Cell Systems

Article

Selective association of short tandem repeats with DNA-binding domains and intrinsically disordered regions of transcription factors

Highlights

- Why short tandem repeats (STRs) are enriched in regulatory DNA regions is unknown
- We measured transcription factor (TF) binding to libraries of 2–5 bp STRs within cells
- TF-binding preferences depend on STR identity
- Msn2 STR preferences depend on its long (>500 aa) intrinsically disordered region

Authors

Matan Vidavski, Sagie Brodsky, Wajd Manadre, Tamar Jana Lang, Vladimir Mindel, Yoav Navon, Naama Barkai

Correspondence

naama.barkai@weizmann.ac.il

In brief

Vidavski et al. explore how short tandem repeats (STRs), frequently found in regulatory DNA, influence transcription factor (TF) binding. Using STR libraries in yeast, they show that TFs bind specific STRs with distinct preferences. For Msn2, this specificity arises from both its DNA-binding domain and a long intrinsically disordered region.

Vidavski et al., 2025, Cell Systems *16*, 101349
August 20, 2025
https://doi.org/10.1016/j.cels.2025.101349

CellPress

Cell Systems

CellPress

Article

Selective association of short tandem repeats with DNA-binding domains and intrinsically disordered regions of transcription factors

Matan Vidavski,[1,2] Sagie Brodsky,[1,2] Wajd Manadre,[1] Tamar Jana Lang,[1] Vladimir Mindel,[1] Yoav Navon,[1] and Naama Barkai[1,3,*]

[1]Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel
[2]These authors contributed equally
[3]Lead contact
*Correspondence: naama.barkai@weizmann.ac.il
https://doi.org/10.1016/j.cels.2025.101349

SUMMARY

Short tandem repeats (STRs) are enriched in regulatory regions and can bind transcription factors (TFs), as shown for selected examples *in vitro*. Here, we use a library-based assay to systematically measure TF binding to STRs of 2–5 bp units within budding yeast cells. We examined STR binding by four TFs, including Msn2, and further tested six Msn2 mutants, including two that contained only the DNA-binding domain (DBD) or only the 642-aa intrinsically disordered region (IDR). We find substantial STR effects on motif-dependent and motif-independent binding, which varied between TFs. For Msn2, STR association was explained by the DBD binding at motif half-sites and the IDR favoring homopurine-homopyrimidine and AT-rich repeats. TF-preferred STRs are enriched in the human genome but remain at low frequency at yeast promoters. We discuss the implications of our results for understanding the role of STRs and their crosstalk with TF IDRs in regulating TF binding across genomes.

INTRODUCTION

Transcription factors (TFs) bind to gene regulatory regions using DNA-binding domains (DBDs) that recognize specific DNA sequence motifs. However, DBD-preferred motifs are too short and abundant to explain TF locations in genomes: only a minority of motif sites are bound, and TFs are also found at sites containing weak or no motifs.[1–3] How TFs distinguish their genomic targets within the multitude of motif sites available in genomes challenges our understanding of gene regulation at the genomic scale.[4–11]

TFs are enriched in intrinsically disordered regions (IDRs) that can extend to hundreds of residues, in some cases occupying the full TF sequence outside the DBD (non-DBD).[12–19] Recently, we found that IDRs direct TF binding in genomes,[1,20–22] which was perhaps unexpected given the low interaction specificity associated with disordered sequences. Msn2 of budding yeast presents a well-studied example; its >500-aa IDR contains multiple weak and partially redundant determinants that are together required for Msn2 to bind its genomic targets.[20,23] Recent reports questioned the contribution of TF-TF interactions, and of nucleosomes, in directing Msn2 locations across the genome, leaving the molecular basis of IDR-directed binding unclear.[23–25]

In this study, we extend our analysis of IDR-directed binding to short tandem repeats (STRs). STRs of 2–6 bp are abundant polymorphic elements of eukaryotic genomes and their expansion underlies severe degenerative diseases still lacking effective therapy.[26–33] STR toxicity can arise from altered protein function and can also result from STR-induced noncanonical DNA structures that impede DNA replication or repair.[34–39] STRs are enriched in human enhancers and yeast promoters. They were implicated in gene regulation by association studies that linked STR length variations with gene expression.[40–44]

Direct binding of TFs to selected STRs was demonstrated *in vitro*.[45–50] The role of IDRs in the TF-STR association was not addressed systematically, but specific examples have provided some conflicting indications with *in vitro* examples. These findings indicate that the full-length TF binds STRs more weakly than its DBD-only mutant.[50] On the other hand, the EWSR-FLI oncogene is directed to bind (and form new enhancers) at GGAA repeats by the fusion of a disordered protein (EWSR) to the FLI TF.[51–54] FLI therefore presents a case of IDR-directed STR binding, which may be of more general use.

In this work, we tested systematically for TF-STR binding inside cells, focusing on Msn2 as a model for IDR-directed TFs (Figure 1A). We used massive parallel binding assay (MPBA), a method we recently reported for parallel analysis of TF binding to thousands of designed DNA sequences.[55] Using MPBA, we followed the binding of several TFs across 2–5 bp unit STRs. Our data revealed that STRs modulate both the motif-dependent and motif-independent DNA binding in a manner that varied between STRs and between tested TFs. For Msn2, it reveals that both its DBD and its disordered non-DBD contribute to the

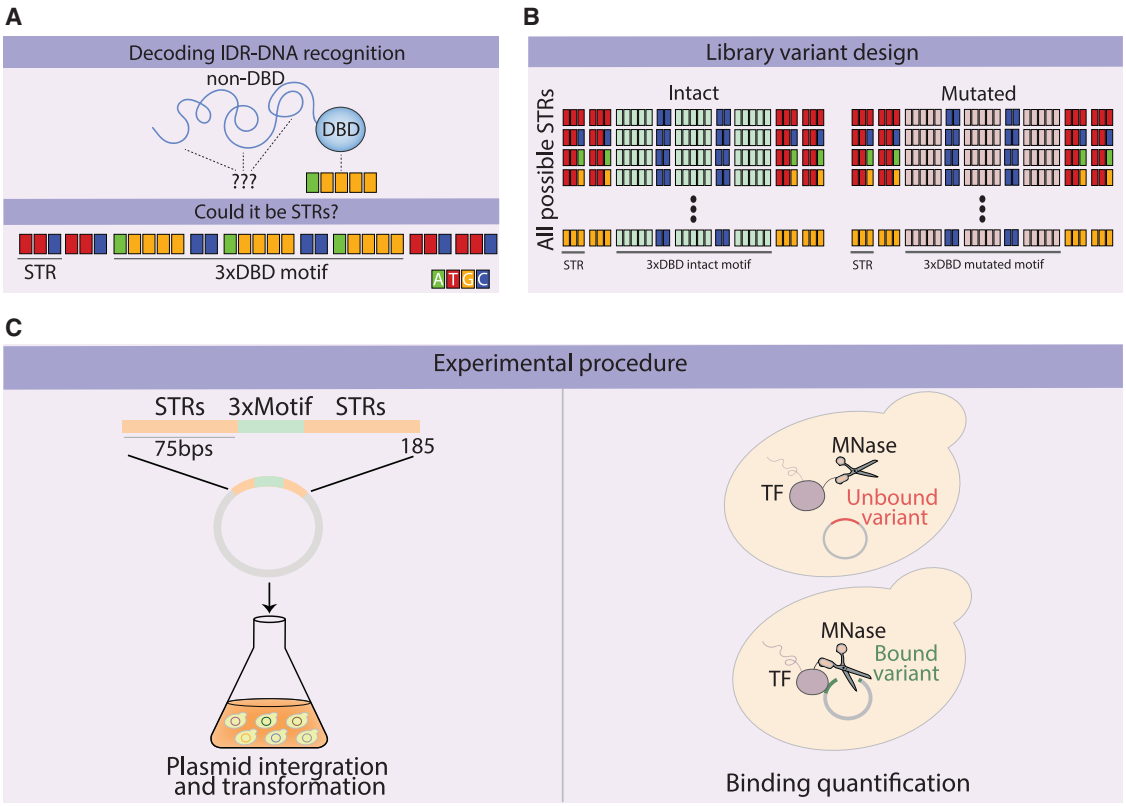

Figure 1. A library-based assay for measuring TF binding across STRs

(A) Decoding IDR-DNA recognition: we asked whether TF IDRs direct binding at STRs.

(B) Design of the STR sequence library: libraries were designed for each TF of interest. Sequences in the libraries were of 185 bp, each including three intact TF-binding motifs or their mutated version at the center. On both sides, the motifs were flanked by 75 bp of a given STR. Depending on the TF tested, our designed libraries covered all possible 2–4 or 2–5 bp repeats.

(C) Measuring TF binding across an STR library using MPBA: the library of designed sequences (Table S1) was processed as a pool. Sequences were integrated into a plasmid and transformed into a yeast strain containing the TF of interest fused to an MNase (Table S2). A short calcium pulse activated the MNase, triggering the cleavage of TF-bound sequences. This cleavage is then quantified using high-throughput sequencing, comparing the PCR-amplified library before and after MNase activation. Finally, each sequence was assigned a binding score corresponding to its depletion in the MNase-activated pool. Note that negative scores correspond to high depletion and therefore strong binding, whereas positive ones correspond to low binding.

Msn2-STR associations: the DBD localizes Msn2 to its motif half-site, while the IDR biases Msn2 binding at homopurine-homopyrimidine and AT-rich repeats. Comparing our TF-STR-binding data to the distributions of STRs in genomes, we find that STR abundance in the human genome correlated with measured TF-binding scores and that TF-bound STRs were also enriched in TF-bound yeast promoters, although remaining of low frequency and low repeat number. We discuss our results within a proposed working model of IDR-directed DNA binding

RESULTS

A library-based assay for measuring TF binding across STRs

We compared TF binding across libraries of STRs using the MPBA.[55] Each sequence in the library included 150 bp STR of 3–5 bp units, at the center of which we positioned three intact or mutated binding motifs of the tested TF (Figure 1B). We measured TF binding across this library as follows. First, the synthesized sequences were integrated as a pool to a designed prepared plasmid and transformed into budding yeast cells that expressed the TF of interest fused to an MNase (Figure 1C; Table S1). Second, cleavage of TF-bound sequences was triggered by a short calcium pulse that activated the fused MNase. Third, using PCR and sequencing we quantified the relative abundance of each sequence within the pre- and post-activated pools, providing a measure for the cleavage-triggered loss of each STR. Finally, a TF-binding score was assigned to each sequence, measuring the loss (normalized fold-change) in abundance, averaged over all independent replicates and equivalent STRs.

Motif binding by Cbf1 and Reb1 is modified by motif-flanking STRs

We tested our approach using Reb1 and Cbf1 as general regulatory factors (GRFs) that show tight and specific motif binding.[56] We applied MPBA[55] to respective STR libraries of 2–4 bp units, retrieving 562–580 sequences (84%–87%), with 390–440 (58%–66%) receiving sufficient coverage for quantifying abundances (Figure S1A). Abundance data were similar

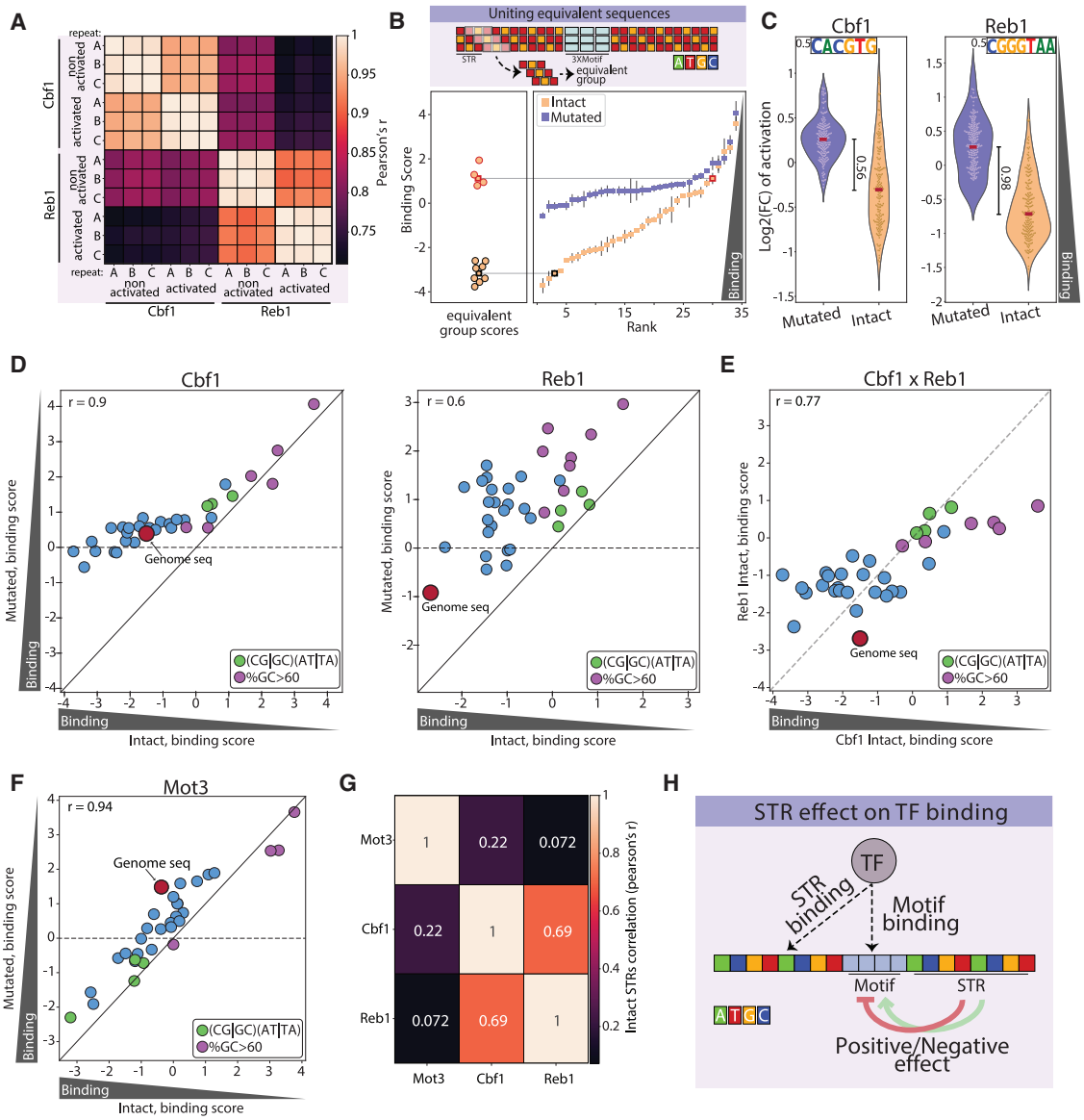

Figure 2. STRs modulate binding preferences of Cbf1 and Reb1

(A) Reproducibility of MPBA sequencing data: MPBA was applied to measure the binding of Cbf1 and Reb1 to respective 2–4 bp STR libraries. The correlation between abundances of sequence reads before and after MNase activation is shown, as indicated. Repeats are assigned letters (A–C). Samples are either MNase activated or non-activated.

(B) Binding scores are reproducible between equivalent STRs: we included in our libraries separated representations for reverse complement and cyclic-related STRs, which all code for essentially the same double-stranded DNA (dsDNA), as shown in the top scheme (e.g., GTG, GGT, TGG, and CCA). For subsequent analysis, sequences within each equivalent set were grouped together, using their median binding score and represented by a single sequence. An example of two such equivalent sets is shown on the bottom left panel. The median score of each set is seen on the bottom right panel. Shown in the bottom right panel are the median Cbf1-binding scores and the standard error within all equivalent groups, ordered by their score rank in the sample ranking. Samples colored and ranked separately for motif-containing and motif-lacking sequences.

(C) Stronger binding at motif-containing sequences: shown are the distributions of binding scores of motif-containing and motif-lacking sequences, as indicated. Included are individual sequences before the grouping of equivalent groups. Outliers, defined by the bottom and top 2.5%, are not presented.

(D) Motif binding of Reb1 and Cbf1 is limited by a similar set of STRs: the scatterplots compare Cbf1-binding (left) and Reb1-binding (right) scores at same-STR sequences containing intact (*x* axis) or mutated (*y* axis) motifs, as indicated. The colors indicate C/G-rich STRs (green), 4-letter STR palindromes (purple), and the endogenous genomic sequence, introduced to the library and plasmid as a control sequence (red).

(E) STR effects show moderate similarity between Cbf1 and Reb1: shown is a comparison of the binding of Reb1 and Cbf1 to same-STR sequences containing their respective motifs. Colors as in (D).

(legend continued on next page)

between the pre- and post-activated pools, reflecting biases in library composition. Of note, our measure of TF binding relied on the fold-change in sequence abundances and was therefore invariant to this bias, although it did limit our ability to test certain STRs, including pure GC or AT ones that were poorly synthesized or lost by perturbed replication[57,58] (Figure S1C). The fold-changes in sequence abundances were reproducible between repeats (c = 0.99) and consistent among STRs related by reverse-complementation or cyclic permutations, which define, essentially, the same DNA sequence (Figures 2A, 2B, and S1B). STR-binding scores were assigned by grouping together all these equivalent sequences and averaging their (normalized) fold-change in abundance (STAR Methods).

Sequences with intact motifs received binding scores that were higher by an average of ~2-fold (Reb1) or 1.5-fold (Cbf1) compared with sequences containing mutated motifs (Figures 2C and 2D). Note that a binding score of 2 corresponds to a 4-fold-reduced abundance, corresponding to a loss of ~80% of respective sequences. Motif effects varied between STRs; for example, motifs that were embedded within palindromic STRs (e.g., "ACGT") failed to increase TF binding. Most motif-lacking sequences received moderate binding scores, except notable low binding scores at GC-rich STRs (>60%). Residual binding to these STRs that lacked a motif was not explained by partial similarity to the known Cbf1 or Reb1 motifs (Figure S1D). For Cbf1, this binding to motif-lacking STRs was correlated with respective motif-containing STRs (r = 0.9; Figure 2D), and these were partially reproduced in the Reb1-mutated-motif data (r = 0.77; Figure 2E). Finally, and perhaps unexpectedly, sequences replacing the motif-flanking STRs by genomic regions flanking Reb1- or Cbf1-bound sites received top (Reb1)- or high (Cbf1)-binding scores (Figures 2D and 2E).

We also added Mot3 to our initial testing as a single-specific TF showing strong motif binding.[55] Contrasting the GRFs, Mot3-STR data were dominated by STR rather than motif effects (Figures 2F and S1D–S1H), and these were distinct from STRs favored by Reb1 or Cbf1 (r = 0.22 and 0.07; Figure 2G); The "TGCA" palindromic repeat, for example, was poorly bound by both Reb1 and Cbf1 but received a high Mot3-binding score, likely reflecting its high overlap with the Mot3 motif (STR—ATGCA, binding motif—AGG[TC]A; Figure S1D). Together, we conclude that MPBA can compare TF binding across STRs, revealing motif-dependent and motif-independent effects with STR preferences that differ depending on TF and repeat identities (Figure 2H).

Msn2 binds to specific STRs

Our study was motivated by our interest in IDRs of TFs, asking whether these disordered regions contribute to selective TF-STR association. For this, we turned to Msn2 as a model for a TF whose genomic binding is guided by its IDR (Figures 1A and 3A). We synthesized a library of 2–5 bp STRs flanking three intact or mutated Msn2 motifs. Processing these through MPBA using eight partially overlapping pools, we achieved sufficient coverage for 1,280 (67%) sequences (Figures S2A and S2B). The (fold) change in library composition following MNase activation was reproducible between repeats and between equivalent STRs, and these were grouped together and averaged to define STR-binding scores, as above (Figures 3B and 3C). Notably, free-MNase control gave only a weak binding, and its STR preferences showed no similarity with these of Msn2 (c = 0.28; Figures 3D, S2C, and S2D). As for Msn2, a genomic sequence included in the library was among the top-bound sequences (Figure 3E).

Binding scores at motif-containing sequences were higher by an average of ~1.3-fold (Figures 3E and S2E). By comparison, binding scores varied by ~4-fold across STRs. These STR effects were largely similar between sequences containing or lacking the Msn2 motif (r = 0.89; Figures 3F and 3G) and were not due to the immediate motif-flanking sequences as they were largely diminished in libraries modulating only the motif vicinity (Figure S5).

We previously found that Msn2 acts as a major recruiter of Med15, a component of the mediator coactivator.[24,25] First, when tested across the genome, Med15 localizes to Msn2-bound promoters. Further, when deleting Msn2 and its Msn4 homologs, Med15 was lost from Msn2-bound promoters.[24,25] We reasoned that we could therefore use Med15 to validate the Msn2-STR binding, predicting that Med15 would associate with Msn2-favored STRs when tested in wild-type cells but would lose this association if tested in cells deleted of Msn2/4. Indeed, as predicted, Med15 localized to Msn2-favored STRs (c = 0.72) but had lost these preferences in Msn2/4-deleted cells (c = −0.455). Together, these results support a specific Msn2-STR association that is sufficiently strong for cofactor recruitment.

Msn2-STR association depends on its DBD and on its disordered non-DBD

We asked whether Msn2 binds to STRs through its DBD only or whether its disordered non-DBD contributes to this binding. For this, we applied MPBA to Msn2 mutants containing only the DBD or only the 642-aa non-DBD region (Figure 4A). Both mutants were over-expressed, as we previously noted weak binding of natively expressed DBD to MPBA libraries.[55] STR-binding data of both mutants showed high reproducibility (Figures S3A–S3D), with binding signals that were stronger and distinct from free MNase (Figure S3E). The genomic sequence was again among the top-scoring sequences (Figure 4E).

The DBD and non-DBD STR-binding scores varied across STRs and were tightly correlated between STRs that contained or lacked the Msn2 motif (Figures 4C and 4D; r = 0.83 DBD, 0.93 non-DBD). Here, also, STR effects are not due to the immediate motif vicinity (Figure S5). As expected, the DBD localized favorably at motif-containing sequences (Figures 4B and 4C). Also,

(F) STR effects dominate the Mot3 binding pattern: shown is a comparison of the Mot3-binding scores at same-STR sequences containing intact or mutated motifs (left, as in D).
(G) STR preferences differ between Mot3 and Reb1 or Cbf1: shown is the correlation between all STR sequences containing intact motifs of the indicated TFs.
(H) STR effects on TF binding: STRs could modulate binding levels of TF motifs, as in the cases of Cbf1 and Reb1. Alternatively, TFs can bind STR sequences independently of their motif state, as in the case of Mot3.

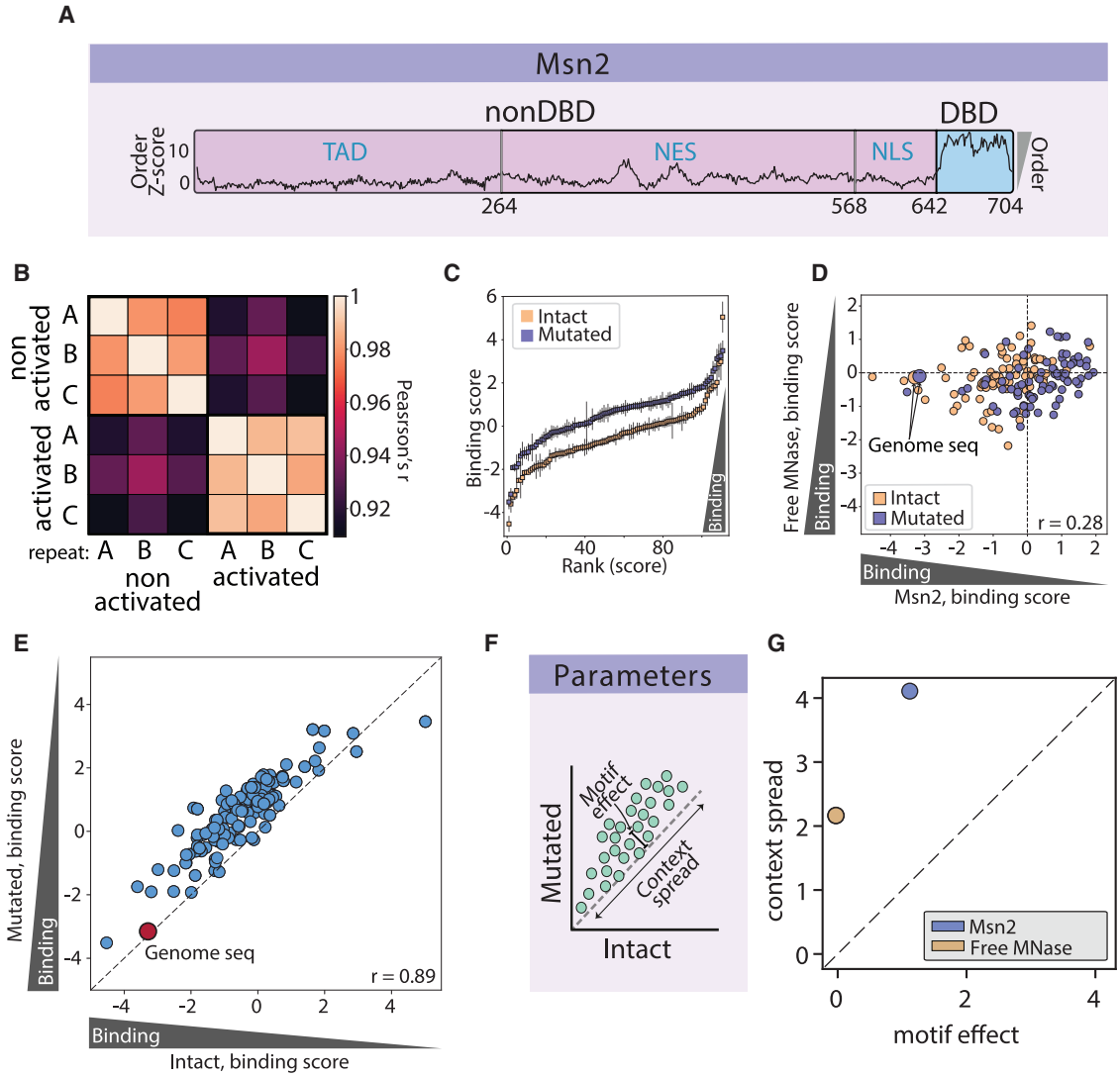

Figure 3. Msn2 binds specific STRs independent of its motif

(A) Msn2 is a highly disordered TF: shown is the predicted disorder of the Msn2 sequence, as calculated using ADOPT.[59] Locations of the main functional domains are indicated, including the DBD, nuclear export and localization signals (NES and NLS), and the transcription activation domain (TAD).

(B and C) Reproducibility of Msn2-binding scores: shown in (B) is the correlation between repeats and time points, as in Figure 2A. The binding scores of equivalent STRs, as in Figure 2B, are shown in (C). Note the stronger binding at motif-containing sequences.

(D) Msn2-binding scores show no similarity to those of a free-MNase control: the binding of an over-expressed free MNase was tested on the Msn2 STR library. The scatterplot compares the binding scores of each STR, as obtained in cells bearing an Msn2-MNase fusion (*x* axis) and ones bearing a free MNase (*y* axis). Sequences containing intact or mutated motifs are colored orange and purple, respectively.

(E–G) Msn2-binding scores are dominated by STR rather than motif effects: the scatterplot (E) compares Msn2-binding scores at sequences containing intact and mutated motifs. STR and motif effects, as defined in (F), are summarized in (G).

as expected, this bias for the DBD-favored motif was significantly weaker for the non-DBD. Of note, the residual motif bias seen for non-DBD was dependent on Msn4 (Figure S4E), which, otherwise, had a limited effect on the non-DBD binding across STRs (Figure S4B; c = 0.8). We conclude from both the DBD and the non-DBD that localization is reproducible across STRs.

STR preferences differed between the DBD and non-DBD (r = 0.54; Figure 4E), and both were partially correlated with the intact Msn2 (r = 0.63 DBD, 0.68 non-DBD; Figure 4E). Therefore, the DBD and non-DBD both contribute to Msn2-STR association. To further validate this non-DBD contribution, we tested three Msn2 non-DBD mutants that we previously found to alter Msn2 binding across the genome[23] (N→H, DEKR→N, and LIV→Y, each changing 94, 90, and 94 aa within the non-DBD). As predicted, these mutations altered STR preferences, with the N→H being the most effective (c = 0.1, when comparing motif-containing sequences; Figures 4F–4I). We conclude that both the DBD and the non-DBD contribute to Msn2-STRs association.

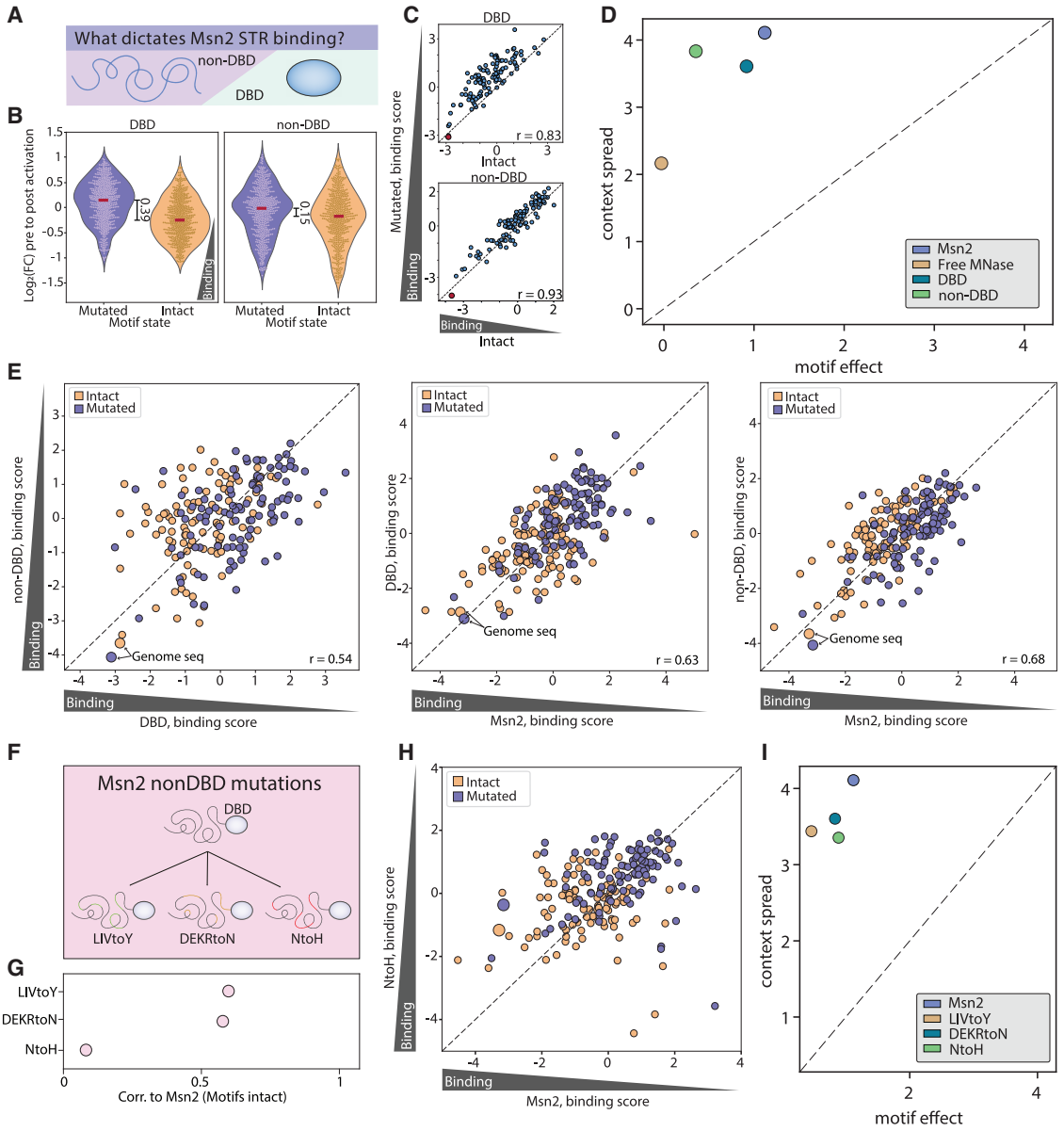

Figure 4. Msn2 STR preferences are equally dependent on its DBD and its non-DBD

(A) Msn2 binding at various STRs could be dictated by its DBD or could also be influenced by its non-DBD: a scheme.

(B) The DBD favors sequences with intact motifs: shown are the distributions of binding scores of motif-containing and motif-lacking sequences for the DBD (left) and non-DBD (right). Included are individual sequences before the grouping of equivalent groups. Outliers, defined by the bottom and top 2.5%, are not presented.

(C and D) Binding of the DBD and the non-DBD is dominated by STR effects: the scatterplots in (C) compare the binding scores of sequences with intact (*x* axis) or mutated (*y* axis) Msn2 motifs for the DBD (top) and non-DBD (bottom), with a red circle representing the genomic control sequence chosen from an Msn2-bound regulatory region. A summary of this data is shown in (D), as in Figures 3F and 3G.

(E) STR preferences of the DBD and non-DBD capture distinct aspects of Msn2: the scatterplots compare the binding scores of the DBD and those of the non-DBD (left), as well as the score received for Msn2 and both its mutants, as indicated (middle and right). Sequences containing intact or mutated motifs are colored orange and purple, respectively.

(F) IDR mutations introduced into Msn2: a scheme describing IDR mutations tested and presented in the following plots. The 3 IDR mutations replace groups of amino acids in the IDR as mentioned.

(G) IDR mutations change STR preferences: correlation of Msn2 and its IDR mutations for STR binding.

(H) IDR amino acid switch shift STR preferences: STR-binding scores of Msn2 (*x* axis) vs. N to H IDR mutant of Msn2 (*y* axis). Sequences containing intact or mutated motifs are colored orange and purple, respectively.

(I) STR and motif effects remain similar: a summary of motif and STR (context) effects for Msn2 and its IDR mutants, as described in Figure 3F and shown in (D) and Figure 3G above.

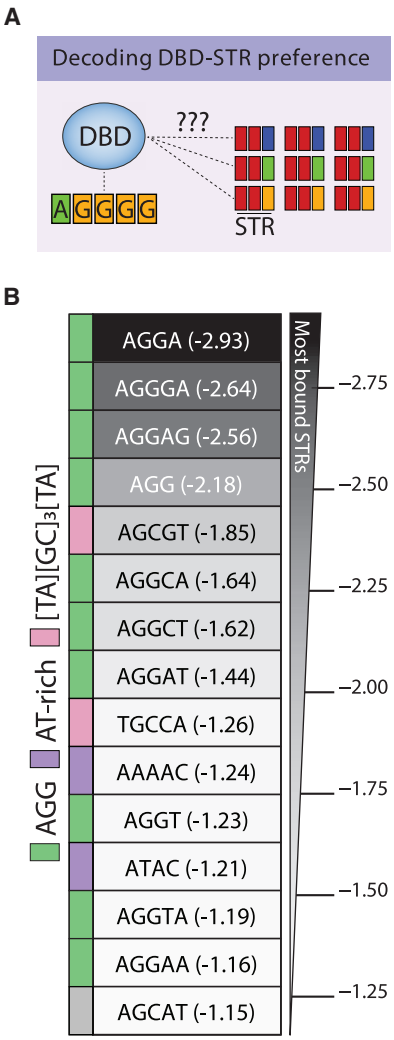

Figure 5. Low-affinity DBD binding to motif half-sites explains its STRs preferences

(A) Identity of DBD-favored STRs may be informative for their binding mechanism: a scheme.

(B) DBD-favored STRs: shown are the top 15 DBD-bound STRs. Presented scores are the average of each STR over its motif-containing and motif-lacking sequences. Colors indicate different STR groups—AGG-containing (green), [TA](G/C)$_3$[TA] containing (pink), and AT-rich (purple).

(C and D) AGG is the most enriched triplet within DBD-favored STRs: each possible triplet of bases was assigned a score depending on the binding scores assigned to all STRs in which it appears, by taking the median of these scores (C, STAR Methods). Scores were measured separately for sequences containing intact or mutated motifs and compared in the scatterplot (D).

(E) A model for direct DBD-AGG binding: a scheme. See the text for details.

DBD-STR association is explained by binding to Msn2 motif half-sites, whereas the non-DBD favors AT-rich and homopurine-homopyrimidine repeats

The Msn2 DBD is a member of the C_2H_2 zinc-finger family (Figure 5A). We noted that the first three letters of its motif, **AGG**GG, are found in the four top-bound STRs as well as the 10/15 top-bound ones (Figure 5B). Other highly bound STRs include two with C/G triplets and AT-rich ones. Further, a systematic analysis identified AGG as the most enriched triplet among DBD-bound STRs, whereas CGC was the second-top within motif-lacking sequences (Figures 5C and 5D). Notably, AGG was previously suggested to bind DBDs of the Msn2 family in an altered docking geometry.[60] Therefore, DBD is likely associated with STR through low-affinity binding to its partial motif (Figure 5E).

The AGG triplet was also found among the top Msn2-bound STRs, although at a lower frequency (5/15 top-bound STRs; Figure 6A). In all five cases, the AGG was followed by adenosine ("A"), and four of these were among the top 15 STRs bound by the DBD. This DBD-related bias in Msn2 preferences was also seen when comparing Msn2 and its non-DBD preferences across STRs that contained AGG, those that contained [CG]$_3$, or those that were AT-rich (Figure 6B). More generally, Msn2 and its non-DBD-favored STRs of high AT content (r = 0.48 and 0.4; Figures 6C and 6D), whereas the DBD showed no such preference (r = −0.1). Notably, the N→H mutant reversed this bias, preferring STRs of high (>60%) GC content (Figure 6D).

AT-bias accounted for much of the shared STR preferences of the non-DBD and Msn2 but did not explain all top-bound STRs. In the case of the non-DBD, these included homopurine-homopyrimidine stretches of varying AT content, with the most favored ones being AGGAG, AG, AAAGG, AGA, and GGAA (Figures 6A–6C). We conclude that although the DBD localizes at AGG triplets, the non-DBD appears to bias Msn2 binding to repeats of homopurine-homopyrimidine and AT-rich stretches (Figure 6E).

TF-favored STRs are abundant in the human genome and in TF-bound yeast promoters

Unexpectedly, we noted that STRs favoring TF binding in our data are highly abundant in the human genome. These two unrelated measures were, in fact, correlated, with maximal correlation seen for the non-DBD (~0.53) contrasting no correlation for free MNase (shown significant for Msn2 non-DBD; Figures 7A, 7B, and S7). For example, the TF-disfavoring STR palindromes (e.g., ACGT) are depleted from the human genome,[50] whereas the favorably bound [AT] or homopurine-homopyrimidine are of high abundance.[50]

Although many factors could lead to this correlation, we still found it interesting and asked whether it is also present in budding yeast. We found that 263 (17%) STRs of 2–6 bp at 80% accuracy are promoter localized, and these are enriched in TF-bound promoters and dominated by homopurine-homopyrimidine and AT-rich STRs (Figure 7C), mirroring IDR-STR

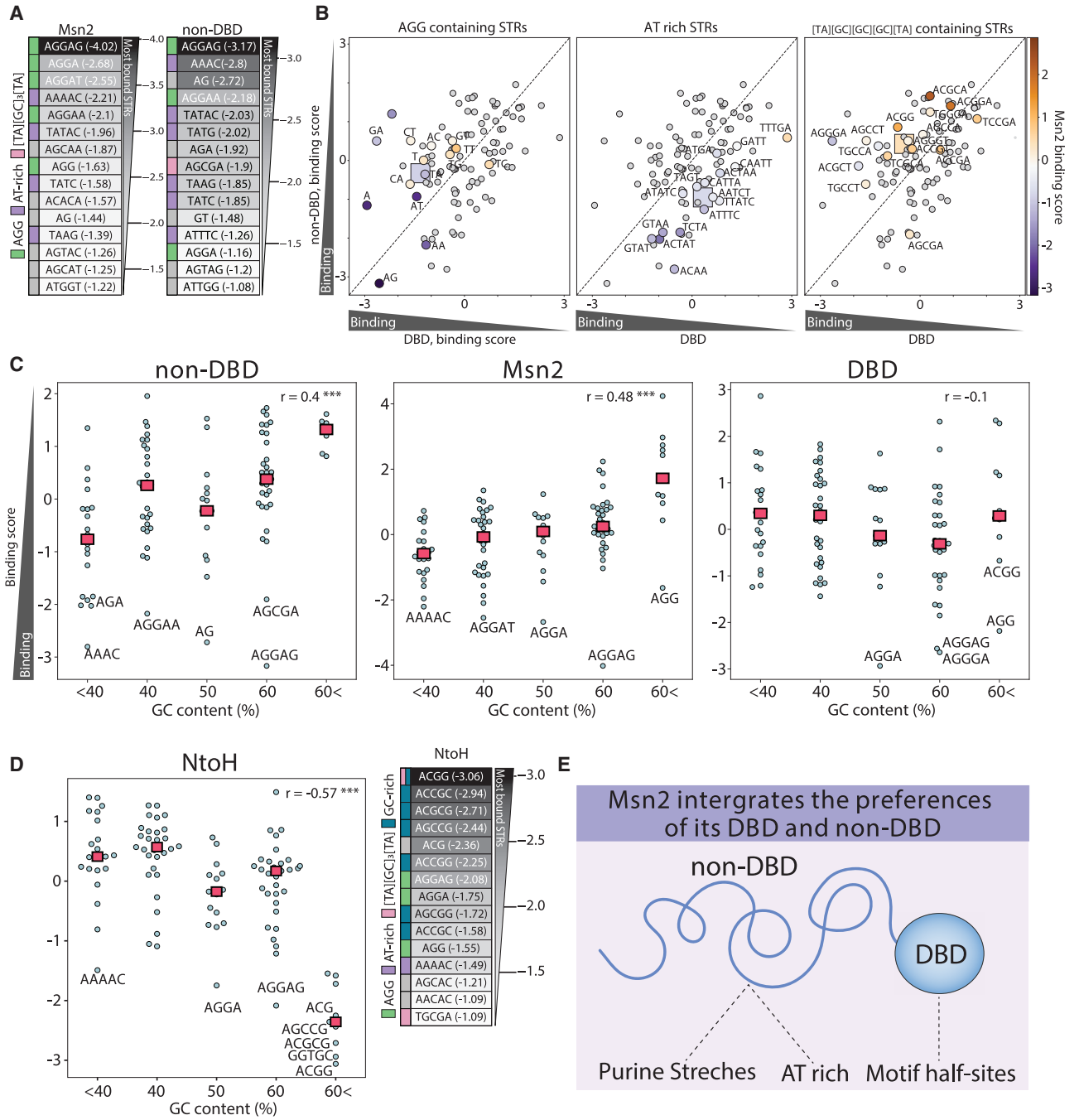

Figure 6. The non-DBD biases Msn2 binding toward high AT content and homopurine-homopyrimidine STRs

(A) STRs favored by Msn2 and its disordered non-DBD: shown are the 15 top-bound STRs for Msn2 (left) and its non-DBD (right). Presented scores are the average of each STR over its motif-containing and motif-lacking sequences. Colors indicate different STR groups—AGG-containing (green), [TA] $(G/C)_3$[TA] containing (pink), AT-rich (purple), and GC-rich (blue).

(B) The non-DBD bias Msn2 binding among the DBD-preferred STRs: the three scatterplots compare the average STR-binding scores received over motif-containing and motif-lacking sequences for the Msn2 non-DBD (*x* axis) and DBD (*y* axis) across different STRs. In each of the three plots, colors indicate the Msn2-binding scores at three selected sub groups of STRs, as indicated: AGG-containing (left, annotated by the STR extension following the AGG triplet), STRs of high AT content (middle), and [TA] $(C/G)_3$ [TA] containing STRs (right). Squares indicate the average score of each colored group of STRs.

(C) Homopurine-homopyrimidine STRs show strong non-DBD binding independent of their AT content: STRs were separated into groups based on their GC content. Shown are the binding scores of the three tested factors, highlighting the strongly bound repeats. Presented scores are the average of each STR over motif-containing and motif-lacking sequences. Squares represent the median of each of the indicated GC content groups. Pearson's correlation *p* values are Msn2: 1.337e−7, Msn2 DBD: 0.2698, and Msn2 non-DBD: 2.89e−5.

(legend continued on next page)

preferences. Notwithstanding this enrichment of TF-favoring STRs in TF-bound promoters (Figure 7D), their frequency remained low (<7%), even at the top TF-occupied ones. Related to that, we noted that top-bound repeats, including AGG, were not enriched in bound promoters or next to Msn2-binding sites and TF-bound promoters were of a low [AT] content.

STR effects depend on the tandem repeat arrangement

The low frequency of TF-bound repeats in gene promoters raised for us the question of whether TF-STR association observed in our assay results from the direct binding to the presented repeats or is a consequence of their tandem arrangements. To examine this further, we constructed new libraries in which we inserted non-tandem repeats into three different backbone sequences. These libraries contained all 3–4 bp repeats and also included several 5 bp repeats that showed strong binding in our earlier experiments. Note that the number of repeat units was reduced by 2-fold because we retained the same overall sequence lengths and that, by design, the flanking regions differed between the backbones. This led to weaker MPBA binding signals for all three factors tested: Msn2 and its DBD and non-DBD mutants.

To overcome the increased noise and inability to average over cyclic-related repeats, which were no longer equivalent, we averaged the binding data over the three promoter backbones (see STAR Methods). Most repeats that showed top binding by Msn2 and its DBD were explained as full or partial motifs (14 and 11 of the top 15, respectively; Figure S5), including preferences for the AGG triplet that explained DBD-STR binding. The non-DBD, by contrast, showed limited binding at these motif-associated repeats (binding score of −0.06 to the full motif and −0.02 to the half motif), including homopurine or AT-rich ones. Instead, the non-DBD localized to repeats of pure or high GC content, which were absent from the tandem STR data, including alternating [GC]. Accordingly, repeat preferences were well correlated between tandem and non-tandem arrangements in the case of DBD (c = 0.42–0.48), intermediate for the full Msn2, and lost for the non-DBD (c = 0). Together, we conclude that STR effects on TF binding depend not only on repeats of low-affinity binding sites but also on their tandem arrangement.

DISCUSSION

There is much interest in defining DNA features that direct TF binding across genomes. IDRs of TFs direct genome binding through multiplicity of weak, redundant determinants spread across their sequence. We hypothesized that these IDRs recognize similar redundant short DNA sequences of weak effects. A recent study further suggested to us that these are IDR-recognized DNA sequences, as the authors have traced TF-STR association to low-affinity binding of the individual repeat units.[50] We therefore decided to screen STRs for those that are recognized by TFs and their IDRs. In this, we noted that, in addition to presenting multitudes of low-affinity binding sites, STRs could influence TF binding through induced changes in DNA shape.[34–39,61]

We focused on Msn2 as a model for IDR-dependent TF. Msn2 showed selective preference for binding certain STRs, and this preference was explained by complementary contributions of its DBD and its disordered non-DBD. When tested individually, the DBD and non-DBD localized across STRs with comparable strengths and spread, yet they differed in their STR preferences. Further, STR preferences of both the DBD and the non-DBD were similarly mirrored in the Msn2 STR binding pattern.

The DBD of Msn2 localized at STRs that included its motif half-site, AGG. This is consistent with previous *in vitro* results in which DBD-STR binding was explained by low-affinity binding at individual repeats.[50] By contrast, STR preferences of the disordered non-DBD were not explained by any short sequences but instead revealed a general bias toward AT-rich STRs and certain homopurine-homopyrimidine stretches. Based on these results, we find it less likely that IDR preferences are explained by low-affinity binding to individual repeats. Rather, we favor the alternative possibility that the non-DBD recognizes some repeat-induced global alteration in DNA helical shape, perhaps being generally more accessible for TF binding. Indeed, when testing binding across a library of non-tandem repeats, the DBD retained similar repeat preferences as seen for the STR, whereas repeat preferences of the non-DBD differed between tandem and non-tandem repeat arrangements. Together, these results support the notion that the IDR associates with STRs indirectly, through their effect on DNA helical structure, a possibility that will be explored in future studies.

We unexpectedly noted that TF-favored STRs were highly abundant within the human genome. This correlation between STR-binding scores and genomic abundance was highest for the Msn2 non-DBD (r = 0.53). Although this may indicate evolutionary selection for IDR-bound STRs, we find it unlikely given the low frequency of STRs in yeast promoters. Rather, we hypothesize that this link between TF binding and genomic abundance might reflect some STR property that independently influences both properties. Consistent with that, others have shown that abundances of STRs correlate with a decreased likelihood of DNA polymerase stalling[62] or folding into stable hairpins or quadruplex,[63] both of which indicate reduced stability of the DNA helix that could influence IDR (or DBD) binding, for example, by sequences potentiating low base-stacking interactions.[64]

Contrasting the mammalian genome, STRs are not widely present in the compact yeast genome. Within promoters, STRs show 2.2-fold enrichment in TF-bound promoters relative to TF-unbound ones. STRs are therefore insufficient to explain the preferred IDR (or TF) binding at these promoters. It is still notable, however, that STRs found in TF-bound promoters are compatible with the IDR-preferred features, including a pronounced bias for homopurine-homopyrimidine and AT-rich repeats. These repeats may, therefore, promote IDR binding in

(D) IDR N to H Msn2 mutant shifts STR preferences toward GC-rich sequences: binding scores of the Msn2 N to H mutant (*y* axis) separated by the GC content of the STR (*x* axis). Pearsons's correlation $p = 3e{-}10$. (Left, same as in C.) Msn2 N to H mutant top 15 bound sequences (right).
(E) The Msn2 STR preferences are guided by both its DBD and non-DBD: while the DBD directs Msn2 to STRs containing its preferred motif half-sites, its disordered non-DBD biases the binding of homopurine-homopyrimidine and AT-rich STRs.

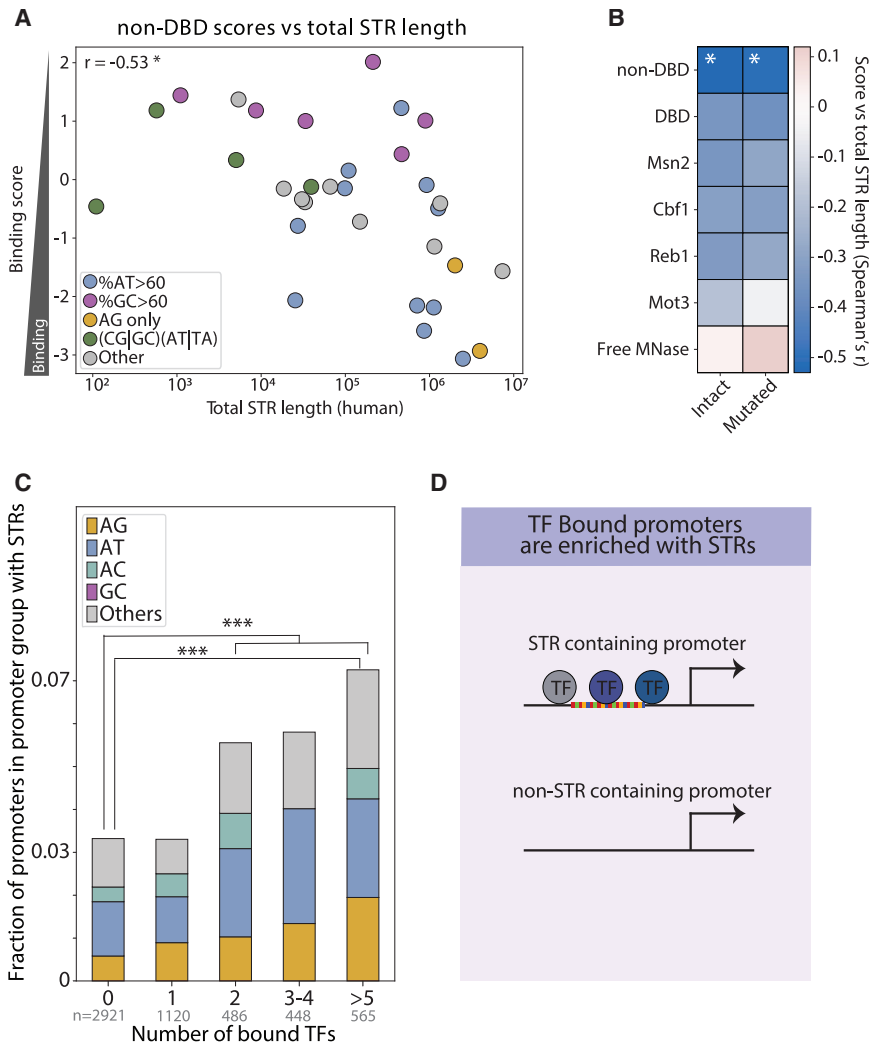

Figure 7. TF-bound STRs are abundant in the human genome and in TF-bound yeast promoters

(A and B) Favorable TF STRs are abundant in the human genome: shown in (A) is the binding score received for the Msn2 non-DBD as a function of the total STR length found in the human genome. Colors indicate different STR types, as indicated; Spearman's correlation value = (−0.53), $p = 0.002$. Shown in (B) is the correlation between the scores received for each tested factor and the total STR length in the human genome. Note the lack of correlation of the free-MNase control. p values (corrected) are reported in Figure S7.

(C and D) TF-bound yeast promoters are enriched with IDR-preferred STRs: shown in (C) is the fraction of yeast promoters containing STRs, separated by the number of bound TFs, as defined by our lab binding compendium (STAR Methods). Note the enrichment of AT-rich and homopurine-homopyrimidine STRs, favored by the disordered Msn2 non-DBD.

In (C), a Fisher exact test was done, comparing promoters bound by 0 TFs to (1) promoters bound by 5 TFs or more ($p = 4.76e{-5}$) and (2) promoters bound by 2 TFs or more ($p = 1.02e{-5}$). Shown in (D) is a scheme summarizing the results.

promoters where they are present, whereas, in other promoters, IDR recognition might be enabled by non-repeated sequences potentiating the same property, e.g., changes in DNA helical structure or DNA breathing. Although consistent with our findings, further studies are required to test this working model.

RESOURCE AVAILABILITY

Lead contact

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

Materials availability

All materials generated in this study are available from the lead contact upon request.

Data and code availability

- All original data have been deposited at https://github.com/sagieb/context and are publicly available as of the date of publication. DOIs are listed in the key resources table.
- All original code has been deposited at https://github.com/sagieb/ and is publicly available as of the date of publication. DOI: https://doi.org/10.5281/zenodo.15069135.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

We would like to thank Connor Horton for the help and data. We would also like to thank our lab members for generating a great scientific atmosphere and fruitful discussion. This project was funded by the ISF, the ERC, and the Minerva Center.

AUTHOR CONTRIBUTIONS

Conceptualization, M.V., S.B., T.J.L., and N.B.; methodology, M.V., S.B., T.J.L., and N.B.; investigation, M.V., S.B., W.M., V.M., and N.B.; visualization, M.V., S.B., and N.B.; funding acquisition, N.B.; project administration, M.V., S.B., and N.B.; supervision, N.B.; writing, M.V., S.B., and N.B.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

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

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cels.2025.101349.

Received: March 7, 2024
Revised: December 3, 2024
Accepted: July 3, 2025
Published: July 29, 2025

REFERENCES

1. Jana, T., Brodsky, S., and Barkai, N. (2021). Speed-Specificity Trade-Offs in the Transcription Factors Search for Their Genomic Binding Sites. Trends Genet. *37*, 421–432. https://doi.org/10.1016/j.tig.2020.12.001.
2. Neph, S., Vierstra, J., Stergachis, A.B., Reynolds, A.P., Haugen, E., Vernot, B., Thurman, R.E., John, S., Sandstrom, R., Johnson, A.K., et al. (2012). An expansive human regulatory lexicon encoded in transcription factor footprints. Nature *489*, 83–90. https://doi.org/10.1038/nature11212.
3. Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T.W., Greven, M.C., Pierce, B.G., Dong, X., Kundaje, A., Cheng, Y., et al. (2012). Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Res. *22*, 1798–1812. https://doi.org/10.1101/gr.139105.112.
4. 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. https://doi.org/10.1016/j.gde.2017.02.007.
5. Todeschini, A.L., Georges, A., and Veitia, R.A. (2014). Transcription factors: specific DNA binding and specific gene regulation. Trends Genet. *30*, 211–219. https://doi.org/10.1016/j.tig.2014.04.002.
6. Pan, Y., Tsai, C.-J., Ma, B., and Nussinov, R. (2010). Mechanisms of transcription factor selectivity. Trends Genet. *26*, 75–83. https://doi.org/10.1016/j.tig.2009.12.003.
7. Kribelbauer, J.F., Rastogi, C., Bussemaker, H.J., and Mann, R.S. (2019). Low-Affinity Binding Sites and the Transcription Factor Specificity Paradox in Eukaryotes. Annu. Rev. Cell Dev. Biol. *35*, 357–379. https://doi.org/10.1146/annurev-cellbio-100617-062719.
8. Badis, G., Berger, M.F., Philippakis, A.A., Talukder, S., Gehrke, A.R., Jaeger, S.A., Chan, E.T., Metzler, G., Vedenko, A., Chen, X., et al. (2009). Diversity and complexity in DNA recognition by transcription factors. Science *324*, 1720–1723. https://doi.org/10.1126/science.1162327.
9. Weirauch, M.T., Yang, A., Albu, M., Cote, A.G., Montenegro-Montero, A., Drewe, P., Najafabadi, H.S., Lambert, S.A., Mann, I., Cook, K., et al. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. Cell *158*, 1431–1443. https://doi.org/10.1016/j.cell.2014.08.009.
10. Wunderlich, Z., and Mirny, L.A. (2008). Spatial effects on the speed and reliability of protein–DNA search. Nucleic Acids Res. *36*, 3570–3578. https://doi.org/10.1093/nar/gkn173.
11. Staller, M.V. (2022). Transcription factors perform a 2-step search of the nucleus. Genetics *222*, iyac111. https://doi.org/10.1093/genetics/iyac111.
12. Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F., and Jones, D.T. (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J. Mol. Biol. *337*, 635–645. https://doi.org/10.1016/j.jmb.2004.02.002.
13. 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. https://doi.org/10.1021/bi0602718.
14. 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. https://doi.org/10.1016/j.jmb.2006.04.016.
15. Peng, Z., Mizianty, M.J., and Kurgan, L. (2014). Genome-scale prediction of proteins with long intrinsically disordered regions. Proteins *82*, 145–158. https://doi.org/10.1002/prot.24348.
16. van der Lee, R., Buljan, M., Lang, B., Weatheritt, R.J., Daughdrill, G.W., Dunker, A.K., Fuxreiter, M., Gough, J., Gsponer, J., Jones, D.T., et al. (2014). Classification of Intrinsically Disordered Regions and Proteins. Chem. Rev. *114*, 6589–6631. https://doi.org/10.1021/cr400525m.
17. Skupien-Rabian, B., Jankowska, U., Swiderska, B., Lukasiewicz, S., Ryszawy, D., Dziedzicka-Wasylewska, M., and Kedracka-Krok, S. (2016). Proteomic and bioinformatic analysis of a nuclear intrinsically disordered proteome. J. Proteomics *130*, 76–84. https://doi.org/10.1016/j.jprot.2015.09.004.
18. Wang, C., Uversky, V.N., and Kurgan, L. (2016). Disordered nucleiome: Abundance of intrinsic disorder in the DNA- and RNA-binding proteins in 1121 species from Eukaryota, Bacteria and Archaea. Proteomics *16*, 1486–1498. https://doi.org/10.1002/pmic.201500177.
19. Holehouse, A.S., and Kragelund, B.B. (2024). The molecular basis for cellular function of intrinsically disordered protein regions. Nat. Rev. Mol. Cell Biol. *25*, 187–211. https://doi.org/10.1038/s41580-023-00673-0.
20. Brodsky, S., Jana, T., Mittelman, K., Chapal, M., Kumar, D.K., Carmi, M., and Barkai, N. (2020). Intrinsically Disordered Regions Direct Transcription Factor In Vivo Binding Specificity. Mol. Cell *79*, 459–471.e4. https://doi.org/10.1016/j.molcel.2020.05.032.
21. Brodsky, S., Jana, T., and Barkai, N. (2021). Order through disorder: The role of intrinsically disordered regions in transcription factor binding specificity. Curr. Opin. Struct. Biol. *71*, 110–115. https://doi.org/10.1016/j.sbi.2021.06.011.
22. Kumar, D.K., Jonas, F., Jana, T., Brodsky, S., Carmi, M., and Barkai, N. (2023). Complementary strategies for directing in vivo transcription factor binding through DNA binding domains and intrinsically disordered regions. Mol. Cell *83*, 1462–1473.e5. https://doi.org/10.1016/j.molcel.2023.04.002.
23. Jonas, F., Carmi, M., Krupkin, B., Steinberger, J., Brodsky, S., Jana, T., and Barkai, N. (2023). The molecular grammar of protein disorder guiding genome-binding locations. Nucleic Acids Res. *51*, 4831–4844. https://doi.org/10.1093/nar/gkad184.
24. Lupo, O., Kumar, D.K., Livne, R., Chappleboim, M., Levy, I., and Barkai, N. (2023). The architecture of binding cooperativity between densely bound transcription factors. Cell Syst. *14*, 732–745.e5. https://doi.org/10.1016/j.cels.2023.06.010.
25. Mindel, V., Brodsky, S., Cohen, A., Manadre, W., Jonas, F., Carmi, M., and Barkai, N. (2024). Intrinsically disordered regions of the Msn2 transcription factor encode multiple functions using interwoven sequence grammars. Nucleic Acids Res. *52*, 2260–2272. https://doi.org/10.1093/nar/gkad1191.

26. Weber, J.L., and Wong, C. (1993). Mutation of human short tandem repeats. Hum. Mol. Genet. *2*, 1123–1128. https://doi.org/10.1093/hmg/2.8.1123.

27. Kruglyak, S., Durrett, R.T., Schug, M.D., and Aquadro, C.F. (1998). Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc. Natl. Acad. Sci. USA *95*, 10774–10778. https://doi.org/10.1073/pnas.95.18.10774.

28. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921. https://doi.org/10.1038/35057062.

29. Ellegren, H. (2004). Microsatellites: simple sequences with complex evolution. Nat. Rev. Genet. *5*, 435–445. https://doi.org/10.1038/nrg1348.

30. Sharma, P.C., Grover, A., and Kahl, G. (2007). Mining microsatellites in eukaryotic genomes. Trends Biotechnol. *25*, 490–498. https://doi.org/10.1016/j.tibtech.2007.07.013.

31. Gemayel, R., Vinces, M.D., Legendre, M., and Verstrepen, K.J. (2010). Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annu. Rev. Genet. *44*, 445–477. https://doi.org/10.1146/annurev-genet-072610-155046.

32. Bilgin Sonay, T.B., Carvalho, T., Robinson, M.D., Greminger, M.P., Krützen, M., Comas, D., Highnam, G., Mittelman, D., Sharp, A., Marques-Bonet, T., and Wagner, A. (2015). Tandem repeat variation in human and great ape populations and its impact on gene expression divergence. Genome Res. *25*, 1591–1599. https://doi.org/10.1101/gr.190868.115.

33. Ziaei Jam, H., Li, Y., DeVito, R., Mousavi, N., Ma, N., Lujumba, I., Adam, Y., Maksimov, M., Huang, B., Dolzhenko, E., et al. (2023). A deep population reference panel of tandem repeat variation. Nat. Commun. *14*, 6711. https://doi.org/10.1038/s41467-023-42278-3.

34. Contente, A., Dittmer, A., Koch, M.C., Roth, J., and Dobbelstein, M. (2002). A polymorphic microsatellite that mediates induction of PIG3 by p53. Nat. Genet. *30*, 315–320. https://doi.org/10.1038/ng836.

35. Mirkin, S.M. (2007). Expandable DNA repeats and human disease. Nature *447*, 932–940. https://doi.org/10.1038/nature05977.

36. Hannan, A.J. (2018). Tandem repeats mediating genetic plasticity in health and disease. Nat. Rev. Genet. *19*, 286–298. https://doi.org/10.1038/nrg.2017.115.

37. Sulovari, A., Li, R., Audano, P.A., Porubsky, D., Vollger, M.R., Logsdon, G.A., Human Genome Structural Variation Consortium, Warren, W.C., Pollen, A.A., Chaisson, M.J.P., et al. (2019). Human-specific tandem repeat expansion and differential gene expression during primate evolution. Proc. Natl. Acad. Sci. USA *116*, 23243–23253. https://doi.org/10.1073/pnas.1912175116.

38. Khristich, A.N., and Mirkin, S.M. (2020). On the wrong DNA track: Molecular mechanisms of repeat-mediated genome instability. J. Biol. Chem. *295*, 4134–4170. https://doi.org/10.1074/jbc.REV119.007678.

39. Malik, I., Kelley, C.P., Wang, E.T., and Todd, P.K. (2021). Molecular mechanisms underlying nucleotide repeat expansion disorders. Nat. Rev. Mol. Cell Biol. *22*, 589–607. https://doi.org/10.1038/s41580-021-00382-6.

40. Vinces, M.D., Legendre, M., Caldara, M., Hagihara, M., and Verstrepen, K.J. (2009). Unstable tandem repeats in promoters confer transcriptional evolvability. Science *324*, 1213–1216. https://doi.org/10.1126/science.1170097.

41. Sawaya, S., Bagshaw, A., Buschiazzo, E., Kumar, P., Chowdhury, S., Black, M.A., and Gemmell, N. (2013). Microsatellite tandem repeats are abundant in human promoters and are associated with regulatory elements. PLoS One *8*, e54710. https://doi.org/10.1371/journal.pone.0054710.

42. Yáñez-Cuna, J.O., Arnold, C.D., Stampfel, G., Boryń, L.M., Gerlach, D., Rath, M., and Stark, A. (2014). Dissection of thousands of cell type-specific enhancers identifies dinucleotide repeat motifs as general enhancer features. Genome Res. *24*, 1147–1156. https://doi.org/10.1101/gr.169243.113.

43. Gymrek, M., Willems, T., Guilmatre, A., Zeng, H., Markus, B., Georgiev, S., Daly, M.J., Price, A.L., Pritchard, J.K., Sharp, A.J., and Erlich, Y. (2016). Abundant contribution of short tandem repeats to gene expression variation in humans. Nat. Genet. *48*, 22–29. https://doi.org/10.1038/ng.3461.

44. Quilez, J., Guilmatre, A., Garg, P., Highnam, G., Gymrek, M., Erlich, Y., Joshi, R.S., Mittelman, D., and Sharp, A.J. (2016). Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans. Nucleic Acids Res. *44*, 3750–3762. https://doi.org/10.1093/nar/gkw219.

45. Willems, R., Paul, A., Van Der Heide, H.G., ter Avest, A.R., and Mooi, F.R. (1990). Fimbrial phase variation in Bordetella pertussis: a novel mechanism for transcriptional regulation. EMBO J. *9*, 2803–2809. https://doi.org/10.1002/j.1460-2075.1990.tb07468.x.

46. Iglesias, A.R., Kindlund, E., Tammi, M., and Wadelius, C. (2004). Some microsatellites may act as novel polymorphic cis-regulatory elements through transcription factor binding. Gene *341*, 149–165. https://doi.org/10.1016/j.gene.2004.06.035.

47. Martin, P., Makepeace, K., Hill, S.A., Hood, D.W., and Moxon, E.R. (2005). Microsatellite instability regulates transcription factor binding and gene expression. Proc. Natl. Acad. Sci. USA *102*, 3800–3804. https://doi.org/10.1073/pnas.0406805102.

48. Afek, A., Schipper, J.L., Horton, J., Gordân, R., and Lukatsky, D.B. (2014). Protein−DNA binding in the absence of specific base-pair recognition. Proc. Natl. Acad. Sci. USA *111*, 17140–17145. https://doi.org/10.1073/pnas.1410569111.

49. Afek, A., Cohen, H., Barber-Zucker, S., Gordân, R., and Lukatsky, D.B. (2015). Nonconsensus Protein Binding to Repetitive DNA Sequence Elements Significantly Affects Eukaryotic Genomes. PLoS Comput. Biol. *11*, e1004429. https://doi.org/10.1371/journal.pcbi.1004429.

50. Horton, C.A., Alexandari, A.M., Hayes, M.G.B., Marklund, E., Schaepe, J.M., Aditham, A.K., Shah, N., Suzuki, P.H., Shrikumar, A., Afek, A., et al. (2023). Short tandem repeats bind transcription factors to tune eukaryotic gene expression. Science *381*, eadd1250. https://doi.org/10.1126/science.add1250.

51. Grünewald, T.G.P., Bernard, V., Gilardi-Hebenstreit, P., Raynal, V., Surdez, D., Aynaud, M.M., Mirabeau, O., Cidre-Aranaz, F., Tirode, F., Zaidi, S., et al. (2015). Chimeric EWSR1-FLI1 regulates the Ewing sarcoma susceptibility gene EGR2 via a GGAA microsatellite. Nat. Genet. *47*, 1073–1078. https://doi.org/10.1038/ng.3363.

52. Gangwal, K., Sankar, S., Hollenhorst, P.C., Kinsey, M., Haroldsen, S.C., Shah, A.A., Boucher, K.M., Watkins, W.S., Jorde, L.B., Graves, B.J., and Lessnick, S.L. (2008). Microsatellites as EWS/FLI response elements in Ewing's sarcoma. Proc. Natl. Acad. Sci. USA *105*, 10149–10154. https://doi.org/10.1073/pnas.0801073105.

53. Patel, M., Simon, J.M., Iglesia, M.D., Wu, S.B., McFadden, A.W., Lieb, J.D., and Davis, I.J. (2012). Tumor-specific retargeting of an oncogenic transcription factor chimera results in dysregulation of chromatin and transcription. Genome Res. *22*, 259–270. https://doi.org/10.1101/gr.125666.111.

54. Boulay, G., Sandoval, G.J., Riggi, N., Iyer, S., Buisson, R., Naigles, B., Awad, M.E., Rengarajan, S., Volorio, A., McBride, M.J., et al. (2017). Cancer-Specific Retargeting of BAF Complexes by a Prion-like Domain. Cell *171*, 163–178.e19. https://doi.org/10.1016/j.cell.2017.07.036.

55. Lang, T.J., Brodsky, S., Manadre, W., Vidavski, M., Valinsky, G., Mindel, V., Ilan, G., Carmi, M., and Barkai, N. (2024). Massively Parallel Binding Assay (MPBA) reveals limited transcription factor binding cooperativity, challenging models of specificity. Preprint at bioRxiv. https://doi.org/10.1101/2024.06.26.600749.

56. Rossi, M.J., Lai, W.K.M., and Pugh, B.F. (2018). Genome-wide determinants of sequence-specific DNA binding of general regulatory factors. Genome Res. *28*, 497–508. https://doi.org/10.1101/gr.229518.117.

57. Kaushal, S., and Freudenreich, C.H. (2019). The role of fork stalling and DNA structures in causing chromosome fragility. Genes Chromosomes Cancer *58*, 270–283. https://doi.org/10.1002/gcc.22721.

58. Walsh, E., Wang, X., Lee, M.Y., and Eckert, K.A. (2013). Mechanism of Replicative DNA Polymerase Delta Pausing and a Potential Role for DNA Polymerase Kappa in Common Fragile Site Replication. J. Mol. Biol. *425*, 232–243. https://doi.org/10.1016/j.jmb.2012.11.016.

59. Redl, I., Fisicaro, C., Dutton, O., Hoffmann, F., Henderson, L., Owens, B.M.J., Heberling, M., Paci, E., and Tamiola, K. (2023). ADOPT: intrinsic protein

disorder prediction through deep bidirectional transformers. NAR Genom. Bioinform. *5*, lqad041. https://doi.org/10.1093/nargab/lqad041.

60. 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. https://doi.org/10.1016/j.molcel.2014.06.019.
61. Duy, D.L., and Kim, N. (2023). Yeast transcription factor MSN2 binds to G4 DNA. Nucleic Acids Res. *51*, 9643–9657. https://doi.org/10.1093/nar/gkad684.
62. Murat, P., Guilbaud, G., and Sale, J.E. (2020). DNA polymerase stalling at structured DNA constrains the expansion of short tandem repeats. Genome Biol. *21*, 209. https://doi.org/10.1186/s13059-020-02124-x.
63. Bacolla, A., Larson, J.E., Collins, J.R., Li, J., Milosavljevic, A., Stenson, P. D., Cooper, D.N., and Wells, R.D. (2008). Abundance and length of simple repeats in vertebrate genomes are determined by their structural properties. Genome Res. *18*, 1545–1553. https://doi.org/10.1101/gr.078303.108.
64. Nordén, B., and Takahashi, M. (2020). Understanding Rad51 function is a prerequisite for progress in cancer research. QRB Discov. *1*, e9. https://doi.org/10.1017/qrd.2020.13.
65. Bar-Ziv, R., Brodsky, S., Chapal, M., and Barkai, N. (2020). Transcription Factor Binding to Replicated DNA. Cell Rep. *30*, 3989–3995.e4. https://doi.org/10.1016/j.celrep.2020.02.114.
66. Chappleboim, M., Naveh-Tassa, S., Carmi, M., Levy, Y., and Barkai, N. (2024). Ordered and disordered regions of the origin recognition complex direct differential in vivo binding at distinct motif sequences. Nucleic Acids Res. *52*, 5720–5731. https://doi.org/10.1093/nar/gkae249.
67. Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., and Zhang, F. (2013). Multiplex Genome Engineering using CRISPR/Cas Systems. Science *339*, 819–823. https://doi.org/10.1126/science.1231143.
68. Anand, R., Memisoglu, G., and Haber, J. (2017). Cas9-mediated gene editing in *Saccharomyces cerevisiae*. Protocol Exchange. https://doi.org/10.1038/protex.2017.021a.
69. Mölder, F., Jablonski, K.P., Letcher, B., Hall, M.B., Tomkins-Tinch, C.H., Sochat, V., Forster, J., Lee, S., Twardziok, S.O., Kanitz, A., et al. (2021). Sustainable data analysis with Snakemake. F1000Res *10*, 33. https://doi.org/10.12688/f1000research.29032.2.
70. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. j. *17*, 3. https://doi.org/10.14806/ej.17.1.200.
71. Lindgreen, S. (2012). AdapterRemoval: easy cleaning of next-generation sequencing reads. BMC Res. Notes *5*, 337. https://doi.org/10.1186/1756-0500-5-337.
72. Foat, B.C., Tepper, R.G., and Bussemaker, H.J. (2008). TransfactomeDB: a resource for exploring the nucleotide sequence specificity and condition-specific regulatory activity of trans-acting factors. Nucleic Acids Res. *36*, D125–D131. https://doi.org/10.1093/nar/gkm828.
73. 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. https://doi.org/10.1186/1471-2105-7-113.
74. Chong, Y.T., Koh, J.L.Y., Friesen, H., Duffy, S.K., Cox, M.J., Moses, A., Moffat, J., Boone, C., and Andrews, B.J. (2015a). Yeast Proteome Dynamics from single cell imaging and automated analysis. Cell *161*, 1413–1424. https://doi.org/10.1016/j.cell.2015.04.051.
75. de Godoy, L.M.F., Olsen, J.V., Cox, J., Nielsen, M.L., Hubner, N.C., Fröhlich, F., Walther, T.C., and Mann, M. (2008). Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. Nature *455*, 1251–1254. https://doi.org/10.1038/nature07341.
76. Tkach, J.M., Yimit, A., Lee, A.Y., Riffle, M., Costanzo, M., Jaschob, D., Hendry, J.A., Ou, J., Moffat, J., Boone, C., et al. (2012). Dissecting DNA damage response pathways by analysing protein localization and abundance changes during DNA replication stress. Nat. Cell Biol. *14*, 966–976. https://doi.org/10.1038/ncb2549.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and virus strains		
E. cloni Electrocompetent Cells	Biosearch technologies	LC601172
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
SPRI beads AMPure XP	Beckman Coulter	Cat#A63881
Digitonin	Sigma Aldrich	Cat#300410
Spermine	Sigma Aldrich	Cat# S3256-5G
Spermidine	Sigma Aldrich	Cat# S026
T4 DNA ligase	NEB	M0202S
10x T4 DNA Ligase buffer	NEB	B0202S
Phusion High-Fidelity PCR Master Mix	NEB	M0531S
KAPA Hifi DNA polymerase	Roche	07958927001
Critical commercial assays		
HiYield Plasmid Maxi Kit	RBC Bioscience	YPM25
Deposited data		
Raw and Processed Data Generated in this Study	this study	https://doi.org/10.5281/zenodo.15069135
Raw Data of STR Length in the Human Genome	Horton et al.[50]	https://doi.org/10.1126/science.add1250.
Experimental models: Organisms/strains		
Yeast	This study	Table S2
Oligonucleotides		
Oligos used to produce plasmid libraries	This study	Table S1
Software and algorithms		
Original code used for this study	this study	https://doi.org/10.5281/zenodo.15069135
MATLAB	MathWorks	N/A
Jupyther Notebook	Project Jupyter	N/A

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

All experiments were done on wild type *S. Cerevisiae*. We used the BY4741 strain, of genotype MATa his3-Δ1 leu2-Δ0 lys2-Δ0 met15-Δ0 ura3-Δ0. Cells were stored frozen, in 80% glycerol - 20% YPD solution, at -80°c. Cells were grown in 30°c, either on agar plates (YPD or SD-ura) or in liquid media (YPD or SD-ura).

We used E. cloni Electrocompetent Cells (Biosearch technologies, LC601172) for plasmid amplification. Cells were stored in -80°c and grown in LB media in 37°c.

Sequence design

All library sequences used in this study are found in Table S1, including the genomic control sequence of each transcription factor (TF). All library sequences were structured as follows: Unique eight base barcode – 75 bps of upstream STRs – edited promoter sequence (including 3 TF preferred motif) – 75 bases of downstream STR. The promoters were chosen based on ChEC-seq binding data collected in our lab for each of the studied TFs.[20,22,65] Each such chosen region contains a bound motif of the relevant TF. A region of 35 base pairs around the motif was copied from the genome, and two additional motifs were inserted into it, one upstream and one downstream, five bases away from the central motif (The original genomic bases were replaced rather than shifted). To avoid generation of restriction sites of the enzymes used to prepare our libraries, a nucleotide was inserted between each library building block. In the mutated state, all three motifs had a point mutation in the middle of the motif. Common to all libraries were 34 sequences: 16 STRs of the length of 2 and a genomic sequence, one set with all 3 motifs intact and another where all are mutated. In the genomic sequence, if an extra motif was found outside the central 3 motifs it was mutated.

Sequence design: Non-tandem STRs

We used the same backbone described above of barcode – upstream sequence – 3 central motifs – downstream sequence, for the non - tandem STRs libraries. Here, to create the sequences, we copied 185 bases from genomic promoters of yeast, shown to bind Msn2 in former lab experiments.[20] STRs of 3-5 bases were inserted to all sequences. STRs were inserted to the selected sequences in 1 of 2 designs: a. in equal amounts upstream and downstream of the main 3 motifs, 7 STRs on either side of it, with constant spacing of 9 - 12 bases (“embedded”), b. 4 times on either side of the main 3 motifs, in clusters of 2 repeats (“flanking”).

DNA digestion and ligation

Libraries were ordered from Agilent (CAT: G7220A#230). Sequences were amplified using PCR with a Herculase II enzyme (Agilent CAT: 600677, 16 cycles, program as in protocol). Three reactions were performed for each library to lower PCR biases. The PCR products were run on an electrophoresis gel and the correct sized band was cut out and cleaned. Both our plasmid (see Table S1) and the PCR products were cut using the SexAI and AscI restriction enzymes (Thermo Fisher, CATs: FD2114 and FD1894), following the manufacturer recommendation for reaction conditions. The plasmid (Table S1, "plasmid sequence") also went through 5’ dephosphorylation using FastAP (Thermo Fisher, CAT: EF0651) to prevent self ligation. We cleaned the restriction reaction products using a QIAQuick PCR purification kit (Qiagen, CAT: 28004). The libraries and plasmids were ligated using a Fast-Link ligation kit (Lucigen, CAT: LK6201H). Then, the ligated plasmids were cleaned using a QIAQuick Minelute PCR purification kit (Qiagen, CAT: 28106).

Bacterial transformation and plasmid extraction

We transformed the plasmids into E. cloni Electrocompetent Cells (Biosearch technologies, LC601172) following the manufacturer’s protocol. Following the transformation, 250ml of warm LB were added to the cells to allow over-night growth in 37°c. We extracted the plasmids from the bacteria using the NucleoBond Xtra Maxi kit (Machery-Nagel, 740414.10), following manufacturer’s protocol.

Yeast transformation

We transformed the plasmids into yeast BY4741 strains, of genotype MATa his3-Δ1 leu2-Δ0 lys2-Δ0 met15-Δ0 ura3-Δ0 with a TF-MNase fusion (Table S2). We used the LiAc/Salmon Sperm DNA/PEG method: Yeast were plated on a YPD plate and one colony was taken and grown in liquid YPD to saturation over night. The yeast were diluted in the morning, 250μl of yeast to 12.5ml of YPD and grew for four more hours. The cells were washed with double distilled water (DDW) and LiAc 100 mM, before resuspension in a mix of 33% PEG 3350, 100 mM LiAc, single stranded salmon sperm DNA and 10 μg of plasmid DNA. The cells were incubated in 30°c for 30 minutes, before a 42°c heat shock of 30 minutes. The cells were pelleted and resuspended in 1 mL of SD – ura media. A small volume, 1/500 of the cells, was plated to estimate transformants numbers and the rest was grown in 30°c shaking, for minimum of 60 hours until reaching stationary phase.

Yeast genetic modification

Over expression strains were made using the following protocol, as described before.[66] Yeast strains were thawed from a frozen stock, plated on a YPD plate and incubated at 30°c over night. Selected colonies were picked and grown in liquid YPD at 30°c, shaking. DNA was edited using the CRISPR-Cas9 system.[67] Next, a PCR amplified repair DNA flanked by 50-bp homology region was transformed together with a bRA89 plasmid, containing a Cas9 and the specific guide-RNA (from James Haber, Addgene plasmid no. 100950). The locus specific guide-RNA was ligated into the bRA89 plasmid as described before.[68] Yeast colonies that were found positive for the bRA89 went through plasmid loss process, by growing in YPD followed by screening for colonies that lost their hygromycin resistance. In over expression strains, the Tdh3 promoter was inserted in place of the Msn2 promoter. In the Msn2 DBD strain, the non – DBD (IDR) region of the coding region was deleted and vice versa for the non – DBD strain.

For the Msn2 IDR mutant strains (LIV to Y, DEKR to N and N to H), synthetic DNA sequences were designed in-silico, were codon optimized for yeast and were ordered from Twist Bioscience. Next, PCR amplified sequences were transformed using CRISPR, as explained in detail in Jonas et al.[23]

Msn2 non – DBD OE with a deletion of Msn4 was made by taking the Msn2 non – DBD OE strain described earlier. CRISPR was used to delete Msn4 from the genome.

METHOD DETAILS

Massive parallel binding assay experimental protocol

Each sample of library transformed stationary phase yeast was diluted to 30ml of SD –URA to reach OD 2-4 over night. The MPBA method requires a comparison of sequence frequencies before and after MNase activation. Therefore, we started our experiment by splitting the yeast, using one half for DNA extraction (non-activated samples) and the other half for ChEC protocol up to proteinase K digestion (activated samples), followed by DNA extraction. Taking the non–activated samples, we pelleted (1 min and 1,500 g) 15ml of OD600 4 yeast cells and resuspended them in 1ml SD –URA before dividing the volume to three 1.5ml low binding tubes. Then, the cells were pelleted again in a centrifuge (1 min and 17,000g) and the media was discarded. The activated samples were used for the ChEC protocol: 15ml of OD 2-4 yeast were pelleted (1 min and 1,500 g), resuspended in 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

moved to a deep 96-well plate. Cells went through two more washes with the same buffer (pelleting the cells again in 1 min and 1,500 g and resuspending thoroughly in 500μl buffer A). The cells were then pelleted and resuspended in 150μl buffer A with 0.1% digitonin. Next, the cells were transferred to a 96-well plate (PCR-96-FLT-C, Axygen) and incubated at 30°c for 5 min for permeabilization. $CaCl_2$ was added to a concentration of 2 mM and the MNase was activated for 180 sec. MNase activity was stopped with the addition of 100μl stop buffer (400 mM NaCl, 20 mM EDTA, 4 mM EGTA and 1% SDS) to 100μl of each sample. Proteinase K was added and the samples were incubated in 30°c for 30 min. For DNA exactraction We used the MasterPure Yeast DNA Purification Kit (Lucigen Corporation, MPY80200) and the kit protocol with some modifications. Samples were resuspended in 300μl of lysis buffer with an addition of RNAse A and incubated at 65°c for 15 minutes. The lysate was transferred to LoBind micro-centrifuge tubes (Eppendorf, 022431021) containing 0.5mm Zirconium Oxide beads (ZrOB05, Next Advance) and were mixed using the Bullet Blender 24 (Next Advance) for 3 min on power 8. The samples were cooled on ice and the bottom of the tube was pierced with a hot metal syringe. Using a centrifuge, tubes were spinned on low speed on top of a second set of tubes and the lysates were collected. Then, 300μl of MPC buffer were added to each sample and mixed using a vortex. All samples were put in a centrifuge (10min, 17,000g), and supernatant was collected and inserted into 500μl iso-propanol. Iso-propanol tubes were centrifuged (10m, 17,000g) to precipitate DNA. Finally, the Iso-propanol was removed and the pellets were washed with 70% ethanol. Ethanol was removed and excess liquid was air dried. 30ul of TE buffer was added to elute DNA.

Library preparation

Samples were cleaned using X1.5 SPRI (AMPure XP, A63881), resuspended in 30μl of elution buffer (10mM Tris-HCl, pH8) and diluted 1:20 in DDW. Diluted DNA was used as a template for a PCR reaction, where DNA was amplified and barcodes were added to each sample. The PCR products were run on an electrophoresis gel to ensure DNA amplification. Samples were cleaned using X0.5 reverse SPRI. Sample concentration was measured using Qubit™ Flex Fluorometer (Invitrogen), and samples were pooled while contributing equal DNA amounts. Pooled samples are SPRI cleaned X0.9. Illumina indices and machine adaptors were added to the samples using a PCR reaction with 1ng of pooled DNA and KAPA HiFi HotStart ReadyMix (Roche, KK2602). Samples were X1 SPRI cleaned and were measured in Qubit and tape station (Agilent) to determine DNA concentration and library quality.

Sequencing: Libraries were sequenced using NovaSeq 6000. The runs were made with the SP100 kit (20040719), parameters: R1-61 cycles Index1-8 cycles Index2-8 cycles R2-61 cycles. To increase complexity, 5% PhiX DNA was added to each run.

Demultiplexing

Our pipeline was built using SnakeMake.[69] Briefly, the forward library primers were removed using cutadapt.[70] Then, AdapterRemoval[71] was used to demultiplex each 32-sample pool into separate samples. Next, each read was assigned to a library variant based on both ends using cutadapt. Reads that did not contain a proper STR on both sides were discarded. Finally, the number of reads assigned to each variant was counted.

Further steps of the Massive parallel binding assay are listed under quantification and statistical analysis below (normalization and confidence threshold, binding score assignment, combining reverse complement and permutated sequences).

Context spread and motif effect

The effect meters were measured in the following way (Figures 3G and 4D). Context spread is the average of the data spread of intact sequences and the data spread of mutated sequences. The data spread was calculated based on the 95th percentile minus the 5th percentile of the data. The motif effect is the median of the differences between all intact and mutated sequences.

Triplet scores

The score of each triplet was the median of the scores of all STRs containing this triplet (Figure 5D). For example, ATTTC and TCCTT (when repeated becomes TCCTTTCCTT) both contain the triplet TTT, among other STRs. The calculated score was the median of the scores of the two STRs above and the rest of the STRs that contain TTT.

AGG-N

We first calculated, for each STR, the average of motif-containing and motif-lacking sequences (Figure 6D). Then, each sequence was assigned to a one of the following sequence groups, according to that sequence content: AGGA, AGGT, AGGC, AGGG, or non-AGG-containing sequences. For example, the sequence AGGAT would be assigned to the AGGA group.

STR abundance in the human genome

The data was taken from Horton et al.[50] and can be found in Table S4 (Figure 7A).

Fraction of STRs in yeast promoters

Yeast STR data was taken from Horton et al.[50] and can be found in Table S5. Promoter annotations can be found in Table S6. We defined promoter STRs as those with >80% match, and at least two repeats within the boundaries of the promoter. Repeats composed of one nucleotide (e.g., AAAAAAAA) were discarded. To define the number of TFs bound to each promoter we used our lab data set (Table S7), which contains, for each TF, the z-score transformed sum of signal received on each promoter. For each promoter, we counted the number of TFs that received a binding score higher than 3 (in units of z-score). Promoters were

divided to groups by number of TFs that bind them. Shown in Figure 7C y - axis is the fraction of promoters from each promoter group with STRs.

TF seq-logos

Data for motif logos was taken from YeTFaSCo (yetfasco.ccbr.utoronto.ca). The data chosen for each motif (Cbf1,[72] Reb1,[73] Mot3[73]) was the one with the highest "Total Score" for a TF, apart from Mot3. In Mot3, the motif chosen had the second highest "Total score" as motifs ranked 2–4 were similar and different from the first-ranked motif logo. Positions with a score lower than 0.5 were discarded.

Protein abundance levels

Protein abundance data was taken from Chong et al.,[74] de Godoy et al.,[75] and Tkach et al.[76] FACS data and RNA expression data were collected in our lab and found in Table S8.

QUANTIFICATION AND STATISTICAL ANALYSIS

Earlier steps of the Massive parallel binding assay are described under method details above.

Normalization and confidence threshold

For each biological sample we produced 3 technical repeats, of which we kept at least 2 repeats. We manually removed repeats showing low correlation to the other two. For each sample, the number of reads was normalized to reach a total read count of 10 million. Then, these values were log2 transformed (Table S1). At this step, we discarded sequences with low read counts as follows: For each sequence, the median read count of all repeats of the non-activated samples had to be greater than 8. The intact and mutated variants of each STR sequence were considered as a pair, and therefore, if one of them was beneath the read count threshold, both were discarded. Sequences in which no reads were detected in at least one sample were also removed from further analysis.

Binding score assignment

The median score of the non-activated repeats was subtracted from the score of each activated repeat. The average of this subtraction was the binding score assigned to each sequence. For comparison between different sequence libraries, we normalized the scores within each library using its median and median absolute deviation (MAD). First, the library median of each sample was subtracted, and then, the result was divided by the MAD.

Combining reverse complement and permutated sequences

We considered equivalent STRs as repeats and took a median of all equivalent sequence scores. For example, the STRs GTG, GGT, and ACC represent the same sequence by being a cyclic repeat of the basic unit GTG and a reverse compliment of it. When combining equivalent sequences, we reduced noise by discarding sequence groups with a standard error greater than 1. Equivalent groups with only one sequence passing the confidence threshold were discarded. We also discarded outlier sequence groups (n=2 for Mot3 and n=1 for free MNase) that showed scores x54-x81 greater (Mot3) and x32 greater (free MNase) than the median score. Results in Table S3. In non - tandem STR libraries, we combined sequences of same STR over three different genomic backbones, using the group median. Sequences that were reverse complementing were considered as the same sequence group for this purpose. Note that when grouping non - tandem STRs, cyclic STRs were not grouped together (GTG, GGT, TGG) as they were not tandem arranged in this construct.

Cell Systems, Volume *16*

Supplemental information

Selective association of short tandem repeats with DNA-binding domains and intrinsically disordered regions of transcription factors

Matan Vidavski, Sagie Brodsky, Wajd Manadre, Tamar Jana Lang, Vladimir Mindel, Yoav Navon, and Naama Barkai

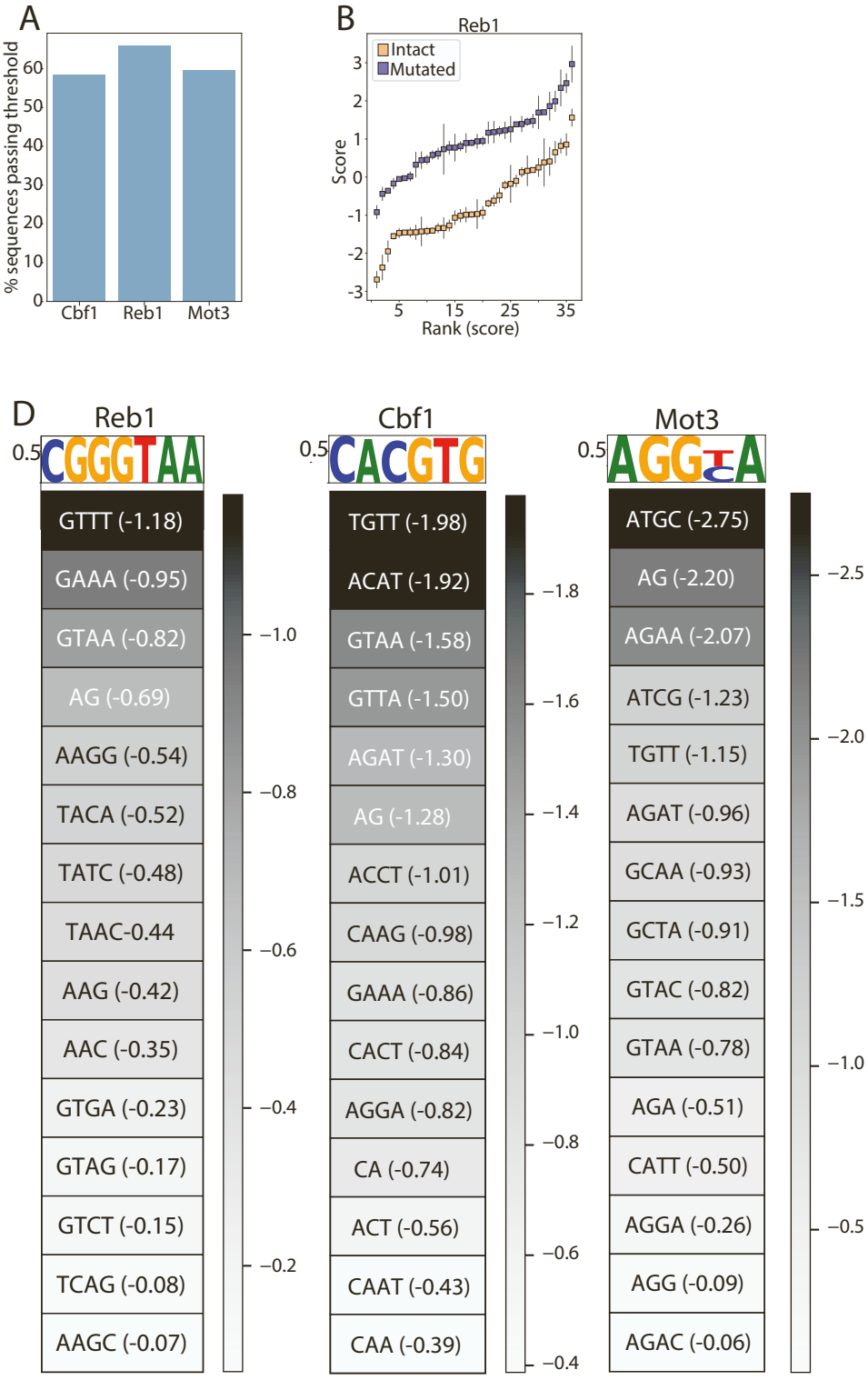
A
% sequences passing threshold
Cbf1
Reb1
Mot3
B
Reb1
Intact
Mutated
Score
Rank (score)
C
Cbf1
Pure GC
Pure AT
All
Density
Reb1
D
Reb1
Cbf1
Mot3
GTTT (-1.18)
GAAA (-0.95)
GTAA (-0.82)
AG (-0.69)
AAGG (-0.54)
TACA (-0.52)
TATC (-0.48)
TAAC-0.44
AAG (-0.42)
AAC (-0.35)
GTGA (-0.23)
GTAG (-0.17)
GTCT (-0.15)
TCAG (-0.08)
AAGC (-0.07)
TGTT (-1.98)
ACAT (-1.92)
GTAA (-1.58)
GTTA (-1.50)
AGAT (-1.30)
AG (-1.28)
ACCT (-1.01)
CAAG (-0.98)
GAAA (-0.86)
CACT (-0.84)
AGGA (-0.82)
CA (-0.74)
ACT (-0.56)
CAAT (-0.43)
CAA (-0.39)
ATGC (-2.75)
AG (-2.20)
AGAA (-2.07)
ATCG (-1.23)
TGTT (-1.15)
AGAT (-0.96)
GCAA (-0.93)
GCTA (-0.91)
GTAC (-0.82)
GTAA (-0.78)
AGA (-0.51)
CATT (-0.50)
AGGA (-0.26)
AGG (-0.09)
AGAC (-0.06)
E
Mot3
Pure AT
All
Density
F
Mot3
Intact
Mutated
Score
Rank (score)
G
Mot3 repeats
non activated
activated
Pearson's r
0.94
H
Mot3
Log2(Fold change)
Mutated
Intact

Figure S1. Library properties and differential STR preferences of the tested TFs.

A. *Detection of the majority of possible STRs:* Shown is the percentage of STR sequences in our library passing the defined confidence threshold (methods). All sequences passing this threshold were included in further analysis.

B. *Reserve complements and cyclic permutations show high reproducibility:* Shown for the Reb1 libraries are the scores assigned for each group of similar sequences. Sequences are colored based on their motif state, and the error bars represent the standard error (STD).

C. *Pure GC and AT sequences are absent from our libraries:* Shown are the distributions of the number of log2 normalized reads for the indicated STR sequences in the libraries of Reb1 and Cbf1.

D. *Differential STR preferences of the tested TFs*: For each of the indicated TFs, shown on the top is its preferred motif and on the bottom are the top 15 bound STRs together with their assigned score. The STRs are sorted from most (top) to least (bottom) preferred.

E. Same as (C) for Mot3.

F. Same as (B) for Mot3

G. *High reproducibility in Mot3 libraries:* The correlation values between the number of log2 normalized reads received in each technical repeat in both time points for libraries transformed into a Mot3-MNase yeast strain are shown.

H. *Mot3 preference is dictated by the identity of STRs rather than the presence of its motif*: The Log2 fold-change from time 0 (prior to MNase activation) of all STRs is shown. Sequences are divided by the state of the three centered motifs. Outliers, defined as the bottom 2.5% and top 2.5%, are not presented.

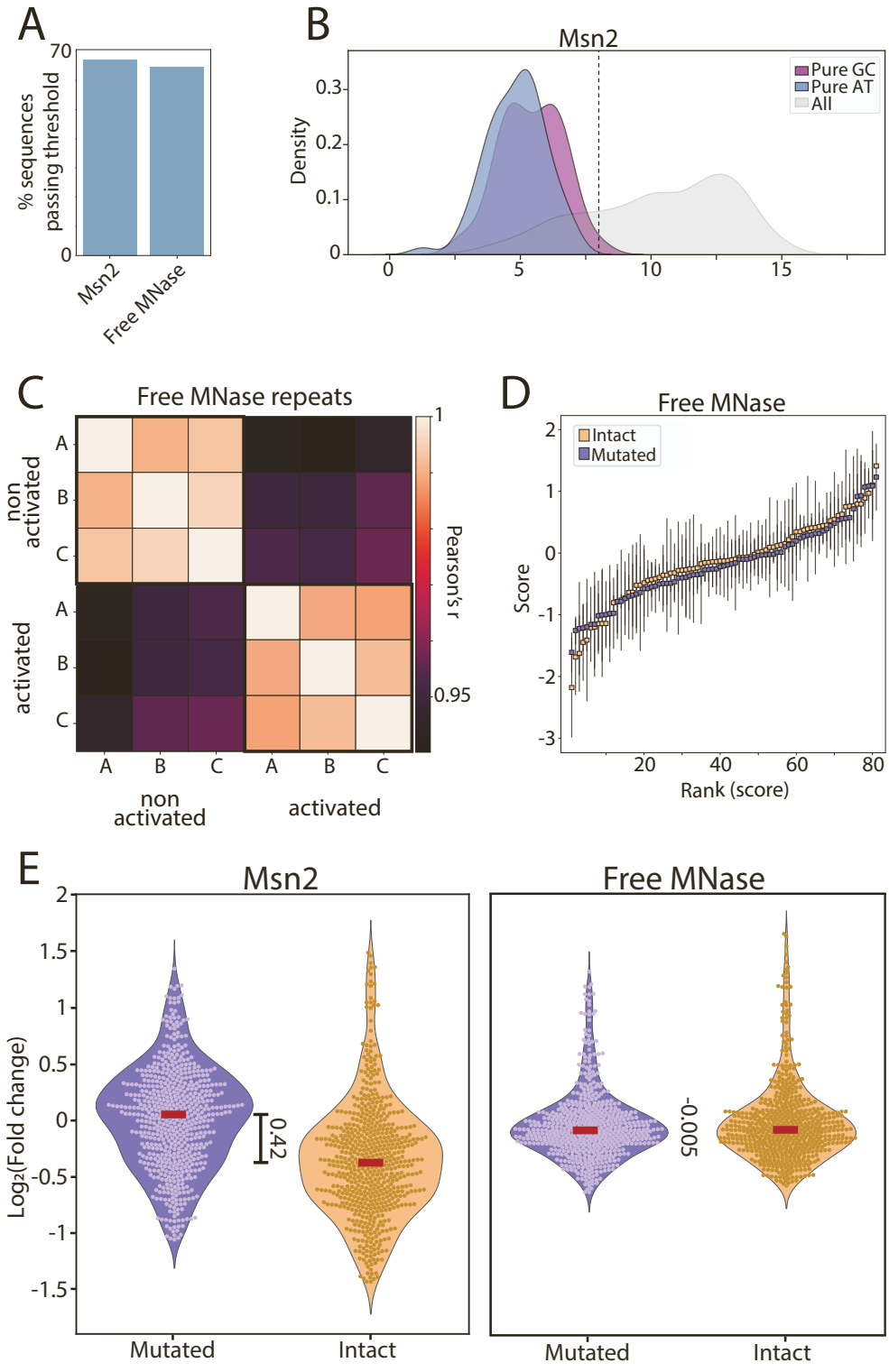
A
70
% sequences passing threshold
0
Msn2
Free MNase
B
Msn2
Pure GC
Pure AT
All
Density
0.3
0.2
0.1
0
0
5
10
15
C
Free MNase repeats
non activated
activated
A
B
C
Pearson's r
1
0.95
non activated
activated
D
Free MNase
Intact
Mutated
Score
2
1
0
-1
-2
-3
20
40
60
80
Rank (score)
E
Msn2
Free MNase
Log2(Fold change)
2
1.5
1
0.5
0
-0.5
-1
-1.5
0.42
-0.005
Mutated
Intact
Mutated
Intact

Figure S2. While showing high reproducibility, a strain bearing a free MNase shows no similarity in STR preference to Msn2.

A. *Number of STRs passing threshold:* Same as in Figure S1A for the indicated strains.

B. *Absence of pure GC and AT sequences in the Msn2 STR library:* Same presentation as Figure S1C.

C. *Free MNase reproducibility:* The correlations between time points and repeats are shown, presentation as in Figure S1G.

D. Same as figure S1B for the free MNase.

E. *Msn2 shows preferences to sequences in which its preferred motifs are intact:* The Log2 fold-change from time 0 (prior to MNase activation) of all STRs are shown for a strain bearing Msn2-MNase (left) and a free MNase (right). Sequences are divided by the state of the three centered Msn2 motifs. Outliers, defined as the bottom 2.5% and top 2.5%, are not presented.

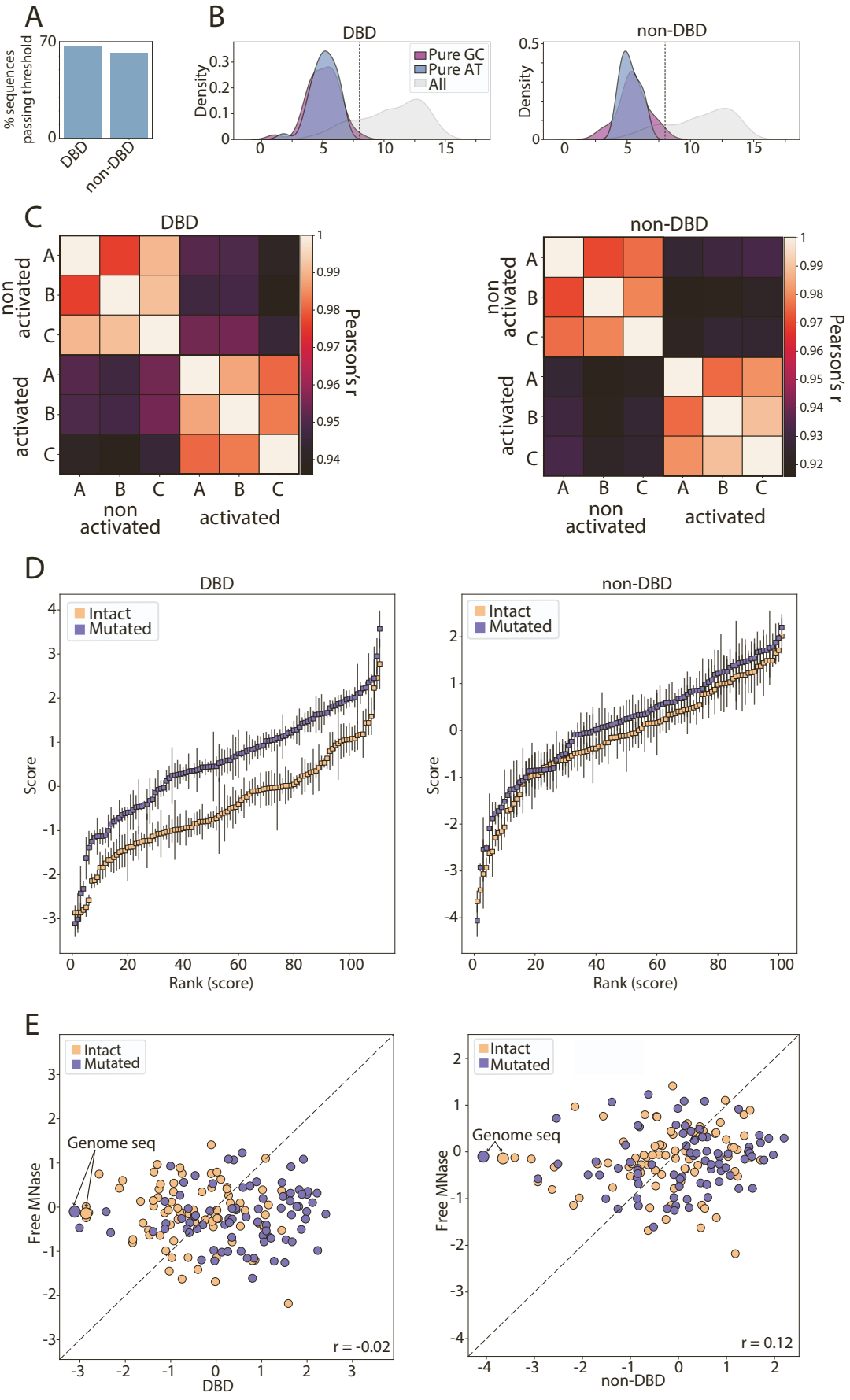
A
% sequences passing threshold
70
0
DBD
non-DBD
B
DBD
non-DBD
Pure GC
Pure AT
All
Density
C
DBD
non-DBD
non activated
activated
A
B
C
Pearson's r
D
DBD
non-DBD
Intact
Mutated
Score
Rank (score)
E
Intact
Mutated
Genome seq
Free MNase
r = -0.02
r = 0.12
DBD
non-DBD

Figure S3. Quality and reproducibility of Msn2 libraries measured in the Msn2 DBD and non-DBD strains.

A. *Number of STRs passing threshold:* Same as in Figure S1A for the indicated strains.

B. *Absence of pure GC and AT sequences:* Same presentation as Figure S1C.

C-D. *The Msn2 DBD and non-DBD show high reproducibility:* Shown in (C) are the correlations between time points and repeats, as in Figure S1G, for the Msn2 DBD (left) and non-DBD (right). Shown in (D) are the scores assigned for each group of similar sequences, including reverse complements and cyclic permutations, for the DBD (left) and non-DBD (right). Sequences are colored based on their motif state, and the error bars represent the standard error (STD).

E. *The Msn2 DBD and non-DBD show no resemblance to the free MNase control:* Shown are the scores received for each STR in the Msn2 library for the DBD (left, x-axis) and the non-DBD (right, x-axis) compared to the free MNase (y-axis). Sequences are colored based on the state of the three central Msn2 motifs.

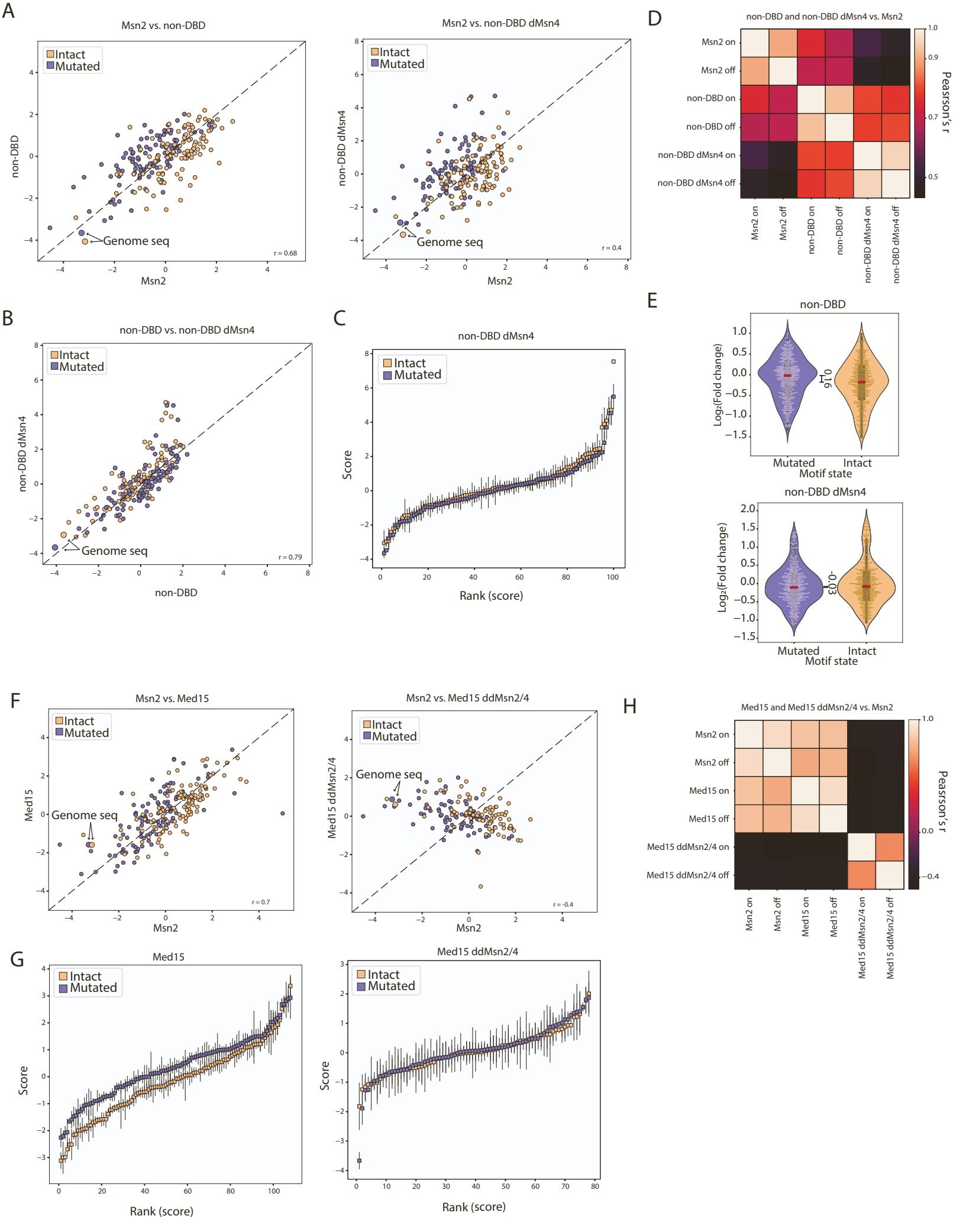
A
Msn2 vs. non-DBD
Intact
Mutated
Genome seq
r = 0.68
non-DBD
Msn2
Msn2 vs. non-DBD dMsn4
Intact
Mutated
Genome seq
r = 0.4
non-DBD dMsn4
Msn2
D
non-DBD and non-DBD dMsn4 vs. Msn2
Msn2 on
Msn2 off
non-DBD on
non-DBD off
non-DBD dMsn4 on
non-DBD dMsn4 off
Pearson's r
B
non-DBD vs. non-DBD dMsn4
Intact
Mutated
Genome seq
r = 0.79
non-DBD dMsn4
non-DBD
C
non-DBD dMsn4
Intact
Mutated
Score
Rank (score)
E
non-DBD
Log2(Fold change)
0.16
Mutated
Intact
Motif state
non-DBD dMsn4
Log2(Fold change)
-0.03
Mutated
Intact
Motif state
F
Msn2 vs. Med15
Intact
Mutated
Genome seq
r = 0.7
Med15
Msn2
Msn2 vs. Med15 ddMsn2/4
Intact
Mutated
Genome seq
r = -0.4
Med15 ddMsn2/4
Msn2
H
Med15 and Med15 ddMsn2/4 vs. Msn2
Msn2 on
Msn2 off
Med15 on
Med15 off
Med15 ddMsn2/4 on
Med15 ddMsn2/4 off
Peasson's r
G
Med15
Intact
Mutated
Score
Rank (score)
Med15 ddMsn2/4
Intact
Mutated
Score
Rank (score)

Figure S4: Msn2 non – DBD retains its preferences and Med15 loses its preferences in the background of Msn2 and Msn4 double deletion.

A. *Similarity in STR binding between Full Msn2 and its non – DBD, even when Msn4 is deleted:* scatter shows STR binding of Msn2 (x-axis) and its non – DBD (y-axis) with native expression of the Msn2 paralog Msn4 (left) and when Msn4 is depleted (right). In both non – DBD strains the full Msn2 is removed from the genome. Sequences containing intact or mutated motifs are colored by orange and purple, respectively.

B. *Deletion of Msn2 and Msn4 do not change the non – DBD STR preferences:* Scatter shows STR binding of the non – DBD in a background including Msn4 (x-axis) and in a background of deletion of Msn4 (y-axis). Sequences containing intact or mutated motifs are colored orange and purple, respectively.

C. *The non – DBD does not favor motif containing over motif lacking sequences:* STR scoring for intact motif (orange) and mutated motif (purple) containing STRs, in the dMsn4 non – DBD. Scores are for STR groups, as in Fig 2B.

D. *The non – DBD dMsn4 correlates more with the non – DBD but not with the full Msn2:* correlation heatmap of the full Msn2, the non – DBD and the non – DBD dMsn4. Heatmap shows both intact ("on") and mutated ("off") motif containing STRs on separate rows, as indicated on heatmap labels. Non – DBD in dMsn4 remains similar to the non – DBD but loses correlation with Msn2.

E. *Motif binding is lost when Msn4 is deleted:* Shown are the distributions of binding scores of motif-containing and motif-lacking sequences for the non – DBD (top) and the non – DBD in dMsn4 (bottom). Included are individual sequences before the grouping of equivalent groups. Outliers, defined by the bottom and top 2.5%, are not presented. Same as in Fig. 4B

F. *Med15 loses similarity to the Msn2 in background of Msn2 and Msn4 double deletion:* scatter shows STR binding of Msn2 (x-axis) and Med15 (y-axis) with native expression of the Msn2 paralog Msn4 (left) and when Msn4 is depleted (right). Sequences containing intact or mutated motifs are colored orange and purple, respectively.

G. *Med15 motif binding is lost when Msn2 and Msn4 are deleted:* STR scoring for intact motif (orange) and mutated motif (purple) containing STRs, in Med15 (left) and in Med15 ddMsn2/4 (right). Scores are for STR groups, as in Fig 2B.

H. *Med15 similarity to Msn2 is lost in double deletion of Msn2 and Msn4:* correlation heatmap of the full Msn2, Med15 and Med15 ddMsn2/4. Heatmap shows both intact ("on") and mutated ("off") motif containing STRs on separate rows, as indicated on heatmap labels. Correlation of Med15 and Msn2 decreases when Msn2 and Msn4 are deleted, while correlation between Med15 and Med15 ddMsn2/4 remains high.

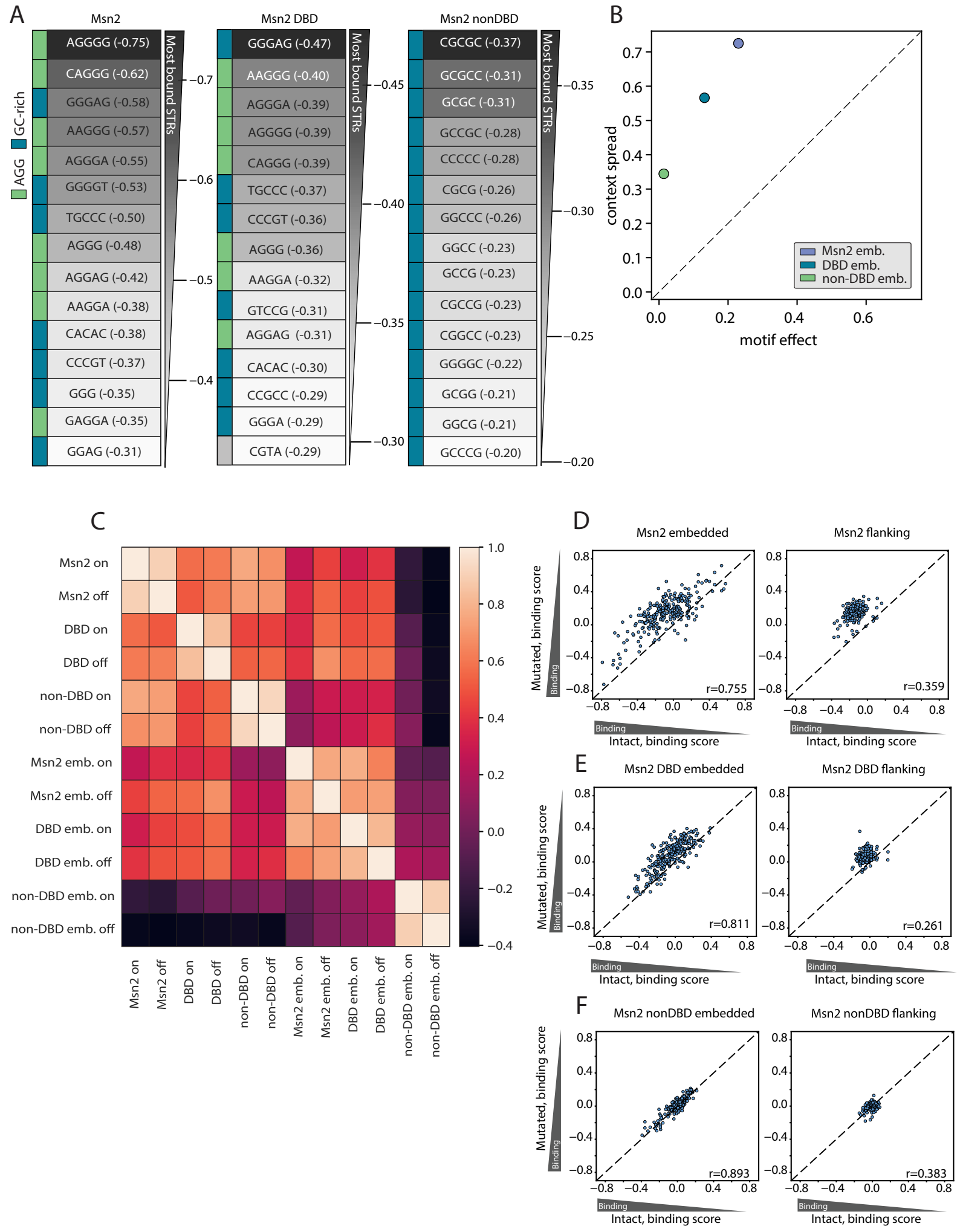
A
Msn2
Msn2 DBD
Msn2 nonDBD
GC-rich
AGG
Most bound STRs
AGGGG (-0.75)
CAGGG (-0.62)
GGGAG (-0.58)
AAGGG (-0.57)
AGGGA (-0.55)
GGGGT (-0.53)
TGCCC (-0.50)
AGGG (-0.48)
AGGAG (-0.42)
AAGGA (-0.38)
CACAC (-0.38)
CCCGT (-0.37)
GGG (-0.35)
GAGGA (-0.35)
GGAG (-0.31)
GGGAG (-0.47)
AAGGG (-0.40)
AGGGA (-0.39)
AGGGG (-0.39)
CAGGG (-0.39)
TGCCC (-0.37)
CCCGT (-0.36)
AGGG (-0.36)
AAGGA (-0.32)
GTCCG (-0.31)
AGGAG (-0.31)
CACAC (-0.30)
CCGCC (-0.29)
GGGA (-0.29)
CGTA (-0.29)
CGCGC (-0.37)
GCGCC (-0.31)
GCGC (-0.31)
GCCGC (-0.28)
CCCCC (-0.28)
CGCG (-0.26)
GGCCC (-0.26)
GGCC (-0.23)
GCCG (-0.23)
CGCCG (-0.23)
CGGCC (-0.23)
GGGGC (-0.22)
GCGG (-0.21)
GGCG (-0.21)
GCCCG (-0.20)
B
context spread
motif effect
Msn2 emb.
DBD emb.
non-DBD emb.
C
Msn2 on
Msn2 off
DBD on
DBD off
non-DBD on
non-DBD off
Msn2 emb. on
Msn2 emb. off
DBD emb. on
DBD emb. off
non-DBD emb. on
non-DBD emb. off
D
Msn2 embedded
Msn2 flanking
Mutated, binding score
Intact, binding score
r=0.755
r=0.359
E
Msn2 DBD embedded
Msn2 DBD flanking
r=0.811
r=0.261
F
Msn2 nonDBD embedded
Msn2 nonDBD flanking
r=0.893
r=0.383

Figure S5: Arrangement of STRs in non - tandem manner effects TF binding

A. *STR preferences shift towards AGG:* top 15 STR bound by Msn2, Msn2 DBD and Msn2 non - DBD, in a library of non-tandem arranged STRs (embedded). Score is the average score of intact and mutated sequences. STR groups are colored as in Fig. 6: AGG containing (green), [TA](G/C)$_3$[TA] containing (pink), AT-rich (purple) and GC-rich (blue).

B. *STR binding meters for embedded STRs:* data summary for the Msn2, the DBD and the non – DBD in context of the non – tandem (embedded) STRs. Same meters as in Fig. 3F-G.

C. *STR preferences alter when STRs are arranged in non - tandem repeats:* correlation heatmap of the full Msn2, the DBD and the non – DBD, with libraries of either tandem repeats of STRs or non – tandem repeats STRs (embedded, "emb.") in a backbone of a native promoter sequence. Heatmap shows both intact ("on") and mutated ("off") motif containing STRs on separate rows, as indicated on heatmap labels.

D-F. *Range of STR effects on binding changes with STR positioning:* A scatter plot comparing binding scores of intact and mutated sequences in Msn2 (top), Msn2 DBD (middle) and Msn2 non – DBD (bottom). Binding scores when STRs are <u>embedded</u> throughout the sequence, with even spacing (left). Binding scores when STRs are clustered upstream and downstream to the central motifs ("flanking", right).

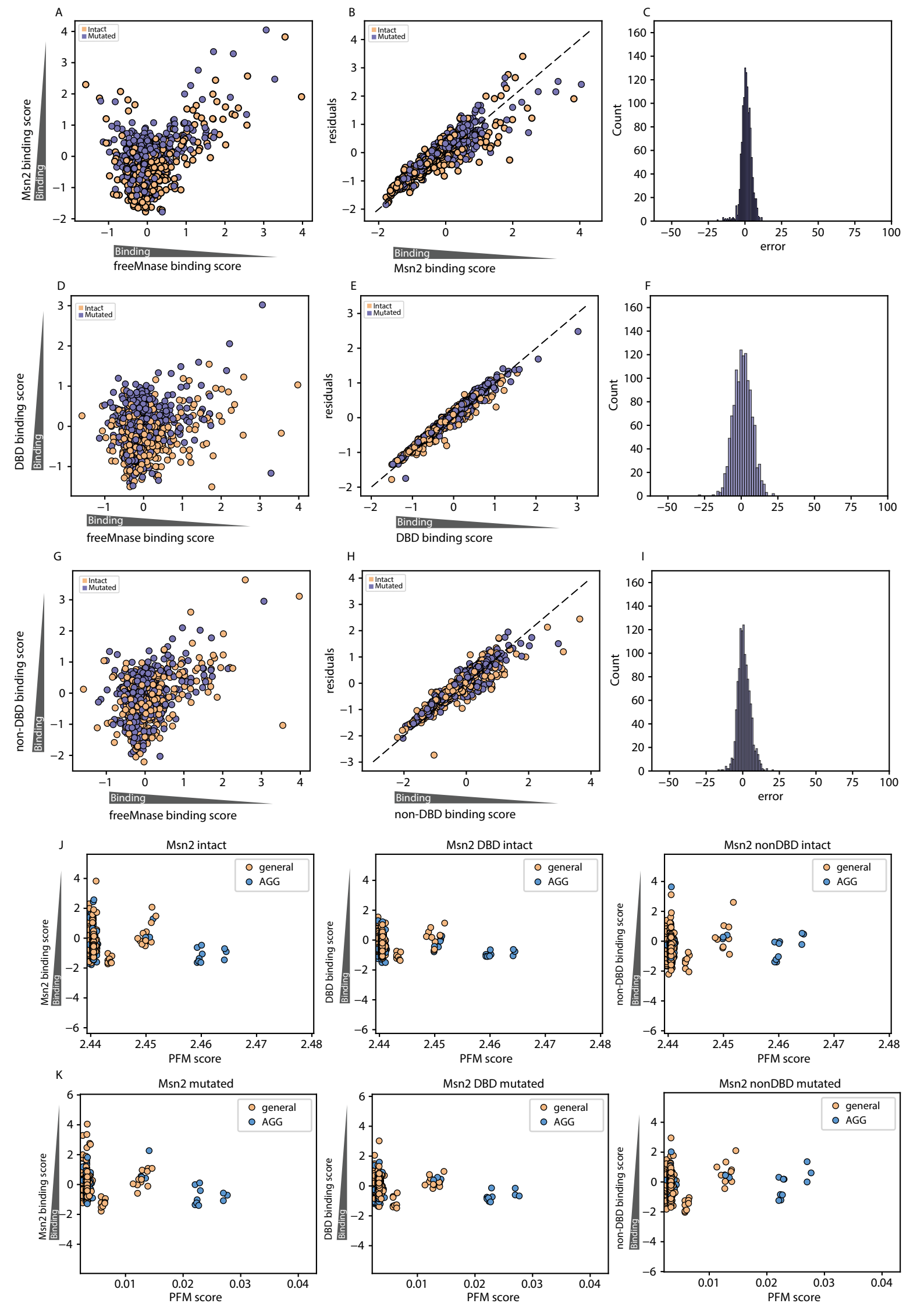
A
Intact
Mutated
Msn2 binding score
Binding
freeMnase binding score
B
residuals
Msn2 binding score
C
Count
error
D
DBD binding score
freeMnase binding score
E
DBD binding score
F
G
non-DBD binding score
freeMnase binding score
H
non-DBD binding score
I
J
Msn2 intact
Msn2 DBD intact
Msn2 nonDBD intact
general
AGG
PFM score
K
Msn2 mutated
Msn2 DBD mutated
Msn2 nonDBD mutated

Figure S6: Residual activity of Mnase and PFM scoring of Msn2 binding prefrences

A. *Variance of STR binding preferences:* scatter of the binding scores of free Mnase (x axis) and of Msn2 (y axis) to the STR library. This data was used for the linear fit described in B.

B. *Spread of Liner Fit Residuals:* Scatter of the residuals (y axis) obtained from a linear fit of Msn2 and free Mnase binding scores, vs. Msn2 binding score (x axis). The data used for the linear fit is presented in A.

C. *Spread of Liner Fit Errors:* The distribution of errors obtained from a linear fit of Msn2 and free Mnase binding scores. The data used for the linear fit is presented in A.

D-F are the same as A-C, for Msn2 DBD rather than Msn2 full length protein.

G-I are the same as A-C, for Msn2 non - DBD rather than Msn2 full length protein.

J. *STR similarity to the Msn2 motif, intact sequences:* scatter shows the binding scores (y axis, for the indicated TF) vs. STR similarity to the Msn2 motif (x axis), scored using a Msn2 motif PFM[1]. The score on the x axis is the sum of products of the STR and the PFM when running the STR over the PFM. Orange was used for any sequence, blue was used for STRs containing an AGG within their sequence, a Msn2 half motif, shown to enable binding (see fig. 5).

K. same as J, here for sequences with mutated central motifs.

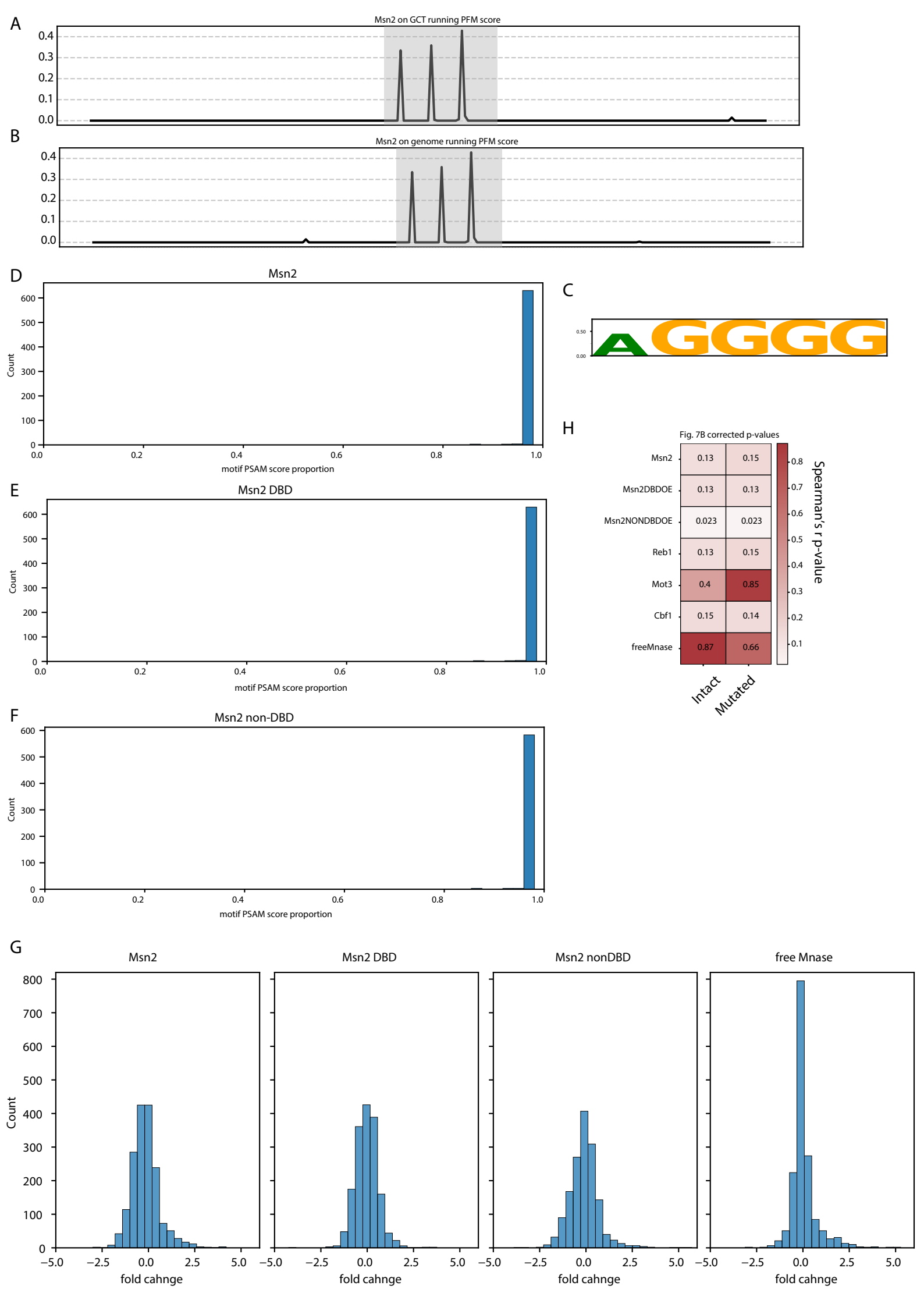
A
Msn2 on GCT running PFM score
B
Msn2 on genome running PFM score
D
Msn2
Count
motif PSAM score proportion
C
E
Msn2 DBD
F
Msn2 non-DBD
H
Fig. 7B corrected p-values
Msn2
Msn2DBDOE
Msn2NONDBDOE
Reb1
Mot3
Cbf1
freeMnase
Intact
Mutated
Spearman's r p-value
G
Msn2
Msn2 DBD
Msn2 nonDBD
free Mnase
fold cahnge

Figure S7: PSAM scoring of motifs and backbone sequences in Msn2 libraries

A. *Predicted motif importance in Msn2 binding:* PSAM score^2 of Msn2 backbone with a random STR (GCT). Running PSAM score was computed for each 7 base window of the backbone used for STR experiments of Msn2. Msn2 three motifs (see Methods for backbone structure) are highlighted. PSAM was obtained from Lee & Bussemaker. [S1]

B. same as B, here for the genome backbone (see Methods) of Msn2.

C. *Msn2 Consensus motif:* Msn2 binding motif logo. Presented are bases with p>0.48. Obtained from YETFASCO, published as Badis et al. [S2]

D. *Motif score accounts for the whole sequence score:* Distribution of PSAM scores proportions in Msn2 libraries. Motif PSAM score as a proportion of the whole sequence PSAM score was computed for each sequence in the STR library (sequences with intact central motifs) of Msn2.

E. same as D, for Msn2 DBD.

F. same as D, for Msn2 non – DBD.
Note that the PSAM score for intact / mutated motifs (AGGGG / AGAGG) is 108.985.

G. *A Wide Range of Binding Scores:* Msn2's distribution of fold – change scores in the STR library. Panels, from left to right: Msn2, Msn2 DBD, Msn2 non – DBD, free Mnase. Data is mid-processed fold change values, before applying a threshold of log2(FC)>8.

H. Correlation p-values, corrected for multiple comparisons, of the correlations presented on Fig. 7B.


RNA expression
Log2(fold chagne)
14
12
10
8
6
4
Msn2
Msn2 DBD
Msn2 non DBD
Tdh3

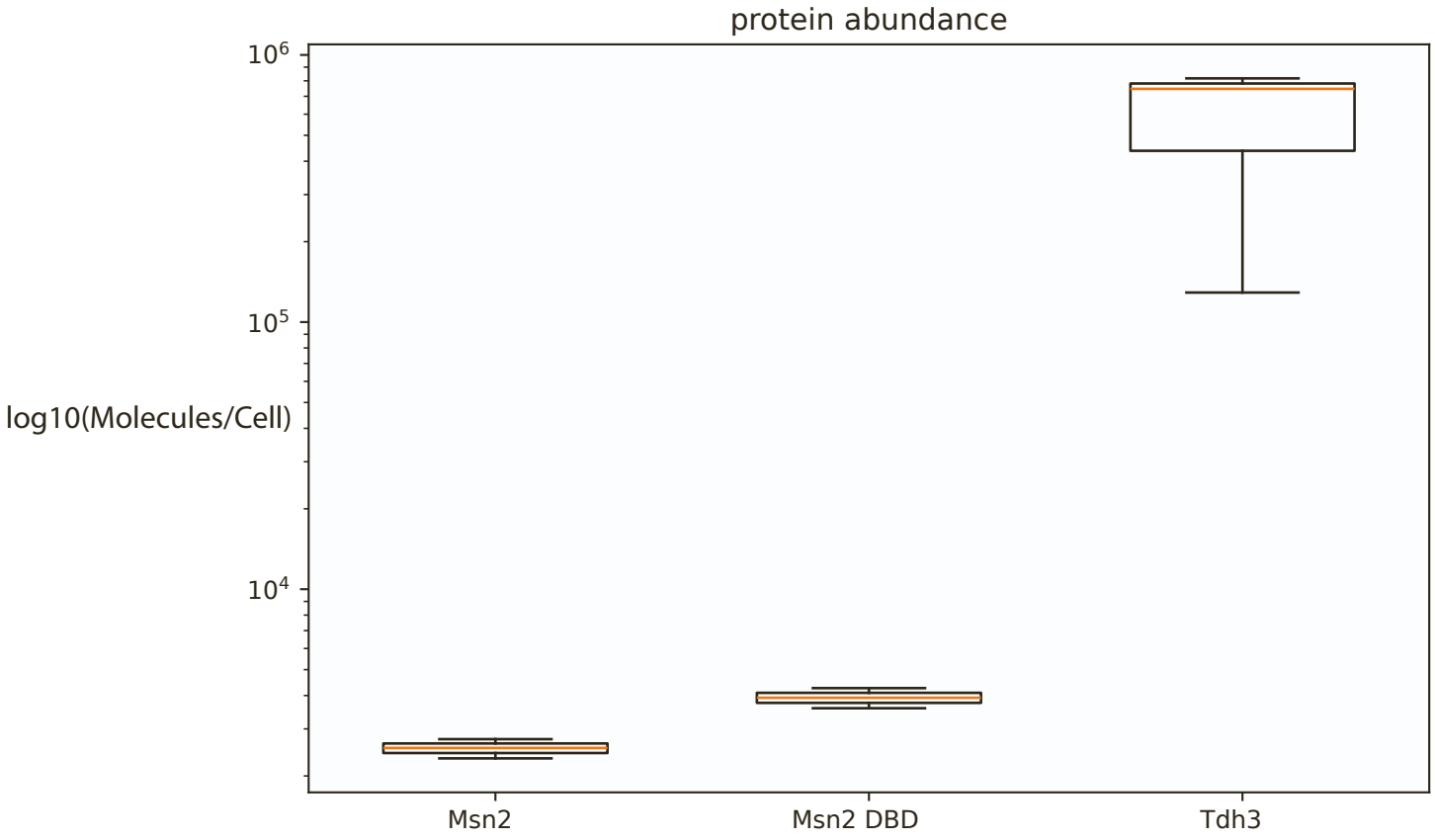
protein abundance
log10(Molecules/Cell)
10^6
10^5
10^4
Msn2
Msn2 DBD
Tdh3

Figure S8: RNA expression and Protein Abundance Data and Estimates

A. *RNA expression:* RNA expression as fold change for Msn2, Msn2 DBD and Msn2 non DBD, expressed under the Msn2 native promoter. RNA expression data is found in supp. table 8.

B. *Protein Abundance:* Protein abundance levels (molecules/cell) based on published data (see Methods) and on FACS data.

Supplemental References

S1. Lee, E., & Bussemaker, H. J. (2010). Identifying the genetic determinants of transcription factor activity. *Molecular Systems Biology*, *6*(1). https://doi.org/10.1038/msb.2010.64

S2. YETFASCO, Badis G, Chan ET, van Bakel H, Pena-Castillo L, Tillo D, Tsui K, Carlson CD, Gossett AJ, Hasinoff MJ, Warren CL, Gebbia M, Talukder S, Yang A, Mnaimneh S, Terterov D, Coburn D, Li Yeo A, Yeo ZX, Clarke ND, Lieb JD, Ansari AZ, Nislow C, Hughes TR. A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. Mol Cell. 2008 Dec 26; 32(6): 878-87.).