AI for genomics

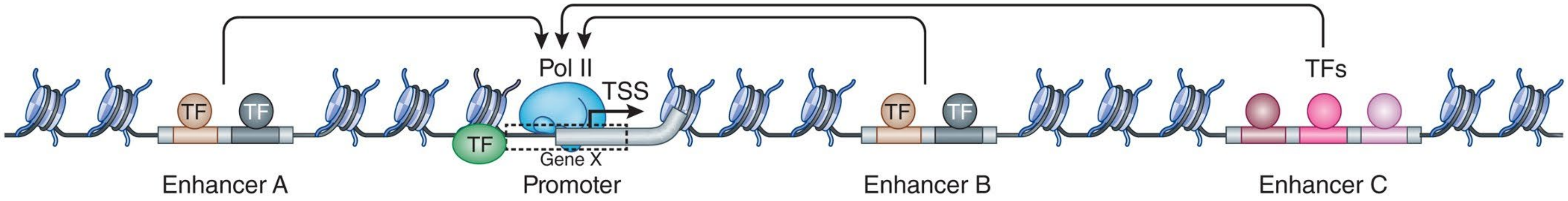
Lieke Michielsen

Department of Human Genetics, LUMC

Delft Bioinformatics Lab, TU Delft

Understanding the non-coding genome

- Only 1.5% of the genome is protein-coding
- Non-coding genome is important for transcriptional control

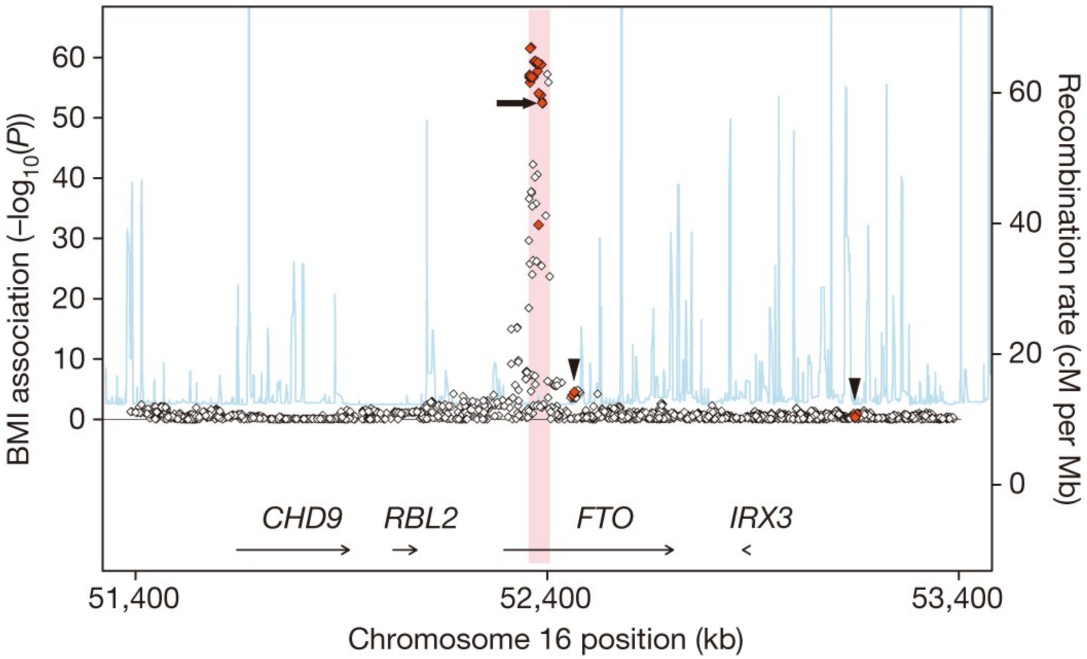


Elkon & Agami, 2017 (https://www.nature.com/articles/nbt.3863)

Understanding the non-coding genome

- ~95% of GWAS variants fall in the non-coding region
  - Which variant is causal?
  - Which gene is affected?
  - Which cell type or tissue is affected?

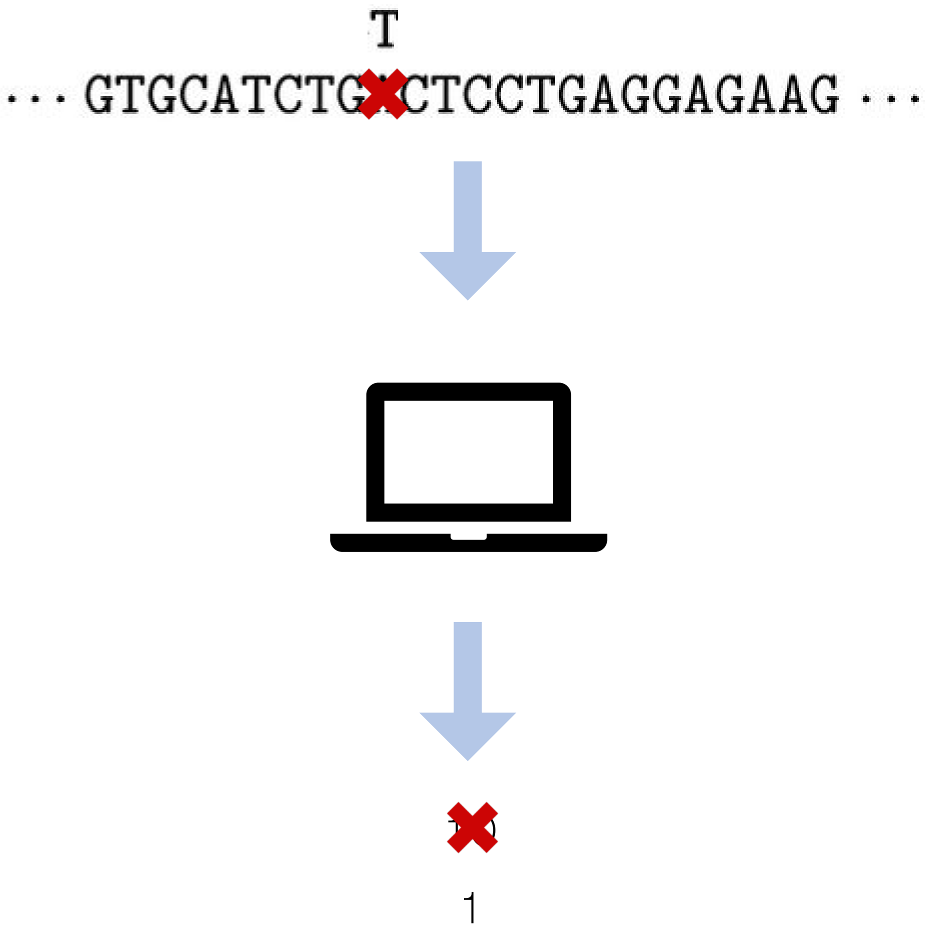
- For eQTLs we know the gene and cell type or tissue, but still suffers from LD



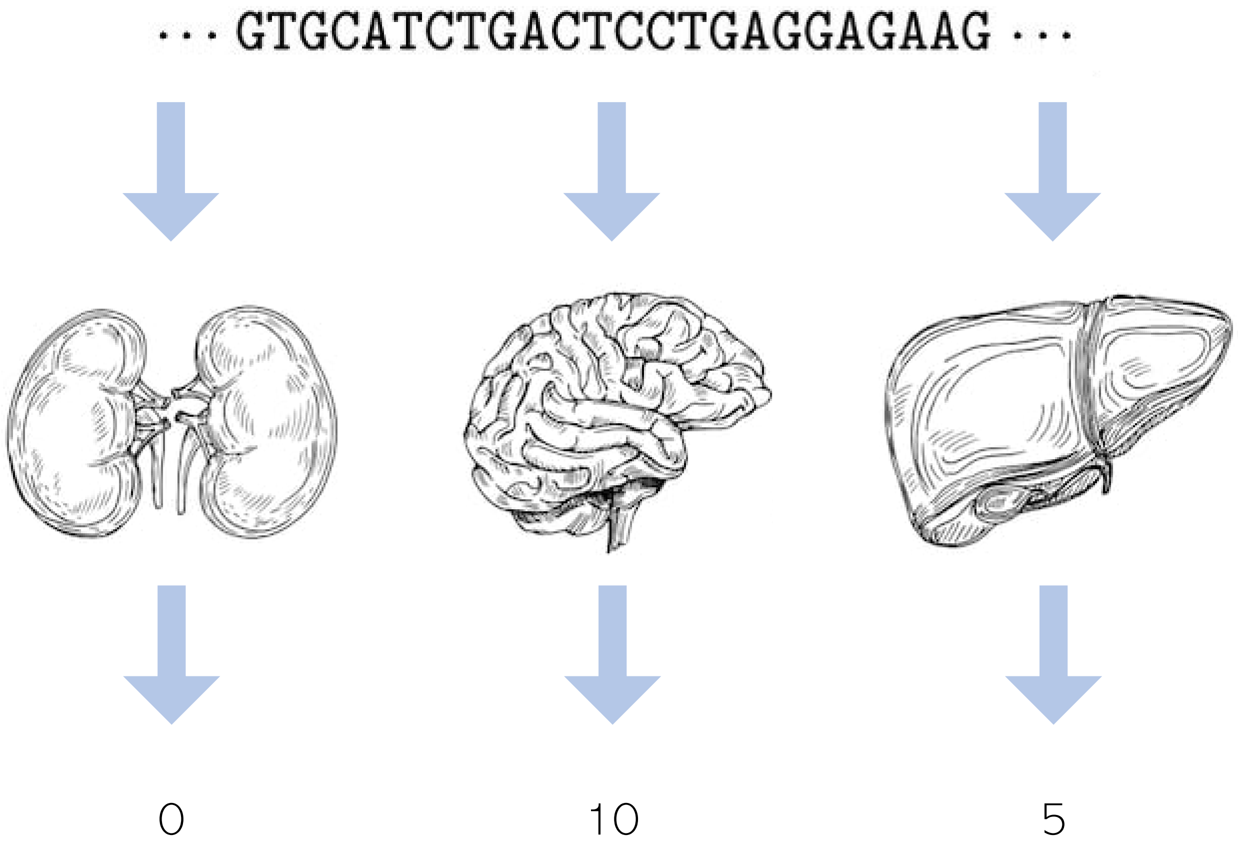
Smemo et al., 2014 (https://www.nature.com/articles/nature13138)

AI for genomics

- Train machine learning models to predict genomic features using the DNA sequence
    - TF binding sites
    - Chromatin accessibility
    - Histone modifications
    - Gene expression
    - ….



Tissue-specific predictions



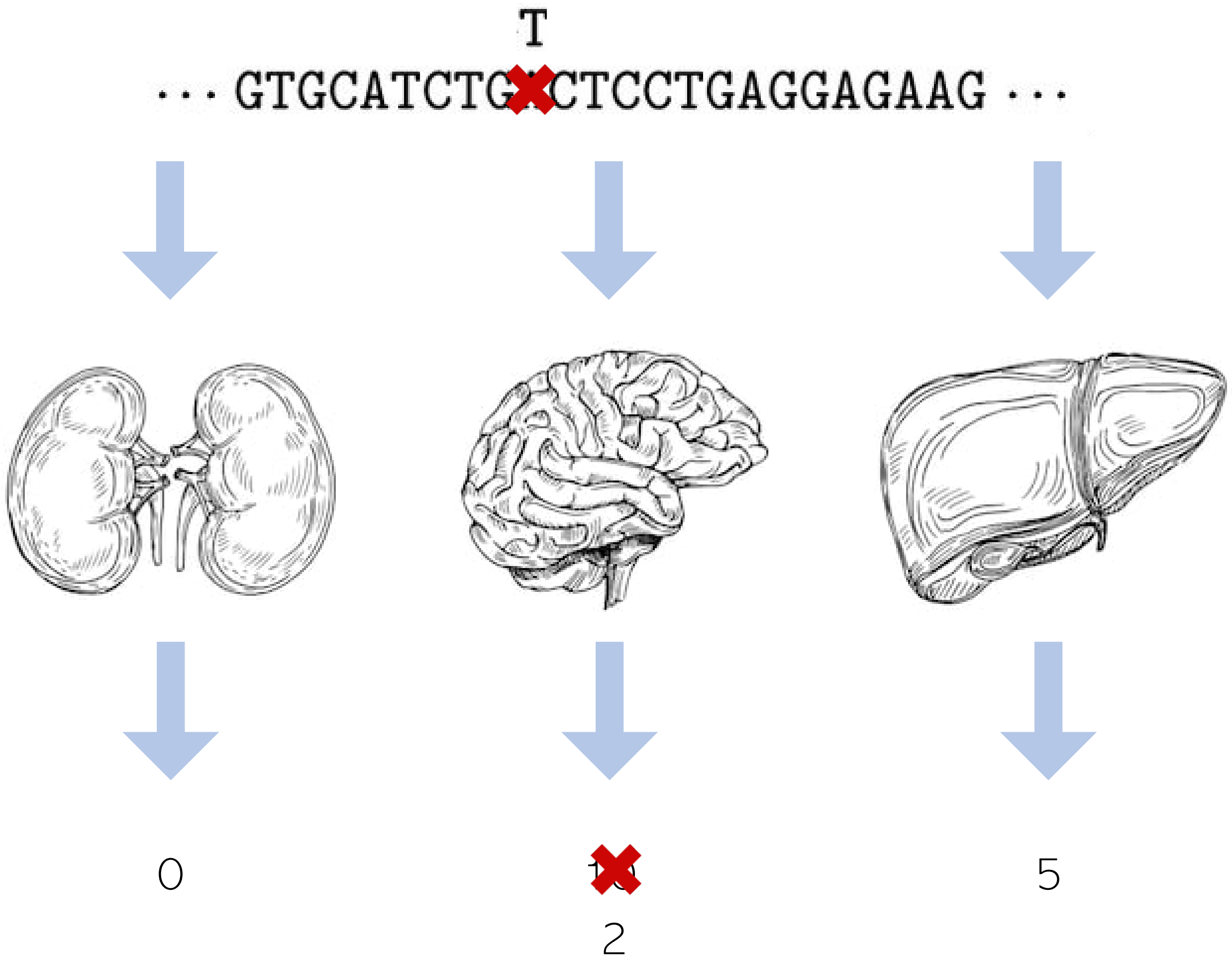
· · · GTGCATCTGACTCCTGAGGAGAAG · · ·

0

10

5

Tissue-specific predictions



T

··· GTGCATCTG✖CTCCTGAGGAGAAG ···

0

~~10~~

2

5

Outline

- Modeling the local sequence
    - Predicting TF binding sites
    - Predicting other genomic features
- Modeling long-range interactions
- Model interpretation

Outline

- Modeling the local sequence
  - Predicting TF binding sites
  - Predicting other genomic features
- Modeling long-range interactions
- Model interpretation

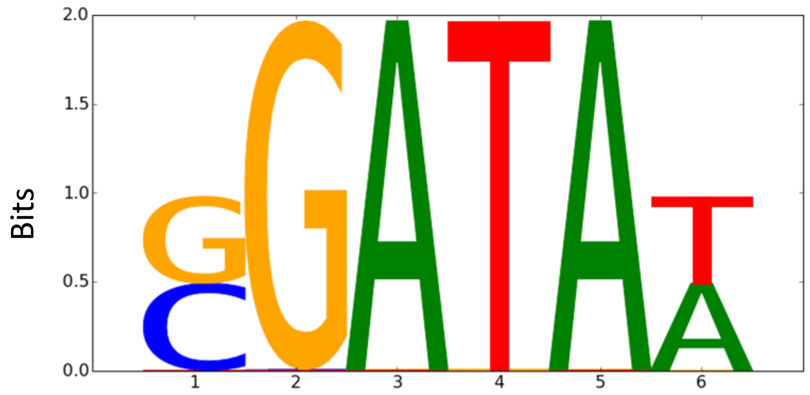
Sequence motifs

```
GGATAA
CGATAA
CGATAT
GGATAT
```

Set of aligned sequences bound by TF

A	0	0	1	0	1	0.5
C	0.5	0	0	0	0	0
G	0.5	1	0	0	0	0
T	0	0	0	1	0	0.5

Position weight matrix (PWM)



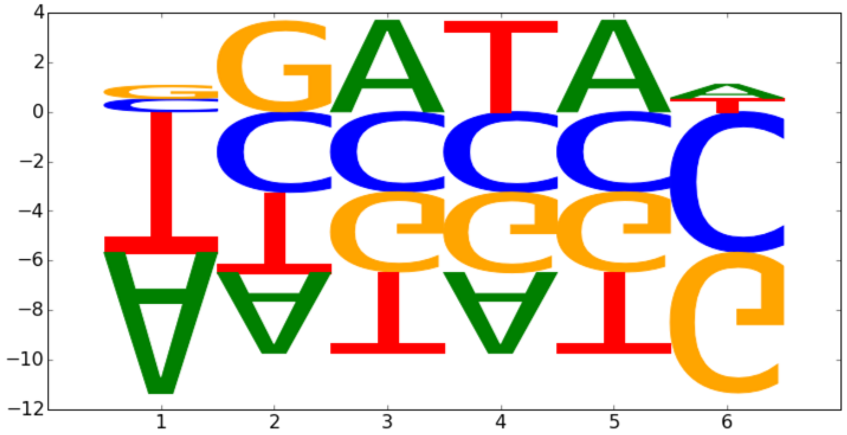
Sequence logo

Sequence motifs

- Position-specific scoring matrix (PSSM): accounting for genomic background nucleotide distribution

A	-5.7	-3.2	3.7	-3.2	3.7	0.5
C	0.5	-3.2	-3.2	-3.2	-3.2	-5.7
G	0.5	3.7	-3.2	-3.2	-3.2	-5.7
T	-5.7	-3.2	-3.2	3.7	-3.2	0.5

$$\log_2 \frac{p_i(x_i = a_i)}{p_{bg}(x_i = a_i)}$$



PSSM logo

Convolution: scoring a sequence with a PSSM

Motif match scores
$(W * x)$

Scoring weights
(W)

A	-5.7	-3.2	3.7	-3.2	3.7	0.5
C	0.5	-3.2	-3.2	-3.2	-3.2	-5.7
G	0.5	3.7	-3.2	-3.2	-3.2	-5.7
T	-5.7	-3.2	-3.2	3.7	-3.2	0.5

One-hot encoding
(X)

■	□	□	□	■	□	□	■	□	□	□	■	□	■	■
□	□	□	■	□	□	□	□	■	■	□	□	□	□	□
□	□	■	□	□	□	□	□	□	□	■	□	□	□	□
□	■	□	□	□	■	■	□	□	□	□	□	■	□	□
A	T	G	C	A	T	T	A	C	C	G	A	T	A	A

Input sequence

Convolution: scoring a sequence with a PSSM

Motif match scores
$(W \ * \ x)$

-11.1									

Scoring weights
(W)

A	-5.7	-3.2	3.7	-3.2	3.7	0.5
C	0.5	-3.2	-3.2	-3.2	-3.2	-5.7
G	0.5	3.7	-3.2	-3.2	-3.2	-5.7
T	-5.7	-3.2	-3.2	3.7	-3.2	0.5

$$\begin{gathered}-5.7*1+0.5*0+0.5*0-5.7*0+\\-3.2*0-3.2*0+3.7*0-3.2*1+\\\ldots\\0.5*0-5.7*0-5.7*0+0.5*1=-11.1\end{gathered}$$

One-hot encoding
(X)

	■	□	□	□	■	□	□	■	□	□	□	■	□	■	■
	□	□	□	■	□	□	□	□	■	■	□	□	□	□	□
	□	□	■	□	□	□	□	□	□	□	■	□	□	□	□
	□	■	□	□	□	■	■	□	□	□	□	□	■	□	□
Input sequence	A	T	G	C	A	T	T	A	C	C	G	A	T	A	A

Convolution: scoring a sequence with a PSSM

Motif match scores
$(W * x)$

-11.1	-11.1								

Scoring weights
(W)

A	-5.7	-3.2	3.7	-3.2	3.7	0.5
C	0.5	-3.2	-3.2	-3.2	-3.2	-5.7
G	0.5	3.7	-3.2	-3.2	-3.2	-5.7
T	-5.7	-3.2	-3.2	3.7	-3.2	0.5

One-hot encoding
(X)

Input sequence

■	□	□	□	■	□	□	■	□	□	□	■	□	■	■
□	□	□	■	□	□	□	□	■	■	□	□	□	□	□
□	□	■	□	□	□	□	□	□	□	■	□	□	□	□
□	■	□	□	□	■	■	□	□	□	□	□	■	□	□
A	T	G	C	A	T	T	A	C	C	G	A	T	A	A

Convolution: scoring a sequence with a PSSM

Motif match scores $(W * x)$

-11.1	-11.1	2.0							

Scoring weights (W)

A	-5.7	-3.2	**3.7**	-3.2	3.7	**0.5**
C	0.5	**-3.2**	-3.2	-3.2	-3.2	-5.7
G	**0.5**	3.7	-3.2	-3.2	-3.2	-5.7
T	-5.7	-3.2	-3.2	**3.7**	**-3.2**	0.5

One-hot encoding (X)

	■				■			■				■		■	■
				■					■	■					
			■								■				
		■				■	■						■		
Input sequence	A	T	G	C	A	T	T	A	C	C	G	A	T	A	A

Convolution: scoring a sequence with a PSSM

Motif match scores
$(W * x)$

-11.1	-11.1	2.0	-4.2						

Scoring weights
(W)

A	-5.7	**-3.2**	3.7	-3.2	**3.7**	0.5
C	**0.5**	-3.2	-3.2	-3.2	-3.2	**-5.7**
G	0.5	3.7	-3.2	-3.2	-3.2	-5.7
T	-5.7	-3.2	**-3.2**	**3.7**	-3.2	0.5

One-hot encoding
(X)

Input sequence

	■	□	□	□	■	□	□	■	□	□	□	■	□	■	■
	□	□	□	■	□	□	□	□	■	■	□	□	□	□	□
	□	□	■	□	□	□	□	□	□	□	■	□	□	□	□
	□	■	□	□	□	■	■	□	□	□	□	□	■	□	□
Input sequence	A	T	G	C	A	T	T	A	C	C	G	A	T	A	A

Convolution: scoring a sequence with a PSSM

Motif match scores
$(W * x)$

-11.1	-11.1	2.0	-4.2	-24.2	-17.3	-18.0	-11.1	-11.8	15.8

Scoring weights
(W)

A	-5.7	-3.2	3.7	-3.2	3.7	0.5
C	0.5	-3.2	-3.2	-3.2	-3.2	-5.7
G	0.5	3.7	-3.2	-3.2	-3.2	-5.7
T	-5.7	-3.2	-3.2	3.7	-3.2	0.5

One-hot encoding
(X)

Input sequence

■	□	□	□	■	□	□	■	□	□	□	■	□	■	■
□	□	□	■	□	□	□	□	■	■	□	□	□	□	□
□	□	■	□	□	□	□	□	□	□	■	□	□	□	□
□	■	□	□	□	■	■	□	□	□	□	□	■	□	□
A	T	G	C	A	T	T	A	C	C	G	A	T	A	A

Convolution: scoring a sequence with a PSSM

Thresholding the scores

0	0	2.0	0	0	0	0	0	0	15.9

Motif match scores $(W * x)$

-11.1	-11.1	2.0	-4.2	-24.2	-17.3	-18.0	-11.1	-11.8	15.8

Scoring weights (W)

A	-5.7	-3.2	3.7	-3.2	3.7	0.5
C	0.5	-3.2	-3.2	-3.2	-3.2	-5.7
G	0.5	3.7	-3.2	-3.2	-3.2	-5.7
T	-5.7	-3.2	-3.2	3.7	-3.2	0.5

One-hot encoding (X)	■				■			■				■		■	■
				■					■	■					
			■								■				
		■				■	■						■		
Input sequence	A	T	G	C	A	T	T	A	C	C	G	A	T	A	A

What if the PSSM is unknown?

Measured binding sites (targets)

0	0	1	0	0	0	0	0	0	1

Motif match scores
$(W * x)$

Scoring weights
(W)

A	w_1	w_2	w_3	w_4	w_5	w_6
C	w_7	w_8	w_9	w_{10}	w_{11}	w_{12}
G	w_{13}	w_{14}	w_{15}	w_{16}	w_{17}	w_{18}
T	w_{19}	w_{20}	w_{21}	w_{22}	w_{23}	w_{24}

One-hot encoding
(X)

■	□	□	□	■	□	□	■	□	□	□	■	□	■	■
□	□	□	■	□	□	□	□	■	■	□	□	□	□	□
□	□	■	□	□	□	□	□	□	□	■	□	□	□	□
□	■	□	□	□	■	■	□	□	□	□	□	■	□	□

Input sequence A T G C A T T A C C G A T A A

What if the PSSM is unknown?

Measured binding sites (targets)

0	0	1	0	0	0	0	0	0	1

Motif match scores
$(W * x)$

Scoring weights
(W)

A	-1.5	-1.4	-1.3	1.3	0.5	-0.6
C	1.1	-0.8	0.7	-0.3	0.4	-0.9
G	-1.0	-0.1	0.0	1.5	-1.4	1.1
T	1.0	-0.5	-0.9	-1.1	-1.0	-0.3

One-hot encoding
(X)

■				■			■				■		■	■
			■					■	■					
		■								■				
	■				■	■						■		

Input sequence

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

1. Randomly initialize W

What if the PSSM is unknown?

Measured binding sites (targets)

0	0	1	0	0	0	0	0	0	1

Motif match scores
$(W * x)$

-2.1	1.6	-5.8	-2.7	-2.1	0.4	-2.0	0.1	0.0	-1.5

Scoring weights
(W)

A	-1.5	-1.4	-1.3	1.3	0.5	-0.6
C	1.1	-0.8	0.7	-0.3	0.4	-0.9
G	-1.0	-0.1	0.0	1.5	-1.4	1.1
T	1.0	-0.5	-0.9	-1.1	-1.0	-0.3

1. Randomly initialize W
2. Convolution between input sequences and scoring weights

One-hot encoding
(X)

■	□	□	□	■	□	□	■	□	□	□	■	□	■	■
□	□	□	■	□	□	□	□	■	■	□	□	□	□	□
□	□	■	□	□	□	□	□	□	□	■	□	□	□	□
□	■	□	□	□	■	■	□	□	□	□	□	■	□	□

Input sequence A T G C A T T A C C G A T A A

What if the PSSM is unknown?

Measured binding sites (targets)

0	0	1	0	0	0	0	0	0	1

Motif match scores
$(W * x)$

-2.1	1.6	-5.8	-2.7	-2.1	0.4	-2.0	0.1	0.0	-1.5

Scoring weights
(W)

A	-1.5	-1.4	-1.3	1.3	0.5	-0.6
C	1.1	-0.8	0.7	-0.3	0.4	-0.9
G	-1.0	-0.1	0.0	1.5	-1.4	1.1
T	1.0	-0.5	-0.9	-1.1	-1.0	-0.3

1. Randomly initialize W
2. Convolution between input sequences and scoring weights
3. Activation function

One-hot encoding
(X)

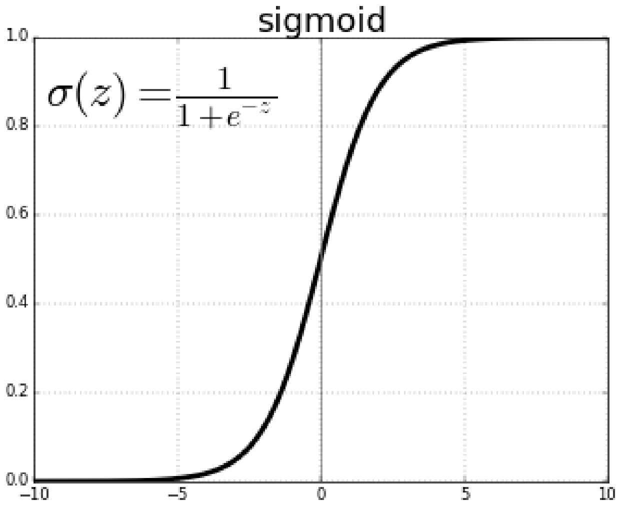
■	□	□	□	■	□	□	■	□	□	□	■	□	■	■
□	□	□	■	□	□	□	□	■	■	□	□	□	□	□
□	□	■	□	□	□	□	□	□	□	■	□	□	□	□
□	■	□	□	□	■	■	□	□	□	□	□	■	□	□

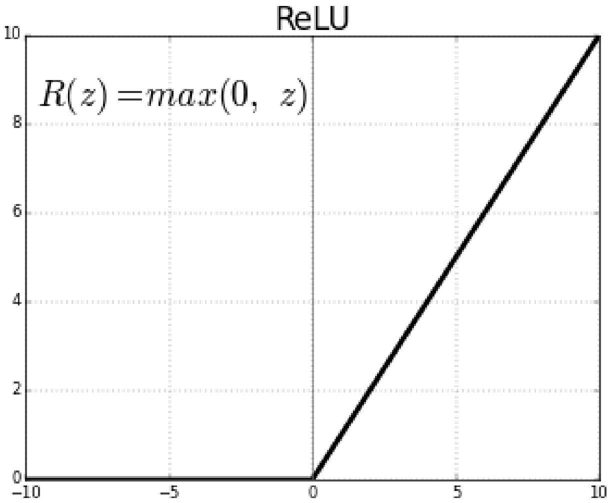
Input sequence A T G C A T T A C C G A T A A

Activation function

- Motif match scores can be very high positive and negative numbers
- Targets are binarized (binding yes/no) or positive (how often binding was measured)

- Activation function maps the scores to the correct range





What if the PSSM is unknown?

Measured binding sites (targets)

0	0	1	0	0	0	0	0	0	1

Motif match scores
$(W * x)$

0.1	0.83	0.0	0.1	0.1	0.6	0.1	0.5	0.5	0.2

Scoring weights
(W)

A	-1.5	-1.4	-1.3	1.3	0.5	-0.6
C	1.1	-0.8	0.7	-0.3	0.4	-0.9
G	-1.0	-0.1	0.0	1.5	-1.4	1.1
T	1.0	-0.5	-0.9	-1.1	-1.0	-0.3

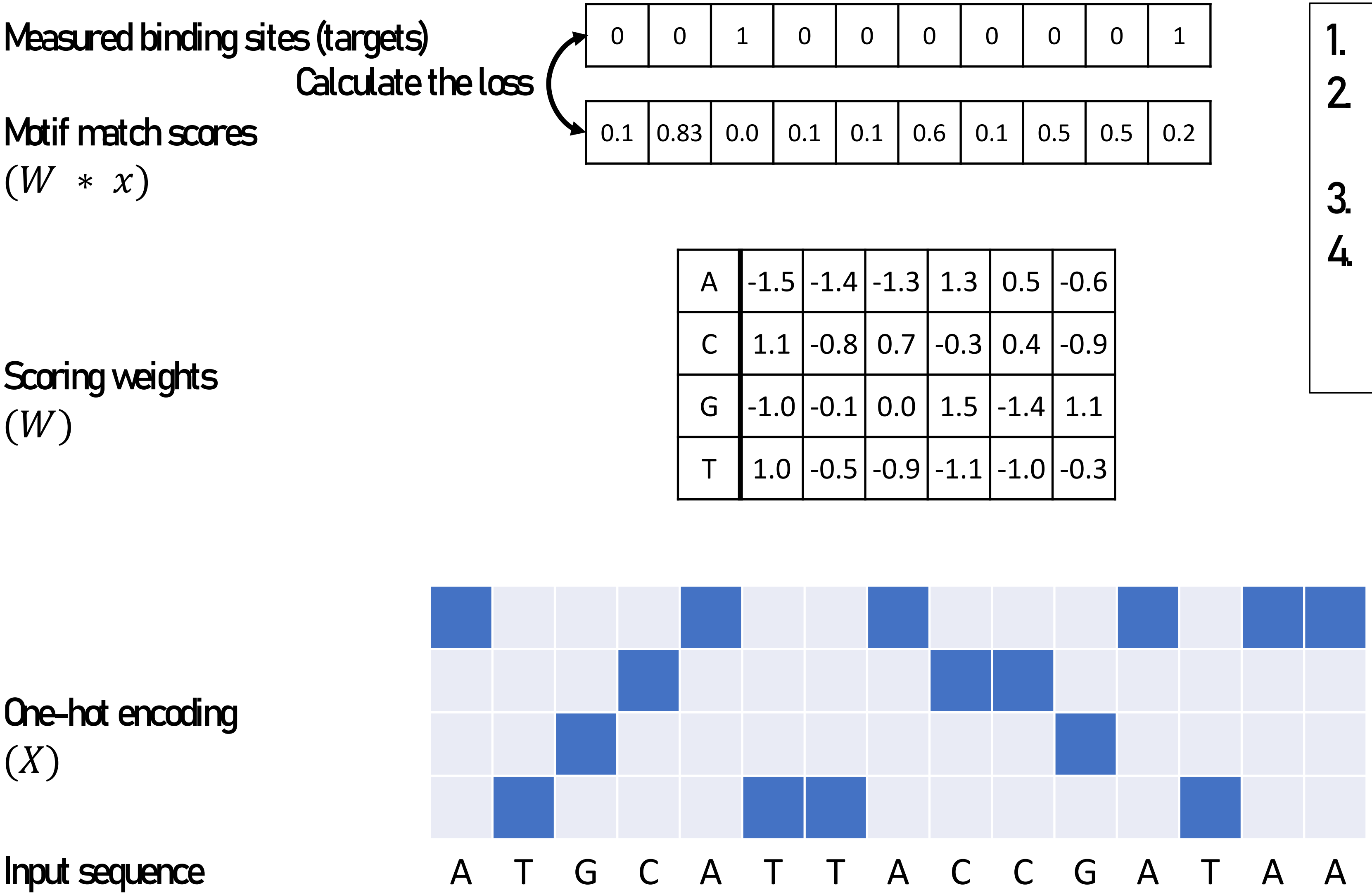
One-hot encoding
(X)

■	□	□	□	■	□	□	■	□	□	□	■	□	■	■
□	□	□	■	□	□	□	□	■	■	□	□	□	□	□
□	□	■	□	□	□	□	□	□	□	■	□	□	□	□
□	■	□	□	□	■	■	□	□	□	□	□	■	□	□

Input sequence A T G C A T T A C C G A T A A

1. Randomly initialize W
2. Convolution between input sequences and scoring weights
3. Activation function

What if the PSSM is unknown?



Measured binding sites (targets)

0	0	1	0	0	0	0	0	0	1

Calculate the loss

Motif match scores
$(W * x)$

0.1	0.83	0.0	0.1	0.1	0.6	0.1	0.5	0.5	0.2

Scoring weights
(W)

A	-1.5	-1.4	-1.3	1.3	0.5	-0.6
C	1.1	-0.8	0.7	-0.3	0.4	-0.9
G	-1.0	-0.1	0.0	1.5	-1.4	1.1
T	1.0	-0.5	-0.9	-1.1	-1.0	-0.3

One-hot encoding
(X)

Input sequence

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

1. Randomly initialize W
2. Convolution between input sequences and scoring weights
3. Activation function
4. Compare predictions to targets

What if the PSSM is unknown?

Measured binding sites (targets)

0	0	1	0	0	0	0	0	0	1

Motif match scores
$(W * x)$

0.1	0.6	0	0	0	0.1	0	0.3	0.3	0.7

Scoring weights
(W)

A	-1.6	-1.9	-1.0	0.6	0.8	-0.1
C	1.2	-1.1	0.2	-0.9	-0.1	-1.7
G	-0.9	0.5	-0.2	0.6	-1.5	0.2
T	0.5	-1.0	-1.2	-0.5	-1.4	0.1

One-hot encoding
(X)

■	□	□	□	■	□	□	■	□	□	□	■	□	■	■
□	□	□	■	□	□	□	□	■	■	□	□	□	□	□
□	□	■	□	□	□	□	□	□	□	■	□	□	□	□
□	■	□	□	□	■	■	□	□	□	□	□	■	□	□

Input sequence A T G C A T T A C C G A T A A

1. Randomly initialize W
2. Convolution between input sequences and scoring weights
3. Activation function
4. Compare predictions to targets
5. Update W
6. Repeat steps 2-5 until convergence

What if the PSSM is unknown?

Measured binding sites (targets)

0	0	1	0	0	0	0	0	0	1

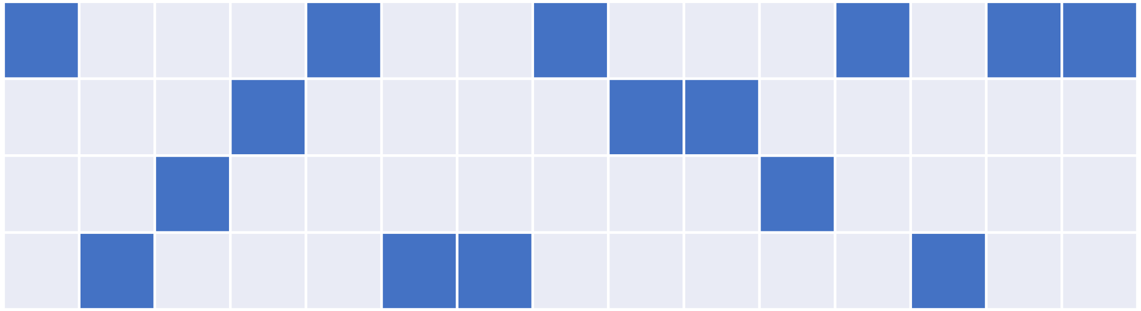
Motif match scores
$(W * x)$

0	0	0.9	0	0	0	0	0	0	1

Scoring weights
(W)

A	-5.7	-3.2	3.7	-3.2	3.7	0.5
C	0.5	-3.2	-3.2	-3.2	-3.2	-5.7
G	0.5	3.7	-3.2	-3.2	-3.2	-5.7
T	-5.7	-3.2	-3.2	3.7	-3.2	0.5

One-hot encoding
(X)



Input sequence A T G C A T T A C C G A T A A

1. Randomly initialize W
2. Convolution between input sequences and scoring weights
3. Activation function
4. Compare predictions to targets
5. Update W
6. Repeat steps 2-5 until convergence

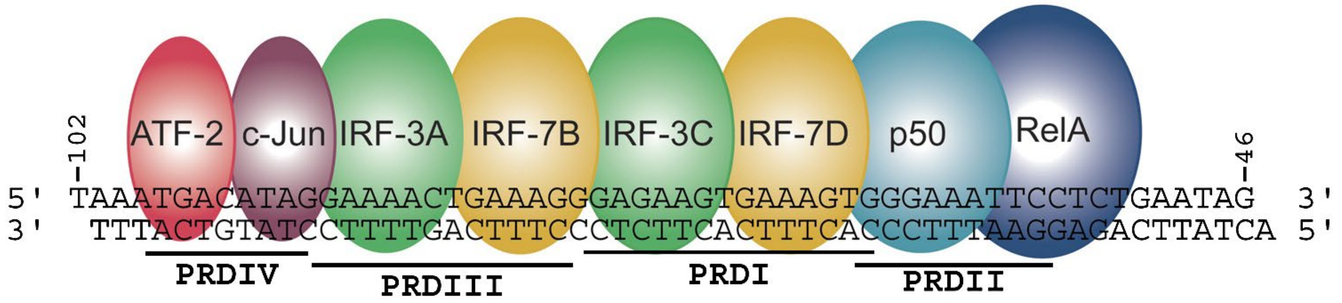
What if the PSSM is unknown?

- Example of a very small convolutional neural network (CNN)

- Train this CNN using the reference genome
- Divide reference genome in bins of ~128bp
- Divide bins into train – validation – test set
  - Train: learn the weights of the filter
  - Validation: optimize hyperparameters of the network
  - Test: evaluate the performance of your model

Usually, it's a bit more complex...

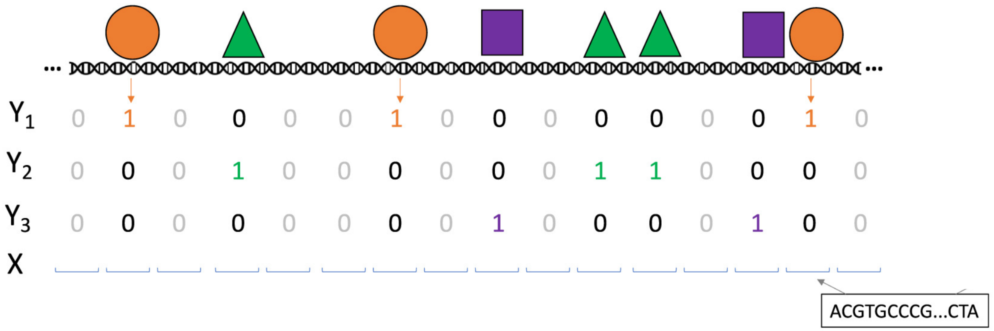
- TFs prefer to bind together
- Spacing between the motifs is important


ATF-2
c-Jun
IRF-3A
IRF-7B
IRF-3C
IRF-7D
p50
RelA
−102
−46
5' TAAATGACATAGGAAAACTGAAAGGGAGAAGTGAAAGTGGGAAATTCCTCTGAATAG 3'
3' TTTACTGTATCCTTTTGACTTTCCCTCTTCACTTTCACCCTTTAAGGAGACTTATCA 5'
PRDIV
PRDIII
PRDI
PRDII

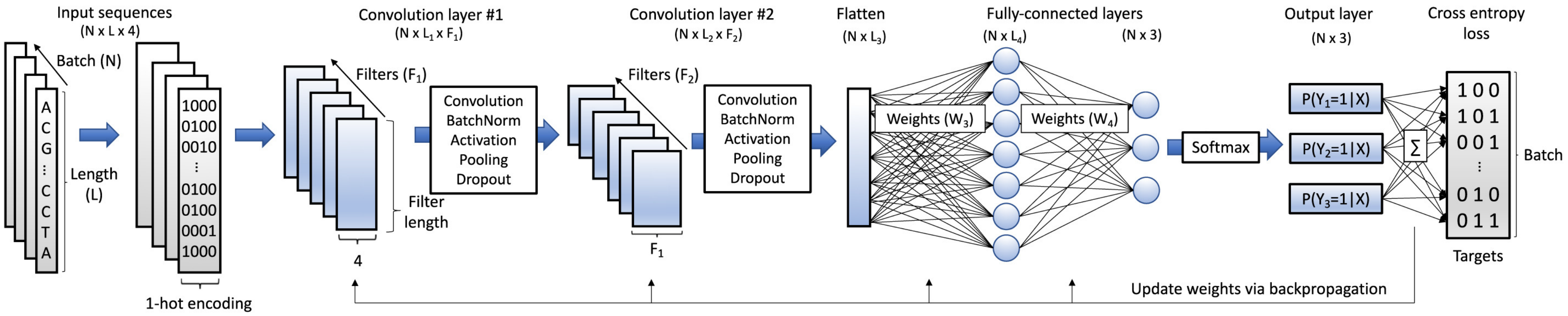
Panne et al., 2014 (https://www.cell.com/fulltext/S0092-8674%2807%2900656-3)

Deep convolutional neural networks

- Multitask learning: predict binding sites for multiple TF simultaneously



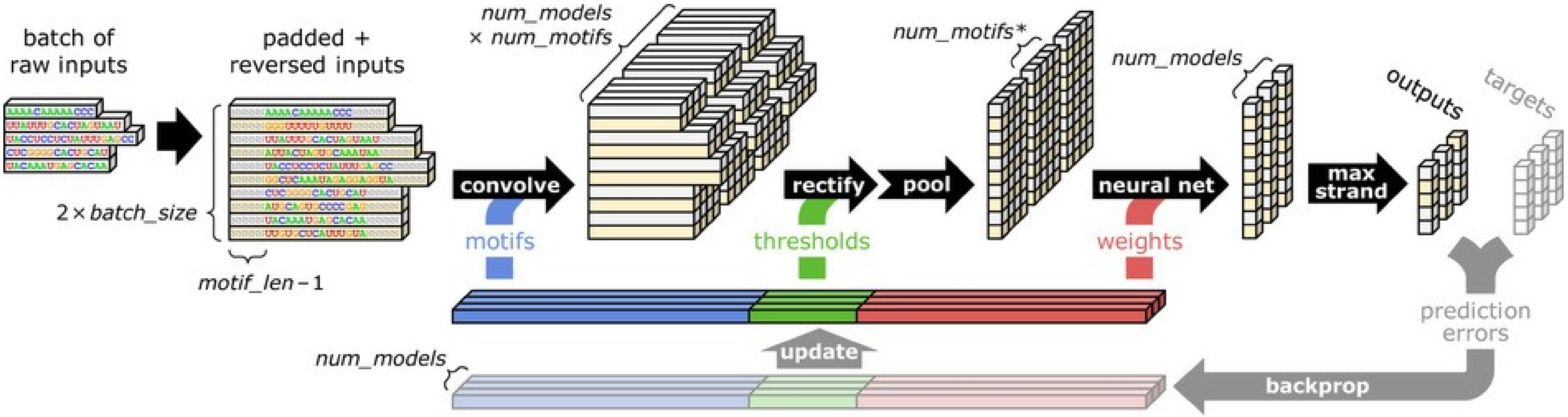
Y1 0 1 0 0 0 0 1 0 0 0 0 0 0 0 1 0
Y2 0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 0
Y3 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0
X
ACGTGCCCG...CTA



Koo & Ploenzke, 2020 (https://www.sciencedirect.com/science/article/pii/S2452310020300032?via%3Dihub)

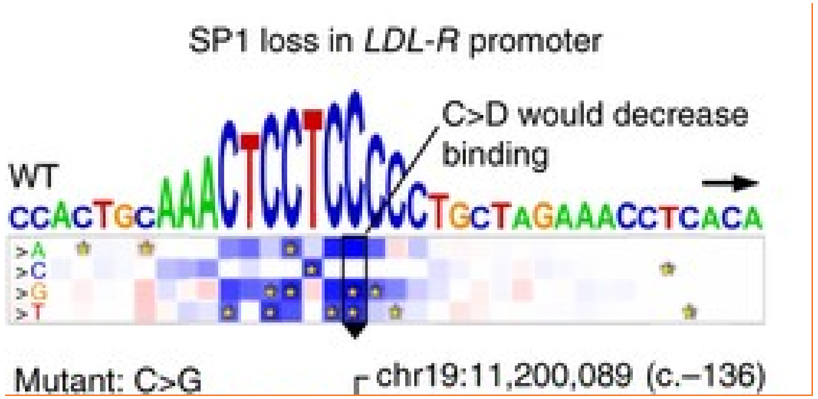
DeepBind

- Input: sequence of 14-101 bp
- Output: predicted binding for 538 TFs and 194 RNA binding proteins



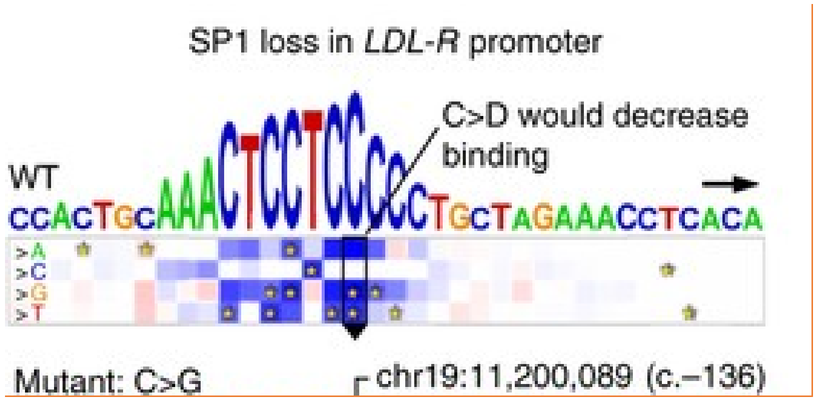
Alipanahi et al., 2015 (https://www.nature.com/articles/nbt.3300)

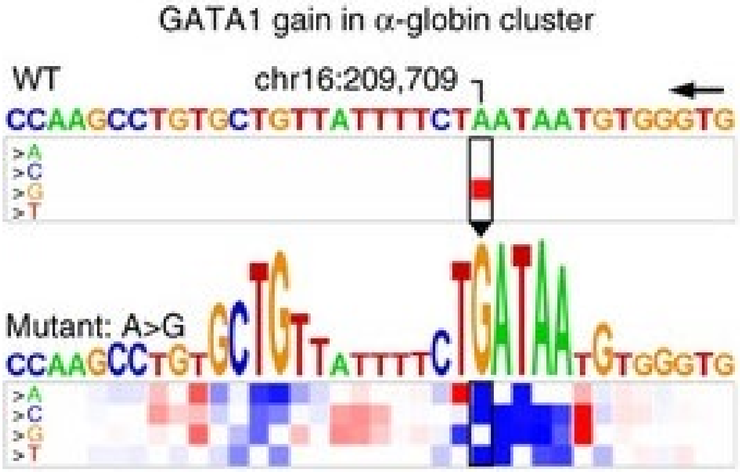
Analysing potentially disease-causing variants


SP1 loss in *LDL-R* promoter
C>D would decrease binding
WT
CCACTGCAAACTCCTCCCCCTGCTAGAAACCTCACA
>A
>C
>G
>T
Mutant: C>G
chr19:11,200,089 (c.–136)

Alipanahi et al., 2015 (https://www.nature.com/articles/nbt.3300)

Analysing potentially disease-causing variants


SP1 loss in *LDL-R* promoter
C>D would decrease binding
WT
CCACTGCAAACTCCTCCCCCTGCTAGAAACCTCACA
>A
>C
>G
>T
Mutant: C>G
chr19:11,200,089 (c.–136)


GATA1 gain in α-globin cluster
WT
chr16:209,709
CCAAGCCTGTGCTGTTATTTTCTAATAATGTGGGTG
>A
>C
>G
>T
Mutant: A>G
CCAAGCCTGTGCTGTTATTTTCTGATAATGTGGGTG
>A
>C
>G
>T

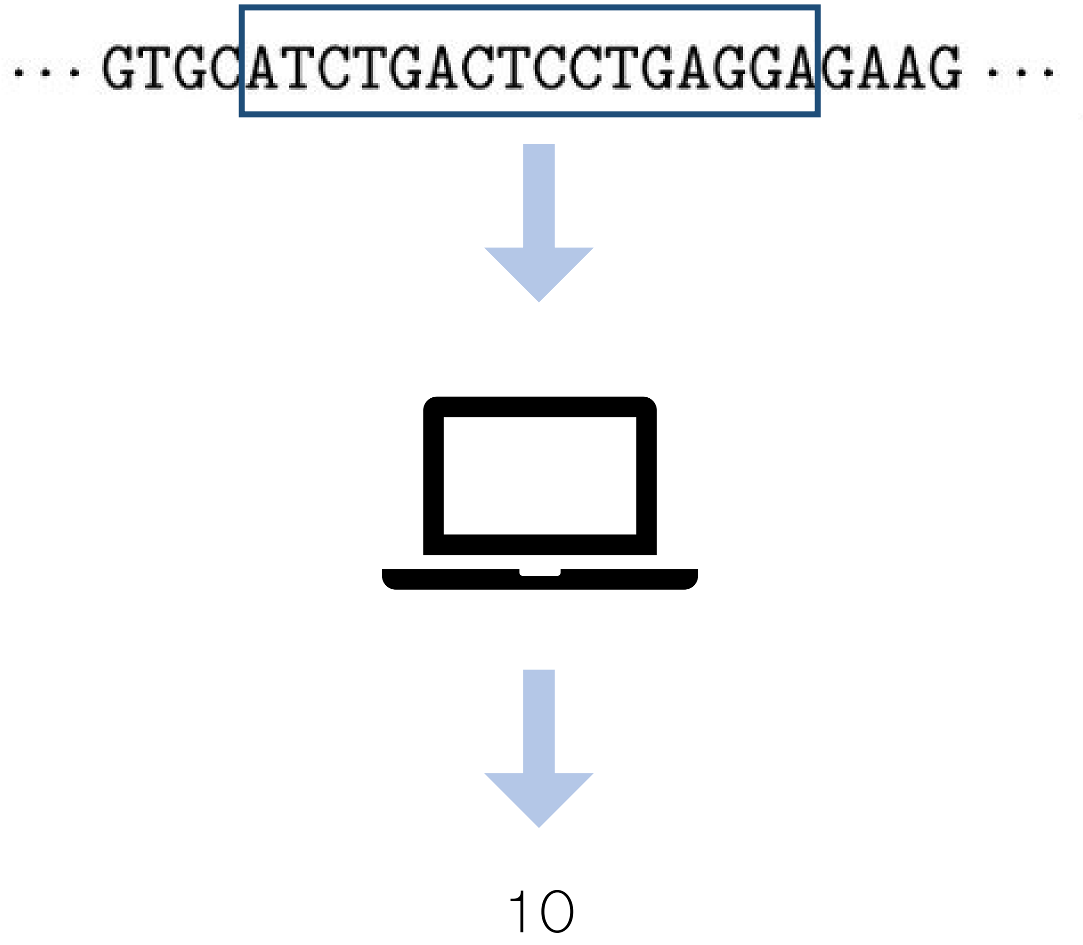
Alipanahi et al., 2015 (https://www.nature.com/articles/nbt.3300)

Outline

- Modeling the local sequence
  - Predicting TF binding sites
  - Predicting other genomic features
- Modeling long-range interactions
- Model interpretation

Predicting other genomic features

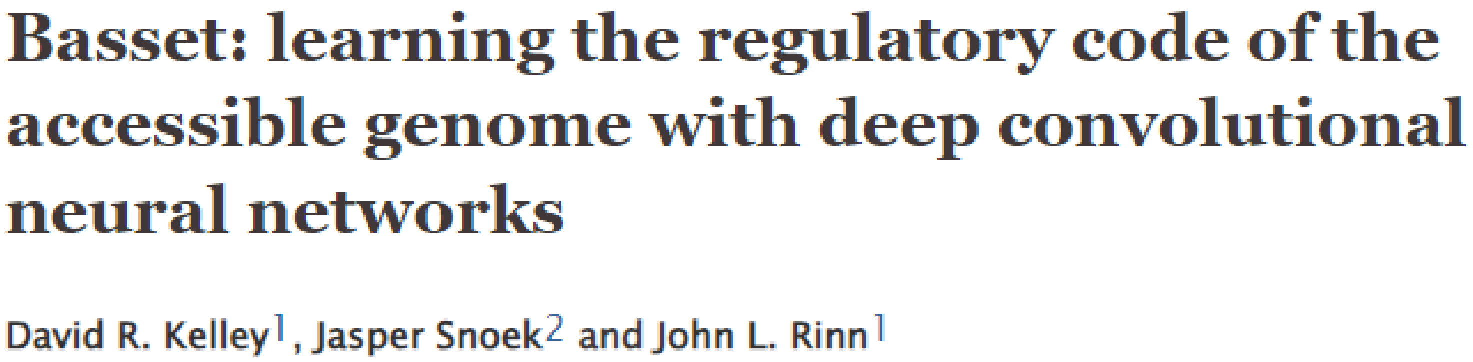
- Is there a TF binding site in this bin?
- Is this part of the DNA accessible?
- Are there histone modifications here?



··· GTGCATCTGACTCCTGAGGAGAAG ···

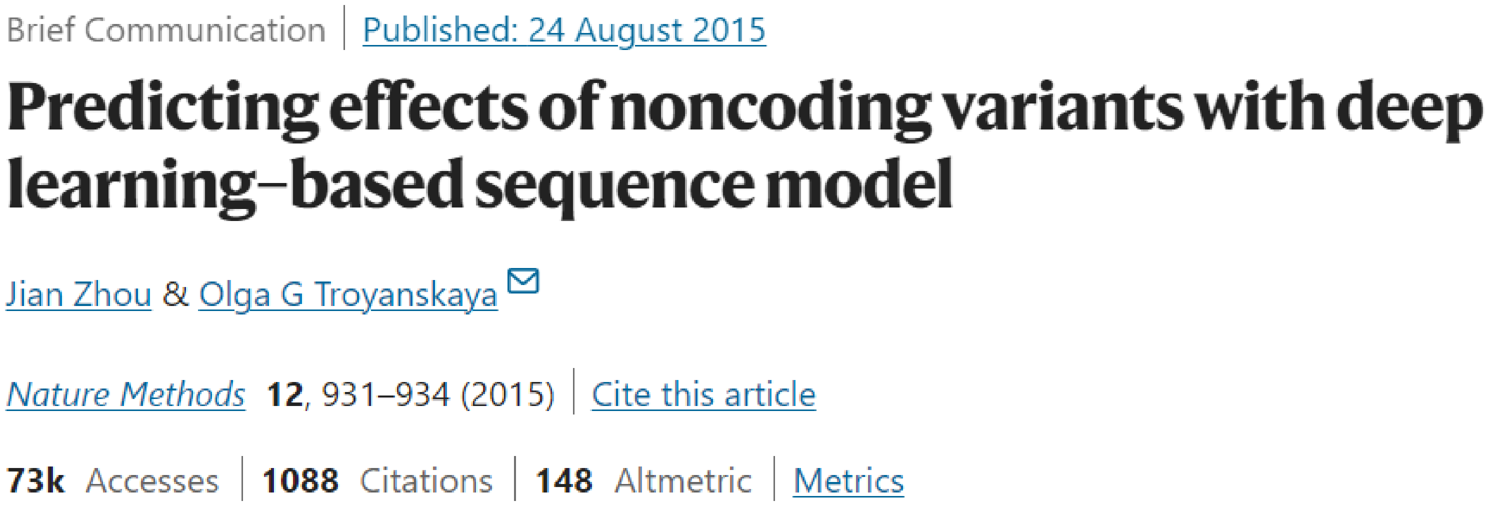
10

Predicting chromatin-related features



Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks

David R. Kelley[1], Jasper Snoek[2] and John L. Rinn[1]



Brief Communication | Published: 24 August 2015

Predicting effects of noncoding variants with deep learning–based sequence model

Jian Zhou & Olga G Troyanskaya

Nature Methods **12**, 931–934 (2015) | Cite this article

73k Accesses | **1088** Citations | **148** Altmetric | Metrics

Kelley et al., 2016 (https://genome.cshlp.org/content/26/7/990), Zhou & Troyanskaya, 2015 (https://www.nature.com/articles/nmeth.3547)

Predicting chromatin-related features

Basset

- Input sequence: 600bp
- Bin size: 600bp

- Predict DNase-seq peaks in 164 tissues/cell types

DeepSEA

- Input sequence: 1000 bp
- Bin size: 200 bp

- Predict 919 chromatin-related features measured in different cell types
  - TF binding sites
  - DNase
  - Histone marks

Outline

- Modeling the local sequence
  - Predicting TF binding sites
  - Predicting other genomic features
- Modeling long-range interactions
- Model interpretation

Downside of previous models

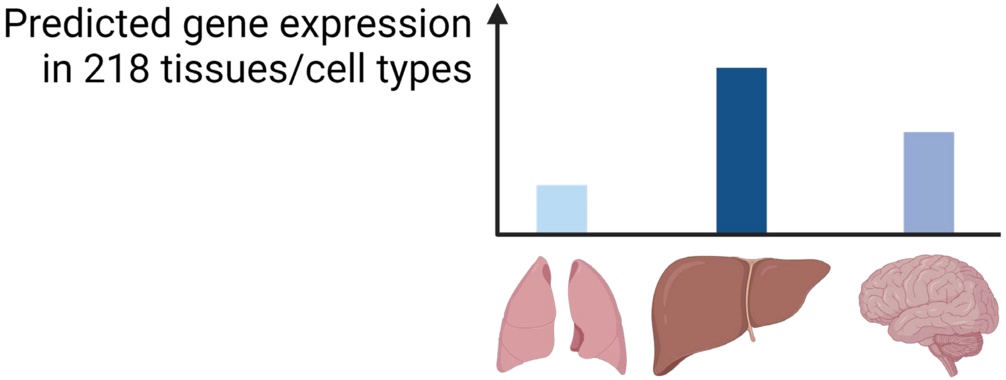
- We know the predicted effect of a mutation on a local genomic feature
- Long-range interactions are important (i.e. the DNA folds)

- What is the effect on gene expression?

ExPecto

Input sequence: 40kb, centered around TSS

··· GTGCATCTGACTCCTGAGGAGAAG ···



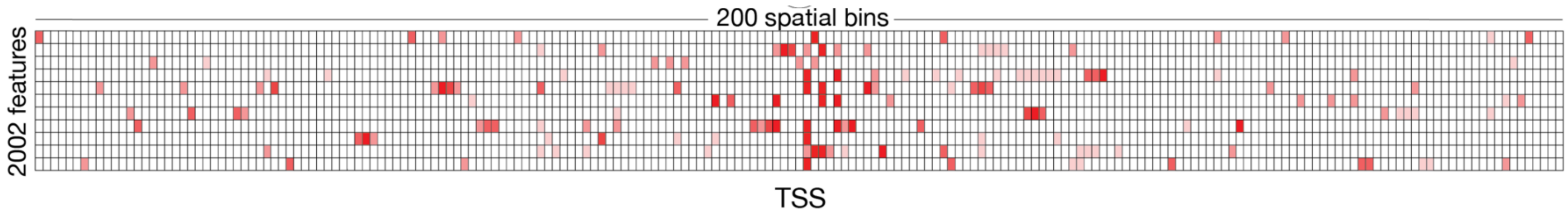
Zhou et al., 2018 (https://www.nature.com/articles/s41588-018-0160-6)

ExPecto

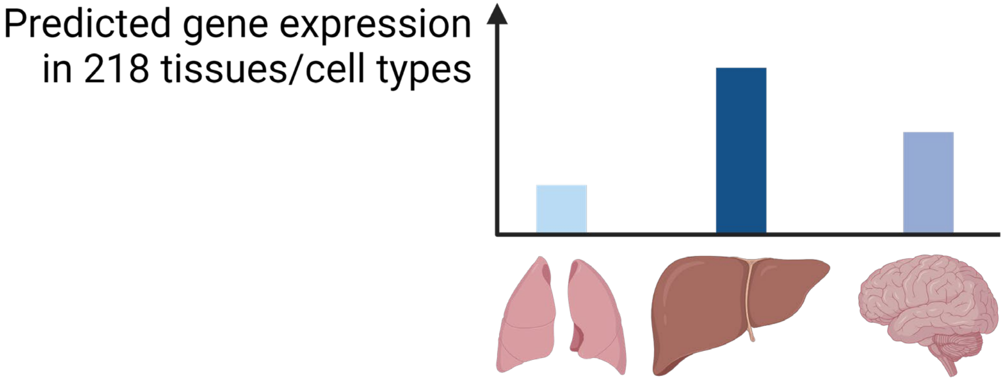
Input sequence: 40kb, centered around TSS

··· GTGCATCTGACTCCTGAGGAGAAG ···

Use DeepSEA to predict 2002 genomic features for 200bp bins



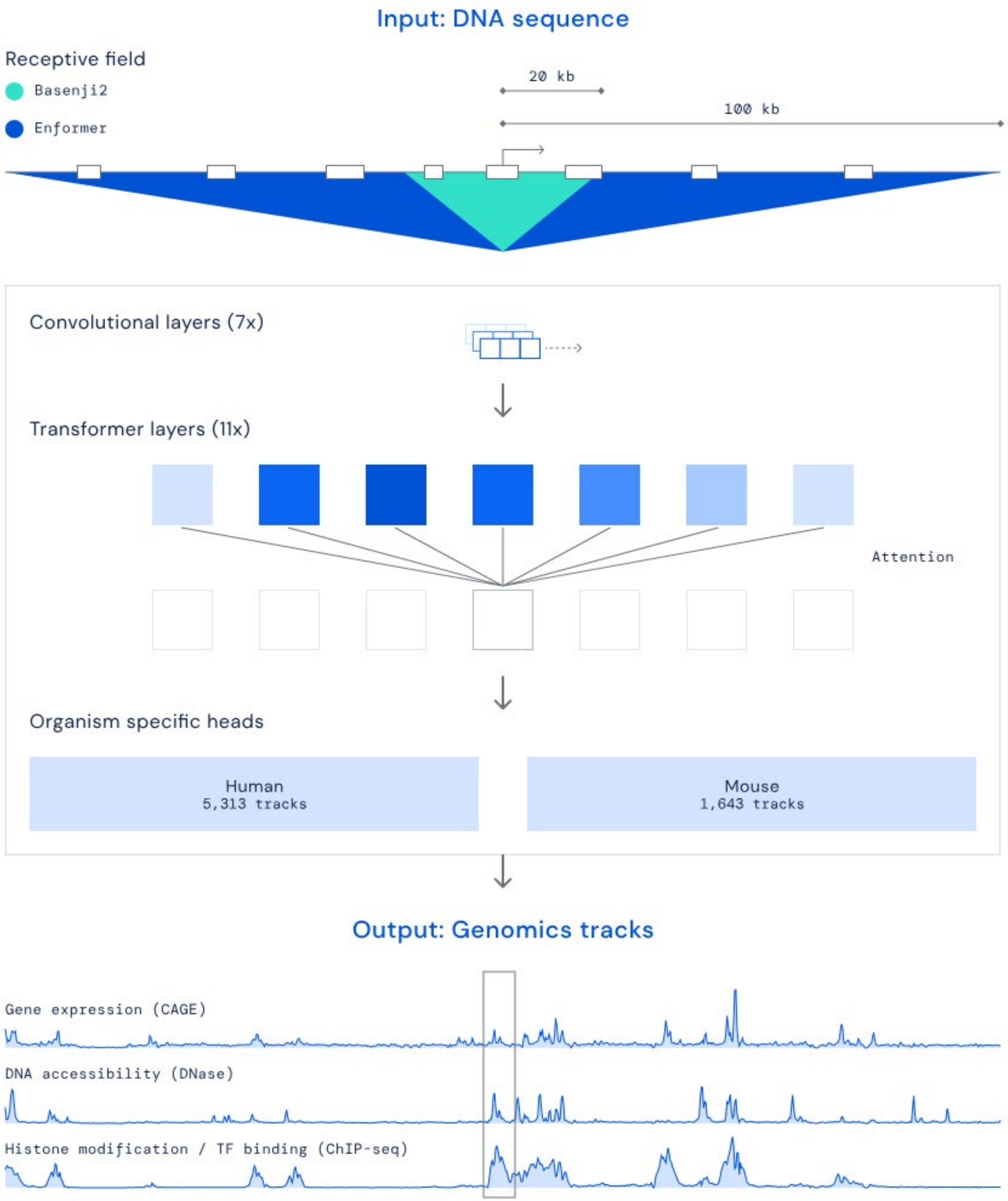
Train linear model to predict gene expression



Zhou et al., 2018 (https://www.nature.com/articles/s41588-018-0160-6)

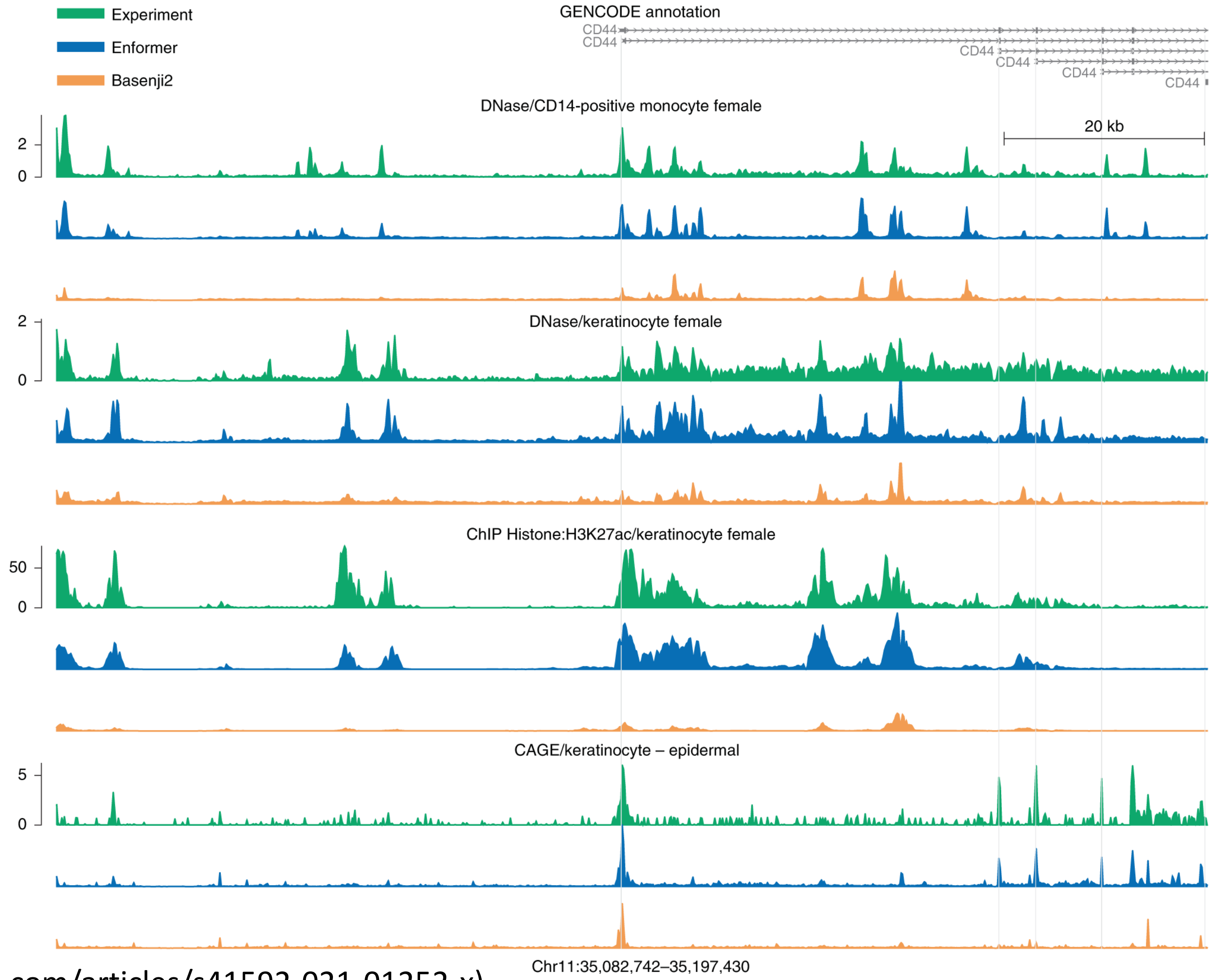
Enformer

- Predicts reads for 128bp bins
- Trained on data from human and mouse



Avsec et al., 2021 (https://www.nature.com/articles/s41592-021-01252-x)

Enformer

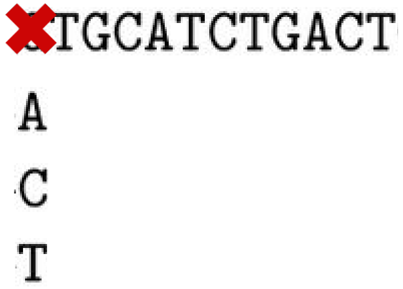


Avsec et al., 2021 (https://www.nature.com/articles/s41592-021-01252-x)

Outline

- Modeling the local sequence
    - Predicting TF binding sites
    - Predicting other genomic features
- Modeling long-range interactions
- Model interpretation

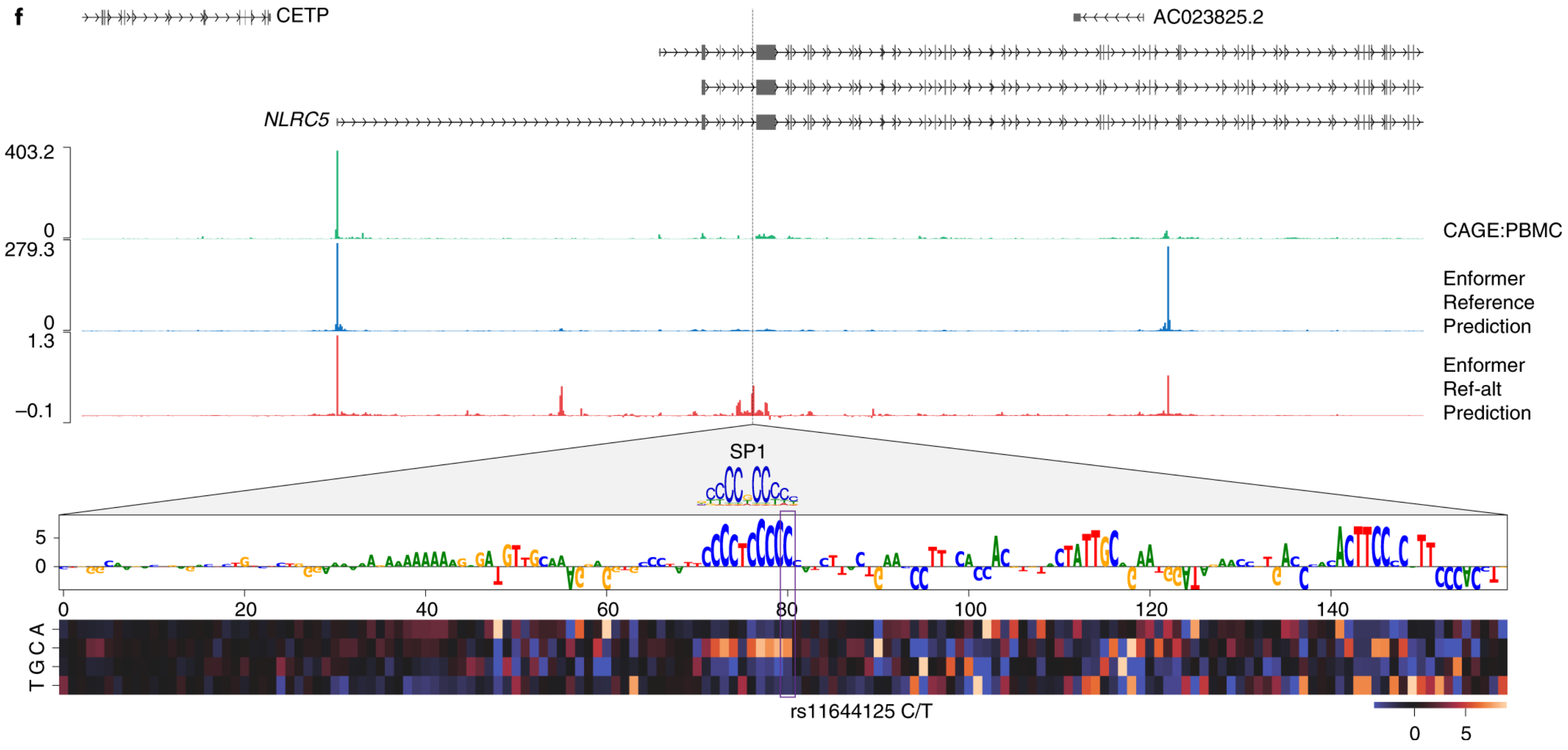
In-silico saturation mutagenesis



- ISM score:
  - $ISM_{g,i,n} = y_{pred,g,i,n} - \frac{1}{4}\sum_{m \in \{A,C,G,T\}} y_{pred,g,i,m}$
  - Gene g
  - Position i
  - Nucleotide n

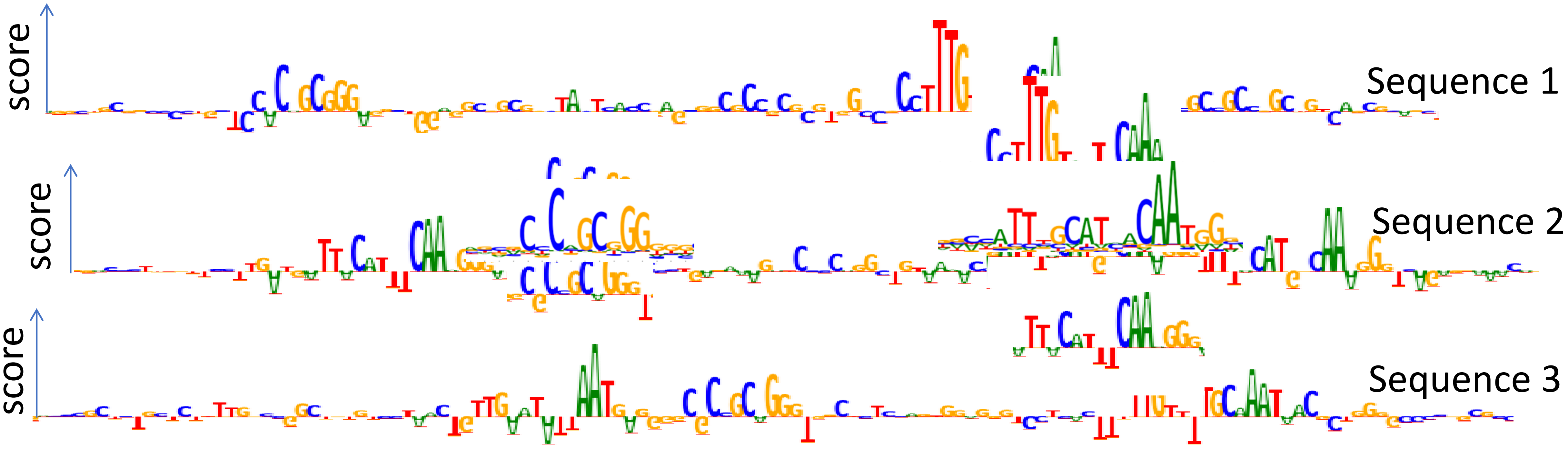
	G	A	C	T
y_{pred}	0.98			

In-silico saturation mutagenesis



Avsec et al., 2021 (https://www.nature.com/articles/s41592-021-01252-x)

TF-MoDISCo: TF Motif Discovery from Importance Scores



Anshul Kundaje (https://canvas.stanford.edu/courses/174437/files)

Drawbacks of discussed models

- Tissues are heterogeneous → ideally we need a model for every cell type
- How to make predictions across cell types?
    - Can be solved by adding DNase as input

- Personalized predictions
    - All models are trained on the reference genome only

- Learning the effect of distal enhancers

Summary

- Deep learning models can be used to predict genomic features directly from the DNA sequence

- Advantage of using these sequence-to-prediction models
  - Improving our general understanding of biological processes in a cell
  - Predicting the effect of (non-coding) variants

- Similar models exist for the RNA sequence (e.g. predicting alternative splicing)

Useful resources

- Course material
  - Coursera deep learning course (https://www.coursera.org/specializations/deep-learning)
  - Seq2expr course Stanford (https://canvas.stanford.edu/courses/174437/files)

- AI to genomics methods
  - DeepBind (https://www.nature.com/articles/nbt.3300)
  - Basset (https://genome.cshlp.org/content/26/7/990), basenji (https://genome.cshlp.org/content/28/5/739), enformer (https://www.nature.com/articles/s41592-021-01252-x), borzoi (https://www.biorxiv.org/content/10.1101/2023.08.30.555582v1)
  - DeepSEA (https://www.nature.com/articles/nmeth.3547), Sei (https://www.nature.com/articles/s41588-022-01102-2), ExPecto (https://www.nature.com/articles/s41588-018-0160-6), ExPectoSC (https://www.sciencedirect.com/science/article/pii/S2667237523002242?via%3Dihub)
  - TF-MoDisco (https://arxiv.org/abs/1811.00416)
  - BigRNA (https://www.biorxiv.org/content/10.1101/2023.09.20.558508v1)

- Benchmarking papers/discussing current limitations
  - Distal enhancers (https://genomebiology.biomedcentral.com/articles/10.1186/s13059-023-02899-9)
  - Personalized predictions (https://www.nature.com/articles/s41588-023-01574-w, https://www.nature.com/articles/s41588-023-01524-6)