Comparative and Evolutionary Analysis of the Bacterial Homologous Recombination Systems

Eduardo P. C. Rocha1,2*, Emmanuel Cornet2, Bénédicte Michel3

1 Unité Génétique des Génomes Bactériens, Institut Pasteur, Paris, France, 2 Atelier de Bioinformatique, Université Pierre et Marie Curie, Paris, France, 3 Laboratoire de Génétique Microbienne, Institut National de la Recherche Agronomique, Jouy en Josas, France

Homologous recombination is a housekeeping process involved in the maintenance of chromosome integrity and generation of genetic variability. Although detailed biochemical studies have described the mechanism of action of its components in model organisms, there is no recent extensive assessment of this knowledge, using comparative genomics and taking advantage of available experimental data on recombination. Using comparative genomics, we assessed the diversity of recombination processes among bacteria, and simulations suggest that we missed very few homologs. The work included the identification of orthologs and the analysis of their evolutionary history and genomic context. Some genes, for proteins such as RecA, the resolvases, and RecR, were found to be nearly ubiquitous, suggesting that the large majority of bacterial genomes are capable of homologous recombination. Yet many genomes show incomplete sets of presynaptic systems, with RecFOR being more frequent than RecBCD/AddAB. There is a significant pattern of co-occurrence between these systems and antirecombinant proteins such as the ones of mismatch repair and SbcB, but no significant association with nonhomologous end joining, which seems rare in bacteria. Surprisingly, a large number of genomes in which homologous recombination has been reported lack many of the enzymes involved in the presynaptic systems. The lack of obvious correlation between the presence of characterized presynaptic genes and experimental data on the frequency of recombination suggests the existence of still-unknown presynaptic mechanisms in bacteria. It also indicates that, at the moment, the assessment of the intrinsic stability or recombination isolation of bacteria in most cases cannot be inferred from the identification of known recombination proteins in the genomes.

Synopsis
Genomes evolve mostly by modifications involving large pieces of genetic material (DNA). Exchanges of chromosome pieces between different organisms as well as intragenomic movements of DNA regions are the result of a process named homologous recombination. The central actor of this process, the RecA protein, is amazingly conserved from bacteria to human. In addition to its role in the generation of genetic variability, homologous recombination is also the guardian of genome integrity, as it acts to repair DNA damage. RecA-catalyzed DNA exchange (synapse) is facilitated by the action of presynaptic enzymes and completed by postsynaptic enzymes (resolvases). In addition, some enzymes counteract RecA. Here, the researchers assess the diversity of recombination proteins among 117 different bacterial species. They find that resolvases are nearly as ubiquitous and as well conserved at the sequence level as RecA. This suggests that the large majority of bacterial genomes are capable of homologous recombination. Presynaptic systems are less ubiquitous, and there is no obvious correlation between their presence and experimental data on the frequency of recombination. However, there is a significant pattern of co-occurrence between these systems and antirecombinant proteins.

Figure 1. Models for the Mode of Action of the Main Homologous Recombination Proteins in E. coli at ssDNA Gaps or dsDNA Ends

Enzymes of known biochemical activities are shown. The presynaptic steps result in the formation of a RecA filament. At gaps, this step requires RecJ, RecF, RecO, and RecR. In E. coli, RecBCD or the RecFOR pathway (Figure 1). Both pathways work to provide a ssDNA molecule coated with RecA to allow the invasion of a homologous molecule [13,15]. RecBCD promotes the repair of double-stranded DNA (dsDNA) breaks, whereas RecFOR is involved in the repair of ssDNA gaps. In the RecBCD pathway, all the required functions—helicase, nuclease, and RecA loading—are assembled in a single holoenzyme [16]. RecBCD binds to dsDNA ends, unwinds, and degrades DNA until it encounters a γ site. The activity of RecBCD is modified at χ, where it starts producing ssDNA and loading RecA [17]. RecF, RecO, and RecR bind gapped ssDNA and displace the SSB proteins to allow RecA coating. There is evidence for interactions between RecR and either RecF or RecO, but not for the existence of a tricomponent complex [18,19]. The RecJ ssDNA exonuclease acts in concert with RecFOR to enlarge the ssDNA region when needed. Strand exchange is catalyzed by RecA [20], a multifunctional protein also involved in the regulation of the SOS response and in the activity of polymerases that facilitate replication across DNA lesions [21]. In E. coli, the joint molecules formed by RecA are resolved by either the RuvABC complex or, in an unknown way, by the action of the RecG helicase. The RuvAB and RuvC proteins catalyze the branch migration and the resolution of Holliday junction recombination intermediates, respectively. These three proteins are thought to interact in a resolvosome complex, in which a RuvABC-junction complex tracks along DNA, with RuvC able to scan for cleavable sequences as the DNA passes through (Figure 1). Finally, replication is directly linked to the recombination process during double-strand break repair, as a viable recombinant is only obtained if the recombination intermediate is used to initiate replication, via the action of the PriA protein [22,23]. Conversely, recombination proteins participate in replication progression as, for example, RecFOR and RecA are required for the resumption of a normal replication rate after treatment with a DNA-damaging agent, and RecBC is required for the viability of several replication mutants [2].

Evidence is accumulating that other bacteria use different proteins for some recombination steps. For example, in firmicutes, RecBCD is replaced by the analogous complex AddAB (named RecAB in streptococci and lactococci) [24,25], and there is evidence indicating that a functional χ site is present in these genomes, albeit variable in size and...
composition [26]. In these genomes, RecA also replaces RuvC [27]. The frequency of homologous recombination is diminished by the action of other proteins. The general mismatch repair system (MutS/LH in *E. coli*) antagonizes homologous recombination between nonidentical DNA sequences by blocking the RecA-mediated strand exchange process if mismatches are present [28]. Hence, the mismatch repair system prevents recombination between homologous sequences and has an important role in defining bacterial species barriers [29]. The helicase II, UvrD, also acts as an antirecombinant, possibly by unwinding the paired DNA recombinant intermediates [30], or by displacing RecA from ssDNA [31]. On the other hand, UvrD can stimulate RecA-driven branch migration and can participate in the RecFOR pathway [32]. Finally, in *recBC* mutant cells, RecFOR can initiate recombination from DNA double-strand ends that have a single-strand extension, but only when SbcB, a ssDNA-specific 3′ → 5′ exonuclease, is inactivated. When present, this nuclease prevents RecFOR action by removing the 3′ extremity on which RecA could be loaded; in addition, the growth of *recBC sbcB* mutants requires the inactivation of the SbcCD proteins for unknown reasons [1,33]. Antirecombinant proteins must be taken into account when assessing the potential recombination machinery of bacteria, as it can be evaluated from genome sequences.

An extended assessment of the proteins involved in DNA repair followed the publication of the first genome sequences [34]. This pioneering work showed that genes implicated in homologous recombination are not homogeneously distributed among bacterial species. Unfortunately, no equivalent extensive work has been done recently that focuses precisely on homologous recombination and takes advantage of the nearly 200 completely sequenced genomes. Yet different sets of recombination-related genes have been found among some bacterial groups [35–38]. We have thus tried to assess the distribution of homologous recombination genes in complete genomes, using a large set of tools involving sequence and phylogenetic analysis [34,39], as well as colocalization data. This type of analysis presupposes that recombination proteins are ancient enough to have diverged from one or a few proteins for which we know the function for at least one element in the family. Although recombination is probably a very old process, our data suggest that some genes may have been missed because they are not yet functionally characterized. A further assumption of our analysis is that sequence similarity will remain strong enough to allow finding these genes by sequence similarity. We make some simulations that suggest that few genes are likely to have been lost if sequence divergence follows the pattern (but not necessarily the rate) of RecA. The use of genomic context should also reduce this problem. Finally, this analysis also supposes that orthologs have similar functions. Although this is usually assumed, proteins with multiple functions may have gained or lost part of them during evolution. For example, the role of RecA in SOS is unused—possibly lost—in the bacteria that lack this response. After establishing the repertoire of genes, we evaluate their co-occurrence, evolutionary rate, and colocalization, taking into account their functional association in known pathways. This was then put into relation with the evolutionary history of genes and the assessment of the experimental evidence for recombination.

**Results/Discussion**

**Introductory Remarks**

As described in Materials and Methods, we first applied an automatic methodology to find candidate orthologs of genes implicated in homologous recombination. The analysis started from genomes for which experimental evidence was available for the function of the genes. This typically included not only *E. coli* and *B. subtilis* but also much less studied bacteria such as mollicutes (for RuvAB [40]), actinobacteria (for Ku [41]), or others. Naturally, when an ortholog was found in a phylogenetic group, it was used to search for further orthologs within the group. Second, we made a more detailed analysis by searching for InterPro domains and making FASTA searches; and by taking into account phylogenetic analyses and information on gene colocalization. Using these diverse sources of information, we were able to list candidate homologous recombination genes in 117 genomes (Figure 2). Some genes are highly conserved in sequence and nearly ubiquitous. For these genes, the methods we used are very reliable and provide uniformly consistent results. However, for some less ubiquitous, fast-evolving, or poorly characterized genes, we found sometimes either inconsistent similarity or weak hits, e.g., similarity smaller than 40%, FASTA hits with *E* ~10−5, matches with a nonspecific motif or with large variation in protein length. Under these conditions, and when no reliable close ortholog is available, it is hazardous to confidently predict orthology. Hence, we conservatively regard these genes as “putative” orthologs. For some proteins, e.g., RecO and RecX, the list of putatives is relatively large.

**RecA and Resolvases Are Nearly Ubiquitous Genes**

No homologous recombination gene is present in all bacterial genomes. However, many genes are widespread among all or nearly all groups and are extremely frequent within each group (Figure 2). RecA is absent only in the several genomes of *Buchnera* and *Blochmania* and presents frame shifts in Onion Yellows (OY) phytoplasma. The near ubiquity of RecA matches well with its preeminent role in homologous recombination and has been previously noticed [34,42,43]. Its absence among intracellular bacteria has also been widely documented [36,44–46]. Unsurprisingly, bacteria lacking RecA have very few other recombination proteins. Several proteins are almost as frequent as RecA. The genes coding for the RuvAB Holliday junction branch migration complex always co-occur and are absent from the genomes that lack RecA and from only two genomes where RecA is present, *Wigglesworthia* Gb and *Aquifex aeolicus* (Figures 2 and 3). Although they lack RuvAB, these two genomes contain a RecG ortholog—another Holliday junction branch migrating helicase. The gene for RecG is also very frequent, absent only from all mollicutes and all Chlamydiaceae as well as from *Desulfovibrio vulgaris*.

Some proteins are believed to be functional analogs, although they apparently lack a common evolutionary history (i.e., they are not orthologs). RuvAB in *E. coli* forms a complex with the resolvase RuvC. RuvC is less ubiquitous than RuvAB, which is explained by its functional replacement by the analog RecU in firmicutes and mollicutes [27]. Our data indicate that only ten genomes lack both RuvC and RecU (this includes the genomes that lack RecA; Table 1). In these rare
cases, the resolvase function may be provided by YqgF [47], which is only absent from seven genomes. However, our data suggest that RuvC/RecU and YqgF are not simple functional analogs because they co-occur in the large majority of genomes. In addition, a resolvase activity of the YqgF proteins has not yet been demonstrated either in vitro or in vivo. The function performed by resolution proteins may also be carried out by prophage-encoded proteins [48].

PriA is nearly ubiquitous and is only absent in genomes of some intracellular endosymbionts, *Deinococcus radiodurans*, *A. aeolicus*, and from most genomes of mollicutes. Among actinobacteria, there is a putative ortholog of PriA that is smaller and very divergent. With the exception of *Candidatus Blochmania floridanus* (which lacks RecA), all genomes with AddAB or RecBCD (the presynaptic proteins that act at double-strand ends) have PriA. In conclusion, RecA, branch migration systems, and resolvases, and to a lesser extent the protein that couples recombination and replication PriA, are present in nearly all the bacterial genomes (Table 1).

The RecBCD and AddAB Presynaptic Recombination Proteins

RecBCD provides another example of complementary distribution of similar but nonorthologous systems. The AddAB proteins (and their orthologs RexAB) replace RecBCD in firmicutes and in most β- and α-proteobacteria. AddAB is almost ubiquitous among these groups, as it is missing only in *Bacillus halodurans*, *Neisseria meningitidis*, and *Chromobacterium violaceum*—these having RecBCD instead. A recent work analyzed a homolog of AddA in proteobacteria and confirmed its role in the repair of double-strand breaks [49]. Although AddA and AddB closely co-occur in most genomes, the AddB gene of *B. subtilis* has no significant similarity with the ones of proteobacteria (*E* > 0.01 for FASTA hits, <25% identity on a global alignment). Because AddB is slightly more conserved than AddA among firmicutes (see following), one would expect the AddB protein of proteobacteria to have significant similarity with the AddB protein of firmicutes if it shared a common evolutionary

---

**Figure 2.** Probable Presence, Putative, and Unlikely Presence of Recombination-Associated Genes in the Studied Genomes
Black indicates presence is probably, grey indicates putative presence, and white indicates presence is unlikely. F indicates that the gene is present in the genome but contains frame shifts (genes with known programmed frame shifts, introns, and inteins are indicated, as regular genes, in black).

DOI: 10.1371/journal.pgen.0010015.g002

**Figure 3.** Distribution of Some Recombination Genes in a Phylogenetic Tree of Bacteria

Tree adapted from [99]. The position of *Pirellula* and *Fusobacterium* are still unclear [100].

DOI: 10.1371/journal.pgen.0010015.g003
history. Hence, the AddB proteins of the two clades may be functional analogs but not orthologs. This is consistent with recent data indicating that AddA shares stronger resemblance with RecB than AddB does with RecC, reflecting a more central role for the function of RecB/AddA in the complex (M. El Karoui, personal communication).

Genes coding for proteins that participate in complexes tend to systematically co-occur in genomes. This is the case for AddAB, RuvAB, RuvAB/RuvC(RecU), SbcCD, and MutS1L (see following). A major exception to this trend is the frequent occurrence of a RecD protein when RecD is absent, in mollicutes, firmicutes, Des. vulgaris, and some groups of proteobacteria (see Figure 2). For example, among actinobacteria, the Mycobacterium (with RecBC) and the two Streptomyces (without) are on opposite sides of the tree, and a similar contrast is found in 6-proteobacteria, where Geo bacter sulfurreducens has RecBC and Des. vulgaris does not. In some genomes, such as Chlamydiae, there are multiple copies of RecD, typically one in each side of the tree. The analysis of the protein sequences of the two groups of RecD shows a major difference between RecD1 (a protein systemically associated with RecBC) and RecD2 (a protein present in genomes lacking RecBC). Within each RecD group, one can identify most of the major phylogenetic groups of bacteria. For example, among actinobacteria, the Mycobacterium (with RecBC) and the two Streptomyces (without) are on opposite sides of the tree, and a similar contrast is found in 6-proteobacteria, where Geo bacter sulfurreducens has RecBC and Des. vulgaris does not. In some genomes, such as Chlamydiae, there are multiple copies of RecD, typically one in each side of the tree. The analysis of the protein sequences of the two groups of RecD shows a major difference between them. RecD2 contains an N-terminus extension including a domain identified as RuVA domain 2-like in InterPro that is absent from RecD1. This domain is also present in UvrC and is essential for the 5' incision in the prokaryotic nucleotide excision repair process [50]. The RecD2 protein of D. radiodurans, the only one biochemically studied, is a DNA helicase with a low processivity and a yet-unidentified role [51].

Finally, some bacteria have a functional nonhomologous end-joining mechanism (NHEJ), allowing the repair of dsDNA breaks [52]. Contrary to homologous recombination, NHEJ does not require sequence homology—only complementary ends. The key factors of NHEJ are a Ku protein that binds to the termini of the double-strand breaks and has the bridging activity, and a ligase that ligates the termini. Our results indicate that NHEJ genes are present in few bacteria (Ku is present in 24 genomes out of the 117), with no particular phylogenetic trend, as they are found in firmicutes, actino-

Table 1. Genomes Missing Key Components of the Homologous Recombination

<table>
<thead>
<tr>
<th>RecA</th>
<th>Holiday Junction Processing—RecG and RuvAB</th>
<th>Holiday Junction Resolvases—RuvC and RecU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Buchnera</td>
<td>Buchnera</td>
<td>Aquilax aequosus a</td>
</tr>
<tr>
<td>Candidatus Blochmania</td>
<td>Candidatus Blochmania</td>
<td>Candidatus Blochmania b</td>
</tr>
<tr>
<td>OY phytoplasma (P)</td>
<td>OY phytoplasma</td>
<td>Borrelia burgdorferi a</td>
</tr>
<tr>
<td>Clostridium b</td>
<td>Clostridium b</td>
<td>Clostridium b</td>
</tr>
<tr>
<td>M. genitalium b</td>
<td>M. genitalium b</td>
<td></td>
</tr>
<tr>
<td>M. mycoides (O. mycoides)</td>
<td>M. mycoides (O. mycoides)</td>
<td>OY phytoplasma b</td>
</tr>
</tbody>
</table>

Listed in each column are the genomes missing the components given at the top of each column.

aPresence of putative prophage resolvase [47].
bPresence of putative ubiquitous resolvase Yqf.

cDOI: 10.1371/journal.pgen.0010015.t001

bacteria, and several groups of proteobacteria (see Figure 2). As indicated previously [53,54], the two genes tend to co-occur contiguously in genomes, probably constituting an operon. In some bacteria, we found many copies of the Ku ligase genes. For example, Agrobacterium tumefaciens contains six copies of the Ku gene and eight copies of the ligase, and Bradyrhizobium japonicum contains four copies of the Ku gene and two copies of the ligase. Thus, in these genomes, Ku has probably a very important role. We then tested the patterns of co-occurrence of NHEJ and RecBCD/AddAB to see whether the presence of one could compensate for the absence of the other (as both act to repair double-strand breaks). We found these systems to co-occur independently ($p = 0.6, \chi^2$ test). NHEJ is the major pathway for repairing DNA double-strand breaks in mammalian cells, whereas homologous recombination is so in yeast [55]. Because most bacterial genomes lack NHEJ, homologous recombination also appears to be the major repair pathway acting on such lesions in bacteria.

The RecFOR Presynaptic Proteins

Whereas the RecB, RecC, and RecD polypeptides form a stable active complex, in the RecFOR pathway, there are interactions between some of the elements but no stable complexes between the three proteins. Interestingly, the RecBCD/AddAB and RecFOR proteins, instead of showing a complementary pattern of co-occurrence, tend to co-occur more frequently than expected ($p < 0.001, \chi^2$ test). This

Figure 4. Unrooted Phylogenetic Tree of the RecD Protein

The dotted line separates genomes containing RecBCD from the ones containing only RecD. The tree was constructed using Tree-Puzzle, using the JTTþI model with eight classes [74]. Bootstraps were made using SEQBOOT and CONSENSE from the PHYLIP package [101]. C. acetobutylicum, Clostridium acetobutylicum; C. tepidum, Chlorobium tepidum; C. violaceum, Chromobacterium violaceum; D. vulgaris, Desulfovibrio vulgaris; E. carotovora, Erwinia carotovora; E. faecalis, Enterococcus faecalis; L. plantarum, Lactobacillus plantarum; L. lactis, Lactococcus lactis; P. multocida, Pasteurella multocida; M. mobile, Mycoplasma mobile; M. flororum, Mesoplasma flororum; M. mycoides, Mycoplasma mycoides; M. pulmonis, Mycoplasma pulmonis; P. maritima, Prochlorococcus maritima; S. enterica, Salmonella enterica; S. oneidensis, Shewanella oneidensis; X. fastidiosa, Xylella fastidiosa.

DOI: 10.1371/journal.pgen.0010015.g004
means that if RecBCD/AddAB is present (absent), then RecFOR is more likely to be present (absent), which probably reflects the specificity of these two systems on complementary types of lesions (see Figure 1).

Although RecF historically served as a reference for this pathway, it is absent from 29 genomes and is the least frequent protein in the set (see Figure 2). At the other extreme, RecR is the most frequent, being absent from only 19 genomes. In agreement with RecR being present in the two active complexes RecOR and RecR [18,19], there is no single occurrence of RecO or RecF when RecR is absent.

In *E. coli*, the RecJ exonuclease acts during gap repair to enlarge the ssDNA region for RecFOR binding [56]. Recj is absent from the species that lack RecA and from the mollicutes and the mycobacteria, which may use an alternative exonuclease. RecQ is absent from 48 genomes, in agreement with the observation that the RecQ helicase is required in *E. coli* for RecFOR-mediated recombination only in a *recBC sbeB sbeCD* mutant [57].

Recombination without Presynaptic Recombination Proteins?

Our analysis indicates that certain bacterial genomes lack most presynaptic recombination proteins (see Figure 2). One possibility is that these genomes lack homologous recombination altogether. This may be the case for some species lacking nearly all homologous recombination proteins, such as all *Buchnera*, or the OY phytoplasma (Table 1). However, for the genomes containing RecA and resolvases, this is most unlikely. We therefore made an extensive analysis of the literature and selected genomes lacking most presynaptic proteins but for which there is evidence for homologous recombination (Table 2). Such evidence comes from experimental studies of the homologous recombination processes or experimental studies that have used homologous recombination to engineer/inactivate genes, and from multilocus sequence typing data that indicate a population structure driven by frequent recombination. One also typically assumes that natural transformation is used for recombination repair or gene acquisition, which suggests that competent bacteria should have some type of homologous recombination [4,58].

It is surprising that highly recombining genomes, such as *Helicobacter pylori* [59–61] or *Streptomyces coelicolor* [62] lack a large fraction of the presynaptic proteins. One should note that with the exception of both *Streptomyces*, these genomes also lack NHEJ, and many also code for antirecombinants, such as MutS2. This suggests that either presynaptic proteins are dispensable for efficient homologous recombination in some genomes or other, unknown systems, exist in these genomes. The first hypothesis is supported by data indicating that some *E. coli recA* mutations (*RecA* P67W, *RecA*441, *RecA*730, and *RecA*803) can displace SSB proteins much more efficiently than the wild-type, and thus function in the absence of presynaptic proteins [63]. However, if some genomes lack presynaptic functions because their RecA protein is able to efficiently bind SSB-covered DNA, it is not through one of the studied RecA mutations in *E. coli*, because we did not find any of these mutations in natural genomes. Furthermore, it remains to be understood how organisms lacking presynaptic functions could control RecA activity to avoid its improper fixation to any ssDNA (e.g., on the template of the replicating lagging strand). Yet-unidentified presynaptic systems may exist in these genomes. Recombination presynaptic functions are fulfilled in eukaryotes by proteins that have no homology with *E. coli* proteins, in spite of their capacity to facilitate the binding to DNA of their cognate RecA homolog [64].

Proteins That Antagonize Homologous Recombination

Another way of increasing the frequency of homologous recombination without making changes in the recombination machinery is to eliminate the function of antirecombinant proteins. We tested whether there are associations between the losses of presynaptic systems and the losses of antirecombinant proteins, such as UvrD, MutS1L, MutS2, and SbeB genes. UvrD is nearly ubiquitous. The presence of MutS1L correlates with the presence of RecBCD/AddAB and RecFOR (*RecBCD/AddAB*: observed 102, expected 69; Re-
cFOR; observed 91, expected 80; both \( p < 0.005 \), Pearson’s exact test). This suggests that a lower activity of RecA in the absence of presynaptic systems can be compensated for by the loss of the mismatch repair system. Contrary to MutS1, MutS2 is not involved in mismatch repair and suppresses homologous recombination between identical sequences, in addition to homeologous recombination, in *H. pylori* [60]. However, no significant association was found between the presence or absence of MutS2 and that of the presynaptic systems. As the *H. pylori* enzyme is the only MutS2 that has been studied in detail so far, it is possible that the antirecombination property of this MutS2 protein is specific for this species.

SbcB, which in RecBCD backgrounds prevents the repair of double-strand breaks by RecFOR, has a statistically significant pattern of co-occurrence and co-omission with RecBCD/AddAB (observed 63, expected 53, \( p < 0.01 \), Pearson’s exact test), but not with RecFOR (\( p > 0.1 \), same test). In fact, only one of the bacteria lacking RecBCD/AddAB contains SbcB. This indicates that the absence (presence) of RecBCD/AddAB is correlated with the absence (presence) of this antirecombinant gene, which may allow RecFOR to efficiently repair double-strand breaks in RecBCD/AddAB backgrounds. SbcCD is much more frequent than SbcB and also co-occurs with RecBCD/AddAB (observed 64, expected 52, \( p < 0.01 \), Pearson’s exact test). However, the role of SbcCD in homologous recombination is unclear.

Colocalization of Genes

Genes involved in a common mechanism tend to be tightly coregulated and, for this reason, clustered in the genome [65]. We have therefore searched for the colocalization of these genes among our set of genomes. With few exceptions, we found that only the recombination genes that are part of stable complexes are systematically clustered. The addAB genes colocalize in 20 of 21 co-occurrences among firmicutes, the exception being *Clostridium tetani*. Among proteobacteria these genes are together in 13 of 13 genomes. The three genes for RecBCD were found to colocalize in 28 of their 31 co-occurrences. RuvA and RuvB colocalized in 77 of 111 co-occurrences, with exceptions including all chlamydiae, all cyanobacteria, all \( \varepsilon \)-proteobacteria, all streptococci, all bacteroides, and most spirochetes, as well as a few phylogenetically dispersed genomes. RuvA, RuvB, and RuvC colocalized in 45 of 78 co-occurrences of the three genes. In firmicutes and mollicutes, RuvC is replaced by RecU, but this gene only colocalizes with RuvAB in two genomes (*Mycoplasma genitalium* and *M. pneumoniae*). Thus, RecU and RuvC are very different in this respect. YqgF was rarely found close to other recombination genes. The two key genes for NHEJ (Ku and the ligase) were found together in 19 of 24 genomes. Naturally, as for the co-occurrence of genes in genomes, the closeness of their co-occurrence is influenced by the phylogenetic distribution of the available genomes. Close occurrence of genes in highly sampled clades, e.g., firmicutes or proteobacteria, will be more preeminent than in clades with few available sequences.

RecA and RecX are close in many genomes and are partly coexpressed in *E. coli* [66]. In some bacteria, the overexpression of RecA is toxic in the absence of RecX, and in vitro, RecX modulates the action of RecA by blocking the extension of the RecA filament [67]. However, although in *E. coli* RecX inhibits the action of RecA [68], in *Neisseria gonorrhoeae* its inactivation leads to a decrease in homologous recombination [66]. Expanding previous observations [69], we found that 35 of the 37 co-occurrences of bona fide orthologs of recX colocalize with recA. The exceptions are *N. meningitidis* and *Photobacterium luminescens*. In contrast, very few genes among the more distantly related, putative recX orthologs are physically close to recA genes. In particular, the putative recX of firmicutes are systematically far in the chromosome from recA. The proteins coded by these genes are larger and less than 40\% similar to the RecX from *E. coli* and from actinobacteria. It is thus uncertain whether they perform the same function. However, RecX also shows large relative variations in length among well-characterized orthologs (e.g., among \( \gamma \)-proteobacteria the *E. coli* protein has 166 residues, whereas in *Yersinia pestis* it has 188, and in *Shewanella oneidensis* it has 123). It has been suggested that the uncoupling between recA and the putative recX in *N. gonorrhoeae* and *B. subtilis* could be associated with their competence for natural transformation [66]. However, such uncoupling is a characteristic of all firmicutes, not specifically of the competent ones, and it is not found in other competent bacteria such as *Haemophilus influenzae* or *H. pylori* (which lacks RecX).

Although recF, recR, and recO do not colocalize, both recF and recO often colocalize with genes coding for replication proteins. Many genomes have an operon close to the replication origin containing four genes: *dnaA* (involved in replication initiation), *dnaN* (\( \beta \)-clamp of the DNA polymerase III), *recF*, and *gyrB* (DNA gyrase) [70]. Among the 86 occurrences of recF, it is close to *dnaA* in 54, close to *dnaN* in 58, and close to *gyrB* in 52. The four genes are together in 40 genomes. Finally, the *dnaX* gene, which encodes both the \( \gamma \) and \( \delta \) subunits of *E. coli* DNA polymerase III, is close to recR in *E. coli*, and the genes are partially cotranscribed [71]. Among the 97 genomes containing *dnaX* and recR, the genes colocalize in 65. These results indicate that instead of clustering together, recombination genes that are not part of stable complexes are often colocalized with genes involved in replication. The linkage between genes of these two cellular processes is certainly associated with the role of homologous recombination in repairing DNA lesions that block DNA synthesis [72,73].

Relative Evolutionary Rates of the Proteins

The substitution rate of proteins is the result of the interplay between mutation and functional constraints. Hence, if one discounts horizontal gene transfer, the differences in substitution rates between proteins should reflect their relative tolerance to change (i.e., they should be associated with the fraction of changes that allows maintaining the function). To assess the relative tolerance of each recombination protein to changes, we computed evolutionary distances within the sets of all bona fide orthologs, using Tree-Puzzle [74]. We then used RecA as the reference protein because of its near ubiquity and slow evolutionary rate [42]. The regression analyses of the substitution rates of each protein as a function of the substitution rate of RecA showed one single group in which RecA evolves faster—the mollicutes (data not shown). We have thus not used these points in the regressions. All other proteins were then compared to RecA, and we found a considerable diversity among the different proteins in terms of substitution rates.
A more developed version of this method has recently been proposed to find horizontal gene transfer between distant taxa [75]. Using our data, we found very little evidence of such events (data not shown). RuvB has evolved almost as slowly as RecA (16% faster), whereas some proteins have evolved a little faster, such as RecR (+68%) and RecU (+100%). However, most proteins have evolved much faster than RecA. Among these, there is a group of proteins that has evolved between 4.0 and 4.5 times faster than RecA and that includes RecB, RecD, RecX, AddA, AddB, YqgF, and RecO. Because RecD is divided in two groups, these data only include the RecD proteins that are in the group of genomes containing RecBC (i.e., RecD1).

The proteins of the RecFOR pathway have a peculiar evolutionary pattern. In addition to being present with very different frequency, with RecF being more frequently absent than RecR or RecO, they also show remarkably different substitution rates, with high conservation for RecR, lower conservation for RecF, and among the lowest conservation for RecO (Figure 5). This may be the result of the double participation of RecR in interactions with RecO and RecF, which would increase the constraints on its evolution. The crystal structure of the *D. radiodurans* RecR protein reveals the existence of a ring-shaped tetramer, theoretically able to encircle dsDNA [76]. This particular clamp-like structure may also have contributed to the high level of conservation of the protein.

It's interesting to note that among the fastest-evolving proteins, some are nearly ubiquitous (RecD and YqgF), and some are much rarer (RecB and AddAB). This suggests that few proteins have been missed in the analysis as a result of excessive sequence divergence. We made a set of simulations to assess this problem more precisely. We allowed protein sequences to evolve according to the evolutionary model of RecA, but at a different relative rates (see Materials and Methods). This analysis showed that only proteins evolving more than four times faster than RecA are expected to be missed in our similarity searches at this evolutionary distance and using our 40% similarity criterion (Figure 6). Even for proteins evolving 5.5 times faster than RecA, in none of our 100 simulations would we miss more than six orthologs. These orthologs were systematically in the fast-evolving mollicutes clade. Naturally, this is an oversimplification of the evolution of proteins, because proteins evolve in a changing context, and this may change their relative rates of evolution. In addition, these analyses do not take into account that insertions and deletions may be more frequent in some proteins than in others. Yet they indicate that few homologous genes are expected to have been lost in the present analysis as a result of excessive sequence divergence.

Conclusion

The presynaptic role of RecBCD and RecFOR and the branch migration activity of RuvAB and RecG suggest functional redundancy, whereas, in contrast, the patterns of co-occurrence of these systems agree with the experimental works indicating complementary, and not redundant, roles for these proteins. Interestingly, this work also indicates that the RecFOR pathway may be more conspicuously important among bacteria than RecBCD, as it is significantly more frequent. RecR is the most conserved of the three proteins, and understanding how recombination is promoted in the organisms that encode a RecR homolog but do not have RecF or RecO would help understand the functioning of these recombination mediator proteins. The associations of *redR* and *redF* with genes involved in replication are often conserved, suggesting that the close association between

---

**Figure 5.** Regression Lines of the Substitution Rates of the Recombination Proteins Plotted against the Substitution Rates of RecA

RecA is the slowest and most ubiquitous of these proteins and its substitution rates are the x-axis of the plot (dashed lines indicate the RecA identity line). The regression was forced to pass through zero at intercept, and the slopes of the lines indicate the relative rapidity of the protein’s evolution relative to that of RecA (varies from 1.3 for RuvB to 4.4 to RecB); the points associated with mollicutes were removed because we found these to evolve proportionally faster in RecA than in the other proteins.

DOI: 10.1371/journal.pgen.0010015.g005

**Figure 6.** Results of the Simulations of Protein Evolution following the Phylogenetic Tree of RecA Using the JTT + I Model with Eight Classes and an a of 0.59

We used Seq-Gen to evolve protein sequences with regularly spaced scaling factors (100 experiments for each scaling factor) and analyzed for each experience which sequences showed less than 40% similarity.

DOI: 10.1371/journal.pgen.0010015.g006
replication and recombination observed in _E. coli_ is common to most bacteria.

A central tenet of current genomic studies is the possibility of associating gene content with phenotype variation. Because the abundance of repeats in genomes correlates well with rearrangement rates and with the capacity of generating genetic variation [8,9], and because repeats are cause and consequence of recombination processes, one could expect an association between the repertoire of recombination genes and the number of repeats. We were unable to observe such a correlation. Indeed, except for genomes lacking RecA and resolvases (which are stable, have few repeats, and possibly lack homologous recombination), bacteria known to recombine frequently may either have a complete repertoire of known recombination genes or lack a substantial part of it. A striking example of the latter is provided by _H. pylori_ [77], which is highly recombinogenic, although it lacks most presynaptic proteins and has antirecombinants such as UvrD and MutS2. In addition, at the intraspecies level, the differences in the population structure do not correlate with the genome content in recombination proteins. For example, serogroup A of _N. meningitidis_ is mostly clonal, contrary to the majority of the others [78]. However, we found that both serotypes A [79] and B [80] have the same almost complete repertoire of homologous recombination proteins. Hence, associations between stability of a genome and the lack of some recombination proteins, as was proposed for _Bifidobacterium longum_ [81] and _Corynebacterium_ species [38], must be viewed with exceptional care before experimental confirmation.

The reasons for this lack of simple association between genotype and phenotype are probably multiple. Orthologs do not necessarily have the same exact functionalities and are likely to have different levels of activity. For example, presynaptic systems may be less necessary if the affinity of RecA for ssDNA is higher. The frequency of recombination events may also depend on the implication of recombination proteins in different cellular processes. For example, the coupling of recombination and replication may depend on the replication machinery and on the frequency of replication arrest. Specific genetic regulatory systems may also lead to different rates of recombination. For example, the onset of competence may be differently related in various organisms with cell growth and with the level of expression of recombination enzymes. Also, equivalent cellular processes may be associated with different enzymatic systems. For example, in neisserial species and _E. coli_, transformation-associated recombination takes place through the RecBCD pathway, whereas in _B. subtilis_, chromosomal transformation decreases 2.5-fold in a _recO_ mutant [82], and in streptococci, AddAB is not involved in chromosome transformation [83], possibly because in competent firmicutes only ssDNA enters the cell. In contrast, in the competent _Helicobacter_ and _Campylobacter_ species, all these genes but RecR are absent. One could also expect that recombination activity is also constrained by ecological factors. Endosymbionts live in very protected environments, and this, associated with reductive genome evolution, has led to the loss of recombination functions [36,37]. However, apart from this case, we could not find any other obvious association between lifestyle and the presence or absence of recombination proteins, which once again is in agreement with the inherent housekeeping role of homologous recombination.

This housekeeping role of homologous recombination is probably also why we found little evidence of horizontal transfer among these genes. Genes implicated in the generation of genetic variation tend to be frequently horizontally transferred [84,85], but not housekeeping genes involved in managing genetic information [86]. Interestingly, multilocus sequence data also indicate that RecA rarely recombines among strains of the same species [87,88]. This does not mean that horizontal transfer is altogether absent. Such events are the most parsimonious explanation for the existence of some analogous replacements, such as AddAB among proteobacteria or RuvC in _Thermoanaerobacter tengcongensis_. They are also probably responsible for the sporadic occurrence of NHEJ in different phylogenetic groups. In addition, given the frequency of prophage sequences in bacterial genomes [89], and the many phage-encoded recombination systems, recombination genes of known phage origin, which have not been included in this study, may also play a role in the variations of recombination mechanisms.

Our study defines a core of recombination genes encoding for proteins nearly ubiquitous in bacterial species. These include the genes that encode RecA (which has a homolog among eukaryotes), RuvAB, RecR, RuvC/RecU, and to a minor extent RecG, RecN, RecJ, and PriA. These genes are present in nearly all bacterial groups and show little horizontal transfer. This justifies the use of such proteins as phylogenetic markers [43]. Their widespread distribution demonstrates their importance in bacteria and justifies the emphasis on their detailed biochemical and functional study.

**Materials and Methods**

Data. We analyzed the number of 117 different bacterial species (see Figure 2), taken from GenBank Genomes (ftp://ftp.ncbi.nih.gov/genomes/Bacterial). The list of proteins related to homologous recombination was taken from the literature [13,24] and included RecA, RecB, RecC, RecD, RecF, RecG, RecJ, RecN, RecO, RecQ, RecR, RecU, RecX, RecA, RecB, RecD, MutS1, MutS2, MutL, and UvrD; the putative resolvase YggF [47]; and the Ku and ligase genes responsible for nonhomologous end joining in some bacteria [52].

Assignment of orthology. One should note that many recombination genes belong to large protein families, such as helicases [90] or nuclease [47]. Hence, simple sequence similarity is not an indication of orthology. Assignment of orthology followed an automated step and then manual curation. The automatic method was the following. We started from the protein in _E. coli_ (except for AddAB, MutS2, Ku, and RecU, where we started from _B. subtilis_) and searched for orthology in all other genomes. Genes were regarded as potential orthologs if they were bidirectional best hits with at least 40% similarity in sequence and their sequences were less than 30% different in length. The alignments were done using an adapted version of the Needleman-Wunsch algorithm (global alignment), in which the nonaligned edges of the largest sequence are not penalized [91], using the matrix BLOSUM60 and typical gap penalties. For comparison, we also made FASTA searches, because they allow for the detection of more local similarities [92]. Then we took the less similar protein hit, respecting the previously cited conditions as a query, and relaunched the analysis on the entire set of genomes with the same parameters. The proteins resulting from the intersection of these lists were temporarily regarded as bona fide orthologs. The other proteins were put together with the ones showing significant FASTA hits (E < 10⁻⁴⁷) on the other genomes, as well as the ones originally annotated as orthologs (but not respecting the above conditions). We then
References


11. Meriem El Karoui, Ivan Matic, Vincent Daubin, and two anonymous reviewers provided important comments and criticism on the manuscript. Alain Blanchard and Pascal Sirand-Pugnet provided important input and thoughts on recombination in mollicutes.

Competing interests. The authors have declared that no competing interests exist.

Author contributions. EPCR conceived and designed the experiments. EPCR and BM performed the experiments. EPCR, EC, and BM analyzed the data. EPCR and EC prepared the paper.
Recombination Systems in Bacteria


