Determinants of differential DNA binding specificity between closely related transcription factors

Ning Shen[1,2], Jingkang Zhao[2,3], Raluca Gordan[2,4]

[1]*Department of Pharmacology and Cancer Biology.* [2]*Center for Genomic and Computational Biology,* [3]*Program in Computational Biology and Bioinformatics,* [4]*Department of Biostatistics and Bioinformatics, Duke University*

ABSTRACT

Most eukaryotic transcription factors (TFs) are part of large protein families, with several TF family members (i.e. paralogous TFs) being expressed at the same time in the cell but targeting different sets of genes and performing different regulatory functions. Closely related TFs, with amino acid similarity of 70% or more in the DNA binding domain (DBD), are generally believed to have identical DNA binding specificities. However, their in vivo genomic binding patterns are markedly different. Currently, we do not have a good understanding of the general molecular mechanisms by which TFs with highly similar DBDs select distinct in vivo targets.

We show that, in general, closely related TFs interact differently with their putative genomic targets even in vitro, in the absence of any additional factors. Our study is focused on eleven paralogous factors from 4 protein families: bHLH, E2F, ETS, and RUNX. For each pair of related TFs, we used genomic-context protein-binding microarray (gcPBM) assays to compare the binding affinities of the two factors for ~25,000 putative genomic binding sites in vitro, in a cell-free environment where only naked DNA and purified TF are present. We find that for most pairs of paralogous TFs, the two factors interact differently with their genomic sites in vitro, despite having identical PWMs. The only two exceptions were: (1) E2F1 and E2F3, which play similar regulatory roles and can partially substitute for each other in the cell, and (2) Runx1 and Runx2, which are typically not expressed at the same time in the cell; in addition, Runx2 is known to compensate for the loss of Runx1 in leukemia cells, which is consistent with our finding that these paralogous TFs have identical specificities. To identify DNA sites with differential specificities for paralogous TFs, we developed a weighted regression-based approach that leverages measurements of experimental noise derived from replicate gcPBM experiments.

The way in which paralogous TFs differ is specific to each protein family: E2F1 and E2F4 prefer the same core GCGC/GCGG and differ in their flanking preferences for medium to high affinity sites, ETS factors ETS1 and ELK1 different in specificity for medium and low affinity sites, while bHLH factors c-Myc and Mad1 prefer different flanks for their highest affinity site CACGTG, and differ significantly in their affinity for alternative cores CACATG/CACGCG. Overall, we find that differences in genomic binding specificity between paralogous TFs are due both to direct recognition of DNA bases in the core binding sites (i.e. base readout), and to indirect recognition of different structural features in the flanking regions (i.e. shape readout). We use linear support vector models to identify sequence and structural determinants of specificity differences between paralogous TFs.

Importantly, the differences in intrinsic binding preferences between paralogous TFs, as identified in vitro by gcPBM, can partly explain differential in vivo binding, measured by ChIP-seq. While we cannot expect in vitro specificities to completely explain the in vivo binding patterns of paralogous TFs, our work shows that the intrinsic preferences of TFs for their genomic sites represent an important mechanism by which closely related factors achieve their regulatory specificity.