Evolution of transcription factor binding through sequence variations and turnover of binding sites

Gat Krieger[1*], Offir Lupo[1*], Patricia Wittkopp[2] and Naama Barkai[1]†

[1]Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel
[2]Department of Ecology and Evolutionary Biology, Department of Molecular, Cellular, and Developmental Biology, University of Michigan, Ann Arbor, MI, USA
[*] Equal contribution
†Corresponding author: naama.barkai@weizmann.ac.il

Abstract

Variations in noncoding regulatory sequences play a central role in evolution. Interpreting such variations, however, remains difficult even in the context of defined attributes such as transcription factor (TF) binding sites. Here, we systematically link variations in *cis*-regulatory sequences to TF binding by profiling the allele-specific binding of 27 TFs expressed in a yeast hybrid, in which two related genomes are present within the same nucleus. TFs localize preferentially to sites containing their known consensus motifs but occupy only a small fraction of the motif-coding sites available within the genomes. Differential binding of TFs to the orthologous alleles was well explained by variations that alter motif sequence, while differences in chromatin accessibility between alleles were of little apparent effect. Motif variations that abolished binding when present in one allele only, were still bound when present in both alleles, suggesting evolutionary compensation, with a potential role for sequence conservation at the motif's vicinity. At the level of the full promoter, we identify cases of binding site turnover, where binding sites are reciprocally gained and lost, yet most interspecific differences remained uncompensated. Our results demonstrate the flexibility of TFs to bind imprecise motifs and the fast evolution of TF binding sites between related species.

Introduction

Changes in gene expression play a key role in cellular adaptation, physiology and development. Guiding these changes are transcription factors (TFs) that bind DNA at sequence motifs allowing activation or repression of gene transcription. Understanding

how TF binding diverges between species is therefore central for understanding how gene regulation evolves.

TFs contain DNA-binding domains (DBDs) that bind with high affinity to short DNA sequence motifs (typically 6-12 base pairs). Sequence variations leading to the emergence or disappearance of binding motifs may therefore drive regulatory divergence by changing TF binding. Previous studies examined for such functional variations by comparing TF binding between related species (Borneman et al., 2007; Bradley et al., 2010; Paris et al., 2013; Schmidt et al., 2010; Stefflova et al., 2013; Wilson et al., 2008), between human individuals (Kasowski et al., 2010; Kilpinen et al., 2013; Maurano et al., 2015) or between alleles of heterozygous cells (Reddy et al., 2012). It was proven difficult, however, to relate the measured changes in TF binding with variations in motif sequence. In their analysis of allele-specific binding of 25 human TFs, Reddy et al. concluded that only 12% of differentially-bound sites were associated with variations in known binding sequences (Reddy et al., 2012). Similarly, studies comparing binding of six TFs between two Drosophila species revealed only modest correlation between inter-specific differences in binding and sequence variations in known motifs (Bradley et al., 2010).

The difficulty of associating inter-species differences in TF binding with variations in *cis*-regulatory sequences mirrors the difficulty in predicting TF binding sites. Indeed, motif preference remains a poor indicator for TF binding *in vivo,* primarily because TFs typically bind at only a small subset of motif-containing sites found in genomes. TF binding could therefore evolve in *cis* not only through the emergence or disappearance of binding motifs, but also through variations in DNA accessibility. Examples for such *cis*-variations include changes in nucleosome positioning (Mirny, 2010; Sun et al., 2015), variations affecting binding of a cooperating TF (Avsec et al., 2021; Stefflova et al., 2013), or variations in promoter regions surrounding the motif, perhaps recognized by TF regions outside the DBD (Brodsky et al., 2020).

In this work, we systematically associated variations in known TF binding motifs to changes in TF binding, by mapping allele-specific binding of 27 TFs within an inter-specific yeast hybrid. The hybrid's nucleus contains two related parental genomes. By applying allele-specific mapping, we could directly compare TF binding to the two

genomes, while ensuring a uniform trans-regulatory environment (Emerson et al., 2010; Floc'hlay et al., 2021; Hill et al., 2021; Krieger et al., 2020; Lupo et al., 2021; Metzger et al., 2016; Tirosh et al., 2009; Wong et al., 2017; Yang et al., 2021). Our analysis examined the contribution of two types of variations, in sequence motifs and in chromatin accessibility, to divergence of TF binding at individual binding sites. We further examined evolutionary changes of TF binding at the full promoter level, by distinguishing cases of compensated binding site turnover (in which loss of a binding site is compensated by gain of an adjacent binding site) from cases of an uncompensated gain/loss. Finally, by capitalizing on the hundreds of sequence variations in motif-containing sites between the genomes, we defined the effective cost of each binding site mutation *in vivo*, linking this cost with sequence conservation at the motif's vicinity. Our results highlight key aspects in the evolution of TF binding between closely related yeast species.

Results

Mapping allele-specific binding of 27 transcription factors within an interspecies hybrid

To examine systematically the effect of *cis* variation on TF binding, we generated F_1 hybrids by mating two closely related budding yeast species: *S. cerevisiae* and *S. paradoxus.* These two species diverged approximately five million years ago, and largely retained gene identity and synteny. Sequence identity reaches ~90% in coding regions and ~75% in promoters (Scannell et al., 2011; Yue et al., 2017). Both species genomes are highly compact with short intergenic regions (200 - 400 bp) that function primarily as gene promoters. Others and us have previously used this hybrid as a model for studying the principles of regulatory evolution (Artieri and Fraser, 2014; Emerson et al., 2010; Krieger et al., 2020; Lupo et al., 2021; Metzger et al., 2016; Tirosh et al., 2009; Weiss et al., 2018).

We selected 27 TFs of five protein families (Table S1). All selected TFs are of known function, and their motif preferences were previously described through *in vitro* and *in vivo* experiments (De Boer and Hughes, 2012; Sandelin et al., 2004). We mapped the localization of these TFs along the orthologous hybrid genomes using chromatin endonuclease cleavage followed by sequencing (ChEC-seq) (Zentner et al., 2015). For

this, each TF was fused to a MNase, allowing us to trigger DNA cleavage at the close vicinity of the TF binding site using short (30 seconds) Ca^{2+} pulse. The short DNA fragments were extracted and sequenced. We found this method to give a highly reproducible and spatially-resolved TF binding maps (Bar-Ziv et al., 2020; Brodsky et al., 2020; Gera et al., 2021; Lupo et al., 2021). We previously observed that orthologous TF proteins bind to similar locations in the hybrid genome and at a similar level, by profiling both the *S. cerevisiae* TF orthologue and the *S. paradoxus* orthologue in separate experiments (Lupo et al., 2021). This observation was consistent with the generally slow evolution of TF preferences (Carroll, 2005). We therefore profiled only the *S. cerevisiae* orthologue, examining how its binding differs between the two alleles (Figure 1 A). Notably, comparing our binding profile with six published datasets revealed high consistency in promoter binding, peak binding and preferred motifs (Supplemental note 1, Figure S16, Figure S17, Table S4, Table S5). Binding signals defined by ChEC-seq were largely restricted to promoter regions, as expected (Figure S1 A).

Next, we compared TF binding between the two orthologues. For this, we distinguished first the overall signal obtained throughout each promoter and second, the locations of individual binding sites within promoters. Focusing first on the level of gene promoters, we find high conservation: in all TFs, promoter binding pattern was correlated between the two alleles (Pearson correlation coefficient, R, ranging from 0.75-0.96). Correlation between experimental replicates was significantly higher (R ranging from 0.9 to 0.99), supporting the reproducibility of our data, and suggesting that some allelic differences do exist (Figure 1B). Of note, correlation between different TFs was much lower (Average R = 0.12) (Figure S1 B). Other measures of correlation were also examined (Supplemental note 3, Figure S 19); however, we found the Pearson linear correlation coefficient to be most appropriate because most TFs bind a small number of targets.

Examining individual promoters, we noted various patterns of conservation and divergence (Figure 1). In the case of Swi5, for example, promoter binding profiles were highly similar between orthologues (R = 0.96), and this similarity extended when examining binding peaks at highly bound promoters (e.g. *SIC1*, Figure 1C, experimental replicates in Figure S1 C-E). In other cases, overall promoter binding was conserved between the two alleles, yet the distribution of binding peaks varied along specific promoters, implying on binding-site turnover (Dermitzakis and Clark, 2009; Ludwig et al.,

2000; Moses et al., 2006). For example, Tbf1 showed similar overall binding to the *CTH1*/*GIR2* promoter in the two orthologous alleles, yet, the precise binding pattern differed, and this variation was linked to a change in the location of the Tbf1 motif within the two orthologous promoters (Figure 1 D, blue box). We also observed cases of divergence, in which overall binding differed between the two orthologous promoters, as exemplified by the binding of Reb1 to the *GSP2* promoter, which was significantly stronger at the *S. paradoxus* allele. Also here, differential binding correlated with the presence of an additional Reb1 motif in *S. paradoxus* allele but not in the *S. cerevisiae* allele (Figure 1 E). We conclude that while TF binding remains largely invariant at the resolution of the full promoter, cases of *cis* divergence at the level of individual TF binding sites are readily identified.

Transcription factors bind a selected subset of motif-containing sites within the two genomes

Promoter regions in budding yeast are typically 200 - 400 base pairs (bp) long (Figure S 1 F, Kristiansson et al., 2009), while individual binding sites contain only 6-12 bp. To examine whether our data can define TF binding at a resolution that is compatible with individual binding sites, we first observed the binding signal around motif-containing sites (Figure 2 A), referring to the known *in vitro* motif of each TF as curated in the YetFasco (De Boer and Hughes, 2012) and JASPAR (Sandelin et al., 2004) databases. For most TFs, these *in vitro* defined motifs agreed well with *de novo* motifs defined from our data by either enrichment of 7-mer sequences around bound sites or the MEME-ChIP algorithm (Machanick and Bailey, 2011) (Table S1, Supplemental note 1).

Considering first the Reb1 TF, we find binding signal at locations containing its *in vitro* motif, as expected. Binding, however, was restricted to only ~30% of motif sites found within promoters (Figure 2 A). In this analysis, we estimated the significance of TF binding relative to a set of random sites within promoters and defined a binding threshold at 95% of random site distribution (Figure 2 B), resulting in 2063 Reb1-bound sites. Binding level at motif sites was moderately correlated (R = 0.25) with the motif p-value as defined by FIMO (Grant et al., 2011). Consistent with the expected pattern of this

method, MNase-cleavage signal peaked at the motif boundaries and was depleted from within the motif itself, the latter is protected from cleavage by the bound TF (Figure 2 A). Results for other TFs were similar, although varied in details depending on TF identity: such as the cleavage symmetry around the motif and the width of the cleavage-protected region (Figure 2 C, full profiles in Figure S2). These details perhaps reflect differences in TF mobility on the DNA (Suter, 2020), motif-specificity and the size of the protein or protein-complex bound to the DNA. An example for the latter factor may be Hap4, where the protected area appears significantly larger (30 bases) than the known motif (7 bases) (Figure 2 C), perhaps indicating its binding as a subunit of the larger Hap2/3/4/5 transcriptional activation complex (McNabb and Pinto, 2005). We conclude that ChEC-seq allows mapping of individual binding locations with high resolution.

To map individual binding sites we used an available peak-calling algorithm (Materials and methods). Peak locations were largely consistent with previously published data (Supplemental note 1, Figure S16). Notably, a considerable fraction (0.2 – 0.6) of reads were mapped to peaks (Figure S4). Overall, 28% of the peaks were associated with the known *in vitro* motif, here referred to as binding sites, in which peaks and motifs are less than 30 bases apart (Figure S5 A). The percentage of peaks associated with the known *in vitro* motifs (all motif realizations with FIMO p-value < 0.001) ranged from 8% to 62% between the different TFs. This fraction was 2.5 times higher than the fraction of random sites that reside next to an *in vitro* motif (averaging over all TFs), resembling high motif-association of peaks in our data. We observe high specificity of TFs to their binding sites, with low overlap between TFs (2% - 4%, Figure S18 D) and no typical binding pattern at binding sites of an unrelated TF (Figure S3).

To estimate the level of systematic noise in our data, namely binding peaks that are not due to TF binding, we tested the ChEC-seq profile of an endogenously expressed free-MNase (Supplemental note 2, Figure S18). Only 2.5% - 4% of TF binding peaks overlapped with free-MNase peaks, representing the false-positive rate of the method and agreeing with previous estimations (Zentner et al., 2015). Therefore, the high number of peaks that were not motif-associated, were also not bound by a free-MNase, and may indicate functional binding events. Such events might result for example, from recruitment by interacting TF, or from protein regions outside the DBD that interact with DNA. As the sequence basis of these binding events is not characterized, we decided to

focus our analysis on binding peaks containing the known *in vitro* motifs, representing 8% to 62% of TF-specific peaks (Figure S5 A).

Differential TF binding to the two hybrid alleles correlates with variations in motif sequence, while differences in motif accessibility play a minor role

TFs could gain new binding sites through at least two mechanisms (Figure 3 A). First, mutations could change the accessibility of the DNA in regions containing a motif site, for example through sequence mutations causing a nucleosome-depleted region. Second, new motifs could emerge by mutations within accessible region. As these two processes occur in parallel, their prevalence may vary depending on the motif type and the processes governing TF specificity, namely its attraction to only a subset of its motif-containing sites.

To distinguish between these two mechanisms of divergence, we focused on TF binding to motif-containing sites. To enable comparison of orthologous binding sites, we locally aligned orthologous promoters, and compared sequence and binding level over the aligned coordinates (materials and methods). We distinguish between sites where both orthologues contain the corresponding motif (conserved between species; common) and sites where the motif is found in only one species' genome (diverged between species; unique). Overall, 36% of sites were classified as common sites, 36% as *cerevisiae*-unique, and 27% as *paradoxus*-unique (Figure 3 B, Figure S6 A). As expected, the fraction of common sites bound by the respective TFs was, on average, twice as high as that of the unique sites (Figure S6 B).

A significant fraction of common sites remained unbound in both genomes, implying that these sites are likely to be inaccessible for TF binding (Figure 3 B, Figure S6 C). Indeed, for most TFs (e.g. Reb1), nucleosome occupancy at unbound sites was higher than that at bound ones (Figure 3 C, Figure S7), while other TFs (e.g. the stress-related TF Sok2) were preferentially bound at sites that are often nucleosome-occupied, as reported before (Kaplan et al., 2009). On average, most TF were preferentially bound at sites of low nucleosome occupancy, prompting us to ask whether changes in nucleosome occupancy between alleles could also explain the divergence of binding. For this, we

asked whether cases in which both orthologous alleles contain a motif (common motif), yet only one of these alleles is in fact bound, might result from differential nucleosome occupancy of the two alleles. We find that in these cases, and also in cases of unique motif sites, nucleosomes are equally positioned in the bound and non-bound orthologues (Figure 3 C, Figure S7). This suggests that the differences in DNA accessibility, at least as reflected by nucleosome occupancy, play a minor role in the divergence of TF binding preferences.

To examine whether differential TF binding correlates with the emergence or loss of the binding motif, we focused on cases of unique sites, in which the motif is present in only one of the alleles. We asked whether, in these cases, the allele containing the motif is more likely to be bound than the one that lacks the motif. This was indeed the case (Figure 3 D): in 25/27 TF in our set, we observed high correspondence between the allele containing the motif and the one bound by the respective TFs. Together, these data suggest that divergence in TF binding due to changes in DNA accessibility are less frequent compared to these caused by the emergence or loss of a binding motif.

TF binding to an imperfect motif depends on the genomic context

Our analysis so far focused on cases where TF binding was lost or gained in one of the genomes. Next, we considered also quantitative changes, where TFs bound the two alleles but at different levels. Such quantitative differences in the allele-specific TF binding were in fact quite common and accounted for the majority of binding changes (Figure S8). We asked whether these quantitative differences could be explained by sequence variations within the binding motif. For this, we focused on binding peaks that contain the associated motif in at least one of the genome, and considered the sequence variations within the motif site and in its immediate surroundings. To further focus the analysis, we considered first cases of unique alternative alleles, in which one orthologue has the consensus motif (as defined *in vitro*), while the second orthologue has a one-letter variant either within the motif itself, or in its flanking region (five bases upstream / downstream the core motif). Comparing TF binding occupancy at the two orthologues allowed quantifying the average cost (reduction in TF binding) of each deviation from consensus (Figure 4 A).

For Reb1, deviations from consensus in the core motif had a strong impact on binding (Figure 4 A). This sensitivity to deviation from the consensus motif differed between TFs and was further dependent on the position and the precise alternative (Figure 4 C). In fact, some TFs remained largely insensitive to single-letter variations (e.g. Skn7, Gcr2, Stb3), while others showed greatly reduced binding (e.g. Rap1, Tbf1, Pho4). In some cases, variations in sequences flanking the known motif were also of apparent consequences: in the case of Reb1, for example, a “T” at position -1 was associated with 100 folds reduction in binding (Figure 4 A). Of note, Reb1 protein was shown to bind the DNA base at the -1 position, and a “T” at that position was predicted to distort DNA shape (Jaiswal et al., 2016; Rossi et al., 2018). This effect outside of the core motif, however, was the exception, as in most cases, variants of apparent effect were restricted to the motif itself, suggesting little contribution from the immediate motif-flanking region.

Our analysis therefore supports the notion that sequence variations within the known *cis* regulatory motif reduce binding in a manner that depends on the TF and the precise sequence alternative. We next asked whether these same deviations from the consensus motif exert a similar cost on binding also when conserved in both species’ genomes (common alternative). Here, we reasoned that deviations from the consensus that appear in both species’ genomes have been preserved by selection, and may therefore reflect the need for lower binding, or, alternatively be compensated by contributions from adjacent sequences. For this analysis, we examined sites in which both alleles contain a motif variant that differs in the same one letter from the consensus motif (common alternative) and asked whether binding to these sites is weaker than binding to the consensus motif, as found in other locations in the genome. For Reb1, the apparent cost of common alternatives (Figure 4 B) was considerably lower than the cost of unique alternatives (Figure 4 A, the average binding fold change between consensus to alternative in common alternative sites is 1.54, where in unique alternative sites it is 3). This same result extended to the majority of other TFs: the same alternative led to higher apparent cost when appearing in only one of the alleles (Figure 4 C) than when appearing in both alleles (Figure 4 D). Of note, the same effect was seen also when comparing unique alternative sites to consensus sites found elsewhere in the genome (Figure S9 B). As a control, we validated that the consensus allele at unique sites is bound at the same level as sites of conserved consensus (Figure S9 C).

Together, our results above support the notion that region-specific effects beyond the motif sequence act to modulate TF binding (Dror et al., 2015; Levo et al., 2015). This could occur through changes in motif accessibility, positioning of the motif within the promoter, or interaction with other DNA-bound co-factors. To test for such compensatory effects, we examined the sequence conservation between *S. cerevisiae* and *S. paradoxus* orthologues at the motif vicinity, postulating that if the surrounding region contributes to motif binding it will remain conserved between the similarly bound alleles (Figure 4 E). Focusing first on Reb1, we find that, as expected, common consensus and common alternative sites were in almost full conservation at the motif region (variation seen is a result of short INDELs). By contrast, sequence conservation decreased in the immediate vicinity of the motif in common consensus sites and in unique alternative sites but stayed relatively high in common alternative sites. The same pattern repeated when examining sequence conservation of seven yeast species (phastCons score, Siepel et al., 2005), where common alternative sites showed higher conservation also at the motif region, and the conservation at their flanking region was higher than that of random-site background (Figure 4 F). Nucleosomes were equally depleted in both types of alternative sites, but depletion was deeper at common consensus sites (Figure S10). This pattern repeated also for Abf1, Rap1 and Tbf1 transcription factors, but was not apparent in other factors, in which conservation did not drop at the immediate motif vicinity (Figure S10). A related observation was reported before for CTCF binding in human, where motif mutations were associated with a strong reduction in binding when appeared in a non-conserved genomic region, but the same mutations showed only a minor effect when appeared in a highly conserved genomic region (Spivakov et al., 2012).

Overall, we find that when a weak motif appears in only one species, it diminishes binding, but when it is species-conserved it allows high level of binding. For specific TFs, the latter appear in highly conserved genomic regions that perhaps compensate for the motif weakness.

Sequence variation in the motif predicts TF binding variation

Our analysis revealed that, for most TFs, interspecific variations in the core motif reduce DNA binding. To examine whether these differences in sequence are sufficient for predicting binding variations, we devised three linear predictive models. The first two models use a motif score as a single predictor: The first is based on the known *in vitro*-derived motif (PWM score) and the second is based on 7-mer sequence enrichment from our data (7-mer motif score). As seen in Figure 5 A, variation in the 7-mer motif score was highly correlated with binding variation of Reb1 between the two alleles. The third model is a multivariate predictor that combines the two aforementioned motif scores and additional features: GC content at 15 bases flanking the motif, nucleosome occupancy at 300 bp centered at the motif, sequence conservation score (phastCons, Siepel et al., 2005) and distance of the peak from the closest transcription start site (Figure 5 B). To distinguish between cases of differential binding that resulted from motif-variation and cases that resulted from other changes, we applied the models on different subsets of the data: all peaks, peaks associated with no motif, peaks associated with a motif and lastly peaks associated with a non-conserved motif (Figure S11 A).

Using the multivariate model, we could explain, on average, 35% of the variance in binding to the two alleles at peaks associated with a non-conserved motif (Figure 5 B). Percentage of variability explained ranged between TFs, with $R^2 = 0.03$ for Hms2 to $R^2 = 0.75$ for Ace2. In the majority of the TFs we examined (18/27), variability in binding to the two orthologues was well explained by sequence variation in the motif and its immediate surroundings ($R^2 > 0.3$), where in three cases sequence-based prediction exceeded R^2 of 0.5. Expanding the prediction to all motif-containing peaks (motif is either conserved or non-conserved) resulted in a somewhat lower predicting power ($0.02 < R^2 < 0.66$, median = 0.32, Figure S11 A). Of note, only 19% of the differentially bound peaks (with more than two-fold change between alleles) were associated with a non-conserved motif, consistent with previous results (Reddy et al., 2012). However, when accounting only for motif-associated peaks, most (60%) of differentially bound peaks were associated with a non-conserved motif (Figure S11 C). Limiting the prediction to peaks with no motif resulted in no predicting power ($0.04 < R^2 < 0.26$, median = 0.11), hence the additional features added to the model beyond the motif, could not explain binding variability in the absence of a motif (Figure S11 A).

In most cases (21/27 TFs), the log-ratio of motif score derived from the data was the best predictor of binding variation. Notable exceptions include Abf1 and Rgt1 where the 7-mer score did not capture the full motif (Table S1), likely because of the large gap between the two parts of Abf1 motif and the long A-stretch of Rgt1 motif. Difference in GC content was predictive only for Skn7 peaks, in correlation with the GC-rich motif of this TF (Figure S11 B). Other features had no predicting power (Figure S11 B). Transforming the features and the predicted change in binding to absolute level (i.e. predicting how big the change is, regardless of its direction) resulted in lower R^2 values and therefore lower prediction power (Figure S11 D).

Based on the studied factors, we conclude that the variation in motif sequence within binding sites is a strong predicator of binding variation for the majority of TFs, reaching 35% explained variation on average.

Gains and losses of binding sites are more common than binding-site turnover

Our analysis so far focused on TF binding at individual binding peaks. We next revisited the integration of binding peaks within the context of the full promoter. Specifically, we wished to define the prevalence of binding-site turnover, whereby, for example, a loss of a binding site in one location along a promoter is compensated by the gain of a binding site at an adjacent location within the same promoter.

To characterize cases of binding site turnover we classified promoters into four classes: 1) Conserved promoters: cases in which all binding sites are common to both alleles (53% of promoters, Figure 6 A), 2) Turnover promoters: cases in which binding sites appear in both orthologues, but on different locations along the promoter, suggesting reciprocal gain or loss of binding motifs (7% of promoters, Figure 6 B). 3) Unbalanced promoters: cases in which one or more binding sites are allele-specific, but other binding sites remain conserved (9% of promoters, Figure 6 C), and 4) Fully unbalanced promoters: cases in which only one of the alleles is bound by the TF (27% of promoters, Figure 6 D). Cases of conserved motifs that are bound at only one allele were considered as “not defined” (5% of promoters). As in previous analyzes, we also

considered only peaks that reside next to a strong binding motif (with FIMO p-value < 0.001).

We built a custom algorithm for promoter classification, which takes a list of peaks as input, classifies motifs, binding sites and promoters into the abovementioned classes (Figure S12 A). To assess the algorithm performance, we manually defined 220 promoters, including up to 20 cases of each class for three TFs, and compared our manual classification with the algorithm output per promoter class. We observe mean sensitivity of 82% and mean specificity of 94% across the different classes (Figure S12 B, C).

We find that the conserved and fully unbalanced classes were the largest promoter classes, consisting of 53% and 27% of all examined promoters, respectively (Figure 6 E, left). Conserved promoters are bound to a higher level (59% of the signal, summing across all TFs) as compared to fully unbalanced promoters (15% of the signal; considering the more highly bound allele (Figure 6 E, right). Among the different TFs there was little variation in the proportion of promoter classes and their binding levels (Figure S13 A). These trends generally repeated also when elevating the minimal peak threshold, although the proportion of conserved promoters increased with increased threshold (Figure S13 B). We note that the different promoter classes are bound to different extents, on average, although our classifier does not take the total promoter binding level into account: the turnover and unbalanced classes were bound at levels twice as high as the conserved promoters (Figure S13 C). This higher binding reflected a larger number of binding sites in these classes, while binding at individual sites was at a similar level (Figure S13 C).

Binding site turnover is a result of reciprocal gains and losses of binding sites. The distance between the turning-over binding sites on the aligned sequence coordinates could be long as in *CDC5* promoter (35 bp) or short, as in *YBL055C* promoter where the two motifs overlap but appear on different strands (Figure 6 B, Figure 6 F). Examining the full set of TFs, we find that most turning-over binding sites appear in close proximity (median distance = 20 bp) and in 37% of these the distance is ten base pairs or less (Figure S14). Specific examples of short-distance binding site turnover are presented in Figure 6 F for Reb1- and Ace2-bound promoters. In the case of Reb1, peaks appear

upstream to the motif in divergent directions corresponding to the appearance of the motif on different strands, and in the case of Ace2 the signal-depleted area aligns with the motif location of each orthologue. This resembles evolutionary conservation of binding site location in the presence of sequence divergence.

Another aspect of binding site turnover would be buffering of differential binding: while binding sites appear on different locations in the two orthologues, the total promoter binding should remain similar. To test that we plotted the correlation between orthologues, summing either on total promoters or on individual peaks (Figure 6 G). Indeed, correlation coefficients were higher on promoters than on peaks. For a control, we summed the binding signal along increasing genomic bins (30, 100, 300 and 1000 bp) and found a strong shift in correlation between the bins of 30 and 100 bp (individual peaks are 20 bp wide), resembling higher divergence in peak binding relative to promoter binding (Figure S15 C, D). Experimental repeats showed high correlation ($R > 0.9$) in all examined bins, in most TFs (23 of 27), and along motif-associated peaks (Figure S15 B). To examine this shift in correlation in more detail, we repeated the same analysis but separately for the four promoter class (Figure 6 H). Namely, for each TF, we considered each time only a single promoter class and examined the shift in orthologous correlation at the promoter level to the correlation on the individual peak level. Indeed, turnover promoters showed the largest shift in promoter-peak correlation (significantly different than the shift in conserved promoters, p-value = 0.003, Tukey's-honestly significant difference test), regardless of these being a relatively small fraction of total promoters (Figure 6 I). The shift observed in unbalanced promoters results from the presence of an additional conserved motif, which increase the correlation at the promoter level.

To conclude, we find that half (53%) of bound promoters retained a fully conserved set of binding sites. Among diverged promoters, a large fraction is bound in only one of the alleles (27%), and a small fraction (7%) shows compensation through binding site turnover, likely indicating functional conservation despite of sequence divergence.

Discussion

Understanding the sequence determinants of transcription factor binding in living cells is a major challenge. In this study, we profiled *in vivo* TF binding in a yeast hybrid that contains two related genomes within the same nucleus. The high sequence divergence in regulatory regions between these two genomes (~25%) provides a wide range of sequence variation that can be examined in parallel. This enabled us to measure how multitudes of sequence variations affect TF binding within a genomic context.

Differential binding can result from differential use of a common pool of potential binding sites containing the binding motif, or from gain/loss of sequence motifs. Our data supports the second scenario, as we found that most conserved motif sites were either bound or unbound in both genomes, while differentially bound sites were associated with a sequence variation within the binding motifs. Further, we detected little, if any, differences in nucleosome positioning at sites that are differentially bound. Therefore in yeast, sequence evolution in regulatory regions appears to occur more readily than changes in chromatin accessibility at TF binding sites, in agreement with previous studies (Tirosh et al., 2010; Tsankov et al., 2010). The ability of a TF to bind its motif-containing site was also shown to depend on DNA shape features (Abe et al., 2015; Zhou et al., 2015), however these features did not separate between *in vivo* bound and unbound motif-containing sites (Zentner et al., 2015) and therefore were not examined in this study.

TFs bind strongly at sites containing their consensus motif, compatible with sequence motifs defined *in vitro.* We expected to observe binding also at sites containing imprecise motifs, for example sites containing one alternative base, but that this binding would be, on average, lower than binding to the consensus motif. This was indeed the case when an alternative base was present in one allele only, while the second allele carried the consensus motif. Notably, however, we find that genomic sites containing the same alternative allele in both orthologous genomes were bound at almost the same extent as sites containing the consensus motif. A perfect consensus sequence might not always be the best for the organism in terms of fitness at each site, however. Non-consensus, or low-affinity binding sites are in fact widespread in the yeast genome (Tanay, 2006), and were shown to be important in fly and mouse development (Crocker et al., 2016; Rowan et al., 2010; Scardigli et al., 2003).

We have further shown that imprecise (weak) motifs of certain TFs, which appear in both orthologous alleles, commonly reside in regions of exceptionally high sequence conservation. This observation is in contrast to the cases of consensus sites, which are often found as islands of conservation within sequence-diverged regions. This may be related to a previous report showing that polymorphisms in CTCF motifs have greater effects on TF binding when they appear in sequence-diverged regions than in sequence-conserved regions (Spivakov et al., 2012). The authors attribute this effect to co-factors that allow CTCF to bind at imprecise motifs. In our case, such binding partners are known for Rap1 (Tornow et al., 1993) but have not been described for the other TFs. We find it more likely that this effect is related to chromatin, as the TFs showing this effect (Abf1, Reb1 and Rap1) all act in the regulation of nucleosome positioning (Rhee and Pugh, 2011). Therefore, we speculate that the presence of other TFs at nearby sites stabilizes the binding of these TFs to non-optimal motif sites (Mirny, 2010). Another possibility involves DNA interactions through the non-DNA-binding domain parts of the protein (Brodsky et al., 2020). To conclude, we suggest that non-consensus sites are bound to a high level and are species-conserved due to a local sequence compensation.

In the majority of the TFs we examined, variability in binding to the two orthologues was well explained by sequence variation in the motif and its immediate surroundings. Previous studies addressing the problem of predicting binding variation from sequence variation, reported on a generally limited predicting power for differential TF binding (Bradley et al., 2010; Halow et al., 2021; He et al., 2011b; Reddy et al., 2012; Stefflova et al., 2013; Zheng et al., 2010). In an influential study for the field, Reddy et al. measured TF binding of in human heterozygous cell lines, and reported that only 12% of differentially-bound sites were associated with sequence variations in known binding motifs (Reddy et al., 2012). Here, we report a similar fraction (19%), however, when considering only motif-associated peaks, we find that most (60%) of the differentially bound sites were associated with a sequence variation in the motif. Further, using quantitative models, we show that variability in motif score is the best predictor for variability in TF binding, whereby other features had limited contribution. Our improved prediction may result from the use of the ChEC-seq method which provides high-resolution mapping of TF binding. In addition, the use of F1 hybrids for this work allows profiling both orthologous genomes in the same cell and thus reduces both technical and *trans*-driven variations that can reduce power and hamper interspecific comparisons.

When classifying evolutionary changes in TF binding at the promoter level, we find that most of the bound promoters involve unbalanced gain or loss of binding sites, whereas only 7% of the bound promoters show evidence of compensation by binding site turnover. This result is in agreement with reports from other model organisms, including the Zeste TF in Drosophila (Moses et al., 2006), liver-specific TFs in mice (Stefflova et al., 2013) and individual TFs in yeast (Borneman et al., 2007). Furthermore, higher rates of positive and purifying selection compared to compensatory neutral evolution were modeled for Drosophila enhancers (He et al., 2011a), but not for yeast promoters (Mustonen et al., 2008). The flexibility and high rate of binding site evolution suggests that in many cases binding site loss or addition is not deleterious. Overall, we find that TF binding evolves through gains and losses of binding sites, with quantitative changes in binding level being highly predictable from sequence variation within the motif.

To conclude, in this study we report on two linked observations: 1) imprecise but conserved motifs are bound to a high level by TFs, and 2) the observation of short-distance binding sites turnover, where binding localization is conserved despite of sequence divergence. These observations demonstrate the fast and flexible evolution of TF binding sites between related species, and we expect to see these phenomena in other organisms as well.

Materials and Methods

Yeast strains

Yeast strains in this study were constructed on the background of *S. cerevisiae* BY4741 and *S. paradoxus* CBS432 (OS142) and their hybrid. For ChEC-seq, transcription factors were tagged with MNase on their C-terminus, by amplifying the MNase-KanMX cassette from the pGZ108 plasmid, a gift from Steven Henikoff. Strains that were previously generated in our lab were based on transformation of BY4741 with MNase-KanMX cassette, with an ORF-MNase linker of 33 amino acids (Bar-Ziv et al., 2020; Brodsky et al., 2020; Lupo et al., 2021). In this study, strains were generated on the background of the C-SWAT library (Meurer et al., 2018) a gift from Maya Schuldiner. In these strains, the MNase-KanMX cassette was inserted between L3 and L4 linkers, with an ORF-MNase linker of 15 amino acids.

free-MNase strain contains MNase from the pGZ108 plasmid, without any linker, under the *TDH3* promoter, integrated into the *MSN2* genomic locus.
Transformations to *S. cerevisiae* were done using the traditional LiAc/SS DNA/PEG method (Gietz et al., 1995). Transformations to *S. paradoxus* were done using SORB-competent cells (Bleuven et al., 2019). Strains are listed in Table S2, primers are listed in Table S3.

ChEC-seq

ChEC-seq experiments were performed as described previously (Zentner et al., 2015) with modifications. In this study, replicates are biological replicates, starting from separate overnight starters of the same strain. Each TF was profiled in at least two replicates. Cultures were grown overnight to saturation in YPD media and diluted into 5 mL of fresh YPD media to reach OD_{600} of 4 the following morning after ~10 divisions. Cultures were pelleted at 1500 g and resuspended in 1 mL Buffer A (15 mM Tris pH 7.5, 80 mM KCl, 0.1 mM EGTA, 0.2 mM spermine, 0.5 mM spermidine, 1× Roche cOmplete EDTA-free mini protease inhibitors, 1 mM PMSF), and then transferred to DNA low-bind tubes (Eppendorf 022431021). Cells were washed twice more in 500 µL Buffer A, pelleted, and resuspended in 150 µL Buffer A containing 0.1% digitonin. Then, cells were transferred to an Eppendorf 96-well plate (Eppendorf 951020401) for permeabilization (30°C for 5 min). $CaCl_2$ was added to a final concentration of 2 mM for 30 seconds. Next, 100 µL of stop buffer (400 mM NaCl, 20 mM EDTA, 4 mM EGTA and 1% SDS) were mixed with 100 µL of sample. Proteinase K was then added (5 µL of 20 mg/ml), and incubated at 55°C for 30 min. Nucleic acids were extracted with an equal volume (200 µL) of ultrapure phenol/chloroform/isoamyl alcohol, and ethanol-precipitated (at −80°C for > 1 hour) with 2.5 volumes of cold EtOH 96%, 45 µg Glycoblue and sodium acetate to a final concentration of 20 mM. DNA was centrifuged (4°C for 10 min), washed with EtOH 70% and treated with RNase A in a final concentration of 2.5 mg/ml (37°C for 20 min), followed by another round of DNA cleanup and ethanol precipitation. To enrich for small DNA fragments, reverse 0.8×SPRI clean-up (right-side size selection) was carried out, following by isopropanol SPRI (left-side size selection) of 1.8×SPRI and 5.4× isopropanol. DNA samples were eluted in 20 µL 0.1×TE.
Library preparation was done similarly to a published protocol (Skene and Henikoff, 2017) with specific modifications. End-repair and A-tailing (ERA) of the small DNA fragments was done the following: [T4 DNA ligase buffer (10×), dNTPs (10 mM), ATP

(10 mM), 50% PEG 4000, T4 PNK (6 U), T4 DNA Pol (0.3 U), Taq DNA Pol (0.1 µL per sample) with 14.6 µL sample] with the PCR protocol: [lid heated to 60°C , 12°C 15', 37°C 15', 58°C 45']. Samples were cleaned with reverse 0.5×SPRI followed by left-side isopropanol SPRI: 1.3×SPRI (with the previous step reaches to 1.8×SPRI) and 5.4× isopropanol. Indexed adaptors (Blecher-Gonen et al., 2013) were ligated to the DNA using quick ligase (2000 U/µL ,2 µL per sample) and quick ligase buffer (2×), in 20°C for 15 minutes. Clean-up was done: 1.2×SPRI (left-side) followed by addition of 1.2×HXN buffer (24 µL 5 M NaCl, 19.2 µL 50% PEG 8000 and 4.8 µL H2O per sample), reaching 1.6× SPRI. Library amplification was carried out with library-specific enrichment primers [23 µL sample DNA, 2 µL enrichment primers, 25 µL KAPA Hifi PCR mix] with the following PCR protocol: [98°C 45'' + 16 cycles (98°C 15'' + 60°C 15'') + 72°C 1']. PCR products were cleaned with left-side 1×SPRI. Library concentration was measured with Q-bit, library size distribution was measured with TapeStation. Libraries were sequenced on Illumina NovaSeq and NextSeq500 machines, with a 51-bases paired end reads.

Computational analysis

Programs

Programs used for read alignment are indicated below. Downstream analyzes were originally implemented in MATLAB 2019 and in R 3.6.3 (R Core Team, 2013). Online programs of the MEME suite were used as well.

Read alignment

FASTQ reads were trimmed from adaptors with cutadapt (Martin, 2011), then aligned to a the hybrid genome using Bowtie 2 (Langmead and Salzberg, 2012) with the parameters: [-p8 --local --very-sensitive --trim-to 40 --dovetail --score-min G,16,8]. The hybrid genome is a concatenation of the genomes of *S. cerevisiae* S288c (R64-1-1/sacCer3) and *S. paradoxus* CBS432 (Yue et al., 2017), including the mitochondrial genomes. Bowtie 2 reports on one (or zero) alignments per read, therefore a given read was mapped only once, to one of the parental genomes. Reads with zero mismatches were 94% - 95% of total mapped reads, in three representative samples. Genome coverage of the 5' end of reads was generated using SAMtools (Li et al., 2009) and BEDtools (Quinlan and Hall, 2010), with the genomcov parameters: [-5 -fs 1].

ChEC-seq data normalization and processing

Raw genome coverage counts were divided by the total number of reads and multiplied by 10^7. Gene promoters were defined using two published datasets of *S. cerevisiae* transcription start sites (Park et al., 2014; Pelechano et al., 2013), where the version with the shorter 5' UTR, in which the TSS is upstream of the start codon, was selected per gene. *S. paradoxus* TSS were defined for orthologous genes using the 5' UTR lengths defined for *S. cerevisiae*. For both genomes, location of specific TSS was manually-edited based on functional genomic data. Promoters were defined as intergenic regions 400 bps upstream to the TSS or to the position where a promoter meets another transcript. Promoters were defined for 5105 out of 6701 genes.

Motif enrichment

Motif enrichment was performed using two methods:

Motif score: As in Brodsky et al., 2020, all possible sequences of length k (*k*-mers) were given a numerical index (16384 possibilities of 7-mers), where each nucleotide in the hybrid genome was indexed accordingly. To compute motif score of a given sample, ChEC-seq signal was smoothed (moving average of 20 nucleotides) and the averaged signal for each *k*-mer was then calculated across all of its occurrences in all promoters. Motif scores of TFs were based on 7-mer sequences.

MEME-ChIP: Sequences of 60 bp centered at top peaks (98% bootstrap level) were extracted per TF, and were used as an input for MEME-ChIP (Machanick and Bailey, 2011), with Yestract (Teixeira et al., 2006) and JASPAR (Sandelin et al., 2004) as reference databases.

Probability weight matrices (PWMs)

In vitro PWMs were collected from the public databases YetFasco (De Boer and Hughes, 2012) and JASPAR (Sandelin et al., 2004) and are listed in Table S1. To allocate significant realizations of these motifs in the hybrid genome we used FIMO (Grant et al., 2011), with the *in vitro* PWM and aligned hybrid genome as input, with significance threshold of p-value < 0.001.

Data-driven PWMs of the different TFs were generated based on the top 20 7-mer sequences of each factor, as in Brodsky et al., 2020.

Aligned genome coordinates

To directly compare ChEC-seq signal and sequence variation between the hybrid alleles we aligned orthologous gene promoters and ordered the genomic data accordingly, as done previously (Venkataram and Fay, 2010). Specifically, we extracted 5105 orthologous and locally aligned their upstream intergenic region with MATLAB function [swalign] with a gap-opening penalty of 10, gap-extension penalty of 0.5 and 'NT' alphabet. This resulted in a reduced, comparable genome of 2,544,708 million base-pairs.

Peak calling

Peaks were called from smoothed ChEC-seq profiles (5' end of reads, 20 bases moving average) using Matalb [findpeaks] function with the following parameters: 'MinPeakHeight' was defined from the data, 'MinPeakProminence' was equal to 'MinPeakHeight', 'MinPeakDistance' was 20 bases, 'MinPeakWidth' was ten bases. As the basal signal level was higher in promoters with high peaks, only peaks that exceed the 90th percentile of their promoter signal were selected. 'MinPeakHeight' definition was the 95 percentile of signal at random sites on promoters. Peak tables are provided as supplementary Table S6.

Peak-motif association

The highest motif score was located at a range of 60 bp centered at the peak. Orthologous peaks that were separated by less than ten bases were unified into a single peak location. Peaks further than 800 from any TSS were filtered out from further analysis.

Position-specific mutation cost

In order to measure the binding cost due to mutations at specific positions of the motif, the following analysis was carried out: peaks were aligned relative to the location of their maximal motif score. Then, for each peak, the motif sequence of the better-scored allele was aligned to the motif PWM, for sequence comparison. Alignment to the PWM was based on the product of probabilities (P) of all positions. To allow flexibility, the minimal PWM-score allowed sequence variability at positions with maximal probability (P) < 0.7. This way Reb1 sites of TTACCCG and TTACCCT were both allowed. Sequence substitution was analyzed relative to the motif consensus sequence, which is the maximal PWM-scoring sequence. To find sequences with an alternative allele, an iterated algorithm was implemented: in each iteration a certain position of the motif is

"mutated" so the nucleotide probabilities in that position equals 0.25. Figure 5 summarizes this analysis, where the average $\log_2$ ratio of alternative to consensus is shown.

Prediction of TF binding variation

Multiple linear models were analyzed in R 3.6.3 (R Core Team, 2013). Relative feature importance was analyzed using RelaImpo package (Grömping, 2006).

Data access

All raw and processed sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE196451.

The FASTQ data generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA700498.

Code is shared in: https://github.com/GatKrieger/TFhybrid.

Acknowledgements

We thank the Barkai lab members, especially to Felix Jonas, Michal Chapal and Sagie Brodsky for their help with experiments, useful discussions and constructive comments on the manuscript.

Funding

This project was supported by the U.S. National Science Foundation– U.S. Israel Binational Science Foundation–Molecular and Cellular Biosciences (NSF-BSF-MCB) (2019625), the Israel Science Foundation (ISF) (1738/15), and the Minerva center (AZ 57 46 9407 65).

Competing interest statement

The authors declare no competing interests.

Author contributions

G.K. and N.B. designed the research. G.K. and O.L. performed experiments. G.K analyzed the data. All authors contributed to the writing of the paper.

735

736 **References**

737 Abe, N., Dror, I., Yang, L., Slattery, M., Zhou, T., Bussemaker, H.J., Rohs, R., and
738 Mann, R.S. (2015). Deconvolving the recognition of DNA shape from sequence. Cell
739 *161*, 307–318.
740 Artieri, C.G., and Fraser, H.B. (2014). Evolution at two levels of gene expression in
741 yeast. 411–421.
742 Avsec, Ž., Weilert, M., Shrikumar, A., Krueger, S., Alexandari, A., Dalal, K., Fropf, R.,
743 McAnany, C., Gagneur, J., Kundaje, A., et al. (2021). Base-resolution models of
744 transcription-factor binding reveal soft motif syntax. Nat. Genet.
745 Bar-Ziv, R., Brodsky, S., Chapal, M., and Barkai, N. (2020). Transcription Factor Binding
746 to Replicated DNA. Cell Rep. *30*, 3989-3995.e4.
747 Blecher-Gonen, R., Barnett-Itzhaki, Z., Jaitin, D., Amann-Zalcenstein, D., Lara-Astiaso,
748 D., and Amit, I. (2013). High-throughput chromatin immunoprecipitation for genome-wide
749 mapping of in vivo protein-DNA interactions and epigenomic states. Nat. Protoc. *8*, 539–
750 554.
751 Bleuven, C., Dubé, A.K., Nguyen, G.Q., Gagnon-Arsenault, I., Martin, H., and Landry,
752 C.R. (2019). A collection of barcoded natural isolates of Saccharomyces paradoxus to
753 study microbial evolutionary ecology. Microbiologyopen *8*, 1–13.
754 De Boer, C.G., and Hughes, T.R. (2012). YeTFaSCo: A database of evaluated yeast
755 transcription factor sequence specificities. Nucleic Acids Res. *40*.
756 Borneman, A.R., Gianoulis, T.A., Zhang, Z.D., Yu, H., Rozowsky, J., Seringhaus, M.R.,
757 Wang, L.Y., Gerstein, M., and Snyder, M. (2007). Divergence of Transcription Factor
758 Binding Sites Across Related Yeast Species. Science (80-.). *317*, 815–819.
759 Bradley, R.K., Li, X.-Y., Trapnell, C., Davidson, S., Pachter, L., Chu, H.C., Tonkin, L.A.,
760 Biggin, M.D., and Eisen, M.B. (2010). Binding Site Turnover Produces Pervasive
761 Quantitative Changes in Transcription Factor Binding between Closely Related
762 Drosophila Species. PLOS Biol. *8*, e1000343.
763 Brodsky, S., Jana, T., Mittelman, K., Chapal, M., Kumar, D.K., Carmi, M., and Barkai, N.
764 (2020). Intrinsically Disordered Regions Direct Transcription Factor In Vivo Binding
765 Specificity. Mol. Cell *79*, 459-471.e4.
766 Carroll, S.B. (2005). Evolution at Two Levels: On Genes and Form. PLOS Biol. *3*, e245.
767 Crocker, J., Preger-Ben Noon, E., and Stern, D.L. (2016). The Soft Touch: Low-Affinity

768 Transcription Factor Binding Sites in Development and Evolution (Elsevier Inc.).
769 Dermitzakis, E.T., and Clark, A.G. (2009). Evolution of transcription factor binding sites
770 in mammalian gene regulatory regions: Handling counterintuitive results. J. Mol. Evol.
771 *68*, 654–664.
772 Dror, I., Golan, T., Levy, C., Rohs, R., and Mandel-Gutfreund, Y. (2015). A widespread
773 role of the motif environment in transcription factor binding across diverse protein
774 families. Genome Res. *25*, 1268–1280.
775 Emerson, J.J., Hsieh, L.C., Sung, H.M., Wang, T.Y., Huang, C.J., Lu, H.H.S., Lu, M.Y.J.,
776 Wu, S.H., and Li, W.H. (2010). Natural selection on cis and trans regulation in yeasts.
777 Genome Res. *20*, 826–836.
778 Floc'hlay, S., Wong, E.S., Zhao, B., Viales, R.R., Thomas-Chollier, M., Thieffry, D.,
779 Garfield, D.A., and Furlong, E.E.M. (2021). Cis-acting variation is common across
780 regulatory layers but is often buffered during embryonic development. Genome Res. *31*,
781 211–224.
782 Fordyce, P.M., Gerber, D., Tran, D., Zheng, J., Li, H., DeRisi, J.L., and Quake, S.R.
783 (2010). De novo identification and biophysical characterization of transcription-factor
784 binding sites with microfluidic affinity analysis. Nat. Biotechnol. *28*, 970–975.
785 Gera, T., Jonas, F., More, R., and Barkai, N. (2021). Evolution of binding preferences
786 among whole-genome duplicated transcription factors. BioRxiv 2021.07.27.453962.
787 Gietz, R.D., Schiestl, R.H., Willems, A.R., and Woods, R.A. (1995). Studies on the
788 transformation of intact yeast cells by the LiAc/SS□DNA/PEG procedure. Yeast *11*, 355–
789 360.
790 Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a
791 given motif. Bioinformatics *27*, 1017–1018.
792 Grömping, U. (2006). Relative importance for linear regression in R: the package
793 relaimpo. J. Stat. Softw. *17*, 1–27.
794 Halow, J.M., Byron, R., Hogan, M.S., Ordoñez, R., Groudine, M., Bender, M.A.,
795 Stamatoyannopoulos, J.A., and Maurano, M.T. (2021). Tissue context determines the
796 penetrance of regulatory DNA variation. Nat. Commun. *12*.
797 He, B.Z., Holloway, A.K., Maerkl, S.J., and Kreitman, M. (2011a). Does positive selection
798 drive transcription factor binding site turnover? a test with drosophila cis-regulatory
799 modules. PLoS Genet. *7*.
800 He, Q., Bardet, A.F., Patton, B., Purvis, J., Johnston, J., Paulson, A., Gogol, M., Stark,
801 A., and Zeitlinger, J. (2011b). High conservation of transcription factor binding and

802 evidence for combinatorial regulation across six Drosophila species. Nat. Genet. *43*,
803 414–421.
804 Hill, M.S., Vande Zande, P., and Wittkopp, P.J. (2021). Molecular and evolutionary
805 processes generating variation in gene expression. Nat. Rev. Genet. *22*, 203–215.
806 Jaiswal, R., Choudhury, M., Zaman, S., Singh, S., Santosh, V., Bastia, D., and
807 Escalante, C.R. (2016). Functional architecture of the Reb1-Ter complex of
808 Schizosaccharomyces pombe. Proc. Natl. Acad. Sci. U. S. A. *113*, E2267–E2276.
809 Kaplan, N., Moore, I.K., Fondufe-Mittendorf, Y., Gossett, A.J., Tillo, D., Field, Y.,
810 LeProust, E.M., Hughes, T.R., Lieb, J.D., Widom, J., et al. (2009). The DNA-encoded
811 nucleosome organization of a eukaryotic genome. Nature *458*, 362–366.
812 Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S.M.,
813 Habegger, L., Rozowsky, J., Shi, M., Urban, A.E., et al. (2010). Variation in transcription
814 factor binding among humans. Science (80-.). *328*, 232–235.
815 Kilpinen, H., Waszak, S.M., Gschwind, A.R., Raghav, S.K., Witwicki, R.M., Orioli, A.,
816 Migliavacca, E., Wiederkehr, M., Gutierrez-arcelus, M., Panousis, N.I., et al. (2013).
817 Coordinated Effects of Sequence Variation on DNA Binding, Chromatin Structure, and
818 Transcription. Science (80-.). *342*, 744–747.
819 Krieger, G., Lupo, O., Levy, A.A., and Barkai, N. (2020). Independent evolution of
820 transcript abundance and gene regulatory dynamics. Genome Res. *30*, 1000–1011.
821 Kristiansson, E., Thorsen, M., Tamás, M.J., and Nerman, O. (2009). Evolutionary forces
822 act on promoter length: identification of enriched cis-regulatory elements. Mol. Biol.
823 Evol. *26*, 1299–1307.
824 Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2.
825 Nat. Methods *9*, 357–359.
826 Levo, M., Zalckvar, E., Sharon, E., Machado, A.C.D., Kalma, Y., Lotan-Pompan, M.,
827 Weinberger, A., Yakhini, Z., Rohs, R., and Segal, E. (2015). Unraveling determinants of
828 transcription factor binding outside the core binding site. Genome Res. *25*, 1410.
829 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G.,
830 Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and
831 SAMtools. Bioinformatics *25*, 2078–2079.
832 Ludwig, M.Z., Bergman, C., Patel, N.H., and KreLtman, M. (2000). Evidence for
833 stabilizing selection in a eukaryotic enhancer element. Nature *403*, 564–567.
834 Lupo, O., Krieger, G., Jonas, F., and Barkai, N. (2021). Accumulation of cis- and trans-
835 regulatory variations is associated with phenotypic divergence of a complex trait

836 between yeast species. G3 Genes|Genomes|Genetics *11*, jkab016.
837 Machanick, P., and Bailey, T.L. (2011). MEME-ChIP: motif analysis of large DNA
838 datasets. Bioinformatics *27*, 1696–1697.
839 Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput
840 sequencing reads. EMBnet. J. *17*, 10–12.
841 Maurano, M.T., Haugen, E., Sandstrom, R., Vierstra, J., Shafer, A., Kaul, R., and
842 Stamatoyannopoulos, J.A. (2015). Large-scale identification of sequence variants
843 influencing human transcription factor occupancy in vivo. Nat. Genet. *47*, 1393–1401.
844 McNabb, D.S., and Pinto, I. (2005). Assembly of the Hap2p/Hap3p/Hap4p/Hap5p-DNA
845 Complex in Saccharomyces cerevisiae. Eukaryot. Cell *4*, 1829–1839.
846 Metzger, B.P.H., Duveau, F., Yuan, D.C., Tryban, S., Yang, B., and Wittkopp, P.J.
847 (2016). Contrasting Frequencies and Effects of cis- and trans-Regulatory Mutations
848 Affecting Gene Expression. Mol. Biol. Evol. *33*, 1131–1146.
849 Meurer, M., Duan, Y., Sass, E., Kats, I., Herbst, K., Buchmuller, B.C., Dederer, V.,
850 Huber, F., Kirrmaier, D., Štefl, M., et al. (2018). Genome-wide C-SWAT library for high-
851 throughput yeast genome tagging. Nat. Methods *15*, 598–600.
852 Mirny, L.A. (2010). Nucleosome-mediated cooperativity between transcription factors.
853 Proc. Natl. Acad. Sci. U. S. A. *107*, 22534–22539.
854 Moses, A.M., Pollard, D.A., Nix, D.A., Iyer, V.N., Li, X.-Y., Biggin, M.D., and Eisen, M.B.
855 (2006). Large-Scale Turnover of Functional Transcription Factor Binding Sites in
856 Drosophila. PLOS Comput. Biol. *2*, e130.
857 Mustonen, V., Kinney, J., Callan, C.G., and Lässig, M. (2008). Energy-dependent
858 fitness: A quantitative model for the evolution of yeast transcription factor binding sites.
859 Proc. Natl. Acad. Sci. U. S. A. *105*, 12376–12381.
860 Paris, M., Kaplan, T., Li, X.Y., Villalta, J.E., Lott, S.E., and Eisen, M.B. (2013). Extensive
861 Divergence of Transcription Factor Binding in Drosophila Embryos with Highly
862 Conserved Gene Expression. PLoS Genet. *9*.
863 Park, D., Morris, A.R., Battenhouse, A., and Iyer, V.R. (2014). Simultaneous mapping of
864 transcript ends at single-nucleotide resolution and identification of widespread promoter-
865 associated non-coding RNA governed by TATA elements. Nucleic Acids Res. *42*, 3736–
866 3749.
867 Pelechano, V., Wei, W., and Steinmetz, L.M. (2013). Extensive transcriptional
868 heterogeneity revealed by isoform profiling. Nature *497*, 127–131.
869 Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing

genomic features. Bioinformatics *26*, 841–842.

Reddy, T.E., Gertz, J., Pauli, F., Kucera, K.S., Varley, K.E., Newberry, K.M., Marinov, G.K., Mortazavi, A., Williams, B.A., Song, L., et al. (2012). Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. Genome Res. *22*, 860–869.

Rhee, H.S., and Pugh, B.F. (2011). Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell *147*, 1408–1419.

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.

Rowan, S., Siggers, T., Lachke, S.A., Yue, Y., Bulyk, M.L., and Maas, R.L. (2010). Precise temporal control of the eye regulatory gene Pax6 via enhancer-binding site affinity. Genes Dev. *24*, 980–985.

Sandelin, A., Alkema, W., Engström, P., Wasserman, W.W., and Lenhard, B. (2004). JASPAR: an open□access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res. *32*, D91–D94.

Scannell, D.R., Zill, O.A., Rokas, A., Payen, C., Dunham, M.J., Eisen, M.B., Rine, J., Johnston, M., and Hittinger, C.T. (2011). The Awesome Power of Yeast Evolutionary Genetics: New Genome Sequences and Strain Resources for the Saccharomyces sensu stricto Genus. G3 (Bethesda). *1*, 11–25.

Scardigli, R., Ba□umer, N., Gruss, P., Guillemot, F., and Le Roux, I. (2003). Direct and concentration-dependent regulation of the proneural gene Neurogenin2 by Pax6.

Schmidt, D., Wilson, M.D., Ballester, B., Schwalie, P.C., Brown, G.D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C.P., Mackay, S., et al. (2010). Five-Vertebrate ChIP-seq Reveals the Evolutionary Dynamics of Transcription Factor Binding. Science (80-.). *328*, 1036–1040.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., and Richards, S. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. *15*, 1034–1050.

Skene, P.J., and Henikoff, S. (2017). An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. Elife *6*, 1–35.

Spivakov, M., Akhtar, J., Kheradpour, P., Beal, K., Girardot, C., Koscielny, G., Herrero, J., Kellis, M., Furlong, E.E.M., and Birney, E. (2012). Analysis of variation at transcription

factor binding sites in Drosophila and humans. Genome Biol. *13*, R49.

Stefflova, K., Thybert, D., Wilson, M.D., Streeter, I., Aleksic, J., Karagianni, P., Brazma, A., Adams, D.J., Talianidis, I., Marioni, J.C., et al. (2013). Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. Cell *154*, 530–540.

Sun, Y., Nien, C.-Y., Chen, K., Liu, H.-Y., Johnston, J., Zeitlinger, J., and Rushlow, C. (2015). Zelda overcomes the high intrinsic nucleosome barrier at enhancers during Drosophila zygotic genome activation. Genome Res. *25*, 1703–1714.

Suter, D.M. (2020). Transcription factors and DNA play hide and seek. Trends Cell Biol. *30*, 491–500.

Tanay, A. (2006). Extensive low-affinity transcriptional interactions in the yeast genome. Genome Res. *16*, 962–972.

Team, R.C. (2013). R: A language and environment for statistical computing.

Teixeira, M.C., Monteiro, P., Jain, P., Tenreiro, S., Fernandes, A.R., Mira, N.P., Alenquer, M., Freitas, A.T., Oliveira, A.L., and Sá-Correia, I. (2006). The YEASTRACT database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae. Nucleic Acids Res. *34*, D446–D451.

Tirosh, I., Reikhav, S., Levy, A.A., and Barkai, N. (2009). A Yeast Hybrid Provides Insight into the Evolution of Gene Expression Regulation. Science (80-.). *324*, 659–662.

Tirosh, I., Sigal, N., and Barkai, N. (2010). Divergence of nucleosome positioning between two closely related yeast species: genetic basis and functional consequences. Mol. Syst. Biol. *6*, 365.

Tornow, J., Zeng, X., Gao, W., and Santangelo, G.M. (1993). GCR1, a transcriptional activator in Saccharomyces cerevisiae, complexes with RAP1 and can function without its DNA binding domain. EMBO J. *12*, 2431–2437.

Tsankov, A.M., Thompson, D.A., Socha, A., Regev, A., and Rando, O.J. (2010). The Role of Nucleosome Positioning in the Evolution of Gene Regulation. PLOS Biol. *8*, e1000414.

Venkataram, S., and Fay, J.C. (2010). Is transcription factor binding site turnover a sufficient explanation for cis-regulatory sequence divergence? Genome Biol. Evol. *2*, 851–858.

Weiss, C. V, Roop, J.I., Hackley, R.K., Chuong, J.N., Grigoriev, I. V, Arkin, A.P., Skerker, J.M., and Brem, R.B. (2018). Genetic dissection of interspecific differences in yeast thermotolerance. Nat. Genet. *50*, 1501.

Wilson, M.D., Barbosa-Morais, N.L., Schmidt, D., Conboy, C.M., Vanes, L., Tybulewicz, V.L.J., Fisher, E.M.C., Tavaré, S., and Odom, D.T. (2008). Species-Specific Transcription in Mice Carrying Human Chromosome 21. *322*, 434–438.

Wong, E.S., Schmitt, B.M., Kazachenka, A., Thybert, D., Redmond, A., Connor, F., Rayner, T.F., Feig, C., Ferguson-Smith, A.C., Marioni, J.C., et al. (2017). Interplay of cis and trans mechanisms driving transcription factor binding and gene expression evolution. Nat. Commun. *8*.

Yang, M.G., Ling, E., Cowley, C.J., Greenberg, M.E., and Vierbuchen, T. (2021). Characterization of sequence determinants of enhancer function using natural genetic variation.

Yue, J.-X., Li, J., Aigrain, L., Hallin, J., Persson, K., Oliver, K., Bergström, A., Coupland, P., Warringer, J., Lagomarsino, M.C., et al. (2017). Contrasting evolutionary genome dynamics between domesticated and wild yeasts. Nat. Genet. *49*, 913–924.

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

Zheng, W., Zhao, H., Mancera, E., Steinmetz, L.M., and Snyder, M. (2010). Genetic analysis of variation in transcription factor binding in yeast. Nature *464*, 1187–1191.

Zhou, T., Shen, N., Yang, L., Abe, N., Horton, J., Mann, R.S., Bussemaker, H.J., Gordân, R., and Rohs, R. (2015). Quantitative modeling of transcription factor binding specificities using DNA shape. Proc. Natl. Acad. Sci. U. S. A. *112*, 4654–4659.

Figure legends

Figure 1: Experimental system to profile cis-variation in transcription factor binding. A) Scheme of the experimental system. *S. cerevisiae* strains where a TF was fused to a MNase (illustrated as scissors) were crossed with a WT *S. paradoxus* strain to form a hybrid, on which the ChEC-seq method was applied to profile *in vivo* TF binding. Orthologous promoters harbor sequence variations (red and blue lines) and differential binding level. B) Global similarity in orthologous binding of the 27 TFs examined here. Presented is the Pearson correlation coefficient of sum of signal on all yeast promoters (6701 promoters), between experimental replicates (*S. cerevisiae* genome in red, *S. paradoxus* genome in blue) and between orthologous genomes (purple). Data is the mean and standard deviation of two to five repeats. Right: Promoter binding of three

TFs, each point is the sum of signal on a specific gene promoter. C-D) Examples for TF binding to orthologous promoters. C) Conservation of Swi5 binding to *SIC1* promoter, *S. cerevisiae* orthologue (upper panel) and *S. paradoxus* orthologue (lower panel). Chec-seq signal is the 5' end of reads, presented in purple. Nucleosome occupancy data of the hybrid (Tirosh et al., 2010) are presented as gray lines. Transcription start sites are presented in gray dashed-lines (Park et al., 2014; Pelechano et al., 2013). For Swi5, CCAGC motif sequences are marked in blue (plus strand) and black (minus strand) boxes. ORFs are presented as gray boxes. D) Binding site turnover of Tbf1 to *GIR2*/*CTH1* promoter. The blue box marks the region of binding site turnover, in which the Tbf1 motif appears on the plus strand in the *S. cerevisiae* allele (ACCTA), and the same motif realization appears on the minus strand in the *S. paradoxus* allele (TAGGT), where motif sequences partially overlap. Consensus motif of Tbf1 is [C/A]CCTA. E) Divergence in Reb1 binding to *GSP2* promoter. Annotation similar as in D, Reb1 consensus motif is TTACCC[G/T].

Figure 2: Transcription factors bind a selected subset of motif-coding sites. A) Reb1 binding signal to motif-coding sites (potential binding sites). Top: in vitro motif of Reb1 (Fordyce et al., 2010). Middle: average ChEC-seq signal (5' end of reads) in a logarithmic scale, bound sites presented in purple, non-bound sites presented in gray. Bottom: heatmap of ChEC-seq signal at 7115 sites that contain the Reb1 motif in both hybrid genomes. Right: motif p-value according to FIMO (Grant et al., 2011). B) Binding level distribution of Reb1 in Reb1 motif sites and in random sites. Binding level threshold was set as the 95% of random site distribution. This threshold defines the bound sites indicated in A. C) Signal around motif sites of all examined TFs, at bound (top row) and non-bound (bottom row) sites. Boxes indicate motif size. Full profiles are presented in Figure S2.

Figure 3: Differential TF binding to the two alleles correlates with variations in motif sequence, while differences in motif accessibility play a minor role. A) Suggested mechanisms for TF binding evolution. Left: motif site is conserved, but in the unbound allele it is occupied by a nucleosome and therefore not accessible for TF binding. Right: motif site is lost due to a single nucleotide variation. Nucleosomes are illustrated in yellow, TF: purple oval, motif site: blue box, nucleotide variation: gray stripe.

B) Proportion of motif sites (left) and proportion of bound sites among the common motif sites (right). Left: proportion of motif sites than are common to both orthologues, and sites that appear only in a certain orthologue (cer- or par-unique) among all *in vitro* defined motif sites of the full set of 27 TFs (62,970 are common, 63,455 cer-unique, 46,440 par-unique, see Figure S6 for TF-specific proportions). Right: proportion of binding to common motif sites.
C) Nucleosome occupancy does not explain differential binding between orthologues. Presented are nucleosome occupancy profiles averaged over motif sites, centered at the binding motif of Reb1 (top panels) and Sok2 (lower panels). Left: all motif-coding sites, divided to bound sites (purple) and non-bound sites (gray). Middle: Common motif sites which show diverged binding. Nucleosomes at bound allele in purple, nucleosomes at non-bound allele in gray. Right: Unique motif sites with diverged binding. Nucleosomes at the bound allele, which harbors a motif, in purple, nucleosomes at the non-bound and motif-less allele in gray (profiles of all TFs in Figure S7).
D) Binding to unique-motif sites, with biased binding to the motif-containing allele. Left: percent of bound sites. Middle: odds ratio of Fisher's exact test, full black circles indicate significant comparisons (p-value < 0.05, FDR corrected). Right: Number of unique-motif sites.

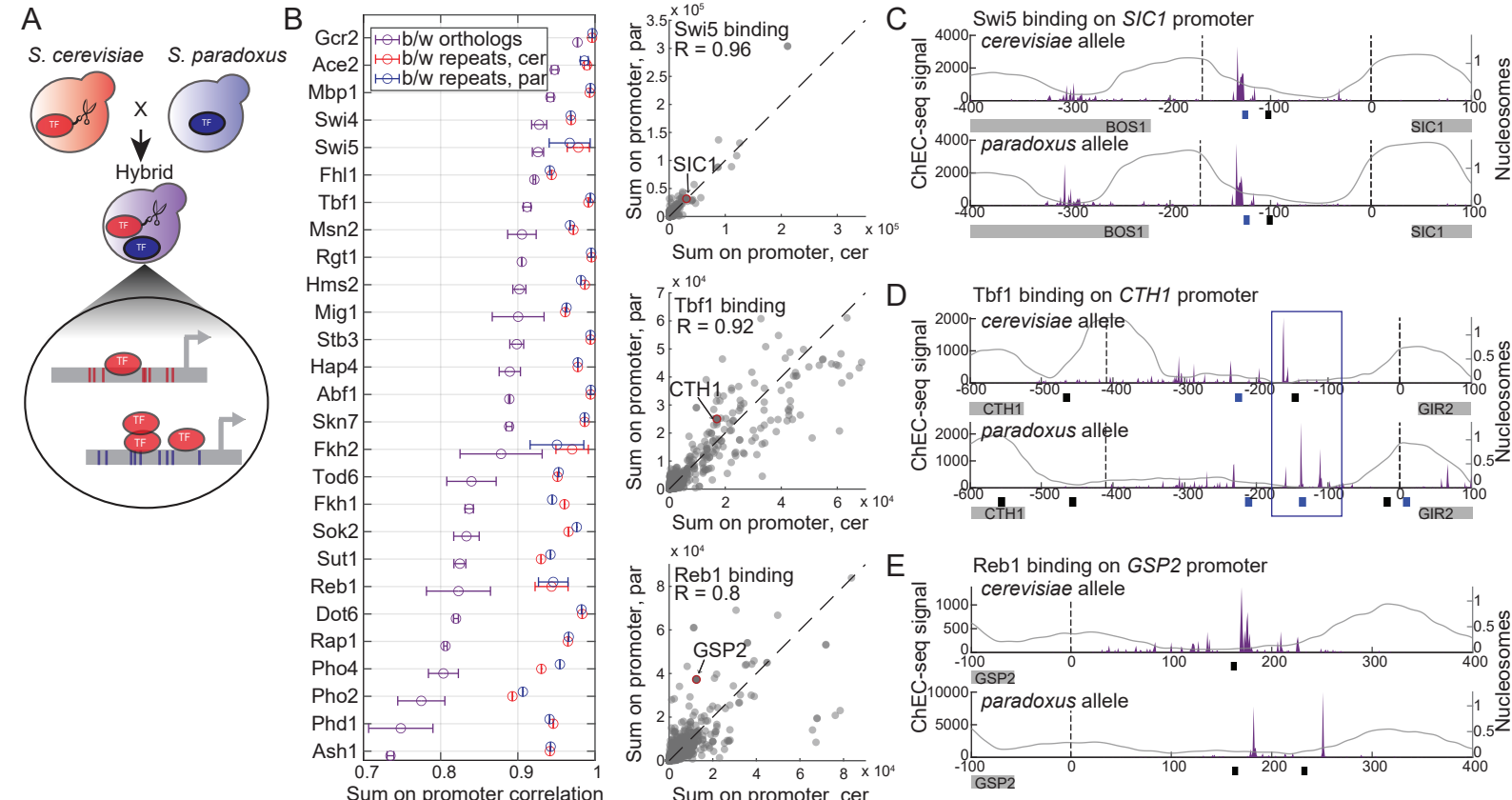
Figure 4: The cost of *cis*-regulatory mutations on TF binding. A) The mutation cost of Reb1, measured at unique alternative sites where one orthologue has the consensus motif (as defined *in vitro*), while the second orthologue has a one-letter variant (alternative). Each dot represents the mean of at least two sites (see Figure S9 B for the number of sites). Seqlogo of *in vitro* motif of Reb1 (Fordyce et al., 2010) is presented on top. B) Mutation cost of Reb1, measured for common alternative sites, where both orthologues have the same one-letter variant. These sites are compared to common consensus sites found elsewhere in the genome. C) Costs of unique alternatives for 22 TFs. The heatmap represents change in binding as in A, here the four rows stand for the four nucleotides A, C, G, T. Red box represents the consensus allele. Minimal two sites, gray color represents missing data. Bases flanking the motif have no consensus sequence therefore the computation was done relative to the most common nucleotide. D) Cost of common alternatives for 22 TFs, as in C. E) Common alternative binding sites are found at conserved genomic regions. Shown is the sequence conservation between *S. cerevisiae* and *S. paradoxus* orthologues (same nucleotide = 1, different

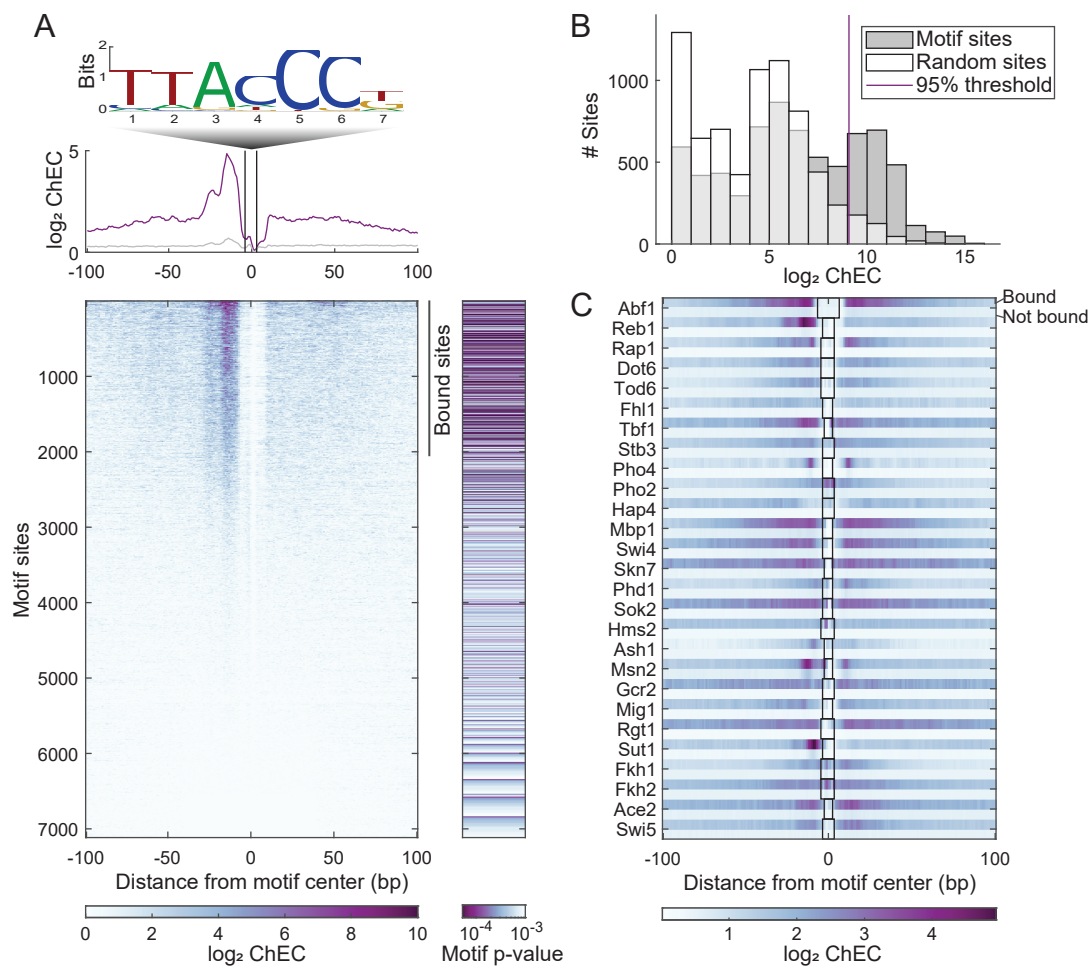
nucleotide/ INDEL = 0) at Reb1 binding sites of type: common alternative, unique alternative and common consensus, as well as in random sites at promoters. Shown is the mean signal per group. The number of sites in each group is indicated in parenthesis. F) Common alternative sites are conserved through the yeast lineage. The phastCons conservation score (Siepel et al., 2005) is shown for the three Reb1 site groups as in E.

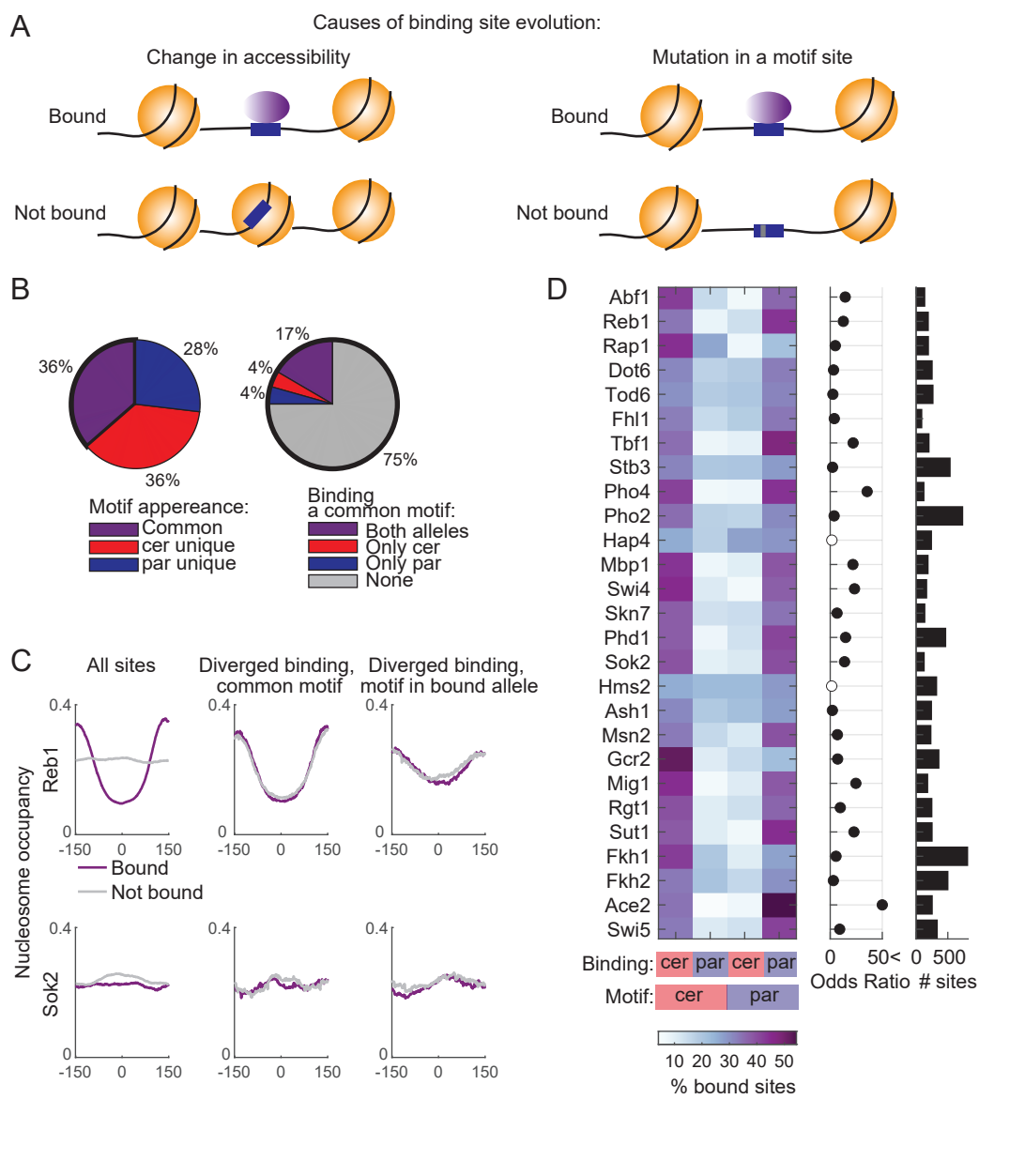
Figure 5: Sequence variation predicts DNA binding variation. A) Change in motif score predicts variation in Reb1 binding. Shown is the $\log_2$-ratio of for 7-mer motif scores (x-axis) and ChEC-seq signal (y-axis) between *S. cerevisiae* and *S. paradoxus* orthologues, at peaks associated with a non-conserved Reb1 motif (i.e., a different motif sequence appears in each orthologue). Sequence alignment at two specific sites are presented, *S. cerevisiae* orthologue on the upper row and *S. paradoxus* orthologue in the lower row, Reb1 motif written in red. B) Linear models predict binding variation at peaks associated with non-conserved motifs. Shown is the percent of explained variability (R^2 * 100) for each TF, using three models: in vitro PWM score, 7-mer motif score derived from our data, and a compilation of these with another four predictors (see text). Prediction for other peak categories are presented in Figure S11 A.

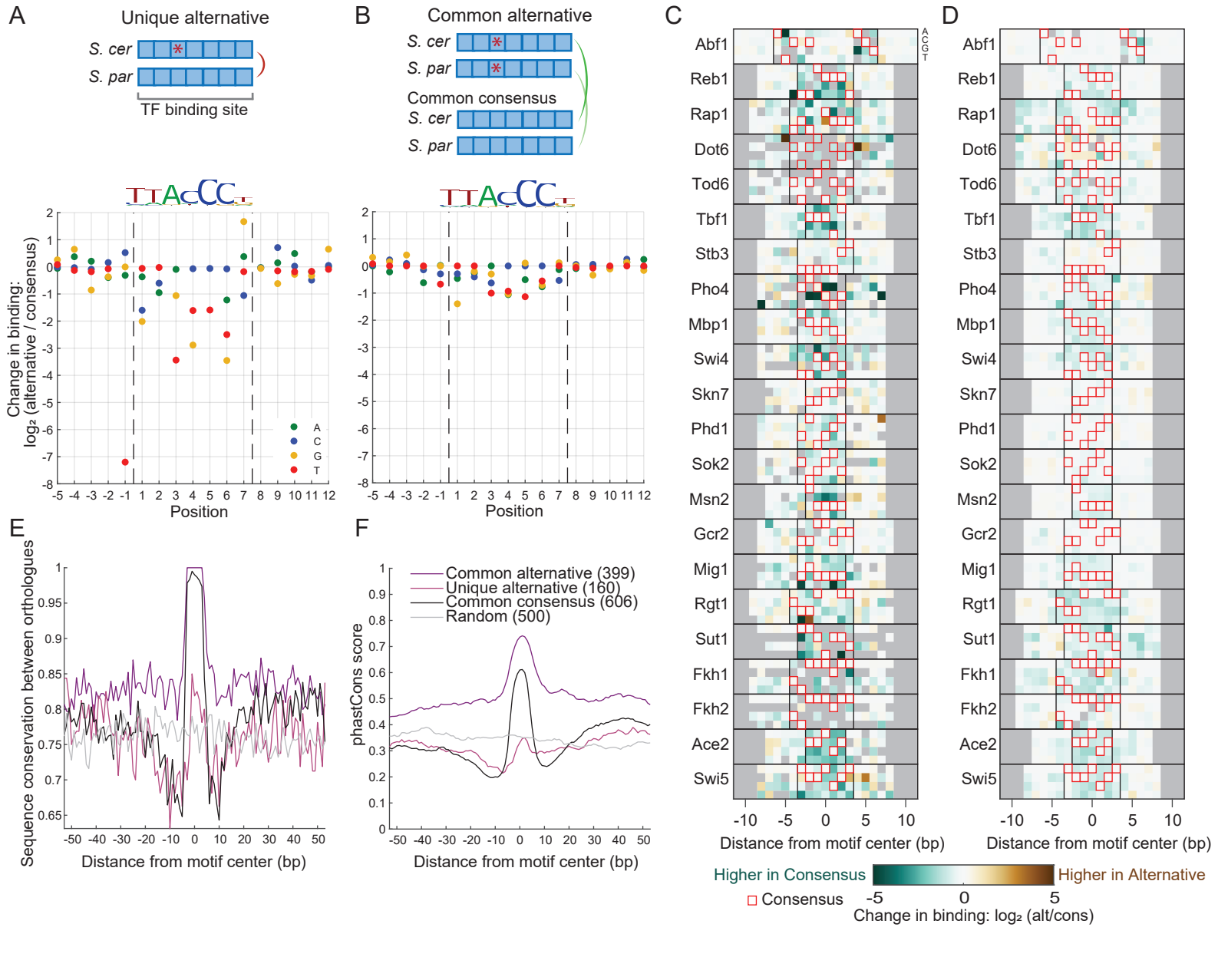
Figure 6: Promoter evolution and binding site turnover. A-D) Four evolutionary classes of TF binding variation. Schemes are shown in the left most panels, genome-browser snapshots of Reb1-bound promoters are shown as examples on the middle and right panel. Sequence alignment between orthologues is presented beneath each examples, X-axis represent location on the promoter relative to TSS. figure legend beneath panel D. A) Conserved: all binding sites are species-conserved. B) Turnover: reciprocal gain and loss of binding sites. C) Unbalanced: species-specific sites along with conserved sites. D) Fully unbalanced: Binding sites appear in only one orthologue. E) Proportion of number of promoters and binding signal per promoter class. Shown is the distribution of the full set of TFs, proportion per TF is presented in Figure S13 A. Binding signal refers to the ChEC-seq signal in the higher-bound orthologue, normalized by the total signal in that orthologue. F) Examples of short-distance binding site turnover. Shown are the binding signal in both alleles, and the sequence alignment in the lower panel. Boxes mark motifs, arrows mark the motif's strand. G) Correlation on promoters is

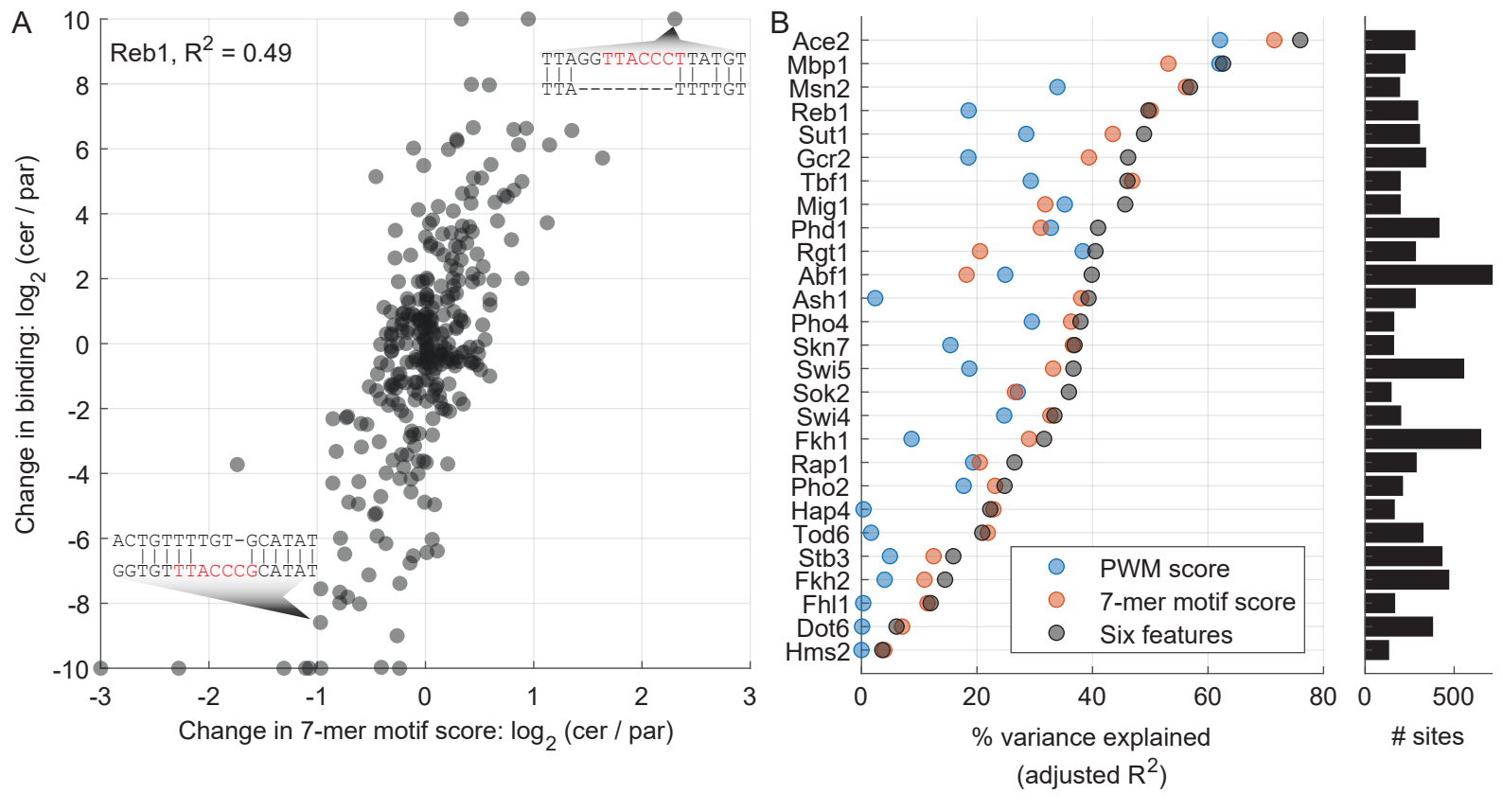
higher than correlation on peaks when comparing orthologues. Shown are correlation coefficients between orthologues, over motif-associated peaks (y-axis) and promoters (x-axis), among all promoters. Here, we summed the binding signal only on peaks within peak-containing promoters, but the correlation coefficients were quantitatively similar to those obtained from the more simplistic approach of summing up the signal over the full promoter, as presented in Figure 1 B (see Figure S15 A for comparison). H) The shift in promoter correlation vs. peak correlation is more apparent at turnover and unbalanced promoters. Shown is the correlation between orthologues, summing over promoters (x-axis) and over motif-associated peaks (y-axis) as in G, per promoter class. I) Turnover promoters show a higher promoter similarity despite of lower peak similarity. Shown are the differences between correlation on promoters to correlation on peaks for the different promoter classes. Each dot represents a TF, letters represent statistically distinguished groups after Tukey's-honestly significant difference test.

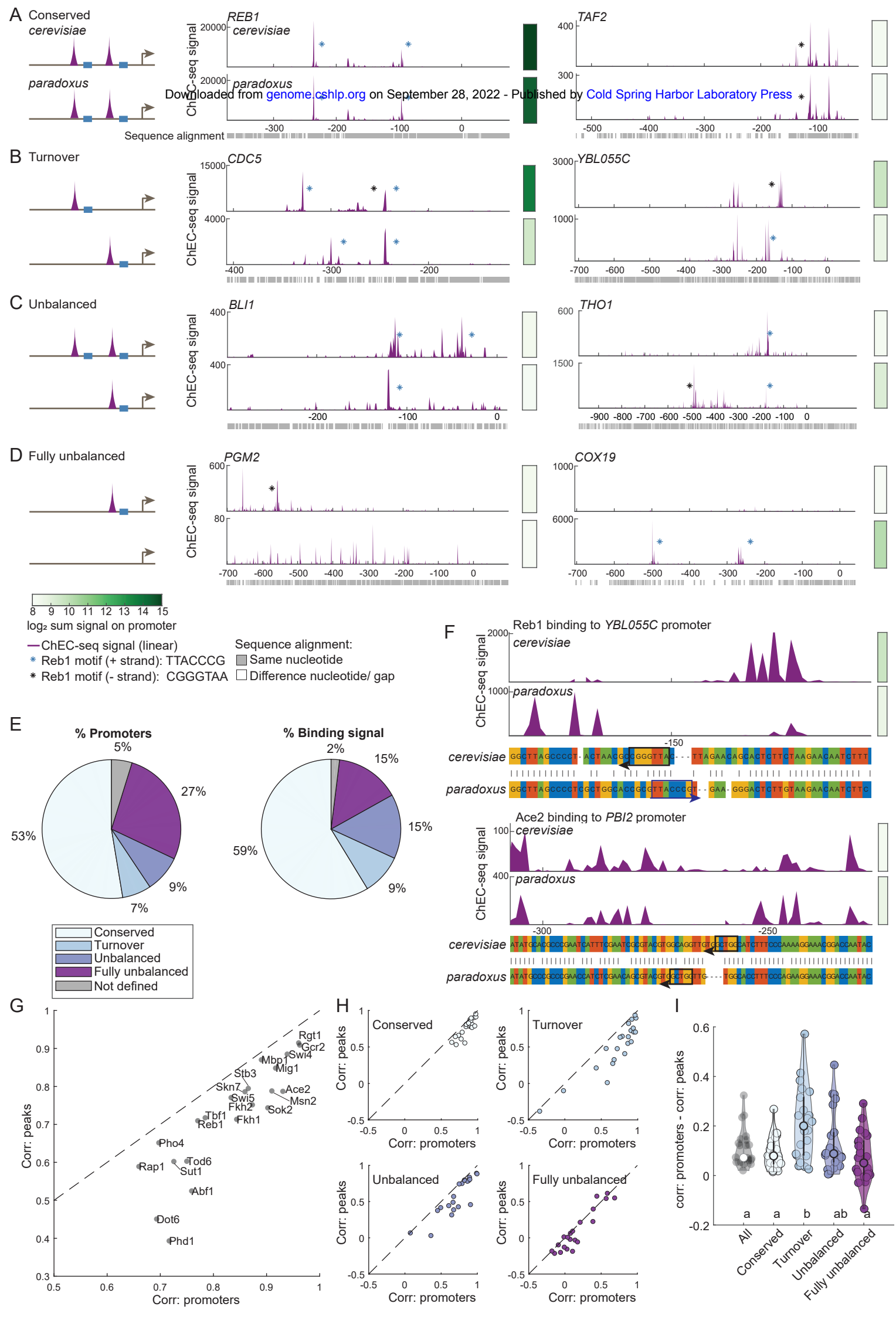

A
S. cerevisiae
S. paradoxus
X
Hybrid
TF
B
b/w orthologs
b/w repeats, cer
b/w repeats, par
Gcr2
Ace2
Mbp1
Swi4
Swi5
Fhl1
Tbf1
Msn2
Rgt1
Hms2
Mig1
Stb3
Hap4
Abf1
Skn7
Fkh2
Tod6
Fkh1
Sok2
Sut1
Reb1
Dot6
Rap1
Pho4
Pho2
Phd1
Ash1
0.7
0.8
0.9
1
Sum on promoter correlation
Swi5 binding
R = 0.96
SIC1
Tbf1 binding
R = 0.92
CTH1
Reb1 binding
R = 0.8
GSP2
Sum on promoter, par
Sum on promoter, cer
C
Swi5 binding on SIC1 promoter
cerevisiae allele
paradoxus allele
ChEC-seq signal
Nucleosomes
BOS1
SIC1
D
Tbf1 binding on CTH1 promoter
CTH1
GIR2
E
Reb1 binding on GSP2 promoter
GSP2


A
Bits
log₂ ChEC
Motif sites
Bound sites
Distance from motif center (bp)
log₂ ChEC
Motif p-value
B
Motif sites
Random sites
95% threshold
Sites
log₂ ChEC
C
Bound
Not bound
Abf1
Reb1
Rap1
Dot6
Tod6
Fhl1
Tbf1
Stb3
Pho4
Pho2
Hap4
Mbp1
Swi4
Skn7
Phd1
Sok2
Hms2
Ash1
Msn2
Gcr2
Mig1
Rgt1
Sut1
Fkh1
Fkh2
Ace2
Swi5
Distance from motif center (bp)
log₂ ChEC


A
Causes of binding site evolution:
Change in accessibility
Mutation in a motif site
Bound
Not bound
Bound
Not bound
B
36%
28%
36%
17%
4%
4%
75%
Motif appeerance:
Common
cer unique
par unique
Binding
a common motif:
Both alleles
Only cer
Only par
None
C
All sites
Diverged binding,
common motif
Diverged binding,
motif in bound allele
Nucleosome occupancy
Reb1
Sok2
0.4
0
-150
0
150
Bound
Not bound
D
Abf1
Reb1
Rap1
Dot6
Tod6
Fhl1
Tbf1
Stb3
Pho4
Pho2
Hap4
Mbp1
Swi4
Skn7
Phd1
Sok2
Hms2
Ash1
Msn2
Gcr2
Mig1
Rgt1
Sut1
Fkh1
Fkh2
Ace2
Swi5
Binding:
cer
par
cer
par
Motif:
cer
par
0
50<
Odds Ratio
0
500
sites
10
20
30
40
50
% bound sites


A
Unique alternative
S. cer
S. par
TF binding site
B
Common alternative
S. cer
S. par
Common consensus
S. cer
S. par
Change in binding:
log₂ (alternative / consensus)
A
C
G
T
Position
C
D
Abf1
Reb1
Rap1
Dot6
Tod6
Tbf1
Stb3
Pho4
Mbp1
Swi4
Skn7
Phd1
Sok2
Msn2
Gcr2
Mig1
Rgt1
Sut1
Fkh1
Fkh2
Ace2
Swi5
Distance from motif center (bp)
Higher in Consensus
Higher in Alternative
Consensus
Change in binding: log₂ (alt/cons)
E
Sequence conservation between orthologues
Distance from motif center (bp)
F
phastCons score
Common alternative (399)
Unique alternative (160)
Common consensus (606)
Random (500)
Distance from motif center (bp)


A
Reb1, R2 = 0.49
TTAGGTTACCCTTATGT
TTA--------TTTTGT
ACTGTTTTGT-GCATAT
GGTGTTTACCCGCATAT
Change in binding: log2 (cer / par)
Change in 7-mer motif score: log2 (cer / par)
B
Ace2
Mbp1
Msn2
Reb1
Sut1
Gcr2
Tbf1
Mig1
Phd1
Rgt1
Abf1
Ash1
Pho4
Skn7
Swi5
Sok2
Swi4
Fkh1
Rap1
Pho2
Hap4
Tod6
Stb3
Fkh2
Fhl1
Dot6
Hms2
PWM score
7-mer motif score
Six features
% variance explained
(adjusted R2)
sites



Evolution of transcription factor binding through sequence variations and turnover of binding sites

Gat Krieger, Offir Lupo, Patricia Wittkopp, et al.

Genome Res. published online May 26, 2022
Access the most recent version at doi:10.1101/gr.276715.122