Origins and evolution of the recA/RAD51 gene family: Evidence for ancient gene duplication and endosymbiotic gene transfer

Zhenguo Lin*†, Hongzhi Kong‡, Masatoshi Nei*§, and Hong Ma*†§

*Department of Biology and the Institute of Molecular Evolutionary Genetics and †Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, PA 16802; and ‡State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Xiangshan, Beijing 100093, China

Contributed by Masatoshi Nei, May 22, 2006

The bacterial recA gene and its eukaryotic homolog RAD51 are important for DNA repair, homologous recombination, and genome stability. Members of the recA/RAD51 family have functions that have differentiated during evolution. However, the evolutionary history and relationships of these members remains unclear. Homolog searches in prokaryotes and eukaryotes indicated that most eubacteria contain only one recA. However, many archaeal species have two recA/RAD51 homologs (RAD A and RADB), and eukaryotes possess multiple members (RAD51, RAD51B, RAD51C, RAD51D, DM C1, XRCC2, XRCC3, and recA). Phylogenetic analyses indicated that the recA/RAD51 family can be divided into three subfamilies: (i) RADα, with highly conserved functions; (ii) RADβ, with relatively divergent functions; and (iii) recA, functioning in eubacteria and eukaryotic organelles. The RADα and RADβ subfamilies each contain archaeal and eukaryotic members, suggesting that a gene duplication occurred before the archaea/eukaryote split. In the RADα subfamily, eukaryotic RAD51 and DM C1 genes formed two separate monophyletic groups when archaeal RAD A genes were used as an outgroup. This result suggests that another duplication event occurred in the early stage of eukaryotic evolution, producing the DM C1 clade with meiosis-specific genes. The RADβ subfamily has a basal archaeal clade and five eukaryotic clades, suggesting that four eukaryotic duplication events occurred before animals and plants diverged. The eukaryotic recA genes were detected in plants and protists and showed strikingly high levels of sequence similarity to recA genes from proteobacteria or cyanobacteria. These results suggest that endosymbiotic transfer of recA genes occurred from mitochondria and chloroplasts to nuclear genomes of ancestral eukaryotes.

DNA double-strand breaks (DSBs) can occur either spontaneously during DNA replication or by exogenous DNA-damaging agents. Efficient repair of DSBs is critical for genomic stability and cellular viability. A major DSB repair pathway is homologous recombination, which is also critical for meiosis and generation of genetic diversity. Among the best known recombination genes are the E. coli recA gene and its eukaryotic homologs RAD51s (2, 3). recA encodes a DNA-dependent ATPase that binds to single-stranded DNA and promotes strand invasion and exchange between homologous DNA molecules (4). The two eukaryotic recA homologs, RAD51 and DM C1, were first discovered in the budding yeast Saccharomyces cerevisiae and are structurally and functionally similar to the E. coli recA gene (5, 6).

Homologs of recA and RAD51 have then been identified in many prokaryotes and eukaryotes. In eubacteria, only one recA gene has been previously reported in each species (7). Unlike eubacteria, several archaeal species have two recA/RAD51-like genes, called RAD A and RAD B (Table 1, which is published as supporting information on the PNAS web site) (8, 9). Among eukaryotes, the budding yeast and the fission yeast Schizosaccharomyces pombe contain four RAD51-like genes (RAD51, DM C1, RAD55/hp55, and RAD57/hp57) (5, 6, 10, 11). In vertebrate animals and plants, there are usually seven different RAD51-like genes: RAD51, RAD51B, RAD51C, RAD51D, DM C1, XRCC2, and XRCC3 (Table 1) (12, 13). In addition, the flowering plants Arabidopsis thaliana and rice (Oryza sativa) each possess four conserved recA-like genes (refs. 14 and 15, as well as this article) that have higher levels of sequence similarity with eubacterial recA genes than with the eukaryotic RAD51-like genes.

Eukaryotic RAD51-like genes play important roles in homologous recombination, maintaining chromosomal integrity in both the mitotic and meiotic cell cycles (12, 16). For example, disruption of the mouse RAD51 gene can lead to cell death and embryo inviability (17). The DM C1 gene is specifically required for meiotic recombination in yeast, plants, and animals (12). Similarly, the RAD51, RAD51C, and XRCC3 homologs in Arabidopsis are also essential for meiosis (12, 15). Although all RAD51-like genes promote homologous recombination, they may have distinct functions.

Previous studies suggested that the RAD51 and DM C1 genes were generated by an ancient gene duplication in the common ancestor of all eukaryotes (8, 19). However, the evolutionary relationships of most recA/RAD51 family members remain unclear. To elucidate this question, we conducted extensive searches for recA/RAD51-like genes from public databases and performed detailed phylogenetic analyses of the genes identified. In this paper, we present the results of these studies and propose a model of the evolutionary history of the entire group of recA/RAD51 genes based on our findings.

Results

recA/RAD51-Like Genes. We performed BLAST searches for recA/ RAD51-like genes from various organisms, especially from the species whose genomes have been completely sequenced (Table 1). We found only one recA sequence in each eubacterial species, except in two species whose recA has recently duplicated (20, 21). Two recA/RAD51 family members, RAD A and RAD B, were found in many archaeal species such as Archaeoglobus fulgidus and Pyrococcus abyssi. However, only RAD A genes were found in some other archaeal groups (Table 1). In the budding and fission yeasts, the four known genes, RAD51, DM C1, RAD55/hp55, and RAD57/hp57, were recovered (5, 6, 10, 11). RAD55 and RAD57 are highly divergent from each other and from RAD51 and DM C1. Flowering plants, vertebrate animals, and sea urchin (Strongylocentrotus purpuratus) each have seven RAD51-like genes (RAD51, DM C1, RAD51B, RAD51C, RAD51D, XRCC2, and XRCC3). DM C1 was not present in the nematode (Caenorhabditis elegans) and the fruit

Conflict of interest statement: No conflicts declared.

Abbreviations: ML, maximum likelihood; NJ, neighbor-joining.

†To whom correspondence may be addressed. E-mail: nxm2@psu.edu or hxm16@psu.edu.

© 2006 by The National Academy of Sciences of the USA
proteins also contain a N-terminal domain of RAD51 and DMC1 (24, 25). Plant RecA terminal domain that is absent in other members. The RecA motif, which is a nonspecific DNA-binding domain. In contrast, regions of RADA, RAD51, and DMC1 have a modified HhH conserved consensus motifs, Walker A and Walker B, which are present in ATPases and confer ATP binding and hydrolysis activities (22). In addition, some member proteins have additional conserved N-terminal and/or C-terminal domains (Fig. 1). For example, archaeal RADA and eukaryotic RAD51, DMC1, RAD51B, RAD51C, RAD51D, and XRC3 proteins contain an N-terminal domain that is absent in RecA proteins. The N-terminal regions of RADA, RAD51, and DMC1 have a modified HhH motif, which is a nonspecific DNA-binding domain. In contrast, eubacterial and plant RecA proteins contain a conserved C-terminal domain that is absent in other members. The RecA C-terminal domains also bind to double-stranded DNA (23) and are similar in function to, but distinct in sequence from, the N-terminal domain of RAD51 and DMC1 (24, 25). Plant RecA proteins also contain a ~70-aa N-terminal region that is different from the RADA/RAD51 N-terminal regions. At least some plant RecA proteins contain organelle-targeting peptides (14, 15).

**Fig. 1.** Schematic diagram of domain structures of representative RecA/RAD51-like proteins, drawn to scale. Domain names are indicated in the figure.

**Fly** (Drosophila melanogaster) but was detected in silkworm (Bombyx mori). In addition, recA-like genes were found in plants such as *A. thaliana*, rice, slime mold (*Dictyostelium discoideum*), green algae, brown algae, and red algae but not in archaea, fungi, or animals (Table 1).

Multiple protein sequence alignment showed that all predicted RecA/RAD51-like proteins share a highly conserved central domain with ~230 aa, which we named here as the RecA/RAD51 domain (Fig. 1). In the RecA/RAD51 domain, there are two highly conserved consensus motifs, Walker A and Walker B, which are present in ATPases and confer ATP binding and hydrolysis activities (22). In addition, some member proteins have additional conserved N-terminal and/or C-terminal domains (Fig. 1). For example, archaeal RADA and eukaryotic RAD51, DMC1, RAD51B, RAD51C, RAD51D, and XRC3 proteins contain an N-terminal domain that is absent in RecA proteins. The N-terminal regions of RADA, RAD51, and DMC1 have a modified HhH motif, which is a nonspecific DNA-binding domain. In contrast, eubacterial and plant RecA proteins contain a conserved C-terminal domain that is absent in other members. The RecA C-terminal domains also bind to double-stranded DNA (23) and are similar in function to, but distinct in sequence from, the N-terminal domain of RAD51 and DMC1 (24, 25). Plant RecA proteins also contain a ~70-aa N-terminal region that is different from the RADA/RAD51 N-terminal regions. At least some plant RecA proteins contain organelle-targeting peptides (14, 15).

**Ancient Duplication Events in the recA/RAD51 Gene Family.** To investigate the evolutionary history of the recA/RAD5 family gene, we conducted phylogenetic analyses by using RecA/RAD51-like protein sequences from representative species of eubacteria, archaea, and eukaryotes whose genomes have been sequenced (see Materials and Methods). In this study, the neighbor-joining (NJ) and maximum likelihood (ML) methods were used to construct phylogenetic trees. These two methods gave essentially the same tree except for some minor details. Our results indicate that the recA/RAD5 family members can be divided into three major groups, designated as the recA, RADα, and RADβ subfamilies (Fig. 2A and Fig. 6, which is published as supporting information on the PNAS web site). The recA subfamily includes members from eubacteria, plants, and protists, whereas the RADα and RADβ subfamilies each contain genes from archaea and eukaryotes. In the RADα subfamily, RAD51 and DMC1 genes form two separate monophyletic groups, each of which contains plant, fungal, and animal genes. The archaeal RADα genes form a clade separate from the combined group of RAD51s and DMC1s. Therefore, it is likely that RAD51 and DMC1 genes were derived from a eukaryotic RAD4 gene by gene duplication before the divergence of plants from fungi and animals. Similarly, in the RADβ subfamily, each of the RAD51C, XRC3, RAD51B, RAD51D, and XRC2 genes form a monophyletic group that includes genes from plants and animals, whereas the archaeal RADβs form a separate basal clade. This topology is supported by multiple analyses (Figs. 6–10, which are published as supporting information on the PNAS web site) and suggests that these five eukaryotic RAD51-like genes were derived from a single ancestral RADβ gene by successive duplication events, all of which occurred before the divergence of plants from fungi and animals.

To better understand the evolutionary relationships of lineages within the RADα and RADβ subfamilies, additional phylogenetic analyses were performed by using only RADα and RADβ subfamily sequences. The phylogenetic tree (Fig. 2B) showed that RADα and RADβ genes form two separate groups (100% bootstrap support). The evolutionary relationships among the members of these two subfamilies are identical to those in the previous tree (Fig. 2A). However, the bootstrap supports for each clade were significantly improved in the RADα/RADβ tree (Fig. 2B). In the RADβ subfamily, the RAD51C group is the first to separate among five eukaryotic RADβ genes, followed by the XRC3C group. RAD51D and XRC2C genes formed two well-supported sister groups that seem to have emerged most recently among the eukaryotic RADβ genes.

**Accelerated Evolution of Some recA/RAD51-Like Genes.** Fig. 2 shows that the RADα subfamily genes are highly conserved and have evolved at a much lower rate than the RADβ subfamily genes, possibly reflecting the conserved functions of the RADα genes in recombination. In the RADβ subfamily, the XRC2C and RAD51D genes have evolved at a much higher rate than the genes in the other clades.

The tree topology for each group of genes was congruent with the species phylogeny except in the RAD51 lineage, which contains two genes from *C. elegans* and Caenorhabditis briggsae. When these genes were included in the phylogenetic analysis, they formed a basal clade outside the plant, animal, and fungal RAD51 genes (Fig. 3). This anomalous tree topology was apparently because the two worm genes have evolved very rapidly. Rapid evolution can be seen from the matrix of amino acid sequence identity for the ten representative species used (Table 2, which is published as supporting information on the PNAS web site). This table shows that the sequence identity is always low when the two worm genes are compared to those from the five eukaryotic groups (Table 2). The two *Drosophila* genes also evolved significantly faster than other non-worm genes according to the phylogenetic test of rate differences (results not shown) (26).

In addition to RAD51, *D. melanogaster* has four other RAD51-like genes, Spindle-B (Spn-B), Spindle-D (Spn-D), CG2412, and CG6318. It was suggested that Spn-B and Spn-D are related to RAD51C and XRC3C, respectively (27), and that these two genes have evolved significantly faster than their orthologs in vertebrate animals and plants, which is in agreement with our results. In addition, our analysis suggests that CG2412 and CG6318 are orthologous to RAD51D and XRC2C, respectively (Fig. 8). Moreover, a putative gene in *C. elegans* (NP.498799) was shown to belong to the RAD51D group (data not shown).

Because the budding yeast and fission yeast RAD55 and RAD57 genes are highly divergent, their evolutionary relationships with other RAD51 family members are difficult to determine. Our results suggest that RAD57 and RAD55 are orthologous to XRC3C and XRC2C, respectively (Fig. 9).

**Two Different Origins of Eukaryotic recA Genes.** As shown in Fig. 2A, two Arabidopsis recA genes, AtrecA1 and AtrecA3, cluster with eubacterial recA genes rather than with eukaryotic RAD51-like genes, suggesting that the plant recA genes might have evolved through a mechanism different from that of the RAD51-like genes. To examine the relationships of eukaryotic and eubacterial recA genes more closely, we conducted another phylogenetic analysis by using all available eukaryotic recA genes (Table 1). We also included recA
genes from 23 eubacterial species, which represent six main taxonomic groups. As expected, the recA genes from these six groups of eubacteria formed six distinct clades (Fig. 4). Interestingly, the eukaryotic recA genes formed two separate groups. One group, including all recA1 genes from plants, green algae, red algae, and brown algae, cluster with cyanobacteria recA with strong support. The other group, composed of recA2, recA3, and recA4 genes from flowering plants and the D. discoideum recA, grouped together with proteobacteria recAs. Our results suggest that the eukaryotic recA genes have two different origins: cyanobacteria and proteobacteria.

Discussion

Ancient Duplication of RAD51-Like Genes, Functional Divergence, and Origin of Meiosis. Our phylogenetic analysis indicates that eubacterial and eukaryotic recA genes form a single clade, whereas the remaining genes, referred to as RAD51-like genes, form two separate groups (RADα and RADβ), each of which contains both archaeal and eukaryotic members. If we accept the idea that eukaryotes and archaea shared a common ancestor, the RADα and RADβ groups are likely to have been generated by gene duplication that predated the divergence of archaea and eukaryotes. In addition, subsequent gene duplication events in early eukaryotes generated seven major groups that are maintained in both animals and plants. Gene duplication allows one copy to maintain the existing function and the other to gain a new function. Specifically, the genes in the RADα subfamily are important for homologous recombination and DNA repair, which are similar to the eubacterial recA functions (12, 13, 17, 28, 29), suggesting that these genes have maintained the original function.

However, duplications of RAD51-like genes are likely to have produced major functional innovations that are critical for the success of eukaryotes. In the RADα lineage, further gene duplication occurred before the divergence of eukaryotes and generated the RAD51 and DMC1 genes. RAD51 is important for a general function in homologous recombination during both somatic DNA repair and meiosis, whereas DMC1 acts exclusively during meiosis where gene function has been tested. In meiosis, recombination between homologous chromosomes is of central importance for the association and proper segregation of homologous chromosomes, and the function of DMC1 is likely to promote the recombination between homologous chromosomes, rather than sister chromatids (5, 30). Therefore, the “birth” of DMC1 might have directly contributed to the origin of meiosis and sexual reproduction in eukaryotes. DMC1 was found in several protists, including Giardia, which is among the earliest divergent protists (31), providing evidence that meiosis originated before the divergence of extant eukaryotes, as previously proposed (32), and asexuality among eukaryotes is a derived
insects, such as silkworm, have retained DMC1 and RAD51 chromosomal interactions are related to the rapid evolution of meiotic recombination and its relationship with other recombinational processes in turn may have influenced the genetic complexity and genome stability of eukaryotic organisms. Furthermore, the fact that members of multiple lineages, including the meiosis-specific DMC1 clade, are critical for meiosis and that members of these clades are found in protists, even the earliest-divergent ones, suggest that meiosis originated very early in the eukaryotic history.

**Acquisition of recA Genes by Eukaryotic Nuclear Genomes by Means of Endosymbiotic Gene Transfer.** It has been proposed that mitochondria and chloroplasts were incorporated into eukaryotic cells from proteobacteria and cyanobacteria progenitors, respectively, through endosymbiotic events (31, 40). In fact, many genes of these euubacterial origins have become eukaryotic nuclear genes, some of which encode proteins functioning in mitochondria and chloroplasts (40). Previously, a limited analysis of the Arabidopsis recA1 and recA2 genes suggested that both are most closely related to cyanobacteria (15). However, the extensive
analysis shown here strongly suggests that eukaryotic recA genes were derived from two different eubacterial origins, proteobacteria and cyanobacteria (Figs. 4 and 5). The N-terminal region of the Arabidopsis RecA1 protein contains a putative plastid transist peptide, and the protein was detected in the chloroplast (14). The close relationship of Arabidopsis recA1 and its plants and algae orthologs with cyanobacterial recAcs supports the idea that the chloroplast evolved from a cyanobacterion-like endosymbiont in the ancestors of photosynthetic eukaryotes (31).

In addition, the Arabidopsis RecA2 protein contains a predicted mitochondrial-targeting peptide and is localized to the mitochondrion (15). Besides plants, the animal-like protist D. discoideum also has a proteobacterial-like recA (Fig. 4). This finding supports the idea that a eukaryotic recA originated before the animal–plant split, although the possibility of horizontal gene transfer after the split cannot be ruled out. Moreover, Arabidopsis, poplar, rice, and maize contain three proteobacterial-like recA genes, which can be divided into two subgroups, the recA2 group and recA3/recA4 group (Fig. 4). This result suggests that a gene duplication event and subsequent functional divergence occurred on proteobacterial-like recA genes in the early history of flowering plants. These results are consistent with the hypothesis that the mitochondrion was derived from an endosymbiotic relative of proteobacteria (31, 40).

The presence of one recA in the animal-like protist D. discoideum, but the absence of a recA gene in animal and fungi genomes, suggests that recA might be lost in animals and fungi after the endosymbiotic origin of mitochondria. The genomes of animal and fungi mitochondrial are much smaller (~16 kb in animals and ~50 kb in fungi) and contain fewer genes than those of plant mitochondria and chloroplasts (several hundred kb) (41). Furthermore, animal mitochondrial DNA evolves much more rapidly (42–44) than plant mitochondrial and chloroplast genes (45, 46). It has been suggested that the acceleration of molecular evolution in the small genome of endosymbiotic bacteria, Buchnera, is mainly because of enhanced mutation rate (47). It was shown that the E. coli RecA protein functions in the chloroplast of the green algae Chlamydomonas to regulate recombination in a way similar to that in E. coli (48). Therefore, the existence of RecA-mediated recombination could be a major reason for the maintenance of large plant chloroplast and mitochondrion genomes, because the efficient DNA repair by homologous recombination would reduce the deleterious effects of mutations. Conversely, the mitochondrial recA might have been lost in the ancestor of animals and fungi. The loss of recA from animal and fungal genomes might have resulted in a reduction of the integrity and size of their mitochondrial genomes. Alternatively, mitochondrial genome reduction might have proceeded and allowed the loss of the mitochondrial recA gene in animals and fungi.

A Model for the Evolutionary History of the recA/RADS1 Gene Family.

On the basis of the results obtained here, we propose a plausible scenario of the evolutionary history of the recA/RADS1 gene family (Fig. 5). In this model, all recA/RADS1-like genes evolved from a single common ancestor by gene duplication, gene loss, and endosymbiotic gene transfer. The duplication of an ancient recA gene before the divergence of archaea and eukaryotes gave rise to two lineages of RADS1-like genes, RADα and RADβ, whereas the recA-like gene has been maintained as a single-copy gene in eubacteria (except for some species). In archaea, RADA and RADB are maintained as single-copy genes after the separation from eukaryotes, with a possible loss of RADB in some lineages. In the eukaryotic lineage, both RADα and RADβ experienced additional gene duplication events before the divergence of major eukaryotic groups. Gene duplication produced the RADS1 and DMCI genes from RADα, whereas, in the RADβ subfamily, it generated the RADS1C, XRCC3, RADS1B, RADS1D, and XRCC2 genes successively. DMCI was apparently
lost from some insect and nematode species. Some RADβ genes were also lost from several fungal and invertebrate lineages. The eukaryotic recA genes originated from proteobacteria and cyanobacteria. By 2006, the mitochon-ionic-derived recA gene experienced further duplications in flowering plants but was lost in the ancestor of animals and fungi. This model provides a basis for the functional conservation of homologous recombination.

Materials and Methods

Data Retrieval. Protein sequences of the E. coli recA and human RAD51, DMC1, RAD51C, RAD51B, RAD51D, XRCC2, and XRCC3 genes were retrieved from the National Center for Biotechnology Information (NCBI) database and were used as queries for gene search using BLAST, TBLASTN, and PSI-BLAST for recA/RAD51-like genes from NCBI databases, with e value 1e-5 as the cutoff. One hundred forty-five published or previously predicted sequences from representative organisms of eubacteria, archaee, and eukaryotes were selected (Table 1). Thirty-two recA/RAD51-like genes were predicted from genomic sequences based on sequence similarities (Table 1). The sequences are available upon request.

Sequence Alignment. Preliminary multiple sequence alignments were carried out by using CLUSTALX 1.8 (49) and MUSCLE V.3.52 (50). In the CLUSTALX alignment, we used BLOSUM series as the protein extension penalties. The default parameter setting was used in the CLUSTALX alignment. We used BLOSUM series as the protein functionally conserved homologous recombination.

Phylogenetic Analyses. We constructed NJ trees using MEGA 3.0 (51) and ML trees by using PHYML V.2.4 (53). The reliability of interior branches was assessed with 1,000 bootstrap resamplings by using “pairwise deletion option” of amino acid sequences with gamma parameters (unless otherwise indicated). Gamma parameter values were estimated by using PHYML software. ML analyses were performed by using PHYML with 1,000 bootstrap resamplings. Here the Jones, Taylor, and Thornton (JTT) model for amino acid sequences and gamma parameters were used. We did not use maximum-parsimony methods because this method tends to yield unreliable results when highly divergent sequences were included. Tree files were viewed by using MEGA. NJ trees are shown with bootstrap values for NJ and ML analyses (first and second values, respectively), unless otherwise indicated.

We thank L. Hao, B. Hedges, W. Hu, W. Li, W. Martin, A. Rooney, W. Pawlowski, Z. Yang, J. Klein, and L. Zahn for helpful comments and discussion. This work was supported by National Institutes of Health Grant GM63871–01 (to H.M.) and GM020293 (to M.N.) and National Natural Science Foundation of China Grant 30121003 (to Song Ge, Chinese Academy of Sciences).