Evolutionary divergence of intrinsic and *trans*-regulated nucleosome positioning sequences reveals plastic rules for chromatin organization

Alex Tsankov,[1,2] Yoshimi Yanagisawa,[3] Nicholas Rhind,[3,6] Aviv Regev,[1,2,4,5,6] and Oliver J. Rando[3,5,6]

[1]*Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA;* [2]*Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts 02140, USA;* [3]*Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, Massachusetts 01605, USA;* [4]*Howard Hughes Medical Institute, Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02140, USA*

The packaging of eukaryotic genomes into nucleosomes plays critical roles in chromatin organization and gene regulation. Studies in *Saccharomyces cerevisiae* indicate that nucleosome occupancy is partially encoded by intrinsic antinucleosomal DNA sequences, such as poly(A) sequences, as well as by binding sites for *trans*-acting factors that can evict nucleosomes, such as Reb1 and the Rsc3/30 complex. Here, we use genome-wide nucleosome occupancy maps in 13 *Ascomycota* fungi to discover large-scale evolutionary reprogramming of both intrinsic and *trans* determinants of chromatin structure. We find that poly(G)s act as intrinsic antinucleosomal sequences, comparable to the known function of poly(A)s, but that the abundance of poly(G)s has diverged greatly between species, obscuring their antinucleosomal effect in low-poly(G) species such as *S. cerevisiae*. We also develop a computational method that uses nucleosome occupancy maps for discovering *trans*-acting general regulatory factor (GRF) binding sites. Our approach reveals that the specific sequences bound by GRFs have diverged substantially across evolution, corresponding to a number of major evolutionary transitions in the repertoire of GRFs. We experimentally validate a proposed evolutionary transition from Cbf1 as a major GRF in pre-whole-genome duplication (WGD) yeasts to Reb1 in post-WGD yeasts. We further show that the mating type switch-activating protein Sap1 is a GRF in *S. pombe*, demonstrating the general applicability of our approach. Our results reveal that the underlying mechanisms that determine in vivo chromatin organization have diverged and that comparative genomics can help discover new determinants of chromatin organization.

[Supplemental material is available for this article.]

The eukaryotic genome is packaged into nucleosomes, which consist of 147 base pairs wrapped around an octamer of histone proteins. The genomic locations of nuclesomes play a critical role in cellular processes that involve DNA (Kornberg and Lorch 1999; Radman-Livaja and Rando 2010). Genome-wide mapping of nucleosome positions in budding yeast shows that most genes contain a long nucleosome-depleted region in their proximal promoter, commonly called the "nucleosome-free region" (NFR) (Rando and Ahmad 2007; Jiang and Pugh 2009; Rando and Chang 2009). Nucleosomes immediately downstream from the NFR are generally well positioned, with positioning decaying with increasing distance into gene bodies (Kornberg 1981; Kornberg and Stryer 1988; Yuan et al. 2005; Mavrich et al. 2008; Mobius and Gerland 2010; Vaillant et al. 2010).

Several mechanisms have been proposed for establishing nucleosome locations in vivo, largely based on observations in the model organism *S. cerevisiae*. DNA sequence can thermodynamically favor or repel histone binding. Most importantly, AT-rich antinucleosomal sequences, such as poly(A) tracts, are the most predictive intrinsic sequence signals for establishing chromatin structure in vivo (Drew and Travers 1985; Sekinger et al. 2005; Ioshikhes et al. 2006; Peckham et al. 2007; Kaplan et al. 2008; Yuan and Liu 2008; Zhang et al. 2009). Poly(A) sequences are enriched in yeast nucleosome-free regions (Yuan et al. 2005) and are depleted of histones in nucleosome reconstitution experiments in vitro (Kaplan et al. 2008; Zhang et al. 2009). Nucleosome depletion over poly(A) elements scales with homopolymer length (Field et al. 2008) and, in general, nucleosome depletion in vivo correlates with overall AT% (Tillo and Hughes 2009).

In addition, *trans*-acting proteins can move or evict nucleosomes, thereby overcoming sequence preferences (Whitehouse et al. 2007; Clapier and Cairns 2009). A key class of sequence-directed *trans* factors are "general regulatory factors" (GRFs), highly abundant, sequence-specific DNA-binding proteins that cause nucleosome eviction in vivo. Loss of GRFs in vivo results in increased nucleosome occupancy over their binding sites. GRFs may function by directly competing with nucleosomes for binding to DNA, or by recruiting the RSC chromatin remodeling complex (Yu and Morse 1999; Yarragudi et al. 2004; Raisner et al. 2005; Badis et al. 2008; Hartley and Madhani 2009; Ganapathi et al. 2010).

Here, we use genome-wide nucleosome mapping from 13 *Hemiascomycota* fungi (12 recently published profiles [Tsankov et al. 2010] and new data for *S. pombe*) to gain a broader understanding of the mechanisms that establish chromatin structure. We identify

[5]**These authors contributed equally to this work.**
[6]**Corresponding authors.**
E-mail nick.rhind@umassmed.edu.
E-mail aregev@broad.mit.edu.
E-mail oliver.rando@umassmed.edu.
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.122267.111. Freely available online through the *Genome Research* Open Access option.

large-scale evolutionary reprogramming in the determinants of chromatin organization and use it to discover new intrinsic and *trans*-regulated nucleosome positioning sequences. We find that poly(G) tracts act intrinsically to direct nucleosome depletion in a manner analogous to poly(A) tracts. In addition, by identifying sequences that are specifically nucleosome-depleted in vivo but not in vitro, we computationally infer the putative binding motifs for GRFs, two of which we validate experimentally. Together, these results provide a broad perspective on the relationship between genome evolution and chromatin establishment, and provide a model system for understanding the mechanistic basis for GRF function.

Results

Conservation and divergence in nucleosome positioning sequences

To study the evolution of sequence characteristics underlying nucleosome occupancy across organisms, we used genome-wide nucleosome position maps in 13 *Ascomycota* species (Fig. 1A), generated by Illumina sequencing of mononucleosomal DNA isolated from mid-log cultures. For each species' chromatin map, we determined the normalized nucleosome occupancy per base pair, correcting for differences in sequencing depth and MNase digestion level (see Methods). In order to investigate the sequence characteristics underlying nucleosome–DNA interactions, we calculated the extent of relative nucleosome depletion over all 7-mer sequences (see Methods) in the genome of each species (Fig. 1B).

We found substantial conservation in the degree of nucleosome depletion over most sequences, with some notable exceptions. Most (89%) 7-mers varied little (variance <0.5) in their nucleosome depletion across all species (7.3% expected by chance, Binomial $P < 10^{-300}$, Methods). AT-rich, intrinsic antinucleosome sequences (Drew and Travers 1985; Iyer and Struhl 1995; Sekinger et al. 2005; Kaplan et al. 2008; Zhang et al. 2009) were highly depleted of nucleosomes in vivo in all species, although the level of depletion was variable across species (Fig. 1B,C). Interestingly, of the 480 7-mers that were nucleosome depleted (mean occupancy <−0.75) in at least one species, 188 (39%) were not particularly nucleosome depleted in *S. cerevisiae* (mean occupancy >−0.25 both in vivo and in vitro). Thus, models that predict nucleosome occupancy based on sequence must be applied to other species with caution (as noted in Lantermann et al. 2010), as these models depend on N-mers that are nucleosome depleted in *S. cerevisiae* (Kaplan et al. 2008). This observation highlights the importance of understanding the factors that determine nucleosome positioning across a wide variety of organisms.

PolyGs act as global intrinsic antinucleosomal sequences

Among the differentially depleted 7-mers, poly(G)-rich sequences (Fig. 1B, purple bar), such as GGGGGGG (G7), were strongly depleted of nucleosomes in vivo in some of the species (e.g., *S. bayanus*, *S. castellii*, *C. glabrata*, *D. hansenii*), but not in the model organism *S. cerevisiae* and its closest relatives (Fig. 1B,C). Classic studies on the *HIS3* promoter showed that poly(G) can substitute for poly(A) as an antinucleosomal sequence (Drew and Travers 1985; Iyer and Struhl 1995), suggesting that poly(G)s may function as global intrinsic antinucleosomal sequences. Supporting this hypothesis, we found that G7 is nucleosome depleted when *C. albicans* genomic DNA is assembled into nucleosomes in vitro (Fig. 1C; Field et al. 2009). In vitro nucleosome depletion over both poly(G)s and poly(A)s scales with homopolymeric run length (Fig. 1D). Interestingly, nucleosome depletion increases with length to a slightly greater extent over poly(G)s than over poly(A)s (Fig. 1D), suggesting that poly(G)s may repel nucleosomes more efficiently. In addition, poly(G)s reside predominantly in intergenic regions (Fig. 1E), as is the case for poly(A) tracts (Iyer and Struhl 1995).

In vivo, poly(G) elements were highly depleted of nucleosomes in a number of yeast species in a phylogenetically coherent way (Fig. 1C,F). Poly(G)s of various lengths were significantly depleted of nucleosomes in three species that diverged successively post-WGD (*S. castelli*, the human pathogen *C. glabrata*, and *S. bayanus*). Poly(G)s were also depleted in *C. albicans* (Fig. 1F), although the extent of nucleosome depletion was much greater in vitro than in vivo, suggesting that *trans*-acting factor(s) may play an active role in occluding poly(G) tracts in vivo in this species.

The evolutionary changes in the relative dominance of poly(A) or poly(G) tracts as antinucleosomal sequences may be related to larger trends in the evolution of genome sequence composition. The abundance of poly(A) and poly(G) sequences of various lengths varies considerably across the 13 genomes (Fig. 1E; Supplemental Table 1), especially at intergenic regions. Furthermore, the presence of G7 and A7 sequences in NFRs varies significantly between species, with poly(A)s being more abundant in most species, including *C. albicans* (Supplemental Table 2). While it may appear counter-intuitive that nucleosome depletion across an intrinsic sequence can vary between species, this may be due to the fact that G7 can occur as part of a longer (G8, G9, etc.) tract, and hence, the average depletion over G7 sequences partially reflects the distribution of longer poly(G) elements in the genome of interest. Indeed, most species with abundant long poly(G) sequences (*S. bayanus*, *C. glabrata*, *S. castellii*) exhibit strong nucleosome depletion over G7 sequences. The two exceptions [*Y. lipolytica* and *C. albicans*, which carry many poly(G) elements that are not particularly nucleosome depleted in vivo], most likely result from in vivo regulation of poly(G) exposure (see below). Finally, the overall abundance of long poly(A) and poly(G) sequences at promoters is positively correlated with the median NFR widths in all species ($R = 0.735$, $P = 0.0028$) (Fig. 1G). Specifically, species with fewer intrinsic antinucleosomal sequences, such as *Kluyveromyces waltii*, have shorter NFRs on average. Thus, the genome composition of intrinsic antinucleosomal sequences can impact global characteristics of a species' chromatin organization.

Identifying motifs for *trans*-acting chromatin regulators

Other novel sequences were nucleosome depleted in some of our species in vivo, but not in either the *S. cerevisiae* or the *C. albicans* in vitro reconstitutions (Fig. 1B), suggesting that they are binding sites for *trans*-acting regulators. Previous studies in *S. cerevisiae* (Yarragudi et al. 2007; Kaplan et al. 2008; Clapier and Cairns 2009) and our own analysis across 12 species (Tsankov et al. 2010) have confirmed that such sequences can correspond to binding sites for known general regulatory factors (GRFs), such as Reb1 (Fig. 1B, orange; Badis et al. 2008; Zhu et al. 2009) and the Rsc3/30 components of the RSC chromatin remodeling complex (Fig. 1B, green; Badis et al. 2008; Clapier and Cairns 2009; Zhu et al. 2009).

To systematically identify binding sites of GRFs, we developed a new computational procedure (Fig. 2A; see Methods). In the first step, our method clusters nucleosome-depleted 7-mers based on sequence similarity. In the next step, it finds a parsimonious number of position specific scoring matrices (PSSMs) that represent consensus DNA-binding sequences for potential GRFs. This pro-

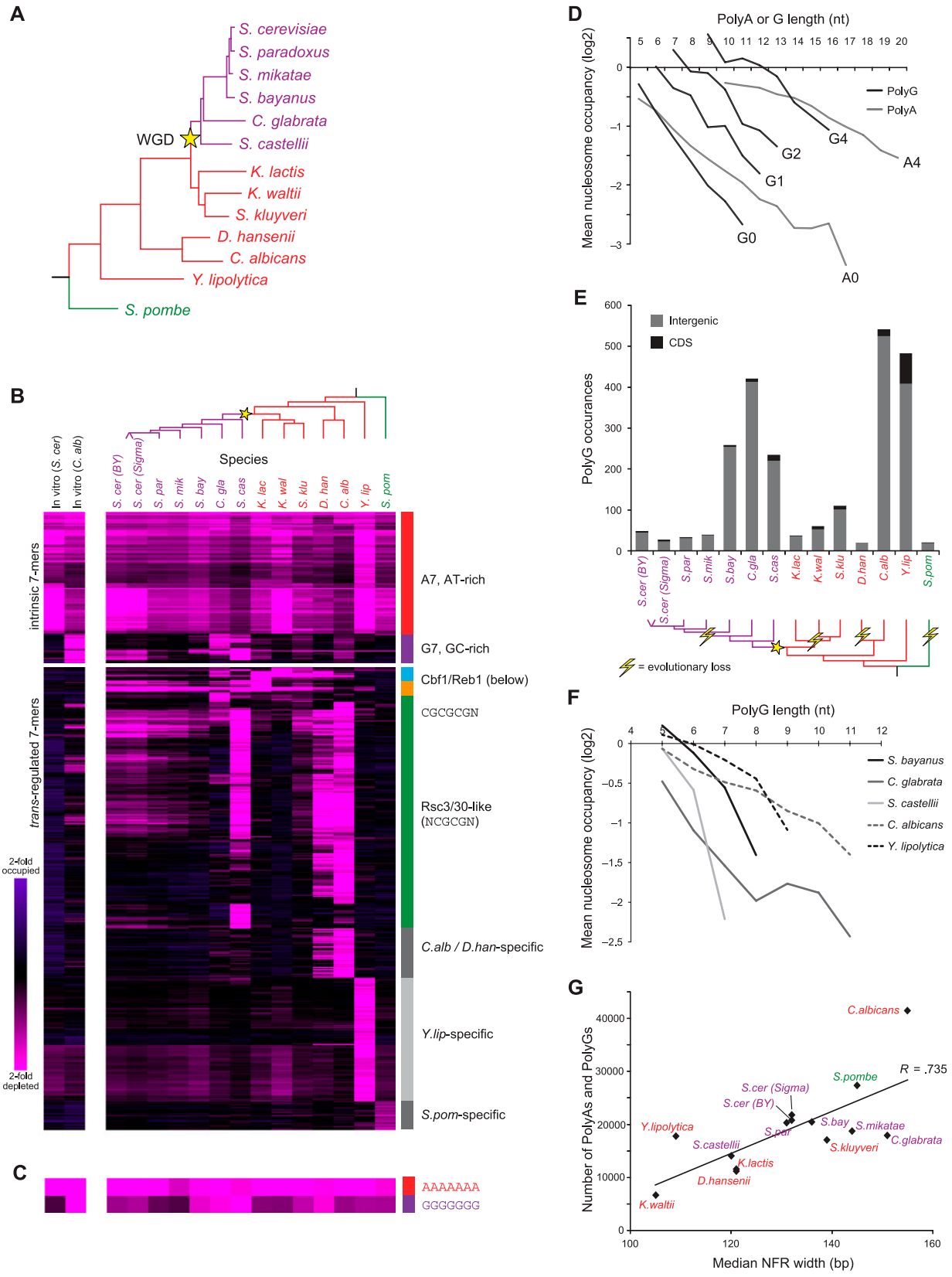

Figure 1. (Legend on next page)

cedure automatically predicts the correct (known) PSSMs of GRFs in *S. cerevisiae*, including Reb1 and Rsc3/30 (Fig. 2A).

We uncovered several novel candidate GRF binding sites that act in one or more species, but not in the model organism *S. cerevisiae* (Fig. 2B-F; Supplemental Table 3). We focus on three main findings: (1) the CACGTGA motif (Fig. 2B) that serves as the binding site for Cbf1 in *S. cerevisiae* and *C. albicans* (Harbison et al. 2004; Badis et al. 2008; Hogues et al. 2008; Zhu et al. 2009) and is strongly nucleosome depleted in a subset of pre-WGD species; (2) two PSSMs associated with nucleosome depletion only in *Y. lipolytica* (Fig. 2D,E), and one in *C. albicans* (Fig. 2C); and (3) a PSSM from *S. pombe* similar to the binding site for Sap1, a protein involved in DNA replication and recombination (Fig. 2F).

Cbfl acts as a global GRF in *C. albicans*

We first tested our hypothesis that Cbf1 is a GRF in a pre-WGD species, but not in a post-WGD species. We began by ruling out several alternative hypotheses for depletion over CACGTGA sequences in pre-WGD species. First, we found that the binding site CACGTGA is similarly depleted of nucleosomes in *S. cerevisiae* grown in glucose (above) and in ethanol (Kaplan et al. 2008), where Cbf1 up-regulates respiration genes (Fig. 3A; Lavoie et al. 2010). Thus, nucleosome depletion in pre-WGD species is not simply a consequence of the higher expression of respiration genes in these species (Conant and Wolfe 2007). Second, the nucleosome depletion over CACGTGA sequences in pre-WGD species could have been an artifact of genome organization if these motifs were located in these species closer to intrinsic antinucleosomal sequences, such as poly(A) or poly(G). However, Cbf1 sites were nucleosome occupied in the in vitro data from *C. albicans*, but were nucleosome depleted in vivo, ruling out this possibility (Fig. 3A).

Finally, we noted that several related *S. cerevisiae* proteins, including Cbf1, Pho4, Rtg3, and Tye7, bind variants of the CACGTGA motif (Badis et al. 2008; Zhu et al. 2009). In order to determine whether Cbf1 or some other protein acts as a GRF specifically in pre-WGD species, we therefore measured nucleosome positions across the genome in *S. cerevisiae* and *C. albicans* strains lacking Cbf1 (Fig. 3; Biswas et al. 2003; Lavoie et al. 2010). We chose *C. albicans* over other pre-WGD species (which exhibit an even more prominent role for Cbf1), since it is a conservative test of our hypothesis, and since strong GRF proteins are often essential, complicating genetic analysis.

Our data show that Cbf1 significantly affects promoter nucleosome occupancy in *C. albicans*, but much less so in *S. cerevisiae*. First, in *C. albicans,* Cbf1 sites are enriched within the NFRs that become most occluded when comparing occupancy between a *cbf1Δ* strain and the wild-type (Fig. 3B). Such changes are much weaker in *S. cerevisiae* (Supplemental Fig. 1A). Second, in *C. albicans*, the only two 7-mers that become substantially occluded in the *cbf1Δ* strain compared with the wild-type are the experimentally validated binding sites for Cbf1 (Hogues et al. 2008; Lavoie et al. 2010), CACGTGA and ACGTGAC (Fig. 3C, blue squares). In contrast, nucleosome occupancy of no particular 7-mer was substantially affected by *CBF1* deletion in *S. cerevisiae* (Fig. 3D). Finally, average nucleosome occupancy across all intergenic CACGTGA occurrences increases by 67% in the *cbf1Δ* strain in *C. albicans* (paired *t*-test $P = 4 \cdot 10^{-47}$) (Fig. 3E; Supplemental Fig. 1C) compared with a much smaller, albeit significant increase of 19% in *S. cerevisiae* (paired *t*-test $P = 5 \cdot 10^{-5}$; Supplemental Fig. 1B,D). Together, these results confirm that Cbf1 functions globally as a GRF in *C. albicans* through the binding site CACGTGA, and has largely lost this function in *S. cerevisiae*.

Antagonistic relationship between GRFs and antinucleosomal sequences

Interestingly, GC-rich 7-mers become significantly depleted of nucleosomes when Cbf1 is removed (Fig. 3C), and poly(G) elements are remarkably more nucleosome depleted (Fig. 3F) in *C. albicans* strains lacking Cbf1 (where 38% of poly(G)s are located in NFRs) than in WT strains (19% in NFRs). Indeed, poly(G) and poly(A) sequences of all lengths become more nucleosome depleted in the *C. albicans cbf1Δ* strain compared with the wild-type (Fig. 3G,H), while this was not the case in *S. cerevisiae* (Fig. 3I; Supplemental Fig. 1E). This observation is surprising, since poly(A)s and GRF binding sites have been shown to cooperate in establishing NFRs at certain genes in *S. cerevisiae* (Lascaris et al. 2000; Raisner et al. 2005). This result supports our hypothesis (Fig. 1C) that the degree of global in vivo nucleosome occupancy of the intrinsic antinucleosomal G7 sequence is also affected by species-specific *trans*-regulators. The mechanistic basis for this antagonistic coupling between Cbf1 and polyGs is still unknown, since we find that they do not generally co-occur at the same promoters (Supplemental Fig. 2A), indicating that poly(G) coverage in wild-type cells does not result from Cbf1-mediated nucleosome sliding.

To test whether this tradeoff between GRFs and intrinsic antinucleosomal sequences is a general property of GRFs in other species, we analyzed nucleosome occupancy at poly(A) sequences using published maps in *S. cerevisiae* strains depleted for the GRFs Abf1, Reb1, and Rsc3 (Badis et al. 2008). Indeed, poly(A) sequences become more nucleosome depleted in all three GRF mutants when compared with wild-type strains (Fig. 4). As with Cbf1 and poly(G), this appears not to be a consequence of juxtaposed sites for the

Figure 1. PolyG is an instrinsic antinucleosomal sequence element. (*A*) Phylogeny of species included in this study (adapted from Wapinski et al. 2007). Whole-genome duplication (WGD) event is marked by the yellow star. Species names are colored to denote major phylogenetic groups. (*B*) Nucleosome occupancy over 7-mers. All 7-mer sequences (rows) with mean $\log_2$ occupancy <−0.75 (there are no 7-mers with occupancy >0.75) in at least one species (columns) are shown across all species for in vivo data (*righthand* 14 columns) (Tsankov et al. 2010; Weiner et al. 2010) and for in vitro reconstitution experiments (*lefthand* two columns) (Kaplan et al. 2008; Field et al. 2009). Sequences are clustered by their nucleosome occupancy profiles and specific clusters are marked on *right*. (Pink) depleted; (violet) occupied. (*C*) Data for 7-mers AAAAAAA and GGGGGGG, as in *B*. (*D*) poly(G) sequences affect nucleosome depletion in vitro in a similar manner to poly(A) sequences. Shown is the average $\log_2$ nucleosome occupancy (*y*-axis) from in vitro reconstitution of *C. albicans* genomic DNA (Field et al. 2009) for poly(A) and poly(G) sequences of various lengths (*x*-axis). Ai (Gi) refers to poly(A) [poly(G)] sequences with *i* mismatches (e.g., A0, no mismatches; A4, four mismatches). (*E*) Abundance and locations of poly(G) sequences in each species. (*Top*) Shown are values for sequences of strength 4 or greater (see Methods). (Gray) Positioned within intergenic region; (black) positioned within coding sequence (CDS). (*Bottom*) A phylogenetic reconstruction of evolutionary losses (lightning bolt) of abundance in poly(G) sequences along the phylogeny. (*F*) poly(G) elements are also nucleosome depleted in vivo in several yeast species. Shown are the mean in vivo $\log_2$ nucleosome occupancies (*y*-axis) for poly(G) sequences (no mismatches) of different lengths (*x*-axis) in several species. (*G*) Median NFR width in each species correlates with abundance of antinucleosomal tracts in its genome. Shown is the median NFR width (Tsankov et al. 2010) (*x*-axis) for each species (◆) vs. the total number of poly(A) and poly(G) sequences of strength 2 or greater in that species (*y*-axis). Line represents the best linear fit. Species names are colored as in *A*.

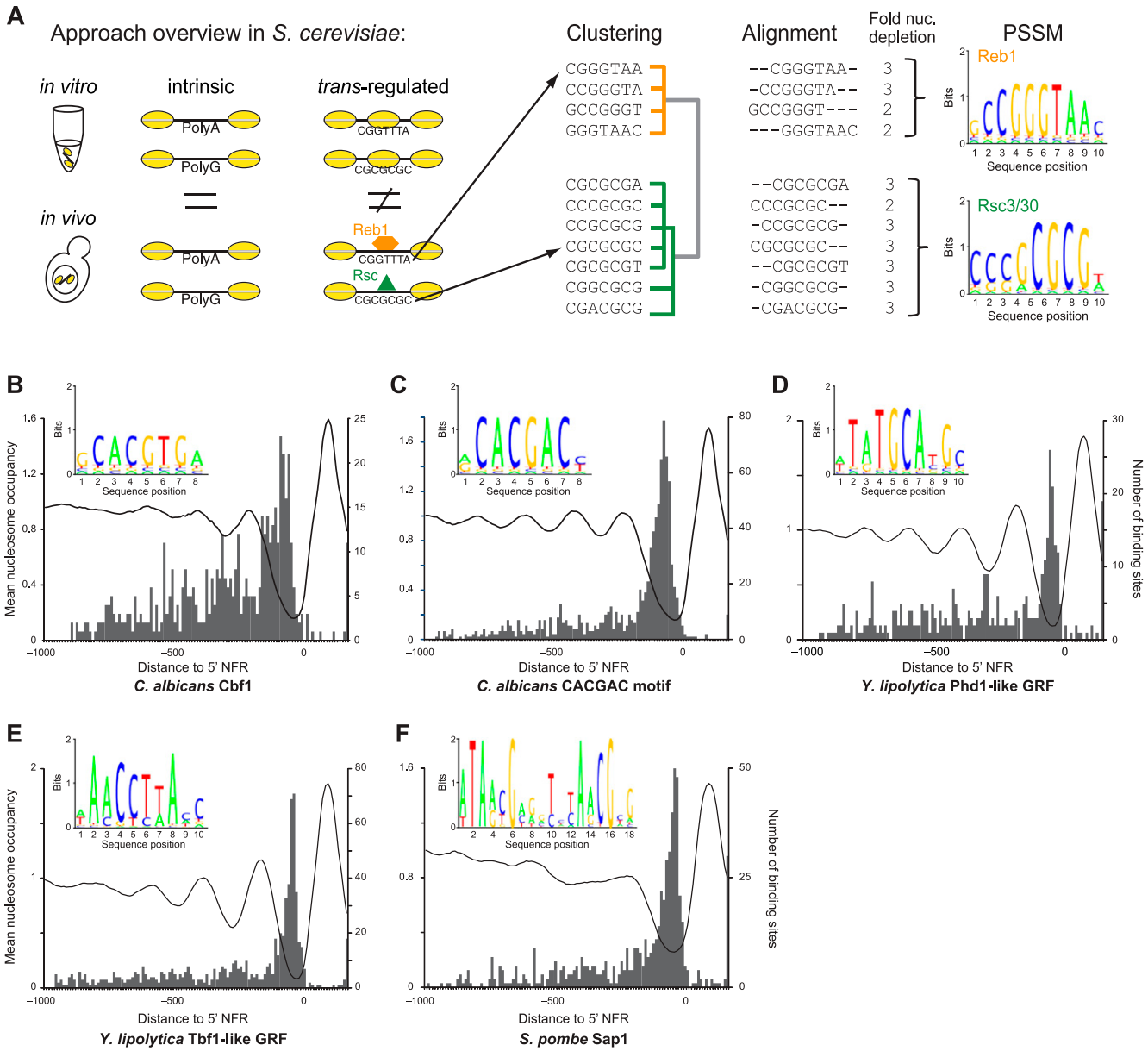

Figure 2. Identification of novel general regulatory factor (GRF) motifs. (*A*) Overview of GRF motif discovery approach. Nucleosome-depleted sequences are first classified as intrinsic (*left*) if they evict nucleosomes in vitro and in vivo, or as *trans*-regulated (*right*) if they evict nucleosomes more strongly in vivo. *Trans*-regulated nucleosome positioning sequences in each genome are then clustered based on similarity, aligned, and combined into a position specific scoring matrix (PSSM). Shown are the results for *S. cerevisiae*, where the algorithm outputs the known PSSMs of chromatin regulators Reb1 and Rsc3/30. (*B–E*) Predicted binding sites for GRFs in *C. albicans* (*B*,*C*), and *Y. lipolytica* (*D*,*E*). Shown are the sequence logos of the PSSMs (*insets*) for the sites learned by our approach from 7-mers depleted in each species. For *Y. lipolytica*, the names of *S. cerevisiae* proteins with similar sequence specificity are displayed. Each graph shows the average normalized nucleosome occupancy (*left y*-axis) in promoters with significant matches to the PSSM (black curve, aligned by a gene's +1 nucleosome), as well as the location (gray bars) and number (*right y*-axis) of binding-site locations. As observed with GRFs in other species (Tsankov et al. 2010), these binding sites are almost entirely (>89%) NFR-localized. (*F*) Known Sap1 motif (Ghazvini et al. 1995) is identified as a GRF site in *S. pombe*. As in *B–E* but for *S. pombe* data.

GRF and poly(A)s, as we find that the majority of poly(A) sequences that become more nucleosome depleted in GRF mutants are located at promoters without sites for the GRF in question (Supplemental Fig. 2; data not shown). This suggests a general antagonistic relationship between GRFs and intrinsic antinucleosomal sequences, reminiscent of the homeostatic maintenance of a fixed amount of accessible DNA observed in histone H1 mutants (Woodcock et al. 2006).

Predicted binding sites for novel GRFs

We next explored the two PSSMs in *Y. lipolytica*, TATGCAtG and AACCTt/aA, and the PSSM CACGAC in *C. albicans*, which are enriched in NFRs (>91%) (Fig. 2D,E). We have found no candidate binding protein for the *C. albicans* CACGAC motif (Fig. 2C). Conversely, we found that the two PSSMs in *Y. lipolytica* are similar to the known binding sites for *S. cerevisiae* proteins Phd1 (aTGCAtg) (Badis

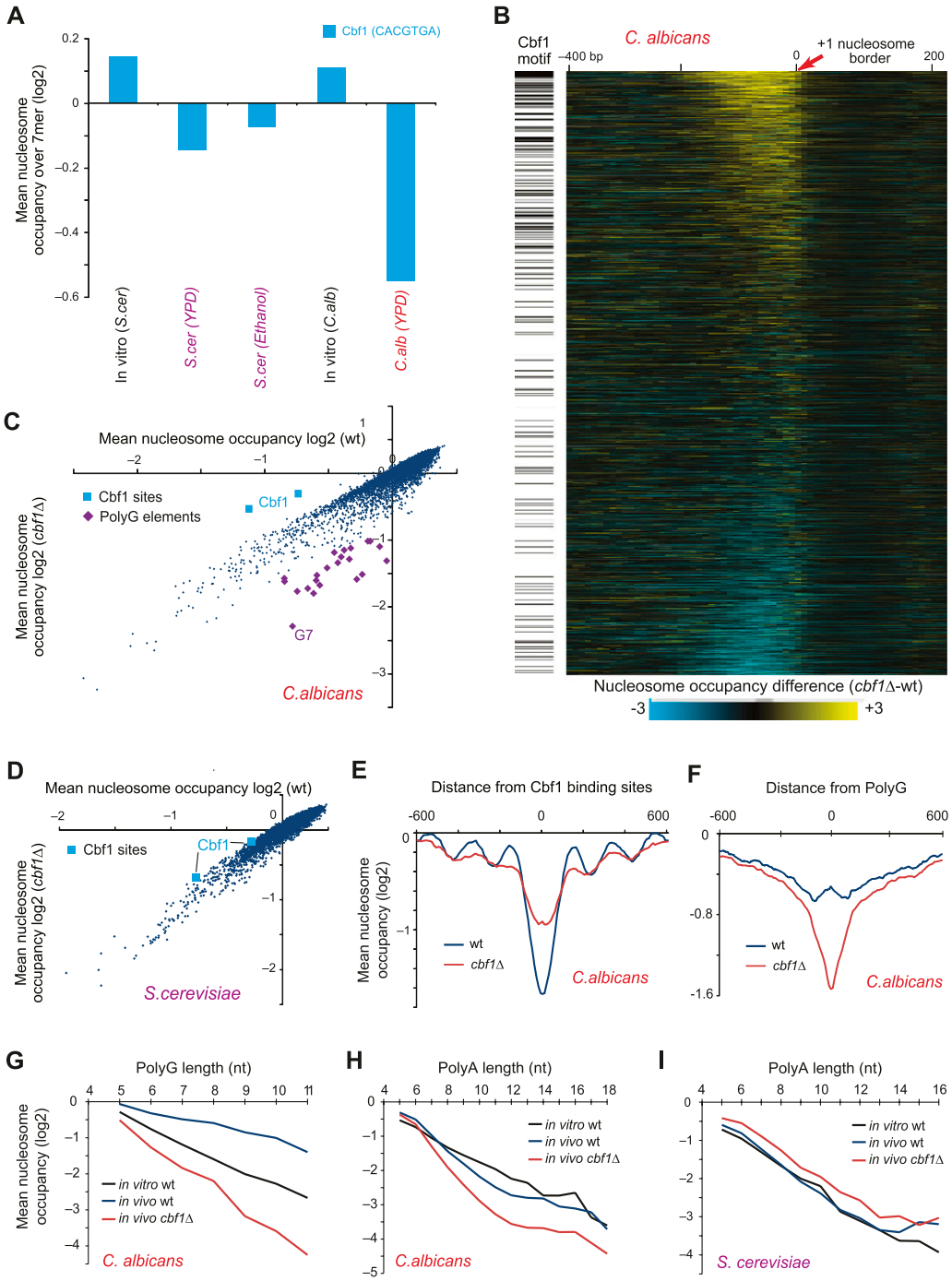

Figure 3. Cbf1 acts as a GRF in *C. albicans*. (*A*) Nucleosome depletion at the Cbf1 binding site in *C. albicans* is not a result of respiratory growth or genomic organization. Mean nucleosome occupancy at Cbf1 sites (*y*-axis) is shown for the indicated conditions (Kaplan et al. 2008; Field et al. 2009) and species. (*B*) Cbf1 deletion in *C. albicans* increases nucleosome occupancy at Cbf1 motifs. (*Left*) Genes with significant matches to the Cbf1-binding site (black). (*Right*) Difference in nucleosome abundance at each gene in *C. albicans* (rows) between *cbf1Δ* and wild-type strains. Genes are aligned by the +1 nucleosome/NFR boundary (0, red arrow) and ranked from gain (*top*, yellow) to loss (*bottom*, blue) in nucleosome occupancy over their NFR. (*C*) Cbf1-binding sites are the only 7-mers with increased mean nucleosome occupancy in *C. albicans cbf1Δ* strains compared with wild type. Shown is the mean nucleosome occupancy ($\log_2$) for each 7-mer in the wild-type (*x*-axis) and the *cbf1Δ* strain (*y*-axis). Cbf1 binding sites are indicated as blue squares, poly(G) sequences as purple diamonds. (*D*) As in *C*, but for *S. cerevisiae*. (*E*) Increased nucleosome occupancy in CACGTGA Cbf1 sites in *C. albicans* in the *cbf1Δ* strain. Shown is the average $\log_2$ nucleosome occupancy (*y*-axis) at all genes with a CACGTGA Cbf1 motif match in their promoter in wild-type (blue) and *cbf1Δ* (red) strains. Genes are aligned by the location of the CACGTGA Cbf1 motif (position 0 on the *x*-axis). (*F*) Poly(G) sequences are more nucleosome depleted in a *cbf1Δ* strain. Shown are average nucleosome occupancy values (*y*-axis) centered on all poly(G) elements of strength of 2 or greater (0 on the *x*-axis) for *cbf1Δ* (red) and wild-type (blue) strains in *C. albicans*. (*G–I*) Cbf1 deletion affects nucleosome occupancy in vivo at intrinsic poly(G) and poly(A) sequences in *C. albicans*, but not in *S. cerevisiae*. Shown are mean nucleosome occupancy levels ($\log_2$, *y*-axis) for poly(G) (*G*) and poly(A) (*H*,*I*) sequences of different length (*x*-axis) in *C. albicans* (*G*,*H*) and *S. cerevisiae* (*I*) for wild-type (blue), in vitro (black), and *cbf1Δ* (red) experiments.

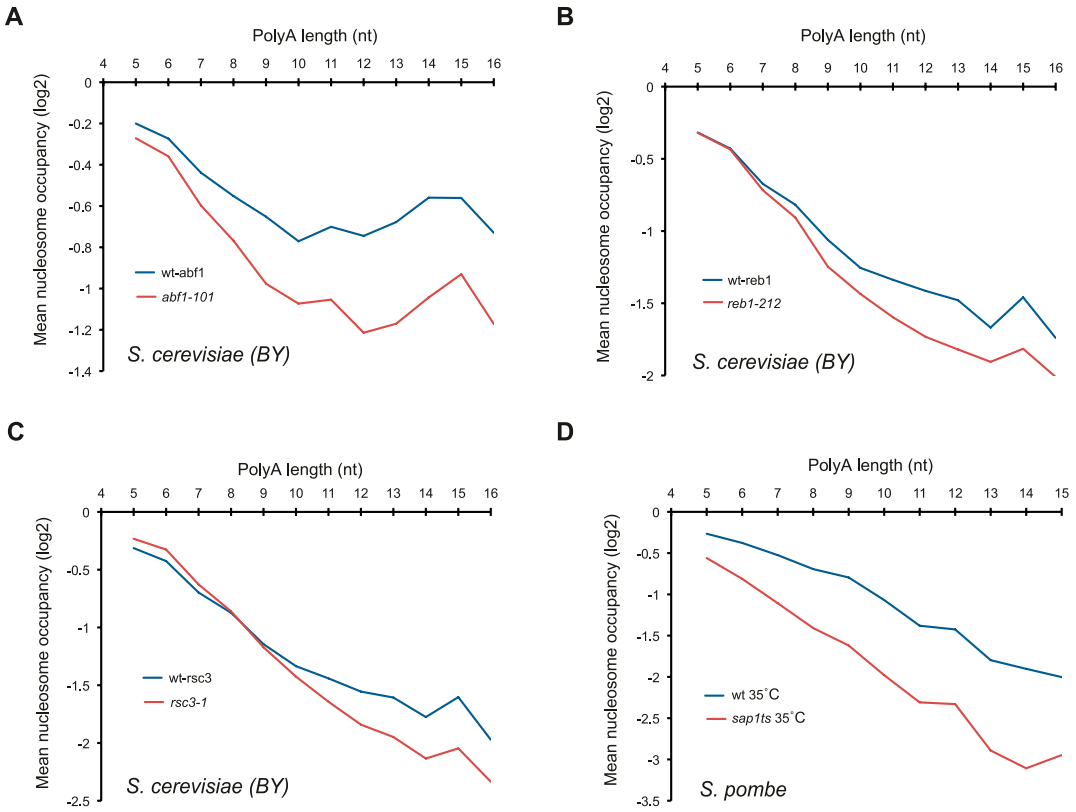

Figure 4. GRF deletion affects occupancy at intrinsic poly(*A*) sequences in *S. cerevisiae* and *S. pombe*. Shown are mean nucleosome occupancy levels ($\log_2$, *y*-axis) for poly(A) sequences of different length (*x*-axis) in wild-type cells (blue) and in red: (*A*) *abf1-101* strain in *S. cerevisiae*; (*B*) *reb1-212* strain in *S. cerevisiae*; (*C*) *rsc3-1* strain in *S. cerevisiae*; and (*D*) *sap1*^{ts} strain at restrictive temperature (35°C) in *S. pombe*.

et al. 2008; Zhu et al. 2009) and Tbf1 (AACCCTAa), both of which have orthologs in *Y. lipolytica* (Wapinski et al. 2007). Notably, although Tbf1 has been associated with the regulation of ribosomal protein genes in *Y. lipolytica* and other species (Hogues et al. 2008; Lavoie et al. 2010) (e.g., *C. albicans*, *D. hansenii*, *S. pombe*), its binding site is specifically nucleosome depleted in *Y. lipolytica* (Fig. 1B), suggesting an additional function in this species as a GRF.

Sap1 is a *trans*-regulator of chromatin in *S. pombe*

The predicted PSSM in *S. pombe* (Fig. 2F) corresponds to the DNA-binding site for Switch Activating Protein 1 (Sap1), a protein involved in DNA replication and mating-type switching (Ghazvini et al. 1995). Although Sap1 is an important, well-studied DNA-binding protein in *S. pombe,* its possible role as a GRF has not been previously reported, although prior nucleosome mapping identified the 5mer TAACG (contained within the Sap1 PSSM) as nucleosome depleted in vivo (Lantermann et al. 2010).

We experimentally validated the role of Sap1 in nucleosome eviction by mapping nucleosomes in a *sap1*^{ts} strain (Noguchi and Noguchi 2007) grown at a restrictive temperature (35°C). First, Sap1 sites are strongly enriched at those NFRs that become nucleosome occluded in a *sap1*^{ts} compared with a wild-type strain grown at restrictive temperatures (Fig. 5A). Second, nucleosome occupancy at Sap1 binding sites increases by 150% in the *sap1*^{ts} strain compared with wild-type (paired *t*-test $P < 10^{-300}$) (Fig. 5B), demonstrating Sap1's role in nucleosome eviction in vivo. These results were neither affected by changes in the MNase digestion level, nor of the increased temperature used to inactivate the temperature-sensitive mutant (Supplemental Fig. 3A). Third, two of the three 5-mers that are substantially more occluded at the restrictive temperature in the *sap1*^{ts} strain compared with wild-type correspond to different variants of the Sap1 half sites (TAACG and TAGCG) (Fig. 5C; Supplemental Fig. 3B). Interestingly, the third 5-mer is GGGGG, suggesting another coupling between poly(G)s and GRFs, but with an opposite behavior than that observed for Cbf1–poly(G) in *C. albicans*. Similar results are obtained when analyzing 7-mers (Supplemental Fig. 3C). Together, these results validate a causal role for the Sap1 protein in nucleosome eviction in vivo.

Discussion

Our study finds and validates remarkable plasticity in the intrinsic and *trans* determinants of nucleosome occupancy across species. We find that poly(G)s play a major role as global intrinsic anti-nucleosomal sequences in some species, and their effect has largely escaped attention in the literature due to their low abundance in the model organism *S. cerevisiae*. Poly(G) tracts were also nucleosome depleted in the nematode *C. elegans* (Supplemental Fig. 4A), suggesting that the intrinsic nucleosome depletion observed over poly(G)s extends to metazoans (Valouev et al. 2008).

Our combined experimental and computational approach provides a principled way to detect novel candidate GRFs (Supplemental Table 3) in any species based on that species' genome and a nucleosome occupancy map. This approach is also applicable to metazoans, where we have uncovered a likely binding site for an

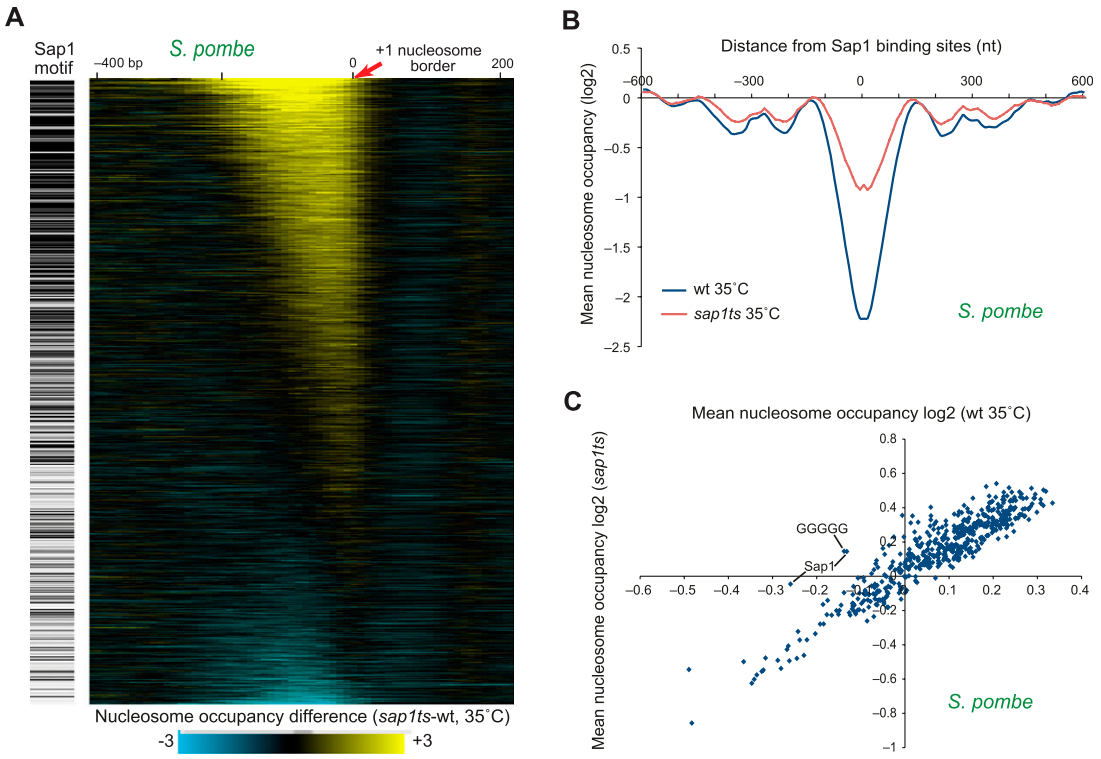

Figure 5. Sap1 acts as a GRF in *S. pombe*. (*A*) Sap1 deletion in *S. pombe* results in increased nucleosome occupancy at Sap1 motifs. (*Left*) Genes with significant matches to the Sap1 binding half-site (black). (*Right*) Difference in nucleosome abundance at each gene in *S. pombe* (rows) between $sap1^{ts}$ and wild-type strains (both at restrictive temperatures). Genes are aligned by the +1 nucleosome/NFR boundary (0, red arrow) and ranked from gain (*top*, yellow) to loss (*bottom*, blue) in nucleosome occupancy over their NFR. (*B*) Increased nucleosome occupancy at Sap1 half-sites in a $sap1^{ts}$ strain in *S. pombe*. Shown are $\log_2$ nucleosome occupancy averages at all genes with a significant Sap1 motif match in their upstream promoter for a $sap1^{ts}$ strain (red) and a wild-type strain (blue) grown in restrictive temperatures (35°C). Genes are aligned by the location of the Sap1 motif. Nucleosome occupancy increases over Sap1 sites is characteristic of GRF activity. (*C*) Increased nucleosome occupancy in 5-mers, reflecting the Sap1 half-sites in a $sap1^{ts}$ strain compared with wild-type (both at restrictive temperature, 35°C). Shown is the mean nucleosome occupancy ($\log_2$) for each 5-mer in the wild-type (*x*-axis) and the $sap1^{ts}$ strain (*y*-axis). Sap1 half-sites are labeled. The only additional site with increased occupancy is the intrinsic sequence GGGGG.

antinucleosomal GRF (Subirana and Messeguer 2010) in the nematode *C. elegans* (CGGCAAAT, Supplemental Fig. 4B). We experimentally validated several new GRFs that we predicted from our analyses. Most notably, we find that Cbf1 acts as a GRF through the CACGTG site (Fig. 3) in several pre-WGD species including *C. albicans*, but has largely lost this function in post-WGD species with the concurrent emergence of Reb1 as a GRF in these species. Inspection of protein sequences for orthologs with or without GRF activity yield no perfect correlates with GRF activity, although we note a strong trend for proteins with GRF activity to have low-complexity domains predicted to be "intrinsically disordered" (Fuxreiter et al. 2008), often rich in glutamine or asparagine.

What evolutionary pressures led to this coordinated change in GRF function from Cbf1 to Reb1? While many scenarios could affect the suite of transcription factors that function as GRFs, we note one interesting possibility here. Specifically, in *S. cerevisiae*, Cbf1 plays a role in sequence-directed establishment of the mononucleosomal point centromere. Conversely, species such as *S. pombe* and *C. albicans* have regional centromeres, consisting of longer stretches of multiple centromeric nucleosomes. While the nature of centromeres is not known for every species in this phylogeny, we note that most of the species where Cbf1 acts as a GRF use a regional centromere. Whatever the selective pressures that lead to the development of sequence-directed point centromeres (see Malik and Henikoff 2009), we imagine that the antinucleosomal effects of GRF activity could potentially be incompatible with the use of point centromeres for chromosome segregation, thus exerting pressure on Cbf1 to lose GRF activity.

We also discover that the well-studied protein Sap1 is the major *trans*-acting evictor of nucleosomes in *S. pombe*. This new role for Sap1 suggests that coupling of GRFs and mating loci regulation is a common feature in yeasts, seperated by over a billion years of evolution. However, it is important to note that Abf1 in *S. cerevisiae* and Reb1 in *K. lactis* both play roles in the establishment of silencing at the mating loci, whereas Sap1's role in *S. pombe* appears instead to be to activate mating type switching via its effects on DNA polymerase pausing (Ghazvini et al. 1995).

Future studies will be able to utilize orthologous proteins from those species whose GRF ensembles differ, to identify how related proteins can gain or lose the ability to evict nucleosomes. Together, our results demonstrate the evolutionary plasticity of the mapping from genomic sequence to chromatin architecture, and establish a toolbox for mechanistic dissection of GRF function.

Methods

Strains

We used the following wild-type strains in the study: *Saccharomyces cerevisiae*, BY4741, *Saccharomyces cerevisiae*, Sigma1278b L5366,

Saccharomyces paradoxus, NRRL Y-17217, *Saccharomyces mikatae*, IFO1815, *Saccharomyces bayanus*, NRRL Y-11845, *Candida glabrata*, CLIB 138, *Saccharomyces castellii*, NRRL Y-12630, *Kluyveromyces lactis*, CLIB 209, *Kluyveromyces waltii*, NCYC 2644, *Saccharomyces kluyveryii*, NRRL 12651, *Debaryomyces hansenii*, NCYC 2572, *Candida albicans*, SC 5314, *Yarrowia lipolytica*, CLIB 89, and *Schizosaccharomyces pombe*, 972h.

We used the following mutant strains in the study: *Candida albicans*, CBF1M5A (Biswas et al. 2003) (Δcbf1::FRT/Δcbf1::URA3), *Schizosaccharomyces pombe sap1*^{ts}, ENY1125 (h+ leu1-32 ura4-D18 sap1-48ts-3FLAG(kanMX6)) (Noguchi and Noguchi 2007).

Growth conditions

S. cerevisiae and *C. albicans cbf1*Δ strains were grown in our rich medium with the following formulation: Yeast extract (1.5%), Peptone (1%), Dextrose (2%), SC Amino Acid mix (Sunrise Science) 2 g/liter, Adenine 100 mg/L, Tryptophan 100 mg/L, Uracil 100 mg/L. This medium was designed to mitigate differences in growth rates between species, and was used for the 12 species in Tsankov et al. (2010). For the Sap1 experiments, wild-type *S. pombe* strains were grown in YES at 30°C (Forsburg and Rhind 2006) and also at a restrictive temperature of 35°C. The *sap1-ts* strain was grown in YES at a permissive temperature of 25°C and a restrictive temperature of 35°C.

Preparation of nucleosomal DNA and Illumina sequencing

S. cerevisiae and *C. albicans cbf1*Δ strains were handled as previously described (Tsankov et al. 2010). For the Sap1 experiments in *S. pombe*, overnight cultures for each species were grown to mid-log phase (~OD 0.75) in 250 mL of medium. Nucleosomal DNA isolation was carried out as previously described (Lantermann et al. 2010) with the following slight modifications. Cells were spheroplasted with zymolase and Novozym for 45 min. MNase digestion levels for all samples were uniformly chosen to contain a slightly visible trinucleosome band. Mononucleosomes were size selected on a gel and purified using BioRad Freeze-N-Squeeze tubes, followed by phenol-chloroform extraction. Selected DNA was prepared for sequencing using the standard Illumina protocol that includes blunt ending, adaptor ligation, PCR amplification, and final size selection plus gel purification (Shivaswamy et al. 2008; Tsankov et al. 2010; Weiner et al. 2010). Libraries were sequenced on an Illumina 1G Analyzer to generate 36-bp single-end reads.

Sequencing read alignment and data post-processing

As previously described (Tsankov et al. 2010), we used BLAT (Kent 2002) to map sequence reads from each experiment to the corresponding reference genome, keeping only reads that mapped to a unique location and allowing for up to four mismatches. Each uniquely mapped read was then extended to a length of 100 bp (extending by 147 bp does not affect our biological conclusions). To generate a genomic nucleosome occupancy landscape, we summed all extended reads covering each base pair. We then masked all repetitive regions along each track, defining repetitive regions as locations in the genome that cannot be uniquely defined by the length of a read (36 bp). Only the start position for the 36-bp repetitive region was masked on both the Watson and the Crick strand, indicating that we cannot identify nucleosome ends uniquely at that location in the genome. We also masked all regions of nucleosome occupancy greater than 10 times the mean occupancy to remove outlier effects that occur in places such as the rDNA locus. To normalize for sequencing depth for each genomic nucleosome track, we divided the occupancy at each location by the mean nucleosome occupancy per base pair. These normalized maps were used to generate the average nucleosome occupancy plots at gene promoters containing binding sites of different GRFs (Fig. 2).

Detection of nucleosome positions

To infer the location of nucleosomes from the data, we used a Parzen window approach similar to that previously described (Albert et al. 2007; Shivaswamy et al. 2008). Our modified approach (Tsankov et al. 2010) uses three parameters—the average DNA fragment length, the standard deviation of the Parzen window, and the maximum allowable overlap between nucleosomes. To estimate the mean DNA fragment length in each experiment, we shifted reads from one strand and then correlated them with the reads of the opposite strand. For each species, we observed a peak in the cross-correlation at a shift between 127 and 153 bp, which we used to estimate the mean DNA fragment length per experiment. We chose a standard deviation of the Parzen window of 30 bp for all species, since it closely matched the observed standard deviation around the cross-correlation peak of each experiment. Finally, we set the maximum allowable overlap between nucleosomes to 20 bp. We then shifted all read start locations by half of the mean DNA fragment length in the direction toward the dyad of the nucleosome they represent. Our approach places a normal distribution with a standard deviation of 30 bp at each read's shifted locations. Summing all individual curves for all loci leads to a smoothed probability landscape of nucleosome occupancy. We next identify all peaks along the landscape, which represent nucleosome centers. The algorithm then places nucleosomes along the genome in the order of decreasing peak heights (greedy approach) and iteratively masks out these regions to prevent more than a 20 bp overlap between nucleosomes.

Finding 5′ and 3′ NFRs

As previously described (Tsankov et al. 2010), we define 5′ and 3′ nucleosome-free regions (NFRs) as the linker DNA of "significant length" closest to the 5′ and 3′ end of each gene, respectively. To find NFRs, we first created a nucleosome call landscape for each genome, normalized for sequencing depth in the same manner as the nucleosome occupancy maps (above). NFR boundaries were often obscured by very low occupancy nucleosome calls. We therefore removed all nucleosome calls with occupancy <40% of the average nucleosome occupancy from the map. We searched for 5′ or 3′ NFRs within 1000 bases upstream/downstream of the 5′ or 3′ end of each gene, truncated when neighboring ORFs overlapped this region. We then defined an NFR as the linker DNA longer than 60 bp closest to the 5′ or 3′ end of each gene. If no linker longer than 60 bp was found in this search, we defined the NFR as the first linker from the 5′ or 3′ end. Our method was highly predictive of transcription start sites (TSSs) in *S. cerevisiae* (Xu et al. 2009)—the NFR boundary closest to the 5′ end of the gene was able to predict 84% of TSSs within 50 bp. Linker lengths of 50 or 70 bp and occupancy thresholds of 30% or 50% produced highly similar results (data not shown).

N-mer analysis

Prior to analysis, we $\log_2$-transformed the normalized nucleosome occupancy data (data post-processing, above), subtracted the mean, and divided by the standard deviation. Hence, the global nucleosome occupancy data for each species is approximately normal with zero mean and unit variance. We also used the same procedure for processing published in vitro data (Kaplan et al. 2008). These log-normalized maps were also used to generate the average

nucleosome occupancy plots comparing wild-type and mutant strains for Cbf1 and Sap1, and for characterizing poly(A)s and poly(G)s.

For each N-mer, we define the in vivo depletion score as the mean $-\log_2$ normalized nucleosome occupancy across all instances, and all instances of the reverse complement. We also defined the depletion score relative to in vitro as power 2 of the difference between the in vivo depletion scores in each species and the in vitro depletion scores in *S. cerevisiae* (also repeated for in vitro data from *C. albicans*) (Field et al. 2009). The analysis was done for n = 5, 6, 7, and 8.

To quantify the variation of 7-mers across species, we first standardized the depletion scores to have a zero mean and unit variance per species. Histograms of 7-mer depletion scores for all species closely resembled a normal distribution (data not shown). We then measured the variance for each 7-mer across all 14 strains displayed in Figure 1B. For 14 random samples from independent and identically distributed (IID) normal distributions, a variance of <0.5 is expected to occur only 7% of the time. In contrast, we observed that 89% of 7-mers have a variance <0.5. To estimate the significance of this finding, we used a Binomial RV with the success rate parameter p equal to 0.07 and number of trials n equal to 8192 possible nonredundant 7-mers. To calculate a *P*-value, we measured the probability that 7254 7-mers (89%) or more had a variance of <0.5 by summing the right tail of the distribution that is ≥7254 successes.

Characterizing poly(dAT) and poly(dCG) tracts

To annotate all poly(dAT) tracts in each species and determine their nucleosome repelling strength we used an approach similar to a previously described one (Field et al. 2008). For each species' genome, we annotated all poly(A) or poly(T) tracts of length L of 5 bp or more. We define the depletion score for a tract of length L as the mean of the $-\log_2$ normalized nucleosome occupancy across all instances of that length. This was calculated both using in vitro data from *C. albicans* (Field et al. 2009) and the in vivo data from each species. For long poly(dAT) tracts with very few occurrences in a given genome, we noticed variability in the depletion score, likely due to small sample size. To mitigate this problem, we fit a line for depletion scores versus L using a weighted linear least squares fit with weights proportional to the number of occurrences for tracts of length L. We then used the line as an estimate for long tracts with fewer than 100 occurrences in a given genome. We iterated this procedure for all maximal poly(dAT) tracts with k allowed mismatches, k = 1,. . .,20. The depletion score increases linearly with L for tracts with different k, confirming that a linear fit is appropriate (Tsankov et al. 2010).

To aggregate all nonoverlapping poly(dAT) tracts within a given genome, we first discretized the strengths for each L. We define the fold depletion score of all tracts of length L as power 2 of the depletion score. We then quantized all poly(dAT) tract fold depletion scores to the highest fold depletion level exceeding 2, 4, 8, 16, and 32. For example, a tract with a depletion score of 3.5 is $2^{3.5}$ = 11.3-fold depleted in nucleosomes relative to average, and would be assigned a fold depletion score of 8. We next iterated over all poly(dAT) tracts with mismatches k = 0,. . .,20, replacing overlapping tracts only if the tract with more mismatches had a higher quantized fold depletion score. To annotate poly(dCG) tracts, we used the same approach as above, but now treating consecutive sequences of Cs and Gs as homopolymeric tracts and other, interrupting nucleotides as mismatches.

Our technique makes three slight modifications to a previously published method (Field et al. 2008). First, we used in vitro data from *C. albicans* (Field et al. 2009), instead of in vivo data from *S. cerevisiae*, allowing us to more accurately estimate the intrinsic nucleosome repelling strength of poly(dAT) elements and to characterize the intrinsic antinucleosomal properties of poly(dCG) tracts, which are very rare in the *S. cerevisiae* genome. Second, we calculated nucleosome depletion in the log domain, since depletion scales log-linearly with the length L of poly(dAT) tracts, and log-linear scaling allows us to estimate the in vitro depletion of rare homopolymers.

K-mer clustering and motif discovery

To discover GRF motifs in different species, we first restricted our search to all N-mers that had a depletion score relative to in vitro data that is greater than 2 (compared with both *S. cerevisae* and *C. albicans* in vitro data). We also included the top 20 7-mers for species with <20 7-mers that passed this threshold. To construct PSSMs, we first calculated a similarity matrix between all possible pairs of these 7-mers. The similarity was measured using the dot product of the best possible ungapped alignment between two 7-mers, allowing for reverse complements. We defined the similarity between two 7-mers $\vec{x}$ and $\vec{y}$ as:

$$S(\vec{x},\vec{y}) = M - \frac{1}{4}L,$$

where L is the length of the alignment and M is the number of matches. To group similar 7-mers, the similarity was then converted to a distance by subtracting $S(\vec{x},\vec{y})$from the self-similarity, $S(\vec{x},\vec{x})$, as follows:

$$d(\vec{x},\vec{y}) = S(\vec{x},\vec{x}) - S(\vec{x},\vec{y}).$$

The distances between all 7-mers were then clustered using average-linkage hierarchical clustering. For all species, we grouped subtrees of 7-mers into clusters by visual inspection, allowing for at most one alignment error or mismatch between the two most similar 7-mers in a cluster. Heuristically, a distance for cutting trees was around 2. Clusters of less than three elements were removed from consideration.

We then performed progressive multiple alignment for all 7-mers within each cluster. We used the NUC44 scoring matrix and computed the average score for two matched residues (S_m). Opening gaps within 7-mers was not allowed. Gaps flanking the 7-mers were penalized as $S_m/3$, as it produced a good tradeoff for penalizing mismatches between two residues versus one residue and a terminal gap.

To form PSSMs, letters in each position of the alignment were summed, weighted by their depletion score relative to in vitro. Therefore, 7-mers with a higher depletion score contributed more to the PSSM. To prevent overfitting, we inserted pseudocounts of 0.5 for each entry in the PSSM, equivalent to adding an extra, non-informational 7-mer with a depletion score relative to in vitro of 2.

Promoter TF motif scanning

Promoter sequences for each gene were defined as 1000 bases upstream, truncated when neighboring ORFs overlapped with this region. To find the location of binding sites for the new GRFs (in *C. albicans*, *Y. lipolytica*, *S. pombe,* and other species) we identified motif targets via the TestMOTIF software program (Barash et al. 2005) using a 3-order Markov background model estimated from the entire set of promoters per genome and the PSSMs for each candidate GRF (below). We considered all motif instances with a *P*-value <0.05 as significant, and limited the number of promoters with significant sites to the top 1000. This upper bound was chosen

to exceed the maximal number of promoters bound (866, $P < 0.05$) by any transcription factor in *S. cerevisiae*, as measured by ChIP–chip (Harbison et al. 2004).

Scoring motif occurrences in NFRs

To measure the affinity for GRF sites at NFRs of the wild-type and mutant strains, we conducted the following analysis (as previously described) (Tsankov et al. 2010). We represent each motif of length L by a position-specific scoring matrix (PSSM) P, or the probability distribution $P(S_1,\ldots, S_L)$ of that motif occurring over any sequence $S_1\ldots S_L$. This is a standard approximation to a factors binding energy for sequence $S_1\ldots S_L$. We also learned the 0th-order Markov background probability distribution $B(S_1,\ldots, S_L)$ for each sequence $S_1\ldots S_L$, set to the frequency of the four nucleotides in the promoter regions of a given species. We calculate $A(P,S)$, a motif's affinity score for an NFR sequence S, by summing the contributions of $P(S_1,\ldots, S_L)/B(S_1,\ldots, S_L)$ over all allowable positions k in S as follows:

$$A(P,S)=\sum_k \frac{P(S_k,\ldots,S_{k+L-1})}{B(S_k,\ldots,S_{k+L-1})}=\sum_k\prod_{j=1}^{L}\frac{p(S_{k+j-1},j)}{b(S_{k+j-1})}.$$

Here, $b(S_{k+j-1})$ is the background probability of the nucleotide S_{k+j-1} of sequence S, and $p(S_{k+j-1},j)$ is the probability for nucleotide S_{k+j-1} in position j of the motif's PSSM. For the results in this study, we combined the contributions of both forward and reverse strands of each NFR. Also, normalizing the affinity by the length of each NFR sequence did not affect our results significantly.

Data access

Published nucleosome data for the 12 species in Tsankov et al. (2010) is available at the NCBI Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo) under accession no. GSE21960. Nucleosome data for *S. cerevisiae* and *C. albicans cbf1*D mutants, and for *S. pombe* wild-type and *sap1*^{ts} strains, have been deposited at GEO, accession no. GSE28839.

Acknowledgments

We thank Ido Amit and Nadia Tsankova for comments on this manuscript, Leslie Gaffney for help with figures, H. Lavoie and M. Whiteway for the gift of the *C. albicans cbf1Δ*, J. Mellor for the gift of the *S. cerevisiae cbf1Δ*, Eishi Noguchi for the gift of the *S. pombe sap1*^{ts}, and Nir Friedman for many helpful discussions. A.T. was supported by the NSF Graduate Research Fellowship. Work was supported by the Human Frontiers Science Program, the Howard Hughes Medical Institute, a Career Award at the Scientific Interface from the Burroughs Wellcome Fund, an NIH PIONEER award, the Broad Institute, and a Sloan Fellowship (A.R.), by American Cancer Society grant GMC-113639 (N.R.), and by NIGMS grant GM079205 and the Burroughs Wellcome Fund (O.J.R.).

References

Albert I, Mavrich TN, Tomsho LP, Qi J, Zanton SJ, Schuster SC, Pugh BF. 2007. Translational and rotational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome. *Nature* **446:** 572–576.

Badis G, Chan ET, van Bakel H, Pena-Castillo L, Tillo D, Tsui K, Carlson CD, Gossett AJ, Hasinoff MJ, Warren CL, et al. 2008. A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. *Mol Cell* **32:** 878–887.

Barash Y, Elidan G, Kaplan T, Friedman N. 2005. CIS: compound importance sampling method for protein-DNA binding site p-value estimation. *Bioinformatics* **21:** 596–600.

Biswas K, Rieger KJ, Morschhauser J. 2003. Functional characterization of CaCBF1, the *Candida albicans* homolog of centromere binding factor 1. *Gene* **323:** 43–55.

Clapier CR, Cairns BR. 2009. The biology of chromatin remodeling complexes. *Annu Rev Biochem* **78:** 273–304.

Conant GC, Wolfe KH. 2007. Increased glycolytic flux as an outcome of whole-genome duplication in yeast. *Mol Syst Biol* **3:** 129. doi: 10.1038/msb 4100270.

Drew HR, Travers AA. 1985. DNA bending and its relation to nucleosome positioning. *J Mol Biol* **186:** 773–790.

Field Y, Kaplan N, Fondufe-Mittendorf Y, Moore IK, Sharon E, Lubling Y, Widom J, Segal E. 2008. Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. *PLoS Comput Biol* **4:** e1000216. doi: 10.1371/journal.pcbi.1000216.

Field Y, Fondufe-Mittendorf Y, Moore IK, Mieczkowski P, Kaplan N, Lubling Y, Lieb JD, Widom J, Segal E. 2009. Gene expression divergence in yeast is coupled to evolution of DNA-encoded nucleosome organization. *Nat Genet* **41:** 438–445.

Forsburg SL, Rhind N. 2006. Basic methods for fission yeast. *Yeast* **23:** 173–183.

Fuxreiter M, Tompa P, Simon I, Uversky VN, Hansen JC, Asturias FJ. 2008. Malleable machines take shape in eukaryotic transcriptional regulation. *Nat Chem Biol* **4:** 728–737.

Ganapathi M, Palumbo MJ, Ansari SA, He Q, Tsui K, Nislow C, Morse RH. 2010. Extensive role of the general regulatory factors, Abf1 and Rap1, in determining genome-wide chromatin structure in budding yeast. *Nucleic Acids Res* **39:** 2032–2044.

Ghazvini M, Ribes V, Arcangioli B. 1995. The essential DNA-binding protein sap1 of *Schizosaccharomyces pombe* contains two independent oligomerization interfaces that dictate the relative orientation of the DNA-binding domain. *Mol Cell Biol* **15:** 4939–4946.

Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, et al. 2004. Transcriptional regulatory code of a eukaryotic genome. *Nature* **431:** 99–104.

Hartley PD, Madhani HD. 2009. Mechanisms that specify promoter nucleosome location and identity. *Cell* **137:** 445–458.

Hogues H, Lavoie H, Sellam A, Mangos M, Roemer T, Purisima E, Nantel A, Whiteway M. 2008. Transcription factor substitution during the evolution of fungal ribosome regulation. *Mol Cell* **29:** 552–562.

Ioshikhes IP, Albert I, Zanton SJ, Pugh BF. 2006. Nucleosome positions predicted through comparative genomics. *Nat Genet* **38:** 1210–1215.

Iyer V, Struhl K. 1995. Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. *EMBO J* **14:** 2570–2579.

Jiang C, Pugh BF. 2009. Nucleosome positioning and gene regulation: advances through genomics. *Nat Rev Genet* **10:** 161–172.

Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, Leproust EM, Hughes TR, Lieb JD, Widom J, et al. 2008. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458:** 362–366.

Kent WJ. 2002. BLAT–the BLAST-like alignment tool. *Genome Res* **12:** 656–664.

Kornberg R. 1981. The location of nucleosomes in chromatin: specific or statistical. *Nature* **292:** 579–580.

Kornberg RD, Lorch Y. 1999. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* **98:** 285–294.

Kornberg RD, Stryer L. 1988. Statistical distributions of nucleosomes: nonrandom locations by a stochastic mechanism. *Nucleic Acids Res* **16:** 6677–6690.

Lantermann AB, Straub T, Stralfors A, Yuan GC, Ekwall K, Korber P. 2010. *Schizosaccharomyces pombe* genome-wide nucleosome mapping reveals positioning mechanisms distinct from those of *Saccharomyces cerevisiae*. *Nat Struct Mol Biol* **17:** 251–257.

Lascaris RF, Groot E, Hoen PB, Mager WH, Planta RJ. 2000. Different roles for abf1p and a T-rich promoter element in nucleosome organization of the yeast RPS28A gene. *Nucleic Acids Res* **28:** 1390–1396.

Lavoie H, Hogues H, Mallick J, Sellam A, Nantel A, Whiteway M. 2010. Evolutionary tinkering with conserved components of a transcriptional regulatory network. *PLoS Biol* **8:** e1000329. doi: 10.1371/journal.pbio.1000329.

Malik HS, Henikoff S. 2009. Major evolutionary transitions in centromere complexity. *Cell* **138:** 1067–1082.

Mavrich TN, Ioshikhes IP, Venters BJ, Jiang C, Tomsho LP, Qi J, Schuster SC, Albert I, Pugh BF. 2008. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res* **18:** 1073–1083.

Mobius W, Gerland U. 2010. Quantitative test of the barrier nucleosome model for statistical positioning of nucleosomes up- and downstream of transcription start sites. *PLoS Comput Biol* **6:** doi: 10.1371/journal.pcbi.1000891.

Noguchi C, Noguchi E. 2007. Sap1 promotes the association of the replication fork protection complex with chromatin and is involved in the replication checkpoint in *Schizosaccharomyces pombe*. *Genetics* **175:** 553–566.
Peckham HE, Thurman RE, Fu Y, Stamatoyannopoulos JA, Noble WS, Struhl K, Weng Z. 2007. Nucleosome positioning signals in genomic DNA. *Genome Res* **17:** 1170–1177.
Radman-Livaja M, Rando OJ. 2010. Nucleosome positioning: how is it established, and why does it matter? *Dev Biol* **339:** 258–266.
Raisner RM, Hartley PD, Meneghini MD, Bao MZ, Liu CL, Schreiber SL, Rando OJ, Madhani HD. 2005. Histone variant H2A.Z marks the 5′ ends of both active and inactive genes in euchromatin. *Cell* **123:** 233–248.
Rando OJ, Ahmad K. 2007. Rules and regulation in the primary structure of chromatin. *Curr Opin Cell Biol* **19:** 250–256.
Rando OJ, Chang HY. 2009. Genome-wide views of chromatin structure. *Annu Rev Biochem* **78:** 245–271.
Segal E, Widom J. 2009. What controls nucleosome positions? *Trends Genet* **25:** 335–343.
Sekinger EA, Moqtaderi Z, Struhl K. 2005. Intrinsic histone-DNA interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast. *Mol Cell* **18:** 735–748.
Shivaswamy S, Bhinge A, Zhao Y, Jones S, Hirst M, Iyer VR. 2008. Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. *PLoS Biol* **6:** e65. doi: 10.1371/journal.pbio.0060065.
Subirana JA, Messeguer X. 2010. The most frequent short sequences in non-coding DNA. *Nucleic Acids Res* **38:** 1172–1181.
Tillo D, Hughes TR. 2009. G+C content dominates intrinsic nucleosome occupancy. *BMC Bioinformatics* **10:** 442. doi: 10.1186/1471-2105-10-442.
Tsankov AM, Thompson DA, Socha A, Regev A, Rando OJ. 2010. The role of nucleosome positioning in the evolution of gene regulation. *PLoS Biol* **8:** e1000414. doi: 10.1371/journal.pbio.1000414.
Vaillant C, Palmeira L, Chevereau G, Audit B, d'Aubenton-Carafa Y, Thermes C, Arneodo A. 2010. A novel strategy of transcription regulation by intragenic nucleosome ordering. *Genome Res* **20:** 59–67.
Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, Peckham H, Zeng K, Malek JA, Costa G, McKernan K, et al. 2008. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome Res* **18:** 1051–1063.
Wapinski I, Pfeffer A, Friedman N, Regev A. 2007. Natural history and evolutionary principles of gene duplication in fungi. *Nature* **449:** 54–61.
Weiner A, Hughes A, Yassour M, Rando OJ, Friedman N. 2010. High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. *Genome Res* **20:** 90–100.
Whitehouse I, Rando OJ, Delrow J, Tsukiyama T. 2007. Chromatin remodelling at promoters suppresses antisense transcription. *Nature* **450:** 1031–1035.
Woodcock CL, Skoultchi AI, Fan Y. 2006. Role of linker histone in chromatin structure and function: H1 stoichiometry and nucleosome repeat length. *Chromosome Res* **14:** 17–25.
Xu Z, Wei W, Gagneur J, Perocchi F, Clauder-Munster S, Camblong J, Guffanti E, Stutz F, Huber W, Steinmetz LM. 2009. Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457:** 1033–1037.
Yarragudi A, Miyake T, Li R, Morse RH. 2004. Comparison of ABF1 and RAP1 in chromatin opening and transactivator potentiation in the budding yeast *Saccharomyces cerevisiae*. *Mol Cell Biol* **24:** 9152–9164.
Yarragudi A, Parfrey LW, Morse RH. 2007. Genome-wide analysis of transcriptional dependence and probable target sites for Abf1 and Rap1 in *Saccharomyces cerevisiae*. *Nucleic Acids Res* **35:** 193–202.
Yu L, Morse RH. 1999. Chromatin opening and transactivator potentiation by RAP1 in *Saccharomyces cerevisiae*. *Mol Cell Biol* **19:** 5279–5288.
Yuan GC, Liu JS. 2008. Genomic sequence is highly predictive of local nucleosome depletion. *PLoS Comput Biol* **4:** e13. doi: 10.1371/journal.pcbi.0040013.
Yuan GC, Liu YJ, Dion MF, Slack MD, Wu LF, Altschuler SJ, Rando OJ. 2005. Genome-scale identification of nucleosome positions in S. cerevisiae. *Science* **309:** 626–630.
Zhang Y, Moqtaderi Z, Rattner BP, Euskirchen G, Snyder M, Kadonaga JT, Liu XS, Struhl K. 2009. Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo. *Nat Struct Mol Biol* **16:** 847–852.
Zhu C, Byers KJ, McCord RP, Shi Z, Berger MF, Newburger DE, Saulrieta K, Smith Z, Shah MV, Radhakrishnan M, et al. 2009. High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res* **19:** 556–566.

Received February 15, 2011; accepted in revised form July 5, 2011.