Analysis of Chromatin Structure Lecture 18

David K. Gifford

Massachusetts Institute of Technology

Today's Narrative Arc

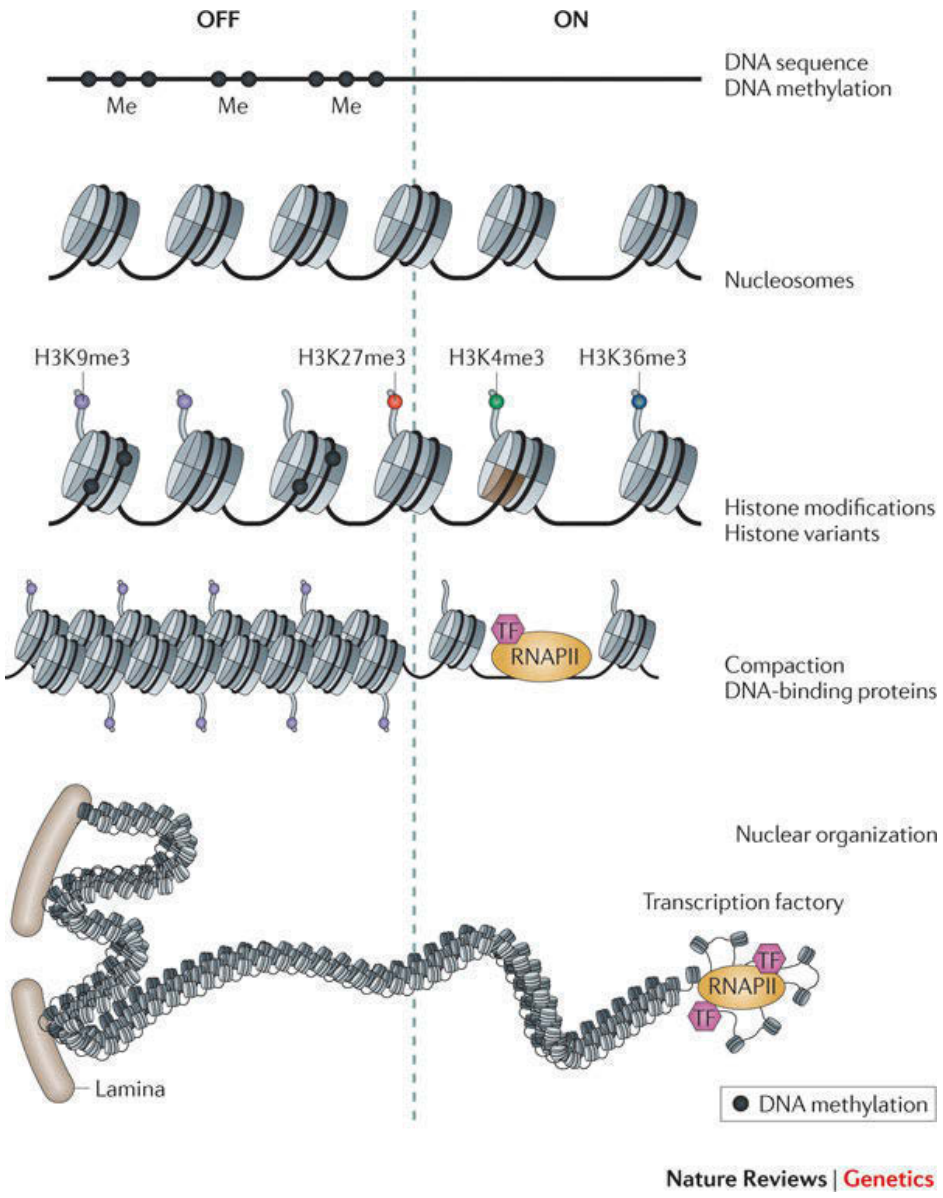
1. Using computational methods we can break the epigenetic "code" that describes the function and state of genome elements. Epigenetic state regulates gene function without changing primary DNA sequence. Epigenetic state includes histone marks, DNA methylation, and chromatin openness.
2. We can estimate the protein occupancy of the genome and discover pioneer factors with DNase-seq via computational methods.
3. We can map enhancers to their regulatory targets with the computational analysis of ChIA-PET data (and similar technologies)

Today's Computational Methods

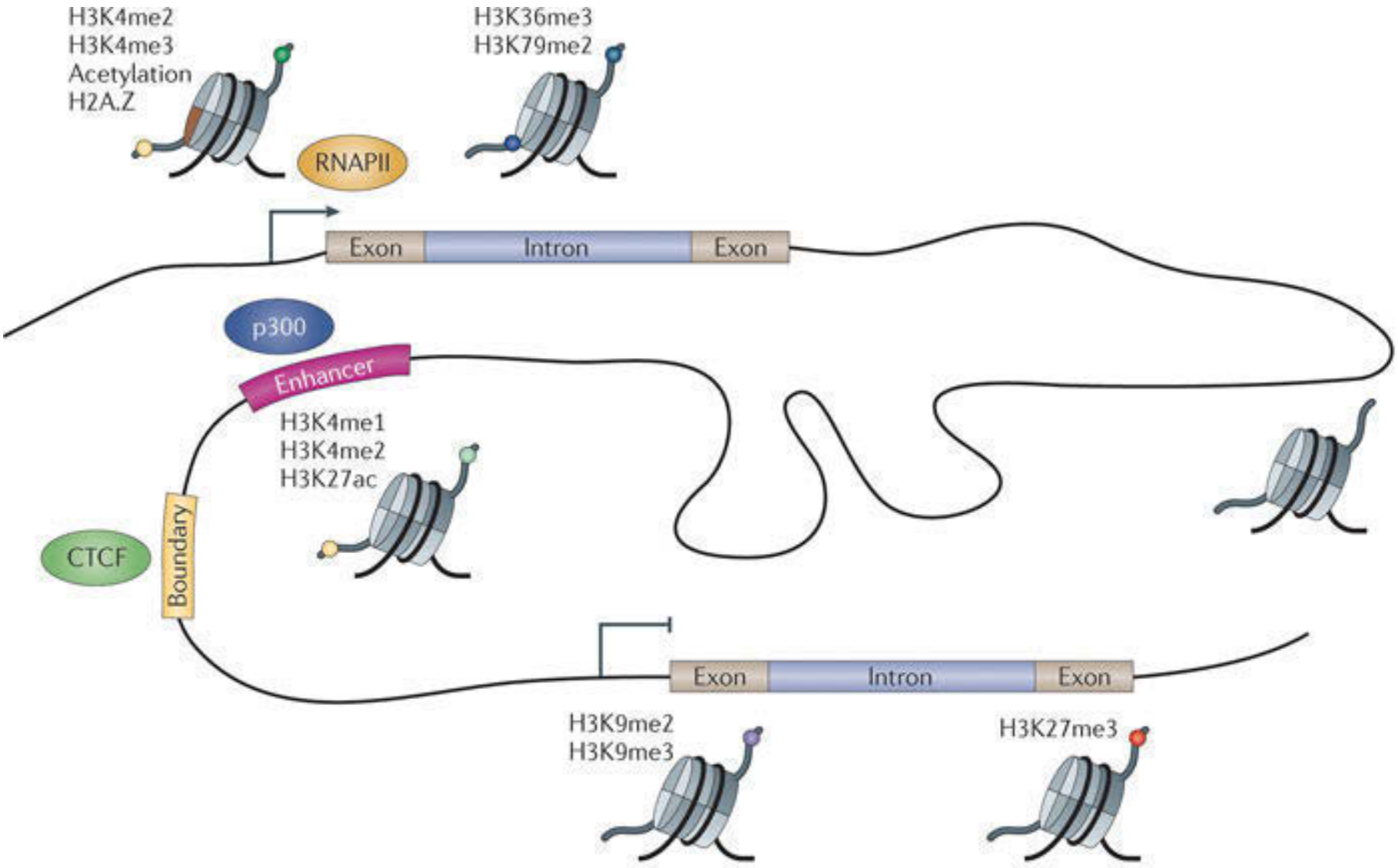
1. Dynamic Bayesian Networks
2. Factor binding classification using a log likelihood ratio
3. Hypergeometric distribution

Chromatin organization has multiple structural layers and organizes chromatin into “domains”

Both DNA methylation and chromatin marks contain important functional information

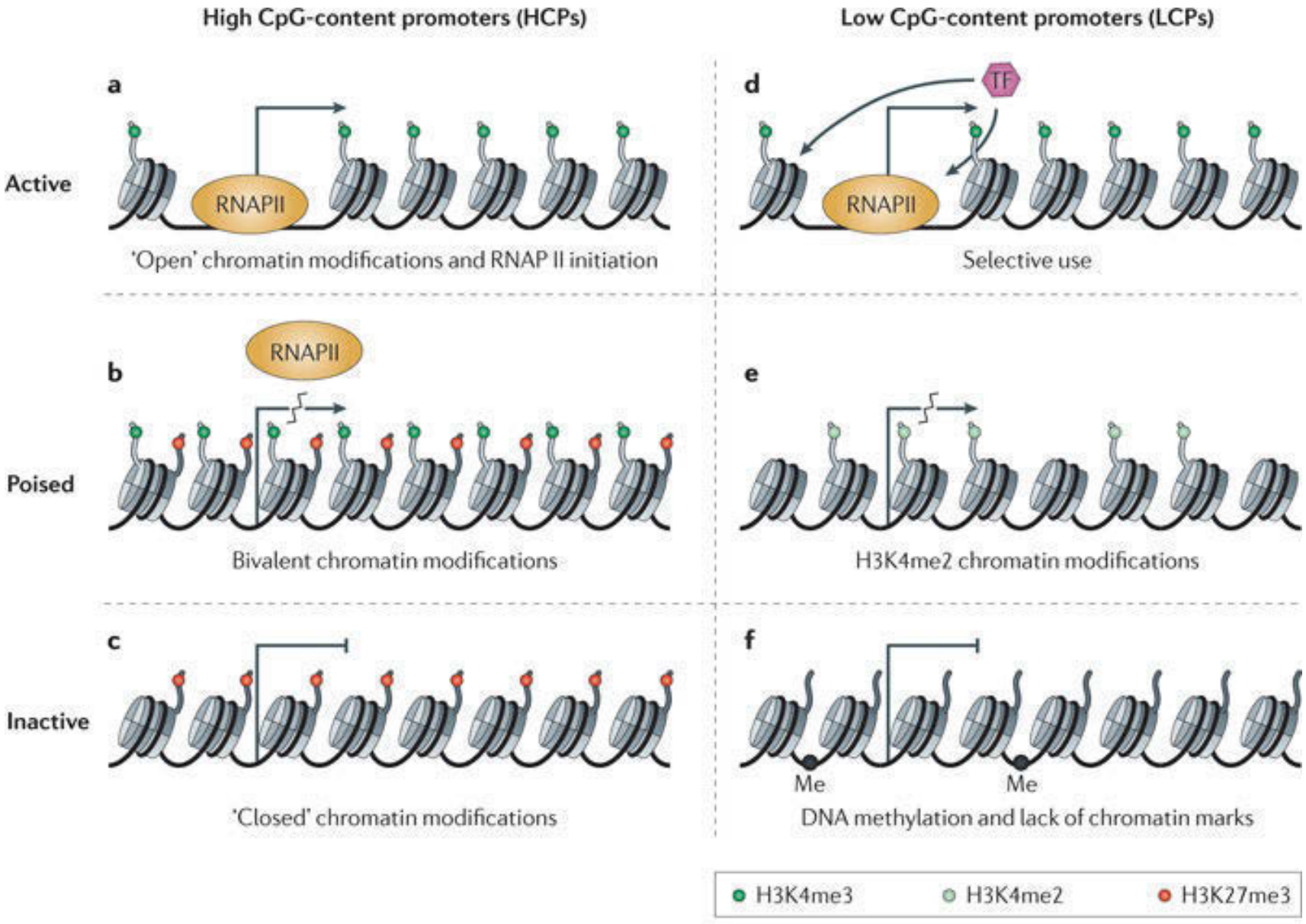


Source: Zhou, Vicky W., Alon Goren, et al. "Charting Histone Modifications and the Functional Organization of Mammalian Genomes." *Nature Reviews Genetics* 12, no. 1 (2010): 7-18.

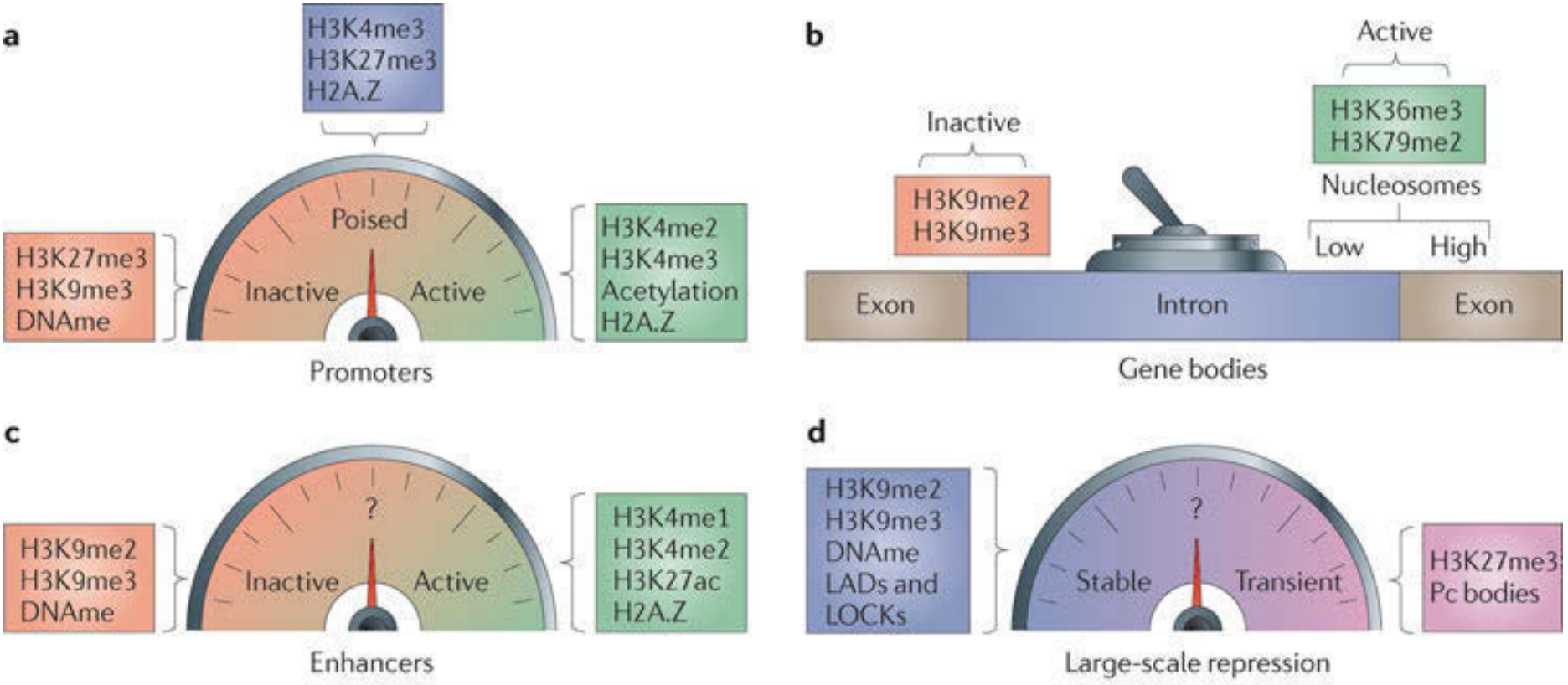


Source: Zhou, Vicky W., Alon Goren, et al. "Charting Histone Modifications and the Functional Organization of Mammalian Genomes." *Nature Reviews Genetics* 12, no. 1 (2010): 7-18.

Nature Reviews | Genetics



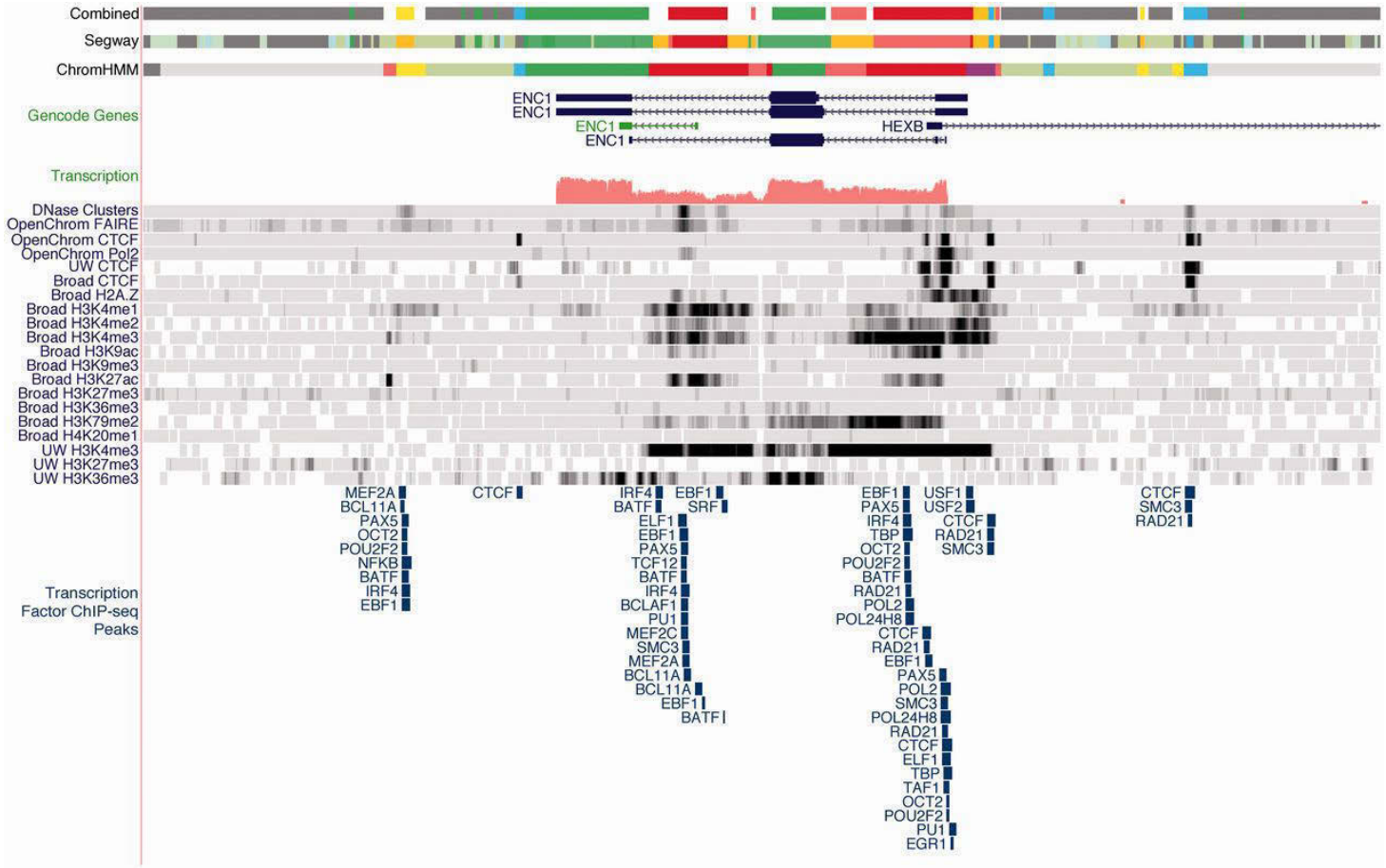
Nature Reviews | Genetics
Source: Zhou, Vicky W., Alon Goren, et al. "Charting Histone Modifications and the Functional Organization of Mammalian Genomes." *Nature Reviews Genetics* 12, no. 1 (2010): 7-18.



Source: Zhou, Vicky W., Alon Goren, et al. "Charting Histone Modifications and the Functional Organization of Mammalian Genomes." *Nature Reviews Genetics* 12, no. 1 (2010): 7-18.

What is the Histone Code?

View of the ENC1 locus on the minus strand using the ENCODE GM12878 segmentations.



Source: Hoffman, Michael M., Jason Ernst, et al. "Integrative Annotation of Chromatin Elements from ENCODE Data." *Nucleic Acids Research* (2012): gks1284.

Hoffman M M et al. Nucl. Acids Res. 2013;41:827-841

Nucleic Acids Research

Ideas for chromatin track analysis

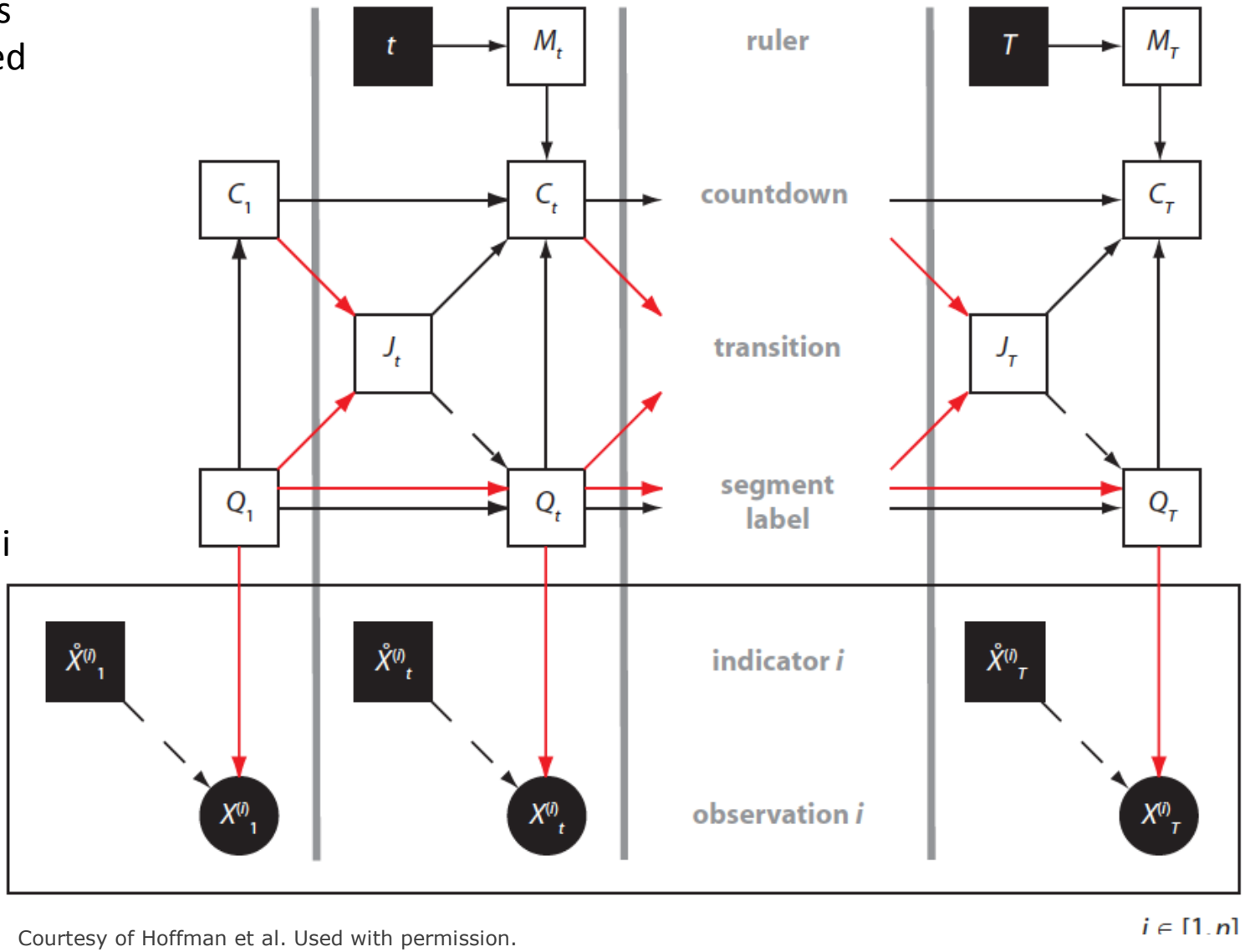
- Hidden Markov Model (ChromHMM)
- Dynamic Bayesian Network (Segway)
  - Bayesian Network that models data sampled at intervals. Still a directed acyclic graph (DAG).
  - Can learn model with Graphical Model Toolkit (GMTK)
  - Can incorporate relationships between variables and handle missing data
  - 1bp analysis resolution

Segway Dynamic Bayesian Network

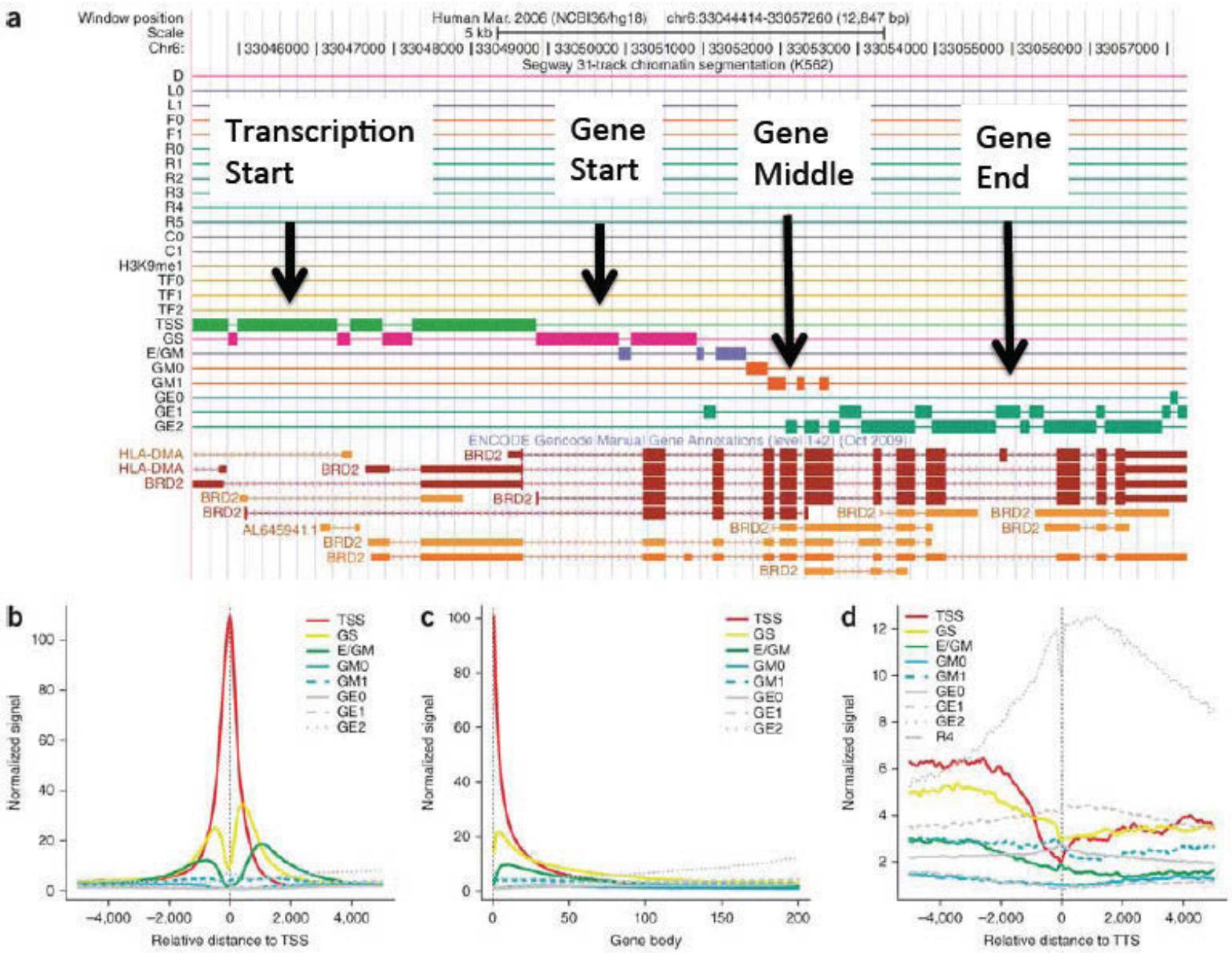
Black Nodes are observed

Training on 1% of genome with GMTK

Analysis on entire genome with Viterbi



Courtesy of Hoffman et al. Used with permission.
Source: Hoffman, Michael M., Orion J. Buske, et al. "Segway: Simultaneous Segmentation of Multiple Functional Genomics Data Sets with Heterogeneous Patterns of Missing Data."



Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Hoffman, Michael M., Orion J. Buske, et al. "Unsupervised Pattern Discovery in Human Chromatin Structure through Genomic Segmentation." *Nature Methods* 9, no. 5 (2012): 473-6.

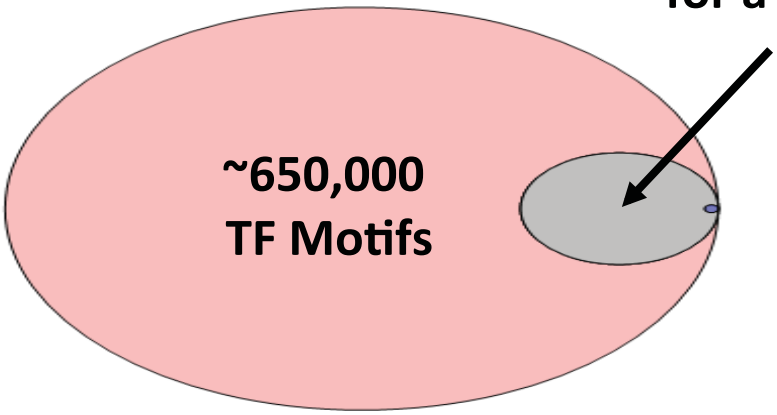
Today's Narrative Arc

1. We can break the epigenetic "code" that describes the function and state of genome elements using computational methods. Epigenetic state regulates gene function without changing primary DNA sequence. Epigenetic state includes histone marks, DNA methylation, and chromatin openness.
2. **We can estimate the protein occupancy of the genome and discover pioneer factors with DNase-seq via computational methods.**
3. We can map enhancers to their regulatory targets with the computational analysis of ChIA-PET data (and similar technologies)

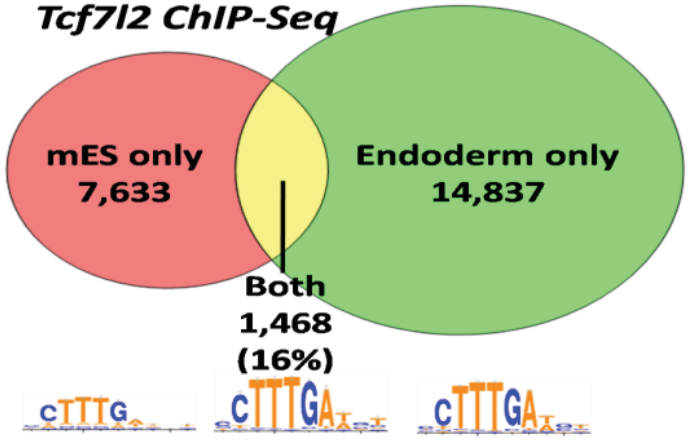
What are the rules governing how a TF chooses its genomic binding sites?

Motifs are insufficient to predict binding

~50,000 binding sites for a typical TF

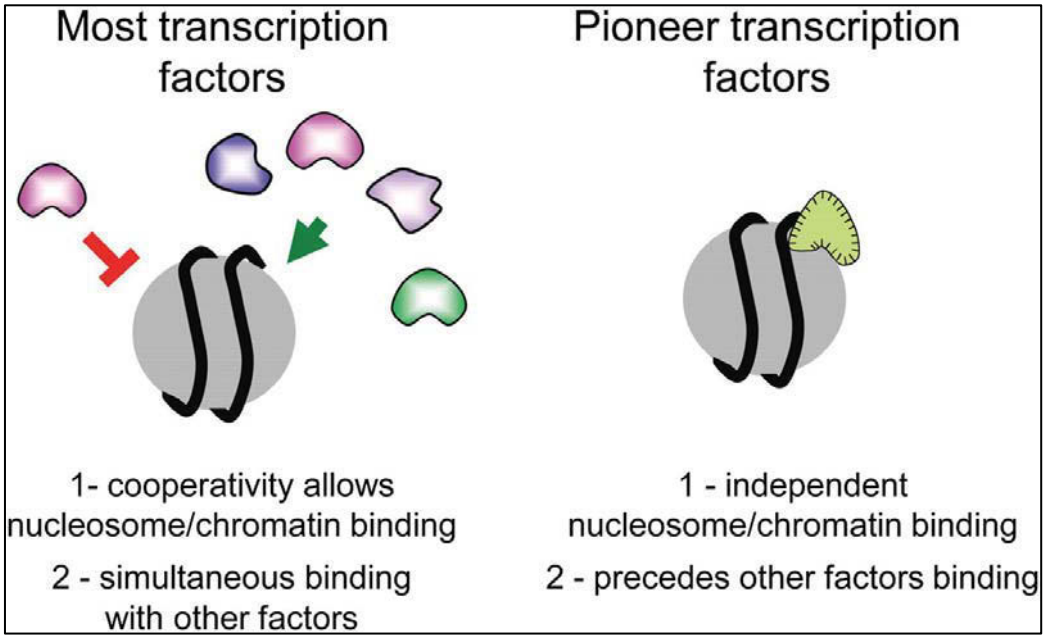


Binding sites change across time



Pioneer Transcription Factors (TFs) are special

- Pioneer TFs bind target sites regardless of chromatin state
  - FoxA, histone mimic [Gualdi 1996]
  - iPS reprogramming factors pioneer (Klf, Oct, Sox) [Soufi 2011]
  - Determined via in-depth molecular biological study



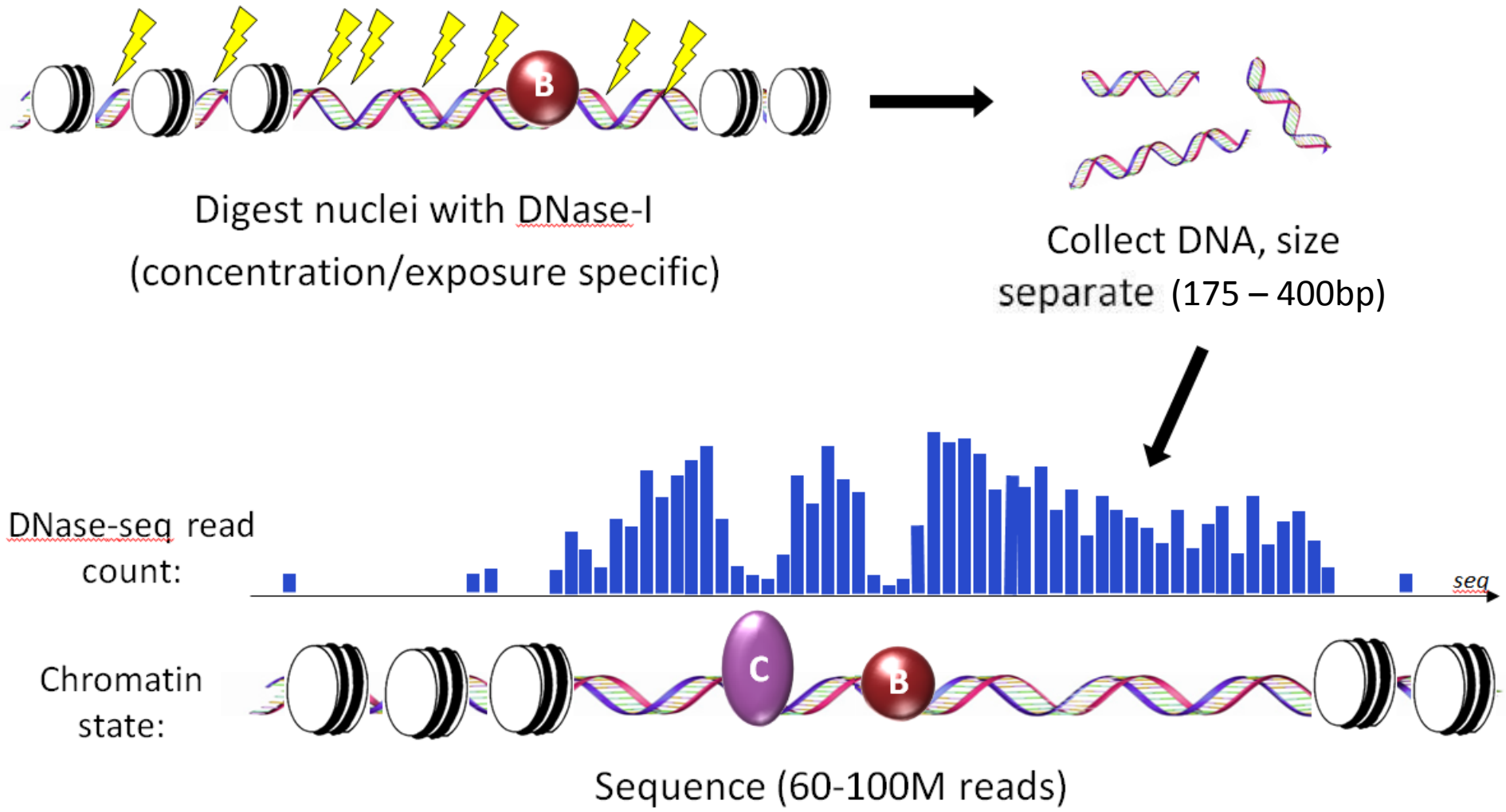
Source: Zaret, Kenneth S., and Jason S. Carroll. "Pioneer Transcription Factors: Establishing Competence for Gene Expression." *Genes & Development* 25, no. 21 (2011): 2227-41.

Zaret 2011

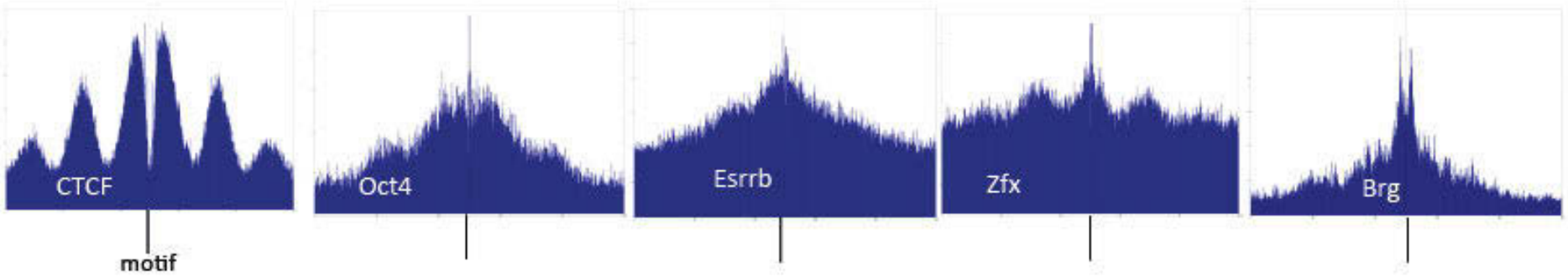
Overview of Results

- Claim 1: Protein Interaction Quantitation (PIQ) accurately predicts transcription factor (TF) binding from DNase-seq data
- Claim 2: PIQ can identify pioneer factors that regulate proximal chromatin opening and TF binding
- Claim 3: Certain pioneer TFs are directional
- Claim 4: Settler factors follow pioneer factor binding and loss of pioneer binding causes chromatin to return to a closed state

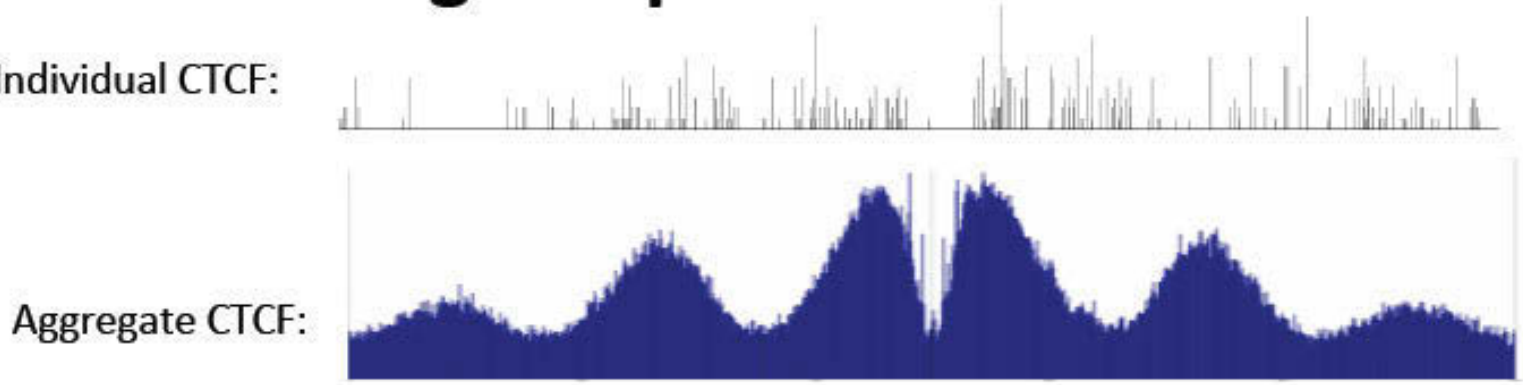
DNase-seq reveals genome protection profiles



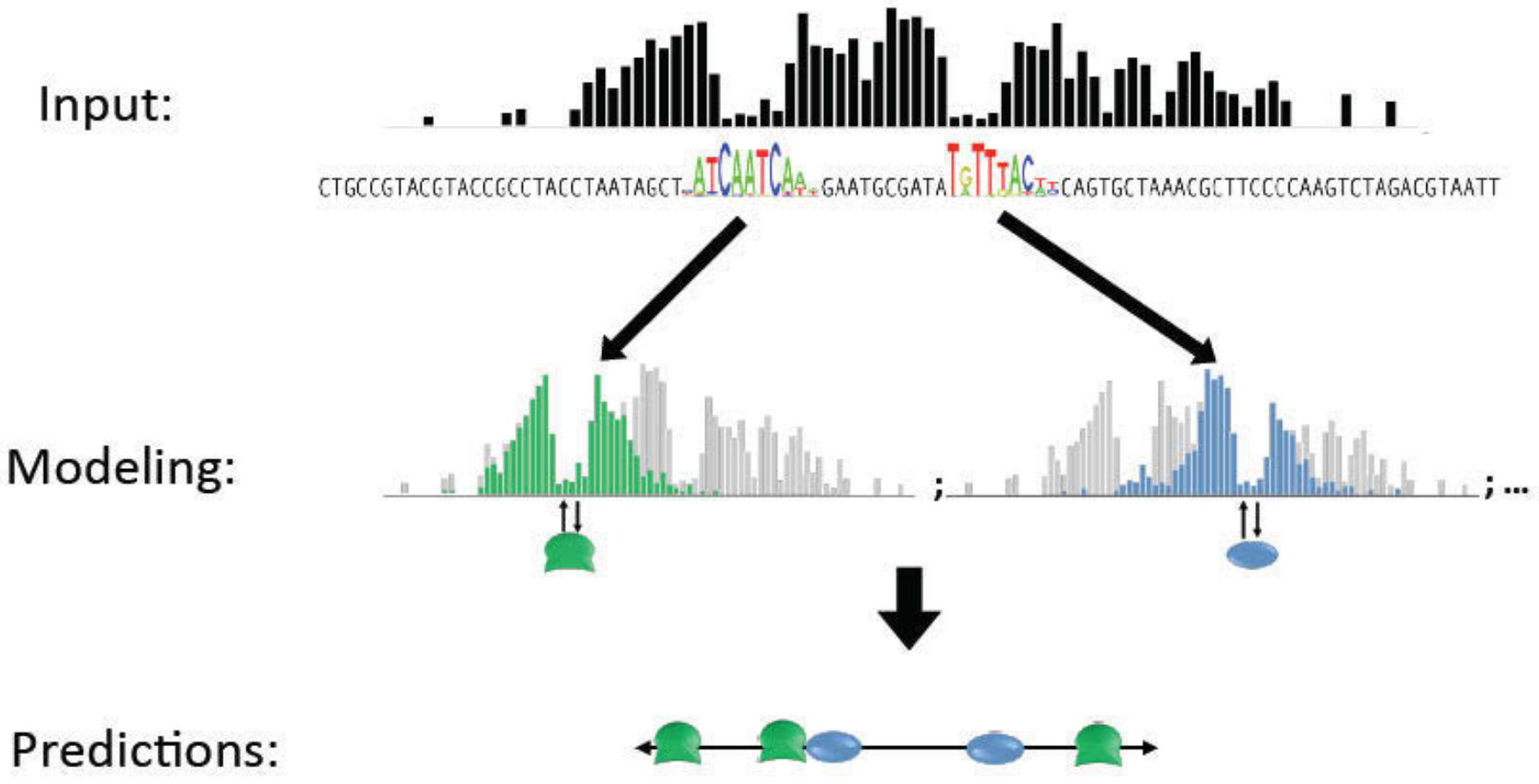
Bound factors leave distinct DNase-seq profiles



Individual binding site prediction is difficult



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



Motivation and Design goals for PIQ

1. Resistance to low coverage and noisy data
2. Integrate multiple experiments
3. Scalability to whole genome with thousands of motifs.

Use a Poisson-Gaussian Process
Near linear time approximate inference

4. High spatial accuracy through motifs
5. Robust worst case behavior

Use a monotone prior to incorporate side information

Poisson-GP model estimates the unoccupied genome

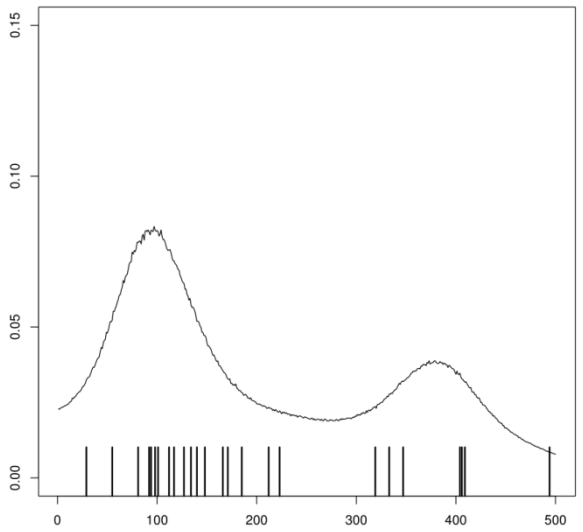
- Our model is a compound distribution (MVN is multivariate normal)

$$c_i \sim Poisson(\exp(\lambda_i))$$

$$\lambda \sim MVN(\mu_0, \Sigma)$$

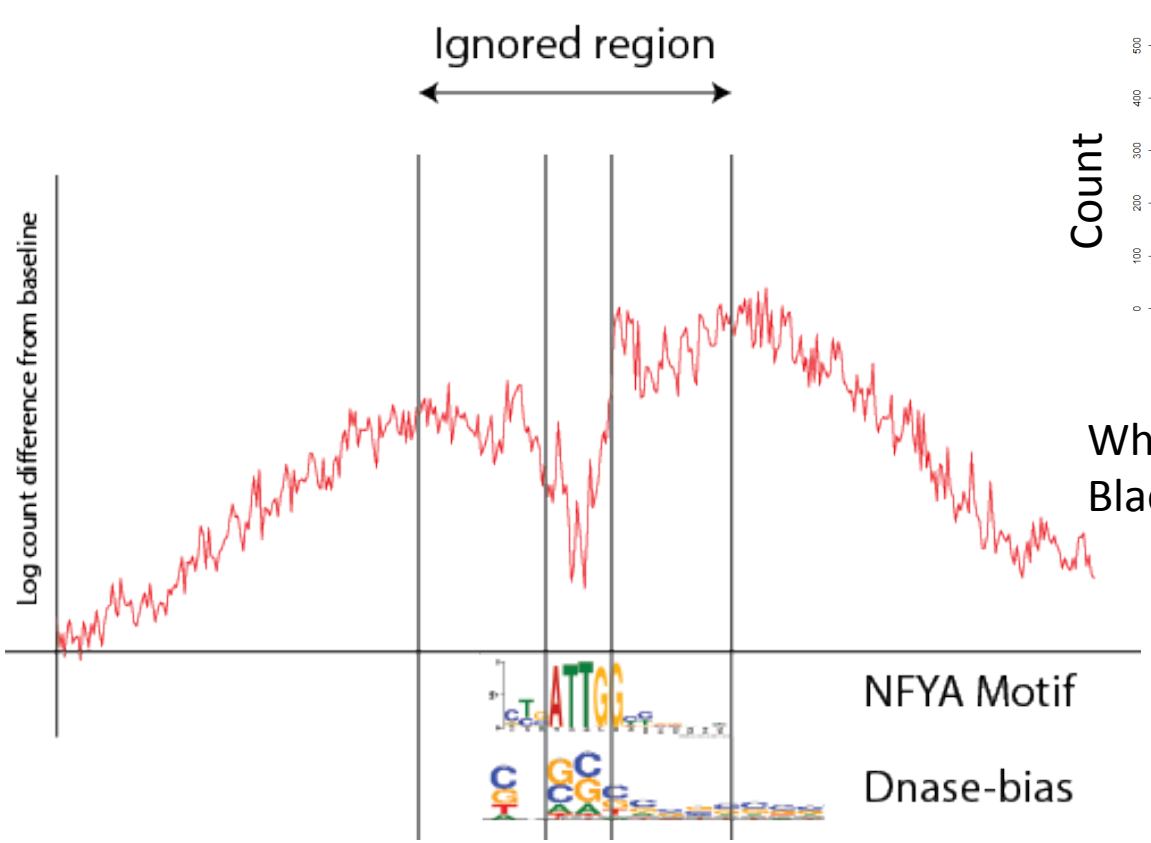
- We give the correlation between bases a special stationary structure

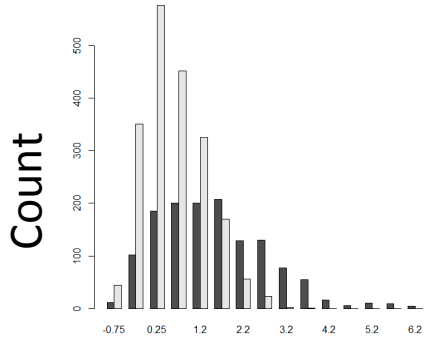
$$\Sigma_{i,j} = \sigma_0^2 \text{cor}(i-j)$$



Profiles are tested for significance to eliminate motif proximal DNase-bias

- Our test statistic is the absolute deviation of log-rates outside the motif match and its flank.
- The strongest DNase profiles, and those we focused on in our work all have effects far outside the motif match region.





White – random k-mers
Black – known TF motifs

PIQ model of TF binding

- The genome is modeled as the sum of smooth terms (λ_i) and factor specific terms.

$$c_i \sim Poisson(\exp(\lambda_i + \delta_j \gamma_i))$$

$$\lambda \sim MVN(\mu_0, \Sigma)$$

- γ_i is the factor specific profile,
- δ_j is a binding indicator.
- Each factor's binding is calculated as a log-likelihood ratio after adjusting for effects of nearby factor profiles.
- Profiles are estimated via the E-M algorithm.

Likelihood ratio testing for TF binding

$$P(c_i|I_j,\mu_i) = \int_{-\infty}^{\infty} Pois(c_i|exp(x + I_j\beta_{i-j}))N(x|E[\mu_i], \mathrm{Var}[\mu_i])$$

$$P(I_j|c_i,\mu_i) = \frac{\prod_i P(c_i|I_j = 1,\mu_i)}{\prod_i P(c_i|I_j = 0,\mu_i)}$$

- β_{i-j} is the factor specific profile,
- I_j is a binding indicator.
- Each factor's binding is calculated as a log-likelihood ratio after adjusting for effects of nearby factor profiles.

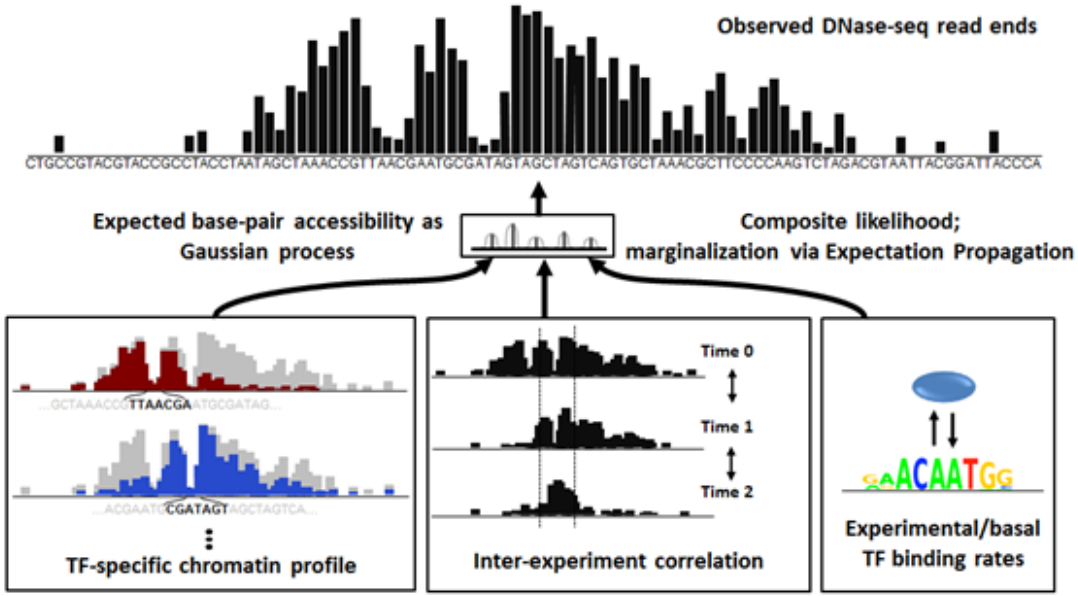
Add priors to log likelihood ratio and compare sum to null distribution for significance

$$L_j = f_j + g_j + \log P\left(I_j \mid c, \mu\right)$$

- f_j is a monotone motif prior
- g_j is a monotone count prior.
- Result is a rank list of calls; binary calls are made with a null distribution at p = 0.01

Integrated model

1. Robust model of TF binding that models overlapping profiles
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.



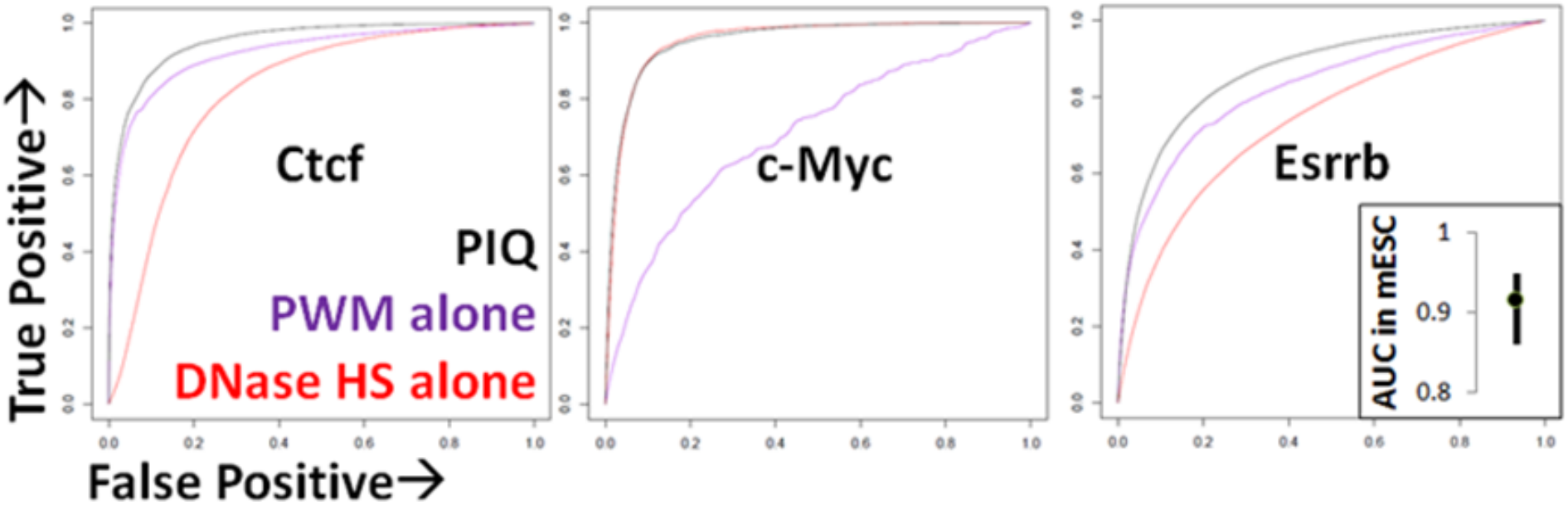
Source: Hashimoto, Tatsunori Benjamin. "Computation Identification of Transcription Factor Binding Using DNase-seq." PhD diss., Massachusetts Institute of Technology, 2014.

PIQ computational scaling

Attribute	Typical usage	Asymptotic scaling
Size of genome in bases	2.8 billion	N
Number of motifs	1500 Motifs	L
Number of experiments	10 experiments	K
Window size	400 bases	W
Number of CPUs	80 CPUs	M
Runtime	~ 1 day clock time ~ 80 days CPU time	$O(NLK/M + W^3K/M + K^3)$
Memory	~ 2Gb / CPU	$O(W^2K + K^2)$

Claim 1: PIQ is highly accurate at predicting TF binding

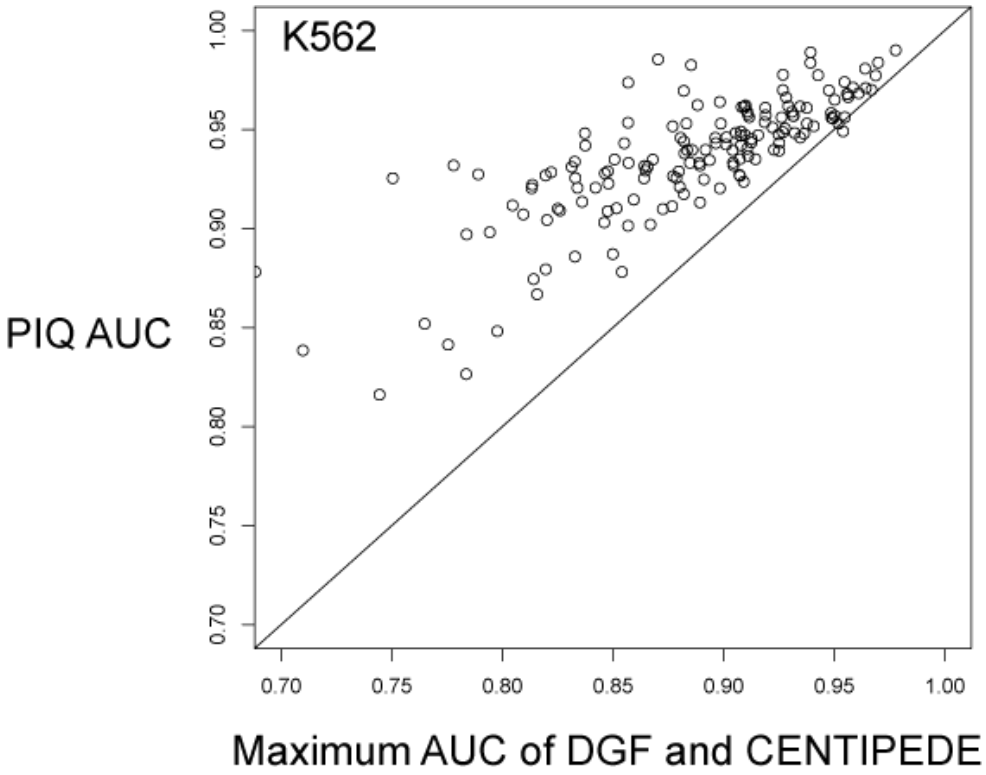
Receiver operating characteristic (ROC) curves show PIQ matches closely with ChIP-seq data.



Source: Sherwood, Richard I., Tatsunori Hashimoto, et al. "Discovery of Directional and Nondirectional Pioneer Transcription Factors by Modeling DNase Profile Magnitude and Shape." *Nature Biotechnology* 32, no. 2 (2014): 171-8.

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

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

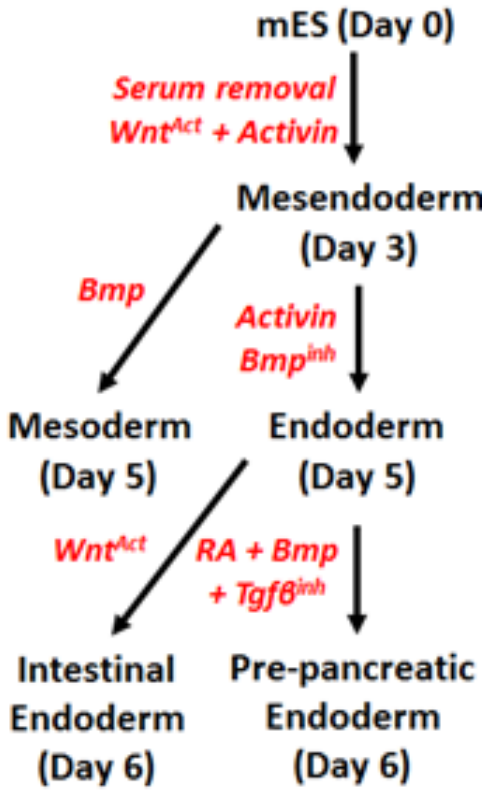


Claim 1: PIQ is highly accurate at predicting TF binding

- For certain factors concordance with ChIP-bound sites is AUC 0.9+
- If a factor is detectable via Dnase-seq PIQ shows high ppv (70%) with good coverage (50%)
- A factor is Dnase-detectable if
  - It has a strong binding motif
  - Binds in DNase-accessible regions
  - Has strong DNA binding affinity to protect from Dnase
- Of 302 ENCODE K562 ChIPs, 75 were strongly Dnase detectable.

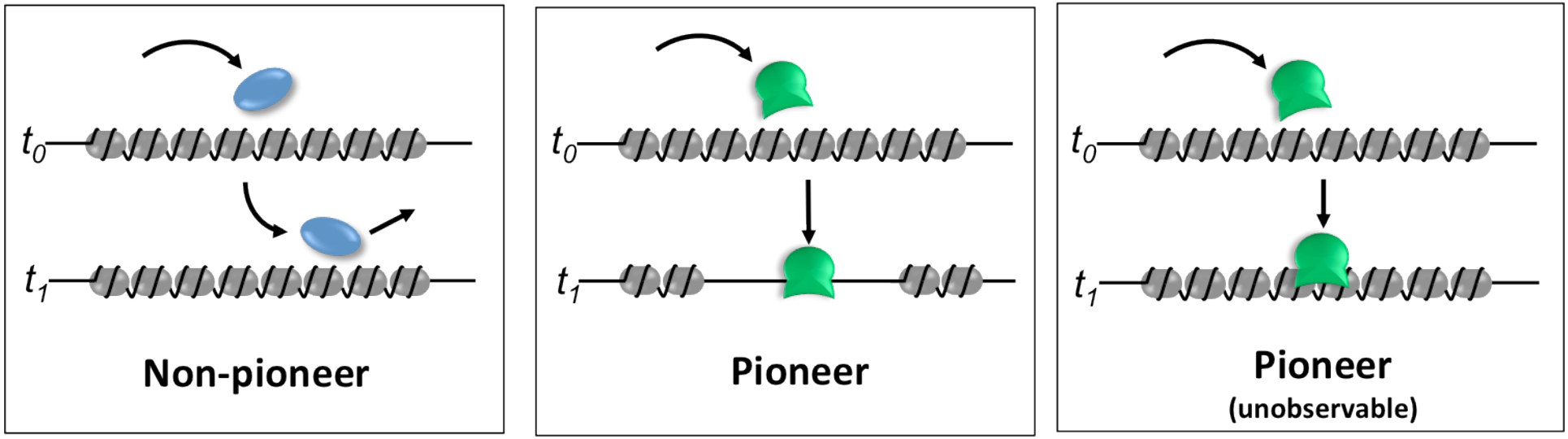
Claim 2: PIQ can identify pioneer factors that regulate proximal chromatin and binding

A typical TF has motif matches to hundreds of thousands of locations in the genome; why are only a few thousand motifs bound?



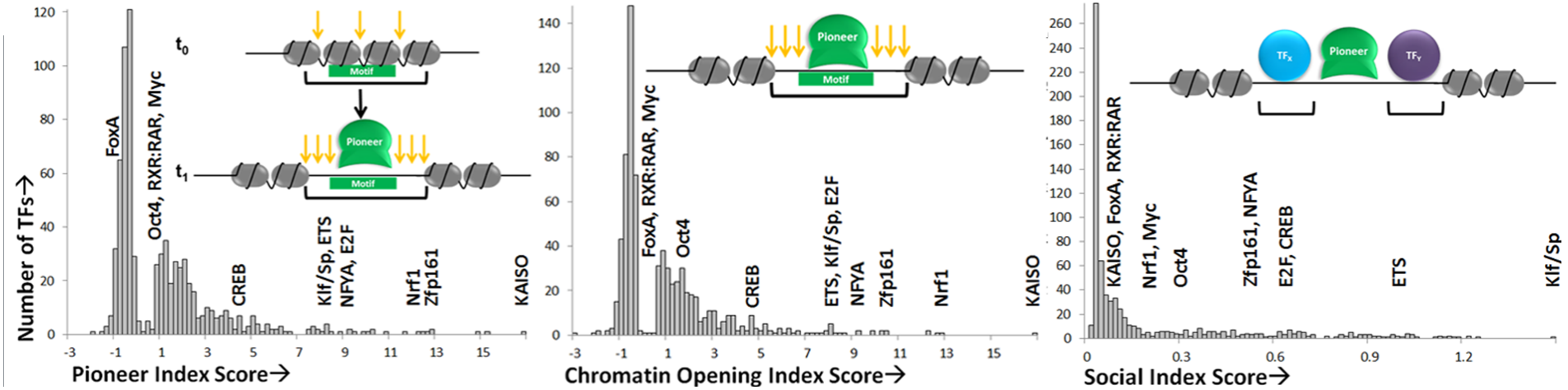
Systematic pioneer identification using PIQ

- Observe chromatin state change over time around all bound and unbound TF sequence motifs
  - Requires DNase-hypersensitive binding site
  - Requires TF sequence-specificity
- e.g. if at t_0 chromatin is inaccessible:



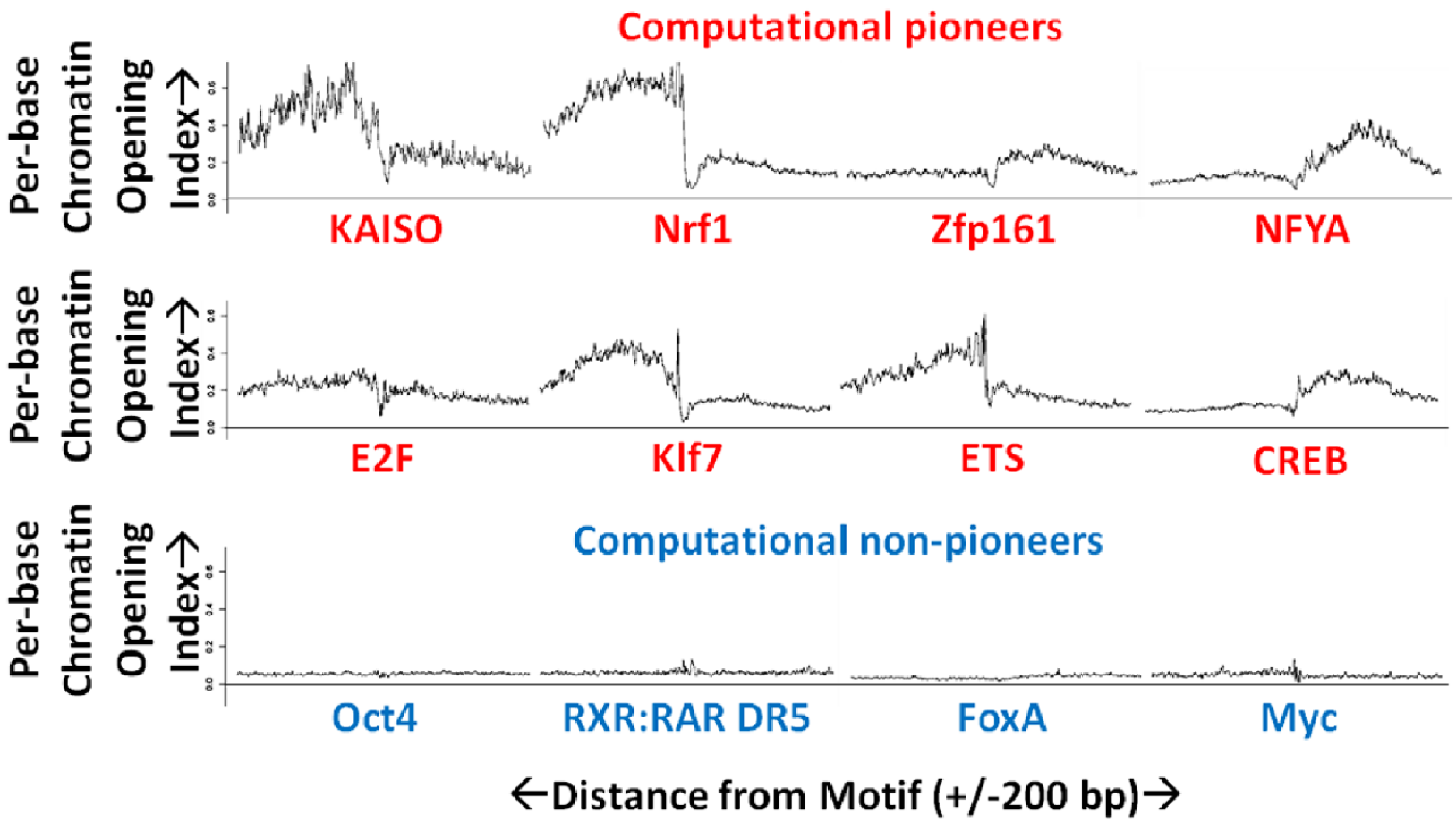
Claim 2: PIQ can identify pioneer factors that regulate proximal chromatin and binding

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



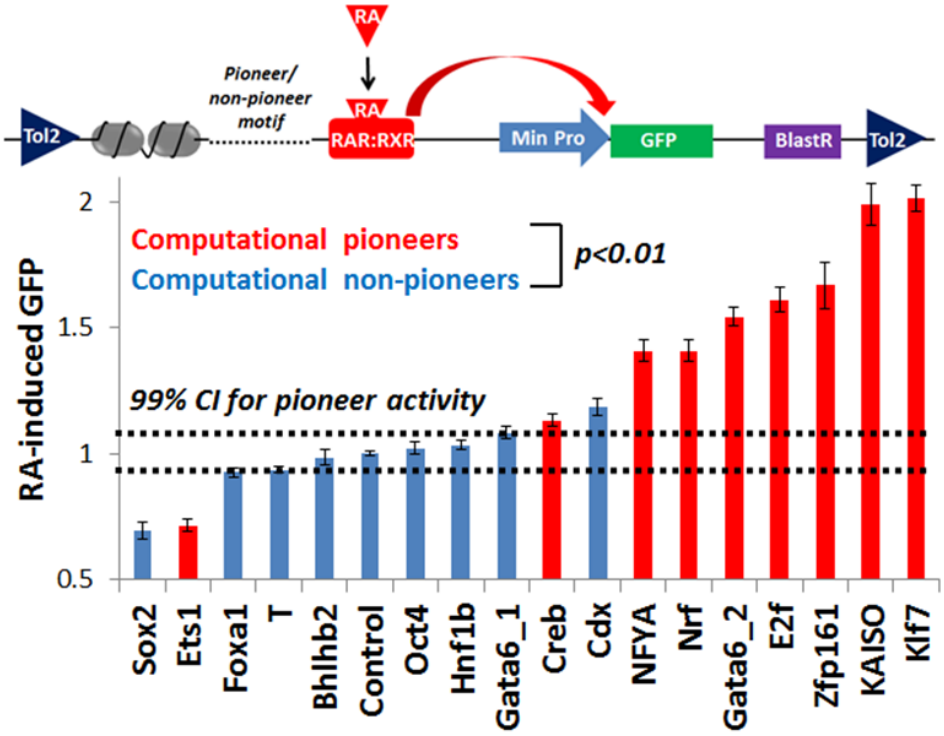
Source: Sherwood, Richard I., Tatsunori Hashimoto, et al. "Discovery of Directional and Nondirectional Pioneer Transcription Factors by Modeling DNase Profile Magnitude and Shape." *Nature Biotechnology* 32, no. 2 (2014): 171-8.

Pioneer TFs have identifiable profiles (n=120)



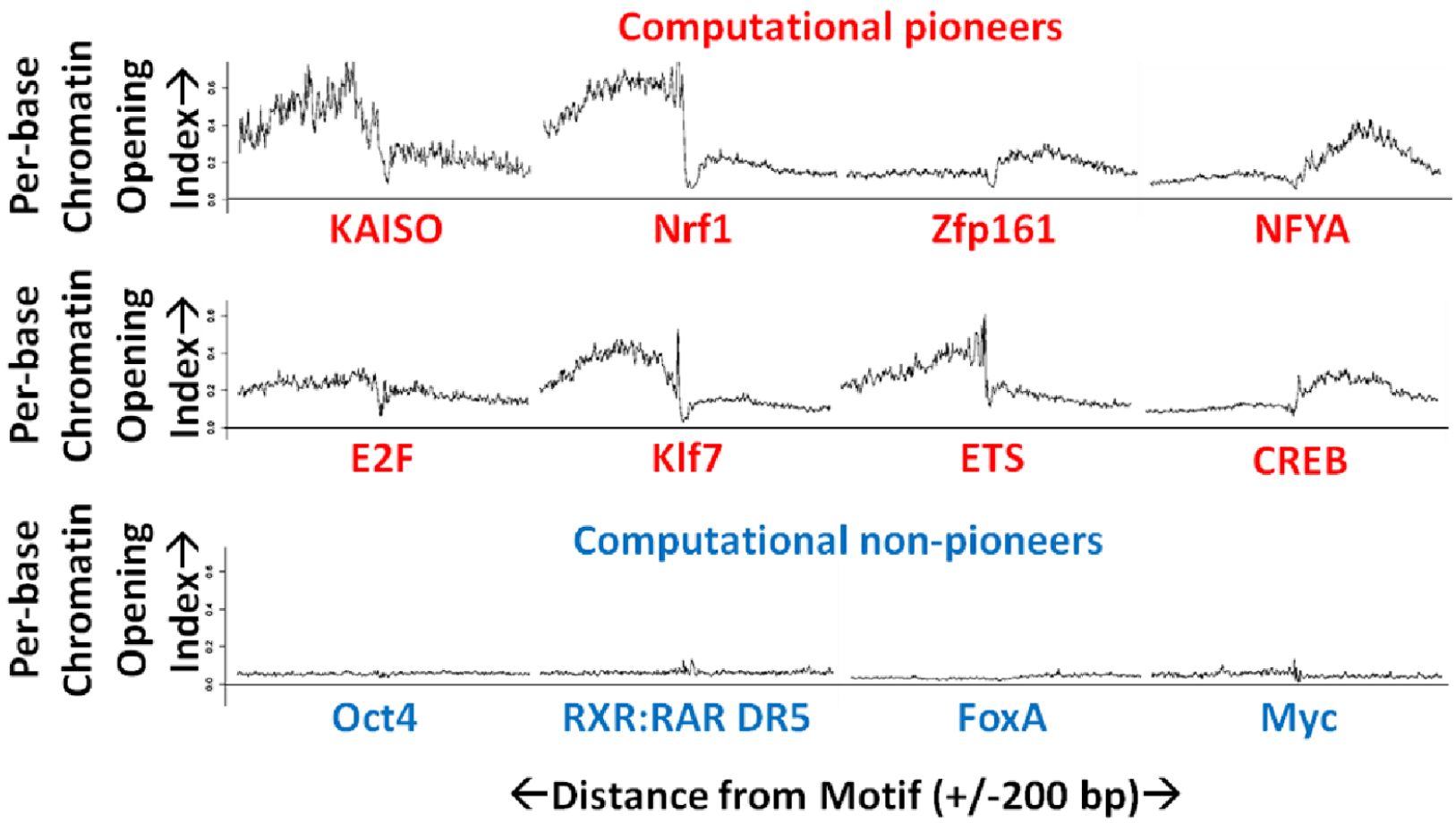
In vitro reporter assays recapitulate computational predictions

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



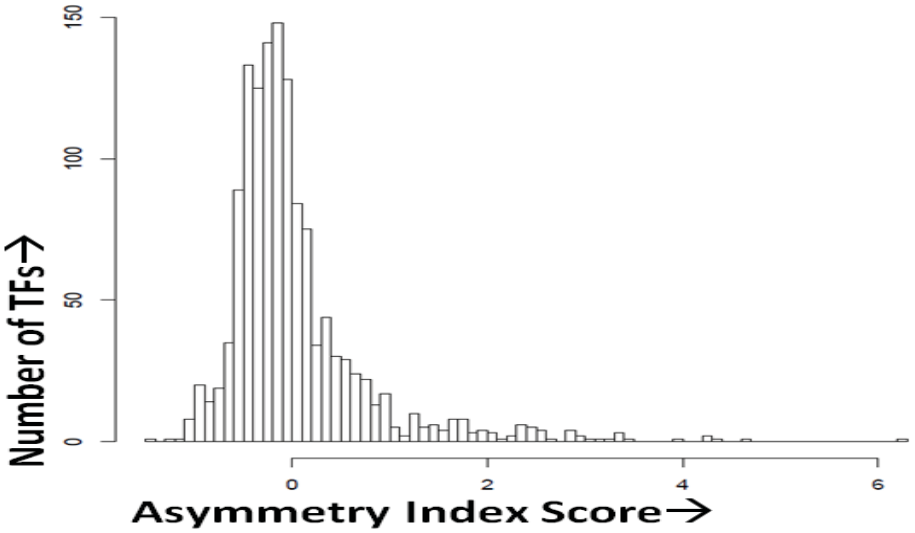
Source: Sherwood, Richard I., Tatsunori Hashimoto, et al. "Discovery of Directional and Nondirectional Pioneer Transcription Factors by Modeling DNase Profile Magnitude and Shape." *Nature Biotechnology* 32, no. 2 (2014): 171-8.

Pioneer TFs have identifiable profiles

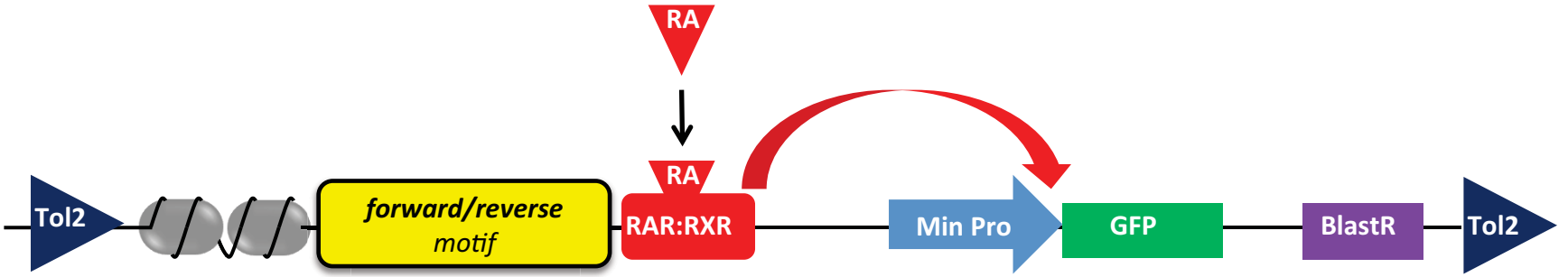


Claim 3: Certain pioneer TFs are directional

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

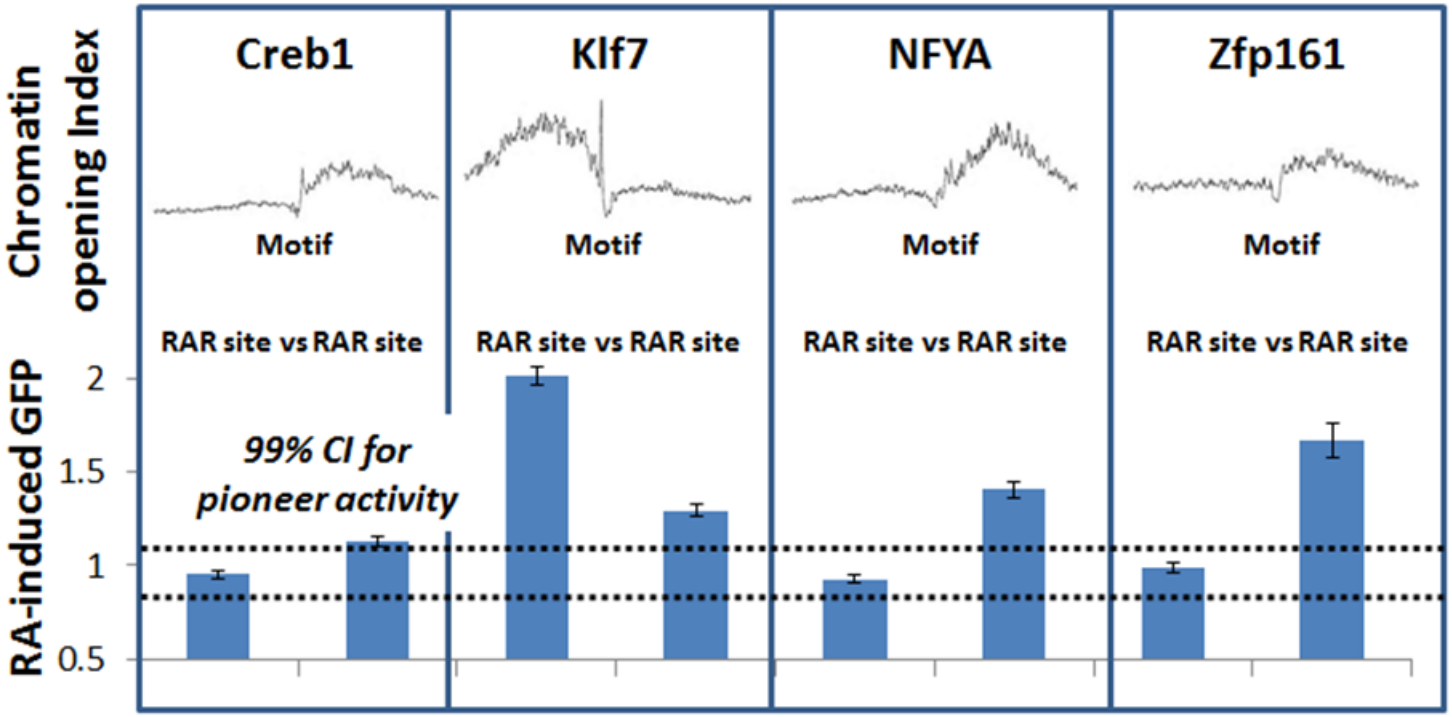


- Biological validation by testing both motif orientations



Claim 3: Certain pioneer TFs are directional

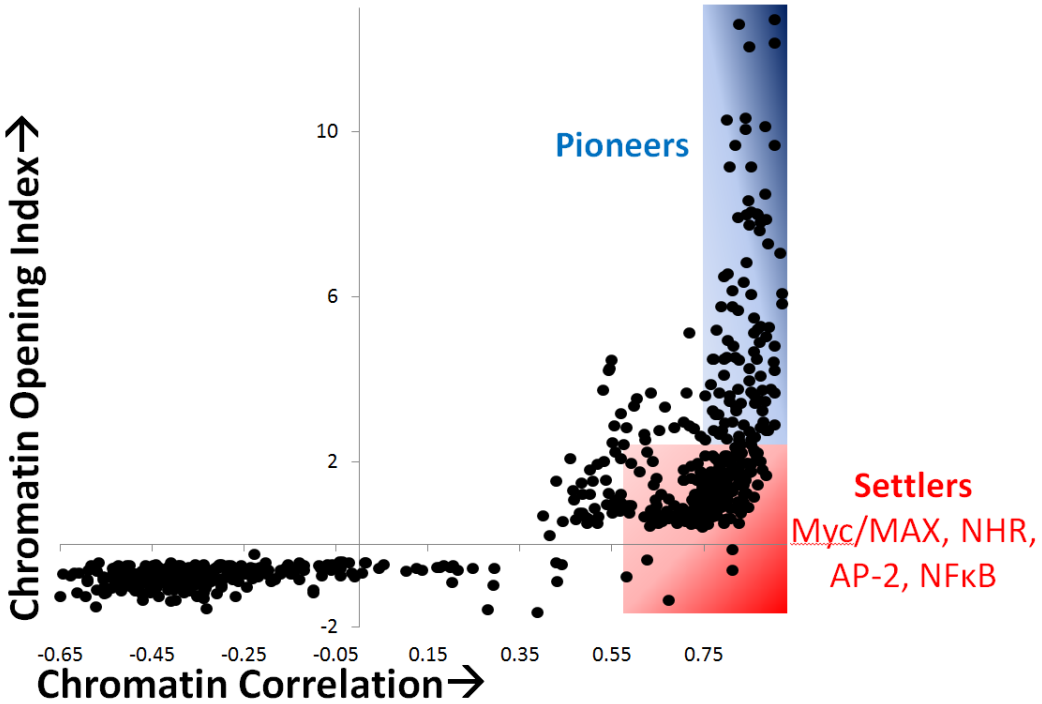
Orienting the motif direction in the reporter recapitulates expected directional behaviors.



Source: Sherwood, Richard I., Tatsunori Hashimoto, et al. "Discovery of Directional and Nondirectional Pioneer Transcription Factors by Modeling DNase Profile Magnitude and Shape." *Nature Biotechnology* 32, no. 2 (2014): 171-8.

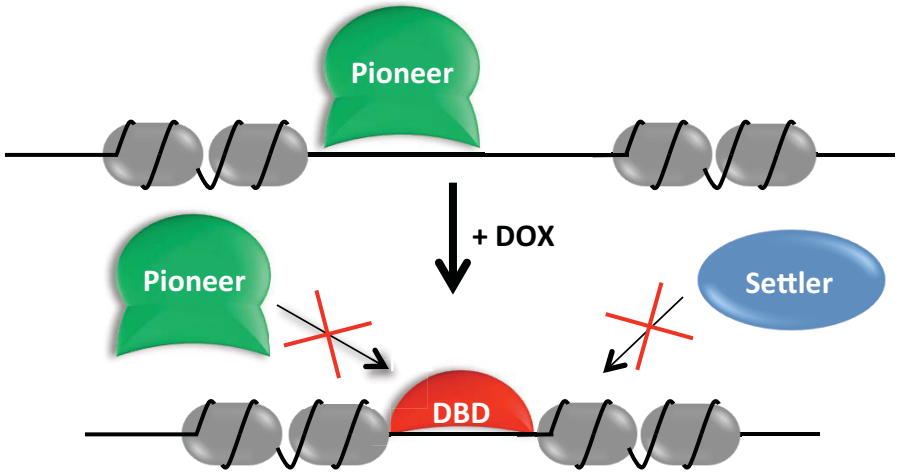
Claim 4: Settlers factors follow pioneer factor binding and loss of pioneer binding causes chromatin to return to a closed state

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



We validate pioneer/settler model via a dominant-negative competition assay

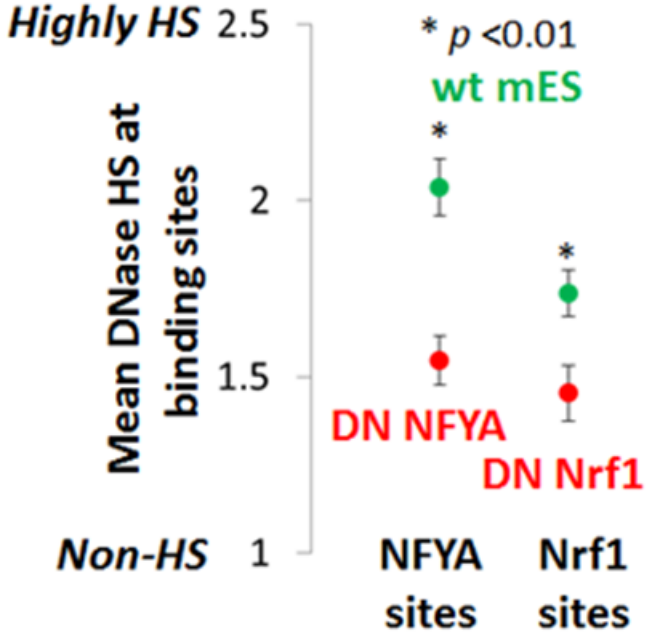
- Construct pioneer DBD protein that retains no pioneering function
- Induction of DBD protein competes for genomic binding, reducing local chromatin accessibility settlers rely on
- Compare proximal chromatin openness
- Compare ChIP levels for neighboring settler binding



Source: Sherwood, Richard I., Tatsunori Hashimoto, et al. "Discovery of Directional and Nondirectional Pioneer Transcription Factors by Modeling DNase Profile Magnitude and Shape." *Nature Biotechnology* 32, no. 2 (2014): 171-8.

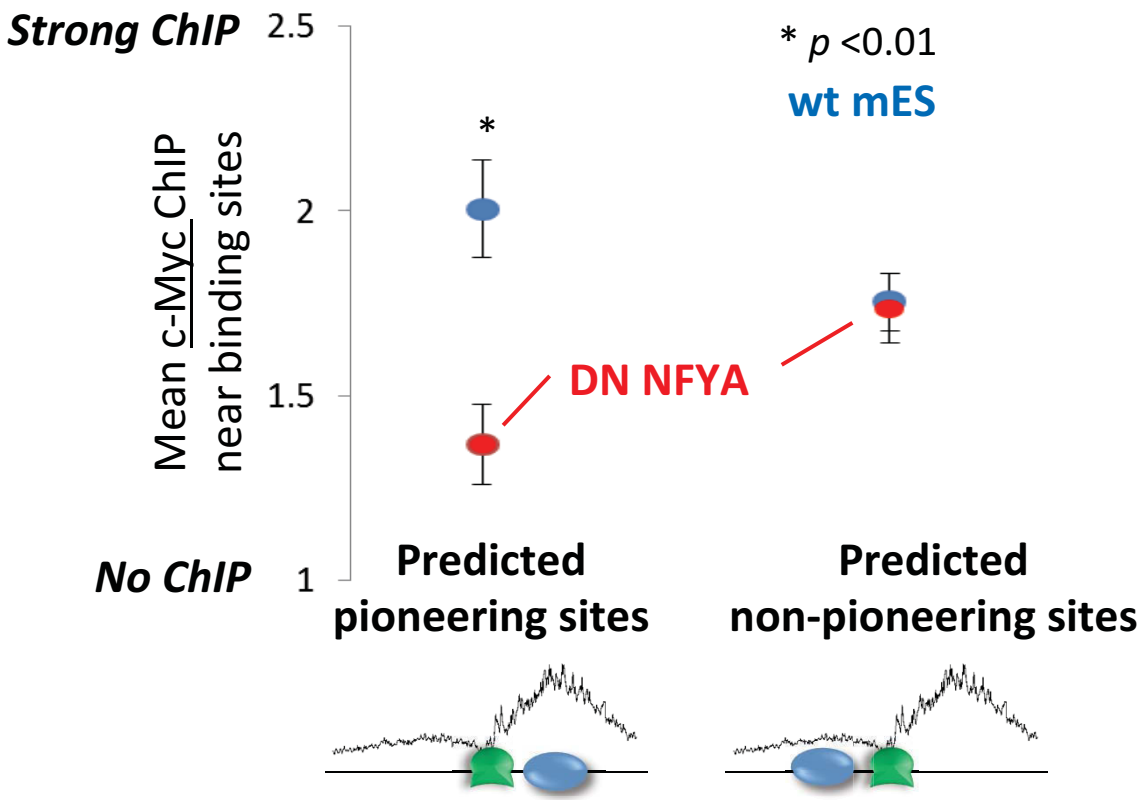
Dominant negative pioneers reduce proximal DNase HS

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.

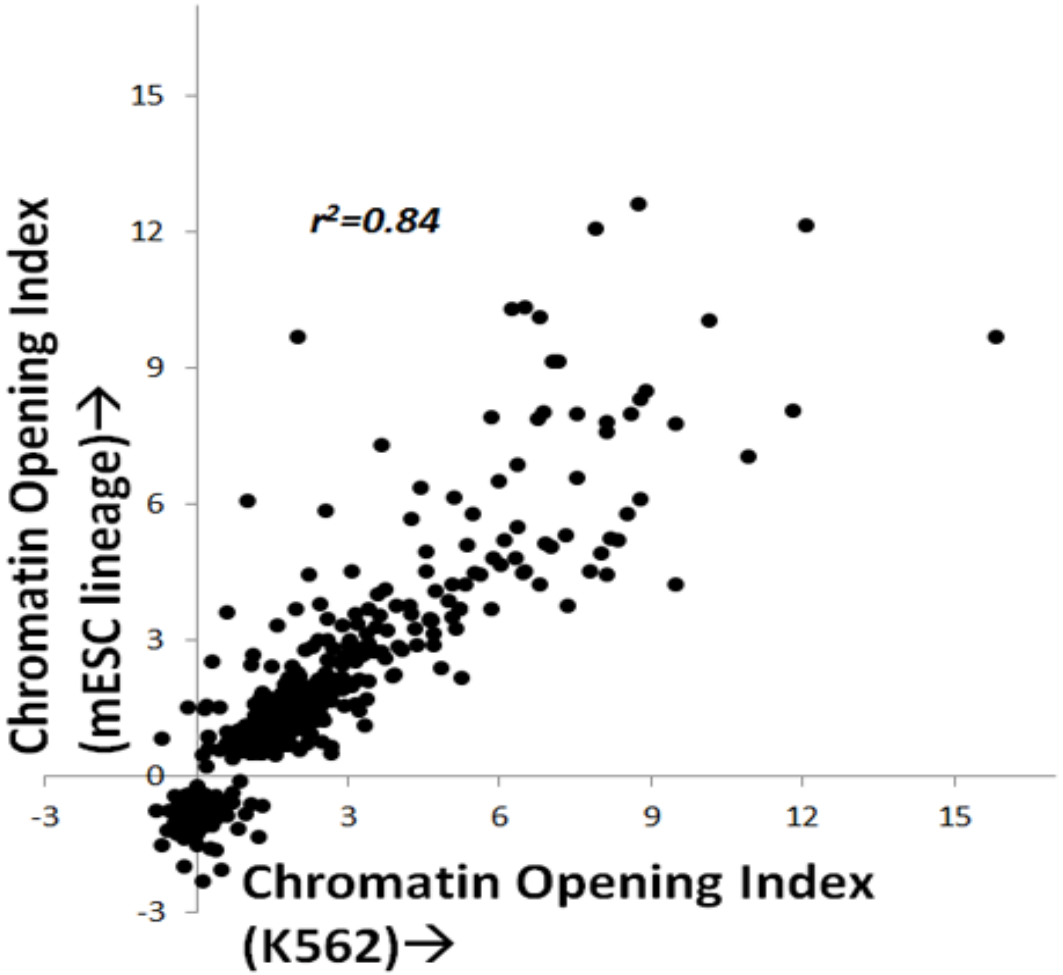


Dominant negative NFYA reduces binding of downstream c-Myc

- Relative c-Myc ChIP-qPCR over genomic binding regions with either NFYA / c-Myc or c-Myc / NFYA to test asymmetry



Pioneers appear to be conserved between human/mouse



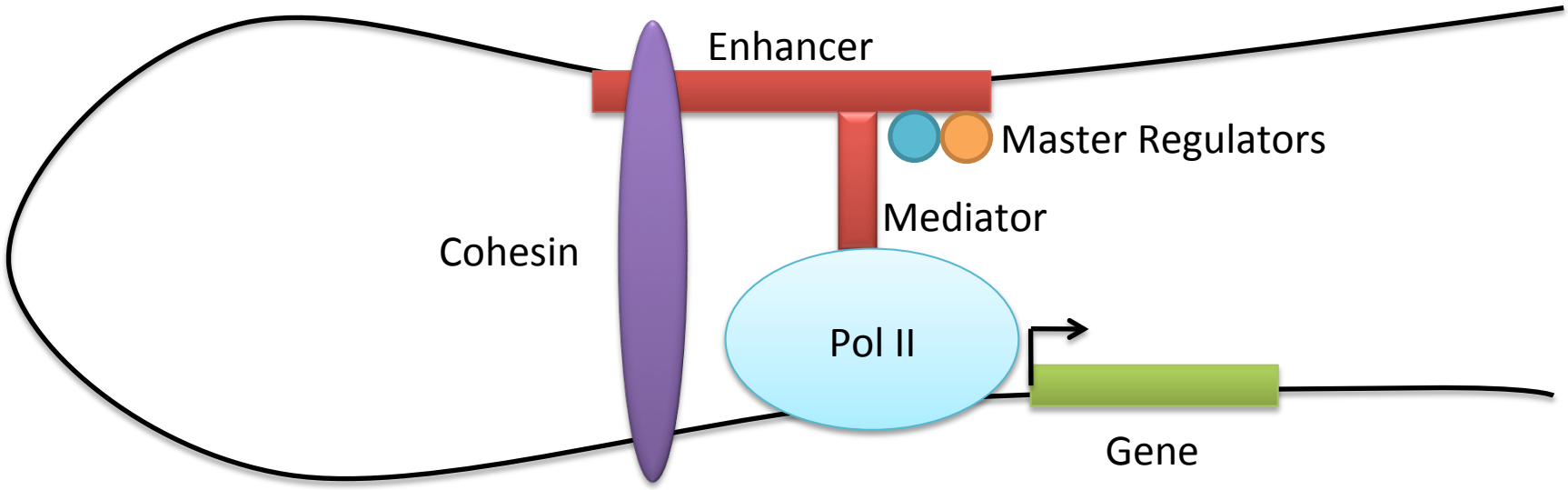
Overview of Results

- PIQ is highly accurate at predicting transcription factor (TF) binding from DNase-seq data
- PIQ can identify pioneer factors regulate proximal chromatin opening and TF binding
- Certain pioneer TFs are directional
- Settlers factors follow pioneer factor binding and loss of pioneer binding causes chromatin to return to a closed state

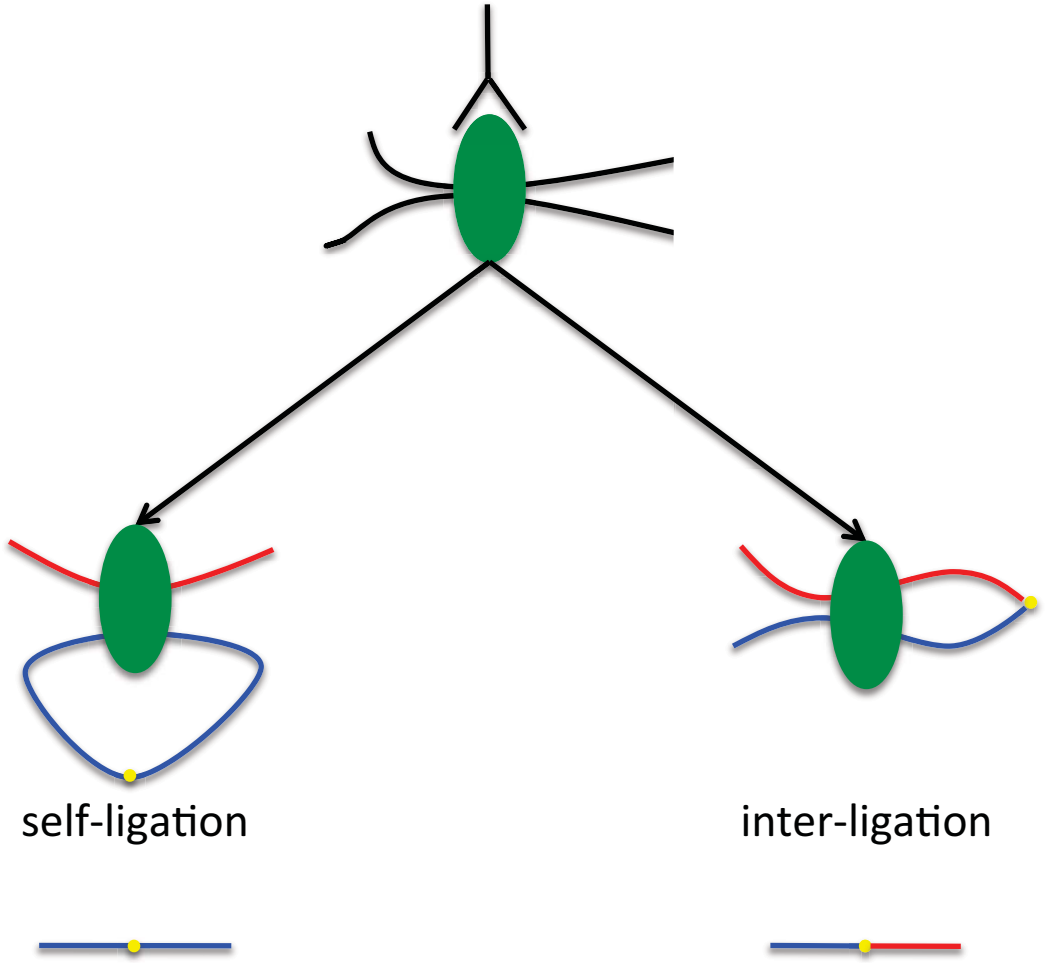
Today's Narrative Arc

1. We can break the epigenetic "code" that describes the function and state of genome elements using computational methods. Epigenetic state regulates gene function without changing primary DNA sequence. Epigenetic state includes histone marks, DNA methylation, and chromatin openness.
2. We can estimate the protein occupancy of the genome and discover pioneer factors with DNase-seq via computational methods.
3. **We can map enhancers to their regulatory targets with the computational analysis of ChIA-PET data (and similar technologies)**

Enhancers regulate distal target genes by genome looping

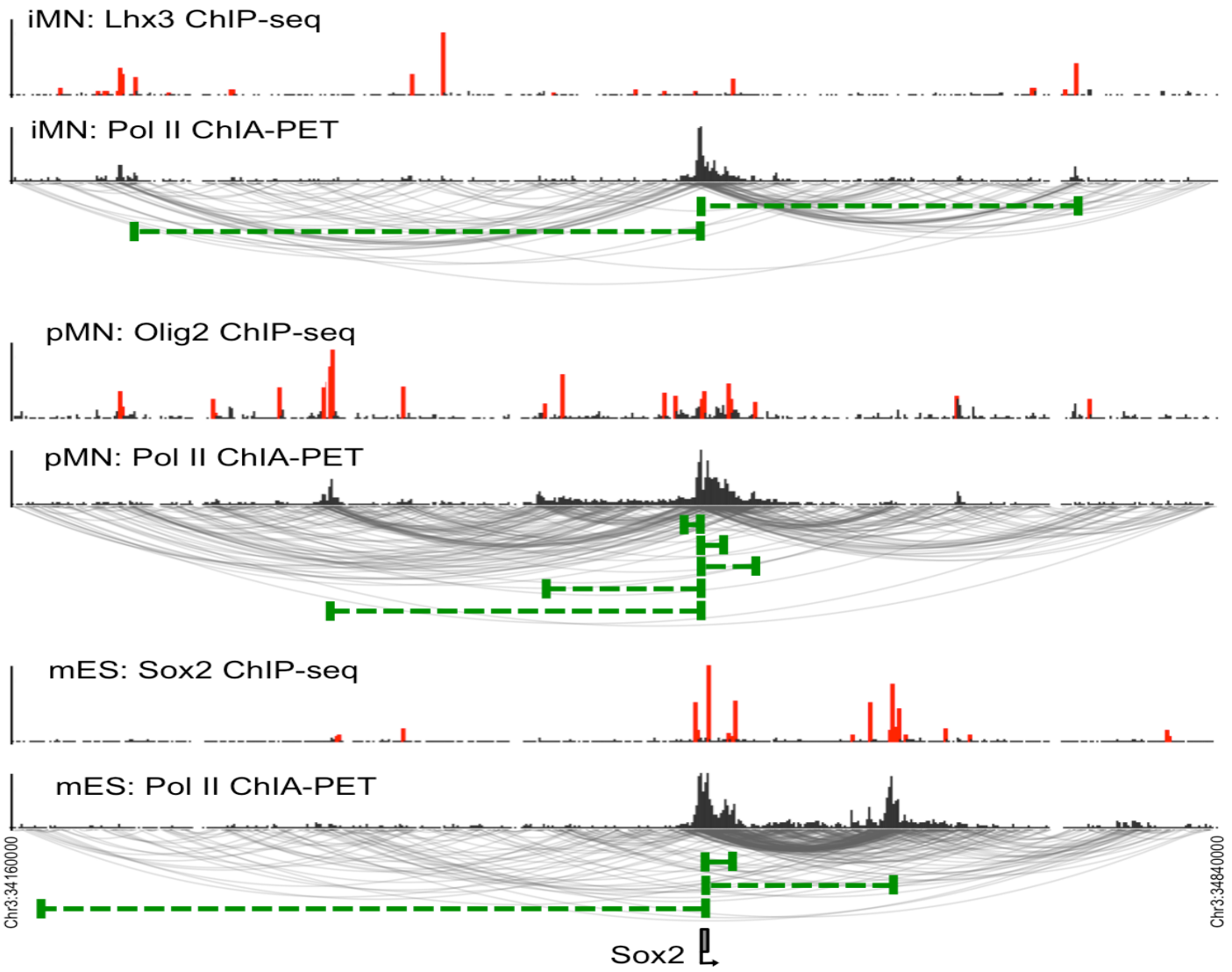


ChIA-PET protocol - After IP of RNA Pol II, sonnication, and ligation, ligation products are sequenced



The ChIA-PET Protocol

ChIA-PET discovered enhancer linkages



The significance of observing *I* inter-ligation events between two binding events *A* and *B* can be calculated using a hypergeometric test

Let $I_{A,B}$ be the number of inter-ligation events between binding events A and B. Let c_A and c_B be the number of ligation event ends associated with A and B, respectively. Let N be the total number of ligation events ends. The null hypothesis assumes that each ligation event end has an equal probability of ligating with any other end. Then, under the null hypothesis:

$$P(I_{A,B}|N, c_A, c_B) = \frac{\binom{c_A}{I_{A,B}}\binom{N-c_A}{c_B-I_{A,B}}}{\binom{N}{c_B}}$$

$$p = \sum_{i=I_{A,B}}^{\min\{c_A,c_B\}} P(i|N, c_A, c_B)$$

Issues with ChIA-PET

1. High false negative rate. Libraries produced are not complex enough to permit further discovery by additional sequencing.
2. Specific to a protein (RNA Polymerase II in our example)
3. Hi-C and derivatives may solve these problems eventually

Estimating total events from overlap

Imagine we perform two biological replicates of an experiment and obtain 1000 events in each, of which 900 are identical

We can use a hypergeometric model to infer how many possible events exist (*N*) given two sample sizes (*m* and *n*) and an overlap (*k*):

$$\hat{N} = \underset{N}{\operatorname{argmax}} \left[P(X = k; N, m, n) \right]$$

Using this model, we predict ~1100 total events

Approximate closed form solution for total number of events

- The ML estimate of *N* is approximately:

$$\hat{N}(m,n,k) = \frac{mn}{k}$$

- One way to see this is by using the normal approximation of the binomial approximation to the hypergeometric distribution:

$$P(X=k;N,m,n) \approx \text{Binomial}\left(X=k;n=n,p=\frac{m}{N}\right)$$
$$\approx \text{Normal}\left(X=k;\mu=\frac{mn}{N},\sigma^2=\frac{mn}{N}(1-\frac{m}{N})\right)$$

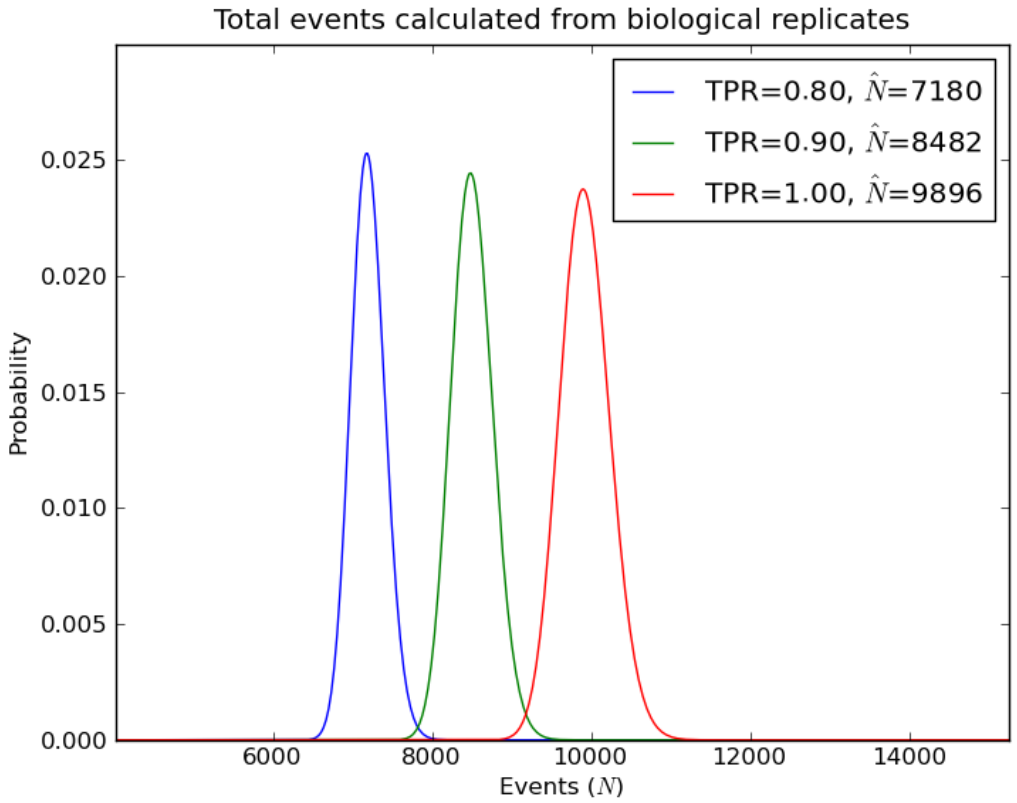
Allowing for false positive events

- What if some events in each replicate are false positives? Then we will overestimate the total event count
- We can assume that overlapping (shared) events are true positives and that (1 – *f*) of the remaining events are false negatives, where f is the true positive rate (TPR)
- This approximation lets us update *m* and *n* and apply the same model:

$$m' = (1 - f)(m - k) + k$$
$$n' = (1 - f)(n - k) + k$$

A higher true positive rate estimates more total events with a fixed overlap

- Replicate A had 3811 events, replicate B had 1384 events
- The overlap was 533 events
- Likelihood plots versus *N* for several true positive rates (TPR):



Today's Narrative Arc

1. Using computational methods we can break the epigenetic "code" that describes the function and state of genome elements. Epigenetic state regulates gene function without changing primary DNA sequence. Epigenetic state includes histone marks, DNA methylation, and chromatin openness.
2. We can estimate the protein occupancy of the genome and discover pioneer factors with DNase-seq via computational methods.
3. We can map enhancers to their regulatory targets with the computational analysis of ChIA-PET data (and similar technologies)

Today's Computational Methods

1. Dynamic Bayesian Networks
2. Factor binding classification using a log likelihood ratio
3. Hypergeometric distribution

FIN

7.91J / 20.490J / 20.390J / 7.36J / 6.802J / 6.874J / HST.506J Foundations of Computational and Systems Biology
Spring 2014