COMMENTARY

Mathematical description of eukaryotic chromosome replication

Huilin Li[a] and Michael E. O'Donnell[b,c,1]

The DNA genome must be completely duplicated with exquisite accuracy before a cell divides. The origin of replication (the place replication starts) is a single unique DNA sequence in a bacterial genome (1). By contrast, eukaryotic chromosomes have numerous initiation start sites, but these sites are not defined by a particular sequence and they change location in each cell cycle (2, 3). How such a vital process as DNA replication is orchestrated in seemingly random fashion is a mystery. In PNAS, Kelly and Callegari (4) devise a simple mathematical model that largely describes global chromosome replication dynamics in the fission yeast *Saccharomyces pombe*, using extensive global datasets from Kaykov and Nurse (5) and from Daigaku et al. (6) of the Carr laboratory. Kelly and Callegari's model requires few parameters and assumes that selection of initiation sites is stochastic. Their mathematical modeling depends on two main features: (*i*) AT-rich DNA to which the *S. pombe* origin recognition complex (ORC) binds, and (*ii*) DNA that is outside transcription units. The ability to describe the global landscape of replication over the *S. pombe* genome gives hope that the approach may apply to higher eukaryotes such as ourselves.

The quest to identify replication origin sequences in eukaryotes has a long and torturous history (3). Over three decades, many laboratories have tried to locate the elusive origin sequence in mammals, but this has often led to conflicting conclusions and hotly debated results. A main region of study was the 55-kb intergenic region between the dihydrofolate reductase (DHFR) and 2BE2121 loci of Chinese hamster ovary cells, a region that amplifies in response to methotrexate treatment (7). Dissection of this initiation region mostly resulted in less and less frequent initiations, although a few regions appeared to hold promise, including a region as small as 500 bp (8). While the search for defined origin sequences in eukaryotes finally came up empty-handed, it reinforced an emerging view that eukaryotic initiation-site selection and timing of firing is a stochastic process.

Defined origins do exist in one of the smallest eukaryotes, the budding yeast *Saccharomyces cerevisiae* (1, 2), but exactly which origins are used in a given cell cycle and the mechanisms that determine when they fire is not yet understood. The discovery of defined autonomously replicating origin sequences in budding yeast has led to a detailed understanding of the biochemistry of origin activation. The six-subunit ORC was first isolated by Bell and Stillman (9). Since then, contributions of many laboratories over two dozen years has led to the detailed mechanistic picture in which ORC, along with Cdc6 and Cdt1, loads two head-to-head stacked Mcm2-7 rings onto DNA in G1 phase, referred to as a pre-replicative complex (pre-RC) (10, 11). At the G1-to-S transition, the pre-RC is activated by several proteins and two kinases to assemble Cdc45 and GINS onto the minichromosome maintenance (MCM) protein complexes to form the active Cdc45/Mcm2-7/GINS (CMG) helicase identified by Ilves et al. (12) and Moyer et al. (13), both of the Botchan laboratory. The two CMG helicases generate bidirectional replication forks.

Dividing replication into two phases explained the "licensing" phenomenon identified by Blow and Laskey (14). Thus, PreRCs can only be assembled, or licensed, in G1 phase (i.e., the PreRC), and can only be fired in S phase, explaining how replication of a chromosome with numerous start sites is limited to

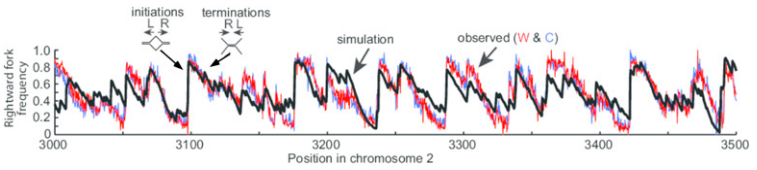

Fig. 1. Example of Pu-seq data and mathematical model fit. The black line is the mathematical model fit to experimental Pu-seq data [red, Watson (W) strand (Pol ε); blue Crick (C) strand (Pol δ)]. Rightward fork frequency increases to the right (R) of an initiation site, and decreases to the right of a termination site. L, left. Adapted from figure 4 of ref. 4.

[a]Structural Biology Program, Van Andel Research Institute, Grand Rapids, MI 49503; [b]Program of Biochemistry and Structural Biology, The Rockefeller University, New York, NY 10065; and [c]Howard Hughes Medical Institute, The Rockefeller University, New York, NY 10065

Author contributions: H.L. and M.E.O. wrote the paper.

The authors declare no conflict of interest.

See companion article on page 4973.

[1]To whom correspondence should be addressed. Email: odonnel@rockefeller.edu.

Published online February 19, 2019.

only once per cell cycle. For the job of DNA synthesis per se, eukaryotes use two different DNA polymerases (Pols). Studies by Burgers and Kunkel (15) (and their collaborators) using mutant Pols revealed that Pol ε replicates the leading strand and Pol δ replicates the lagging strand. Importantly, while other eukaryotes lack well-defined origins, they contain the same replication machinery as budding yeast. However, the global dynamics that underlie replication of chromosomes was still not understood for any eukaryote.

The fission yeast *S. pombe* is an attractive model to study replication because it has a small (13.6-Mb) genome dispersed on three chromosomes, yet its replication pattern is like that in *Xenopus*, mice, and humans. Thus, *S. pombe* lacks defined origins and initiates in AT-rich sites that are used inefficiently and, when selected, fire stochastically. Chuang and Kelly (16) previously identified that *S. pombe* ORC subunit 4 (Orc4) has nine AT-hook domains, strongly biasing it to AT-rich sequences. To gain insight into the global replication dynamics in *S. pombe* (i.e., the problem of origin selection and timing), Kelly and Callegari (4) utilize two extensive global datasets from refs. 5 and 6.

In the study by Kaykov and Nurse (5), single-molecule DNA combing was used to map the locations of initiation sites and their timing over the entire genome. Importantly, the study greatly advanced DNA combing techniques to examine individual DNA molecules of great lengths, up to 5 Mb, thus observing numerous origins in one DNA molecule. Consistent with stochastic events, examination of the same sequence from different cells showed distinct patterns of initiation. In fact, descendants from the same cell gave different patterns of origin selection/timing, indicating that inherited epigenetic marks do not dictate the *S. pombe* replication program. Interestingly, the frequency of initiation sites gradually increased fourfold over the first half of S phase, revealing that S phase is not a sharp transition from G1, but instead is a more gradual process. Initiation sites were mapped to 2-kb resolution, predicting about 1,200 initiation sites per cell, with an average distance of about 11 kb between them. While exciting, the new data did not explain the underlying mechanism of initiation-site selection and their firing time.

Daigaku et al. (6) were interested in the usage of Pol ε and Pol δ over the genome and confirmed that Pol ε and Pol δ are largely confined to leading and lagging strands, respectively, over the entire genome (with minor exceptions). To perform global mapping of Pol usage, they developed a Pol usage sequencing (Pu-seq) method to map global use of Pols δ and ε during replication. The Pu-seq technique uses certain mutants of each Pol that incorporate ribo-NTPs at elevated frequency, followed by alkaline cleavage at ribo-NMP incorporation sites that were mapped at nucleotide resolution by high-throughput sequencing. The data were binned to 300 bp, giving a very high-resolution map of global genome replication.

Kelly and Callegari (4) combine these global genome datasets along with key insights from other studies. A main lead was data indicating that initiation sites occur outside transcription units. For example, transcription had been inferred to prevent sites of replication initiation, and one of the earlier such studies placed a Gal promotor near an efficient origin in *S. cerevisiae*, which showed that transcription interferes with origin activity (17). Moreover, it had been noted that mammalian initiation sites in the classic DHFR locus did not occur in the body of the transcription unit, but upon inactivating the DHFR promotor, initiation sites were now observed to occur in the body of the gene (7). Moreover, the Pol usage data by Daigaku et al. (6) revealed that in *S. pombe*, many AT-rich sites that should bind ORC do not initiate if they are in transcription units. Thus, Kelly and Callegari (4) incorporate into the mathematical model that transcription interferes with pre-RCs, probably by either displacing pre-RCs or preventing their assembly. The model further incorporates AT-rich enhancement of *S. pombe* ORC (SpORC) binding in a probability distribution function that was optimized by the global experimental datasets. The model accurately describes the Pu-seq data in chromosome 2 at high resolution (Fig. 1).

The report in PNAS by Kelly and Callegari demonstrates that the complex task of selecting replication initiation sites and timing of their firing in *S. pombe* can be described by a simple stochastic mathematical model with surprisingly few variables.

There are interesting predictions of the mathematical model, some expected and others more surprising. It has long been known that cells contain many more MCM proteins than origins, often referred to as the "MCM paradox." Excess MCM proteins have been assumed to form pre-RCs that do not fire but are held in reserve to initiate synthesis for replication forks that stop prematurely in S phase. Consistent with this, the model suggests that excess pre-RCs may be loaded but are knocked off DNA by replication forks or transcription before they get a chance to fire. The model also predicts that very late replicating sequences extend beyond the normal S phase and may explain DNA synthesis in G2 phase as a longer time that is needed to complete replication of regions with a scarcity of pre-RCs (4).

The model further suggests that SpORC evolved to bind extragenic regions, which are typically rich in AT sequence. This suggests that similar evolutionary pressures may also apply to ORCs of higher eukaryotes. Although higher eukaryotes lack the AT hook in Orc4, some metazoan ORC subunits contain domains that bind histone modifications; thus, particular chromatin states may factor into higher eukaryotic replication dynamics (3, 18). One may expect that several other variables will be needed to explain global replication dynamics of higher eukaryotes, considering their greater complexity compared with yeast. For example, a particular metazoan cell line has a reproducible initiation-site profile, but different developmentally derived cells show quite different profiles (18), suggesting that 3D physical structure of the chromatin, transcription profile, or other developmental changes may be important determinants of initiation-site selection and timing. A classic case is *Xenopus*, in which global transcription is suppressed in embryos and potential initiation sites are uniformly distributed (19), but changes at the midblastula stage result in nonuniform and stochastic initiation sites.

Of possible medical importance, mutations are associated with late-replicating regions (20). Thus, mathematical prediction of late-replicating regions may promote understanding of some types of genomic instability that lead to pathological states, including cancer. Interestingly, earlier studies showed that in *S. pombe*, repair of DNA lesions can occur by homologous recombination, a relatively error-free process, but this process is shut down in G2 phase, and repair shifts to error-prone processes like translesion DNA Pols (21) which, if general, may help explain mutagenesis during late replication.

In overview, the report in PNAS by Kelly and Callegari (4) demonstrates that the complex task of selecting replication

initiation sites and timing of their firing in *S. pombe* can be described by a simple stochastic mathematical model with surprisingly few variables and, thus, provides a view that the stochastic replication program of higher eukaryotic cells may also be understood and described by mathematical modeling in the future.

1 O'Donnell M, Langston L, Stillman B (2013) Principles and concepts of DNA replication in bacteria, archaea, and eukarya. *Cold Spring Harb Perspect Biol* 5:a010108.
2 Bleichert F, Botchan MR, Berger JM (2017) Mechanisms for initiating cellular DNA replication. *Science* 355:eaah6317.
3 Prioleau MN, MacAlpine DM (2016) DNA replication origins—Where do we begin? *Genes Dev* 30:1683–1697.
4 Kelly T, Callegari AJ (2019) Dynamics of DNA replication in a eukaryotic cell. *Proc Natl Acad Sci USA* 116:4973–4982.
5 Kaykov A, Nurse P (2015) The spatial and temporal organization of origin firing during the S-phase of fission yeast. *Genome Res* 25:391–401.
6 Daigaku Y, et al. (2015) A global profile of replicative polymerase usage. *Nat Struct Mol Biol* 22:192–198.
7 Hamlin JL, Mesner LD, Dijkwel PA (2010) A winding road to origin discovery. *Chromosome Res* 18:45–61.
8 Burhans WC, Vassilev LT, Caddle MS, Heintz NH, DePamphilis ML (1990) Identification of an origin of bidirectional DNA replication in mammalian chromosomes. *Cell* 62:955–965.
9 Bell SP, Stillman B (1992) ATP-dependent recognition of eukaryotic origins of DNA replication by a multiprotein complex. *Nature* 357:128–134.
10 Evrin C, et al. (2009) A double-hexameric MCM2-7 complex is loaded onto origin DNA during licensing of eukaryotic DNA replication. *Proc Natl Acad Sci USA* 106:20240–20245.
11 Remus D, et al. (2009) Concerted loading of Mcm2-7 double hexamers around DNA during DNA replication origin licensing. *Cell* 139:719–730.
12 Ilves I, Petojevic T, Pesavento JJ, Botchan MR (2010) Activation of the MCM2-7 helicase by association with Cdc45 and GINS proteins. *Mol Cell* 37:247–258.
13 Moyer SE, Lewis PW, Botchan MR (2006) Isolation of the Cdc45/Mcm2-7/GINS (CMG) complex, a candidate for the eukaryotic DNA replication fork helicase. *Proc Natl Acad Sci USA* 103:10236–10241.
14 Blow JJ, Laskey RA (1988) A role for the nuclear envelope in controlling DNA replication within the cell cycle. *Nature* 332:546–548.
15 Burgers PMJ, Kunkel TA (2017) Eukaryotic DNA replication fork. *Annu Rev Biochem* 86:417–438.
16 Chuang RY, Kelly TJ (1999) The fission yeast homologue of Orc4p binds to replication origin DNA via multiple AT-hooks. *Proc Natl Acad Sci USA* 96:2656–2661.
17 Snyder M, Sapolsky RJ, Davis RW (1988) Transcription interferes with elements important for chromosome maintenance in Saccharomyces cerevisiae. *Mol Cell Biol* 8:2184–2194.
18 Nordman J, Orr-Weaver TL (2012) Regulation of DNA replication during development. *Development* 139:455–464.
19 Mahbubani HM, Paull T, Elder JK, Blow JJ (1992) DNA replication initiates at multiple sites on plasmid DNA in Xenopus egg extracts. *Nucleic Acids Res* 20:1457–1462.
20 Liu L, De S, Michor F (2013) DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes. *Nat Commun* 4:1502.
21 Callegari AJ, Kelly TJ (2016) Coordination of DNA damage tolerance mechanisms with cell cycle progression in fission yeast. *Cell Cycle* 15:261–273.