STANFORD
COMPUTER SCIENCE

STANFORD | Genetics
SCHOOL of MEDICINE

Deep learning oracles for genomic discovery

Kundaje lab

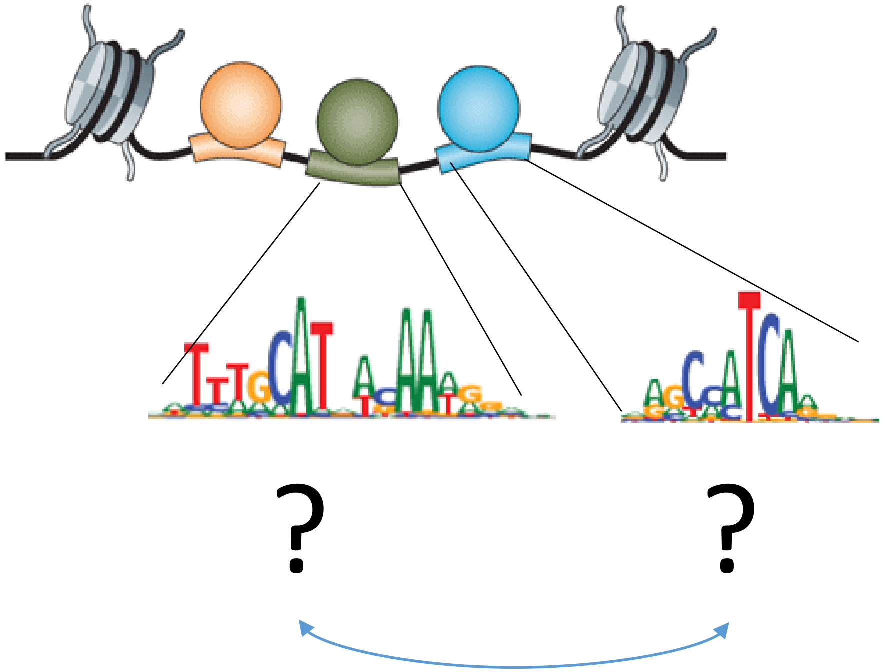
Genetics, Computer Science

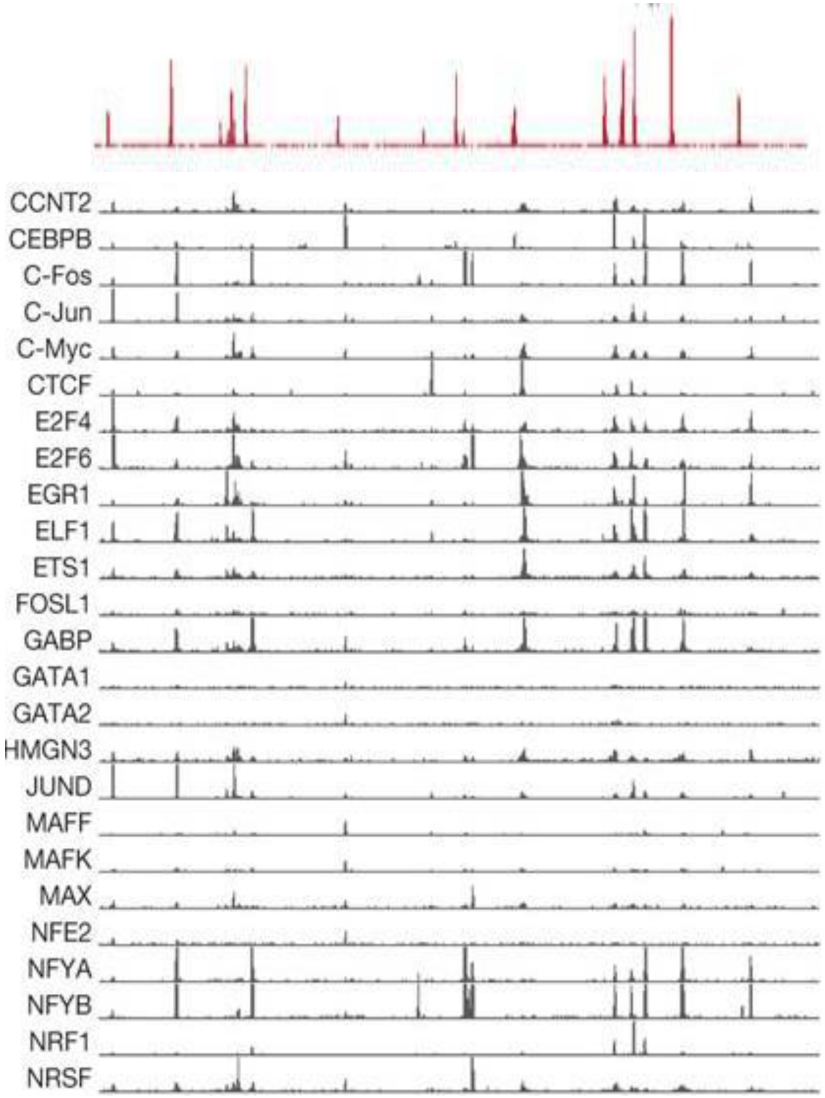
Stanford University

http://anshul.kundaje.net

Deciphering functional DNA words and their syntax in regulatory DNA

Syntax: Rules of arrangement, preferred spacing, orientation, interactions between works





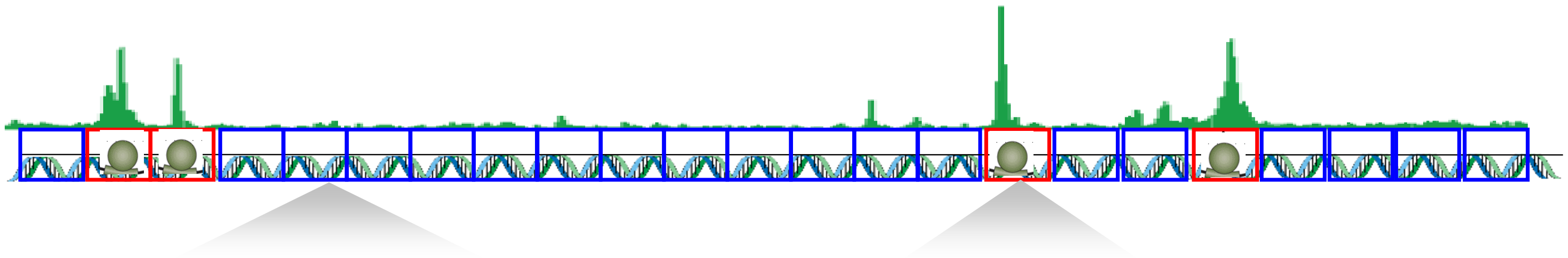
chromatin accessibility (ATAC-seq / DNase-seq)

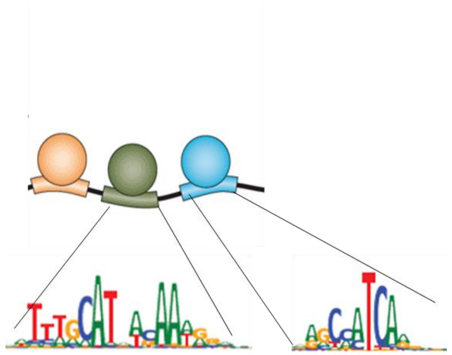
Protein-DNA binding maps (ChIP-seq, ChIP-exo)

Adapted from Thurman et al 2012

Predictive model of regulatory DNA

Genome-wide protein-DNA binding map





...GACTTGAAACGGCATTG...

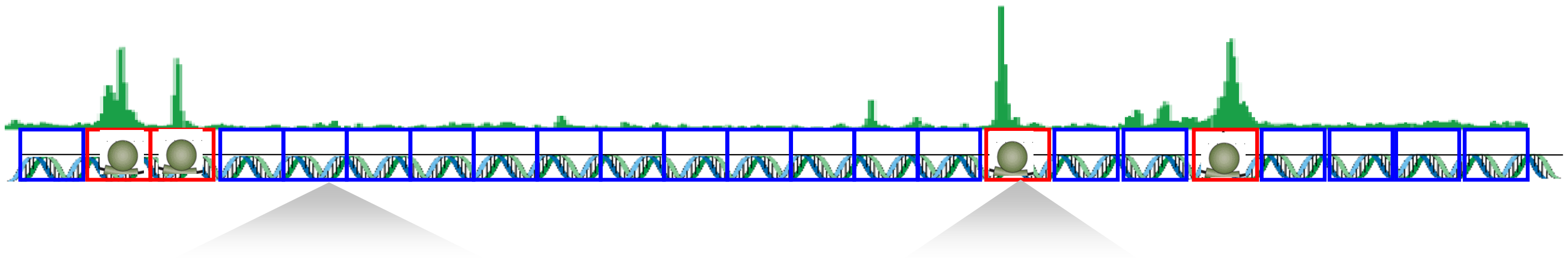
Inactive (0) (0.3)

...GACAGATAATGCATTGA...

Active (+1) (20.2)

Predictive model of regulatory DNA

Genome-wide protein-DNA binding map

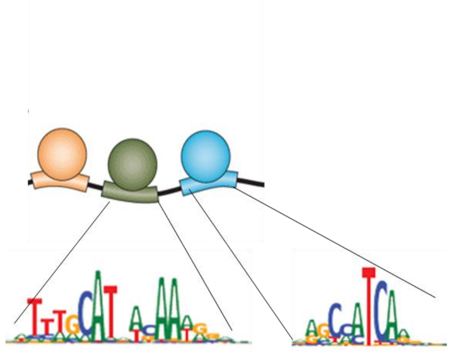


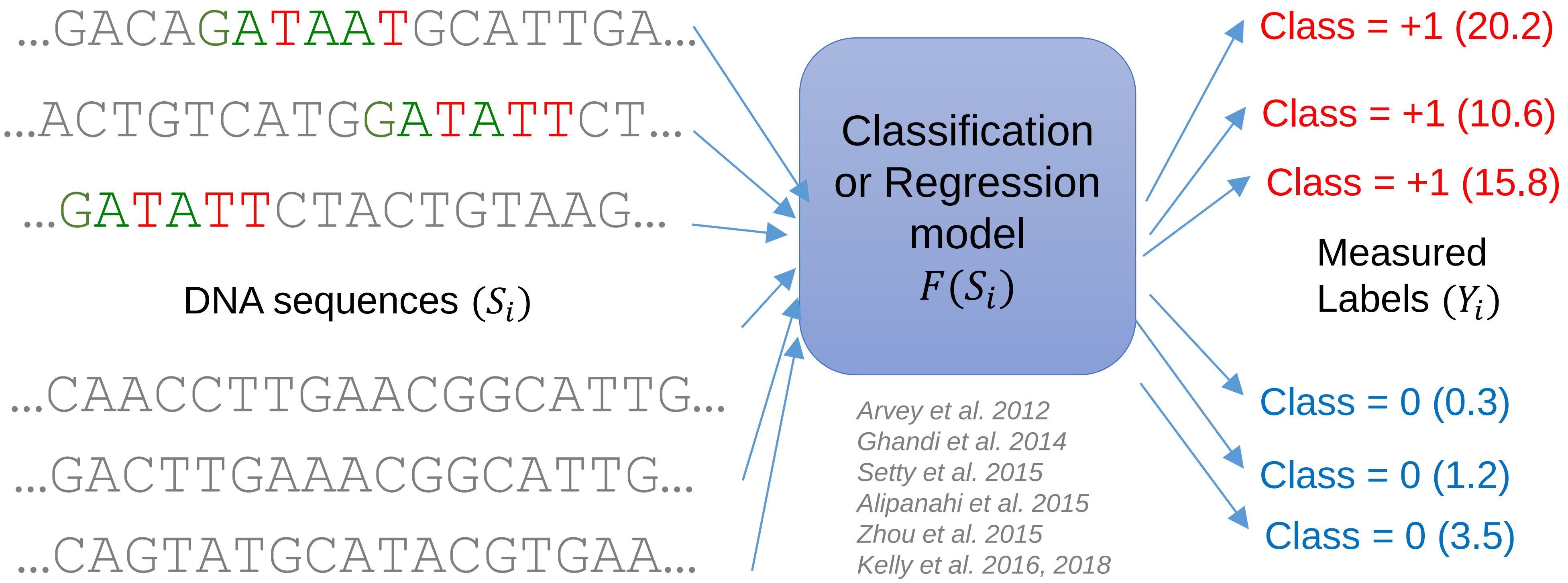
...GACTTGAAACGGCATTG...

Inactive (0) (0.3)

...GACAGATAATGCATTGA...

Active (+1) (20.2)





...GACAGATAATGCATTGA...

...ACTGTCATGGATATTCT...

...GATATTCTACTGTAAG...

DNA sequences (S_i)

...CAACCTTGAACGGCATTG...

...GACTTGAAACGGCATTG...

...CAGTATGCATACGTGAA...

Classification or Regression model $F(S_i)$

Class = +1 (20.2)

Class = +1 (10.6)

Class = +1 (15.8)

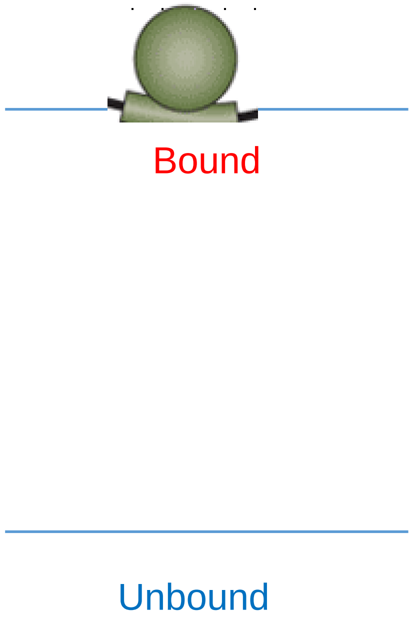
Measured Labels (Y_i)

Class = 0 (0.3)

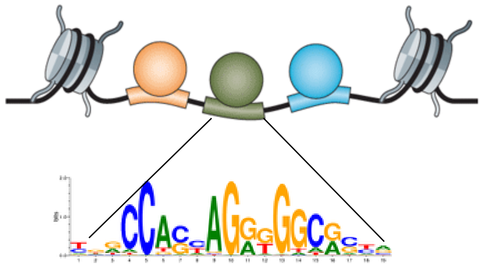
Class = 0 (1.2)

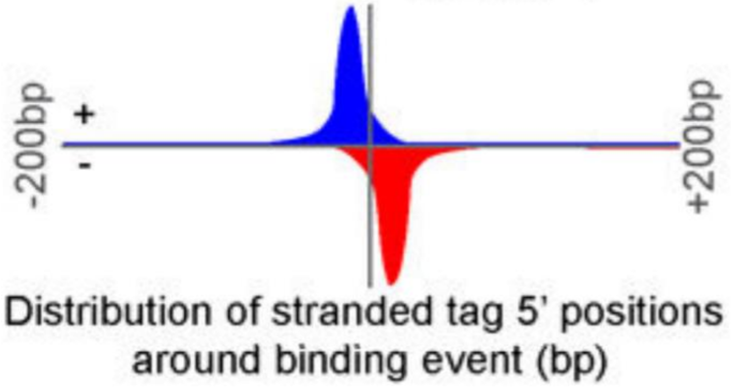
Class = 0 (3.5)

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

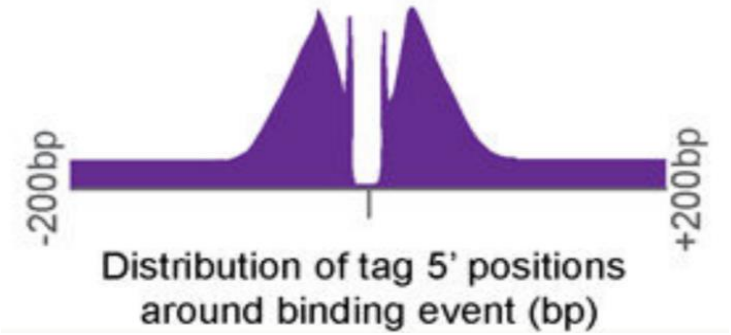


High-resolution ‘shapes’ and ‘spans‘ of TF and chromatin profiles capture exquisite information about protein-DNA contacts




-200bp
+
-
+200bp
Distribution of stranded tag 5’ positions
around binding event (bp)

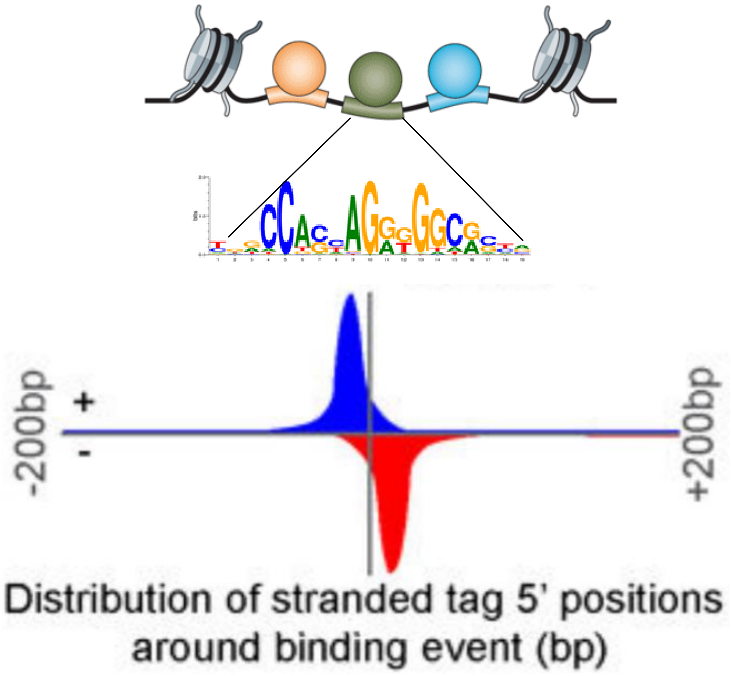
Protein-DNA binding expt.



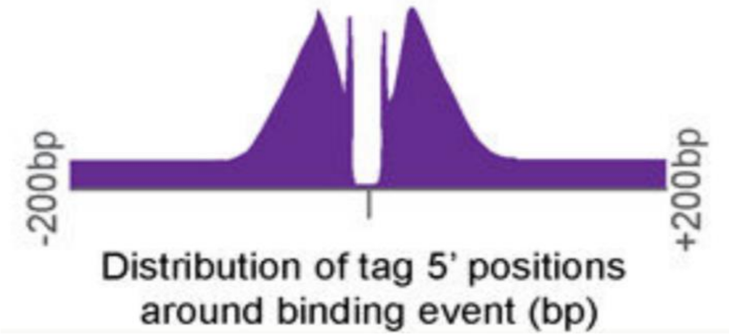
DNA accessibility experiments

https://doi.org/10.3109/10409238.2015.1051505

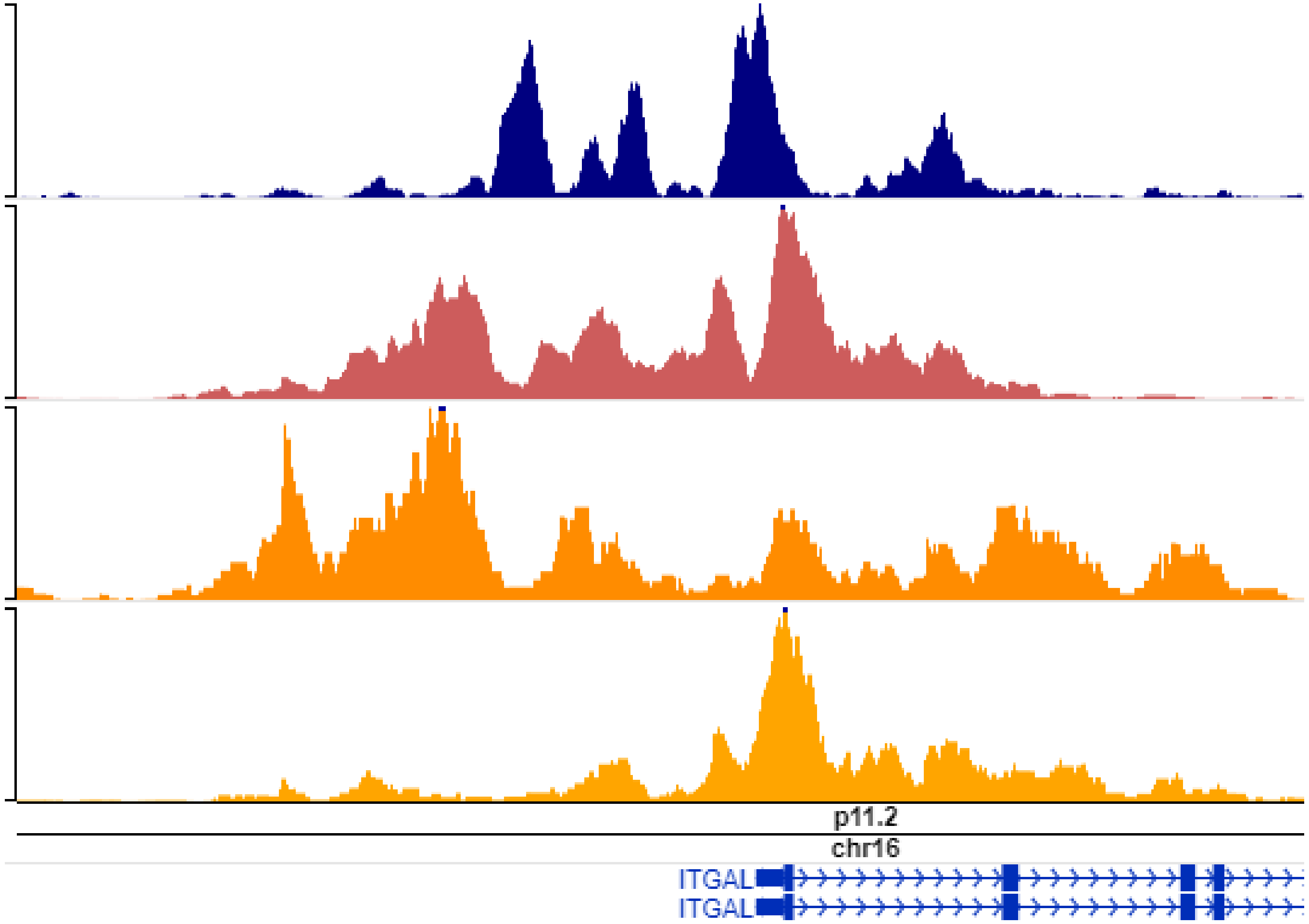
High-resolution ‘shapes’ and ‘spans‘ of TF and chromatin profiles capture exquisite information about protein-DNA contacts


-200bp
+
-
+200bp
Distribution of stranded tag 5' positions around binding event (bp)

Protein-DNA binding expt.



DNA accessibility experiments



DNase-seq
ATAC-seq

H3K27ac
ChIP-seq

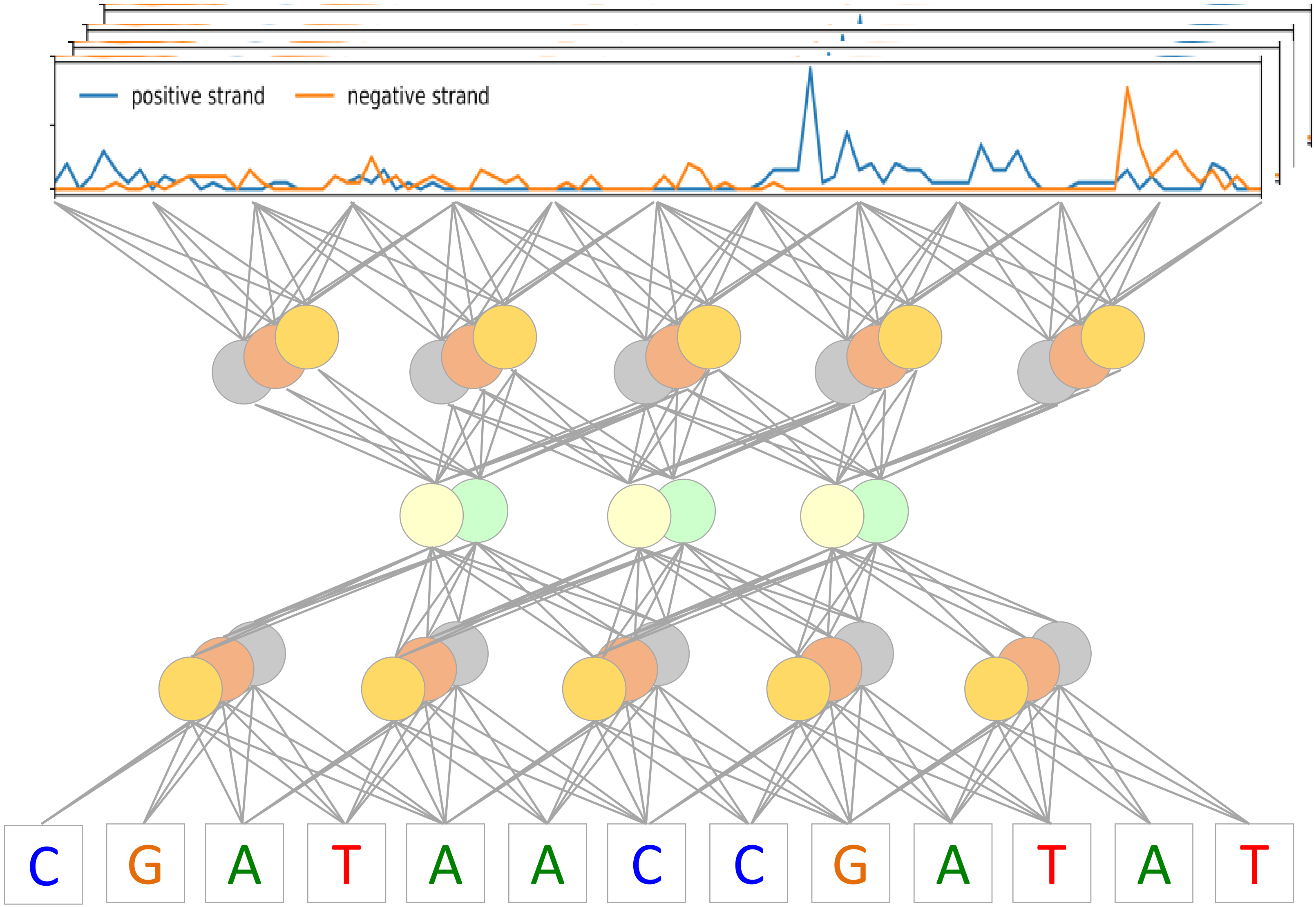
H3K4me1
ChIP-seq

H3K4me3
ChIP-seq

https://doi.org/10.3109/10409238.2015.1051505

BPNet : Sequence to base-res. TF binding profiles

Total reads + base-resolution probability profile (1 kb)



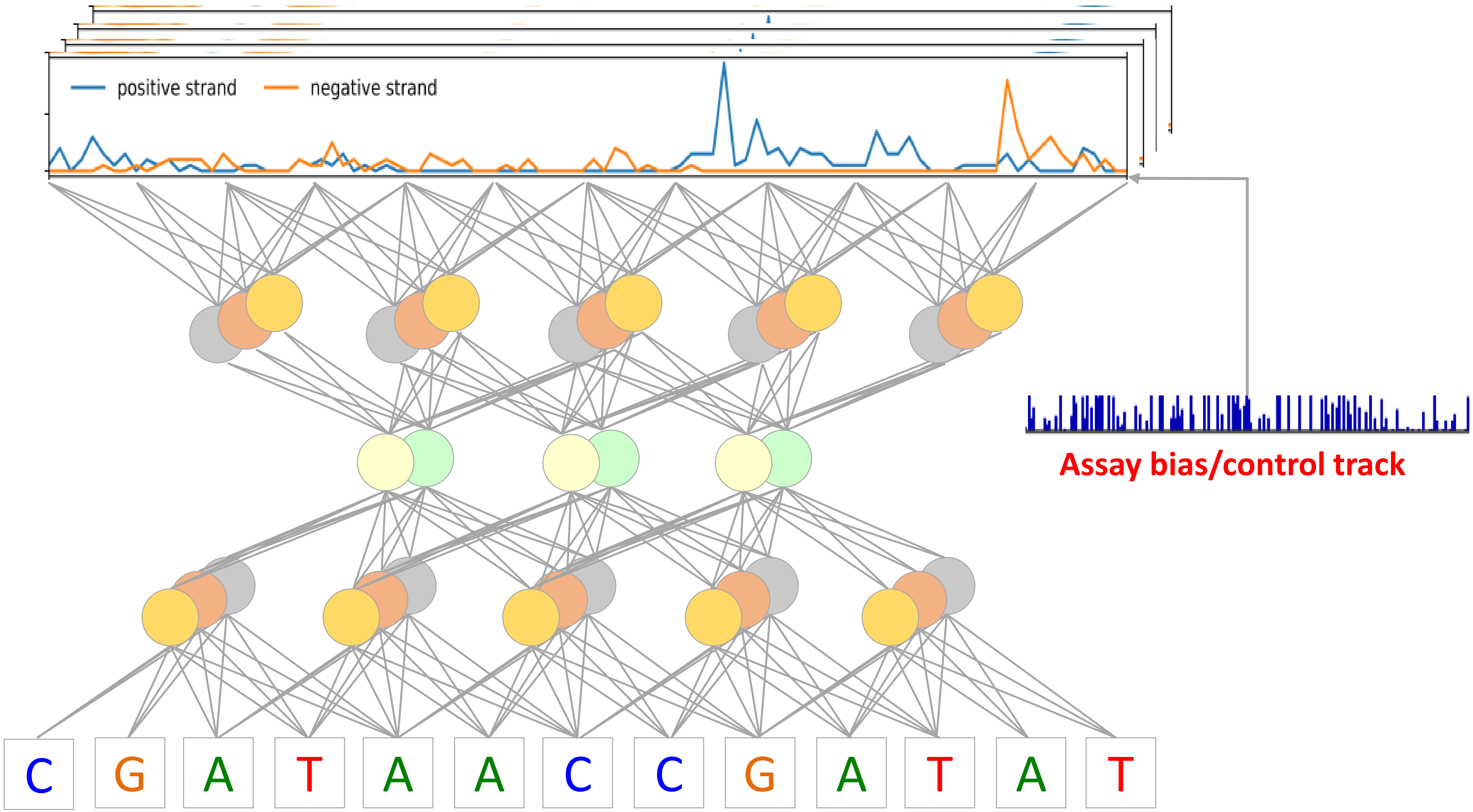
Sequence windows (2 kb)

Ziga Avsec

Avsec et al. 2021 Nature Genetics

BPNet : Sequence to base-res. TF binding profiles

Total reads + base-resolution probability profile (1 kb)



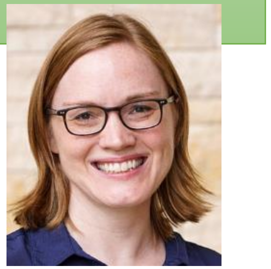
Assay bias/control track

Sequence windows (2 kb)

Ziga Avsec

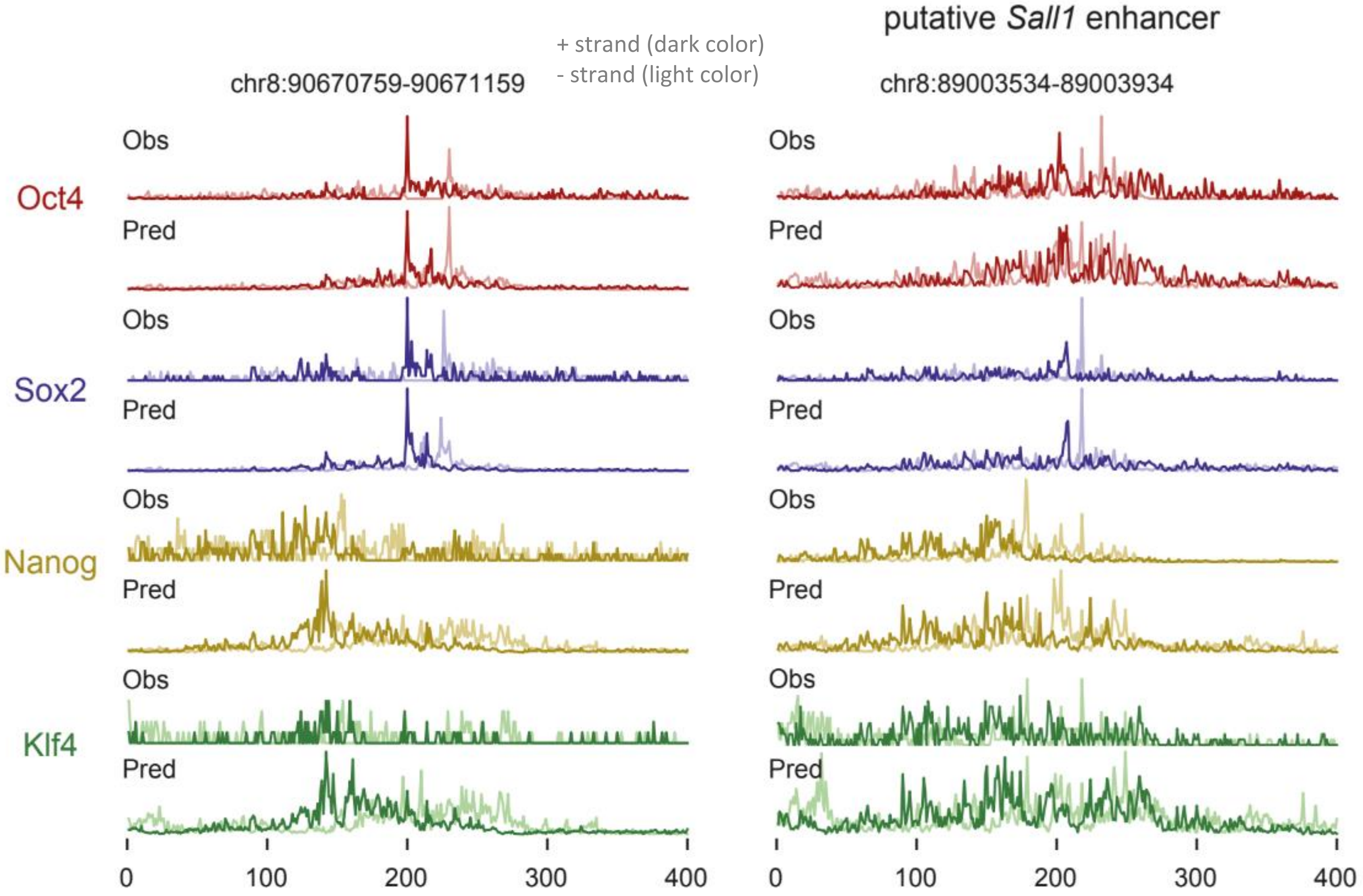
Avsec et al. 2021 Nature Genetics

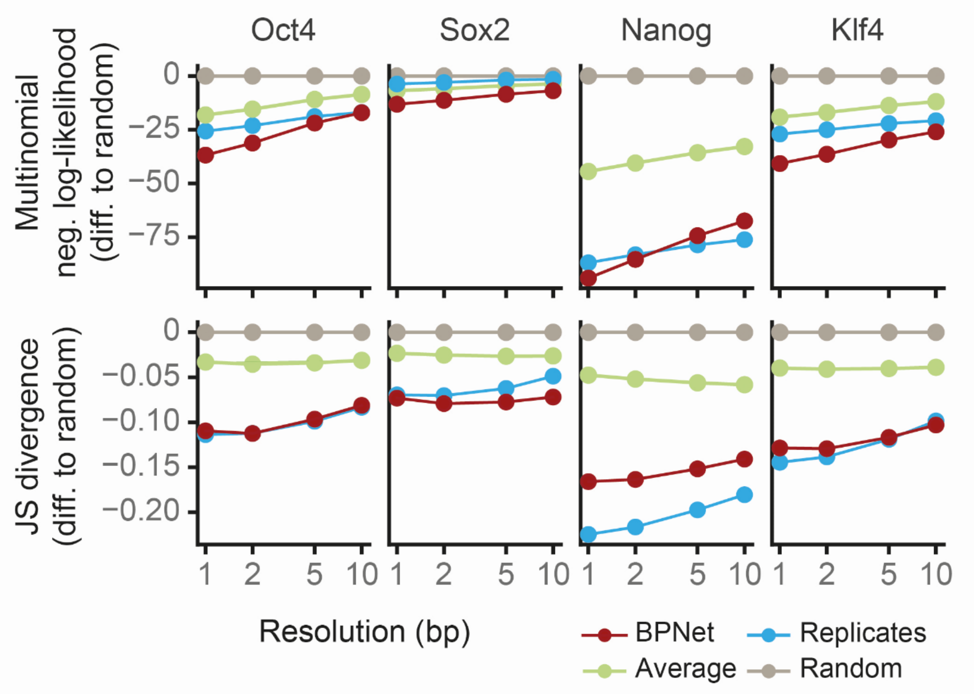
BPNet predicts base resolution protein-DNA binding profiles with unprecedented accuracy (on par with replicate concordance)



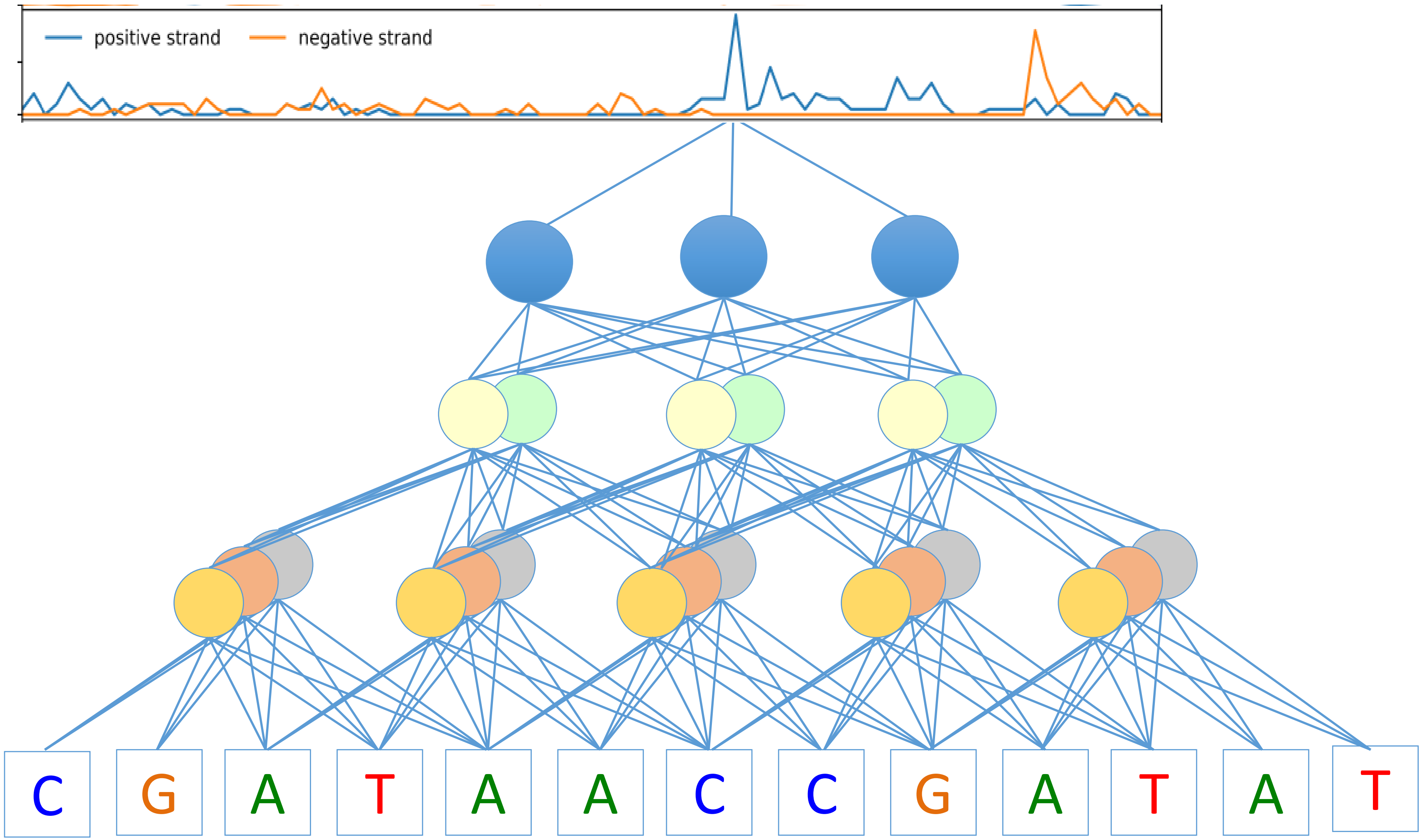
Julia Zeitlinger lab

Oct4, Sox2, Nanog and Klf4 in mESCs





DeepLIFT: Inferring predictive nucleotides in any sequence

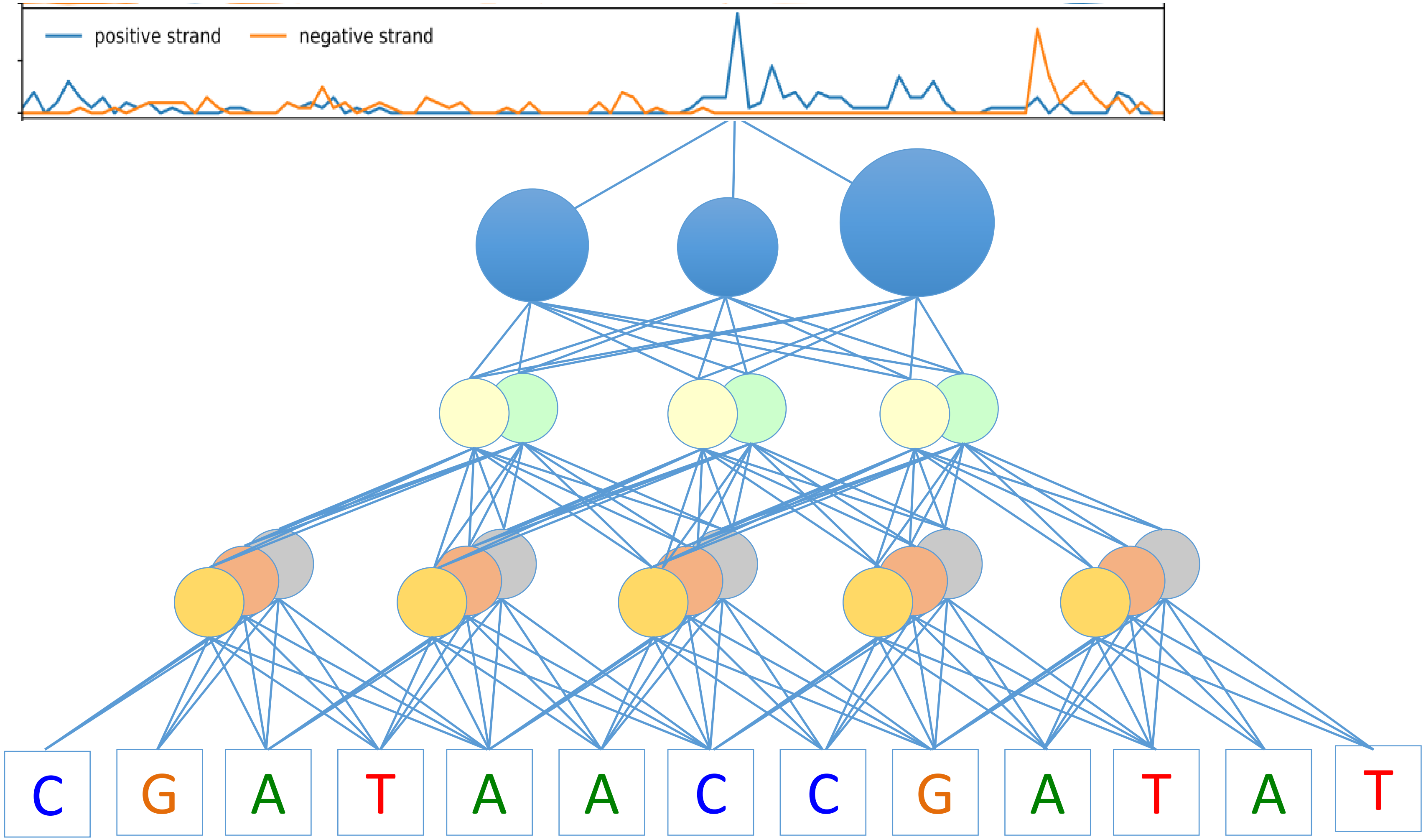


Avanti Shrikumar

Alex Tseng

Shrikumar et al. 2017 ICML
Shrikumar et al. 2019 ISMB
Tseng et al. 2020 NeurIPS
Greenside et al. 2018, ECCB

DeepLIFT: Inferring predictive nucleotides in any sequence

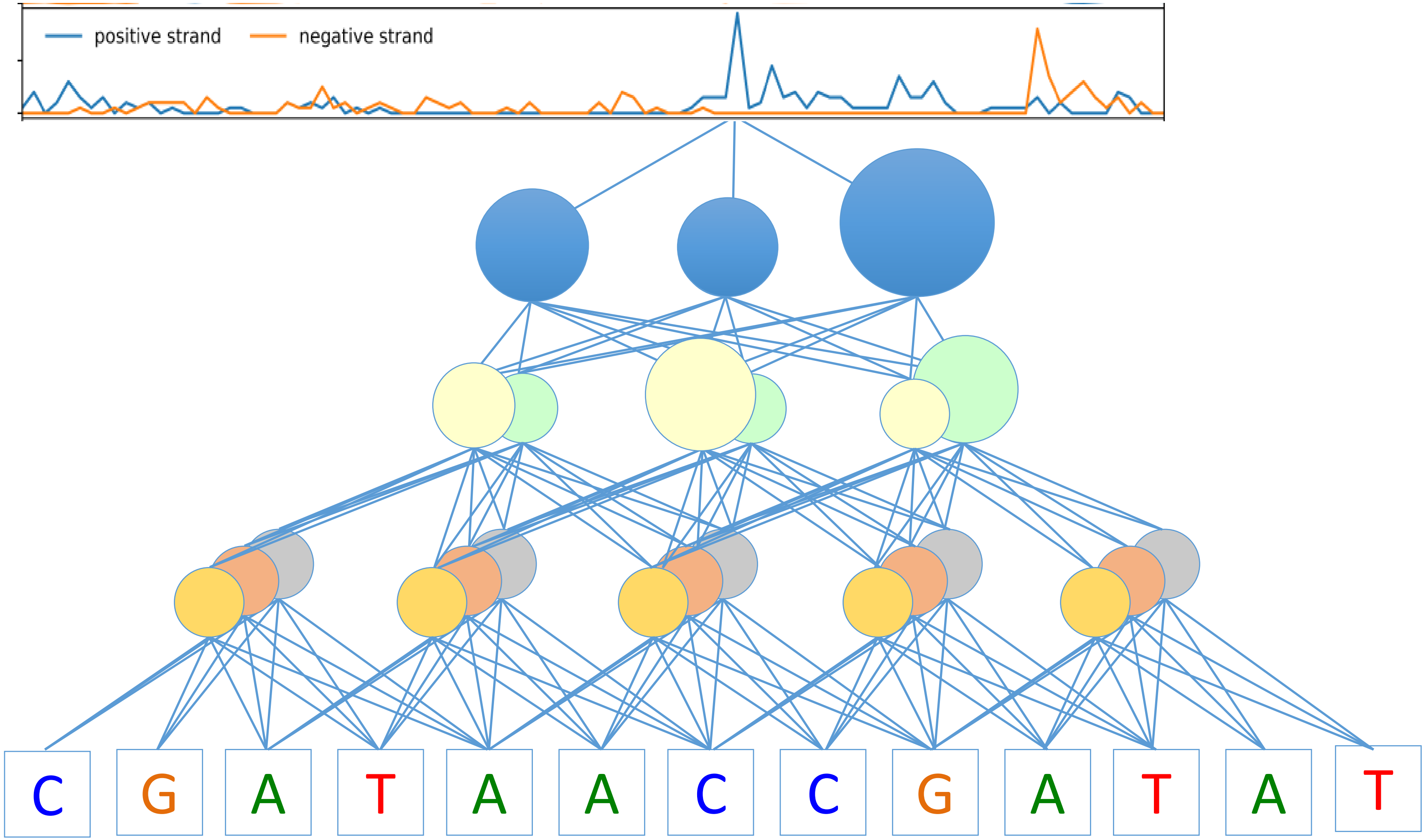


Avanti Shrikumar

Alex Tseng

Shrikumar et al. 2017 ICML
Shrikumar et al. 2019 ISMB
Tseng et al. 2020 NeurIPS
Greenside et al. 2018, ECCB

DeepLIFT: Inferring predictive nucleotides in any sequence

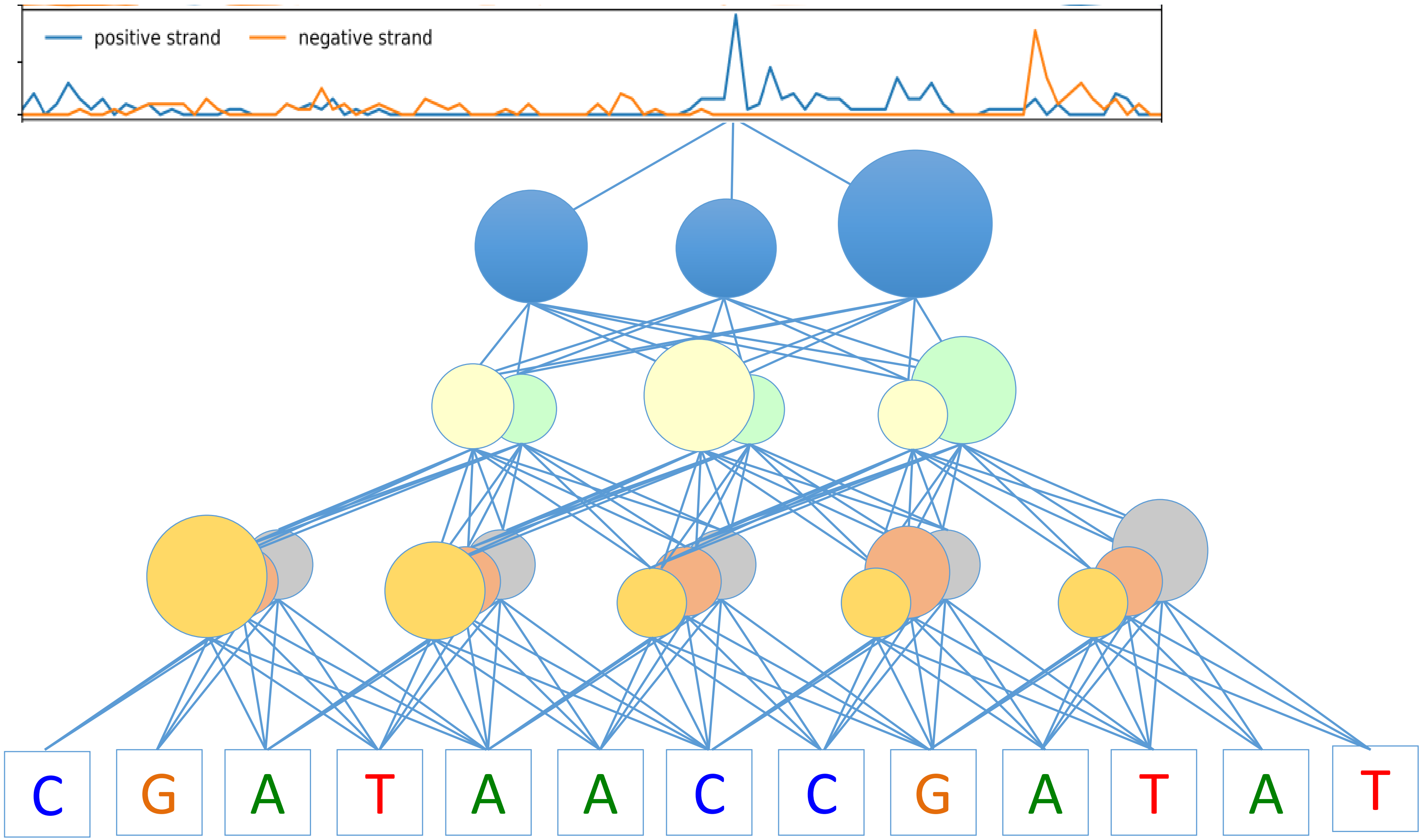


Avanti Shrikumar

Alex Tseng

Shrikumar et al. 2017 ICML
Shrikumar et al. 2019 ISMB
Tseng et al. 2020 NeurIPS
Greenside et al. 2018, ECCB

DeepLIFT: Inferring predictive nucleotides in any sequence

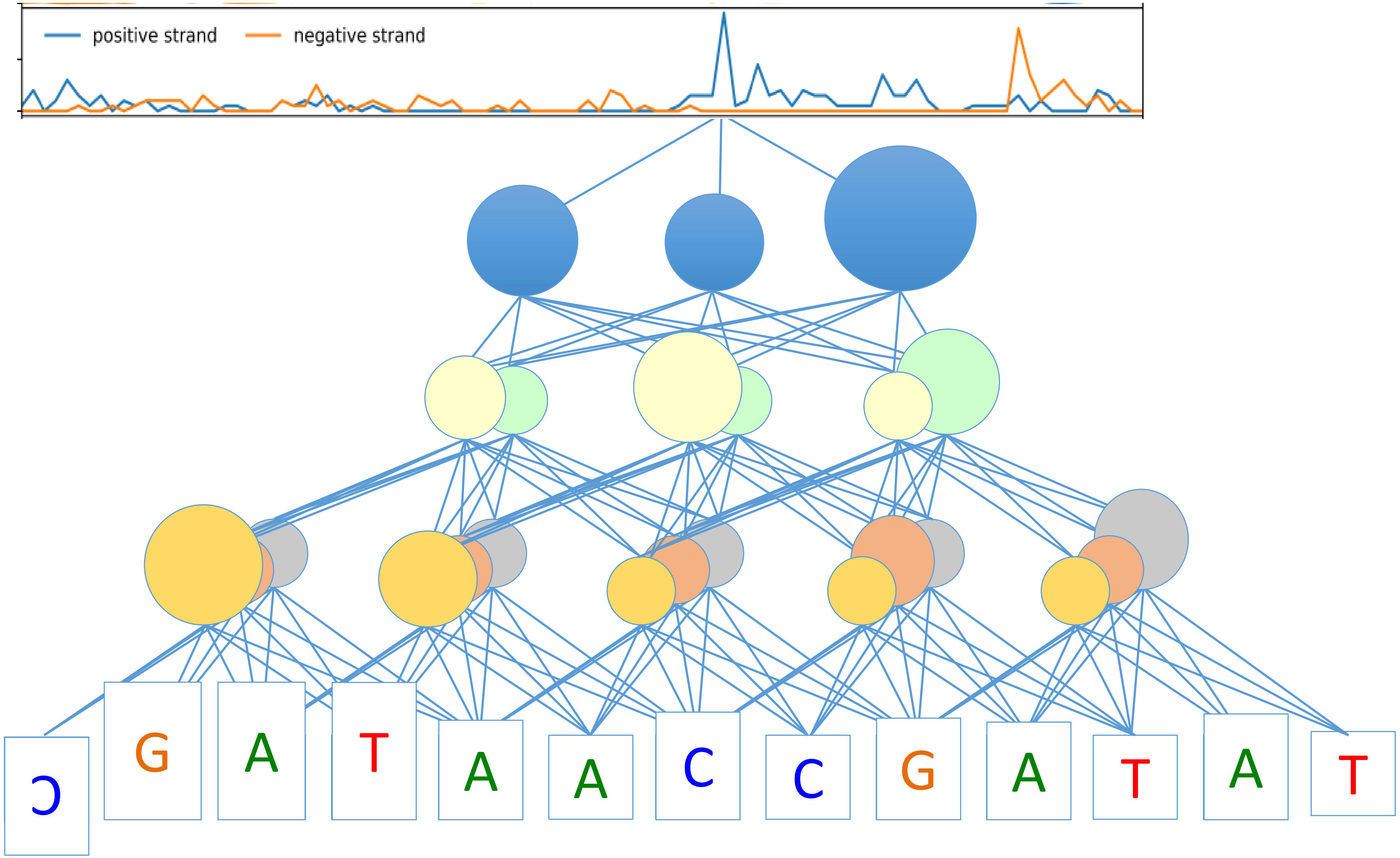


Avanti Shrikumar

Alex Tseng

Shrikumar et al. 2017 ICML
Shrikumar et al. 2019 ISMB
Tseng et al. 2020 NeurIPS
Greenside et al. 2018, ECCB

DeepLIFT: Inferring predictive nucleotides in any sequence

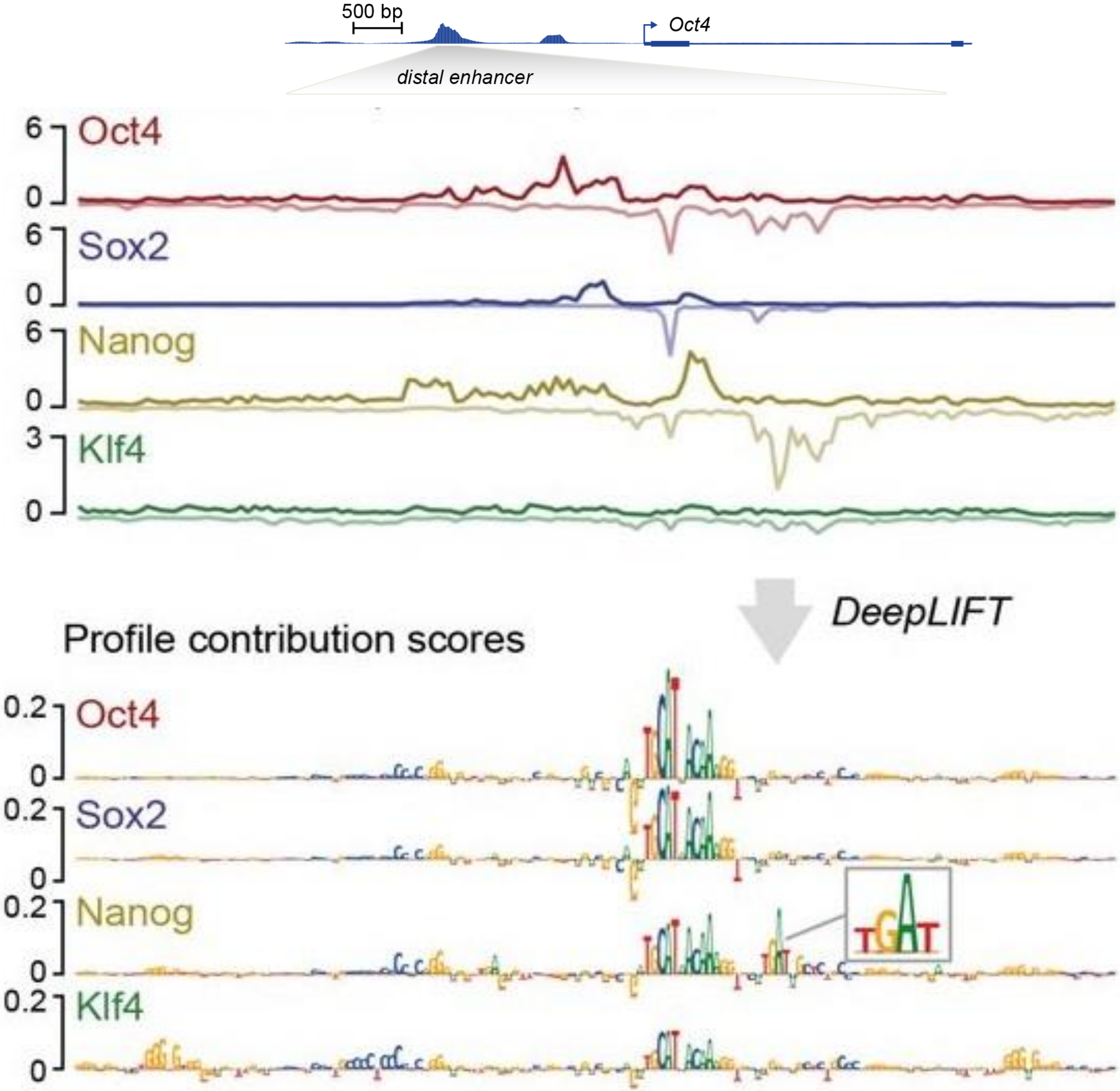


Avanti Shrikumar

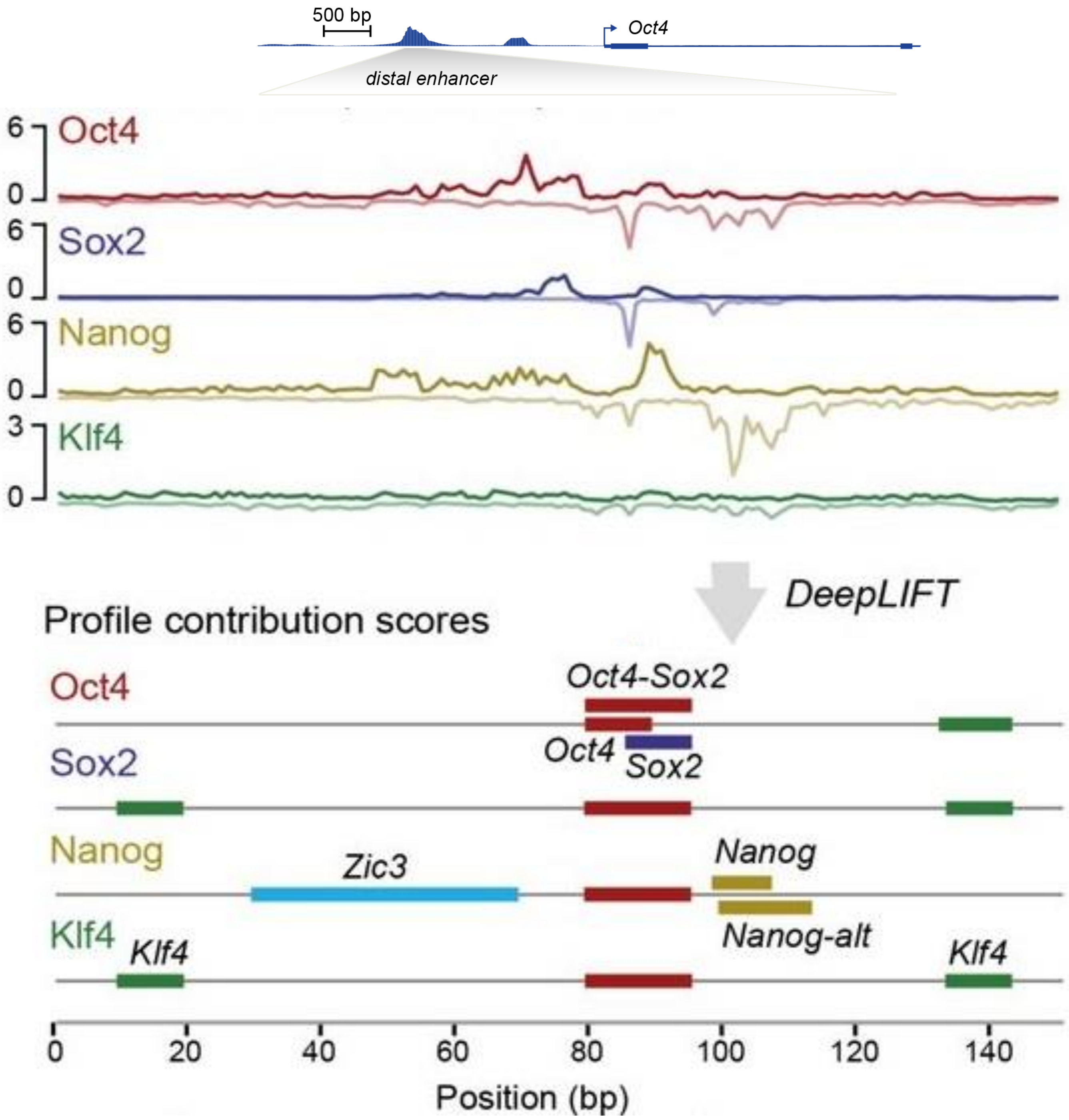
Alex Tseng

Shrikumar et al. 2017 ICML
Shrikumar et al. 2019 ISMB
Tseng et al. 2020 NeurIPS
Greenside et al. 2018, ECCB

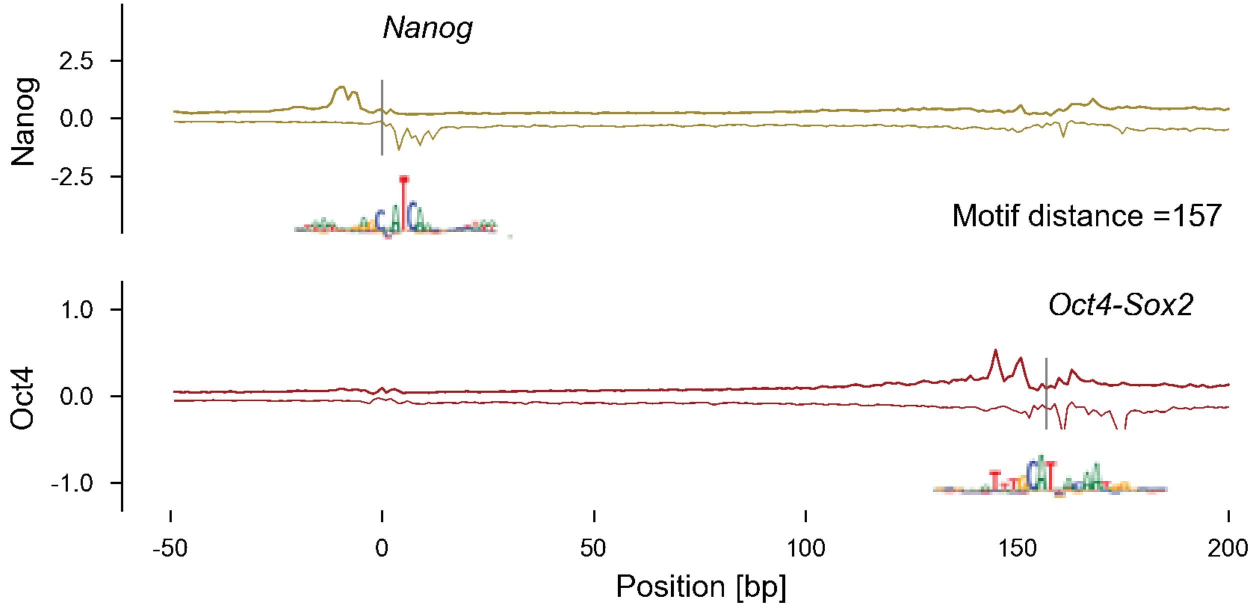
mESCs

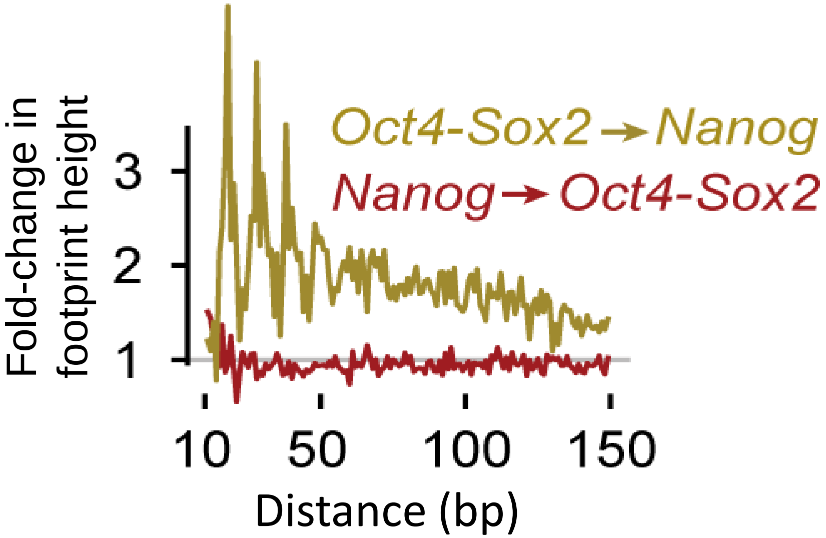

500 bp
Oct4
distal enhancer
6
0
Oct4
Sox2
Nanog
3
Klf4
DeepLIFT
Profile contribution scores
0.2
TGAT

mESCs

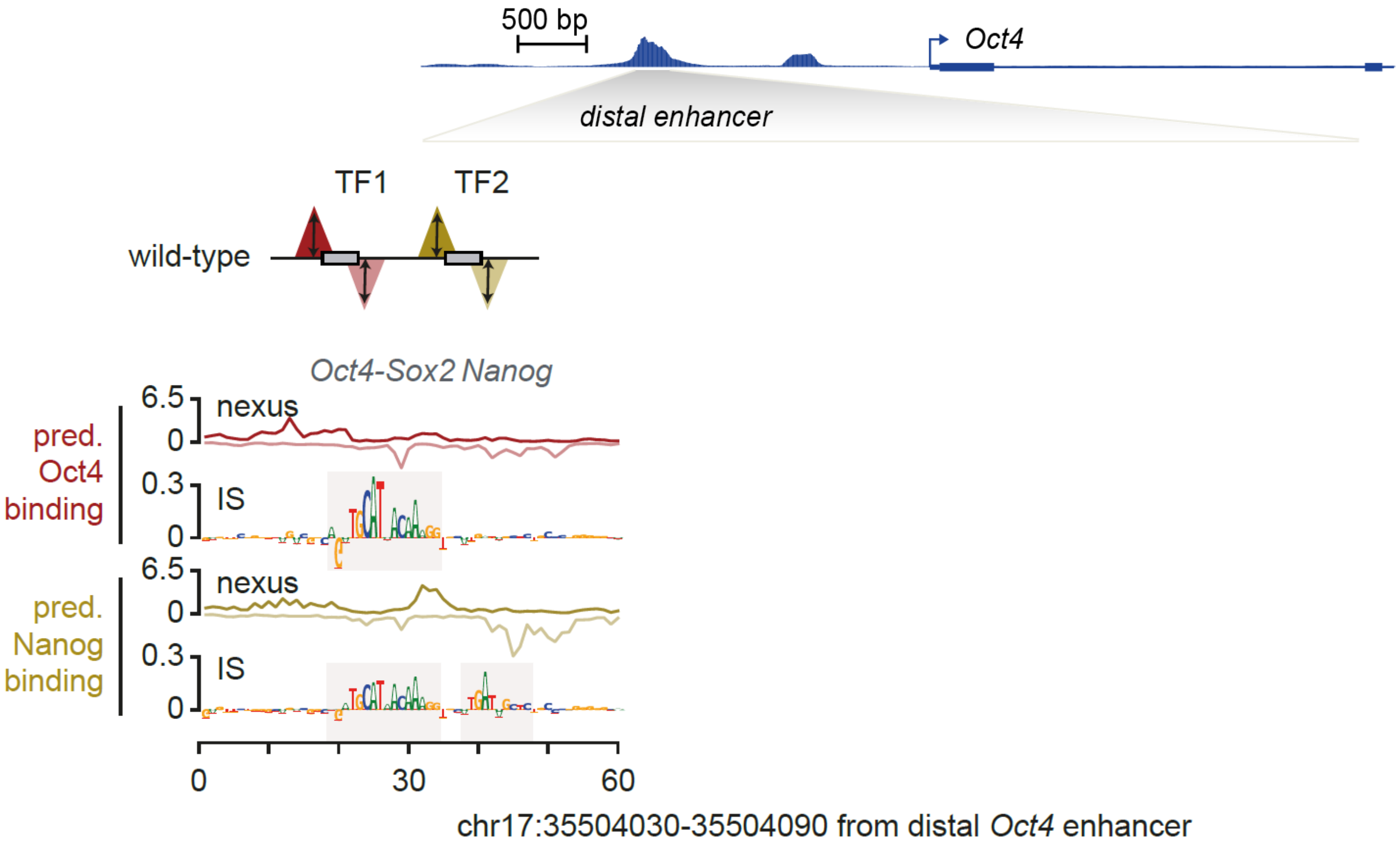

500 bp
Oct4
distal enhancer
6
0
Oct4
6
0
Sox2
6
0
Nanog
3
0
Klf4
DeepLIFT
Profile contribution scores
Oct4
Oct4-Sox2
Oct4
Sox2
Sox2
Nanog
Zic3
Nanog
Nanog-alt
Klf4
Klf4
Klf4
0
20
40
60
80
100
120
140
Position (bp)

Deciphering syntax dependent TF cooperativity with synthetic designed sequences

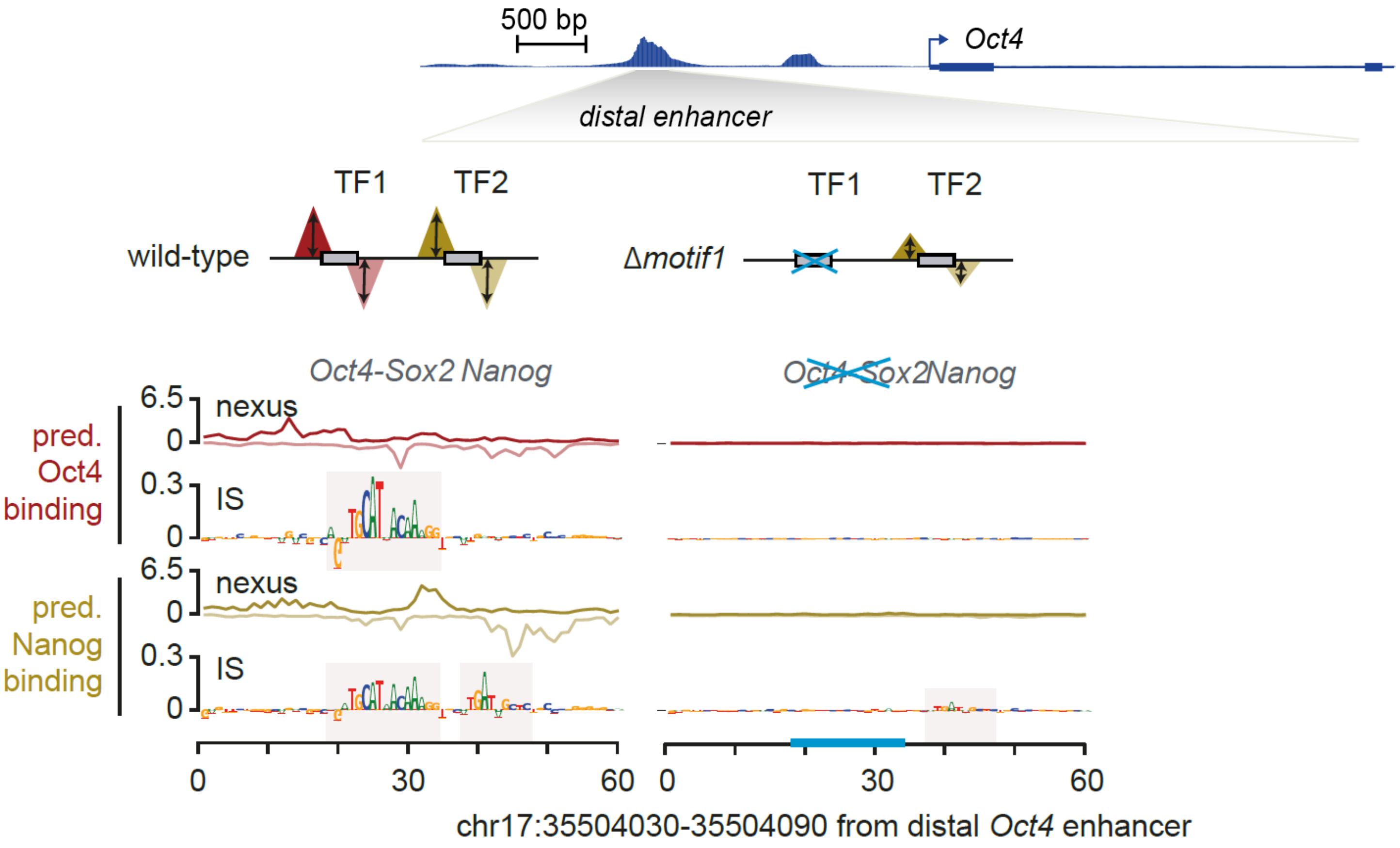




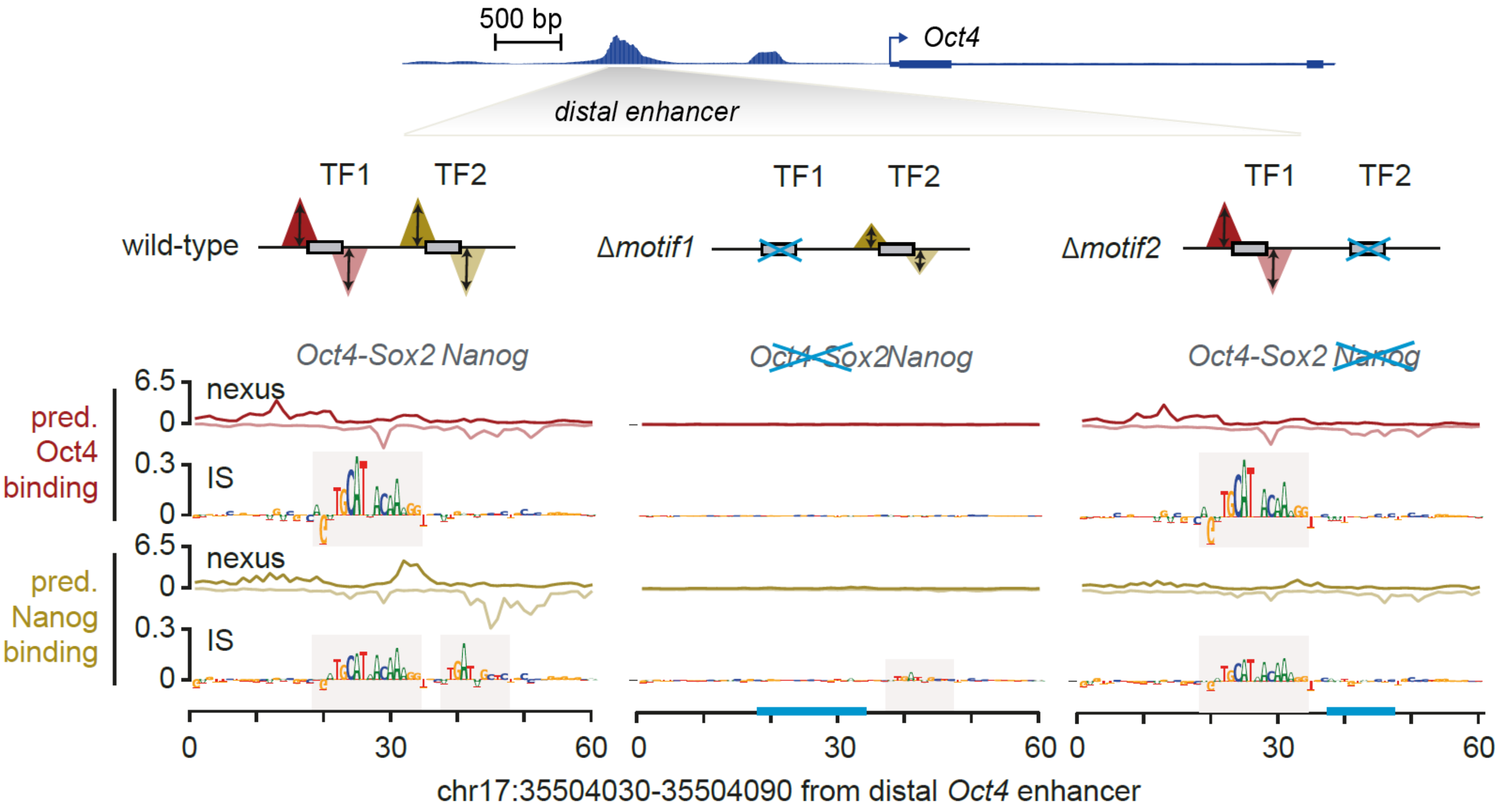
Deciphering syntax dependent TF cooperativity with *in-silico* genome editing



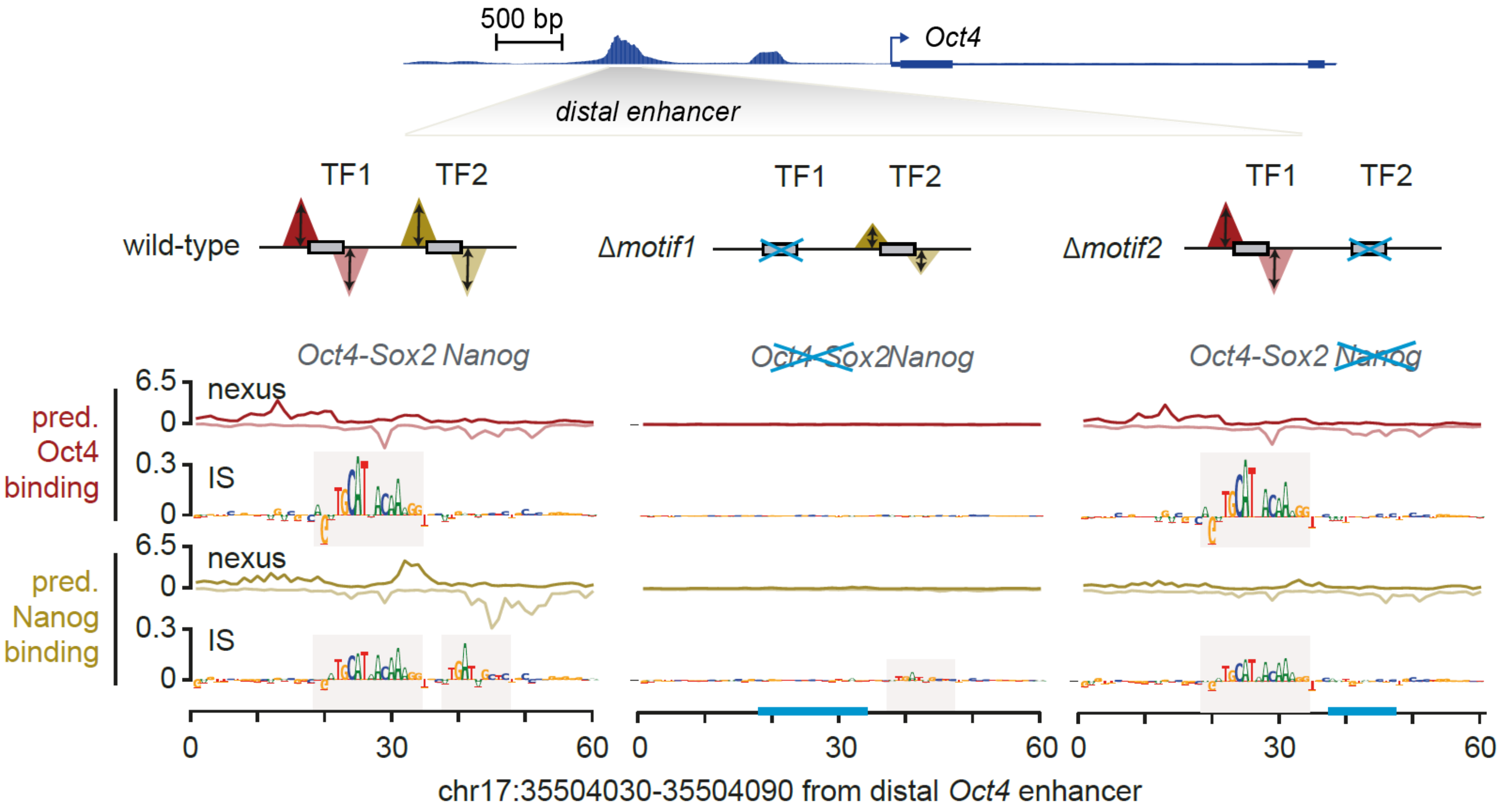
Deciphering syntax dependent TF cooperativity with *in-silico* genome editing

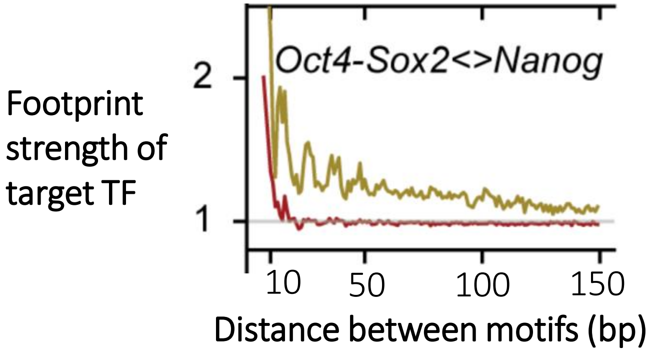


Deciphering syntax dependent TF cooperativity with *in-silico* genome editing

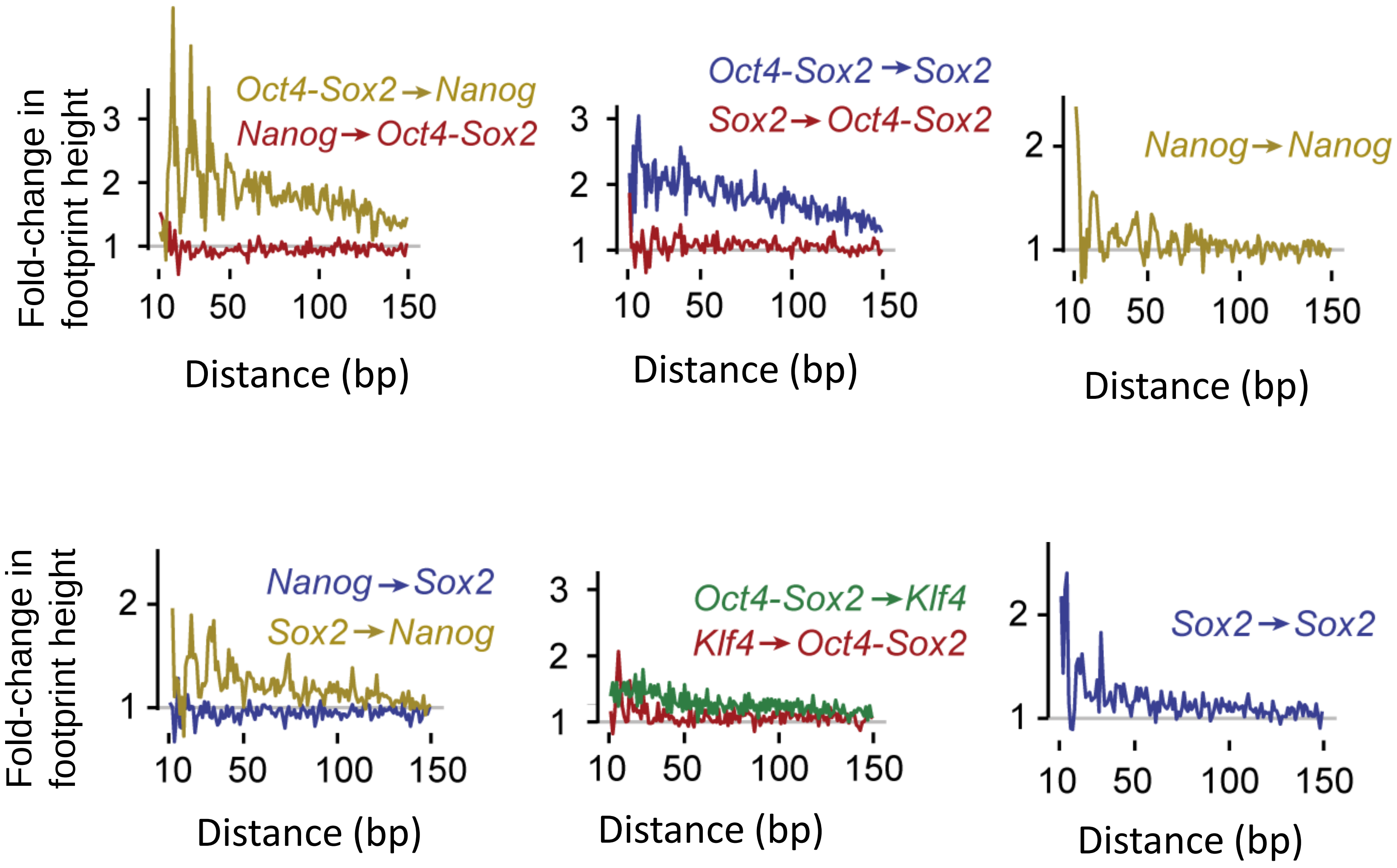


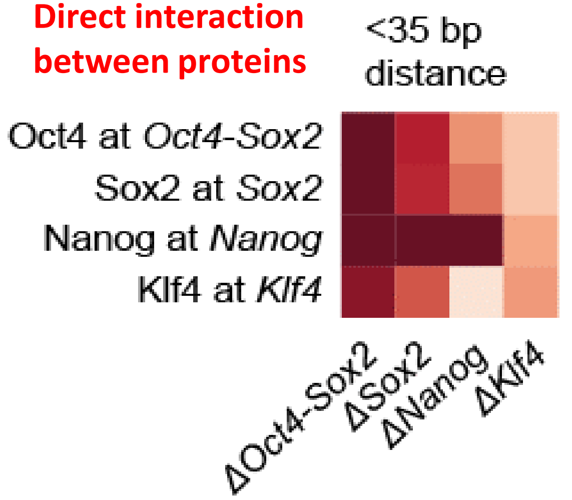
Deciphering syntax dependent TF cooperativity with *in-silico* genome editing

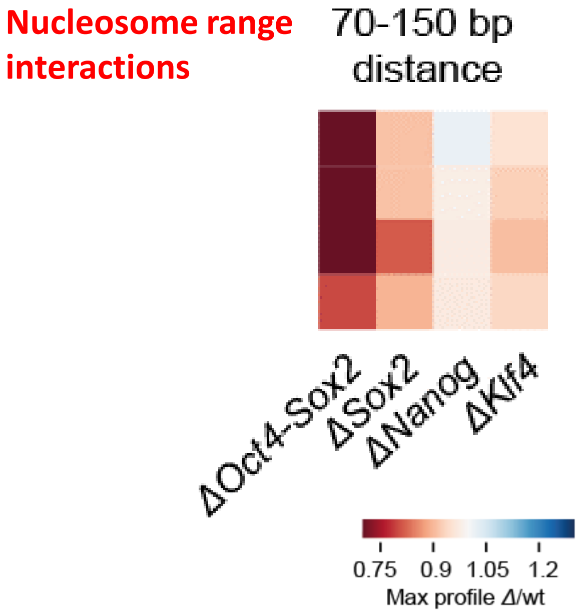




Distance dependent motif syntax rules of asymmetric directional cooperativity

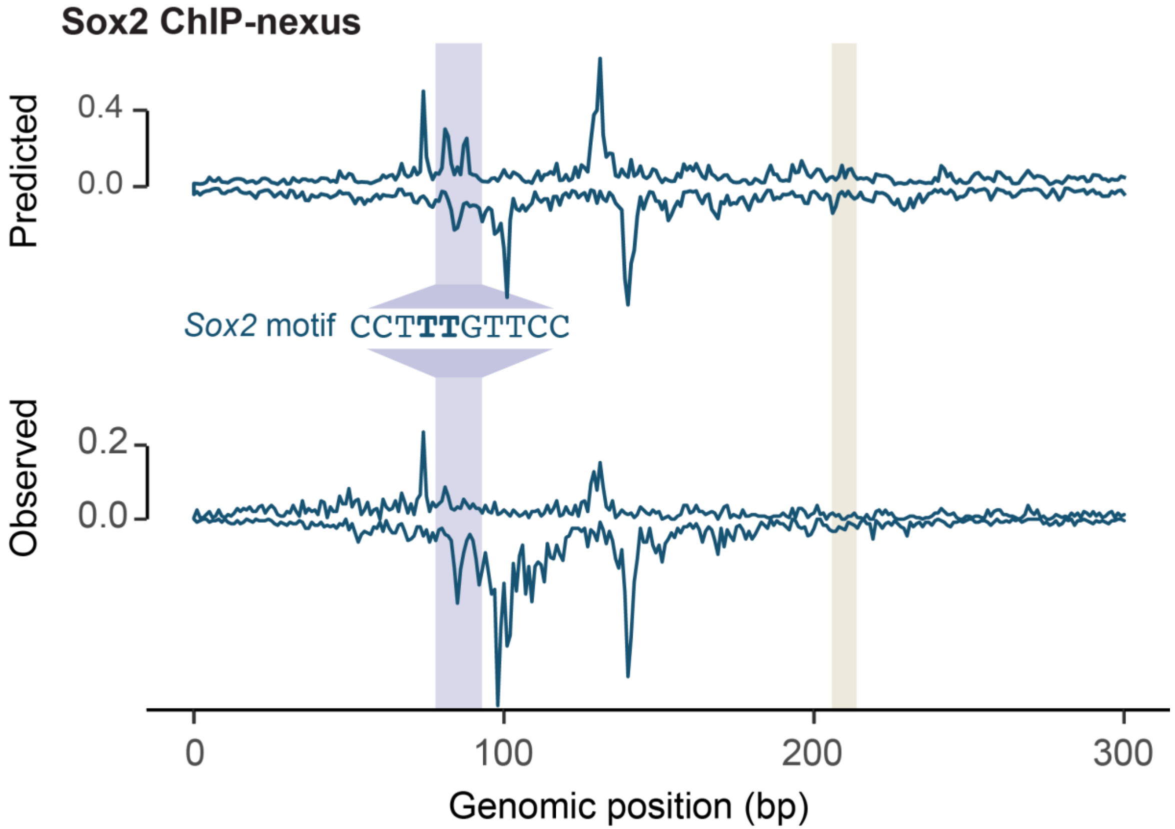



Direct interaction between proteins


Nucleosome range interactions

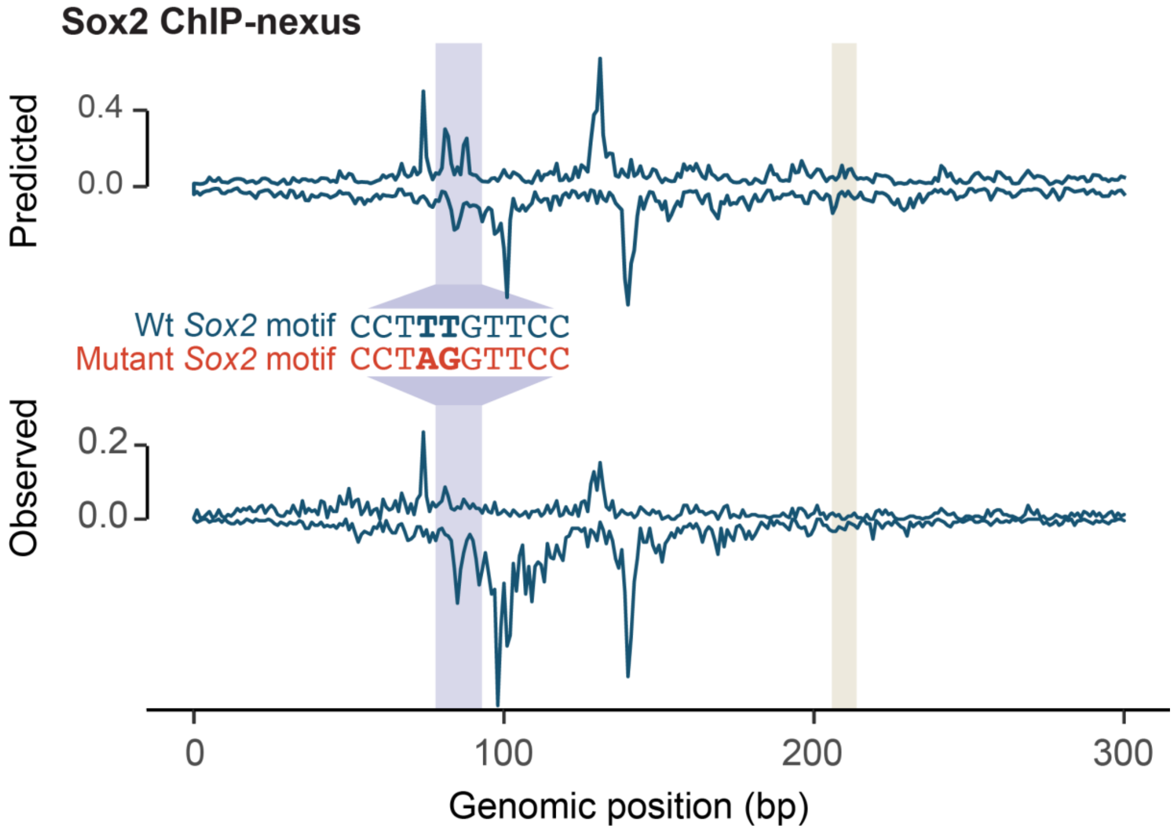
Validated with CRISPR/Cas9 syntax editing experiments

CRISPR mutations validate motif syntax Nanog <> Sox2



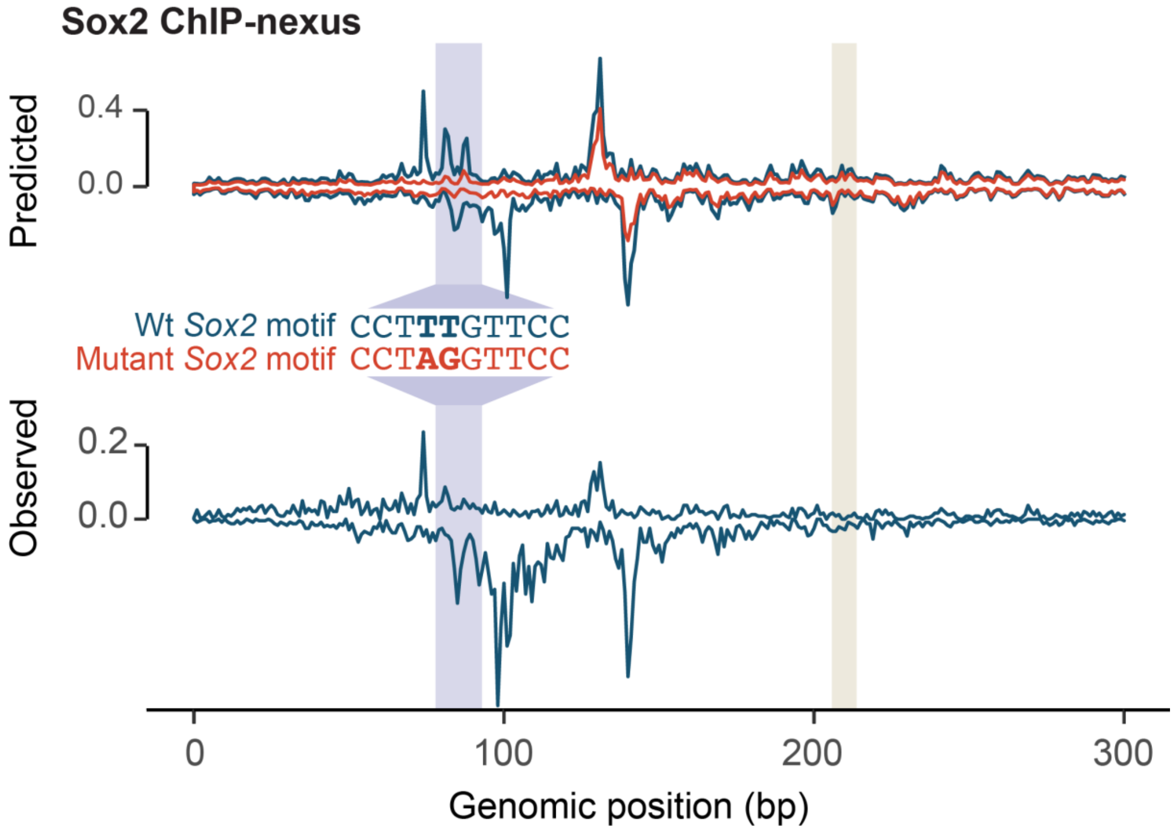
Julia Zeitlinger, Sabrina Krueger, Melanie Weilert

CRISPR mutations validate motif syntax Nanog <> Sox2



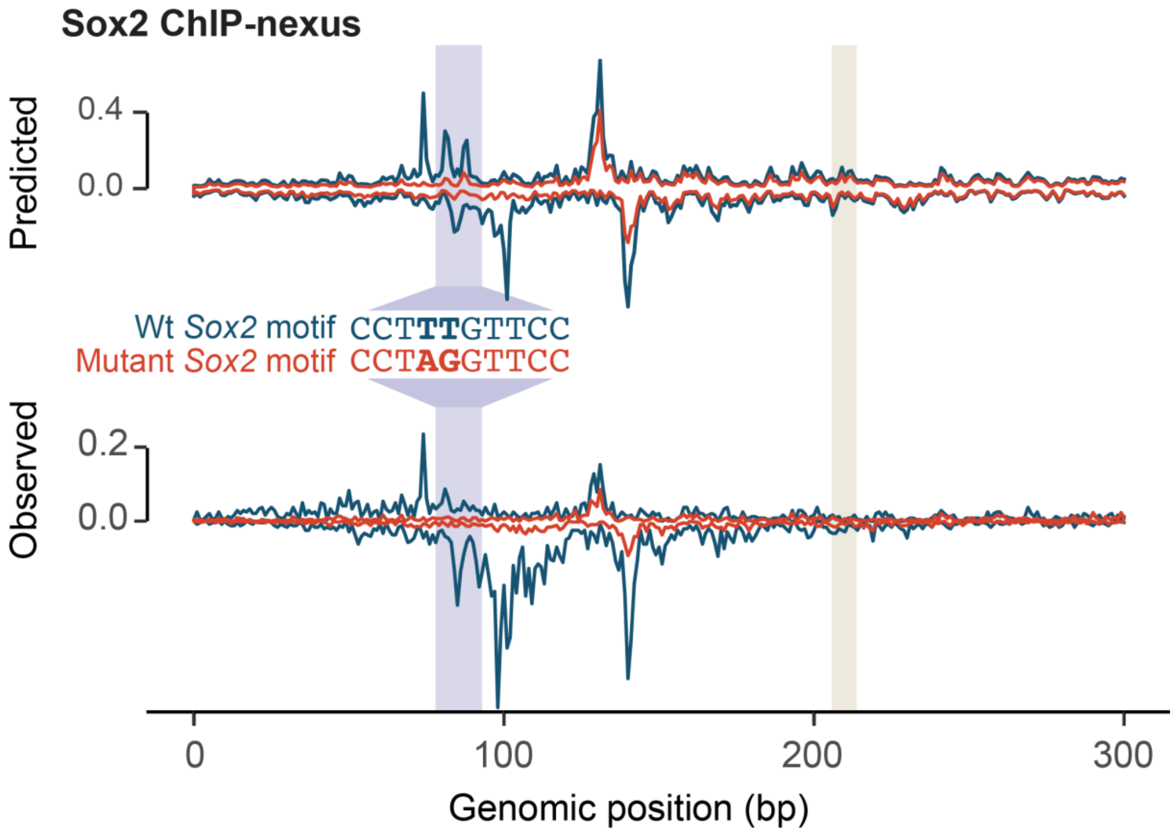
Julia Zeitlinger, Sabrina Krueger, Melanie Weilert

CRISPR mutations validate motif syntax Nanog <> Sox2



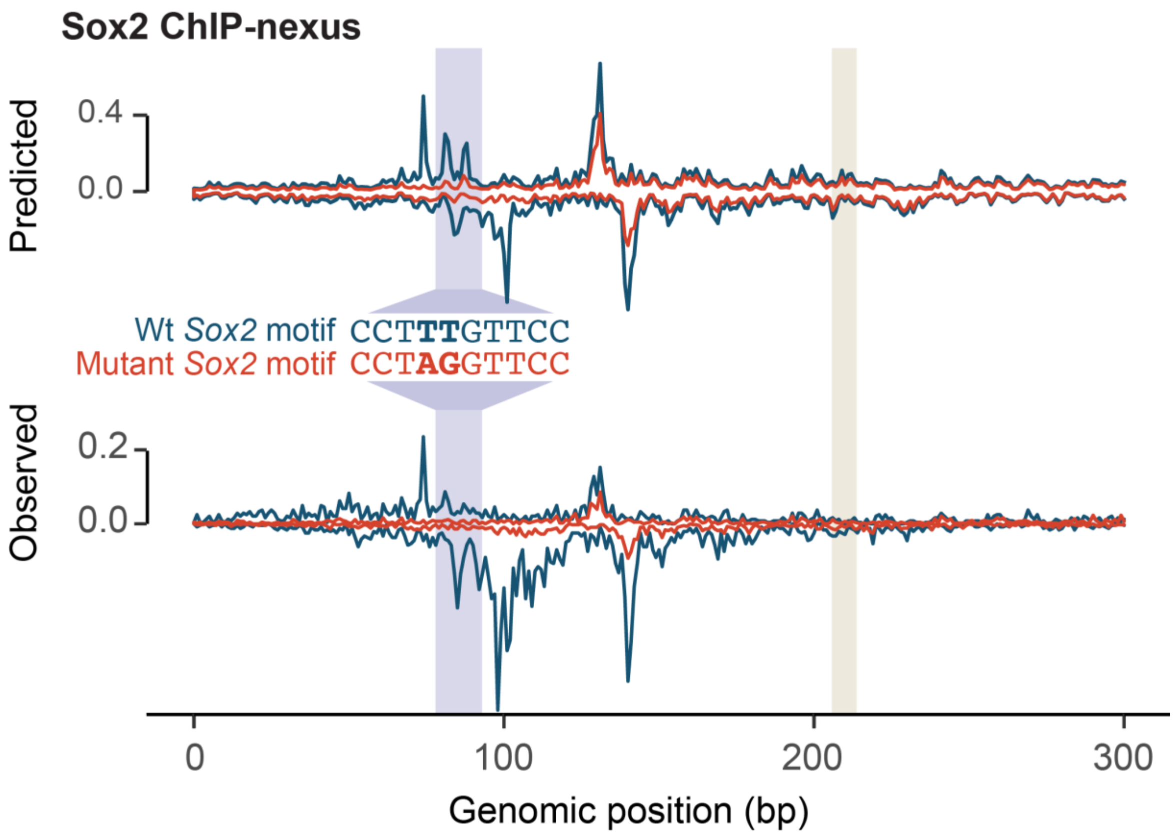
Julia Zeitlinger, Sabrina Krueger, Melanie Weilert

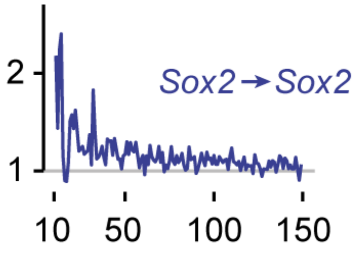
CRISPR mutations validate motif syntax Nanog <> Sox2



Julia Zeitlinger, Sabrina Krueger, Melanie Weilert

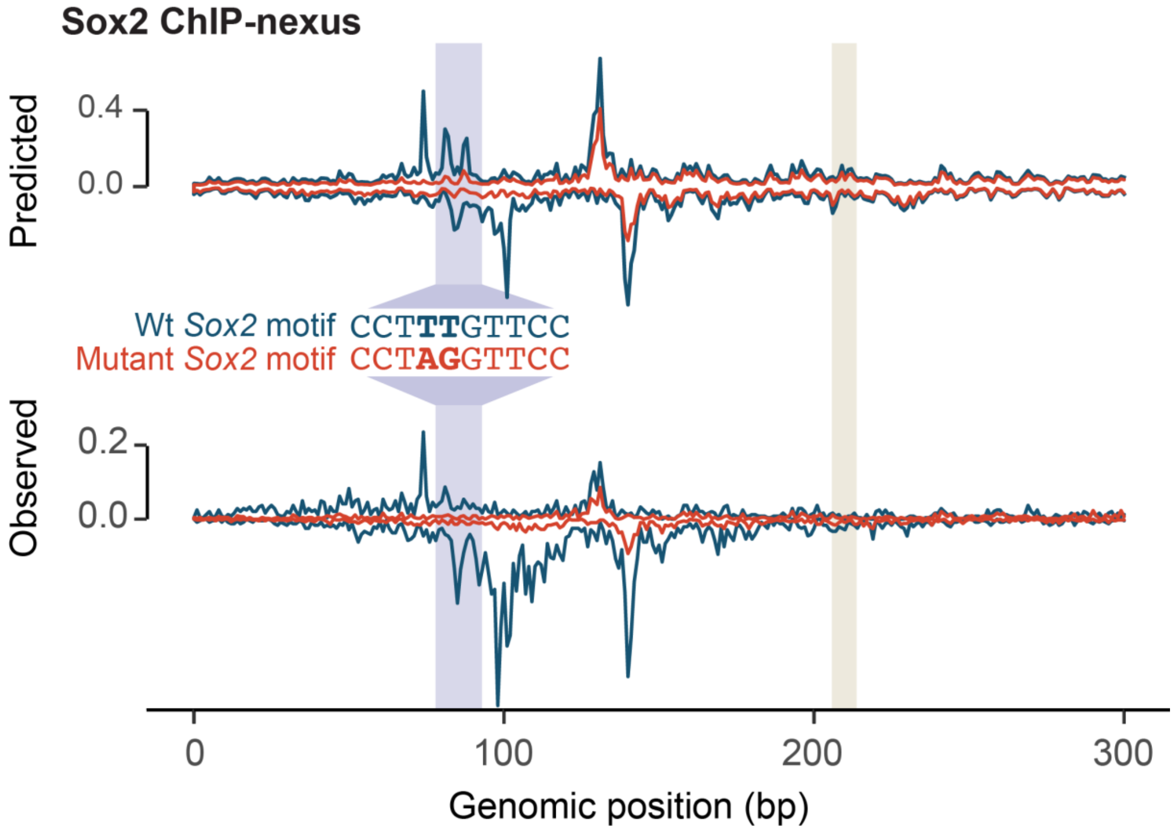
CRISPR mutations validate motif syntax Nanog <> Sox2

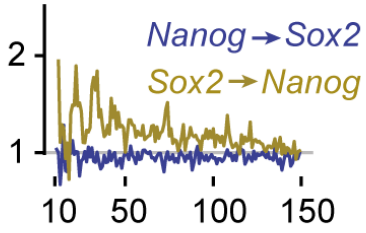


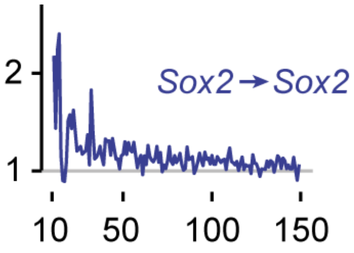


Julia Zeitlinger, Sabrina Krueger, Melanie Weilert

CRISPR mutations validate motif syntax Nanog <> Sox2

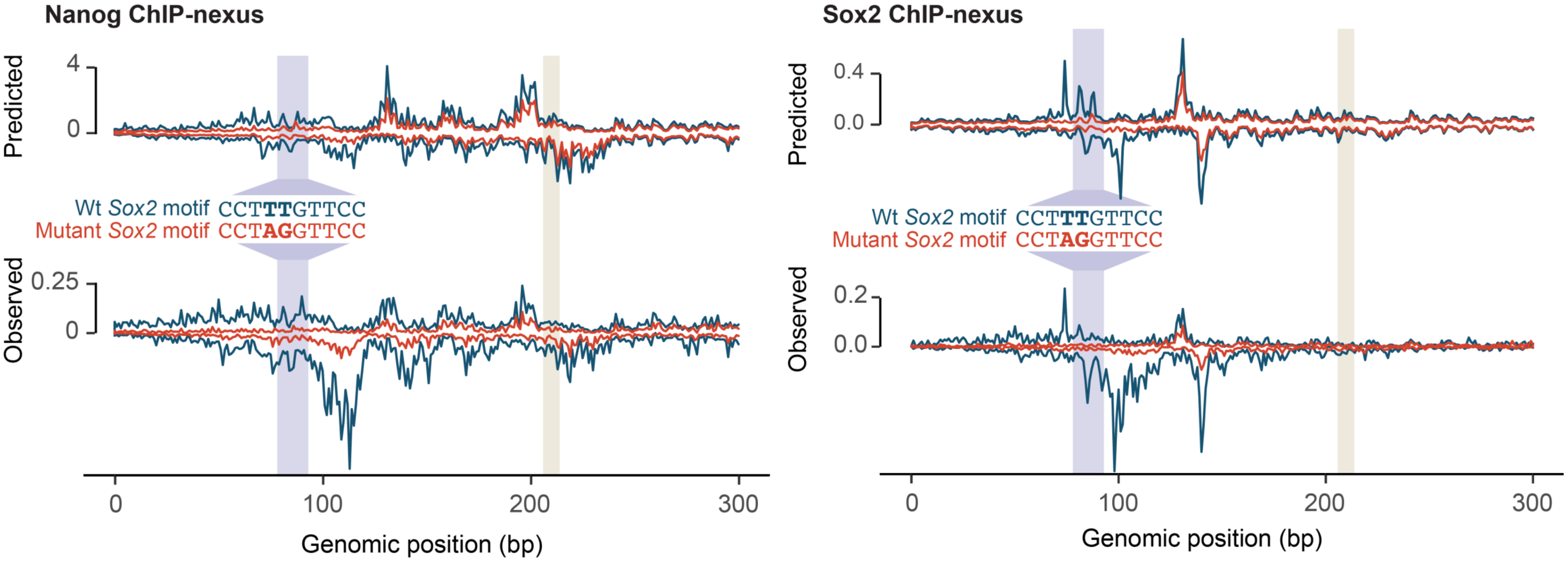


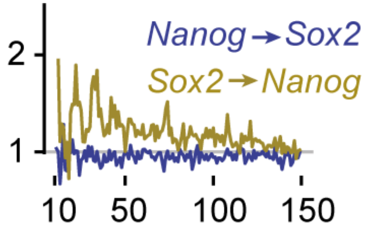


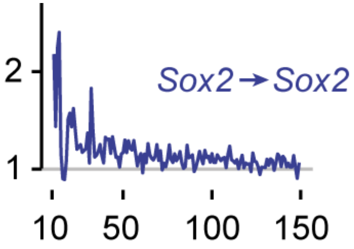


Julia Zeitlinger, Sabrina Krueger, Melanie Weilert

CRISPR mutations validate motif syntax Nanog <> Sox2

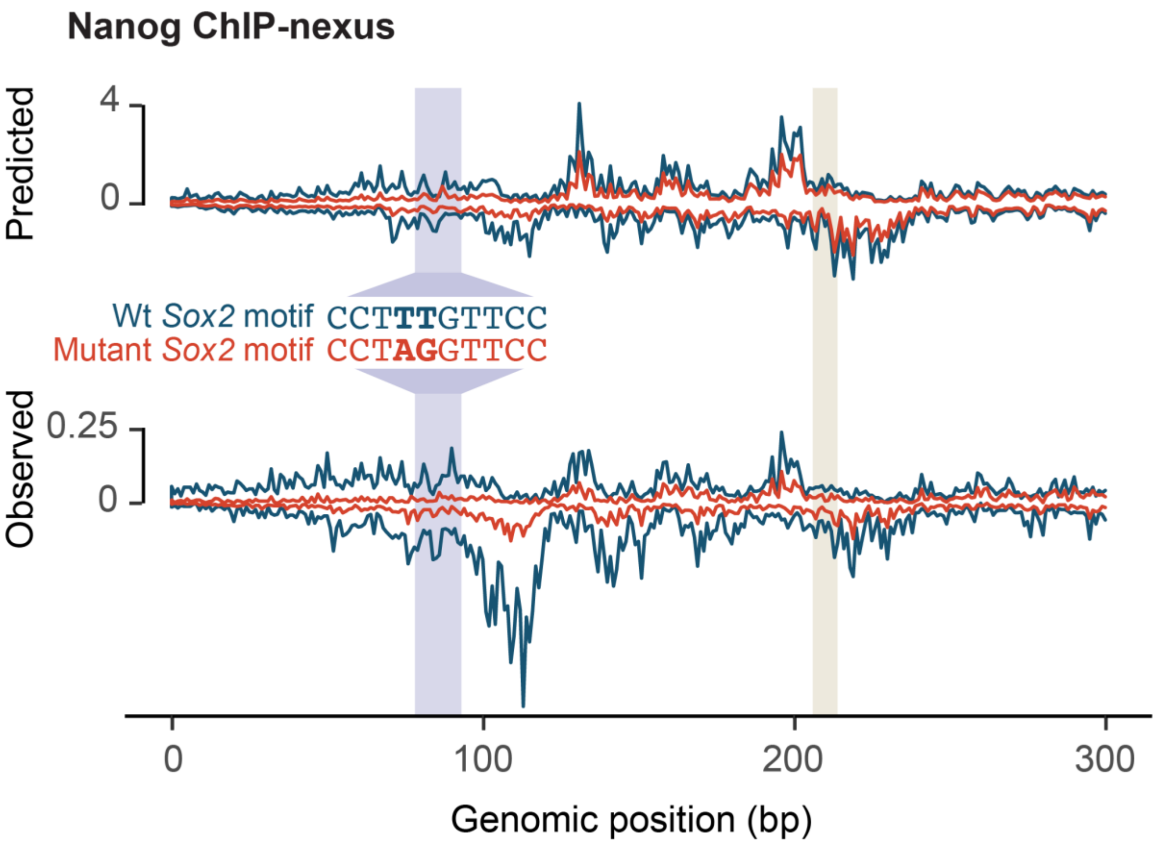


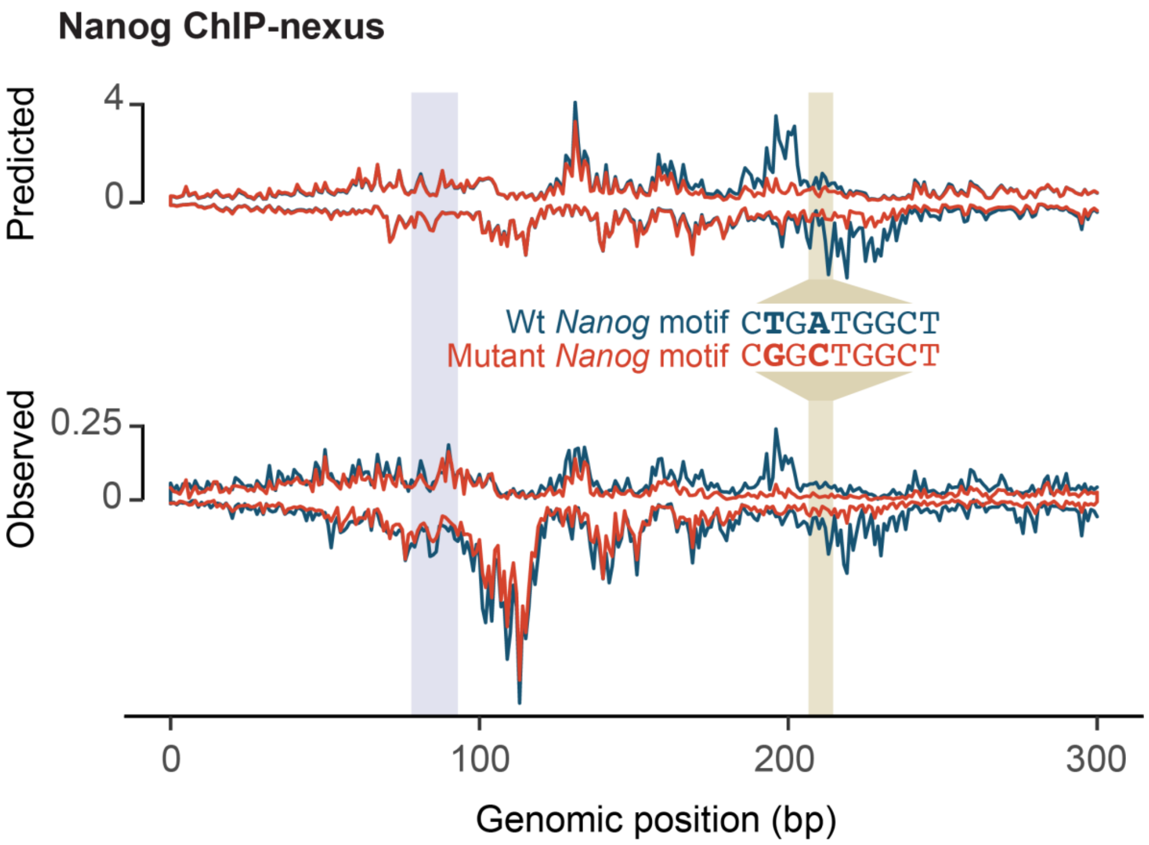


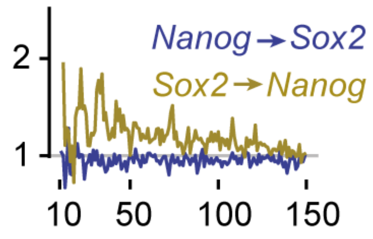


Julia Zeitlinger, Sabrina Krueger, Melanie Weilert

CRISPR mutations validate motif syntax Nanog <> Sox2



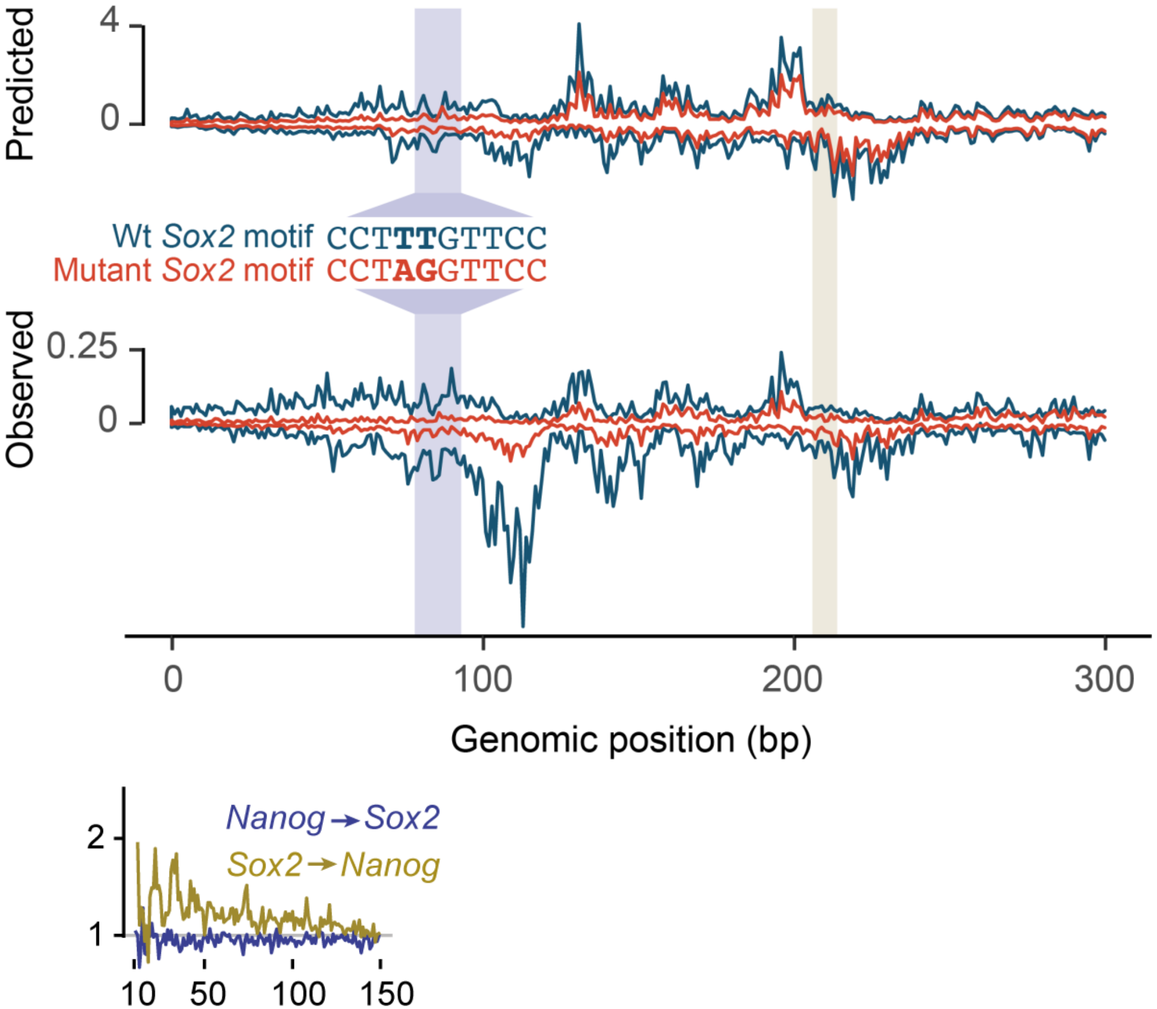




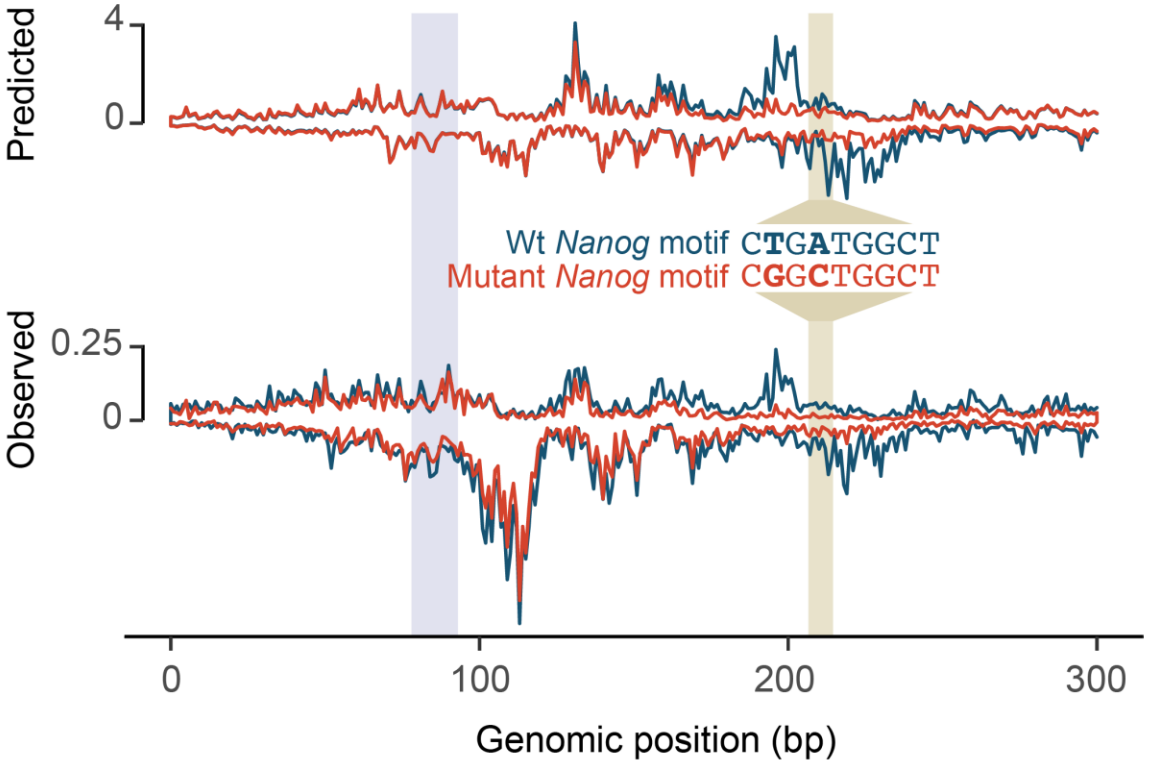
Julia Zeitlinger, Sabrina Krueger, Melanie Weilert

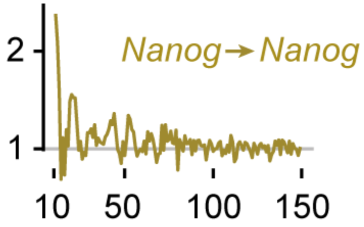
CRISPR mutations validate motif syntax Nanog <> Sox2

Nanog ChIP-nexus



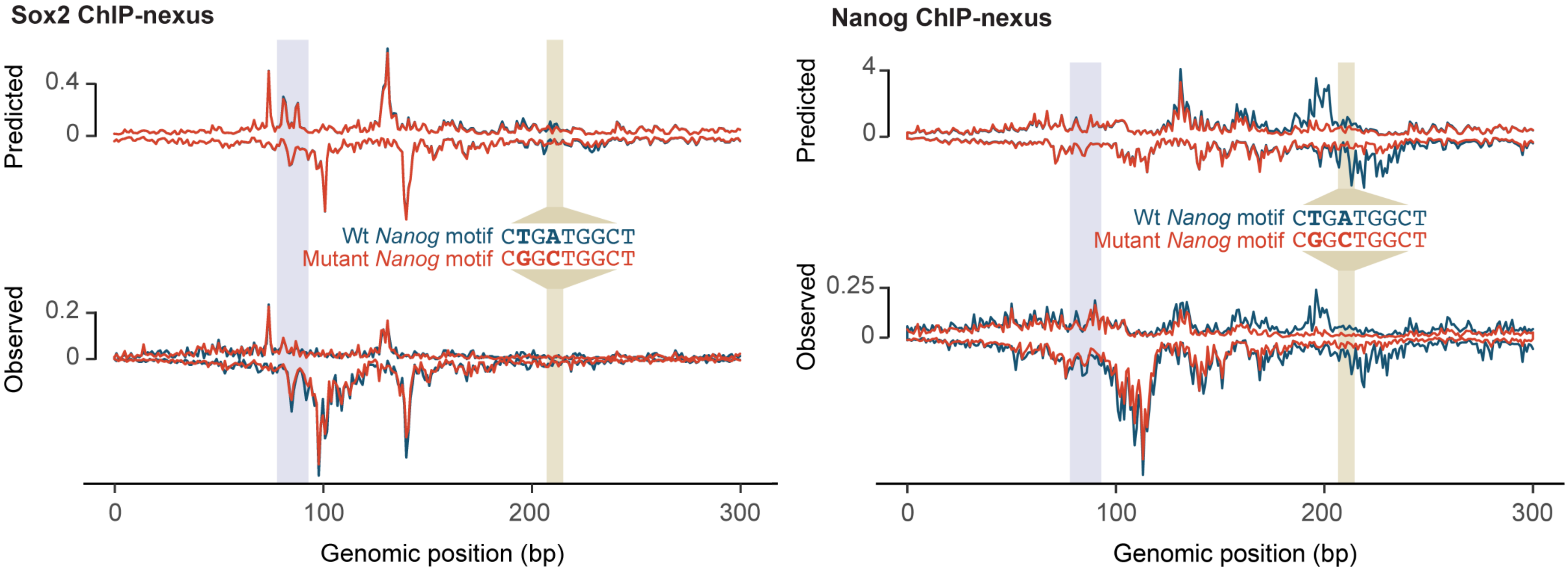
Nanog ChIP-nexus

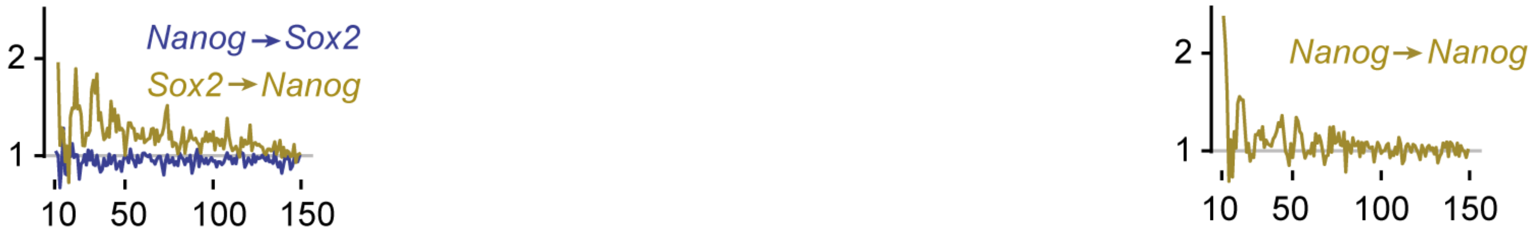




Julia Zeitlinger, Sabrina Krueger, Melanie Weilert

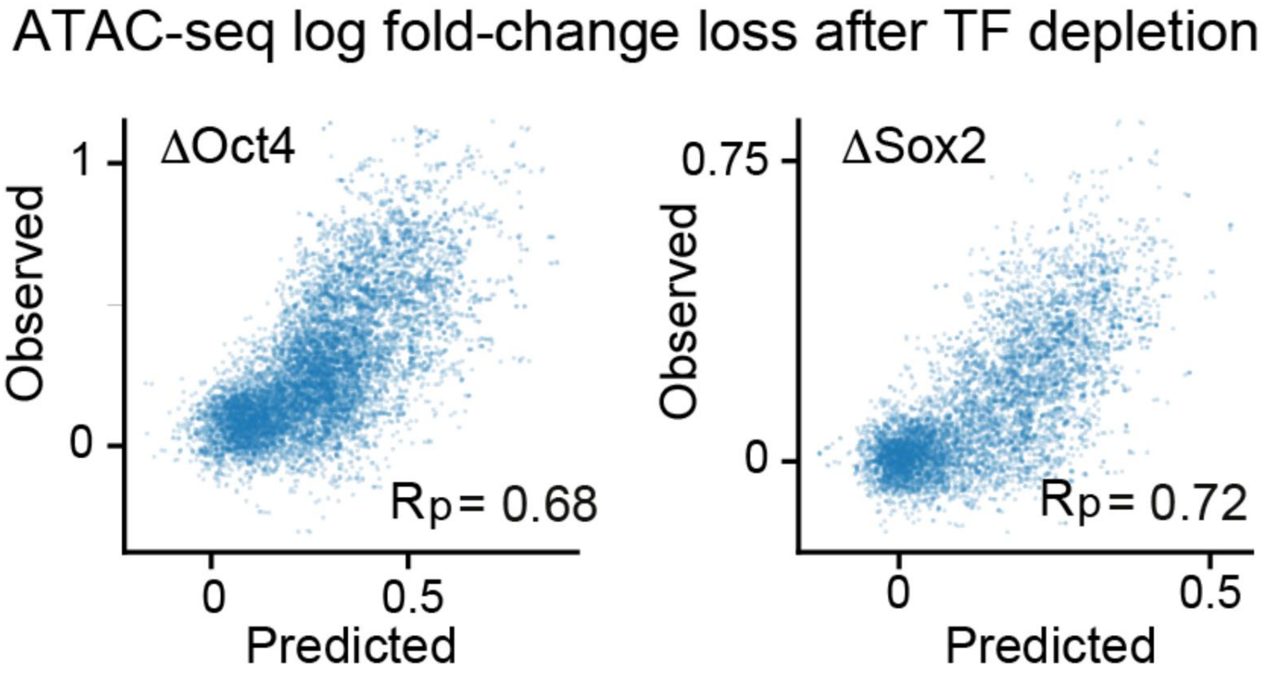
CRISPR mutations validate motif syntax Nanog <> Sox2





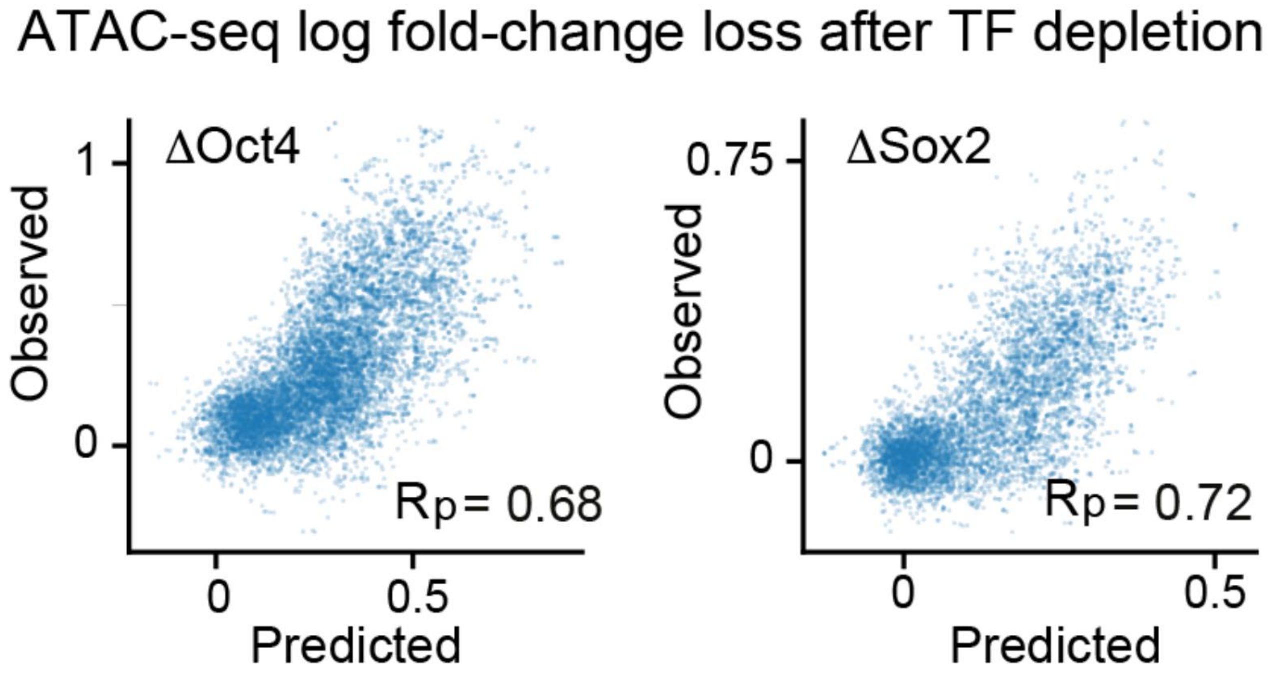
Julia Zeitlinger, Sabrina Krueger, Melanie Weilert

Binding syntax is predictive of differential accessibility after TF depletion & reporter expression

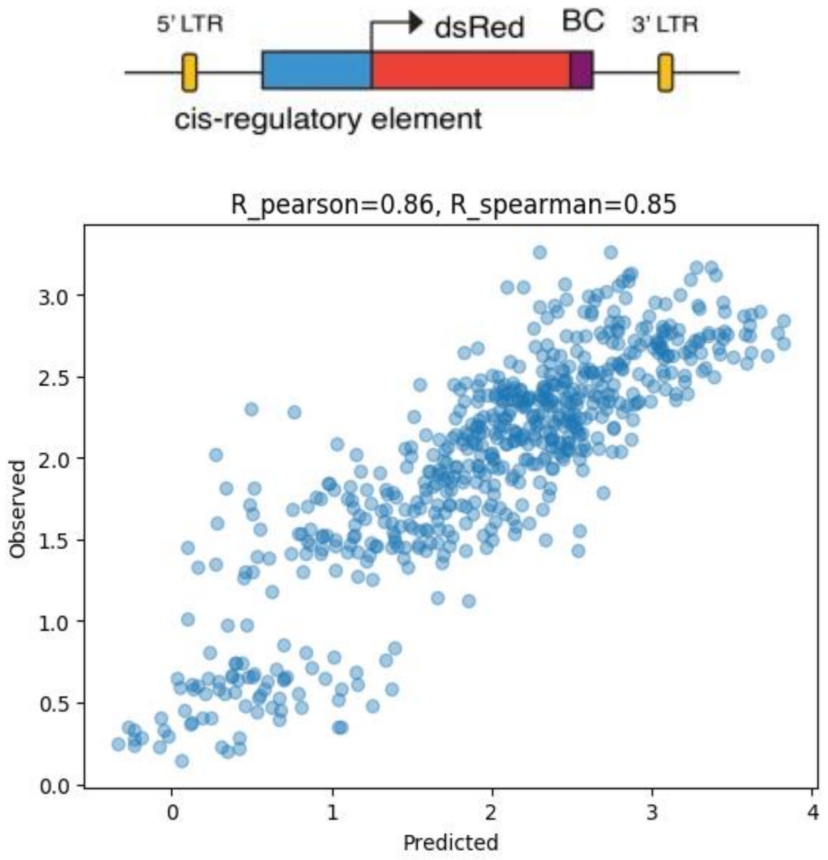


(Independent previously published data from Friman et al. 2019)

Binding syntax is predictive of differential accessibility after TF depletion & reporter expression

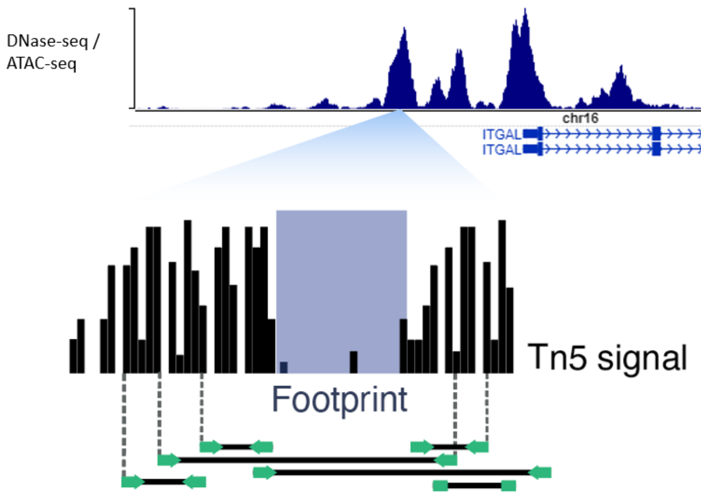


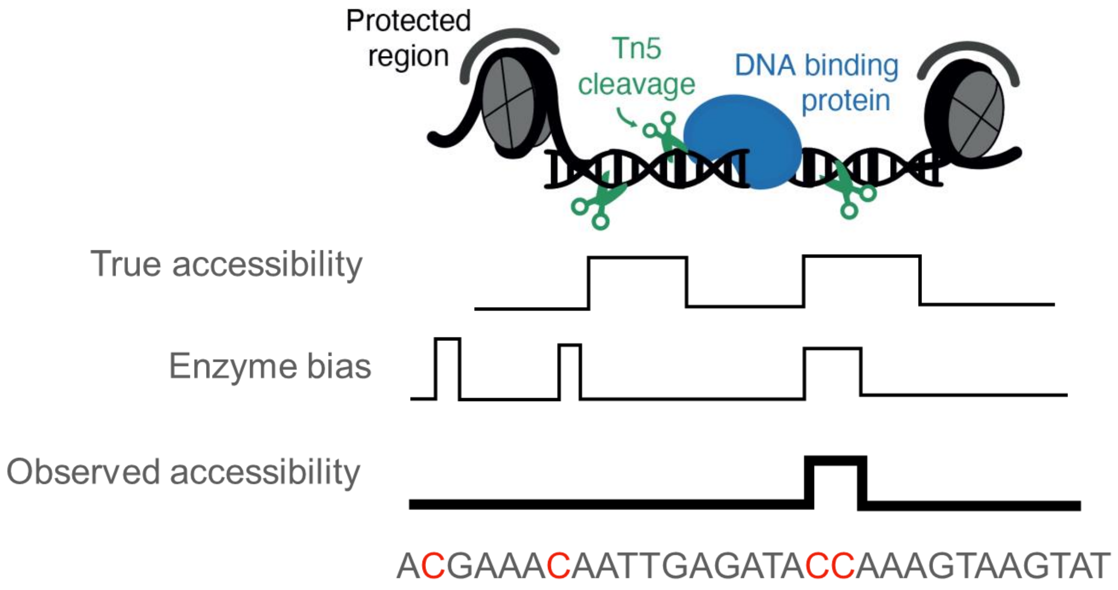
(Independent previously published data from Friman et al. 2019)

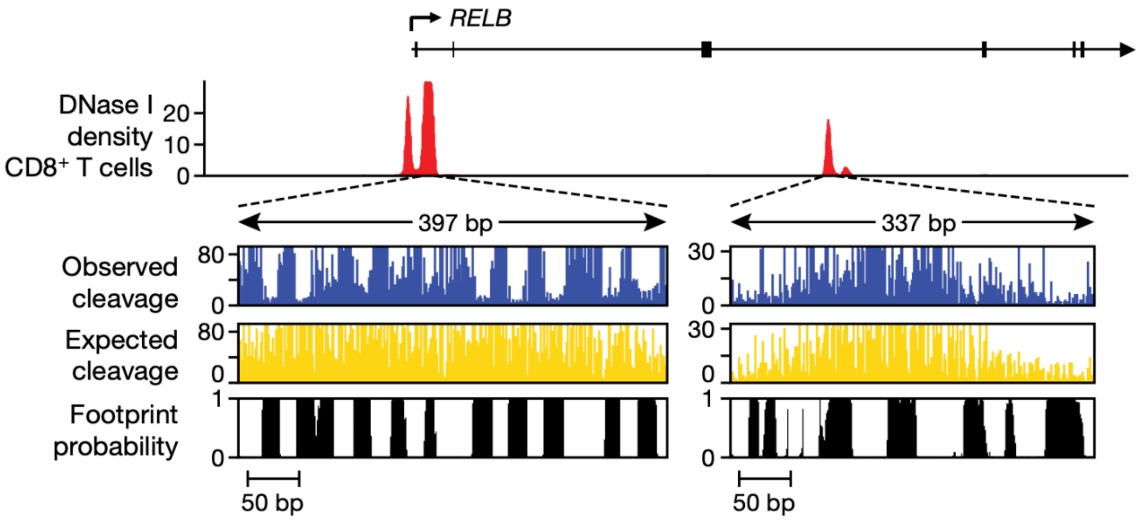


(Independent published MPRA data from King, Maricque, Cohen 2018)

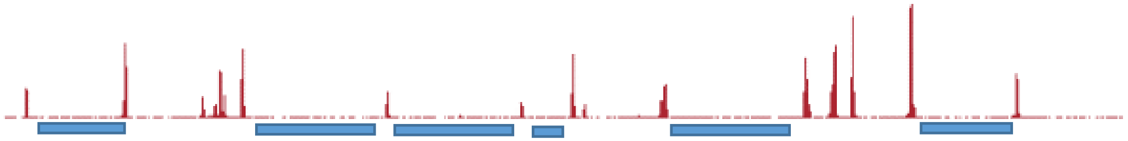
Modeling ATAC-seq / DNase-seq profiles (enzyme bias affects footprints)

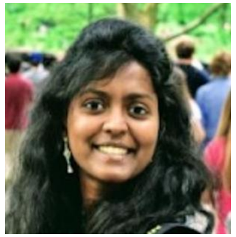






How to estimate Tn5 / DNase bias?

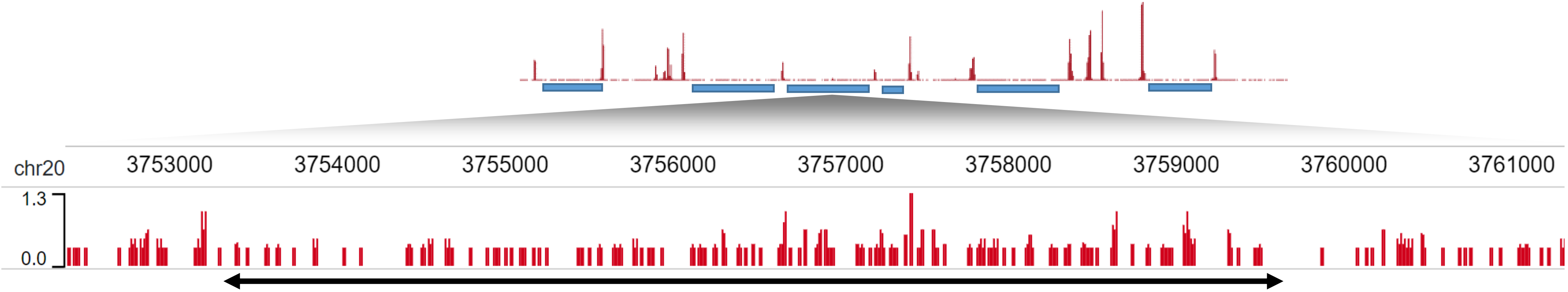




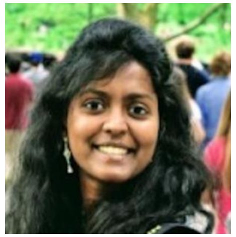
Anusri Pampari

Anna Shcherbina

How to estimate Tn5 / DNase bias?



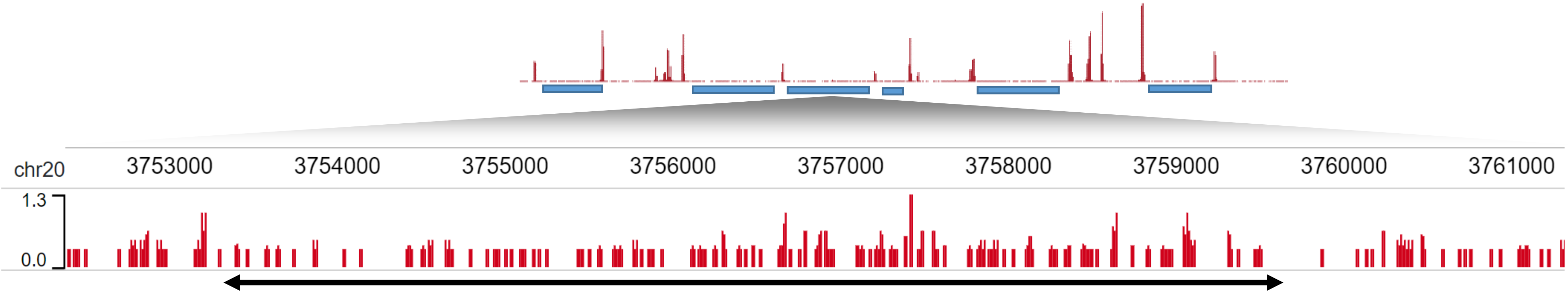
Read distribution in background regions is a function of enzyme bias



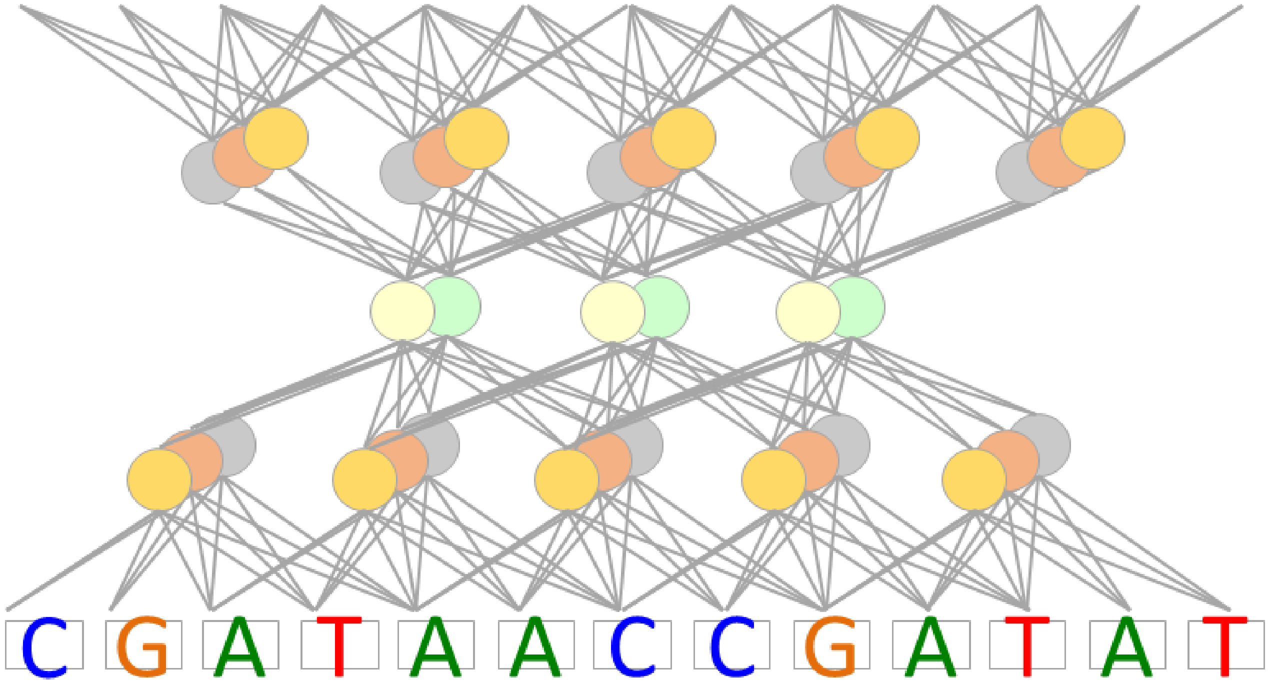
Anusri Pampari

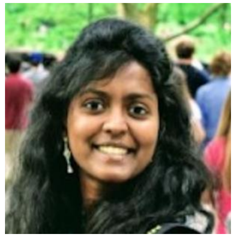
Anna Shcherbina

How to estimate Tn5 / DNase bias?



Read distribution in background regions is a function of enzyme bias

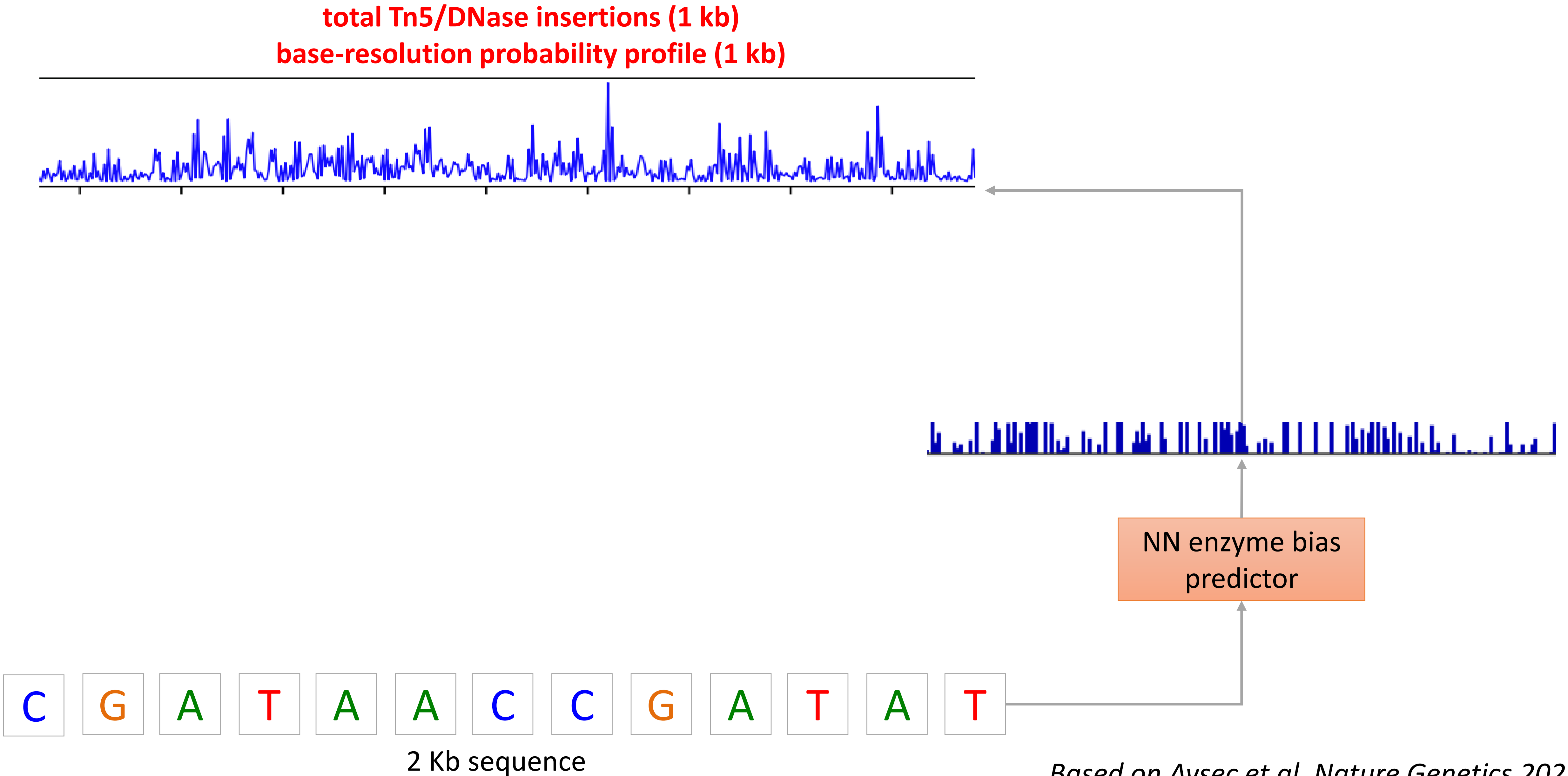




Anusri Pampari

Anna Shcherbina

ChromBPNet: Sequence to base-res chromatin accessibility profiles



total Tn5/DNase insertions (1 kb)

base-resolution probability profile (1 kb)

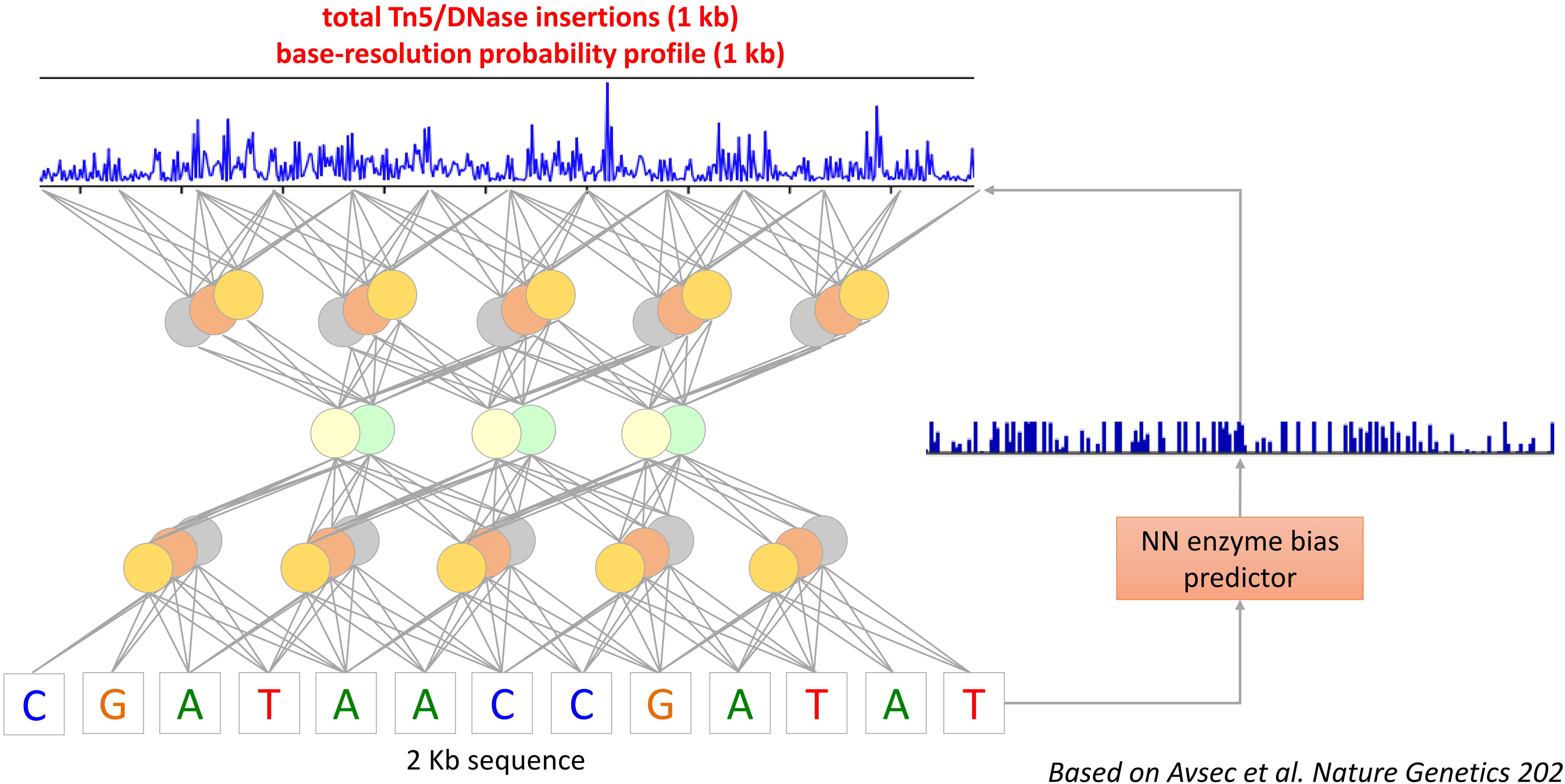
NN enzyme bias predictor

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

2 Kb sequence

Based on Avsec et al. Nature Genetics 2021

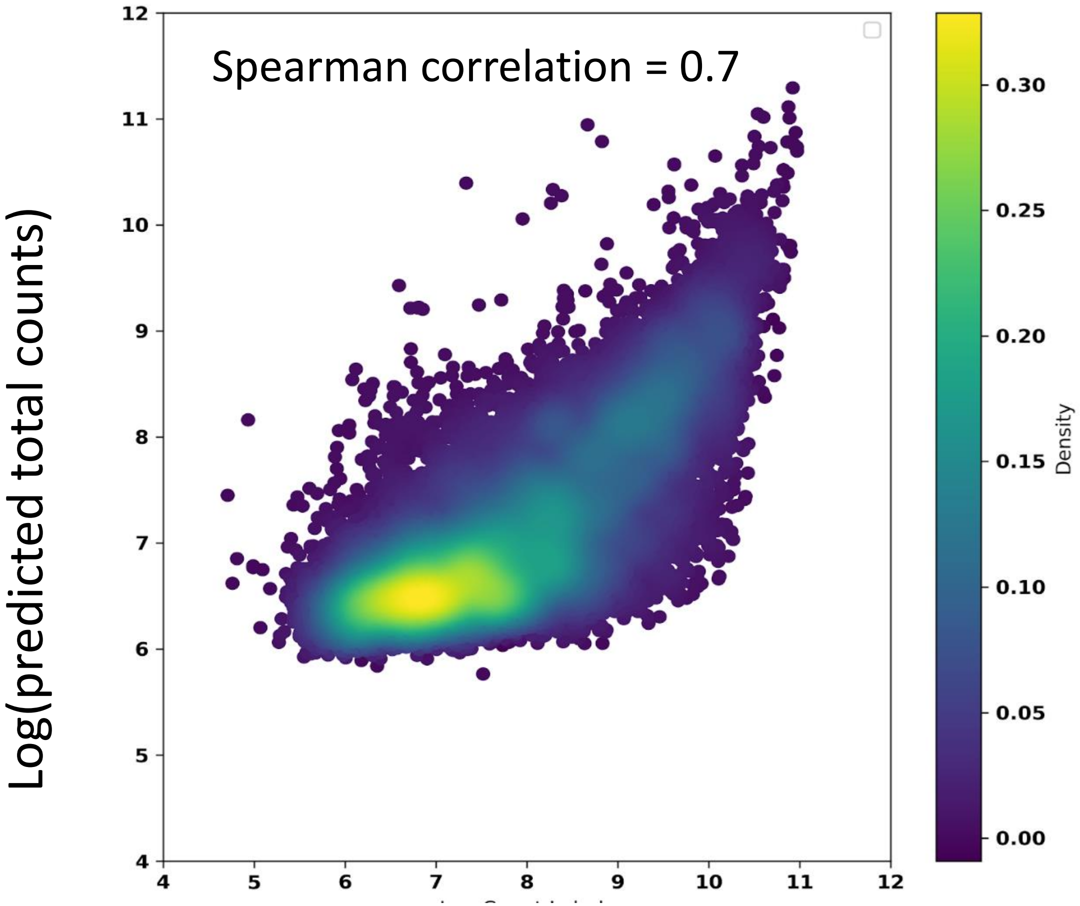
ChromBPNet: Sequence to base-res chromatin accessibility profiles


total Tn5/DNase insertions (1 kb)
base-resolution probability profile (1 kb)
NN enzyme bias predictor
C
G
A
T
A
A
C
C
G
A
T
A
T
2 Kb sequence

Based on Avsec et al. Nature Genetics 2021

Prediction performance (held-out chromosomes)

Total counts prediction performance

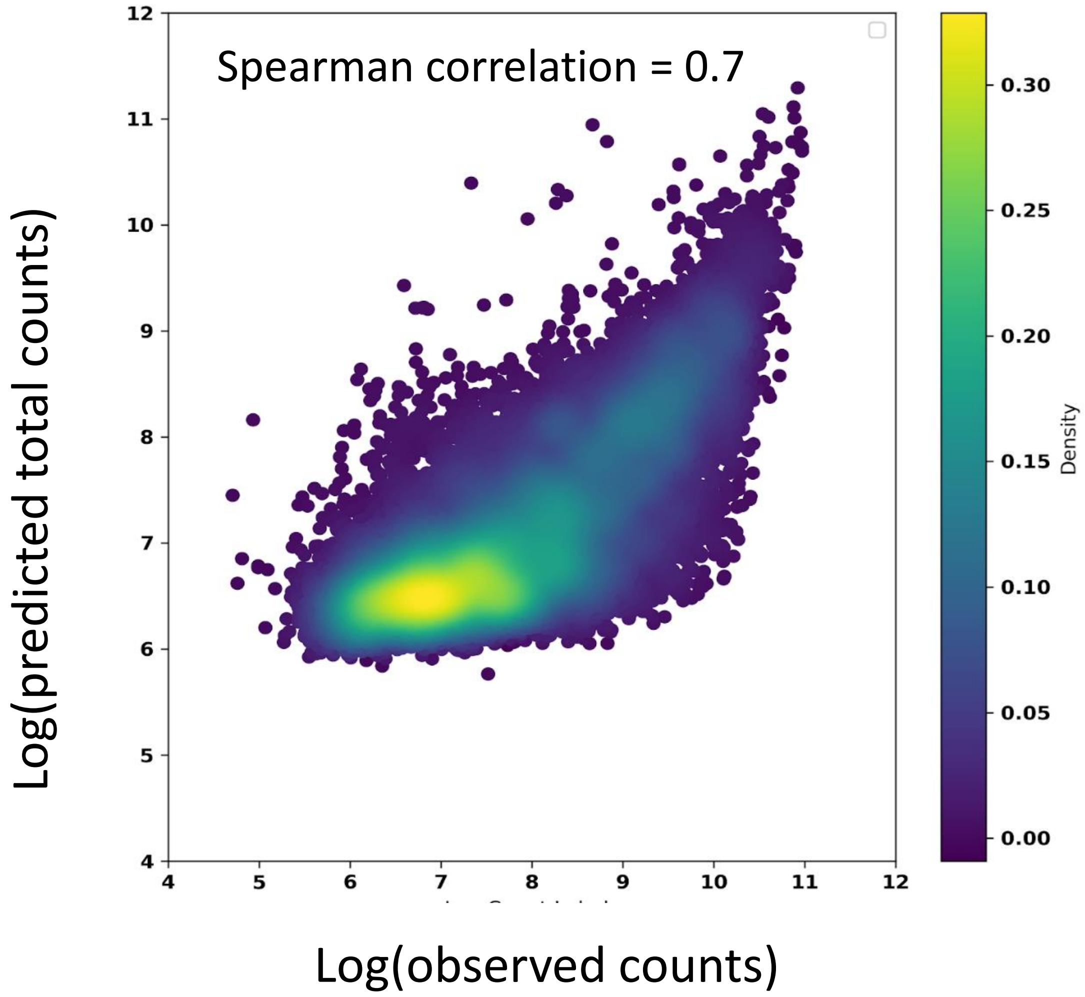


Log(predicted total counts)

Log(observed counts)

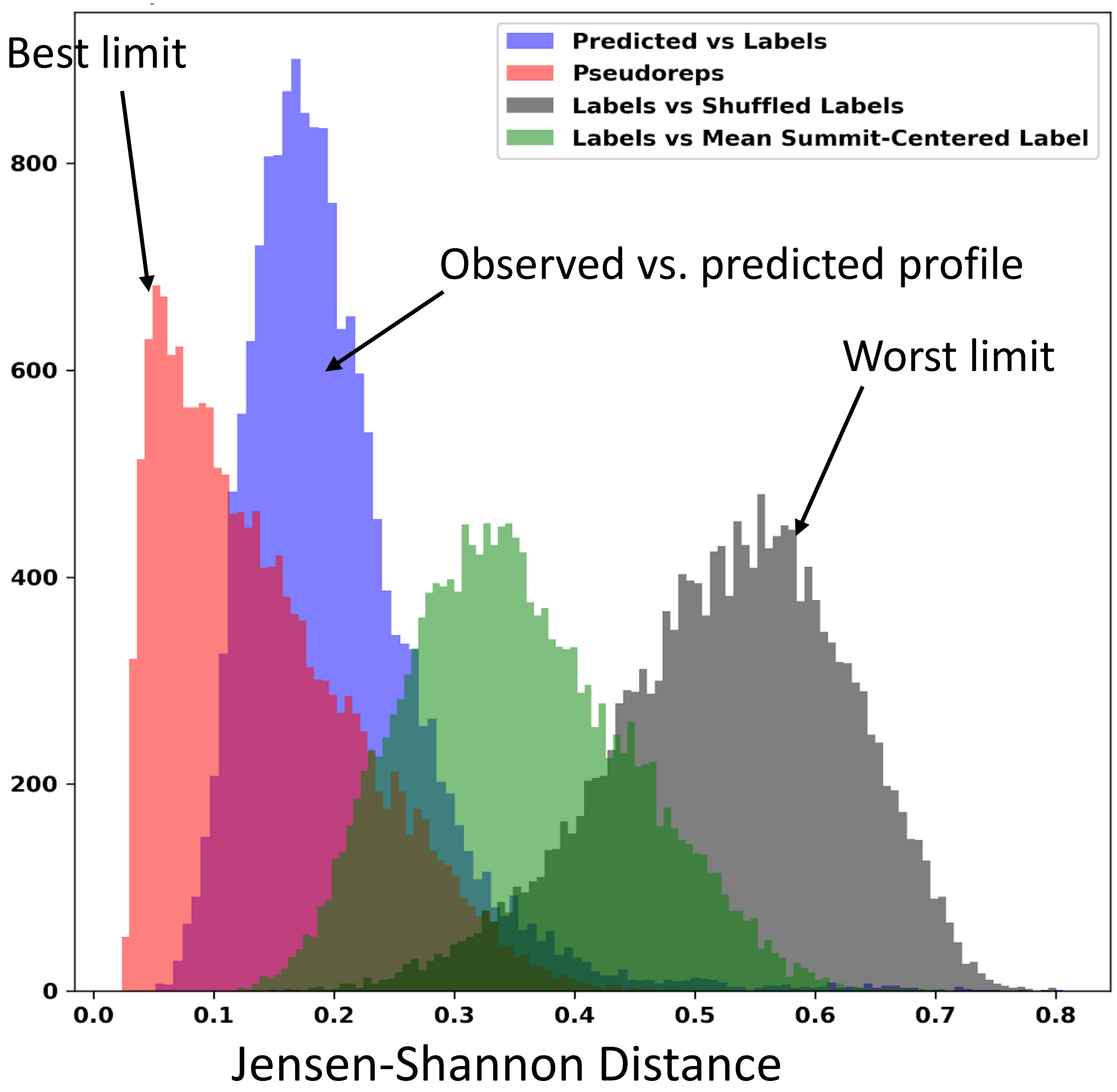
Prediction performance (held-out chromosomes)

Total counts prediction performance



Spearman correlation = 0.7

Profile prediction performance



Best limit

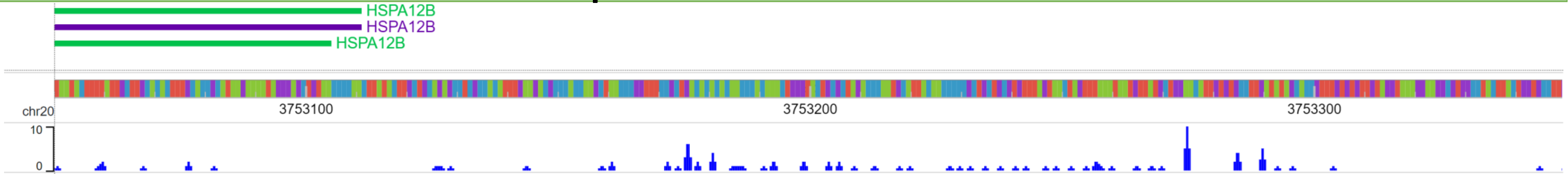
Observed vs. predicted profile

Worst limit

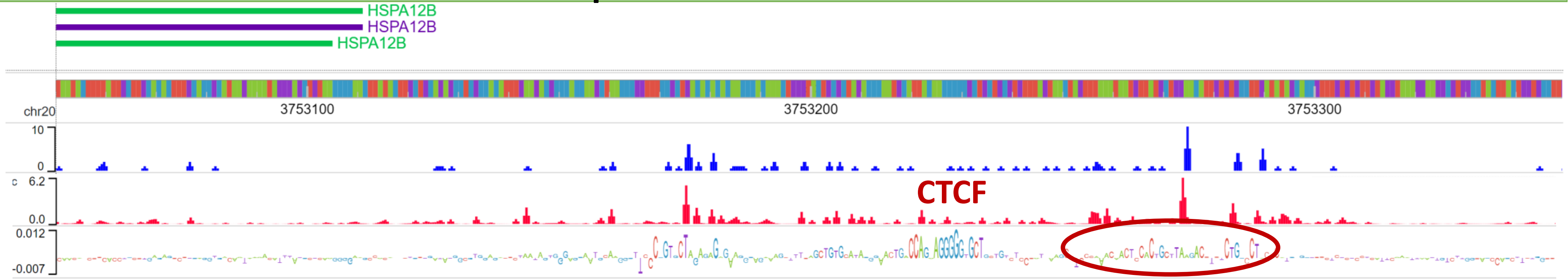
Denoised base-resolution bias-corrected chromatin accessibility footprints & de-biased sequence features

Denoised base-resolution bias-corrected chromatin accessibility footprints & de-biased sequence features

Observed track



Denoised base-resolution bias-corrected chromatin accessibility footprints & de-biased sequence features



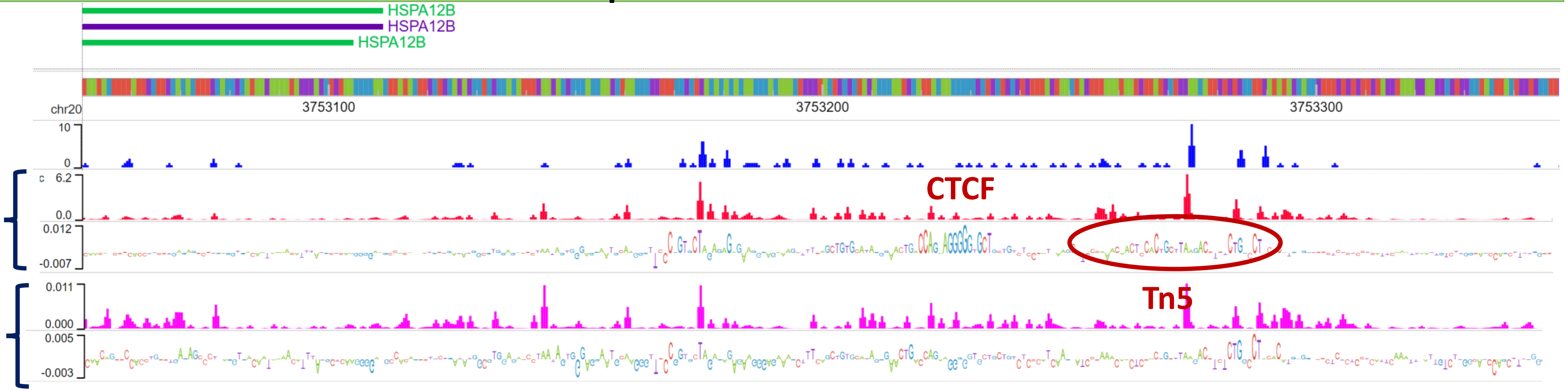
Observed track

ChromBPNet without Bias correction

CTCF

Tn5

Denoised base-resolution bias-corrected chromatin accessibility footprints & de-biased sequence features

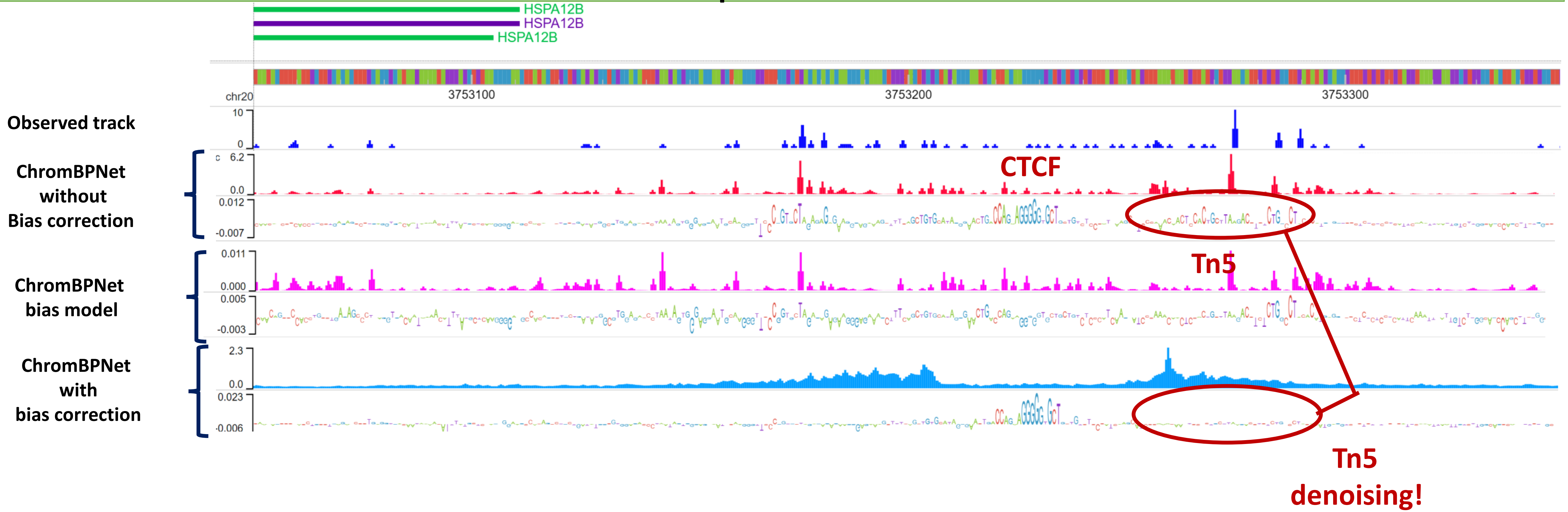


Observed track

ChromBPNet without Bias correction

ChromBPNet bias model

Denoised base-resolution bias-corrected chromatin accessibility footprints & de-biased sequence features



Observed track

ChromBPNet without Bias correction

ChromBPNet bias model

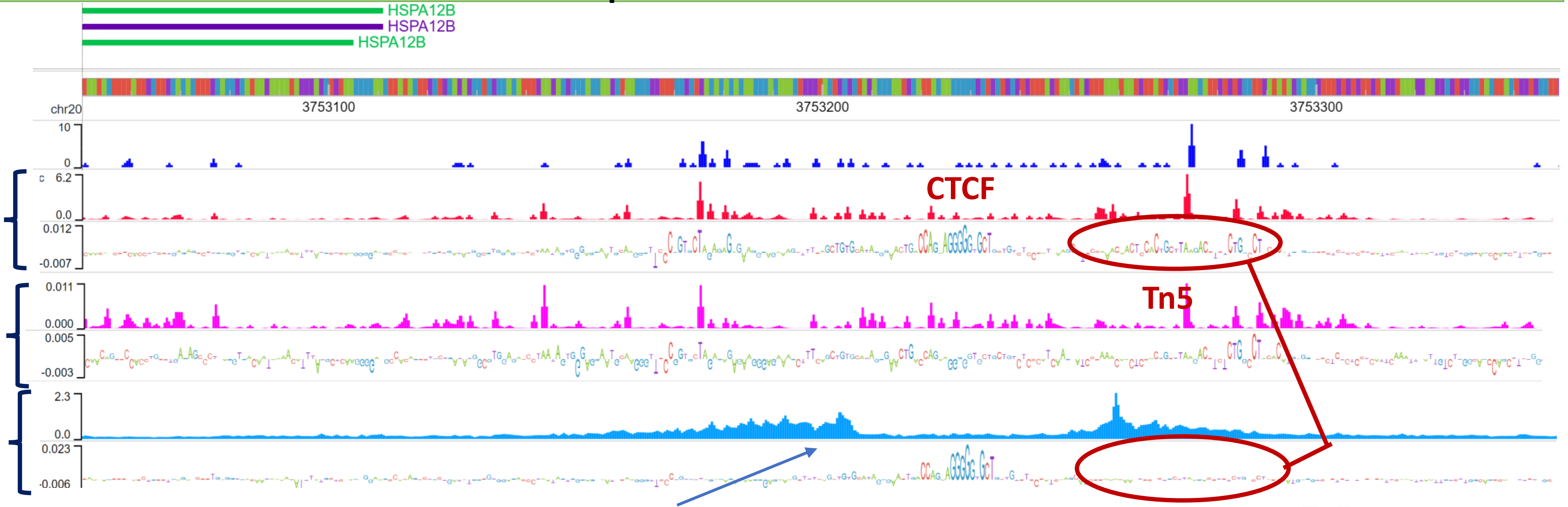
ChromBPNet with bias correction

CTCF

Tn5

Tn5 denoising!

Denoised base-resolution bias-corrected chromatin accessibility footprints & de-biased sequence features



Observed track

ChromBPNet without Bias correction

ChromBPNet bias model

ChromBPNet with bias correction

CTCF

Tn5

CTCF Footprint

Tn5 denoising!

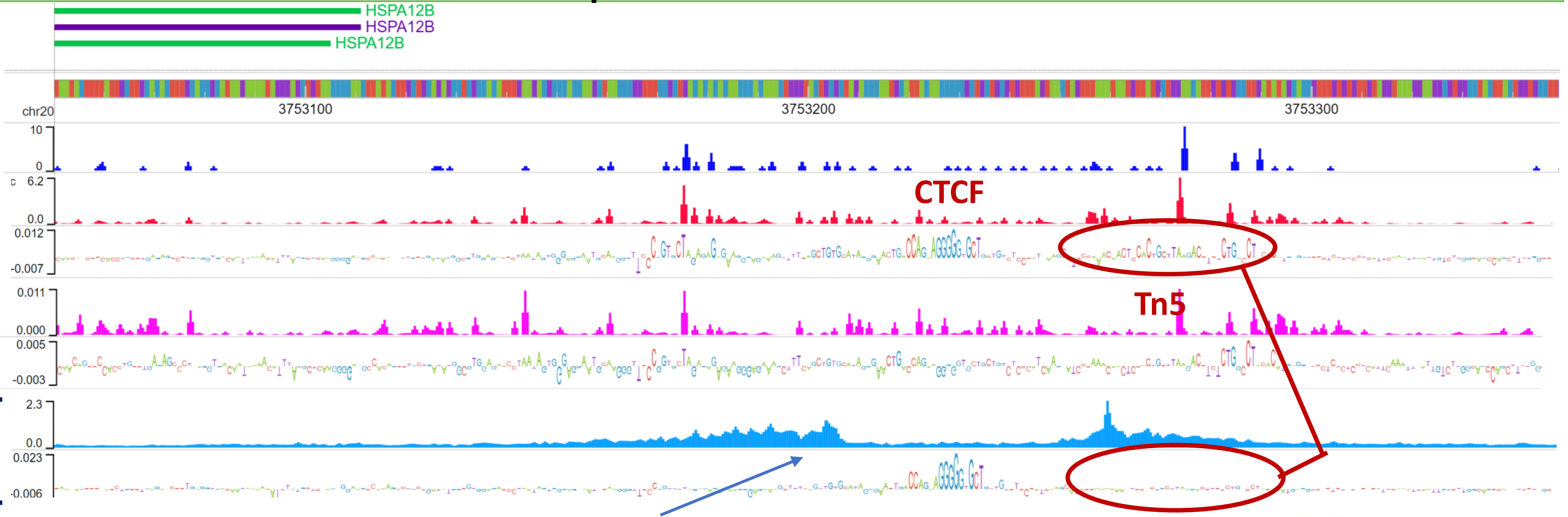
Denoised base-resolution bias-corrected chromatin accessibility footprints & de-biased sequence features

Observed track

ChromBPNet without Bias correction

ChromBPNet bias model

ChromBPNet with bias correction



CTCF

Tn5

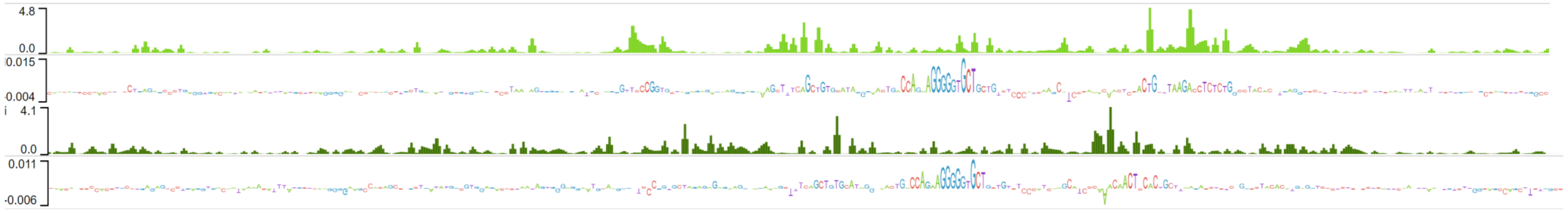
Tn5 denoising!

CTCF Footprint

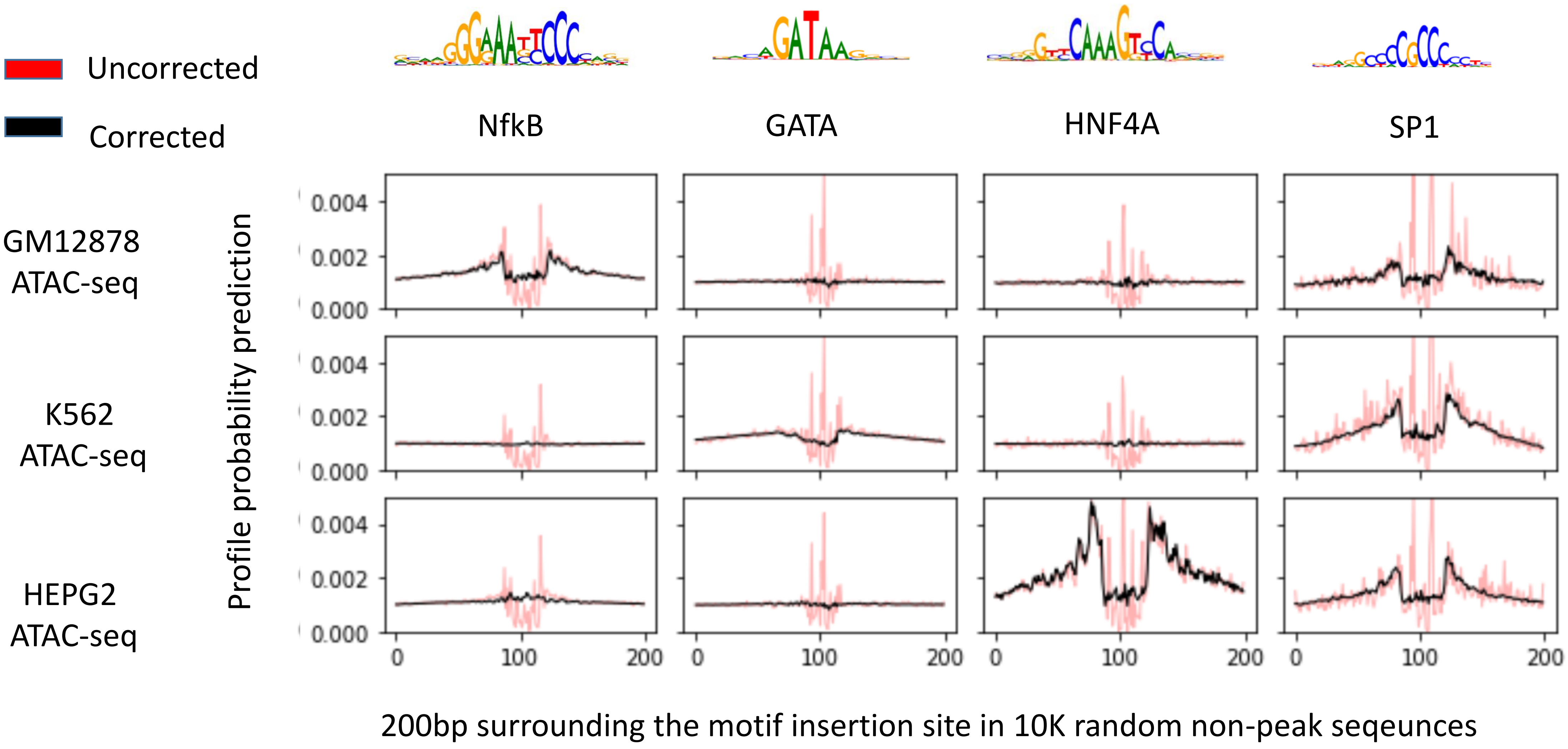
Using existing work (HINT-ATAC and TOBIAS)

ChromBPNet with HINT-ATAC correction

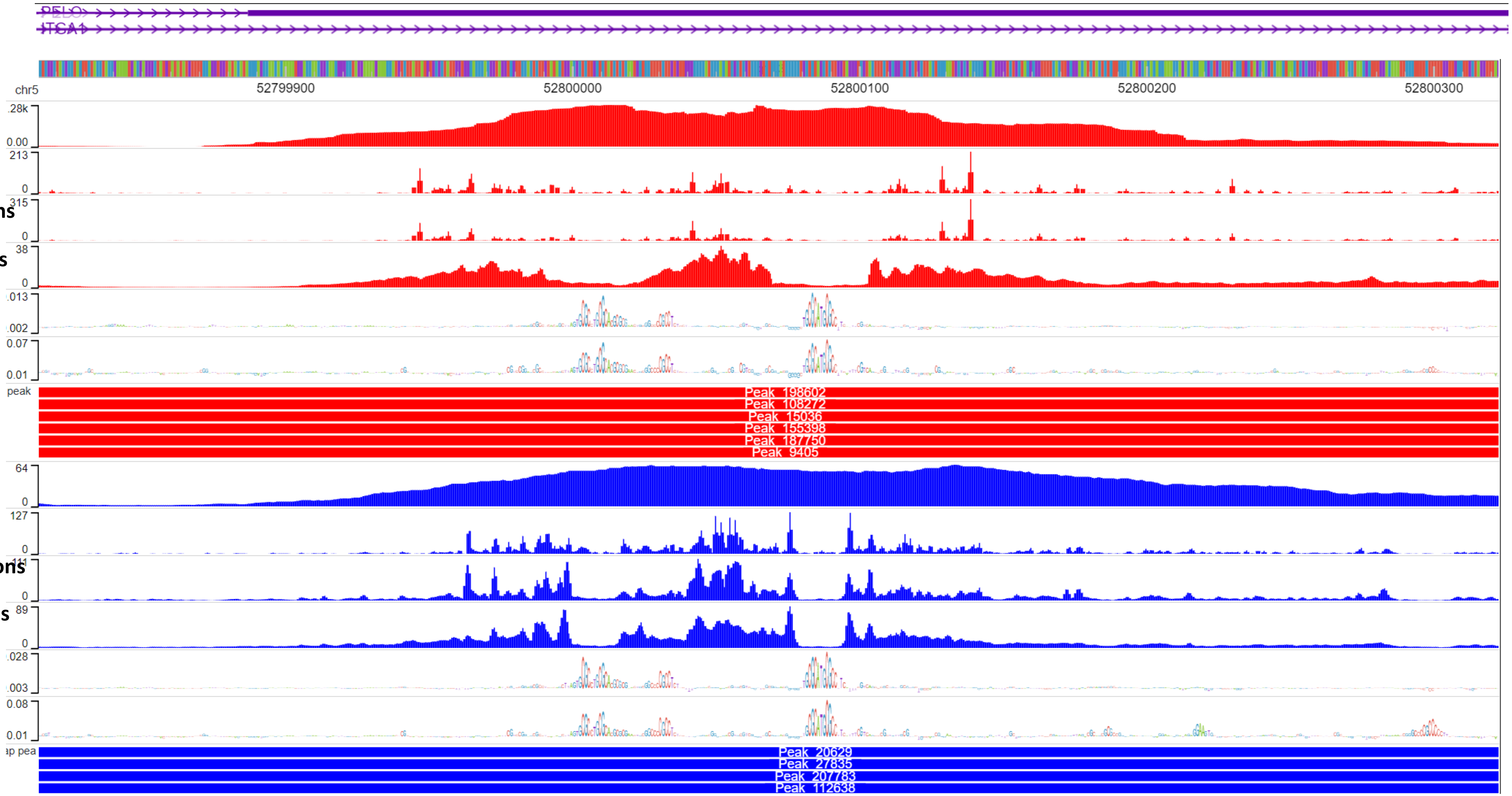
ChromBPNet With TOBIAS correction



ChromBPNet can predict marginal footprints of cell-type specific TFs



Similar sequence syntax derived from DNase-seq and ATAC-seq data



ATAC pvalue track

ATAC observed track

ATAC uncorrected predictions

ATAC corrected predictions

ATAC profile contribution

ATAC count contribution

ATAC overlap peak set

DNASE pvalue track

DNASE observed track

DNASE uncorrected predictions

DNASE corrected predictions

DNASE profile contribution

DNASE count contribution

DNASE overlap peak set

High fidelity denoising, imputation and interpretations at different read coverages

Beta-globin locus in K562

High fidelity denoising, imputation and interpretations at different read coverages

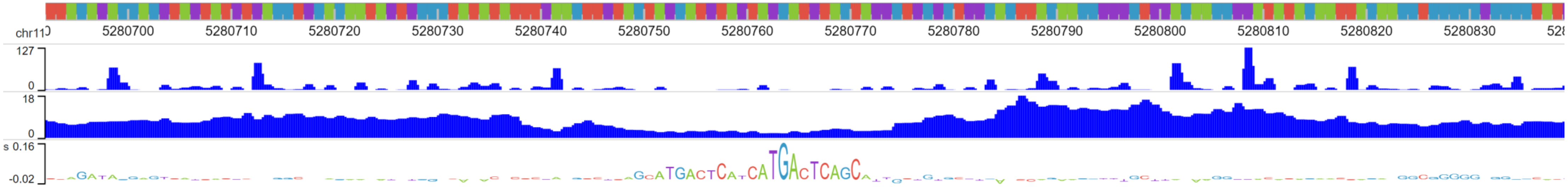
Beta-globin locus in K562

500M

Observed track

Bias corrected track

Contribution scores



High fidelity denoising, imputation and interpretations at different read coverages

Beta-globin locus in K562

500M

Observed track

Bias corrected track

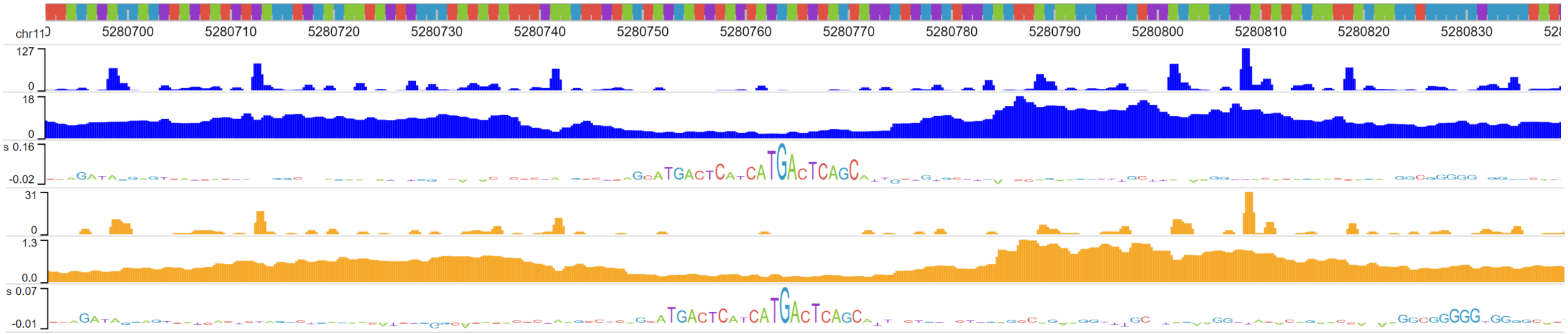
Contribution scores

100M

Observed track

Bias corrected track

Contribution scores



High fidelity denoising, imputation and interpretations at different read coverages

Beta-globin locus in K562

500M

Observed track

Bias corrected track

Contribution scores

100M

Observed track

Bias corrected track

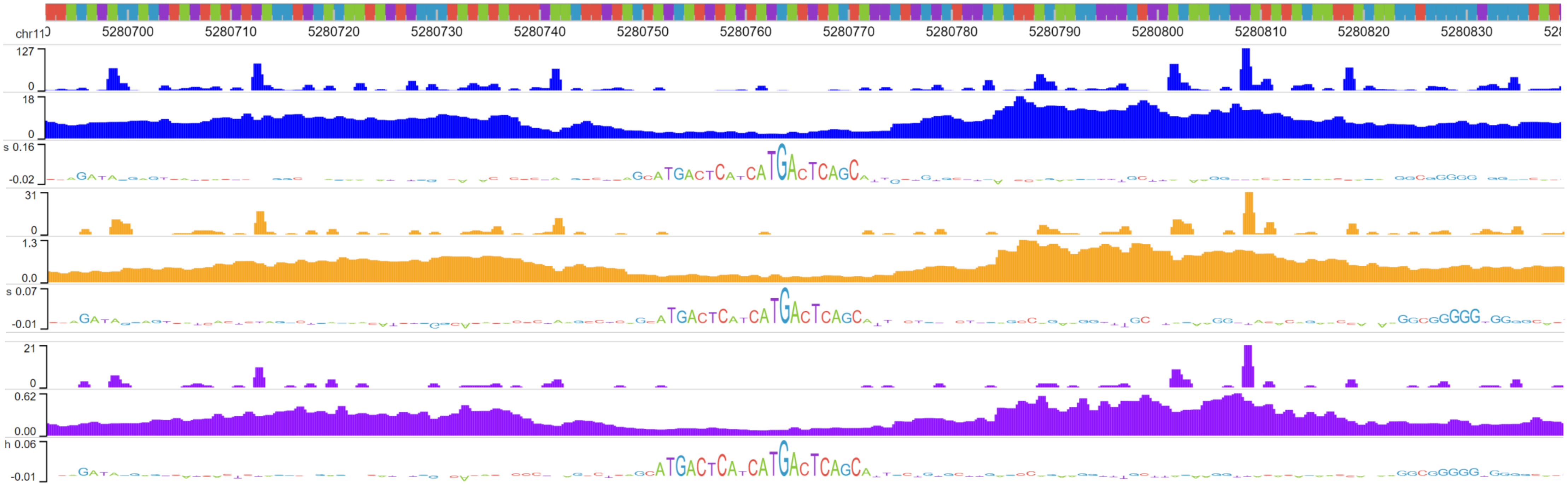
Contribution scores

50M

Observed track

Bias corrected track

Contribution scores



High fidelity denoising, imputation and interpretations at different read coverages

Beta-globin locus in K562

500M

Observed track

Bias corrected track

Contribution scores

100M

Observed track

Bias corrected track

Contribution scores

50M

Observed track

Bias corrected track

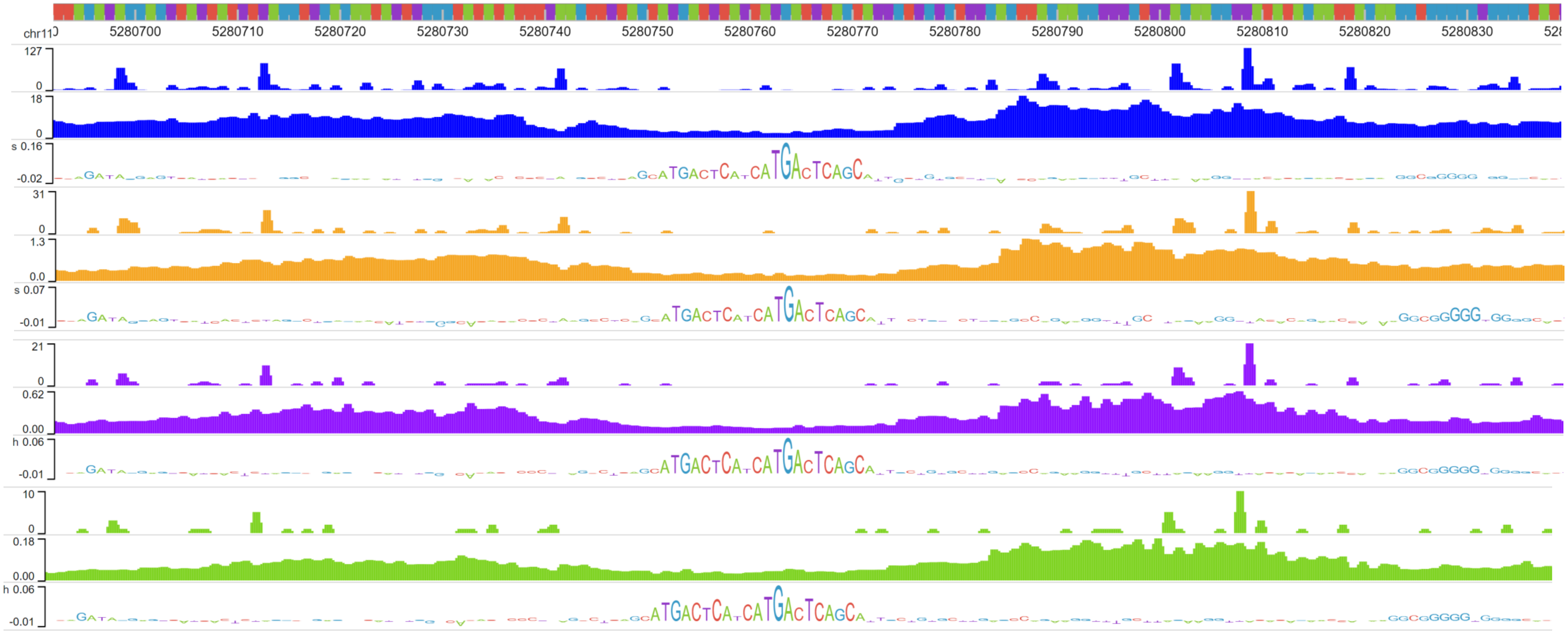
Contribution scores

25M

Observed track

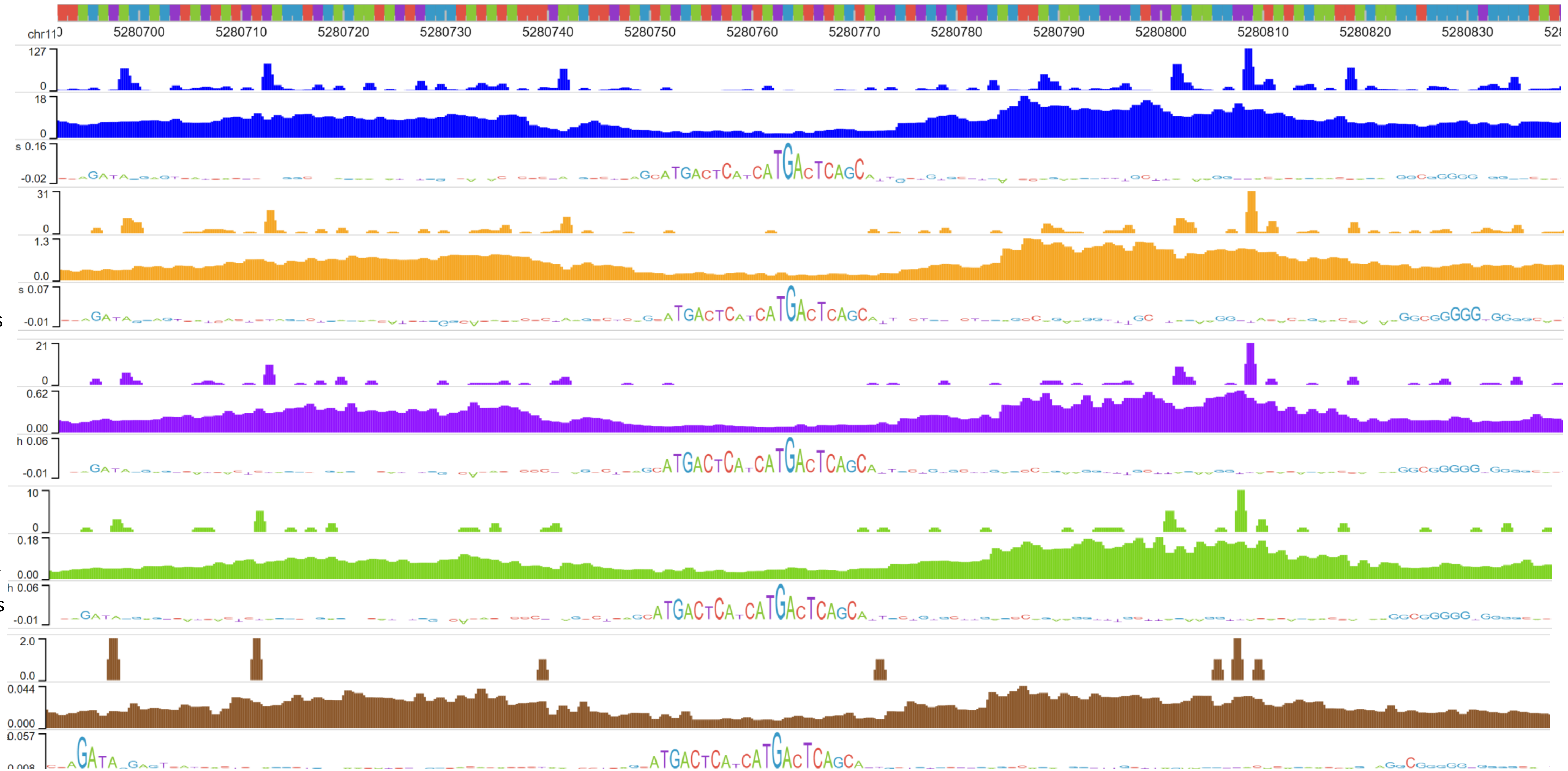
Bias corrected track

Contribution scores

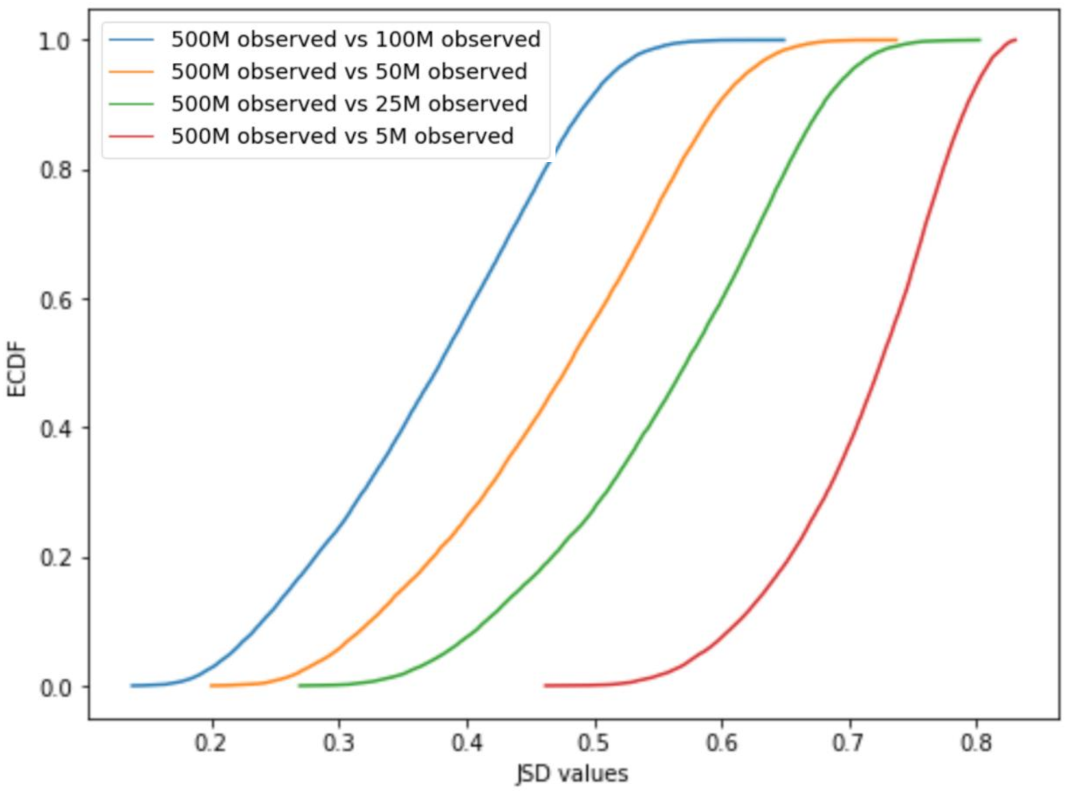


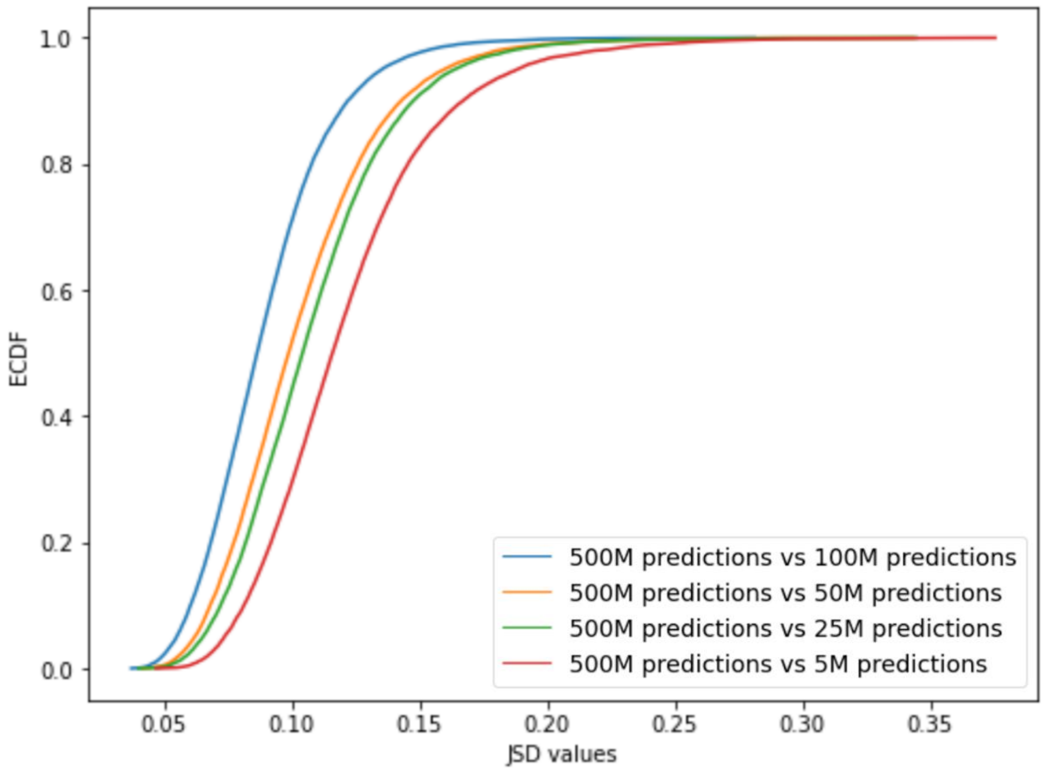
High fidelity denoising, imputation and interpretations at different read coverages

Beta-globin locus in K562



ChromBPNet predicted tracks are substantially similar compared to observed tracks at different read depths



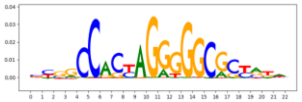


Using 500M as ground truth we compare degradation in signal quality at different read depths

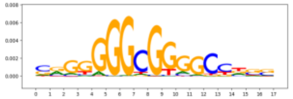
ChromBPNet predicts substantially similar profiles compared to the observed tracks

High fidelity marginal footprinting in K562 at different read depths

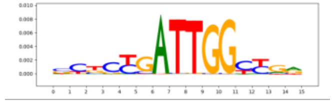
High fidelity marginal footprinting in K562 at different read depths



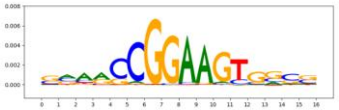
CTCF



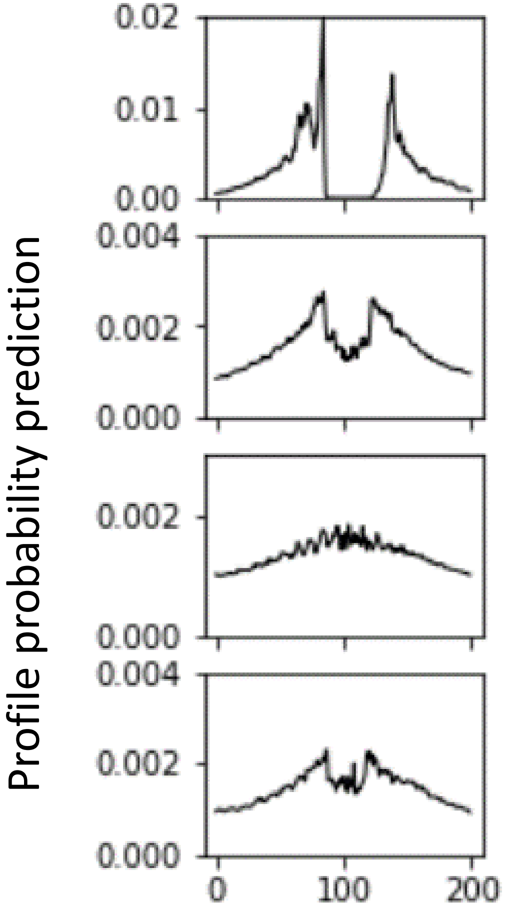
SPI1



NFYB



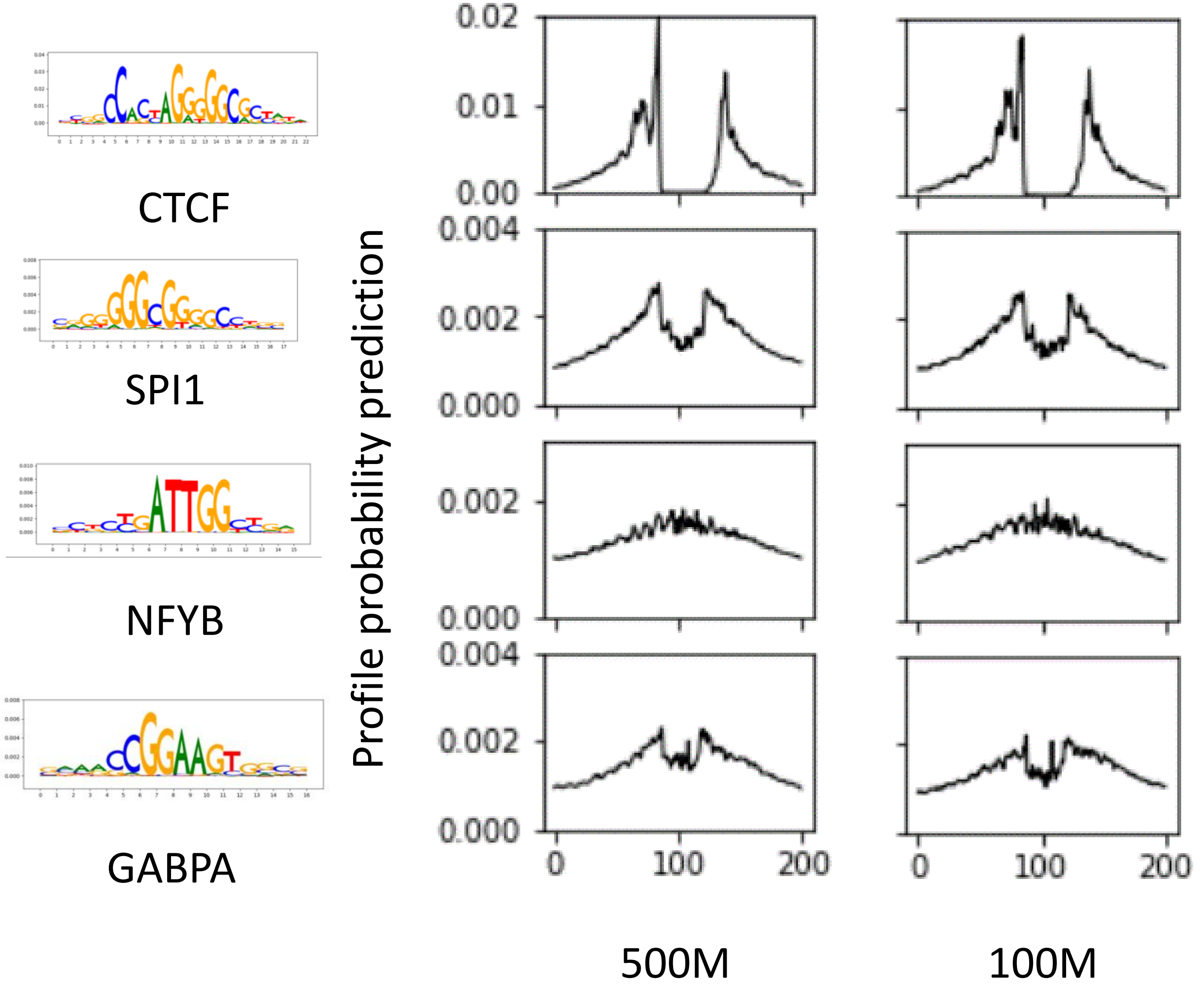
GABPA



500M

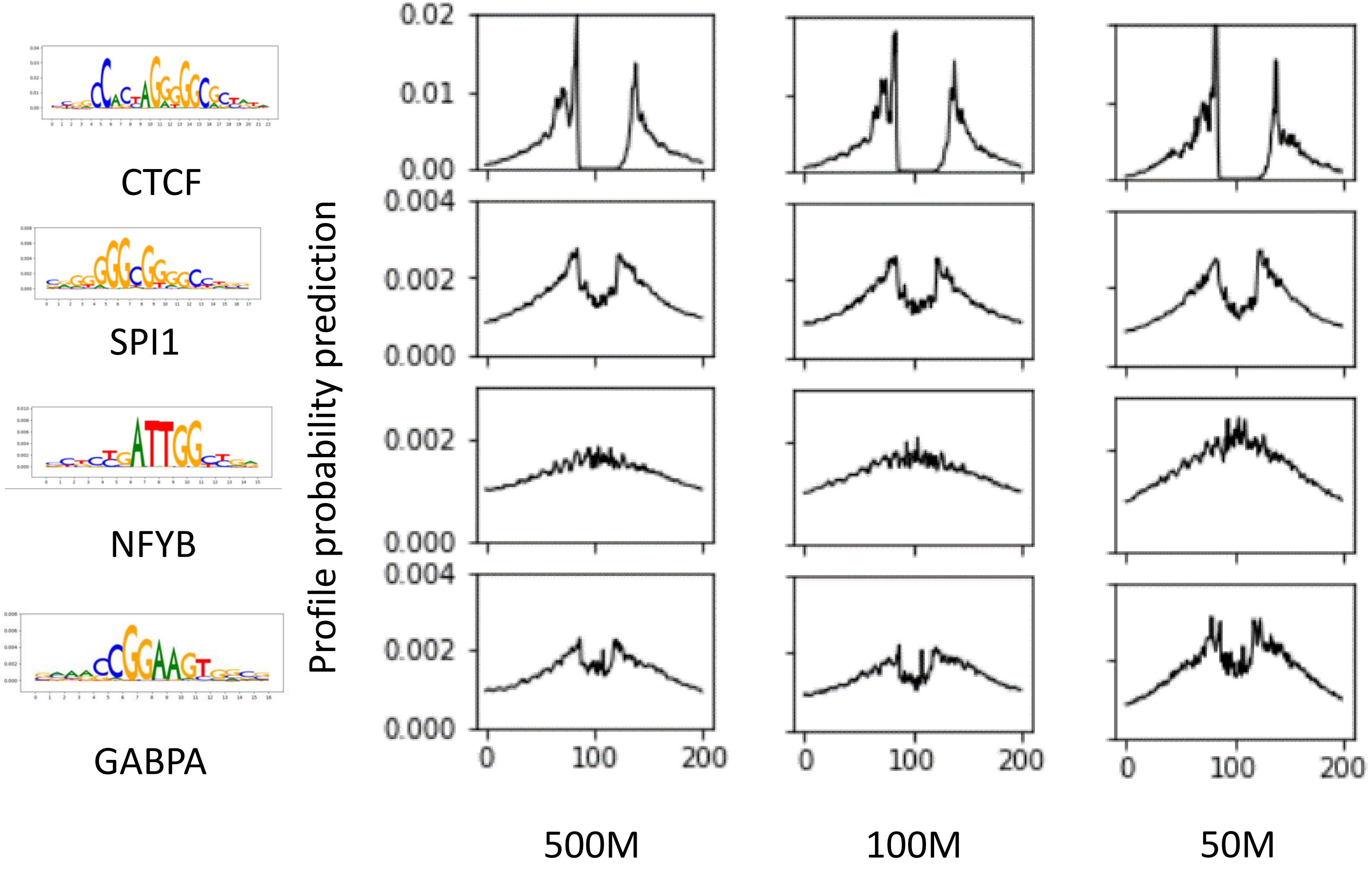
200bp surrounding the motif insertion site

High fidelity marginal footprinting in K562 at different read depths



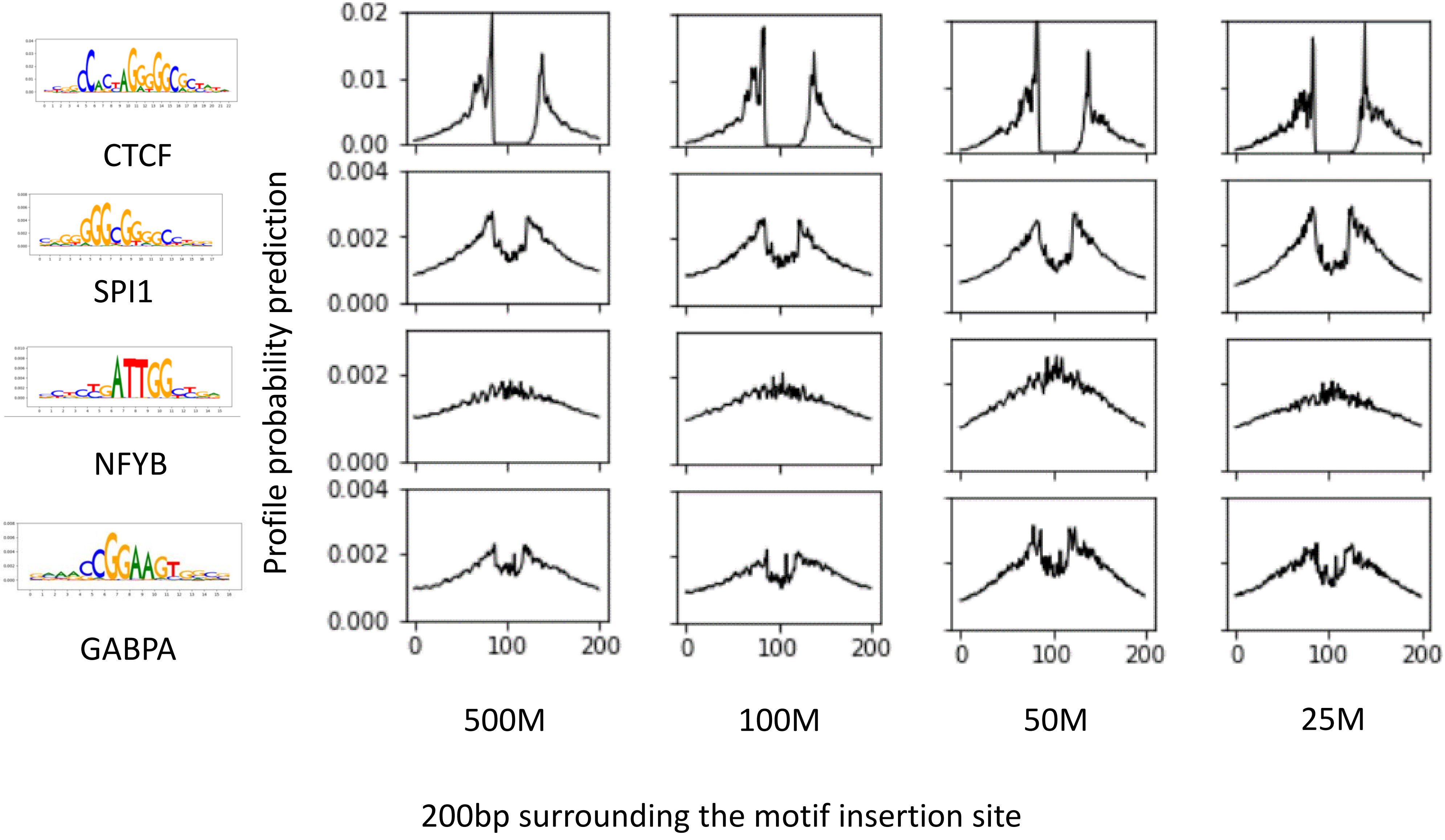
200bp surrounding the motif insertion site

High fidelity marginal footprinting in K562 at different read depths

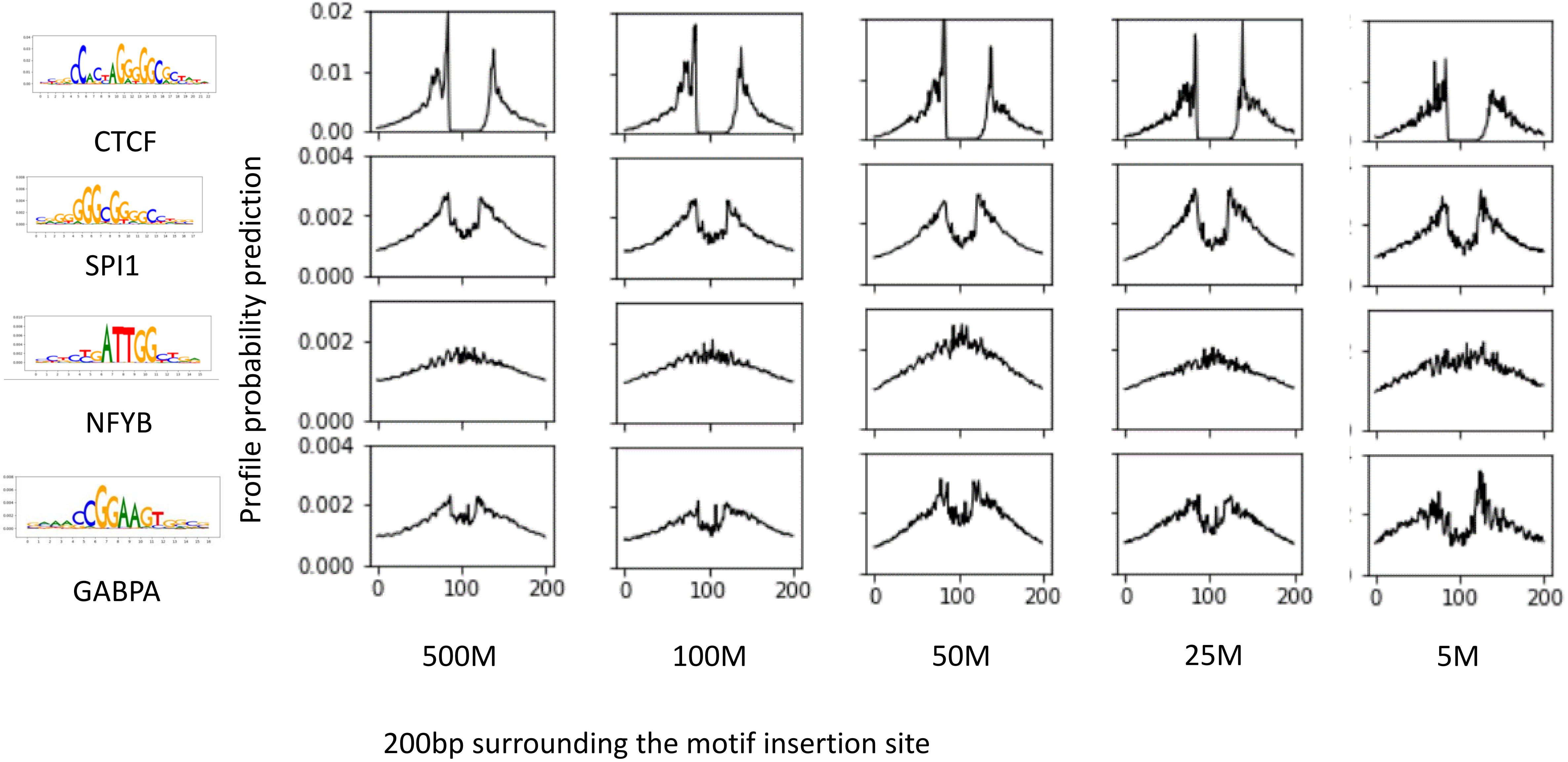


200bp surrounding the motif insertion site

High fidelity marginal footprinting in K562 at different read depths



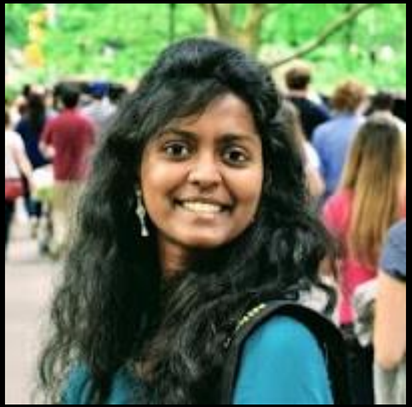
High fidelity marginal footprinting in K562 at different read depths



Model driven prioritization and interpretation of non-coding genetic variation

Anna Shcherbina

Soumya Kundu



Anusri Pampari

Laksshman Sundaram

Large proportion of disease-associated genetic loci are non-coding

Benign

.......ACTGATCG**C**AATCG.......

.......ACTGATCG**G**AATCG.......

Risk

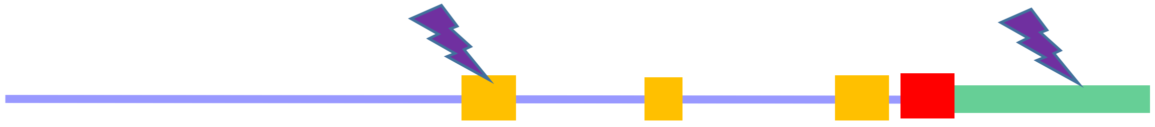
Large proportion of disease-associated genetic loci are non-coding

Benign

…….ACTGATCG**C**AATCG…….

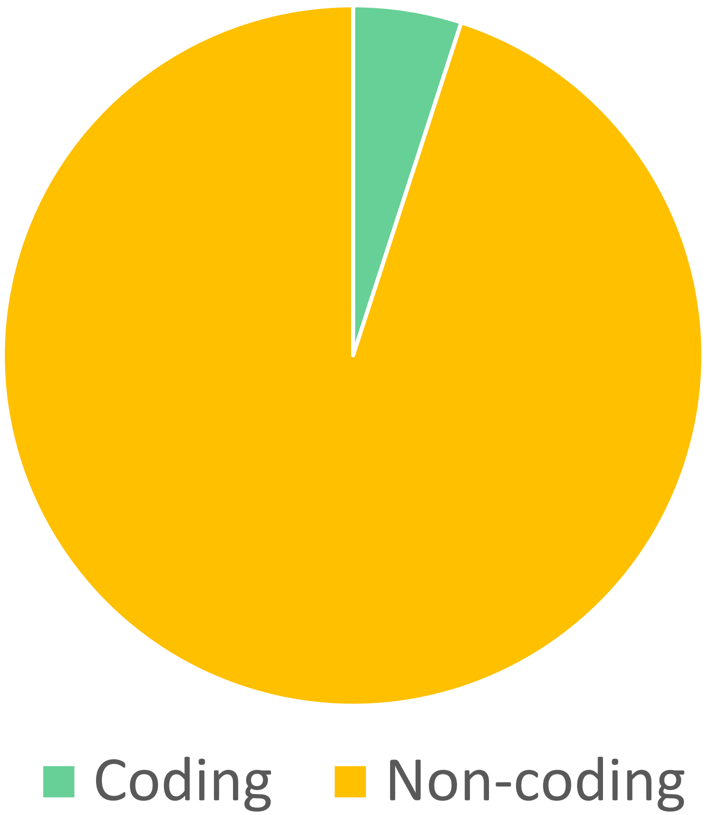
…….ACTGATCG**G**AATCG…….

Risk



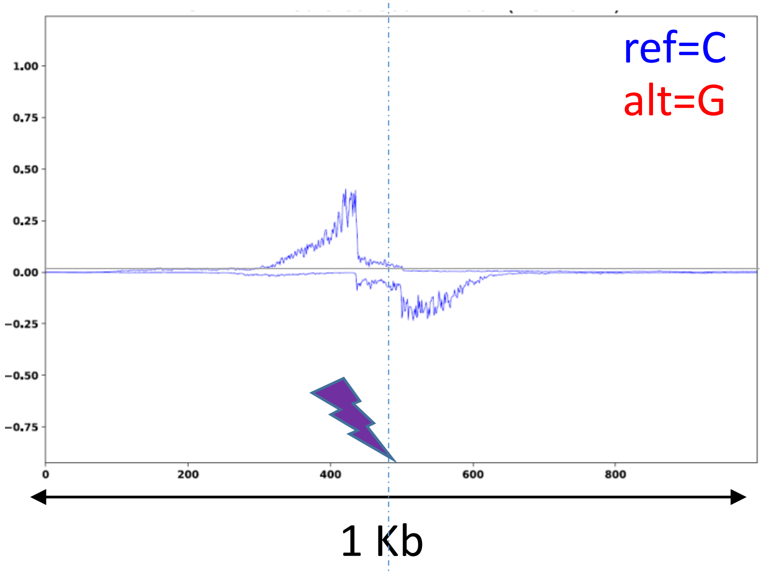
Control elements
(**Non-coding variants**)

Gene
(Coding variant)

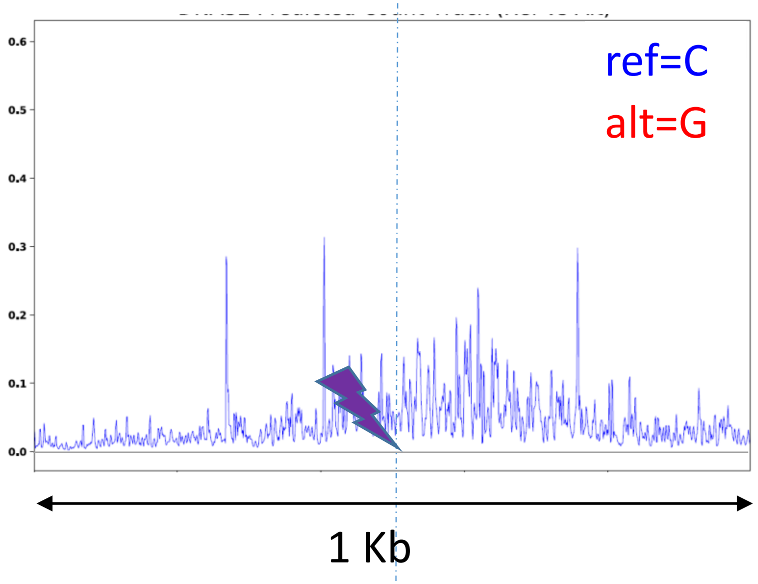


BPNet/ChromBPNet can predict variants influencing regulatory activity

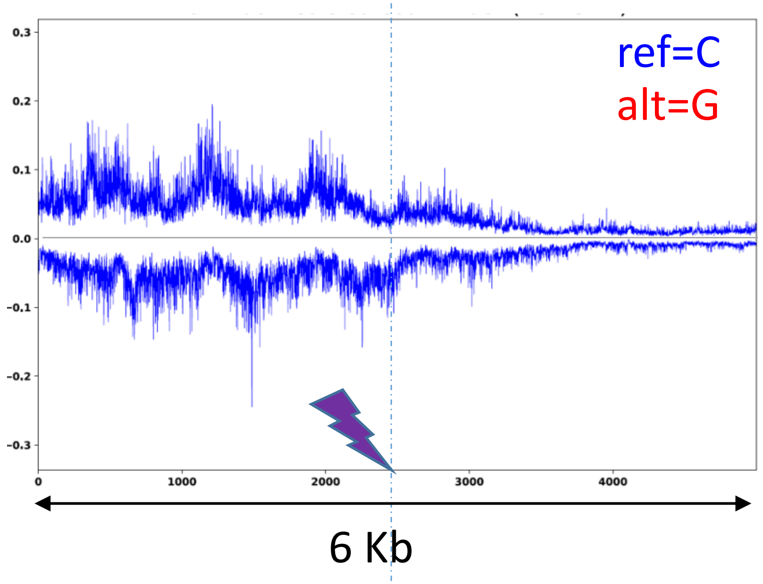
Predicted SPI1 protein-DNA binding



Predicted chromatin accessibility

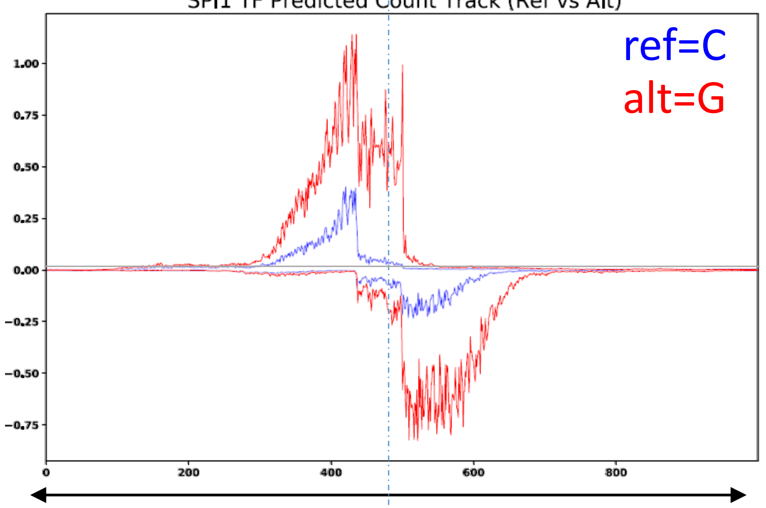


Predicted histone mark (H3K27ac)



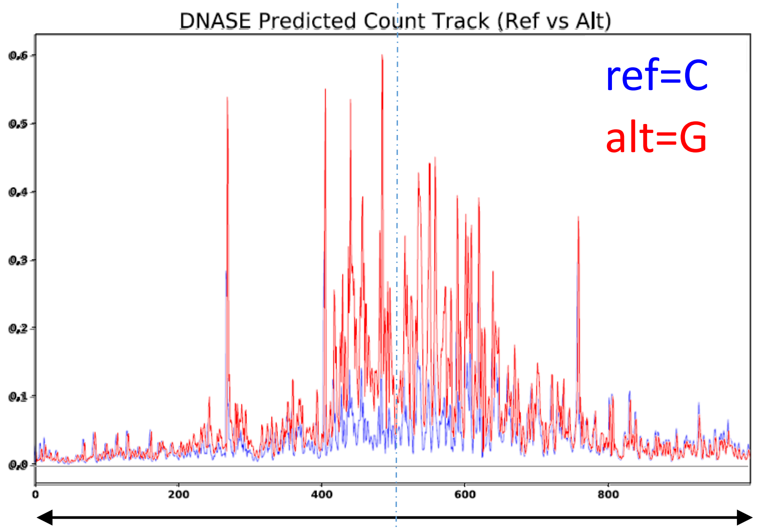
BPNet/ChromBPNet can predict variants influencing regulatory activity

Predicted SPI1 TF ChIP-seq



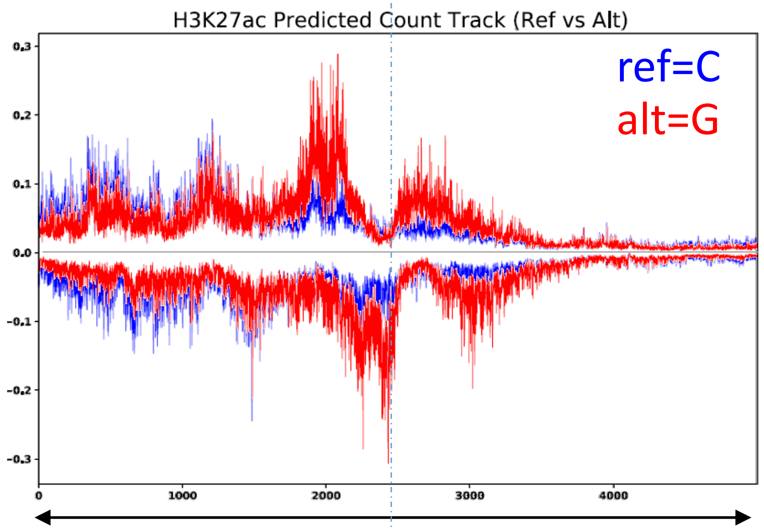
1 Kb

Predicted DNase-seq



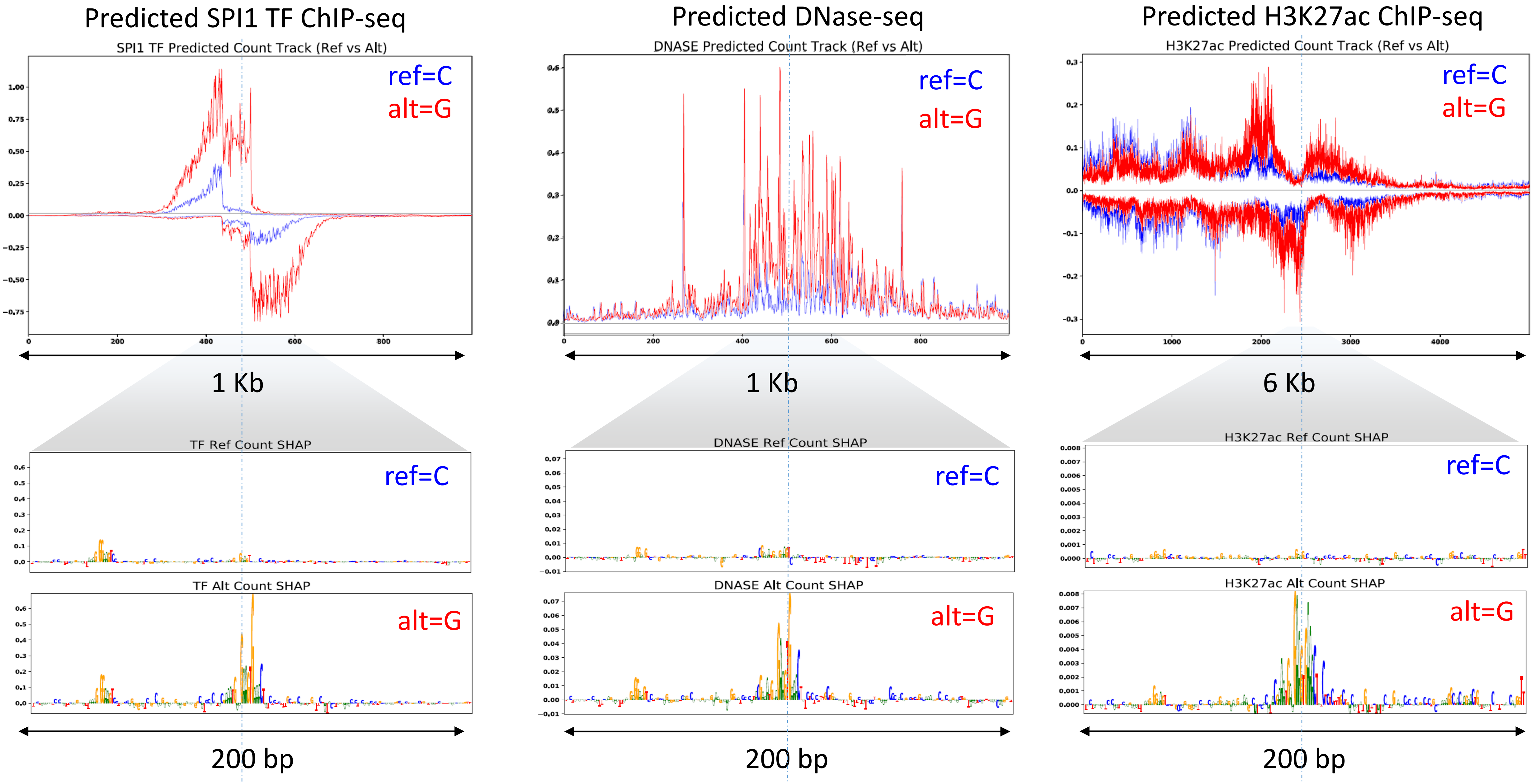
1 Kb

Predicted H3K27ac ChIP-seq



6 Kb

BPNet/ChromBPNet can interpret variants influencing regulatory activity



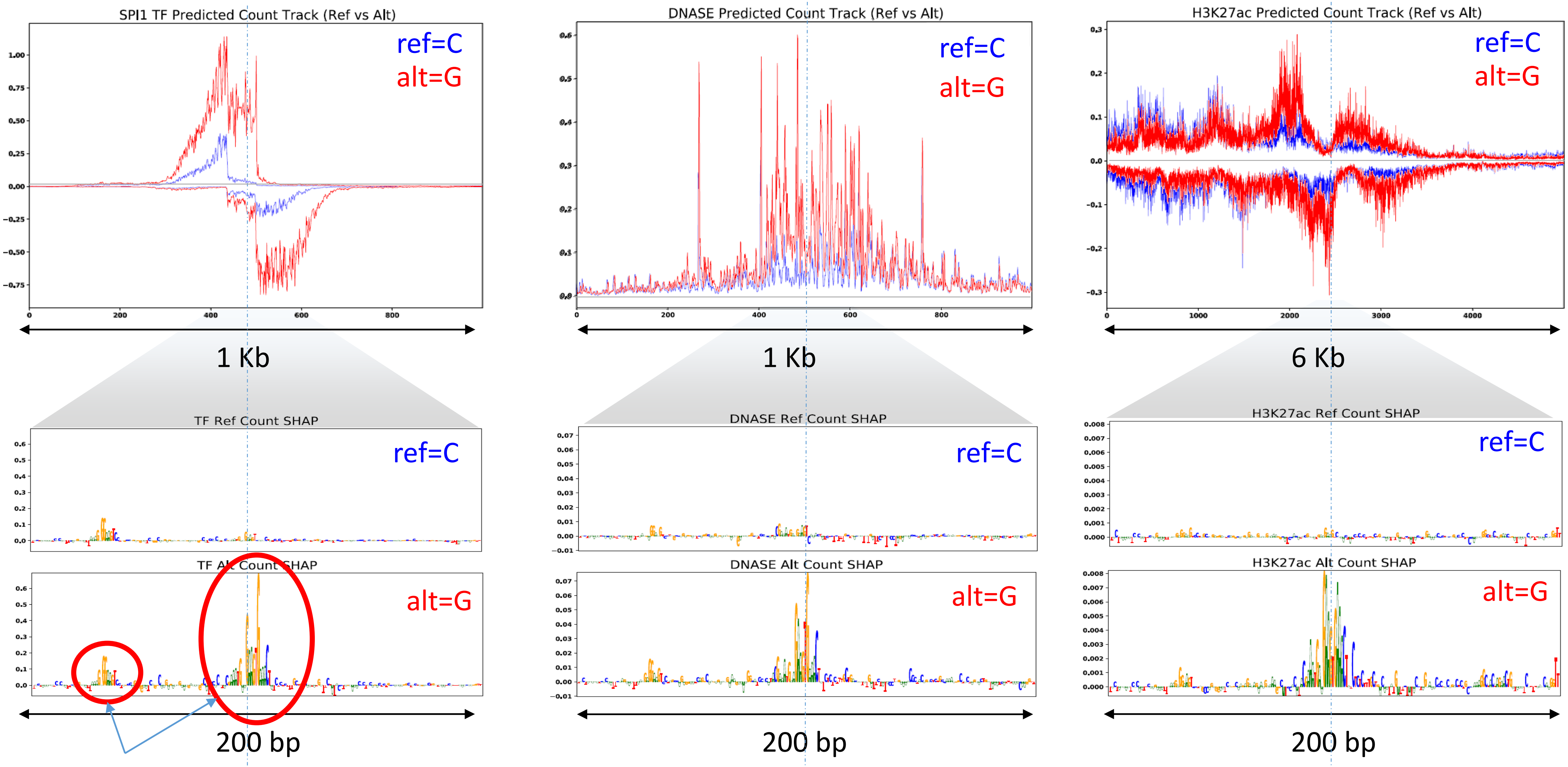
Model interpretation predicts sequence drivers

BPNet/ChromBPNet can interpret variants influencing regulatory activity

Predicted SPI1 TF ChIP-seq

Predicted DNase-seq

Predicted H3K27ac ChIP-seq



SPI1 motifs

Model interpretation predicts sequence drivers

Variant Effect Scoring with ChromBPNet

- ChromBPNet has two heads counts and profiles

- Variant effect scoring with counts head

$$log(counts_{alt}) - log(counts_{ref})$$

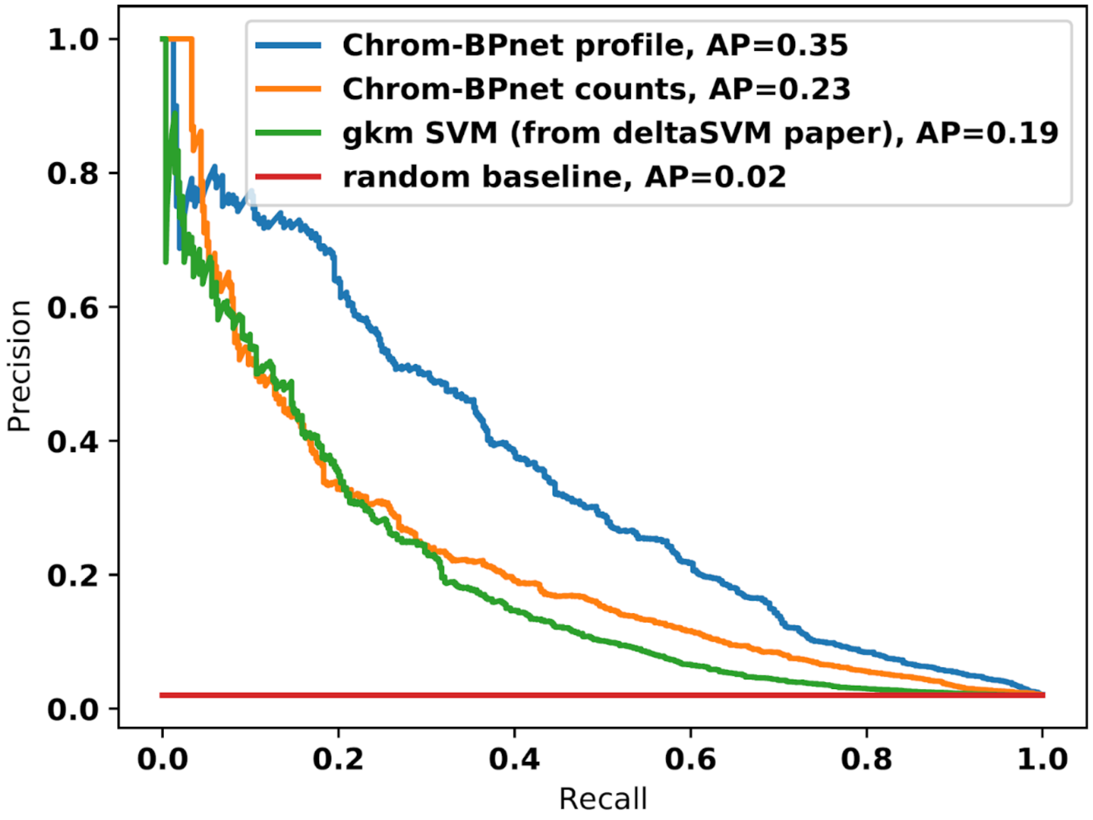
- Variant effect scoring with profile head

$$JensenShanon(Profile_{alt,} Profile_{ref}) * Sign(log(counts_{alt}) - log(counts_{ref}))$$

Dnase-seq ChromBPNet outperforms deltaSVM for predicting dsQTLs in LCLs

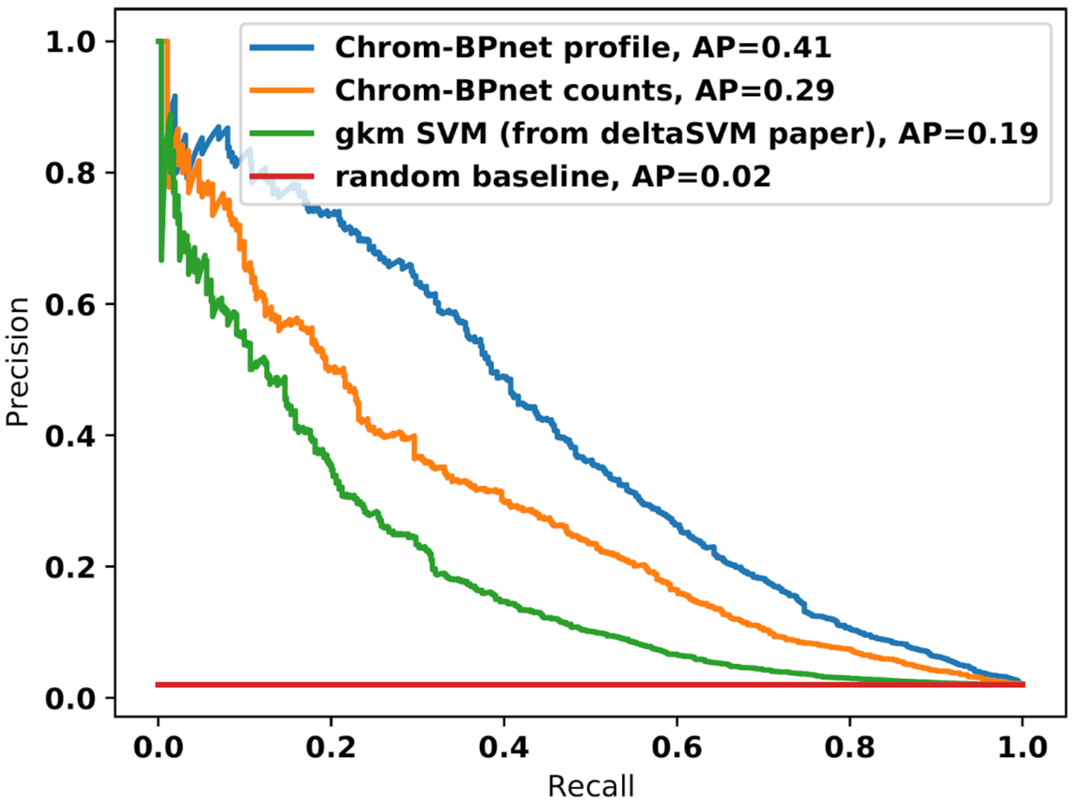
GM12878 DNASE-seq model

85M read depth



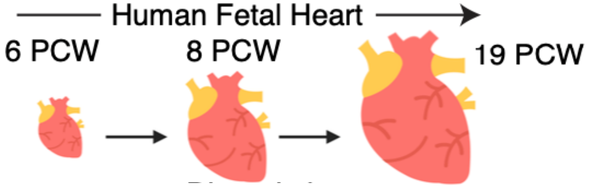
GM12878 ATAC-seq model

175M read depth



dsQTLs: Degner et al 2012

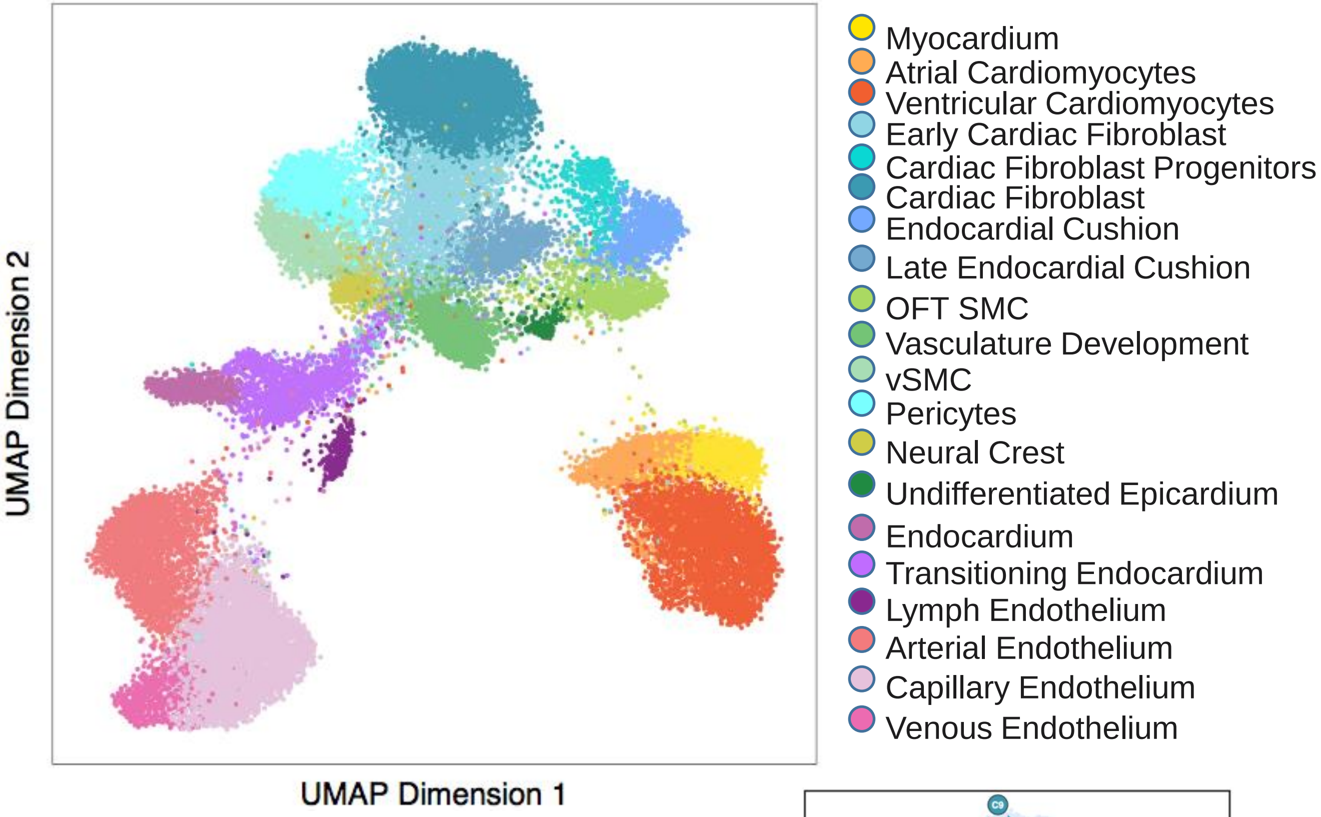
Single cell chromatin dynamics during human cardiogenesis

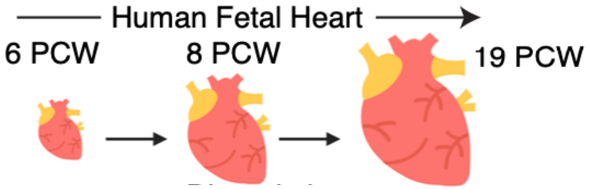


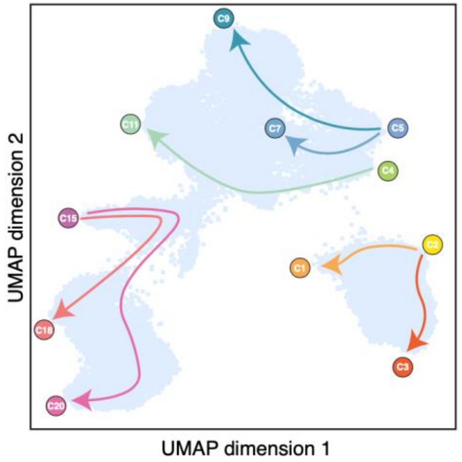
Sundaram*, Ameen*, et al. In review

Laksshman
Sundaram

Single cell chromatin dynamics during human cardiogenesis



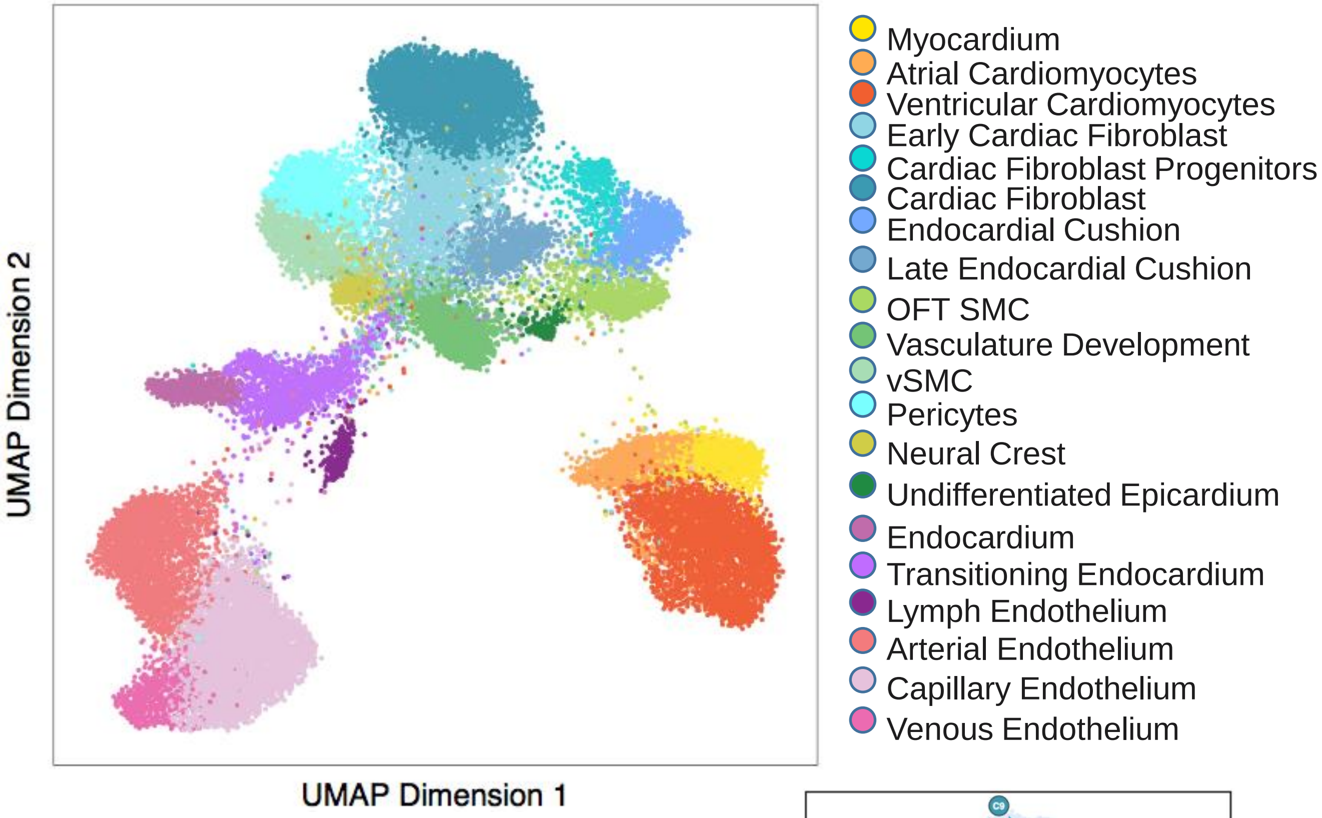


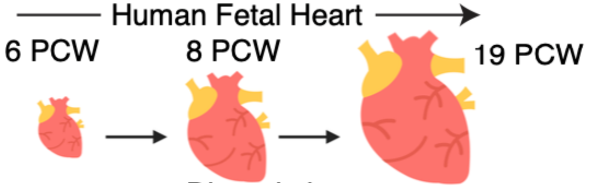


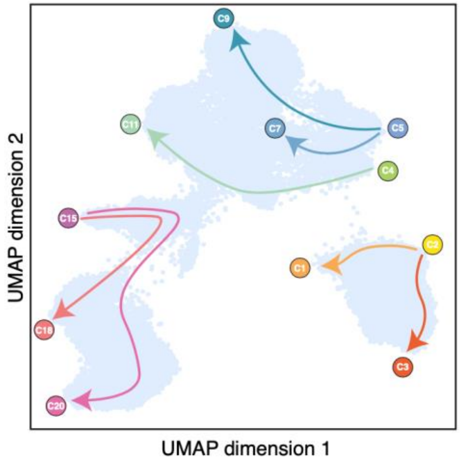
Laksshman
Sundaram

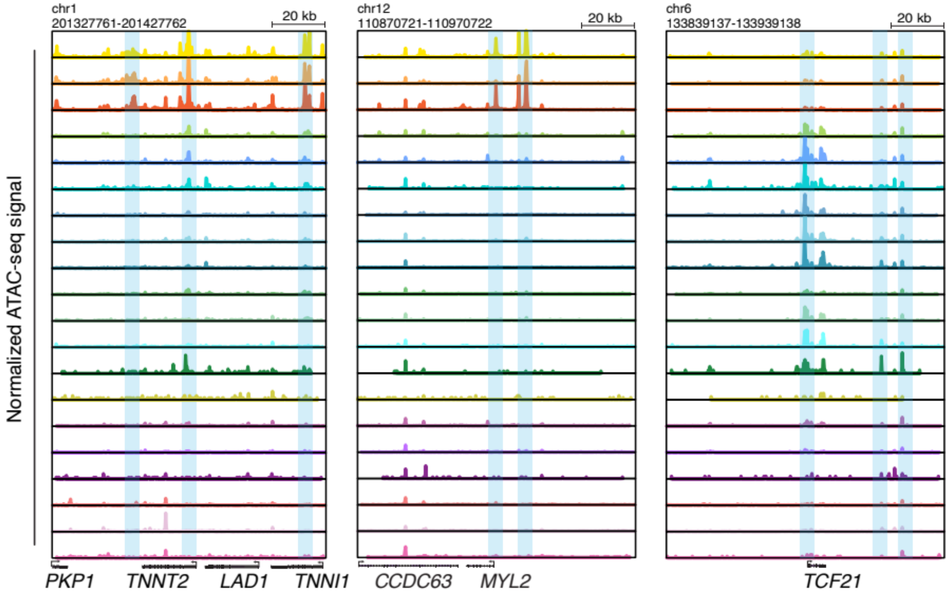
Sundaram*, Ameen*, et al. In review

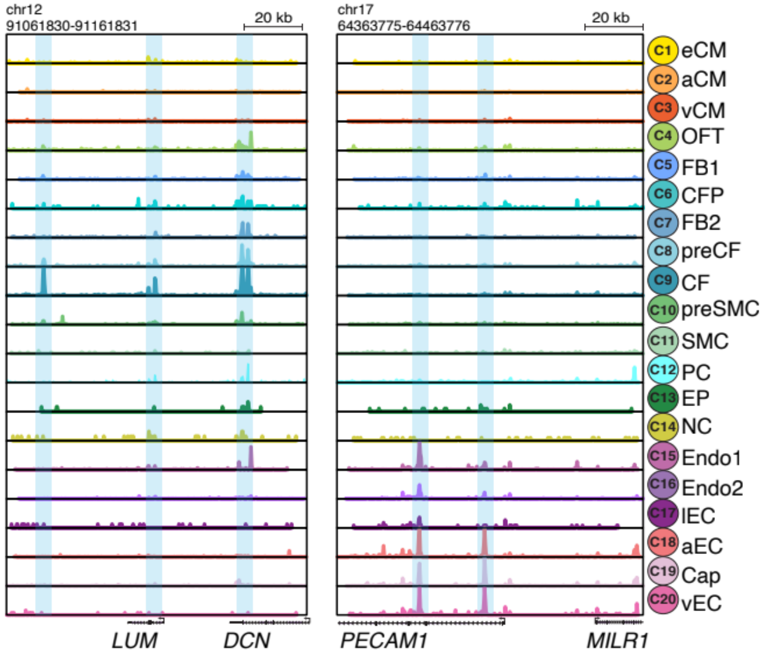
Single cell chromatin dynamics during human cardiogenesis







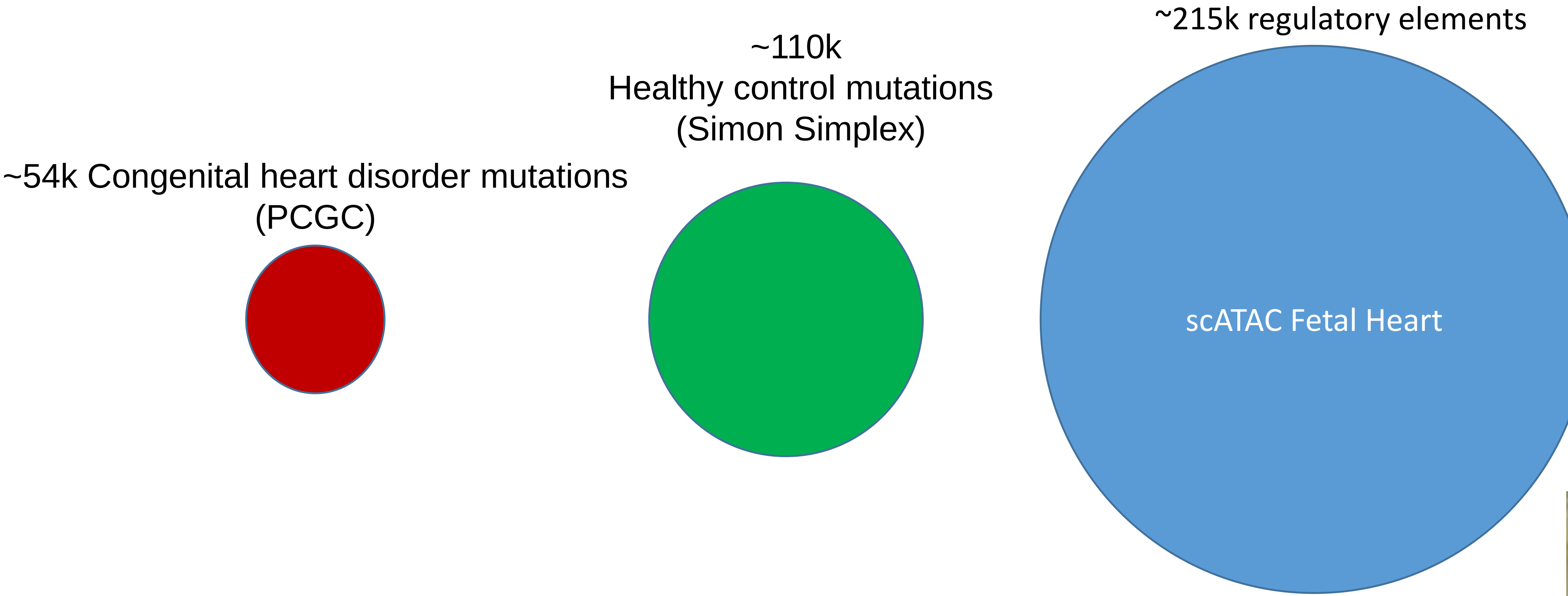




Laksshman Sundaram

Sundaram*, Ameen*, et al. In review

Prioritizing de-novo mutations in congenital heart disease with cell-type resolved regulatory map of fetal heart



~54k Congenital heart disorder mutations (PCGC)

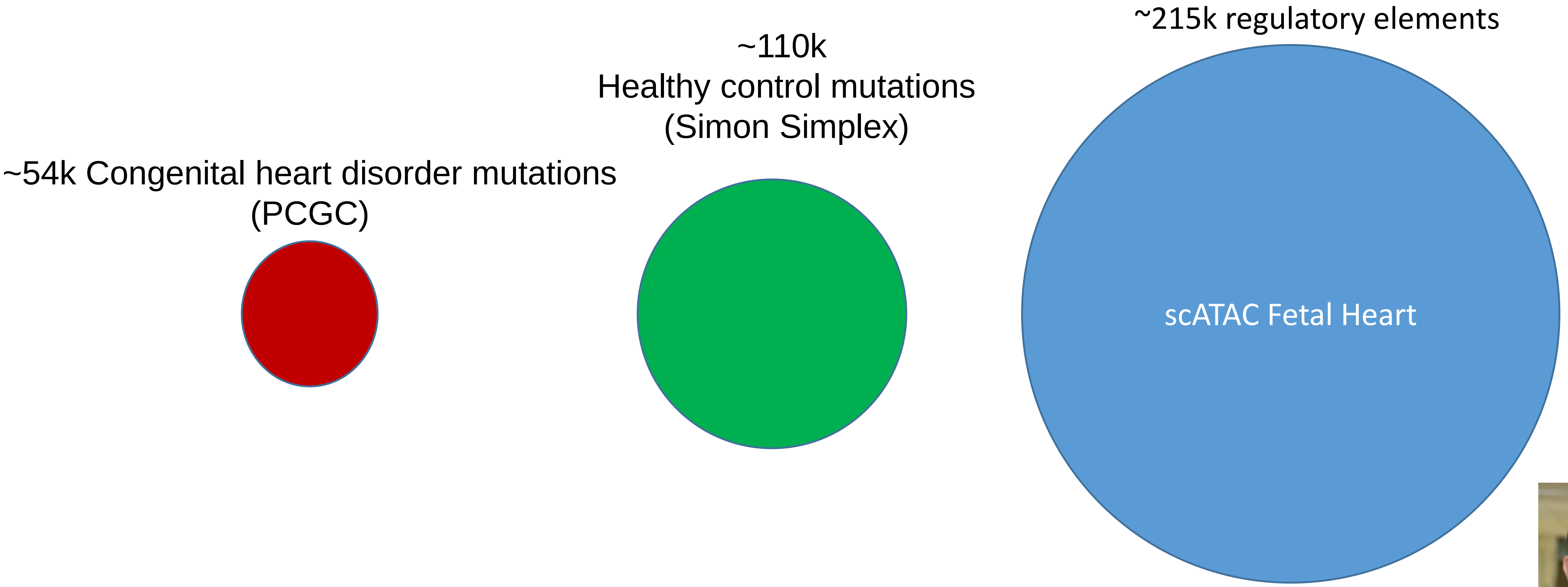
~110k Healthy control mutations (Simon Simplex)

~215k regulatory elements

scATAC Fetal Heart

Laksshman Sundaram

Prioritizing de-novo mutations in congenital heart disease with cell-type resolved regulatory map of fetal heart



~54k Congenital heart disorder mutations
(PCGC)

~110k
Healthy control mutations
(Simon Simplex)

~215k regulatory elements

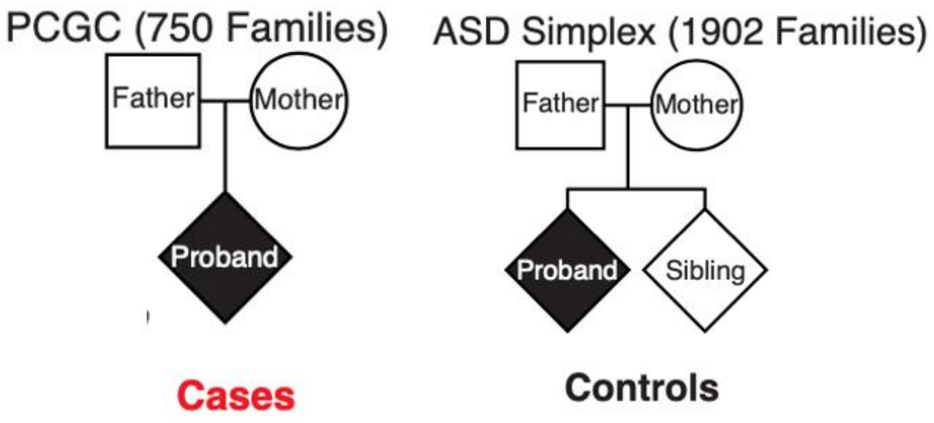
scATAC Fetal Heart

No enrichment of CHD mutations in all/cell type resolved scATAC-seq peaks!

Laksshman
Sundaram

Prioritizing mutations with cell-type resolved ChromBPNet models

Laksshman
Sundaram



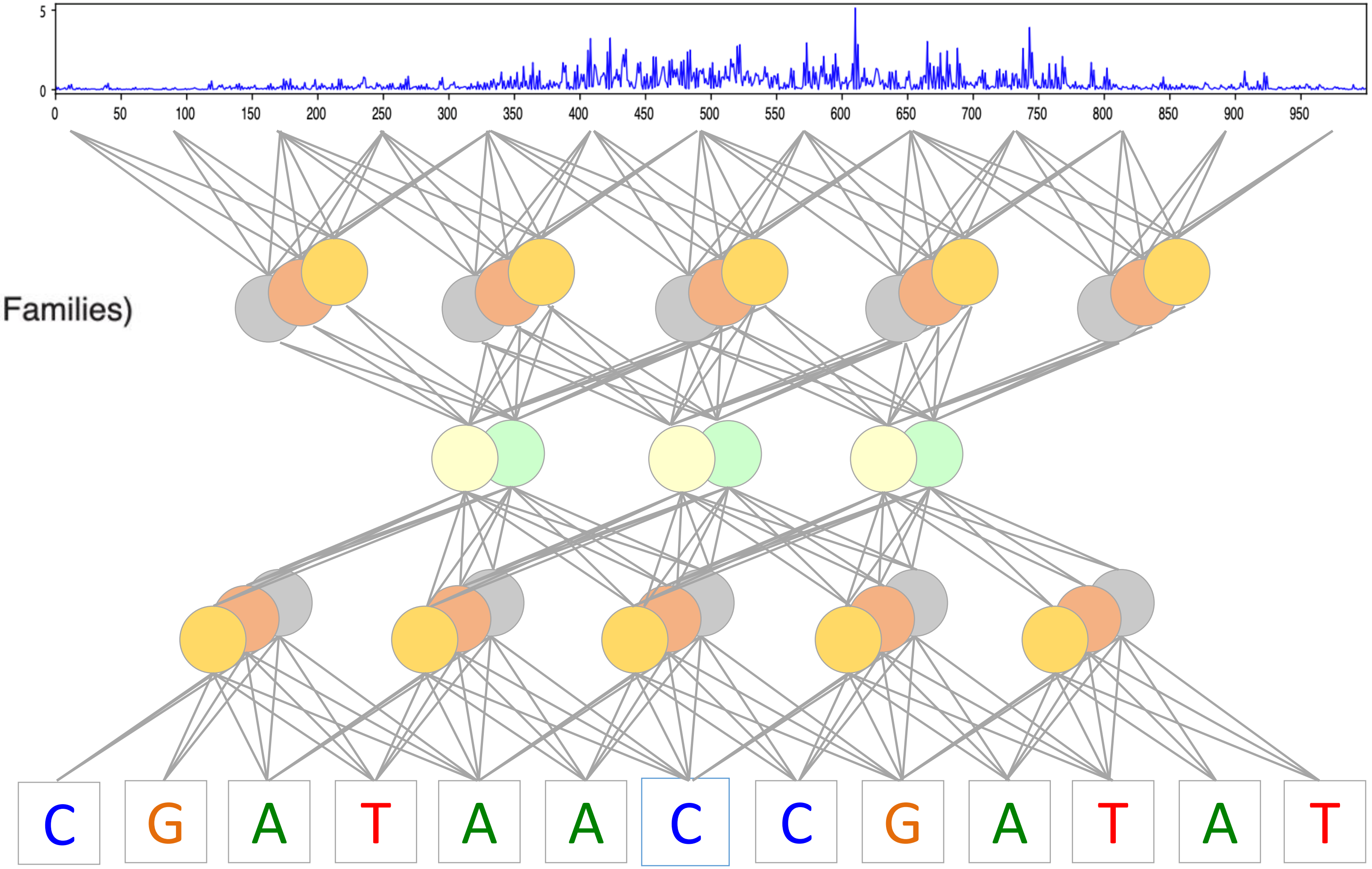
PCGC (750 Families)

ASD Simplex (1902 Families)

Cases

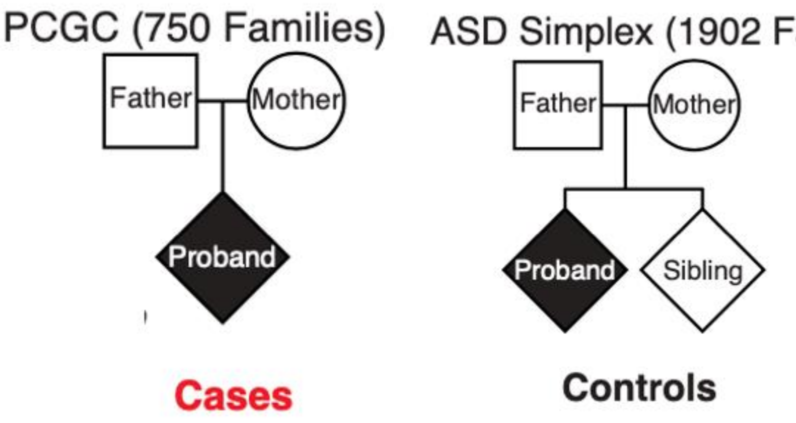
Controls

De novo non-coding mutations

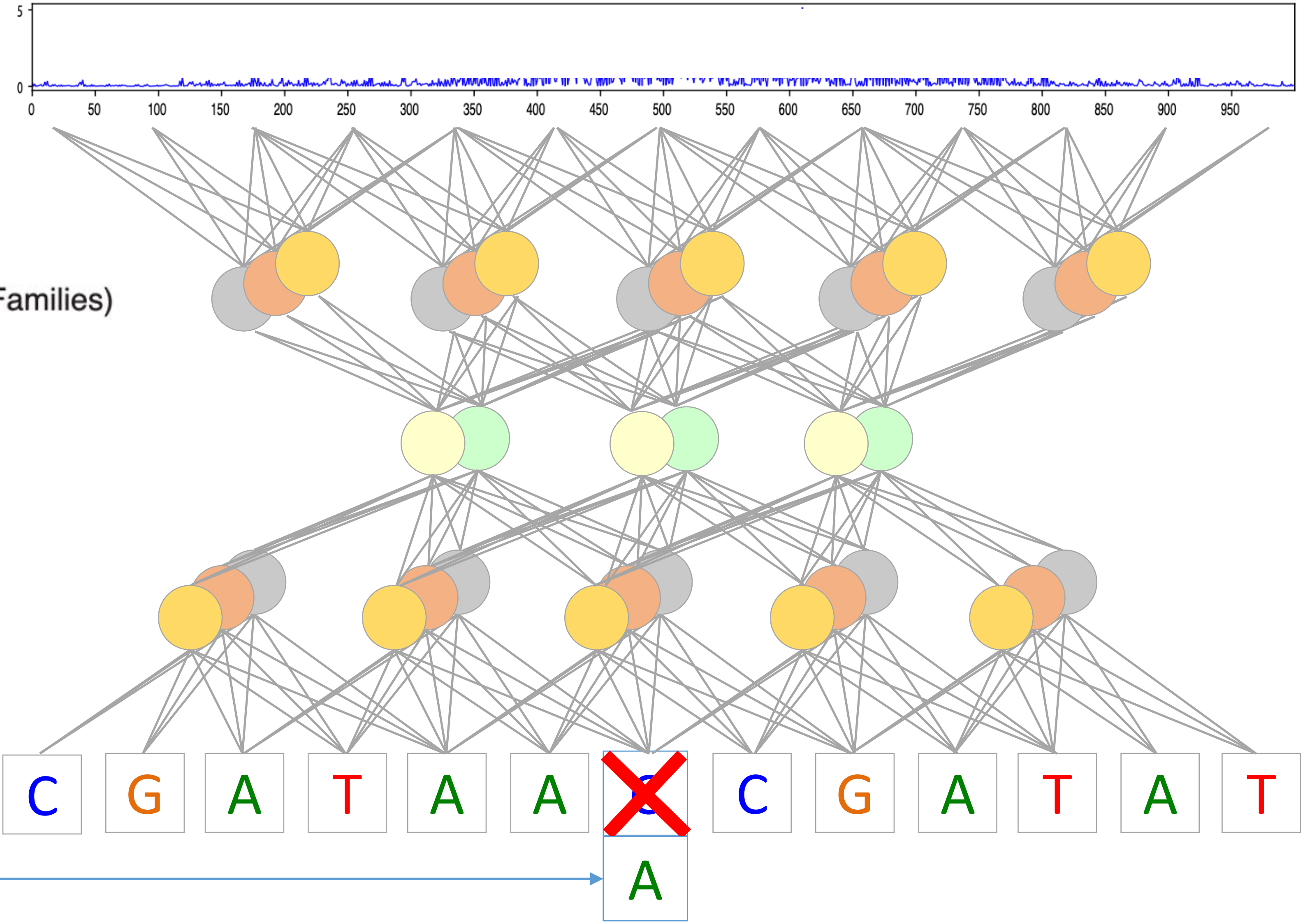


Prioritizing mutations with cell-type resolved ChromBPNet models

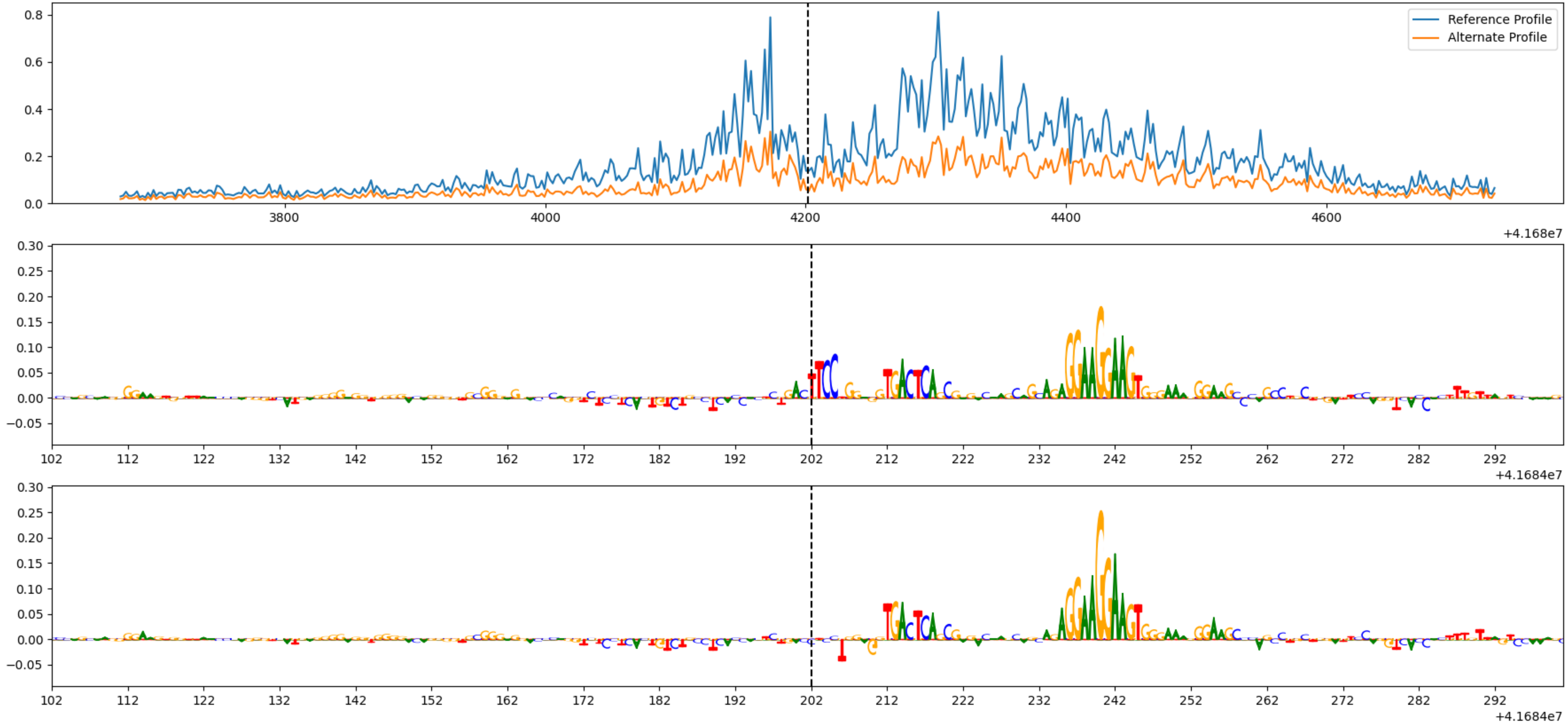
Laksshman
Sundaram



De novo non-coding mutations



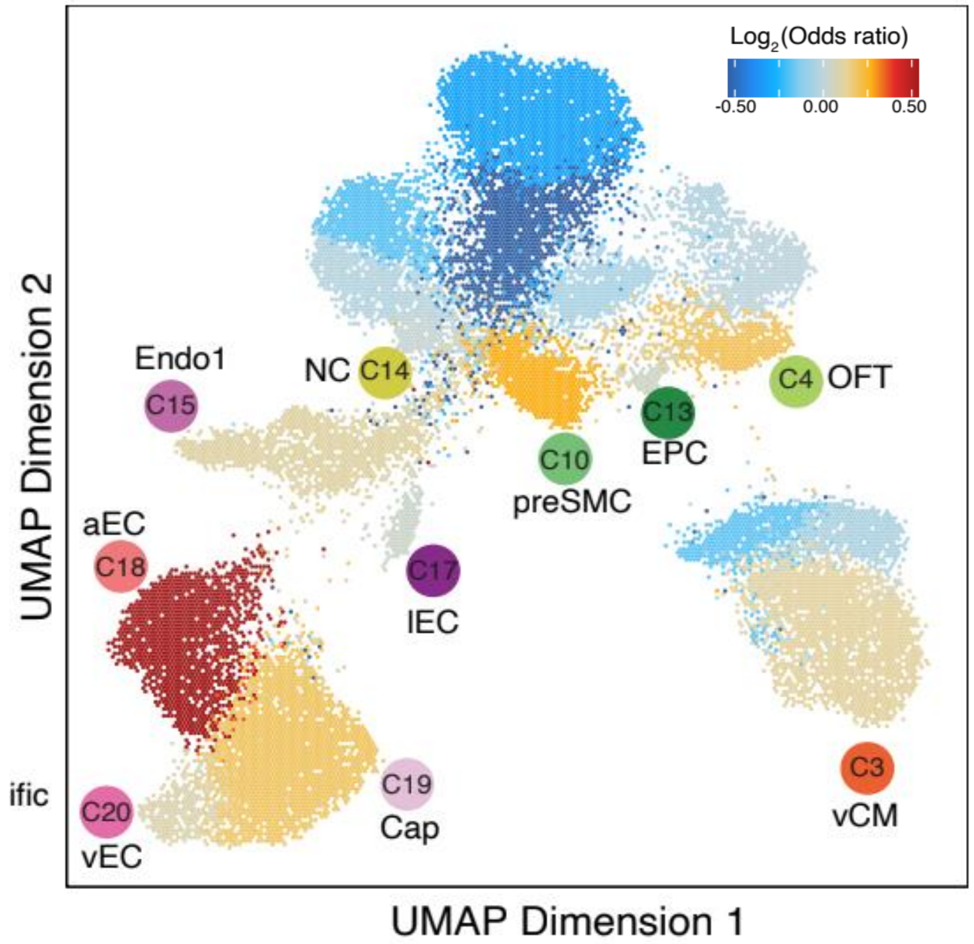
Eg: CHD case mutation affecting accessibility of enhancer in Art/Cap endothelial cells



Laksshman Sundaram

Mutation disrupts an ETS/ELK/ETV family motif

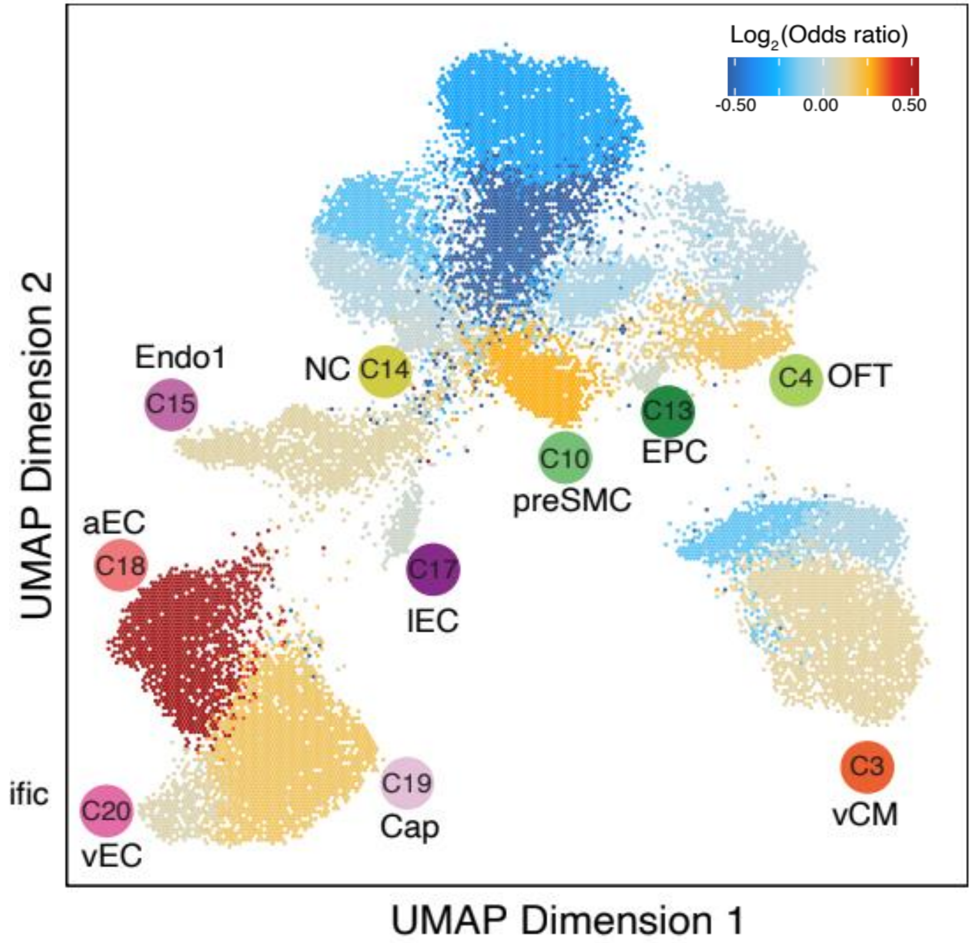
Cell states enriched for prioritized *de novo* non-coding mutations in CHD

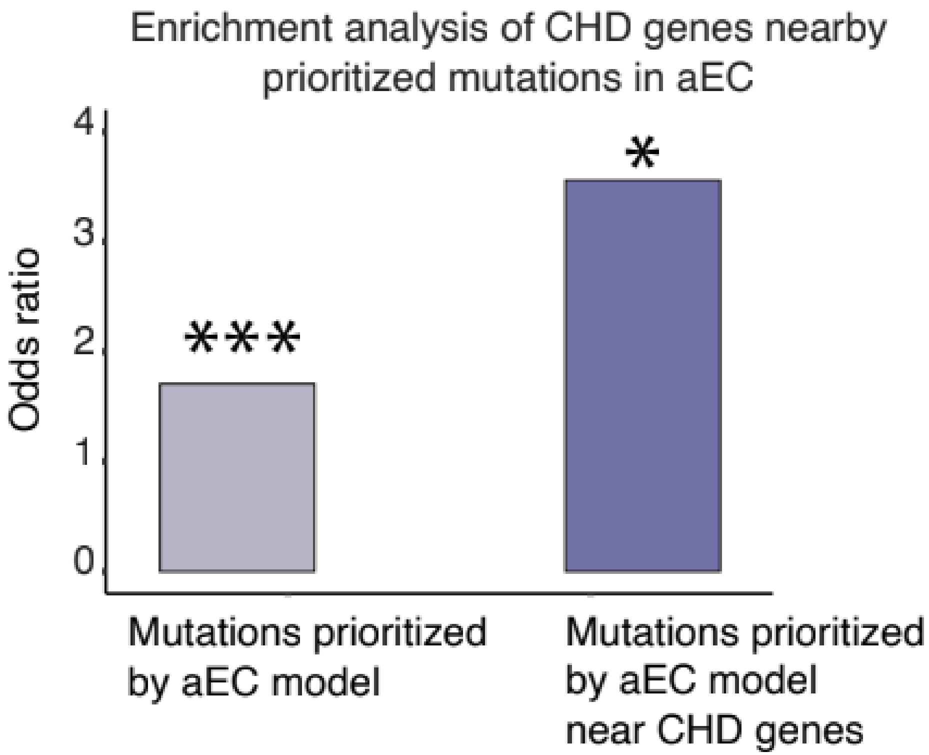


<u>Arterial & Capillary endothelial cells</u> are most significantly enriched for CHD mutations (structural defects)

Laksshman Sundaram

Cell states enriched for prioritized *de novo* non-coding mutations in CHD

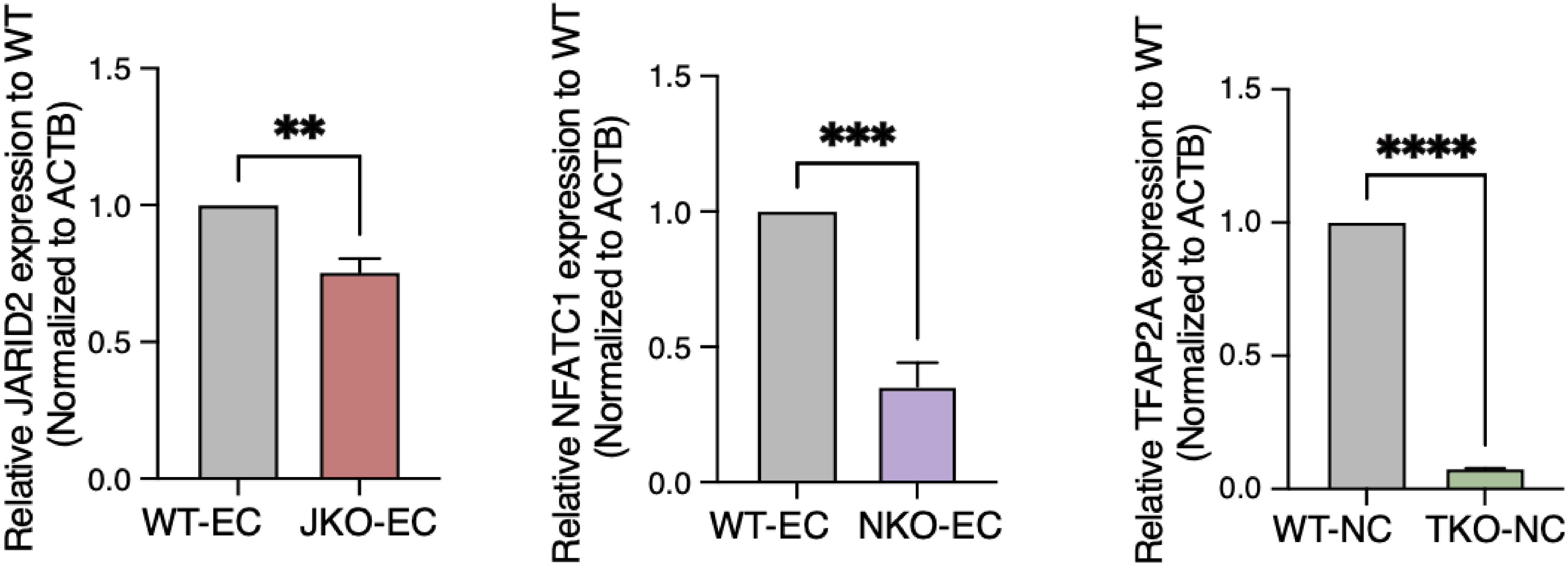




Laksshman Sundaram

<u>Arterial & Capillary endothelial cells</u> are most significantly enriched for CHD mutations (structural defects)

CRISPR experiments confirm downstream gene targets of enhancers containing prioritized CHD mutations



Summary

- Base-resolution neural networks can learn very accurate models of regulatory DNA sequence from bulk and single cell regulatory profiling experiments
- Can be queried to decipher novel subtle sequence syntax properties
- Can be used to decipher regulatory genetic variation
- Can be used to prioritize likely causal variants in GWAS loci and de-novo non-coding mutations
- Can be used to design precise genome editing experiments
- <u>Foundation of *in-silico* platforms for biological discovery, hypothesis generation & model-driven iterative expt. design</u>

STANFORD COMPUTER SCIENCE

Kundaje lab

STANFORD SCHOOL of MEDICINE | Genetics

Daniel Kim (BMI)

Kelly Cochran (CS)

Soumya Kundu (CS)

Surag Nair (CS)

Maxim Zaslavsky (CS)

Vivek Ramalingam (Postdoc)

Caleb Lareau (Postdoc)

Akshay Balsubramani (Postdoc)

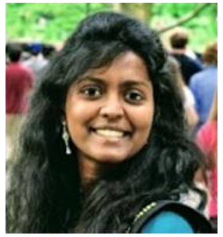
Georgi Marinov (Postdoc)

Alex Tseng (CS)

Amr Alexandari (CS)

Abhimanyu Banerjee (Physics)

Laksshman Sundaram (CS)


Anusri Pampari (CS)

Kristy Mualim (Bioinformatician)

Jacob Schreiber (Postdoc)

Mahfuza Sharmin (Postdoc)

Eran Kotler (Postdoc)

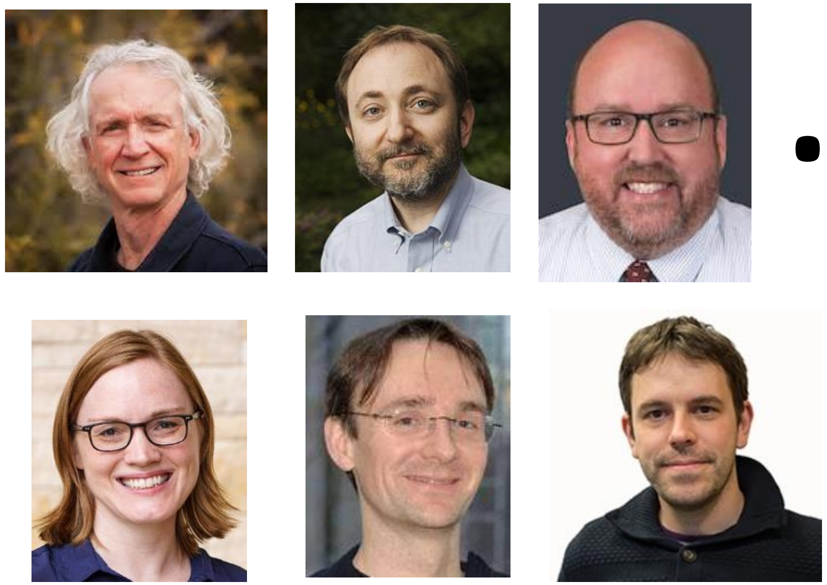
Zahoor Zafrulla (ML engineer)

Jin Wook Lee (Software engineer)

STANFORD
COMPUTER SCIENCE

STANFORD SCHOOL of MEDICINE | Genetics

Collaborator labs



....

ENCODE

ROADMAP epigenomics PROJECT

Funding

ALFRED P. SLOAN FOUNDATION 1934

1R01HG009674

1U01HG009431

1U24HG009446

R01ES02500902

1DP2OD022870

NIH DIRECTOR'S NEW INNOVATOR AWARD