# Computational Systems Biology
# Deep Learning in the Life Sciences

6.802 6.874 20.390 20.490 HST.506

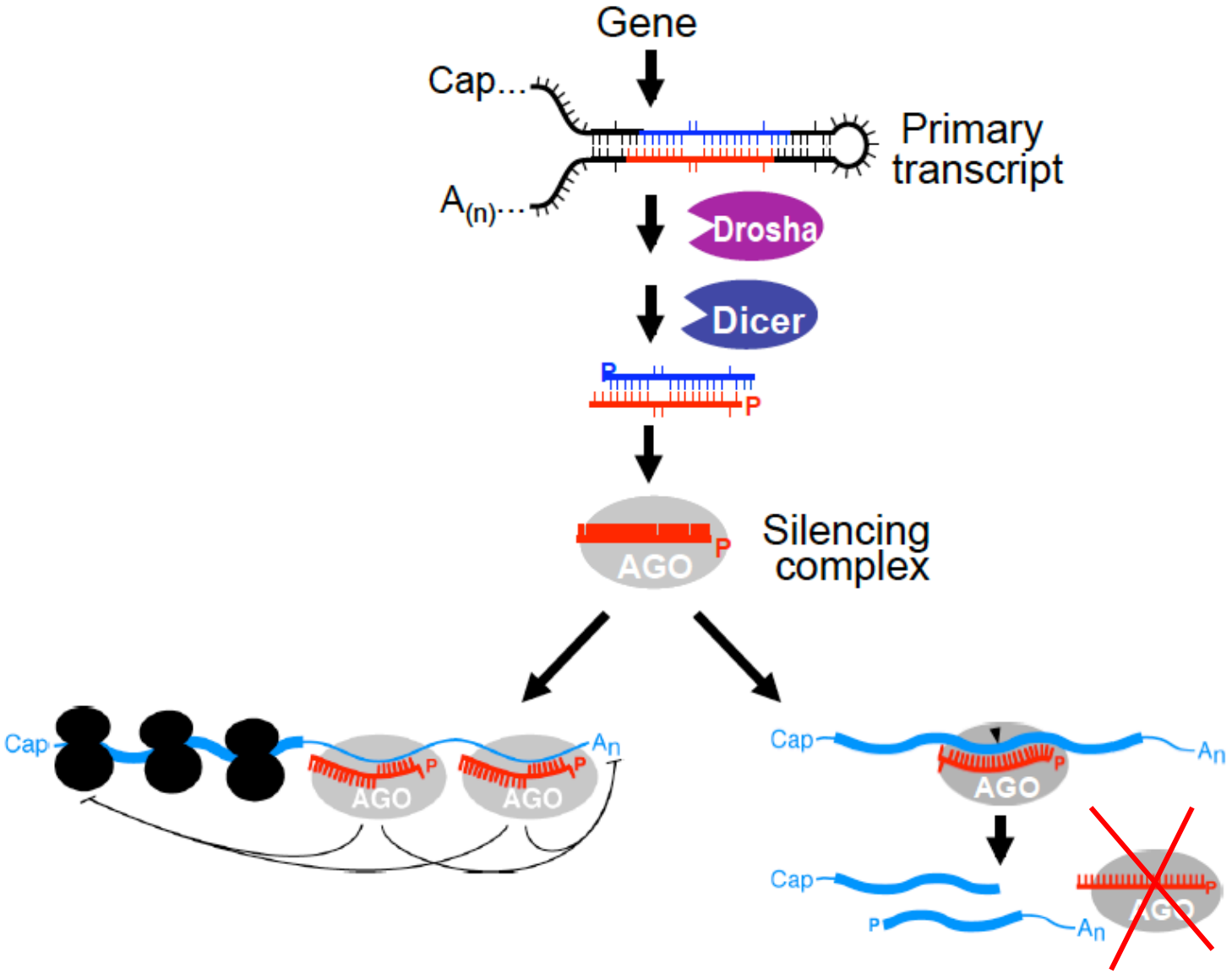
David Gifford
Lecture 9

March 7, 2018

## Regulatory programs
## Chromatin accessibility


http://mit6874.github.io

# miRNAs regulate gene translation




https://www.frontiersin.org/articles/10.3389/fendo.2018.00402/full

# What's on tap today!

- Regression trees for modeling gene expression
- Chromatin accessibility can reveal TF binding
- Predicting chromatin accessibility
  - Deep learning methods (Bassett)
  - Log linear model (SCM)

# What you should know

- How to build a regression tree
- How to penalize a likelihood to produce a BIC score to compare models
- Definition of a Bayes factor
- How to solve a log linear model
- How to interpret DNase-seq and ATAC-seq data
- What controls chromatin accessibility

# We can penalize a likelihood based upon the complexity of a model

k – Number of parameters in the model

n - Number of observations in X

$$BIC_M = \log \mathcal{L}_M(X|\hat{\Theta}) - \frac{k}{2} \log n$$

# A BIC score is an approximation to integration over parameter space

Bayes' theorem:

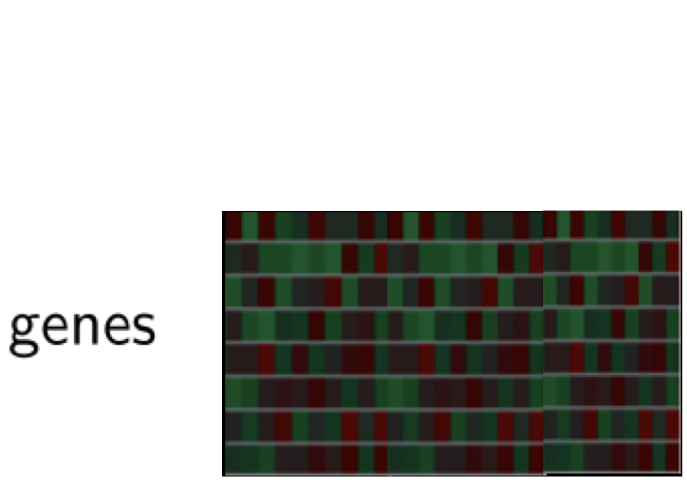
$$\Pr(M|D) = \frac{\Pr(D|M)\Pr(M)}{\Pr(D)}.$$

Bayes' Factor K

$$K = \frac{\Pr(D|M_1)}{\Pr(D|M_2)} = \frac{\int \Pr(\theta_1|M_1)\Pr(D|\theta_1, M_1)\,d\theta_1}{\int \Pr(\theta_2|M_2)\Pr(D|\theta_2, M_2)\,d\theta_2} = \frac{\Pr(M_1|D)}{\Pr(M_2|D)}\frac{\Pr(M_2)}{\Pr(M_1)}.$$

# Regression trees are a modular approach for modeling RNA expression

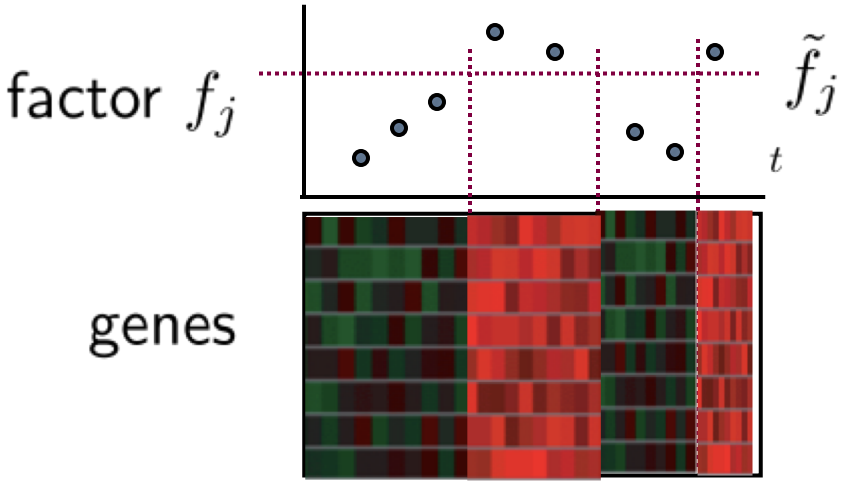
# Competing "programs"




$$N(x_{it}; \hat{\mu}, \hat{\sigma}^2)$$

BIC score =

$$l(X; \hat{\theta}) - \tfrac{2}{2} \log(nm)$$




$$f_{jt} \leq \tilde{f}_j \qquad f_{jt} > \tilde{f}_j$$

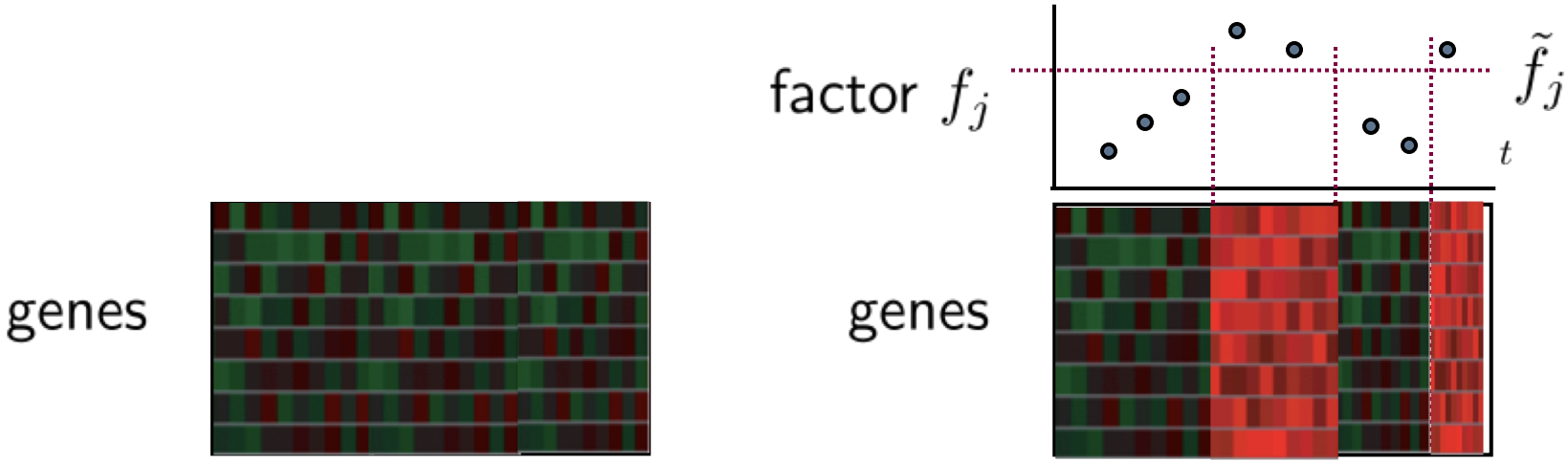
$$N(x_{it}; \hat{\mu}_1, \hat{\sigma}_1^2) \qquad N(x_{it}; \hat{\mu}_2, \hat{\sigma}_2^2)$$

BIC score =

$$l(X|f_j; \hat{\Theta}) - \tfrac{5}{2} \log(nm)$$

Red – high expression ("hot")
Green – low expression

# The algorithm




$N(x_{it}; \hat{\mu}, \hat{\sigma}^2)$

BIC score $=$
$l(X; \hat{\theta}) - \frac{2}{2} \log(nm)$

$f_{jt} \leq \tilde{f}_j$ $\quad$ $f_{jt} > \tilde{f}_j$

$N(x_{it}; \hat{\mu}_1, \hat{\sigma}_1^2)$ $\quad$ $N(x_{it}; \hat{\mu}_2, \hat{\sigma}_2^2)$

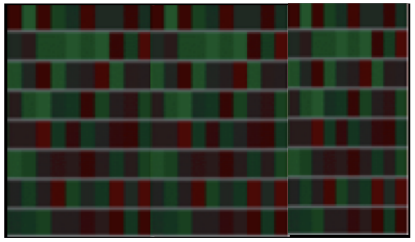
BIC score $=$
$l(X|f_j; \hat{\Theta}) - \frac{5}{2} \log(nm)$

Algorithm:

1) Calculate BIC scores with and without factor
2) If null model wins, stop, otherwise recurse

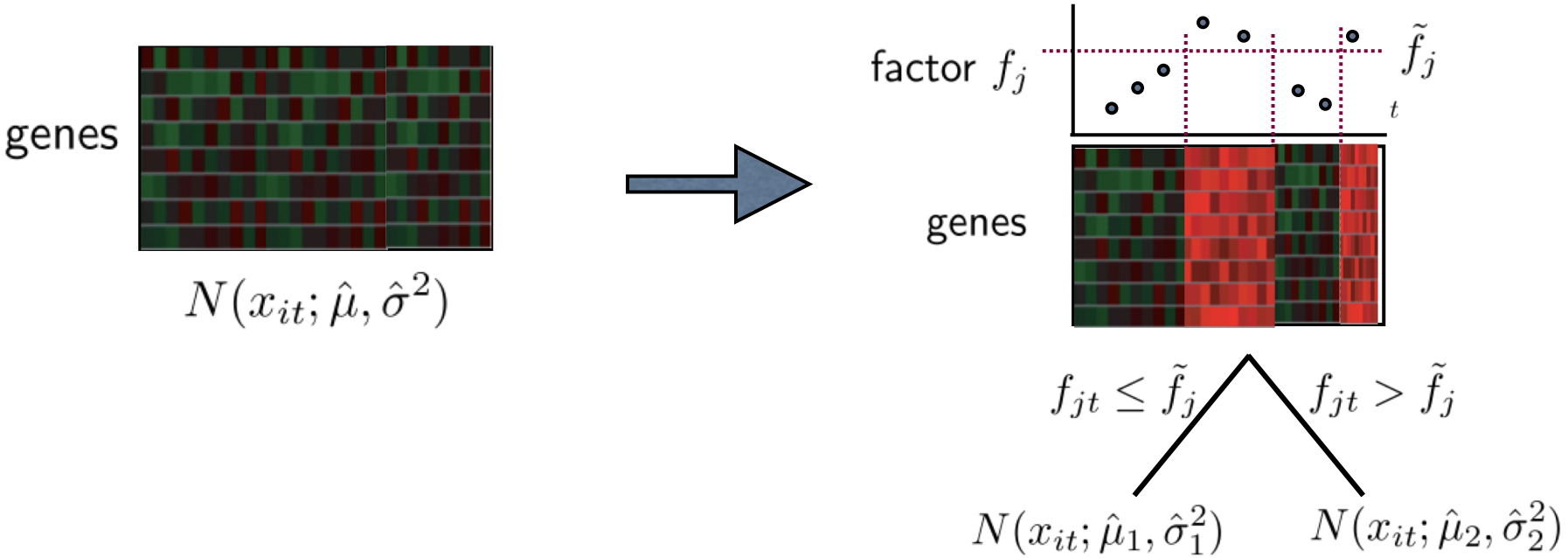
# Recursive estimation

genes

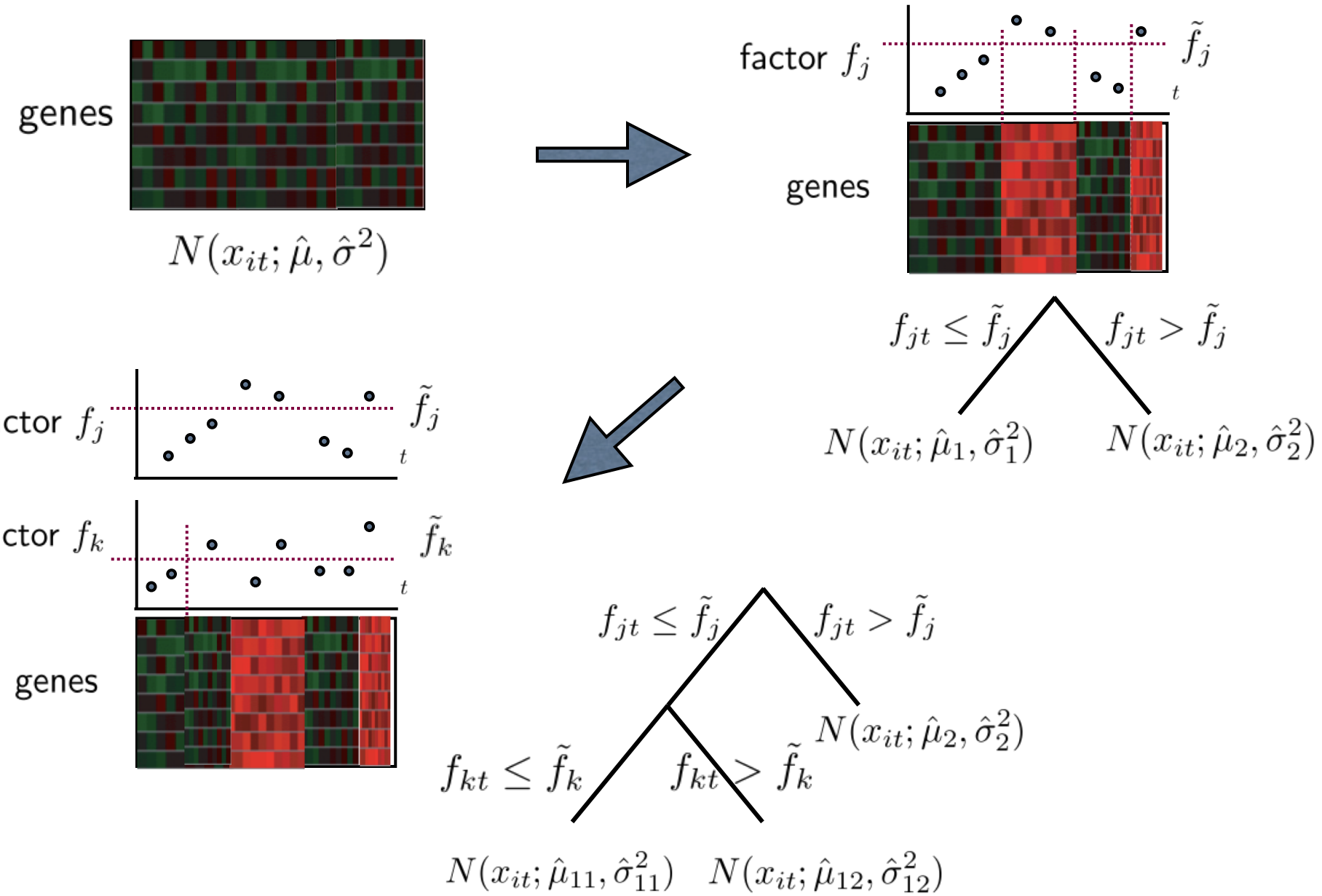


$$N(x_{it}; \hat{\mu}, \hat{\sigma}^2)$$

# Recursive estimation

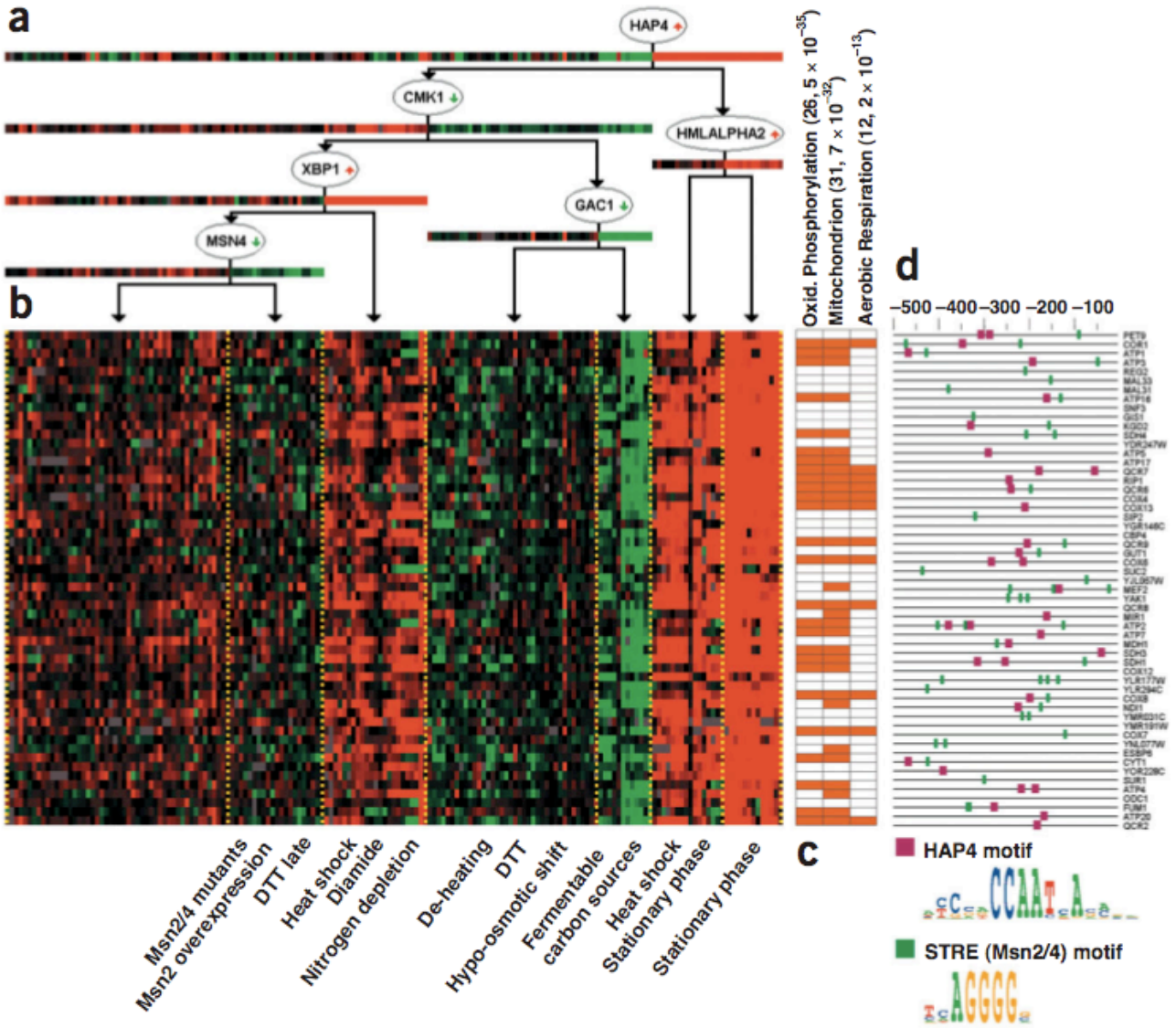


# Recursive estimation

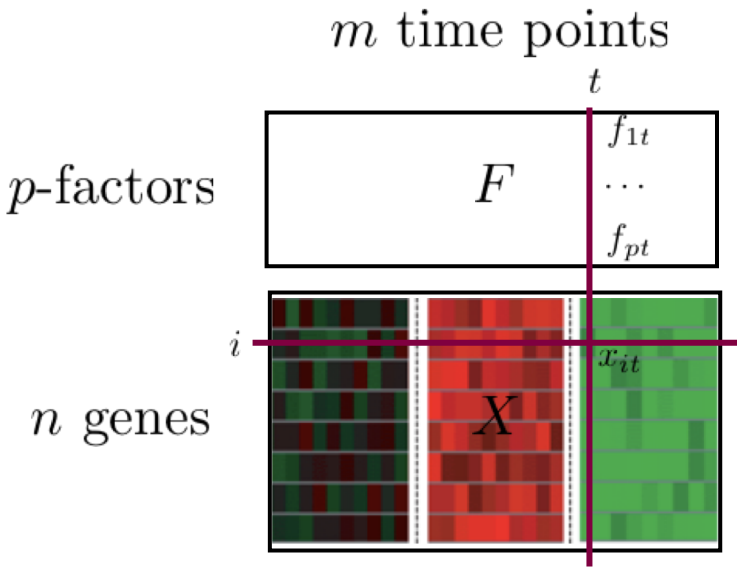


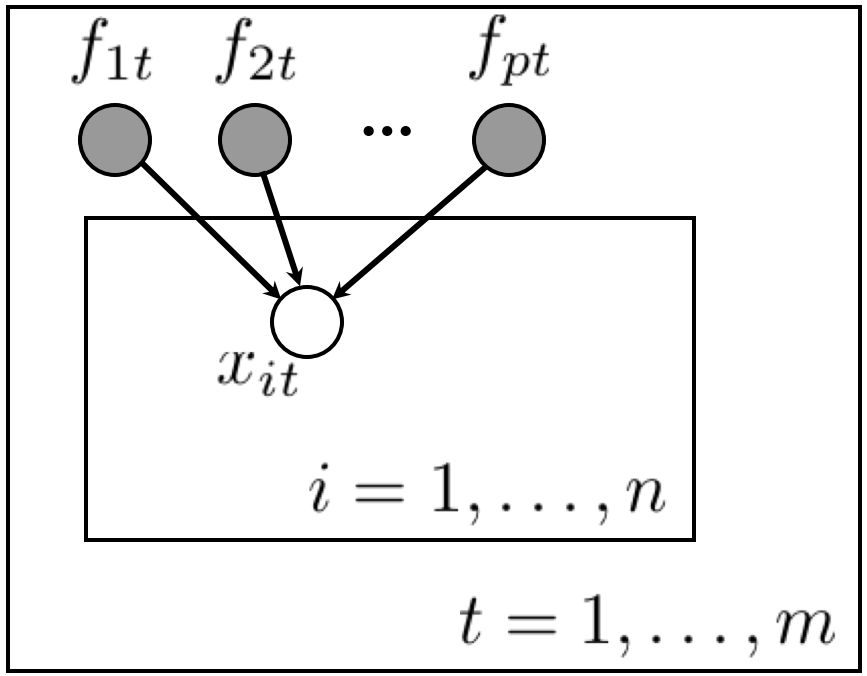
# Expression programs

- An example expression program found in yeast



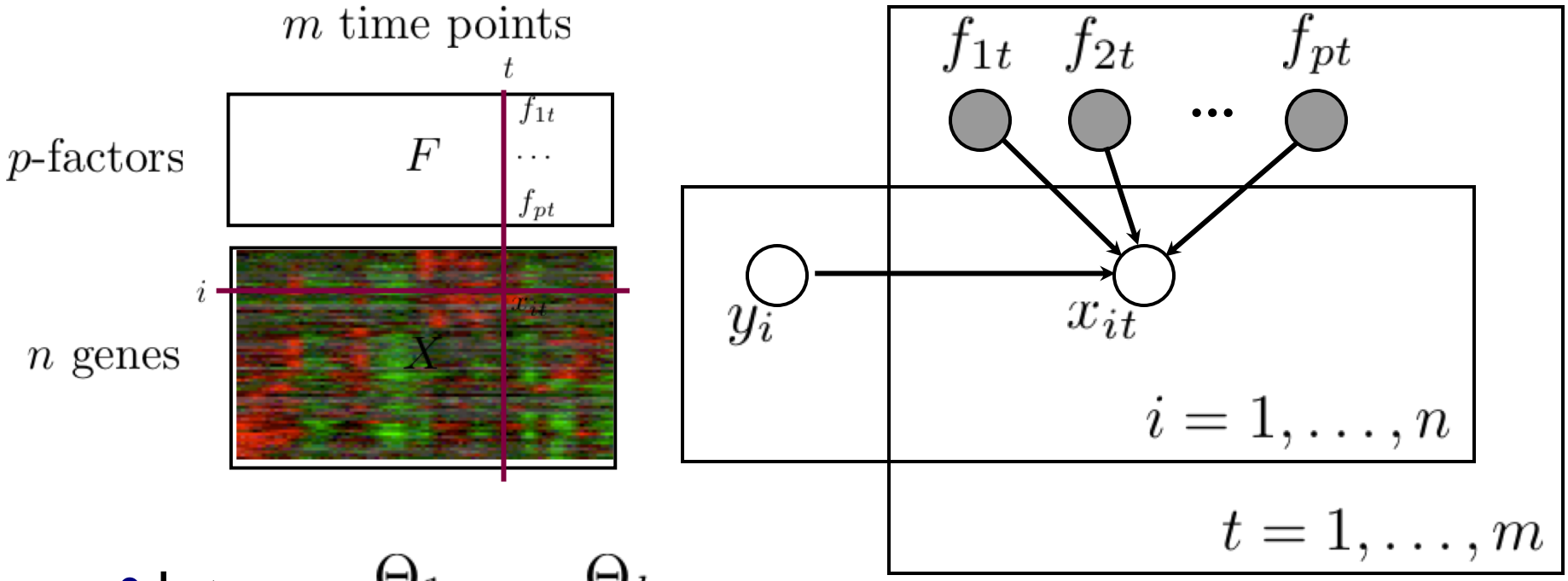
# A single regression tree: sampling model




$$P(X|F,\Theta_1) = \prod_{t=1}^{m} \left[ \prod_{i=1}^{n} P(x_{it}|f_{1t},\ldots,f_{pt},\Theta_1) \right]$$

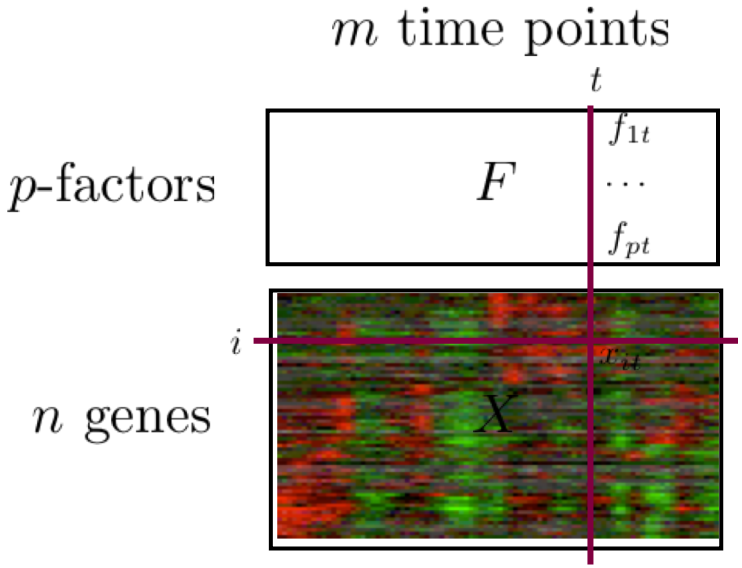
# A mixture of trees: sampling model

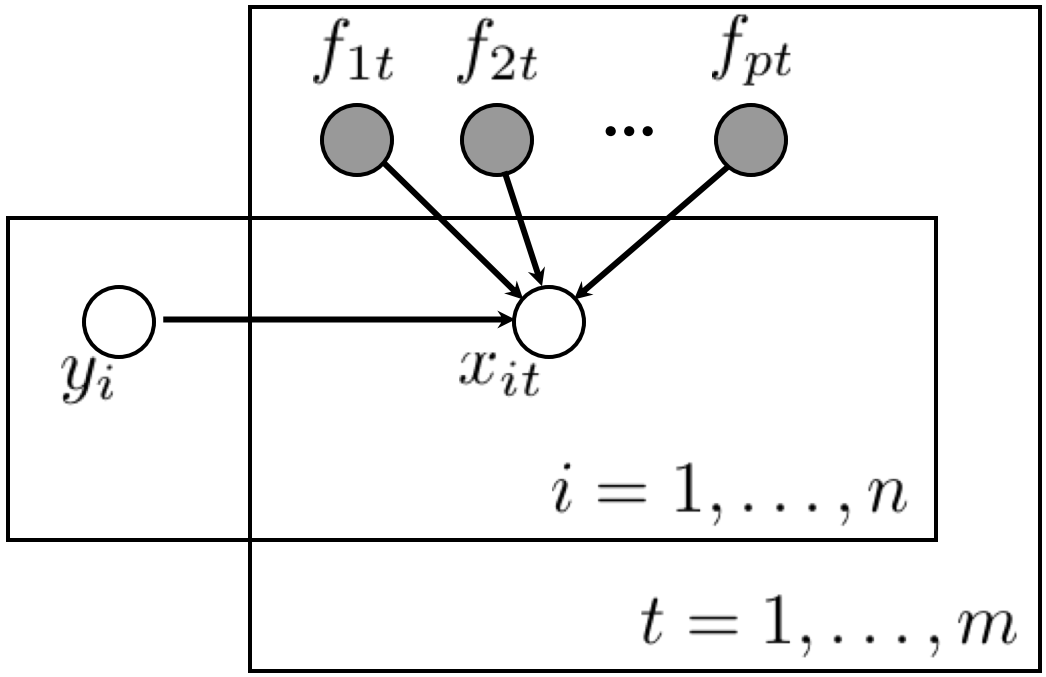



- k trees $\Theta_1, \ldots, \Theta_k$

$$P(x_{it}|f_{1t}, \ldots, f_{pt}, y) = P(x_{it}|f_{1t}, \ldots, f_{pt}, \Theta_y)$$

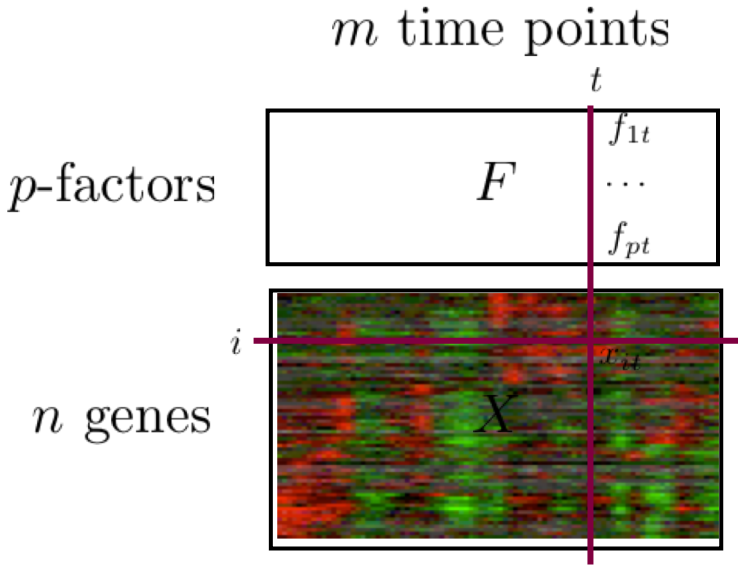
# A mixture of trees: sampling model

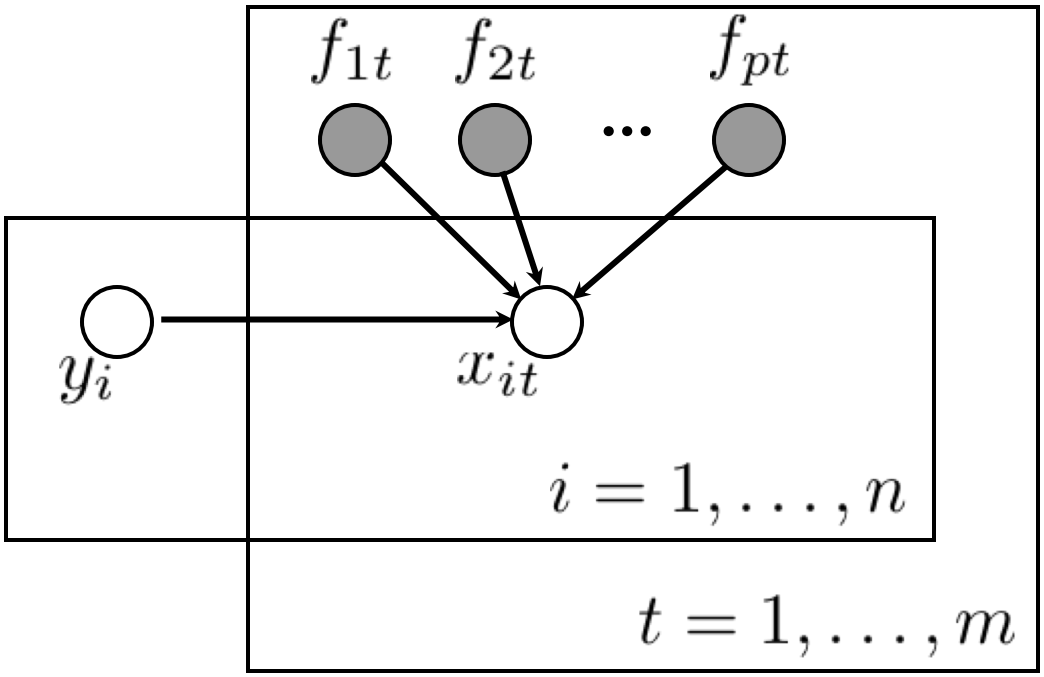



$$P(X|F,\Theta) = \prod_{i=1}^{n} \sum_{y_i=1}^{k} P(y_i) \prod_{t=1}^{m} P(x_{it}|f_{1t},\ldots,f_{pt},\Theta_{y_i})$$

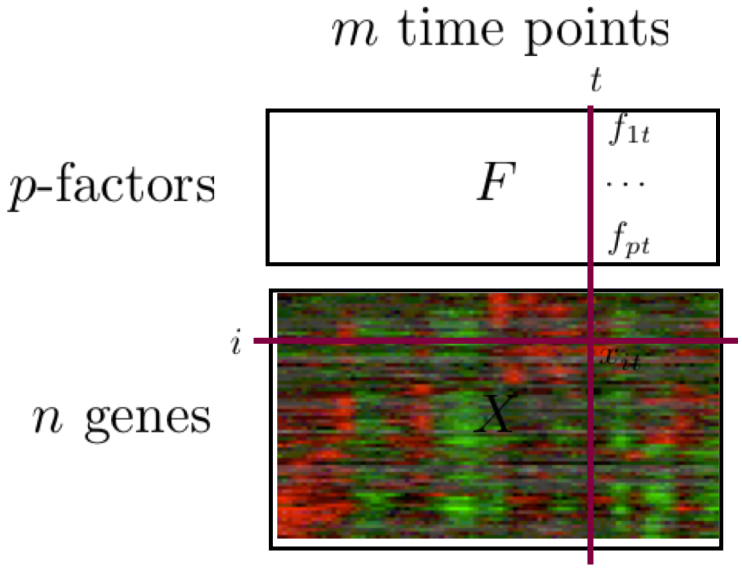
# A mixture of trees: sampling model

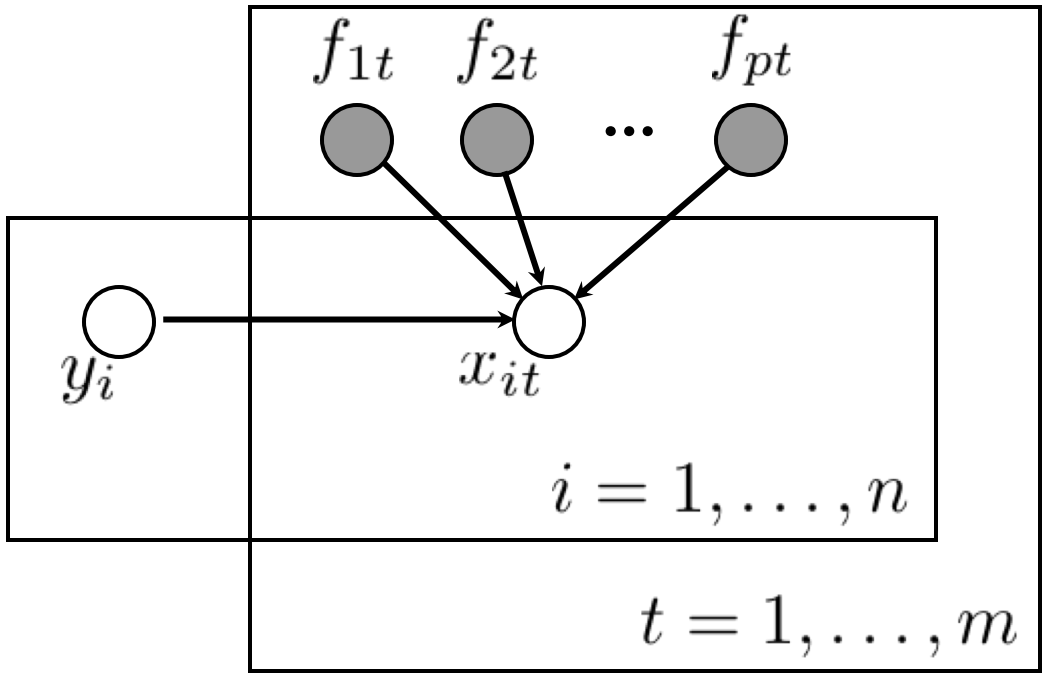



$$P(X|F,\Theta) = \prod_{i=1}^{n} \sum_{y_i=1}^{k} P(y_i) \prod_{t=1}^{m} P(x_{it}|f_{1t},\ldots,f_{pt},\Theta_{y_i})$$

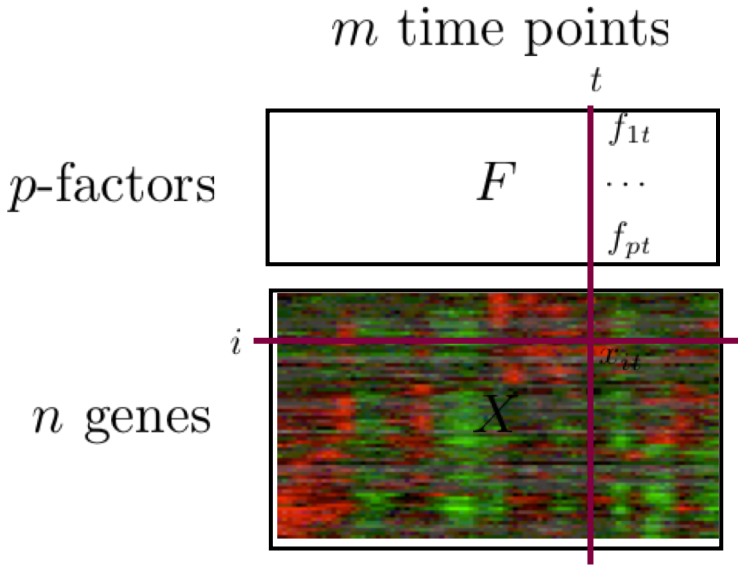
# A mixture of trees: sampling model

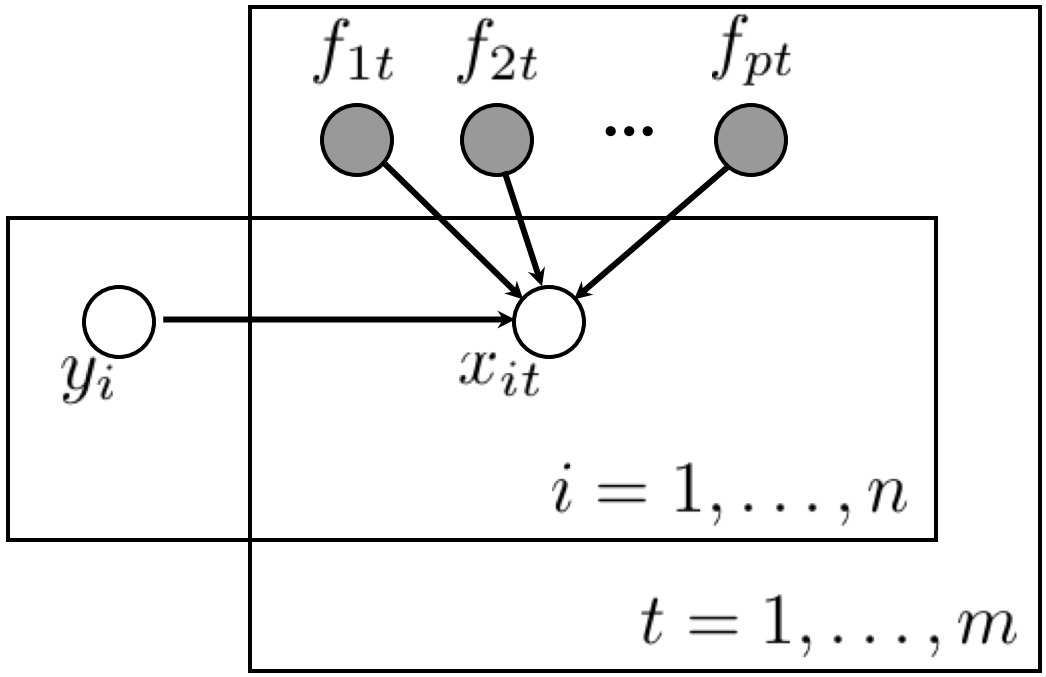



$$P(X|F,\Theta) = \prod_{i=1}^{n} \sum_{y_i=1}^{k} P(y_i) \prod_{t=1}^{m} P(x_{it}|f_{1t},\ldots,f_{pt},\Theta_{y_i})$$

# A mixture of trees: sampling model

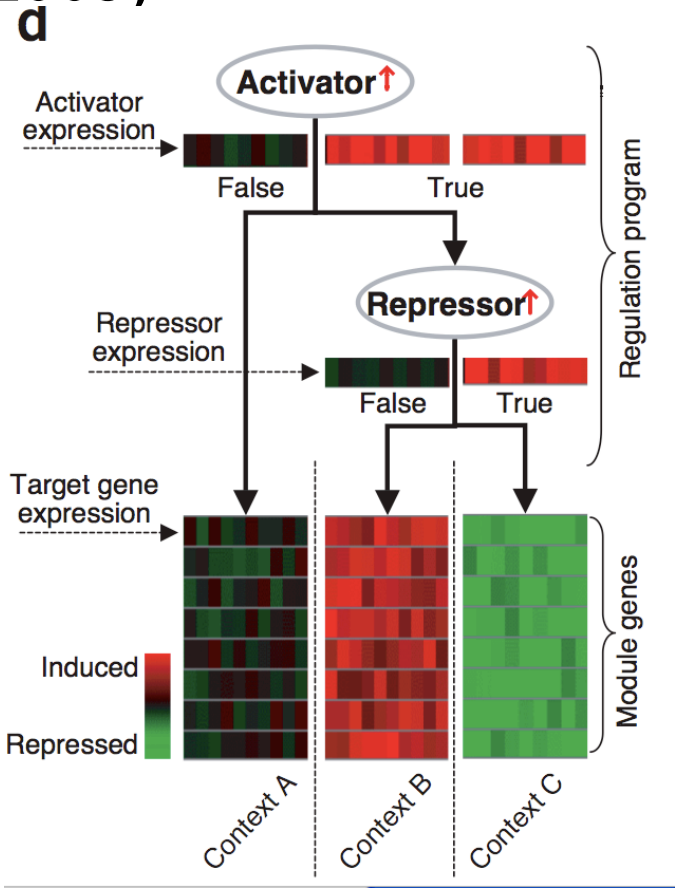



$$P(X|F,\Theta) = \prod_{i=1}^{n} \sum_{y_i=1}^{k} P(y_i) P(\underline{x}_i|F,\Theta_{y_i})$$

# Finding expression programs

- We can jointly search for transcriptionally regulated subsets of genes and the corresponding programs (Segal et al., 2003)



# EM algorithm

- We model the program assignments and expression profiles as

$$P(y_1, \ldots, y_n, \underline{x}_1, \ldots, \underline{x}_n | F, \Theta) = \prod_{i=1}^{n} P(y_i)\, P(\underline{x}_i | F, \Theta_{y_i})$$

- We can optimize the parameters of this model via the EM algorithm

(0) cluster genes into $k$ clusters
(1) re-estimate regression trees for each cluster (program)
(2) (softly) re-assign each gene to the best program

$$p(y = j|i) \propto P(y) P(\underline{x}_i | F, \Theta_y)$$

# Regression trees as programs

- Each regression tree can be represented as an "if-then-else" program

Cond. distribution $P(x|f_1, \ldots, f_p) = P(x|F)$

if $f_j \leq \tilde{f}_j$

$x \sim N(x; \mu_1, \sigma_1^2)$

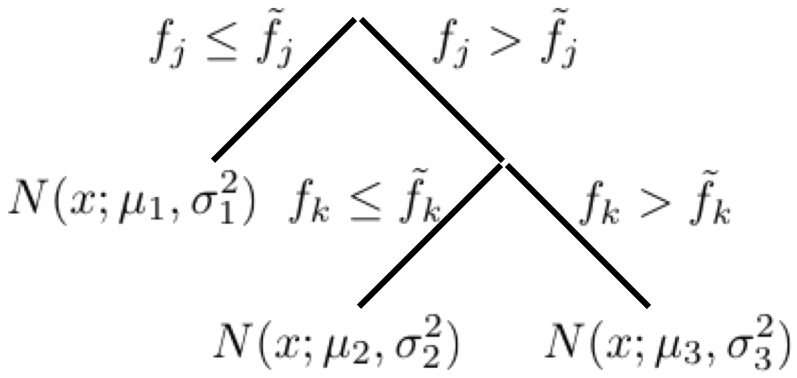
else

if $f_k \leq \tilde{f}_k$

$x \sim N(x; \mu_2, \sigma_2^2)$

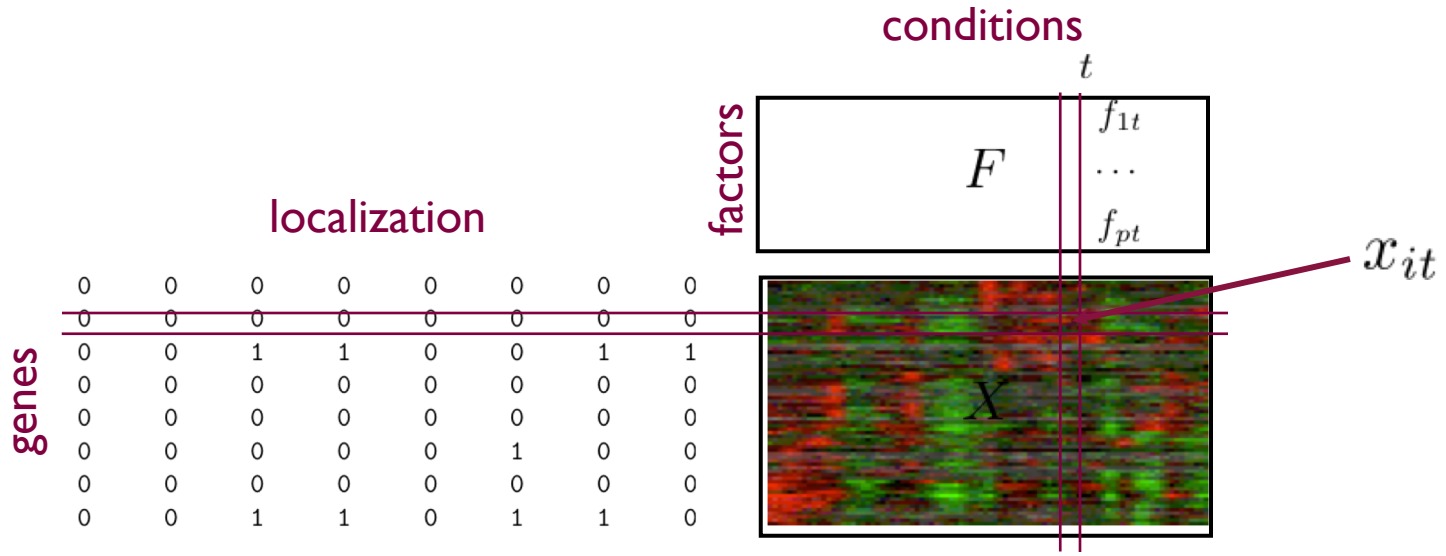
else

$x \sim N(x; \mu_3, \sigma_3^2)$



# Extensions...

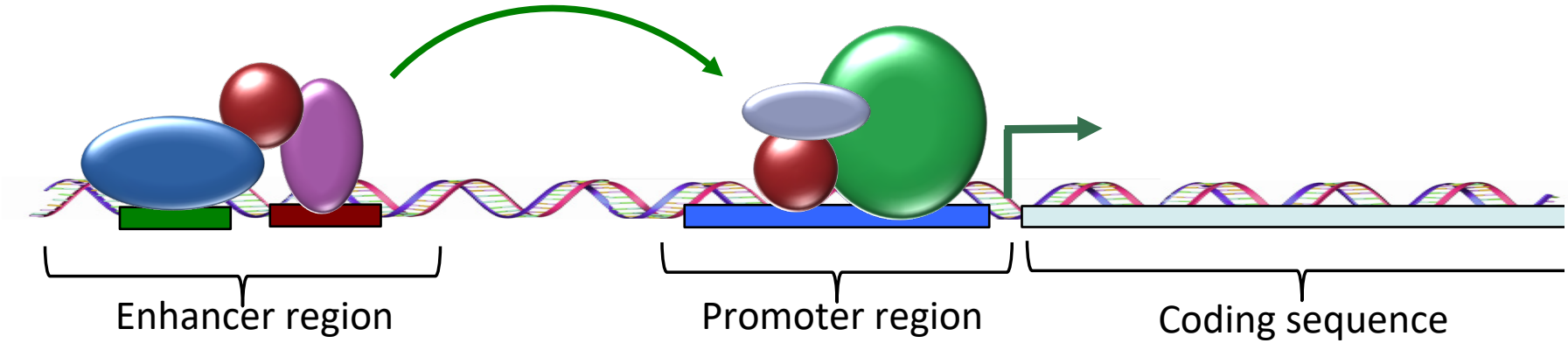
- The basic regression trees formulation permits us to estimate various types of "programs"
  - e,g, TF expression & localization, etc.



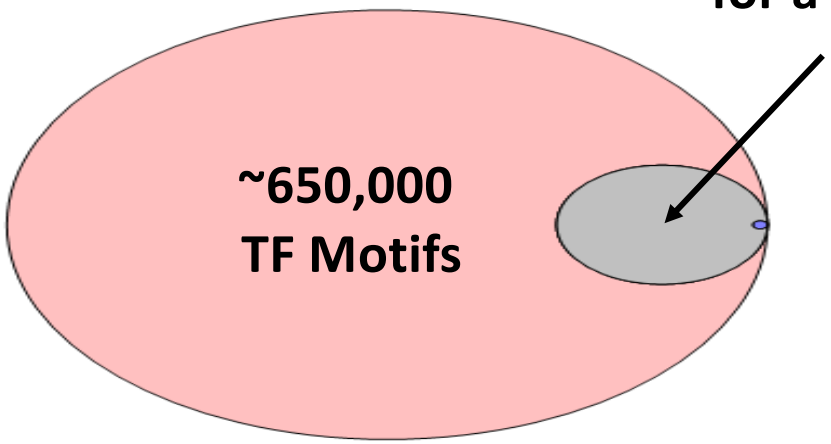
# Chromatin accessibility can reveal TF binding

Sherwood, RI, et al. "**Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape"** *Nat. Biotech 2014.*

# How do transcription factors control activation of cell-type-specific promoters and enhancers?

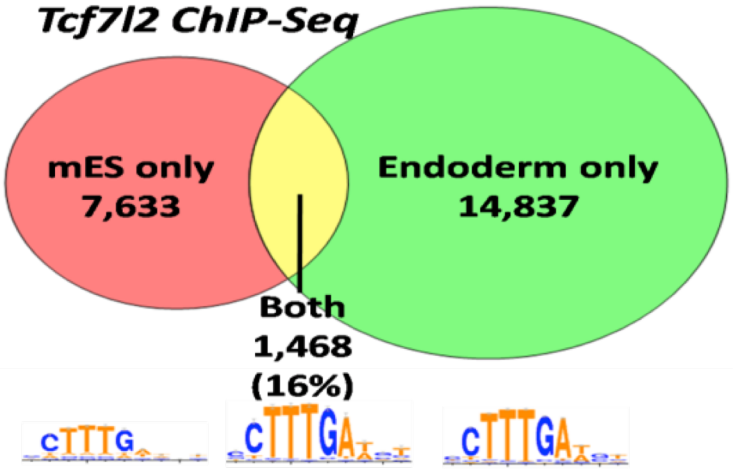


# Motifs are insufficient to predict TF binding

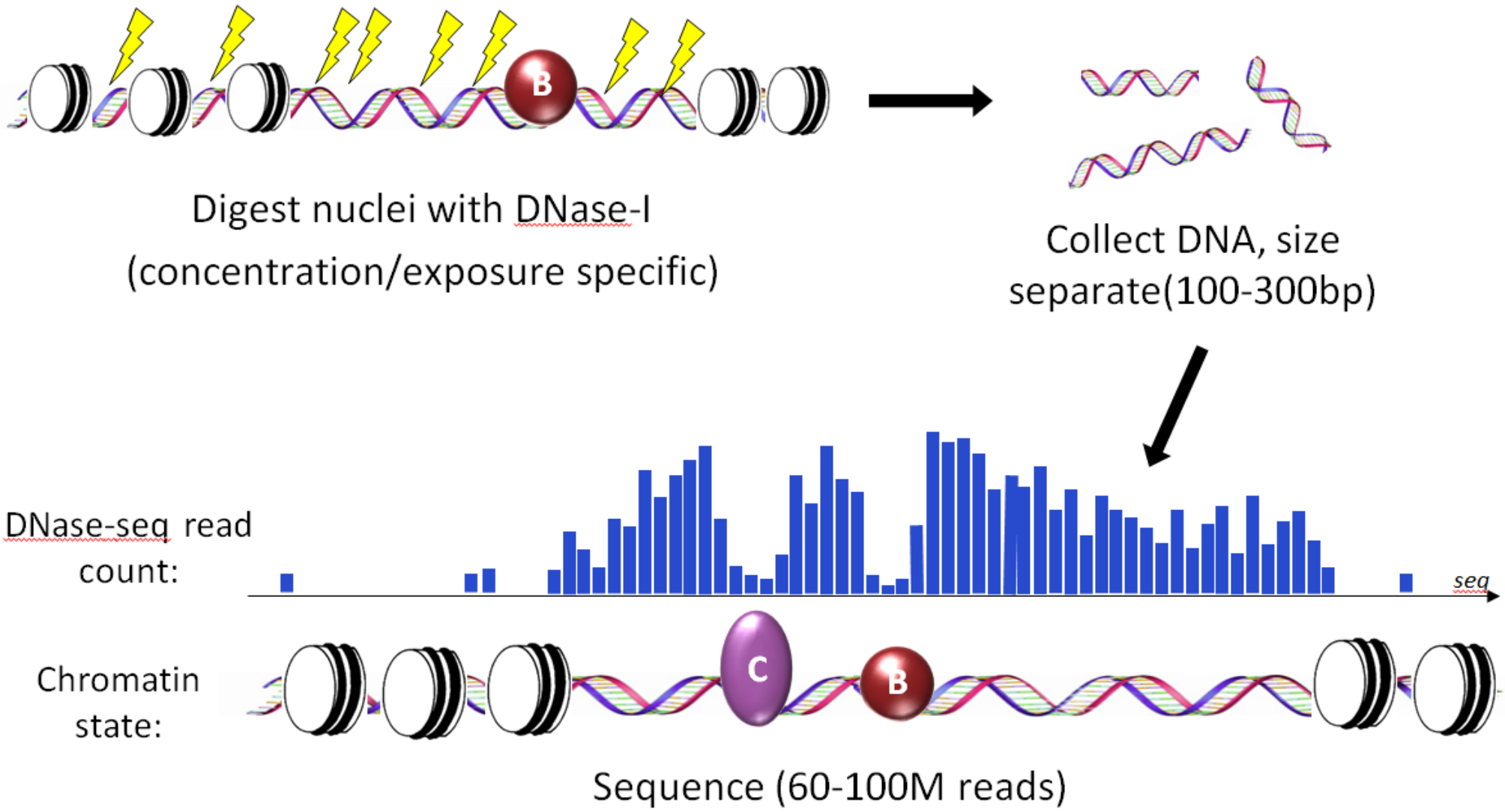



~50,000 binding sites for a typical TF

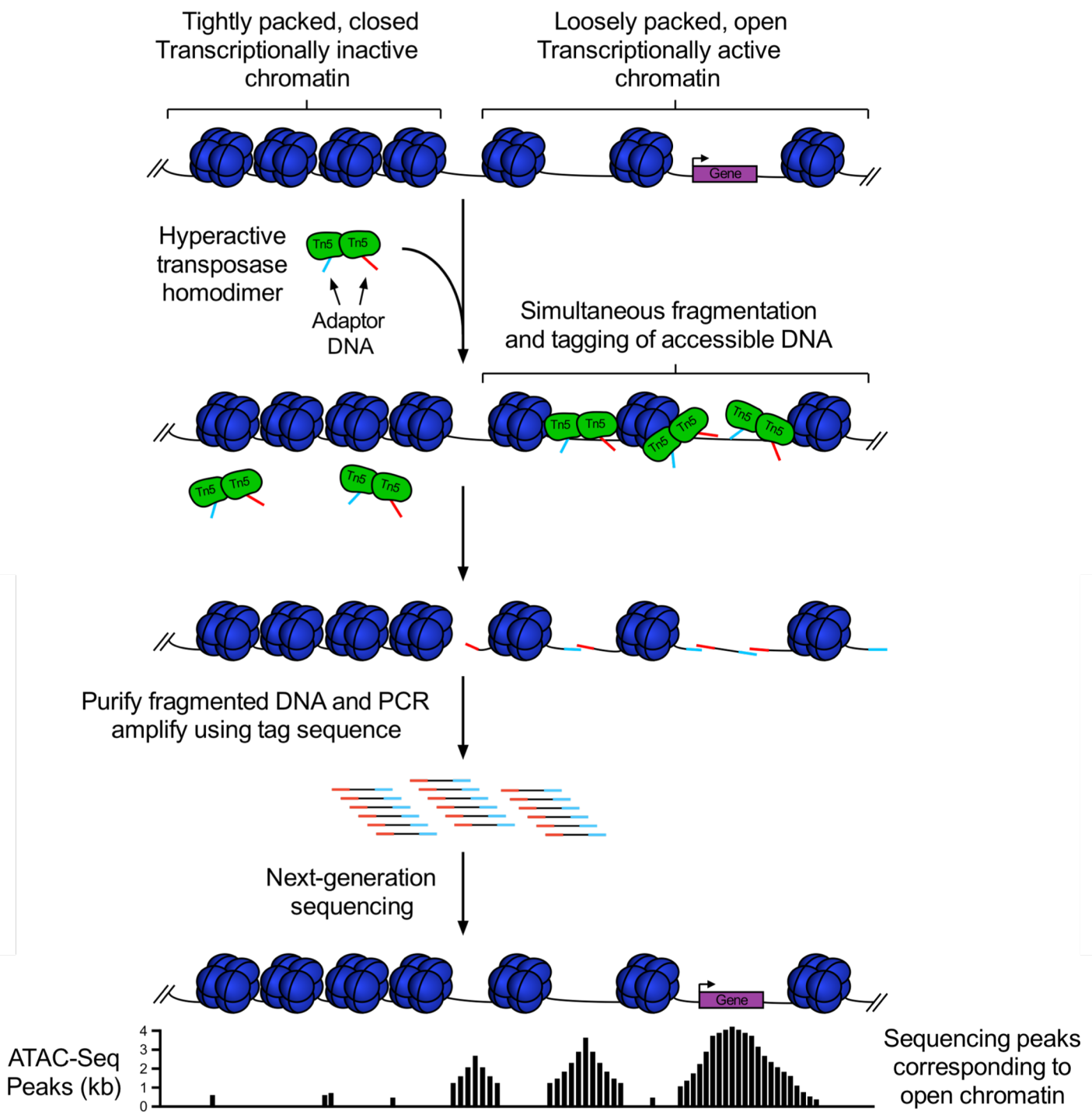
**Binding sites change across time**



# DNase-seq reveals genome protection profiles

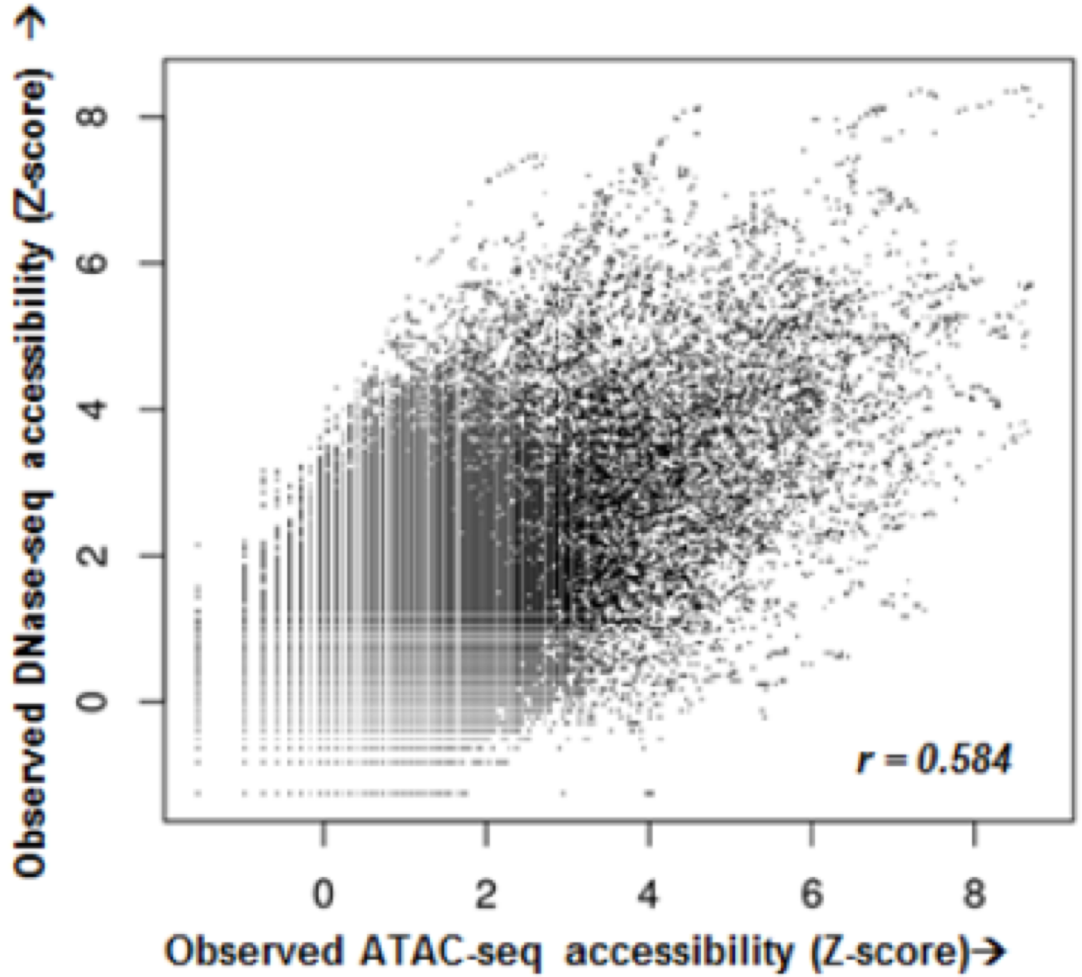


# ATAC-seq


Tightly packed, closed
Transcriptionally inactive
chromatin
Loosely packed, open
Transcriptionally active
chromatin
Gene
Hyperactive
transposase
homodimer
Tn5
Tn5
Adaptor
DNA
Simultaneous fragmentation
and tagging of accessible DNA
Purify fragmented DNA and PCR
amplify using tag sequence
Next-generation
sequencing
Gene
ATAC-Seq
Peaks (kb)
4
3
2
1
0
Sequencing peaks
corresponding to
open chromatin

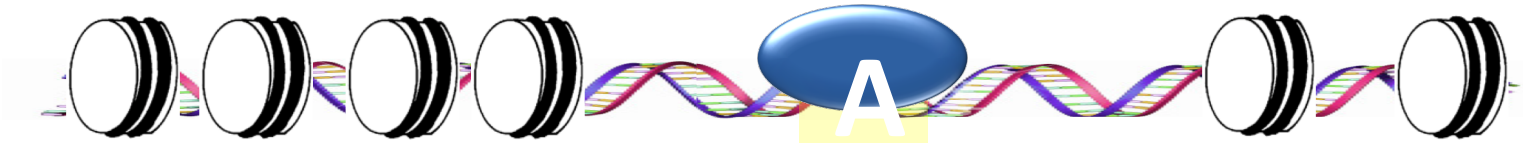
# ATAC-seq and DNase-seq are not identical

GM12878, Chr. 14,
Each point is accessibility in a 2 kb window

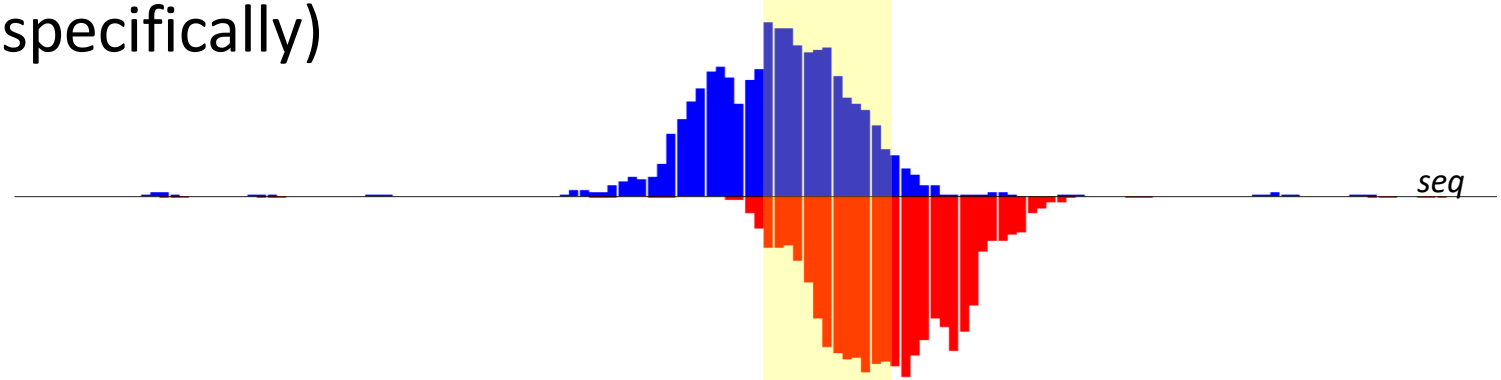



Hashimoto TB, et al. "**A Synergistic DNA Logic Predicts Genome-wide Chromatin Accessibility"** *Genome Research* 2016

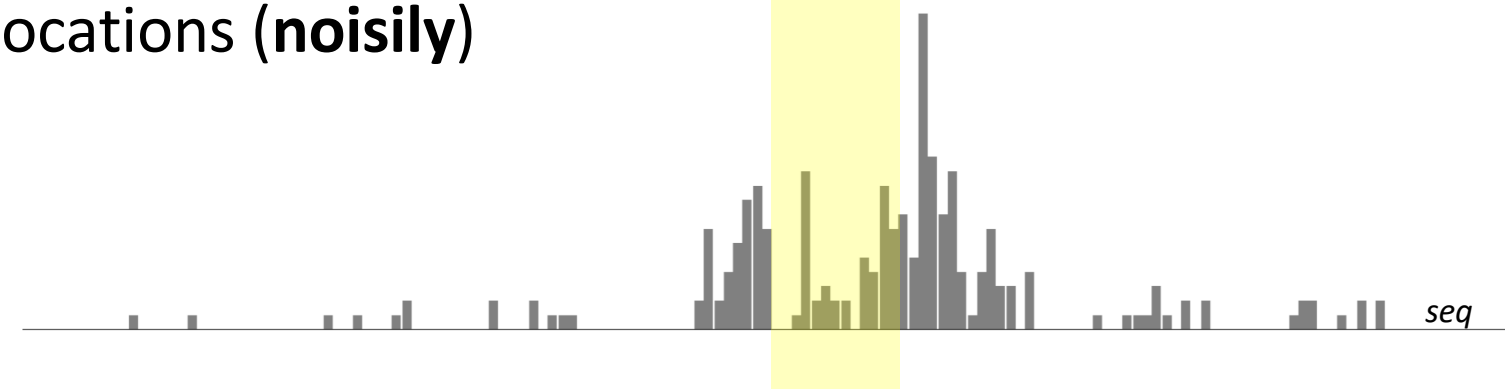
# Dnase-seq is less defined evidence than ChIP-seq



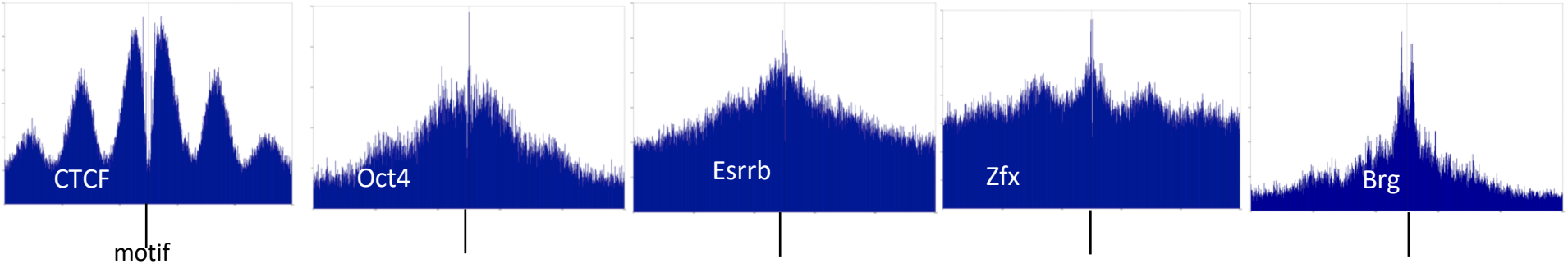

ChIP-seq reports **TF-binding** locations regions (specifically)




DNase-seq reports proximal **TF-non-binding** locations (**noisily**)

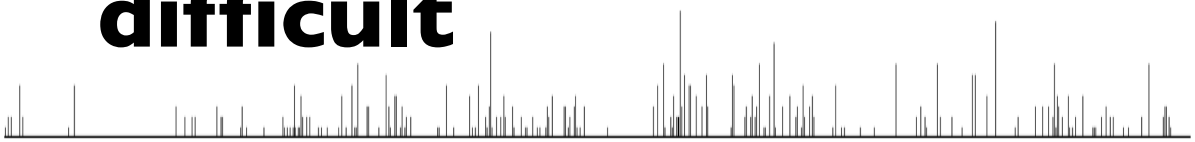


# Bound factors leave distinct DNase-seq profiles

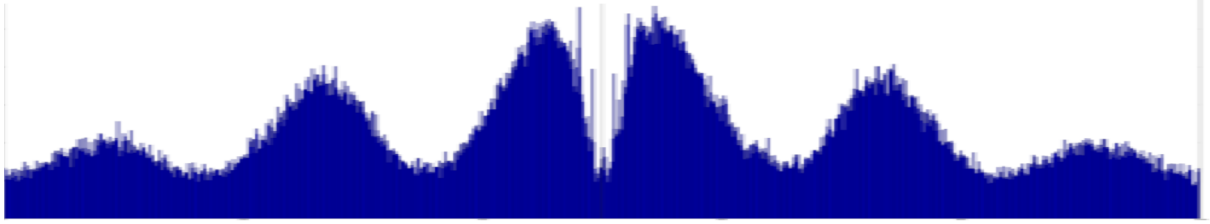



# Individual binding site prediction is difficult

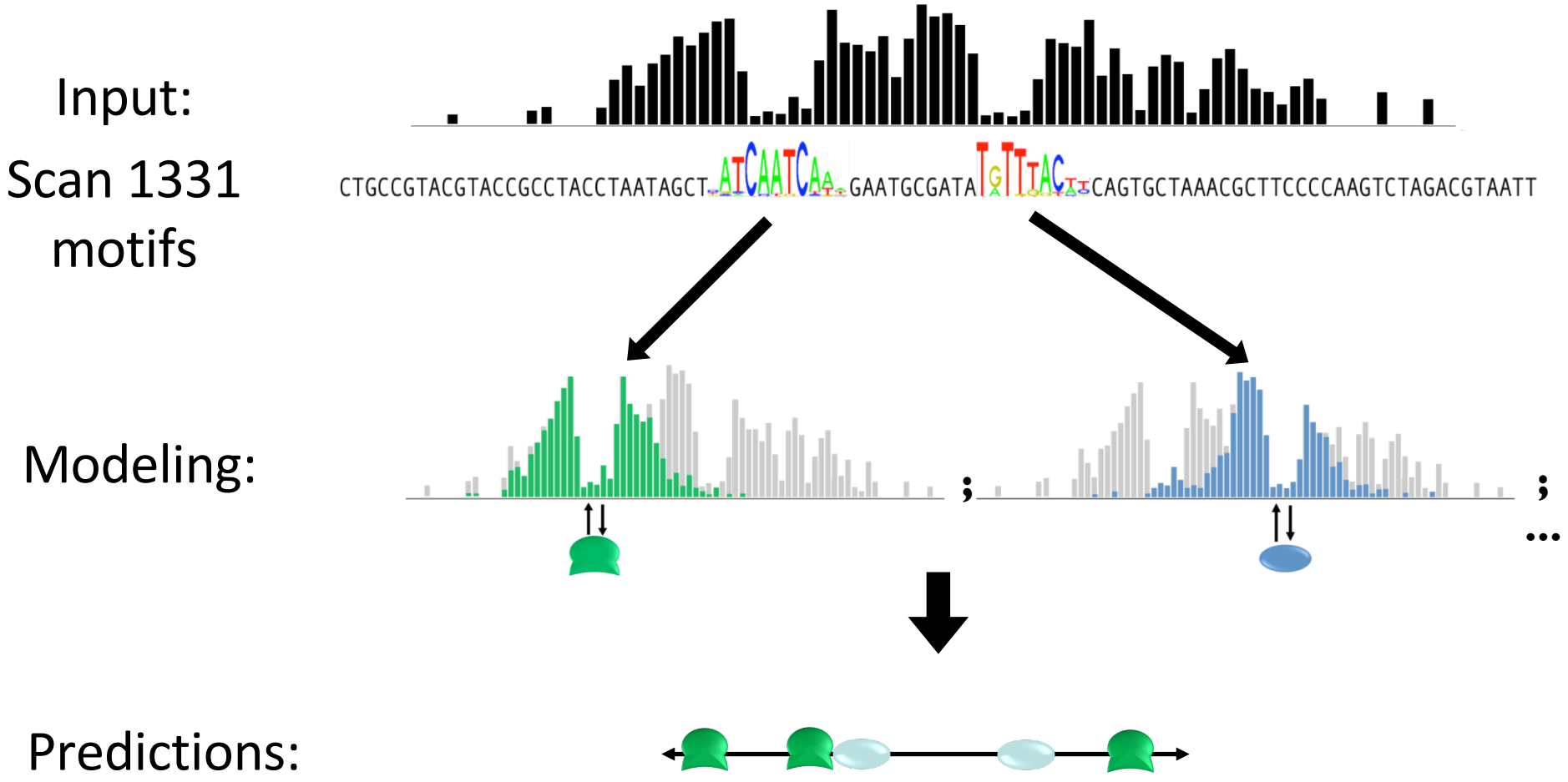
Individual CTCF:



Aggregate CTCF:

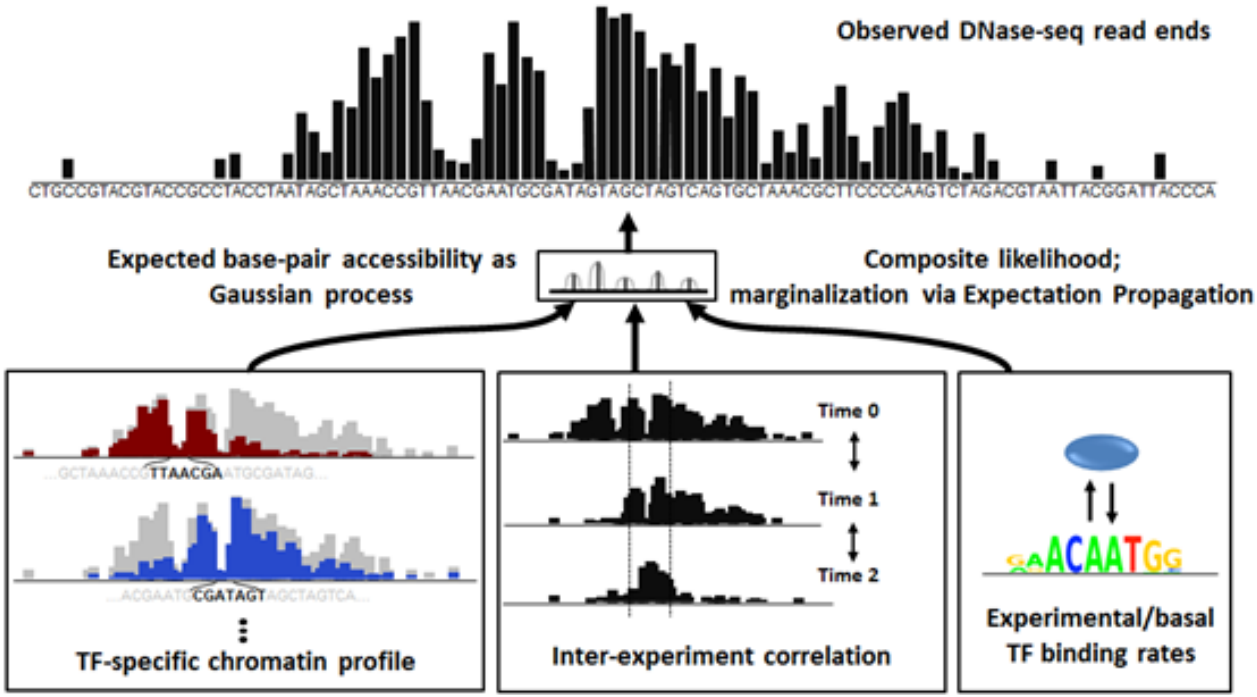


# *PIQ*: algorithm to predictively model TF binding from DNase-seq + Sequence



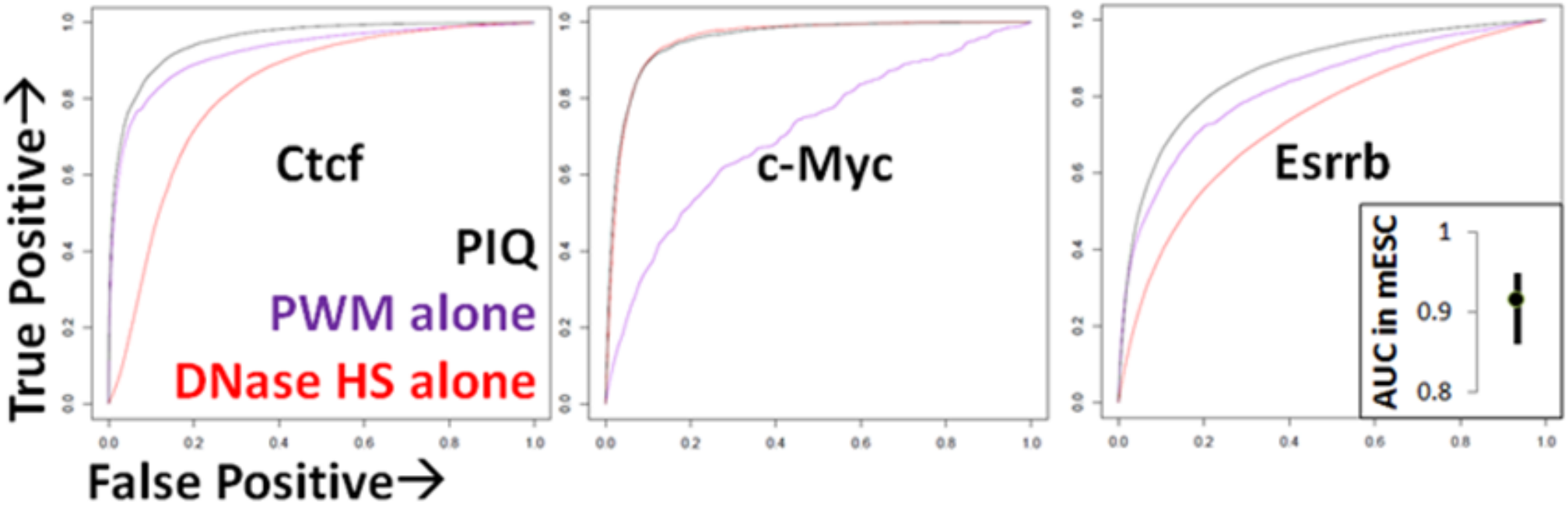
# PIQ weighs evidence to compute binding

1. Robust model of TF binding that models overlapping footprints
2. Model of the unoccupied genome using a Gaussian Process that captures inter-experiment and base correlations.
3. Better motif models that captures nonlinear PWM to binding effects.



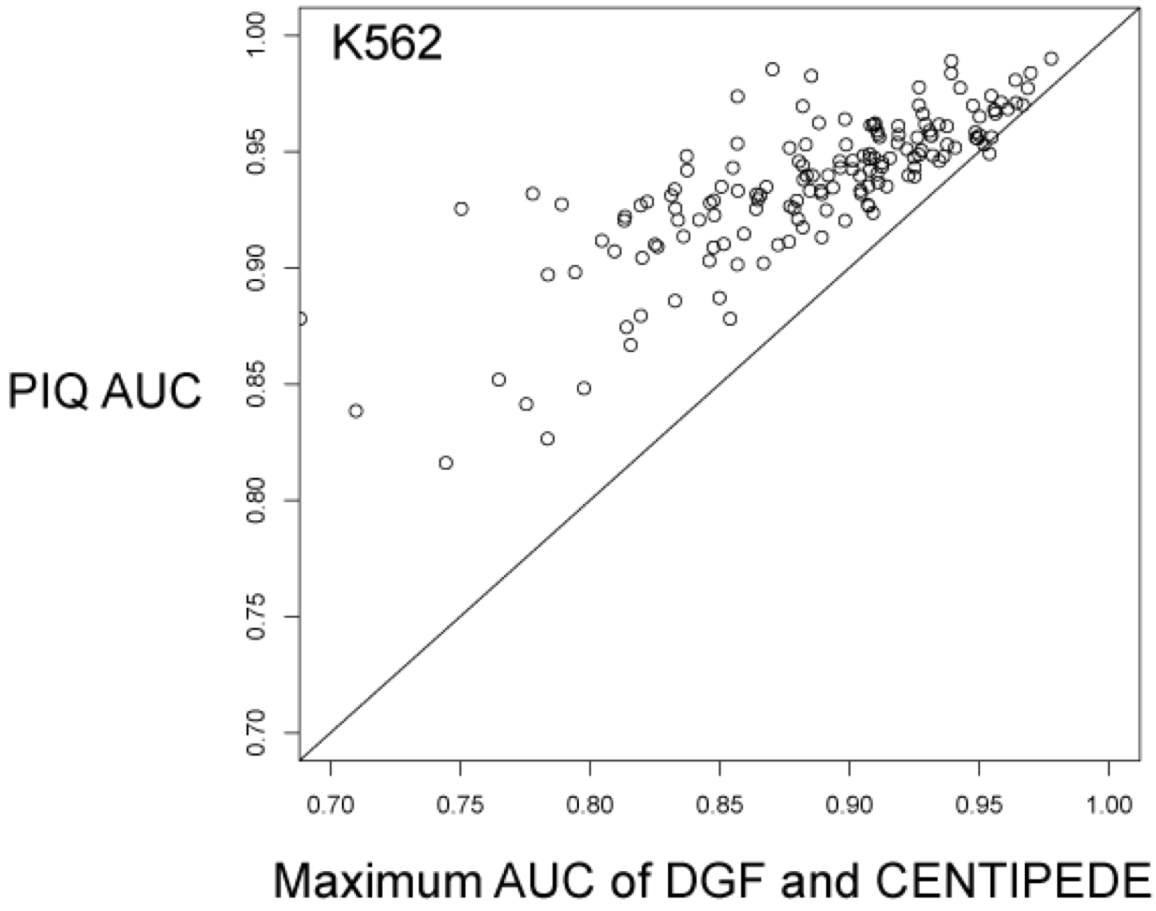
# PIQ accurately predicts mESC TF binding

Receiver operating characteristic (ROC) curves show PIQ matches closely with ChIP-seq data.



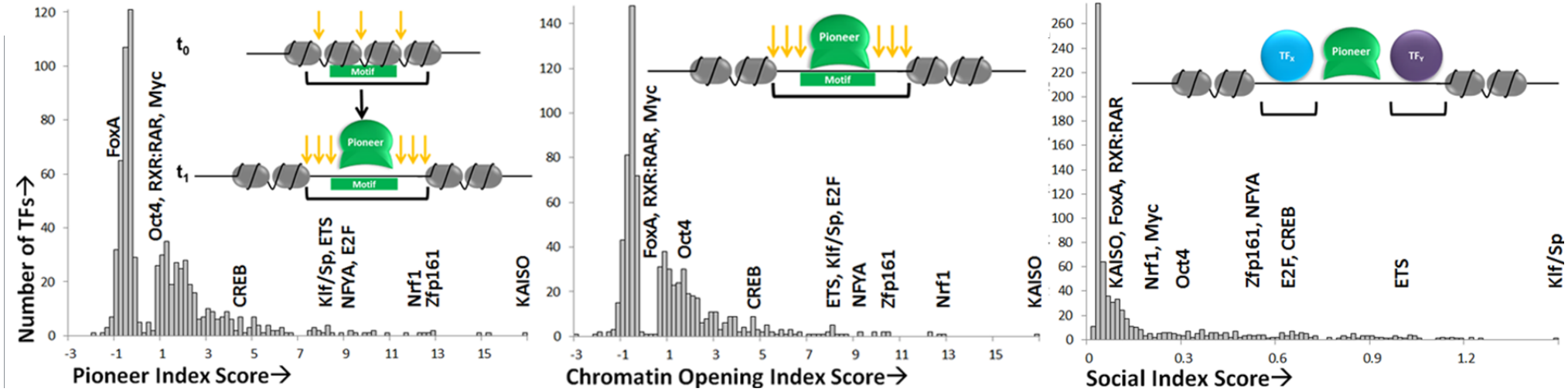
# PIQ outperforms existing methods when predicting binding for 313 ENCODE ChIP-seq experiments

PIQ ( .93 Mean AUC); Centipede (.87); DGF (.65)

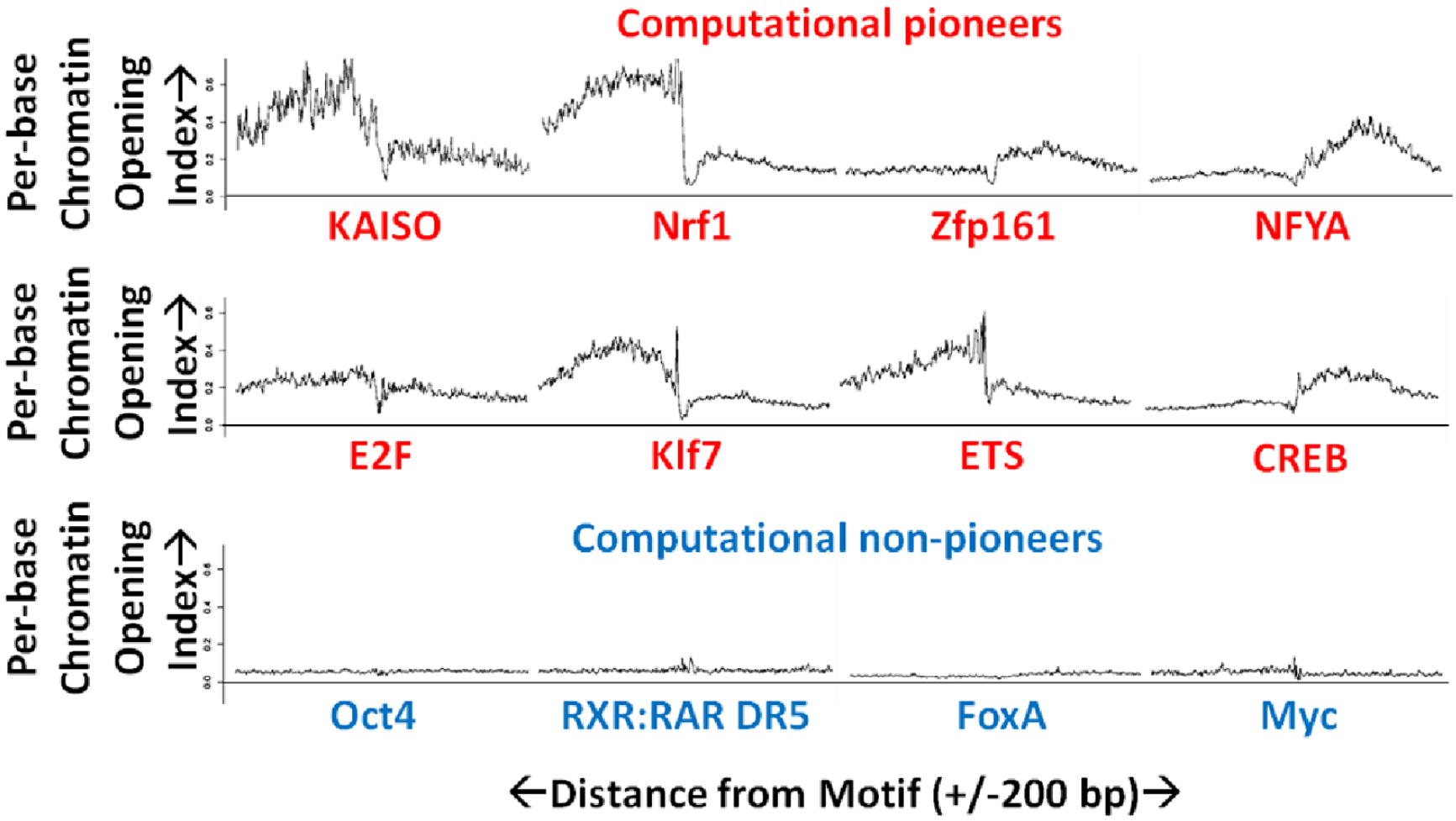


# Pioneer factors are distinct from other factors and regulate proximal chromatin and binding

Three separate metrics (differential chromatin, static chromatin, cobinding) show several factors that are consistent pioneers.

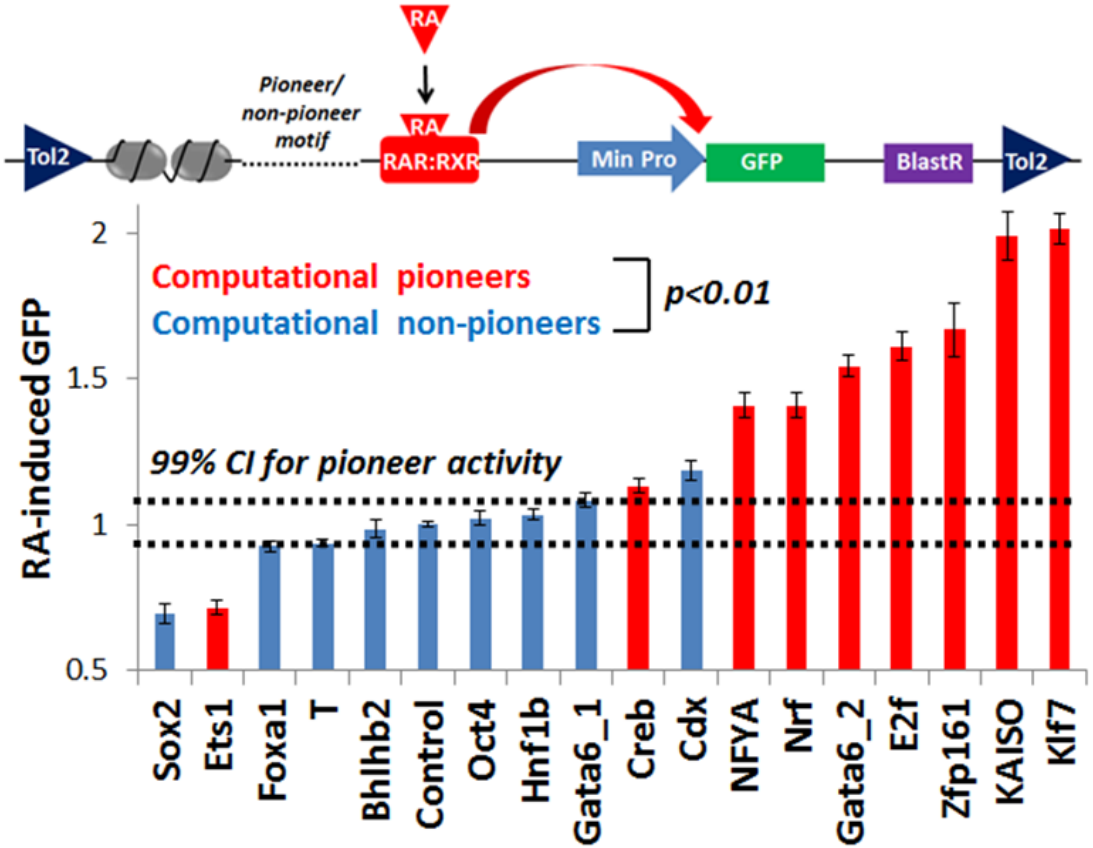


# Pioneer TFs have identifiable profiles

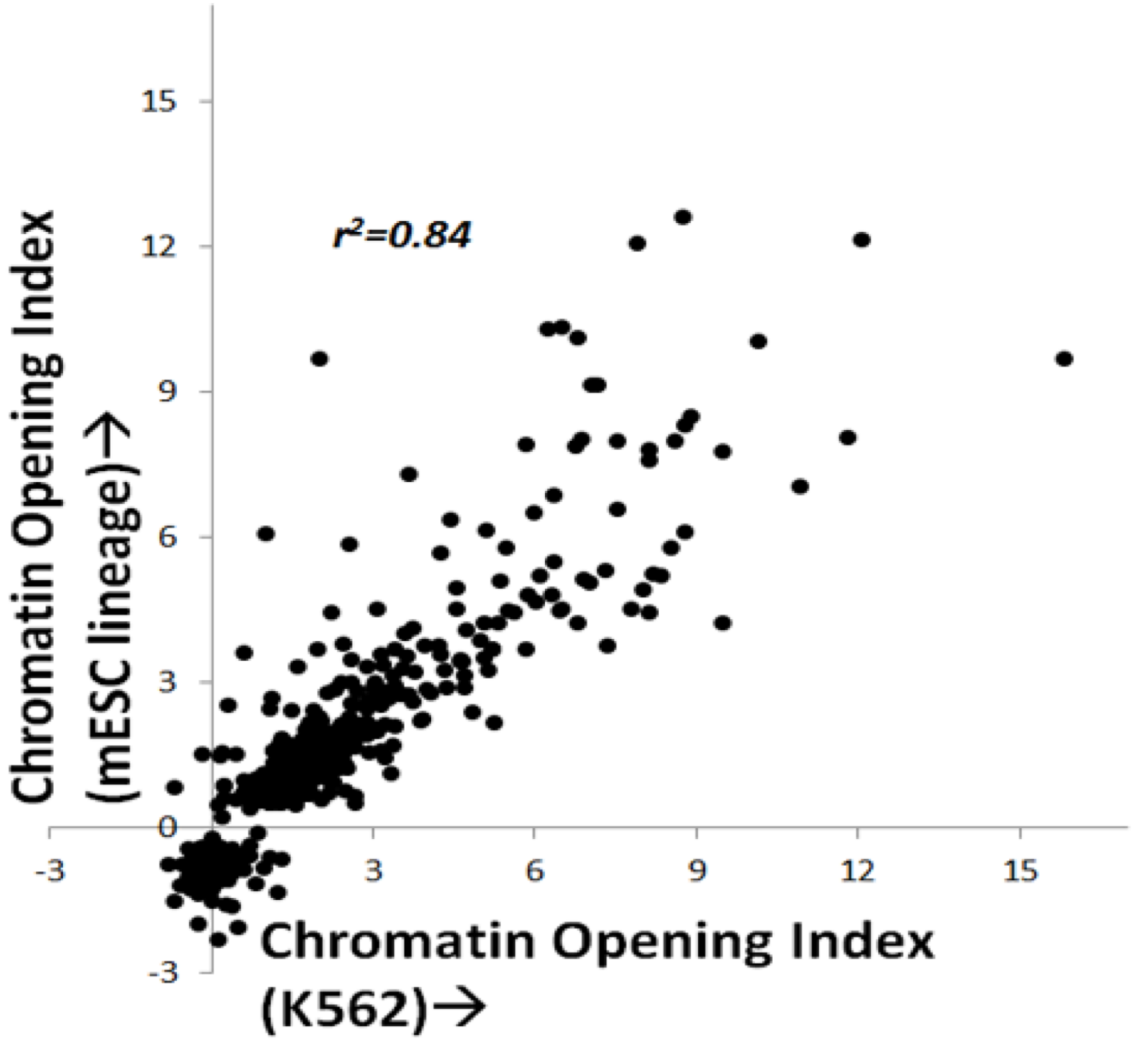


# In vitro reporter assays recapitulate computational predictions

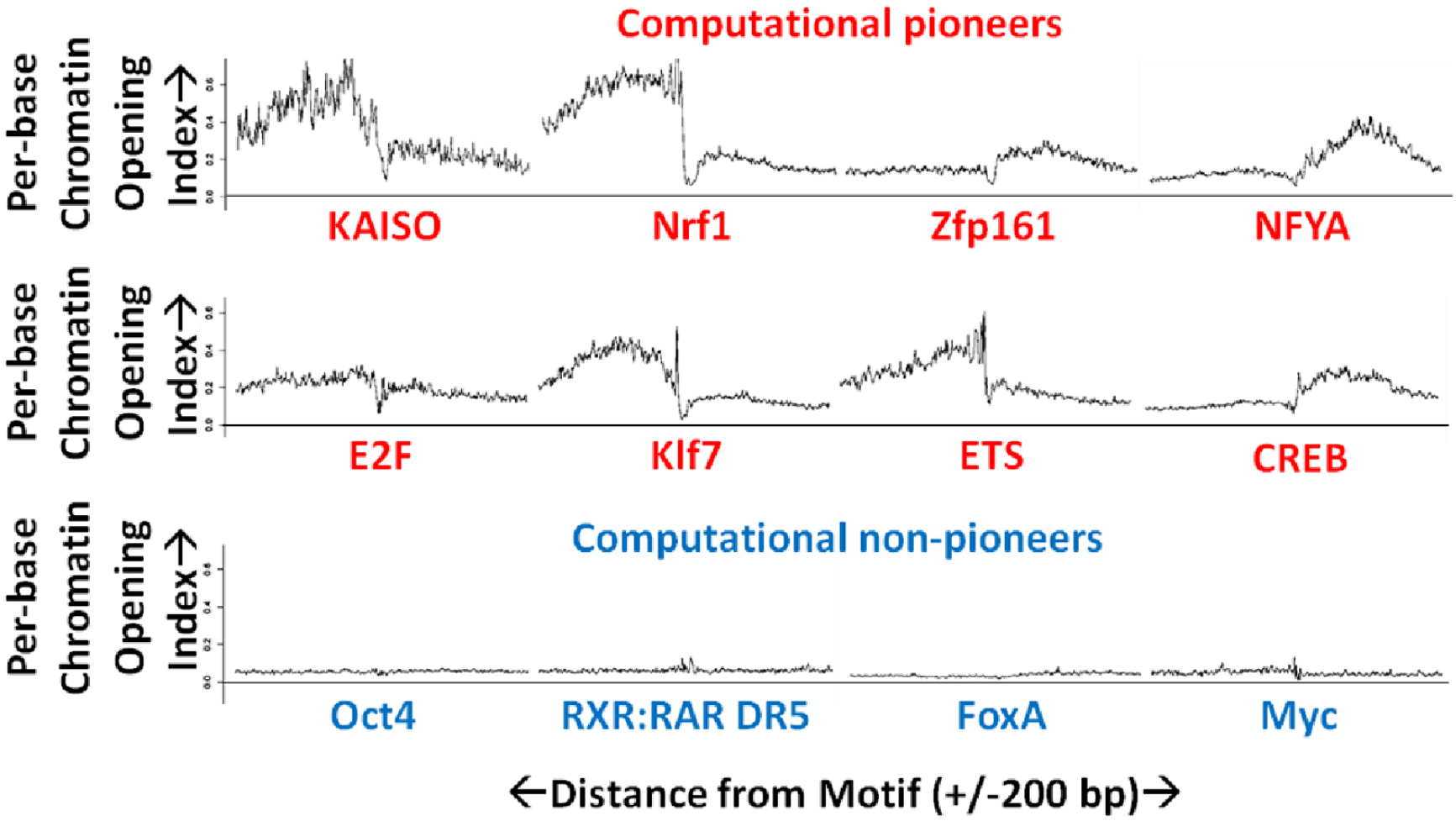
Using a Tol2 based GFP reporter, we confirm finding that these pioneers create new enhancers.



# Pioneers appear to be conserved between human/mouse

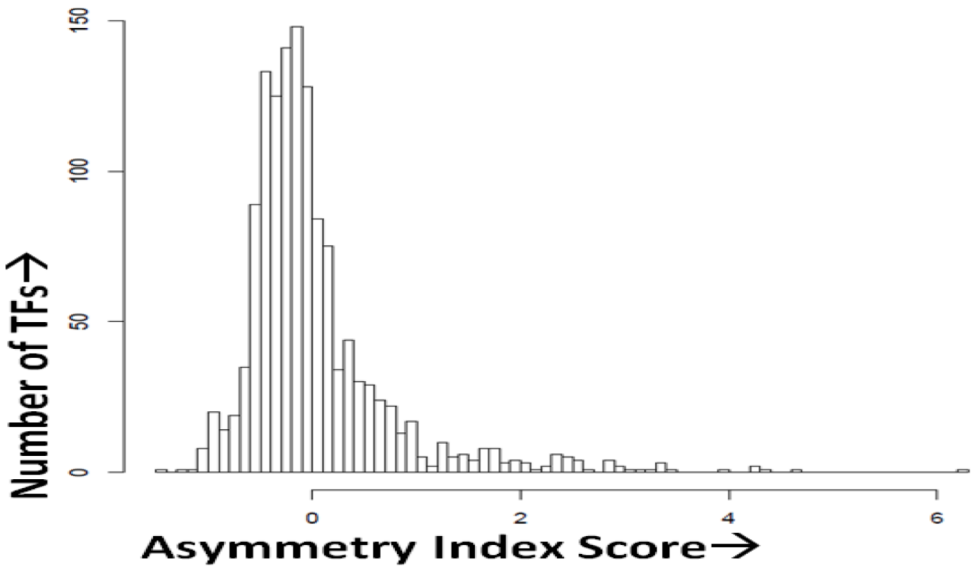


# Pioneer TFs have identifiable profiles

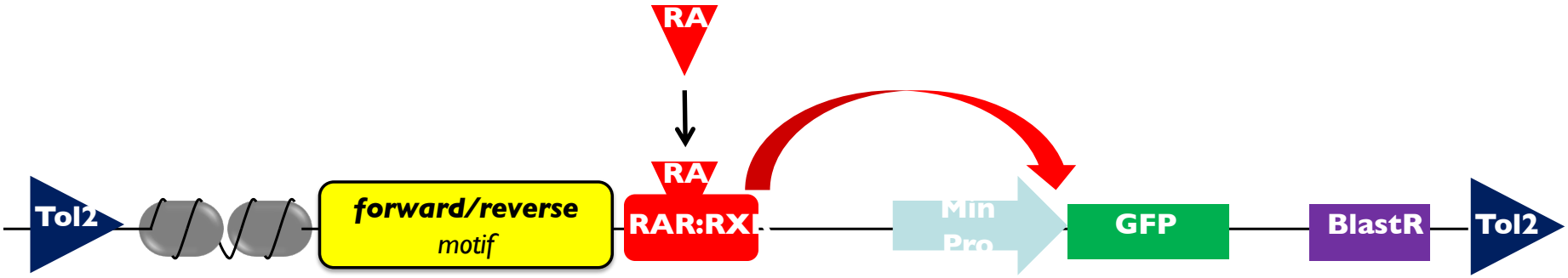


# Certain pioneer TFs are directional

- We define asymmetry index as the expected change between left and right sides in (squared) chromatin opening index score

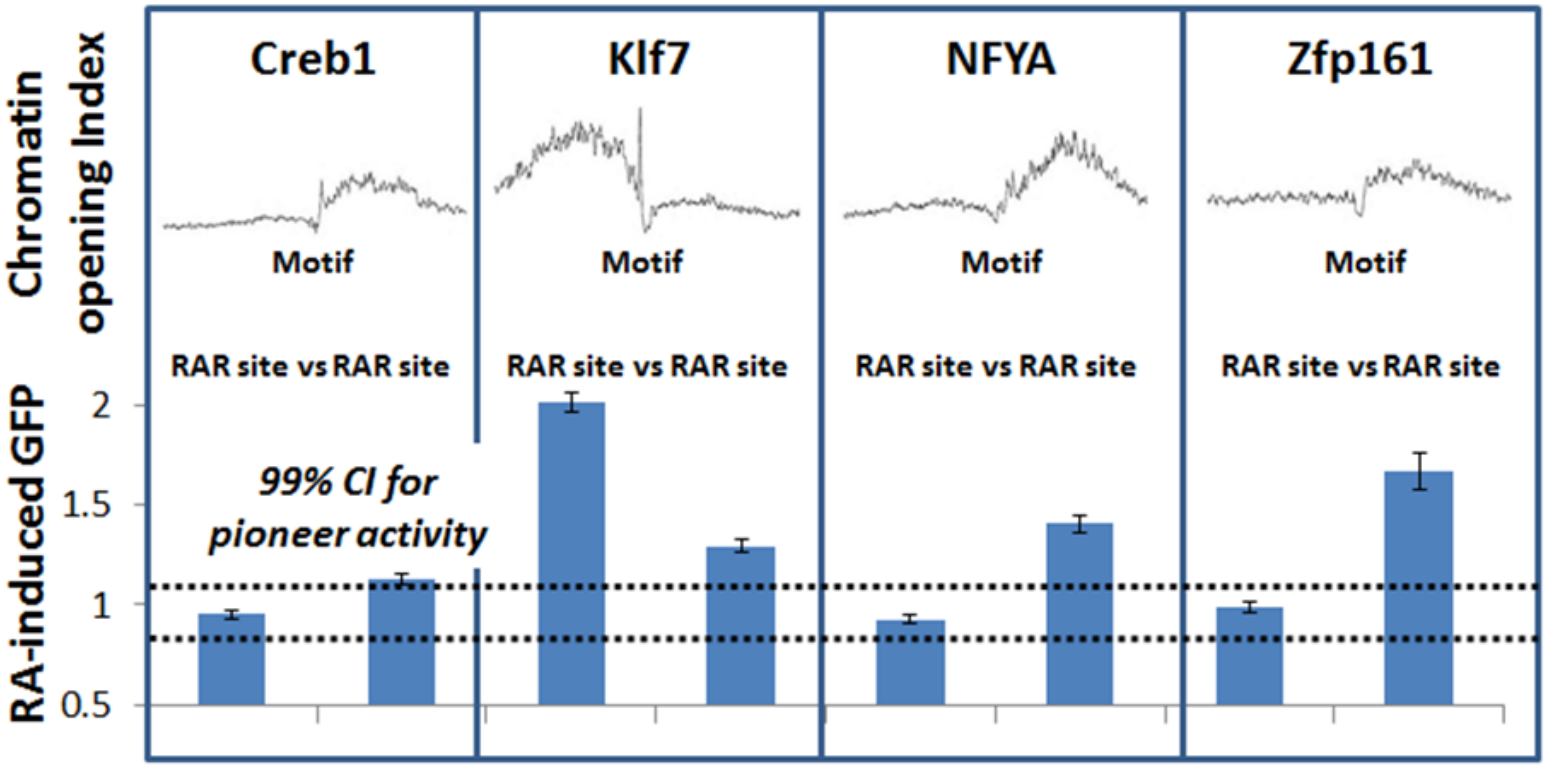



- Biological validation by testing both motif orientations



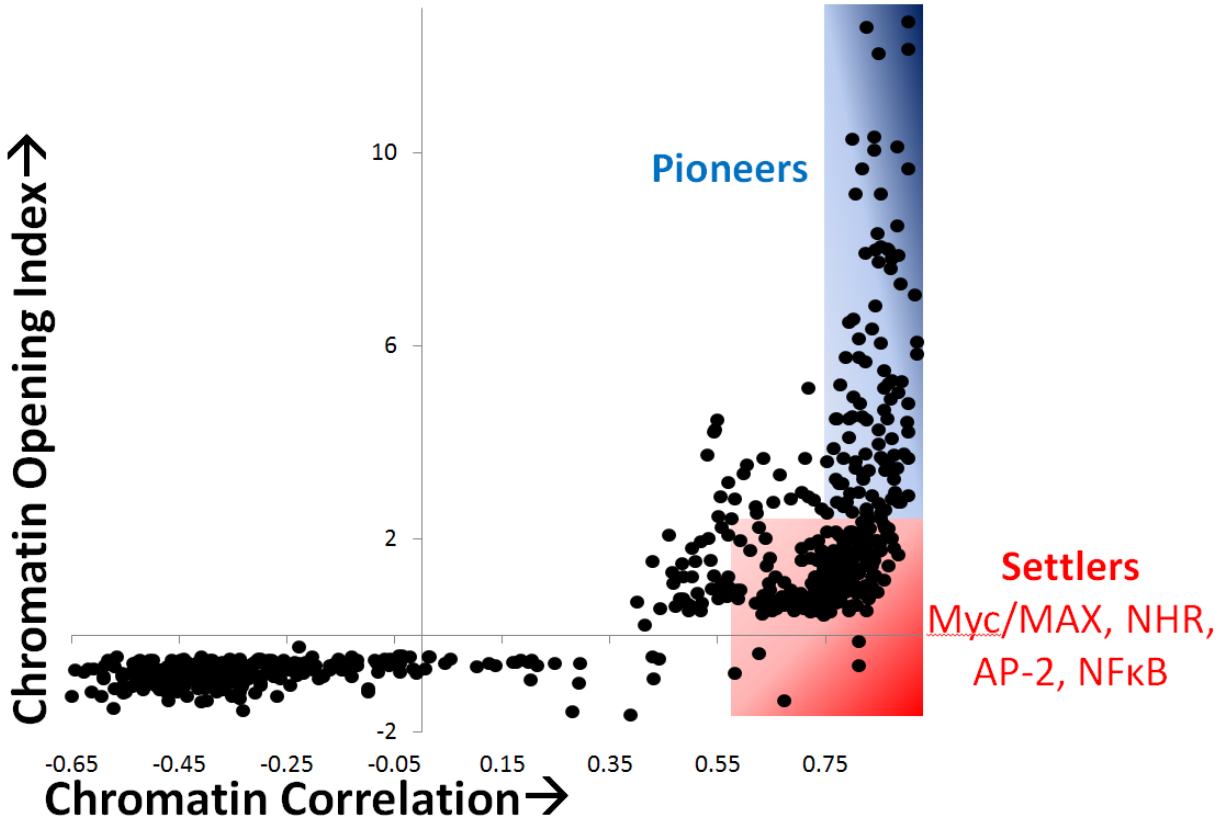
# Certain pioneer TFs are directional

Orienting the motif direction in the reporter recapitulates expected directional behaviors.

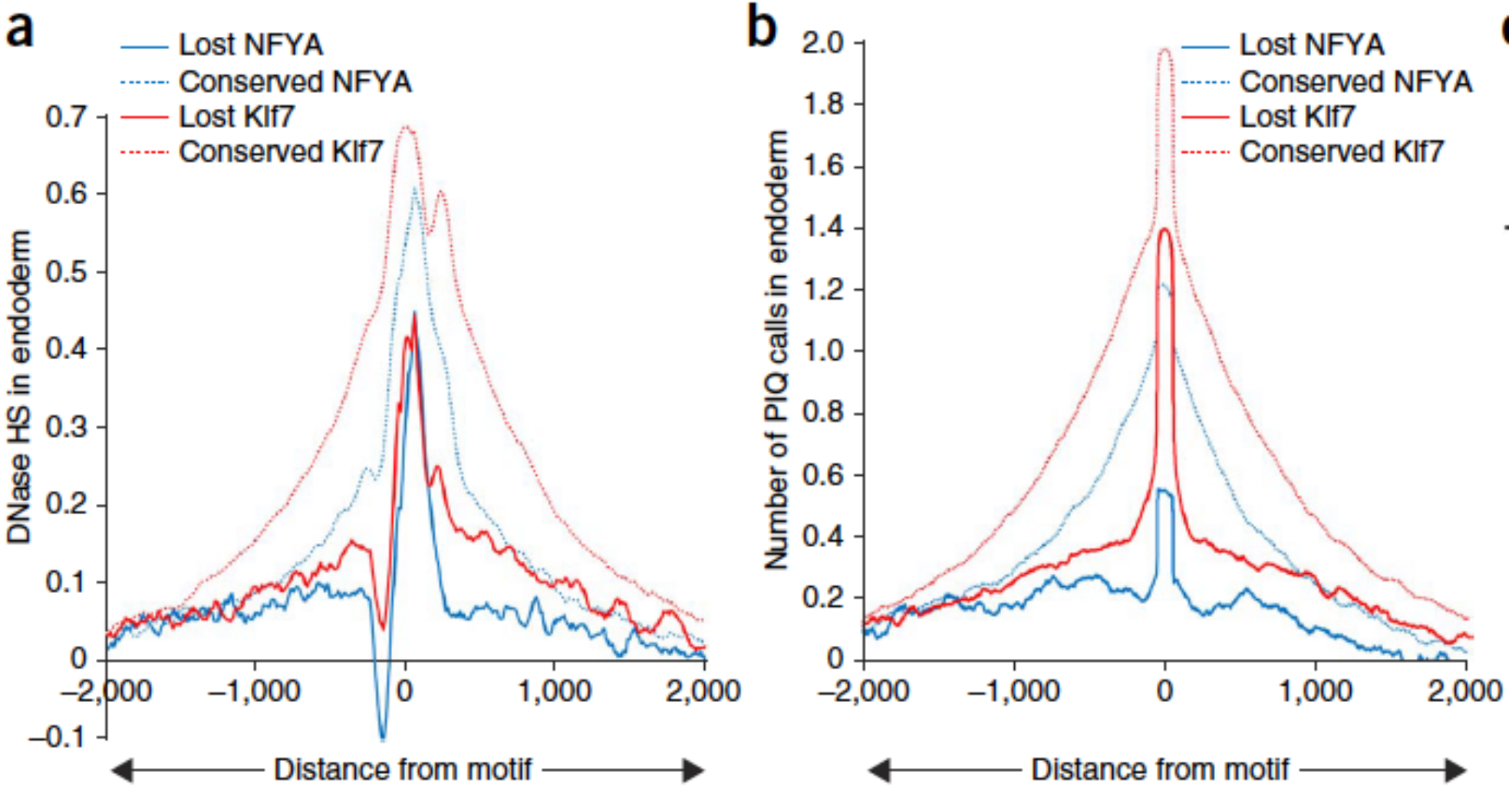


# Settler factors follow pioneer factor binding and loss of pioneer binding causes chromatin to return to a closed state

Pioneers (chromatin opening and dependent) are rare and distinct, while there exists a class of chromatin dependent, but non-opening factors.

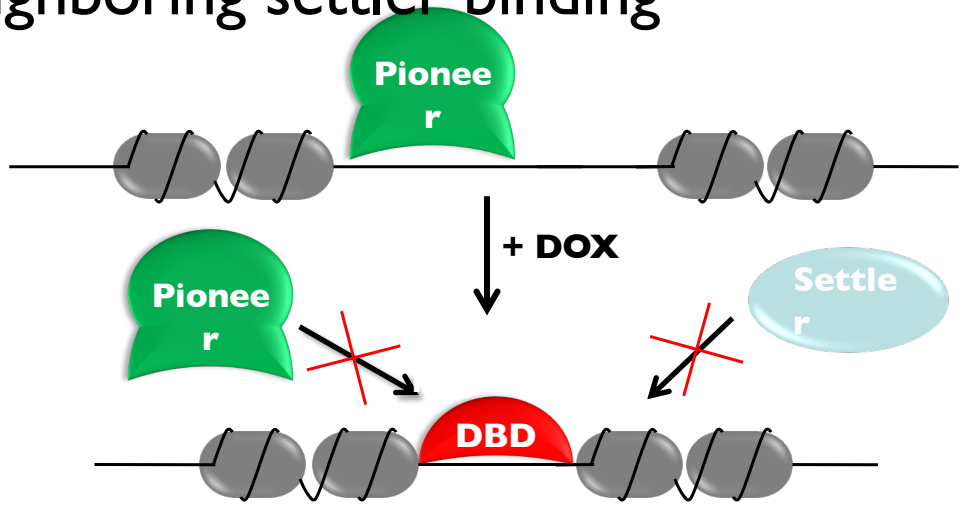


# Loss of pioneer binding causes chromatin to return to closed states



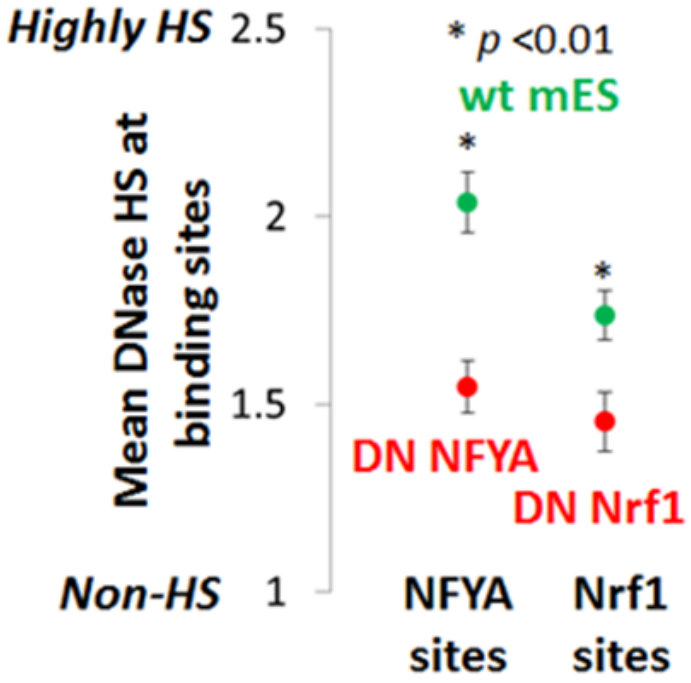
# Validate pioneer/settler model via dominant-negative competition assay

- Construct pioneer DBD protein that retains no pioneering function
- Induction of DBD protein competes for genomic binding, reducing local chromatin accessibility settlers rely on
- Compare proximal chromatin openness
- Compare ChIP levels for neighboring settler binding

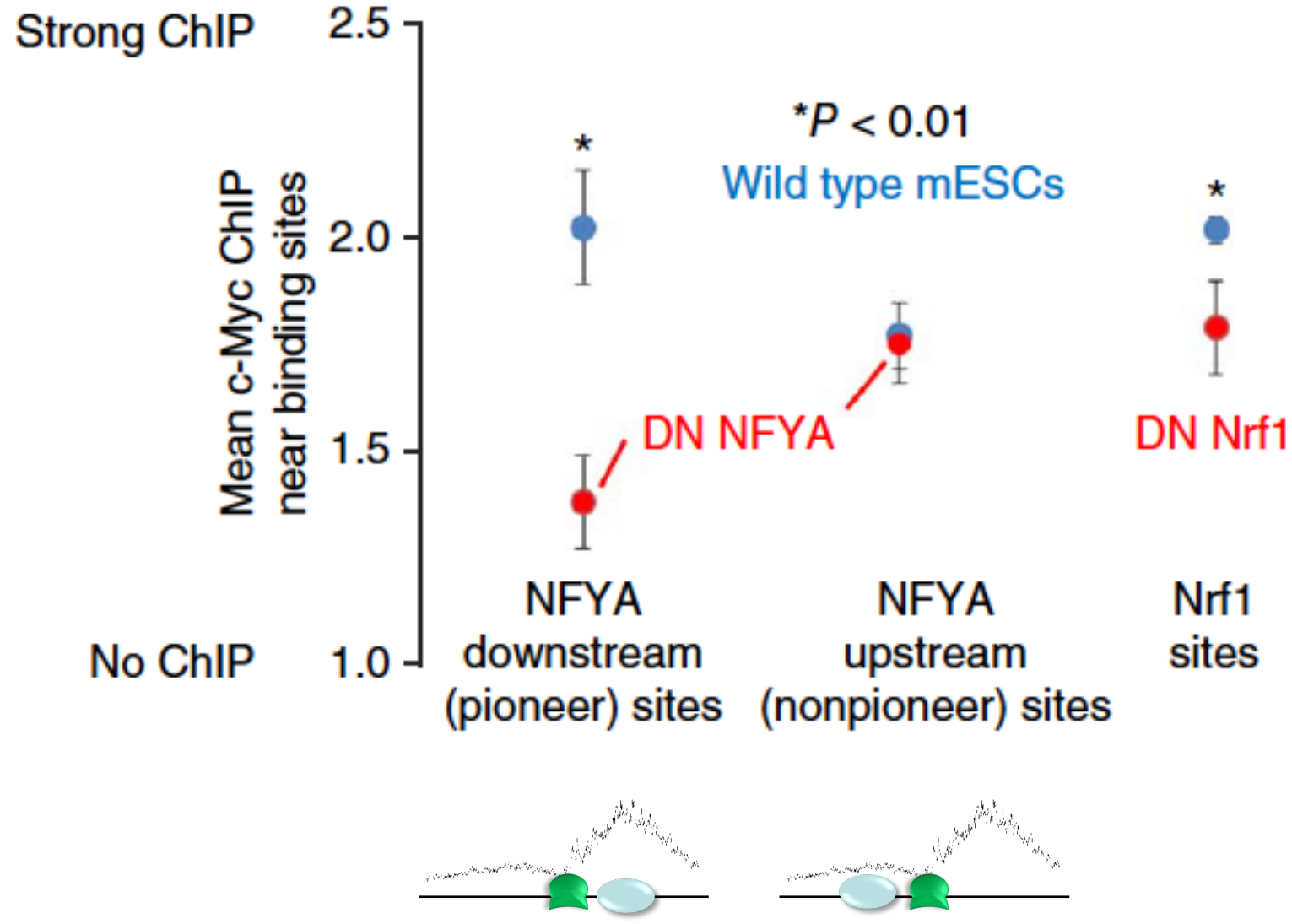


# Dominant negative pioneers reduce proximal DNase HS

We created dominant negative versions of the NFYA and Nrf1 pioneers and measured DNase accessibility at native NFYA and Nrf1 sites after induction of dominant negatives.



# Dominant negative pioneers reduce proximal binding of c-Myc



# Chromatin accessibly influences transcription factor binding

- Modeling accessibility profiles yields binding predictions and pioneer factor discovery
- Asymmetric accessibility is induced by *directional pioneers*
- The binding of *settler factors* can be enabled by proximal pioneer factor binding

Sherwood, RI, et al. "**Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape"** *Nat. Biotech 2014.*

# How genome sequence determines cell-type specific chromatin accessibility

# Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks.

David R. Kelley
Jasper Snoek
John L. Rinn

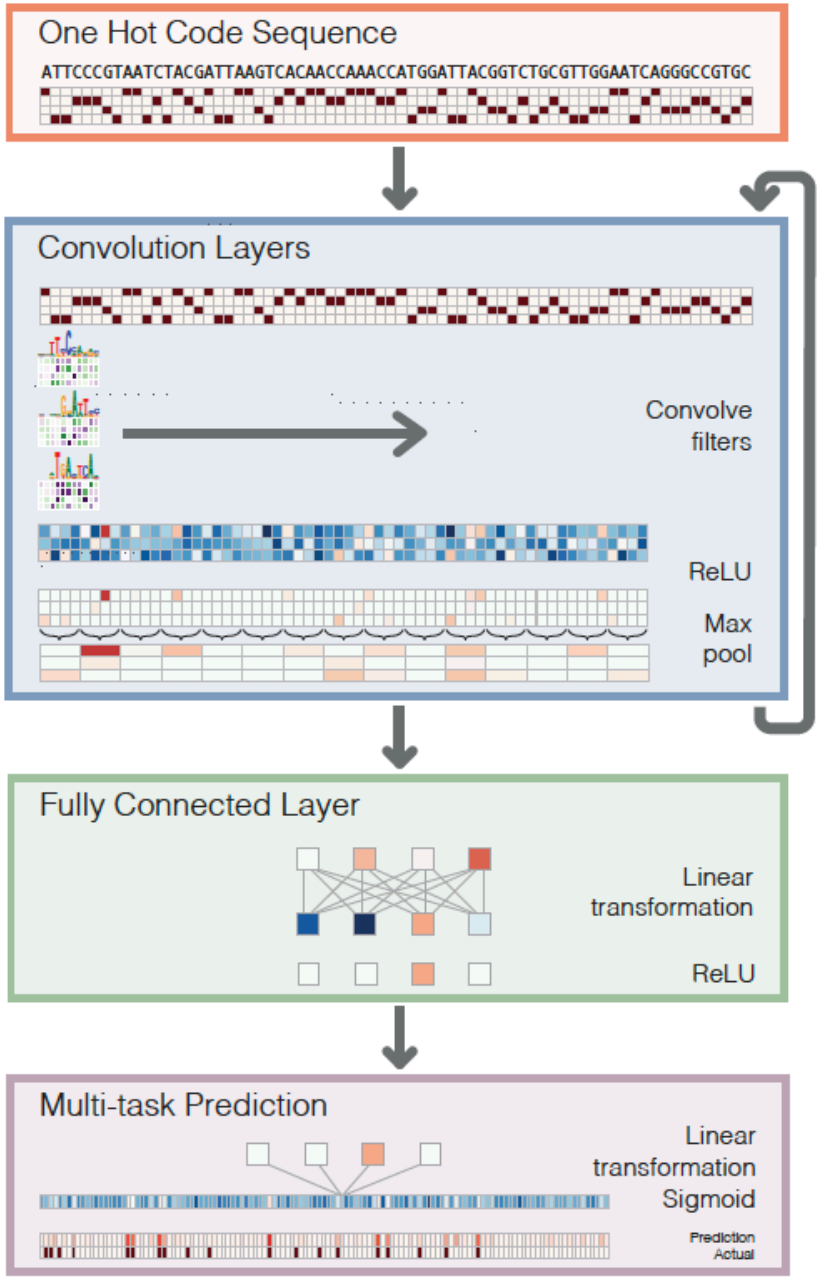
Genome Research, March

# Bassett architecture for accessibility prediction

Input:
600 bp

1.9 million
training
examples

Output:
168 bits

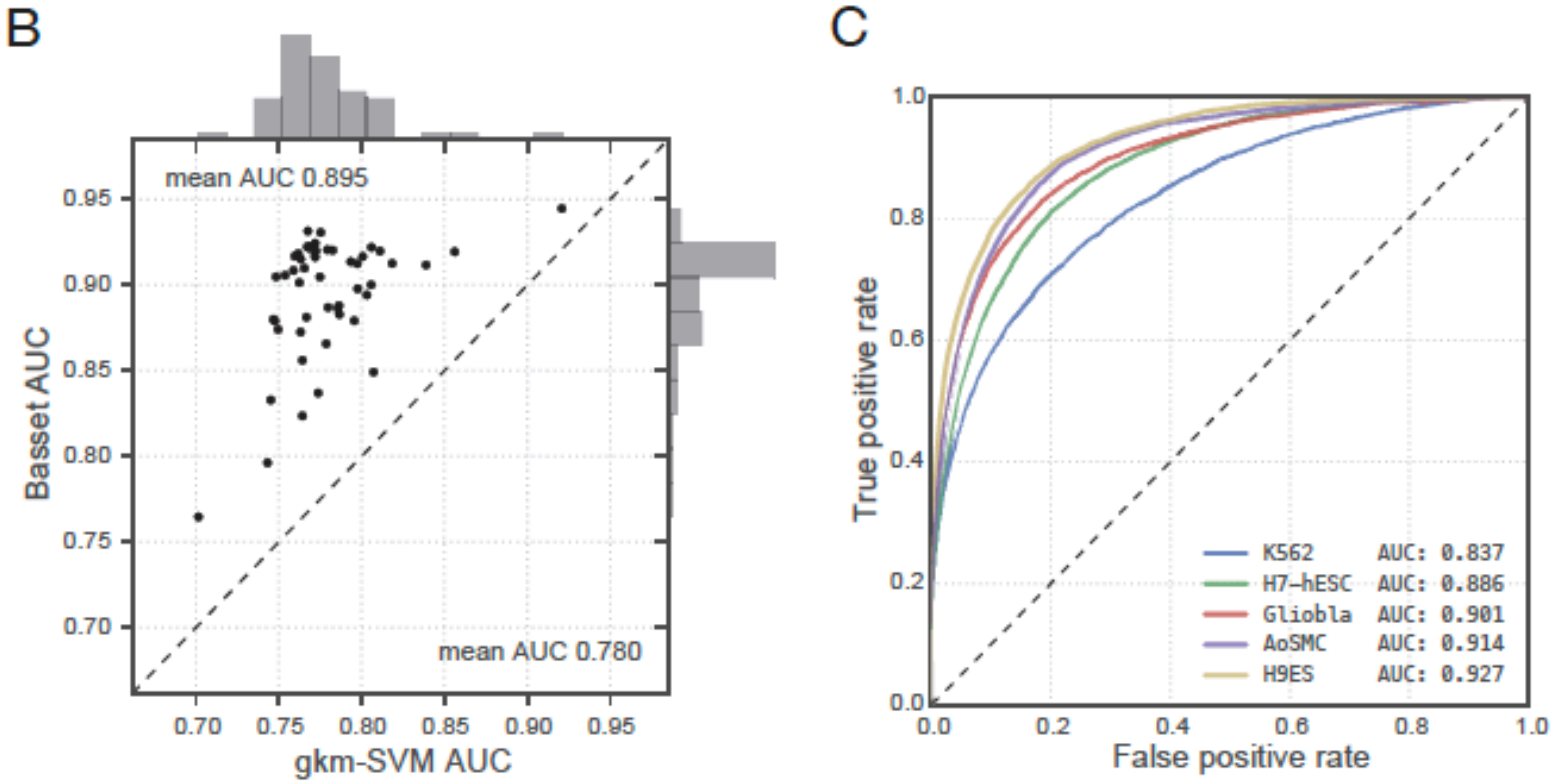



300 filters
3 conv layers
3 FC layers

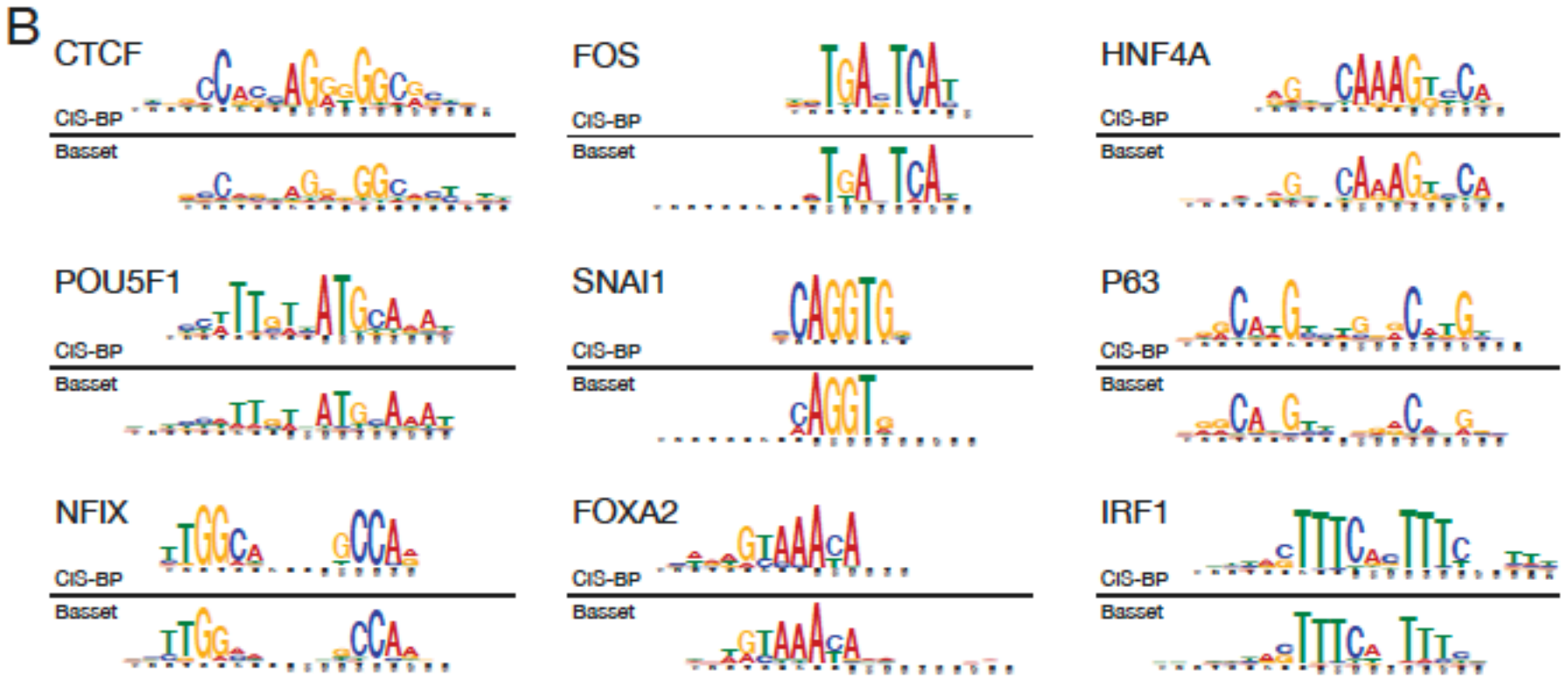
3 fully connected
layers

168 outputs
(1 per cell type)

# Bassett AUC performance vs. gkm-SVM

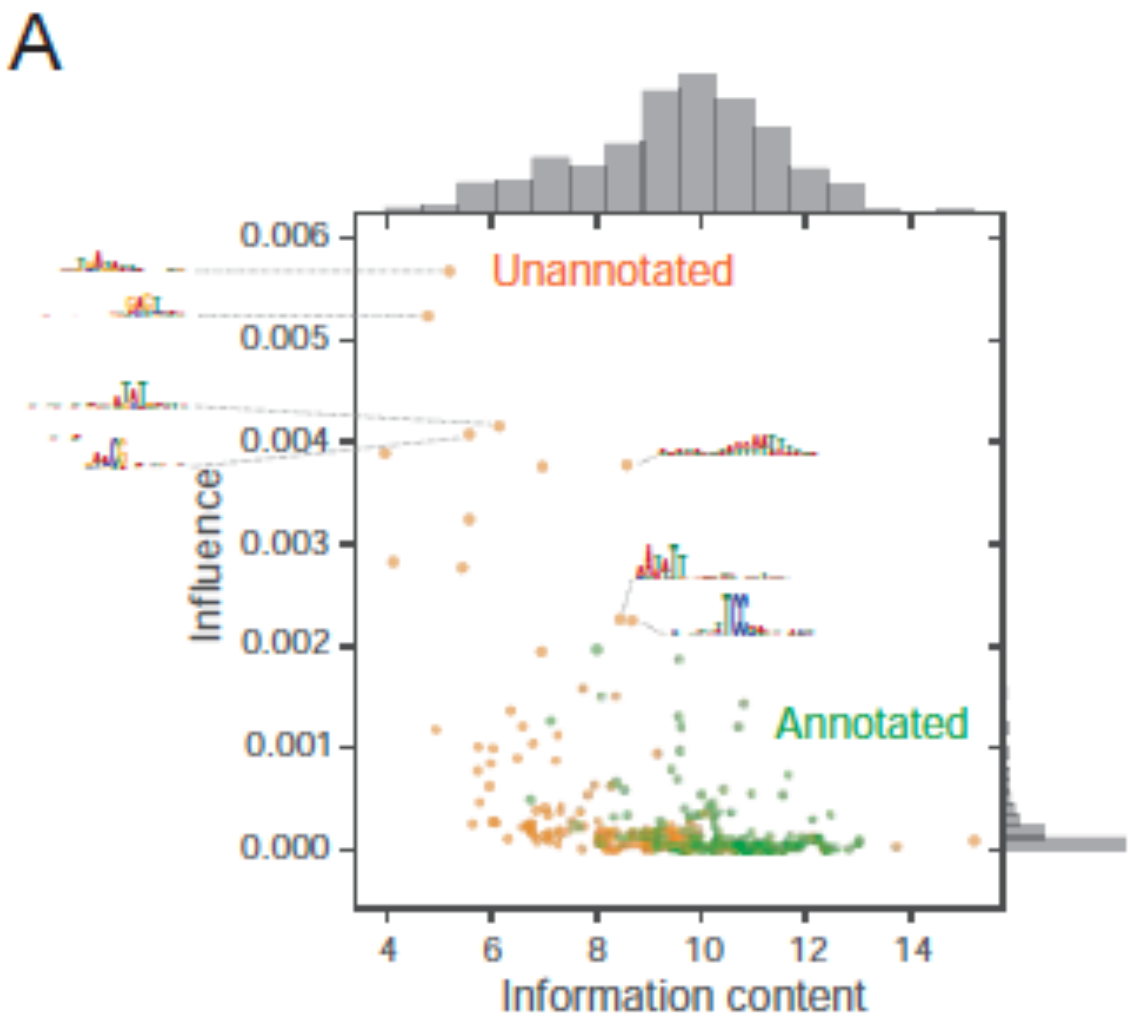


# 45% of filter derived motifs are found in the CIS-BP database

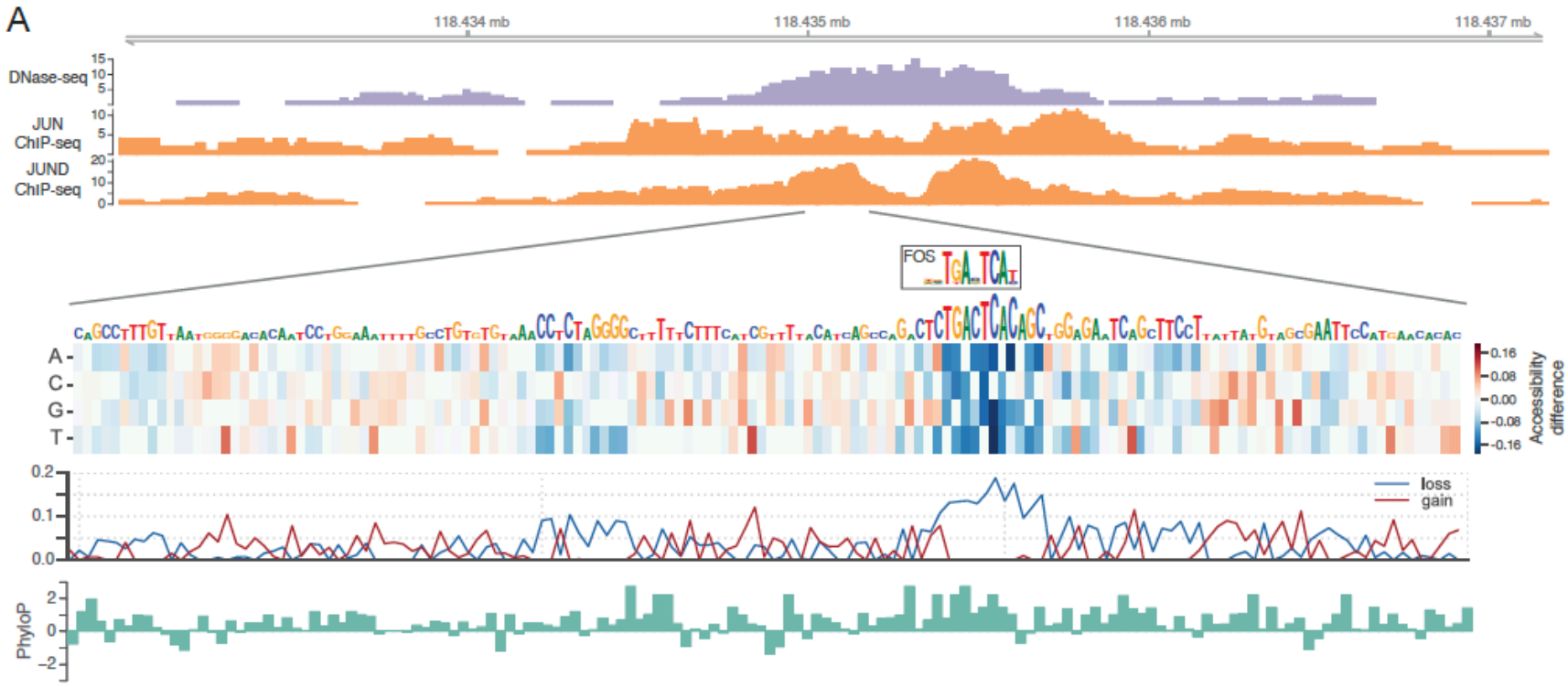



Motifs created by clustering matching input sequences and computing PWM

# Motif derived from filters with more information tend to be annotated



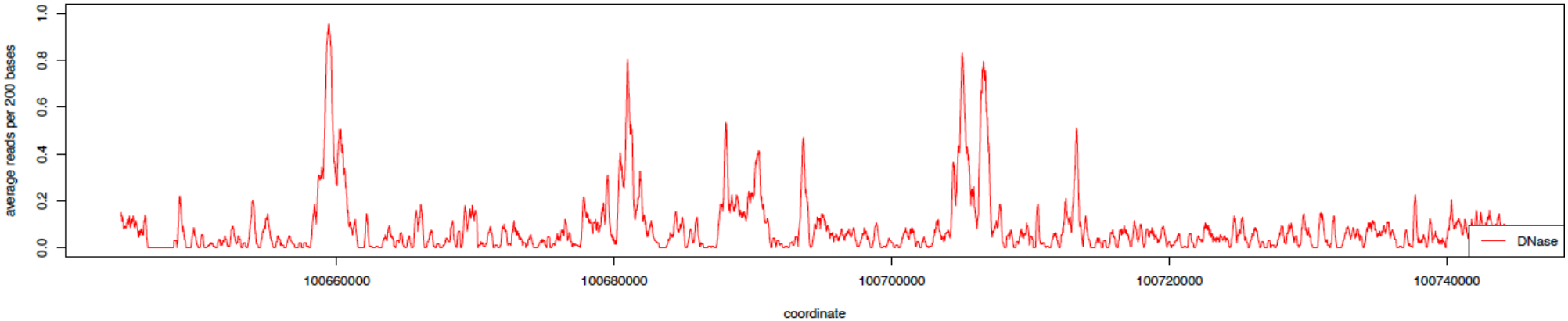
# Computational saturation mutagenesis of an AP-1 site reveals loss of accessibility



# How genome sequence determines cell-type specific chromatin accessibility

Hashimoto TB, et al. "**A Synergistic DNA Logic Predicts Genome-wide Chromatin Accessibility**" *Genome Research* 2016

# Can we predict chromatin accessibility directly from DNA sequence?




DNase-seq data across a 100 kilobase window
(Chromosome 14 K562 cells)

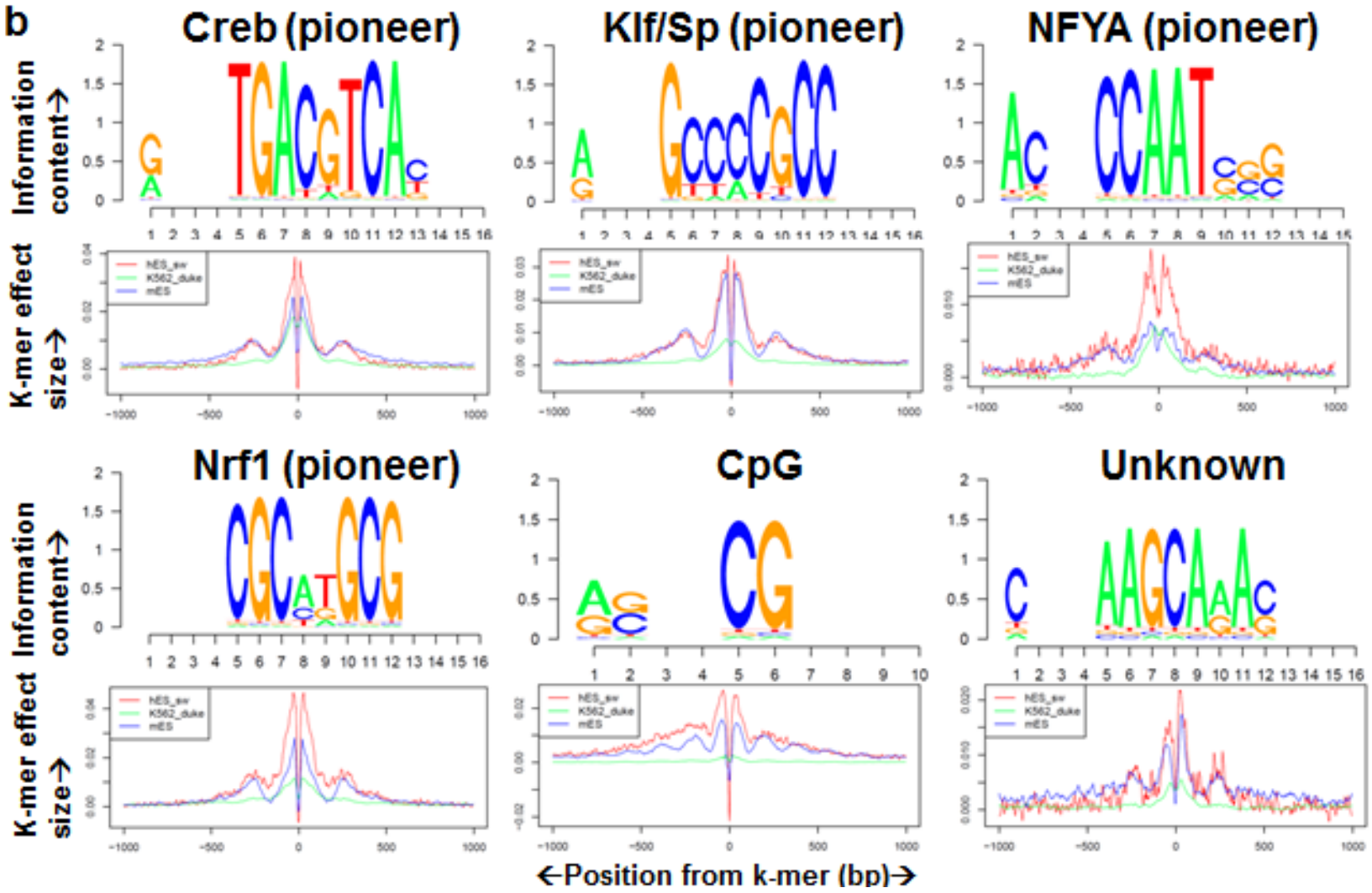
Motivation –

1. Understand the fundamental biology of chromatin accessibility
2. Predict how genomic variants change chromatin accessibility

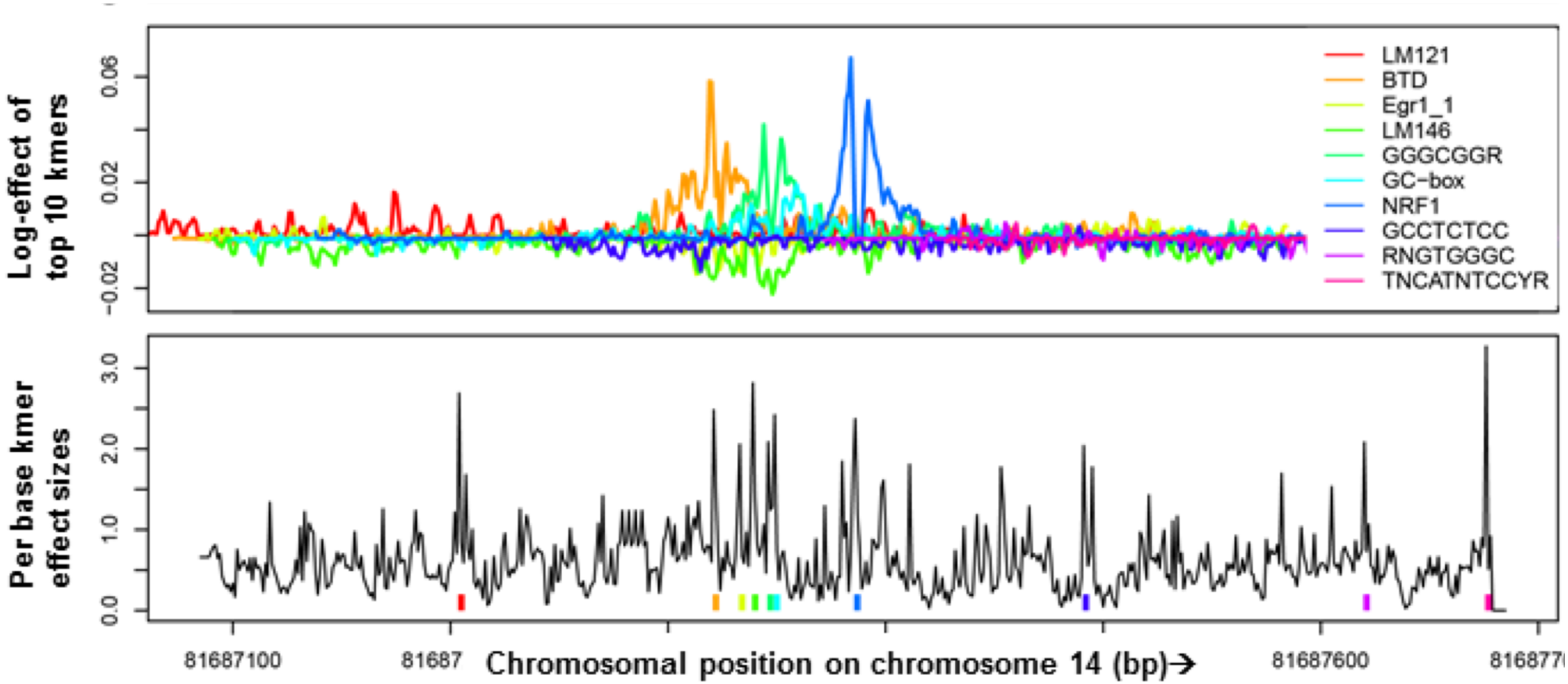
# Can we discover DNA "code words" encoding chromatin accessibility?

- The DNA "code words" encoding chromatin accessibility can be represented by k-mers (k <= 8)
- K-mers affect chromatin accessibility locally within +/- 1 kb with a fixed spatial profile
- A particular k-mer produces the same effect wherever it occurs

# Chromatin accessibility arises from interactions, largely among pioneer TFs



# The Synergistic Chromatin Model (SCM) is a K-mer model



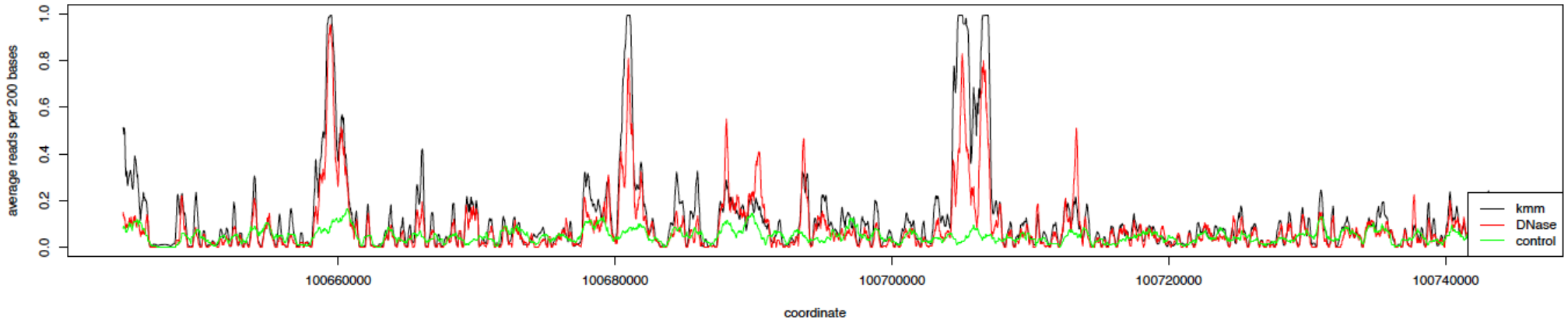

~40,000 K-mers in model

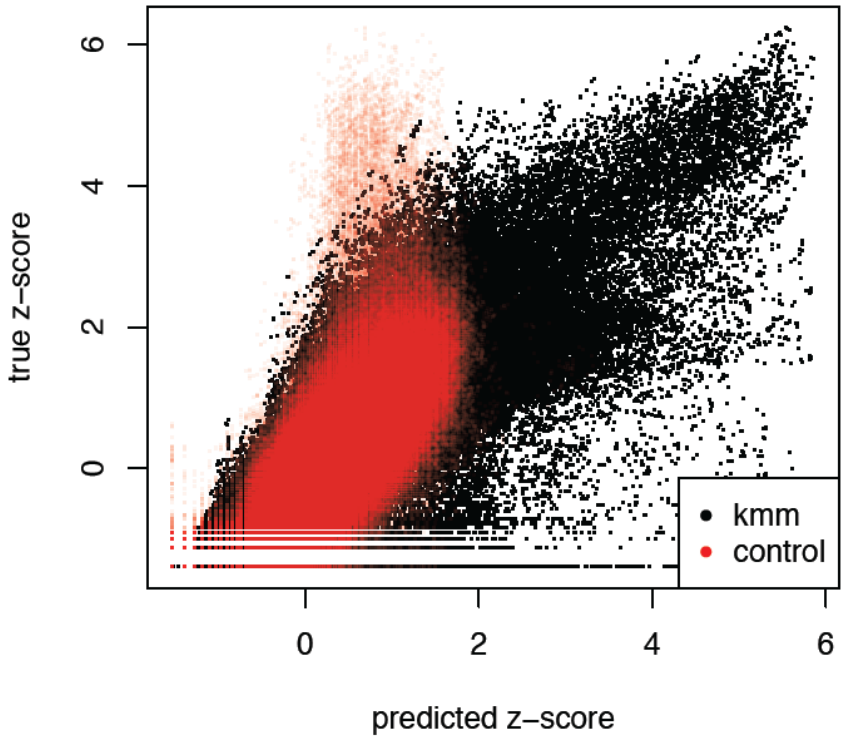
~5,000,000 parameters

543 iterations * 360 seconds / iteration * 40 cores

= ~ 90 days

# Training on K562 DNase-seq data from chromosomes 1 – 13 predicts chromosome 14 (black line)

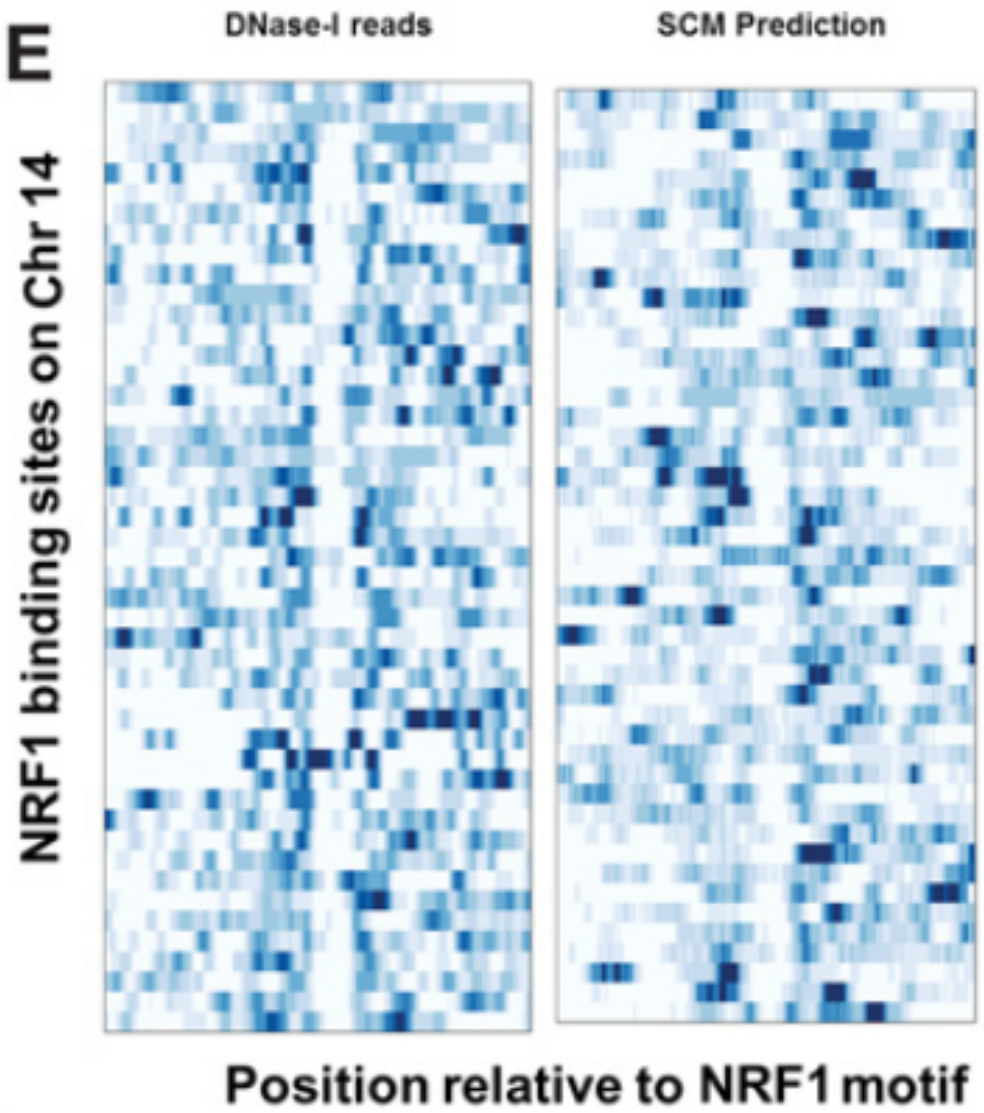



**KMM $R^2$ 0.80**

**Control $R^2$ 0.47**

# SCM predicts accessibility data from a NRF1 binding site



# SCM is solved by regularized Poisson regression

$\theta^k$ describes the accessibility profiles for k-mers
$c_i$ is the observed read count at base i
$\lambda_i$ is the predicted count at base i

$$\max_{\theta}\left(\sum_i c_i \log(\lambda_i) - \lambda_i\right) - \eta \sum \left|\theta^k\right|_1.$$

The intermediate variables $\lambda$ are defined by:

$$\lambda_i = \exp\left[\left(\sum_{k\in[1..K]} \sum_{j\in[-M,M-1]} \theta^k_{(g^k_{i+j},-j)}\right) - \theta_0\right].$$

How can we learn $\theta$?

Recall convex functions can be solved by gradient methods

$c(x) = Mx + b$

$c(x) = e^{c1(x)}$

$c(x) = c1(x) + c2(x)$

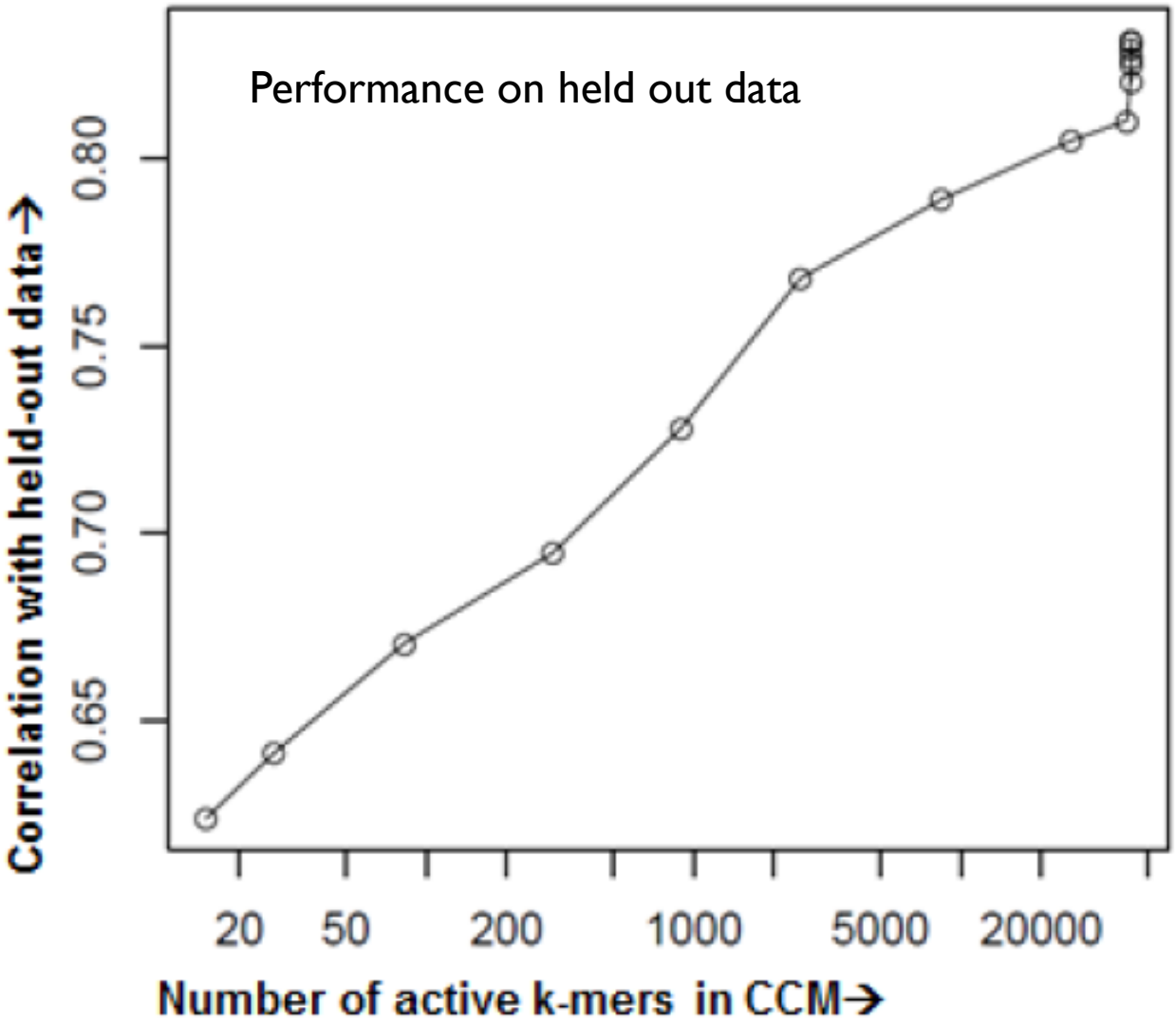
$c(x) = xp$

$c(x) = |x|$

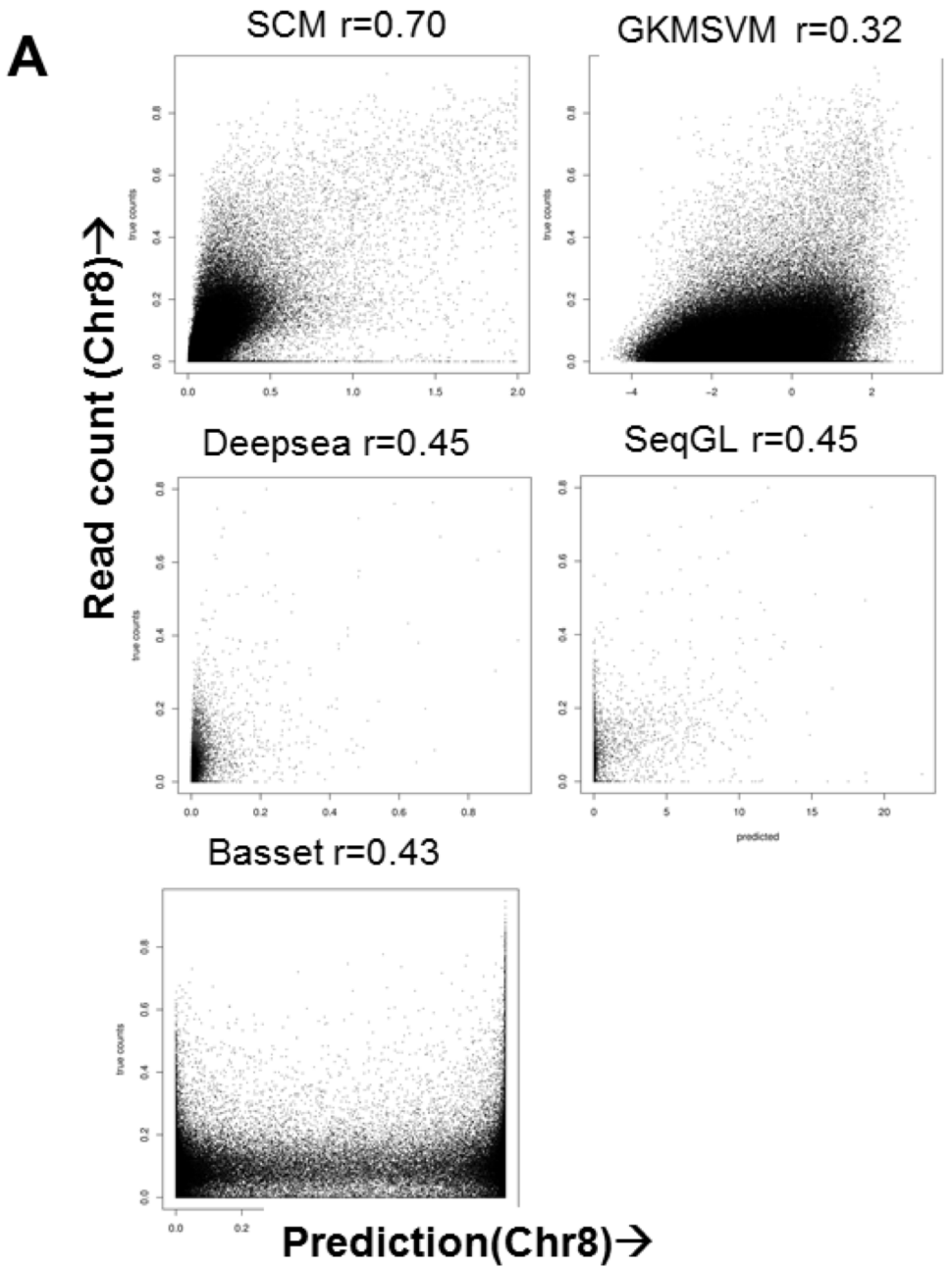
$c(x) = x \log x$

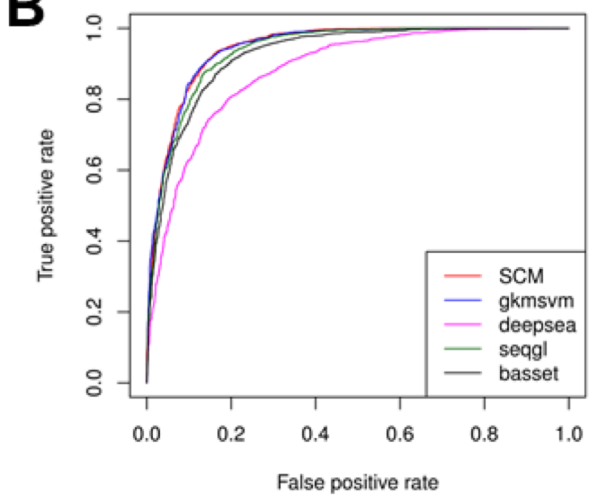
$c(x) = \max( c1(x), c2(x) )$

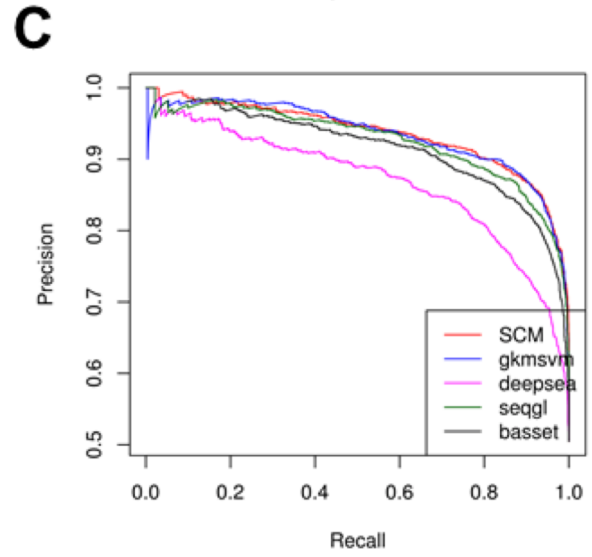
# SCM model performance improves with more k-mer features



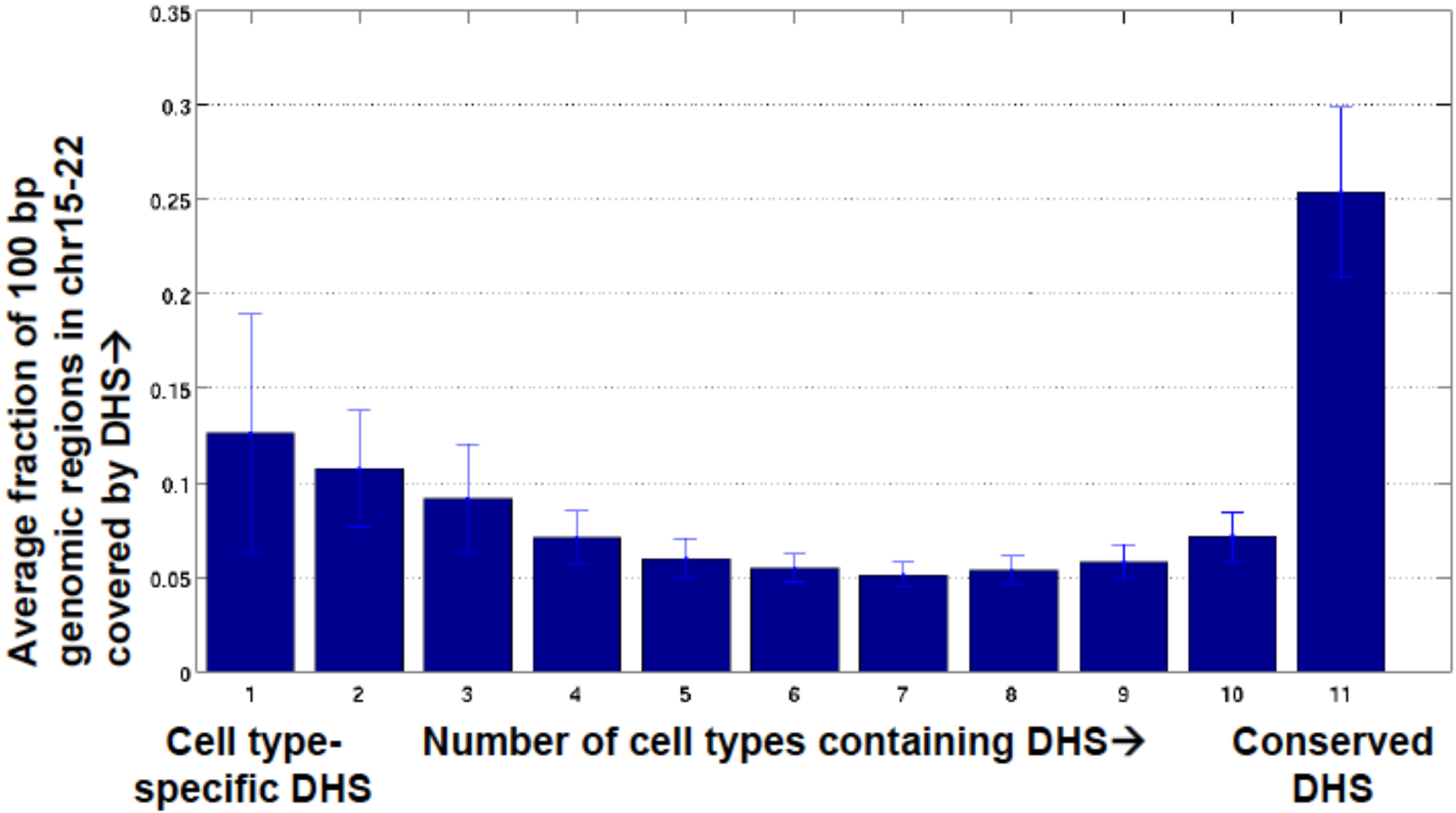
# SCM outperforms contemporary models at predicting chromatin accessibility from sequence (K562)



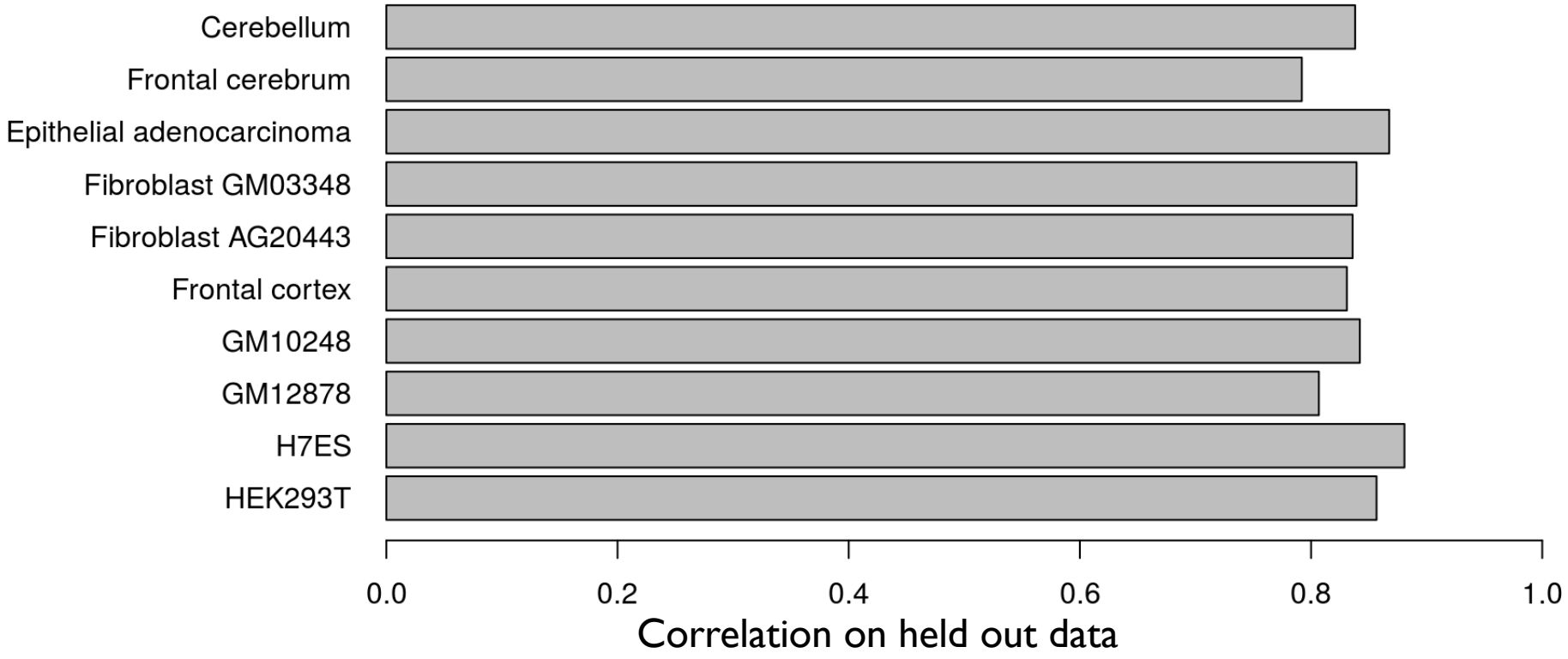



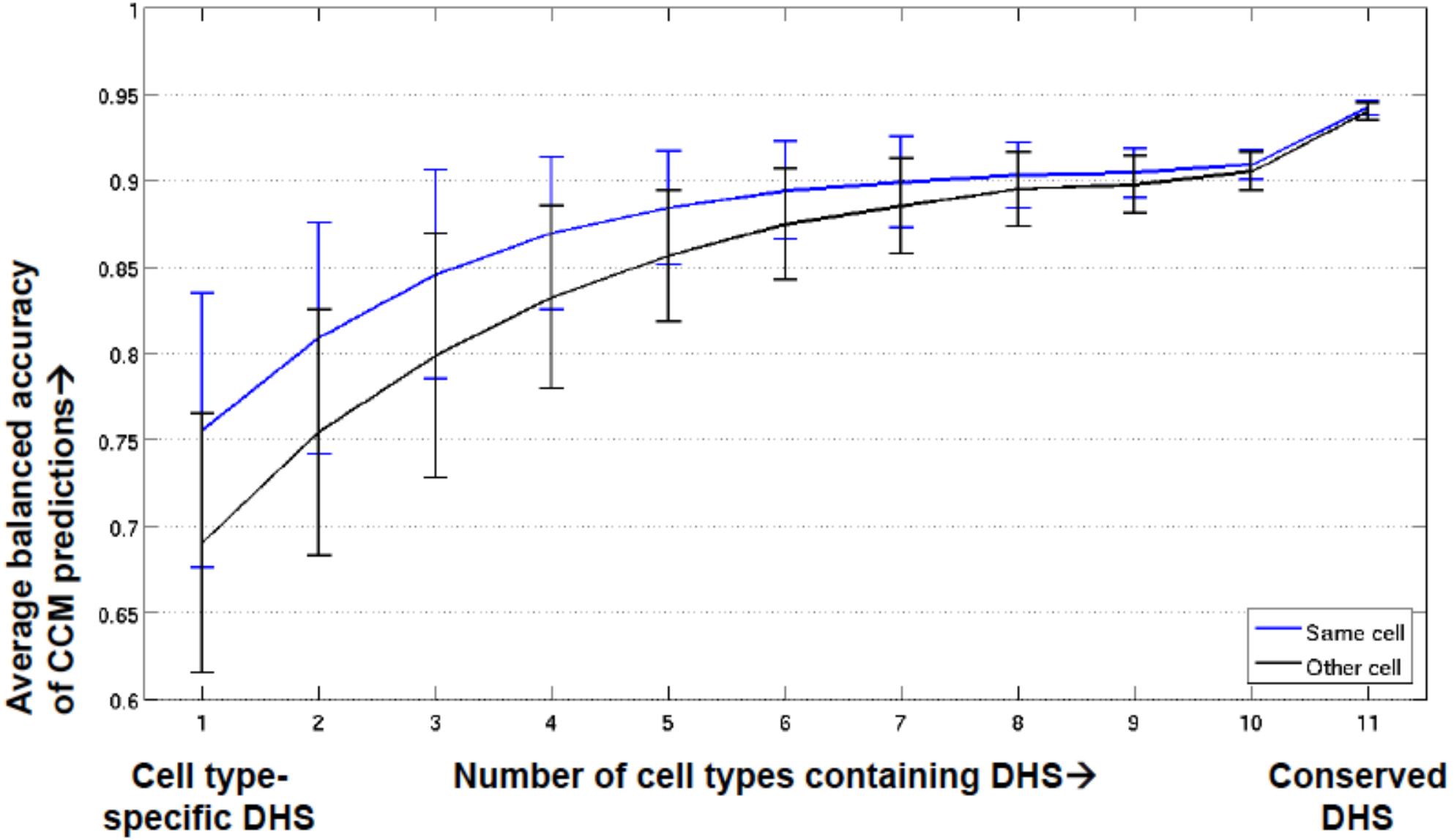
# Accessibility contains cell type specific and cell type independent components (11 cell types, Chr 15-22)



# SCM models have similar predictive power for other cell types

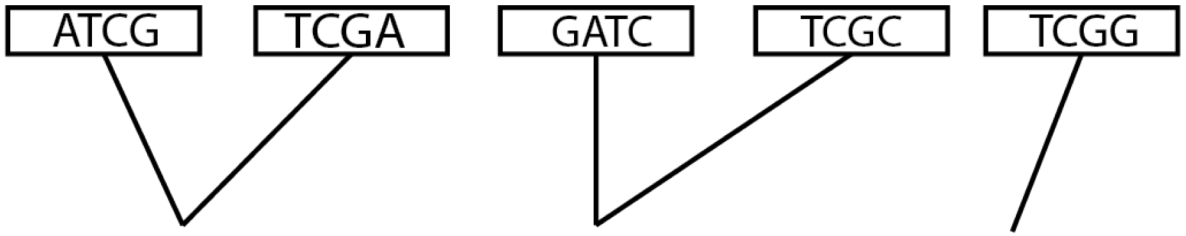


# SCM model trained on ES data performs better on shared DNase hot spots (Chr 15 – 22)



# We created synthetic "phrases" each of which contains k-mers that are similar in chromatin opening score

All K-mers



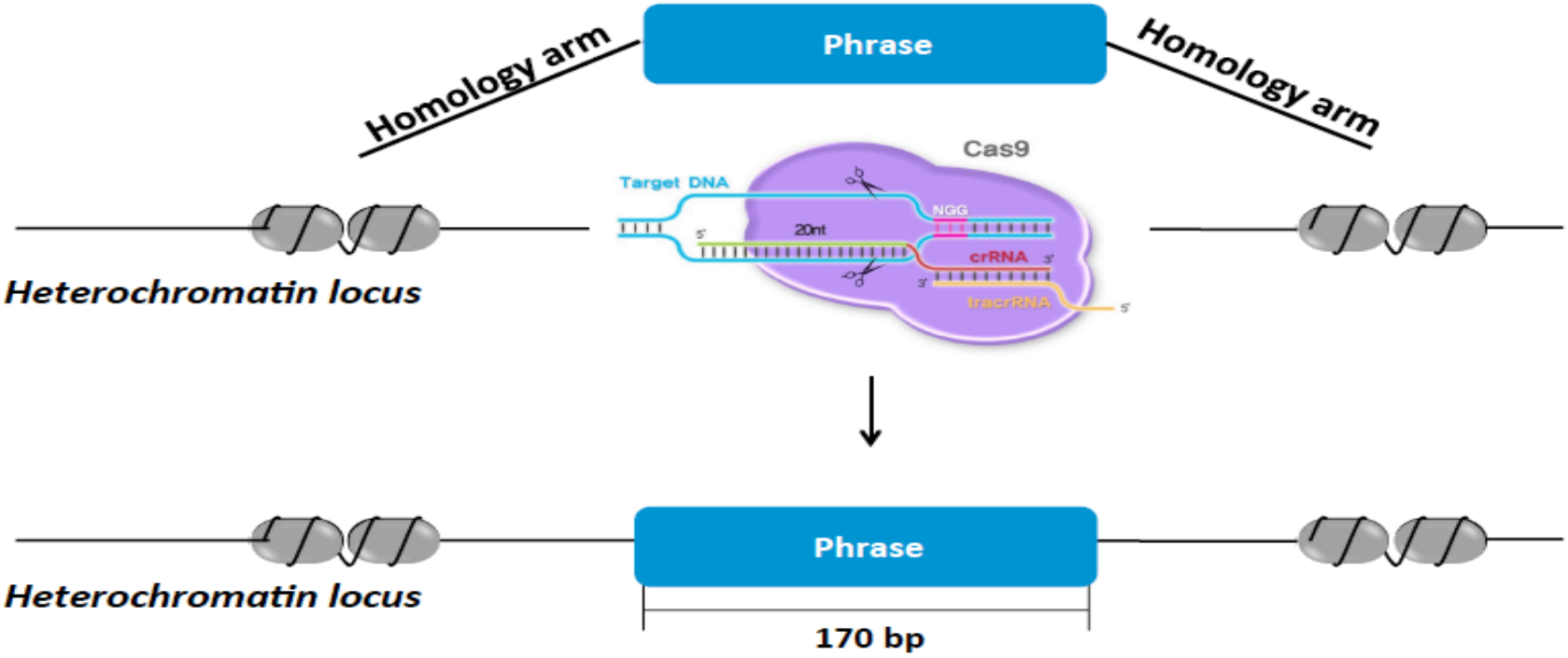

Sort into classes


Construct Debrujin Graph


Biased random walk

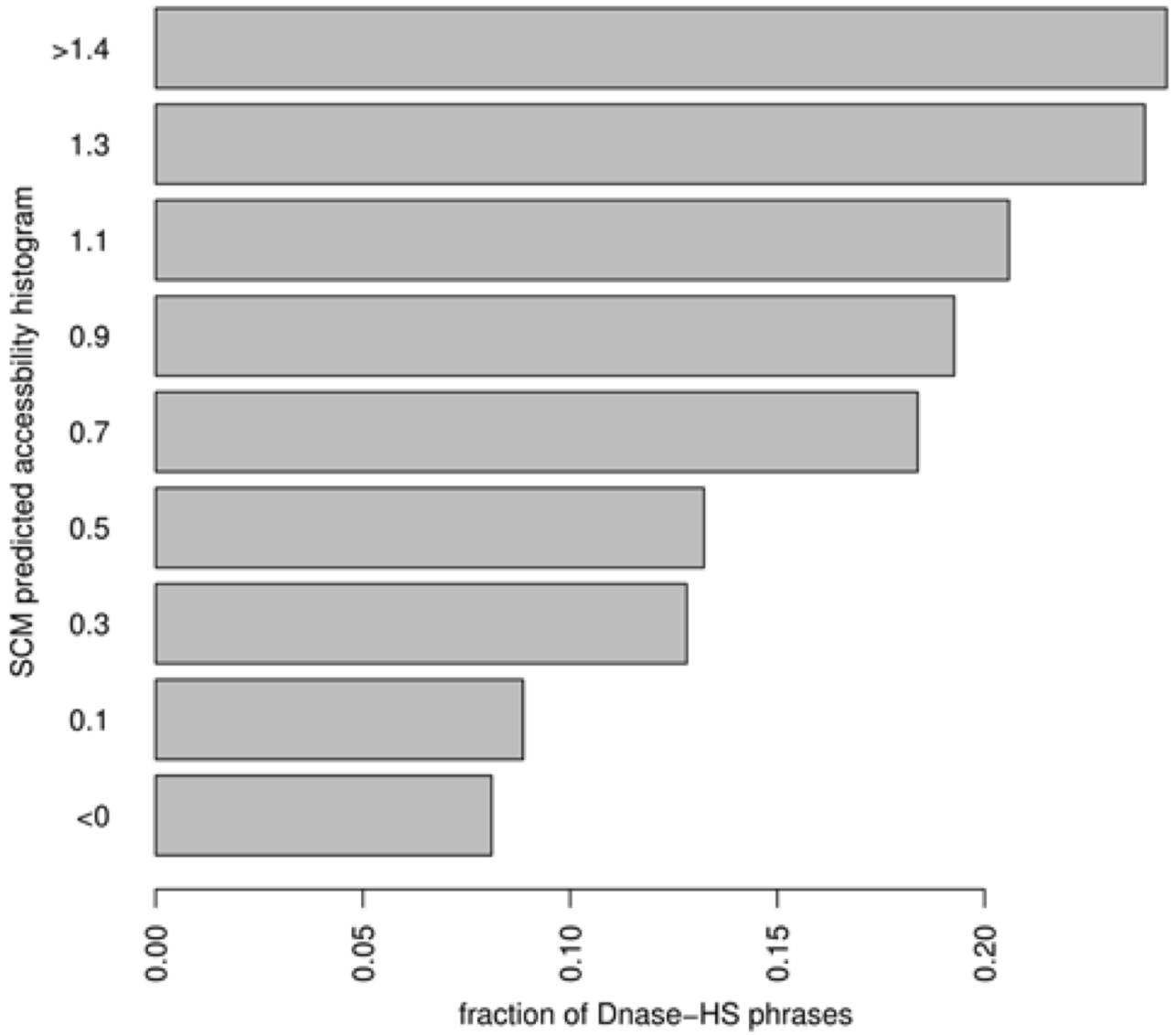
1. GATCGC
2. GATCGA

# Single Locus Oligonucleotide Transfer
# >6,000 designed phrases into a chromosomal locus




| Heterochromatin locus | A | B | C |
|---|---|---|---|
| % alleles with phrase integration | 35 | 15 | 15 |
| # unique integrations | 350,000 | 150,000 | 150,000 |

# Predicted accessibility matches measured accessibility



# FIN - Thank You