Many transcription factors recognize DNA shape

Katie Pollard

Gladstone Institutes

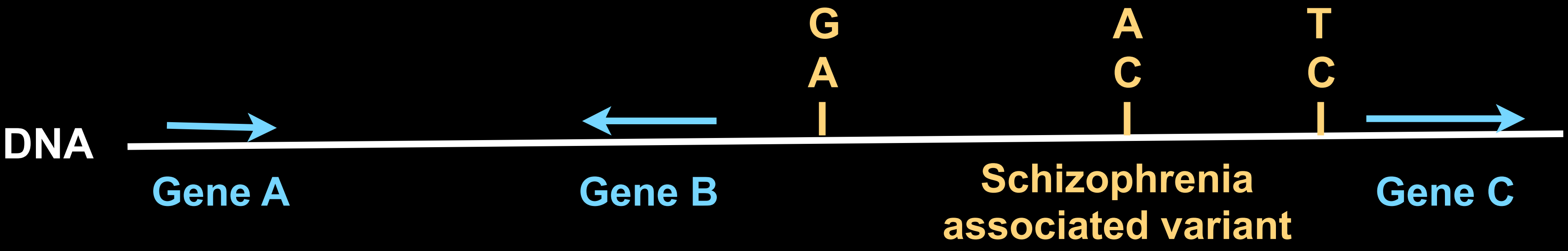
UCSF Division of Biostatistics, Institute for Human Genetics, and Institute for Computational Health Sciences

ENCODE Users Meeting - Stanford, CA

June 10, 2016

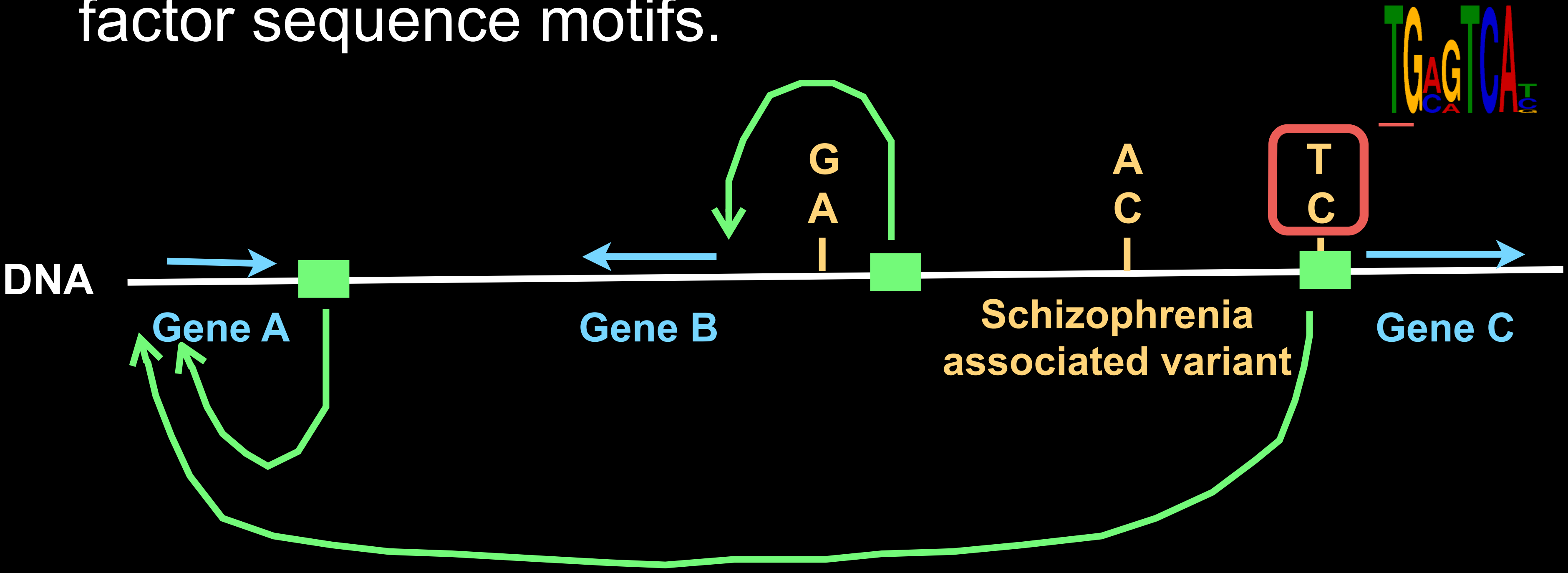
Most disease associated mutations are outside coding regions

Hypothesis 1: Non-coding variants alter transcription factor sequence motifs.



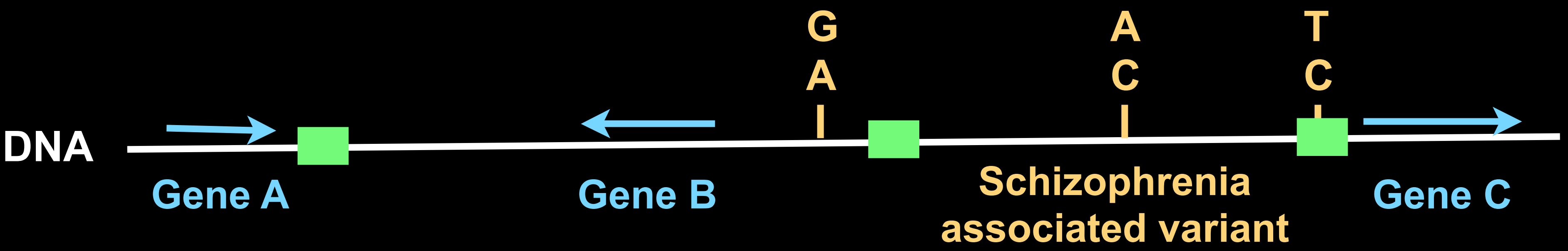
Most disease associated mutations are outside coding regions

Hypothesis 1: Non-coding variants alter transcription factor sequence motifs.



Approach: Map variants to correct pathways by predicting enhancers and their target genes. Score variants for changes in binding affinity.

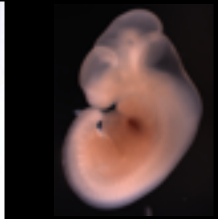
EnhancerFinder distinguishes biologically active enhancers



Training Data

VISTA Enhancers

Yes



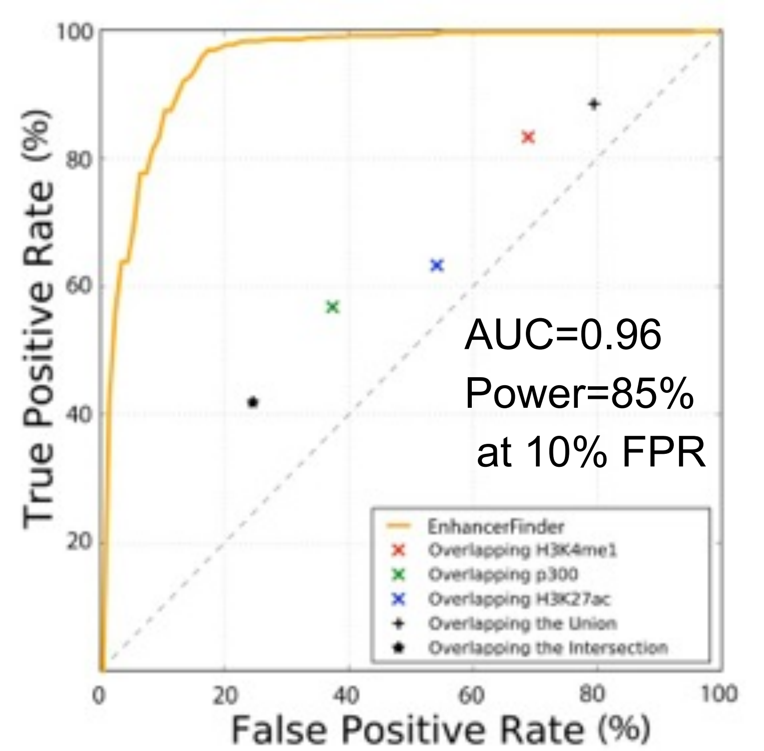
No

Functional Genomics

DNA Sequences

`AAAA,AAAC,AAAG,AAAT,..`

Performance on held out data

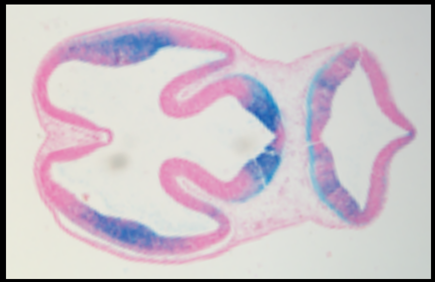


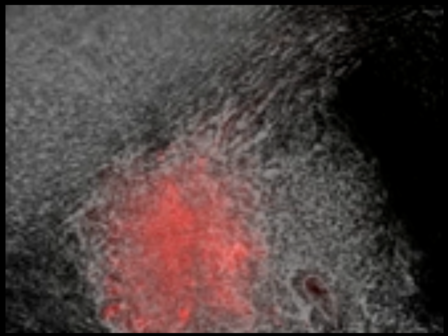
AUC=0.96

Power=85%

at 10% FPR

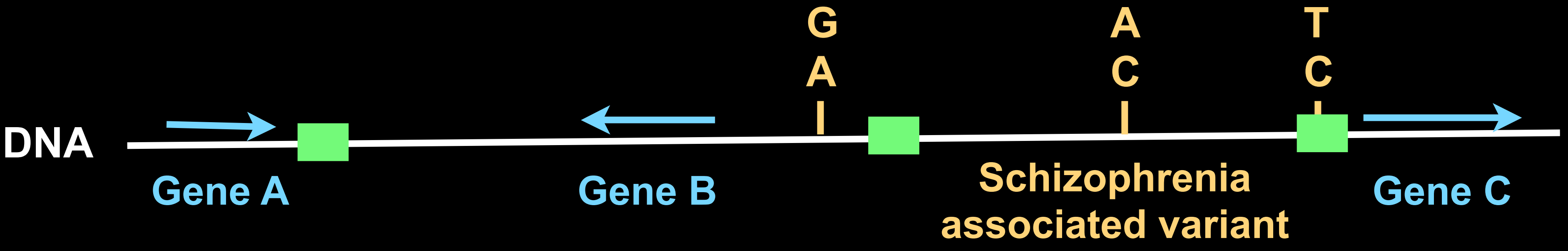
80% validate in vivo





Erwin et al. (2014)
Capra et al. (2014)

EnhancerFinder distinguishes biologically active enhancers



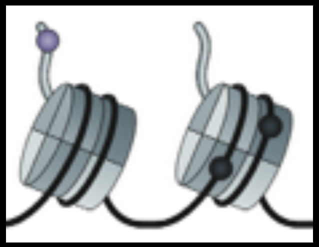
Training Data

VISTA Enhancers

Yes

No

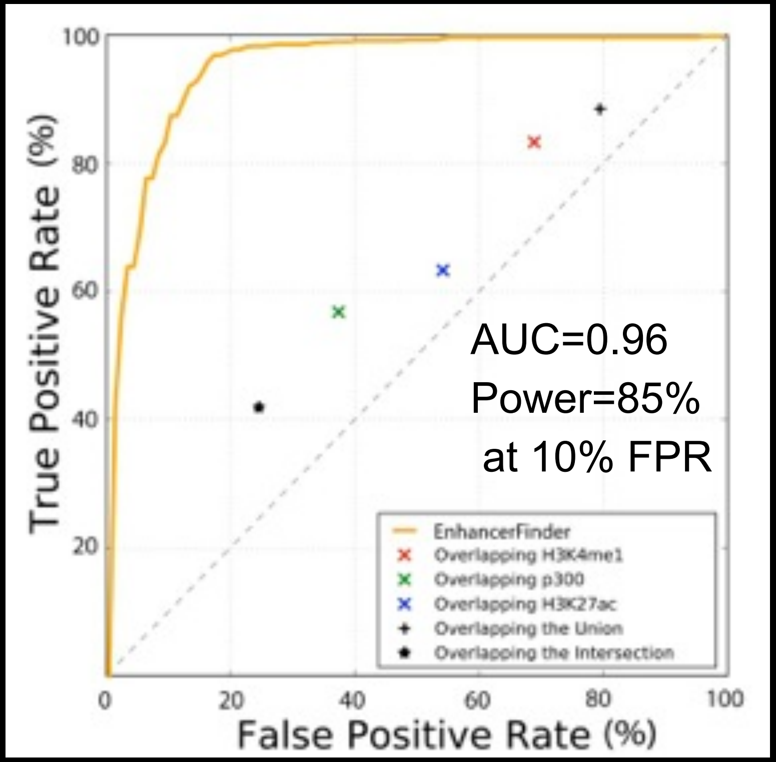
Functional Genomics



DNA Sequences

`AAAA,AAAC,AAAG,AAAT,..`

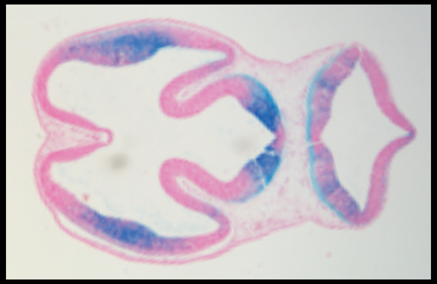
Performance on held out data

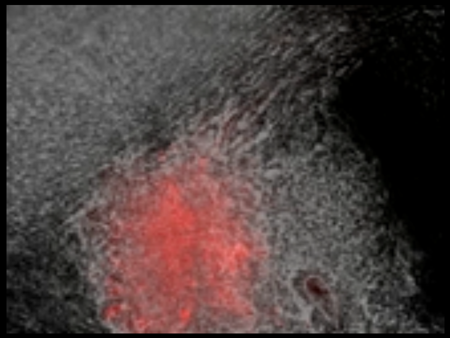


AUC=0.96

Power=85%
at 10% FPR

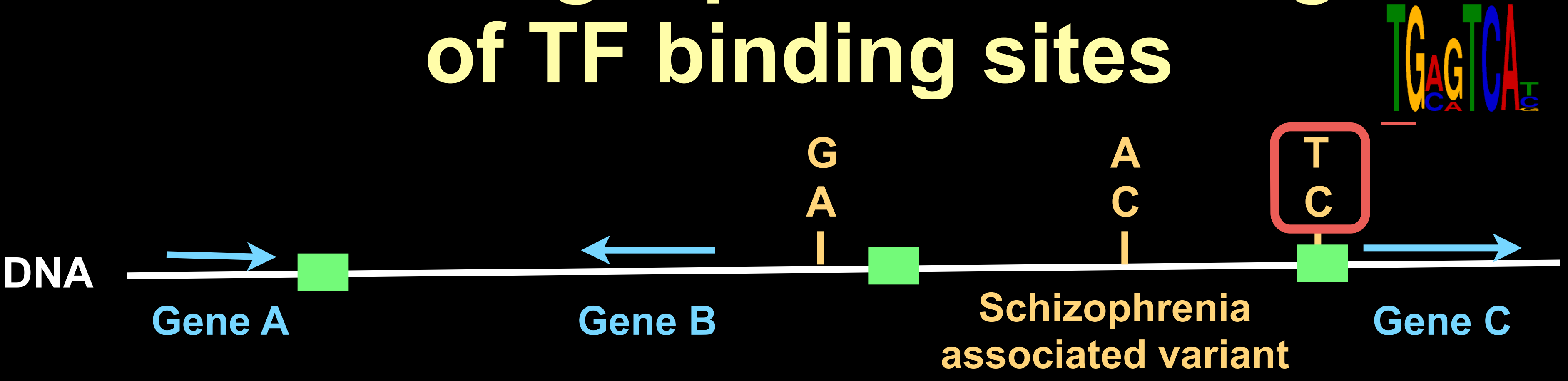
80% validate in vivo



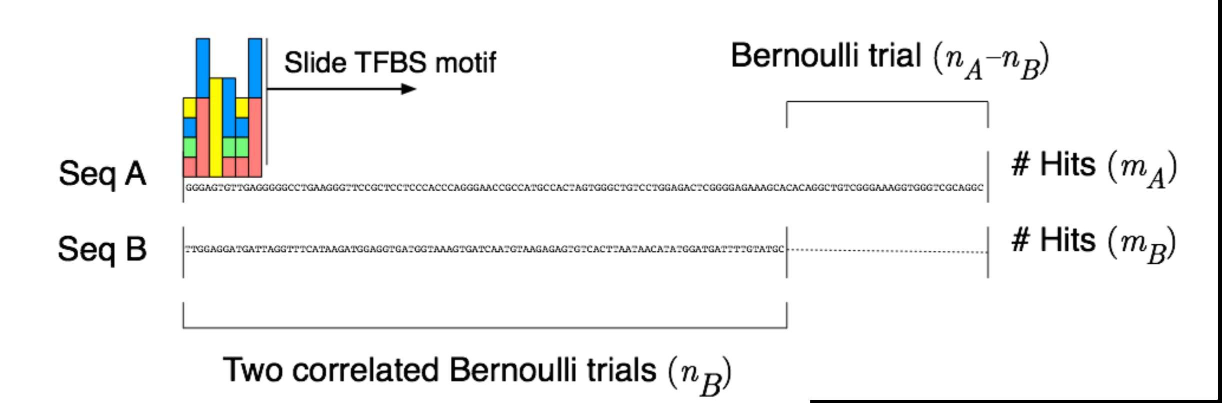


Erwin et al. (2014)
Capra et al. (2014)

MotifDiverge quantifies loss/gain of TF binding sites



Statistical model for TFBS evolution with turnover

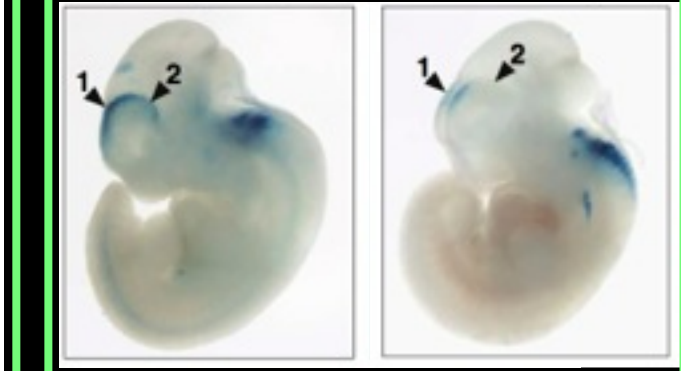

Slide TFBS motif
Bernoulli trial $(n_A - n_B)$
Seq A
GGGAGTGTTGAGGGGGCCTGAAGGGTTCCGCTCCTCCCACCCAGGGAACCGCCATGCCACTAGTGGGCTGTCCTGGAGACTCGGGGAGAAAGCACACAGGCTGTCGGGAAAGGTGGGTCGCAGGC
# Hits (m_A)
Seq B
TTGGAGGATGATTAGGTTTCATAAGATGGAGGTGATGGTAAAGTGATCAATGTAAGAGAGTGTCACTTAATAACATATGGATGATTTTGTATGC
# Hits (m_B)
Two correlated Bernoulli trials (n_B)

$$P(N_{xy} = n_{xy}) = \tag{6}$$
$$\begin{cases} \sum_{j=0}^{k_x - k_y} P_s(N_1 = n_{xy} - j) Bin(N_2 = j) & \text{for} \quad k_x \geq k_y \\ \sum_{j=0}^{k_y - k_x} P_s(N_1 = n_{xy} + j) Bin(N_2 = j) & \text{for} \quad k_x < k_y, \end{cases}$$

P-value for net change in binding

- One or many TFs
- Alignment-free
- Evolutionary model
- Motif specific

Predicts change of function

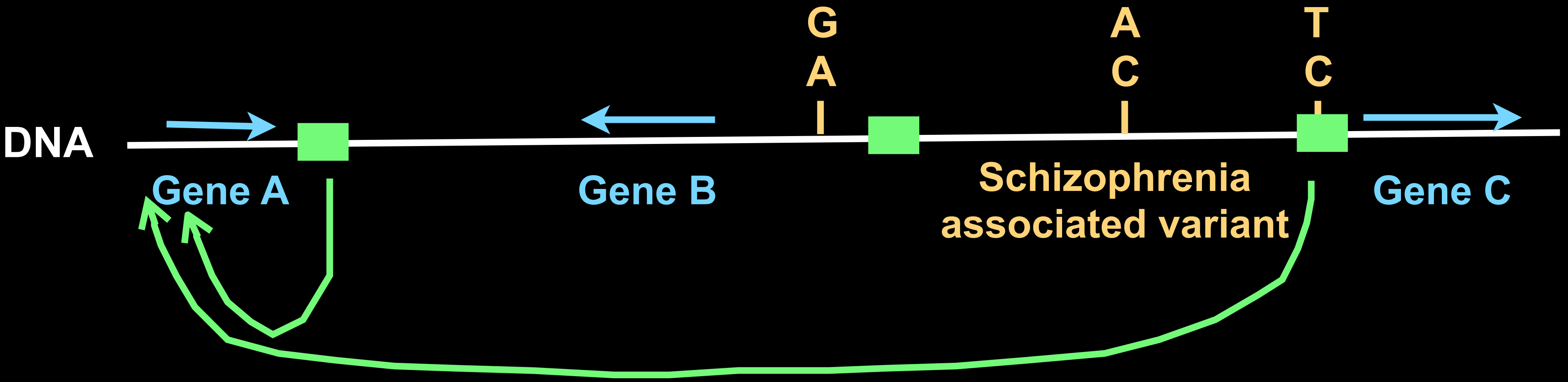


Detects loss/gain of function mutations with high accuracy

- Better than conservation scores
- In vivo and MPRAs in cell lines

Ritter et al. (2010)
Kostka et al. (2015)

TargetFinder maps distal regulatory elements to genes



DNA

Gene A

Gene B

G
A

A
C

T
C

Schizophrenia associated variant

Gene C

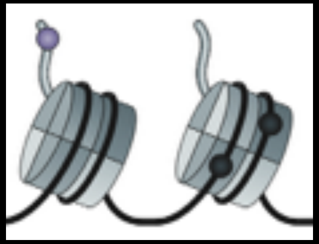
Training Data

Active enhancers

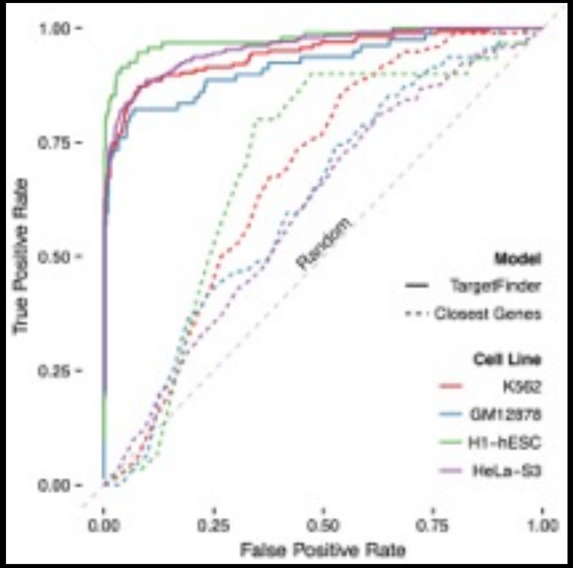
Expressed genes

Hi-C interactions

Functional Genomics



Closest gene usually wrong



Reveals distinct genomic signature of looping DNA

- Heterochromatin on loop
- Cohesin within 6Kb of enhancer and promoter but not mid-loop
- TFs bound with CTCF improve predictions

Whalen et al. (2016)

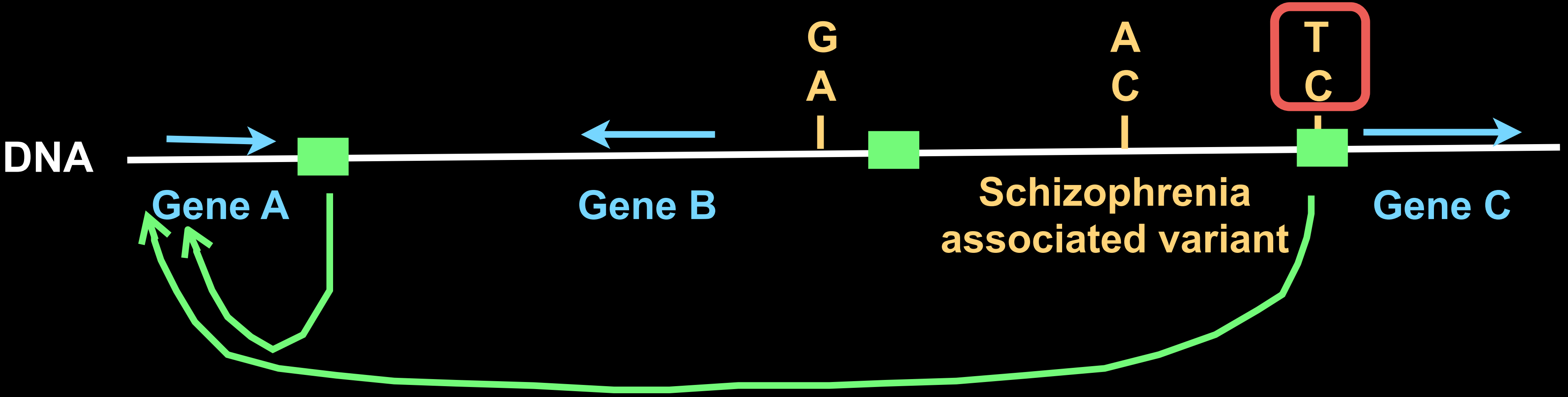
Summary and Challenges

- Machine-learning on biologically validated enhancers identifies non-coding variants most likely to affect gene regulation and the targeted genes.
  - Massive integration of functional genomics data enables cell type specific predictions
  - Many enhancer-like regions are minimally active and not consistently looping to a target gene

Summary and Challenges

- Machine-learning on biologically validated enhancers identifies non-coding variants most likely to affect gene regulation and the targeted genes.
  - Massive integration of functional genomics data enables cell type specific predictions
  - Many enhancer-like regions are minimally active and not consistently looping to a target gene
- But much remains to be explained…
  - Functional variants outside enhancers

TargetFinder maps distal regulatory elements to genes



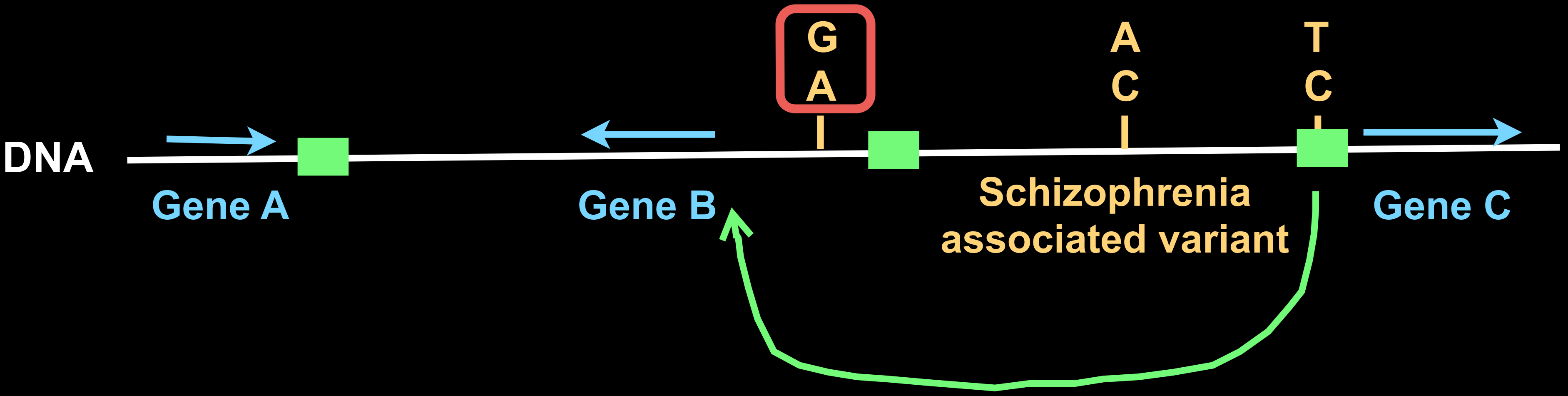
Hypothesis 2: Non-coding variants alter binding sites of structural proteins and chromatin modifiers.

Reveals distinct genomic signature of looping DNA

- Heterochromatin on loop
- Cohesin within 6Kb of enhancer and promoter but not mid-loop
- TFs bound with cohesin improve predictions

Whalen et al. (2016)

TargetFinder maps distal regulatory elements to genes



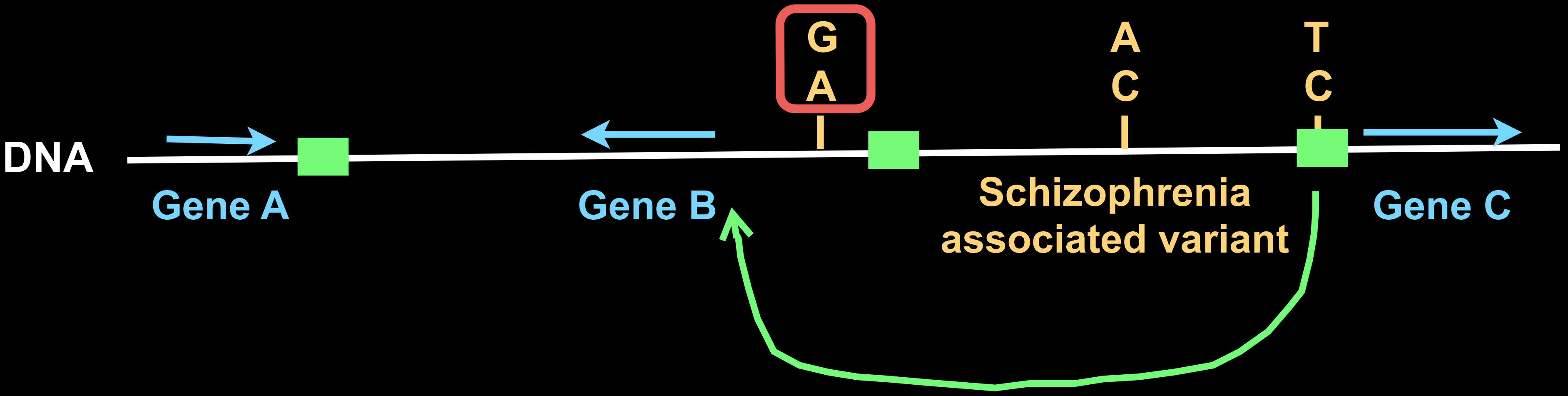
Hypothesis 2: Non-coding variants alter binding sites of structural proteins and chromatin modifiers.

Reveals distinct genomic signature of looping DNA

- Heterochromatin on loop
- Cohesin within 6Kb of enhancer and promoter but not mid-loop
- TFs bound with cohesin improve predictions

Whalen et al. (2016)

TargetFinder maps distal regulatory elements to genes



Hypothesis 2: Non-coding variants alter binding sites of structural proteins and chromatin modifiers.

Approach: CRISPR edit sites identified by TargetFinder, then test chromatin and expression.

Reveals distinct genomic signature of looping DNA

- Heterochromatin on loop
- Cohesin within 6Kb of enhancer and promoter but not mid-loop
- TFs bound with cohesin improve predictions

Whalen et al. (2016)

Summary and Challenges

- Machine-learning on biologically validated enhancers identifies non-coding variants most likely to affect gene regulation and the targeted genes.
  - Massive integration of functional genomics data enables cell type specific predictions
  - Many enhancer-like regions are minimally active and not consistently looping to a target gene
- But much remains to be explained…
  - Functional variants outside enhancers
  - Enhancer variants outside sequence motifs

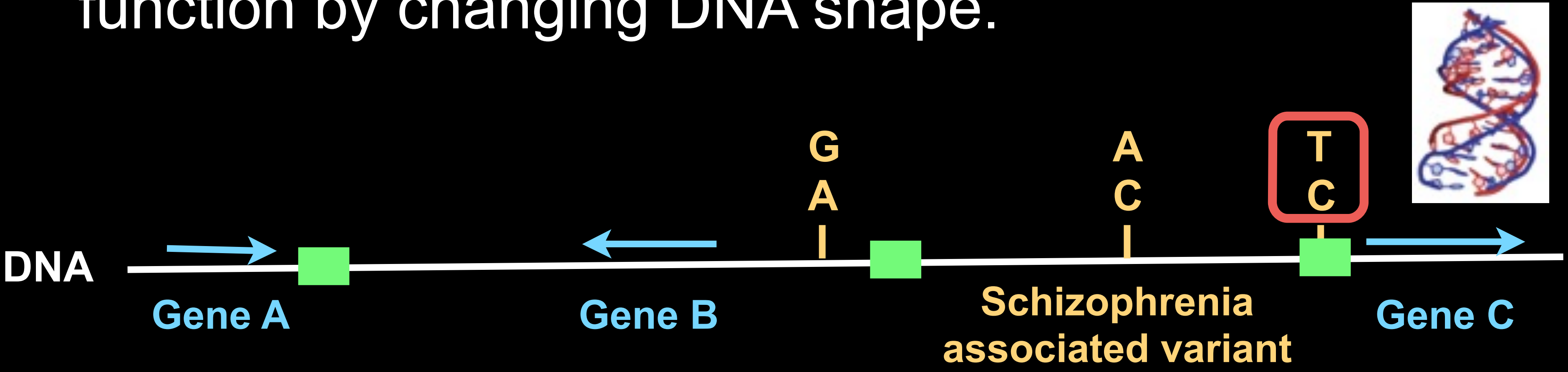
Summary and Challenges

- Machine-learning on biologically validated enhancers identifies non-coding variants most likely to affect gene regulation and the targeted genes.
  - Massive integration of functional genomics data enables cell type specific predictions
  - Many enhancer-like regions are minimally active and not consistently looping to a target gene
- But much remains to be explained…
  - Functional variants outside enhancers
  - Enhancer variants outside sequence motifs

For a typical ENCODE TF 23% of the top 2000 ChIP-seq peaks have no sequence motif (range = 1%-63%)

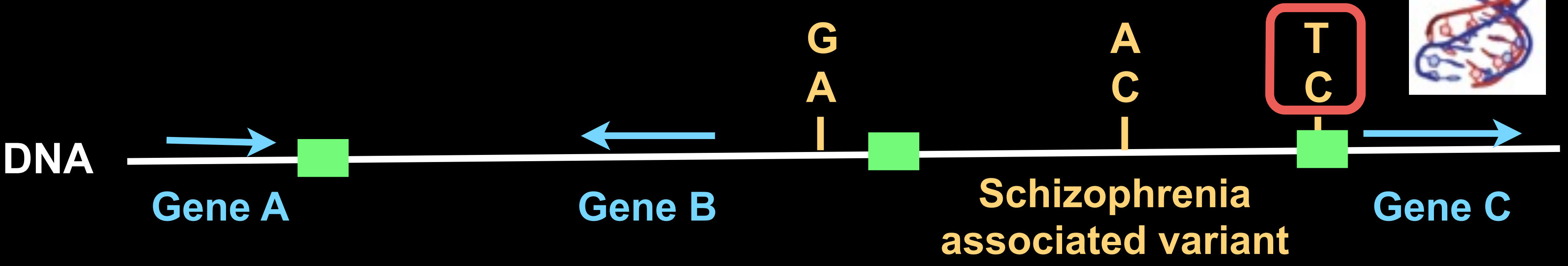
Many enhancer mutations are outside known or de novo sequence motifs

Hypothesis 3: Non-coding variants alter enhancer function by changing DNA shape.



Many enhancer mutations are outside known or de novo sequence motifs

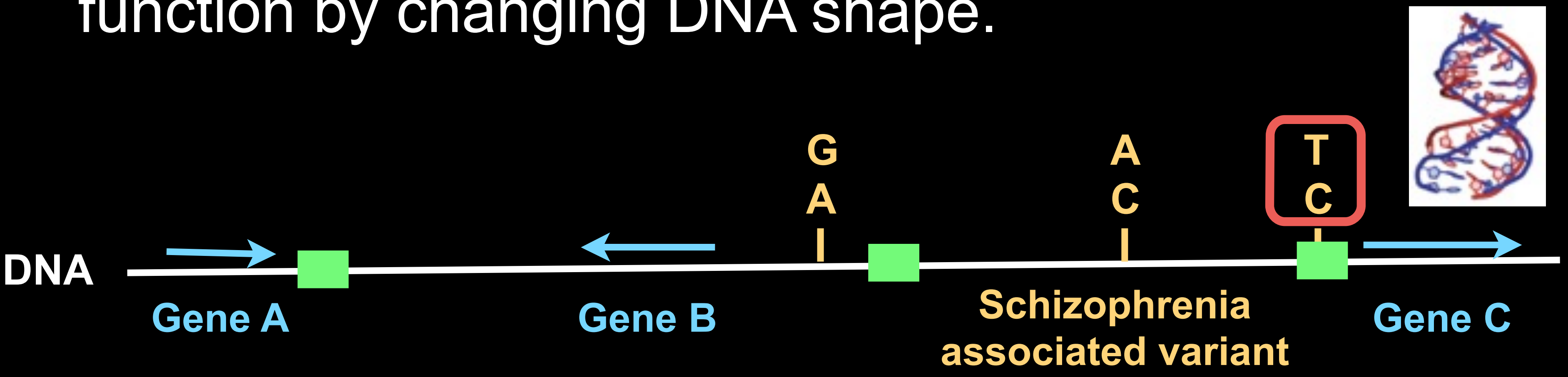
Hypothesis 3: Non-coding variants alter enhancer function by changing DNA shape.



- TFs can recognize shape in addition to sequence.
- DNA shape differentiates similar sequence motifs.
- Distinct sequences can encode same shape.

Many enhancer mutations are outside known or de novo sequence motifs

Hypothesis 3: Non-coding variants alter enhancer function by changing DNA shape.



DNA

Gene A

Gene B

G
A

A
C

T
C

Schizophrenia associated variant

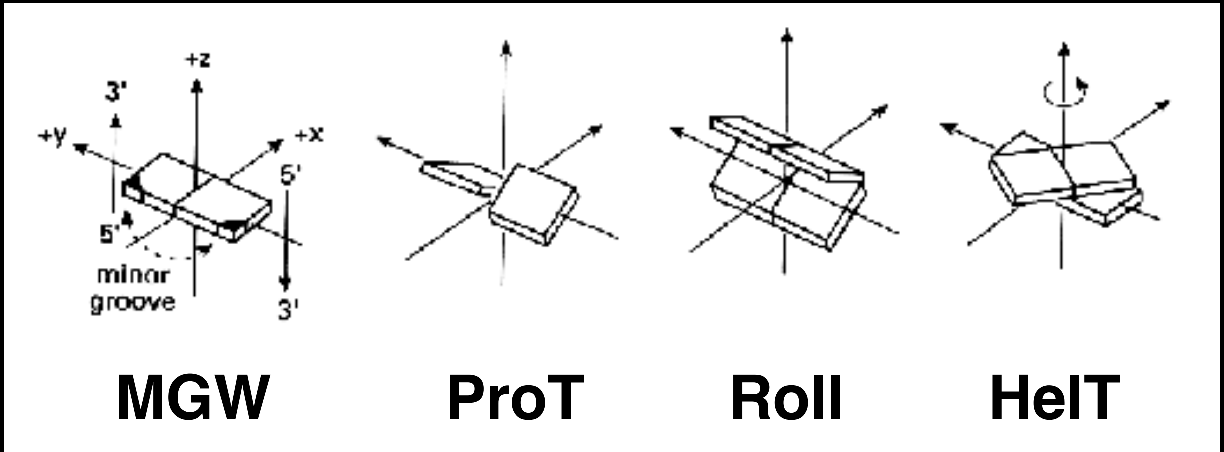
Gene C

- TFs can recognize shape in addition to sequence.
- DNA shape differentiates similar sequence motifs.
- Distinct sequences can encode same shape.

Approach: Algorithm to learn **shape motifs** de novo for all ENCODE TFs, predict shape motif hits in ChIP-seq peaks, compare to sequence motifs

De novo shape motif discovery

1. Estimate DNA structure: **`DNAshape`** (Zhou et al. 2013)
   - Maps 5-mer sequences to structural features.
   - Based on molecular dynamics simulations.



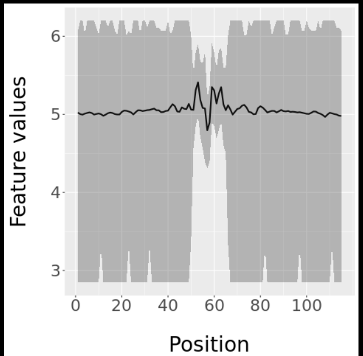
[Figures from Blackburn and Gait, 1996]

De novo shape motif discovery

1. Estimate DNA structure: **`DNAshape`** (Zhou et al. 2013)
   - Maps 5-mer sequences to structural features
   - Based on molecular dynamics simulations
2. Learn TF shape motifs: Search ChIP-seq peaks for windows with similar values of a shape feature.
   - Gibbs sampling with scores ~ $\exp(\sum D_{ij})$
   - Vary window size 5-25bp
   - Train on 1000 of top 2000 peaks for ~250 TFs

De novo shape motif discovery

1. Estimate DNA structure: **`DNAshape`** (Zhou et al. 2013)
   - Maps 5-mer sequences to structural features
   - Based on molecular dynamics simulations
2. Learn TF shape motifs: Search ChIP-seq peaks for windows with similar values of a shape feature.
   - Gibbs sampling with scores ~ exp(∑ D_{ij})
   - Vary window size 5-25bp
   - Train on 1000 of top 2000 peaks for ~250 TFs
3. Call hits: Scan ChIP-seq peaks with shape motifs.
   - Null distribution on distance from mean shape feature value at each position



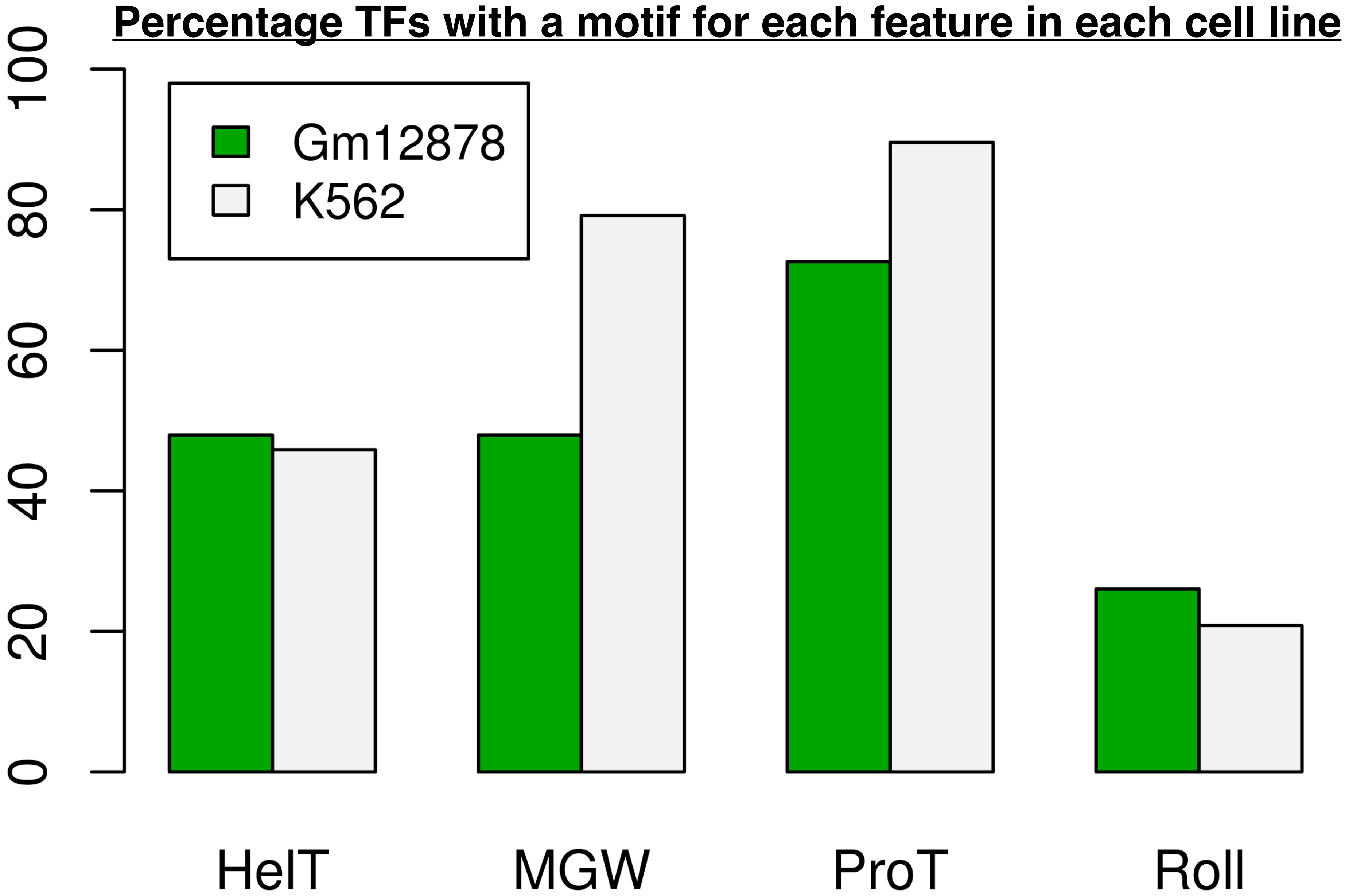
De novo shape motif discovery

1. Estimate DNA structure: **DNAshape** (Zhou et al. 2013)
   - Maps 5-mer sequences to structural features
   - Based on molecular dynamics simulations
2. Learn TF shape motifs: Search ChIP-seq peaks for windows with similar values of a shape feature.
   - Gibbs sampling with scores ~ $\exp(\sum D_{ij})$
   - Vary window size 5-25bp
   - Train on 1000 of top 2000 peaks for ~250 TFs
3. Call hits: Scan ChIP-seq peaks with shape motifs.
   - Null distribution on distance from mean shape feature value at each position
   - Apply to remaining 1000 of top 2000 peaks and flanking non-peak regions for each TF

De novo shape motif discovery

1. Estimate DNA structure: **`DNAshape`** (Zhou et al. 2013)
   - Maps 5-mer sequences to structural features
   - Based on molecular dynamics simulations
2. Learn TF shape motifs: Search ChIP-seq peaks for windows with similar values of a shape feature.
   - Gibbs sampling with scores ~ exp($\sum D_{ij}$)
   - Vary window size 5-25bp
   - Train on 1000 of top 2000 peaks for ~250 TFs
3. Call hits: Scan ChIP-seq peaks with shape motifs.
   - Null distribution on distance from mean shape feature value at each position
   - Apply to remaining 1000 of top 2000 peaks and flanking non-peak regions for each TF
4. Enrichment test: Hypergeometric p-value.

Shape motifs are common



Shape complements sequence motifs

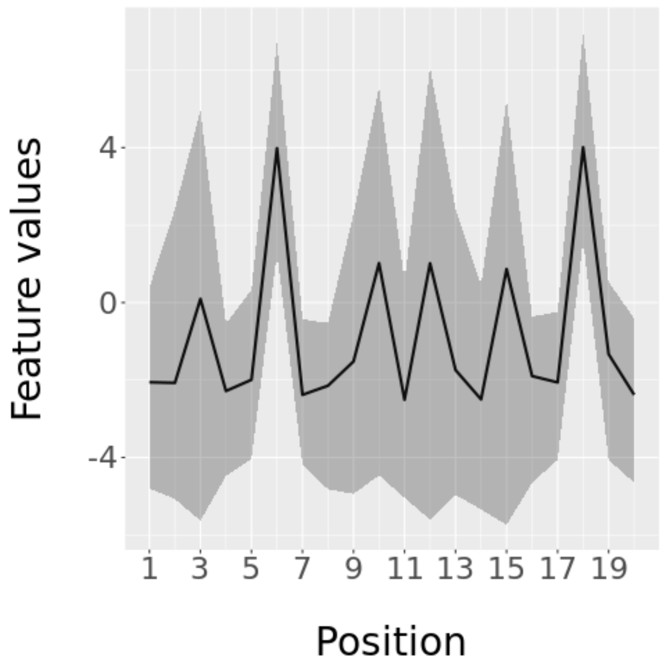
- Most peaks without sequence motifs have at least one shape motif. It is typically at the peak center.

Shape complements sequence motifs

- Most peaks without sequence motifs have at least one shape motif. It is typically at the peak center.
- ~25% of peaks have sequence and shape motifs.

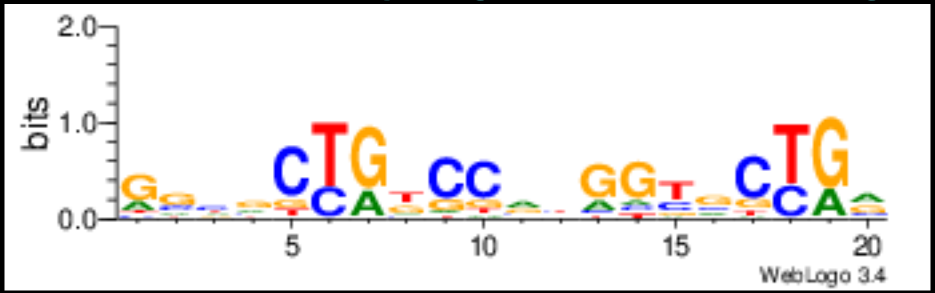
Shape complements sequence motifs

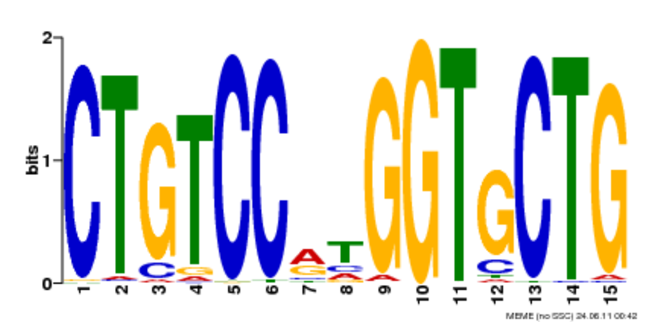
- Most peaks without sequence motifs have at least one shape motif. It is typically at the peak center.
- ~25% of peaks have sequence and shape motifs.
  - These can be similar,



Nrsf Roll motif in K562

Underlying sequence logo

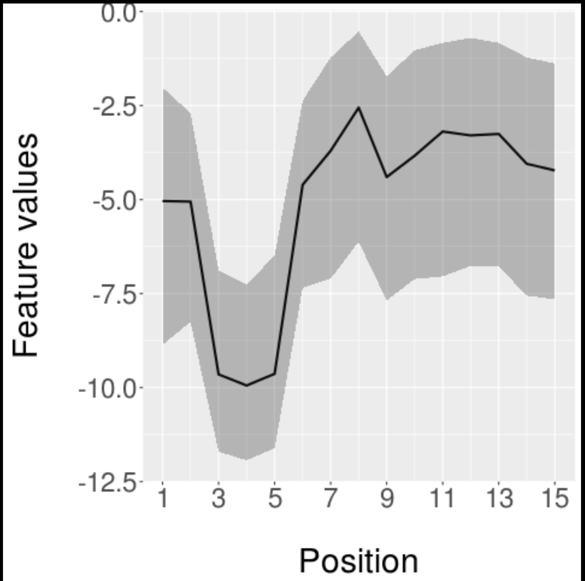




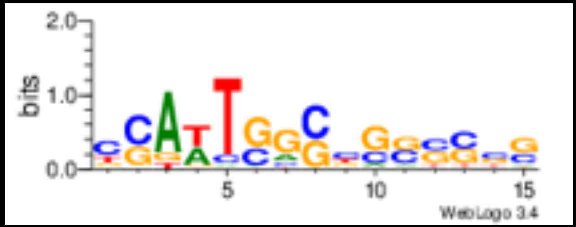
Nrsf FactorBook sequence motif

Shape motifs are complementary

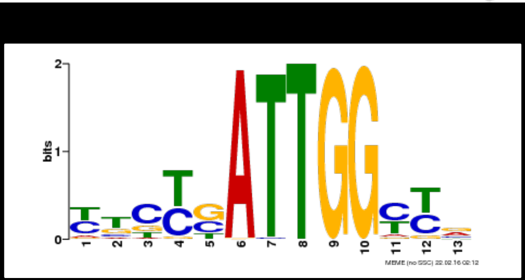
- Most peaks without sequence motifs have at least one shape motif. It is typically at the peak center.
- Many peaks have sequence and shape motifs.
  - These can be similar,
  - Extensions or refinements of one another,



Cfos ProT motif in K562



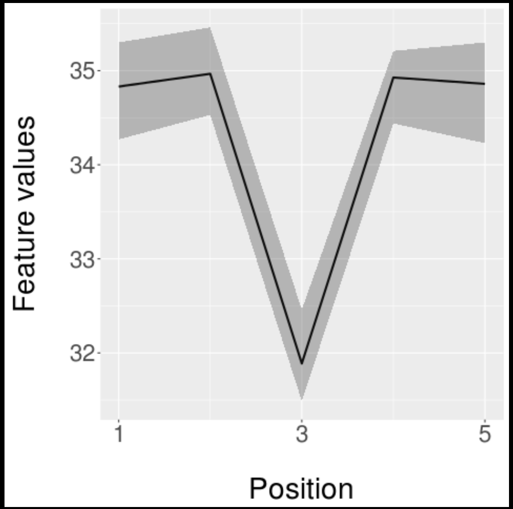
Underlying sequence



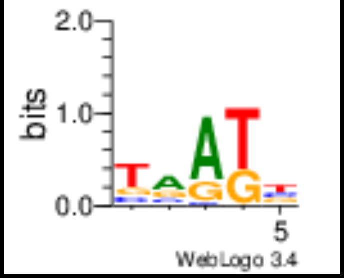
FactorBook sequence motif

Shape motifs are complementary

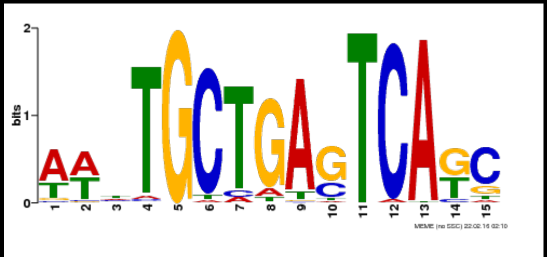
- Most peaks without sequence motifs have at least one shape motif. It is typically at the peak center.
- Many peaks have sequence and shape motifs.
  - These can be similar,
  - Extensions or refinements of one another,
  - Or very different



Maff HelT motif in K562



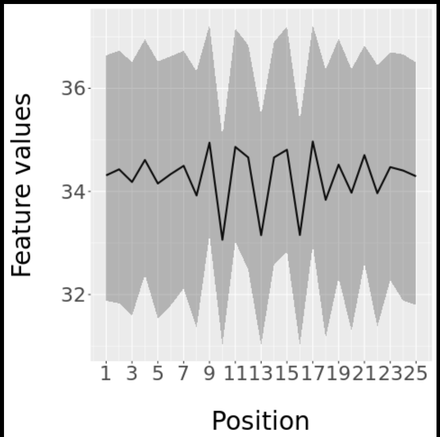
Underlying sequence



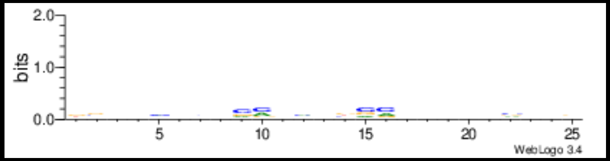
FactorBook sequence motif

Shape motifs are complementary

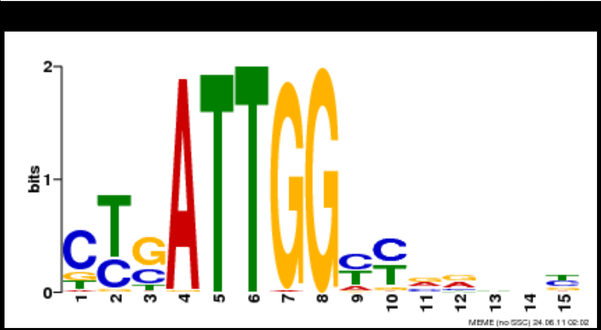
- Most peaks without sequence motifs have at least one shape motif. It is typically at the peak center.
- Many peaks have sequence and shape motifs.
  - These can be similar,
  - Extensions or refinements of one another,
  - Or very different
- Shape motifs can flank sequence motifs



Nfya HelT motif in K562



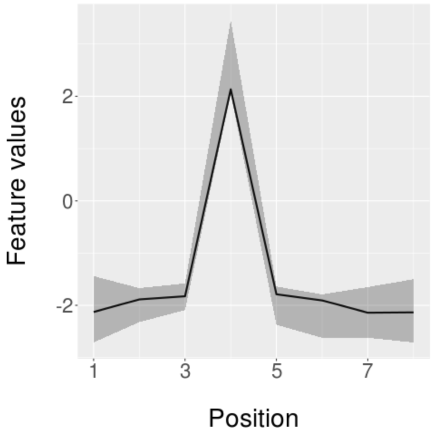
Underlying sequence



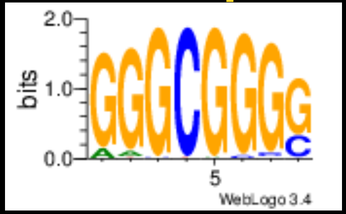
FactorBook sequence motif is 3bp upstream

Shape motifs are complementary

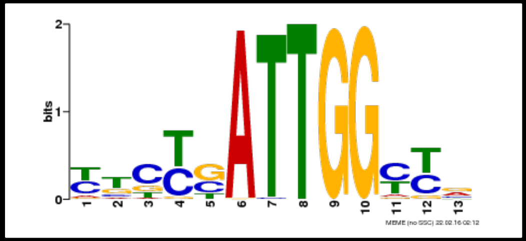
- Most peaks without sequence motifs have at least one shape motif. It is typically at the peak center.
- Many peaks have sequence and shape motifs.
  - These can be similar,
  - Extensions or refinements of one another,
  - Or very different
- Shape motifs can flank sequence motifs



Cfos Roll motif in K562



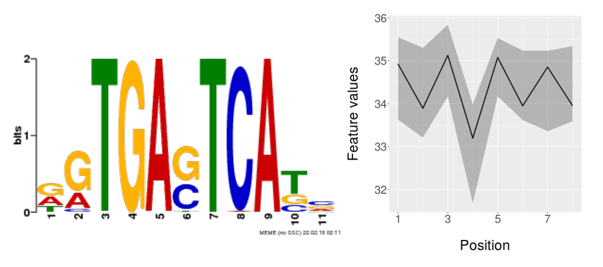
Underlying sequence



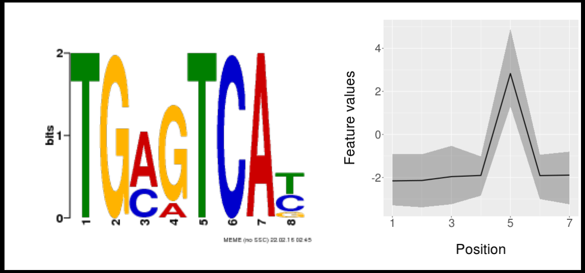
FactorBook sequence motif is 30bp away

Shape motifs are complementary

- Most peaks without sequence motifs have at least one shape motif. It is typically at the peak center.
- Many peaks have sequence and shape motifs.
  - These can be similar,
  - Extensions or refinements of one another,
  - Or very different
- Shape motifs can flank sequence motifs
- Shape motifs can differ between TFs with similar sequence motifs and/or the same protein fold.



Fosl1 has a HelT motif



Atf3 has a Roll motif

Ongoing Work

- Hierarchical or mixture model of TF binding with sequence and shape motifs
  - Decompose sequence motifs by shape types
  - Spectrum of recognition modes

Ongoing Work

- Hierarchical or mixture model of TF binding with sequence and shape motifs
  - Decompose sequence motifs by shape types
  - Spectrum of recognition modes
- Shape motifs in different contexts
  - Co-factors and complexes
  - Weak ChIP-seq peaks

Ongoing Work

- Hierarchical or mixture model of TF binding with sequence and shape motifs
  - Decompose sequence motifs by shape types
  - Spectrum of recognition modes
- Shape motifs in different contexts
  - Co-factors and complexes
  - Weak ChIP-seq peaks
- Role of shape in ectopic binding of TFs when co-factors are absent [Luna-Zurita et al. 2016]

Ongoing Work

- Hierarchical or mixture model of TF binding with sequence and shape motifs
  - Decompose sequence motifs by shape types
  - Spectrum of recognition modes
- Shape motifs in different contexts
  - Co-factors and complexes
  - Weak ChIP-seq peaks
- Role of shape in ectopic binding of TFs when co-factors are absent [Luna-Zurita et al. 2016]
- Evolutionary modeling of DNA shape
  - Conservation of shape without sequence
  - Scoring SNPs for effects on shape motifs

Collaborators

EnhancerFinder
Tony Capra
Gen Haliburton

DNA Shape
Hassan Samee

TargetFinder
Rebecca Truty
Sean Whalen

MotifDiverge
Dennis Kostka

Functional Assays
Hane Ryu
Alex Pollen
Nadav Ahituv
Arnold Kriegstein

Funding from NIMH, NIGMS, NHLBI, PhRMA Foundation, Gladstone Institutes