Cell, Vol. 120, 167–168, January 28, 2005, Copyright ©2005 by Elsevier Inc. DOI 10.1016/j.cell.2004.11.031

Letter to the Editor

Eukaryotic/Archaeal Primase and MCM Proteins Encoded in a Bacteriophage Genome

During the evolution of life on Earth, two distinct DNA replication machineries have emerged: that shared by the archaea and eukaryotes and that of bacteria. DNA replication is fundamental to the life of all cells, so this dichotomous evolutionary distribution is surprising. Here, we describe the identification of a protein with homology to eukaryotic DNA primase and MCM encoded within a prophage that is integrated in the genome of the bacterium *Bacillus cereus*.

Despite mechanistic similarities, the core components of the bacterial and archaeal/eukaryotic DNA replication machineries possess little primary sequence homology (Edgell and Doolittle, 1997; Leipe et al., 1999; Forterre, 1999). A nonorthologous gene displacement event has been proposed to account for this dichotomy, with the original genes having been replaced by nonhomologous counterparts. In the replicon takeover hypothesis, Forterre has suggested that a viral origin for these proteins may explain this puzzling gap in the evolutionary tree (Forterre, 1999). However, to date there has been no direct evidence for bacteriophage harboring primary sequence homologs of core archaeal/eukaryotic replication proteins or archaeal/eukaryal viruses with bacterial replication-associated genes. In this light, it is worthwhile to note that PCNA, the sliding clamp of archaea and eukaryotes, structurally resembles the sliding clamp of bacteriophage in the T4 family. However, in contrast to the situation we describe below, there is no significant primary sequence similarity between the T4 clamp and PCNA (Shamoo and Steitz, 1999; Moarefi et al., 2000).

During a database search for homologs of the archaeal/eukaryotic replicative helicase, the MCM complex (Bell and Dutta, 2002), we identified a gene encoding an MCM-related protein in the genome of the bacterium *Bacillus cereus* ATCC 14579 (Ivanova et al., 2003). Significantly, this gene (BC1863) is encoded within an integrated phage (phBC6A51) in the *B. cereus* genome (Ivanova et al., 2003). Furthermore, the MCM-related gene is located within an apparent polycistronic transcription unit, immediately upstream of a phage-encoded DNA polymerase (Figure 1). The C-terminal half of the protein is homologous to the AAA^+ domain of the MCMs (a search of the NCBI Conserved Domain Database [Marchler-Bauer et al., 2002] gave a probability value of $2 \times e^{-10}$ compared to the MCM2 family; see also Supplemental Figure S1 at http://www.cell.com/cgi/content/full/cgi/120/2/167/DC1/). Furthermore, iterative searching using the PsiBlast program (Altschul et al., 1997) revealed that the N-terminal quarter of the protein contains a region homologous to the catalytic subunit of the archaeal/eukaryotic DNA primase (Supplemental Figure S2; Frick and Richardson, 2001). This remarkable organization of both primase and MCM helicase motifs together in one protein has not previously been described for archaea and eukaryotes. Interestingly, this modular nature is reminiscent of the T7 phage primase-helicase protein. However, the T7 primase and helicase domains are related to bacterial DnaG and DnaB, respectively (Frick and Richardson, 2001).

How might the phage have acquired this gene? One possibility is by a relatively recent horizontal transfer event. There is now substantial evidence for lateral gene transfer between the prokaryotic domains of life. For example, up to 24% of the genes in the genome of the bacterium *Thermatoga maritima* have closest homologs in archaea rather than bacteria (Nelson et al., 1999), suggesting that almost a quarter of the *T. maritima* genome's coding potential may have been derived via lateral gene transfer from archaea. Crucially, however, while such observations provide support for ongoing exchange of a range of metabolic effector genes between archaea and bacteria, there has been no evidence to date for exchange of the core information processing machineries, including those of DNA replication and transcription, between life's domains.

The existence in early evolution of either an archaeal virus with a replication machinery that gave rise to that of present day bacteria or, as we describe here, a bacteriophage with archaeal/eukaryotic-like proteins is a cornerstone of the replicon takeover hypothesis (Forterre, 1999). Therefore, an enticing alternative explanation for the origin of the primase-MCM is that it represents an ancient progenitor family. If this is the case, then a potential source for the genes found in eukaryotes and archaea is immediately evident. A scenario could be envisaged in which such a phage integrated into the genome of an early bacterial-like organism at the time of branching of bacterial and archaeal/eukaryotic lineages. It is conceivable that the nature of this integration event led to the cell being dependent upon the phage replication machinery, resulting in the bifurcation of the replication machineries that we see today.

Finally, in addition to the implications for the origin and ongoing evolution of the primase and MCM genes discussed above, we note that the presence of the primase-MCM in the context of a phage will provide an invaluable genetic tool for the dissection of the function of these central replication proteins.

Acknowledgments

This work was supported by the Medical Research Council. We thank Ron Laskey for invaluable discussions.

Adam T. McGeoch and Stephen D. Bell*
Medical Research Council Cancer Cell Unit
Hutchison MRC Research Centre
Hills Road
Cambridge
CB2 2XZ
United Kingdom

*Correspondence: sb419@hutchison-mrc.cam.ac.uk

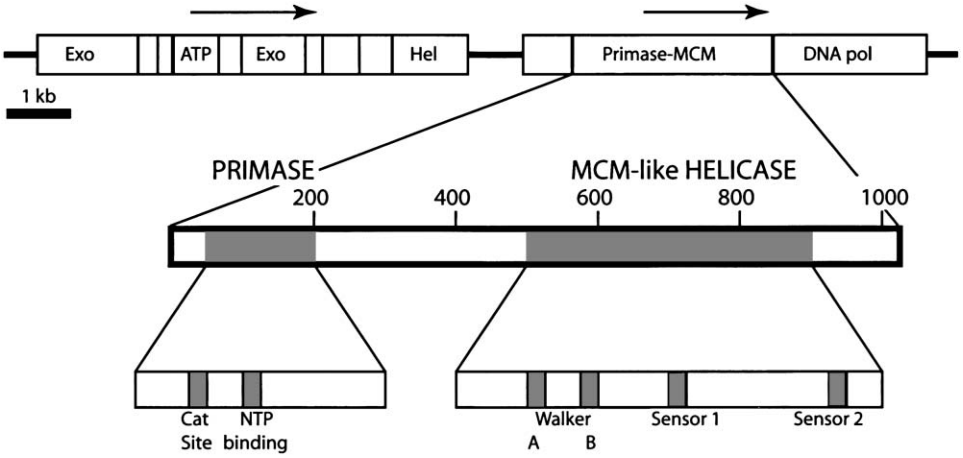

Figure 1. Organization of the Phage-Encoded Primase MCM Gene in *B. Cereus*

The top of the figure illustrates part of the integrated prophage phBC6A51 in the *B. cereus* genome (Ivanova et al., 2003). Arrows indicate the direction of transcription. Open reading frames are indicated by open rectangles, ORFs with clear homologs are named, Exo are exonucleases of the SbcC and SbcD families, ATP is a putative ATPase, Hel is a DEAD box containing putative DNA or RNA helicase, DNA pol is a DNA polymerase, and the primase-MCM is also indicated. All other ORFs encode phage proteins of unknown function.

A diagram of the primary sequence elements of the predicted translation product of the primase-MCM (open reading frame BC1863; Ivanova et al., 2003) is shown beneath the ORF diagram; numbers indicate positions in amino acids. Regions homologous to archaeal/eukaryotic DNA primase catalytic subunit and MCM are indicated by gray rectangles. These regions are expanded in the bottom of the figure. Putative catalytic site aspartates and NTP binding residues of the primase domain are indicated, as are the key features of the MCM AAA^{+} nucleotide binding domain.

Selected Reading

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Nucleic Acids Res. *25*, 3389–3402.

Bell, S.P., and Dutta, A. (2002). Annu. Rev Biochem. *71*, 333–374.

Edgell, D.R., and Doolittle, W.F. (1997). Cell *89*, 995–998.

Forterre, P. (1999). Mol. Microbiol. *33*, 457–465.

Frick, D.N., and Richardson, C.C. (2001). Annu. Rev. Biochem. *70*, 39–80.

Ivanova, N., Sorokin, A., Anderson, I., Galleron, N., Candelon, B., Kapatral, V., Bhattacharyya, A., Reznik, G., Mikhailova, N., and Lapidus, A. (2003). Nature *423*, 87–91.

Leipe, D.D., Aravind, L., and Koonin, E.V. (1999). Nucleic Acids Res. *27*, 3389–3401.

Marchler-Bauer, A., Panchenko, A.R., Shoemaker, B.A., Thiessen, P.A., Geer, L.Y., and Bryant, S.H. (2002). Nucleic Acids Res. *30*, 281–283.

Moarefi, I., Jeruzalmi, D., Turner, J., O'Donnell, M., and Kuriyan, J. (2000). J. Mol. Biol. *296*, 1215–1223.

Nelson, K.E., Clayton, R.A., Gill, S.R., Gwinn, M.L., Dodson, R.J., Haft, D.H., Hickey, E.K., Peterson, J.D., Nelson, W.C., and Ketchum, K.A. (1999). Nature *399*, 323–329.

Shamoo, Y., and Steitz, T.A. (1999). Cell *99*, 155–166.