# Supplementary Information

# A high-resolution protein architecture of the budding yeast genome

Matthew J. Rossi[1], Prashant K. Kuntala[1], William K.M. Lai[1,2], Naomi Yamada[1], Nitika Badjatia[1], Chitvan Mittal[1,2], Guray Kuzu[1], Kylie Bocklund[1], Nina P. Farrell[1], Thomas R. Blanda[1], Joshua D. Mairose[1], Ann V. Basting[1], Katelyn S. Mistretta[1], David J. Rocco[1], Emily S. Perkinson[1], Gretta D. Kellogg[1,2], Shaun Mahony[1], B. Franklin Pugh[1,2]*

[1]Center for Eukaryotic Gene Regulation, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, Pennsylvania 16802, USA

[1,2]Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York, 14853, USA

*Correspondence: fp265@cornell.edu

# Supplementary Notes

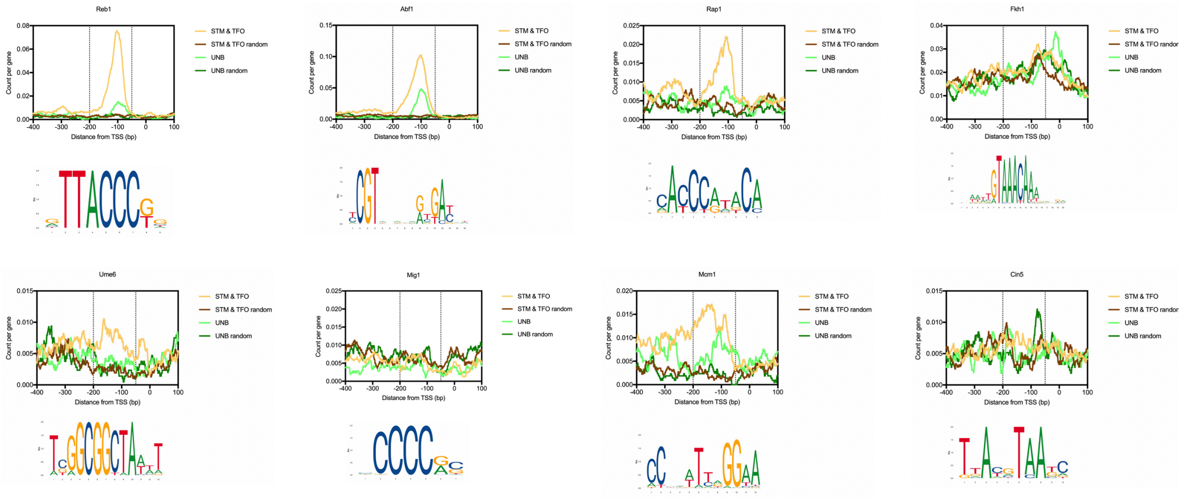
**Evidence that most UNB genes are unlikely to be bound by TFs or STM cofactors**

It is plausible that the 78 TFs that were detected by ChIP-exo at STM/TFO promoters were actually present at UNB promoters but fell below our threshold of detection (i.e., potentially false negatives). Alternatively, their binding might be environmentally condition-specific, and thus gone undetected in the rich media used in this study. It is plausible that other TFs that were not detectable at all or not assayed were actually bound to UNB promoters. We assess these potential false negatives in multiple ways.

**First**, if we assume that cofactors can be recruited only by site-specifically bound TFs, then promoters that contain cofactors but lack TFs, are likely to represent false negatives (i.e., missed TF detection). That is, we can use cofactor binding as a proxy for missing TF binding. We identified 74 targets as being cofactors (Supplementary Data $3_{1K}$). Using a liberal threshold, 20% of all promoters have at least one of these cofactors detectable by ChExMix in YPD at 25˚C. Only 10% of them (2% of all promoters) also lack a detectable TF. Thus, using the most relaxed criteria of having only a single cofactor event out of 74 possible cofactor events, we can expect that no more than 10% of all promoters have an undetected TF.

**Second**, we considered the use of a TF's cognate motif as a proxy for detecting missed TFs. We reasoned that the cognate motifs for these TFs should be enriched at canonical locations in promoter regions. JASPAR lists 66 families of TF motifs (http://jaspar.genereg.net/matrix-clusters/fungi/?detail=true; see Supplementary Data $2_{11}$), and so we searched for each. However, since most TFs individually bind to <30 genes site-specifically (which is <0.5% of all genes), we were concerned that motif occurrence would not be distinguishable from random background occurrences. We therefore empirically assessed the limits of motif detection by scanning promoter regions of sets of genes that had a recorded site-specific binding protein (i.e., STM and TFO, N=2,767; RP was excluded), thereby serving as positive controls or a gauge of sensitivity. These genes were scanned with each of the 66 motif families (using a PWM in FIMO)[1]. We also scanned with scrambled version of the motif, to serve as negative controls. The main motifs that were detectably enriched at their canonical location relative to its scrambled control were associated with Reb1, Abf1, Rap1, and Mcm1 (Supplementary Fig. 1). These represent the more abundant promoter motifs that also have higher motif complexity. The failure to convincingly detect known abundant motifs for many TFs (e.g. Cin5, Fkh1, and Tbf1) indicates that there is insufficient sensitivity to use motifs as a proxy for de novo prediction of TF binding for those that bind a small number of genes. Nonetheless, using Reb1 as an extreme example, we found ~2% of all UNB promoters had an unbound motif for Reb1 (and thus potential false negatives for TFO classification), compared to ~20% of all other genes having a bound Reb1 motif (i.e., the expected maximum at UNB promoters). Thus, the potential false negative rate for Reb1 as being incorrectly classified as UNB was no more than 10% (20%/2%). For those TF motifs that we could detect at STM and TFO promoters, we did not significantly detect them at UNB promoters, indicating that there is unlikely to be an abundance of UNB promoters that can bind TFs.

The one exceptional motif was poly(dA:dT) (generally defined as at least 5 A's in a row), which is abundant at UNB genes and involved in nucleosome organization. No TF motifs in JASPAR have poly(dA:dT). Four TFs (Azf1, Stb3, Sfp1, and Ntd80) have a motif that consists of 4-5 A's in a row, but together they only bind <5% of all genes.



Very large cluster that contains Tbf1 (cluster 1)

**Supplementary Fig. 1** | Scanning for TF motifs in UNB promoters. Shown are the results of performing a FIMO search on UNB promoter regions using nucleotide position weight matrix (PWM) for the 66 motif families in the JASPAR database. Representatives for eight of the most abundant TFs are shown. A weblogo of the PWM used in the search is presented below each plot. For each motif family, the “STM & TFO” promoters served as positive controls. TF motifs are expected to be concentrated 50-200 bp upstream of TSS. Scrambled versions of the PWM (“random”) serve as the negative controls. None of the 66 motif families were found to be enriched in UNB promoters relative to negative controls.

**Third**, we used MEME[2] to search UNB promoters for de novo motifs, but found no enriched motifs beyond poly(dA:dT) tracts. We looked downstream of the TSS, as one report has suggested for the Gcn4 TF[3]. However, we found no motif enrichment relative to negative controls.

Taken together, our assessment of the false negatives (missed TF binding) at UNB promoters is not likely to be above 10%. Therefore, we conclude that most UNB promoters lack TF/cofactor binding, and likely will not ever achieve such binding under any condition.

For TFO genes, <25% contain a bound TF that is more typically found at STM promoters. These TFO promoters may have been algorithmically misclassified, possibly falling below the threshold of STM cofactor detection, or being condition-specific. With ~10% uncertainty of UNB and ~25% uncertainty of TFO, we estimate that 65-70% of all genes evolved a constitutive promoter architecture that lacks TF/STM-cofactor interactions.

1 Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017-1018 (2011).

2 Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* **2**, 28-36 (1994).

3 Rawal, Y. *et al.* Gcn4 Binding in Coding Regions Can Activate Internal and Canonical 5' Promoters in Yeast. *Mol Cell* **70**, 297-311 e294 (2018).