Machinery of DNA Replication

Nina Y. Yao¹ and Michael E. O’Donnell¹,²
¹Structural and Molecular Biology Program,
The Rockefeller University,
New York, NY, USA
²Howard Hughes Medical Institute,
The Rockefeller University,
New York, NY, USA

Introduction

The DNA genome must be duplicated with great precision before a cell divides in order to maintain the species. DNA replication is performed by numerous proteins that work together as a dynamic machine, referred to as a “replisome” (O’Donnell et al. 2013). The core components of replisome machines are present in cellular domains of life, Bacteria, Archaea, and Eukarya. These core replisome components are a helicase, DNA polymerases, RNA primase, the sliding clamp, and a clamp loader. The helicase of replisome machines is first assembled onto DNA at origins of replication by initiation proteins, after which the helicase couples ATP hydrolysis to unwind DNA and form replication forks, the Y-shaped DNA structure where the DNA strands are separated and duplicated (Bell and Labib 2016; O’Donnell et al. 2013). Primase associates with the helicase to generate short “RNA primers” that the DNA polymerases extend by matching complementary bases (i.e., dNTPs) to each separated strand, using them as templates for the new daughter duplexes. Sliding clamps are ring-shaped proteins that encircle primed sites and directly bind the DNA polymerases, holding them to DNA in a mobile fashion for the high processivity required to replicate long genomes. The clamps are placed around DNA by a clamp loader that couples ATP to open and close the clamps around DNA. All cells also contain a single-strand DNA (ssDNA)-binding protein that protects single-strand (ss) DNA from nuclease attack during replication. In addition to these core replisome factors, eukaryotic replisomes contain numerous ancillary factors involved in regulating replisomes in response to cellular conditions and to enable the replisome machine to partially disassemble and pass nucleosomes (Bell and Labib 2016).

DNA polymerases can only extend DNA in one direction, 5’-3’, because the dNTP substrates are activated at their 5’ position. Therefore, given that the DNA strands of double-strand (ds) DNA are antiparallel, only one DNA polymerase can extend DNA continuously in the direction of...
the fork, referred to as the “leading strand.” The antiparallel “lagging strand” must be extended in the opposite direction of the moving fork, and this entails sequential RNA priming and DNA synthesis to form a series of Okazaki fragments. The Okazaki fragments are then stitched together by ligase after removal of RNA primers, a 5’-3’ exonuclease. The overall process is referred to as “semidiscontinuous” replication.

In a twist of nature, three central players of the replisome, helicase, primase, and DNA polymerase, share no amino acid homology between bacteria and eukaryotes/archaea, indicating that the modern-day replication process evolved after the evolutionary split of bacteria and eukaryotes (see Table 1). Thus, replisome machines display different attributes among bacteria and eukaryotes (Yao and O’Donnell 2016). This review will briefly discuss each of the core replisome components and then provide an overview of the organization of the replisome machine in bacteria and eukaryotes.

**DNA Helicase**

All cellular replicative helicases are hexamer rings that encircle and translocate along one strand, acting as an ATP-driven wedge to split the duplex (Fig. 1) (O’Donnell and Li 2018). Despite the lack of sequence homology, the individual helicase subunits of bacteria and eukaryotes are composed of two globular domains, and the ATP site is in the C-terminal domain (Fig. 1a). The two-domain “dumbbell” shape of each subunit gives the hexamer the appearance of two stacked rings, an N-tier and a C-tier (Fig. 1b). The bacterial replicative helicase (e.g., *E. coli* DnaB) encircles the lagging strand and tracks 3’-5’ (Fig. 1b). As DnaB tracks along ssDNA, it adopts a spiral shape in which each subunit binds two adjacent phosphates of the DNA backbone (Itsathitphaisarn et al. 2012). ATP hydrolysis causes the bottom subunit to release the DNA and leapfrog the other subunits to take a position at the top of the spiral, as process known as “staircasing” (Fig. 1c). Circular permutation of this process enables helicase tracking along ssDNA without helicase rotation (Lyubimov et al. 2011). The ATPase motors in the C-domains are based on the RecA fold and track on ssDNA in front of the N-tier (O’Donnell and Li 2018).

The eukaryotic helicase contains six different, but homologous, MCM (minichromosome maintenance) subunits and forms a MCM2–7 heterohexamer that encircles the leading strand (Fig. 1b). However, MCM2–7 requires tight association with Cdc45 and the GINS tetramer (Sld5, Psf1, Psf2, Psf3) to become an active helicase and is therefore referred to as CMG (Cdc6, MCM2–7, GINS) (Bleichert et al. 2017; Ilves et al. 2010). Two CMG helicases are assembled on DNA at an origin in a complex process involving factors that do not travel with replication forks (Bell and Labib 2016; O’Donnell et al. 2013). The ATP sites are located in the C-domain of the MCM2–7 subunits, but unlike its bacterial counterpart, CMG encircles and tracks along the leading strand and moves N-tier first (Fig. 1b) (Georgescu et al. 2017). Moreover, the MCM ATP sites are formed by an AAA+ fold (ATPases associated with diverse cellular activities). Whether CMG movement occurs by the staircase, mechanism is not yet certain, as mutation of some ATP sites has very little effect on its activity (Ilves et al. 2010), and alternative mechanisms have been proposed (O’Donnell and Li 2018).

### Machinery of DNA Replication, Table 1  The core enzymes of cellular replisomes

<table>
<thead>
<tr>
<th>Protein</th>
<th>Folding pattern bacteria</th>
<th>Folding pattern eukaryotes</th>
<th>Common ancestor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Helicase</td>
<td>RecA fold (DnaB)</td>
<td>AAA+ fold (MCM2–7)</td>
<td>No</td>
</tr>
<tr>
<td>Primase</td>
<td>Toprim (DnaG)</td>
<td>DnaX family (Pol a-primase)</td>
<td>No</td>
</tr>
<tr>
<td>DNA Pol</td>
<td>C-family (Pol III)</td>
<td>B-family (Pols ε and δ)</td>
<td>No</td>
</tr>
<tr>
<td>Clamp loader</td>
<td>AAA+ pentamer</td>
<td>AAA+ pentamer</td>
<td>Yes</td>
</tr>
<tr>
<td>Sliding clamp</td>
<td>6-domain ring (beta)</td>
<td>6-domain ring (PCNA)</td>
<td>Yes</td>
</tr>
</tbody>
</table>

The Machinery of DNA replication, Table 1 presents the core enzymes of cellular replisomes.
DNA Polymerase

There exist many types of DNA polymerases for replication and repair (Yang and Gao 2018). Replicative DNA polymerases (Pol), and some repair Pols, contain a proofreading 3’-5’ exonuclease site that corrects rare misinsertion errors by the polymerase active site. There are several families of DNA polymerases that share no sequence homology, but two families are responsible for replication: C-family Pol III in bacteria and B-family Pol ε and Pol δ in eukaryotes (Burgers and Kunkel 2017; O’Donnell et al. 2013; Yang and Gao 2018). The replicative polymerase of the
**E. coli** bacterium is the Pol III core, a three-subunit complex composed of the C-family α subunit DNA polymerase, the 3′-5′ proofreading ε subunit, and θ subunit. Within α is a PHP domain (polymerase and histidinol phosphatase), which in some bacteria contains a proofreading exonuclease activity, while other α subunits, like **E. coli** Pol III core, recruit a separate 3′-5′ exonuclease subunit, and its PHP domain has become a structural element without activity (Barros et al. 2013). Eukaryotic cells use two different replicative Pols, both in the B-family; Pol ε acts on the leading strand, and Pol δ synthesizes the lagging strand (reviewed in Burgers and Kunkel 2017; Yang and Gao 2018). The Pol and exonuclease activities are on the same chain in both Pol ε and Pol δ. Pol ε also contains the Dpb2,3,4 subunits, but only Dpb2 is essential (Pursell and Kunkel 2008). Pol δ also contains two accessory subunits, Pol31 which is essential and Pol32 which can be deleted (Burgers and Kunkel 2017). Functions of the accessory subunits in Pols ε and δ are not yet well understood (Burgers and Kunkel 2017).

All DNA polymerases have the shape of a right hand (Fig. 2a, b), with fingers, palm, and thumb domain (Yang and Gao 2018). Thus while primary amino acid sequences of Pols of different Pol families may not be homologous, their basic shape is the same. The most conserved domain is the palm, containing the catalytic center. The thumb helps bind the duplex, and the fingers bind the dNTP and close down onto the palm to make a chamber with the dimensions of a base pair. If the base pair is incorrect, it does not fit into the dimensions of the chamber, slowing the catalytic step and giving time for the fingers to release the dNTP and try again. There are three acidic residues in the palm domain that coordinate metal ions and perform the chemistry of dNTP incorporation (Yang and Gao 2018). The architecture of this domain varies with different families, but the 3-metal ion chemistry is presumed to be the same (Yang and Gao 2018).
The metals activate the 3’ primer terminus to attack the \( \alpha \)-phosphate of an incoming dNTP, incorporating a dNMP while the pyrophosphate departs.

If an incorrect nucleotide is added to the chain, the polymerase is slow to add a new dNTP to the mismatched 3’ terminus. This provides time for movement of the mismatch to the exonuclease site, requiring about 3–4 base pairs of spontaneous unwinding. Once the mismatch is removed, the 3’ terminus is rebound at the palm for correct insertion. Error frequencies of replicative polymerases are about one in a million dNMP incorporation events, and the exonuclease adds another 10–100-fold overall accuracy by removal of incorrectly added dNMP.

Primase

DNA polymerases cannot perform the initial condensation of two NTPs; this step is specific to RNA polymerases. Thus DNA Pols require a pre-existing primed site. Primed sites are formed by an enzyme called primase which is limited in the length of RNA that it can produce (reviewed in Bell 2019). Bacterial primase (e.g., DnaG in *E. coli*) is a single subunit enzyme, and it interacts transiently with the helicase to form a RNA primer about a dozen nucleotides in length. Sequence and structural studies find that bacterial primase shares homology to topoisomerases, referred to as a Toprim domain (Podobnik et al. 2000). The need to bind helicase for activity ensures that RNA primers are formed at a replication fork junction (Fig. 3).

The eukaryotic primase is the four-subunit Pol \( \alpha \)-primase (reviewed in Perera et al. 2013). The largest subunit, Pol1, is a DNA polymerase, and the smaller subunits are Pol12, Pri1, and Pri2. The Pri subunits of Pol \( \alpha \)-primase contain the RNA priming activity and generate RNA primers of 6–8 nucleotides. The RNA primer is internally transferred to Pol1, which extends the primer with dNTPs to form a hybrid RNA-DNA primer of about 20–25 nucleotides. Pol \( \alpha \)-primase lacks a 3'-5' exonuclease, and accordingly DNA synthesis lacks the high fidelity of Pols \( \varepsilon \) and \( \delta \). The RNA and DNA synthesized by Pol \( \alpha \)-primase are removed by the FEN1 nuclease or the Dna2 nuclease. Studies in the millisecond time range show that Fen1 efficiently collaborates with Pol \( \delta \) on PCNA to remove RNA-DNA made by Pol \( \alpha \)-primase (Stodola and Burgers 2016). Deep sequencing studies indicate that Okazaki fragment length is determined by nucleosome deposition (about 150 bp) that occurs rapidly as Okazaki fragments are produced (Smith and Whitehouse 2012).

Sliding Clamps and Clamp Loader

The sliding clamp was discovered for its function in DNA replication (Stukenberg et al. 1991), but it is now known that sliding clamps are utilized by many processes besides replication, including various repair reactions, cell cycle kinases, and translesion synthesis (Georgescu et al. 2015a). The *E. coli* \( \beta \) sliding clamp was the first protein known to functionally encircle DNA (Stukenberg et al. 1991). But we now know that many DNA metabolic enzymes encircle DNA, including replicative helicases, origin binding factors, and some topoisomerases.

**Sliding Clamps and Clamp Loader**

The sliding clamp was discovered for its function in DNA replication (Stukenberg et al. 1991), but it is now known that sliding clamps are utilized by many processes besides replication, including various repair reactions, cell cycle kinases, and translesion synthesis (Georgescu et al. 2015a). The *E. coli* \( \beta \) sliding clamp was the first protein known to functionally encircle DNA (Stukenberg et al. 1991). But we now know that many DNA metabolic enzymes encircle DNA, including replicative helicases, origin binding factors, and some topoisomerases.

---

**Machinery of DNA Replication, Fig. 3** Pol \( \alpha \)-primase makes a hybrid RNA-DNA primer. Left: the Pri subunits form a short RNA primer. Middle: the RNA primer is transferred to the DNA polymerase subunit (Pol1) without enzyme dissociation. Right: DNA Pol1 extends the RNA to form a 20–25 RNA-DNA primer.
The bacterial β clamp monomer contains three domains with the same folding pattern, implying they derived from gene duplications, and it dimerizes to form the six-domain ring (Fig. 4a) (Kong et al. 1992). The eukaryotic PCNA contains two domains per monomer and trimerizes to form a six-domain ring, predicted from the β structure (Jeruzalmi et al. 2002; Kong et al. 1992) (Fig. 4a). Hence sliding clamps were likely present in the last universal cellular ancestor, unlike the Pol, helicase, and primase.

The clamp loader is composed of five essential subunits in all cells. In the bacterial clamp loader, three subunits (tau) are identical, while in eukaryotes all five subunits are encoded by different genes (Jeruzalmi et al. 2002). All five subunits contain an N-terminal two-domain AAA+ module for ATP binding (Kelch et al. 2011). In both bacteria and the eukaryotic RFC (replication factor C) clamp loader, the subunits are arranged as a circular pentamer through tight association of their C-domains, forming a cap, or collar from which the AAA+ modules are suspended (Kelch et al. 2011). The ATP sites are located at subunit interfaces in which one subunit binds the ATP and the adjacent subunit contributes a catalytic arginine finger, required for ATP hydrolysis. This interfacial location of ATP
sites enables coordinated action of all the subunits in the clamp loading process (Fig. 4). Clamp loaders bind ATP to open their respective clamp, but hydrolysis does not occur until both clamp and DNA are bound.

Clamp loaders bind DNA within a central chamber, and the DNA scaffolds an architectural change in the clamp loader in which the subunits achieve a spiral pitch similar to dsDNA (Fig. 4). The conformation change upon DNA binding brings catalytic residues within the clamp loader into proper geometry for ATP hydrolysis. Upon ATP hydrolysis the clamp loader ejects from the clamp, enabling the clamp to snap shut around dsDNA. Specificity for the ss/ds junction is the result of the closed cap that is formed by the C-domains. Thus, the DNA inside the clamp loader must make a sharp bend to exit the gap in the side of the clamp loader, located between two of the subunits. The sharp bend underlies how the clamp loader derives specificity for a primed junction because the ssDNA at the primed junction has the requisite flexibility to enable the dsDNA to fit into the center of the clamp loader and ssDNA to bend out the side of the clamp loader.

Replisome Architecture

Bacterial Replisome

It was long been thought that the replisome would contain two DNA polymerases, one for each strand, and that the antiparallel lagging strand would form a DNA loop for each Okazaki fragment (Fig. 5). Indeed DNA looping is observed in the *E. coli* and model phage systems (Yao and O’Donnell 2008). In an unexpected twist, the *E. coli* replisome has three DNA Pol III cores, not just two (McInerney et al. 2007) (Fig. 5a). The three copies of tau subunit in the clamp loader contain an extended C-terminal region that projects out of the clamp loader collar to bind both DnaB and Pol III core (Lewis et al. 2016). Hence, the clamp loader organizes the replisome, connecting three Pol III cores to one
hexamer helicase (Fig. 5a). Considering DNA only has two strands, one may question the value of the third Pol III core. Single-molecule studies comparing replisomes having either 2 or 3 Pol III cores show the 3-Pol replisome is more processive and better at completion of Okazaki fragments compared to the 2-Pol form (Georgescu et al. 2011). Hence, the third Pol provides a tighter grip of the replisome to DNA and a greater capability to complete Okazaki fragments.

The picture of a sliding clamp holding DNA polymerase to DNA fits nicely for the continuous leading strand, but the lagging strand is made as a series of 1–2 kb Okazaki fragments (O’Donnell et al. 2013). At a fork rate of 1 kb/s, the lagging Pol must rapidly recycle from the end of an Okazaki fragment to a new RNA primer made by primase every second or two. Two processes account for rapid Pol recycling. First, Pol III core rapidly ejects from its β clamp upon finishing a DNA fragment and recycles to a new primed site (O’Donnell 1987). This process is referred to as “collision release” because Pol III ejects from β-DNA upon colliding with the 5′ terminus of a previous Okazaki fragment. Once released, the Pol III core rebinds a new clamp loaded onto the next RNA primer synthesized at the fork. The second process is called “signal release” because the Pol releases from the clamp before competing an Okazaki fragment, suggesting it is signaled in some way to release from the beta clamp prematurely (Benkovic and Spiering 2017). The nature of the signal is unclear, but torsional strain by two connected Pols traveling helical paths is sufficiently strong to disconnect the Pol-clamp complex before completing replication (Kurth et al. 2013). Collision and signal release generalizes to the T4 phage (Benkovic and Spiering 2017).

It was long thought that the leading and lagging strand Pols of the replisome were coordinated, such that if one stopped, the other would also stop. However, these views were dispelled by ensemble and single-molecule studies showing that leading synthesis could proceed in the absence of lagging strand synthesis (McInerney and O’Donnell 2004; Yao et al. 2009). Thus, the Pols are physically coupled but not functionally coupled (Lewis et al. 2016).

Eukaryotic Replisome

Eukaryotic replisomes move about 25 bp/s and form Okazaki fragments of about 150 bp in length (Bell and Labib 2016; O’Donnell et al. 2013). While Pols ε and δ were known to be required for replication, the strands they operated upon only recently came to light from genetic studies by Tom Kunkel, Peter Burgers, and their collaborators using active site mutants that leave mutagenic signatures during function (Burgers and Kunkel 2017). These studies showed that Pol ε replicates the leading strand and Pol δ replicates the lagging strand.

The architecture of the eukaryotic replisome is coming into focus from EM structural studies of reconstituted replisome machines (Pellegrini and Costa 2016; Sun et al. 2015), and genetic and biochemical studies have determined the process by which Pols ε and δ are targeted to their respective strands (Burgers and Kunkel 2017). Pol ε binds CMG directly, explaining how it targets the leading strand (Langston et al. 2014). Unexpectedly, Pol ε could not extend lagging strand primers, and further studies demonstrated this is due to the intracellular concentrations of factors in which RFC blocks Pol ε function with PCNA, unless Pol ε is bound to CMG (Schauer and O’Donnell 2017). By contrast, Pol δ is not inhibited by RFC and is fully active on the lagging strand but is slow and distributive on the leading strand (Georgescu et al. 2015b). Pol δ, like E. coli Pol III, ejects from PCNA-DNA upon colliding with slow CMG (Schauer and O’Donnell 2017). Both Pol ε and Pol δ rapidly take over primed sites from Pol α-primase, regardless of which strand Pol α-primase is on, followed by replacement by the correct Pol, thereby ensuring that Pol α-primase only synthesizes primed sites (Georgescu et al. 2015b). Hence, Pol α, δ, ε asymmetry is now understood by discovery of these several rules of polymerase engagement.

CryoEM analysis of CMG revealed an atomic model of its 11-subunit structure bound to a replication fork, revealing that CMG tracks on DNA N-tier first, opposite the bacterial helicase (Georgescu et al. 2017). Negative stain EM, combined with cross-linking mass spectrometry, showed that Pol ε binds the
C-tier of CMG (Sun et al. 2015). Similar studies showed that Pol α-primase binds the Ctf4 homotrimer (Simon et al. 2014) located on the opposite side of CMG from Pol ε (Sun et al. 2015). Hence, the CMG helicase is sandwiched by the leading and lagging strand DNA polymerases (Fig. 5b). The leading strand proceeds through the central channel of CMG and reaches Pol ε-PCNA on the C-side of CMG. The deflected strand off the N-tier of CMG is immediately accessible to the Pol α-primase that resides there. Priming is then followed by use of Pol δ-PCNA for Okazaki fragment fill-in. It is interesting to note that a helicase that has one Pol above it and another Pol below it is a different replisome arrangement from textbook illustrations that show both DNA Pols trail the helicase.

Future Directions

Numerous questions remain for future studies that have important application to genomic instability and thus cancer and other pathological conditions. There are many ancillary factors that travel with replisomes not discussed here, but understanding their function is on the short term horizon. The process by which replisomes interface with DNA repair processes is still ill-understood. A prime example is fork regression repair (Bhat and Cortez 2018). In eukaryotes fork regression is important to genomic integrity. Among the enzymes involved are Rad51 and BRCA2, enzymes of double-strand break repair, central to genomic stability and cancer. Upon fork encounter with a leading strand blocking lesion, the fork is proposed to reverse, involving BRCA2 and Rad51, along with a host of other factors. However, the field of fork reversal has yet to study the effect of replication enzymes on this important process. For example, do all forks reverse, even enzymatically active forks? If the only forks that reverse are those stopped by a lesion, what prevents the fork reversal enzymes at normal forks? Also, every nucleosome is encountered by a replisome, and mutants in Pol ε, Pol α, Ctf4, and MCM2 reveal they are involved in maintenance of the epigenome, implying a role in differentiation. How epigenetic marks on nucleosomes bound to the parental duplex are transferred to the daughter strands is a new area of study. There are numerous other questions, including how replisomes deal with DNA bound proteins and transcribing RNA polymerase. The new techniques of cryoEM and of new high resolution light microscopy in cells are enormous new tools for discovery of new structures and how these machines function in a cellular context. The future is bright with promise, and many new questions have emerged from the recent and rapid advances in this critical field of genome replication.

References

Georgescu R, Langston L, O’Donnell M (2015a) A proposal: evolution of PCNA’s role as a marker of newly replicated DNA. DNA Repair (Amst) 29.4:15


