Title (88 chars < 120): Punctuated evolution and transitional hybrid network in an ancestral cell cycle of fungi

Authors: Edgar M. Medina[1,2], Jonathan J. Turner[3], Raluca Gordân[2,4], Jan M. Skotheim[3], and Nicolas E. Buchler[1,2]

[1] Department of Biology, Duke University, Durham, NC, 27708, USA
[2] Center for Genomic and Computational Biology, Duke University, Durham, NC, 27708, USA
[3] Department of Biology, Stanford University, Stanford, CA, 94305, USA
[4] Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, 27710, USA

Contact: nicolas.buchler@duke.edu

Abstract: (134 words < 150)

Although cell cycle control is an ancient, conserved, and essential process, some core animal and fungal cell cycle regulators share no more sequence identity than non-homologous proteins. Here, we show that evolution along the fungal lineage was punctuated by the early acquisition and entrainment of the SBF transcription factor through horizontal gene transfer. Cell cycle evolution in the fungal ancestor then proceeded through a hybrid network containing both SBF and its ancestral animal counterpart E2F, which is still maintained in many basal fungi. We hypothesize that a virally-derived SBF may have initially hijacked cell cycle control by activating transcription via the *cis*-regulatory elements targeted by the ancestral cell cycle regulator E2F, much like extant viral oncogenes. Consistent with this hypothesis, we show that SBF can regulate promoters with E2F binding sites in budding yeast.

Impact statement (30 words): Cell cycle network evolution in a fungal ancestor was punctuated by the arrival of a viral DNA-binding protein that was incorporated into the G1/S regulatory network controlling cell cycle entry.

Keywords: cell cycle / evolution / fungi / chytrid / virus / horizontal gene transfer

INTRODUCTION

The networks regulating cell division in yeasts and animals are highly similar in both physiological function and network structure (Figure 1) (Cross et al., 2011; Doonan and Kitsios, 2009). For example, the cell cycle controls proliferation in response to a variety of internal and external signals during the G1 phase, between cell division and DNA replication. These input signals, including cell growth, are integrated into a gradual increase in cyclin dependent kinase (Cdk) activity, which triggers a feedback loop at the basis of the all-or-none irreversible decision to proliferate (Bertoli et al., 2013).

Many of the molecular mechanisms underlying G1 regulation are highly conserved. In animal cells, Cyclin D, in complex with either Cdk4 or Cdk6, initiates cell cycle entry by phosphorylating the retinoblastoma protein, pRb. This begins the inactivation of pRb and the concomitant activation of the E2F transcription factors that induce transcription of downstream cyclins E and A, which complete the inhibition of pRb thereby forming a positive feedback loop (Bertoli et al., 2013). Similarly, in budding yeast, the G1 cyclin Cln3-Cdk1 complex initiates the transition by phosphorylating and partially inactivating Whi5, an inhibitor of the SBF transcription factor (Costanzo et al., 2004; de Bruin et al., 2004; Nasmyth and Dirick, 1991; Ogas et al., 1991; Sidorova and Breeden, 1993). This allows for SBF-dependent transcription of the downstream G1 cyclins *CLN1* and *CLN2*, which also inactivate Whi5 to complete a positive feedback loop (Skotheim et al., 2008). Thus, both the biochemical function of G1 regulators and their specific targets are highly conserved (Figure 1).

Many of the individual proteins performing identical roles are unlikely to be true orthologs, *i.e.*, it cannot be inferred from sequence identity that the proteins evolved from a common ancestral gene. In yeast, a single cyclin-dependent kinase, Cdk1, binds distinct cyclin partners to perform all the functions of three non-orthologous animal Cdks (Cdk2, 4 and 6) during cell cycle entry (Liu and Kipreos, 2000). Furthermore, no member of the transcription factor complex SBF-Whi5 exhibits amino acid sequence identity or structural similarity to any member of the E2F-pRb complex (Cross et al., 2011; Hasan et al., 2013; Taylor et al., 1997). Finally, Cdk inhibitors such as Sic1 and p27 play analogous roles in yeast and mammals despite a

total lack of sequence identity (Cross et al., 2011). Taken together, these examples imply significant evolution of cell cycle regulatory proteins in fungi and/or animals while the network topology remains largely intact. Although identification of network topology is restricted to a few model organisms and is not as broad as sequence analysis, the similar network topology in budding yeast and animals suggests that this feature is more conserved than the constituent regulatory proteins (Cross et al., 2011; Doonan and Kitsios, 2009).

The shared presence of E2F-pRb within plants (Archaeaplastida) and animal (Metazoa) lineages would suggest that this regulatory complex, rather than the fungal SBF-Whi5 complex, was present in the last eukaryotic common ancestor (Cao et al., 2010; Doonan and Kitsios, 2009; Fang et al., 2006; Harashima et al., 2013). The divergence of G1 regulator sequences is surprising because fungi and animals are more closely related to one another than either is to plants. This fungal-metazoan difference raises the question as to where the fungal components came from. Fungal components could either be rapidly evolved ancestral regulators or have a distinct evolutionary history, which would suggest convergent evolution of regulatory networks.

To address this question, we examine conserved and divergent features of eukaryotic cell cycle regulation. In contrast to previous work that considered a protein family of cell cycle regulators in isolation (Cao et al., 2010; 2014; Eme et al., 2011; Gunbin et al., 2011; Ma et al., 2013; Wang et al., 2004), we studied the evolutionary history of an entire regulatory network across hundreds of species. We examined a greater number of genomes covering most of eukaryotic diversity, including Excavata, Haptophyta, Cryptophyta, SAR (Stramenopiles, Alveolata, Rhizaria), Archaeplastida (plants), Amoebozoa, Apusozoa and the Opisthokonta (animals and fungi). This survey allowed us to estimate the cell cycle repertoire of the last eukaryotic common ancestor (LECA), a prerequisite to clarifying the evolutionary transitions of the cell cycle components of both animals and fungi.

Our results indicate that LECA likely had complex cell cycle regulation involving at least one Cdk, multiple cyclin families, activating and inhibitory E2F transcription factors, and pRb-family pocket proteins. Identifying the LECA repertoire helps establish that the emergence of SBF-Whi5 is abrupt and distinguishes fungi from all

other eukaryotes. We show that basal fungi can have both ancestral E2F-pRb and fungal SBF-Whi5 components. Thus, fungal evolution appears to have proceeded through a hybrid network before abruptly losing the ancestral components in the lineage leading to Dikarya. This supports the hypothesis that network structure, rather than the individual components, has been conserved through the transition to fungi and argues against the case of convergent evolution.

Our data confirm that SBF shows homology to KilA-N, a poorly characterized domain present in prokaryotic and eukaryotic DNA viruses. Thus, SBF is not derived from E2F (its functional analog) and likely emerged through horizontal gene transfer in the fungal ancestor. We show that SBF can regulate promoters with E2F binding sites in budding yeast. We then then use high-throughput *in vitro* binding assay data to elucidate the shared nucleotide preferences of E2F and SBF for DNA binding. These data suggest that a viral SBF may have initially hijacked cell cycle control, activating transcription via the *cis*-regulatory elements targeted by the ancestral cell cycle regulator E2F, much like extant viral oncogenes.

RESULTS

Reconstruction of complex cell cycle control in the ancestral eukaryote

Recent work shows that the last eukaryotic common ancestor (LECA) already had a complex repertoire of protein families (Dacks and Field, 2007; Eichinger et al., 2005; Merchant et al., 2007). Indeed, all sequenced eukaryotic lineages have lost entire gene families that were present in LECA (Fritz-Laylin et al., 2010). In contrast to the growing consensus that LECA had an extensive repertoire of proteins, the prevailing view of the cell cycle in LECA is that it was based on a simple oscillator constructed with relatively few components (Coudreuse and Nurse, 2010; Nasmyth, 1995). According to the 'simple' LECA cell cycle model, an ancestral oscillation in Cyclin B-Cdk1 activity drove periodic DNA replication and DNA segregation, while other aspects of cell cycle regulation, such as G1 control, may have subsequently evolved in specific lineages. The model was motivated by the fact that Cdk activity of a single Cyclin B is sufficient to drive embryonic cell cycles in frogs (Murray and Kirschner, 1989) and fission yeast (Stern and Nurse, 1996), and that many yeast G1 regulators have no eukaryotic orthologs (Figure 1).

To determine the complexity of LECA cell cycle regulation, we examined hundreds of diverse eukaryotic genomes (Fungi, Amoebozoa, Excavata, SAR, Haptophyta, Cryptophyta). We first built sensitive profile Hidden Markov Models (Eddy, 2011) for each of the gene families of cell cycle regulators from model organisms *Arabidopsis thaliana, Homo sapiens, Schizosaccharomyces pombe,* and *Saccharomyces cerevisiae.* These HHMs were then used to query the sequenced eukaryotic genomes for homologs of both fungal and animal cell cycle regulators (Figures 2-3; see Methods). Phylogenetic analyses were performed on the detected homologs for accurate sub-family assignment of the regulators and inference of their evolutionary history (see Methods). If LECA regulation were simple, we would expect little conservation beyond the Cyclin B-Cdk1 mitotic regulatory module. However, if LECA regulation were more complex, we would expect to see broad conservation of a wider variety of regulators.

While we did not find either of the fungal regulators (SBF and Whi5) outside of Fungi, we did find proteins that regulate the animal cell cycle in Archaeplastida, Amoebozoa, SAR, Haptophyta, Cryptophyta, Excavata or Metazoa,. For example, the cyclin sub-families (A, B, D, and E) known to regulate the cell cycle in metazoans (for cyclin phylogeny see Figure 3-figure supplement 1) are found across the major branches of eukaryotes. We also found examples of all three sub-families of E2F transcription factors (E2F1-6, DP, E2F7/8) and the pRb family of pocket proteins (for E2F/DP and pRb phylogeny see Figure 3-figure supplement 2 and Figure 3-figure supplement 3). Nearly all species contain the APC specificity subunits Cdc20 and Cdh1/Fzr1, which regulate exit from mitosis and maintain low Cdk activity in G1 (for Cdc20-family APC phylogeny see Figure 3-figure supplement 4). Taken together, these data indicate that LECA cell cycle regulation was based on multiple cyclin families, as well as regulation by the APC complex and members of the pRb and E2F families. More broadly, our phylogenetic analyses tend to place the fungal regulators as sister groups to the metazoan regulators, as would be expected from the known eukaryotic species tree. These phylogenies are in agreement with the hypothesis that many fungal and metazoan regulators were vertically inherited from an opisthokont ancestor rather than loss of these regulators in fungi followed by secondary acquisition through horizontal gene transfer.

Members of the Cdk1-3 family (i.e. CdkA in plants) are also broadly conserved across eukaryotes, suggesting they were the primary LECA cell cycle Cdks (for CDK phylogeny see Figure 3-figure supplement 5). Other cell cycle Cdk families in animal (Cdk4/6) and plant (CdkB) are thought to be specific to those lineages. However, we found CdkB in Stramenopiles, which may have arrived via horizontal transfer during an ancient secondary endosymbiosis with algae as previously suggested (Cavalier-Smith, 1999). We excluded from our analysis other families of cyclin-Cdks, *e.g.,* Cdk7-9, which regulate transcription and RNA processing, and Cdk5 (yeast Pho85), which regulate cell polarity, nutrient regulation, and contribute to cell cycle regulation in yeast (Cao et al., 2014; Guo and Stiller, 2004; Ma et al., 2013; Moffat and Andrews, 2004). While interesting and important, an extensive examination of these cyclin-Cdk families is beyond the scope of this work.

A hybrid E2F-pRb-SBF-Whi5 network on the path of fungal evolution

To identify the possible origins of fungal SBF and Whi5, we searched for sequence homologs across eukaryotic genomes (see Methods). We were unable to find any eukaryotic homologs of SBF or Whi5 outside of fungi, with a few exceptions related to DNA viruses that we discuss later. SBF is an important member of a larger family of winged helix-turn-helix transcription factors in fungi that includes Xbp1, Bqt4, and the APSES family (Acm1, Phd1, Sok2, Egf1, StuA). The emergence of the new fungal components SBF, which includes the large APSES family, and Whi5 is abrupt and occurs near the split of basal fungi from metazoans (Figures 4-5; for SBF only, SBF+APSES, and Whi5 phylogeny see Figure 5-figure supplement 1, Figure 5-figure supplement 2, and Figure 5-figure supplement 3). The precise location remains unclear because we have only 1 Nuclearid genome (*Fonticula alba*) and because Microsporidia are fast-evolving fungal parasites with reduced genomes (Cuomo et al., 2012). Interestingly, the new regulators (SBF and Whi5) and ancestral regulators (E2F and Rb) co-exist broadly across basal fungi and the lineages formerly known as "zygomycetes". Zoosporic, basal fungi such as Chytrids (*e.g.*, *Spizellomyces punctatus*) can have both fungal and animal cell cycle regulators, which likely represents the ancestral fungal hybrid network. SBF-Whi5 in budding yeast plays a similar role to E2F-pRb in animals, which suggests that these pathways were functionally redundant in an ancestral hybrid network. This redundancy would lead to the evolutionary instability of the hybrid network and could explain why different

constellations of components are present in the extant zygomycetes and basal fungi. For example, the zygomycetes have lost pRb while retaining E2F, which was then abruptly lost in the transition to Dikarya. However, with the possible exception of Microsporidia, all fungi have retained SBF and never completely reverted back to the original ancestral state; see Figure 5.

The SBF and E2F family of transcription factors are unlikely to be orthologs.

A simple scenario to explain the emergence of a hybrid network would be gene duplication of the E2F pathway followed by rapid sequence evolution to create a partially redundant SBF pathway. To this end, we scrutinized the highly conserved E2F and SBF DNA-binding domains to detect any sequence and structural homology. We used the Pfam HMMER model of the E2F/DP DNA-binding domain and SBF DNA-binding domain (KilA-N.hmm), which is homologous to the KilA-N domain (Iyer et al., 2002). We used KilA-N.hmm from Pfam for remote homology detection of SBF+APSES because it was trained on a diverse set of KilA DNA-binding domains across bacterial DNA viruses, eukaryotic DNA viruses, and fungal SBF+APSES proteins. Thus, it should be a more sensitive HMM model to detect remote KilA-N homologues in other eukaryotic genomes. Our controls (*H. sapiens* genome, Fig. 6A and *S. cerevisiae* genome, Fig. 6B) demonstrate that E2F_TDP.hmm is specific to E2F/DP and that KilA-N.hmm is specific to SBF+APSES. We show that genomes with hybrid network (*S. punctatus*, Fig. 6C, and other basal fungi with both transcription factors, Fig. 6D) have both E2F/DP and SBF+APSES. E2F_TDP.hmm never hits an SBF+APSES transcription factor and KilA-N.hmm never hits an E2F transcription factor (i.e. there are no scores on the diagonal of the panels in Figure 6). Thus, there is no misclassification by the Pfam HMM models. These data suggest that SBF and other KilA-N domains have no more sequence identity than non-homologous proteins to E2F/DP. We find that non-fungal genomes only hit E2F/DP (Fig. 6E) with the notable exception of *Trichomonas vaginalis*, the only non-fungal genome with E2F/DP and KilA-N homologs (Fig. 6F). We will discuss the case of *T. vaginalis* in the next section.

To test the possibility that SBF was a gene duplication of E2F/DP and evolution was so rapid that sequence identity was lost, but structural and functional homology to E2F/DP was maintained, we looked for possible evidence of structural homology. The

DNA-binding domains of SBF/MBF (Taylor et al., 2000; Xu et al., 1997) and E2F/DP (Zheng et al., 1999) are structurally classified as members of the winged-helix-turn-helix (wHTH) family, which is found in both prokaryotes and eukaryotes (Aravind and Koonin, 1999; Aravind et al., 2005; Gajiwala and Burley, 2000). Although the DNA-binding domains of E2F/DP and SBF/MBF are both classified as wHTH proteins, they show important differences in overall structure and mode of protein-DNA complex formation that lead us to conclude that it is highly unlikely that they are orthologs.

Many wHTH transcription factors, including the E2F/DP family, have a 'recognition helix' that interacts with the major or minor grooves of the DNA. The E2F/DP family has an RRXYD DNA-recognition motif in its helix that is invariant within the E2F/DP family and is responsible for interacting with the conserved, core GCGC motif (Zheng et al., 1999) (see Figure 7A: red structure). The RRXYD recognition motif is strikingly conserved in E2F/DP across all eukaryotes, including the E2F/DP proteins uncovered in basal fungi (Figure 7B, left). The first solved SBF/MBF crystal structure, Mbp1 from *S. cerevisiae* in the absence of DNA, suggested Mbp1 recognizes its MCB (Mlu I cell cycle box, ACGCGT) binding site via a recognition helix (Taylor et al., 1997; Xu et al., 1997). However, a recent crystal structure of PCG2, an SBF/MBF homolog in the rice blast fungus *Magnaporthe oryzae*, in complex with its MCB binding site does not support this proposed mode of DNA binding (Liu et al., 2015). In striking contrast to many wHTH structures, in which the recognition helix is the mediator of DNA binding specificity, the wing of PCG2 binds to the minor groove to recognize the MCB binding site. The two glutamines in the wing (Q82, Q89) are the key elements that recognize the core MCB binding motif CGCG (Figure 7A, blue structure). Family-specific conservation in the DNA-binding domain is observed for all members of the SBF and APSES family, including basal fungal sequences (Figure 7B, right). In summary, the incongruences in sequence, structure, and mode of DNA-interaction between E2F/DP and SBF/MBF families strongly suggest that SBF is not derived from E2F.

Viral origin and evolution of the fungal SBF and APSES family

Since SBF is unlikely to be orthologous to the E2F family of transcription factors, we considered the straightforward alternative. Previous work has shown that the DNA-binding domain of the APSES and SBF proteins is homologous to a viral KilA-N domain (Iyer et al., 2002). KilA-N is a member of a core set of "viral hallmark genes" found across diverse DNA viruses that infect eubacteria, archaea, and eukaryotes (Koonin et al., 2006). Outside the fungal SBF/APSES sub-family, little is known about

the KilA-N domain structure, its DNA-binding recognition sequence, and function (Brick et al., 1998). The wide distribution of DNA viruses and KilA-N across the three domains of life suggests that the fungal ancestor likely acquired SBF via horizontal gene transfer rather than the other way around.

To broaden the scope of analysis beyond the eukaryotic genomes that we studied, we carefully surveyed all KilA-N domains detected by the Pfam database. The majority of known KilA-N domains (weighted by species, not the number of sequences) are found in prokaryotes (85%) with a smaller fraction (10%) found in eukaryotes and a smaller fraction found in DNA viruses (5%). The KilA-N domains in prokaryotes appear to be either integrated by or derived from prokaryotic DNA viruses (i.e. bacteriophage), and thus, we will treat them as such. Within the eukaryotes, all known KilA-N domains are found in fungal genomes with three notable exceptions.

The first exception is *Trichomonas vaginalis*, a parasitic excavate with 1000+ KilA-N domains (Figure 6F). The *T. vaginalis* KilA-N domains have top blast hits to prokaryotic and eukaryotic DNA viruses, e.g. Mimivirus, a large double-stranded DNA virus of the Nucleo-Cytoplasmic Large DNA Viruses (Yutin et al., 2009). Mimiviruses are giant viruses known to infect simple eukaryotic hosts, such as *Acanthamoeba* and possibly other eukaryotes (Abrahão et al., 2014; Raoult and Forterre, 2008). The second and third exceptions are found in two insects, *Acyrthosiphon pisum* ('pea aphid') and *Rhodius prolixus* ('triatomid bug'). The one KilA-N domain in *A. pisum* genome has a top blast hit to eukaryotic DNA viruses (e.g. Invertebrate Iridescent Virus 6). The three KilA-N domains in *R. prolixus* have top blast hits to prokaryotic DNA viruses (e.g. Enterobacteria phage P1). The diverse and sparse distribution of KilA-N domains throughout the eukaryotic genomes is consistent with their horizontal gene transfer into hosts from eukaryotic DNA viruses and/or via engulfed bacteria that were infected with prokaryotic DNA viruses. In fact, the horizontal transfer of genes between Mimivirus and their eukaryotic host, or the prokaryotic parasites within the host, has been shown to be a more frequent event that previously thought (Moreira and Brochier-Armanet, 2008)

To gain further insight into the possible evolutionary origins of the SBF subfamily via horizontal gene transfer, we aligned diverse KilA-N sequences from the Uniprot and

PFAM database to the KilA-N domain of our most basal fungal SBF+APSES sequences (Zoosporic fungi ("chytrids") and "Zygomycetes") and built a phylogenetic tree (Figure 8). There are three major phylogenetic lineages of KilA-N domains: those found in eukaryotic viruses, prokaryotic viruses, and the fungal SBF+APSES family. Our results show that the fungal SBF family is monophyletic and is strongly supported by multiple phylogenetic support metrics. This suggests a single HGT event as the most likely scenario that established the SBF+APSES family in a fungal ancestor. However, our current phylogeny is unable to distinguish whether the SBF family arrived in a fungal ancestor through a eukaryotic virus or a phage-infected bacterium. Structural and functional characterization of existing viral KilA-N domains could help distinguish between these two hypotheses.

SBF ancestor could regulate E2F-target genes

Of all the members of the SBF+APSES family, the most likely candidate to be a "founding" TF is SBF, as it is the only member present in all fungi (Figure 5). In budding yeast and other fungi, SBF functions in G1/S cell cycle regulation and binds a consensus site CGCGAA (Gordân et al., 2011), which overlaps with the consensus site GCGSSAAA for the E2F family (Rabinovich et al., 2008); see Figure 9A. The APSES regulators, Xbp1, and MBF in budding yeast bind TGCA, TCGA, ACGCGT motifs, respectively. A viral origin of the SBF+APSES family—with the founding member involved in cell cycle control—suggests the hypothesis that perhaps the founder TF functioned like a DNA tumor virus protein and hijacked cell cycle control to promote proliferation.

For the viral TF (SBF) to hijack cell cycle control in the fungal ancestor, it must have been able to both bind E2F regulatory regions and then activate the expression of genes under E2F in a cell cycle-regulated fashion. The overlap between the conserved E2F and SBF consensus sites suggests that ancestral SBF could bind E2F regulatory regions (Figure 9A). However, a single base pair substitution in the SBF motif can reduce gene expression by up to ~95% (Andrews and Moore, 1992) and flanking regions outside the core are often important for binding affinity and gene expression (Nutiu et al., 2011). To our knowledge, no one has directly measured the extent to which animal E2F and yeast SBF bind similar sites either *in vivo* or *in vitro*. To first test whether yeast SBF can bind a canonical E2F binding site, we inserted consensus E2F binding sites in the budding yeast genome. The hijacking hypothesis

would be supported *in vivo* if E2F binding sites could generate SBF-dependent cell cycle regulated gene expression. We used the well-studied *CLN2* promoter, which has three binding sites for SBF (SCB, Swi6-dependent cell cycle box) in a nucleosome-depleted region (Figure 9B-9C). Removal of these SCB sites is known to eliminate cell cycle-dependent gene expression (Bai et al., 2010). In our experiment, the change in the SBF site was much more substantial than a single base pair mutation. We replaced the complete SBF sites (TCACGAAA) of *CLN2* (Koch et al., 1996) with a known E2F binding site (GCGCGAAA) from the promoters of the histone gene cluster in mammals (Rabinovich et al., 2008). We observed significant oscillations in GFP expression, which were coordinated with the cell cycle. Importantly, the amplitude of these oscillations was dependent on the budding yeast SBF (Swi4), but not MBF (Mbp1), and disappeared when the 3 binding sites were removed (Figure 9C-9D). This experiment demonstrates that budding yeast SBF can bind E2F-like sites, despite the fact that Dikarya lost ancestral E2F hundreds of millions of years ago.

There are, of course, other possible E2F DNA binding sites that we could have used in our experiment; we picked this one because it is a well-characterized E2F binding site. To further explore the overlap in sequence specificity, we analyzed data from high-throughput protein-binding microarray (PBM) assays (Afek et al., 2014; Badis et al., 2008) of human E2F (E2F1) and budding yeast SBF (Swi4). PBM assays measure, in a single experiment, the binding of recombinant proteins to tens of thousands of synthetic DNA sequences, guaranteed to cover all possible 10-bp DNA sequences in a maximally compact representation (each 10-mer occurs once and only once). We used these PBM data to generate DNA motifs for E2F and SBF (Berger et al., 2006), and to compute, for each possible 8-bp DNA sequence, an enrichment score (or E-score) that reflects the specificity of the protein for that 8-mer. E-scores vary between -0.5 and +0.5, with larger values corresponding to higher affinity binding sites (Berger et al., 2006). As shown in Figure 10A, E2F1 and Swi4 can bind a set of common motifs. For example, the E2F binding site variant that we tested in budding yeast (GCGCGAAA, highlighted in red), is one of the sites commonly bound *in vitro* by E2F and SBF.

Most notably, the *in vitro* PBM data show that there are specific motifs that can be bound only by E2F or only by SBF. To identify the key nucleotide differences

between E2F-only and SBF-only binding, we created motifs of E2F-only and SBF-only sites. The consensus E2F-only (NNSGCGSN) and SBF-only (NNCRCGNN) motifs indicate that differential specificity between E2F1 and SBF is mediated by the nucleotides in the 3rd and 4th positions (underlined) before the invariant CG at the 5th and 6th positions. E2F has a strict preference for G in the 4th position, where as SBF has a strict preference for C in the 3rd position (Figure 10A).

We then scanned the promoters of known E2F target genes from the human genome (*CCNE1*, *E2F1*, *EZH2*) with our empirically-defined DNA binding sites from PBM assays (Afek et al., 2014; Badis et al., 2008) to predict putative E2F-only, SBF-only, and common sites (Figure 10B). As expected, there are many predicted E2F-only and common (E2F & SBF) sites that could be bound by E2F in these known target genes. However, we could also find many potential SBF-only binding sites in these same promoters. We then extended our analysis to 290 known E2F target genes in the human genome to test the generality of SBF cross-binding to E2F sites (Figure 10C). Most E2F target promoters could be bound by SBF. Taken together, this set of experiments lends support to the hijacking hypothesis, where an ancestral SBF may have taken control of several E2F-regulated genes.

DISCUSSION

Cell division is an essential process that has been occurring in an uninterrupted chain for billions of years. Thus, one expects strong conservation in the regulatory network controlling the eukaryotic cell division cycle. Consistent with this idea, cell cycle network structure is highly similar in budding yeast and animal cells. However, many components performing similar functions, such as the SBF and E2F transcription factors, lack sequence identity, suggesting a significant degree of evolution or independent origin. To identify axes of conservation and evolution in eukaryotic cell cycle regulation, we examined a large number of genome sequences in Archaeplastida, Amoebozoa, SAR, Haptophyta, Cryptophyta, Excavata, Metazoa and Fungi. Across eukaryotes, we found a large number of proteins homologous to metazoan rather than fungal G1/S regulators. Our analysis indicates that the last eukaryotic common ancestor likely had complex cell cycle regulation based on Cdk1, Cyclins D, E, A and B, E2F, pRb and APC family proteins.

In contrast, SBF was not present in the last common eukaryotic ancestor, and abruptly emerged, with its regulator Whi5, in fungi likely due to the co-option of a viral KilA-N protein at the base of the fungal lineage. The origin of Whi5 is unclear because we found no homologs outside of fungi. Whi5 is a mostly unstructured protein, which, like pRb, recruits transcriptional inhibitor proteins to specific sites on DNA via transcription factor binding (Huang et al., 2009; Wang et al., 2009). The relatively simple structure of Whi5 suggests that it may have been subsequently co-opted as a phosphopeptide to entrain SBF activity to cell cycle regulated changes in Cdk activity (Figure 11).

The replacement of E2F-Rb with SBF-Whi5 at the core of the cell cycle along the fungal lineage raises the question as to how such a drastic change to fundamental regulatory network could evolve. One answer can be found in the evolution of transcription factors. When the function of an essential transcription factor does change, it often leaves behind a core part of its regulon for another factor (Brown et al., 2009; Gasch et al., 2004; Lavoie et al., 2009). This process of handing off transcription factor function has been observed to proceed through an intermediate state, present in some extant genomes, in which both factors perform the function (Tanay et al., 2005). The logic of proceeding through an intermediate state has been well-documented for the regulation of genes expressed only in yeast of mating type **a** (asgs) (Tsong et al., 2006). In the ancestral yeast and many extant species, asgs expression is activated by a protein only present in **a** cells, while in other yeasts, expression is repressed by a protein only present in α cells and **a**/α diploids. The replacement of the ancestral positive regulation by negative regulation occurred via yeast that contained both systems illustrating how an essential function can evolve through a hybrid state (Baker et al., 2012).

Clearly, something similar happened during cell cycle evolution. It appears that the replacement of the E2F-pRb transcription regulatory complex with the SBF-Whi5 complex proceeded via a hybrid intermediate that preserved its function. In the hybrid intermediate, E2F-Rb and SBF-Whi5 may have evolved to be parallel pathways whose functions overlapped to such an extent that the previously essential E2F-Rb pathway could be lost in the transition to Dikarya. Interestingly, many basal fungi (e.g. Chytrids) have preserved rather than lost this hybrid intermediate, which suggests that each pathway may have specialized functions. Chytrids exhibit both

454 animal (e.g. centrioles, flagella, amoeboid movement) and fungal features (e.g. cell
455 walls, hyphal morphology) whose synthesis needs to be coordinated with cell
456 division. The preservation of the hybrid network in chytrids could then be explained if
457 animal-like and fungal features are regulated by the E2F-Rb and SBF-Whi5 pathways
458 respectively (Figure 11). Once Fungi lost many of the ancestral animal-like features
459 (flagellated cells, amoeboid movement) during the emergence of the "zygomycetes"
460 or Dikarya, the ancestral E2F-pRb components could have evolved new functions or
461 have been lost.
462
463 The origin of the hybrid network at the base of Fungi is abrupt and may have been
464 initiated by the arrival of SBF via virus. Many tumor viruses activate cell
465 proliferation. For example, the DNA tumor viruses Adenovirus and SV40 highjack cell
466 proliferation in part by activating the expression of E2F-dependent genes by binding
467 pRb to disrupt inhibition of E2F (DeCaprio, 2009). While the specific mechanisms
468 may differ, when SBF entered the fungal ancestor cell it might have activated the
469 transcription of E2F target genes. Rather than inhibiting the inhibitor of E2F, SBF
470 may have directly competed for E2F binding sites with transcriptionally inactive E2F-
471 Rb complexes (Figure 11). Consistent with this model, we have shown here that SBF
472 can directly regulate gene expression in budding yeast via a consensus E2F binding
473 site. Thus, the cooption of a viral protein generated a hybrid network to ultimately
474 facilitate dramatic evolution of the core cell cycle network in fungi.

475 **MATERIALS AND METHODS**

476 **Identification of potential protein family homologs.**
477 We used Profile-Hidden Markov Models (profile-HMMs) to detect homologs for each of
478 the families studied, using the HMMER 3 package (Eddy, 2011). Profile-HMMs are
479 sensitive tools for remote homology detection. Starting with a set of diverse yet
480 reliable protein homologs is fundamental for detecting remote protein homology and
481 avoiding "model poisoning" (Johnson et al., 2010). To this end, we used reliable
482 training-set homologs from the cell cycle model organisms *Arabidopsis thaliana,*
483 *Homo sapiens, Schizosaccharomyces pombe,* and *Saccharomyces cerevisiae*, to build
484 the profile-HMMs used to detect homologs. Our profile-HMM search used a stringent
485 e-value threshold of 1e-10 to detect putative homologs in the "best" filtered protein
486 sets (where available) of our 100+ eukaryotic genomes (see Supplementary File 1A
487 for genome details). All putative homologs recovered through a profile-HMM search

were further validated (or rejected) using an iterative search algorithm (Jackhmmer) against the annotated SwissProt database using the HMMER web server (Finn et al., 2011).

Our profile-HMM for E2F/DP family only detects E2F or DP, where as our profile-HMM for SBF/MBF family only detects SBF/MBF (or APSES). The same protein was never identified by both profile-HMMs because the sequence profiles and the structure are non-homologous. In the case of basal fungi, which have both E2F/DP and SBF/MBF, all proteins classified as an E2F/DP had clear homology to E2F or DP (see alignment in Figure 7B) and all proteins that we classified as SBF/MBF had clear homology to SBF/MBF (see alignment in Figure 7B).

Phylogenetic-based classification of protein homologs in sub-families.

A phylogenetic analysis and classification was built in four stages. In the first stage, we used MAFFT-L-INS-i (-maxiterate 1000) to align the sequences of eukaryotic protein family members (Katoh and Standley, 2013). We then used probabilistic alignment masking using ZORRO (Wu et al., 2012) to create different datasets with varying score thresholds. Next, we used ProtTest 3 to determine the empirical amino-acid evolutionary model that best fit each of our protein datasets using several criteria: Akaike Information Criterion, corrected Akaike Information Criterion, Bayesian Information Criterion and Decision Theory (Darriba et al., 2011). Last, for each dataset and its best-fitting model, we ran different phylogenetic programs that use maximum-likelihood methods with different algorithmic approximations (RAxML and PhyML) and Bayesian inference methods (PhyloBayes-MPI) to reconstruct the phylogenetic relationships between proteins.

For RAxML analyses, the best likelihood tree was obtained from five independent maximum likelihood runs started from randomized parsimony trees using the empirical evolutionary model provided by ProtTest. We assessed branch support via rapid bootstrapping (RBS) with 100 pseudo-replicates. PhyML 3.0 phylogenetic trees were obtained from five independent randomized starting neighbor-joining trees (RAND) using the best topology from both NNI and SPR moves. Non-parametric Shimodaira-Hasegawa-like approximate likelihood ratio tests (SH-aLRTs) and parametric *à la Bayes* aLRTs (aBayes) were calculated to determine branch support from two independent PhyML 3.0 runs. For Bayesian inference we used PhyloBayes

(rather than the more frequently used MrBayes) because it allows for site-specific amino-acid substitution frequencies, which better models the level of heterogeneity seen in real protein data (Lartillot and Philippe, 2004; Lartillot et al., 2009). We performed Phylobayes analyses by running three independent chains under CAT and the exchange rate provided by ProtTest 3 (e.g. CAT-LG), four discrete gamma categories, and with sampling every 10 cycles. Proper mixing was initially confirmed with Tracer v1.6 (2014). The first 1000 samples were discarded as burn-in, and convergence was assessed using bipartition frequencies and summary statistics provided by bpcomp and tracecomp from Phylobayes. These were visually inspected with an R version of AWTY (https://github.com/danlwarren/RWTY) (Nylander et al., 2008). The best phylogenies are shown in Figure 3-figure supplement 1-5 and Figure 5-figure supplement 1-3, and were used to tentatively classify sequences into sub-families and create Figures 2-5, 8.

We note that the confidence of each node in the phylogenetic trees was assessed using multiple, but complementary support metrics: (1) posterior probability for the Bayesian inference, (2) rapid bootstrap support (Stamatakis, 2006; Stamatakis et al., 2008) for RAxML, and (3) non-parametric Shimodaira-Hasegawa-like approximate likelihood ratio tests (SH-aLRTs) and parametric *à la Bayes* aLRTs (aBayes) for PhyML. These different support metrics complement each other in their advantages and drawbacks. SH-aLRT is conservative enough to avoid high false positive rates but performs better compared to bootstrapping (Guindon et al., 2010; Simmons and Norton, 2014). aBayes is powerful compared to non-parametric tests, but has a tendency to increase false-positive rates under serious model violations, something that can be balanced with SH-aLRTs (Anisimova and Gascuel, 2006; Anisimova et al., 2011).

Strain construction:

Our *CLN2pr-GFP-CLN2PEST* constructs were all derived from pLB02-0mer (described in (Bai et al., 2010) and obtained from Lucy Bai). To create pLB02-CLN2, a synthetic DNA fragment (IDT, Coralville, IA) encompassing a region of the *CLN2* promoter from 1,130 bp to 481 bp upstream of the *CLN2* ORF was digested with BamHI and SphI and ligated into pLB02-0mer digested with the same enzymes. To create pLB02-E2F, which contains E2F binding sites, the same procedure was applied to a

version of the promoter fragment in which the SCBs at 606bp, 581bp, and 538bp upstream of the ORF were replaced with the E2F binding site consensus sequence GCGCGAAA (Thalmeier et al., 1989). All these plasmids were linearized at the BbsI restriction site in the *CLN2* promoter and transformed. Both *swi4Δ* and *mbp1Δ* strains containing pLB02-0mer, pLB02-Cln2, pLB02-E2F fluorescent expression reporters were produced by mating lab stocks using standard methods. JE103 was a kind gift from Dr. Jennifer Ewald. Plasmids and strains are listed in Supplementary File 1B and 1C, respectively.

Imaging and analysis:

Imaging proceeded essentially as described in Bean *et al*., 2006. Briefly, early log-phase cells were pre-grown in SCD and gently sonicated and spotted onto a SCD agarose pad (at 1.5%), which was inverted onto a coverslip. This was incubated on a heated stage on a Zeiss Observer Z.1 while automated imaging occurred (3 minute intervals, 100-300 ms fluorescence exposures). Single-cell time-lapse fluorescence intensity measurements were obtained using software described in (Doncic and Skotheim, 2013; Doncic et al., 2011), and oscillation amplitudes were obtained manually from the resulting traces. The single-cell fluorescence intensity traces used mean cellular intensity with the median intensity of the entire field of view subtracted, to control for any fluctuations in fluorescent background. The resulting measurements were analyzed in R.

Bioinformatic analysis of protein binding microarrays and human promoters:

Universal PBM data for Swi4 was downloaded from the cis-BP database (Weirauch et al., 2014). We used the PBM 8-mer E-scores reported in cis-BP for the data set M0093_1.02:Badis08:SWI4_4482.1_ArrayB. Universal PBM data for E2F1 (Afek et al., 2014) is available in the GEO database (accession number GSE61854, probes UnivV9_*). E2F1 8-mer E-scores were computed using the Universal Protein Binding Microarray (PBM) Analysis Suite (Berger and Bulyk, 2009). An E-score cutoff of 0.37 was used to call SBF and E2F binding sites. This cutoff corresponds to a false positive discovery rate of 0.001 (Badis et al., 2009). To generate DNA motifs for E2F-only, SBF-only, and common sites, we used the Priority software (Gordân et al., 2010) with a uniform prior to align the 8-mers with E-score > 0.37 for E2F only, SBF only, or both E2F and SBF, respectively (Figure 10A). Promoter sequences (1000bp

upstream of transcription start) for known E2F targets (Eser et al., 2011) were retrieved from *Homo sapiens* genome as provided by UCSC (hg38) (BSgenome.Hsapiens.UCSC.hg38) and the annotation package org.Hs.eg.db from Bioconductor (v3.2) (Gentleman et al., 2004; Huber et al., 2015). Only regions for which at least two consecutive sliding windows of 8-mers (1 nucleotide step; 7 overlapping nucleotides) were high scoring (E-score ≥ 0.37) were called as potential SBF or E2F binding regions. Overlapping or "common" binding regions between E2F and SBF were defined as regions that, regardless of their difference in length (nested or partially overlapping), overlapped in at least one full 8-mer (Figure 10B-C).

ACKNOWLEDGEMENTS

We thank F. Cross and A. Robinson-Mosher for extensive discussions, and J. Heitman, D. Lew, J. Nevins, S. Rubin for critical comments on the manuscript. We thank J. Stajich for providing access to the Bioinformatics cluster at the Institute for Integrative Genome Biology at UC Riverside supported by the College of Natural and Agricultural Sciences, the National Science Foundation, and Alfred P Sloan Foundation.

References:

Abrahão, J.S., Dornas, F.P., Silva, L.C.F., Almeida, G.M., Boratto, P.V.M., Colson, P., La Scola, B., and Kroon, E.G. (2014). Acanthamoeba polyphaga mimivirus and other giant viruses: an open field to outstanding discoveries. Virol. J. *11*, 120.

Adl, S.M., Simpson, A.G.B., Lane, C.E., Lukeš, J., Bass, D., Bowser, S.S., Brown, M.W., Burki, F., Dunthorn, M., Hampl, V., et al. (2012). The revised classification of eukaryotes. Journal of Eukaryotic Microbiology *59*, 429–493.

Afek, A., Schipper, J.L., Horton, J., Gordân, R., and Lukatsky, D.B. (2014). Protein-DNA binding in the absence of specific base-pair recognition. Proc. Natl. Acad. Sci. U.S.a. *111*, 17140–17145.

Andrews, B.J., and Moore, L. (1992). Mutational analysis of a DNA sequence involved in linking gene expression to the cell cycle. Biochem. Cell Biol. *70*, 1073–1080.

Anisimova, M., and Gascuel, O. (2006). Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. Systematic Biology *55*, 539–552.

Anisimova, M., Gil, M., Dufayard, J.-F., Dessimoz, C., and Gascuel, O. (2011). Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. Systematic Biology *60*, 685–699.

629 Aravind, L., and Koonin, E.V. (1999). DNA-binding proteins and evolution of
630 transcription regulation in the archaea. Nucleic Acids Res *27*, 4658–4670.

631 Aravind, L., Anantharaman, V., Balaji, S., Babu, M.M., and Iyer, L.M. (2005). The
632 many faces of the helix-turn-helix domain: Transcription regulation and beyond⋆.
633 FEMS Microbiol Rev *29*, 231–262.

634 Badis, G., Berger, M.F., Philippakis, A.A., Talukder, S., Gehrke, A.R., Jaeger, S.A.,
635 Chan, E.T., Metzler, G., Vedenko, A., Chen, X., et al. (2009). Diversity and
636 complexity in DNA recognition by transcription factors. Science *324*, 1720–1723.

637 Badis, G., Chan, E.T., van Bakel, H., Pena-Castillo, L., Tillo, D., Tsui, K., Carlson,
638 C.D., Gossett, A.J., Hasinoff, M.J., Warren, C.L., et al. (2008). A library of yeast
639 transcription factor motifs reveals a widespread function for Rsc3 in targeting
640 nucleosome exclusion at promoters. Mol Cell *32*, 878–887.

641 Bai, L., Charvin, G., Siggia, E.D., and Cross, F.R. (2010). Nucleosome-Depleted
642 Regions in Cell-Cycle-Regulated Promoters Ensure Reliable Gene Expression in Every
643 Cell Cycle. Dev Cell *18*, 544–555.

644 Baker, C.R., Booth, L.N., Sorrells, T.R., and Johnson, A.D. (2012). Protein
645 modularity, cooperative binding, and hybrid regulatory states underlie transcriptional
646 network diversification. Cell *151*, 80–95.

647 Berger, M.F., and Bulyk, M.L. (2009). Universal protein-binding microarrays for the
648 comprehensive characterization of the DNA-binding specificities of transcription
649 factors. Nat Protoc *4*, 393–411.

650 Bertoli, C., Skotheim, J.M., and de Bruin, R.A.M. (2013). Control of cell cycle
651 transcription during G1 and S phases. Nat Rev Mol Cell Biol *14*, 518–528.

652 Brick, D.J., Burke, R.D., Schiff, L., and Upton, C. (1998). Shope fibroma virus RING
653 finger protein N1R binds DNA and inhibits apoptosis. Virology *249*, 42–51.

654 Brown, V., Sabina, J., and Johnston, M. (2009). Specialized Sugar Sensing in Diverse
655 Fungi. Curr Biol.

656 Cao, L., Chen, F., Yang, X., Xu, W., Xie, J., and Yu, L. (2014). Phylogenetic analysis
657 of CDK and cyclin proteins in premetazoan lineages. BMC Evol Biol *14*, 1–16.

658 Cao, L., Peng, B., Yao, L., Zhang, X., Sun, K., Yang, X., and Yu, L. (2010). The
659 ancient function of RB-E2F pathway: insights from its evolutionary history. Biol
660 Direct *5*, 55–.

661 Cavalier-Smith, T. (1999). Principles of Protein and Lipid Targeting in Secondary
662 Symbiogenesis: Euglenoid, Dinoflagellate, and Sporozoan Plastid Origins and the
663 Eukaryote Family Tree. Journal of Eukaryotic Microbiology *46*, 347–366.

664 Costanzo, M., Nishikawa, J.L., Tang, X., Millman, J.S., Schub, O., Breitkreuz, K.,
665 Dewar, D., Rupes, I., Andrews, B., and Tyers, M. (2004). CDK activity antagonizes
666 Whi5, an inhibitor of G1/S transcription in yeast. Cell *117*, 899–913.

667 Coudreuse, D., and Nurse, P. (2010). Driving the cell cycle with a minimal CDK

668 control network. Nature *468*, 1074–1079.

669 Cross, F.R., Buchler, N.E., and Skotheim, J.M. (2011). Evolution of networks and
670 sequences in eukaryotic cell cycle control. Philosophical Transactions of the Royal
671 Society B: Biological Sciences *366*, 3532–3544.

672 Cuomo, C.A., Desjardins, C.A., Bakowski, M.A., Goldberg, J., Ma, A.T., Becnel, J.J.,
673 Didier, E.S., Fan, L., Heiman, D.I., Levin, J.Z., et al. (2012). Microsporidian genome
674 analysis reveals evolutionary strategies for obligate intracellular growth. Genome Res
675 *22*, 2478–2488.

676 Dacks, J.B., and Field, M.C. (2007). Evolution of the eukaryotic membrane-trafficking
677 system: origin, tempo and mode. J Cell Sci *120*, 2977–2985.

678 Darriba, D., Taboada, G.L., Doallo, R., and Posada, D. (2011). ProtTest 3: fast
679 selection of best-fit models of protein evolution. Bioinformatics *27*, 1164–1165.

680 de Bruin, R.A.M., McDonald, W.H., Kalashnikova, T.I., Yates, J., and Wittenberg, C.
681 (2004). Cln3 activates G1-specific transcription via phosphorylation of the SBF bound
682 repressor Whi5. Cell *117*, 887–898.

683 DeCaprio, J.A. (2009). How the Rb tumor suppressor structure and function was
684 revealed by the study of Adenovirus and SV40. Virology *384*, 274–284.

685 Doncic, A., and Skotheim, J.M. (2013). Feedforward Regulation Ensures Stability and
686 Rapid Reversibility of a Cellular State. Mol Cell.

687 Doncic, A., Falleur-Fettig, M., and Skotheim, J.M. (2011). Distinct interactions select
688 and maintain a specific cell fate. Mol Cell *43*, 528–539.

689 Doonan, J.H., and Kitsios, G. (2009). Functional evolution of cyclin-dependent
690 kinases. Mol Biotechnol *42*, 14–29.

691 Eddy, S.R. (2011). Accelerated profile HMM searches. PLoS Comput Biol *7*,
692 e1002195.

693 Eichinger, L., Pachebat, J.A., Glöckner, G., Rajandream, M.-A., Sucgang, R.,
694 Berriman, M., Song, J., Olsen, R., Szafranski, K., Xu, Q., et al. (2005). The genome
695 of the social amoeba Dictyostelium discoideum. Nature *435*, 43–57.

696 Eme, L., Trilles, A., Moreira, D., and Brochier-Armanet, C. (2011). The phylogenomic
697 analysis of the anaphase promoting complex and its targets points to complex and
698 modern-like control of the cell cycle in the last common ancestor of eukaryotes. BMC
699 Evol Biol *11*, 265.

700 Eser, U., Falleur-Fettig, M., Johnson, A., and Skotheim, J.M. (2011). Commitment to
701 a Cellular Transition Precedes Genome-wide Transcriptional Change. Mol Cell *43*,
702 515–527.

703 Fang, S.-C., de los Reyes, C., and Umen, J.G. (2006). Cell size checkpoint control by
704 the retinoblastoma tumor suppressor pathway. PLoS Genet *2*, e167.

705 Finn, R.D., Clements, J., and Eddy, S.R. (2011). HMMER web server: interactive

706 sequence similarity searching. Nucleic Acids Res *39*, W29–W37.

707 Fritz-Laylin, L.K., Prochnik, S.E., Ginger, M.L., Dacks, J.B., Carpenter, M.L., Field,
708 M.C., Kuo, A., Paredez, A., Chapman, J., Pham, J., et al. (2010). The genome of
709 Naegleria gruberi illuminates early eukaryotic versatility. Cell *140*, 631–642.

710 Gajiwala, K.S., and Burley, S.K. (2000). Winged helix proteins. Curr Opin Struct Biol
711 *10*, 110–116.

712 Gasch, A.P., Moses, A.M., Chiang, D.Y., Fraser, H.B., Berardini, M., and Eisen, M.B.
713 (2004). Conservation and evolution of cis-regulatory systems in ascomycete fungi.
714 PLoS Biol *2*, e398.

715 Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis,
716 B., Gautier, L., Ge, Y., Gentry, J., et al. (2004). Bioconductor: open software
717 development for computational biology and bioinformatics. Genome Biol *5*, R80.

718 Gordân, R., Murphy, K.F., McCord, R.P., Zhu, C., Vedenko, A., and Bulyk, M.L.
719 (2011). Curated collection of yeast transcription factor DNA binding specificity data
720 reveals novel structural and gene regulatory insights. Genome Biol *12*, R125.

721 Gordân, R., Narlikar, L., and Hartemink, A.J. (2010). Finding regulatory DNA motifs
722 using alignment-free evolutionary conservation information. Nucleic Acids Res *38*,
723 e90–e90.

724 Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O.
725 (2010). New algorithms and methods to estimate maximum-likelihood phylogenies:
726 assessing the performance of PhyML 3.0. Systematic Biology *59*, 307–321.

727 Gunbin, K.V., Suslov, V.V., Turnaev, I.I., Afonnikov, D.A., and Kolchanov, N.A.
728 (2011). Molecular evolution of cyclin proteins in animals and fungi. BMC Evol Biol *11*,
729 224.

730 Guo, Z., and Stiller, J.W. (2004). Comparative genomics of cyclin-dependent kinases
731 suggest co-evolution of the RNAP II C-terminal domain and CTD-directed CDKs. BMC
732 Genomics *5*, 69.

733 Hallmann, A. (2009). Retinoblastoma-related proteins in lower eukaryotes. Commun
734 Integr Biol *2*, 538–544.

735 Harashima, H., Dissmeyer, N., and Schnittger, A. (2013). Cell cycle control across
736 the eukaryotic kingdom. Trends Cell Biol *23*, 345–356.

737 Hasan, M.M., Brocca, S., Sacco, E., Spinelli, M., Papaleo, E., Lambrughi, M.,
738 Alberghina, L., and Vanoni, M. (2013). A comparative study of Whi5 and
739 retinoblastoma proteins: from sequence and structure analysis to intracellular
740 networks. Front Physiol *4*, 315.

741 Huang, D., Kaluarachchi, S., Van Dyk, D., Friesen, H., Sopko, R., Ye, W., Bastajian,
742 N., Moffat, J., Sassi, H., Costanzo, M., et al. (2009). Dual regulation by pairs of
743 cyclin-dependent protein kinases and histone deacetylases controls G1 transcription
744 in budding yeast. PLoS Biol *7*, e1000188.

745 Huber, W., Carey, V.J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B.S.,
746 Bravo, H.C., Davis, S., Gatto, L., Girke, T., et al. (2015). Orchestrating high-
747 throughput genomic analysis with Bioconductor. Nat Methods *12*, 115–121.

748 Iyer, L.M., Koonin, E.V., and Aravind, L. (2002). Extensive domain shuffling in
749 transcription regulators of DNA viruses and implications for the origin of fungal
750 APSES transcription factors. Genome Biol *3*, RESEARCH0012.

751 Johnson, L.S., Eddy, S.R., and Portugaly, E. (2010). Hidden Markov model speed
752 heuristic and iterative HMM search procedure. BMC Bioinformatics *11*, 431.

753 Katoh, K., and Standley, D.M. (2013). MAFFT Multiple Sequence Alignment Software
754 Version 7: Improvements in Performance and Usability. Mol Biol Evol *30*, 772–780.

755 Koch, C., Schleiffer, A., Ammerer, G., and Nasmyth, K. (1996). Switching
756 transcription on and off during the yeast cell cycle: Cln/Cdc28 kinases activate bound
757 transcription factor SBF (Swi4/Swi6) at start, whereas Clb/Cdc28 kinases displace it
758 from the promoter in G2. Genes Dev *10*, 129–141.

759 Koonin, E.V., Senkevich, T.G., and Dolja, V.V. (2006). The ancient Virus World and
760 evolution of cells. Biol Direct *1*, 29.

761 Lartillot, N., and Philippe, H. (2004). A Bayesian mixture model for across-site
762 heterogeneities in the amino-acid replacement process. Mol Biol Evol *21*, 1095–1109.

763 Lartillot, N., Lepage, T., and Blanquart, S. (2009). PhyloBayes 3: a Bayesian
764 software package for phylogenetic reconstruction and molecular dating.
765 Bioinformatics *25*, 2286–2288.

766 Lavoie, H., Hogues, H., and Whiteway, M. (2009). Rearrangements of the
767 transcriptional regulatory networks of metabolic pathways in fungi. Curr Opin
768 Microbiol *12*, 655–663.

769 Liu, J., and Kipreos, E.T. (2000). Evolution of cyclin-dependent kinases (CDKs) and
770 CDK-activating kinases (CAKs): differential conservation of CAKs in yeast and
771 metazoa. Mol Biol Evol *17*, 1061–1074.

772 Liu, J., Huang, J., Zhao, Y., Liu, H., Wang, D., Yang, J., Zhao, W., Taylor, I.A., and
773 Peng, Y.-L. (2015). Structural basis of DNA recognition by PCG2 reveals a novel DNA
774 binding mode for winged helix-turn-helix domains. Nucleic Acids Res *43*, 1231–1240.

775 Ma, Z., Wu, Y., Jin, J., Yan, J., Kuang, S., Zhou, M., Zhang, Y., and Guo, A.-Y.
776 (2013). Phylogenetic analysis reveals the evolution and diversification of cyclins in
777 eukaryotes. Molecular Phylogenetics and Evolution *66*, 1002–1010.

778 Merchant, S.S., Prochnik, S.E., Vallon, O., Harris, E.H., Karpowicz, S.J., Witman,
779 G.B., Terry, A., Salamov, A., Fritz-Laylin, L.K., Maréchal-Drouard, L., et al. (2007).
780 The Chlamydomonas genome reveals the evolution of key animal and plant
781 functions. Science *318*, 245–250.

782 Moffat, J., and Andrews, B. (2004). Late-G1 cyclin-CDK activity is essential for
783 control of cell morphogenesis in budding yeast. Nat Cell Biol *6*, 59–66.

784 Moreira, D., and Brochier-Armanet, C. (2008). Giant viruses, giant chimeras: the
785 multiple evolutionary histories of Mimivirus genes. BMC Evol Biol *8*, 12.

786 Murray, A.W., and Kirschner, M.W. (1989). Cyclin synthesis drives the early
787 embryonic cell cycle. Nature *339*, 275–280.

788 Nasmyth, K. (1995). Evolution of the cell cycle. Philos Trans R Soc Lond, B, Biol Sci
789 *349*, 271–281.

790 Nasmyth, K., and Dirick, L. (1991). The role of SWI4 and SWI6 in the activity of G1
791 cyclins in yeast. Cell *66*, 995–1013.

792 Nutiu, R., Friedman, R.C., Luo, S., Khrebtukova, I., Silva, D., Li, R., Zhang, L.,
793 Schroth, G.P., and Burge, C.B. (2011). Direct measurement of DNA affinity
794 landscapes on a high-throughput sequencing instrument. Nat Biotechnol *29*, 659–
795 664.

796 Nylander, J.A.A., Wilgenbusch, J.C., Warren, D.L., and Swofford, D.L. (2008). AWTY
797 (are we there yet?): a system for graphical exploration of MCMC convergence in
798 Bayesian phylogenetics. Bioinformatics *24*, 581–583.

799 Ogas, J., Andrews, B.J., and Herskowitz, I. (1991). Transcriptional activation of
800 CLN1, CLN2, and a putative new G1 cyclin (HCS26) by SWI4, a positive regulator of
801 G1-specific transcription. Cell *66*, 1015–1026.

802 Rabinovich, A., Jin, V.X., Rabinovich, R., Xu, X., and Farnham, P.J. (2008). E2F in
803 vivo binding specificity: Comparison of consensus versus nonconsensus binding sites.
804 Genome Res *18*, 1763–1777.

805 Raoult, D., and Forterre, P. (2008). Redefining viruses: lessons from Mimivirus. Nat
806 Rev Microbiol *6*, 315–319.

807 Sidorova, J., and Breeden, L. (1993). Analysis of the SWI4/SWI6 protein complex,
808 which directs G1/S-specific transcription in Saccharomyces cerevisiae. Mol Cell Biol
809 *13*, 1069–1077.

810 Simmons, M.P., and Norton, A.P. (2014). Divergent maximum-likelihood-branch-
811 support values for polytomies. Molecular Phylogenetics and Evolution *73*, 87–96.

812 Skotheim, J.M., Di Talia, S., Siggia, E.D., and Cross, F.R. (2008). Positive feedback
813 of G1 cyclins ensures coherent cell cycle entry. Nature *454*, 291–296.

814 Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic
815 analyses with thousands of taxa and mixed models. Bioinformatics *22*, 2688–2690.

816 Stamatakis, A., Hoover, P., and Rougemont, J. (2008). A rapid bootstrap algorithm
817 for the RAxML Web servers. Systematic Biology *57*, 758–771.

818 Stern, B., and Nurse, P. (1996). A quantitative model for the cdc2 control of S phase
819 and mitosis in fission yeast. Trends Genet *12*, 345–350.

820 Tanay, A., Regev, A., and Shamir, R. (2005). Conservation and evolvability in
821 regulatory networks: The evolution of ribosomal regulation in yeast. Proc Natl Acad

822 Sci USA.

823 Taylor, I.A., Treiber, M.K., Olivi, L., and Smerdon, S.J. (1997). The X-ray structure of
824 the DNA-binding domain from the Saccharomyces cerevisiae cell-cycle transcription
825 factor Mbp1 at 2.1 A resolution. J Mol Biol *272*, 1–8.

826 Thalmeier, K., Synovzik, H., and Mertz, R. (1989). Nuclear factor E2F mediates basic
827 transcription and trans-activation by E1a of the human MYC promoter. Genes & ….

828 Travesa, A., Kalashnikova, T.I., de Bruin, R.A.M., Cass, S.R., Chahwan, C., Lee, D.E.,
829 Lowndes, N.F., and Wittenberg, C. (2013). Repression of G1/S Transcription Is
830 Mediated via Interaction of the GTB Motifs of Nrm1 and Whi5 with Swi6. Mol Cell Biol
831 *33*, 1476–1486.

832 Tsong, A.E., Tuch, B.B., Li, H., and Johnson, A.D. (2006). Evolution of alternative
833 transcriptional circuits with identical logic. Nature *443*, 415–420.

834 van den Heuvel, S., and Dyson, N.J. (2008). Conserved functions of the pRB and E2F
835 families. Nat Rev Mol Cell Biol *9*, 713–724.

836 Wang, G., Kong, H., Sun, Y., Zhang, X., Zhang, W., Altman, N., DePamphilis, C.W.,
837 and Ma, H. (2004). Genome-wide analysis of the cyclin family in Arabidopsis and
838 comparative phylogenetic analysis of plant cyclin-like proteins. Plant Physiol *135*,
839 1084–1099.

840 Wang, H., Carey, L.B., Cai, Y., Wijnen, H., and Futcher, B. (2009). Recruitment of
841 Cln3 cyclin to promoters controls cell cycle entry via histone deacetylase and other
842 targets. PLoS Biol *7*, e1000189.

843 Weirauch, M.T., Yang, A., Albu, M., Cote, A.G., Montenegro-Montero, A., Drewe, P.,
844 Najafabadi, H.S., Lambert, S.A., Mann, I., Cook, K., et al. (2014). Determination and
845 inference of eukaryotic transcription factor sequence specificity. Cell *158*, 1431–
846 1443.

847 Wu, M., Chatterji, S., and Eisen, J.A. (2012). Accounting for alignment uncertainty in
848 phylogenomics. PLoS ONE *7*, e30288.

849 Xu, R.M., Koch, C., Liu, Y., Horton, J.R., Knapp, D., Nasmyth, K., and Cheng, X.
850 (1997). Crystal structure of the DNA-binding domain of Mbp1, a transcription factor
851 important in cell-cycle control of DNA synthesis. Structure *5*, 349–358.

852 Yutin, N., Wolf, Y.I., Raoult, D., and Koonin, E.V. (2009). Eukaryotic large nucleo-
853 cytoplasmic DNA viruses: clusters of orthologous genes and reconstruction of viral
854 genome evolution. Virol. J. *6*, 223.

855 Zheng, N., Fraenkel, E., Pabo, C.O., and Pavletich, N.P. (1999). Structural basis of
856 DNA recognition by the heterodimeric cell cycle transcription factor E2F-DP. Genes
857 Dev *13*, 666–674.

858

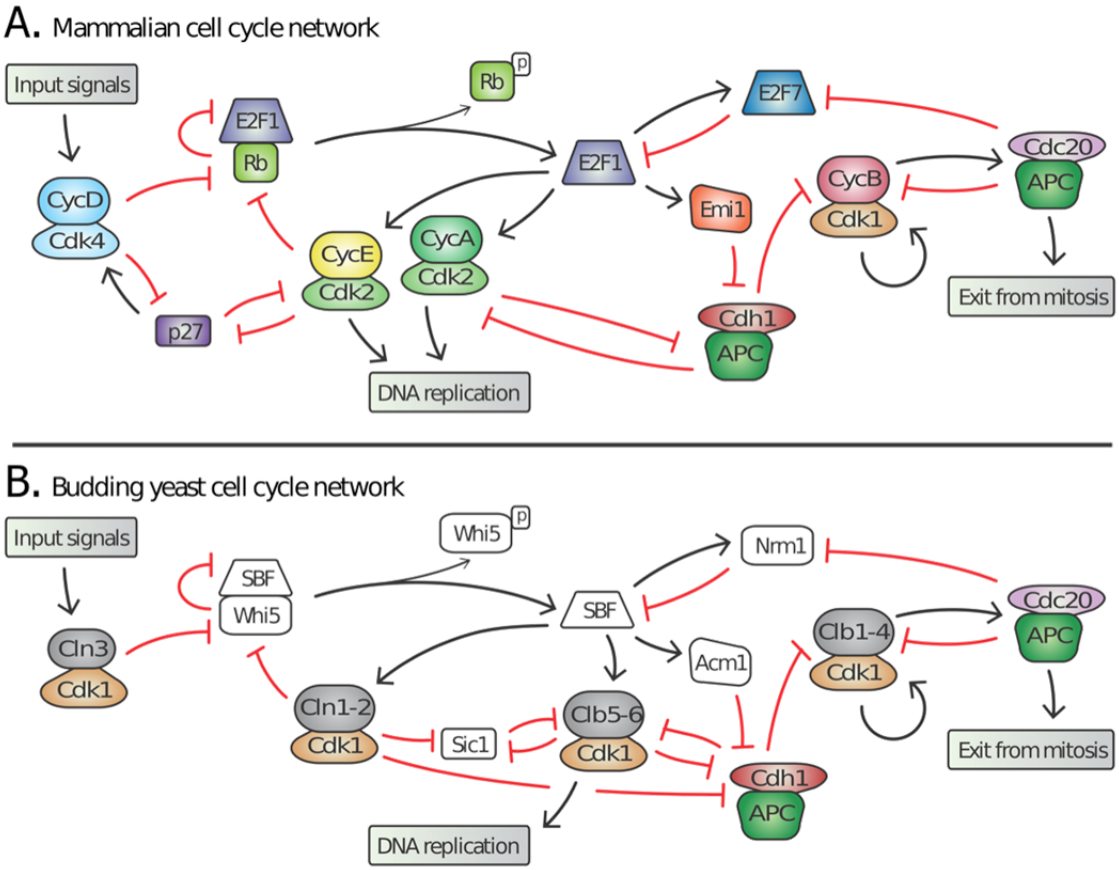


Figure 1. Topology of G1/S regulatory network in mammals and budding yeast is conserved, yet many regulators exhibit no detectable sequence homology. Schematic diagram illustrating the extensive similarities between animal (A) and budding yeast (B) G1/S cell cycle control networks. Similar coloring denotes members of a similar family or sub-family. Fungal components colored white denote proteins with no identifiable animal orthologs.

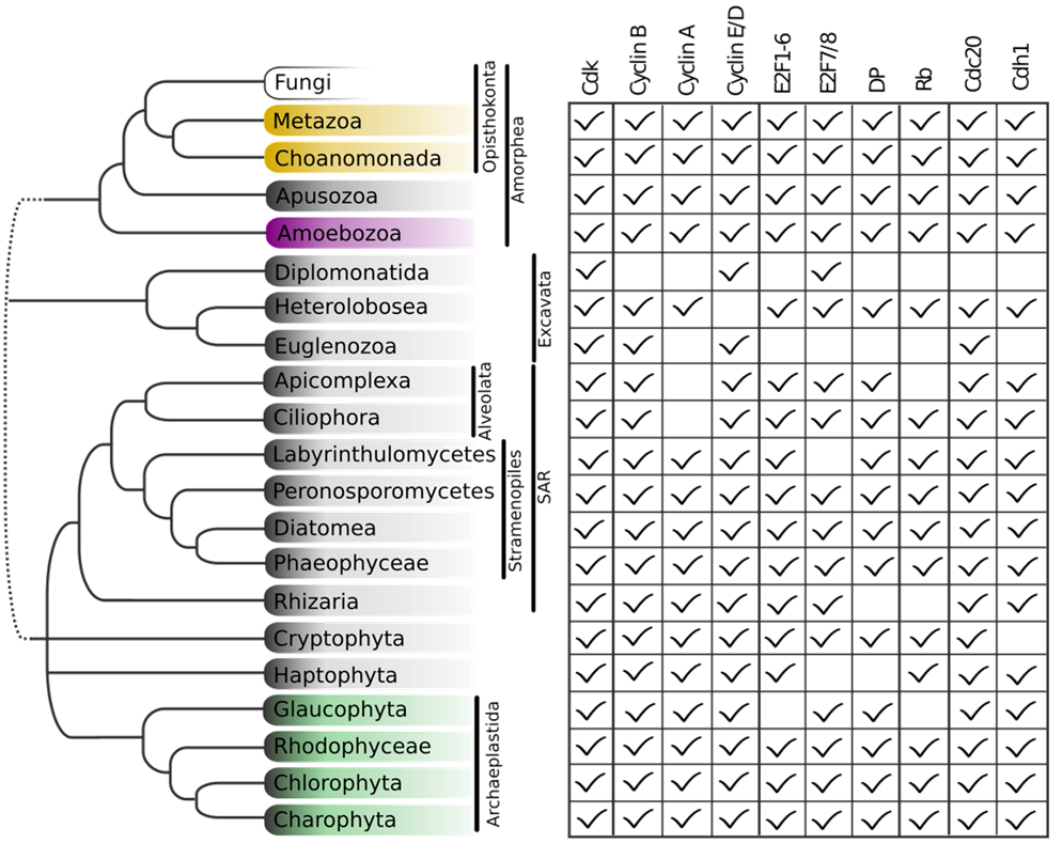


Figure 2. Animal and plant G1/S regulatory network was present in the last eukaryotic common ancestor. Distribution of cell cycle regulators across the eukaryotic species tree (Adl et al., 2012). Animals (Metazoa) and yeasts (Fungi) are sister groups (Opisthokonta), and are distantly related to plants (Charophyta), which are members of the Archaeplastida. Check marks indicate the presence of at least one member of a protein family in at least one sequenced species from the corresponding group. See Figure 3 for a complete list of components in all species analyzed.

			(CycB3)		(CycO/F/J/G/I)						
	(H. sapiens)	Cdk1-4,6	CycB	CycA	CycE/D	E2F1-6	E2F7-8	DP	Rb	Cdc20	Fzr1
Metazoa	*H. sapiens (human)*	5	3	2	13	6	2	3	3	2	1
	G. gallus (chicken)	4	2	2	13	6	2	2	3	2	1
	D. rerio (zebrafish)	5	3	2	12	5	2	3	3	1	2
	B. floridae (lancelet)	5	1	1	6	2	2	1	2	1	2
	C. intestinalis (sea squirt)	5	1	2	5		1	1	1	1	1
	S. purpuratus (urchin)	3	2	1	6	2		1	3	2	1
	L. gigantea (sea snail)	3	2	1	7	2	1	1	2	2	1
	D. melanogaster (fly)	3	2	1	3	2		1	2	1	2
	C. elegans (nematode)	3	4	1	2	2	1	1	1	1	1
	N. vectensis (anemone)	3	3	1	5	5	3	1	1	2	1
	A. queenslandica (sponge)	2	1	1	5	2	1	1	3	2	1
	T. adhaerens (placozoa)	3	2	1	3	2	1	1	2	1	1
Choanomonada	*Sphaeroforma arctica*		1		3	2		2	1	1	1
	Capsaspora owczarzaki		1	1	4	1	1	1	1	1	1
	Monosiga brevicollis	1	1	1	2	1	1	1	1	1	1
	Salpingoeca rosetta	2	1	1	2	2	1	1	2	1	1
Apusozoa	*Thecamonas trahens*	1	1	1	3	1	1	1	1	1	1
Amoebozoa	*Dictyostelium discoideum*	1	1	1	1	1		1	1	1	1
	Dictyostelium purpureum	1	1	1	1	1		1	1	1	1
	Dictyostelium fasciculatum	1	1	1	1	1	1	1	1	1	1
	Polysphondylium pallidum	1	1	1		1	2		1	1	1
	Entamoeba histolytica	6	2		3						2
	Entamoeba nuttalli	6	1		1						2
Diplomonatida	*Giardia intestinalis*	2			2		1				
Heterolobosea	*Naegleria gruberi*	3	3	3		1	1	1	1	3	1
Euglenozoa	*Trypanosoma brucei*	3	1		2					3	
	Leishmania major	3	2		1					2	
	Leishmania donovani	2	2		1					2	
Apicomplexa	*Symbiodinium minutum*	7	2							1	1
	Plasmodium falciparum	2									1
	Cryptosporidium muris	3			2	1	1	1			1
	Toxoplasma gondii	2									1
Ciliophora	*Tetrahymena thermophila*	7	5		11	2	2	3	2	1	8
Labyrinthulomycetes	*Aurantiochytrium limacinum*	3	1	1	3			1	1	1	1
	Schizochytrium aggregatum	3	1	1	2					1	1
	Aplanochytrium kerguelense	3	1	1	2	1		1	1	2	1
Peronosporomycetes	*Phytophthora infestans*	5	1	1	3	1	1	1	1	1	1
	Pythium ultimatum	1	1	1	4	1	1	1	1	1	1
	Hyaloperonospora parasitica	3		2	4			2	1	1	
	Saprolegnia parasitica	3	1	2	4	1	1	2	1	4	1
Diatomea	*Fragilariopsis cylindrus*	2	4	1	10	1	1	1		2	1
	Phaeodactylum tricornutum	2	1	1	10	2	1	1	1	2	1
	Thalassiosira pseudonana	2	3	1	26	1	1	1	1	2	2
Eustigmatales	*Nannochloropsis gaditana*	3	2	1	2			1	1	1	
Phaeophyceae	*Ectocarpus siliculosus*	3	2	1	2	1	1	1	1	2	1
Pelagophyceae	*Aureococcus anophaefferens*	2	2		2	1		1		2	1
Rhizaria	*Bigelowiella natans*	1	2	1	3	1	2			2	1
Cryptophyta	*Guillardia theta*	4	3	1	3	1	1	2	1	3	
Haptophyta	*Emiliania huxleyi*	2	2	3	1	2			1	6	2
Glaucophyta	*Cyanophora paradoxa*	1	1	1	1		1	1		1	2
Rhodophyceae	*Porphyridium cruentum*	3	1	1	1	1	1	1		1	1
	Cyanidioschyzon merolae	2		1	2	1	1	2	1	1	1
Chlorophyta	*Ostreococcus tauri*	2	1	1	2	1	1	1	1	1	1
	Ostreococcus lucimarinus	2	1	1	2	1	1	1	1	1	1
	Micromonas pusilla	2	1	1	2	1	1	1	1	1	1
	Coccomyxa subellipsoidea	4	1	1	1	1	1	1	1	1	1
	Volvox carteri	3	1	1	4	1		1	1	1	1
	Chlamydomonas reinhardtii	3	1	1	4	1		1	1	1	1
Charophyta	*Physcomitrella patens*	9	2	8	2	3	2	2	3	6	4
	Selaginella moellendorffii	3	1	3	3	2	1	1	2	3	2
	Brachypodium distachyon	5	5	6	14	3	2	2	2	5	2
	Oryza sativa	5	6	6	12	4	2	2	2	3	2
	Arabidopsis thaliana	5	11	10	11	3	3	2	1	6	3
	(A. thaliana)	CdkA/B	CycB	CycA	CycD	E2FA-C	DEL	DP	RBR	Cdc20	Fzr1
					(CycSDS)						

Figure 3. Comparative genomic data of G1/S regulators across eukaryotes. We developed profile-HMMs to detect cell division cycle regulators in eukaryotic genomes. For each cell cycle regulatory family (*e.g.*, cyclins), we used molecular phylogeny to classify eukaryotic sequences into sub-families (*e.g.*, Cyclins B, Cyclin A, Cyclins E/D). See Methods for details and Figure 3-supplement 1 (Cyclin), Figure 3-supplement 2 (E2F/DP), Figure 3-supplement 3 (pRb), Figure 3-supplement 4 (Cdc20-family), and Figure 3-supplement 5 (CDK) for final phylogenies. Each entry lists the number of sub-family members (column) for each eukaryotic genome (row). Grey rows list the sub-family gene names in *H. sapiens* and *A. thaliana*. Additional cyclin sub-family members are listed in parentheses.

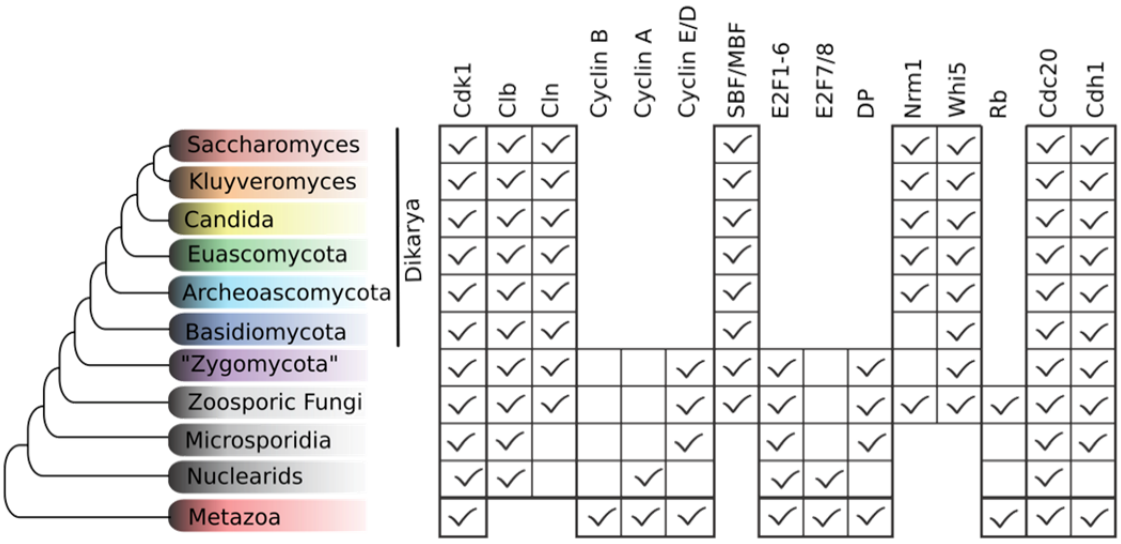
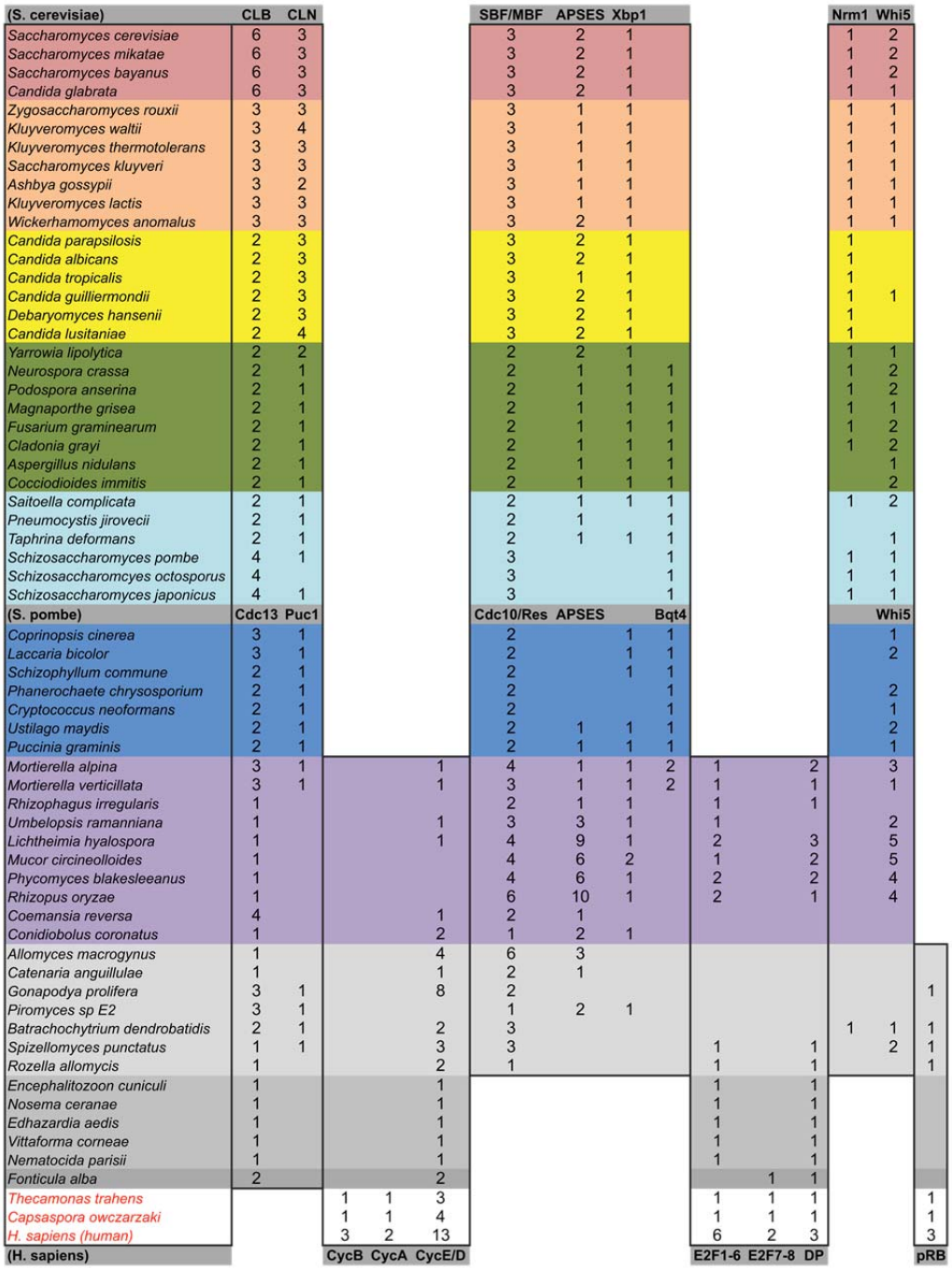


Figure 4. Fungal ancestor evolved novel G1/S regulators, which eventually replaced ancestral cyclins, transcription factors, and inhibitors in Dikarya. Basal fungi and "Zygomycota" contain hybrid networks comprised of both ancestral and fungal specific cell cycle regulators. Check marks indicate the presence of at least one member of a protein family in at least one sequenced species from the group. Cells are omitted (rather than left unchecked) when a family is completely absent from a clade. See Figure 5 for a complete list of components in all fungal species analyzed.



(S. cerevisiae)	CLB	CLN				SBF/MBF	APSES	Xbp1					Nrm1	Whi5	
Saccharomyces cerevisiae	6	3				3	2	1					1	2	
Saccharomyces mikatae	6	3				3	2	1					1	2	
Saccharomyces bayanus	6	3				3	2	1					1	2	
Candida glabrata	6	3				3	2	1					1	1	
Zygosaccharomyces rouxii	3	3				3	1	1					1	1	
Kluyveromyces waltii	3	4				3	1	1					1	1	
Kluyveromyces thermotolerans	3	3				3	1	1					1	1	
Saccharomyces kluyveri	3	3				3	1	1					1	1	
Ashbya gossypii	3	2				3	1	1					1	1	
Kluyveromyces lactis	3	3				3	1	1					1	1	
Wickerhamomyces anomalus	3	3				3	2	1					1	1	
Candida parapsilosis	2	3				3	2	1					1		
Candida albicans	2	3				3	2	1					1		
Candida tropicalis	2	3				3	1	1					1		
Candida guilliermondii	2	3				3	2	1					1	1	
Debaryomyces hansenii	2	3				3	2	1					1		
Candida lusitaniae	2	4				3	2	1					1		
Yarrowia lipolytica	2	2				2	2	1					1	1	
Neurospora crassa	2	1				2	1	1	1				1	2	
Podospora anserina	2	1				2	1	1	1				1	2	
Magnaporthe grisea	2	1				2	1	1	1				1	1	
Fusarium graminearum	2	1				2	1	1	1				1	2	
Cladonia grayi	2	1				2	1	1	1				1	2	
Aspergillus nidulans	2	1				2	1	1	1					1	
Cocciodioides immitis	2	1				2	1	1	1					2	
Saitoella complicata	2	1				2	1	1	1				1	2	
Pneumocystis jirovecii	2	1				2	1		1						
Taphrina deformans	2	1				2	1	1	1					1	
Schizosaccharomyces pombe	4	1				3			1				1	1	
Schizosaccharomcyes octosporus	4					3			1				1	1	
Schizosaccharomyces japonicus	4	1				3			1				1	1	
(S. pombe)	**Cdc13**	**Puc1**				**Cdc10/Res**	**APSES**		**Bqt4**					**Whi5**	
Coprinopsis cinerea	3	1				2			1	1				1	
Laccaria bicolor	3	1				2			1	1				2	
Schizophyllum commune	2	1				2			1	1					
Phanerochaete chrysosporium	2	1				2				1				2	
Cryptococcus neoformans	2	1				2				1				1	
Ustilago maydis	2	1				2	1	1	1					2	
Puccinia graminis	2	1				2	1	1	1					1	
Mortierella alpina	3	1			1	4	1	1	2	1		2		3	
Mortierella verticillata	3	1			1	3	1	1	2	1		1		1	
Rhizophagus irregularis	1					2	1	1		1		1			
Umbelopsis ramanniana	1				1	3	3	1		1				2	
Lichtheimia hyalospora	1				1	4	9	1		2		3		5	
Mucor circineolloides	1					4	6	2		1		2		5	
Phycomyces blakesleeanus	1					4	6	1		2		2		4	
Rhizopus oryzae	1					6	10	1		2		1		4	
Coemansia reversa	4				1	2	1								
Conidiobolus coronatus	1				2	1	2	1							
Allomyces macrogynus	1				4	6	3								
Catenaria anguillulae	1				1	2	1								
Gonapodya prolifera	3	1			8	2									1
Piromyces sp E2	3	1				1	2	1							
Batrachochytrium dendrobatidis	2	1			2	3							1	1	1
Spizellomyces punctatus	1	1			3	3				1		1		2	1
Rozella allomycis	1				2	1				1		1			1
Encephalitozoon cuniculi	1				1					1		1			
Nosema ceranae	1				1					1		1			
Edhazardia aedis	1				1					1		1			
Vittaforma corneae	1				1					1		1			
Nematocida parisii	1				1					1		1			
Fonticula alba	2				2						1	1			
Thecamonas trahens			1	1	3					1	1	1			1
Capsaspora owczarzaki			1	1	4					1	1	1			1
H. sapiens (human)			3	2	13					6	2	3			3
(H. sapiens)			**CycB**	**CycA**	**CycE/D**					**E2F1-6**	**E2F7-8**	**DP**			**pRB**

Figure 5. Comparative genomic data of G1/S regulators across fungi. We developed profile-HMMs to detect fungal-specific cell division cycle regulators in eukaryotic genomes. For each cell cycle regulatory family (*e.g.*, SBF+APSES), we used molecular phylogeny to classify eukaryotic sequences into sub-families (*e.g.*, SBF/MBF, APSES, Xbp1, Bqt4). Corresponding table shows the number of regulators of each class for all species analyzed. See Methods for details and Figure 5-supplement 1 (SBF only), Figure 5-supplement 2 (SBF+APSES), and Figure 5-supplement 3 (Whi5) for final phylogenies. Grey rows list the sub-family gene names in *S. cerevisiae*, *S. pombe*, and *H. sapiens*

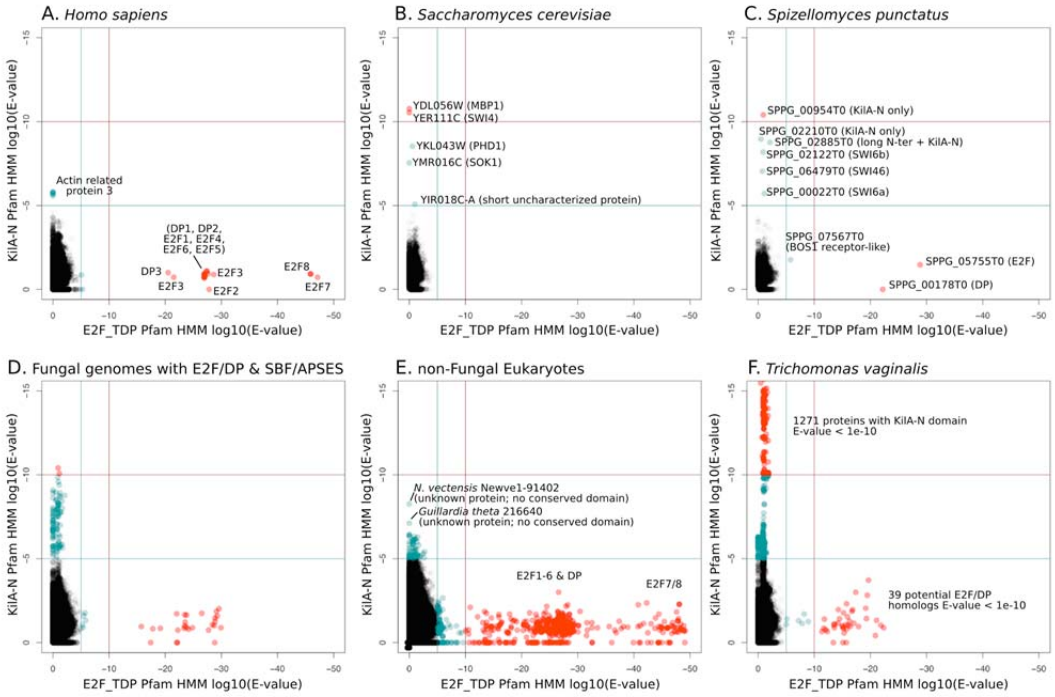


Figure 6. E2F/DP and KilA-N HMMs detect different sequences. We used the Pfam HMMER model of the E2F/DP DNA-binding domain (E2F_TDP.hmm) and the SBF+APSES DNA-binding domain (KilA-N.hmm). We used KilA-N.hmm for remote homology detection of SBF-like proteins because pSMRC.hmm (our HMM model trained on fungal SBFs to detect SBF; see Figure 5 - supplement 1) includes two well-conserved fungal ankyrin repeats in addition to the KilA-N DNA binding domain. Every protein in the query genome (listed at top) was scored using hmmsearch with E2F/DP HMM (x-axis) and KilA-N HMM (y-axis). All scores below 1E-5 (i.e., marginally significant) are blue and those below 1E-10 (i.e. highly significant) are red. All hits with E-values between 1e-5 and 1e-10 were further validated (or rejected) using an iterative search algorithm (Jackhmmer) against the annotated SwissProt database using the HMMER web server (Finn et al., 2011). We then inspected these sequences manually for key conserved KilA-N residues. (A) *Homo sapiens* only has E2F/DP, (B) *Saccharomyces cerevisiae* only has KilA-N (i.e. SBF, MBF, APSES). (C) *Spizellomyces punctatus* and (D) other basal fungi have both E2F/DP and KilA-N. (E) All the non-fungal eukaryote genomes that we surveyed only have E2F/DP. (F) *Trichomonas vaginalis* is one of the few eukaryotes outside of fungi that has both E2F/DP and KilA-N. The E2F/DP HMM and KilA-N HMM always have orthogonal hits (i.e. no protein in our dataset significantly hits both HMMs).

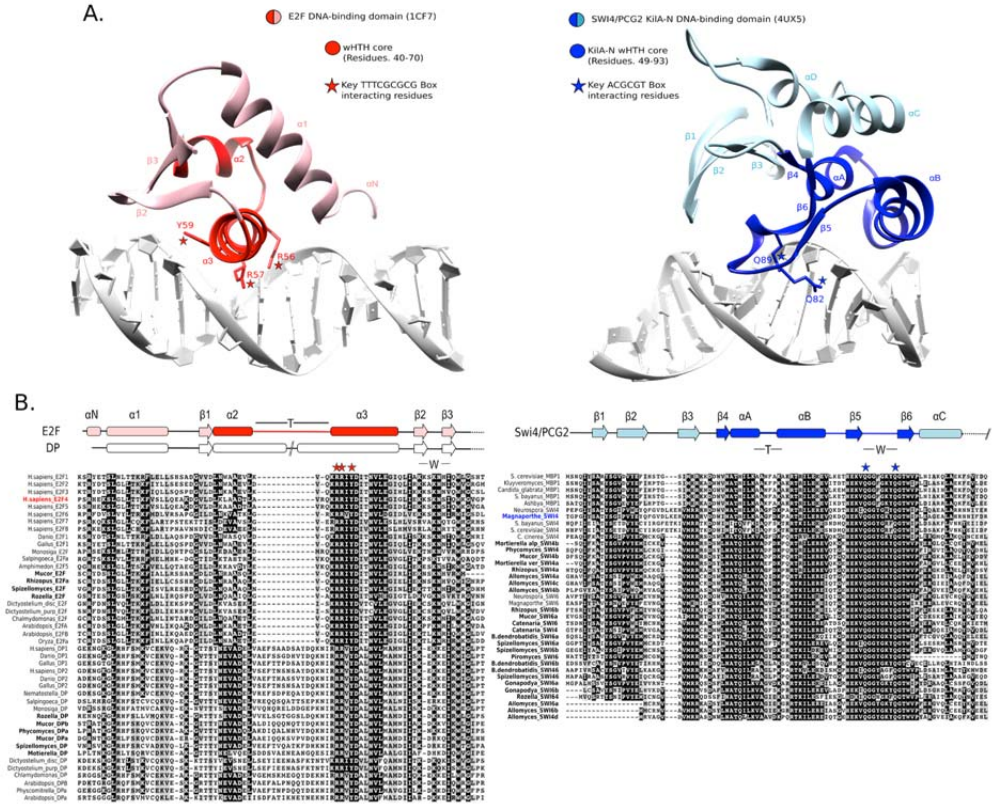


Figure 7. E2F and SBF show incongruences in sequence, structure, and mode of DNA binding. (A) Although both proteins share a winged helix-turn-helix (wHTH) domain, the E2F/DP and SBF/MBF superfamilies do not exhibit significant sequence identity or structural similarity to suggest a common recent evolutionary origin according to CATH or SCOP databases. Furthermore, each wHTH has a different mechanism of interaction with DNA: the arginine and tyrosine side-chains of recognition helix-3 of E2F (E2F4 from *Homo sapiens* (Zheng et al., 1999)) interact with specific CG nucleotides, where as the glutamine side-chains of the "wing" of SBF/MBF (PCG2 from *Magnaporthe oryzae* (Liu et al., 2014)) interact with specific CG nucleotides. (B) Sequence alignment of the DNA binding domain of representative eukaryotic E2F/DP (left) and fungal SBF/MBF (right). The corresponding secondary structure is above the sequence alignment. Evolutionary conserved residues of sequence aligned DNA binding domains are highlighted in black. Bold sequence names correspond to E2F/DP and SBF/MBF sequences from basal fungi. Colored sequence names correspond to sequences of the structures shown in panel A. PDB IDs for the structures used are shown in parentheses. W = wing; T= turn.

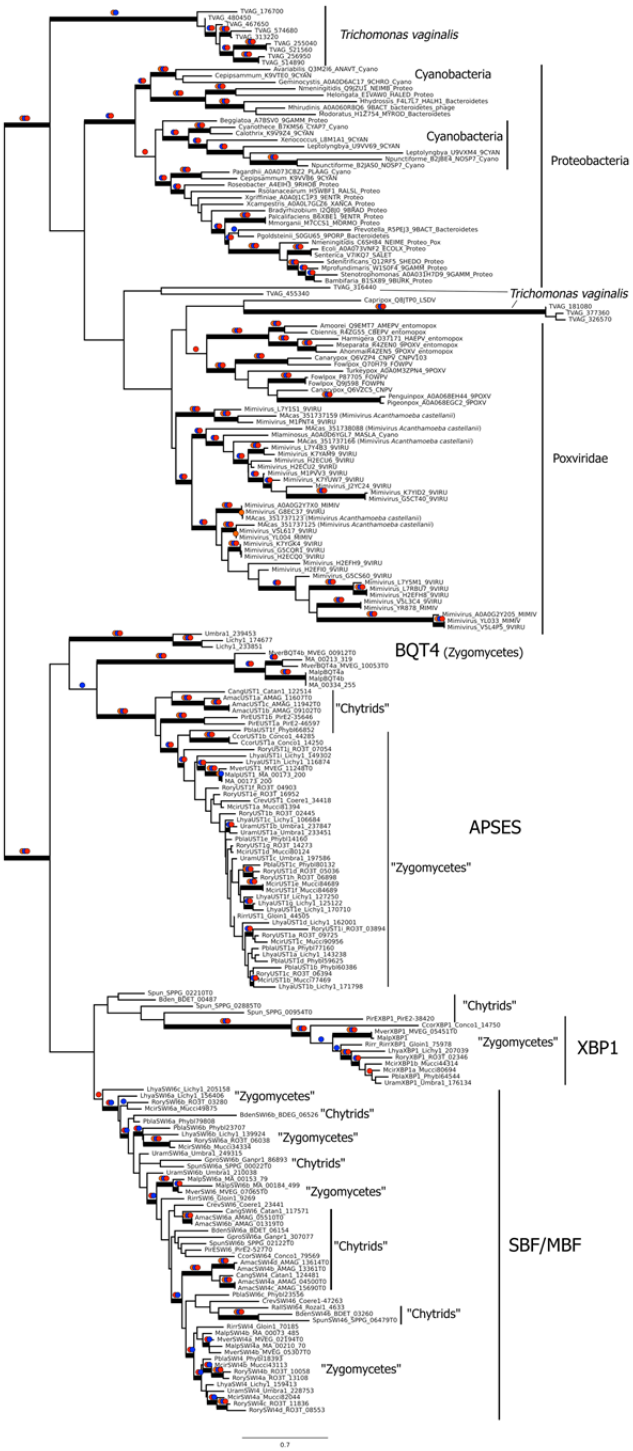


Figure 8. Viral origin of yeast cell cycle transcription factor SBF. Maximum likelihood unrooted phylogenetic tree depicting relationships of fungal SBF-family proteins, KilA-N domains in prokaryotic and eukaryotic DNA viruses. The original dataset was manually pruned to remove long-branches and problematic lineages. Our reduced KilA-N dataset has a total of 219 sequences, 130 positions. Confidence at nodes was assessed with multiple support metrics using different phylogenetic programs under LG model of evolution. Colored dots in branches indicate corresponding branch supports (red dots: PhyML aBayes ≥ 0.9; blue dots: PhyML

SH-aLRT ≥ 0.80; orange: RAxML RBS ≥ 70%). Thick branches indicate significant support by at least two metrics, one parametric and one non-parametric; Scale bar in substitutions per site; (See Methods).

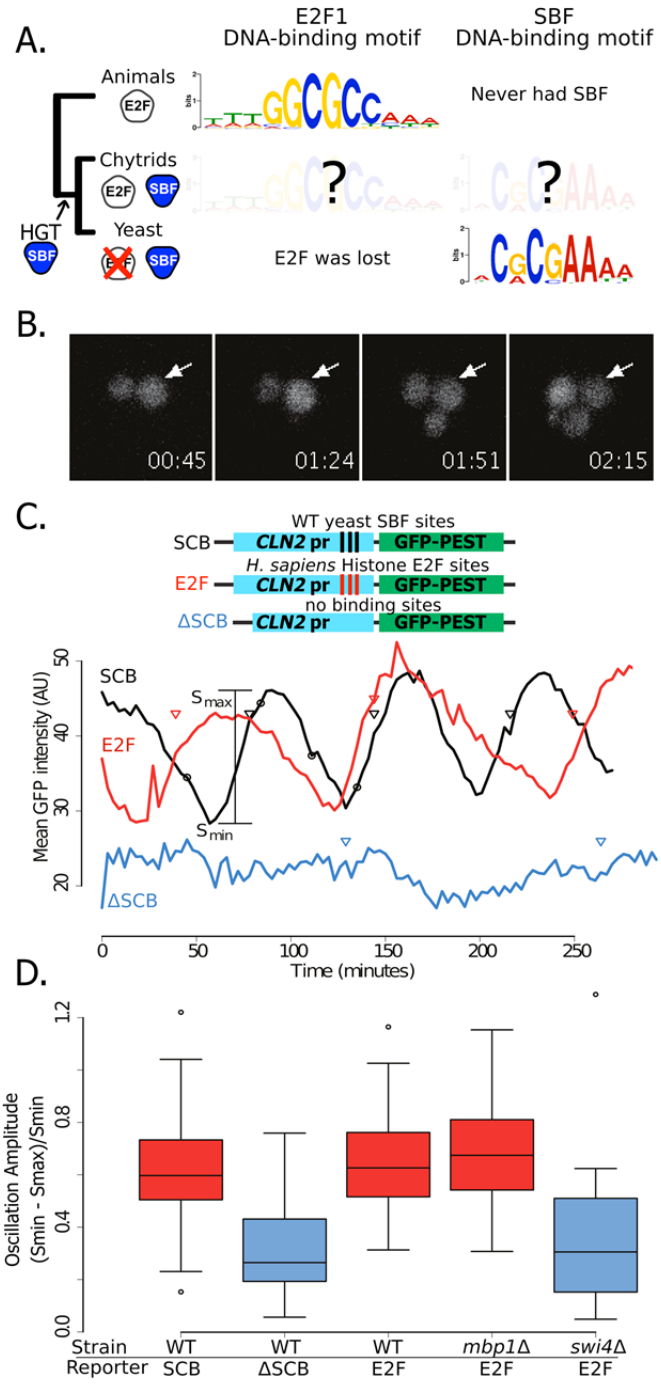
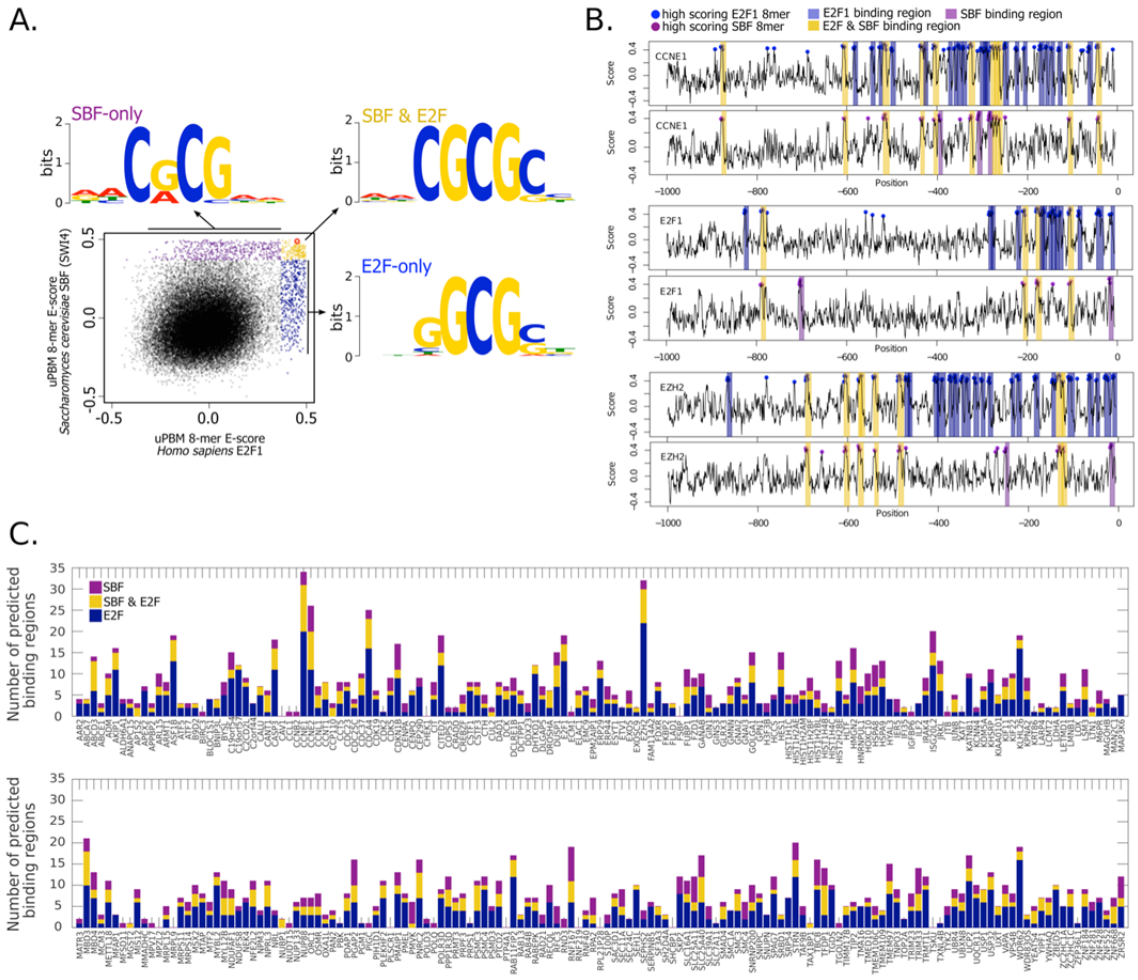


Figure 9. Yeast cell cycle transcription factor SBF can regulate cell cycle-dependent transcription via E2F binding sites in vivo (A) Phylogenetic tree of animals, chytrids, yeast labelled with E2F, SBF or both transcription factors (TF) if present in their genomes. The known DNA-binding motifs of animal E2F (E2F1) and yeast SBF (Swi4) were taken from the JASPAR database, where as Chytrid E2F and SBF motifs are unknown. (B) Fluorescence images of cells expressing a destabilized GFP from the SBF-regulated *CLN2* promoter. (C) Oscillation of a transcriptional reporter in budding yeast. Characteristic time series of GFP expression from a *CLN2* promoter (SCB), a *CLN2* promoter where the SBF binding sites were deleted (ΔSCB), or a *CLN2* promoter where the SBF binding sites were replaced with E2F binding sites

from the human gene cluster promoters (E2F). Oscillation amplitudes were quantified by scaling the mean fluorescence intensity difference from peak to trough divided by the trough intensity $(S_{max} - S_{min})/S_{min}$. Circles denote time points corresponding to (b). Triangles denote budding events. (D) Distribution of oscillation amplitudes for different genotypes and GFP reporters. *swi4Δ* and *mbp1Δ* strains have deletions of the SBF and MBF DNA-binding domain subunits respectively. t-test comparisons within and across red and blue categories yield p-values > 0.3 or < 0.01 respectively. Boxes contain 25^{th}, median and 75^{th} percentiles, while whiskers extend to 1.5 times this interquartile range.



985

986 **Figure 10: High-throughput DNA binding data for yeast SBF and human E2F**
987 **shows that SBF could bind many E2F-regulated promoters in the human**
988 **genome** (A) Plot of *in vitro* protein binding microarray 8-mer E-scores for *Homo*
989 *sapiens* E2F1 (Afek et al., 2014) versus *S. cerevisiae* SBF protein Swi4 (Badis et al.,
990 2008). All 8-mer motifs colored (E-score > 0.37) are considered significant targets
991 with a false positive discovery rate of 0.001 (Badis et al., 2009). Yellow are common
992 8-mer motifs bound by both E2F1 and SBF, blue are E2F-only motifs, and purple are
993 SBF-only motifs. The E2F motif from histone cluster promoters used in Figure 9 is
994 circled in red. (B) E2F1 (top) and SBF (bottom) PBM motifs were used to scan the
995 proximal (1000bp) promoters of E2F-regulated promoters (CCNE1, E2F1, and EZH2).
996 Promoter regions with a significant hit (8-mer E-score > 0.37) to a E2F or SBF motif
997 have blue or purple dot, respectively. Predicted TF-binding regions were defined as
998 at least 2 consecutive overlapping 8-mers (7 nucleotide overlap) and shaded as blue
999 (E2F binding region) or purple (SBF-binding region). Common or E2F & SBF binding
1000 regions were colored yellow and defined as regions that overlap in at least one full 8-
1001 mer (C) Summary of E2F-only regions (blue), SBF-only regions (purple), and E2F
1002 and SBF co-regulated regions (yellow) for a set of 290 E2F-regulated promoters
1003 (Eser et al., 2011).

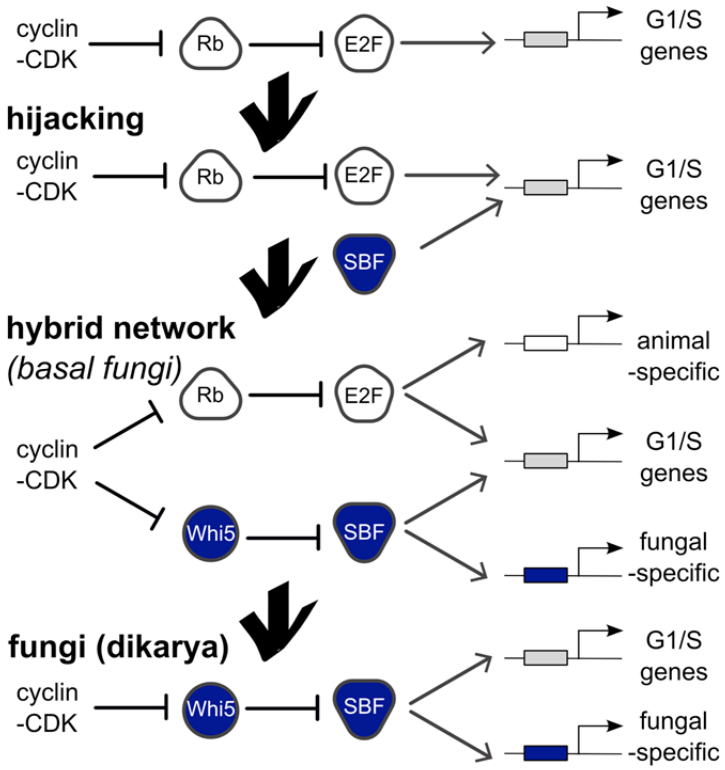


Figure 11. Punctuated evolution of a conserved regulatory network. Evolution can replace components in an essential pathway by proceeding through a hybrid intermediate. Once established, the hybrid network can evolve dramatically and lose previously essential regulators, while sometimes retaining the original network topology. We hypothesize that SBF may have hijacked the cell cycle of a fungal ancestor by binding cis-regulatory DNA sites of E2F and activating expression of G1/S genes, thus promoting cell cycle entry. Cell cycle hijacking in a fungal ancestor was followed by evolution of Whi5 to inhibit SBF and Whi5 was subsequently entrained to upstream cell cycle control through phospho-regulation by old or new cyclin-CDKs to create a hybrid network with parallel pathways. The hybrid network likely provided redundant control of the G1/S regulatory network, which could explain the eventual loss of E2F and its replacement by the SBF pathway in more derived fungi (Dikarya). Interestingly, Chytrids have hybrid networks and are transitional species because they exhibit animal-like features of the opisthokont ancestor (centrioles, flagella) and fungal-like features (cell wall, polarized growth). We hypothesize that E2F and SBF also bind and regulate a subset of animal-specific and fungal-specific G1/S genes, which could help explain the preservation of the hybrid network in Chytrids. Ancestral SBF expanded to create an entire family of transcription factors (APSES) that regulate fungal-specific traits such as sporulation, differentiation, morphogenesis, and virulence.

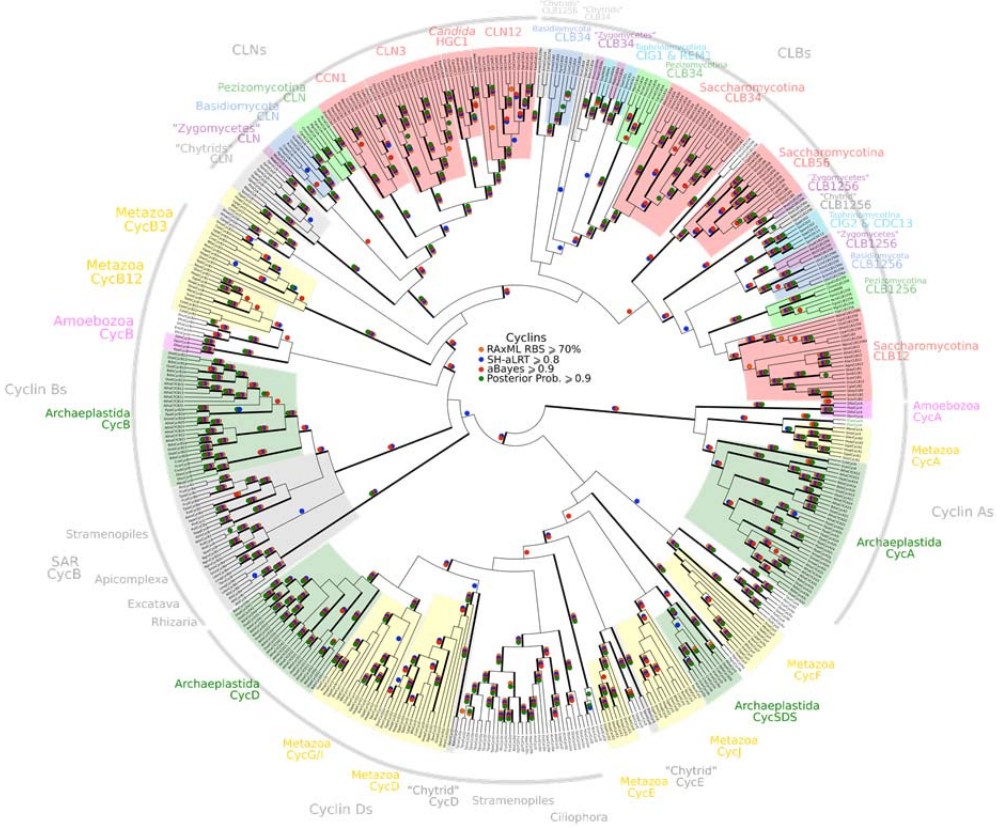


Figure 3 - supplement 1: Reduced phylogeny of eukaryotic cell cycle cyclins. The cell division cycle (CDC) cyclin family consists of several sub-families with a well-characterized cyclin box: (1) CycA (*H. sapiens, A. thaliana*), (2) CycB and CLB (*H. sapiens, A. thaliana, S. cerevisiae, S. pombe*), (3) CycD (*H. sapiens, A. thaliana*), (4) CycE (*H. sapiens*), and (5) CLN (*S. cerevisiae, S. pombe*). We combined CycA, CycB, CycD, CycE, CLB, and CLN sequences from *H. sapiens* (10 cyclins)*, A. thaliana* (13 cyclins)*, S. cerevisiae* (9 cyclins)*,* and *S. pombe* (5 cyclins) to create a eukaryotic CDC cyclin profile-HMM (pCYC.hmm) following the procedure outlined in the Methods. Our HMM profile was sensitive enough to discover known, but uncharacterized cyclin sub-families (CycO, CycF, CycG, CycI, CycJ, CycSDS) as *bona fide* CDC cyclins. A domain threshold of E-20 was used to identify potential CDC cyclin homologs. We first made a phylogeny of all cyclins to classify them. This dataset was then manually pruned to remove long-branches and problematic lineages. Our reduced CDC cyclin dataset has a total of 499 sequences. Columns with the top 10% Zorro score (496 positions) were used in our alignment. Confidence at nodes was assessed with multiple support metrics using different phylogenetic programs under LG model of evolution (aBayes and SH-aLRT metrics with PhyML, RBS with RAxML, Bayesian Posterior Probability with Phylobayes (23,163 sampled trees, meandiff= 0.01, maxdiff= 0.5)). Colored dots in branches indicate corresponding branch supports. Thick branches indicate significant support by at least two metrics, one parametric and one non-parametric; branch support thresholds are shown in the center of the figure; see Methods.

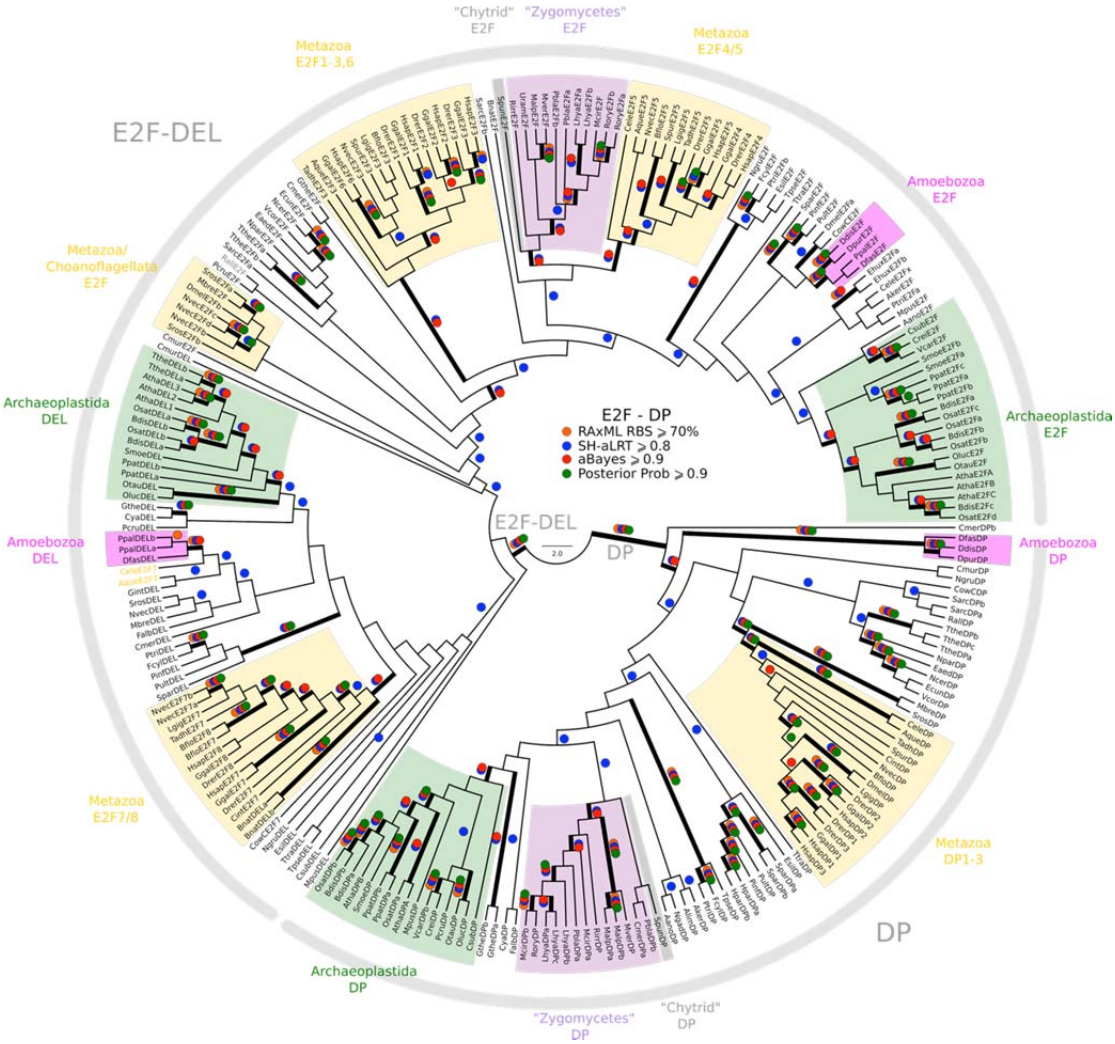


Figure 3 - supplement 2: Phylogeny of eukaryotic E2F/DP transcription factors. E2F-DP is a winged helix-turn-helix DNA-binding domain that is conserved across eukaryotes (van den Heuvel and Dyson, 2008). There are three sub-families within the E2F-DP family: (1) the E2F subfamily, (2) the E2F7-8/DEL subfamily, and (3) the DP subfamily. The E2F family consists of E2F1-6 (*H. sapiens*) and E2FA-C (*A. thaliana*). The E2F7-8/DEL family consists of E2F7-8 (*H. sapiens*) and DEL1-3 (*A. thaliana*). The DP family consists of DP1-3 (*H. sapiens*) and DPA-B (*A. thaliana*). The members of E2F form heterodimers with DP, whereas the DEL family has two DNA-binding domains and does not require DP to bind DNA. We used the E2F_TDP.hmm profile from PFAM to uncover members of the E2F/DP family across eukaryotes. A domain threshold of E-10 was used to identify potential E2F/DP homologs. Our E2F/DP dataset (pE2FDP.fasta) has 248 sequences. Columns with the top 8% Zorro score (284 positions) were used in our alignment. Confidence at nodes was assessed with multiple support metrics using different phylogenetic programs under LG model of evolution (aBayes and SH-aLRT metrics with PhyML, RBS with RAxML, Bayesian Posterior Probability with Phylobayes (53,009 sampled trees, meandiff=0.0064, maxdiff=0.18)). Colored dots in branches indicate corresponding branch supports. Thick branches indicate significant support by at least two metrics, one parametric and one non-parametric; branch support thresholds are shown in the center of the figure; see Methods.

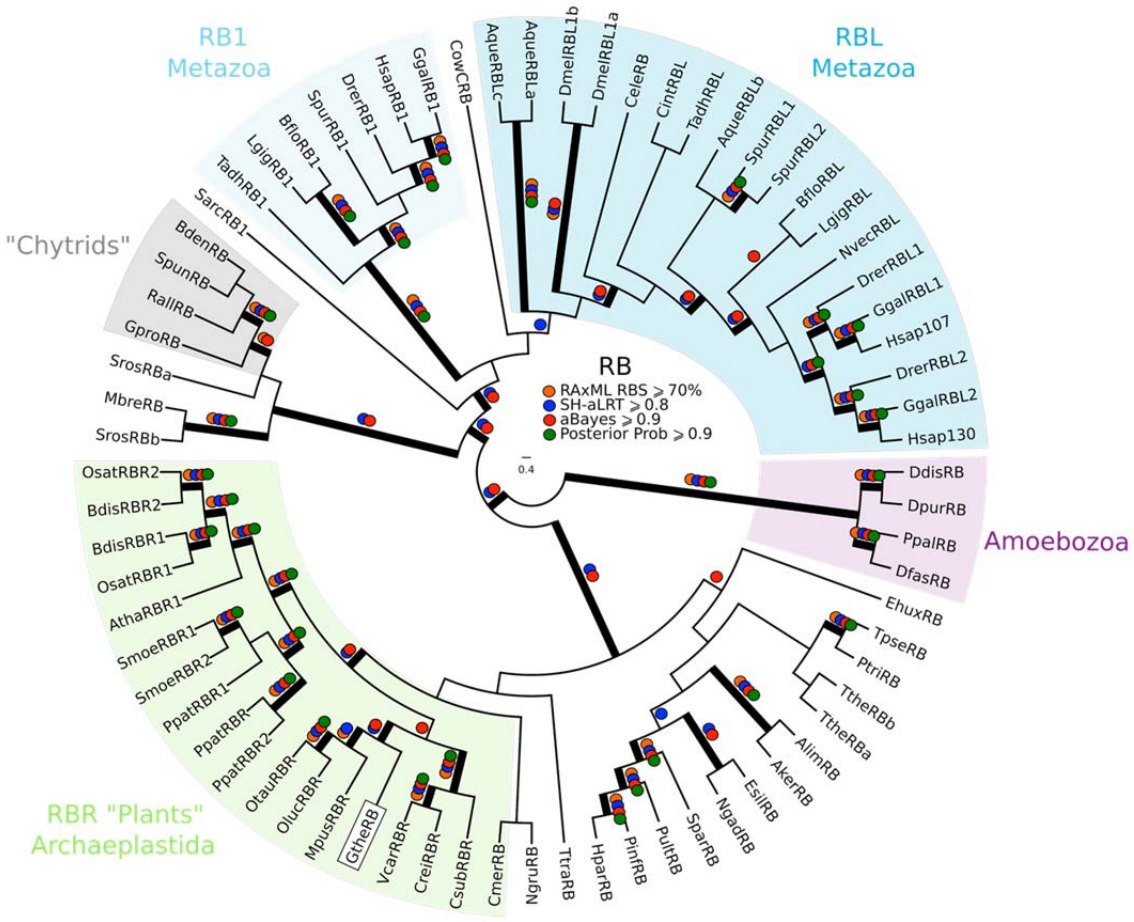


Figure 3 - supplement 3: Phylogeny of eukaryotic Rb inhibitors. *H. sapiens* has Rb1, RBL1 (p107), and RBL2 (p130), and *A. thaliana* has RBR1. The model fungi *S. cerevisiae* and *S. pombe* do not have any obvious retinoblastoma pocket proteins. We needed more eukaryotic sequences to create a robust HMM profile (pRb.hmm) for the pRb family. Based on the pRb sequences collected in (Hallmann, 2009), we built a profile-HMM using putative pRb homologs from *H. sapiens, G. gallus, C. intestinalis, D. melanogaster, C. elegans, N. vectensis, T. adhaerens* (metazoa)*; B. dendrobatidis* (fungi); *D. discoideum, D. purpureum, T. pseudonana, P. tricornutum, N. gruberi, E. huxleyi* (protists)*; C. merolae, O. tauri, O. lucimarinus, M. pusilla, V. carteri, C. reinhartdii, P. patens, S. moellendorfii, A. thaliana* (plants). A domain threshold of E-20 was used to identify pRB homologs. Our pRB dataset (pRb.fasta) has 72 sequences. Columns with the top 15% Zorro score (566 positions) were used in our alignment. Confidence at nodes was assessed with multiple support metrics using different phylogenetic programs under LG model of evolution (aBayes and SH-aLRT metrics with PhyML, RBS with RAxML, Bayesian Posterior Probability with Phylobayes (23,219 sampled trees, meandiff=0.0035, maxdiff=0.067)). Colored dots in branches indicate corresponding branch supports. Thick branches indicate significant support by at least two metrics, one parametric and one non-parametric; branch support thresholds are shown in the center of the figure; see Methods.

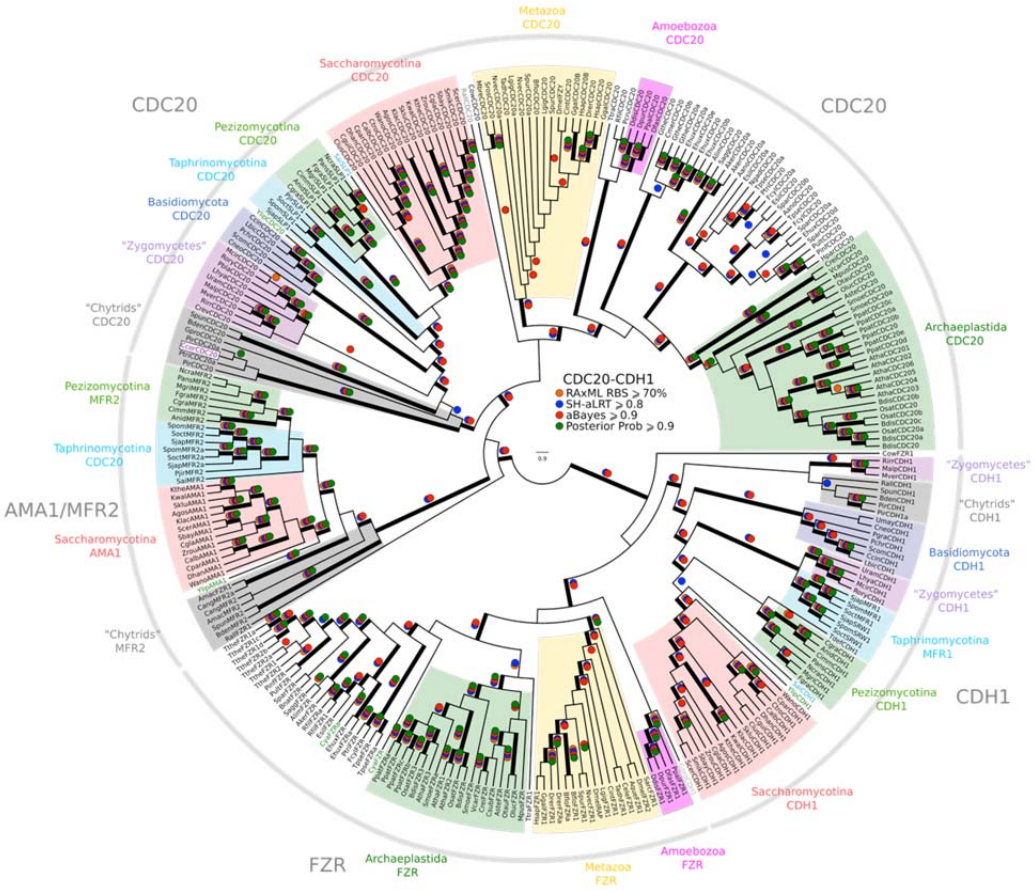


Figure 3 - supplement 4: Reduced phylogeny of eukaryotic Cdc20-family APC regulators. We combined CDC20 and CDH1/FZR1 sequences from *H. sapiens* (3 members), *A. thaliana* (9 members), and *S. cerevisiae* (3 members) to create a eukaryotic CDC20-family APC regulator profile-HMM (pCDC20.hmm) following the procedure outlined in the Methods. A domain threshold of E-50 was used to identify CDC20 homologs. Our pCDC20 dataset (pCDC20.fasta) has 350 sequences. This dataset was manually pruned to remove long-branches and problematic lineages. Our reduced CDC20 dataset has a total of 289 sequences. Columns with the top 20% Zorro score (608 positions) were used in our alignment. Confidence at nodes was assessed with multiple support metrics using different phylogenetic programs under LG model of evolution (aBayes and SH-aLRT metrics with PhyML, RBS with RAxML, Bayesian Posterior Probability with Phylobayes (13,638 sampled trees, meandiff=0.015, maxdiff=0.37)). Colored dots in branches indicate corresponding branch supports. Thick branches indicate significant support by at least two metrics, one parametric and one non-parametric; branch support thresholds are shown in the center of the figure; see Methods.

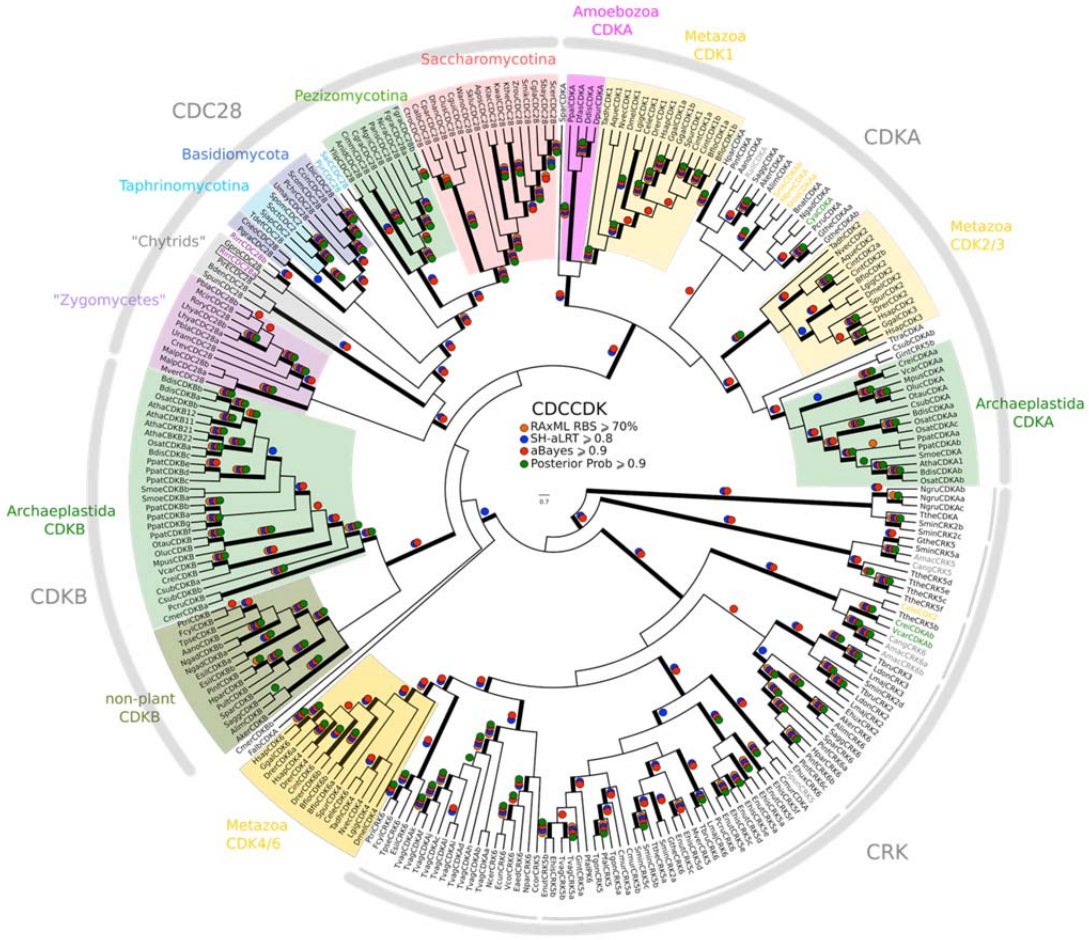


Figure 3 - supplement 5: Phylogeny of eukaryotic cyclin-dependent kinases. To create a profile-HMM (pCDCCDK.hmm) for eukaryotic cell cycle CDK, we combined Cdk1-3, Cdk4, Cdk6 sequences from *H. sapiens,* CdkA and CdkB from *A. thaliana,* Cdc28 from *S. cerevisiae,* and Cdc2 from *S. pombe*. A domain threshold of E-20 was used to identify CDK homologs. Our cell cycle CDK dataset (pCDCCDK.fasta) has 272 sequences. Columns with the top 15% Zorro score (473 positions) were used in our alignment. Confidence at nodes was assessed with multiple support metrics using different phylogenetic programs under LG model of evolution (aBayes and SH-aLRT metrics with PhyML, RBS with RAxML, Bayesian Posterior Probability with Phylobayes (28,193 sampled trees, meandiff=0.015, maxdiff=0.53)). Colored dots in branches indicate corresponding branch supports. Thick branches indicate significant support by at least two metrics, one parametric and one non-parametric; branch support thresholds are shown in the center of the figure; see Methods.

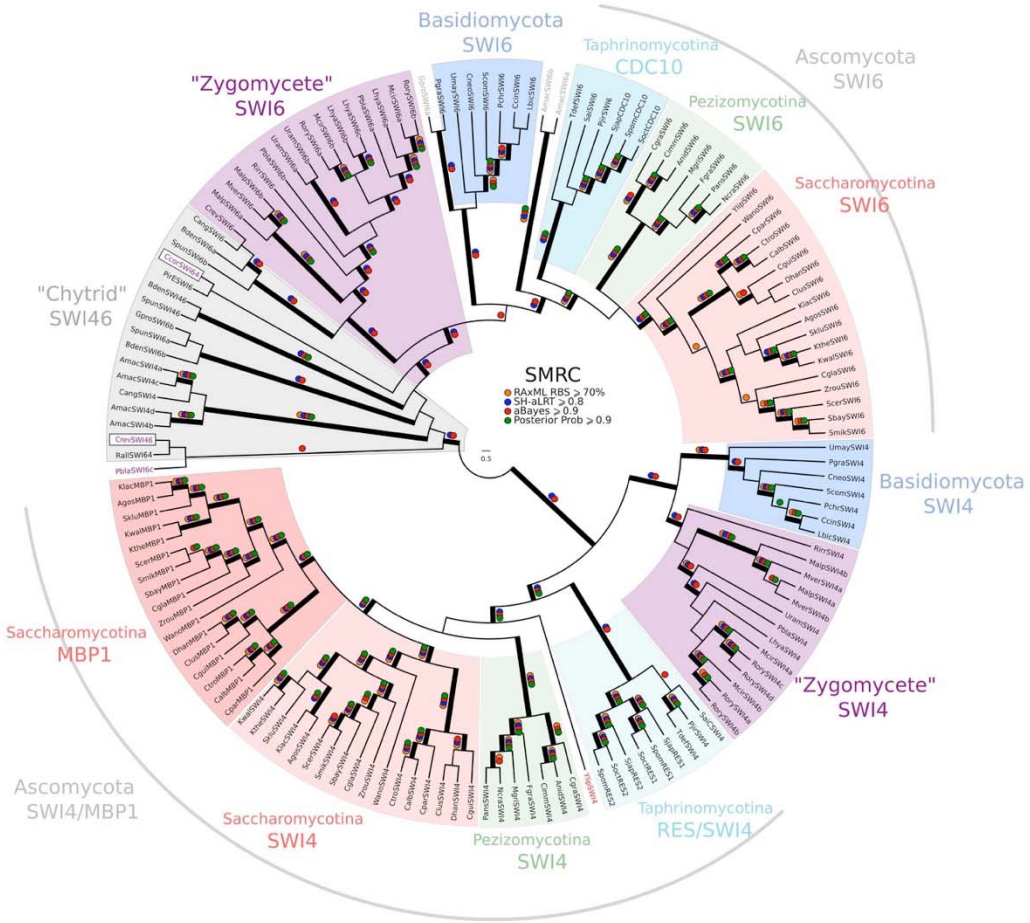


Figure 5 - supplement 1: Phylogeny of fungal SBF transcription factors. SBF and MBF are transcription factors that regulate G1/S transcription in budding and fission yeast. To detect SMRC (Swi4/6 Mbp1 Res1/2 Cdc10) across fungi, we built a sensitive profile-HMM (pSMRC.hmm) by combining well-characterized SMRC sequences from *S. cerevisiae, C. albicans, N. crassa, A. nidulans,* and *S. pombe*. A domain threshold of E-10 was used to identify SMRC homologs. Our SMRC dataset (pSMRC.fasta) has 147 sequences. Columns with the top 20% Zorro score (709 positions) were used in our alignment. Confidence at nodes was assessed with multiple support metrics using different phylogenetic programs under LG model of evolution (aBayes and SH-aLRT metrics with PhyML, RBS with RAxML, Bayesian Posterior Probability with Phylobayes (19,457 sampled trees, meandiff=0.0056, maxdiff=0.145)). Colored dots in branches indicate corresponding branch supports. Thick branches indicate significant support by at least two metrics, one parametric and one non-parametric; branch support thresholds are shown in the center of the figure; see Methods.

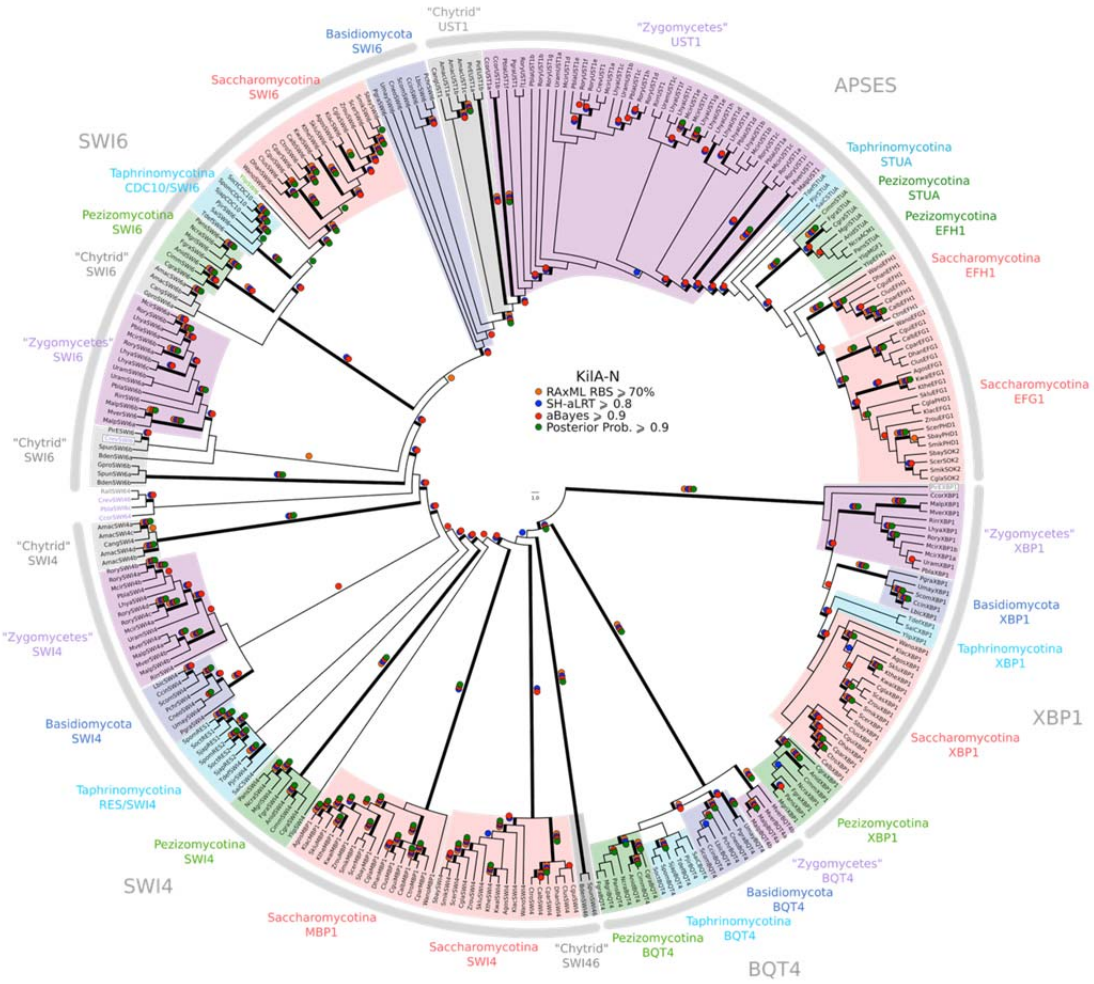


Figure 5 - supplement 2: Phylogeny of fungal SBF and APSES transcription factors. SBF and APSES transcription factors (Asm1, Phd1, Sok2, Efg1, StuA) share a common DNA-binding domain (KilA-N), which is derived from DNA viruses. During our search for SBF and APSES homologs, we consistently detected two additional fungal sub-families with homology to KilA-N: XBP1 (family name taken from *S. cerevisiae*) and BQT4 (family name taken from *S. pombe*). To detect APSES, XBP1, and BQT4 homologs, we built profile-HMMs (APSES.hmm, XBP1.hmm, and BQT4.hmm) by combining APSES, XBP1, and BQT4 homologs from *S. cerevisiae, C. albicans, N. crassa, A. nidulans,* and *S. pombe*. A domain threshold of E-10 was used to identify APSES, XBP1, and BQT4 homologs. Our final dataset (pKILA.fasta) contains all fungal KILA sub-families (SBF, APSES, XBP1, BQT4) and has a total of 301 sequences. Columns with the top 10% Zorro score (447 positions) were used in our alignment. Confidence at nodes was assessed with multiple support metrics using different phylogenetic programs under LG model of evolution (aBayes and SH-aLRT metrics with PhyML, RBS with RAxML, Bayesian Posterior Probability with Phylobayes (15,251 sampled trees, meandiff=0.012, maxdiff=0.25)). Colored dots in branches indicate corresponding branch supports. Thick branches indicate significant support by at least two metrics, one parametric and one non-parametric; branch support thresholds are shown in the center of the figure; see Methods.

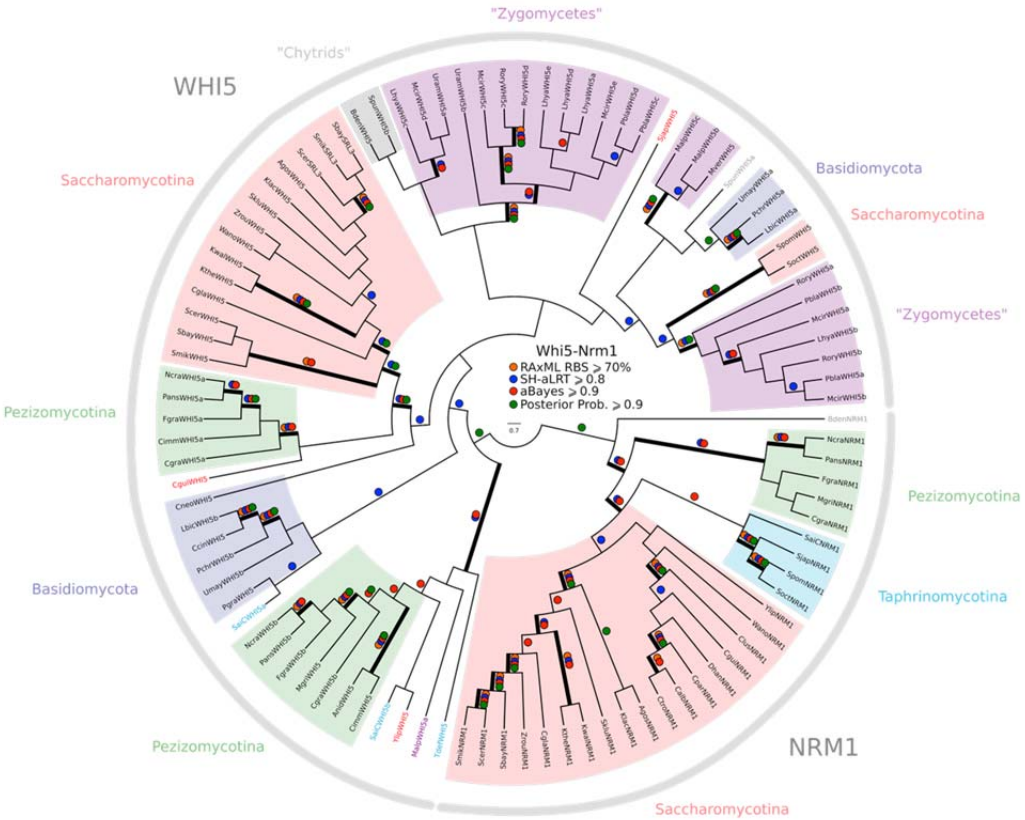


Figure 5 - supplement 3: Phylogeny of fungal Whi5 inhibitors. WHI5 and NRM1 are a yeast-specific protein family that has been identified and functionally characterized across *S. cerevisiae*, *C. albicans*, and *S. pombe.* Both WHI5 and NRM1 are fast evolving proteins. There is a small conserved region of 25 amino-acids (known as the GTB domain) that is responsible for interacting with Swi6/Cdc10 (Travesa et al., 2013). Unfortunately, the Whi5.hmm profile from PFAM is unable to detect an SRL3 paralogue in *S. cerevisiae* or the NRM1 orthologues in *A. gossypii* or *C. albicans.* We built a more sensitive profile-HMM (pWHI5.hmm) by combining WHI5/NRM1 sequences across ascomycetes (including SRL3 from Saccharomyces genomes and NRM1 from Candida genomes). A domain threshold of E-05 was used to identify WHI5 homologs. Our WHI5 dataset (pWHI5.fasta) has 98 sequences. Columns with the top 15% Zorro score (260 positions) were used in our alignment. Confidence at nodes was assessed with multiple support metrics using different phylogenetic programs under the LG model of evolution (aBayes and SH-aLRT metrics with PhyML, RBS with RAxML, Bayesian Posterior Probability with Phylobayes (77,696 sampled trees, meandiff=0.0068, maxdiff=0.11)). Colored dots in branches indicate corresponding branch supports. Thick branches indicate significant support by at least two metrics, one parametric and one non-parametric; branch support thresholds are shown in the center of the figure; see Methods.

Supplementary File 1A: List of eukaryotic genomes. We downloaded and analyzed the following annotated genomes using the "best" filtered protein sets when available. We gratefully acknowledge the Broad Institute, the DOE Joint Genome Institute, Génolevures, PlantGDB, SaccharomycesGD, AshbyaGD, DictyBase, JCV Institute, Sanger Institute, TetrahymenaGD, PythiumGD, AmoebaDB, NannochloroposisGD, OrcAE, TriTryDB, GiardiaDB, TrichDB, CyanophoraDB, and CyanidioschizonDB for making their annotated genomes publicly available. We especially thank D. Armaleo, I. Grigoriev, T. Jeffries, J. Spatafora, S. Baker, J. Collier, and T. Mock for allowing us to use their unpublished data.

Supplementary File 1B: Plasmids.

Supplementary File 1C: Strains. All yeast strains were derived from W303 and constructed using standard methods.

Supplementary File 1D: Protein sequences of cell cycle regulators. All protein sequences from different genome that were used to create molecular phylogenies can be downloaded as FASTA files.