MAX-PLANCK-INSTITUTE
FOR BIOPHYSICAL CHEMISTRY

Bayesian Markov models consistently surpass PWMs at predicting regulatory motifs

30 years TRANSFAC

Goettingen
8. March 2018

Johannes Söding, MPI for Biophysical Chemistry, Göttingen

Methods and research topics of the Söding lab

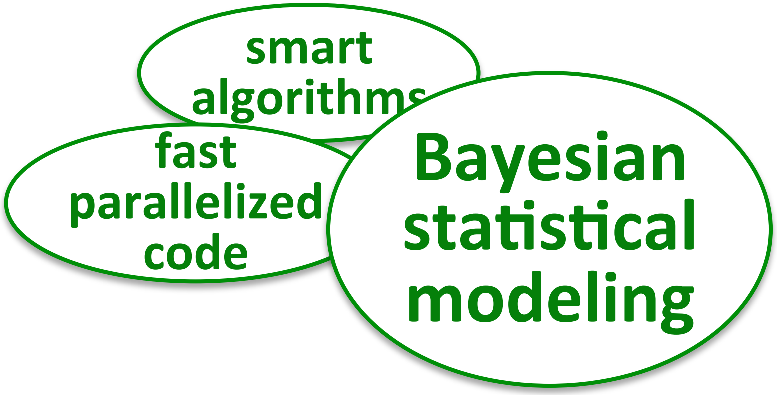
Computational metagenomics

- Fast & deep sequence searching and clustering
- Protein-level assembly
- Large-scale binning & X-assembly
- Gene prediction

↔

Protein function & structure

- Coevolution analysis
- Viral metagenomics (Virus-X)
- Functional module discovery in massive metagenomic data



smart algorithms

fast parallelized code

Bayesian statistical modeling

(Post-) Transcriptional regulation

- Regulatory motif discovery
- RNA-protein binding cooperat.
- PAR-CLIP data analysis
- NGS data analysis

Systems medicine of complex diseases

- Risc locus prediction
- Risk variant fine-mapping
- Discovery of drug targets from GWAS & eQTL data

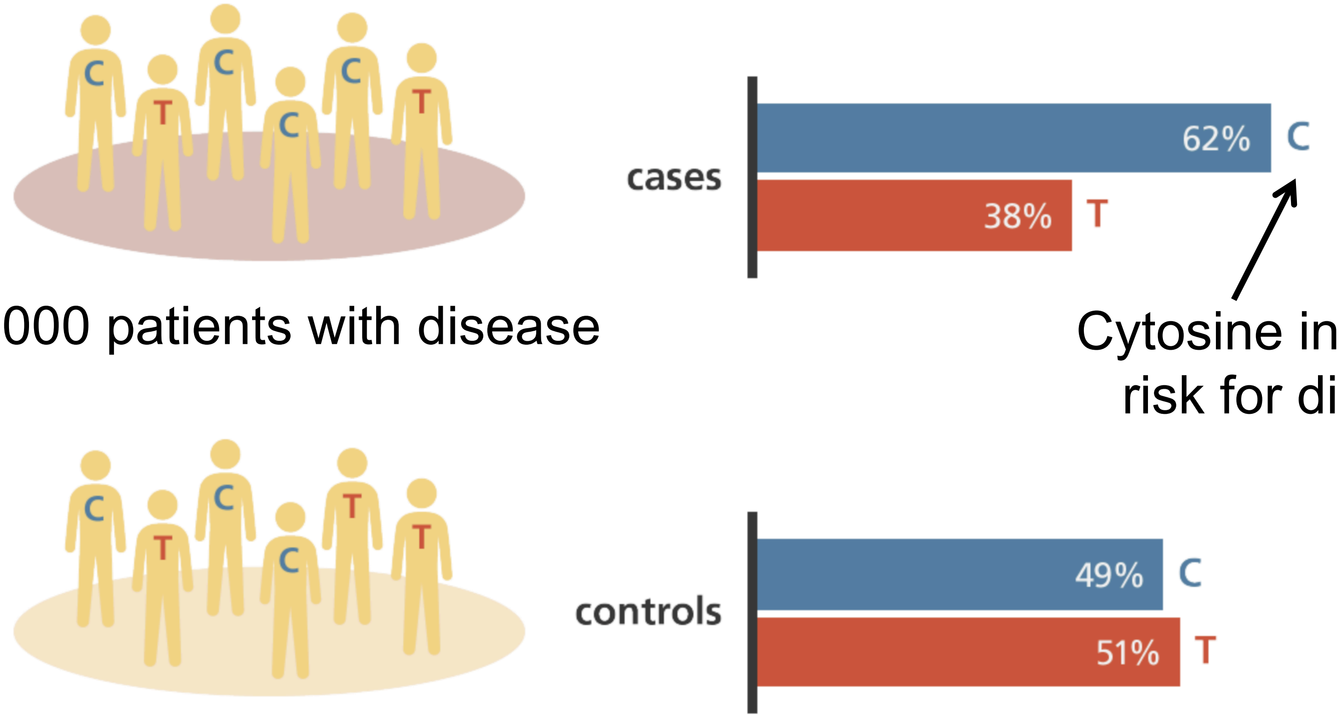
Single-cell transcriptomics

- Reconstruction of cellular lineage trees
- Denoising scRNA-seq data

Why are we interested in transcriptional regulation?

Genetic causes of common diseases linked to dysregulation of gene networks

(SNP = position in genome with variation in population)



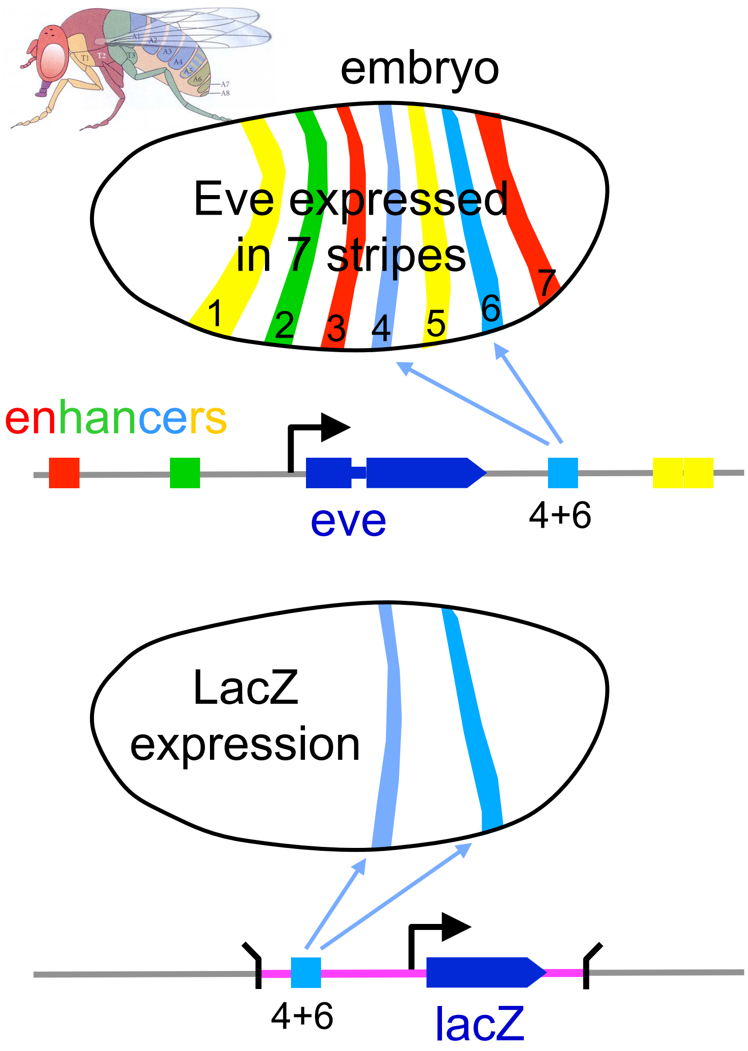
10,000 patients with disease

Cytosine increases risk for disease

10,000 healthy people

- **≥ 90% of causal SNPs non-coding**
- These SNPs **disrupt transcription factor binding sites** and thereby influence the expression of target genes

How is an organism encoded in its genome?

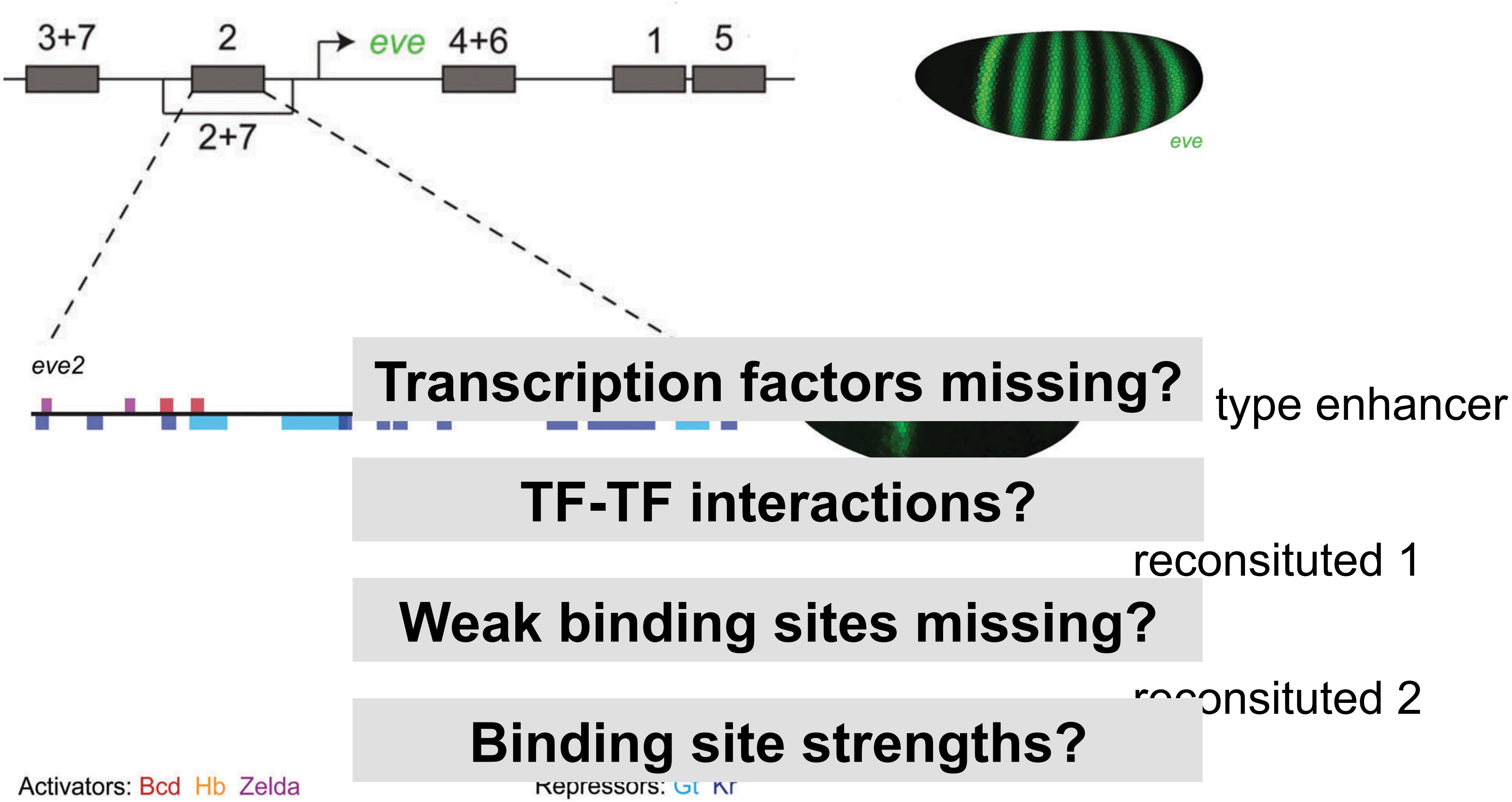


Genomically inserted
4+6 enhancer with lacZ reporter

- Genomes contain all information for a single cell to develop into a complex organism and to survive and procreate
- Genomes are molecular programs, which are read by transcription factors binding to specific DNA sequences.
- Transcription rates are the result of complex molecular computations at promoters and enhancers
- We want to understand and predict these molecular computations

Are we there yet?

(What I cannot create, I do not understand)



Transcription factors missing?

TF-TF interactions?

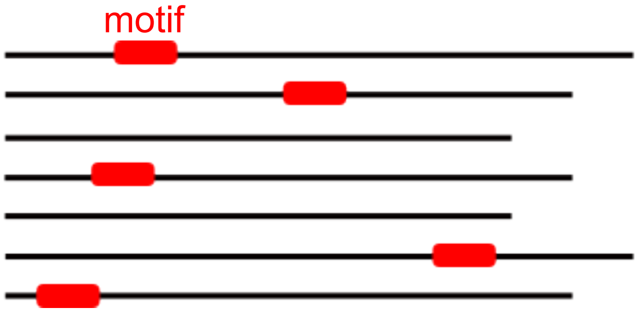
Weak binding sites missing?

Binding site strengths?

We cannot reconstitute even best-studied enhancers with designed sequences

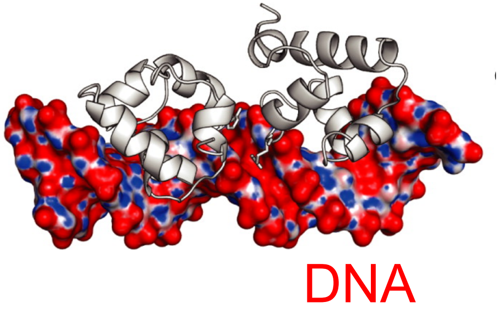
BJ Vincent, J Estrada & AH DePace, Integrative Biol. 2016

De-novo motif discovery: binding site motifs for TFs enriched in set of sequences



DNA

- **ChIP-seq**
- **SELEX-seq**
- **Protein binding microarrays** (PBMs)
- DNase-seq, FAIRE-seq, ATAC-seq: open chromatin
- CAGE, RACE: transcription start sites
- RNA-seq: co-expressed genes
- Hi-C, ChIA-PET: loops and 3D nuclear structure
- ...



RNA

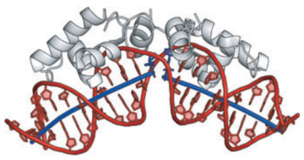
- PAR-CLIP, ICLIP: RNAs bound by RNA-binding factors
- SELEX-seq: binding motifs
- RNA-seq: 3 ‘UTRs of co-expressed genes

Postion weight matrices (PWMs) assume independence of nucleotides within site

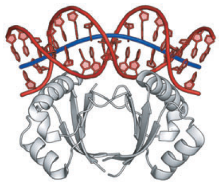
But how important are correlations among nucleotides in regulatory motifs?

Correlations between neighboring nucleotides:

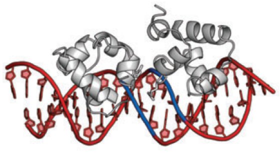
- Shape readout of DNA



bend



kink



minor groove width

- Multiple (sequence-dependent) binding modes
- Variable spacers between half-sites
- Complex combination of motifs at varying distances, e.g. through multiple DNA binding domains, collaborative binding etc.

Diversity and Complexity in DNA Recognition by Transcription Factors

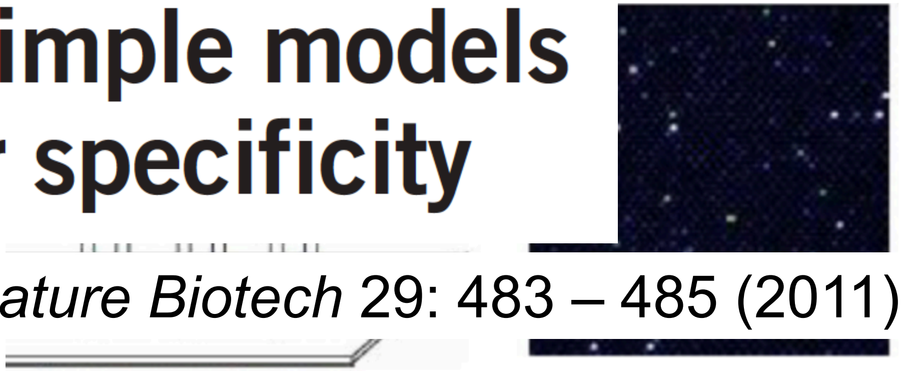
Gwenael Badis,[1]* Michael F. Berger,[2,3]* Anthony A. Philippakis,[2,3,4]* Shaheynoor Talukder,[1,5]* Andrew R. Gehrke,[2]* Savina A. Jaeger,[2]* Esther T. Chan,[5]* Genita Metzler,[6] Anastasia Vedenko,[7] Xiaoyu Chen,[1] Hanna Kuznetsov,[6] Chi-Fong Wang,[8] David Coburn,[1] Daniel E. Newburger,[2] Quaid Morris,[1,5,9,10] Timothy R. Hughes,[1,5,10]† Martha L. Bulyk[2,3,4,11]†

Quantitative analysis demonstrates most transcription factors require only simple models of specificity

Y Zhao & G Stormo, *Nature Biotech* 29: 480 – 483 (2011).

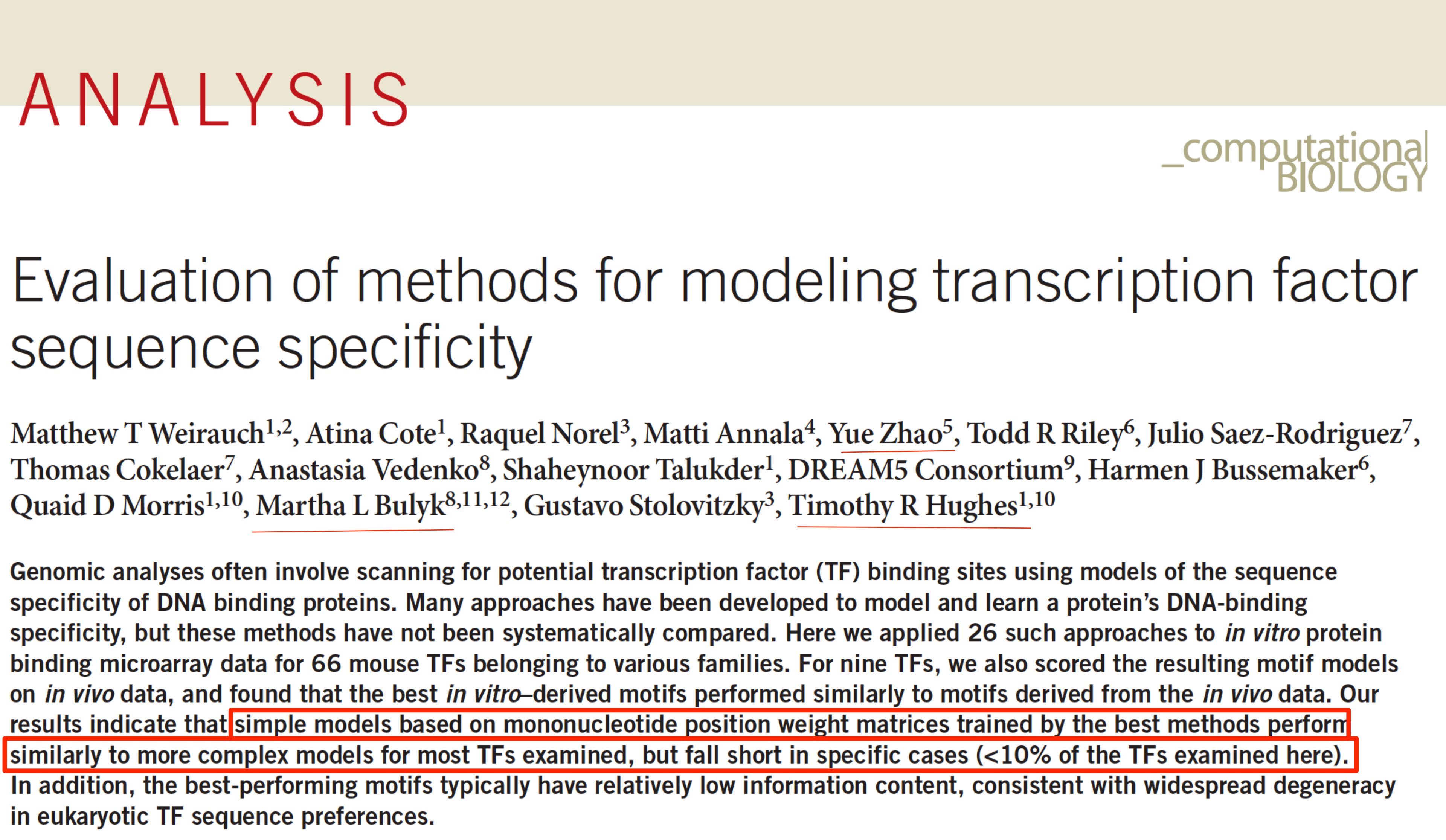
…

Jury remains out on simple models of transcription factor specificity



Q Morris, ML Bulyk, TR Hughes, *Nature Biotech* 29: 483 – 485 (2011).

Protein binding microarray



ANALYSIS

_computational BIOLOGY

Evaluation of methods for modeling transcription factor sequence specificity

Matthew T Weirauch[1,2], Atina Cote[1], Raquel Norel[3], Matti Annala[4], Yue Zhao[5], Todd R Riley[6], Julio Saez-Rodriguez[7], Thomas Cokelaer[7], Anastasia Vedenko[8], Shaheynoor Talukder[1], DREAM5 Consortium[9], Harmen J Bussemaker[6], Quaid D Morris[1,10], Martha L Bulyk[8,11,12], Gustavo Stolovitzky[3], Timothy R Hughes[1,10]

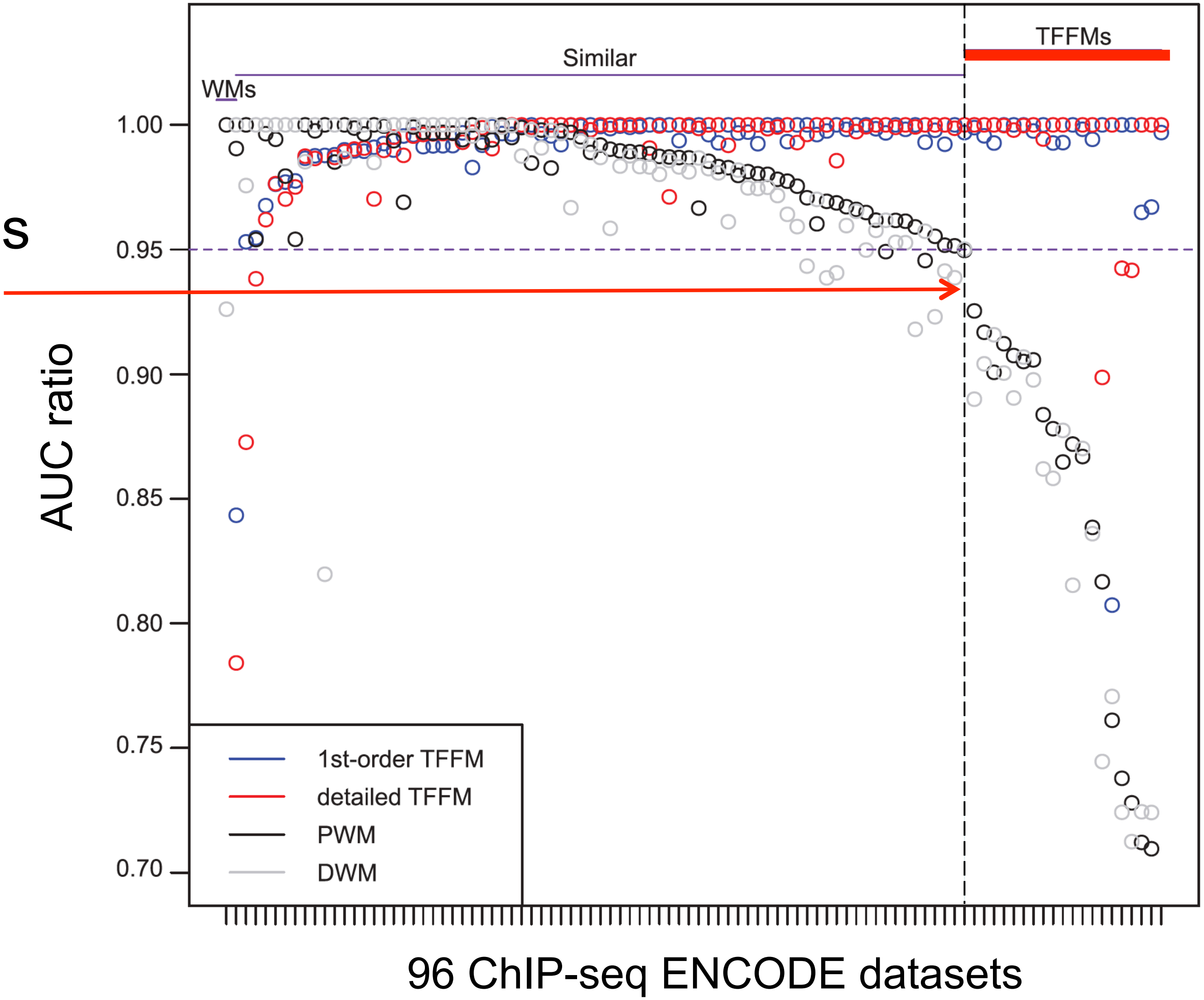
Genomic analyses often involve scanning for potential transcription factor (TF) binding sites using models of the sequence specificity of DNA binding proteins. Many approaches have been developed to model and learn a protein's DNA-binding specificity, but these methods have not been systematically compared. Here we applied 26 such approaches to *in vitro* protein binding microarray data for 66 mouse TFs belonging to various families. For nine TFs, we also scored the resulting motif models on *in vivo* data, and found that the best *in vitro*–derived motifs performed similarly to motifs derived from the *in vivo* data. Our results indicate that simple models based on mononucleotide position weight matrices trained by the best methods perform similarly to more complex models for most TFs examined, but fall short in specific cases (<10% of the TFs examined here). In addition, the best-performing motifs typically have relatively low information content, consistent with widespread degeneracy in eukaryotic TF sequence preferences.

Weirauch et al., Nature Biotechnology (2013)

The Next Generation of Transcription Factor Binding Site Prediction

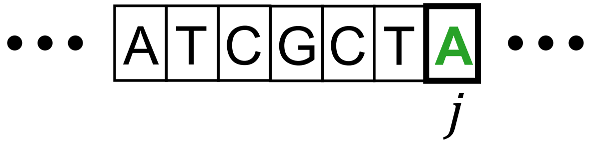
Anthony Mathelier*, **Wyeth W. Wasserman*** *PLoS Comput Biol* 9, e1003214 (2013)

TFFMs (1st order Markov models) perform >5% better than PWMs in 20% of cases



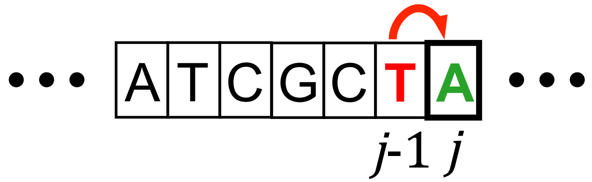
Markov Models (MMs)
model correlations among nucleotides

k'th order MM: probability depends on *k* previous nucleotides



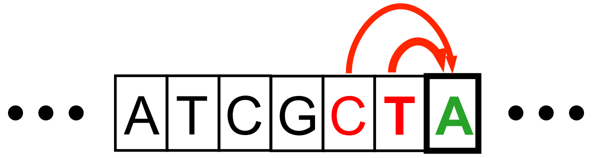
$$\mathrm{Score}(x_1 \dots x_W) = \sum_{j=1}^{W} \log \frac{p_j(x_j)}{p_{\mathrm{bg}}(x_j)}$$

0th order, PWM



$$\mathrm{Score}(x_1 \dots x_W) = \sum_{j=1}^{W} \log \frac{p_j(x_j | x_{j-1})}{p_{\mathrm{bg}}(x_j | x_{j-1})}$$

1st order

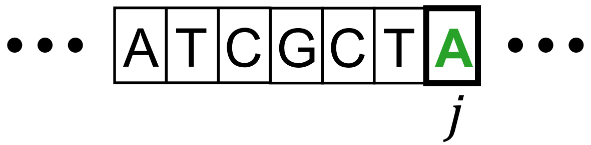


$$\mathrm{Score}(x_1 \dots x_W) = \sum_{j=1}^{W} \log \frac{p_j(x_j | x_{j-1}, x_{j-2})}{p_{\mathrm{bg}}(x_j | x_{j-1}, x_{j-2})}$$

$\vdots$

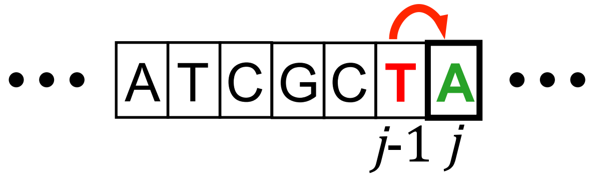
Markov Models (MMs) model correlations among nucleotides

k’th order MM: probability depends on *k* previous nucleotides



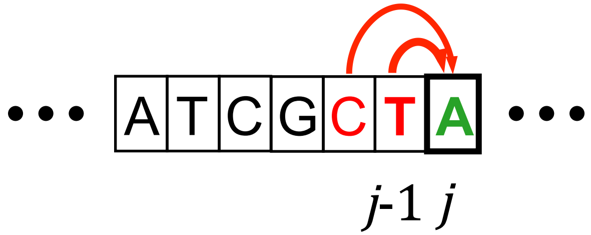
$$p_j(A) = \frac{\#(A)}{\#\ \text{seqs}}$$

0th order, PWM



$$p_j(A|T) = \frac{\#(TA)}{\#(T)}$$

1st order



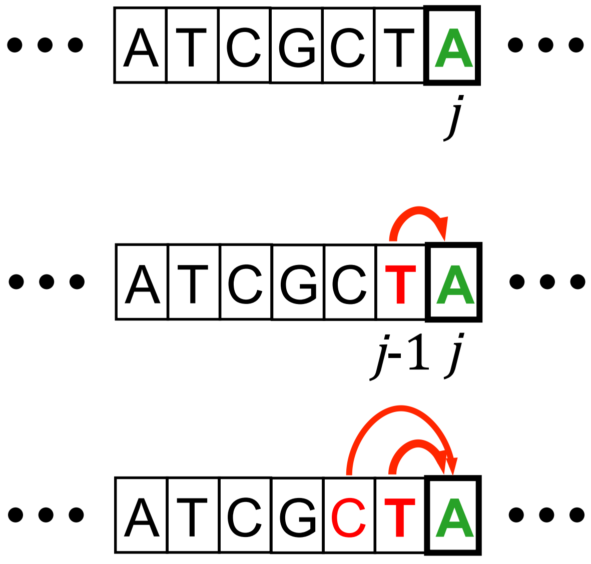
$$p_j(A|CT) = \frac{\#(CTA)}{\#(CT)}$$

2nd order

⋮

Markov Models (MMs) model correlations among nucleotides

k'th order MM: probability depends on *k* previous nucleotides



• • • A T C G C T A • • •
j
• • • A T C G C T A • • •
j-1 j
• • • A T C G C T A • • •

$$p_j(\mathbf{A}) = \frac{\#(\mathbf{A}) + 1}{\#\text{ seqs} + 4}$$

0th order, PWM

Pseudo-counts

$$p_j(\mathbf{A}|\mathbf{T}) = \frac{\#(\mathbf{TA}) + 1}{\#(\mathbf{T}) + 4}$$

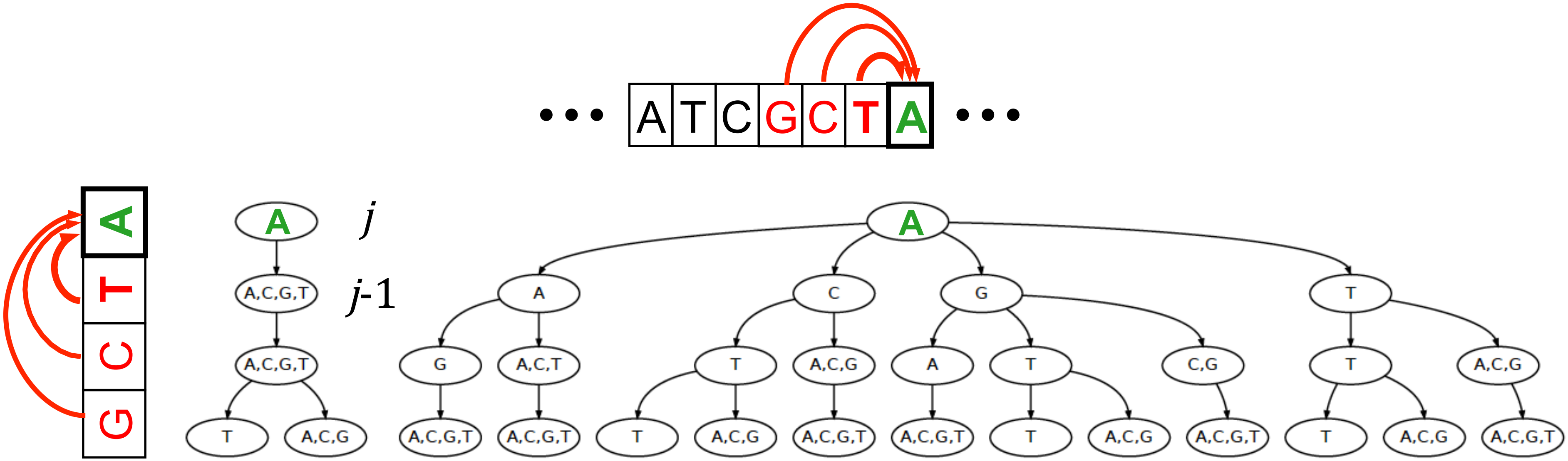
1st order

$$p_j(\mathbf{A}|\mathrm{CT}) = \frac{\#(\mathrm{C}\mathbf{TA}) + 1}{\#(\mathrm{CT}) + 4}$$

2nd order

For order *k* one needs **~100 ×4^{k+1}** sequences (!) to learn probabilities with 10% relative accuracy

Many higher-order models prune the dependency graph and pool contexts



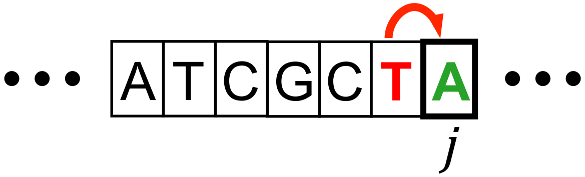
A	0.0307	0.2080	0.0147	0.1679	0.0005	0.9664	0.0214	0.0076	0.0002	0.9829	0.0278	0.0003	0.1435	0.1014
C	0.8782	0.0063	0.0147	0.8243	0.9985	0.0112	0.0214	0.0076	0.9994	0.0057	0.9166	0.9278	0.0017	0.0519
G	0.0307	0.4097	0.9559	0.0039	0.0005	0.0112	0.9357	0.0076	0.0002	0.0057	0.0278	0.0716	0.4274	0.3491
T	0.0604	0.3760	0.0147	0.0039	0.0005	0.0112	0.0214	0.9772	0.0002	0.0057	0.0278	0.0003	0.4274	0.4977

inhomogeneous
parsimonious
Markov model

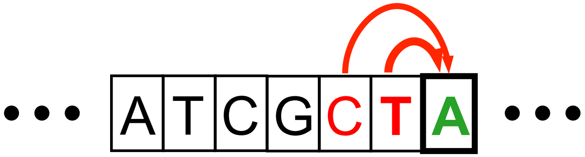
inhomogeneous variable-order Markov model

Optimization requires comparing very many discrete tree topologies
⇒ Slow and challenging to train (model comparison)
⇒ Cannot discover motifs de-novo, require pre-aligned motif sequences

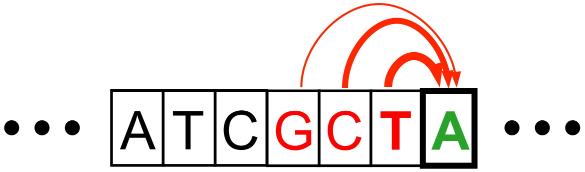
We use pseudocounts from lower-order!



$$p_j(A|T) = \frac{\#(TA) + 20\, p_j(A)}{\#(T) + 20}$$



$$p_j(A|CT) = \frac{\#(CTA) + 60\, p_j(A|T)}{\#(CT) + 60}$$



$$p_j(A|GCT) = \frac{\#(GCTA) + 180\, p_j(A|CT)}{\#(GCT) + 180}$$

$$\vdots$$

Pseudo-counts

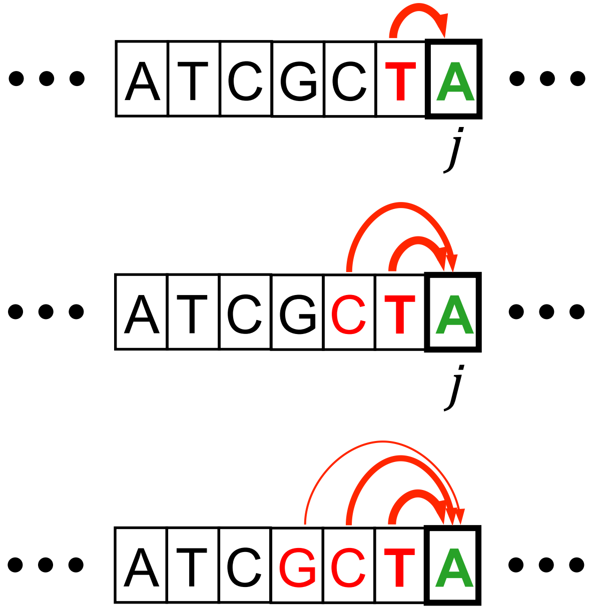
Siebert and Soding, NAR 2016

We use pseudocounts from lower-order!

If many counts for k-mer

⇒ counts dominate over pseudocounts

⇒ use maximum likelihood estimate



$$p_j(\mathbf{A}|\mathbf{T}) = \frac{400 + 20\, p_j(\mathbf{A})}{500 + 20}$$

$$p_j(\mathbf{A}|\mathbf{CT}) = \frac{360 + 60 \times 0.8}{400 + 60}$$

$$p_j(\mathbf{A}|\mathrm{G}\mathbf{CT}) = \frac{340 + 180 \times 0.9}{350 + 180}$$

$$\vdots$$

Pseudo-counts

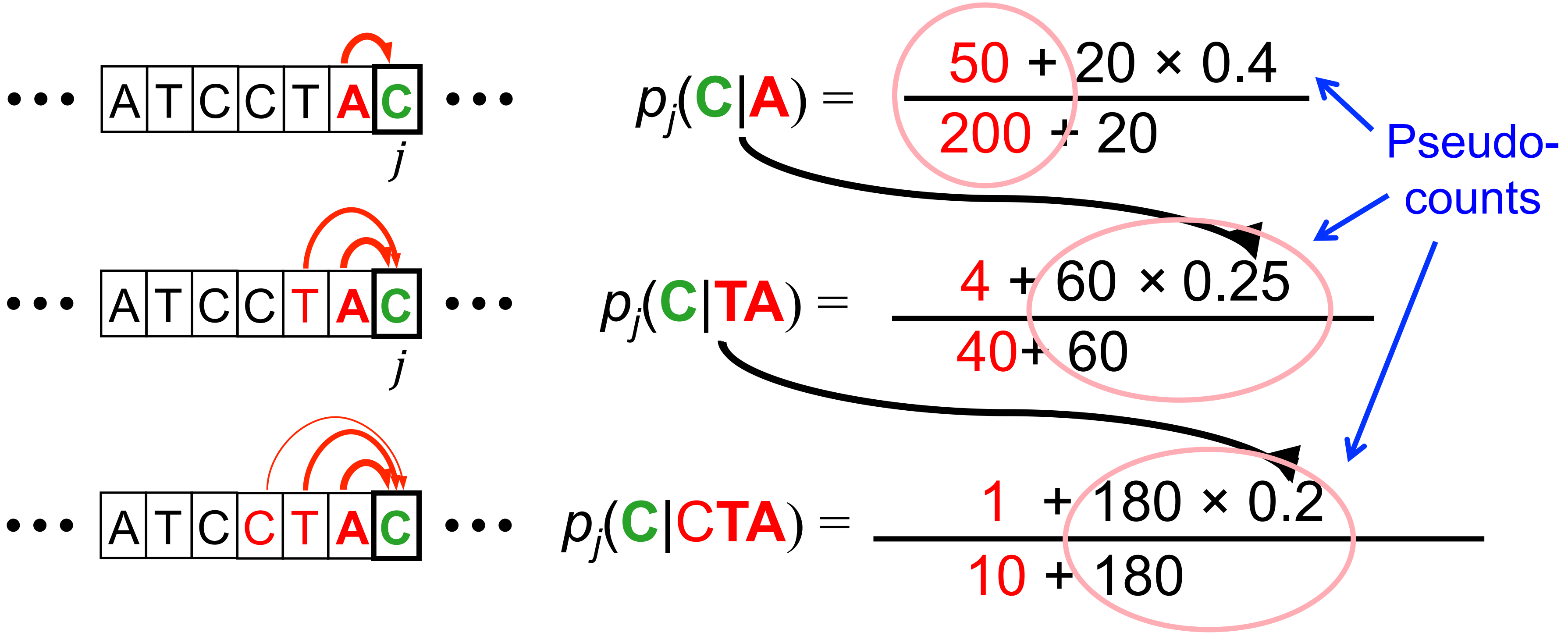
Siebert and Soding, NAR 2016

We use pseudocounts from lower-order!

If few counts for k-mer

⇒ pseudocounts dominate over counts

⇒ fall back on lower-order estimate



$$p_j(\mathbf{C}|\mathbf{A}) = \frac{50 + 20 \times 0.4}{200 + 20}$$

$$p_j(\mathbf{C}|\mathbf{TA}) = \frac{4 + 60 \times 0.25}{40 + 60}$$

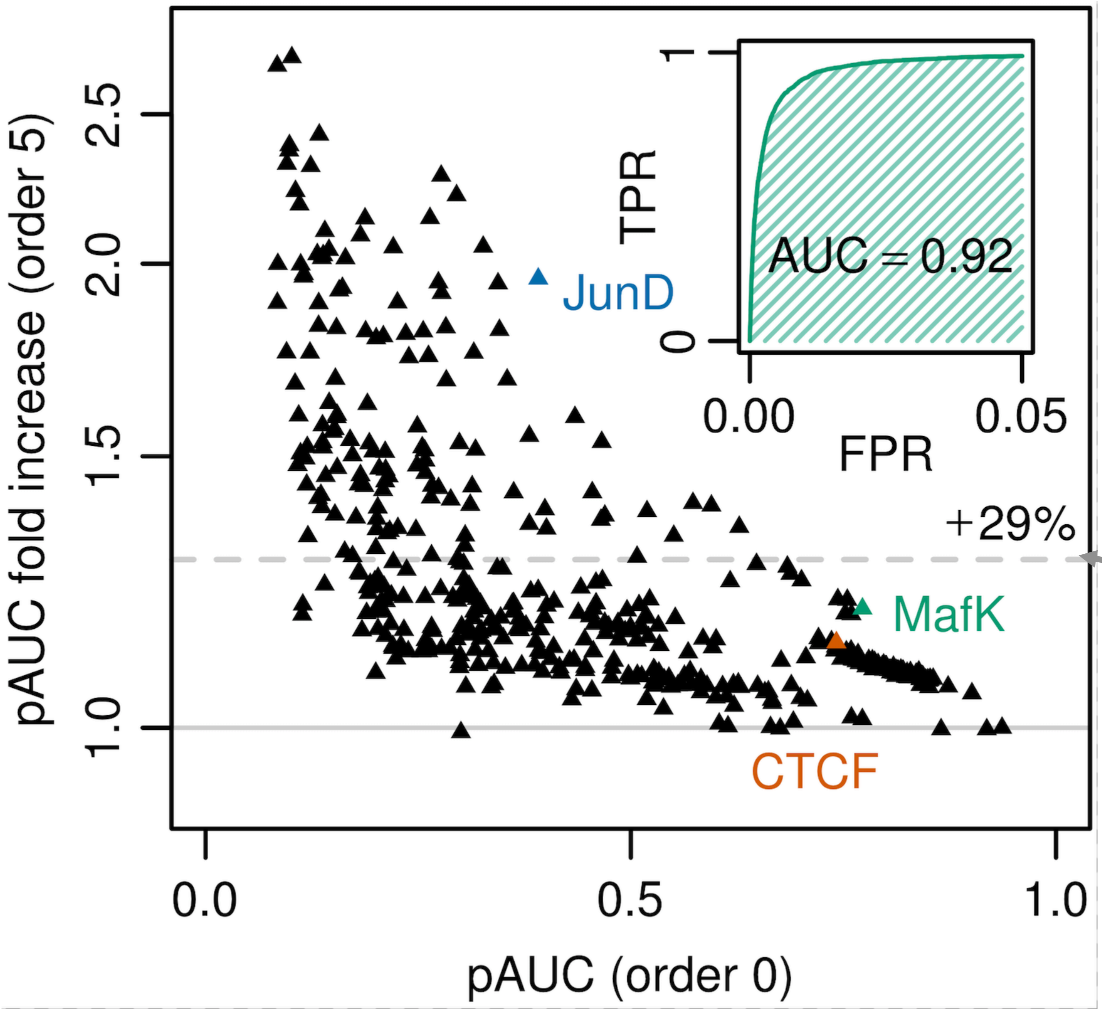
$$p_j(\mathbf{C}|\mathbf{CTA}) = \frac{1 + 180 \times 0.2}{10 + 180}$$

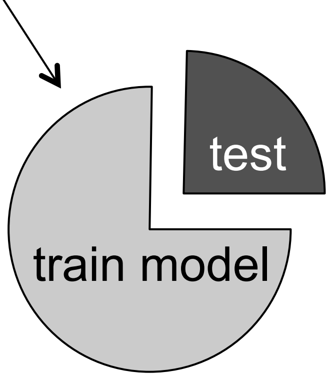
Bayesian Markov models **only learn parameters for which enough information exists to estimate accurately!**

No need for optimizing discrete dependency trees.

5th order BaMMs learn binding motifs from ChIP-seq better than PWMs

Increase of partial Area under ROC curve at 5% FPs (pAUC)
for each of 446 ENCODE ChIP-seq datasets (4-fold cross-validated)



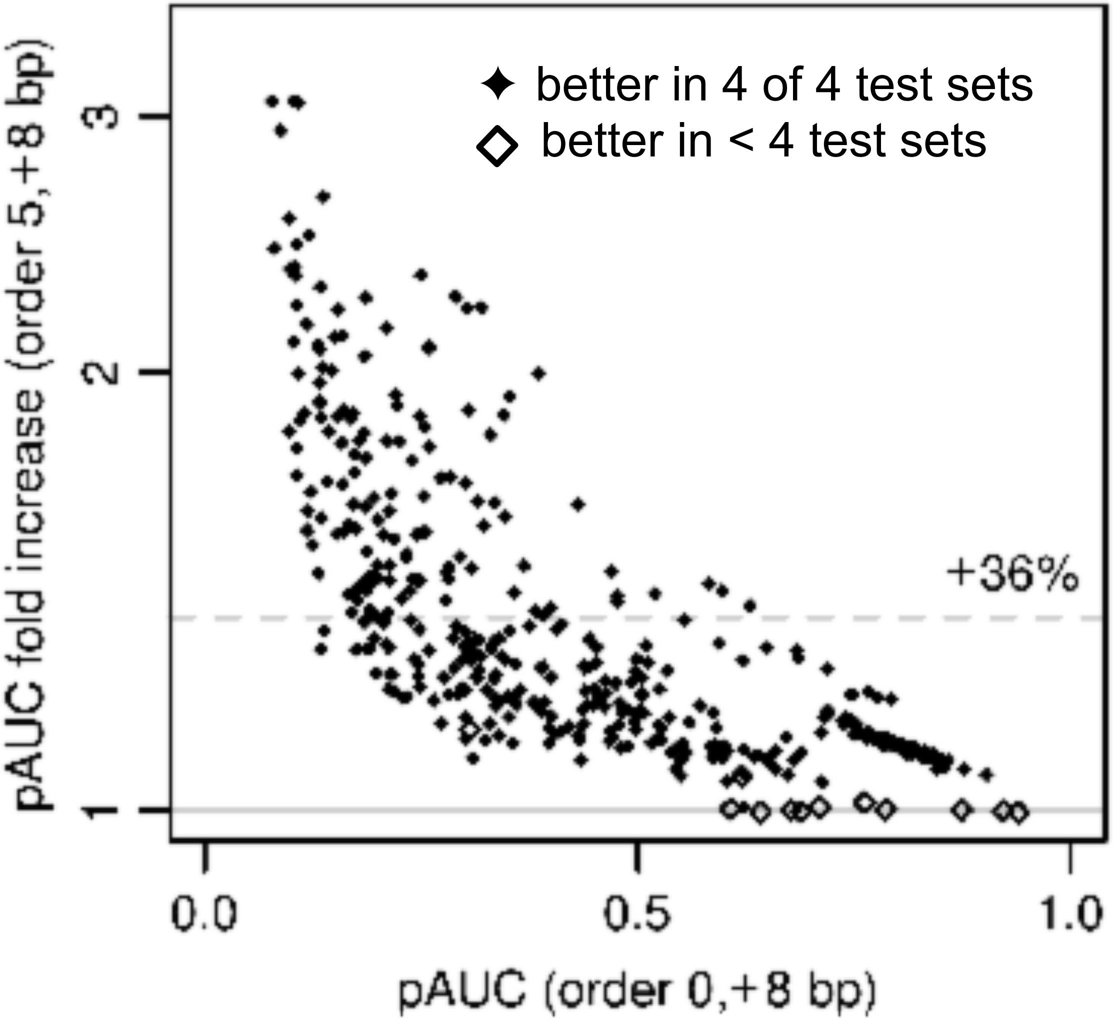


100:1 neg:pos,
neg. sequences from
2nd order hom. model

average pAUC
improvement

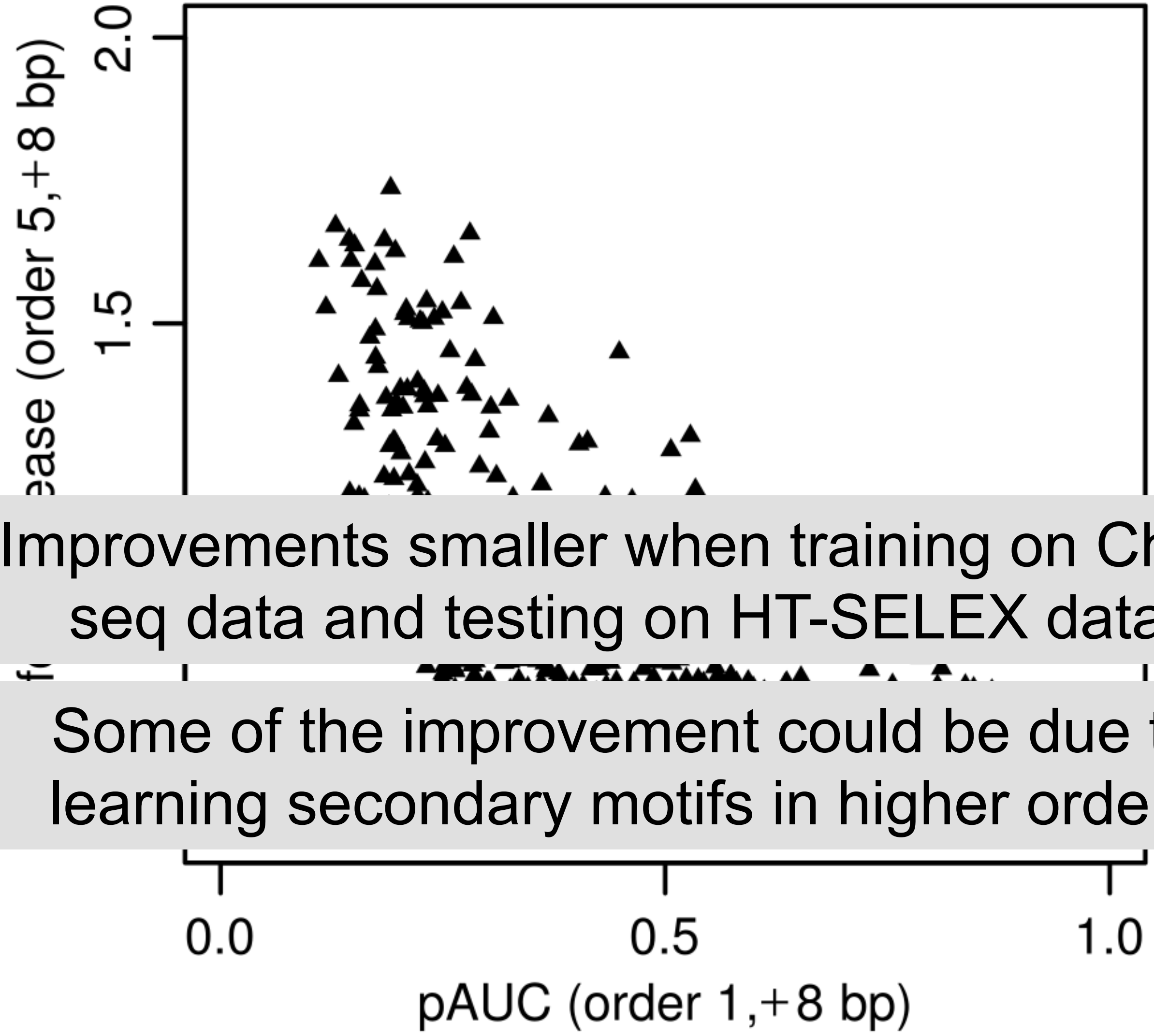
Gains of 5th order BaMMs over PWMs grow when including flanking nucleotides

Increase of pAUC on 446 ENCODE ChIP-seq sets for +8bp-extended models



5th order BaMMs achieve sizeable gains even over 1st order BaMMs

Increase of pAUC on 446 ENCODE ChIP-seq sets for +8bp-extended models

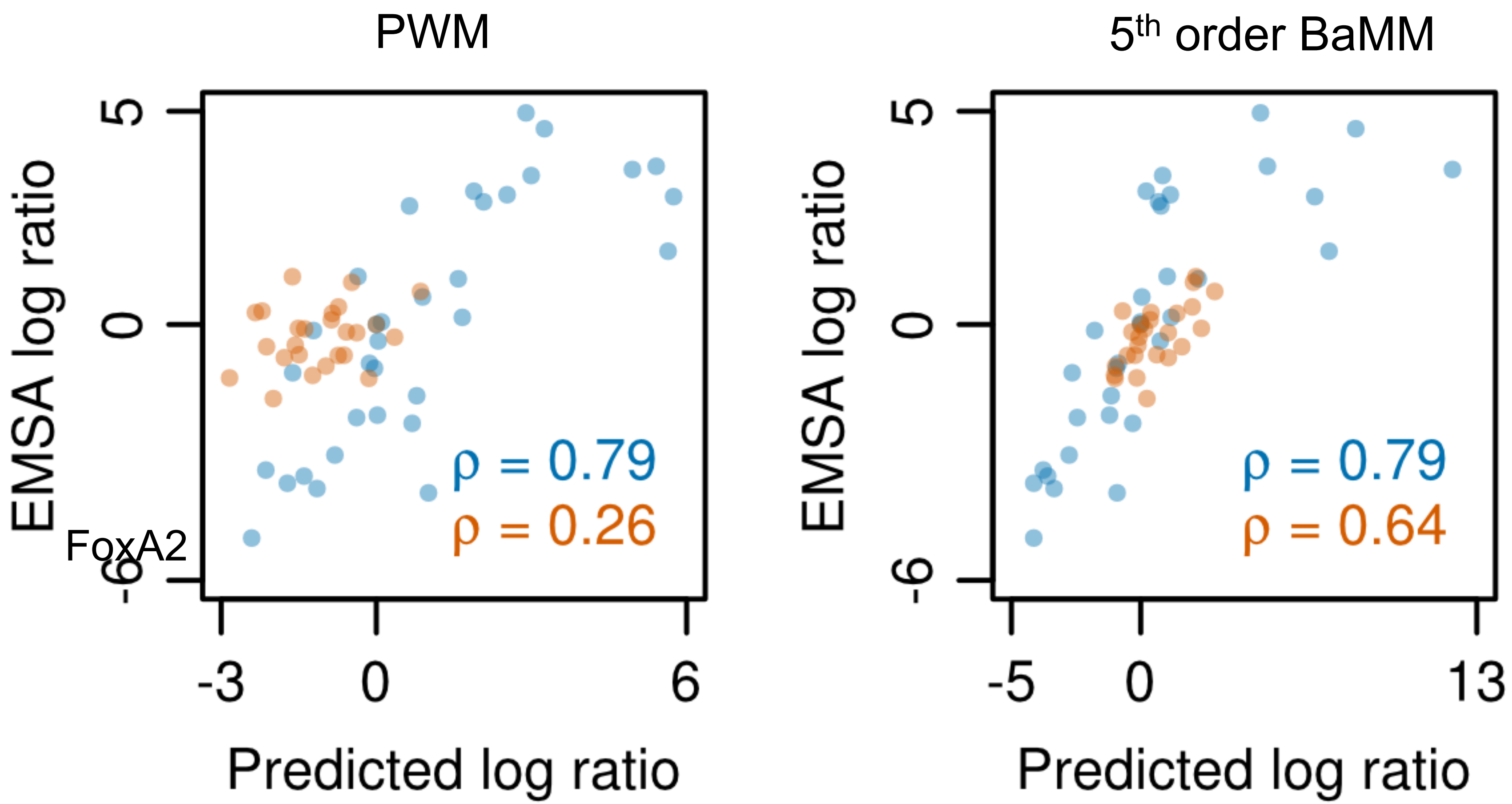


Improvements smaller when training on ChIP-seq data and testing on HT-SELEX data!

Some of the improvement could be due to learning secondary motifs in higher orders

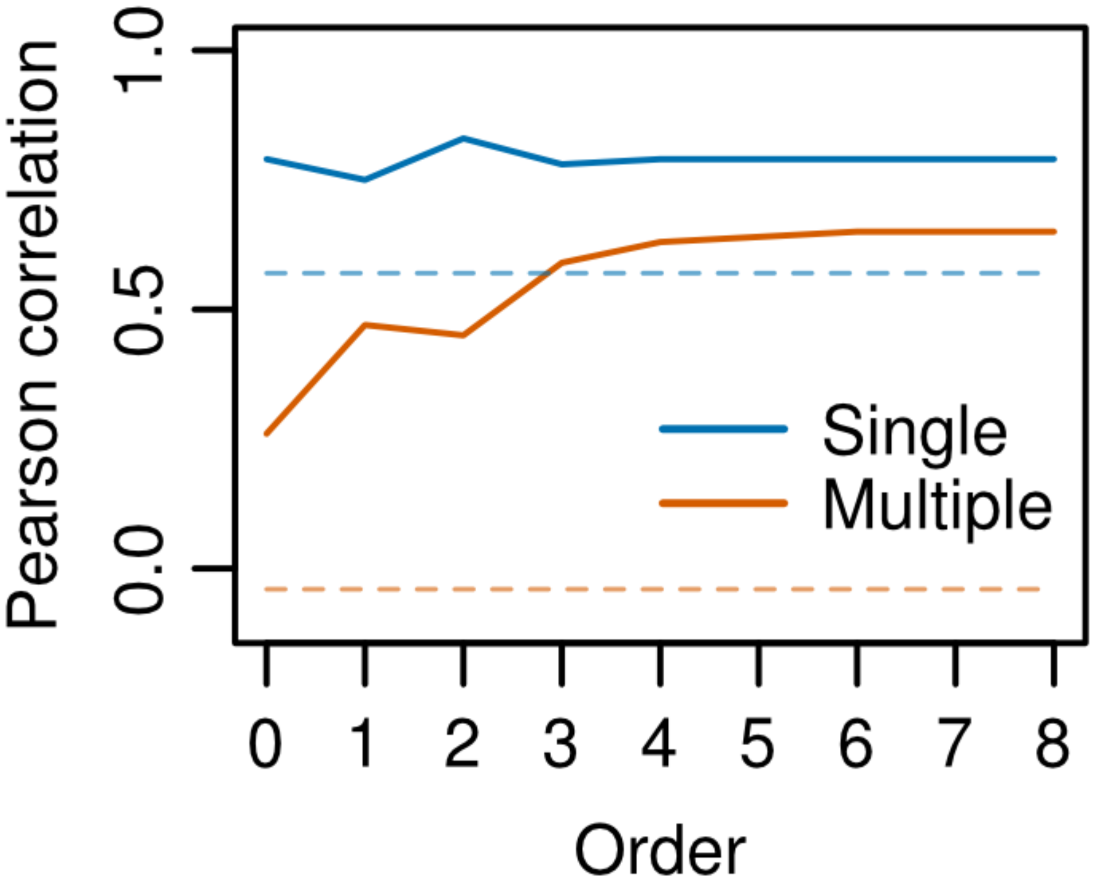
Klf4 motifs trained on ChIP-seq, tested on EMSA affinities of mutated binding sites

single mutation from consensus
double mutation from consensus

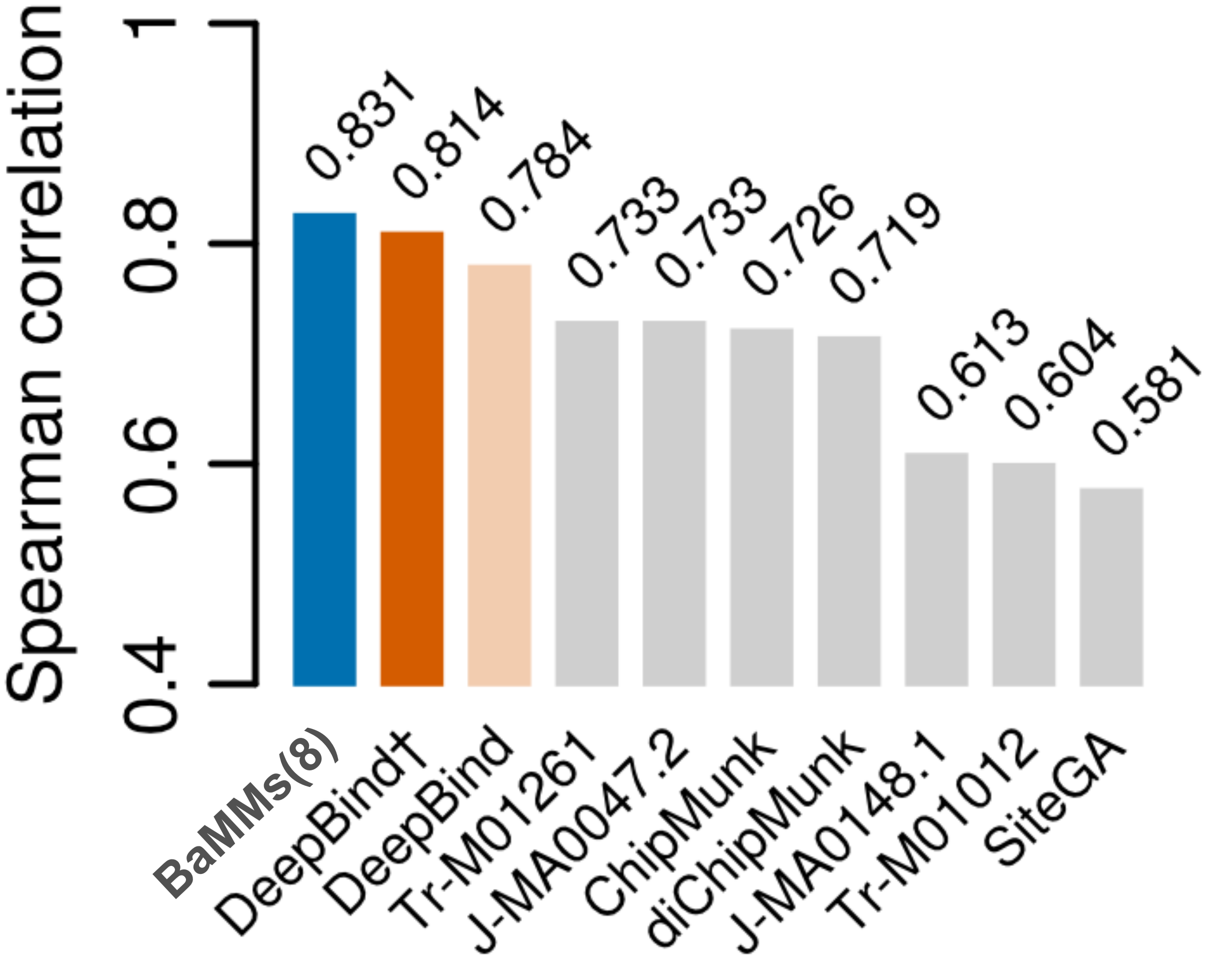


Klf4 motifs trained on ChIP-seq, tested on EMSA affinities of mutated binding sites

single mutation from consensus
double mutation from consensus

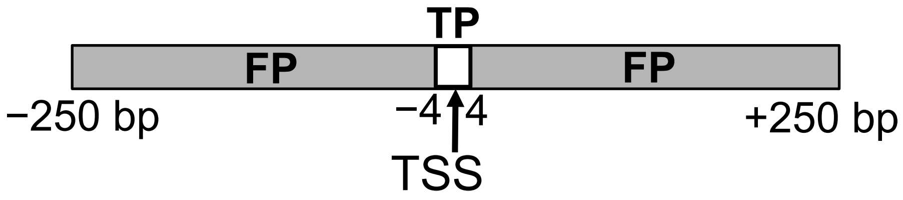


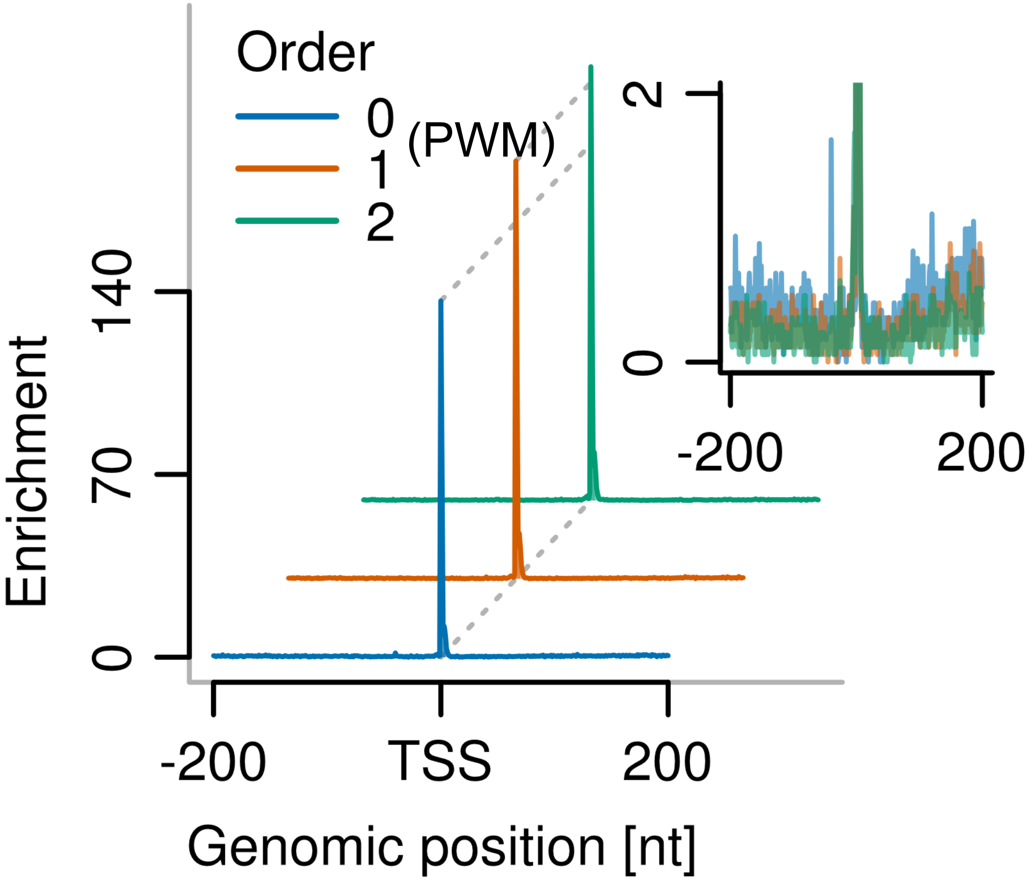
FoxA2 motifs trained on ChIP-seq tested on EMSA affinities of mutated binding sites

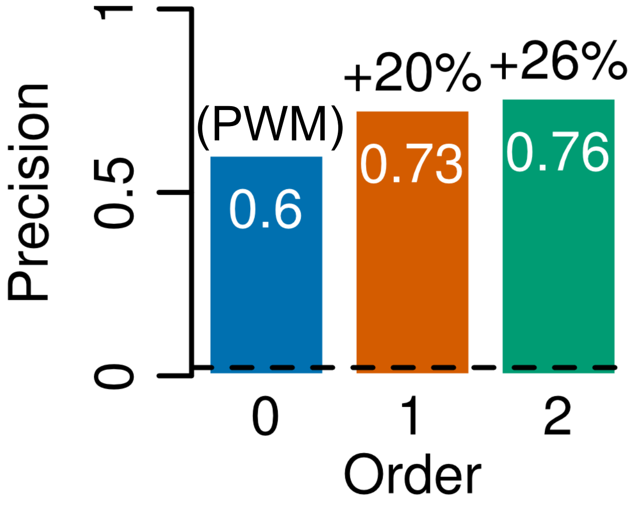


Detecting fly narrow-peak TSSs

(CAGE data from Adelman lab)

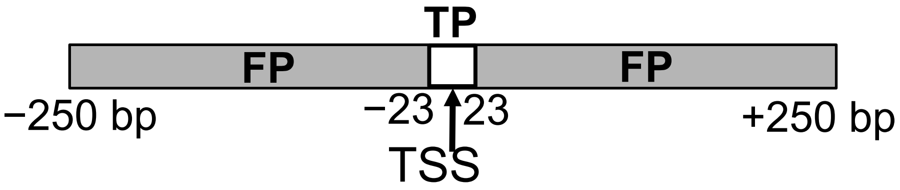






Detecting fly broad-peak TSSs

(CAGE data from Adelman lab)

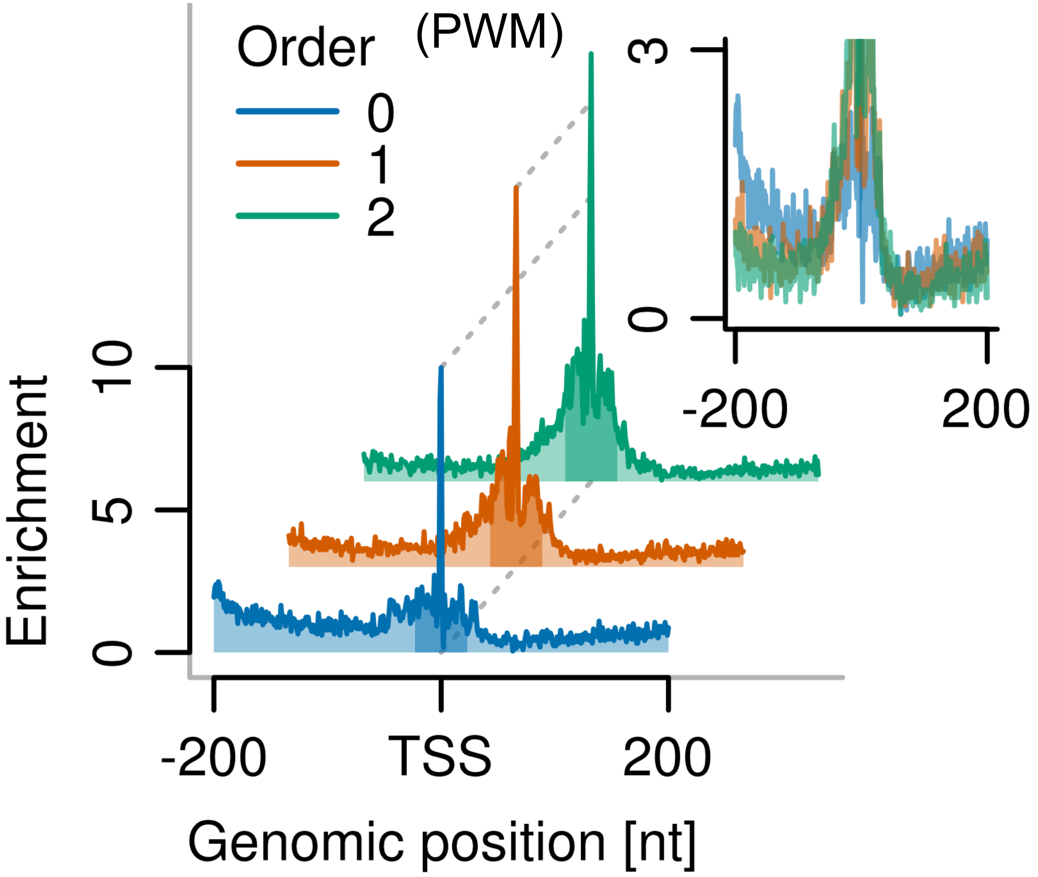


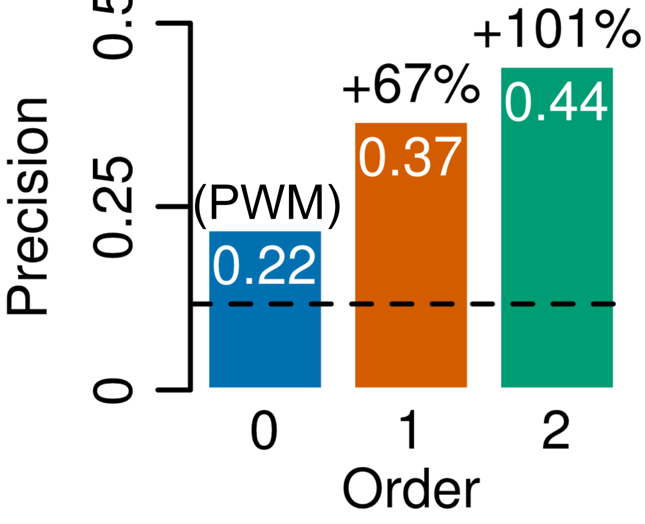
TP

FP FP

−250 bp −23 23 +250 bp

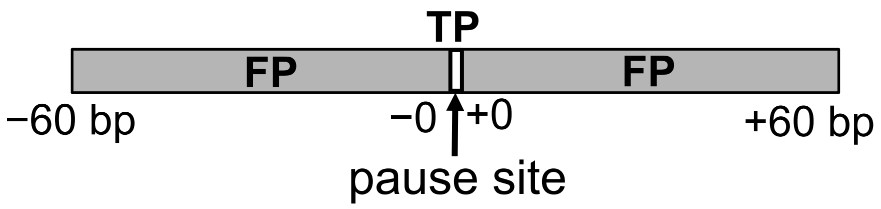
TSS

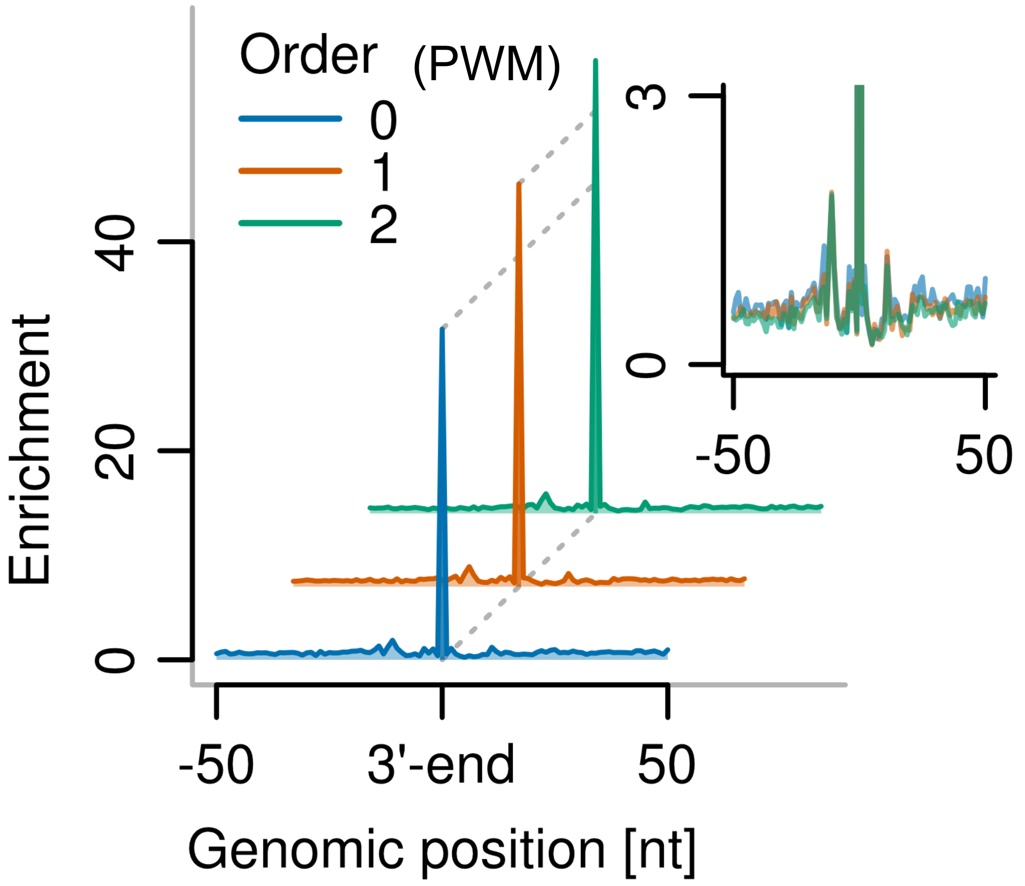


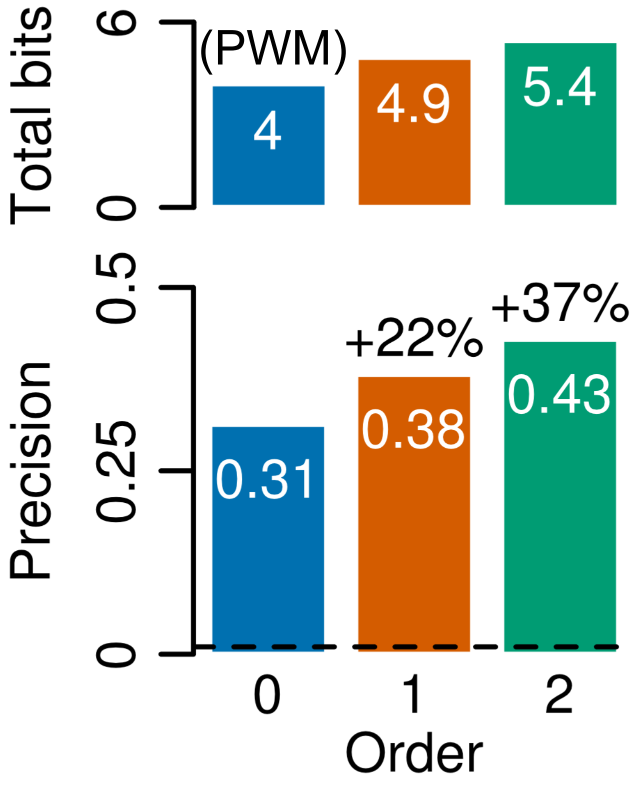


Pause sites in E. coli

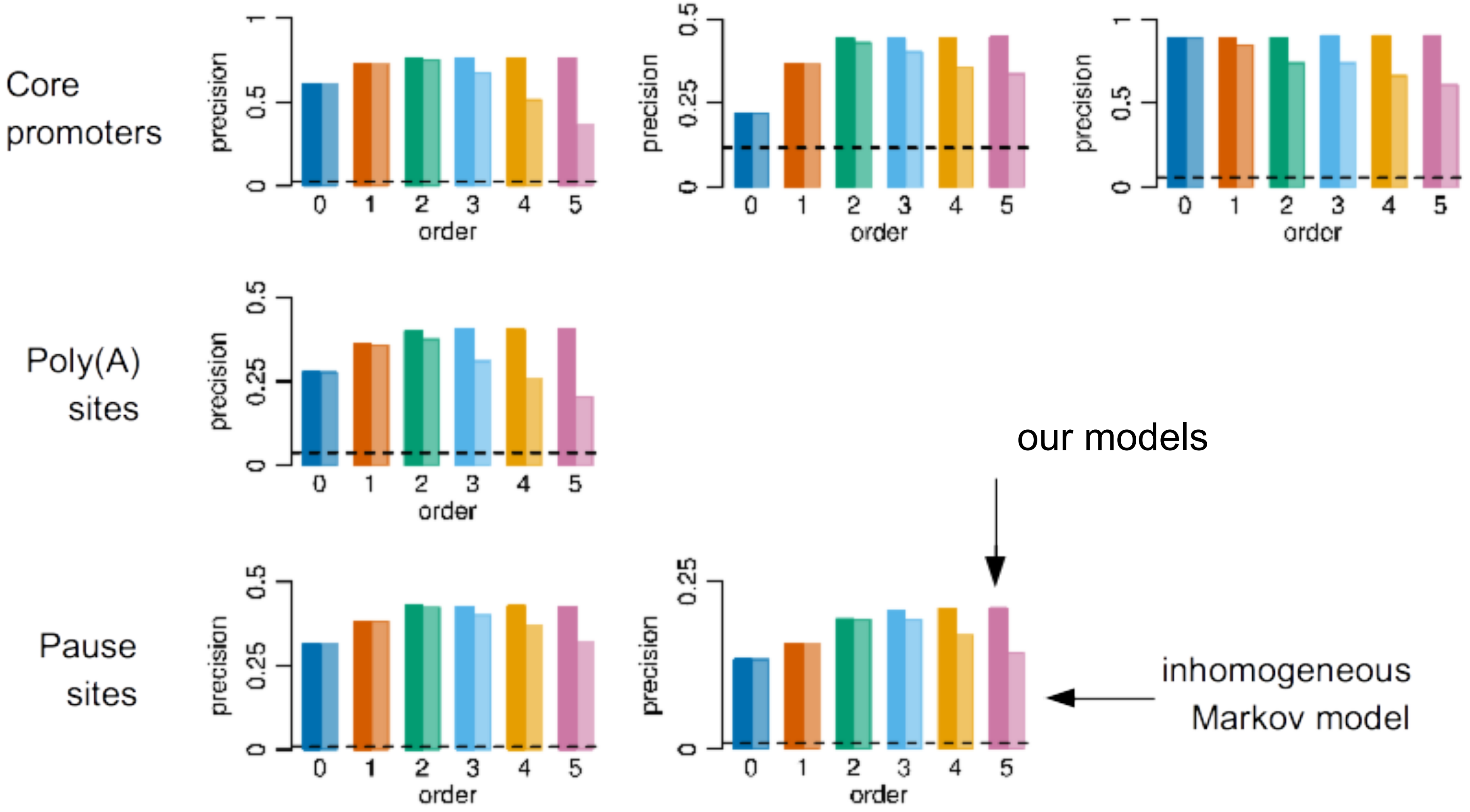
(data from Landick lab)







BaMMs are robust to overtraining



Core promoters

Poly(A) sites

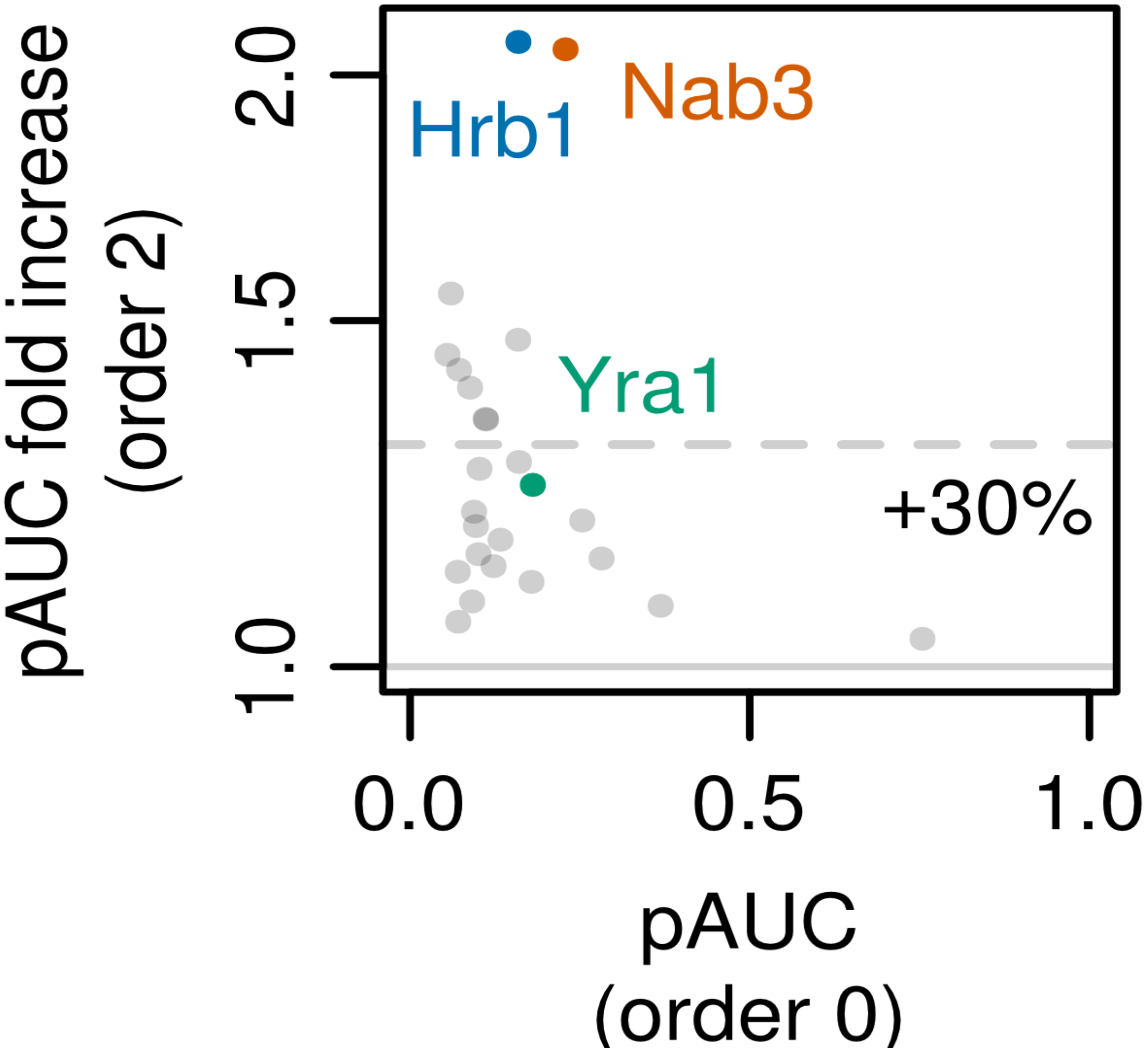
Pause sites

our models

inhomogeneous Markov model

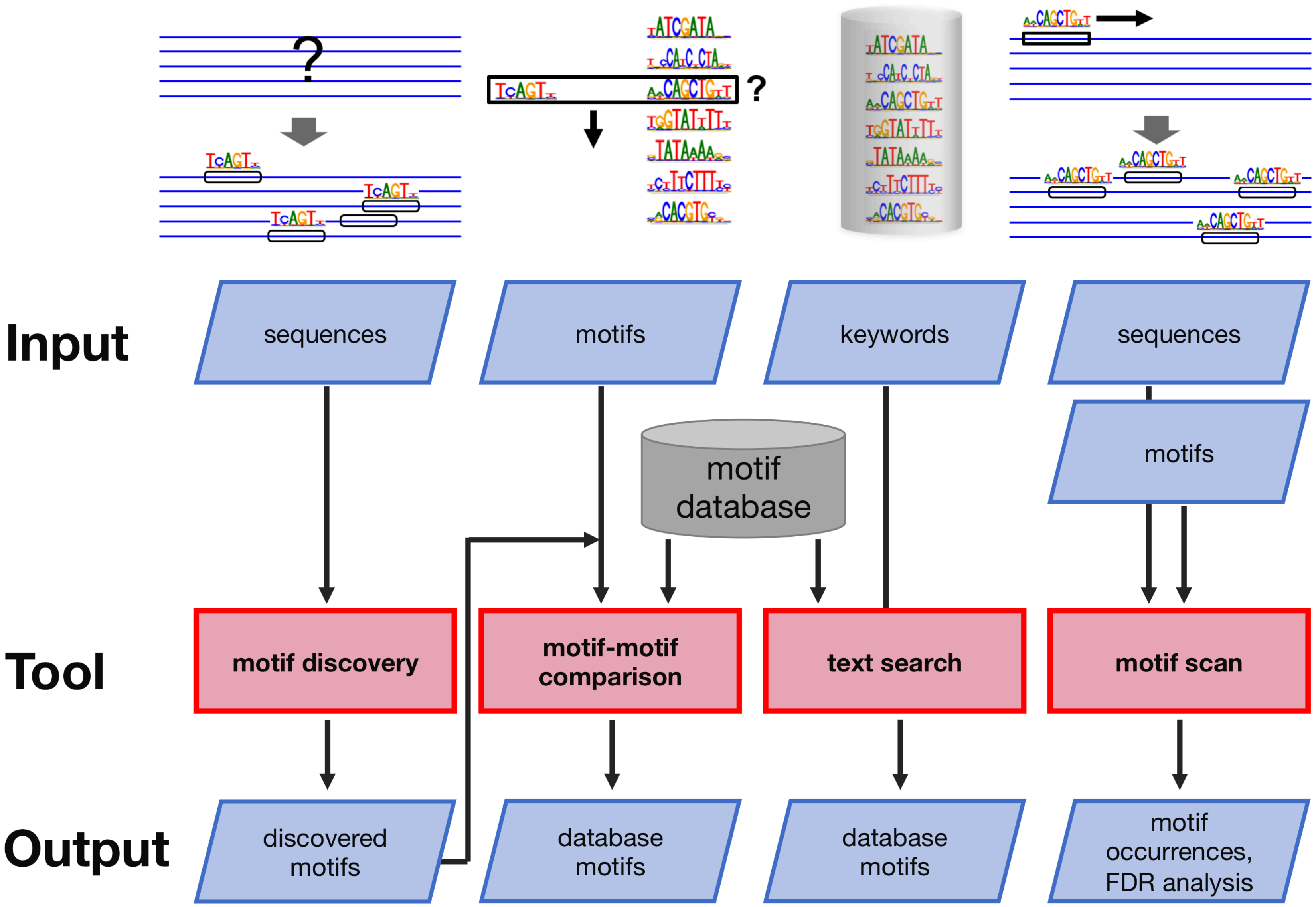
RNA-binding sites from PAR-CLIP: higher order vs. PWM

(data from Cramer lab, Goettingen)



BaMMmotif server offers 4 tools

https://bammmotif.mpibpc.mpg.de



Input

sequences

motifs

keywords

sequences

motifs

motif database

Tool

motif discovery

motif-motif comparison

text search

motif scan

Output

discovered motifs

database motifs

database motifs

motif occurrences, FDR analysis

How to assess motif models?

- E-values / P-values for enrichment of motif occurrences in input set vs. background set are popular

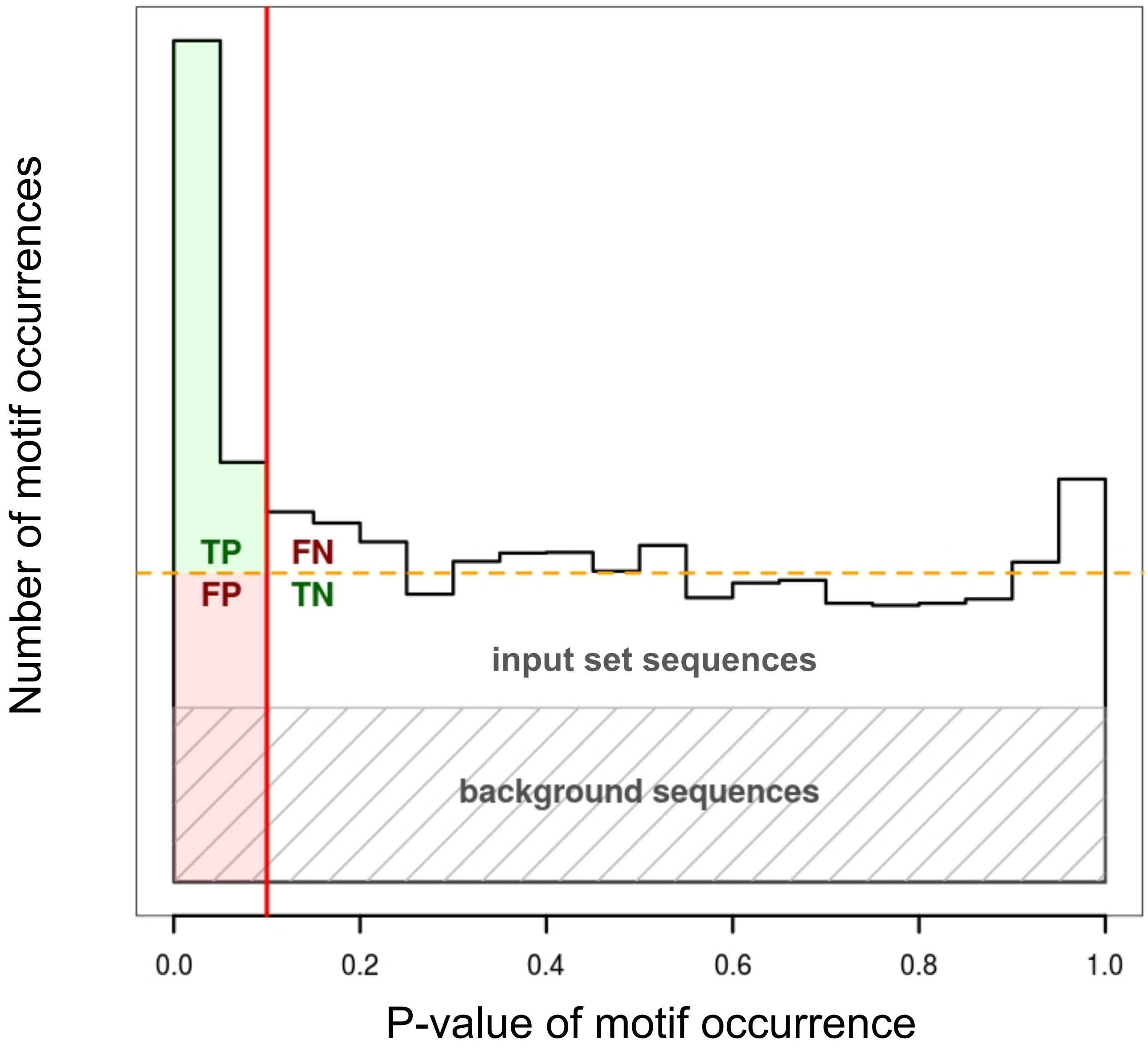
- But P-values can be very significant for motifs without biological relevance, e.g. input set is large and background set is not 100% realistic

- We need a quality measure that informs us about how well the model will identify binding sites in unseen datasets

- The demands on model specificity depend on the expected ratio of positive to negative sequences! E.g. ChIP-seq: ~1:1, scanning promoters: 1:100

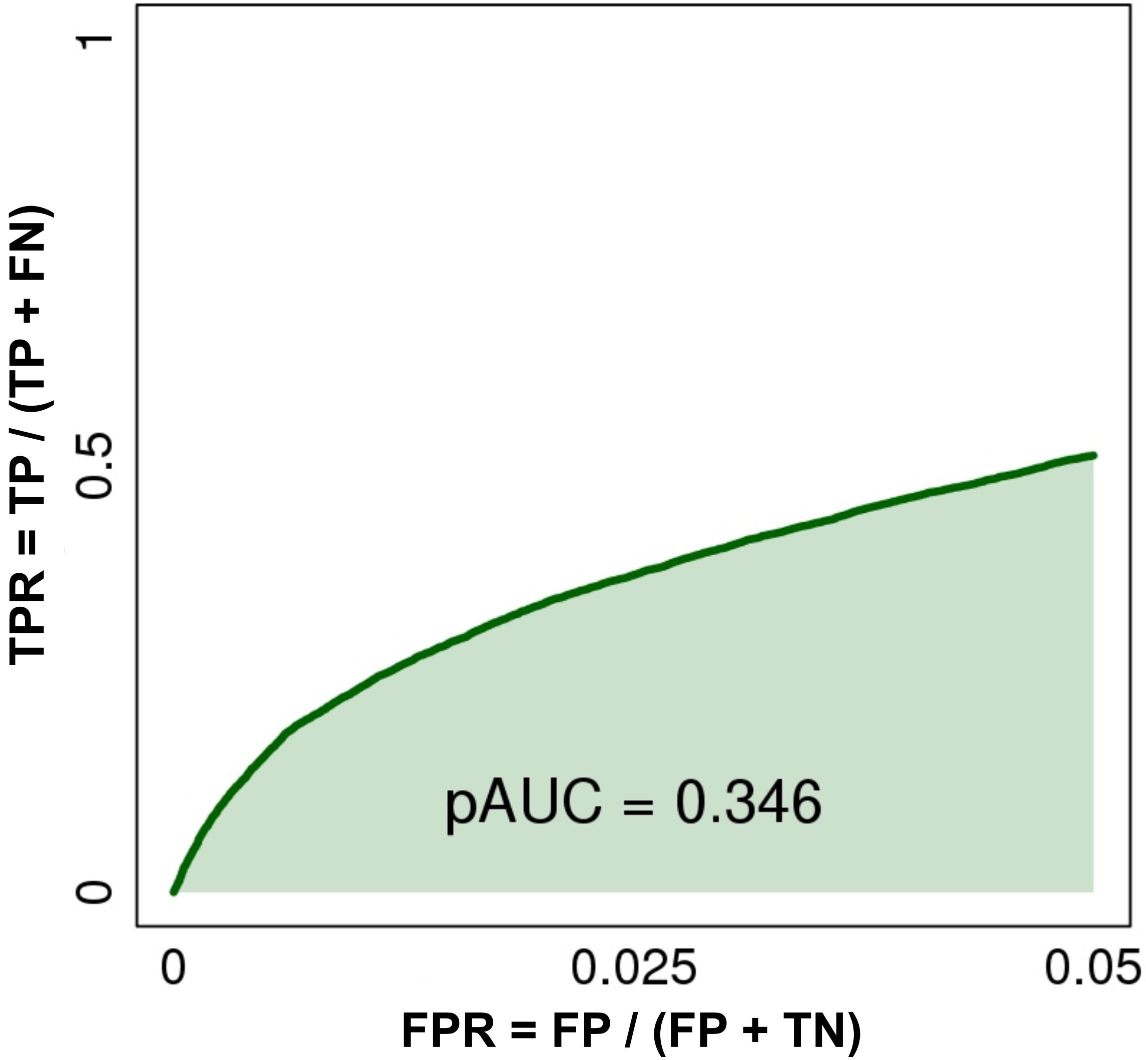
How to assess motif models?

Use ROC-like analysis



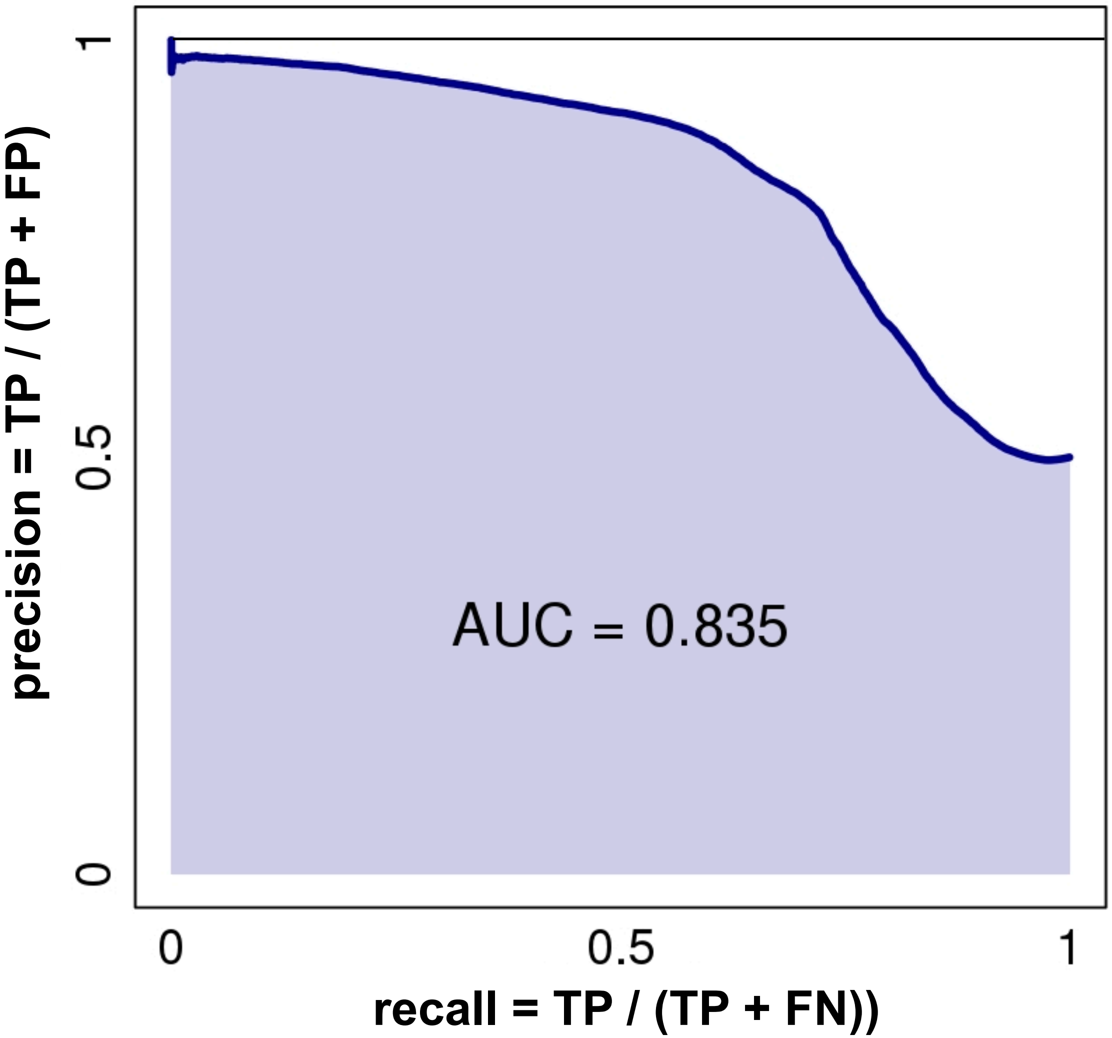
Partial ROC curve?

Relevant range of false positive rate (FPR) depends on expected pos:neg ratio!



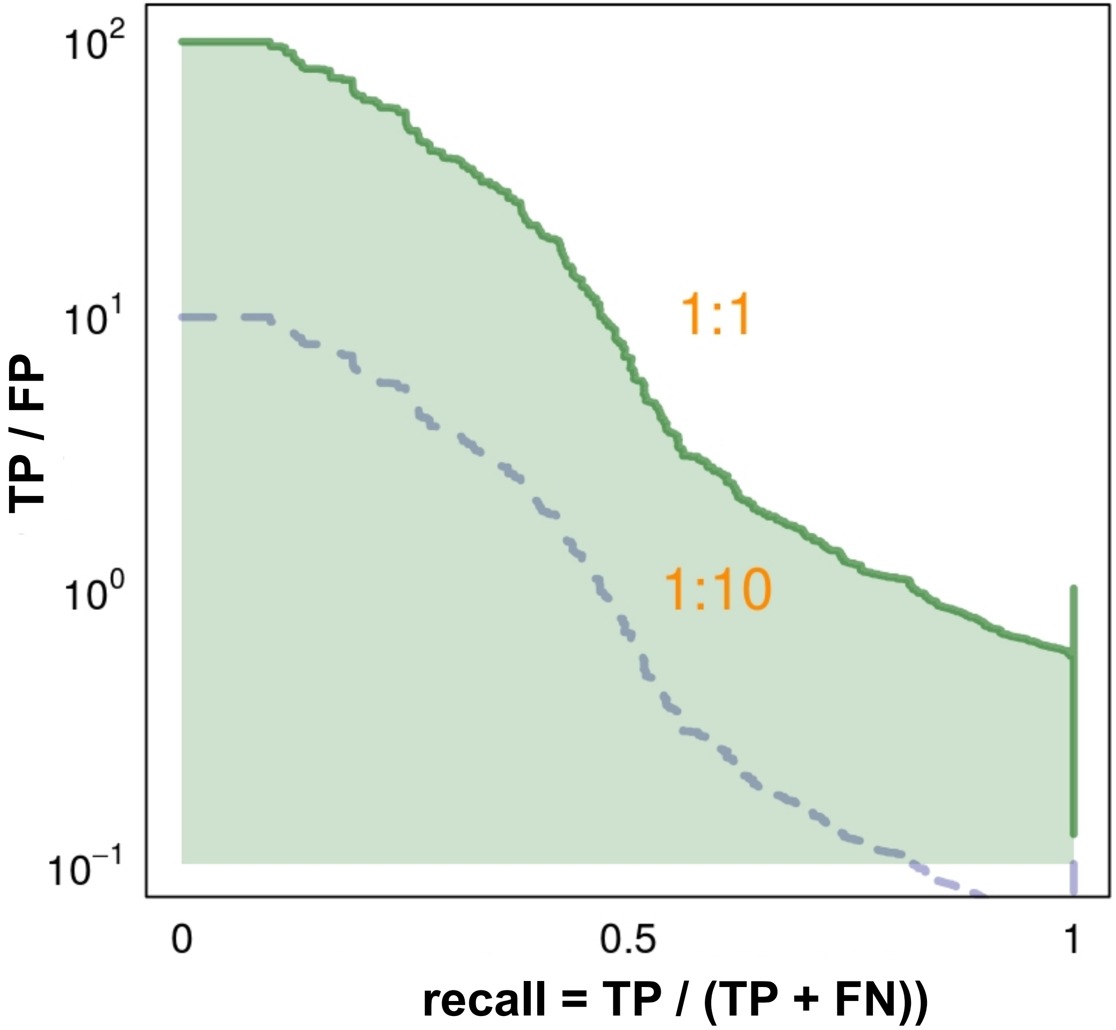
Precision-recall curve?

Relevant range of precision depends on expected pos:neg ratio!



TP/FP ratio-recall curve!

- Covers entire relevant range of precision or FDR (log scale)
- Different ratios pos:neg simply result in shifted curves



precision = 10/11 →

precision = 0.5 →

precision = 1/11 →

Summary (main part)

- Higher-order correlations significantly contribute to the binding specificity of most transcription factors
- Modeling higher correlations in a Bayesian framework can significantly improve predictions on 97% of tested factors, by +36% in pAUC on average. BUT: how much improvements due to learning >1 model in ChIP-seq data?
- BaMMs are very robust. They never overtrained
- BaMMs excel in learning complex motif architectures
- We should move from PWMs to higher-order models
- BaMMmotif server at https://bammmotif.mpibpc.mpg.de

MAX PLANCK INSTITUTE
FOR BIOPHYSICAL CHEMISTRY
(KARL FRIEDRICH BONHOEFFER INSTITUTE)

Thank you for listening!