Inference of 3D regulatory interactions from 2D genomic data

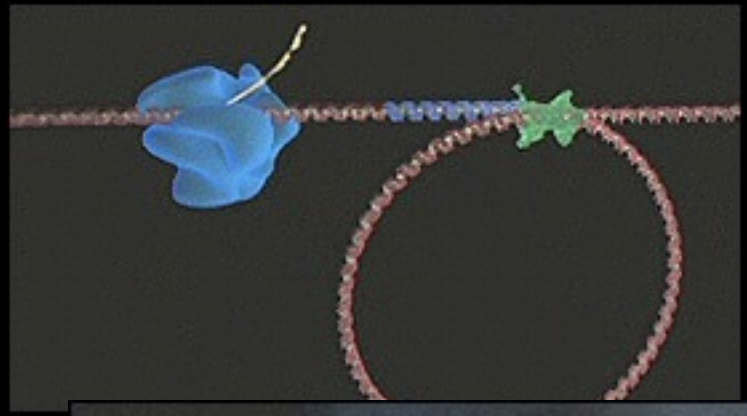
Katie Pollard

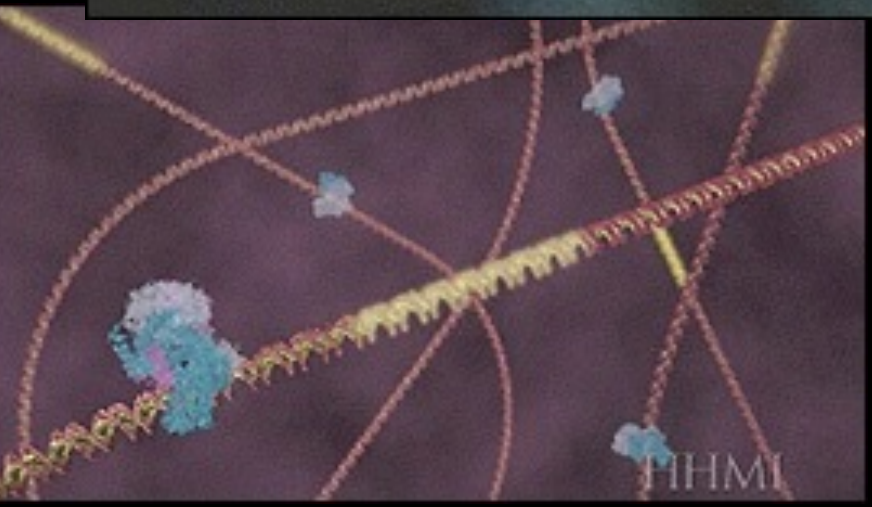
**Gladstone Institutes, Institute for Human Genetics,
Division of Biostatistics - UCSF**

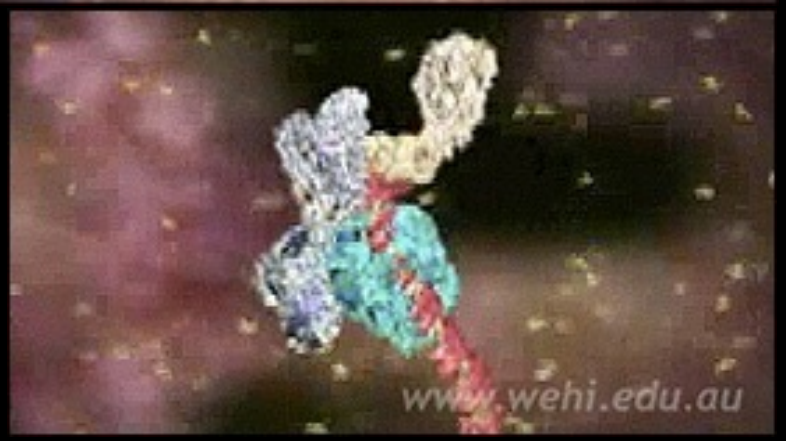
ENCODE Users Meeting

Bolger Center - June 30, 2015

Eukaryotic gene regulation is 3D and complex







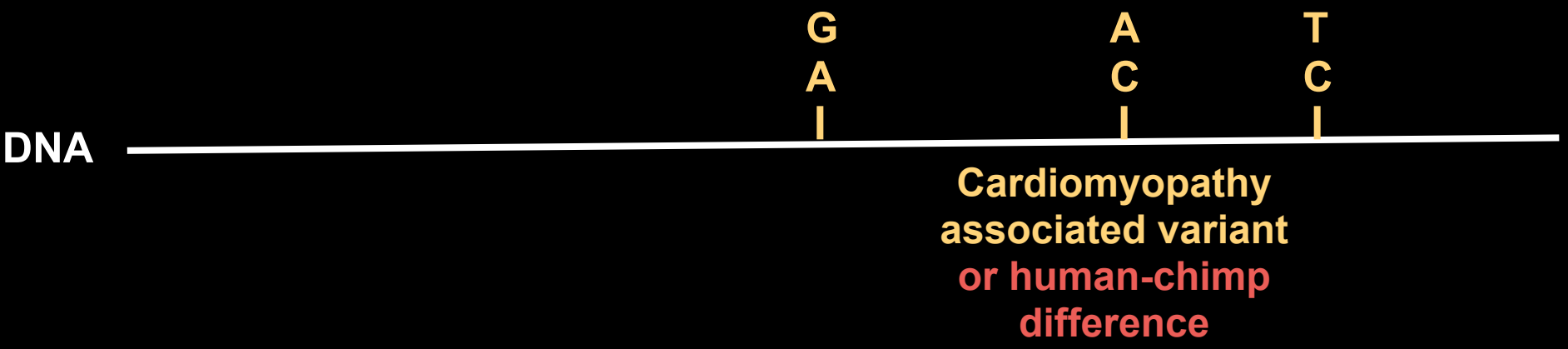
Eric Keller

Drew Berry

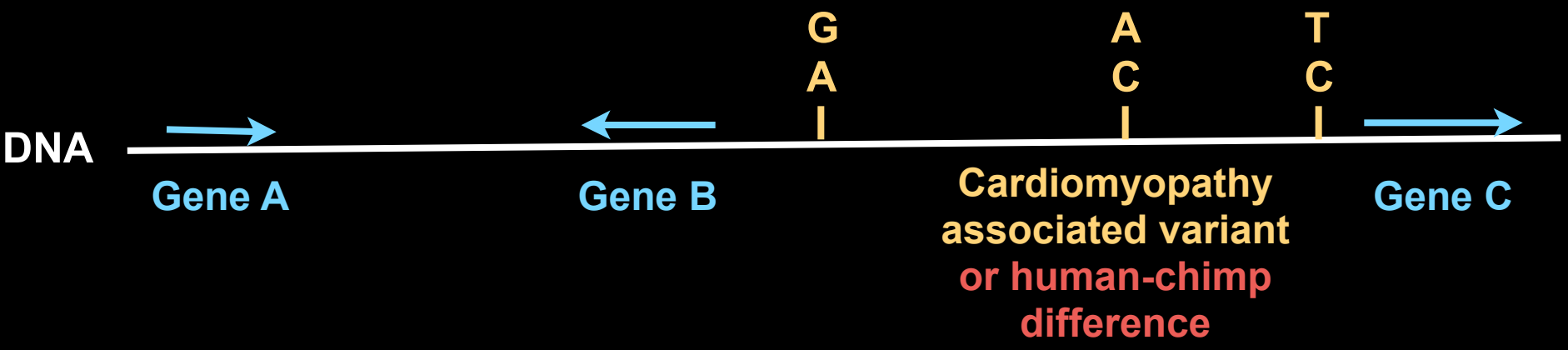
Most phenotype associated mutations are outside coding regions

Most phenotype associated mutations are outside coding regions

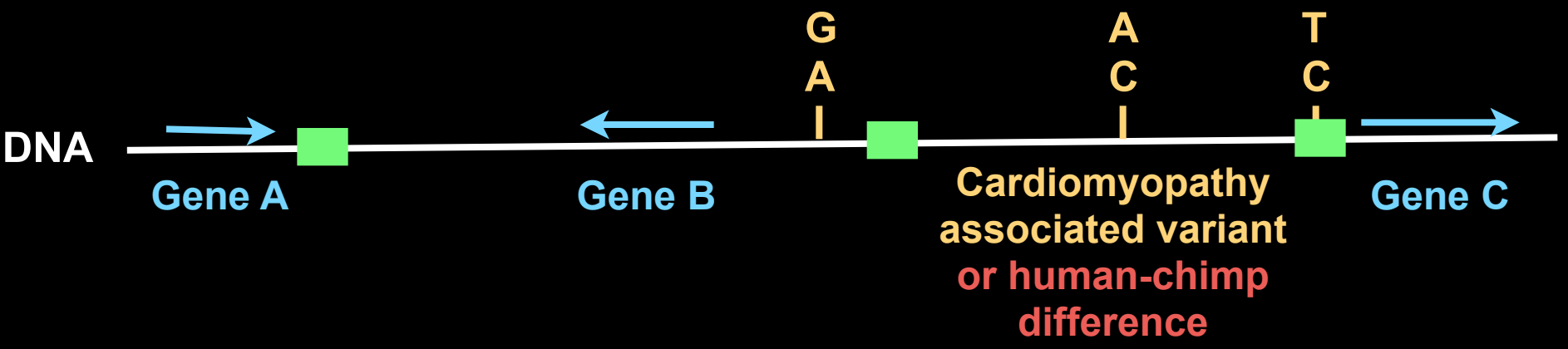
Most phenotype associated mutations are outside coding regions



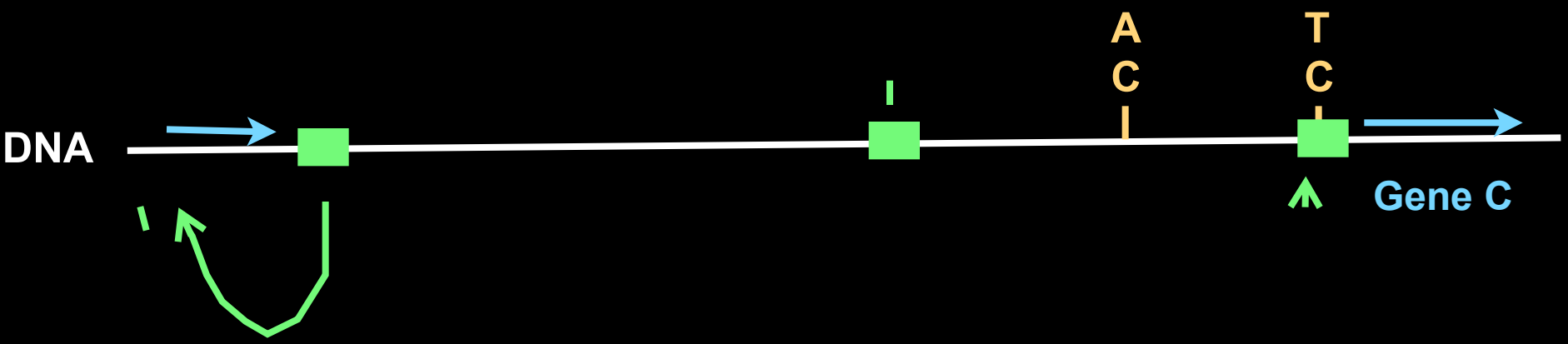
Most phenotype associated mutations are outside coding regions



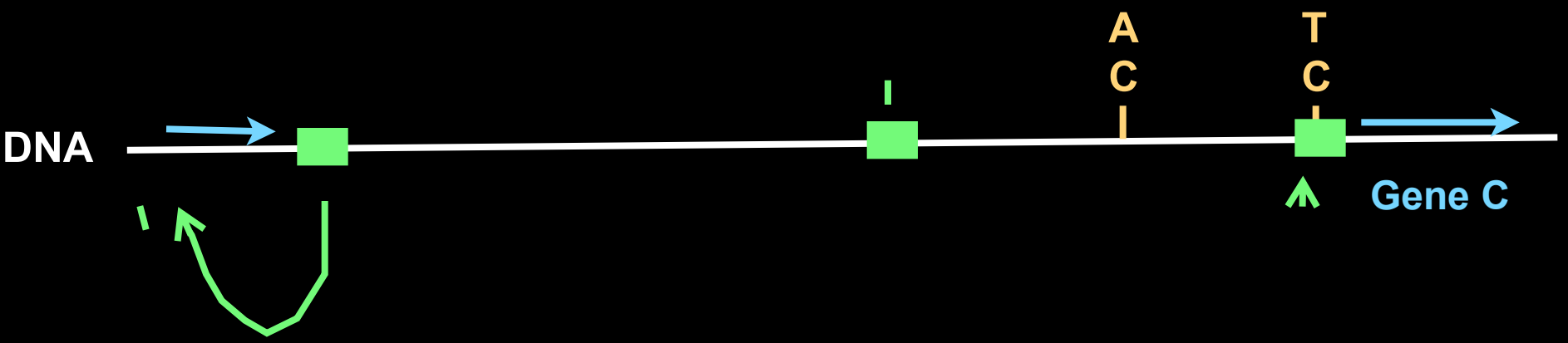
Most phenotype associated mutations are outside coding regions



Most phenotype associated mutations are outside coding regions

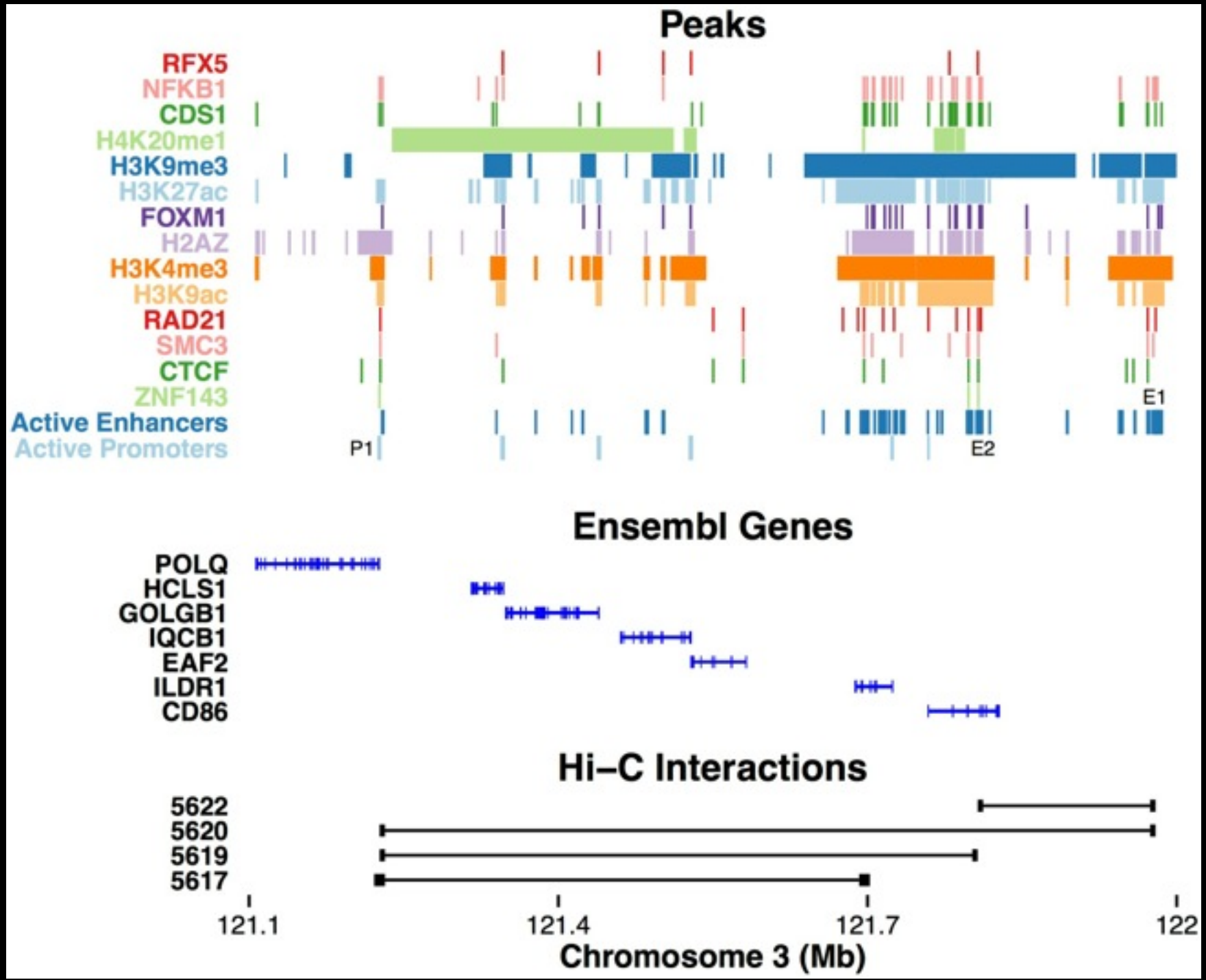


Most phenotype associated mutations are outside coding regions



- **Enhancers:** individual ChIP-seq data sets identify <50% of known enhancers, plus many false positives
- **Gene targets:** closest gene is right ~10% of time

Can we reconstruct 3D interactions between enhancers and promoters from 2D genomic data?


Peaks
RFX5
NFKB1
CDS1
H4K20me1
H3K9me3
H3K27ac
FOXM1
H2AZ
H3K4me3
H3K9ac
RAD21
SMC3
CTCF
ZNF143
Active Enhancers
Active Promoters
P1
E1
E2
Ensembl Genes
POLQ
HCLS1
GOLGB1
IQCB1
EAF2
ILDR1
CD86
Hi–C Interactions
5622
5620
5619
5617
121.1
121.4
121.7
122
Chromosome 3 (Mb)

TargetFinder: Training

Teach a machine learning algorithm to discriminate true versus false enhancer-promoter interactions based on their features.

Training Data

Active enhancer expressed gene

Positives = Hi-C +

Negatives = Hi-C -

Rao et al 2014, 1-Kb resolution

TargetFinder: Training

Teach a machine learning algorithm to discriminate true versus false enhancer-promoter interactions based on their features.

Training Data

Active enhancer expressed gene

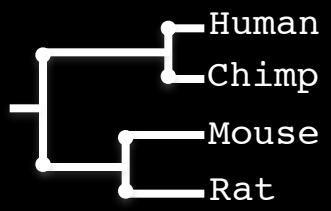
Positives = Hi-C +

Negatives = Hi-C -

Rao et al 2014, 1-Kb resolution

Features

Evolutionary Conservation



Conserved synteny of enhancer and promoter

TargetFinder: Training

Teach a machine learning algorithm to discriminate true versus false enhancer-promoter interactions based on their features.

Training Data

Active enhancer expressed gene

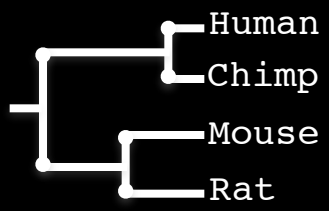
Positives = Hi-C +

Negatives = Hi-C -

Rao et al 2014, 1-Kb resolution

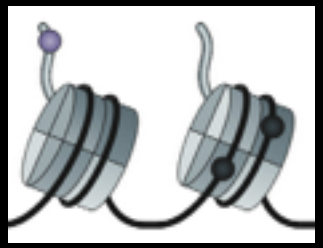
Features

Evolutionary Conservation



Conserved synteny of enhancer and promoter

Functional Genomics



ChIA-PET

RNA-seq

ChIP-seq (TFs, histones)

enhancer/promoter/window

ChromHMM & Segway

TargetFinder: Training

Teach a machine learning algorithm to discriminate true versus false enhancer-promoter interactions based on their features.

Training Data

Active enhancer expressed gene

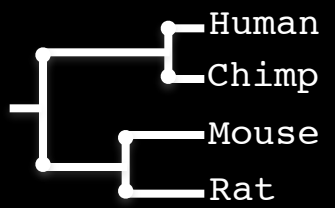
Positives = Hi-C +

Negatives = Hi-C -

Rao et al 2014, 1-Kb resolution

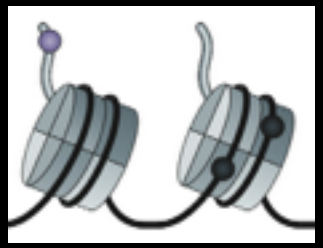
Features

Evolutionary Conservation



Conserved synteny of enhancer and promoter

Functional Genomics



ChIA-PET
RNA-seq
ChIP-seq (TFs, histones)
enhancer/promoter/window
ChromHMM & Segway

Sequence Annotations

```
AAAA,AAAC,AAAG,AAAT,
AACA,AACC,AACG,AACT,
AAGA,AAGC,AAGG,AAGT,
AATA,AATC,AATG,AATT,
ACAA,ACAC,ACAG,ACAT,
...
```

K-mer correlation
Annotated functions and pathways of gene and enhancer bound TFs

TargetFinder: Training

Teach a machine learning algorithm to discriminate true versus false enhancer-promoter interactions based on their features.

Training Data

Active enhancer expressed gene

Positives = Hi-C +

Negatives = Hi-C -

Rao et al 2014, 1-Kb resolution

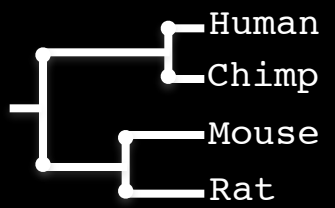
Computational Algorithm

Decision trees: good for interacting features

Ensemble learning: build many imperfect classifiers and combine them to improve prediction accuracy

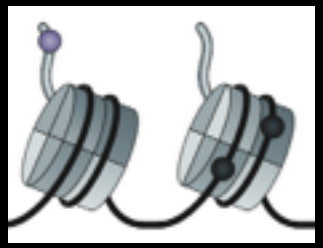
Features

Evolutionary Conservation



Conserved synteny of enhancer and promoter

Functional Genomics



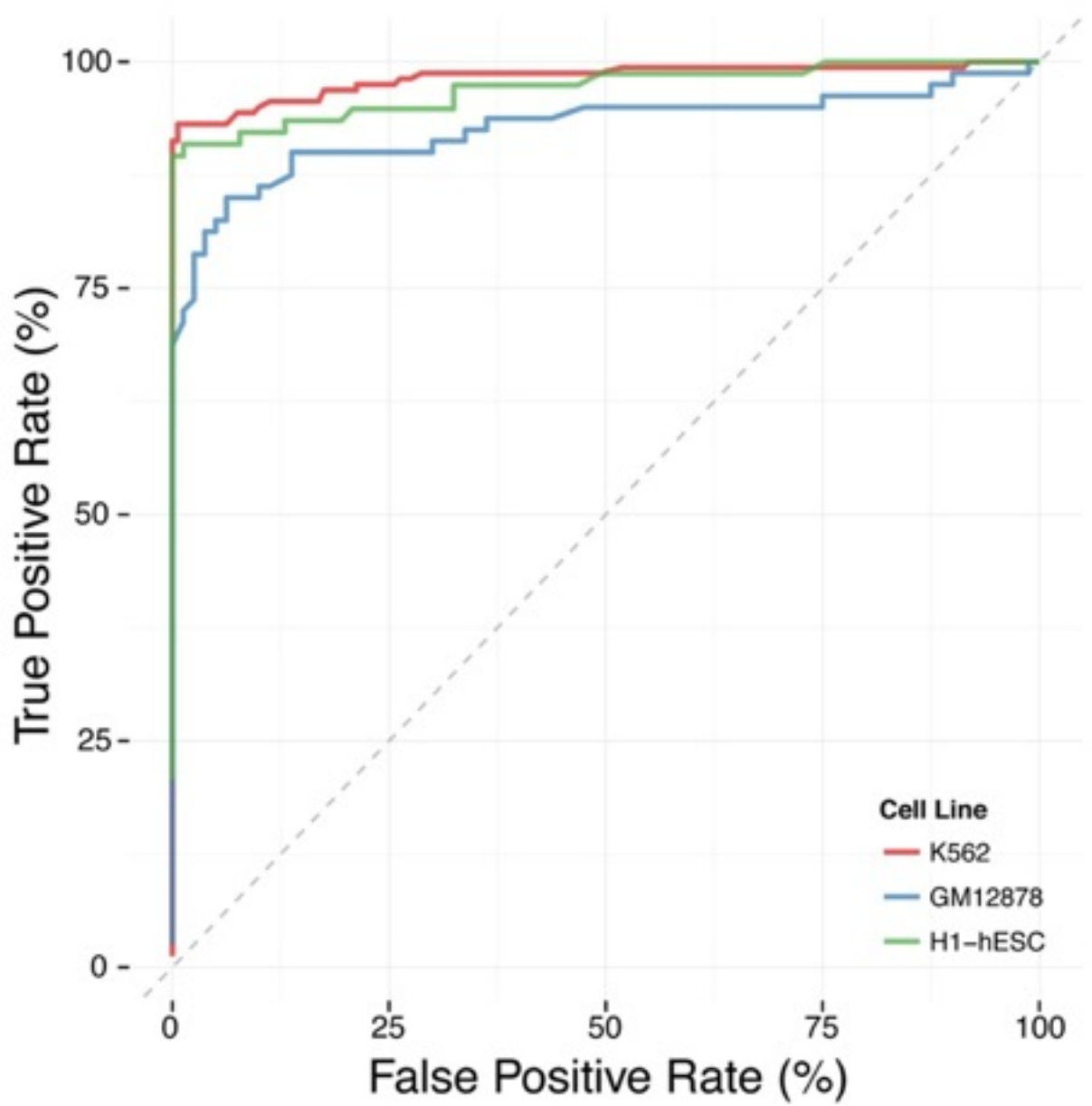
ChIA-PET
RNA-seq
ChIP-seq (TFs, histones)
enhancer/promoter/window
ChromHMM & Segway

Sequence Annotations

```
AAAA,AAAC,AAAG,AAAT,
AACA,AACC,AACG,AACT,
AAGA,AAGC,AAGG,AAGT,
AATA,AATC,AATG,AATT,
ACAA,ACAC,ACAG,ACAT,
...
```

K-mer correlation
Annotated functions and pathways of gene and enhancer bound TFs

TargetFinder: Performance



AUC=0.94-0.96

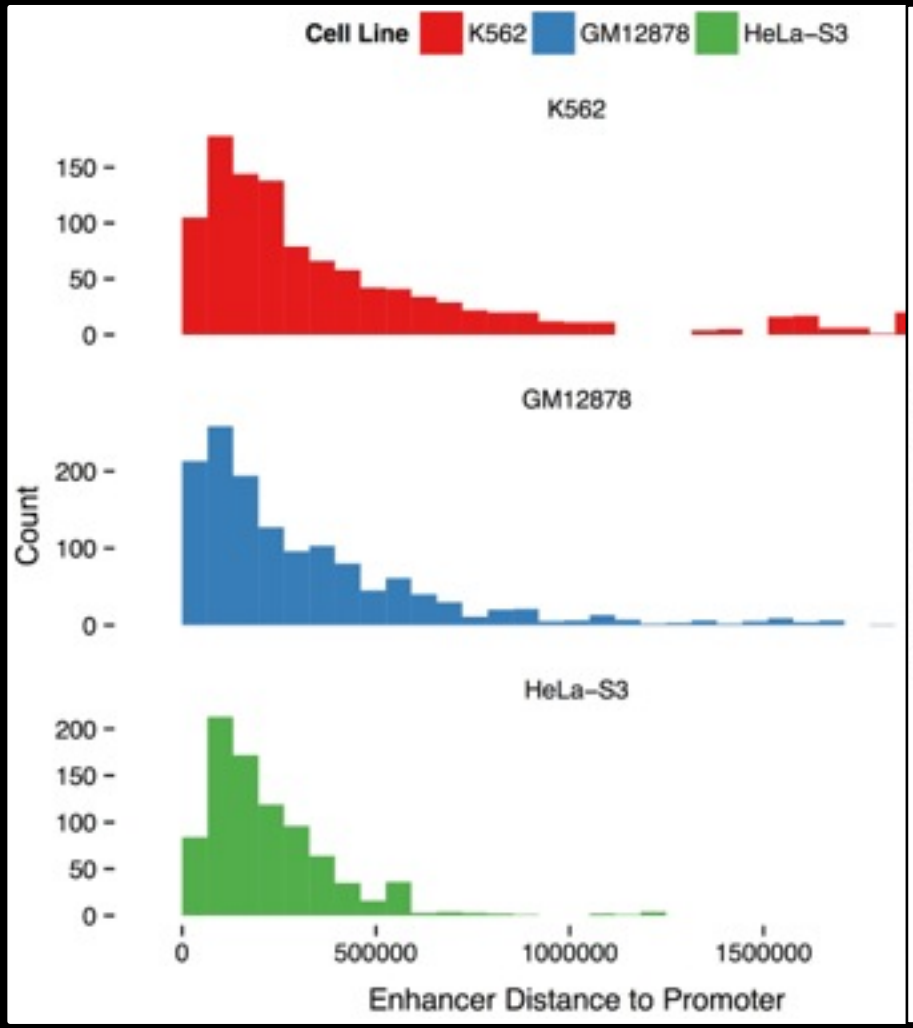
Precision =90-95%

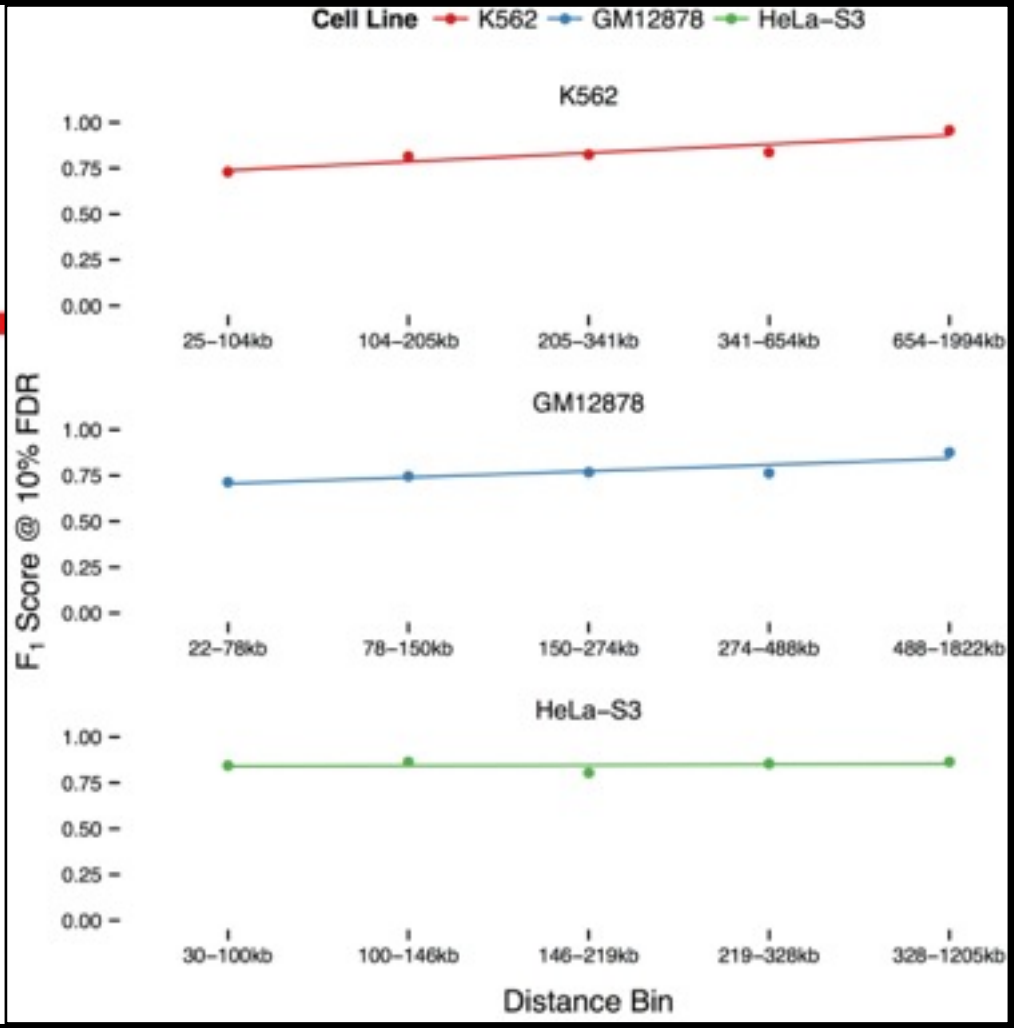
Recall=76-83%

Power=85-89% at 10% FPR

Significantly better than random and logistic regression

TargetFinder performs well at very long distances





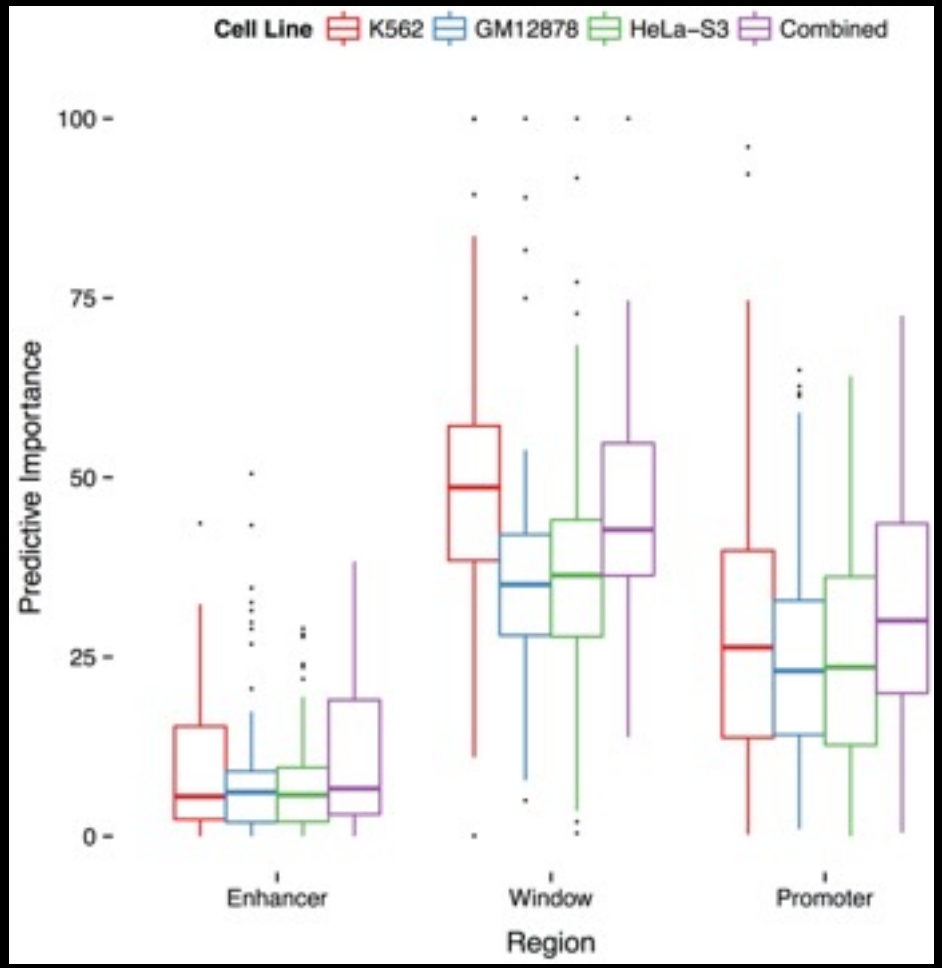
TargetFinder: Feature Importance

TargetFinder: Feature Importance

Most predictive features mark the *window between* the enhancer and the promoter

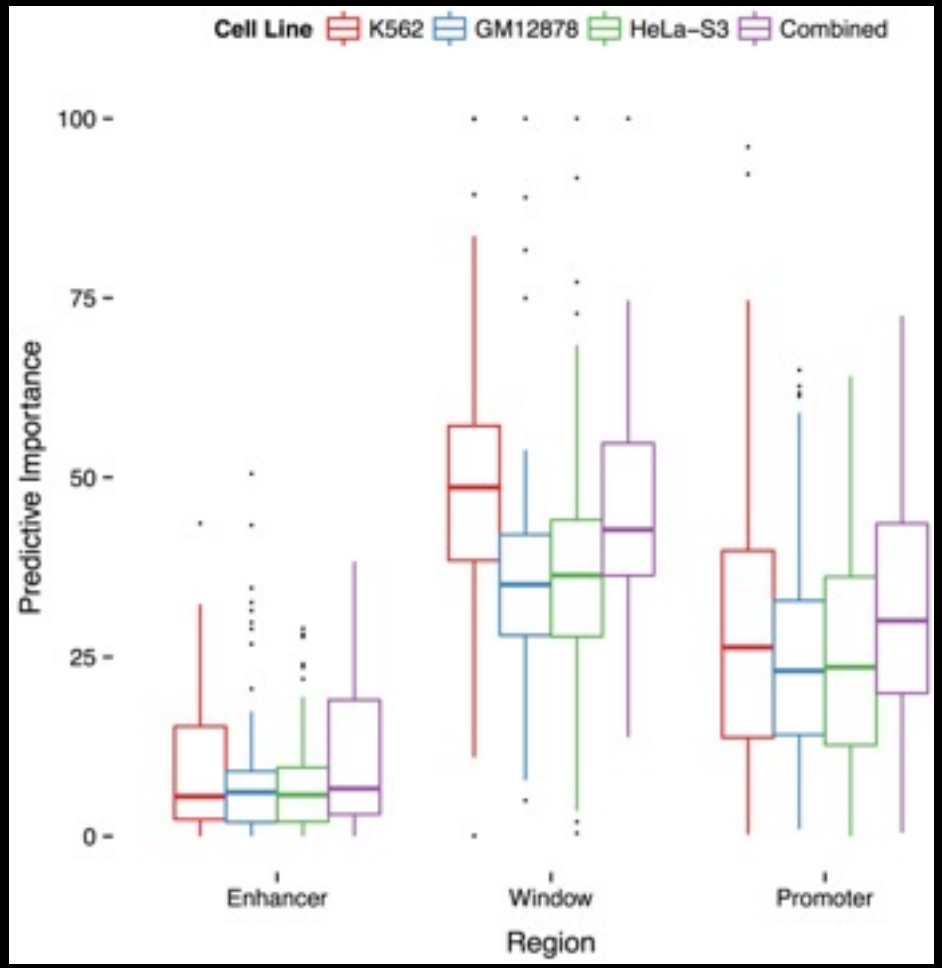
TargetFinder: Feature Importance

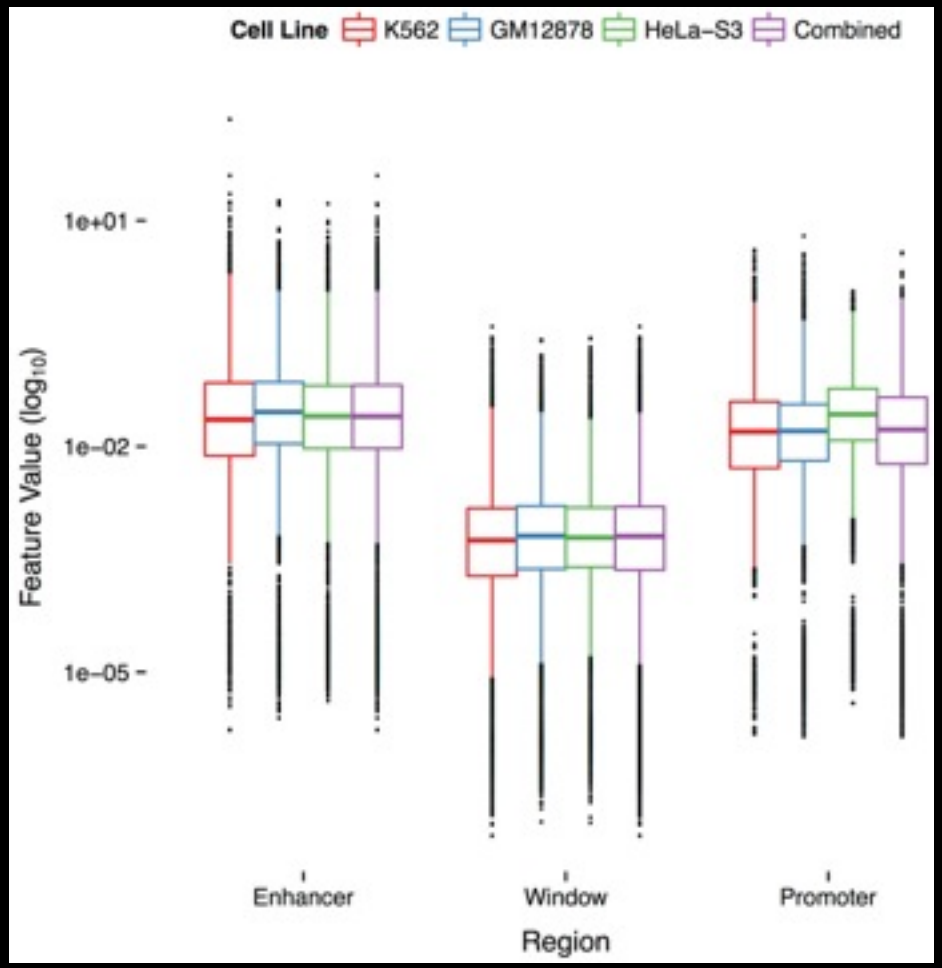
Most predictive features mark the *window between* the enhancer and the promoter



TargetFinder: Feature Importance

Most predictive features mark the *window between* the enhancer and the promoter





TargetFinder: Feature Importance

TargetFinder: Feature Importance

Most useful features for prediction are TF and histone marks in the *window between* the enhancer and the promoter

- **True interactions**
  - Enhancer-associated proteins: P300, JUN, TFs
  - Marks of heterochromatin, lack of DNA methylation
  - Marks of paused or poised RNA polymerase

TargetFinder: Feature Importance

Most useful features for prediction are TF and histone marks in the *window between* the enhancer and the promoter

- **True interactions**
  - Enhancer-associated proteins: P300, JUN, TFs
  - Marks of heterochromatin, lack of DNA methylation
  - Marks of paused or poised RNA polymerase
- **False interactions**
  - Cohesin complex: CTCF, RAD21, SMC3, ZNF143
  - Histone marks of open chromatin and elongation
  - Marks of active promoters and gene bodies

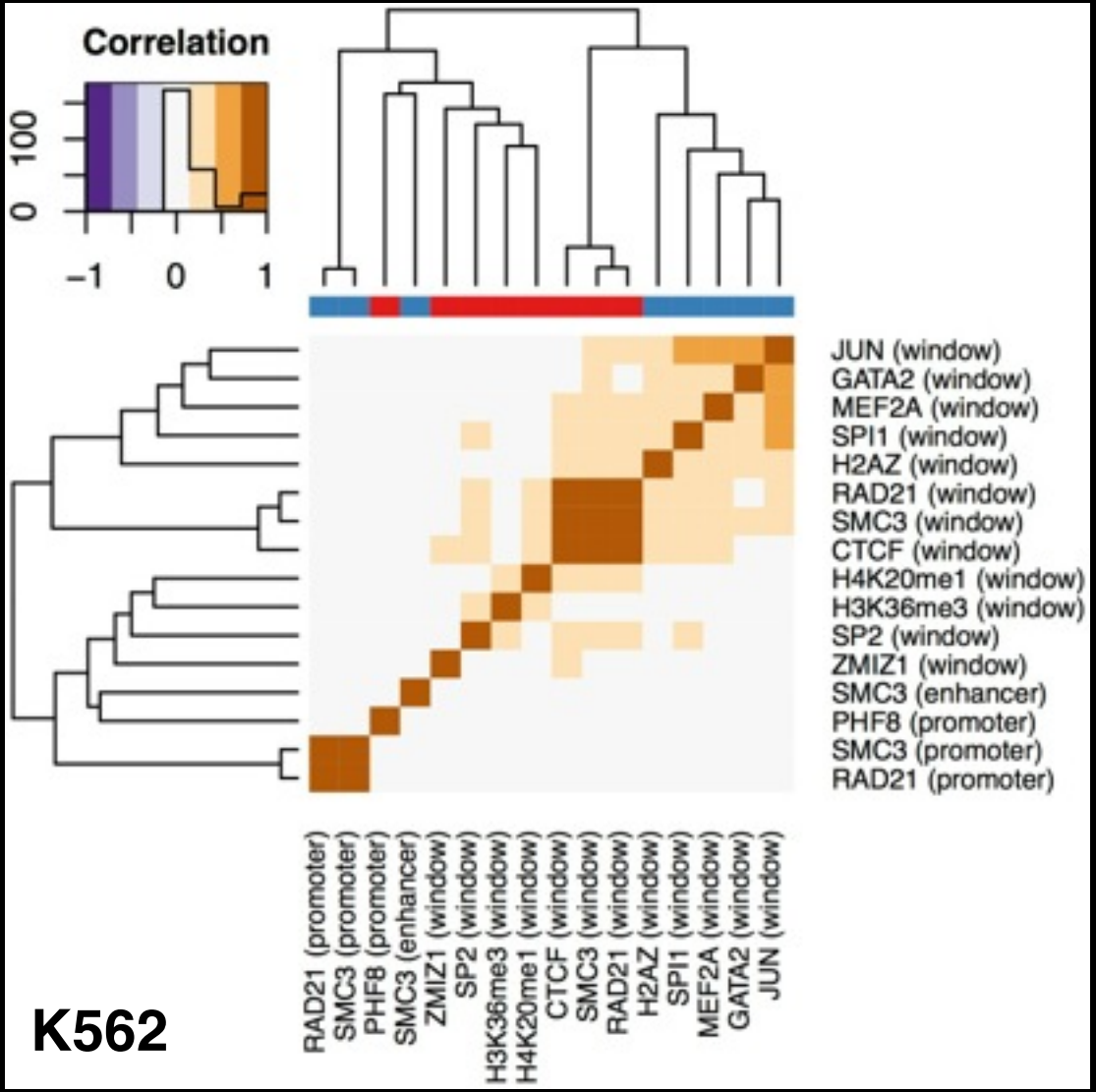
TargetFinder: Feature Importance

Most useful features for prediction are TF and histone marks in the *window between* the enhancer and the promoter

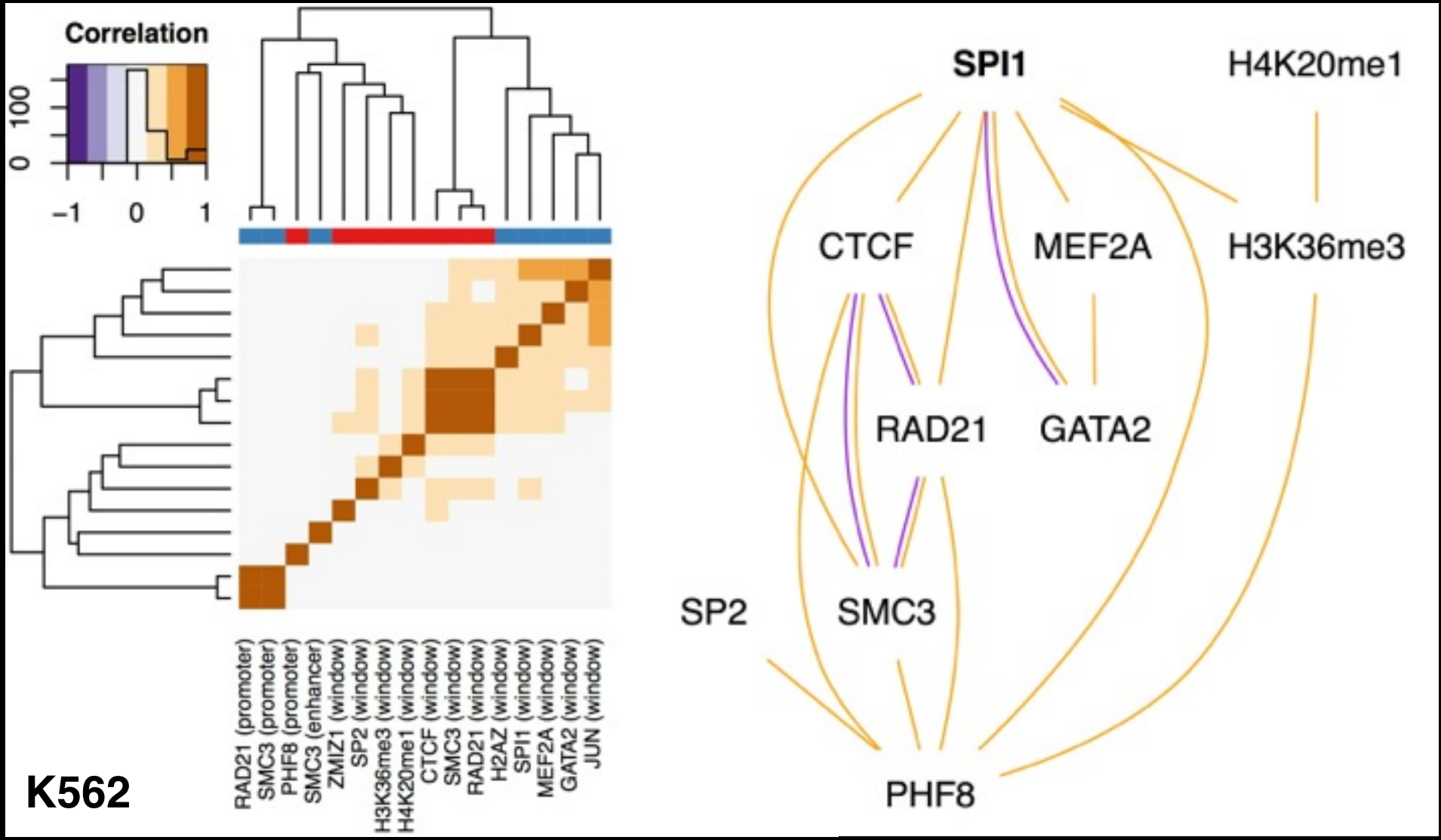
- **True interactions**
  - Enhancer-associated proteins: P300, JUN, TFs
  - Marks of heterochromatin, lack of DNA methylation
  - Marks of paused or poised RNA polymerase
- **False interactions**
  - Cohesin complex: CTCF, RAD21, SMC3, ZNF143
  - Histone marks of open chromatin and elongation
  - Marks of active promoters and gene bodies

Many “window” features have a different meaning when marking promoters and enhancers (e.g., cohesin)

Predictive features colocate and form complexes



Predictive features colocate and form complexes



Can TargetFinder work outside ENCODE cell lines?

Can TargetFinder work outside ENCODE cell lines?

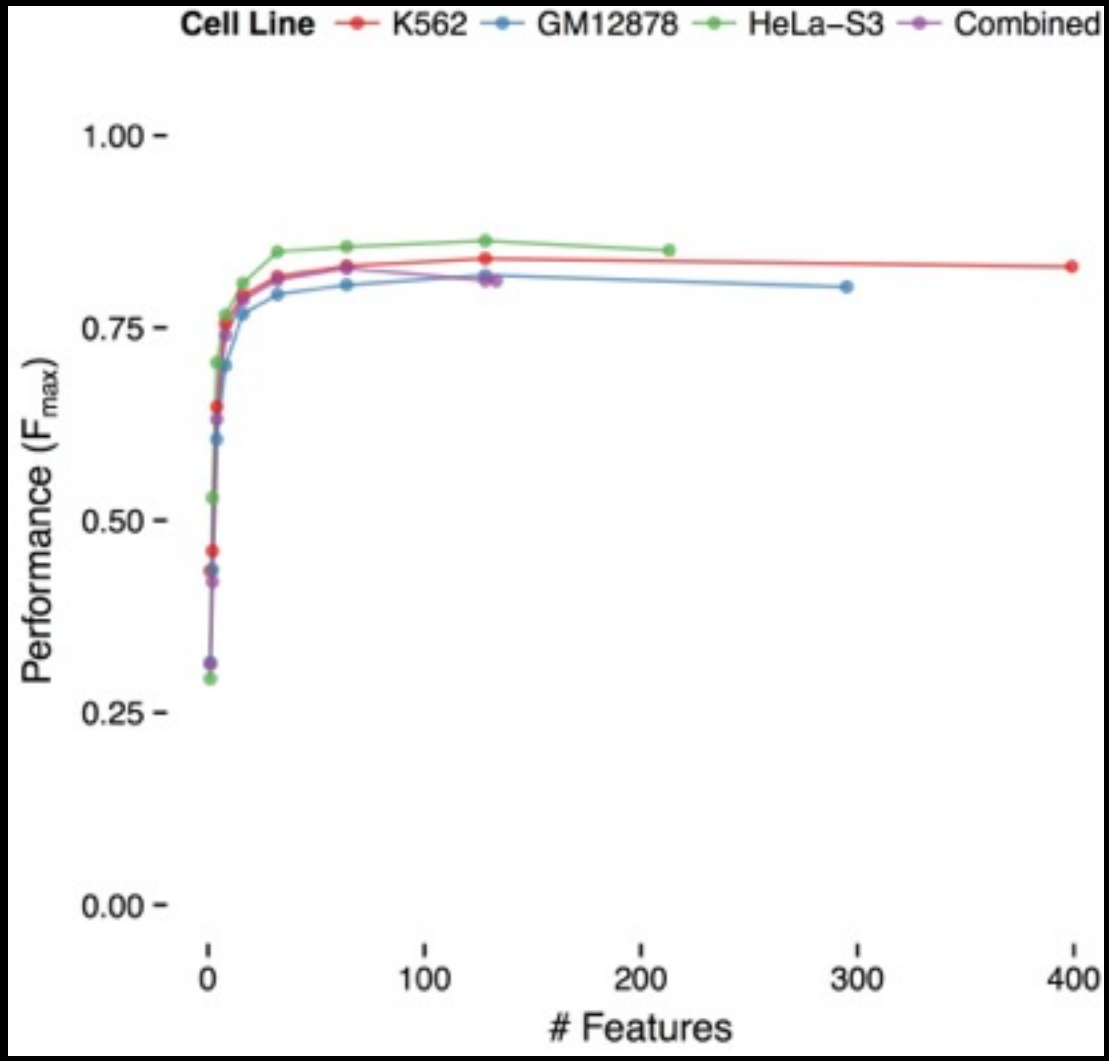
What is a minimal set of experiments for accurate prediction?

Can TargetFinder work outside ENCODE cell lines?

What is a minimal set of experiments for accurate prediction?

optimal: 16+

minimal: 8



Can TargetFinder work outside ENCODE cell lines?

Can TargetFinder work outside ENCODE cell lines?

Test if models generalize across cell types

Can TargetFinder work outside ENCODE cell lines?

Test if models generalize across cell types

EVALUATE

TRAIN Fmax values	GM12878	K562	HeLa-S3	HUVEC
GM12878	0.83	0.40	0.43	0.39
K562	0.46	0.85	0.45	0.44
HeLa-S3	0.43	0.38	0.88	0.41
HUVEC	0.39	0.40	0.38	-

Can TargetFinder work outside ENCODE cell lines?

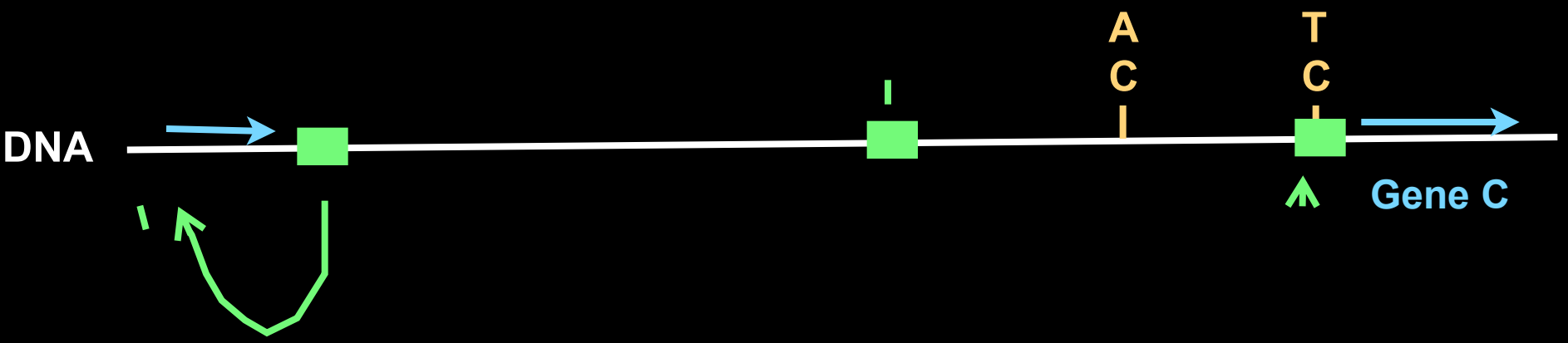
Test if models generalize across cell types

EVALUATE

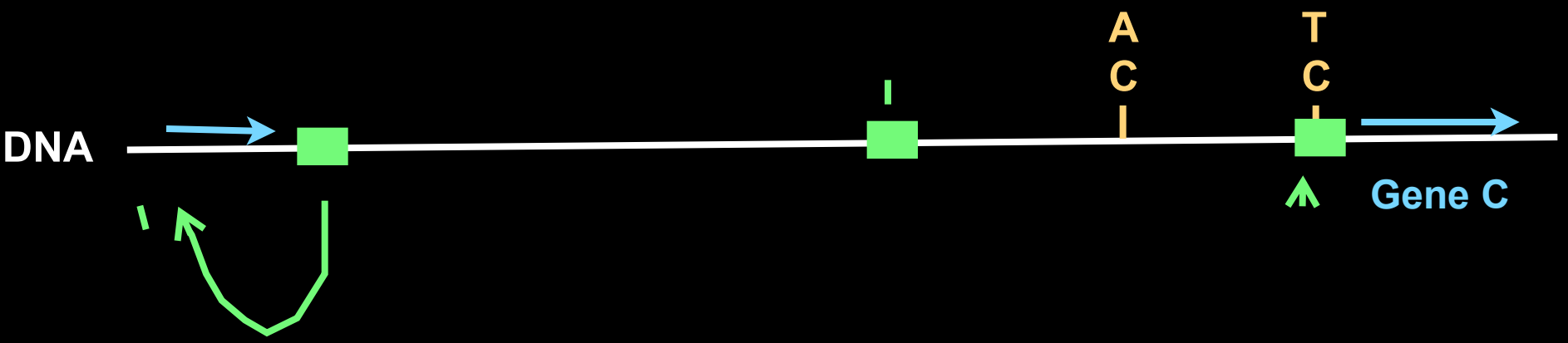
TRAIN Fmax values	GM12878	K562	HeLa-S3	HUVEC
GM12878	0.83	0.40	0.43	0.39
K562	0.46	0.85	0.45	0.44
HeLa-S3	0.43	0.38	0.88	0.41
HUVEC	0.39	0.40	0.38	-

Expect ~35% precision and 55% recall on a new cell type with ~10 ChIP-seq datasets

TargetFinder accurately annotates enhancer-promoter pairs



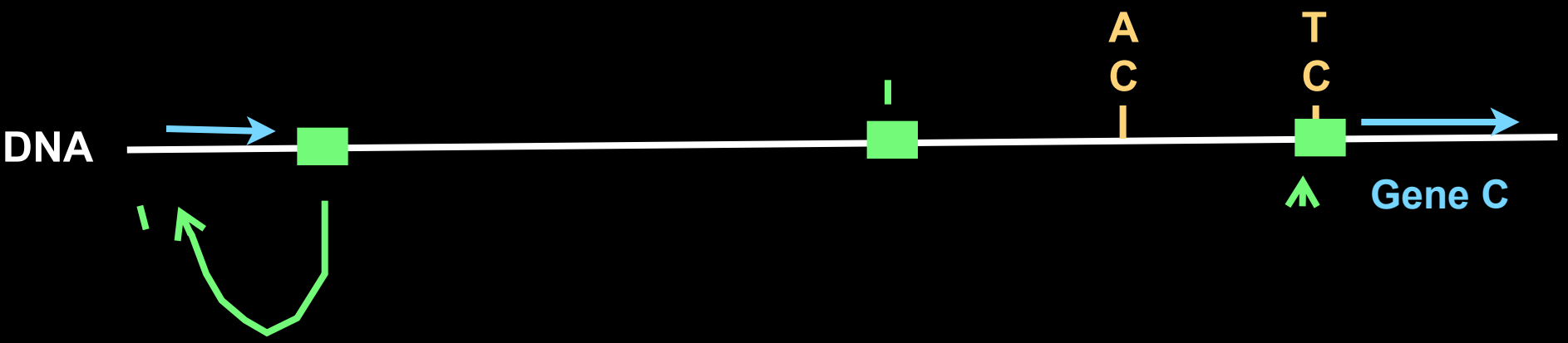
TargetFinder accurately annotates enhancer-promoter pairs



Massive data integration improves prediction

- **Closest gene**
  - Usually fails to identify the right promoter
  - Many false positives

TargetFinder accurately annotates enhancer-promoter pairs



Massive data integration improves prediction

- **Closest gene**
  - Usually fails to identify the right promoter
  - Many false positives
- **TargetFinder**
  - Identifies 95-90% of known pairs (55% with less data)
  - Few false positives

Which human genome sequences function as long-range enhancers?

EnhancerFinder: Training

Teach a machine learning algorithm to identify developmental enhancers active in different tissues based on their features.

Training Data

VISTA Browser


+


–

EnhancerFinder: Training

Teach a machine learning algorithm to identify developmental enhancers active in different tissues based on their features.

Training Data

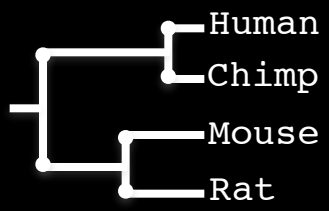
VISTA Browser


+


−

Features

Evolutionary Conservation



EnhancerFinder: Training

Teach a machine learning algorithm to identify developmental enhancers active in different tissues based on their features.

Training Data

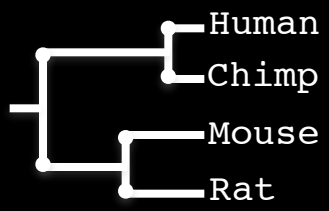
VISTA Browser


+

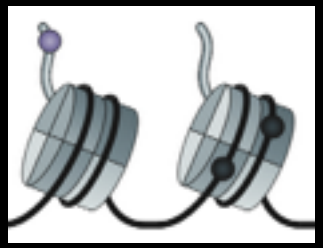

−

Features

Evolutionary Conservation



Functional Genomics



ChIP-seq (TFs, histones)
DNase Hypersensitivity
ENCODE
Epigenomics Roadmap
Bench-to-Bassinet

EnhancerFinder: Training

Teach a machine learning algorithm to identify developmental enhancers active in different tissues based on their features.

Training Data

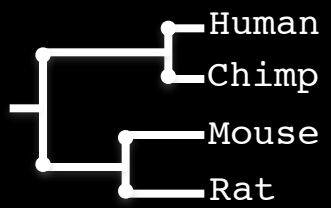
VISTA Browser


+

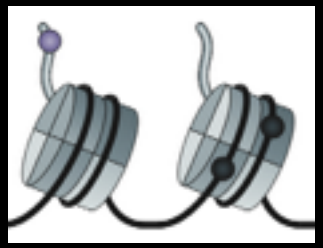

−

Features

Evolutionary Conservation



Functional Genomics



ChIP-seq (TFs, histones)
DNase Hypersensitivity
ENCODE
Epigenomics Roadmap
Bench-to-Bassinet

DNA Sequence Motifs

```
AAAA,AAAC,AAAG,AAAT,
AACA,AACC,AACG,AACT,
AAGA,AAGC,AAGG,AAGT,
AATA,AATC,AATG,AATT,
ACAA,ACAC,ACAG,ACAT,
...
```

short k-mers
known TF motifs

EnhancerFinder: Training

Teach a machine learning algorithm to identify developmental enhancers active in different tissues based on their features.

Training Data

VISTA Browser


+


–

Computational Algorithm

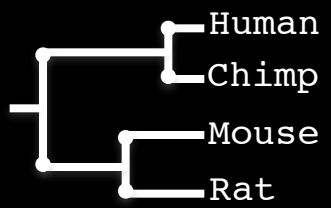
Support vector machine: separates 2 groups

Multi-kernel: good for combining heterogeneous data types with different weights

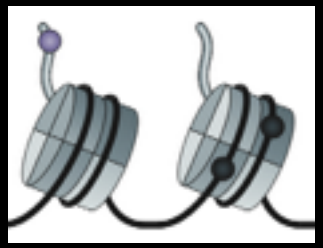
$$f(x) = \sum_{i=1}^{N} \alpha_i \sum_{j=1}^{M} \beta_j k_j(x, x_i) + b$$

Features

Evolutionary Conservation



Functional Genomics



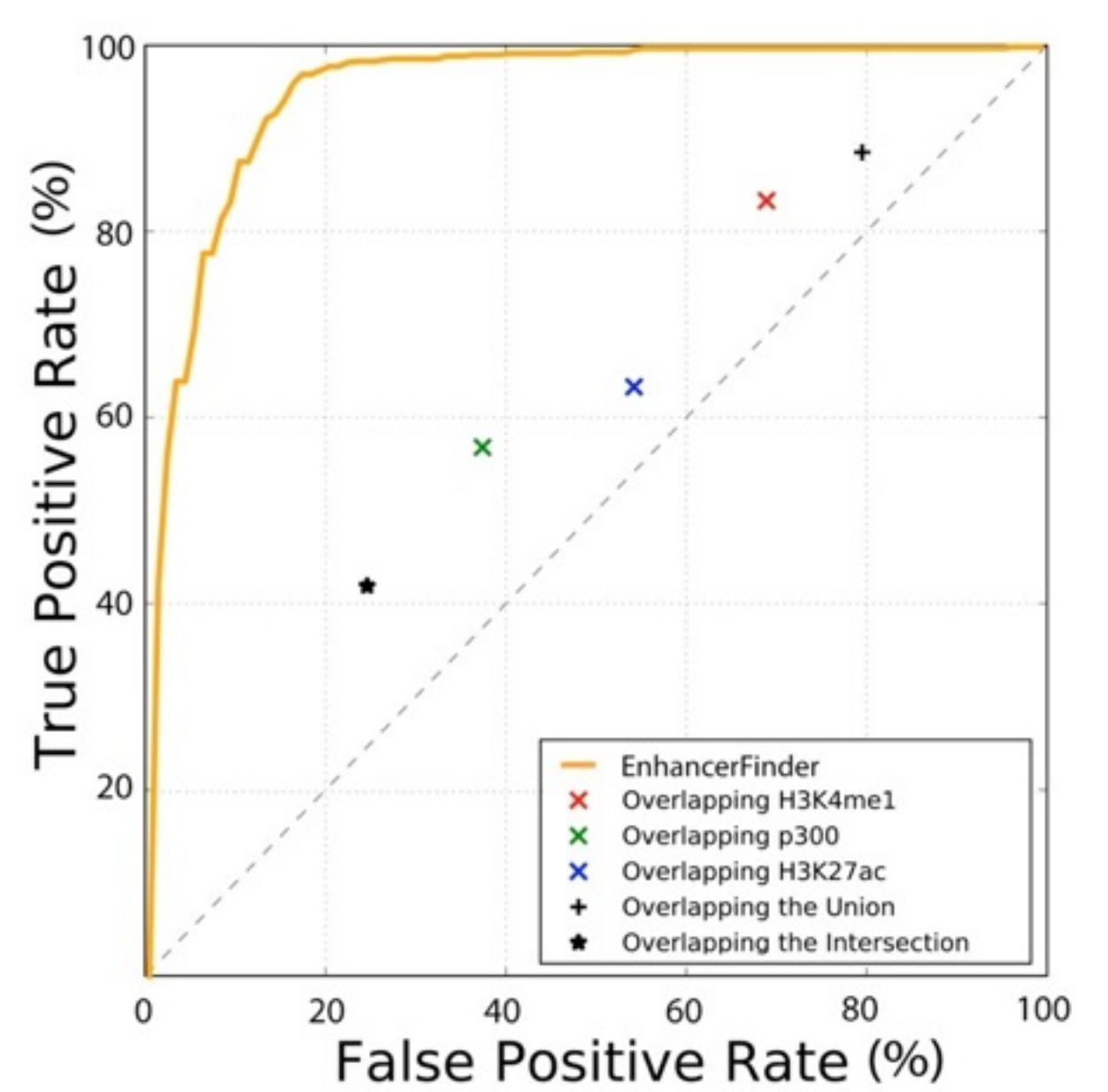
ChIP-seq (TFs, histones)
DNase Hypersensitivity
ENCODE
Epigenomics Roadmap
Bench-to-Bassinet

DNA Sequence Motifs

```
AAAA,AAAC,AAAG,AAAT,
AACA,AACC,AACG,AACT,
AAGA,AAGC,AAGG,AAGT,
AATA,AATC,AATG,AATT,
ACAA,ACAC,ACAG,ACAT,
...
```

short k-mers
known TF motifs

EnhancerFinder: Performance



AUC=0.96
Power=85%
at 10% FPR,
Recall=85% at
93% Precision

FDR ~10-50%
Significantly
better than
other methods

>80% in vivo
validation rate

Erwin et al. (2014)
PLoS Comp Bio

EnhancerFinder: Predictions

Erwin et al. (2014) PLoS Comp Bio, Capra et al. (2014) PTRSB

EnhancerFinder: Predictions

- 84,301 developmental enhancer predictions
  - Cover 2% of the human genome
  - Nearby genes have high expression and annotated functions in the relevant fetal tissue
  - Significant overlap with disease mutations
  - Cluster around developmental transcription factors and signaling genes

Erwin et al. (2014) PLoS Comp Bio, Capra et al. (2014) PTRSB

EnhancerFinder: Predictions

- 84,301 developmental enhancer predictions
  - Cover 2% of the human genome
  - Nearby genes have high expression and annotated functions in the relevant fetal tissue
  - Significant overlap with disease mutations
  - Cluster around developmental transcription factors and signaling genes

- 239 predictions overlap a Human Accelerated Region (33% of HARs), 25/30 validated *in vivo*

Erwin et al. (2014) PLoS Comp Bio, Capra et al. (2014) PTRSB

EnhancerFinder: Predictions

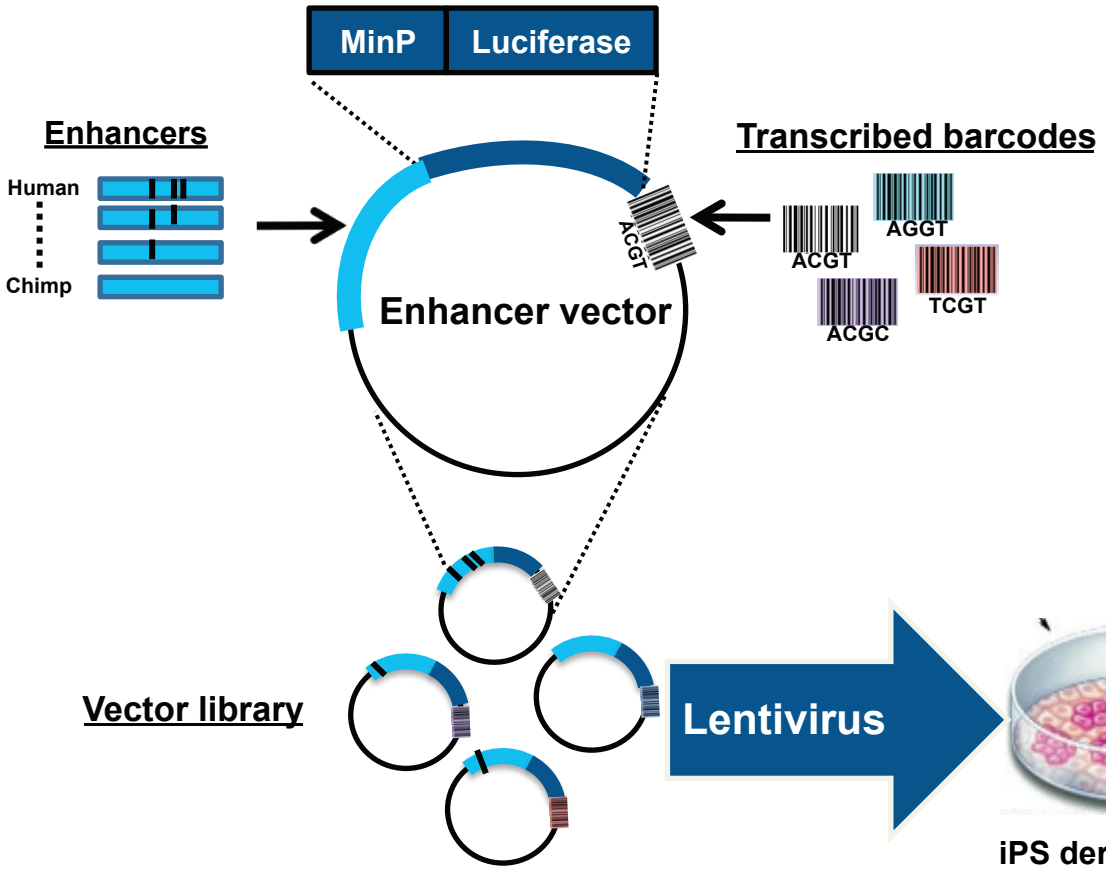
- 84,301 developmental enhancer predictions
  - Cover 2% of the human genome
  - Nearby genes have high expression and annotated functions in the relevant fetal tissue
  - Significant overlap with disease mutations
  - Cluster around developmental transcription factors and signaling genes

- 239 predictions overlap a Human Accelerated Region (33% of HARs), 25/30 validated *in vivo*

- Identify sites with fitness effects (Gulko et al 2015)

Erwin et al. (2014) PLoS Comp Bio, Capra et al. (2014) PTRSB

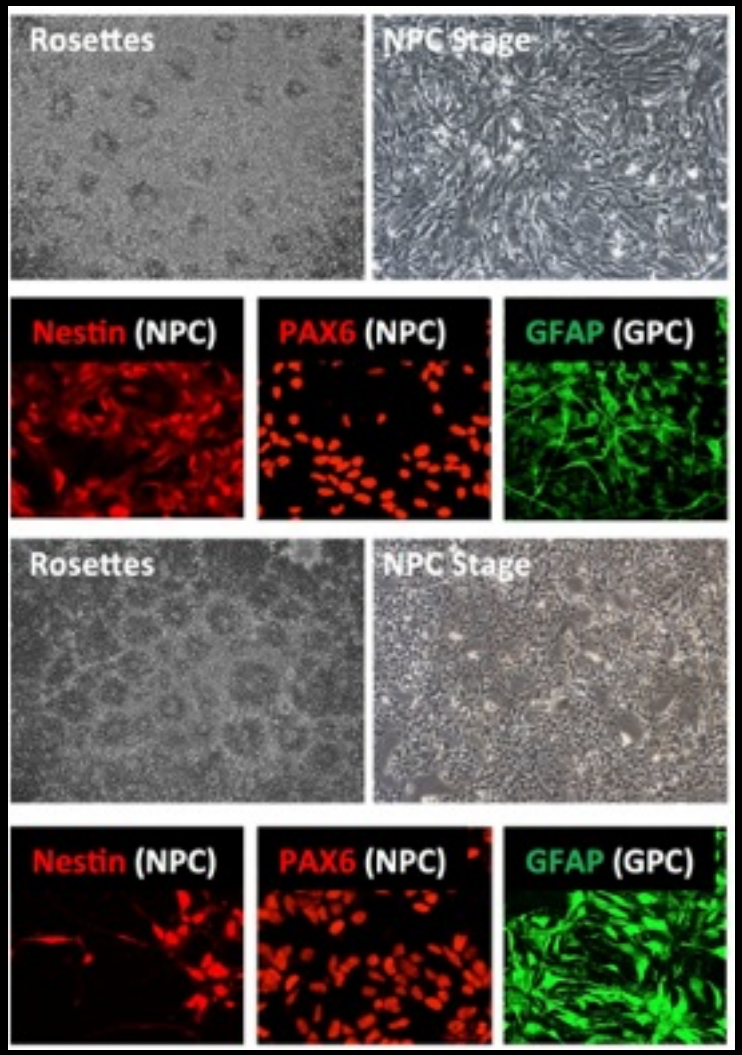
Massively Parallel Reporter Assays and capture Hi-C for validation

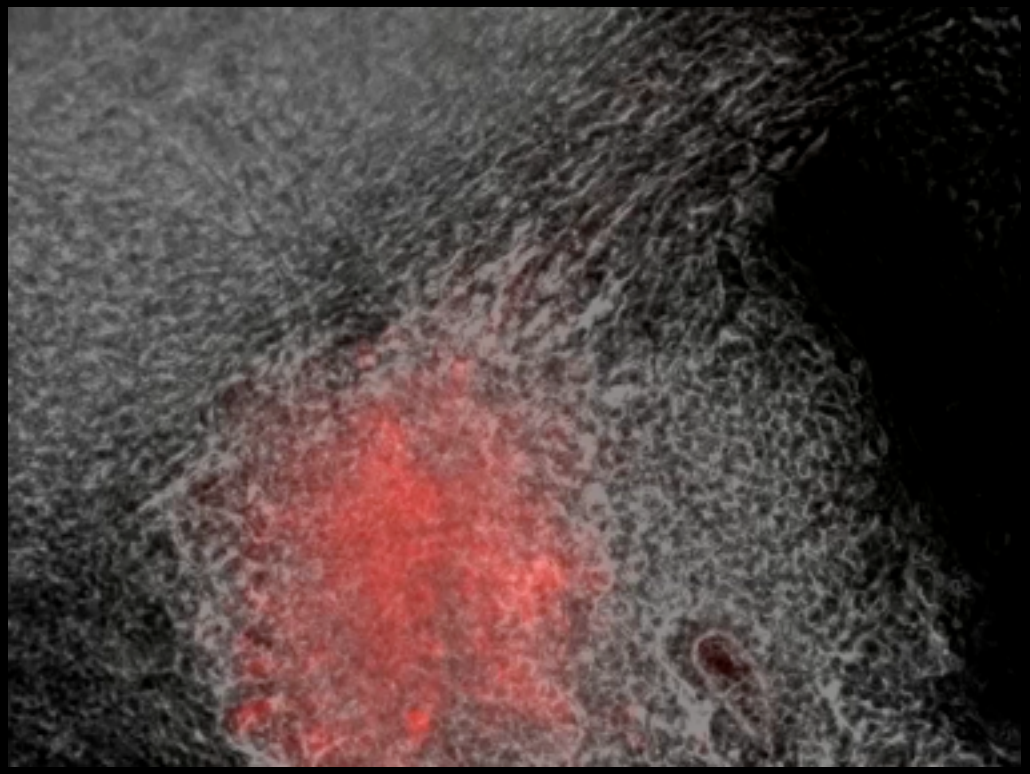


- Test >12,000 170bp enhancers in parallel
- Quantitative activity assayed via RNA-seq
- Compare genotypes
- Human vs. chimp
- Disease SNPs

Hane Ryu, Nadav Ahituv, Jay Shendure, Yin Shen

Induced pluripotent stem cell derived neuronal and cardiac lines



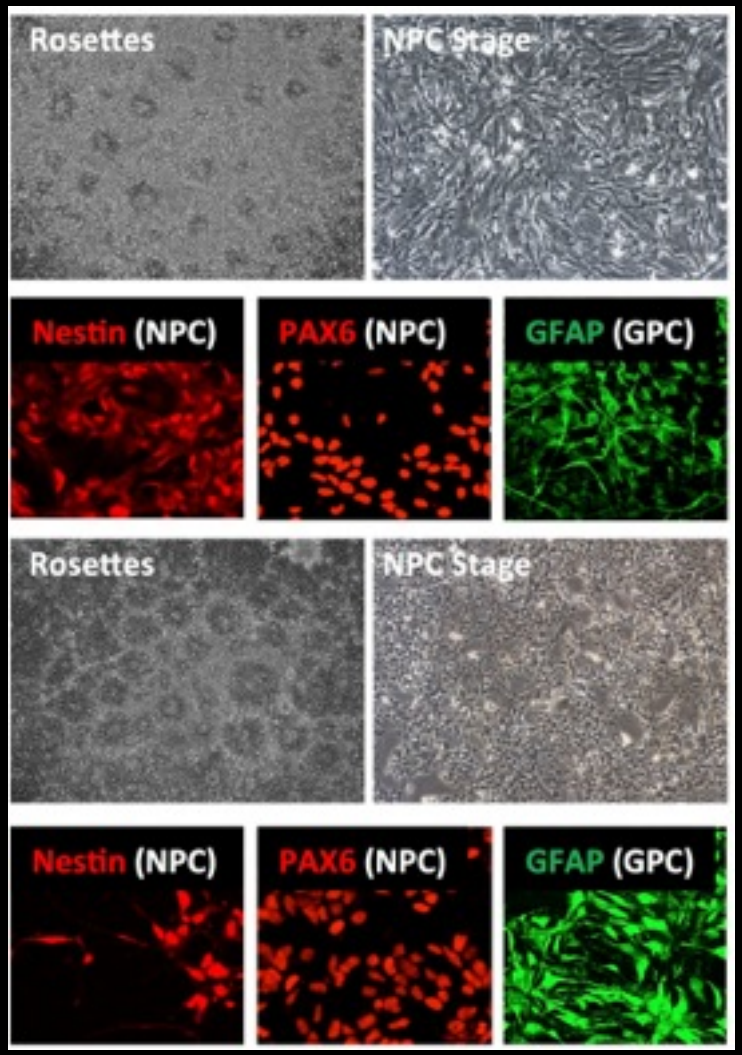


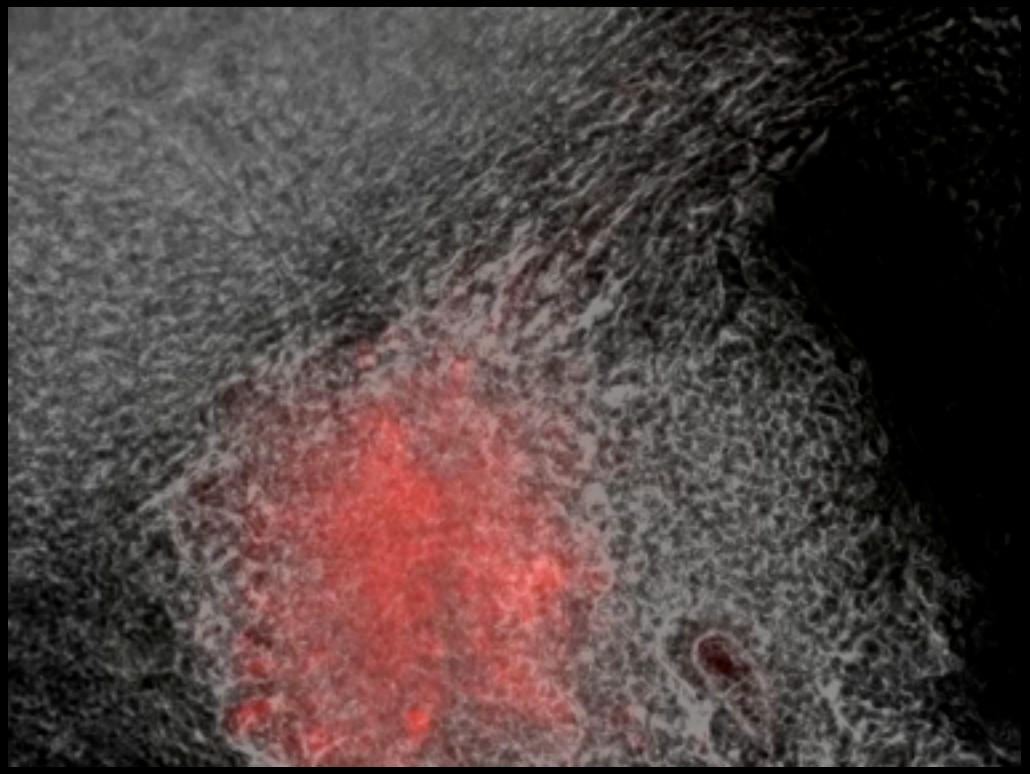
Human iPSC derived cardiomyocytes

Hane Ryu, Alex Pollen, Nadav Ahituv, Arnold Kriegstein

Bruce Conklin

Induced pluripotent stem cell derived neuronal and cardiac lines





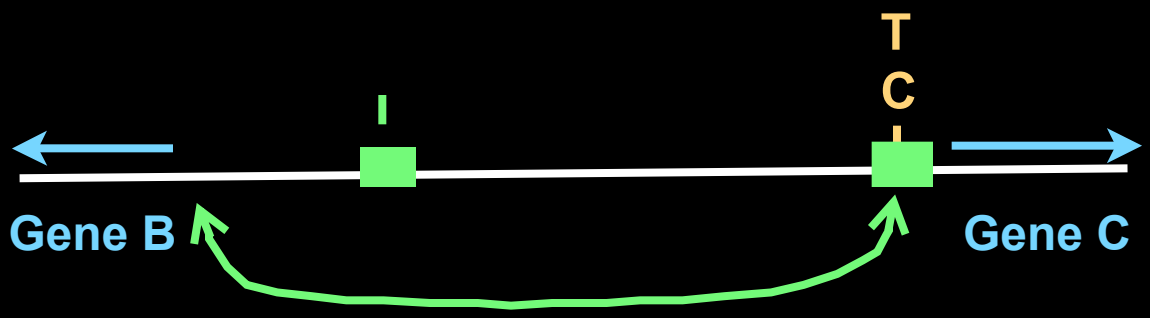
Human iPSC derived cardiomyocytes

Hane Ryu, Alex Pollen, Nadav Ahituv, Arnold Kriegstein

Bruce Conklin

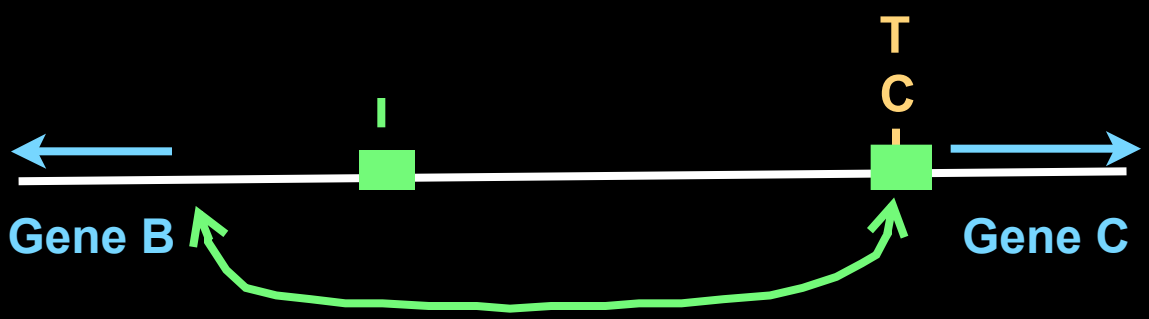
Discovering the role of enhancers in human biology and evolution

Discovering the role of enhancers in human biology and evolution

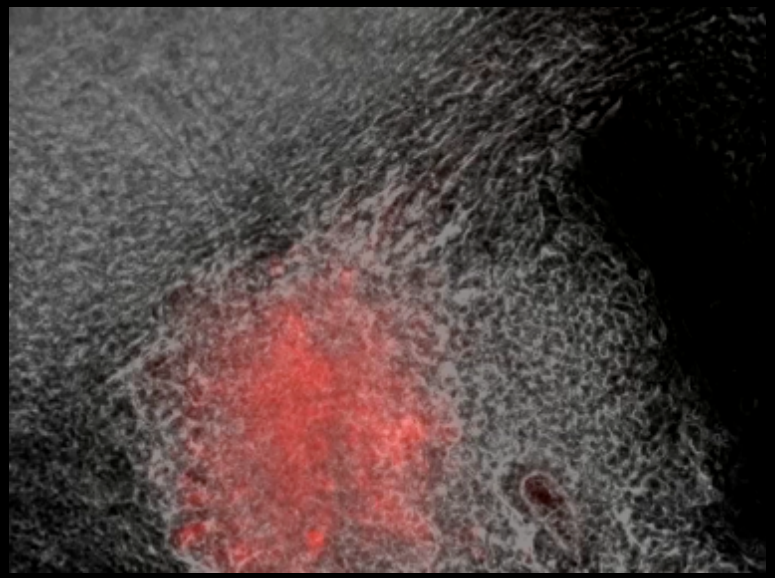


Computational Characterization

Discovering the role of enhancers in human biology and evolution

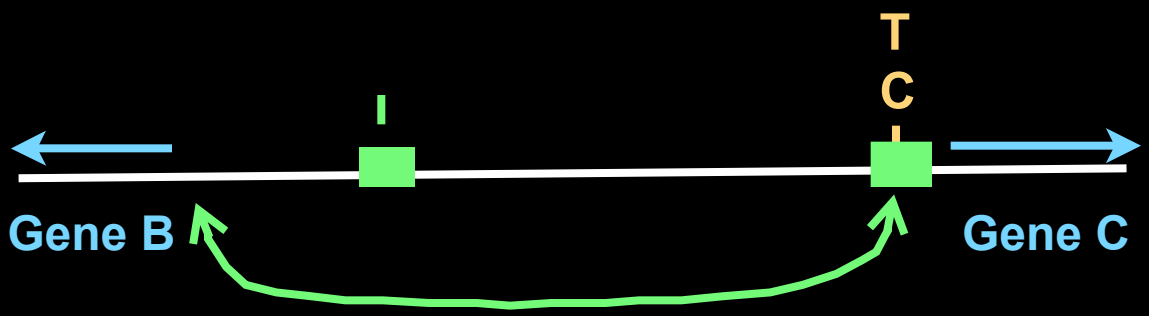


Computational Characterization

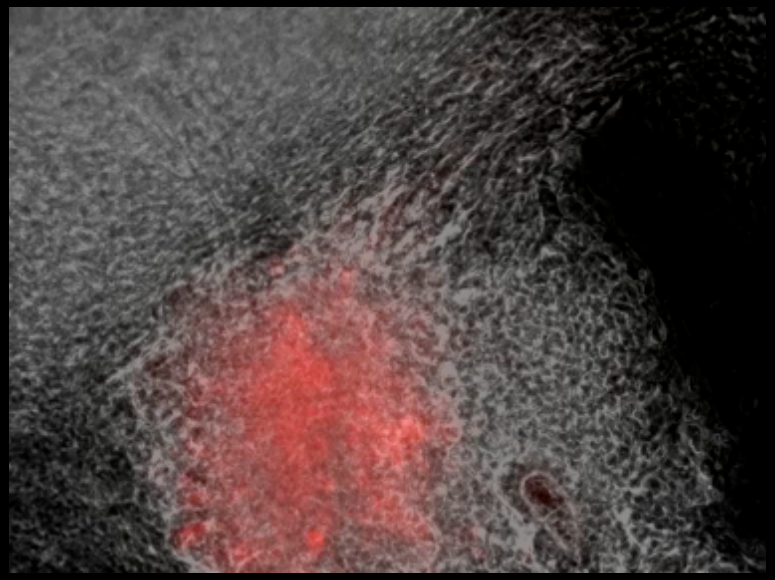


iPSC based screening

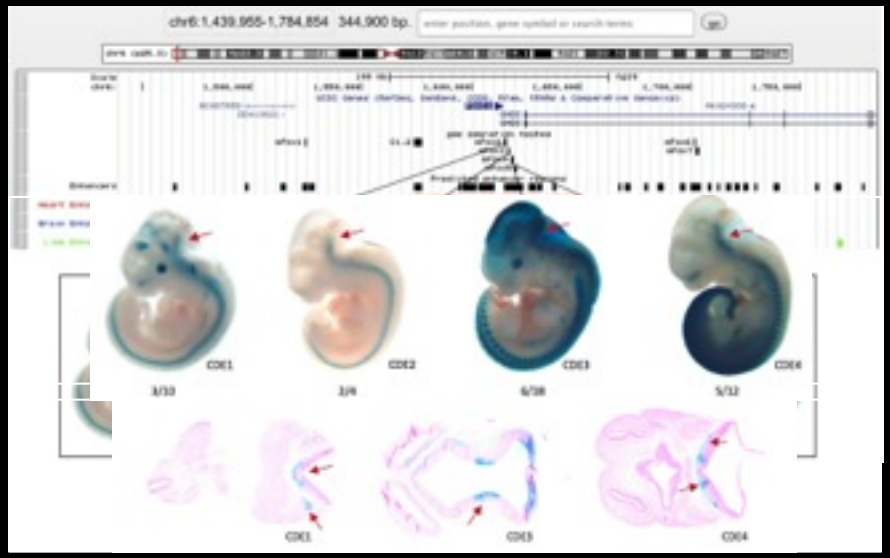
Discovering the role of enhancers in human biology and evolution



Computational Characterization



iPSC based screening



In vivo molecular studies

Collaborators

<u>EnhancerFinder</u>
Gen Haliburton
Tony Capra
Dennis Kostka
John Rubenstein

<u>TargetFinder</u>
Sean Whalen
Rebecca Truty
Benoit Bruneau

<u>MotifDiverge</u>
Dennis Kostka
Tara Friedrich
Deb Ritter
Jeff Chuang

Funding from NHLBI, PhRMA Foundation, Gladstone Institutes