STANFORD
COMPUTER SCIENCE

Stanford
MEDICINE | Department of Genetics

Decoding sequence syntax of gene regulation and functional genetic variation

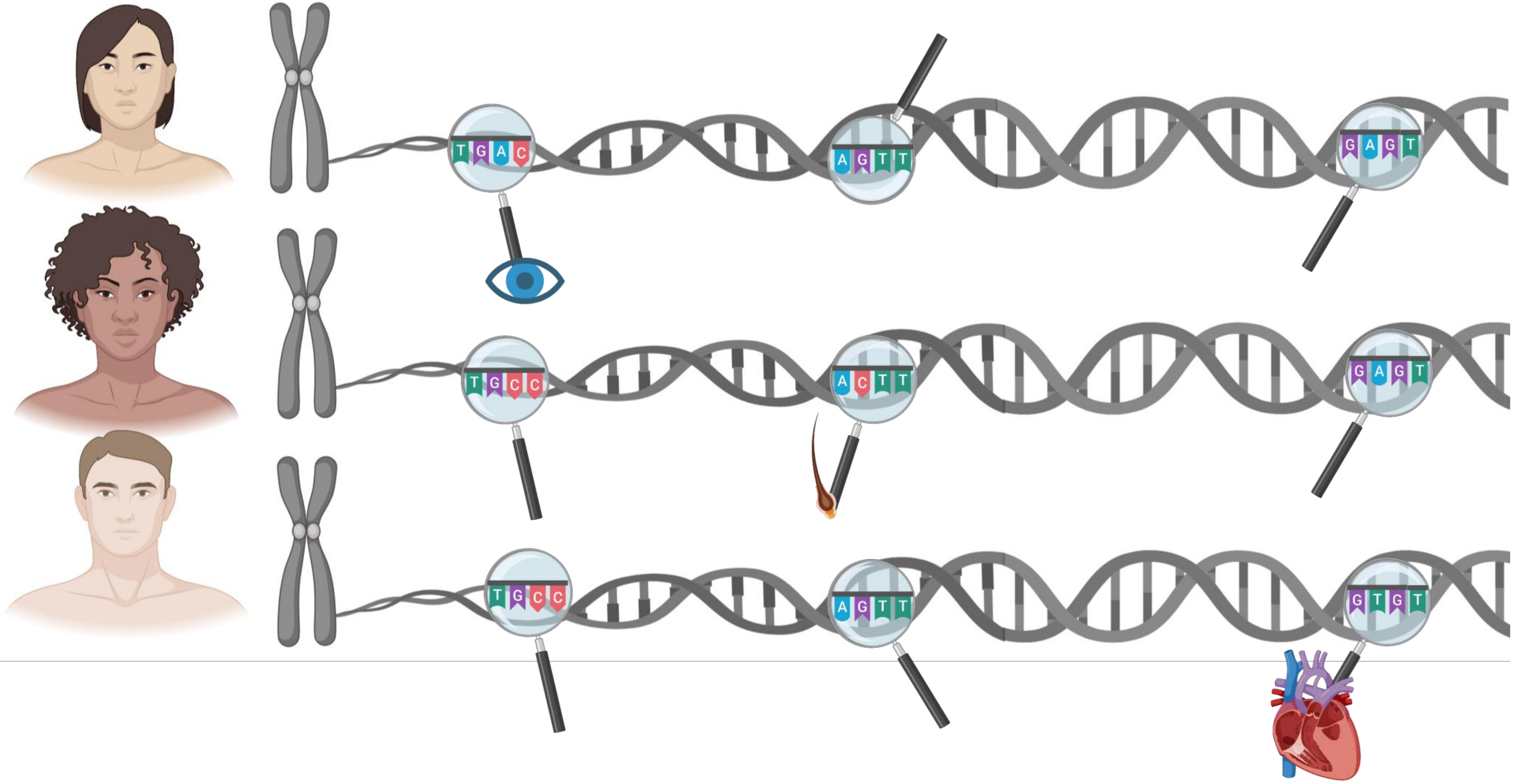
Kundaje lab

Twitter:@anshulkundaje

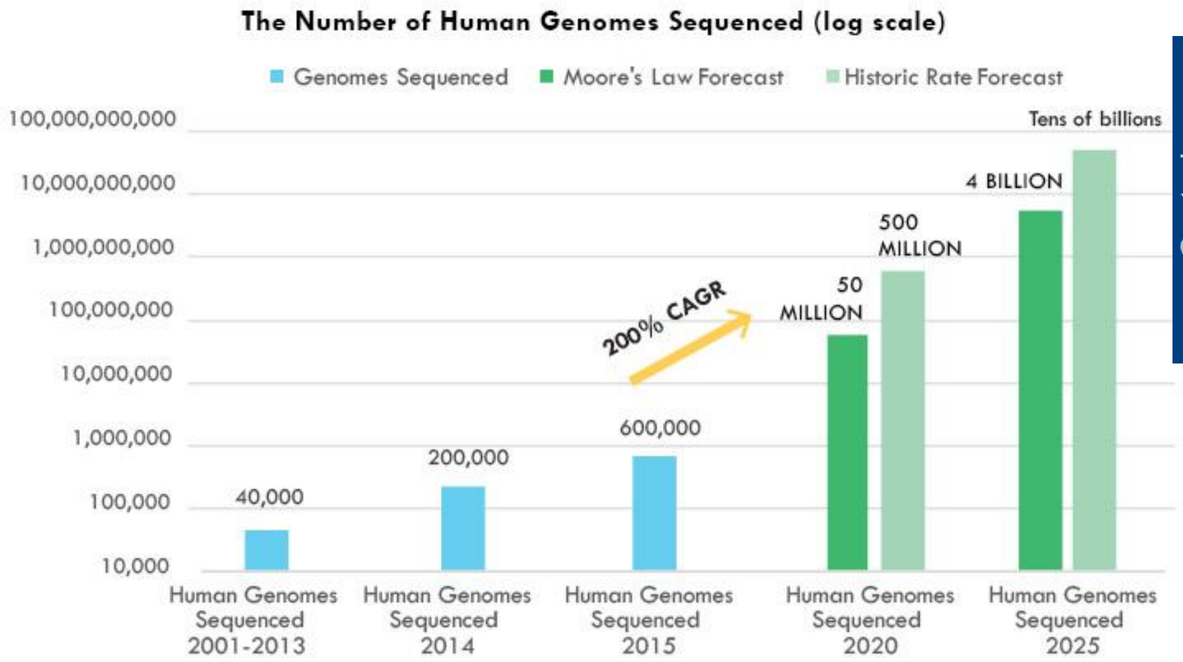
Website: http://anshul.kundaje.net

Genetic variants associated with traits and diseases

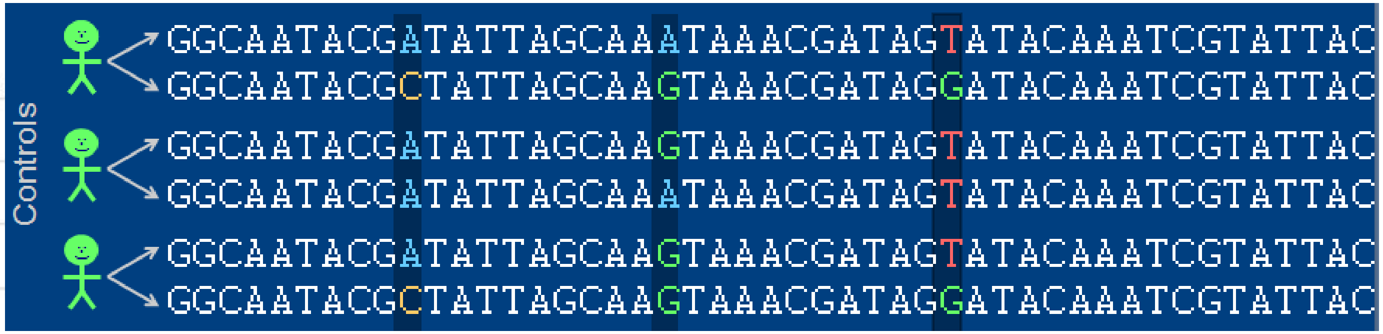
Genetic variants associated with traits and diseases



Population sequencing to identify disease-associated genetic variants



Source: National Human Genome Research Institute (NHGRI), ARK Investment Management LLC


Controls
GGCAATACGATATTAGCAAATAAACGATAGTATACAATCGTATTAC
GGCAATACGCTATTAGCAAGTAAACGATAGGATACAATCGTATTAC
GGCAATACGATATTAGCAAGTAAACGATAGTATACAATCGTATTAC
GGCAATACGATATTAGCAAATAAACGATAGTATACAATCGTATTAC
GGCAATACGATATTAGCAAGTAAACGATAGTATACAATCGTATTAC
GGCAATACGCTATTAGCAAGTAAACGATAGGATACAATCGTATTAC

GA II
1.6 billion bp per day
(2008)

GA IIx
5 billion bp per day
(2009)

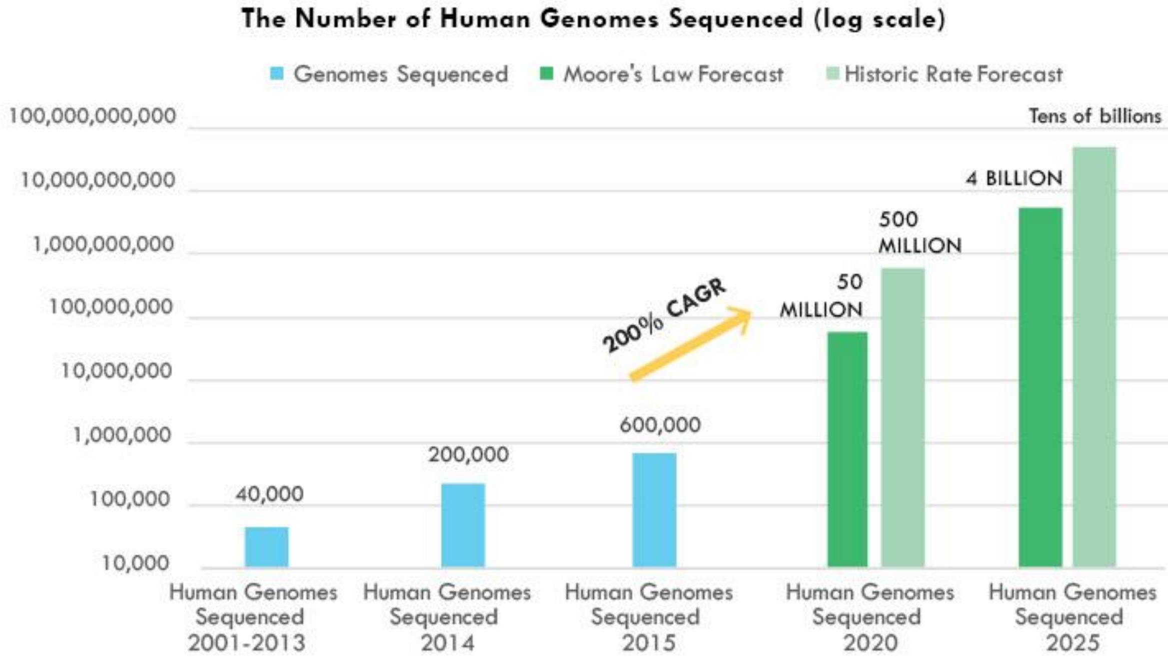
HiSeq 2500
60 billion bp per day
(2012)

Images: www.illumina.com/systems
Numbers: www.politigenomics.com/next-generation-sequencing-informatics
Dates: Illumina press releases

Oxford Nanopore technology

Millions of common and rare genetic variants found in human population

Population sequencing to identify disease-associated genetic variants



Source: National Human Genome Research Institute (NHGRI), ARK Investment Management LLC

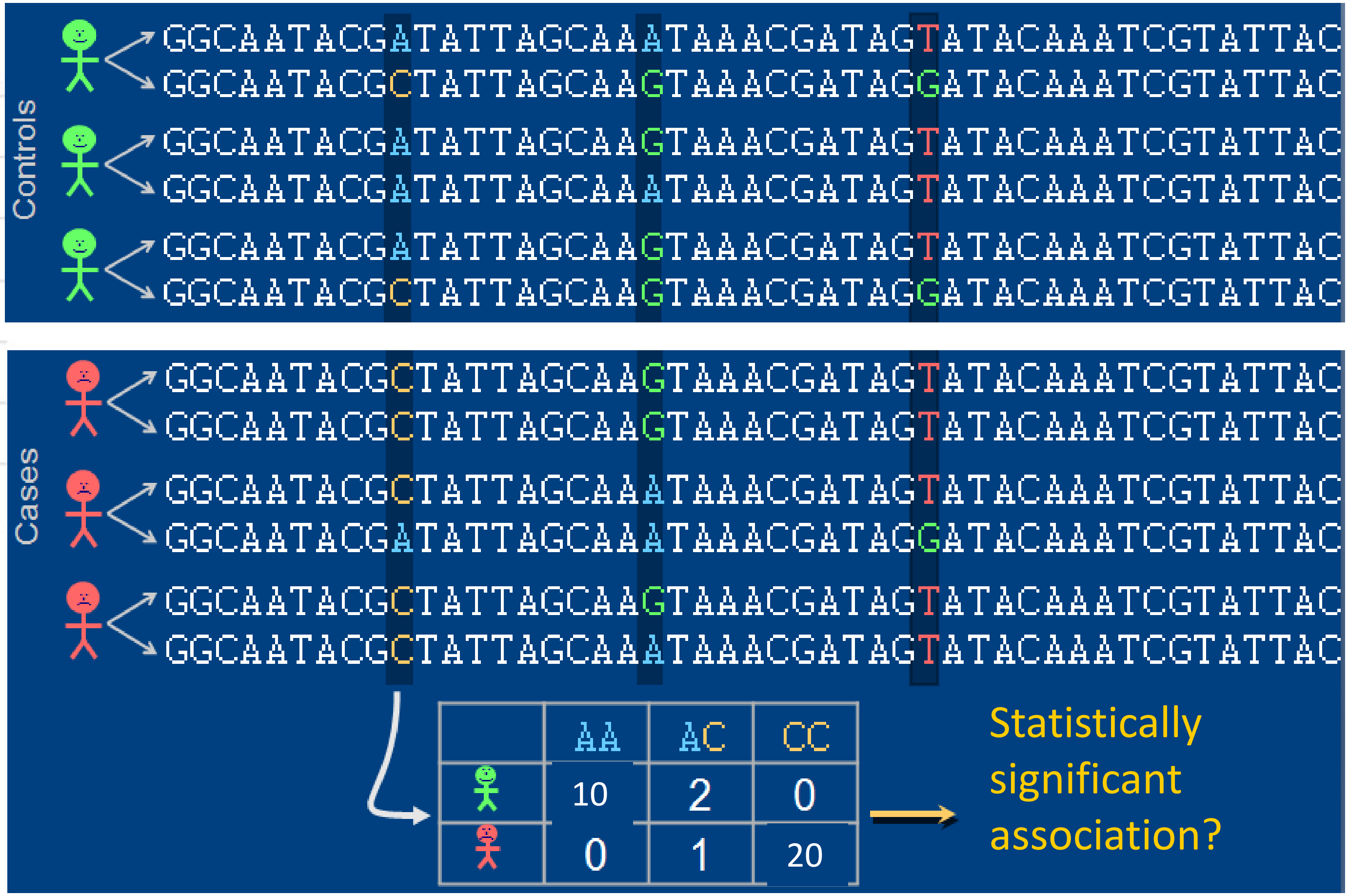
GA II
1.6 billion bp per day
(2008)

GA IIx
5 billion bp per day
(2009)

HiSeq 2500
60 billion bp per day
(2012)

Oxford Nanopore
technology

Images: www.illumina.com/systems
Numbers: www.politigenomics.com/next-generation-sequencing-informatics
Dates: Illumina press releases



Controls

```
GGCAATACGATATTAGCAAATAAACGATAGTATACAAATCGTATTAC
GGCAATACGCTATTAGCAAGTAAACGATAGGATACAAATCGTATTAC
GGCAATACGATATTAGCAAGTAAACGATAGTATACAAATCGTATTAC
GGCAATACGATATTAGCAAATAAACGATAGTATACAAATCGTATTAC
GGCAATACGATATTAGCAAGTAAACGATAGTATACAAATCGTATTAC
GGCAATACGCTATTAGCAAGTAAACGATAGGATACAAATCGTATTAC
```

Cases

```
GGCAATACGCTATTAGCAAGTAAACGATAGTATACAAATCGTATTAC
GGCAATACGCTATTAGCAAGTAAACGATAGTATACAAATCGTATTAC
GGCAATACGCTATTAGCAAATAAACGATAGTATACAAATCGTATTAC
GGCAATACGATATTAGCAAATAAACGATAGGATACAAATCGTATTAC
GGCAATACGCTATTAGCAAGTAAACGATAGTATACAAATCGTATTAC
GGCAATACGCTATTAGCAAATAAACGATAGTATACAAATCGTATTAC
```

	AA	AC	CC
🙂	10	2	0
☹	0	1	20

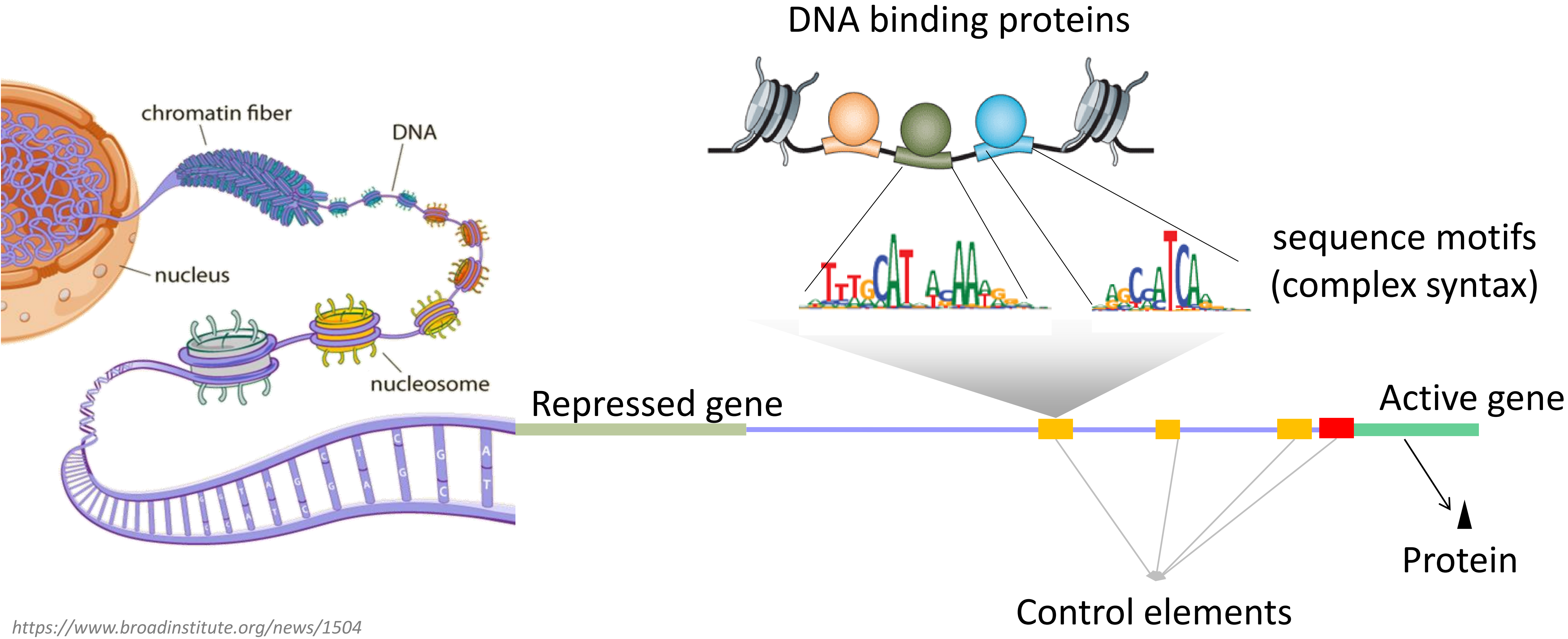
Statistically significant association?

Millions of common and rare genetic variants found in human population

TGCCAAGCAGCAAAGTTTTGCTGCTGTTTATTTTTGTAGCTCTTACTATATTCT
ACTTTTACCATTGAAAATATTGAGGAAGTTATTTATATTTCTATTTTTTATATAT
TATATATTTTATGTATTTTAATATTACTATTACACATAATTATTTTTTATATATATGA
AGTACCAATGACTTCCTTTTCCA[illegible]AGCAATAATGAAATTTCACAGTATGAAA
ATGGAAGAAATCAATAAAATTATACGTGACCTGTGGCGAAGTACCTATCGTG
GACAAGGTGAGTACCATGGTGTATCACAAATGCTCTTTCCAAAGCCCTCTCC
GCAGCTCTTCCCCTTATGACCTCTCATCATGCCAGCATTACCTCCCTGGACCC
CTTTCTAAGCATGTCTTTGAGATTTTCTAAGAATTCTTATCTTGGCAACATCTT
GTAGCAAGAAAATGTAAAGTTTTCTGTTCCAGAGCCTAACAGGACTTACATA
TTTGACTGCAGTAGGCATTATATTTAGCTGATGACATAATAGGTTCTGTCATA
GTGTAGATAGGGATAAGCCAAAATGCAATAAGAAAAACCATCCAGAGGAA
ACTCTTTTTTTTTTCTTTTTCTTTTTTTTTTTTCCAGATGGAGTCTCGCACTTC
TCTGTCACCCGGGCTGGAGCGCAGTGGTGCAATCTTGGCTCACTGCAACCT
CCACCTCCTGGGTTCAGGTGATTCTCCCACCTCAGCCTCCCGAGTAGTAGCT
GGAATTACAGGTGCGCGCTCCCACACCTGGCTAATTTTTTGTATTCTTAGTA
GAGATGGGGTTTCACCATGTTGGCCAGGCTGGTCTCAAACTCCTGCCCTCA
GGTGATCTGCCCACCTTGGCCTCCCAGTGTTGGGTTTACAGGCGTGAGCCA
CCGCGCCTGGCCTGGAGGAAACTCTTAACAGGGAAACTAAGAAAGAGTTG
AGGCTGAGGAACTGGGGCATCTGGGTTGCTTCTGGCCAGACCACCAGGCT
CTTGAATCCTCCCAGCCAGAGAAAGAGTTTCCACACCAGCCATTGTTTTCCT
CTGGTAATGTCAGCCTCATCTGTTGTTCCTAGGCTTACTTGATATGTTTGTAA
ATGACAAAAGGCTACAGAGCATAGGTTCCTCTAAAATATTCTTCTTCCTGTGT
CAGATATTGAATACATAGAAATACGGTCTGATGCCGATGAAAATGTATCAGCT
TCTGATAAAAGGCGGAATTATAACTACCGAGTGGTGATGCTGAAGGGAGAC
ACAGCCTTGGATATGCGAGGACGATGCAGTGCTGGACAAAAGGCAGGTAT
CTCAAAAGCCTGGGGAGCCAACTCACCCAAGTAACTGAAAGAGAGAAACA
AACATCAGTGCAGTGGAAGCACCCAAGGCTACACCTGAATGGTGGGAAGC
TCTTTGCTGCTATATAAAATGAATCAGGCTCAGCTACTATTATT …………

What is the functional consequence of genetic variants?

Functional components of the human genome



DNA binding proteins

sequence motifs (complex syntax)

Repressed gene

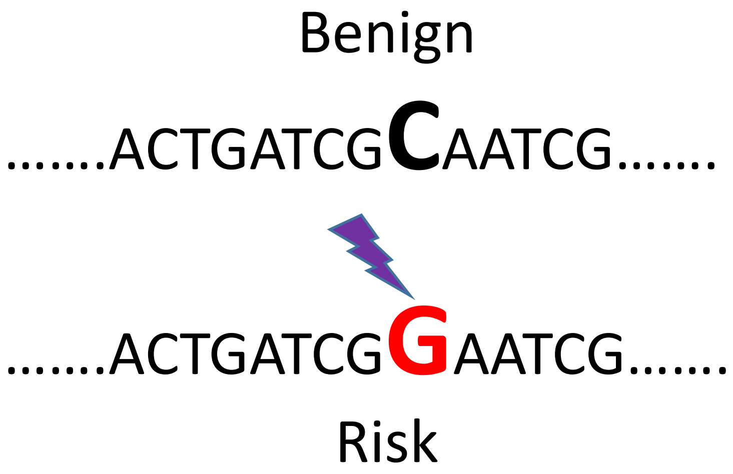
Active gene

Protein

Control elements

https://www.broadinstitute.org/news/1504

>95% of disease variants are not disrupting protein coding gene regions



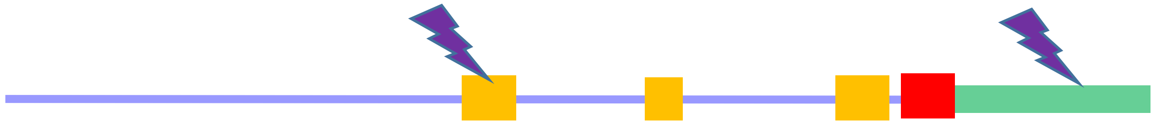
Benign

…….ACTGATCG**C**AATCG…….

…….ACTGATCG**G**AATCG…….

Risk

>95% of disease variants are not disrupting protein coding gene regions



Control elements
(**<u>Non-coding variants</u>**)

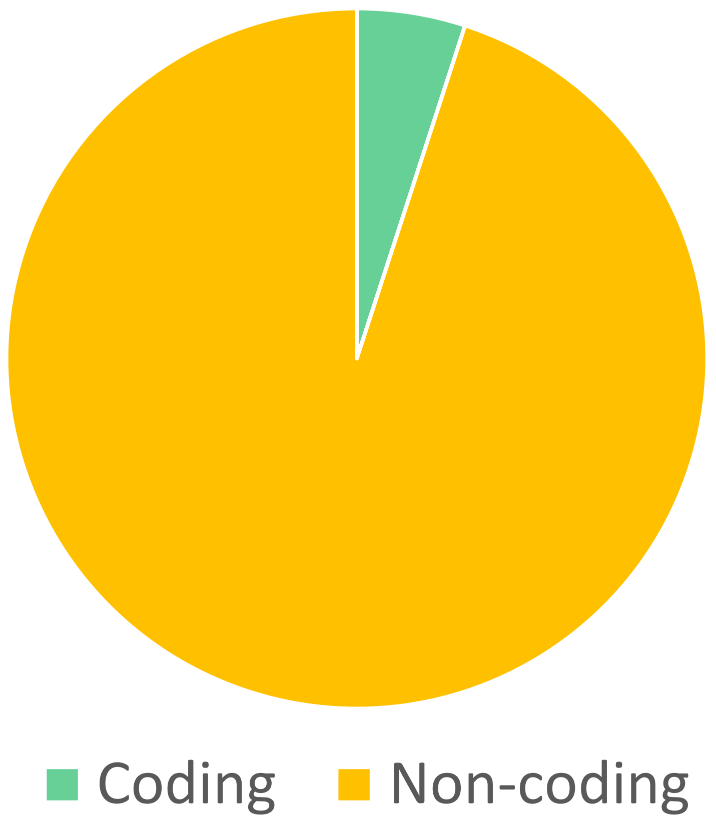
Gene
(Coding variant)

Benign

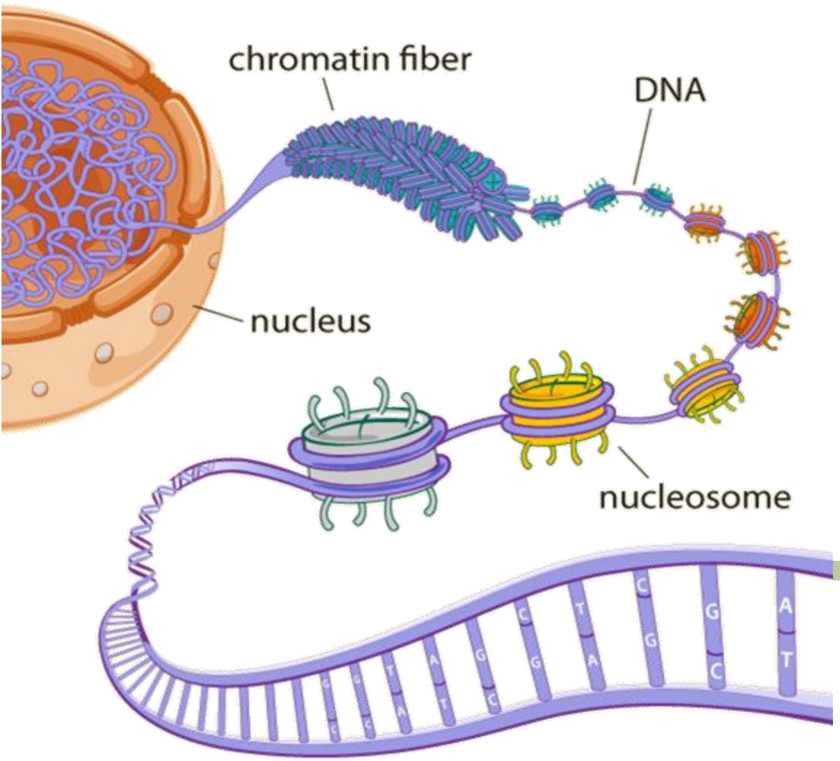
…….ACTGATCG**C**AATCG…….

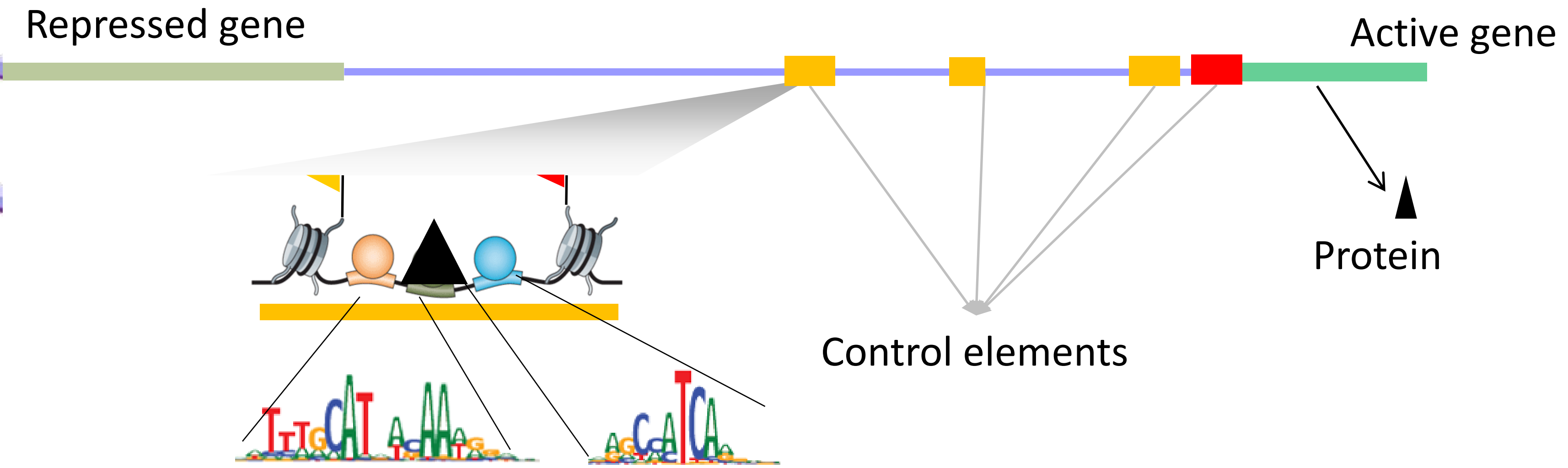
…….ACTGATCG**G**AATCG…….

Risk



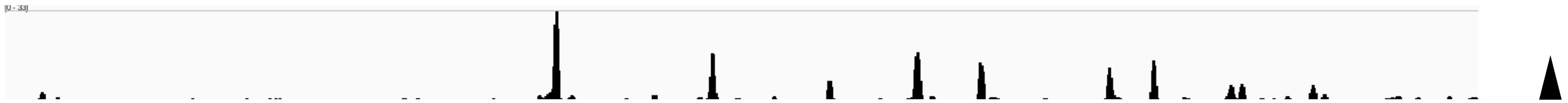
Molecular mapping of functional components of the genome

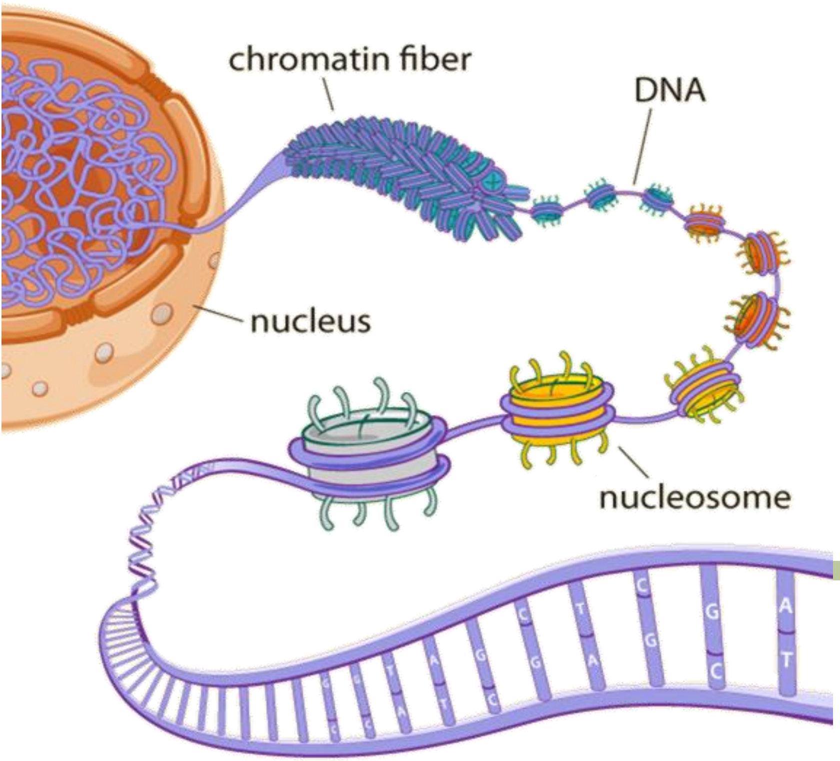




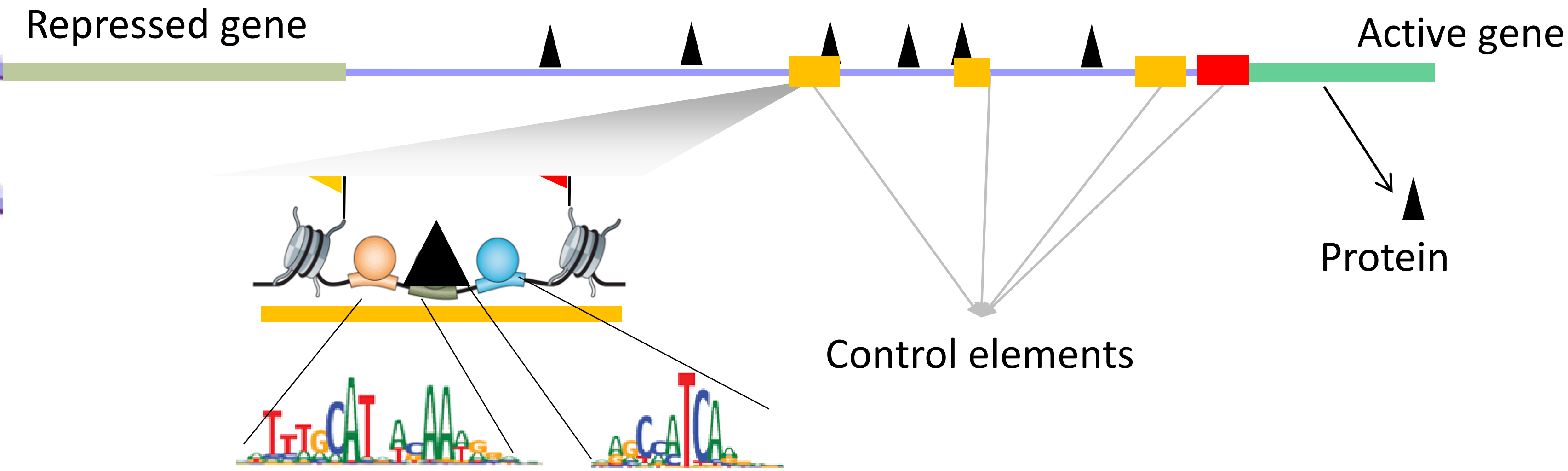
https://www.broadinstitute.org/news/1504

Molecular mapping of functional components of the genome

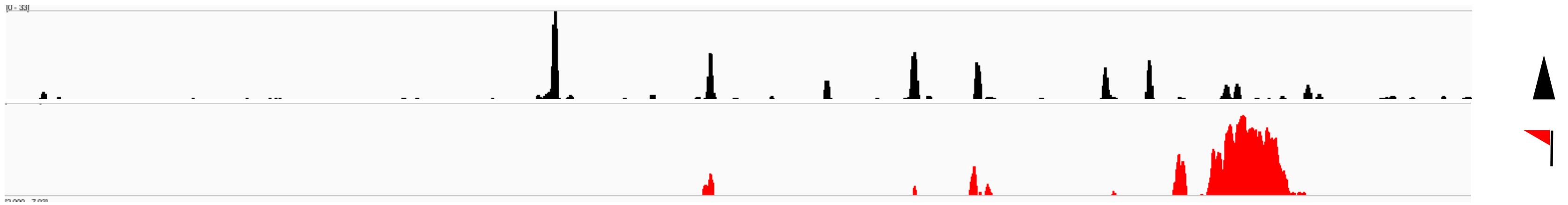


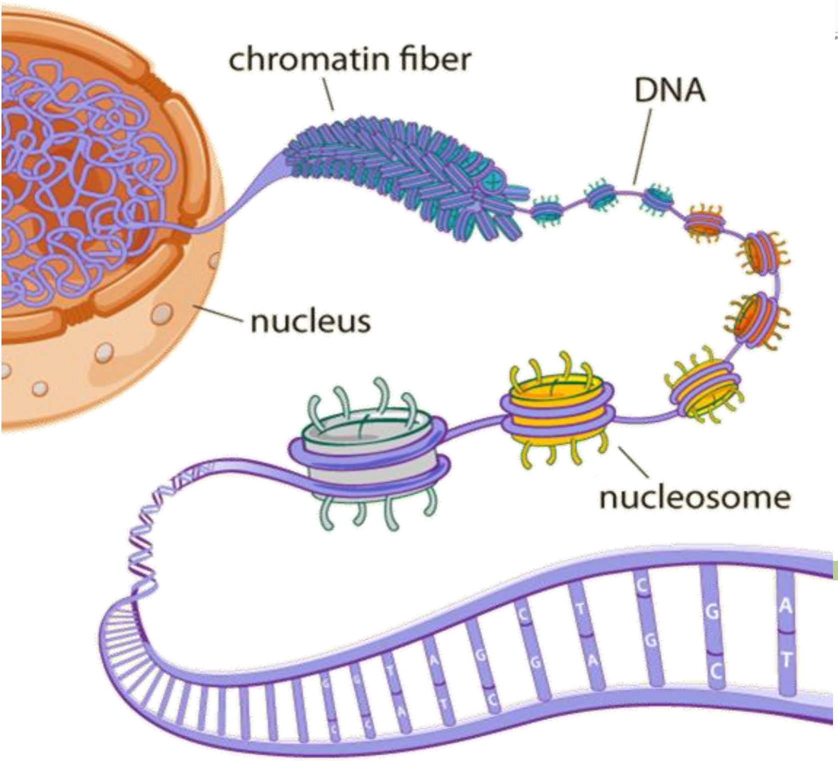


https://www.broadinstitute.org/news/1504

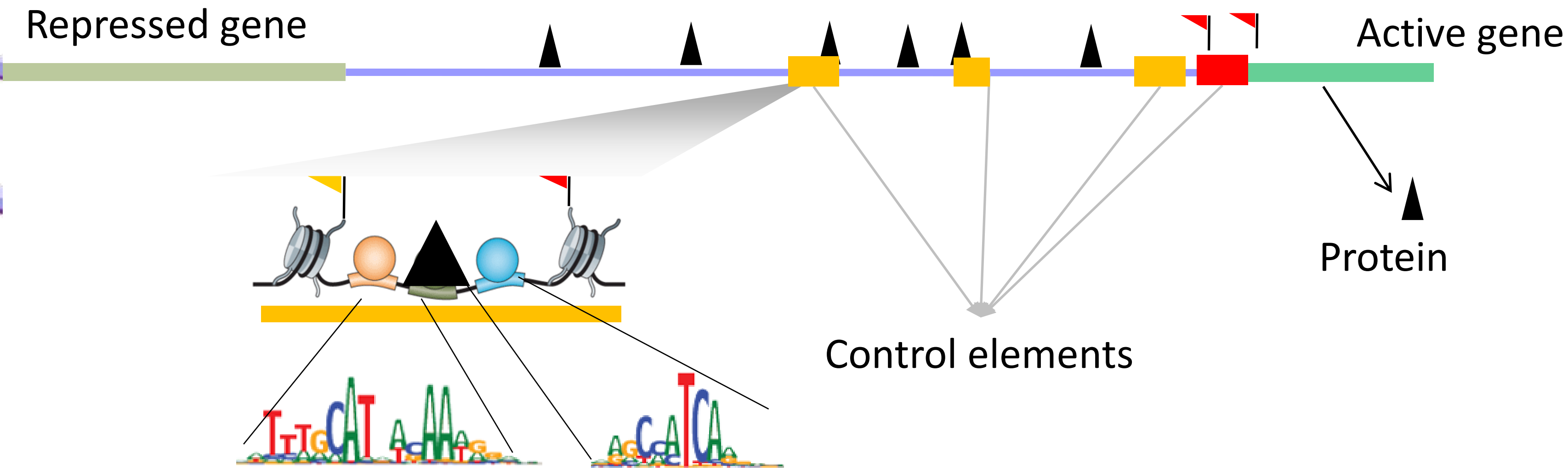


Molecular mapping of functional components of the genome

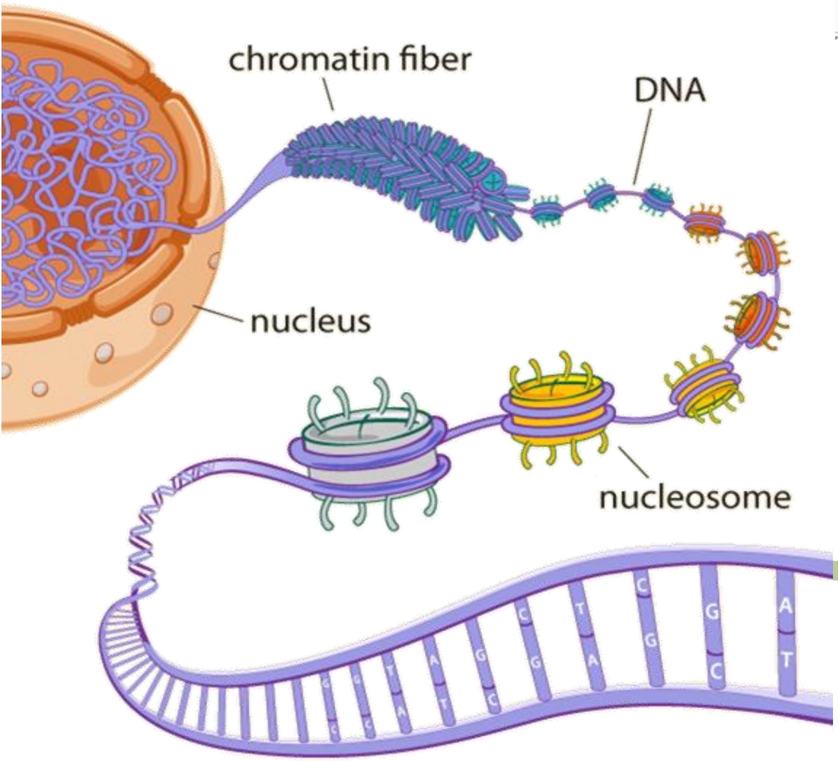




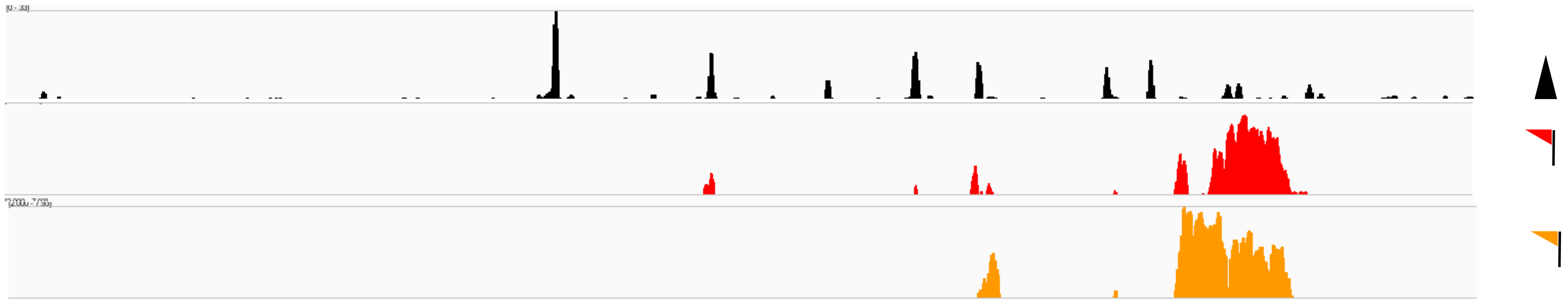
https://www.broadinstitute.org/news/1504

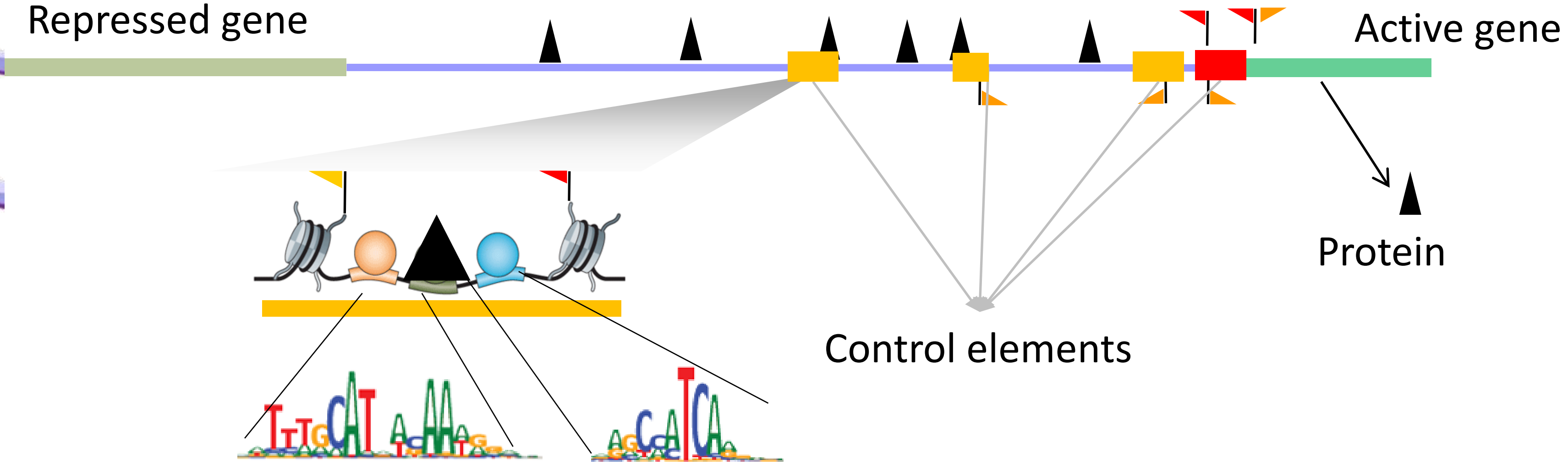


Molecular mapping of functional components of the genome

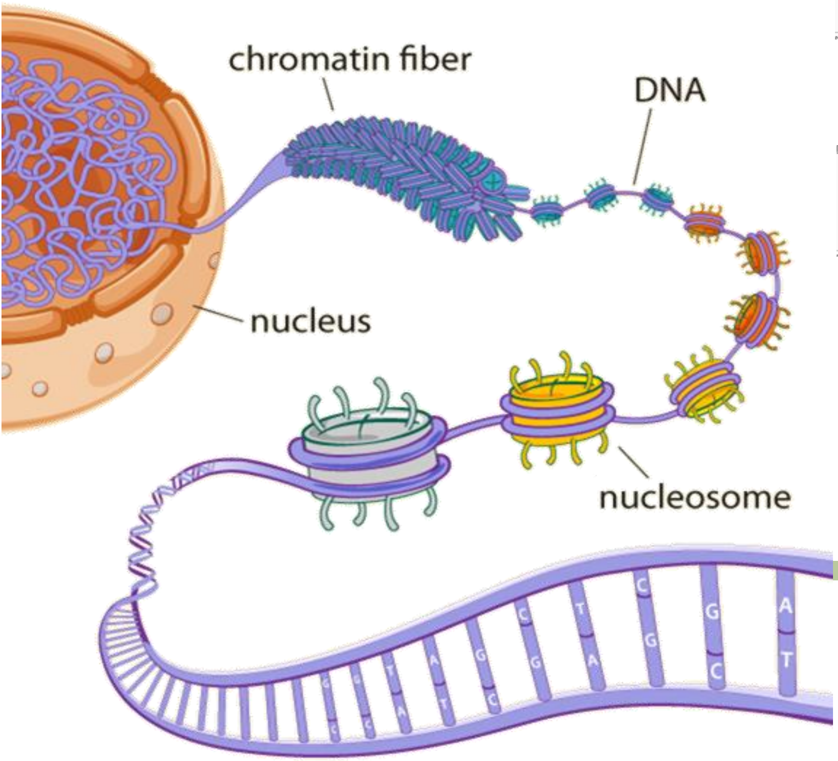


https://www.broadinstitute.org/news/1504

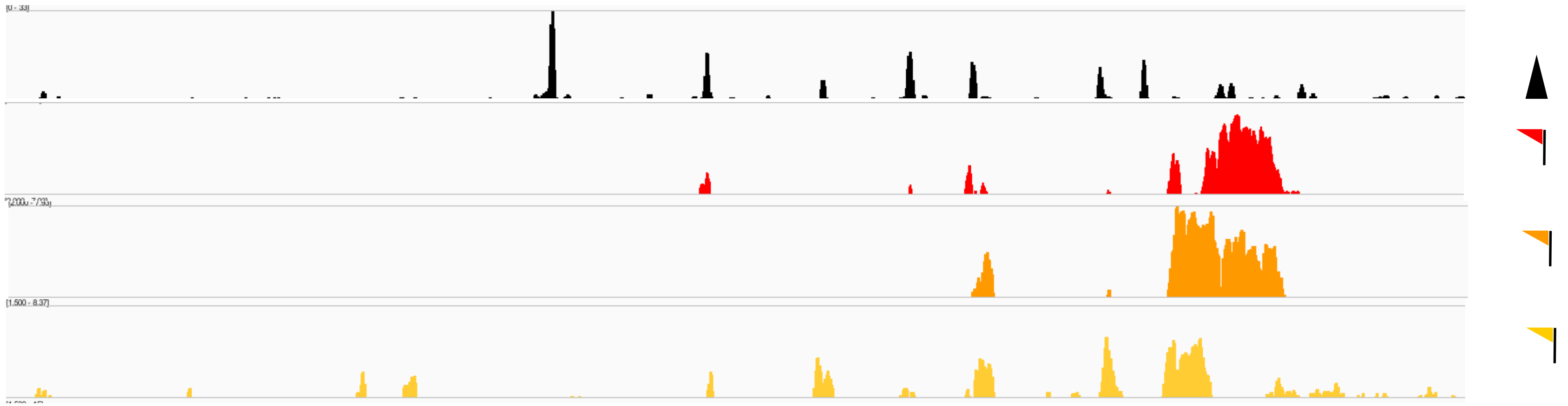


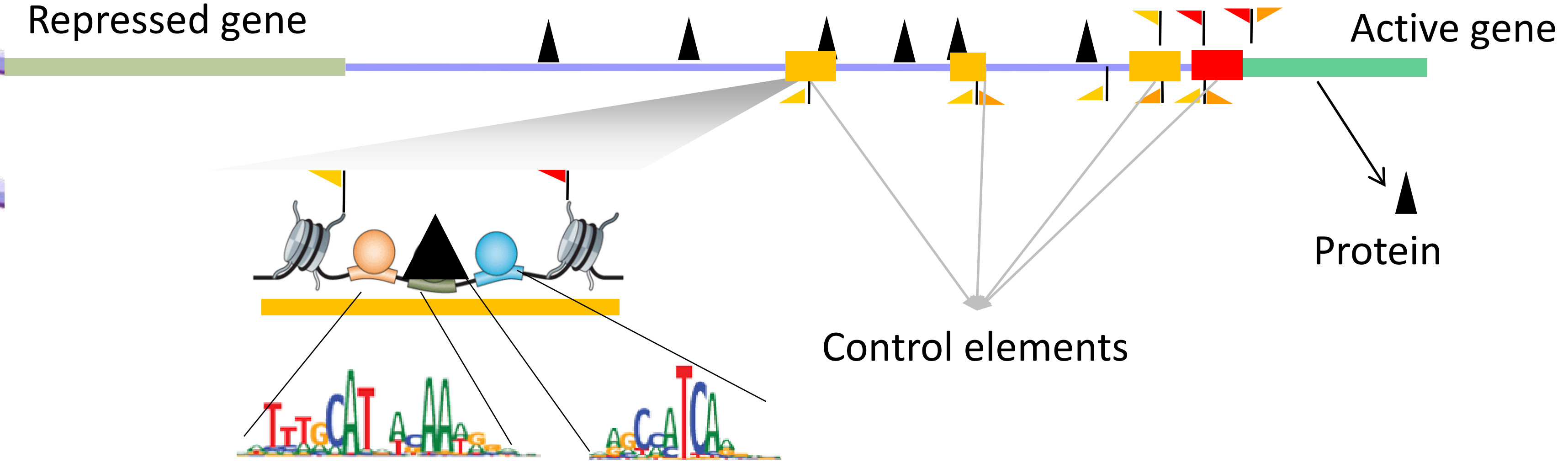


Molecular mapping of functional components of the genome

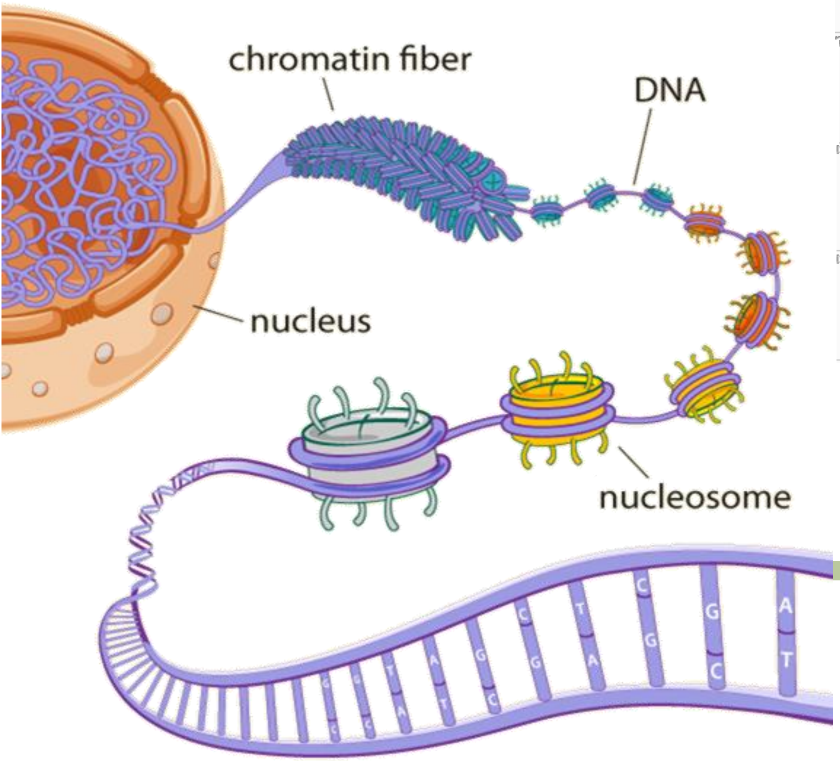


https://www.broadinstitute.org/news/1504

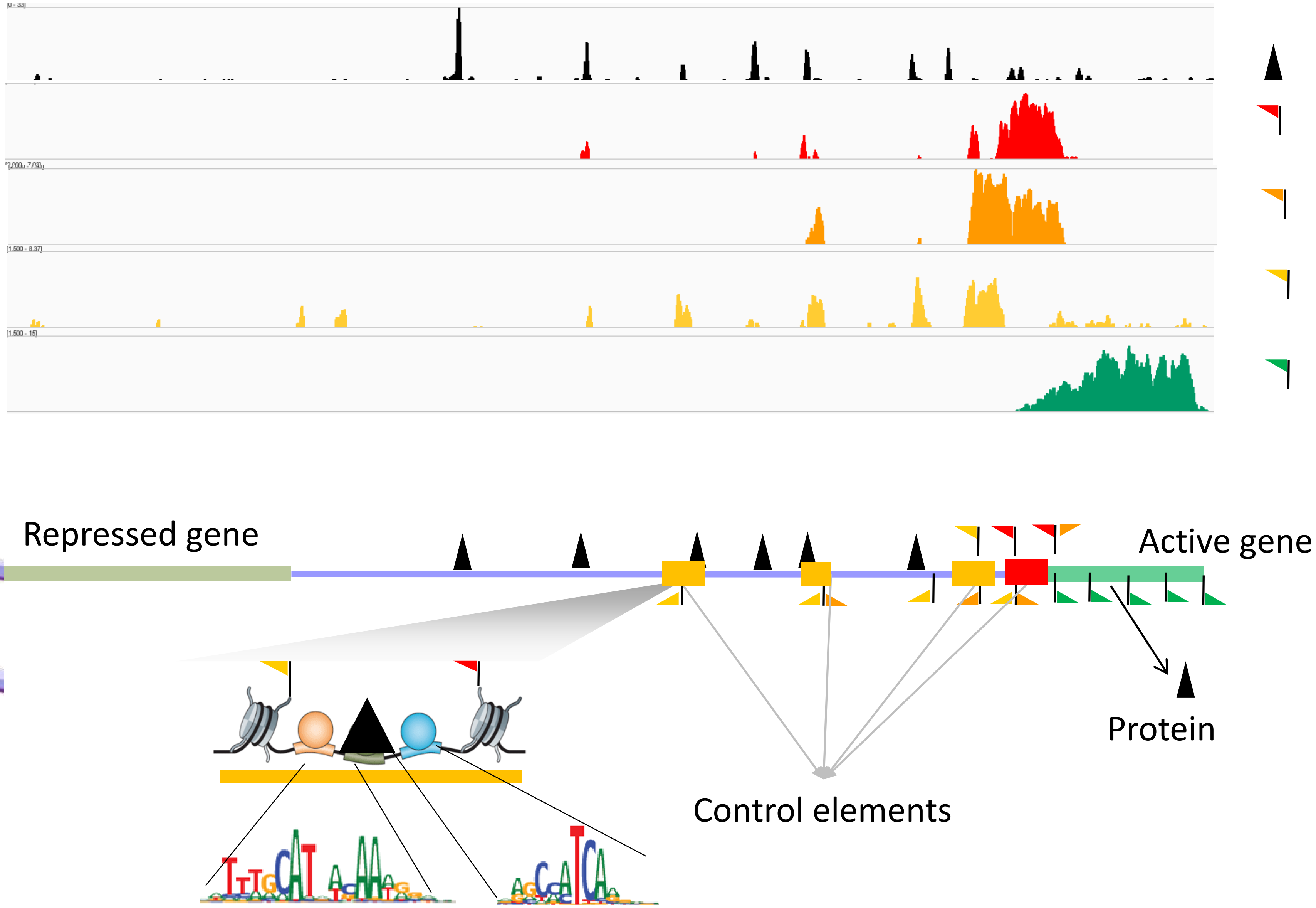




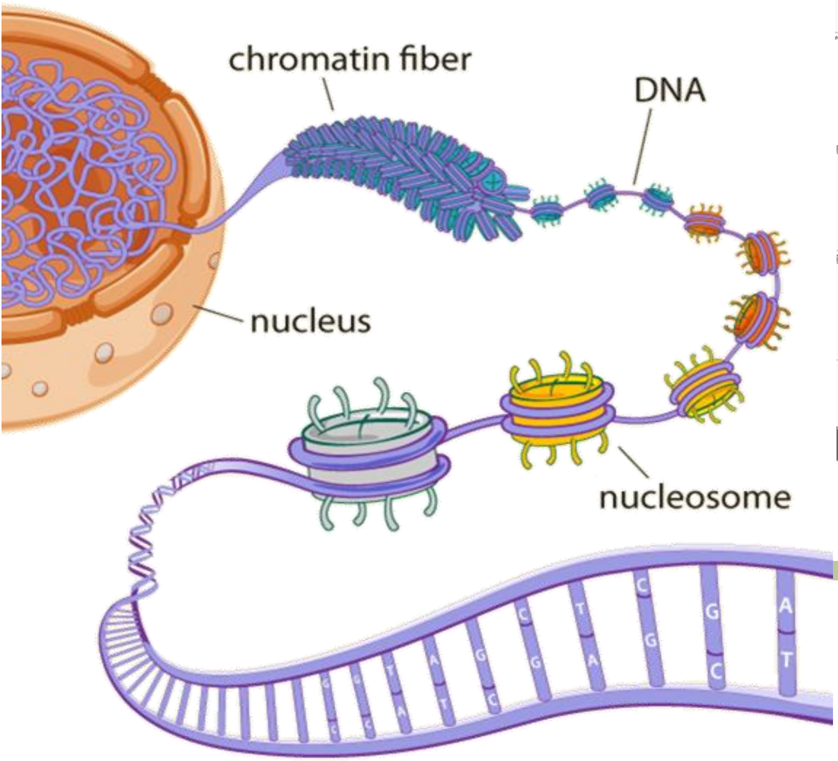
Molecular mapping of functional components of the genome



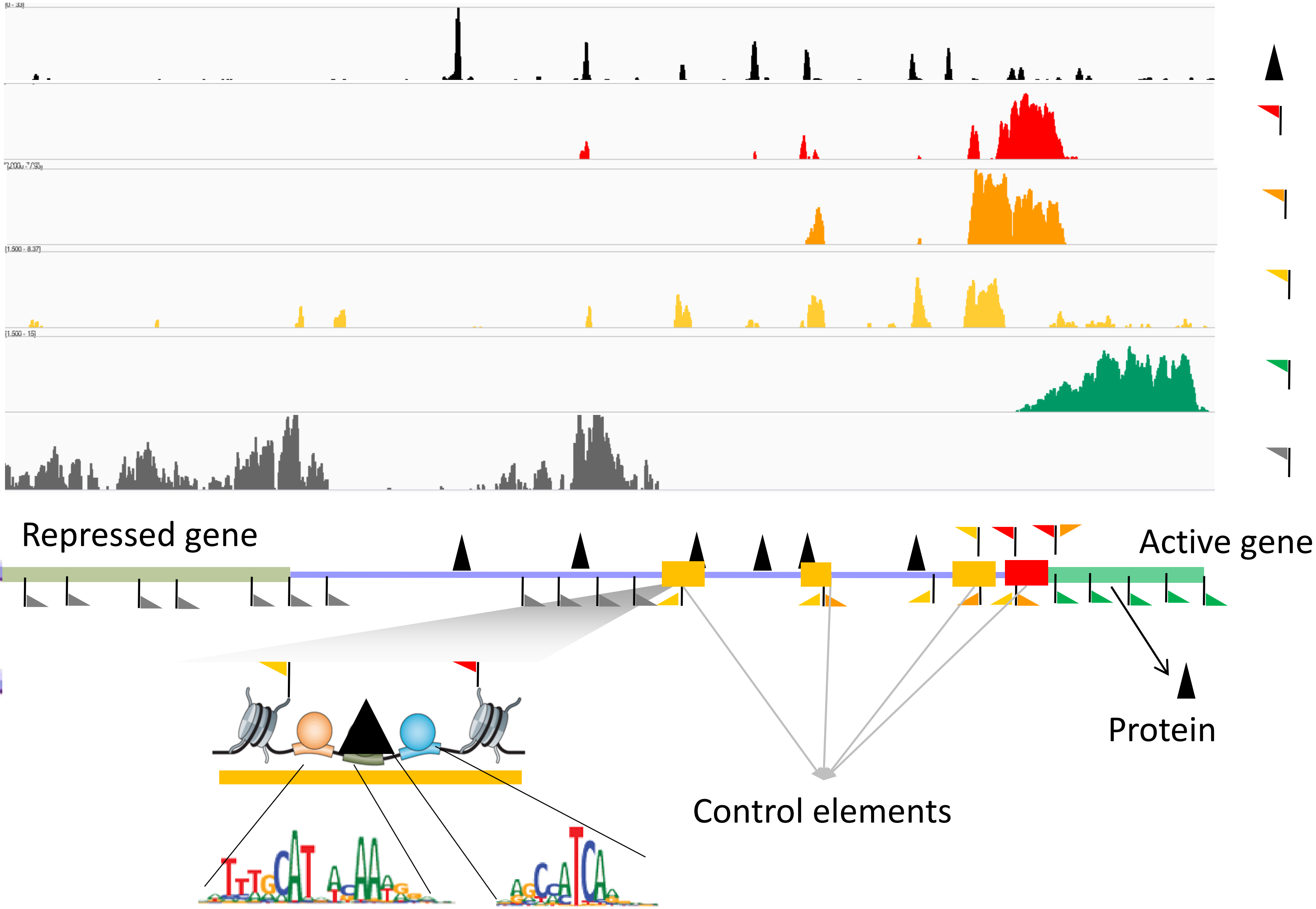
https://www.broadinstitute.org/news/1504



Molecular mapping of functional components of the genome

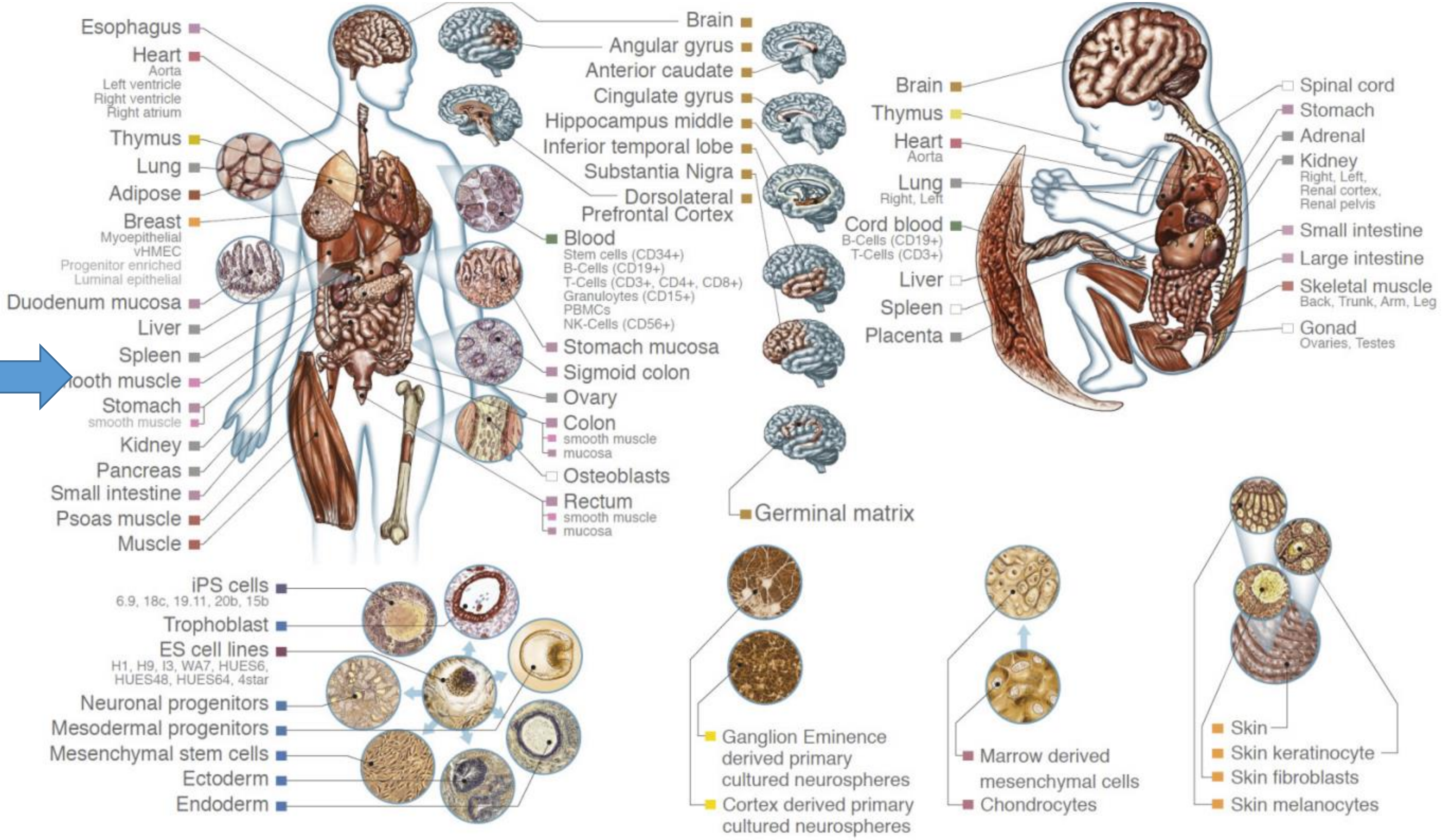


https://www.broadinstitute.org/news/1504



One genome ⇔ many cell types

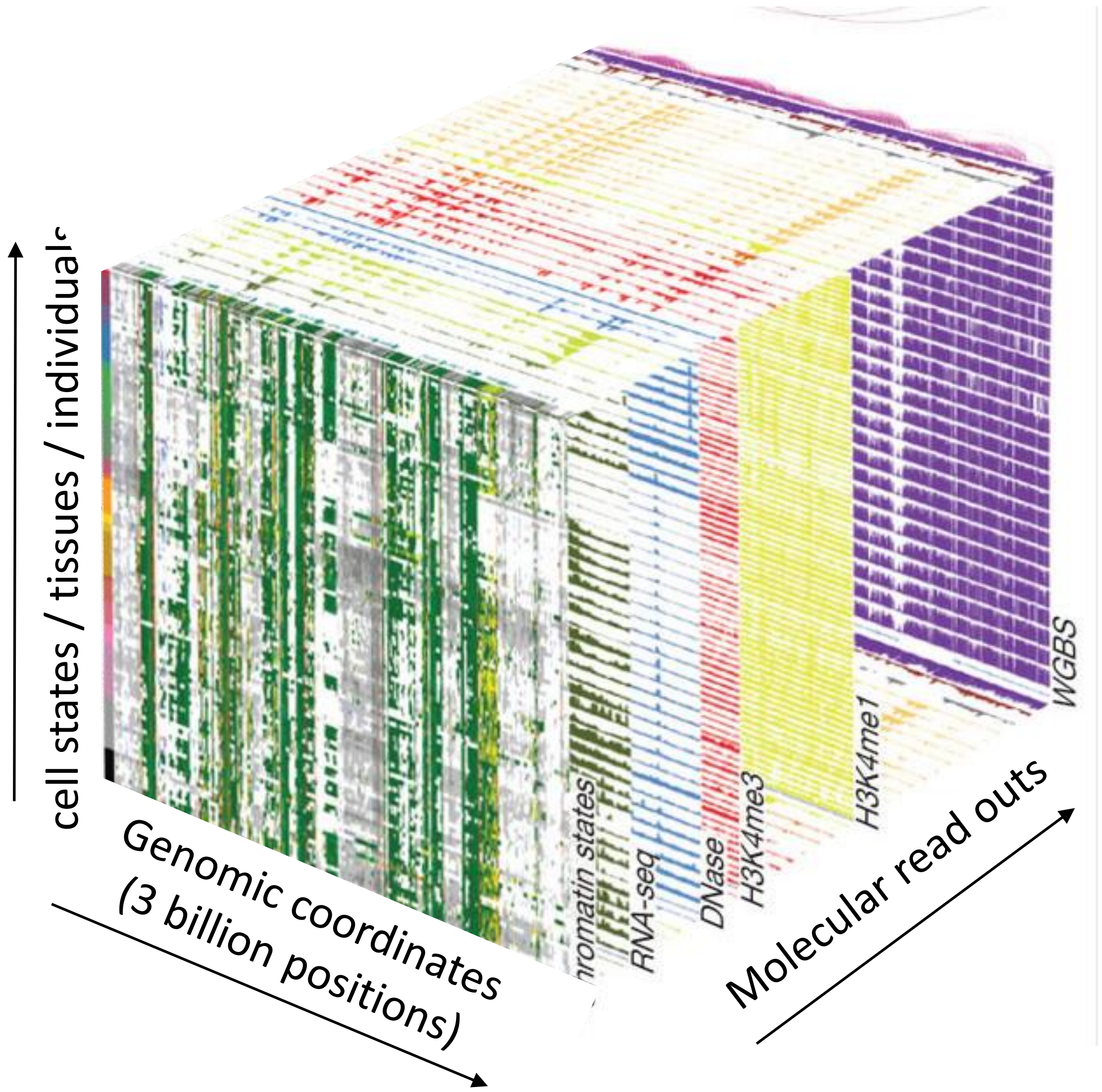
ACCAGTTACGACGG
TCAGGGTACTGATA
CCCCAAACCGTTGA
CCGCATTTACAGAC
GGGGTTTGGGTTTT
GCCCCACACAGGTA
CGTTAGCTACTGGT
TTAGCAATTTACCG
TTACAACGTTTACA
GGGTTACGGTTGGG
ATTTGAAAAAAAGT
TTGAGTTGGTTTTT
TCACGGTAGAACGT
ACCTTACAAA…………



http://www.roadmapepigenomics.org/

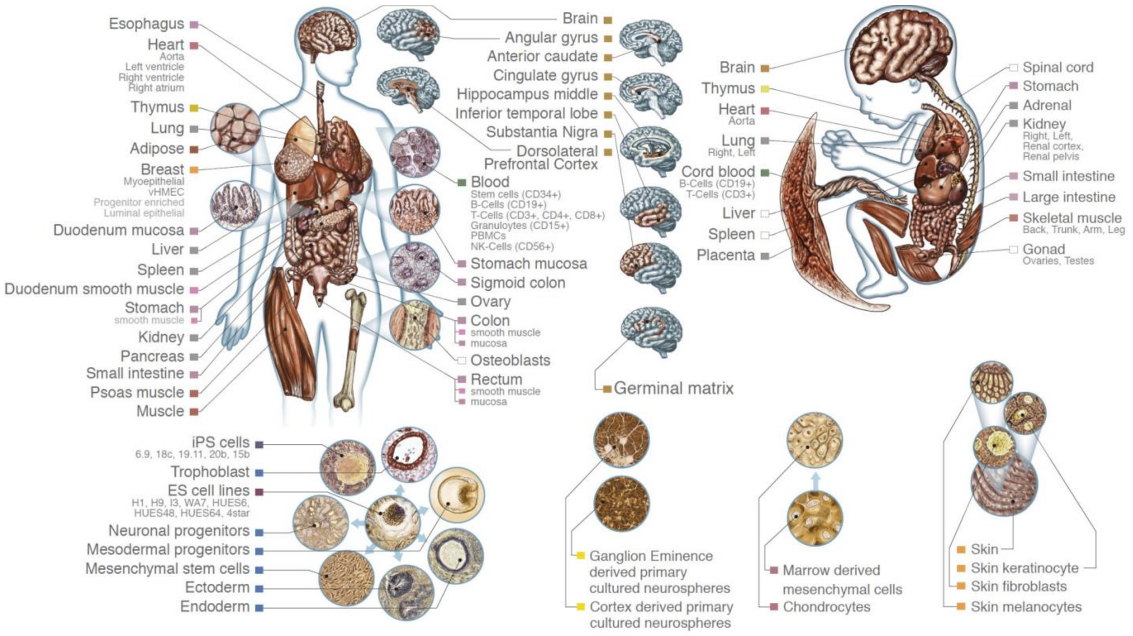
ENCODE

ROADMAP epigenomics PROJECT

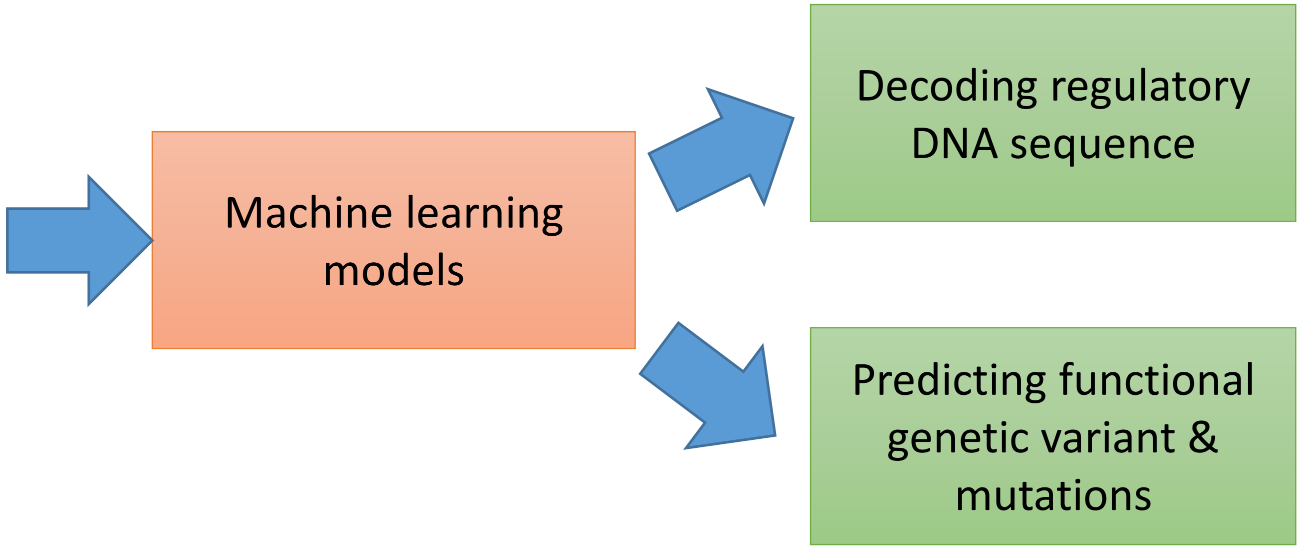


Dunham, Kundaje et al. 2012 Nature

Kundaje et al. 2015 Nature



100s of Cell-Types/Tissues



Machine learning models

Decoding regulatory DNA sequence

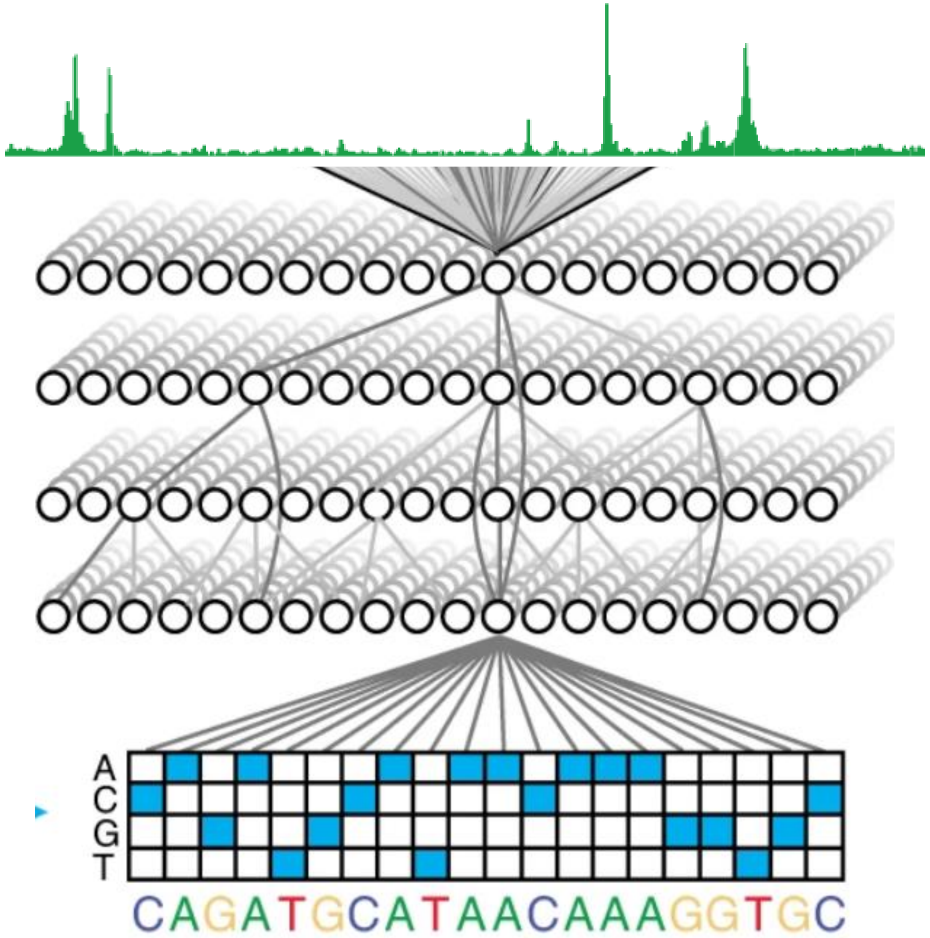
Predicting functional genetic variant & mutations

Deep learning framework for decoding regulatory DNA

Ziga Avsec

Anusri Pampari

Anna Shcherbina



BPNet

(maps sequence to base-resolution profiles)

One model for every expt.

Avsec et al. 2021, Nature Genetics
Shrikumar et al. 2017, ICML
Tseng et al. 2020, NeurIPS
Nair et al, 2022, Bioinformatics
Schreiber et al. 2022, Biorxiv

Avanti Shrikumar

Alex Tseng

Surag Nair

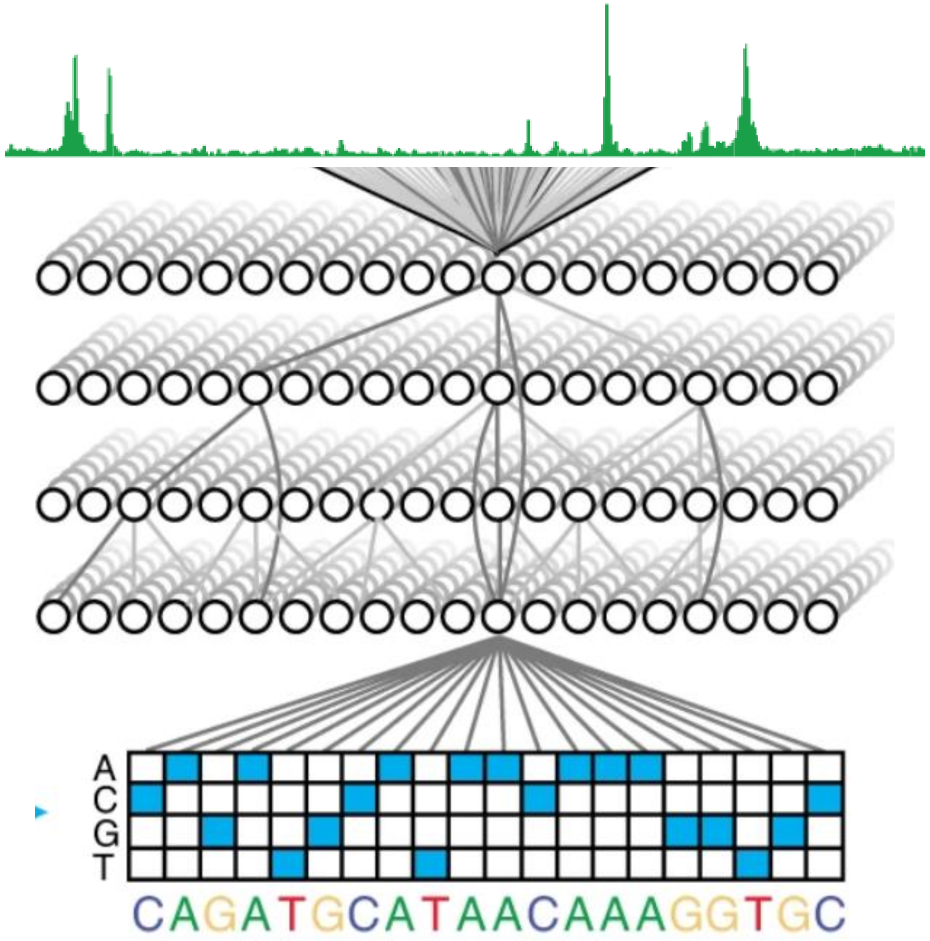
Jacob Schreiber

Deep learning framework for decoding regulatory DNA

Ziga Avsec

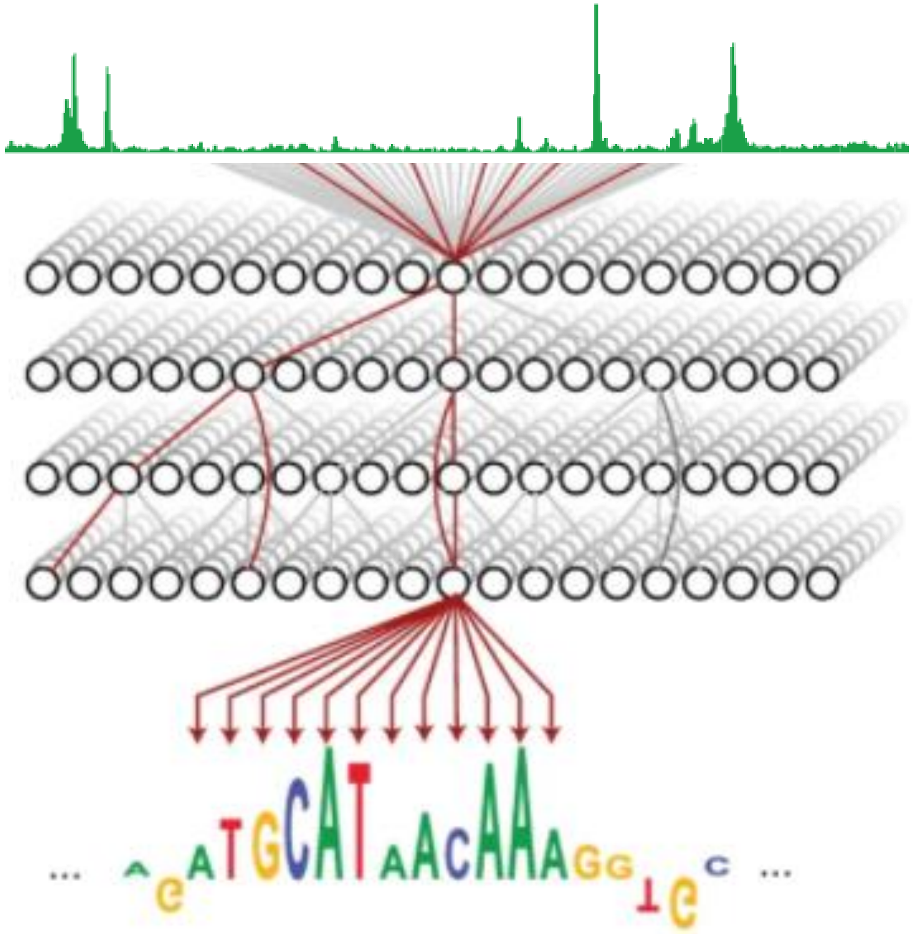
Anusri Pampari

Anna Shcherbina



BPNet

(maps sequence to base-resolution profiles)
One model for every expt.



DeepLIFT, FastISM, Yuzu, MoDISCo

(infers contribution of every base in each control sequence thru lens of model)

Avanti Shrikumar

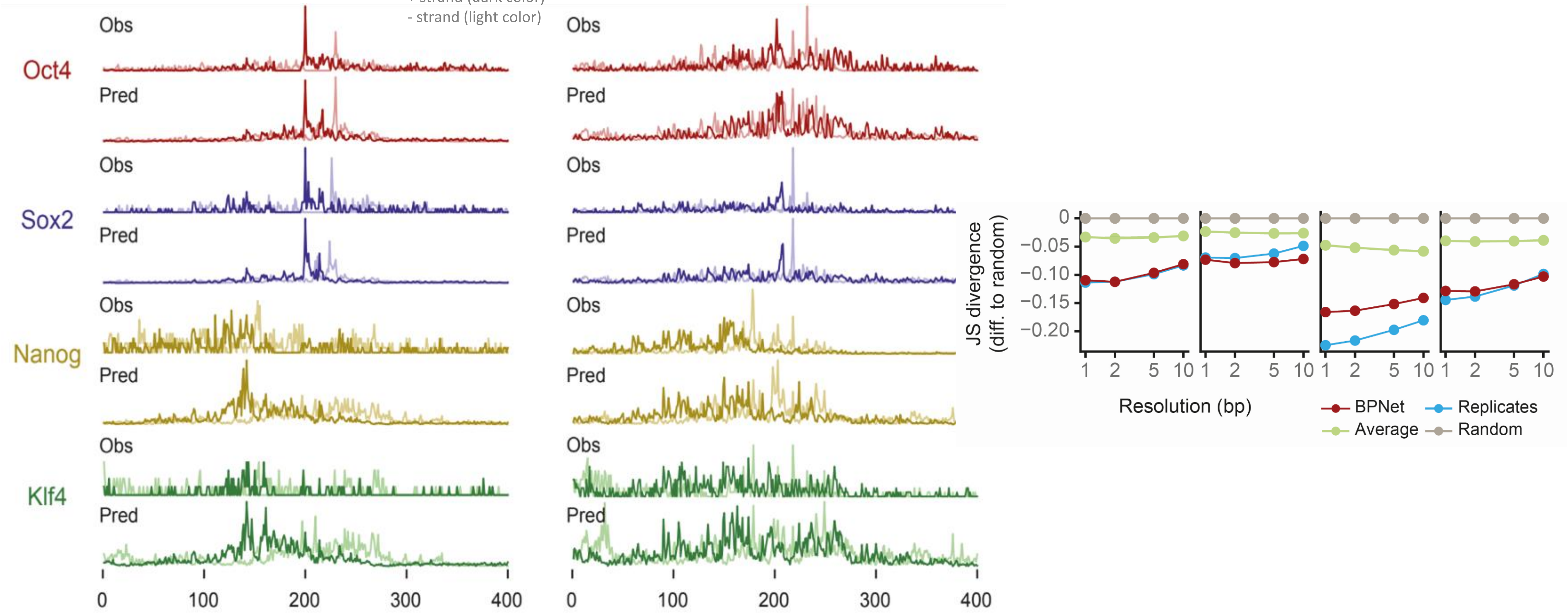
Alex Tseng

Surag Nair

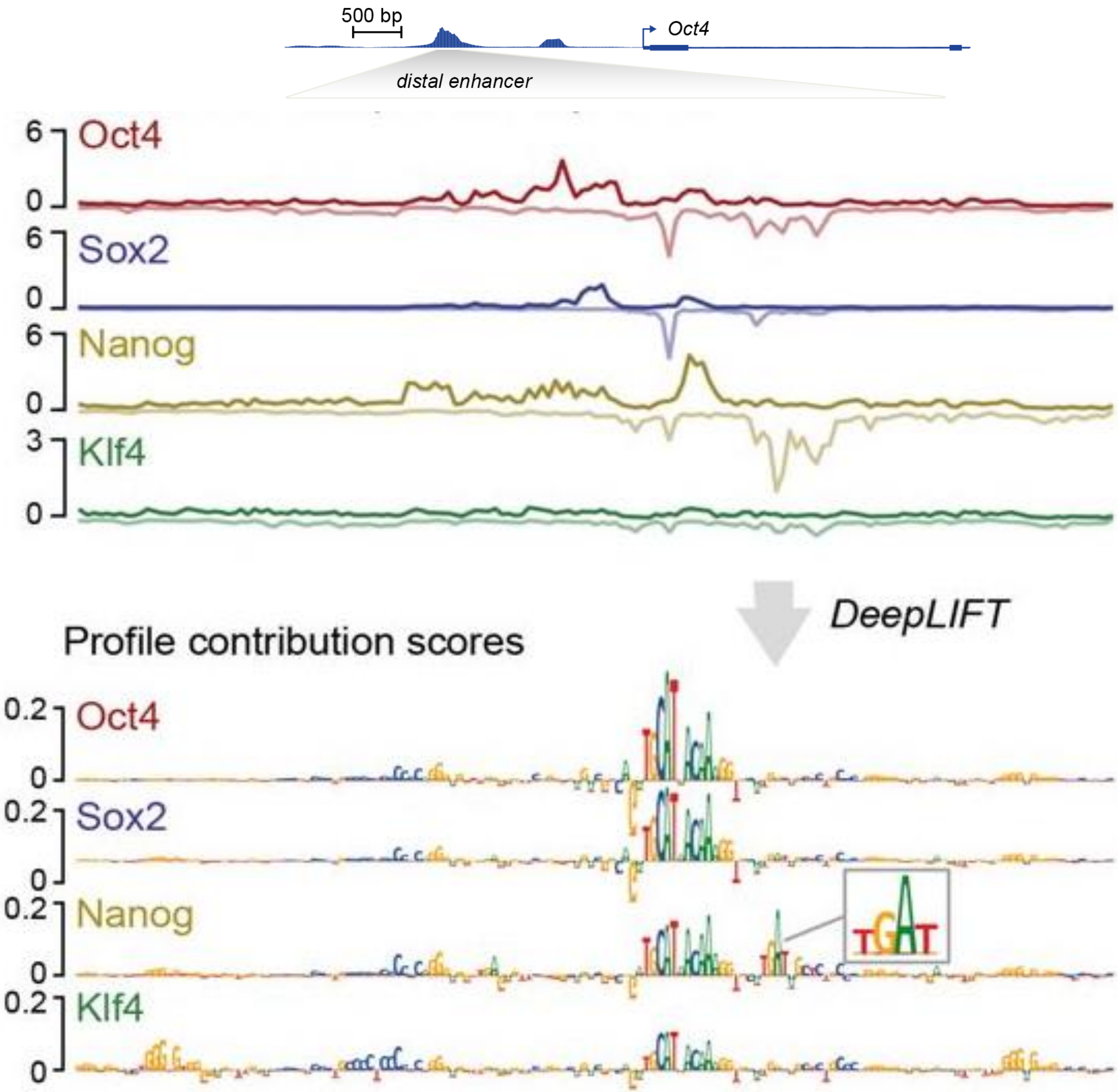
Jacob Schreiber

Avsec et al. 2021, Nature Genetics
Shrikumar et al. 2017, ICML
Tseng et al. 2020, NeurIPS
Nair et al, 2022, Bioinformatics
Schreiber et al. 2022, Biorxiv

BPNet maps DNA sequence to base-resolution molecular profiles with unprecedented accuracy (on par with concordance between replicate experiments)



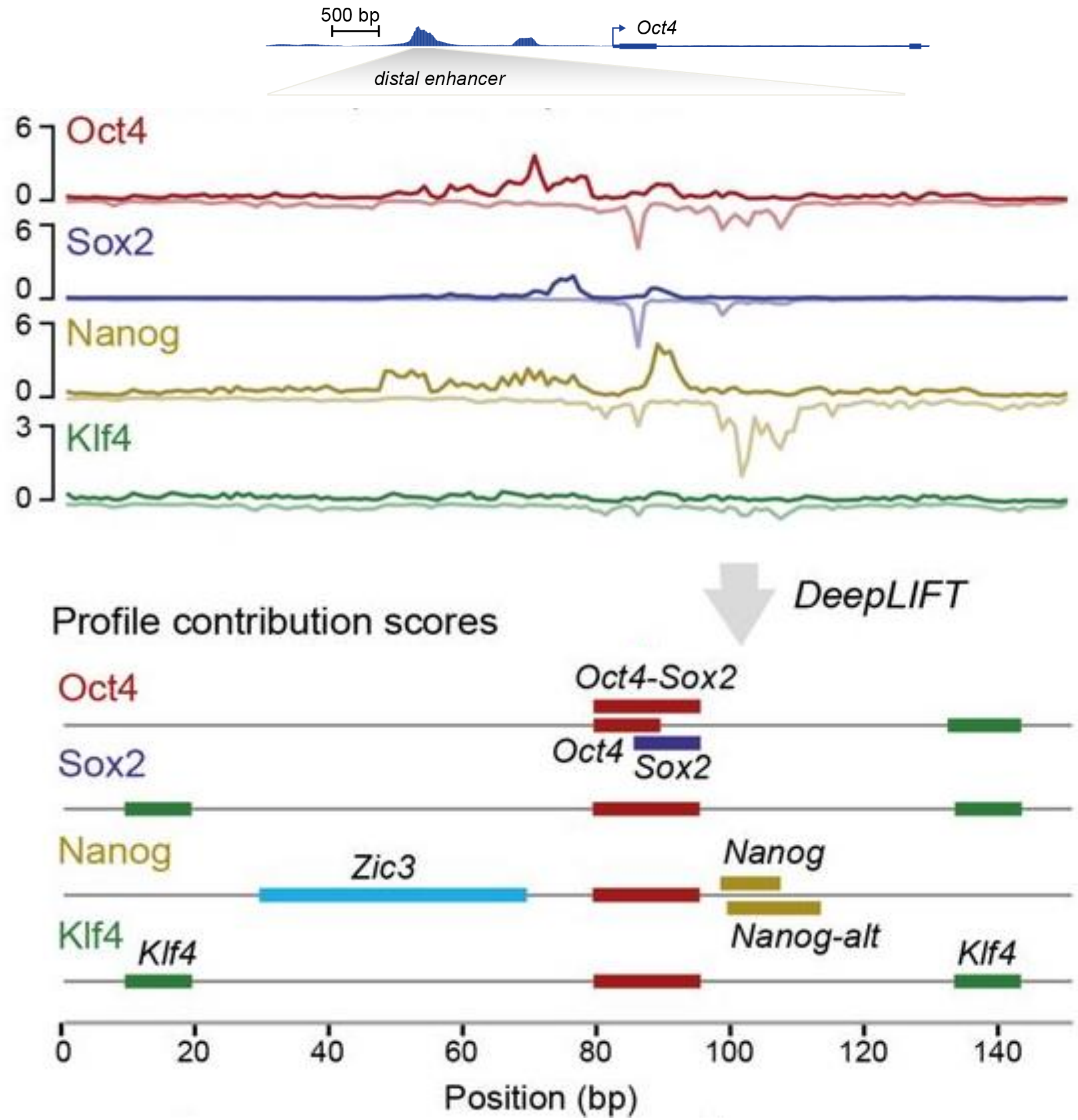
+ strand (dark color)
- strand (light color)



Avanti Shrikumar

Alex Tseng

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

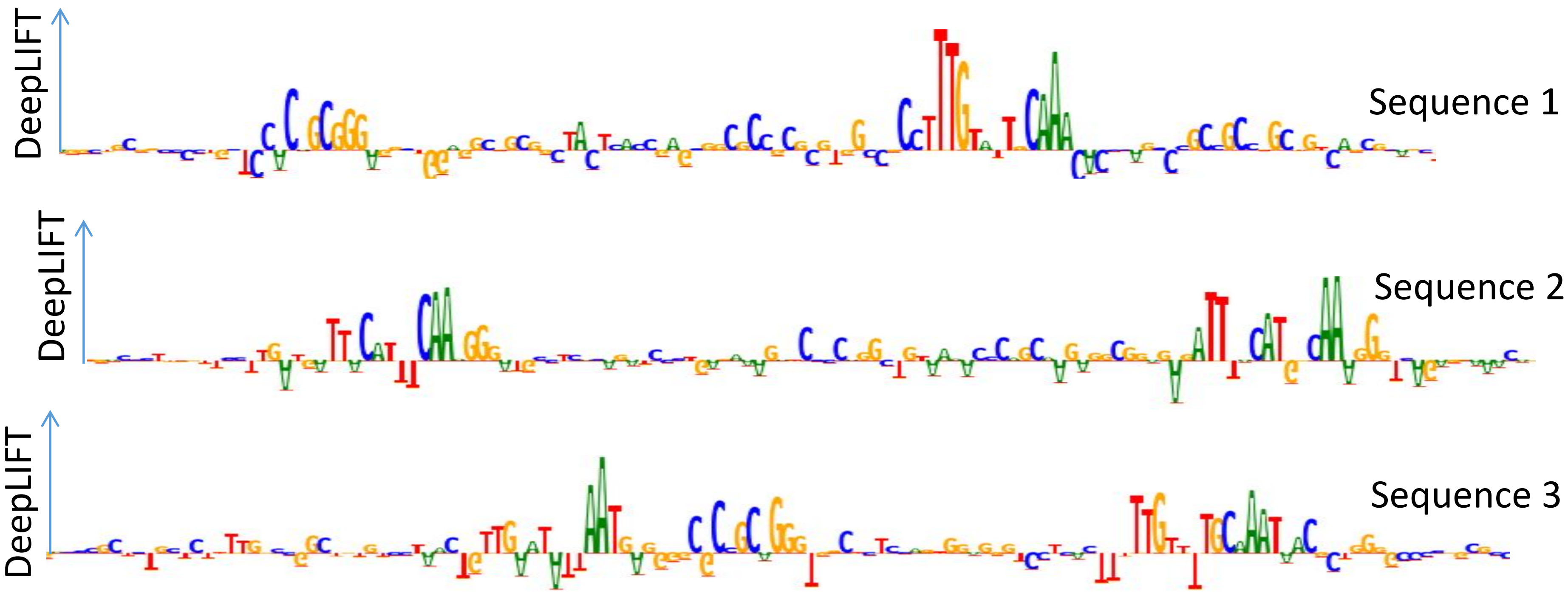


Avanti Shrikumar

Alex Tseng

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

TF-MoDISCO: Consolidate predictive subsequences into non-redundant motif representations



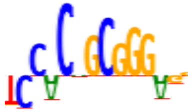
Avanti Shrikumar

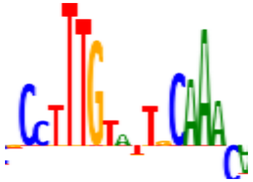
Alex Tseng

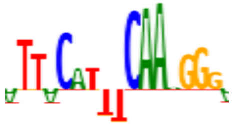
Shrikumar et al. 2018

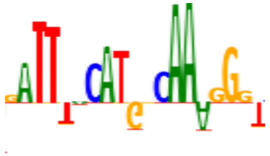
Avsec et al. Nature Genetics 2021

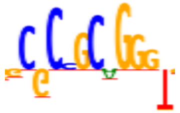
TF-MoDISCO: Consolidate predictive subsequences into non-redundant motif representations









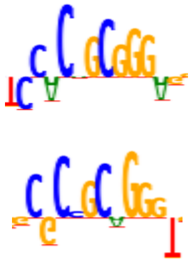


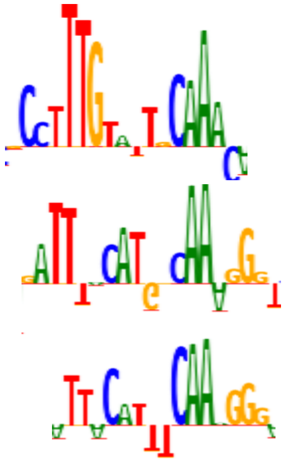
Avanti Shrikumar

Alex Tseng

Shrikumar et al. 2018
Avsec et al. Nature Genetics 2021

TF-MoDISCO: Consolidate predictive subsequences into non-redundant motif representations



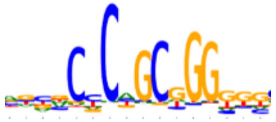


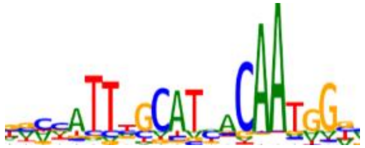
Avanti Shrikumar

Alex Tseng

Shrikumar et al. 2018
Avsec et al. Nature Genetics 2021

TF-MoDISCO: Consolidate predictive subsequences into non-redundant motif representations



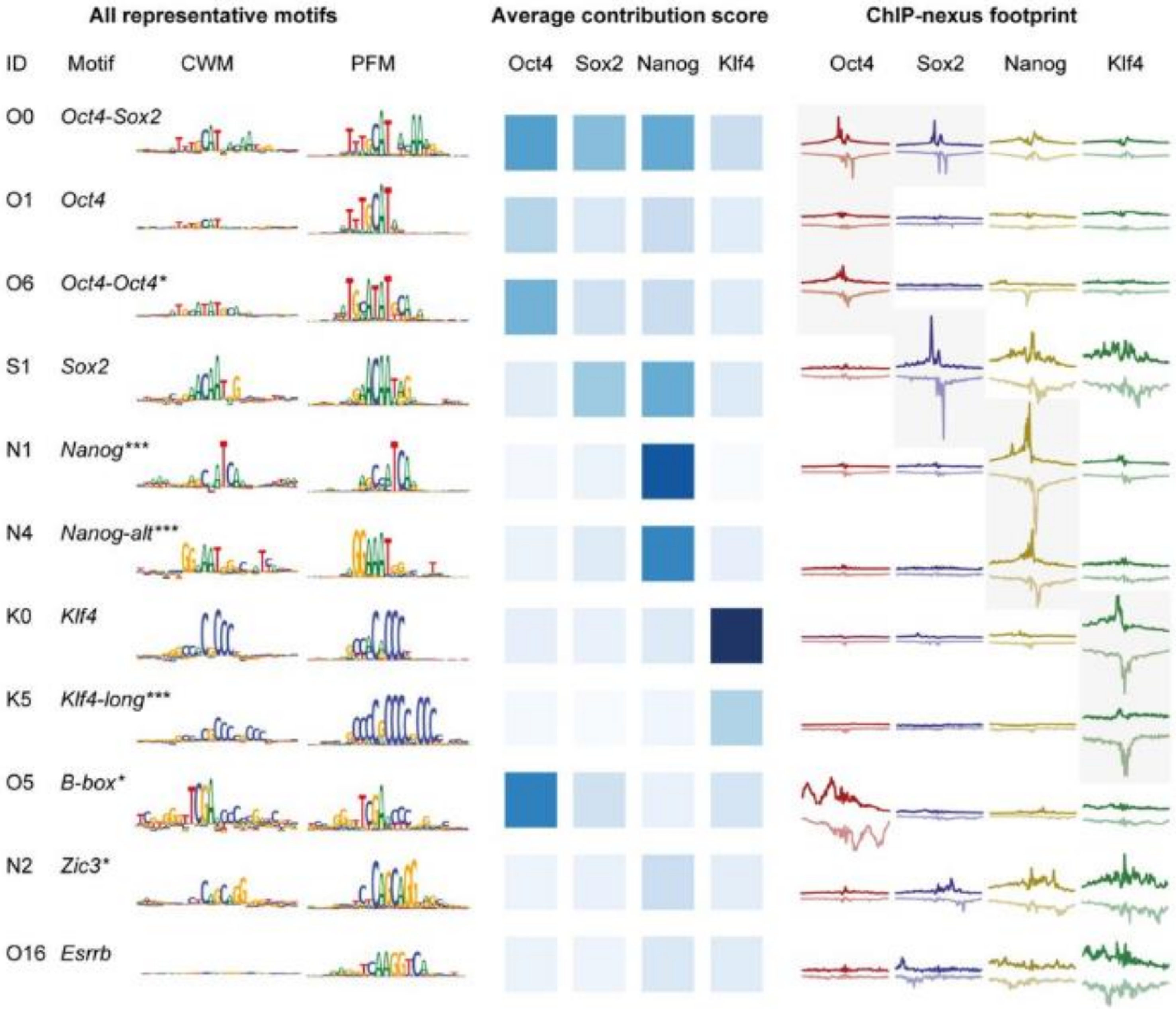


Avanti Shrikumar

Alex Tseng

Shrikumar et al. 2018
Avsec et al. Nature Genetics 2021

Complex repertoire of motifs due to cooperative binding



50 motifs for 4 proteins!

Syntax discovery using *in-silico* perturbations

Use BPNet model as in-silico oracle to perform perturbation experiments

1) On synthetic sequences

Syntax discovery using *in-silico* perturbations

Use BPNet model as in-silico oracle to perform perturbation experiments

1) On synthetic sequences

Syntax discovery using *in-silico* perturbations

Use BPNet model as in-silico oracle to perform perturbation experiments

1) On synthetic sequences

2) By mutating genomic sequences

Syntax discovery using *in-silico* perturbations

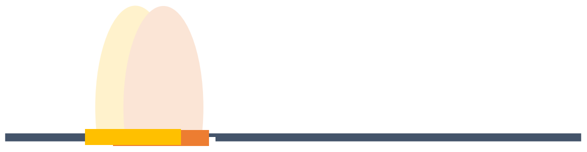
Use BPNet model as in-silico oracle to perform perturbation experiments

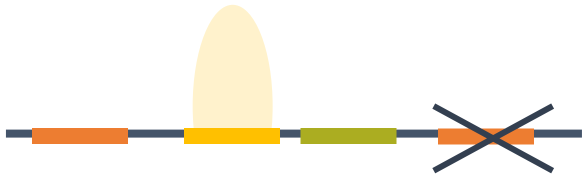
1) On synthetic sequences

2) By mutating genomic sequences

Syntax discovery using *in-silico* perturbations

Use BPNet model as in-silico oracle to perform perturbation experiments



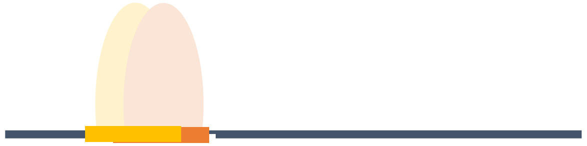


1) On synthetic sequences

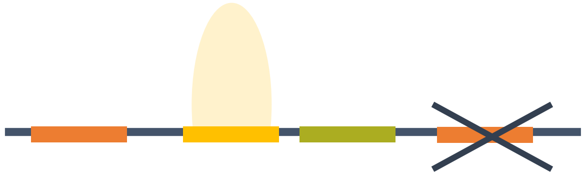
2) By mutating genomic sequences

Syntax discovery using *in-silico* perturbations

Use BPNet model as in-silico oracle to perform perturbation experiments



1) On synthetic sequences

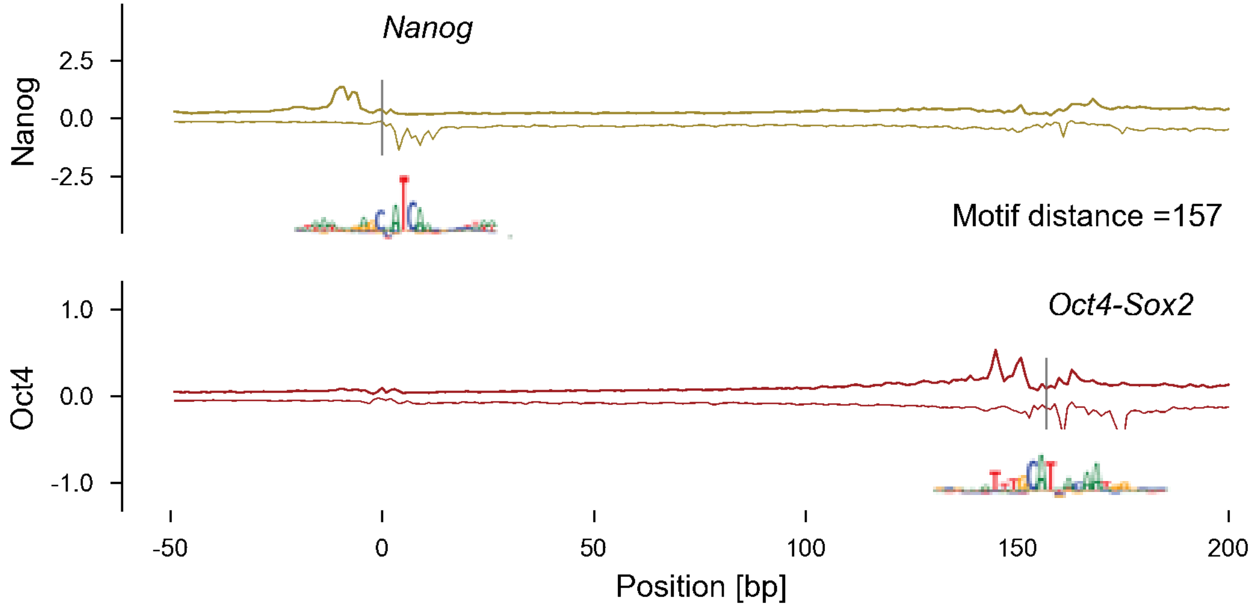


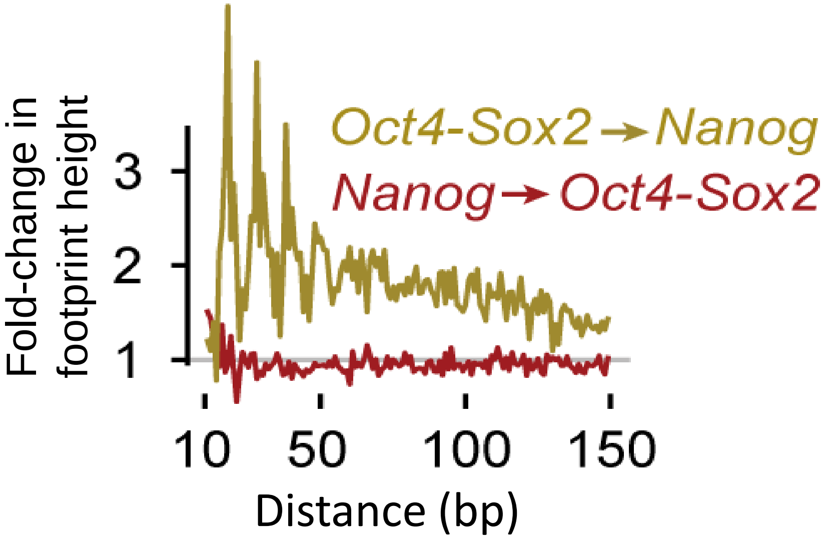
2) By mutating genomic sequences

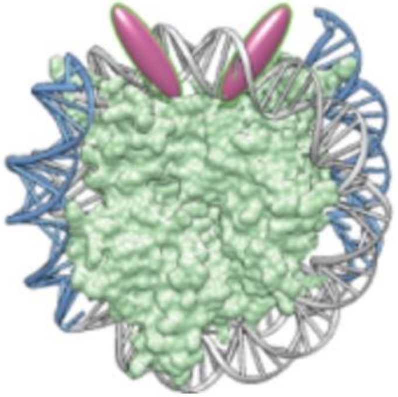
***In silico* biochemistry**

***In silico* genetics**

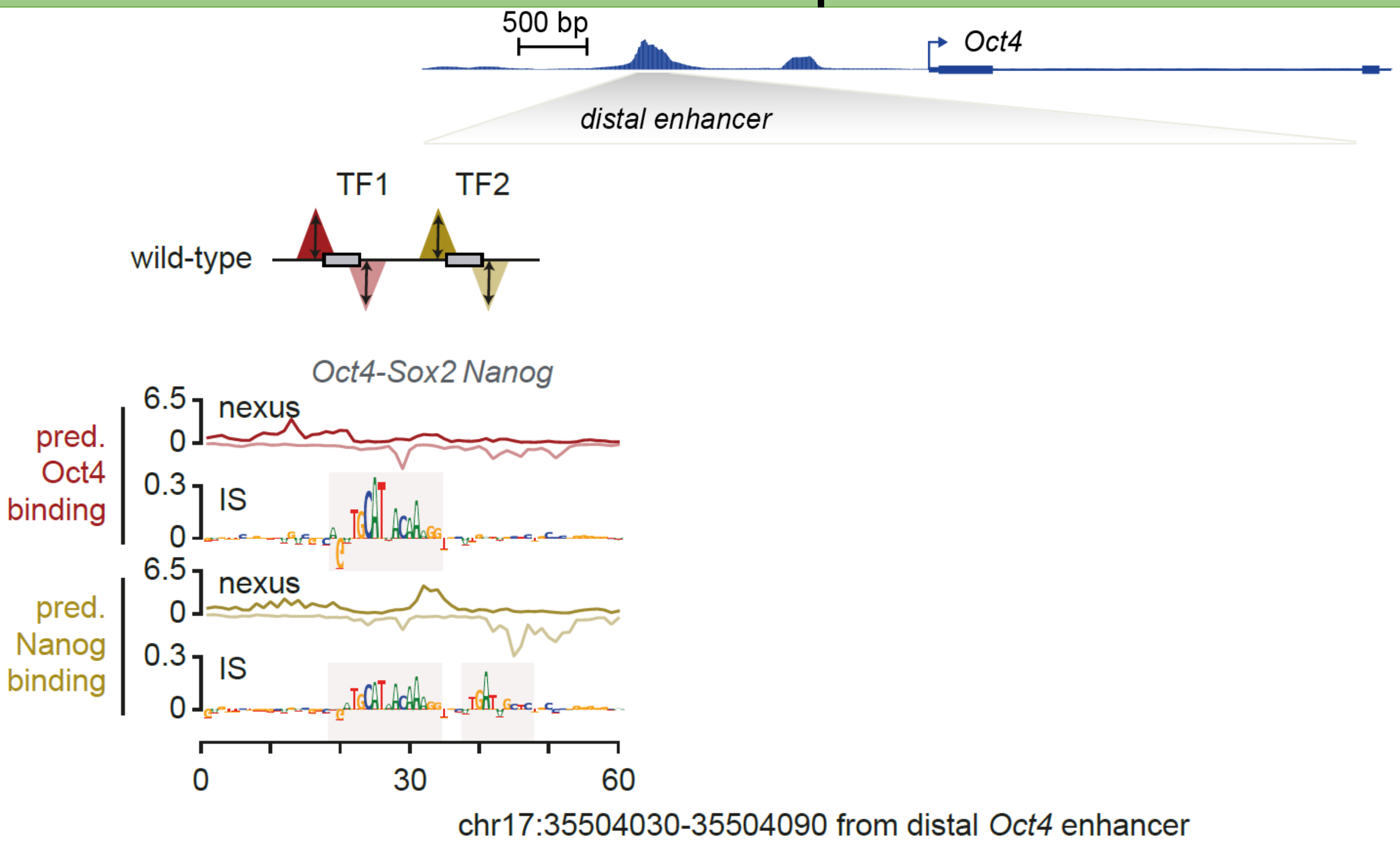
In-silico reporters: Designing synthetic sequences to query models to reveal syntax



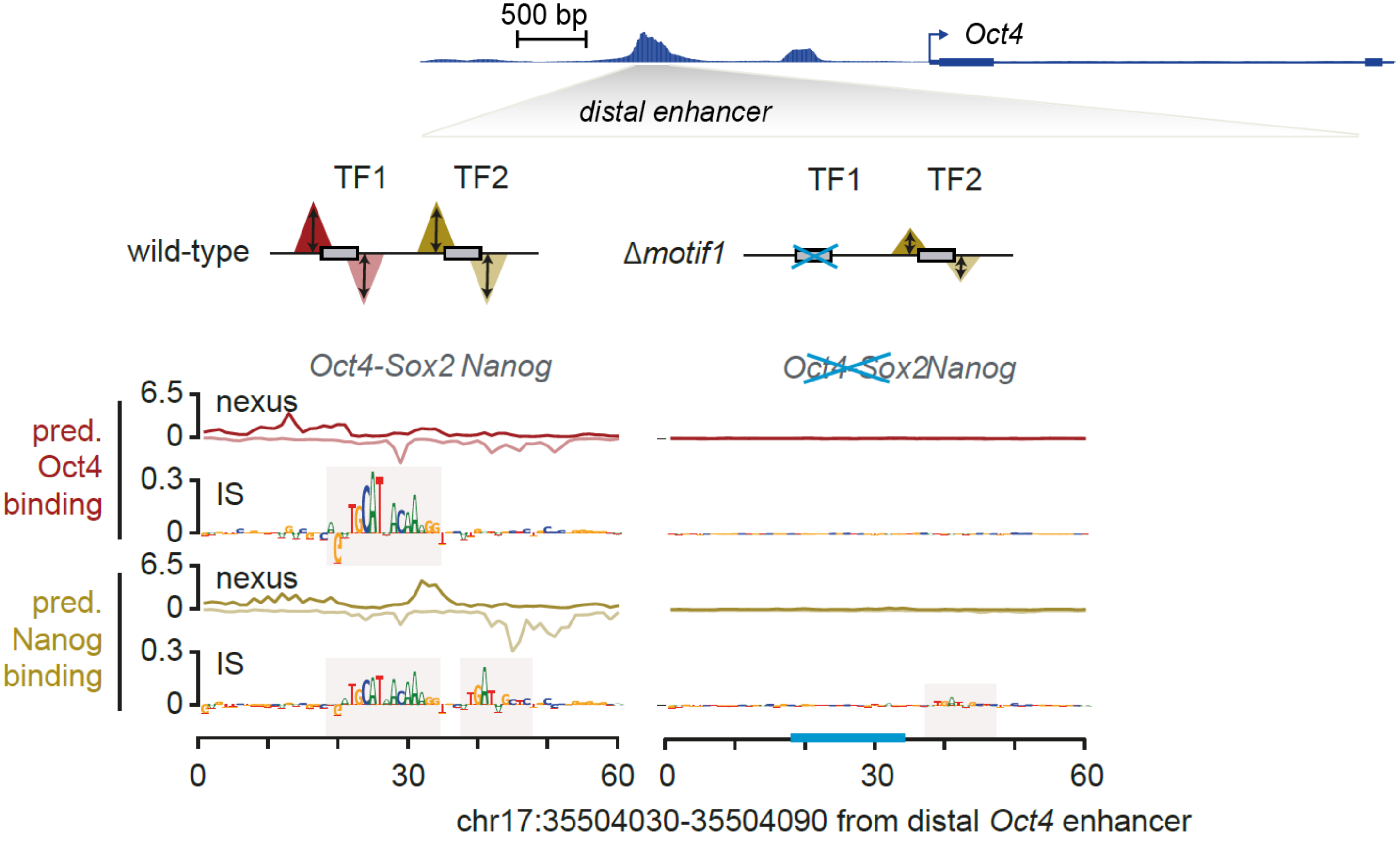




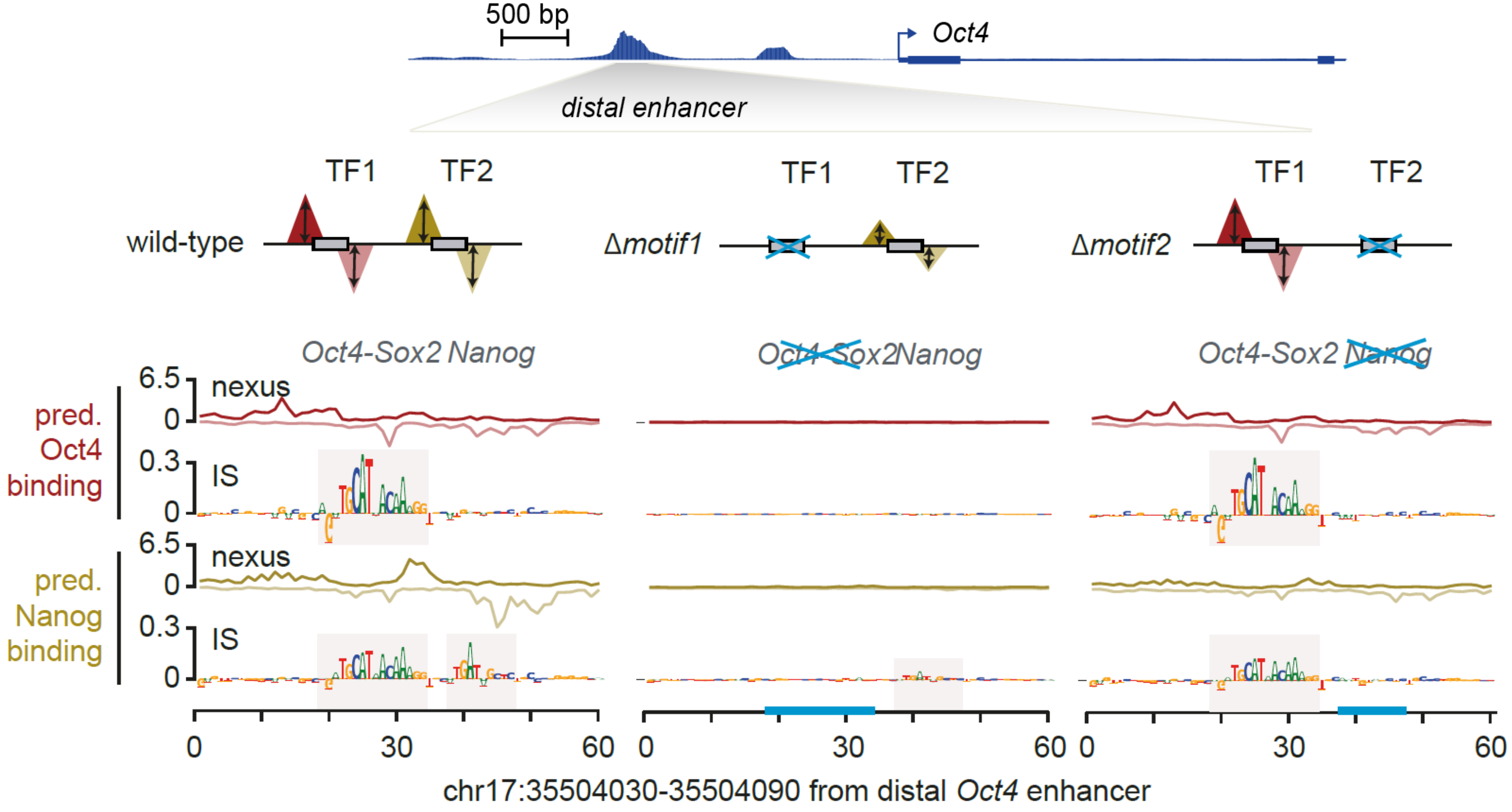
in-silico genome editing: Deciphering syntax by perturbing genomic sequences



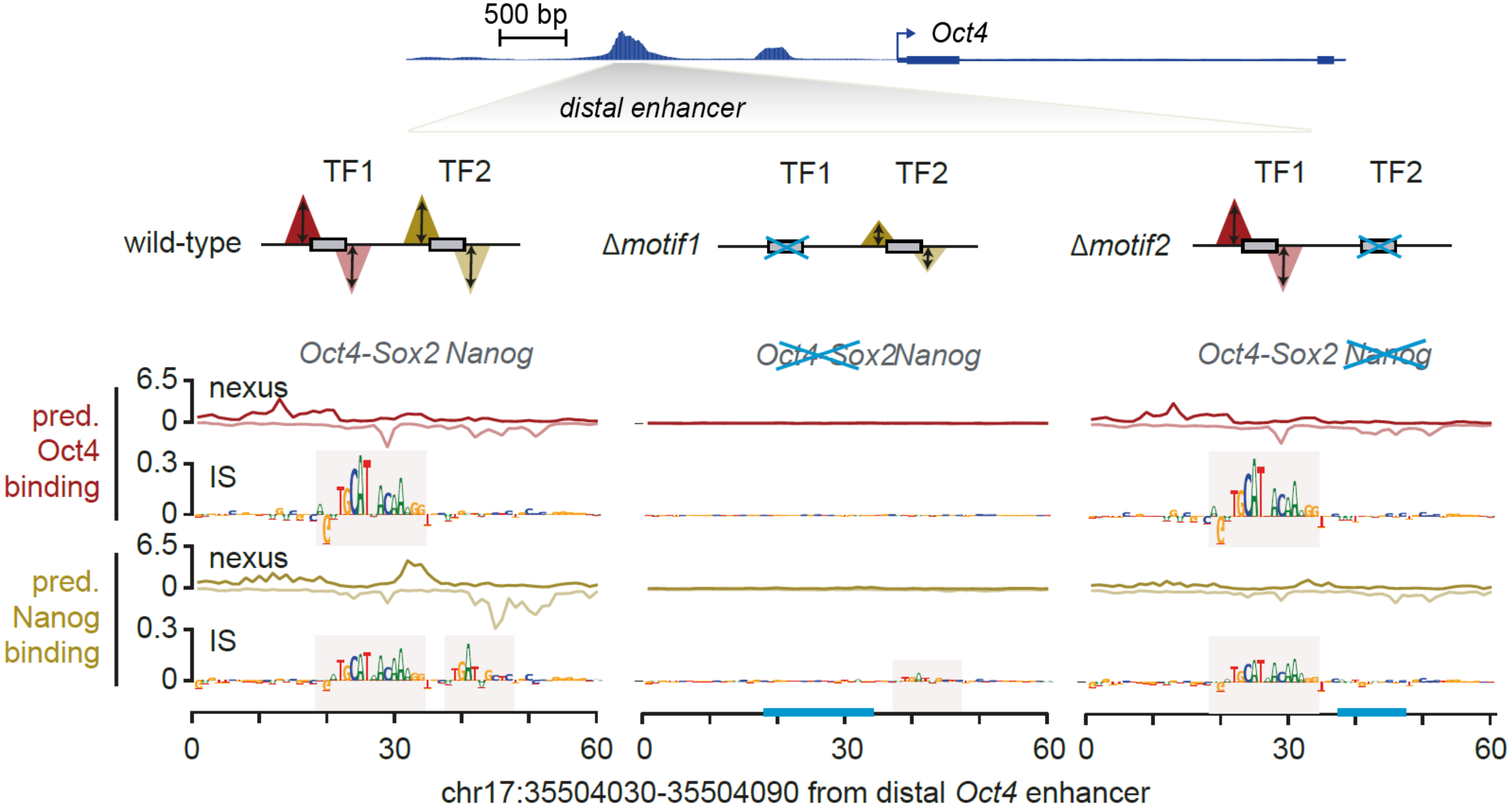
in-silico genome editing: Deciphering syntax by perturbing genomic sequences

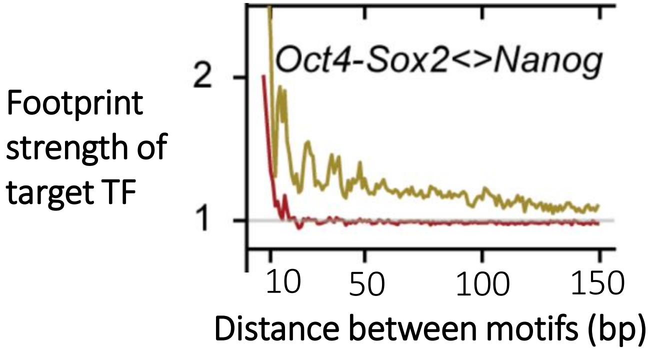


in-silico genome editing: Deciphering syntax by perturbing genomic sequences

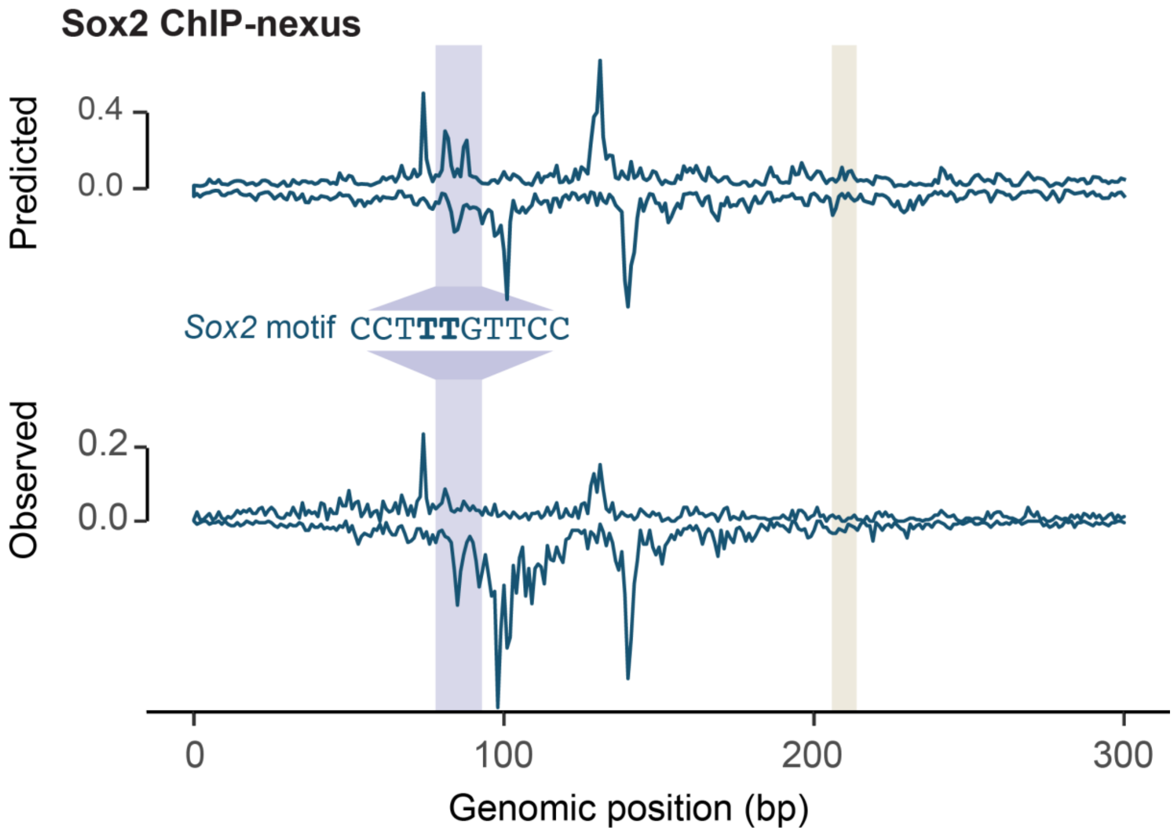


in-silico genome editing: Deciphering syntax by perturbing genomic sequences



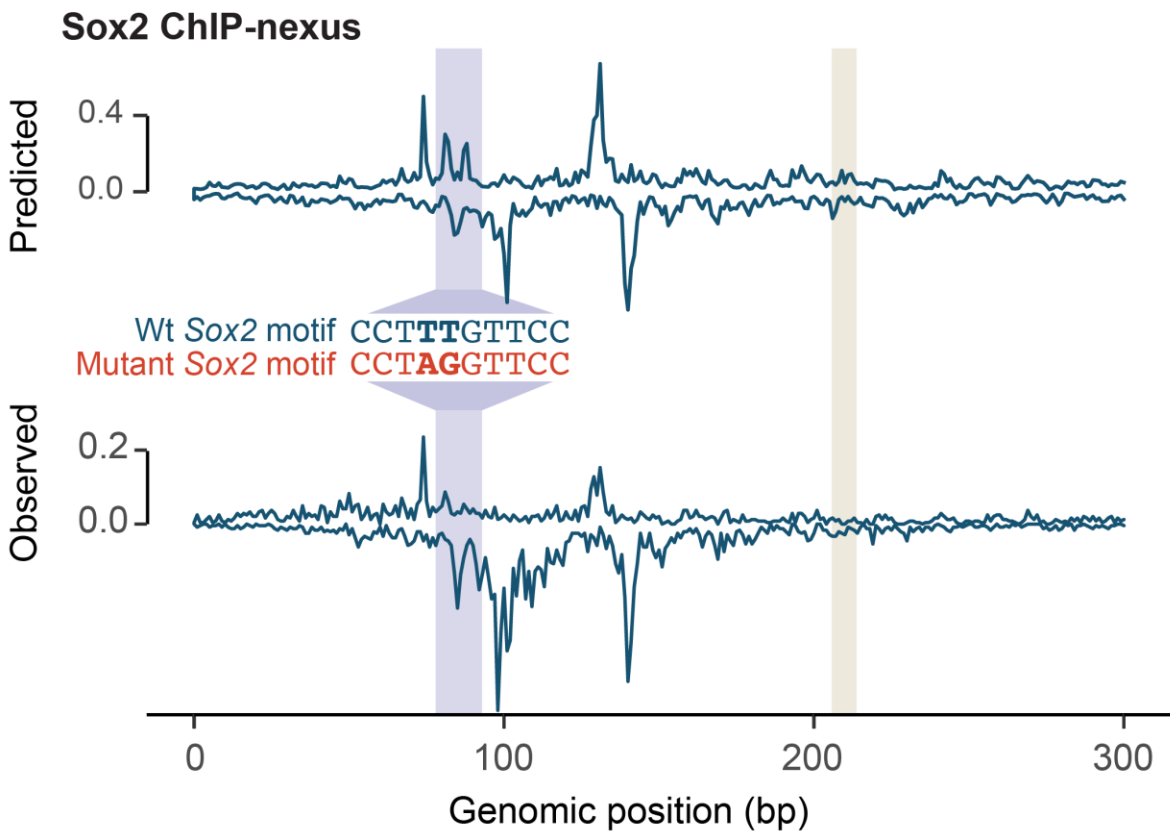


Designing CRISPR experiments to validate motif syntax



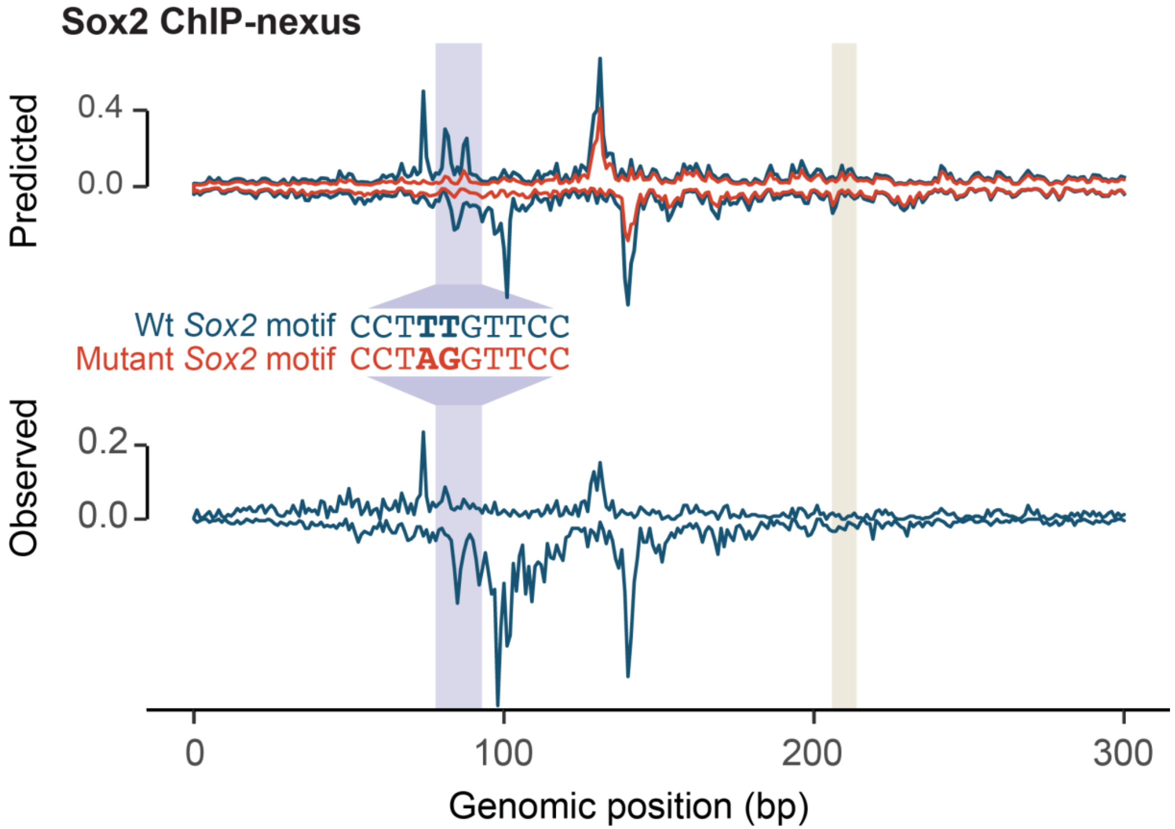
Julia Zeitlinger, Sabrina Krueger, Melanie Weilert

Designing CRISPR experiments to validate motif syntax



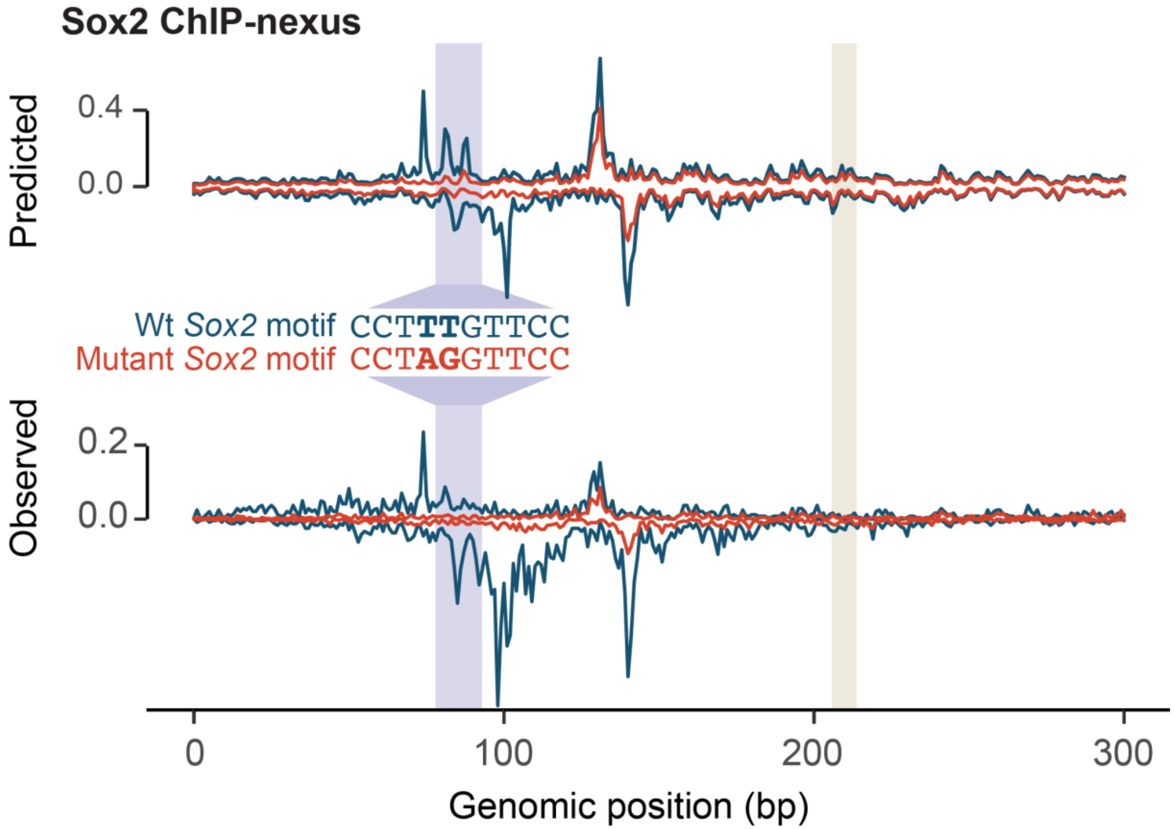
Julia Zeitlinger, Sabrina Krueger, Melanie Weilert

Designing CRISPR experiments to validate motif syntax



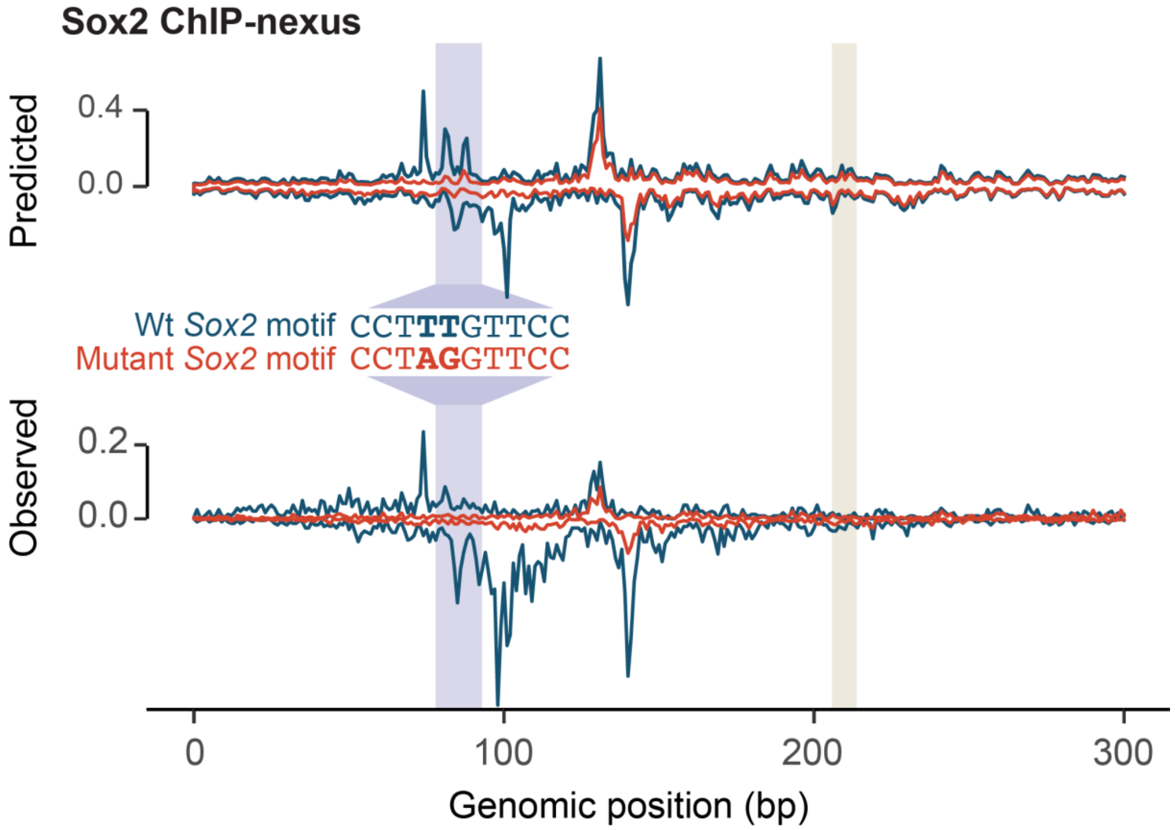
Julia Zeitlinger, Sabrina Krueger, Melanie Weilert

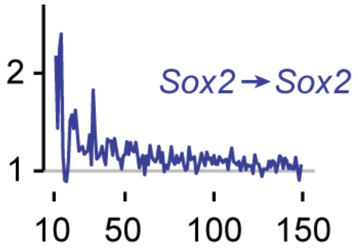
Designing CRISPR experiments to validate motif syntax



Julia Zeitlinger, Sabrina Krueger, Melanie Weilert

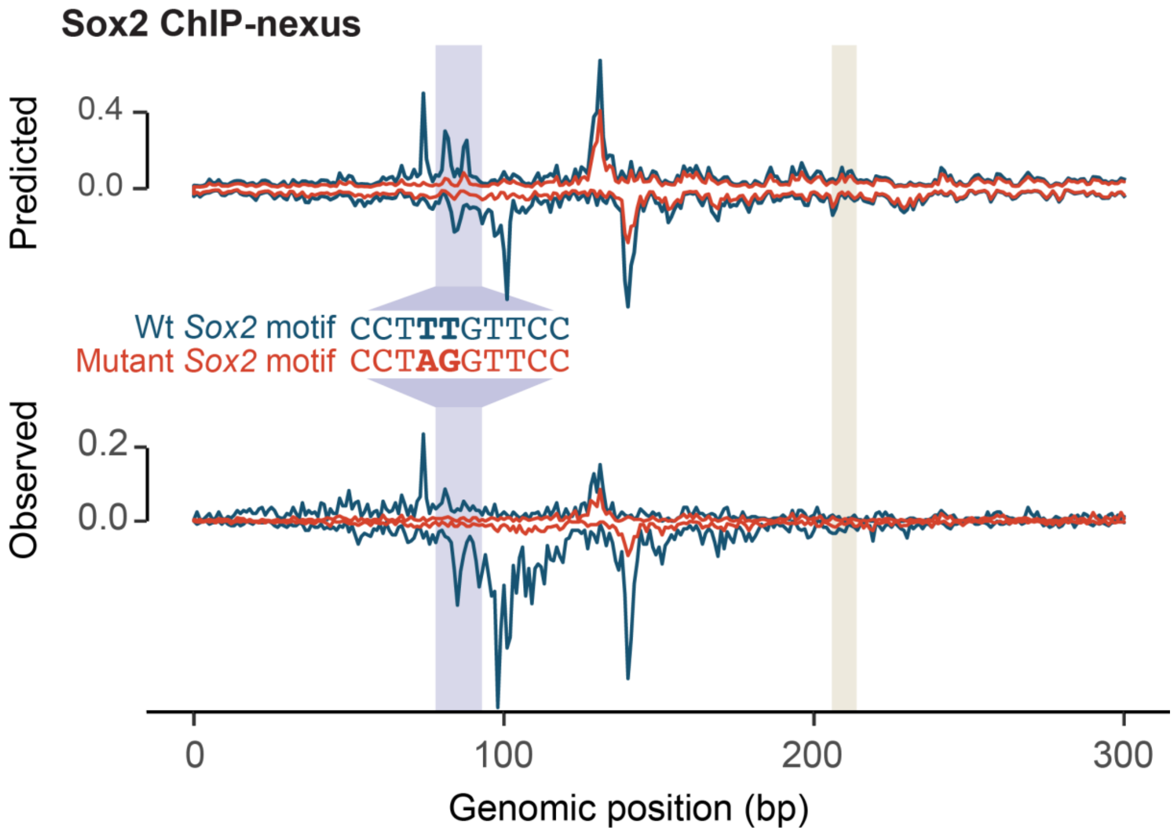
Designing CRISPR experiments to validate motif syntax

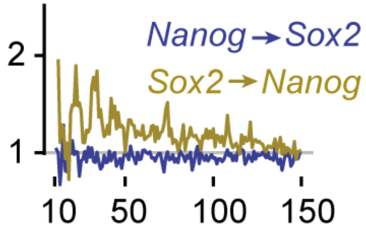


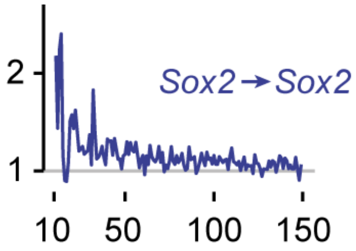


Julia Zeitlinger, Sabrina Krueger, Melanie Weilert

Designing CRISPR experiments to validate motif syntax

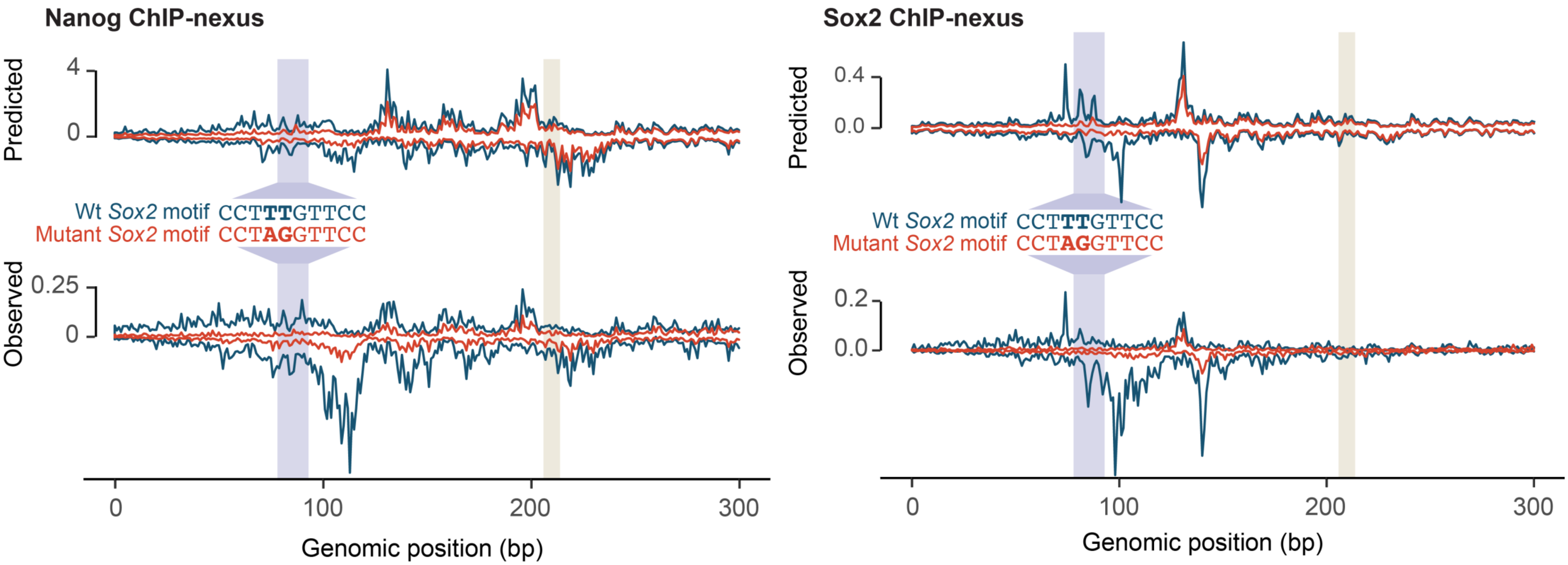


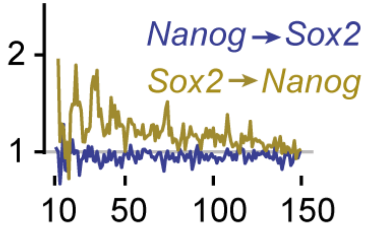


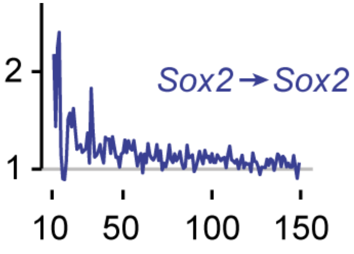


Julia Zeitlinger, Sabrina Krueger, Melanie Weilert

Designing CRISPR experiments to validate motif syntax

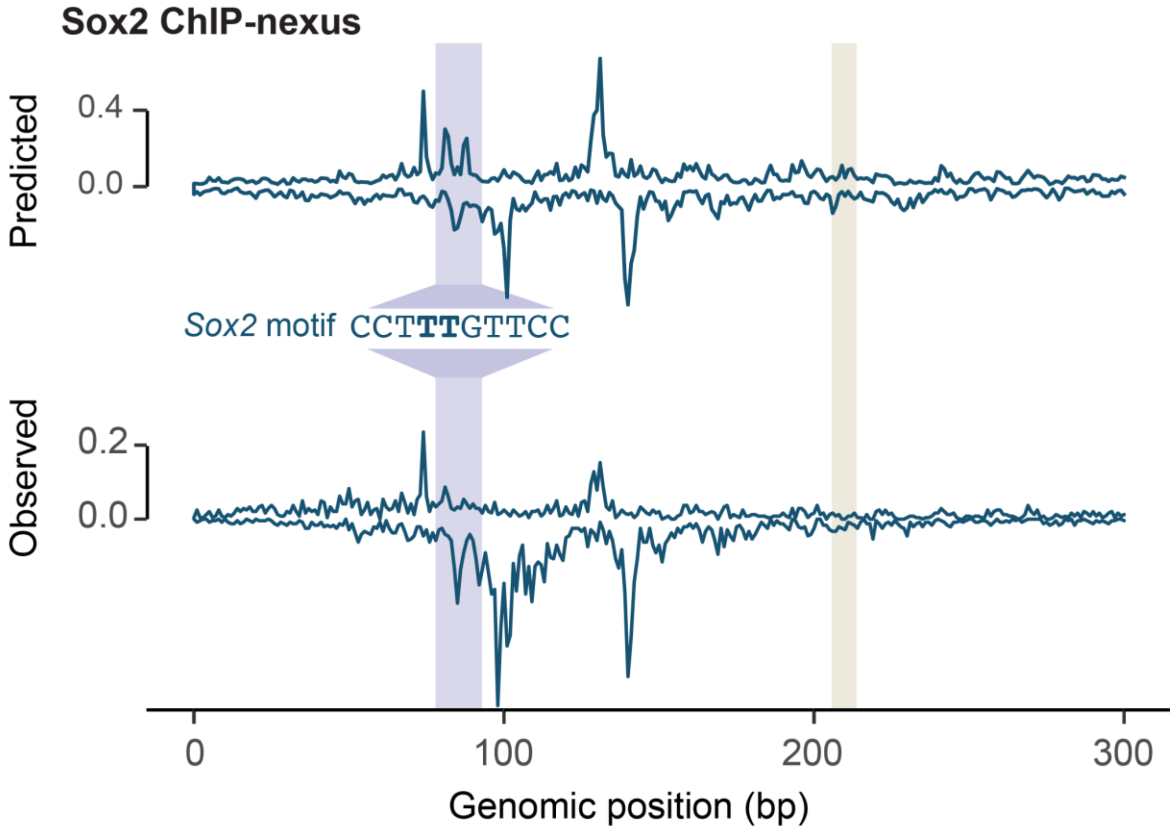






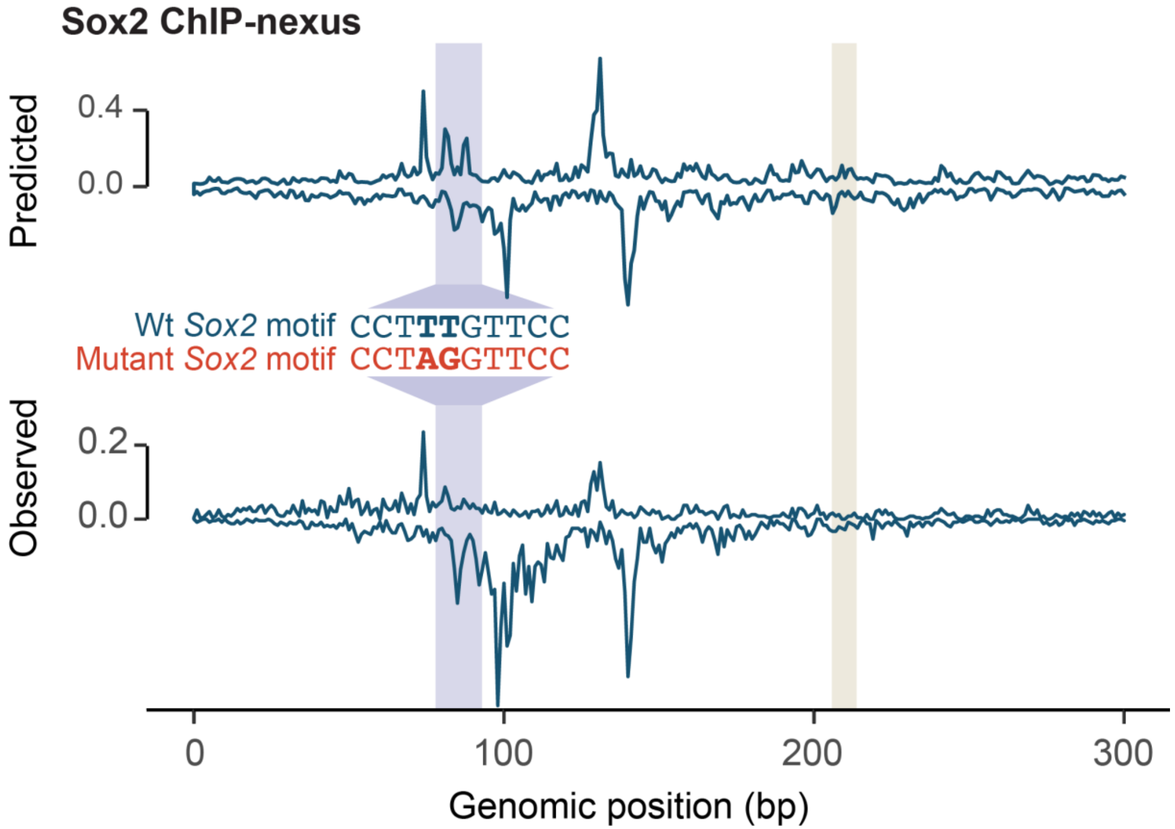
Julia Zeitlinger, Sabrina Krueger, Melanie Weilert

Designing CRISPR experiments to validate motif syntax



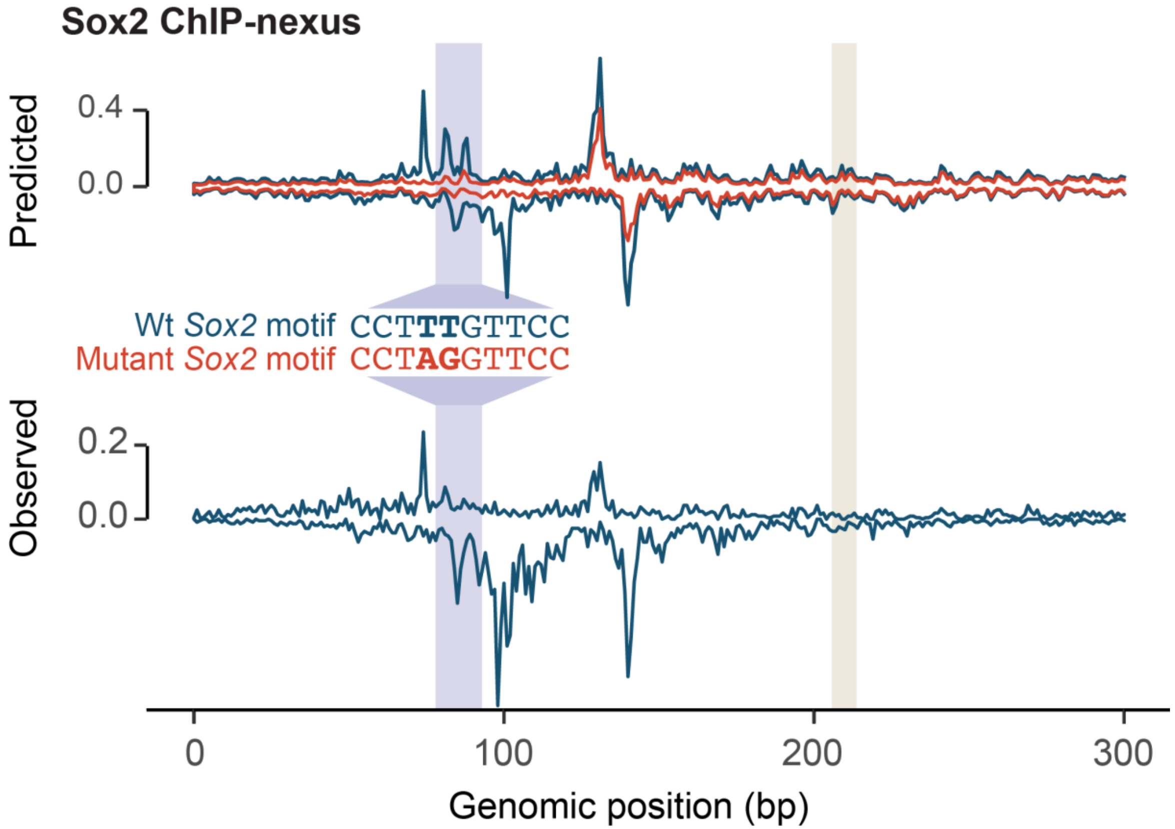
Julia Zeitlinger, Sabrina Krueger, Melanie Weilert

Designing CRISPR experiments to validate motif syntax



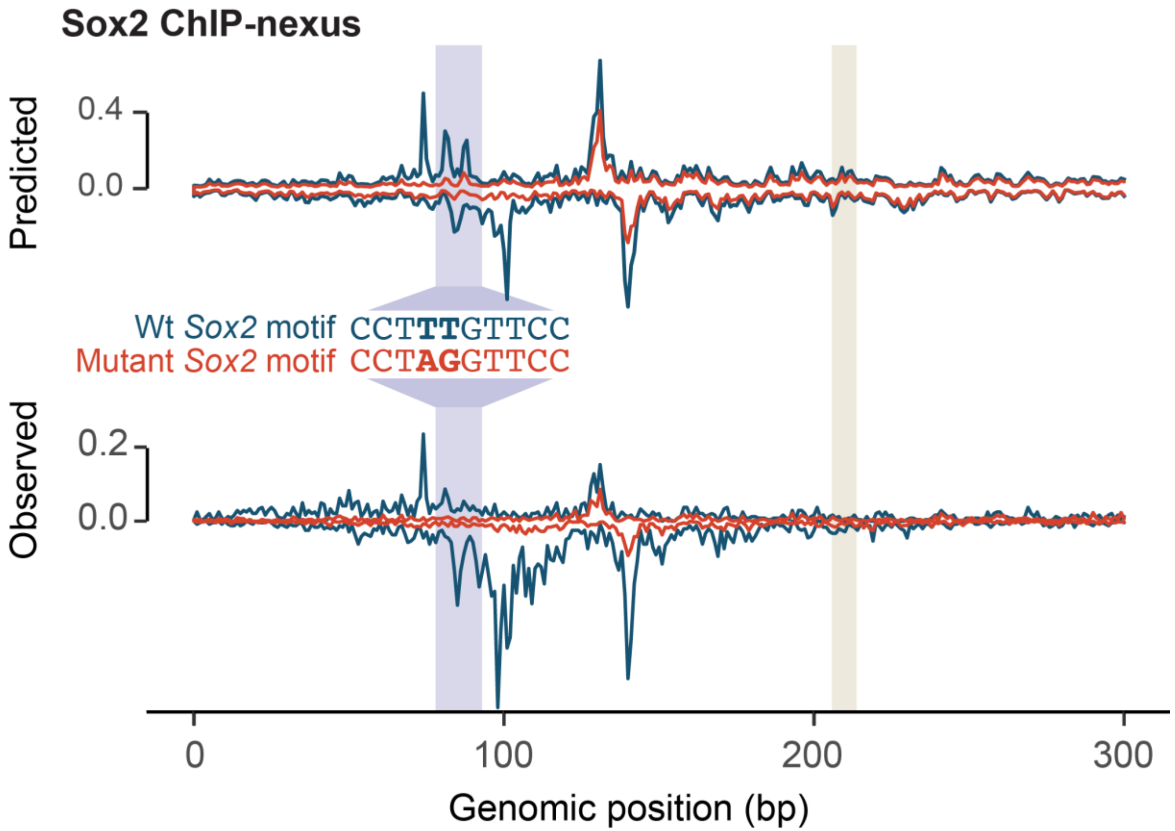
Julia Zeitlinger, Sabrina Krueger, Melanie Weilert

Designing CRISPR experiments to validate motif syntax



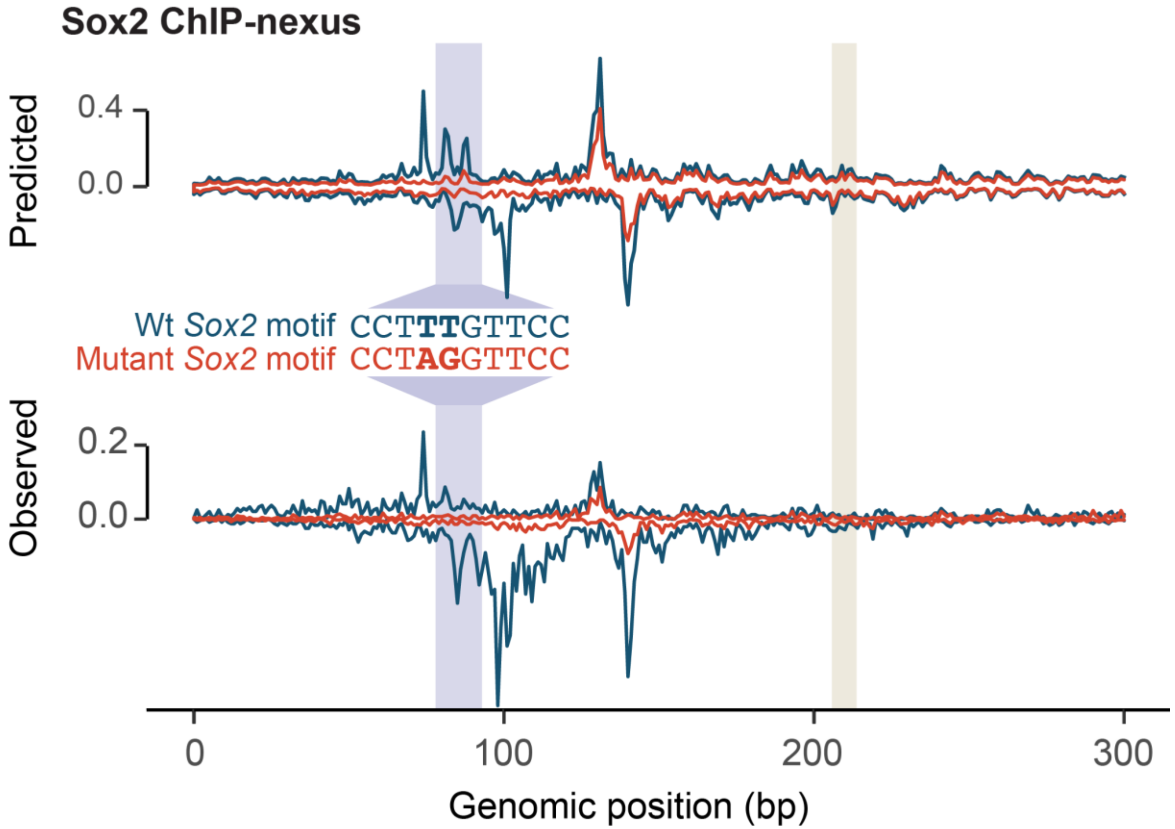
Julia Zeitlinger, Sabrina Krueger, Melanie Weilert

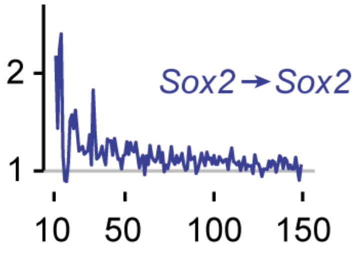
Designing CRISPR experiments to validate motif syntax



Julia Zeitlinger, Sabrina Krueger, Melanie Weilert

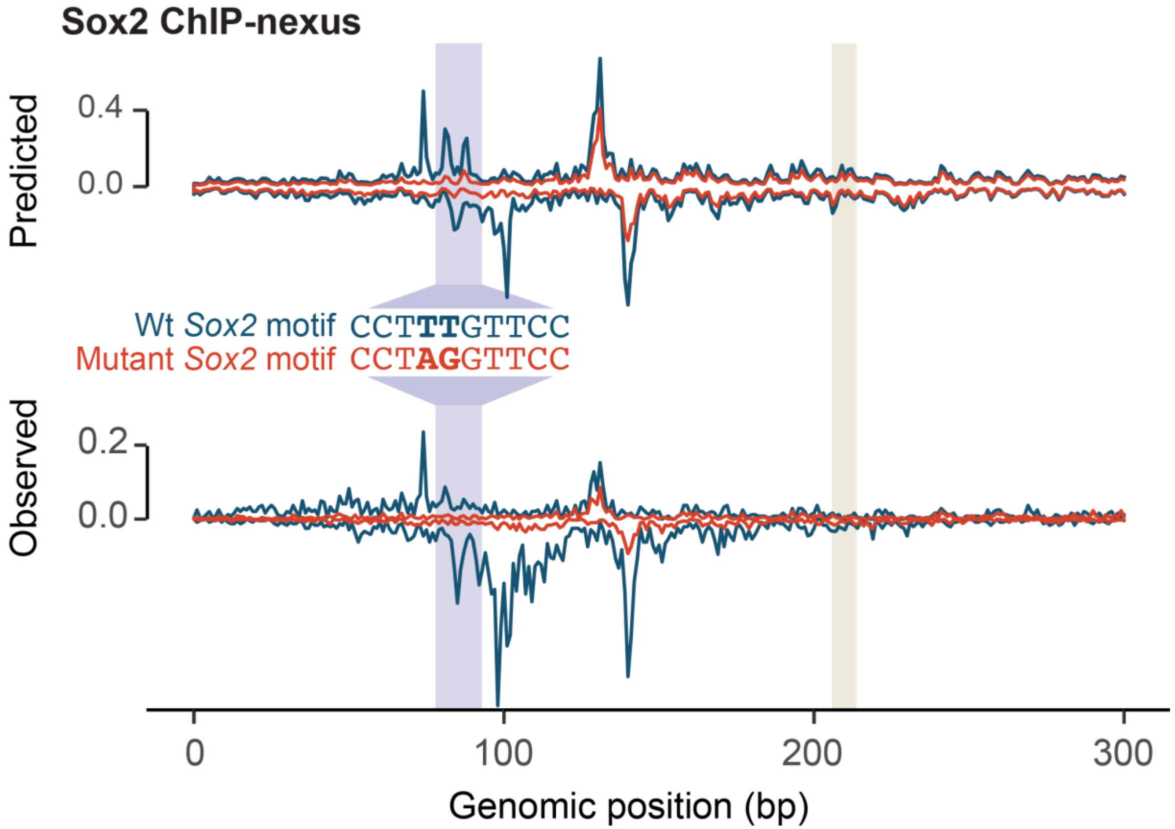
Designing CRISPR experiments to validate motif syntax

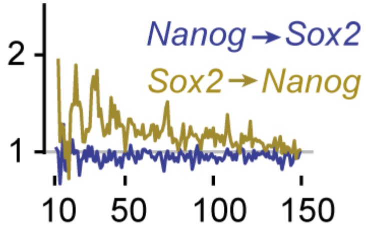


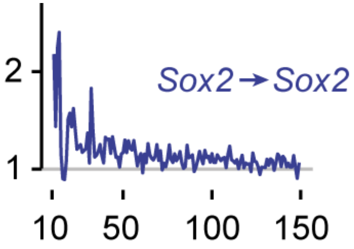


Julia Zeitlinger, Sabrina Krueger, Melanie Weilert

Designing CRISPR experiments to validate motif syntax

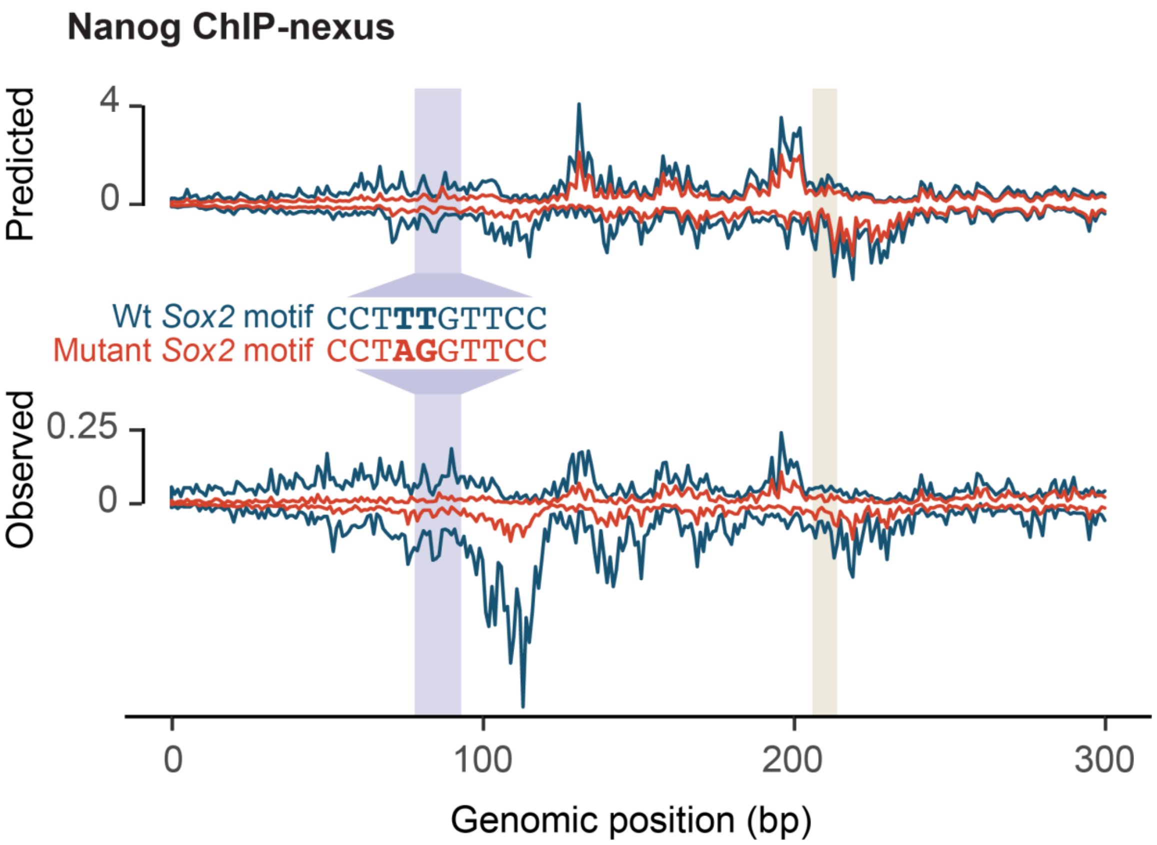


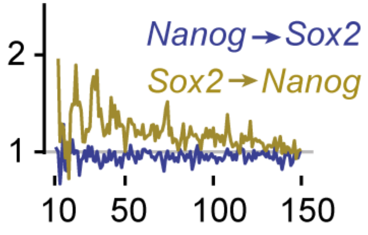


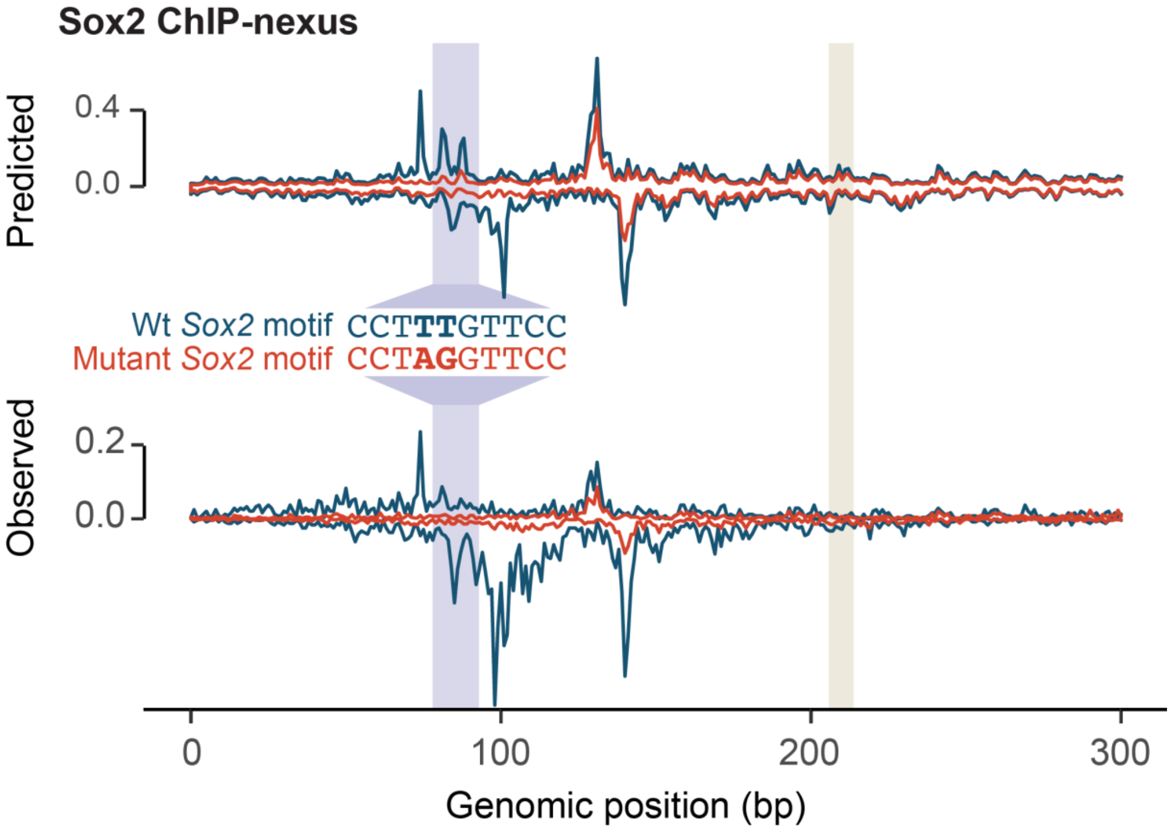


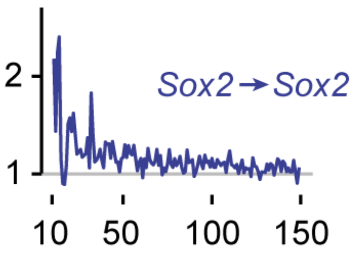
Julia Zeitlinger, Sabrina Krueger, Melanie Weilert

Designing CRISPR experiments to validate motif syntax





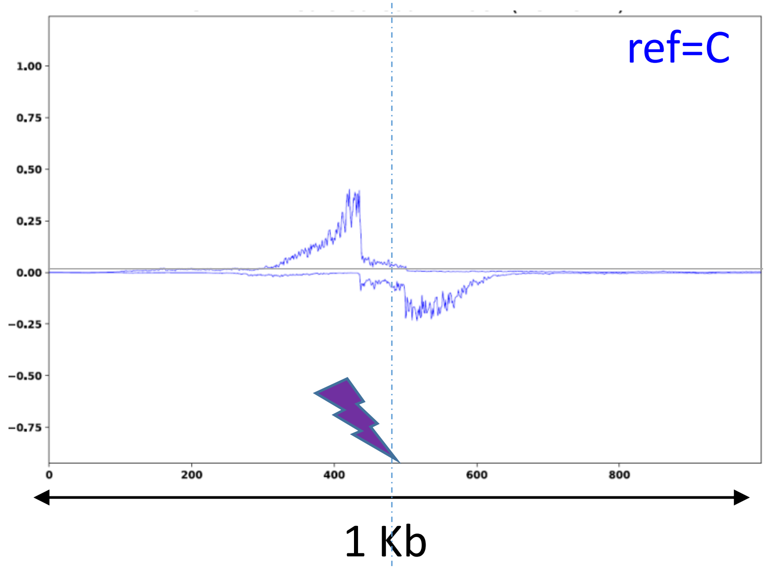




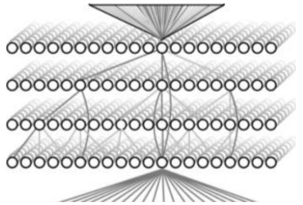
Julia Zeitlinger, Sabrina Krueger, Melanie Weilert

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

Predicted molecular profile of protein-DNA binding



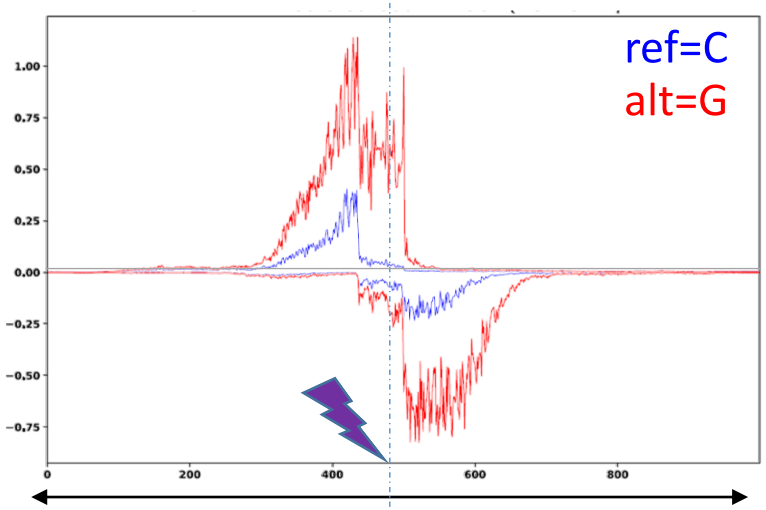
PredictedSignal



……ACTGAT**C**GCAATCG…….

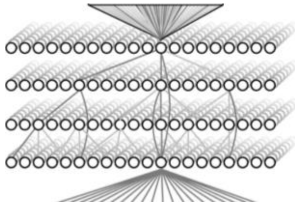
In-silico mutagenesis: Predict effect of genetic variant on molecular activity

Predicted molecular profile of protein-DNA binding



1 Kb

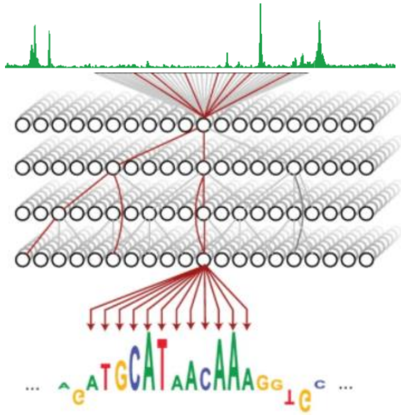
ΔPredictedSignal



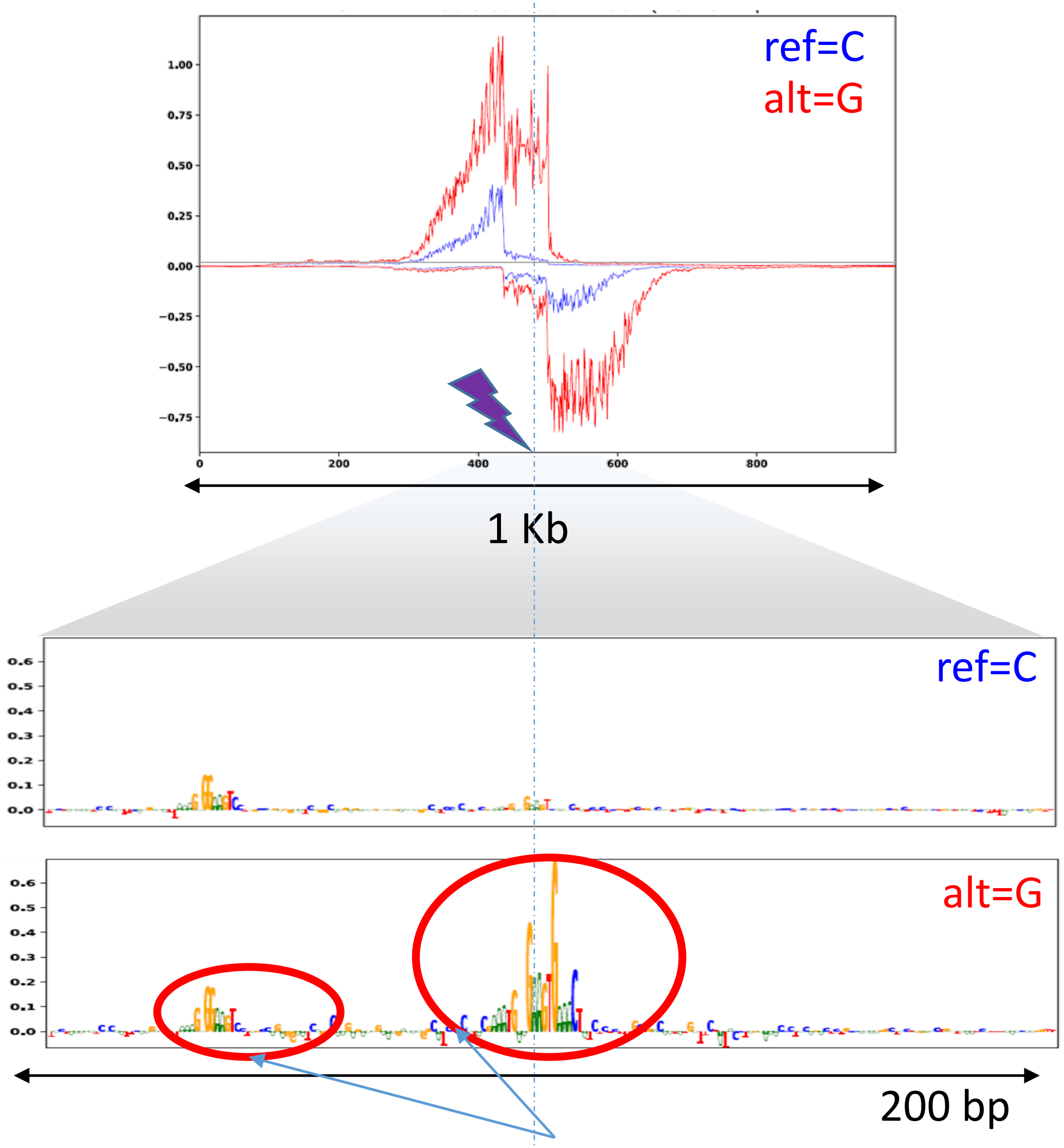
……ACTGAT**C**GCAATCG…….

……ACTGAT**G**GCAATCG…….

Interpret disrupted predictive sequence syntax



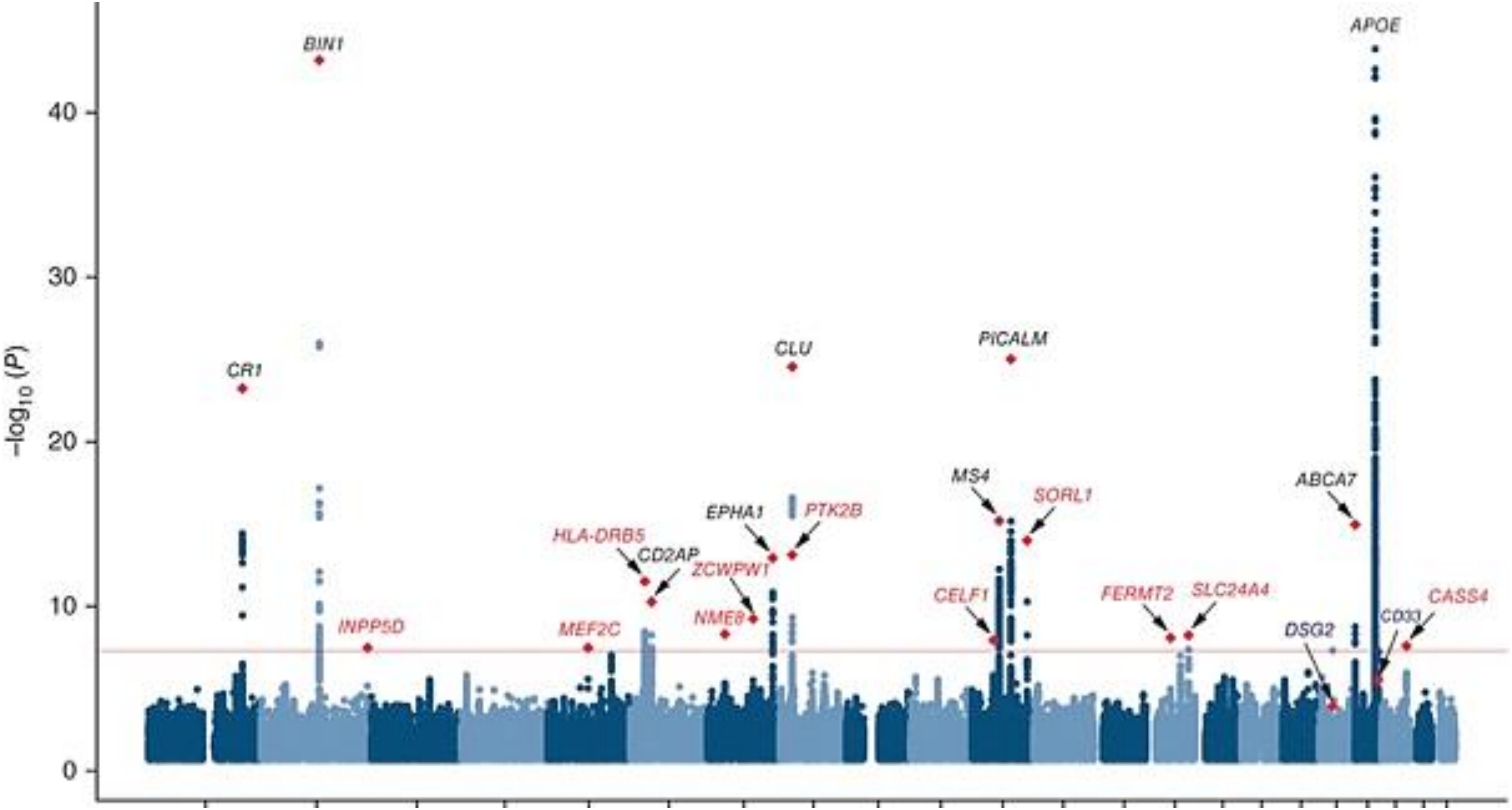
Predicted molecular profile of protein-DNA binding



Sequence binding motifs of SPI1 DNA binding protein

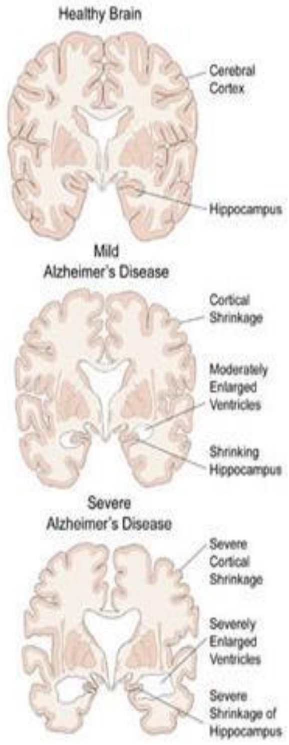
Genetic loci associated with Alzheimer's disease

Statistical significance of association



(Lambert et al., Nat. Genet., 2013)

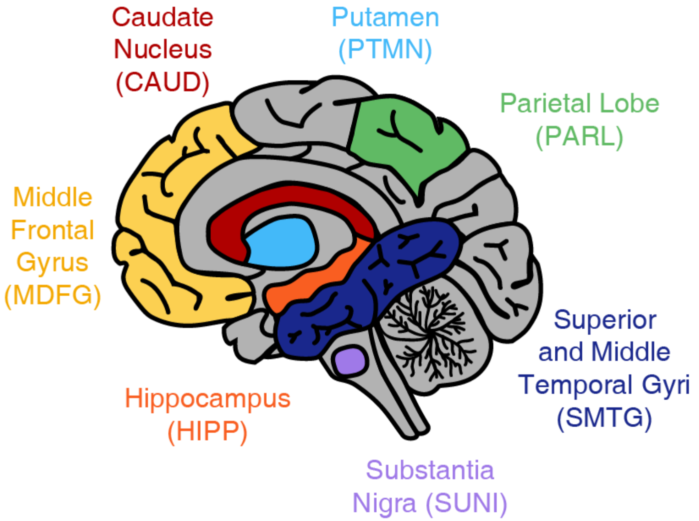
Genomic position ⟶



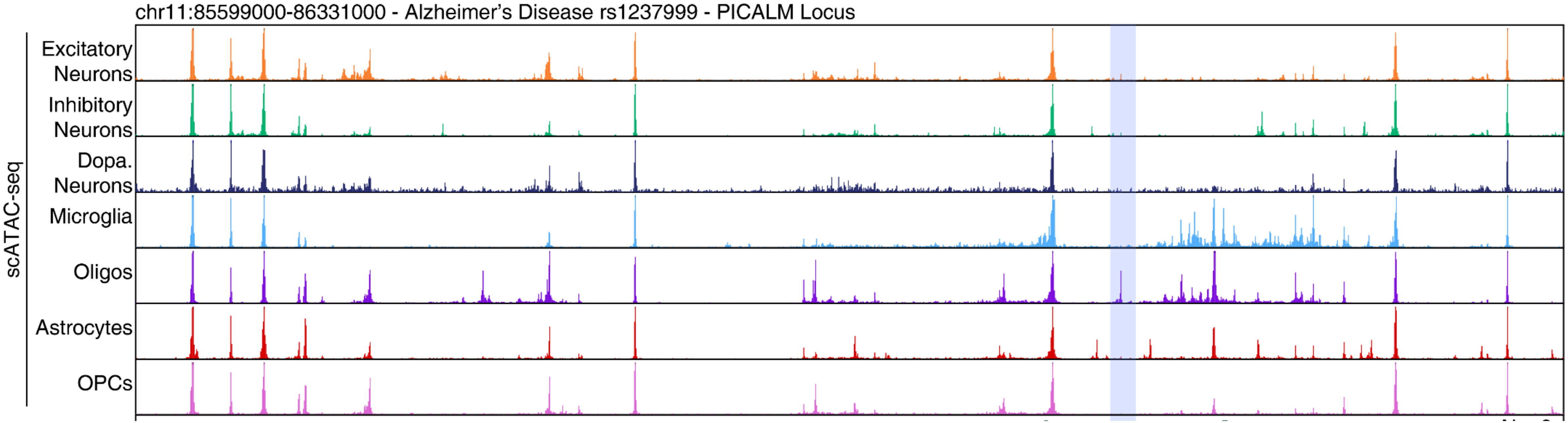
Healthy Brain
Cerebral Cortex
Hippocampus
Mild Alzheimer's Disease
Cortical Shrinkage
Moderately Enlarged Ventricles
Shrinking Hippocampus
Severe Alzheimer's Disease
Severe Cortical Shrinkage
Severely Enlarged Ventricles
Severe Shrinkage of Hippocampus

Illustration by Bob Morreale, American Health Assistance Foundation

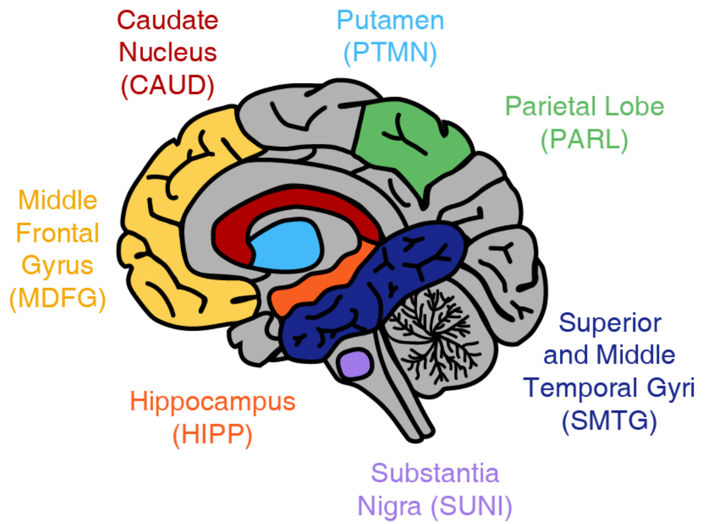
Molecular profiling of cell types in the brain



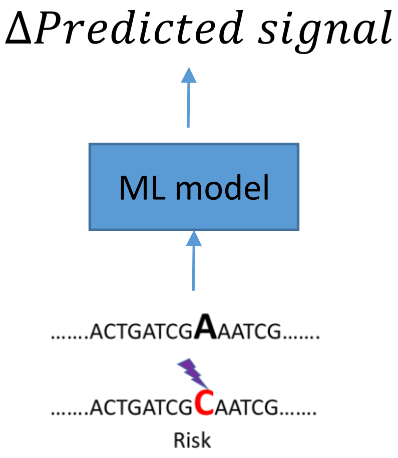
Corces et al. 2020, Nature Genetics


chr11:85599000-86331000 - Alzheimer's Disease rs1237999 - PICALM Locus

Molecular profiling of cell types in the brain



Corces et al. 2020, Nature Genetics



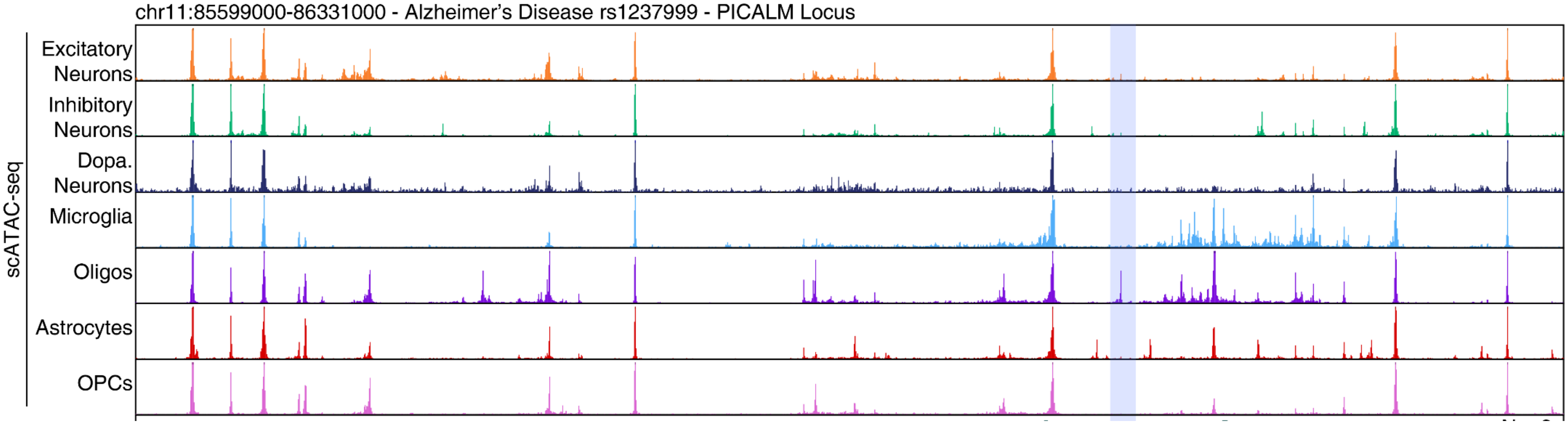
$\Delta Predicted\ signal$

ML model

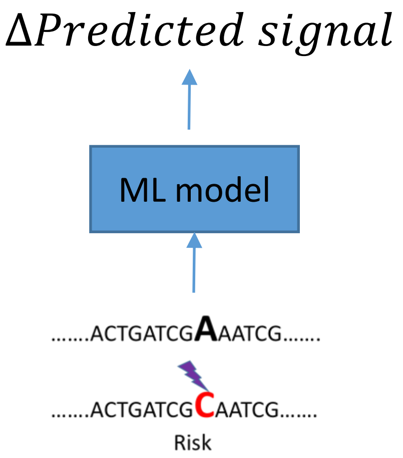
.......ACTGATCG**A**AATCG.......

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

Risk



Predicting and interpreting causal AD variants



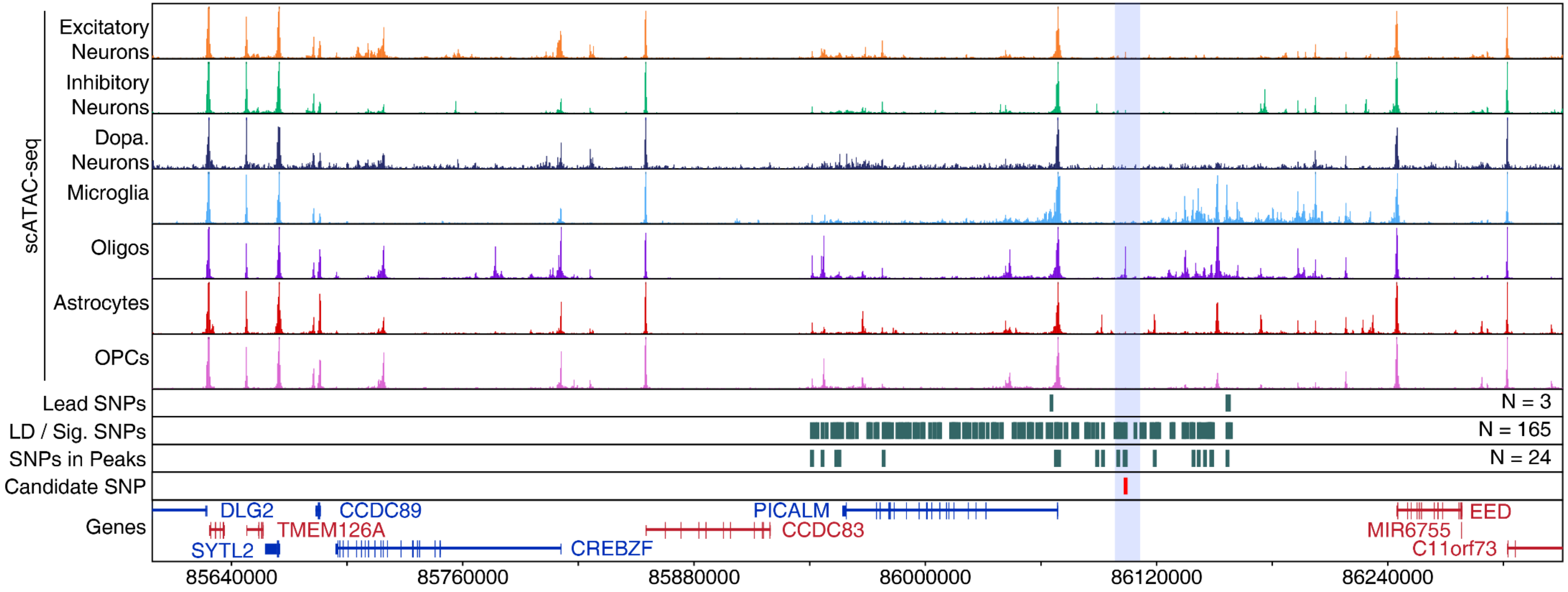
$\Delta Predicted\ signal$

ML model

.......ACTGATCG**A**AATCG.......

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

Risk

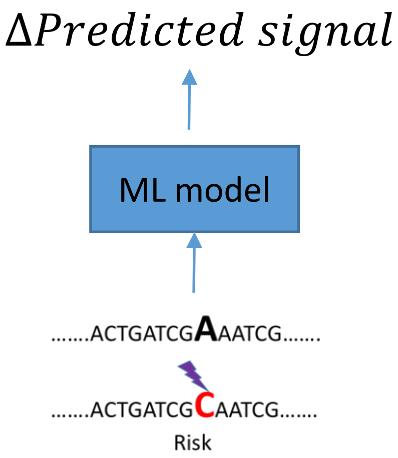


chr11:85599000-86331000 - Alzheimer’s Disease rs1237999 - PICALM Locus

Anna Shcherbina

Soumya Kundu

Predicting and interpreting causal AD variants



$\Delta Predicted\ signal$

ML model

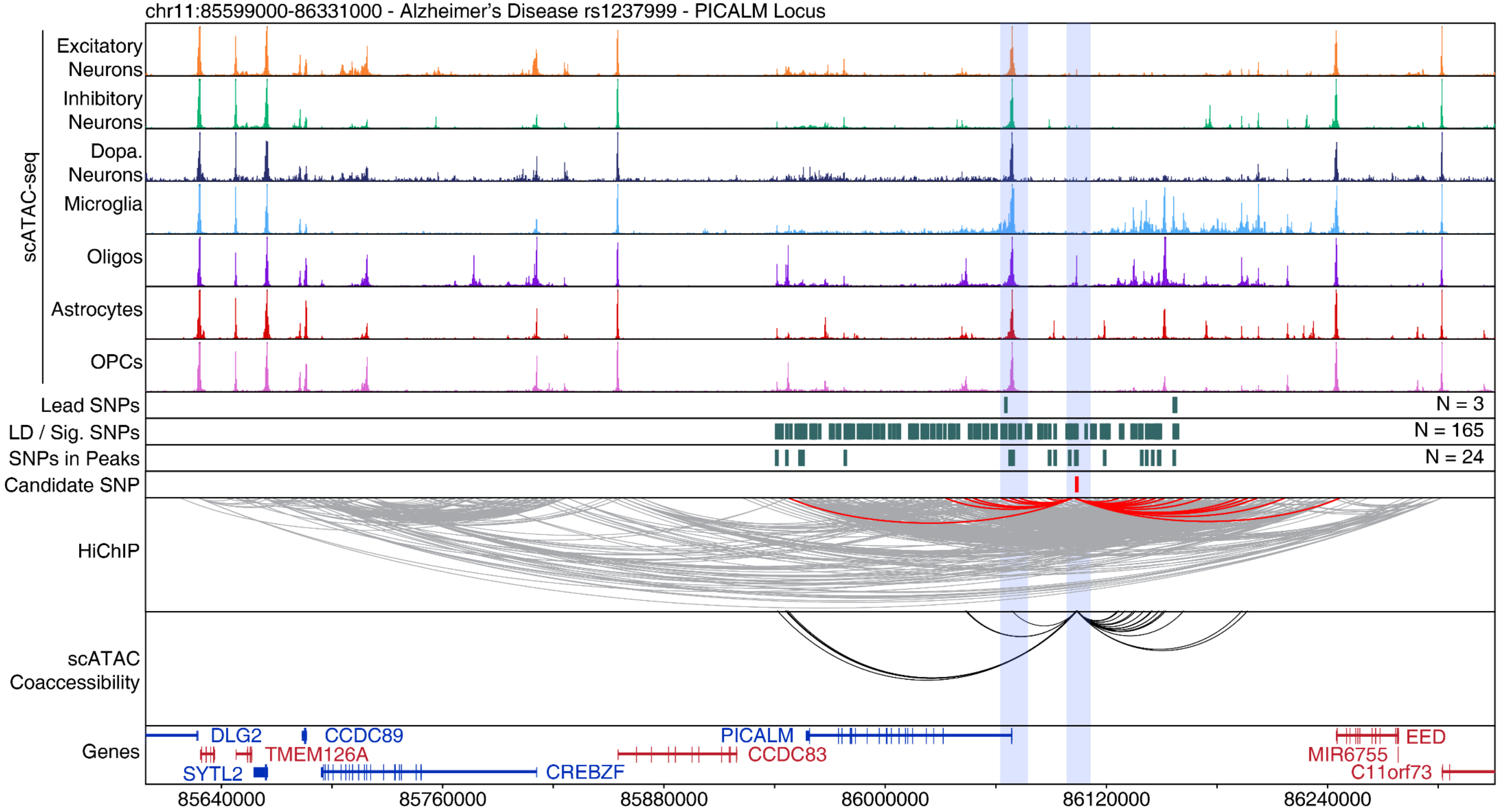
.......ACTGATCG**A**AATCG.......

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

Risk

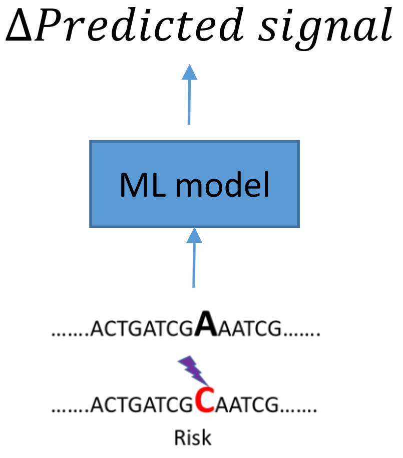
Anna Shcherbina

Soumya Kundu



chr11:85599000-86331000 - Alzheimer's Disease rs1237999 - PICALM Locus

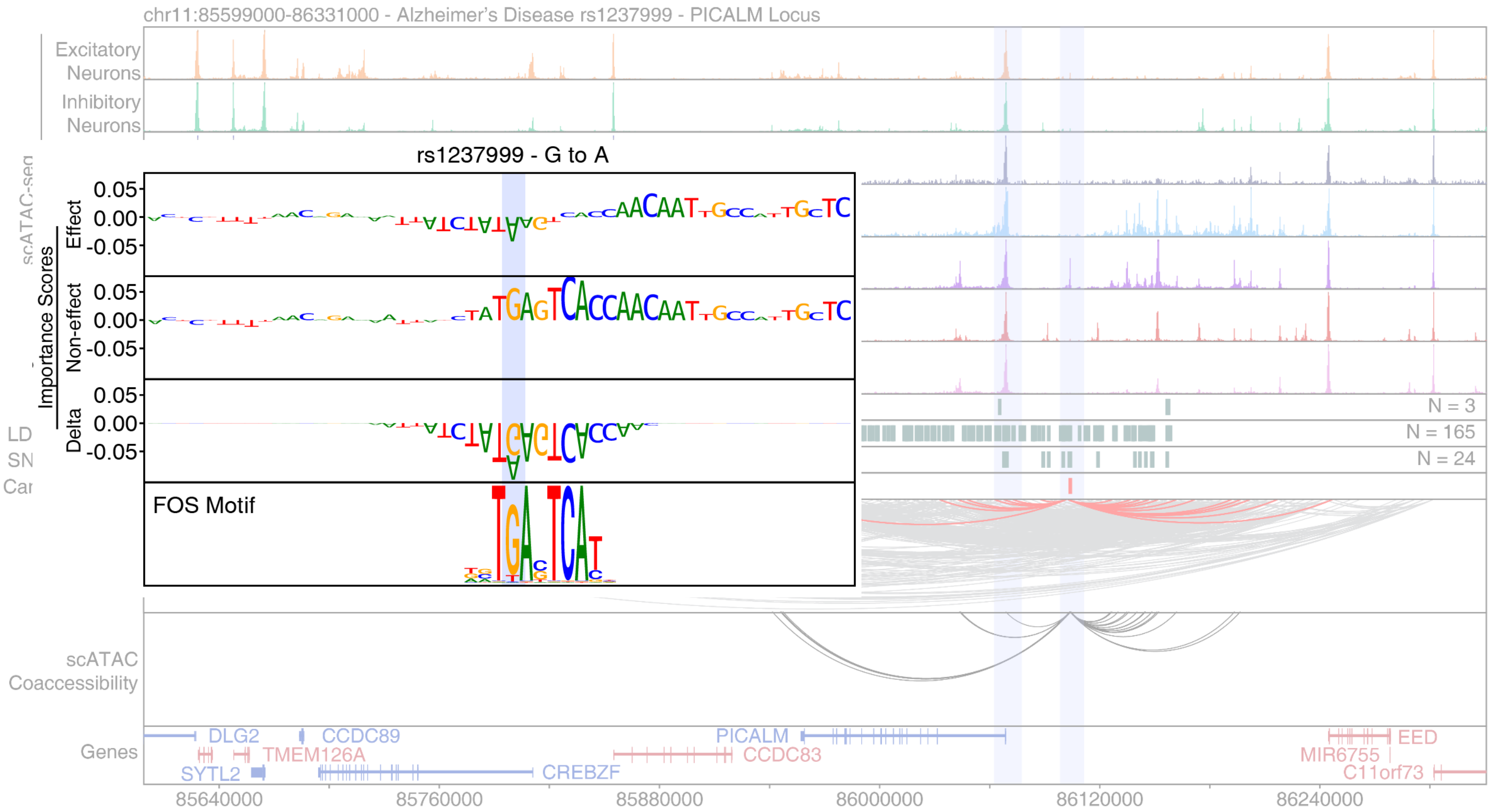
Predicting and interpreting causal AD variants



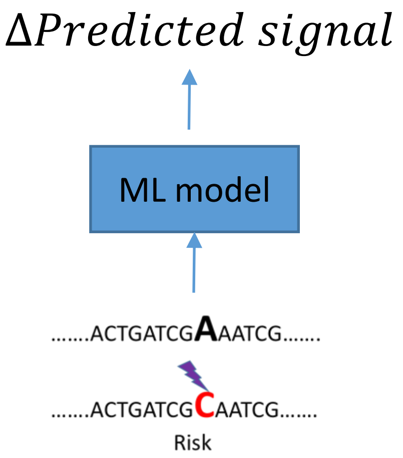
$\Delta Predicted\ signal$
ML model
.......ACTGATCGAAATCG.......
.......ACTGATCGCAATCG.......
Risk

Anna Shcherbina

Soumya Kundu



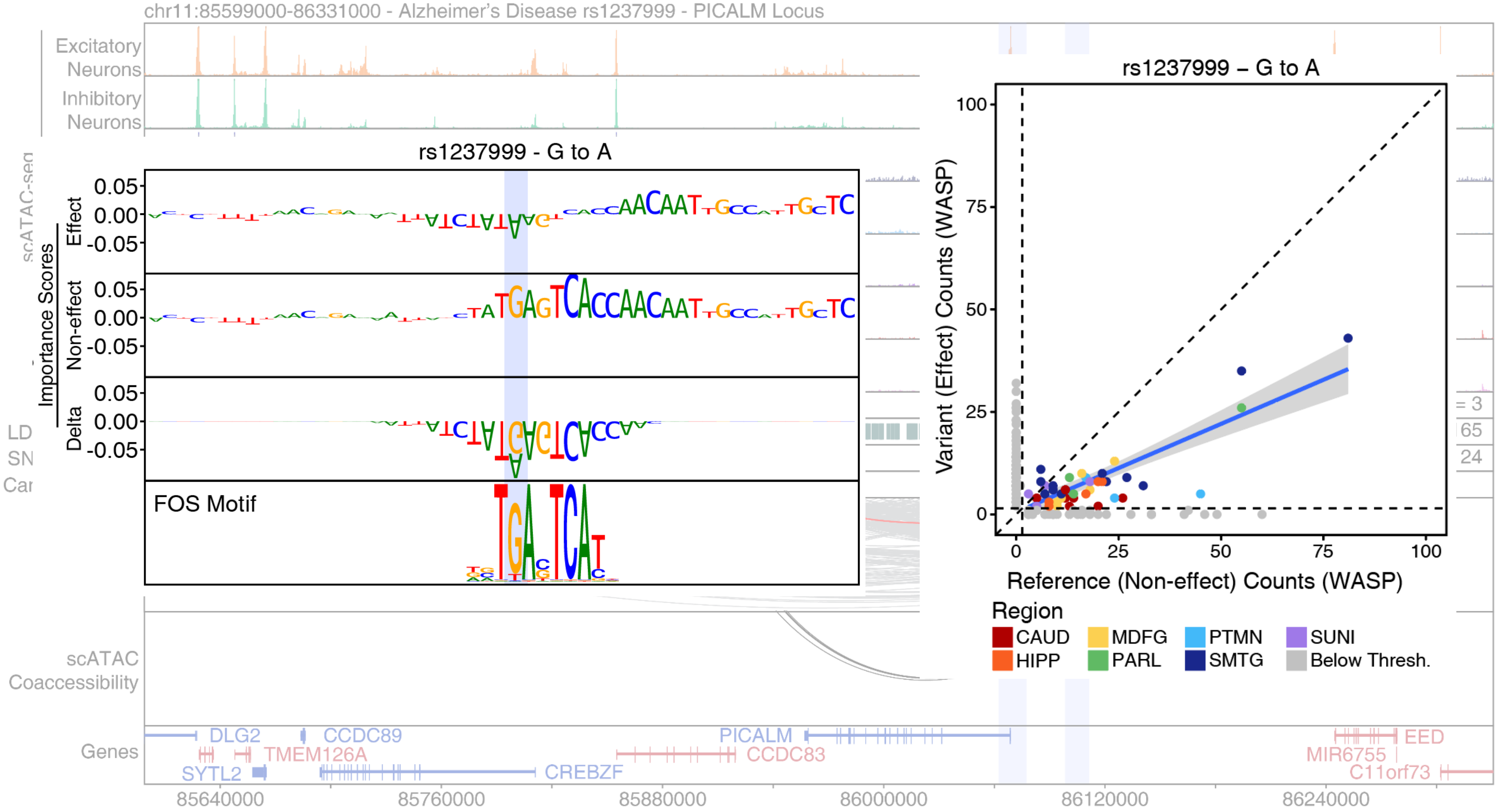
Predicting and interpreting causal AD variants



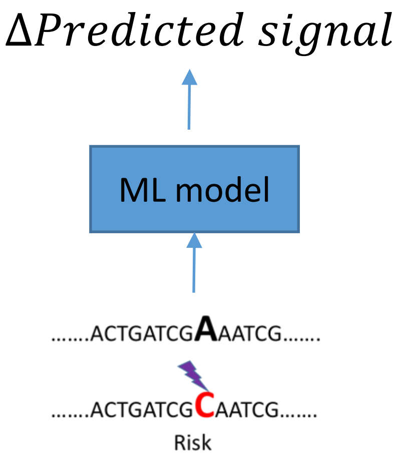
$\Delta Predicted\ signal$
ML model
.......ACTGATCGAAATCG.......
.......ACTGATCGCAATCG.......
Risk

Anna Shcherbina

Soumya Kundu



Predicting and interpreting causal AD variants



ΔPredicted signal

ML model

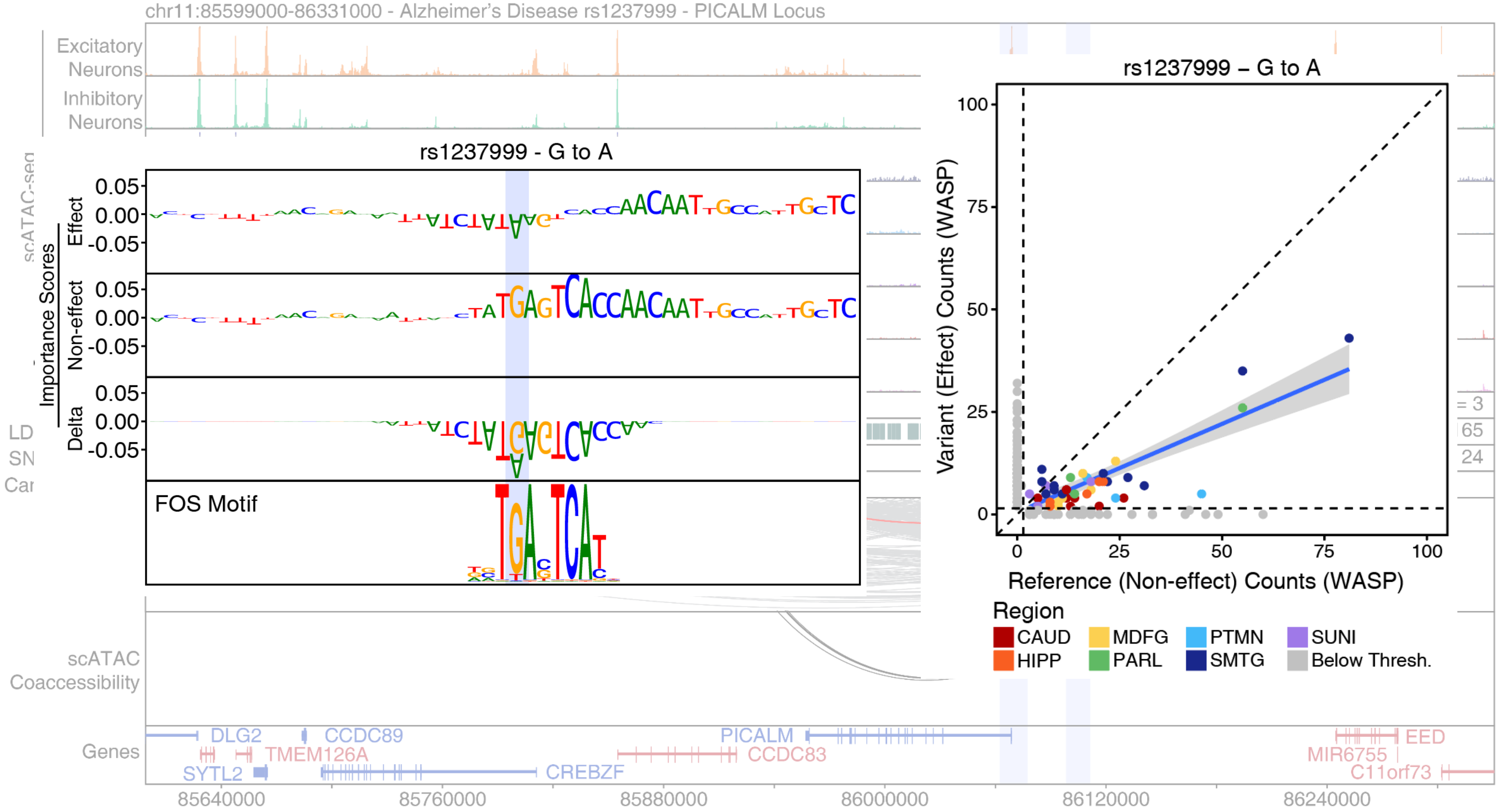
.......ACTGATCGAAATCG.......

.......ACTGATCGCAATCG.......

Risk

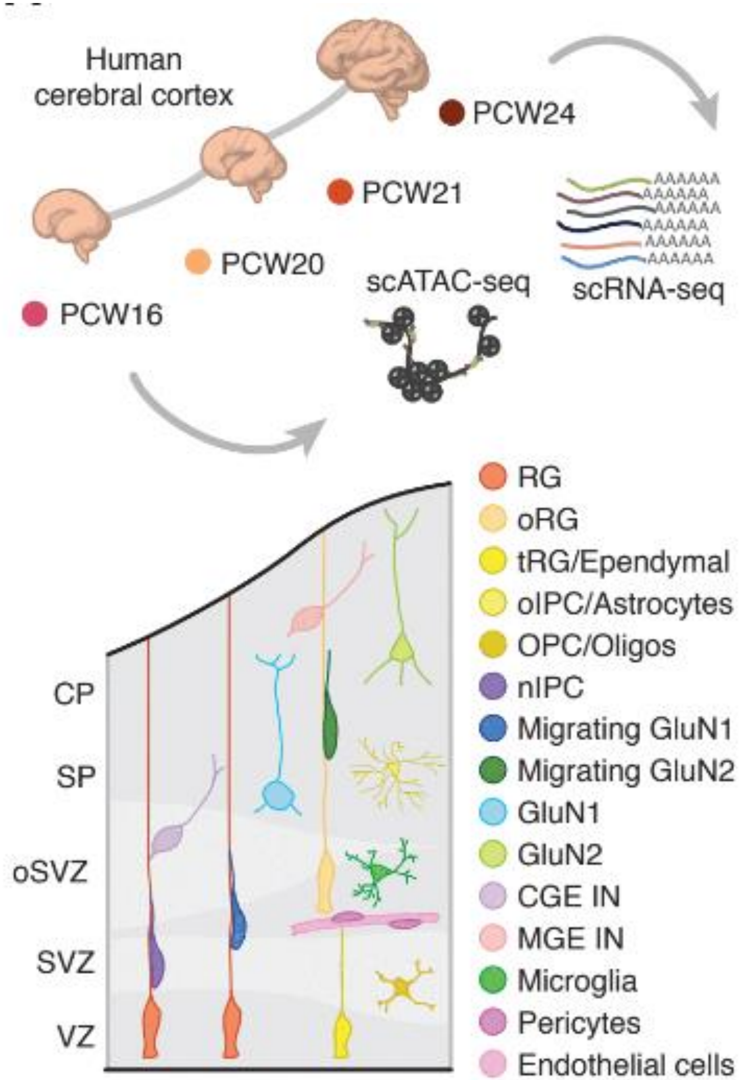
Anna Shcherbina

Soumya Kundu



Genetic variant rs1237999 disrupts a sequence motif of the FOS protein in a control element of the PICALM gene in oligodendrocyte cells in the brain

Predicting *de-novo* mutations in Autism

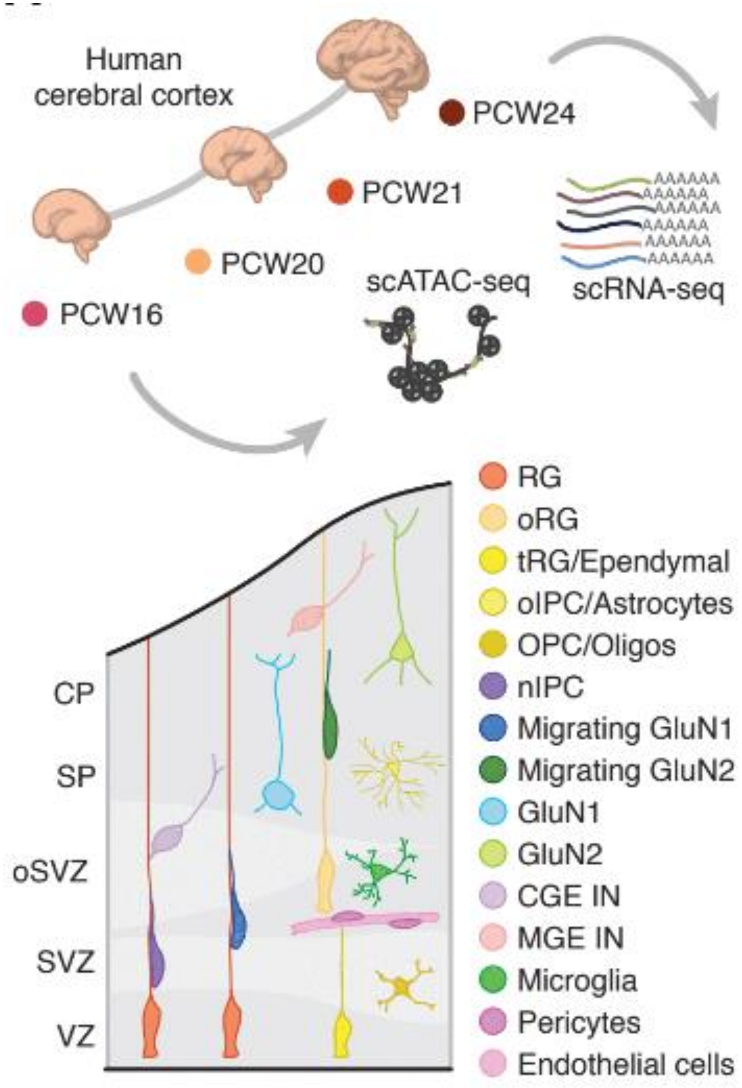


Laksshman
Sundaram

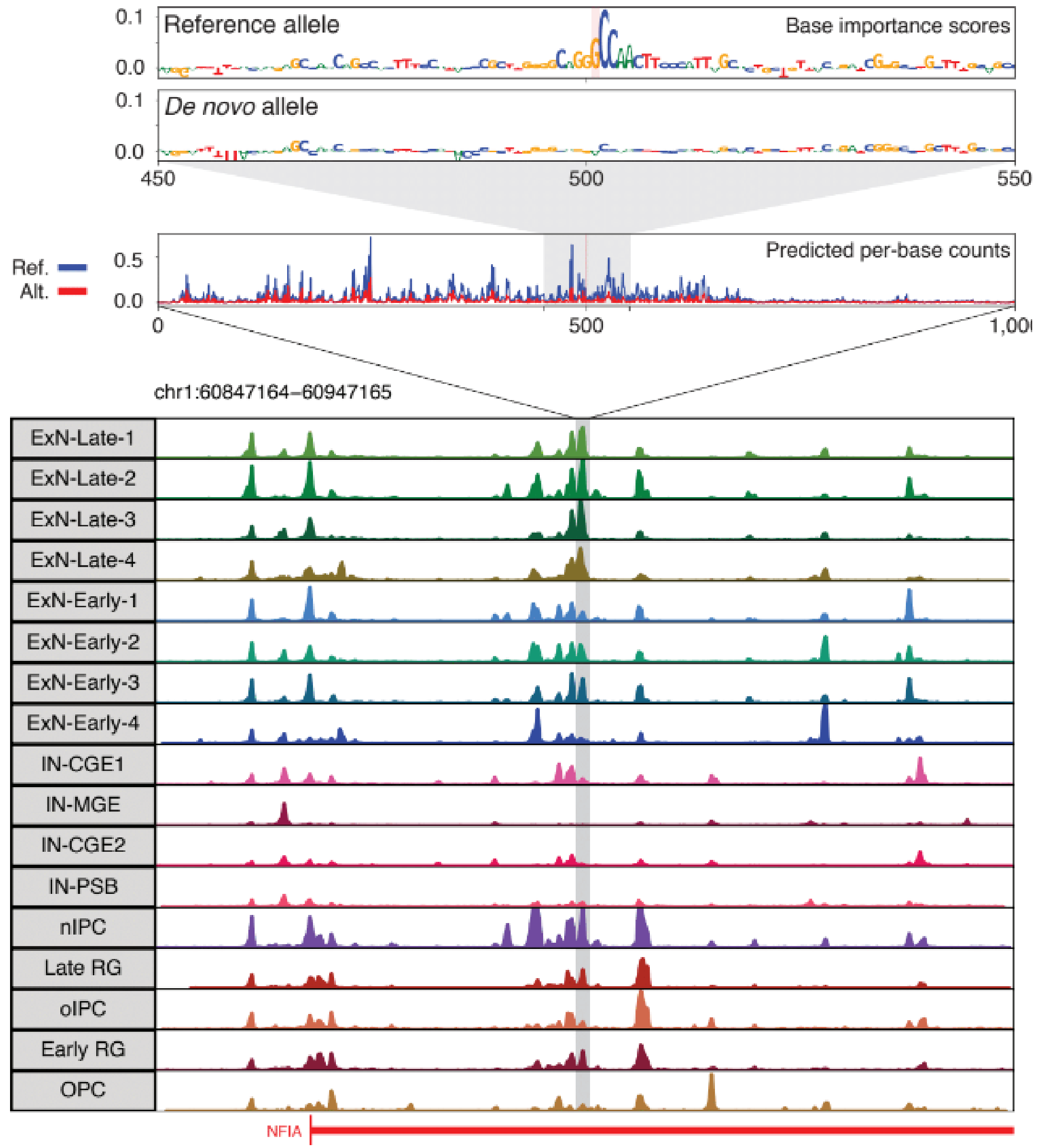
with Greenleaf, Pasca labs ***Trevino et al. 2021, Cell***

Prediction: Mutation disrupts NFIA motif in control element of NFIA gene in glutamatergic neurons

Predicting *de-novo* mutations in Autism



Laksshman Sundaram

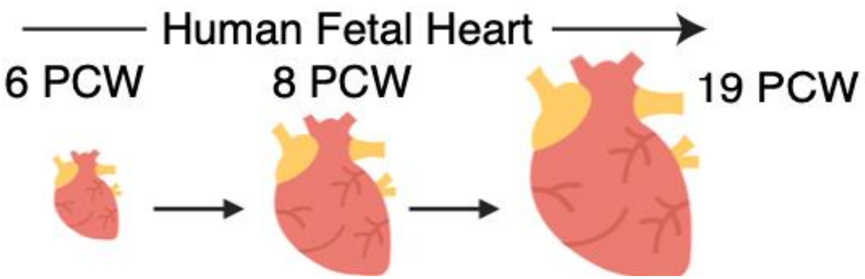


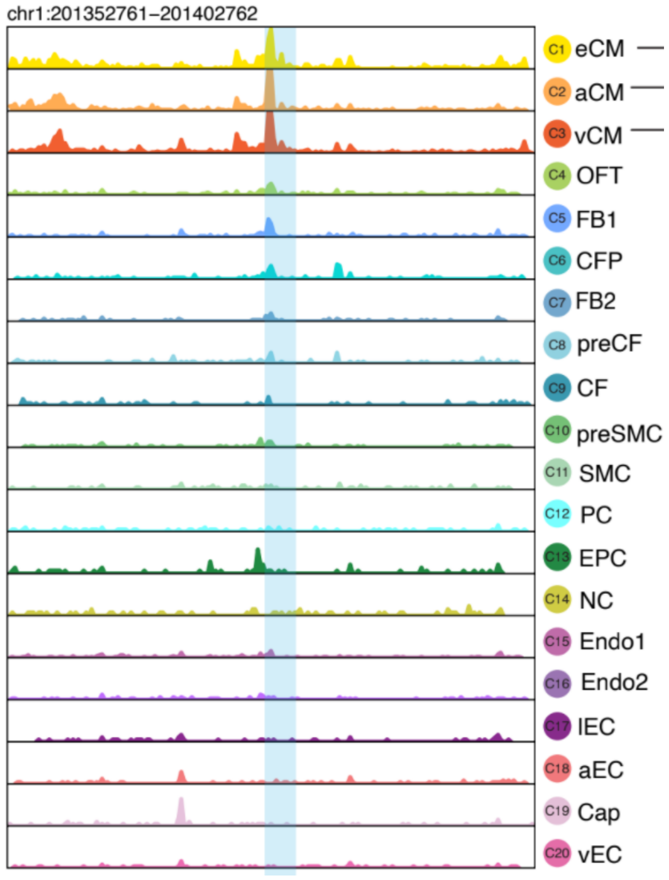
Prediction: Mutation disrupts NFIA motif in control element of NFIA gene in glutamatergic neurons

with Greenleaf, Pasca labs

Trevino et al. 2021, Cell

Predicting *de-novo* mutations in congenital heart disease



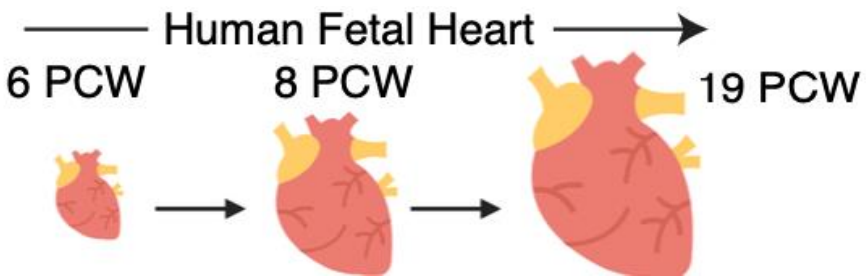


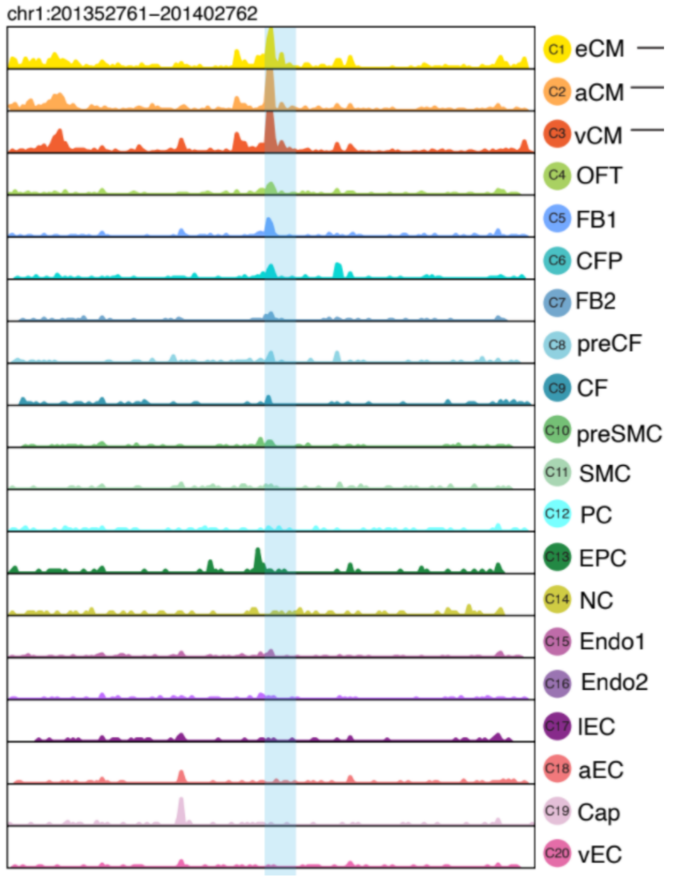
Laksshman Sundaram

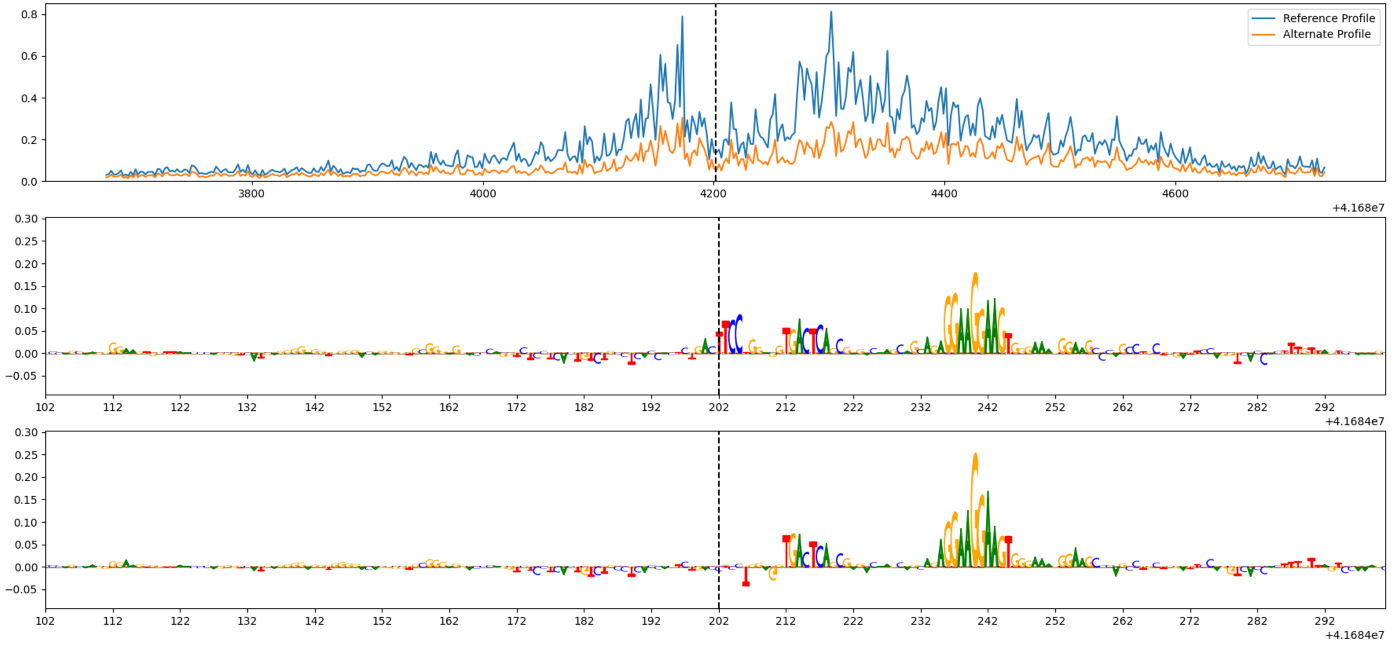
Mo Ameen

with Wang, Karakikes, Quertermous, Greenleaf labs

Predicting *de-novo* mutations in congenital heart disease







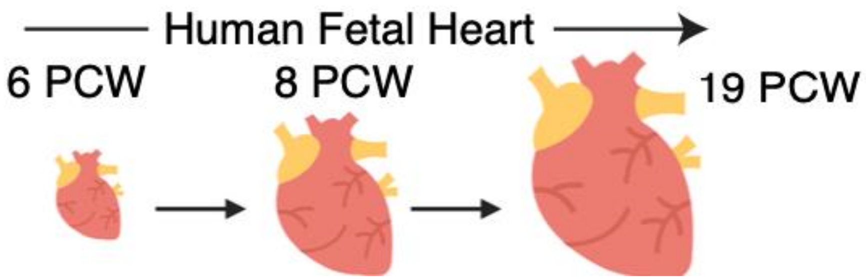
Laksshman Sundaram

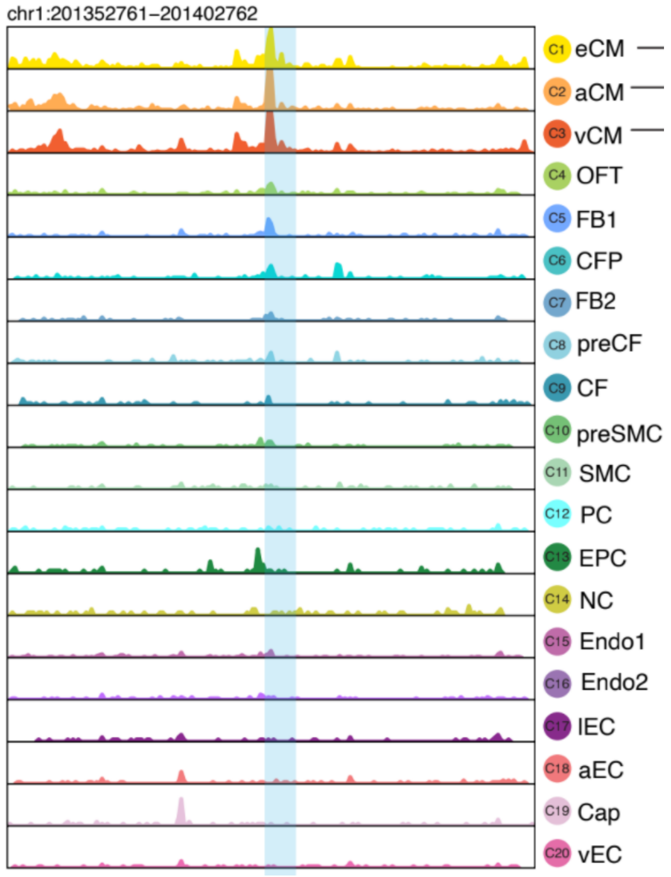
Mo Ameen

Prediction: Mutation disrupts ETV motif in control element active in arterial endothelial cells

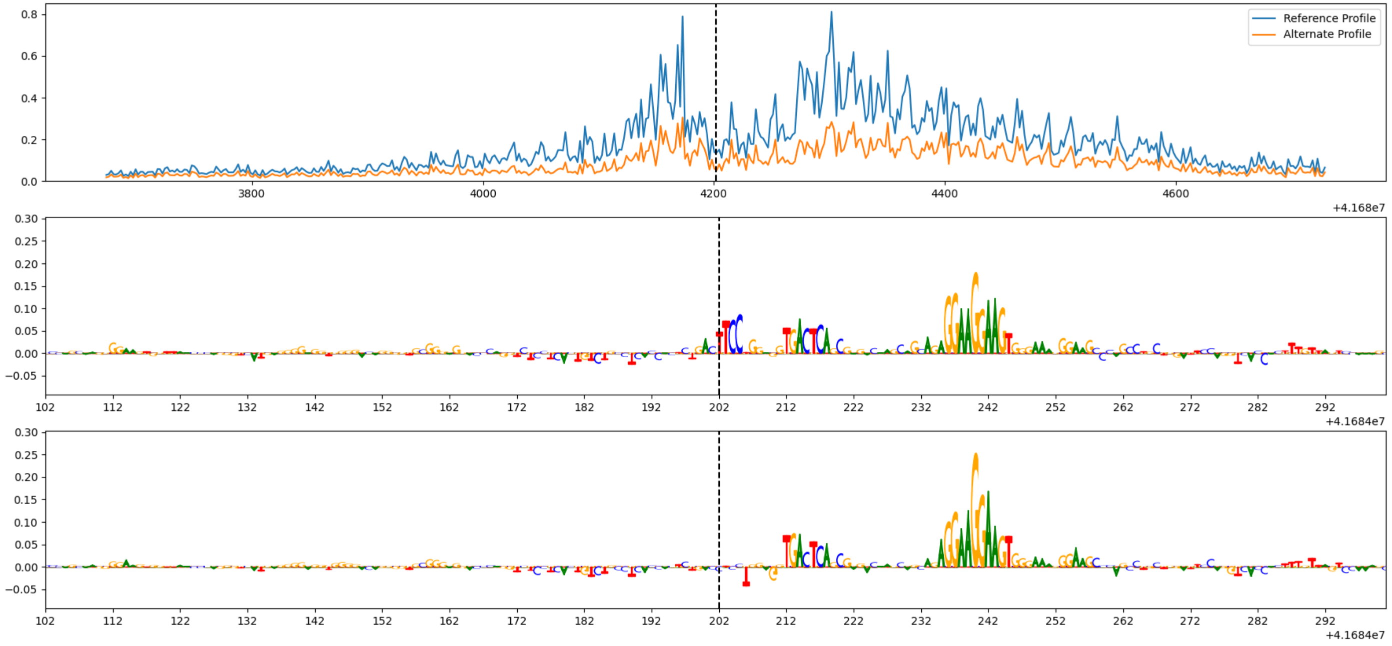
with Wang, Karakikes, Quertermous, Greenleaf labs

Predicting *de-novo* mutations in congenital heart disease





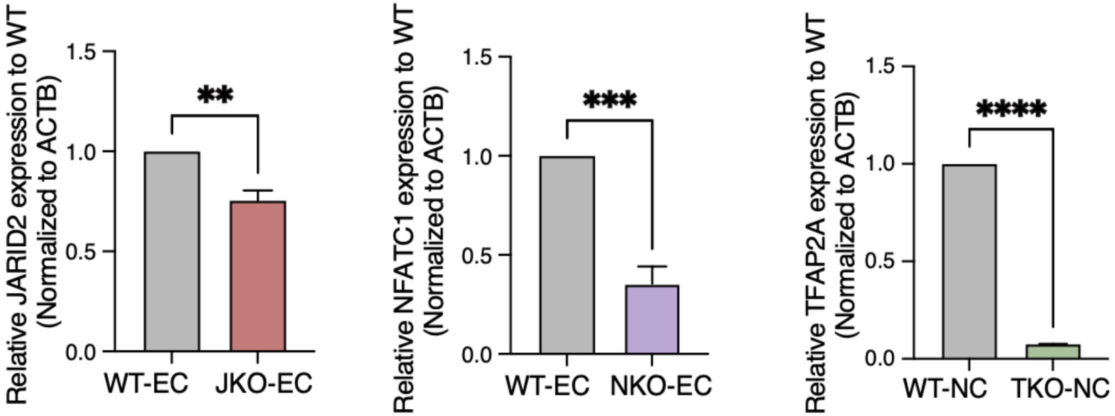
with Wang, Karakikes, Quertermous, Greenleaf labs



Laksshman Sundaram

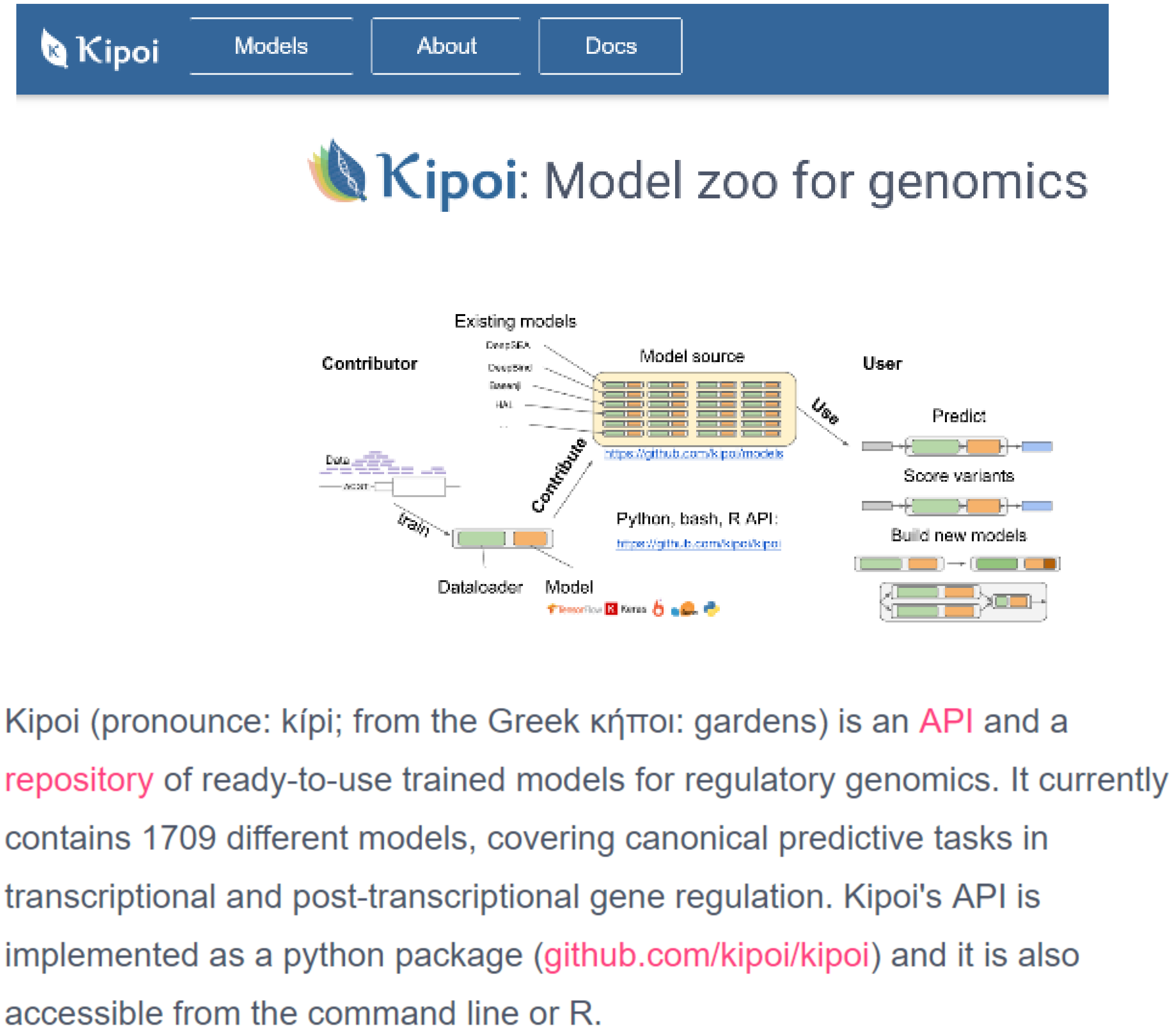
Mo Ameen

Prediction: Mutation disrupts ETV motif in control element active in arterial endothelial cells



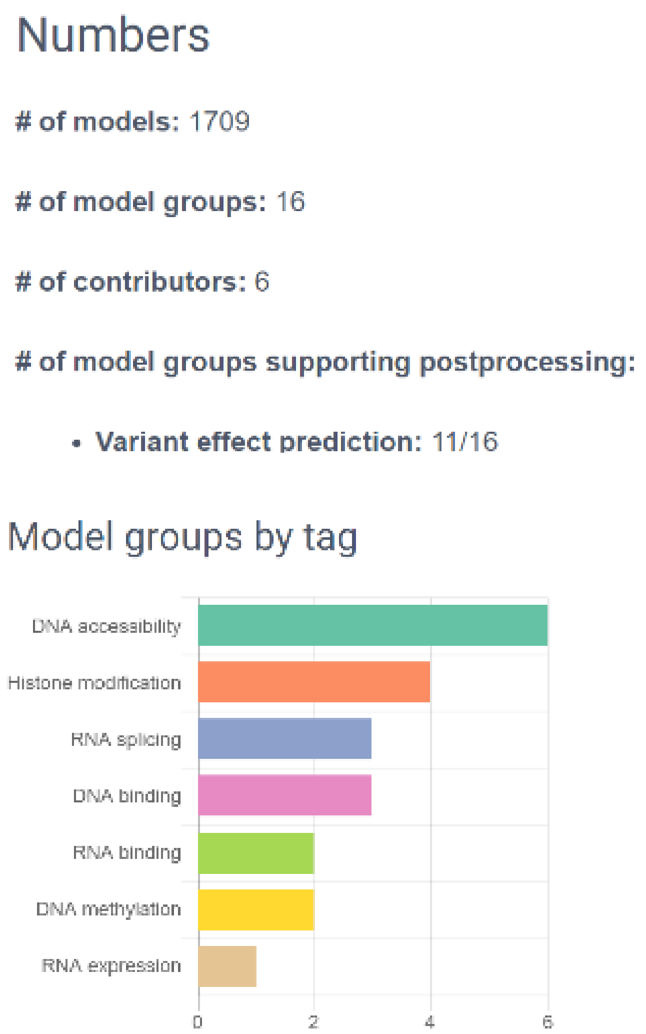
CRISPR/Cas9 experiments validate downstream target genes

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



Kipoi: Model zoo for genomics

Kipoi (pronounce: kípi; from the Greek κήποι: 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

Avsec et al. 2019 Nature Biotech

Summary

- Large-scale molecular profiling datasets => decipher genome function
- Neural networks can map DNA sequence to molecular profiles with unprecedented accuracy
- Models can be interpreted to decipher functional DNA letters, words and syntax
- Models can be used to decipher disease-associated mutations
- Predictions are validated by genome editing experiments
- Predictions can provide clues for therapeutic interventions

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)

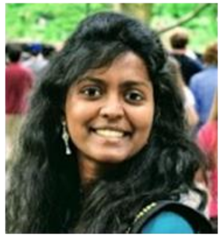
Georgi Marinov (Postdoc)

Alex Tseng (CS)

Amr Alexandari (CS)

Abhimanyu Banerjee (Physics)

Laksshman Sundaram (CS)


Anusri Pampari (CS)

Kristy Mualim (Bioinformatician)

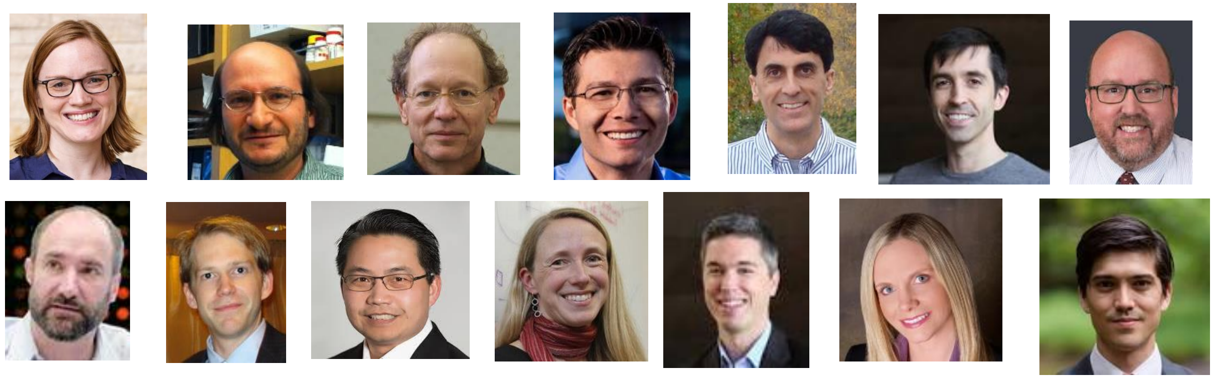
Jacob Schreiber (Postdoc)

Mahfuza Sharmin (Postdoc)

Eran Kotler (Postdoc)

Zahoor Zafrulla (ML engineer)

Collaborators



Funding

ALFRED P. SLOAN FOUNDATION 1934

1R01HG009674

1U01HG009431

1U24HG009446

ENCODE

R01ES02500902

ROADMAP epigenomics PROJECT

1DP2OD022870

NIH DIRECTOR'S NEW INNOVATOR AWARD