

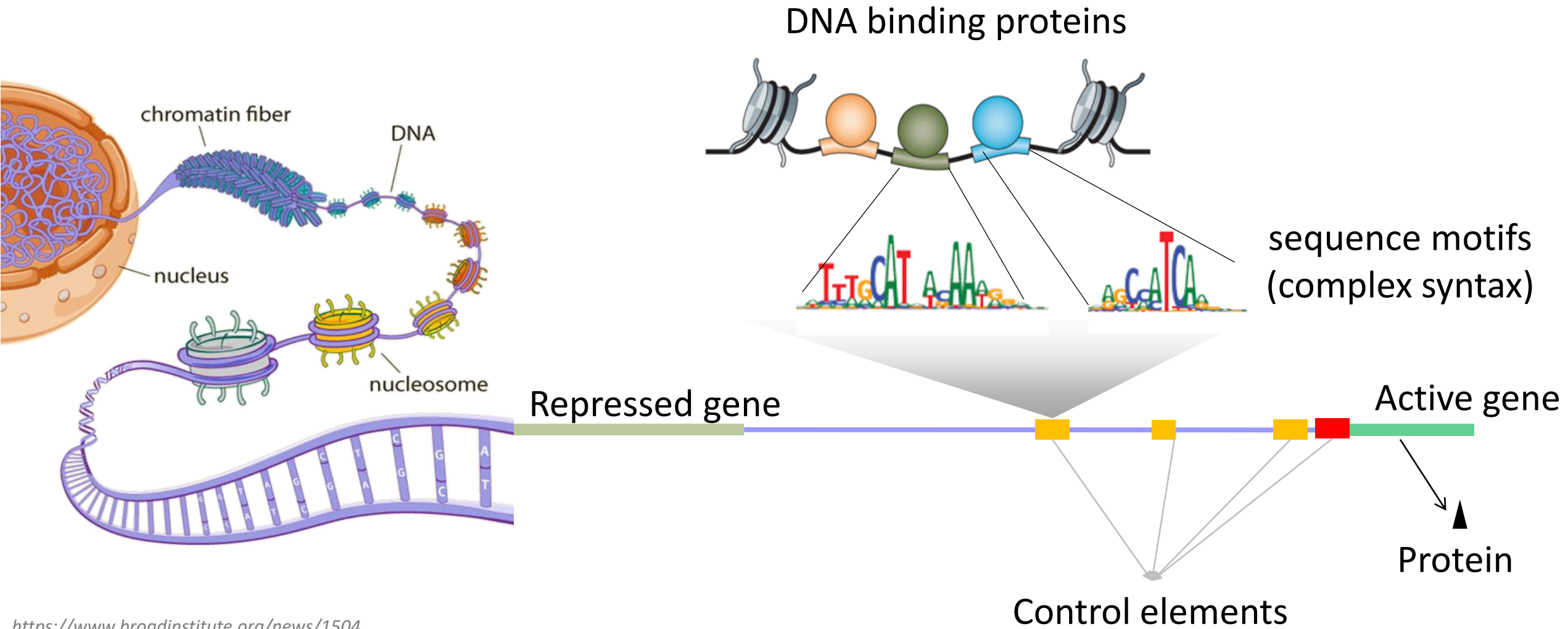
Debugging genomic profiling experiments & predictive models with interpretation tools

Anshul Kundaje

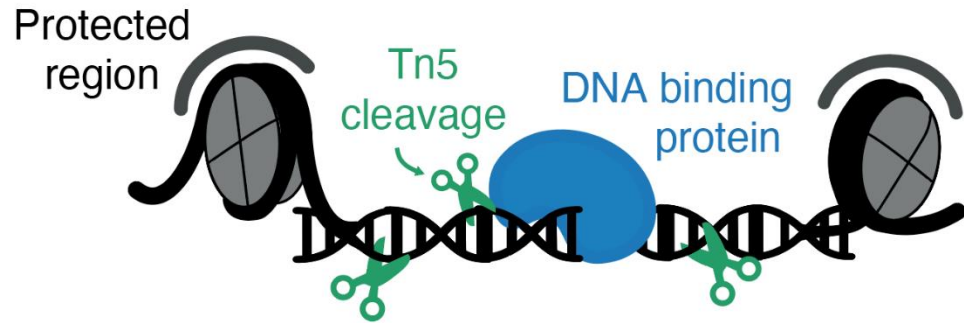
Twitter: @anshulkundaje

Github: <https://github.com/kundajelab/>

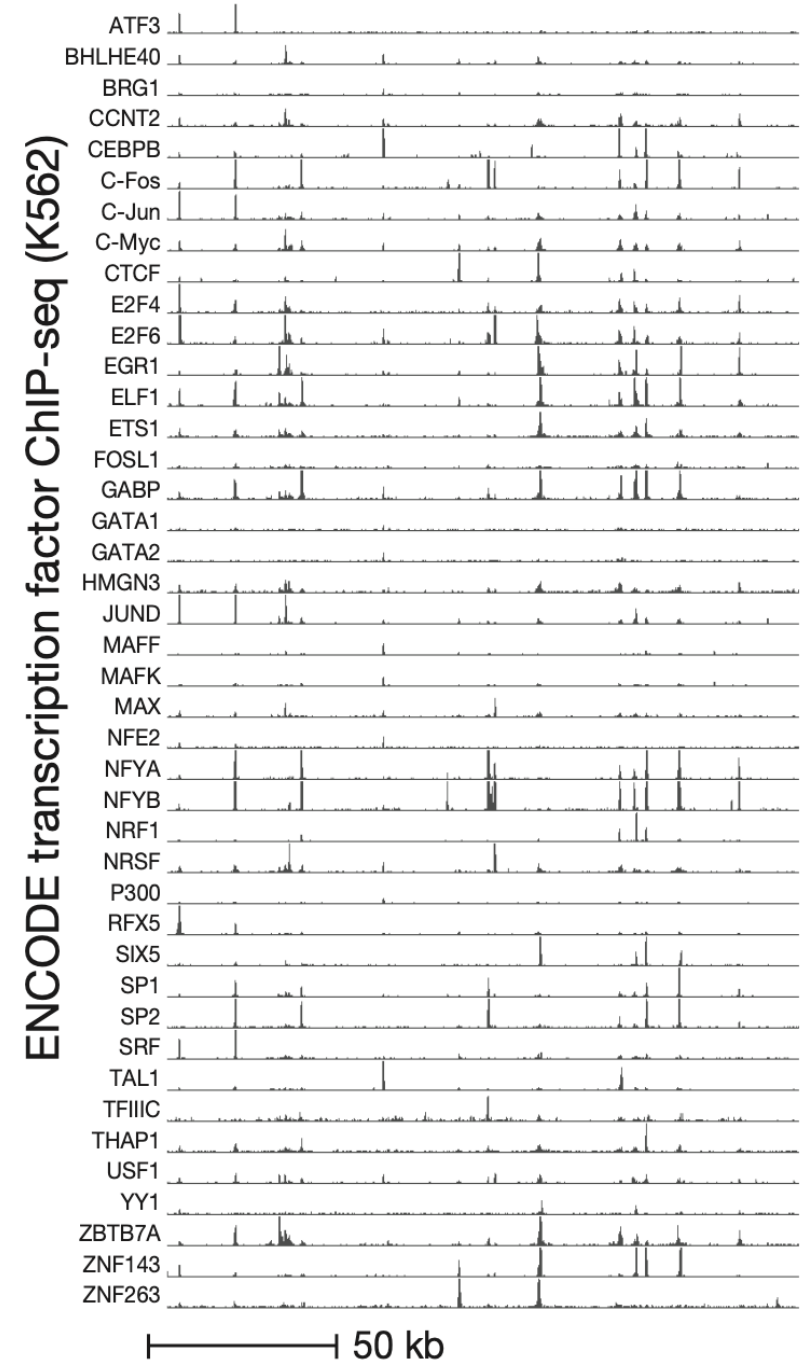
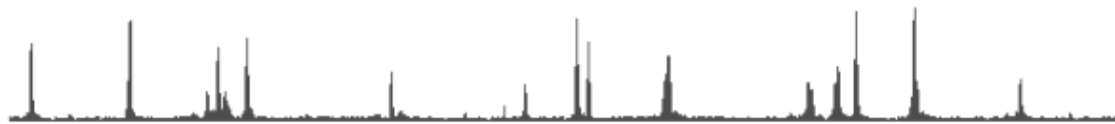
Functional components of the human genome



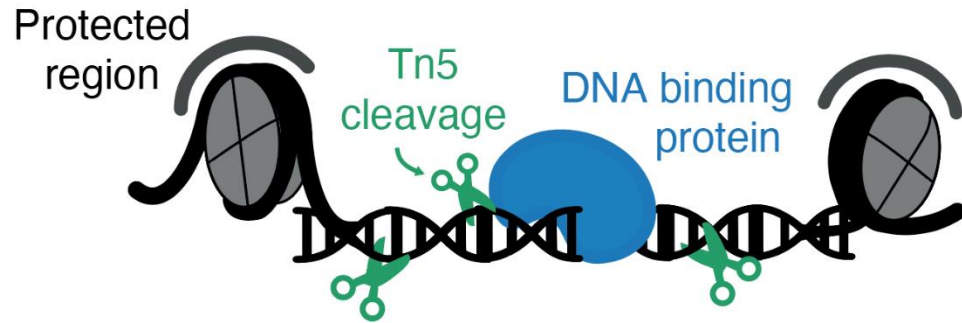
Profiling regulatory DNA



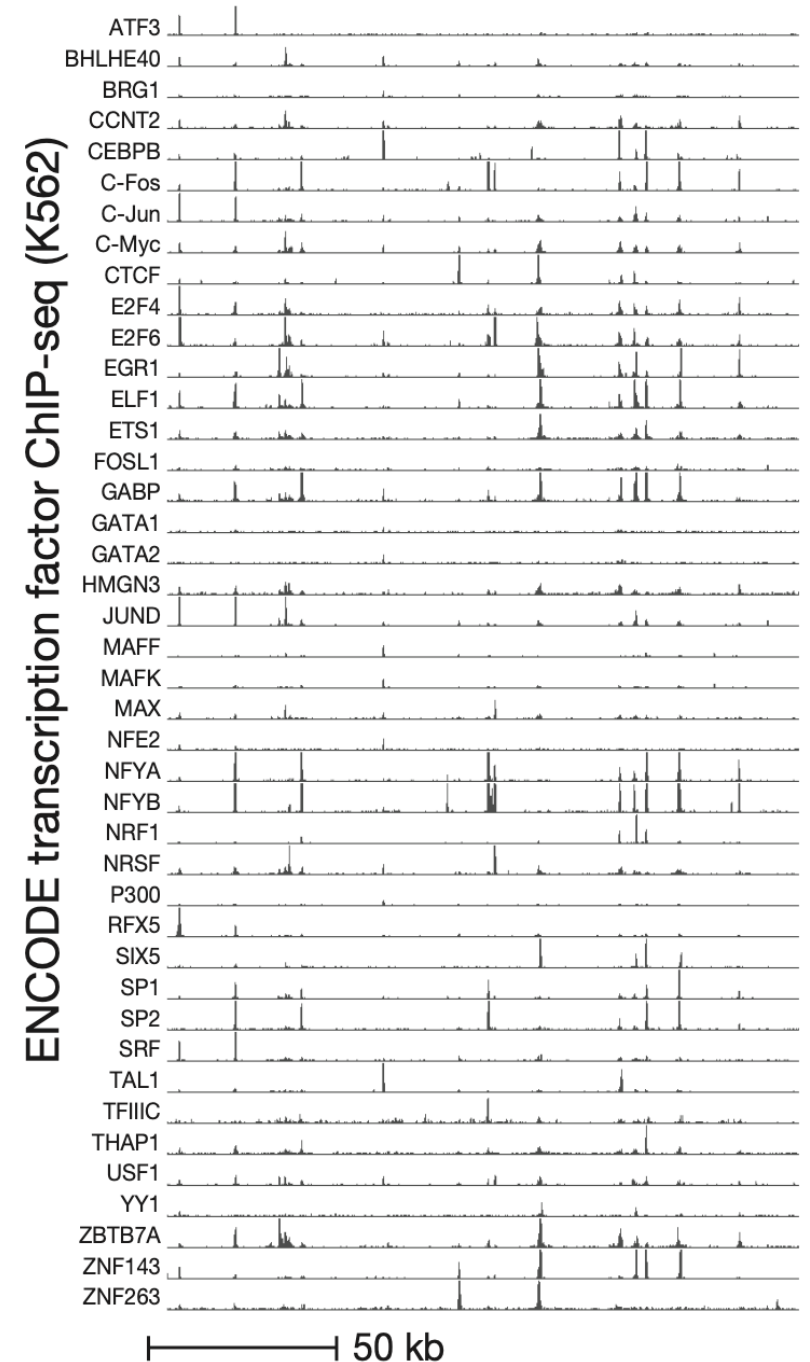
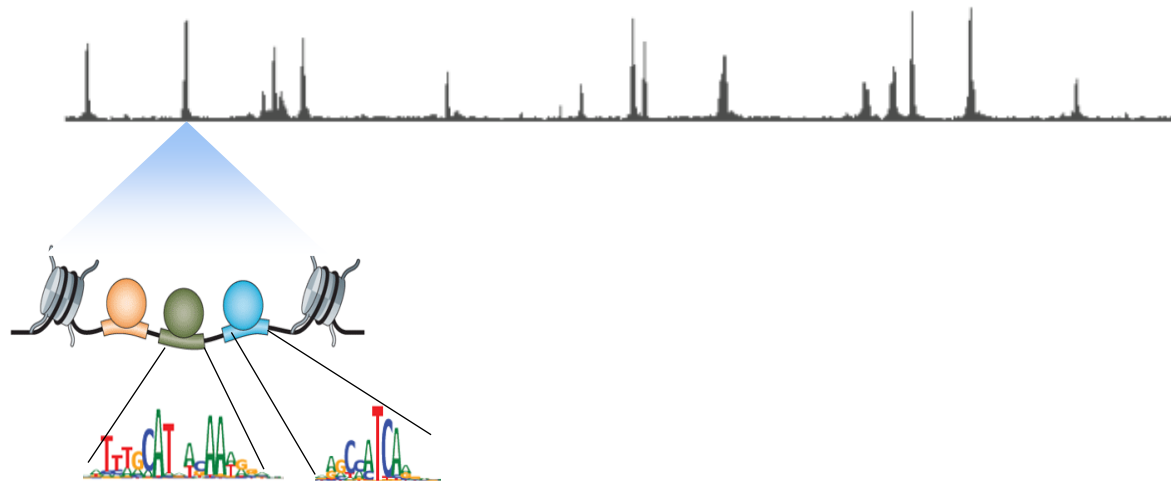
TF ChIP-seq / ATAC-seq / scATAC-seq / DNase-seq / PRO-cap



Profiling regulatory DNA

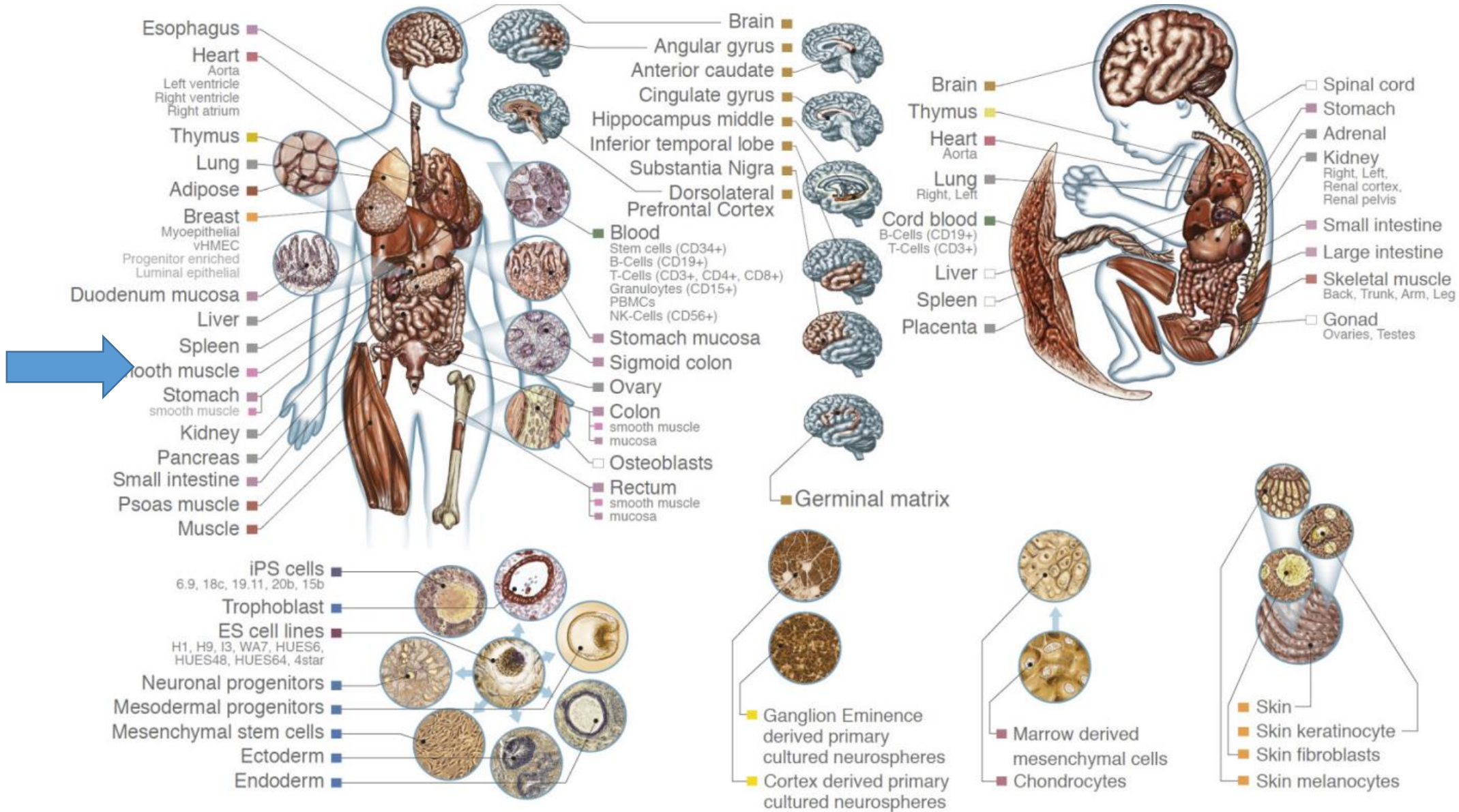


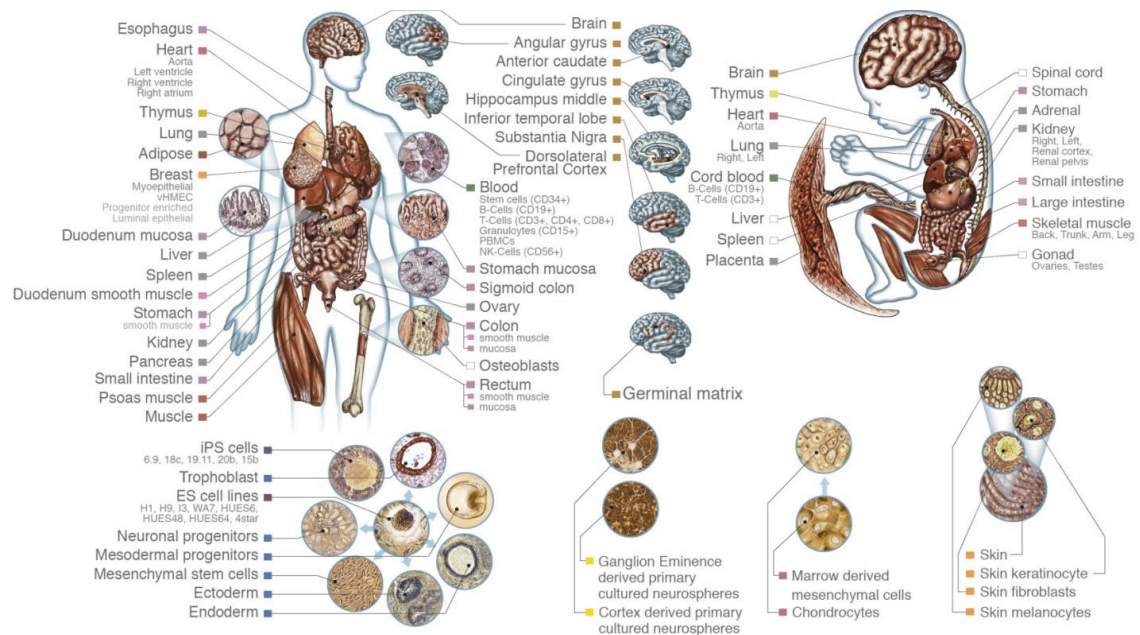
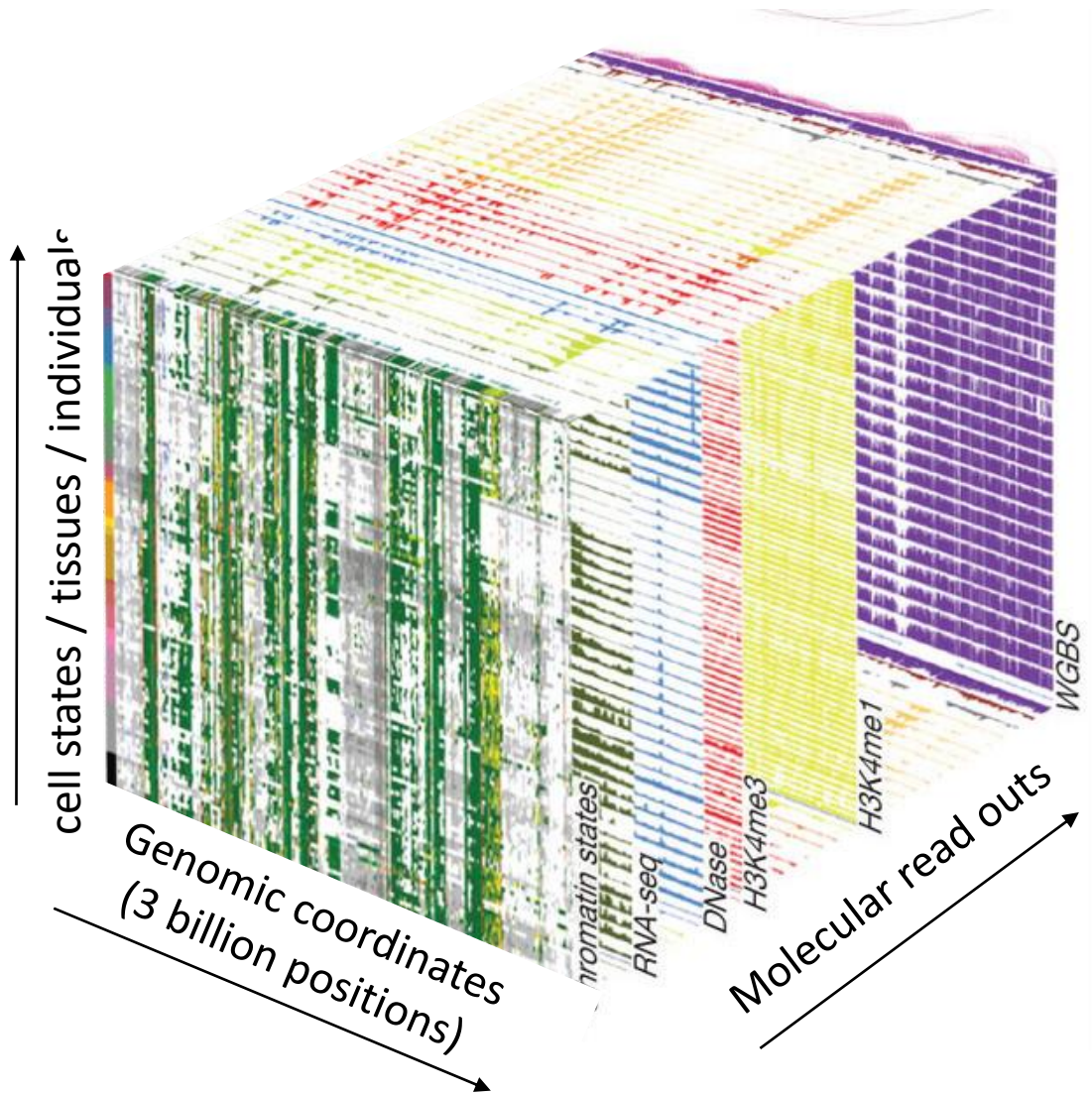
TF ChIP-seq / ATAC-seq / scATAC-seq / DNase-seq / PRO-cap



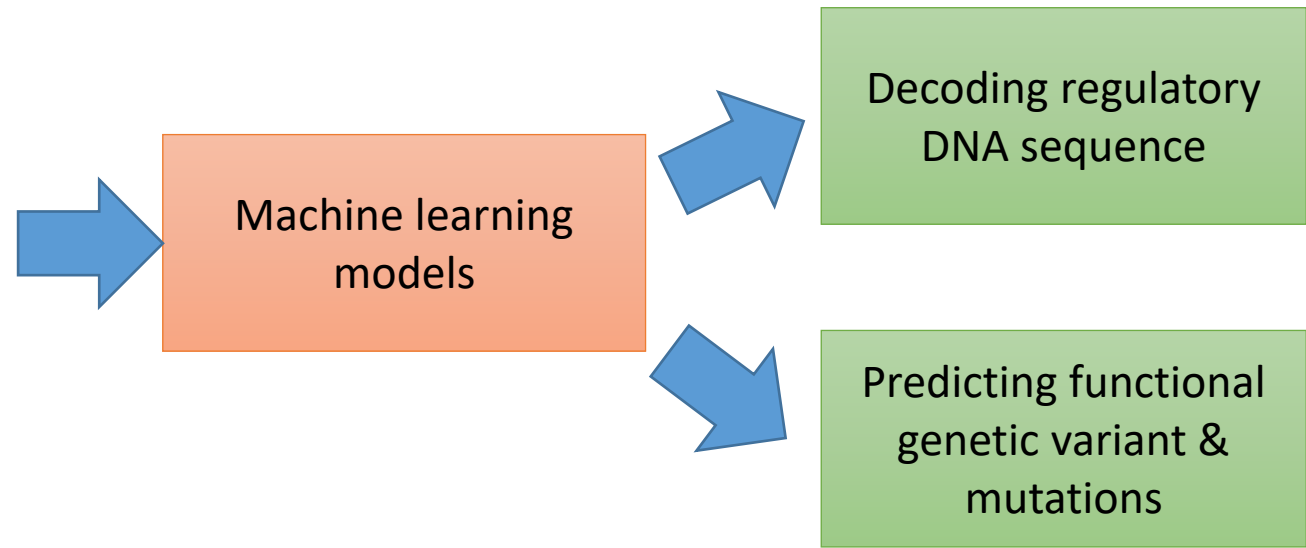
One genome ⇔ many cell types

ACCAGTTACGACGG
 TCAGGGTACTGATA
 CCCCAAACCGTTGA
 CCGCATTACAGAC
 GGGGTTTGGGTTTT
 GCCCCACACAGGTA
 CGTTAGCTACTGGT
 TTAGCAATTTACCG
 TTACAACGTTTACA
 GGGTTACGGTTGGG
 ATTTGAAAAAAGT
 TTGAGTTGGTTTTT
 TCACGGTAGAACGT
 ACCTTACAAA.....





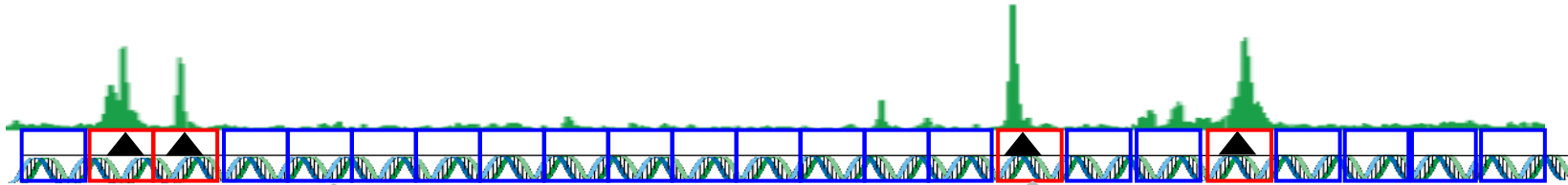
100s of Cell-Types/Tissues



Dunham, Kundaje et al. 2012 Nature
Kundaje et al. 2015 Nature

Predictive sequence models of chromatin profiles

DNase-seq / ATAC-seq data / TF or histone ChIP-seq profiles



...GACTTGAAACGGCATTG...
No Peak (0) (0.3)

...GACAGATAATGCATTGA...
Peak (+1) (20.2)

...GACAGATAATGCATTGA...

...ACTGTCATGGATAATTCT...

...GATAATTCTACTGTAAG...

DNA sequences (S_i)

...CAACCTTGAACGGCATTG...

...GACTTGAAACGGCATTG...

...CAGTATGCATACGTGAA...

Classification
or Regression
model
 $F(S_i)$

Arvey et al. 2012
Ghandi et al. 2014
Setty et al. 2015
Alipanahi et al. 2015
Zhou et al. 2015
Kelly et al. 2016, 2018
Avsec et al. 2021

Class = +1 (20.2)

Class = +1 (10.6)

Class = +1 (15.8)

Measured
Labels (Y_i)

Class = 0 (0.3)

Class = 0 (1.2)

Class = 0 (3.5)



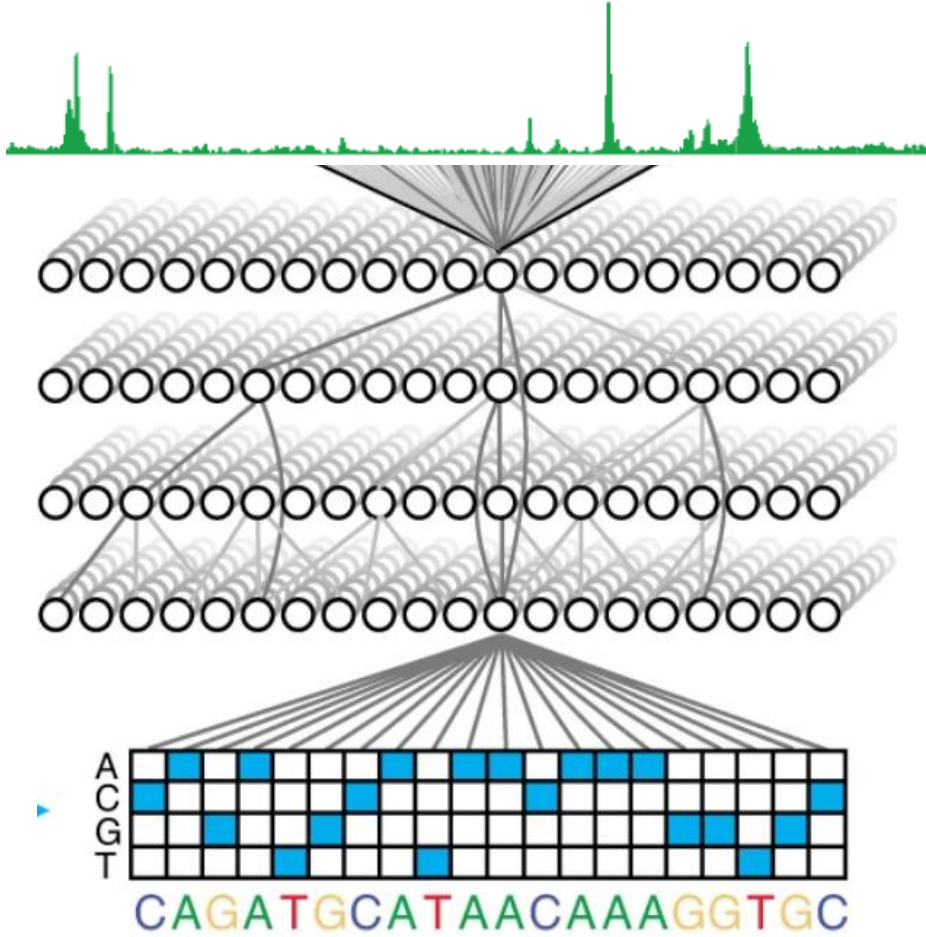
Peak



No peak

Base-resolution deep learning models of regulatory DNA

maps sequence to base-resolution coverage profiles



One model for every expt.

Base-resolution deep learning models of regulatory DNA

BPNet

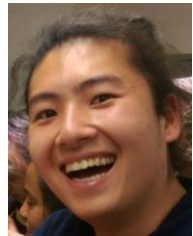
(TF Binding)

TF CHIP-exo, CHIP-seq, CUT&RUN

<https://github.com/kundajelab/bpnet>



Ziga Avsec



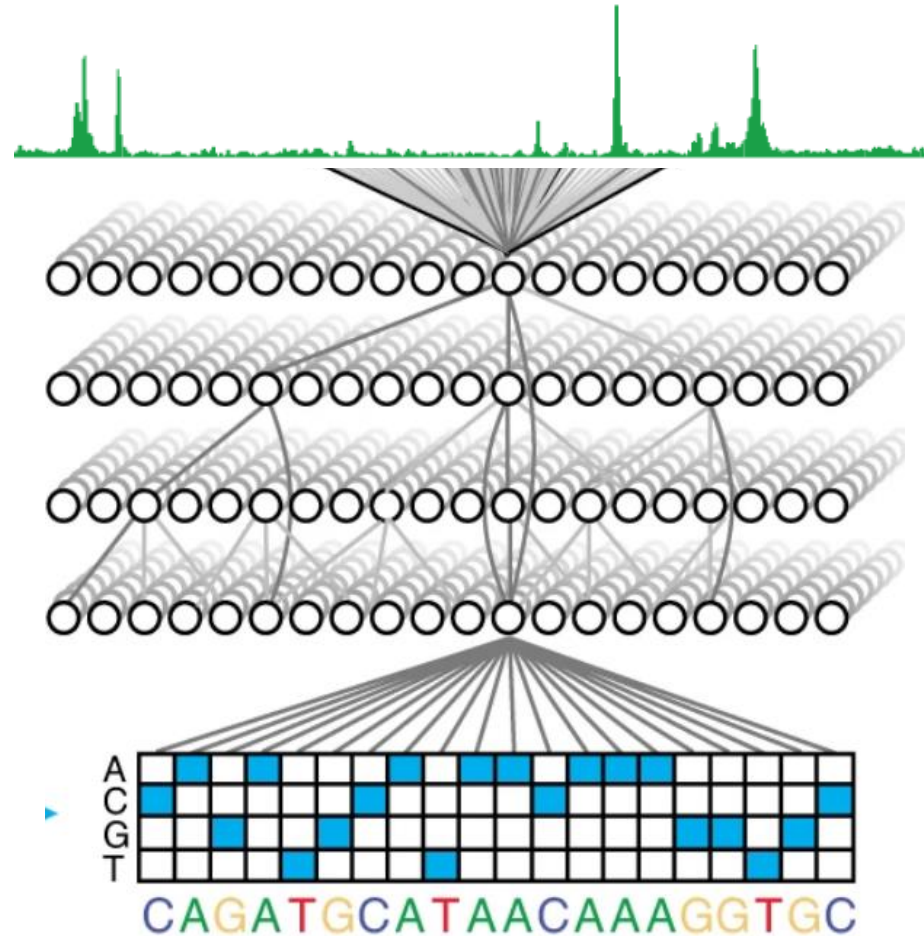
Alex Tseng



Vivek Ramalingam
(Postdoc)

Avsec et al. 2021, Nature Genetics

maps sequence to base-resolution
coverage profiles



One model for every expt.

Base-resolution deep learning models of regulatory DNA

BPNet

(TF Binding)

TF CHIP-exo, CHIP-seq, CUT&RUN

<https://github.com/kundajelab/bpnet>



Ziga Avsec



Alex Tseng



Vivek Ramalingam
(Postdoc)

Avsec et al. 2021, Nature Genetics

ChromBPNet

(Chromatin Accessibility)

ATAC-seq, DNase-seq, scATAC-seq

<http://github.com/kundajelab/chrombpnet>



Anusri Pampari

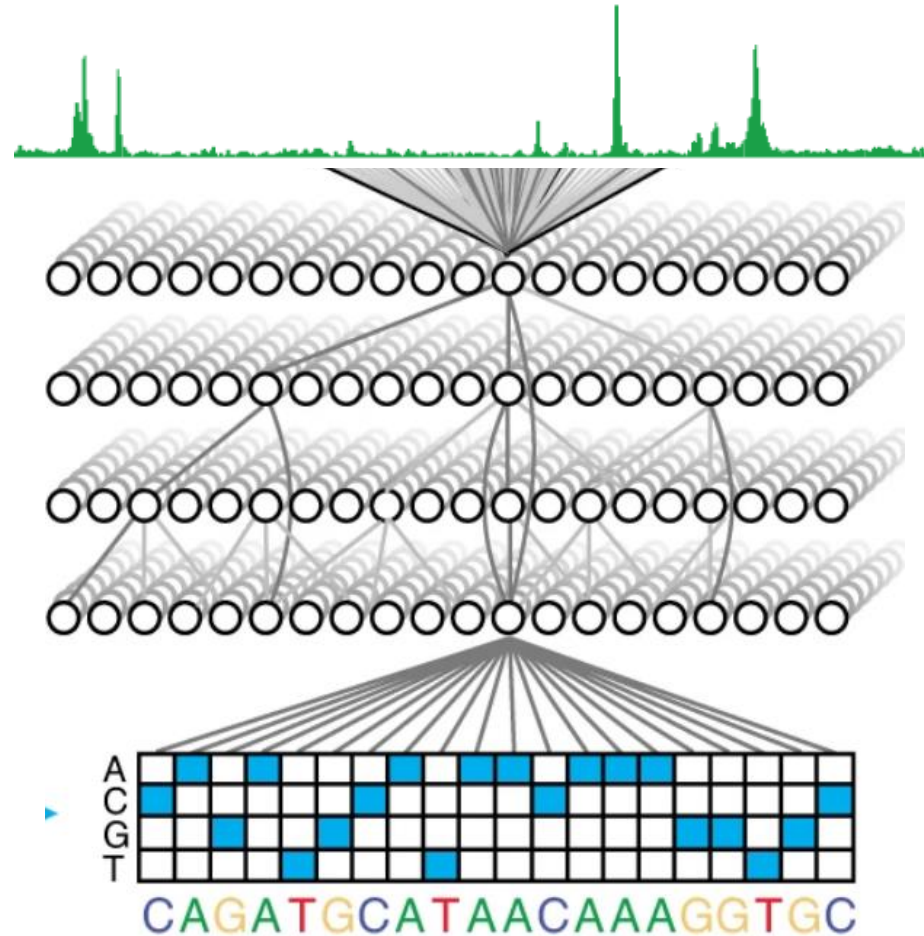


Anna Shcherbina



Surag Nair

maps sequence to base-resolution
coverage profiles



One model for every expt.

Base-resolution deep learning models of regulatory DNA

BPNet

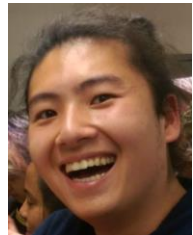
(TF Binding)

TF CHIP-exo, CHIP-seq, CUT&RUN

<https://github.com/kundajelab/bpnet>



Ziga Avsec



Alex Tseng



Vivek Ramalingam
(Postdoc)

Avsec et al. 2021, Nature Genetics

ChromBPNet

(Chromatin Accessibility)

ATAC-seq, DNase-seq, scATAC-seq

<http://github.com/kundajelab/chrombpnet>



Anusri Pampari

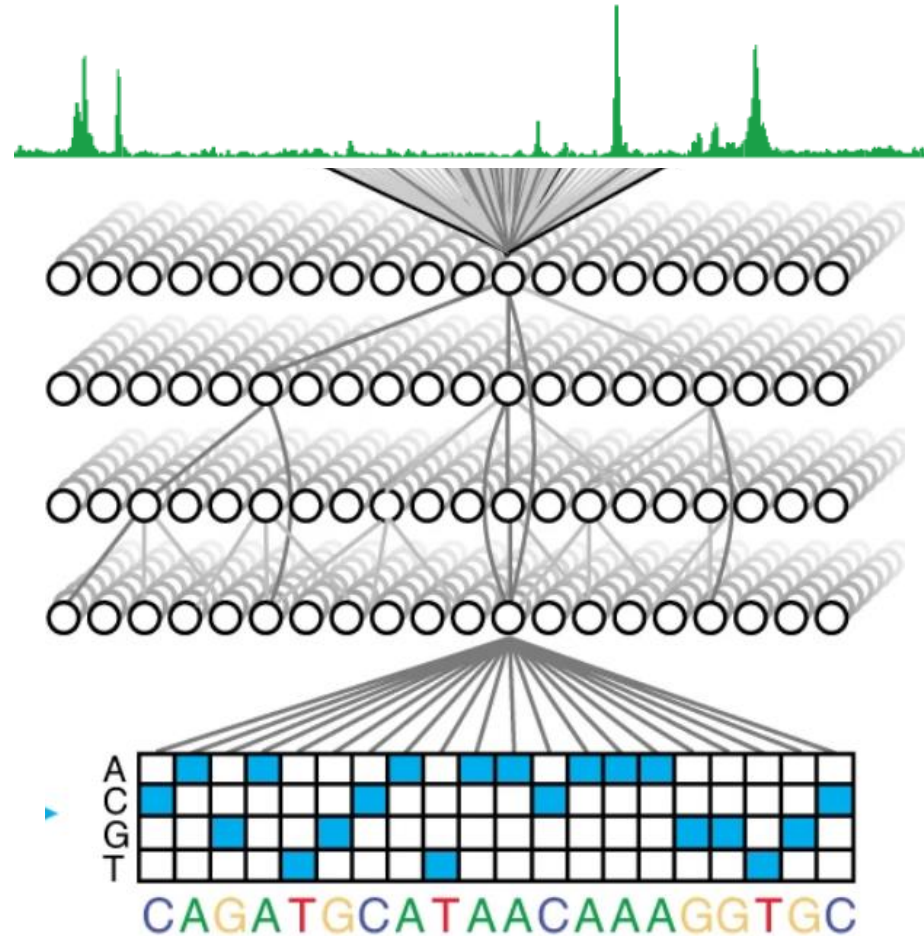


Anna Shcherbina



Surag Nair

maps sequence to base-resolution
coverage profiles



One model for every expt.

ProCapNet

(Nascent & steady state Tx)

PRO-cap, CAGE, RAMPAGE

https://github.com/kundajelab/nascent_RNA_models



Kelly Cochran

Base-resolution deep learning models of regulatory DNA

BPNet

(TF Binding)

TF CHIP-exo, CHIP-seq, CUT&RUN

<https://github.com/kundajelab/bpnet>



Ziga Avsec



Alex Tseng



Vivek Ramalingam
(Postdoc)

Avsec et al. 2021, Nature Genetics

ChromBPNet

(Chromatin Accessibility)

ATAC-seq, DNase-seq, scATAC-seq

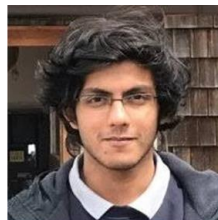
<http://github.com/kundajelab/chrombpnet>



Anusri Pampari

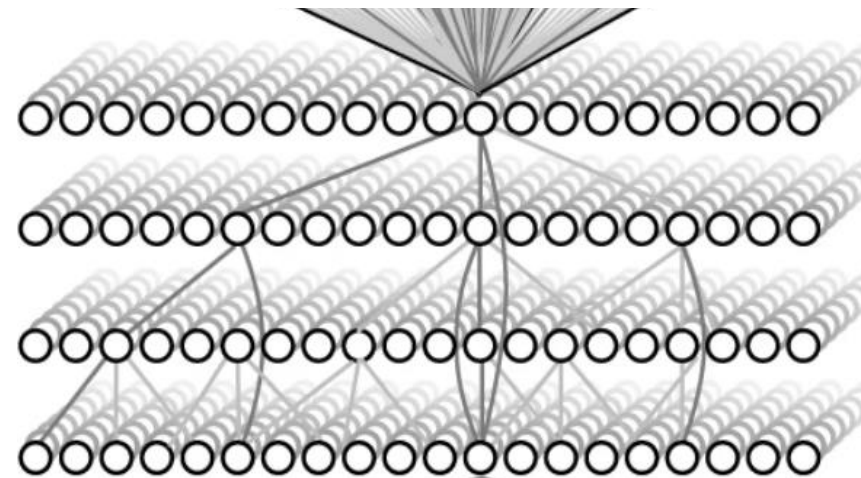


Anna Shcherbina



Surag Nair

maps sequence to base-resolution
coverage profiles



One model for every expt.

ProCapNet

(Nascent & steady state Tx)

PRO-cap, CAGE, RAMPAGE

https://github.com/kundajelab/nascent_RNA_models



Kelly Cochran

ReporterNet

(High throughput reporter assays)

MPRA, STARR-seq, ATAC-STARR-seq

(Coming soon!)



Ziwei Chen

Interpret predictive sequence features in reg. DNA via model

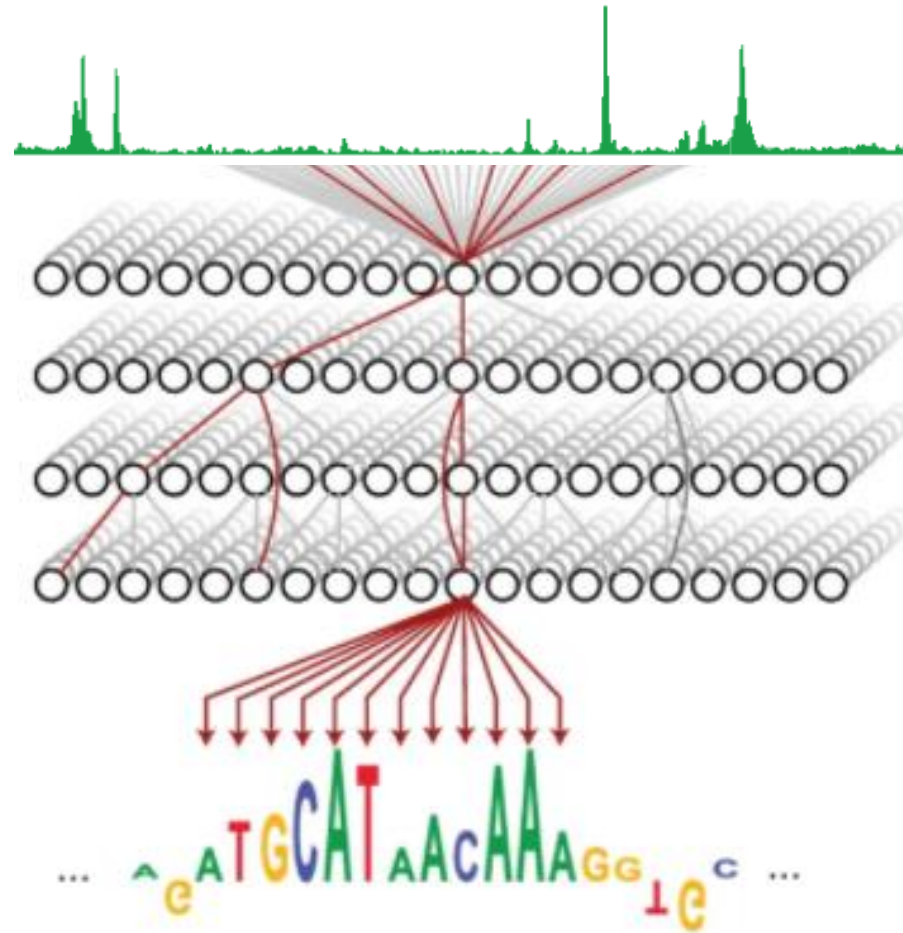
infers contribution of every base in any query sequence through lens of model

DeepLIFT



Avanti Shrikumar

<https://github.com/kundajelab/deeplift>



FastISM



Surag Nair

<https://github.com/kundajelab/fastism>

Yuzu



Jacob Schreiber

<https://github.com/kundajelab/yuzu>

Shrikumar et al. 2017, ICML

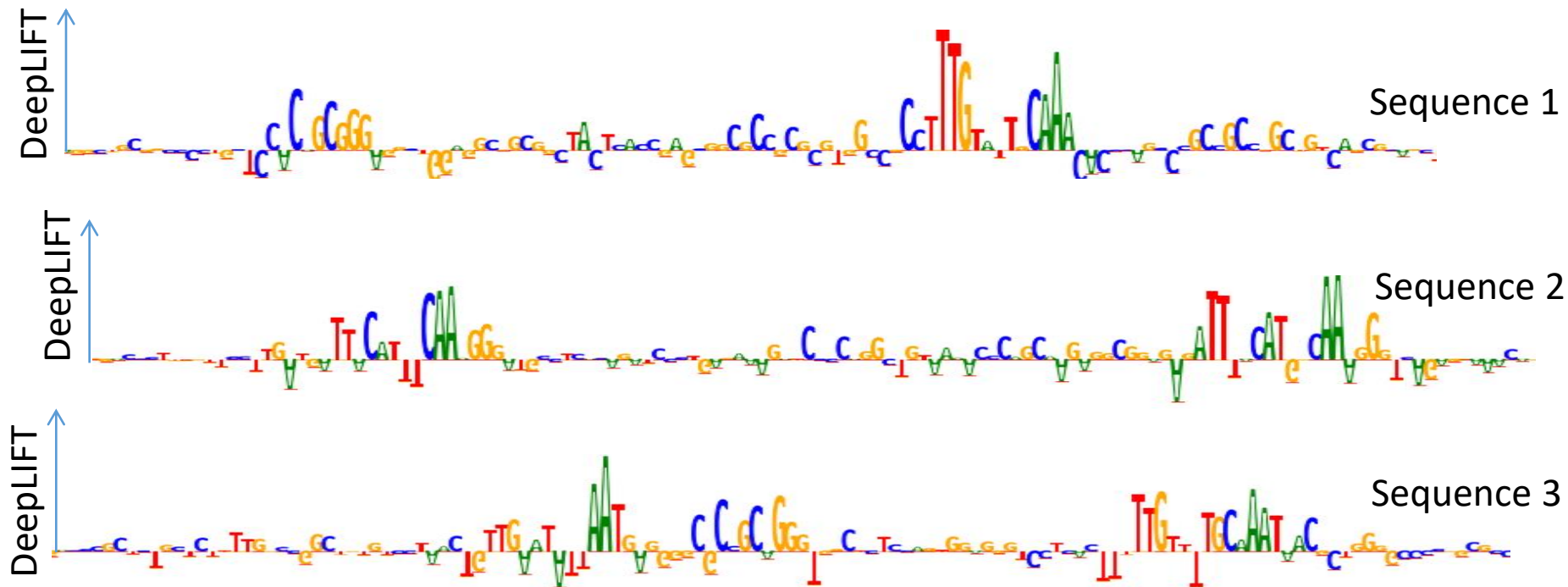
Tseng et al. 2020, NeurIPS

Nair et al, 2022, Bioinformatics

Schreiber et al. 2022, Biorxiv

Summarize predictive motif patterns genome-wide

TF-MODISCO & FiNeMo



Avanti Shrikumar



Alex Tseng



Jacob Schreiber

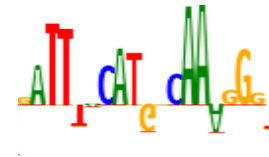
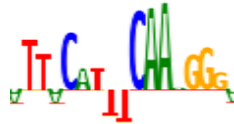
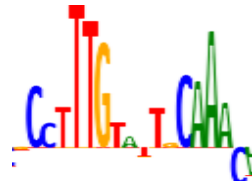


Austin Wang

<https://github.com/jmschrei/tfmodisco-lite>
https://github.com/austintwang/finemo_gpu

Summarize predictive motif patterns genome-wide

TF-MODISCO & FiNeMo



Avanti Shrikumar



Alex Tseng



Jacob Schreiber



Austin Wang

Summarize predictive motif patterns genome-wide

TF-MODISCO & FiNeMo



Avanti Shrikumar



Alex Tseng



Jacob Schreiber



Austin Wang

<https://github.com/jmschrei/tfmodisco-lite>
https://github.com/austintwang/finemo_gpu

Summarize predictive motif patterns genome-wide

TF-MODISCO & FiNeMo



Avanti Shrikumar



Alex Tseng



Jacob Schreiber

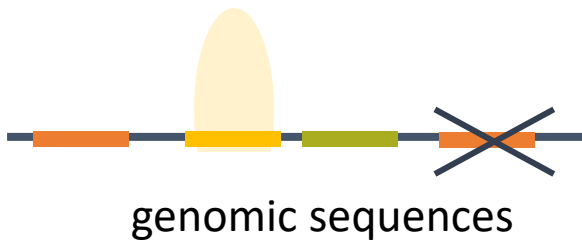


Austin Wang

<https://github.com/jmschrei/tfmodisco-lite>
https://github.com/austintwang/finemo_gpu

In-silico perturbation framework for causal discovery of syntax & variation

Combinatorial perturbation screens
on synthetic & genomic sequences



Ziga Avsec



Daniel Kim



Vivek Ramalingam
(Postdoc)

Avsec et al. 2021, Nature Genetics

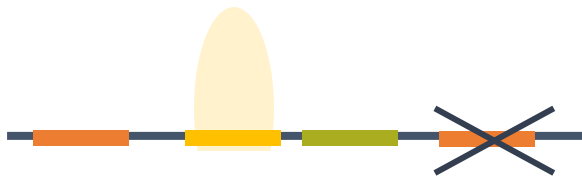
Kim et al. 2021, Nature Genetics

In-silico perturbation framework for causal discovery of syntax & variation

Combinatorial perturbation screens on synthetic & genomic sequences



synthetic sequences



genomic sequences



Ziga Avsec



Daniel Kim

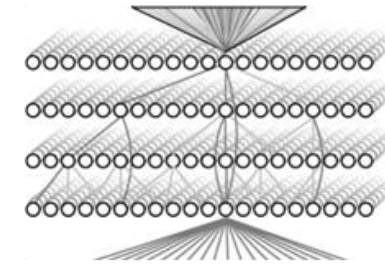


Vivek Ramalingam
(Postdoc)

Avsec et al. 2021, Nature Genetics
Kim et al. 2021, Nature Genetics

Variant effect screens
(Common, rare, SNVs, indels)

Δ PredictedSignal

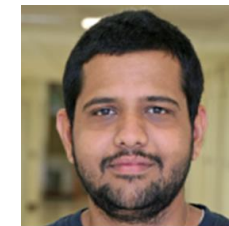


.....ACTGAT **C** GCAATCG.....

.....ACTGAT **G** GCAATCG.....



Soumya Kundu



Lakshman
Sundaram



Ziwei Chen

Debugging, de-biasing & reconciling chromatin accessibility data with ChromBPNet

ChromBPNet

(Chromatin Accessibility)

ATAC-seq, DNase-seq, scATAC-seq

<http://github.com/kundajelab/chrombpnet>



Anusri Pampari

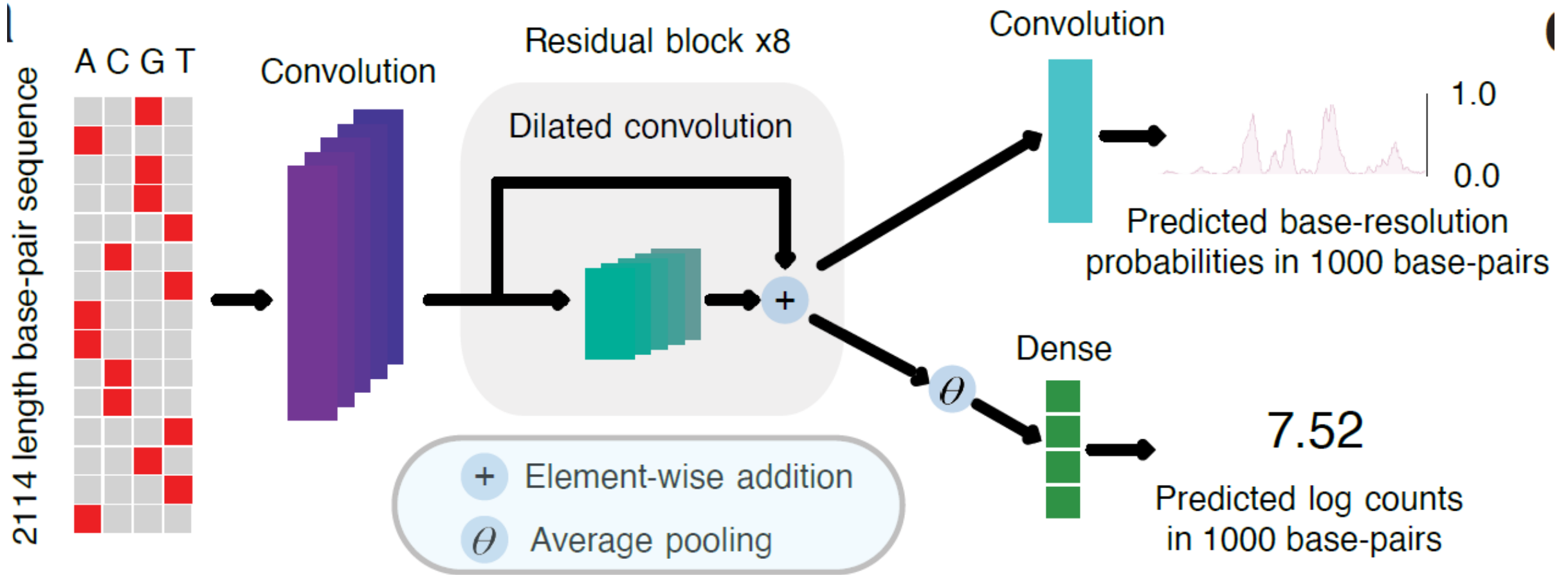


Anna Shcherbina



Surag Nair

ChromBPNet: Sequence to base-pair chromatin accessibility profiles



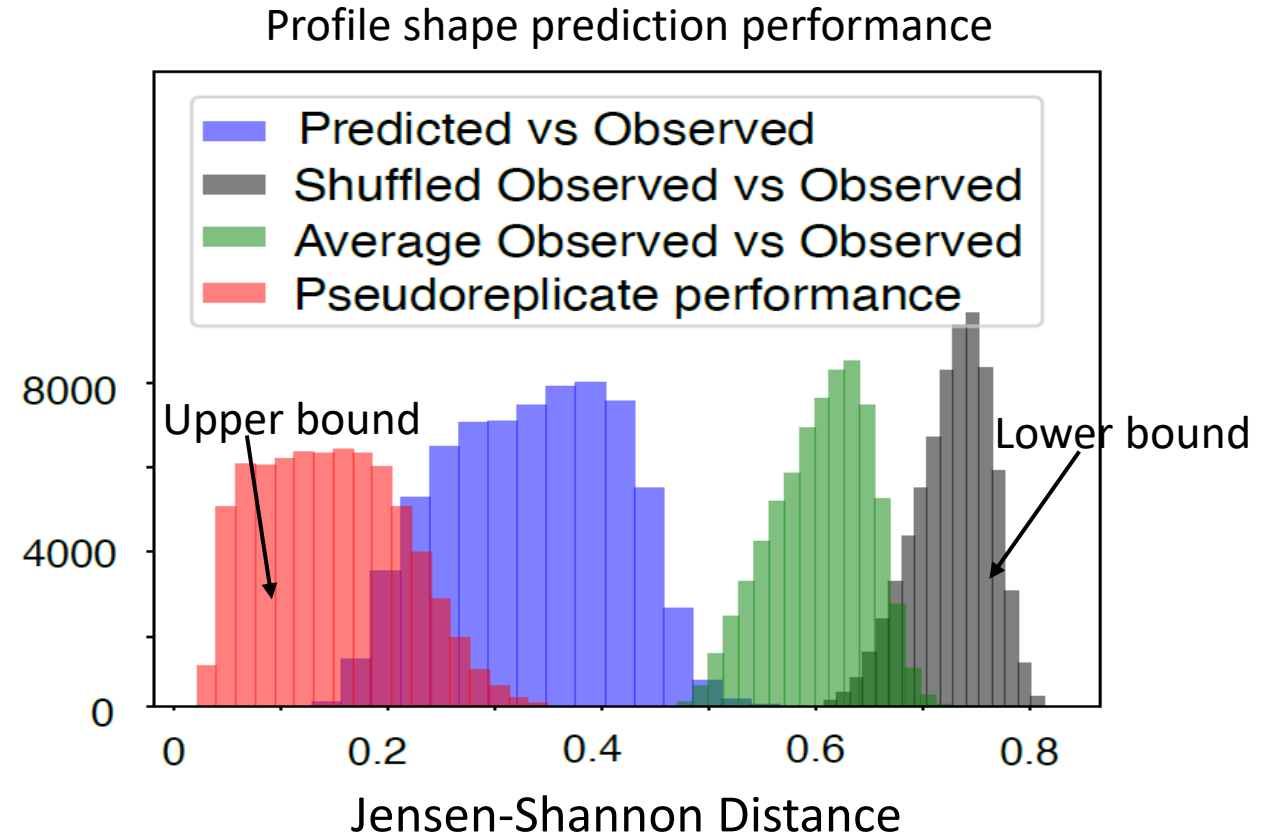
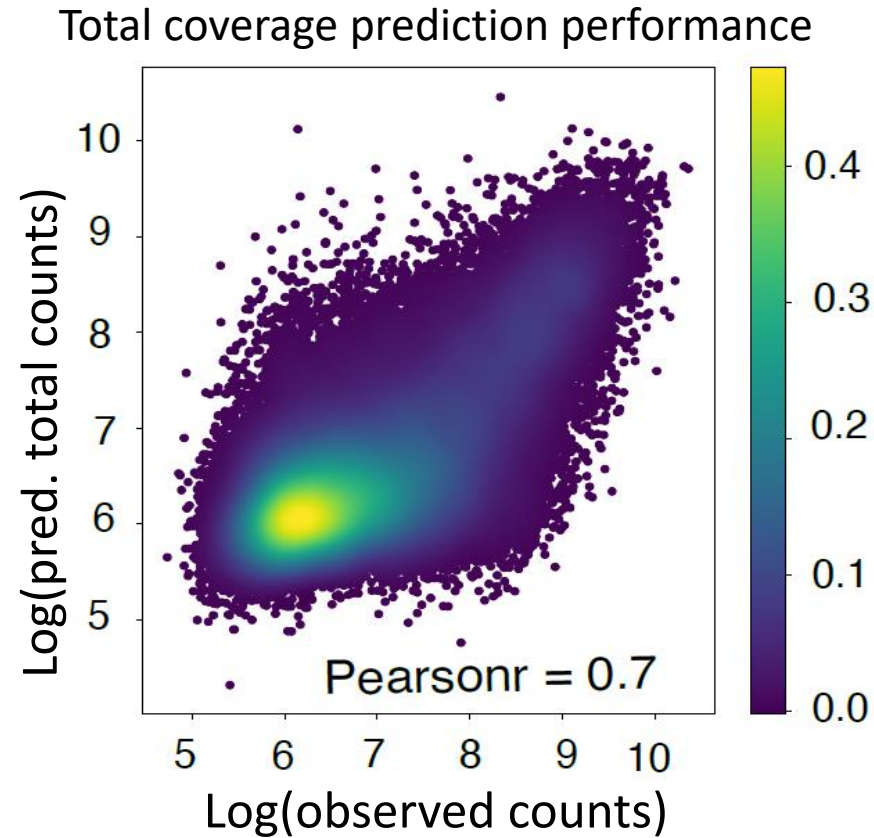
Anusri Pampari



Anna Shcherbina

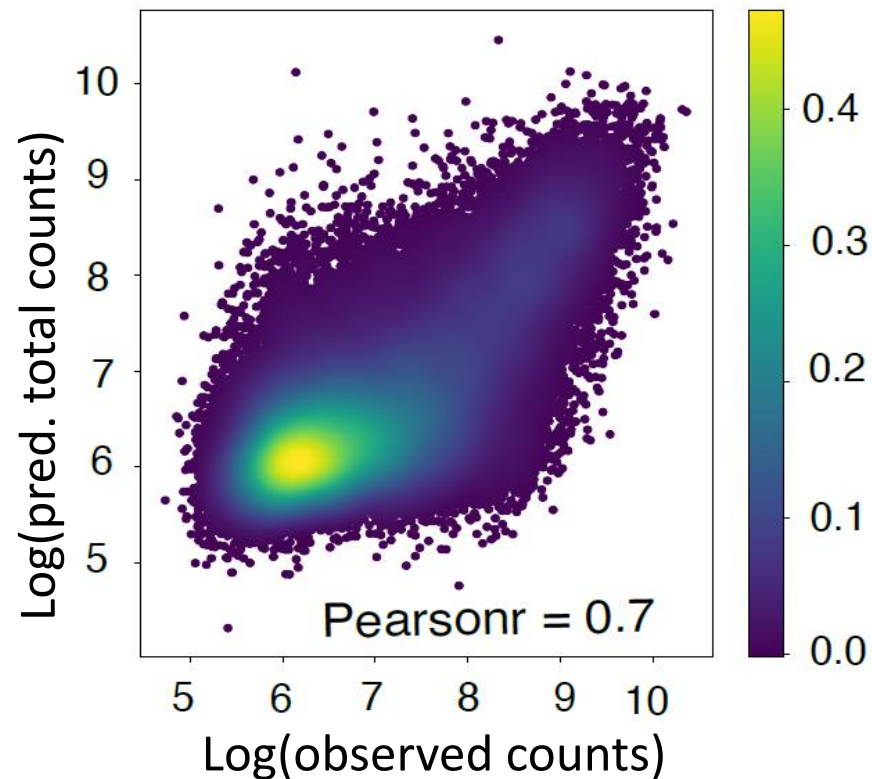
Fig: Jacob Schreiber

ChromBPNet: accurate prediction of ATAC-seq profiles from sequence

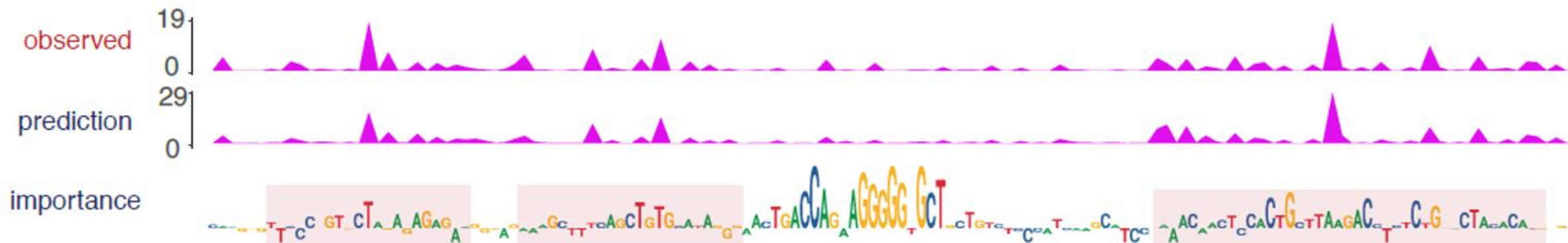
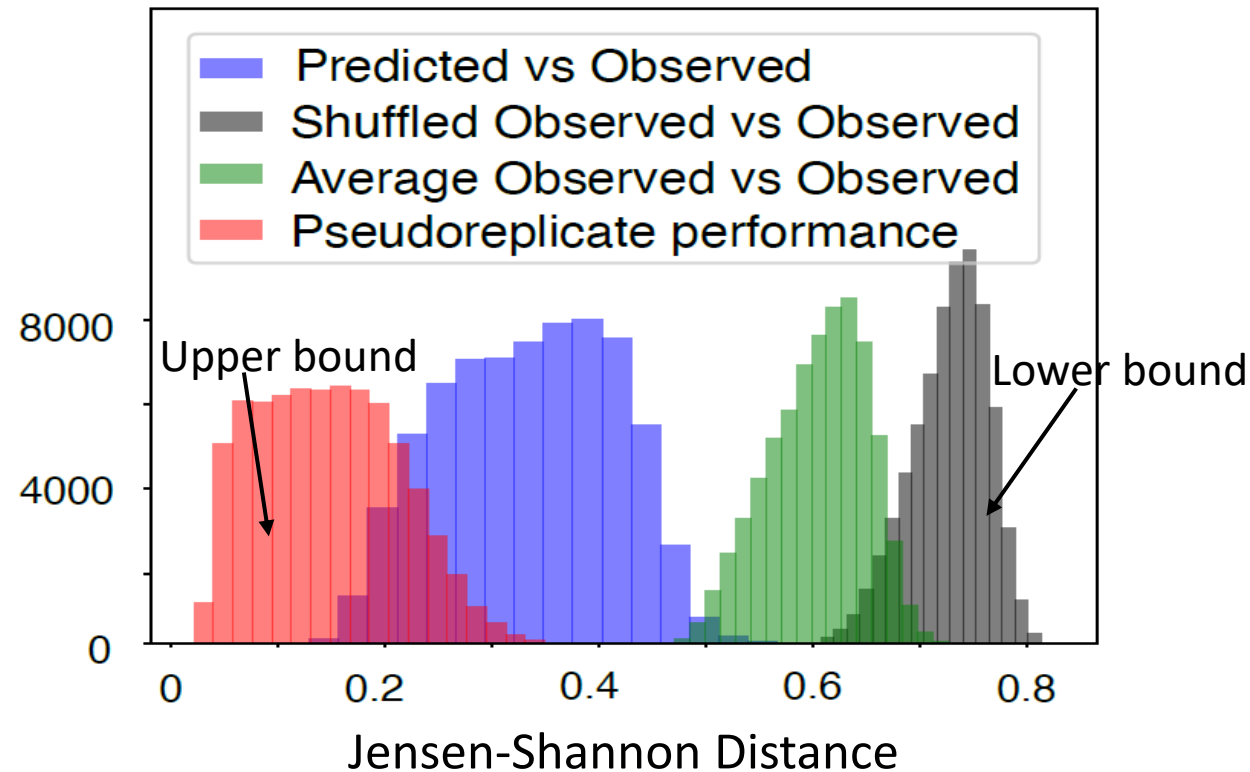


ChromBPNet: accurate prediction of ATAC-seq profiles from sequence

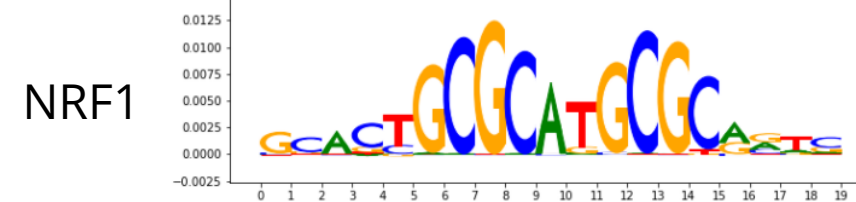
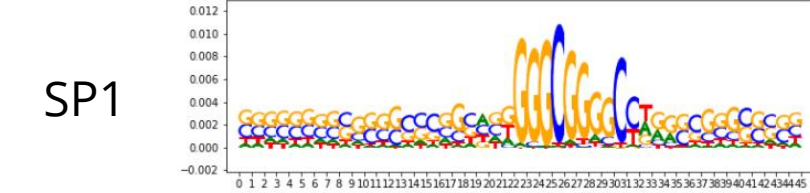
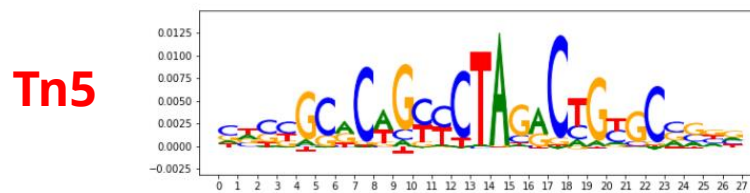
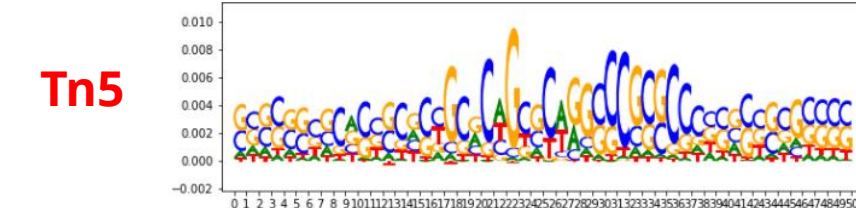
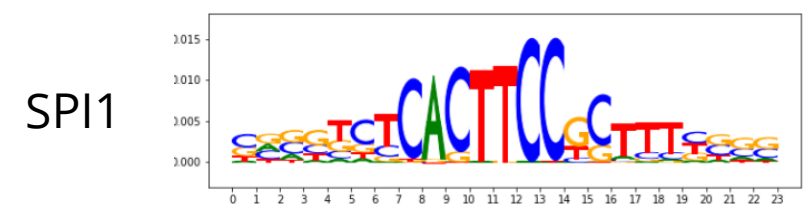
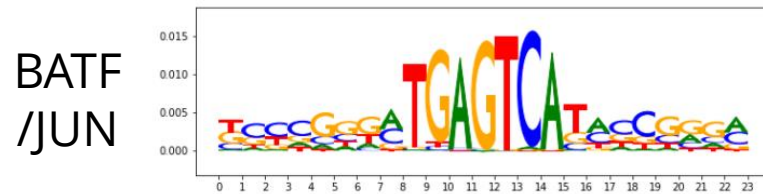
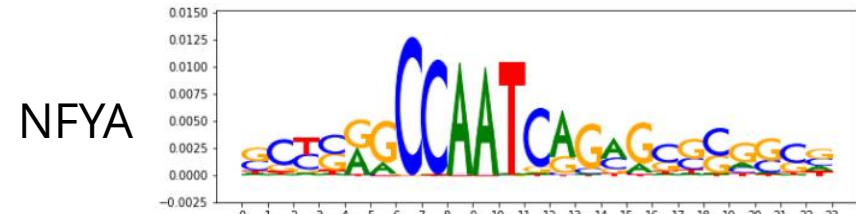
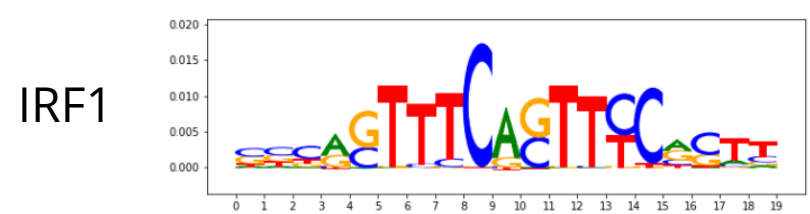
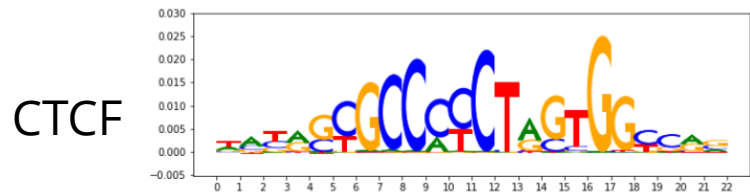
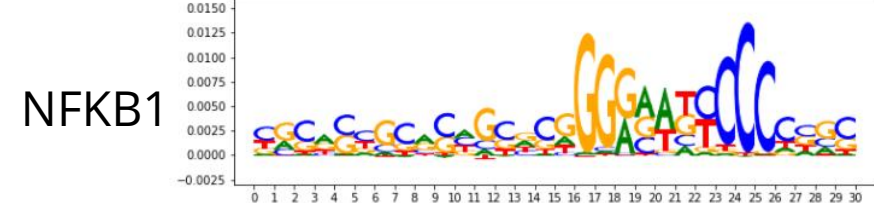
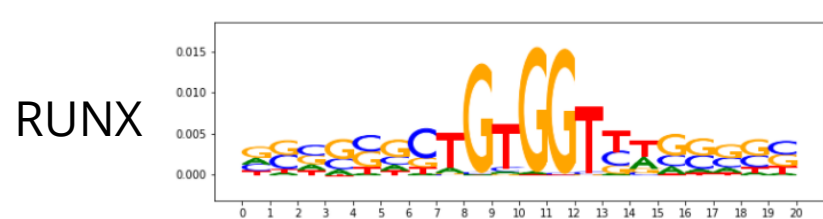
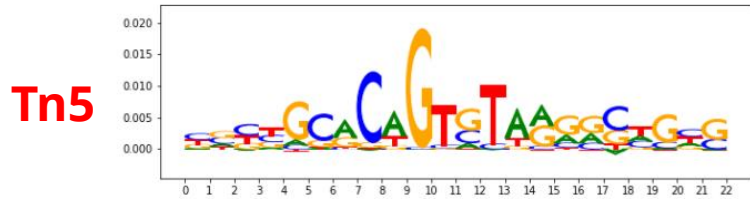
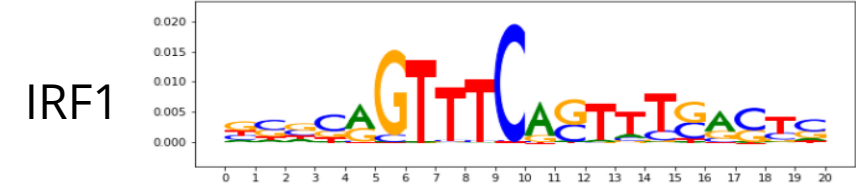
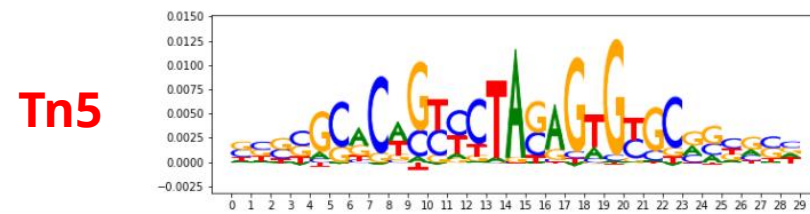
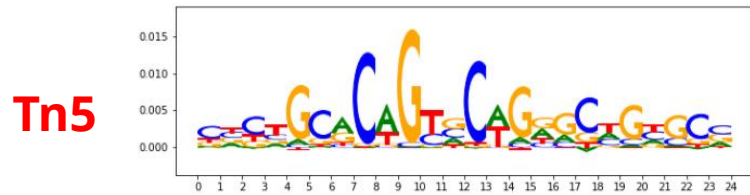
Total coverage prediction performance



Profile shape prediction performance

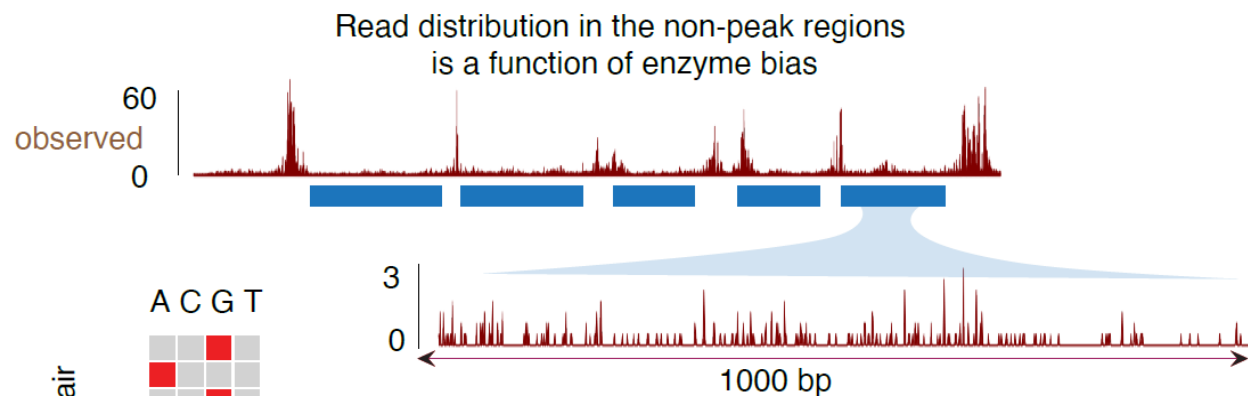


Motifs learned by model are heavily corrupted by Tn5 (ATAC-seq enzyme) sequence bias!



How to correct bias?

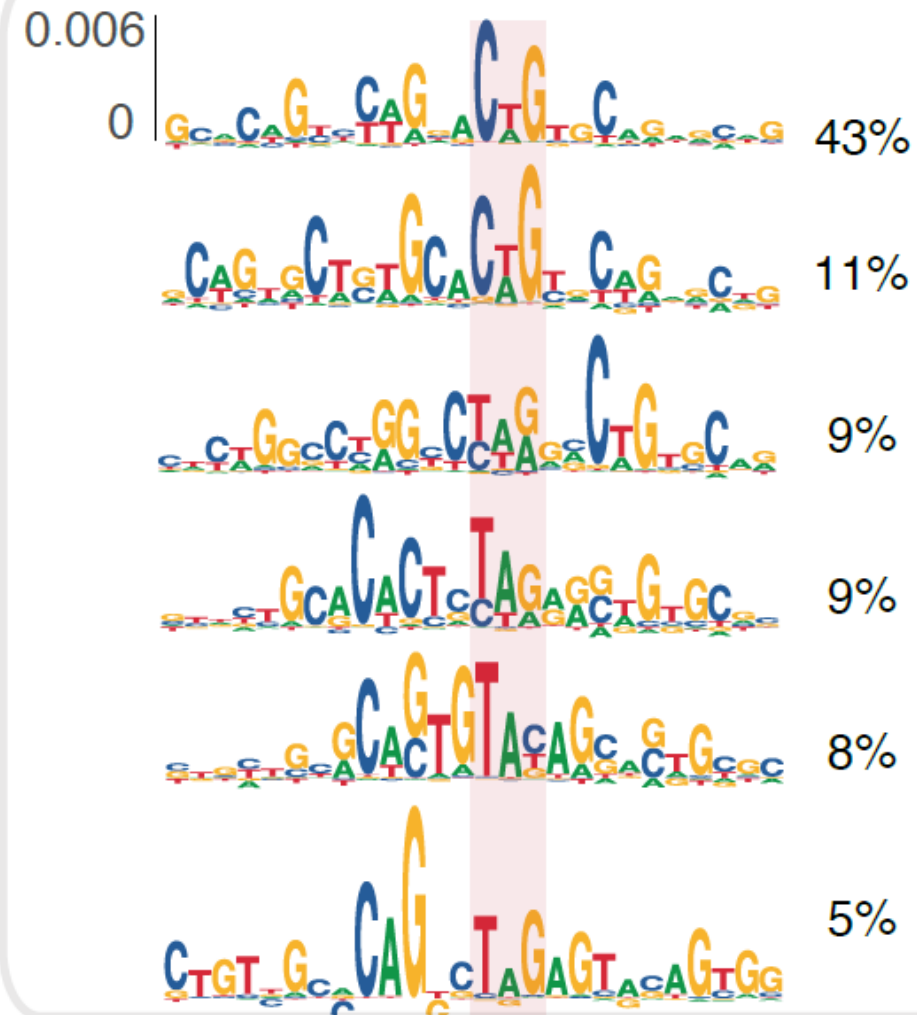
Use neural network to learn Tn5 bias from chromatin background



2114 length base-pair sequence

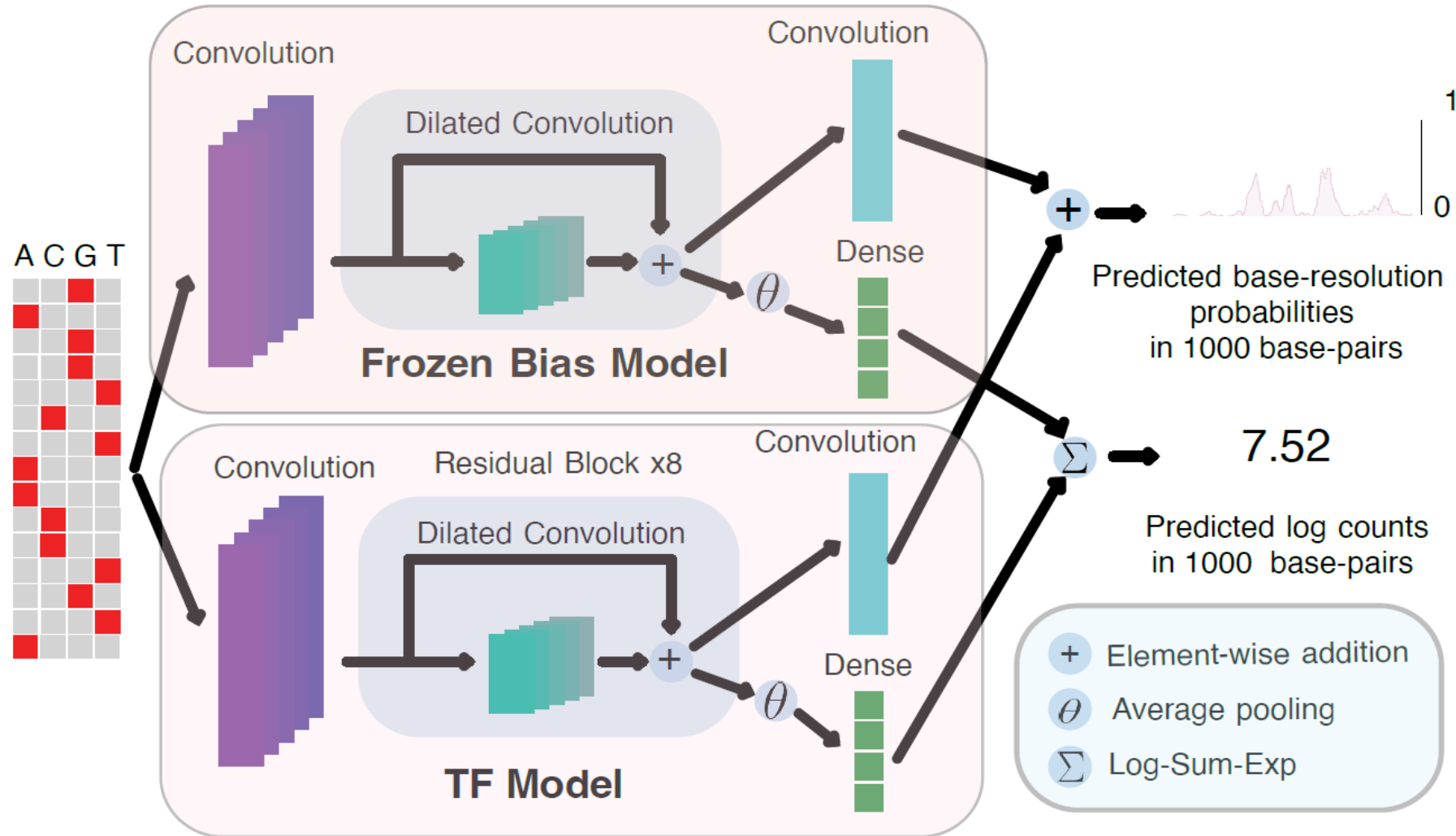


BPNet Model
(128 filters, 4 dilation layers)
Bias Model



How to correct bias?

“Bias-factorized” ChromBPNet model



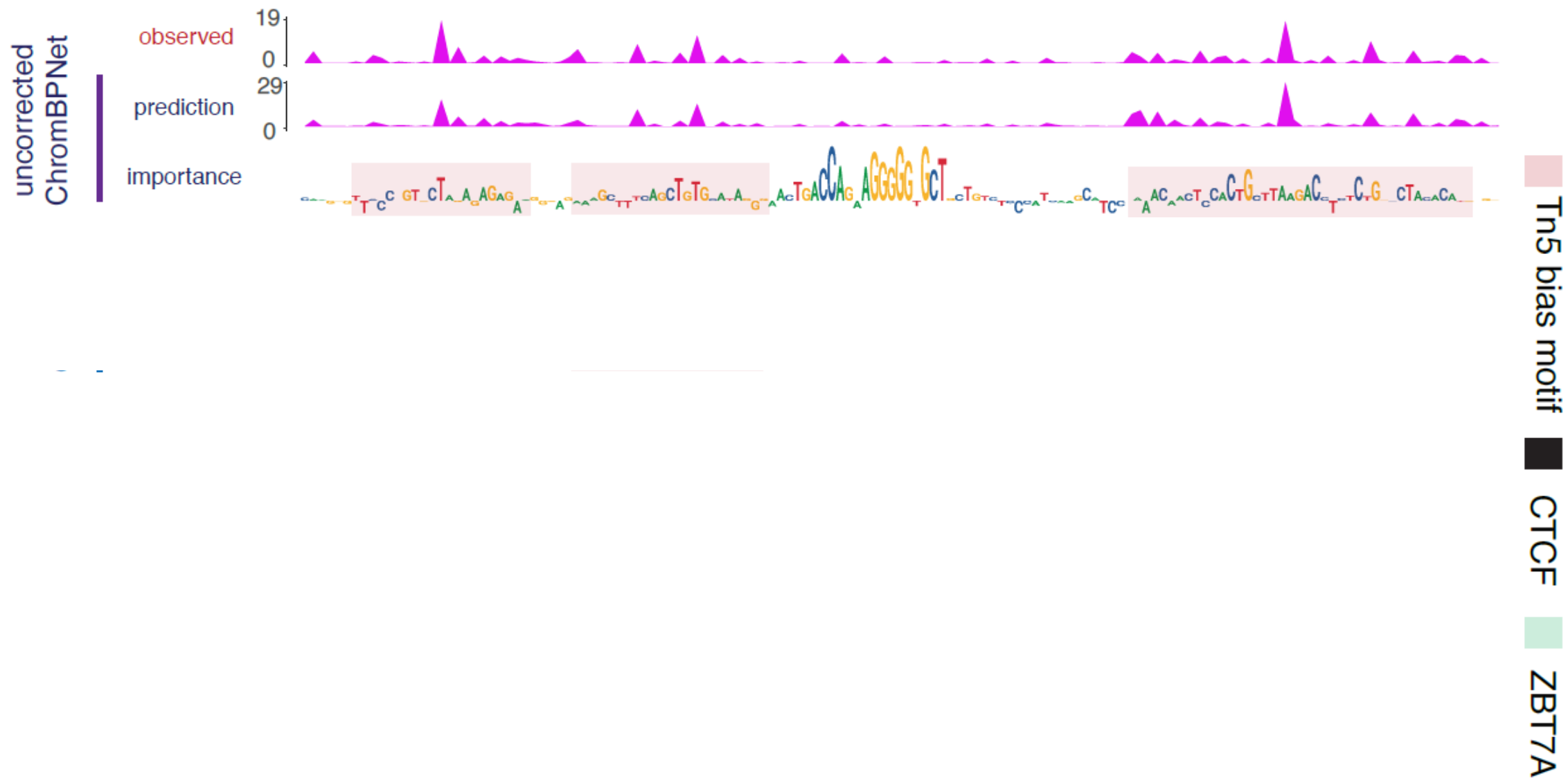
Anusri Pampari

Fig: Jacob Schreiber

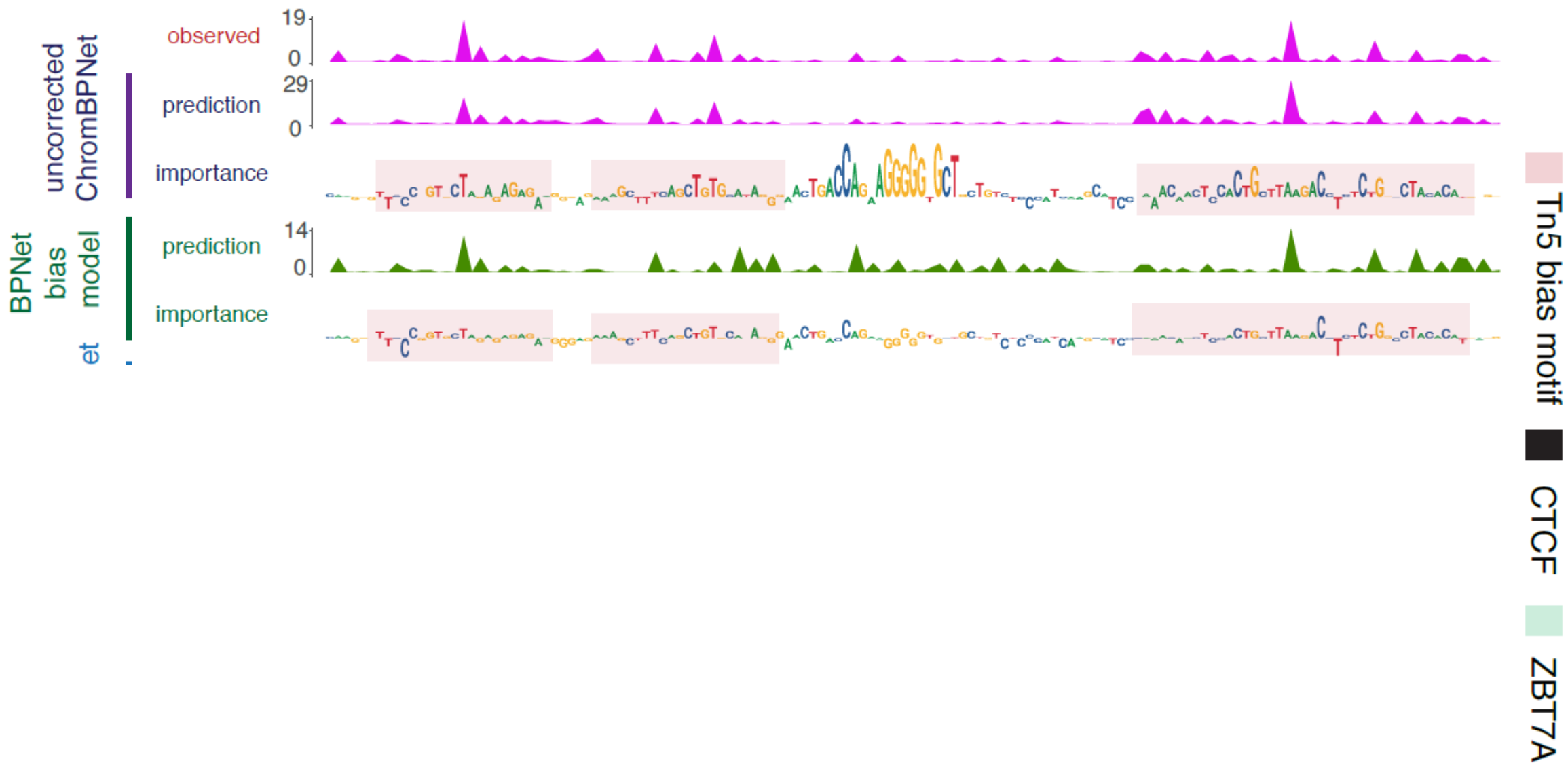


Anna Shcherbina

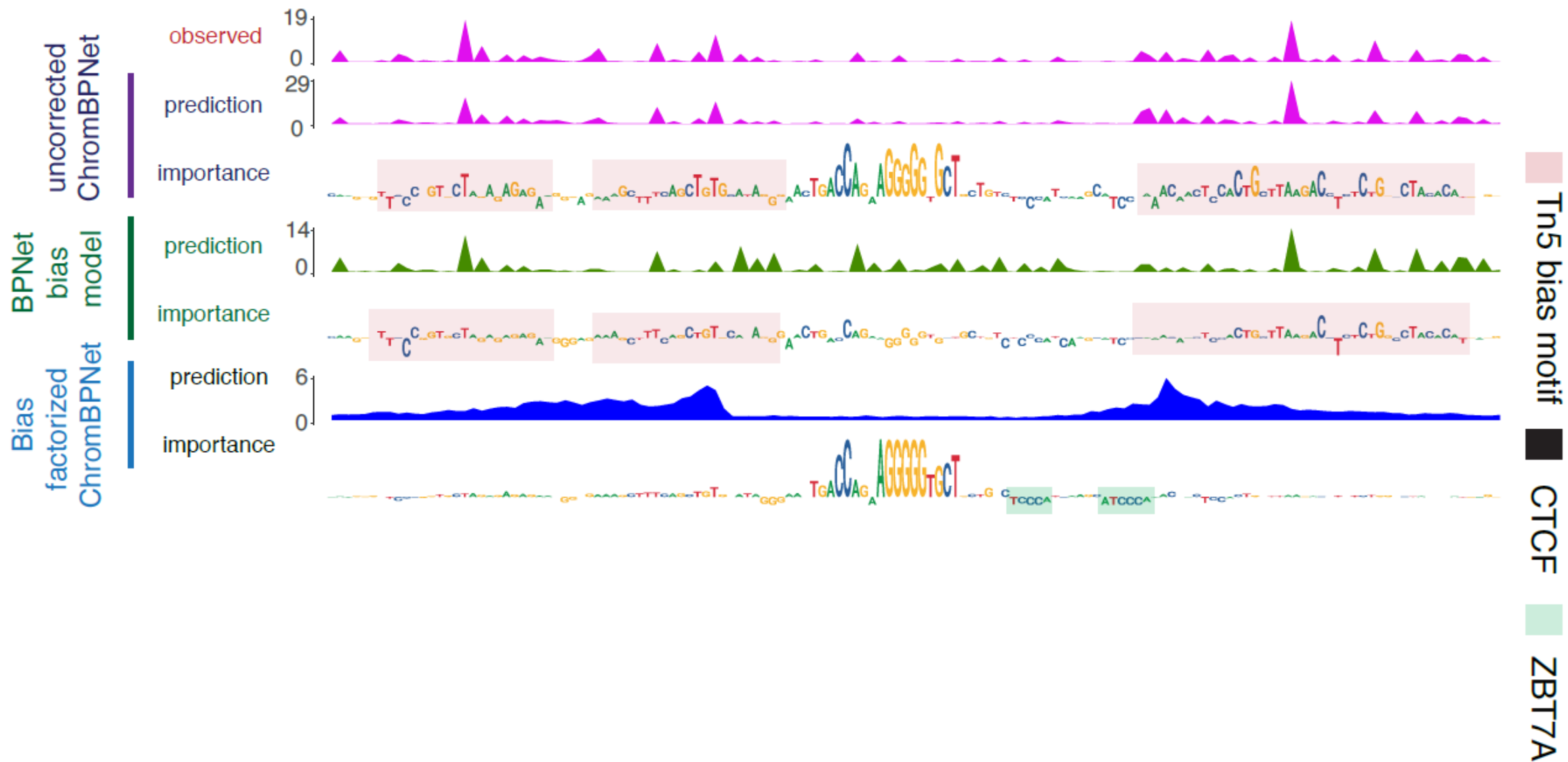
ChromBPNet provides denoising and imputation of footprints at individual loci



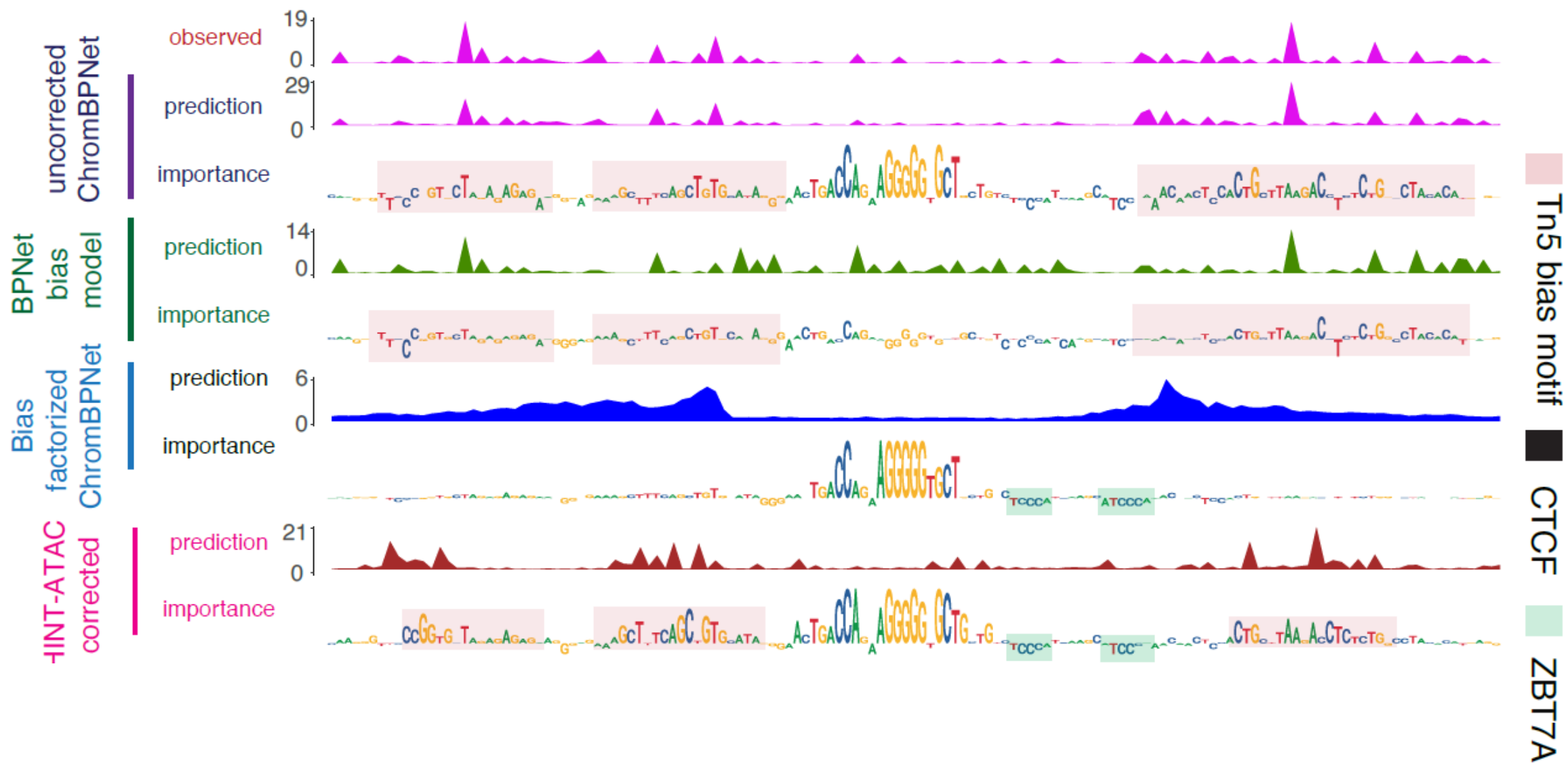
ChromBPNet provides denoising and imputation of footprints at individual loci



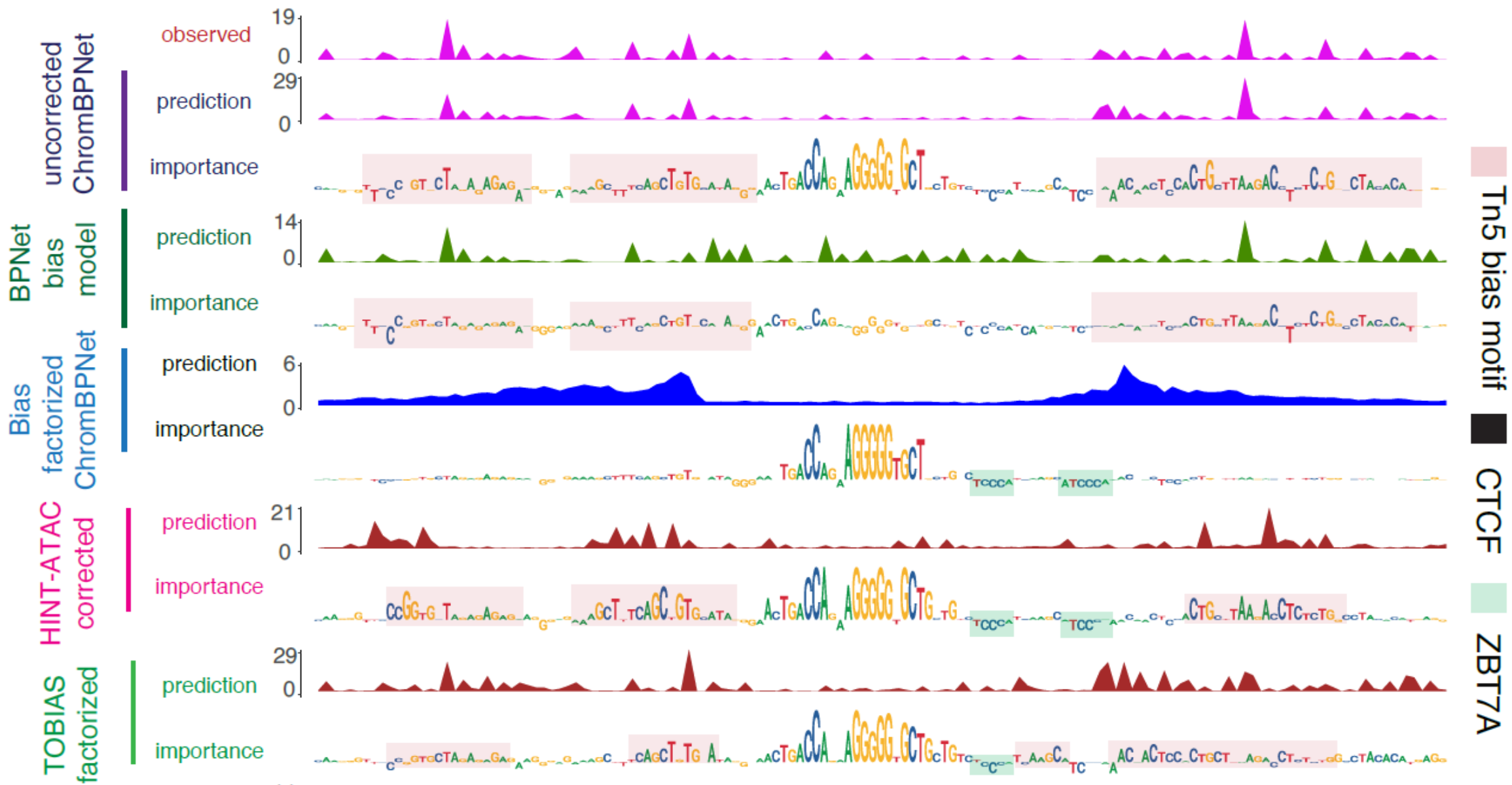
ChromBPNet provides denoising and imputation of footprints at individual loci



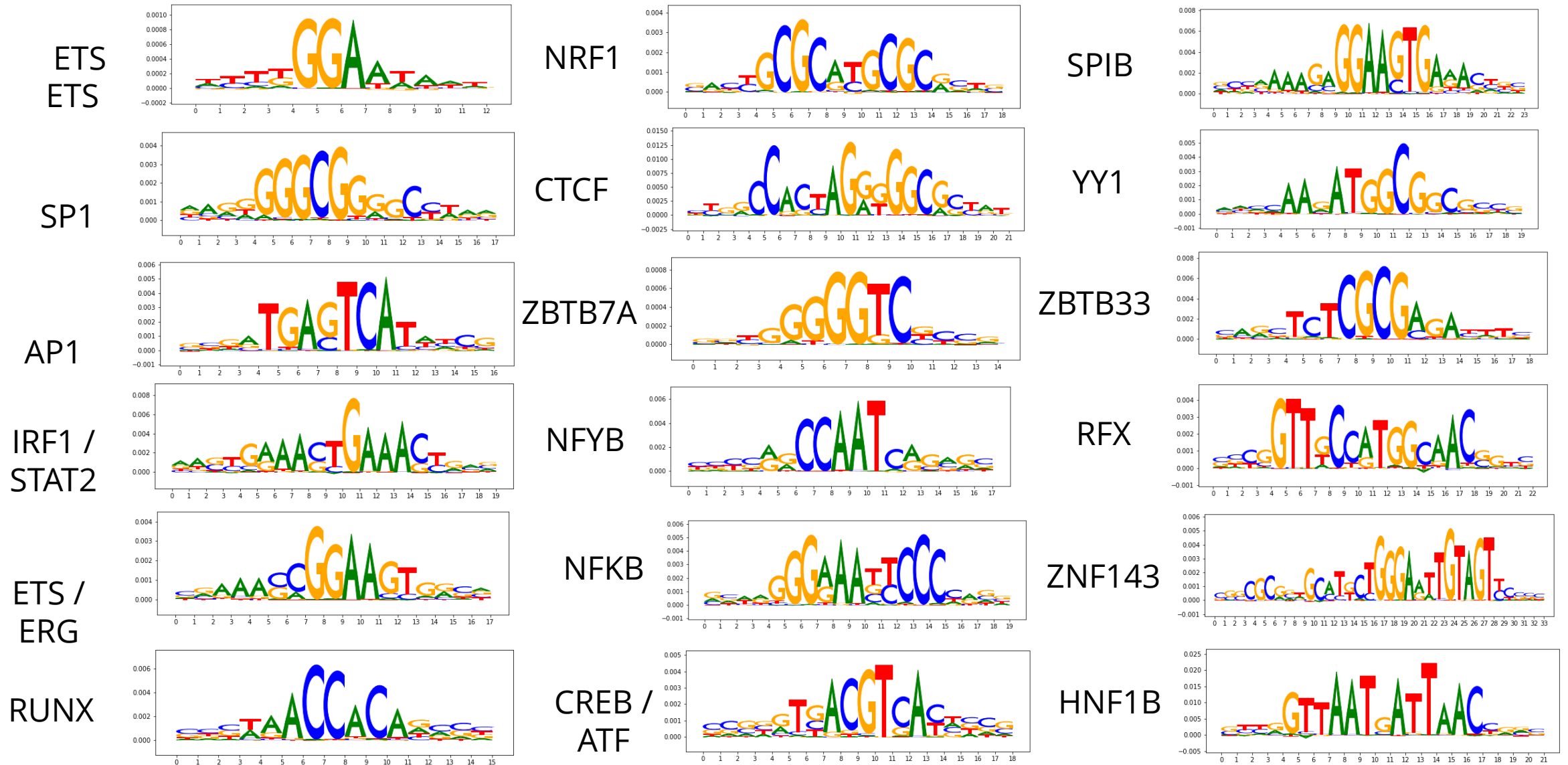
ChromBPNet provides denoising and imputation of footprints at individual loci



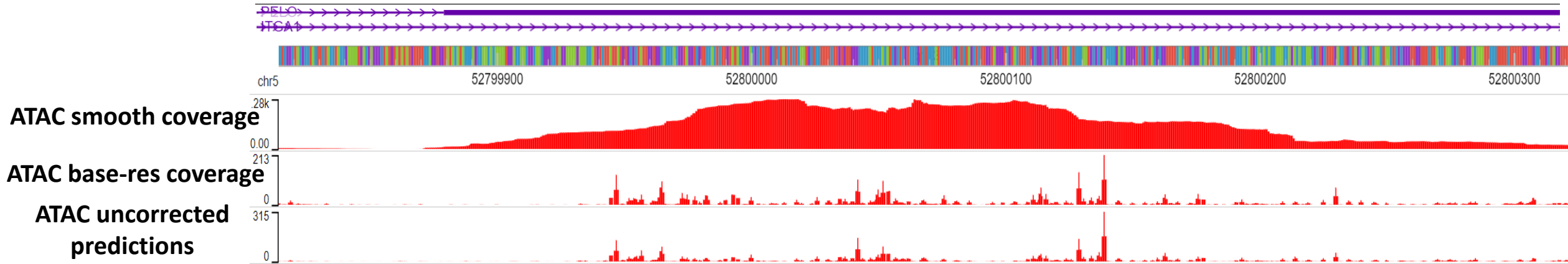
ChromBPNet provides denoising and imputation of footprints at individual loci



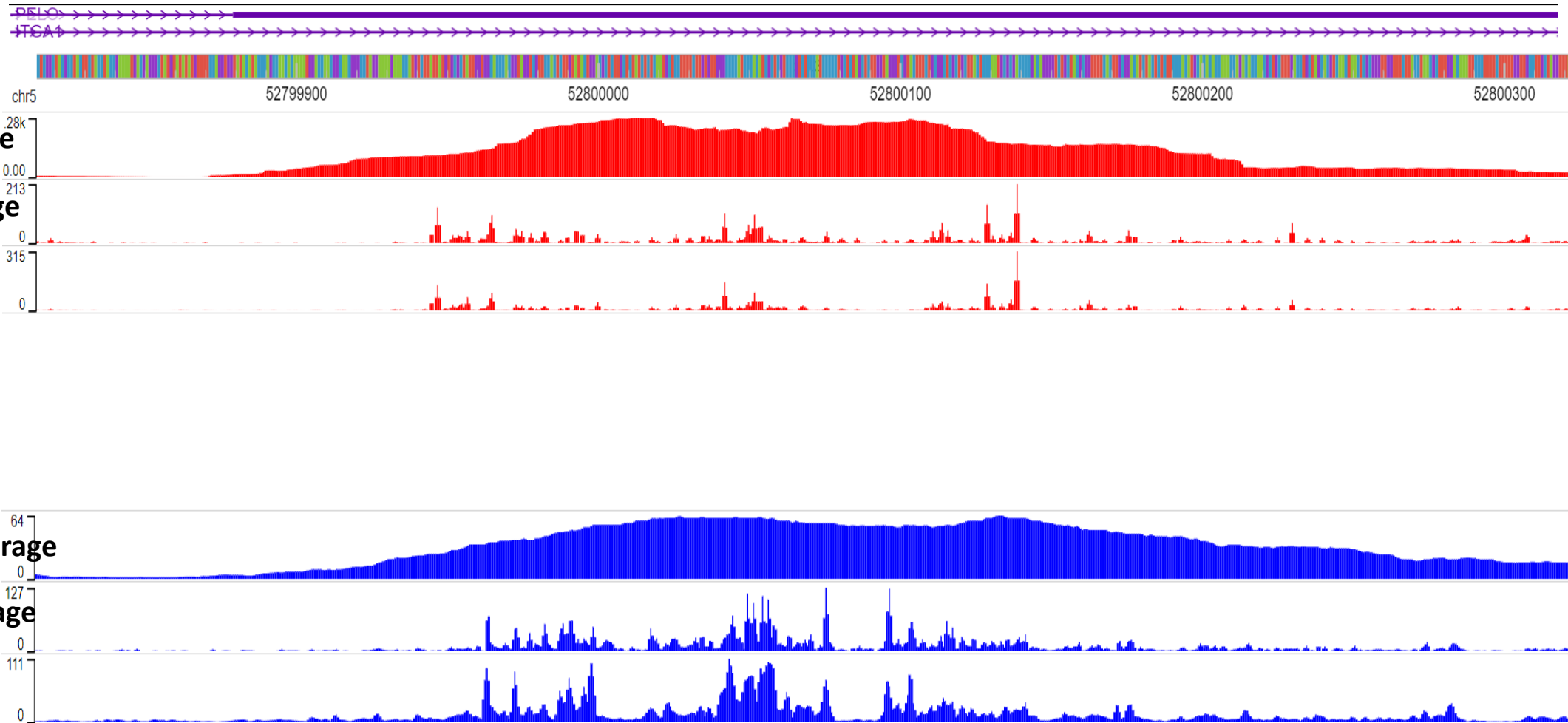
Motifs learned by bias-factorized ChromBPNet fully corrects Tn5 bias



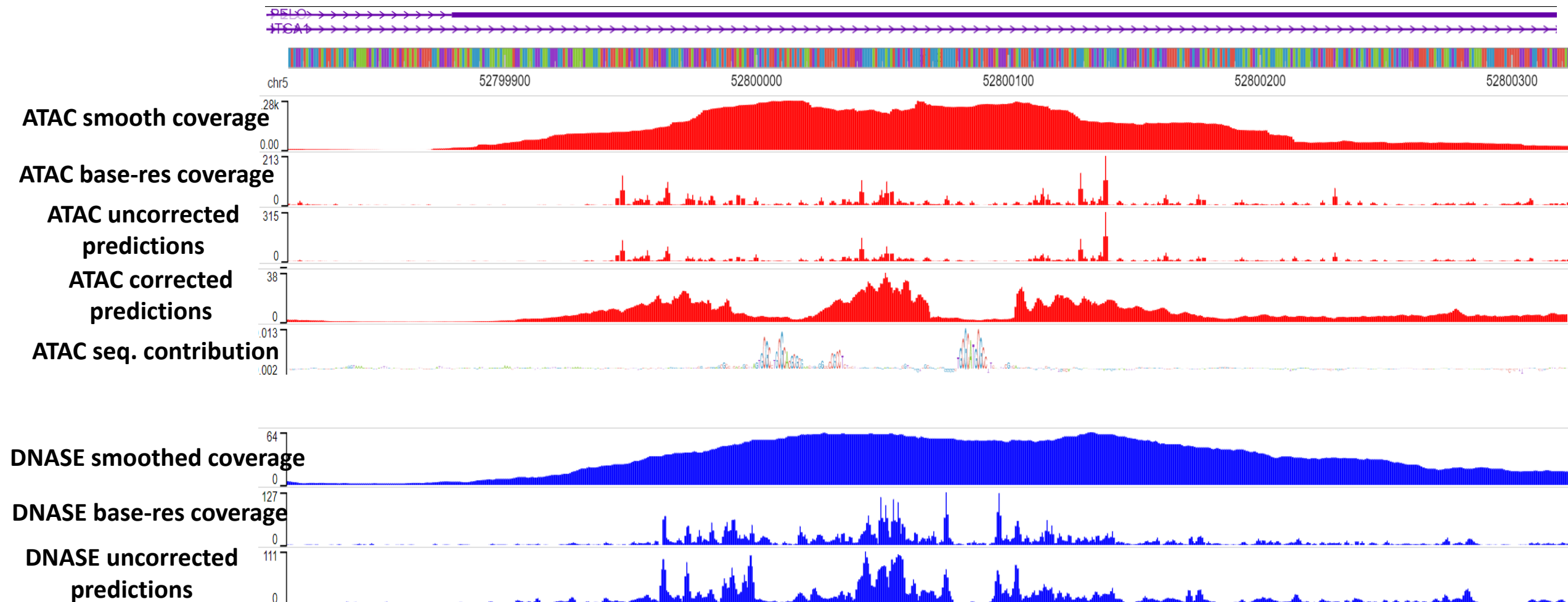
ChromBPNet reconciles DNase-seq and ATAC-seq experiments after bias correction



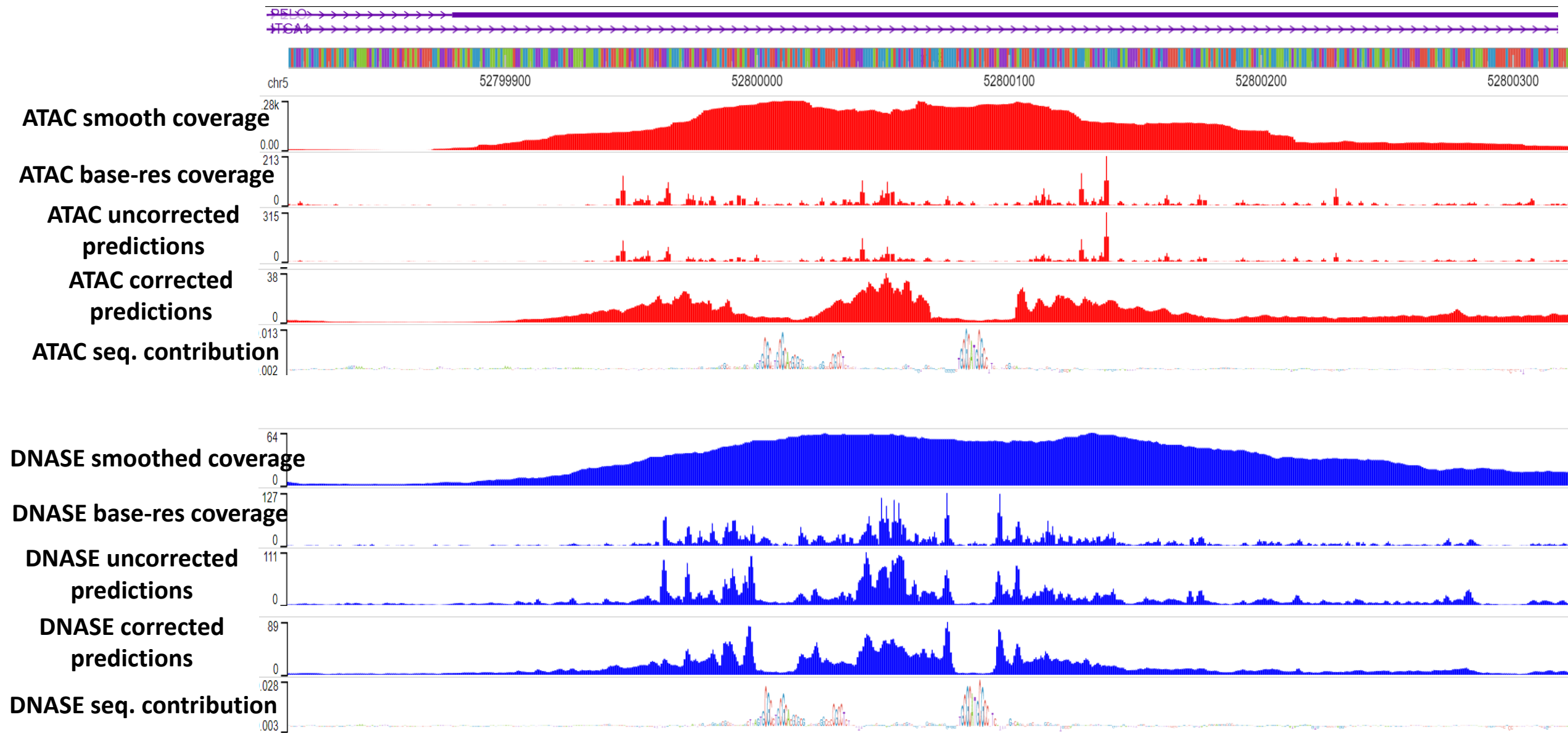
ChromBPNet reconciles DNase-seq and ATAC-seq experiments after bias correction



ChromBPNet reconciles DNase-seq and ATAC-seq experiments after bias correction

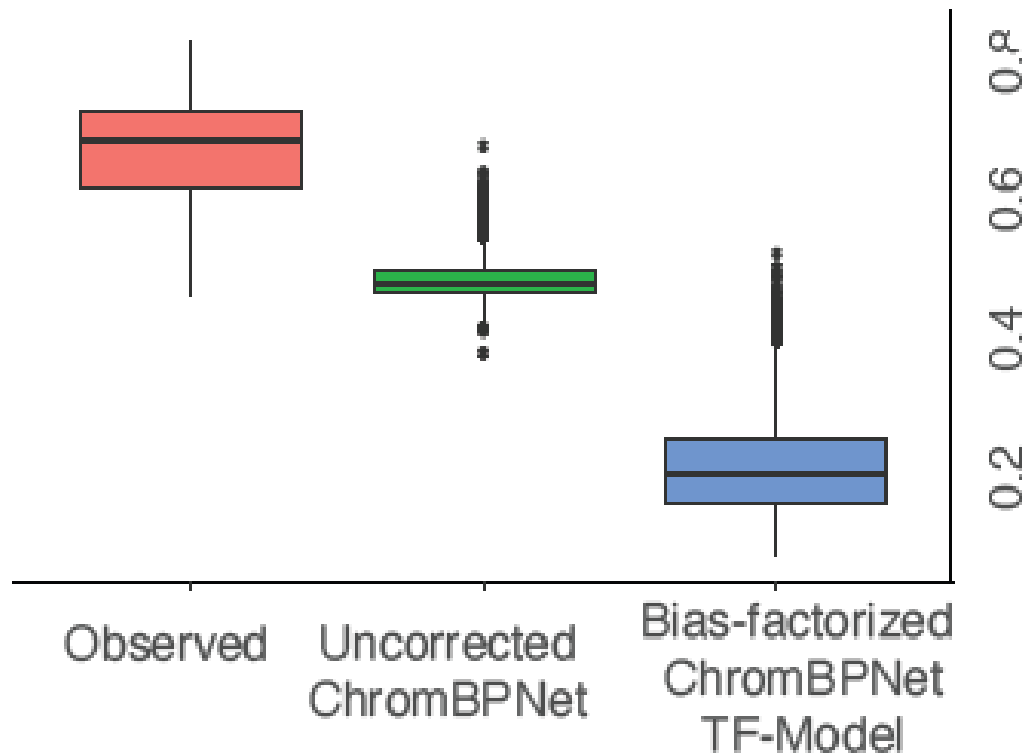


ChromBPNet reconciles DNase-seq and ATAC-seq experiments after bias correction

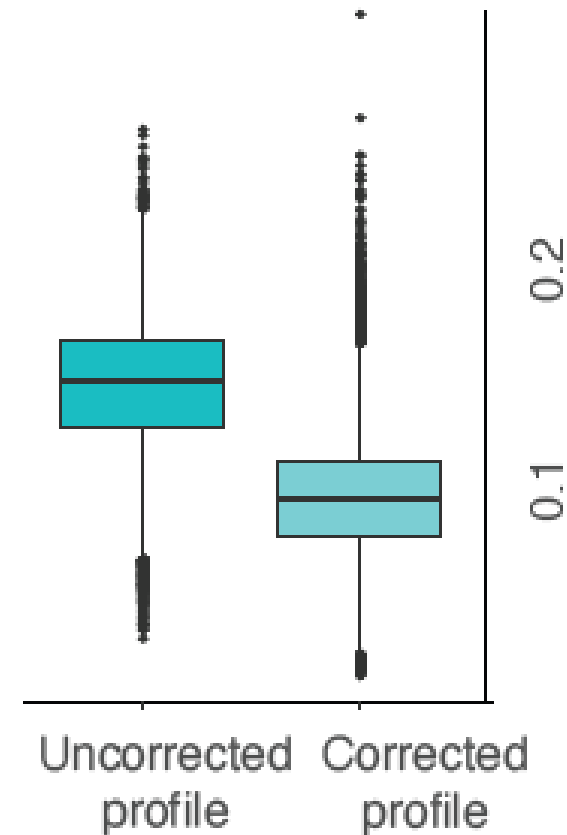


Bias correction reduces differences between DNase-seq & ATAC-seq profiles and contribution scores

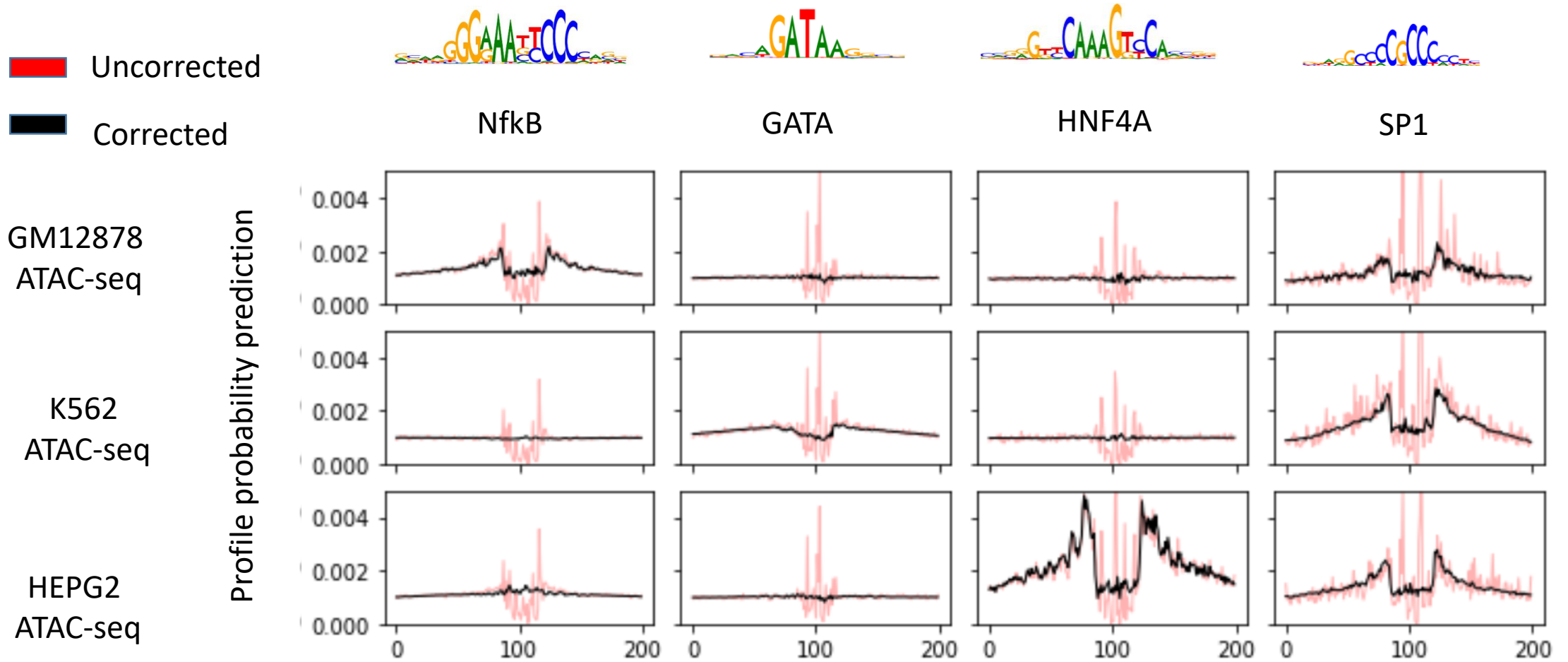
Jensen Shannon Distance between matched pairs of DNase-seq and ATAC-seq profiles



Jensen Shannon Distance between matched pairs of DeepLIFT score profiles from DNase-seq and ATAC-seq models



ChromBPNet dramatically improves cell-type specificity of TF footprints





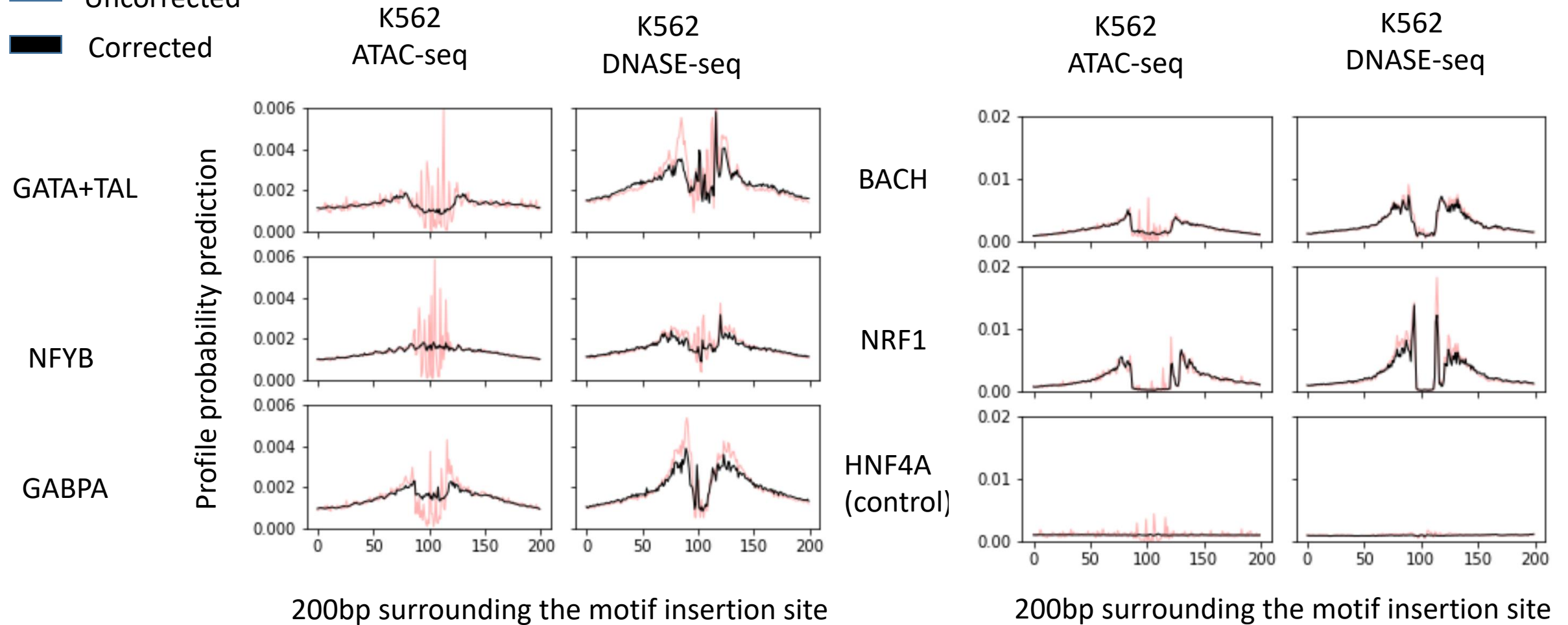
200bp surrounding the motif insertion site in 10K random non-peak sequences



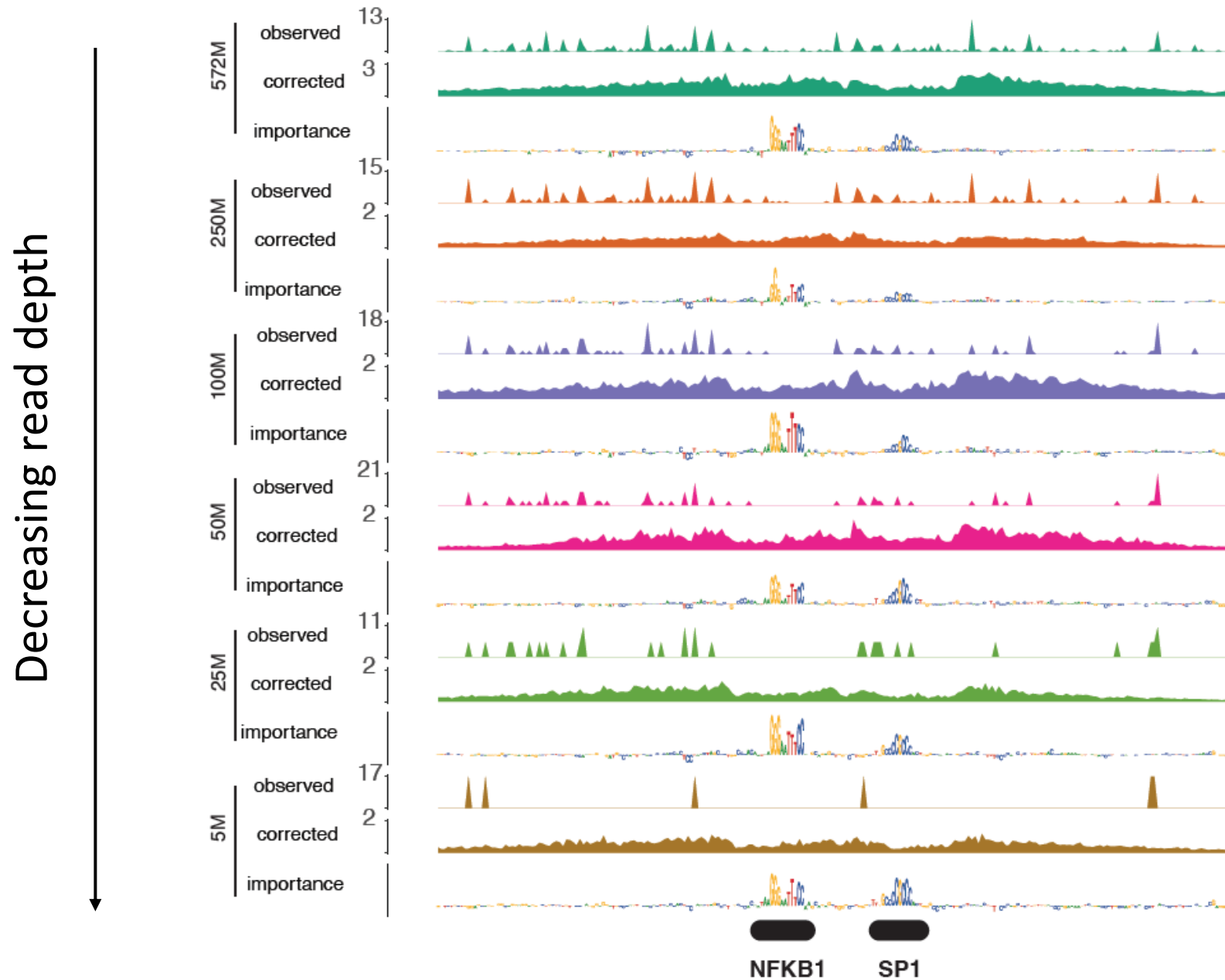
Anusri Pampari

ChromBPNet allows systematic comparison of Dnase-seq & ATAC-seq footprints

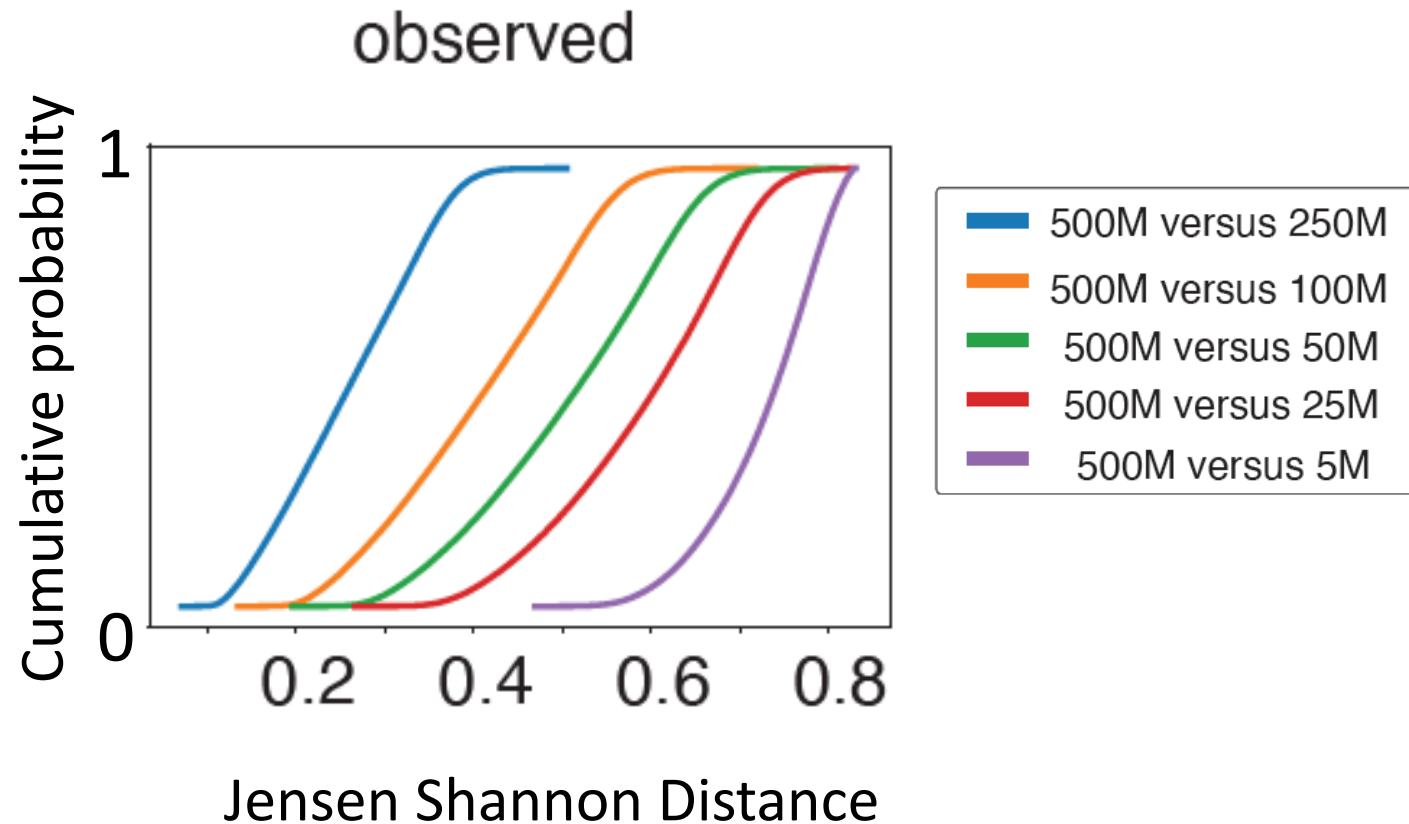
 Uncorrected
 Corrected



High fidelity denoising, imputation and interpretations at low read coverage

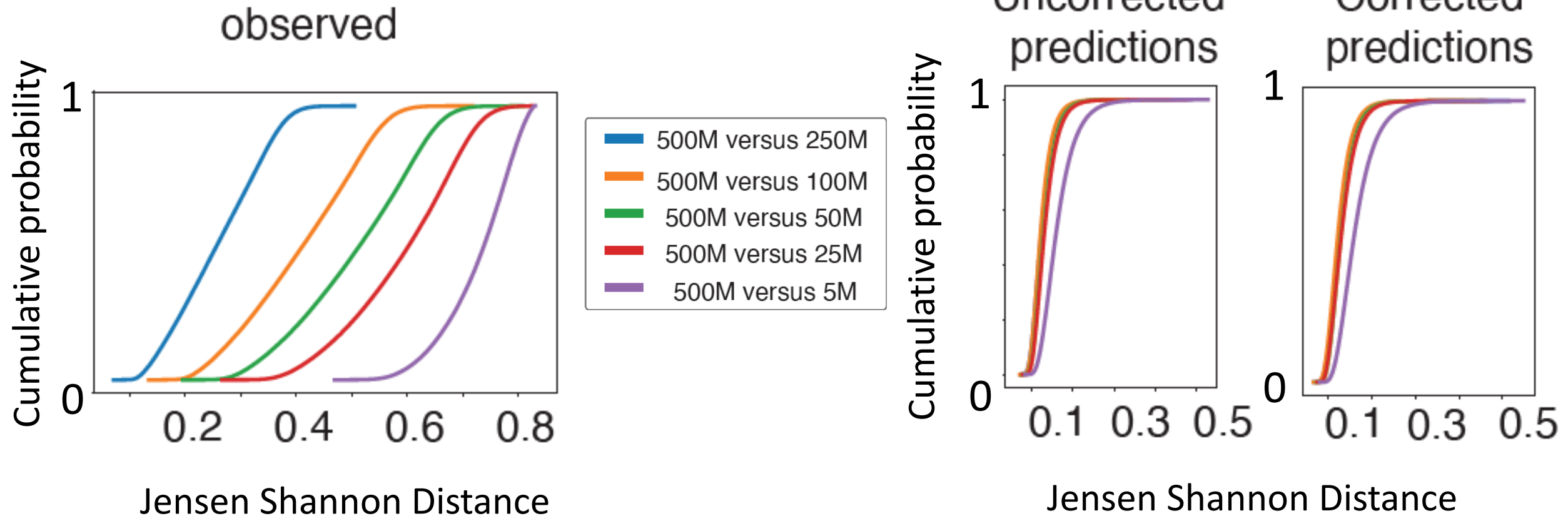


ChromBPNet accurately denoises and imputes signal from low coverage data



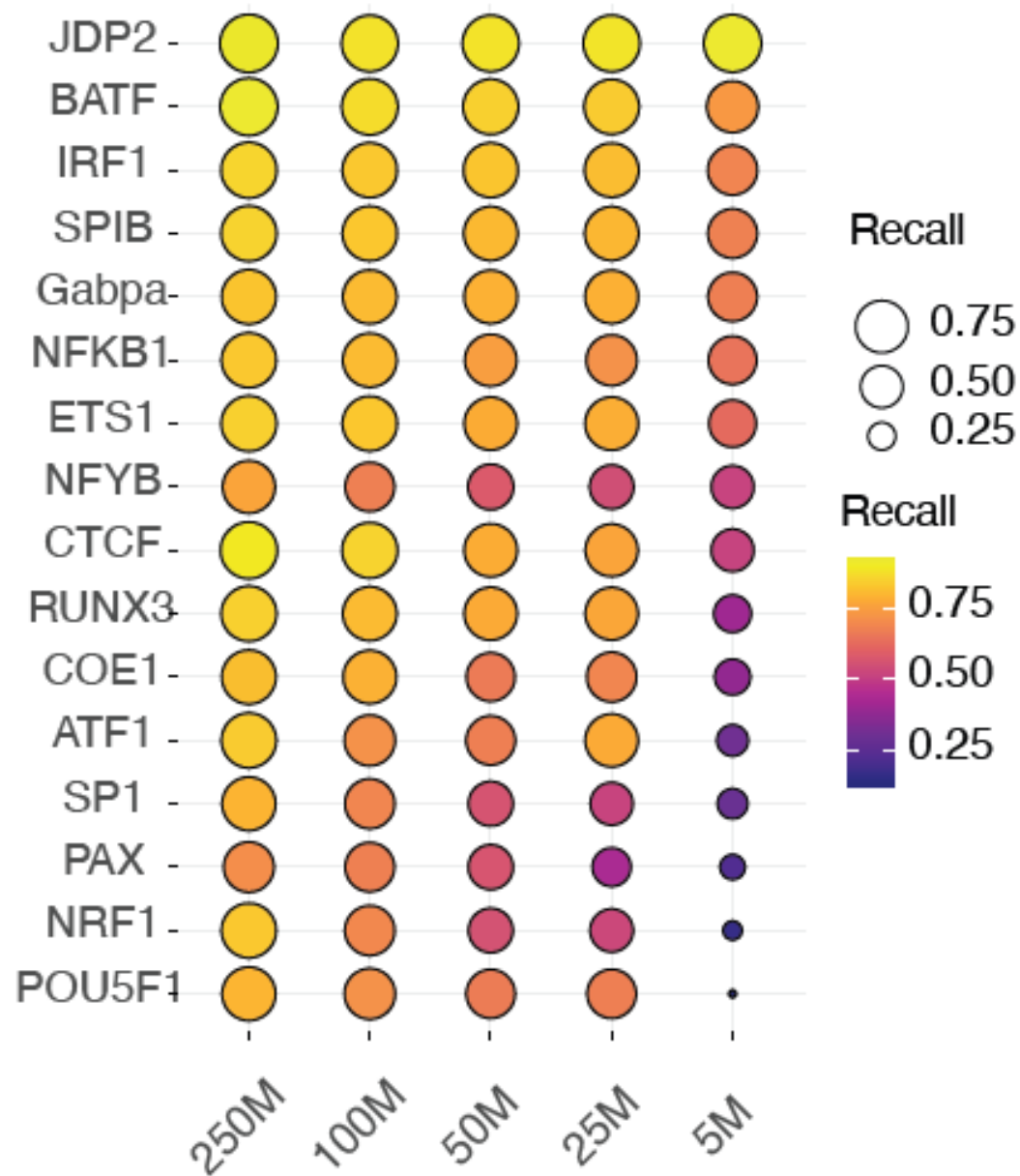
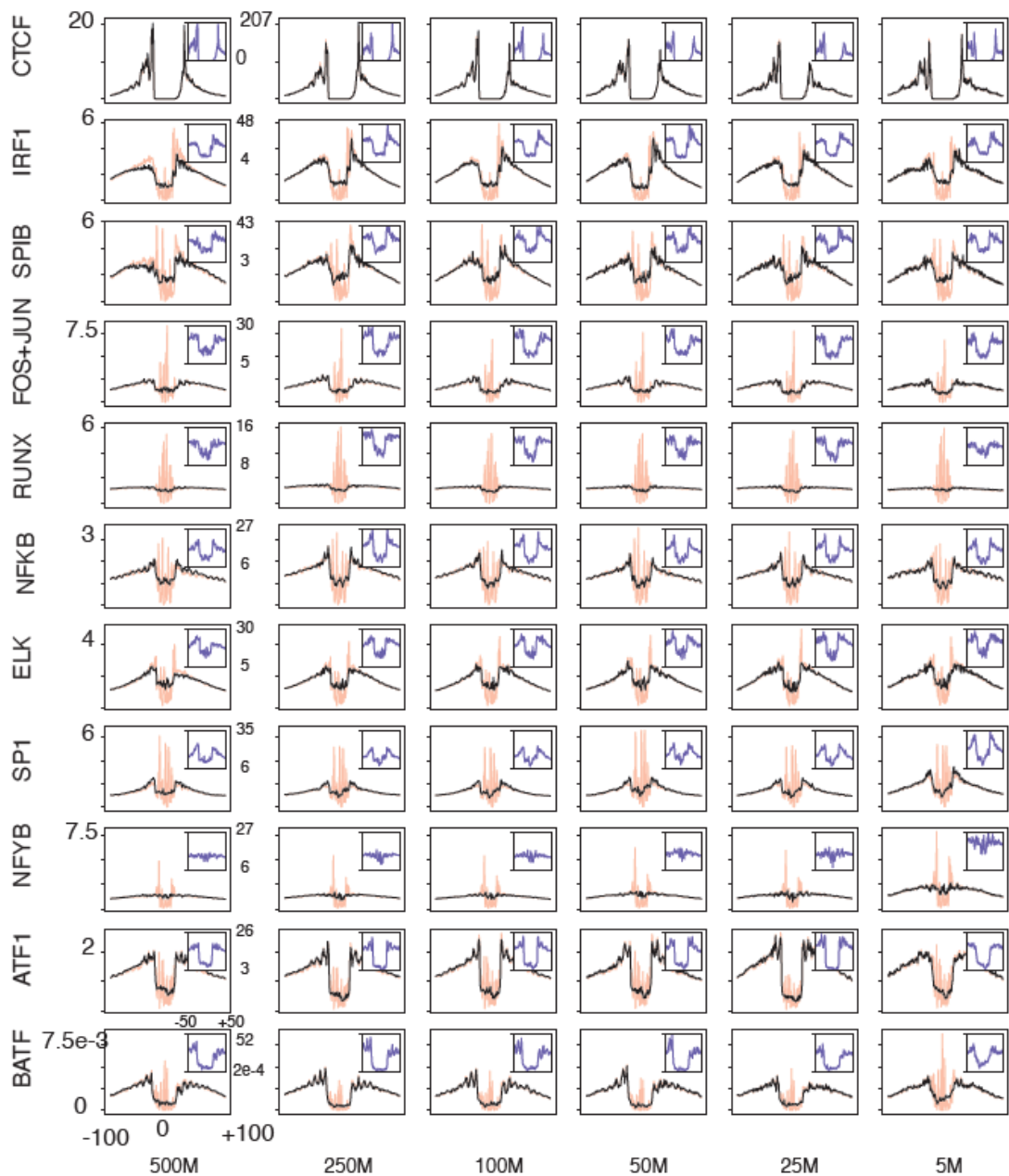
Using 500M as ground truth we compare degradation in observed and predicted signal profiles at different read depths

ChromBPNet accurately denoises and imputes signal from low coverage data



Using 500M as ground truth we compare degradation in observed and predicted signal profiles at different read depths

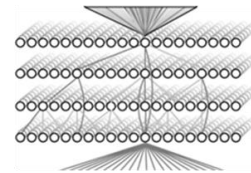
High fidelity marginal footprints & motif instance detection at low read depths



Screening genetic variants for regulatory effects

Variant effect screens
(Common, rare, SNVs, indels)

Δ PredictedSignal



.....ACTGAT **C**GCAATCG.....
.....ACTGAT **G**GCAATCG.....



Soumya Kundu



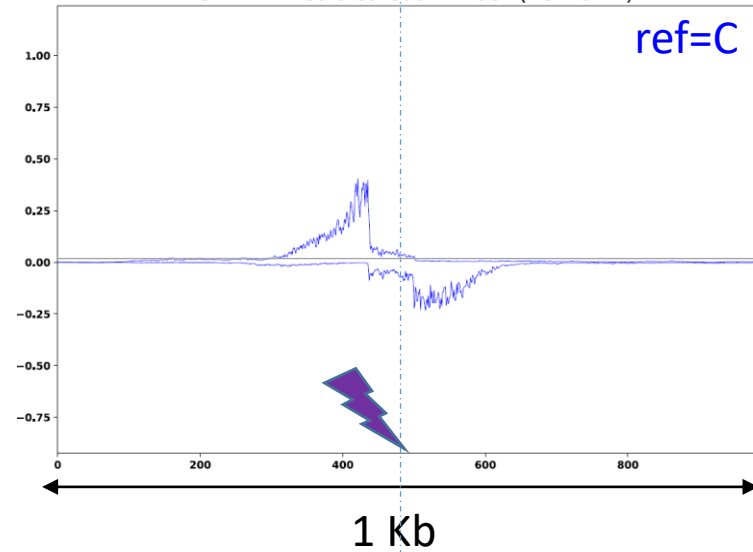
Lakshman
Sundaram



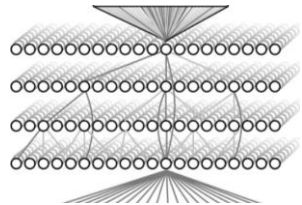
Ziwei Chen

In-silico mutagenesis: Predict effect of genetic variant on molecular activity

Predicted molecular profile of protein-DNA binding



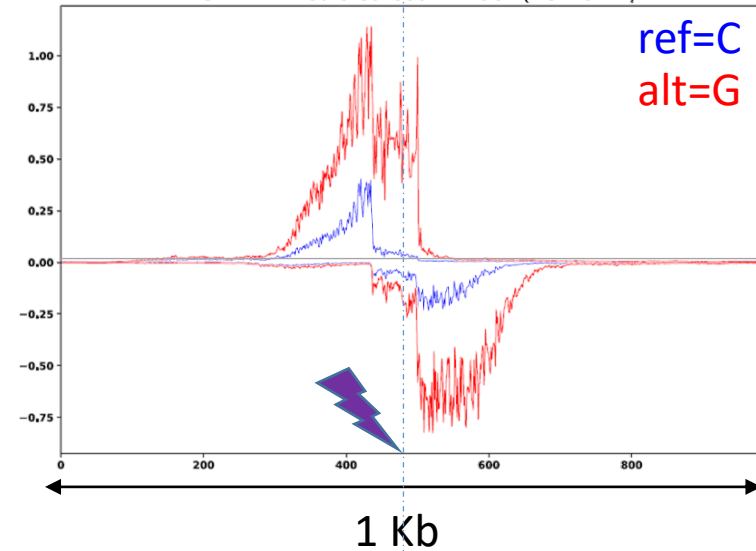
PredictedSignal



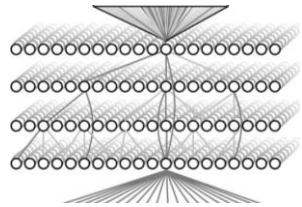
.....ACTGAT **C**GCAATCG.....

In-silico mutagenesis: Predict effect of genetic variant on molecular activity

Predicted molecular profile of protein-DNA binding



Δ PredictedSignal

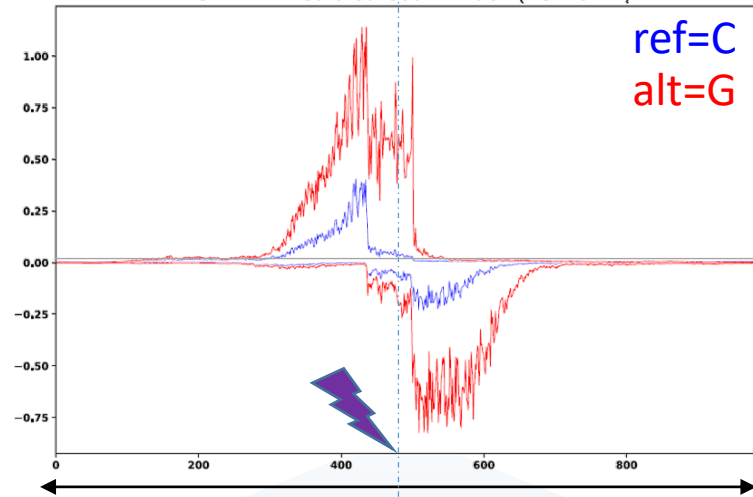


.....ACTGAT **C** GCAATCG.....

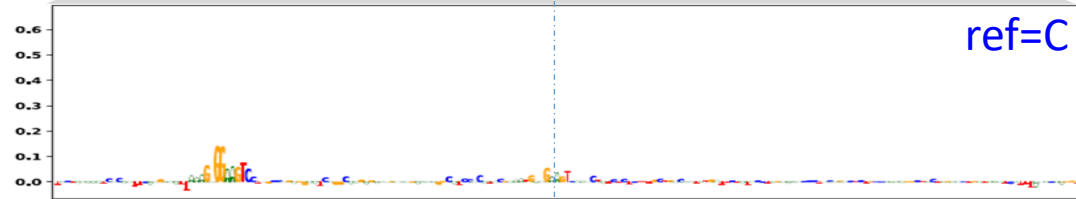
.....ACTGAT **G** GCAATCG.....

Interpret disrupted predictive sequence syntax

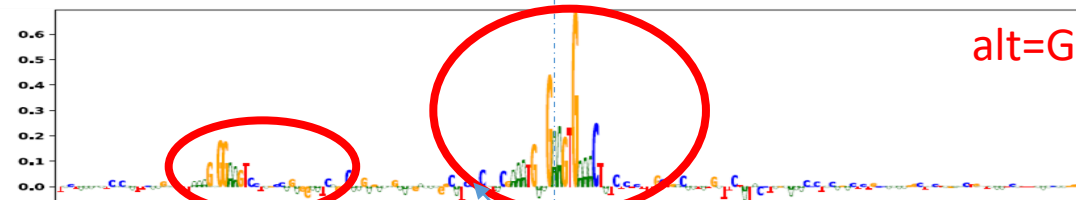
Predicted molecular profile of protein-DNA binding



1 Kb



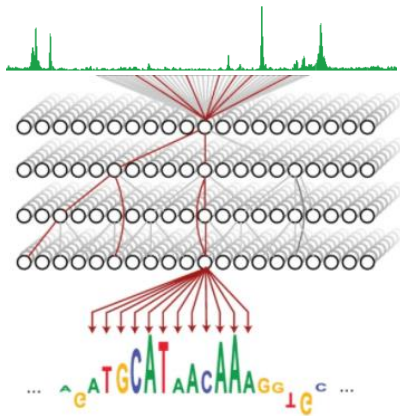
ref=C



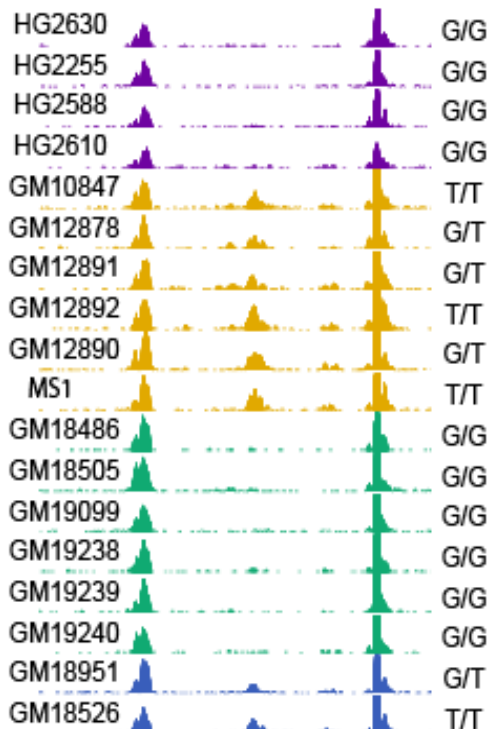
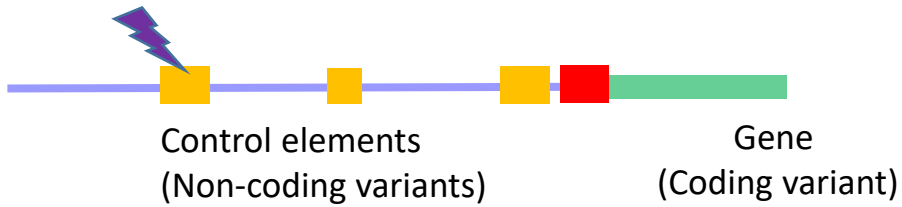
alt=G

200 bp

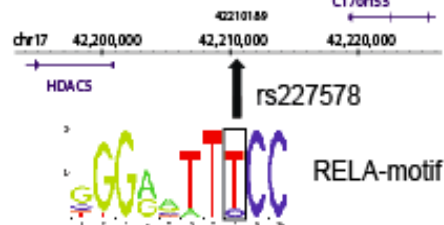
Sequence binding motifs of SPI1 DNA binding protein



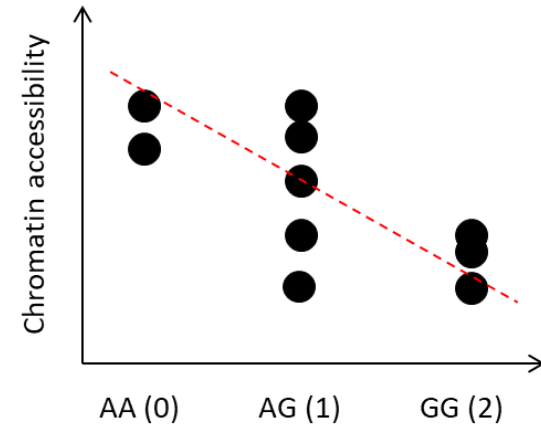
Molecular quantitative trait loci (QTLs): Identifying genetic variation associated with variation in molecular activity (chromatin, expression etc)



Genetic variation



1. Sequence genomes of 100-1000s of individuals
2. Obtain tissue of interest from ALL these individuals
3. Perform molecular profiling experiments in ALL individuals
4. Perform statistical association of each variant with variation of molecular activity (expression, chromatin accessibility) of each element (gene, regulatory element)
5. Output: QTLs = Variants with statistically significant association with molecular activity



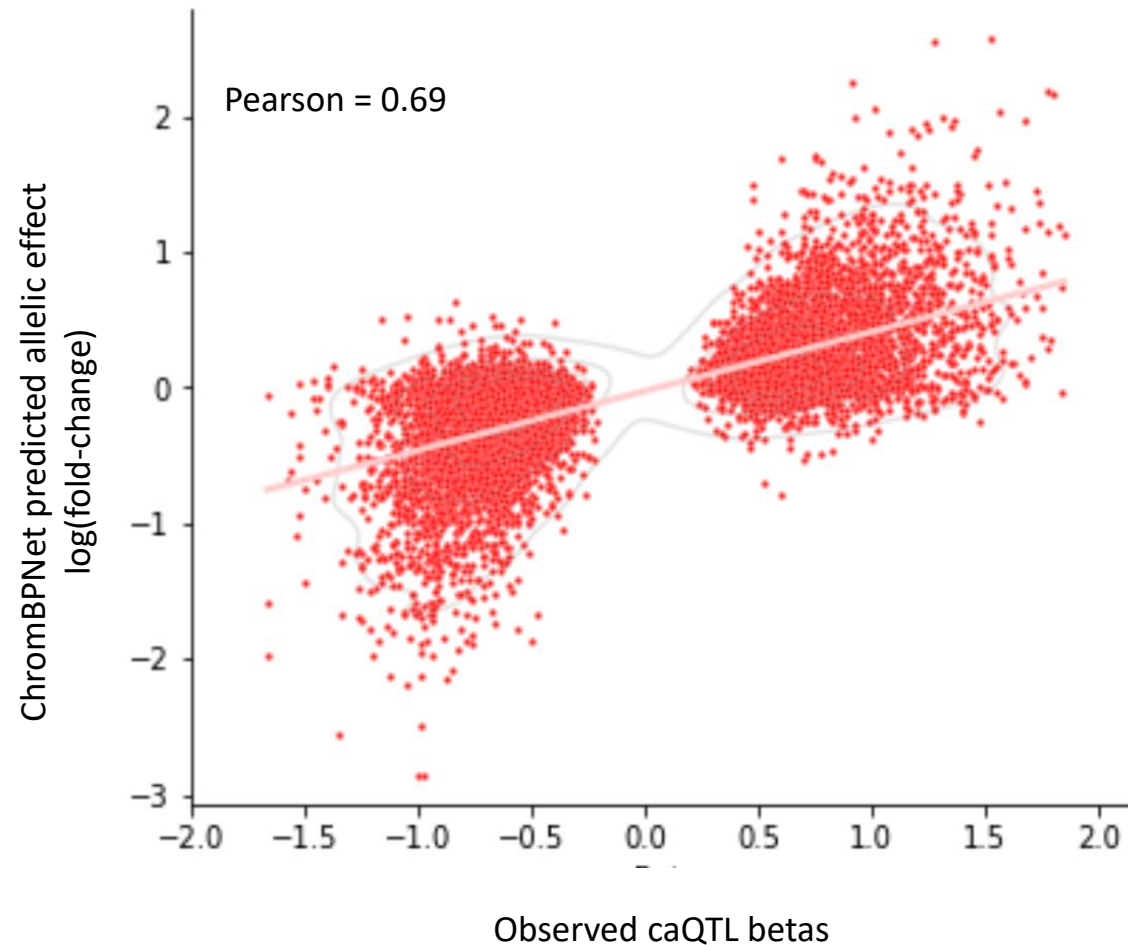
$$Y = \beta X + \epsilon$$

Limitations of this approach:

1. Very cumbersome and expensive for each tissue / cell type
2. Difficult to access some tissues / cell types in 100s of individuals

Prediction of effect sizes of variants measured in diverse African cohort using model trained on a European reference dataset

11,098 caQTLs in LCLs from diverse African populations

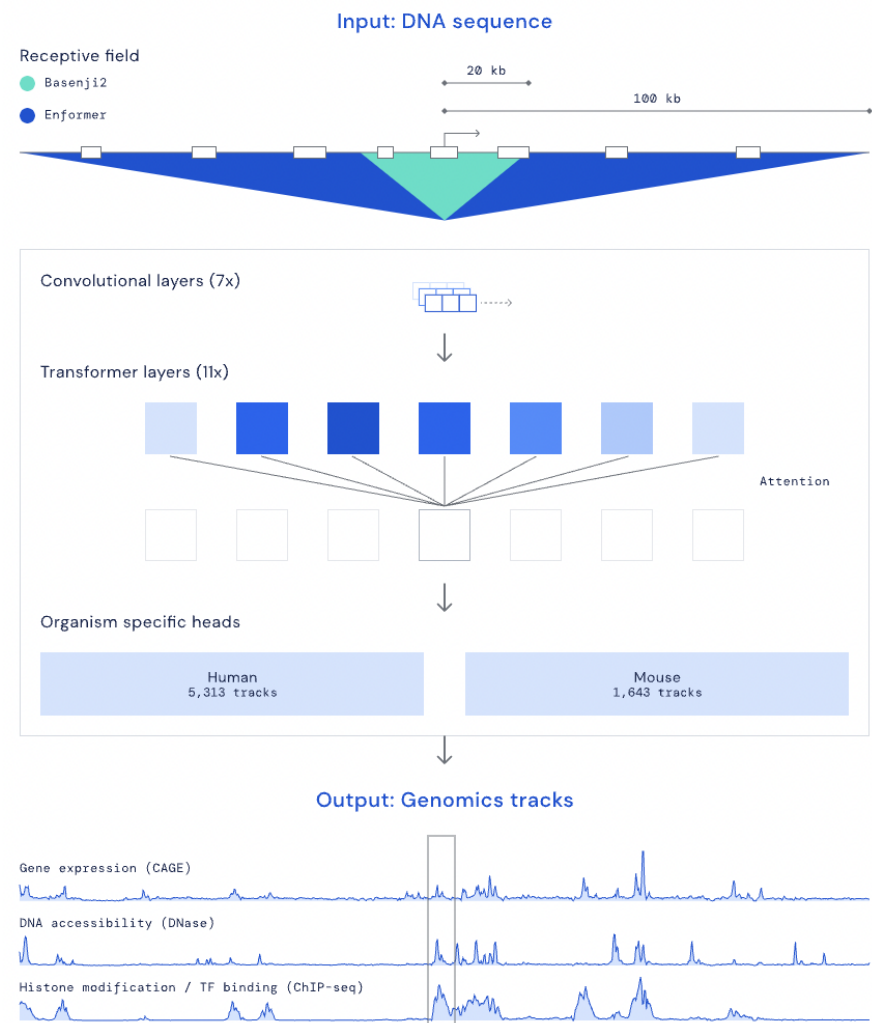
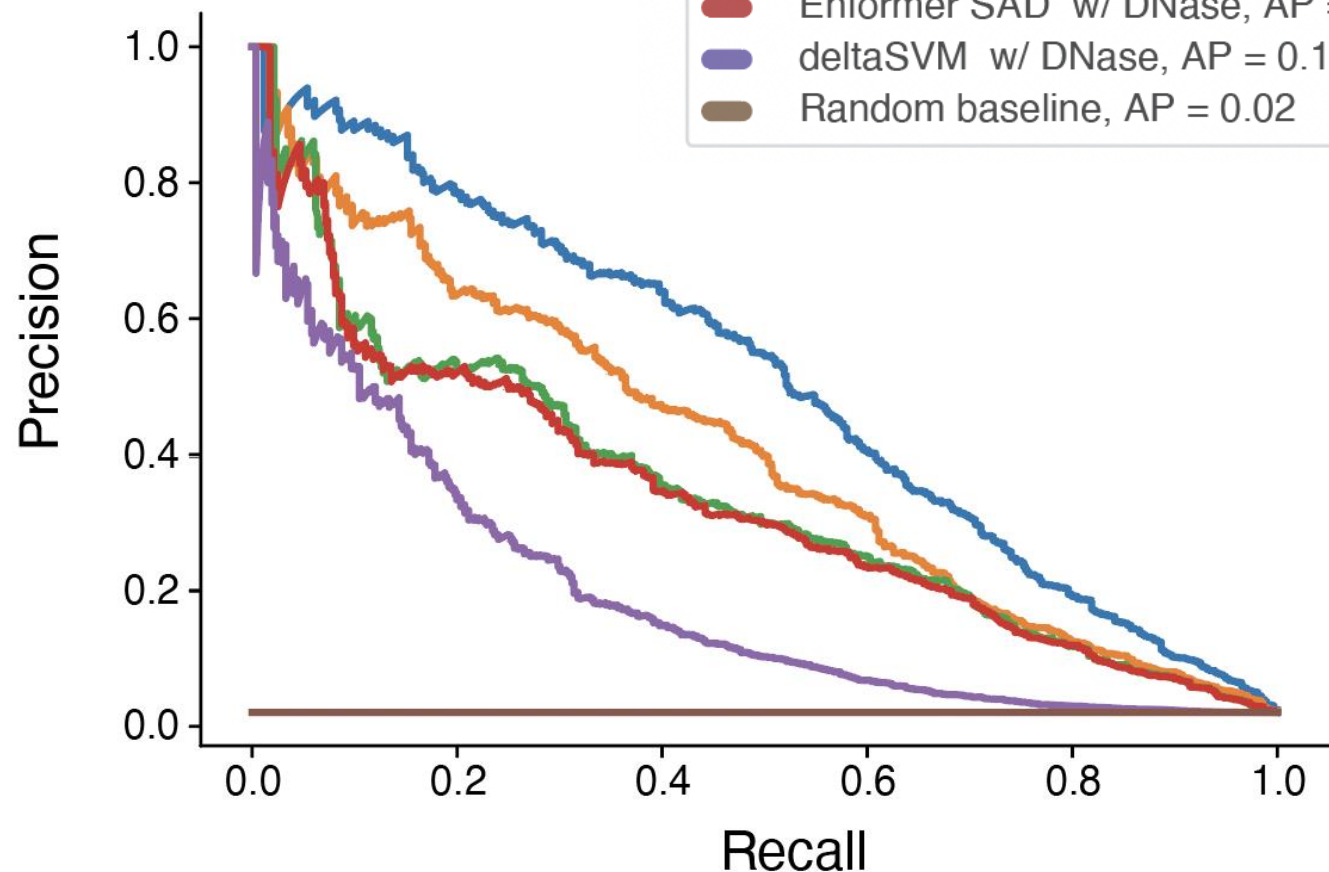


ChromBPNet model trained on a single reference ATAC-seq dataset of European ancestry

ChromBPNet outperforms other models for predicting variants affecting chromatin accessibility

Average precision (AP) w/
predicted effect sizes from

- ChromBPNet w/ ATAC, AP = 0.51
- ChromBPNet w/ DNase, AP = 0.41
- Enformer SAR w/ DNase, AP = 0.34
- Enformer SAD w/ DNase, AP = 0.33
- deltaSVM w/ DNase, AP = 0.19
- Random baseline, AP = 0.02



deltaSVM: Lee et al. 2015
Enformer: Avsec et al. 2021

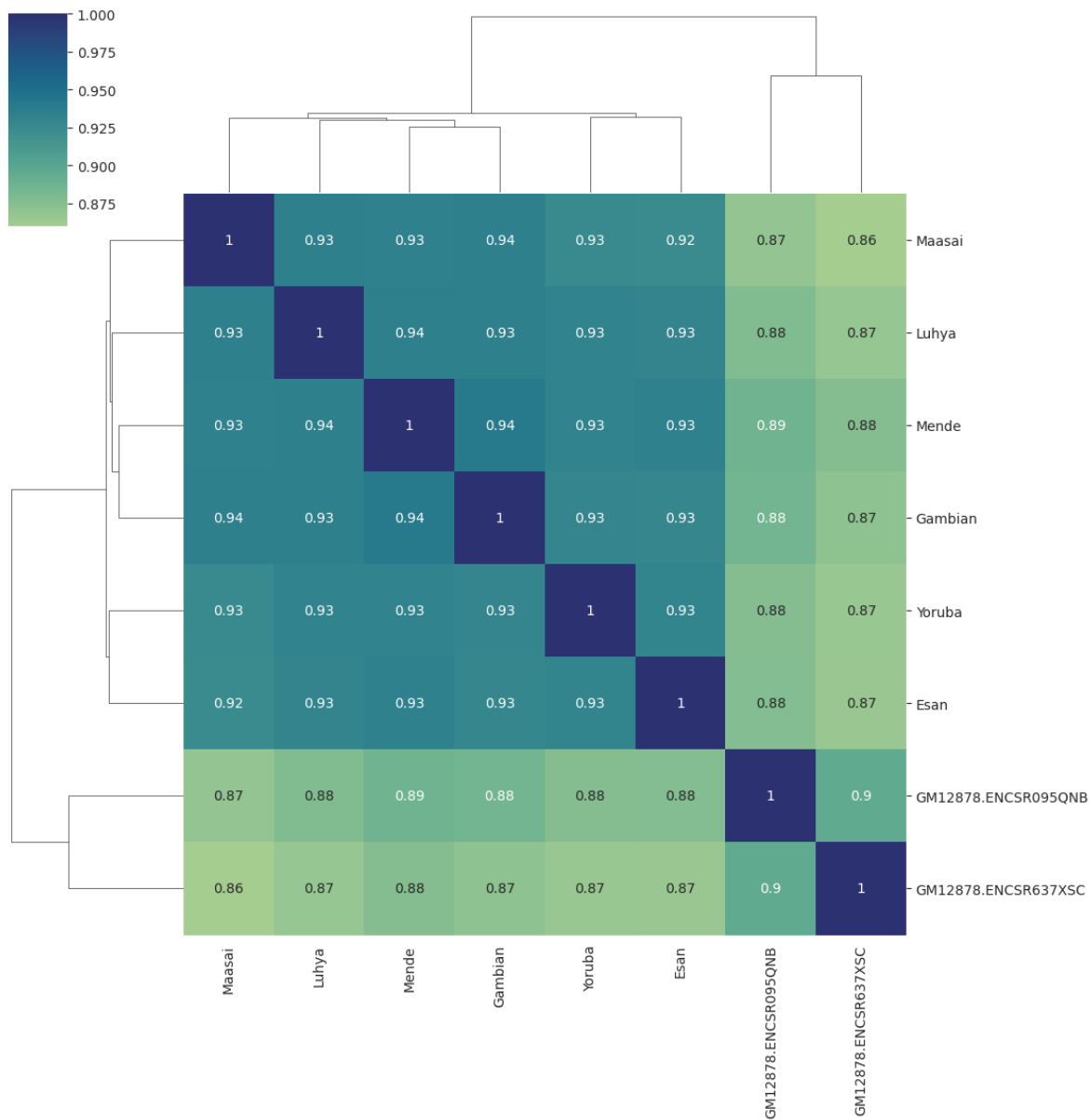
ChromBPNet substantially outperforms “DNA language models” (zero-shot, probed and fine-tuned) for variant effect prediction

Dataset	Model	Zero-Shot AUROC		Probed		Fine-tuned		<i>Ab initio</i>	
		Likelihood	Embedding	Pearson r	AUROC	Pearson r	AUROC	Pearson r	AUROC
African	DNABERT-2	-	0.480	-0.003	0.482	0.184	0.616	-	-
	GENA-LM	-	0.508	0.005	<u>0.508</u>	0.201	0.604	-	-
	HyenaDNA	0.486	0.515	0.036	0.488	<u>0.265</u>	0.611	-	-
	NT	<u>0.525</u>	<u>0.519</u>	<u>0.041</u>	0.503	0.230	<u>0.623</u>	-	-
	ChromBPNet	-	-	-	-	-	-	0.671	0.772
Yoruban	DNABERT-2	-	0.505	<u>0.102</u>	0.476	0.473	0.631	-	-
	GENA-LM	-	0.501	-0.052	0.397	0.414	0.628	-	-
	HyenaDNA	0.436	0.515	0.022	0.466	0.503	0.573	-	-
	NT	<u>0.469</u>	<u>0.613</u>	0.055	<u>0.555</u>	<u>0.507</u>	<u>0.670</u>	-	-
	ChromBPNet	-	-	-	-	-	-	0.738	0.892

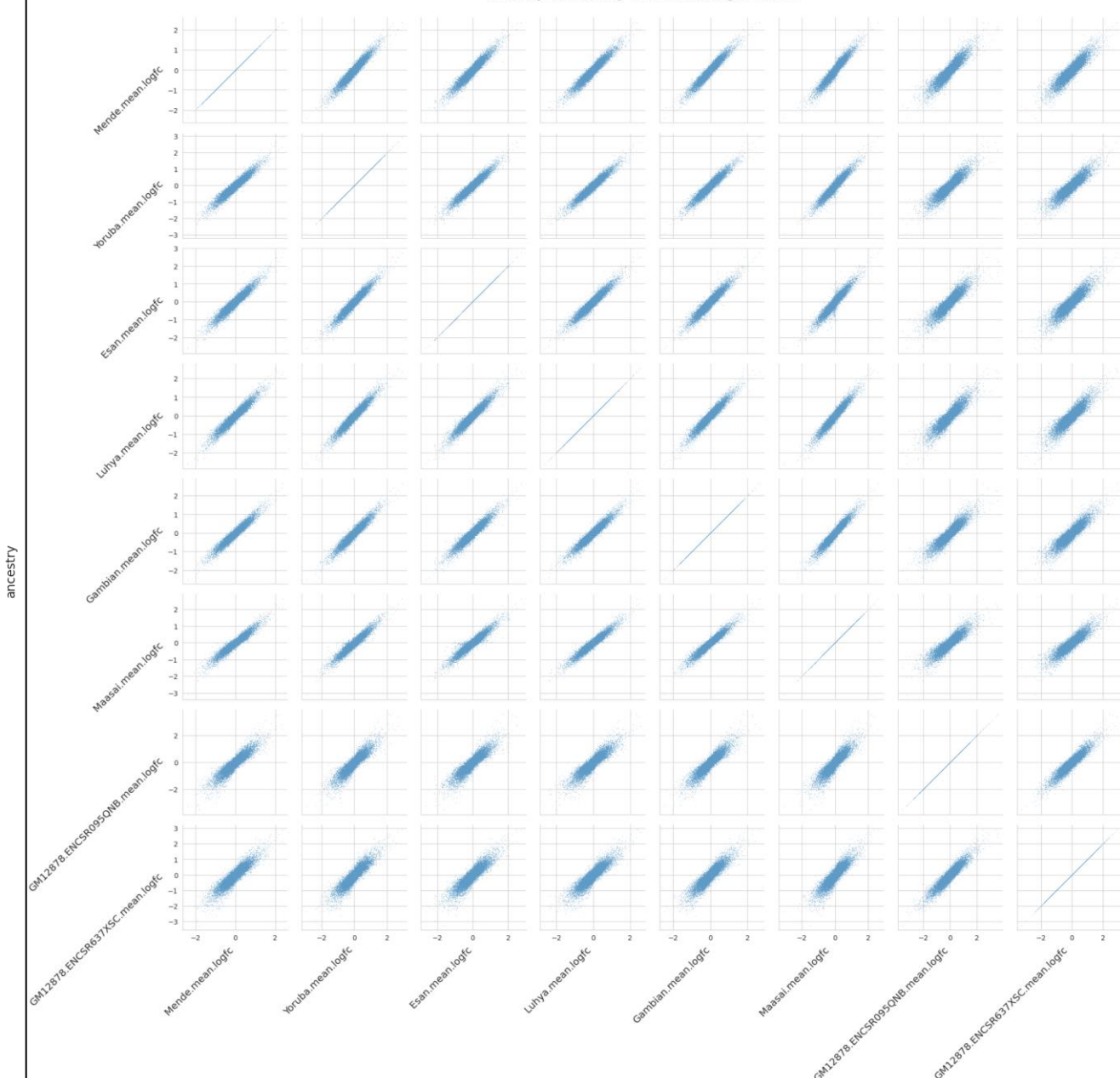
DNA-LMs are probed and fine tuned to predict chromatin accessibility profiles genome-wide (exactly the same data that ChromBPNet is trained on)

ChromBPNet models trained on European LCL ATAC-seq reference generalizes to African caQTLs

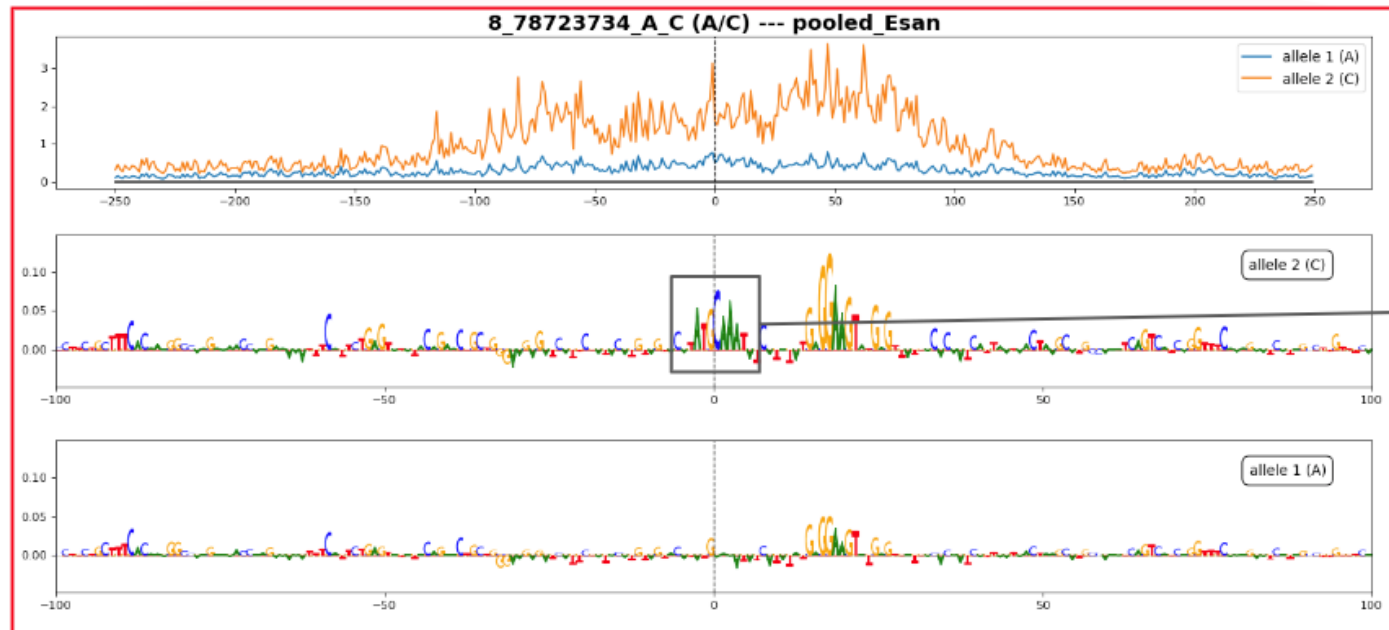
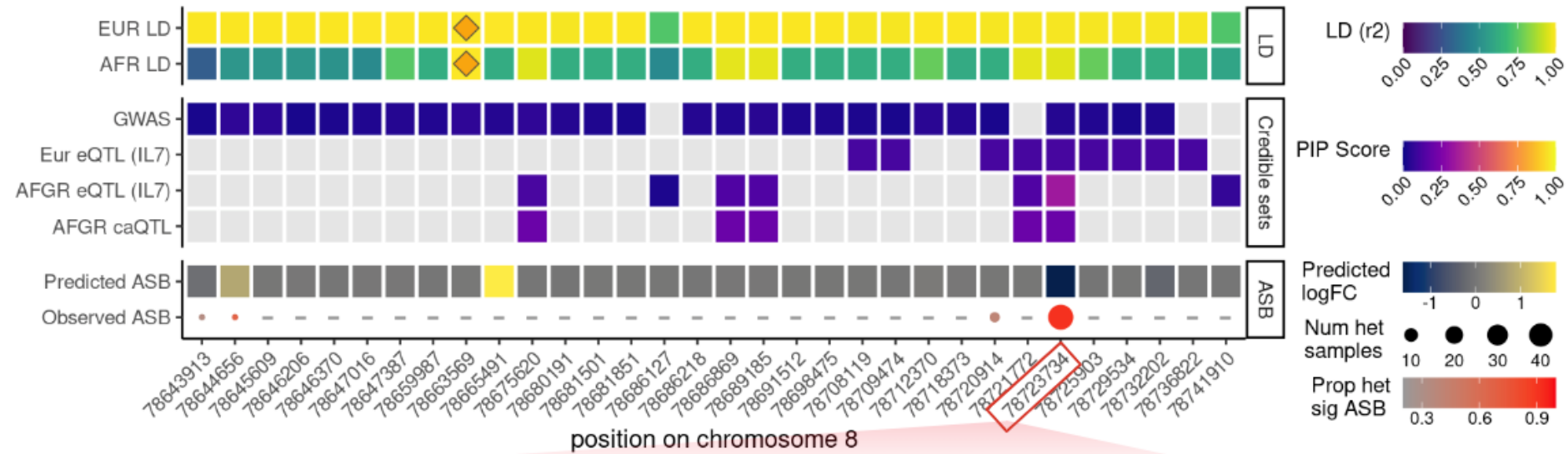
Ancestry vs. Ancestry ChromBPNet LogFC Pearson Correlation



Ancestry vs. Ancestry ChromBPNet LogFC Scores

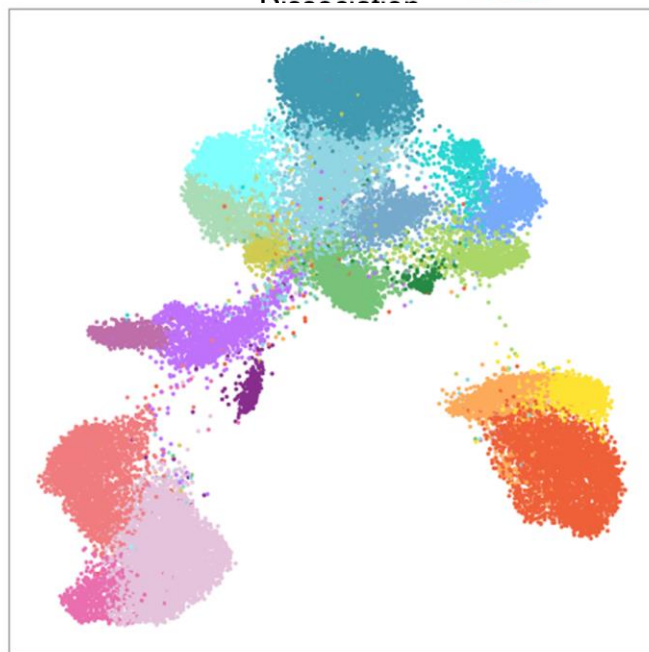
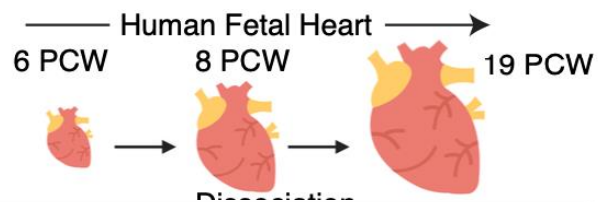


Fine mapping functional variant in multiple sclerosis GWAS locus (IL7)



**OCT family
binding site motif**

Predicting *de-novo* non-coding variants in congenital heart disease from fetal heart scATAC-seq



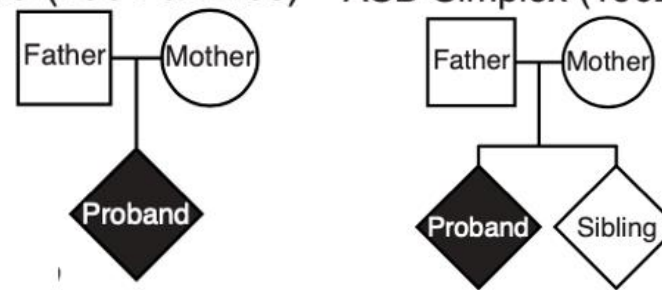
- Myocardium
- Atrial Cardiomyocytes
- Ventricular Cardiomyocytes
- Early Cardiac Fibroblast
- Cardiac Fibroblast Progenitors
- Cardiac Fibroblast
- Endocardial Cushion
- Late Endocardial Cushion
- OFT SMC
- Vasculature Development
- vSMC
- Pericytes
- Neural Crest
- Undifferentiated Epicardium
- Endocardium
- Transitioning Endocardium
- Lymph Endothelium
- Arterial Endothelium
- Capillary Endothelium
- Venous Endothelium



☞ Lakshman Sundaram (CS)

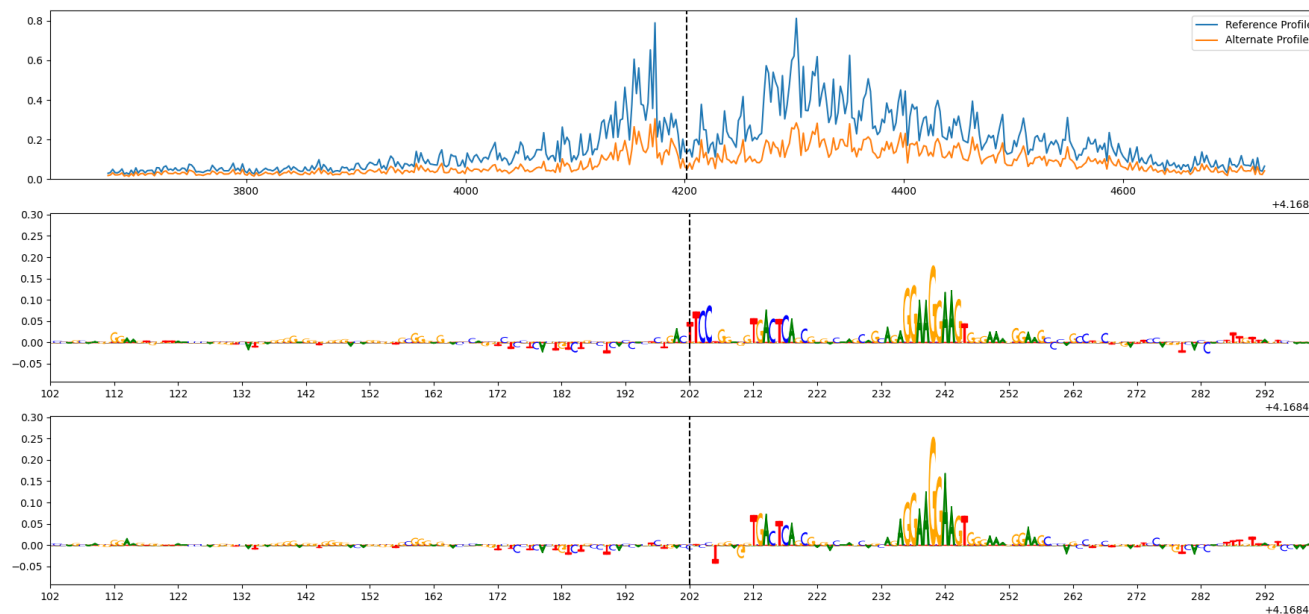
Ameen et al. 2022, Cell

PCGC (750 Families) ASD Simplex (1902 Families)



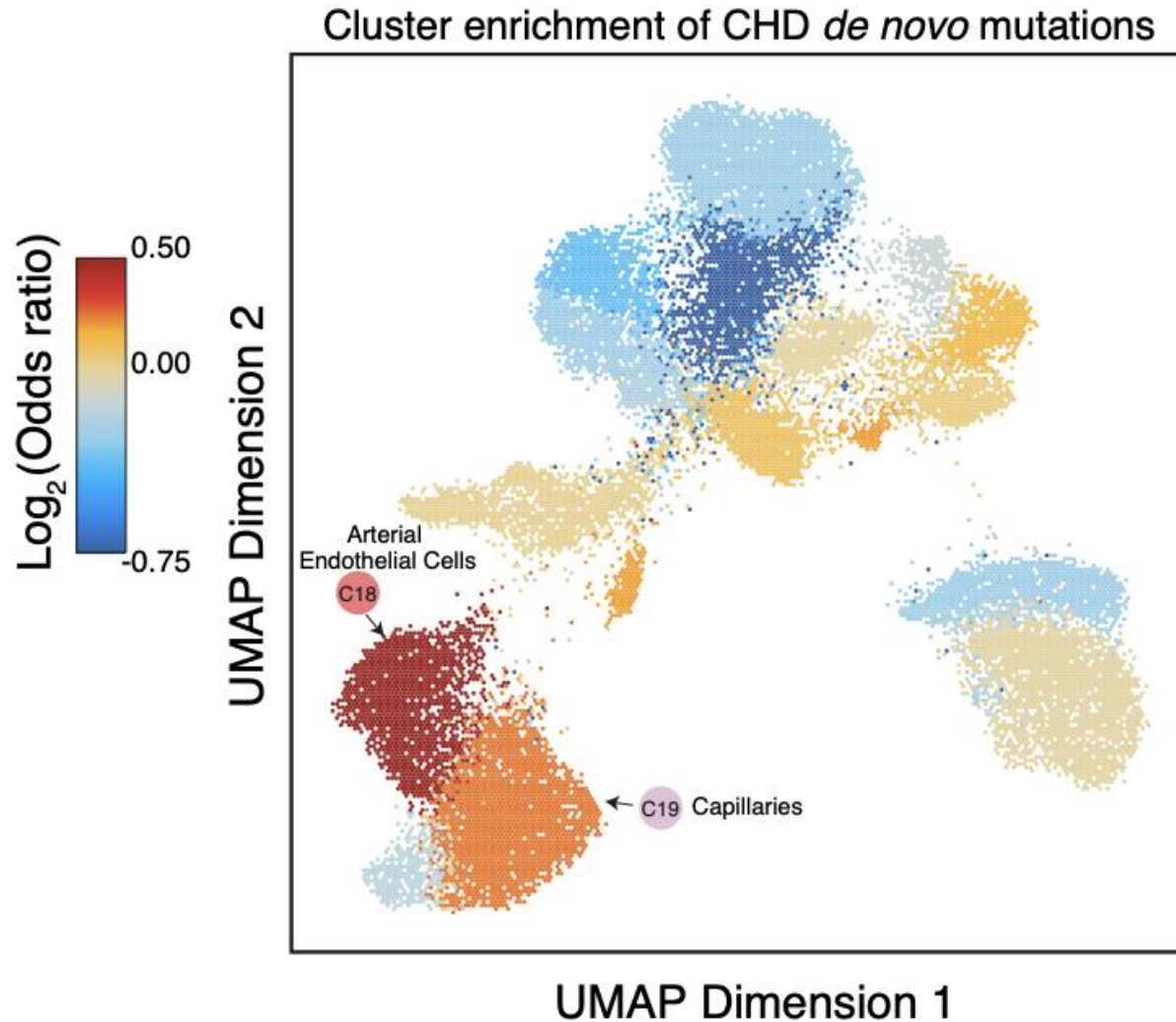
Cases

Controls



Mutation disrupts an ETS/ELK/ETV family motif

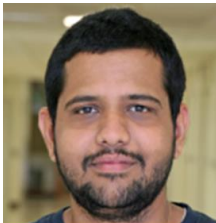
Arterial endothelial cells are enriched for prioritized *de novo* CHD non-coding mutations



Cell types ranked for disease enrichment

Provides a window into developmental timepoints critical for Congenital heart disease

Arteries and Capillaries the most enriched cell types

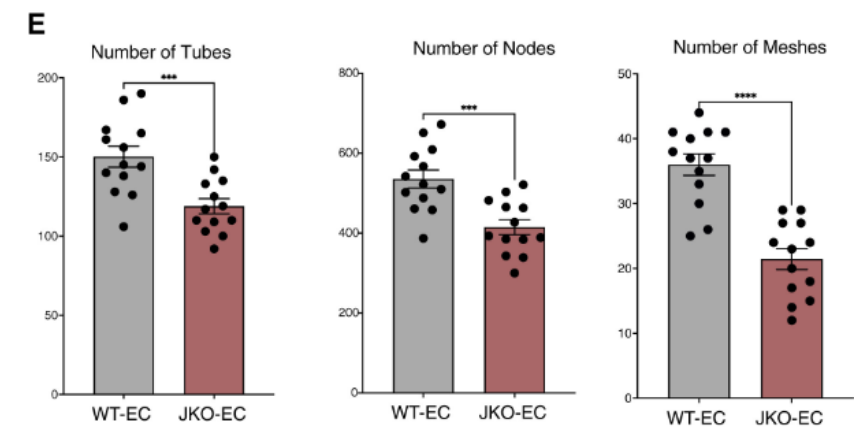
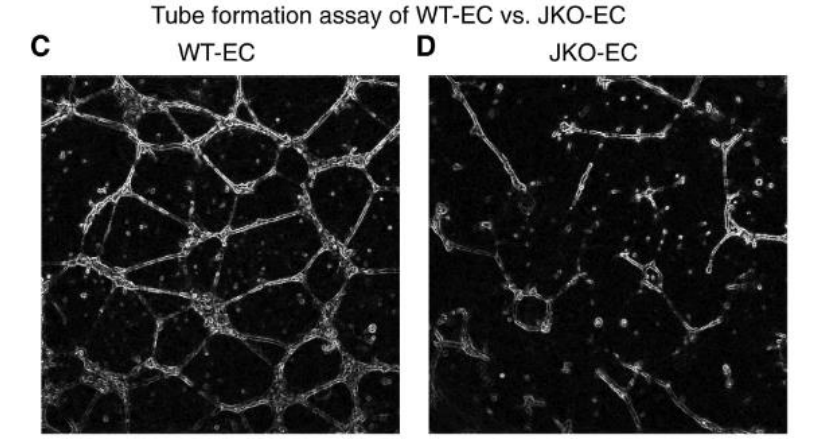
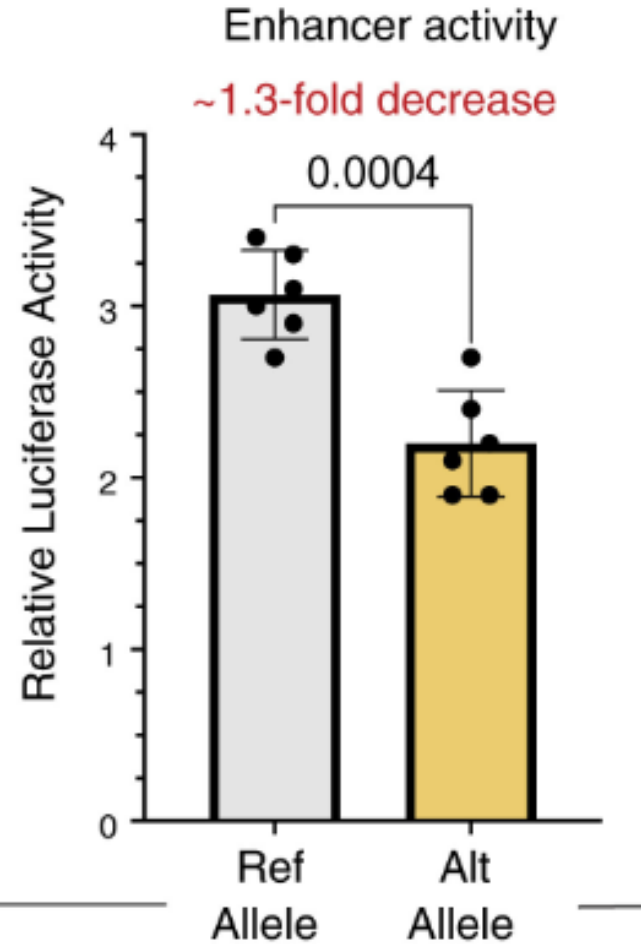
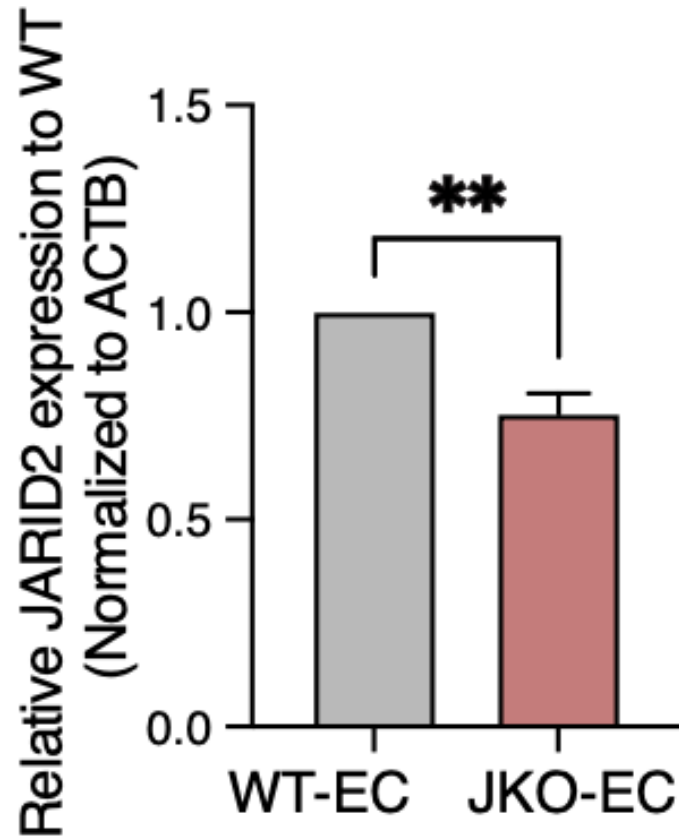


Lakshman Sundaram

CRISPR, Luciferase, phenotypic assay supports prioritized CHD variant in the JARID2 locus in aECs



Mo Ameen



Summary

- Neural networks can accurately model cell context specific regulatory profiles at base-res. from DNA sequence
- Models can be interpreted to decipher complex sequence syntax
- Assay biases can be detected, learned & corrected improving concordance between related assays, imputing missing signal from sparse profiles and reveal causal biological sequence drivers of activity
- Models can predict & interpret counterfactual effects of regulatory genetic variants
- Scale is not everything: Small models can outperform massive models

Acknowledgements

Kundaje lab



ENCODE

Stam lab
Greenleaf lab
Snyder lab
Tehwey lab
Reddy lab
Shendure lab
Ahituv lab

Funding

1R01HG009674

1U01HG009431

1U24HG009446



R01ES02500902



1DP2OD022870



Ziga Avsec



Julia Zeitlinger

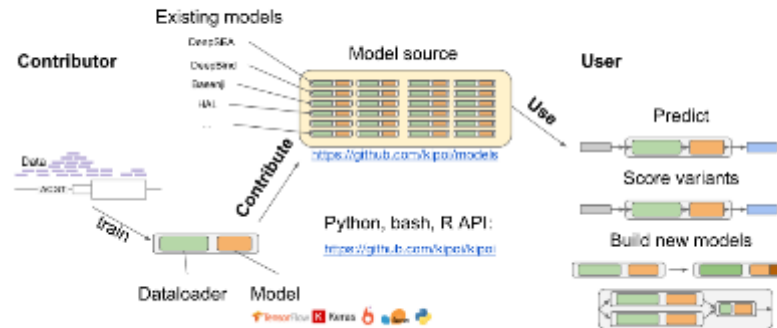
Democratizing ML for genomics: <http://kipoi.org/>



Ziga Avsec



Kipoi: Model zoo for genomics



Kipoi (pronounce: kίpi; from the Greek κίπτοι: gardens) is an **API** and a **repository** of ready-to-use trained models for regulatory genomics. It currently contains 1709 different models, covering canonical predictive tasks in transcriptional and post-transcriptional gene regulation. Kipoi's API is implemented as a python package (github.com/kipoi/kipoi) and it is also accessible from the command line or R.

Numbers

of models: 1709

of model groups: 16

of contributors: 6

of model groups supporting postprocessing:

- Variant effect prediction: 11/16

Model groups by tag

