Epigenetics - Predicting TF binding with DNase-Seq and PIQ

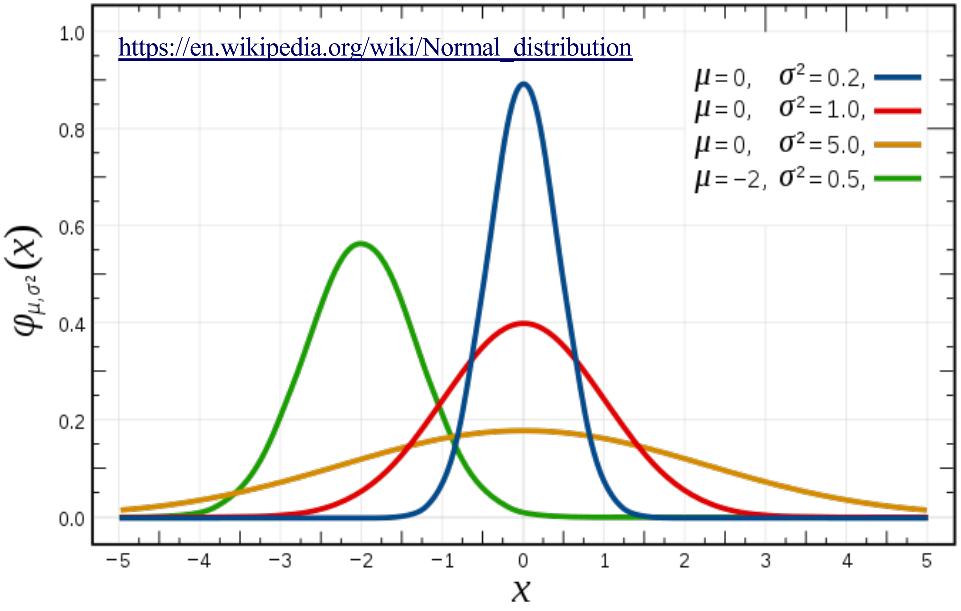
BMI/CS 776
www.biostat.wisc.edu/bmi776/
Spring 2020

Daifeng Wang
daifeng.wang@wisc.edu

Gaussian distribution

- A random variable, $x \sim \mathcal{N}(\mu, \sigma^2)$

$$\mathrm{pdf}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



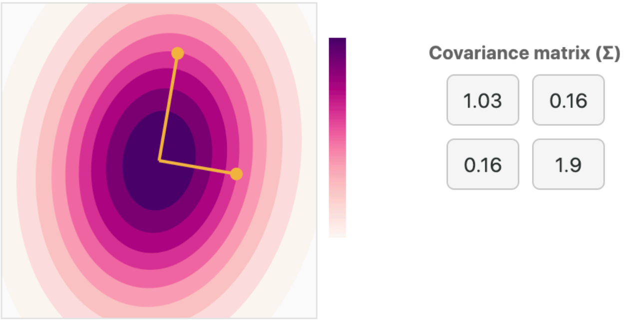
https://en.wikipedia.org/wiki/Normal_distribution

- *X* is # of mapped reads at a position
  - μ is average reads, σ^2 show how reads fluctuate from average across regions

Multivariate Gaussian distributions

- Multiple random variables
  - $\vec{x} = [x_1\ x_2\ \dots\ x_n]^T \sim \mathcal{N}(\vec{\mu}, \boldsymbol{\Sigma})$
  - $\text{pdf}(\vec{x}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^T \boldsymbol{\Sigma}^{-1}(\vec{x}-\vec{\mu})}$
  - Covariance matrix

    $\boldsymbol{\Sigma} = E[(\vec{x} - \vec{\mu})\ (\vec{x} - \vec{\mu})^T]$



https://distill.pub/2019/visual-exploration-gaussian-processes/#Multivariate

- # of reads at Position i and Position j
  - $[x_i\ x_j] \sim \mathcal{N}([\mu_i\ \mu_j], \begin{bmatrix} \sigma_i^2 & E[(x_i - \mu_i)(x_j - \mu_j)] \\ E[(xi - \mu_i)(xj - \mu_j)] & \sigma_j^2 \end{bmatrix})$

Kernel function for covariance

- Covariance measures “similarity” of x_i and x_j
  - $k(i,j) = E[(xi - \mu_i)(xj - \mu_j)]$
- Replace by other kernel functions defining covariance
  - Radial Basis Function (RBF)

$$k_{RBF}(i,j) = \sigma^2 \exp(-\frac{(i-j)2}{2l^2})$$

- Also, mean functions $\mu(i), \mu(j)$

Gaussian process (GP)

- A stochastic process with mean function $\mu(.)$ and covariance function $k(.,.)$ so that any finite set of multi-variates $[x_1\ x_2\ \dots\ x_n]$ is from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{K})$
  - $\boldsymbol{\mu}$ is *n*-dimension vector with i^{th} element = $\mu(i)$
  - $\boldsymbol{K}$ is a symmetric matrix (*n* x *n*) and $\boldsymbol{K}_{i,j} = k(i,j)$
- $x_{(.)} \sim \boldsymbol{\mathcal{GP}}(\mu(.), k(.,.))$
  - Infinite number of random variables, $x_1\ x_2\ \dots$

Gaussian process regression

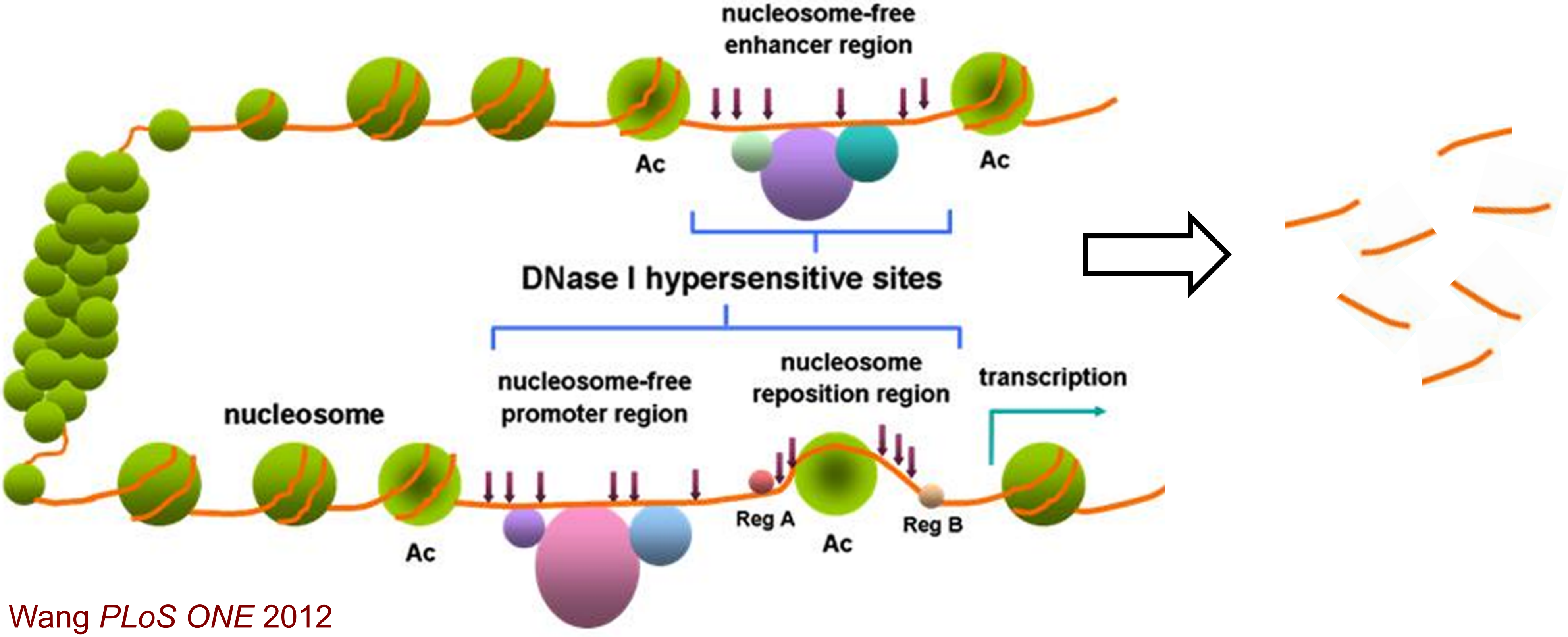
- $x(i)$ is a regression function to predict # of reads y_i on position i
  - $y_i = x(i) + \varepsilon_i$, where ε_i is noise ~ $\mathcal{N}(0, \sigma^2)$
- $\boldsymbol{\mathcal{GP}}(0,\ k(.,.))$ as prior for regression function to predict a distribution of y
  - Use training data *S*= $\{p,\ y_p\}$, $p \in \{1,2, \ldots\}$, predict posterior distribution(y_q|*S*,*T*) ~$\mathcal{N}(\vec{\mu}^{\,*}, \boldsymbol{\Sigma}^{*})$ from testing data *T*=$\{q,\ y_q\}$

$$\vec{\mu}^{\,*} = \mathbf{K}(\vec{p},\vec{q})(\mathbf{K}(\vec{p},\vec{p}) + \sigma^2\mathbf{I})^{-1}\overrightarrow{y_p}$$

$$\boldsymbol{\Sigma}^{*} = \mathbf{K}(\vec{q},\vec{q}) + \sigma^2\mathbf{I} - \mathbf{K}(\vec{p},\vec{q})(\mathbf{K}(\vec{p},\vec{p}) + \sigma^2\mathbf{I})^{-1}\mathbf{K}(\vec{p},\vec{q})$$

DNase I hypersensitive sites

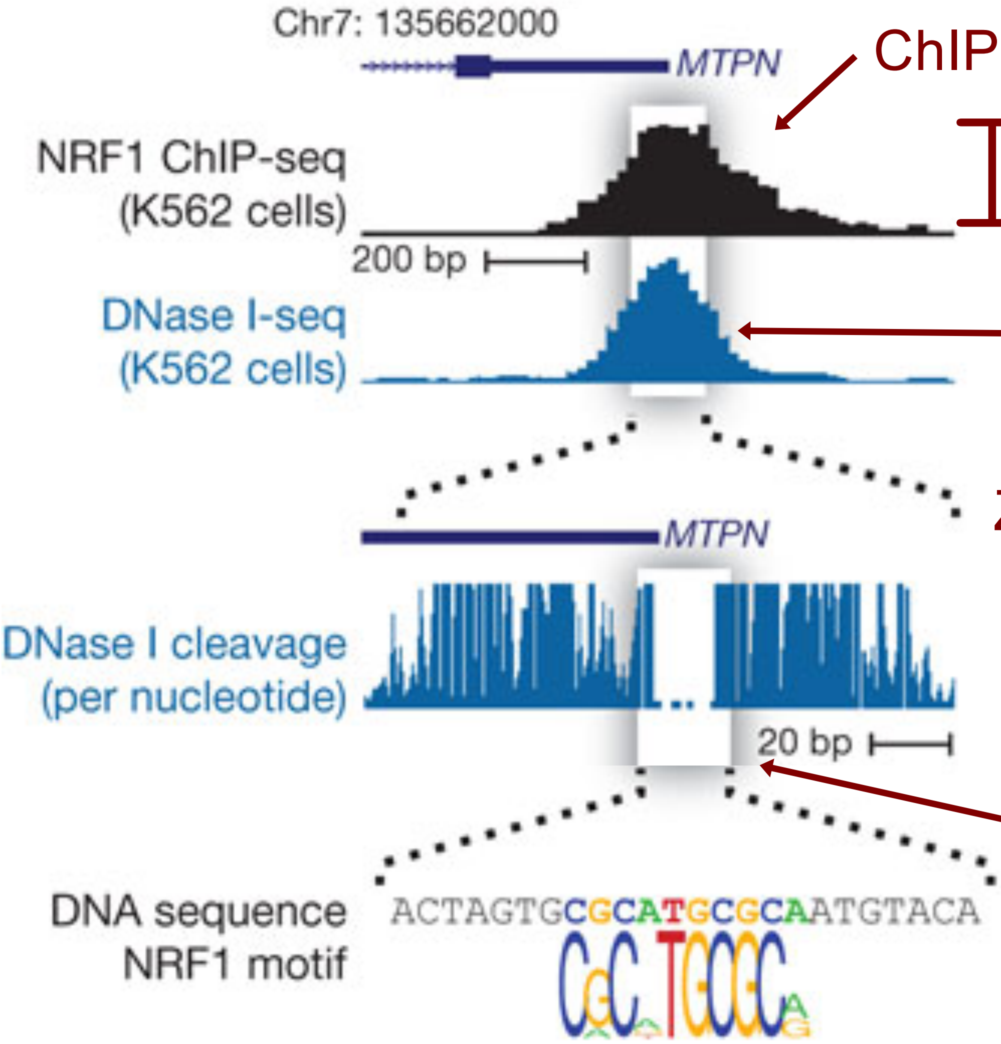
- Arrows indicate DNase I cleavage sites
- Obtain short reads that we map to the genome



Wang *PLoS ONE* 2012

DNase I footprints

- Distribution of mapped reads is informative of open chromatin and specific TF binding sites



ChIP-Seq peak

Read depth at each position

Nucleosome free
“open” chromatin

Zoom in

TF binding prevents
DNase cleavage leaving
Dnase I “footprint”, only
consider 5′ end

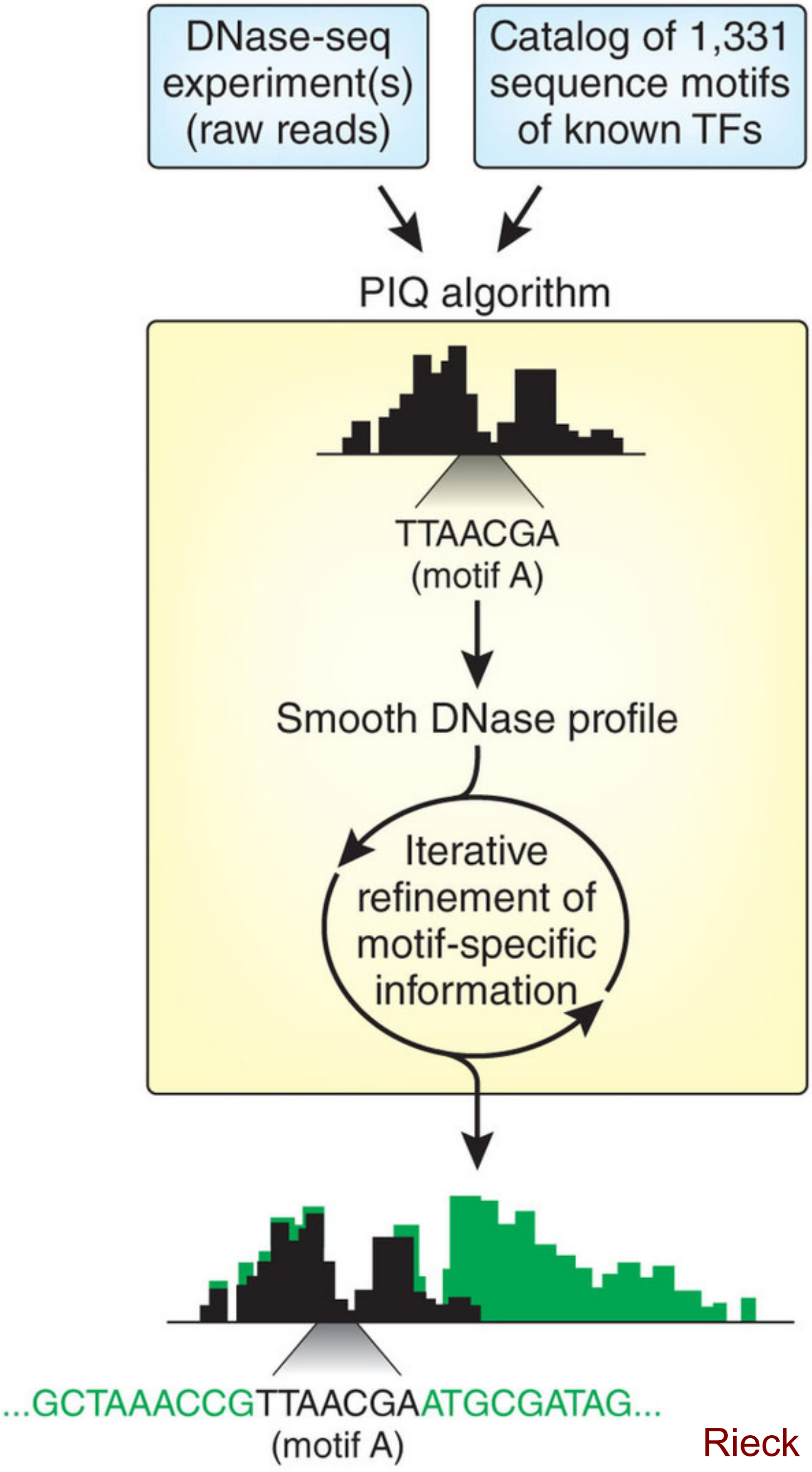
Neph *Nature* 2012

DNase I footprints to TF binding predictions

- DNase footprints suggest that ***some*** TF binds that location
- We want to know ***which*** TF binds that location
- Two ideas:
  - Search for DNase footprint patterns, then match TF motifs
  - Search for motif matches in genome, then model proximal DNase-Seq reads

We'll consider this approach for TF/motif specific effects

Protein Interaction Quantification (PIQ)



DNase-seq experiment(s) (raw reads)

Catalog of 1,331 sequence motifs of known TFs

PIQ algorithm

TTAACGA (motif A)

Smooth DNase profile

Iterative refinement of motif-specific information

...GCTAAACCGTTAACGAATGCGATAG...

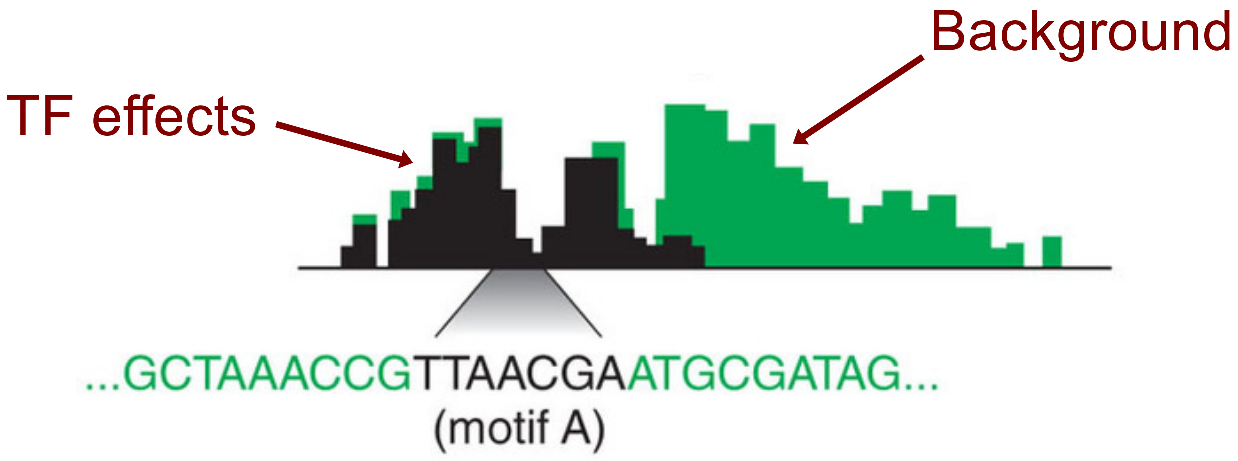
(motif A)

- Sherwood et al. *Nature Biotechnology* 2014
- **Given**: TF motifs and DNase-Seq reads
- **Do**: Predict binding sites of each TF

Rieck and Wright *Nature Biotechnology* 2014

PIQ main idea

- With no TF binding, DNase-Seq reads come from some background distribution
- TF binding changes read density in a *TF-specific* way

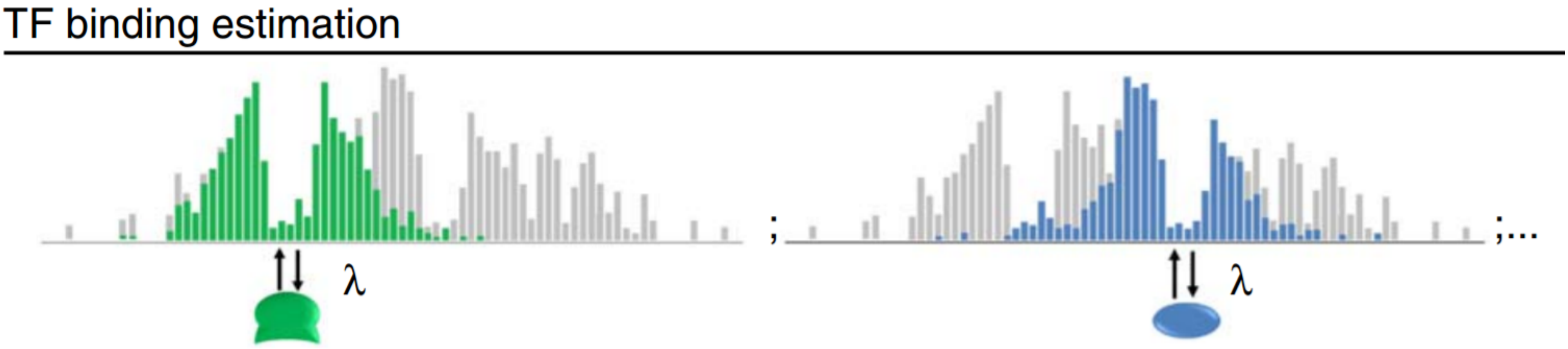


TF effects

Background

PIQ main idea

- Shape of DNase peak and footprint depend on the TF



TF A

TF B

Sherwood *Nature Biotechnology* 2014

Gaussian processes

- Can model and smooth sequential data
- Bayesian approach
- Jupyter notebook demonstration

PIQ features

- We’ll discuss
  - Modeling the DNase-Seq background distribution
  - How TF binding impacts that distribution
  - Priors on TF binding
  - Single experiment/strand, single factor
- We’ll skip
  - Modeling multiple replicates or conditions, cross-experiment and cross-strand effects
  - Expectation propagation, iteratively approximating probability distributions
  - TF hierarchy: pioneers, settlers, migrants

Algorithm preview

- Identify candidate binding sites with PWMs
- Build a probabilistic model of the DNase-Seq reads
- Estimate TF binding effects
- Estimate which candidate binding sites are bound
- Predict pioneer, settler, and migrant TFs

DNase-Seq background

- Each replicate is noisy, don’t want to over-interpret this noise
    - Only counting density of 5′ ends of reads

- Manage two competing objectives
    - Smooth some of the noise
    - Don’t destroy base pair resolution signal

Raw Dnase-seq reads from GP

- Log-read rate per base *u* from a Gaussian Process $\mathcal{N}(\vec{\mu}_0, \boldsymbol{\Sigma})$
  - Positions *i* and *j*: u_i and u_j , $\boldsymbol{\Sigma}_{i,j} = \sigma_0 k(|i-j|)$
  - *e.g.*, *k* is correlation

- # of reads (read counts) x_i at Position *I*
  - $x_i \sim$Poisson(exp(u_i))

- Estimate a background GP(μ_0, σ_0, k, $\boldsymbol{\Sigma}^{-1}$)
  - Supplement C.5

TF-specific DNase profile

- Adjust the log-read rate by a TF-specific effect at binding sites

DNase profile
for factor *l*

Whether site
m is bound

$$\widehat{\mu_{i,l}} = \mu_i + \begin{cases} \beta_{i-y_m,l} & |i-y_m| \le W\ and\ I_m = 1 \\ 0 & otherwise \end{cases}$$

DNase log-read
rate adjusted for
binding of factor *l*

Midpoint location
of binding site *m*

Window size

DNase log-read rate
at position *i* from
Gaussian process

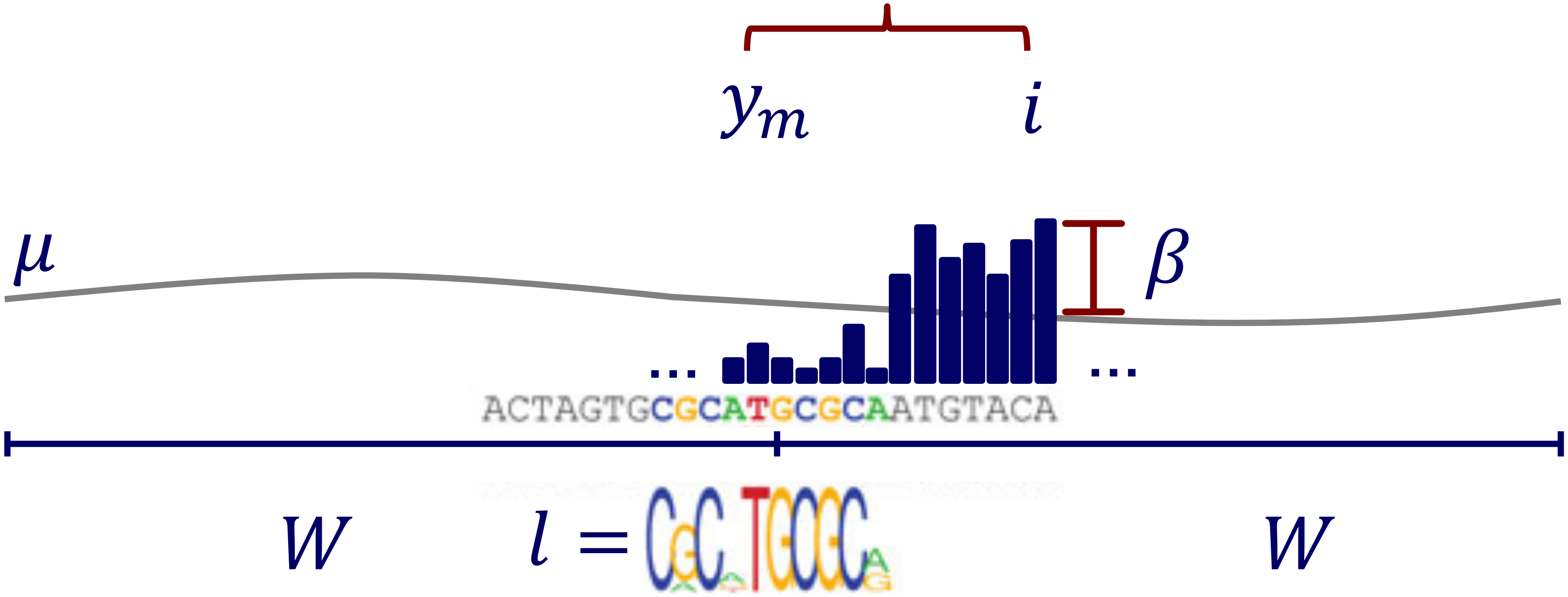
TF DNase profile

- DNase profiles represented as a vector for each TF

DNase profile for factor *l*

$$\widehat{\mu_{i,l}} = \mu_i + \begin{cases} \beta_{i-y_m,l} & |i-y_m| \le W\ and\ I_m = 1 \\ 0 & otherwise \end{cases}$$

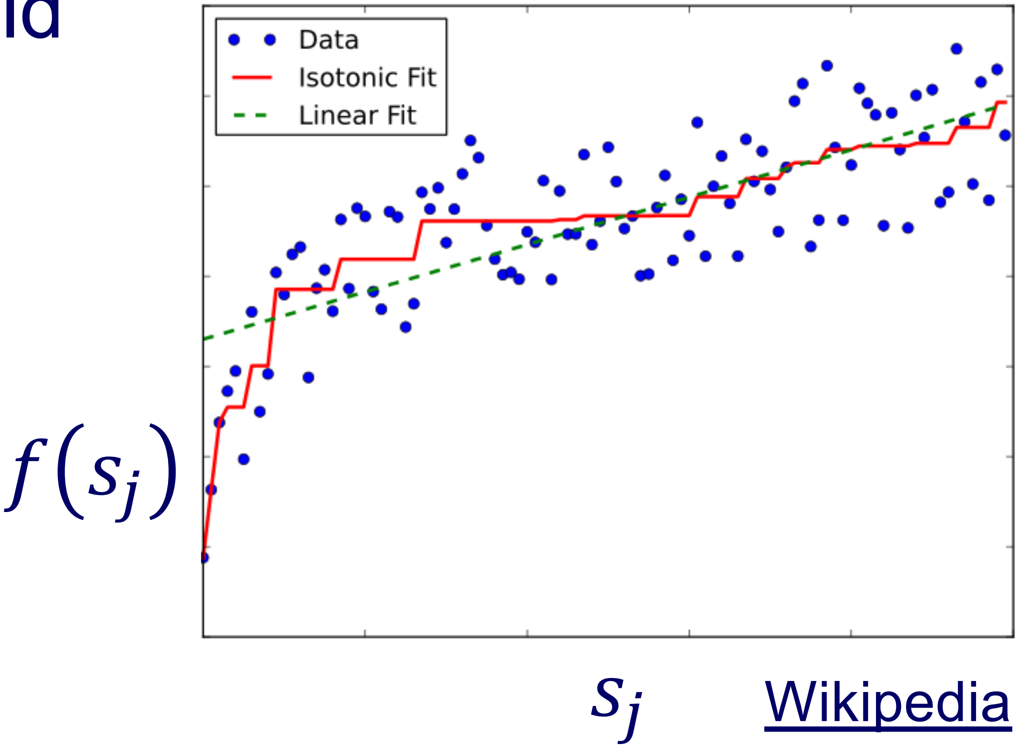
Can't be too far apart



Priors on TF binding

- TF binding event I_j should be more likely when
  - motif score s_j is high
  - DNase counts c_j are high (around matched motif)
- Isotonic (monotonic) regression

Example only, not realistic data



Wikipedia

$$\log(P(I_j = 1)) = f(s_j) + g(c_j)$$

Estimate Gaussian Process posterior

- Given background, read counts c_i and TF binding event I_j
  - Estimate Mean E[u_i| c_i] and variance Var[u_i| c_i]
- Non-binding sites by expectation propagation
- Binding sites by TF-specific effect model

Estimate binding sites

- Given posterior mean and variance E[*u*] and Var[*u*] per base
  - Estimate L_j=odd ratio(Prob(bound at j)/Prob(not bound at j)=$f_j + g_j + logit(p_j)$
  - p_j is determined by P(counts | binding or not, posterior *u*)
- Given L_j, s_j, c_j, and update priors f & g by least-square monotone regression

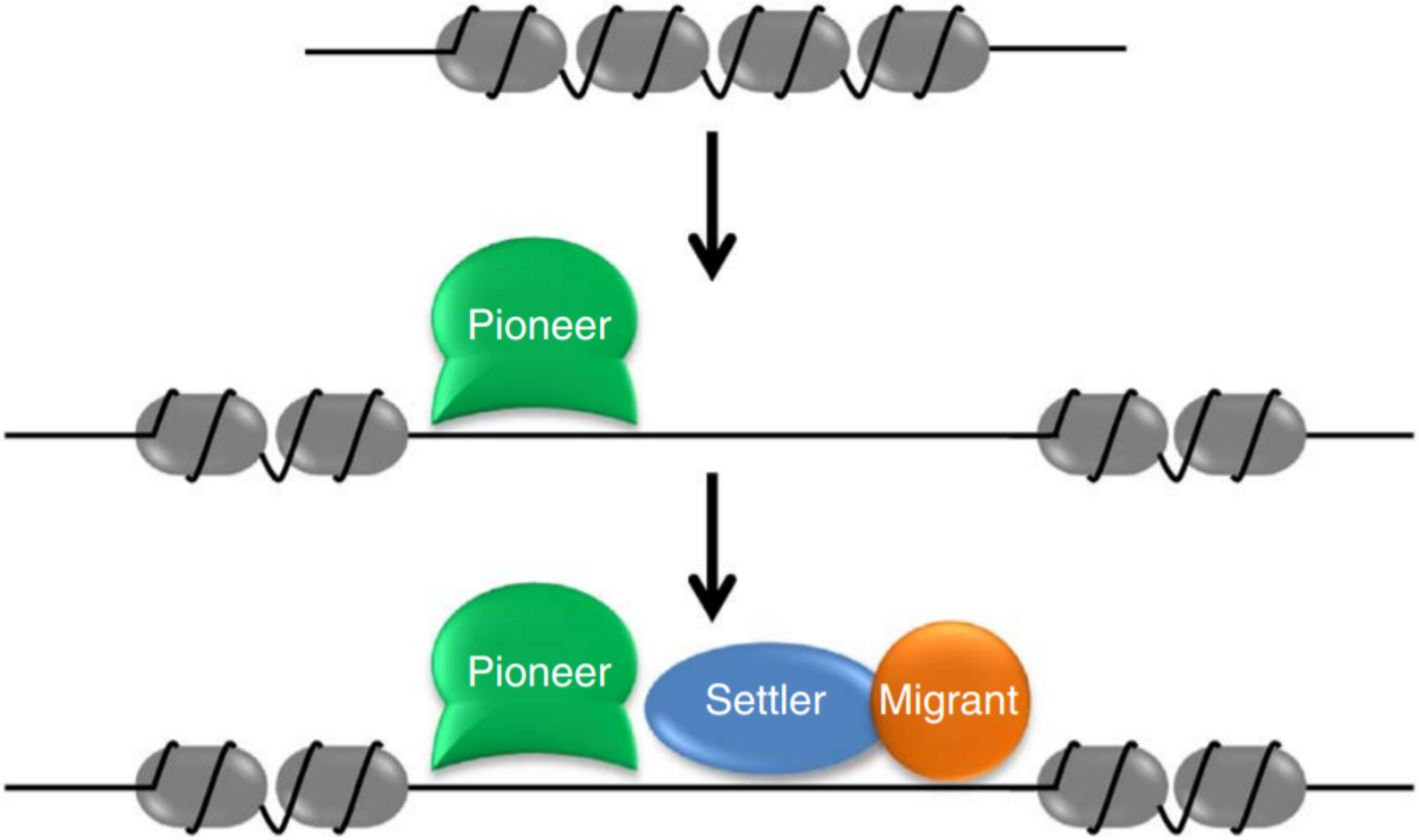
Full algorithm

- **Given**: TF motifs and DNase-Seq reads
- **Do**: Predict binding sites of each TF

- Identify candidate binding sites with PWMs
- Fit Gaussian process parameters for background
- Estimate TF binding effects $\beta_{i-j,l}$
  - using the top 10000 scoring motifs as bound sites
- Iterate until parameters converge
  - Estimate Gaussian process posterior with expectation propagation
  - Estimate expectation of which candidate binding sites are bound
  - Update monotonic regression functions for binding priors

TF binding hierarchy

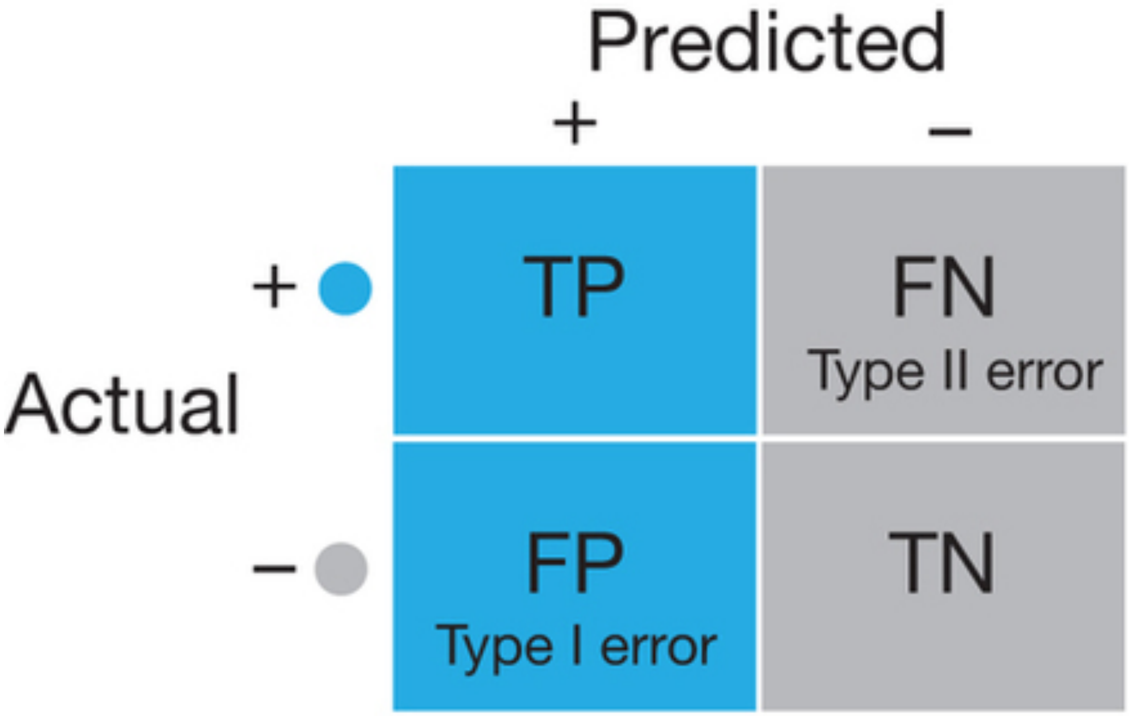
- Pioneer, settler, and migrant TFs



Sherwood *Nature Biotechnology* 2014

Evaluation: confusion matrix

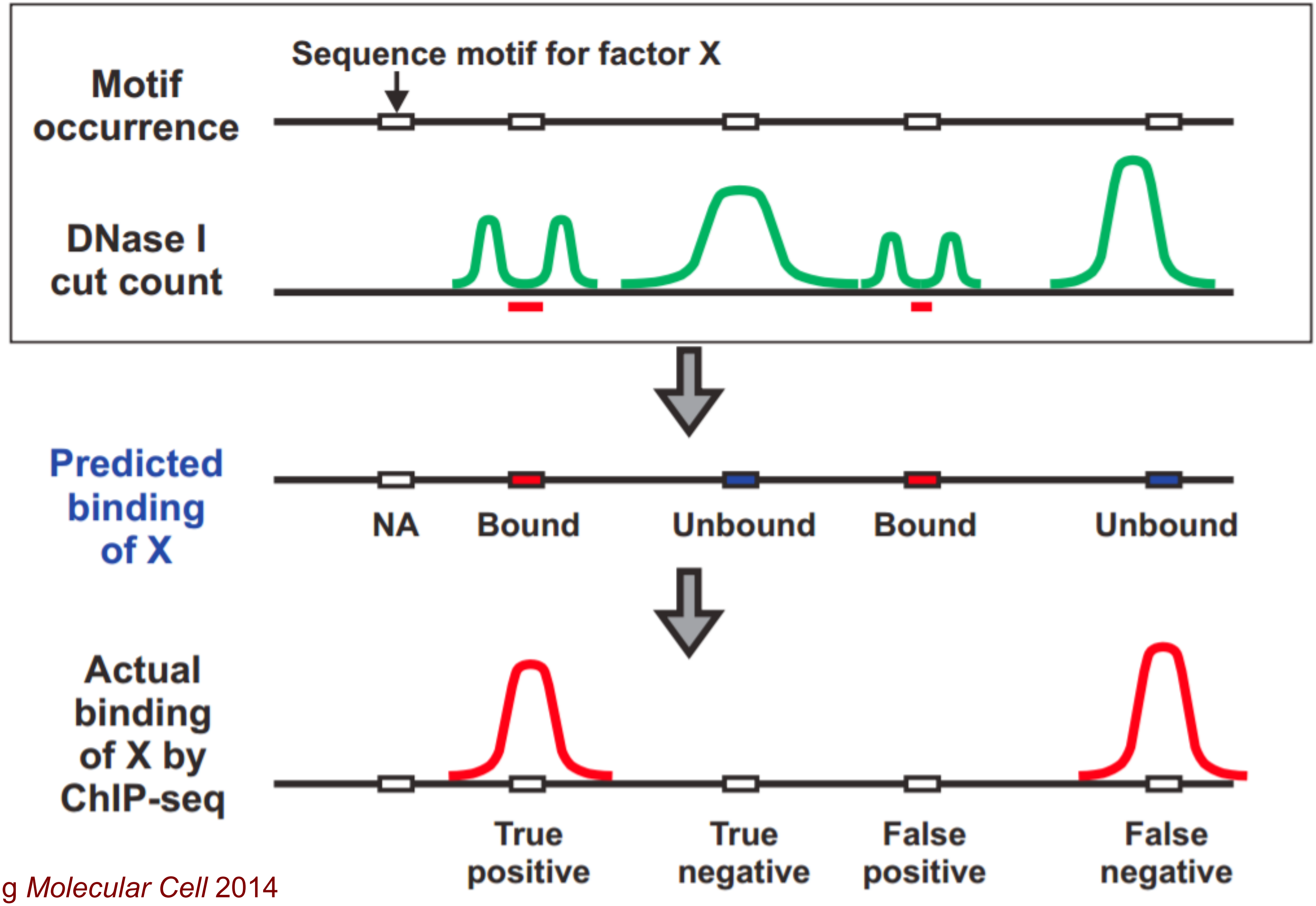
- Compare predictions to actual ground truth (gold standard)



		Predicted	
		+	–
Actual	+ ●	TP	FN Type II error
	– ●	FP Type I error	TN

Lever *Nature Methods* 2016

Evaluation: ChIP-Seq gold standard



Sung *Molecular Cell* 2014

Evaluation: ROC curve

- Calculate **r**eceiver **o**perating **c**haracteristic curve (ROC)
- True Positive Rate versus False Positive Rate
- Summarize with **a**rea **u**nder **ROC** curve (AUROC)

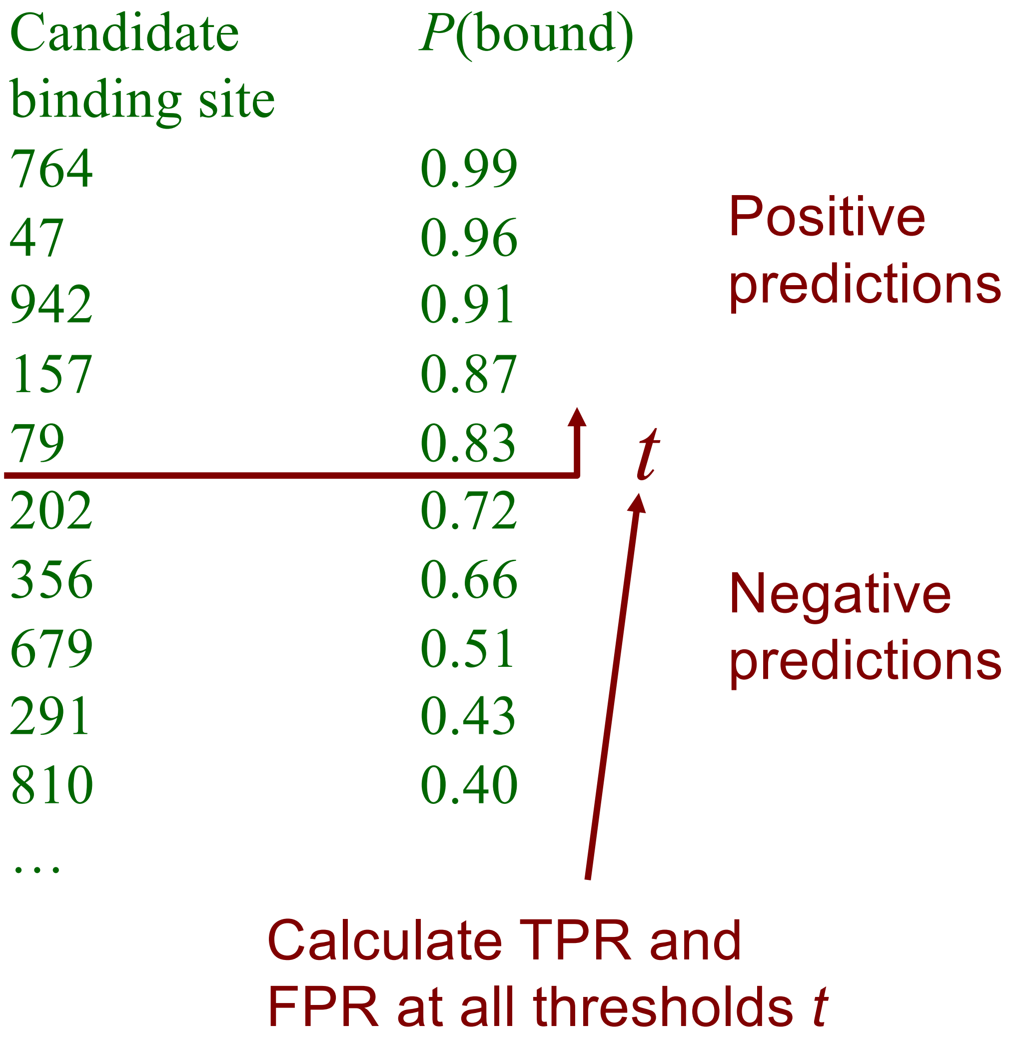
$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN}$$

Includes true negatives
Reason to prefer precision-recall for class imbalanced data

Evaluation: ROC curve

- TPR and FPR are defined for a **set** of positive predictions
- Need to threshold continuous predictions
- Rank predictions
- ROC curve assesses all thresholds



Candidate binding site	P(bound)
764	0.99
47	0.96
942	0.91
157	0.87
79	0.83
202	0.72
356	0.66
679	0.51
291	0.43
810	0.40
…	

Positive predictions

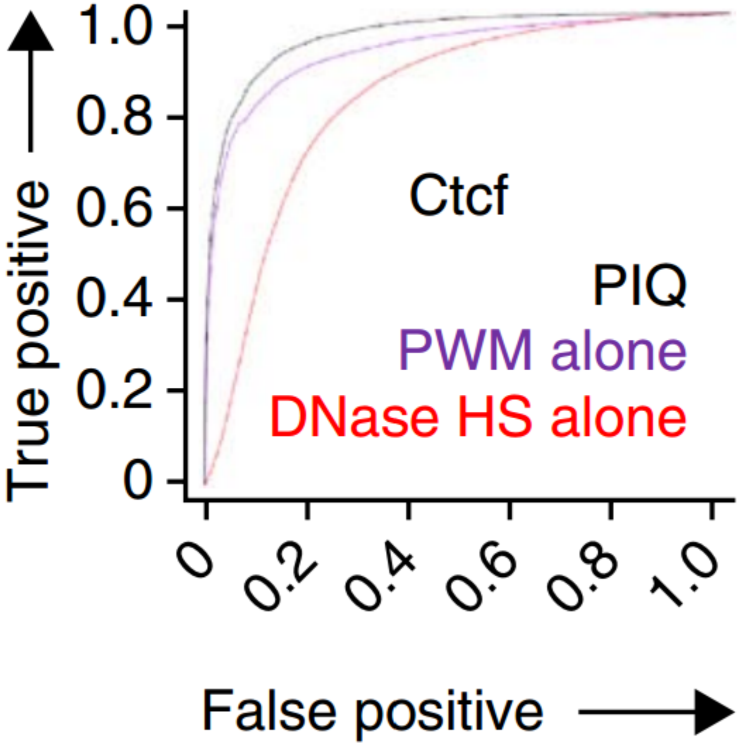
t

Negative predictions

Calculate TPR and FPR at all thresholds t

PIQ ROC curve for mouse Ctcf

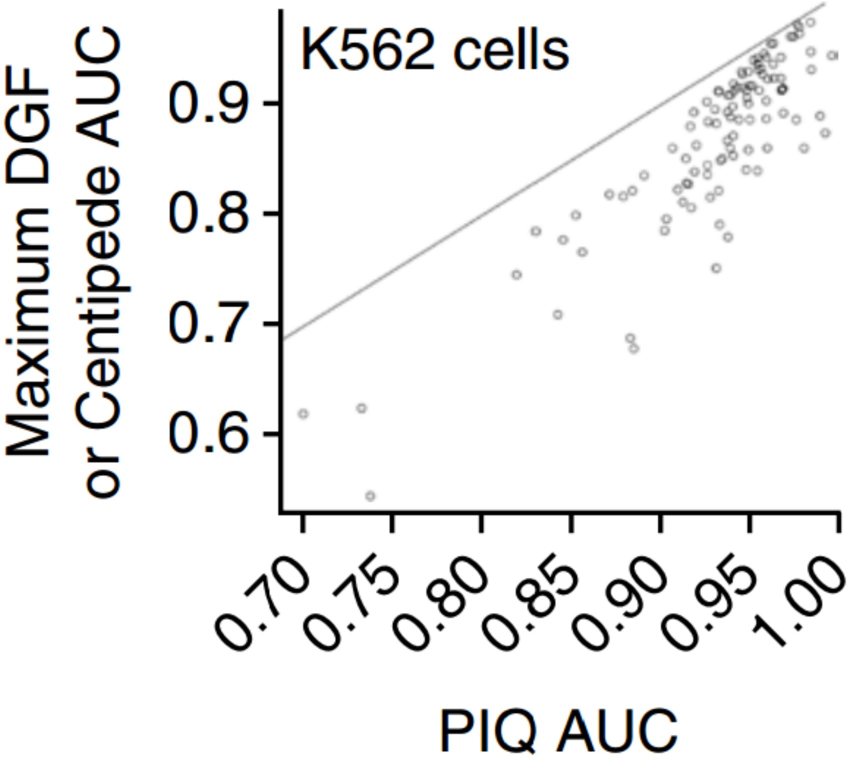
- Compare predictions to ChIP-Seq
- Full PIQ model improves upon motifs or DNase alone



Sherwood *Nature Biotechnology* 2014

PIQ evaluation

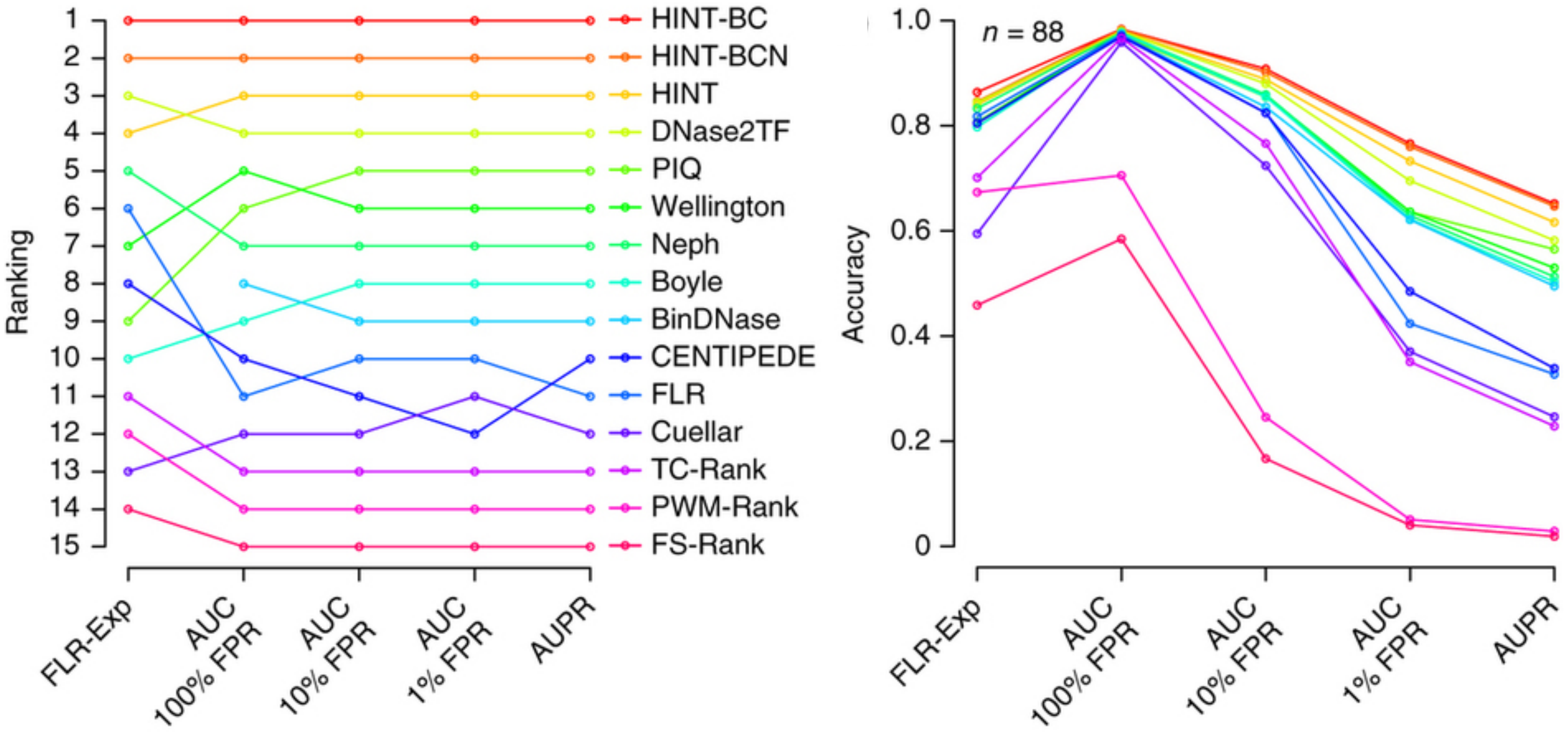
- Compare to two standard methods
  - 303 ChIP-Seq experiments in K562 cells
  - Centipede, digital genomic footprinting
- Compare AUROC
  - PIQ has very high AUROC
  - Mean 0.93
  - Corresponds to recovering median of 50% of binding sites



Sherwood *Nature Biotechnology* 2014

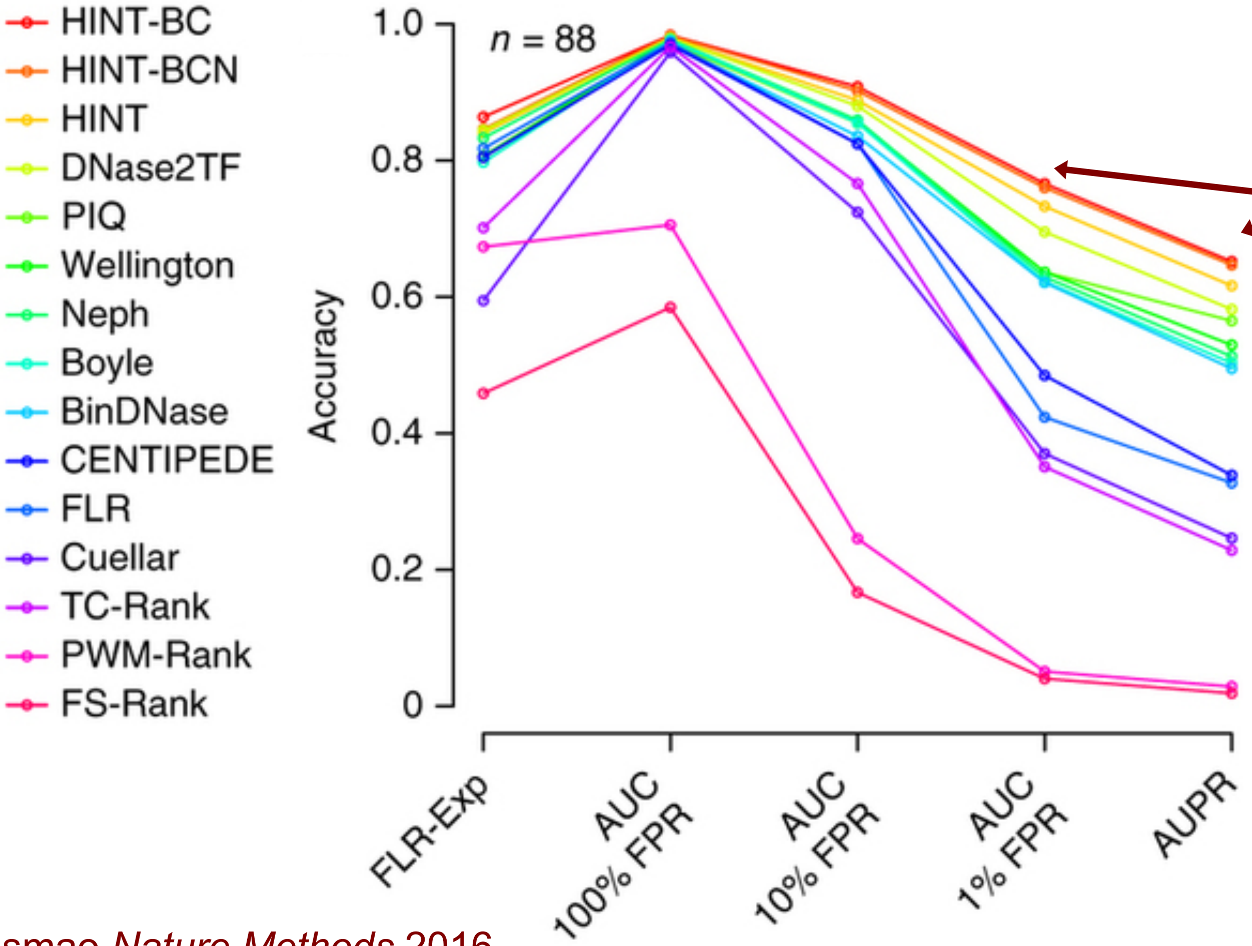
DNase-Seq benchmarking

- PIQ among top methods in large scale DNase benchmarking study
- HMM-based model HINT was top performer



Gusmao *Nature Methods* 2016

Downside of AUROC for genome-wide evaluations



Almost all methods look equally good when using full ROC curve AUROC close to 1.0

Precision-recall curve or truncated ROC curve differentiate methods

Gusmao *Nature Methods* 2016

PIQ summary

- Smooth noisy DNase-Seq data without imposing too much structure
- Combine DNase-Seq and motifs to predict condition-specific binding sites
- Supports replicates and multiple related conditions (e.g. time series)