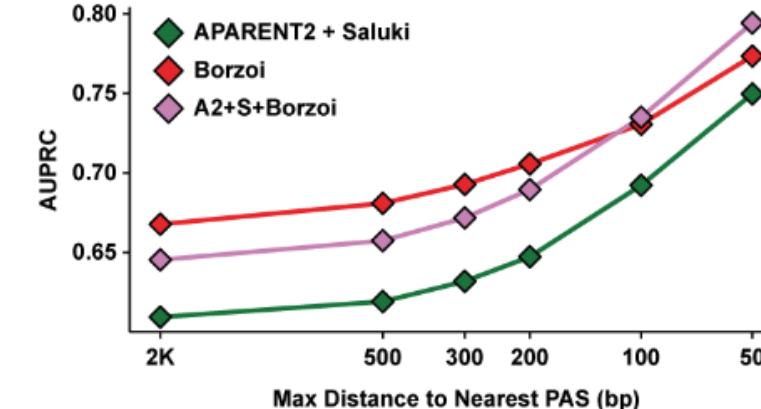
G



H



I

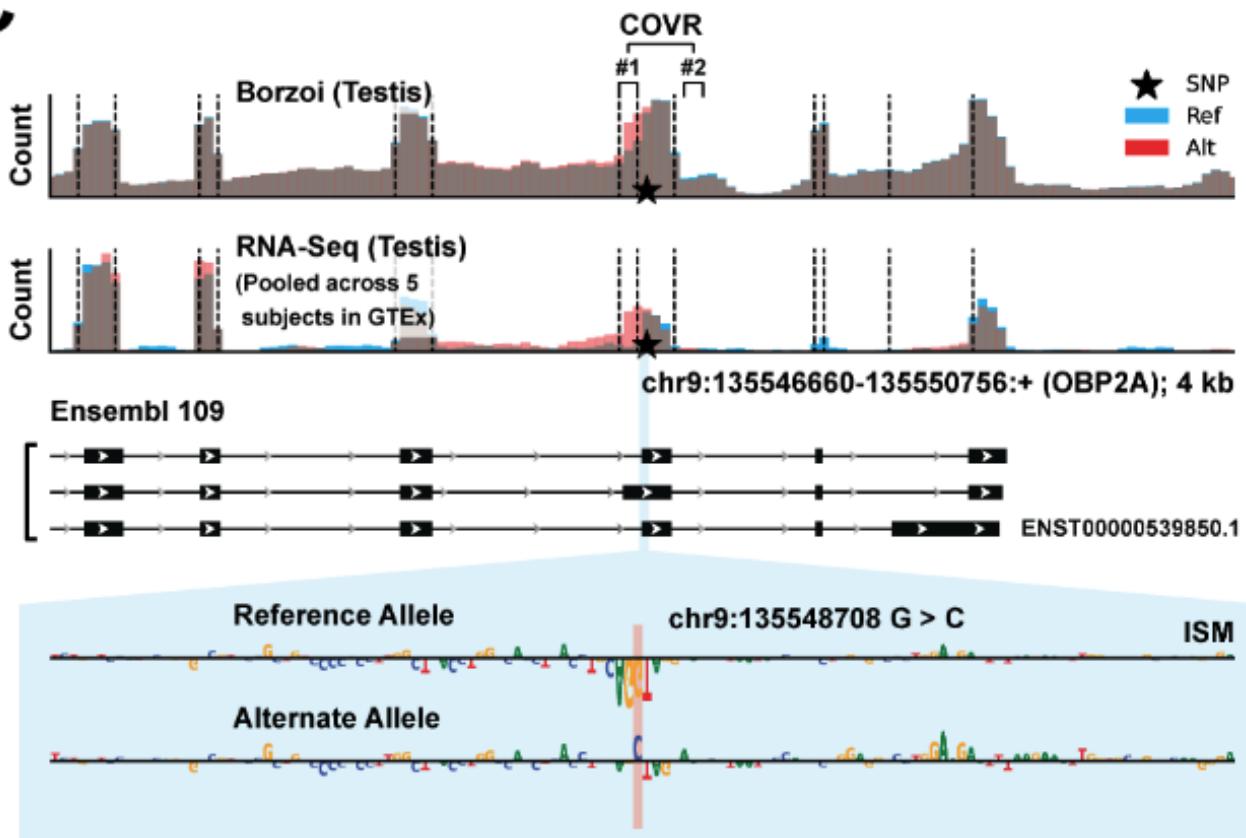


<https://github.com/stephenslab/susieR>

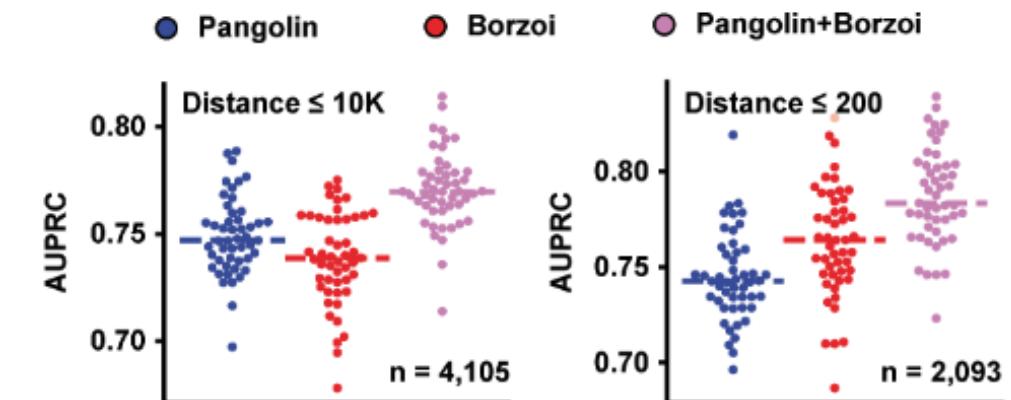
Zou et al. (2022). Fine-mapping from summary data with the “Sum of Single Effects” model. PLoS genetics

# Splicing QTL

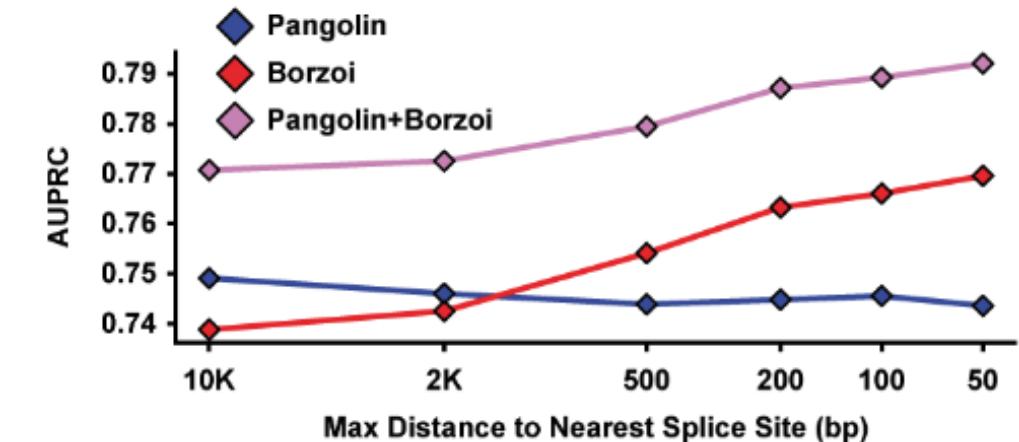
C



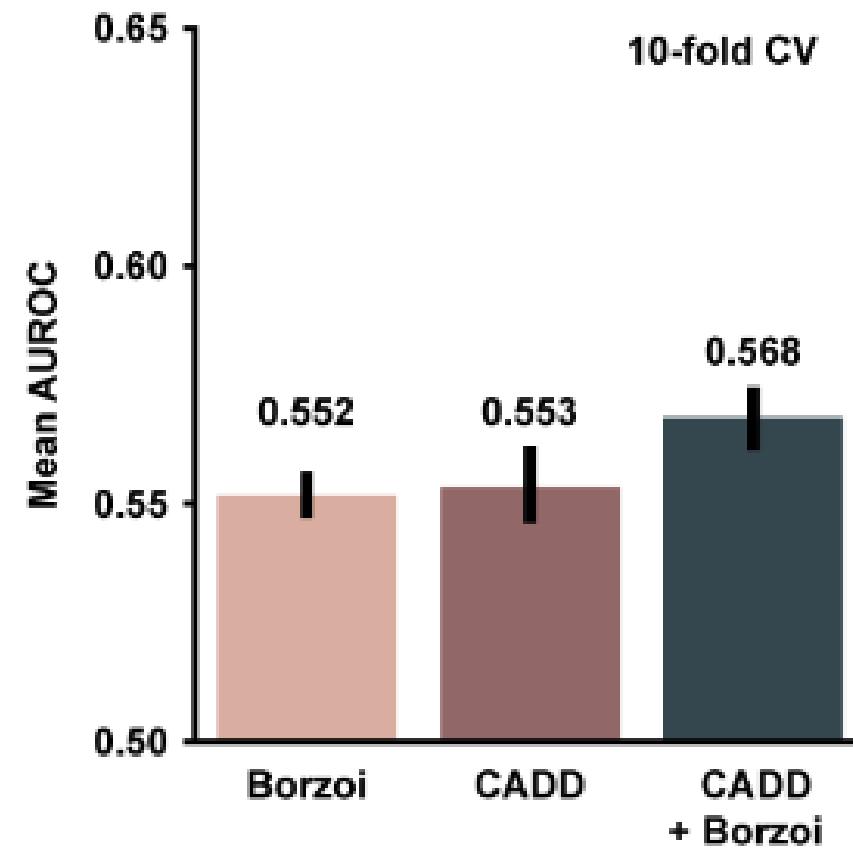
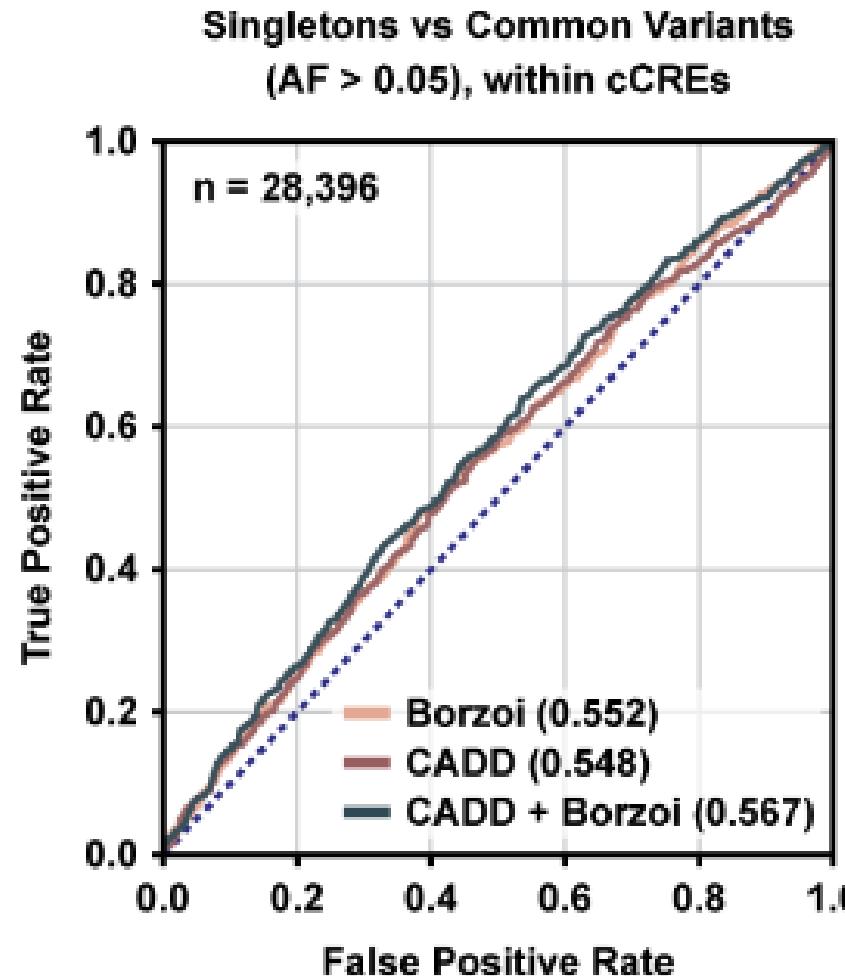
D



E



# Classifying singleton variants from common variation



# Summary

- Introduced the new model, Borzoi, to predict RNA-Seq data, offering significant improvements in transcriptional regulation benchmarks compared to Enformer.
- Borzoi also allows us to study splicing and polyadenylation at near state-of-the-art levels.
- The variant effect predictions from Borzoi are shown in the paper. Goals are to gain more insights from GWAS.
- The goal is to empower these studies with more precise predictions, facilitating discoveries in common diseases

# Next Study Group

- **Date and Time:** November 5<sup>th</sup> , 12:00 pm - 1:00 pm.
- **Presenter:** Mahler Revsine
- **Location:** Room 228 at Malone or on zoom
- **Paper:** HyenaDNA: long-range genomic sequence modeling at single nucleotide resolution
  - <https://arxiv.org/abs/2306.15794>

- We will share slides in the slack channel [#deep-learning-reading-group](#)
- We encourage you to register for presentation
- We will share the schedule
- See you all on November 5<sup>th</sup> !