Epigenetics - Predicting TF binding with DNase-Seq and PIQ

BMI/CS 776

www.biostat.wisc.edu/bmi776/

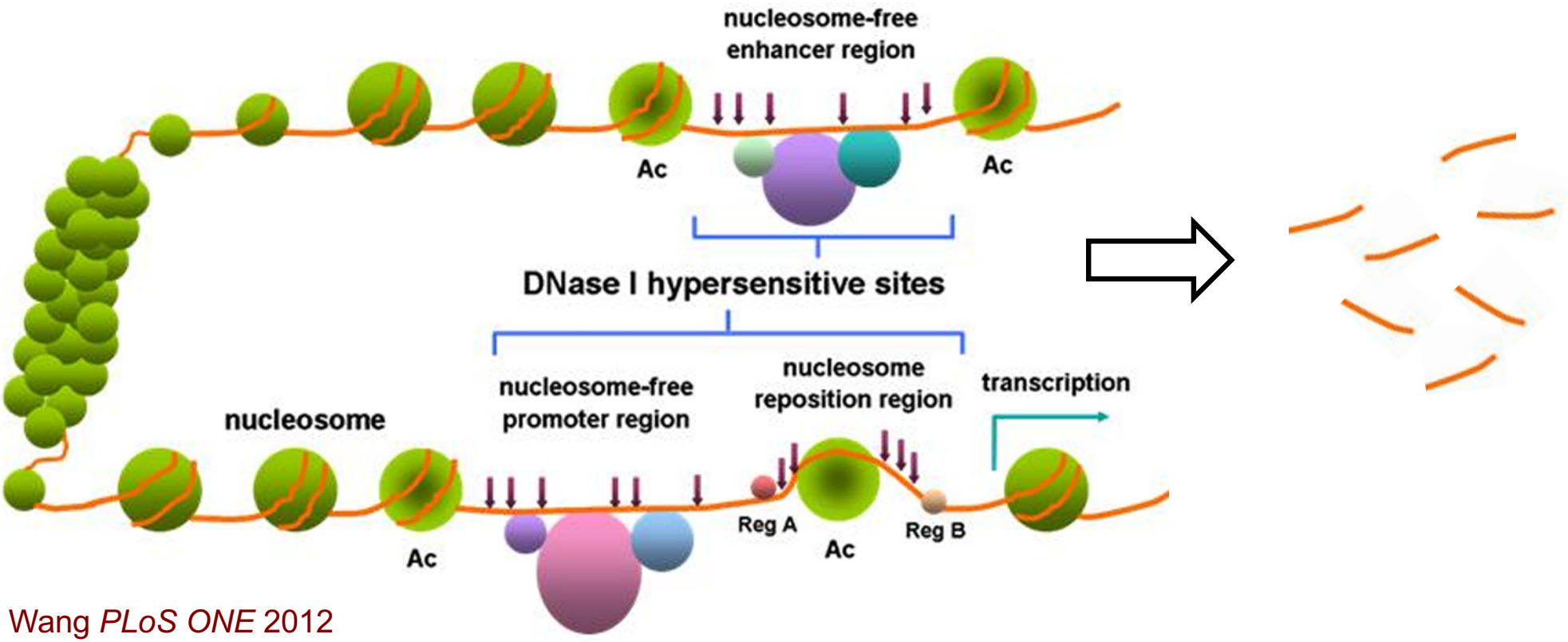
Spring 2019

Colin Dewey

colin.dewey@wisc.edu

DNase I hypersensitive sites

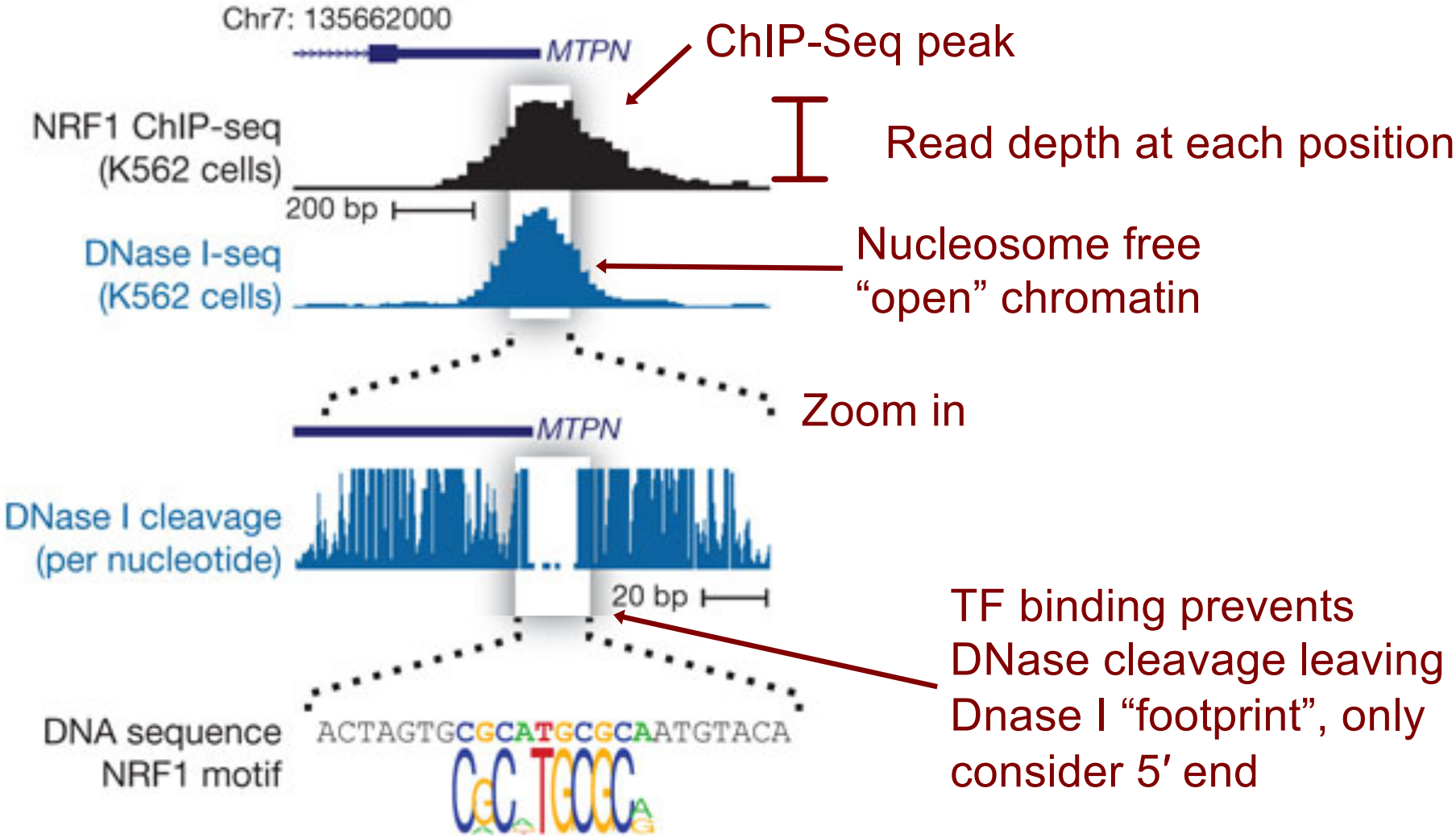
- Arrows indicate DNase I cleavage sites
- Obtain short reads that we map to the genome



Wang *PLoS ONE* 2012

DNase I footprints

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



Neph *Nature* 2012

DNase I footprints to TF binding predictions

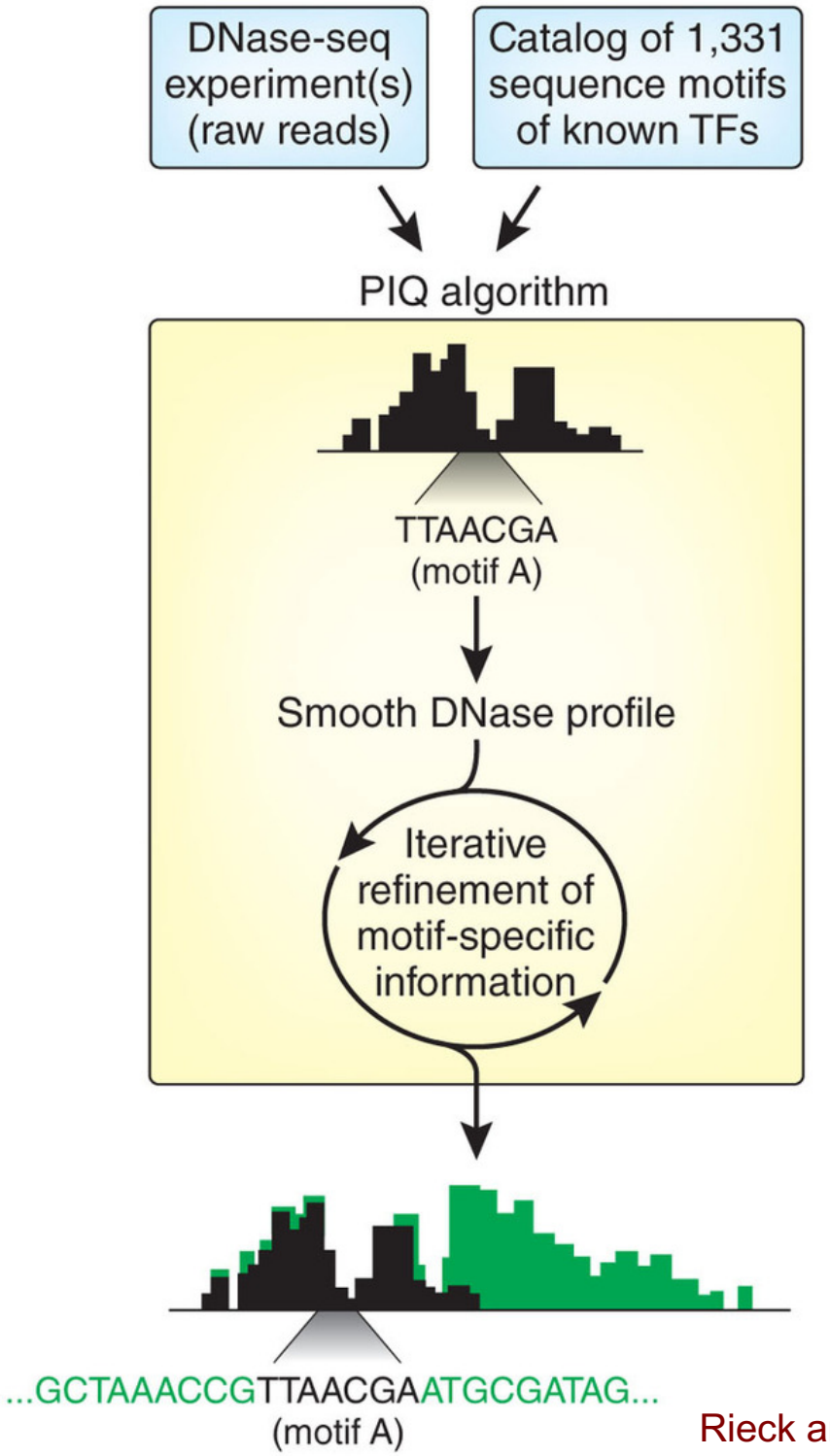
- DNase footprints suggest that ***some*** TF binds that location

- 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

Protein Interaction Quantification (PIQ)



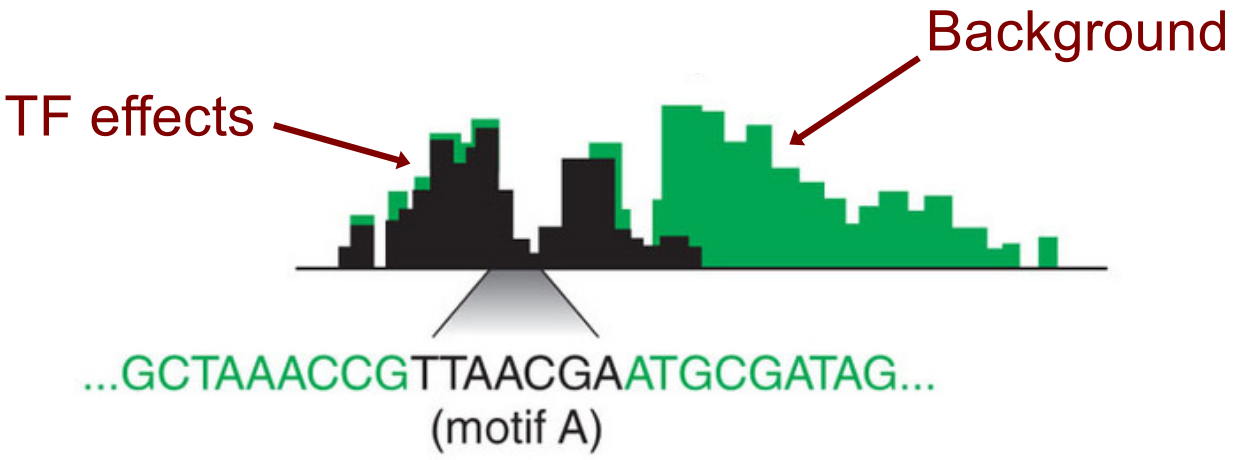
- Sherwood et al. *Nature Biotechnology* 2014
- **Given**: TF motifs and DNase-Seq reads
- **Do**: Predict binding sites of each TF

Rieck and Wright *Nature Biotechnology* 2014

PIQ main idea

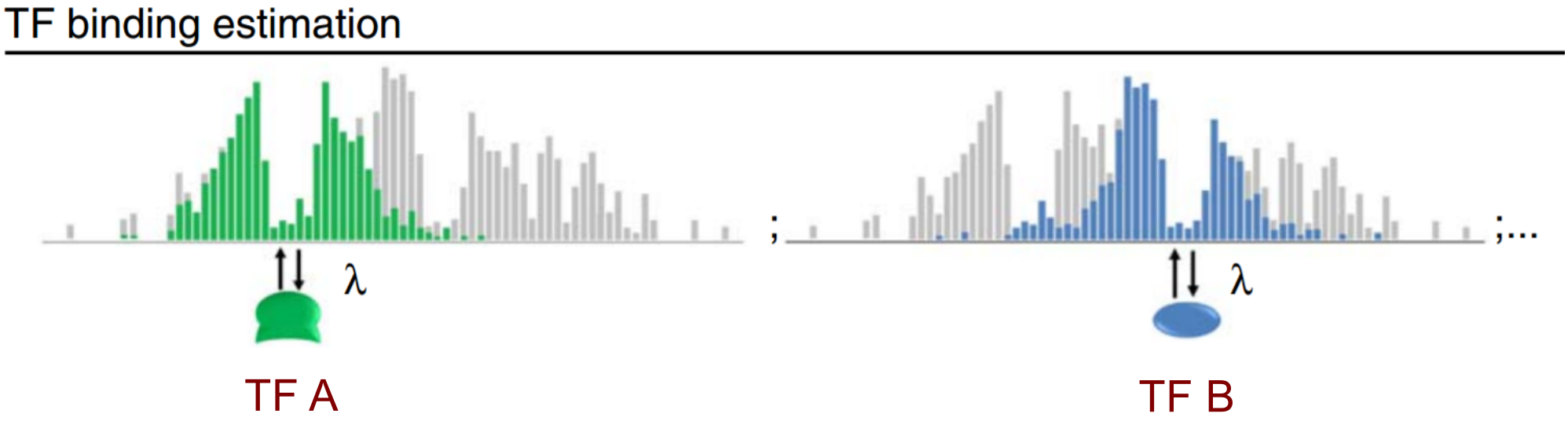
- With no TF binding, DNase-Seq reads come from some background distribution

- TF binding changes read density in a *TF-specific* way



PIQ main idea

- Shape of DNase peak and footprint depend on the TF



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

- We'll skip
  - Modeling multiple replicates or conditions, cross-experiment and cross-strand effects
  - Expectation propagation
  - 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

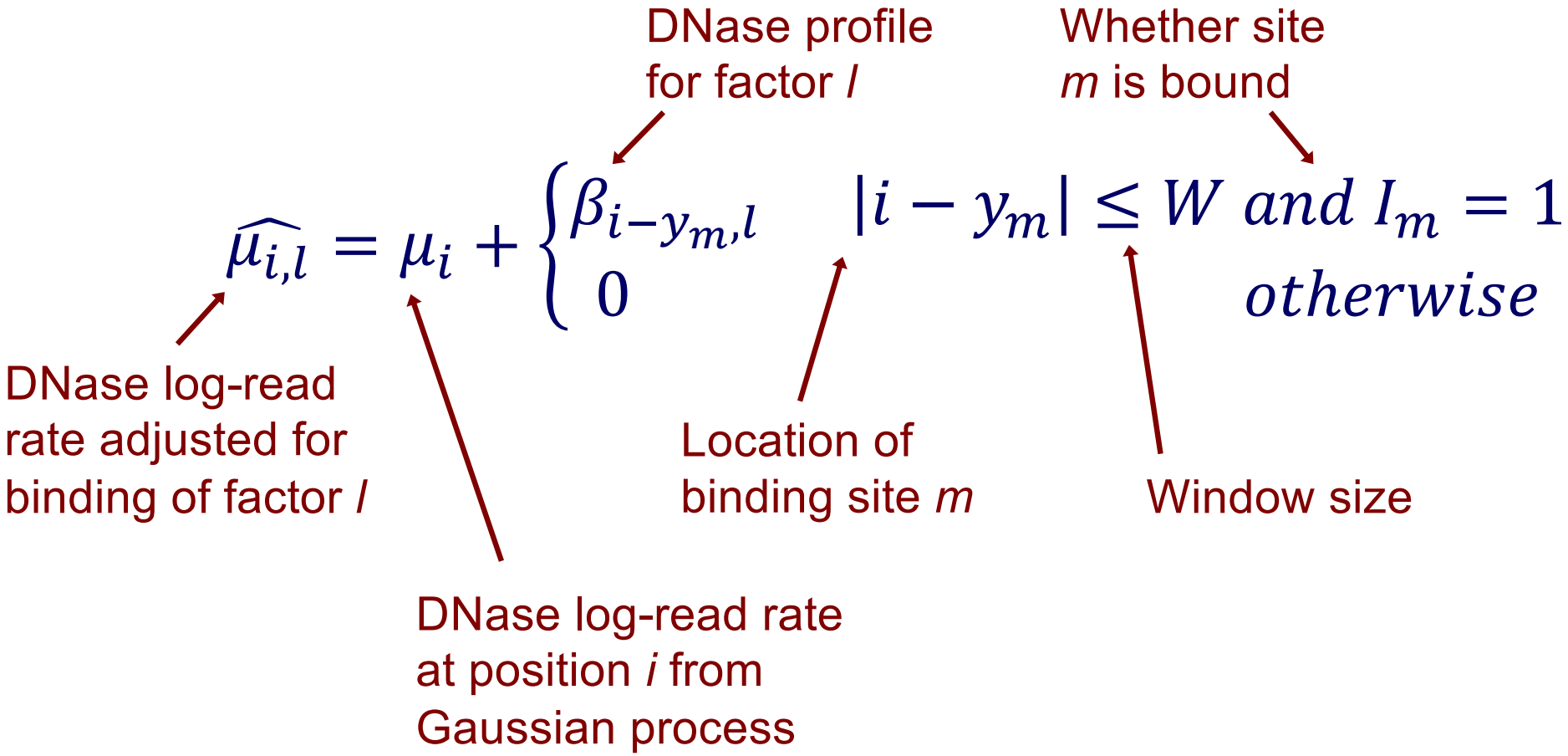
- Each replicate is noisy, don't want to over-interpret this noise
  - Only counting density of 5′ ends of reads
- Manage two competing objectives
  - Smooth some of the noise
  - Don't destroy base pair resolution signal

Gaussian processes

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

TF DNase profile

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



TF DNase profile

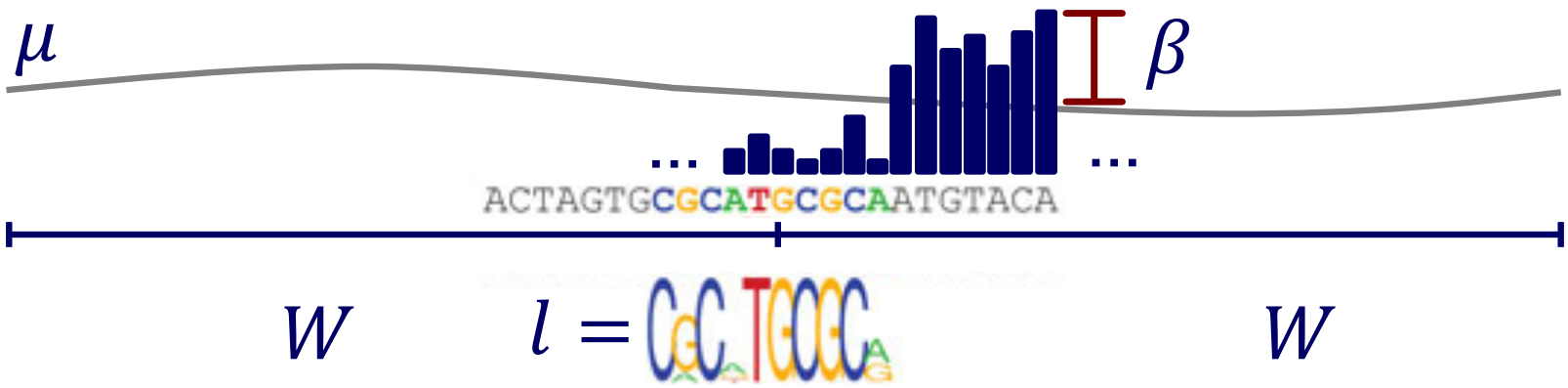
- DNase profiles represented as a vector for each TF

DNase profile for factor *l*

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

Can't be too far apart

y_m i

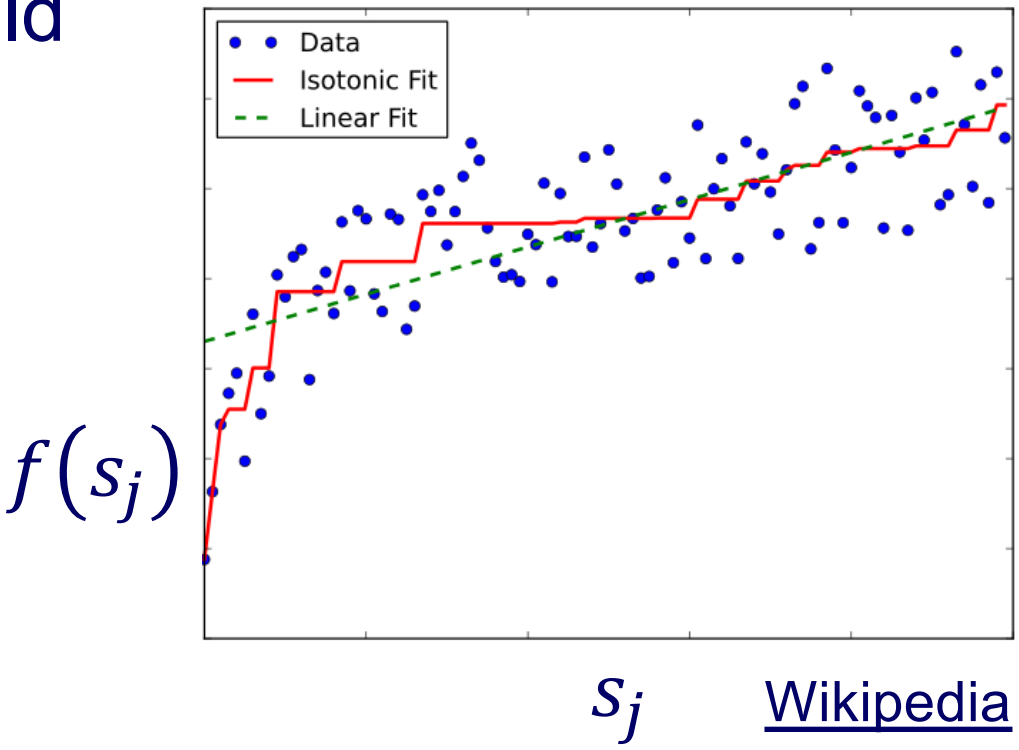


Priors on TF binding

- TF binding event I_j should be more likely when
  - motif score s_j is high
  - DNase counts c_j are high

- Isotonic (monotonic) regression

Example only, not realistic data



Wikipedia

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

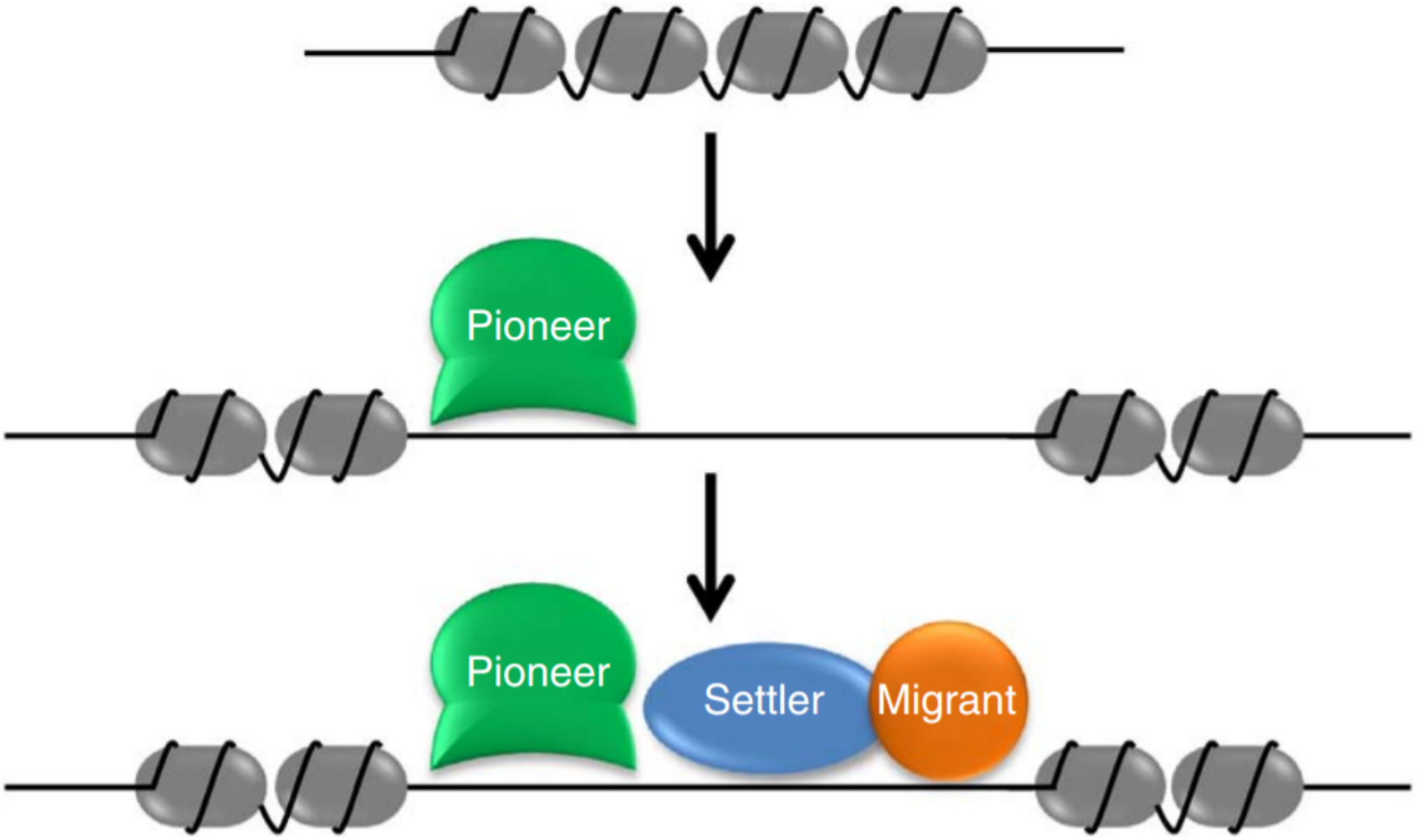
Full algorithm

- **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 $\beta_{i-j,l}$
- 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

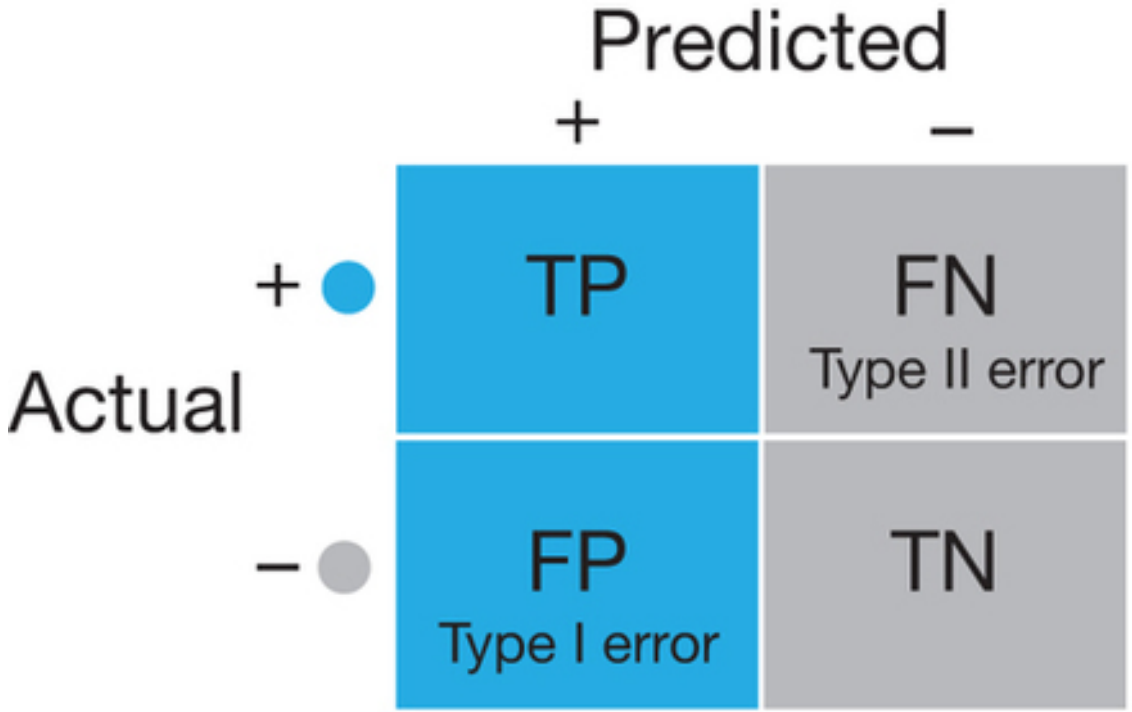
- Pioneer, settler, and migrant TFs



Sherwood *Nature Biotechnology* 2014

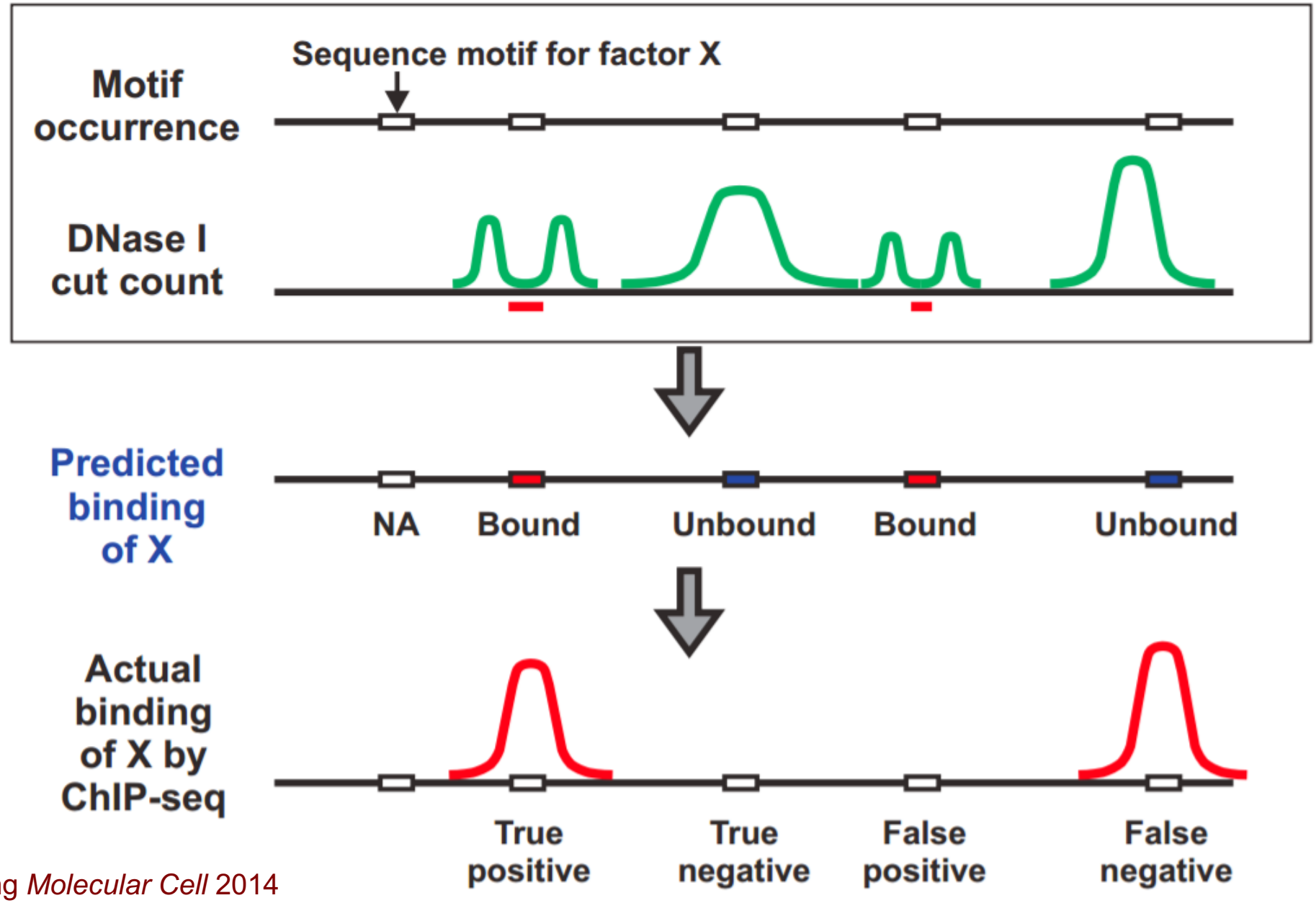
Evaluation: confusion matrix

- Compare predictions to actual ground truth (gold standard)



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)

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

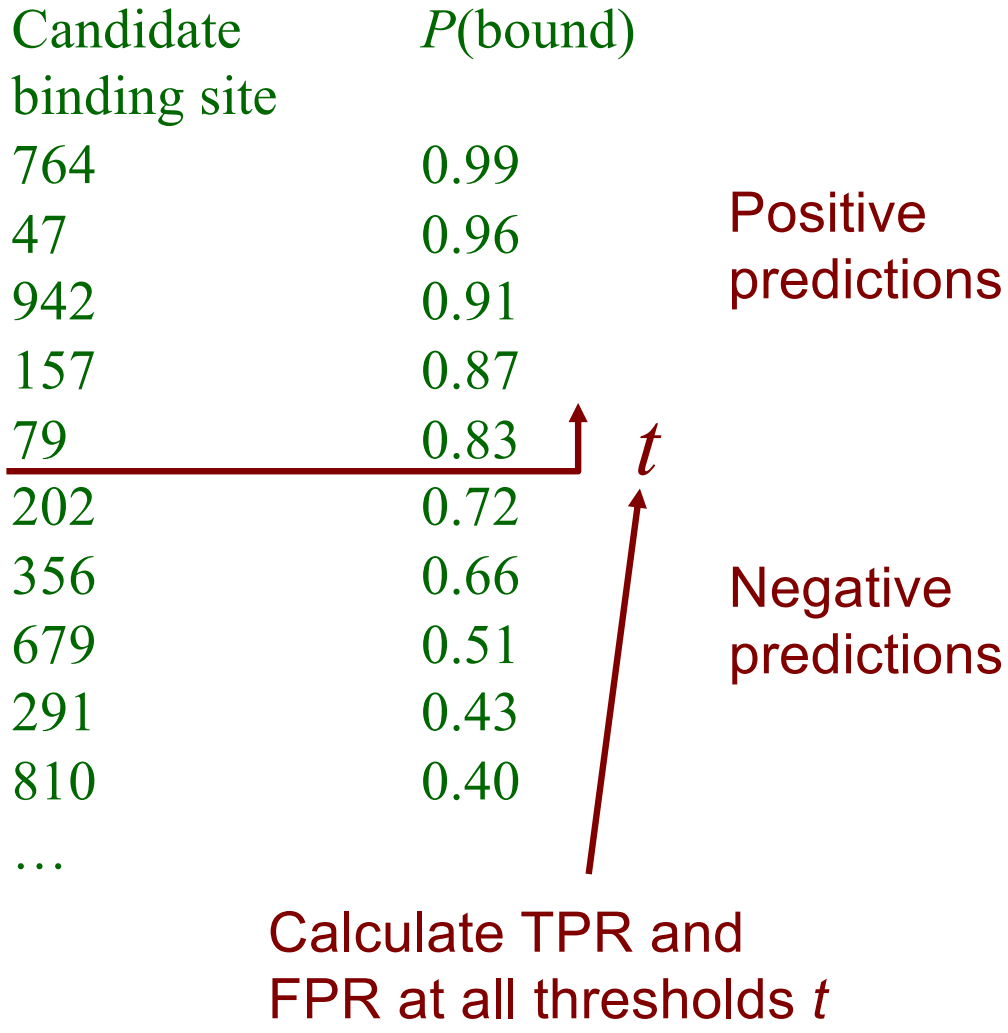
$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN}$$

Includes true negatives

Reason to prefer precision-recall for class imbalanced data

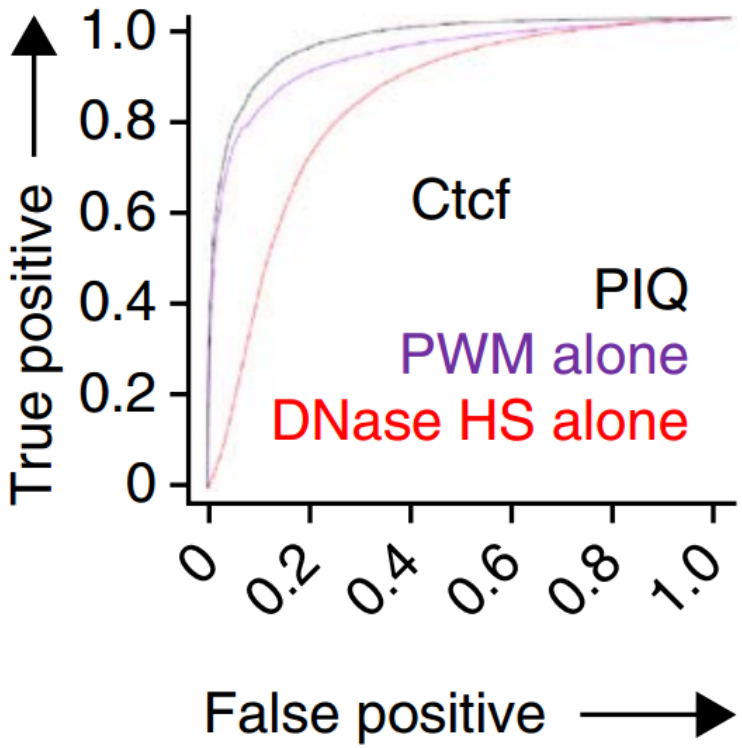
Evaluation: ROC curve

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



PIQ ROC curve for mouse Ctcf

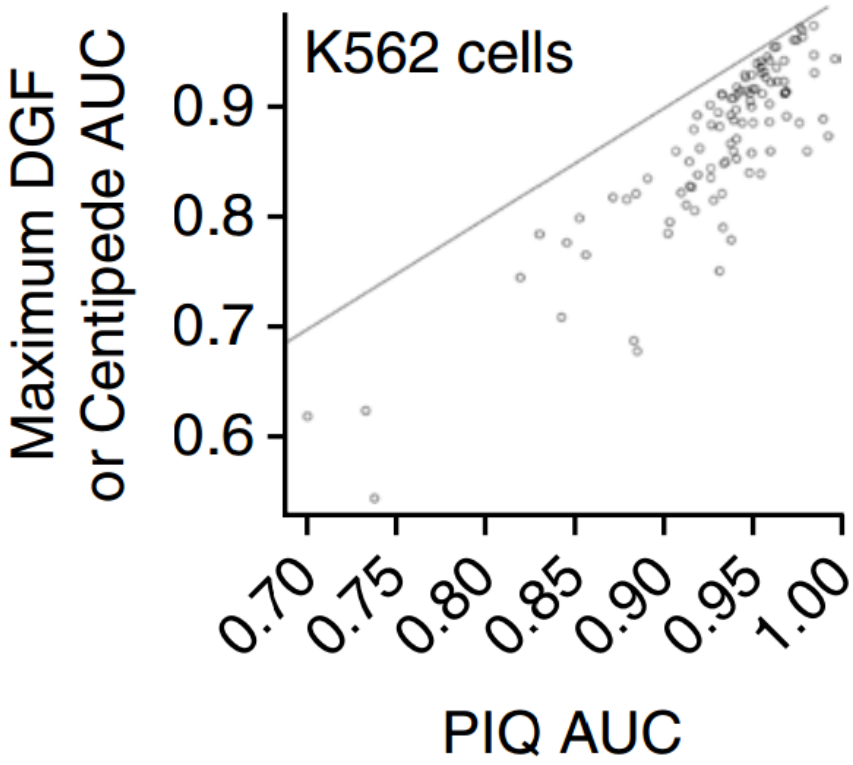
- Compare predictions to ChIP-Seq
- Full PIQ model improves upon motifs or DNase alone



Sherwood *Nature Biotechnology* 2014

PIQ evaluation

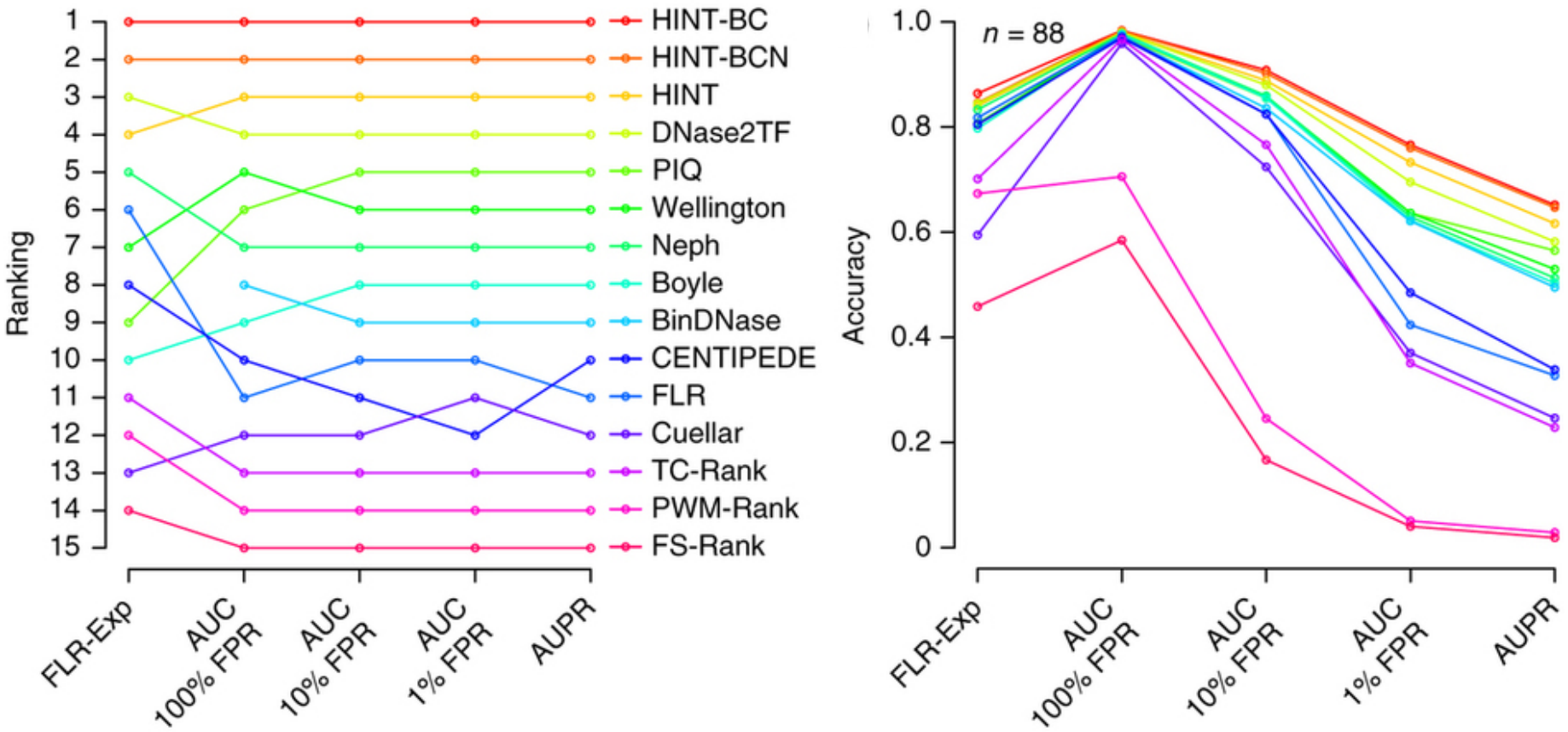
- Compare to two standard methods
  - 303 ChIP-Seq experiments in K562 cells
  - Centipede, digital genomic footprinting
- Compare AUROC
  - PIQ has very high AUROC
  - Mean 0.93
  - Corresponds to recovering median of 50% of binding sites



Sherwood *Nature Biotechnology* 2014

DNase-Seq benchmarking

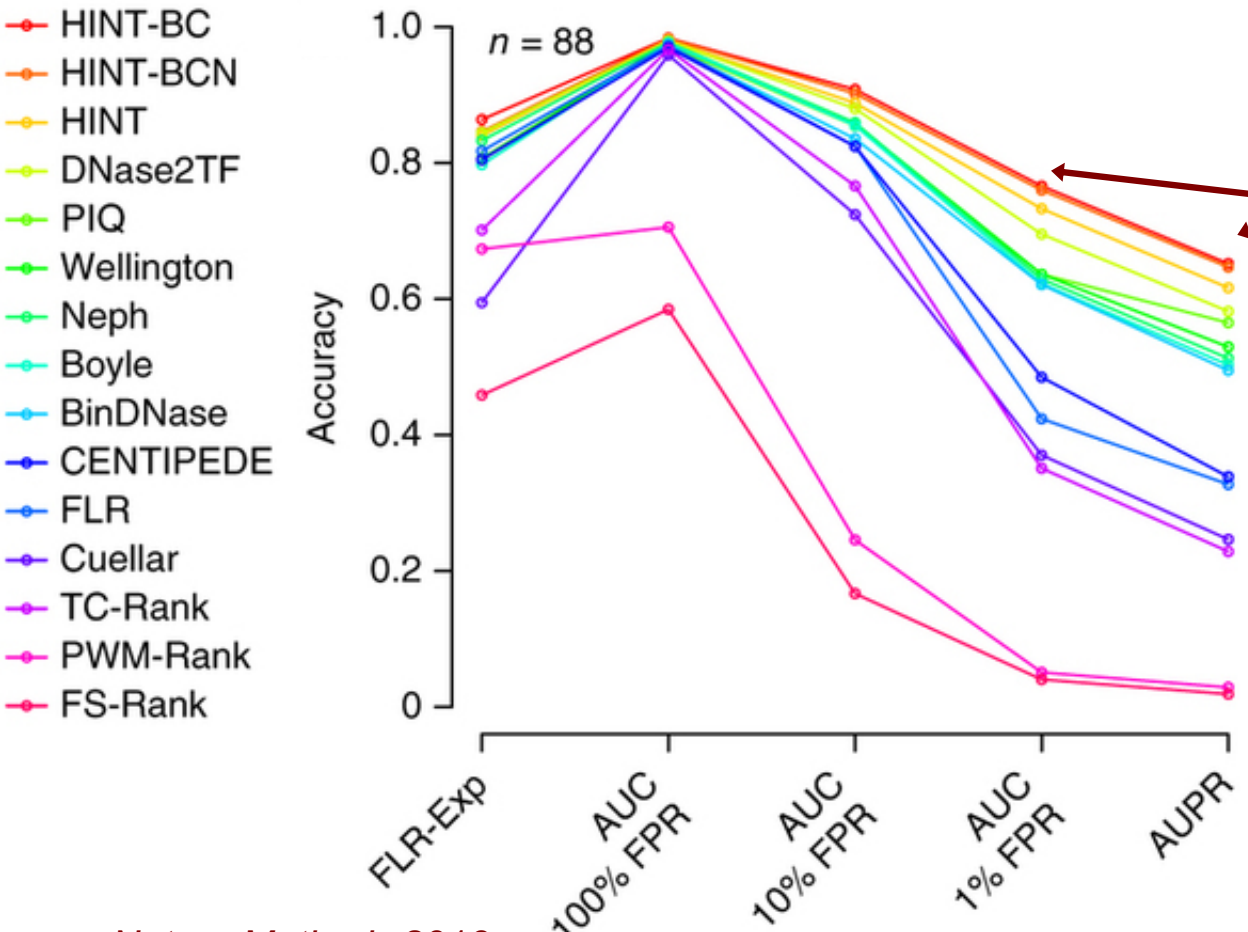
- PIQ among top methods in large scale DNase benchmarking study
- HMM-based model HINT was top performer



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)