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

Authors: Edgar M. Medina¹,², Jonathan J. Turner³, Jan M. Skotheim³, and Nicolas E. Buchler¹,²

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

Contact: nicolas.buchler@duke.edu

Abstract: (130 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. 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 viral 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 (15-30 words): Cell cycle network evolution in a fungal ancestor was punctuated by arrival of a viral DNA-binding protein that was permanently incorporated into the G1/S regulatory network controlling cell cycle entry.

Keywords: cell cycle / evolution / fungi / chytrid / virus
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. While 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 indicates 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.

Here, 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. 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. The LECA repertoire helps establish that the emergence of SBF-Whi5 is abrupt and distinguishes fungi from all other eukaryotes.
We also 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, but not the individual components, has been conserved through the transition to fungi and argues against the case of convergent evolution.

Finally, 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. That SBF can still regulate promoters with E2F binding sites in budding yeast suggests 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 proteins (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 LECA cell cycle 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,
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 infer 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.

We did not find either of the fungal regulators, SBF and Whi5, in any Archaeplastida, Amoebozoa, SAR, Haptophyta, Cryptophyta, Excavata or Metazoa. However, we did find the cyclin sub-families (A, B, D, and E) known to regulate the cell cycle in metazoans across all major branches of eukaryotes (for cyclin phylogeny see Figure 3-figure supplement 1). 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.

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 Cdns (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 suggested by (Cavalier-Smith, 1999). We excluded from our analysis other families of cyclin-Cdns including 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 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 one exception that we discuss in the next section. 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 and 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". At one extreme, Chytridiomycota (e.g., Spizellomyces punctatus) can have both fungal and animal cell cycle regulators, which likely represents the ancestral hybrid network. However, this hybrid network was evolutionarily unstable, as 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. The instability of the hybrid network suggests some functional redundancy between the new SBF-Whi5 and ancient E2F-Rb pathway.

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

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 sequence, overall structure, and
mode of protein-DNA complex formation that lead us to conclude that it is highly unlikely that they are orthologs.

The strongest arguments against SBF-E2F orthology are based on structural biology. 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 6A: 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 6B, left). In contrast, the first solved SBF/MBF 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). A recent structure of PCG2, a 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 6A, 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 6B, 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 therefore sought an alternative evolutionary origin. SBF is a member of a larger family of transcription factors in yeast that includes Xbp1, Bqt4, and the APSES family (Acm1, Phd1, Sok2, Egf1, StuA); see Figure 5. 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.

We were unable to find ancestral SBF-like proteins in other eukaryotes—with the notable exception of *Trichomonas vaginalis*, a parasitic excavate that has over 100 homologs. These potential SBF homologs were compared to the Uniprot protein database, and their best sequence matches were to Mimivirus, a large double-stranded DNA virus of the Nucleo-Cytoplasmic Large DNA Viruses (Yutin et al., 2009). Upon closer inspection, both fungal SBF sequences and *T. vaginalis* Mimivirus sequences share the viral KilA-N domain. To gain insight into the possible evolutionary origins of the SBF subfamily, we aligned all KilA-N sequences from the Uniprot and PFAM database (Finn et al., 2014) to all our fungal SBF sequences and built a phylogenetic tree (Figure 7). There are three major phylogenetic groups of KilA-N domains: those found in eukaryotic viruses, prokaryotic viruses, and the fungal SBF family. 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. However, our phylogeny does show that *T. vaginalis* obtained its KilA-N proteins independently from the fungal ancestor because its KilA-N domains are clearly placed within the eukaryotic DNA viruses (e.g., Mimivirus) rather than with the fungal subfamily (Figure 7).

**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. 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). 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. 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 test whether yeast SBF can bind a canonical E2F binding site, we inserted consensus E2F binding sites in the budding yeast genome; see Figure 8. 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. Removal of these SCB sites is known to eliminate cell cycle-dependent gene expression (Bai et al., 2010). We changed all three SCBs to the E2F binding site variant GCGCGAAA known to regulate 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, but not MBF, and disappeared when the 3 binding sites were removed. Taken together, this set of in vivo experiments lends support to the hijacking hypothesis, where an ancestral SBF took 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 central 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 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 9).

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 \( \alpha \) cells and a/\( \alpha \) 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 have preserved rather than lost this hybrid intermediate, which suggests that
each pathway may have specialized functions. Chytrids exhibit both animal (e.g. centrioles, flagella, amoeboid movement) and fungal features (e.g. cell walls, hyphal morphology) whose synthesis needs to be coordinated with cell division. The preservation of the hybrid network in chytrids could then be explained if animal and fungal features are regulated by the E2F-Rb and SBF-Whi5 pathways respectively.

The origin of the hybrid network at the base of Fungi is abrupt and may have been initiated by the arrival of SBF via virus. Many tumor viruses activate cell proliferation. For example, the DNA tumor viruses Adenovirus and SV40 highjack cell proliferation in part by activating the expression of E2F-dependent genes by binding pRb to disrupt inhibition of E2F (DeCaprio, 2009). While the specific mechanisms may differ, when SBF entered the fungal ancestor cell it might have activated the transcription of E2F target genes. Rather than inhibiting the inhibitor of E2F, SBF may have directly competed for E2F binding sites with transcriptionally inactive Rb-E2F complexes (Figure 9). Consistent with this model, we have shown here that SBF can directly regulate gene expression in budding yeast via a consensus E2F binding site. Thus, the cooption of a viral protein generated a hybrid network to ultimately facilitate dramatic evolution of the core cell cycle network in fungi.

**MATERIALS AND METHODS**

**Identification of potential protein family homologs.**

We used Profile-Hidden Markov Models (profile-HMMs) to detect homologs for each of the families studied, using the HMMER 3 package (Eddy, 2011). Profile-HMMs are sensitive tools for remote homology detection. Starting with a set of diverse yet reliable protein homologs is fundamental for detecting remote protein homology and avoiding “model poisoning” (Johnson et al., 2010). To this end, we used reliable training-set homologs from the cell cycle model organisms *Arabidopsis thaliana*, *Homo sapiens*, *Schizosaccharomyces pombe*, and *Saccharomyces cerevisiae*, to build the profile-HMMs used to detect homologs. Our profile-HMM search used a stringent e-value threshold of 1e-10 to detect putative homologs in the "best" filtered protein sets (where available) of our 100+ eukaryotic genomes (see Supplementary File 1A 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, whereas 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 different. 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 new Figure 6B) and all proteins that we classified as SBF/MBF had clear homology to SBF/MBF (see alignment in new Figure 6B). Thus, we have not misclassified E2Fs or SBF/MBFs in basal fungi and we have not reached the limit of homology detection.

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 (-maxiter 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 (Rambaut et al., 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.

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 via 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.

**ACKNOWLEDGEMENTS**

We thank F. Cross and A. Robinson-Mosher for extensive discussions, and R. Gordan, 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:**


Bioinformatics 25, 2286–2288.


CLN1, CLN2, and a putative new G1 cyclin (HCS26) by SWI4, a positive regulator of G1-specific transcription. Cell 66, 1015–1026.


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.
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. See Figure 5 for a complete list of components in all fungal species analyzed.
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 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 7. Viral origin of yeast cell cycle transcription factor SBF. Maximum Likelihood unrooted phylogenetic tree depicting relationships of fungal SBF-family proteins (blue), KilA-N domains in prokaryotic (black, green), and eukaryotic DNA viruses (orange, red). The particular origin of the HGT event that originated the SBF/APSES family is unclear. We found no significant rapid bootstrapping support (RBS ≥80%) for internal branches. Inference was performed under GTR+G model of evolution. Scale bar in substitutions per site; (See Methods).
Figure 8. Yeast cell cycle transcription factor SBF can regulate cell cycle-dependent transcription via E2F binding sites (A) Fluorescence images of cells expressing a destabilized GFP from the SBF-regulated CLN2 promoter. (B) 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 consensus E2F binding sites (E2F). Oscillation amplitudes were quantified by scaling the mean fluorescence intensity difference from peak to trough divided by the trough intensity \( \frac{S_{\text{max}} - S_{\text{min}}}{S_{\text{min}}} \). Circles denote time points corresponding to (b). Triangles denote budding events. (C) 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 25th, median and 75th percentiles, while whiskers extend to 1.5 times this interquartile range.
Figure 9. 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.
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. reinhardtii*, *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 SL-arLT 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 $E=0.05$ 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.