Multi-state identification of transcription factor binding sites from DNase-seq data

Aleksander Jankowski

March 27, 2018

UNIVERSITY OF WARSAW

Genome Institute of Singapore

Multi-state identification of transcription factor binding sites from DNase-seq data

Aleksander Jankowski

March 27, 2018

Jerzy Tiuryn

UNIVERSITY OF WARSAW

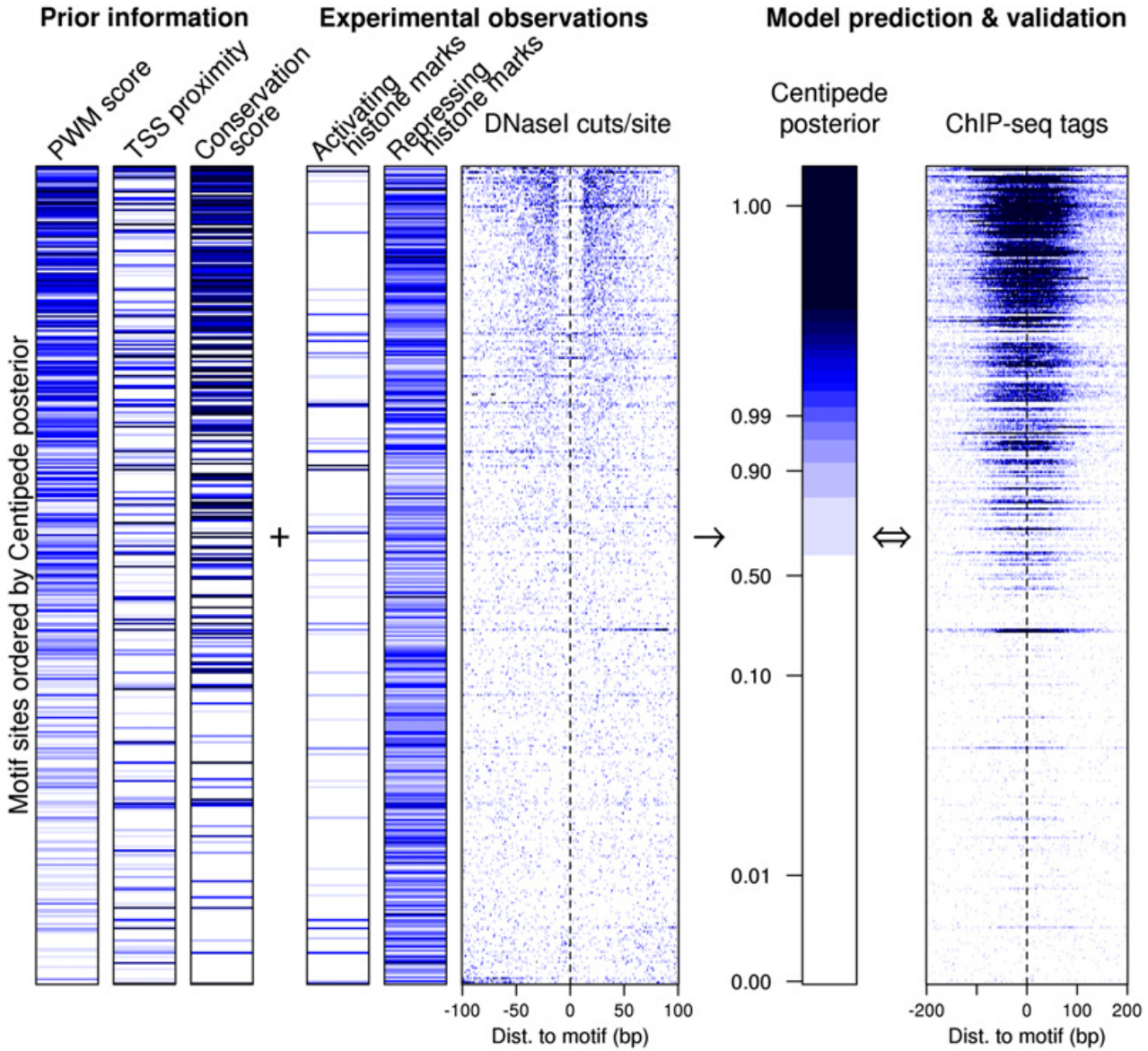
Genome Institute of Singapore
A*STAR

Shyam Prabhakar

Introduction

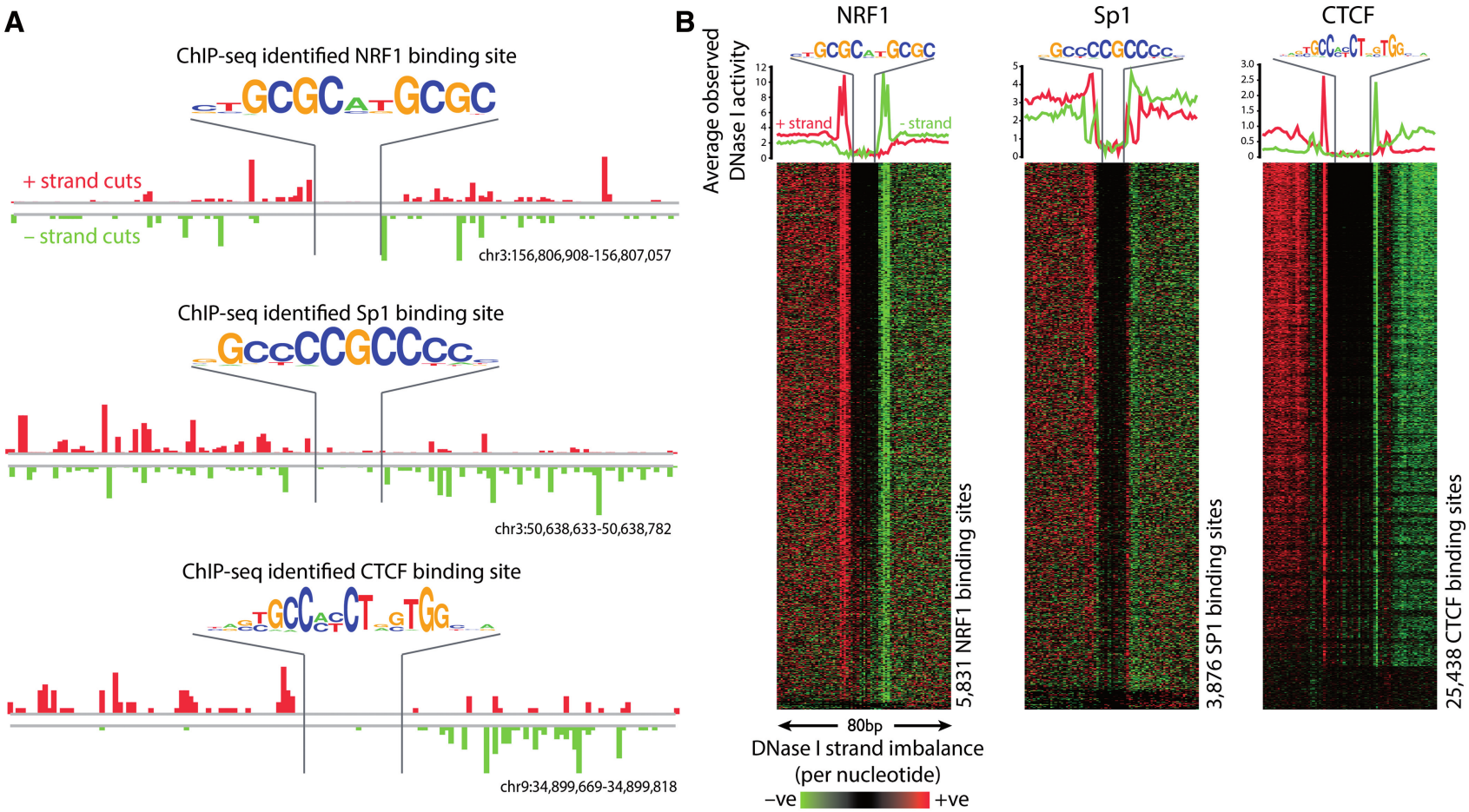
- Our goal is to identify individual transcription factor (TF) binding sites from genome sequence information and cell-type–specific experimental data, such as DNase-seq.
- We present Romulus (Jankowski et al., *Bioinformatics* 2016), a novel computational method for this purpose.
- Romulus combines the strengths of previous approaches, and improves robustness by reducing the number of free parameters in the model by an order of magnitude.

Previous approach: CENTIPEDE



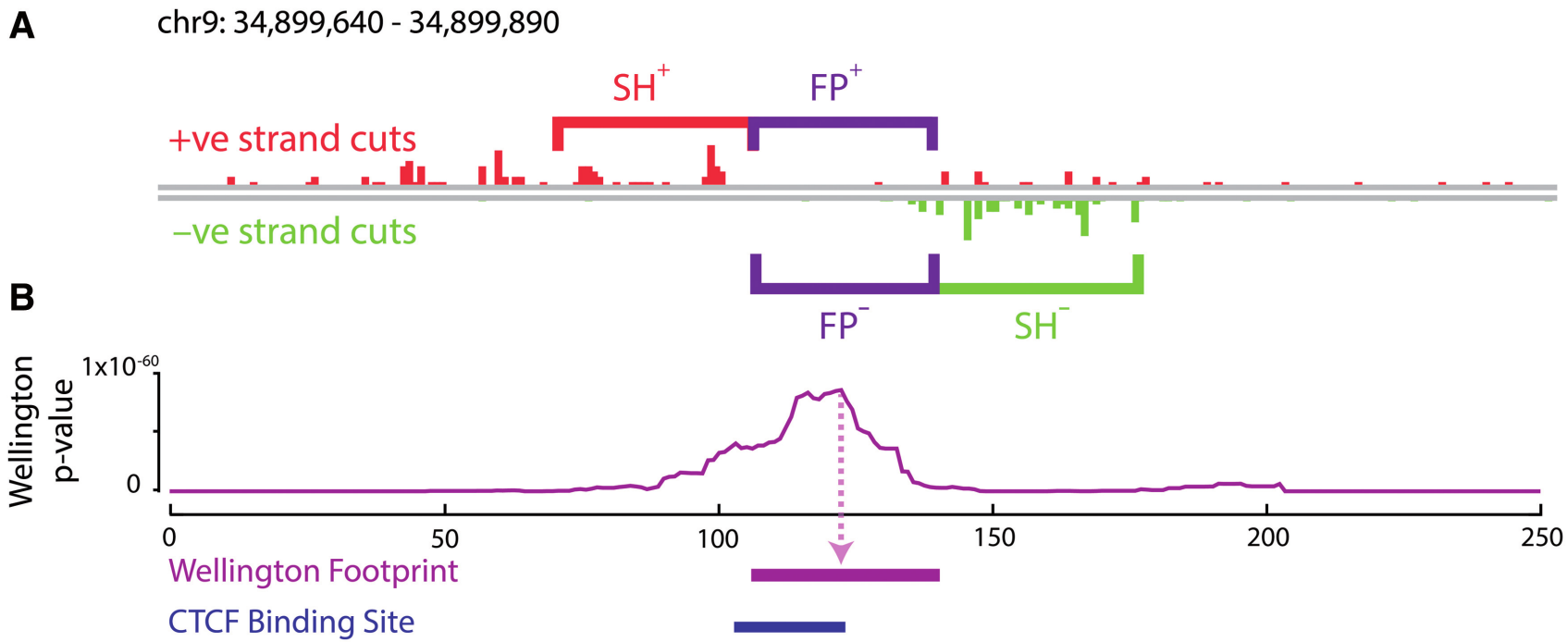
Pique-Regi et al., *Genome Res.* 2011

Previous approach: Wellington



Piper et al., *Nucleic Acids Res.* 2013

Previous approach: Wellington

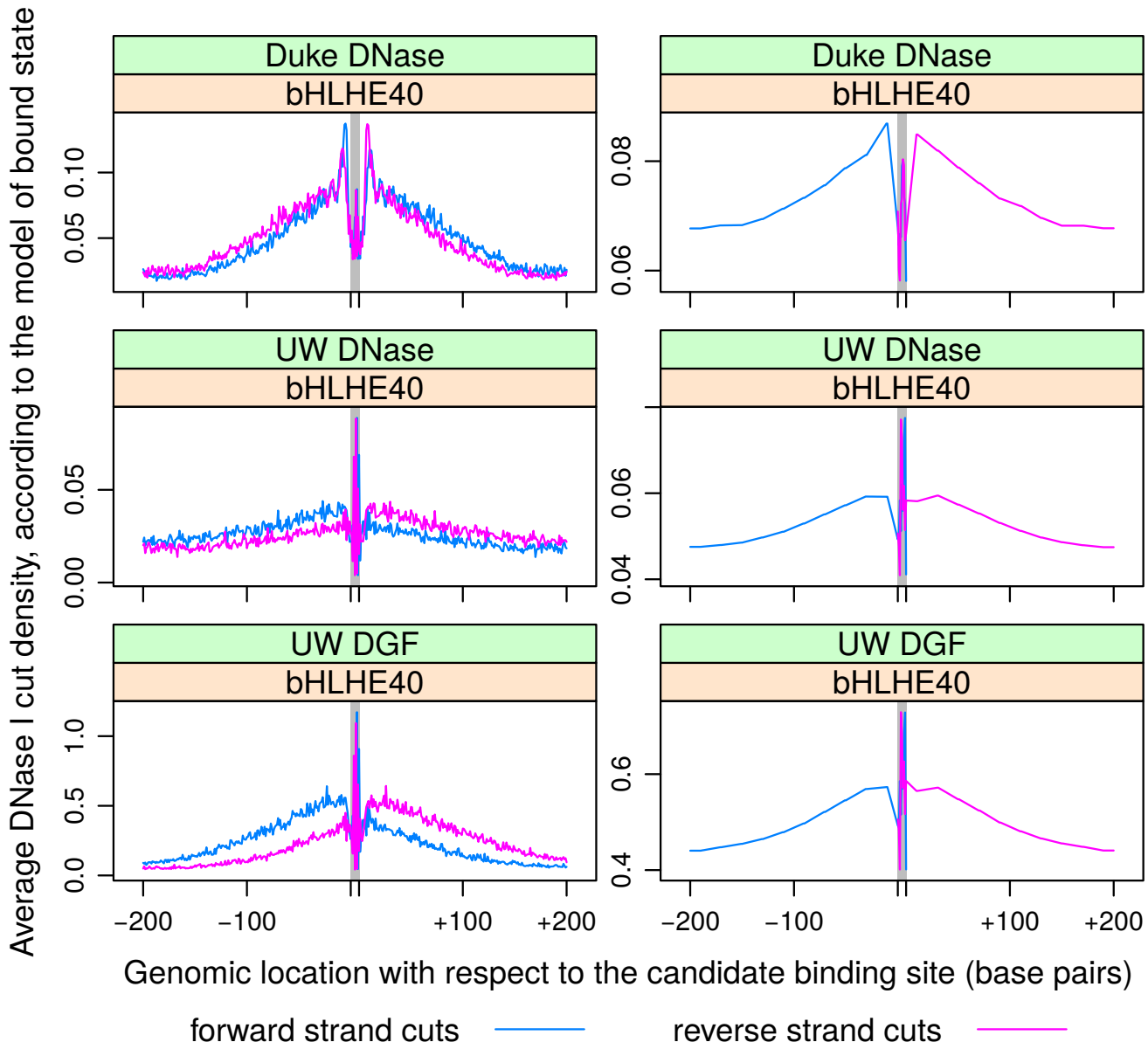


Piper et al., *Nucleic Acids Res.* 2013

Romulus approach

- For a given TF, we first identify candidate binding sites that have reasonable sequence affinity, using a position weight matrix.
- We employ an Expectation-Maximization-based approach to simultaneously learn the DNase I cut profiles and classify the binding sites as bound or unbound.
- Our method is unique by allowing for multiple bound states for a single TF, differing in their cut profile and overall affinity for DNase I cuts.
- We achieve robustness by grouping the DNase I cuts into bins, according to their location and strand.

Example DNase I cut profiles: CENTIPEDE vs. Romulus



Prior probabilities of TF binding: two-state case

The prior component captures the genomic sequence and other prior (i.e. independent of cell type or conditions) characteristics of the candidate binding site for a given TF.

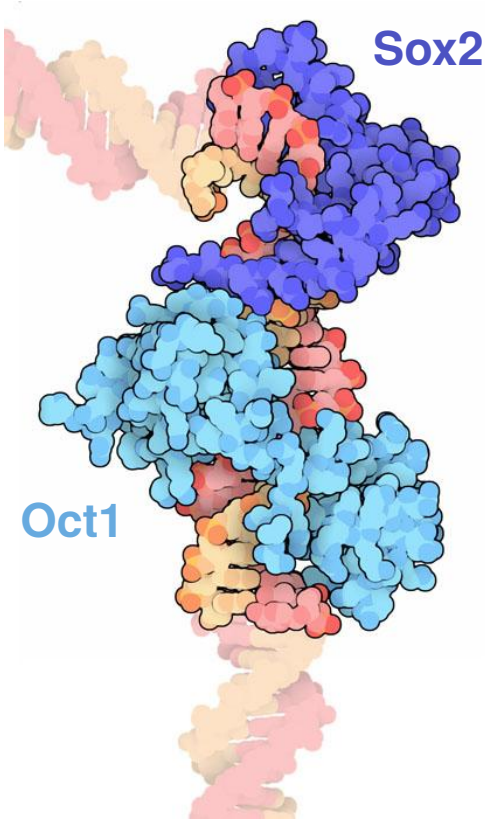
Let $x_i^{(j)}$ be the value of the j-th prior characteristic ($1 \leq j \leq J$) for genomic instance i.

In the simplest case, where motif instance i can be either "bound" or "unbound", we apply a logistic model:

$$\frac{P(Z_i = 1)}{P(Z_i = 0)} = \exp\left(\beta_0 + \beta_1 \cdot x_i^{(1)} + \beta_2 \cdot x_i^{(2)} + \beta_3 \cdot x_i^{(3)} + \dots\right)$$

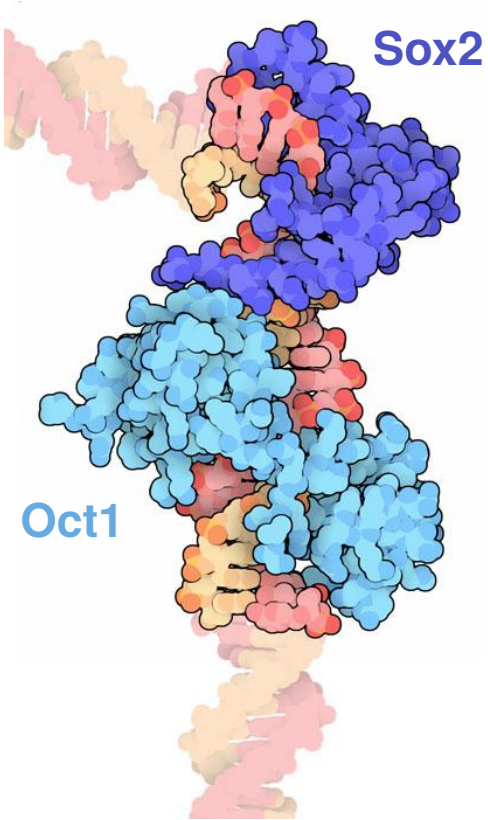
Here, $Z_i = 1$ indicates that the i-th motif instance is bound, whereas $Z_i = 0$ indicates that it remains unbound.

Why we should consider multiple binding modes?

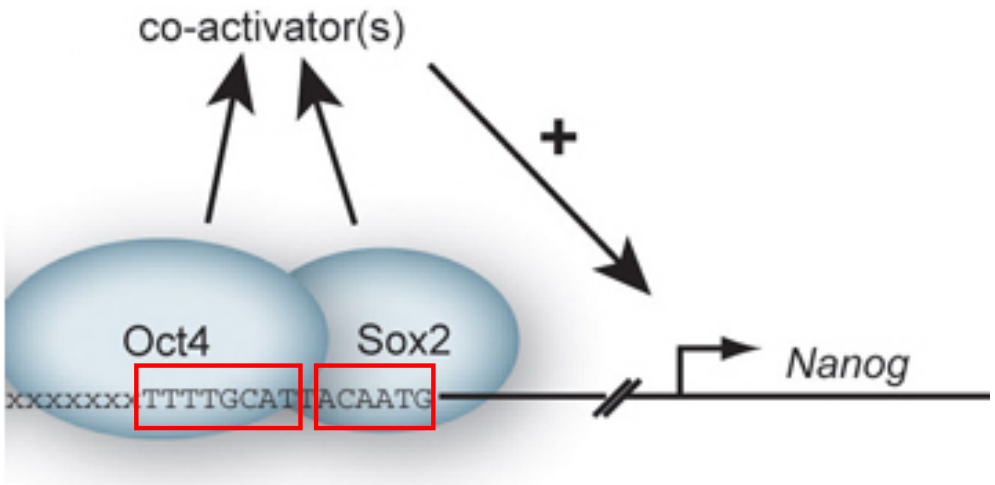


RCSB Protein Data Bank

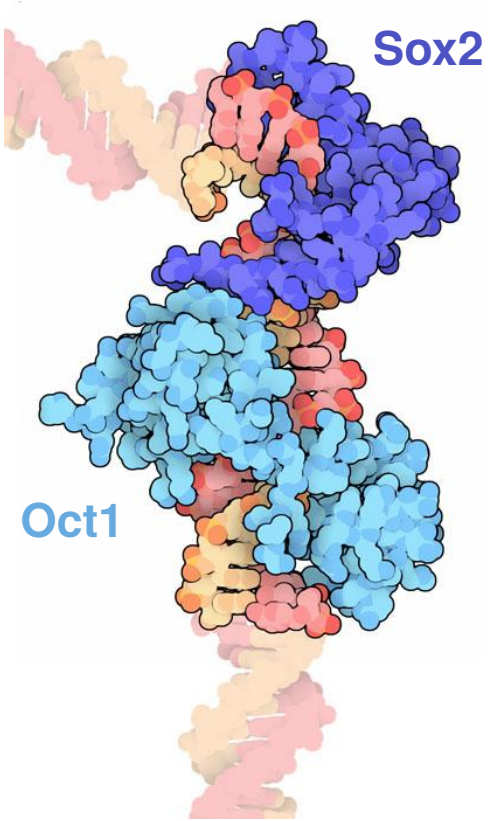
Why we should consider multiple binding modes?



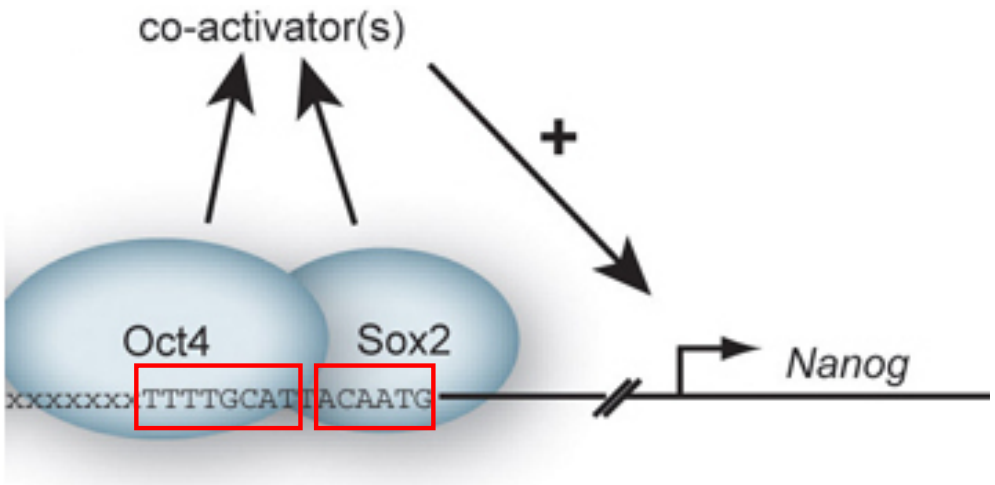
RCSB Protein Data Bank



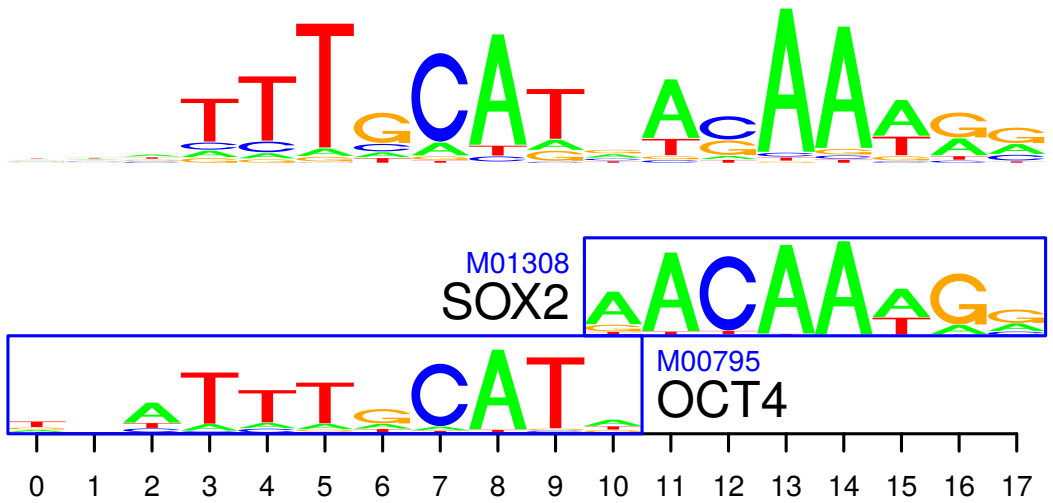
Why we should consider multiple binding modes?



RCSB Protein Data Bank



van den Berg et al., Mol Cell Biol. 2008 Oct;28(19):5986-95.



Prior probabilities of TF binding: general case

To model the prior probabilities in general case of multiple binding modes ($k = 1, \ldots$), we apply a logistic model against the unbound "pivot" case ($k = 0$):

$$\frac{P(Z_i = k)}{P(Z_i = 0)} = \exp\left(\beta_0^{(k)} + \beta_1^{(k)}\gamma_1^{(k)} \cdot x_i^{(1)} + \beta_2^{(k)}\gamma_2^{(k)} \cdot x_i^{(2)} + \beta_3^{(k)}\gamma_3^{(k)} \cdot x_i^{(3)} + \ldots\right)$$

where the indicators $\gamma_j^{(k)} \in \{0, 1\}$ specify whether the prior characteristic $x_i^{(j)}$ should be taken into account in the k-th binding mode.

Prior probabilities of TF binding: cooperative binding

Consider a TF that manifests one or more cooperative binding modes ($k = 2, \ldots, K + 1$), with well-defined structures of the underlying motif complexes.

The prior characteristic for these partner motif instances are calculated no matter how favorable they may be for binding, and are included in the sequence $x_i^{(j)}$.

The monomer binding mode ($k = 1$) should be characterized only by the characteristics referring to the primary motif instance. Hence, $\gamma_j^{(1)} = 0$ for all the characteristics j referring to any of the partner motifs.

The dimer binding modes ($k = 2, \ldots, K + 1$) should have indicators $\gamma_j^{(k)}$ ensuring that only the characteristics specific to the primary motif instance and to the partner motif instances within the motif complex k are taken into account.

Chromatin state component of Romulus model

The probability of observing a given distribution of DNase I cuts on a given strand is calculated as a product of negative binomial and multinomial components:

$$P((\mathrm{DNase}_{i,j})_j \mid Z_i = k) = \mathrm{NegativeBinomial}(\mathrm{DNaseSum}_i^{(k)} \mid p^{(k)}, r^{(k)}) \cdot \cdot \mathrm{Multinomial}((\mathrm{DNaseBin}_{i,b}^{(k)})_b \mid \mathrm{DNaseSum}_i^{(k)}, (\lambda_b^{(k)})_b) \cdot \quad (1)$$

where $\mathrm{DNaseSum}_i^{(k)}$ is the number of DNase I cuts within 200 bp from the motif complex, and $\mathrm{DNaseBin}_{i,b}^{(k)}$ is the number of DNase I cuts in b-th bin, where $b = 1, \ldots, B$.

We impose an additional constraint: the multinomial coefficients $\lambda_b^{(0)}$ are proportional to the bin sizes, i.e. there is no positional preference for DNase I cuts in the unbound case.

Expectation-Maximization approach

to simultaneously learn model parameters and classify the binding sites

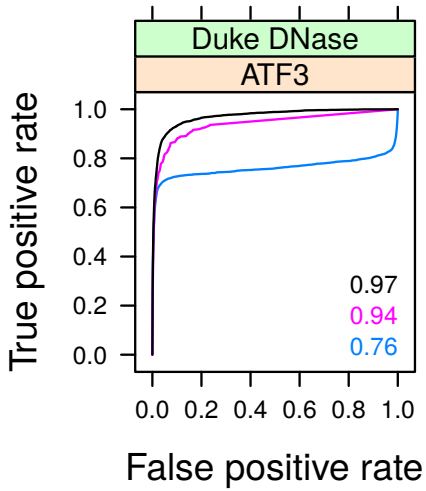
- Expectation (“E”): for each site, estimate its likelihood to be bound (possibly considering multiple binding modes).
- Maximization (“M”): for each binding mode, estimate its parameters (defining the prior component, DNase I cut profile, and the total number of DNase I cuts).
- We iterate the ExpectationMaximization procedure, in each iteration getting a revised vector of parameters, until the posterior probabilities do not change by more than 0.001.

Benchmarking approach

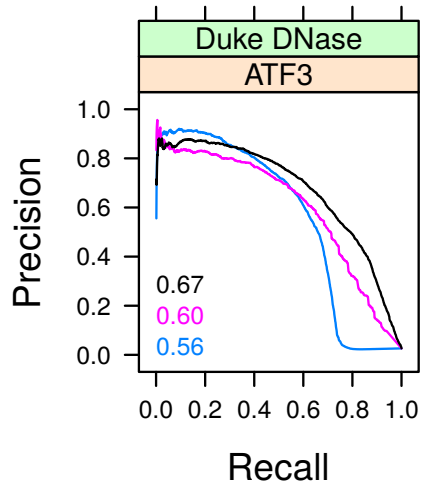
- We systematically benchmarked Romulus along with CENTIPEDE and Wellington.
- We applied all the methods in an unsupervised manner to DNase-seq data from three ENCODE sources:
  - “single hit” protocol: Duke DNase
  - “double hit” protocol: University of Washington (UW) DNase and UW Digital Genomic Footprinting (DGF).
- From each of the DNase-seq data sources, we consider three human cell lines: A549 (lung adenocarcinoma epithelial), HepG2 (hepatocellular carcinoma) and K562 (leukemia).
- To validate the predictions, we used 39 ChIP-seq datasets from ENCODE to define genuine TF binding sites. Note that no ChIP-seq data were used for training.

Benchmarking statistics

$$\text{True positive rate} = \frac{TP}{TP+FN}$$



$$\text{Precision} = \frac{TP}{TP+FP}$$



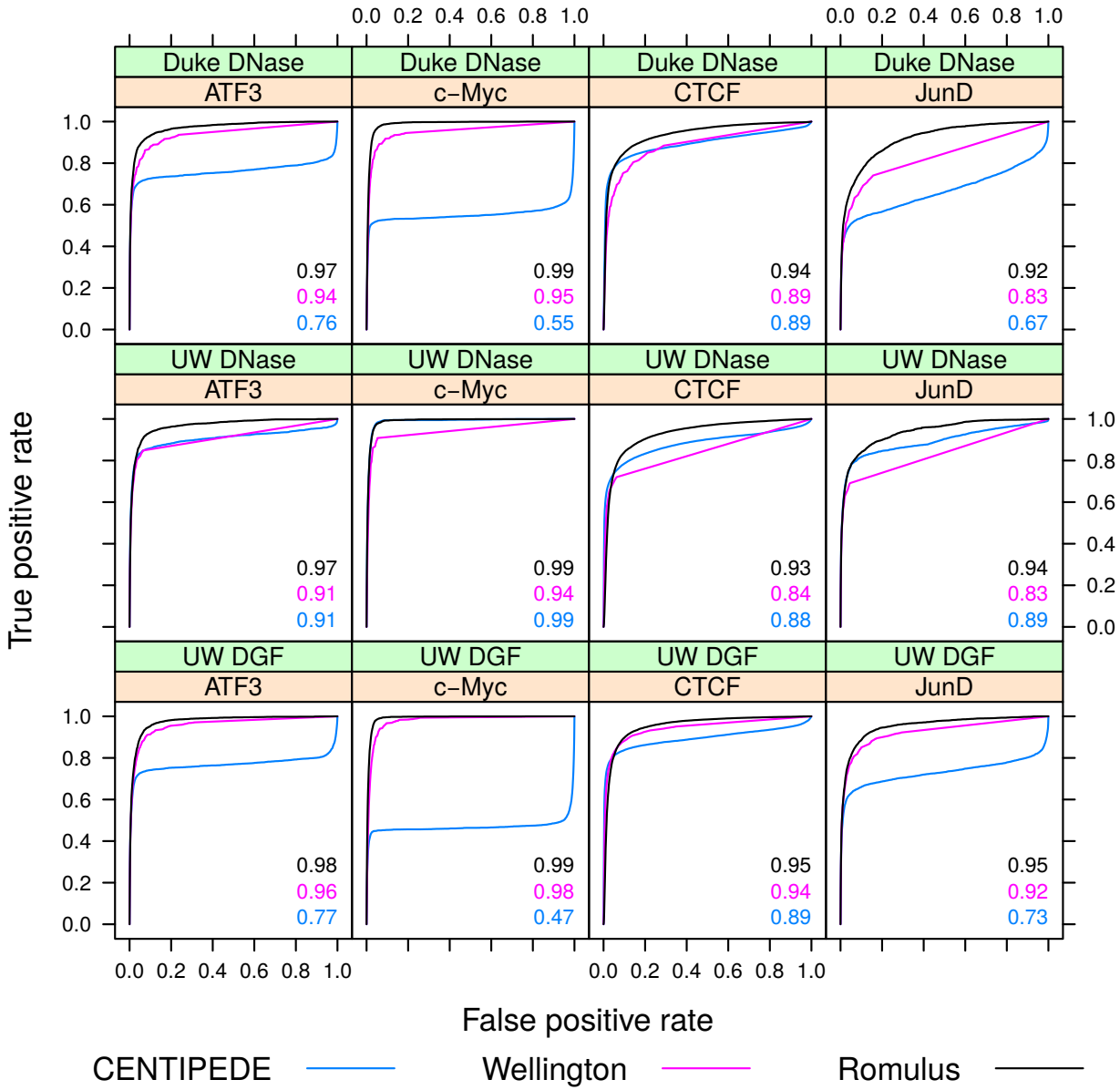
$$\text{False positive rate} = \frac{FP}{FP+TN}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

CENTIPEDE Wellington Romulus

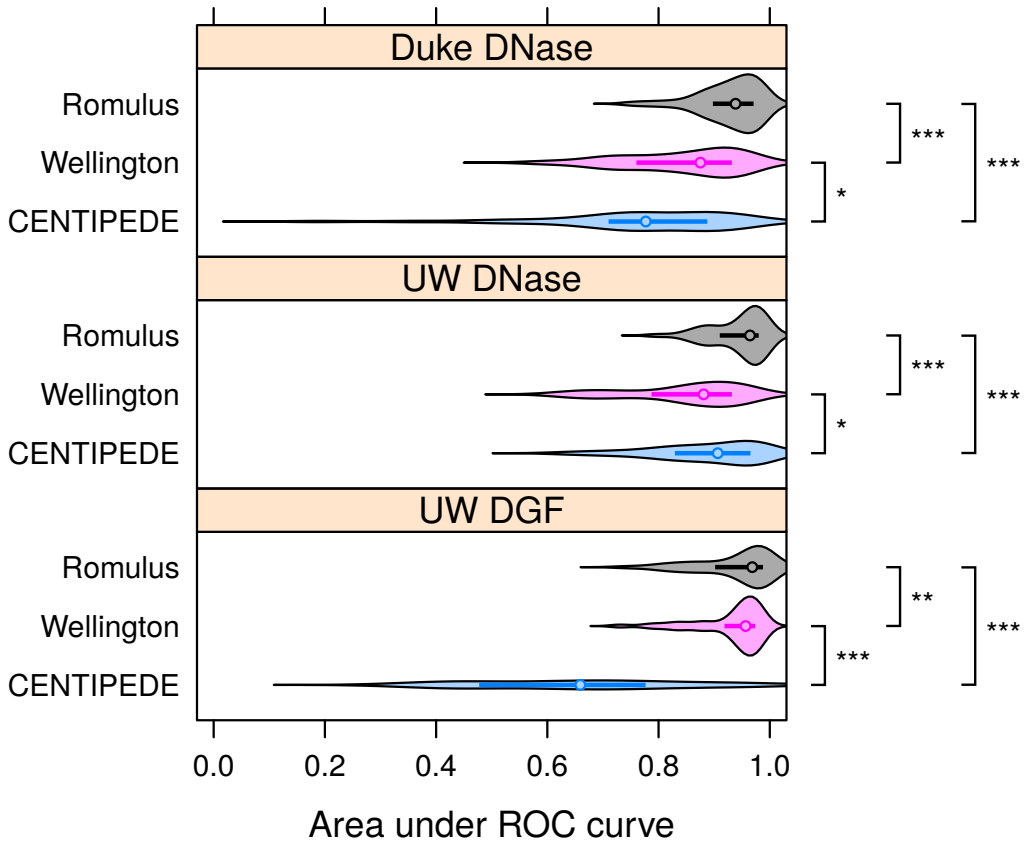
Romulus systematically outperforms existing methods

as measured by Area under Receiver Operating Characteristic curves



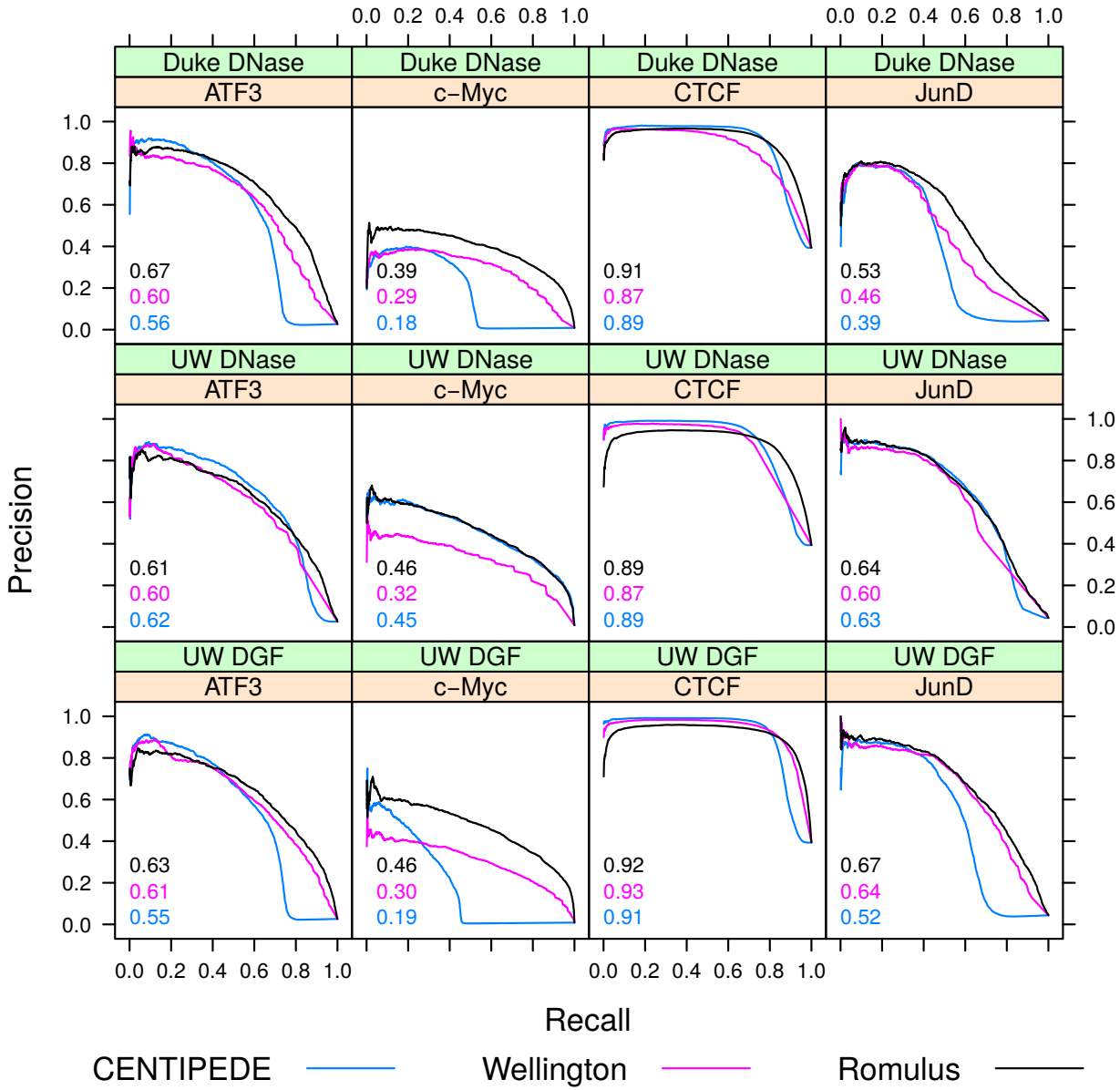
Romulus systematically outperforms existing methods

as measured by Area under Receiver Operating Characteristic curves



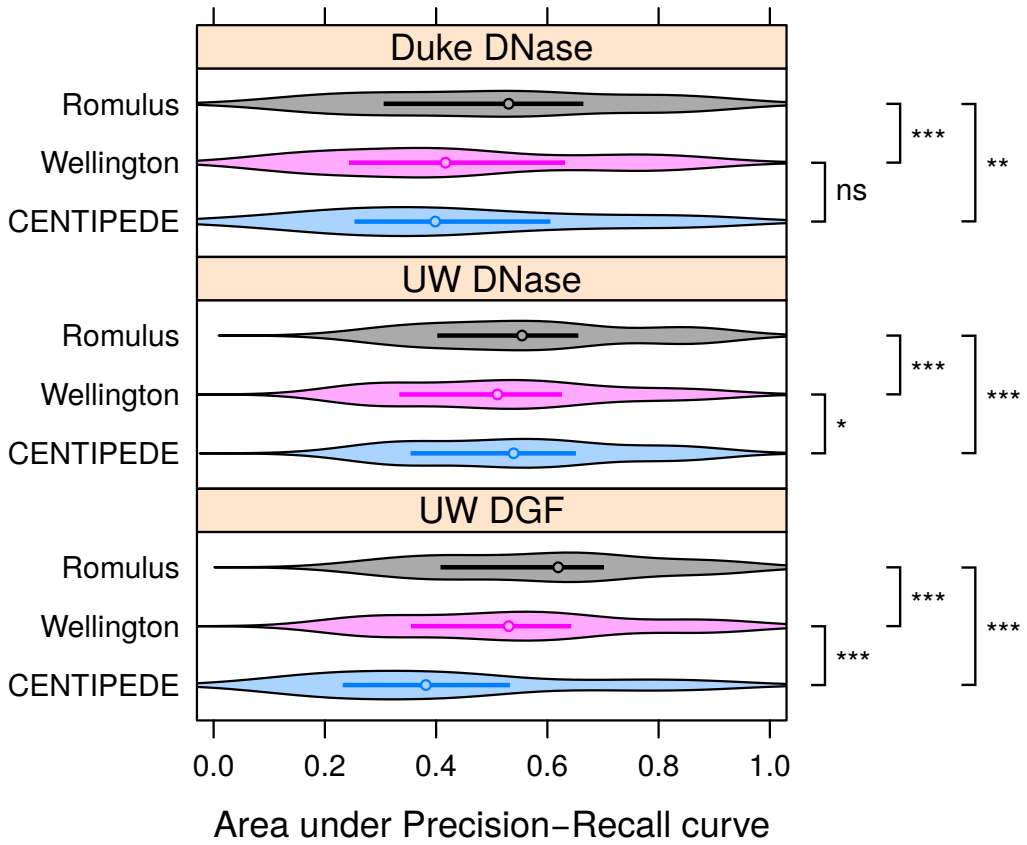
Romulus systematically outperforms existing methods

as measured by Area under Precision-Recall curves



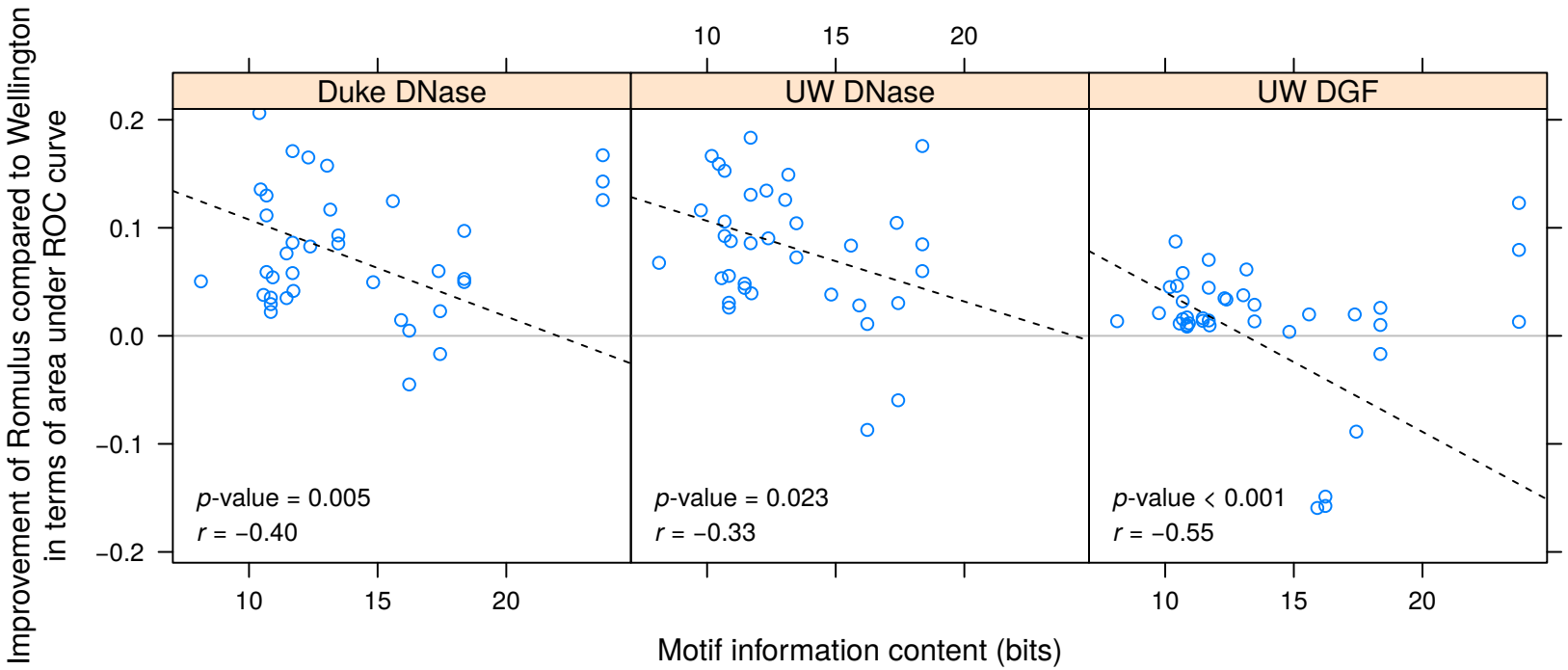
Romulus systematically outperforms existing methods

as measured by Area under Precision-Recall curves



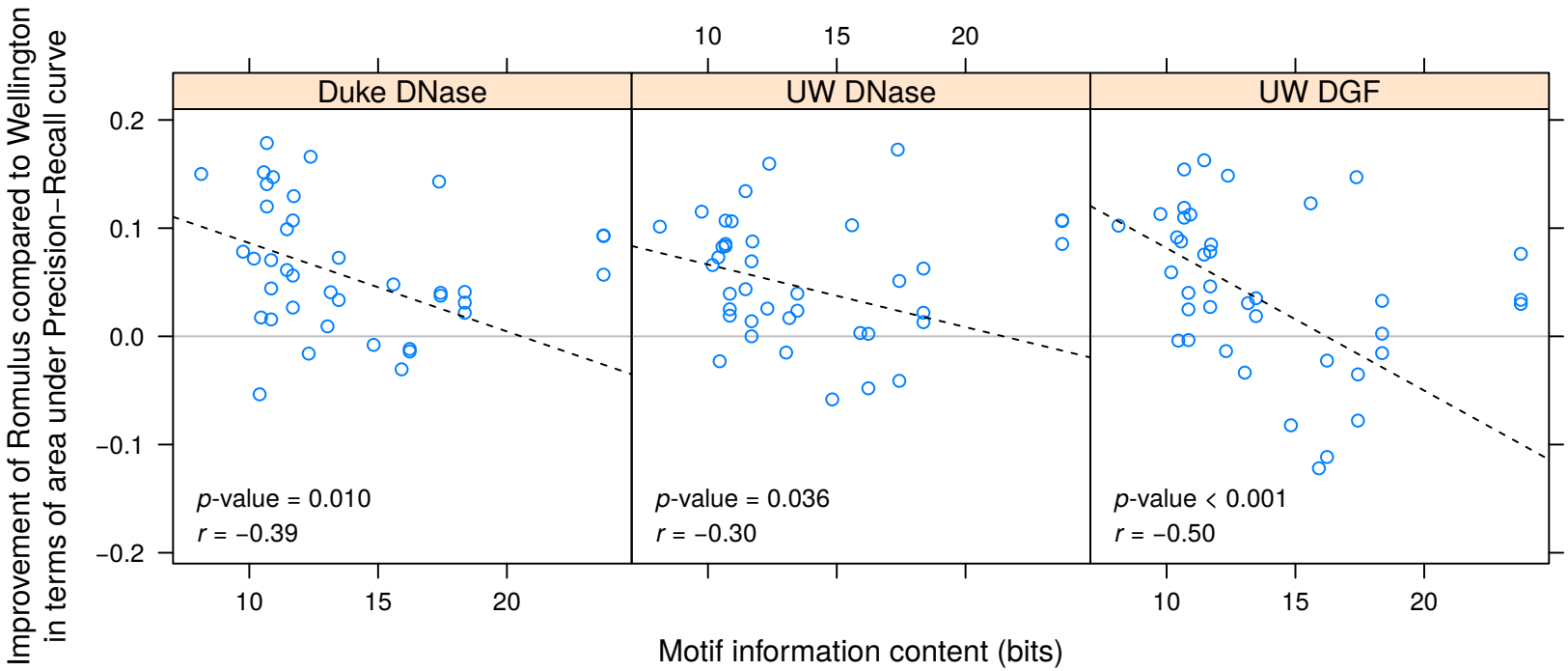
Gain of predictive power of Romulus over Wellington

is significantly higher for TFs with low-information-content motifs



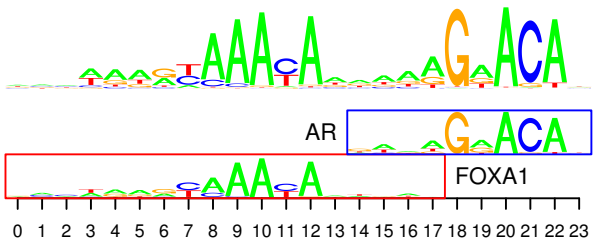
Gain of predictive power of Romulus over Wellington

is significantly higher for TFs with low-information-content motifs

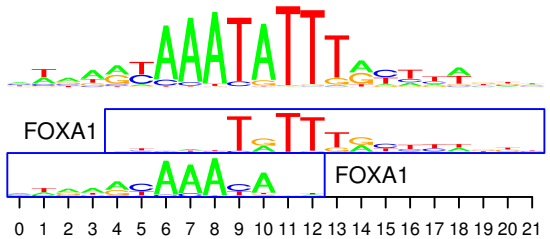


Predicted FOXA1 dimer interactions in LNCaP cells

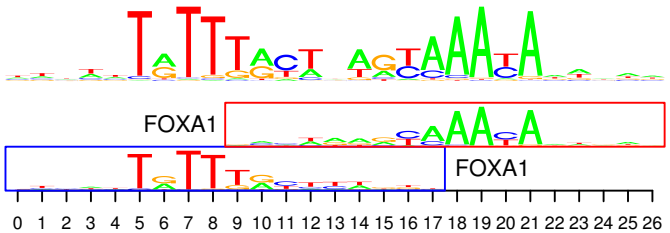
A LNCaP
885 instances
p–value: 2.91×10^{-135}



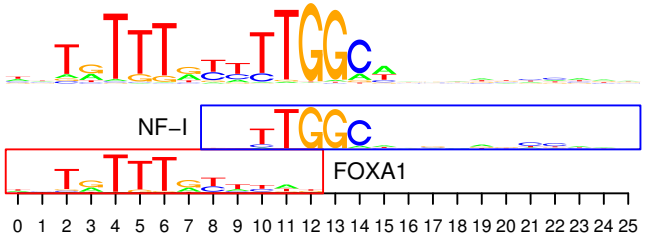
B LNCaP + MCF–7
1592 + 523 = 2115 instances
p–value: 5.14×10^{-93}



C LNCaP
650 instances
p–value: 8.78×10^{-18}

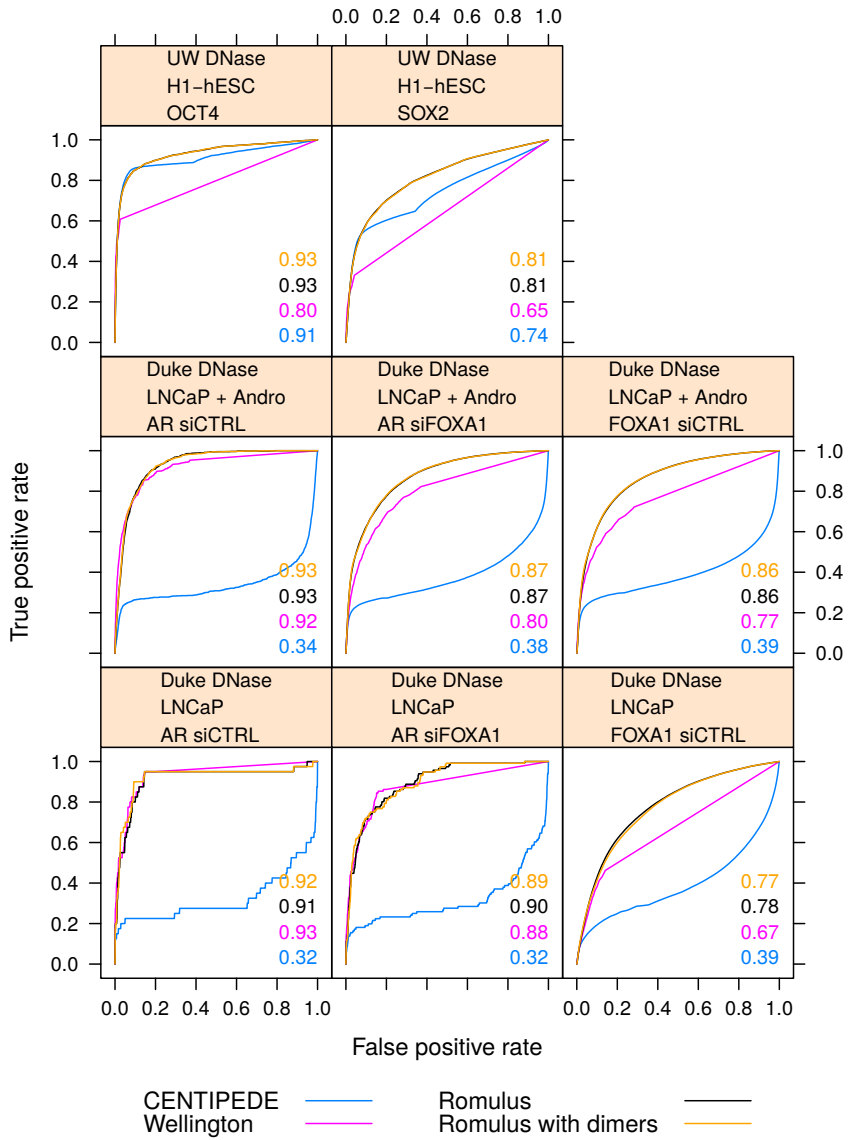


D LNCaP
1062 instances
p–value: 6.35×10^{-15}

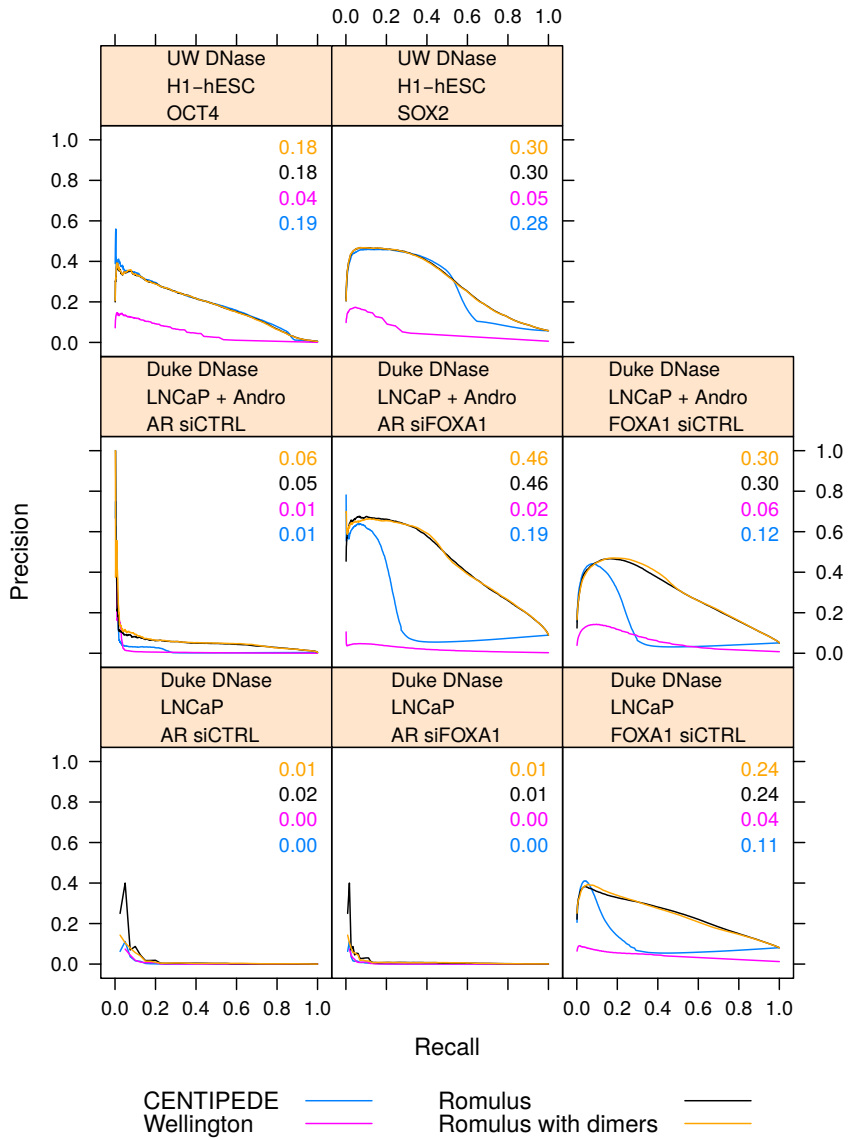


Jankowski et al., *Genome Res.* 2013

Knowledge of TF dimerization modes does not improve the prediction of individual TF binding sites

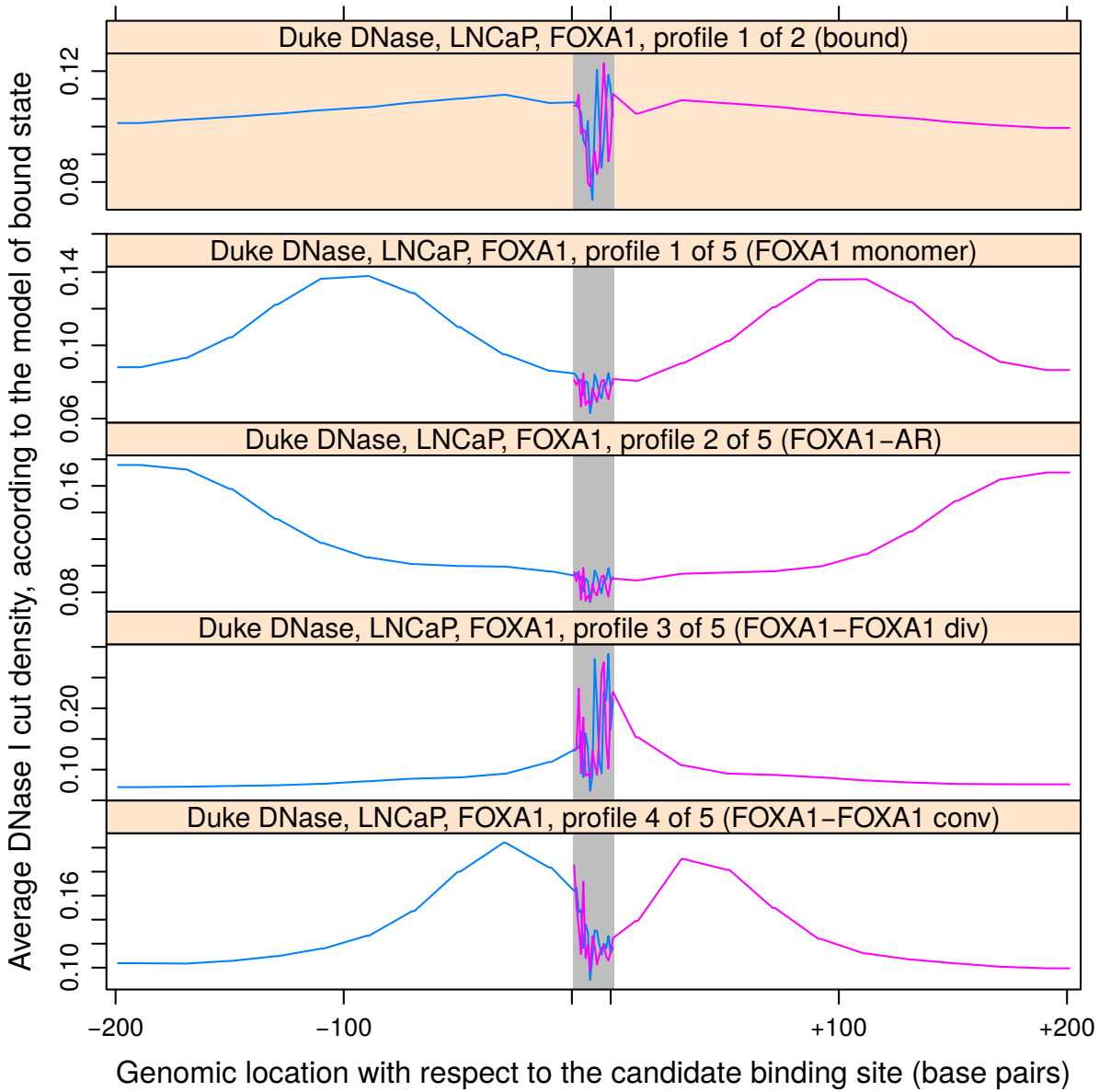


Knowledge of TF dimerization modes does not improve the prediction of individual TF binding sites



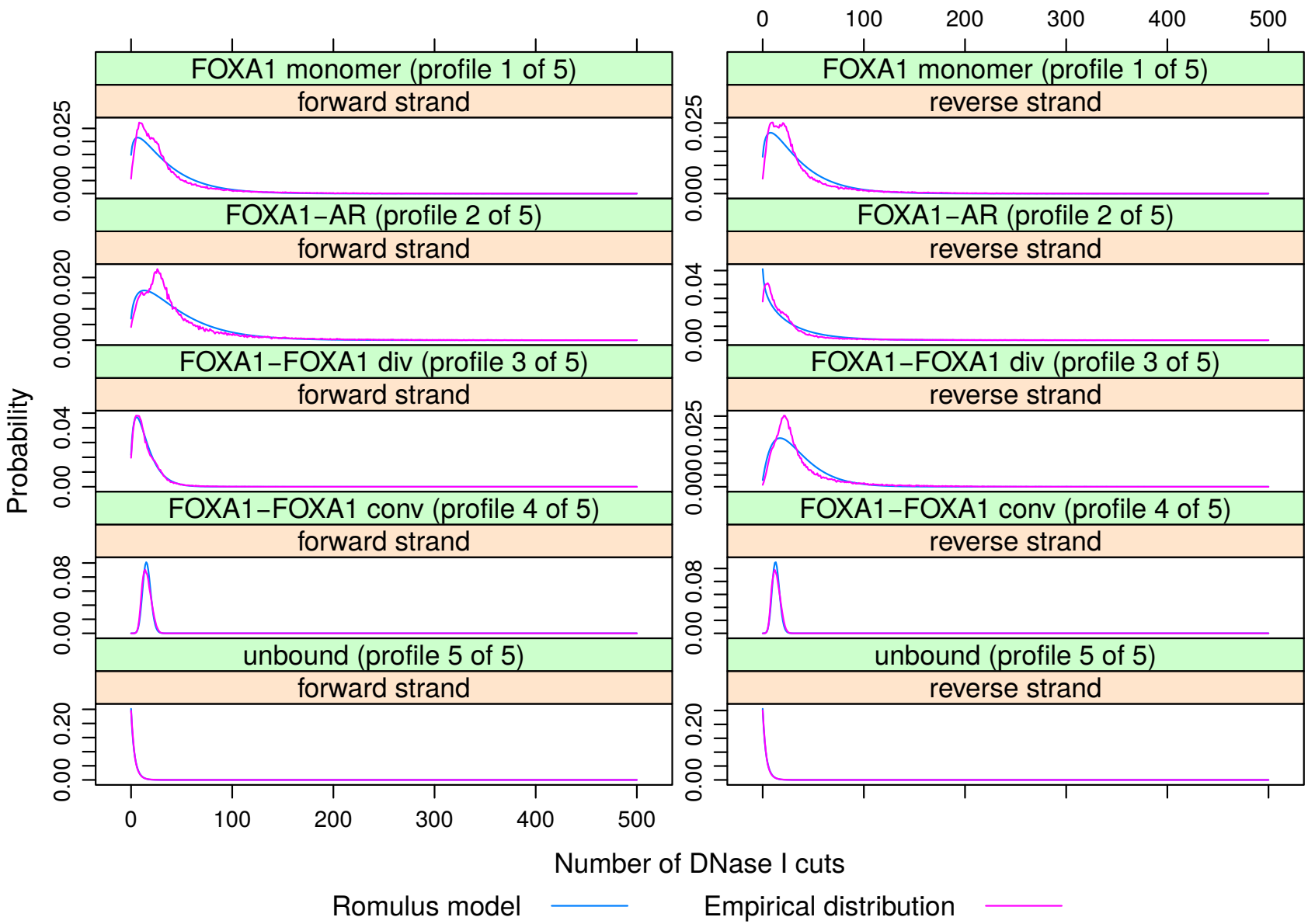
Romulus models differ between the binding modes

yet their inclusion does not improve the prediction of individual TF binding sites



Romulus models differ between the binding modes

yet their inclusion does not improve the prediction of individual TF binding sites



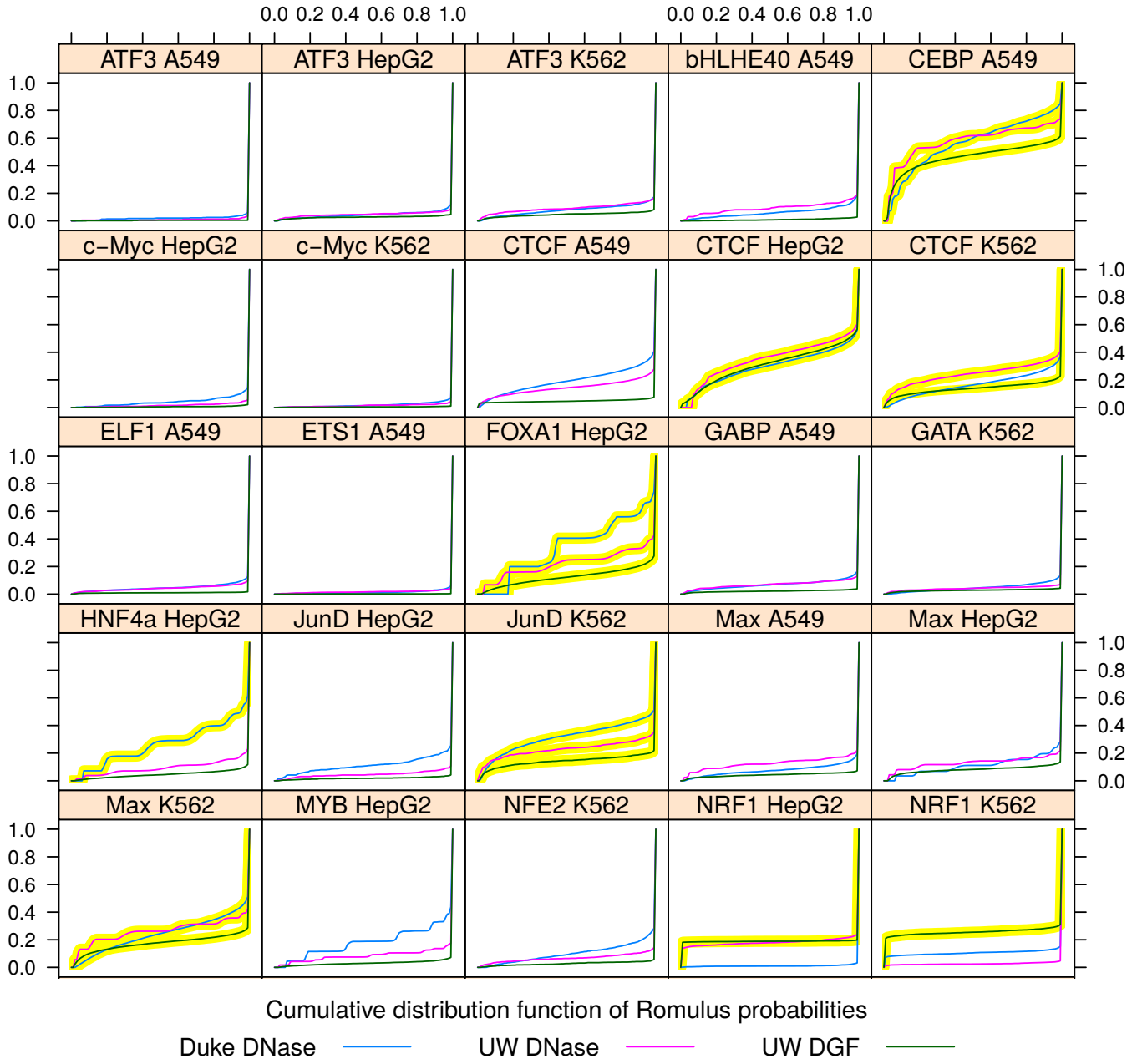
Intermezzo: what makes a good model?

- A mathematical model describes a system using mathematical concepts.
- If the model successfully captures part of the real world, then the model is *realistic*.
- “All models are wrong, but some are useful.” (George E.P. Box)
- A good model is *predictive*, i.e. deals reasonably well with extrapolating into the unknown.
- Even better, an *explanatory* model tells how to intervene with the system to alter the outcome in a desired manner.
- What could we learn from the cases when our model fails?

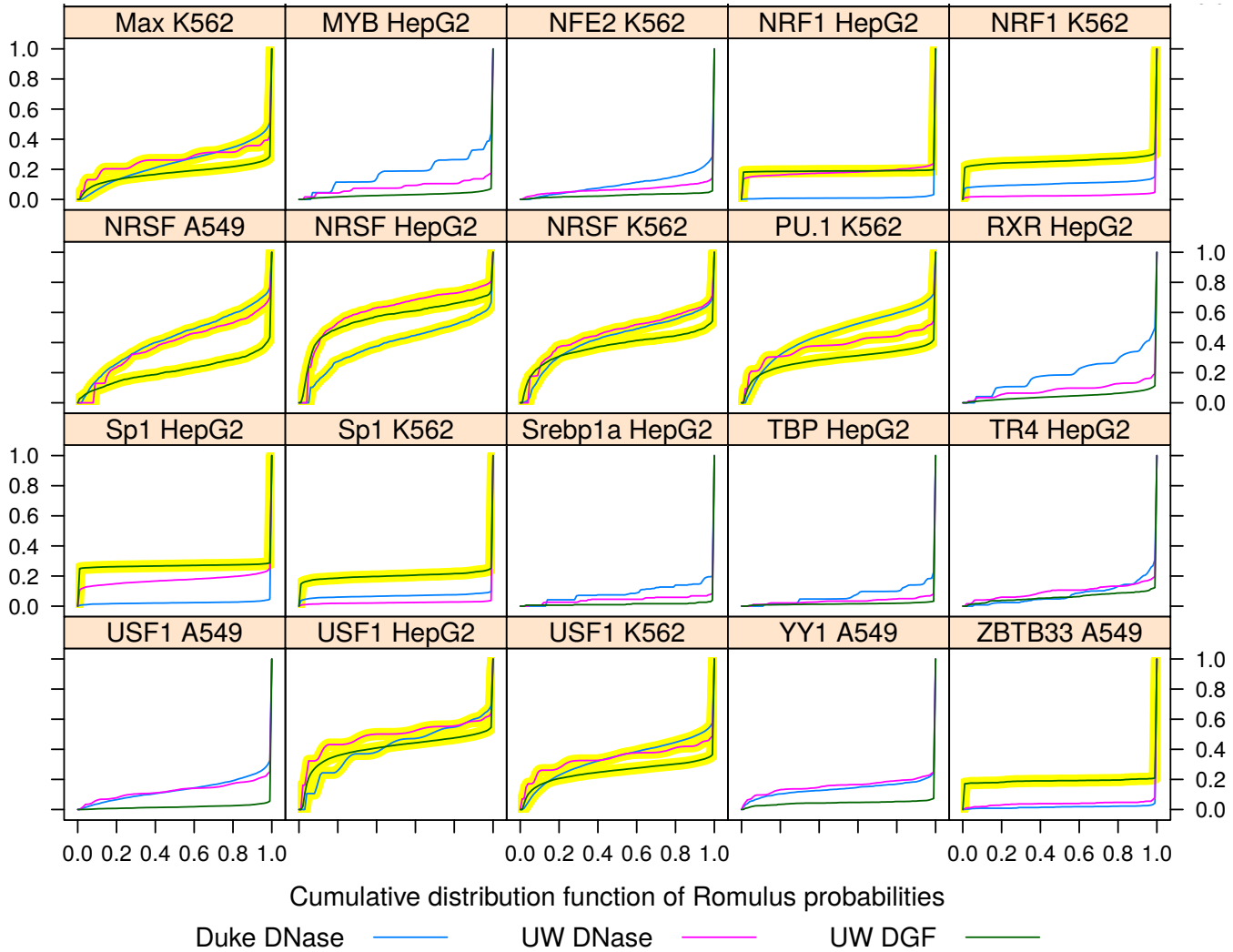
Discrepancies between predictions and actual binding

- Some TFs are able to bind closed chromatin, in violation of the assumptions of Romulus and other algorithms.
- In such a situation of binding to nucleosomal DNA, the way Romulus model accounts for the local chromatin openness profile is not necessarily appropriate.
- To quantify this discrepancy, we limited the scope to the bound motif instances according to the ChIP-seq data, and considered the probabilities of the chromatin state component in the Romulus model.
- We then plotted the cumulative distribution functions of these probabilities.

Cumulative distribution of Romulus chromatin state



Cumulative distribution of Romulus chromatin state

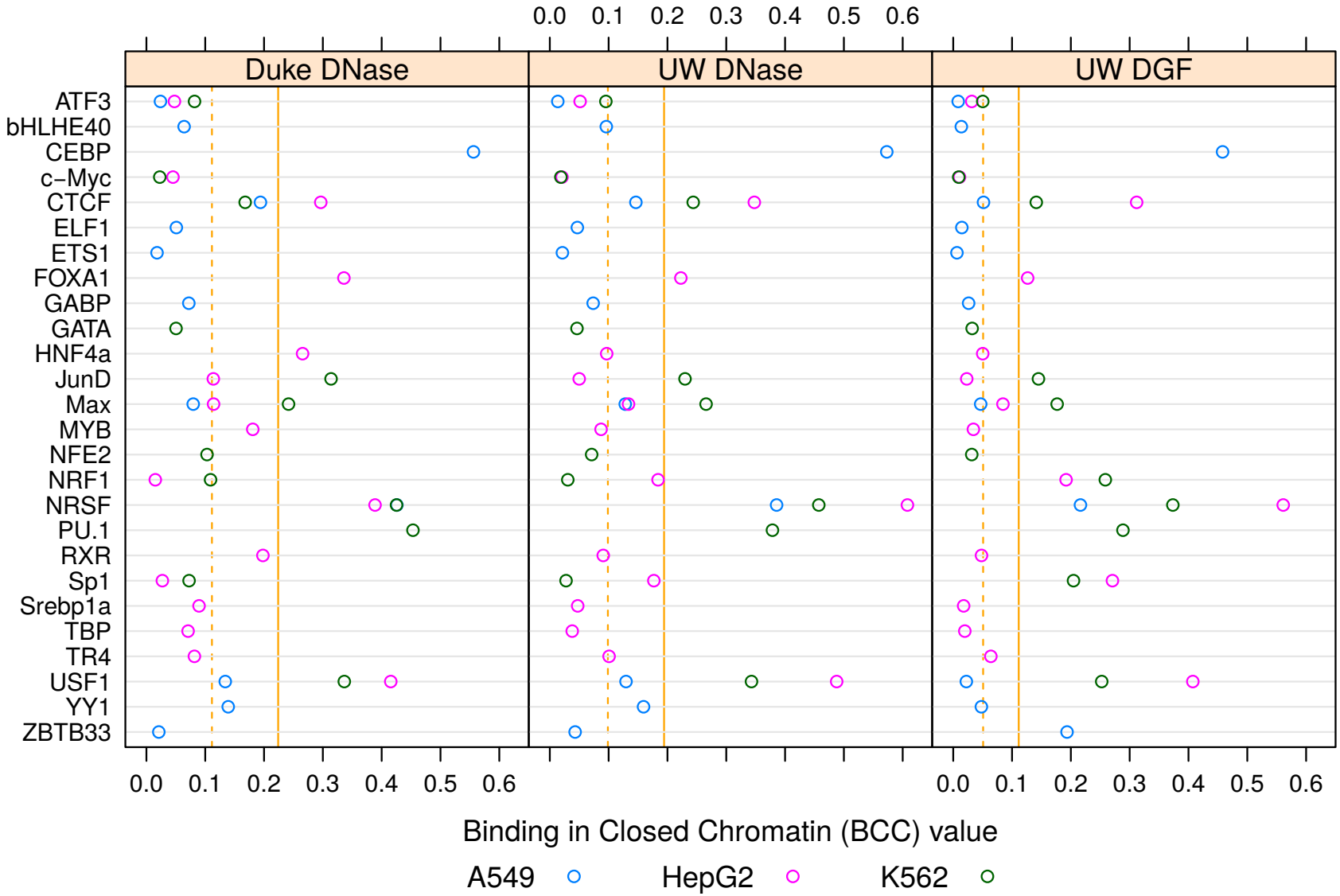


Binding in Closed Chromatin

as a quantitative predictor of pioneer factor activity

- To quantify the amount of TF binding that takes place in loci without a pronounced local chromatin openness signal, we introduce Binding in Closed Chromatin (BCC), as the Area-Under-Curve of the cumulative distribution function described before.
- Note that we take only the chromatin state component, and exclude the prior (genomic sequence) component.
- We focused on the TFs that had a BCC value, in at least one case, more than one MAD (median absolute deviation) above the median.

Binding in Closed Chromatin values



Summary

- Our method, Romulus, combines the benefits of CENTIPEDE and Wellington, and significantly outperforms them, regardless of the DNase-seq protocol used.
- The advantage of Romulus was observed especially when applied to binding site prediction for low-information-content motifs.
- The inclusion of these additional states for the known TF dimers did not yield an increase in predictive power.
- We introduce Binding in Closed Chromatin (BCC) as a quantitative measure of TF pioneer factor activity. Uniquely, this measure quantifies a defining feature of pioneer factors, namely their ability to bind closed chromatin.