Identification and *in silico* analysis of the origin recognition complex in the human fungal pathogen *Candida albicans*

Sreedevi Padmanabhan[1], Kaustuv Sanyal[2*], Dharani Dhar Dubey[1*]

[1]Molecular Biology Laboratory, Department of Biotechnology, Veer Bahadur Singh Purvanchal University, Jaunpur 222 003, Uttar Pradesh, India, [2]Molecular Mycology Laboratory, Molecular Biology and Genetics Unit, Jawaharlal Nehru Centre for Advanced Scientific Research, Jakkur, Bangalore 560 064, India.

* Corresponding author

Email address: sanyal@jncasr.ac.in; dddubey2003@gmail.com

Telephone number: 91-80-22082878; 91-9453362949

Fax number: 91-80-22082766

Abstract

DNA replication in eukaryotes is initiated by the orchestrated assembly and association of initiator proteins (heterohexameric Origin Recognition Complex, ORC) on the replication origins. These functionally conserved proteins play significant roles in diverse cellular processes besides their central role in ignition of DNA replication at origins. While *Candida albicans*, a major human fungal pathogen, is an ascomycetous, asexual, diploid budding yeast but it is significantly diverged from a much better studied model organism *Saccharomyces cerevisiae.* The components of the DNA replication machinery in *C. albicans* remain largely uncharacterized. Identification of factors required for DNA replication is essential for understanding the evolution of the DNA replication machinery. We identified the putative ORC homologs in *C. albicans* and determined their relatedness with those of other eukaryotes including several yeast species. Our extensive *in silico* studies demonstrate that the domain architecture of CaORC proteins share similarities with the ORC proteins of *S. cerevisiae*. We dissect the domain organization of ORC (trans-acting factors) proteins that seem to associate with DNA replication origins in *C. albicans.* We present a model of the 3D structure of CaORC4 to gain further insights of this protein's function.

Introduction

DNA replication in eukaryotes is initiated by the orchestrated assembly and association of initiator proteins on the replication origins. The hunt for initiator proteins in higher eukaryotes picked up pace after the discovery of the Origin Recognition Complex (ORC) comprising of six protein subunits of the ORC1-6 complex in budding yeast[1]. Extensive studies in other organisms showed that the initiator ORC proteins are functionally conserved in all eukaryotes and the association of ORC proteins with DNA replication origins is critical for initiation of DNA replication, a fundamental process of life. The replicators, occupied by ORC proteins, fire asynchronously in S phase. Replicators in different organisms have widely variable DNA sequence requirements. In some organisms no obvious DNA sequence requirements could be detected. The sequential assembly of the pre-replication complex (pre-RC) proteins on the origins is mediated by ORC. ORC orthologs have been identified in many eukaryotes like *Schizosaccharomyces pombe, Drosophila melanogaster, Xenopus laevis,* and humans. Genetic and biochemical investigations demonstrate the ORC proteins of these organisms to be essential for DNA replication initiation[2]. Although the replication origins in higher eukaryotes do not share a consensus sequence as in bacteria or budding yeast, the proteins that are recruited to origins in most metazoans are similar to those in bacteria and yeast suggesting replication-associated proteins are evolutionarily conserved[3-4]. ORC-mediated ATP hydrolysis is essential for recruiting MCM (Mini Chromosome Maintenance) proteins[5] which subsequently act as helicases and unwind DNA double helix to facilitate initiation of DNA replication.

Besides playing a central role in DNA replication initiation at discrete origin sites, ORC proteins are also involved in a variety of cellular processes like heterochromatin formation, transcriptional regulation, S-phase checkpoint regulation, mitotic chromosome assembly, sister chromatid cohesion, cytokinesis, ribosome biogenesis and tissue specific gene regulation. ORC mutations are seen in various human diseases[6] such as Meier–Gorlin syndrome[7, 8], EBV (Epstein–Barr virus)-infected diseases[9], American trypanosomiasis and African trypanosomiasis[10].

There has been a wide prevalence of yeast infections over the years with *Candida* species that can cause superficial to fatal systemic infections. These fungal infections can be fatal for immunocompromised individuals where the mortality rate is even higher[11]. Availability of a handful antifungal drugs and frequent isolation of drug-resistant isolates led to complications in disease management and treatment procedures[12]. Hence, to circumvent this malicious infection is to find species-specific new drug targets to develop more effective and safer antifungals. *C. albicans* is one such opportunistic fungal pathogens which is an asexual, diploid, budding yeast[13-14]. Protein complexes involved in the DNA replication of *C. albicans* are not characterized. As DNA replication is a rate limiting step in the propagation of the yeast and not many anti-fungal drugs are available to curb *Candida* infection, in this study we sought to identify and dissect the domain architecture of CaORC proteins with an aim to provide clues on their evolutionary conservation/diversification across various species to explore their potential as species-specific drug targets.

Results

Identification of preRC genes in *C. albicans* genome by *in silico* analysis

First, we identified the homologs of the preRC complex in *C. albicans*, determined relatedness of these proteins present in other species, compared the various domains such as the BAH domain, AAA+, AT-hook and Walker motifs in the ORC proteins of a number of species and predicted the structure of ORC4 in *C. albicans*, CaORC4. The CaORC1-6 genes were identified by a BLAST search with *S. cerevisiae* ScORC1-6 as the query sequences against the *C. albicans* genome database (CGD) (http://www.candidagenome.org/cgi-bin/compute/blast-sgd.pl) (Table 1 and Table 2).

ClustalW2 is a DNA or protein multiple sequence alignment program for multiple sequences[15]. We performed a pair-wise amino acid sequence alignment of the CaORC proteins with those of *S. cerevisiae, S. pombe*, *Drosophila, Xenopus*, Mouse and humans individually and their respective clustalW scores are tabulated (Table 3). Although, in general, the CaORC proteins show limited sequence similarities with their counterparts in various species, CaORC1, 2 and 6 show maximum similarities to their *S. cerevisiae* counterparts while CaORC3, 4 and 5 appear to be more similar to those of mammals which is evident from the phylogenetic map (Figure 1A).

BAH domain in ORC proteins

The Expasy PROSITE consists of documentation entries describing protein domains, families and functional sites as well as associated patterns and profiles to identify them. The Expasy PROSITE tool predicts the presence of an evolutionarily conserved BAH_domain spanning the region between 44th and 179th amino acids at the N-terminal of CaORC1 (Figure 1B). The BAH domain is involved in protein-protein interactions and has been found to be important in DNA methylation, replication and transcriptional regulation[16].

AAA+ domains in CaORC proteins

ATPases associated with various cellular activities (AAA+) domains[17-18] are those that are activated by ATP binding and inactivated by ATP hydrolysis[19-23]. The ATPase activity is indispensable for the origin-ORC association and henceforth for the establishment of the pre-initiation complex. Preventing ORC ATP hydrolysis inhibits repeated MCM2-7 loading[5]. CaORC1 and CaORC4, each contains a consensus AAA+ domain (420-571 a.a. in CaORC1; 139-318 a.a. in CaORC4) (Figure 1B, 1C and 1D), which belongs to the AAA+ family that is pivotal to the initiation of eukaryotic DNA replication. There is an amino acid residue Tyr^{174} in human ORC4 (Tyr^{232} in *S. cerevisiae*) that is found between the Walker B motif and sensor I of the AAA+ domain which may be responsible for interacting with a conserved arginine residue on an adjacent helix structure of ORC4[2, 6, 21,23-26]. This residue is present in CaORC4 (Tyr^{273}) too probably doing a similar function.

Although the ScORC1-ORC5 all have AAA+ domains, there is a variation among the subunits with respect to the catalytic core motifs within the Walker A and B motif regions both within and between the species. It is reported with experimental evidence that only the ScORC1 and ScORC5 can bind ATP, of which only ScORC1 has a perfect signature to the Walker B motif. By consensus, it is considered that the ORC1 would be the prime ATPase of all the eukaryotes examined so far[25, 26]. In metazoans too, although the ORC1, ORC4 and ORC5 bind ATP, the Walker A signature is found to have perfect match with ORC4. A similar pattern is observed in CaORC proteins too demonstrating their close homology with the higher eukaryotes. The Walker A motifs in ORC5 seem to be diverged (Table 4).

Walker A and B motifs in CaORC proteins

The motif GXXXXGKT (X, any residue) is a common nucleotide binding fold in the α- and β-subunits of F1-ATPase, myosin and other ATP-requiring enzymes[27]. This motif is present in the shape of a loop around nucleotides and utilizes its highly conserved residues of lysine and threonine to bind to their phosphate oxygen atoms. This consensus sequence of GXXXXGKT(S), with serine substituting threonine in some cases, is more popularly known as the Walker loop or P-loop (phosphate binding loop). The Walker B motif with the consensus sequence hhhhDE (a negatively charged residue followed by a stretch of hydrophobic a.a,,h) is essential for ATP hydrolysis. The Walker motifs are present in CaORC1, CaORC4 and CaORC5

(Walker B is absent in CaORC5) (Table 4). Besides CaORC1, the perfect signature of the Walker motif is found in CaORC4 with a putative Walker A motif (147-153 a.a) and a putative Walker B motif (410-426 a.a) the amino acid sequences for which are shown in Table 5. These motif signatures seem to be more closely related to the metazoan/higher eukaryotic sequences.

AT-hook motifs in CaORC proteins

AT hooks are DNA-binding motifs with a preference for A/T rich regions. These motifs are found in a variety of proteins, including the high mobility group (HMG) proteins. The AT-hook is a small motif which has a typical sequence pattern centered on a glycine-arginine-proline (GRP) tripeptide [28,29]. The importance of this short conserved sequence is that it is necessary and sufficient for binding DNA and ori-ORC association. CaORC2 is found to have an AT-hook motif (182-197a.a, Figure 1D, Table 6) indicative of its propensity to bind origins.

PIP motif in CaORC proteins

A conserved Proliferating Cell Nuclear Antigen (PCNA) binding motif called the PCNA-interacting protein (PIP) box (QXXMXXFFFY) is found in the CaORC1 protein (524-536 a.a). Of the CaORC proteins, the PIP box is found to be unique to CaORC1.

MOD1 motif in CaORC3

Two independent domains in human ORC3, a coiled-coil domain at the N terminus and a second region containing a MOD1-interacting region (MIR; 213-218aa)[30], were found to be directly bound to the heterochromatin protein, HP1α[31]. A conserved peptide motif named MIR (MOD1 interacting region - PXVHH) which is essential for their interaction with MOD1, a serotonin-gated chloride channel that modulates locomotory behavior in *C. elegans*[32] is found in CaORC3 protein (435-448 a.a).

Although the CaORC proteins share less amino acid sequence homology with the ORC proteins of *S. cerevisiae* and *S. pombe*, the other ORC associated proteins (MCM proteins) seem to have higher homology (Table 7). Interestingly, the predicted molecular weights of the ORC complexes are equal in these three yeasts (Table 8).

PEST motif in CaORC proteins

A PEST sequence is a peptide sequence that is rich in proline (P), glutamic acid (E), serine (S), and threonine (T). This sequence is associated with proteins that have a short intracellular half-life; hence, it is hypothesized that the PEST sequence acts as a signal peptide for protein degradation. CaORC2 (130-172 a.a.) and CaORC3 (6-33 a.a.) contain PEST motif. Analysis of PEST signals in human and mouse ORC proteins suggests that only ORC1 is targeted for ubiquitination which is likely to hold good for all mammals[33]. The domains of CaORC proteins are compared with other eukaryotes and are tabulated in Table 6 and compared with *S. cervevisiae* in Figure 1D. Recent studies have shown the evolution of the phospho regulation pattern in replication proteins of various yeast species including *C. albicans*[34].

Evolutionary relationships of ORC proteins

Molecular Evolutionary Genetics Analysis (MEGA) is an integrated tool for conducting sequence alignment, inferring phylogenetic trees, estimating divergence times, mining online databases, estimating rates of molecular evolution, inferring ancestral sequences, and testing evolutionary hypotheses[35]. The evolutionary history was inferred using the Neighbor-Joining method[36]. The optimal tree with the sum of branch length = 29.06450731 is shown in Figure 1A. The ORC1, ORC2 and ORC5 proteins from yeast to humans are found to have common nodes. Subsequently, the ORC proteins from the related species of *C. albicans* in the CTG clade were also compared and a phylogenetic tree was constructed (Figure 2A). The time tree demonstrates the diversification rate of these ORC proteins across the species of which ORC1 and ORC4 seem to be older than their counterparts (Figure 2B). In order to understand the sequence identity of the ORC sequences across various yeast species, Sequenceserver (http://blast.wei.wisc.edu/)[37,38] was used across 86 publicly available yeast genomes (Figure 3; Supplementary Table 1).

Structure prediction of CaORC proteins

Prediction of secondary structure using Phyre

Over the past few decades, a number of computational tools for protein structure prediction have been developed. The **p**rotein **h**omolog**y**/analogy **r**ecognition **e**ngine (**Phyre**) is one of the widely used structure prediction systems providing a simple interface to results. The Phyre server (http://www.imperial.ac.uk/phyre) uses a library of known protein structures taken from the Structural Classification of Proteins (SCOP) database[39] and augmented with newer depositions in the Protein Data Bank (PDB)[40]. The sequence of each of these structures is scanned against a non-redundant sequence database and a profile is generated and deposited in the 'fold library'. The known and predicted secondary structure of these proteins is also stored in the fold library. The popular web servers for fold recognition are Phyre, I-TASSER, SAM-T06, HHpred.

We used I-TASSER (Iterative Threading ASSEmbly Refinement[41]) for structure prediction of CaORC proteins. Of all the CaORC proteins, CaORC4 was found to be one of the putative candidates for further fine refinement studies of the protein structure due to its higher Cscore (combined measure, See Methods section) which indicates a better confidence in predicting the function using the template (Table 9). Hence, we proceeded for predicting the structure of CaORC4 using Phyre.

Secondary structure and disorder prediction for CaORC4

The query sequence (CaORC4p) is scanned against the non-redundant sequence database and a profile is constructed. Five iterations of PSI-BLAST are used to gather both close and remote sequence homologs. The PSI-BLAST provides a means of detecting distance relationships between proteins. The pair-wise alignments generated by PSI-BLAST are combined into a single alignment with the query sequence as the master. The secondary structure of CaORC4p is predicted following profile construction.

Three independent secondary structure prediction programs are used in Phyre: Psi-Pred1[42], SSPro[43] and JNet[44]. The output of each program is in the form of a three-state prediction: alpha helix, beta strand and coil. Each of these three programs provides a confidence value at each position of the query for each of the three secondary structure states. These confidence values are averaged and a final, consensus prediction is calculated and displayed beneath the individual predictions.

Fold recognition for CaORC4

The profile and secondary structure of CaORC4 is then scanned against the fold library using a profile–profile alignment algorithm detailed in[45]. This alignment process returns a score on which the alignments are ranked. These scores are fitted to an extreme value distribution to generate an E-value. The top ten highest scoring alignments are then used to construct full 3D models of the CaORC4p (Figure 4A and Figure 4B).

Interactions of pre-RC proteins – SMART prediction

SMART (Simple Modular Architecture Research Tool) is a web-based tool (http://smart.embl.de/) that allows rapid identification and annotation of protein domains and the analysis of protein domain architectures. This provides the complete set of protein descriptions allowing users to quickly find relevant information[46-47]. The predicted functional partners of the preRC proteins in *C. albicans* are enlisted in the Table 10 and are also shown schematically in the Figure 4C-H. In short, it is evident that although the size of the individual proteins in the ORC complex across diverse yeast species is varied, the whole complex constitutes to ~412 KDa (Table 8). Our *in silico* analysis suggests that although CaORC proteins share less sequence homology with yeasts, Drosophila, Xenopus, mouse and humans (Table 7), some of the characteristic functional motifs are retained in them (Figure 1D, Table 6). CaORC1 is found to have the BAH domain and the PIP motif, CaORC2 has an AT-hook motif, and CaORC3 has a MOD1-interacting region (MIR). The AAA ATPase is found in CaORC1 and CaORC4 and the PEST motif is found in CaORC2 and CaORC3. We used Phyre to predict the secondary structure and modeled the 3D structure of CaORC4 with walker A and B motifs and arginine finger motif. We used SMART predictions to check the putative interactive partners of CaORC proteins of which CaORC4 was found to have no direct interaction with any other CaORC protein.

Discussion

CaORC proteins (CaORC1-6) and their associated proteins were identified by a BLAST analysis using the *S. cerevisiae* proteins as the query sequences in the Candida Genome Database (CGD). The phylogenetic analysis suggests that in spite of limited amino acid sequence similarity with their counterparts in other organisms, the CaORC proteins share most of the functional domains with them. Interestingly, the amino acid sequences of CaORC1, 2 and 4 share higher degree of similarities than CaORC3, 5 and 6 to those of *S. cerevisiae*. CaORC1, 4 and 5 tend to be homologous to the mammalian counterparts. Moreover, the CaORC proteins are also compared across CTG clade and other yeast species to provide a robust roadmap for further comparative yeast subphylum analysis (Figure 2 and Figure 3). Of the other preRC components compared here, Cdt1 has no apparent homolog in *C. albicans*, whereas, all other pre-RC proteins such as Cdc6 and Mcm2-7are very similar to their counterparts in other yeasts.

The main function of ORC proteins is to associate specifically with origins and recruit other factors including Cdc6 and MCM2-7 to form the preRC. In *S. cerevisiae*, the origins have a conserved stretch of 11 bp, the ARS consensus sequence, ACS, which is essential for ORC binding and origin activity. The replication origins of *C. albicans* (based on limited available data[48-51]) appear to be similar to those of *S. pombe* and other higher eukaryotes in having no such consensus sequence. In *S. pombe,* ORC4 binds with AT-rich origins via its 9 AT-hook motifs. Moreover, the ORC-origin binding might be affected both by intrinsic factors such as the DNA sequence that marks the ORC binding site and by extrinsic factors such as the chromatin component that marks both the histone and non-histone proteins. The absence of conserved sequences in *C. albicans* origins[50] along with our *in silico* analysis suggests that the CaORC-origin interactions would be largely chromatin dependent. In the genome-wide studies for identification of replication origins in *C. albicans* by ChIP-microarray based approach using an antibody against *S. cerevisiae* ORC complex, low nucleosome occupancy has been shown as conserved landmark of replication origins in *C. albicans*[51].

The BAH module found in several chromatin-associated proteins play important roles in gene silencing, replication and transcriptional regulation by promoting protein-protein interaction[12]. The BAH domain of humanORC1 has been shown to bind to H4K20me2[52] and abrogation of this binding causes impaired ORC1 loading onto origins, and cell cycle progression. The BAH domain present in CaORC1 along with the highly conserved basic residues (K-362 and R-367)[53] in its AAA domain is likely to play a key role in ORC-origin binding in *C. albicans*.

The AAA+ domains present in different ORC subunits (ORC1 and 5 in *S. cerevisiae* and ORC1, 4, and 5 in metazoans) are important for the assembly of ORC at origins and those in Cdc6 are critical for the loading of the MCM proteins (Table 6). Like metazoans, the CaORC subunits 1, 4, and 5 and CaCdc6 containing AAA+ domains are likely to be engaged in ORC assembly and consequent MCM recruitment although a perfect match to the Walker A and B motifs are present only in CaORC4 (Table 4 and 5). In all tested organisms, ORC1 has been found to be the major ATPase required for ORC assembly at origins. Experimental evidence would be required to find out if some or all of these subunits are involved in ATP binding and hydrolysis in *C. albicans*. Cdt1 helps in Cdc6 recruitment to origin bound ORC and is important in cell cycle regulation of preRC assembly at origins and limiting replication to a single round per cell cycle. The absence of a Cdt1 homolog in *C. albicans* suggests that this important task may be accomplished by a different mechanism/factor (Table 6 and Table 7). The unique presence of the PEST motif in CaORC2 and CaORC3 indicates that these components might be degraded in a cell cycle specific manner facilitating ORC turnover. The unique sequence of nine copies of AT hook motifs present in SpORC4 are critical for origin binding of ORC4 which is ATP-independent in *S. pombe*[24]. In *S. cerevisiae*, the origin binding of ORC is ATP-dependent and the presence of single DNA-binding AT-hook motif (PRKRGRPRK) is identified in the disordered regions of ScORC2[19]. The presence of the small AT-hook motif in CaORC2 to be another plausible motif for origin binding and their role in replication remains highly speculative. Similarly, it remains elusive as to whether the presence of MIR domain in CaORC3 has any role in silencing by binding to any heterochromatin component like HP1.

In *S. cerevisiae*, the ScORC proteins associate with origins in a sequence-dependent manner. Only ORC1, ORC2, ORC4 and ORC5 appear to make direct contacts to A and B1 domains of the replication origin[33, 54-55]. ScORC3 helps in forming the stable complex without directly binding to the DNA whereas ORC6 does not bind to the DNA but helps in recruiting multiple Cdt1 molecules[56-59]. The situation is very different in *Drosophila* cells where DNA replication initiates at many sites, which are probably sequence independent, throughout the genome at the same time[60]. In contrast to ScORC6, which is not required for DNA binding, DmORC6 is required for the DNA binding of DmORC and is an integral part of the DmORC complex[61]. The DmORC6 alone has DNA binding activity, likely due to the predicted TFIIB-like DNA binding domain in the smallest subunit[62]. DmORC binds DNA with little sequence specificity. ORC proteins generally require ATP to interact specifically with origin DNA (except in the case of SpORC). In all the species studied so far, ORC1, ORC4 and ORC5 contain potential ATP binding sites. ATP hydrolysis by ORCs to regulate DNA binding is well studied in ScORCs and DmORCs[26, 59]. The *in silico* predictions by Beltrao and colleagues[34] showed the increasing probability of CaORC2, CaORC4 and CaORC6 proteins to be phosphorylated by Cdc28, a cyclin dependent protein kinase.

We were able to build the 3D protein structure of CaORC4 only whereas the other CaORC proteins did not have good homology with the known PDB (Protein Data Bank) structures. From our *in silico* analysis of interactive studies, it is evident that CaORC3, CaORC5 and CaORC6 do not interact with the other ORC counterparts. It is possible that only CaORC1, CaORC2 and CaORC4 would be involved in DNA binding during the process of DNA replication and the other counterparts may aid in tethering or in conformational organization. CaCdc6 and Cdc54, the apparent common binding partners of CaORC1, CaORC2 and CaORC4 and many MCMs are also predicted to play important role(s) in preRC assembly and functioning. We also find a potential ATP binding site in CaORC4 which might help in the regulation of origin binding. The mode of ORC assembly at origins in *C. albicans* might be different from that in other yeasts. The *in silico* detection of the presence of AAA+ ATPase and Walker motifs in CaORC4 and its likely interaction with MCM proteins suggest that CaORC4 might be involved in stable binding to origin DNA and loading MCM proteins to origins. While possibilities of a physical association between CaORC4 and other CaORC proteins were not obvious, the role of some unknown factors mediating ORC assembly in *C. albicans* is not ruled out. CDC6, CDC54 and MCM proteins interact with CaORC1, CaORC2 and CaORC4. In absence of a direct interaction of CaORC4 with other ORC counterparts, these proteins might be mediating interaction between them. Moreover, the absence of Cdt1 in *C. albicans* might provide an additional role for CaORC4.

Recent studies demonstrate that besides the involvement of specific proteins that control DNA replication, some enzymes with primary functions that are involved in various other processes can also play a vital role in the regulation of genome duplication. There seems to be a direct link between central carbon metabolism and DNA replication regulation from prokaryotes [63-65] to eukaryotes including humans[66-67]. A recent analysis[66] demonstrates that partial silencing of genes encoding for the glycolytic and TCA enzymes affects the entry of human fibroblasts into the S-phase. It is also reported that ScORC proteins interact with some of the metabolic genes that are associated with replication origins[68]. One such example is the hexokinase (HXK2) gene which at a decreased level causes substantial impairment in DNA synthesis. Our preliminary reports from Co-IP studies (data not shown) also showed an interaction of CaORC4 with HXK2 by which it is speculated that CaORC4 might play a role in the regulation of central carbon metabolism besides its cardinal role of DNA replication. This can be further supported by the induced expression of CaORC4 in response to alpha pheromone in SpiderM medium[69].

From the above observations, we hypothesize that CaORC4 might be less tightly associated with the core preRC complex but involved in cell cycle regulation and DNA checkpoint activation. It is quite possible that CaORC4 may not be bound to chromatin throughout the cell cycle as seen in *Drosophila* and yeast [33]. Recent studies advocate a concerted interaction between ORCs, nucleosomes and replication origin DNA that stabilizes ORC-origin binding in yeast. The atomic force microscope (AFM) studies show that ORC establishes its origin interaction by binding to both nucleosome-free origin DNA and neighboring nucleosomes that are species-specific[70].

Recent reports suggest that Drg1, an AAA-ATPase protein is the potential target for the drug diazaborine. This drug is demonstrated to block ribosome biogenesis in yeast[71]. Similarly, a valosin containing AAA-ATPase protein, P97 is found to be a therapeutic target for CB-5083 in the cancer treatment[72]. A study on Trypanosoma ORC has raised possibilities on identifying novel drug targets demonstrating the drug potential of the pre-replication machinery[73].

Our *in silico* studies would form the basis for understanding the domain architecture and further characterization of CaORC proteins which can be validated by *in vitro* studies. It may ultimately provide clues about the potential drug targets helping curb Candida infection at the step of DNA replication.

Methods

Annotation of *C. albicans* pre-RC genes

The genome of *C. albicans* (http://www.candidagenome.org/) was searched for homologs of pre-RC complex genes using BLAST[74]. Alignment of pre-RC gene sequences from Candida and its homologs in other eukaryotic organisms was carried out using the ClustalW algorithm[15].

The pairwise ClustalW scores are calculated by the number of identities between the two sequences, divided by the alignment length in terms of percentage.

Phylogenetic analysis

Phylogenetic analysis was performed with the MEGA4 program[75].

***In silico* analysis**

The putative protein sequences whose theoretical characteristics were obtained using several programs in the ExPASy (Expert Protein Analysis System) server of the Swiss Institute of Bioinformatics (www.expasy.ch/tools/). Protein sequences were entered into MotifScan (pattern searches), ProDOM (protein domain identification), Interpro (protein domain and pattern search identification), NetPhos (prediction sites for phosphorylation) and PESTfind (identification of PEST sequences), SMART (prediction of protein domain architecture) and Phyre (secondary structure prediction). To determine the sequence identity of CaORC across 86 diverse publicly available yeast databases, a TBLASTN was performed in the Sequenceserver (http://blast.wei.wisc.edu/) with CaORC proteins as the query sequence[37-38] and the percent identity was plotted against the species using Graphpad Prism[76].

Phyre structure prediction parameters

Cscore^{GO} is a combined measure for evaluating global and local similarity between query and template protein. This score ranges from 0-1 where a higher value indicates a better confidence in predicting the function using the template. Cscore^{LB} is the confidence score of predicted binding site of the protein with values ranging between 0-1. Higher the score more reliable is the ligand binding prediction.

References

1. Bell, S. P. & Stillman, B. ATP-dependent recognition of eukaryotic origins of DNA replication by a multiprotein complex. *Nature* **357**, 128-134, doi:10.1038/357128a0 (1992).
2. Bell, S. P. & Dutta, A. DNA replication in eukaryotic cells. *Annu Rev Biochem* **71**, 333-374, doi:10.1146/annurev.biochem.71.110601.135425 (2002).
3. DePamphilis M. L. Replication origins in metazoan chromosomes: factor fiction? *Bioessays* **21**, 5-16, doi: 10.1002/(SICI)1521-1878(199901)21:1<5::AID-BIES2>3.0.CO;2-6 (1999).
4. Gilbert D. M. Making sense of eukaryotic DNA replication origins. *Science* **294**, 96-100, doi:10.1126/science.1061724 (2001).
5. Bowers, J. L., Randell, J. C. W., Chen, S. Y. & Bell, S. P. ATP hydrolysis by ORC catalyzes reiterative Mcm2-7 assembly at a defined origin of replication. *Mol Cell* **16**, 967-978, doi:DOI 10.1016/j.molcel.2004.11.038 (2004).
6. Shen, Z. The origin recognition complex in human diseases. *Bioscience Rep* **33**, 475-483, doi:ARTN e00044 10.1042/BSR20130036 (2013).
7. Bicknell, L. S. et al. Mutations in the pre-replication complex cause Meier-Gorlin syndrome. *Nat Genet* **43**, 356-U156, doi:10.1038/ng.775 (2011).
8. Bicknell, L. S. et al. Mutations in ORC1, encoding the largest subunit of the origin recognition complex, cause microcephalic primordial dwarfism resembling Meier-Gorlin syndrome. *Nat Genet* **43**, 350-U103, doi:10.1038/ng.776 (2011).
9. Tao Q, Young L.S, Woodman C.B et al. Epstein-Barr virus (EBV) and its associated human cancers - Genetics, epigenetics, pathobiology and novel therapeutics. *Frontiers in Bioscience* **11**,2672-2713 (2006).
10. Dang, H. Q. & Li, Z. The Cdc45.Mcm2-7.GINS protein complex in trypanosomes regulates DNA replication and interacts with two Orc1-like proteins in the origin recognition complex. *J Biol Chem* **286**, 32424-32435, doi:10.1074/jbc.M111.240143 (2011).
11. Low, C. Y. & Rotstein, C. Emerging fungal infections in immunocompromised patients. *F1000 Med Rep* **3**, 14, doi:10.3410/M3-14 (2011).
12. Whaley, S. G. *et al.* Azole Antifungal Resistance in Candida albicans and Emerging Non-albicans Candida Species. *Front Microbiol* **7**, doi:ARTN 2173 10.3389/fmicb.2016.02173 (2017).
13. Riggsby, W. S., Torres-Bauza, L. J., Wills, J. W. & Townes, T. M. DNA content, kinetic complexity, and the ploidy question in Candida albicans. *Mol Cell Biol* **2**, 853-862 (1982).
14. Kabir, M. A., Hussain, M. A. & Ahmad, Z. Candida albicans: A Model Organism for Studying Fungal Pathogens. *ISRN Microbiol*, **538694**, doi:10.5402/2012/538694 (2012).
15. Thompson, J.D, Higgins, D.G, Gibson T.J . CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**,4673-4680 (1994).

16. Callebaut, I., Courvalin, J. C. & Mornon, J. P. The BAH (bromo-adjacent homology) domain: a link between DNA methylation, replication and transcriptional regulation. *Febs Lett* **446**, 189-193, doi:Doi 10.1016/S0014-5793(99)00132-5 (1999).
17. Duderstadt, K. E. & Berger, J. M. AAA plus ATPases in the initiation of DNA replication. *Crit Rev Biochem Mol* **43**, 163-187, doi:10.1080/10409230802058296 (2008).
18. Wigley, D. B. ORC proteins: marking the start. *Curr Opin Struc Biol* **19**, 72-78, doi:10.1016/j.sbi.2008.12.010 (2009).
19. Duncker, B. P., Chesnokov, I. N. & McConkey, B. J. The origin recognition complex protein family. *Genome Biol* **10**, doi:ARTN 214 10.1186/gb-2009-10-3-214 (2009).
20. Kawakami, H. & Katayama, T. DnaA, ORC, and Cdc6: similarity beyond the domains of life and diversity. *Biochem Cell Biol* **88**, 49-62, doi:10.1139/O09-154 (2010).
21. Bell SP. The origin recognition complex: from simple origins to complex functions. *Genes Dev*. **16**, 659-672, doi: 10.1101/gad.969602 (2002).
22. Guernsey, D. L. et al. Mutations in origin recognition complex gene ORC4 cause Meier-Gorlin syndrome. *Nat Genet* **43**, 360-364, doi:10.1038/ng.777 (2011).
23. Iyer, L. M., Leipe, D. D., Koonin, E. V. & Aravind, L. Evolutionary history and higher order classification of AAA+ ATPases. *J Struct Biol* **146**, 11-31, doi:10.1016/j.jsb.2003.10.010 (2004).
24. Chuang, R. Y. & Kelly, T. J. The fission yeast homologue of Orc4p binds to replication origin DNA via multiple AT-hooks. *Proc Natl Acad Sci U S A* **96**, 2656-2661 (1999).
25. Speck, C., Chen, Z., Li, H. & Stillman, B. ATPase-dependent cooperative binding of ORC and Cdc6 to origin DNA. *Nat Struct Mol Biol* **12**, 965-971, doi:10.1038/nsmb1002 (2005).
26. Klemm, R. D., Austin, R. J. & Bell, S. P. Coordinate binding of ATP and origin DNA regulates the ATPase activity of the origin recognition complex. *Cell* **88**, 493-502 (1997).
27. Walker, J. E., Saraste, M., Runswick, M. J. & Gay, N. J. Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *EMBO J* **1**, 945-951 (1982).
28. Reeves, R. & Nissen, M. S. The A.T-DNA-binding domain of mammalian high mobility group I chromosomal proteins. A novel peptide motif for recognizing DNA structure. *J Biol Chem* **265**, 8573-8582 (1990).
29. Aravind, L. & Landsman, D. AT-hook motifs identified in a wide variety of DNA-binding proteins. *Nucleic Acids Res* **26**, 4413-4421 (1998).
30. Murzina, N., Verreault, A., Laue, E. & Stillman, B. Heterochromatin dynamics in mouse cells: Interaction between chromatin assembly factor 1 and HP1 proteins. *Mol Cell* **4**, 529-540, doi:Doi 10.1016/S1097-2765(00)80204-X (1999).
31. Prasanth, S. G., Shen, Z., Prasanth, K. V. & Stillman, B. Human origin recognition complex is essential for HP1 binding to chromatin and heterochromatin organization. *P Natl Acad Sci USA* **107**, 15093-15098, doi:10.1073/pnas.1009945107 (2010).
32. Ranganathan, R., Cannon, S. C. & Horvitz, H. R. MOD-1 is a serotonin-gated chloride channel that modulates locomotory behaviour in C. elegans. *Nature* **408**, 470-475 (2000).
33. Li, C. J. & DePamphilis, M. L. Mammalian Orc1 protein is selectively released from chromatin and ubiquitinated during the S-to-M transition in the cell division cycle. *Mol Cell Biol* **22**, 105-116, doi:Doi 10.1128/Mcb.22.1.105-116.2002 (2002).

34. Beltrao, P. et al. Evolution of Phosphoregulation: Comparison of Phosphorylation Patterns across Yeast Species. *Plos Biol* **7**, doi:ARTN e1000134 10.1371/journal.pbio.1000134 (2009).
35. Kumar, S., Tamura, K. & Nei, M. Mega - Molecular Evolutionary Genetics Analysis Software for Microcomputers. *Comput Appl Biosci* **10**, 189-191 (1994).
36. Saitou, N. & Nei, M. The Neighbor-Joining Method - a New Method for Reconstructing Phylogenetic Trees. *Mol Biol Evol* **4**, 406-425 (1987).
37. Priyam, A., B. J. Woodcroft, V. Rai, A. Munagala, I. Moghul et al., Sequenceserver: a modern graphical user interface for custom BLAST databases bioRxiv http://biorxiv.org/lookup/doi/10.1101/033142(2015).
38. Shen, X. X.et al.Reconstructing the backbone of the saccharomycotina yeast phylogeny using genome-scale data. *Genes Genomes Genetics* **6**, 3927-3939 (2016).
39. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. Scop - a Structural Classification of Proteins Database for the Investigation of Sequences and Structures. *J Mol Biol* **247**, 536-540, doi:Doi 10.1016/S0022-2836(05)80134-2 (1995).
40. Berman, H. M. et al. The Protein Data Bank and the challenge of structural genomics. *Nat Struct Biol* **7**, 957-959, doi:Doi 10.1038/80734 (2000).
41. Zhang, Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* **9**, 40, doi:10.1186/1471-2105-9-40 (2008).
42. McGuffin, L. J., Bryson, K. & Jones, D. T. The PSIPRED protein structure prediction server. *Bioinformatics* **16**, 404-405 (2000).
43. Pollastri, G., Przybylski, D., Rost, B. & Baldi, P. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins* **47**, 228-235, doi:10.1002/prot.10082 (2002).
44. Cole, C.J.D., Barber, J.D., Barton, G.J. The Jpred 3 secondary structure prediction server. *Nucleic Acids Res* **36**, W197-201,doi: 10.1093/nar/gkn238 (2008).
45. Bennet-Lovsey, R.M., Herbert, A.D., Sternberg, M.J. et al. Exploring the extremes of sequence / structure space with ensemble fold recognition in the program Phyre. *Proteins : Structure, Function and Bioinformatics* **70**, 611-625, doi: 10.1002/prot.21688 (2008).
46. Schultz, J., Milpetz, F., Bork, P. & Ponting, C. P. SMART, a simple modular architecture research tool: Identification of signaling domains. *P Natl Acad Sci USA* **95**, 5857-5864, doi:DOI 10.1073/pnas.95.11.5857 (1998).
47. Letunic, I., Doerks, T. & Bork, P. SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Research* **40**, D302-D305, doi:10.1093/nar/gkr931 (2012).
48. Cannon, R. D., Jenkinson, H. F. & Shepherd, M. G. Isolation and nucleotide sequence of an autonomously replicating sequence (ARS) element functional in Candida albicans and Saccharomyces cerevisiae. *Mol Gen Genet* **221**, 210-218 (1990).
49. Beckerman, J., Chibana, H., Turner, J. & Magee, P. T. Single-copy IMH3 allele is sufficient to confer resistance to mycophenolic acid in Candida albicans and to mediate transformation of clinical candida species. *Infect Immun* **69**, 108-114, doi:Doi 10.1128/Iai.69.1.108-114.2001 (2001).
50. Mitra, S., Gomez-Raja, J., Larriba, G., Dubey, D. D. & Sanyal, K. Rad51-Rad52 Mediated Maintenance of Centromeric Chromatin in Candida albicans. *Plos Genet* **10**, doi:ARTN e1004344 10.1371/journal.pgen.1004344 (2014).

51. Tsai, H. J. *et al.* Origin replication complex binding, nucleosome depletion patterns, and a primary sequence motif can predict origins of replication in a genome with epigenetic centromeres. *MBio* **5**, e01703-01714, doi:10.1128/mBio.01703-14 (2014).
52. Kuo, A.J., Song, J., Cheung, P et al. The BAH domain of ORC1 links H4K20me2 to DNA replication licensing and Meier-Gorlin syndrome, *Nature* **484**,115-119, doi: 10.1038/nature10956 (2012).
53. Kawakami, H., Ohashi, E., Kanamoto, S., Tsurimoto, T. & Katayama, T. Specific binding of eukaryotic ORC to DNA replication origins depends on highly conserved basic residues. *Sci Rep-Uk* **5**, doi:ARTN 14929 10.1038/srep14929 (2015).
54. Clarey, M. G. et al. Nucleotide-dependent conformational changes in the DnaA-like core of the origin recognition complex. *Nature Structural & Molecular Biology* **13**, 684-690, doi:10.1038/nsmb1121 (2006).
55. Lee, D. G. & Bell, S. P. Architecture of the yeast origin recognition complex bound to origins of DNA replication. *Mol Cell Biol* **17**, 7159-7168 (1997).
56. Asano, T., Makise, M., Takehara, M. & Mizushima, T. Interaction between ORC and Cdt1p of Saccharomyces cerevisiae. *Fems Yeast Res* **7**, 1256-1262, doi:10.1111/j.1567-1364.2007.00299.x (2007).
57. Chen, S., Bell, S.P. CDK prevents Mcm2-7 helicase loading by inhibiting Cdt1 interaction with Orc6. *Genes Dev* **25**, 363-372, doi: 10.1101/gad.2011511 (2011).
58. Chen S, de Vries MA, Bell SP. Orc6 is required for dynamic recruitment of Cdt1 during repeated Mcm2-7 loading. *Genes Dev* **21**, 2897-2907, doi: 10.1101/gad.1596807 (2007).
59. Takara, T. J. & Bell, S. P. Multiple Cdt1 molecules act at each origin to load replication-competent Mcm2-7 helicases. *Embo Journal* **30**, 4885-4896, doi:10.1038/emboj.2011.394 (2011).
60. Blumenthal, A. B., Kriegstein, H. J. & Hogness, D. S. The units of DNA replication in Drosophila melanogaster chromosomes. *Cold Spring Harb Symp Quant Biol* **38**, 205-223 (1974).
61. Chesnokov, I., Remus, D. & Botchan, M. Functional analysis of mutant and wild-type Drosophila origin recognition complex. *Proc Natl Acad Sci U S A* **98**, 11997-12002, doi:10.1073/pnas.211342798 (2001).
62. Liu, S. X. et al. Structural analysis of human Orc6 protein reveals a homology with transcription factor TFIIB. *P Natl Acad Sci USA* **108**, 7373-7378, doi:10.1073/pnas.1013676108 (2011).
63. Janniere, L. et al. Genetic Evidence for a Link Between Glycolysis and DNA Replication. *Plos One* **2**, doi:ARTN e447 10.1371/journal.pone.0000447 (2007).
64. Maciag, M., Nowicki, D., Janniere, L., Szalewska-Palasz, A. & Wegrzyn, G. Genetic response to metabolic fluctuations: correlation between central carbon metabolism and DNA replication in Escherichia coli. *Microb Cell Fact* **10**, doi:Artn 19 10.1186/1475-2859-10-19 (2011).
65. Maciag-Dorszynska, M., Ignatowska, M., Janniere, L., Wegrzyn, G. & Szalewska-Palasz, A. Mutations in central carbon metabolism genes suppress defects in nucleoid position and cell division of replication mutants in Escherichia coli. *Gene* **503**, 31-35, doi:10.1016/j.gene.2012.04.066 (2012).
66. Konieczna, A., Szczepanska, A., Sawiuk, K., Lyzen, R. & Wegrzyn, G. Enzymes of the central carbon metabolism: Are they linkers between transcription, DNA replication, and carcinogenesis? *Med Hypotheses* **84**, 58-67, doi:10.1016/j.mehy.2014.11.016 (2015).

67. Lincet, H. & Icard, P. How do glycolytic enzymes favour cancer cell proliferation by nonmetabolic functions? *Oncogene* **34**, 3751-3759, doi:10.1038/onc.2014.320 (2015).
68. Shor, E. et al. The Origin Recognition Complex Interacts with a Subset of Metabolic Genes Tightly Linked to Origins of Replication. *Plos Genet* **5**, doi:ARTN e1000755 10.1371/journal.pgen.1000755 (2009).
69. Bennett, R. J. & Johnson, A. D. The role of nutrient regulation and the Gpa2 protein in the mating pheromone response of C. albicans. *Mol Microbiol* **62**, 100-119, doi:10.1111/j.1365-2958.2006.05367.x (2006).
70. Hizume, K., Yagura, M. & Araki, H. Concerted interaction between origin recognition complex (ORC), nucleosomes and replication origin DNA ensures stable ORC-origin binding. *Genes Cells* **18**, 764-779, doi:10.1111/gtc.12073 (2013).
71. Loibl, M., Klein, I., Prattes, M et al. The drug diazaborine blocks ribosome biogenesis by inhibiting the AAA-ATPase Drg1, *J Biol Chem* **289**, 3913-3922, doi: 10.1074/jbc.M113.536110 (2014).
72. Anderson, D.J, Moigne, R.L, Djakovic, S et al. Targeting the AAA ATPase p97 as an approach to treat cancer through disruption of protein homeostasis, *Cancer Cell* **28**, 653-665, doi: 10.1016/j.ccell.2015.10.002 (2015).
73. Calderano, S. G., de Melo Godoy, P. D., da Cunha, J. P. & Elias, M. C. Trypanosome prereplication machinery: a potential new target for an old problem. *Enzyme Res* **2011**, 518258, doi:10.4061/2011/518258 (2011).
74. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402 (1997).
75. Tamura, K., Dudley, J., Nei, M. & Kumar, S. MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol Biol Evol* **24**, 1596-1599, doi:10.1093/molbev/msm092 (2007).
76. Mary, L. S. GraphPad Prism, data analysis, and scientific graphing. *J. Chem. Inf. Comput. Sci* **37**, 411-412 doi: 10.1021/ci960402j(1997).
77. Zuckerkandl, E., Pauling, L. Molecules as documents of evolutionary history. *J Theor Biol* **8**, 357-366 (1965).
78. Coletta, A. et al. Low-complexity regions within protein sequences have position-dependent roles. *Bmc Syst Biol* **4**, doi:Artn 43 10.1186/1752-0509-4-43 (2010).
79. Jones, D.T., Taylor, W.R., Thornton, J.M. The rapid generation of mutation data matrices from protein sequences, *Comput Appl Biosci* **8**, 275-282(1992).
80. Tamura, K. et al. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0, *Mol Biol Evol* **30**, 2725-2729, doi: 10.1093/molbev/mst197 (2013).
81. Tamura, K. et al. Estimating divergence times in large molecular phylogenies, *Proc Natl Acad Sci USA* **109**,19333-19338, doi: 10.1073/pnas.1213199109 (2012).

Acknowledgements

This work was supported by Department of Biotechnology to KS and DDD (BT/PR13724/BRB/10/782/2010). The award of direct Senior Research Fellowships to SP from

Council of Scientific and Industrial Research (9/1014(0001)2K10-EMR-I) is greatly acknowledged. We thank Dr. E.J.Woo, Korea for the 3D structural studies.

Author Contributions

S.P. performed experiments, analyzed data and wrote the paper. K.S. and D.D. designed the study, analyzed data, and wrote the paper.

Competing Interests

The authors declare that they have no competing interests.

Figure legends

Figure. 1. Evolutionary relationship of CaORC proteins with other species and comparative domain architecture of CaORC and ScORC proteins. (A) Phylogram of ORC proteins. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Poisson correction method[77] and are in the units of the number of amino acid substitutions per site. All positions containing gaps and missing data were eliminated from the dataset (Complete deletion option). There were a total of 116 positions in the final dataset. Phylogenetic analyses were conducted in MEGA4[35]. (**B**) The SMART (Simple Modular Architecture Research Tool) prediction shows the presence of the BAH domain spanning between 44th and 179th amino acids at the N-terminal of CaORC1 and (**C**) The AAA+ domain in CaORC4 protein, the purple box represents the low complexity region (LCR). The LCR may be involved in flexible binding associated with specific functions but also that their positions within a sequence may be important in determining both their binding properties and their biological

roles [78]. (**D**) Comparative domain architecture of ORC proteins in *S. cerevisiae* and *C. albicans.* The red box denotes the BAH domain, the grey box is the AAA+ domain, cyan bar represents the AT-hook motif, black bar represents the Walker motifs, dark blue bar represents the PIP motif, yellow bar represents the MIR motif and the green bar represents the PEST motif.

Figure. 2. ORC phylogeny in CTG clade. (A) Molecular Phylogenetic analysis of ORC proteins in the CTG clade by Maximum Likelihood method. The evolutionary history was inferred by using the Maximum Likelihood method based on the JTT matrix-based model[75]. The tree with the highest log likelihood (-14518.9956) is shown. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using a JTT model, and then selecting the topology with superior log likelihood value. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 29 amino acid sequences. All positions containing gaps and missing data were eliminated. There were a total of 237 positions in the final dataset. Evolutionary analyses were conducted in MEGA6 [80]. **(B)** The time tree molecular Phylogenetic analysis of ORC proteins in the CTG clade by the Maximum Likelihood method. The timetree shown was generated using the RealTime method [81]. Divergence times for all branching points in the topology were calculated using the Maximum Likelihood method based on the JTT matrix-based model[79]. The estimated log likelihood value of the topology shown is -14518.9956. The tree is drawn to scale, with branch lengths measured in the relative number of substitutions per site. The analysis involved 29 amino acid sequences. All positions containing gaps and missing data were eliminated. There were a total of 237 positions in the final dataset. Evolutionary analyses were conducted in MEGA6[80].

Figure. 3. Comparative profile of percentage identity of CaORC proteins with other yeasts. The hexameric ORC complex containing ORC1-6 protein sequences are compared individually with diverse yeast species whose genome database is publicly available[37-38] and their percent identity is plotted using Graphpad Prism[76]. **(A)** Percent identity of 104 hits of ORC1 sequences. **(B)** Percent identity of 86 hits of ORC2 sequences. **(C)** Percent identity of 90 hits of ORC3 sequences **(D).** Percent identity of 104 hits of ORC4 sequences. **(E)**Percent identity of 88 hits of ORC5 sequences. **(F)** Percent identity of 69 hits of ORC6 sequences.

Figure. 4. 3D model of CaORC4 and putative interactors of CaORC proteins. (A) 3D model of CaORC4 with DNA. (**B**) 3D model of CaORC4 with Walker A bound to ATP sphere, Walker B and R finger motifs. (**C-H**) Protein interaction map of the *C. albicans* pre-RC including CaORC1, CaORC2, CaORC4, CaMCM2, CaMCM3, CaMCM5 (CDC46) respectively (Table 10). The bright red circle is the query protein. The interaction map of CaMCM4 and CaMCM6 are the same as CaMCM3.

Table 1. Putative pre-RC proteins coded by the *C. albicans* genome

Protein	**ORF#**	**Chr#**	**Protein**	**ORF#**	**Chr#**
CaORC1	Orf19.3000	1	CaMCM2	Orf19.4354	R
CaORC2	Orf19.5358	2	CaMCM3	Orf19.1901	2
CaORC3	Orf19.6942	3	CaMCM4	Orf19.3761	1
CaORC4	Orf19.4221	5	CaMCM5	Orf19.5487	2
CaORC5	Orf19.2369	R	CaMCM6	Orf19.2611	R

CaORC6	Orf19.3289	1	CaMCM7	Orf19.202	2
CaCDC6	Orf19.5242	1			

Table 2. Comparison of putative CaORC sequences with ORC sequences of *S. cerevisiae* and *S. pombe*. The systematic names, ORF and protein length along with isoelectric pH of the ORC1-6 in *S. cerevisiae*, *S. pombe* and *C. albicans*.

Gene	*S.cerevisiae*				*S.pombe*				*C.albicans*			
	Systematic name	Length (bp)	Protein length (a.a)	pI	Systematic name	Length (bp)	Protein length (a.a)	pI	Systematic name	Length (bp)	Protein length (a.a)	pI
ORC1	YML065W	2745	914	5.52	SPBC29A10.15	2124	709	7	ORF19.3000	2418	805	5.99
ORC2	YBR060C	1863	620	9.45	SPBC685.09	1608	535	5.51	ORF19.5358	2067	688	8.26
ORC3	YLL004W	1851	616	5.27	SPAC3H1.01C	2073	690	5.59	ORF19.6942	2049	682	5.32
ORC4	YPR162C	1590	529	6.39	SPBP23A10.13	2919	972	9.31	ORF19.4221	1695	564	6.19
ORC5	YNL261W	1440	479	5.64	SPBC646.14C	1368	455	8.83	ORF19.2369	1491	496	6.22
ORC6	YHR118C	1308	435	8.16	SPBC2A9.12	795	264	8.48	ORF19.3289	1092	363	9.2

Table 3.Pairwise alignment results of CaORC proteins with other eukaryotes

Protein Name	Clustal W scores						Length of protein (a.a)						
	Ca vs Sc	***Ca vs Sp***	***Ca vs Dm***	***Ca vs Xl***	***Ca vs Mm***	***Ca vs Hs***	***Ca***	***Sc***	***Sp***	***Dm***	***Xl***	***Mm***	***Hs***
ORC1	26	26	20	20	23	20	805	914	709	924	886	840	861
ORC2	25	21	19	16	20	19	688	620	535	618	558	576	577
ORC3	18	14	13	14	18	17	682	616	690	721	709	715	712
ORC4	25	24	21	23	25	23	564	529	972	459	432	433	436
ORC5	22	21	20	17	23	24	496	479	455	460	448	435	435
ORC6	17	12	6	15	9	11	363	435	264	257	225	262	252

Sc – Saccharomyces cerevisiae; Sp – Schizosaccharomyces pombe; Ca – Candida albicans; Dm – Drosophila melanogaster; Xl – Xenopus laevis; Mm – Mus musculus; Hs – Homo sapiens

Table 4. Comparison of Walker A and Walker B motifs of CaORC proteins with other species

Protein name	**Organism**	**Walker A motif** **(**GXXXXGKT/S)	**Walker B motif** (hhDE)
ORC1	*S. cerevisiae*	GTPGVGKT	LLDE
	S. pombe	GTPGTGKT	LMDE
	C. albicans	GVPGMGKT	LMDE
	C.elegans	GVPGTGKT	LI DE
	D. melanogaster	GVPGTGKT	LVDE
	M. musculus	GVPGTGKT	LVDE
	H. sapiens	GVPGTGKT	LVDE
ORC4	*S. cerevisiae*	GPRQSYKT	IFDE
	S. pombe	GPRGSGKS	VLEE
	C. albicans	GPRSSGKT	SDLE
	C.elegans	GERNCGRE	LVRD
	D. melanogaster	GPRGSGKT	ILEE
	M. musculus	GPRGSGKT	ILDE
	H. sapiens	GPRGSGKT	ILDE
ORC5	*S. cerevisiae*	GYSGTGKT	
	S. pombe	GVASTAKT	
	C. albicans	GYKSIGKT	
	C.elegans	GEDGSGRS	
	D. melanogaster	GHSGTGKT	
	M. musculus	GHTASGKT	
	H. sapiens	GHTASGKT	

Table 5. Putative signature of Walker motifs in CaORC proteins

Motif name	Functions	Sequence	Motif and sequence position in CaORC proteins
Walker A motif	Motif associated with phosphate binding	GXXXXGK	GVPGMGK (428-434) – CaORC1 GPRSSGK (147-153) – CaORC4 GYKSIGK (44-50) – CaORC5
Walker B motif	Essential for ATP hydrolysis	(R/K)XXXGXXXL/VhhhhD	RKPLVILMDE (506-515) – CaORC1 RTTGSNGVQDLVTSLSD (410-426) – CaORC4

Table 6. Domains of *C.albicans* ORC proteins compared with other eukaryotes

Protein	*C.albicans*	*S.cerevisiae*	*S.pombe*	*D.melanogaster*	*X.laevis*	*M.musculus*	*H.sapiens*	*A.thaliana*
ORC1	BAH domain, PIP motif, AAA ATPase, Walker A & B motifs	BAH domain, AAA ATPase	BAH domain	BAH domain, AAA ATPase	BAH domain, AAA ATPase, PEST motif	BAH domain, AAA ATPase, PEST motif	BAH domain, AAA ATPase, PEST motif	BAH domain, PHD zinc finger, AAA ATPase, PEST motif
ORC2	AT hook, PEST motif	AT hook	Not determined	No hits	No hits	No hits	No hits	PEST motif
ORC3	MIR, PEST motif	ND	Not determined	AAA ATPase (P loop)	Not determined	No hits	MIR	Domain 1 Cullins, PEST motif
ORC4	AAA ATPase, Walker A & B motifs	No hits	AT hook	AAA ATPase (P loop)	AAA ATPase (P loop)	AAA ATPase (P loop)	AAA ATPase (P loop)	AAA ATPase
ORC5	WalkerA motif	AAA ATPase (P loop)	Not determined	AAA ATPase (P loop)	Not determined	AAA ATPase (P loop)	AAA ATPase (P loop)	AAA ATPase, PEST motif
ORC6	No hits	No hits	Not determined	No hits	Not determined	No hits	No hits	No hits
CDC6	AAA ATPase	AAA ATPase	Not determined	Not determined	AAA ATPase	AAA ATPase	AAA ATPase	AAA ATPase
CDT1	No hits	Not determined	Not determined	No hits	No hits	No hits	No hits	PEST motif

Table 7. Comparison of the Clustal W scores and lengths of the ORC associated proteins in *S. cerevisiae* and *S. pombe* with *C. albicans*.

Protein Name	Clustal W scores			Length of protein (a.a)		
	Ca* vs *Sc	***Ca* vs *Sp***	***Sc* vs *Sp***	***Ca***	***Sc***	***Sp***
CDC6	27	10	8	480	513	1086
CDT1	NA	NA	11	NA	604	444
MCM2	67	58	60	903	868	830
MCM3	56	49	49	878	971	879
MCM4	62	56	56	912	933	911
MCM5	67	60	61	728	775	720
MCM6	65	55	53	880	1017	892
MCM7	60	58	57	781	845	760

Sc – Saccharomyces cerevisiae; Sp – Schizosaccharomyces pombe; Ca – Candida albicans

NA-Not applicable

Table 8. Predicted molecular weight of the ORC proteins in *C. albicans, S. cerevisiae* and *S. pombe*

ORC proteins	M.W in *S.cerevisiae* (in KDa)	M.W in *S.pombe* (in KDa)	M.W in *C.albicans* (in KDa)
ORC1	120	80	91
ORC2	72	61	78.6
ORC3	62	80	79.2
ORC4	56	108	64
ORC5	53	52	57
ORC6	50	31	41
Total	**~412**	**~412**	**~412**

Table 9. Cscore values of CaORC proteins

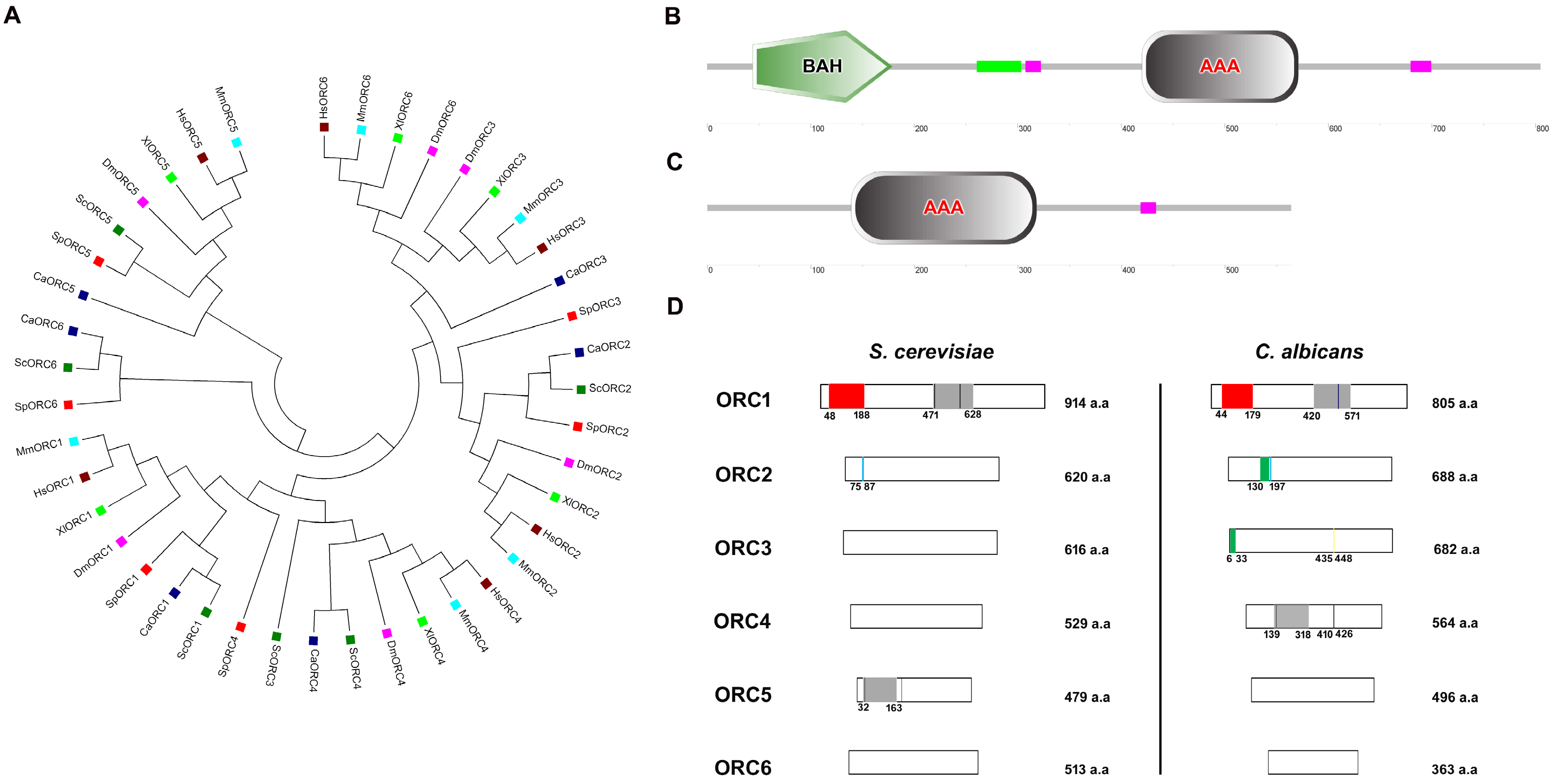
ORC proteins	**Cscore^{GO}**	**Cscore^{LB}**
ORC1	0.24	0.41
ORC2	0.25	0.02
ORC3	0.16	0.01
ORC4	0.29	0.6
ORC5	0.29	0.58
ORC6	0.21	0.01

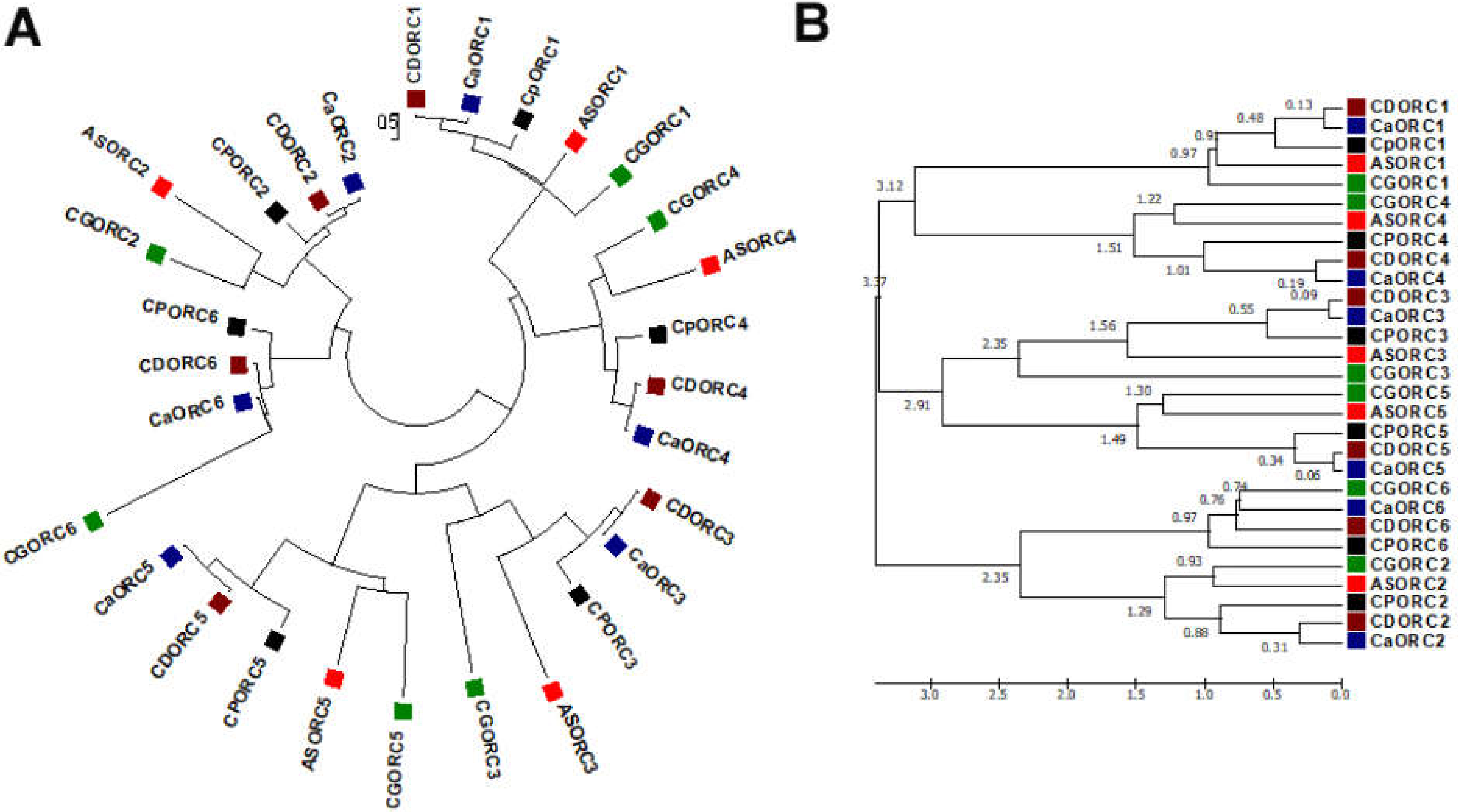
Table 10. SMART predictions of pre-RC proteins' interactions

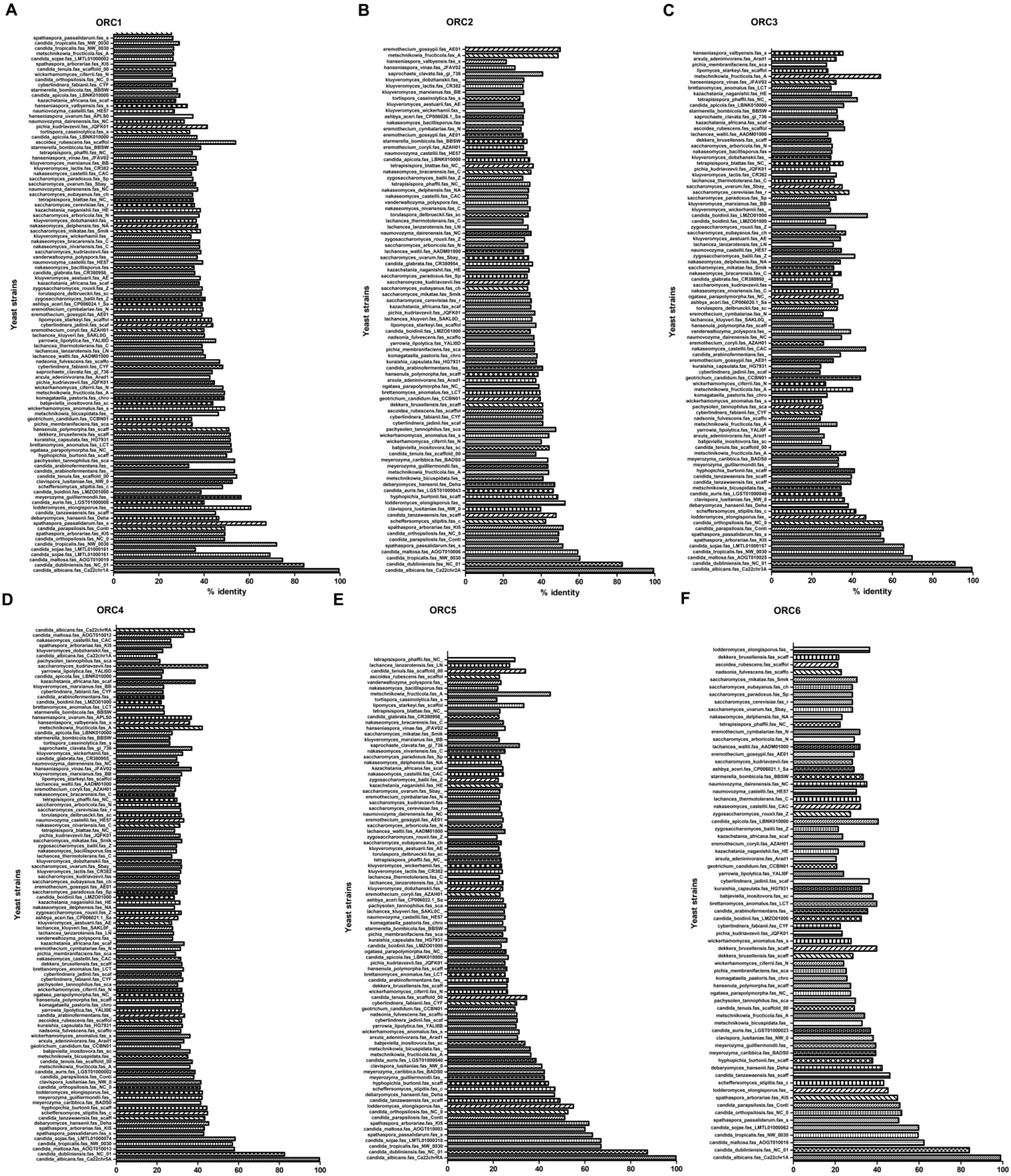
Pre-RC protein	**Domains / motifs**	**Putative interacting partners**
CAORC1	BAH & AAA	ORC2,CDC6,CDC46,**CDC54,MCM2,MCM3,MCM6, CAWG05_985**,POL12,LRO1
CAORC2		ORC1,CDC6,CDC46,**CDC54**,CDC45,CDC7,**MCM2, MCM3,MCM6,CAWG05_985**
CAORC3		-
CAORC4	AAA	**CDC54**,CDC6,**MCM2,MCM3,MCM6,CAWG05_985**
CAORC5		-
CAORC6		-
CACDC6	AAA	-
CAMCM2	MCM	ORC1,ORC2,CDC45,CDC46,**CDC54**,CDC7,**MCM3, MCM6,CAWG05_985**,RFA1
CAMCM3	AAA / MCM	ORC1,ORC2,ORC4,CDC45,CDC46,**CDC54**,CDC7,**MCM2, MCM3,MCM6, CAWG05_985**
CAMCM4 / CDC54	MCM	ORC1,ORC2,CDC45,CDC46,**CDC54**,CDC7,**MCM2, MCM3,MCM6, CAWG05_985**
CAMCM5 / CDC46	AAA / MCM	CDC45,**CDC54**,CDC7,**MCM2,MCM3,MCM6, CAWG05_985**,PRI1, POL30, RFA1

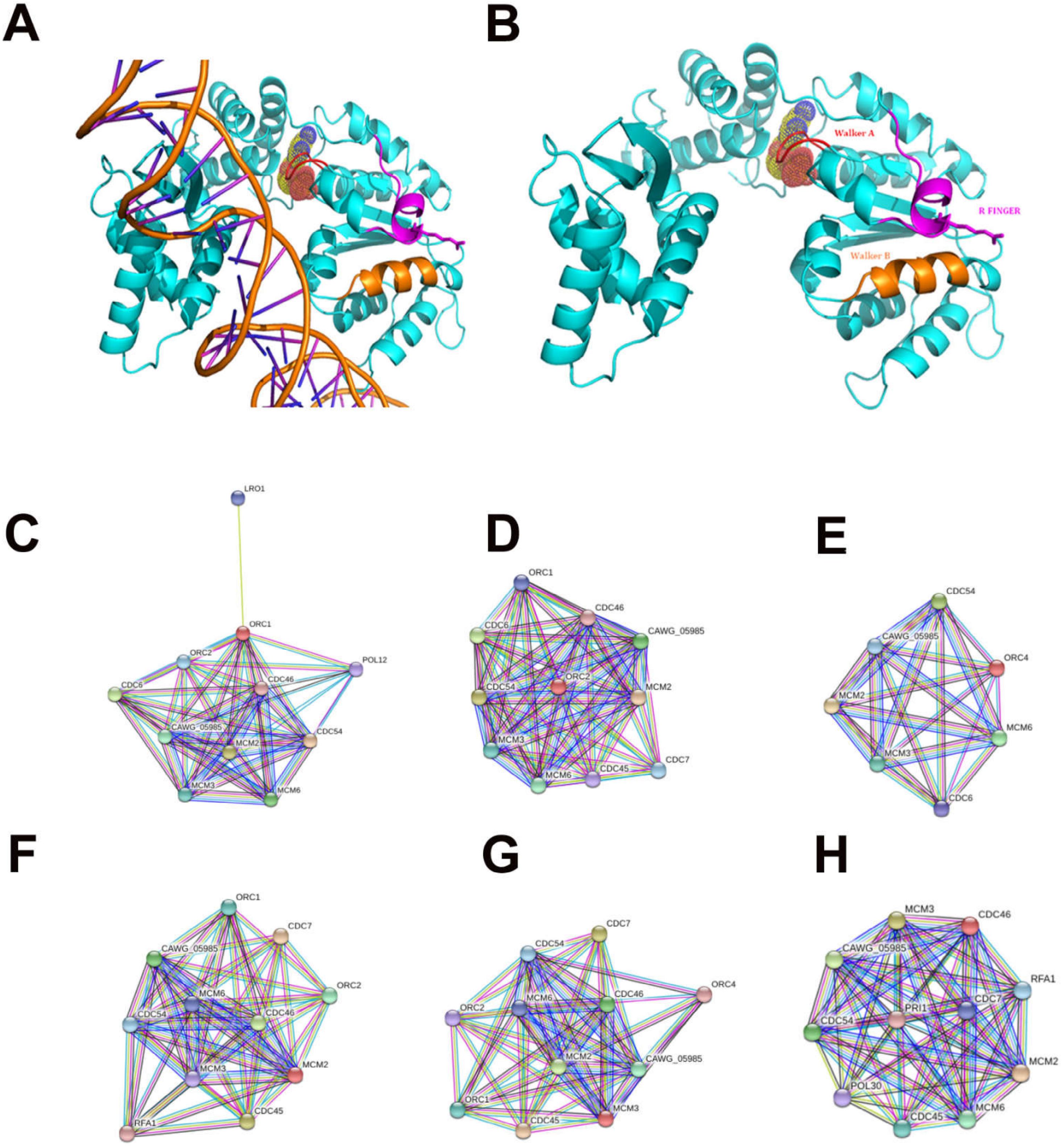
CAMCM6	MCM	ORC1,ORC2,ORC4,CDC45,CDC46,**CDC54**,CDC7,**MCM2, MCM3,MCM6, CAWG05_985**
CAMCM7	AAA / MCM	-

Note : CAWG05_985 = MCM7


A
MmORC5
HsORC5
XlORC5
DmORC5
ScORC5
SpORC5
CaORC5
CaORC6
ScORC6
SpORC6
MmORC1
HsORC1
XlORC1
DmORC1
SpORC1
CaORC1
ScORC1
SpORC4
ScORC3
CaORC4
ScORC4
DmORC4
XlORC4
MmORC4
HsORC4
MmORC2
HsORC2
XlORC2
DmORC2
SpORC2
ScORC2
CaORC2
SpORC3
CaORC3
HsORC3
MmORC3
XlORC3
DmORC3
DmORC6
XlORC6
MmORC6
HsORC6
B
BAH
AAA
0
100
200
300
400
500
600
700
800
C
AAA
0
100
200
300
400
500
D
S. cerevisiae
C. albicans
ORC1
48
188
471
628
914 a.a
44
179
420
571
805 a.a
ORC2
75 87
620 a.a
130
197
688 a.a
ORC3
616 a.a
6 33
435 448
682 a.a
ORC4
529 a.a
139
318
410 426
564 a.a
ORC5
32
163
479 a.a
496 a.a
ORC6
513 a.a
363 a.a


A
0.5
CDORC1
CaORC1
CpORC1
ASORC1
CGORC1
CGORC4
ASORC4
CpORC4
CDORC4
CaORC4
CDORC3
CaORC3
CPORC3
ASORC3
CGORC3
CGORC5
ASORC5
CPORC5
CDORC5
CaORC5
CGORC6
CaORC6
CDORC6
CPORC6
CGORC2
ASORC2
CPORC2
CDORC2
CaORC2
B
0.13
0.48
0.91
0.97
3.12
1.22
1.51
1.01
0.19
3.37
0.09
0.55
1.56
2.35
1.30
2.91
1.49
0.34
0.06
0.74
0.76
0.97
0.93
2.35
1.29
0.88
0.31
CDORC1
CaORC1
CpORC1
ASORC1
CGORC1
CGORC4
ASORC4
CPORC4
CDORC4
CaORC4
CDORC3
CaORC3
CPORC3
ASORC3
CGORC3
CGORC5
ASORC5
CPORC5
CDORC5
CaORC5
CGORC6
CaORC6
CDORC6
CPORC6
CGORC2
ASORC2
CPORC2
CDORC2
CaORC2
3.0
2.5
2.0
1.5
1.0
0.5
0.0


A
ORC1
B
ORC2
C
ORC3
D
ORC4
E
ORC5
F
ORC6
Yeast strains
% identity
0
20
40
60
80
100


A
B
Walker A
R FINGER
Walker B
C
LRO1
ORC1
ORC2
POL12
CDC6
CDC46
CAWG_05985
MCM2
CDC54
MCM3
MCM6
D
ORC1
CDC46
CDC6
CAWG_05985
CDC54
ORC2
MCM2
MCM3
CDC7
MCM6
CDC45
E
CDC54
CAWG_05985
ORC4
MCM2
MCM6
MCM3
CDC6
F
ORC1
CDC7
CAWG_05985
ORC2
MCM6
CDC54
CDC46
MCM3
MCM2
CDC45
RFA1
G
CDC7
CDC54
ORC4
MCM6
CDC46
ORC2
MCM2
CAWG_05985
ORC1
CDC45
MCM3
H
MCM3
CDC46
CAWG_05985
RFA1
CDC7
PRI1
CDC54
MCM2
POL30
CDC45
MCM6