JOURNAL OF BACTERIOLOGY, Aug. 2009, p. 5076–5084 Vol. 191, No. 16
0021-9193/09/$08.00+0 doi:10.1128/JB.00254-09

Phylogenetic Ubiquity and Shuffling of the Bacterial RecBCD and AddAB Recombination Complexes▽†

Gareth A. Cromie*

Division of Basic Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109-1024

Received 25 February 2009/Accepted 8 June 2009

RecBCD and AddAB are bacterial enzymes that share similar helicase and nuclease activities and initiate repair of DNA double-strand breaks by homologous recombination. Examination of the phylogenetic distribution of AddAB and RecBCD revealed that one or the other complex is present in most sequenced bacteria. In addition, horizontal gene transfer (HGT) events involving *addAB* and *recBCD* appear to be common, with the genes encoding one complex frequently replacing those encoding the other. HGT may also explain the unexpected identification of archaeal *addAB* genes. More than 85% of *addAB* and *recBCD* genes are clustered on the genome, suggesting operon structures. A few organisms, including the *Mycobacteria*, encode multiple copies of these complexes of either the same or mixed classes. The possibility that the enzymatic activities of the AddAB and RecBCD enzymes promote their horizontal transfer is discussed, and the distribution of AddAB/RecBCD is compared to that of the RecU/RuvC resolvases. Finally, it appears that two sequence motifs, the Walker A box involved in ATP binding and an iron-sulfur-cysteine cluster, are present only in subsets of AddB proteins, suggesting the existence of mechanistically distinct classes of AddB.

Homologous recombination is central to the repair of DNA damaged by single-strand (SS) and double-strand (DS) breaks and gaps. Such discontinuities can occur after exposure to exogenous agents such as ionizing radiation but also when DNA replication forks break and as intermediates in DNA repair processes. Homologous recombination can also rearrange genetic information, making it important in processes such as phase variation of bacterial outer membrane proteins (for an example, see reference 2).

Homologous recombination is highly conserved among viruses, bacteria, archaea, and eukaryotes. There are three recognizable stages (for reviews, see references 25 and 35). In the first stage, presynapsis, the DNA substrate is processed to give a SS region coated with strand exchange protein(s). In bacteria, this protein is RecA. In the second stage, synapsis, the protein-DNA complex pairs with its complementary homologous DNA target, displacing the other strand at the target site and forming a joint molecule. In the final stage, postsynapsis, DNA replication fills in any gaps in the joint molecule, which is then resolved to give recombinant DNA products.

In bacteria, there appear to be two different presynaptic pathways that use either the related AddAB or RecBCD holoenzymes (Fig. 1) or, alternatively, the RecFOR proteins. RecFOR proteins appear to operate on SS gaps to ensure RecA is loaded there, whereas RecBCD and AddAB act on DNA DS breaks (DSBs), processing them to yield SS 3′ DNA ends coated with RecA (3, 43).

The RecBCD complex has most thoroughly been studied in *Escherichia coli*, a member of the *Gamma* class of the phylum *Proteobacteria*. RecB and RecD are superfamily I helicases, and RecB is also a nuclease. After the RecBCD complex loads onto a DS DNA end, the RecB and RecD helicases separate the strands, with the slower RecB helicase traveling on one strand in the 3′-to-5′ direction and the faster RecD helicase traveling on the other strand from 5′ to 3′ (38). During translocation, *E. coli* RecBCD recognizes a specific DNA sequence, Chi (5′-GCTGGTGG-3′), perhaps via the RecC protein, which appears to be an inactivated helicase (Fig. 1) (32, 34). Chi is a recombination hot spot that switches the complex to a recombinogenic mode, producing a 3′ SS end at the Chi site via the nuclease activity of RecB (for reviews, see references 25 and 35). In addition to producing a 3′ SS "tail" needed for recombination, the RecBCD complex actively loads the RecA strand exchange protein onto this DNA (5). In *E. coli*, the RecBCD complex is responsible for essentially all recombinational repair of DSBs.

Bacillus subtilis and some other bacteria lack RecBCD but possess a different complex with many of the same activities. This complex is composed of the AddA and AddB proteins (22). Like RecB, AddA is a superfamily I helicase and a nuclease (18, 21), organized with the same domain structure as RecB (Fig. 1). Similar to RecC, AddB appears to be an inactivated helicase, but it also possesses a nuclease domain similar to those of RecB and AddA (Fig. 1) (2, 11, 43). Like RecBCD, AddAB is an ATP-dependent helicase that acts as a nuclease in conjunction with its helicase activity, with the AddA nuclease acting on the 3′-end strand and the AddB nuclease on the 5′-end strand (43). Also like RecBCD, AddAB recognizes Chi-like control sequences, for instance, 5′-GCGCGTG-3′ in *Lactococcus lactis* and 5′-AGCGG-3′ in *B. subtilis* (7, 12). The AddAB complex is required for recombinational repair of DSBs (1, 2, 44), but it is not known if, like RecBCD, it actively loads RecA.

Both RecBCD and AddAB are important for bacterial pathogenicity. *addAB* mutations reduce the ability of *Helicobacter pylori* to colonize mouse stomachs, and *recBC* mutants of

* Mailing address: Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, P.O. Box 19024, Seattle, WA 98109. Phone: (206) 667-4431. Fax: (206) 667-6497. E-mail: gcromie@fhcrc.org.

† Supplemental material for this article may be found at http://jb.asm.org/.

▽ Published ahead of print on 19 June 2009.

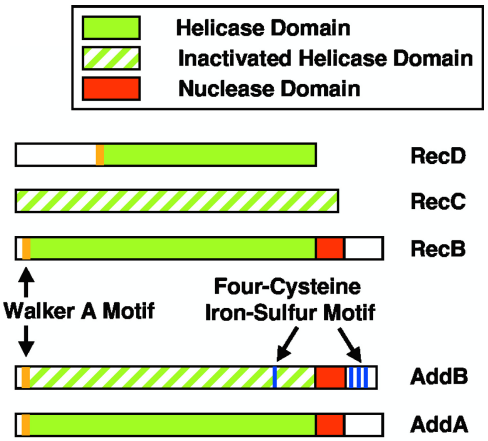


FIG. 1. Structure of the RecBCD and AddAB proteins. The RecB, RecD, and AddA proteins include a helicase domain (solid green) with the canonical six-helicase motifs of helicase superfamily I. The most N-terminal of these motifs is the Walker A box (orange). The RecC and AddB proteins include an inactivated helicase domain (striped green). The RecB, AddA, and AddB proteins additionally possess similar short nuclease domains toward their C termini (red). Some, but not all, AddB proteins possess an N-terminal Walker A and/or a mostly C-terminal iron-sulfur motif made up of four cysteines (blue).

Salmonella enterica serovar Typhimurium are severely compromised for infection and killing (2, 8).

Using a strict set of criteria, Rocha et al. (30) examined the distribution across bacteria of the RecBCD and AddAB enzymes, along with other recombination and DNA repair proteins. They determined that the AddAB proteins are ubiquitous in some taxa and the RecBCD proteins in others, while yet other taxa have examples of both complexes in different species. In addition, they classified several species as lacking both the AddAB and RecBCD complexes. Recently, *addAB* genes from *Epsilonproteobacteria* were identified (2, 27, 40). Previously, all examined members of this proteobacterial class had been classified by Rocha et al. as lacking every component of the RecBCD and AddAB systems. This finding suggests that the stringent criteria used by Rocha et al. (30) might have led them to miss other examples of AddAB and RecBCD from sequenced bacteria. On this basis, and to gain a fuller understanding of the importance of the AddAB and RecBCD enzymes, their phylogenetic distribution was reexamined.

MATERIALS AND METHODS

Identification of RecBCD and AddAB proteins. RecB, AddA, and AddB all possess long helicase or inactivated helicase domains and short nuclease domains. The helicase domains are similar to those found in several other widely distributed proteins, including UvrD and Rep. However, the nuclease domains found in RecB, AddA, and AddB, which are essential for the action of these recombination proteins, are absent from the related helicases. Any proteins scoring highly in alignments with RecB, AddA, or AddB that lack a nuclease domain are therefore very unlikely to be functional analogs of the recombination proteins. These facts were used as the basis of a two-step identification process designed to avoid collecting and aligning large numbers of other helicase proteins, such as Rep, that clearly lack the nuclease domains of RecB, AddA, and AddB. Analogous two-step processes were used to identify RecC and RecD1 sequences (RecD1 is the variant of RecD associated with RecB and RecC; RecD2 is not [30]).

Position-specific scoring matrices (PSSMs) were generated for RecB, RecC, and RecD1 starting with the *E. coli* protein sequences, for AddA and AddB starting with the *B. subtilis* sequences, and for RecD2 starting with the *Deinococcus radiodurans* RecD2 sequence (41). In each case, PSI-BLAST (http://blast.ncbi.nlm.nih.gov) iterations versus all predicted proteins from the bacterial domain were carried out using the "nr" database with default parameters in order to identify the highest-scoring significant hit (E value, <0.05) from each major taxonomic group shown in Fig. 2. Hits were validated as true homologs based on their annotations and apparent similarity of operon structure with other genes of the *recBCD* or *addAB* systems and by manual inspection and comparison of their alignments (using MAFFT version 5 [20] via the Jalview software package [42], which uses the "—auto" option for MAFFT) to the original *E. coli*, *B. subtilis*, or *D. radiodurans* sequences. No more than one sequence from each taxon was used in each iteration for construction of the PSSM. Iterations were stopped when no new taxa provided validated hits, and the final PSSMs were saved. The core sequences used to generate these PSSMs are shown in Table S1 in the supplemental material. This approach was used to provide greater sensitivity than is possible using BLAST with single sequences, while allowing a manageable number of core sequences to be validated for the PSSM.

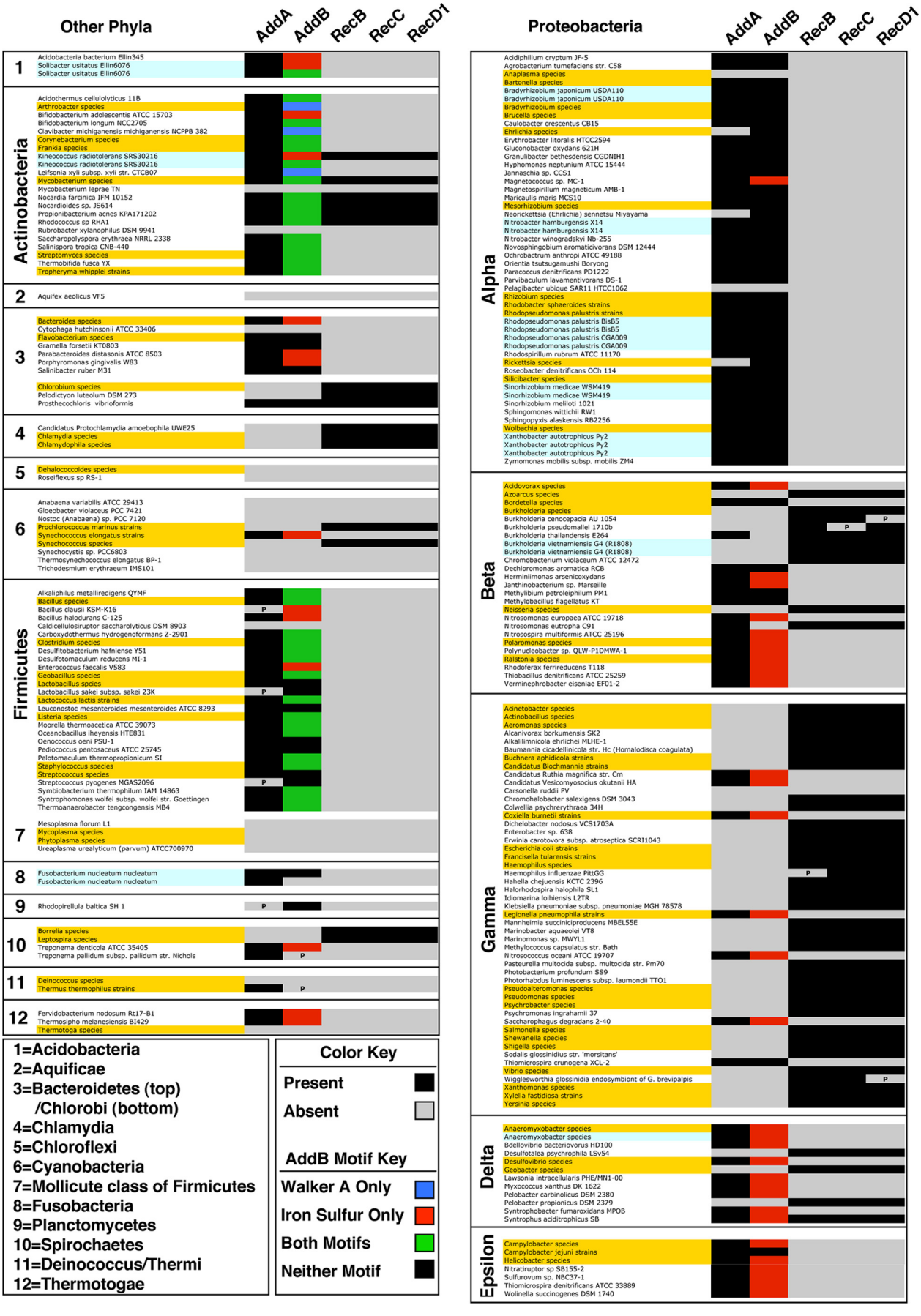
The *E. coli* and *B. subtilis* RecBC and AddAB protein sequences were then used, along with the corresponding core PSSMs, to perform a PSI-BLAST search of each major bacterial taxon in turn. This was done against the "nr" database of predicted protein sequences, using default parameters. All hits with E values of <0.05, coming from the 513 fully sequenced organisms considered in this study, were taken as putative homologs of the query sequence, providing they were $>50\%$ of the length of the query sequence. This additional threshold should identify only those proteins that align to at least a substantial fraction of the long helicase/inactivated helicase domains. To be on the final list of accepted RecC sequences, the alignment of the sequences to that of *E. coli* RecC was required to cover $>40\%$ of the length of the *E. coli* RecC protein.

To get the final list of accepted RecB and AddAB sequences, the proteins from the PSI-BLAST sweeps were filtered for nuclease domains by scoring against RecB, AddA, and AddB nuclease hidden Markov models (HMMs). Proteins that scored significantly ($E < 0.05$) against the matching nuclease HMM and less well against the other two nuclease HMMs were accepted as homologs. The nuclease HMMs were constructed using the HMMer program (version 2.2; http://hmmer.janelia.org/) from the set of core AddA, AddB, and RecB sequences that were first aligned using MAFFT and were then edited to include only the nuclease regions of the alignments (for *E. coli* RecB, amino acids 1060 to 1114; for *B. subtilis* AddA, amino acids 1156 to 1204; for *B. subtilis* AddB, amino acids 941 to 985). HMMer was also used to score each putative sequence against each HMM.

In a small number of cases, species were observed where only AddA or AddB passed the HMM nuclease filter. In these cases, the most significant hits from the initial PSI-BLAST search for the missing class were reexamined. These proteins were aligned, using MAFFT, to the full set of proteins of the class in question. Good quality alignment, particularly to the nuclease domain, was then accepted instead of the HMM nuclease step in assigning homology. Similarly, in a small number of species, two AddA proteins were identified but no AddB was identified, or vice versa. In these cases, the relative (significant) scores of the two proteins against the AddA and AddB PSSMs and alignment to each full set of AddA and AddB proteins were used to assign protein identity instead of the HMM nuclease step. Generally, the annotation, protein length, and chromosomal position of the relevant genes (e.g., the newly assigned AddB protein is encoded by the gene adjacent to *addA*) supported the reassignments.

Detection of the RecD component of the RecBCD holoenzyme was complicated by the existence of two distinct families of RecD proteins, one associated with RecBC ("RecD1") and one not ("RecD2"). The N-terminal regions of RecD1 and RecD2 sequences differ (30). Again, a two-step process was used to identify RecD1 sequences, starting with low stringency RecD1 PSI-BLAST followed by scoring against RecD1 and RecD2 HMMs.

To identify putative RecD1 proteins, the *E. coli* RecD1 sequence was used along with the RecD1 PSSM for a PSI-BLAST search against each major bacterial taxon in turn. For each organism possessing RecBC sequences, as determined above, the highest scoring significant ($E < 0.05$) match to RecD1, where the alignment with *E. coli* RecD1 was $>40\%$ of the length of the *E. coli* RecD1 sequence, was taken as a putative RecD1. To get the final list of accepted RecD1 proteins, putative RecD1 sequences were scored against RecD1 and RecD2 HMMs. These HMMs were generated, using HMMer, from full-length alignments of the RecD1 and RecD2 core sets of proteins constructed using MAFFT. Sequences were accepted as RecD1 proteins if they scored significantly ($E < 0.05$) against the RecD1 HMM and less well against the RecD2 HMM. If the highest-scoring RecD1 PSI-BLAST match from an organism failed the HMM test, then the next highest significant PSI-BLAST match was tested. The final list of RecD1 sequences was well supported by the apparent operon organization



(chromosomal clustering) of the corresponding genes with those encoding RecB and RecC.

For the *Chlamydiae* and some of the *Cyanobacteria*, the HMM step of the RecD1 assignment worked poorly. Therefore, the HMM step was replaced by manual examination and comparison of the alignments of putative RecD1 sequences to the RecD1 and RecD2 core alignments. Only sequences aligning well to those of RecD1 and failing to align to the diagnostic N-terminal extension of the RecD2 sequences (30) were accepted as RecD1.

To identify archaeal AddA and AddB sequences, PSSMs were constructed, using the archaeal sequences that passed the bacterial AddA and AddB PSI-BLAST and nuclease HMM tests. The highest-scoring sequences, AddA from *Halobacterium salinarum* and AddB from *Methanoculleus marisnigri* JR1, were then used in an archaeal PSI-BLAST search along with the corresponding PSSMs. Sequences scoring significantly in this test ($E < 0.05$) were then scored against the bacterial nuclease HMMs, as above.

Identification of RecU and RuvC proteins. To identify RecU and RuvC sequences, the *E. coli* K12 RuvC and the *B. subtilis* RecU sequences were used for BLAST (http://blast.ncbi.nlm.nih.gov) searches against the "nr" database of predicted protein sequences restricted to the bacterial domain. Default parameters were used, and all hits with E values of <0.05 were retrieved, provided they came from one of the 513 fully sequenced genomes considered here. These proteins were then aligned, using MAFFT, as above. The putative RuvC and RecU proteins had consistent sizes, with each possessing a single highly conserved domain covering almost the full length of each protein sequence. In more than 95% of cases, no more than a single protein was identified in each bacterial genome.

Phylogenetic trees. Phylogenetic trees were constructed from alignments of each full set of MAFFT-aligned sequences. For AddAB and RecBCD, the sequences of the individual subunits were concatenated before alignment to improve the phylogenetic signal. In cases where more than one holoenzyme was encoded in a single genome, gene organization was used to determine which sequences were concatenated. All of these alignments had average pairwise amino acid identities greater than 20%, sufficiently high for the generation of reliable phylogenetic trees (39). Trees were constructed using PhyML (17), using the JTT (Jones-Taylor-Thornton) matrix model for multiple amino acid substitutions and an aLRT (approximate likelihood ratio test) SH-like branch support to provide branch reliability estimates. The resulting trees were analyzed using the MEGA software package (version 4.0) (37).

RESULTS AND DISCUSSION

The AddAB and RecBCD enzymes are very widely distributed across the bacterial domain. AddAB and RecBCD proteins were identified as described in Materials and Methods. The distribution of the two protein complexes across 513 sequenced bacterial genomes is shown in Fig. 2 and also in Fig. S1 in the supplemental material. These proteins are found in at least some organisms of every major taxonomic group examined except for the phyla *Aquificae* and *Chloroflexi* and the *Mollicutes* class of phylum *Firmicutes*. Among the 513 organisms examined, *recBCD* and/or *addAB* genes are present in 474 (92%), *addAB* genes are found in 292, *recBCD* genes are found in 206, and genes encoding proteins from both complexes are found in 24 (see below). The frequent occurrence of either the AddAB or RecBCD complexes in bacteria but the infrequent occurrence of both complexes in the same species supports the idea that these are alternative protein machines carrying out essentially the same important function.

In the analysis above, when any gene encoding a component of a complex was identified, genes encoding all of the other components were also almost always identified (>90%). There are 292 organisms that possess at least one of the AddAB proteins and 206 organisms that possess at least one of the RecBCD proteins. Among these, both the *addA* and *addB* genes are present in 270 of 292 organisms (92%), and equal numbers of *recB*, *recC*, and *recD* genes were observed in 200 of 206 organisms (97%). In many cases, the "missing" gene could be identified by examination of the chromosomal region around its partner (Fig. 2; see also Fig. S1 in the supplemental material). Along with the apparent operon structure of the *recBCD* and *addAB* genes (below), these high cooccurrence frequencies suggest that the methods for detecting RecBCD and AddAB proteins used here are both sensitive and discriminating.

Thirty-nine of the organisms examined here appear to lack every component of the AddAB and RecBCD protein complexes. Of these, 17 are members of the *Mollicutes* class of the phylum *Firmicutes*, whose members lack a cell wall and primarily exist as intracellular parasites. Isolated examples of species apparently lacking all of these proteins are also found in several other taxa, and several of these isolated examples (e.g., the gammaproteobacterium *Carsonella ruddii*, the alphaproteobacterial *Anaplasma* spp., and the actinobacterium *Mycobacterium leprae*) are also obligate intracellular symbionts or pathogens with small genomes. It has previously been noted that such organisms lack a number of recombination proteins, perhaps due to the extremely stable and sequestered environments in which they exist (30, 33, 36). However, as previously noted, while the intracellular gammaproteobacteria *Buchnera* and "*Candidatus* Blochmannia" appear to lack RecA and other recombination proteins, they possess RecBCD (16, 30, 36). This finding suggests that their RecBCD complexes act solely as helicase/nucleases and not as recombinases. A cell viability role independent of recombination is suggested for RecBCD by the lower viability of *recB* and *recC* null mutants than that of *recA* mutants in *E. coli* (9), perhaps due to RecBCD nuclease suppression of harmful rolling-circle replication.

"Shuffling" of the AddAB and RecBCD complexes. The distribution of the AddAB and RecBCD proteins is difficult to reconcile with a simple vertical transmission model. When a recent phylogenetic tree of bacterial species (13) is examined, both AddAB and RecBCD proteins occur in many lineages, but very rarely in the same organism (Fig. 2 and 3). Although the deeper branching order and the position of the root of the bacterial tree (Fig. 3) are still under debate (for an example,

FIG. 2. Wide distribution of the AddAB and RecBCD proteins across 513 fully sequenced bacterial genomes. The 513 fully genome-sequenced bacteria considered here were split into 18 major taxa. Closely related strains or species have been grouped (yellow background), and species with multiple copies of *addAB* or *recBCD* genes are indicated by an additional line (pale blue background). The presence of genes encoding AddA, RecB, RecC, and RecD1 and identified by my methods is indicated by a black box. Boxes for AddB indicate the presence of both the Walker A and iron-sulfur motifs (green), neither motif (black), the Walker A box only (blue), or the iron-sulfur motif only (red). In all cases, a gray box indicates absence of the corresponding gene. Note that *Magnetococcus* species strain MC-1, here grouped with the *Alphaproteobacteria*, is now designated as an "unclassified proteobacterium." Examples for which no protein meets the criteria used here but for which a potential homolog is identified by other means are shown as gray boxes marked "P" (see Fig. S1 in the supplemental material).

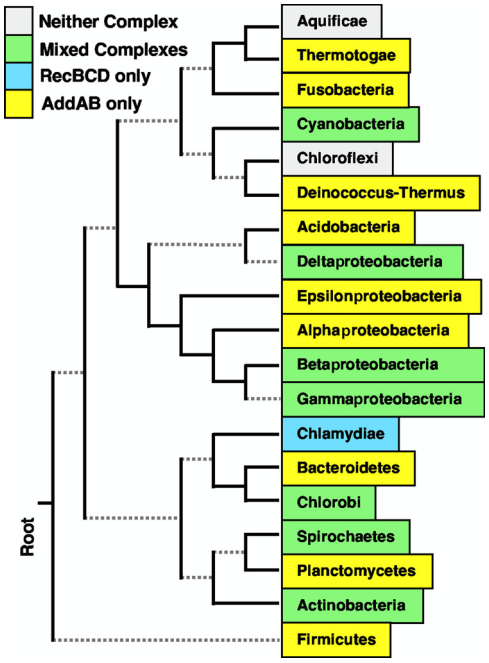


FIG. 3. Distribution of AddAB and RecBCD across the bacterial tree argues for multiple displacements of one complex by the other via horizontal transfer of genes. A cladogram (from reference 13) of the bacterial taxa from Fig. 2 is color coded to indicate the occurrence of the AddAB and RecBCD complexes. When both complexes are found in a taxon (green boxes), they can occur within the same, or different, species. Branches supported by bootstrap values of less than 80% are indicated by dashed lines.

see reference 10, 13), each complex often occurs in different species within clades that are well supported by 16S rRNA (http://www.arb-silva.de/) and "core gene" trees (for an example, see reference 13). This distribution strongly implies multiple horizontal gene transfer (HGT) events, which are known to occur at high frequencies in bacteria (reviewed in reference 24). However, the situation with AddAB and RecBCD is unusual in that it is an example of nonorthologous gene displacement (23) by HGT, with acquisition of the genes encoding one protein complex and loss of those encoding the other rather than, for example, simple replacement of *addAB* genes with foreign *addAB* sequences. This process results in phylogenetic "shuffling" of the AddAB and RecBCD complexes. The usual chromosomal clustering of *recBCD* and *addAB* genes (see below) would facilitate HGT, as single modest segments of DNA could carry the genes encoding each full complex (26).

Mechanistically, loss of the genes encoding one enzyme (e.g., AddAB) and acquisition of those encoding the other enzyme (e.g., RecBCD) might occur in either order. Selective pressure on bacterial genome size could result in the loss of one complex after the acquisition of an additional, essentially redundant, set of genes. Alternatively, loss of active AddAB or RecBCD function (by gene deletion or mutation) could precede the acquisition of the genes encoding the other complex and might in fact promote this acquisition (see below).

Multiple AddAB and RecBCD complexes are encoded within single genomes. The RecBCD and AddAB protein complexes appear to carry out essentially the same range of recombination and DNA helicase and nuclease tasks. It was therefore surprising to find several organisms whose genomes encode both sets of proteins or include multiple genes encoding the same class of proteins (Fig. 2). The *Mycobacterium* spp. and some other species of *Actinobacteria* possess both AddAB and RecBCD, as do the chlorobium *Prosthecochloris vibrioformis* and the deltaproteobacterium *Syntrophus aciditrophicus* SB. In addition, several genomes possess multiple copies of the *addAB* genes; for instance, the alphaproteobacterium *Xanthobacter autotrophicus* Py2 possesses three such pairs of genes. Only one of the genomes studied, that of the betaproteobacterium *Burkholderia vietnamiensis* G4, encodes multiple copies of the RecBCD complex. The organisms in which a single genome encodes multiple AddAB, or both AddAB and RecBCD, complexes appear to have received the "additional" sets of genes by HGT (see below). It is not clear what the functions of the apparently redundant individual AddAB and RecBCD protein complexes are in these organisms.

Clustering of *addAB* genes and of *recBCD* genes. The clustering, relative order, and orientation of the *recBCD* genes are highly conserved and suggest an operon structure. In *E. coli*, the *recBCD* genes are all clustered together on the same strand in the order C[X]BD, where X is the gene encoding the PtrA peptidase. It appears that in *E. coli* these genes are not organized into a single operon (4). However, the *ptrA* gene is associated with the *recBCD* genes only in organisms closely related to *E. coli*, such as *Salmonella* and *Yersinia* species. The most common organization of the *recBCD* genes (where all three genes are identified) is as a cluster on one strand with the order CBD and no interrupting genes (114 of 201; 57%), the next most common organization being C[X]BD (22%; mostly *E. coli* and related species) (see Fig. S1 in the supplemental material). If a *recBCD* gene cluster is defined more broadly as having the relative order CBD occurring on the same strand and with no more than three genes (including *recB*) between *recC* and *recD*, then the *recBCD* genes are clustered 86% of the time. An operon structure would allow coregulation of RecB, RecC, and RecD expression. Insertion of the *ptrA* gene in *E. coli* and related organisms may have split an ancestral *recCBD* operon.

In *B. subtilis*, the *addAB* genes form an operon, with the genes adjacent on one strand and transcribed in the order *addBA* (22). Like the organization of the *recBCD* genes, this organization is highly conserved, with 241 of 282 (85%) *addAB* gene pairs having the same organization as in *B. subtilis* (see Fig. S1 in the supplemental material). Among members of the phyla *Thermotogae* and *Bacteroidetes* and among some *Deltaproteobacteria*, the *addA* and *addB* genes are still clustered and transcribed in the same direction, but their order is reversed (i.e., *addAB*). Using the criterion that *addA* and *addB* genes are clustered (in either relative order) if they occur on one strand and separated by no more than three genes, *addAB* pairs are clustered 91% of the time.

AddAB proteins are found in a few archaea, but not in eukaryotes. To my knowledge, there are no reports of RecBCD or AddAB proteins occurring in organisms other than bacteria. However, a PSI-BLAST search using the *B. subtilis* AddA and AddB sequences and corresponding bacterial PSSMs identified a small number of archaeal sequences that also passed the nuclease HMM scoring step. No eukary-

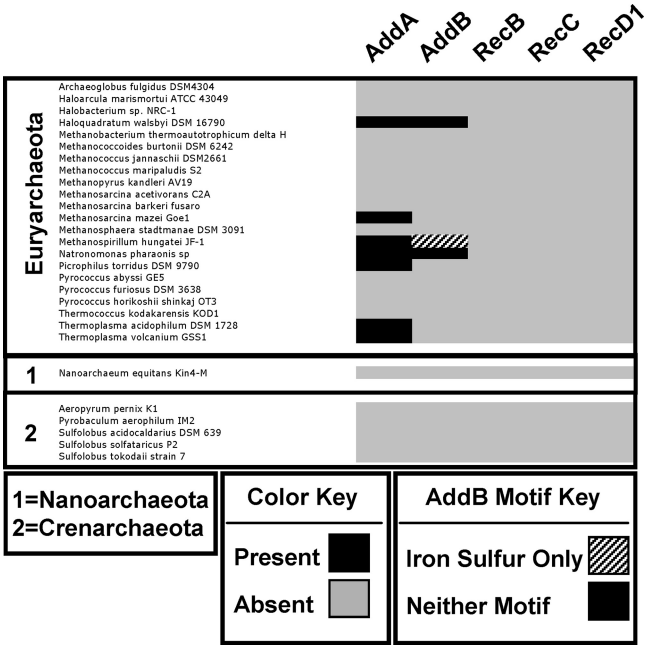


FIG. 4. Distribution of AddAB in 28 sequenced archaea. The 28 fully genome-sequenced archaeal species considered here were split into three major taxa. The presence or absence of genes encoding AddA, AddB, RecB, RecC, and RecD1, determined by my methods, is indicated. Note that only AddB sequences with the iron-sulfur motif alone or sequences lacking both the Walker A and iron-sulfur motifs are shown.

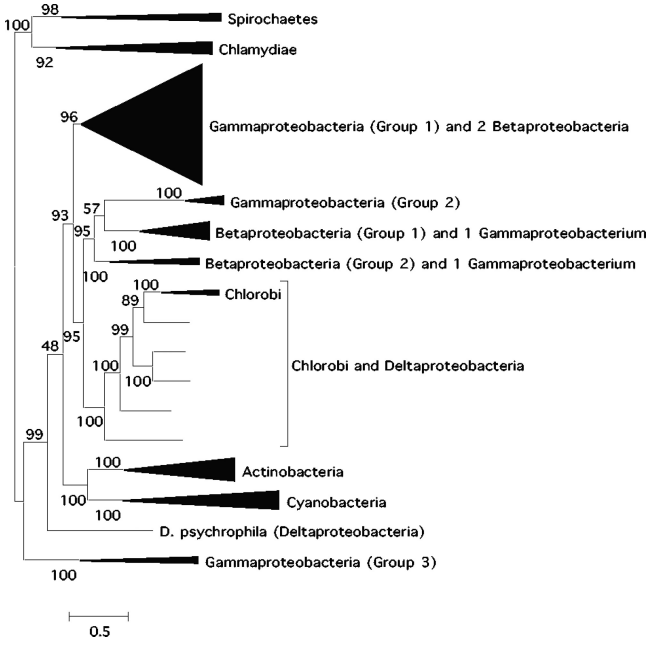


FIG. 5. Evidence for HGT in the RecBCD phylogenetic tree. An unrooted maximum likelihood tree was produced for the bacterial RecBCD complexes, with individual subunits concatenated to form single sequences for each complex. aLRT SH-like branch support values are shown. The scale bar indicates the number of amino acid substitutions per site.

otic sequences were identified. The archaeal sequences were then used to construct PSSMs and carry out PSI-BLAST searches against the archaeal domain, followed by scoring against the bacterial nuclease HMMs to assign homology (see Materials and Methods).

Among 28 fully sequenced archaea, AddAB protein pairs are restricted to three organisms (Fig. 4). In all three, the *addA* and *addB* genes are adjacent and on the same strand in the order *addBA*, the order most common in bacteria (see above). A small number of additional species possess a putative AddA, but no AddB, protein. The limited distribution of the archaeal AddAB proteins is consistent with HGT of the *addAB* genes into these organisms from bacteria. HGT between these domains has been observed before (6, 14, 28).

No archaeal sequences passed the criteria for assignment as RecB or RecC orthologs. For each of the RecB and RecC proteins, only one eukaryotic sequence (from *Oryzae sativa*) passed the scoring thresholds. These sequences were very similar (>80% identical) to RecBC sequences from the gammaproteobacterium *Serratia proteamaculans* and may represent bacterial contamination in sequencing.

Phylogenetic trees of AddAB and RecBCD sequences suggest multiple horizontal transfers of *addAB* and *recBCD* genes. To further examine the evolutionary relationships among the AddAB and RecBCD proteins, maximum likelihood phylogenetic trees were generated, using alignments of the full set of concatenated AddA and AddB sequences and concatenated RecB, RecC, and RecD1 sequences (see Materials and Methods). The interacting nature of the protein subunits and their genomic clustering strongly suggest that each full set of genes is inherited as a unit, so that concatenated sequences can be used to maximize the information available for tree construction.

Examination of the RecBCD tree shows that sequences from the phyla *Actinobacteria*, *Chlamydiae*, *Chlorobi*, *Cyanobacteria*, and *Spirochaetes* each cluster to form monophyletic groups, with high levels of associated branch confidence (Fig. 5; see also Fig. S2 in the supplemental material). These patterns are consistent with vertical transmission of the *recBCD* genes within these taxa or with intrataxon HGT. However, sequences from the *Proteobacteria* do not form a monophyletic group, indicating HGT.

The AddAB tree shows that among phyla where more than two organisms possess AddAB sequences, these sequences cluster to form monophyletic groups only among the *Actinobacteria*. Sequences from the *Bacteroidetes*, *Firmicutes*, and *Proteobacteria* clearly do not form monophyletic groups (Fig. 6; see also Fig. S3 in the supplemental material).

In organisms where more than one set of *addAB* genes are present on the genome, only one protein sequence clusters with the AddAB sequences from other members of the same taxon. This is also true for RecBCD sequences from the single organism found to have two sets of *recBCD* genes, the betaproteobacterium *B. vietnamiensis* G4. The "secondary" AddAB sequences are highly diverged from the first sequences and fail to form monophyletic groups with them (Fig. 6; see also Fig. S3 in the supplemental material). This suggests that these secondary enzyme complexes are the result of HGT events of *addAB* sequences between diverse species and not of duplications of

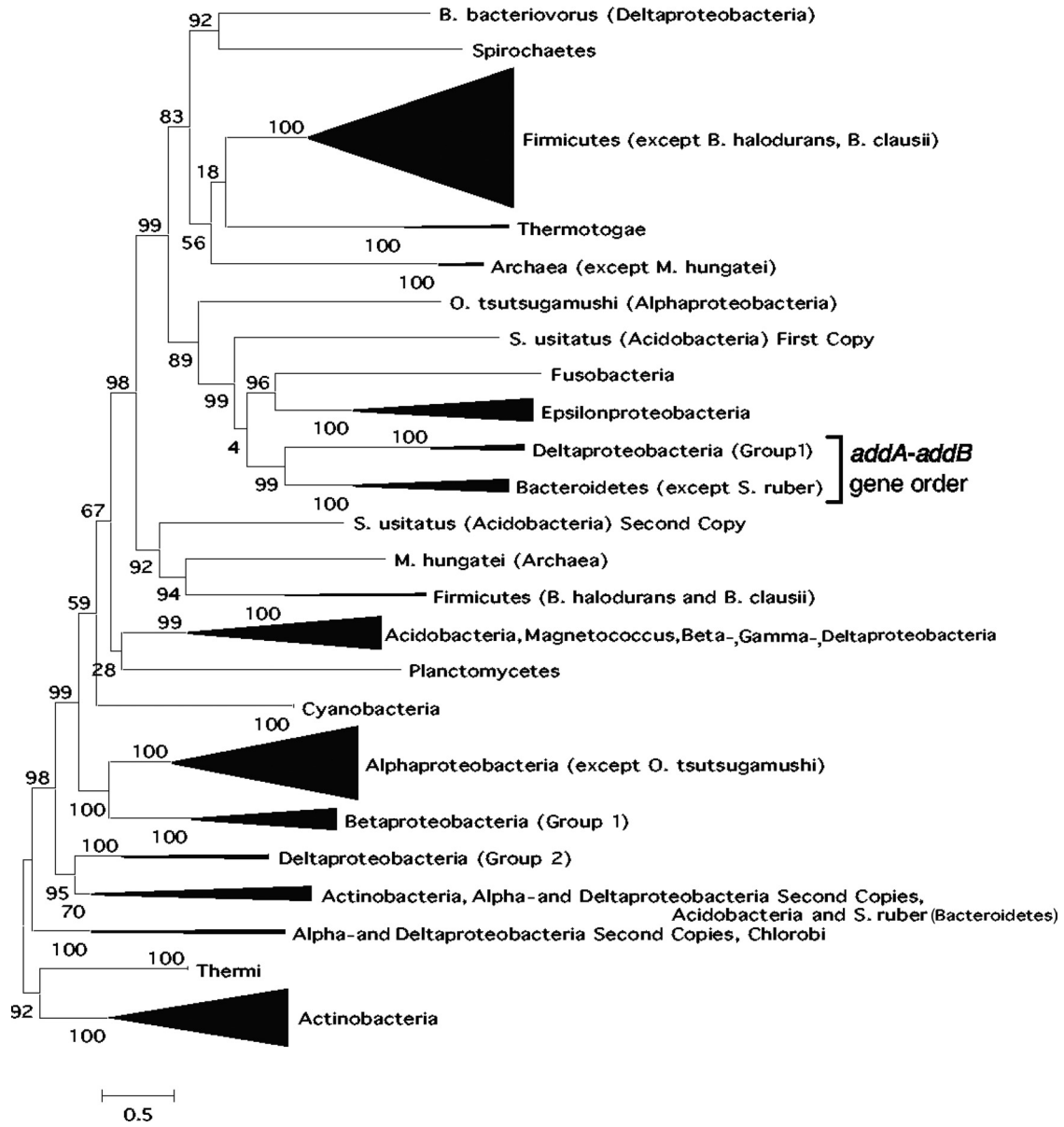


FIG. 6. Evidence for HGT in the AddAB phylogenetic tree. An unrooted maximum likelihood tree was produced for the bacterial and archaeal AddAB complexes, with individual subunits concatenated to form single sequences for each complex. aLRT SH-like branch support values are shown. The scale bar indicates the number of amino acid substitutions per site.

the existing "primary" *addAB* genes. The same conclusion can be drawn for the two sets of *recBCD* sequences from *B. vietnamiensis* G4 (Fig. 5; see also Fig. S2 in the supplemental material).

Both AddAB and RecBCD sequences occur within each class in the *Beta-*, *Gamma-*, and *Deltaproteobacteria*, suggesting HGT shuffling of the two complexes (Fig. 2 and 3). In addition, both the AddAB and the RecBCD sequences from each of these taxa fail to form monophyletic groups, suggesting that HGT within each protein class, as well as between classes, is particularly common in these proteobacterial groups (Fig. 5 and 6; see also Fig. S2 and S3 in the supplemental material). For example, one of the two main groupings of AddA and AddB sequences from *Deltaproteobacteria* come from organisms where the gene order is *addA-addB* ("AB") rather than the more common "BA" (associated with the other deltaproteobacterial sequences). The deltaproteobacterial "AB" group sequences are most closely related to those from the majority of the *Bacteroidetes*, which share the unusual "AB" gene organization (Fig. 6; see also Fig. S1 and S3 in the supplemental material). This strongly supports a more common ancestor for the AddA and AddB sequences from these *Bacteroidetes* and the "AB-ordered" *Deltaproteobacteria* than for the AddA and AddB sequences from the two groups of *Deltaproteobacteria*. Similarly, the phylogenetic tree of RecBCD sequences suggests that *recBCD* genes from *Gammaproteobacteria* have undergone HGT into two *Betaproteobacteria*, and *recBCD* genes from *Betaproteobacteria* have undergone HGT into two *Gam-*

maproteobacteria (Fig. 5; see also Fig. S2 in the supplemental material).

The phylogenetic trees of AddA and of AddB sequences are also inconsistent with a monophyletic group of archaeal AddAB sequences (Fig. 6; see also Fig. S3 in the supplemental material). Instead, the trees suggest HGT of bacterial *addAB* genes into the archaea on more than one occasion.

Distribution of the Walker A box and iron-sulfur cluster in AddB proteins. The AddB protein appears to be an inactivated helicase (with a nuclease domain), and AddB sequences from *Epsilonproteobacteria* lack the Walker A motif needed for ATP binding and helicase activity in AddA, RecB, RecD, and other active helicases (2). In contrast, the *B. subtilis* AddB protein does contain the Walker A motif, located at the extreme N terminus of the protein (22), as it is in AddA and RecB (Fig. 1). Given these two different observations, all AddB sequences were classified either as possessing or lacking the Walker A sequence, based on whether they matched, at their N termini, (minimally) three of the four defined positions in the conserved GXXXXGK[T/S] motif (http://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi?uid=108546).

The majority of AddB sequences lack the Walker A motif (Fig. 2), which is limited to the phyla *Actinobacteria*, *Firmicutes*, and *Acidobacteria*. The *Firmicutes* are split into two large groups, with *Bacillus*, *Clostridium*, *Staphylococcus*, and other species having the Walker A box and *Streptococcus*, *Lactobacillus*, and other species lacking it.

In addition to a Walker A motif, the *B. subtilis* AddB protein contains a four-cysteine motif believed to form an iron-sulfur cluster, with three of these four cysteines clustered in the C terminus of the protein as the motif CXXCXXXXXC (M. Dillingham, personal communication) (Fig. 1). Such iron-sulfur motifs coordinate iron atoms through binding of the metal to the sulfur present in the cysteine residues, often to promote electron transfer or catalysis. Consistent with the presence of the iron-sulfur motif, AddB from *B. subtilis* binds iron atoms (Dillingham, personal communication). The motif, defined as conservation of (minimally) three of the four cysteine residues, is found in at least some AddB sequences from most major taxa, with the *Alphaproteobacteria* being a notable exception (Fig. 2).

AddB sequences essentially fall into three classes: those with both iron-sulfur and Walker A motifs, those with neither, and those with the iron-sulfur motif only. Only four organisms, all *Actinobacteria*, have AddB proteins with the Walker A motif and no iron-sulfur cluster (Fig. 2).

What are the mechanistic roles of these two motifs? The role of the Walker A motif in AddB proteins is unclear. In *B. subtilis*, the iron-sulfur cluster is essential for DNA end binding (Dillingham, personal communication), but AddB proteins from organisms that lack the iron-sulfur motif must still be proficient in this activity. Therefore, it appears that there are mechanistically distinct classes of AddB proteins, and hence, AddAB complexes.

Do *addAB* and *recBCD* genes promote their own horizontal transfer? The AddAB and RecBCD proteins are helicase-nuclease complexes that promote recombination, and it is tempting to speculate that these functions are relevant to the horizontal transfer of *addAB* and *recBCD* genes. In *E. coli*, the nuclease activity of the RecBCD complex has been shown to degrade incoming unprotected bacteriophage DNA, blocking infection (for an example, see reference 29), although in other organisms, RecBCD and AddAB do not prevent acquisition of DNA taken up by natural competence mechanisms (2, 15, 19, 27). In general, however, nuclease degradation of incoming DNA could discourage HGT, while recombination could integrate an incoming foreign DNA fragment, given sufficient homology with the recipient genome. Therefore, if a genome loses functional *addAB* or *recBCD* genes, the reacquisition of either set of those genes via horizontal transfer (presumably increasing organism fitness) could be promoted indirectly by the increased stability of incoming foreign DNA and possibly directly by the recombinase action of the RecBCD or AddAB complexes expressed from the incoming DNA.

In light of the model above, the distribution of the RuvC and RecU Holliday junction resolvase proteins should make a good comparison to that of the RecBCD and AddAB systems. Like RecBCD/AddAB, the RuvC/RecU proteins are very widely conserved and carry out the same function (30). In fact, *B. subtilis* RecU can substitute for RuvC in *E. coli* cells (31), which suggests that the two proteins could successfully be phylogenetically shuffled, like AddAB and RecBCD. However, because RuvC/RecU cannot degrade DNA entering a cell and because *ruvC* and *recU* mutants are less recombination defective than *recBCD* or *addAB* mutants, the model described above to explain HGT and the shuffling of RecBCD and AddAB should not apply to RuvC/RecU.

When the phylogenetic distribution of the RecU and RuvC proteins is examined, it can be seen that among the fully sequenced genomes examined here there is no evidence of replacement of one protein by the other. Instead, RecU is essentially ubiquitous among the *Mollicutes* and *Bacilli* classes of the *Firmicutes* (see Fig. S4 and S5 in the supplemental material), which is consistent with vertical transmission from a common ancestor (13). RuvC is found in almost all nonfirmicutes bacteria and some members of the class *Clostridia* of the phylum *Firmicutes* (see Fig. S4 and S5 in the supplemental material). In two members of the class *Clostridia*, both RuvC and RecU have been detected, and two copies of RecU are also found in a small number of the members of the class *Bacilli*. However, while RuvC and RecU are almost ubiquitous among bacteria, no RuvC proteins are found in the classes *Mollicutes* or *Bacilli* and no RecU proteins are found outside the phylum *Firmicutes* (see Fig. S5 in the supplemental material). This finding is in strong contrast to the frequent phylogenetic shuffling of the RecBCD and AddAB proteins (Fig. 3).

Examination of phylogenetic trees for RuvC and RecU and comparison with species level trees also suggest less frequent HGT of these proteins than of RecBCD and AddAB (see Fig. S6, S7, S8, and S9 in the supplemental material). For AddAB and RecBCD, the phyla *Acidobacteria*, *Actinobacteria*, *Bacteroidetes*, *Chlorobi*, *Cyanobacteria*, *Proteobacteria*, and *Spirochaetes* show evidence of HGT, either through the presence of both complexes within a single phylum or the failure of either the AddAB or RecBCD sequences to form monophyletic groups in their respective trees. The only phylum in which there are more than two sequenced organisms that possess AddAB/RecBCD proteins and in which there is no evidence of HGT is the phylum *Chlamydiae*. In contrast, there is no evidence for HGT of RecU sequences between the phylum *Fir-*

micutes and other phyla, and even within this phylum there is no evidence of HGT of RecU sequences between the classes *Mollicutes* and *Bacilli*. While there is evidence of HGT of RuvC sequences in some phyla, among the *Bacteroidetes*, *Chlorobi*, *Cyanobacteria*, *Deinococcus-Thermus*, and *Thermotogae*, RuvC sequences form monophyletic groups. Similarly, the minimum number of inferred HGT events needed to explain the phylogenetic tree of RuvC sequences from the phylum *Proteobacteria* is much lower than the number needed to explain the patterns of AddAB and RecBCD inheritance within this phylum. Therefore, as expected, the pattern of shuffling and frequent HGT seen for the AddAB and RecBCD proteins was not observed with RuvC and RecU, which is consistent with the proposed model for AddAB/RecBCD HGT described above.

ACKNOWLEDGMENTS

I thank Janet Young, Susan Amundsen, and Gerry Smith for comments on the manuscript.

This research was supported by research grant GM031693 from the National Institutes of Health to Gerry Smith.

REFERENCES

1. **Alonso, J. C., A. C. Stiege, and G. Luder.** 1993. Genetic recombination in *Bacillus subtilis* 168: effect of *recN*, *recF*, *recH* and *addAB* mutations on DNA repair and recombination. Mol. Gen. Genet. **239:**129–136.
2. **Amundsen, S. K., J. Fero, L. M. Hansen, G. A. Cromie, J. V. Solnick, G. R. Smith, and N. R. Salama.** 2008. *Helicobacter pylori* AddAB helicase-nuclease and RecA promote recombination-related DNA repair and survival during stomach colonization. Mol. Microbiol. **69:**994–1007.
3. **Amundsen, S. K., and G. R. Smith.** 2003. Interchangeable parts of the *Escherichia coli* recombination machinery. Cell **112:**741–744.
4. **Amundsen, S. K., A. F. Taylor, A. M. Chaudhury, and G. R. Smith.** 1986. *recD*: the gene for an essential third subunit of exonuclease V. Proc. Natl. Acad. Sci. USA **83:**5558–5562.
5. **Anderson, D. G., and S. C. Kowalczykowski.** 1997. The translocating RecBCD enzyme stimulates recombination by directing RecA protein onto ssDNA in a χ-regulated manner. Cell **90:**77–86.
6. **Aravind, L., R. L. Tatusov, Y. I. Wolf, D. R. Walker, and E. V. Koonin.** 1998. Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. Trends Genet. **14:**442–444.
7. **Biswas, I., E. Maguin, S. D. Ehrlich, and A. Gruss.** 1995. A 7-base-pair sequence protects DNA from exonucleolytic degradation in *Lactococcus lactis*. Proc. Natl. Acad. Sci. USA **92:**2244–2248.
8. **Cano, D. A., M. G. Pucciarelli, F. Garcia-del Portillo, and J. Casadesus.** 2002. Role of the RecBCD recombination pathway in *Salmonella* virulence. J. Bacteriol. **184:**592–595.
9. **Capaldo, F. N., G. Ramsey, and S. D. Barbour.** 1974. Analysis of the growth of recombination-deficient strains of *Escherichia coli* K-12. J. Bacteriol. **118:** 242–249.
10. **Cavalier-Smith, T.** 2006. Rooting the tree of life by transition analyses. Biol. Direct **1:**19.
11. **Chédin, F., and S. C. Kowalczykowski.** 2002. A novel family of regulated helicases/nucleases from gram-positive bacteria: insights into the initiation of DNA recombination. Mol. Microbiol. **43:**823–834.
12. **Chédin, F., P. Noirot, V. Biaudet, and S. D. Ehrlich.** 1998. A five-nucleotide sequence protects DNA from exonucleolytic degradation by AddAB, the RecBCD analogue of *Bacillus subtilis*. Mol. Microbiol. **29:**1369–1377.
13. **Ciccarelli, F. D., T. Doerks, C. von Mering, C. J. Creevey, B. Snel, and P. Bork.** 2006. Toward automatic reconstruction of a highly resolved tree of life. Science **311:**1283–1287.
14. **Faguy, D. M., and W. F. Doolittle.** 2000. Horizontal transfer of catalase-peroxidase genes between archaea and pathogenic bacteria. Trends Genet. **16:**196–197.
15. **Fernández, S., S. Ayora, and J. C. Alonso.** 2000. *Bacillus subtilis* homologous recombination: genes and products. Res. Microbiol. **151:**481–486.
16. **Gil, R., F. J. Silva, E. Zientz, F. Delmotte, F. Gonzalez-Candelas, A. Latorre, C. Rausell, J. Kamerbeek, J. Gadau, B. Holldobler, R. C. van Ham, R. Gross, and A. Moya.** 2003. The genome sequence of *Blochmannia floridanus*: comparative analysis of reduced genomes. Proc. Natl. Acad. Sci. USA **100:**9388–9393.
17. **Guindon, S., and O. Gascuel.** 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. **52:**696–704.
18. **Haijema, B. J., G. Venema, and J. Kooistra.** 1996. The C terminus of the AddA subunit of the *Bacillus subtilis* ATP-dependent DNase is required for the ATP-dependent exonuclease activity but not for the helicase activity. J. Bacteriol. **178:**5086–5091.
19. **Halpern, D., A. Gruss, J. P. Claverys, and M. El-Karoui.** 2004. *rexAB* mutants in *Streptococcus pneumoniae*. Microbiology **150:**2409–2414.
20. **Katoh, K., K. Misawa, K. Kuma, and T. Miyata.** 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. **30:**3059–3066.
21. **Kooistra, J., B. J. Haijema, A. Hesseling-Meinders, and G. Venema.** 1997. A conserved helicase motif of the AddA subunit of the *Bacillus subtilis* ATP-dependent nuclease (AddAB) is essential for DNA repair and recombination. Mol. Microbiol. **23:**137–149.
22. **Kooistra, J., and G. Venema.** 1991. Cloning, sequencing, and expression of *Bacillus subtilis* genes involved in ATP-dependent nuclease synthesis. J. Bacteriol. **173:**3644–3655.
23. **Koonin, E. V., A. R. Mushegian, and P. Bork.** 1996. Non-orthologous gene displacement. Trends Genet. **12:**334–336.
24. **Koonin, E. V., and Y. I. Wolf.** 2008. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic Acids Res. **36:** 6688–6719.
25. **Kowalczykowski, S. C.** 2000. Initiation of genetic recombination and recombination-dependent replication. Trends Biochem. Sci. **25:**156–165.
26. **Lawrence, J. G.** 1997. Selfish operons and speciation by gene transfer. Trends Microbiol. **5:**355–359.
27. **Marsin, S., A. Mathieu, T. Kortulewski, R. Guerois, and J. P. Radicella.** 2008. Unveiling novel RecO distant orthologues involved in homologous recombination. PLoS Genet. **4:**e1000146.
28. **Nelson, K. E., R. A. Clayton, S. R. Gill, M. L. Gwinn, R. J. Dodson, D. H. Haft, E. K. Hickey, J. D. Peterson, W. C. Nelson, K. A. Ketchum, L. McDonald, T. R. Utterback, J. A. Malek, K. D. Linher, M. M. Garrett, A. M. Stewart, M. D. Cotton, M. S. Pratt, C. A. Phillips, D. Richardson, J. Heidelberg, G. G. Sutton, R. D. Fleischmann, J. A. Eisen, O. White, S. L. Salzberg, H. O. Smith, J. C. Venter, and C. M. Fraser.** 1999. Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. Nature **399:**323–329.
29. **Oliver, D. B., and E. B. Goldberg.** 1977. Protection of parental T4 DNA from a restriction exonuclease by the product of gene 2. J. Mol. Biol. **116:**877–881.
30. **Rocha, E. P., E. Cornet, and B. Michel.** 2005. Comparative and evolutionary analysis of the bacterial homologous recombination systems. PLoS Genet. **1:**e15.
31. **Sanchez, H., D. Kidane, P. Reed, F. A. Curtis, M. C. Cozar, P. L. Graumann, G. J. Sharples, and J. C. Alonso.** 2005. The RuvAB branch migration translocase and RecU Holliday junction resolvase are required for double-stranded DNA break repair in *Bacillus subtilis*. Genetics **171:**873–883.
32. **Schultz, D. W., A. F. Taylor, and G. R. Smith.** 1983. *Escherichia coli* RecBC pseudorevertants lacking chi recombinational hotspot activity. J. Bacteriol. **155:**664–680.
33. **Silva, F. J., A. Latorre, and A. Moya.** 2003. Why are the genomes of endosymbiotic bacteria so stable? Trends Genet. **19:**176–180.
34. **Singleton, M. R., M. S. Dillingham, M. Gaudier, S. C. Kowalczykowski, and D. B. Wigley.** 2004. Crystal structure of RecBCD enzyme reveals a machine for processing DNA breaks. Nature **432:**187–193.
35. **Smith, G. R.** 2001. Homologous recombination near and far from DNA breaks: alternative roles and contrasting views. Annu. Rev. Genet. **35:**243–274.
36. **Tamas, I., L. Klasson, B. Canback, A. K. Naslund, A. S. Eriksson, J. J. Wernegreen, J. P. Sandstrom, N. A. Moran, and S. G. Andersson.** 2002. 50 million years of genomic stasis in endosymbiotic bacteria. Science **296:**2376–2379.
37. **Tamura, K., J. Dudley, M. Nei, and S. Kumar.** 2007. MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol. Biol. Evol. **24:**1596–1599.
38. **Taylor, A. F., and G. R. Smith.** 2003. RecBCD enzyme is a DNA helicase with fast and slow motors of opposite polarity. Nature **423:**889–893.
39. **Thompson, J. D., F. Plewniak, and O. Poch.** 1999. A comprehensive comparison of multiple sequence alignment programs. Nucleic Acids Res. **27:** 2682–2690.
40. **Wang, G., and R. J. Maier.** 2009. A RecB-like helicase in *Helicobacter pylori* is important for DNA repair and host colonization. Infect. Immun. **77:**286–291.
41. **Wang, J., and D. A. Julin.** 2004. DNA helicase activity of the RecD protein from *Deinococcus radiodurans*. J. Biol. Chem. **279:**52024–52032.
42. **Waterhouse, A. M., J. B. Procter, D. M. Martin, M. Clamp, and G. J. Barton.** 2009. Jalview version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics **25:**1189–1191.
43. **Yeeles, J. T., and M. S. Dillingham.** 2007. A dual-nuclease mechanism for DNA break processing by AddAB-type helicase-nucleases. J. Mol. Biol. **371:**66–78.
44. **Zuñiga-Castillo, J., D. Romero, and J. M. Martinez-Salazar.** 2004. The recombination genes *addAB* are not restricted to gram-positive bacteria: genetic analysis of the recombination initiation enzymes RecF and AddAB in *Rhizobium etli*. J. Bacteriol. **186:**7905–7913.