bioRxiv preprint doi: https://doi.org/10.1101/2020.07.04.187286; this version posted July 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Evolution of DNA Replication Origin Specification and Gene Silencing Mechanisms

Hu, Y.[1,2], Tareen, A.[3], Sheu, Y-J.[1], Ireland, W. T.[5], Speck, C.[6], Li, H.[7], Joshua-Tor, L. [1,4], Kinney, J. B.[3] and Stillman, B.[1]*

Author Affiliations:

1. Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA.
2. Program in Molecular and Cell Biology, Stony Brook University, Stony Brook, NY 11794, USA.
3. Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA
4. W. M. Keck Structural Biology Laboratory, Howard Hughes Medical Institute, Cold Spring Harbor, NY, 11724, USA.
5. Department of Physics, California Institute of Technology, Pasadena, California, USA
6. DNA Replication Group, Institute of Clinical Sciences, Faculty of Medicine, Imperial College London, London W12 0NN, United Kingdom
7. Structural Biology Program, Van Andel Institute, Grand Rapids, MI 49503, USA

* Corresponding Author stillman@cshl.edu

Abstract

DNA replication in eukaryotic cells initiates from chromosomal locations, called replication origins, that bind the Origin Recognition Complex (ORC) prior to S phase. Origin establishment is guided by well-defined DNA sequence motifs in *Saccharomyces cerevisiae* and some other budding yeasts, but most eukaryotes lack sequence-specific origins. At present, the mechanistic and evolutionary reasons for this difference are unclear. A 3.9 Å structure of *S. cerevisiae* ORC-Cdc6-Cdt1-Mcm2-7 (OCCM) bound to origin DNA revealed, among other things, that a loop within Orc2 inserts into a DNA minor groove and an α-helix within Orc4 inserts into a DNA major groove[1]. We show that this Orc4 α-helix mediates the sequence-specificity of origins in *S. cerevisiae*. Specifically, mutations were identified within this α-helix that alter the sequence-dependent activity of individual origins as well as change global genomic origin firing patterns. This was accomplished using a massively parallel origin selection assay analyzed using a custom mutual-information-based modeling approach and a separate analysis of whole-genome replication profiling and statistics. Interestingly, the sequence specificity of DNA replication initiation, as mediated by the Orc4 α-helix, has evolved in close conjunction with the gain of ORC-Sir4-mediated gene silencing and the loss of RNA interference.

Main

In the budding yeast *S. cerevisiae*, replication origins are specified by DNA sequence motifs that comprise an essential A element (about 11nt in length) and multiple B elements[2]. Such sequences enable the replication of extrachromosomal plasmids and are thus termed autonomously replicating sequences (ARSs)[3]. By contrast, replication origins are sequence non-specific in plants and animals, in the fission yeast *Schizosaccharomyces pombe*, and even in many other budding yeasts and fungi[4]. Nevertheless, the proteins involved in the initiation of DNA replication are highly conserved. In eukaryotes, the six subunit ORC complex (comprising Orc1-6) assembles on DNA prior to S-phase[5]. ORC then recruits Cell Division

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.04.187286; this version posted July 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Cycle 6 (Cdc6), chromatin licensing and DNA replication factor 1 (Cdt1), and the replication helicase subunits Mcm2-7 to form a pre-Replicative Complex (pre-RC)[6]. In prior work, a structure of a pre-RC assembly intermediate containing the *S. cerevisiae* ORC-Cdc6-Cdt1-Mcm2-7 (OCCM) bound to origin DNA (*ARS1*) was determined at ~3.9 Å by cryo-electron microscopy[1]. This structure revealed multiple OCCM-DNA interactions, including an Orc4 α-helix inserted into the DNA major groove and an Orc2 loop inserted into the minor groove (Fig. 1a). These interactions were subsequently confirmed by a higher resolution ORC-DNA structure[7]. We note that a lysine-rich region of Orc1 interacts with DNA in this latter structure but not in the OCCM, suggesting considerable plasticity in origin recognition during pre-RC assembly.

Interestingly, the Orc4 α-helix and Orc2 loop have evolved in a manner that parallels the evolution of origin sequence specificity. Sequence alignments suggest that these features have been acquired in a sub-group of *Saccharomyces*-related budding yeasts, but are absent in all other eukaryotes including other budding yeasts, other fungi (including *S. pombe*), plants, and animals (Fig. 1c, Extended Data Fig. 1a). High resolution structures of Human[8] and *Drosophila*[9] ORC show the lack of the Orc4 α-helix and Orc2 loop (Fig. 1b, Extended Data Fig. 1b). The Orc4 α-helix and Orc2 loop are present but diverged in some other budding yeasts, such as *Kluyveromyces lactis*, which has sequence-specific origins that exhibit a DNA sequence motif that differs from the *S. cerevisiae* motif.

These observations suggest that the Orc4 α-helix and/or Orc2 loop might play key roles in origin sequence specificity. To investigate which specific residues might be involved, 32 individual Orc4 α-helix mutants and 7 Orc2 loop mutants were examined using plasmid shuffle assay (Fig. 1d-e, Extended Data Fig. 2, 3; see Methods). The Orc2 loop mutants were either lethal, had strong defects, or had little effect (Fig. 1e, Extended Data Fig. 3). Deletion of the Orc4 α-helix or its replacement with the 13-amino acid *K. lactis* Orc4-α-helix were lethal. Other Orc4 α-helix mutants led to different levels of growth deficiency (Fig. 1d, Extended Data Fig. 2). Based on the growth deficiency phenotype, nine viable Orc4 mutants were chosen for further detailed analysis. To perform the genetics in a complete manner, two conservative mutations (orc4$^{F485Y,\ Y486F}$ and orc4^{R478K}) were chosen for comparison. The wild type and Orc4 mutants were tagged at the amino terminus (NTAP-tag) and integrated into the genome as the sole Orc4 subunit that formed a functional ORC (Extended Data Fig. 4). Some strains with the integrated version of the mutant Orc4 (strains G, Extended Data Fig. 6) proliferated far better compared to strains that relied on a mutant Orc4 subunit that was expressed from a single origin minichromosome (strains P, Extended Data Fig. 6). Five of these mutants were created to investigate the FY residues at positions 485-486, which have evolved to IQ in *K. lactis*. Specifically, orc4^{Y486Q}, orc4^{F485I}, and orc4$^{F485I,\ Y486Q}$ were used to study the effects of these evolutionary changes, while orc4$^{F485A,\ Y486A}$ and orc4$^{F485Y,\ Y486F}$ (a conservative swap of aromatic amino acids) were used to investigate these residues more generally. The other four mutants, orc4^{R478A}, orc4^{R478K}, orc4^{N489A}, and orc4^{N489W}, were chosen to investigate the roles of R^{478} and N^{489}, two conserved residues at opposite ends of the α-helix that we predicted to mediate both protein-protein contacts and contacts with DNA backbone phosphates. Some of these Orc4 mutants exhibited slower growth rates (Fig. 1f) and slowed passage through S phase and mitosis (Extended Data Fig. 5). To better understand the effects of these mutations on origin activity and specificity, we performed two complementary, but independent deep-sequencing-based assays: massively parallel origin mutagenesis on plasmids and genome-wide DNA replication profiling.

Massively parallel origin mutagenesis. To quantify the sequence-dependent activity of ORC at specific origins of interest, we performed a massively parallel origin selection assay (MPOS assay) on two different origins in wild type and nine yeast strains harboring the Orc4 variants. 150 base-pairs of either *ARS1* (also

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.04.187286; this version posted July 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

known as *ARS416*) or *ARS317* DNA were synthesized at a 15% per-nucleotide substitution rate and cloned into plasmids that carried a selective marker[10,11] (Fig. 2a). These two plasmid libraries were then separately transfected into the ten strains of yeast described above, and the mutated ARSs that remained after multiple cell divisions were sequenced. A custom motif inference algorithm, based on mutual information maximization, was then applied to these sequence data and used to infer quantitative motifs describing origin activity in each strain. This algorithm proved to be essential for the correct analysis of the data (see below).

Some mutants, such as orc4$^{F485Y, Y486F}$ and orc4^{R478K}, yielded motifs very similar to WT (Fig. 2b for *ARS1*, Extended Data Fig. 7a for *ARS317*). Other strains, such as the orc4^{N489A}, orc4^{N489W}, orc4^{R478A} and orc4$^{F485I, Y486Q}$ retained a far less diverse set of mutant ARSs and yielded noisier motifs that, nevertheless, remained relatively similar to the WT ARS consensus sequence (Fig. 2b for *ARS1*, Extended Data Fig. 7a for *ARS317*). However, two Orc4 α-helix mutants exhibited robust changes to their ARS motifs: in both the orc4$^{F485A, Y486A}$ and orc4^{Y486Q} mutants, and for both the *ARS1* and *ARS317* experiments (Fig. 2b and c, Extended Data Fig. 7), two dinucleotides present in the WT consensus sequence motif at position 29-30 changed. In the *ARS1* experiments, motif A/T G/T has been switched to T/A C/T in the orc4$^{F485A, Y486A}$ strain and switched to T/A T/G in the orc4^{Y486Q} strain (Fig. 2c). Substantial changes were observed at the same position in the *ARS317* MPOS assay (Extended Data Fig. 7b). A Principal Component Analysis (PCA) of the motifs inferred from multiple biological replicates confirmed that reproducible changes in motifs indeed resulted from the mutations in question (Extended Data Fig. 8a). A quantitative analysis of the mutual-information-based motif inference method (IM) compared to the standard enrichment ratio calculation (EM), showed that the new mutual-information-based motif inference method was essential for resolving these mutation-dependent changes in origin specificity (Extended Data Fig. 8).

The structural basis for origin sequence specification. Observations using the ~3Å high-resolution structure[7] of Orc4 α-helix on DNA (Fig. 2d, Supplementary Video 1) can further rationalize the mutation-dependent change in origin specificity. We suggest that Y486 interacts with the DNA base C/A30 (Fig. 2f), which is on the complementary strand from the A/T G/T dinucleotide whose readout is altered in the orc4$^{F485A, Y486A}$ and orc4^{Y486Q} mutants (Figure 2c, purple box) in a face-to-edge T-type π interaction, as often seen in protein-DNA interfaces[12]. In addition, F485 sits against a hydrophobic stretch comprised of the methyl groups emanation from a run of T's, T26-29 (Fig. 2e).

At each end of the Orc4 insertion α-helix are amino acids that we suggest provide affinity for ORC to DNA as well as help position this sequence-reading α-helix correctly in the major groove. R478 interacts with A487 on the α-helix as well as the adjacent V475 and could have an alternative conformation whereby contacts DNA phosphate (Fig. 2e and f). Even a conservative amino acid substitution R478K was slightly deficient (Fig. 1f and Extended Data Table 1) and had a subtle change in the genome wide origin firing pattern (Extended Data Fig. 11d and Fig. 3c) indicating that there is an additional role for the R478 side group that the charge and length of the lysine side group cannot fully replace the arginine at this location. Likewise, N489 that sits at the opposite end of the α-helix would be in range of a DNA phosphate contact and even a DNA base contact with T28 upon minor adjustments of the model that are well within the EM density[7] (Fig. 2e). Mutation of either amino acid had the largest, non-lethal effect (Fig. 1) and thus they could play key roles in both positioning the α-helix in the major groove and contributing to affinity of ORC to DNA.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.04.187286; this version posted July 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Genome-wide replication origin profiling. To investigate the mutation-dependent origin usage changes in natural genomic replication origin profiling, cells were arrested in G1 phase and released into S phase in the presence of hydroxyurea (HU). HU treatment restricts (via checkpoint signaling) origin firing to those origins that normally become active in early S phase, and prevents the activation of origins that normally fire later. The addition of 5-Ethynyl-2'-deoxyuridine (EdU), followed by purification of EdU-labeled DNA and high-throughput DNA sequencing, was then used to map the locations and activities of early origins throughout the genome (Extended Data Fig. 9 and Fig. 3c). Mrc1 is a replication fork associated protein and mediator of intra-S phase checkpoint signaling. All origins fired, as expected, when the *MRC1* gene was deleted. We observed more origins firing than have been confirmed in oriDB[13]. This Orc4^{WT} *mrc1*Δ profile thus reveals a maximal set of possible origins against which to compare the replication profiles of NTAP tagged Orc4 integrated strains. As expected, only a subset of the maximal set of possible origins fired in wild type (WT) cells containing the NTAP-tagged Orc4^{WT} protein. The genome-wide replication origin profiles were reproducible in biological replicates (Extended Data Fig. 9b for WT and *mrc1*Δ; Fig. 3c and Extended Data Fig. 9c for all mutant Orc4 strains).

Only two completely "de novo" replication origin locations were found in genomic origin firing profiles of all nine orc4 mutant strains, which were neither predicted to be ARS locations in OriDB nor exist in the Orc4^{WT} *mrc1*Δ profile. One was found in orc4$^{F485I, Y486Q}$ (Fig. 3a) and another was found in orc4$^{F485A, Y486A}$ strain (Fig. 3b). Indeed, we didn't expect to see a dramatic change in de novo origin locations because the mutations we made are only single or double point-mutations and would not be expected to create tremendous amount of "de novo" replication origin locations. However, extensive changes in the Orc4 α-helix, such as orc4$^{Klα\text{-}helix}$ or orc4$^{Δα\text{-}helix}$, were lethal and therefore could not be assessed for "de novo" origins.

Despite the very few "de novo" replication origin locations, genomic the origin firing pattern changed considerably in some strains (Extended Data Fig. 11b-j). There were numerous origins that are active in both *mrc1*Δ and the wild-type strain (aka. early origins) but were specifically repressed in the orc4 mutant strains (Fig. 3f and Extended Data Fig. 11a-j, orange arrow direction). At the same time, there were numerous origins that were inactive in the wild-type strain but were activated in the *orc4* mutant strains (Fig. 3e and h, green arrows and Extended Data Fig. 11a-j, green arrow direction). The activation or repression pattern is mutant dependent. The chromosome IV profiles for the nine *orc4* mutant strains was used as an example chromosome to show more details (Fig. 3c, Extended Data Fig. 9c). While some origins, either those that normally fire early or late, did not change (Fig. 3c, g and i, black arrows), other origins changed their firing pattern in Orc4 mutant strains. For example, an active origin in WT and most of the mutant strains was not active in the orc4$^{F485I, Y486Q}$ and orc4$^{F485A, Y486A}$ strains (Fig. 3c and f, green arrow). In contrast, many origins that normally do not fire in HU in WT became active in multiple mutants (Fig. 3c, e and h, green arrows). The firing pattern was mutant dependent. Interestingly, two origins on Chr. IV were active only in the orc4$^{F485I, Y486Q}$ mutant strain, in which the IQ are the amino acids that exist in *K. lactis* (Fig. 3c, e and i, red arrows). When, however, the conservative orc4$^{F485Y, Y486F}$ double mutant was analyzed, the firing pattern was like WT.

There are chromatin and chromosome location context factors that were suggested in previous studies to play roles in controlling the origin timing[14,15]. Our genome-wide statistics result supports this idea. Origin firing peak heights generally do not correlate with how well they match the ACS motifs (Extended Data Fig. 12, section A). Only when the origin sequence recognition is reduced/disturbed, such as in the orc4 F485 and Y486 mutants, the origin locations that originally can have high origin activities are then

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.04.187286; this version posted July 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

severely affected (Extended Data Fig. 12, section B). The orc4 R478 and N489 mutations had a slower growth phenotype and this is likely due to reduced origin licensing, consistent with the proposal that the Orc4 α-helix that is not positioned correctly due to amino acid changes at each end of the α-helix, making it more difficult to stably position within the DNA major groove. As a consequence, poor DNA interaction would lead to overall much fewer origins that become active (Extended Data Fig. 11 q-t support this idea) resulting in larger replicon size (i.e., greater inter-origin distance) and a slower doubling time (Fig. 1f). These mutants would be forced to use late origins under HU to survive. This is also consistent with the observation that the peak widths on average for these slow growing mutants was larger than WT because the smaller number of origins utilized would not be restrained by rate limiting replication factors, which are known to exist [16,17]. We suggest that the mutations at Orc4 R478 and N489 would not cause sequence specificity changes but instead general inefficient binding to the origins. Indeed, the correlation between the origin firing peak heights from genome-wide origin data and the sequence motif matching quality scores from the MPOS *ARS* selection assays did not correlate well in the Orc4 R478 and N489 mutant strains compared to those strains (e.g., orc4$^{F484I,\ Y486Q}$ and orc4$^{F485A,\ F486A}$) that change DNA sequence specificity (Extended Data Fig. 12, section C). We suggest that mutations that affect the affinity and sequence specificity of ORC for origin DNA, coupled with the known influence of chromatin context on origin timing[14,15], combine to dramatically change which origins are utilized in the genome.

The recruitment of MCM2-7 to replication origins was analyzed by Mcm2 chromatin-immunoprecipitation (ChIP) in G1 phase (Fig. 3d, Extended Data Fig. 10b). The ChIP-Mcm2 results were reproducible in biological replicates (Extended Data Fig. 10a) and correspond well to the replication origin profiles (Fig. 3), although it is known that Mcm2-7 can move on chromosomes once loaded [18]. A particularly interesting case is the origin switch in the orc4$^{F485I,\ Y486Q}$ mutant strain (Fig. 3d and e, two red arrows at right), where the Mcm2 binding also switched, albeit not completely. Another interesting case is the inactive late origins in HU in WT that have lower or no ChIP-Mcm2 signal in orc4$^{F485I,\ Y486Q}$ mutant strain are not active origins in the mutant strain (Fig. 3e and h, blue arrows). Combined, these results suggest that the Orc4 α-helix is a major determinant of origin utilization in the genome.

The DNA sequence statistical analysis for genome-wide origin firing pattern changes. We examined the DNA sequences predicted to be ACS from OriDB[13,37] under each EdU peak and performed a genome-wide statistical analysis. The result shows that, specifically in the Y486 mutant strains (orc4$^{F485I,\ Y486Q}$, orc4$^{F485A,\ Y486A}$ and orc4^{Y486Q}), the origin firing peak heights were significantly reduced when the dinucleotide sequence "AG" at the position corresponding to the ARS consensus sequence (ACS) 29-30 nucleotides in the motifs that we determined from MPOS assay (Fig. 2c, Fig. 4a, c-e). Hereafter, these positions in the genomic ACS were defined as "position 29-30" for easier reference. The Orc4^{WT} strain, together with the rest of the mutant strains, have relatively equal origin firing peak height regardless of the origin dinucleotide sequence at position 29-30 (Fig. 4a-b, Extended Data Fig. 13). This, both the MPOS assay data (Fig. 2c) and the analysis of the ACS in the origins used in the genome, show that Orc4 F485 and Y486, especially Y486, are essential for recognizing origins with the "AG" dinucleotide sequence at position 29-30 These mutants effectively reduce the chances of utilizing origins with the "AG" sequence. These data show that the ORC4 α-helix defines the sequence specificity and hence location of active origins in the genome.

Co-evolution of DNA Replication Origin Specification and Gene Silencing Mechanisms. The data using both whole genome analysis and the MPOS assays demonstrate that the ORC4 α-helix contribute to selection of origin sequences in the yeast genome, as predicted by the structure of the OCCM[1] and

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.04.187286; this version posted July 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

ORC[7] on origin DNA. The conservation of the α-helix and loop is restricted to a small clade of *Saccharomyces*-related species and where it has been determined, corresponds to the origins of DNA replication having a demonstrable consensus sequence (Raguraman, M.K and Liachko, I. in Kaplan[19]) (Fig. 5).

It is known that in some budding yeasts, such as *S. cerevisiae* and *K. lactis*, ORC, functioning with Silent Information Regulator (SIR) proteins, is also required in transcriptional gene silencing of mating type loci, rDNA and telomeres[20-22]. Evolutionally, Sir2 and Sir4 preceded the acquisition of Sir1 and Sir3 (which is related to Orc1) in the ORC-Sir4-mediated transcriptional gene silencing pathway. In *K. lactis*, Sir4 binds directly to Orc1 but in S. cerevisiae, Sir1 binds to Orc1 and Sir4 binds to Sir3 that arose from Orc1 as a result of whole genome duplication[22] (WGD, Fig. 5). In both species, Sir2 is required, as its histone deacetylase activity is essential for the gene silencing function. Interestingly, Sir4 binds to Esc1 which is located in the nuclear envelope, tethering the silent loci to the nuclear periphery[23]. Of relevance here is that Sir4 is related in structure to nuclear lamins, which are present in most eukaryotes but are absent in yeast[24].

We observed a very interesting co-evolution of origin sequence specification (Fig. 5, first three columns) and gene silencing (Fig. 5, remaining columns). The acquisition of sequence specific origins, the Orc4 α-helix and the Orc2 loop (Fig. 5, blue shadow) correlated precisely with the acquisition of ORC-Sir4-mediated transcriptional gene silencing (Fig. 5, yellow shadow).

Dicer is an RNase III family member and a key mediator in the RNA interference (RNAi) pathway, which has been shown to control gene silencing by transcriptional and post-transcriptional mechanisms[25]. However, Dicer, but not other components of the RNAi pathway, has an RNAi-independent role in *S. pombe* in the termination of transcription at replication stress sites[26]. This may contribute to alleviation of R-loop mediated conflicts between DNA replication and transcription, particularly in repeated sequences and heterochromatin. The vast majority of eukaryotes that lack sequence-specific origins, including plants, animals and the majority of fungi including yeast have vast repeated sequences and heterochromatin and thus need RNAi[25] (Fig. 5, orange shadow) or Dicer's RNAi-independent role in maintaining genome stability, particularly if origin locations are stochastic, as has been shown in *S. pombe*[27].

Budding yeasts that lack sequence-specific origins, such as the pathogenic yeast *Candida albicans* and the industrial yeast *Yarrowia lipolytica* that can metabolize unusual hydrocarbons, have lost, or are in the process of losing RNAi[28,29]. Some retain a non-canonical Dicer (Dcr*) that has an RNase III domain and has been shown in *C. albicans* to exhibit RNAi to silence transposable elements and sub-telomeric repeated sequences[28]. They lack both the Orc4 α-helix and the Orc2 loop and do not have demonstratable sequence-specific origins. In this context, *Y. lipolytica* is an interesting case since it has lost Dicer and Argonaute (Ago) and lacks ORC-Sir4 silencing and sequence-specific origins. *Y. lipolytica* has dispersed rDNA gene clusters that are sub-telomeric and has a relatively low gene density (one gene per 3.3kb), far lower than the gene density found in *S. cerevisiae* (one gene per 2kb)[30]. It also uses Tay1, a TRF-like protein for telomeric and sub-telomeric gene silencing, which is more similar to the human shelterin complex mechanism[31]. Moreover, it is a heterothallic yeast, which does not switch its mating type and thus lacks silent mating type loci[30]. We suggest and that *Y. lipolytica* may be a useful specifies to investigate origin location and sequence specificity and we are studying replication patterning in this species.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.04.187286; this version posted July 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

The budding yeasts that have acquired sequence specific origins and ORC-Sir4-mediated gene silencing system are likely to have lost RNAi completely, albeit some retained the non-canonical Dicer (Dcr*). One budding yeast, *T. delbrueckii*, has ORC-Sir4 silencing and has retained Dcr* and Ago, but the latter are not involved in transcriptional gene silencing[29]. As species lost RNAi with a concomitant reduction in repeated sequences in the genome, including centromere associated repeated sequences, we suggest that in the *Saccharomyces*-related, ORC-Sir4 containing budding yeast that ORC evolved to bind DNA in a sequence specific manner, providing a mechanism to locate origins of DNA replication in intergenic regions[6]. Such a location would help maintain genome stability by reducing the possibility of conflicts between DNA replication and transcription, including the formation of R-loops[32]. The remaining repeated sequences in these species, such as the silent mating type loci in homothallic yeast, have evolved to be protected from loss by recombination and be transcriptionally silenced by an ORC-Sir4 dependent recruitment of the histone deacetylase Sir2. It is possible that in *Y. lipolytica*, Sir2 binds directly to ORC and thus bypasses the requirement for the other SIR proteins.

In *S. pombe*, Orc4 has an AT-hook DNA binding domain at its amino-terminus that localizes initiation of DNA replication to AT-rich sequences in the genome[33] even though replication origin utilization throughout the genome is known to be stochastic[27]. We found similar sequences are present in Orc4 in many fungi, including *Neurospora crassa*. Since the AT-hook sequences and the Orc4 α-helix and Orc2 loop are absent in other fungi, animals and plants, they must have an alternative mechanism of specifying origin location, a topic of major interest.

ORC is involved in maintenance of heterochromatin in *Drosophila* and Human, via an interaction between ORC1 and the heterochromatin protein HP1[34,35]. Furthermore, ORC in Human cells is also involved in repression of transcription of the CCNE1 gene encoding Cyclin E via interactions with the histone methyltransferase SUV39H1 and the Retinoblastoma tumor suppressor protein (Rb)[36]. Thus ORC-dependent gene silencing may exist outside of species that have acquired Sir4.

Main References

1. Yuan, Z. *et al.* Structural basis of Mcm2-7 replicative helicase loading by ORC-Cdc6 and Cdt1. *Nat Struct Mol Biol* **24,** 316–324 (2017).
2. Marahrens, Y. & Stillman, B. A yeast chromosomal origin of DNA replication defined by multiple functional elements. *Science* **255,** 817–823 (1992).
3. Stinchcomb, D. T., Struhl, K. & Davis, R. W. Isolation and characterisation of a yeast chromosomal replicator. *Nature* **282,** 39–43 (1979).
4. Prioleau, M.-N. & MacAlpine, D. M. DNA replication origins-where do we begin? *Genes Dev* **30,** 1683–1697 (2016).
5. Bell, S. P. & Stillman, B. ATP-dependent recognition of eukaryotic origins of DNA replication by a multiprotein complex. *Nature* **357,** 128–134 (1992).
6. Bell, S. P. & Labib, K. Chromosome Duplication in Saccharomyces cerevisiae. *Genetics* **203,** 1027–1067 (2016).
7. Li, N. *et al.* Structure of the origin recognition complex bound to DNA replication origin. *Nature* **355,** 1–22 (2018).
8. Tocilj, A. *et al.* Structure of the active form of human origin recognition complex and its ATPase motor module. *eLife* **6,** 1822 (2017).
9. Bleichert, F. & Berger, J. M. Crystal structure of the eukaryotic origin recognition complex. *Nature* **519,** 321–326 (2015).

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.04.187286; this version posted July 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

10. Liachko, I., Youngblood, R. A., Keich, U. & Dunham, M. J. High-resolution mapping, characterization, and optimization of autonomously replicating sequences in yeast. *Genome Res* **23,** 698–704 (2013).
11. Hoggard, T. *et al.* High Throughput Analyses of Budding Yeast ARSs Reveal New DNA Elements Capable of Conferring Centromere-Independent Plasmid Propagation. *G3 (Bethesda)* **6,** 993–1012 (2016).
12. Wilson, K. A., Kellie, J. L. & Wetmore, S. D. DNA-protein π-interactions in nature: abundance, structure, composition and strength of contacts between aromatic amino acids and DNA nucleobases or deoxyribose sugar. *Nucleic Acids Res* **42,** 6726–6741 (2014).
13. Siow, C. C., Nieduszynska, S. R., Müller, C. A. & Nieduszynski, C. A. OriDB, the DNA replication origin database updated and extended. *Nucleic Acids Res* **40,** D682–6 (2012).
14. Stevenson, J. B. & Gottschling, D. E. Telomeric chromatin modulates replication timing near chromosome ends. *Genes Dev* **13,** 146–151 (1999).
15. Soriano, I., Morafraile, E. C., Vázquez, E., Antequera, F. & Segurado, M. Different nucleosomal architectures at early and late replicating origins in Saccharomyces cerevisiae. *BMC Genomics* **15,** 791 (2014).
16. Tanaka, S., Nakato, R., Katou, Y., Shirahige, K. & Araki, H. Origin Association of Sld3, Sld7,and Cdc45 Proteins Is a Key Stepfor Determination of Origin-Firing Timing. *Curr Biol* **21,** 2055–2063 (2011).
17. Lynch, K. L., Alvino, G. M., Kwan, E. X., Brewer, B. J. & Raghuraman, M. K. The effects of manipulating levels of replication initiation factors on origin firing efficiency in yeast. *PLoS Genet* **15,** e1008430 (2019).
18. Gros, J. *et al.* Post-licensing Specification of Eukaryotic Replication Origins by Facilitated Mcm2-7 Sliding along DNA. *Mol cell* **60,** 797–807 (2015).
19. Kaplan, D. L. *The Initiation of DNA Replication in Eukaryotes*. (Springer, 2016). doi:10.1007/978-3-319-24696-3
20. Bell, S. P., Kobayashi, R. & Stillman, B. Yeast origin recognition complex functions in transcription silencing and DNA replication. *Science* **262,** 1844–1849 (1993).
21. Fox, C. A., Loo, S. & Dillin, A. The origin recognition complex has essential functions in transcriptional silencing and chromosomal replication. *Genes Dev* **9,** 911–924 (1995).
22. Hickman, M. A. & Rusche, L. N. Transcriptional silencing functions of the yeast protein Orc1/Sir3 subfunctionalized after gene duplication. *Proceedings of the National Academy of Sciences* **107,** 19384–19389 (2010).
23. Grunstein, M. & Gasser, S. M. Epigenetics in Saccharomyces cerevisiae. *Cold Spring Harbor Perspectives in Biology* **5,** a017491 (2013).
24. Diffley, J. F. & Stillman, B. Transcriptional silencing and lamins. *Nature* **342,** 24–24 (1989).
25. Martienssen, R. & Moazed, D. RNAi and heterochromatin assembly. *Cold Spring Harbor ...* **7,** a019323 (2015).
26. Castel, S. E. *et al.* Dicer promotes transcription termination at sites of replication stress to maintain genome stability. *Cell* **159,** 572–583 (2014).
27. Dynamics of DNA replication in a eukaryotic cell. *Proceedings of the National Academy of Sciences* **116,** 4973–4982 (2019).
28. Drinnenberg, I. A. *et al.* RNAi in budding yeast. *Science* **326,** 544–550 (2009).
29. Ellahi, A. & Rine, J. Evolution and Functional Trajectory of Sir1 in Gene Silencing. *Mol Cell Biol* **36,** 1164–1179 (2016).
30. Yarrowia lipolytica. *Yeast* **29,** 409–418 (2012).

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.04.187286; this version posted July 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

31. Kramara, J. *et al.* Tay1 protein, a novel telomere binding factor from Yarrowia lipolytica. *Journal of Biological Chemistry* **285,** 38078–38092 (2010).
32. Hamperl, S., Bocek, M. J., Saldivar, J. C., Swigut, T. & Cimprich, K. A. Transcription-Replication Conflict Orientation Modulates R-Loop Levels and Activates Distinct DNA Damage Responses. *Cell* **170,** 774–786 (2017).
33. Chuang, R. Y. & Kelly, T. J. The fission yeast homologue of Orc4p binds to replication origin DNA via multiple AT-hooks. *Proc Natl Acad Sci USA* **96,** 2656–2661 (1999).
34. Pak, D. T. *et al.* Association of the origin recognition complex with heterochromatin and HP1 in higher eukaryotes. *Cell* **91,** 311–323 (1997).
35. Prasanth, S. G., Shen, Z., Prasanth, K. V. & Stillman, B. Human origin recognition complex is essential for HP1 binding to chromatin and heterochromatin organization. *Proc Natl Acad Sci USA* **107,** 15093–15098 (2010).
36. Hossain, M. & Stillman, B. Opposing roles for DNA replication initiator proteins ORC1 and CDC6 in control of Cyclin E gene transcription. *eLife* **5,** 10.7554–eLife.12785 (2016).
37. Nieduszynski, C. A., Knox, Y. & Donaldson, A. D. Genome-wide identification of replication origins in yeast by comparative genomics. *Genes Dev* **20,** 1874–1879 (2006).

Figure and Table Legends

Fig. 1 | DNA interacting Orc4 α-helix and Orc2 loop are essential. a, Top-view of ORC-Cdc6 structure encircling an origin DNA with Orc1-6 and Cdc6 indicated. Recolored from previous cryo-EM work[1] OCCM structure (PDB code 5udb) and Orc4 α-helix and Orc2 loop that interact with DNA are colored in red. **b,** Orc4 structure superposition among Human Orc4 in blue (from PDB code 5uj7), *Drosophila* Orc4 in grey (from PDB code 4xgc) and *S. cerevisiae* Orc4 in salmon (from PDB code 5udb). Orc4 α-helix that interacts with DNA is colored in red. **c,** Multiple sequence alignment of Orc4 among representing eukaryotic species as indicated. Orc4 α-helix region indicated with species that don't have sequence specific origins shadowed in blue and species that sequence specific origins exist shadowed in pink. **d-e,** Maps of mutant viability phenotype from plasmid shuffle assay: Orc4 α-helix in a helical wheel (**d**) and Orc2 loop in connected line (**e**). Mutant deficiency phenotypes (Extended Data Fig. 2 and 3) are summarized in color codes as indicated. Amino acids indicated in one-letter abbreviation. Different mutant types are indicated with different shapes. **f,** Growth curves of NTAP-Orc4 integrated strains (see Methods) in YPD with initiation OD600 at 0.05 at 30˚C to measure OD600 at different time points.

Fig. 2 | Selected origin sequence changes following a Massively Parallel Origin Selection (MPOS) assay. a, Schematic diagram of MPOS assay. **b,** ARS motifs for Orc4-integrated variants at A and B1 elements generated using an *ARS1 (ARS416)* variant library. See methods for how motifs are graphically rendered. **c,** Magnified view of the A element region in **b** from $Orc4^{WT}$, $orc4^{F485A, Y486A}$, $orc4^{Y486Q}$ strains. Dark purple rectangles indicate the major changes at positions 29-30 in the Orc4 mutant strains. **d,** Top-view of Orc4 α-helix insertion from ORC-DNA structure at 3Å (PDB code 5zr1) positioned in the DNA major groove. F485, Y486, N489 and R478 interact with DNA in base-specific (specificity) and base-nonspecific (affinity) manner. **e-f,** same as in **d,** but view in different angles. Red asterisks denote the base-specific interactions between amino acid and DNA base. Blue asterisks denote the base-nonspecific interaction between amino acid and DNA phosphate backbone. Green asterisks denote the interaction between amino acids. Prime symbols denote bases on the opposite strand. Bases numbering denotes the positions in logo (see **b**). **e** shows the hydrophobic interaction between F485 and T-rich region T26-T29, base-specific interaction between N489 and T28, base-nonspecific interaction between N489 and phosphate backbone of T28, and R478 interaction with V475. **f** shows the aromatic edge-face interaction

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.04.187286; this version posted July 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

between Y486 and A29'on the opposite strand, hydrophobic interaction between Y486 and C30', base-nonspecific interaction between R478 and phosphate backbone of A27', and R478 interaction with A487 and V475.

Fig. 3 | Orc4 α-helix mutants change the pattern of origin firing and MCM binding. Genome-wide origin firing profile, α-factor blocked and released into S phase in 200mM hydroxyurea (HU) for 90 mins. **a**, the completely new origin location ChrXII: 95,827-117,318 in orc4$^{F485I,\ Y486Q}$ strain. **b**, shows the completely new origin location ChrV: 248,069-251,910 in orc4$^{F485A,\ Y486A}$ strain. **c,** Whole genome replication profiles. Chromosome IV (ChrIV) is shown as a representation. Strains are in the order of shorter to longer doubling time (Extended Data Table 3) from top to bottom. **d-i,** ChIP profile of MCM (anti-Mcm2) of NTAP-Orc4 integrated strains at ChrIV in G1 phase and it's comparison to replication origins profile. orc4$^{F485I,\ Y486Q}$ strain is used as an example of NTAP-Orc4 mutant strains to compare with Orc4^{WT} strain in whole-chromosome view (**d**) and zoom-in views (**e-i**). Green arrows indicate the example locations of origins with firing pattern changes in the orc4$^{F485I,\ Y486Q}$ strain. At these locations, similar changes were also observed in other mutants (see **c**). Black arrows indicate the example locations of origins with firing pattern that remained the same in orc4$^{F485I,\ Y486Q}$ strain. At these locations, origins with firing pattern also remained the same in other mutants (see **c**). Red arrows indicate the example locations of origin firing pattern changes are unique in the orc4$^{F485I,\ Y486Q}$ strain but not in other mutant strains (see **c**). Blue arrows indicate the examples of inactive late origins in HU in WT that have lower or no ChIP-Mcm2 signal in orc4$^{F485I,\ Y486Q}$ mutant strain and are not active origins in the mutant strain.

Fig. 4 | Genomic origin firing pattern changes are sequence specific. DNA sequences under each origin replication peaks that are predicted to match the ARS consensus sequence (ACS) were obtained from OriDB[13,37]. Genome-wide statistical analysis was performed to check the dependence of origin firing peak height on dinucleotide identity at position 29-30 that correspond to logo derived from the MPOS data (Fig. 2c). P-values that correspond to a one-way ANOVA test. Asterisks denotation: *$p<0.05$, **$p<0.01$, ***$p<0.001$. Dinucleotides occurring in 3 or less annotated ACSs were removed prior to this analysis; a total of 211 ACSs were analyzed, while 9 were removed. **a,** Genome-wide statistical analysis results for all ten strains. **b,** box plots for Orc4 wild-type and three of the *orc4* mutant strains. Y-axis shows genomic origin firing peak heights in $\log_{10}$. Each dot denotes an annotated ACS. Box plots elements: the minimum height, first (lower) quartile, median, third (upper) quartile, and maximum height. Diamond denotes outliers that exhibited aberrantly large values. The "AG" dinucleotide at position 29-30 in the motif logo is not utilized in mutants that change Y486 to alanine or glutamine.

Fig. 5 | Co-evolution of DNA Replication Origin Specification and Gene Silencing. Phylogenetic tree is drawn based on whole genome vicinity and is in unscaled branches. Whole Genome Duplication (WGD) event is indicated in the tree. Table inspired by previous studies on gene silencing[29]. "+" indicates the exist of genes, "-" indicates the absent of genes, "*" indicates exist of noncanonical dicer. Orange box: species that use RNAi for gene silencing. Yellow box: species that use ORC and Sir proteins for gene silencing. Blue box: species that have sequence specific origins.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.04.187286; this version posted July 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Methods

Yeast genetic methods and strain construction

Yeast strains generated in this study (described in Supplementary Methods Table 1) were derived from W303-1a (*MATa ade2-1 can1-100 his3-11,15 leu2-3,112 trp1-1 ura3-1*).

The YB51 (*orc4*Δ*::TRP1* + p*ORC4/URA3*) strain was used for plasmid shuffle assay (see Methods, Plasmid shuffle assay). A PCR-based gene deletion strategy was used for disrupting endogenous Orc4 with *TRP1* and a *URA3*-containing plasmid (pRS416) carrying wildtype *ORC4* gene is used as complement.

The Orc4 site-directed mutation constructs-containing plasmids were used for plasmid shuffle assay (see Methods, Plasmid shuffle assay). Based on a CEN-based *LEU2*-containing plasmid constructs (pRS415) carrying wildtype *ORC4* gene, Orc4 site-directed mutation constructs were created using PCR mutagenesis strategy, confirmed by DNA sequencing (see Methods, Plasmid shuffle assay).

The NTAP-Orc4 integrated yeast strains were used for phenotype characterization assays, including the genome-wide DNA replication origin profile analyses, chromatin-immunoprecipitation and massively parallel origin mutagenesis and selection assay. NTAP-Orc4 integrated strains were derived from YB1588 (*MATa orc4*Δ*::TRP1 bar1*Δ*::TRP1 LEU2::BrdU-Inc* + p*ORC4/URA3*), which is a meiotic product of a diploid strain obtained by crossing YB51 (*MATα orc4*Δ*::TRP1* + p*ORC4/URA3*) and YB1549 (*MATa bar1Δ::TRP1 LEU2::BrdU-Inc*). YB1549 was derived from YS2251 (*MATa bar1Δ::TRP1*) by inserting a BrdU-INC cassette[38] with *LEU2* to facilitate EdU incorporation.

NTAP-Orc4 construct was used for NTAP-Orc4 integrated strains construction, which was generated using a PCR based strategy with tag coming from pBS1761[39] (purchased from Euroscarf). The construct is inserted into *his3* locus of YB1588 using CRISPR/Cas9 system[40]. Then, the plasmid containing Cas9 gene was dropped off by non-selective culture and tested for loss of plasmid marker. Subsequently, the plasmid carrying wildtype Orc4 gene was dropped off by counter selecting on 5 fluoroorotic acid 5-FOA plates for loss of *URA3*. The loss of p*ORC4*/*URA3* were confirmed by PCR and sequencing in combination with phenotypic assessment.

Plasmid shuffle assay

The Orc4 α-helix mutants were screened for function in vivo using plasmid shuffle assay similar to previously described[41]. The Orc4 site-directed mutation constructs-containing plasmids with LEU2 marker were transformed into YB51 (*orc4*Δ*::TRP1* + p*ORC4/URA3*) and selected on SC-Leu-Ura plates. The transformants were isolated, grown in YPD overnight, and spotted onto 5-FOA plates with 10-fold serial dilutions starting from $1.5x10^7$ cells to select for loss of URA3 plasmid carrying the wild-type Orc4. As control, the same dilutions were spotted on YPD plates. YPD or 5-FOA plates were cultured under 30˚C, or 25˚C or 37˚C to test their cold or temperature sensitivity.

Cell extract preparation, immunoprecipitation, immunoblot analysis and antibodies

Whole cell extraction from NTAP-Orc4 integrated strains (see Methods, Yeast genetic methods and strain construction) was prepared as previously described[41]. Cell extracts were analyzed for protein concentrations. Immunoprecipitation procedures were performed by mixing ~1.6 mg of total proteins and 30 µl of the IgG Sepharose 6 Fast Flow beads (GE Healthcare, Cat# 17-0969-01) at 4°C for 2h and precipitating the NTAP tagged Orc4. The beads were washed extensively with EBX buffer (recipe same

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.04.187286; this version posted July 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

as previously published[41]) and boiled in 30ul loading sample buffer (cite cold spring harbor recipe online). Proteins from immunoprecipitation (IP) and cell extract (as IP input) were fractionated by SDS-10% PAGE and transferred to nitrocellulose membrane. Immunoblot analysis was performed using antibodies against Orc4 (SB12) used at 1:2000 dilution and Orc1 (SB13) used at 1:1000 dilution and TBS with 0.05% Tween 20 was used for preparing blocking and washing solutions.

Cell growth, block, synchronization and flow cytometry analysis

Exponentially growing yeast cells (~10^7 cells/mL) in YPD were synchronized in G1 with 25 ng/mL of α-factor (*bar1*Δ strains are used in this study) for 3h at 30°C. To release from G1 arrest, cells were collected by filtration and promptly washed twice on the filter using one culture volume of H_2O and then resuspended into YPD medium. 1ml of cells was collected at different time points by adding sodium azide to final concentration at 0.1%. Cells are quickly centrifuged, resuspended with 400ul H_2O and fixed by adding 1ml 100% ethanol and rotate overnight at 4˚C. Cells then are quickly centrifuged, washed one time with H_2O, resuspended in 250ul RNaseA (Sigma-Aldrich) solution (2mg/ml), incubated for 4h in a 37˚C shaker and then sonicated using a Tekmar Sonic Disruptor with 630-0418 Tapered Microtip for 2 cycles of pulse for 1 second "ON", 1 second "OFF" at amplitude setting 22-25%. Proteinase K (Sigma-Aldrich) solution was added to final concentration at 1mg/ml and incubate for 1h in 50˚C in Eppendorf Thermomixer R Mixer, 1.5ml Block with speed at 750rpm. Cells were then quickly centrifuged, resuspend in 50mM Tris PH7.5 with SYBR green I (Thermo Fisher) diluted at 1:10,000 ratio and filtered through strainer cap tubes (Corning™ Falcon™ Test Tube with Cell Strainer Snap Cap). BD LSRFortessa Dual Special Order System instrument and BD FACSDiva Software Version 8.0.1 Firmware Version 1.4 (BD LSRFortessa) were used to collect the data by measuring SYBR green signal. Same number of yeast cells data (30,000 events per run) were collected for each sample. FlowJo Version 10.6.1 was used to analyze the data and no gating strategy used.

Massively parallel origin selection (MPOS) assay

Both the *ARS1* (*ARS416*) and HMR-E (*ARS317*) libraries used ARS sequences 150bp in length and synthesized with a 15% mutation rate at each position. Variant ARSs were cloned in bulk into a *HIS3*-containing plasmid. The libraries were then transformed into NTAP-Orc4 integrated yeast strains (see Methods, Yeast genetic methods and strain construction). The transformed cells were plated on SC-his plate, grew in 30˚C incubator, washed off from plates when saturated, inoculated into SC-his medium and shook at 30˚C shaker till it reached saturation to harvest. ARS-containing plasmid DNA was isolated and PCR-amplified and ligated with custom inline barcodes (Supplementary Method Table 3), quantified, pooled, and submitted for sequencing. Computational analyses of MPOS assay data are described below (see Methods, Computational analyses of MPOS data). DNA sequencing data were submitted to the Sequence Read Archive database (see Methods, Data and code availability).

Computational analyses of MPOS data

Processing of MPOS data. Illumina reads from the MPOS experiments were analyze using custom Python scripts. The output of this pipeline was, for each library or selected sample, a list of variant ARS sequences with each sequence assigned a corresponding read count. These lists were used as input to both the ER and IM motif modeling algorithms described below

Matrix models for ARS motifs. Our motif modeling effort aimed to predict the activity of a variant ARS based on its DNA sequence. Specifically, we sought a mathematical function $a(s)$ that quantifies the

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.04.187286; this version posted July 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

activity of an arbitrary input ARS DNA sequence s. We assumed this function could be represented by a matrix model[42], i.e.,

$$a(s) = \sum_b \sum_l \theta_{bl} s_{bl}$$

where $b = A, C, G, T$ indexes the four DNA bases, $l = 1, 2, \ldots, L$ indexes nucleotide positions, the sequence $s = \{s_{bl}\}$ is represented by a $4 \times L$ matrix of indicator variables ($s_{bl} = 1$ if base b occurs at position l; $s_{bl} = 0$ otherwise), and $\theta = \{\theta_{bl}\}$ is a $4 \times L$ matrix of model parameters that must be inferred from data. All inferred motifs were limited to sequences of length $L = 50$ encompasing both the A and B1 elements of the assayed ARSs. To facilitate the comparison of motifs to one another, both visually and through PCA analysis, motif parameters were centered and rescaled via the transformation $\theta_{bl} \rightarrow \theta'_{bl}/C$ where $\theta'_{bl} = \theta_{bl} - \frac{1}{4}\sum_{b'} \theta_{b'l}$ and $C = \sqrt{\sum_b \sum_l \theta'^2_{bl}}$.

Sequence logos. Sequence logos were generated by Logomaker [43]. In these representations of motif parameters, the value of θ_{bl} is represented by the height of character b at position l (or negative that height if the character is drawn below the x-axis).

Principal component analysis (PCA). The PCAs shown in Extended Data Figure S10 were performed on inferred motifs as follows. The motif parameters θ were first centered and normalized as described above. Each parameter matrix was then unrolled into a $4L \times 1$ vector, where $L = 50$. Standard PCA analysis was then performed on different motif subsets, as shown in panel a.

Motif inference. The inference of motif parameters was performed using a second-generation version of the MPAthic software package[44]. MPAthic enables motif inference using either enrichment ratios (ER) or information maximization (IM).

ER inference is the standard way of computing sequence motifs from massively parallel selection experiments[45]. Here, parameter values θ^{ER} are given by

$$\theta_{bl}^{\mathrm{ER}} = \log_2 \frac{f_{bl}^{\mathrm{selected}}}{f_{bl}^{\mathrm{library}}},$$

where $f_{bl}^{\mathrm{library}}$ is the fraction of sequences in the initial ARS library that have base b at position l, and $f_{bl}^{\mathrm{selected}}$ is defined similarly for selected ARS sequences. These fractions were computed using a pseudocount of 1.

IM inference seeks to identify parameters θ^{IM} that maximize the mutual information between the predicted activity of an assayed ARS sequence and the sample that sequence was observed in. Specifically, one aims to maximize

$$I(\theta) = \sum_{\mathrm{samples}} p(\mathrm{sample}) \int da\, p(a|\mathrm{sample}) \log_2 \frac{p(a|\mathrm{sample})}{p(a)}$$

where "sample" indicates either the library sample or the selected sample, and $p(a|\mathrm{sample})$ is the distribution of activities assigned to the sequences in that sample by a motif with parameters θ. For a given

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.04.187286; this version posted July 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

choice of θ, the distribution $p(a|\text{sample})$ was computed as in Kinney et al. [46]: the activities a in both samples combined were sorted, replaced by their ranks, and binned into 1000 equipopulated bins; for each sample, the distribution of sequence counts across bins was then smoothed using Gaussian kernel having a standard deviation of 20 bins. The marginal probability was subsequently computed as $p(a) = \sum_{\text{samples}} p(\text{sample})p(a|\text{sample})$. The optimal values θ^{IM} were identified using a Metropolis Monte Carlo simulation in which each θ was assigned relative probability of $2^{NI(\theta)}$, where N is the number of read counts in both the library sample and selected samples combined. Each Monte Carlo run was initiated at random parameter values then carried out for 25,000 steps. Each reported motif resulted from averaging together the end-points of five independent Monte Carlo runs. This IM inference strategy closely followed the one described previously[46]. The present work is the first to show that, as predicted from previous theoretical arguments[47], IM motif inference removes systematic experiment-to-experiment variation that confounds ER motif inference.

Genome-wide replication origin profile analysis

Isolation and preparation of DNA for genome-wide replication origin profile analysis is similar to previously described[48]. Yeast cells were synchronized in G1 with α-factor and were released into medium containing 0.2 mg/mL pronase E, 0.2 M HU, and 0.5 mM EdU. Cells were collected by centrifugation at 90 mins after release into S phase. Genomic DNA was extracted and fragmented. EdU-genomic DNA was then biotinylated using the Click reaction and purified using Streptavidin T1 magnetic beads (Invitrogen). Libraries for Illumina sequencing were constructed using TruSeq ChIP Library Preparation Kit (Illumina). Computational analyses of sequencing data are described below (see Methods, Computational analyses of replication origin profile and ChIP-seq data). DNA sequencing data were submitted to the Sequence Read Archive database (see Methods, Data and code availability).

Chromatin immunoprecipitation

The ChIP-seq for Orc1 and Mcm2 were performed as described[49] with modification. About 10^9 synchronized yeast cells were fixed with 1% formaldehyde for 15 min at room temperature (RT), then quenched with 130 mM glycine for 5 min at RT, harvested by centrifugation, washed twice with TBS (50 mM Tris.HCl pH 7.6, 150 mM NaCl), and flash frozen. Cell pellets were resuspended in 600 µl lysis buffer (50 mM HEPES-KOH pH 7.5, 150 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% Na-Deoxycholate, 0.1% SDS, 1 mM PMSF, protease inhibitor tablet (Roche)), and disrupted by bead beating using multi-tube vortex (Baxter Scientific Products SP Multi-Tube Vortexer S8215-1) for 12-15 × 30s at maximum setting. Cell extracts were collected and sonicated using Bioruptor (UCD-200, Diagenode) for 38 cycles of pulse for 30 seconds "ON", 30 seconds "OFF" at amplitude setting High (H). The extract was centrifuged for 5 min at 14,000 rpm. The soluble chromatin was used for IP. Antibody against Mcm2 (mcm228) was preincubated with washed Dynabeads Protein A/G. For each immunoprecipitation, 80 µl antibody-coupled beads was added to soluble chromatin. Samples were incubated overnight at 4°C with rotation, after which the beads were collected on magnetic stands, and washed 3 times with 1 ml lysis buffer and once with 1 ml TE, and eluted with 250 µl preheated buffer (50 mM Tris.HCl pH 8.0, 10 mM EDTA, 1% SDS) at 65°C for 15 min. Immunoprecipitated samples were incubated overnight at 65°C to reverse crosslink, and treated with 50 µg RNase A at 37°C for 1 hr. 5µl proteinase K (Roche) was added and incubation was continued at 55°C for 1 hr. Samples were purified using MinElute PCR purification kit (Qiagen). Libraries for Illumina sequencing were constructed using TruSeq ChIP Library Preparation Kit (Illumina). Computational analyses of sequencing data are described below (see Methods, Computational analyses of replication origin profile and ChIP-seq data).

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.04.187286; this version posted July 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Computational analyses of replication origin profile and ChIP-seq data

Illumina reads from the genome-wide replication origin profiling and ChIP-seq experiments were mapped to the *S. cerevisiae* S288C genome using BWA, after which pileup files were created using SAMtools (http://www.htslib.org). Pileup counts were then smoothed via convolution with a uniform kernel of width 5000 bp (for replication origin profiles) or 300 bp (for ChIP-seq). To normalize the profiles relative to one another, we computed the number of reads bounding 99.5% of positions within each profile and divided the entire profile by this number.

Data availability

Raw Illumina reads are available at [SRA ACCESSION NUMBER].

Code availability

Processed data files, analysis scripts, and scripts used for figure generation are available at [GITHUB URL].

Note for reviewers: We will provide all code used for data analysis in Github as open source software upon acceptance. It will be freely available on Github. SRA accession number will be publicly available upon revision of the paper. In the meantime, all raw data, processed data and custom codes have currently been uploaded onto an unstructured repository:
MPOS assay data link: labshare.cshl.edu/shares/stillmanlab/www-data/yixinhu/MPOS/
Origin firing EdU profile data link: labshare.cshl.edu/shares/stillmanlab/www-data/yixinhu/EdU/
ChIP-seq data link: labshare.cshl.edu/shares/stillmanlab/www-data/yixinhu/ChIP-seq/
Custom codes link: labshare.cshl.edu/shares/stillmanlab/www-data/yixinhu/Code/

Method References / Additional References

38. Viggiani, C. J. & Aparicio, O. M. New vectors for simplified construction of BrdU-Incorporating strains of *Saccharomyces cerevisiae*. *Yeast* **23,** 1045–1051 (2006).
39. Puig, O. *et al.* The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods* **24,** 218–229 (2001).
40. Anand, R., Memisoglu, G. & Haber, J. Cas9-mediated gene editing in Saccharomyces cerevisiae. (2017). doi:10.1038/protex.2017.021a
41. Sheu, Y.-J. & Stillman, B. The Dbf4-Cdc7 kinase promotes S phase by alleviating an inhibitory activity in Mcm4. *Nature* **463,** 113–117 (2010).
42. Kinney, J. B. & McCandlish, D. M. Massively Parallel Assays and Quantitative Sequence-Function Relationships. *Annu Rev Genomics Hum Genet* **20,** 99–127 (2019).
43. Tareen, A. & Kinney, J. B. Logomaker: beautiful sequence logos in Python. *Bioinformatics* **36,** 2272–2274 (2020).
44. Ireland, W. T. & Kinney, J. B. MPAthic: Quantitative Modeling of Sequence-Function Relationships for massively parallel assays. *bioRxiv* **17,** doi:10.1101–054676 (2016).
45. Stormo, G. D. Modeling the specificity of protein-DNA interactions. *Quant Biol* **1,** 115–130 (2013).
46. Kinney, J. B., Murugan, A., Callan, C. G. & Cox, E. C. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proceedings of the National Academy of Sciences* **107,** 9158–9163 (2010).

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.04.187286; this version posted July 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

47. Atwal, G. S. & Kinney, J. K. Learning quantitative sequence–function relationships from massively parallel experiments. *J. Stat. Phys.* **162,** 1203–1243 (2016).
48. Sheu, Y. J., Kinney, J. B., Lengronne, A., Pasero, P. & Stillman, B. Domain within the helicase subunit Mcm4 integrates multiple kinase signals to control DNA replication initiation and fork progression. *Proceedings of the National Academy of Sciences* **111,** E1899–E1908 (2014).
49. Behrouzi, R. *et al.* Heterochromatin assembly by interrupted Sir3 bridges across neighboring nucleosomes. *eLife* **5,** 209 (2016).
50. Sheu, Y.-J., Kinney, J. B. & Stillman, B. Concerted activities of Mcm4, Sld3, and Dbf4 in control of origin activation and DNA replication fork progression. *Genome Res* **26,** 315–330 (2016).

Acknowledgements

We thank Jennifer Shapp and Kevin Chen for help with experiments. This work was supported by NIH grants R01GM45436 and P01CA13106 to B.S. and R35GM133777 to J.B.K. L. J-T. is an Investigator of the Howard Hughes Medical Institute. The Cold Spring Harbor Laboratory NextGen Sequencing Cancer Center Shared Resource is supported by grant P30 CA045508.

Author Contributions

B.S., J.B.K., Y.H., Y-J. S., L. J-T, C.S., and H.L. designed the experiments, A.T, W.T.I. and J.B.K. developed the mutual information maximization algorithm. Y.H., J.B.K. and Y-J. S. performed the experiments. Y.H., J.B.K. and B.S. wrote the paper. All authors analyzed aspects of the data.

Competing Interest Declaration

There are no competing interests.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.04.187286; this version posted July 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

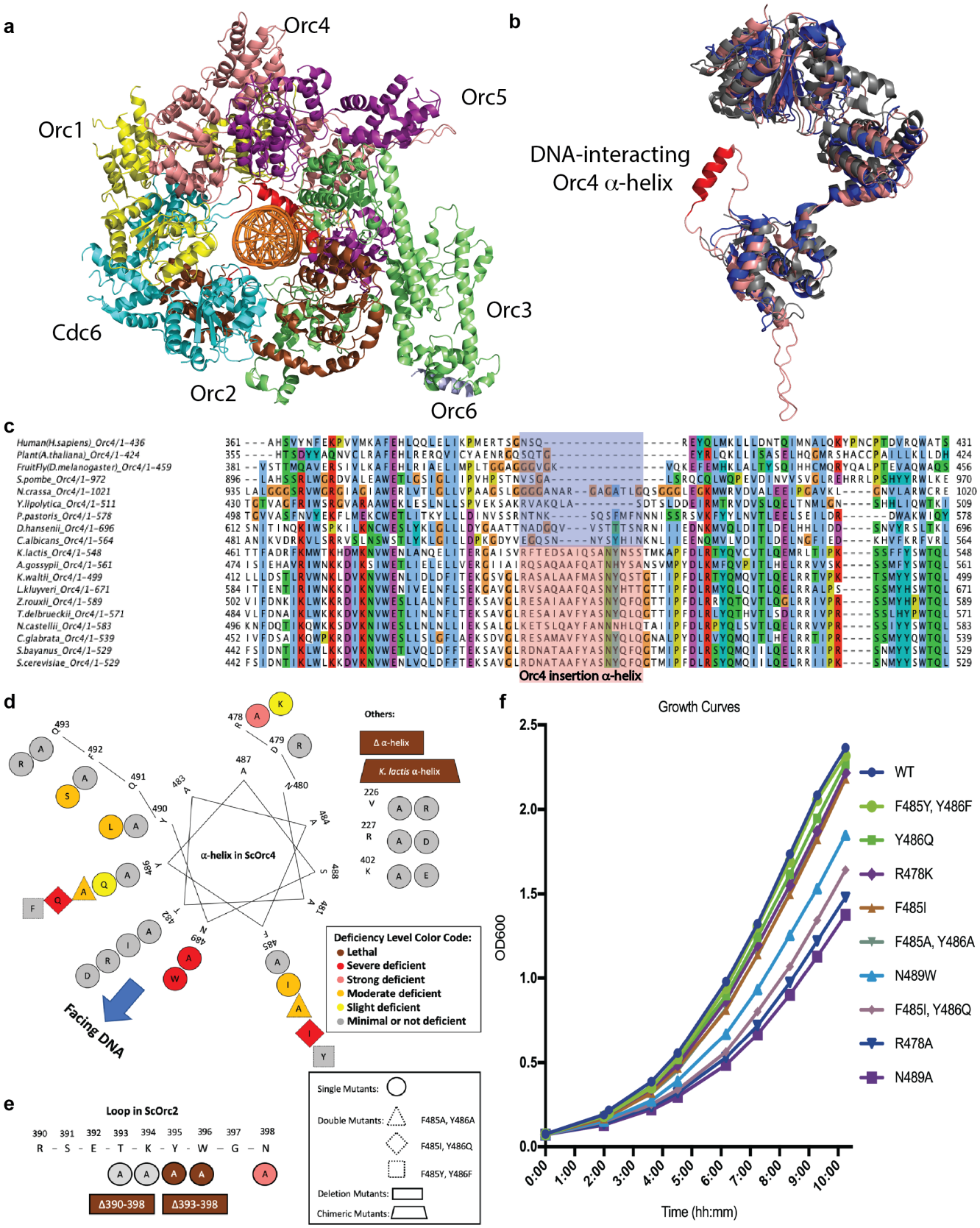


Fig. 1

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.04.187286; this version posted July 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

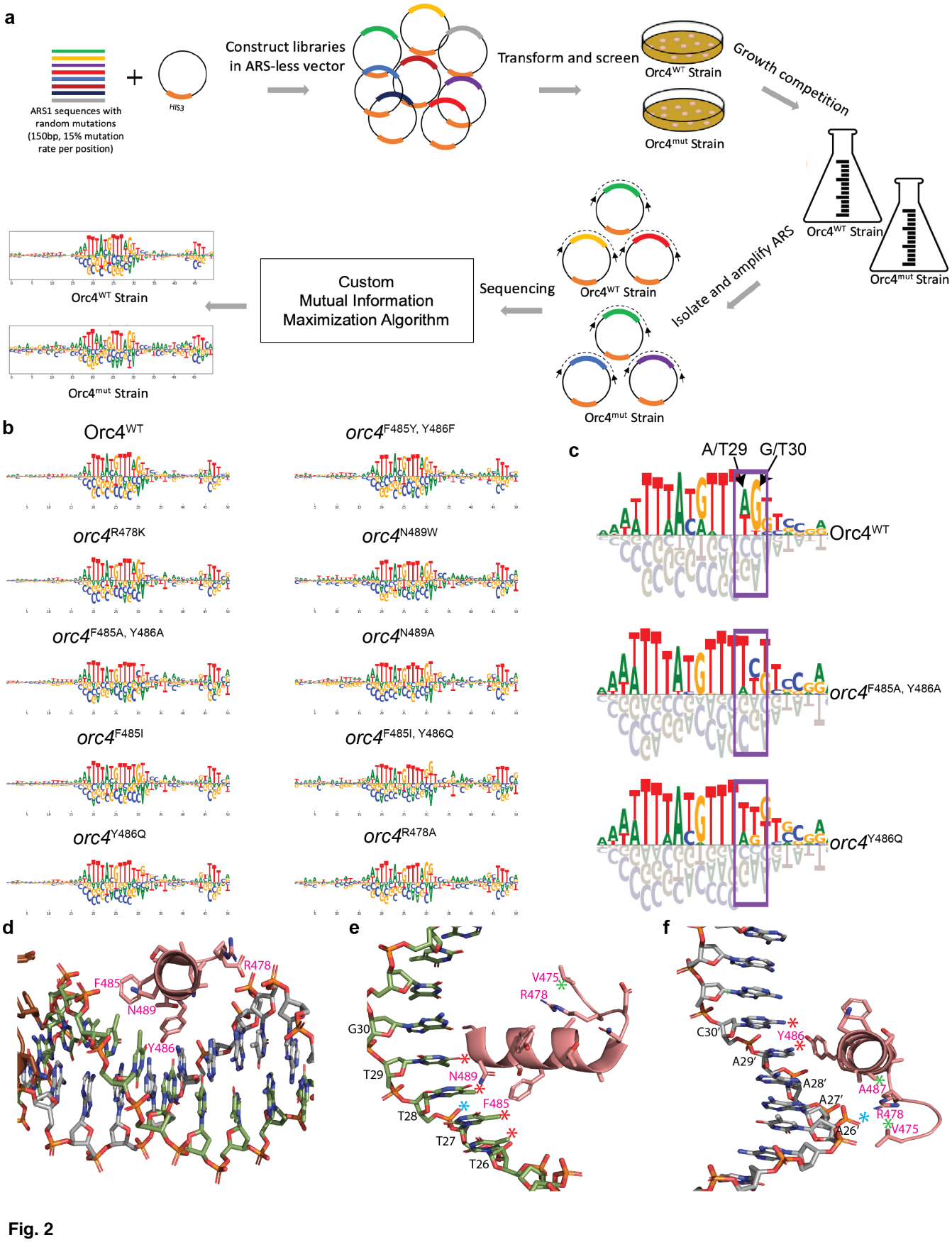


Fig. 2

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.04.187286; this version posted July 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

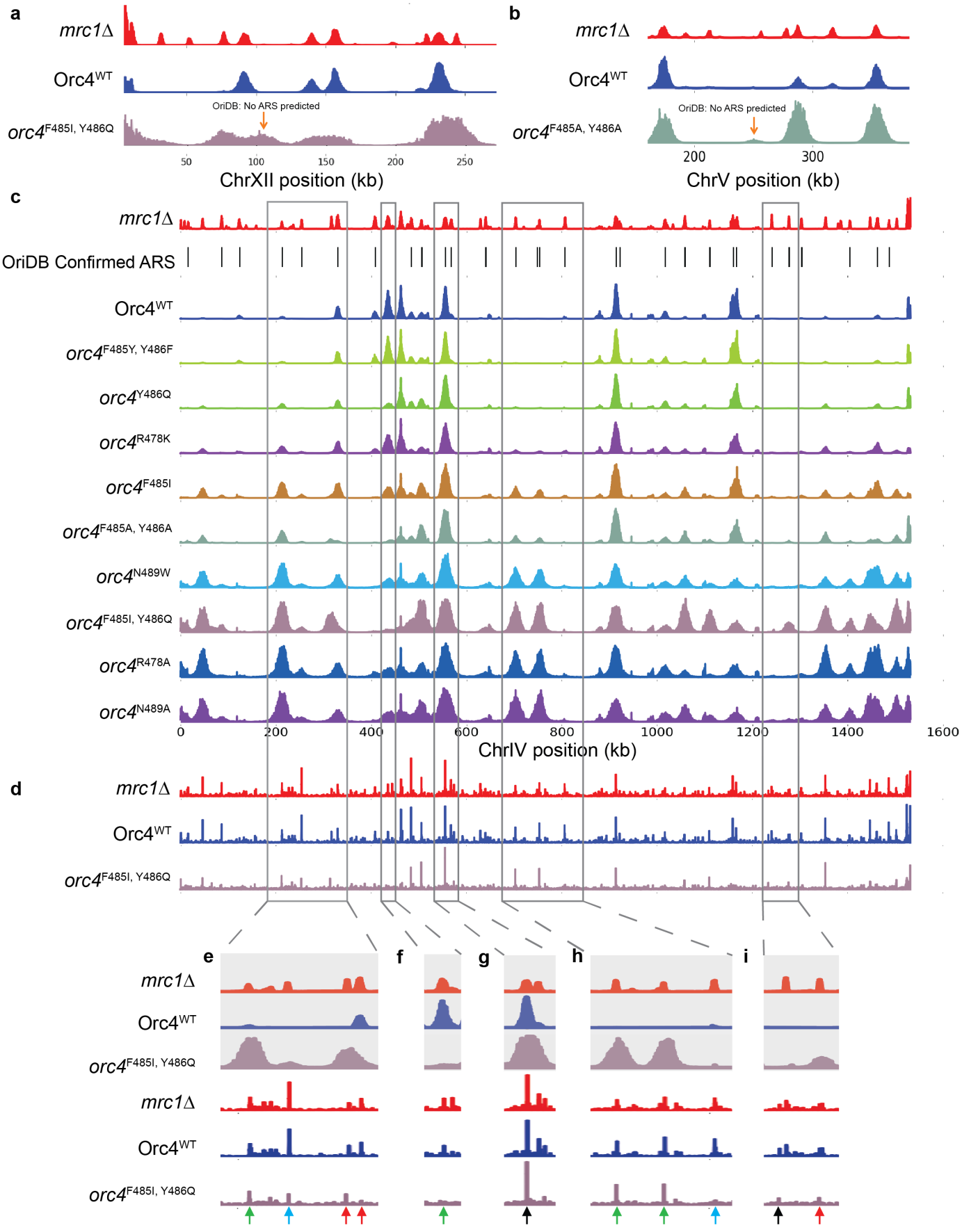


Fig. 3

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.04.187286; this version posted July 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

a

Orc4 Strains	P-value	Symbols
WT	0.73385	
F485Y, Y486F	0.706186	
F485I	0.718865	
Y486Q	0.0348792	*
F485I, Y486Q	2.84918e-05	***
F485A, Y486A	5.75003e-06	***
R478K	0.397443	
R478A	0.804511	
N489A	0.430692	
N489W	0.579688	

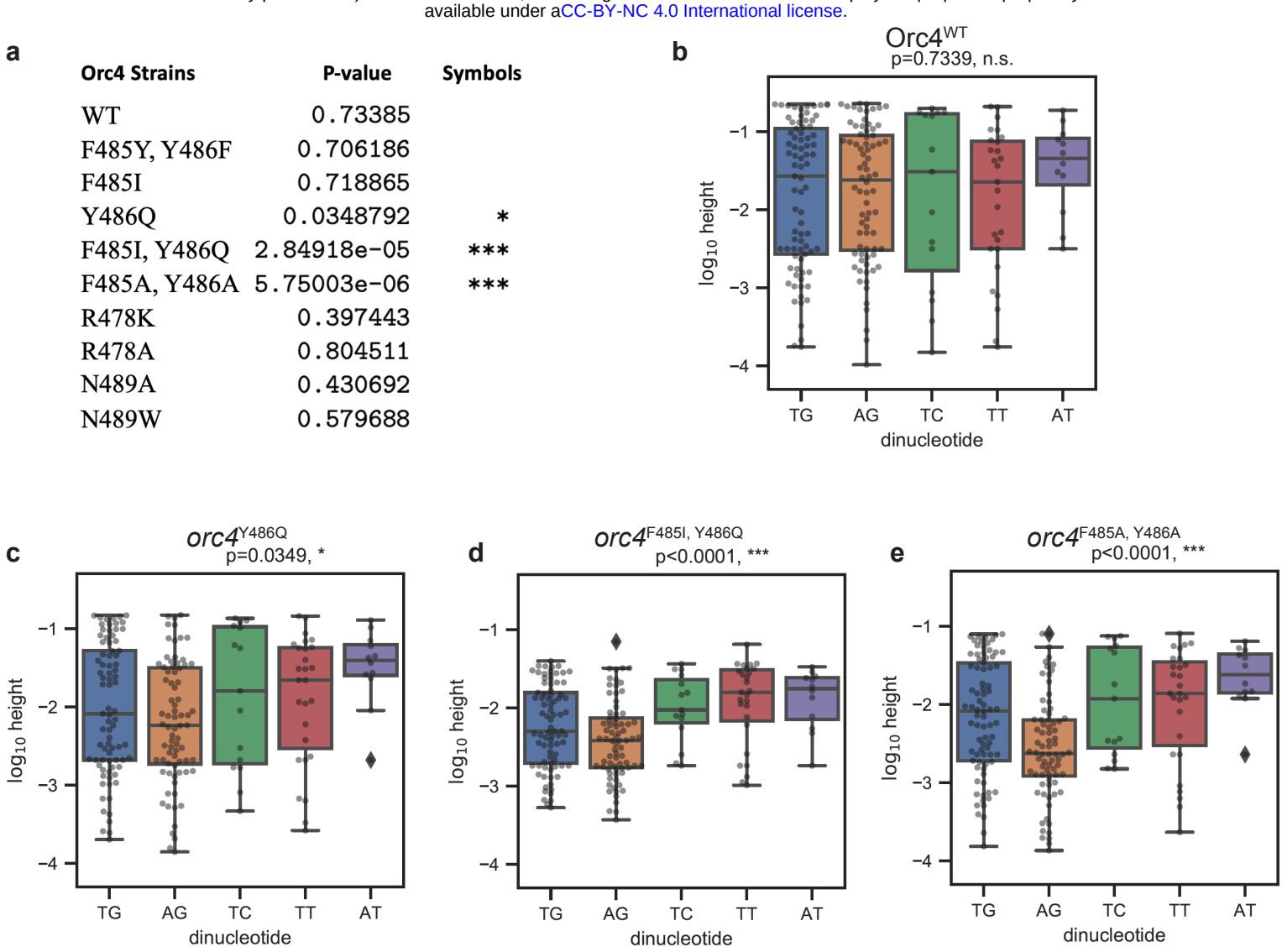
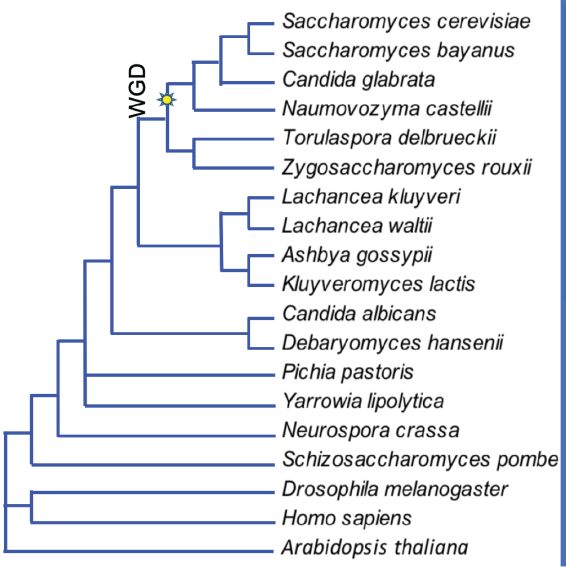


Figure 4

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.04.187286; this version posted July 4, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



Tree in unscaled branches

Orc4 a-helix	Orc2 loop	Defined Origin	Orc1	Sir3	Sir2	Hst1	Sir1	Sir4	Esc1	Dcr	Ago	RdRP	HP1
+	+	+	+	+	+	+	+	+	+	-	-	-	-
+	+	+	+	+	+	+	+	+	+	*	-	-	-
+	+	+	+	+	+	+	-	+	+	-	-	-	-
+	+	+	+	+	+	+	+	+	+	*	+	-	-
+	+	+	+	-	+	-	-	+	+	*	+	+	-
+	+	+	+	-	+	-	-	+	+	-	-	-	-
+	+	+	+	-	+	-	-	+	+	-	-	-	-
+	+	+	+	-	+	-	-	+	+	-	-	-	-
+	+	+	+	-	+	-	-	+	+	-	-	-	-
+	+	+	+	-	+	-	-	+	+	-	-	-	-
-	-	-	+	-	+	-	-	-	-	*	+	-	-
-	-	-	+	-	+	-	-	-	-	*	-	-	-
-	-	-	+	-	+	-	-	-	-	*	-	-	-
-	-	-	+	-	+	-	-	-	-	-	-	-	-
-	-	-	+	-	+	-	-	-	-	+	+	+	+
-	-	-	+	-	+	-	-	-	-	+	+	+	+
-	-	-	+	-	+	-	-	-	-	+	+	+	+
-	-	-	+	-	+	-	-	-	-	+	+	+	+
-	-	-	+	-	+	-	-	-	-	+	+	+	+

* Non-canonical Dicer that contains an RNase III domain (like RNT1)

Fig. 5