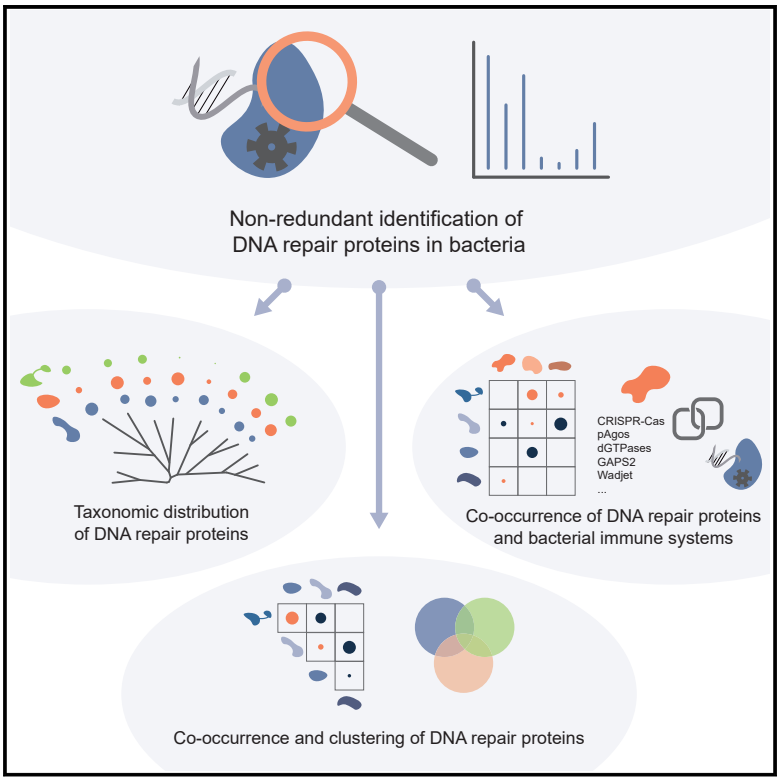
Cell Reports

Article

Distribution of bacterial DNA repair proteins and their co-occurrence with immune systems

Graphical abstract



Authors

Sumanth K. Mutte, Patrick Barendse, Pilar Bobadilla Ugarte, Daan C. Swarts

Correspondence

daan.swarts@wur.nl

In brief

Mutte and Barendse et al. developed a search strategy to accurately identify bacterial DNA repair proteins. This enabled systematic analyses of their taxonomic distribution, co-occurrence, genomic clustering, and co-occurrence with bacterial immune systems, offering an updated perspective on the distribution of DNA repair systems and their connection to immune systems.

Highlights

- A novel strategy enables stringent identification of bacterial DNA repair proteins
- An updated taxonomic distribution of bacterial DNA repair systems is provided
- RecBC(D), AddAB, and AdnAB (but not Rec(F)OR) are mostly clustered on the genome
- Co-occurrence of DNA repair proteins and bacterial immune systems is analyzed

Mutte et al., 2025, Cell Reports *44*, 115110
January 28, 2025
https://doi.org/10.1016/j.celrep.2024.115110

CellPress

Cell Reports

CellPress
OPEN ACCESS

Article

Distribution of bacterial DNA repair proteins and their co-occurrence with immune systems

Sumanth K. Mutte,[1,2,3] Patrick Barendse,[1,3] Pilar Bobadilla Ugarte,[1] and Daan C. Swarts[1,4,*]
[1]Laboratory of Biochemistry, Wageningen University, 6708 WE Wageningen, the Netherlands
[2]MyGen Informatics, 6706 JE Wageningen, the Netherlands
[3]These authors contributed equally
[4]Lead contact
*Correspondence: daan.swarts@wur.nl
https://doi.org/10.1016/j.celrep.2024.115110

SUMMARY

Bacteria encode various DNA repair pathways to maintain genome integrity. However, the high degree of homology between DNA repair proteins or their domains hampers accurate identification. Here, we describe a stringent search strategy to identify DNA repair proteins and provide a systematic analysis of taxonomic distribution and co-occurrence of DNA repair proteins involved in RecA-dependent homologous recombination. Our results reveal the widespread presence of RecA, SSB, and RecOR proteins and phyla-specific distribution for the DNA repair complexes RecBCD, AddAB, and AdnAB. Furthermore, we report co-occurrences of DNA repair proteins with immune systems, including specific CRISPR-Cas subtypes, prokaryotic Argonautes (pAgos), dGTPases, GAPS2, and Wadjet. Our results imply that while certain DNA repair proteins and immune systems might function in conjunction, no immune system strictly relies on a specific DNA repair protein. As such, these findings offer an updated perspective on the distribution of DNA repair systems and their connection to immune systems in bacteria.

INTRODUCTION

In bacteria, genomic DNA can be damaged by both endogenous and exogenous processes. Endogenous processes include stalled replication forks, metabolic/radical products, and nucleases,[1–3] whereas exogenous processes include UV radiation and other DNA-damaging agents, including antibiotics, chemicals, and acids.[1,4,5] As genomic DNA encodes essential genes, maintaining its integrity is vital. To this end, bacteria encode diverse DNA repair pathways.[6–8] Mutations of single DNA nucleotides are generally repaired by base excision repair, nucleotide excision repair, or mismatch repair pathways.[9,10] In contrast, double-stranded DNA breaks (DSBs) and other DNA lesions are generally repaired by either non-homologous end joining (NHEJ) or homologous recombination (HR) pathways.[7,11] Alternatively, DSBs can be repaired through microhomology-mediated end joining (MMEJ)[12] and single-strand annealing (SSA)[13] pathways, of which proteins involved partially overlap with NHEJ or HR pathways. NHEJ involves the removal of damaged bases, error-prone resynthesis, and direct ligation of the free DNA ends.[14,15] While NHEJ is the main mode of DSB repair in eukaryotes, only 22% of bacteria encode homologs of known NHEJ proteins.[16,17] Instead, DSBs are mostly repaired through HR pathways in bacteria.[11,18] During HR, a DNA template homologous to the damaged DNA is used to repair the DSB, which allows mutation-free DNA repair.[18,19] As most bacteria harbor multiple copies of their genome (or duplicated genomes are present during replication), homologous DNA templates required for HR are generally readily present in bacteria.[20]

Distinct bacterial HR pathways exist, each of which involve proteins that can also be used in other distinct DNA repair pathways.[18] Most bacterial HR pathways involve the protein RecA (i.e., RecA-dependent HR pathways). While RecA-independent HR pathways also exist, their activity has only been observed under specific circumstances (e.g., in bacteria without RecA, at [short/inverted] direct repeats and for specific types of DNA damage)[13,21–23] and are therefore not discussed further here. RecA-dependent HR pathways rely on the formation of a single-stranded (ss)DNA fragment to which RecA binds. Consequentially, ssDNA-bound RecA forms a presynaptic filament that places the ssDNA in a helical B-form conformation, which allows the ssDNA to invade homologous dsDNA sequences.[24–26] This facilitates strand exchange, after which HR can be completed through branch migration and resolving the Holliday junction by various proteins (e.g., RecG, RecQ, Rus, and/or RuvABC[27–29]). Preceding strand invasion, RecA relies on one of various DNA repair pathways for the formation of and/or RecA loading onto ssDNA fragments. Well-studied bacterial DNA repair pathways involved in ssDNA generation and RecA loading include Rec(F)OR, RecBCD, AddAB, and AdnAB (Figure 1).

The RecFOR pathway is typified by RecF, RecO, and RecR (Figure 1A).[31–35] The RecFOR pathway is involved in the repair of ssDNA gaps that can occur during DNA replication[31] and in plasmid recombination and conjugation.[35] Such ssDNA gaps are sometimes extended by the nuclease RecJ, in cooperation with the helicase RecQ,[36,37] after which the ssDNA is coated

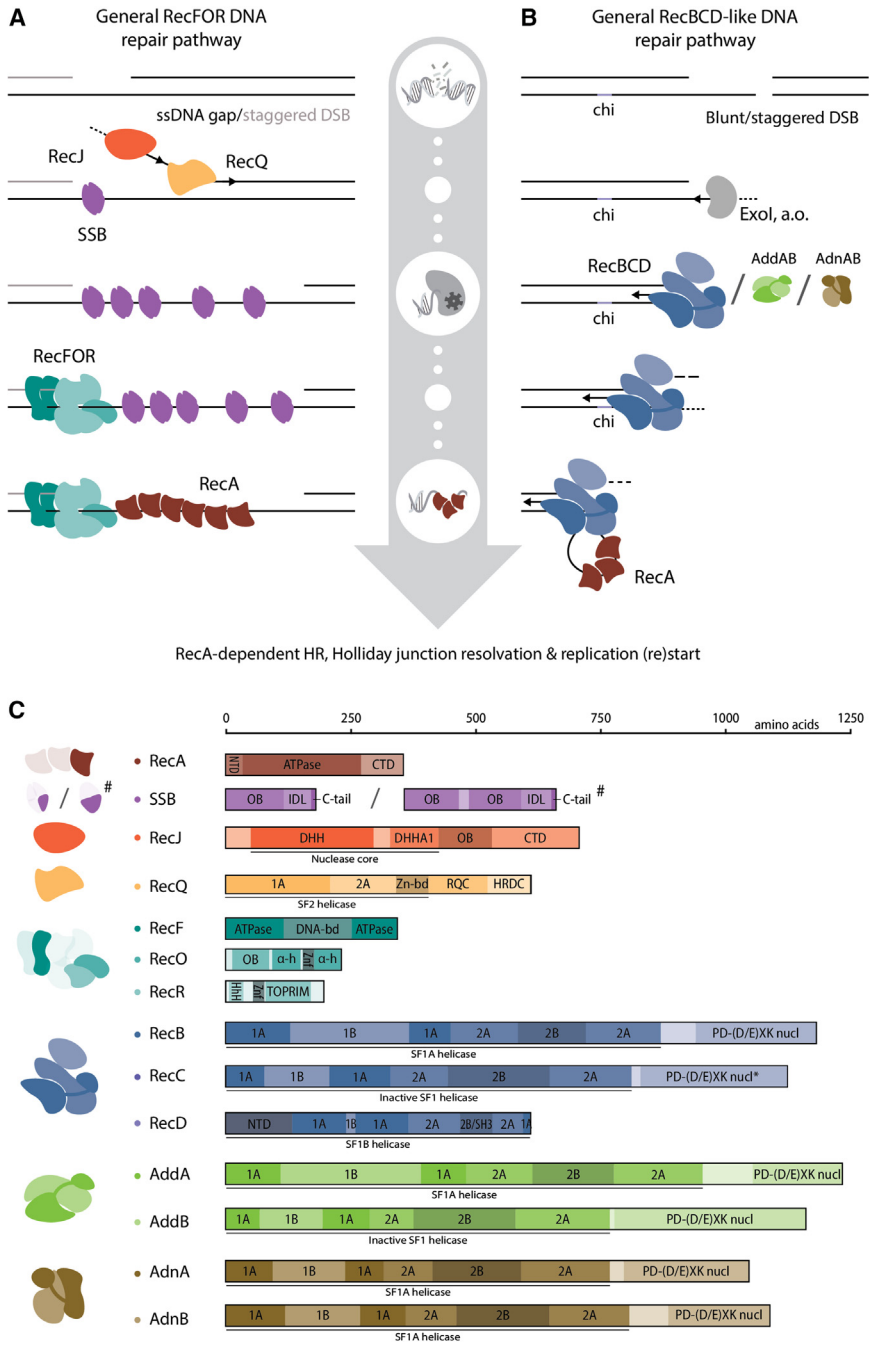


Figure 1. Start of homologous recombination pathways in bacteria and domain architecture of proteins involved

(A) The general RecFOR repair pathway as described for *T. thermophilus*, *D. radiodurans*, and *E. coli*. RecQ helicase and RecJ nuclease widen the ssDNA gap or ssDNA overhang at a ssDNA gap or DSB, respectively. SSB binds exposed ssDNA. The RecFOR complex binds ssDNA-dsDNA junctions, or RecOR binds SSB-coated ssDNA, and replaces SSB with RecA.

(B) The general RecBCD-like pathway as described for *E. coli*. Blunt-ended DSBs are processed by nucleases like ExoI to form blunt-ended DSBs. The RecBCD complex (or AddAB in *B. subtilis* or AdnAB in *M. smegmatis*) assembles on the blunt-ended DSB and processes the dsDNA by unwinding and cleavage of ssDNA. Once a chi site is bound, cleavage of the chi-site-containing strand is inhibited, after which that strand is looped out, and RecA is loaded onto it. In Rec(F)OR, RecBCD, AddAB, and AdnAB pathways, RecA-loaded ssDNA forms a synaptic filament that can invade a dsDNA template. Upon base pairing between complementary strands, a (double) Holliday junction is formed. RuvABC resolves the Holliday junction, after which DNA polymerase fills in the ssDNA gaps.

(C) Domain architecture of proteins that initiate homologous recombination in bacteria. NTD, N-terminal domain; CTD, C-terminal domain; OB fold, oligonucleotide-binding fold; IDL, intrinsically disordered linker; C-tail, C-terminal tail; DHHA1, DHH associated 1; 1A/2A, RecA-like domains 1/2; 1B/2B, domains inserted in RecA-like domains 1A/2A; SF1A/SF1B/SF2 helicase, superfamily 1/2 A/B helicases; Zn-bd, zinc-binding domain; RQC, RecQ C-terminal domain; HRDC, helicase and RNaseD C-terminal domain; DNA-bd, DNA binding domain; α-h, α helical domain; Znf, zinc finger; HhH, helix-hairpin-helix domain; TOPRIM, topoisomerase-primase domain; PD-(D/E)XK nucl, nuclease domain with conserved catalytic motif PD-(D/E)XK; SH3, SRC homology domain 3. *, catalytically inactive. #In most bacterial clades, SSB contains one OB fold and forms a homotetramer, whereas in the Deinococcota, SSB contains two OB folds and forms homodimers.[30] See also Figure S1 for catalytic residues and motifs of archetype proteins.

by ssDNA-binding protein (SSB).[35,37] After the DNA replication machinery stalls at these ssDNA gaps, the RecFOR complex assembles on the dsDNA/ssDNA transition[38,39]: a RecF dimer assembles on the dsDNA, a tetrameric RecR ring clamps the ssDNA, and RecO binds ssDNA on the RecF-distal side of the RecR ring.[33,37] In the absence of RecF, a RecOR complex can also assemble in the middle of an ssDNA gap or on a staggered DSB *in vitro*, where the RecF dimer is replaced by an additional RecO monomer.[40] The bound Rec(F)OR complex interacts with SSB and stimulates loading of RecA onto SSB-coated ssDNA.[41]

Beyond repairing ssDNA gaps, the RecFOR pathway also repairs DSBs when more efficient DSB repair pathways are absent.[36] In that case, RecQ and/or RecJ process blunt or staggered DSBs to generate stretches of ssDNA.[36,37]

The RecBCD pathway is typified by the RecBCD helicase-nuclease complex (Figure 1B). RecBCD binds blunt-end DSBs, which can be generated from staggered-end DSBs by nucleases such as RecJ, SbcCD, and ExoI.[7] The dsDNA is pulled through RecBCD by the helicase domains of RecB and RecD, while the RecB nuclease domain asymmetrically cleaves both unwound

DNA strands that exit the RecBCD complex.[42,43] RecBCD-mediated dsDNA degradation continues until RecC binds a chi (crossover hotspot instigator) site, a short sequence motif found in the genome. Chi site binding changes the activity of RecBCD: while degradation of the 5′ end DNA strand is enhanced, degradation of the 3′ end is attenuated, and the formed 3′ ssDNA strand is bulged out of the RecBCD complex.[44,45] RecA is loaded onto this 3′ ssDNA,[44] after which strand exchange takes place, and DNA repair is completed. It has recently been suggested that RecBCD also functions in ssDNA gap repair, blurring the lines between distinct DNA repair pathways.[46]

The AddAB pathway is typified by the AddAB helicase-nuclease complex (Figure 1C).[42] Akin to RecBCD, the AddAB complex recognizes and processes blunt-end DSBs, recognizes a (distinct) chi site, and generates a 3′ ssDNA strand onto which RecA is loaded.[47,48] Although the proteins that comprise the RecBCD and AddAB complexes are homologous and function in a similar fashion, major differences between the complexes exist (Figure 1C): RecBCD consists of three subunits (RecB, RecC, and RecD), in which RecB is a helicase-nuclease, RecC is a helicase-nuclease of which both domains are inactive, and RecD is a helicase (and lacks a nuclease domain).[44] In contrast, AddAB consists of two subunits only (AddA and AddB), each of which contains a helicase domain (inactive in AddB) and an active nuclease domain.[48] Furthermore, RecBCD unwinds the DNA strands at a different rate (causing ssDNA loop formation in front of the complex), while AddAB translocates both DNA strands at the same rate and therefore requires SSB to prevent cleavage product reannealing.[42,45,47,49] The AdnAB repair complex is closely homologous to AddAB.[50] However, in contrast to AddAB, in AdnAB, both subunits (AdnA and AdnB) contain functional helicase and nuclease domains.[42] Akin to RecBCD, AdnAB does not require SSB for processive strand unwinding, but no chi site has been identified.[51–54]

Beyond DNA repair, RecBCD and AddAB have also been implicated in prokaryotic immunity.[44,55–64] RecBCD can directly degrade phage linear dsDNA genomes or replication intermediates that have exposed dsDNA ends.[55,56] As such, phages have developed various inhibitors to counteract RecBCD activity.[65–68] Furthermore, the products of RecBCD- and/or AddAB-mediated invader DNA degradation result in the formation of DNA degradation products that indirectly or directly guide other prokaryotic immune systems, including CRISPR-Cas systems[57–59] and prokaryotic Argonaute (pAgo) proteins.[60–64] This suggests that these immune systems rely on DNA repair complexes, that they act in conjunction, or that synergistic effects between these DNA repair complexes and immune systems exist. As such, analyses of the co-occurrence of DNA repair proteins and these and other prokaryotic immune systems might shed light on their functional co-dependence and the mechanisms underlying prokaryotic immunity. However, the high degree of homology between DNA repair proteins or domains thereof (Figure 1C) makes the accurate identification of DNA repair proteins challenging. While previous studies have shed light on the distribution of proteins involved in HR,[69–73] we found that available TIGR/HMM profiles do not accurately distinguish different DNA repair enzymes, hindering subsequent analyses.

Here, we describe a unique search strategy to identify and classify bacterial DNA repair proteins, involving new HMM profiles with query sequences from diverse phyla, followed by reverse HMM score-based filtering. We used this stringent search strategy to identify DNA repair proteins in the RefSeq database, which facilitated accurate analyses of their taxonomic distribution, (co-)occurrence, and genomic clustering over the bacterial phyla. Furthermore, by identifying prokaryotic immune systems using DefenseFinder and analyzing their co-occurrence with DNA repair proteins, we confirm existing associations with CRISPR-Cas systems and find many novel correlations including pAgos, dGTPases, GAPS2, and Wadjet. As such, this work provides an updated view on the distribution of DNA repair proteins in bacteria, extends our insights into their genomic clustering, and sheds light on the co-occurrence between DNA repair proteins and prokaryotic immune systems.

RESULTS

Accurate identification and classification of DNA repair proteins

To perform phylogenetic and taxonomic distribution analyses of DNA repair proteins, accurate identification and classification are required. In earlier studies, several HMM profiles have been developed to identify the homologs that belong to specific protein families.[70,71,74] Using these publicly available HMM profiles, we attempted to identify DNA repair proteins in the representative scaffold and whole genomes of the RefSeq database (accessed on January 16, 2021, 7,249 genomes). However, the presence of certain DNA repair proteins appeared highly overestimated (e.g., 27,589 RecB hits for TIGR00609, >3 copies per genome; Table S1). Furthermore, multiple proteins were identified by multiple individual HMM profiles. This indicates that search constraints are too lenient for accurate identification and classification, probably due to the high degree of homology and similar domain architecture of distinct DNA repair proteins (Figure 1C). To generate new HMM profiles enabling more accurate identification, for each DNA repair protein, we selected $\geq$10 sequences (11–17) from highly different phylogenetic groups (Figure 2A). However, HMM searches using the newly constructed HMM profiles still resulted in the overrepresentation of various proteins (e.g., 27,846 hits for RecB, >3 copies per genome; Figure 2B), and >20,000 proteins were identified by multiple HMM profiles (e.g., 90,698 proteins were identified as RecB, RecD, AddA, AddB, AdnA, and AdnB; Figure S2A). This underscores the challenge of classifying highly homologous DNA repair proteins. These results imply that further filtering is required to facilitate accurate identification of DNA repair proteins.

To obtain accurate classification of the DNA repair proteins, for each protein identified, reciprocal HMM searches were performed using the new HMM profiles developed (Figure 2A). Using the reciprocal search strategy, query proteins were first identified using the different HMM search profiles (i.e., for each type of DNA repair protein, an HMM profile exists, and each of these profiles is used to identify a pool of DNA repair proteins). Subsequently, for each identified DNA repair protein, an HMM score was generated for each of the HMM search profiles (i.e., for

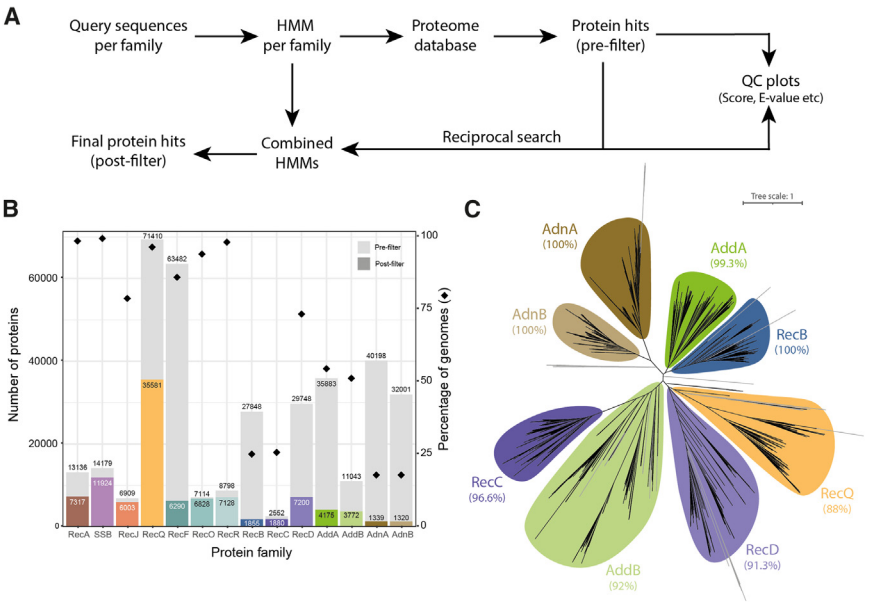


Figure 2. Accurate identification of DNA repair proteins and phylogenetic validation of helicases

(A) Schematic representation of the pipeline used for DNA repair protein identification and classification. 11–17 sequences per protein family were selected from diverse bacterial phyla to construct HMM profiles. These profiles were used to search the proteomes from 7,249 RefSeq genomes. The obtained (overlapping; pre-filter) hits were subjected to a reciprocal search using our collection of HMM profiles, after which proteins were classified using the highest scoring HMM profile (post-filter).

(B) Number of proteins identified per protein family pre-filter (gray) and post-filter (colored, primary y axis). Diamonds (◆, secondary y axis) indicate the percentage of genomes that contain the corresponding protein (post-filter).

(C) Unrooted phylogenetic tree of the SF1 and SF2 helicase domains of DNA repair proteins RecQ, RecB, RecC, RecD, AddA, AddB, AdnA, and AdnB. Clades are labeled and colored according to the protein classifying this clade, and the percentages indicate the proteins correctly classified in this clade. Proteins with classifications that do not follow the phylogeny are colored gray. See also Figure S2 for proteins identified by multiple HMM profiles and Figure S3 for a midpoint-rooted tree with protein annotations and corresponding HMM scores.

each type of DNA repair protein). After that, for each DNA repair protein, the different HMM scores were ranked, and the protein was assigned the identification of the HMM search profile that generated the highest HMM score.

The reciprocal HMM searches facilitated a classification of DNA repair proteins for which no overlapping classification exists (Figures 2B and S2B). RecA and SSB are present in ~99% of the genomes (Figure 2B). These findings underscore the central role of these SSBs in different DNA repair pathways.[26,37,42] Similarly, RecJ and RecQ are present in 79% and 97% of genomes, respectively (Figure 2B). The proteins RecO and RecR are identified in >94% of all genomes, whereas RecF is identified in 86% of all genomes (Figure 2B). In line with the *in vitro* activity of RecOR without RecF,[75] this implies that RecF is not essential in Rec(F)OR pathways. RecB and RecC are found in 25%–26% of all bacterial genomes, while RecD is found in 73% (Figure 2B). This confirms previous analyses that most RecD proteins have a function outside the context of RecBCD.[76,77] Finally, AddA and AddB are found in 55% and 51% of genomes, whereas AdnA and AdnB are present in 18% of genomes (Figure 2B).

To verify the classification and obtain insights about the shared ancestry of DNA repair proteins, we performed phylogenetic analysis of the SF1 helicase domain-containing DNA repair proteins (RecB, RecC, RecD, AddA, AddB, AdnA, AdnB) and SF2 helicase RecQ (Figure 1C). 96.3% of proteins cluster in distinct phylogenetic clades according to their classification, which endorses the accuracy of our method (Figures 2C and S3). The small fraction of proteins for which the assigned classification does not match the phylogeny generally have a low score for all reverse HMM searches and form (sub)clades with long branch lengths, which indicates they represent distantly related proteins (Figures 2C and S3). While increasing the cutoff score could limit these inaccurate identifications, it would also lower the number of correctly identified proteins. The phylogeny reveals that the AddA-RecB, AddB-RecC, and AdnA-AdnB pairs have common ancestors (Figure 2C),[41] in contrast to an earlier study in which it was suggested that AddAB and RecBCD proteins evolved from AdnAB proteins.[78] RecD and RecQ form individual clades, and within the RecQ clade, there is a clear distinction of two subclades (Figure 2C). The current identification and classification strategy allows for further analysis of the distribution of these DNA repair proteins in bacteria.

Global taxonomic distribution of DNA repair proteins

To obtain a clear image of the taxonomic distribution of DNA repair proteins across the bacterial phyla, we analyzed the abundance of DNA repair proteins in the different bacterial phyla (Figure 3). We limited our analysis to phyla with ≥6 representative genomes in the representative scaffold and whole genomes of the RefSeq database and pooled the remaining genomes as "other." In accordance with their high abundance, RecA, SSB, and RecQ are present in all bacterial phyla (Figure 3; Table S2). All three RecFOR proteins are present in most bacterial clades (Figure 3), while Acidobacteriota, Aquificota, Mycoplasmatota, Nitrospirota, Planctomycetota, and Thermodesulfobacteriota mostly encode only RecOR and lack RecF. RecFOR proteins are almost completely absent in Thermotogota (which form a separate clade between Gracilicutes and Terrabacteria[79,80]). Given that Thermotogota undergo extensive DNA exchange[81,82] and that certain Thermotogota species are naturally competent,[83] a more divergent Rec(F)OR pathway or another analogous DNA repair system might exist in Thermotogota.

Despite RecBCD being arguably the most-studied bacterial HR complex,[42–45,68,84] RecB and RecC are (almost) completely absent in 12 out of 21 phyla included in our analysis, including bacteria belonging to the Terrabacteria and DST (Deinococcota,

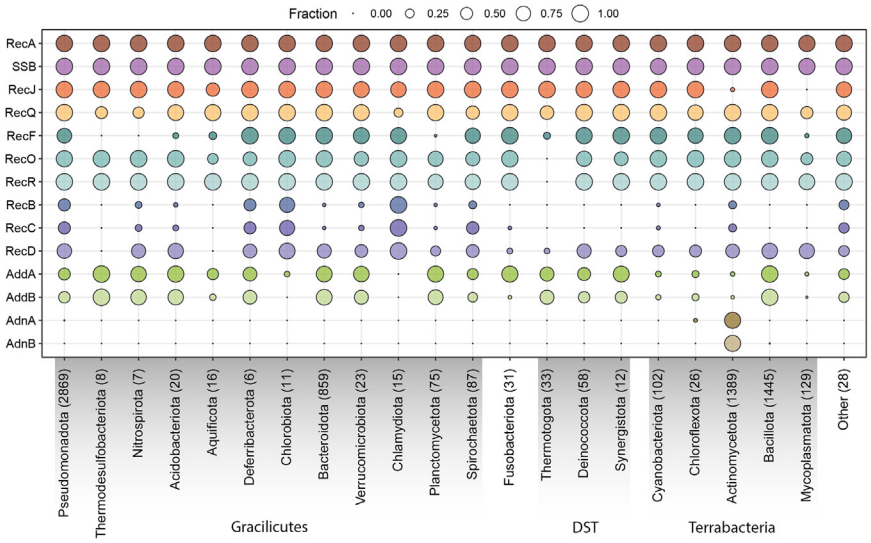


Figure 3. Taxonomic distribution of DNA repair proteins

Bubble plot indicating the fraction of genomes that encode specific DNA repair proteins per bacterial phyla (ordered according to bacterial phylogeny[79,80]). Phyla with <6 genomes were grouped as "other."[79,80] Numbers next to the phyla names indicate the number of genomes present in the total dataset of 7,249 genomes.

Synergistota, and Thermotogota) groups. RecBCD is present in >50% of bacteria that belong to the Gracilicute group Chlamydiota, Chlorobiota, Deferribacterota, and Pseudomonadota phyla (Figure 3; Table S2). In the remaining five Gracilicute phyla, RecB and RecC are present in ≤20% of genomes. Several bacterial phyla that do not encode RecBC do encode RecD (e.g., Bacillota, Chloroflexota, Deinococcota, and Mycoplasmatota), which indicates the existence of the RecD variant "RecD2," as previously indicated[76,77] (Figure 3).

In general, AddAB is (sometimes sparsely) distributed over all phyla except Chlamydiota, and AddA is slightly more widespread than AddB (Figure 3). In line with previous literature, AdnA and AdnB are 99.96% specific to Actinomycetota.[71] Various phyla, including the Terrabacteria clades Chloroflexota, Cyanobacteriota, and Mycoplasmatota, encode (almost) no RecBC, AddAB, and/or AndAB complexes. RecJ is found ubiquitously in all phyla except Actinomycetota and Mycoplasmatota, which do encode RecOR (Mycoplasmatota) or RecFOR and AdnAB (in Actinomycetota) (Figure 3). This implies that HR occurs independently of RecJ in these phyla or that they encode more divergent RecJ homologs or analogous proteins with a similar function. Combined, our analysis reveals phyla-specific patterns of DNA repair proteins, of which certain findings invite further exploration.

Co-occurrence and co-encoding of DNA repair proteins

Rec(F)OR, RecBCD, AddAB, and AdnAB form complexes in certain bacterial species.[37,42,43,47,51] Principal-components analysis (PCA) confirms that proteins of the DNA repair complexes RecFOR, RecBC, AddAB, and AdnAB are distributed accordingly (Figure 4A). To understand if co-occurrence (or exclusion) also exists between these and other DNA repair proteins, the co-occurrence of all DNA repair proteins was investigated (Figures 4B and 4C). The presence of Rec(F)OR-coding genes is strongly correlated, with representation of all three proteins in 83% of the genomes, while 11% of the genomes encode RecOR but not RecF (Figures 4B and 4C). Also, the presence of AdnA/AdnB, AddA/AddB, and RecB/RecC protein combinations is strongly correlated (Figure 4B). As RecD is often encoded in genomes independently of RecBC (in 65% of the genomes), it does not cluster with RecBC in the PCA (Figure 4A), and the correlation for RecD co-occurrence with RecBC is weak (Figure 4B). However, RecBC is rarely found in the absence of RecD (Figure 4C).

The presence of AddAB is negatively correlated with both of the other DSB repair complexes, RecBC(D) and AdnAB (Figures 4B and 4C), which suggests that they are mutually exclusive. As expected due to their high abundance, proteins that occur in >90% of genomes (SSB, RecA, RecR, RecQ, RecO) mostly show weak correlations with other proteins (Figure 4B). However, the presence of RecQ is positively correlated with the presence of RecFOR and RecBCD proteins, while a slightly negative correlation is observed for the presence of RecQ and AddAB proteins (Figure 4B). While a functional link exists between RecQ and RecFOR (Figure 1A), to our knowledge, no such links have been established between RecQ and RecBCD. The negative correlation of RecQ with AddAB could be attributed to the complementary roles of RecQ and RecJ in processing DSBs (followed by the RecFOR pathway)[36,85]; the presence of RecJ is positively correlated with AddAB (Figure 4B), which suggests that AddAB and RecJ function in conjunction. Combined, these results confirm the co-occurrence of known DNA repair proteins (Rec(F)OR, RecBCD, AddAB, and AdnAB) and reveal co-occurrence patterns that have not been identified before.

In bacteria, genes encoding proteins that function in conjunction sometimes cluster in the genome, for example in operons.[86] To determine if the genomic clustering of genes encoding DNA repair proteins is conserved, we analyzed the genomic distance between the genes encoding complex-forming DNA repair proteins (Figure 4D). Despite their linked functionality and omnipresence, Rec(F)OR-coding genes rarely cluster (Figures 4D and S4), as described previously.[69] In contrast, 90% of RecBCD-coding genes are found in clusters, mostly in Pseudomonadota, Deferribacterota, and Chlorobiota (Figures 4D and 4E). In contrast, in Chlamydiota, RecBCD-coding genes are present in all studied genomes (Figure 3) but rarely clustered (Figure 4D). AddA- and AddB-coding genes cluster in 87% of AddAB-encoding genomes (Figure 4D) distributed across most phyla (Figure 4E), whereas RecJ, the occurrence of which is strongly positively correlated with AddAB (Figure 4B), is rarely encoded in proximity

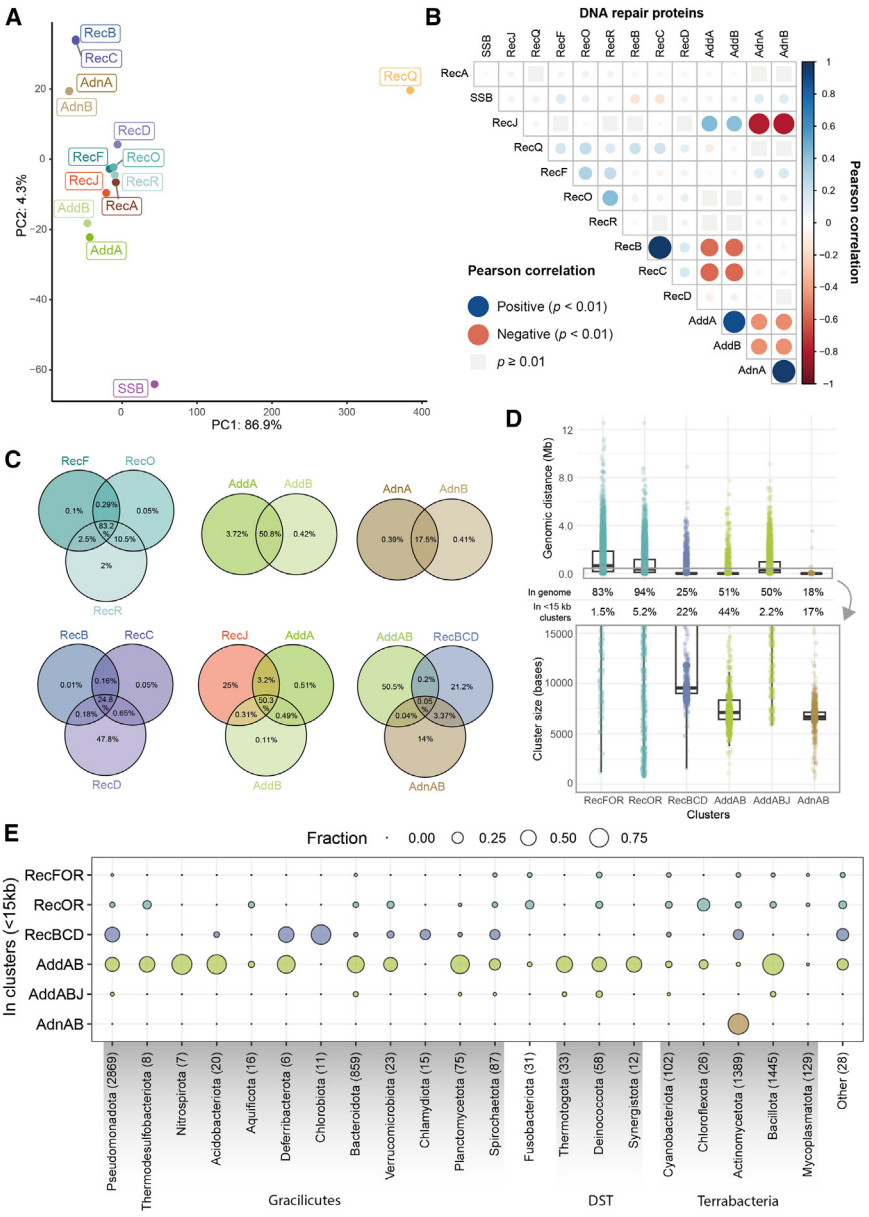


Figure 4. Co-occurrence and genomic clustering of DNA repair proteins

(A) PCA plot indicating similarity in the distribution of DNA repair proteins across 7,249 bacterial genomes with two most-important PCs (PC1: 86.9% and PC2: 4.3%). Proteins close together indicate a similar overall distribution of these proteins.

(B) Pairwise Pearson correlation of the presence of each DNA repair protein across 7,249 bacterial genomes. Red/blue circles represent negative/positive correlations, with size and color intensity indicating correlation strength. Gray squares indicate insignificant correlations ($p \geq 0.01$).

(C) Venn diagrams indicating the percentage of genomes having co-occurrence of DNA repair proteins for RecFOR, RecBCD, AddAB (with and without RecJ), and AdnAB. In addition, a Venn diagram indicating the co-occurrence of complete AddAB, RecBCD, and AdnAB genomic clusters is shown. Percentages indicate the genomes with the specified proteins compared to the total number of genomes (7,249) considered.

(D) Genomic distance between genes encoding co-occurring DNA repair proteins. If three genes are considered (RecFOR, RecBCD, AddABJ), then the maximum distance is shown. Genes are considered clustered when the maximum distance is <15 kb (see bottom). See also Figure S4 for each combination and individual cluster sizes.

(E) Taxonomic distribution of DNA repair protein clusters <15 kb in bacterial phyla (ordered according to bacterial phylogeny[79,80]). Numbers next to the phyla names indicate the number of genomes present in the total dataset of 7,249 genomes. Phyla with less than 6 genomes are grouped as "other."

to AddA or AddB (Figure 4D). The Actinomycetota AdnAB pair exists in clusters in 97% of the AdnAB-encoding genomes (Figures 4D and 4E). Together, these results signify that repair proteins involved in a complex or functional in the same pathway are found in a cluster or putative operon (AddAB, AdnAB, and RecBCD) or are encoded independently (Rec(F)OR).

DNA repair proteins and genome size

Bacterial genomes widely vary in genome size, with bacteria from certain phyla having smaller genomes than the others (Figure S5). There is a strict correlation between genome size and proteome size (Pearson correlation coefficient [R] = 0.99; Figure 5A), and it has previously been established that a positive correlation also exists between genome size and the number of prokaryotic immune systems encoded.[87,88] To determine if a correlation exists between genome size and the number of DNA repair proteins encoded, we compared the genome size to the total number of DNA repair proteins. While the number of DNA repair proteins encoded increases up to a genome size of ~5 Mb, it plateaus at ~15–16 DNA repair proteins encoded in larger genomes (Figure 5B). RecQ is diverse and overrepresented in our dataset (with 35,581 copies identified in 6,998 genomes, >5 copies per genome; Figures 2B and 2C). Therefore, we hypothesized that removing RecQ from the analysis would clarify the analysis. When RecQ is excluded, the number of DNA repair proteins plateaus at 9–10 in genomes of ~2.5 Mb and larger (Figure 5B). Pearson correlation analysis confirms that a positive correlation between genome size and DNA repair system abundance exists in genomes <2.5 Mb (R = 0.56) but that this correlation is lost in genomes >2.5 Mb (R = 0.069). This suggests that larger genomes do not necessarily require a larger set of canonical DNA repair proteins to maintain genome integrity.

Next, we analyzed whether specific DNA repair proteins are more abundant in genomes within a specific genome size range

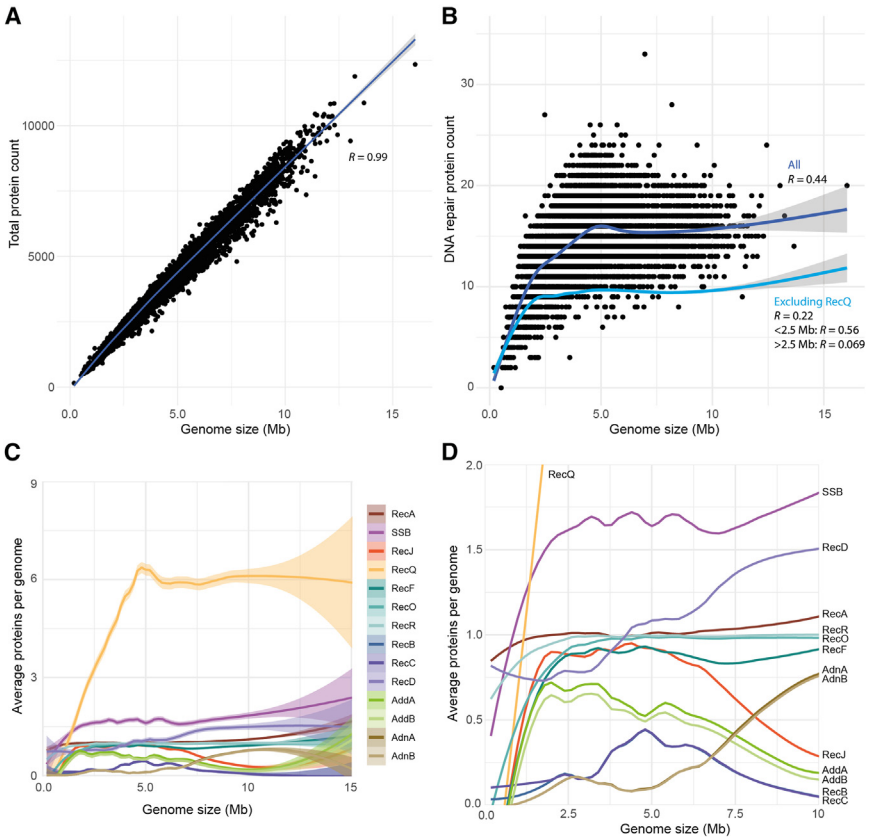


Figure 5. DNA repair proteins and genome size

(A) The total number of proteins in a genome in relation to genome size. The line indicates the average with a 95% confidence interval (shaded). Pearson correlation coefficient $R = 0.99$.

(B) The total number of DNA repair proteins in a genome in relation to genome size. The dark blue line indicates the average with 95% confidence interval (shaded) and $R = 0.44$. The light blue line indicates the average of all proteins excluding RecQ. Pearson correlation calculated for all genome sizes, $R = 0.22$; for genomes <2.5 Mb, $R = 0.56$; and for genomes >2.5 Mb, $R = 0.069$.

(C) The number of individual DNA repair proteins in a genome in relation to genome size. The lines indicate averages with 95% confidence intervals (shaded) of individual DNA repair proteins.

(D) Close-up of (C) but excluding confidence intervals. See also Figure S5 for the distribution of genome sizes per bacterial phylum.

(Figures 5C and 5D). Whereas genomes <1 Mb encode RecA, SSB, and RecR, the other DNA repair proteins only sporadically occur in these small genomes (Figures 5C and 5D). Furthermore, AddAB is most abundant in genomes that are ~1.5–3.5 Mb, whereas RecBCD is most abundant in genomes that are ~3.5–6.5 Mb in size (Figure 5D). In line with Actinomycetota having relatively large genomes (Figure S5), AdnAB is most often found in genomes >5 Mb. This suggests that specific proteins (RecA, SSB, and RecR) are essential and selected for even in small genomes, while DSB repair complexes become more important in genomes with a size beyond 1 Mb (Figure 5D). Furthermore, while the few genomes encoding multiple DSB repair complexes (RecBCD, AddAB, and AdnAB) are large genomes (Figure S5; Table S2), larger genomes do not necessarily encode multiple DNA repair complexes, possibly because their activities would compete with each other.

Co-occurrence of DNA repair proteins and prokaryotic immune systems

Various prokaryotic immune systems, which protect prokaryotes against mobile genetic elements, including phages and plasmids, function in conjunction with DNA repair proteins.[57–59,63,65,89,90] This includes retron Ec48, which senses (phage-mediated) RecBCD inhibition and triggers abortive infection,[65] and type I-E, I-F, II-A, and III-A CRISPR-Cas systems, whose spacer acquisition is enhanced by RecBCD or AddAB.[57–59,89] In addition to CRISPR-Cas, the products of RecBCD- or AddAB-mediated dsDNA degradation products are used by pAgo systems as small ssDNA guides.[62,63]

If CRISPR-Cas and pAgo systems, and possibly other immune systems, indeed function in conjunction with DNA repair proteins, then this suggests that correlations between their occurrences could exist. Indeed, previous bioinformatics analyses confirm that the presence of certain CRISPR-Cas subtypes correlates with DNA repair proteins.[71] While such analyses might also expose putative functional relations between other immune systems and DNA repair proteins, such analyses have not yet been extended to other prokaryotic immune systems. Here, we use our classification of DNA repair proteins to investigate their co-occurrences with immune systems as identified by DefenseFinder in whole-genome species representatives in the RefSeq database (>2,300 genomes; Figure 6A).[91]

For the co-occurrence of DNA repair proteins and CRISPR-Cas systems, several positive correlations ($R > 0.3$, $p < 0.001$) can be observed (Figure 6B; Data S1A). The presence of type I-E CRISPR-Cas systems is positively correlated with AdnAB ($R > 0.41$) and negatively correlated with the presence of AddAB ($R < -0.25$) and RecJ ($R = 0.36$; Figure 6B). This suggests that although RecBCD and RecJ play a role in naive and primed spacer acquisition in the *E. coli* type I-E CRISPR-Cas system,[57,89] this is not necessarily the case in other bacteria (Figure 6B; Data S1A). We hypothesize that in bacteria lacking RecBCD and/or RecJ, other nucleases contribute to spacer acquisition, for example ExoVII or AdnAB. In contrast to *E. coli* RecBCD, chi site recognition has not been shown for AdnAB, which suggests that it is not essential for CRISPR adaptation.[57,89,92] The presence of type I-F CRISPR-Cas systems is positively correlated with that of RecB and RecC ($R = 0.48$ for both) and negatively correlated with that of AddAB ($R < -0.26$; Figure 6B). This suggests that RecBC might play a role in spacer acquisition for type I-F CRISPR-Cas systems.

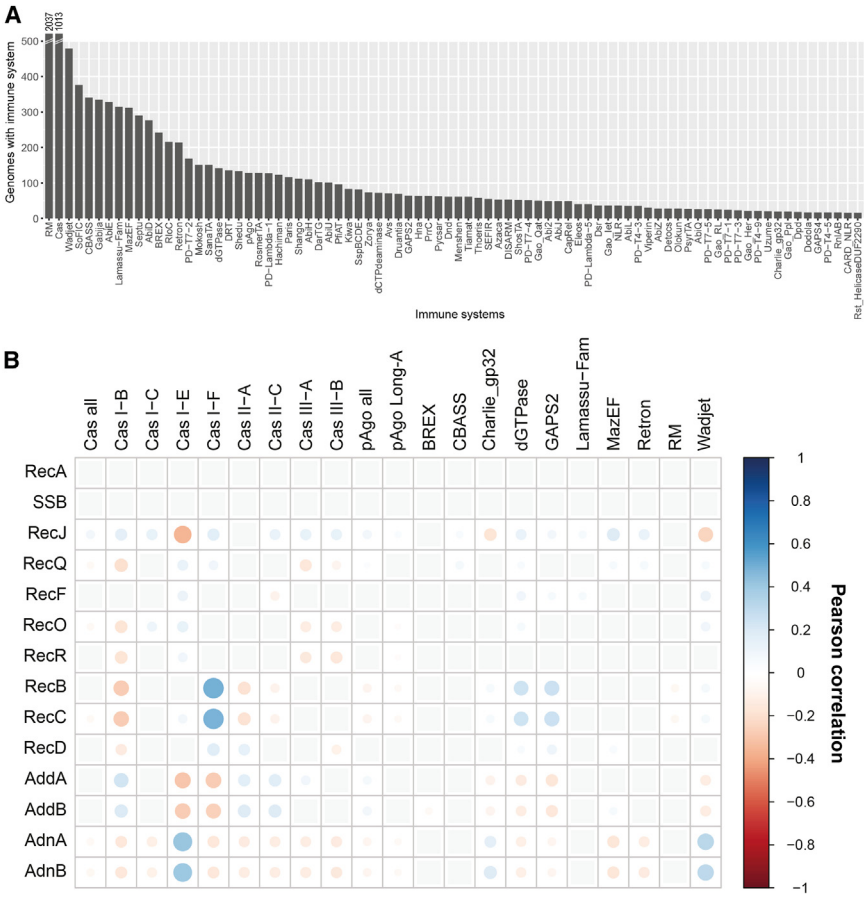


Figure 6. Co-occurrence of DNA repair proteins and bacterial immune systems

(A) Immune systems identified in our dataset using DefenseFinder.[87,91] Systems identified in <15 genomes are not shown.

(B) Pairwise Pearson correlations for selected co-occurring DNA repair proteins and immune systems. Red/blue circles represent negative/positive correlations, with size and color intensity indicating correlation strength. Gray squares indicate insignificant correlations ($p \geq 0.01$). For absolute co-occurrences and more correlations, see also Data S1A (CRISPR-Cas subtypes), Data S1B (pAgo subtypes), and Data S1C (all immune systems). For a full list showing what genomes contain which DNA repair proteins and immune systems, see Table S3.

The presence of type II-A CRISPR-Cas systems is positively correlated with the presence of AddAB ($R > 0.16$; Figure 6B), and their functional co-dependance is confirmed by the *Staphylococcus aureus* type II-A CRISPR-Cas system, which requires AddAB for efficient spacer acquisition.[58] Similarly, spacer acquisition by the *Staphylococcus epidermidis* type III-A CRISPR-Cas system is also enhanced by AddAB,[59] in line with the positive correlation of the presence of AddA with type III-A CRISPR systems ($R = 0.09$; Figure 6B). When considering each phylum separately, the overall co-occurrences get stronger, and additional phyla-specific co-occurrences become apparent (Data S1A). For example, in Pseudomonadota, positive correlations between the presence of RecF and type I-B and I-F CRISPR-Cas systems and RecO with type I-F CRISPR-Cas systems exist, while in Bacillota, a negative correlation between the presence of AddB and type III-B CRISPR-Cas systems exists (Data S1A). Combined, these results reveal that there is no strict co-dependence of CRISPR-Cas subtypes on distinct DNA repair proteins, but they function in conjunction with specific preferences within phyla.

Despite RecBCD and AddAB having been implicated in generating small DNA guides for certain pAgos from the long-A pAgo clade,[62,63] a previous analysis suggests that RecBCD and AddAB are present in only 47% of genomes encoding catalytically active long-A pAgos.[63] In line with that observation, no moderate/strong correlations ($R > 0.10$) between the presence of any of the pAgo subtypes (including long A) and DNA repair proteins are observed across all phyla (Figure 6B; Data S1B). Also within phyla, no strong correlations (based on system presence in >10 genomes) are observed (Data S1B). Instead, a weak positive correlation between the presence of pAgos with RecJ, AddA, and AddB exists ($R = 0.09$, 0.09, and 0.08, respectively), and a weak negative correlation between the presence of pAgos with RecBC and AdnAB exists (all $R = -0.08$) (Figure 6B). This suggests that while (long-A) pAgos might be able to utilize the products of DNA repair proteins as guides,[62,63] they are functionally not strictly dependent on them.

Beyond these correlations, the analysis uncovers hitherto unreported (but mostly weak) correlations for the presence of DNA repair proteins and Wadjet, MazEF, dGTPases, GAPS2, Charlie_gp32, and more (Figure 6B; Data S1C). Wadjet is an SMC protein complex that senses DNA topology (small circular vs. linear/long circular) and a nuclease that cleaves upon complex stalling.[93,94] In line with previous observations that Wadjet is typically found in Actinomycetota,[91] its presence is positively correlated with the presence of AdnAB ($R \geq 0.30$) and negatively correlated with RecJ ($R = -0.25$; Figure 6B). The toxin-antitoxin system MazEF, which a.o. senses DNA damage,[95] is positively correlated with RecJ ($R = 0.18$) and negatively correlated with AdnAB ($R = -0.16$). The dGTPases, which deplete cellular dGTP and are suggested to slow down phage replication,[96] are weakly positively correlated with the presence of RecBC ($R = 0.24$), RecJ ($R = 0.14$), and RecF ($R = 0.09$) and weakly negatively correlated with AddAB ($R \geq -0.12$; Figure 6B). Also, the occurrence of the GAPS2 system is positively correlated with RecB and RecC ($R \geq 0.25$) and negatively correlated with AddA and AddB ($R \geq -0.16$; Figure 6B). GAPS2 defends against phages through an unknown mechanism[97] and contains a BRCT (BRCA1 C terminus) domain, which is generally involved

in DNA repair, recombination, cell cycle control, and protein-protein interactions.[98,99] Also, for various other immune systems of which the mechanisms are unknown, a positive correlation with the presence of one or more DNA repair proteins is observed. This includes AbiH, SoFIC, AbiJ, AbiN, Azaca, Druantia, Gao_Qat, NLR, and PD-Lambda1. It is possible that their functional mechanisms are linked to the activity of DNA repair proteins. Furthermore, 44 out of 64 GAPS2 systems were found together with a dGTPase (R = 0.44; Data S1C), which suggests possible synergy between these systems. These and other correlations provide leads for further investigation into synergies between immune systems and DNA repair proteins (Figure 6B; Data S1C).

DISCUSSION

Bacteria encode a wide array of DNA repair proteins, but which DNA repair proteins co-occur and/or are mutually exclusive has remained largely unclear, possibly due to challenges in accurate classification. In this study, we present an enhanced method for the classification of DNA repair proteins through reverse search score classification from various phylogenetic groups (Figure 2). This has facilitated in-depth analysis of the taxonomic distribution (Figure 3), co-occurrence, and genomic clustering of DNA repair proteins (Figure 4) and their abundance in relation to genome size (Figure 5). Finally, co-occurrence analysis between DNA repair proteins and prokaryotic immune systems was performed (Figure 6).

By analyzing the abundance of DNA repair proteins across non-redundant genome databases, we show that RecBCD is less conserved than previously thought,[69–71] as it occurs only (and often sporadically) in 9 out of 21 of the studied bacterial phyla (Figure 3). Most other bacteria encode either AddAB (most common) or AdnAB (specific to Actinomycetota). Yet, bacteria from other clades, particularly bacteria from various Terrabacterial clades, do not encode any of these canonical DSB repair complexes. The small genome size of Mycoplasmatota (many intracellular symbionts[100]) and Aquificota (Figure S5) could explain the lack of these DSB repair complexes, as was described previously.[72] However, this does not explain the absence of DSB repair complexes in Cyanobacteriota and Chloroflexota (Figures 3 and S5).[101,102] Possibly, in these species, RecA self-loading[103] is sufficient to stimulate HR without auxiliary proteins, or possibly other proteins mediate DSB repair, for example RecQ/RecS and RecJ together with RecFOR.[36] We can also not rule out that certain bacteria contain alternative DNA repair proteins that facilitate DSB repair, for example more remote homologs of RecBCD/AddAB (akin to AdnAB in Actinomycetota) or non-related DNA repair proteins with analogous functionality.

Our analysis confirms the co-distribution (Figures 4B and 4C) and clustering (Figure 4D) of proteins that form the DSB repair complexes AddAB, AdnAB, and RecBCD (although RecD also exists outside the context of RecBC as stand-alone RecD2[76]). Although RecO and RecR occur more frequently with each other than with RecF (consistent with RecOR activity not strictly relying on RecF), neither RecFOR nor RecOR was found to cluster on the genome. Instead, RecFOR proteins have been shown to be part of operons that also encode proteins involved in replication (e.g., DnaA-DnaN-RecF and RecR-DnaX).[69] This aligns with recent insights into the spatiotemporal localization of RecFOR and RecF with the replisome, which is suggested to guide the RecFOR assembly toward ssDNA gaps resulting from replication.[31,38,39,104,105] Although RecJ is implicated in RecFOR pathways (Figure 1A), there is no positive correlation between the presence of RecJ and any of the RecFOR components (Figure 4A). Instead, a positive correlation between the presence of RecJ and AddAB (but not with AdnAB) is observed, indicating a versatile role for RecJ in DNA repair alongside its involvement in mismatch repair, base excision repair, and other DNA repair pathways.[106–108]

Exploration of the co-occurrence of bacterial immune systems and DNA repair systems confirms earlier observed correlations (e.g., for CRISPR-Cas systems and RecBCD[71]; Data S1A) but also reveals numerous novel correlations (Figure 6B; Data S1), which suggests that these DNA repair proteins and immune systems might function in conjunction. While pAgo proteins might be guided by RecBCD/AddAB-generated DNA fragments, we found no strong correlations between long-A pAgos and RecBCD or AddAB (Figure 6B). This suggests that DNA-guided pAgos might also rely on other mechanisms, including chopping,[64,109,110] or other nucleases that degrade invader DNA. Of note, Cyanobacteriota lack RecBCD and AddAB, but many Cyanobacterial long-A pAgos are co-encoded with putative Cas4 family nucleases (part of the PD-(D/E)XK nuclease superfamily[111] also encompassing AddAB, RecBC, and AdnAB; Figure 1C).[112,113]

The presence of various other immune systems, including dGTPases, GAPS2, Wadjet, MazEF, and Charlie_gp32, is positively correlated with DNA repair proteins (Figure 6B). While it remains unknown whether any functional relevance underlies these correlations, these and other co-occurrences (e.g., those observed for CRISPR-Cas systems) are never strict. Therefore, in support of the notion that prokaryotic immune systems are generally extensively transferred between bacteria and archaea from distinct phylogenetic clades,[114,115] we conclude that immune systems never exclusively rely on specific DNA repair proteins but that, in certain species, the immune and DNA repair might function in conjunction.

Beyond DNA repair proteins being important for maintaining genome integrity and function as immune systems, mobile genetic elements exploit or inhibit DNA repair proteins for natural transformation or recombination.[116–118] As such, DNA repair proteins influence the spread of mobile genetic elements and horizontal gene transfer and facilitate genome diversification. Given that prokaryotic immune systems are regularly transferred by horizontal gene transfer,[114,115] the observed associations of immune systems with DNA repair proteins could also be the result of enhanced horizontal gene transfer in that host or limited horizontal gene transfer in others. Finally, DNA repair proteins have been repurposed for genome engineering techniques.[119–121] As such, understanding the distribution of recombinatorial DNA repair systems can improve our knowledge of gene dissemination and might contribute to enhancing genome engineering techniques.

Limitations of the study

Our method accurately identifies commonly studied DNA repair proteins that initiate HR and can be extended to other DNA repair

proteins for a more comprehensive view of the DNA repair landscape. However, our analysis is limited by the diversity of available whole-genome sequences in the RefSeq database. As a result, in our analyses, Pseudomonadota, Actinomycetota, and Bacillota are overrepresented. Analyses on curated databases with a set of genomes that are equally distributed could facilitate a fairer distribution analyses.

We would like to emphasize that functional redundancy between DNA repair proteins might exist; a specific immune system might function in conjunction with RecBCD in one bacterium but with AddAB in another. Furthermore, correlations involving epistatic groups (e.g., RecBC and AddAB) may not always imply functional (in)compatibility; alternatively, such correlations could result from the presence/absence of another protein with functional (in)compatibility.

Certain correlations observed might be caused by a skewed phyletic distribution. This is, for example, the case for various correlations involving AdnAB observed in Actinomycetota. This could mean that the presence of AdnAB and certain immune systems are correlated simply because they are limited to the same phylum. However, since DNA replication proteins, and especially immune systems, are often horizontally transferred,[79,80,114,115] these restricted correlations could imply a functional dependence.

Additionally, analyzing data at the phylum level increases the strength and significance of correlations between the presence of DNA repair proteins and immune systems, but it reduces the overall biological relevance of these correlations. These co-occurrences will not only depend on the specific genetic context (i.e., what other proteins are encoded), but also the ecological environment of the species should be considered (e.g., intracellular symbionts might be less exposed to invading DNA and/or less exposed to DNA-damaging agents). As such, the correlations discovered in our analysis provide a solid *in silico* basis that will require further experimental investigation to reveal the putative mechanisms underlying the observed correlations.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Daan C. Swarts (daan.swarts@wur.nl).

Materials availability

This study did not generate physical materials.

Data and code availability

- Correlation data and other data underlying the figures are publicly available through GitHub: https://github.com/sumanthmutte/DNArepairProteins.
- All original code is publicly available through GitHub: https://github.com/sumanthmutte/DNArepairProteins.
- Any additional information required to reanalyze the data reported in this work is available from the lead contact upon request.

ACKNOWLEDGMENTS

We thank members of the Swarts lab and the Laboratory of Biochemistry at Wageningen University for valuable discussion. We thank Aude Bernheim for insightful discussion of preliminary data. This work was supported by a European Research Council (ERC) starting grant (ERC-2020-STG 948783) to D.C.S.

AUTHOR CONTRIBUTIONS

Conceptualization, P.B., P.B.U., and D.C.S.; methodology and investigation, S.K.M.; formal analysis, S.K.M. and P.B.; writing – original draft, S.K.M. and P.B.; writing – review & editing, S.K.M., P.B., and D.C.S.; supervision and funding acquisition, D.C.S.

DECLARATION OF INTERESTS

S.K.M. is the founder and CEO of MyGen Informatics.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- METHOD DETAILS
  - Protein selection, HMM profiles and database search for homologs
  - Filtering criteria and homolog selection
  - Phylogeny of helicase domains
  - Principal component analysis (PCA) and correlation of DNA repair proteins
  - Genomic clustering of gene families
  - Genome size correlations
  - Correlation of DNA repair proteins with bacterial immune systems
- QUANTIFICATION AND STATISTICAL ANALYSIS
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2024.115110.

Received: July 29, 2024
Revised: October 20, 2024
Accepted: December 3, 2024

REFERENCES

1. Hoeijmakers, J.H. (2001). Genome maintenance mechanisms for preventing cancer. Nature *411*, 366–374. https://doi.org/10.1038/35077232.
2. Lindahl, T. (1993). Instability and decay of the primary structure of DNA. Nature *362*, 709–715. https://doi.org/10.1038/362709a0.
3. Cadet, J., Berger, M., Douki, T., and Ravanat, J.-L. (1997). Oxidative damage to DNA: Formation, measurement, and biological significance. In Reviews of Physiology Biochemistry and Pharmacology, *131* (Springer Berlin Heidelberg), pp. 1–87. https://doi.org/10.1007/3-540-61992-5_5.
4. Gates, K.S. (2007). The Chemical Reactions of DNA Damage and Degradation. In Reviews of Reactive Intermediate Chemistry (Wiley)), pp. 333–378. https://doi.org/10.1002/9780470120828.ch8.
5. Rastogi, R.P., Richa, Kumar, A., Kumar, A., Tyagi, M.B., and Sinha, R.P. (2010). Molecular Mechanisms of Ultraviolet Radiation-Induced DNA Damage and Repair. J. Nucleic Acids *2010*, 592980. https://doi.org/10.4061/2010/592980.
6. Shuman, S., and Glickman, M.S. (2007). Bacterial DNA repair by non-homologous end joining. Nat. Rev. Microbiol. *5*, 852–861. https://doi.org/10.1038/nrmicro1768.
7. Sinha, A.K., Possoz, C., and Leach, D.R.F. (2020). The Roles of Bacterial DNA Double-Strand Break Repair Proteins in Chromosomal DNA

Replication. FEMS Microbiol. Rev. *44*, 351–368. https://doi.org/10.1093/femsre/fuaa009.

8. Morita, R., Nakane, S., Shimada, A., Inoue, M., Iino, H., Wakamatsu, T., Fukui, K., Nakagawa, N., Masui, R., and Kuramitsu, S. (2010). Molecular Mechanisms of the Whole DNA Repair System: A Comparison of Bacterial and Eukaryotic Systems. J. Nucleic Acids *2010*, 179594. https://doi.org/10.4061/2010/179594.
9. Goosen, N., and Moolenaar, G.F. (2008). Repair of UV damage in bacteria. DNA Repair *7*, 353–379. https://doi.org/10.1016/j.dnarep.2007.09.002.
10. Hsieh, P. (2001). Molecular mechanisms of DNA mismatch repair. Mutat. Res. *486*, 71–87. https://doi.org/10.1016/S0921-8777(01)00088-X.
11. Alonso, J.C., Cardenas, P.P., Sanchez, H., Hejna, J., Suzuki, Y., and Takeyasu, K. (2013). Early steps of double-strand break repair in Bacillus subtilis. DNA Repair *12*, 162–176. https://doi.org/10.1016/j.dnarep.2012.12.005.
12. Sfeir, A., Tijsterman, M., and McVey, M. (2024). Microhomology-Mediated End-Joining Chronicles: Tracing the Evolutionary Footprints of Genome Protection. Annu. Rev. Cell Dev. Biol. *40*, 195–218. https://doi.org/10.1146/annurev-cellbio-111822-014426.
13. Ithurbide, S., Bentchikou, E., Coste, G., Bost, B., Servant, P., and Sommer, S. (2015). Single Strand Annealing Plays a Major Role in RecA-Independent Recombination between Repeated Sequences in the Radioresistant Deinococcus radiodurans Bacterium. PLoS Genet. *11*, e1005636. https://doi.org/10.1371/journal.pgen.1005636.
14. Bertrand, C., Thibessard, A., Bruand, C., Lecointe, F., and Leblond, P. (2019). Bacterial NHEJ: a never ending story. Mol. Microbiol. *111*, 1139–1151. https://doi.org/10.1111/mmi.14218.
15. Brissett, N.C., and Doherty, A.J. (2009). Repairing DNA double-strand breaks by the prokaryotic non-homologous end-joining pathway. Biochem. Soc. Trans. *37*, 539–545. https://doi.org/10.1042/BST0370539.
16. Sharda, M., Badrinarayanan, A., and Seshasayee, A.S.N. (2020). Evolutionary and Comparative Analysis of Bacterial Nonhomologous End Joining Repair. Genome Biol. Evol. *12*, 2450–2466. https://doi.org/10.1093/gbe/evaa223.
17. McGovern, S., Baconnais, S., Roblin, P., Nicolas, P., Drevet, P., Simonson, H., Piétrement, O., Charbonnier, J.-B., Le Cam, E., Noirot, P., and Lecointe, F. (2016). C-terminal region of bacterial Ku controls DNA bridging, DNA threading and recruitment of DNA ligase D for double strand breaks repair. Nucleic Acids Res. *44*, 4785–4806. https://doi.org/10.1093/nar/gkw149.
18. Wyman, C., Ristic, D., and Kanaar, R. (2004). Homologous recombination-mediated double-strand break repair. DNA Repair *3*, 827–833. https://doi.org/10.1016/j.dnarep.2004.03.037.
19. Wiktor, J., Gynnå, A.H., Leroy, P., Larsson, J., Coceano, G., Testa, I., and Elf, J. (2021). RecA finds homologous DNA by reduced dimensionality search. Nature *597*, 426–429. https://doi.org/10.1038/s41586-021-03877-6.
20. Soppa, J. (2017). Polyploidy and community structure. Nat. Microbiol. *2*, 16261. https://doi.org/10.1038/nmicrobiol.2016.261.
21. Jain, K., Wood, E.A., Romero, Z.J., and Cox, M.M. (2021). RecA-independent recombination: Dependence on the *Escherichia coli* RarA protein. Mol. Microbiol. *115*, 1122–1137. https://doi.org/10.1111/mmi.14655.
22. Lovett, S.T., Hurley, R.L., Sutera, V.A., Aubuchon, R.H., and Lebedeva, M.A. (2002). Crossing Over Between Regions of Limited Homology in *Escherichia coli* : RecA-Dependent and RecA-Independent Pathways. Genetics *160*, 851–859. https://doi.org/10.1093/genetics/160.3.851.
23. Lovett, S.T. (2017). Template-switching during replication fork repair in bacteria. DNA Repair *56*, 118–128. https://doi.org/10.1016/j.dnarep.2017.06.014.
24. Yang, H., Zhou, C., Dhar, A., and Pavletich, N.P. (2020). Mechanism of strand exchange from RecA–DNA synaptic and D-loop structures. Nature *586*, 801–806. https://doi.org/10.1038/s41586-020-2820-9.
25. Chen, Z., Yang, H., and Pavletich, N.P. (2008). Mechanism of homologous recombination from the RecA–ssDNA/dsDNA structures. Nature *453*, 489–494. https://doi.org/10.1038/nature06971.
26. Liu, J., Ehmsen, K.T., Heyer, W.-D., and Morrical, S.W. (2011). Presynaptic filament dynamics in homologous recombination and DNA repair. Crit. Rev. Biochem. Mol. Biol. *46*, 240–270. https://doi.org/10.3109/10409238.2011.576007.
27. Hiom, K. (2009). DNA Repair: Common Approaches to Fixing Double-Strand Breaks. Curr. Biol. *19*, R523–R525. https://doi.org/10.1016/j.cub.2009.06.009.
28. Liu, Y., and West, S.C. (2004). Happy Hollidays: 40th anniversary of the Holliday junction. Nat. Rev. Mol. Cell Biol. *5*, 937–944. https://doi.org/10.1038/nrm1502.
29. Nautiyal, A., and Thakur, M. (2024). Prokaryotic DNA Crossroads: Holliday Junction Formation and Resolution. ACS Omega *9*, 12515–12538. https://doi.org/10.1021/acsomega.3c09866.
30. Oliveira, M.T., and Ciesielski, G.L. (2021). The Essential, Ubiquitous Single-Stranded DNA-Binding Proteins. In Methods in Molecular Biology (Humana Press Inc.), pp. 1–21. https://doi.org/10.1007/978-1-0716-1290-3_1.
31. Henrikus, S.S., Henry, C., Ghodke, H., Wood, E.A., Mbele, N., Saxena, R., Basu, U., van Oijen, A.M., Cox, M.M., and Robinson, A. (2019). RecFOR epistasis group: RecF and RecO have distinct localizations and functions in *Escherichia coli*. Nucleic Acids Res. *47*, 2946–2965. https://doi.org/10.1093/nar/gkz003.
32. Pham, P., Wood, E.A., Cox, M.M., and Goodman, M.F. (2023). RecA and SSB genome-wide distribution in ssDNA gaps and ends in *Escherichia coli*. Nucleic Acids Res. *51*, 5527–5546. https://doi.org/10.1093/nar/gkad263.
33. Chaudhary, S.K., Elayappan, M., Jeyakanthan, J., and Kanagaraj, S. (2020). Structural and functional characterization of oligomeric states of proteins in RecFOR pathway. Int. J. Biol. Macromol. *163*, 943–953. https://doi.org/10.1016/j.ijbiomac.2020.07.062.
34. Satoh, K., Kikuchi, M., Ishaque, A.M., Ohba, H., Yamada, M., Tejima, K., Onodera, T., and Narumi, I. (2012). The role of Deinococcus radiodurans RecFOR proteins in homologous recombination. DNA Repair *11*, 410–418. https://doi.org/10.1016/j.dnarep.2012.01.008.
35. Morimatsu, K., and Kowalczykowski, S.C. (2003). RecFOR Proteins Load RecA Protein onto Gapped DNA to Accelerate DNA Strand Exchange. Mol. Cell *11*, 1337–1347. https://doi.org/10.1016/S1097-2765(03)00188-6.
36. Morimatsu, K., and Kowalczykowski, S.C. (2014). RecQ helicase and RecJ nuclease provide complementary functions to resect DNA for homologous recombination. Proc. Natl. Acad. Sci. USA *111*, E5133–E5142. https://doi.org/10.1073/pnas.1420009111.
37. Nirwal, S., Czarnocki-Cieciura, M., Chaudhary, A., Zajko, W., Skowronek, K., Chamera, S., Figiel, M., and Nowotny, M. (2023). Mechanism of RecF–RecO–RecR cooperation in bacterial homologous recombination. Nat. Struct. Mol. Biol. *30*, 650–660. https://doi.org/10.1038/s41594-023-00967-z.
38. Henry, C., Mbele, N., and Cox, M.M. (2023). RecF protein targeting to postreplication (daughter strand) gaps I: DNA binding by RecF and RecFR. Nucleic Acids Res. *51*, 5699–5713. https://doi.org/10.1093/nar/gkad311.
39. Henry, C., Kaur, G., Cherry, M.E., Henrikus, S.S., Bonde, N.J., Sharma, N., Beyer, H.A., Wood, E.A., Chitteni-Pattu, S., van Oijen, A.M., et al. (2023). RecF protein targeting to post-replication (daughter strand) gaps II: RecF interaction with replisomes. Nucleic Acids Res. *51*, 5714–5742. https://doi.org/10.1093/nar/gkad310.
40. Radzimanowski, J., Dehez, F., Round, A., Bidon-Chanal, A., McSweeney, S., and Timmins, J. (2013). An 'open' structure of the RecOR complex supports ssDNA binding within the core of the complex. Nucleic Acids Res. *41*, 7972–7986. https://doi.org/10.1093/nar/gkt572.

41. Umezu, K., Chi, N.W., and Kolodner, R.D. (1993). Biochemical interaction of the Escherichia coli RecF, RecO, and RecR proteins with RecA protein and single-stranded DNA binding protein. Proc. Natl. Acad. Sci. USA *90*, 3875–3879. https://doi.org/10.1073/pnas.90.9.3875.

42. Wigley, D.B. (2013). Bacterial DNA repair: recent insights into the mechanism of RecBCD, AddAB and AdnAB. Nat. Rev. Microbiol. *11*, 9–13. https://doi.org/10.1038/nrmicro2917.

43. Singleton, M.R., Dillingham, M.S., Gaudier, M., Kowalczykowski, S.C., and Wigley, D.B. (2004). Crystal structure of RecBCD enzyme reveals a machine for processing DNA breaks. Nature *432*, 187–193. https://doi.org/10.1038/nature02988.

44. Dillingham, M.S., and Kowalczykowski, S.C. (2008). RecBCD Enzyme and the Repair of Double-Stranded DNA Breaks. Microbiol. Mol. Biol. Rev. *72*, 642–671. https://doi.org/10.1128/MMBR.00020-08.

45. Gaydar, V., Zananiri, R., Saied, L., Dvir, O., Kaplan, A., and Henn, A. (2024). Communication between DNA and nucleotide binding sites facilitates stepping by the RecBCD helicase. Nucleic Acids Res. *52*, 3911–3923. https://doi.org/10.1093/nar/gkae108.

46. Pagès, V. (2016). Single-strand gap repair involves both RecF and RecBCD pathways. Curr. Genet. *62*, 519–521. https://doi.org/10.1007/s00294-016-0575-5.

47. Saikrishnan, K., Yeeles, J.T., Gilhooly, N.S., Krajewski, W.W., Dillingham, M.S., and Wigley, D.B. (2012). Insights into Chi recognition from the structure of an AddAB-type helicase-nuclease complex. EMBO J. *31*, 1568–1578. https://doi.org/10.1038/emboj.2012.9.

48. Krajewski, W.W., Fu, X., Wilkinson, M., Cronin, N.B., Dillingham, M.S., and Wigley, D.B. (2014). Structural basis for translocation by AddAB helicase–nuclease and its arrest at χ sites. Nature *508*, 416–419. https://doi.org/10.1038/nature13037.

49. Yeeles, J.T.P., van Aelst, K., Dillingham, M.S., and Moreno-Herrero, F. (2011). Recombination Hotspots and Single-Stranded DNA Binding Proteins Couple DNA Translocation to DNA Unwinding by the AddAB Helicase-Nuclease. Mol. Cell *42*, 806–816. https://doi.org/10.1016/j.molcel.2011.04.012.

50. Sinha, K.M., Unciuleac, M.-C., Glickman, M.S., and Shuman, S. (2009). AdnAB: a new DSB-resecting motor–nuclease from mycobacteria. Genes Dev. *23*, 1423–1437. https://doi.org/10.1101/gad.1805709.

51. Jia, N., Unciuleac, M.C., Xue, C., Greene, E.C., Patel, D.J., and Shuman, S. (2019). Structures and single-molecule analysis of bacterial motor nuclease AdnAB illuminate the mechanism of DNA double-strand break resection. Proc. Natl. Acad. Sci. USA *116*, 24507–24516. https://doi.org/10.1073/pnas.1913546116.

52. Gupta, R., Unciuleac, M.-C., Shuman, S., and Glickman, M.S. (2017). Homologous recombination mediated by the mycobacterial AdnAB helicase without end resection by the AdnAB nucleases. Nucleic Acids Res. *45*, 762–774. https://doi.org/10.1093/nar/gkw1130.

53. Wang, B.-B., Xu, J.-Z., Zhang, F., Liu, S., Liu, J., and Zhang, W.-G. (2022). Review of DNA repair enzymes in bacteria: With a major focus on AddAB and RecBCD. DNA Repair *118*, 103389. https://doi.org/10.1016/j.dnarep.2022.103389.

54. Unciuleac, M.-C., and Shuman, S. (2010). Double Strand Break Unwinding and Resection by the Mycobacterial Helicase-Nuclease AdnAB in the Presence of Single Strand DNA-binding Protein (SSB). J. Biol. Chem. *285*, 34319–34329. https://doi.org/10.1074/jbc.M110.162925.

55. Benzinger, R., Enquist, L.W., and Skalka, A. (1975). Transfection of Escherichia coli spheroplasts. V. Activity of recBC nuclease in rec+ and rec minus spheroplasts measured with different forms of bacteriophage DNA. J. Virol. *15*, 861–871. https://doi.org/10.1128/jvi.15.4.861-871.1975.

56. Behme, M.T., Lilley, G.D., and Ebisuzaki, K. (1976). Postinfection control by bacteriophage T4 of Escherichia coli recBC nuclease activity. J. Virol. *18*, 20–25. https://doi.org/10.1128/jvi.18.1.20-25.1976.

57. Levy, A., Goren, M.G., Yosef, I., Auster, O., Manor, M., Amitai, G., Edgar, R., Qimron, U., and Sorek, R. (2015). CRISPR adaptation biases explain preference for acquisition of foreign DNA. Nature *520*, 505–510. https://doi.org/10.1038/nature14302.

58. Modell, J.W., Jiang, W., and Marraffini, L.A. (2017). CRISPR–Cas systems exploit viral DNA injection to establish and maintain adaptive immunity. Nature *544*, 101–104. https://doi.org/10.1038/nature21719.

59. Aviram, N., Thornal, A.N., Zeevi, D., and Marraffini, L.A. (2022). Different modes of spacer acquisition by the *Staphylococcus epidermidis* type III-A CRISPR-Cas system. Nucleic Acids Res. *50*, 1661–1672. https://doi.org/10.1093/nar/gkab1299.

60. Esyunina, D., Okhtienko, A., Olina, A., Panteleev, V., Prostova, M., Aravin, A.A., and Kulbachinskiy, A. (2023). Specific targeting of plasmids with Argonaute enables genome editing. Nucleic Acids Res. *51*, 4086–4099. https://doi.org/10.1093/nar/gkad191.

61. Huang, S., Wang, K., and Mayo, S.L. (2023). Genome manipulation by guide-directed Argonaute cleavage. Nucleic Acids Res. *51*, 4078–4085. https://doi.org/10.1093/nar/gkad188.

62. Jolly, S.M., Gainetdinov, I., Jouravleva, K., Zhang, H., Strittmatter, L., Bailey, S.M., Hendricks, G.M., Dhabaria, A., Ueberheide, B., and Zamore, P.D. (2020). Thermus thermophilus Argonaute Functions in the Completion of DNA Replication. Cell *182*, 1545–1559.e18. https://doi.org/10.1016/j.cell.2020.07.036.

63. Kuzmenko, A., Oguienko, A., Esyunina, D., Yudin, D., Petrova, M., Kudinova, A., Maslova, O., Ninova, M., Ryazansky, S., Leach, D., et al. (2020). DNA targeting and interference by a bacterial Argonaute nuclease. Nature *587*, 632–637. https://doi.org/10.1038/s41586-020-2605-1.

64. Bobadilla Ugarte, P., Barendse, P., and Swarts, D.C. (2023). Argonaute proteins confer immunity in all domains of life. Curr. Opin. Microbiol. *74*, 102313. https://doi.org/10.1016/j.mib.2023.102313.

65. Millman, A., Bernheim, A., Stokar-Avihail, A., Fedorenko, T., Voichek, M., Leavitt, A., Oppenheimer-Shaanan, Y., and Sorek, R. (2020). Bacterial Retrons Function In Anti-Phage Defense. Cell *183*, 1551–1561.e12. https://doi.org/10.1016/j.cell.2020.09.065.

66. Murphy, K.C. (1991). Lambda Gam protein inhibits the helicase and chi-stimulated recombination activities of Escherichia coli RecBCD enzyme. J. Bacteriol. *173*, 5808–5821. https://doi.org/10.1128/jb.173.18.5808-5821.1991.

67. Murphy, K.C. (2000). Bacteriophage P22 Abc2 protein binds to RecC increases the 5′ strand nicking activity of RecBCD and together with λ Bet, promotes Chi-independent recombination 1 1Edited by M. Gottesman. J. Mol. Biol. *296*, 385–401. https://doi.org/10.1006/jmbi.1999.3486.

68. Wilkinson, M., Wilkinson, O.J., Feyerherm, C., Fletcher, E.E., Wigley, D.B., and Dillingham, M.S. (2022). Structures of RecBCD in complex with phage-encoded inhibitor proteins reveal distinctive strategies for evasion of a bacterial immunity hub. Elife *11*, e83409. https://doi.org/10.7554/eLife.83409.

69. Rocha, E.P.C., Cornet, E., and Michel, B. (2005). Comparative and Evolutionary Analysis of the Bacterial Homologous Recombination Systems. PLoS Genet. *1*, e15. https://doi.org/10.1371/journal.pgen.0010015.

70. Cromie, G.A. (2009). Phylogenetic Ubiquity and Shuffling of the Bacterial RecBCD and AddAB Recombination Complexes. J. Bacteriol. *191*, 5076–5084. https://doi.org/10.1128/JB.00254-09.

71. Bernheim, A., Bikard, D., Touchon, M., and Rocha, E.P.C. (2019). A matter of background: DNA repair pathways as a possible cause for the sparse distribution of CRISPR-Cas systems in bacteria. Philos. Trans. R. Soc. Lond. B Biol. Sci. *374*, 20180088. https://doi.org/10.1098/rstb.2018.0088.

72. Garcia-Gonzalez, A., Vicens, L., Alicea, M., and Massey, S.E. (2013). The distribution of recombination repair genes is linked to information content in bacteria. Gene *528*, 295–303. https://doi.org/10.1016/j.gene.2013.05.082.

73. Gurung, D., and Blumenthal, R.M. (2020). Distribution of RecBCD and AddAB recombination-associated genes among bacteria in 33 phyla. Microbiology (N. Y.) *166*, 1047–1064. https://doi.org/10.1099/mic.0.000980.

74. Li, W., O'Neill, K.R., Haft, D.H., DiCuccio, M., Chetvernin, V., Badretdin, A., Coulouris, G., Chitsaz, F., Derbyshire, M.K., Durkin, A.S., et al. (2021). RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. Nucleic Acids Res. *49*, D1020–D1028. https://doi.org/10.1093/nar/gkaa1105.

75. Sakai, A., and Cox, M.M. (2009). RecFOR and RecOR as Distinct RecA Loading Pathways. J. Biol. Chem. *284*, 3264–3272. https://doi.org/10.1074/jbc.M807220200.

76. Montague, M., Barnes, C., Smith, H.O., Chuang, R.-Y., and Vashee, S. (2009). The Evolution of RecD Outside of the RecBCD Complex. J. Mol. Evol. *69*, 360–371. https://doi.org/10.1007/s00239-009-9290-x.

77. Ramos, C., Hernández-Tamayo, R., López-Sanz, M., Carrasco, B., Serrano, E., Alonso, J.C., Graumann, P.L., and Ayora, S. (2022). The RecD2 helicase balances RecA activities. Nucleic Acids Res. *50*, 3432–3444. https://doi.org/10.1093/nar/gkac131.

78. Unciuleac, M.-C., and Shuman, S. (2010). Characterization of the Mycobacterial AdnAB DNA Motor Provides Insights into the Evolution of Bacterial Motor-Nuclease Machines. J. Biol. Chem. *285*, 2632–2641. https://doi.org/10.1074/jbc.M109.076133.

79. Coleman, G.A., Davín, A.A., Mahendrarajah, T.A., Szánthó, L.L., Spang, A., Hugenholtz, P., Szöllősi, G.J., and Williams, T.A. (2021). A rooted phylogeny resolves early bacterial evolution. Science *372*, eabe0511. https://doi.org/10.1126/science.abe0511.

80. Williams, T.A., Davin, A.A., Morel, B., Szánthó, L.L., Spang, A., Stamatakis, A., Hugenholtz, P., and Szöllősi, G.J. (2023). The power and limitations of species tree-aware phylogenetics. Preprint at: bioRxiv. https://doi.org/10.1101/2023.03.17.533068

81. Zannoni, D., and De, R. (2014). In Microbial BioEnergy: Hydrogen Production, D. Zannoni and R. De Philippis, eds. (Netherlands: Springer). https://doi.org/10.1007/978-94-017-8554-9.

82. Nesbø, C.L., S Swithers, K., Dahle, H., Haverkamp, T.H.A., Birkeland, N.-K., Sokolova, T., Kublanov, I., and Zhaxybayeva, O. (2015). Evidence for extensive gene flow and *Thermotoga* subpopulations in subsurface and marine environments. ISME J. *9*, 1532–1542. https://doi.org/10.1038/ismej.2014.238.

83. Han, D., Xu, H., Puranik, R., and Xu, Z. (2014). Natural transformation of Thermotoga sp. strain RQ7. BMC Biotechnol. *14*, 39. https://doi.org/10.1186/1472-6750-14-39.

84. Cheng, K., Wilkinson, M., Chaban, Y., and Wigley, D.B. (2020). A conformational switch in response to Chi converts RecBCD from phage destruction to DNA repair. Nat. Struct. Mol. Biol. *27*, 71–77. https://doi.org/10.1038/s41594-019-0355-2.

85. Sanchez, H., Kidane, D., Castillo Cozar, M., Graumann, P.L., and Alonso, J.C. (2006). Recruitment of *Bacillus subtilis* RecN to DNA Double-Strand Breaks in the Absence of DNA End Processing. J. Bacteriol. *188*, 353–360. https://doi.org/10.1128/JB.188.2.353-360.2006.

86. Osbourn, A.E., and Field, B. (2009). Operons. Cell. Mol. Life Sci. *66*, 3755–3775. https://doi.org/10.1007/s00018-009-0114-3.

87. Tesson, F., Hervé, A., Mordret, E., Touchon, M., d'Humières, C., Cury, J., and Bernheim, A. (2022). Systematic and quantitative view of the antiviral arsenal of prokaryotes. Nat. Commun. *13*, 2561. https://doi.org/10.1038/s41467-022-30269-9.

88. Olijslager, L.H., Weijers, D., and Swarts, D.C. (2024). Distribution of specific prokaryotic immune systems correlates with host optimal growth temperature. NAR Genom. Bioinform. *6*, 105. https://doi.org/10.1093/nargab/lqae105.

89. Shiriaeva, A.A., Kuznedelov, K., Fedorov, I., Musharova, O., Khvostikov, T., Tsoy, Y., Kurilovich, E., Smith, G.R., Semenova, E., and Severinov, K. (2022). Host nucleases generate prespacers for primed adaptation in the *E. coli* type I-E CRISPR-Cas system. Sci. Adv. *8*, 8650. https://doi.org/10.1126/sciadv.abn8650.

90. Heussler, G.E., Miller, J.L., Price, C.E., Collins, A.J., and O'Toole, G.A. (2016). Requirements for Pseudomonas aeruginosa Type I-F CRISPR-Cas Adaptation Determined Using a Biofilm Enrichment Assay. J. Bacteriol. *198*, 3080–3090. https://doi.org/10.1128/JB.00458-16.

91. Tesson, F., Planel, R., Egorov, A.A., Georjon, H., Vaysset, H., Brancotte, B., Neron, B., Mordret, E., Atkinson, G.C., Bernheim, A., et al. (2024). A Comprehensive Resource for Exploring Antiphage Defense: DefenseFinder Webservice, Wiki and Databases. Preprint at: Cold Spring Harbor Laboratory. https://doi.org/10.1101/2024.01.25.577194

92. Radovčić, M., Killelea, T., Savitskaya, E., Wettstein, L., Bolt, E.L., and Ivančić-Baće, I. (2018). CRISPR–Cas adaptation in Escherichia coli requires RecBCD helicase but not nuclease activity, is independent of homologous recombination, and is antagonized by 5′ ssDNA exonucleases. Nucleic Acids Res. *46*, 10173–10183. https://doi.org/10.1093/nar/gky799.

93. Deep, A., Gu, Y., Gao, Y.-Q., Ego, K.M., Herzik, M.A., Zhou, H., and Corbett, K.D. (2022). The SMC-family Wadjet complex protects bacteria from plasmid transformation by recognition and cleavage of closed-circular DNA. Mol. Cell *82*, 4145–4159.e7. https://doi.org/10.1016/j.molcel.2022.09.008.

94. Shaltiel, I.A., Datta, S., Lecomte, L., Hassler, M., Kschonsak, M., Bravo, S., Stober, C., Ormanns, J., Eustermann, S., and Haering, C.H. (2022). A hold-and-feed mechanism drives directional DNA loop extrusion by condensin. Science *376*, 1087–1094. https://doi.org/10.1126/science.abm4012.

95. Ramisetty, B.C.M., Natarajan, B., and Santhosh, R.S. (2015). *mazEF*-mediated programmed cell death in bacteria: "What is this? Crit. Rev. Microbiol. *41*, 89–100. https://doi.org/10.3109/1040841X.2013.804030.

96. Tal, N., Millman, A., Stokar-Avihail, A., Fedorenko, T., Leavitt, A., Melamed, S., Yirmiya, E., Avraham, C., Brandis, A., Mehlman, T., et al. (2022). Bacteria deplete deoxynucleotides to defend against bacteriophage infection. Nat. Microbiol. *7*, 1200–1209. https://doi.org/10.1038/s41564-022-01158-0.

97. Mahata, T., Kanarek, K., Goren, M.G., Rameshkumar, M.R., Bosis, E., Qimron, U., and Salomon, D. (2023). Gamma-Mobile-Trio systems define a new class of mobile elements rich in bacterial defensive and offensive tools. Preprint at: Cold Spring Harbor Laboratory. https://doi.org/10.1101/2023.03.28.534373

98. Peña-Guerrero, J., Fernández-Rubio, C., García-Sosa, A.T., and Nguewa, P.A. (2023). BRCT Domains: Structure, Functions, and Implications in Disease—New Therapeutic Targets for Innovative Drug Discovery against Infections. Pharmaceutics *15*, 1839. https://doi.org/10.3390/pharmaceutics15071839.

99. Muhseena, N.,K., Mathukkada, S., Das, S.P., and Laha, S. (2021). The repair gene BACH1 - a potential oncogene. Oncol Rev *15*, 519. https://doi.org/10.4081/oncol.2021.519.

100. Rottem, S. (2002). Invasion of Mycoplasmas into and Fusion with Host Cells. In Molecular Biology and Pathogenicity of Mycoplasmas, S. Razin and R. Herrmann, eds. (Springer US), pp. 391–401. https://doi.org/10.1007/0-306-47606-1_17.

101. Whitton, B.A., and Potts, M. (2012). Introduction to the Cyanobacteria. In Ecology of Cyanobacteria II (Springer Netherlands), pp. 1–13. https://doi.org/10.1007/978-94-007-3855-3_1.

102. Speirs, L.B.M., Rice, D.T.F., Petrovski, S., and Seviour, R.J. (2019). The Phylogeny, Biodiversity, and Ecology of the Chloroflexi in Activated Sludge. Front. Microbiol. *10*, 2015. https://doi.org/10.3389/fmicb.2019.02015.

103. Galletto, R., Amitani, I., Baskin, R.J., and Kowalczykowski, S.C. (2006). Direct observation of individual RecA filaments assembling on single DNA molecules. Nature *443*, 875–878. https://doi.org/10.1038/nature05197.

104. Macián, F., Pérez-Roger, I., and Armengod, M.E. (1994). An improved vector system for constructing transcriptional lacZ fusions: analysis of regulation of the dnaA, dnaN, recF and gyrB genes of Escherichia coli. Gene *145*, 17–24. https://doi.org/10.1016/0378-1119(94)90317-4.

105. Kidane, D., Sanchez, H., Alonso, J.C., and Graumann, P.L. (2004). Visualization of DNA double-strand break repair in live bacteria reveals dynamic recruitment of *Bacillus subtilis* RecF, RecO and RecN proteins to distinct sites on the nucleoids. Mol. Microbiol. *52*, 1627–1639. https://doi.org/10.1111/j.1365-2958.2004.04102.x.

106. Cheng, K., Xu, Y., Chen, X., Lu, H., He, Y., Wang, L., and Hua, Y. (2020). Participation of RecJ in the base excision repair pathway of Deinococcus radiodurans. Nucleic Acids Res. *48*, 9859–9871. https://doi.org/10.1093/nar/gkaa714.

107. Harms, K., Schön, V., Kickstein, E., and Wackernagel, W. (2007). The RecJ DNase strongly suppresses genomic integration of short but not long foreign DNA fragments by homology-facilitated illegitimate recombination during transformation of *Acinetobacter baylyi*. Mol. Microbiol. *64*, 691–702. https://doi.org/10.1111/j.1365-2958.2007.05692.x.

108. Burdett, V., Baitinger, C., Viswanathan, M., Lovett, S.T., and Modrich, P. (2001). *In vivo* requirement for RecJ, ExoVII, ExoI, and ExoX in methyl-directed mismatch repair. Proc. Natl. Acad. Sci. USA *98*, 6765–6770. https://doi.org/10.1073/pnas.121183298.

109. Zander, A., Willkomm, S., Ofer, S., van Wolferen, M., Egert, L., Buchmeier, S., Stöckl, S., Tinnefeld, P., Schneider, S., Klingl, A., et al. (2017). Guide-independent DNA cleavage by archaeal Argonaute from Methanocaldococcus jannaschii. Nat. Microbiol. *2*, 17034. https://doi.org/10.1038/nmicrobiol.2017.34.

110. Swarts, D.C., Szczepaniak, M., Sheng, G., Chandradoss, S.D., Zhu, Y., Timmers, E.M., Zhang, Y., Zhao, H., Lou, J., Wang, Y., et al. (2017). Autonomous Generation and Loading of DNA Guides by Bacterial Argonaute. Mol. Cell *65*, 985–998.e6. https://doi.org/10.1016/j.molcel.2017.01.033.

111. Steczkiewicz, K., Muszewska, A., Knizewski, L., Rychlewski, L., and Ginalski, K. (2012). Sequence, structure and functional diversity of PD-(D/E)XK phosphodiesterase superfamily. Nucleic Acids Res. *40*, 7016–7045. https://doi.org/10.1093/nar/gks382.

112. Ryazansky, S., Kulbachinskiy, A., and Aravin, A.A. (2018). The Expanded Universe of Prokaryotic Argonaute Proteins. mBio *9*, e01935-18. https://doi.org/10.1128/mBio.01935-18.

113. Swarts, D.C., Makarova, K., Wang, Y., Nakanishi, K., Ketting, R.F., Koonin, E.V., Patel, D.J., and van der Oost, J. (2014). The evolutionary journey of Argonaute proteins. Nat. Struct. Mol. Biol. *21*, 743–753. https://doi.org/10.1038/nsmb.2879.

114. Koonin, E.V., Makarova, K.S., Wolf, Y.I., and Krupovic, M. (2020). Evolutionary entanglement of mobile genetic elements and host defence systems: guns for hire. Nat. Rev. Genet. *21*, 119–131. https://doi.org/10.1038/s41576-019-0172-9.

115. Rocha, E.P.C., and Bikard, D. (2022). Microbial defenses against mobile genetic elements and viruses: Who defends whom from what? PLoS Biol. *20*, e3001514. https://doi.org/10.1371/journal.pbio.3001514.

116. Frost, L.S., Leplae, R., Summers, A.O., and Toussaint, A. (2005). Mobile genetic elements: the agents of open source evolution. Nat. Rev. Microbiol. *3*, 722–732. https://doi.org/10.1038/nrmicro1235.

117. Bobay, L.-M., Touchon, M., and Rocha, E.P.C. (2013). Manipulating or Superseding Host Recombination Functions: A Dilemma That Shapes Phage Evolvability. PLoS Genet. *9*, e1003825. https://doi.org/10.1371/journal.pgen.1003825.

118. Choi, W., Jang, S., and Harshey, R.M. (2014). Mu transpososome and RecBCD nuclease collaborate in the repair of simple Mu insertions. Proc. Natl. Acad. Sci. USA *111*, 14112–14117. https://doi.org/10.1073/pnas.1407562111.

119. Zheng, L., Tan, Y., Hu, Y., Shen, J., Qu, Z., Chen, X., Ho, C.L., Leung, E.L.-H., Zhao, W., and Dai, L. (2022). CRISPR/Cas-Based Genome Editing for Human Gut Commensal *Bacteroides* Species. ACS Synth. Biol. *11*, 464–472. https://doi.org/10.1021/acssynbio.1c00543.

120. Vento, J.M., Crook, N., and Beisel, C.L. (2019). Barriers to genome editing with CRISPR in bacteria. J. Ind. Microbiol. Biotechnol. *46*, 1327–1341. https://doi.org/10.1007/s10295-019-02195-1.

121. Hoshijima, K., Jurynec, M.J., and Grunwald, D.J. (2016). Precise genome editing by homologous recombination. In Methods in Cell Biology (Academic Press), pp. 121–147. https://doi.org/10.1016/bs.mcb.2016.04.008.

122. O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. *44*, D733–D745. https://doi.org/10.1093/nar/gkv1189.

123. Katoh, K., and Standley, D.M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol. Biol. Evol. *30*, 772–780. https://doi.org/10.1093/molbev/mst010.

124. Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972–1973. https://doi.org/10.1093/bioinformatics/btp348.

125. Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol. Biol. Evol. *32*, 268–274. https://doi.org/10.1093/molbev/msu300.

126. Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., and Jermiin, L.S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods *14*, 587–589. https://doi.org/10.1038/nmeth.4285.

127. Letunic, I., and Bork, P. (2024). Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. Nucleic Acids Res. *52*, W78–W82. https://doi.org/10.1093/nar/gkae268.

128. Koopal, B., Potocnik, A., Mutte, S.K., Aparicio-Maldonado, C., Lindhoud, S., Vervoort, J.J.M., Brouns, S.J.J., and Swarts, D.C. (2022). Short prokaryotic Argonaute systems trigger cell death upon detection of invading DNA. Cell *185*, 1471–1486.e19. https://doi.org/10.1016/j.cell.2022.03.012.

129. Mao, X., Ma, Q., Liu, B., Chen, X., Zhang, H., and Xu, Y. (2015). Revisiting operons: an analysis of the landscape of transcriptional units in E. coli. BMC Bioinf. *16*, 356. https://doi.org/10.1186/s12859-015-0805-8.

130. Taboada, B., Estrada, K., Ciria, R., and Merino, E. (2018). Operon-mapper: a web server for precise operon identification in bacterial and archaeal genomes. Bioinformatics *34*, 4118–4120. https://doi.org/10.1093/bioinformatics/bty496.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
NCBI RefSeq database (downloaded on 16-01-2021)	O'Leary et al.[122]	https://www.ncbi.nlm.nih.gov/refseq/
DefenseFinder (data from run on RefSeq)	Tesson et al.[91]	https://defensefinder.mdmlab.fr/
Correlation data and other intermediate data underlying this study	This study	https://github.com/sumanthmutte/DNArepairProteins
Software and algorithms		
HMMER v3.3	Eddy et al.	http://hmmer.org/
Clustal Omega v1.2.4	EBI	https://www.ebi.ac.uk/jdispatcher/msa/clustalo
MAFFT v7.505	Katoh et al.[123]	https://mafft.cbrc.jp/alignment/software/
trimAl v1.4	Capella-Gutierrez et al.[124]	https://vicfero.github.io/trimal/
IQtree v2.2.0	Nguyen et al.[125]	http://www.iqtree.org/
ModelFinder	Kalyaanamoorthy et al.[126]	http://www.iqtree.org/ModelFinder/
iTOL v6	Letunic et al.[127]	https://itol.embl.de/
Base R and packages (tidyverse, UpSetR, ggplot2, ggrepel, corrplot, ggvenn)	R Core Team	https://www.r-project.org/
R scripts	This study	https://github.com/sumanthmutte/DNArepairProteins

METHOD DETAILS

Protein selection, HMM profiles and database search for homologs

For each of the 14 protein families of interest (AddA, AddB, AdnA, AdnB, RecA, RecB, RecC, RecD, RecF, RecO, RecR, RecJ, RecQ, and SSB) 11 to 17 protein sequences were collected from the NCBI RefSeq proteins database.[122] The selection process included the criteria that the proteins belonged to diverse phyla of bacteria, and preferably proteins were picked that have been experimentally verified. If not enough experimentally verified proteins could be selected, they were picked from phyla with many hits with previous HMM profiles (e.g., from Bernheim et al., 2019).[71] Protein identifiers along with species names, per protein family, are provided in the GitHub repository. One HMM profile for each protein family was generated with 'hmmbuild' in HMMER (hmmer.org; v3.3) using the sequence alignment made with default parameters of Clustal Omega (v1.2.4; https://www.ebi.ac.uk/jdispatcher/msa/clustalo). These HMM profiles were searched in the proteomes from NCBI RefSeq database (downloaded on 16-01-2021) with a pipeline search strategy that we developed earlier.[128] Refer to the GitHub repository for the details of 7249 genomes. Briefly, proteomes of 7249 genomes were searched for homologs using 'hmmsearch' (HMMER v3.3) with additional parameters "–notextw –E 0.001". We then extracted all the hits i.e., possible homologs per family, calculated the sequence length and checked for the distribution using the in-house developed R scripts (https://github.com/sumanthmutte/DNArepairProteins). For comparison purposes, we also identified protein homologs using the existing HMM profiles for some protein families (AddAB & RecBCD) that were taken from previous studies (Table S1).[70,71]

Filtering criteria and homolog selection

Given the high overlap in the retrieved homologs across multiple protein families, unique and specific filtering criteria were needed to identify the true homologs of each family. In a previous study,[128] length and HMM-score based filters were used, but here these two filters were inefficient to identify true homologs. Instead, reciprocal HMM score based filtering was used to classify homologs into corresponding families. First, the 'FoldDifference' is calculated by dividing the reciprocal HMM score of the top hit with the highest hit score while the 'AbsoluteDifference' is calculate by subtracting these two values. A protein is assigned or classified to the protein family of the top hit, when score is above the median of all the top hits or when the FoldDifference is above 1.5 & AbsoluteDifference is greater than the median calculated earlier. A protein is not assigned to any protein family when the top hit HMM score is less than the median, otherwise considered to have 'multiple' hits and not classified into a specific family. A brief summary of the pipeline along with the number of homologous sequences identified before and after the filtering are provided in Figure 2A and 2B. The filtered (or passed) hits from each family were imported into the R environment for further processing through 'tidyverse' package in R. Upset plots were made using the 'UpSetR' package while the bubble plots showing the phylum specific distribution of each of these protein

families is made using 'ggplot' package. R scripts used for data processing are available in the GitHub repository. The taxonomy information for all the 7249 genomes under consideration is obtained from the NCBI Taxonomy database https://www.ncbi.nlm.nih.gov/taxonomy.

Phylogeny of helicase domains

To build a phylogenetic tree of the helicase domains of various protein families (AddAB, AdnAB and RecBCDQ), we randomly selected protein sequences of 150 'post-filter' hits that are homologs of each family. All the collected protein sequences were aligned using 'genafpair' (E-INS-i) algorithm of MAFFT with a maximum of 1000 iterations (v7.505).[123] Further the positions (aligned columns) with more than 70% gaps were removed using trimal (v1.4).[124] Phylogenetic tree was built using IQtree (v2.2.0)[125] with 'LG + F + R10' as the model of evolution selected using in-built 'ModelFinder'[126] with a maximum of 1000 rapid bootstrap. Phylogenetic tree was visualized in iTOL (v6).[127]

Principal component analysis (PCA) and correlation of DNA repair proteins

Number of homologs per genome and per gene family have been calculated and has been used to calculate the principal components using 'prcomp' function in base R. 'ggplot' and 'ggrepel' packages have been used to plot the PCA and add the labels, respectively. The same input data has been used to make a Pearson correlation plot of individual gene families using 'corrplot' function in 'corrplot' package of R. A p-value of 0.01 was used as the significance cutoff in all the correlation analyses.

Genomic clustering of gene families

For each genomic cluster under consideration (AddAB, AddABJ, AdnAB, RecBCD, RecFOR, RecOR and RecJQ), we calculated the distribution of these protein pairs and classified them based on total cluster size taken as the sum of each gene size plus the intergenic length. Gene size and intergenic length information is retrieved from the genomic features file supplied with each RefSeq genome from NCBI. Since the third quartile of every known cluster was less than 15 kb and operon sizes are below 15 kb,[129,130] a gene set is considered to be in a genomic cluster if the size of the cluster is less than 15 kb. Venn diagrams were made using 'ggvenn' package and the boxplots were made using 'ggplot2' package in R.

Genome size correlations

Genome size and the total protein count were taken from the RefSeq database along with the proteome sequences from the genomic features file. Pearson correlation for the total protein count and the DNA repair protein count against the genome size was calculated using methods described above. The average number of proteins per genome was plotted with smoothing using Loess method with 95% confidence intervals.

Correlation of DNA repair proteins with bacterial immune systems

Data for the presence of immune systems in the bacterial genomes from the RefSeq database has been downloaded from DefenseFinder (accessed on 28-June-2024; https://defensefinder.mdmlab.fr/wiki/refseq.[87,91] Out of 172333 genomes available in DefenseFinder, we selected 2324 genomes that are included in our study for the downstream correlation analysis. Pearson correlation ($p < 0.01$) for the presence (or absence) of DNA repair proteins and the immune systems is calculated irrespective of the number of proteins identified per protein or the system in that genome. Correlation with specific subtypes was performed only for the CRISPR-Cas system. For the pAgo subtype correlations, we utilized the data from a previous study which was performed on the same dataset as the DNA repair proteins (7249 genomes).[128]

QUANTIFICATION AND STATISTICAL ANALYSIS

Different 'n' values used specifically in certain analysis were clearly mentioned in the plots or the figure legends. The overall percentage calculations were made using the number of genomes (n) used in the study i.e., 7249. Correlations were calculate using the 'Pearson' method with a statistical significance (p-value) cut-off of 0.01, unless otherwise specified in the specific method or the figures and the figure legends.

ADDITIONAL RESOURCES

All the scripts and necessary data files with additional information are available through GitHub: https://github.com/sumanthmutte/DNArepairProteins.