

34. S. Miyamoto, P. A. Kollman, *J. Comp. Chem.* **13**, 952 (1992).
35. T. Darden, D. York, L. Pedersen, *J. Chem. Phys.* **98**, 10089 (1993).
36. F. Zhu, E. Tajkhorshid, K. Schulten, *FEBS Lett.* **504**, 212 (2001).
37. O. S. Smart, J. G. Neduvellil, X. Wang, B. A. Wallace, M. S. P. Sansom, *J. Mol. Graphics* **14**, 354 (1996).
38. E. Espinosa, E. Molins, C. Lecomte, *Chem. Phys. Lett.* **285**, 170 (1998).
39. R. M. Esnouf, *J. Mol. Graphics Model.* **15**, 132 (1997).
40. E. A. Merritt, D. J. Bacon, *Methods Enzymol.* **277**, 505 (1997).
41. I. Klapper, R. Hagstrom, R. Fine, K. Sharp, B. Honig, *Proteins Struct. Funct. Genet.* **1**, 47 (1986).
42. Free-energy profiles $G_{PMF}(z) = -k_B T \ln n(z)$ (PMF,

potential of mean force) have been obtained from the residence frequency $n(z)$ along the pore axis z . Rotational water entropies [plotted as entropic contribution TS , where $T = 300$ K, to the free-energy profile (kJ/mol)] were calculated from the Eulerian angle distribution with respect to the pore, $\rho(\theta) = \rho(\theta_1, \theta_2, \theta_3)$, obtained from the trajectories of all water molecules passing through the pore region, $S = k_B \int d^3 \theta \rho(\theta) \ln \rho(\theta) \approx k_B (\sum_i n_i \ln n_i) / \sum_i n_i$, using suitable equally sized Eulerian angle volume elements, i , and the number, n_i , of configurations within these volume elements. Correlated motions of adjacent hydrogen-bonded water molecules along the pore axis were quantified with a correlation coefficient $c = \langle \Delta z_1 \Delta z_2 \rangle / (\langle \Delta z_1 \Delta z_1 \rangle \langle \Delta z_2 \Delta z_2 \rangle)^{1/2}$, where $\Delta z_i = z_i(t +$

$\Delta t) - z_i(t)$ is the displacement of the two water molecules $i = 1, 2$ in the pore axis direction during a $\Delta t = 10$ ps interval. Pore dimensions were evaluated with the HOLE program (37). All profiles were averaged over multiple MD snapshots.

43. We thank P. Agre, A. Engel, and R. Stroud for stimulating discussions; P. Deen, A. Engel, Y. Fujiyoshi, T. Heimburg, R. Böckmann, and G. Vriend for carefully reading the manuscript; and P. Tieleman for providing force-field parameters for POPE and an equilibrated POPE conformation. B.L.d.G. was supported by the BIOTECH program of the European Union (grants BIO4-CT98-0024 and QLRT 2000/00778).

9 September 2001; accepted 1 November 2001

Genome-Wide Distribution of ORC and MCM Proteins in *S. cerevisiae*: High-Resolution Mapping of Replication Origins

John J. Wyrick,^{1,2*} Jennifer G. Aparicio,^{3*} Ting Chen,³
John D. Barnett,² Ezra G. Jennings,^{1,2} Richard A. Young,^{1,2}
Stephen P. Bell,^{1,4} Oscar M. Aparicio^{3†}

DNA replication origins are fundamental to chromosome organization and duplication, but understanding of these elements is limited because only a small fraction of these sites have been identified in eukaryotic genomes. Origin Recognition Complex (ORC) and minichromosome maintenance (MCM) proteins form prereplicative complexes at origins of replication. Using these proteins as molecular landmarks for origins, we identified ORC- and MCM-bound sites throughout the yeast genome. Four hundred twenty-nine sites in the yeast genome were predicted to contain replication origins, and ~80% of the loci identified on chromosome X demonstrated origin function. A substantial fraction of the predicted origins are associated with repetitive DNA sequences, including subtelomeric elements (X and Y') and transposable element-associated sequences (long terminal repeats). These findings identify the global set of yeast replication origins and open avenues of investigation into the role(s) ORC and MCM proteins play in chromosomal architecture and dynamics.

In eukaryotic cells, chromosome replication initiates from DNA loci called origins of replication that are distributed along chromosomes. In *Saccharomyces cerevisiae*, a DNA sequence that functions as an origin is termed an autonomously replicating sequence (ARS) (1–3). An 11–base pair (bp) ARS consensus sequence (ACS) is essential for replication initiation and is recognized by the eukaryotic replication initiator, the Origin Recognition Complex (ORC)

(4–7). Additional sequences, including an A/T-rich region, are required for origin function (7, 8). Although the presence of a match to the ACS is required for ARS activity, it is not sufficient, and the majority of matches to the ACS in the genome do not have ARS activity (9). Furthermore, ARS activity varies depending on chromosomal position, suggesting that local chromatin structure influences ARS function (1). As a result of these properties, the chromosome-wide identification of origins of replication has been a labor-intensive task (10). To date, the ARSs on only 2 of the 16 chromosomes (III and VI), representing about 5% of the genome, have been mapped at high resolution and their chromosomal activities characterized (11–15).

Initiation of DNA replication is regulated through the ORC-dependent recruitment of minichromosome maintenance (MCM) proteins to the ORC-origin complexes during G₁ phase of the cell cycle (16). As the binding of ORC

and MCM proteins occurs at or very near the origin, we determined the genome-wide locations of ORC- and MCM-binding sites to identify the positions of potential DNA replication origins across the *S. cerevisiae* genome. Chromatin immunoprecipitation (ChIP) was used to enrich for protein-bound DNA that was labeled and hybridized to DNA microarrays in triplicate (17). In this study, we used DNA microarrays that included probes to most yeast intergenic sequences and open reading frames (ORFs) for a total of 12,158 loci (18). Analysis of the hybridization data generated an average binding ratio (fluorescent intensity of enriched versus unenriched DNA) along with a confidence interval for binding (P value) of the protein of interest with each DNA sequence present on the arrays (17). We reasoned that loci exhibiting association with both ORC and MCM proteins would represent sites of bona fide origins of replication. We performed five independent experiments to identify the binding sites of Orc1p, Mcm3p, Mcm4p, and Mcm7p and the entire ORC complex (using a polyclonal antibody recognizing all six ORC subunits, α -Orc1-6p) (18).

We identified 707 binding sites for Mcm3p, 719 for Mcm4p, 671 for Mcm7p, 568 for Orc1p, and 531 for Orc1-6p (each at $P \leq 0.025$) (18). Extensive overlap occurred among these binding sites, particularly among those associated with the MCM proteins. For example, 443 sites bound all three of the MCM proteins, and 206 additional sites bound two of the three MCM proteins (these 649 MCM-binding sites represent 477 nonadjacent sites). Only a few binding sites for each MCM protein (42, 73, and 46 for Mcm3p, Mcm4p, and Mcm7p, respectively) resided at isolated loci (defined as loci where neither the identified locus nor immediately adjacent loci exhibited binding to any of the other proteins tested). Seventy-five sites that bound all three MCM proteins ($P < 0.025$) failed to show ORC binding in either experiment ($P < 0.10$), and 12 sites that bound ORC in both experiments ($P < 0.025$) failed to bind any MCM proteins ($P < 0.10$). The majority of genomic sites tested (90%) did not show ORC or MCM association in any of the experiments.

To determine whether the identified loci

¹Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. ²Whitehead Institute for Biomedical Research, Cambridge, MA 02142, USA. ³Program in Molecular and Computational Biology, University of Southern California, Los Angeles, CA 90089–1340, USA. ⁴Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

*These authors contributed equally to this work.

†To whom correspondence should be addressed. E-mail: oapari@usc.edu

were associated with origins of replication, we compared the locations of ORC- and MCM-binding sites with the positions of previously identified ARSs (Fig. 1). We used these comparisons to develop a stringent set of criteria to maximize the number of accurately predicted sites while minimizing false positives. We discovered a high degree of correlation between the number of MCM and ORC proteins identified at particular sites (low *P* values) and the positions of known ARSs (Fig. 1). We established *P* value thresholds to identify potential origins in the genome based primarily on the colocalization of ORC and/or MCM protein binding sites and augmented by giving additional value to a locus if an immediately adjacent locus showed binding to an MCM or ORC protein (18). Detection of binding to loci adjacent to the actual site of an origin was anticipated and frequently observed because of the size distribution of DNA enriched by ChIP (19). Application of these criteria to our data resulted in the identification of 22 of 25 (88%) known ARSs on chromosomes III and VI (Fig. 1). Sixteen of the 22 ARSs were identified within the probe corresponding to the ARS, and six were identified by an adjacent probe (in four of these six cases, the exact probe corresponding to the ARS was not present on the arrays). Five additional loci on chromosomes III and VI were identified that do not contain an ARS. Because the average length of the probe DNA on the microarrays was ~600 bp, these results illustrate that our method can accurately identify the position of ARSs to a resolution of 1 kb or less (18).

Application of the above criteria for the identification of potential origins to the entire genome located 644 sites representing 429 non-adjacent loci termed proposed ARS (pro-ARS) (Fig. 2 and Web table 2) (18). We tested for ARS function of the 29 pro-ARSs on chromosome X and identified 23 previously unknown ARSs (Fig. 3) (18). Two of these ARSs were at sites that bound MCM but not ORC proteins. Six of the pro-ARS loci showed no ARS activity. Thus, 79% of the loci on chromosome X that we identified contained an ARS and false positives were easily eliminated (Fig. 3). That only three known origins on chromosomes III and VI were not identified suggests that we have identified the great majority of origins throughout the genome.

The pro-ARSs (and ARSs) typically lie in intergenic regions, suggesting that transcription and origin function are disfavored. All the newly identified ARSs on chromosome X lie in intergenic regions (Fig. 3). There is a size bias to intergenic regions harboring potential origins. Whereas 18, 27, and 38% of all the total intergenic regions were shorter than 200, 250, and 300 bp, respectively, only 8, 14, and 20% of the pro-ARS intergenic loci were of these respective sizes. In accord, the average length of all intergenic probes was 477 bp, whereas the average length of pro-ARS and known ARS-

containing intergenics was 649 and 715 bp, respectively. This characteristic suggests an additional space requirement for binding of ORC and MCM proteins and associated chromatin structure in these intergenic regions that also must accommodate the binding of proteins regulating expression of the adjacent genes (20).

We discovered that the potential origins are not randomly distributed, with clustering of up to five pro-ARSs occurring in telomere-adjacent regions (within 20 kbp of either end) (Fig. 2). Some of these pro-ARSs are likely to be associated with the subtelomeric X and Y' DNA elements (21). However, the large number of pro-ARSs identified in these regions was unexpected as most telomeres contain only one X and/or one Y' element. We confirmed the ARS

function of all 11 pro-ARS sequences near the telomeres of chromosomes VI and X, indicating that the observed distribution of pro-ARS sequences near telomeres was not due to cross hybridization with probes representing repeated elements in telomeric regions (18). Three of the four ARSs near the left telomere of chromosome VI were weak, and all three weak ARSs identified on chromosome X were near other ARSs (Web table 4) (18). The dominance of one origin's activity over a proximal adjacent origin may allow accumulation of mutations that reduce the potential activity of the inactive origin (22, 23). Moreover, some of these weak ARSs may represent ORC-binding sites that participate in the establishment of chromatin domains, perhaps re-

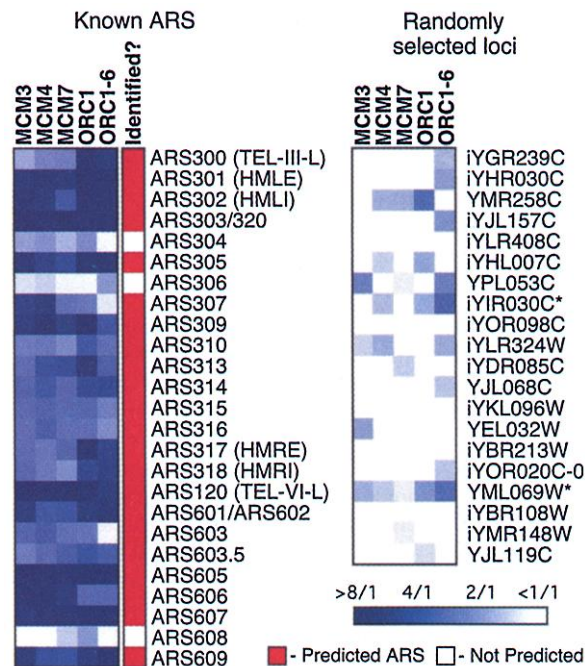


Fig. 1. ORC and MCM binding to previously identified replication origins. Average binding ratios (blue/white) of ORC and MCM proteins to the known ARS-containing loci on chromosomes III and VI (ARS308 and ARS604 were not present on the arrays) and some randomly selected loci are shown. Random selection was accomplished with the "randbetween" function in Excel. The "i" preceding the locus name indicates the intergenic region to the right of the gene. Asterisks indicate randomly selected loci adjacent to or within 1 kb of a predicted origin. Data for other known origins are available in Web table 1 (18).

Table 1. Potential replication origins frequently colocalize with sequences associated with transposable elements. Analysis is described in (18).

Feature set	Total intergenics	Intergenic blocks	Expected overlaps	Actual overlaps	<i>P</i> values
Intergenic probes	6286	6137			
Pro-ARS (intergenics)	324	324			
tRNA	277	254	14.7	49	1.7E-13
Ty1-5	50	47	2.7	11	8.9E-05
Ty1 (delta)	31	30	1.7	6	6.5E-03
Ty2 (delta)	13	13	0.75	3	3.3E-02
Ty3 (sigma)	2	2	0.12	0	8.9E-01
Ty4 (tau)	3	3	0.17	1	1.5E-01
Ty5 (omega)	1	1	0.058	1	5.4E-02
LTR (all)	333	210	12.1	47	3.0E-15
LTR-delta	254	180	10.4	35	4.8E-10
LTR-sigma	43	40	2.3	8	1.9E-03
LTR-tau	29	28	1.6	11	8.7E-07
LTR-omega	8	7	0.40	3	7.3E-03
Independent tRNA	96	93	5.4	12	5.3E-03
Independent LTR	40	34	2.0	7	3.0E-03

REPORTS

nucleating a structure at repeating intervals. ORC binds only a subset of the many dispersed ACSs throughout the genome, whereas it appears to bind with high frequency to repetitive DNA elements in tran-

scriptionally silent heterochromatin in yeast and *Drosophila* (24, 25).

The genomic scale and high resolution of our analysis allowed determination of the relationship between the locations of potential rep-

lication origins and specific chromosomal features. Only 28 pro-ARSs localized to intergenic regions upstream of ~800 cell cycle-regulated genes, arguing against a role for replication origins in the regulation of most of these genes

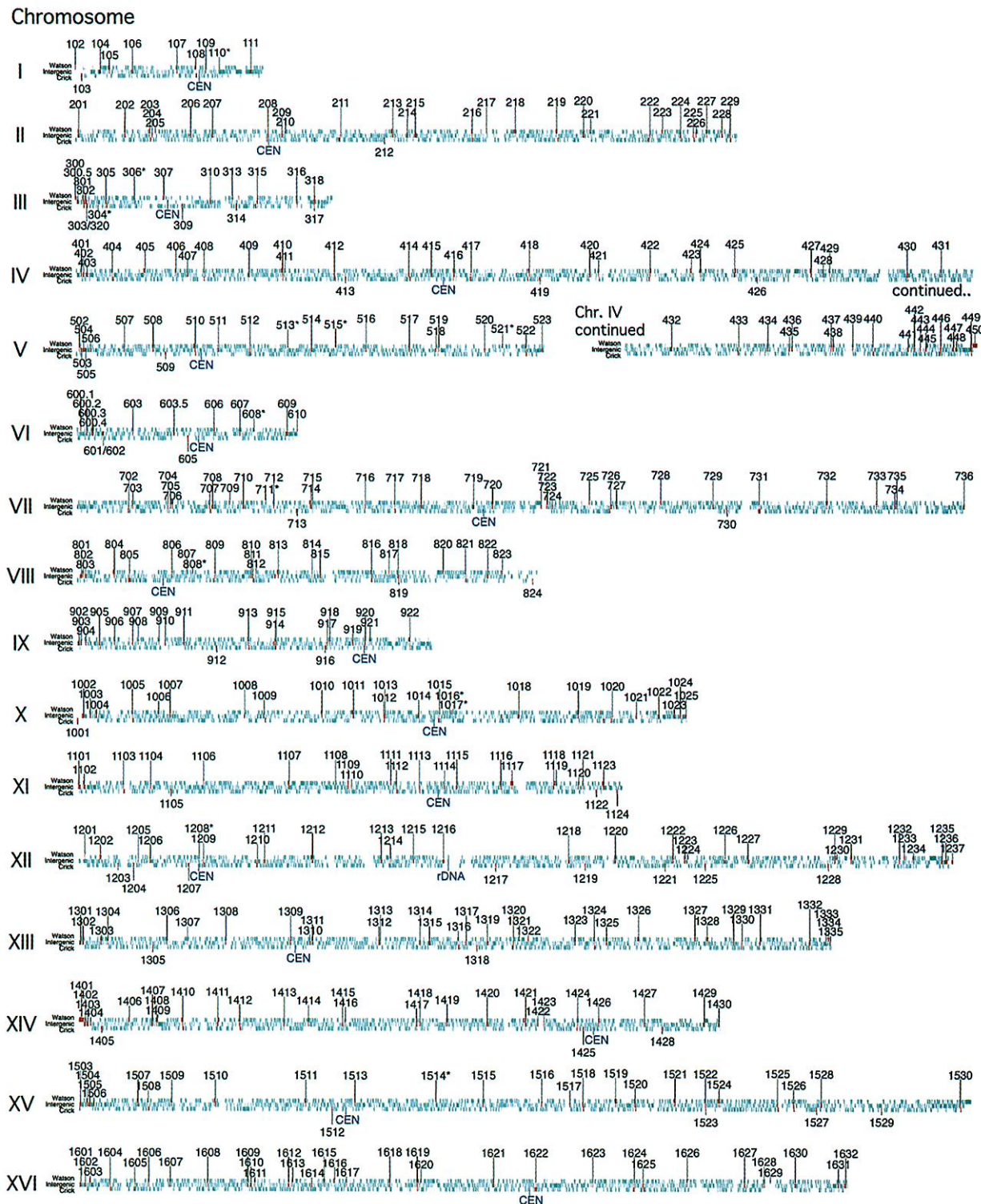
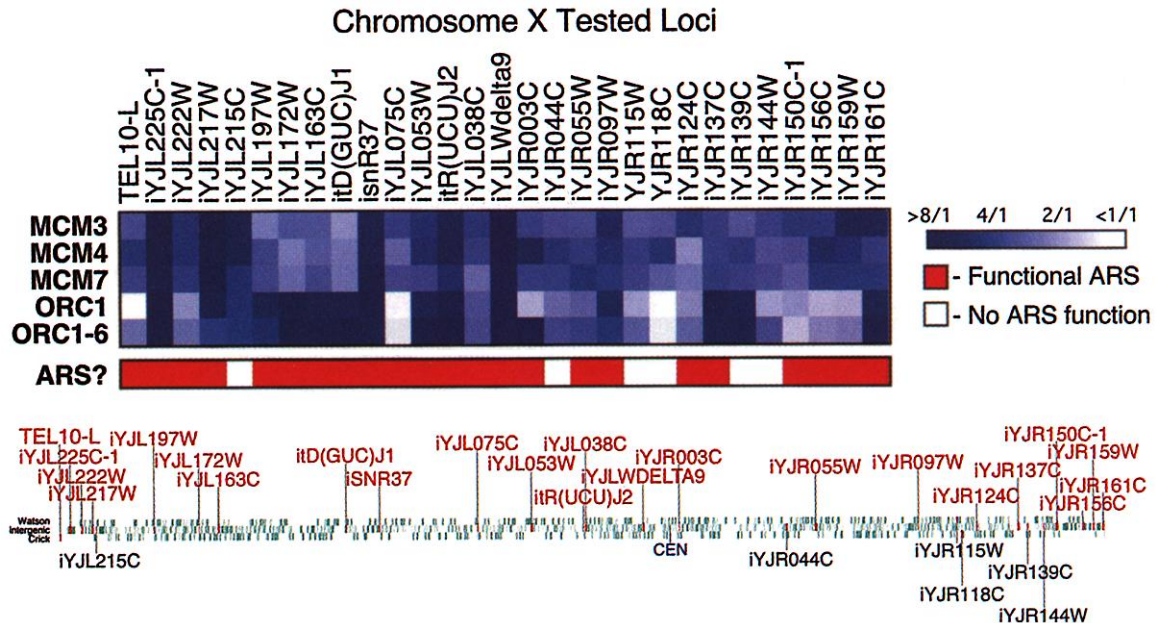


Fig. 2. Genome-wide location of potential replication origins. The genomic position of each probe present on the arrays is plotted to scale as a green bar (Web table 3) (18). The predicted origin-containing loci (pro-ARS) are plotted to scale as a red bar and named systematically (Web table 2) (18). Variations in width and

apparent intensities of green or red color reflect different probe lengths, not hybridization ratios. Probes to Watson and Crick ORFs are plotted on the top and bottom rows; intergenic sequences are plotted on the center rows. Asterisks indicate known ARSs that were not identified.

Fig. 3. ORC and MCM binding to chromosome X identify potential replication origins. The average binding ratios (blue/white) of ORC and MCM proteins to predicted origin-containing loci on chromosome X are shown. ARS activity is indicated in red. The positions of the tested loci are graphically represented as in Fig. 2. Locus names with no ARS function are in black.



(Web table 5) (18, 26). We also compared colocalizations of pro-ARSs with transposable elements (Ty), long-terminal repeats (LTR) (both solo and Ty-associated), and tRNAs (Table 1). Pro-ARSs and each of these elements colocalized within the same intergenic region four times more frequently than expected on a random basis. Because Ty, LTR, and tRNA sequences frequently colocalize with each other, the correlation with potential origins may derive from only one of these chromosomal elements or a combination (27). The highest level of correlation occurred with LTRs, including LTRs not associated with a tRNA or Ty element (Table 1, independent LTR). LTRs contain transcription initiation and termination signals that may establish chromatin domains that influence the function of replication origins. Future analyses of the functional interactions between replication origins and transposable elements should provide important insights into the regulation of origin function, genetic rearrangements, and the molecular evolution of eukaryotic genomes.

An important question is how replication origins are regulated such that distinct and characteristic initiation frequencies of origins occur along chromosomes. We investigated whether any relationship exists between the association of ORC and MCM proteins with potential origins and their function as replication origins at their chromosomal loci and observed that inefficient origins (except for silencer-associated origins) appeared to elude identification (Fig. 1, ARS304 and ARS608). Reanalysis by ChIP (with specific primers instead of DNA microarrays) of ORC and MCM binding to representative efficient and inefficient origins revealed a strong correlation between ORC binding and efficient origin activation (Web fig. 1) (18). We cannot exclude the possibility that chromatin

structures at these weak origins may sterically inhibit ChIP analysis; nevertheless, the use of a polyclonal antibody as well as different, tagged proteins moderates this concern. The low level of ORC and MCM binding at these weak origins suggests that we predominantly have identified active origins throughout the genome. However, we observed higher levels of ORC and MCM binding at the inefficient, telomere-associated and *HM* silencer-associated origins (Fig. 1).

Analysis of genome-wide replication timing in yeast has been reported recently, allowing approximation of the positions of efficient replication origins (28). The ChIP-based method used here identified the majority of origins found in the previous report and complements and extends that study by providing direct, high-resolution mapping of potential origins, including an additional subset of inefficient origins. The distinct characteristics and distributions of these ORC- and MCM-binding sites have revealed extensive associations with repetitive DNA elements with implications for other important chromosomal functions such as chromatin organization and genome stability. The success of this approach in *S. cerevisiae* suggests that similar approaches can be used to identify potential origins in other organisms. The roles of ORC and the MCM proteins are conserved in multicellular eukaryotes, where origin structure and location has been much more difficult to characterize by conventional methods (16, 29).

References and Notes

1. C. S. Newlon, in *DNA Replication in Eukaryotic Cells*, M. L. DePamphilis, Ed. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1996), pp. 873–914.
2. D. T. Stinchcomb, K. Struhl, R. W. Davis, *Nature* **282**, 39 (1979).
3. C. L. Hsiao, J. Carbon, *Proc. Natl. Acad. Sci. U.S.A.* **76**, 3829 (1979).

4. S. P. Bell, B. Stillman, *Nature* **357**, 128 (1992).
5. J. R. Broach et al., *Cold Spring Harb. Symp. Quant. Biol.* **47**, 1165 (1983).
6. A. Deshpande, C. Newlon, *Mol. Cell. Biol.* **12**, 4305 (1992).
7. Y. Marahrens, B. Stillman, *Science* **255**, 817 (1992).
8. R. Y. Huang, D. Kowalski, *EMBO J.* **12**, 4521 (1993).
9. C. S. Newlon, J. F. Theis, *Curr. Opin. Genet. Dev.* **3**, 752 (1993).
10. K. L. Friedman, M. K. Raghuraman, W. L. Fangman, B. J. Brewer, *J. Cell Sci. Suppl.* **19**, 51 (1995).
11. K. L. Friedman, B. J. Brewer, W. L. Fangman, *Genes Cells* **2**, 667 (1997).
12. C. S. Newlon et al., *Genetics* **129**, 343 (1991).
13. A. Poloumienko, A. Dershowitz, J. De, C. S. Newlon, *Mol. Biol. Cell* **12**, 3317 (2001).
14. K. Shirahige, T. Iwasaki, M. B. Rashid, N. Ogasawara, H. Yoshikawa, *Mol. Cell. Biol.* **13**, 5043 (1993).
15. M. Yamashita et al., *Genes Cells* **2**, 655 (1997).
16. T. J. Kelly, G. W. Brown, *Annu. Rev. Biochem.* **69**, 829 (2000).
17. B. Ren et al., *Science* **290**, 2306 (2000).
18. Supplementary materials are available on Science Online at www.sciencemag.org/cgi/content/full/294/5550/2357/DC1.
19. O. M. Aparicio, in *Current Protocols in Molecular Biology*, F. M. Ausubel et al., Eds. (Wiley, New York, 1999), vol. 4, pp. 21.3.1–21.3.12.
20. J. R. Lipford, S. P. Bell, *Mol. Cell* **7**, 21 (2001).
21. E. J. Louis, *Yeast* **11**, 1553 (1995).
22. B. J. Brewer, *Curr. Opin. Genet. Dev.* **4**, 196 (1994).
23. Y. Marahrens, B. Stillman, *EMBO J.* **13**, 3395 (1994).
24. S. Loo, J. Rine, *Annu. Rev. Cell Dev. Biol.* **11**, 519 (1995).
25. D. T. Pak et al., *Cell* **91**, 311 (1997).
26. P. T. Spellman et al., *Mol. Biol. Cell* **9**, 3273 (1998).
27. J. D. Boeke, S. B. Sandmeyer, in *The Molecular and Cellular Biology of the Yeast Saccharomyces*, J. R. Broach, J. R. Pringle, E. W. Jones, Eds. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1991), vol. 1, pp. 193–262.
28. M. K. Raghuraman et al., *Science* **294**, 115 (2001).
29. D. M. Gilbert, *Science* **294**, 96 (2001).
30. We thank D. Gottschling and J. Tower for critical reading of the manuscript, M. K. Raghuraman for sharing results before publication, and F. Ge, N. Hananet, N. Mehran, S. Rozen, and T. Volkert for technical assistance. This work was supported by NSF (J.J.W.), Howard Hughes Medical Institute (S.P.B.), NIH (grants GM52339 to S.P.B. and GM34365 to R.A.Y.), Corning, and Burroughs-Wellcome Fund (O.M.A.).

7 September 2001; accepted 30 October 2001