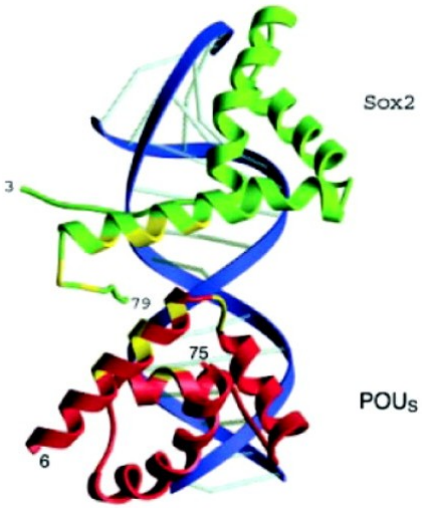
Predicting cell type-specific transcription factor cooperative binding

Aleksander Jankowski

October 19, 2011

Genome Institute of Singapore

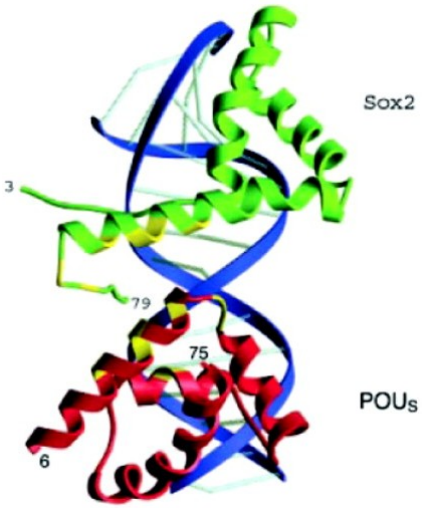
Motivation and introduction



Chambers
and Tomlinson,
Development 136,
2311-2322.

- Cooperative binding of transcription factors is essential for the regulation of gene expression.
- Some cooperativities are defining features of specific cell types, like the OCT-SOX cooperativity in embryonic stem cells.

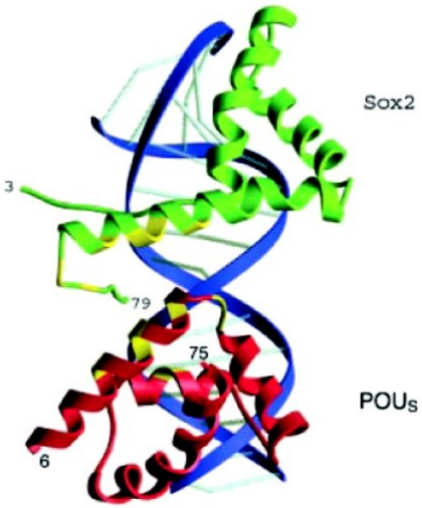
Motivation and introduction



Chambers and Tomlinson, *Development* 136, 2311-2322.

- Cooperative binding of transcription factors is essential for the regulation of gene expression.
- Some cooperativities are defining features of specific cell types, like the OCT-SOX cooperativity in embryonic stem cells.
- How to predict cooperative binding of transcription factors (TFs)?

Motivation and introduction



Chambers and Tomlinson, *Development* 136, 2311-2322.

- Cooperative binding of transcription factors is essential for the regulation of gene expression.
- Some cooperativities are defining features of specific cell types, like the OCT-SOX cooperativity in embryonic stem cells.
- How to predict cooperative binding of transcription factors (TFs)?
- How many cooperating pairs of TFs are there in the genome? In which cell types they interact?
- Could we systematically answer these questions?

Strategy

- Pique-Regi *et al.* (Genome Res., 2011) have shown the following "lemma":

  DNase-seq + position weight matrices $\rightarrow$ genome-wide, cell type-specific prediction of transcription factor binding.

Strategy

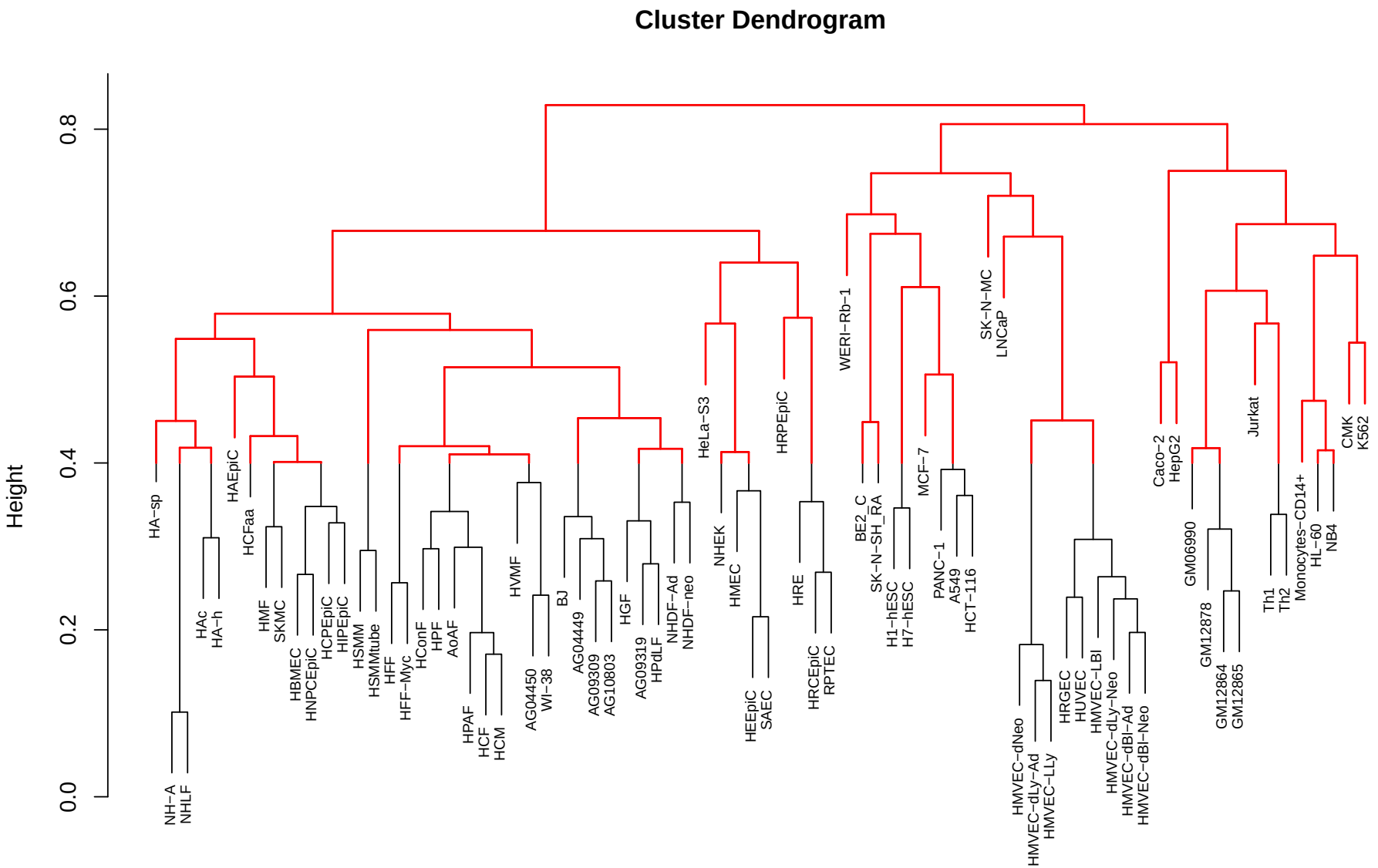
- Pique-Regi *et al.* (Genome Res., 2011) have shown the following "lemma":

  DNase-seq + position weight matrices $\rightarrow$ genome-wide, cell type-specific prediction of transcription factor binding.

- Our approach:

  DNase-seq + position weight matrices $\rightarrow$ genome-wide, cell type-specific prediction of transcription factor cooperative binding.

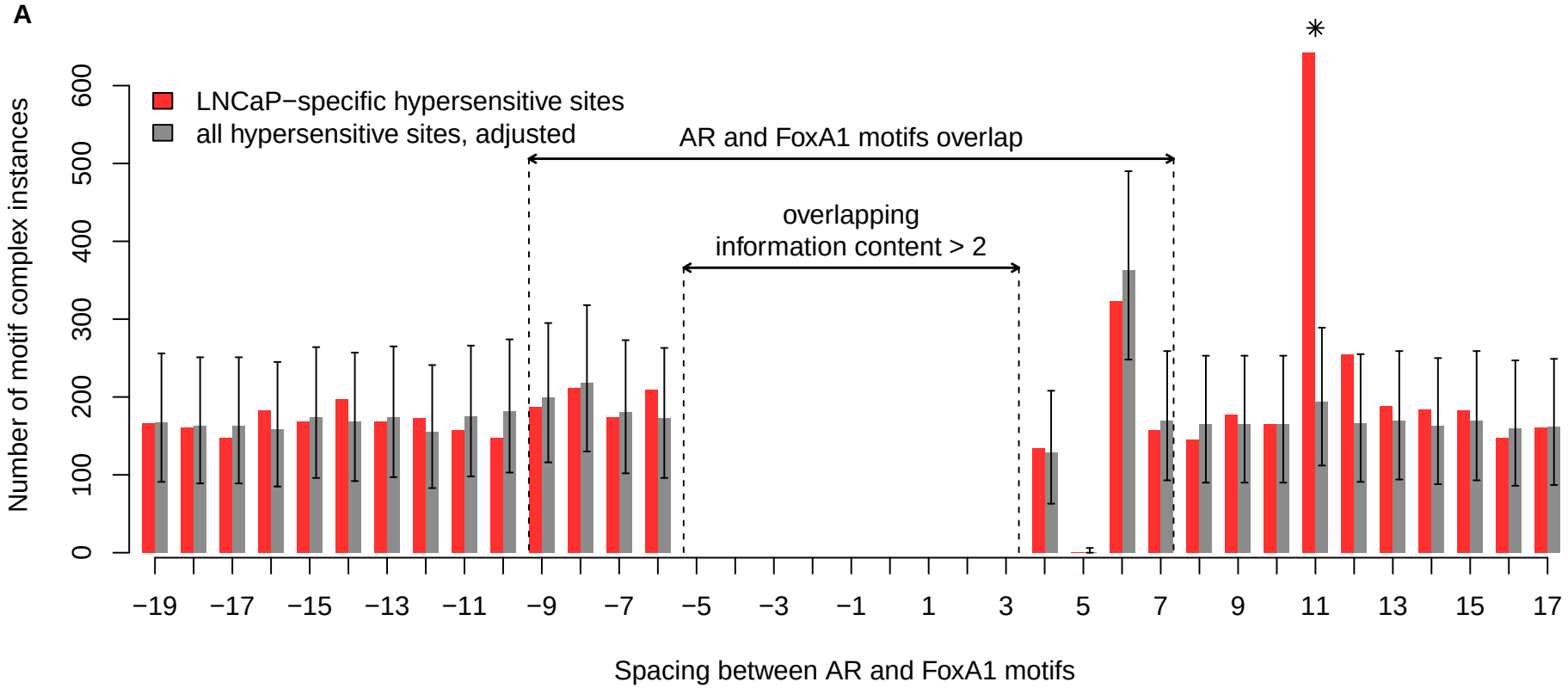
Clustering of cell types



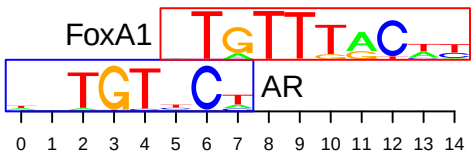
DNase-seq datasets by John Stam lab, University of Washington.

Example result: AR-FoxA1 motif complex in LNCaP cells

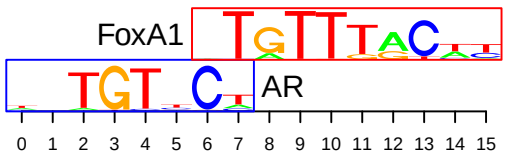
The position of AR motif is fixed,
we are considering different positions of FoxA1 motif.



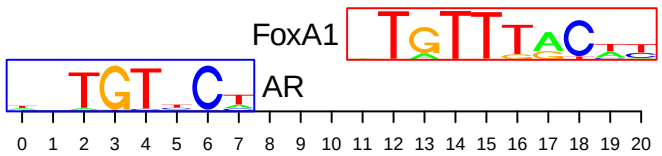
B spacing = 5, motif conflict



C spacing = 6, motif similarity



D spacing = 11, predicted TF cooperativity



Methods in brief

- For each cell type, pair of motifs $(\mathcal{M}_1, \mathcal{M}_2)$, and their fixed mutual orientation and spacing, calculate:
  - number of motif complex occurrences:
    - M_{12} – in the cell type-specific hypersensitive regions
    - m_{12} – in all hypersensitive regions.
  - number of possible motif complex binding sites:
    - N_{12} – in the cell type-specific hypersensitive regions
    - n_{12} – in all hypersensitive regions.
  - analogically, numbers of individual motif occurrences (M_1, M_2, m_1, m_2) and their possible binding sites (N_1, N_2, n_1, n_2).
- Test statistics: Bernoulli schema with success probability

$$p = \frac{f_1}{b_1}\frac{f_2}{b_2} \cdot \frac{m_{12}}{n_{12}}, \quad \text{where } f_i = \frac{M_i}{N_i} \text{ and } b_i = \frac{m_i}{n_i}.$$

- We calculate the p-value as the probability of observing at least M_{12} successes in N_{12} trials.

Methods in brief

- For each cell type, pair of motifs $(\mathcal{M}_1, \mathcal{M}_2)$, and their fixed mutual orientation and spacing, calculate:
  - number of motif complex occurrences:
    - M_{12} – in the cell type-specific hypersensitive regions
    - m_{12} – in all hypersensitive regions.
  - number of possible motif complex binding sites:
    - N_{12} – in the cell type-specific hypersensitive regions
    - n_{12} – in all hypersensitive regions.
  - analogically, numbers of individual motif occurrences (M_1, M_2, m_1, m_2) and their possible binding sites (N_1, N_2, n_1, n_2).
- Test statistics: Bernoulli schema with success probability

$$p = \frac{f_1}{b_1}\frac{f_2}{b_2} \cdot \frac{m_{12}}{n_{12}},$$ where $f_i = \frac{M_i}{N_i}$ and $b_i = \frac{m_i}{n_i}$.

- We calculate the p-value as the probability of observing at least M_{12} successes in N_{12} trials.

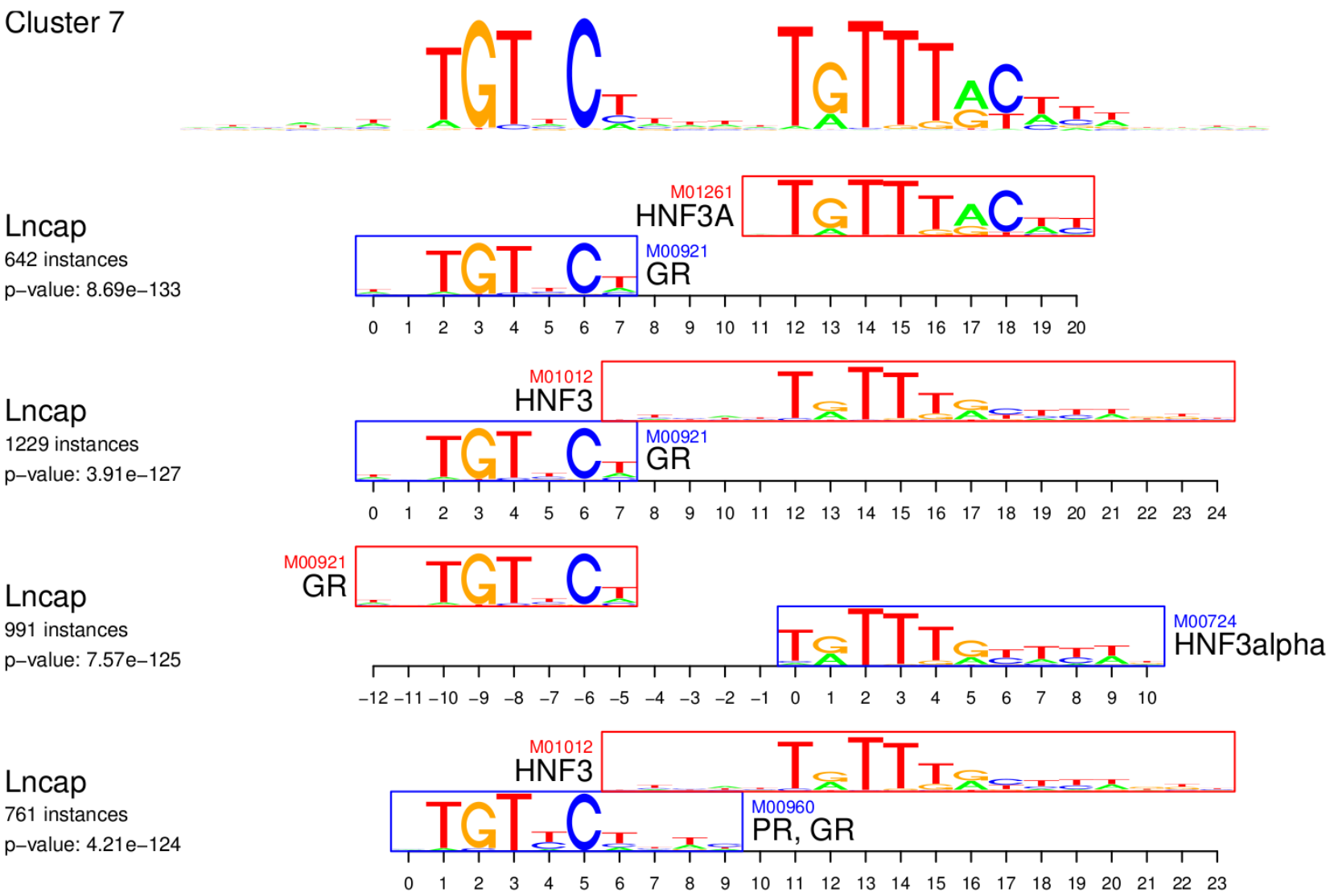
Results overview

- Approximate number of hypothesis considered:

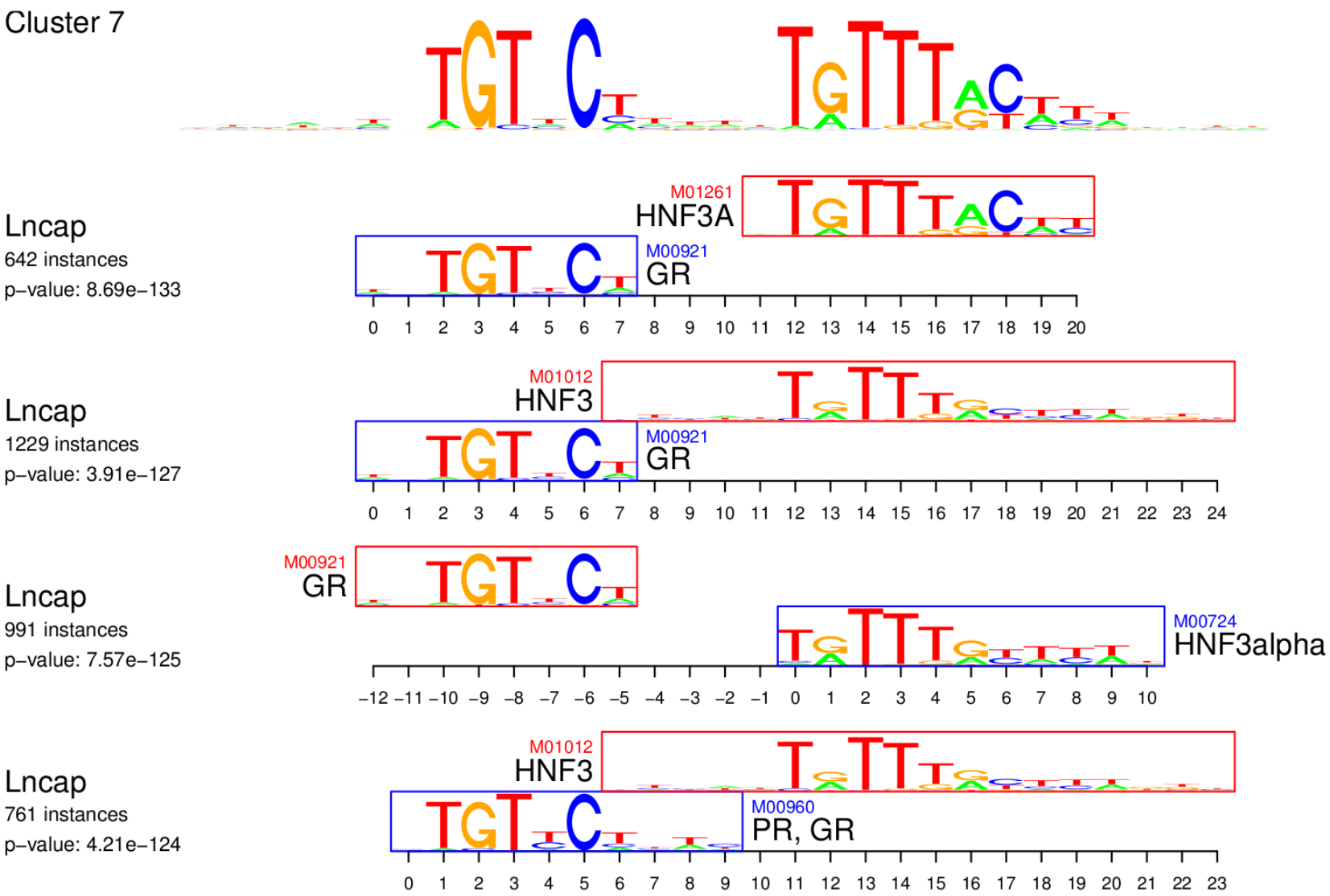
$$\frac{1000 \text{ motifs} \cdot 1000 \text{ motifs}}{2} \cdot 40 \text{ cell types} \cdot 2 \cdot 40 \text{ spacings}.$$

- We found 4520 overrepresented motif complexes.
- To account for the redundancy of the motif database, we clustered them into 460 cell type-specific predictions.
- We also provide a dataset of their 270 713 genomic locations.

AR-FoxA1 cooperativity in LNCaP cells



AR-FoxA1 cooperativity in LNCaP cells

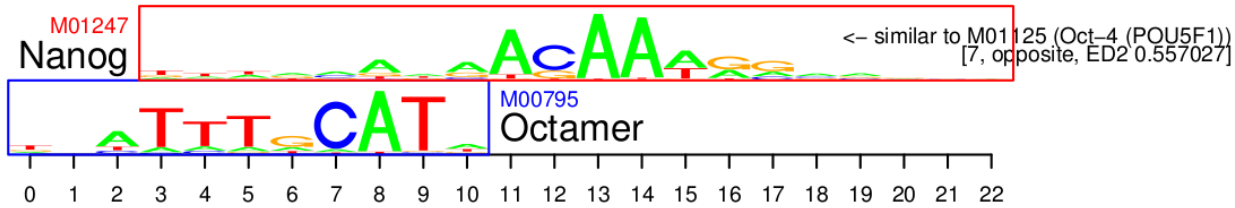


Cooperativity reported recently (Wang *et al.*, Nature, 2011).

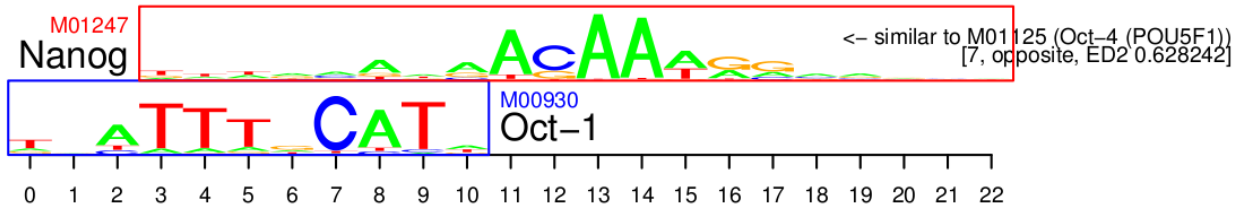
OCT-SOX cooperativity in embryonic stem cells

Cluster 3

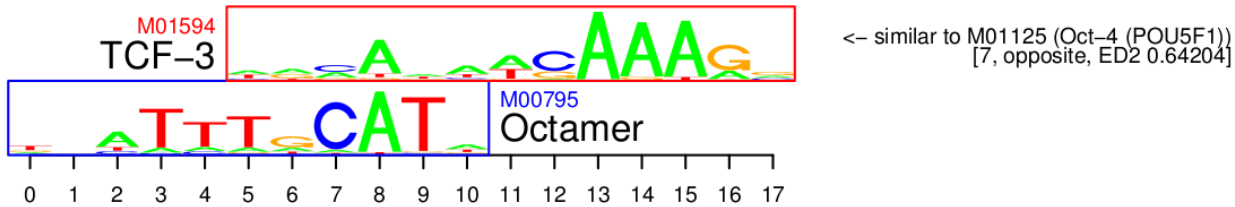
H1hescH7es
1636 instances
p–value: 1.25e–233



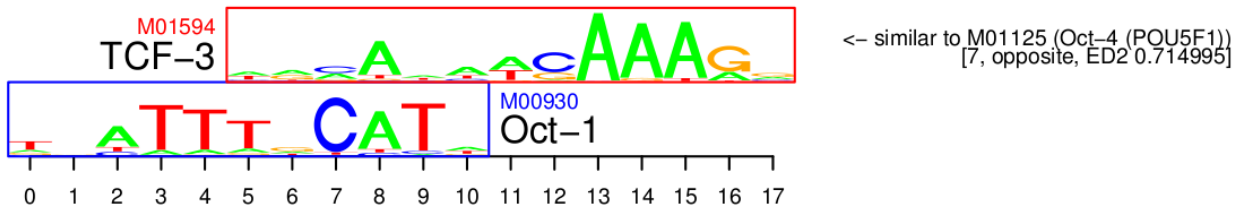
H1hescH7es
1413 instances
p–value: 4.03e–206



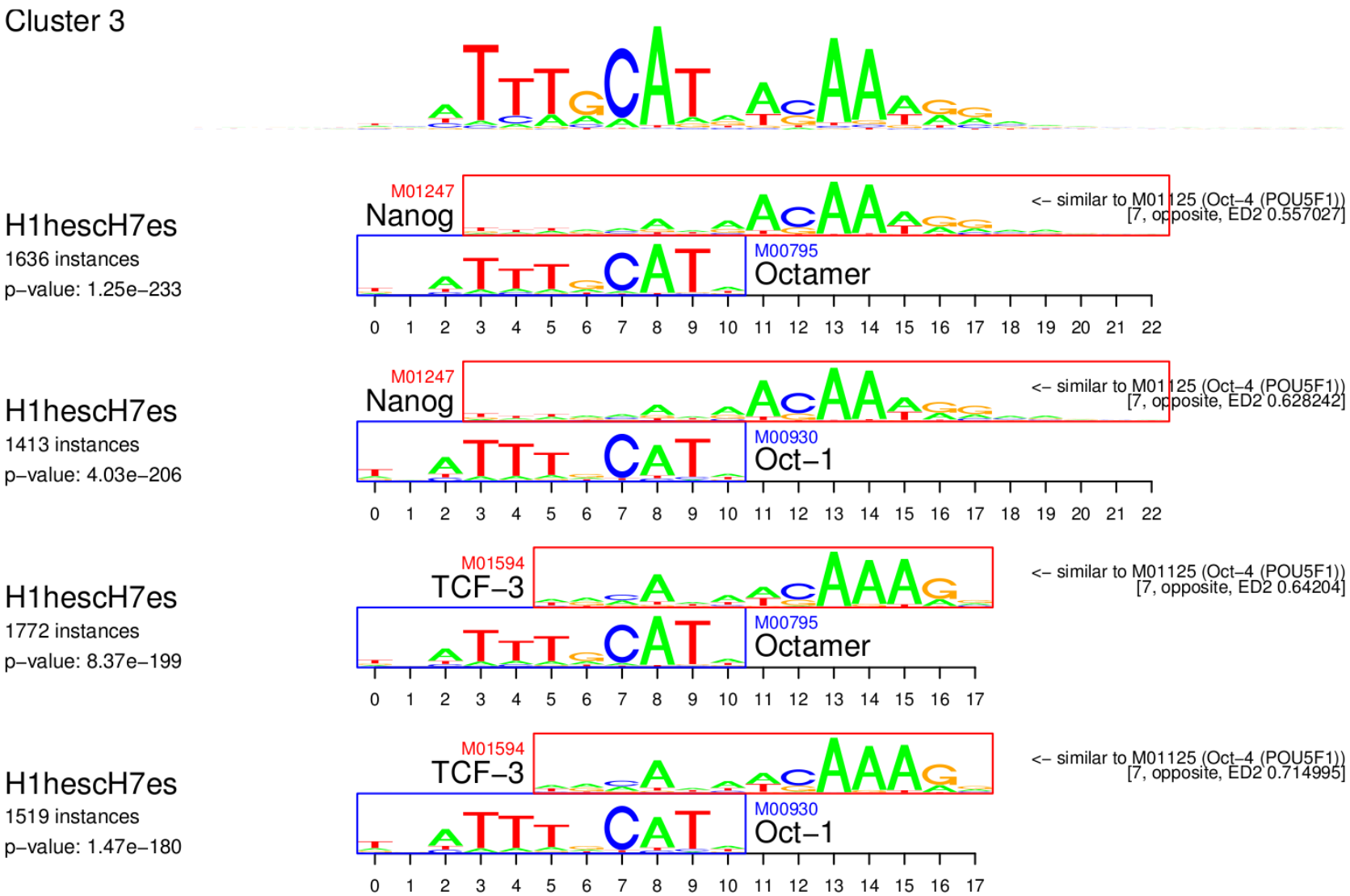
H1hescH7es
1772 instances
p–value: 8.37e–199



H1hescH7es
1519 instances
p–value: 1.47e–180

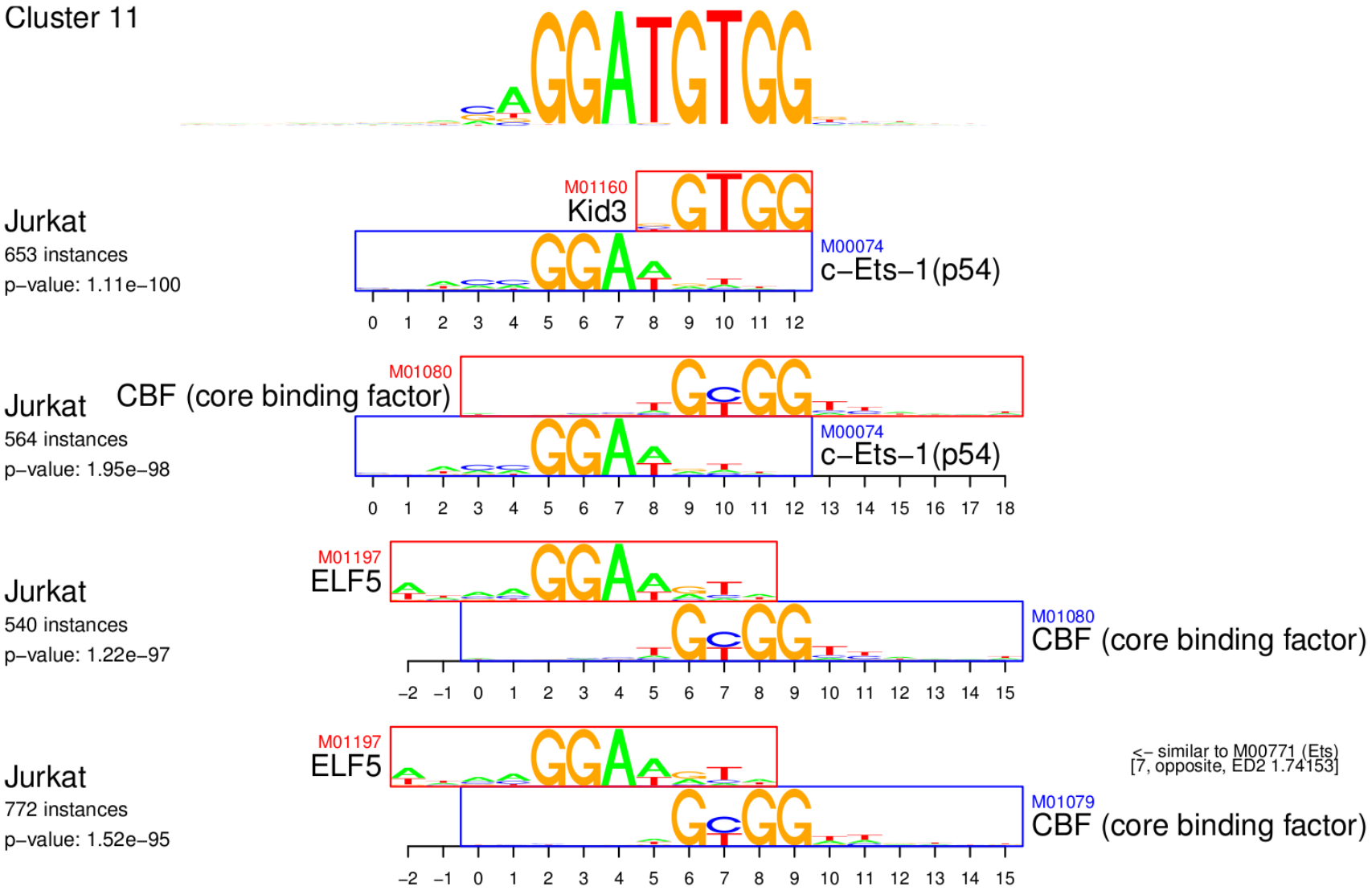


OCT-SOX cooperativity in embryonic stem cells

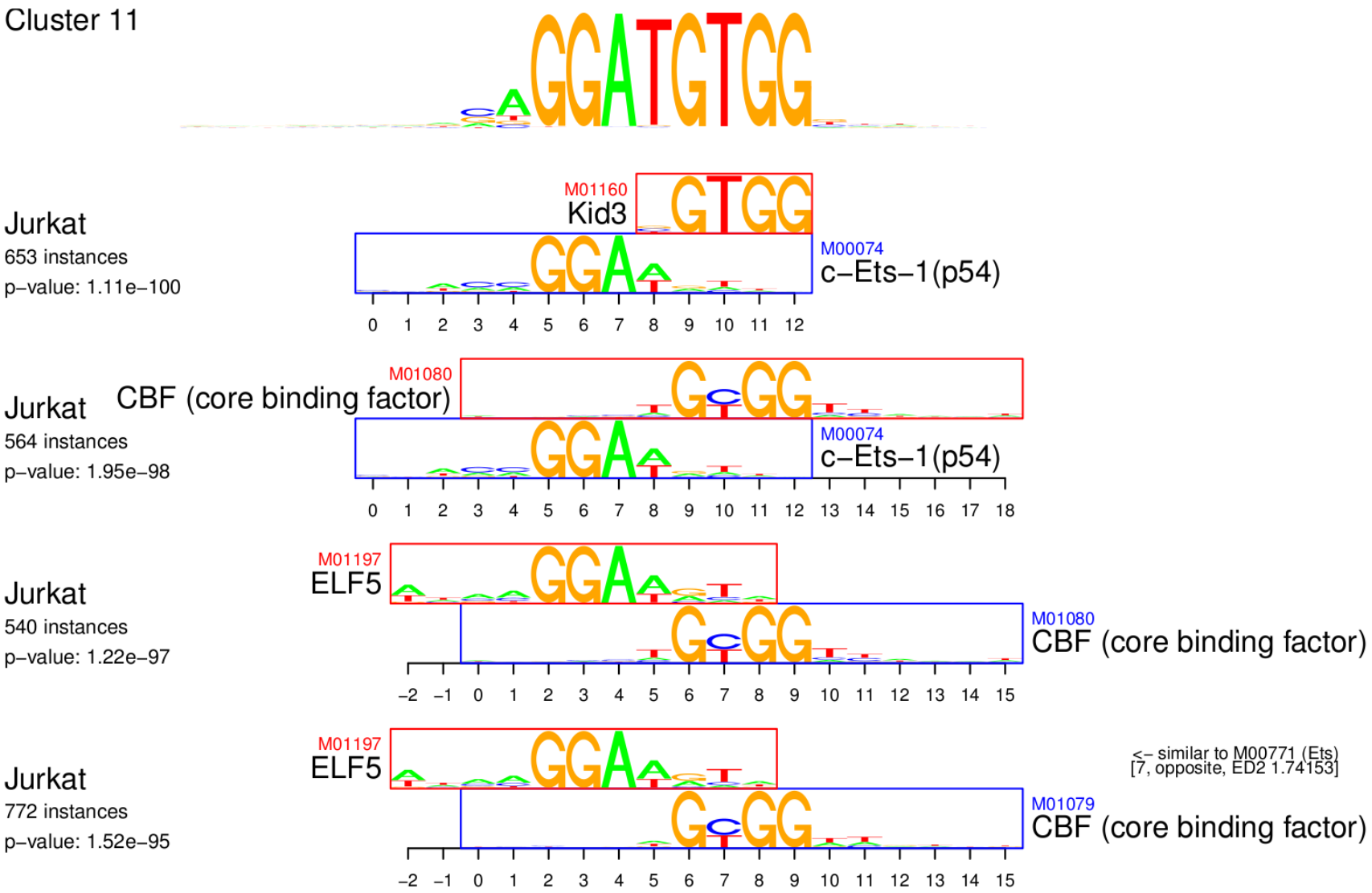


Well-recognized cooperativity, essential for maintaining pluripotent state in embryonic stem cells (Chen *et al.*, Cell, 2008).

ETS-CBF cooperativity in Jurkat cells

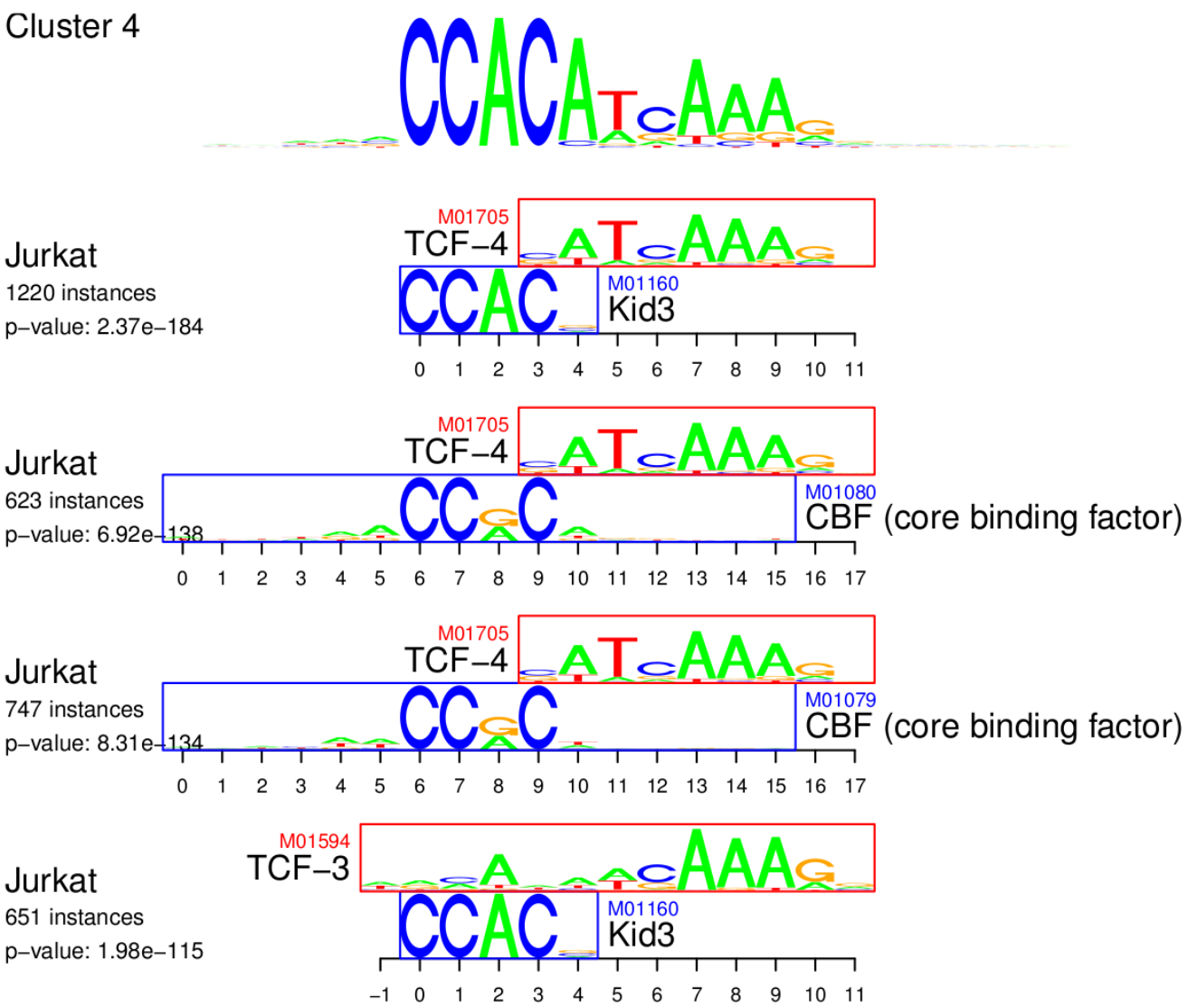


ETS-CBF cooperativity in Jurkat cells

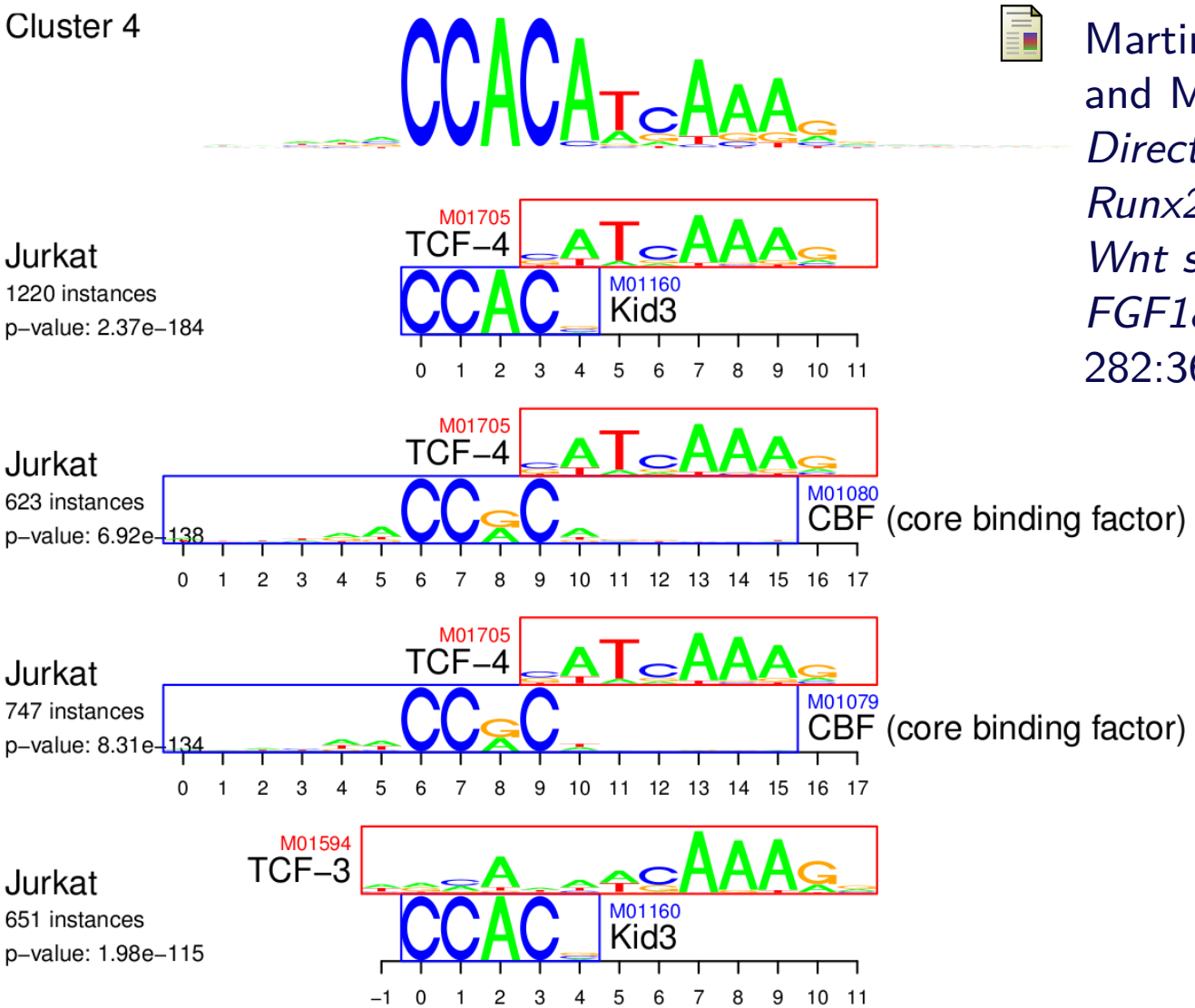


Cooperativity reported previously (Hollenhorst *et al.*, PLoS Genet., 2009).

TCF-CBF cooperativity in Jurkat cells



TCF-CBF cooperativity in Jurkat cells

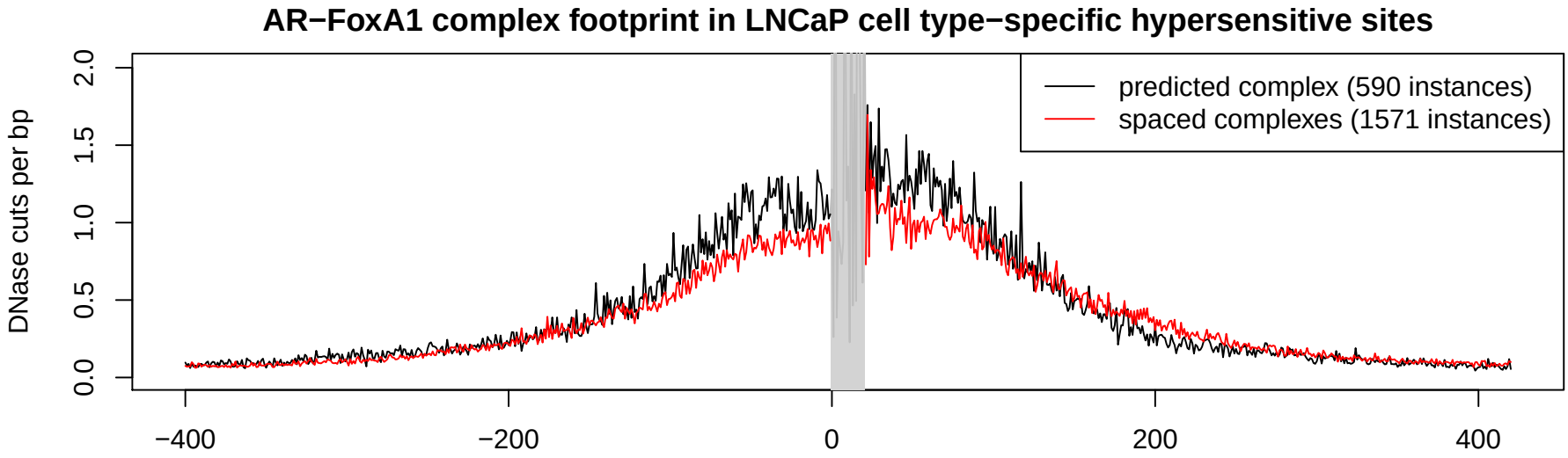


Martina I. Reinhold and Michael C. Naski, *Direct interactions of Runx2 and canonical Wnt signaling induce FGF18*. J. Biol. Chem. 282:3653-3663, 2007.

Cooperativity prediction with high biological relevance.

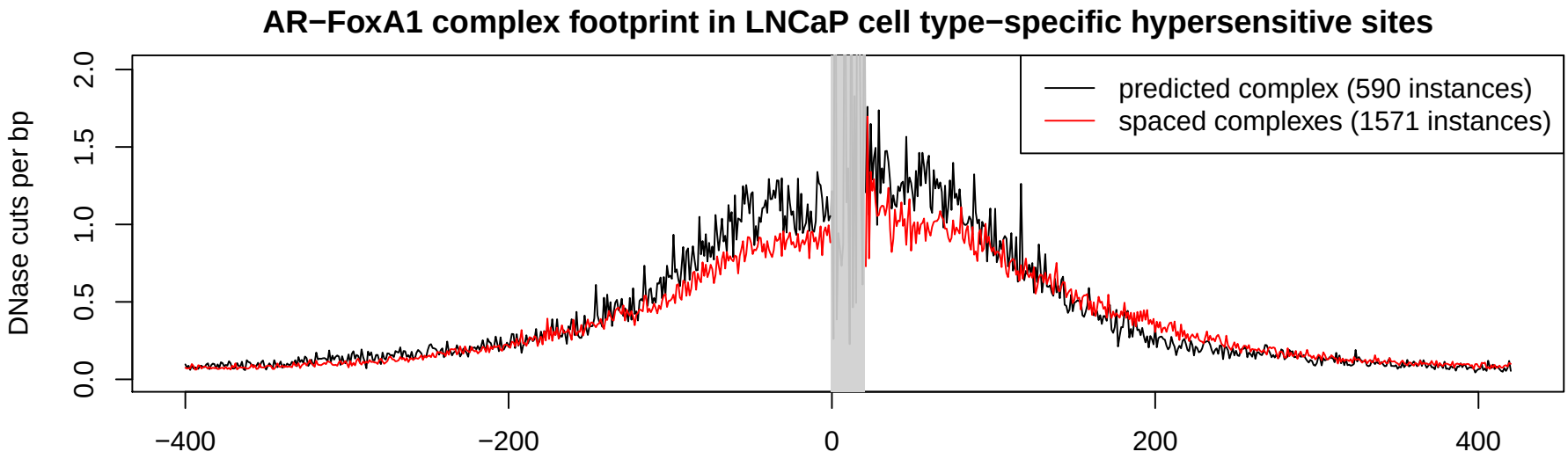
Validation using different aspect of DNase-seq data

- Instead of limiting to the hypersensitive (DNase cut-enriched) regions, we use the actual DNase cut density near motif complex instances.
- We compared the number of DNase cuts ($\pm$100 bp) between
  - instances of predicted motif complex
  - instances of the motif complex consisting of the same two motifs, but with slightly increased spacing (spaced complexes)



Validation using different aspect of DNase-seq data

- Instead of limiting to the hypersensitive (DNase cut-enriched) regions, we use the actual DNase cut density near motif complex instances.
- We compared the number of DNase cuts ($\pm$100 bp) between
  - instances of predicted motif complex
  - instances of the motif complex consisting of the same two motifs, but with slightly increased spacing (spaced complexes)



- 48% of our predictions are called significant, given 5% FDR.

Validation with the atlas of TF interactions

- To verify our predictions in a systematic manner, we confront them with the atlas of combinatorial transcriptional regulation in man (Ravasi *et al.*, 2010).
- The atlas contains 5238 interactions between 1988 human TFs, detected by mammalian two-hybrid assays.
- It is not certain how applicable the atlas is to the cellular *in vivo* conditions.

Validation with the atlas of TF interactions

- To verify our predictions in a systematic manner, we confront them with the atlas of combinatorial transcriptional regulation in man (Ravasi *et al.*, 2010).
- The atlas contains 5238 interactions between 1988 human TFs, detected by mammalian two-hybrid assays.
- It is not certain how applicable the atlas is to the cellular *in vivo* conditions.
- The comparison shows that 15% of our predictions are confirmed by the atlas (p-value $< 10^{-69}$).

Summary

- DNase-seq datasets are sufficient to identify cell type-specific transcription factor direct cooperativity.
- We found 460 cell type-specific TF cooperative interactions, vast majority of them are novel.
- Our predictions were corroborated by the DNase cut density and mammalian two-hybrid assays.

Summary

- DNase-seq datasets are sufficient to identify cell type-specific transcription factor direct cooperativity.
- We found 460 cell type-specific TF cooperative interactions, vast majority of them are novel.
- Our predictions were corroborated by the DNase cut density and mammalian two-hybrid assays.
- Direct cooperativity of transcription factors seems to be a widespread mechanism.

Acknowledgements

Ewa Szczurek
Jerzy Tiuryn
University of Warsaw

Shyam Prabhakar
Genome Institute of Singapore

John Stamatoyannopoulos Lab
University of Washington
ENCODE Project