Conservation of a DNA replication motif among phylogenetically distant budding yeast species

Haniam Maria[1], Shivali Kapoor[1], Tao Liu[2], and Laura N. Rusche[1*]

[1] Department of Biological Sciences, State University of New York at Buffalo, Buffalo NY, 14260, USA

[2] Department of Biostatistics and Bioinformatics, Roswell Park Comprehensive Cancer Center, Buffalo NY, 14203

*Author for Correspondence: Laura Rusche, Department of Biological Sciences, State University of New York at Buffalo, Buffalo NY, Telephone: 716-645-5198, Email: lrusche@buffalo.edu

Abstract

Eukaryotic DNA replication begins at genomic loci termed origins, which are bound by the Origin Recognition Complex (ORC). Although ORC is conserved across species, the sequence composition of origins is more varied. In the budding yeast *Saccharomyces cerevisiae*, the ORC binding motif consists of an A/T-rich 17 bp "extended ACS" sequence adjacent to a B1 element composed of two 3 bp motifs. Similar sequences occur at origins in closely related species, but it is not clear when this type of replication origin arose and whether it pre-dated a whole genome duplication that occurred around 100 million years ago in the budding yeast lineage. To address these questions, we identified the ORC binding sequences in the non-duplicated species *Torulaspora delbrueckii*. We used chromatin immunoprecipitation followed by sequencing and identified 190 ORC binding sites distributed across the eight *T. delbrueckii* chromosomes. Using these sites, we identified an ORC binding motif that is nearly identical to the known motif in *S. cerevisiae*. We also found that the *T. delbrueckii* ORC binding sites function as origins in *T. delbrueckii* when cloned onto a plasmid and that the motif is required for plasmid replication. Finally, we compared an *S. cerevisiae* origin with two *T. delbrueckii* ORC-binding sites and found that they conferred similar stabilities to a plasmid. These results reveal that the ORC-binding motif arose prior to the whole genome duplication and has been maintained for over 100 million years.

Key Words: autonomously replicating sequence (ARS), ARS consensus sequence (ACS), *Torulaspora delbrueckii*, replication origin, whole genome duplication (WGD), Origin Recognition Complex (ORC)

Significance

DNA replication is the process by which the genome doubles in preparation for cell division. In organisms such as animals, plants, and fungi, replication starts at multiple locations throughout the genome, ensuring efficient doubling. These initiation sites are defined differently across species, and in budding yeast, which are single-celled fungi, a specific DNA sequence is required. This study reveals that this DNA sequence can be remarkably stable over evolutionary time. We found that the same sequence is used in two yeasts, *Saccharomyces cerevisiae* and *Torulaspora delbrueckii*, whose last common ancestor existed over 100 million years ago.

Introduction

DNA replication is an essential step in the cell cycle that is required for inheritance of genetic information. Replication begins at genomic loci termed origins of replication, which are demarcated by the origin recognition complex (ORC) in eukaryotes. ORC is a six-subunit protein complex that binds directly to DNA (Bell and Stillman 1992). Although ORC is conserved across eukaryotes, the DNA sequence to which it binds varies. For example, in the fission yeast *Schizosaccharomyces pombe,* ORC binds stretches of A/T-rich sequences in a stochastic manner (Dai, et al. 2005; Patel, et al. 2006; Segurado, et al. 2003), whereas in the budding yeast *Saccharomyces cerevisiae*, ORC binds a specific A/T-rich consensus sequence (Theis and Newlon 1997). Although this ORC binding motif is conserved among closely related *Saccharomyces* species (Müller and Nieduszynski 2012), it is not known when it arose or how stable it has been over evolutionary time.

In the budding yeast *S. cerevisiae*, replication origins were initially identified by their ability to support replication of plasmids and were consequently termed autonomously replicating sequences (ARSs) (Chan and Tye 1980; Hsiao and Carbon 1979; Struhl, et al. 1979). ARSs in *S. cerevisiae* are characterized by a 17 bp A/T-rich motif termed the extended ARS Consensus Sequence (ACS) (Theis and Newlon 1997; Theis, et al. 1999) coupled with an adjacent element termed B1 (Marahrens and Stillman 1992). The entire consensus sequence spans 33 bp and is termed the ORC-ACS (Eaton, et al. 2010; Xu, et al. 2006). Among other related yeast species whose origins have been examined, *Candida glabrata* has a 17 bp extended ACS similar to that of *S. cerevisiae* and a weak signature of the B1 element (Descorps-Declère, et al. 2015). In contrast, the more evolutionarily distant budding yeast species *Lachancea kluyveri* and *Lachancea waltii* have similar but shorter ACS-like elements, and a B1 element with a different consensus sequence (Di Rienzi, et al. 2012; Liachko, et al. 2011). Another budding yeast, *Kluyveromyces lactis*, has considerably different origins, with an unusually long 50 bp ACS composed of short patches of A/T nucleotides (Liachko, et al. 2010). However, the evolutionary distance between these species and *S. cerevisiae* limits our understanding of how and when replication origins have evolved. To enhance evolutionary

comparisons, we have now identified ORC binding sites in a budding yeast species, *Torulaspora delbrueckii,* that belongs to a lineage of budding yeast species whose origins of replication have not been explored (Figure 1A).

T. delbrueckii, like *Lachancea* and *Kluyveromyces* species, diverged from *Saccharomyces* species before a whole genome duplication (Shen, et al. 2018). This duplication is thought to have occurred more than 100 million years ago through the hybridization of two yeast species -- one from the *Zygosaccharomyces* and *Torulaspora* (ZT) clade, to which *T. delbrueckii* belongs, and one from the *Kluyveromyces*, *Lachancea*, and *Eremothecium* (KLE) clade (Marcet-Houben and Gabaldon 2015). Although the doubled genome originally had equal portions from the two parents, subsequent pseudogenization and gene loss was biased, such that duplicated species like *S. cerevisiae* retain more genes from the ZT lineage than the KLE lineage (Marcet-Houben and Gabaldon 2015). Although replication origins have been identified in species from the KLE clade, origins have not been studied in any species from the ZT clade. Identifying the origins in both parental lineages involved in the ancient hybridization will improve our understanding of how origin selection was affected by this event.

A good proxy for the locations of origins are the ORC binding sites because DNA replication begins with the ATP-dependent binding of ORC to the origin DNA. Therefore, to determine the locations of origins of replication throughout the *T. delbrueckii* genome, we coupled chromatin immunoprecipitation of three subunits of the ORC complex with high throughput DNA sequencing. We identified 190 ORC binding sites across the eight chromosomes in the *T. delbrueckii* genome. We then used these 190 sequences to predict an ORC binding motif. We identified a conserved motif that is strikingly similar to the *S. cerevisiae* ORC-ACS. We further found that several representative ORC binding sites function as ARSs when placed on a plasmid and that the conserved motif is required for this ARS function. Finally, we compared the stabilities of plasmids bearing an *S. cerevisiae* origin and two *T. delbrueckii* ORC-binding sites and found that they were similar. Overall, this study reveals that the consensus sequence that supports ORC binding and DNA replication in *S. cerevisiae* pre-dates the whole genome duplication and was likely contributed by one of the two parents.

Results

Identification of *T. delbrueckii* ORC binding sites. In eukaryotes, the first step of DNA replication is ATP-dependent association of the six-subunit origin recognition complex (ORC) with chromosomal replication origins. To identify regions of the *T. delbrueckii* genome that are potential replication origins, we mapped ORC binding sites across the genome using chromatin immunoprecipitation followed by high throughput sequencing (ChIP-Seq). We immunoprecipitated chromatin fragments associated with three epitope tagged subunits of the ORC complex, Orc1, Orc2 and Orc4 (Figure 1B). Orc1 had a 3XMYC epitope tag, whereas Orc2 and Orc4 had V5 epitope tags. Using two different tags could reduce false signal due to non-specific immunoprecipitation. We also included a mock immunoprecipitation from an untagged strain. Sequence reads from each ChIP sample were aligned to the *T. delbrueckii* genome (Figure 1C), and then one million randomly selected high quality reads were used to identify sites of higher enrichment (peaks) using MACS2 (Zhang, et al. 2008). We identified a total of 216 Orc1 binding sites, 264 Orc2 binding sites and 227 Orc4 binding sites that were present in replicate ChIP-Seq samples from two independently tagged strains for each subunit (Supplemental Figure 1). Finally, we identified 190 sites associated with all three ORC subunits (Figure 1D). These ORC binding sites averaged 670 bp in length and had p-values of 10^{-3} or better (Supplemental Table 1). Binding sites were named systematically based on their location, including the chromosome number and linear coordinate divided by 1000.

Locations of *T. delbrueckii* ORC binding sites. The genome of *T. delbrueckii* is ~9.22 Mb and is distributed across eight chromosomes with 4,972 open reading frames (ORFs) (Gomez-Angulo, et al. 2015; Gordon, et al. 2011). To determine how replication origins are distributed across the genome, we plotted the positions of the 190 *T. delbrueckii* ORC binding sites (Figure 2, Supplemental Figure 2). As described in more detail below, most ORC binding sites were intergenic. The average distance between sites was 47 kbp. We also examined the proximity of ORC binding sites to features reported to be near origins in other yeast. Specifically, centromeres (Descorps-Declère, et al. 2015; Di Rienzi, et al. 2012; Koren, et al. 2011;

Müller and Nieduszynski 2017) and histone genes (Müller and Nieduszynski 2017; Raghuraman, et al. 2001) replicate early in other yeast species. Seven of the eight *T. delbrueckii* centromeres are within 11.4 kbp of an ORC binding site, and *CEN8* is 23.6 kbp away. The histone genes are an average of 3.5 kbp and no more than 9.8 kbp from an ORC binding site.

Identification of motif associated with *T. delbrueckii* ORC binding sites. Origins of replication in *S. cerevisiae* and other budding yeast contain a consensus sequence that is essential for ORC binding and origin function. In *S. cerevisiae,* the core 11 bp sequence is termed the ACS (ARS (Autonomously Replicating Sequence) Consensus Sequence) (Theis and Newlon 1997; Theis, et al. 1999), and the full 33 bp sequence is the ORC-ACS (Eaton, et al. 2010; Xu, et al. 2006). To identify a functionally similar sequence required for ORC binding in *T. delbrueckii*, we searched for a consensus motif among the 190 ORC binding sites identified by ChIP-Seq. We first used the motif finder MEME (Bailey, et al. 2009) without specifying the motif length and obtained a 17 bp conserved motif that was present in 187 of 190 ORC binding sequences (Figure 3A). Interestingly, this motif is nearly identical to the published extended ACS for *S. cerevisiae* and *C. glabrata* (Descorps-Declère, et al. 2015; Eaton, et al. 2010) but less similar to motifs from non-duplicated yeast in the KLE clade (Figure 3B). For example, *L. kluyveri,* has a shorter 9 bp ACS (Liachko, et al. 2011), and *L. waltii* has a 13 bp ACS (Di Rienzi, et al. 2012).

We next fixed the motif length to 33 bp to investigate if there is a B1-like element in *T. delbrueckii*, as in *S. cerevisiae*. Indeed, when a longer motif was specified, MEME returned a sequence that included the extended ACS as well as two shorter motifs resembling the *S. cerevisiae* B1 element (Figure 3B). This motif occurred in 188 of the 190 ORC binding sites when a first order Markov model of the background was built. One binding site, TdARS_VI_530, did not have a match to the consensus motif in this analysis, but did have a match when a second order Markov model of the background was built. For another binding site, III_1257, no match to the conserved motif was detected, although inspection of the genome browser indicated that ORC was associated with this site (Supplemental Figure 4A). As described below, this site had low replication activity on a plasmid. Overall, the motif we detected was more similar to the *S.*

cerevisiae ORC-ACS than motifs described in non-duplicated species from the KLE clade. In these species, the consensus sequence of the B1 element is longer than in *S. cerevisiae* and *T. delbrueckii*. In contrast, the duplicated species *C. glabrata* has a weak match to the B1 element (Descorps-Declère, et al. 2015). Thus, the motif associated with *T. delbrueckii* ORC binding sites is highly similar to the ORC-ACS motif known in *S. cerevisiae*, suggesting that this sequence arose in a common ancestor of the two species and prior to the genome duplication.

A tyrosine in Orc4 that contacts DNA is found in species with an extended ACS. Cryo-EM studies of the *S. cerevisiae* pre-replication complex indicate that a loop in Orc2 and an alpha-helix in Orc4 make direct contacts with DNA in the extended ACS (Yuan, et al. 2017). Moreover, genetic studies pinpoint tyrosine 486 in ScOrc4 as conferring specificity for nucleotides 14 and 15 in the ACS (Hu, et al. 2020). These positions harbor A/T and G/T in *S. cerevisiae*, *C. glabrata*, and *T. delbrueckii*, but no particular nucleotide is found at these positions in the KLE ORC binding sites (Figure 3B). To determine whether having defined nucleotides at these positions within the ORC binding motif (Figure 3B: red boxed nucleotides (*S. cerevisiae* positions 14 and 15 and *T. delbrueckii* positions 12 and 13)) correlates with the presence of this tyrosine, we aligned the Orc4 sequences from the same species (Figure 3C). Indeed, the tyrosine residue occurred in *S. cerevisiae*, *C. glabrata*, and *T. delbrueckii* but not in the KLE species.

ORC binding motifs coincided with the peak of ORC enrichment. If the ACS-like motif we detected by MEME is the ORC recognition site, it should coincide with the peak of ORC binding at each site. We therefore generated aggregate plots and heat maps of the ChIP-Seq signals from the three ORC subunits centered around the 33 bp motif (Figure 4C). For this analysis, the motifs were all oriented the same way. As predicted, the peak of ORC binding did coincide with the motif. Interestingly, we observed that the maximum enrichment was at slightly different positions for the three subunits. The peak for Orc1 was ~60 bp upstream of the center of the motif (nucleotide 17), for Orc4 it was ~20 bp upstream, and for Orc2 it was ~10 bp downstream of the center of the motif. This result is consistent with crosslinking studies of *S.*

cerevisiae ORC that show a differential arrangement of the ORC subunits across the ACS (Lee and Bell 1997).

***T. delbrueckii* ORC binding motifs were preferentially located in intergenic regions.** In other yeasts, replication origins are primarily located in intergenic regions (Descorps-Declère, et al. 2015; Di Rienzi, et al. 2012; Liachko, et al. 2010; Nieduszynski, et al. 2007). To determine if this is also true in *T. delbrueckii*, we determined the positions of the 189 ORC binding motifs relative to the flanking genes. We found a strong preference for intergenic regions, with 170 of 189 motifs (90%) located within an intergenic region (Figure 4A). The average length of these 170 intergenic regions was 784 bp. In contrast, intergenic regions without an ORC binding site were shorter, with an average length of 357 bp (Figure 4B). We also examined the orientations of the genes flanking intergenic regions with ORC binding sites (Figure 4A). There was a slight bias for ORC-binding motifs to be positioned with one gene transcribing towards the motif and the other transcribing away from it. This arrangement was observed for 98 out of 170 motifs (57.6%), which is slightly higher than the 50% expected with no bias. Approximately equal numbers of motifs were located between divergently transcribed genes (37) and convergently transcribing genes (34). Thus, most of the *T. delbrueckii* ORC-binding sites are located within intergenic regions, where they are less likely to interfere with transcription, but there is no favored orientation of the flanking genes.

ORC binding sites enabled plasmids to replicate. An autonomously replicating sequence (ARS) is a sequence that is sufficient to allow a plasmid to replicate. To test whether the *T. delbrueckii* ORC binding sites are functional ARSs, we cloned six representative binding sites into a plasmid lacking an origin and assessed whether the plasmid was stable in *T. delbrueckii*. These six ORC binding sites (TdARS_VII_22, 69, 228, 525, 681, and 734) were 330 to 580 bp in length, and each had a fold enrichment of three or more for all three ORC subunits (Figure 5). We focused on ORC binding sites on chromosome VII because we wanted to identify ORC-binding sites adjacent to the cryptic mating type locus *HMRa1* (unpublished data). The ORC binding sites were cloned onto a plasmid with no other origin sequences, a hygromycin

(antibiotic) resistance gene, and a *T. delbrueckii* centromere to ensure segregation. We also created a similar plasmid containing *S. cerevisiae* ARS209, which is known to have ARS activity in *T. delbrueckii* (Ellahi and Rine 2016). Once plasmids were taken up by *T. delbrueckii* cells, efficient replication was required for growth on medium with hygromycin; therefore, only plasmids bearing a functional ARS yielded robust colonies (Figure 6A). We observed that all six ORC binding sites had strong ARS function, producing robust colonies similar to those produced with the *S. cerevisiae* ARS. In contrast, a control plasmid lacking an ORC binding site produced only tiny colonies. To quantify this difference, we measured the average colony area (Figure 6B) and found that colonies produced with plasmids containing strong ORC binding sites were ≥ 1 mm^2, whereas the control plasmid yielded colonies of less than 0.09 mm^2.

Conserved motif associated with ORC-binding sites was required for ARS activity. In *S. cerevisiae* and other yeast, the ACS is required for ARS activity. To determine whether the motif we detected using MEME behaves as an ACS, we deleted the consensus sequence from TdARS_VII_681, generating TdARS_VII_681Δ (Figure 5B). Plasmids carrying this modified ORC binding site produced only tiny colonies, significantly smaller than those colonies harboring the plasmid with an intact motif (Figure 6A, B). We also tested the one ORC-binding site that did not have a match to the consensus motif in our MEME analysis, site III_1257. All three ORC subunits were present at position III_1257 (Supplemental Figure 4A). However, plasmids carrying site III_1257 produced significantly smaller colonies compared to plasmids bearing an ORC-binding site with the motif (Supplemental Figures 4B and 4C). These findings indicate that the detected motif is required for replication and therefore behaves as an ACS.

Sequences matching the ACS but lacking ORC binding had no ARS activity. Although the ACS is required for DNA replication in *S. cerevisiae* and other yeasts, it is not sufficient for origin activity. In fact, the *S. cerevisiae* genome contains many more matches to the ACS than ORC binding sites or functional replication origins (Belsky, et al. 2015). This observation prompted us to test whether sequences that match the ACS motif but lack ORC binding would display ARS activity. To do so, we cloned three genetic loci

with potential ACSs but low or questionable ORC enrichment. One site, VII_55, had a match to the ORC-binding motif using FIMO (Bailey, et al. 2009), but no ORC enrichment. Another site, VII_591, had low enrichment (≤3) of Orc2 and no enrichment for Orc1 and Orc4. A third site, VII_6, was subtelomeric and had a broad peak of Orc1 that likely reflects sub-telomeric heterochromatin. However, Orc2 and Orc4 had a different pattern of association, with narrow peaks of low enrichment (≤3) (Supplemental Figure 3). To determine whether VII_6 represents a subtelomeric origin, we included it in the ARS assay. Plasmids containing the three ACS-like sequences produced only tiny colonies with an average area of less than 0.09 mm^2 (Figures 6A and B). This finding indicates that the identified ACS motif alone is not sufficient for replication and that additional features are needed to recruit ORC and promote DNA replication.

Plasmids bearing *S. cerevisiae* and *T. delbrueckii* ARSs were lost at similar rates in *T. delbrueckii*. The predicted *T. delbrueckii* ACS is similar to, but not exactly the same as, the *S. cerevisiae* ACS (Figure 3), which could indicate slight differences in the specificity of ORC in the two species. To compare the abilities of ScARS209 and *T. delbrueckii* origins to promote replication, we used a plasmid loss assay (Marahrens and Stillman 1992), which is more sensitive than the transformation assay. We compared the relative stabilities of plasmids bearing TdARS_VII_228 and TdARS_VII_681, which have high fold-enrichment of ORC subunits (Figure 5B), with a plasmid bearing *Sc*ARS209. To measure the retention of the plasmid, yeast containing plasmids were grown for ~15 generations in non-selective medium (YPD), and the fraction of hygromycin-resistant cells was calculated at the start and end of the non-selective growth. Both the *T. delbrueckii* and the *S. cerevisiae* ARS-containing plasmids showed a similar extent of plasmid loss (Figure 6C), consistent with the sequence features necessary for origin function being similar in the two species. As an aside, we note that the loss rate for the plasmid we used in *T. delbrueckii* is higher than for the plasmids that have been studied in *S. cerevisiae* (Marahrens and Stillman 1992). This higher loss rate could be due to differences in centromere-mediated segregation.

Discussion

In this study, we mapped ORC binding sites and characterized their consensus sequences in the non-duplicated budding yeast *Torulaspora delbrueckii*. Using ChIP-Seq, we identified binding sites for three of the six subunits of the origin recognition complex (ORC). Out of the 216 Orc1, 264 Orc2 and 227 Orc4 binding sites, 190 sites were associated with all three subunits. We also found a consensus motif in 189 of the 190 ORC binding sites. In addition, we found that six of six tested ORC binding sites with a consensus motif had autonomous replication function on a plasmid, supporting the expectation that the ORC binding sites are genomic origins of replication.

Our approach of using ChIP-Seq to identify potential origins of replication was less labor-intensive than screening genomic libraries for ARS activity and homed in on consensus sequences more precisely than measuring replication timing across the genome. This ChIP-Seq strategy takes advantage of the foundational role of ORC in recruiting other proteins to the pre-replication complex. In addition, we used two different epitope tags to reduce technical biases. Validating this approach, all six ORC binding sites tested for ARS function supported plasmid replication. In contrast, sites that matched the consensus sequence but had low or no association of the three subunits showed no ARS function. For example, VII_55, had a match to the ORC-binding motif but no ORC enrichment, consequently, had no replication function on a plasmid.

The locations of ORC-binding sites across the *T. delbrueckii* genome are consistent with observations in previously studied budding yeasts. For example, in many yeast species, centromeres and histone genes replicate early and hence are near origins (Descorps-Declère, et al. 2015; Di Rienzi, et al. 2012; Koren, et al. 2011; Müller and Nieduszynski 2017). Early replication of the centromeres may promote kinetochore assembly or proper orientation of sister chromatids (McCarroll and Fangman 1988). Similarly, early replication of histone genes may enhance production of histones prior to duplication of the remaining genome (Mei, et al. 2017). Consistent with observations in other yeast, we found that *T. delbrueckii* ORC binding sites are close to centromeres and histone genes. Each of the two copies of histone H2A, H2B, H3,

and H4 genes are located within 9.8 kbp of ORC binding sites. Similarly, seven of the eight *T. delbrueckii* centromeres were within 11.4 kbp of an ORC binding site. *CEN8* is farther from the nearest ORC binding site (23.6 kbp). An examination of the ChIP-Seq read depth in the vicinity of *CEN8* did not reveal an ORC binding site that had been missed. Thus, it is either not essential for centromeres to be close to origins, or *CEN8*, which has not been examined functionally, may not be the actual centromere. Although our work did not explicitly examine replication timing, these findings are consistent with the importance of early replication for centromere and histone gene function.

The average distance between adjacent ORC binding sites in the *T. delbrueckii* genome is approximately 47 kbp, which falls in the middle of the range for budding yeasts. The average distance is shorter for the 269 origins in *S. cerevisiae* (30 kbp), longer for the 124 origins in *K. lactis* (70 kbp), but about the same for the 195 origins in *L. waltii* (52 kbp) and the 275 origins in *C. glabrata* (45 kbp). Interestingly, we identified fewer ORC binding sites in *T. delbrueckii* (190) than there are in *S. cerevisiae* (269). Given that only a fraction of origins are used in any given cell division in *S. cerevisiae* (Friedman, et al. 1997), it would be interesting to explore the fraction of active origins and replicon size in *T. delbrueckii*. Alternatively, the difference in origin number and spacing may indicate that additional origins are yet to be discovered in *T. delbrueckii*, which has not been as extensively studied as *S. cerevisiae*.

We also found that a majority (90%) of the *T. delbrueckii* ORC-binding sites occur in intergenic regions, although there was no bias for a particular orientation of flanking genes. This lack of bias is similar to the non-duplicated species *K. lactis* (Liachko, et al. 2010), but different from *S. cerevisiae* (Nieduszynski, et al. 2007) and *L. waltii* (Di Rienzi, et al. 2012) where ARSs are biased for intergenic regions between convergently and divergently transcribed genes, respectively. We also found that intergenic regions with ORC-binding sites are slightly larger than those without. This arrangement could favor replication initiation by reducing transcription interference with ORC binding.

We noticed in aggregate plots that the peaks of the three ORC subunits were slightly offset from one another with respect to the consensus motif. The peak of Orc1 was the most 5' relative to the T-rich strand, and Orc2 was the most 3'. This offset binding is similar to previously reported crosslinking positions

(Lee and Bell 1997). Moreover, it is possible that Orc1 contacts DNA upstream of the ORC-ACS motif through its nucleosome-binding BAH domain. However, the shifted peak of Orc1 could also be a technical artifact, as the Orc1 ChIP-Seq data was collected at a different time using a different antibody. In addition, single-end sequencing was used for Orc1 and paired-end sequencing was used Orc2 and Orc4. Therefore, whether the skewed enrichment of the three ORC subunits is a functional aspect of ORC-DNA interaction requires further investigation.

The consensus sequence we found to be associated with ORC binding sites was essential but not sufficient for replication of a plasmid. In the ARS assay, deletion of the 17 bp ACS-like motif resulted in small colony sizes, reflecting little replication of the plasmid. Additionally, an ORC binding site which had no consensus ORC binding motif replicated poorly. On the other hand, matches to the 17 bp consensus sequence that were not occupied by ORC produced only tiny colonies, indicating poor or no ARS function. Hence, there are most likely additional features beyond the consensus motif that are important for ORC binding and DNA replication. In *S. cerevisiae*, additional B elements downstream of the ORC-ACS are required for origin function. For example, A-rich B2 elements facilitate helicase loading and unwinding (Lipford and Bell 2001; Zou and Stillman 2000), and the B3 element recruits Abf1, thereby positioning nucleosomes (Lipford and Bell 2001). Presumably, similar sequences exist near ORC binding sites in *T. delbrueckii*.

Although specific sequences define replication origins in all Saccharomycetaceae yeasts that have been examined, the consensus sequence varies across species. This variability is particularly pronounced among the non-duplicated yeast from the *Kluyveromyces*, *Lachancea, Eremothecium* (KLE) clade. A comparative study of ten *Lachancea* species revealed a variety of motifs at replication origins (Agier, et al. 2018), and *K. lactis* has an unusually long 50 bp ARS which barely resembles the consensus motifs from other budding yeast (Liachko, et al. 2010). In contrast *T. delbrueckii,* belonging to the *Zygosaccharomyces, Torulaspora* (ZT) clade, has a motif that is highly similar to the ORC-ACS found in duplicated species such as *S. cerevisiae* and *C. glabrata*. We also observed that an *S. cerevisiae* origin and two *T. delbrueckii* origins confer similar stability to a plasmid in *T. delbrueckii* cells. It is striking that the ORC-ACS motif is

so similar in *T. delbrueckii* and *S. cerevisiae*, which diverged over 100 million years ago, whereas origin motifs have shifted on a shorter time scale among KLE species.

The similarity between the ORC-ACS motifs in *S. cerevisiae* and *T. delbrueckii* raises questions about the early events in the whole genome duplication. In the initial hybridization event, it seems likely that the ZT parent contributed replication origins of the ORC-ACS type, whereas the KLE parent contributed another type of origin. Did having two different mechanisms for recognizing origins lead to genomic instability? How was this conflict ultimately resolved in favor of the ORC-ACS type of origin? One possibility is that ORC subunits from the ZT parent were retained, favoring use of the ORC-ACS sequence. Once the ORC proteins favored the ORC-ACS type of origin, DNA derived from the KLE parent may have been replicated less efficiently than DNA derived from the ZT parent. Did this poor replication mimic DNA damage and trigger recombination between the homologous KLE and ZT sequences? Did this lead to loss of heterozygosity in favor of the efficiently replicated ZT sequences? If so, this uneven replication may have contributed to duplicated species retaining more genetic material from the ZT parent compared to the KLE parent (Marcet-Houben and Gabaldon 2015).

In summary, this work reveals that origin sequences can be stable over long periods of evolutionary time and sets the stage for future studies on how origin selection mechanisms impact DNA retention after hybridization events.

Materials and Methods

Tree construction

To construct the species tree, we selected 10 protein-coding genes (Supplemental Table 2) found in single copy across the 9 species considered (Aguileta, et al. 2008). Their amino acid sequences were obtained from the yeast gene order browser (Byrne and Wolfe 2005) and the candida gene order browser (Maguire, et al. 2013) and then concatenated, resulting in a combined alignment of 9925 positions. The concatenated sequences were aligned in MAFFT v7.475 using E-INS-i with default settings (Katoh and Standley 2013). MAFFT alignment was input into IQTree (Trifinopoulos, et al. 2016) to determine the best amino acid substitution model: LG+F+I+G4 (Hoang, et al. 2018). Finally, a maximum likelihood phylogenetic tree was constructed using PhyML v3.1 using the LG+F+I+G4 model as implemented in Seaview v5.0.4 (Gouy, et al. 2010). Node support was assessed with approximate likelihood ratio tests (aLRT). Subtree pruning and regrafting (SPR) was used as a search algorithm with five random starts.

Yeast growth and transformation

Yeast strains used in this study (Supplemental Table 3) were derived from JRY10156 (Ellahi and Rine 2016), a *Torulaspora delbrueckii MATα ura3Δ0 trp3-1* strain descended from NRRL Y-866. Yeast were grown at 30˚C in YPD medium containing 1% yeast extract, 2% peptone, and 2% glucose. *T. delbrueckii* cells were transformed using electroporation (Hickman and Rusche 2009). Briefly, cells were harvested at an optical density (OD_{600}) ~1 and resuspended at 15 OD/ml in YPD containing 25 mM DTT and 20 mM HEPES, pH 8. Cells were shaken for 30 minutes at 30°, collected, and washed in electroporation buffer (10 mM tris pH 7.5, 270 mM sucrose, 1 mM LiOAc). Cells were then resuspended at 100 OD/ml in electroporation buffer. Electroporation was conducted in 0.2 mm cuvettes using 50 µL cells at 1000 V, 300 Ω, and 25 µF. After electroporation, cells recovered 15 minutes in chilled YPD and then three hours at 30° before plating on selective medium (YPD containing 400 µg/mL hygromycin). For strain construction, 100 ng of linear DNA and 1 µL of 10 mg/mL sheared salmon sperm DNA (AM9680, Invitrogen) were used in

a volume no more than 5 µL. For the transformation frequency assay (Clyne and Kelly 1997), JRY10156 was transformed with 5 ng of each plasmid to be tested.

Plasmid and yeast strain construction

Plasmids used in this study are listed in Supplemental Tables 4 and 5. Plasmids for tagging ORC subunits were derived from pRS414 (Sikorski and Hieter 1989). *T. delbrueckii ORC1* (*TDEL0D00910*), *ORC2* (*TDEL0D03510*), and *ORC4* (*TDEL0H00550*) reading frames along with ~500 bp flanking sequence were amplified from genomic DNA (JRY10156) and ligated into pRS414. To tag Orc2 and Orc4, the V5 epitope tag and hygromycin resistance marker were amplified from plasmid pFA6a-6xGLY-V5-hphMX4 (Funakoshi and Hochstrasser 2009) and stitched into the 3' end of each gene. For Orc1, the 3XMYC tag was inserted in the linker between the BAH and AAA+ domains, and the hygromycin resistance marker was added downstream of the gene. The 3XMYC epitope tag was amplified from plasmid pWZV87 (Brand, et al. 1985), and the hygromycin resistance gene was amplified from pFA6a-6xGLY-V5-hphMX4 (Funakoshi and Hochstrasser 2009). To integrate Orc1-3XMYC, Orc2-V5, and Orc4-V5 into *T. delbrueckii*, tagging constructs were excised from the plasmids by restriction digestion and used to transform JRY10156 (Ellahi and Rine 2016). To confirm an intact reading frame, the tagged genes were sequenced, and to confirm protein expression, immunoblots were used.

Plasmids for the transformation frequency assay were derived from pRS41H (Ellahi and Rine 2016), which has a *T. delbrueckii* centromere (*CEN3*), an *S. cerevisiae* origin (ARS209), and a hygromycin resistance gene. To facilitate insertion of potential origins, the *S. cerevisiae* ARS was deleted and replaced with SphI and NheI cut sites, generating pLR1275. Next, ORC binding sites to be tested for ARS activity were amplified from *T. delbrueckii* genomic DNA (JRY10156) using primers with flanking SphI and NheI sites. *S. cerevisiae* ARS209 was also amplified from pRS41H using primers with flanking SphI and NheI sites. III_1257 was amplified using primers with flanking SphI and XbaI sites. Digested PCR products were ligated into plasmid pLR1275 using NEB instant sticky-end ligase master mix (NEB M0370S).

385 Additionally, the 17 bp ORC binding motif was deleted from pLR1278 and replaced with XbaI cut site
386 using PCR-mediated mutagenesis to generate pLR1287.
387
388 **Immunoblotting**
389 Immunoblots were performed as previously described (Hickman and Rusche 2010). Cells were
390 grown to an OD_{600} of ~1.0, and fixed with a final concentration of 10% TCA. For lysis, 40 OD equivalents
391 of fixed cells were vortexed 5 min in the presence of silica beads (0.5 mm dia. #11079105z, BioSpec
392 Products) in 40 µl lysis buffer (10 mM HEPES pH 7.9, 150 mM KCl, 1.5 mM $MgCl_2$, 0.5 mM DTT, 10%
393 glycerol, 5 µg/ml chymostatin, 2 µg/ml pepstatin A, 156 µg/ml benzamidine, 35.2 µg/ml TPCK, 174 µg/ml
394 PMSF, and 1× Complete Protease Inhibitor (Roche)). Proteins were denatured by the addition of 1/3 volume
395 3X sample buffer (30% glycerol, 15% β-mercaptoethanol, 6% SDS, 200 mM Tris pH 6.8, 0.08 mg/ml
396 bromophenol blue) and incubation at 95° for 5 min. Finally, samples were clarified by centrifugation.
397 Aliquots (10 µl) of protein extract were resolved on a 7.5% acrylamide gel, transferred to membrane
398 (Amersham 45004008), and probed to detect Orc1 protein (anti-MYC, Millipore 06-549), Orc2 or Orc4
399 proteins (anti-V5, Millipore AB3792) or Pgk1 (ab113687, Abcam) as a loading control.
400
401 **Transformation frequency and plasmid loss assay**
402 For the transformation frequency assay, *T. delbrueckii* cells were transformed with 5 ng plasmid
403 and plated on selective medium (YPD + 400 µg/mL hygromycin). Three replicate transformations were
404 performed, and colony images were taken after two days growth at 30˚C. The colony areas were determined
405 for all the colonies from three replicate transformation plates using Image J and then averaged (Schneider,
406 et al. 2012).
407 For the plasmid loss assay, three single colonies were used to inoculate 5 ml YPD + hygromycin
408 and grown for 22 hours to an OD_{600} of approximately 13. Aliquots of each culture were plated onto three
409 selective (YPD + hygromycin) and nonselective (YPD) plates to determine the fraction of cells resistant to
410 hygromycin and hence containing the plasmid. The culture was also diluted to OD_{600} =0.0003 in

nonselective (YPD) medium and grown for 30 hours to an OD_{600} of approximately 13. The fraction of cells containing plasmid was then determined as described above.

Chromatin immunoprecipitation and sequencing

Chromatin immunoprecipitation was performed as previously described (Rusche and Rine 2001). Cells were grown in YPD at 30˚C, harvested at an OD_{600} of ~1, and cross-linked for 1 hour in 1% formaldehyde. For immunoprecipitation, 4 µl of anti MYC antibody (Millipore 06-549) or anti V5 antibody (Millipore AB3792) were used. The immunoprecipitation was conducted with 10 µL of Protein A agarose beads in the absence of BSA and salmon sperm DNA. Library preparation (Takara Bio SMARTer ThruPlex DNA seq kit) and sample barcoding was done at the Next-Generation Sequencing facility at University at Buffalo. The samples were then sequenced on an Illumina 1.9 sequencer using 76 bp single-end sequencing for Orc1-3XMYC (LRY3204 and LRY3211) and 151 bp paired-end sequencing for Orc2-V5 and Orc4-V5 (LRY3142, LRY3143, LRY3138 and LRY3139).

Bioinformatics analysis pipeline

Prior to sequence alignment, the paired-end sequence reads were processed to remove 3' adapter sequences ("`-a AGATCGGAAGAGCACACGTCTGAACTCCAGTCA -A AGATCGGAAGAGCGTCGTGTAGGGAAA GAGTGT`") using Cutadapt/2.10 (Martin 2011) with the parameters "--nextseq-trim=20, -m 10". Trimmed paired-end reads and raw single-end reads were aligned to the *T. delbrueckii* genome (CBS1146 (GCA_000243375.1)) (Gordon, et al. 2011) using Bowtie2/2.3.4.1 (Langmead, et al. 2009). Samtools/1.9 (Li, et al. 2009) was used to convert .sam output files from Bowtie2/2.3.4.1 to .bam, -sorted.bam and index files. To identify ORC binding sites, **M**odel-based **A**nalysis of **C**hIP-**S**eq, MACS2/2.2.7.1 (Zhang, et al. 2008) was used to call peaks using one million randomly selected high quality reads with mock IP as control for Orc2 and Orc4 and genomic input as control for Orc1 samples; For all samples, --nomodel was used, --extsize 200 was used for single-end reads only, and -f BEDPE was used for paired-end reads only. To determine intersecting peaks for Orc1, Orc2, and Orc4, Bedtools/2.23.0 (Quinlan and Hall 2010) was used

with the parameters "-f 0.4, -r" which requires at least 40% overlap among all peaks. Finally, to determine the ORC binding motif, MEME Suite 5.3.0 (**M**ultiple **E**m for **M**otif **E**licitation) (Bailey, et al. 2009) was used with the 190 ORC-binding sites as input.

To compare the ORC-binding motifs of various species, the coordinates of replications origins were retrieved from published literature for *S. cerevisiae* (Belsky, et al. 2015), *K. lactis* (Liachko, et al. 2010), *C. glabrata* (Descorps-Declère, et al. 2015), *L. kluyveri* (Liachko, et al. 2011) and *L. waltii* (Di Rienzi, et al. 2012). Reference genomes for *S. cerevisiae* (GCF_000146045.2), *K. lactis* (GCF_000002515.2), *C. glabrata* (GCF_000002545.3), and *L. kluyveri* (GCA_000149225.1) were downloaded from NCBI. The reference genome for *L. waltii* was provided by M. K. Raghuraman (Di Rienzi, et al. 2009). MEME (Bailey, et al. 2009) was used to search for consensus motifs in 269 *S. cerevisiae*, 124 *K. lactis*, 275 *C. glabrata*, 84 *L. kluyveri* and 194 *L. waltii* sequences. The search assumed the motif is present zero or once per sequence (the ZOOPS model) and enriched over the background. A first order Markov model of the background was built for *T. delbrueckii*, a second order Markov model of the background was built for *S. cerevisiae* and *L. waltii*, a fourth order Markov model of the background was built for *K. lactis* and *L. kluyveri* , and a zero order Markov model of the background was built for *C. glabrata*. A 33 bp motif was specified for *S. cerevisiae*, *C. glabrata*, *L. waltii,* and *T. delbrueckii*, a 9 bp motif was specified for *L. kluyveri*, and a 50 bp motif was specified for *K. lactis*.

Aggregate plot and heat maps for ChIP-Seq signals for Orc1, Orc2 and Orc4 were generated using DeepTools/3.5.0 (Ramirez, et al. 2014). For this analysis, .bed files of two replicates for each ORC subunit were combined using MACS/2.2.7.1. DeepTools/3.5.0 'computeMatrix reference-point' was used to calculate scores per genome region using the parameters "--sortRegions no --binSize 10". For both aggregate plot and heat map the ChIP-Seq signals from the three ORC subunits were centered around the 33 bp motif generated from MEME and extended 1 kb base pairs upstream and downstream of the motif.

Data availability

The data underlying this article, including raw Illumina data for Orc1/2/4 ChIP-sequencing were uploaded at the NCBI GEO database (Accession number: GSE165127).

Acknowledgements

We thank Aigerim Nugmanova for assistance in making plasmids, Michael Buck for suggestions on ChIP-seq data analysis, Jasper Rine and Aisha Ellahi for strains and plasmids, M. K. Raghuraman for providing the *L. waltii* genome assembly, Derek Taylor for assistance on the phylogenetic tree construction and members of the Rusche lab for suggestions and support. Bioinformatic analysis was supported by the Center for Computational Research at the University at Buffalo (http://hdl.handle.net/10477/79221). This work was supported by the National Science Foundation (MCB 1615367) and the Mark Diamond Research Fund of the Graduate Student Association at the University at Buffalo, the State University of New York.

References

Agier N, et al. 2018. The evolution of the temporal program of genome replication. Nature Communications 9: 2199. doi: 10.1038/s41467-018-04628-4

Aguileta G, et al. 2008. Assessing the performance of single-copy genes for recovering robust phylogenies. Syst Biol 57: 613-627. doi: 10.1080/10635150802306527

Bailey TL, et al. 2009. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 37: W202-208. doi: 10.1093/nar/gkp335

Bell SP, Stillman B 1992. ATP-dependent recognition of eukaryotic origins of DNA replication by a multiprotein complex. Nature 357: 128-134. doi: 10.1038/357128a0

Belsky JA, MacAlpine HK, Lubelsky Y, Hartemink AJ, MacAlpine DM 2015. Genome-wide chromatin footprinting reveals changes in replication origin architecture induced by pre-RC assembly. Genes Dev 29: 212-224. doi: 10.1101/gad.247924.114

Brand AH, Breeden L, Abraham J, Sternglanz R, Nasmyth K 1985. Characterization of a "silencer" in yeast: a DNA sequence with properties opposite to those of a transcriptional enhancer. Cell 41: 41-48. doi: 10.1016/0092-8674(85)90059-5

Byrne KP, Wolfe KH 2005. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. Genome Res 15: 1456-1461. doi: 10.1101/gr.3672305

Chan CS, Tye BK 1980. Autonomously replicating sequences in Saccharomyces cerevisiae. Proceedings of the National Academy of Sciences 77: 6329-6333. doi: 10.1073/pnas.77.11.6329

Clyne RK, Kelly TJ 1997. Identification of autonomously replicating sequence (ARS) elements in eukaryotic cells. Methods 13: 221-233. doi: 10.1006/meth.1997.0522

Dai J, Chuang R-Y, Kelly TJ 2005. DNA replication origins in the *Schizosaccharomyces pombe* genome. Proceedings of the National Academy of Sciences of the United States of America 102: 337-342. doi: 10.1073/pnas.0408811102
Descorps-Declère S, et al. 2015. Genome-wide replication landscape of *Candida glabrata*. BMC Biology 13: 69. doi: 10.1186/s12915-015-0177-6
Di Rienzi SC, Collingwood D, Raghuraman MK, Brewer BJ 2009. Fragile genomic sites are associated with origins of replication. Genome Biol Evol 1: 350-363. doi: 10.1093/gbe/evp034
Di Rienzi SC, et al. 2012. Maintaining replication origins in the face of genomic change. Genome Res 22: 1940-1952. doi: 10.1101/gr.138248.112
Eaton ML, Galani K, Kang S, Bell SP, MacAlpine DM 2010. Conserved nucleosome positioning defines replication origins. Genes Dev 24: 748-753. doi: 10.1101/gad.1913210
Ellahi A, Rine J 2016. Evolution and Functional Trajectory of Sir1 in Gene Silencing. Mol Cell Biol 36: 1164-1179. doi: 10.1128/mcb.01013-15
Friedman KL, Brewer BJ, Fangman WL 1997. Replication profile of *Saccharomyces cerevisiae* chromosome VI. Genes to cells : devoted to molecular & cellular mechanisms 2: 667-678. doi: 10.1046/j.1365-2443.1997.1520350.x
Funakoshi M, Hochstrasser M 2009. Small epitope-linker modules for PCR-based C-terminal tagging in *Saccharomyces cerevisiae*. Yeast 26: 185-192. doi: 10.1002/yea.1658
Gomez-Angulo J, et al. 2015. Genome Sequence of *Torulaspora delbrueckii* NRRL Y-50541, Isolated from Mezcal Fermentation. Genome Announc 3. doi: 10.1128/genomeA.00438-15
Gordon JL, et al. 2011. Evolutionary erosion of yeast sex chromosomes by mating-type switching accidents. Proc Natl Acad Sci U S A 108: 20024-20029. doi: 10.1073/pnas.1112808108
Gouy M, Guindon S, Gascuel O 2010. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol 27: 221-224. doi: 10.1093/molbev/msp259
Hickman MA, Rusche LN 2009. The Sir2-Sum1 Complex Represses Transcription Using Both Promoter-Specific and Long-Range Mechanisms to Regulate Cell Identity and Sexual Cycle in the Yeast *Kluyveromyces lactis*. PLoS Genet 5: e1000710. doi: 10.1371/journal.pgen.1000710
Hickman MA, Rusche LN 2010. Transcriptional silencing functions of the yeast protein Orc1/Sir3 subfunctionalized after gene duplication. Proc Natl Acad Sci U S A 107: 19384-19389. doi: 10.1073/pnas.1006436107
Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol Biol Evol 35: 518-522. doi: 10.1093/molbev/msx281
Hsiao CL, Carbon J 1979. High-frequency transformation of yeast by plasmids containing the cloned yeast ARG4 gene. Proc Natl Acad Sci U S A 76: 3829-3833. doi: 10.1073/pnas.76.8.3829
Hu Y, et al. 2020. Evolution of DNA replication origin specification and gene silencing mechanisms. Nat Commun 11: 5175. doi: 10.1038/s41467-020-18964-x
Katoh K, Standley DM 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30: 772-780. doi: 10.1093/molbev/mst010
Koren A, et al. 2011. Correction: Epigenetically-Inherited Centromere and Neocentromere DNA Replicates Earliest in S-Phase. PLoS Genet 7: 10.1371/annotation/d1374e1378b1374eb-2385-1374a1346-1938e-1319bbae1374fcf1389. doi: 10.1371/annotation/d4e8b4eb-2385-4a46-938e-19bbae4fcf89
Langmead B, Trapnell C, Pop M, Salzberg SL 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10: R25. doi: 10.1186/gb-2009-10-3-r25
Lee DG, Bell SP 1997. Architecture of the yeast origin recognition complex bound to origins of DNA replication. Mol Cell Biol 17: 7159-7168. doi: 10.1128/mcb.17.12.7159

Li H, et al. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079. doi: 10.1093/bioinformatics/btp352
Liachko I, et al. 2010. A comprehensive genome-wide map of autonomously replicating sequences in a naive genome. PLoS Genet 6: e1000946. doi: 10.1371/journal.pgen.1000946
Liachko I, et al. 2011. Novel features of ARS selection in budding yeast *Lachancea kluyveri*. BMC Genomics 12: 633. doi: 10.1186/1471-2164-12-633
Lipford JR, Bell SP 2001. Nucleosomes positioned by ORC facilitate the initiation of DNA replication. Mol Cell 7: 21-30. doi: 10.1016/s1097-2765(01)00151-4
Maguire SL, et al. 2013. Comparative genome analysis and gene finding in Candida species using CGOB. Mol Biol Evol 30: 1281-1291. doi: 10.1093/molbev/mst042
Marahrens Y, Stillman B 1992. A yeast chromosomal origin of DNA replication defined by multiple functional elements. Science 255: 817-823. doi: 10.1126/science.1536007
Marcet-Houben M, Gabaldon T 2015. Beyond the Whole-Genome Duplication: Phylogenetic Evidence for an Ancient Interspecies Hybridization in the Baker's Yeast Lineage. PLoS Biol 13: e1002220. doi: 10.1371/journal.pbio.1002220
Martin M 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. 2011 17: 3. doi: 10.14806/ej.17.1.200
McCarroll RM, Fangman WL 1988. Time of replication of yeast centromeres and telomeres. Cell 54: 505-513. doi: https://doi.org/10.1016/0092-8674(88)90072-4
Mei Q, et al. 2017. Regulation of DNA replication-coupled histone gene expression. Oncotarget 8: 95005-95022. doi: 10.18632/oncotarget.21887
Müller CA, Nieduszynski CA 2012. Conservation of replication timing reveals global and local regulation of replication origin activity. Genome Res 22: 1953-1962. doi: 10.1101/gr.139477.112
Müller CA, Nieduszynski CA 2017. DNA replication timing influences gene expression level. The Journal of cell biology 216: 1907-1914. doi: 10.1083/jcb.201701061
Nieduszynski CA, Hiraga S-i, Ak P, Benham CJ, Donaldson AD 2007. OriDB: a DNA replication origin database. Nucleic Acids Res 35: D40-D46. doi: 10.1093/nar/gkl758
Patel PK, Arcangioli B, Baker SP, Bensimon A, Rhind N 2006. DNA replication origins fire stochastically in fission yeast. Mol Biol Cell 17: 308-316. doi: 10.1091/mbc.e05-07-0657
Quinlan AR, Hall IM 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841-842. doi: 10.1093/bioinformatics/btq033
Raghuraman MK, et al. 2001. Replication dynamics of the yeast genome. Science 294: 115-121. doi: 10.1126/science.294.5540.115
Ramirez F, Dündar F, Peter S, Grüning B, Manke T 2014. DeepTools: A flexible platform for exploring deep-sequencing data. Nucleic Acids Res 42. doi: 10.1093/nar/gku365
Rusche LN, Rine J 2001. Conversion of a gene-specific repressor to a regional silencer. Genes Dev 15: 955-967. doi: 10.1101/gad.873601
Schneider CA, Rasband WS, Eliceiri KW 2012. NIH Image to ImageJ: 25 years of image analysis. Nature Methods 9: 671-675. doi: 10.1038/nmeth.2089
Segurado M, de Luis A, Antequera F 2003. Genome-wide distribution of DNA replication origins at A+T-rich islands in *Schizosaccharomyces pombe*. EMBO Rep 4: 1048-1053. doi: 10.1038/sj.embor.embor7400008
Shen XX, et al. 2018. Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum. Cell 175: 1533-1545.e1520. doi: 10.1016/j.cell.2018.10.023
Sikorski RS, Hieter P 1989. A system of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in *Saccharomyces cerevisiae*. Genetics 122: 19-27.

588 Struhl K, Stinchcomb DT, Scherer S, Davis RW 1979. High-frequency transformation of yeast:
589 autonomous replication of hybrid DNA molecules. Proc Natl Acad Sci U S A 76: 1035-1039. doi:
590 10.1073/pnas.76.3.1035
591 Theis JF, Newlon CS 1997. The ARS309 chromosomal replicator of *Saccharomyces cerevisiae* depends on
592 an exceptional ARS consensus sequence. Proc Natl Acad Sci U S A 94: 10786-10791. doi:
593 10.1073/pnas.94.20.10786
594 Theis JF, Yang C, Schaefer CB, Newlon CS 1999. DNA sequence and functional analysis of homologous
595 ARS elements of *Saccharomyces cerevisiae* and *S. carlsbergensis*. Genetics 152: 943-952.
596 Trifinopoulos J, Nguyen LT, von Haeseler A, Minh BQ 2016. W-IQ-TREE: a fast online phylogenetic tool
597 for maximum likelihood analysis. Nucleic Acids Res 44: W232-235. doi: 10.1093/nar/gkw256
598 Xu W, Aparicio JG, Aparicio OM, Tavaré S 2006. Genome-wide mapping of ORC and Mcm2p binding sites
599 on tiling arrays and identification of essential ARS consensus sequences in S. cerevisiae. BMC Genomics
600 7: 276. doi: 10.1186/1471-2164-7-276
601 Yuan Z, et al. 2017. Structural basis of Mcm2-7 replicative helicase loading by ORC-Cdc6 and Cdt1.
602 Nature structural & molecular biology 24: 316-324. doi: 10.1038/nsmb.3372
603 Zhang Y, et al. 2008. Model-based Analysis of ChIP-Seq (MACS). Genome Biology 9: R137. doi:
604 10.1186/gb-2008-9-9-r137
605 Zou L, Stillman B 2000. Assembly of a complex containing Cdc45p, replication protein A, and Mcm2p at
606 replication origins controlled by S-phase cyclin-dependent kinases and Cdc7p-Dbf4p kinase. Mol Cell Biol
607 20: 3086-3096. doi: 10.1128/mcb.20.9.3086-3096.2000

608

609 Figures and Legends

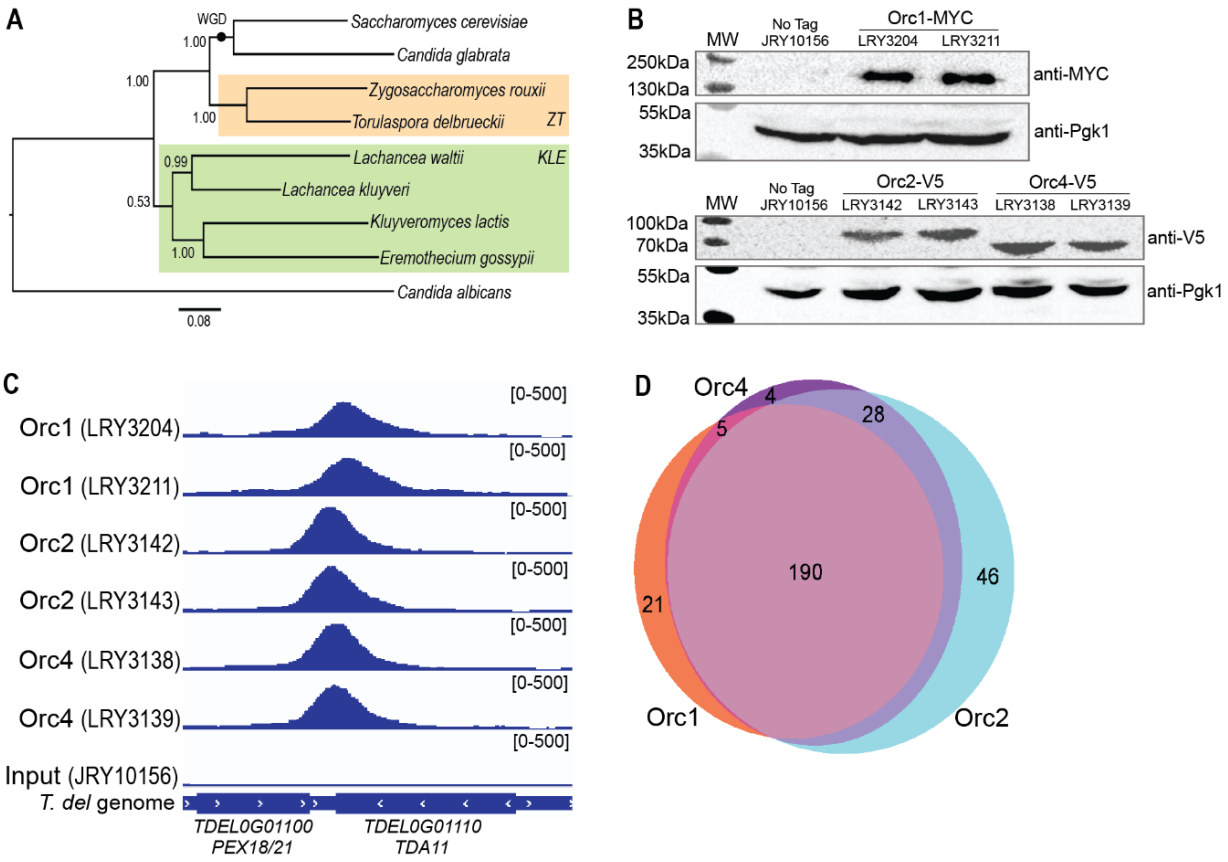


Fig. 1. **Identification of Orc1, Orc2 and Orc4 binding sites in *T. delbrueckii*. (A)** Evolutionary relationship of the species discussed in the paper. A species tree was built using a maximum likelihood approach on a concatenated alignment of 10 single copy genes. Node support values (aLRT, approximate likelihood ratio test) are shown. WGD denotes the whole genome duplication (WGD). *ZT* denotes the *Zygosaccharomyces* and *Torulaspora* clade (tan box), and KLE denotes the *Kluyveromyces*, *Lachancea*, and *Eremothecium* clade (green box). **(B)** Immunoblot of tagged ORC subunits TdOrc1-3XMYC (LRY3204, LRY3211), TdOrc2-V5 (LRY3142, LRY3143) and TdOrc4-V5 (LRY3138, LRY3139). The untagged parent strain (JRY10156) was included as a control. Endogenous 3-phosphoglycerate kinase (Pgk1) was detected as a loading control. **(C)** A representative ORC binding site on Chromosome VII, ARS_VII_228, is depicted. Each line represents the read depth of sequences generated by ChIP-Seq from

one tagged ORC strain. The bottom line shows the positions of genes flanking the ORC binding site. **(D)** Venn diagram indicating the overlap between 216 Orc1 binding sites (coral), 264 Orc2 binding sites (blue), and 227 Orc4 binding sites (purple).

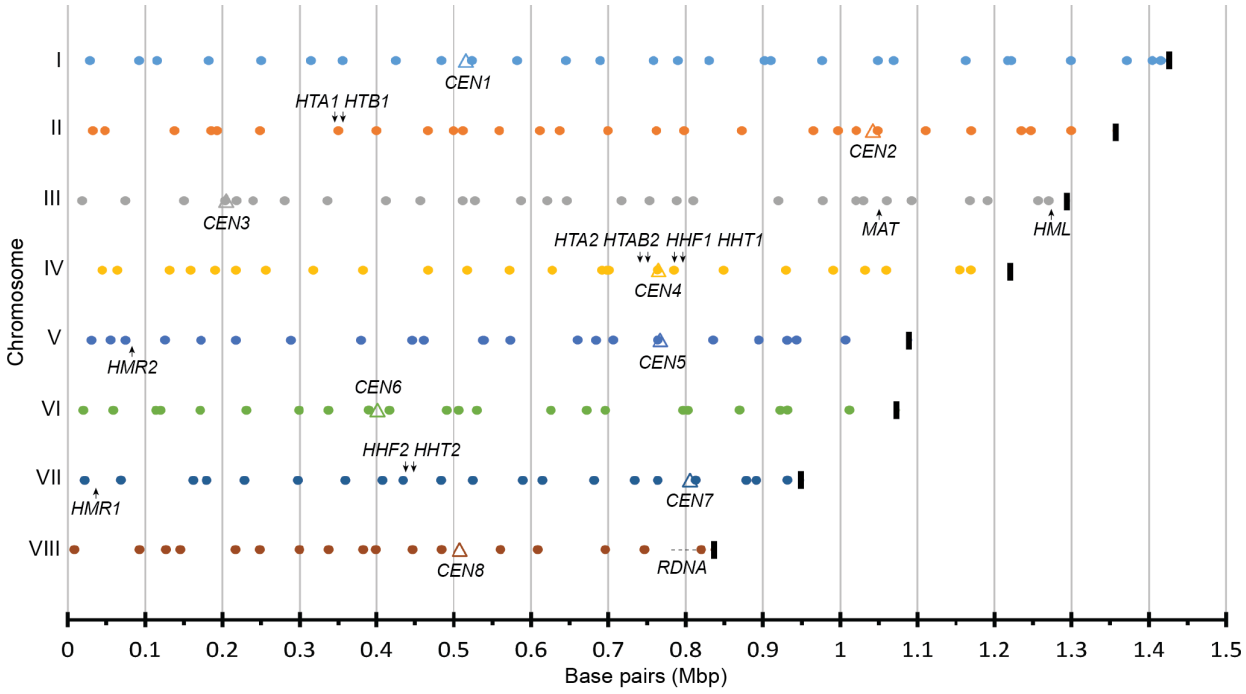
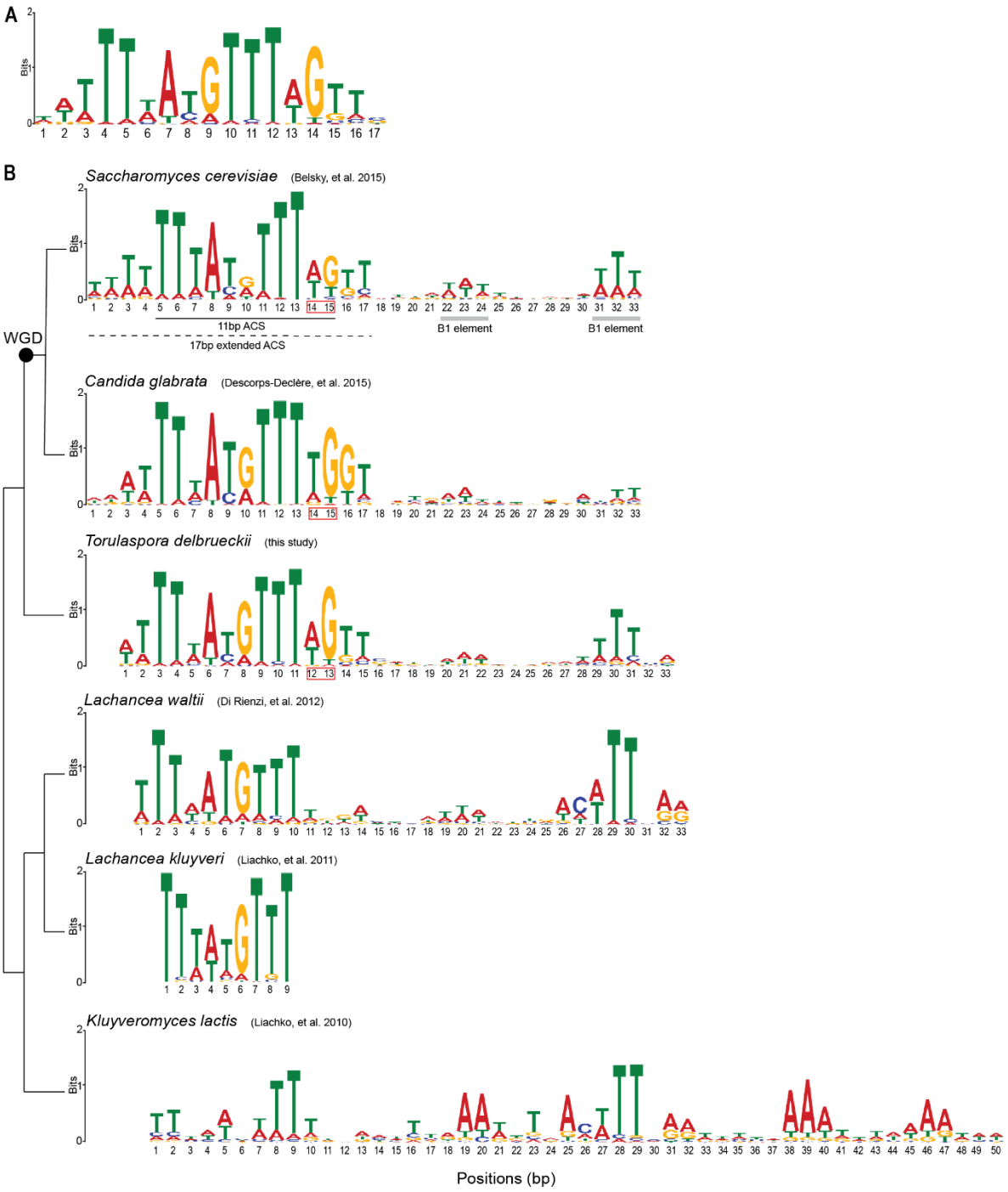


Fig. 2. Genome-wide distribution of ORC binding sites. Dots represent the 190 ORC binding sites on each of the eight *T. delbrueckii* chromosomes. Thick vertical black lines indicate the right end of each chromosome, and Δ indicates the centromere. The histone genes, *HTA*, *HTB*, *HHF*, and *HHT*, are indicated with black arrows. The mating type loci, *MAT*, *HML*, and *HMR*, and the rDNA repeat locus (dashed line) are genomic landmarks shown for reference. Every major unit on the scale represents 100,000 base pairs.

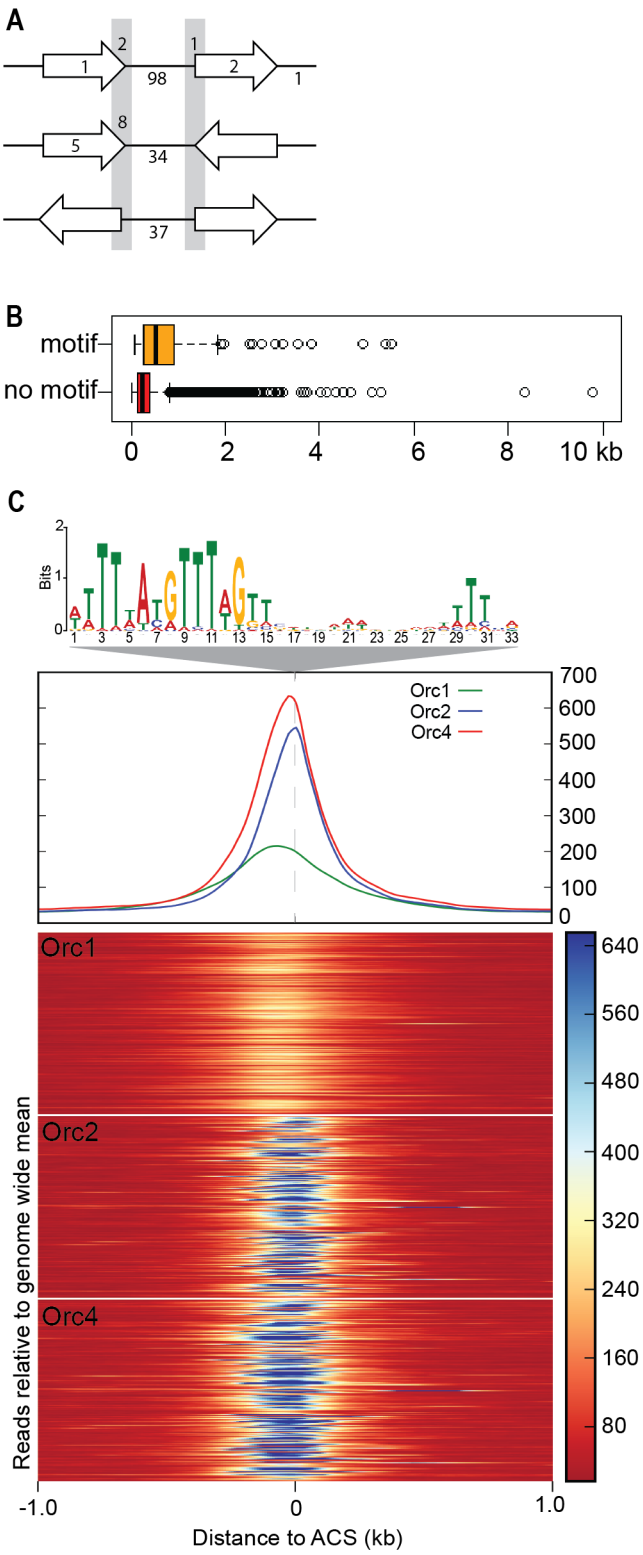


C

Orc4 insertion α-helix

```
S.cerevisiae_Orc4   455 DVKNVWENLVQLDFFTEKSAVGLRDNATAAFYASNYQFQGTMIPFDLRSYQM 506
C.glabrata_Orc4     465 DIKNVWESLLSLGFLAEKSDVGLRESAMAVFYASNYQLQGNALPYDLRVYQM 516
T.delbrueckii_Orc4  497 DVKNVWETLINLNFLTEKSAVGLRESAIAVFYASNYQLQGTTIPFDLRLYQT 548
L.waltii_Orc4       425 DVKNVWENLIDLDFITEKGSVGLRASALAAFMASNYQSTGTIIPFDLRTYQM 476
L.kluyveri_Orc4     597 DIKNVWENIVELNFLSEKGAVGLRVSAQAAFQASNYHTTGTTIPFDLRMYQL 648
K.lactis_Orc4       474 DMKNVWENLANQELITERGAISVRFTEDSAIQSANYNSSTMKAPFDLRTYQV 525
                        *:*****.: .  :::*:. :.:* .  :.: ::**:      *:*** **
```

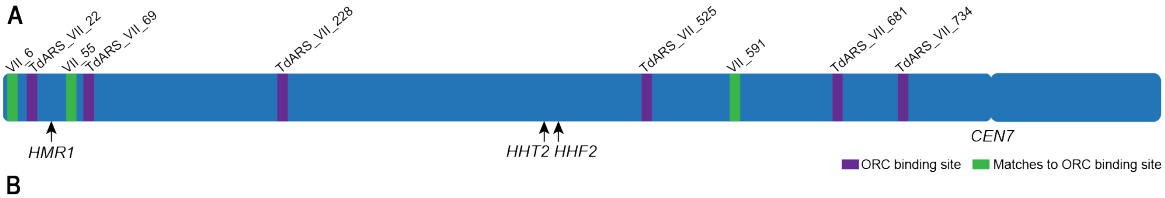
Fig. 3. Comparison of sequence motifs associated with *T. delbrueckii* ORC binding sites and other known origins of replication. (A) A 17bp motif was found in 187 of 190 *T. delbrueckii* ORC binding sites by MEME (Bailey, et al. 2009) when the motif length was not specified. **(B)** A longer motif was found in 189 of 190 predicted *T. delbrueckii* ORC binding sites when the motif length was fixed to 33. This motif is compared to known motifs from *S. cerevisiae* (Belsky, et al. 2015), *C. glabrata* (Descorps-Declère, et al. 2015), *L. kluyveri* (Liachko, et al. 2011), *K. lactis* (Liachko, et al. 2011) and *L. waltii* (Di Rienzi, et al. 2012). The cladogram representing the relationships of the species (Marcet-Houben and Gabaldon 2015) is not to scale. WGD denotes the whole genome duplication. For *S. cerevisiae*, the 11 bp ACS is underlined by a solid black line, the extended ACS /A element is underlined by a black dotted line, and the B1 elements are underlined by gray boxes. **(C)** Alignment of a portion of Orc4 from the same species. The DNA-binding alpha helix is highlighted grey, and the tyrosine (Y) that contacts nucleotides 14 and 15 in *S. cerevisiae* and likely nucleotides 12 and 13 in *T. delbrueckii* (red boxed in 3B) (Hu, et al. 2020) is highlighted red.



622 **Fig. 4. The ORC-binding motif is preferentially located between genes and coincides with the peak of**

623 **ORC binding. (A)** Position and number of ORC binding motifs relative to the flanking genes. The arrows

indicate the transcription direction of the genes. Motifs included within the vertical grey bars span the junction between the gene and the intergenic region. One ORC binding motif occurs beteween the last gene and the end of the chromosome (top line). **(B)** Box and whisker plot representing the sizes (kb) of intergenic regions with and without an ORC binding site. Each box shows the 25th, 50th and 75th percentile. The vertical line in each box represents the median size. The whiskers show 10th and 90th percentiles. **(C)** Aggregate plot (middle) and heat maps (bottom) of ChIP-Seq signals for Orc1, Orc2 and Orc4 are shown centered on the 33 bp motif (top) generated using MEME. Data for each binding site is oriented with the ACS to the left and B1 element to the right. The y-axis for the aggregrate plot and heat maps represents average read depth of sequences generated by ChIP-Seq. The x axis represents ChIP-Seq signals across 1 kb upstream and downstream of the 33 bp motif.



B

ARS	Genomic location	Fold Enrichment			ORC binding site (17bp motif generated by MEME analysis)
		Orc1	Orc2	Orc4	
VII_6	NC_016507.1:6009	9.68	2.69	2.33	GGCGAAATCT ATTTTTTTTTTAATTTC GCTGCGATTA
TdARS_VII_22	NC_016507.1:21702	4.59	2.79	9.23	TGAAAAAGTA CAATTGATGTTTAGTAC TCCAAATCTT
VII_55	NC_016507.1:55285	-	-	-	AAACAGTCAT TTGAAACATTTAGTATG ACTGTCTTTT
TdARS_VII_69	NC_016507.1:68800	8.27	6.15	17.88	TTATTTCTTA AATTTAATGTTTTATTG GTGTAAGCAG
TdARS_VII_228	NC_016507.1:228278	12.72	15.86	32.94	TACTACACTA ATTTTTATGTTTTGTTC AAAAAATTGG
TdARS_VII_525	NC_016507.1:524812	3.72	21.20	10.11	TCTTGCATTG CATTTGAGGTTTAGTTG GGAATTATAG
VII_591	NC_016507.1:591055	-	2.25	-	GAGCCTGGGG TATTAATGTTTTGGCCT TCGAAGCTTT
TdARS_VII_681	NC_016507.1:681431	10.97	8.64	56.04	CTTACTTGCT TATTTTATATTTAGTTG ACAGATACAT
TdARS_VII_681Δ	NC_016507.1:681431	N.D.	N.D.	N.D.	CTTACTTGCT ACAGATACAT
TdARS_VII_734	NC_016507.1:734153	3.93	4.82	11.34	TGTTTTTTAT ATCTTGACGTTTAGTTG TTGGTTTTGG

Fig. 5. ORC-binding sites on chromosome VII tested for ARS activity. (A) Diagram of *T. delbrueckii* chromosome VII. Purple bars represents the six (out of 21) ORC binding sites on chromosome VII that were tested for ARS activity. Green bars represent three sites with matches to the motif but low or no

enrichment of ORC subunits. The black arrows indicate the position of the cryptic mating type locus, *HMR1*, and the histone genes, *HHT2* and *HHF2*. **(B)** For the nine sites that were tested, the table summarizes the genomic location, fold enrichment of Orc1, Orc2, and Orc4, and the match to the 17 bp consensus sequence identified by MEME (proposed ACS; purple).

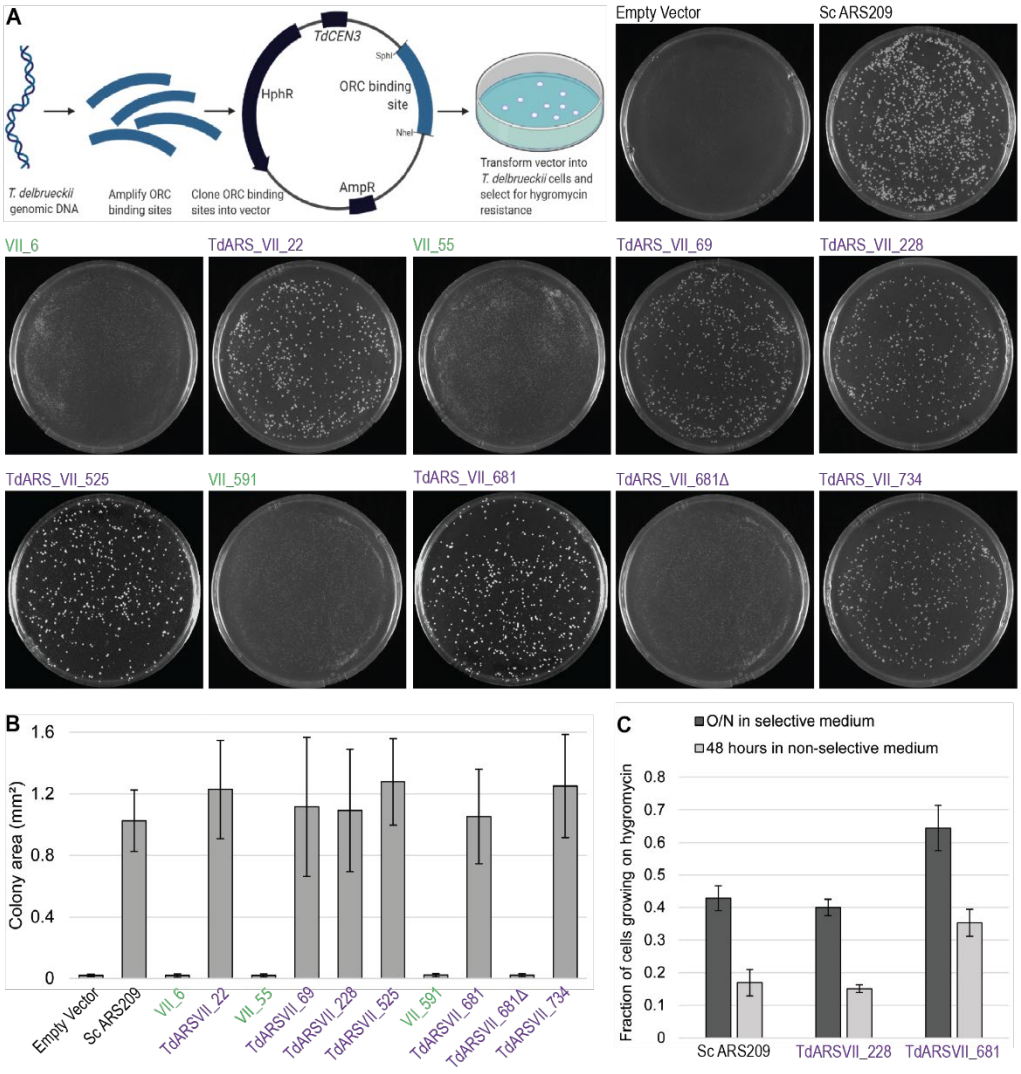


Fig. 6. ORC binding sites act as autonomously replicating sequences (ARSs) on plasmids. (A) To test for ARS activity, a hygromycin-sensitive *T. delbrueckii* strain (JRY10156) was transformed with plasmids bearing various ORC binding sites and a hygromycin-resistance gene. Hygromycin-resistant colonies were imaged after two days. Controls were plasmids lacking an ARS (empty vector) and bearing *Sc*ARS209, a

645 *S. cerevisiae* origin known to function in *T. delbrueckii*. **(B)** The average area of the colonies shown in
646 panel A was calculated using Image J (Schneider, et al. 2012). Error bars represent standard deviations
647 between three replicate experiments. **(C)** Plasmid loss assays were performed on TdARS_VII_228,
648 TdARS_VII_681, and ScARS209. The fraction of cells able to grow on hygromycin, and therefore
649 containing plasmids with the hygromycin-resistance gene, was determined before and after growth in non-
650 selective medium (YPD) for 30 hours. Error bars represent the standard error of the mean of three replicate
651 experiments.