Genome-wide studies of replication origins in *Candida albicans* reveal their conserved and distinct features

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Hung-Ji Tsai

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Judith Berman, PhD, Duncan Clarke, PhD

April 2013

Acknowledgement

I would like to acknowledge my thesis advisor, Judith Berman, for her fully support in these years. Her unending passion for science and work ethnic inspires me and keeps me motivated. I could not have gone such far withour her encouragement and help. I would also likes to thank my co-advisor, Duncan Clarke, for letting me explore my scientific interest. Next, I would like to thanks to my thesis committee: Deanna Koepp, Dan Voytas, Tom Neufeld, and Anja Bielinsky for their helpful opinions and understanding.

Thank you to all the present and past members of the Berman lab, especially Laura Burrack, who is always willing to listen and help from the beginning to the last day of my PhD career. I also want to thank to Josh Baller, Matt Anderson, Meleah Hickman, Ben Harrison, Mark McClellan, Shelly Clancy, and Maryam Gerami-Nejad for all the help in science, life and any other trivial matters, which all mean a lot to me. I deeply appreciate my wonderful labmates.

Finally, I want to thank to my family for being supportive and understanding. I also like to thank to my lovely fiancée, Man-Shun (Sarah) Fu for her company and taking care of me.

Abstract

Faithful DNA replication is required for genome inheritance during cell division. To ensure complete duplication of the entire genome, the initiation of replication occurs at multiple loci along the chromosomes. The determinants to regulate where and when DNA replication initiates in higher eukaryotes are little known. Not only the properties of nucleotide sequences, but also the dynamic chromatin structure are highly controlled to define replication origins. The major goal in my doctoral thesis is to understand the features of replication origins and their impact on genome organization in the pathogenic yeast *Candida albicans*.

I combined computational genomics and experimental approaches to address the following research objectives. First, I investigated the distinct replication features of centromeres. Centromere is a specialized chromosomal locus required for chromosome segregation during cell division. Importantly, *Candida* centromeres are epigenetically defined regional centromeres, similar to centromeres in higher eukaryotes. I found that centromeres constitutively replicate first on each chromosome and this early replication event is linked to the epigenetic nature of centromeres. Furthermore, aligning ORC binding sites with conserved nucleosome depletion patterns throughout the genome revealed the locations of potential chromosomal origins. Strikingly, origin DNA conferred ARS (autonomously replication sequences) function on a linear plasmid vector. Thus, I performed a genome-wide ARS screen to identify the consensus sequences (ACS) for ARS function, and I identified a unique 15 bp ACS motif. This motif is required for origin activity on the plasmid, and it functions in the

chromosomal context when associated with appropriately positioned nucleosomes. Thus, despite the presence of a regional, sequence-independent centromere, *C. albicans* requires a specific sequence motif for replication origin function.

Taken together, I comprehensively mapped the origins of replication in the *C. albicans* genome and characterized their conserved and distinct features. Importantly, the discovery of the ACS motif allowed us to develop the first plasmid shuttle vector for use as a genetic tool in *C. albicans*. All work here facilitates the study of *C. albicans* as a promising model organism for understanding fungal pathogenesis and eukaryotic genome organization.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

General Introduction

DNA encodes the genetic material in every living cell. DNA determines the distinct characteristics of an organism and is transferred from parents to offspring. Each time a cell divides, the DNA must faithfully replicate and segregate to maintain the integrity of the genome. However, intrinsic and extrinsic factors can stochastically confer mutations, from a single nucleotide change to a gain or loss of a whole chromosome, during cell division. On an evolutionary scale, cells must balance the conservation of genetic information with the promotion of genetic diversity. Genetic changes passed from parents to progeny can lead to detrimental outcomes, but also can potentiate cellular evolution. This dissertation focuses on the genetic elements necessary for genome inheritance by studying mechanisms of initiation of DNA replication and the link between DNA replication and chromosome segregation. In particular, I investigated the relationship between centromeres and replication origins and characterized the features of replication origins in the yeast *Candida albicans*. I found that two classes of replication origins co-exist in the *C. albicans* genome: centromeric origins are epigenetically-inherited with distinct timing of activation; chromosomal origins have definitive sequence motifs for origin activity and activate only within a conserved chromosomal context.

1.1. *Candida albicans*

1.1.1. *Clinical relevance*

C. albicans resides in the gastrointestinal tract, oral and vaginal mucosa as a commensal organism without disease phenotypes in most healthy humans.

However, *C. albicans* is also an opportunistic fungal pathogen in immuno-compromised patients and can cause mucosal infections and even life-threatening systematic infections (Mavor et al., 2005; reviewed in 2002). If *C. albicans* invades the blood stream, it can colonize the liver, heart, and kidneys. *C. albicans* is the most prevalent human fungal pathogen, and *Candidiasis,* caused mostly by *C. albicans,* is the fourth most common nosocomial infection in US hospitals (Pfaller and Diekema, 2007). The excess hospital costs resulting from Candidiasis to be up to 320 million US dollars annually (Banerjee et al., 1991; Miller et al., 2001; Trick et al., 2002; Wisplinghoff et al., 2004). While treatment options for *C. albicans* infections exist, treatment failure and the development of resistance to antifungals occurs frequently. Thus, new strategies to treat *C. albicans* infections are urgently needed.

1.1.2. *Molecular mechanisms of antifungal drug resistance*

One factor inhibiting the development of novel antifungal drug targets is that *C. albicans*, like its human host, is a eukaryote and shares many conserved biological processes with its host cells. The shared properties of the eukaryote genomes makes the development of therapies difficult for *C. albicans* as well as for other fungal infections because of deleterious side effects from antifungal treatment. Additionally, the emergence of antifungal drug resistance is frequently found during the treatment of *C. albicans* infections (e.g., : (Siikala et al., 2010)). Antifungal drug resistance in *C. albicans* occurs through different, and potentially synergistic, mechanisms including genetic alterations and abnormal expression

of targeting proteins. For instance, resistance to azoles, a family of antifungals that targets the cell membrane, can arise if mutations appear in the azole-target gene *ERG11*, which is essential for ergostrol biosynthesis (Marichal et al., 1999; Sanglard et al., 1998; White, 1997a). Mutation of the transcription factor *TAC1* leads to resistance via constitutive expression of the multidrug transporters Cdr1 and Cdr2 responsible for efflux of antifungal drug out of cells (Coste et al., 2006; Sanglard et al., 1997; 1995; Siikala et al., 2010; White, 1997b). The echinochandin class of antifungals results in the destruction of the cell wall; mutations in the echinocandin target gene *FKS1*, which encodes the subunit of β-1, 3-glucan synthase, may render *C. albicans* resistant (Lee et al., 2012).

In addition to point mutations, other mechanisms of antifungal resistance have been identified recently. For instance, loss of heterozygosity, gross chromosomal rearrangements, and aneuploidy are now known as causes of drug resistance phenotypes. Point mutations, as described previously, alter the function of drug targeting genes. Loss of heterozygosity (LOH) may reveal existing mutations by decreasing allelic diversity. For example: the *MRR1* gene with homologous gain-of-function point mutations, P683S and G997V, was identified in clinical isolates with a drug-resistant phenotype (Franz et al., 1998; Morschhäuser et al., 2007). In addition, gross chromosomal rearrangements and aneuploidy have been identified in many *C. albicans* clinical isolates and confer azole resistance by increasing the copy number of drug target genes (Selmecki et al., 2006; 2008). The development of antifungal resistance occurs following selection of mutations and genome structural changes. These mutations and

genome-scale alterations result from failures in DNA replication and chromosome segregation. Therefore, to better understand mechanisms of drug resistance in *C. albicans* it is important to study its genome structure.

1.1.3. *Candida albicans genome*

The first *C. albicans* shotgun genome sequencing was achieved in 1996, and the assembled genome was published and displayed in the *Candida* Genome Database (www.candidagenome.org) (Jones, 2004). *C. albicans* has a 14.3 Mb genome size and contains eight chromosome pairs, called chromosome 1 to 7 and chromosome R. 6202 open reading frames (ORFs) in total have been identified. Interestingly, *C. albicans* is most frequently a highly heterozygous diploid. Comparison of the chromosome sequences reveals that 81% of the genome is heterozygous (Jones, 2004; van het Hoog et al., 2007).

Until recently*, C. albicans* has been considered an obligate diploid species, which lacks a complete sexual cycle. While a complete sexual cycle with meiosis has not been observed (Bennett and Johnson, 2003), a parasexual cycle that results in the production of haploids, diploids and tetraploids occurs in *C. albicans* to promote genetic diversity (Bennett and Johnson, 2003; Forche et al., 2008; Hickman et al., 2013). To initiate the parasexual cycle, *C. albicans* undergoes a white-opaque phase switch (Bennett and Johnson, 2003). White cells are round and form dome-shaped colonies, while opaque cells are more elongated and form flatter-shaped colonies. The white-opaque switch is regulated by transcriptional regulatory proteins encoded at the *MTL* (Mating type-

like) loci, as well as a network of interconnected transcription factors (Zordan et al., 2007). Importantly, the opaque cells have 10^6 times higher mating efficiency than the white cells. Mating of two haploid cells also requires switching to the opaque phase (Bennett and Johnson, 2003). In the parasexual cycle, a process known as 'concerted chromosome loss' reduces the number of chromosomes per cell from the tetraploid state to a near-diploid state generating progeny with a diverse range of genetic material through either gain or loss of chromosomes (Forche et al., 2008). A similar 'concerted chromosome loss' process is thought to produce haploids from diploid parents (Hickman et al., 2013)

Chromosomal shuffling events potentiate *C. albicans* to survive in adverse conditions, especially since *C. albicans* is well known to tolerate aneuploidy either *in vivo* (in host) or *in vitro* (in the laboratory) (reviewed in (Rustchenko, 2007; Selmecki et al., 2006; 2010)). Thus, under circumstances where *C. albicans* in a host environment encounters treatment with antifungal drugs, the aberrant karyotype from the parasexual cycle may render the cells drug-resistant *in vivo* (Forche et al., 2008). Even in the absence of drug treatment, commensal *C. albicans* are constantly challenged by the host immune system. In response to this stressful condition, the parasexual cycle may also potentiate the ability of *C. albicans* to survive in its host.

Aneuploidy and gross chromosomal rearrangements are commonly found in clinical *C. albicans* isolates with a drug resistant phenotype. The most striking example is the isochromosome 5 segmental aneuploidy consisting of two left arms of chromosome 5. Isochromosome (5L) confers azole resistance (Selmecki

et al., 2006; 2008), and more than 20 % of azole resistant strains have this specific aneuploidy event. In addition to the i(5L), ~50% of azole-resistant strains carry a least one extra-chromosome. The prevalence of aneuploidy in *C. albicans* emphasizes the importance of investigating the basic molecular mechanisms resulting in ploidy changes. Furthermore, since aneuploidy and polyploidy are hallmarks of cancer cells (Storchová and Pellman, 2004), understanding ploidy changes in the *C. albicans* genome may contribute to an understanding of cancer biology. A significant question is how the *C. albicans* genome faithfully replicates and segregates to daughter cells. The two critical elements responsible for these processes are replication origins and centromeres.

1.2. Initiation of DNA replication

Two fundamental processes are required for the proper inheritance of genetic information from the parents to the progeny: high fidelity DNA replication and proper segregation of the duplicated chromosomes. Failures during either step of the cell cycle can result in severe genome instability including aneuploidy (reviewed in (Holland and Cleveland, 2012)).

In 1922, while genes were still called “ultramicroscopic bodies”, Muller suggested that genes must be faithfully replicated; otherwise mutations can pass through to the daughter cells (Muller, 1922). After the discovery of double-stranded DNA in 1953, the mechanisms of DNA replication have been intensively

investigated. Jacob *et al* in 1963 proposed that there are both *trans*- and *cis*-acting regulators that facilitate the initiation of chromosome replication (Jacob et al., 1963). Every step of replication, initiation, elongation, and termination is crucial to maintain the fidelity of the genome. Defects in the process of DNA replication can result in genome instability and eventually lead to disease phenotypes, including predisposal to cancer (reviewed in (Masai et al., 2010)). Perhaps due to the physiological significance of this process, the basic replication machinery is highly conserved across all eukaryotes.

1.2.1. *Replication initiation factors*

To initiate DNA replication, a DNA double helix must open to allow the DNA synthesis machinery to copy each DNA strand. These unwound sites of DNA, termed 'replication origins', are recognized by conserved replication factors. There are three steps to initiate origin firing: first, recognition of the origin, second, assembly of a pre-replication complex (pre-RC), and last, activation of the pre-RC (reviewed in (Masai et al., 2010)). These three processes must occur only once at an origin per cell cycle. Many of the protein factors necessary for the initiation of DNA replication are well conserved in all eukaryotes. A critical component is the origin recognition complex (ORC), which first binds to origin DNA at late M phase (Bell and Stillman, 1992; Bell et al., 1993; Foss et al., 1993; Micklem et al., 1993). ORC has six protein subunits. ORC1, 2, 4, and 5 directly bind to DNA, and ORC3 and 6, which do not cross-link to DNA, are required for

ORC function in the budding yeast *S. cerevisiae* (reviewed in (Bell, 2002). In *Drosophila*, all six ORC subunits bind origin DNA. Importantly ORC requires ATP to bind origin DNA. Specifically, ATP binding of ORC1 has been shown to be important for the interaction of ORC with DNA (Chesnokov et al., 2001). ORC1, 4 and 5 all have potential ATP binding sites (Klemm et al., 1997). After origin DNA recognition, Cdc6, Cdt1, and the MCM2-7 complex (minichromosome maintenance protein) are sequentially assembled onto origin DNA at late M to early G1 phase to form the pre-RC complex and then to fire the origin (reviewed in (Branzei and Foiani, 2010)). More work has been done in *S. cerevisiae* since it is the simplest eukaryotic organism and amenable to many genetic manipulations to study conserved biological processes.

1.2.2. *Replication origins*

Sequence features:

In contrast to the high level of conservation of the eukaryotic replication machinery, origin DNA sequences differ considerably in different species. Origins in *S. cerevisiae* have been well characterized. Struhl and Stinchcomb *et al*. began the first step of understanding the locations and sequences of the chromosomal “replicators” in 1979 by identifying chromosomal replicators using episomal plasmids (Stinchcomb et al., 1979). The so-called replicators, replication origins, were defined by the ability to support autonomous replication (Struhl et al., 1979). Thus replicator DNA was isolated from the *S. cerevisiae*

genome and called “ARS” (autonomously replicating sequence) (Chan and Tye, 1980; Stinchcomb et al., 1979).

ORC specifically binds to a degenerate 11-17 bp AT-rich conserved sequence (5′-[A/T]TTTAT[A/G]TTT[A/T]-3′), termed the ARS consensus sequence (ACS) (Bell and Stillman, 1992; Broach et al., 1983; Diffley and Cocker, 1992). In addition to the ACS motif, downstream modular elements, called B elements, were found while studying *ARS1* (Marahrens and Stillman, 1992). B elements (B1, B2, B3, and B4) have no well-defined sequences, but B1 has been proposed to be an enhancer for ACS recruiting of ORC (Lee and Bell, 1997; Rao and Stillman, 1995; Rao et al., 1994). The entire region containing the B elements spans around 100 bp and has an intrinsic helical instability to facilitate origin initiation. The B3 element contains a binding site for Abf1, a replication factor with additional roles as a transcription factor. Interestingly, the B3 element sequence can also be replaced by other easily unwound sequences (reviewed in (Marahrens and Stillman, 1992; Newlon and Theis, 1993)). Together, origins in budding yeast are strongly ‘hard-wired’, meaning that origin initiation strongly relies on the primary sequence context, because the ACS is a critical motif for origin activity. However, the degenerate 11 bp ACS motif matches over 12,000 loci throughout the genome and only ~400 origins actively fire during the *S. cerevisiae* cell cycle (Nieduszynski, 2006). Thus the ACS alone is a poor predicator of origin location. Other determinants, beyond the sequence, contribute to origin selection.

In the fission yeast *Schizosaccharomyces pombe*, Clyne *et al.* performed analyses similar to those done in *S. cerevisiae* to identify origin sequences (Clyne and Kelly, 1995). Yet ARSs in *S. pombe* that are capable of supporting autonomous replication of episomal plasmids do not have an ACS motif. Instead, extended 0.5-1 kb sequences with multiple AT-hook motifs were identified as ORC binding sites (Chuang and Kelly, 1999; Maundrell et al., 1988). *S. pombe* ORC4, which contains an N-terminal AT-tract DNA binding motif, can directly bind the AT- rich motifs for origin initiation (Lee et al., 2001). When the AT-content of an origin is reduced, origin initiation is delayed (Okuno et al., 1999), indicating that the AT-rich feature is important for origin function. Similar to the high abundance of the ACS motif in the *S. cerevisiae* genome, in *S. pombe* these AT-rich islands can be found ubiquitously (Dai, 2005; Segurado et al., 2003). From AT-rich islands randomly chosen bioinformatically from the genome, 90% of selected AT-rich islands were detected with active replicating intermediates, indicative of functional origins (Segurado et al., 2003). Thus, *S. pombe* origins may have less constraints on the composition of origin sequences.

In another yeast species, *Kluyveromyces lactis,* which is evolutionarily related to *S. cerevisiae,* a unique 50 bp ACS motif was identified using a plasmid-based ARS screen (Liachko et al., 2010). This ACS motif contains the auxiliary B element within the 50 bp motif. Taken together, studies from different yeasts suggest that the existence of replication origins is crucial for genome

maintenance, but the sequence of individual origins can be diverse and rapidly evolve.

Many studies that identified replication origins in yeast species have exploited plasmid assays to determine origin sequences conferring ARS function. Unfortunately, the ARS assay did not work well in multicellular organisms, due to an apparent lack of specificity of pre-RC binding to DNA in higher eukaryotes (Krysan et al., 1993). This also resulted in a lack of usable plasmid shuttle vectors in most eukaryotes. Even though recent genome-scale methods and sequencing technology have advanced the progress of origin studies in these higher eukaryotes, origin sequence features remain elusive. Several lines of evidence suggest origins are correlated with transcription and are located at gene promoters (Cayrou et al., 2011; Karnani et al., 2010). Specifically, transcription start sites (TSS) may serve as hot spots for origin initiation.

The most well characterized origin in humans is located within the Lamin B2 gene. It contains AT-rich sequences and a CpG island within a 1 kb region (Abdurashidova et al., 2000; 2003). CpG islands are ~200 bp sequences with abundant CG sites and also are frequently found at gene promoters (Aladjem, 2007). Moreover, the β-globin and Mycs gene are also identified as origin initiation zones (Aladjem et al., 1998; Liu et al., 2003; Waltz et al., 1996). In the Chinese hamster, replication origins initiate within the 55 kb intergenic locus downstream of the amplified DHFR (dihydrofolate reductase) locus and between two converged transcriptional units. Multiple origins can actively fire at different loci within the 55 kb origin zone (Dijkwel and Hamlin, 1995). Interestingly, a CpG

island sequence was not found in the DHFR or beta-globin origin zone, but AT-rich sequences are present, suggesting that these AT-rich sequences are preferential initiation sites. Moreover, Altman *et al* showed that the sequence of the lamin B2 origin is sufficient for origin initiation at ectopic locations. Additionally, the AT-rich sequences from the human lamin B2 gene can also function at the Chinese hamster origin zone, suggesting similar sequences are functional for origin replication among mammalian species (Altman and Fanning, 2004). Although there is no definitive sequence for origin initiation in metazoans, the results obtained to date imply that sequence structures, such as ability to for G-quartets, may contribute to origin function (Cayrou et al., 2011).

Chromatin features:

Since sequence features to define origins in metazoans remain loosely defined, local chromatin structures may also determine origin selection during the cell cycle (reviewed in (Masai et al., 2010)). Origins localize to regions with open chromatin. In *S. cerevisiae*, ORC not only binds nucleosome free regions (NFRs) but also facilitates nucleosome positioning. Lipford *et al*. performed both in *vivo* and in *vitro* assays to show origin activity depends on the ORC dependent nucleosomal array adjacent to the origin (Lipford and Bell, 2001). Most interestingly, the disruption of positioned nucleosomes affects pre-RC loading, rather than ORC binding, at origins. Specifically, properly positioned nucleosomes around the ACS are required for origin function. The nucleosome positioning at NFRs may allow the pre-replication complex (pre-RC) to have enough space to load onto chromosome DNA at the right sequence position

(Eaton et al., 2010a; Lipford and Bell, 2001). As we described above, the ACS motif appears much more frequently than *bona fide* active origins in the *S. cerevisiae* genome. Therefore, beyond the sequence level, local chromatin structure is a signature of origin function and selection.

The “open” or “closed” state of chromatin can be changed by histone modifications. The absence of Rpd3, a histone deacetylase, can alter the replication timing of late origins such that they fire earlier in S-phase (Knott et al., 2009; Vogelauer et al., 2002). Sir2, another histone deacetylase, also has been shown to negatively regulate origin activity in a subset of *S. cerevisiae* origins (Crampton et al., 2008). Histone acetylation is an open chromatin mark and thus contributes to the formation of NFRs for origin function at ACS loci. In studies using plasmids carrying *ARS1*, acetylation of histone H3 and H4 specifically around the *ARS1* origin can facilitate origin activity (Unnikrishnan et al., 2010). Moreover, in metazoans, gene promoters serving as a preferential origin initiation region also have high acetylation levels or other histone modifications that indicate open chromatin (reviewed in (Aladjem, 2007)). Chromatin modifiers involved in the link between transcription and origin will be important to study the epigenetic regulation of replication origins in higher eukaryotes.

Recently, genome-scale approaches combined with high-throughput assays have been intensively used to map the origins in different species (Gilbert, 2010). Either protein components of the replication machinery or the known nucleic acid features were used as landmarks of origins in these studies. Chromatin immunoprecipitation using antibodies against pre-RC complex

members, combined with microarray or next generation sequencing technology has been used extensively to specify origins (Gilbert, 2010; 2012; Ryba et al., 2011). In addition, the copy number of replicated DNA can be monitored because the regions close to the replication origin have double the DNA copy number in S-phase relative to regions farther from origins (Hansen et al., 2010; Koren et al., 2010a; Schübeler et al., 2002).

1.2.3. *Autonomously replicating plasmids*

Early studies of origins in *S. cerevisiae* were based on the plasmid system. From the identification of primary origin sequence to mapping chromatin features at origins (Stinchcomb et al., 1979; Unnikrishnan et al., 2010), episomal plasmids carrying ARS have provided important insights into replication. Plasmid shuttle vectors are one of the most important molecular tools for genetic manipulation in biology. In 1952, Joshua Lederberg first introduced the term “plasmid” (Lederberg, 1952). Plasmids are usually small, circular double-stranded DNA fragments and are commonly found in bacteria, archaea, and many eukaryotic organisms. In yeasts, the 2 μm circle was the first well-documented extra-chromosomal DNA to be stably maintained in the nucleus. The 2 μm circle plasmid is composed of two inverted repeats, four genes, and, of course, an autonomously replicating sequence (ARS) (Futcher, 1988). Interestingly, the extra-chromosomal 2 μm circle, like the chromosomes, initiates replication only once per cell cycle, but during this single round of replication it can accumulate up to 50-100 extra copies. Thus, the 2 μm circle has been modified to be an expression vector used in yeast genetics. This was the starting point to establish

a powerful molecular tool for genetic analyses in yeasts.

The minimal *cis*-acting element required for the maintenance of the plasmid circle is an origin of replication. While an ARS replicates the plasmid faithfully with chromosomes during S-phase, plasmids with only ARS regions segregate randomly and have high loss rates during cell division. Thus, additional elements are required for the stability of plasmids. The addition of a selectable marker gene and/or a centromere dramatically increase plasmid stability (Clarke and Carbon, 1980). Different selectable markers have been used, such as the metabolic genes *URA3* and *LEU2* in auxotrophic laboratory yeasts, and the antibiotic resistance genes *NAT* (nourseothricin) and hygromycin B (Chee and Haase, 2012; Krügel et al., 1988). Importantly, in *S. cerevisiae*, centromeres are “hard wired” with consensus sequences and, thus the 150- 200 bp region containing the centromere DNA can be cloned on the plasmid to facilitate assembly of the segregation machinery (reviewed in (Cleveland et al., 2003)). Together, the selectable marker and the centromere on the plasmid vector enable plasmid maintenance. When a plasmid vector carries the ARS, the centromere and a selectable marker, the copy number of the plasmid is 1-2 copies per cell. Additionally, integrating plasmid vectors relying on homologous sequences on the vector to insert DNA at a specific genomic locus via recombination have been developed (Chee and Haase, 2012; Parent et al., 1985).

Yeast plasmid vectors have been used extensively in genetic manipulations; for instance, protein expression, gene disruption, and

complementation. In the early 1980's, the most remarkable discovery using plasmid vectors was the identification of ARSs in *the S. cerevisiae* genome. The sequences first identified were: *ARS1* around the *TRP1* gene (Tschumper and Carbon, 1980), *ARS2* near the *ARG4* gene (Hsiao and Carbon, 1979), and *ARS3* adjacent to the *SUP11* gene (Feldmann et al., 1981). Furthermore, Chan and Tye isolated the ARSs from the yeast genome based on high transformation efficiency of plasmids containing fragments from a genomic library (Chan and Tye, 1980). This was the first genome-wide mapping of *cis*-acting elements for origins of replication. Later, linear plasmids mimicking natural linear chromosomes were developed and shown to be stably segregated during both mitotic and meiotic cell cycles (Dani and Zakian, 1983; Murray and Szostak, 1983).

Circular plasmid vectors have been developed for other yeast species as well. In 1980, the structures of centromere and replication origins were still unknown in *C. albicans*. Cannon *et al* used the conventional method from Chan and Tye's work to clone genomic DNA fragments for isolation of functional ARSs from the *C. albicans* genome (Cannon et al., 1990). Two *C. albicans* ARSs were identified; they further proposed an 11 bp AT rich sequence could be the consensus sequences for ARS function. However, the isolated *C. albicans* ARS sequences have not been comprehensively characterized or further addressed. Later, two other studies isolated two DNA fragments potentially containing ARS function on a conventional plasmid based on high yeast transformation efficiency (Beckerman et al., 2001; Herreros et al., 1992). One DNA fragment is a chimeric

DNA from three different regions of the genome; the other is a gene encoding IMP dehydrogenase, *IMH3*. A major caveat of these studies is that circular plasmid DNA in *C. albicans* rapidly integrates into the genome after the process of yeast transformation, despite the fact that origin DNA should have the ability to maintain autonomous replication of plasmids. None of the previous studies presented direct evidence of origin function in those DNA fragments. Thus, prior to my studies, no reliable plasmid system was establilshed as a genetic tool for *C. albicans*.

1.3. Centromeres

After the entire genome is replicated in S-phase, duplicated chromosomes need to segregate equally into daughter nuclei for genome integrity and conservation. Large-scale chromosome abnormalities may result from the failure of faithful chromosome segregation (reviewed in (Verdaasdonk and Bloom, 2011)). In the late 1880s, Walther Flemming first named this specific chromosomal locus as a “centromere”, meaning a densely packed constriction of the mitotic chromosome identified by cytology (Flemming, 1880). During mitosis, the centromere, a specialized DNA locus present in only one copy on each chromosome, serves as the attachment site for spindle microtubules. Centromeres assemble a large proteinaceous complex termed the kinetochore, which interacts with mitotic spindle and directs chromosome segregation. The structure of the centromere differs extensively among different species.

1.3.1. *Centromere structure*

In *S. cerevisiae*, centromeres are called “point centromeres” because they are small (<200 bp), defined by DNA sequence elements, and structurally simple. Point centromeres are also found in other yeasts within the *Saccharomycetaceae* family including other *Saccharomyces* species as well as *K. lactis* and *Candida glabrata*. The primary sequences of the *S. cerevisiae* centromere contain three functional domains; centromere DNA elements: CDEI (8 bp), CDEII (78-86 bp), and CDEIII (26 bp). CDEII is AT-rich (~90%), and CDEI and CDEII contain binding sites for centromere binding proteins (Clarke and Carbon, 1983; Fitzgerald-Hayes et al., 1982; Hieter et al., 1985; Keith and Fitzgerald-Hayes, 2000). The centromere DNA region interacts with a centromere-specific histone H3 variant, Cse4, that replaces the canonical H3-containing nucleosomes. Highly phased nucleosome arrays flank the centromere ((Chen et al., 2000), reviewed in (Allshire and Karpen, 2008)). Cse4 is the yeast homolog of a human kinetochore protein CENP-A and it interacts primarily with the CDEII element. The Ndc10 protein is another factor essential for centromere function in organisms with point centromeres; its sequence-specific interaction with CDEIII specifies assembly at centromere DNA sequence elements (Jiang et al., 1993). Importantly, centromere function is determined by specific protein-DNA interactions that result in kinetochore assembly. Modification of histone subunits, such as H4, influences centromere function (Black et al., 2007; Sullivan and Karpen, 2004). Additionally, many other centromere binding proteins, including Scm3, Mif2, and Cbf1-Cbf3

provide a broad interacting surface for the attachment of the outer kinetochore complex (reviewed in (Allshire and Karpen, 2008; Verdaasdonk and Bloom, 2011)).

More work on the detailed chromatin structures surrounding the core Cse4/CENP-A has been done in fission yeast *S. pombe*. Compared to small, point centromeres in *S. cerevisiae*, *S. pombe* centromeres are much larger at 40-100 kb in size and are comprised of mostly inverted repeats. No consensus sequence is found at the central core region (Hahnenberger et al., 1991; Niwa et al., 1989). This “regional centromere” structure is similar to centromere structure in higher eukaryotes, such as *Drosophila* and humans, although the centromere structure in metazoans is much larger and more repetitive. Thus, *S. pombe* centromeres are considered a model for studying the structure of the mammalian centromeres (Allshire and Karpen, 2008). In addition, the most important feature of *S. pombe* centromeres is that heterochromatin assembles at centromeres and is highly important in establishing centromeric chromatin and centromere function. Specifically, histone H3 methylated at lysine 9 (H3K9) recruits Swi6, the homolog of the vertebrate HP1 (Heterochromatin Protein 1), to centromeres. RNA interference (RNAi) initiated from transcription of the repeats is essential for heterochromatin assembly (Folco et al., 2008).

The DNA structure of the *S. pombe* centromere consists of a central core region with complex flanking inverted repeat motifs, L (otrL ,*imrL*), and *ortR, imrR* which are arrayed with variable arrangement among strains. Notably an engineered minimal centromere consists of a central core (cnt1, cnt2, and cnt3)

and repeat sequences for a total of 12 kb DNA (Pidoux and Allshire, 2004); however function of the engineered centromere can be maintained with internal deletions of the repeats (Baum et al., 1994). The flexible requirement of repeat sequences at the centromeres indicates the functional redundancy of those sequence elements; however, the repetitive sequences facilitate centromere activation (Ngan and Clarke, 1997). Regional centromeres composed of repetitive sequences are conserved in most eukaryotes. For instance, centromeres in *Arabidopsis thaliana* contain 180 bp repeats, retroelements, and middle repetitive sequences (Copenhaver et al., 1999).

Mammalian centromeres are also regional centromeres consisting of large arrays of repetitive satellite DNA spanning from 250 kb to over 1.5 Mb (reviewed in (Black and Cleveland, 2011)). Due to the repetitive nature of the centromere, sequence assembly is challenging and the centromeres remain incompletely assembled in the human genome. Human centromeres are comprised of alpha satellite sequences containing tandem arrays of a 171 bp repeating unit with more than 30,000 copies (Willard, 1985). Similar to *S. pombe* and *Drosophila*, alpha satellite DNA does not directly specify centromere location, but facilitates centromere assembly via binding of the kinetochore protein CENP-B to sequence elements within the alpha satellite repeat (Okada et al., 2007). CENP-A nucleosomes are the essential determinant of a functional centromere in humans. CENP-A binding at centromere DNA is co-localized at the inner kinetochore plate with CENP-C to form a specialized structure that recruits

functional components of the kinetochore (reviewed in (Allshire and Karpen, 2008; Burrack and Berman, 2012a)). Chromatin, rather than DNA sequences, determines centromere function and identity in most eukaryotes. Similar to replication origins, the centromere binding proteins are mostly evolutionarily conserved, but no hard-wired conserved DNA exists to specify precise sites of CENP-A assembly in most eukaryotes.

1.3.2. *Neocentromeres*

In species with epigenetically defined regional centromeres, centromere DNA sequences by themselves cannot specify the centromere location and function. Evidence supporting the model that centromere sequence is not sufficient for kinetochore formation comes from studies of dicentric chromosomes containing two regions of centromere DNA. Stimpson *et al*. have shown that one of the two centromeres on dicentric chromosomes can be inactivated such that a dicentric chromosome has only one functional centromere locus, even though both sets of alpha satellite sequences remain at least partially intact (Stimpson et al., 2012). Additionally, the formation of 'neocentromeres' provides evidence that centromere sequence is not necessary for centromere function (reviewed in (Burrack and Berman, 2012b; Stimpson and Sullivan, 2010)).

Many experimental systems have shown that ectopic neocentromeres can form if the native centromere is inactivated. In *Drosophila*, when chromosome breakage is induced by γ-radiation, neocentromeres can form within the pericentric region (Maggert and Karpen, 2001). Moreover, CENP-A is sufficient to

recruit kinetochore assembly to ectopic chromosomal loci. Similar experiments have been done in *S. pombe*; centromere deletion resulted in neocentromere formation adjacent to telomeres (Ishii et al., 2008). Neocentromere formation in both *Drosophila* and *S. pombe* occurs at the borders of euchromatin and heterochromatin. However, the genomic features necessary for neocentromere formation are still not understood.

In human, more than 100 different neocentromeres have been identified on at least 20 chromosomes (Marshall et al., 2008). Neocentromere formation in humans may be induced via large-scale chromosomal rearrangements and chromosomal fusions. The abnormal genome rearrangements resulting in neocentromere formatin are indicative of genome instability. Neocentromere formation is also associated with cancers, such as non-Hodgkins lymphoma, acute myeloid leukemia, liposarcomas, and lung cancer (Amor and Choo, 2002).

While *S. pombe* is considered an ideal model to study epigenetic centromeres, the presence of 40-100kb repeats makes dissection of centromere function challenging. *C. albicans* has much smaller regional centromeres with epigenetic features. *C. albicans* centromeres have a 3-5 kb core region, and the sequences of each of the eight centromeres is unique (Sanyal et al., 2004). Interestingly, compared to other eukaryotes with regional centromeres, *C. albicans* centromeres have less repetitive DNA. The core regions of centromeres are flanked by repetitive sequences, but the type and length of the repeats differs at each centromere. Importantly, the small centromeres are epigenetically defined by Cse4 binding. Moreover, neither pericentric heterochromatin around

the *C. albicans* centromeric core region, nor a known homolog of HP1 and histone H3 lysine 9 methylation (or its modifier), has been identified in *C. albicans*. Thus, *C. albicans* has a small epigenetic regional centromere without highly repetitive sequences or a requirement for heterochromatin. The simple structure of centromeres in *C. albicans* is more amenable for genetic manipulations (Burrack and Berman, 2012a). Remarkably, Ketel *et al* deleted the native centromere 5 core region in *C. albicans* and were able to detect neocentromere formation (Ketel et al., 2009). Two distinct classes of neocentromeres, proximal neocentromeres and distal neocentromeres were identified. Proximal neocentromeres formed adjacent to the native centromere locus, which is similar to the neocentromere in *Drosophila*; distal neocentromeres formed at ectopic chromosomal loci with no apparent specialized genomic features.

1.4. Research objectives

The genome plasticity *of C. albicans* as well as its clinical importance makes us eager to understand its genomic structure. In my dissertation work, I investigated fundamental genetic elements in the genome: replication origins and centromeres. Here, the major goal was to investigate origin organization in the *C. alibcans* genome, and thus, to further understand the basic mechanisms resulting in the highly dynamic genome of *C. albicans*.

In chapter 2, I discovered the unexpected link between centromeres and replication origins in *C. albicans*. In replication timing profile experiments in *C. albicans*, the centromere region is the first place to replicate on each chromosome. I demonstrated that origins are located within the core region of centromeres and that they fire efficiently during the cell cycle. Surprisingly, origin function is linked to the epigenetically defined centromere. Kinetochore proteins recruit replication proteins to fire an origin. This is the first evidence of a epigenetic centromere specificying origin function and epigenetically programming the replication timing. Additionally, centromeric DNA is the first well-characterized origin in *C. albicans*.

In Chapter 3, I extended my studies on origins and developed a novel linear plasmid system, which can autonomously replicate in *C. albicans* cells. I also re-visited previously identified origins and conserved origins to understand how the plasmid behaves in *C. albicans*. This is a significant breakthrough for the *C. albicans* community and makes molecular studies more feasible in *C. albicans.*

In Chapter 4, I further mapped origin locations in the *C. albicans* genome utilizing multiple genome-scale approaches combined with bioinformatics methods. I successfully identified both active and potential origins in *C. albicans*. Moreover, I combined a traditional ARS screening method with high-throughout technology using a new plasmid system to discover sequence features of *C. albicans* origins. The analysis revealed a 15 bp ACS motif for origin activity. Thus, my work on origin characterization showed for the first time that a eukaryote with

epigenetically-defined regional centromeres has primary sequences required for origin function.

In this thesis, I have comprehensively characterized replication origins in the *C. albicans* genome. *C. albicans* has some genome features similar to highly sequence-defined organisms such as *S. cerevisiae*, and other genome features, such as epigenetically-defined regional centromeres and the ability to form neocentromeres, that more closely resemble genome features of higher eukaryotes. My studies indicate that *C. albicans* has a unique placement on the evolutionary tree, and thus, is a promising organism for further studies of genome dynamics.

Chapter 2.

Centromeric origins are associated with epigenetically-inherited centromere function in *Candida albicans*

This report was originally published as,

Koren A, **Tsai H-J**, Tirosh I, Burrack LS, Barkai N, et al. (2010) Epigenetically-Inherited Centromere and Neocentromere DNA Replicates Earliest in S-Phase. PLoS Genet 6(8): e1001068. doi:10.1371/journal.pgen.1001068

Introduction

Centromeres are specialized loci on each chromosome and are essential for proper chromosome segregation during cell division. Lack of a functional centromere can cause chromosomes to become unstable and result in gain-or-loss of chromosomes. The centromere is unique because it needs to be distinguished from other chromosome loci, especially as a recognition site for assembly of the kinetochore complex (reviewed in (Allshire and Karpen, 2008; Malik and Henikoff, 2009; Morris and Moazed, 2007)). In the budding yeast *S. cerevisiae*, primary DNA consensus sequences are sufficient to define centromere location; however, most eukaryotic centromeres are regional centromeres not defined at the DNA sequence level, but instead are inherited epigenetically. It is paradoxical that over evolutionary time, centromeric DNA sequences differ among closely related species while the major centromeric proteins/kinetochore remain conserved (Malik and Henikoff, 2009). Specifically, a unique, conserved histone H3 variant, termed CENP-A/CenH3 (Cse4 in yeasts), defines eukaryotic centromeres epigenetically. As CENP-A/Cse4 forms a functional centromere, the position can be inherited for many generations at the chromosome epigenetically (Black and Cleveland, 2011). If a native centromere fails to function, in rare cases a neocentromere can form at an ectopic locus to maintain the chromosome by establishing CENP-A/Cse4 binding and kinetochore assembly (Burrack and Berman, 2012b; Ishii et al., 2008; Ketel et al., 2009; Maggert and Karpen, 2001). The mechanisms by which centromere and

neocentromere positions are stably maintained and determined remain elusive. Several models have been proposed. For instance, Cse4 can bind ectopic loci at a chromosome before a ubiquitin-mediated proteolysis mechanism degrades Cse4 bound to chromosome arms to preserve Cse4 binding exclusively at centromeres (Ranjitkar et al., 2010). Another interesting model from metazoans is that centromeric DNA replicates with a distinct timing and that drives Cse4 deposition along with the replication timing (Dunleavy et al., 2009; Foltz et al., 2009; Mellone et al., 2011). However, the replication timing of centromeres remains poorly understood due to the lack of fine resolution to accurately identify the sites of replication initiation. Moreover, most epigenetically inherited centromeres have a complicated, repetitive structure. For example, in humans, the centromeres span several hundred Kb to a few Mb in length and contain repetitive DNA sequences that have not been fully assembled in the human genome (Zeitlin, 2010). Thus, the functional correlation between replication initiation timing and centromere identity has been puzzling.

C. albicans has epigenetically inherited centromeres similar to most eukaryotes, and its regional centromeres are short (3-5 kb) without complicated repetitive sequences (Sanyal et al., 2004). Intriguingly, when the *C. albicans* centromere is deleted, neocentromeres can form at ectopic loci (Ketel et al., 2009). Similar to human neocentromeres, *C. albicans* neocentromeres are found at many locations on the chromosome arms and form independently of canonical heterochromatin. Since the *C. albicans* genome has been fully sequenced,

including the centromere region, and shares many epigenetic features with higher eukaryotes, *C. albicans* has become an attractive model organism to study how the replication initiation coordinates with centromere function. Furthermore, neocentromere formation permits the separation of the functional centromere from the genome context of the native centromere.

In Koren et a*l.,* we profiled replication timing throughout the entire genome of *C. albicans* (Koren et al., 2010b). The replication timing profile was based on the comparison of DNA copy number at different genomic regions. Asynchronized log phase cultures were sorted by flow cytometry into G1 and S-phase fractions to separate the cells into unreplicated and replicating pools, respectively. DNA from G1 and S-phase fractions was extracted, labeled with different fluorescent dyes and hybridized onto genomic microarrays (Koren et al., 2010a). Thus, the difference in DNA copy number at each genomic region during S-phase relative to G1 was revealed; regions with higher copy numbers of DNA represented relatively early replicating regions. Thus, peaks along a chromosome indicated the locations of replication origins (Figure 2.1).

Approximately 150 peaks were described in Koren *et al* (Figure 2.1, Table 2.1). Remarkably, the earliest initiation of replication on all eight chromosomes occurred near centromeres. For all chromosomes, the estimated origin positions were on average 13 kb distal from centromeres. Thus, the centromeric region is the first place to replicate while cells enter S-phase. A weaker correlation

between early replication and centromeres has also been observed in *S. cerevisiae,* but this is the first observation in eukaryotes with strong correlation between centromere location and replication timing. Furthermore, Koren *et al.* analyzed replication timing of centromeres in *S. pombe* and found that centromeres also replicate early in fission yeast (Koren et al., 2010b). Together, these results suggest that early replication of centromeres may be a general feature of fungi.

In this chapter, I further characterized the mechanisms functionally linking centromere function and early replication using biochemical approaches. My work also indicated that centromeric origins are distinct and regulated differentially from origins at chromosome arms.

Materials and Methods

Yeast strains and growth condition

Yeast strains used in this chapter were *Candida albicans* SC5314 and RM1000 #2 (YJB7617). Strains were stored as stocks in 50% glycerol at -80 °C and grown in YPAD medium (1% yeast extract, 1% peptone, 2% glucose, 1.5% agar) at 30°C

Chromatin immunoprecipitation (ChIP)

The procedures of chromatin immunoprecipitation were modified from Ezhkova *et al*. and Ketel et al. specifically for ORC or MCM ChIP experiments. Yeast

strains were cultured overnight in 30 °C and then were diluted to an OD_{600} of 0.2 in YPAD medium. Log-phase cells were harvested at OD_{600} of ~1 after 3 hours growing time. Cells were fixed in 1% formaldehyde for 20 minutes at 30 °C and 125 mM for 5 minutes at room temperature. Cells were spun down at 4°C, 2500-3000 rpm for 5 minutes, and then the cell pellets were stored or taken to the next step after discarding the supernatant. Each cell pellet was suspended in 450 µl lysis buffer (50 m*M* HEPES, pH 7.5, 140 mM sodium chloride, 1 m*M* EDTA, pH 8.0, 1% Triton X-100, 0.1% DOC, 0.1% sodium dodecyl sulfate, Roche proteinase inhibitor cocktails), and then equal volume of acid-washed glass beads added (approximately 500 µl). The cell pellets were vortexed either on a bead-beater for 1 minute each time, 6 times (one minute on, one minute off) or on a vortex for 10 minutes each time, 4 times (10 minutes on, 2 minutes off). The lysed cells were seen as broken cell shapes or phase-dark under the microscope. Puncture the bottom of the 1.7 ml tubes with needle and the tubes were placed into new tubes to collect the cell lysates by spinning down with full speed for 10 second. The collected cell pellets were vortexed again for 30 second in 4 °C, and then the cloudy supernatants were collected. The supernatants were placed on ice and sonicated to yield DNA fragments of 500-1000 bp (4 times of sonication, 15 second per time, output power 20 %). Soluble lysate was collected after insoluble matter was precipitated twice by centrifugation in 4°C, 10 minutes each time. The volume of collected soluble lysate was adjusted to 2 mL by lysis buffer and 5 % of it was collected as input lysate, which was stored at -80°C until the reverse-crosslinking step. The rest of

the cell lysate was split into two reactions with and without antibodies, respectively. After the reaction was incubated at 4°C overnight, protein A-agarose beads were added for another 3-4 hours incubation at 4°C. Then, the reaction was washed subsequently with cold lysis buffer, high-salt washing buffer (50 m*M* HEPES, pH 7.5, 500 mM sodium chloride, 1 m*M* EDTA, pH 8.0, 1% Triton X-100, 0.1% DOC, 0.1% sodium dodecyl sulfate), DOC buffer (10 m*M* Tris-HCl, pH 8.0, 0.25 *M* lithium chloride, 0.5% Nonidet- P40 (NP-40), 0.5% sodium deoxycholate (DOC), 1 m*M* ethylenediamine- tetraacetic acid (EDTA, pH 8.0)), and TE (10 m*M* Tris-HCl, pH 8.0, 1 m*M* EDTA, pH 8.0) in 4°C, 5 minutes for each washing step. DNA were eluted by adding 100 TES buffer (50 m*M* Tris-HCl, pH 8.0, 10 m*M* EDTA, pH 8.0, 1% SDS) twice in 65°C, 10 min for each time. Total of 200 µl eluted DNA was incubated overnight at 65°C and followed by RNase A and proteinase K treatment. Proteins were removed by phenol-chloroform/isoamyl alcohol/chloroform extraction. Samples were ethanol precipitated with 20 ug of glycogen overnight at 4°C. Finally, DNA was collected by centrifugation, washed in 70% ethanol and resuspended in 50-80 µl TE, depending on the input protein concentration.

Quantitative PCR

ChIP samples were analyzed by quantitative PCR using the Roche LightCycler 480 system. The reactions were performed in the following conditions: pre-incubation at 95°C for 10 minutes, 50 cycles of 95°C for 10 seconds and 57°C for 15 seconds, and a final cooling at 40°C for 10 seconds. Primers and Probes used in the reactions were designed by the online software, Universal Probe

Library Assay Design Center (Roche) and are listed in Table 2.2. Analyses were performed using LightCycler 480 software release 1.5.0 sp3. Relative enrichment ratio was calculated as amplification of immuoprecipitated DNA relative to total cell lysate (input).

DNA two-dimensional gel electrophoresis (DNA 2D gel)

The procedures of DNA 2D gel electrophoresis were modified from Leizmeier et al. and personal communications with Bonita Brewer's lab.

DNA preparation:

Yeast cells were grown overnight (OD_{600} of ~25) and subcultured to an OD_{600} of 0.8 in pre-warmed YPAD medium in 30°C for 40 minutes or 2 hr. Sodium azide was added to 0.1% and quickly placed on ice. Ice-cold EDTA was added immediately to final 25 mM before spinning cells at 4°C, 7000 rpm for 5 minutes. Cell pellet was washed by ice water before adding 0.2 ml ice-cold NIB buffer (17% glycerol, 50 mM MOPS buffer, 150 mM potassium acetate, 2 mM magnesium chloride, 0.5 mM spermidine, and 0.15 mM spermine; pH is adjusted to 7.2) to resuspend cell pellets in 15 ml tube. With 0.2 ml NIB, cell pellets can be frozen at -20°C or continued to the next step. 0.9 ml glass beads were added, and placed on a vortex to lyse the cells; vortexing 5-10 times for 2 minutes (2 minutes on; 1 minute on ice) or until cells were 75% lysed by seeing broken cells shape or phase-dark cells under the microscope. 0.6 ml NIB was added into the reaction to remove the supernatant to another tube. The step was repeated on ice to get 1.3 ml of cell lysates. Nuclei were spun down by centrifugation with full

speed at 4°C for 30 minutes, and then pellets were resuspended in 0.3 ml TEN buffer (50 mM Tris, 50 mM EDTA, 100 mM NaCl). RNase A was added on ice for 5 minutes before proteinase K was added with 12 µl of 25% Sarkosyl. The reactions were mixed gently by inverting and transferred to 37°C for 30 minute incubation. The lysed nuclei were spun down for 5 minutes, full speed at 4°C. The supernatants needs to be transferred gently with a cut pipette tip, and then DNA extracted by 0.3 ml phenol-chloroform/isoamyl alcohol/chloroform. The mixes were gently mixed by a dozen inversions. The final volume of extracted DNA solution at this step would be around 0.2-0.25 ml. 2 volumes of absolute ethanol containing 0.5 M potassium acetate were added, and then the reaction was slowly mixed by inverting. The reaction was rested on ice for 5-10 minutes until the white DNA spool was visualized at the bottom of the tube. After discarding the alcohol, DNA pellets were washed with 0.3 ml 70% ethanol for 5 minutes at top speed. DNA pellets can be re-dissolved in 75-100 µl of TE at 4 °C overnight.

DNA digestion and electrophoresis

20 µg of extracted DNA were digested using proper restriction enzyme, which can yield a 3-6 kb targeting DNA fragments. 300-400 unit of restriction enzyme was added in final volume 600 µl for 5-6 hours. Digested DNA was ethanol precipitated and was resolved in 18 µl of TE at the end for gel electrophoresis. The first dimension was in 0.35% agarose gel and ran by 1.2 V/cm, 20-22 hours; the second dimension was in 0.95% agarose gel and ran by 5.5V/cm for6-8 hours in 4 °C cold room.

Southern blot

DNA was transferred from agarose gels to nylon membranes (Amersham). Membranes were probed overnight at 65°C (Sigma, PerfectHyb prehybridization buffer). The probes were labeled with alpha-^{32}P-dCTP (Agilent, Prime-it II Random primer labeling kit). The primers used to make probes are listed in Table 2.2. The membranes were exposed to film for 3-5 days at -80°C.

Results

Assessment of origin activity from replication timing profile

The genome-wide replication timing profile provided high spatial resolution for all eight *Candida albicans* chromosomes and indicated that the centromere region replicated earlier than other regions on the chromosome. To confirm replication origin activity, I performed DNA two-dimensional gel electrophoresis to analyze the replicating intermediates at the genomic region of identified peaks from the replication timing profile (Figure 2.2). DNA 2D gel electrophoresis developed by Brewer and Fangman is considered the best approach to physically map origins of replication by distinguishing differential migration of replicating intermediates (Figure 2.2a) (Brewer and Fangman, 1987). A non-linear DNA molecule does not move at the same rate as a linear DNA molecule when both have equal molecular weight/length. The first dimension of agarose gel electrophoresis separates the replicating intermediates by their size; the

second dimension separates the replicating intermediates according to their shape (Figure 2.2b). Then, the DNA fragments of interest are probed by Southern blot to determine if replication intermediates occur within the region. The existence of bubble arcs indicate an actively firing origin. Because replication forks travel quickly, the method requires enrichment of replicating intermediates through the synchronization step of the cell cycle.

On chromosome 5, two peaks identified in the replication timing profile near the centromere tend to fire earlier than other putative origins (Figure 2.3). Thus, we first examined if these two regions serve as origins by DNA 2D gels. However, we cannot synchronize *C. albicans* cells to enrich the replication intermediates using methods commonly used in *S. cerevisiae* due to differences in the mating cycle of *C. albicans* and in its responses to cell cycle arrest. Alternatively, we grew cells overnight and sub-cultured the cells into fresh medium to allow G0/G1 (stationary) cells to enter S-phase. In parallel, cells were treated with Hydroxyurea (HU) to stall replication fork progression to enrich for replicating intermediates. We also examined regions proximal to the replication timing peaks where Y-arc intermediates should be detected if an origin of replication exists within the peak. Unexpectedly, bubble arc intermediates were not detected at any of the three regions (Figure 2.4). The results have been reproduced several times to confirm that no origin activity can be detected by 2D gel electrophoresis at the exact predicted origin regions from the replication timing profile. Nevertheless, at the right chromosome arm region near CEN5, strong Y-arc intermediates were detected indicating that replication initiation

occurs near the centromere even though the other regions I examined were not origin positions.

While the resolution of the raw data for the replication timing was high, smoothing algorithms were necessary to analyze the replication timing on a genome-wide scale and the smoothing algorithms likely compromised the resolution of the chromosome-scale data. An origin position could be anywhere in a 20 kb window surrounding the peaks (Koren et al., 2010b). We ruled out the possibility that those putative origins may not fire efficiently in each cell cycle because the origins mapped in the replication timing profile should be the most early and efficient origins.

The centromere core contains an actively firing origin

We took advantage of the conserved nature of the replication machinery proteins to determine origin locations surrounding the centromere regions. Origin recognition complex (ORC) is a heterohexamer complex targeted to origin positions to recruit replication machinery (Bell and Stillman, 1992). To increase the resolution of origin position mapping, I performed chromatin immunoprecipitation (ChIP) using polyclonal anti-ORC antibodies to identify ORC binding sites as putative replication initiation sites. In parallel, I also performed Cse4 ChIP as Cse4 binding defines the centromere core. Using quantitative PCR, I tiled the entire centromeric region including the putative origins from the replication timing profile and the centromere core. The ChIP experiment had a

resolution of approximately 0.5kb to specify the ORC binding sites. As expected, Cse4 binding was detected at the *CEN5* core region but not at the right arm of CEN5 near the previously identified replication timing peak (Figure 2.5, left). Surprisingly, ORC binding was not detected at the peak region (Chr5R) but was highly enriched at the CEN5 core region (Figure 2.5, right). Thus, the result suggested that *CEN5* core region was bound by ORC as a potential origin. Moreover, it is reasonable to hypothesize that centromere core regions have a distinct position effect and origins near or at centromeres fire earlier and more efficiently than others on a chromosome.

The observation that centromere regions have ORC binding is intriguing because centromeric ORC binding has not been identified in other yeast species. *S. cerevisiae* and other closely-related yeasts have a consensus sequences (*CDEI, II, III*) for centromere identity as well as sequence motifs for origin function but do not have ORC binding at centromere sequences. *C. albicans* has an epigenetic regional centromere and origin sequences have not yet been characterized. To further investigate the role of ORC binding at the centromere in *C. albicans*, I examined ORC binding at the *CEN5* region by tiling with qPCR primers spanning the 3 kb of the centromere core region and 6 kb of flanking region at both arms. Interestingly, ORC binding was highly enriched throughout the entire *CEN5* core region (~3 kb) while the average ORC binding length is less than 1 kb in *S. cerevisiae* (Figure 2.6). Most importantly, ORC binding is coincident with Cse4 binding at the *CEN5* core. The surprising co-localization of

Cse4 (centromere) and ORC (replication origin) indicated that the two processes may be even more closely regulated than I expected initially.

To assess if the centromere core can actively fire as an origin, I performed DNA 2D gels probing the *CEN5* core region and its proximal region. I indeed detected bubble-arc intermediates indicating an active origin within the *CEN5* core region (Figure 2.7, right). Additionally, the *CEN5* proximal region only has Y-arc intermediates and strong X-arc intermediates (Figure 2.7, left). The results suggest that the centromere core contains an active origin and that multiple origins nearby could also fire actively during the cell cycle. Most importantly, the centromeric core region bound by Cse4 can serve as an active origin.

The biochemical approaches presented here suggest that the epigenetically inherited centromere in *C. albicans* has origin activity. This unexpected finding is the first evidence suggesting that an origin at a centromere with ORC binding may not rely on the recognition of origin consensus sequences for origin function. My working hypothesis is that ORC binding as well as origin activity at centromeres might be related to centromere identity in an epigenetic manner. Also, the centromere is the first identified *bona fide* origin in *C. albicans*.

Origin activity at centromere is associated with centromere function

To test our hypothesis that origin activity is linked to the functional centromere on a chromosome, we took advantage of epigenetic features of *C. albicans* regional centromeres. Ketel *et al*. showed that neocentromeres form at ectopic loci when the native centromere is deleted (Ketel et al., 2009). One

neocentromere in a *cen5* deleted strain formed at the 170kb region on chromosome 5 as detected by Cse4 binding. The parent and neocentromere strains provided a unique opportunity to address the order of the dependency between centromere and origin function. The hypothesis is: if origin function at centromeres is determined by centromere function, the neocentromere position (*neoCEN*) should become an origin following neocentromere assembly.

Accordingly, we performed a replication timing profile in the *cen5* deletion strain with the 170 kb neocentromere. Strikingly, the first place to replicate on chromosome 5 shifted to the *neoCEN* position when the native centromere was deleted. (Figure 2.8) This strongly suggested that changing the epigenetically defined centromere location altered the replication temporal program. As we described previously, the replication timing profile only provided us a chromosome scale spatial view of origin locations. To further explore our hypothesis that origin function was linked to centromere identity, I performed ORC ChIP to ask if Cse4 and ORC were co-localized with each other at the neocentromere as was previously observed at the native centromere. Interestingly, when the native centromere was intact (wild type strain), there was no ORC binding at the 170 kb *neoCEN* region (Figure 2.9, upper panel). This indicated that the neocentromere did not form at a pre-existing origin location. When the native centromere was deleted and the neocentromere formed at the 170 kb *neoCEN* position, I detected ORC binding co-localized with Cse4 (Figure 2.9, lower panel). The results suggest that a functional centromere can recruit an origin. This was an exciting observation because in yeast with well-know

replication origin features, such as *S. cerevisiae*, an origin functions depending on definitive sequences and the genome context. In our work, the replication origin at the centromere correlated with an epigenetically defined centromere.

Centromere determinant, Cse4 recruits ORC.

To further ask the question how a functional centromere recruits a replication origin, we used a conditional *PCK1* promoter to manipulate the level of *CSE4* expression (Sanyal and Carbon, 2002). Thus, we could test the effect of altering Cse4 protein levels on the occupancy of ORC at the centromere. When the *CSE4* conditional strain was grown in medium with glucose, the expression of *CSE4* under control of the *PCK1* promoter is repressed. In contrast, when the strain was grown in medium with succinate, the *PCK1-promoter fused CSE4* is over-expressed. I grew the strains, including a wild-type strain, overnight to stationary phase and sub-cultured them into glucose and succinate medium, respectively. Then I performed Cse4 and ORC ChIP to examine binding of these proteins at *CEN5*. Consistent with our hypothesis that a functional centromere recruits an origin, over-expression of *CSE4* resulted in increased levels of ORC binding at the *CEN5* core (Figure 2.10, right) while repression of *CSE4* expression resulted in reduced levels of ORC binding (Figure 2.10, left). Thus, Cse4 recruits centromeric ORC binding either directly or through recruitment of an intermediate factor. Taken together, we conclude that centromere function confers both the presence of an active firing replication origin and the distinct early replication timing during the cell cycle.

Centromeric origin is distinct from others at chromosome arms.

Unlike other yeast species, no origin sequences have previously been defined in *C. albicans*. We identified the epigenetically inherited centromere as an efficient origin, however this origin may be determined differently from other origins on the chromosomal arms and we therefore sought to identify additional non-centromeric origins. We hypothesized that the centromeric origin may be a distinct origin that constitutively fires during the cell cycle where other origins may have separate mechanisms of regulation.

To understand how replication origins are organized in the *C. albicans* genome, we used genome-scale approaches to map origin locations. I performed ORC ChIP-on-chip to specify all the potential origins throughout the entire genome. As expected, all 8 centromeres have strong ORC binding (Figure 2.11). We further used DNA 2D gels to examine origin activity at numbers of ORC binding sites to find actively firing origins (described in the Chapter 4). A specific ORC binding site located at the 246 kb region of chromosome 4 was identified as an active origin during the cell cycle. Thus, we can compare the difference in regulation between the centromeric origins and the newly identified chromosomal origins to test our hypothesis that centromeric origins have distinct regulatory mechanisms.

In metazoans, no definitive sequence motif is required for origin activity. However, transcription start sites are hotspots for origin locations due to the local chromatin structure (Cayrou et al., 2011). To fire an origin, the local chromatin

needs to be in an “open” status to allow replication machinery access to initiation sites. Thus, histone modifications may be involved in determining the local chromatin structure for origin activity. Set1 is a methyltransferase and required for histone H3 lysine 4 methylation in *C. albicans* (Raman et al., 2006). Histone H3K4 methylations are enriched at transcription start sites and 5’ regions of ORFs (Li et al., 2007). To address the question of whether centromere origins are regulated differently than non-centromeric origins, I performed ORC ChIP in wild type and *set1* deletion strains, respectively. Interestingly, at the chromosomal origin region, ORC occupancy was strongly reduced in the *set1* deletion strain (Figure 2.12, bottom) while ORC binding remained the same at the centromere region in both strains (Figure 2.12, top). The result suggests that ORC binding, which is an important determinant of origin function, is regulated by *trans*-acting factors such as Set1. Importantly, centromeric origins are a distinct class of origins and are not affected by Set1. Moreover, I tested if histone H3K4 methylation can be detected within the centromeric region by performing ChIP experiments using anti-H3 and anti-H3K4me3 antibodies, respectively. I tiled the entire centromeric and pericentromeric regions by quantitative PCR and found that the H3 binding and H3K4 methylation levels were low within the centromere core region as well as the inverted repeats (Figure 2.13). This could explain the observation that centromeric ORC binding is not affected by deletion of Set1. The result also suggests that Cse4 and potentially other centromere determinants could be the key factor to promote constitutive, efficient origin activity at centromeres.

Discussion

In this chapter, my work combined with the replication timing analyses in Koren *et al.* provides the first comprehensive analysis of an efficient origin of DNA replication within an epigenetically inherited centromere. I used several approaches to demonstrate that centromeric determinants can influence DNA replication timing.

This provides evidence that two fundamental mechanisms of genome inheritance, genome replication and chromosome segregation, are linked. Centromeres are essential for faithful chromosome segregation. We suggest that the epigenetic inheritance of early centromere replication may be important to ensure the inheritance of centromere identity. After we published this work, Pohl *et al.* found that in *S. cerevisiae* centromeres have a positional effect on origin activity and enhance early replication timing of the pericentric origins (Pohl et al., 2012).

C. albicans has regional centromeres with a different primary centromere sequence on each chromosome. During evolution, centromere DNA sequences are highly divergent among species because the mutation rates at centromeres are higher than those at intergenic regions (Padmanabhan et al., 2008). Despite the lack of sequence homology, Koren *et al*. found that an interesting evolutionary trace pattern appears in all eight *C. albicans* centromeres. An asymmetric GC skew pattern indicating constitutive replication origin activity was detected (Koren et al., 2010b). The GC skew is a pattern found at hundreds of

bacterial, archeal, and viral centromeres in those organisms with only single replication origin in their genome (Sernova and Gelfand, 2008). The skew pattern implies the presence of a replication origin and is the consequence of mutations that accumulate over evolutionary time (Pavlov et al., 2002; Sernova and Gelfand, 2008). In most eukaryotes, this pattern is not detectable because most eukaryotic origins are not constitutively active in each cell cycle. The identification of the GC skew patterns at centromere origins in *C. albicans* provides further evidence that centromeric origins are distinct and constitutive origins.

Combined with the studies in Koren *et al.*, we propose a self-reinforcing, positive feedback loop model (Koren et al., 2010b). Centromere determinants can recruit replication factors at a distinct time while cells start entering S-phase. This early replication event facilitates the recruitment of centromeric determinants such as Cse4 to the earliest replicating locus when Cse4 expression level peaks at the very beginning of S-phase. By comparing a strain with a native centromere with a strain with a neocentromere, we were able to determine that origin recruitment depends on the presence of a functional centromere, but that neocentromere formation does not require a pre-existing origin. In addition, I performed ORC ChIP-chip to map replication origins throughout the entire genome. Over 900 ORC binding sites were identified. I correlated the ORC binding sites with all the neocentromere locations identified in our lab to see if neocentromere formation preferentially occurs at pre-existing origins (over 20 genomic loci can be neocentromere locations; Laura Burrack, personal

communication). We did not observe a significant correlation between neocentromere locations and ORC binding sites. Thus, the centromeric origin may be essential for centromere maintenance, but not initial assembly.

Furthermore, we tested the hypothesis that the centromeric origin is distinct from chromosomal arm origins and that *trans*-acting regulation of these two types of origins differs. Base on our results, we propose that the centromeric origin fires constitutively since Cse4 has to be incorporated in precise timing at each cell cycle. Since *trans*-acting factors that influence chromosome arm origins, such as chromatin modifiers, are not involved in the regulation of centromeric ORC binding, Cse4 recruits ORC constantly in each cell cycle to replicate at the beginning of S-phase. I have performed ORC ChIP-Seq experiments in both wild type and *set1* deletion strains to determine the effect of H3K4me on ORC binding on a genome-wide scale. Interestingly, only a subset of chromosomal ORC binding sites were regulated by Set1 (data not shown). We are now working to investigate additional mechanisms resulting in the differential regulation of centromeric and chromosomal origins.

In conclusion, we propose that the inheritance of centromere position is correlated to the constitutive early firing of centromeric replication origins. Additionally, we suggest that centromeric origins are a distinct class of origins with characteristic epigenetic features, such as Cse4 binding and low levels of H3K4 methylation.

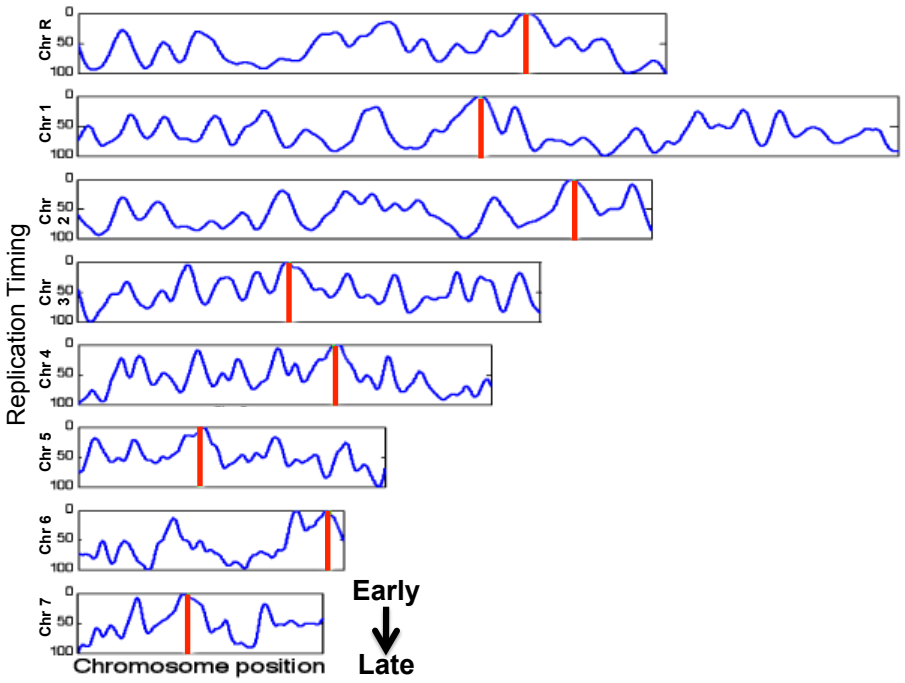
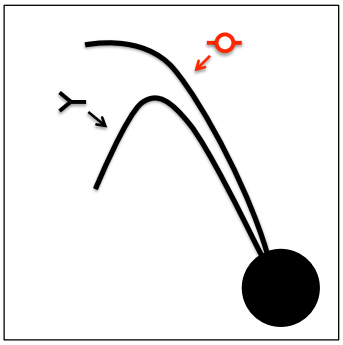


Figure 2.1. Centromere regions replicate first at each chromosome.

Experimental method to measure DNA replication timing has been shown in Koren et al. The earlier a certain locus replicates, the higher its DNA content will be. The higher peak along the chromosome represents earlier initiation timing. Red bars are the locations of centromeres at chromosomes.

A



B

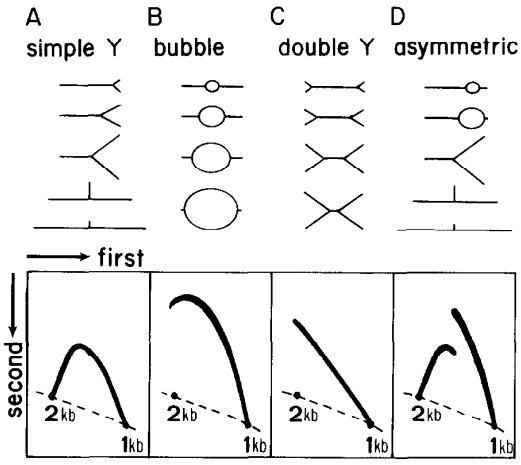


Figure 2.2. DNA 2D gels indicate origin activity.

(A) The first dimension is from left to right; the second dimension is from top to bottom. Bubble arc intermediates represent an initiation site of replication in the fragment (an origin). Y-arc intermediates represent replication forks in the fragment. (B) Cartoons of replication intermediates. The first dimension is to separate the sizes of the fragments; the second dimension is to separate the shapes of the fragments. The diagram is modified from Brewer and Fangman, 1987.

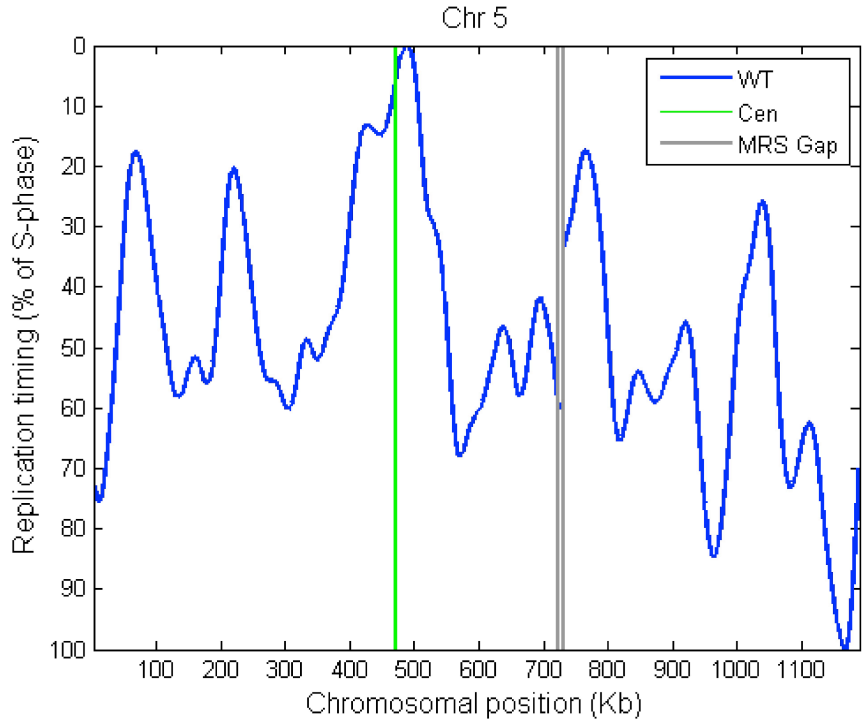


Figure 2.3. Centromere region is the earliest place to replicate at the chromosome 5. The position of *CEN5* is indicated with a green line. The blue line represent the replication timing profile. Two highest peaks are near *CEN5*.

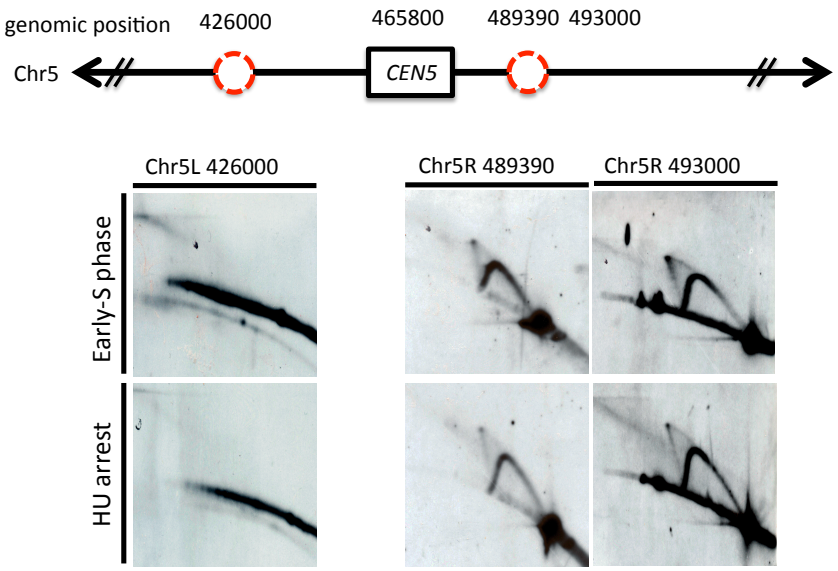


Figure 2.4. Assessment of origin activity at putative origin locations

Putative origins from the replication timing profile were examined by DNA 2D gels. The red circles represent the locations of tested fragments. No bubble arc intermediates were detected

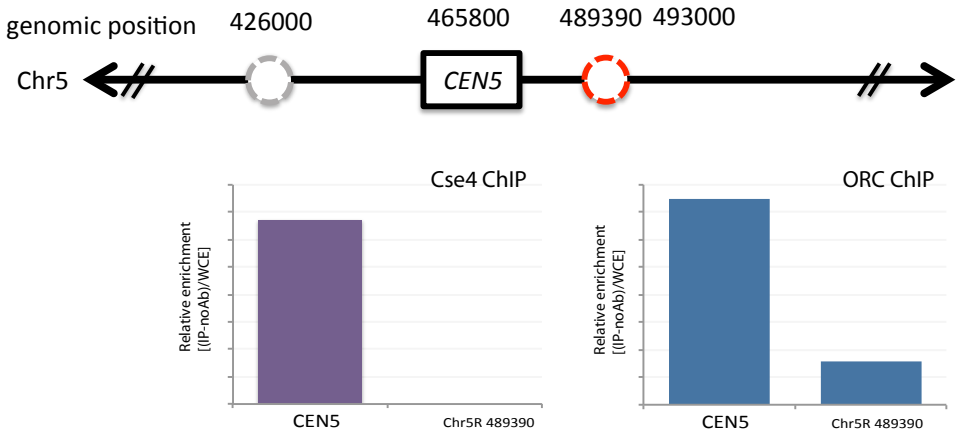


Figure 2.5. ORC binds *CEN5* core region.

ChIP experiments were performed to test the Cse4 and ORC occupancy at *CEN5* and the location of putative origin at the right arm of chromosome 5. Cse4 binds *CEN5* but not the location of putative origin (left panel). ORC also binds *CEN5*, but its occupancy at the location of putative origin is low.

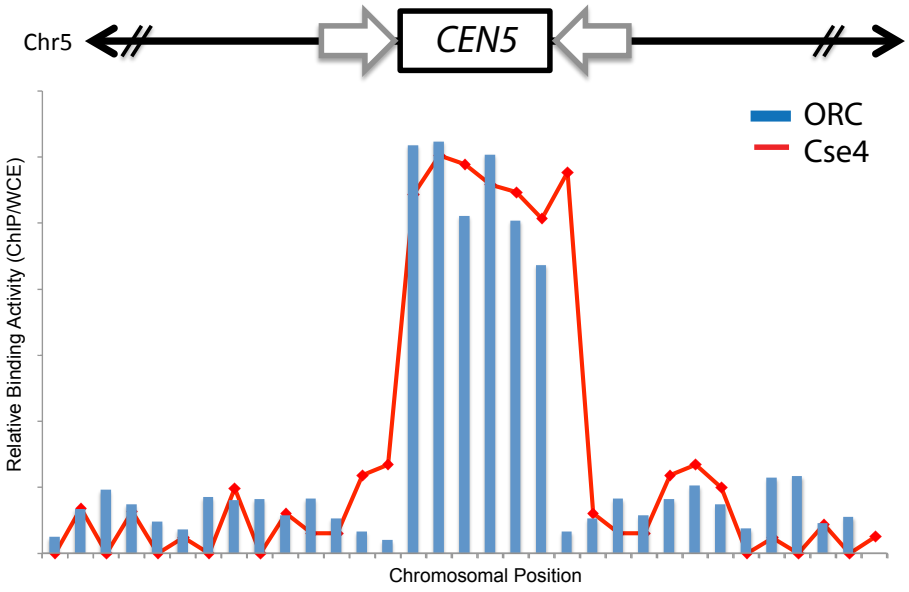


Figure 2.6. ORC binding is enriched and co-localized with Cse4 at *CEN5*.

Blue bars represent ORC bindings. ORC only binds *CEN5* core region but not the proximal regions, and co-localized with Cse4, which is represented by red line.

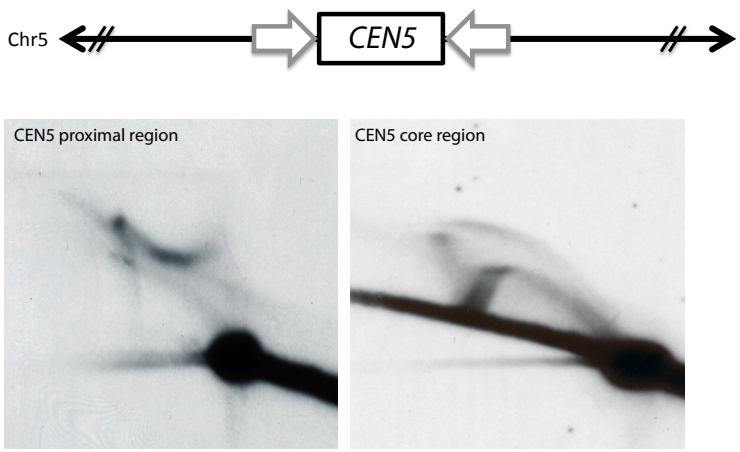


Figure 2.7. *CEN5* core contains origin activity

DNA 2D gels identify replication intermediates at *CEN5* core and its proximal region. Bubble arc intermediates were detected to show *CEN5* is an actively firing origin (right panel). No bubble arc intermediates were detected at *CEN5*-proximal region (left panel).

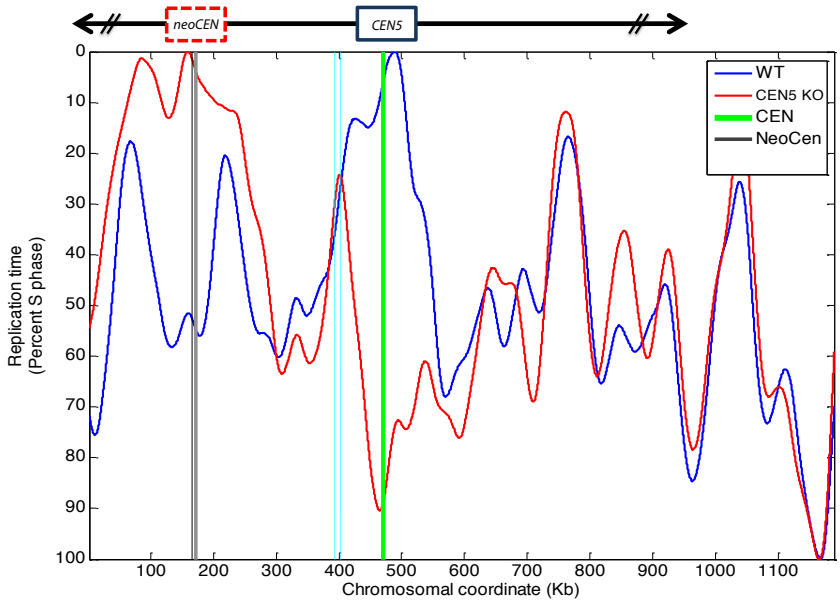


Figure 2.8. Neocentromere locus is the first place to replicate at a *CEN5*-deleted chromosome.

The native centromere (green line) is the earliest replicating region in wild-type cells (blue line). While wild-type *CEN5* is deleted, neocentromere locus at 170 kb (grey line) is the earliest replicating region at the chromosome 5.

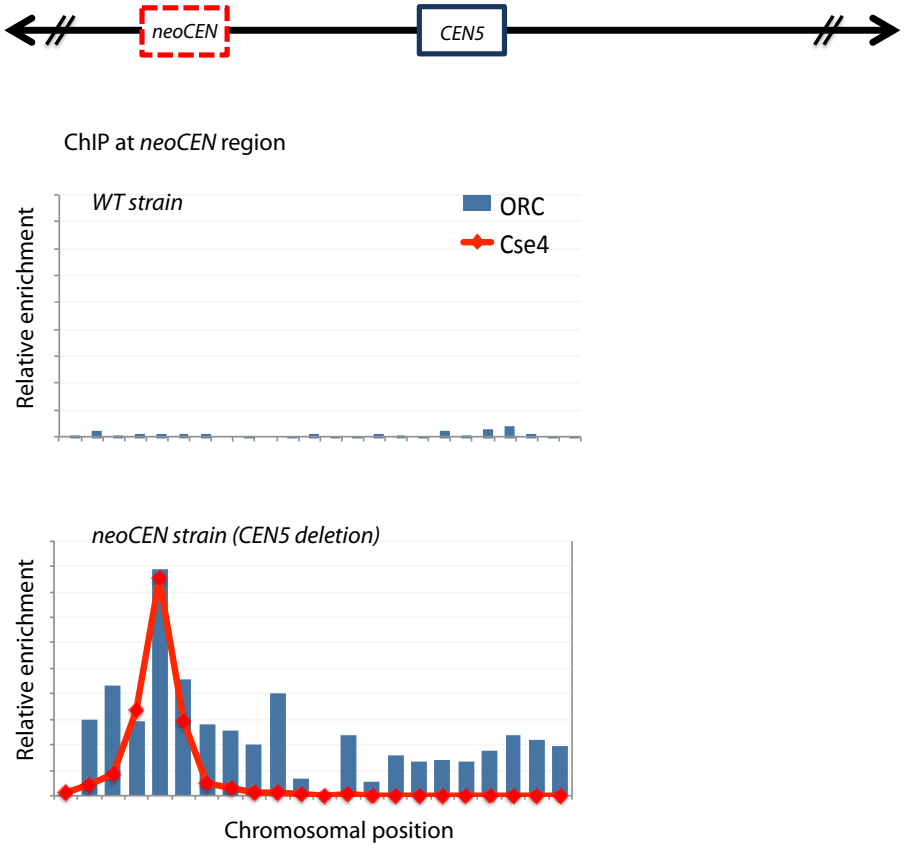


Figure 2.9. Functional neocentromere recruits ORC.

ChIP experiments were performed in a wild-type strain and a neoCEN strain. In wild-type strain (top), ORC and Cse4 bindings were not detected at 170 kb region. While native CEN5 is deleted in neoCEN strain, neocentromere forms at 170 kb region with Cse4 binding (red line) as well as ORC binding (blue bars).

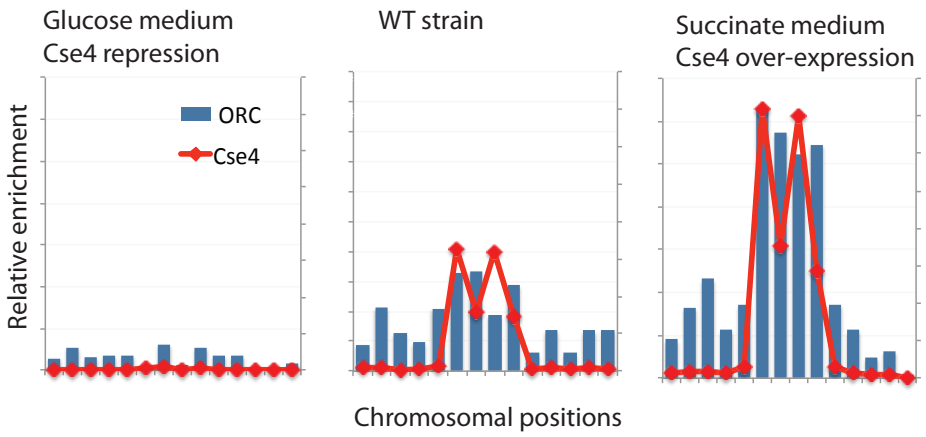


Figure 2.10. Cse4 recruits ORC at centromere core region.

A conditional PCK1 promoter was used to manipulate Cse4 expression level. In a wild-type strain (middle), ORC (blue bars) and Cse4 (red line) are co-localized at *CEN5*. Over-expression of Cse4 protein level resulted in increased levels of ORC binding at the *CEN5* core (right) while repression of Cse4 protein level resulted in reduced levels of ORC binding (left).

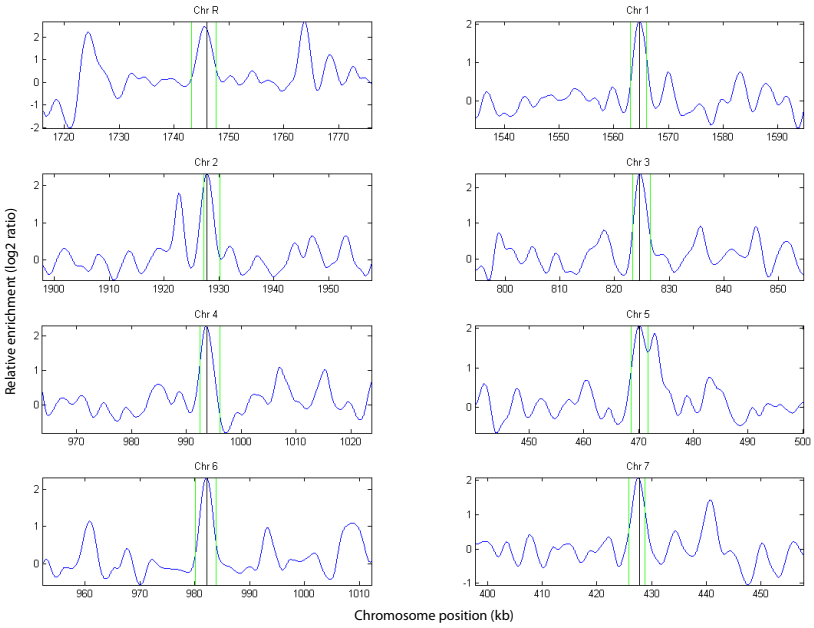


Figure 2.11. ORC binds all 8 centromeres

ORC ChIP-chip was performed. ORC strongly binds all 8 centromeres (blue line). Green lines represent the boarders of Cse4 bindings.

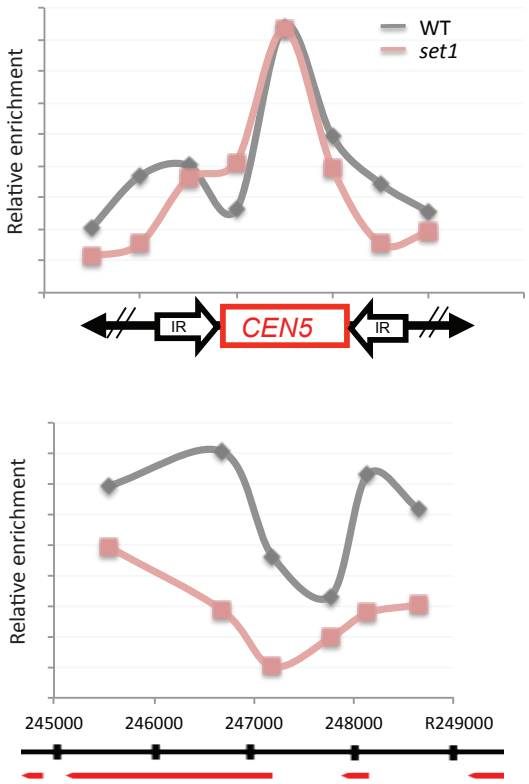


Figure 2.12. Centromeric and chromosomal ORC bindings are regulated differently by Set1.

ORC ChIP experiments were performed in a wild-type (grey line) and a *set1* (dark pink) strain. Centromeric ORC occupancy remains the same in both strains; in contrast, chromosomal ORC occupancy at an origin region is reduced in the *set1* strain.

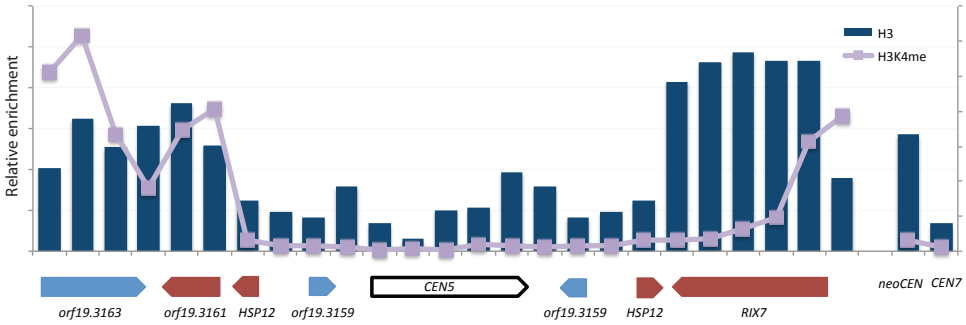


Figure 2.13. H3K4 methylations are absent at *CEN5* core region.

ChIP experiments using anti-H3 and H3K4me3 antibodies were performed, respectively. *CEN5* core region has low occupancy of both histone H3 and H3K4me3.

Table 2.1. Predicted origin locations from replication timing profile (Koren *et al.* 2010)

Chr R	Chr 1	Chr 2	Chr 3	Chr 4	Chr 5	Chr 6	Chr 7
10930	43020	29050	9380	56560	67110	103190	74960
62930	119700	172200	135360	170890	160360	189180	128510
174660	206030	318060	188420	239840	217710	377130	230890
336120	337310	408990	333070	337310	331220	479980	410540*
460470	423370	523760	429380	464530	426000	519500	535930
513800	539890	642960	554250	562500	489390*	617980	576400
586860	724140	793060	590080	621150	638550	681290	726950
662820	884060	964230	695150	776250	692730	715780	790220
786960	927960	1040080	807340*	844100	765230	853950	827650
837270	1116420	1145250	860200	983700	846400	900670	885090
949580	1163200	1234420	947910	1012810*	920790	964240*	920770
1012530	1295950	1325460	1047400	1119930	1038970		
1098270	1385760	1384740	1131610	1204530	1109710		
1167700	1559480*	1443970	1231700	1324190			
1227270	1694570	1620720	1352790	1457910			
1345670	1825610	1745260	1412900	1508120			
1475860	1882050	1914560*	1474870	1583900			
1537220	1964390	2064620	1565460				
1606220	2172090	2090840	1607210				
1733600	2292390	2155290	1717650				
1773400*	2412670						
1882500	2494490						
1919990	2587660						
2004880	2727600						
2049820	2848430						
2162540	3063020						
2241000	3111850						

* centromere-proximal origins

Table 2.2. List of primers in this study

Forward primers	Sequence	Reverse primers	Sequence	UPL probe	Purpose (target)
3902	gaaactccccaacgaatactga	3903	ttcgaaacgatagaaaattcgag	95	CEN5
3904	tgctatatctaatgcccccagt	3905	tgctgttgtatttgactccttctta	131	CEN5
3906	gtgaaccaacagaaactgatcct	3907	tcaaatcaatgttttcagtgaactt	145	CEN5
3908	gaggaggcaatcaacagcac	3909	tcttttgcgaagctccattt	136	CEN5
3910	caaactcgatccaaccttctg	3911	Aagcccgaagatccatcag	143	CEN5
3912	actttccctatctgatgttgcac	3913	Tttgtcaggtcatcatcatttctt	164	CEN5
3914	tcgagtgtaaaagtatgtggattgt	3915	Gaatggacctcatgcaagc	111	CEN5
3916	gtagcaccggtgacaactcc	3917	tgccaaaactgaaattggaa	146	CEN5
3918	gttctctaaatgcctccaagcta	3919	tggtccctagtgttgctatcg	8	CEN5
3920	cagaataacagcaagtcggtca	3921	tgtatggaatgtggattttgttg	132	CEN5
4055	ctagtgaggctgtaaacctcgacttttcaataccg	4057	ttcaacacatttgttacaacttcaaa	90	CEN5
3922	ccctctgttgctgttacttgag	3923	ttttgtataaagcaaggcattgaa	96	CEN5
3924	gtgagaagaaaagtaaatgacttcgat	3925	aaacacttgcaaccaatacagg	144	CEN5
3926	aggtgattgttgcatagtcatttc	3927	ggctataacttctaactggcattgt	106	CEN5
3928	tcacttgaattggaacagaattg	3929	aacaaatgtcaattaaacaacggtat	87	CEN5
3930	gcttctcaattatgtatggaatgtg	3931	cagcaagtcggtcaaaatca	132	CEN5
3933	ccagcatcatcaggctcttta	3934	tgacagagataggatgcgttatg	151	CEN5
3935	ctgcaccagagaagtttccac	3936	cccttacacgcaaatgttga	109	CEN5
3937	atccctctctcttggcactg	3938	caccttggctatcacgtatca	48	CEN5
3939	ctcttgagatctgcaccaaca	3940	ggtgagcatttcaactatagagagc	109	CEN5
3941	agggccgtacaaaagtactcc	3942	gactcccaattttacatccagaa	119	CEN5
3943	ggatcatgctcttcgtctgc	3944	tccatgatttacttcaggagca	108	CEN5
3945	tccttcccatctacagtcttatca	3946	ggaatgagctcagtagagtgtctg	140	CEN5
3993	catgactactaccccataggcttt	3994	tgaagaggttggtgggttttgt	9	CEN7
4678	aaattcttatatcctgcttccataaaa	4679	tgcttcaattcaaactattgctg	128	ORI410
4680	gcatttaaaactgttgatcttttcg	4681	cacagttatcccgcaacctc	52	ORI410
4682	aattgattaattccagtgggtttg	4683	tgactttagcaatgggggttt	145	ORI410
4684	ccaaaataaacaccgcagga	4685	ggtggattattgtgatcttattaggc	138	ORI410
4686	gtgttgctaataaatggtgaaggat	4687	caccatcaccaccaccagta	10	ORI410
4688	cagtgactgatcttgttttaattgg	4689	actgaagaaaatacggaagcaaa	150	ORI410
4690	tggaaaaacaacgatagtcaaaaa	4691	ttagaagagagggttgaatctcg	31	ORI410
3150	CATCCACCAAAGTTTCTCCC	3151	GCCTCATTGACCTTATTAGCC		CEN5L (2D gel)
3152	AAGTTGCTACAGGACAGCAG	3153	CAACGACTCACGATTTGCATG		CEN5R (2Dgel)
3181	ctagtaagcgcctttgaatg	3182	tagtaaattgagagctcagc		CEN5R (2Dgel)
4071	tgaaaccataaatatcgcaaaaga	4072	tgtattgtttaaagttatccagcgagt		
4073	aaagccagaagttcaacaatttatc	4074	ttctttctcttttggtgataattcaa		
4075	gctcgtacttcaactcaaggttc	4076	gctcctgaagcagcaagataa		
4077	ttaccttcggcttgtaatgttg	4078	agagccatcatctggtgcat		

Table 2.2. Continued

Forward primers	Sequence	Reverse primers	Sequence	UPL probe	Purpose (target)
4079	catctcaagtatcttctttgacaactg	4080	gtttggaaactggcaatggt		
4081	gcatcacaacataagcaagca	4082	ctccgcgtgctaaaagagtc		
4083	cgtccagctttttgtttcttc	4084	ttgaacttcggctataaatgaca a		
4085	accaggtttgtgagacatttga	4086	cttcaagaagaaaagaacatttt cag		
4087	gccatcaccaccacattttt	4088	tgtaacggattaattctgcgaata c		
4089	caacttggaggttgtttcttgg	4090	aattctacacttcgtcttcggttta		
4091	cggcatttgctggaaaaata	4092	gaagggccaagagtggagat		
4093	gttgacacagggattttgacg	4094	accaatattaaatttttgactgag ca		
4095	ctgtttcaagtgaatttgcttctagt	4096	gattcactggccgatgatg		
4097	gaaacatcagttagtgagtctgctg	4098	agattcactggcaaacagttca		
4099	gatgccacaactacacagtcagt	4100	gatccattgccagatccact		
4101	caggtgccaacaatggatct	4102	tgatggaacaaccgagagtg		
4103	cgtcggctatggttattcca	4104	cattgaagacaacacactcaag g		
4105	cccatcacgtctttatccattt	4106	agtatatcttgagcagatgaaca atca		
4107	ccgttgaaaaacgaagatgag	4108	tgcttaaatttgttgacgttgtc		
4109	tcttcaaactttcctttcaagaca	4110	tcgcaaaagaatatgaattaact acag		
4111	tttgaaagtgggagcatgaa	4112	gatggaaaaatgttgatggagtc		
4113	cccttctaacccggcatatc	4114	ttagttaactgttccgccagatt		

Chapter 3

A new autonomously replicating plasmid in *Candida albicans* as a potential genetic tool

Introduction

Multiple yeast species, as single-celled eukaryotes, have been effectively used as model organisms to study conserved cellular processes. A critical advantage of using yeast model systems for molecular studies is that the genomes are more amenable to manipulation than in higher eukaryotes. A very important genetic tool in yeast is the plasmid shuttle vector that is autonomously replicating and well maintained in the yeast cell (Parent et al., 1985) and can also be 'shuttled' to *E. coli* for facile cloning manipulations. For instance, the development of plasmid-based yeast artificial chromosomes in *S. cerevisiae* (Dani and Zakian, 1983; Murray and Szostak, 1983) and *S. pombe* provided significant insights into chromosome behavior, revealing genetic elements required for replication and segregation (Steiner and Clarke, 1994). The yeast two-hybrid system uses plasmid-based protein expression and can reveal protein-protein interaction *in vivo*. However, conventional plasmid shuttle vectors do not work in higher eukaryotes such as mammals.

In the budding yeast *S. cerevisiae*, circular plasmid vectors are widely used for genetic complementation and protein expression (Parent et al., 1985). An autonomously replicating sequence (ARS) is an essential component to maintain plasmid vectors in cells (Struhl et al., 1979). ARS consensus sequences were originally identified from plasmid-based ARS screens (Chan and Tye, 1980; Stinchcomb et al., 1979; Struhl et al., 1979). Additionally, centromere sequences can be added to plasmids to increase segregation accuracy. The hard-wired

nature of *S. cerevisiae* origins and centromeres allows a plasmid vector to function autonomously if it contains the right primary DNA sequences. When a plasmid only carries an ARS, it can replicate autonomously but cannot segregate faithfully into daughter cells. By adding centromere sequences onto the plasmid, it becomes stable, and well-maintained at low copy number during cell division (Clarke and Carbon, 1980). Additionally, linear centromere plasmids constructed by adding telomere sequences at the termini of linearized DNA were stably maintained in a manner that resembled natural chromosome maintenance (Dani and Zakian, 1983; Murray and Szostak, 1983; Polizzi et al., 1989). In *S. pombe*, circular ARS plasmids can yield high transformation efficiency with high copy number and are widely used as well (Heyer et al., 1986).

Steiner *et al.* constructed a linear plasmid carrying centromere sequences to study centromere establishment, and the results suggest that the higher order of chromatin structure is required for the establishment of centromere function on a linear plasmid backbone (Steiner and Clarke, 1994). Thus, by mimicking chromosome structure, CEN plasmids are a useful system for studying chromosome properties in *S. pombe*, but are not suitable as a genetic tool.

In *C. albicans*, there have been three autonomously replicating sequences (ARS) identified in previous studies (Cannon et al., 1990; Herreros et al., 1992). However, none of them has been successfully utilized to develop a plasmid shuttle vector. The first two studies mimicked plasmid-based ARS screens in *S. cerevisiae* to identify origin sequences from the *C. albicans* genome. Isolated free plasmids with ARS function carried recombinant chimeric sequences, which

were presumably a product of gross rearrangement between plasmid and genomic sequences (Herreros et al., 1992). Thus, conventional *C. albicans* plasmids from previous studies may frequently integrate into chromosomes and do not act as ARS plasmids in yeast cells. In a third study, Beckerman *et al.* made an interesting observation that a plasmid containing the *IMH3* gene replicates autonomously in *C. albicans* cells (Beckerman et al., 2001). A plasmid containing the *IMH*3 gene, which confers resistance to mycophenolic acid (MPA), has been shown to yield high transformation efficiency and high copy number when MPA concentrations were increased in selectable medium (Du et al., 2004). This was the first well-defined genomic locus conferring ARS function. Moreover, rDNA containing origin activity was isolated naturally as a large circular plasmid in *C. albicans* and also can function as an ARS plasmid. However, consistent with the early studies, these ARS containing plasmids are unstable, with a high frequency of chromosomal integration and a mix of integrated and episomal forms can both exist in *C. albicans* cells.

The basic genetic elements on a stably inherited plasmid shuttle vector are an ARS, a centromere, and a selectable marker. Thus, we aimed to investigate these major components in the *C. albicans* genome. We have performed genome-wide mapping of ORC binding sites and revealed a large number of potential origins. Second, although centromere sequences differ on each chromosome in *C. albicans*, the epigenetic features of centromeres can recruit origin activity at centromeric core regions (Koren et al., 2010b). Thus, the *C. albicans* centromere can serve not only as a centromere but also as an origin.

Finally, prototrophic markers such as *URA3* and *TRP1* are commonly used in auxotrophic yeasts. Previously, our lab constructed a centromere plasmid with a *URA3* marker, and it yielded high transformation efficiency. However, an experimental control plasmid with the *URA3* marker, but without centromere sequences, also yielded yeast transformants. Interestingly, these *C. albicans* transformants contained both integrated and free plasmids without previously characterized ARS elements present, suggesting that the *URA3* gene itself may be acting as weak origin of replication and also as a homologous sequence that targets plasmid recombination with the genome (M. Gerami-Nejad and K. Bouchonville, personal communication).

In this chapter, I expanded my work from chapter 2 and developed an autonomously replicating plasmid system in *C. albicans*. Previously, I have identified centromeric origins and mapped ORC binding sites throughout the entire genome. Here I investigated the origin/ORC binding sequences and further tested them on a plasmid. Two potential origin sequences, a centromere origin and an origin identified by ORC binding, were studied in this chapter. I also revisited the previously identified *IMH3* origin to compare behavior of this origin with the origins I identified. The development of non-integrating plasmids is a significant advancement in the genetic tools available for use in *C. albicans*.

Materials and Method

Yeast strains and growth condition

Yeast strains used in this chapter were all based on *Candida albicans* RM1000 #2 (YJB7617), from Pete Magee lab. Strains were stored as stocks in 50% glycerol at -80 °C and grown in YPAD medium (1% yeast extract, 1% peptone, 2% glucose, 1.5% agar) at 30 °C. The selection medium contained 400 µg of nourseothricin (clonNAT). Yeast transformations were based on the Lithium acetate transformation protocol (Wilson et al. 1999).

Plasmid construction

The linear plasmid backbone was from pMG2218. A telomere cassette was subcloned from ApaI and NotI digested fragments from pMG2193 into the SacII and NotI sites of pMG2379. The construct is p2380, named pLN0 (Figure 3.14, top). A linker for restriction enzyme digestion was made by adding restriction enzyme digestion sites at both end of KanMX. KanMX with additional linkers were amplified (forward primer (5357): CTAGACTAGTGCGGCCGCggccggccaggcccggatccccgggttaattaa; reverse primer (5358): CTAGACTAGTGCGGCCGCggccggccaggccttaattaaGaattcgagctcgtttaaac and then the KanMX on pLN0 was replaced by the KanMX with linkers. The plasmid was named pLN (p2382) (Figure 3.14, middle).

ORI410 DNA was subcloned from the genome by primer 4921: TCCCCGCGGTTGATGATTGGATCGGGTTC and 4922: TCCCCGCGGGGAACATCTGAAATTGGTTC, and then inserted into the SacII site on pGEM-NAT (pMG2339). The resulting plasmid was named p2340. Later the linear cassette on pLN was subcloned by XhoI digestion into XhoI site on p2340. The pLN-ORI410 was initially named p2402.

To construct a linear *IMH3* plasmid, the *IMH3* gene was subcloned from the EcoRI site on pMG2081 into the PvuII site on pLN. The plasmid, pLN-*IMH3* was initially named p2435.

An intergenic region was cloned from the genome by primer 5721: caagagggcgtcGGATTTAGTTCCATTATGG and primer 5722: ctcttggacgccACACGACCTTGGAGTGATC and inserted into BsaH1 (AcyI) sites on pLN. The plasmid was named pLN-nonORC in this chapter and documented as p2408.

CHEF gel analysis

Continuous homogeneous electrophoretic field (CHEF) samples were prepared in agarose plugs as described previously (Selmecki et al. 2008). Cells were washed in 50 mM EDTA and lysed with zymolyase in 1% low melt agarose prepared in a plug mold. Plugs were incubated in LET buffer (0.45 M EDTA pH 8.0, 0.01 M Tris pH 7.5) with 1 % beta-mercaptoethanol in 37°C overnight during zymolyase digestion. Plugs were washed twice with 50 mM EDTA, pH 8.0 and digested by 0.2 mg/L proteinase K in NDS buffer (0.01 M Tris pH 7.5, 0.45 M

EDTA pH 8.0, 1% N-lauroyl sarcosine) at 50°C for at least 48 hours. After washing with 50 mM EDTA, pH 8.0 twice, 30 minutes each time, plugs were used immediately or stored in 50 mM EDTA, pH 8.0 at 4°C.

Plugs were run on a CHEF DRIII electrophoresis system (Bio-Rad) on a 1% Megabase® agarose gel in 0.5x TBE under the following conditions: 11°C, 60-120 second switch for 14 or 18 hours at 6 V/cm with 120° angle to visualize DNA fragments. Autonomous linear plasmids are readily separated from whole chromosomes in these gels.

Southern blot

DNA was transferred from agarose gels to Magnacharge nylon membranes by the following steps. After agarose gels were stained with EtBr, pictures under UV were taken to record the position of each size of chromosomes. Then the gels were embedded in 0.25 M HCl for 15 minutes. After rinsing with water for 1 minute, the gels were gently shaked in 1.5 M HCl, 0.5 M NaOH for 15 minutes, twice. To neutralize the gels, the gels were then gently shaked in 1M Ammonium Acetate, 0.02 M NaOH for 15 minutes, twice. Then the gels and membranes were rinced in 20x SSC for 10 minutes for the transfer reaction. After at least 20 hours transfer reaction, the membranes were cross-linked under UV for the following porbing steps. Membranes were probed overnight at 42°C and detected with anti-Digoxigenin-AP- conjugated antibodies. Probes were prepared by PCR amplification of *NAT* gene sequences using DIG-labeling kit (Roche).

Plasmid copy number

Plasmid copy number was estimated by quantitative PCR. Genomic DNA from yeast transformants was extracted, and for each reaction 50 ng of gDNA was analyzed in a Roche LightCycler 480 Real-Time PCR machine. Plasmid specific primers recognizing the *NAT* gene and genome-specific primers recognizing *CEN7* loci were chosen to measure the plasmid DNA amount (NAT gene forward primer (5807): ttctattttaatcaaatgttagcgtga; reverse primer (5808): tgatattactttctgcgcacttaactt. CEN7 forward primer (3993): catgactactaccccataggcttt; reverse primer (3994): catgactactaccccataggcttt. Ratio from *NAT* gene amplicon was normalized by ratio to the *CEN7* locus amplicon to determine the relative copy number of plasmids.

Plasmid loss assay

Yeast transformants were grown in non- selective medium YPAD for two days, and then were first diluted 10 fold. Cells were diluted in 1:10 ratio for 4 or 8 serial dilutions and for each diluted sample, 5 µl was spotted onto YPAD (non-selective) and YPAD+NAT (selective) plates. The plates were incubated for 2 days or 3 days to compare the growth properties of the colonies.

Results

Circular plasmids integrate at high rates and thus are not practical ARS plasmids in C. albicans

The centromere has been identified as the most efficient origin in the genome. *CEN4* was cloned onto a circular plasmid to examine the ARS function. A large fraction of *C. albicans* circular plasmids rapidly integrated into the genome, although a subset of supercoiled plasmids still remained free (Kelly Bouchonville, Figure 3.1). Surprisingly, the pGEM-*URA3* shuttle vector itself yielded high transformation efficiency without the *CEN4* DNA. Similar to the *CEN4* plasmids, some of the transformants contained free plasmids while others integrated. We reasoned that the *URA3* marker might cause high integration because the *URA3* gene on a circular plasmid has homology to the genome and might direct recombination. It also appeared that *URA3* has a weak ARS function, producing large numbers of small and highly unstable colonies (Kelly Bouchonville, personal communication).

Since we were interested in characterizing sequences with ARS function, the weak origin activity of *URA3* interfered with these assays. Therefore, we switched to using a heterologous selectable marker gene *NAT*, from *Streptomyces noursei*, which confers resistance to the antibiotic nourseothricin.. Instead of using centromeric DNA, which recruits ORC epigenetically, I cloned ORC binding sequences, which may directly attract ORC, to confer ARS activity on the plasmid. I introduced pNAT plasmids with and without a 1.2 kb ORC binding sequences as well as pGEM-*URA3* into *C. albicans* cells. Unlike the *URA3* marker plasmid, the pNAT empty plasmid did not yield any transformants supporting the idea that *URA3* does have a putative weak ARS function (Figure

3.2). Importantly, a pNAT plasmid containing a 1.2 kb ORC binding sequence yielded higher transformation efficiency than the empty vector (Figure 3.2).

The rapid integration of circular plasmids into the *C. albicans* genome complicates analysis of plasmid activity, such that we cannot determine ARS function by yeast transformation efficiency alone. Previously, Kelly Bouchonville found that transformant colony shape might be correlated with the existence of free plasmids in *C. albicans* cells. Consistently, most pGEM-*URA3* transformants have 'nibbled' colony shape indicating that the transformants contains free plasmids that are likely lost at a very high rate (Figure 3.2 bottom panel, right). Although the pNAT-ORC plasmids yielded higher transformation efficiency, most of the transformants have a round colony shape suggesting the plasmids have integrated into the genome during transformation (Figure 3.2, bottom panel, left). I further performed plasmid rescue experiments to extract free plasmids from *C. albicans* cells followed by bacterial transformation. Consistent with the colony shape indicating plasmid integration into the genome, no free pNAT-ORC plasmids were rescued. ORC binding sequences may cause high frequencies of integration due to the sequence homology even though it may have ARS activity.

Linear plasmids with telomere sequences at both termini

Circular plasmids containing either ORC binding sequences or centromere sequences integrate frequently and thus do not autonomously replicate as stable episomal forms. To develop a stably maintained plasmid system, I designed a

linear plasmid vector. In *Histoplasma capsulatum*, circular plasmids cannot autonomously replicate, similar to our findings in *C. albicans*. Interesting, Woods and Goldman constructed a linear plasmid with telomere sequences at both termini (Woods and Goldman, 1992; 1993). The addition of telomere sequences is sufficient to support ARS plasmid maintenance in *Histoplasmia capsulatum*. Thus, I inserted inverted telomere repeats, separated by a kanamycin-resistant gene (*KAN*^{R}), into the pNAT plasmid (Figure 3.3). During plasmid construction, I also added a linker with multiple cloning sites for restriction enzyme digestion between the telomere repeats and the *KAN*^{R} gene. Thus, the plasmid carrying telomeres, sequences for linearization, and the *NAT* gene (pLN) can be linearized to transform into *C. albicans*.

Several DNA fragments were analyzed for ARS function using the pLN linear plasmid (Figure 3.4). Since a functional centromere can recruit origin function, I cloned the entire *CEN4* region with flanking inverted repeats into the pLN plasmid to mimic the structure of an artificial chromosome. Additionally, I mapped ORC binding sites throughout the entire genome by ORC ChIP-chip. Each ORC binding site could be a potential origin. Thus, I cloned a 1.2 kb ORC binding sequence onto the linear plasmid vector. Since the ORC binding site, named *ORI410*, is an actively firing origin in the chromosomal context (see Figure 4.5 in Chapter 4), I can test if the ORC binding sequences can confer ARS activity. Additionally, *IMH3* is a gene that conferred ARS function in previous studies (Beckerman et al., 2001). Although the *IMH3* work was done with a *URA3*-containing plasmid, yeast transformation efficiency of *IMH3*

plasmids increased by addition of MPA selection, suggesting that the *IMH3* gene has a putative function in plasmid maintenance (Du et al., 2004). Additionally, I chose an intergenic non-ORC binding site as a negative control sequence. This 1 kb DNA fragment was from the intergenic region between *orf19.1961* and *orf19.1963*. Since it is not a predicted ARS, the transformants from this control plasmid will determine whether a random DNA sequence can function as an ARS on a plasmid vector and will estimate the frequency of chromosomal integration in each experiment.

Linear plasmids carrying replicators can yield yeast transformants

I have performed several yeast transformation experiments and obtained similar transformation efficiency results in each replicate (Figure 3.5). pLN empty plasmids and pLN-nonORC plasmids always yield 0 or a small number of transfomants, which might be from random chromosomal integrations. pLN plasmids containing *ORI410* sequences yield relatively high transformation efficiency. Also, we did obtain transformants from pLN plasmids carrying the *IMH3* gene, but the efficiency was lower than other ARS containing plasmids. The low efficiency might be due to long homologous sequences causing chromosomal integration and thus only few transfomants with harmless integrations or free plasmids can survive. Transformation of pLN-*CEN4* also resulted in low transformation efficiency. Interestingly, the colony phenotypes (nibbled/round edge) observed following transformation of the circular plasmid were not apparent in the linear plasmid transformants suggesting that the linear

plasmids are maintained better during cell division. Overall, as expected, the ARS containing fragment *ORI410* can confer high transformation efficiency on a linear plasmid. pLN with *CEN4* mimicking an artificial chromosome had lower transformation efficiency, but the number of transformants was higher than the background level (pLN empty and pLN-nonORC).

From our previous work on pGEM-*URA3*, we learned that the transformation efficiency alone may not be a reliable determinant of plasmid behavior in terms of autonomous replication (Kelly Bouchonville, personal communication). Thus, additional approaches must be used to ask if the linear plasmids replicate autonomously. For example, autonomously replicating circular shuttle plasmids can be rescued from yeast genomic DNA and detected by their ability to transform bacteria. However, the linearized plasmids cannot be maintained in *E. coli* because the ends are not recognized. To determine if the *C. albicans* transformants contain autonomously replicating linear plasmids, I performed CHEF gel analysis to separate chromosomes (large size) and linear plasmids (small size) and then analyzed them on Southern blot using a plasmid-specific probe. Thus, I could examine if the linear plasmids act as an autonomously replicating plasmid with or without chromosomal integration.

pLN-CEN4 plasmid mimics an artificial chromosome

First, I tested whether the *pLN-CEN4* plasmid replicated autonomously as an extrachromosomal element. In either *S. cerevisiae* or *S. pombe*, artificial

chromosomes are well developed and characterized. They are stably maintained and can replicate with natural chromosomes during the cell cycle. *In C. albicans*, I cloned the entire regional centromere 4 onto pLN plasmids and hypothesized that *CEN4* DNA can perform both centromere and origin function. I randomly picked nine transformants from a single transformation experiment to perform CHEF-Southern blotting analysis. Five of nine transformants have a band distinguishable from the chromosomes in the EtBr staining (Figure 3.6A, top panel). Furthermore, all five transformants have autonomously replicating linear pLN-*CEN4* plasmids without any chromosomal integration (Figure 3.6B, right panel). This is striking because although the transformation efficiency of pLN-*CEN4* is lower than ARS containing linear plasmids, half of the transformants have a well-maintained artificial chromosome.

To further characterize the features of the pLN-*CEN4* plasmid, I analyzed another two independent transformants containing pLN-*CEN4* plasmids and confirmed the presence of non-chromosomal bands. However, unexpectedly, the size of pLN-*CEN4* in these cells was much larger than the initial size of pLN-*CEN4*. The size of pLN-*CEN4* in cells is consistently around 700 kb in all transformants, and thus, ruling out the idea that they arose through random recombination of linear plasmids. Another possibility was that if pLN-*CEN4* plasmids recircularized in cells, the supercoiled form of plasmids might show slow mobility and become insensitive to exonuclease. Thus, I performed Bal-31 exonuclease assays to ask if these molecules were linear. Indeed, the large bands were sensitive to digestion by Bal-31, indicating that they are linear (data

not shown). Also, the pLN-*CEN4* plasmids in the transfomants were stably maintained in selective medium and lost in non-selective medium (data not shown).

To understand the source of these very large DNA sequences carrying the pLN-*CEN4* sequences, I performed a CGH-microarray on the isolated band to identify all the genomic sequences on this band. This experiment tests whether the telomere sequences of pLN-*CEN4* plasmids lengthened *in vivo* (in which case only plasmid sequences and telomere sequences should hybridize to the band) or whether pLN-*CEN4* integrated into a specific chromosome and yielded large-scale chromosomal truncations. I extracted the pLN-*CEN4* DNA band specifically from the CHEF gel and hybridized the extracted DNA to a whole-genome CGH-microarray using the parental strain DNA as the reference control. We expected that the plasmid DNA should appear as a genomic region with increased copy number. Notably, the copy number of the right arm of chromosome 4 was strongly increased with the breakpoint immediately adjacent to the *CEN4* region (Figure 3.7). These results suggest that pLN-*CEN4* linear plasmids integrated into the homologous *CEN4* chromosomal locus during yeast transformation and break-induced replication generated additional copies of the right arm of chromosome 4 with the pLN-*CEN4* vector backbone. Thus the truncated chromosome 4 can function normally and is well maintained in the selective medium. Since *C. albicans* tolerates aneuploidy quite well (Selmecki et al., 2010), pLN-*CEN4* plasmid might preferentially integrate into its homologous site and replicate with chromosomes via a recombination pathway. Thus, a

centromere, containing the most efficient origin in the genome, might function as an ARS on the linear plasmid, but analysis of the resulting plasmids is complicated by integration of the vector forming an isochromosome.

ORI410 sequence can confer ARS function

Because centromere ARSs were not amenable to analysis on plasmids due the the integration issue described above, and because CENs span 3-5 kb of DNA, we sought to test another, smaller putative ARS region. Accordingly, we tested linear plasmids containing sequences that direct ORC binding. When the 1.2 kb ORC binding sequence (*ORI410)* was tested in circular pNAT plasmids, I did not detect any autonomously replicating plasmids in the transformants. However, because *ORI410* is an actively firing origin in the chromosomal context (see Chapter 4, Figure 4.5). I examined the same *ORI410* sequences on the linear plasmid vector. Strikingly, all 17 randomly selected transformants analyzed by CHEF-Southern blotting analyses, contained free, autonomously replicating plasmids (Figure 3.8A). Remarkably, no chromosomal integration was detected. Importantly, the molecular weight of pLN-*ORI410* plasmid *in vivo* (isolated from cells following transformation, was similar to the size of the original plasmid and was consistent in different transformants. I further selected another two transformants from different yeast transformation experiments for CHEF-Southern blotting analyses. Consistent with earlier results, no chromosomal integration was detected (Figure 3.8B). In additional transformation experiments, chromosomal integrations were sometimes detected in a few transformants (data

not shown). Overall, the linear pLN-*ORI410* conferred ARS function and was well-maintained as an autonomous plasmid in *C. albicans* cells. Interestingly, the existence of an episomal form of linear pLN-*ORI410* plasmids was not correlated with either colony shape phenotype or the colony size.

pLN-ORI410 has features of an ARS plasmid

In *S. cerevisiae*, either circular or linear ARS plasmids can reach high copy numbers in the cells. Thus, I used qPCR to amplify a plasmid specific region (*AMP*^{R} gene) and a genomic locus (*CEN7*) to compare the copy number of pLN-*ORI410* plasmid relative to the copy number of a single genomic locus. The copy number of plasmids was measured based on the relative amount of *CEN7* DNA (2 copies in the genome) and *AMP*^{R} gene. Two transformants tested had 8-12 copies of plasmid DNA in yeast cells (Figure 3.9).

In addition, as a genetic tool, ARS plasmids can be maintained (albeit not perfectly) in selective media, but are lost at high frequency when grown in non-selective medium. This is an alternative approach to test if the pLN plasmids integrate into the chromosomes. Thus, I performed a plasmid loss assay to ask if pLN plasmids are unstable Yeast transformants bearing pLN-*ORI410* plasmids were grown in non-selective medium for two days and then spotted on both selective and non-selective medium plates. Thus, I randomly picked 36 transformants containing pLN-*ORI410* plasmids. After two days of growth in non-selective media, most cells failed to grow on NAT plates (Figure 3.10, bottom panel), indicating that pLN-*ORI410* was lost. Interestingly, the subset of

transformants with chromosomal integrations as well as free linear plasmids shown in CHEF-Southern blotting analyses (data not shown), also lost the ability to confer resistance to nourseothricin when grown for two days in non-selective media. It is tempting to speculate that, in these transformants, chromosomal integration occurred only in a small sub-population, and those transformants may have growth defects due to chromosomal truncations.

Taken together, I have shown that the ORC binding sequence, *ORI410* on the linear plasmid vector can function as an ARS plasmid in *C. albicans cells*, similar to the conventional circular ARS plasmid in *S. cerevisiae*.

IMH3 gene confers weak ARS activity

In previous studies, the *IMH3* gene, which confers resistance to mycophenic acid (MPA), has been shown to have ARS function on a plasmid (Beckerman et al., 2001). Moreover, the copy number of the *IMH3* containing circular plasmids increased significantly when MPA concentration increased (Du et al., 2004). However, the pLN-*IMH3* linear plasmid yielded only a few yeast transformants, and only one transformant from four tested contained autonomously replicating pLN-*IMH3* (Figure 3.11). The low transformation efficiency may result from the long homologous sequences causing higher integration frequencies. I performed CHEF-Southern blotting to analyze the transformant and to compare it topLN-*ORI410*. Interestingly, the band intensity of pLN-*IMH3* was much weaker than pLN-*ORI410* (Figure 3.11), suggesting that the copy number of pLN-*IMH3* is much lower that the copy number of pLN-*ORI410*

plasmid. Thus, either the pLN-*IMH3* plasmid replicated poorly or its stability was low. Together, these results suggest that the *IMH3* gene may be a weak ARS.

In addition, I revisited the idea that with MPA selection, the *IMH3* plasmid copy number would increase. One possibility is that the *IMH3* gene, instead of being an ARS element on a plasmid, can serve as a selectable marker to promote the stability of linear pLN-*IMH3* plasmids. However, I was unable to use MPA to successfully select cells bearing pLN-*IMH3* plasmids in either circular or linear forms (data not shown). For cells under nourseothricin selection, the addition of MPA did slightly increase the copy number of linear pLN-*IMH3* plasmids (Figure 3.12). However, the result was not as significant as reported in previous studies when *IMH3* was on an *URA3* plasmid backbone.

Discussion

It has been over 30 years since scientists started using plasmid shuttle vectors in yeasts. In this chapter, I made a significant breakthrough for *C. albicans* biology by establishing a functional, stable ARS plasmid. In the past few years, several vectors have been designed for use in *C. albicans* for gene expression, epitope-tagging, and gene knockouts. However, previously designed vectors are all designed for specific chromosomal integration sites. For example, our laboratory has recently developed a vector to integrate sequences via homologous recombination into the genome between *orf19.1961* and *orf19.1963*, which is a large intergenic region that was shown to be a 'neutral' integration site

because no phenotypes were observed when the intergenic region was disrupted. Despite these significant advances with integrating vectors, an autonomous plasmid system is a more convenient option for many genetic manipulations, such as complementation of mutations.

ORC binding sites originally identified by genome mapping (Chapter 4) have an average ~1 kb in size. The 1.2 kb *ORI410* region in the *C. albicans* genome was determined to fire actively during the cell cycle and here, we also showed that it functions as an ARS on the linear plasmid vector. However, we saw that the 1.2 kb *ORI410* plasmid still has a certain level of chromosomal integration, likely because of the it's 1.2 kb homology with genomic DNA. As described in the next chapter, I performed two different screens to identify the minimal sequence requirement for ARS function. A 97 bp DNA fragment from the *ORI410* region conferring ARS function was isolated. Importantly, linear plasmids carrying the 97 bp sub-*ORI410* fragment have features of an ARS plasmid, such as plasmid loss under non-selective conditions (Figure 3.13.). Therefore, this smaller, optimized plasmid may minimize chromosomal integration, while maintaining utility as a plasmid vector in *C. albicans.* I will address the properties of these plasmids in more detail in Chapter 4.

The pLN-*CEN4* linear plasmid mimicking an artificial chromosome generated a large-scale chromosomal rearrangement. Although I failed to make an ideal *CEN* plasmid, while investigating the features of the pLN-*CEN4* plasmid,

I found that several critical centromere determinants, Cse4, ORC, and kinetochore protein, Mtw1 can be recruited onto the linear plasmid by plasmid-ChIP experiments (data not shown). *In vivo*, the pLN-*CEN4* plasmid integrated to form a truncated linear chromosome composed of the pLN plasmid backbone and the right arm of the chromosome 4. This truncated chromosome appears to function and segregate as a natural chromosome *in vivo*. However, like other integrating plasmids, it does not have the useful features that make autonomous plasmids desirable genetic tools. To improve the work on development of an artificial chromosome, more experiments on centromere sequence features by high-throughput methods have been intensively performed in the laboratory.

In *C. albicans*, circular plasmids integrate into the genome with high efficiency. This may be because no “strong” ARSs have been identified, or because homologous recombination between the genome and plasmid occurs frequently. Interestingly, the prototrophic markers *URA3* support replication of non-integrated circular plasmids, although the plasmid stability is low (Kelly bouchonville, personal communication). Linear pGEM-*URA3* plasmids also yielded high transformation efficiency in selective medium. A working hypothesis is that the high transformation efficiency and maintenance of some episomal pGEM-*URA3* plasmids may be due to the ARS activity of *URA3.* Consistent with this, the genomic *URA3* locus includes an ORC binding site. In the chromosomal context, transcription and replication are mutually exclusive to avoid replication fork collapses. However, origin activity can exist upstream of an ORF when fork

progression and transcription progress in the same direction (Shor et al., 2009). Evidence that origins exist at promoter regions due to the local chromatin structure has been reported in *S. cerevisiae* and higher eukaryotes (reviewed in (Mechali, 2010)). Consistent with this, pLN-*IMH3* linear plasmids increased copy number when MPA concentration increased. To increase production of Imh3 protein and increase resistance to MPA, either the *IMH3* gene had to be transcribed more efficiently per locus or the copy number of the plasmid had to be increased by replication. My observation that copy number slightly increased suggests that increased copy number of *IMH3* gene may increase the total Imh3 protein level for the resistance to MPA. Thus, when *C. albicans* plasmids use a transcribed unit as an ARS element, either expression of that transcription unit and/or plasmid behavior may be influenced.

In conclusion, I demonstrated that different potential ARS elements function on a newly established linear plasmid system. A linear plasmid containing ORC binding sequences in *C. albicans* can be replicated episomally and act much like *S. cerevisiae* ARS sequences. Thus, I successfully developed a genetic tool for the *Candida* community. Meanwhile, I am working to improve the pLN plasmid vector to have a smaller replication sequence (97 bp mini-ARS) and I aim to develop a non-integrating CEN linear plasmid. Thus, my studies will improve the availability of genetic methods in *C. albicans* and increase the utility of *C. albicans* as a model organism to study genome organization.

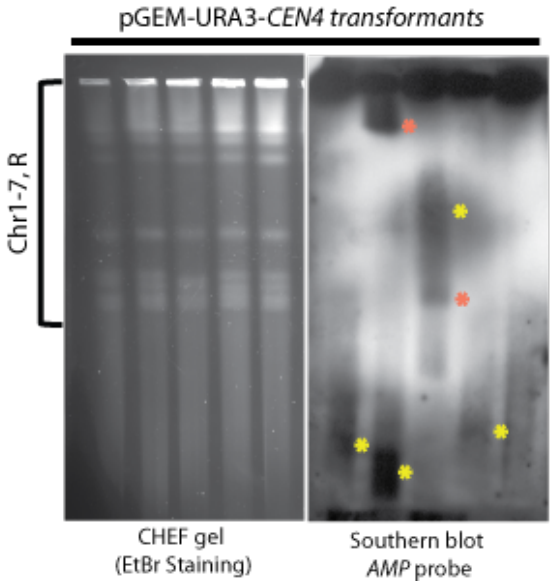


Figure 3.1. CHEF-Southern blotting analyses of circular pGEM-*URA3*-*CEN4* transformants (modified from Kelly Bouchonville, doctoral thesis, 2011)

Five isolates were analyzed in CHEF-Southern blot. The circular pGEM-*URA3*-*CEN4* plasmid in *C. albicans* cells maintained as both supercoiled forms (yellow stars) and chromosomal integrations (red stars).

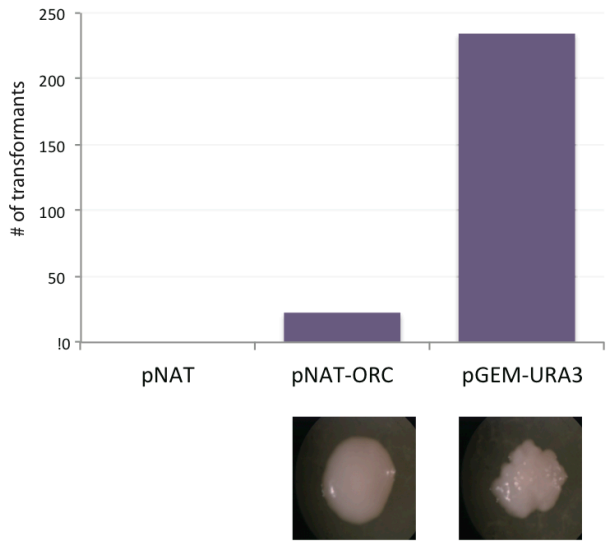


Figure 3.2. Circular pNAT-ORC plasmids yield higher transformation efficiency than the empty pNAT plasmids.

The circular pNAT plasmids containing a 1.2 kb ORC binding sequence yielded 20 transformants while no transformants from the empty pNAT vectors. pGEM-*URA3* plasmids can yield high transformation efficiency even though no replicator was contained. Most of the pNAT-ORC transformants have round colony shape; pGEM-*URA3* transformants have nibbled colony shape.

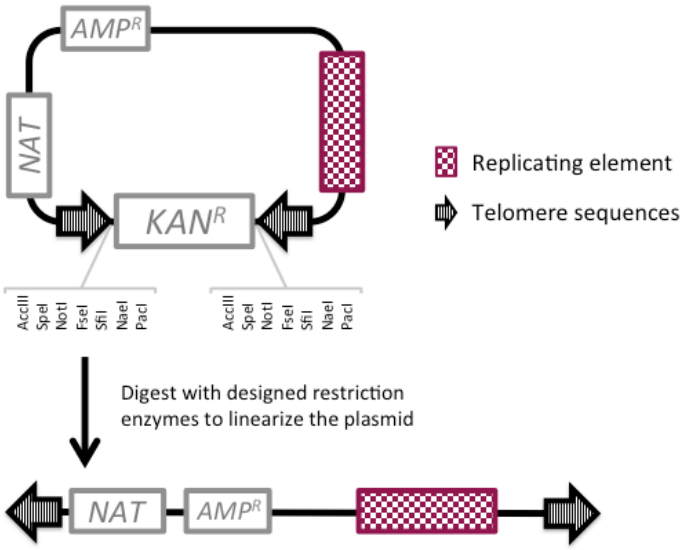


Figure 3.3. Linearization of pLN plasmids.

pLN plasmids include a selectable marker that confers nourseothricin resistance (*NAT)* in *C. albicans*, the beta-lactamase gene that confers resistance to ampicillin (*AMP*^{R}) in *E. coli,* inverted *C. albicans* telomere repeats separated by a gene that confers kanamycin resistance (*KAN*^{R}), flanked by polylinkers carrying multiple restriction enzyme recognition sequences and different potential replicating elements. After the *KAN*^{R} gene is excised (lower panel), the plasmid is linearized with telomere sequences exposed at the termini.

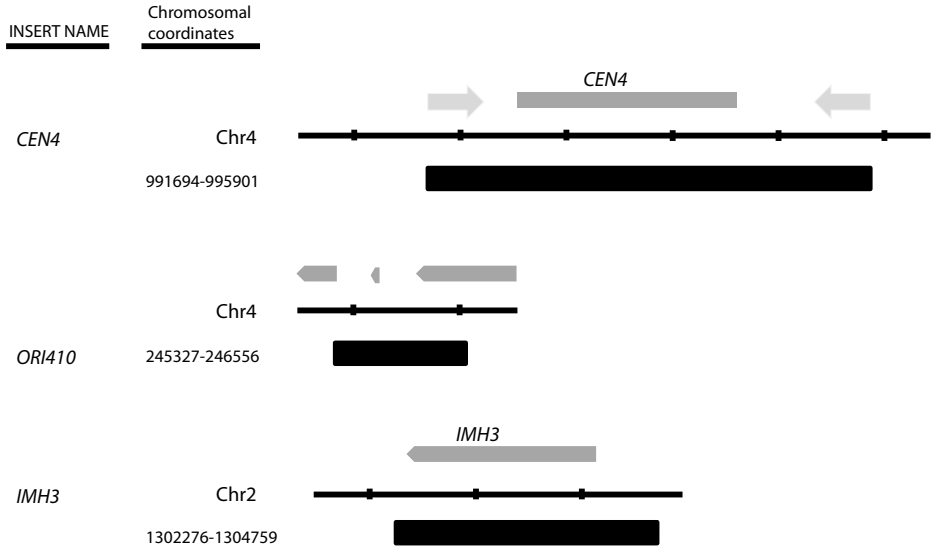


Figure 3.4. Analyzed origin fragments in this study.

CEN4 DNA containing flanking inverted repeats with 4 kb in size was cloned into the pLN plasmid (top). A 1.2 kb ORC binding sequence conferring origin activity in the native genome locus was cloned into the pLN plasmid (middle). This ORC binding site has been shown to be an active origin in the genome, named *ORI410*. A putative origin, *IMH3* gene including its promoter sequence, was cloned into pLN plasmid to test its ARS function (bottom).

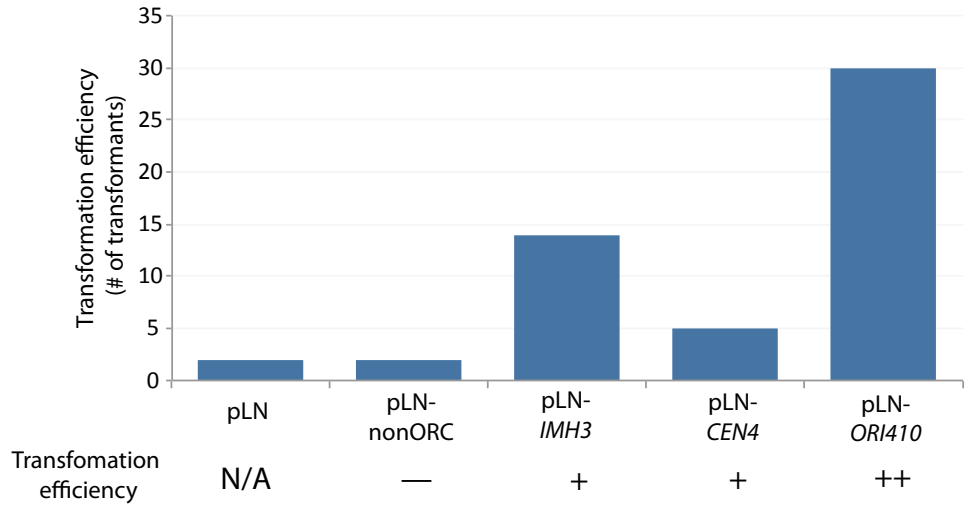


Figure 3.5. Origin DNA yields high transformation efficiency.

Both pLN plasmid vector and pLN carrying a non-ORC binding sequence can yield few transformants, considered as a background level. pLN plasmids carrying *ORI410* DNA yielded high transformation efficiency. pLN-*IMH3* and pLN-*CEN4* have relatively lower transformation efficiency but still higher than the background level.

A B

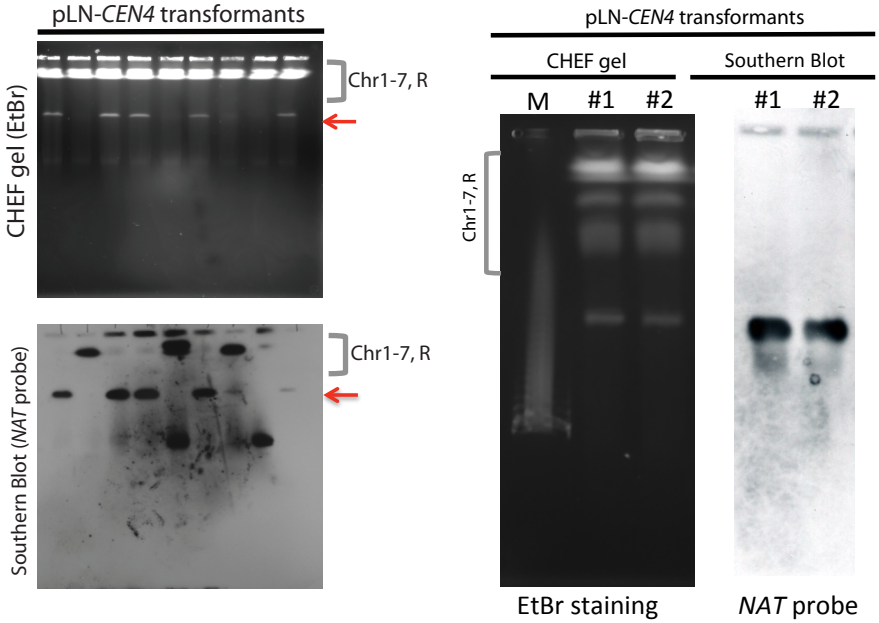


Figure 3.6. CHEF-Southern blotting analyses of pLN-*CEN4* transformants

(A) Five transformants containing pLN-*CEN4* have a distinguishable band without any chromosomal integrations, indicating pLN-*CEN4* is autonomously replicating in *C. albicans* cells. (B) Two transformants were analyzed again in another CHEF-gel. pLN-*CEN4* plasmids *in vivo* have over 700 kb in size.

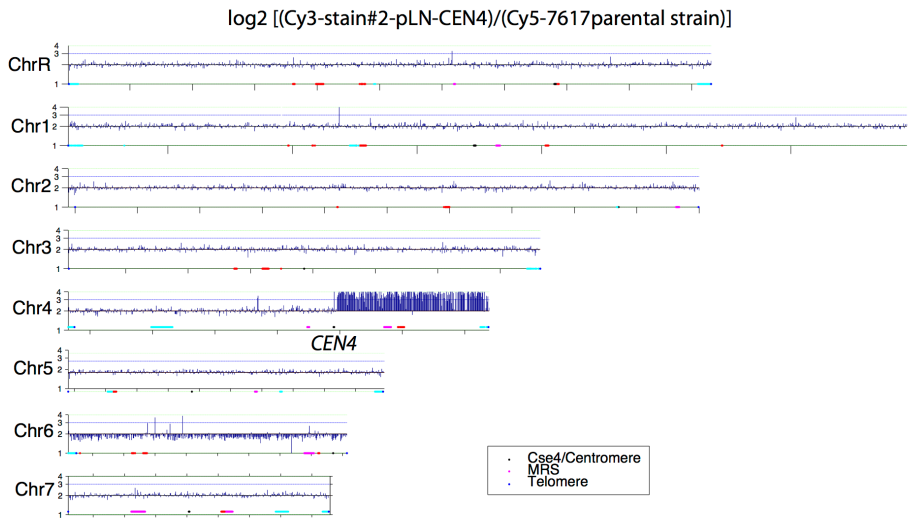


Figure 3.7. CGH analyses indicate DNA copy number in pLN-CEN4 transformants

Genomic DNA from a pLN-CEN4 transfromant and its parent strain was extracted and labeled to be hybridized on the CGH microarray to find the source of high molecular weight of pLN-CEN4 *in vivo*. The entire right arm of chromosome 4 is present in extra copies.

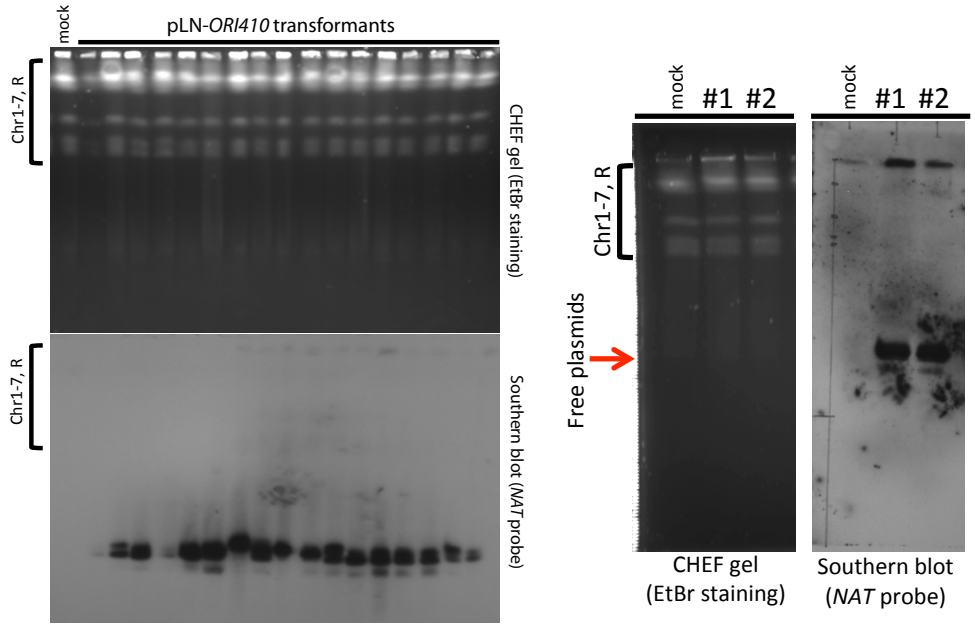


Figure 3.8. CHEF-Southern blotting analyses of pLN-*ORI410* transformants

(A) 17 pLN-*ORI410* transformants were analyzed, and all of them were autonomously maintained in cells without chromosomal integrations. (B) Two of the transformants were analyzed in another CHEF-Southern blotting experiment. The size of pLN-*ORI410* in vivo remains the same.as expected.

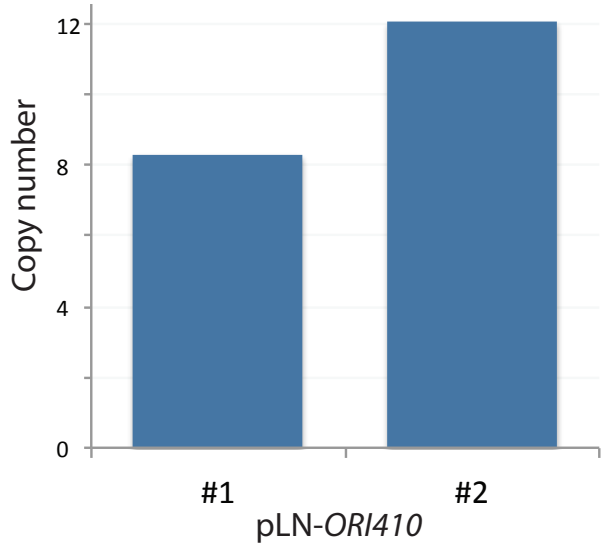


Figure 3.9. Plasmid copy number of pLN-*ORI410* plasmids

The plasmid copy number of pLN-*ORI410* was measured relative to CEN7 locus (2 copies). 8-12 copies of pLN-*ORI410* were estimated in a transformant.

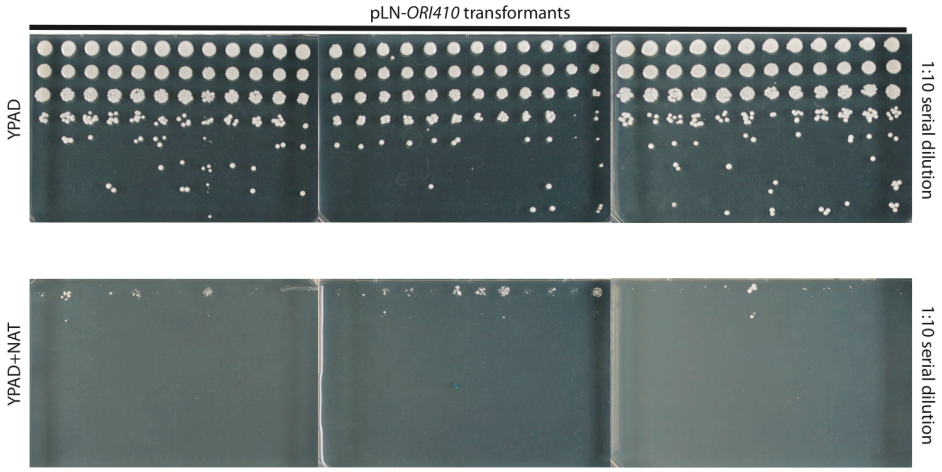


Figure 3.10. pLN-*ORI410* plasmids were maintained under nourseothricin (NAT) selection.

Plasmid loss assay was performed to examine if pLN-*ORI410* plasmids were maintained autonomously. After two days of growth in non-selective media, the transformants failed to grow on YPAD-NAT plates.

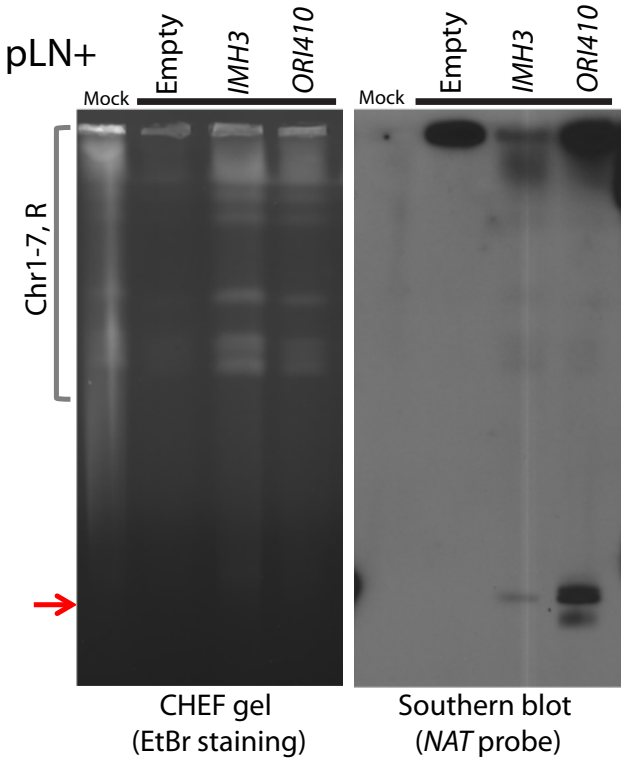


Figure 3.11. *IMH3* gene can confer weak ARS activity.

One transformant from pLN-*IMH3* transformation experiment were analyzed in CHEF-Southern blot. A weak band of plasmid size indicates autonomous replication of pLN-*IMH3* plasmids. The copy number of pLN-*IMH3* was low when compared to the pLN-*ORI410*.

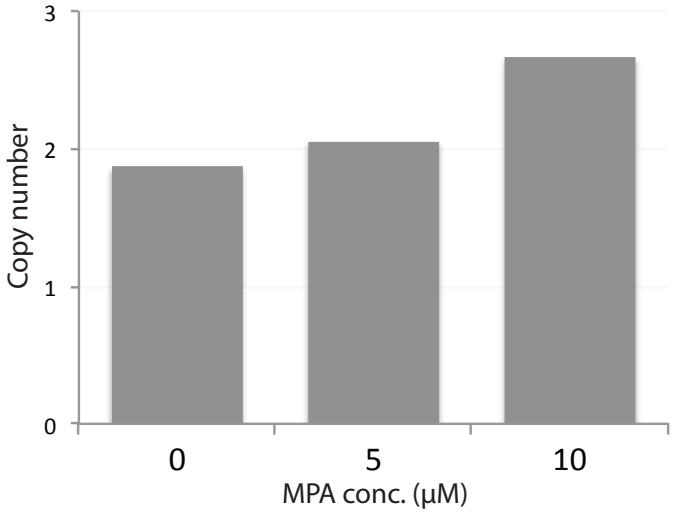


Figure 3.12. The addition of MPA slightly increased the copy number of pLN-*IMH3* plasmids

MPA was added in the nourseothricin medium to keep pLN-*IMH3* maintain in cells. After two days of growth, the copy number of pLN-IMH3 was measured, and the copy number slightly increased while MPA concentration increased.

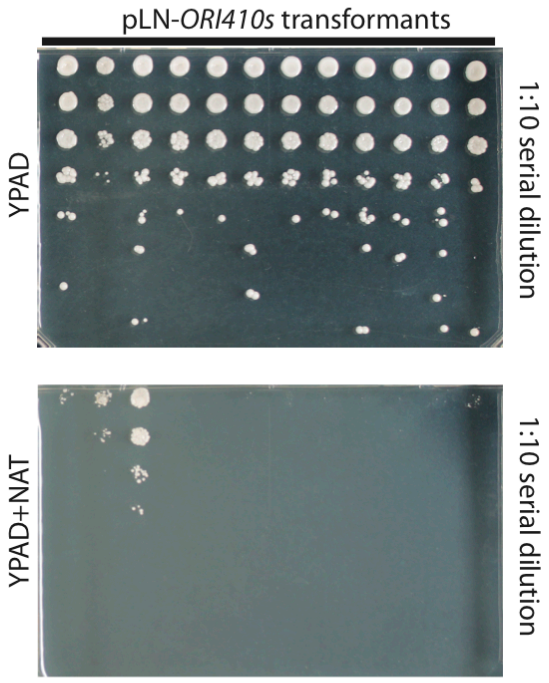


Figure 3.13. pLN plasmids carrying a 97 bp ARS fragment were maintained under nourseothricin (NAT) selection.

The 97 bp ARS fragment was isolated from 'mini-ARS' screen (Chapter 4, Figure 4.7, fragment 's'), and its transformants failed to grow after two days of growth in non-selective media, indicating its autonomous replication on pLN plasmids.

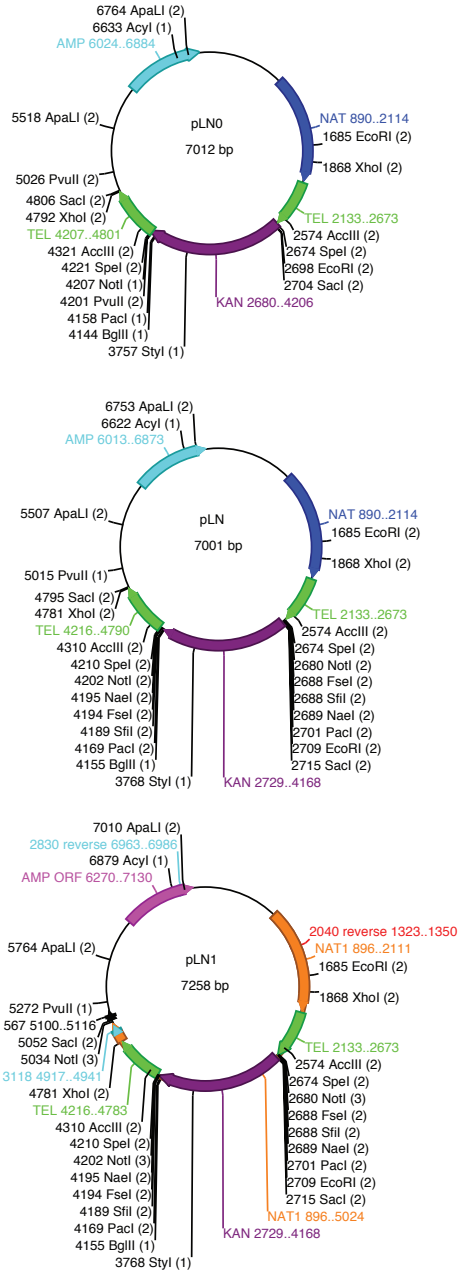


Figure 3.14. Restriction enzyme map of pLINEAR-NAT series plasmids (pLNs).

Chapter 4

Primary sequence and nucleosome depletion patterns specify replication origins in the *C. albicans* genome.

Introduction

Proper inheritance of genetic information from each cell to its progeny requires faithful DNA replication using replication machinery that is evolutionarily conserved among all eukaryotes. Conserved features of genome replication include the requirement for the highly conserved six-protein Origin Recognition Complex (ORC), which binds DNA and can recruit the pre-Replicative complex (pre-RC) and that in turn triggers the initiation of replication (Aladjem, 2007; Bell and Stillman, 1992). However, the mechanisms by which the positions of replication origins are specified remain elusive in most organisms. For example, while ORC binds within origin regions, it also binds to many regions that normally are not active origins, that may be dormant origins or that may not be origins at all.

The nature of DNA sequence requirements for replication origins differs considerably in different eukaryotes: the budding yeast, *Saccharomyces cerevisiae* has a defined 11-17 bp sequence necessary for replication, that, by analogy with the DNA sequence-defined 'point centromeres', can be considered 'point origins'. These 'point origins' function as Autonomously Replicating Sequences (ARSs) by conferring origin activity to extra-chromosomal plasmids (Chan and Tye, 1980; Malik and Henikoff, 2009; Nieduszynski, 2006; Stinchcomb et al., 1979; Theis and Newlon, 1997). Screens for ARS function identified an AT-rich ARS Consensus Sequence (ACS) motif that is necessary, but not sufficient, for origin function. Other flanking sequences, including a B-element composed of

T-rich sequences also contribute to efficient ARS function (Lee and Bell, 1997; Marahrens and Stillman, 1992; Rao and Stillman, 1995). Genome-wide mapping of replication origins in several other yeast species that are closely related to *S. cerevisiae* and have point centromeres, revealed conserved elements responsible for origin function (Di Rienzi et al., 2012; Liachko et al., 2010; 2011; Theis et al., 1999). For example, in *Kluyveromyces lactis*, a 50 bp ARS consensus sequence (*Kl*ACS) is necessary for ARS activity and several isolated ARSs were shown to direct origin firing in the genome. The *K. lactis* motif differs from the *S. cerevisiae* ACS motif, suggesting that the evolutionary constraints on sequences required for origin firing may be weak and that replication origin motif sequences may be difficult to recognize in more distantly related organisms (Mechali, 2010).

Much less is known about origin features in other organisms. Most organisms other than the *Saccharomycetaceae* yeasts, have regional, rather than point centromeres, in which function is inherited epigenetically rather than by virtue of DNA sequence alone (Allshire and Karpen, 2008). In most of these organisms the requirements for origin function have also remained elusive. In fission yeast, replication origins are found in AT-rich regions but no specific consensus motif is evident. These ~ 1 kb long AT-rich sequences are associated with the ORC4 subunit (Clyne and Kelly, 1995; Lee et al., 2001; Xu et al., 2012). In metazoans, such as *Drosophila* and mammals, ORC binds DNA without detectable sequence specificity, and origin DNA may not be defined by a specific sequence (Cayrou et al., 2011; Karnani et al., 2010; MacAlpine et al., 2010),

although origin G-rich repeated elements (OGREs) were recently reported within 67% of Drosophila and 90% of human cells, respectively and G-C-rich sequences that are able to form G-quartet structures also appear to be associated with human origins (Besnard et al., 2012). However, the degree to which these sequences are necessary or sufficient for origin function and the mechanism by which they act to direct origin firing remain to be elucidated (Gilbert, 2012).

In *S. cerevisiae*, all active origins within the genome can function as ARSs on plasmids, but not all ARS sequences can function as an origin in the genome. The features that distinguish ARSs from active origins are a subject of active investigation. Of the thousands of ACS motifs in the genome, only a few hundred are bound by ORC and only some of these are active origins. Furthermore, ORC binding sites in active origins are associated with a pattern of conserved nucleosome positioning, which appears to be important for origin activation in the genome (Eaton et al., 2010a; Lipford and Bell, 2001). Thus, specific nucleosome positioning patterns around the origins may specify which ACS regions bound by ORC do, or do not, serve as active origins in the genome. Based on these observations, I hypothesized that a conserved pattern of nucleosome positioning, combined with the identification of ORC binding sites, may distinguish replication origins in other species regardless of the presence of defined replication origin sequence. I tested this hypothesis in *Saccharomycotina* species using the

pathogenic yeast, *Candida albicans*, an ascomycete distantly related to *S. cerevisiae*.

In chapter 1, replication timing profiles indicated that *C. albicans* centromeres are tightly associated with highly efficient origins that fire first on each chromosome. Importantly, these centromeric origins were epigenetically inherited together with centromere function, demonstrated by the appearance of new, highly efficient origins associated with neocentromeres (Koren et al., 2010b). However, the low resolution of these replication timing profiles did not enable precise localization of non-centromeric origin positions in the *C. albicans* genome. Given that *C. albicans* has regional centromeres, it was not clear whether a specific DNA sequence motif would function as an ARS or origin in *C. albicans*. A *bona fide* ARS or ACS motif in *C. albicans* genome has not been discovered.

Here I combined computational and experimental genome-scale approaches to comprehensively predict replication origin features in the pathogenic yeast *C. albicans*. I combined ORC binding site positions and conserved nucleosome occupancy positions to predict replication origin locations throughout the entire *C. albicans* genome. Non-denaturing two dimensional gel analyses of selected predicted origins provided direct evidence of origin function at several non-centromeric loci in the *C. albicans* genome. In chapter 3, I described the development of a linear plasmid, which is maintained as an autonomously replicating plasmid. Using this linear plasmid system, I determined

the minimum DNA length for necessary for ARS activity and then performed a genome-scale ARS screen, which yielded over 100 additional ARSs and a 15 bp ACS motif. Mutagenesis of this motif inactivated its ARS function, providing support for the idea that a specific defined DNA sequence is necessary for replication origin function in a eukaryote with an epigenetically defined regional centromere.

Materials and Method

Yeast strains and growth condition

Yeast strains used in this chapter were *Candida albicans* SC5314 and RM1000 #2 (YJB7617). Strains were stored as stocks in 50% glycerol at -80 °C and grown in YPAD medium (1% yeast extract, 1% peptone, 2% glucose, 1.5% agar) at 30 °C. The selection medium contains 400 μg of nourseothricin (clonNAT). Yeast transformation were based on the Lithium acetate transformation protocol (Wilson et al., 1999).

Chromatin immunoprecipitation (ChIP)

The procedures of chromatin immunoprecipitation were modified from Ezhkova et al. and Ketel et al. specifically for ORC or MCM ChIP experiments. The antibodies were kindly provided by Dr. Stephen P Bell. Yeast strains were cultured overnight in 30 °C and then were diluted to an OD_{600} of 0.2 in YPAD medium. Log-phase cells were harvested at OD_{600} of ~1 after 3 hours growth.

Cells were fixed in 1% formaldehyde for 20 minutes at 30 °C and 125 mM for 5 minutes at room temperature. Cells were spun down at 4 °C, 2500-3000 rpm for 5 minutes. Cell pellets can be stored or continued to the next step after discarding the supernatant. Each cell pellet was suspended in 450 µl lysis buffer (50 m*M* HEPES, pH 7.5, 140 mM sodium chloride, 1 m*M* EDTA, pH 8.0, 1% Triton X-100, 0.1% DOC, 0.1% sodium dodecyl sulfate, Roche proteinase inhibitor cocktails), and then an equal volume of acid-washed glass beads added (proximally 500 µl). The cell pellets were vortexed either on a bead-beater for 1 minute each time, 6 times (one minute on, one minute off) or on a vortex for 10 minutes each time, 4 times (10 minutes on, 2 minutes off). The lysed cells are seen with a broken cell shape or phase-dark under the microscope. After puncture of the bottom of the 1.7 ml tubes with a needle, the tubes were placed into new tubes to collect the cell lysates by spinning down with full speed for 10 seconds. The collected cell pellets were vortexed again for 30 second in 4 °C, and then the cloudy supernatants were collected. The supernatants were placed on ice and sonicated to yield DNA fragments of 500-1000 bp (4 times of sonication, 15 second per time, output power 20 %). Soluble lysate was collected after insoluble matter was precipitated twice by centrifugation in 4°C, 10 minutes each time. The volume of collected soluble lysate was adjusted to 2 mL by lysis buffer and 5 % of it will be collected as input lysate, which was stored in -80 °C until reverse-crosslink step. The rest of the cell lysate was split into two reactions with and without antibodies, respectively. After each reaction was incubated at 4°C overnight, protein A-agarose beads were added for another 3-4 hours

incubation at 4°C. Then, the reaction was washed subsequently with cold lysis buffer, high-salt washing buffer (50 m*M* HEPES, pH 7.5, 500 mM sodium chloride, 1 m*M* EDTA, pH 8.0, 1% Triton X-100, 0.1% DOC, 0.1% sodium dodecyl sulfate), DOC buffer (10 m*M* Tris-HCl, pH 8.0, 0.25 *M* lithium chloride, 0.5% Nonidet- P40 (NP-40), 0.5% sodium deoxycholate (DOC), 1 m*M* ethylenediamine- tetraacetic acid (EDTA, pH 8.0)), and TE (10 m*M* Tris-HCl, pH 8.0, 1 m*M* EDTA, pH 8.0) at 4°C, 5 minutes for each washing step. DNA was eluted by adding 100 TES buffer (50 m*M* Tris-HCl, pH 8.0, 10 m*M* EDTA, pH 8.0, 1% SDS) twice in 65°C, 10 min for each time. Total 200 µl of eluted DNA was incubated overnight in 65°C and followed by RNase A and proteinase K treatment. Proteins were removed by phenol-chloroform/isoamyl alcohol/chloroform extraction. Samples were ethanol precipitated with 20 ug of glycogen overnight at 4°C. Finally, DNA was collected by centrifugation, washed in 70% ethanol and resuspended in 50-80 µl TE, depending on the input protein concentration.

ChIP-on-chip was performed according to Agilent protocols, and ChIP DNA was labeled and hybridized to custom designed Agilent microarrays (template number available upon request) containing ~70 bp probes spanning the entire genome using a custom designed microarray. ORC binding positions were identified by a hill climbing algorithm (500 bp window) with significant binding considered as signals at least two standard deviations (SD) above the mean of the population.

Nucleosome-ORC association pattern analyses

Proposed origins of replication (*pro*ORIs) were predicted using a Least Absolute Shrinkage and Selection Operator (LASSO) regularized logistic regression model trained on ORC binding as determined by ChIP microarrays combined with published nucleosome occupancy patterns (Tsankov et al., 2010). Model training and prediction was implemented using the glmnet package of the R Statistical Software with a binomial kernal and LASSO regularization (Friedman et al., 2010). For training, established ORI regions in *K. lactis* served as a positive set while random *K. lactis* regions served as a negative set with 6-fold cross-validation. The likelihood of *pro*ORI presence was evaluated at each position in the *C. albicans* genome, using the *K. lactis* trained model, based on a centered 1024bp window.

DNA two-dimensional gel electrophoresis (DNA 2D gel)

The procedures of DNA 2D gel electrophoresis were modified from Leizmeier et al. and personal communications with Bonita Brewer's lab.

DNA preparation:

Yeast cells were grown overnight (OD_{600} of ~25) and subcultured to an OD_{600} of 0.8 in pre-warmed YPAD medium at 30°C for 40 minutes or 2 hr. Sodium azide was added to 0.1% and quickly placed on ice. Ice-cold EDTA was added immediately to final 25 mM before spinning cells at 4°C, 7000 rpm for 5 minutes. Cell pellets were washed with ice water before adding 0.2 ml ice-cold NIB buffer

(17% glycerol, 50 mM MOPS buffer, 150 mM potassium acetate, 2 mM magnesium chloride, 0.5 mM spermidine, and 0.15 mM spermine; pH is adjusted to 7.2) to resuspend cell pellets in 15 ml tube. With 0.2 ml NIB, cell pellets can be frozen at -20°C or continued to the next step. 0.9 ml glass beads were added, and placed on a vortex to lyse the cells; vortexing 5-10 times for 2 minutes (2 minutes on; 1 minute on ice) or until cells were 75% lysed by seeing broken or phase-dark cells under the microscope. 0.6 ml NIB was added into the reaction to remove the supernatant to another tube. The step was repeated on ice to get 1.3 ml of cell lysates. Nuclei were spun down by centrifugation with full speed at 4°C for 30 minutes, and then pellets were resuspended in 0.3 ml TEN buffer (50 mM Tris, 50 mM EDTA, 100 mM NaCl). RNase A was added on ice for 5 minutes before proteinase K was added with 12 µl of 25% Sarkosyl. The reactions were mixed gently by inverting and transferred to 37°C for 30 minute incubation. The lysed nuclei were spun down for 5 minutes, full speed at 4°C. The supernatant needs to be transferred gently with a cut pipette tip, and then DNA was extracted by 0.3 ml phenol-chloroform/isoamyl alcohol/chloroform. The mixes were gently mixed by a dozen inversions. The final volume of extracted DNA solution at this step is around 0.2-0.25 ml. 2 volumes of absolute ethanol containing 0.5 M potassium acetate were added, and then the reaction was slowly mixed by inverting. The reaction was rest on ice for 5-10 minutes until the white DNA spool was visualized at the bottom of the tube. After discarding the alcohol, DNA pellet were washed by 0.3 ml 70% ethanol for 5 minutes at top speed. DNA pellet can be resolved in 75-100 µl of TE at 4 °C overnight.

DNA digestion and electrophoresis

20 µg of extracted DNA were digested using proper restriction enzyme, which can yield a 3-6 kb targeting DNA fragments. 300-400 unit of restriction enzyme was added in final volume 600 µl for 5-6 hours. Digested DNA was ethanol precipitated and was resolved in 18 µl of TE at the end for gel electrophoresis. The first dimension was in 0.35% agarose gel and ran by 1.2 V/cm, 20-22 hours; the second dimension was in 0.95% agarose gel and ran by 5.5V/cm for6-8 hours in 4 °C cold room.

Southern blot

DNA was transferred from agarose gels to nylon membranes (Amersham). Membranes were probed overnight at 65°C (Sigma, PerfectHyb prehybridization buffer). The probes were labeled with alpha-^{32}P-dCTP (Agilent, Prime-it II Random primer labeling kit). The primers used to make probes are listed in Table 2.1. The membranes were exposure to film for 3-5 days at -80°C.

Plasmid constructions

Linear cassette on pLN (p2382) was subcloned by XhoI into XhoI site on pGEM-NAT (p2339). The plasmid was called p2414 initially and then called pLN1 in this Chapter (Chapter 3, Figure 3.14, bottom).

Mutagenesis on pLN-ORI410 (p2402) and its subclone from mini-ARS screen pLN-ORI410s (p2436) were generated by the mismatched PCR amplifications using primer 5936:

TCGACACTTTCAGAAATTGGTTGGCCAAGTGGAAATGGCACATTTCCCTCCT

CGGTTGCCAGTTTATCTAATCTGGAATATTTTGCTAATGATTTTTTGTGTG; and primer 5937: TCGACACACAAAAAATCATTAGCAAAATATTCCAGATTAGATAAACTGGCAA CCGAGGAGGGAAATGTGCCATTTCCACTTGGCCAACCAATTTCTGAAAGTG. Two oligos were annealed together as double-strand DNA and then was inserted into the PvuII site of pLN. The plasmid carrying a mutated ARS was named p2440 and noted as mutACS in this Chapter.

Plasmid library constructions

mini-ARS screen

Three identified *ORIs* (ranging from 1.2 to 3.2 kb in size) were amplified from their genomic loci by PCR, extracted from the agarose gels to prevent the contamination from genomic DNA. The purified PCR products were digested with AluI and treated with alkaline phosphatase (CIP) to prevent the multimerization of small-size fragments. In parallel, the purified PCR products were treated with DNase I and then alkaline phosphatase. Digested DNA from two independent methods of digestion were ran on agarose gels, respectively. The digested DNA fragments between 0.1kb and 1 kb in length were purified from the agarose gels and then ligated into the pLN1 vector that had been digested with PvuII and transformed into competent *E. coli* cells. Three mini-ARS libraries were transformed into yeast and the clones from the mini-ARS library conferring high transformation efficiency were selected for Sanger sequencing reactions.

Genomic ARS library

Purified genomic DNA from *C. albicans* SC5314 was digested with restriction enzyme *Rsa*I and treated with alkaline phosphatase. The digested genomic DNA fragments were ligated into the *Pvu*II site on the pLN1 vector, the ligation products were transformed into competent *E. coli* cells, and the transformed colonies were pooled to prepare plasmids to be used to transform *C. albicans.* The genomic ARS library was transformed into *C. albicans*. Over 1200 individual yeast transformants were picked up to examine nourseothricin resistance and then were pooled together to make CHEF gel plugs. CHEF was performed and plasmid sized DNAs were separated from the chromosomes to be extracted from the gels. Plasmid containing DNA was extracted from gel slices and amplified to subclone back into the PvuII site on pLN1 vector. The enrichment library was made and transformed into *C. albicans*. Over 1000 individual yeast transformants were pooled together for another round of CHEF gels. Plasmids containing DNA were isolated and extracted again, and then amplified inserts were subjected into Illumina Sequencing.

CHEF gel

CHEF plugs were prepared as described previously (Selmecki et al. 2008). Cells were washed in 50 mM EDTA and lysed with zymolyase in 1% low melt agarose prepared in a plug mold. Plugs were incubated in LET buffer (0.45 M EDTA pH 8.0, 0.01 M Tris pH 7.5) with 1 % beta-mercaptoethanol in 37°C overnight during

zymolyase digestion. Plugs were washed twice with 50 mM EDTA, pH 8.0 and digested by 0.2 mg/L proteinase K in NDS buffer (0.01 M Tris pH 7.5, 0.45 M EDTA pH 8.0, 1% N-lauroyl sarcosine) at 50°C for at least 48 hours. After washing with 50 mM EDTA, pH 8.0 twice, 30 mini each time, plugs can be stored in 50 mM EDTA, pH 8.0 at 4°C.

Plugs were run on a CHEF DRIII electrophoresis system (Bio-Rad) on a 1% Megabase agarose gel in 0.5x TBE under the following conditions: 11°C, 60-120 switch for 14 or18 hours at 6 V/cm with 120° angle to visualize DNA fragments. Linear plasmids can be separated from chromosomes.

Results

ORC binds multiple sites in the C. albicans *genome*

ORC is associated with all active origins and thus I mapped ORC binding sites across the *C. albicans* genome by chromatin immunoprecipitation (ChIP) from log-phase cultures using polyclonal anti-ORC complex antibodies followed by microarray hybridization ChIP-chip (Wyrick et al., 2001). A large number (970) peaks above the average binding ratio were mapped with high resolution. Consistent with our previous work, the eight origins within the *C. albicans* regional centromere central cores bound ORC with high affinity.

In total, 393 ORC binding sites were considered significant, based on a binding ratio greater than two standard deviations above the mean (Table 4.1, Figure 4.1). Our previous replication timing profile analyses estimated ~150

active replication origins in the *C. albicans* genome (Koren et al., 2010b). Furthermore, the number of ORC binding sites is generally much larger than the number of active origins (Shor et al., 2009). Thus, I assumed that only a subset of ORC sites in the *C. albicans* genome fire efficiently during a given cell cycle.

In *S. cerevisiae* almost all replication origins are intergenic and the ACS is positioned within a nucleosome free region (NFR) (Berbenetz et al., 2010). Consistent with this, nearly half of the significant ORC binding sites were intergenic (47%) and the nucleosome occupancy patterns within 0.5 kb upstream and downstream of each ORC binding site were associated with regions shown to have low nucleosome occupancy (Figure 1A). These ORC-proximal nucleosome free regions (NFRs) are similar in size to those identified in *S.cerevisiae* (~0.5 kb) and are smaller than those detected in *S. pombe* (5 kb)(Field et al., 2008; Tsankov et al., 2010).

An unexpected result was that *C. albicans* ORC binding exhibited a strong association with tRNA loci: 94% (119/126) of tRNA loci were within 1 kb of the center of all ORC binding peaks (Figure 1B). A subset of tRNAs are associated with ORC sites in *S. cerevisiae* (13%, 38/284) and with ARSs *in K. lactis* (17 %, 22/131). The tRNA-ORC association in *C. albicans* is highly significant (*p-value* < 2.3e-162. Accordingly, this raises the possibility that tRNA-related factors may contribute to ORC binding and/or to origin function in *C. albicans*.

Conserved nucleosome positions predict origin locations.

In *S. cerevisiae*, origin location is known, and has been characterized and catalogued (Siow et al., 2012); ORC binding is necessary, but not sufficient, for origin firing and ORC binding at origins precisely positions the flanking upstream and downstream nucleosomes in the local chromosomal context (Eaton et al., 2010a). Thus, a specific pattern of nucleosome occupancy, characterized by a nucleosome-free region (NFR) surrounded by phased nucleosomes, provides a signature that distinguishes replication origins from DNA sequences containing ACSs but that are not active origins. A similar pattern of nucleosome positioning is seen at *K. lactis* origins (Tsankov et al., 2010). We used a logistic regression model trained on the ORC and nucleosome occupancy patterns associated with *K. lactis* ARSs (Materials and Methods) to predict the positions of *C. albicans* origins (Figure 4.3).

As an indicator of model accuracy, we evaluated the Precision-Recall curves of a *K. lactis* trained model on *K. lactis* ORI predication. Our predictions successfully predicted 30-35% of identified origins in *K. lactis.* Since predications were only based on empirically derived nucleosome occupancy, and not DNA sequence, I then asked whether the known *Kl*ACS motif was present at these predicted *K. lactis* origin sites. Indeed, 33% of the sequences predicted to be *K. lactis* origins have a *Kl*ACS motif falling within a 1 kb window.

We then identified proposed origins of replication (*pro*ORIs) in the *C. albicans* genome by combining computational analyses of the conserved patterns of nucleosome positioning at *K. lactis* origins, together with all 970 ORC binding sites. The computation-based prediction method generated 386 *pro*ORIs

throughout the genome (Table S2). The number of origins on each chromosome was roughly proportional to chromosomal length (Figure 4.4A). As expected, based on the prediction criteria, *pro*ORIs in *C. albicans* are located in nucleosome free regions (Figure 4.4B, left) and tRNA loci are mostly coincident with *pro*ORIs (Figure 4.4B, right). Furthermore, they are biased towards intergenic regions (83%), a predicted genomic feature of replication origins in other organisms (Nieduszynski, 2006).

Some proORIs are active origins in their chromosomal context

Two-dimensional non-denaturing gels (2D gels) are considered direct indicators of origin firing, as they can visualize the “bubble structure”, representing bidirectional replication forks that are actively replicating a given origin (Figure 4.5A) (Brewer and Fangman, 1987). Furthermore, *C. albicans* centromeric origin firing has been detected on 2D gels (Koren et al., 2010b). Here, we first examined origin activity at the 5S ribosomal DNA (rDNA) region, because rDNA is highly conserved throughout yeast species and is present in 55-60 identical tandem units with potential origin activity within each unit (Linskens and Huberman, 1988). Using unsynchronized log phase cultures, I detected activation of the rDNA origin (Figure 4.5A, bottom panel.). Thus, origin firing can be detected within a defined, non-centromeric region of *C. albicans* genomic DNA using 2D gels.

I next chose five other *pro*ORIs for two-dimensional non-denaturing gel analysis. These analyses required higher detection sensitivity because each

genomic locus is present in only two copies per diploid cell. Nonetheless, I detected origin activity at three of these five *pro*ORIs (Fig 4.5B, top panel. *pro*ORI055, *pro*ORI1046, and *pro*ORI410). I also examined the possible origin activity of ORC binding sites that did not have well-positioned nucleosomes and several random genome regions that do not bind ORC. In all of these cases, no bubble arc intermediates were observed (Figure 4.5B low panel right). Thus, the computational predictions of *pro*ORIs revealed at least three *bona fide* origins that fire actively within their native genome context and thus are termed 'ORI' rather than *'pro*ORI'.

Identification of a minimal origin DNA region as a functional ARS

ORI410 can confer ARS function on a linear plasmid vector, as described in Chapter 3. With *bona fide* origins of replication in hand, I was able to search for the minimal length of ORI sequences for origin function. Origins of replication in *S. cerevisiae* and *K. lactis* are generally ~200 bp in length; the ORIs identified were much larger (~1.2 kb), primarily because the prediction algorithm used ~1 kb windows to calculate nucleosome occupancy. To ask if the ARS function within these ORIs could be narrowed down to a smaller region that had origin function, I performed a 'mini-ARS screen' (minimal ARS) (Figure 4.6) (Liachko et al., 2012). Essentially, I amplified fragments of the three ORI DNAs from the genome (1-3 kb regions, Figure 4.5B) by PCR and digested the DNA into smaller fragments using either restriction enzymes such as AluI that recognize 4 bp sequences or using DNase I, which digests DNA randomly. The resulting small

fragments (average ~ 100-200 bp) were inserted at the PvuII site of the pLN plasmid to generate ORI-specific 'mini-ARS libraries'.

The three different mini-ARS libraries, generated from the three ORIs (Fig. 4.7), were screened for high transformation efficiency in *C. albicans*. The minimal size of inserts in transformants ranged from 65 to ~300 bp. For *ORI1046*, the smallest fragments isolated were ~225 bp (Figure 4.7, middle). For *ORI055* and *ORI410* (Figure 4.7, top and bottom panel), DNA fragments from more than one region of the original ORI region were able to confer high transformation efficiency. Of note, the tRNA within *ORI410* was not included in the minimal ARS regions. Taken together, these results indicated that fragments of 100-300 bp in length from replication origins that function in the *C. albicans* genome can confer ARS activity, defined as the ability to replicate autonomously on the linear plasmid backbone. Furthermore, at least in the tested origins, the tRNA within in the origin is not required for ARS activity. The observation that more than one sub-region of origin DNA around an origin could confer ARS activity suggests that chromosomal origins may be composed of more than one region with a high likelihood of initiating DNA replication. Indeed, in *S. cerevisiae*, many efficient origins are associated with multiple ACSs.

Isolating C. albicans ARSs using linear plasmids

The identification of ORIs using origin prediction algorithms that relied upon conserved nucleosome position patterns and ORC binding sites followed by confirmation through the detection of bubble arcs in DNA 2D gels, is a very labor

intensive approach that is not feasible for the detection of large numbers of origins. Once it became clear that origin sequences could confer ARS function, I set out to isolate large numbers of genomic DNA sequences that confer ARS function by performing a genome-wide screen for ARS activity. In the *Saccharomycetaceae* yeasts, plasmid-based ARS screens identify potential origins at the sequence level (Hsiao and Carbon, 1979; Liachko et al., 2010), although not all ARSs are active origins in their native genomic context.

We constructed ARS screening libraries using random fragments of *C. albicans* genomic DNA inserted into plasmid pLN1 and then linearized the resulting libraries to excise the *KAN*^{R} stuffer sequence between the telomere repeats (Figure 4.8). The resulting linearized plasmid libraries were transformed into *C. albicans*. To ensure that I was enriching for plasmids that remained autonomous rather than any DNA that may have integrated into the genome, I performed an additional "ARS enrichment" step (Figure 4.8, step 4-6), in which I pooled all the transformants, separated the episomal plasmids from the chromosomes on CHEF gels; and then extracted the episomal linear plasmids from the gels. Inserts in these plasmids were sub-cloned from the extracted plasmids to produce an 'enriched' ARS library that underwent a second round of screening by linearization and transformation into *C. albicans*. After pooling all the transformants from the enriched libraries, I again isolated the ARS-containing plasmids on CHEF gels, amplified the inserts using plasmid-specific primers and

then analyzed the insert regions using Illumina deep sequencing to identify putative ARSs.

This ARS-enrichment screen identified 127 unique ARSs, from all 8 *C. albicans* chromosomes (Table 4.3.). Strikingly, two ARSs isolated in the screen, *Ca*ARS015 and *Ca*ARS403, overlapped with the mini-ARS fragments shown to be functional in the mini-ARS screen (*ORI055* and *ORI410*, respectively in Figure 4.7, labeled with stars). This result strongly suggests that the ARS screen can identify sequences that have origin function in the native genome based on their plasmid activities.

In addition, to determine if the sequences cloned in the ARS screen were reproducibly able to confer ARS function, two newly identified ARSs were manually sub-cloned back into the pLN1 plasmid. Importantly, both of these ARSs (74 bp *Ca*ARS305 and 229 bp *Ca*ARS021) conferred high transformation efficiency to the pLN1 backbone (Figure 4.9A).

A primary sequence motif for ARS function

Given that defined short DNA fragments conferred ARS activity, I next asked whether these fragments had distinctive primary sequence features (e.g., a consensus sequence). To this end, I compared the 127 ARS sequences isolated in the ARS screen, relative to 1000 negative control sequences using the MEME Suite (Bailey et al., 2009). An AC-rich 15 bp motif was detected in the isolated ARSs (Figure 5b, upper) (*E-value* 1.9e-001). In parallel, I also compared the origins from the prediction (*pro*ORIs), which were identified through

nucleosome-ORC associated patterns prior to the ARS screen (Table 4.2, 386 *pro*ORIs). With the *pro*ORI sequences, a similar AC-rich motif was identified (*E-value* 1.3e-132), albeit with different positional weights for the nucleotide composition (Figure 4.10, middle). Moreover, a similar AC-rich motif (*E-value* 1.6e-154) was also found within the significant ORC binding sites (Figure 4.10, bottom). The similarity of these motifs suggests that all ORC binding sites may be potential ARSs or origins. Importantly, these related motifs were identified in two independent screens—through the plasmid-based ARS assay, which is thought to isolate origin motifs in a sequence-dependent manner, and from the *pro*ORI predictions, which were determined from experiementally determined nucleosome depletion patterns and ORC binding affinity. Furthermore, the overlap of the *Ca*ARSs with three *pro*ORIs, as well as the similarity of the ACS and *pro*ORI consensus motifs, indicates that the sequence-independent algorithms used to predict *pro*ORIs can accurately identify genome regions capable of specifying replication initiation.

The nucleotide composition in the *Ca*ACS motif is neither similar to the well-known 11-17 bp T-rich *Sc*ACS motif in *S. cerevisiae, nor* to the 50 bp *Kl*ACS motif in *K. lactis* nor to the AT-rich character of *S. pombe* origin regions. In addition to the CA-rich motif, I also identified a 21 bp long T-rich motif within the *pro*ORIs (Figure 4.10, bottom). Indeed, the T-rich sequences usually aligned with the centers of the *pro*ORI positions. This is not surprising, given that nucleosome depletion was a feature used to define *pro*ORIs, and T-rich sequences generally provide strong nucleosome exclusion signals (Kaplan et al., 2009). Over half (55

percent) of the predicted *pro*ORIs (213/386) have both primary sequence motifs (AC-rich *Ca*ACS and T-rich, *p-value* < 0.0001) located within a 1 kb window. I hypothesize that this conserved T-rich motif in *pro*ORI sequences may serve as an auxiliary B-element for origin activity in *C. albicans* genome.

In *S. cerevisiae*, origins have apparent A-T asymmetry, resulting from the T-rich *Sc*ACS motif and A-rich B elements. To ask if a nucleotide asymmetry is present at *C. albicans* origins, I examined the relative positions of the AC-rich *Ca*ACS motif and the T-rich sequences found in *pro*ORIs (Figure 4.11). Intriguingly, the AC-rich *Ca*ACS motifs were usually located upstream relative to the position of the *pro*ORIs (Figure 4.11, top) and the T-rich sequences (putative B-elements) were located approximately 100 bp downstream from the *Ca*ACS motif (Figure 4.11, bottom), indicating that A:T asymmetry may exist at *C. albicans* origins as well. Thus, origins appear to be composed of similar required elements (ACS motif and auxiliary elements) even if the primary sequences of these elements differ. Taken together, motif analysis suggests that origins and ARSs in *C. albicans*, as in other yeasts, have primary sequence features either necessary for origin activity directly and/or necessary for shaping local chromatin structure, which in turn facilitates origin firing.

The 15 bp CaACS motif is necessary but not sufficient to fire an origin in the genome context

To test the hypothesis that the 15 bp *Ca*ACS motif is functionally linked to origin activity, I altered the nucleotide composition of the AC-rich motif by site-

directed mutagenesis, changing a central ACC to GGG and TTT (Figure 4.12A) and compared the transformation efficiency of plasmids carrying the mutagenized and wild-type motifs. Importantly, the plasmid carrying the mutated motif yielded significantly fewer transformants than the plasmid with the intact motif (Figure 4.12B, left), consistent with the idea that the *Ca*ACS is necessary for ARS activity. In parallel, I also mutagenized the downstream AC-rich nucleotides and tested for ARS activity. No obvious difference in transformation efficiency between plasmids containing wild type sequences or mutations in the downstream of *Ca*ACS was observed (data not shown). Thus, the 15 bp *Ca*ACS motif identified from isolated ARS sequences is important for origin activity.

To ask if *Ca*ACS is required for pre-RC recruitment, I asked if the mini-chromosome maintenance (MCM) complex was associated with ARS plasmids by ChIP with anti-MCM complex antibodies. MCM was enriched on the plasmid with an intact CaACS motif and, consistent with the yeast transformation experiment, MCM association with the plasmid carrying the mutant *Ca*ACS motif was reduced (Figure 4.11B, right). These results support the idea that the *Ca*ACS motif is required for active plasmid replication.

In *S. cerevisiae*, ACS motifs are also necessary, but not sufficient, to direct origin firing: of the over 12,000 *Sc*ACS motifs, only ~400 are estimated to be *bona fide* origins. Similarly, *Sc*ORC binding is necessary, but not sufficient, for origin firing (Eaton et al., 2010a). Accordingly, I asked whether *Ca*ACS motifs that are also ORC binding sites are sufficient to direct origin firing within their genome context. To answer this question, I used 2D gels to compare the origin

activity at two genomic loci that were identified in the ARS screen. Both loci had similar ORC binding occupancy but one, *Ca*ARS222, was also a predicted origin (*pro*ORI*)* with an associated nucleosome depletion profile, while *Ca*ARS413, was not a *pro*ORI because it was not associated with a nucleosome depleted region despite being within an intergenic region.

Importantly, *Ca*ARS222 (*pro*ORI246) did form a bubble arc (Figure 4.13, left), indicating that it is an active origin. In contrast, intergenic *Ca*ARS413 formed only a Y-arc and no bubble-arc, consistent with the idea that it does not fire in its chromosomal context despite its ability to function as a *Ca*ARS on a linear plasmid (Figure 4.13, right). Thus, as in *S. cerevisiae*, the *Ca*ACS motif and the ability to bind ORC are important for origin activity, but are not sufficient to direct efficient origin initiation in the genome context. Rather, other criteria, including appropriately positioned nucleosomes, are required for origin firing in the genome. Thus, I suggest that, as in other organisms, ACS sequences represent dormant origins, that have the potential to fire but that, in the genome, are likely to be replicated passively (e.g., by a more efficient origin that fires earlier at an adjacent position on the chromosome).

Discussion

Despite early work identifying several possible *C. albicans* ARS sequences on circular plasmids, the work presented here is the first to comprehensively map replication origins and to demonstrate *bona fide* origin

firing from a specific DNA region in *C. albicans*. The combination of computational genomics and experimental approaches identified short genomic DNA regions that formed bubble arc replication intermediates and that could confer autonomous plasmid maintenance to a linear plasmid. Development of a linear plasmid system for *C. albicans* allowed us to demonstrate that origin DNA, but not random DNA, confers replication activity (Chapter 3) and to delimit the minimal DNA fragment size necessary to direct autonomous replication. Based on the ability to detect plasmid replication, a genome-wide screen identified 127 ARSs and revealed a 15 bp *Ca*ACS motif that is necessary for replication function. Additionally, this work has established a linear plasmid tool that will be useful for genetic manipulations of *C. albicans.* Detailed analysis of plasmid components will be described in detail elsewhere (Chapter 3).

This work is the first example in which origin function was predicted without the *a priori* knowledge of an associated primary DNA sequence and was then experimentally demonstrated. In *S. cerevisiae*, nucleosome depleted regions and ORC binding have always been analyzed in the context of the well-characterized ACS (Breier et al., 2004; Nieduszynski, 2006); for *K. lactis* and other related species, ARS activity was determined first and then origins were identified, based on predicted consensus sequences and intergenic positions (Liachko et al., 2010). In higher organisms, very deep sequence analysis has identified many potential origins and derived a predicted origin sequence and structure, but the requirements of this sequence/structure remain to be demonstrated (Besnard et al., 2012). Here, origins of replication were identified

using a combination of ORC binding site identification and algorithms for nucleosome occupancy trained on conserved nucleosome positioning and ORC occupancy at known *K. lactis* origins. Non-dentauring 2D gels identified bubble arc intermediates at three out of five *pro*ORIs tested, and did not detect them at ORC binding sites that lacked the nucleosome depletion pattern. This indicates that a specific nucleosome depletion/occupancy pattern is a conserved feature of origins of replication and that identification of this pattern provides predictive power. A similar nucleosome depletion pattern was also seen in *Drosophila* ORC binding sites (Eaton et al., 2010b). Not all tested *pro*ORIs formed detectable bubble arcs in DNA 2D gels, but based on studies in other organisms, it remains possible that at least some of these *pro*ORIs may be late/inefficient or dormant origins.

The locations of isolated ARSs and of origins are strongly biased towards intergenic regions in *S. cerevisiae* and *K. lactis* (Brewer, 1994; Liachko et al., 2010), which is thought to separate the processes of replication initiation and transcription activation and thereby avoid collisions between the two large complexes that must act on the same DNA. Yet 17% of the *C. albicans* predicted origins were located within ORFs. I presume that many of these are dormant origins. Interestingly, Shor *et al.* found that ORC can bind within highly transcribed ORFs that are enriched for genes that encode proteins involved in central metabolism in *S. cerevisiae* (Shor et al., 2009). This is the case in *C. albicans* as well (Figure 4.14A): ORC tends to bind to highly transcribed ORFs that encode genes involved in metabolism and ORFs adjacent to *pro*ORIs have

higher levels of RNA expression than do random genes (Figure 4.14B). Highly transcribed genes usually have well-positioned nucleosomes with an NFR just 5' of the transcription start site (Rando and Winston, 2012). This pattern is one of several found at *S. cerevisiae* origins (Berbenetz et al., 2010) and because the *pro*ORI prediction algorithm is based on nucleosome depletion patterns that resemble these highly transcribed promoter regions, the *pro*ORI predictions may be biased for transcription start sites as well. Notably, transcription start sites also are hotspots for origin initiation in metazoans. However, in *C. albicans,* as in *S. cerevisiae,* the origin nucleosome positioning pattern at *pro*ORIs is distinct from the pattern at canonical transcription start sites, with the width of NFRs at origins being narrower than that at TSSs (Berbenetz et al., 2010).

Among the well-characterized fungal species *S. cerevisiae*, *K. lactis*, and *S. pombe*, replication origin sequences differ, and therefore replication origins are considered to be evolutionarily dynamic. Consistent with this, origin locations from ARS screens are not syntenic, even though they remain intergenic (Liachko et al., 2010). In contrast, origins nearest to centromeres remain syntenic (Di Rienzi et al., 2012). Origins at or surrounding centromeres differ from other chromosomal origins in that they fire early and/or with high efficiency. For example, in *Saccharomyces* clades, pericentric origins are activated earlier than many other origins on the chromosome arms (Di Rienzi et al., 2012; Pohl et al., 2012). In *C. albicans*, this property of centromeres is even more dramatic; the earliest firing origins are located at the regional centromeres and their function is linked to their epigenetic inheritance (Koren et al., 2010b). All eight of the

centromeric origins were also identified as *pro*ORIs, but I excluded them from further analyses because of their unique properties. The factors that distinguish centromeric origins from chromosomal arm origins remain to be determined.

In general, organisms with epigenetically-inherited regional centromeres have been thought to have epigenetically inherited origin features as well, and to have structural features (e.g., AT-rich regions in *S. pombe* or GC-rich structures in mammals) rather than consensus sequence motifs. A surprising result from this study is that *C. albicans* origins, like those *Saccharomycetaceae* clades, have a primary sequence motif that is necessary for ARS function on the plasmid. Furthermore, I found that a combination of these sequences and a specific nucleosome depletion pattern is necessary for efficient origin activity within the chromosomal context. Thus, despite the evolutionarily dynamic nature of primary sequence motifs at origins, conserved properties of replication origins have been retained over evolutionary time scales.

Given prior difficulties with ARS assays in *C. albicans*, I did not initiate this study with a screen for ARS activity. Rather, only once I found that *pro*ORI DNA fragments could confer origin activity *in vivo* did I test for the ability of genomic DNA to function using newly developed linear plasmids. Linear plasmids may be useful than circular plasmids because single crossovers result in integration of circular plasmids, while if linear plasmids recombine with chromosomal DNA via a single cross over, it would generate acentric fragments that are likely to be lost during mitosis. I assume that strains in which chromosome fragments are lost due to such recombination events would be less fit and thus would be lost from

the population. This would reduce the likelihood that non-ARS linear plasmids fragments would be isolated and would facilitate the identification of sequences that promote autonomous plasmid replication. However, linear plasmids do integrate to some degree, which is why we did the ARS screen by two rounds of CHEF gel purification of linear plasmids to separate them from any integrated plasmids.

Interestingly, tRNA loci were dramatically enriched in the origin predictions; a weaker association between tRNA loci and origins has been observed in *S. cerevisiae* and *K. lactis*. However, at least in ORI1046 and ORI410, the tRNA gene within the ORI was not required for function in the mini-ARS assay (Figure 4d). I suggest that tRNA genes may affect origin firing much like actively transcribed ORFs, given that tRNAs are generally transcribed at high levels. Nonetheless, I cannot rule out the possibility that tRNA are associated with replication origins because of the specific chromatin architecture driven by RNA Pol III transcriptional machinery, including the presence of a well-positioned nucleosome (Iwasaki and Noma, 2012; Noma et al., 2006).

Comparison of the ARSs isolated in our ARS screen with the list of significant ORC binding sites detected by ChIP-chip did not detect a significant level of overlap between these regions. I posit that some ARSs may have the ability to recruit ORC in the context of the plasmid even if they don't do so in the context of their genomic position. Thus, isolated ARSs can direct origin firing on a plasmid but are not sufficient to function in the genome, perhaps because other protein binding and/or positioned nucleosome depletion patterns are necessary

to organize the complex constellation of events that lead to origin firing. This could also be because genomic origins must fire efficiently or else they will be passively replicated while plasmid origins may be less efficient, or fire later, and still be considered functional.

Nonetheless, the unbiased plasmid-based ARS screen enabled the identification of a consensus motif similar to the *pro*ORI and ORC binding site consensus motifs and this motif is required for origin function measured by the ability to direct transformation of linear plasmids and by the ability to recruit the MCM complex, an indicator of successful pre-RC assembly. All three of the AC-rich motifs (*Ca*ACS, *pro*ORIs and ORC binding sites) align well, although the relative weight of the nucleotides differs. In addition, a T-rich sequence is found ~ 100 bp 3' to the CaACS and is often at the center of the region of nucleosome depletion. I propose that the *Ca*ACS may intrinsically assist with accurate nucleosome positioning and thereby may promote a specific nucleosome depletion pattern that facilitates origin firing. While the *Ca*ACS motif does not encode a sequence predicted to promote nucleosome exclusion signal (data not shown), I posit that collaboration between the *Ca*ACS and downstream T-rich motif may create a positioned nucleosome free region, that may facilitate pre-RC complex assembly within the nucleosome depletion region, much like the B-element at *S. cerevisiae* origins. This would then facilitate origin firing in this region.

Despite the 30-35 % precision recall of the *K. lactis* nucleosome pattern algorithm, based on tests *in K. lactis*, the predictive power in *C. albicans* may be

better than the initial computational estimation. Indeed three of five *pro*ORIs had bubble arc intermediates and over 50 % of *pro*ORIs contain the CA-rich ACS motif. It is also possible that the *Ca*ACS in *C. albicans*, compared to the one in *S. cerevisiae* or *K. lactis*, is more prevalent in the genome and thus the *trans*-acting factors (including nucleosome positioning) for origin identity in *C. albicans* have become relatively more important during evolution.

The nucleotide composition of the *Ca*ACS is similar to that of several binding motifs, including those for the Ace2 and Fkh2 transcription factors (data not shown). The *Ca*ACS motif based on the ARS screen is presumably independent from the genome context of the sequences. Thus, our previous concern that *pro*ORI predictions based on nucleosome-ORC associated patterns might drive prediction bias to intergenes or promoters was unnecessary because the existence of transcription factor-like *Ca*ACS from the ARS screen, conducted in parallel, suggests that origins prefer to be intergenric. The similarity between *Ca*ACS motifs and transcriptional factor binding motifs suggests that unexpected functional connections may exist between origin initiation and a specific subset of transcription factors. Intriguingly, in *S. cerevisiae,* Fkh1 and Fkh2 (encoding Forkhead transcription factors) appear to physically cluster early-firing origins spatially within the nucleus, thereby regulating the temporal program of replication initiation in *trans* (Knott et al., 2012). Perhaps in *C. albicans*, there is a relationship between origins and Fkh2, the single Forkhead factor (Bensen et al., 2002), as well. Similarly, perhaps local chromatin structure affects origin-proximal gene expression as well as the timing of replication origin firing during the cell

cycle as it does in *S. cerevisiae* (Knott et al., 2009; Stevenson and Gottschling, 1999; Vogelauer et al., 2002).

In summary, I developed a bioinformatics approach that identified *bona fide* origins by leveraging knowledge about origins in *S. cerevisiae* and *K. lactis*, and then I adapted experimental approaches to comprehensively map the positions of replication origins in *C. albicans*. Importantly, I discovered that regional centromeres, which themselves specify highly efficient, epigenetically inherited (and apparently sequence-independent) origin function, can co-exist with origins on the chromosome arms that utilize primary sequence motifs for replication initiation. The development of linear ARS plasmids provides a useful new tool that, together with the newly discovered haploid *C. albicans* isolates (Hickman et al., 2013), should facilitate the production of libraries for more conventional molecular genetic studies. Finally, my work demonstrates that *C. albicans* is an attractive model organism for studying genome organization. I suggest that insights gained from this work may have implications for higher eukaryotes. It will be especially interesting to determine if origin features discovered here in *C. albicans* will provide new insights into the elusive nature of replication origins in metazoans.

Figure 4.1. Genome-wide map of ORC binding sites

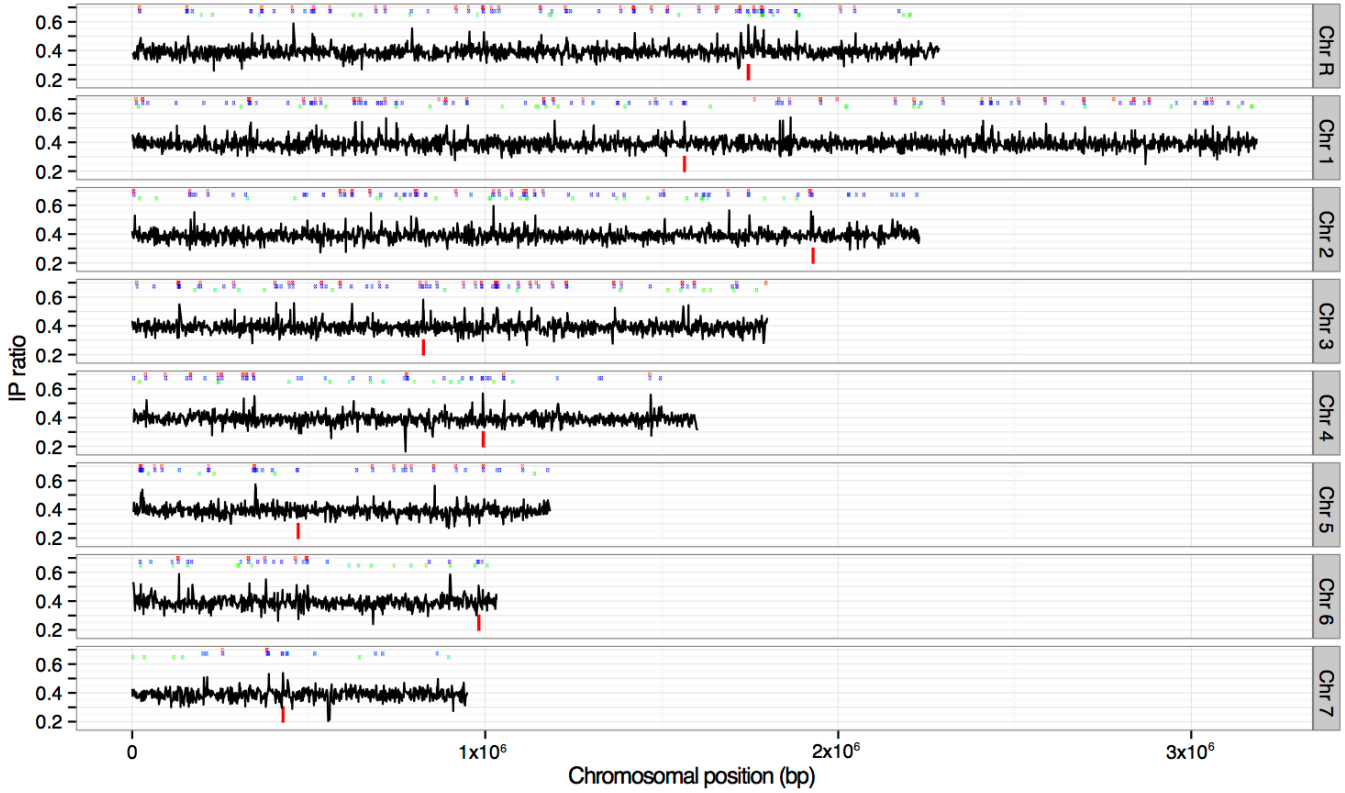


Figure 4.1. Continued.

ORC binding sites were identified by ChIP-on-Chip. Each peak mapped along the length of the 8 *C. albicans* chromosomes represents an ORC binding site. Blue dots denote peaks of the 393 significant ORC binding sites and listed in Table 4.1. Red dots denote predicted origin (*pro*ORI) positions also listed in Table 4.2. Green dots indicate locations of isolated ARS fragments and are listed in Table 4.3. Red vertical lines indicate the positions of centromeres.

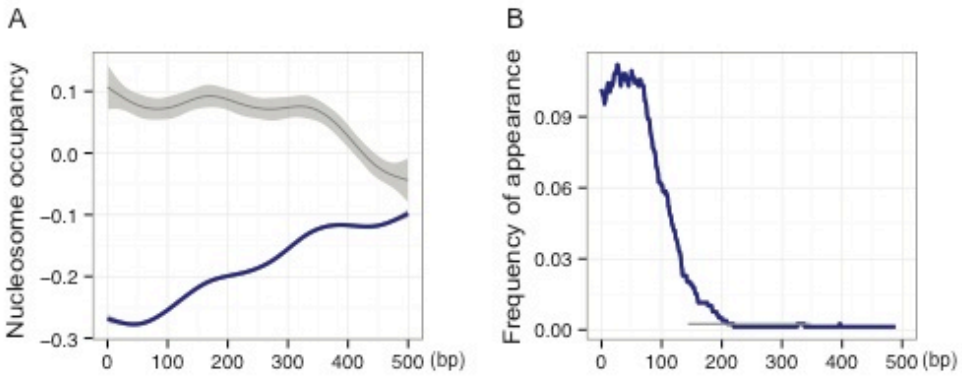


Figure 4.2. ORC binding features

(A) ORC binding sites are associated with regions of low nucleosome occupancy. Significant ORC binding sites were centered. ORC binding sites were analyzed for the likelihood of nucleosome occupancy. The average nucleosome occupancy weighted score is plotted on the Y-axis as a function of distance from the center of the significant ORC binding sites on the X-axis (Blue line). Alignment of random genomic sites (gray line) is also illustrated.
(B) tRNA loci are associated with ORC binding sites. 114 of the 126 *C. albicans* tRNAs are associated with ORC binding sites, with most ORC sites located within 500 bp of a tRNA.

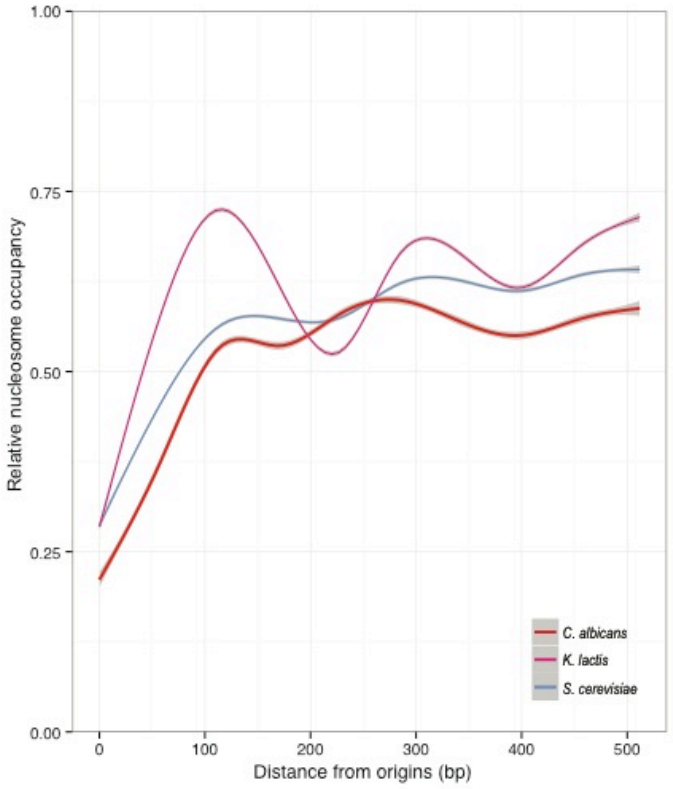
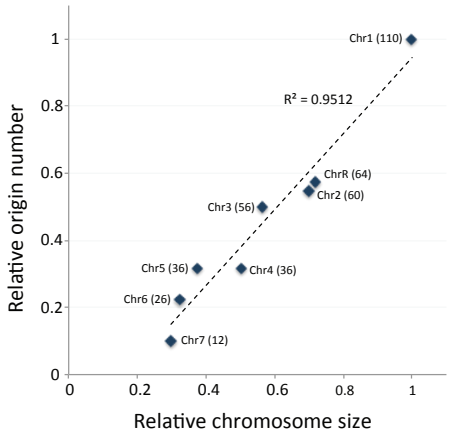


Figure 4.3. Nucleosome occupancy patterns at origins of replication in different yeast species.

Average nucleosome occupancy (Y-axis) as a function of distance from origins of replication (X-axis) in *S. cerevisiae* (light blue) and *K. lactis* (magenta) reveal a similar nucleosome pattern. Nucleosome occupancy at the positions of proposed origins (*pro*ORIs) in *C. albicans* (red) predicted based on a model trained on the *K. lactis* nucleosome occupancy pattern combined with ORC binding sites.

A



B

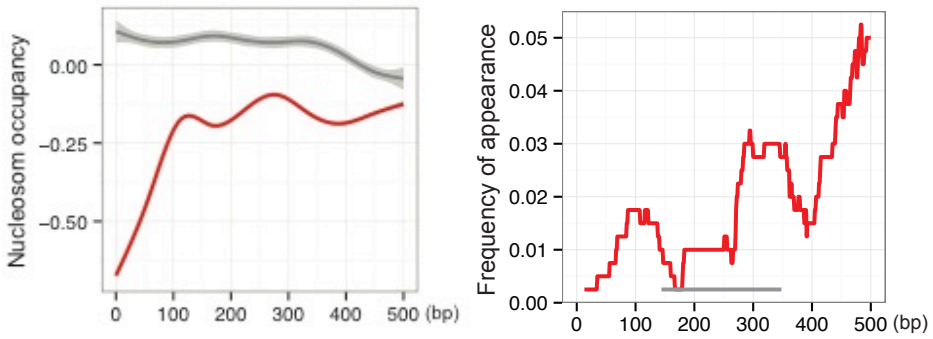


Figure 4.4. Genomic feature of *pro*ORIs

Figure 4.4. Continued.

(A) Total 386 origins were predicted, and the origin number of each chromosome is labeled on the plot. Chromosome size (X-axis, relative to the largest chromosome, Chr1) versus relative origin number on each chromosome (Y-axis, relative to chromosome 1) is roughly proportional. The average distance between *pro*ORIs is 34.9 kb. (B) *pro*ORIs are located at NFRs, *pro*ORI positions are centered (X-axis), and nucleosome occupancy at *pro*ORIs (Y-axis) are low than average ORC binding sites.

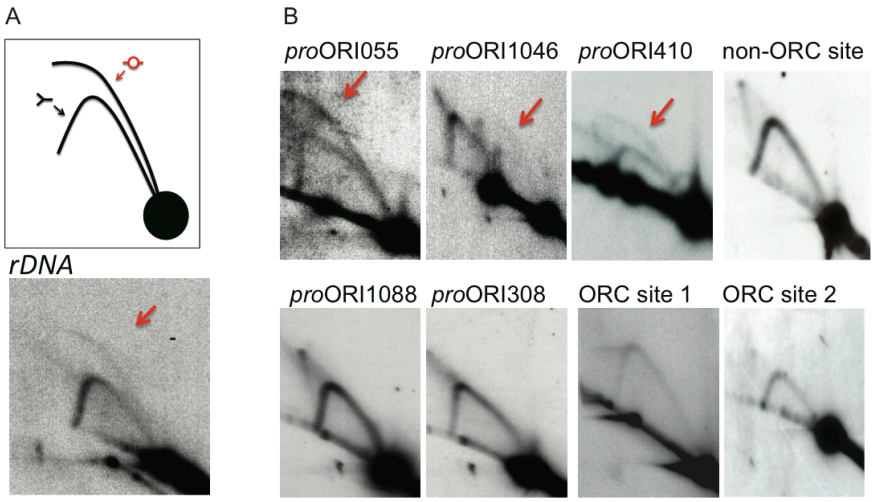


Figure 4.5. Detection of bubble arc replication intermediates within some *pro*ORIs.

(A) DNA 2D gels detect the migration pattern of replicating intermediates. A bubble arc (red) is indicative of a DNA fragment containing an active replication initiation site (upper panel), while a Y-arc is indicative of passive replication from an origin outside the fragment being analyzed. The rDNA repeats include a conserved active origin that fires within a subset of the repeats to generate a bubble arc (bottom panel, red arrow). (B) Three *pro*ORIs form bubble arcs and thus are *bona fide* origins of replication (red arrows), while two did not (lower left two panels). Neither DNA sequences that without an ORC binding site (non-ORC, top right), nor ORC binding sites that are not *pro*ORIs (ORC site, bottom right two panels) produced detectable bubble arcs.

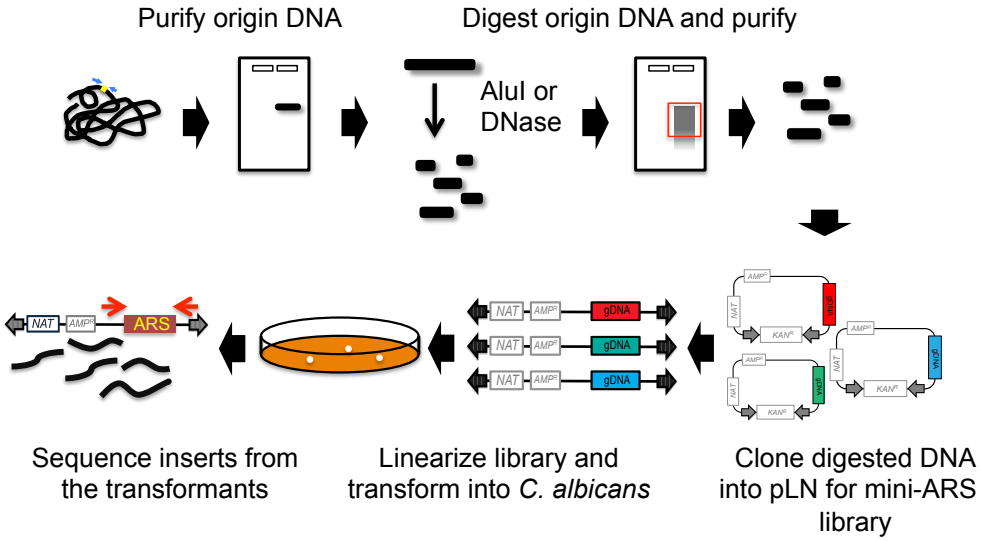


Figure 4.6 Strategy for 'mini-ARS screen'

Origin DNA from *pro*ORI1046 (1.2 kb), *pro*ORI410 (1.2 kb), and proORI055 (3.2 kb) was digested and subcloned into pLN plasmids and then digested or sheared into random size of DNA fragments. DNA fragments around 0.1-1 kb were purified from the agarose gel to be subcloned into pLN plasmids. The libraries were linearized and transformed into *C. albicans*. Then, individual clones which yielded high transformation efficiency, were analyzed by Sanger sequencing.

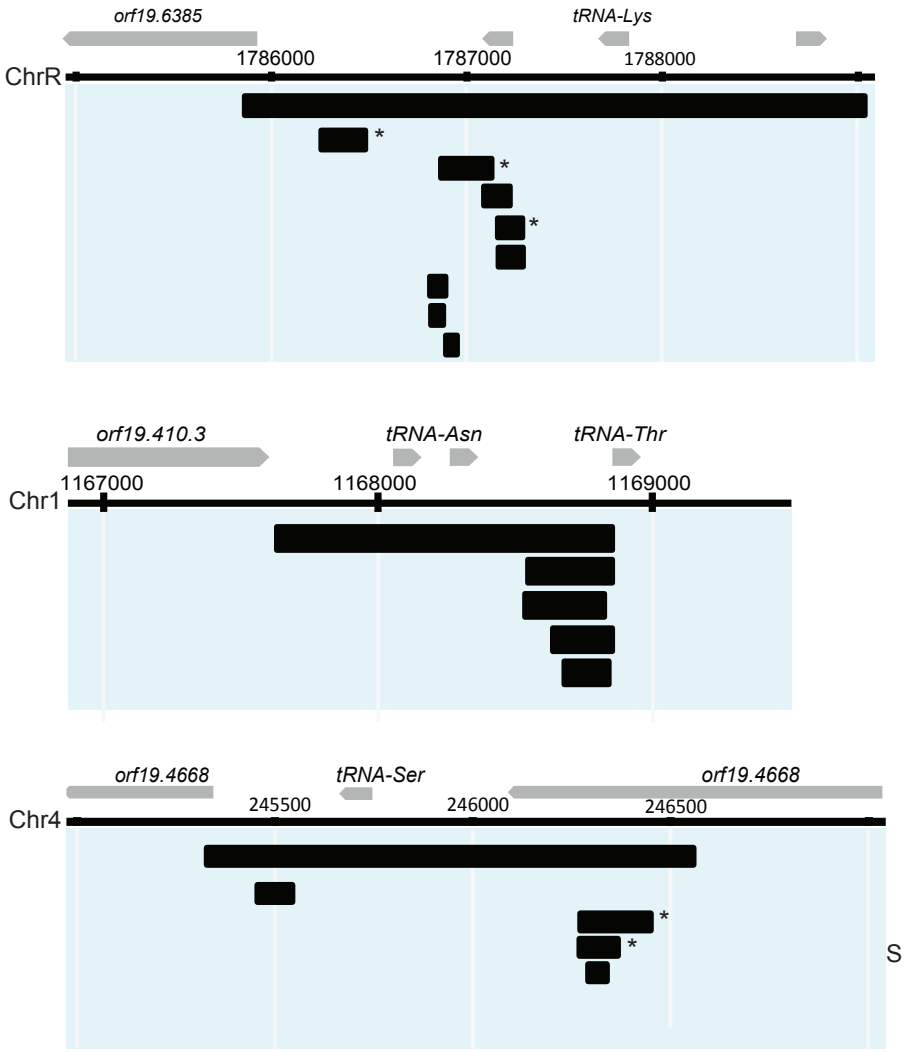


Figure 4.7. Minimal DNA regions required for ARS function

Figure 4.7 continued

Origin DNA from 3.2 bk *pro*ORI055 (top), 1.2 kb *pro*ORI1046 (middle), and 1.2 kb *pro*ORI410 (bottom) was analyzed in mini-ARS screen and identified minimal ARS fragments. Each black bar represents an isolated DNA fragment that yielded relatively high transformation efficiency. Sub-fragments containing ARS function were isolated from each *pro*ORI DNA and verified for function. Stars (*) mark the DNA fragments, which were also isolated from genome-wide ARS screen. Sub-ARS fragment “s” (bottom panel), a 97 bp DNA from *pro*ORI410 was tested for ARS function and used in later experiments.

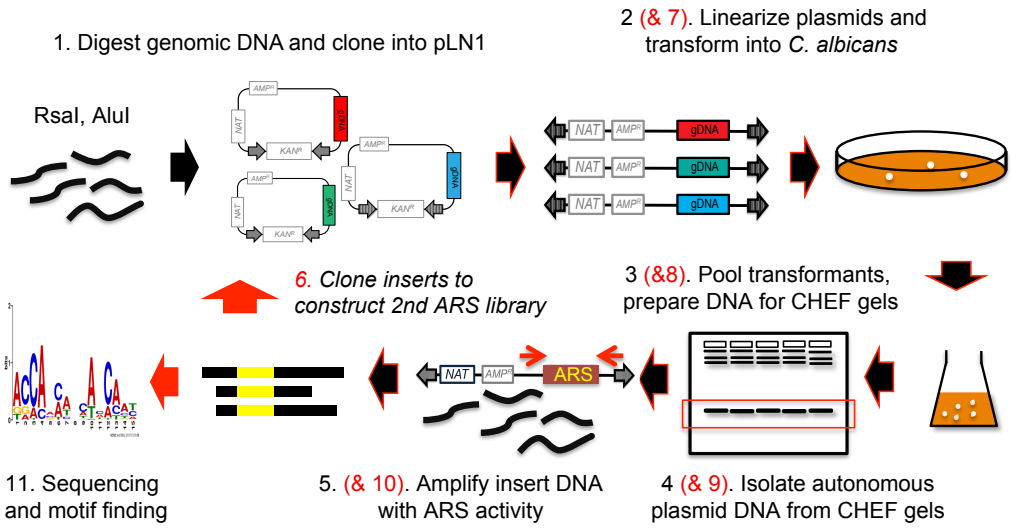


Figure 4.8. Genome-wide ARS screen

Strategy for two-stage enriched genome-wide ARS screen, in which total genomic DNA was digested, cloned into pLN1, transformed into *C. albicans* and then DNA inserts were isolated from transformants. After step 5, a second round of screening (ARS enrichment step, red arrow) was performed (steps 6-9). Ultimately, inserts from the secondary screen were analyzed by Illuminna sequencing and then analyzed for common motifs.

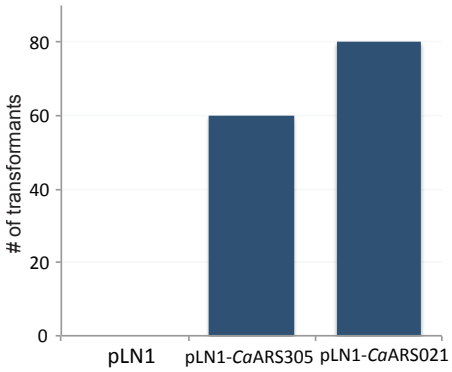


Figure 4.9. Isolated ARSs can yield high transformation efficiency.

Two ARSs originally identified in the ARS enrichment screen (ARS305 and ARS201) were isolated from genomic DNA and cloned into pLN1 at the *Pvu*II site. Both can yielded high transformation efficiency,

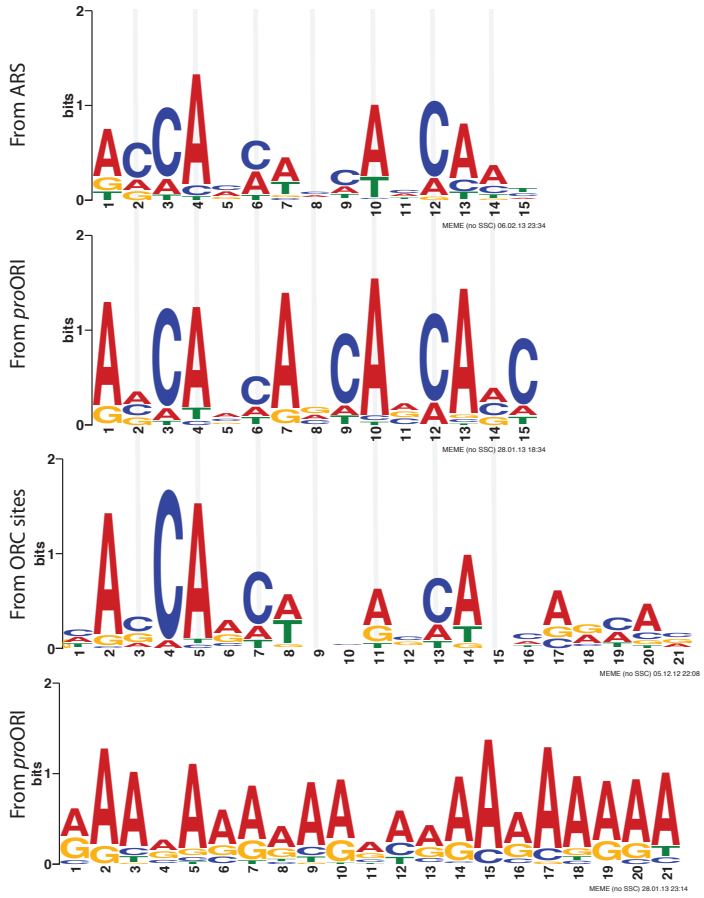


Figure 4.10. A 15 bp *Ca*ACS motif identified from 127 ARSs

The 15 bp *Ca*ACS motif identified from 127 isolated ARSs (top). A similar AC-rich motif was identified from 386 *pro*ORIs (middle), and another related 21 bp motif was identified from the 393 significant ORC binding sites (middle). Furthermore a poly A/T sequence was also identified from *pro*ORIs (bottom).

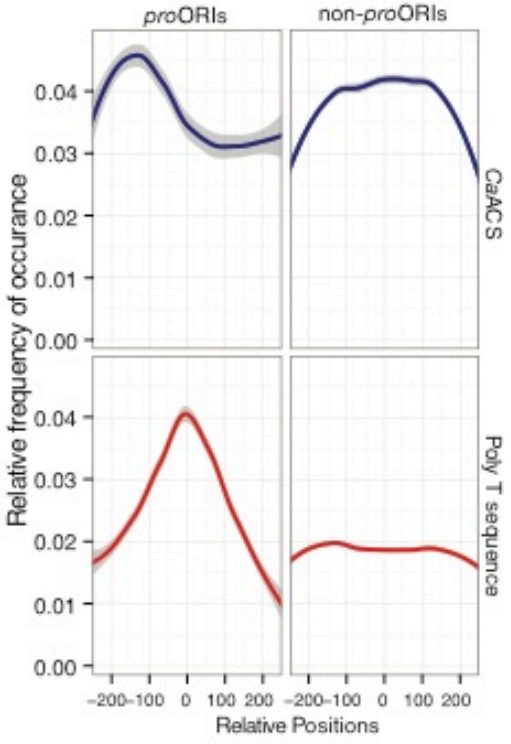


Figure 4.11. Distribution between *Ca*ACS motif and poly A/T sequences

Relative locations between *Ca*ACS motifs and poly A/T sequences at *pro*ORI regions were analyzed, compared to both at non-*pro*ORIs. The CaACS motifs, which are AC-rich, (top) are found to the left of the center of *pro*ORI sequences and a ploy T tract (bottom) is found centered on the *pro*ORIs.

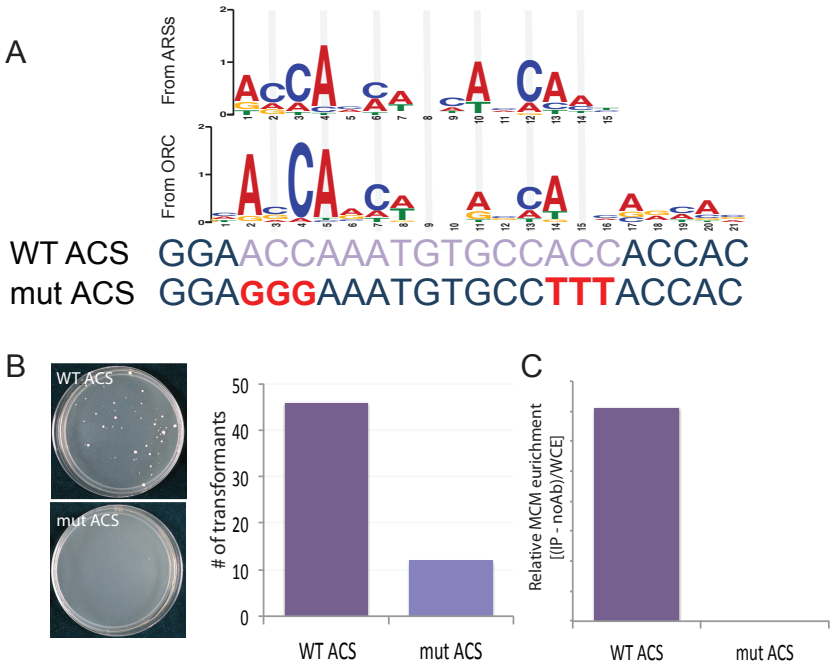


Figure 4.12. Mutagenesis of the *Ca*ACS abolishes origin activity

(A) The 15 bp *Ca*ACS motif (top row), as found in *pro*ORI410, within fragment *ORI410-S* in Figure. 4.7 (middle row), was engineered using site-directed mutagenesis to produce mut ACS (bottom row). (b) Transformation with pLN-*pro*ORI410S carrying the intact *Ca*ACS (WT) yielded many transformants thatn transformation with pLN-proORI410S carrying the mutant *Ca*ACS (mutACS). (c) Relative enrichment of MCM binding was detected with the WTACS plasmid by amplification of the plasmid-specific *NAT* gene locus, while no enrichment of MCM binding was evident with the mutant plasmid.

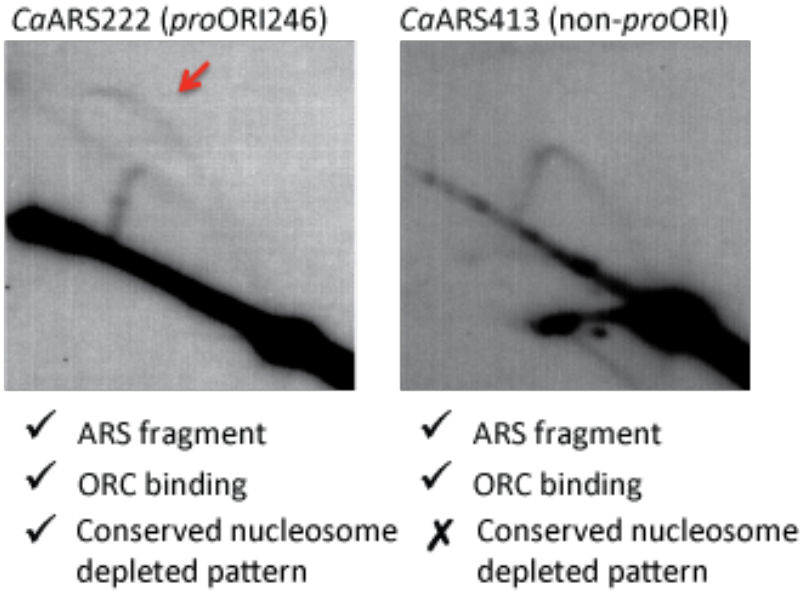
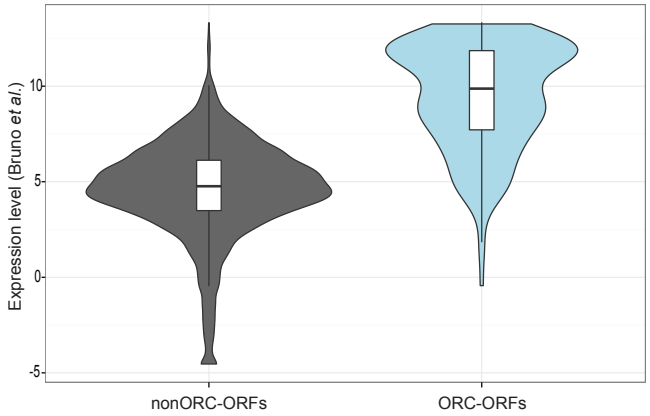


Figure 4.13. *Ca*ACS is essential but not sufficient to fire an origin in the chromosomal context.

Two ARSs isolated from the ARS screen were analyzed by DNA 2D gel for the presence of active origins at their native genomic loci. *Ca*ARS222 (left) was also predicted to be an origin (*pro*ORI246); it produces a bubble arc indicating that it is, indeed, an active origin in the genome. *Ca*ARS413 (right) has a *Ca*ARS motif, but lacks the conserved nucleosome pattern; it produces a Y arc and no bubble arc, indicating that it is replicated passively by adjacent origins.

A



B

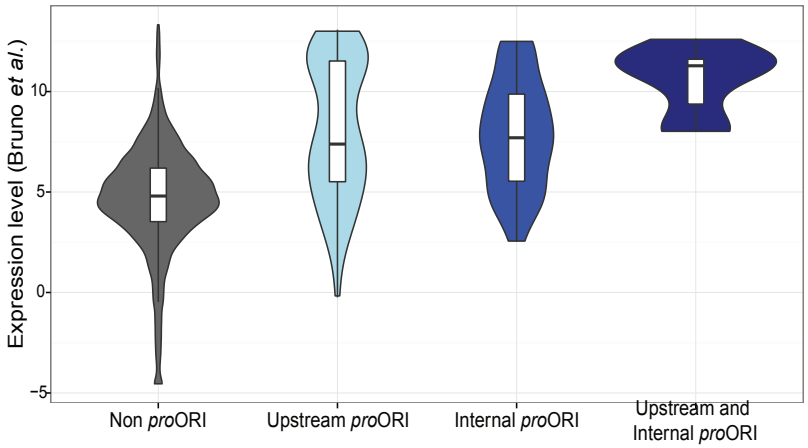


Figure 4.14. ORC/origins are associated with transcribed ORFs

Figure 4.14. Continued.

(A) Approximately half of all ORC binding sites overlapped with ORFs and the expression level of the ORFs associated with ORC are significant higher than the expression level of ORFs not associated with ORC (p-value < 2.2e-16).

(b) Most *pro*ORIs were found in intergenic regions, but some are associated with ORFs. These *pro*ORI-associated ORFs have higher expression levels relative to ORFs that are not associated with a *pro*ORI, whether they contain *pro*ORI sequences upstream, internally, or both upwtream and internally (p-value < 3.04e-14, 2.78e-16, and 1.17e-8, respectively).

Table 4.1. Genome-wide ORC binding sites

Chr	PeakPos	PeakHeight	minPos	maxPos	GenomePos	tRNA
0	22905	1.948711	22012	24405	ORF	NA
0	24261	2.030309	22761	24630	ORF	NA
0	156542	2.135593	155042	158042	ORF	NA
0	157084	1.904776	155584	158584	ORF	NA
0	157687	1.824216	156187	159187	ORF	NA
0	158465	2.347626	156965	159965	ORF	NA
0	172515	2.863247	172150	172849	Intergene	1
0	338047	2.218806	337722	338321	Intergene	NA
0	339629	2.779601	339271	340011	Intergene	1
0	368830	2.311714	368477	370330	ORF	NA
0	369469	2.455653	368477	370969	ORF	NA
0	370292	2.599009	368792	371067	ORF	NA
0	403784	2.47759	403412	404404	Intergene	NA
0	456611	4.163318	456106	457878	Intergene	1
0	491120	1.810473	490621	492620	ORF	NA
0	510726	2.62192	510322	511684	ORF	NA
0	511314	1.905386	510322	511684	ORF	NA
0	515785	2.813639	515336	516105	Intergene	1
0	517096	2.439887	516608	517527	ORF	NA
0	562667	2.091731	562069	564167	ORF	NA
0	563326	2.009138	562069	564826	ORF	NA
0	692974	2.431265	692436	693486	ORF	NA
0	718703	2.393326	718293	719105	Intergene	NA
0	792616	3.510852	791538	794116	Intergene	NA
0	840296	2.311986	839673	840645	ORF	NA
0	920451	2.342193	920195	920982	ORF	NA
0	955426	3.114612	955035	956256	Intergene	NA
0	995993	3.010657	995230	997095	ORF	NA
0	996733	2.936432	995233	997095	ORF	NA
0	1020194	3.64169	1019621	1020680	Intergene	2
0	1029275	1.83653	1027775	1030775	ORF	NA
0	1041690	2.620618	1040951	1042032	ORF	NA
0	1158913	3.640646	1158357	1159465	Intergene	2
0	1190920	3.046675	1190453	1191349	Intergene	2
0	1230200	2.476248	1229201	1231062	Intergene	1
0	1233306	2.605519	1232933	1233754	Intergene	1
0	1243827	2.137722	1243519	1244201	Intergene	1
0	1324372	1.887549	1323800	1324873	Intergene	NA

Table 4.1. Continued.

Chr	PeakPos	PeakHeight	minPos	maxPos	GenomePos	tRNA
0	1386725	1.930979	1386319	1387864	ORF	NA
0	1422419	3.51322	1421967	1423919	Intergene	NA
0	1423442	2.937513	1421967	1424942	Intergene	NA
0	1424412	3.298655	1422912	1425081	Intergene	NA
0	1474229	2.558347	1473821	1474625	Intergene	1
0	1517280	2.059196	1515780	1518780	ORF	NA
0	1519556	2.022115	1518056	1520783	ORF	NA
0	1608170	2.974654	1607753	1608573	Intergene	1
0	1657267	2.66495	1656592	1657836	ORF	NA
0	1713716	2.559364	1713333	1714047	Intergene	1
0	1723239	2.452483	1722612	1724739	Intergene	NA
0	1724479	2.867692	1722979	1725826	ORF	NA
0	1745202	3.907455	1743767	1746702	Intergene	NA
0	1746029	3.003856	1744529	1747529	Intergene	NA
0	1763614	3.476461	1762694	1765109	ORF	NA
0	1764163	3.68721	1762694	1765109	ORF	NA
0	1772799	1.938505	1772287	1773148	Intergene	NA
0	1783936	2.111167	1783416	1785436	ORF	NA
0	1784736	2.327523	1783416	1785850	ORF	NA
0	1787871	2.548975	1786843	1788199	Intergene	1
0	1788823	3.165327	1788450	1789425	Intergene	1
0	1808809	2.206386	1807881	1809276	Intergene	NA
0	1813396	3.3196	1812951	1813875	Intergene	1
0	1831340	1.84409	1830345	1831704	ORF	NA
0	1881144	2.260764	1880259	1881560	Intergene	1
0	1882709	3.28186	1882302	1883275	Intergene	1
0	2009231	3.31666	2008882	2009713	Intergene	1
0	2049271	2.733352	2048929	2049721	Intergene	1
0	2173966	1.798064	2172611	2174667	ORF	NA
1	14599	2.427768	13099	15201	ORF	NA
1	32506	3.164686	31928	32981	Intergene	NA
1	46205	2.108091	45907	46556	Intergene	NA
1	126667	3.979386	126336	127047	Intergene	1
1	204974	3.344459	204516	205313	Intergene	1
1	269449	2.031254	269082	269924	ORF	NA
1	289674	2.700044	289181	290486	Intergene	NA
1	333122	2.485172	332712	334149	ORF	NA
1	333809	2.685477	332712	334149	ORF	NA

Table 4.1. Continued.

Chr	PeakPos	PeakHeight	minPos	maxPos	GenomePos	tRNA
1	336252	3.623867	335594	336756	Intergene	1
1	354339	2.503487	354012	354660	Intergene	1
1	443053	1.807489	442201	444322	ORF	NA
1	488663	2.01033	487163	488999	Intergene	1
1	508471	1.912666	506971	509971	ORF	NA
1	509187	1.956645	507687	510687	ORF	NA
1	510256	1.895777	508756	511756	ORF	NA
1	516221	1.82423	515564	517417	ORF	NA
1	521896	2.940919	521114	522248	ORF	NA
1	536926	2.596398	536484	537298	Intergene	1
1	631111	3.151817	630375	631624	ORF	NA
1	632816	3.130966	632307	633317	ORF	NA
1	643901	3.501125	643307	644361	ORF	NA
1	651034	3.260234	650267	651481	Intergene	NA
1	662731	2.191373	661944	664231	ORF	NA
1	663437	2.446447	661944	664500	ORF	NA
1	697484	1.807664	697236	698984	Intergene	NA
1	698181	2.500148	697236	699226	ORF	NA
1	707694	2.005912	706437	708666	ORF	NA
1	708367	1.848024	706867	708666	ORF	NA
1	719326	3.813631	718892	719640	Intergene	1
1	748385	2.988246	748103	749009	Intergene	1
1	751954	2.171628	751421	752467	ORF	NA
1	766783	2.927186	766325	767175	Intergene	1
1	886935	2.079566	885989	888097	ORF	NA
1	891264	2.814169	890352	892764	ORF	NA
1	911966	2.002098	911248	913466	ORF	NA
1	950421	2.35786	949541	951702	ORF	NA
1	950971	2.818695	949541	951702	ORF	NA
1	1116133	2.6412	1115750	1116575	Intergene	1
1	1168228	3.556744	1167221	1169299	Intergene	2
1	1168892	1.959065	1167392	1169299	Intergene	1
1	1195805	4.068042	1195156	1197232	ORF	NA
1	1246322	2.293212	1245932	1247822	ORF	NA
1	1278927	1.912637	1278555	1280427	ORF	NA
1	1307156	2.638363	1305656	1308656	ORF	NA
1	1376186	1.86876	1375577	1376623	ORF	NA

Table 4.1. Continued.

Chr	PeakPos	PeakHeight	minPos	maxPos	GenomePos	tRNA
1	1378424	2.329692	1377979	1378904	Intergene	NA
1	1409547	2.472853	1409122	1410189	Intergene	NA
1	1468756	2.68305	1468174	1469215	ORF	NA
1	1484110	2.969596	1483685	1484504	Intergene	1
1	1489867	1.970534	1489378	1490230	Intergene	1
1	1528409	1.845097	1527433	1529314	ORF	NA
1	1564200	3.587629	1563414	1565700	Intergene	NA
1	1565097	2.483463	1563597	1566377	Intergene	NA
1	1565678	2.443537	1564178	1566377	Intergene	NA
1	1570040	2.420291	1569445	1570458	Intergene	1
1	1827805	2.15344	1826770	1828948	ORF	NA
1	1835070	3.318947	1834031	1835932	ORF	NA
1	1843678	3.465104	1842987	1844241	ORF	NA
1	1864907	4.075872	1863831	1865681	ORF	NA
1	1952927	3.139985	1952344	1953568	Intergene	NA
1	2164393	2.350185	2164076	2164794	Intergene	1
1	2217947	2.146956	2217700	2218350	Intergene	NA
1	2237179	2.330448	2236368	2237842	ORF	NA
1	2300102	2.673833	2299692	2300547	ORF	NA
1	2407217	2.452213	2406813	2407617	ORF	NA
1	2410508	3.284997	2410088	2411010	Intergene	1
1	2433367	2.257862	2432508	2433820	ORF	NA
1	2434481	2.308522	2433967	2434852	Intergene	NA
1	2436142	2.874213	2435594	2436979	ORF	NA
1	2454177	1.894145	2453728	2454605	ORF	NA
1	2492722	2.650807	2492271	2493243	ORF	NA
1	2513843	1.85749	2513522	2514418	ORF	NA
1	2588803	3.454487	2588297	2589712	Intergene	2
1	2637669	1.849459	2636169	2638814	ORF	NA
1	2698686	2.372837	2698132	2699119	ORF	NA
1	2800899	1.865813	2800301	2801353	Intergene	NA
1	2831276	1.997422	2830815	2831716	Intergene	NA
1	2840591	2.579638	2840049	2841377	ORF	NA
1	2869576	2.231121	2868834	2869900	ORF	NA
1	2883707	2.72213	2883230	2884245	Intergene	NA
1	2960433	1.971645	2960086	2960843	Intergene	NA
1	3012923	2.243576	3012049	3013535	ORF	NA

Table 4.1. Continued.

Chr	PeakPos	PeakHeight	minPos	maxPos	GenomePos	tRNA
1	3039535	1.95095	3038862	3040239	ORF	NA
1	3046609	2.252357	3045885	3048078	ORF	NA
1	3047362	2.339961	3045885	3048078	ORF	NA
1	3061195	3.317998	3060480	3061701	Intergene	1
1	3106962	2.648637	3106446	3107819	ORF	NA
1	3146835	2.20618	3146282	3147435	ORF	NA
2	7508	2.886677	6382	8737	Intergene	1
2	165204	2.633747	164603	166419	ORF	NA
2	166125	2.718398	164625	166419	Intergene	NA
2	176039	3.364998	175291	176550	Intergene	1
2	182722	2.002142	181222	182985	ORF	NA
2	217081	2.362028	216434	217487	Intergene	1
2	287913	2.013841	287579	288539	ORF	NA
2	323533	3.016337	323190	323840	Intergene	1
2	488863	1.925168	487742	490363	ORF	NA
2	490630	1.903365	489130	491161	ORF	NA
2	496186	2.271328	495871	496954	Intergene	1
2	543780	1.846552	542607	545241	ORF	NA
2	544295	1.921622	542795	545241	ORF	NA
2	592499	3.199532	591240	592961	ORF	NA
2	603500	2.044765	603191	604778	Intergene	NA
2	604408	2.488341	603191	604778	ORF	NA
2	624238	2.091345	623748	625738	ORF	NA
2	624757	3.259296	623748	626257	ORF	NA
2	625820	3.781792	624320	626622	ORF	NA
2	676436	3.278419	675959	676994	ORF	NA
2	706486	2.055529	704986	707256	ORF	NA
2	749981	2.05174	749633	751481	ORF	NA
2	771841	2.303483	771347	772132	Intergene	1
2	774620	2.638648	773955	774996	Intergene	NA
2	790304	2.902761	789833	790691	Intergene	2
2	804607	2.510368	804294	805176	ORF	NA
2	806540	2.503463	805736	806810	ORF	NA
2	810425	2.966926	809975	810855	Intergene	2
2	832315	2.807277	831904	832683	ORF	NA
2	834226	1.946648	833851	834613	ORF	NA
2	920812	1.900271	920516	921213	Intergene	NA
2	961541	2.036034	960942	961957	ORF	NA

Table 4.1. Continued.

Chr	PeakPos	PeakHeight	minPos	maxPos	GenomePos	tRNA
2	1023758	4.720856	1022516	1024755	ORF	1
2	1027283	3.677851	1026892	1027824	Intergene	1
2	1042477	1.827547	1042194	1042814	Intergene	1
2	1077589	2.18416	1076806	1077994	Intergene	1
2	1111057	2.01274	1110757	1111671	ORF	NA
2	1112412	2.446178	1112148	1112760	Intergene	1
2	1119260	2.758677	1118673	1120506	ORF	NA
2	1141991	2.858979	1141359	1142686	ORF	NA
2	1143969	3.290099	1143584	1144260	Intergene	1
2	1167427	2.895058	1166880	1167917	ORF	NA
2	1238010	2.027294	1237577	1238336	Intergene	1
2	1303247	1.812337	1301966	1303896	ORF	NA
2	1321505	2.786388	1320934	1321895	Intergene	1
2	1365953	1.930713	1365273	1366223	ORF	NA
2	1395843	2.998747	1395339	1396379	ORF	NA
2	1498081	1.83975	1497572	1498397	Intergene	NA
2	1508105	2.225219	1507164	1509347	ORF	NA
2	1508776	2.550768	1507276	1509347	ORF	NA
2	1600519	2.929742	1600096	1600828	Intergene	1
2	1620380	2.065102	1620011	1620741	Intergene	1
2	1621876	2.243249	1621406	1623216	Intergene	1
2	1633590	1.872489	1633119	1633980	ORF	NA
2	1638007	1.844864	1637413	1639315	Intergene	NA
2	1690910	4.598481	1689869	1691822	ORF	NA
2	1747816	2.423102	1747027	1749174	ORF	NA
2	1748489	3.290098	1747027	1749174	ORF	NA
2	1888149	2.611698	1887785	1888672	Intergene	1
2	1922756	3.894398	1922025	1923415	ORF	NA
2	1927523	3.202286	1926112	1929023	Intergene	NA
2	1928054	3.141677	1926554	1929313	Intergene	NA
2	2031458	2.041621	2030986	2032564	ORF	NA
2	2032173	1.845039	2030986	2032564	ORF	NA
2	2053817	2.390374	2053382	2054651	ORF	NA
2	2075108	2.131691	2074747	2075483	Intergene	NA
2	2154268	2.344081	2152768	2155768	Intergene	NA
2	2167796	2.107081	2167337	2168755	Intergene	1
2	2186826	1.818962	2186402	2187412	ORF	NA
2	2224329	1.88489	2222937	2225212	ORF	NA

Table 4.1. Continued.

Chr	PeakPos	PeakHeight	minPos	maxPos	GenomePos	tRNA
3	16321	2.149253	15992	17411	Intergene	NA
3	63963	1.829037	63623	64642	Intergene	NA
3	133138	3.259839	132883	134638	ORF	NA
3	133697	3.399832	132883	135197	ORF	NA
3	134432	3.712907	132932	135932	ORF	NA
3	135204	2.868264	133704	136636	ORF	NA
3	135928	2.885895	134428	136636	Intergene	NA
3	190135	2.652735	189445	190434	ORF	NA
3	196116	2.893547	195423	196617	Intergene	1
3	262039	2.166302	261776	262276	Intergene	1
3	290878	3.216511	290126	291166	ORF	NA
3	408143	3.503998	407786	408614	Intergene	1
3	412256	2.172059	411967	412667	Intergene	1
3	423899	3.813318	423493	424426	Intergene	1
3	447748	2.956215	447384	448146	Intergene	1
3	458339	3.449439	457902	459227	ORF	NA
3	520794	2.235753	520575	521003	Intergene	NA
3	539069	3.296816	538198	539575	ORF	NA
3	551091	2.777593	550768	551309	Intergene	1
3	592102	3.699895	591520	592907	Intergene	2
3	622736	3.371721	621834	624096	ORF	NA
3	667032	2.098713	665580	668532	ORF	NA
3	681366	2.082497	680985	682788	Intergene	1
3	703193	2.772697	702895	703578	Intergene	1
3	723435	1.831009	722875	724689	ORF	NA
3	818063	2.056122	817316	818347	ORF	NA
3	824345	3.993299	823138	825845	Intergene	NA
3	825397	2.915026	823897	826645	Intergene	NA
3	835750	2.704936	835481	836200	Intergene	1
3	845828	2.668796	845273	846193	Intergene	1
3	866353	2.117454	866056	866641	Intergene	NA
3	941688	2.845581	940243	942353	ORF	NA
3	974998	3.059802	973987	975684	ORF	NA
3	991955	3.260801	991528	992526	ORF	NA
3	993466	2.343292	993170	993782	Intergene	NA
3	1023381	1.810077	1022901	1024283	ORF	NA
3	1031880	2.688292	1030528	1033380	ORF	NA
3	1033003	2.854411	1031503	1034503	ORF	NA

Table 4.1. Continued.

Chr	PeakPos	PeakHeight	minPos	maxPos	GenomePos	tRNA
3	1034452	2.57391	1032952	1035952	ORF	NA
3	1035650	2.565072	1034150	1036353	ORF	NA
3	1039051	3.046392	1038659	1039456	Intergene	1
3	1068660	1.812427	1068202	1069040	ORF	NA
3	1096893	2.980386	1096295	1097582	ORF	NA
3	1130904	2.836899	1130386	1131584	ORF	NA
3	1153543	2.165252	1153078	1153947	ORF	NA
3	1193794	2.156783	1192294	1194217	ORF	NA
3	1232759	2.617114	1232284	1233990	Intergene	2
3	1233587	2.024567	1232284	1233990	Intergene	NA
3	1367332	2.787887	1366815	1368283	ORF	NA
3	1388232	2.775131	1387318	1389732	ORF	NA
3	1469632	2.166012	1469246	1470527	Intergene	1
3	1561114	3.166455	1560461	1561621	Intergene	1
3	1575710	3.166948	1574912	1576322	ORF	NA
3	1595222	2.606847	1594737	1595571	Intergene	1
3	1702547	2.308408	1702016	1703232	ORF	NA
3	1715064	2.029754	1714540	1716353	Intergene	NA
3	1715951	3.156816	1714540	1716353	Intergene	2
4	6379	1.829247	5459	7275	ORF	NA
4	40559	2.720579	39530	41595	ORF	NA
4	97385	3.230501	96987	97733	Intergene	1
4	157047	2.371075	156798	157333	Intergene	1
4	168318	3.324551	167969	168589	Intergene	1
4	208305	2.114345	207875	208603	Intergene	NA
4	245831	2.547063	245483	246513	Intergene	1
4	256877	3.240304	255840	257202	Intergene	NA
4	316085	2.952205	315500	316538	ORF	NA
4	326517	3.215363	326143	327430	Intergene	NA
4	327126	3.753768	326143	327430	Intergene	NA
4	346163	3.45737	345743	347259	ORF	NA
4	347054	2.113977	345743	347259	ORF	NA
4	471093	3.130356	470769	471424	Intergene	NA
4	481170	3.11571	480557	481702	ORF	NA
4	548884	2.380814	548430	549261	ORF	NA
4	627586	2.575335	626934	627924	Intergene	1
4	676180	1.905426	675107	677025	ORF	NA
4	772574	2.593408	772255	772933	Intergene	1

Table 4.1. Continued.

Chr	PeakPos	PeakHeight	minPos	maxPos	GenomePos	tRNA
4	780383	3.276612	779877	781423	Intergene	1
4	781112	2.63207	779877	781423	Intergene	1
4	782666	3.536393	782246	783253	Intergene	2
4	885837	2.004617	885429	886536	ORF	NA
4	937720	3.577388	937083	938275	ORF	NA
4	962837	1.986329	962451	963741	ORF	NA
4	963380	2.57611	962451	963741	ORF	NA
4	993212	3.662442	992364	994712	Intergene	NA
4	993861	3.213261	992364	995342	Intergene	NA
4	995022	2.024649	993522	995342	Intergene	NA
4	1006999	2.799863	1006618	1007428	Intergene	2
4	1015499	2.854312	1014893	1015944	Intergene	1
4	1054339	2.975856	1053741	1054846	ORF	NA
4	1055836	1.842657	1055546	1056201	ORF	NA
4	1206377	2.182445	1204877	1207877	Intergene	NA
4	1325189	2.5286	1324779	1325520	Intergene	1
4	1331650	1.990322	1331342	1332085	Intergene	1
4	1468893	3.494309	1467860	1469997	ORF	NA
4	1498686	1.798958	1498033	1499461	ORF	NA
5	23797	2.215098	23162	25297	ORF	NA
5	25179	2.592402	23679	26679	ORF	NA
5	26163	2.814011	24663	27663	ORF	NA
5	26705	3.0484	25205	28205	ORF	NA
5	28166	2.706366	26666	29666	ORF	NA
5	28865	3.155866	27365	29726	ORF	NA
5	32489	1.828832	30989	32978	ORF	NA
5	66502	2.816219	66067	67029	Intergene	1
5	87133	2.718448	86884	87721	Intergene	NA
5	135427	2.590994	134688	135987	ORF	NA
5	217662	2.001557	217304	217898	Intergene	2
5	220252	2.208925	219749	220857	Intergene	1
5	343610	2.7972	343062	344251	ORF	NA
5	348617	4.028471	347510	350117	ORF	NA
5	349147	3.892384	347647	350647	ORF	NA
5	349693	3.292755	348193	351165	ORF	NA
5	350401	3.448196	348901	351165	ORF	NA
5	355129	1.880865	354793	355908	Intergene	NA
5	371420	1.837191	370492	372643	Intergene	NA

Table 4.1. Continued.

Chr	PeakPos	PeakHeight	minPos	maxPos	GenomePos	tRNA
5	399282	1.858767	398801	399739	ORF	NA
5	469413	2.444965	468221	470913	Intergene	NA
5	470589	3.074422	469089	471645	Intergene	NA
5	638262	2.079735	637757	638945	Intergene	NA
5	683606	2.758845	683129	684079	ORF	NA
5	743972	2.033636	743193	744402	ORF	NA
5	769645	2.716853	769219	770164	Intergene	1
5	776984	2.828027	776623	777456	Intergene	NA
5	793572	2.794848	793161	794017	ORF	NA
5	857007	4.475825	856434	857719	Intergene	NA
5	920520	2.435105	919672	921683	ORF	NA
5	946082	2.60272	945562	947307	ORF	NA
5	997259	2.584267	996880	997705	ORF	NA
5	1035091	2.157411	1034776	1035512	Intergene	1
5	1043562	2.18785	1042062	1044239	ORF	NA
5	1108677	3.325126	1108037	1109333	ORF	NA
5	1179495	1.89777	1178603	1180568	ORF	NA
6	24405	2.610869	23917	24953	ORF	NA
6	54747	2.036829	54120	56043	ORF	NA
6	114949	2.357026	114678	115496	ORF	NA
6	132783	4.090826	131709	133393	ORF	NA
6	159583	2.322507	158083	159967	ORF	NA
6	170585	2.571126	169884	171234	ORF	NA
6	310000	1.978682	309206	310364	Intergene	NA
6	332129	2.747251	331319	332657	ORF	NA
6	357666	3.019816	357222	358303	ORF	NA
6	378892	3.35416	378482	379320	Intergene	1
6	403048	2.843424	402567	403601	ORF	NA
6	425271	2.584037	424874	425944	ORF	NA
6	465618	2.755786	464791	466013	ORF	NA
6	486812	2.309642	486279	487539	ORF	NA
6	496942	2.499527	496106	498442	ORF	NA
6	498372	1.931118	496872	498743	Intergene	NA
6	554567	1.887461	553658	555772	ORF	NA
6	843399	3.198601	843033	843684	Intergene	1
6	900659	4.027694	899159	902159	ORF	NA
6	981061	2.557272	980643	982561	Intergene	NA
6	982388	3.25697	980888	983603	Intergene	NA

Table 4.1. Continued,

Chr	PeakPos	PeakHeight	minPos	maxPos	GenomePos	tRNA
6	993075	2.335477	992610	993435	Intergene	1
7	202931	2.410138	202418	203549	ORF	NA
7	212503	2.472747	211861	213064	ORF	NA
7	259044	3.05541	257886	259579	ORF	NA
7	384682	3.196189	384141	385321	ORF	NA
7	387091	2.916768	386014	388591	ORF	NA
7	388512	2.448702	387012	388888	ORF	NA
7	427355	3.391529	425917	428855	Intergene	NA
7	428121	1.980127	426621	429081	Intergene	NA
7	440319	2.03713	439190	441819	ORF	NA
7	441225	1.958851	439725	441934	ORF	NA
7	519210	1.827737	518732	519533	ORF	NA
7	691483	1.929794	690605	692113	ORF	NA
7	711852	1.863654	710559	713129	ORF	NA
7	866037	2.446696	865678	866442	Intergene	1

Table 4.2. Genome-wide predictions of replication origins

name	Chr	Position	tRNA	location
ORI001	0	23570	NA	ORF
ORI002	0	157818	NA	ORF
ORI003	0	158973	NA	ORF
ORI004	0	160147	NA	ORF
ORI005	0	217055	NA	Intergene
ORI006	0	282154	NA	Intergene
ORI007	0	339066	1	Intergene
ORI008	0	368305	NA	Intergene
ORI009	0	369623	NA	Intergene
ORI010	0	435802	NA	Intergene
ORI011	0	456088	1	Intergene
ORI012	0	494009	NA	Intergene
ORI013	0	510878	NA	ORF
ORI014	0	515133	NA	Intergene
ORI015	0	516373	NA	Intergene
ORI016	0	562763	NA	ORF
ORI017	0	563736	NA	ORF
ORI018	0	600848	NA	ORF
ORI019	0	644253	NA	Intergene
ORI020	0	692225	NA	Intergene
ORI021	0	719169	NA	Intergene
ORI022	0	775351	NA	Intergene
ORI023	0	796293	NA	Intergene
ORI024	0	919984	NA	Intergene
ORI025	0	954793	NA	Intergene
ORI026	0	981850	NA	Intergene
ORI027	0	995232	NA	Intergene
ORI028	0	996237	NA	ORF
ORI029	0	997088	NA	Intergene
ORI030	0	1019838	2	Intergene
ORI031	0	1020768	1	Intergene
ORI032	0	1040925	NA	Intergene
ORI033	0	1070592	NA	Intergene
ORI034	0	1158636	2	Intergene
ORI035	0	1159476	2	Intergene
ORI036	0	1190556	2	Intergene
ORI037	0	1229527	1	Intergene

Table 4.2. Continued.

name	Chr	Position	tRNA	location
ORI038	0	1230050	2	Intergene
ORI039	0	1386193	NA	Intergene
ORI040	0	1421883	NA	Intergene
ORI041	0	1423001	NA	Intergene
ORI042	0	1423905	NA	Intergene
ORI043	0	1473803	1	Intergene
ORI044	0	1474658	1	Intergene
ORI045	0	1519130	NA	Intergene
ORI046	0	1607671	1	Intergene
ORI047	0	1657000	NA	Intergene
ORI048	0	1662451	NA	Intergene
ORI049	0	1723961	NA	ORF
ORI050	0	1724957	NA	ORF
ORI051	0	1725612	NA	Intergene
ORI052	0	1745371	NA	Intergene
ORI053	0	1746887	NA	Intergene
ORI054	0	1763553	NA	ORF
ORI055	0	1783340	NA	Intergene
ORI056	0	1784162	NA	Intergene
ORI057	0	1787352	2	Intergene
ORI058	0	1788266	1	Intergene
ORI059	0	1812834	1	Intergene
ORI060	0	1830931	NA	Intergene
ORI061	0	2008917	1	Intergene
ORI062	0	2009574	1	Intergene
ORI063	0	2010772	NA	Intergene
ORI064	0	2048706	NA	Intergene
ORI1001	1	13714	NA	Intergene
ORI1002	1	31783	NA	Intergene
ORI1003	1	32984	NA	Intergene
ORI1004	1	45776	NA	Intergene
ORI1005	1	264889	NA	Intergene
ORI1006	1	268902	NA	Intergene
ORI1007	1	289359	NA	Intergene
ORI1008	1	332603	NA	Intergene
ORI1009	1	333422	NA	ORF
ORI1010	1	334264	NA	Intergene
ORI1011	1	336053	1	Intergene

Table 4.2. Continued.

name	Chr	Position	tRNA	location
ORI1012	1	336659	1	Intergene
ORI1013	1	353902	1	Intergene
ORI1014	1	488275	1	Intergene
ORI1015	1	505446	NA	Intergene
ORI1016	1	520806	NA	Intergene
ORI1017	1	522472	NA	Intergene
ORI1018	1	536406	1	Intergene
ORI1019	1	626380	NA	Intergene
ORI1020	1	630084	NA	Intergene
ORI1021	1	630854	NA	Intergene
ORI1022	1	631950	NA	Intergene
ORI1023	1	643106	NA	Intergene
ORI1024	1	644339	NA	Intergene
ORI1025	1	649970	NA	Intergene
ORI1026	1	662040	NA	Intergene
ORI1027	1	662923	NA	Intergene
ORI1028	1	663593	NA	Intergene
ORI1029	1	697070	NA	Intergene
ORI1030	1	706142	NA	Intergene
ORI1031	1	707184	NA	ORF
ORI1032	1	718861	1	Intergene
ORI1033	1	748011	1	Intergene
ORI1034	1	751193	NA	Intergene
ORI1035	1	790401	NA	Intergene
ORI1036	1	859316	NA	Intergene
ORI1037	1	871642	NA	Intergene
ORI1038	1	885962	NA	Intergene
ORI1039	1	891075	NA	Intergene
ORI1040	1	891632	NA	Intergene
ORI1041	1	912207	NA	Intergene
ORI1042	1	950643	NA	ORF
ORI1043	1	951618	NA	Intergene
ORI1044	1	1115545	NA	Intergene
ORI1045	1	1167770	2	Intergene
ORI1046	1	1168803	2	Intergene
ORI1047	1	1177344	NA	Intergene
ORI1048	1	1195034	NA	Intergene
ORI1049	1	1196092	NA	ORF

Table 4.2. Continued.

name	Chr	Position	tRNA	location
ORI1050	1	1278559	NA	Intergene
ORI1051	1	1279723	NA	Intergene
ORI1052	1	1280788	NA	Intergene
ORI1053	1	1375567	NA	Intergene
ORI1054	1	1376751	NA	Intergene
ORI1055	1	1377794	NA	Intergene
ORI1056	1	1408939	NA	Intergene
ORI1057	1	1468296	NA	Intergene
ORI1058	1	1469420	NA	Intergene
ORI1059	1	1484772	NA	Intergene
ORI1060	1	1489358	1	Intergene
ORI1061	1	1564576	NA	Intergene
ORI1062	1	1565271	NA	Intergene
ORI1063	1	1570733	NA	Intergene
ORI1064	1	1765693	NA	Intergene
ORI1065	1	1827379	NA	ORF
ORI1066	1	1832589	NA	Intergene
ORI1067	1	1833827	NA	Intergene
ORI1068	1	1834935	NA	Intergene
ORI1069	1	1843978	NA	ORF
ORI1070	1	1864524	NA	Intergene
ORI1071	1	1952311	NA	Intergene
ORI1072	1	1953189	NA	Intergene
ORI1073	1	1977471	NA	Intergene
ORI1074	1	2000003	NA	Intergene
ORI1075	1	2164041	1	Intergene
ORI1076	1	2217427	NA	Intergene
ORI1077	1	2236246	NA	Intergene
ORI1078	1	2299538	NA	Intergene
ORI1079	1	2392264	NA	ORF
ORI1080	1	2407749	NA	Intergene
ORI1081	1	2410070	1	Intergene
ORI1082	1	2410867	1	Intergene
ORI1083	1	2432301	NA	Intergene
ORI1084	1	2433000	NA	ORF
ORI1085	1	2433825	NA	Intergene
ORI1086	1	2435755	NA	Intergene
ORI1087	1	2436381	NA	Intergene

Table 4.2. Continued.

name	Chr	Position	tRNA	locaton
ORI1088	1	2454804	NA	Intergene
ORI1089	1	2492068	NA	Intergene
ORI1090	1	2513438	NA	Intergene
ORI1091	1	2588273	1	Intergene
ORI1092	1	2588834	2	Intergene
ORI1093	1	2698028	NA	Intergene
ORI1094	1	2699382	NA	Intergene
ORI1095	1	2783924	NA	ORF
ORI1096	1	2784650	NA	ORF
ORI1097	1	2830673	NA	Intergene
ORI1098	1	2839737	NA	Intergene
ORI1099	1	2840906	NA	Intergene
ORI1100	1	2868535	NA	Intergene
ORI1101	1	2883209	NA	Intergene
ORI1102	1	2884449	NA	Intergene
ORI1103	1	3012806	NA	ORF
ORI1104	1	3038872	NA	Intergene
ORI1105	1	3046878	NA	Intergene
ORI1106	1	3060479	NA	Intergene
ORI1107	1	3061077	1	Intergene
ORI1108	1	3106691	NA	Intergene
ORI1109	1	3108035	NA	Intergene
ORI1110	1	3147669	NA	Intergene
ORI201	2	6773	NA	ORF
ORI202	2	7327	1	ORF
ORI203	2	8058	NA	Intergene
ORI204	2	165712	NA	ORF
ORI205	2	175514	1	ORF
ORI206	2	182147	NA	ORF
ORI207	2	216659	1	Intergene
ORI208	2	314096	NA	Intergene
ORI209	2	322913	1	Intergene
ORI210	2	496094	1	Intergene
ORI211	2	543865	NA	ORF
ORI212	2	591172	NA	Intergene
ORI213	2	592018	NA	Intergene
ORI214	2	603969	NA	Intergene
ORI215	2	625530	NA	ORF

Table 4.2. Continued.

name	Chr	Position	tRNA	location
ORI216	2	626526	NA	Intergene
ORI217	2	675754	NA	Intergene
ORI218	2	676947	NA	Intergene
ORI219	2	705708	NA	ORF
ORI220	2	751510	NA	ORF
ORI221	2	771504	1	Intergene
ORI222	2	775133	NA	Intergene
ORI223	2	789782	1	ORF
ORI224	2	804199	NA	Intergene
ORI225	2	805327	NA	Intergene
ORI226	2	807018	NA	Intergene
ORI227	2	809814	NA	Intergene
ORI228	2	831594	NA	Intergene
ORI229	2	833748	NA	Intergene
ORI230	2	920194	NA	Intergene
ORI231	2	961221	NA	Intergene
ORI232	2	1026757	1	Intergene
ORI233	2	1032463	NA	Intergene
ORI234	2	1041975	NA	Intergene
ORI235	2	1077201	1	Intergene
ORI236	2	1112023	1	Intergene
ORI237	2	1118552	NA	Intergene
ORI238	2	1119267	NA	ORF
ORI239	2	1141824	NA	Intergene
ORI240	2	1143161	NA	Intergene
ORI241	2	1167084	NA	Intergene
ORI242	2	1302182	NA	Intergene
ORI243	2	1321033	1	ORF
ORI244	2	1395533	NA	Intergene
ORI245	2	1508413	NA	ORF
ORI246	2	1619861	1	Intergene
ORI247	2	1621572	1	Intergene
ORI248	2	1633024	NA	Intergene
ORI249	2	1638401	NA	Intergene
ORI250	2	1690455	NA	Intergene
ORI251	2	1748263	NA	ORF
ORI252	2	1789939	NA	Intergene

Table 4.2. Continued

name	Chr	Position	tRNA	location
ORI253	2	1800222	NA	Intergene
ORI254	2	1888636	1	Intergene
ORI255	2	1921860	NA	Intergene
ORI256	2	1922440	NA	ORF
ORI257	2	1926762	NA	Intergene
ORI258	2	1927968	NA	Intergene
ORI259	2	2167677	1	Intergene
ORI260	2	2168246	2	Intergene
ORI301	3	15819	NA	Intergene
ORI302	3	31679	NA	Intergene
ORI303	3	132816	NA	Intergene
ORI304	3	134088	NA	ORF
ORI305	3	134892	NA	ORF
ORI306	3	135489	NA	ORF
ORI307	3	195628	1	Intergene
ORI308	3	261451	1	Intergene
ORI309	3	290245	NA	Intergene
ORI310	3	407605	1	Intergene
ORI311	3	408707	NA	Intergene
ORI312	3	411794	1	Intergene
ORI313	3	447349	1	Intergene
ORI314	3	457655	NA	Intergene
ORI315	3	458590	NA	Intergene
ORI316	3	538817	NA	Intergene
ORI317	3	539709	NA	Intergene
ORI318	3	591537	1	ORF
ORI319	3	592220	2	Intergene
ORI320	3	702641	NA	Intergene
ORI321	3	818642	NA	Intergene
ORI322	3	824059	NA	Intergene
ORI323	3	824941	NA	Intergene
ORI324	3	835502	1	Intergene
ORI325	3	865880	NA	Intergene
ORI326	3	941265	NA	Intergene
ORI327	3	974963	NA	ORF
ORI328	3	991466	NA	Intergene
ORI329	3	992878	NA	Intergene

Table 4.2. Continued.

name	Chr	Position	tRNA	location
ORI330	3	1022900	NA	Intergene
ORI331	3	1031686	NA	Intergene
ORI332	3	1032559	NA	Intergene
ORI333	3	1033654	NA	Intergene
ORI334	3	1035334	NA	Intergene
ORI335	3	1038783	1	Intergene
ORI336	3	1096455	NA	Intergene
ORI337	3	1130179	NA	Intergene
ORI338	3	1147503	NA	ORF
ORI339	3	1154241	NA	Intergene
ORI340	3	1192938	NA	Intergene
ORI341	3	1197382	NA	Intergene
ORI342	3	1232165	1	Intergene
ORI343	3	1233320	NA	Intergene
ORI344	3	1290770	NA	Intergene
ORI345	3	1366668	NA	Intergene
ORI346	3	1367765	NA	ORF
ORI347	3	1388602	NA	Intergene
ORI348	3	1560752	1	Intergene
ORI349	3	1561755	NA	Intergene
ORI350	3	1575452	NA	ORF
ORI351	3	1594703	1	ORF
ORI352	3	1595802	NA	Intergene
ORI353	3	1701930	NA	Intergene
ORI354	3	1715348	2	Intergene
ORI355	3	1731577	NA	ORF
ORI356	3	1798401	NA	Intergene
ORI401	4	12737	NA	Intergene
ORI402	4	40693	NA	ORF
ORI403	4	41562	NA	Intergene
ORI404	4	96879	1	Intergene
ORI405	4	156547	NA	Intergene
ORI406	4	167823	1	Intergene
ORI407	4	168712	1	ORF
ORI408	4	174706	NA	ORF
ORI409	4	175450	NA	ORF
ORI410	4	245520	1	Intergene

Table 4.2. Continued.

name	Chr	Position	tRNA	location
ORI411	4	255573	NA	Intergene
ORI412	4	256365	NA	ORF
ORI413	4	315679	NA	Intergene
ORI414	4	316917	NA	Intergene
ORI415	4	326632	NA	Intergene
ORI416	4	327618	NA	Intergene
ORI417	4	345928	NA	Intergene
ORI418	4	346617	NA	Intergene
ORI419	4	470463	NA	Intergene
ORI420	4	480446	NA	Intergene
ORI421	4	772175	1	Intergene
ORI422	4	779772	NA	Intergene
ORI423	4	780442	1	Intergene
ORI424	4	782153	1	Intergene
ORI425	4	873749	NA	Intergene
ORI426	4	937207	NA	ORF
ORI427	4	937942	NA	Intergene
ORI428	4	962985	NA	Intergene
ORI429	4	992153	NA	Intergene
ORI430	4	993649	NA	Intergene
ORI431	4	994553	NA	Intergene
ORI432	4	1006711	2	Intergene
ORI433	4	1007582	NA	Intergene
ORI434	4	1055252	NA	Intergene
ORI435	4	1130850	NA	Intergene
ORI436	4	1468763	NA	Intergene
ORI501	5	25670	NA	Intergene
ORI502	5	27592	NA	Intergene
ORI503	5	28384	NA	Intergene
ORI504	5	33364	NA	Intergene
ORI505	5	66103	1	Intergene
ORI506	5	86640	NA	Intergene
ORI507	5	135956	NA	Intergene
ORI508	5	144524	NA	Intergene
ORI509	5	219723	1	Intergene
ORI510	5	342840	NA	Intergene
ORI511	5	343845	NA	Intergene

Table 4.2. Continued.

name	Chr	Position	tRNA	location
ORI512	5	348565	NA	Intergene
ORI513	5	349829	NA	Intergene
ORI514	5	354844	NA	ORF
ORI515	5	370591	NA	Intergene
ORI516	5	468980	NA	Intergene
ORI517	5	469870	NA	Intergene
ORI518	5	519439	NA	Intergene
ORI519	5	555214	NA	Intergene
ORI520	5	637788	NA	Intergene
ORI521	5	682916	NA	Intergene
ORI522	5	743665	NA	ORF
ORI523	5	776455	NA	Intergene
ORI524	5	786274	NA	ORF
ORI525	5	793087	NA	Intergene
ORI526	5	856333	NA	Intergene
ORI527	5	857581	NA	Intergene
ORI528	5	919585	NA	Intergene
ORI529	5	920755	NA	Intergene
ORI530	5	946251	NA	Intergene
ORI531	5	996827	NA	Intergene
ORI532	5	997823	NA	Intergene
ORI533	5	1034532	NA	Intergene
ORI534	5	1108077	NA	ORF
ORI535	5	1178315	NA	Intergene
ORI536	5	1179635	NA	Intergene
ORI601	6	24883	NA	Intergene
ORI602	6	114481	NA	Intergene
ORI603	6	132257	NA	Intergene
ORI604	6	133398	NA	Intergene
ORI605	6	159028	NA	ORF
ORI606	6	331827	NA	ORF
ORI607	6	332849	NA	Intergene
ORI608	6	356948	NA	Intergene
ORI609	6	378278	1	Intergene
ORI610	6	402767	NA	Intergene
ORI611	6	403766	NA	Intergene
ORI612	6	424815	NA	Intergene

Table 4.2. Continued.

name	Chr	Position	tRNA	location
ORI613	6	426260	NA	Intergene
ORI614	6	465319	NA	Intergene
ORI615	6	466253	NA	Intergene
ORI616	6	491357	NA	Intergene
ORI617	6	495909	NA	Intergene
ORI618	6	496770	NA	ORF
ORI619	6	498026	NA	Intergene
ORI620	6	554455	NA	Intergene
ORI621	6	555054	NA	Intergene
ORI622	6	842944	1	Intergene
ORI623	6	843922	NA	Intergene
ORI624	6	874548	NA	Intergene
ORI625	6	981434	NA	Intergene
ORI626	6	982402	NA	Intergene
ORI701	7	202141	NA	Intergene
ORI702	7	211691	NA	Intergene
ORI703	7	257821	NA	Intergene
ORI704	7	258645	NA	ORF
ORI705	7	304103	NA	Intergene
ORI706	7	384326	NA	ORF
ORI707	7	385566	NA	Intergene
ORI708	7	388075	NA	Intergene
ORI709	7	427675	NA	Intergene
ORI710	7	440668	NA	ORF
ORI711	7	690492	NA	Intergene
ORI712	7	865630	1	Intergene

Table 4.3. Genome-wide ARS screen

name	Chr	minPos	maxPos	AvgORC	read depth
ARS001	0	197947	198018	0.650918659	21
ARS002	0	229756	229966	-0.28611725	10069
ARS003	0	361917	362083	-0.08534933	241
ARS004	0	628300	628501	0.536730374	2173
ARS005	0	649217	649347	-0.31801975	2295
ARS006	0	788852	788925	-0.08021849	724
ARS007	0	980497	980674	-0.10857484	1032
ARS008	0	1183573	1183762	0.507603653	449
ARS009	0	1188413	1188855	0.298998215	542
ARS010	0	1550259	1550314	0.222865725	278
ARS011	0	1750736	1750799	0.650214664	7312
ARS012	0	1785122	1785332	2.167823353	5072
ARS013	0	1786504	1786613	-0.61520343	5945
ARS014	0	1786954	1787175	1.248058944	17203
ARS015	0	1787216	1787303	1.211876756	67774
ARS016	0	1788308	1788460	0.330024245	104
ARS017	0	1788763	1788814	3.107717616	138
ARS018	0	1817487	1817676	0.622335013	241
ARS019	0	1889682	1889898	0.538202192	45
ARS020	0	1889898	1889978	0.577730606	1580
ARS021	0	1891305	1891514	0.849796428	38
ARS022	0	1891742	1891852	0.888011931	12
ARS023	0	2187256	2187357	1.068135906	1290
ARS024	0	2204950	2205091	-0.21506911	37398
ARS025	0	2205805	2205920	-1.87650055	1629
ARS101	1	19439	19520	0.514204123	5147
ARS102	1	310700	310881	-0.24768917	114
ARS103	1	477520	477594	-0.67539565	212
ARS104	1	550498	550563	0.162234152	2826
ARS105	1	749294	749467	-0.70487383	815
ARS106	1	845711	845783	0.468596347	12
ARS107	1	1050558	1050641	-0.00962725	44752
ARS108	1	1145771	1145885	0.733575252	762
ARS109	1	1153114	1153219	-0.18505260	1240
ARS110	1	1174022	1174100	-0.46176698	82595
ARS111	1	1208560	1208659	-1.18436858	2922

Table 4.3. Continued.

name	Chr	minPos	maxPos	AvgORC	read depth
ARS112	1	1398119	1398291	0.158109851	681
ARS113	1	1639701	1639757	-0.72425441	37
ARS114	1	2023573	2023627	0.680739613	21552
ARS115	1	2026865	2026935	0.28725076	22915
ARS116	1	2216097	2216244	0.991406593	24180
ARS117	1	2230867	2230986	-0.61367698	3673
ARS118	1	2701225	2701336	0.236510848	4559
ARS119	1	2945288	2945369	0.069819205	35883
ARS120	1	3135889	3136031	-0.19984178	962
ARS121	1	3172420	3172528	0.536755816	64244
ARS122	1	3175034	3175334	0.297872744	1456
ARS201	2	22942	23155	0.704557059	18
ARS202	2	70204	70297	-0.85172366	38115
ARS203	2	328468	328538	0.323100276	113
ARS204	2	462637	462810	0.523495379	165597
ARS205	2	694534	694813	-0.27325460	1976
ARS206	2	704608	704674	0.366182919	7781
ARS207	2	761038	761104	-0.36934256	9905
ARS208	2	862223	862314	-0.78874943	14752
ARS209	2	1013454	1013589	-0.03961313	1209
ARS210	2	1017661	1017781	0.888343476	5985
ARS211	2	1062166	1062244	-0.08653921	5254
ARS212	2	1101279	1101345	0.082392011	40537
ARS213	2	1119157	1119286	2.736576893	93
ARS214	2	1121952	1122250	-0.9425474	22
ARS215	2	1250290	1250360	-0.07079434	1489
ARS216	2	1374955	1375060	0.794089615	120
ARS217	2	1397160	1397223	-0.32363000	2
ARS218	2	1462422	1462500	0.391624476	42925
ARS219	2	1573341	1573580	-0.43450265	190626
ARS220	2	1614690	1614818	0.78011612	5528
ARS221	2	1618227	1618454	0.230007905	933
ARS222	2	1619271	1619395	-0.12994532	29178
ARS223	2	1632378	1632467	0.090135315	28020
ARS224	2	1793627	1793833	-0.01688894	3811
ARS225	2	1848035	1848165	-0.59941619	16688
ARS226	2	1883945	1884015	0.06551185	116978

Table 4.3. Continued.

name	Chr	minPos	maxPos	AvgORC	read depth
ARS301	3	179510	179707	0.026874823	775
ARS302	3	236781	236914	-0.37575965	9614
ARS303	3	304063	304205	-1.61903290	21890
ARS304	3	377009	377126	0.891019119	5051
ARS305	3	569462	569536	0.008292145	76
ARS306	3	887351	887505	0.047001169	65
ARS307	3	1093328	1093426	-0.37393113	616
ARS308	3	1384706	1384881	0.151317313	6039
ARS309	3	1519856	1519916	0.670972665	1564
ARS310	3	1556360	1556465	-0.37302699	3282
ARS311	3	1620864	1621130	0.119417237	6
ARS312	3	1640439	1640506	0.938576874	71511
ARS313	3	1706285	1706360	-0.36357479	39
ARS314	3	1770950	1771058	-0.32350698	25038
ARS401	4	22800	22988	0.365945737	364
ARS402	4	245408	245462	0.264258276	38
ARS403	4	246309	246365	0.822780262	4194
ARS404	4	446246	446359	-1.01320949	17
ARS405	4	562270	562344	-1.61208151	1666
ARS406	4	616508	616663	0.88435437	6865
ARS407	4	677433	677540	0.44172187	122
ARS408	4	713375	713594	-0.09785271	1521
ARS409	4	800840	800938	0.440094273	1959
ARS410	4	855626	855678	-0.77230324	6015
ARS411	4	904315	904383	0.360693012	1151831
ARS412	4	927799	927904	-0.96858513	134
ARS413	4	1026377	1026518	0.56983614	257
ARS414	4	1079924	1080009	-0.63963254	682
ARS501	5	48434	48676	0.54874149	3
ARS502	5	192413	192569	-0.31921322	265
ARS503	5	234086	234215	-1.06718442	23054
ARS504	5	406609	406771	-0.61087149	39
ARS505	5	1141740	1141959	-0.86439491	63
ARS601	6	26259	26333	0.057413559	518
ARS602	6	117475	117592	0.241031403	12393

Table 4.3. Continue.

name	Chr	minPos	maxPos	AvgORC	read depth
ARS603	6	159539	159603	2.303445012	5743
ARS604	6	300903	300975	-0.94753112	551
ARS605	6	304877	305060	0.225550412	4096
ARS606	6	341247	341389	0.4134194	1314
ARS607	6	616544	616614	-1.09092096	13395
ARS608	6	643793	643894	0.223215457	1630
ARS609	6	679880	679964	0.242408728	217
ARS610	6	743927	744011	-1.34910580	62216
ARS611	6	791933	792107	0.156010251	8973
ARS612	6	834649	834799	0.018162978	1996
ARS613	6	903380	903488	1.131664573	13826
ARS614	6	973046	973586	-1.05133097	47
ARS615	6	1007624	1007684	0.849958921	94
ARS701	7	4381	4480	0.718924446	94667
ARS702	7	36686	36745	0.312820917	6337
ARS703	7	119743	120026	0.271510523	156
ARS704	7	145286	145544	-0.32850808	2
ARS705	7	646060	646187	1.026257286	40
ARS706	7	898882	898962	-0.74319007	1375

Chapter 5.

Summary

Faithful genome replication and segregation ensures the transmission of genetic information during cell division. Understanding how cells specify genomic loci to initiate replication is a long-standing question in biology. The genetic elements required for genome inheritance are centromeres and replication origins. Here, my work focused on identifying and characterizing replication origins in a previously uncharacterized yeast genome that has many features shared with higher eukaryotes. Based on genome-wide replication timing analyses in *C. albicans*, we observed early replication events around centromere regions; thus, I started to explore how replication origins are organized in the *C. albicans* genome.

Early replication of centromere regions in S-phase is a shared feature in the evolutionary history of fungi. Prior to our work, Kim and Huberman showed that heterochromatic centromeres in *S. pombe* replicate early in S-phase, indicating that centromeres may be subject to a position effect that regulates replication timing (Kim, 2003). However, *S. pombe* centromeres are repetitive and heterochromatic making it difficult to distinguish specific centromeres and mechanisms. The connections between early replication and centromeres have been elusive due to experimental limitations. We used *C. albicans* as an attractive model organism to study the connection between centromeres and DNA replication because it has regional, epigenetically defined centromeres, which are similar to the majority of eukaryotes, but does not have heterochromatic, complicated sequence structures. Since *C. albicans* shares

epigenetic features of centromeres with higher eukaryotes, we asked if the *C. albicans* origin features are also similar to those of higher eukaryotes. Furthermore, we explored whether the origins at centromeres and origins on the chromosomal arms are activated by different mechanisms.

The overall goal of this dissertation was to map replication origins throughout the genome and to investigate the nature of replication origins. Thus, the first goal was to understand the coordination between genome replication (origins) and segregation (centromeres). We took advantage of multiple genome-scale approaches. Amnon Koren's initial work first revealed a genome-wide replication timing profile prior to my work and showed around 150-200 origins actively firing during the cell cycle (Koren et al., 2010b). The replication timing profile provided a genome-wide view of the patterns of replication origin activity; however, the low resolution could not specify the exact locations of replication origins. I used biochemical approaches to determine that centromeres in *C. albicans* are active origins. Additionally, I found that centromeric origins are distinct, and their function is associated with epigenetically inherited centromeres (Chapter 2). To further specify origin locations, I mapped ORC binding sites throughout the genome (Chapter 2 and 4). I developed a new linear plasmid vector by utilizing centromere DNA and ORC binding sequences as potential ARSs and investigated plasmid behavior. The successful establishment of a plasmid shuttle vector is groundbreaking, as no well-characterized plasmids that are maintained autonomously were previously available for *C. albicans* (Chapter

3). No sequence motifs specifying replication origins had been defined in *C. albicans* and they were not thought to exist because *C. albicans* has regional centromeres. Thus, a sequence-independent and a sequence-dependent method to identify replication origins were performed in parallel. Combining computational and experimental approaches, I mapped replication origins in the *C. albicans* genome and identified a 15 bp ACS motif that is necessary for origin activity (Chapter 4).

Centromeric origins are epigenetically defined and associated with the centromere function.

A centromere is the chromosomal locus where kinetochore assembly occurs to specify chromosome segregation and is found in only one copy per chromosome in most eukaryotes. CENP-A/Cse4 association defines the centromere location. In a select group of yeast species with point centromeres, the centromere is defined by sequence, but in most eukaryotes the centromere is defined epigenetically. Fungal species with epigenetic centromeres and sequence-defined centromeres both have early replication near centromeres. This intriguing correlation connects two major mechanisms for genome inheritance. Koren *et al.* characterized the replication timing profile in *C. albicans* and found that the centromere region not only replicates early, but is the first place to replicate on each chromosome. Based on this work, I examined potential origin locations around centromeres to understand why centromeres replicate

early. Unexpectedly, ORC strongly binds centromere core regions. DNA 2D gels further confirmed the origin activity at centromere core regions. The results provide a direct link between early replication and centromeres. The epigenetic nature of centromere identity in *C. albicans* led us to ask if origin function is also epigenetically inheritable. A neocentromere forms at an ectopic locus of the chromosome when the native centromere is deleted (Ketel et al., 2009). Thus, this neocentromere strain becomes an ideal tool to address the question of whether functional centromeres replicate early. Strikingly, the earliest replicating locus on the chromosome shifted to the location of the neocentromere, a location that was not an origin of replication in the wild type strain. I further investigated the mechanism of this epigenetic correlation and found that the centromeric origin is dependent on the centromere determinant, Cse4. This is the first evidence that origin activity is associated with centromere function.

Interestingly, a sequence feature, GC skew, which associates with replication origins in organisms containing only a single origin per genome, appears at all centromeres in *C. albicans*. The appearance of this fossil sequence feature suggests that the centromeric origin is constitutively activated during every cell cycle. Interestingly, after our exciting findings, a more recent study found that *S. cerevisiae* centromeres have a positional effect that promotes the early replication timing of pericentric origins (Pohl et al., 2012). The centromere position effect can influence the activation timing of origins up to ~20 kb away. Thus, even in *S. cerevisiae*, where centromeres are specificed by a

defined seuqence, novel *trans*-acting factors may regulate the early activation of pericentric origins.

I then compared the centromeric origin with another active origins identified from the genome-wide mapping of ORC binding sites. I discovered that a histone H3 lysine 4 methyltransferase, Set1 regulates ORC occupancy at chromosomal origins but does not affect centromeric ORC binding. Recently, we performed ChIP-seq to map ORC binding sites in both wild type and *set1* deletion strains. When we analyzed the genome-wide set of origins, ORC occupancy at identified origins is dramatically reduced in a *set1* deletion strain but is not reduced at centromeres. The results support the idea that the centromeric origin has properties distinct from those of chromosomal arm origins and that centromeric origins are epigenetically defined constitutive origins.

My work suggests an important future direction for this work--investigating how *trans*-acting factors differentially regulate origin activity. For instance, do epigenetic centromeres also have a centromere position effect? If there is a centromere position effect, ORC occupancy near the centromere may be regulated differently from its occupancy at non-centromeric origins. Moreover, for the subset of Set1-independent ORC binding sites, there may be additional levels of epigenetic regulation that control origin activity. In *S. cerevisiae,* Fkh1/2 (Forkhead transcription factors) established long-range origin clustering to coordinately regulate replication timing by spatial organization to foci where

Cdc45 and other replication factors are concentrated in the nucleus (Knott et al., 2012). When an origin is recruited by an epigenetic centromere, it replicates early. We hypothesize that centromeric origins may co-localize with other early replicating origins or other Set1-independent origins in the nucleus.

An interesting observation from my origin mapping experiments is that the majority of tRNA loci are near strong ORC binding sites. Interestingly, all the active origins identified in DNA 2D gels in this work were near tRNA loci. A major challenge for performing DNA 2D gel analysis of *C. albicans* bubble-arc replicating intermediates is that there is no good protocol for isolating highly synchronized cells. Thus, all active origins identified in this work are likely to be highly efficient origins, including the centromeric origins. This may suggest that tRNA-related *trans*-acting factors affect origin activity. Intriguingly, in *S. pombe*, an RNA PolIII transcription factor, TFIIIC can tether dispersed binding loci, including tRNA loci, to centromeric regions (Iwasaki and Noma, 2012; Noma et al., 2006). I hypothesize that origin activity and/or timing may be regulated by a higher-order organization of chromatin in the nucleus.

Primary sequence and nucleosome-depleted patterns specify replication origins in a genome with epigenetic centromeres.

Centromeres were the first *bona fide* origins identified in the *C. albicans* genome, and they are epigenetically associated with centromere function.

Additionally, we mapped origins on chromosome arms and investigated their features. ORC binding sites are ubiquitous across the genome; however, according to the replication timing profile there are ~200 actively firing origins during the cell cycle. Thus, ORC binding is necessary, but not sufficient to specify an origin to fire. A similar situation also is observed in the *S. cerevisiae* genome; over 12000 ACS motifs are present, but only 400 of them are detected as active in S-phase (Nieduszynski, 2006).

Eaton *et al*. found that origins have a conserved pattern of nucleosome positioning in *S. cerevisiae*, and the positioned nucleosomes can specify origins along with ORC binding (Eaton et al., 2010a). Without prior knowledge of origin sequence motifs, we identified ORC-associated nucleosome-depleted patterns and predicted origins in a sequence-independent manner. I have performed DNA 2D gels to verify if the predicted origins can fire efficiently during the cell cycle. Several predicted origins were indeed identified as active origins. We failed to detect bubble-arc replication intermediates for a few predicted origins; however, we cannot rule out the possibility that these may be still functional, albeit less efficient, origins that we cannot detect due to the insensitivity of the 2D gel electrophoresis technique in *C. albicans*.

One of the obstacles to mapping replication origin function in *C. albicans* was the lack of an autonomous plasmid system. Early studies of origin sequences in yeast used circular plasmids to identify ARSs. Unfortunately,

circular plasmids in *C. albicans,* either with or without origin DNA, cannot be stably maintained because they rapidly integrate into the chromosomal DNA.

I developed a linear plasmid vector to examine the ARS function of origin DNA. Since we predicted origins in a sequence-independent manner, origin DNA was cloned onto the plasmid vectors to ask if the origin DNA conferred ARS function. Strikingly, a linear plasmid vector carrying origin DNA was able to replicate autonomously in *C. albicans* and it remained autonomous for several consecutive passages.

I also constructed 'mini-ARS' libraries to identify the minimal sequences for ARS function from three different active origins. Several ARS-containing fragments spanning 70-200 bp were identified in our screens. Interestingly, multiple small ARS containing fragments could be isolated from individual tested origin DNA regions (1-3 kb). This strongly suggested that a sequence feature such as an ACS motif may confer ARS function.

Based on these findings, I performed a genome-wide ARS screen using the linear plasmid system. From this non-saturating screen, four isolated genomic ARSs overlapped with ARS-containing fragments identified in the independent mini-ARS screen. Moreover, an AC-rich 15 bp ACS motif was identified. This ACS motif is critical for origin activity, as mutagenizing the ACS motif disrupted ARS function.

Surprisingly, I also found an AC-rich motif in origins identified from the sequence-independent prediction method. The nucleotide composition among

the three motifs are similar with a different weight of conservation at each nucleotide. The ACS motif for origin activity appears to be very prevalent in the genome. However, to specify origin function in the chromosomal context, well-positioned nucleosomes are necessary as well. I compared the genomic ARS locations with and without positioned nucleosome patterns using DNA 2D gels and found that the ACS motif is not sufficient to fire an origin. Rather, in the chromosomal context both the ACS motif and a well positioned nucleosome pattern appear to be necessary. Nonetheless, I discovered a specific primary sequence motif that is necessary for origin function in a genome containing epigenetic centromeres.

Origin sequences differ in different species. The origin sequences have evolved rapidly despite strong conservation of the replication factors. For example, *S. cerevisiae* has a 11 bp T-rich motif, and *K. lactis* has a 50 bp ACS motif. In metazoans, origins have been identified at transcription start sites, but definitive sequence motifs are not apparent, with depleted nucleosome patterns and the open chromatin likely specifying origins. Intriguingly, the ACS motif in the *C. albicans* genome has a similar composition to several well-known transcription factor binding motifs, including Fkh2. This suggests that not only the sequence specificity of pre-RC/ORC binding but also *trans*-acting factors may contribute to origin identity.

Future work should readily identify more of the nucleotide reuqirements for the ACS motif. A mutagenesis screening method mimicking the 'mini-ARS' screen should be used to identify those nucleotides critical for origin activity at the sequence level. In parallel, we modifying the current ACS motif to increase ARS activity on the plasmid and to reduce the chromosomal integration frequencies will increase the utility of the plasmids for conventional cloning approaches. Additionally, we are further exploring how the replication origins are organized at both a sequence and chromatin level.

Taken together, I have comprehensively identified and characterized replication origins in the *C. albicans* genome by combining computational and experimental approaches. I identified two distinct classes of origins in the *C. albicans* genome: centromeric origins and chromosomal origins. Centromeric origins are activated earliest during the cell cycle and are epigenetically associated with centromere function and Cse4 chromatin; chromosomal origins have a unique AC-rich motif and conserved patterns of nucleosome positioning to specify origin activity. The identification of origin sequence features also allowed us to develop the first plasmid shuttle vector as a genetic tool for working with *C. albicans*. All the work presented here will facilitate progress in the study of *Candida* biology, making *C. albicans* a promising model organism for studyiing fungal pathogenesis and eukaryotic genome organization.

References

Abdurashidova, G., Deganuto, M., Klima, R., Riva, S., Biamonti, G., Giacca, M., and Falaschi, A. (2000). Start sites of bidirectional DNA synthesis at the human lamin B2 origin. Science *287*, 2023–2026.

Abdurashidova, G., Danailov, M.B., Ochem, A., Triolo, G., Djeliova, V., Radulescu, S., Vindigni, A., Riva, S., and Falaschi, A. (2003). Localization of proteins bound to a replication origin of human DNA along the cell cycle. Embo J *22*, 4294–4303.

Aladjem, M.I., Rodewald, L.W., Kolman, J.L., and Wahl, G.M. (1998). Genetic dissection of a mammalian replicator in the human beta-globin locus. Science *281*, 1005–1009.

Aladjem, M.I. (2007). Replication in context: dynamic regulation of DNA replication patterns in metazoans. Nature Reviews Genetics *8*, 588–600.

Allshire, R.C., and Karpen, G.H. (2008). Epigenetic regulation of centromeric chromatin: old dogs, new tricks? Nature Reviews Genetics *9*, 923–937.

Altman, A.L., and Fanning, E. (2004). Defined sequence modules and an architectural element cooperate to promote initiation at an ectopic mammalian chromosomal replication origin. Mol Cell Biol *24*, 4138–4150.

Amor, D.J., and Choo, K.H.A. (2002). Neocentromeres: role in human disease, evolution, and centromere study. Am. J. Hum. Genet. *71*, 695–714.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res *37*, W202–W208.

Banerjee, S.N., Emori, T.G., Culver, D.H., Gaynes, R.P., Jarvis, W.R., Horan, T., Edwards, J.R., Tolson, J., Henderson, T., and Martone, W.J. (1991). Secular trends in nosocomial primary bloodstream infections in the United States, 1980-1989. National Nosocomial Infections Surveillance System. Am. J. Med. *91*, 86S–89S.

Baum, M., Ngan, V.K., and Clarke, L. (1994). The centromeric K-type repeat and the central core are together sufficient to establish a functional Schizosaccharomyces pombe centromere. Mol Biol Cell *5*, 747–761.

Beckerman, J., Chibana, H., Turner, J., and Magee, P.T. (2001). 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.

Bell, S.P., and Stillman, B. (1992). ATP-dependent recognition of eukaryotic origins of DNA replication by a multiprotein complex. Nature *357*, 128–134.

Bell, S.P., Kobayashi, R., and Stillman, B. (1993). Yeast origin recognition complex functions in transcription silencing and DNA replication. Science *262*, 1844–1849.

Bell, S.P. (2002). The origin recognition complex: from simple origins to complex functions. Genes Dev.

Bennett, R.J., and Johnson, A.D. (2003). Completion of a parasexual cycle in Candida albicans by induced chromosome loss in tetraploid strains. Embo J *22*, 2505–2515.

Bensen, E.S., Filler, S.G., and Berman, J. (2002). A forkhead transcription factor is important for true hyphal as well as yeast morphogenesis in Candida albicans. Eukaryotic Cell *1*, 787–798.

Berbenetz, N.M., Nislow, C., and Brown, G.W. (2010). Diversity of Eukaryotic DNA Replication Origins Revealed by Genome-Wide Analysis of Chromatin Structure. PLoS Genet *6*, e1001092.

Besnard, E., Babled, A., Lapasset, L., Milhavet, O., Parrinello, H., Dantec, C., Marin, J.-M., and Lemaitre, J.-M. (2012). Unraveling cell type-specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. Nat Struct Mol Biol *19*, 837–844.

Black, B.E., and Cleveland, D.W. (2011). Epigenetic centromere propagation and the nature of CENP-a nucleosomes. Cell *144*, 471–479.

Black, B.E., Brock, M.A., Bédard, S., Woods, V.L., and Cleveland, D.W. (2007). An epigenetic mark generated by the incorporation of CENP-A into centromeric nucleosomes. Proc Natl Acad Sci USA *104*, 5008–5013.

Branzei, D., and Foiani, M. (2010). Maintaining genome stability at the replication fork. Nat Rev Mol Cell Biol *11*, 208–219.

Breier, A.M., Chatterji, S., and Cozzarelli, N.R. (2004). Prediction of Saccharomyces cerevisiae replication origins. Genome Biol *5*, R22.

Brewer, B.J. (1994). Intergenic DNA and the sequence requirements for replication initiation in eukaryotes. Curr. Opin. Genet. Dev. *4*, 196–202.

Brewer, B.J., and Fangman, W.L. (1987). The localization of replication origins on ARS plasmids in S. cerevisiae. Cell *51*, 463–471.

Broach, J.R., Li, Y.Y., Feldman, J., Jayaram, M., Abraham, J., Nasmyth, K.A., and Hicks, J.B. (1983). Localization and sequence analysis of yeast origins of DNA replication. Cold Spring Harbor Symposia on Quantitative Biology *47 Pt 2*, 1165–1173.

Burrack, L.S., and Berman, J. (2012a). Flexibility of centromere and kinetochore structures. Trends Genet *28*, 204–212.

Burrack, L.S., and Berman, J. (2012b). Neocentromeres and epigenetically inherited features of centromeres. Chromosome Res *20*, 607–619.

Cannon, R.D., Jenkinson, H.F., and Shepherd, M.G. (1990). Isolation and nucleotide sequence of an autonomously replicating sequence (ARS) element functional in Candida albicans and Saccharomyces cerevisiae. Mol. Gen. Genet. *221*, 210–218.

Cayrou, C., Coulombe, P., Vigneron, A., Stanojcik, S., Ganier, O., Peiffer, I., Rivals, E., Puy, A., Laurent-Chabalier, S., Desprat, R., et al. (2011). Genome-scale analysis of metazoan replication origins reveals their organization in specific but flexible sites defined by conserved features. Genome Res.

Chan, C.S., and Tye, B.K. (1980). Autonomously replicating sequences in Saccharomyces cerevisiae. Proc Natl Acad Sci USA *77*, 6329–6333.

Chee, M.K., and Haase, S.B. (2012). New and Redesigned pRS Plasmid Shuttle Vectors for Genetic Manipulation of Saccharomycescerevisiae. G3 (Bethesda) *2*, 515–526.

Chen, Y., Baker, R.E., Keith, K.C., Harris, K., Stoler, S., and Fitzgerald-Hayes, M. (2000). The N Terminus of the Centromere H3-Like Protein Cse4p Performs an Essential Function Distinct from That of the Histone Fold Domain. … And Cellular Biology.

Chesnokov, I., Remus, D., and Botchan, M. (2001). Functional analysis of mutant and wild-type Drosophila origin recognition complex. Proc Natl Acad Sci USA *98*, 11997–12002.

Chuang, R.Y., and Kelly, T.J. (1999). The fission yeast homologue of Orc4p binds to replication origin DNA via multiple AT-hooks. Proc Natl Acad Sci USA *96*, 2656–2661.

Clarke, L., and Carbon, J. (1980). Isolation of a yeast centromere and construction of functional small circular chromosomes. Nature *287*, 504–509.

Clarke, L., and Carbon, J. (1983). Genomic substitutions of centromeres in Saccharomyces cerevisiae. Nature *305*, 23–28.

Cleveland, D.W., Mao, Y., and Sullivan, K.F. (2003). Centromeres and kinetochores: from epigenetics to mitotic checkpoint signaling. Cell *112*, 407–421.

Clyne, R.K., and Kelly, T.J. (1995). Genetic analysis of an ARS element from the fission yeast Schizosaccharomyces pombe. Embo J *14*, 6348–6357.

Copenhaver, G.P., Nickel, K., Kuromori, T., Benito, M.I., Kaul, S., Lin, X., Bevan, M., Murphy, G., Harris, B., Parnell, L.D., et al. (1999). Genetic definition and sequence analysis of Arabidopsis centromeres. Science *286*, 2468–2474.

Coste, A., Turner, V., Ischer, F., Morschhäuser, J., Forche, A., Selmecki, A., Berman, J., Bille, J., and Sanglard, D. (2006). A mutation in Tac1p, a transcription factor regulating CDR1 and CDR2, is coupled with loss of heterozygosity at chromosome 5 to mediate antifungal resistance in Candida albicans. Genetics *172*, 2139–2156.

Crampton, A., Chang, F., Pappas, D.L., Frisch, R.L., and Weinreich, M. (2008). An ARS element inhibits DNA replication through a SIR2-dependent mechanism. Mol Cell *30*, 156–166.

Dai, J. (2005). DNA replication origins in the Schizosaccharomyces pombe genome. Proc Natl Acad Sci USA *102*, 337–342.

Dani, G.M., and Zakian, V.A. (1983). Mitotic and meiotic stability of linear plasmids in yeast. Proc Natl Acad Sci USA *80*, 3406–3410.

Di Rienzi, S.C., Lindstrom, K.C., Mann, T., Noble, W.S., Raghuraman, M.K., and Brewer, B.J. (2012). Maintaining replication origins in the face of genomic change. Genome Res *22*, 1940–1952.

Diffley, J.F.X., and Cocker, J.H. (1992). Protein-DNA interactions at a yeast replication origin. Nature *357*, 169–172.

Dijkwel, P.A., and Hamlin, J.L. (1995). The Chinese hamster dihydrofolate reductase origin consists of multiple potential nascent-strand start sites. Mol Cell Biol *15*, 3023–3031.

Du, W., Coaker, M., Sobel, J.D., and Akins, R.A. (2004). Shuttle vectors for Candida albicans: control of plasmid copy number and elevated expression of cloned genes. Curr. Genet. *45*, 390–398.

Dunleavy, E.M., Roche, D., Tagami, H., Lacoste, N., Ray-Gallet, D., Nakamura, Y., Daigo, Y., Nakatani, Y., and Almouzni-Pettinotti, G. (2009). HJURP is a cell-cycle-dependent maintenance and deposition factor of CENP-A at centromeres. Cell *137*, 485–497.

Eaton, M.L., Galani, K., Kang, S., Bell, S.P., and MacAlpine, D.M. (2010a). Conserved nucleosome positioning defines replication origins. *24*, 748–753.

Eaton, M.L., Prinz, J.A., MacAlpine, H.K., Tretyakov, G., Kharchenko, P.V., and MacAlpine, D.M. (2010b). Chromatin signatures of the Drosophila replication program. Genome Res.

Feldmann, H., Olah, J., and Friedenreich, H. (1981). Sequence of a yeast DNA fragment containing a chromosomal replicator and a tRNA Glu 3 gene. Nucleic Acids Res *9*, 2949–2959.

Field, Y., Kaplan, N., Fondufe-Mittendorf, Y., Moore, I.K., Sharon, E., Lubling, Y., Widom, J., and Segal, E. (2008). Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. PLoS Comput Biol *4*, e1000216.

Fitzgerald-Hayes, M., Clarke, L., and Carbon, J. (1982). Nucleotide sequence comparisons and functional analysis of yeast centromere DNAs. Cell *29*, 235–244.

Flemming, W. (1880). Beiträge zur Kenntniss der Zelle und Ihrer Lebenserscheinungen. Archiv F. Mikrosk. Anatomie. *18*, 151–259.

Folco, H.D., Pidoux, A.L., Urano, T., and Allshire, R.C. (2008). Heterochromatin and RNAi are required to establish CENP-A chromatin at centromeres. Science *319*, 94–97.

Foltz, D.R., Jansen, L.E.T., Bailey, A.O., Yates, J.R., III, Bassett, E.A., Wood, S., Black, B.E., and Cleveland, D.W. (2009). Centromere-Specific Assembly of CENP-A Nucleosomes Is Mediated by HJURP. Cell *137*, 472–484.

Forche, A., Alby, K., Schaefer, D., Johnson, A.D., Berman, J., and Bennett, R.J. (2008). The parasexual cycle in Candida albicans provides an alternative pathway to meiosis for the formation of recombinant strains. PLoS Biol *6*, e110.

Foss, M., McNally, F., Laurenson, P., and Rine, J. (1993). Origin recognition complex (ORC) in transcriptional silencing and DNA replication in S. cerevisiae. Science *262*, 1838–1844.

Franz, R., Kelly, S.L., Lamb, D.C., Kelly, D.E., Ruhnke, M., and Morschhäuser, J. (1998). Multiple molecular mechanisms contribute to a stepwise development of fluconazole resistance in clinical Candida albicans strains. Antimicrob. Agents Chemother. *42*, 3065–3072.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw *33*, 1–22.

Futcher, A.B. (1988). The 2 μm circle plasmid ofSaccharomyces cerevisiae. Yeast *4*, 27–40.

Gilbert, D.M. (2010). Evaluating genome-scale approaches to eukaryotic DNA replication. Nature Reviews Genetics *11*, 673–684.

Gilbert, D.M. (2012). Replication origins run (ultra) deep. Nat Struct Mol Biol *19*, 740–742.

Hahnenberger, K.M., Carbon, J., and Clarke, L. (1991). Identification of DNA regions required for mitotic and meiotic functions within the centromere of Schizosaccharomyces pombe chromosome I. Mol Cell Biol *11*, 2206–2215.

Hansen, R.S., Thomas, S., Sandstrom, R., Canfield, T.K., Thurman, R.E., Weaver, M., Dorschner, M.O., Gartler, S.M., and Stamatoyannopoulos, J.A. (2010). Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. Proc Natl Acad Sci USA *107*, 139–144.

Herreros, E., García-Sáez, M.I., Nombela, C., and Sánchez, M. (1992). A reorganized Candida albicans DNA sequence promoting homologous non-integrative genetic transformation. Mol Microbiol *6*, 3567–3574.

Heyer, W.D., Sipiczki, M., and Kohli, J. (1986). Replicating plasmids in Schizosaccharomyces pombe: improvement of symmetric segregation by a new genetic element. Mol Cell Biol *6*, 80–89.

Hickman, M.A., Zeng, G., Forche, A., Hirakawa, M.P., Abbey, D., Harrison, B.D., Wang, Y.-M., Su, C.-H., Bennett, R.J., Wang, Y., et al. (2013). The “obligate diploid” Candida albicans forms mating-competent haploids. Nature *494*, 55–59.

Hieter, P., Pridmore, D., Hegemann, J.H., Thomas, M., Davis, R.W., and Philippsen, P. (1985). Functional selection and analysis of yeast centromeric DNA. Cell *42*, 913–921.

Holland, A.J., and Cleveland, D.W. (2012). Losing balance: the origin and impact of aneuploidy in cancer. EMBO Rep *13*, 501–514.

Hsiao, C.L., and Carbon, J. (1979). High-frequency transformation of yeast by plasmids containing the cloned yeast ARG4 gene. Proc Natl Acad Sci USA *76*, 3829–3833.

Ishii, K., Ogiyama, Y., Chikashige, Y., Soejima, S., Masuda, F., Kakuma, T., Hiraoka, Y., and Takahashi, K. (2008). Heterochromatin integrity affects chromosome reorganization after centromere dysfunction. Science *321*, 1088–1091.

Iwasaki, O., and Noma, K.-I. (2012). Global genome organization mediated by

RNA polymerase III-transcribed genes in fission yeast. Gene *493*, 195–200.

Jacob, F., Brenner, S., and Cuzin, F. (1963). On the Regulation of DNA Replication in Bacteria. Cold Spring Harbor Symposia on Quantitative Biology *28*, 329–348.

Jiang, W., Middleton, K., Yoon, H.J., Fouquet, C., and Carbon, J. (1993). An essential yeast protein, CBF5p, binds in vitro to centromeres and microtubules. Mol Cell Biol *13*, 4884–4893.

Jones, T. (2004). The diploid genome sequence of Candida albicans. Proc Natl Acad Sci USA *101*, 7329–7334.

Kaplan, N., Moore, I.K., Fondufe-Mittendorf, Y., Gossett, A.J., Tillo, D., Field, Y., LeProust, E.M., Hughes, T.R., Lieb, J.D., Widom, J., et al. (2009). The DNA-encoded nucleosome organization of a eukaryotic genome. Nature *458*, 362–366.

Karnani, N., Taylor, C.M., Malhotra, A., and Dutta, A. (2010). Genomic study of replication initiation in human chromosomes reveals the influence of transcription regulation and chromatin structure on origin selection. Mol Biol Cell *21*, 393–404.

Keith, K.C., and Fitzgerald-Hayes, M. (2000). CSE4 Genetically Interacts With the Saccharomyces cerevisiae Centromere DNA Elements CDE I and CDE II but Not CDE III: Implications for the Path of the Centromere DNA Around a Cse4p Variant Nucleosome. Genetics.

Ketel, C., Wang, H.S.W., McClellan, M., Bouchonville, K., Selmecki, A., Lahav, T., Gerami-Nejad, M., and Berman, J. (2009). Neocentromeres form efficiently at multiple possible loci in Candida albicans. PLoS Genet *5*, e1000400.

Kim, S.M. (2003). Early-replicating heterochromatin. Genes Dev *17*, 330–335.

Klemm, R.D., Austin, R.J., and Bell, S.P. (1997). Coordinate binding of ATP and origin DNA regulates the ATPase activity of the origin recognition complex. Cell *88*, 493–502.

Knott, S.R.V., Peace, J.M., Ostrow, A.Z., Gan, Y., Rex, A.E., Viggiani, C.J., Tavaré, S., and Aparicio, O.M. (2012). Forkhead Transcription Factors Establish Origin Timing and Long-Range Clustering in S. cerevisiae. Cell *148*, 99–111.

Knott, S.R.V., Viggiani, C.J., Tavaré, S., and Aparicio, O.M. (2009). Genome-wide replication profiles indicate an expansive role for Rpd3L in regulating replication initiation timing or efficiency, and reveal genomic loci of Rpd3 function in Saccharomyces cerevisiae. Genes Dev *23*, 1077–1090.

Koren, A., Soifer, I., and Barkai, N. (2010a). MRC1-dependent scaling of the

budding yeast DNA replication timing program. Genome Res *20*, 781–790.

Koren, A., Tsai, H.-J., Tirosh, I., Burrack, L.S., Barkai, N., and Berman, J. (2010b). Epigenetically-inherited centromere and neocentromere DNA replicates earliest in S-phase. PLoS Genet *6*, e1001068.

Krügel, H., Fiedler, G., Haupt, I., Sarfert, E., and Simon, H. (1988). Analysis of the nourseothricin-resistance gene (nat) of Streptomyces noursei. Gene *62*, 209–217.

Krysan, P.J., Smith, J.G., and Calos, M.P. (1993). Autonomous replication in human cells of multimers of specific human and bacterial DNA sequences. Mol Cell Biol *13*, 2688–2696.

Lederberg, J. (1952). Cell genetics and hereditary symbiosis. Physiol Rev.

Lee, D.G., and Bell, S.P. (1997). Architecture of the yeast origin recognition complex bound to origins of DNA replication. Mol Cell Biol *17*, 7159–7168.

Lee, J.K., Moon, K.Y., Jiang, Y., and Hurwitz, J. (2001). The Schizosaccharomyces pombe origin recognition complex interacts with multiple AT-rich regions of the replication origin DNA by means of the AT-hook domains of the spOrc4 protein. Proc Natl Acad Sci USA *98*, 13589–13594.

Lee, K.K., Maccallum, D.M., Jacobsen, M.D., Walker, L.A., Odds, F.C., Gow, N.A.R., and Munro, C.A. (2012). Elevated cell wall chitin in Candida albicans confers echinocandin resistance in vivo. Antimicrob. Agents Chemother. *56*, 208–217.

Li, B., Carey, M., and Workman, J.L. (2007). The role of chromatin during transcription. Cell *128*, 707–719.

Liachko, I., Youngblood, R.A., Keich, U., and Dunham, M.J. (2012). High-resolution mapping, characterization, and optimization of autonomously replicating sequences in yeast. Genome Res.

Liachko, I., Bhaskar, A., Lee, C., Chung, S.C.C., Tye, B.-K., and Keich, U. (2010). A Comprehensive Genome-Wide Map of Autonomously Replicating Sequences in a Naive Genome. PLoS Genet *6*, e1000946.

Liachko, I., Tanaka, E., Cox, K., Chung, S.C.C., Yang, L., Seher, A., Hallas, L., Cha, E., Kang, G., Pace, H., et al. (2011). Novel features of ARS selection in budding yeast Lachancea kluyveri. BMC Genomics *12*, 633.

Linskens, M.H., and Huberman, J.A. (1988). Organization of replication of ribosomal DNA in Saccharomyces cerevisiae. Mol Cell Biol *8*, 4927–4935.

Lipford, J.R., and Bell, S.P. (2001). Nucleosomes positioned by ORC facilitate the initiation of DNA replication. Mol Cell *7*, 21–30.

Liu, G., Malott, M., and Leffak, M. (2003). Multiple functional elements comprise a Mammalian chromosomal replicator. Mol Cell Biol *23*, 1832–1842.

MacAlpine, H.K., Gordân, R., Powell, S.K., Hartemink, A.J., and MacAlpine, D.M. (2010). Drosophila ORC localizes to open chromatin and marks sites of cohesin complex loading. Genome Res *20*, 201–211.

Maggert, K.A., and Karpen, G.H. (2001). The activation of a neocentromere in Drosophila requires proximity to an endogenous centromere. Genetics *158*, 1615–1628.

Malik, H.S., and Henikoff, S. (2009). Major evolutionary transitions in centromere complexity. Cell *138*, 1067–1082.

Marahrens, Y., and Stillman, B. (1992). A yeast chromosomal origin of DNA replication defined by multiple functional elements. Science *255*, 817–823.

Marichal, P., Koymans, L., Willemsens, S., Bellens, D., Verhasselt, P., Luyten, W., Borgers, M., Ramaekers, F.C.S., Odds, F.C., and Bossche, H.V. (1999). Contribution of mutations in the cytochrome P450 14α-demethylase (Erg11p, Cyp51p) to azole resistance in Candida albicans. ….

Marshall, O.J., Chueh, A.C., Wong, L.H., and Choo, K.H.A. (2008). Neocentromeres: new insights into centromere structure, disease development, and karyotype evolution. Am. J. Hum. Genet. *82*, 261–282.

Masai, H., Matsumoto, S., You, Z., Yoshizawa-Sugata, N., and Oda, M. (2010). Eukaryotic chromosome DNA replication: where, when, and how? Annu Rev Biochem *79*, 89–130.

Maundrell, K., Hutchison, A., and Shall, S. (1988). Sequence analysis of ARS elements in fission yeast. Embo J *7*, 2203–2209.

Mavor, A., Thewes, S., and Hube, B. (2005). Systemic Fungal Infections Caused by Candida Species: Epidemiology, Infection Process and Virulence Attributes. Cdt *6*, 863–874.

Mechali, M. (2010). Eukaryotic DNA replication origins: many choices for appropriate answers. Nat Rev Mol Cell Biol *11*, 728–738.

Mellone, B.G., Grive, K.J., Shteyn, V., Bowers, S.R., Oderberg, I., and Karpen, G.H. (2011). Assembly of Drosophila centromeric chromatin proteins during mitosis. PLoS Genet *7*, e1002068.

Micklem, G., Rowley, A., Harwood, J., Nasmyth, K., and Diffley, J.F. (1993). Yeast origin recognition complex is involved in DNA replication and transcriptional silencing. Nature *366*, 87–89.

Miller, L.G., Hajjeh, R.A., and Edwards, J.E. (2001). Estimating the Cost of Nosocomial Candidemia in the United States. Clinical Infectious Diseases *32*, 1110–1110.

Morris, C.A., and Moazed, D. (2007). Centromere Assembly and Propagation. Cell *128*, 647–650.

Morschhäuser, J., Barker, K.S., Liu, T.T., BlaB-Warmuth, J., Homayouni, R., and Rogers, P.D. (2007). The transcription factor Mrr1p controls expression of the MDR1 efflux pump and mediates multidrug resistance in Candida albicans. PLoS Pathog. *3*, e164.

Muller, H.J. (1922). JSTOR: The American Naturalist, Vol. 56, No. 642 (Jan. - Feb., 1922), pp. 32-50. The American Naturalist.

Murray, A.W., and Szostak, J.W. (1983). Construction of artificial chromosomes in yeast. Nature *305*, 189–193.

Newlon, C.S., and Theis, J.F. (1993). The structure and function of yeast ARS elements. Curr. Opin. Genet. Dev. *3*, 752–758.

Ngan, V.K., and Clarke, L. (1997). The centromere enhancer mediates centromere activation in Schizosaccharomyces pombe. Mol Cell Biol *17*, 3305–3314.

Nieduszynski, C.A. (2006). Genome-wide identification of replication origins in yeast by comparative genomics. Genes Dev *20*, 1874–1879.

Niwa, O., Matsumoto, T., Chikashige, Y., and Yanagida, M. (1989). Characterization of Schizosaccharomyces pombe minichromosome deletion derivatives and a functional allocation of their centromere. Embo J *8*, 3045–3052.

Noma, K.-I., Cam, H.P., Maraia, R.J., and Grewal, S.I.S. (2006). A role for TFIIIC transcription factor complex in genome organization. Cell *125*, 859–872.

Okada, T., Ohzeki, J.-I., Nakano, M., Yoda, K., Brinkley, W.R., Larionov, V., and Masumoto, H. (2007). CENP-B controls centromere formation depending on the chromatin context. Cell *131*, 1287–1300.

Okuno, Y., Satoh, H., Sekiguchi, M., and Masukata, H. (1999). Clustered adenine/thymine stretches are essential for function of a fission yeast replication origin. Mol Cell Biol *19*, 6699–6709.

Padmanabhan, S., Thakur, J., Siddharthan, R., and Sanyal, K. (2008). Rapid evolution of Cse4p-rich centromeric DNA sequences in closely related pathogenic yeasts, Candida albicans and Candida dubliniensis. Proc Natl Acad Sci USA *105*, 19797–19802.

Parent, S.A., Fenimore, C.M., and Bostian, K.A. (1985). Vector systems for the expression, analysis and cloning of DNA sequences in S. cerevisiae. Yeast *1*, 83–138.

Pavlov, Y.I., Newlon, C.S., and Kunkel, T.A. (2002). Yeast origins establish a strand bias for replicational mutagenesis. Mol Cell *10*, 207–213.

Pfaller, M.A., and Diekema, D.J. (2007). Epidemiology of invasive candidiasis: a persistent public health problem. Clin. Microbiol. Rev. *20*, 133–163.

Pidoux, A.L., and Allshire, R.C. (2004). Kinetochore and heterochromatin domains of the fission yeast centromere. Chromosome Res *12*, 521–534.

Pohl, T.J., Brewer, B.J., and Raghuraman, M.K. (2012). Functional centromeres determine the activation time of pericentric origins of DNA replication in Saccharomyces cerevisiae. PLoS Genet *8*, e1002677.

Polizzi, C., Carbon, J., and Clarke, L. (1989). Construction of functional artificial minichromosomes in the fission yeast Schizosaccharomyces pombe. Proceedings of the ….

Raman, S.B., Nguyen, M.H., Zhang, Z., Cheng, S., Jia, H.Y., Weisner, N., Iczkowski, K., and Clancy, C.J. (2006). Candida albicans SET1 encodes a histone 3 lysine 4 methyltransferase that contributes to the pathogenesis of invasive candidiasis. Mol Microbiol *60*, 697–709.

Rando, O.J., and Winston, F. (2012). Chromatin and transcription in yeast. Genetics *190*, 351–387.

Ranjitkar, P., Press, M.O., Yi, X., Baker, R., MacCoss, M.J., and Biggins, S. (2010). An E3 ubiquitin ligase prevents ectopic localization of the centromeric histone H3 variant via the centromere targeting domain. Mol Cell *40*, 455–464.

Rao, H., and Stillman, B. (1995). The origin recognition complex interacts with a bipartite DNA binding site within yeast replicators. Proc Natl Acad Sci USA *92*, 2224–2228.

Rao, H., Marahrens, Y., and Stillman, B. (1994). Functional conservation of multiple elements in yeast chromosomal replicators. Mol Cell Biol *14*, 7643–7651.

Rustchenko, E. (2007). Chromosome instability in Candida albicans. FEMS

Yeast Res. *7*, 2–11.

Ryba, T., Battaglia, D., Pope, B.D., Hiratani, I., and Gilbert, D.M. (2011). Genome-scale analysis of replication timing: from bench to bioinformatics. Nat Protoc *6*, 870–895.

Sanglard, D., Ischer, F., Monod, M., and Bille, J. (1997). Cloning of Candida albicans genes conferring resistance to azole antifungal agents: characterization of CDR2, a new multidrug ABC transporter gene. Microbiology (Reading, Engl.) *143 (Pt 2)*, 405–416.

Sanglard, D., Kuchler, K., Ischer, F., Pagani, J.L., Monod, M., and Bille, J. (1995). Mechanisms of resistance to azole antifungal agents in Candida albicans isolates from AIDS patients involve specific multidrug transporters. Antimicrob. Agents Chemother. *39*, 2378–2386.

Sanglard, D., Ischer, F., Koymans, L., and Bille, J. (1998). Amino Acid Substitutions in the Cytochrome P-450 Lanosterol 14α-Demethylase (CYP51A1) from Azole-Resistant Candida albicans Clinical Isolates Contribute to Resistance to Azole Antifungal Agents.

Sanyal, K., and Carbon, J. (2002). The CENP-A homolog CaCse4p in the pathogenic yeast Candida albicans is a centromere protein essential for chromosome transmission. Proc Natl Acad Sci USA *99*, 12969–12974.

Sanyal, K., Baum, M., and Carbon, J. (2004). Centromeric DNA sequences in the pathogenic yeast Candida albicans are all different and unique. Proc Natl Acad Sci USA *101*, 11374–11379.

Schübeler, D., Scalzo, D., Kooperberg, C., van Steensel, B., Delrow, J., and Groudine, M. (2002). Genome-wide DNA replication profile for Drosophila melanogaster: a link between transcription and replication timing. Nat. Genet. *32*, 438–442.

Segurado, M., de Luis, A., and Antequera, F. (2003). Genome-wide distribution of DNA replication origins at A+T-rich islands in Schizosaccharomyces pombe. EMBO Rep *4*, 1048–1053.

Selmecki, A., Forche, A., and Berman, J. (2006). Aneuploidy and isochromosome formation in drug-resistant Candida albicans. Science *313*, 367–370.

Selmecki, A., Forche, A., and Berman, J. (2010). Genomic plasticity of the human fungal pathogen Candida albicans. Eukaryotic Cell *9*, 991–1008.

Selmecki, A., Gerami-Nejad, M., Paulson, C., Forche, A., and Berman, J. (2008). An isochromosome confers drug resistance in vivo by amplification of two genes,

ERG11 and TAC1. Mol Microbiol *68*, 624–641.

Sernova, N.V., and Gelfand, M.S. (2008). Identification of replication origins in prokaryotic genomes. Brief. Bioinformatics *9*, 376–391.

Shor, E., Warren, C.L., Tietjen, J., Hou, Z., Müller, U., Alborelli, I., Gohard, F.H., Yemm, A.I., Borisov, L., Broach, J.R., et al. (2009). The origin recognition complex interacts with a subset of metabolic genes tightly linked to origins of replication. PLoS Genet *5*, e1000755.

Siikala, E., Rautemaa, R., Richardson, M., Saxen, H., Bowyer, P., and Sanglard, D. (2010). Persistent Candida albicans colonization and molecular mechanisms of azole resistance in autoimmune polyendocrinopathy-candidiasis-ectodermal dystrophy (APECED) patients. J. Antimicrob. Chemother. *65*, 2505–2513.

Siow, C.C., Nieduszynska, S.R., Müller, C.A., and Nieduszynski, C.A. (2012). OriDB, the DNA replication origin database updated and extended. Nucleic Acids Res *40*, D682–D686.

Steiner, N.C., and Clarke, L. (1994). A novel epigenetic effect can alter centromere function in fission yeast. Cell *79*, 865–874.

Stevenson, J.B., and Gottschling, D.E. (1999). Telomeric chromatin modulates replication timing near chromosome ends. Genes Dev *13*, 146–151.

Stimpson, K.M., and Sullivan, B.A. (2010). Epigenomics of centromere assembly and function. Curr Opin Cell Biol *22*, 772–780.

Stimpson, K.M., Matheny, J.E., and Sullivan, B.A. (2012). Dicentric chromosomes: unique models to study centromere function and inactivation. Chromosome Res *20*, 595–605.

Stinchcomb, D.T., Struhl, K., and Davis, R.W. (1979). Isolation and characterisation of a yeast chromosomal replicator. Nature *282*, 39–43.

Storchová, Z., and Pellman, D. (2004). From polyploidy to aneuploidy, genome instability and cancer. Nat Rev Mol Cell Biol *5*, 45–54.

Struhl, K., Stinchcomb, D.T., Scherer, S., and Davis, R.W. (1979). High-frequency transformation of yeast: autonomous replication of hybrid DNA molecules. Proc Natl Acad Sci USA *76*, 1035–1039.

Sullivan, B.A., and Karpen, G.H. (2004). Centromeric chromatin exhibits a histone modification pattern that is distinct from both euchromatin and heterochromatin. Nat Struct Mol Biol *11*, 1076–1083.

Theis, J.F., and Newlon, C.S. (1997). The ARS309 chromosomal replicator of

Saccharomyces cerevisiae depends on an exceptional ARS consensus sequence. Proc Natl Acad Sci USA *94*, 10786–10791.

Theis, J.F., Yang, C., Schaefer, C.B., and Newlon, C.S. (1999). DNA sequence and functional analysis of homologous ARS elements of Saccharomyces cerevisiae and S. carlsbergensis. Genetics *152*, 943–952.

Trick, W.E., Fridkin, S.K., Edwards, J.R., Hajjeh, R.A., Gaynes, R.P., the National Nosocomial Infections Surveillance System Hospitals (2002). Secular Trend of Hospital-Acquired Candidemia among Intensive Care Unit Patients in the United States during 1989–1999. Clinical Infectious Diseases *35*, 627–630.

Tsankov, A.M., Thompson, D.A., Socha, A., Regev, A., and Rando, O.J. (2010). The role of nucleosome positioning in the evolution of gene regulation. PLoS Biol *8*, e1000414.

Tschumper, G., and Carbon, J. (1980). Sequence of a yeast DNA fragment containing a chromosomal replicator and the TRP1 gene. Gene *10*, 157–166.

Unnikrishnan, A., Gafken, P.R., and Tsukiyama, T. (2010). Dynamic changes in histone acetylation regulate origins of DNA replication. Nat Struct Mol Biol *17*, 430–437.

van het Hoog, M., Rast, T.J., Martchenko, M., Grindle, S., Dignard, D., Hogues, H., Cuomo, C., Berriman, M., Scherer, S., Magee, B.B., et al. (2007). Assembly of the Candida albicans genome into sixteen supercontigs aligned on the eight chromosomes. Genome Biol *8*, R52.

Verdaasdonk, J.S., and Bloom, K. (2011). Centromeres: unique chromatin structures that drive chromosome segregation. *12*, 320–332.

Vogelauer, M., Rubbi, L., Lucas, I., Brewer, B.J., and Grunstein, M. (2002). Histone acetylation regulates the time of replication origin firing. Mol Cell *10*, 1223–1233.

Waltz, S.E., Trivedi, A.A., and Leffak, M. (1996). DNA replication initiates non-randomly at multiple sites near the c-myc gene in HeLa cells. Nucleic Acids Res *24*, 1887–1894.

White, T.C. (1997a). The presence of an R467K amino acid substitution and loss of allelic variation correlate with an azole-resistant lanosterol 14alpha demethylase in Candida albicans. Antimicrob. Agents Chemother.

White, T.C. (1997b). Increased mRNA levels of ERG16, CDR, and MDR1 correlate with increases in azole resistance in Candida albicans isolates from a patient infected with human immunodeficiency virus. Antimicrob. Agents Chemother.

Willard, H.F. (1985). Chromosome-specific organization of human alpha satellite DNA. Am. J. Hum. Genet. *37*, 524–532.

Wilson, R.B., Davis, D., and Mitchell, A.P. (1999). Rapid hypothesis testing with Candida albicans through gene disruption with short homology regions. J. Bacteriol. *181*, 1868–1874.

Wisplinghoff, H., Bischoff, T., Tallent, S.M., Seifert, H., Wenzel, R.P., and Edmond, M.B. (2004). Nosocomial bloodstream infections in US hospitals: analysis of 24,179 cases from a prospective nationwide surveillance study. Clinical Infectious Diseases *39*, 309–317.

Woods, J.P., and Goldman, W.E. (1992). In vivo generation of linear plasmids with addition of telomeric sequences by Histoplasma capsulatum. Mol Microbiol *6*, 3603–3610.

Woods, J.P., and Goldman, W.E. (1993). Autonomous replication of foreign DNA in Histoplasma capsulatum: role of native telomeric sequences. J. Bacteriol. *175*, 636–641.

Wyrick, J.J., Aparicio, J.G., Chen, T., Barnett, J.D., Jennings, E.G., Young, R.A., Bell, S.P., and Aparicio, O.M. (2001). Genome-wide distribution of ORC and MCM proteins in S. cerevisiae: high-resolution mapping of replication origins. Science *294*, 2357–2360.

Xu, J., Yanagisawa, Y., Tsankov, A.M., Hart, C., Aoki, K., Kommajosyula, N., Steinmann, K.E., Bochicchio, J., Russ, C., Regev, A., et al. (2012). Genome-wide identification and characterization of replication origins by deep sequencing. Genome Biol *13*, R27.

Zeitlin, S.G. (2010). Centromeres: the wild west of the post-genomic age. Epigenetics *5*, 34–40.

Zordan, R.E., Miller, M.G., Galgoczy, D.J., Tuch, B.B., and Johnson, A.D. (2007). Interlocking transcriptional feedback loops control white-opaque switching in Candida albicans. PLoS Biol *5*, e256.

Berman, J., Sudbery, P.E. (2002). Candida albicans: A molecular revolution built on lessons from budding yeast. *3*, 918–932.