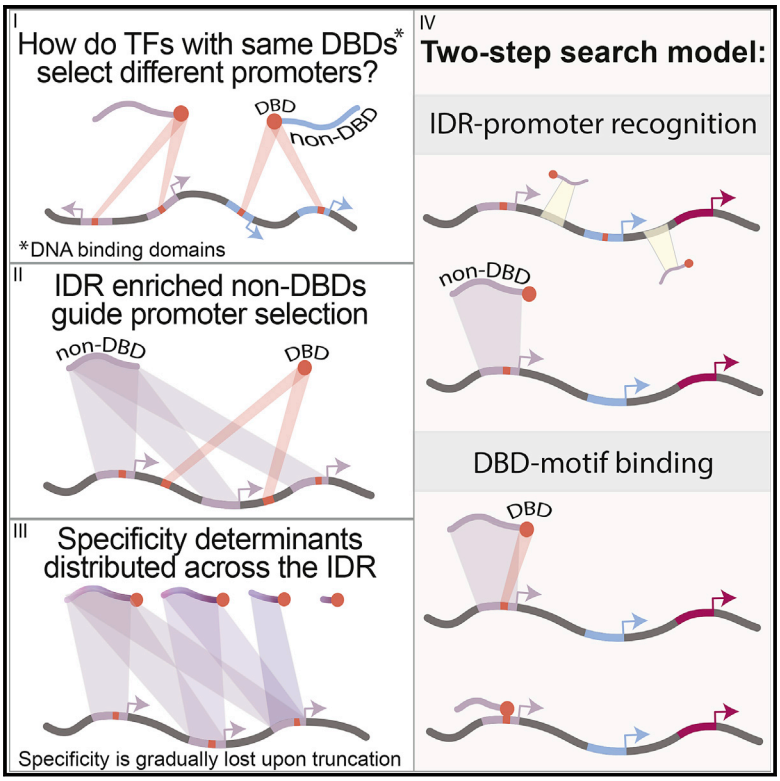
Molecular Cell

Article

Intrinsically Disordered Regions Direct Transcription Factor *In Vivo* Binding Specificity

Graphical Abstract



Authors

Sagie Brodsky, Tamar Jana, Karin Mittelman, Michal Chapal, Divya Krishna Kumar, Miri Carmi, Naama Barkai

Correspondence

naama.barkai@weizmann.ac.il

In Brief

Transcription factors (TFs) bind only a subset of their motif-containing promoters. Brodsky and Jana et al. find that promoter selection is dictated by multiple specificity determinants distributed across long intrinsically disordered regions (IDRs) encoded within the TF sequence. IDR-guided specificity may accelerate binding-site recognition by an initial detection of broad DNA regions.

Highlights

- Transcription factors (TFs) bind a selected subset of their motif-containing promoters
- Long intrinsically disordered regions (IDRs) guide TF promoter selection *in vivo*
- Promoter recognition depends on multiple determinants distributed along the entire IDR
- IDRs may accelerate the TF search for target promoters by recognizing broad DNA regions

Brodsky et al., 2020, Molecular Cell *79*, 1–13
August 6, 2020
https://doi.org/10.1016/j.molcel.2020.05.032

CellPress

Please cite this article in press as: Brodsky et al., Intrinsically Disordered Regions Direct Transcription Factor *In Vivo* Binding Specificity, Molecular Cell (2020), https://doi.org/10.1016/j.molcel.2020.05.032

Molecular Cell

CellPress

Article

Intrinsically Disordered Regions Direct Transcription Factor *In Vivo* Binding Specificity

Sagie Brodsky,[1,3] Tamar Jana,[1,3] Karin Mittelman,[1,2] Michal Chapal,[1] Divya Krishna Kumar,[1] Miri Carmi,[1] and Naama Barkai[1,4,*]

[1]Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel

[2]Present address: School of Molecular Cell Biology and Biotechnology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 6997801, Israel

[3]These authors contributed equally

[4]Lead Contact

*Correspondence: naama.barkai@weizmann.ac.il

https://doi.org/10.1016/j.molcel.2020.05.032

SUMMARY

Transcription factors (TFs) that bind common DNA motifs *in vitro* occupy distinct sets of promoters *in vivo*, raising the question of how binding specificity is achieved. TFs are enriched with intrinsically disordered regions (IDRs). Such regions commonly form promiscuous interactions, yet their unique properties might also benefit specific binding-site selection. We examine this using Msn2 and Yap1, TFs of distinct families that contain long IDRs outside their DNA-binding domains. We find that these IDRs are both necessary and sufficient for localizing to the majority of target promoters. This IDR-directed binding does not depend on any localized domain but results from a multitude of weak determinants distributed throughout the entire IDR sequence. Furthermore, IDR specificity is conserved between distant orthologs, suggesting direct interaction with multiple promoters. We propose that distribution of sensing determinants along extended IDRs accelerates binding-site detection by rapidly localizing TFs to broad DNA regions surrounding these sites.

INTRODUCTION

Transcription factors (TFs) bind with high affinity to short DNA motifs, typically consisting of 6–12 bp. These motifs are highly abundant in the genome, yet most of them remain unoccupied by the respective TFs. The ability of TFs to bind a selected subset of their potential binding sites is important for their *in vivo* function, as it allows for specific binding of their target promoters (Inukai et al., 2017; Lambert et al., 2018). In some genomic regions, binding motifs are generally inaccessible due to the chromatin arrangement (Li et al., 2007). However, inaccessible chromatin cannot explain how related TFs that bind the same motif *in vitro* bind different subsets of this motif-containing sites *in vivo*. Differences in the *in vivo* binding specificity must therefore result from properties encoded within the TF itself. The DNA-binding domain (DBD), for example, could interact with DNA sequences flanking the core motif (Levo and Segal, 2014; Shen et al., 2018). Alternatively, regions outside the DBD could direct the TF to specific promoters by interacting with other DNA-bound co-factors (Morgunova and Taipale, 2017; Shively et al., 2019).

Interactions between proteins depend on geometrical compatibility and noncovalent forces between folded structures. TFs, however, often contain intrinsically disordered regions (IDRs), which are characterized by distinct patterns of interactions (Fuxreiter et al., 2011; Guo et al., 2012; Liu et al., 2006; Minezaki et al., 2006). First, IDRs can form a lock-and-key type of interactions when folding upon binding (Habchi et al., 2014; Wright and Dyson, 2009). Alternatively, disordered domains can interact via long-range electrostatic attractions. This latter type of interactions is often promiscuous (Borgia et al., 2018; Protter et al., 2018) but can still exhibit a wide spectrum of affinities and specificities (Arbesú et al., 2018; Gao et al., 2018). Furthermore, these interactions are commonly multivalent due to the repetitive nature of intrinsically disordered sequences (Tompa, 2003).

The enrichment of IDRs within TFs suggests that properties unique to IDRs may play a role in transcription. One such role may be the formation of phase condensates, which depends on weak multivalent interactions such as those mediated by IDRs (Elbaum-Garfinkle et al., 2015; Wang et al., 2018). In the context of transcription, interactions between IDRs of specific TFs and of Med15, a subunit of the transcriptional mediator, can nucleate condensates that concentrate the transcription apparatus to specific enhancers (Boija et al., 2018; Sabari et al., 2018). Other studies proposed that short IDR segments within TF DBDs can accelerate binding-site search, for example by transferring the DBD between DNA strands (Shoemaker et al., 2000; Vuzman et al., 2010). Finally, it was also proposed that IDRs can directly detect specific DNA sequences, thereby contributing to binding site selection (Guo et al., 2012).

Please cite this article in press as: Brodsky et al., Intrinsically Disordered Regions Direct Transcription Factor *In Vivo* Binding Specificity, Molecular Cell (2020), https://doi.org/10.1016/j.molcel.2020.05.032

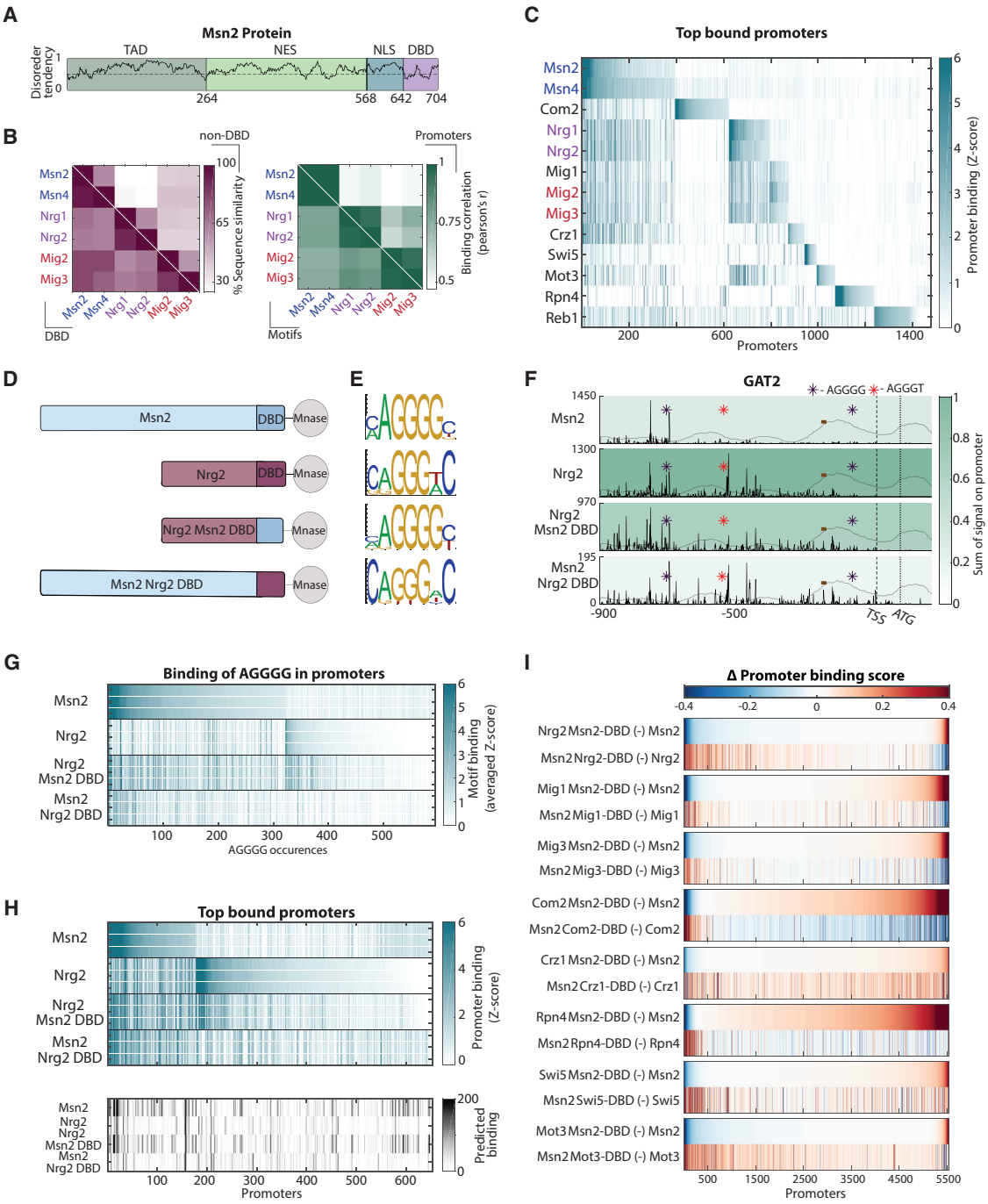


Figure 1. Differences in Binding Profiles of Related TFs Depend on Sequences Outside Their DBD

(A) Predicted disorder tendency of Msn2: the black line indicates the predicted disorder tendency along the Msn2 protein, as calculated by IUPred, with values above 0.5 considered disordered (Dosztányi et al., 2005a, 2005b). Shown are the positions of the transcriptional activation domain (TAD), nuclear export signal (NES), nuclear localization signal (NLS), and DBD (Sadeh et al., 2012).

(B) Similarities between TF paralogs: shown are sequence similarities (left) and binding profiles (right) of three pairs of zinc-finger TF paralogs that arose from the whole genome hybridization (WGH) event that had occurred ~100 million years ago (Wolfe and Shields, 1997; pairs are indicated by different colors). Sequence similarity was calculated separately for the DBDs (left, lower triangle) and the rest of the protein (left, upper triangle). Similarity in binding profiles was quantified by correlation in motif preferences (right, lower triangle) or promoter binding (right, upper triangle). Note that close paralogs have a conserved DBD, prefer the same motifs, and select indistinguishable sets of promoters. Sequences outside the DBDs, however, show little correspondence between close paralogs.

(C) TFs with similar motif preferences bind different promoter sets: promoters bound by at least one of the TFs in our data were selected and ordered by binding strength (see STAR Methods). Shown is the median binding strength (in units of *Z* score) of each TF to each of the 1,450 selected promoters (the number of

(legend continued on next page)

Please cite this article in press as: Brodsky et al., Intrinsically Disordered Regions Direct Transcription Factor *In Vivo* Binding Specificity, Molecular Cell (2020), https://doi.org/10.1016/j.molcel.2020.05.032

Experimental evidence supporting this latter postulated contribution of IDRs to *in vivo* binding is not yet available.

To examine for a possible role of IDRs in directing TF binding-site selection, we considered Msn2 and Yap1, two budding yeast TFs that contain extended IDRs (>500 aa). We found that the DBDs of both TFs are neither sufficient nor required for their localization to the majority of their target promoters. Rather, binding to most target promoters depends additively on a large number of weak and partially redundant determinants distributed throughout their entire IDRs. We also found that this IDR-directed binding remains conserved between distant orthologs that show little sequence similarity. Compensating mutations, therefore, are restricted to the IDR itself, as expected if the IDR directly interacts with multiple promoters, rather than a single recruiting factor. Based on our results, we suggest that IDRs recognize specific DNA regions through the formation of multiple low-affinity interactions with the DNA or the surrounding histones. Therefore, while the DBD interacts with DNA through well-defined localized regions, the non-DBD may comply with the distributed sensing paradigm, with recognition determinants spread throughout its extended IDR. We discuss the possible benefit of this design in allowing rapid detection of targets within large genomes.

RESULTS

DBDs Are Not Sufficient for Explaining Differences in the *In Vivo* Binding Specificity of Related TFs

To examine the possible role of IDRs in guiding the TF binding pattern *in vivo*, we first focused on Msn2, a stress-activated zinc-finger TF that regulates dozens of stress-responsive genes in budding yeast (Gasch et al., 2000; Martínez-Pastor et al., 1996) Msn2 is composed of 704 aa, 62 of which define its DBD, while the rest of its sequence is predicted to be mostly intrinsically disordered (Figure 1A; Table S1). Msn2 binds preferentially to the AGGGG motif both *in vivo* and *in vitro* (MacIsaac et al., 2006; Siggers et al., 2014). Since variants of this motif are bound by additional zinc-finger TFs, associated with different functions, we wished to compare the *in vivo* binding of these TFs. In particular, we asked whether these TFs bind to the same subset or different subsets of AGGGG-containing sites. To this end, we used chromatin endonuclease cleavage followed by sequencing (ChEC-seq; Zentner et al., 2015) and characterized the binding profiles of Msn2 and 11 other zinc-finger TFs, seven of which show similar motif preferences to that of Msn2 *in vitro* (MacIsaac et al., 2006; Persikov and Singh, 2014; Zhao et al., 2009; Figure S1A).

Binding profiles varied between the tested TFs in both motif preferences and promoter selection (Figures 1B, 1C, and S1A–S1D). Furthermore, distinct subsets of AGGGG motif-containing sites were bound by the different TFs, including sites that are not occupied by Msn2 (Figure S1B). The binding profiles of close paralogs, however, were practically identical. Since close paralogs have conserved DBDs but show little sequence similarity outside this domain (Figure 1B), we predicted that the distinct binding profiles of the more distant TFs depend on differences encoded within their DBDs. To test this, we swapped DBDs between the assayed TFs (Figure 1D). This assay confirmed that the preferred DNA-motif is defined by the DBD (Figures 1E, 1F, and S2A). However, contrasting our prediction, the selection of motif-containing target promoters was in fact largely dependent on the non-DBD (Figures 1F–1I, S2B and S2C). For example, swapping the Msn2 DBD with that of Rpn4, still maintained 50% of the Msn2-bound promoters, whereas swapping the non-DBDs abolished binding to 70% of these promoters (Figure S2C). In fact, for all examined cases, sequences outside the DBD are important for directing TFs toward specific subsets of their motif-containing sites. Furthermore, in complementary swapping experiments, swapped TFs tended to gain access to promoters that are lost from the reciprocal swap (Figure 1I), further supporting the role of the non-DBD in directing promoter selection.

The DBD Is Not Required for Msn2 Localization to Its Target Promoters

We next wished to examine more directly whether the DBD is required for targeting Msn2 to its selected binding sites. To this end, we considered two Msn2 mutants, one containing only the DBD, and the other lacking the DBD (non-DBD; Figure 2A). The DBD-only mutant localized to sites containing the Msn2 binding motif, as expected (Figures 2B and 2C). However, it selected a distinct subset of motif-containing sites (Figure 2D). Overall, promoter binding signal was only moderately correlated

repeats for each TF is listed in Table S2). Note the distinct promoter set of Com2, whose DBD is highly similar to that of Msn2/4 (Siggers et al., 2014). Reb1, a myb-family TF, is shown as an outer family control.

(D) Swapping DBDs between TFs: a scheme exemplified by Msn2-Nrg2 DBD swapping.

(E) Motif preferences of swapped TFs are defined by their DBDs: shown are the position weight matrices (PWMs) for the DNA motifs preferentially bound by Msn2, Nrg2, and their DBD-swapped factors (see also Figure S2A and STAR Methods).

(F) Binding to the GAT2 promoter: shown is the median binding signal of the indicated factors to the GAT2 promoter. Gray line indicates nucleosome occupancy. TATA-box is marked by a brown rectangle. The Msn2- and Nrg2-preferred motifs (AGGGG and AGGGT respectively) are also indicated. The overall signal on the promoter is shown as background color, normalized to the top bound promoter of each TF. Note that the overall promoter binding strength of the swapped factors better matches that of their non-DBD donor, while their localization within the promoter better corresponds to their DBD donor.

(G and H) Differences between Nrg2 and Msn2 DBDs do not explain their differential binding patterns: the binding strength of the indicated factors to the top bound ~600 AGGGG motif occurrences within promoters (Z score > 1 in at least one of the factors) is shown in (G), as well as the overall binding strength to 630 target promoters bound by at least one factor (H; see STAR Methods). Three independent repeats are shown for each factor, separated by white lines. Predicted promoter binding strength, based on motif preference, is also shown (bottom panel; see STAR Methods).

(I) Reciprocal loss/gain of binding in complementary swapping experiments: the eight complementary swaps of Msn2 in our dataset were considered, as indicated. First, promoter binding strength of all factors were normalized between 0 and 1. Then, we calculated the change in binding strength of the wild-type TF upon swapping its non-DBD with that of the other TF. These values are plotted in the two indicated lines, ordered by the change value in the upper line. Note that swapped TFs gain access to sites that are lost from their reciprocal swap.

Please cite this article in press as: Brodsky et al., Intrinsically Disordered Regions Direct Transcription Factor *In Vivo* Binding Specificity, Molecular Cell (2020), https://doi.org/10.1016/j.molcel.2020.05.032

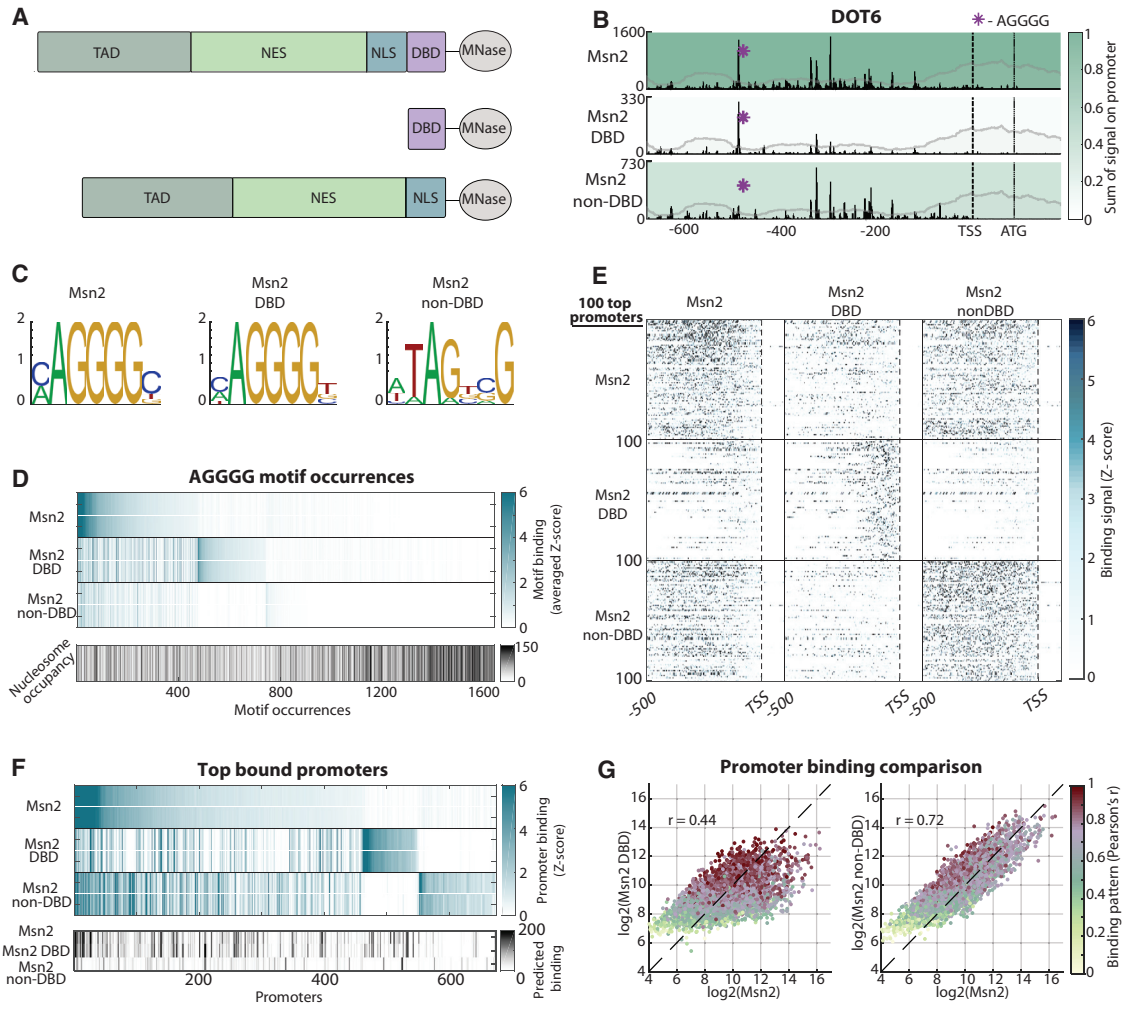


Figure 2. Msn2 Promoter Selection Is Independent of Its DBD

(A) Msn2 mutants used in the analysis. Please see Figures S3A–S3C for similar experiments controlling for the duration of MNase activation.

(B) Binding to the DOT6 promoter: same presentation as in Figure 1F for the indicated factors. Note that deletion of the DBD abolished binding to the Msn2-preferred motif but had a little effect on the total promoter binding.

(C) Localization of Msn2 to its preferred motif requires its DBD: shown are the PWMs summarizing the DNA-motifs preferentially bound by Msn2 and the indicated mutants (see STAR Methods).

(D) The DBD is not sufficient for selecting the Msn2-binding sites *in vivo:* the upper panel shows the binding strength (in units of *Z* score) of the indicated factors to all AGGGG motif occurrences in promoters (~1,600). Two independent repeats are shown for each factor, separated by white lines. Lower panel indicates the nucleosome occupancy around each site (see STAR Methods). Note the loss of motif binding upon DBD deletion, the distinct set of motif-containing sites bound by the DBD-only mutant, and the tendency of all factors to bind regions of relatively low nucleosome occupancy.

(E–G) Msn2 recognizes its target promoters even in the absence of its DBD: the pattern by which each of the indicated factors binds their top target promoters is shown in (E). Each box shows the binding pattern of the TF (indicated on top) to the 100 most bound promoters of the factor indicated on the left. Each row within a box represents one promoter (500 bp), aligned by the transcriptional start site (TSS; dashed line), with color indicating the median binding signal (in units of *Z* score) at a given position. (F) Shows the binding strength of each of the indicated factors to promoters bound by at least one of these factors (see STAR Methods) using the same display as in Figure 1H. White lines distinguish two independent repeats. Binding strength of the factors to all annotated promoters is shown in (G), where each dot represents a promoter. Dots are color coded based on the correlation of binding signal along the promoter. Note that the DBD defines the binding pattern within promoters but contributes little to the overall promoter binding strength.

between the DBD-only mutant and the intact Msn2 (Pearson's r = 0.44).

The DBD of Msn2 is therefore not sufficient for detecting the Msn2 target promoters. We next asked whether the DBD is required for this recognition. As expected, deleting the DBD abolished the Msn2 preference for binding AGGGG-containing sites, and the mutant did not show preference for any other known *cis*-regulatory motif (Figures 2C, 2D, and S3A). Notably, this mutant did localize to most of the Msn2-bound promoters (Figures 2E–2G and S3B–S3D; r = 0.72). Accordingly, while not localizing to AGGGG-containing sites, the non-DBD mutant showed a strong tendency for binding promoters that contain

Please cite this article in press as: Brodsky et al., Intrinsically Disordered Regions Direct Transcription Factor *In Vivo* Binding Specificity, Molecular Cell (2020), https://doi.org/10.1016/j.molcel.2020.05.032

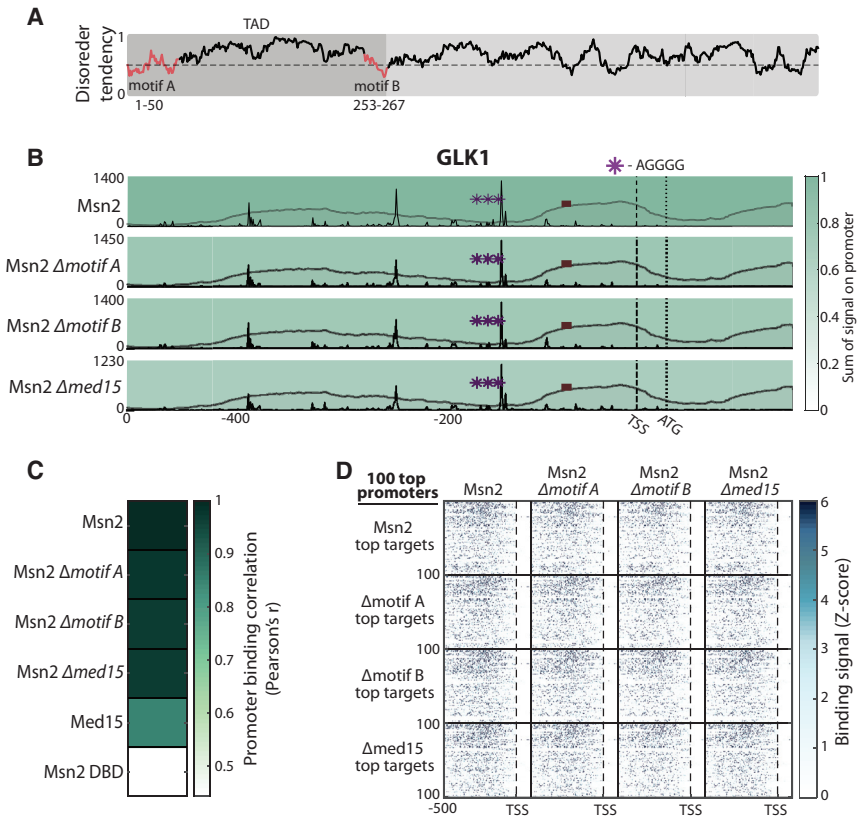


Figure 3. Msn2 Binding Specificity Is Independent of the Transcriptional Mediator

(A) High-complexity sequences within the Msn2 TAD are required for binding the transcriptional mediator: a scheme of Msn2, same presentation as in Figure 1A. Indicated in red are the two structured motifs required for mediator binding (Sadeh et al., 2012).

(B) Msn2 binding to the GLK1 promoter is independent of the transcriptional mediator (same presentation as in Figure 1F for the indicated factors).

(C and D) Binding of Msn2 to its target promoters is independent of the transcriptional mediator: shown in (C) are the promoter binding correlations between the indicated factors and the wild-type Msn2. Binding patterns along the top-bound promoters are shown in (D) (same presentation as in Figure 2E). Note the overlap in binding preferences of Med15 and Msn2.

the AGGGG motif (Figures S3E and S3F). Therefore, the DBD is not required for the localization of Msn2 to the majority of its target promoters.

To verify the specificity of this DBD-independent binding of Msn2, we considered Nrg2, a zinc-finger TF that is similar to Msn2 in terms of its *in vitro* binding preferences (Fordyce et al., 2010; Figure S1A) yet binds a different set of target promoters (Figures 1B-C). Similar to Msn2, deletion of the Nrg2 DBD abolished its ability to bind its preferred motif (Figures S3G and S3H), yet it had a minor effect on its ability to localize to its target promoters (Figure S3I; r = 0.88). Further, differences in binding profiles between the wild-type Msn2 and Nrg2 remained when comparing the respective DBD-deleted mutants (Figure S3J). We conclude that while the DBD defines the precise motif to which Msn2 and Nrg2 bind, it is not required for their localization to the majority of their target promoters. Therefore, for both TFs, sequences outside the DBD direct *in vivo* promoter selection.

Msn2 Binding Specificity Is Independent of Its Interaction with Med15

Our results above show that sequences outside the Msn2 DBD direct its binding specificity *in vivo*. Binding of Msn2 to DNA-bound co-factors may explain such recruitment. A likely candidate is Med15, a component of the transcriptional mediator that was shown to interact with the Msn2 transcriptional activation domain (TAD; Figure 3A; Sadeh et al., 2012). Furthermore, Med15 is required for the incorporation of several TFs into phase condensates (Boija et al., 2018), although this was not shown for Msn2 which does not appear to form condensates upon activation (Chowdhary et al., 2019). By mapping the Med15 binding profile, we found that its promoter selection overlaps with that of Msn2 (Pearson's r = 0.84). Still, deletion of Med15 did not change the *in vivo* binding profile of Msn2, ruling out the possibility that Med15 directs Msn2 to its selected binding sites (Figures 3B–3D).

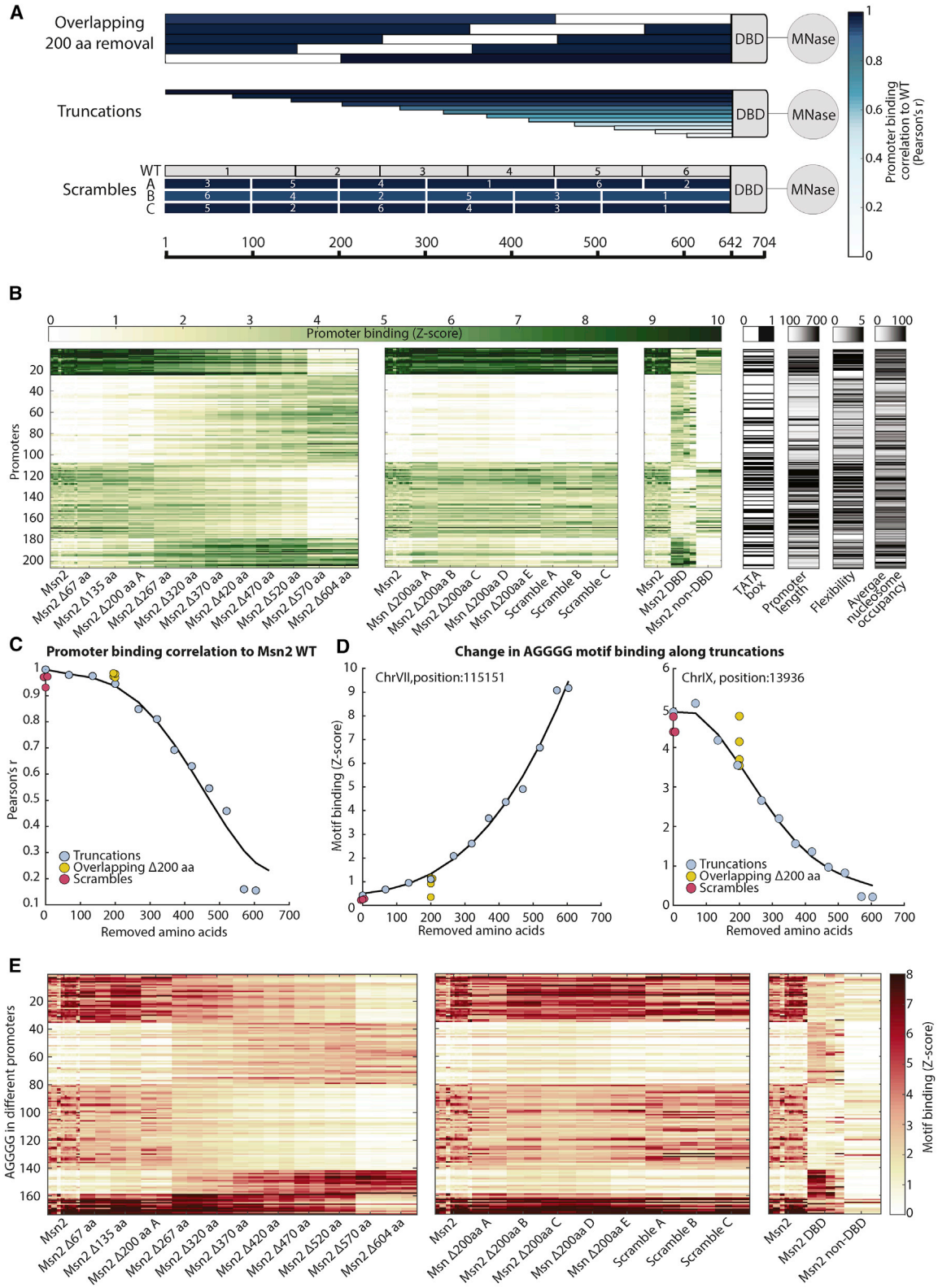
Previous studies detected the Msn2 regions required for its interaction with Med15 by searching for structured motifs embedded within its largely disordered TAD (Sadeh et al., 2012). Two such motifs were identified and were shown to be required both for Med15 binding and Msn2-dependent activation (Figure 3A). We reasoned that these same regions might also interact with other, yet-unidentified co-factors. Mutating either sequence, however, did not have any effect on the binding pattern of Msn2 (Figures 3B–3D). Therefore, neither the transcriptional mediator nor the structured motifs within the TAD of Msn2 are required for its localization to its target promoters.

The Msn2 IDR Directs Promoter Binding by Multiple, Weak, and Partially Redundant Determinants Distributed throughout Its Sequence

To define specific regions within the Msn2 protein that are required for directing its promoter selection, we deleted overlapping segments of 200 aa spanning the entire protein, excluding its DBD (Figure 4A). These deletions had only minor effects on the Msn2 promoter selection (Figures 4A–4C and S4A; Pearson's r = 0.92–0.98).

The ability to remove overlapping regions of 200 aa without altering the genomic binding profile suggests that Msn2 recognizes its specific promoters using multiple redundant determinants. To characterize this further, we generated a series of truncation mutants, in which we sequentially removed ~50 aa from the Msn2 N terminus (excluding its DBD; Figure 4A). The binding profile began to gradually change when removing >200 residues,

Please cite this article in press as: Brodsky et al., Intrinsically Disordered Regions Direct Transcription Factor *In Vivo* Binding Specificity, Molecular Cell (2020), https://doi.org/10.1016/j.molcel.2020.05.032



(legend on next page)

Please cite this article in press as: Brodsky et al., Intrinsically Disordered Regions Direct Transcription Factor *In Vivo* Binding Specificity, Molecular Cell (2020), https://doi.org/10.1016/j.molcel.2020.05.032

with each further truncation having an additional moderate effect (Figures 4A–4C and S4A).

Examining the type of changes introduced by the gradual truncations, we found both loss and gain of binding sites. Binding was gradually lost from most AGGGG-containing sites that were bound by the wild-type Msn2, and, in parallel, a new subset of AGGGG-containing sites became increasingly bound (Figures 4D, 4E, and S4B). Of note, sites that were gained upon truncation showed distinct properties compared to sites that were bound by the wild-type Msn2. The new binding sites gained by the truncations were mostly in short promoters. By contrast, promoters favored by the full protein were longer, enriched with TATA box sequences, occupied by nucleosomes, and highly responsive to perturbations (Figure 4B; statistical analysis in Figure S4C).

To examine whether the order of the IDR sequence contributes to promoter selection, we next scrambled the Msn2 sequence, splitting it into six ~100-aa blocks and reassembling them in different orders (Figure 4A). None of these scrambling perturbations, however, had a significant consequence on Msn2 promoter selection (Figures 4A–4C and S4A; Pearson's r = 0.93–0.97). We conclude that the binding of Msn2 to its target promoters depends on a large number of partially redundant specificity determinants distributed throughout its IDR. Each individual determinant exerts only a weak effect, but their additive contribution allows specific Msn2 binding to its target promoters.

IDR-Directed Promoter Recognition Is Conserved between Distant Orthologs

The finding that multiple sequence elements within the Msn2 IDR are required for directing its promoter selection was surprising to us, since we noted little conservation of this sequence along evolution. This is particularly striking when comparing the conservation of this region with that of the Msn2 DBD, which is highly conserved between orthologs (Figures 5A and 5B). This tight sequence conservation of TF DBDs likely results from their direct interaction with multiple promoters; a mutation perturbing DBD binding is unlikely to be compensated for at the level of its multiple target sites, favoring reversion by a *cis* mutation within the DBD itself. This is in contrast to sequences mediating protein-protein interactions, where compensating mutations can result in the co-evolution of the two proteins.

While showing little sequence conservation, the non-DBDs of the Msn2 orthologs were all predicted to be mostly intrinsically disordered (Figure S5A). To examine the consequence of non-DBD sequence divergence on the Msn2 *in vivo* specificity, we swapped the *S. cerevisiae* Msn2 non-DBD with the corresponding domains taken from its orthologs, or from orthologs of its close paralog, Msn4, which binds to practically identical positions (Figures 1B and 1C). Except for two cases, all orthologs retained the *S. cerevisiae* Msn2 binding patterns (Figures 5C–5E). The large sequence differences between these homologs that maintained the same *in vivo* specificity raised the possibility that it is protein size rather than protein sequence that defines the binding pattern. This, however, appears unlikely, as changes in binding patterns were not associated with changes in size, either when swapping non-DBDs between orthologs or when swapping non-DBDs of different *S. cerevisiae* paralogs (Figures S5B–S5E). Alternatively, conservation of ortholog binding specificity indicates of redundancy in the specificity code, allowing multiple mutations to accumulate in the non-DBD without altering its ability to properly direct promoter selectivity.

Yap1 Selects Its Binding Sites Using the Same Strategy as Msn2

To examine whether our results generalize to other TF families, we considered Yap1, a basic leucine zipper (bZIP) TF, the master regulator of the budding yeast oxidative stress response (Rodrigues-Pousada et al., 2010). Similarly to Msn2, Yap1 also contains a long IDR (>500 aa; Figure 6A). Yap1 promoter selection differs from that of six other bZIP TFs with similar motif preferences (Gordân et al., 2011; Figures S6A and S6B). DBD swapping between these TFs verified that the different binding profiles are not explained by differences in their DBD sequences (Figure S6C). Consistently, truncated Yap1 mutants, containing only its DBD or its non-DBD, confirmed that the DBD is neither sufficient nor required for localizing Yap1 to the majority of its target promoters (Figures S6D–S6H). Therefore, similarly to Msn2, the recruitment of Yap1 to its binding sites depends on sequences in its non-DBD.

Yap1 interacts with another TF, Skn7, which shares many of its target promoters with Yap1 (Lee et al., 1999; Mulford and Fassler, 2011). Accordingly, Yap1 binding sites are enriched with its own preferred motif (TTAGT[A/C]A), and the Skn7 binding motif (GGCCGNC; Figures 6B and S6I). Preventing the Yap1-Skn7 interaction by deleting either Skn7 or a domain in Yap1 required for this interaction (Mulford and Fassler, 2011) abolished the Yap1 localization at Skn7 binding sites (Figures 6B and S6I).

Figure 4. Specificity Determinants Are Distributed across the Entire IDR of Msn2

(A) Msn2 mutant strains: mutants analyzed include deletions of overlapping 200-aa regions spread across the Msn2 sequence (excluding its DBD), gradual ~50-aa N-terminal truncations, and scrambling of 100-aa blocks. Color intensity quantifies the correlation between the promoter binding profiles of the respective mutants to that of the wild-type Msn2.

(B and C) N-terminal truncation gradually shifts Msn2 promoter preferences: Promoters bound by the wild type or at least one of the truncation mutants were chosen (Z score > 3.5) and clustered according to binding strength along truncations (B, green panel). For each factor, we show all independent repeats in our dataset (see Table S2). Presence of TATA-box, length, expression flexibility, and average nucleosome occupancy are shown for each promoter (gray panel). Note that the wild-type protein binds preferentially to long, flexible, TATA-box-containing gene promoters of relatively high nucleosome occupancy (see Figure S4C for statistical analysis). The overall correlations of genome-wide promoter selection between the different mutants and the wild type are shown in (C) (see also Figure S4A). Note that neither scrambling of the Msn2 long IDR nor removal of overlapping 200-aa segments had a significant effect on the Msn2 binding profile.

(D and E) N-terminal truncation results in both gain and loss of binding sites: binding strength (in units of Z score) to two representative AGGGG motifs are shown as a function of the number of amino acids removed from Msn2 (D) (see also Figure S4B). AGGGG-containing sites bound by the wild type or at least one of the truncation mutants (averaged Z score > 3.5) were chosen and clustered according to binding strength along truncations (E). For each factor, all independent repeats are shown (see Table S2).

Please cite this article in press as: Brodsky et al., Intrinsically Disordered Regions Direct Transcription Factor *In Vivo* Binding Specificity, Molecular Cell (2020), https://doi.org/10.1016/j.molcel.2020.05.032

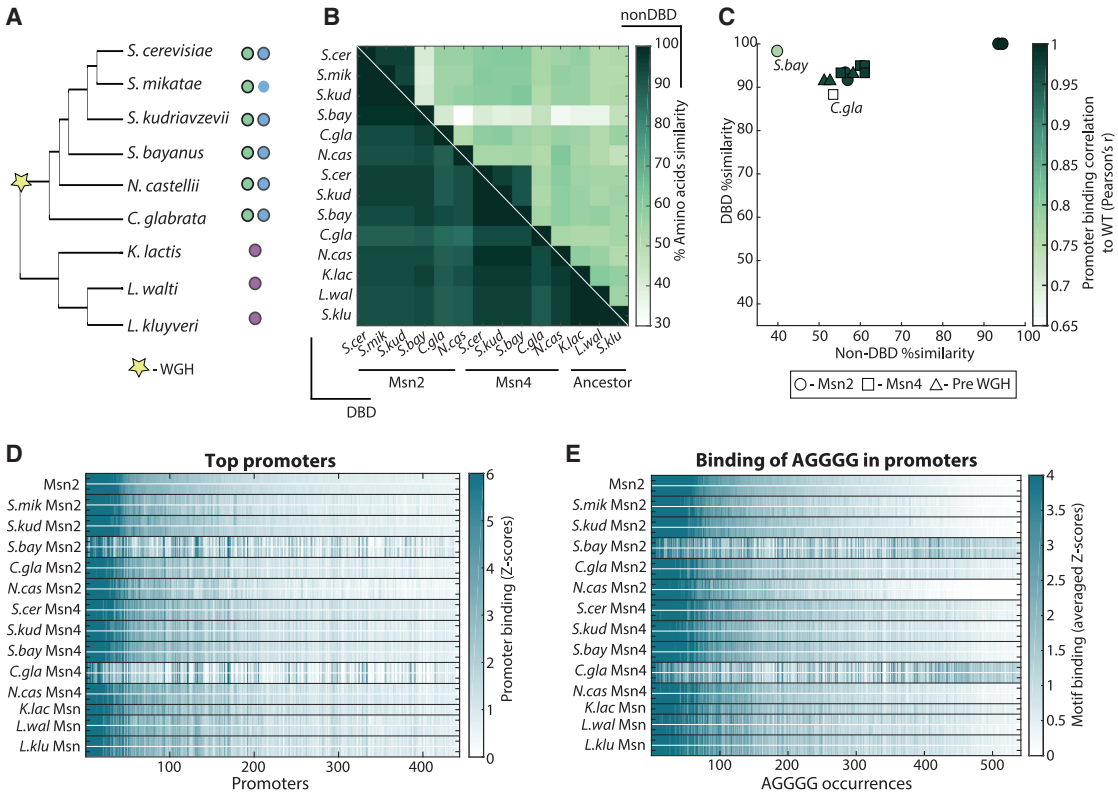


Figure 5. IDR-Directed Specificity Is Conserved between Distant Orthologs

(A) Species chosen for the analysis: orthologs of Msn2, Msn4, and the ancestral Msn are indicated by colors (green, blue, and purple, respectively). The non-DBD of Msn2 was replaced with non-DBDs of the orthologs marked by black circles. The star indicates the WGH event (Wolfe and Shields, 1997). Species phylogenetic tree is based on Shen et al. (2016).

(B) Sequence similarity between Msn2 and Msn4 orthologs: shown are the amino acid similarities measured for DBDs (lower triangle) and the rest of the protein (upper triangle). Note the low sequence similarity of the non-DBDs compared to the highly conserved DBDs. Still, the non-DBD sequences of Msn2 orthologs retained their low-complexity, intrinsically disordered profile (Figure S5A).

(C–E) Distant orthologs maintain conserved promoter preferences: shown in (C) is the promoter selection correlation between the Msn2/Msn4 swapped factors to the *S. cerevisiae* Msn2. The binding strength of all mutants to target promoters bound by at least one factor are shown in (D) (see STAR Methods), while their binding to AGGGG-containing sites is shown in (E) (repeats are separated by white lines). Note that the two orthologs that lost the conserved preferences are of relatively close species (Figure 5A).

Surprisingly, however, these deletions had only a minor effect on the Yap1 promoter selection (Pearson's r = 0.9, 0.91, respectively; Figures 6C and 6D). Therefore, while Skn7 affects the precise binding position of Yap1 within its target promoters, it plays a minor role in recruiting Yap1 to these promoters.

To examine for sequences within Yap1 that are required for its promoter recruitment, we generated a series of C-terminal truncations, subsequently removing ~50-aa segments (excluding the DBD; Figure 6A). Promoter selection was gradually changed; as the Yap1 C terminus became shorter, the association of Yap1 with its target promoters decreased, while binding to a new subset of promoters, preferred by its DBD-only mutant, increased (Figures 6A, 6C–6E, and S7A). Nested removal of 200 aa also had a weak effect on the Yap1 binding profile, with the exception of the 200-aa segment proximal to its DBD (Figure 6A). Therefore, similarly to Msn2, the *in vivo* specificity of Yap1 depends on a multitude of weak sequence determinants distributed throughout its entire IDR.

We noted that the binding profiles of the Yap1 truncation mutants became increasingly similar to that of Yap2, its close paralog (Figure 6D). The Yap2 DBD is highly similar to that of Yap1 (Rodrigues-Pousada et al., 2010), but its non-DBD is shorter (~300 aa) and most of it is predicted to be stably folded (Figure S7B). Truncation of the Yap2 non-DBD revealed a moderate contribution of this domain to the Yap2 specificity, as compared to Yap1 (Figures S7B–S7D). Therefore, while Yap1 is recruited to its target promoters by its non-DBD, Yap2 binding preferences depend mainly on its DBD, with a small additional contribution of sequences outside this domain.

To understand the evolutionary dynamics leading to this divergence in Yap1-Yap2 promoter targeting, we swapped the Yap1 and Yap2 non-DBDs with the corresponding regions of their different orthologs, most of which show little sequence similarity (Figures S7E and S7F). As in the case of Msn2, Yap1 specificity determinants were conserved between distant orthologs, including orthologs from species that have diverged prior to

Please cite this article in press as: Brodsky et al., Intrinsically Disordered Regions Direct Transcription Factor *In Vivo* Binding Specificity, Molecular Cell (2020), https://doi.org/10.1016/j.molcel.2020.05.032

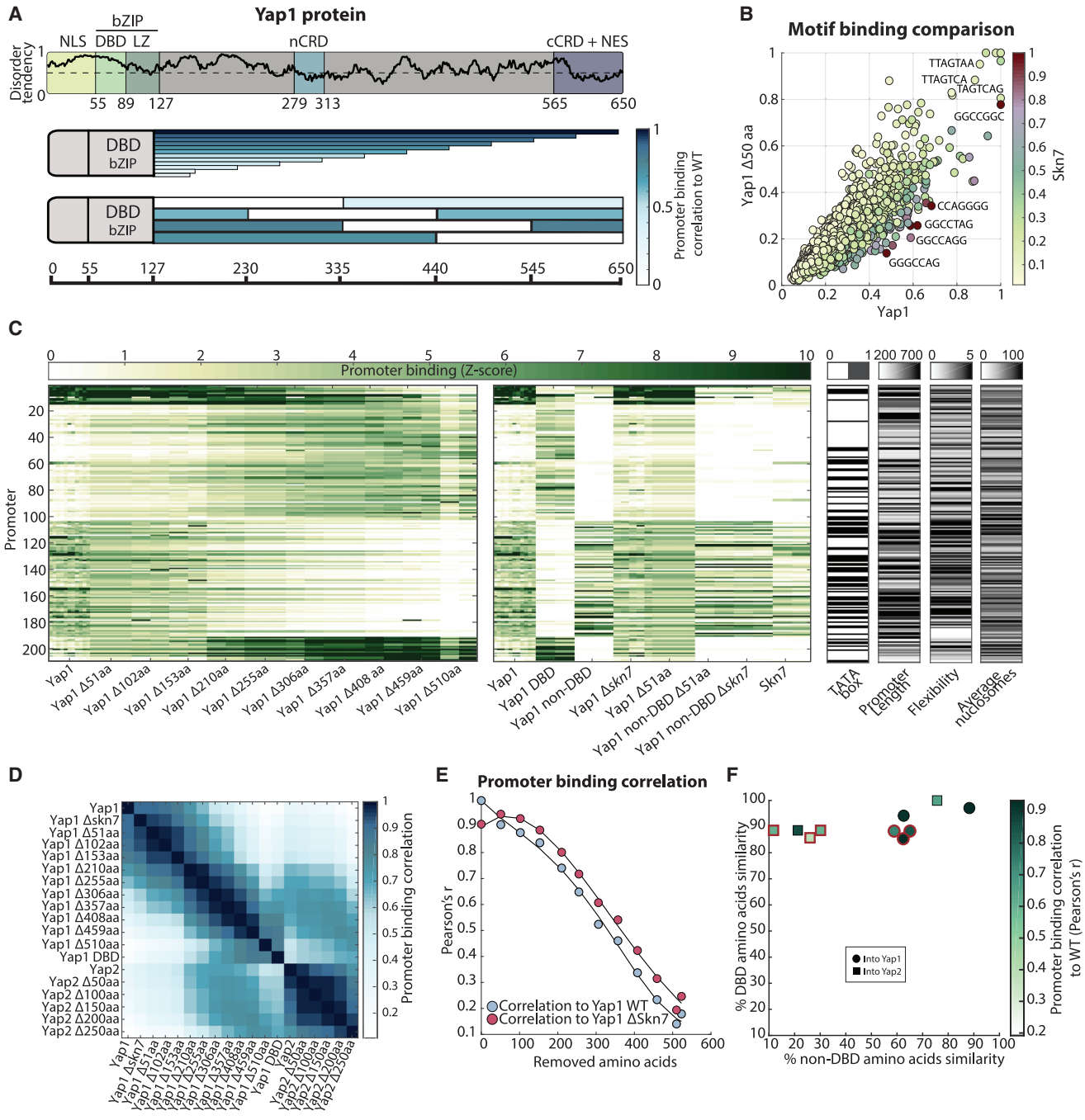


Figure 6. Yap1 Is Directed to Its Target Promoters by Multiple Specificity Determinants Distributed throughout Its IDR

(A) The Yap1 protein sequence is of low complexity: the black line indicates the predicted disorder tendency along the Yap1 protein, as calculated by IUPred (Dosztányi et al., 2005a, 2005b). Indicated are the positions of the NLS, DBD, leucine-zipper, N-terminal cysteine rich domain (n-CRD), C-terminal cysteine rich domain (c-CRD), and NES (Wood et al., 2004). The different truncation mutants used in our analysis are shown, color coded by the correlation between their promoter binding profile and that of the wild-type Yap1 (presentation as in Figure 4A).

(B) Skn7 influences Yap1 motif selection: the tendency to bind at positions containing each 7-mer is shown for Yap1 and its mutant lacking the Skn7 interaction domain (Mulford and Fassler, 2011; STAR Methods). Each motif is color coded based on the extent to which it is bound by Skn7. Note that Yap1 tends to localize at Skn7-preferrred motifs, but this preference is lost when deleting 50 aa required for the Yap1-Skn7 interaction (c-CRD domain; see A). By contrast, Yap1 promoter selection is hardly affected by this deletion (see C–E).

(C–E) C-terminal truncation gradually shifts Yap1 promoter preference: the promoter binding pattern of the Yap1 tested mutants is displayed for selected promoters (C, green panel; presentation as in Figure 4B). Comparison with the binding pattern of other mutants is also shown. Presence of TATA box, length, expression flexibility, and average nucleosome occupancy are shown for each promoter (right, gray panel). The correlations of promoter selection between all

(legend continued on next page)

Please cite this article in press as: Brodsky et al., Intrinsically Disordered Regions Direct Transcription Factor *In Vivo* Binding Specificity, Molecular Cell (2020), https://doi.org/10.1016/j.molcel.2020.05.032

the Yap1/2 duplication event (Figures 6F, S7E, and S7F). Therefore, Yap1 preserved the ancestral promoter recruitment pattern, while Yap2 acquired a new binding profile by losing the IDR-directed specificity. We conclude that, similarly to Msn2, the ability of Yap1 to direct promoter binding through a large number of weak specificity determinants is conserved across long evolutionary distances and between orthologs that show little sequence similarity.

DISCUSSION

TFs often contain low-complexity sequences that code for IDRs. In this study, we examined the role of long (>500 aa) IDRs in directing the *in vivo* binding specificity of the TFs Msn2 and Yap1. Using the ChEC-seq approach, proposed by Henikoff and colleagues (Zentner et al., 2015), we were able to define the genomic binding positions of the intact TFs and of their truncated mutants, including mutants that lack either the non-DBD or the DBD. Of note, other methods, including the commonly used chromatin immunoprecipitation sequencing (ChIP-seq) and the newly described cleavage under targets and tagmentation (CUT&Tag; Kaya-Okur et al., 2019), which we used to verify the ChEC-seq binding profiles of the intact TFs, were not sensitive enough to capture the binding of the TF mutants. The increased sensitivity of ChEC-seq was evident when examining the capacity to detect binding of transcriptionally regulated and motif-containing promoters, not only by the intact TFs but also by their mutants (Data S1). This probably arises from the fact that the ChEC-seq method avoids the use of antibodies and that TF-proximal DNA is cleaved almost immediately upon cell harvesting.

Our study revealed that both Msn2 and Yap1 maintained the ability to recognize their target promoters upon removal of their DBD. Further, we showed that the binding of these TFs to their preferred subset of promoters depends on a large number of specificity determinants distributed throughout their entire IDRs (>500 aa). Each determinant is weak and partially redundant. Still, through their additive contribution, the IDRs direct TF binding to a subset of promoters.

TF IDRs could affect DNA-binding specificity by forming intramolecular interactions with the structured DBD, as was shown for p53 (Krois et al., 2018). In our case, this possibility is refuted, since both Msn2 and Yap1 recognize the majority of their target promoters independently of their DBD. Another possibility is that the IDRs interact with specific DNA-bound proteins that are localized to the same promoters. We do not favor this possibility for two main reasons. First, the *in vivo* specificity was hardly affected by abolishing the interactions with the most promising recruiting candidates—DNA-binding factors whose respective interactions with Msn2 and Yap1 were physically and functionally characterized. While we cannot rule out the existence of alternative recruiting co-factors, our survey of the extensive available literature did not suggest such candidates. Second, the ability of IDRs to direct promoter binding remained conserved across long evolutionary distances and despite high sequence divergence. Compensating mutations are therefore restricted to the IDRs themselves, as expected for domains involved in multi-partner interactions. Indeed, if the IDRs were to interact with a single partner, compensating mutations would likely to have resulted in co-evolution and the loss of cross-species interaction.

We therefore favor the possibility that the IDRs are directly attracted to specific promoters. This attraction could be mediated by electrostatic interactions between the IDR and the DNA sequence, its fold, or the chromatin environment, consistent with the weak and additive nature of the specificity determinants distributed across the IDR. These interactions could be further promoted by geometrical compatibilities between the length or flexibility of the respective nucleotides or amino acids. IDRs can indeed interact with DNA, and while this interaction is often considered nonspecific, high-affinity binding was demonstrated *in vitro,* for example in the case of the Ubx homeodomain protein in *Drosophila* (Liu et al., 2008; Williams et al., 2015). Further studies are required to define the specificity code carried by the IDR sequences and the mechanism through which this interaction is mediated.

Regardless of the mechanism explaining the attraction of TF IDRs to their target promoters, we propose that this attraction accelerates the search of TFs for their binding sites along the DNA. As is well appreciated, the time required for a TF to find its binding sites depends on the size of these sites (Mirny et al., 2009). It had long been realized that a TF that relies solely on 3D diffusion cannot detect its short DNA-binding motifs rapidly enough, so that facilitating mechanisms such as 1D diffusion along the DNA are required (Berg et al., 1981). In *E. coli*, for example, a single lac repressor requires ~6 min to reach its operator, and this is made possible only through a combination of 3D diffusion and 1D DNA sliding (Elf et al., 2007; Li and Xie, 2011). However, even in the simple bacterial genome, diffusion along the DNA is slow, raising the question of whether 1D diffusion can indeed provide the needed acceleration in the longer, more complex, and chromatin-packed eukaryotic genomes (Mirny et al., 2009).

Direct attraction of extended IDRs, composed of hundreds of amino acids, to the DNA or its chromatin environment may provide the needed acceleration. Rather than searching for a binding motif of 6–12 bp, the IDR may be attracted through multivalent weak interactions to specific yet broad DNA regions. The search time could therefore decrease in proportion to the increase of the size of the detected DNA region. Note that in contrast to previous proposals, in which the detected region was effectively increased by invoking periods of 1D diffusion interspacing the 3D-diffusion "jumps", here, the final search for binding sites is restricted to regions where binding is indeed

truncation mutants of Yap1 and Yap2 are shown in (D). The overall correlations in promoter selection between the truncation mutants and Yap1 in wild-type or Skn7-deleted cells are shown in (E). Note that the gradual change in promoter binding along truncations is independent of the Yap1-Skn7 interaction.
(F) Distant Yap1 orthologs maintain conserved promoter preferences: the non-DBDs of Yap1 and Yap2 were replaced with the corresponding regions taken from their orthologs (Figures S7E and S7F). Color code indicates the correlation between the promoter selection of these swapped mutants with that of their respective *S. cerevisiae* protein. Red outline indicates non-DBDs taken from orthologs of species that have diverged prior to the Yap1/2 duplication event.

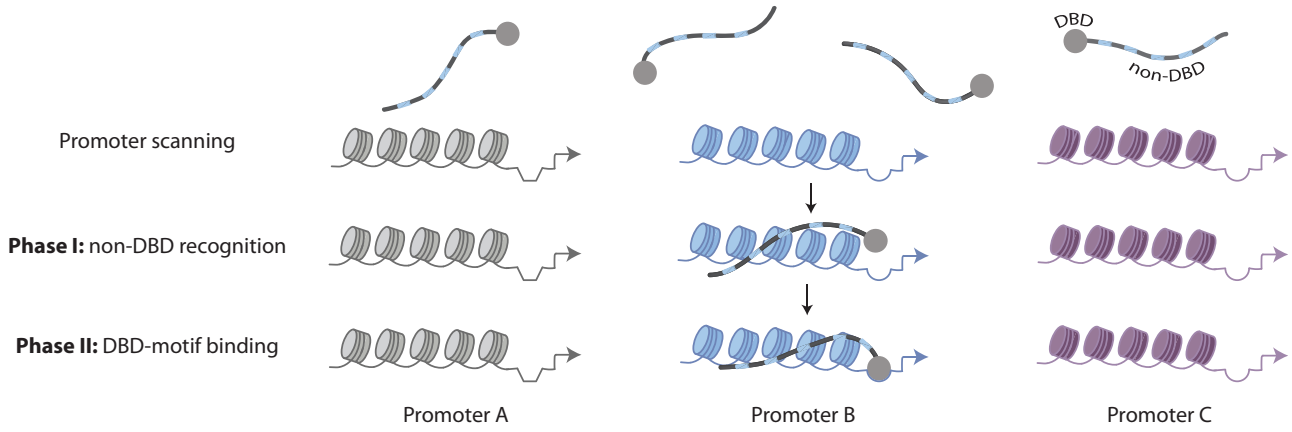


Figure 7. Suggested Model for TF-Promoter Recognition

Two-step search model: we propose that IDR-containing TFs search for their binding sites along the genome through a two-step process. First, the TF localizes to broad DNA regions recognized by its extended IDR. Next, the DBD binds its high-affinity motif present in this region, thereby stabilizing the binding.

required and thereby does not slow down the 3D search (Mirny et al., 2009).

Taken together, we propose that IDRs direct TF binding through a two-step process. First, the IDR localizes the TF to a broad DNA region surrounding the precise binding site. The DBD subsequently recognizes its high-affinity motif present in this region, stabilizing the binding (Figure 7). Considering the potential of this two-step dynamics for accelerating the TF search process, we hypothesize that this strategy of using IDR-guided distributed sensing presents a general way by which *in vivo* specificity is encoded, relevant in particular for TFs that, like Msn2 and Yap1, are required for generating rapid responses to activation cues.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead Contact
  - Materials Availability
  - Data and Code Availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Budding yeast growth, maintenance, and genetic manipulation
- METHOD DETAILS
  - ChEC-Seq experiments
  - MNase-Seq for nucleosome occupancy
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Bioinformatics analysis software
  - ChEC-Seq processing and analysis
  - Target promoter definition
  - Promoter binding probability of swapped strains
  - Motif analysis
  - Probability weight matrices (PWM)
  - Motif sequence clustering
  - Z-scores and binding strength ordering
  - Predicted binding
  - MNase-seq processing and analysis
  - Gene expression flexibility
  - TATA-box
  - Protein alignment

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.molcel.2020.05.032.

ACKNOWLEDGMENTS

We thank Dana Bar-Zvi, Raz Bar-Ziv, Gilad Yaakov, Offir Lupo, Felix Jonas, and Tamar Gera for technical assistance and fruitful discussions. We also thank Yaakov Levy and Sarel Fleishman from the Weizmann institute for scientific consultation. We would like to acknowledge Benny Shilo from the Weizmann Institute and Moshe Yaniv from the Pasteur Institute for commenting on the manuscript. This project was supported by the ISF, BSF, NSF, and the Minerva Center.

AUTHOR CONTRIBUTIONS

S.B., T.J., and N.B. conceived the study and designed experiments; S.B., T.J., K.M., M. Carmi, D.K.K., and M. Chapal performed experiments; S.B. and T.J. analyzed the data; S.B., T.J., and N.B. wrote the manuscript; and N.B. supervised the research.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 29, 2019
Revised: March 10, 2020
Accepted: May 21, 2020
Published: June 16, 2020

REFERENCES

Anand, R., Memisoglu, G., and Haber, J. (2017). Cas9-mediated gene editing in Saccharomyces cerevisiae. Protoc. Exch. Published online April 13, 2017. https://doi.org/10.1038/protex.2017.021a.

Arbesú, M., Iruela, G., Fuentes, H., Teixeira, J.M.C., and Pons, M. (2018). Intramolecular fuzzy interactions involving intrinsically disordered domains. Front. Mol. Biosci. *5*, 39.

Basehoar, A.D., Zanton, S.J., and Pugh, B.F. (2004). Identification and distinct regulation of yeast TATA box-containing genes. Cell *116*, 699–709.

Berg, O.G., Winter, R.B., and von Hippel, P.H. (1981). Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory. Biochemistry *20*, 6929–6948.

Blecher-Gonen, R., Barnett-Itzhaki, Z., Jaitin, D., Amann-Zalcenstein, D., Lara-Astiaso, D., and Amit, I. (2013). High-throughput chromatin immunoprecipitation for genome-wide mapping of in vivo protein-DNA interactions and epigenomic states. Nat. Protoc. *8*, 539–554.

Boija, A., Klein, I.A., Sabari, B.R., Dall'Agnese, A., Coffey, E.L., Zamudio, A.V., Li, C.H., Shrinivas, K., Manteiga, J.C., Hannett, N.M., et al. (2018). Transcription factors activate genes through the phase-separation capacity of their activation domains. Cell *175*, 1842–1855.e16.

Borgia, A., Borgia, M.B., Bugge, K., Kissling, V.M., Heidarsson, P.O., Fernandes, C.B., Sottini, A., Soranno, A., Buholzer, K.J., Nettels, D., et al. (2018). Extreme disorder in an ultrahigh-affinity protein complex. Nature *555*, 61–66.

Byrne, K.P., and Wolfe, K.H. (2005). The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. Genome Res. *15*, 1456–1461.

Chowdhary, S., Kainth, A.S., Pincus, D., and Gross, D.S. (2019). Heat shock factor 1 drives intergenic association of its target gene loci upon heat shock. Cell Rep. *26*, 18–28.e5.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., et al. (2013). Multiplex genome engineering using CRISPR/Cas systems. Science *339*, 819–823.

David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C.J., Bofkin, L., Jones, T., Davis, R.W., and Steinmetz, L.M. (2006). A high-resolution map of transcription in the yeast genome. Proc. Natl. Acad. Sci. USA *103*, 5320–5325.

Dosztányi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005a). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics *21*, 3433–3434.

Dosztányi, Z., Csizmók, V., Tompa, P., and Simon, I. (2005b). The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J. Mol. Biol. *347*, 827–839.

Elbaum-Garfinkle, S., Kim, Y., Szczepaniak, K., Chen, C.C.-H., Eckmann, C.R., Myong, S., and Brangwynne, C.P. (2015). The disordered P granule protein LAF-1 drives phase separation into droplets with tunable viscosity and dynamics. Proc. Natl. Acad. Sci. USA *112*, 7189–7194.

Elf, J., Li, G.-W., and Xie, X.S. (2007). Probing transcription factor dynamics at the single-molecule level in a living cell. Science *316*, 1191–1194.

Fordyce, P.M., Gerber, D., Tran, D., Zheng, J., Li, H., DeRisi, J.L., and Quake, S.R. (2010). De novo identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis. Nat. Biotechnol. *28*, 970–975.

Fuxreiter, M., Simon, I., and Bondos, S. (2011). Dynamic protein-DNA recognition: beyond what can be seen. Trends Biochem. Sci. *36*, 415–423.

Gao, A., Shrinivas, K., Lepeudry, P., Suzuki, H.I., Sharp, P.A., and Chakraborty, A.K. (2018). Evolution of weak cooperative interactions for biological specificity. Proc. Natl. Acad. Sci. USA *115*, E11053–E11060.

Garber, M., Yosef, N., Goren, A., Raychowdhury, R., Thielke, A., Guttman, M., Robinson, J., Minie, B., Chevrier, N., Itzhaki, Z., et al. (2012). A high-throughput chromatin immunoprecipitation approach reveals principles of dynamic gene regulation in mammals. Mol. Cell *47*, 810–822.

Gasch, A.P., Spellman, P.T., Kao, C.M., Carmel-Harel, O., Eisen, M.B., Storz, G., Botstein, D., and Brown, P.O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. Mol. Biol. Cell *11*, 4241–4257.

Gietz, R.D., Schiestl, R.H., Willems, A.R., and Woods, R.A. (1995). Studies on the transformation of intact yeast cells by the LiAc/SS-DNA/PEG procedure. Yeast *11*, 355–360.

Gordân, R., Murphy, K.F., McCord, R.P., Zhu, C., Vedenko, A., and Bulyk, M.L. (2011). Curated collection of yeast transcription factor DNA binding specificity data reveals novel structural and gene regulatory insights. Genome Biol. *12*, R125.

Guo, X., Bulyk, M.L., and Hartemink, A.J. (2012). Intrinsic disorder within and flanking the DNA-binding domains of human transcription factors. Pac. Symp. Biocomput. *2012*, 104–115.

Habchi, J., Tompa, P., Longhi, S., and Uversky, V.N. (2014). Introducing protein intrinsic disorder. Chem. Rev. *114*, 6561–6588.

Hackett, S.R., Baltz, E.A., Coram, M., Wranik, B.J., Kim, G., Baker, A., Fan, M., Hendrickson, D.G., Berndl, M., and McIsaac, R.S. (2020). Learning causal networks using inducible transcription factors and transcriptome-wide time series. Mol. Syst. Biol. *16*, e9174.

Henikoff, J.G., Belsky, J.A., Krassovsky, K., MacAlpine, D.M., and Henikoff, S. (2011). Epigenome characterization at single base-pair resolution. Proc. Natl. Acad. Sci. USA *108*, 18318–18323.

Inukai, S., Kock, K.H., and Bulyk, M.L. (2017). Transcription factor-DNA binding: beyond binding site motifs. Curr. Opin. Genet. Dev. *43*, 110–119.

Kaya-Okur, H.S., Wu, S.J., Codomo, C.A., Pledger, E.S., Bryson, T.D., Henikoff, J.G., Ahmad, K., and Henikoff, S. (2019). CUT&Tag for efficient epigenomic profiling of small samples and single cells. Nat. Commun. *10*, 1930.

Krois, A.S., Dyson, H.J., and Wright, P.E. (2018). Long-range regulation of p53 DNA binding by its intrinsically disordered N-terminal transactivation domain. Proc. Natl. Acad. Sci. USA *115*, E11302–E11310.

Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R., and Weirauch, M.T. (2018). The human transcription factors. Cell *172*, 650–665.

Lee, J., Godon, C., Lagniel, G., Spector, D., Garin, J., Labarre, J., and Toledano, M.B. (1999). Yap1 and Skn7 control two specialized oxidative stress response regulons in yeast. J. Biol. Chem. *274*, 16040–16046.

Levo, M., and Segal, E. (2014). In pursuit of design principles of regulatory sequences. Nat. Rev. Genet. *15*, 453–468.

Li, G.-W., and Xie, X.S. (2011). Central dogma at the single-molecule level in living cells. Nature *475*, 308–315.

Li, B., Carey, M., and Workman, J.L. (2007). The role of chromatin during transcription. Cell *128*, 707–719.

Liu, C.L., Kaplan, T., Kim, M., Buratowski, S., Schreiber, S.L., Friedman, N., and Rando, O.J. (2005). Single-nucleosome mapping of histone modifications in S. cerevisiae. PLoS Biol. *3*, e328.

Liu, J., Perumal, N.B., Oldfield, C.J., Su, E.W., Uversky, V.N., and Dunker, A.K. (2006). Intrinsic disorder in transcription factors. Biochemistry *45*, 6873–6888.

Liu, Y., Matthews, K.S., and Bondos, S.E. (2008). Multiple intrinsically disordered sequences alter DNA binding by the homeodomain of the Drosophila hox protein ultrabithorax. J. Biol. Chem. *283*, 20874–20887.

MacIsaac, K.D., Wang, T., Gordon, D.B., Gifford, D.K., Stormo, G.D., and Fraenkel, E. (2006). An improved map of conserved regulatory sites for Saccharomyces cerevisiae. BMC Bioinformatics *7*, 113.

Martínez-Pastor, M.T., Marchler, G., Schüller, C., Marchler-Bauer, A., Ruis, H., and Estruch, F. (1996). The Saccharomyces cerevisiae zinc finger proteins Msn2p and Msn4p are required for transcriptional induction through the stress response element (STRE). EMBO J. *15*, 2227–2235.

Minezaki, Y., Homma, K., Kinjo, A.R., and Nishikawa, K. (2006). Human transcription factors contain a high fraction of intrinsically disordered regions essential for transcriptional regulation. J. Mol. Biol. *359*, 1137–1149.

Mirny, L., Slutsky, M., Wunderlich, Z., Tafvizi, A., Leith, J., and Kosmrlj, A. (2009). How a protein searches for its site on DNA: the mechanism of facilitated diffusion. J. Phys. A Math. Theor. *42*, 434013.

Morgunova, E., and Taipale, J. (2017). Structural perspective of cooperative transcription factor binding. Curr. Opin. Struct. Biol. *47*, 1–8.

Mulford, K.E., and Fassler, J.S. (2011). Association of the Skn7 and Yap1 transcription factors in the Saccharomyces cerevisiae oxidative stress response. Eukaryot. Cell *10*, 761–769.

Persikov, A.V., and Singh, M. (2014). De novo prediction of DNA-binding specificities for Cys2His2 zinc finger proteins. Nucleic Acids Res. *42*, 97–108.

Protter, D.S.W., Rao, B.S., Van Treeck, B., Lin, Y., Mizoue, L., Rosen, M.K., and Parker, R. (2018). Intrinsically disordered regions can contribute promiscuous interactions to RNP granule assembly. Cell Rep. *22*, 1401–1412.

Rodrigues-Pousada, C., Menezes, R.A., and Pimentel, C. (2010). The Yap family and its role in stress response. Yeast *27*, 245–258.

Sabari, B.R., Dall'Agnese, A., Boija, A., Klein, I.A., Coffey, E.L., Shrinivas, K., Abraham, B.J., Hannett, N.M., Zamudio, A.V., Manteiga, J.C., et al. (2018). Coactivator condensation at super-enhancers links phase separation and gene control. Science *361*, 361.

Sadeh, A., Baran, D., Volokh, M., and Aharoni, A. (2012). Conserved motifs in the Msn2-activating domain are important for Msn2-mediated yeast stress response. J. Cell Sci. *125*, 3333–3342.

Shen, X.X., Zhou, X., Kominek, J., Kurtzman, C.P., Hittinger, C.T., and Rokas, A. (2016). Reconstructing the backbone of the Saccharomycotina yeast phylogeny using genome-scale data. G3 (Bethesda) *6*, 3927–3939.

Shen, N., Zhao, J., Schipper, J.L., Zhang, Y., Bepler, T., Leehr, D., Bradley, J., Horton, J., Lapp, H., and Gordan, R. (2018). Divergence in DNA specificity among paralogous transcription factors contributes to their differential in vivo binding. Cell Syst. *6*, 470–483.e8.

Shively, C.A., Liu, J., Chen, X., Loell, K., and Mitra, R.D. (2019). Homotypic cooperativity and collective binding are determinants of bHLH specificity and function. Proc. Natl. Acad. Sci. USA *116*, 16143–16152.

Shoemaker, B.A., Portman, J.J., and Wolynes, P.G. (2000). Speeding molecular recognition by using the folding funnel: the fly-casting mechanism. Proc. Natl. Acad. Sci. USA *97*, 8868–8873.

Siggers, T., Reddy, J., Barron, B., and Bulyk, M.L. (2014). Diversification of transcription factor paralogs via noncanonical modularity in C2H2 zinc finger DNA binding. Mol. Cell *55*, 640–648.

Tompa, P. (2003). Intrinsically unstructured proteins evolve by repeat expansion. BioEssays *25*, 847–855.

Vuzman, D., Azia, A., and Levy, Y. (2010). Searching DNA via a "Monkey Bar" mechanism: the significance of disordered tails. J. Mol. Biol. *396*, 674–684.

Wang, J., Choi, J.-M., Holehouse, A.S., Lee, H.O., Zhang, X., Jahnel, M., Maharana, S., Lemaitre, R., Pozniakovsky, A., Drechsel, D., et al. (2018). A molecular grammar governing the driving forces for phase separation of prion-like RNA binding proteins. Cell *174*, 688–699.e16.

Williams, D.C., Desai, M., and Ginder, G.D. (2015). A fuzzy DNA binding region in MBD2 recruits the histone deacetylase core complex of NuRD and modifies kinetics of DNA binding. Biophys. J. *108*, 388a.

Wolfe, K.H., and Shields, D.C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. Nature *387*, 708–713.

Wood, M.J., Storz, G., and Tjandra, N. (2004). Structural basis for redox regulation of Yap1 transcription factor localization. Nature *430*, 917–921.

Wright, P.E., and Dyson, H.J. (2009). Linking folding and binding. Curr. Opin. Struct. Biol. *19*, 31–38.

Zentner, G.E., Kasinathan, S., Xin, B., Rohs, R., and Henikoff, S. (2015). ChEC-seq kinetics discriminates transcription factor binding sites by DNA sequence and shape in vivo. Nat. Commun. *6*, 8733.

Zhao, Y., Granas, D., and Stormo, G.D. (2009). Inferring binding energies from selected binding sites. PLoS Comput. Biol. *5*, e1000590.

Please cite this article in press as: Brodsky et al., Intrinsically Disordered Regions Direct Transcription Factor *In Vivo* Binding Specificity, Molecular Cell (2020), https://doi.org/10.1016/j.molcel.2020.05.032

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, Peptides, and Recombinant Proteins		
cOmplete EDTA-free Protease Inhibitor Cocktail	Sigma Aldrich	Cat#11873580001
Proteinase K	Sigma Aldrich	Cat#P2308
RNase A	Sigma Aldrich	Cat#R4875
AMPure XP	Beckman Coulter	Cat#A63881
Glycoblue	Thermo Fisher	Cat# AM9515
Zymolase 100T	Amsbio	Cat#120493-1
IGEPAL CA-630	Sigma Aldrich	Cat#I3021
Micrococcal nuclease (MNase)	Worthington	Cat#LS004797
Digitonin	Sigma Aldrich	Cat#300410
Spermine	Sigma Aldrich	Cat# S3256-5G
Spermidine	Sigma Aldrich	Cat# S0266
Critical Commercial Assays		
HiYield Plasmid Mini Kit	RBC Bioscience	Cat#YPD100
HiYield GEL/PCR DNA Fragments Extraction Kit	RBC Bioscience	Cat#YDF100
Deposited Data		
ChEC-Seq, Mnase-Seq data	This study	BioProject: PRJNA573518 https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA573518
Experimental Models: Organisms/Strains		
Yeast strain information	This study	Table S2
Oligonucleotides		
Primers used for strain creation	This study	Table S3
Recombinant DNA		
pGZ108 (pFA6a-3FLAG-MNase-kanM6)	Addgene	Cat #70231
bRA89 (PGK1-Cas9- HPHMX-BpII)	Addgene	Cat #100950
pAG25 (pFA6-natM6)	Addgene	Cat #35121
Software and Algorithms		
Bowtie2	Johns Hopkins University	http://bowtie-bio.sourceforge.net/index.shtml
MATLAB	MathWorks	https://www.mathworks.com/products/matlab.html
IUPred	Dosztányi et al., 2005a, 2005b	https://iupred.elte.hu/

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Prof. Naama Barkai (naama.barkai@weizmann.ac.il).

Materials Availability

All strains used in this study are available by direct request to the lead contact without any further restrictions.

Data and Code Availability

The ChEC-seq and MNase-Seq datasets reported in this paper are available at the NCBI BioProject database, accession number BioProject: PRJNA573518, https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA573518.

Please cite this article in press as: Brodsky et al., Intrinsically Disordered Regions Direct Transcription Factor *In Vivo* Binding Specificity, Molecular Cell (2020), https://doi.org/10.1016/j.molcel.2020.05.032

CellPress

EXPERIMENTAL MODEL AND SUBJECT DETAILS

All strains used in this study are derived from the wild-type *Saccharomyces cerevisiae* strain, BY4741. Specific genotypes of all strains used in this study are available in Table S2. Growth conditions are specified under each experimental method detailed below.

Budding yeast growth, maintenance, and genetic manipulation

Yeast strains were freshly thawed before experiments from a frozen stock, plated on YPD plates, and grown. Single colonies were picked and grown at 30°C in liquid SD medium. Optical density (OD) measurements are specified in the Method Details section. For genetic manipulation of yeast, BY4741 strain, of genotype MATa his3-Δ1 leu2-Δ0 lys2-Δ0 met15-Δ0 ura3-Δ0, was transformed using the LiAc/SS DNA/PEG method (Gietz et al., 1995). Briefly, a single colony was inoculated in fresh liquid YPD,grown to saturation overnight, diluted into fresh 5 ml YPD and grown to OD_{600} of 0.5. The cells were then washedwith DDW and then with LiAc 100 mM, and resuspended in transformation mix (33% PEG-3350, 100 mM LiAc, single stranded salmon sperm DNA and the DNA oligos intended for transformation). The cells were incubated at 30°C for 30 minutes followed by a 30 minutes heat shock at 42°C. The cells were then plated on YPD plates and grown overnight in 30°C for recovery. In the following day, the cells were replicated to the appropriate selection plate. For ChEC-seq experiments, TFs were C-terminally tagged with an MNase. Yeast cells were transformed with the amplification product of an MNase-Kanamycin cassette from the pGz108 plasmid, a gift from Steven Henikoff, and selected on plates containing G418. Strains including domain swaps, deletions or scrambling, were generated using CRISPR (Cong et al., 2013). To this end, a genomic PCR amplification product or a synthetic oligo was co-transformed as a repair template alongside the bRA89 plasmid, a gift from James Haber, harboring Cas9 and the locus-specific 20 bp guide-RNA. Ligation of the gene-specific guide-RNA into the bRA89 plasmid was performed as previously described (Anand et al., 2017). Following validation of positive clones, the bRA89 plasmid was lost by growth in YPD, followed by selection of colonies that lost the bRA89 Hygromycin resistance. Gene deletion strains used in this study were generated using the amplification product of NatMX6 cassette, derived from the pAG25 plasmid, a gift from John McCusker. All strains generated for this study were verified using PCR and gel electrophoresis followed by DNA sequencing. TF domain annotations used for swapping and partial deletions experiments can be found in Table S1. Detailed lists of strains and oligos are found in Tables S2 and S3 respectively.

METHOD DETAILS

ChEC-Seq experiments

The experiments were performed as described previously (Zentner et al., 2015), with some modifications. Cultures were grown overnight to saturation in SD media and diluted into 5 mL of fresh SD media to reach OD_{600} of 4 the following morning after ~10 divisions. Cultures were pelleted at 1500 g and resuspended in 1 mL Buffer A (15 mM Tris pH 7.5, 80 mM KCl, 0.1 mM EGTA, 0.2 mM spermine, 0.5 mM spermidine, 1 × Roche cOmplete EDTA-free mini protease inhibitors, 1 mM PMSF), and then transferred to DNA low-bind tubes (Eppendorf 022431021). Cells were washed twice more in 500 μL Buffer A, pelleted, and resuspended in 150 μL Buffer A containing 0.1% digitonin. Then, cells were transferred to an Eppendorf 96-well plate (Eppendorf 951020401) for permeabilization (30°C for 5 min). $CaCl_2$ was added to a final concentration of 2 mM. For all experiments presented in the main text, the Mnase was activated for 30 s, except for DBD-deleted strains that were incubated for 60 s (see Figures S3A-C and S6H for a detailed analysis of different MNase activation durations). Next, 100 μL of stop buffer (400 mM NaCl, 20 mM EDTA, 4 mM EGTA and 1% SDS) were mixed with 100 μL of sample. Proteinase K was then added, and incubated at 55°C for 30 min. Nucleic acid extraction was performed as previously described (Zentner et al., 2015), with some modifications in the ethanol precipitation step; In brief, samples were precipitated (at −80°C for > 1 hour) with 2.5 volumes of cold EtOH 96%, 45 μg Glycoblue and sodium acetate to a final concentration of 20 mM. DNA was centrifuged (4°C for 10 min), washed with EtOH 70% and treated with RNase A in a final concentration of 2.5 mg/ml (37°C for 20 min), followed by another round of DNA cleanup and ethanol precipitation. In order to enrich for small DNA fragments, reverse 0.8X SPRI clean-up was carried out. Library preparation was identical to the previously published protocol (Henikoff et al., 2011; Zentner et al., 2015), except for the clean-up steps, which were performed using phenol-chloroform followed by ethanol precipitation as described above (instead of S400 columns). 1X SPRI cleanup was carried out on the indexed (Blecher-Gonen et al., 2013) ChEC amplified libraries, which were then pooled and sequenced on Illumina NextSeq500 for paired end (50 bps for read1 and 15 or 25 bps for read 2). The number of repeats for each strain is indicated in Table S2.

MNase-Seq for nucleosome occupancy

Nucleosome occupancy was measured as previously described (Liu et al., 2005). Briefly, wild-type *S. cerevisiae* BY4741 cells were grown in SD media to saturation overnight, diluted to fresh SD media and grown for 4 divisions at 30°C to the OD_{600} of 4. 10 mL of cells were fixated for 15 min in 1% formaldehyde shaking at room temperature. Glycine was added to a final concentration of 0.125 M for 5 min, shaking at room temperature. Cell pellets were washed, and spheroplasted for 25 minutes in 30°C with zymolase by adding 13 μL buffer Z (1M sorbitol, 50 mM Tris pH 7.4, 1 mM 2-mercaptoethanol and 1 unit zymolase 100T). Spheroplasts were subjected to MNase digestion in 166 μL NP mix, containing 160 μL NP buffer (1M sorbitol, 10 mM Tris pH 4, 50 mM NaCl, 5 mM MgCl2, 1 mM $CaCl_2$, 0.075% IGEPAL CA-630), supplemented with spermidine to a final concentration of 17.5 mM, 2-mercaptoethanol in $1:1.66*10^{-4}$ ratio, Roche cOmplete EDTA-free mini protease inhibitors to a final concentration of x0.12 and 1 unit MNase, for

Please cite this article in press as: Brodsky et al., Intrinsically Disordered Regions Direct Transcription Factor *In Vivo* Binding Specificity, Molecular Cell (2020), https://doi.org/10.1016/j.molcel.2020.05.032

20 minutes in 37°C. MNase treatment was stopped using an equal volume of stop buffer (220mM NaCl, 0.2% SDS, 0.2% sodium deoxycholate, 10mM EDTA, 2% Triton X-100). Cells were treated with RNase (1 μg per sample) at 37°C for 30 min and then with Proteinase K (50 μg per sample) at 37°C for 2 hours. Samples were reverse cross-linked by overnight incubation at 65°C, and DNA was purified using a 2x SPRI cleanup. DNA libraries were indexed (Garber et al., 2012), pooled and pair-end sequenced using the Illumina NextSeq500.

QUANTIFICATION AND STATISTICAL ANALYSIS

Bioinformatics analysis software

Sequencing reads were aligned using Bowtie2 with the parameters specified in each section. Bioinformatics analysis was performed using MATLAB. Disorder tendency was calculated using the IUPred tool (Dosztányi et al., 2005a, 2005b).

ChEC-Seq processing and analysis

Reads were aligned using Bowtie2 (parameters: –best –m 1) to *S. cerevisiae* (reference genome, cerR64). ChEC-Seq tracks, representing the binding of each TF, were calculated by adding +1 to each genomic location corresponding to the first nucleotide in a forward read, or the 50th position corresponding a reverse read. The signal was normalized to a total of 10 million reads, to control for sequencing depth. For promoter analysis, promoters were defined only for genes with an annotated transcript according to David et al. (2006). The length of each promoter was defined as 700 bps upstream to the transcription start site (TSS) or to the position where a promoter meets another transcript. The signal across each promoter was summed and normalized to the maximal promoter length (700 bps) to calculate overall promoter binding for each sample. Meta-gene profiles were generated by averaging the signal of 850 bps over all promoters (from −700 to 150 relative to the TSS).

Target promoter definition

To define the target promoters for each TF, a series of signal thresholds was generated, ranging from 0 to 20,000 normalized reads with intervals of 50. The number of promoters that passed each of the 400 thresholds was calculated for each TF. The threshold for defining the targets for a given TF was set to the point where the number of promoters that pass the threshold is stable (less than a 5 promoters difference) relative to the previous threshold.

Promoter binding probability of swapped strains

For each promoter, the probability to be bound by a given factor (WT, swap containing DBD and swap containing non-DBD) was calculated using the hill function:

$$P(promoter_i) = \frac{sum\ of\ signal\ on\ promoter^n}{sum\ of\ signal\ on\ promoter^n + Target\ threshold^n}$$

n was set to 5 to allow flexibility in the probability definition. Target threshold is the parameter defined in the previous section. The combined probability of each of the possible events was calculated for each promoter (being bound by; 1. WT only 2. WT and swap containing DBD only 3. WT and swap containing non-DBD only 4. All 3 factors). Finally, the averaged probability of each group was calculated as a percentage out of the WT targets.

Motif analysis

For the motif analysis, all possible x-mer sequences were given a numerical index $\frac{(4^x)}{2}$ in total; forward and reverse complement forms of each x-mer were given the same index). Each nucleotide in the yeast genome was indexed according to the x-mer that begins from it. To score each x-mer occurrence, the signal around its mid-position was averaged (20 bp window). To reduce background noise, each position with signal less than 20 normalized reads was set as zero. The averaged signal for each x-mer was then calculated across all of its occurrences in all promoters, and was assigned as its relative binding score.

Probability weight matrices (PWM)

PWMs of the different TFs were generated based on the ten most bound 7-mers of each factor. The sequences were then aligned to the top bound motif using the Needleman-Wunsch local alignment algorithm. Each motif contributed to the PWM based on its relative binding score.

Motif sequence clustering

This analysis was done using the 50 strongest bound motifs. Needleman-Wunsch local alignment algorithm was used to align each motif to all other motifs, in both its forward and reverse complement forms, keeping only the highest alignment score (i.e forward-forward or forward-reverse complement). The clustering algorithm k-means was used to cluster motifs based on alignment scores. This step was repeated with different cluster numbers (2-5) to verify the most representative cluster number for each TF. Finally, consensus PWMs were generated for each cluster separately as described above, taking all of the motifs in the cluster rather than the top 10.

Please cite this article in press as: Brodsky et al., Intrinsically Disordered Regions Direct Transcription Factor *In Vivo* Binding Specificity, Molecular Cell (2020), https://doi.org/10.1016/j.molcel.2020.05.032

Z-scores and binding strength ordering

Z-score distributions were calculated for 1. sum of signal on all promoters of a given TF, 2. signal of all nucleotides positioned within promoters. To calculate the signal around a single motif occurrence, as in Figure 2D, the average Z-score within a window of 20 bps around each motif was calculated. Promoters and motifs (Figures 1C, 2D–2F, S1B, and S6B) were ordered according to the binding strength of the first presented TF, up to the point where its signal goes below 0.5 (in units of Z-score/averaged Z-score for promoters and motifs respectively). From that point, the ordering was done according the binding strength of the next TF (until it goes below 0.5). In Figures 1C and S1B, for WGH duplicate pairs, motifs and promoters were ordered according to the median binding strength of both TFs.

Predicted binding

To predict the binding strength in which a given TF will bind a certain promoter, the 30 strongest motifs of each factor, defined in *motif analysis*, were considered. Each promoter was scored according to the number of occurrences of the top motifs within its sequence (30 points for an occurrence of the top motif and 1 point for an occurrence of the 30^{th} motif). The score was then normalized to the maximal promoter length (700 bps).

MNase-seq processing and analysis

Reads were aligned to cerR64 genome using Bowtie2 paired-end alignment. For the generation of nucleosome occupancy genomic tracks, each nucleotide present within a given read got a +1 value. Reads of each sample were normalized to 10 million. The nucleosome occupancy shown in the figures is the mean signal of two repeats. For calculating the nucleosome occupancy around a single motif occurrence, as in Figure 2D, the signal within a window of 20 bps around each motif was averaged.

Gene expression flexibility

To define the flexibility of each gene, the IDEA data was used (Hackett et al., 2020). This dataset includes the fold-change of each gene in *S. cerevisiae* across multiple data points following an induction of a specific TF (a total of 203 different TFs). For each gene, the flexibility was calculated as the difference between the fold-change of percentile 0.05% to the fold-change of percentile 99.95% over all time points of all TF time courses.

TATA-box

For defining genes with TATA box we looked for the consensus *S. cerevisiae* TATA sequence – TATA[A/T]A[A/T][A/G] (Basehoar et al., 2004) in the 200 bps upstream to the TSS. Only genes with an exact match were considered to have a TATA box.

Protein alignment

Protein sequence alignment was done using Needleman-Wunsch global alignment. The sequences of proteins from the different yeast species were downloaded from the yeast gene order browser (Byrne and Wolfe, 2005).

Molecular Cell, Volume *79*

Supplemental Information

Intrinsically Disordered Regions Direct Transcription Factor *In Vivo* Binding Specificity

Sagie Brodsky, Tamar Jana, Karin Mittelman, Michal Chapal, Divya Krishna Kumar, Miri Carmi, and Naama Barkai

Supplemental Information

Intrinsically disordered regions direct transcription factor *in-vivo* binding specificity

Sagie Brodsky, Tamar Jana, Karin Mittelman, Michal Chapal, Divya Krishna Kumar, Miri Carmi and Naama Barkai

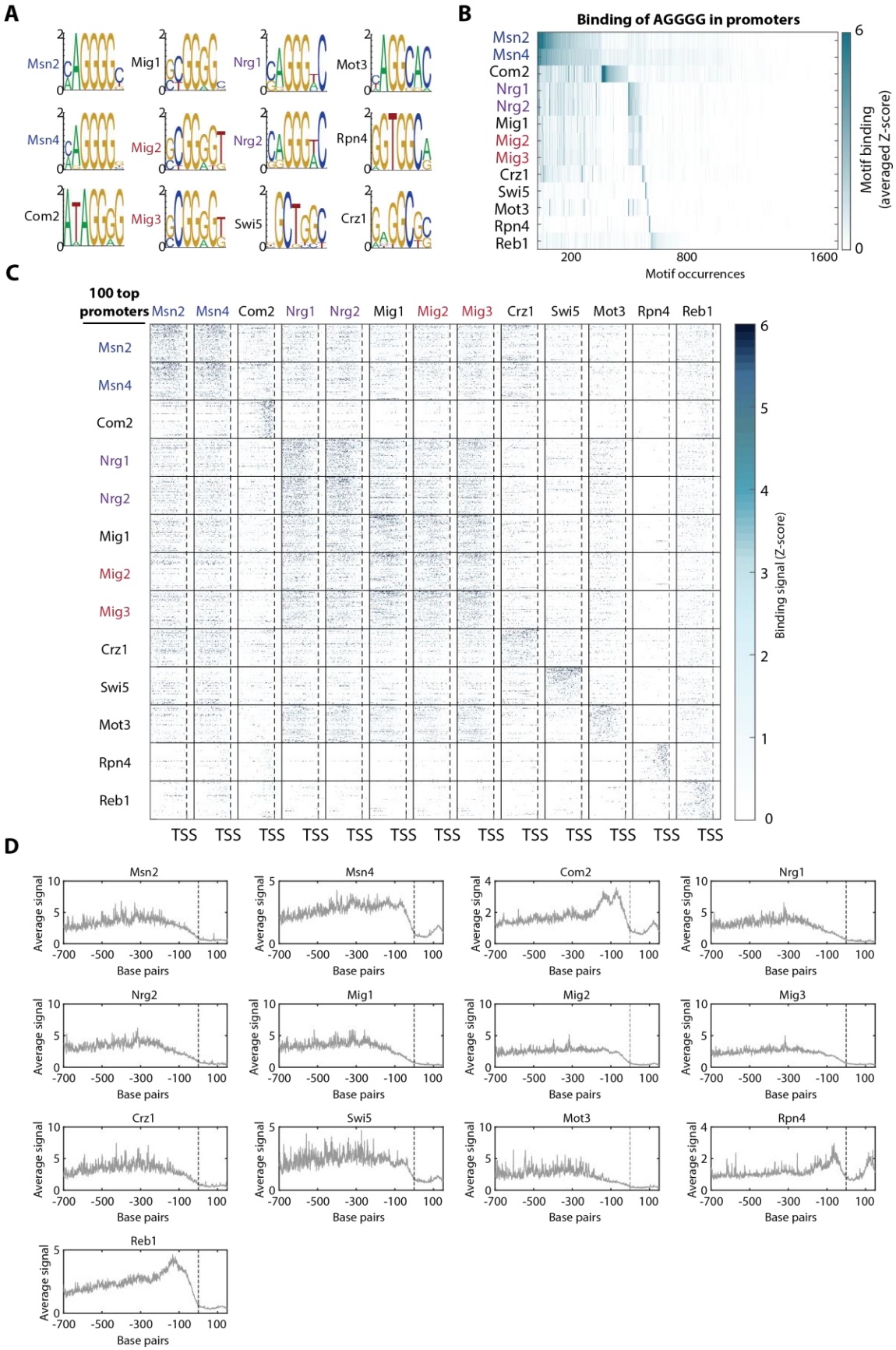

A
Msn2
Mig1
Nrg1
Mot3
Msn4
Mig2
Nrg2
Rpn4
Com2
Mig3
Swi5
Crz1
B
Binding of AGGGG in promoters
Msn2
Msn4
Com2
Nrg1
Nrg2
Mig1
Mig2
Mig3
Crz1
Swi5
Mot3
Rpn4
Reb1
200
800
1600
Motif occurrences
Motif binding (averaged Z-score)
6
0
C
100 top promoters
Msn2 Msn4 Com2 Nrg1 Nrg2 Mig1 Mig2 Mig3 Crz1 Swi5 Mot3 Rpn4 Reb1
TSS
Binding signal (Z-score)
6
5
4
3
2
1
0
D
Msn2
Msn4
Com2
Nrg1
Nrg2
Mig1
Mig2
Mig3
Crz1
Swi5
Mot3
Rpn4
Reb1
Average signal
Base pairs
-700 -500 -300 -100 100

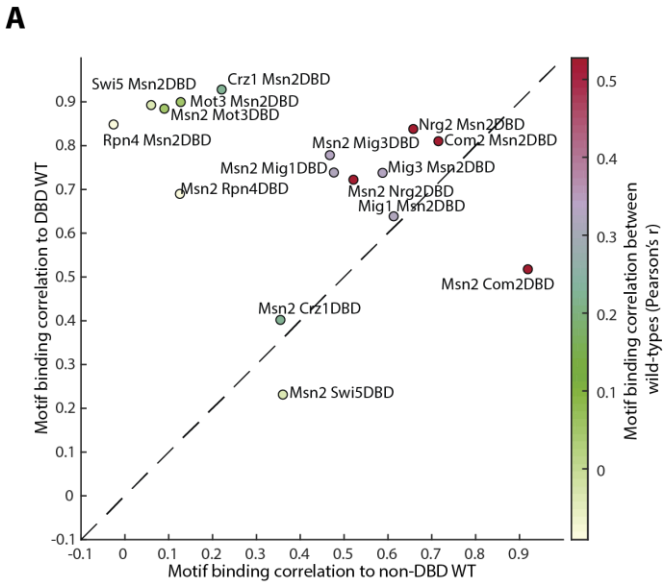
Figure S1: Paralogs of the zinc-finger family show overlapping *in-vivo* motif preferences but distinct promoter selection, Related to Figure 1.

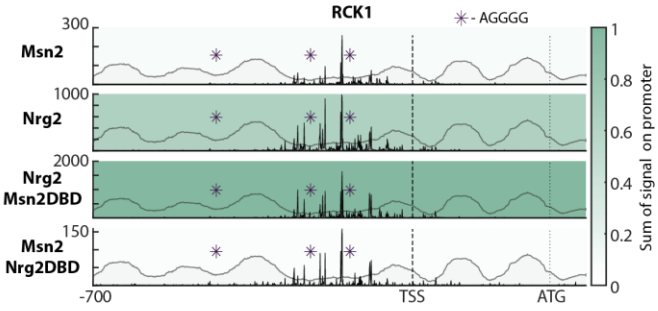
(A) *Motif preferences of the zinc-finger transcription factors (TFs):* Shown are the position weight matrices (PWMs) for the zinc-finger TFs preferred motifs, as defined by their binding profiles (see methods). Note that close paralog pairs (indicated by colors) bind identical motifs, and that most of the TFs (Msn2/4, Nrg1/2, Mig1/2/3 and Com2) bind similar motifs.

(B) *Wild-type zinc-finger TFs bind distinct subsets of AGGGG containing sites:* Shown are all AGGGG motif occurrences within promoters (~1600), ordered according to binding strength (see methods). Close paralog pairs are indicated by colors. Note that Com2, a TF with a highly similar DNA binding domain (DBD) to that of Msn2 (Siggers et al., 2014), binds a distinct set of the AGGGG motif occurrences.

(C) *Promoter binding patterns of the zinc-finger TFs to their top targets:* Each box includes the top 100 promoters bound by the TF indicated on the left. Binding patterns of all other TFs, indicated on top, to these promoters are shown. Each row within a box represents one promoter sequence (500 bp), aligned by the transcriptional start site (TSS; dashed line), with color indicating binding signal (in units of Z-score) at that given position.

(D) *Meta-gene profiles of the zinc-finger TFs:* Profiles were obtained by aligning all promoters by the TSS (dashed line) and averaging the signal (see methods).





% Refrence TF targets bound by swaps

Refrence TF	Donor TF	(i)	(ii)	(iii)	(iv)
Msn2	Com2	34.5	51.3	4.1	10.1
Msn2	Nrg2	48	20.7	19.5	11.7
Msn2	Mig1	39.9	21.4	19.5	19.2
Msn2	Mig3	40	21.5	18.3	20.3
Msn2	Swi5	35.5	23.7	21.6	19.2
Msn2	Mot3	20.8	12.1	38.5	28.6
Msn2	Crz1	39.1	13.6	30.8	16.5
Msn2	Rpn4	12.1	37.6	16.4	33.9
Com2	Msn2	33.1	28.4	9.3	29.2
Nrg2	Msn2	51.1	25.6	15	8.3
Mig1	Msn2	49.2	24.4	18.2	8.2
Mig3	Msn2	58.1	16.7	18.4	6.8
Swi5	Msn2	35.5	7.3	20.7	36.5
Mot3	Msn2	29.4	18.1	30.5	22
Crz1	Msn2	32.6	13.7	20.7	33
Rpn4	Msn2	25.5	5.4	48	21.1

Figure S2: Motif binding preference is dictated by the DBD, while promoter selection is governed by the non-DBD, Related to Figure 1.

(A) *Motif preference is dictated by the DBD*: To assess the relative contribution of the DBD and non-DBD of each TF to motif selection, the top 300 preferred motifs of all wild-type factors were chosen (a total of 1750 unique 7-mers). The motif binding scores (see methods) were calculated for each swap and were correlated to those of the wild-type containing its non-DBD (x-axis) and to the wild-type containing its DBD (y-axis). Note that the majority of the factors are above the diagonal, indicating that the DBDs are dominant in motif selection. Further, color indicates motif preference correlation between the two wild-type TFs. Note that for wild-types showing distinct motif preferences the DBD swapping had an even more pronounced effect on motif selection.

(B) *Wild-type and DBD-swapped Msn2 and Nrg2 binding to the RCK1 promoter:* Binding along the RCK1 promoter, presentation as in Figure 1F, TSS and ATG are indicated. Binding strength is indicated by background color (note also y-scale). Note that Nrg2 binds this promoter stronger than Msn2 even though only the Msn2 preferred motif is present. The swapped strain containing the Nrg2 non-DBD and the Msn2 DBD binds this promoter stronger than both wild-type TFs.

(C) *TF DBDs are not sufficient for directing in-vivo binding:* The table summarizes the consequences of all swapping experiments we performed, as follows. Consider a reference TF and the set of promoters that it binds. We examined whether these promoters remain bound when (**1**) the DBD of this reference TF was replaced with that of a donor TF, or (**2**) when the non-DBD of this reference TF was replaced with that of the same donor. Each promoter was then assigned to one of four groups: (i) remains bound in all cases (irrespectively of swapping), (ii) remains bound when the DBD is replaced but not when the non-DBD is replaced (binding depends on the non-DBD), (iii) remains bound when the non-DBD is replaced but not when DBD is replaced (binding depends on the DBD) and (iv) loses binding upon either swapping (binding depends on both the DBD and the non-DBD). For each TF pair, the fraction of promoters assigned to each of these groups is shown (See Methods section "Promoter binding probability of swapped strains" for more details). Note that the non-DBD plays a significant role in guiding promoter binding in all cases (second column). For example, replacing the Msn2 DBD with the highly similar DBD of Com2 (Siggers et al., 2014), abolished binding of only ~14% of the Msn2-bound promoters (4+10), whereas replacing the non-DBD abolished binding to ~61% of the Msn2-bound promoters (51+10). When the more distant Rpn4 was considered as the Msn2 swapping partner, DBD replacement abolished the binding of ~50% of the Msn2-bound promoters, while non-DBD replacement abolished binding of 70% of these promoters.

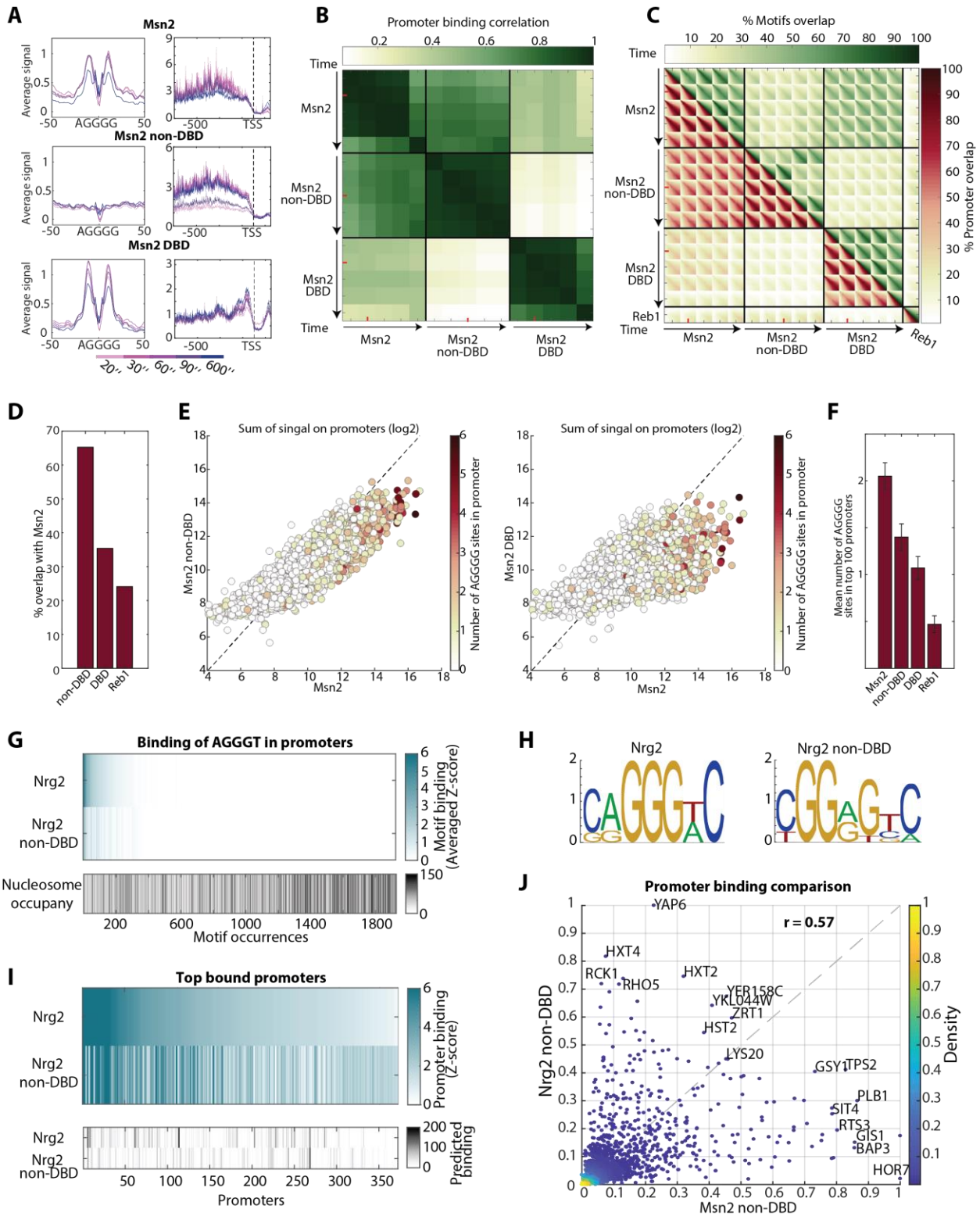

A
Msn2
Msn2 non-DBD
Msn2 DBD
Average signal
AGGGG
TSS
20', 30', 60', 90', 600'
B
Promoter binding correlation
Time
Msn2
Msn2 non-DBD
Msn2 DBD
C
% Motifs overlap
% Promoter overlap
Reb1
D
% overlap with Msn2
non-DBD
DBD
Reb1
E
Sum of singal on promoters (log2)
Msn2 non-DBD
Msn2 DBD
Msn2
Number of AGGGG sites in promoter
F
Mean number of AGGGG sites in top 100 promoters
G
Binding of AGGGT in promoters
Nrg2
Nrg2 non-DBD
Motif binding (Averaged Z-score)
Nucleosome occupany
Motif occurrences
H
Nrg2
Nrg2 non-DBD
I
Top bound promoters
Promoter binding (Z-score)
Predicted binding
Promoters
J
Promoter binding comparison
r = 0.57
YAP6
HXT4
RCK1
RHO5
HXT2
YER158C
YKL044W
ZRT1
HST2
LYS20
GSY1
TPS2
PLB1
SIT4
RTS3
GIS1
BAP3
HOR7
Nrg2 non-DBD
Msn2 non-DBD
Density

Figure S3: While not localizing to the canonical motif, the Msn2 non-DBD preferentially binds to wild-type bound AGGGG containing promoters, Related to Figure 2.

(A) *Motif binding and Metagene profiles are robust to different MNase activation times:* Since ChEC-seq TF binding profiles were shown to be affected by the MNase activation time (Zentner et al., 2015), to control for method specific biases and confirm the validity of our results, we repeated the experiment in a range of Mnase activation durations (20,30,60,90 and 600 seconds). Shown on the left is the binding signal around the Msn2 canonical motif (AGGGG), averaged over all occurrences in promoters. The different lines correspond to different MNase activation times, as indicated. Note the slight gradual decrease in signal for cleavage times exceeding 30 seconds. Metagene profiles are shown on the right (presentation same as Figure S1D).

(B) *Promoter binding pattern:* Binding profiles of the full TFs and their non-DBD and DBD mutants were measured by a sequence of ChEC-seq experiments with varying times of MNase activation (20,30,60,90 and 600 seconds). Shown are the correlations in promoter selection. Each square corresponds to one TF, as indicated, with different lines within this square corresponding to increased MNase cleavage duration. Red lines indicate the cleavage time used for results shown in the main text.

(C) *Overlap in promoter and motif preferences:* Shown are the overlaps of top-bound promoters (lower triangle, red) and preferred 7-mers (upper triangle, green). Since the overlap, defined as the number of common promoters/motifs normalized by the size of the larger group, depends on a threshold, we scanned across a range of thresholds (ranging from 5 to 500), gradually increasing the number of selected promoters/motifs from each dataset. Each square displays this range of overlap values as a function of the number of promoters selected from each of the indicated datasets. Overlap values are shown by color. The diagonal in each square corresponds to cases where equal-sized groups are compared. Red lines indicate the cleavage time used for results shown in the main text.

(D) *The non-DBD binds the majority of Msn2 bound promoters*: Shown is the percentage of overlap of defined binding targets (see methods) between Msn2 and the indicated factors. Overlap was calculated by normalizing to the smaller group.

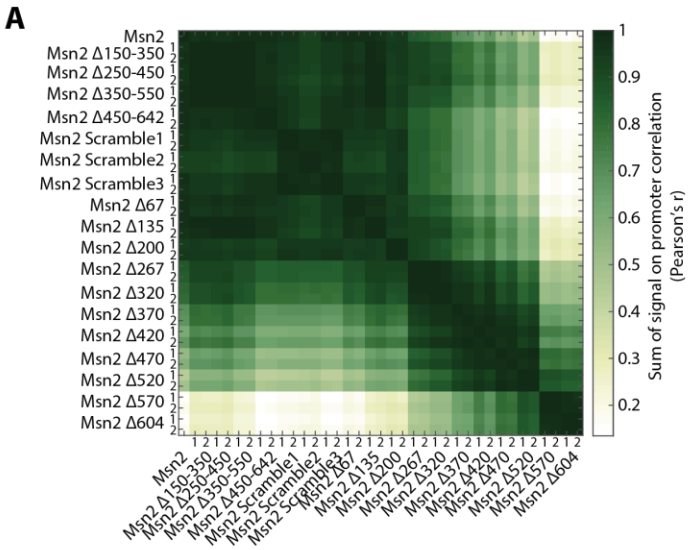
(E) *Similarity between Msn2 and its non-DBD profiles includes binding to AGGGG-containing promoters:* The figure on the left compares the promoter binding strength of Msn2 and its non-DBD. Each dot represents a promoter, the color of each dot indicates the number of AGGGG sites present in this promoter. Comparison with the binding of the DBD-only mutant is also shown (right).

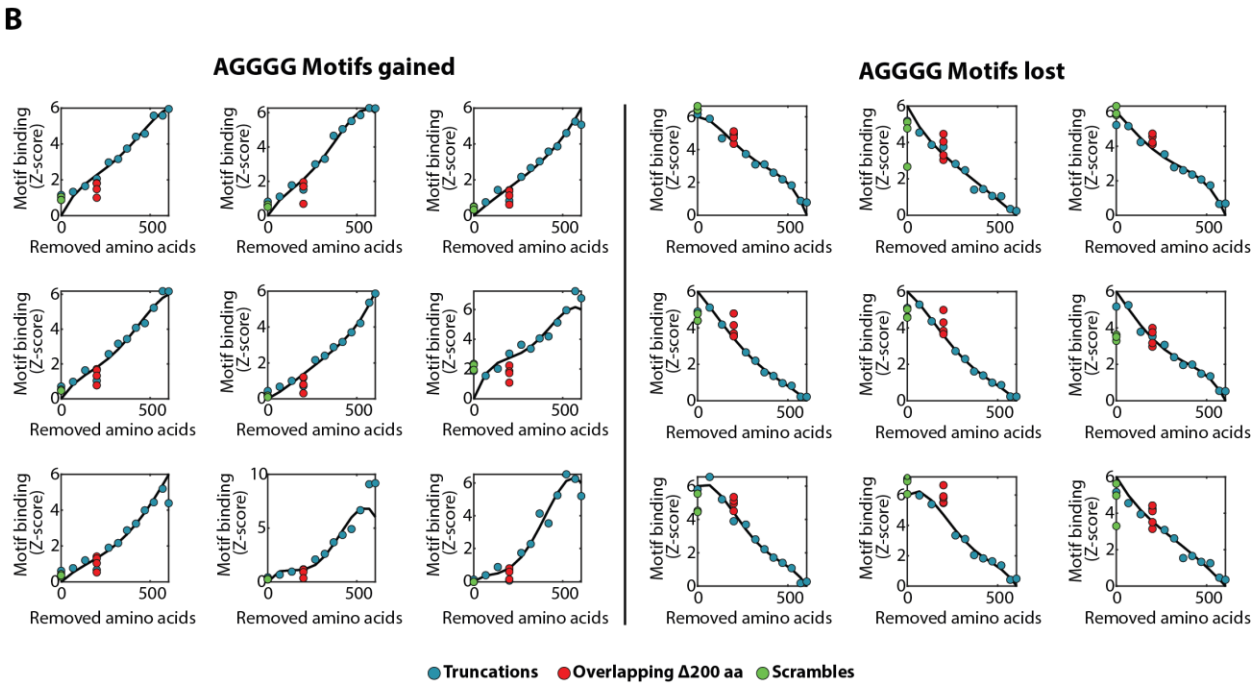
(F) *AGGGG is found within top-bound non-DBD promoters:* Shown is the average number of AGGGG occurrences within the top 100 bound-promoters for each of the indicated factors. Error bars represent the standard error of the mean (SEM).

(G-H) *Nrg2 non-DBD losses the ability to bind the Nrg2 consensus motif:* All occurrences of the preferred Nrg2 binding motif (AGGGT) in promoters are shown in (G) upper panel. Motifs are sorted according to the Nrg2 binding strength (indicated by color, measured in units of Z-score). Lower panel indicates the nucleosome occupancy around each motif occurrence (see methods). PWMs for motifs preferred by Nrg2 and the mutant containing only its non-DBD, as defined by their binding profiles, are shown in (H).

(I) *Nrg2 and its non-DBD mutant bind a highly similar promoter set:* Binding strength (in units of Z-score) is shown in the upper panel for promoters that are bound by at least one of the factors (a total of 370; see methods). Promoters were ordered according to the binding strength of Nrg2. Predicted promoter binding strength, based on motif preference and the promoter sequence, is shown on the bottom panel (see methods).

(J) *The Msn2 and Nrg2 non-DBDs bind distinct promoter sets*: Each dot represents the sum of signal on a given promoter, normalized by the maximal value of each TF. Color code indicates density.


A
Msn2
Msn2 Δ150-350
Msn2 Δ250-450
Msn2 Δ350-550
Msn2 Δ450-642
Msn2 Scramble1
Msn2 Scramble2
Msn2 Scramble3
Msn2 Δ67
Msn2 Δ135
Msn2 Δ200
Msn2 Δ267
Msn2 Δ320
Msn2 Δ370
Msn2 Δ420
Msn2 Δ470
Msn2 Δ520
Msn2 Δ570
Msn2 Δ604
Sum of signal on promoter correlation (Pearson's r)


B
AGGGG Motifs gained
AGGGG Motifs lost
Motif binding (Z-score)
Removed amino acids
Truncations
Overlapping Δ200 aa
Scrambles

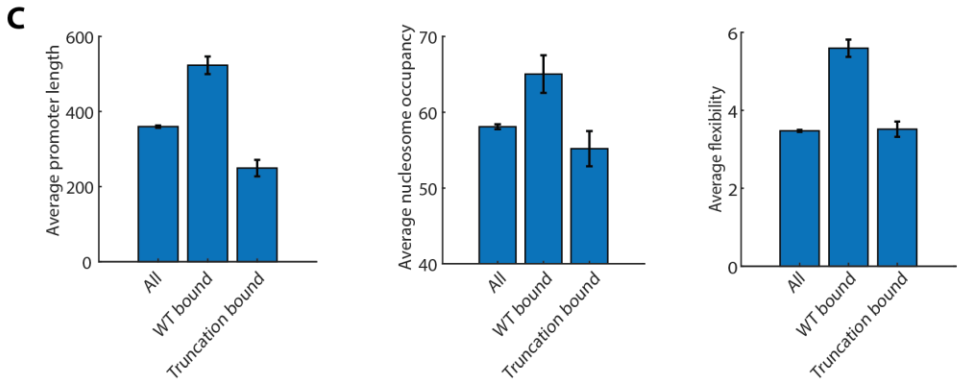

C
Average promoter length
Average nucleosome occupancy
Average flexibility
All
WT bound
Truncation bound

Figure S4: Sequential truncation of Msn2 causes a gradual change in promoter preference, while scrambling or deleting overlapping 200 aa blocks did not have an effect, Related to Figure 4.

(A) *Promoter selection of the different Msn2 mutants:* Shown are the promoter binding correlations between the indicated factors (two repeats for each construct are shown). Sequential N-terminal truncation of Msn2 resulted in a gradual change in promoter preference only when removing more than 200 aa segments, while scrambling or deletion of overlapping 200 aa segments barely had an effect.

(B) *Different AGGGG motif occurrences can be either gained or lost along truncations*: Each box shows a single representative AGGGG motif occurrence. For each motif occurrence, the binding strength (in units of Z-score) of the different mutants is plotted as a function of amino acid removal.

(C) *Wild-type preferred promoters possess unique properties:* Shown are the mean values of promoter lengths (left), average nucleosome occupancy (middle) and average gene flexibility (right; see methods) for the wild-type and truncation preferred promoter clusters shown in Figure 4B. Error bars represent the SEM. Note that truncation mutants tend to bind short promoters ($t(5649) = 4.97$, $p = 6.8\times10^{-7}$), whereas wild-type preferred promoter clusters are long ($t(5635) = 6.88$, $p = 6.6\times10^{-12}$), and enriched with nucleosomes ($t(5635) = 2.78$, $p = 5.4\times10^{-3}$) and TATA box sequences (hypergeometric test, $p = 3.9\times10^{-13}$) compared to all genes with an annotated promoter (methods).

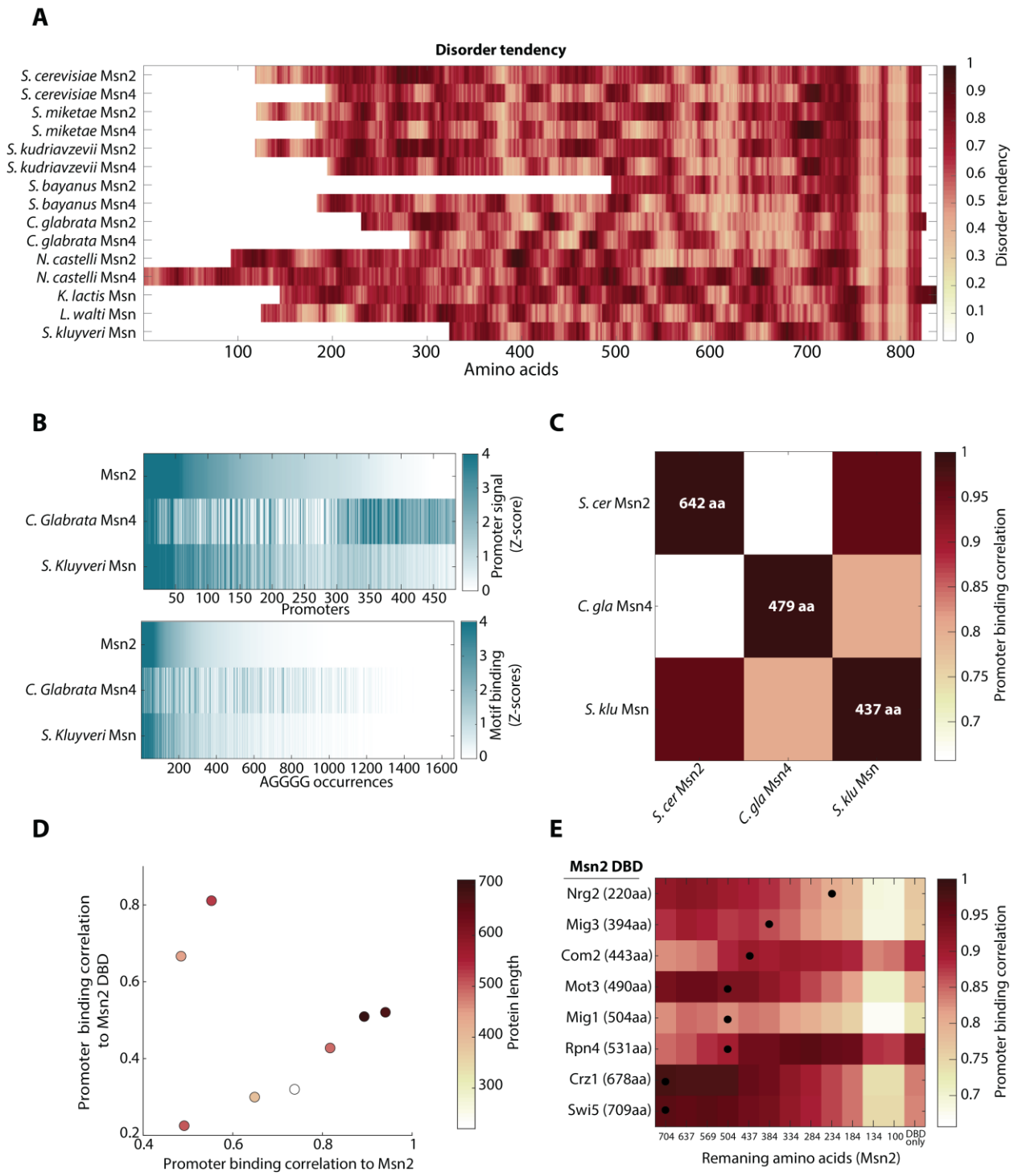

A
Disorder tendency
S. cerevisiae Msn2
S. cerevisiae Msn4
S. miketae Msn2
S. miketae Msn4
S. kudriavzevii Msn2
S. kudriavzevii Msn4
S. bayanus Msn2
S. bayanus Msn4
C. glabrata Msn2
C. glabrata Msn4
N. castelli Msn2
N. castelli Msn4
K. lactis Msn
L. walti Msn
S. kluyveri Msn
100 200 300 400 500 600 700 800
Amino acids
1 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0
Disorder tendency
B
Msn2
C. Glabrata Msn4
S. Kluyveri Msn
50 100 150 200 250 300 350 400 450
Promoters
4 3 2 1 0
Promoter signal (Z-score)
Msn2
C. Glabrata Msn4
S. Kluyveri Msn
200 400 600 800 1000 1200 1400 1600
AGGGG occurrences
4 3 2 1 0
Motif binding (Z-scores)
C
S. cer Msn2
C. gla Msn4
S. klu Msn
642 aa
479 aa
437 aa
S. cer Msn2
C. gla Msn4
S. klu Msn
1 0.95 0.9 0.85 0.8 0.75 0.7
Promoter binding correlation
D
0.8
0.6
0.4
0.2
0.4 0.6 0.8 1
Promoter binding correlation to Msn2
Promoter binding correlation to Msn2 DBD
700 600 500 400 300
Protein length
E
Msn2 DBD
Nrg2 (220aa)
Mig3 (394aa)
Com2 (443aa)
Mot3 (490aa)
Mig1 (504aa)
Rpn4 (531aa)
Crz1 (678aa)
Swi5 (709aa)
704 637 569 504 437 384 334 284 234 184 134 100 DBD only
Remaning amino acids (Msn2)
1 0.95 0.9 0.85 0.8 0.75 0.7
Promoter binding correlation

Figure S5: Protein size is not associated with changes in binding patterns, Related to Figure 5.

(A) *Orthologs of Msn2 and Msn4 contain long IDRs:* Shown are the disorder tendencies of the different orthologs aligned by their highly conserved DBD.

(B-C) *Msn2 orthologs of the same-lengths localize at different locations:* Shown in (B) are the binding strength to promoters (top) or sites containing the AGGGG motif (bottom). Each column represents a promoter or an AGGGG-containing site. Promoters/sites that were bound by at least one of the indicated TFs were chosen, and ordered by the binding strength of the *S. cerevisiae* Msn2 (top line). The two additional profiles correspond to strains in which the Msn2 non-DBD was replaced with the non-DBD of the indicated ortholog. The length of the non-DBDs is 642 aa in *S. cerevisiae*, and ~400 aa in both orthologs. Promoter binding correlations over all annotated promoters, are shown in (C).

(D-E) *Protein size is a poor predictor of similarity in binding pattern:* We considered eight swapped TFs in which we fused the Msn2 DBD to non-DBDs of different zinc-finger TFs. Shown in (D) is the correlation of these different factors with the intact Msn2 or the Msn2 DBD. Each point represents one swap, color-coded by the respective protein size. (E) shows the correlation of the indicated swaps to the series of Msn2 truncation constructs. Protein size is shown, and the similar-sized construct is indicated by a black dot.

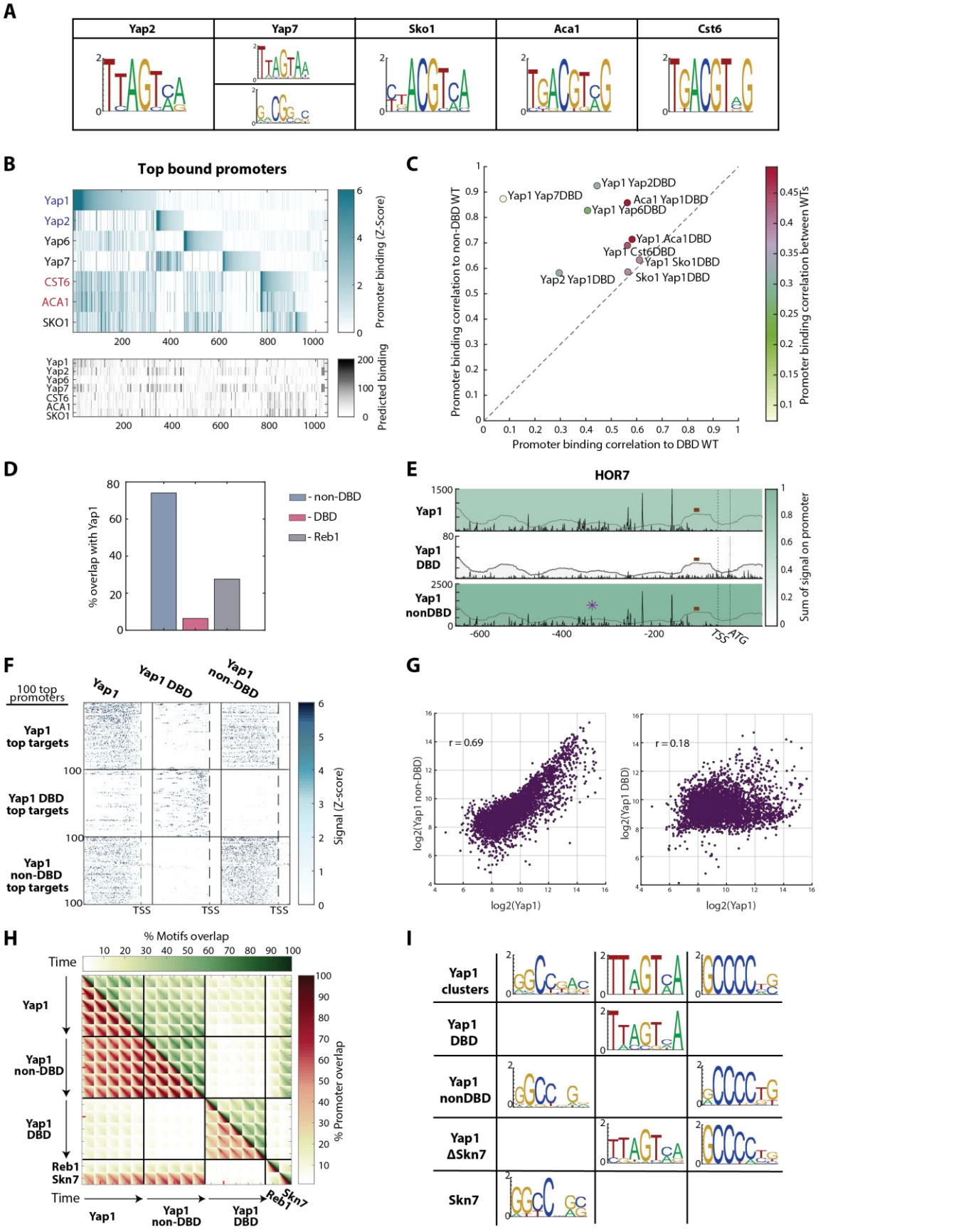

A
Yap2
Yap7
Sko1
Aca1
Cst6
B
Top bound promoters
Yap1
Yap2
Yap6
Yap7
CST6
ACA1
SKO1
Promoter binding (Z-Score)
Predicted binding
C
Yap1 Yap2DBD
Yap1 Yap7DBD
Aca1 Yap1DBD
Yap1 Yap6DBD
Yap1 Aca1DBD
Yap1 Cst6DBD
Yap1 Sko1DBD
Yap2 Yap1DBD
Sko1 Yap1DBD
Promoter binding correlation to non-DBD WT
Promoter binding correlation to DBD WT
Promoter binding correlation between WTs
D
% overlap with Yap1
- non-DBD
- DBD
- Reb1
E
HOR7
Yap1
Yap1 DBD
Yap1 nonDBD
Sum of signal on promoter
TSS
ATG
F
100 top promoters
Yap1
Yap1 DBD
Yap1 non-DBD
Yap1 top targets
Yap1 DBD top targets
Yap1 non-DBD top targets
Signal (Z-score)
TSS
G
r = 0.69
r = 0.18
log2(Yap1 non-DBD)
log2(Yap1 DBD)
log2(Yap1)
H
% Motifs overlap
Time
Yap1
Yap1 non-DBD
Yap1 DBD
Reb1
Skn7
% Promoter overlap
I
Yap1 clusters
Yap1 DBD
Yap1 nonDBD
Yap1 ΔSkn7
Skn7

Figure S6: The Yap1 DBD dictates motif preferences while its non-DBD governs promoter selection, as in the case of Msn2, Related to Figure 6.

(A) Similar *motif preferences of the bZIP TFs:* Shown are the PWMs of the motifs preferred by the different factors.

(B) *Wild-type TFs from the bZIP family bind distinct promoter sets:* Same as Figure 1C for the indicated factors, close paralogs are indicated by colors. The predicted promoter binding, based on motif preferences of each factor, is shown in the lower panel (see methods). Note that one pair of close paralogs, Cst6 and Aca1, shows highly similar promoter preferences, while Yap1 and Yap2, another pair, bind distinct targets.

(C) *Promoter selection in the bZIP family is mainly dictated by the non-DBDs*: The promoter binding profile (sum of signal on promoter, over all promoters) of each swap was correlated to that of the wild-type containing its DBD (x-axis) and the wild-type containing its non-DBD (y-axis). Color indicates the correlation between the two wildtypes. Notably, promoter selection of all swapped strains is more similar to the wild-type containing their non-DBD.

(D) *The non-DBD binds the majority of Yap1 bound promoters*: The percentage of overlap in defined binding targets (see methods) of the TFs indicated by color to the target promoters of Yap1 was calculated by normalizing to the smaller group.

(E) *Binding of Yap1, its DBD and non-DBD mutants to the HOR7 promoter:* Binding along the HOR7 promoter is shown, presentation as in Figure 1F. Note that deletion of the Yap1 DBD barely affects the overall binding strength (sum of signal on promoter is indicated by background color).

(F-G) *Yap1 recognizes its target promoters even in the absence of its DBD:* Binding patterns to the top bound promoters are shown in (F), same presentation as in Figure S1C. Promoter binding strength comparisons between Yap1 and its mutants are shown in (G), each dot represents a promoter.

(H) *Overlap in promoter and motif preferences of Yap1 and its two mutants in different MNase activation durations:* Presentation as in Figure S3C, for MNase activation times of 20,30,60,90 and 600 seconds. Red lines indicate the cleavage time used for results shown in the main text.

(I) *The Binding of Yap1 to its canonical motif depends on its DBD, while binding to two other motifs depend on its non-DBD*: Shown are the PWMs of the indicated factors as defined by their binding profiles after k-means clustering (see methods). Note that one cluster of the Yap1 non-DBD is identical to the Skn7 consensus motif, and is lost when deleting Skn7.

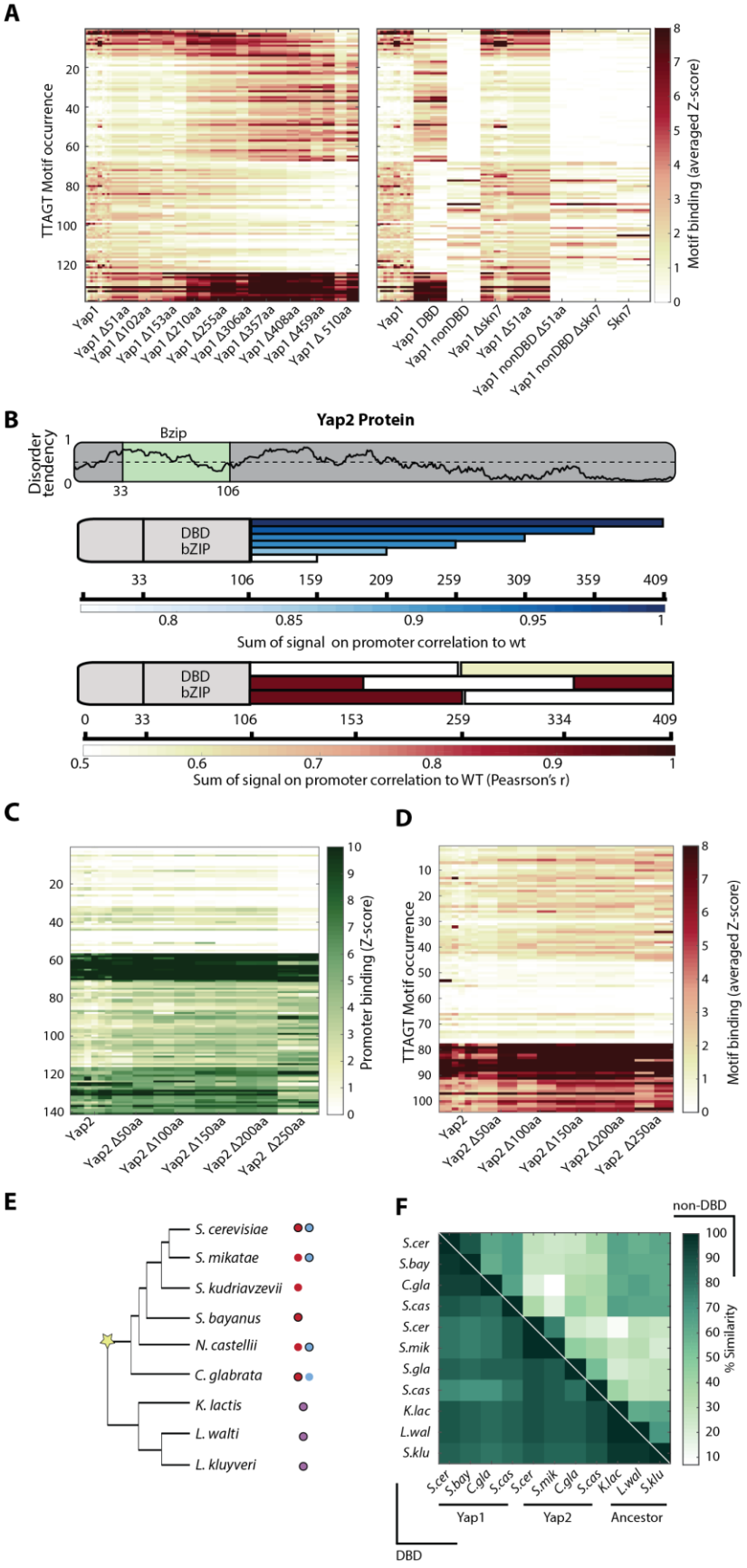

A
TTAGT Motif occurrence
Motif binding (averaged Z-score)
Yap1
Yap1 Δ51aa
Yap1 Δ102aa
Yap1 Δ153aa
Yap1 Δ210aa
Yap1 Δ255aa
Yap1 Δ306aa
Yap1 Δ357aa
Yap1 Δ408aa
Yap1 Δ459aa
Yap1 Δ 510aa
Yap1 DBD
Yap1 nonDBD
Yap1 Δskn7
Yap1 Δ51aa
Yap1 nonDBD Δ51aa
Yap1 nonDBD Δskn7
Skn7
B
Yap2 Protein
Disorder tendency
Bzip
DBD
bZIP
Sum of signal on promoter correlation to wt
Sum of signal on promoter correlation to WT (Peasrson's r)
C
Promoter binding (Z-score)
Yap2
Yap2 Δ50aa
Yap2 Δ100aa
Yap2 Δ150aa
Yap2 Δ200aa
Yap2 Δ250aa
D
TTAGT Motif occurrence
Motif binding (averaged Z-score)
E
S. cerevisiae
S. mikatae
S. kudriavzevii
S. bayanus
N. castellii
C. glabrata
K. lactis
L. walti
L. kluyveri
F
non-DBD
% Similarity
S.cer
S.bay
C.gla
S.cas
S.mik
K.lac
L.wal
S.klu
Yap1
Yap2
Ancestor
DBD

Figure S7: The Yap2 non-DBD has a minor role in promoter selection, Related to Figure 6.

(A) *Yap1 C-terminal truncation results in both loss and gain of binding sites:* Binding strength is quantified by averaged Z-scores (see methods). TTAGT motifs bound by at least one of the indicated factors (averaged Z-score > 3) were chosen and clustered based on their binding pattern along truncation.

(B) *The Yap2 non-DBD barely affects promoter selection:* A scheme of the Yap2 protein is shown on top, the black line indicates the predicted disorder tendency along the Yap2 protein (Dosztanyi et al., 2005a, 2005b). The scheme below shows the mutants used in our analysis. This includes gradual ~50 aa C-terminal truncations and a set of overlapping ~150 aa deletions spread across the Yap2 sequence (excluding its DBD). Color intensity indicates the promoter selection correlation between the respective mutants and Yap2.

(C) *Yap2 C-terminal truncation has a minor effect on promoter selection:* Promoters bound by at least one of the indicated factors (Z-score > 3) were chosen and clustered based on their binding pattern along truncations. Shown is the binding strength of the different truncations to the chosen 140 promoters.

(D) Yap2 *C-terminal truncation results in a minor change of binding site selection:* Same presentation as (A) for Yap2.

(E) *Species chosen for Yap1/Yap2 ortholog analysis:* Yap1 and Yap2 orthologs from each species are indicated in red and blue dots respectively, and the single Yap pre-duplication ancestral variant is shown in purple. The non-DBDs of Yap1 and Yap2 in *S. cerevisiae* were replaced with counterparts taken from the Yap1/Yap2 orthologs that are marked with black circles. The phylogenetic tree shown is based on (Shen et al., 2016). The whole genome hybridization (WGH) event that lead to the emergence of Yap1 and Yap2 is marked with a yellow star.

(F) *Sequence similarity between Yap1 and Yap2 orthologs:* Amino acid sequence similarity was measured separately for the DBD (lower triangle) and for the rest of the protein (upper triangle). Note the high conservation of the DBD, which contrasts the rapid divergence of the non-DBD. Moreover, Yap1 orthologs are more conserved and are more similar to the pre-WGH orthologs than Yap2 orthologs.

Table S1: Domain annotations. Related to STAR Methods.

Factor	DBD	Non-DBD
Saccharomyces cerevisiae Yap1	55-88	89-650
Saccharomyces cerevisiae Yap2	34-67	68-409
Saccharomyces cerevisiae Yap6	212-245	
Saccharomyces cerevisiae Yap7	116-149	
Saccharomyces cerevisiae Aca1	374-408	
Saccharomyces cerevisiae Cst6	416-449	
Saccharomyces cerevisiae Sko1	420-453	
Kluyveromyces waltii Yap1/2	24-56	57-584
Kluyveromyces lactis Yap1/2	43-75	76-583
Saccharomyces kluyveri Yap1/2	48-80	81-567
Candida glabrata Yap1	16-49	50-588
Saccharomyces bayanus Yap1	59-88	89-645
Candida glabrata Yap2	Oct-43	44-486
Saccharomyces mikatae Yap2	18-51	52-393
Saccharomyces castelii Yap2	22-55	56-475
Saccharomyces cerevisiae Msn2	643-702	
Saccharomyces cerevisiae Nrg2	149-209	
Saccharomyces cerevisiae Mig1	34-94	
Saccharomyces cerevisiae Mig3	13-73	
Saccharomyces cerevisiae Com2	385-443	
Saccharomyces cerevisiae Swi5	546-608	
Saccharomyces cerevisiae Crz1	565-623	
Saccharomyces cerevisiae Rpn4	433-511	
Saccharomyces cerevisiae Mot3	342-401	
Saccharomyces cerevisiae Msn2	643-702	1-642
Saccharomyces mikatae Msn2	642-701	1-641
Saccharomyces kudriavzevii Msn2	643-702	1-642
Saccharomyces bayanus Msn2	267-326	1-266
Candida glabrata Msn2	531-590	1-530
Saccharomyces castelii Msn2	761-820	1-760
Saccharomyces cerevisiae Msn4	569-628	1-568
Saccharomyces kudriavzevii Msn4	567-626	1-566
Saccharomyces bayanus Msn4	578-637	1-577
Candida glabrata Msn4	480-539	1-479
Saccharomyces castelii Msn4	669-728	1-668
Kluyveromyces lactis Msn2/4	617-676	1-616
Kluyveromyces waltii Msn2/4	637-696	1-636
Saccharomyces kluyveri Msn2/4	438-497	1-437