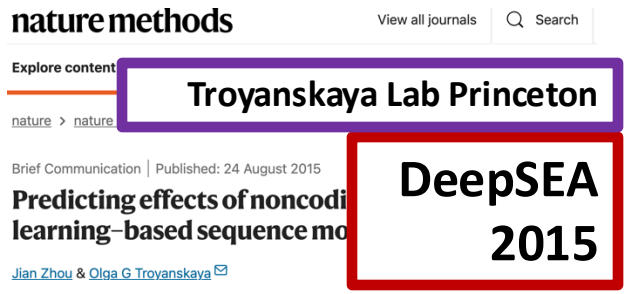
Deep Learning + Genomics Study Group

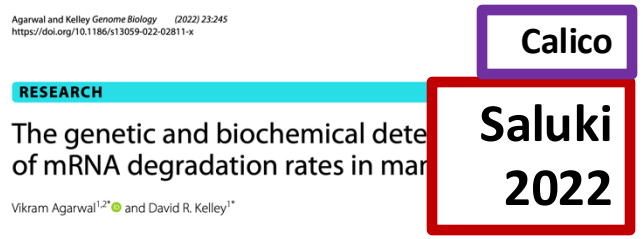
Kuan-Hao Chao / Mahler Revsine

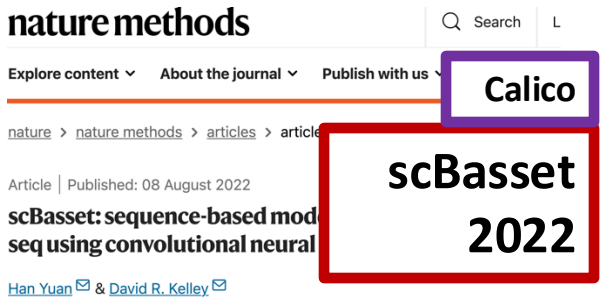
2024.10.22

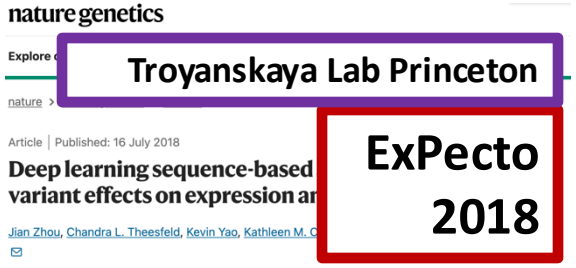
Why are we starting this study group?

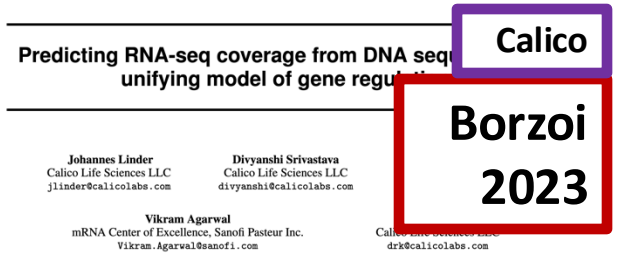
Deep learning-based DNA sequence model



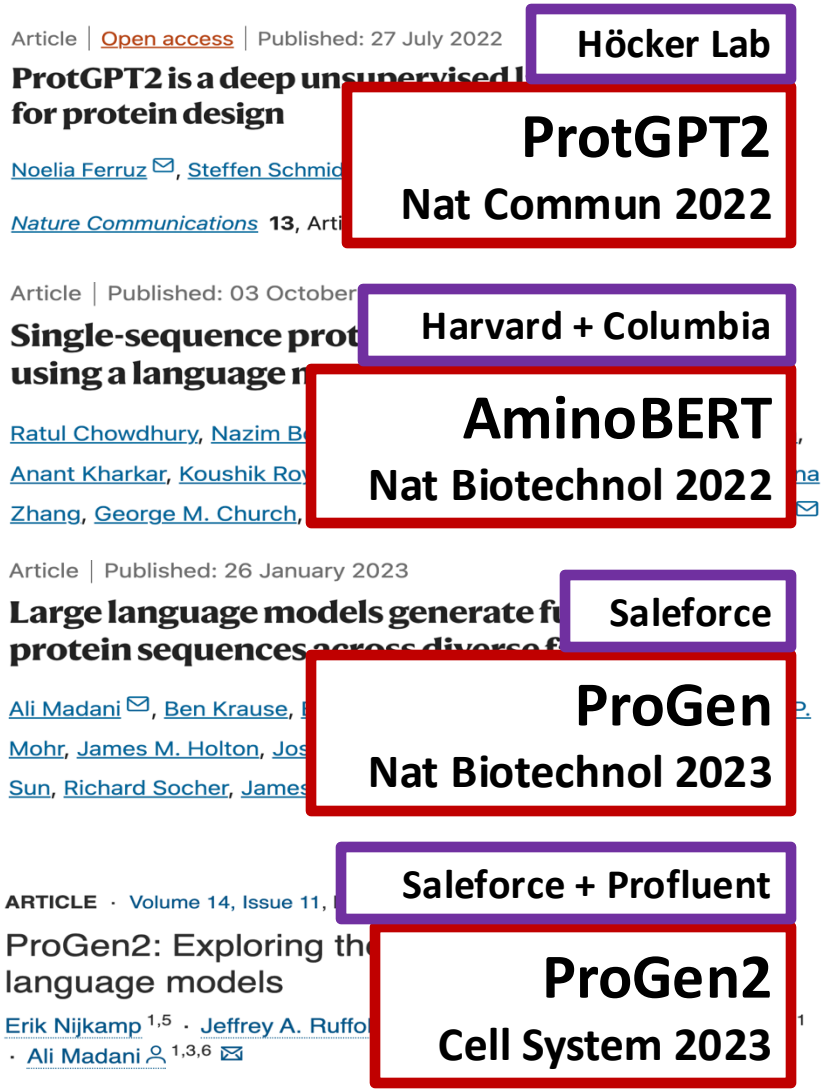


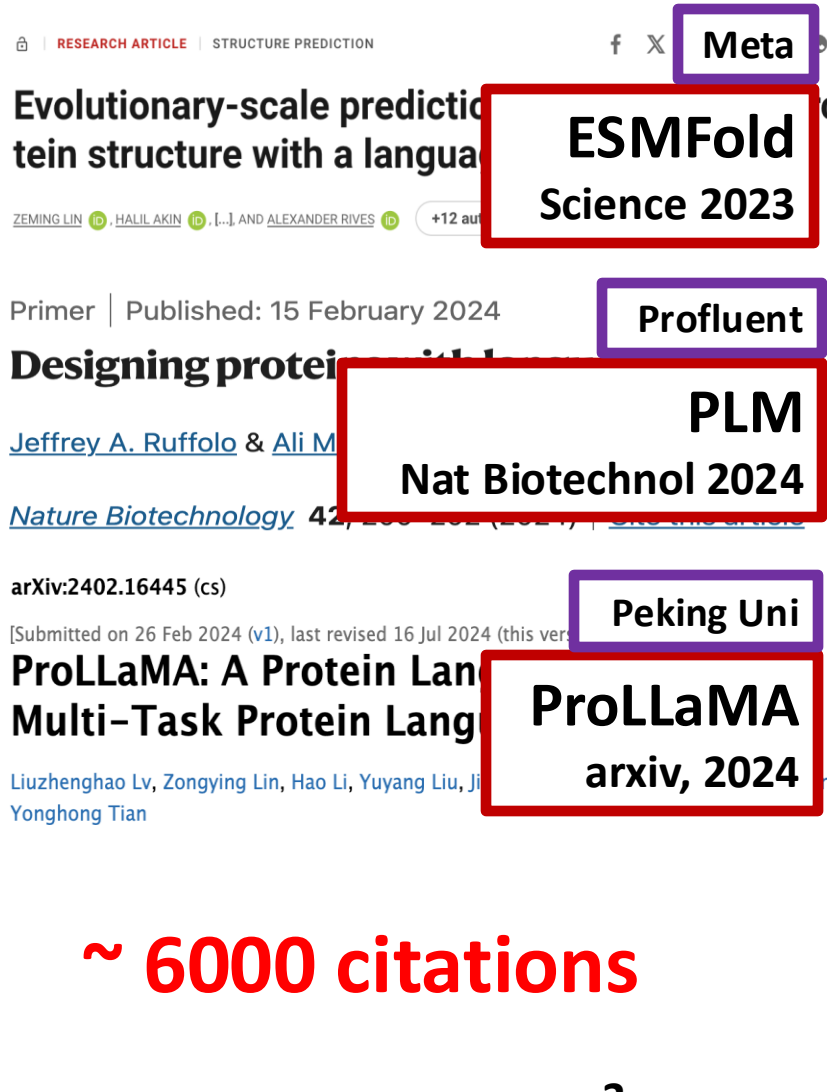






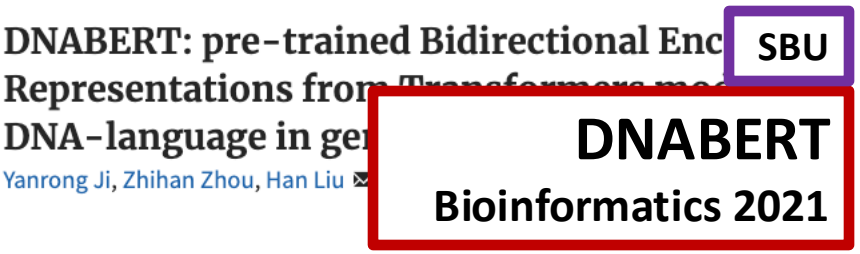
Protein Language model





~ 6000 citations

DNA Language model



SBU

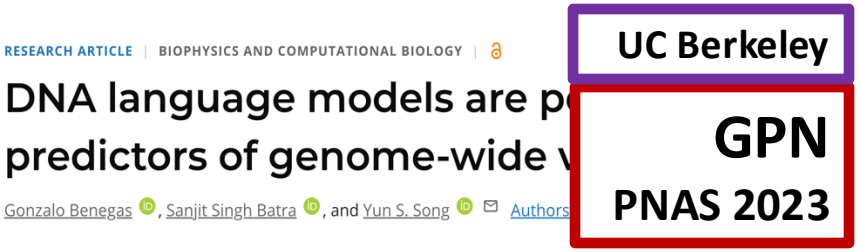
DNABERT
Bioinformatics 2021

SBU + NU

DNABERT-2
ICLR 2024

Stanford + Arc Inst + TogetherAI

Evo
bioRxiv 2024

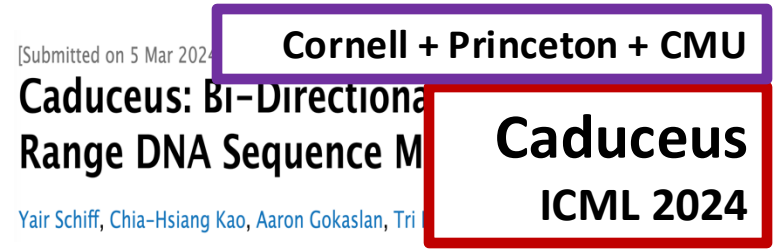


UC Berkeley

GPN
PNAS 2023

TUD

GROVER
Nat Mach Intell 2024



Cornell + Princeton + CMU

Caduceus
ICML 2024

InstaDeep + Nvidia + TUM

Nucleotide Transformer
bioRxiv 2023

Harvard + MIT

Genomic LM
Nat Commun 2024

Cornell + USDA-ARS + Simons

PlantCaduceus
bioRxiv 2024

Stanford

HyenaDNA
NeurIPS 2023

TUM

Species-aware DNA LM
Genom Biol 2024

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 22nd, 12:00 pm - 1:00 pm. We'll release the schedule before the next meeting (November 5th , 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=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/1Ufsju9EwDYyEU82OnT0W5ZkXmrbx2iaP0TTBdesW79k/edit?usp=sharing.

Links

- **Schedule for Fall 2024:**
  - https://docs.google.com/spreadsheets/d/1mRFnzRyX5ThY69i_ne4oGkJJjmoyl4N0d_7tl2tC1Ts/edit?usp=sharing

- **Presenter registration form:**
  - https://docs.google.com/forms/d/e/1FAIpQLSfgNDPOMdSNxO3OT2qdUhoPVETKvxtOxF7T-3dDLJlyilaMsg/viewform?usp=sf_link

Reference: 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

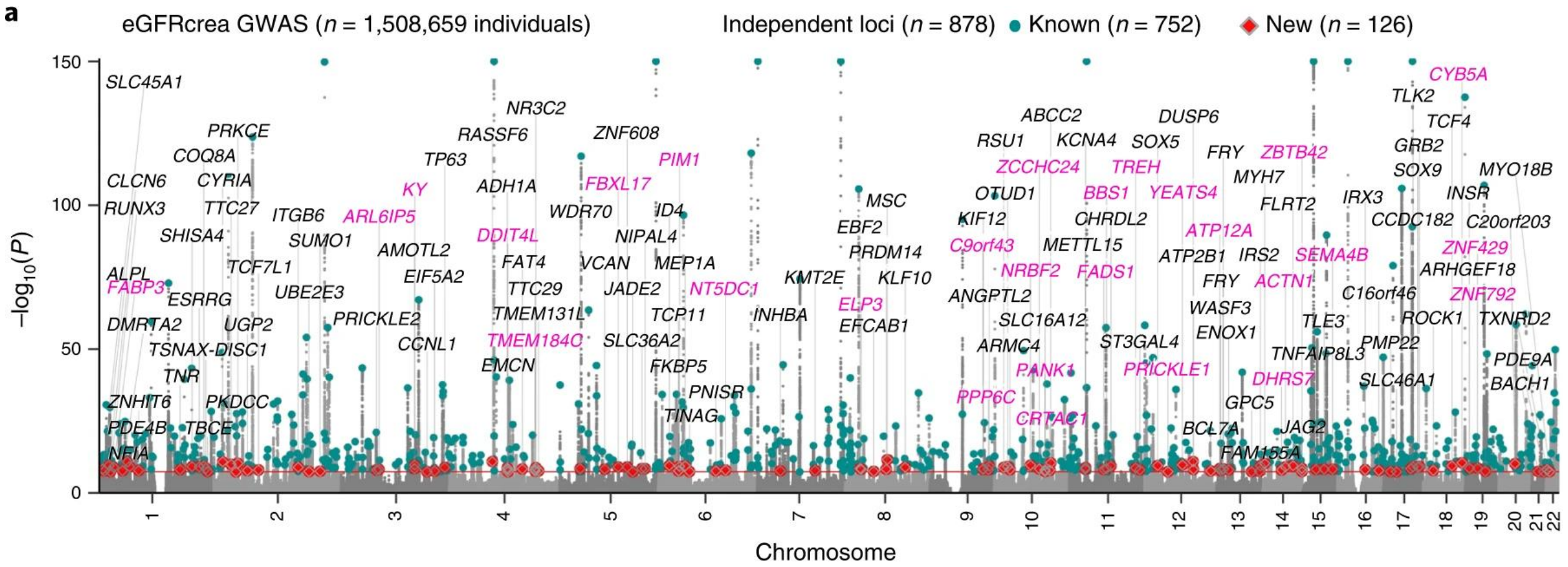
2024.10.22

khchao.com

@KuanHaoChao

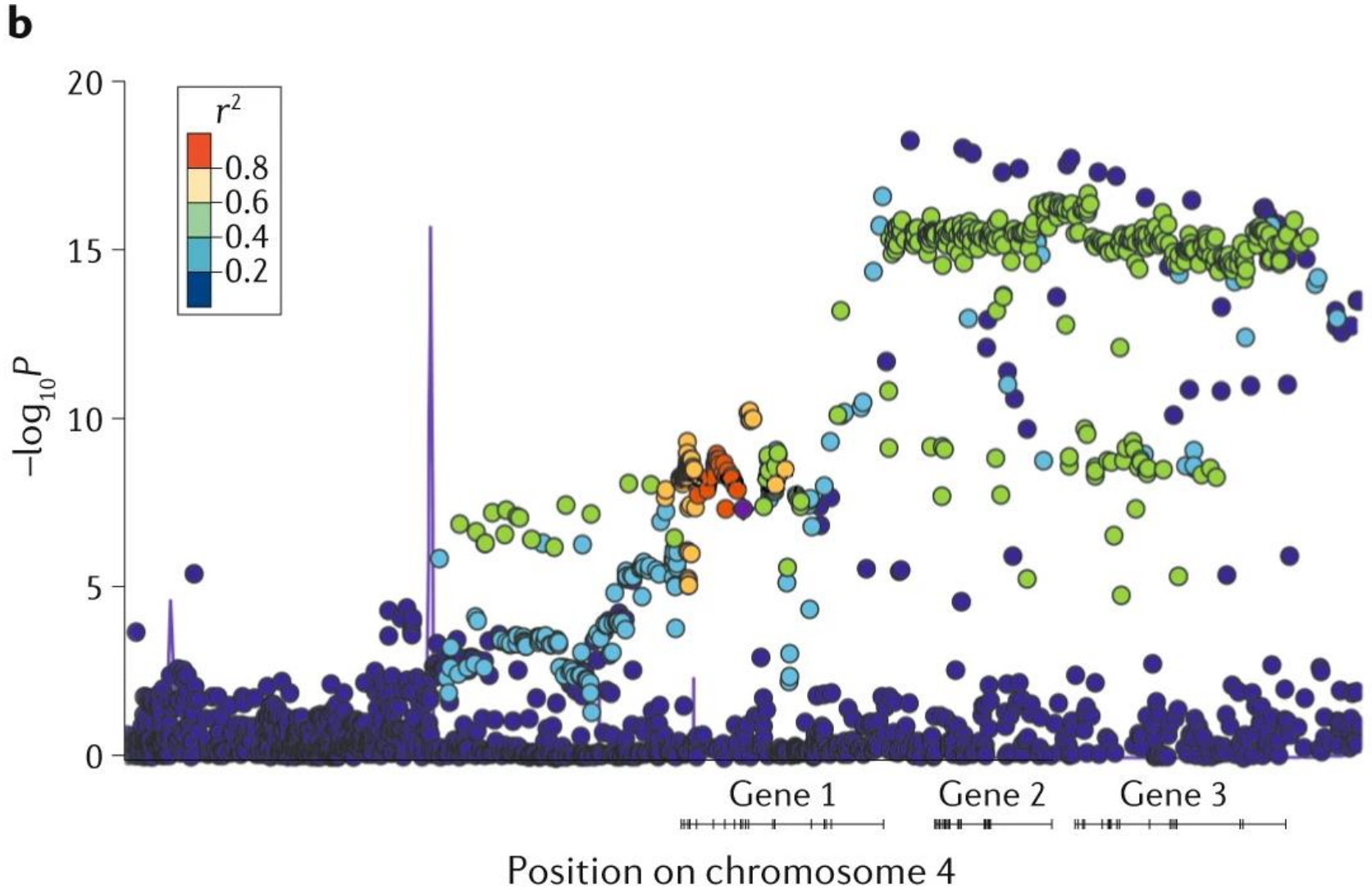
Kuanhao-Chao

GWAS highlights MANY genomic loci



Liu et al. (2022). Epigenomic and transcriptomic analyses define core cell types, genes and targetable mechanisms for kidney disease. Nature genetics

But associations are typically ambiguous

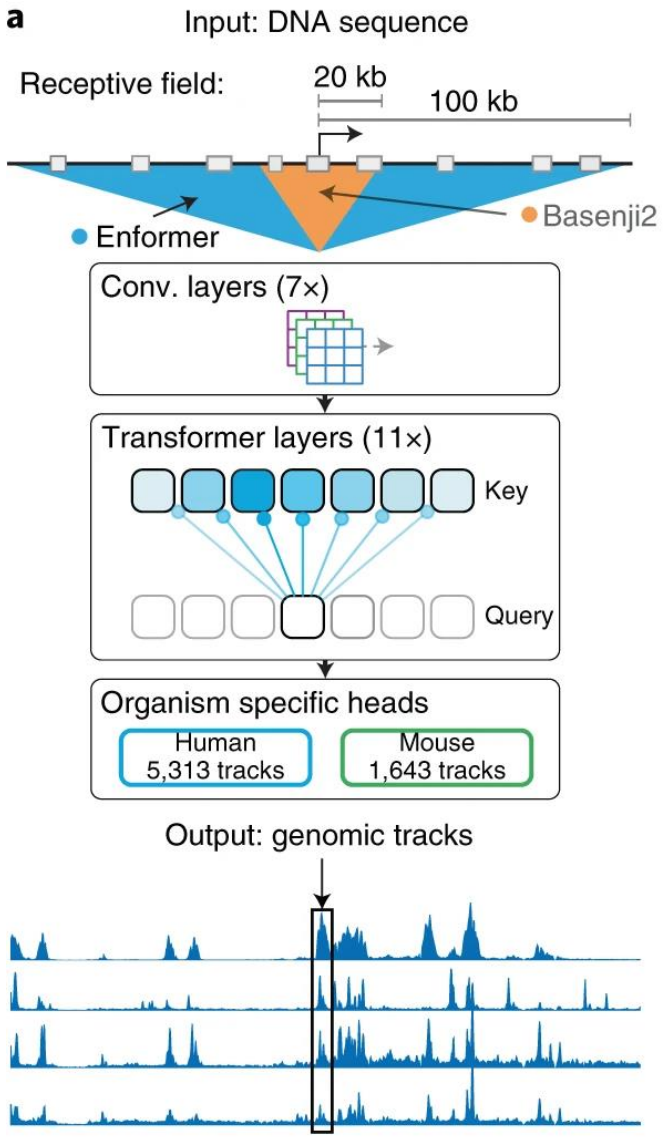


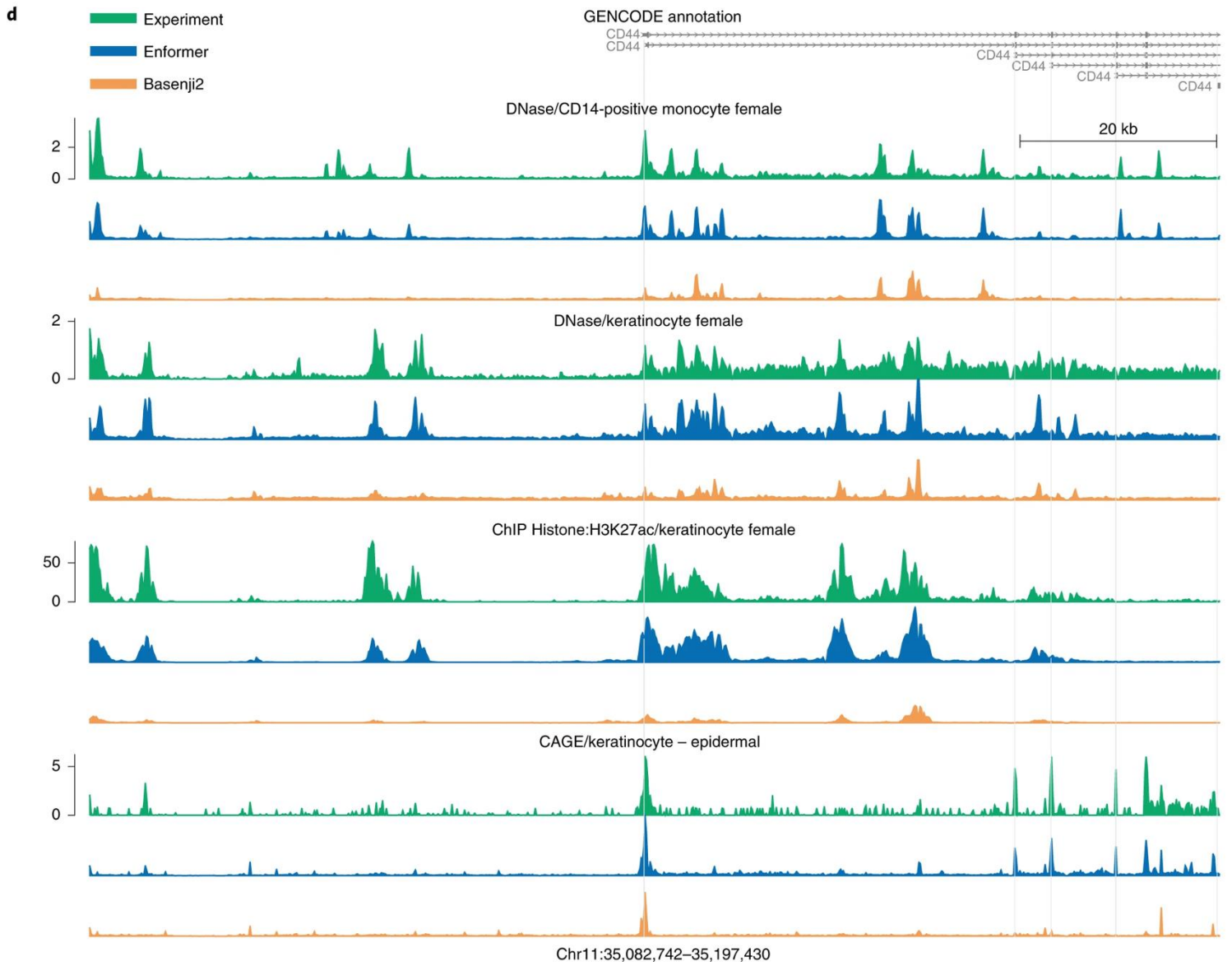
Sullivan et al. Unravelling the complex genetics of common kidney diseases: from variants to mechanisms. Nature Reviews Nephrology

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?

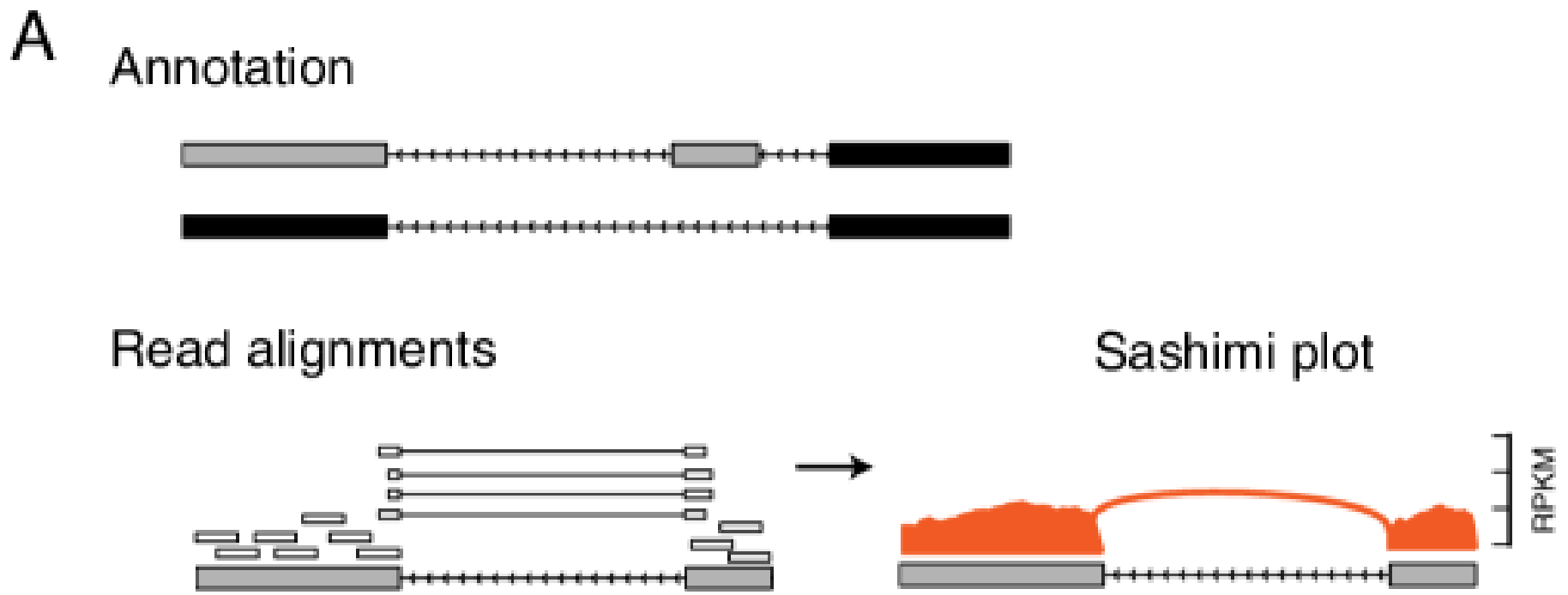
ML problem – predict assays from sequences





Avsec et al. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. Nature methods

Why aren’t you predicting RNA-Seq?



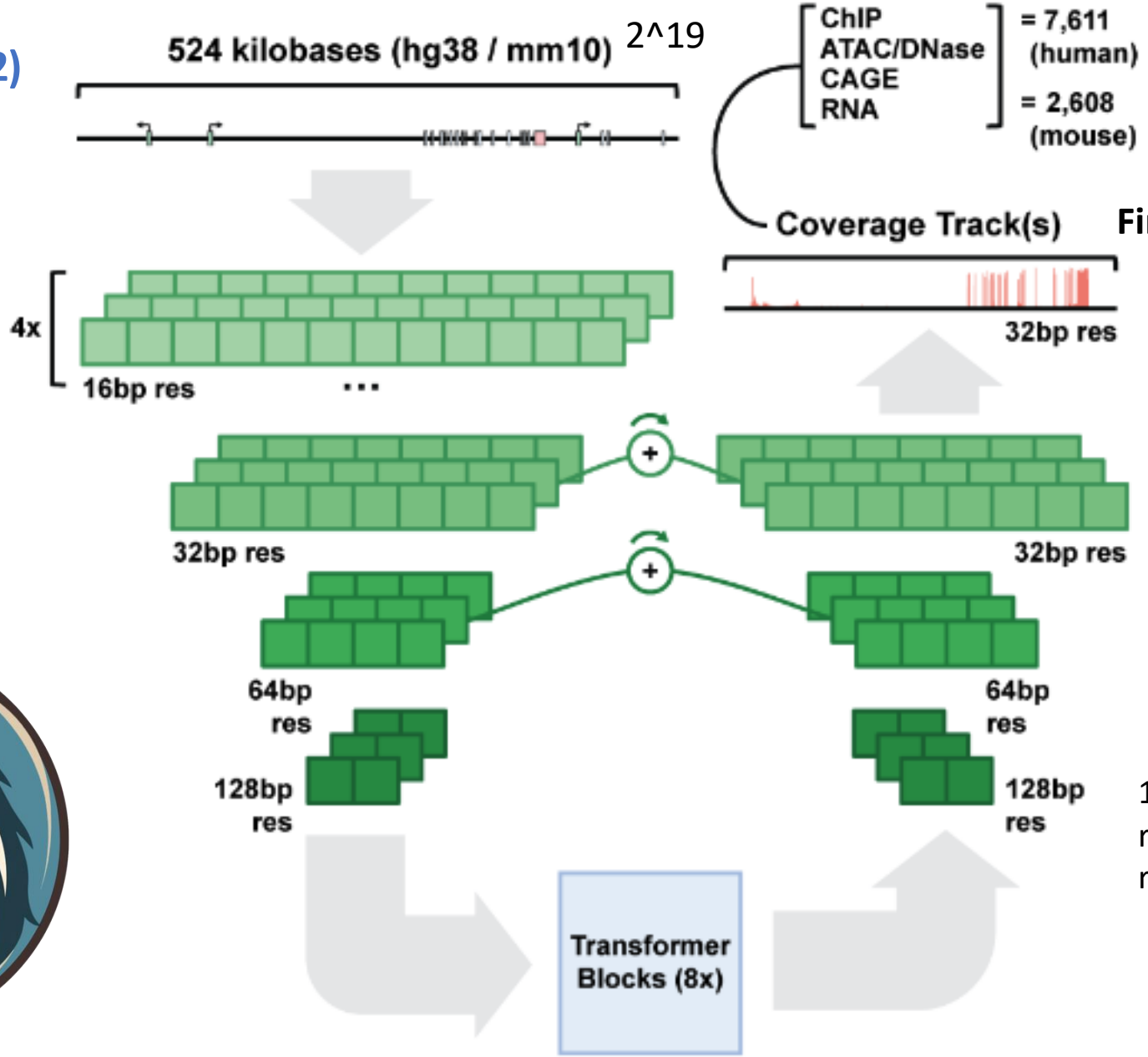
RNA-Seq potential

- **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)

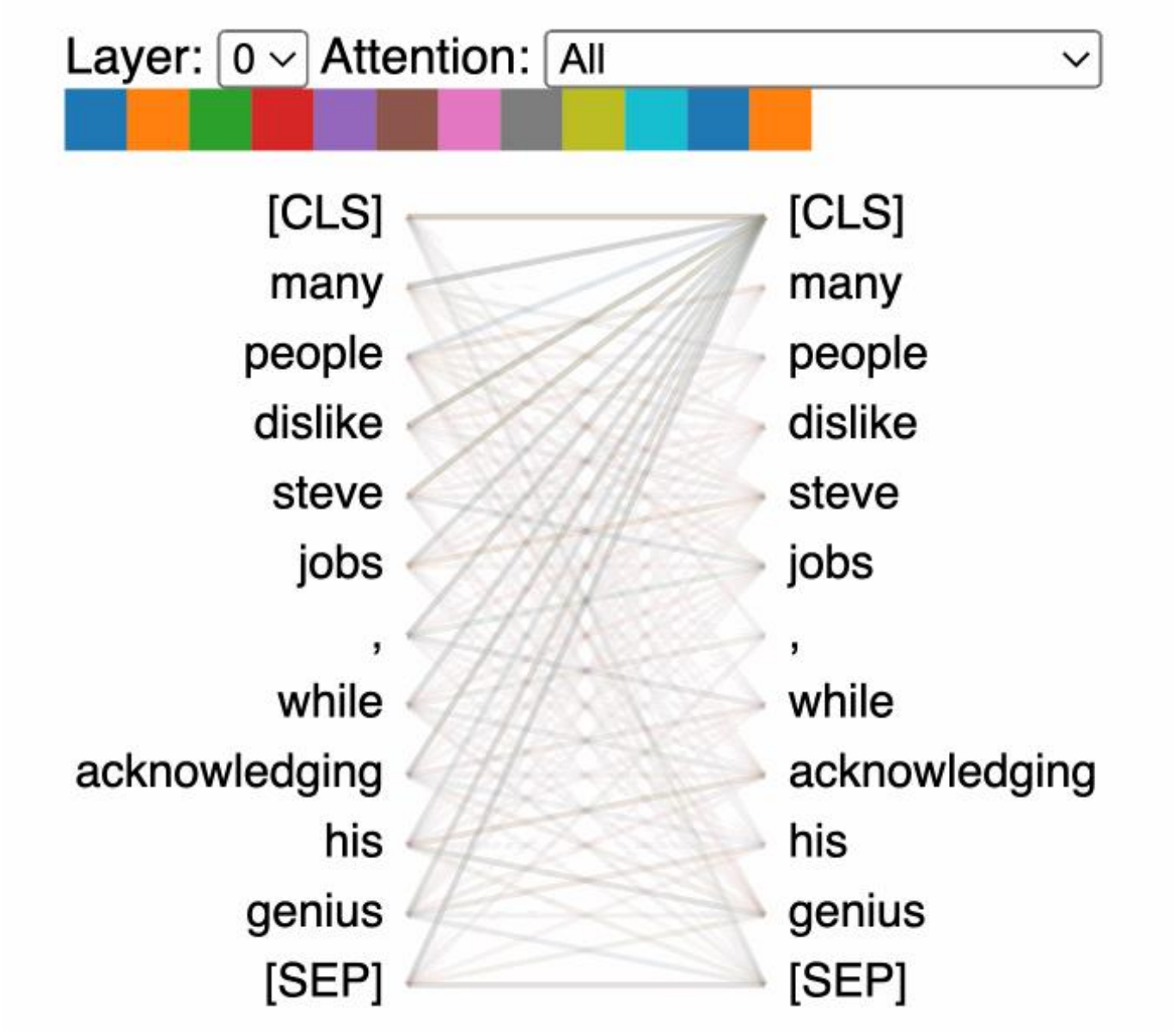
x4



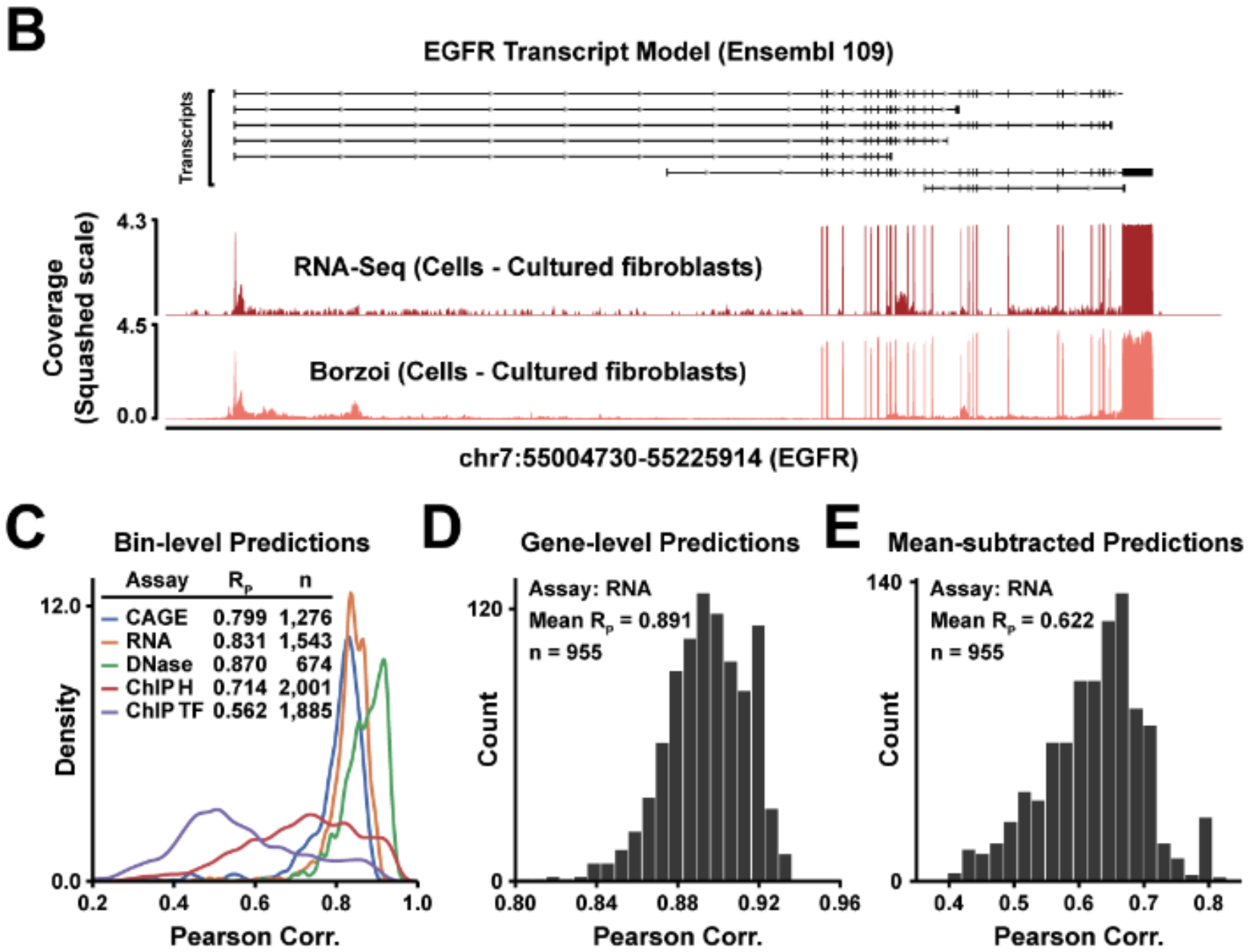
Enformer (128bp)

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

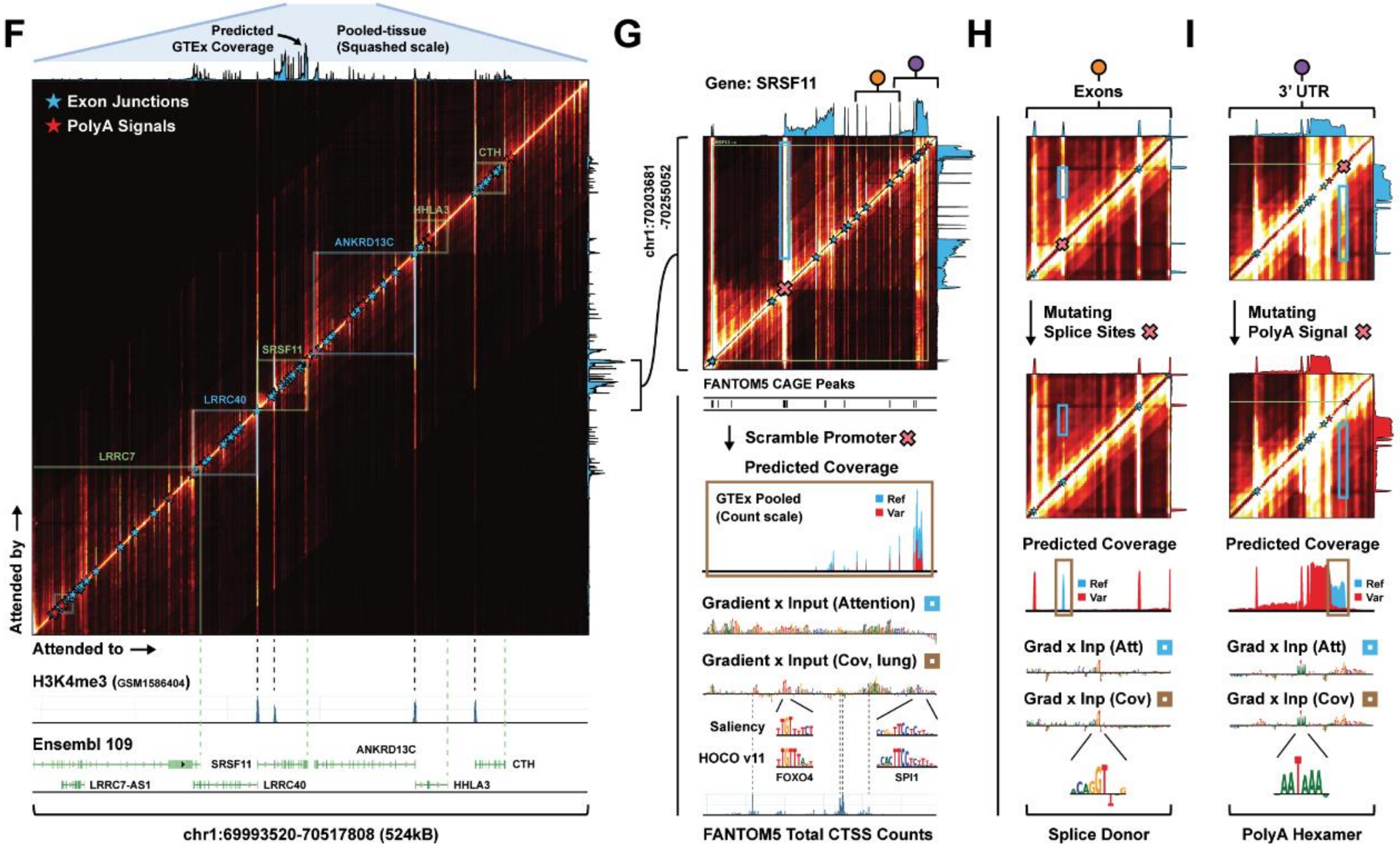
Self-attention



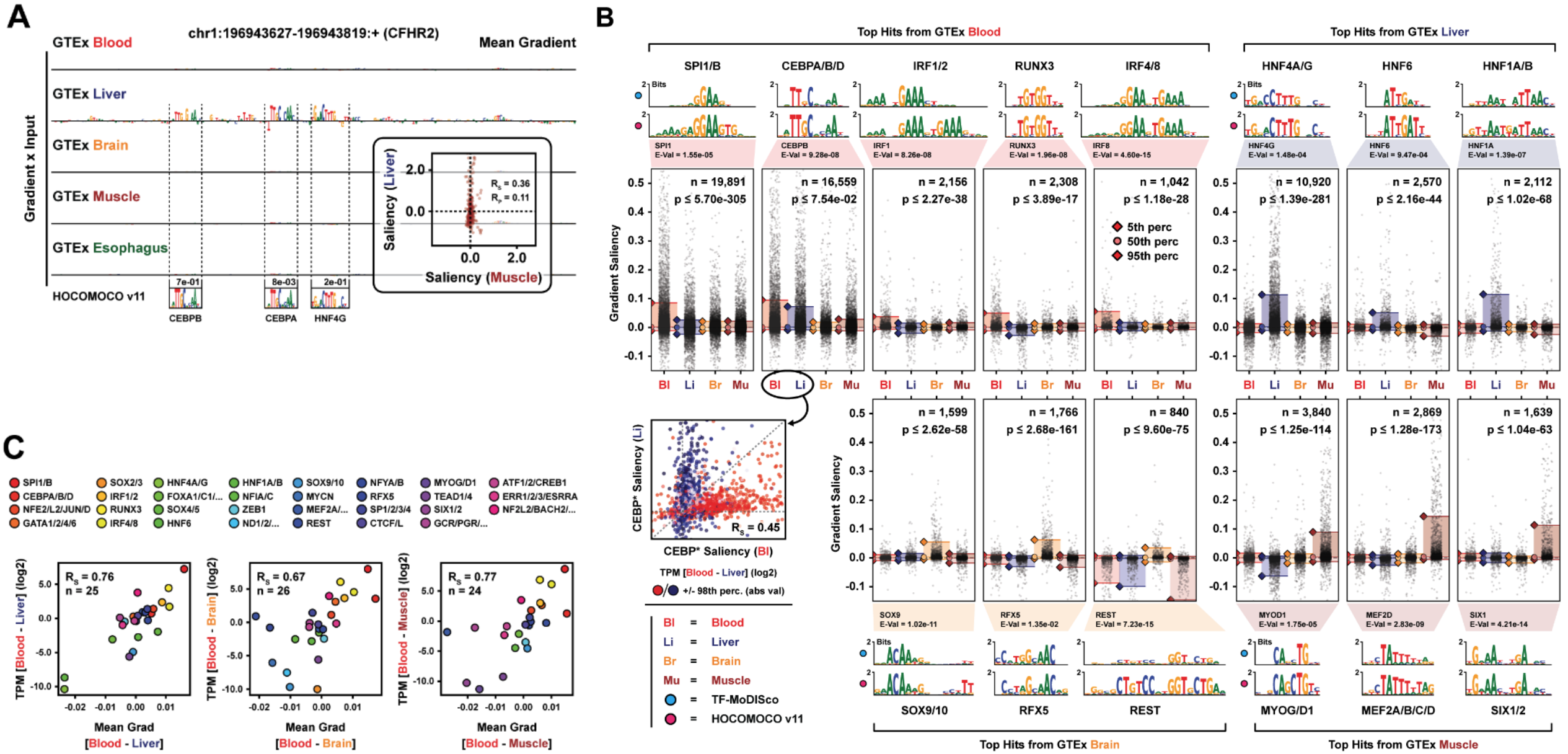
Borzoi predicts RNA-Seq well



Borzoi's attention map interpretation



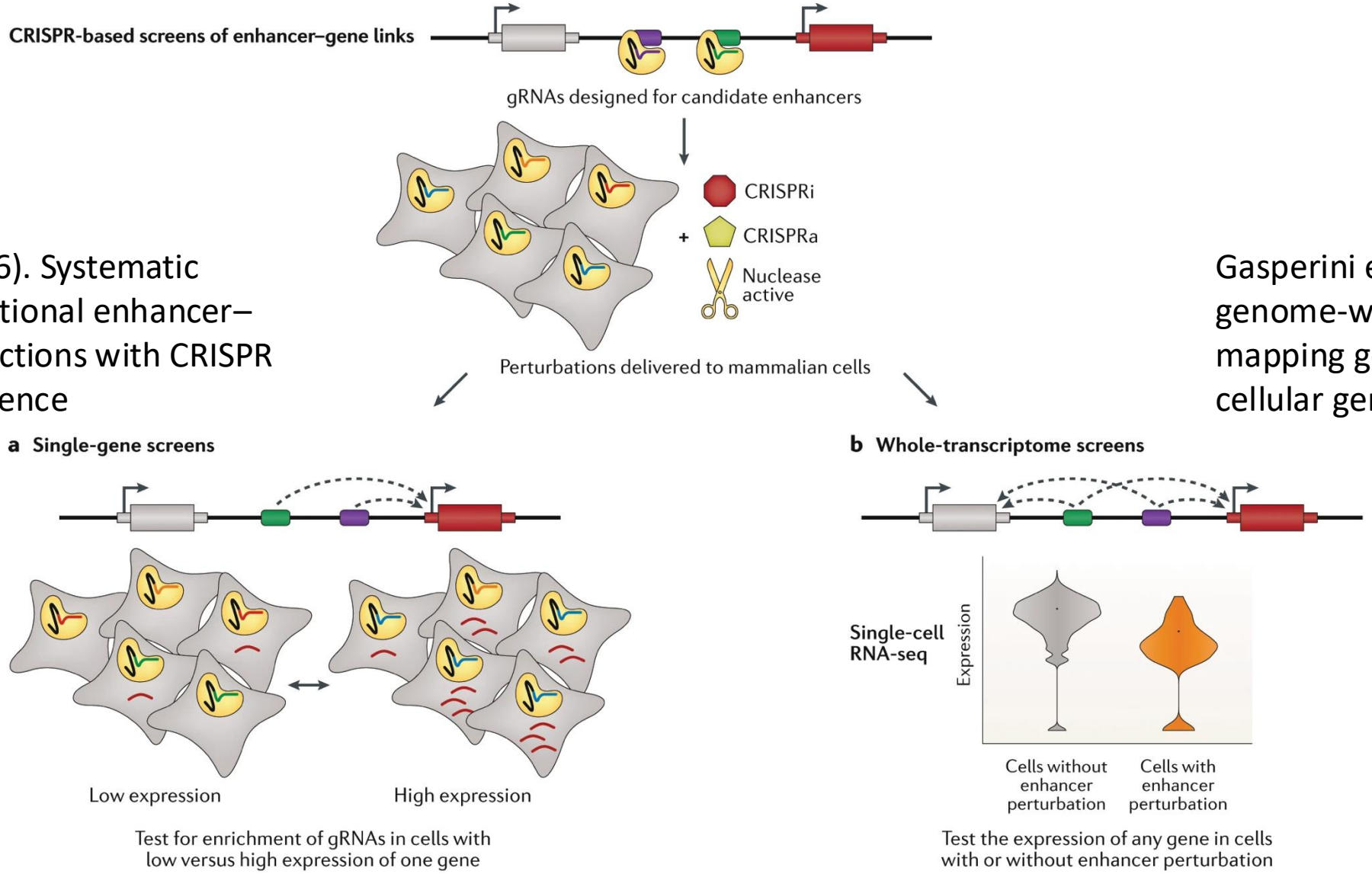
TF motif discovery from Borzoi's gradient x input



Benchmarking long range usage w/CRISPR screens

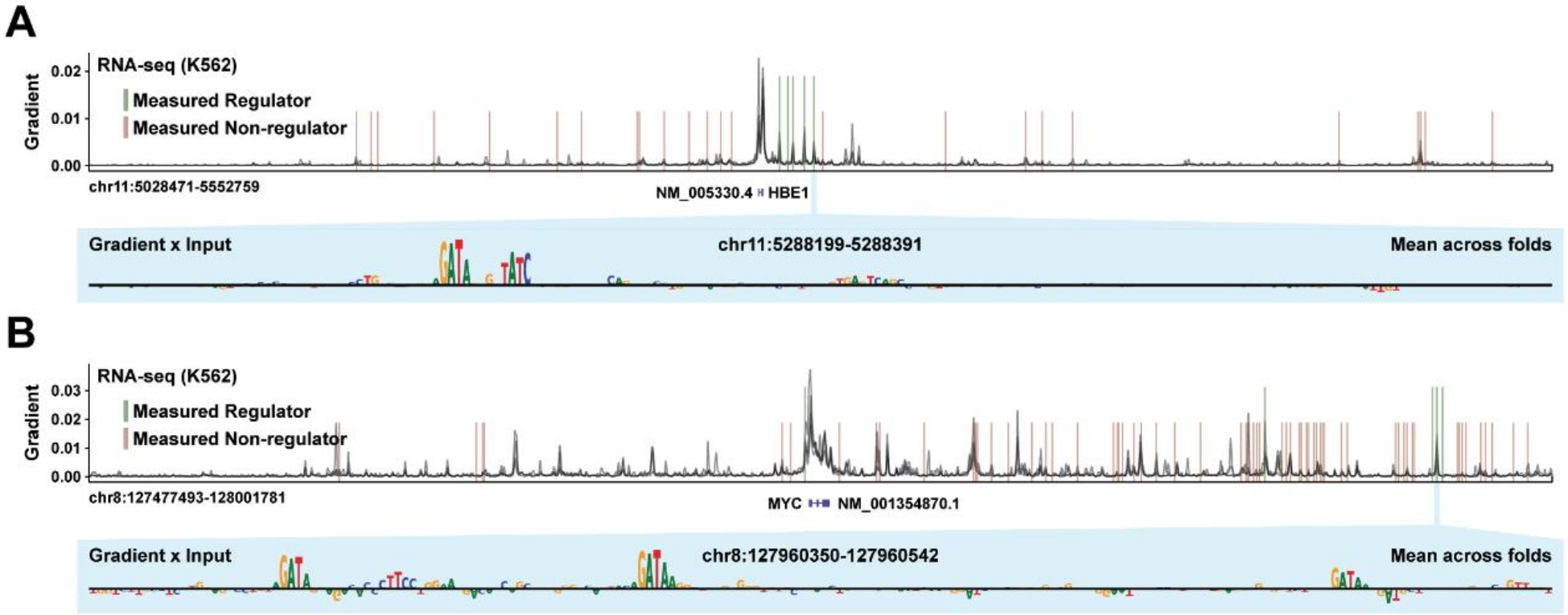
Fulco et al. (2016). Systematic mapping of functional enhancer–promoter connections with CRISPR interference. Science

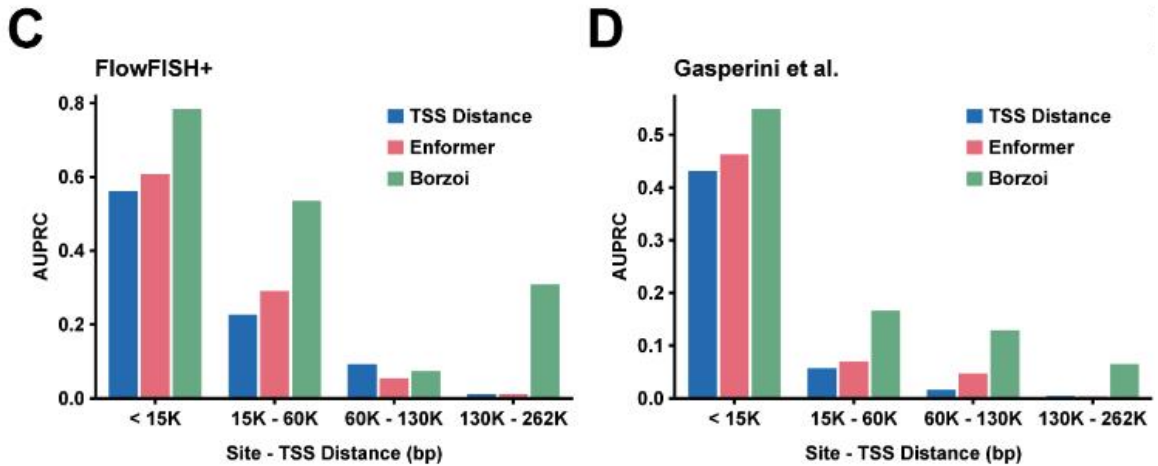
Gasperini et al. (2019). A genome-wide framework for mapping gene regulation via cellular genetic screens. Cell



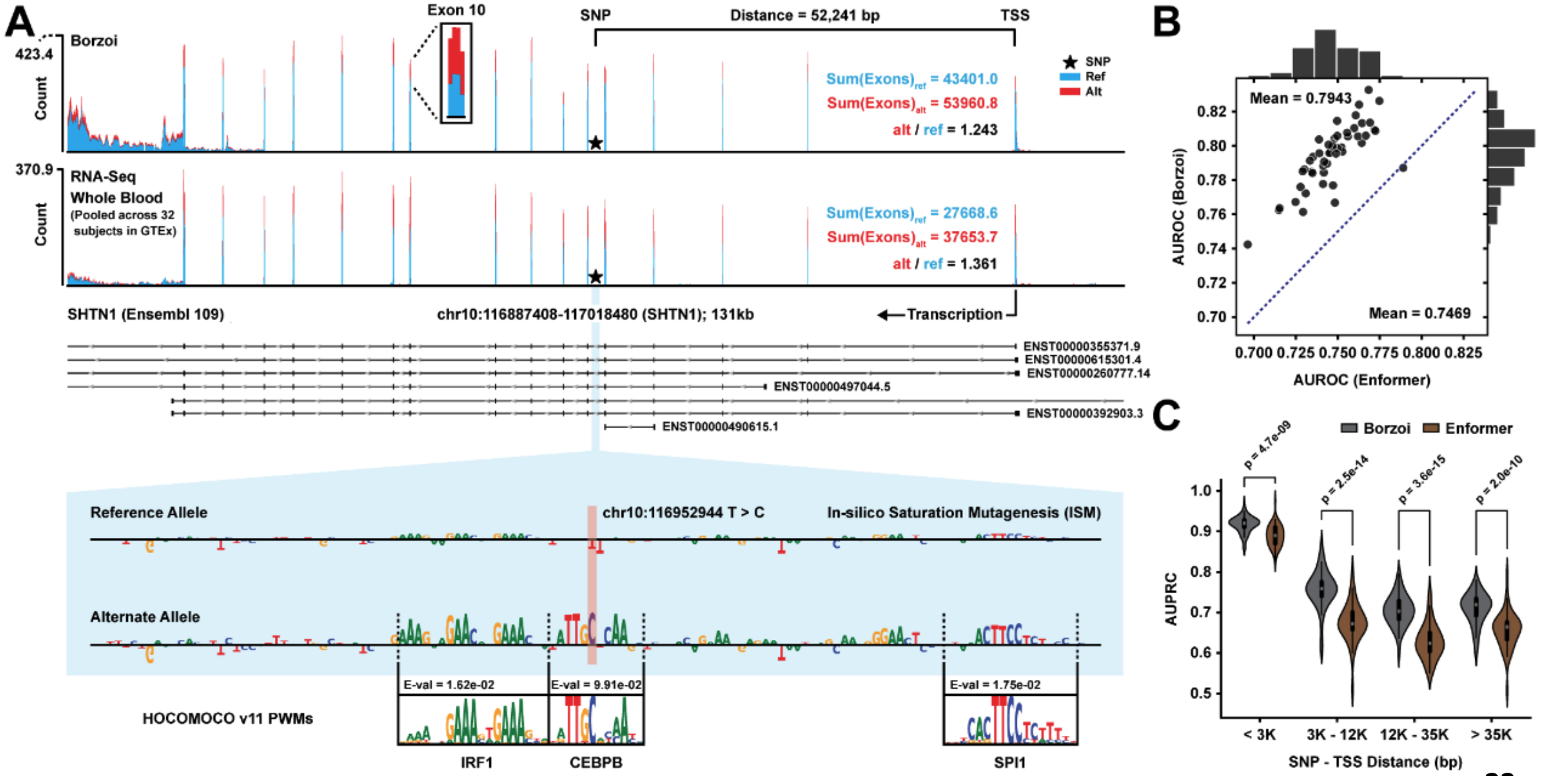
Gasperini, et al. (2020). Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nature Reviews Genetics*

CRISPR enhancer screens





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 ✉

nature genetics

Explore content ∨ About the journal ∨ Publish with us ∨

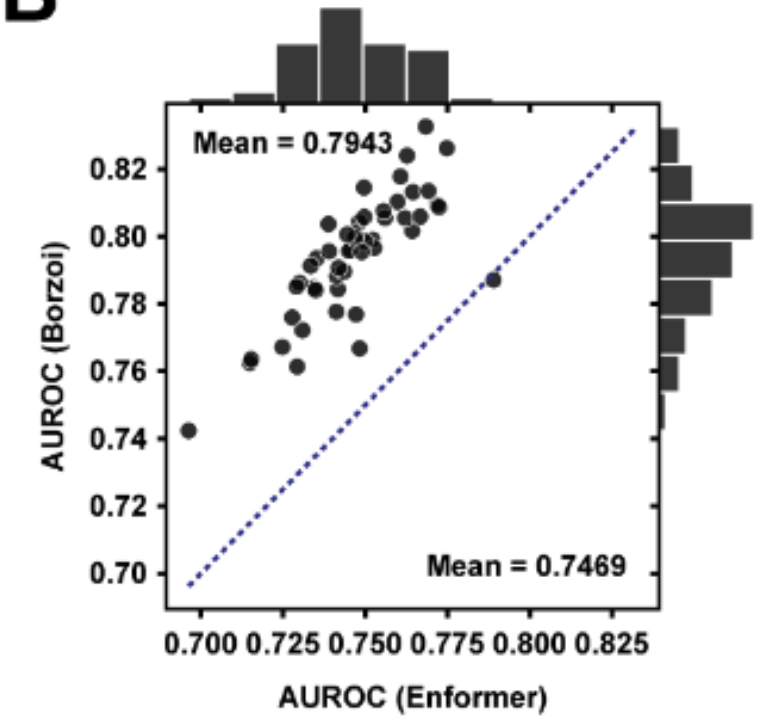
nature > nature genetics > brief communications > article

Brief Communication | Open access | Published: 30 November 2023

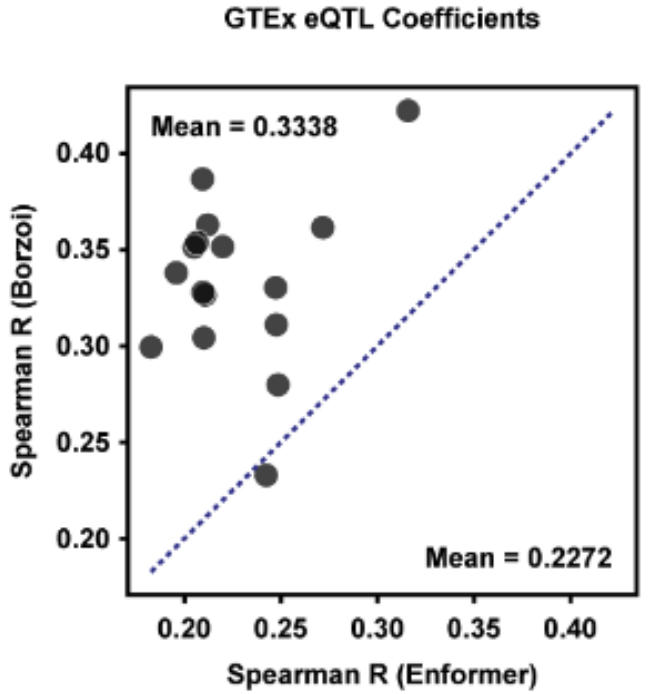
Personal transcriptome variation is poorly explained by current genomic deep learning models

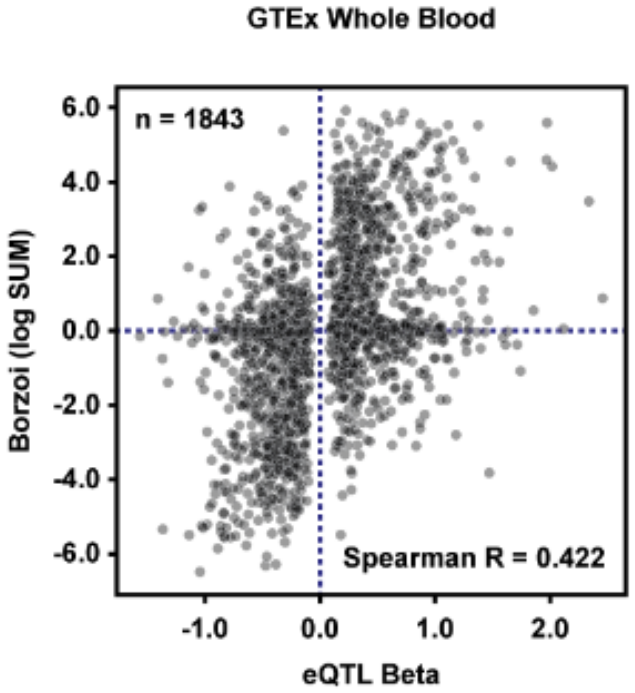
Connie Huang, Richard W. Shuai, Parth Baokar, Ryan Chung, Ruchir Rastogi, Pooja Kathail & Nilah M. Ioannidis ✉

B

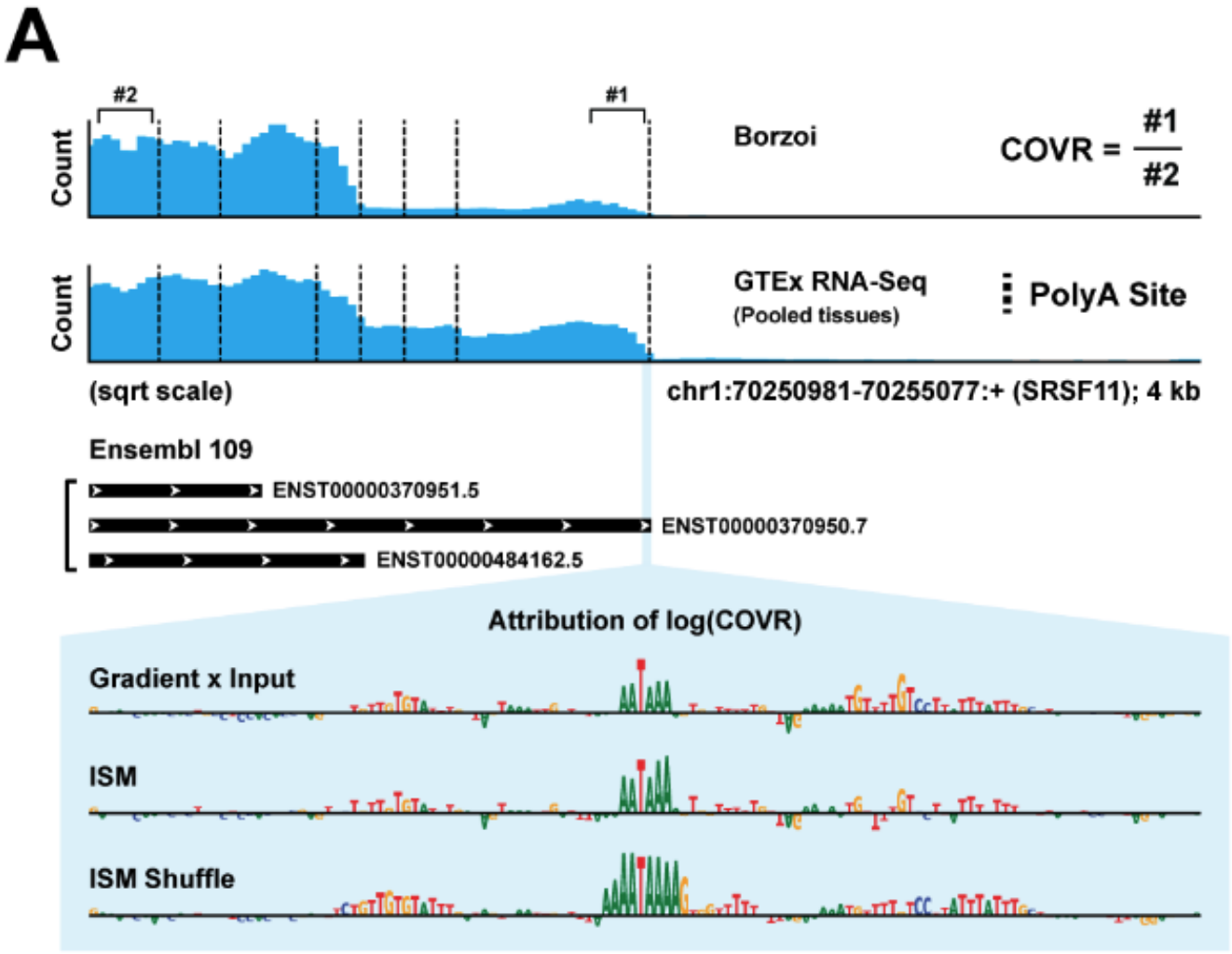


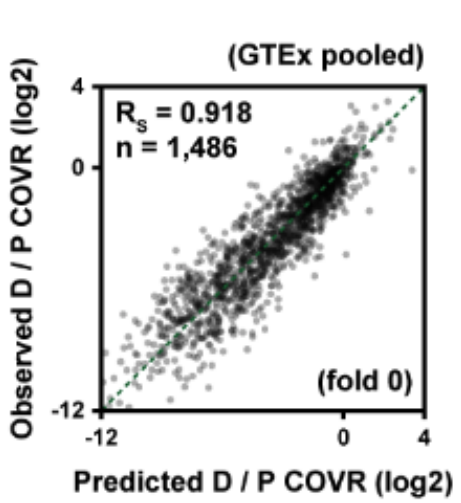
D

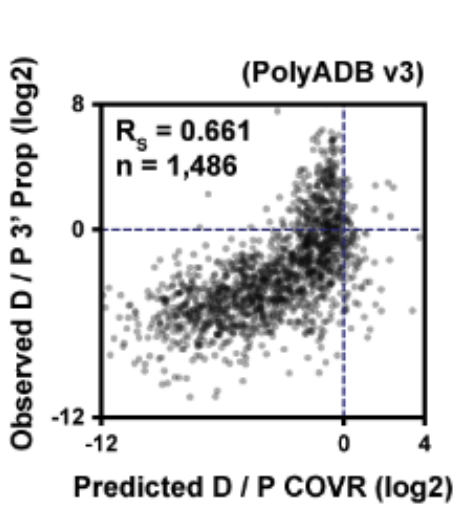


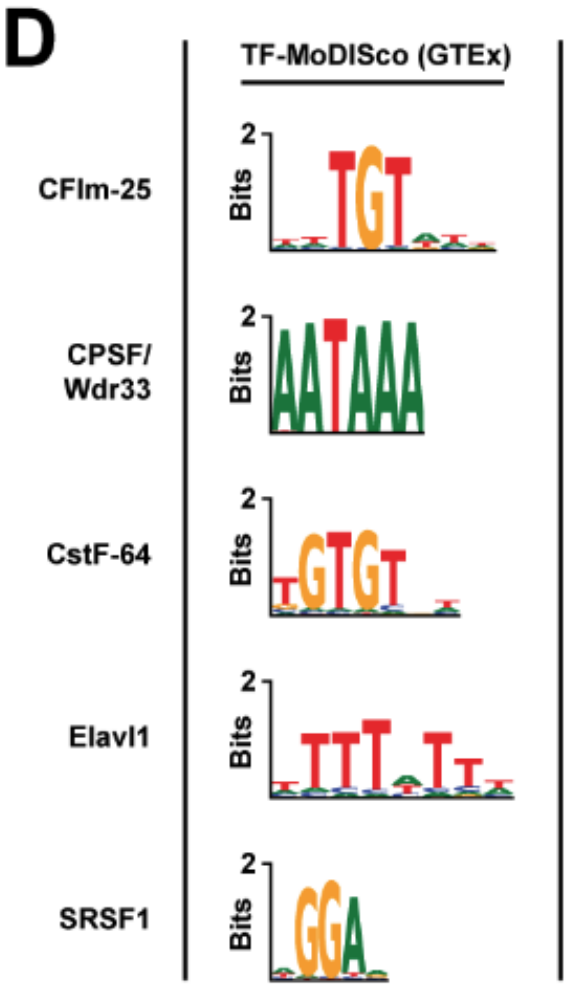


Polyadenylation



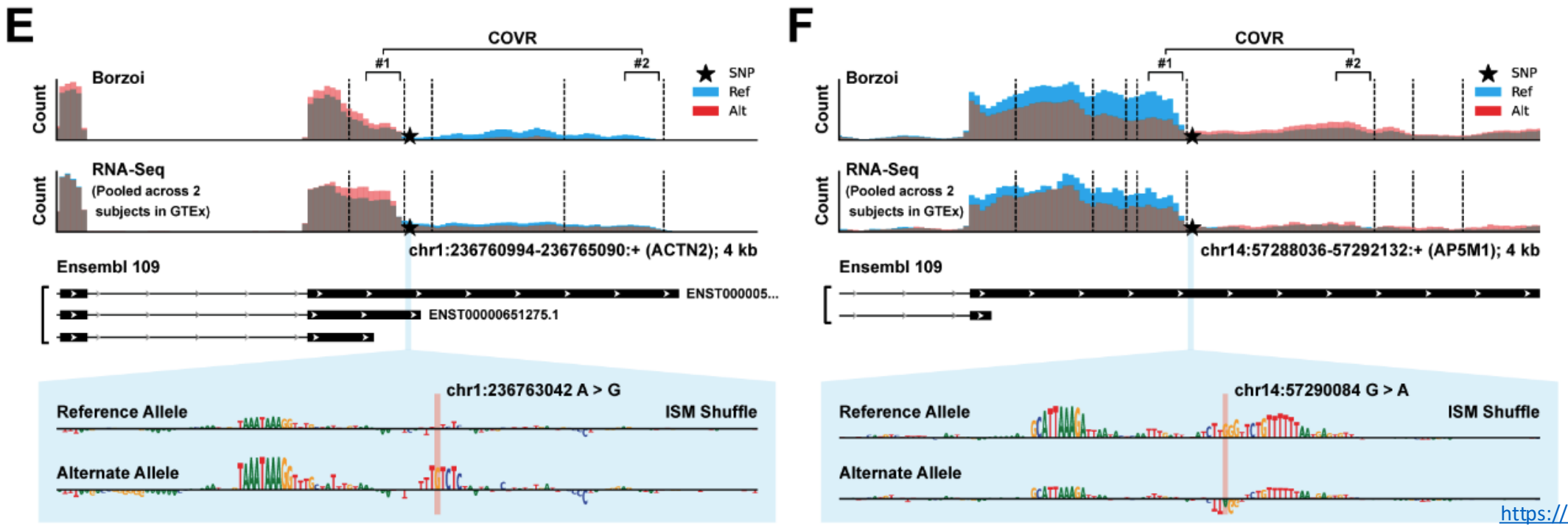






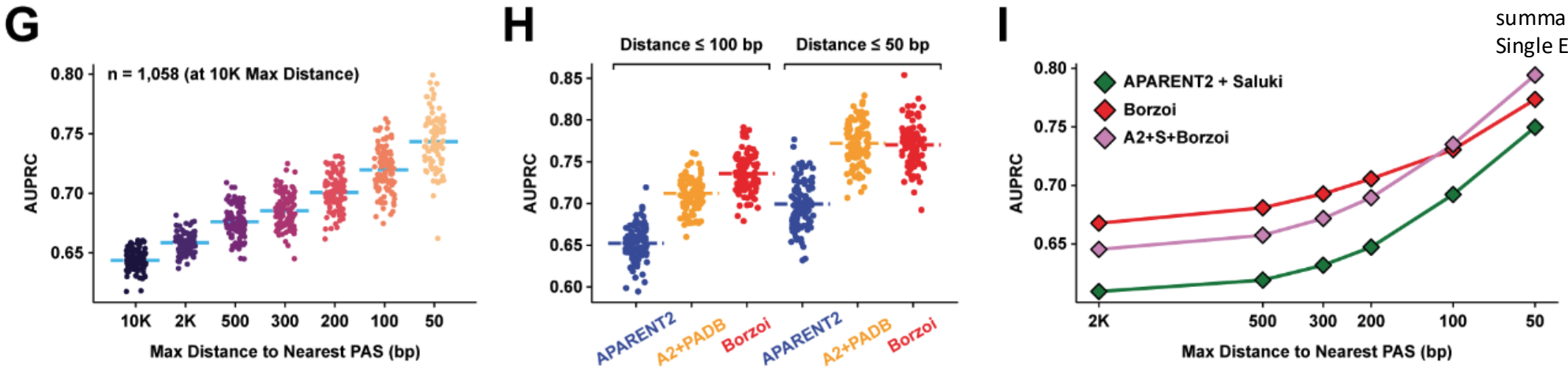
Polyadenylation QTL

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

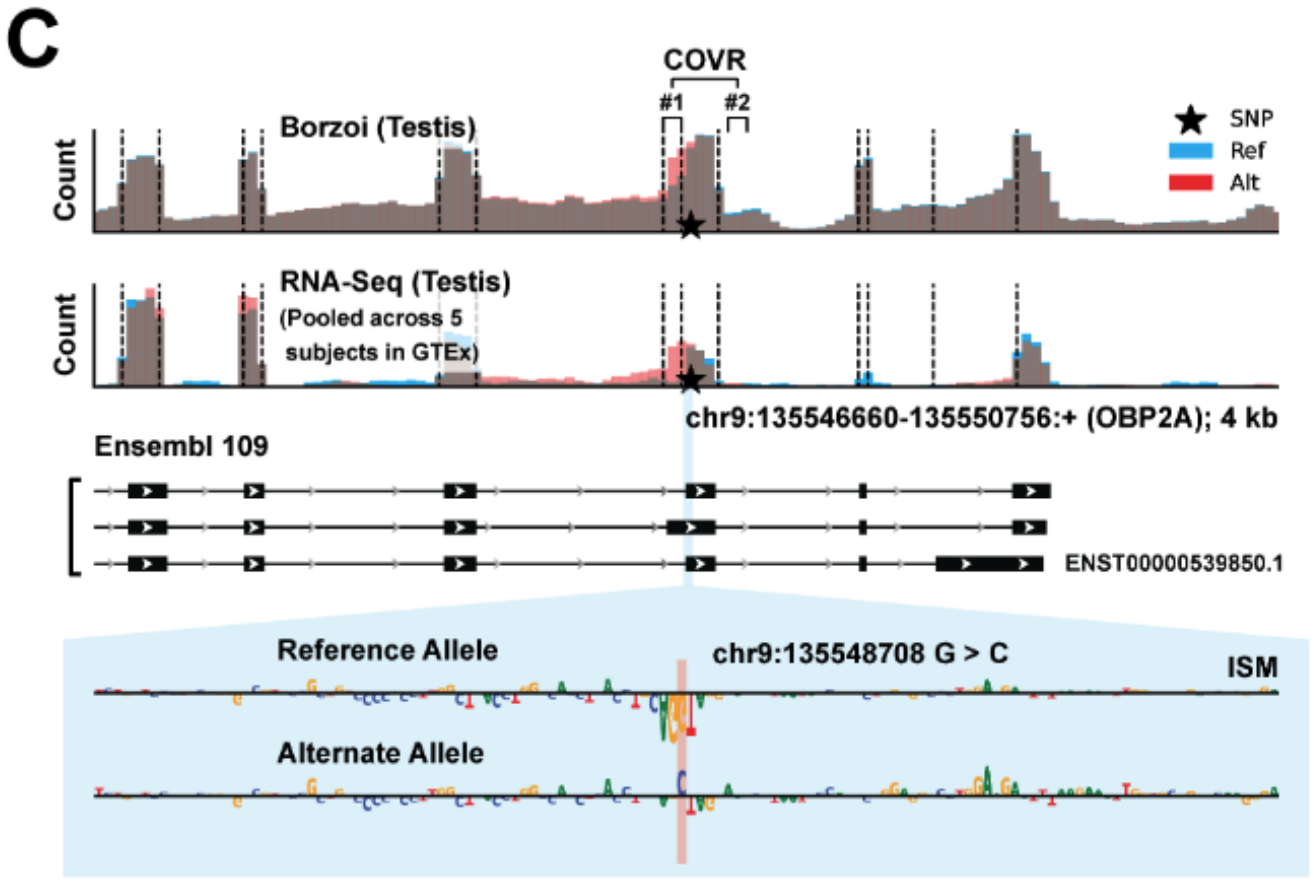


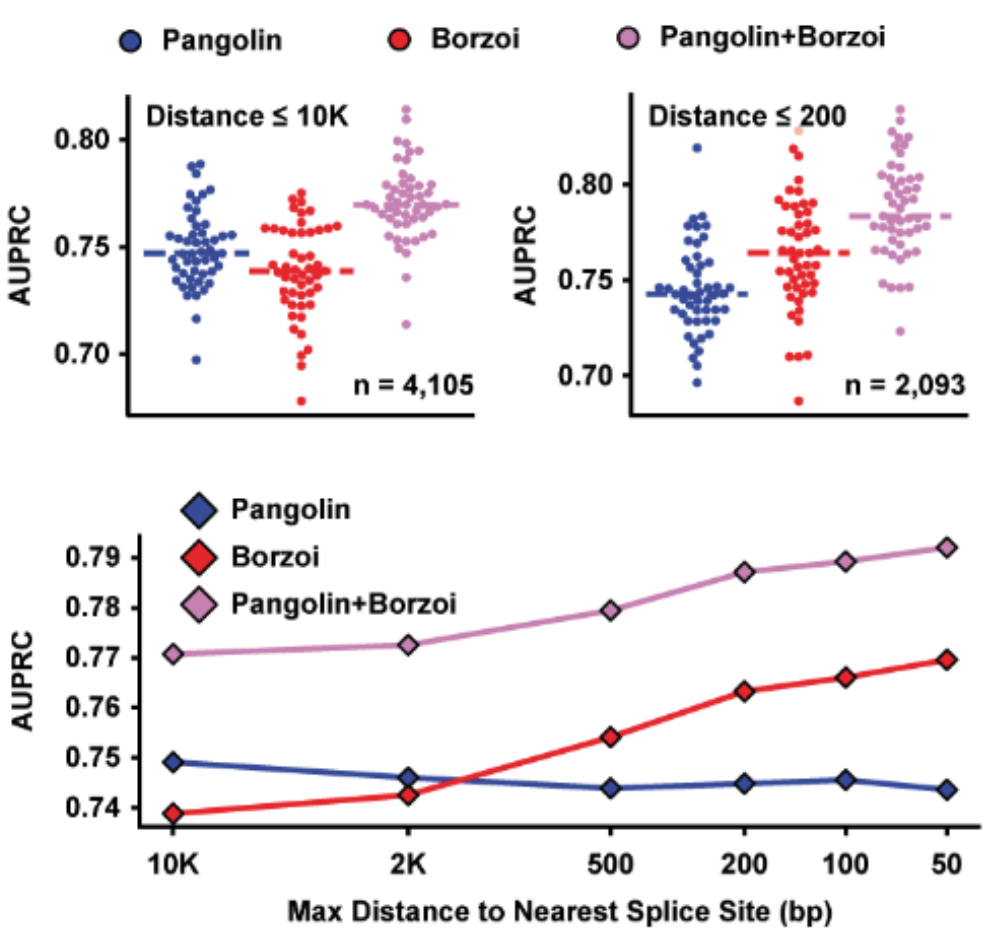
https://github.com/stephenslab/susieR

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

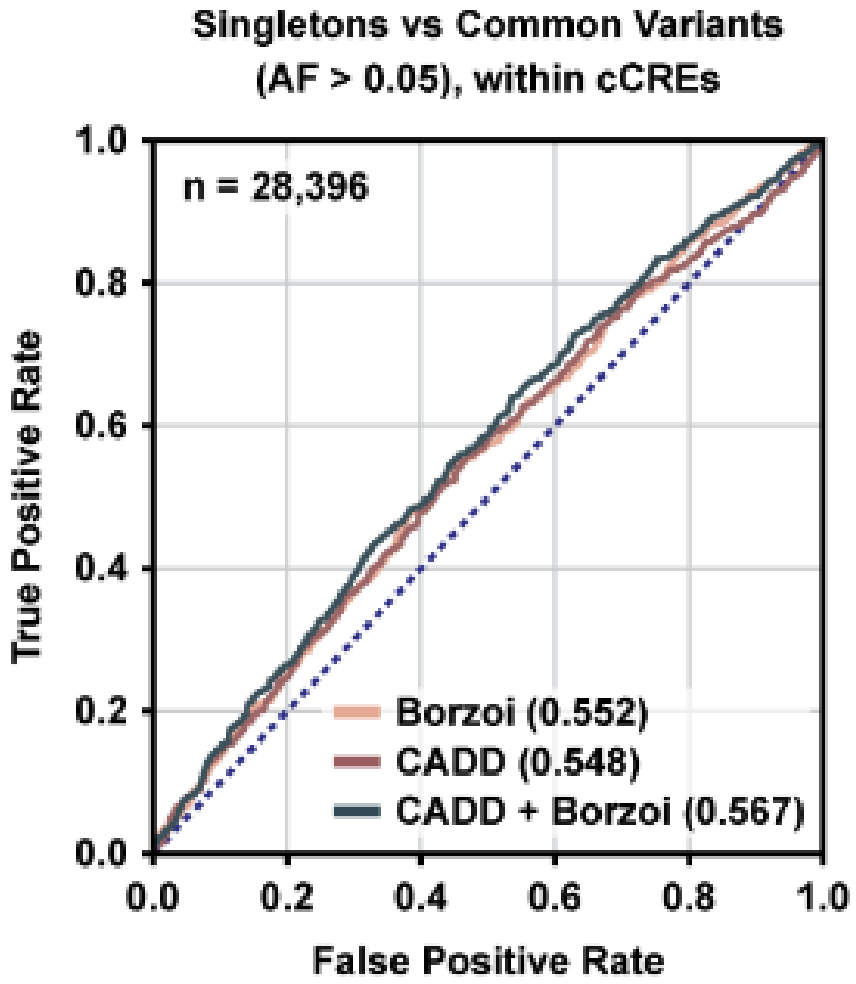


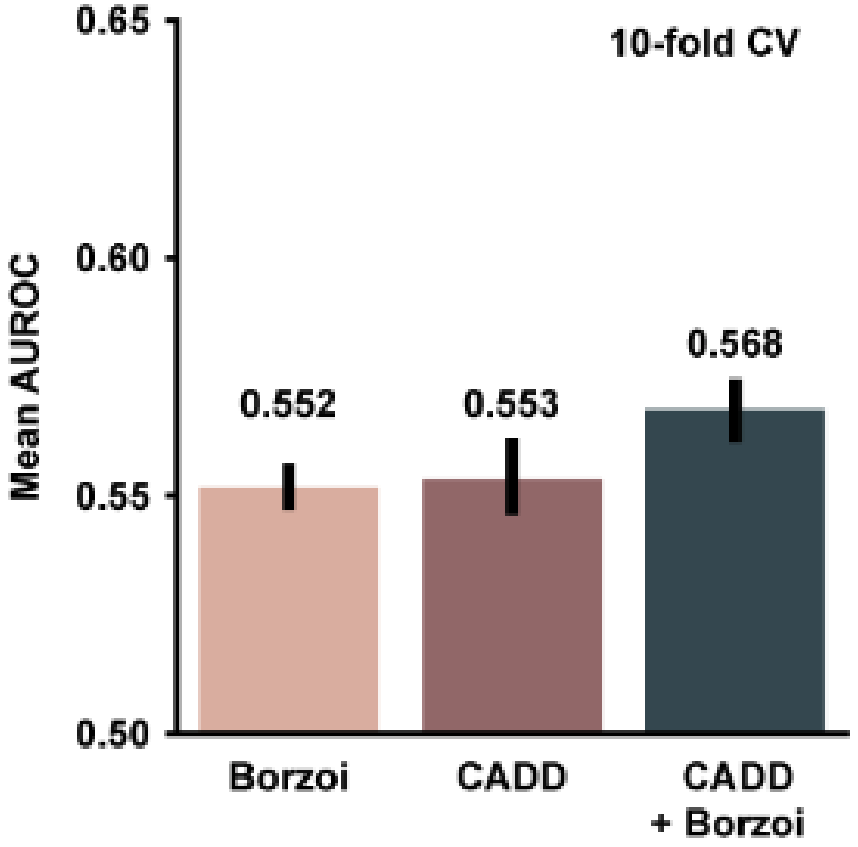
Splicing QTL





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 5th , 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 5th !