Number of transcription factors and binding sites

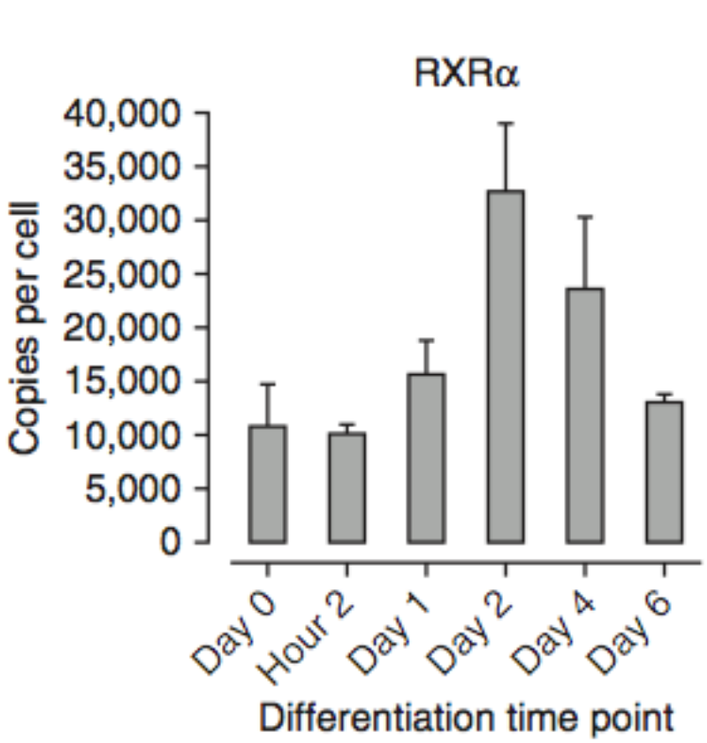
- Mass spectrometry (proteomics) technique to count absolute nb of TF per cell
- ChIP-seq to measure genome-wide (relative) occupancy of same TF
- ChIP-seq of histone modification to characterize accessible vs inaccessible chromatin

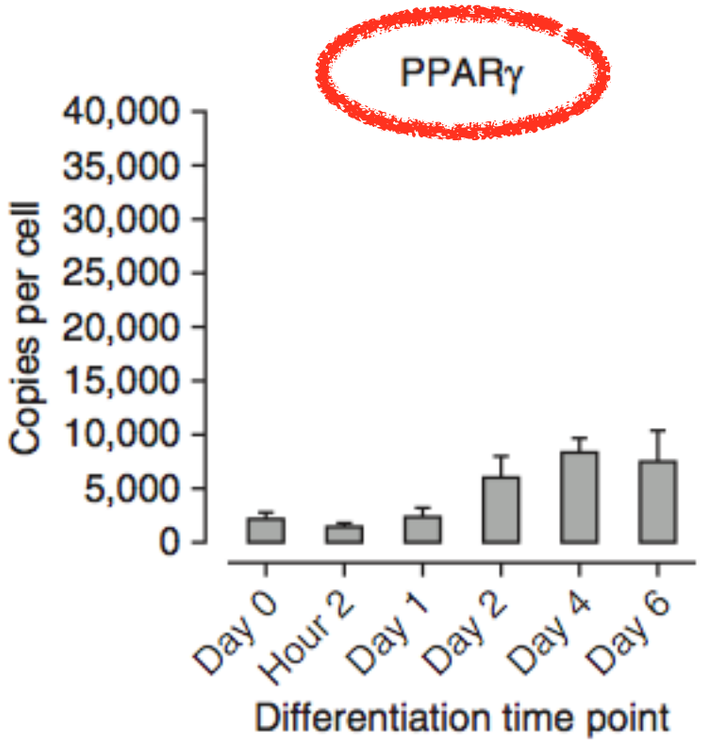
Absolute quantification of transcription factors during cellular differentiation using multiplexed targeted proteomics

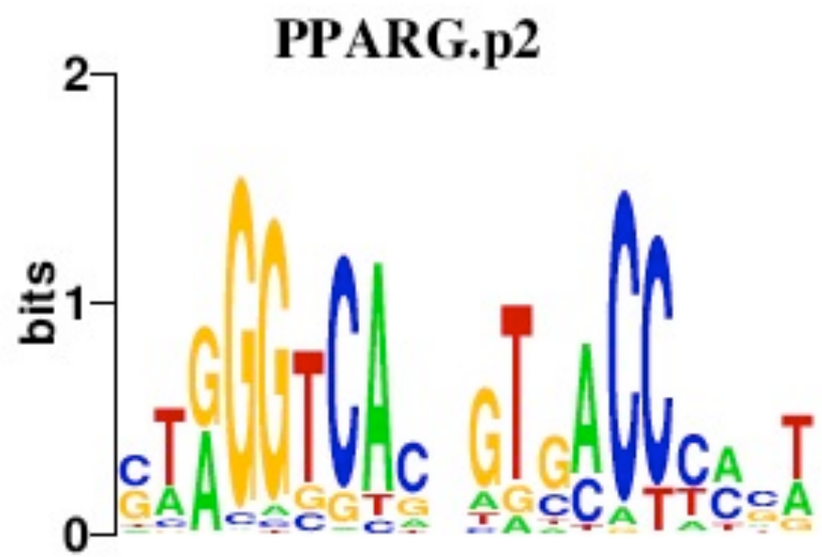
Jovan Simicevic[1,5], Adrien W Schmid[2,5], Paola A Gilardoni[1,5], Benjamin Zoller[3], Sunil K Raghav[1], Irina Krier[1], Carine Gubelmann[1], Frédérique Lisacek[4], Felix Naef[3], Marc Moniatte[2] & Bart Deplancke[1]

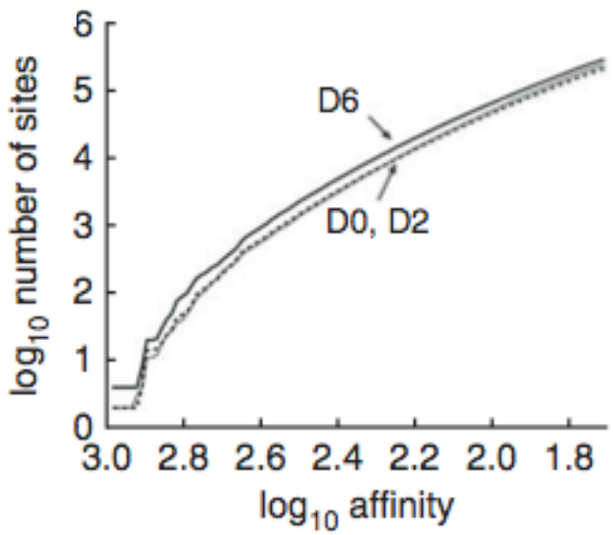
Nat Methods 10, 570–576 (2013)

Number of transcription factors and binding sites

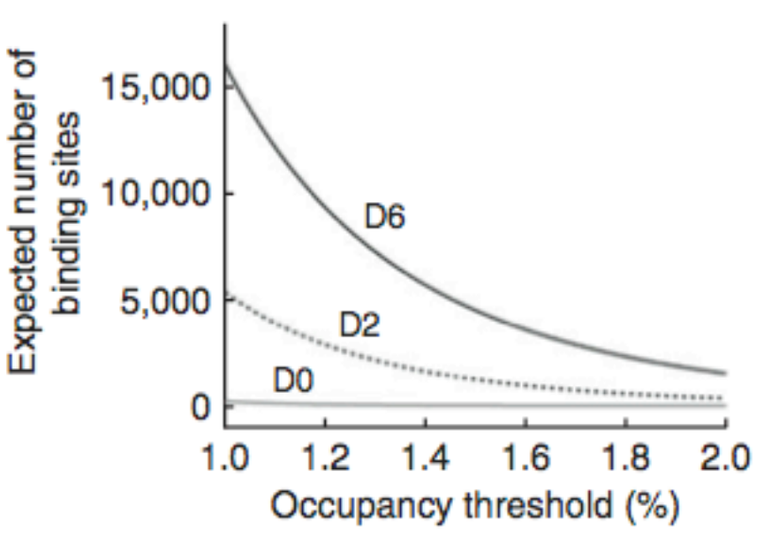




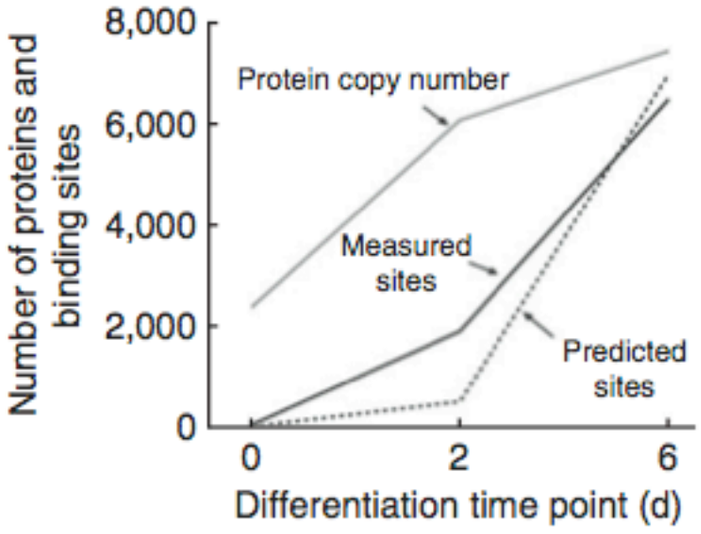




Binding sites in accessible genome



Predicted occupied sites



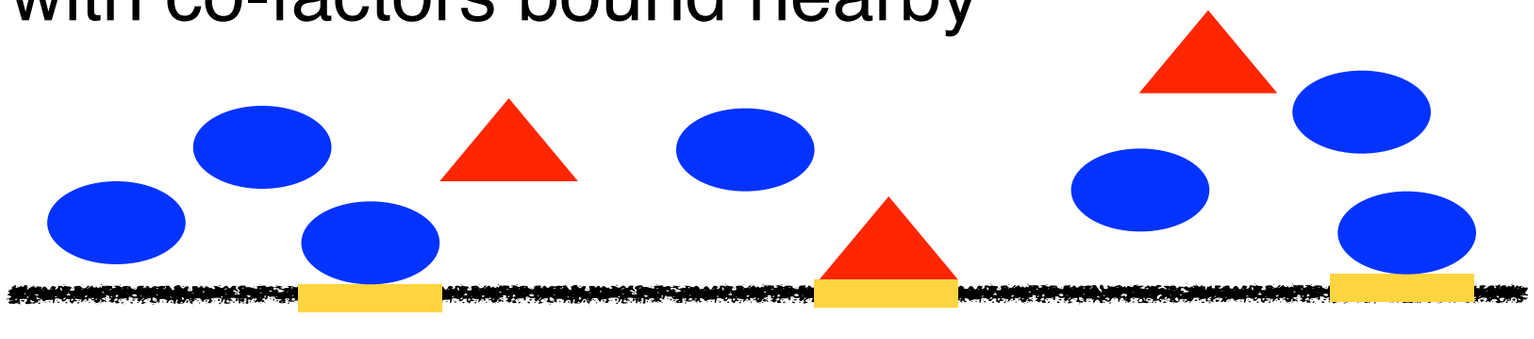
Protein total / bound

What are the limitations of this model?

Model is wrong if occupancy is not just a function of site affinity and protein concentration, for example in case of:

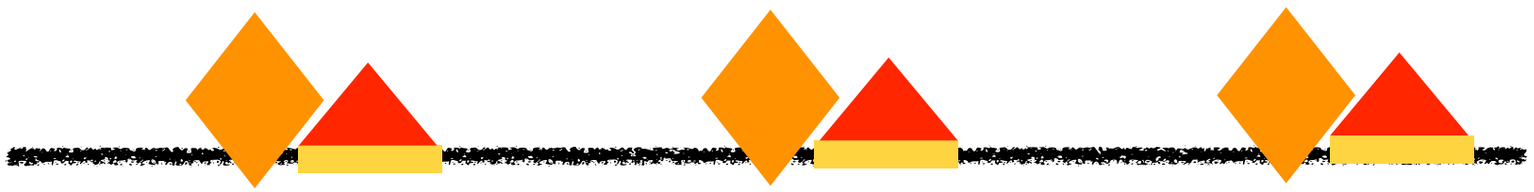
1. Competition with a different factor for similar sites
2. Cooperative interaction with co-factors bound nearby

high concentration of competitor → low occupancy by TF

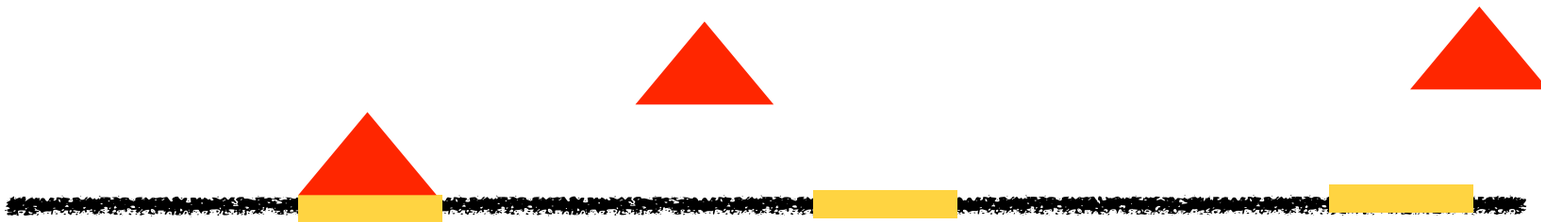


low concentration of competitor → high occupancy by TF

Presence of cooperative co-factor → longer residency time



Depletion of co-factor → shorter residency time



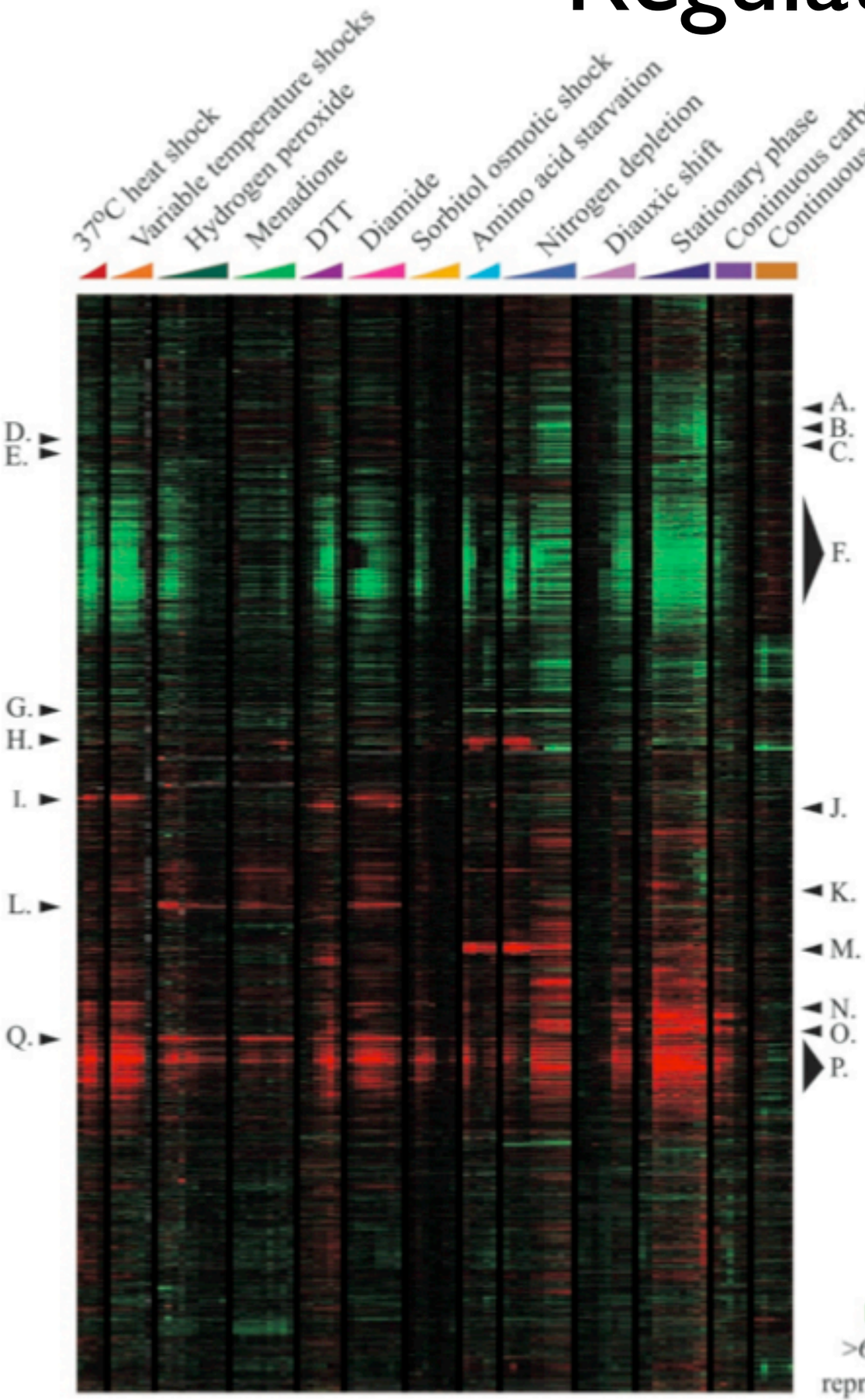
What are the limitations of this model?

We treat every cell like the population average

- Average transcription factor concentration is low, this does not imply that it is low in every cell
- The average site occupancy is a non-linear function of protein concentration: the average occupancy at fluctuating concentrations is not equal to the occupancy at the average concentration

The factor can exist in an active and an inactive chemical form, the measured concentration is not always able to distinguish the two forms

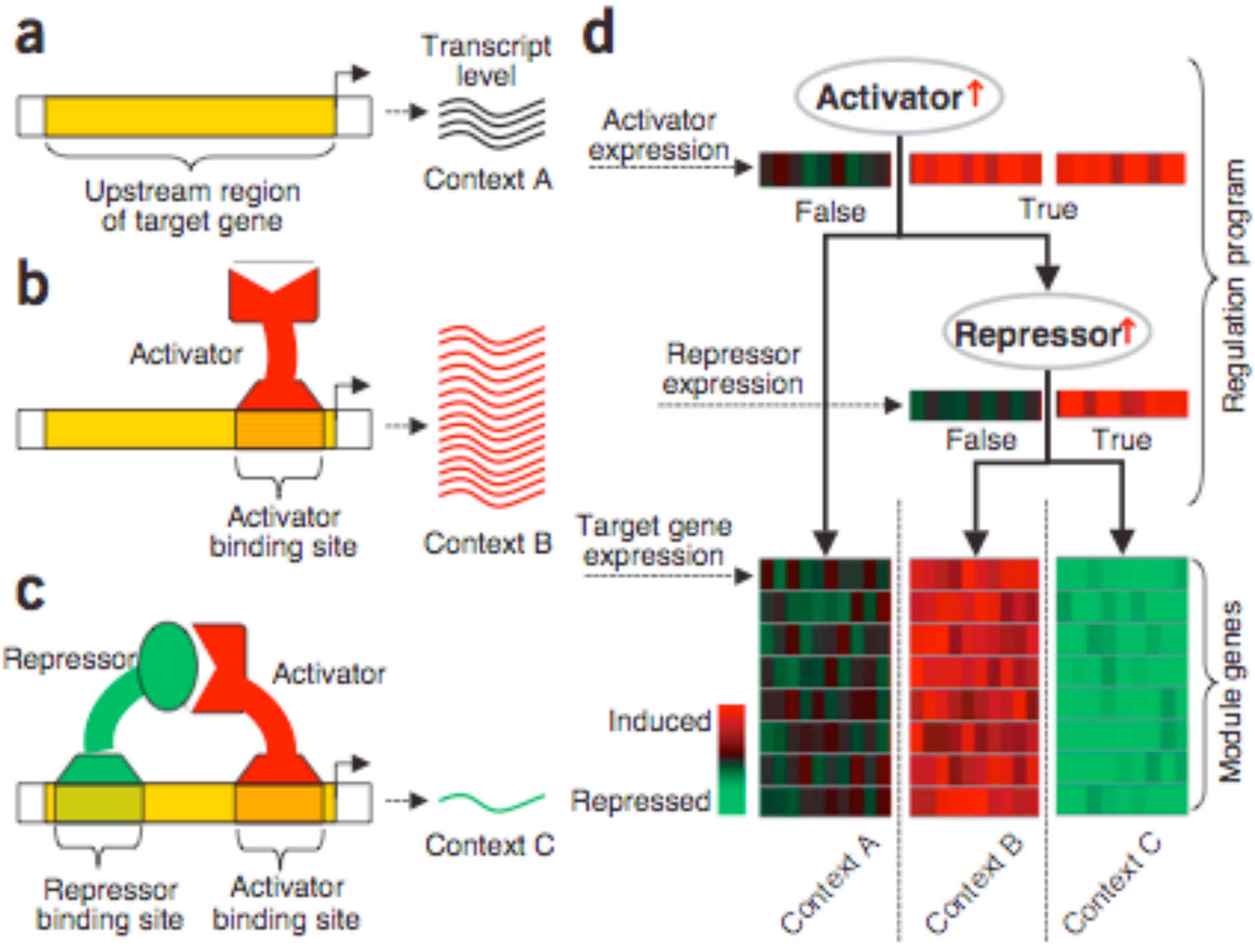
Regulatory modules



- Yeast dataset: 2355 genes, 173 conditions

Segal et al. Nat Genet. 2003

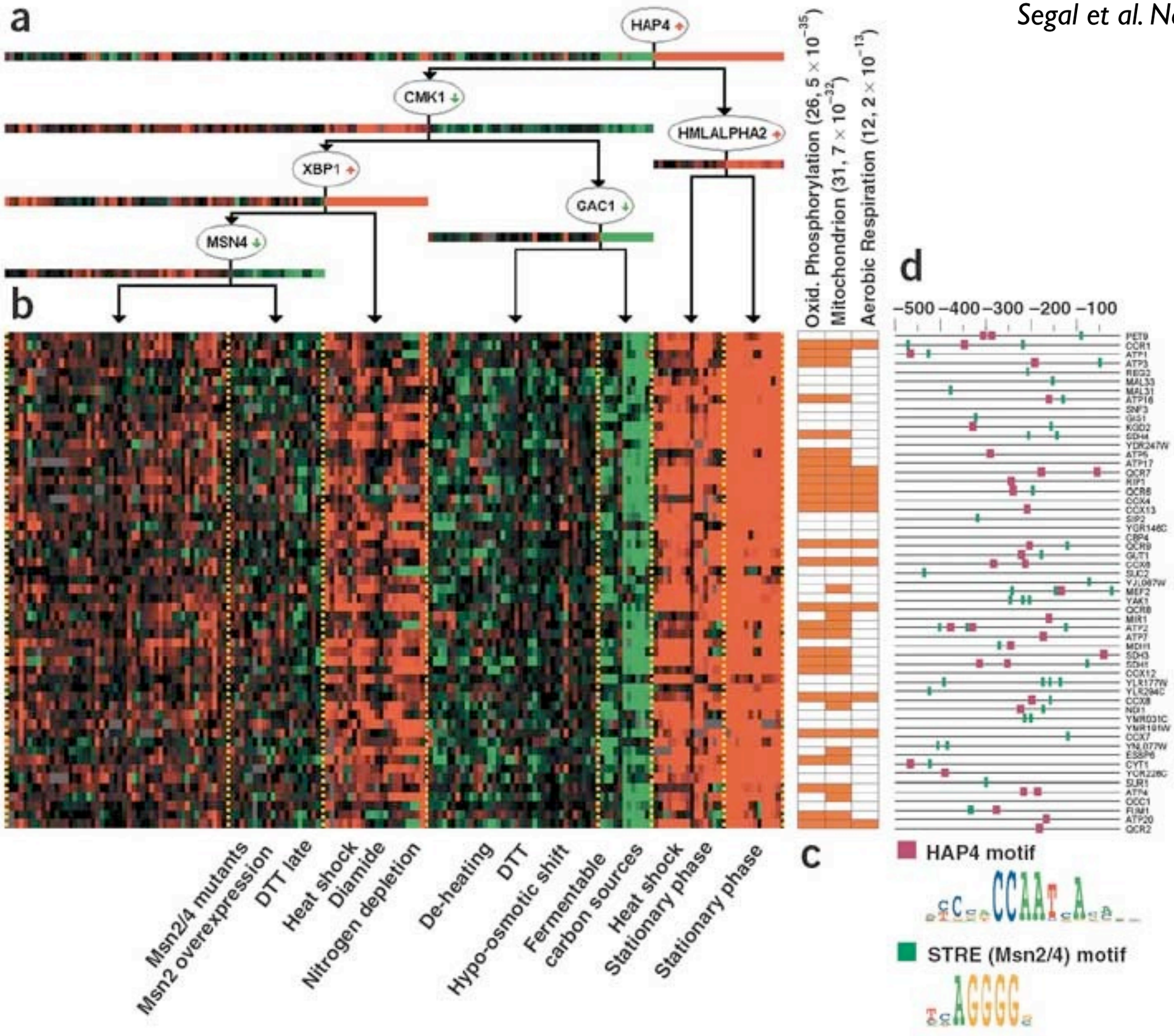
Regulatory modules



Segal et al. Nat Genet. 2003

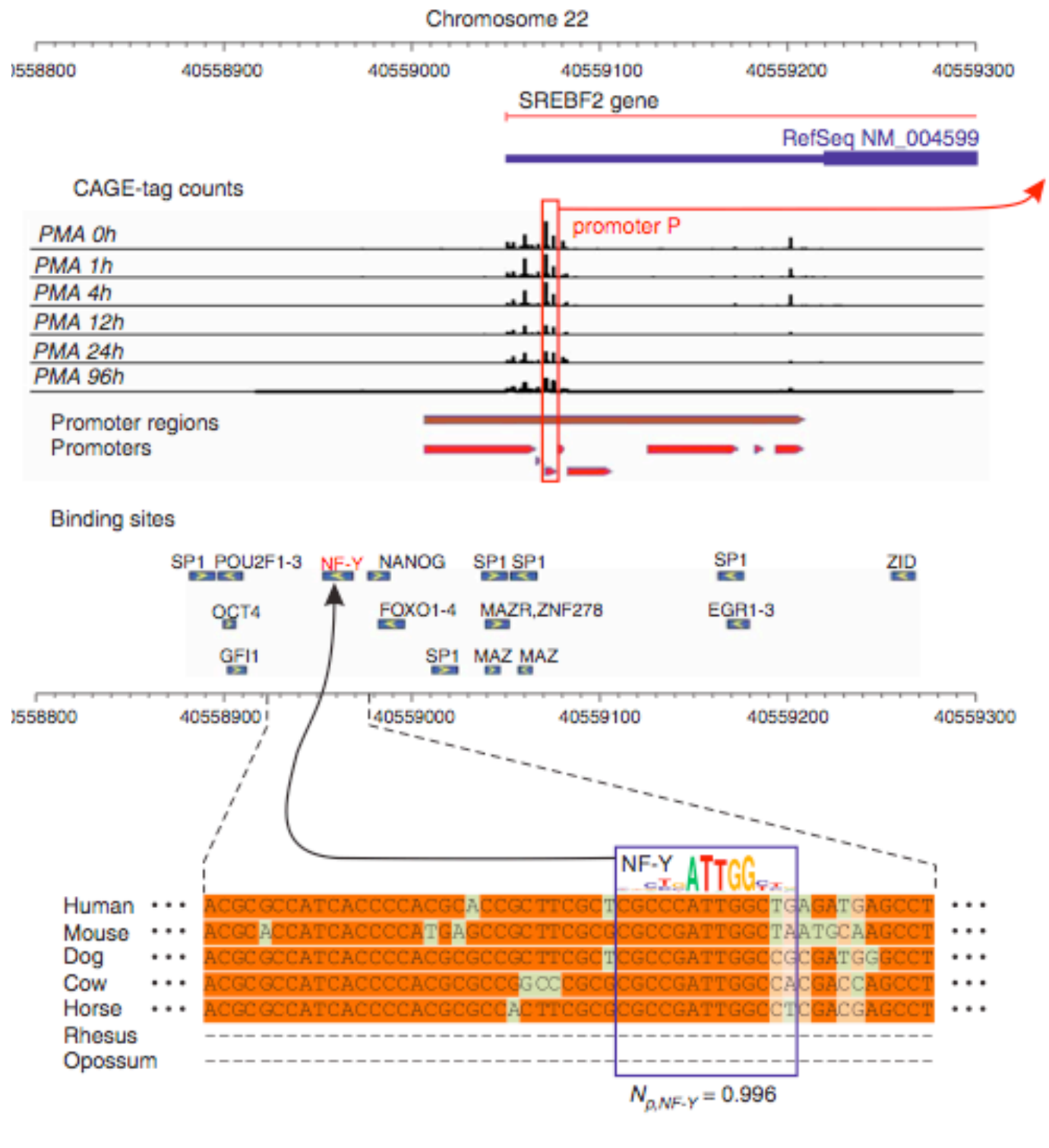
Regulatory modules

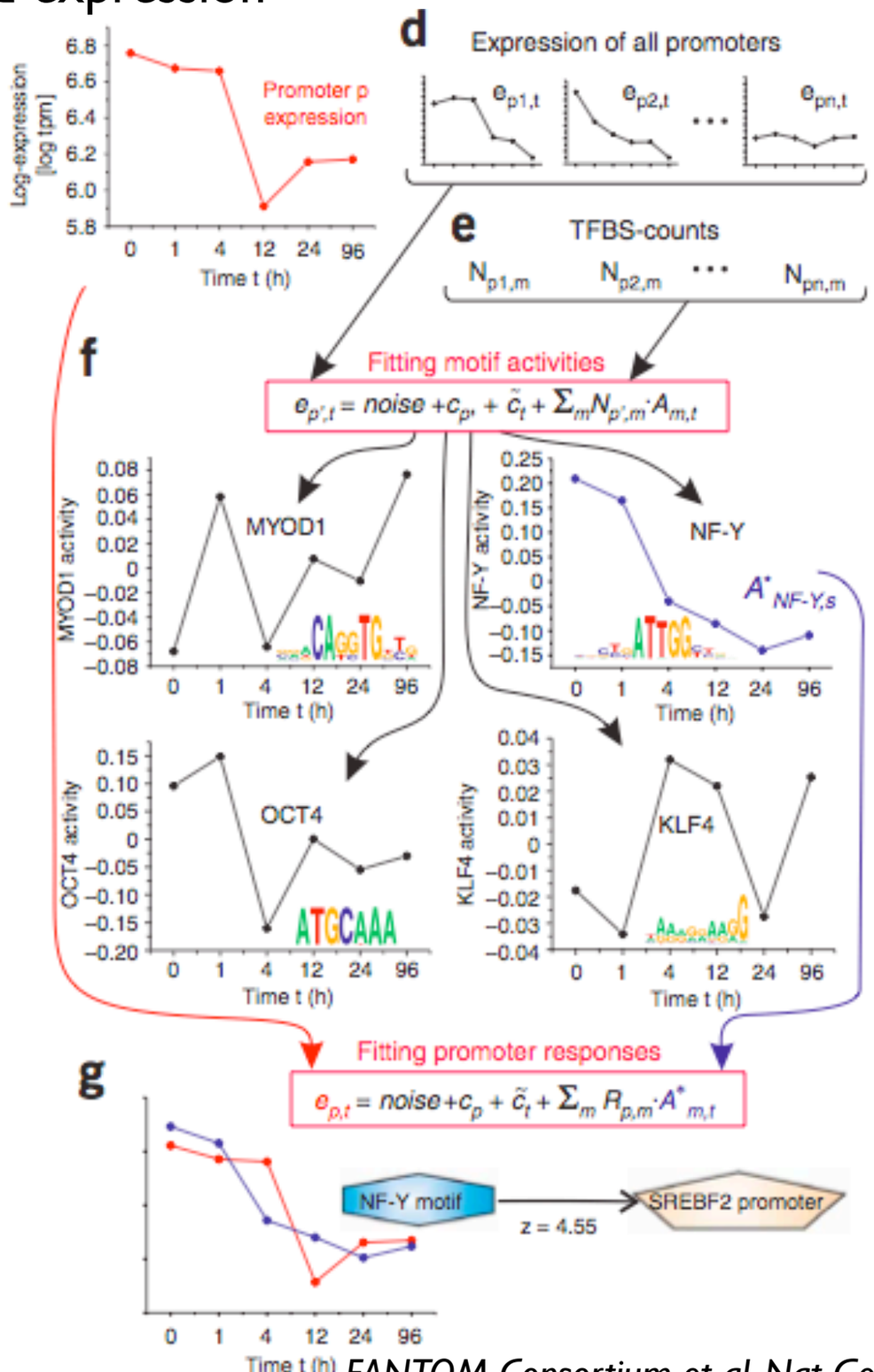
Segal et al. Nat Genet. 2003



MARA

- Motif Activity Response Analysis (E. van Nimwegen, U of Basel)
- Infer regulatory network from motifs, TF activity and expression

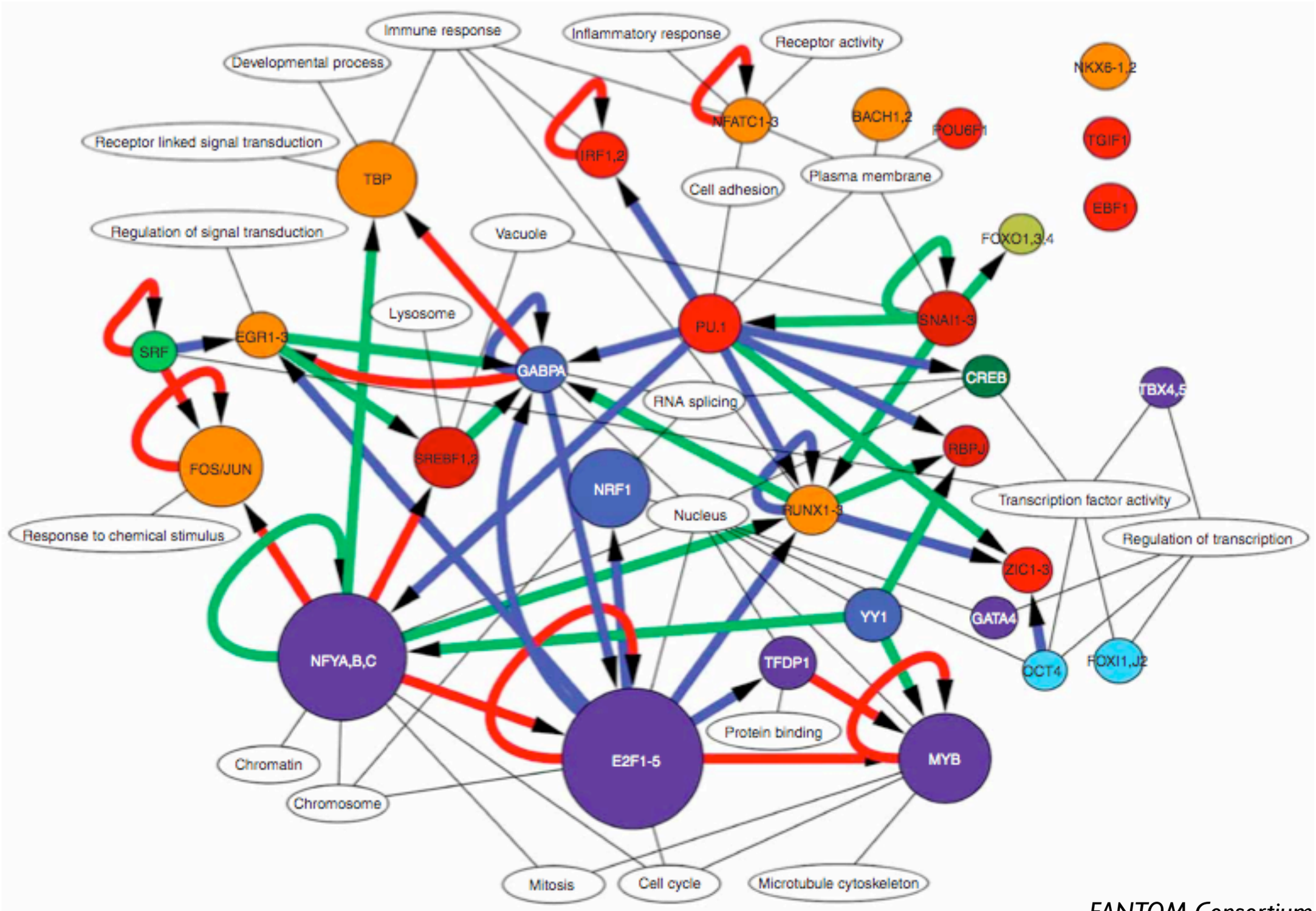




FANTOM Consortium et al. Nat Genet (2009)

MARA

- Motif Activity Response Analysis (E. van Nimwegen, U of Basel)
- Infer regulatory network from motifs, TF activity and expression



FANTOM Consortium et al. Nat Genet (2009)

Fisher combined test

- We collect multiple data of different types from separate experiments (ex.: ChIP-seq, RNA-seq, motif occurrence, etc.)
- We have a statistical procedure for each experiment, providing a p-value for each gene
- Can we gain power from mixing these data?

Fisher combined statistics: $$X^2 = -2\sum_{i=1}^{k}\log_e(p_i),$$

- This has a chi-square distribution with $2k$ degrees of freedom → single global p-value