

Epigenetics - Predicting TF binding with DNase-Seq and PIQ

BMI/CS 776

www.biostat.wisc.edu/bmi776/

Spring 2022

Daifeng Wang

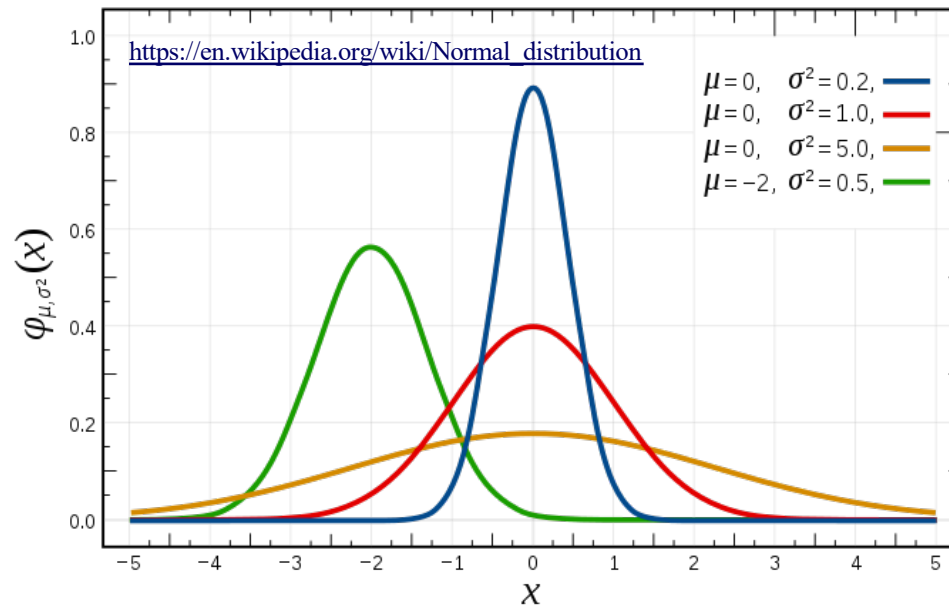
daifeng.wang@wisc.edu

Gaussian distribution

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

pdf(x) =

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



- 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

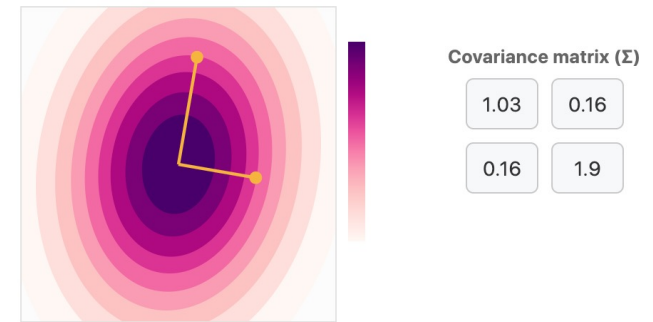
- Multiple random variables

- $\vec{x} = [x_1 \ x_2 \ \dots \ x_n]^T \sim \mathcal{N}(\vec{\mu}, \Sigma)$

- $\text{pdf}(\vec{x}) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(\vec{x} - \vec{\mu})^T \Sigma^{-1} (\vec{x} - \vec{\mu})}$

- Covariance matrix

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



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

- # of reads at Position i and Position j

- $[x_i \ x_j] \sim \mathcal{N}([\mu_i \ \mu_j], \begin{bmatrix} \sigma_i^2 & E[(x_i - \mu_i)(x_j - \mu_j)] \\ E[(x_i - \mu_i)(x_j - \mu_j)] & \sigma_j^2 \end{bmatrix})$

Kernel function for covariance

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

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

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

Gaussian process (GP)

- A stochastic process with mean function $\mu(\cdot)$ and covariance function $k(\cdot, \cdot)$ 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 ($n \times n$) and $\boldsymbol{K}_{i,j} = k(i, j)$
- $x_{(\cdot)} \sim \mathcal{GP}(\mu(\cdot), k(\cdot, \cdot))$
 - Infinite number of random variables, $x_1 \ x_2 \ \dots$

Gaussian process regression

- $f(i)$ is a regression function to predict # of reads x_i on position i
 - $x_i = f(i) + \varepsilon_i$, where ε_i is noise $\sim \mathcal{N}(0, \sigma^2)$
- $\mathcal{GP}(0, k(. , .))$ as prior for regression function to predict a distribution of x
 - Observed data $S = \{p, x_p\}$, $p \in \{1, 2, \dots\}$
 - New data $T = \{q, x_q\}$
 - To predict posterior $P(x_q | x_p, p, q) \sim \mathcal{N}(\vec{\mu}^*, \Sigma^*)$

Gaussian process regression

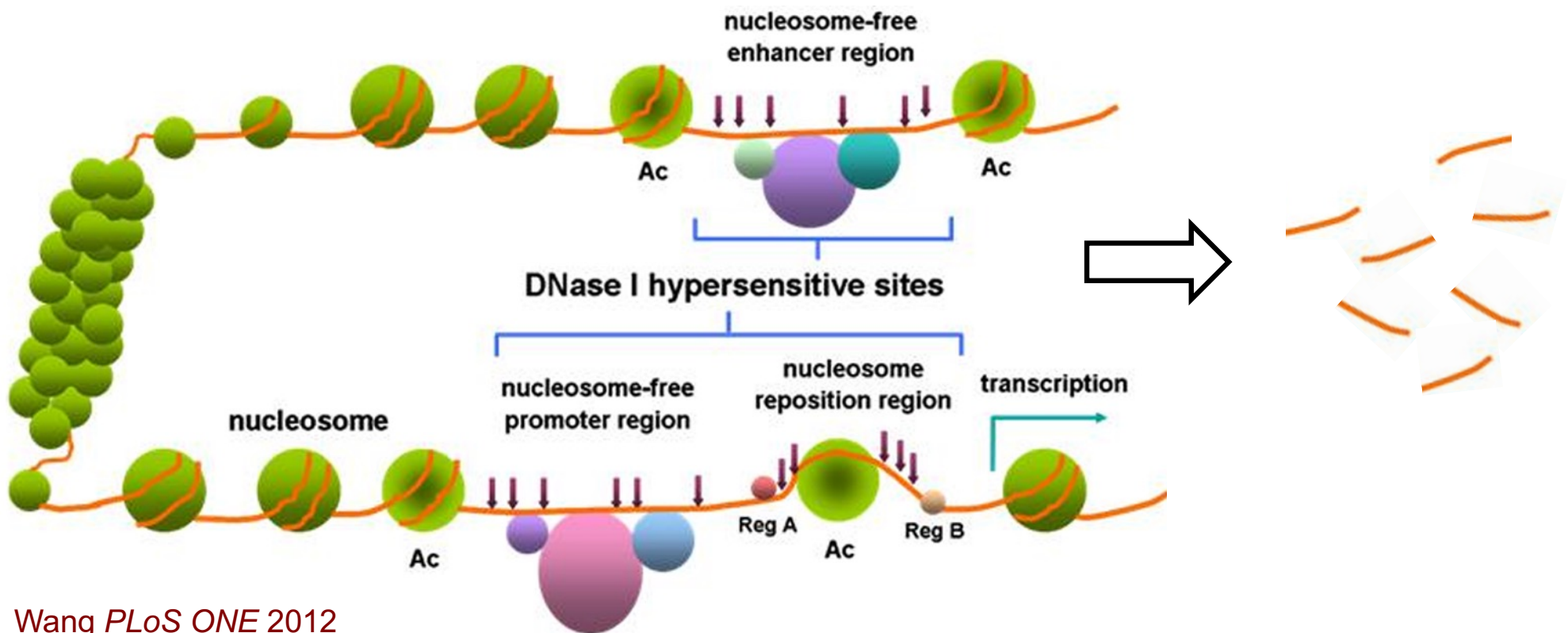
- $\mathcal{GP}(0, k(. , .))$ as prior for regression function to predict a distribution of x
 - Joint probability $P(x_p, x_q | p, q)$ by \mathcal{GP}
 - $P(x_q | x_p, p, q) \sim P(x_p, x_q | p, q) / P(x_p | p)$ by Bayes rule
 - $\vec{\mu}^* = \mathbf{K}(\vec{p}, \vec{q})(\mathbf{K}(\vec{p}, \vec{p}) + \sigma^2 \mathbf{I})^{-1} \vec{x}_p$
 $\Sigma^* = \mathbf{K}(\vec{q}, \vec{q}) + \sigma^2 \mathbf{I} - \mathbf{K}(\vec{p}, \vec{q})(\mathbf{K}(\vec{p}, \vec{p}) + \sigma^2 \mathbf{I})^{-1} \mathbf{K}(\vec{p}, \vec{q})$

Gaussian processes

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

DNase I hypersensitive sites

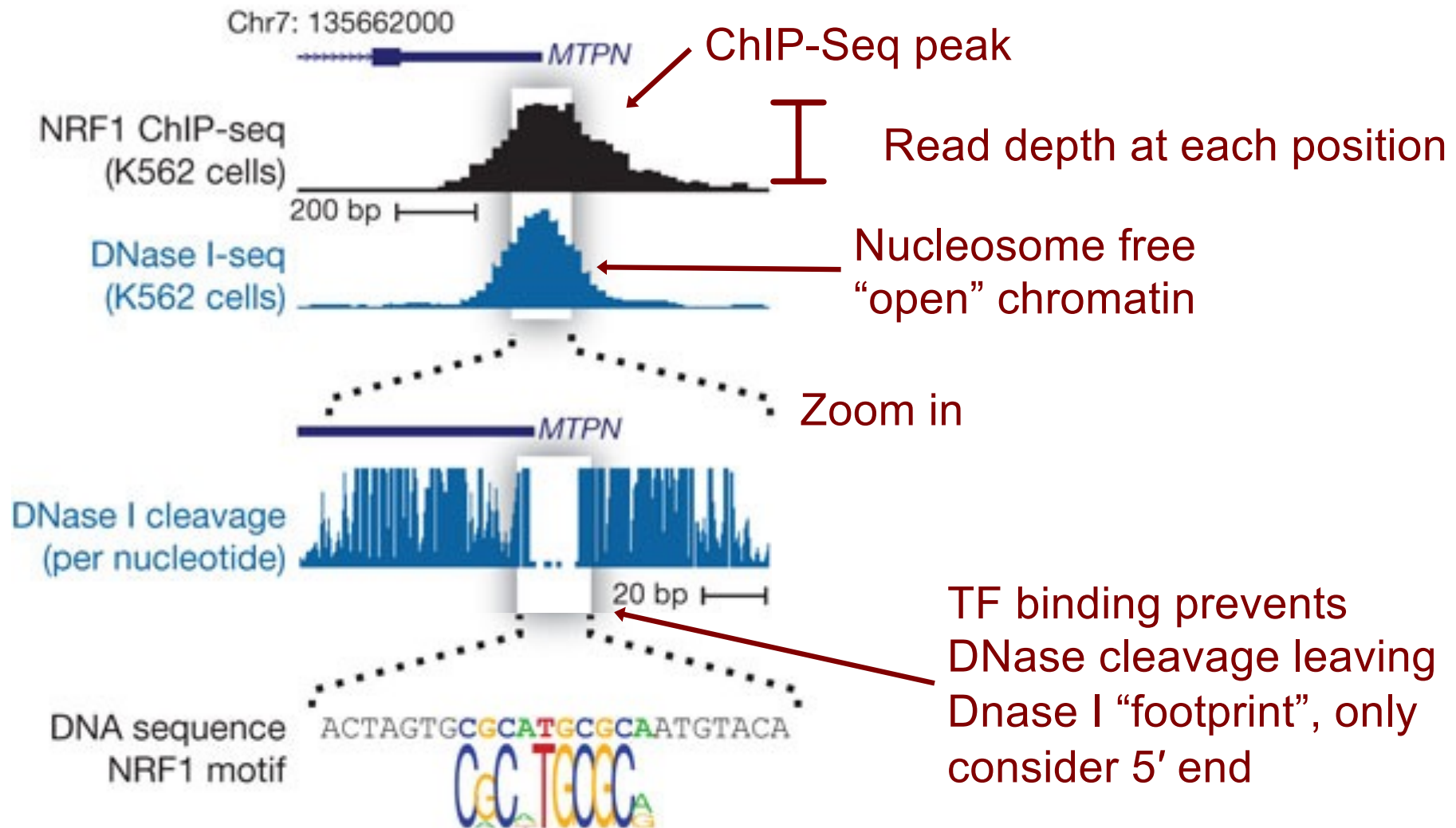
- Arrows indicate DNase I cleavage sites
- Obtain short reads that we map to the genome



Wang *PLoS ONE* 2012

DNase I footprints

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

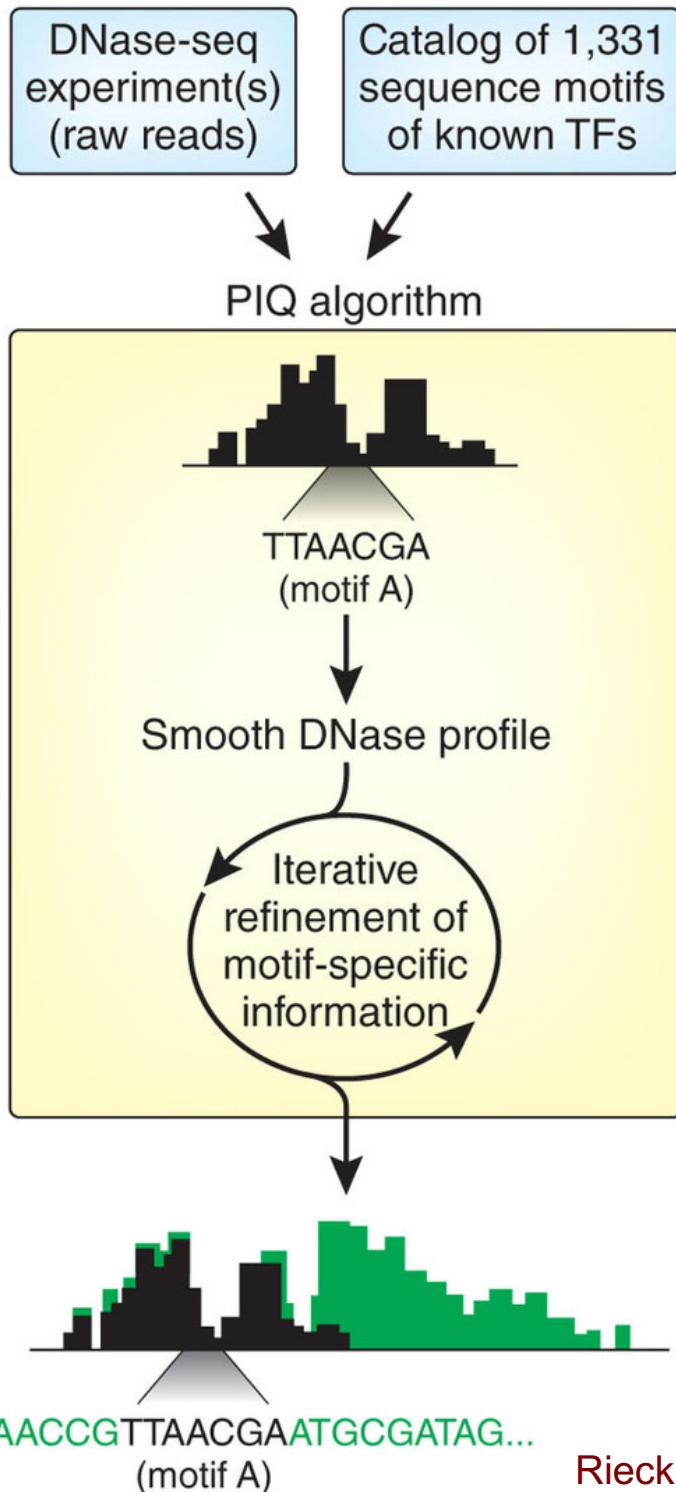


DNase I footprints to TF binding predictions

- 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

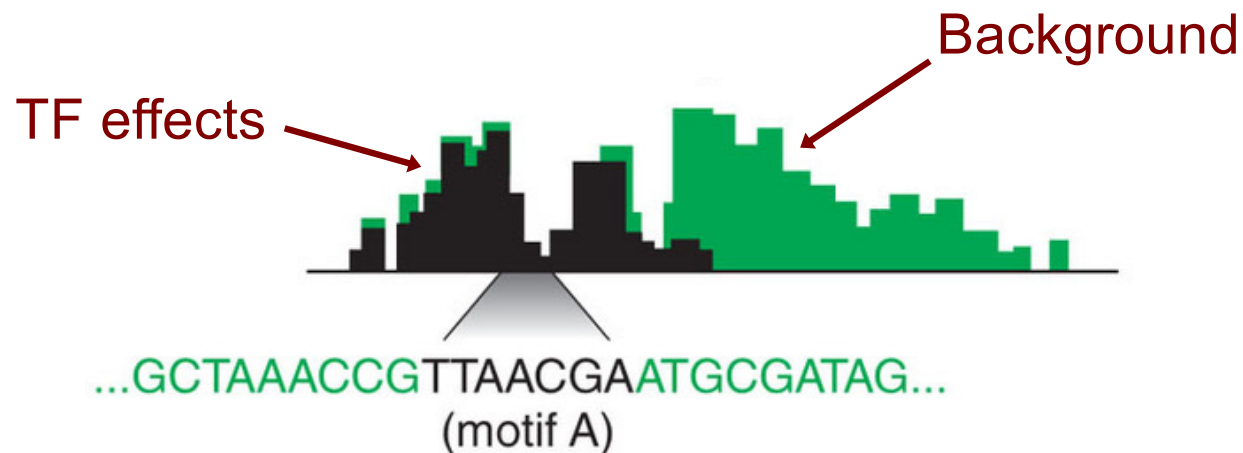
Protein Interaction Quantification (PIQ)



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

PIQ main idea

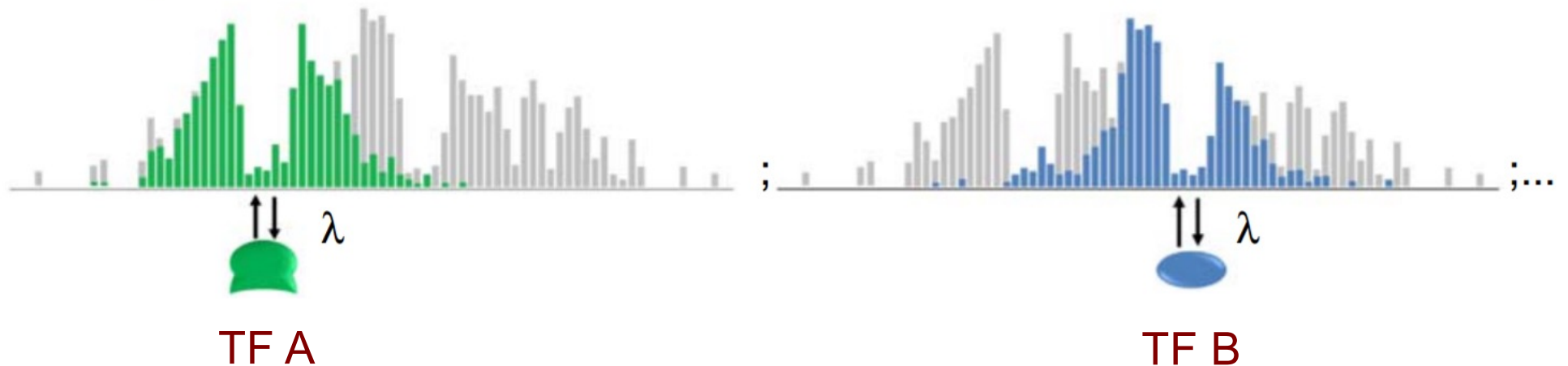
- With no TF binding, DNase-Seq reads come from some background distribution
- TF binding changes read density in a *TF-specific* way



PIQ main idea

- Shape of DNase peak and footprint depend on the TF

TF binding estimation



Sherwood *Nature Biotechnology* 2014

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 $\mathcal{N}(\vec{\mu}_0, \Sigma)$
 - Positions i and j : u_i and u_j , $\Sigma_{i,j} = \sigma_0 k(|i - j|)$
 - e.g., k is correlation
- # of reads (read counts) c_i at Position i
 - $c_i \sim \text{Poisson}(\exp(u_i))$
- Estimate a background GP($\mu_0, \sigma_0, k, \Sigma^{-1}$)
 - Supplement C.5

TF-specific DNase profile

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

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

Diagram illustrating the TF-specific DNase profile adjustment:

- $\widehat{\mu}_{i,l}$: DNase log-read rate adjusted for binding of factor l
- μ_i : DNase log-read rate at position i from Gaussian process
- $\beta_{i-y_m,l}$: DNase profile for factor l
- $|i - y_m| \leq W$: Window size
- $I_m = 1$: Whether site m is bound
- y_m : Midpoint location of binding site m

TF DNase profile

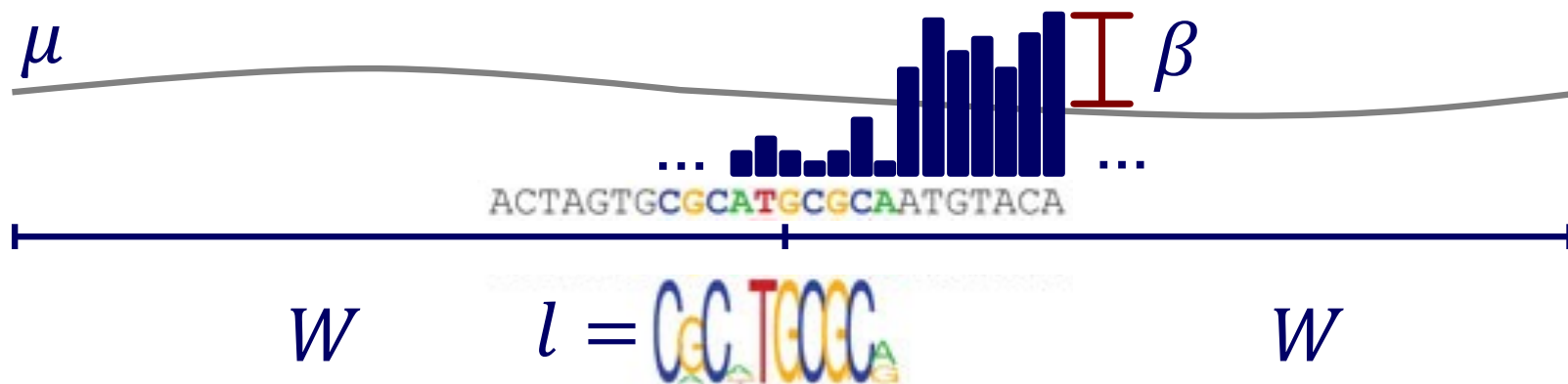
- DNase profiles represented as a vector for each TF

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

DNase profile for factor l

Can't be too far apart

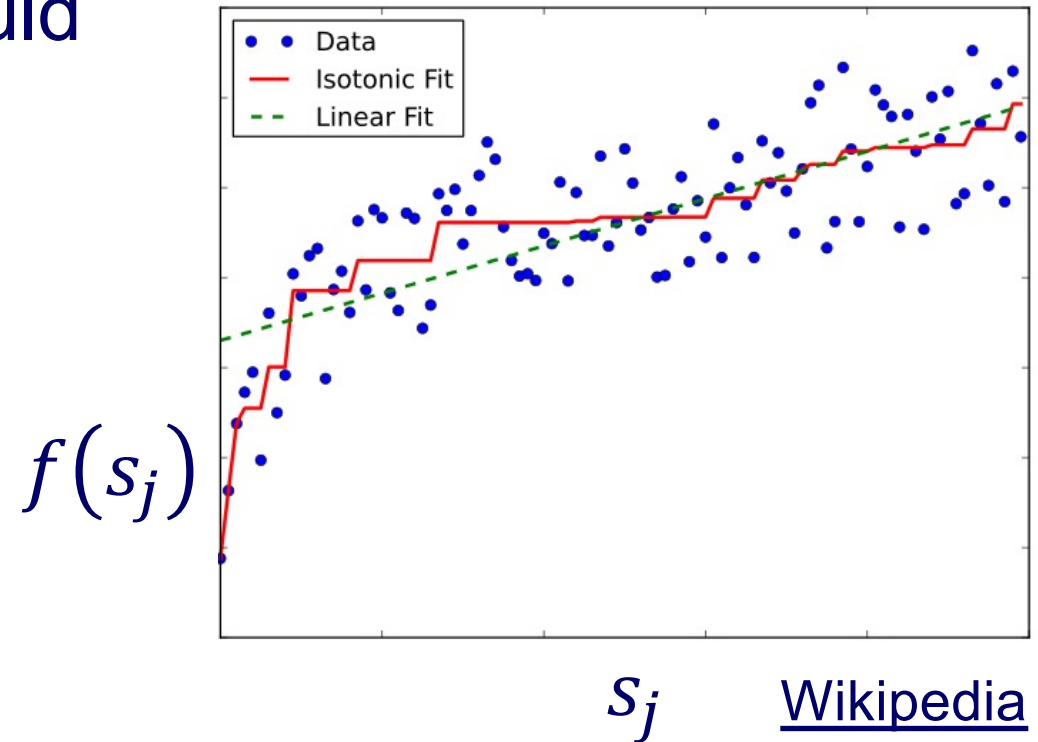
y_m i



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



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

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 $\text{Var}[u_i | c_i]$
- Non-binding sites by expectation propagation
- Binding sites by TF-specific effect model

Estimate binding sites

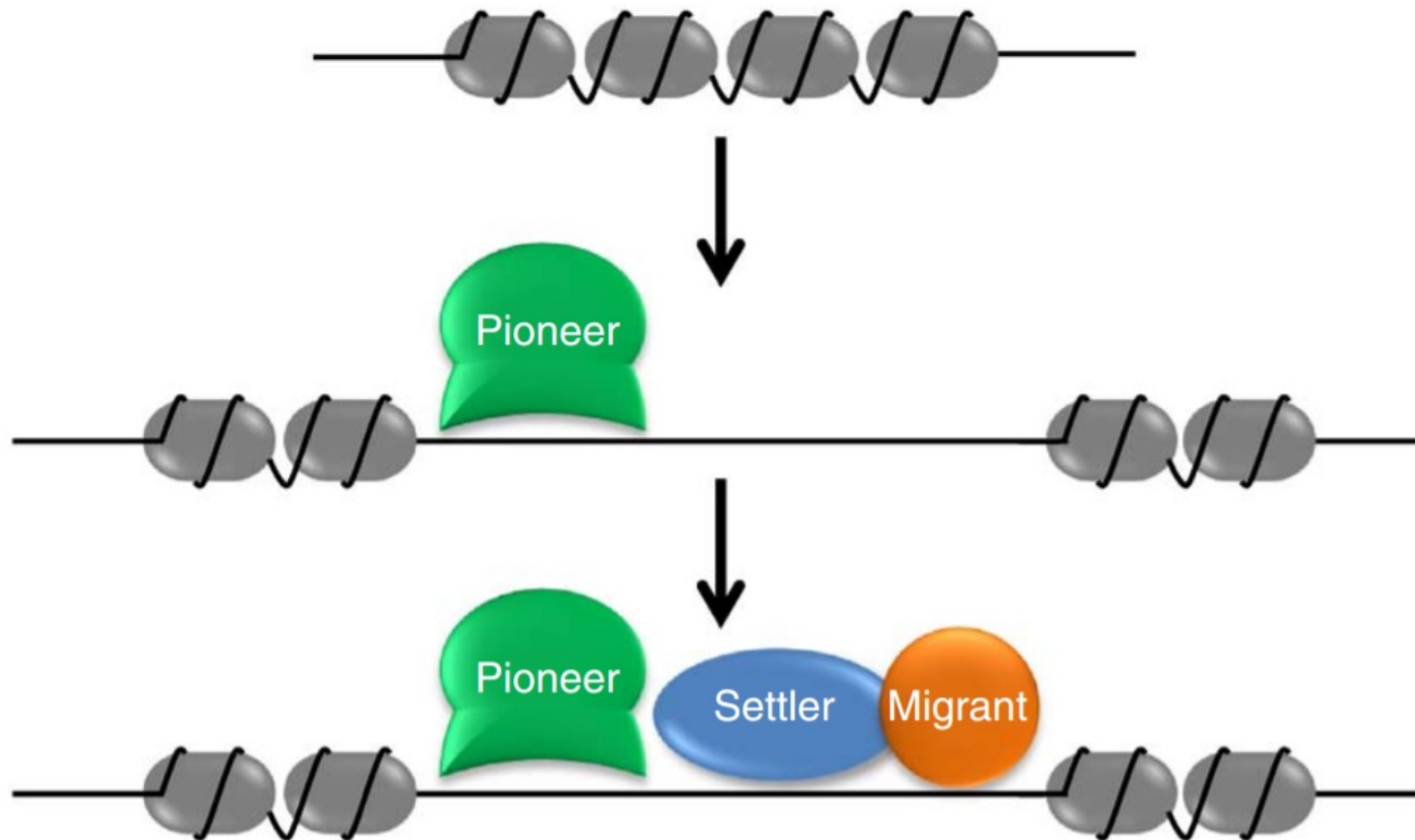
- Given posterior mean and variance $E[u]$ and $\text{Var}[u]$ per base
 - Estimate $L_j = \text{odds ratio}(\text{Prob}(\text{bound at } j) / \text{Prob}(\text{not bound at } j)) = f_j + g_j + \text{logit}(p_j)$
 - p_j is determined by $P(\text{counts} \mid \text{binding or not, posterior } u)$
- Given 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 (via probability)
- Identify candidate binding sites with PWMs
- Fit Gaussian process parameters for background
- Calculate TF binding effects $\beta_{i-j,l}$
 - using the top 10000 scoring motifs as bound sites
- Iterate until parameters converge
 - Estimate Gaussian process posterior (Slide #22)
 - Estimate expectation of which candidate binding sites are bound (Slide #23)
 - Update monotonic regression functions for binding priors (Slide #21)
- Supplement Page 8

TF binding hierarchy

- Pioneer, settler, and migrant TFs



Sherwood Nature Biotechnology 2014

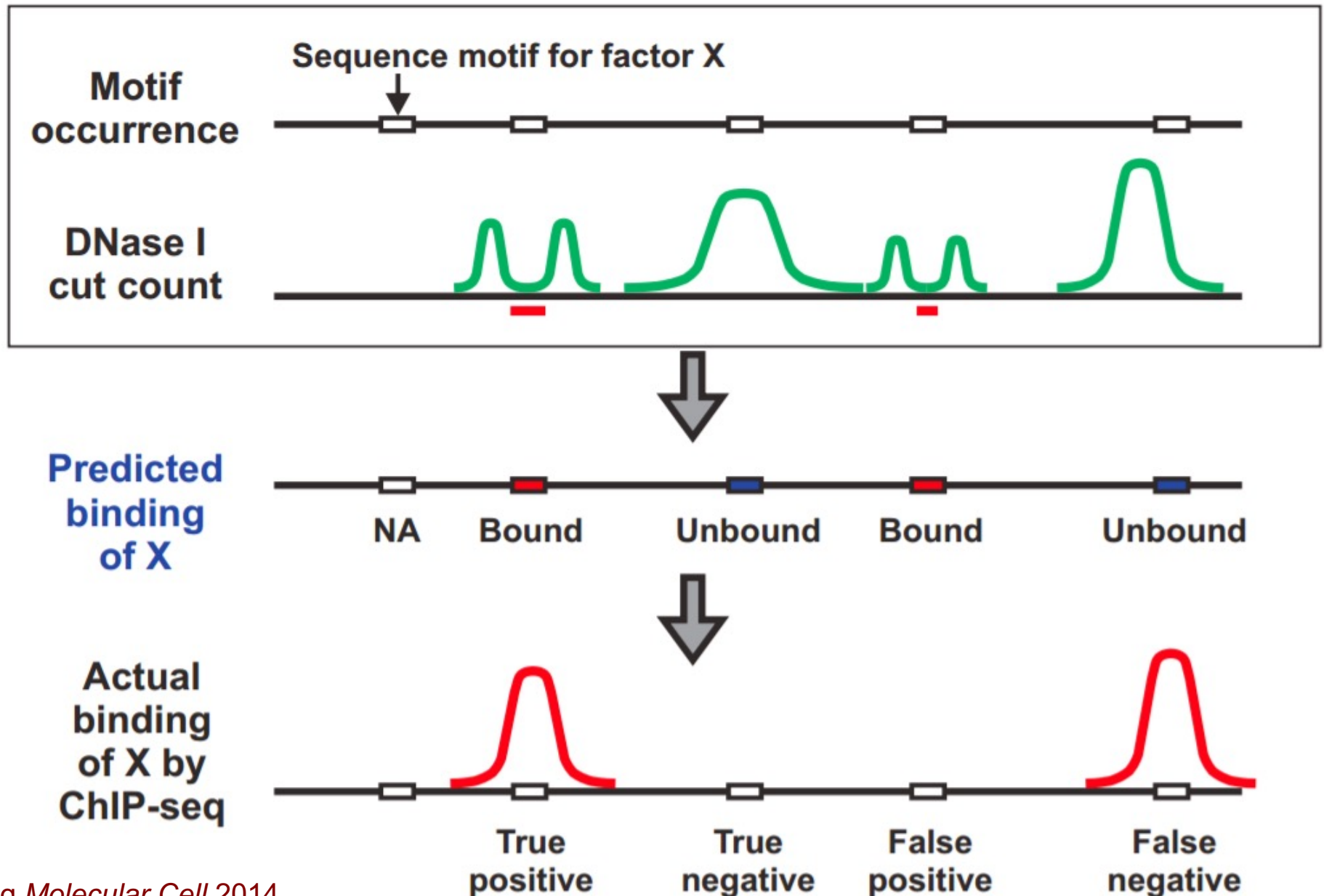
Evaluation: confusion matrix

- 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



Evaluation: ROC curve

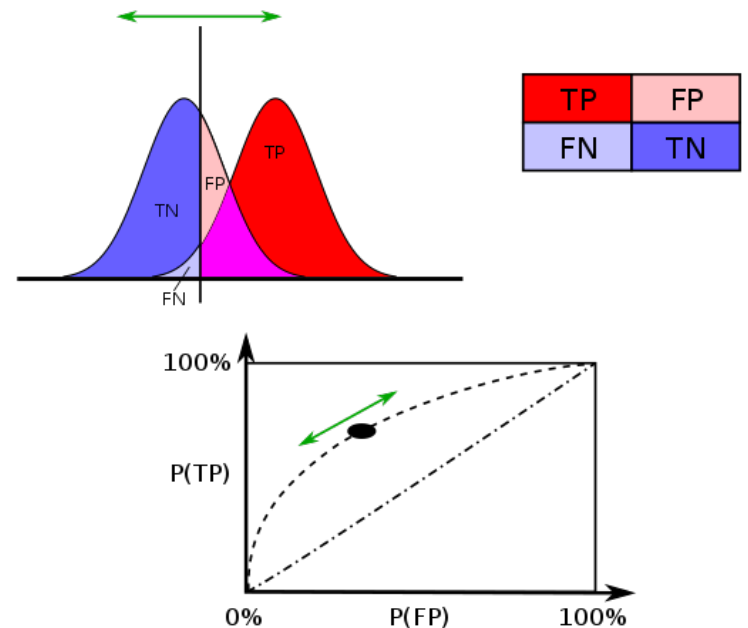
- Calculate receiver operating characteristic curve (ROC)
- True Positive Rate(TPR) versus False Positive Rate (FPR)
- Summarize with **area under ROC curve (AUROC)**

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

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

Includes true negatives

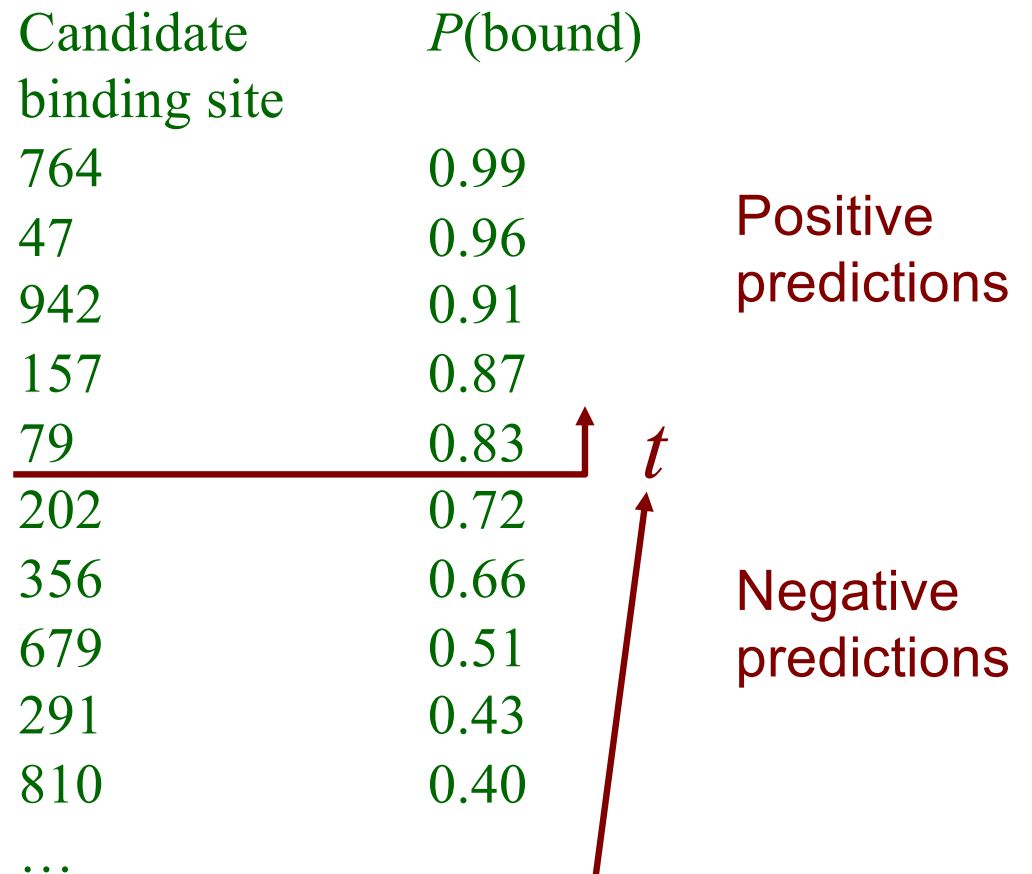
Reason to prefer precision-recall for class imbalanced data



https://en.wikipedia.org/wiki/Receiver_operating_characteristic#/media/File:ROC_curves.svg

Evaluation: ROC curve

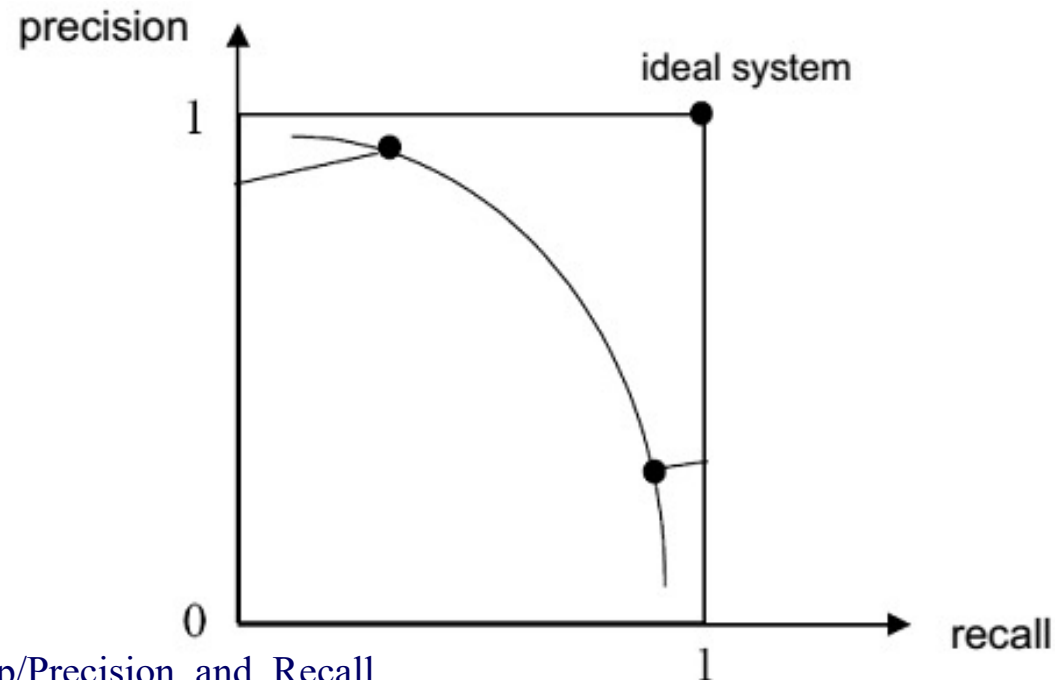
- TPR and FPR are defined for a **set** of positive predictions
- Need to threshold continuous predictions
- Rank predictions
- ROC curve assesses all thresholds



Calculate TPR and FPR at all thresholds t

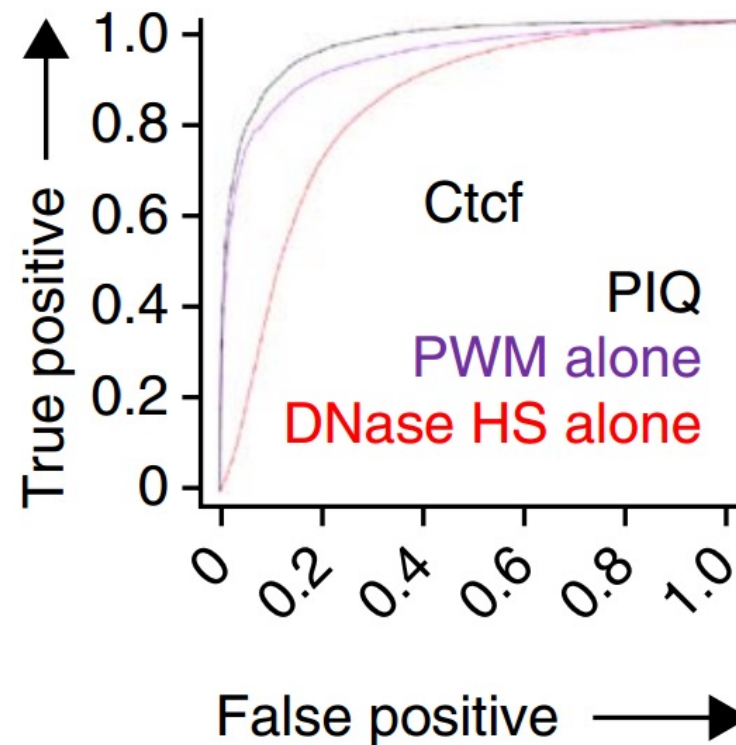
Precision-Recall Curve

- Precision = $TP/(TP+FP)$
- Recall = $TP/(TP+FN) = TPR$
- <https://www.datascienceblog.net/post/machine-learning/interpreting-roc-curves-auc/>



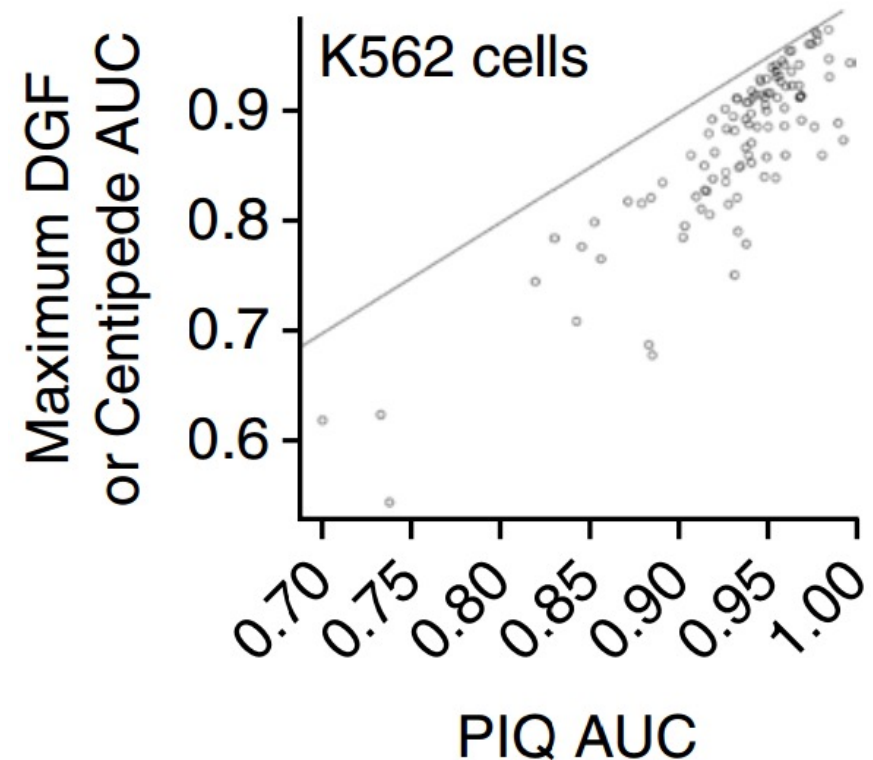
PIQ ROC curve for mouse Ctcf

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



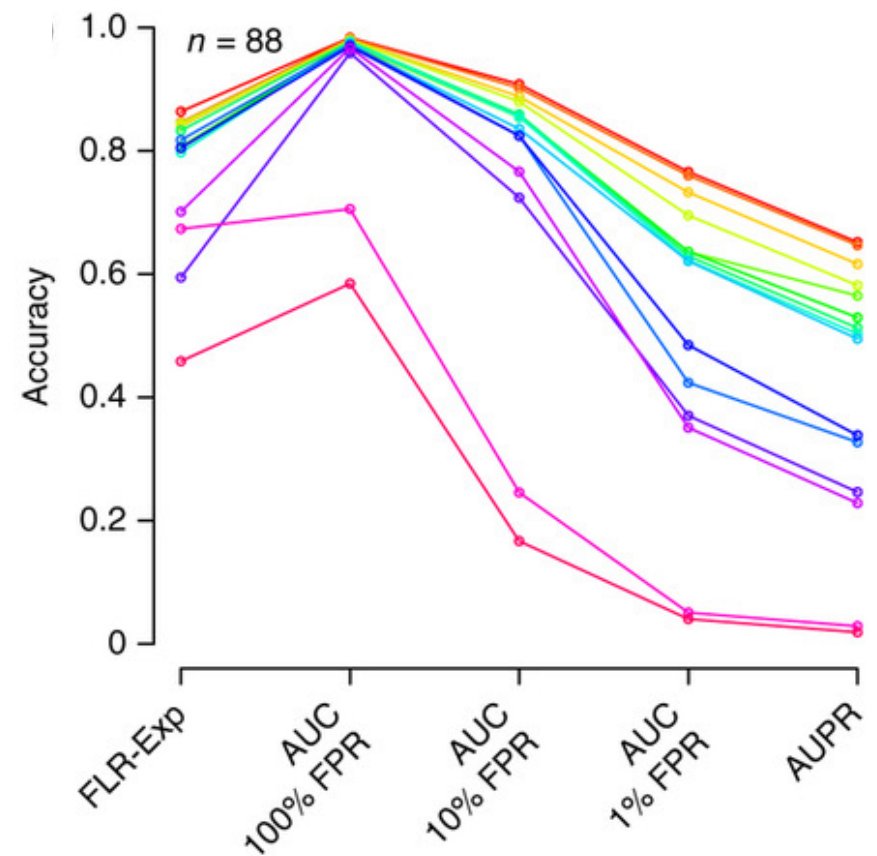
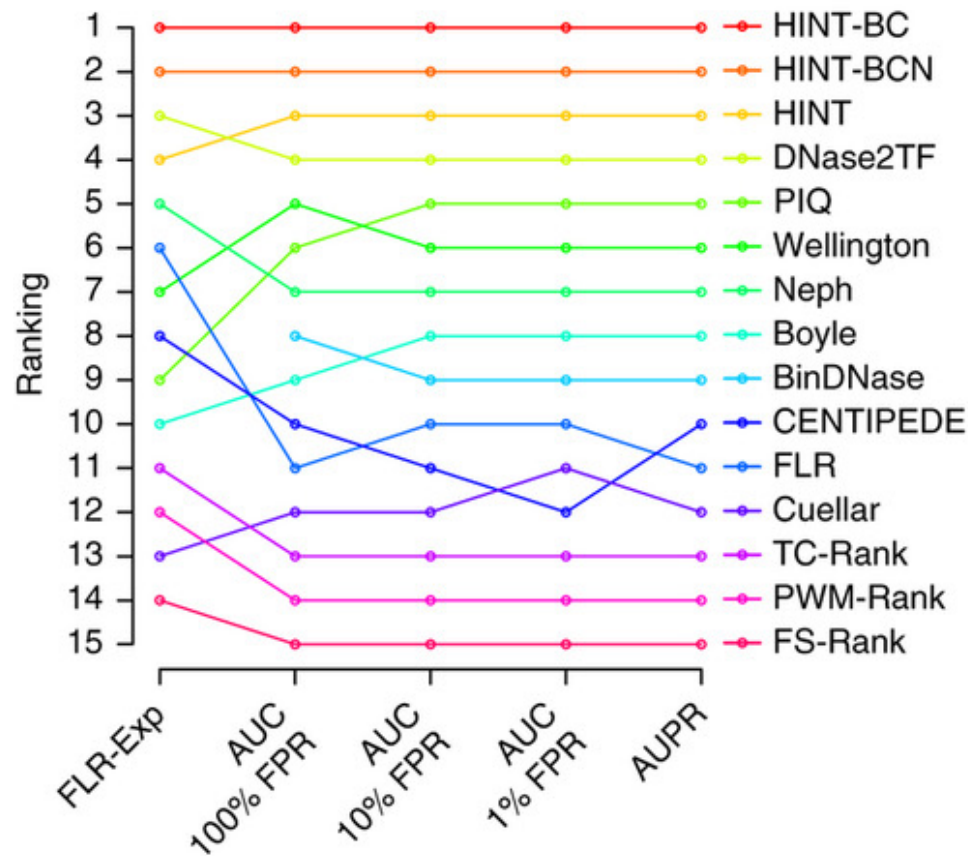
PIQ evaluation

- 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

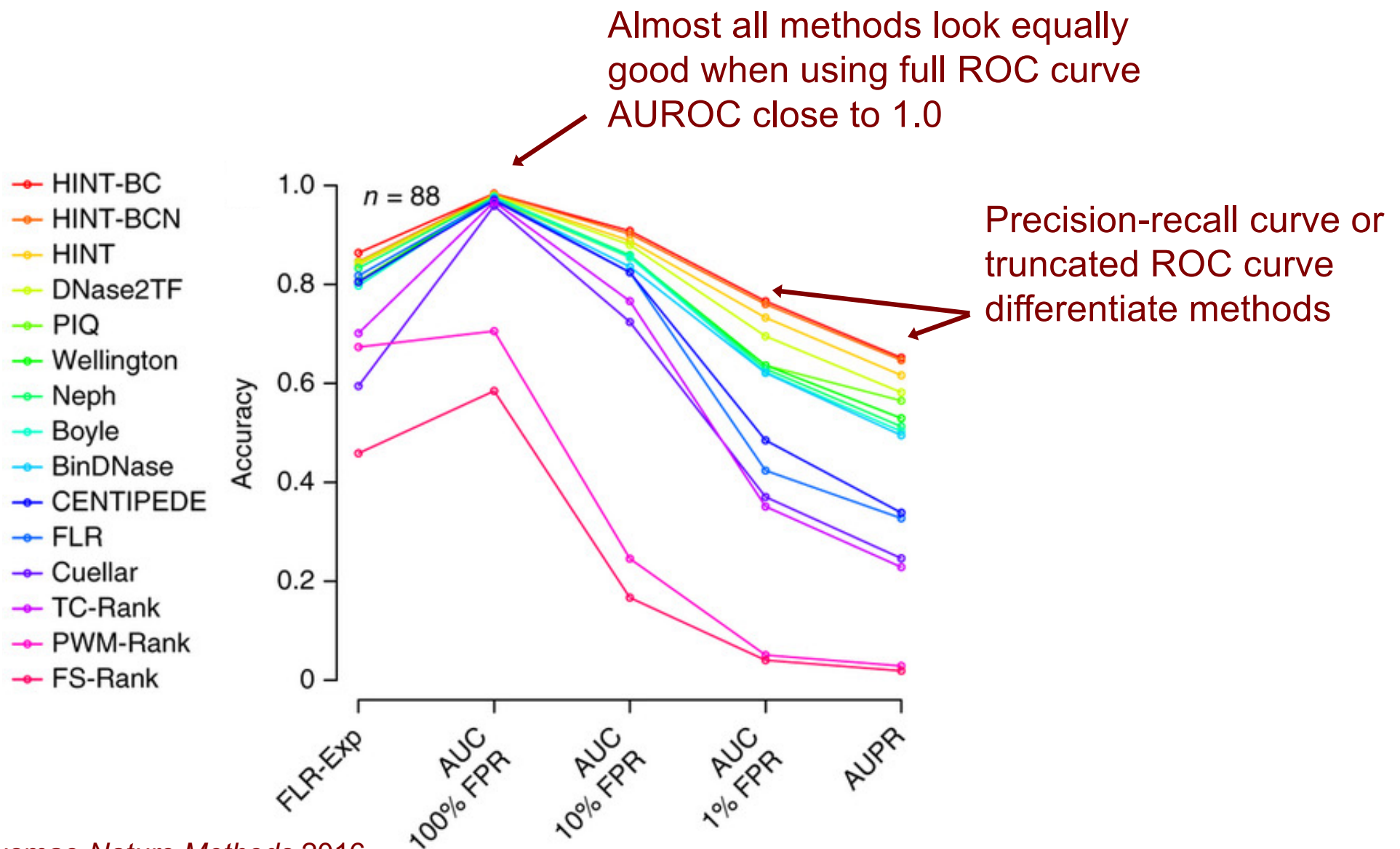


DNase-Seq benchmarking

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



Downside of AUROC for genome-wide evaluations



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)