The dif/Xer Recombination Systems in Proteobacteria

Christophe Carnoy1,2,3*, Claude-Alain Roten4,5

1 INSERM U801, Lille, France, 2 Univ Lille Nord de France, Lille, France, 3 UDSL, Faculté des Sciences Pharmaceutiques et Biologiques, Lille, France, 4 Department of Fundamental Microbiology (DMF), Biology Building, Biophore, Quartier de la Sorge, Lausanne University, Lausanne-Dorigny, Switzerland, 5 Institute of Microbiology (IMUL), University Hospital Center of Lausanne (CHUV), Lausanne University, Lausanne, Switzerland

Abstract

In E. coli, 10 to 15% of growing bacteria produce dimeric chromosomes during DNA replication. These dimers are resolved by XerC and XerD, two tyrosine recombinases that target the 28-nucleotide motif (dif) associated with the chromosome’s replication terminus. In streptococci and lactococci, an alternative system is composed of a unique, Xer-like recombinase (XerS) genetically linked to a dif-like motif (difSL) located at the replication terminus. Preliminary observations have suggested that the dif/Xer system is commonly found in bacteria with circular chromosomes but that assumption has not been confirmed in an exhaustive analysis. The aim of the present study was to extensively characterize the dif/Xer system in the proteobacteria, since this taxon accounts for the majority of genomes sequenced to date. To that end, we analyzed 234 chromosomes from 156 proteobacterial species and showed that most species (87.8%) harbor XerC and XerD-like recombinases and a dif-related sequence which (i) is located in non-coding sequences, (ii) is close to the replication terminus (as defined by the cumulative GC skew) (iii) has a palindromic structure, (iv) is encoded by a low G+C content and (v) contains a highly conserved XerD binding site. However, not all proteobacteria display this dif/XerCD system. Indeed, a sub-group of pathogenic e-proteobacteria (including Helicobacter sp and Campylobacter sp) harbors a different recombination system, composed of a single recombinase (XerH) which is phylogenetically distinct from the other Xer recombinases and a motif (difSL) sharing homologies with difSL. Furthermore, no homologs to dif or Xer recombinases could be detected in small endosymbiont genomes or in certain bacteria with larger chromosomes like the Legionellales. This raises the question of the presence of other chromosomal deconcatenation systems in these species. Our study highlights the complexity of dif/Xer recombinase systems in proteobacteria and paves the way for systematic detection of these components in prokaryotes.

Citation: Carnoy C, Roten C-A (2009) The dif/Xer Recombination Systems in Proteobacteria. PLoS ONE 4(9): e6531. doi:10.1371/journal.pone.0006531

Editor: Niyaz Ahmed, University of Hyderabad, India

Received May 13, 2009; Accepted June 25, 2009; Published September 3, 2009

Copyright: © 2009 Carnoy, Roten. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: Funding organization: INSERM, Univ Lille Nord de France and University of Lausanne. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

* E-mail: christophe.carnoy@univ-lille2.fr

Introduction

In bacteria, DNA replication of circular chromosomes can generate, by homologous recombination, concatenated chromosomes that affect cell viability. In Escherichia coli, resolution of chromosomal concatenates requires a site-specific recombination system involving two tyrosine recombinases (XerC and XerD) associated with FtsK, the DNA translocase involved in bacterial division [1,2,3]. Chromosomal deconcatenation occurs at a specific chromosome site referred to as dif, for the “deletion-induced filamentation”, a phenotype observed in E. coli strains which are either deficient in XerD or XerC recombinases or lack the dif sequence [4,5]. The 28-nt dif sequence is a palindromic motif composed of two inverted repeats (each of which is specifically targeted by one of the two Xer recombinases) separated by a central hexanucleotide. The E. coli dif sequence is located opposite the origin of chromosomal replication, i.e. near the chromosome terminus at the junction of oppositely polarized DNA sequence elements in a 30 kb-region called the dif activity zone (DAZ) [6,7,8,9].

The Xer recombinase system was originally described for E. coli plasmids [10,11] but is not restricted to this bacterial species, since homologous systems have been functionally characterized in Bacillus subtilis, Haemophilus influenzae, Xanthomonas campestris, Caulobacter crescentus and Vibrio cholerae [12,13,14,15,16]. Xer-related recombinases have also been detected by sequence homology or DNA hybridization in many bacterial taxa and some archaeal species [17,18,19,20]. Homologs to E. coli dif sequences have been found in other proteobacteria, firmicutes and actinobacteria [15,21,22,23,16], suggesting the universality of the dif/Xer system in the bacterial kingdom. Recently, an unconventional single Xer-like recombinase targeting an atypical dif sequence was described in streptococci and lactococci [23].

In addition to its role in chromosome dimer resolution, the dif locus may be involved in the integration/excision of exogenous DNA. For instance, the filamentous phages CTXΦ and VGΦ in Vibrio cholerae, f237 in V. parahaemoliticus, CUS-1 in E. coli 0188:K1:H7, YpfΦ in Yersinia pestis and Ci16-v1 and ΦL in Xanthomonas campestris all integrate into the host chromosome at the dif site [24,25,26,27,28,29,30]. The mechanism of prophage genome integration has been described in detail in V. cholerae CTXΦ, the filamentous phage containing the cholera toxin-encoding gene [31,32]. Recently, Val et al. showed that after appropriate folding, CTXΦ’s single-stranded phage DNA forms a dif-like structure that irreversibly recombines with the bacterial dif by using host XerG and XerD recombinases [32]. This clearly demonstrates that dif is a preferential integration site for single-stranded filamentous phages displaying dif-like motifs on their...
genome. Other large genetic DNA elements also target the *dif* sequence, as evidenced by integration of the 57-kb gonococcal genetic island (GGI, containing a type IV secretion system) into the *Neisseria* chromosome [33,34]. Taken as a whole, these studies strongly suggest that the *dif* sequence is a preferential site for exogenous DNA integration and thus contributes to genome evolution in general and to virulence gene acquisition in particular. Moreover, *dif*′s natural ability to integrate exogenous DNA has been used to deliver genes of biotechnological interest to the bacterial chromosome [35,36].

Despite the publication of many bacterial genome sequences (440 had been sequenced as of January 1st, 2007) with over half of these from proteobacteria, no exhaustive analysis of the *dif*Xer system has yet been undertaken. As the *dif* sites do not appear in GenBank’s genome annotation, we developed a strategy for systematically identifying *dif*-related sequences in proteobacterial chromosomes by combining similarity search tools (BLAST and YASS) with genometric methods (cumulative GC-skew analyses). In contrast to initial assumptions in the literature, we found that not all proteobacterial circular chromosomes feature a *dif*/Xer system and that a *ɛ*-proteobacteria sub-group harbors an atypical *dif*/Xer system, indicating heterogeneity of this recombination system in proteobacteria. This report represents the first comprehensive analysis of the *dif* motifs and of their associated recombinases and should facilitate the identification of related recombination systems in prokaryotes.

**Results**

The *dif*-related sequences are highly conserved among the proteobacteria

To detect *dif* homologs in proteobacterial chromosomes, we developed an in silico approach based on (i) homology of the candidate with the experimentally characterized proteobacterial *dif* sequences in *E. coli, C. crescentus, X. campestris, V. cholerae* and *H. influenzae* or with a related sequence found in a close taxon, (ii) location of the putative sequence near the chromosome terminus, as defined by the cumulative GC skew analysis, (iii) presence in different strains of the same species, and (iv) presence of a single copy of the *dif* candidate within the chromosome. Using this strategy, 234 chromosomes from 156 proteobacterial species were analyzed (Table 1 and Table S1). *dif* homologs were found in 87.2% of the chromosomes (204 out of 234) and in 87.8% (137 out of 156) of the species. A *dif*-related sequence was detected in all *β* and *δ*-species and in 97.7% (44 out of 45) and 82.8% (53 out of 64) of the *α* and *γ*-species, respectively. Surprisingly, only 1 out of 8 *ɛ*-proteobacterial species (12.5%) harbored a *dif*-related sequence. Lastly, one unclassified proteobacterium (*Magnetococcus* sp.) also displayed a *dif* homolog.

To avoid redundancy, the first-published chromosome sequence in a species was considered to be representative. Thus, of the 204 *dif* sequences that we characterized, 161 were considered to be representative of the different proteobacterial taxa and were therefore used to define a consensus sequence (Figure 1 and Table S2). The two undecanucleotides (11-mers) corresponding to the XerC and XerD binding sites were designated in this study as *dif*\(^{XerC}\) and *dif*\(^{XerD}\), respectively, whereas the central hexanucleotide between the two Xer binding sites was named as *dif*\(^{cons}\) (Figure 1A). Analysis of the consensus revealed that the *dif*\(^{XerD}\) site is better conserved than the *dif*\(^{XerC}\) site and that within both *dif*\(^{XerC/D}\) boxes, the most conserved region is located in the inner part, near the central region. Regarding *dif*\(^{XerD}\), the adenine residue at position 25 of the 25-nt *dif* sequence is highly conserved, whereas the nucleotides at positions 23 and 24 are more variable (Figure 1A). Within the less conserved nucleotides in *dif*\(^{cons}\), the residue at position 13 (i.e. the second in the hexanucleotide) is the most variable, compared with the other five. Furthermore, the degree of variability upstream and downstream of the 25-mer consensus sequence is high, indicating that the *dif*-related sequences are located in different genetic environments (Figure 1A).

Since *dif*\(^{XerC}\) is more variable than *dif*\(^{XerD}\), we then wondered whether XerC recombinases would be less well conserved than the XerD proteins. To answer this question, a phylogenetic analysis based on the amino acid sequences was performed on both recombinases in the 48 species which were held to be representative of the proteobacterial taxa (Table 1 and Figure 2). Firstly, our analysis revealed a clustering of the Xer recombinases that confirms the taxonomic organization proposed by Olsen et al [37] (i.e. clustering of the *γ* and *β* groups on one hand and the *δ*, *ɛ* and *α* groups on the other). Second, comparison of the XerC and XerD phylogenetic trees revealed greater branch lengths in XerC’s phylogeny than in XerD’s (Figure 2). This clearly indicates greater divergence between the XerC recombinases than between the XerD proteins. The higher variability of the proteobacterial XerC recombinases might thus explain the higher degree of sequence variability for the *dif*\(^{XerC}\) site. This observation strongly suggests co-evolution of the Xer recombinases and their related-*dif* sequences. The greater degree of conservation of XerC relative to XerD might be constrained by the direct interaction of XerD (but not XerC) with the highly conserved translocase FtsK [38]. Thus, evolutionary changes in XerD and consequently in *dif*\(^{XerD}\), might have been limited by the conservation of FtsK.

**Intra-species variations in *dif*-related sequences**

Although the *dif*-related sequences are highly conserved within a given species, differences in the *dif* sequences were observed between strains. To evaluate any intra-species variations, we compared the *dif*-related sequences in the 21 multi-strain *dif*\(^{+}\) species (Table S1) and calculated the degree of variability at each nucleotide position in the *dif* locus (Figure 1B). This analysis again revealed that the *dif*\(^{XerD}\) site is best conserved and that intra-species differences are located at the *dif*\(^{XerC}\) and *dif*\(^{XerD}\) outer ends (nucleotides 1 to 4 and 27–28, respectively). Surprisingly, with regard to the high nucleotide variability of the *dif*\(^{cons}\) in the consensus sequence (Figure 1A), this region displays low intra-species variability. This observation clearly indicates that *dif*\(^{cons}\) is well conserved within strains of the same species but weakly conserved between species.

**Variations in *dif*-related sequences in multi-chromosome bacteria**

In *α*, *β* and *γ*-proteobacteria, some species contain two or three chromosomes, with each (except in *Agrobacterium tumefaciens*) displaying one *dif*-related sequence (Table S3). Comparison of the sequences in given species indicated that (i) each chromosome harbors a distinct *dif* sequence, (ii) the main differences were found in the *dif*\(^{cons}\) region and (iii) the *dif*\(^{XerD}\) region was less variable than the *dif*\(^{XerC}\) region (Figure 1B). Interestingly, the genes encoding the XerC and XerD tyrosine recombinases were always found as single copies, within the largest chromosome (Table S3) indicating that one couple of Xer recombinases interacts with two (or even three) distinct *dif* sequences in multi-chromosome bacteria and confirming the recent report by Val et al.[16]. This observation suggests that the recombinase / *dif* interaction allows some degree of variability - especially for XerC / *dif*\(^{XerC}\). It is noteworthy that nucleotide positions 5 and 8 to 11
Table 1. Genome and dif features of a representative panel of proteobacteria.

| Species                          | Genome features | putative dif features |
|----------------------------------|-----------------|-----------------------|
|                                  | size (bp)       | G+C content | Maximum CGC skew | Nucleotide sequence (1) | position on genome | G+C content | distance from GCG skew | intergenic location (2) |
| **α-proteobacteria**             |                 |             |                  |                         |                  |             |                     |                          |
| Caulobacteraceales               |                 |             |                  |                         |                  |             |                     |                          |
| Caulobacter crescentus CB15      | 4016947         | 0.672       | 1930040          | AAGATCGAATTTTTGAAATTTATGAAAGT | 1946380         | 0.250       | 14159               | yes                       |
| Rhizobiales                      |                 |             |                  |                         |                  |             |                     |                          |
| Agrobacterium tumefaciens str. C58 chr. circular | 2841490       | 0.593       | 1485983          | TAAATCGATAAGATATTTGAAATTTT | 1478815         | 0.250       | 7168                | yes                       |
| Bartonella quintana str. Toulouse | 1581384       | 0.387       | 724981           | AAATCGATAATTTATGAGAATAA | 720906         | 0.179       | 4075                | yes                       |
| Bradyrhizobium japonicum USDA 110 | 9105828      | 0.640       | 4893406          | GATTGCCATAGTTATTGGAATAT | 4996172         | 0.286       | 102766              | yes                       |
| Brucella melitensis 16M chr. I   | 2117136        | 0.571       | 955452           | TAAATCGATAATTTATGAGAATAA | 954740         | 0.321       | 712                 | yes                       |
| Brucella melitensis 16M chr. II  | 1177785        | 0.573       | 757557           | AAATCGATAATTTATGGAACAT | 758183         | 0.214       | 626                 | yes                       |
| Mesorhizobium loti MAF303099     | 7036071        | 0.627       | 203543           | AAATCGATAATTTATGGAACAT | 299619         | 0.321       | 96076               | yes                       |
| **β-proteobacteria**             |                 |             |                  |                         |                  |             |                     |                          |
| Rhodobacteraceales               |                 |             |                  |                         |                  |             |                     |                          |
| Rhodobacter sphaeroides 2.4.1 chr. 1 | 3188599   | 0.690       | 1435088          | GATTGCCATAATCTTATATTCTT | 1436843       | 0.321       | 1755                | yes                       |
| Rhodobacter sphaeroides 2.4.1 chr. 2 | 943016       | 0.690       | 399979           | TATCTCATAAGCAAGATTATCAA | 371575         | 0.250       | 28044               | yes                       |
| Rhodospirillales                 |                 |             |                  |                         |                  |             |                     |                          |
| Magnetospirillum magneticum AMB-1 | 4967148      | 0.650       | 2616185          | CGTGGCCATAATATCATGACAA | 2610339       | 0.393       | 5846                | yes                       |
| Rickettsiales                    |                 |             |                  |                         |                  |             |                     |                          |
| Ehrlichia ruminantium str. Welgevonden | 1516355 | 0.274       | 746761           | ATATTACAATGCTATTGAAATAT | 747982         | 0.143       | 18779               | yes                       |
| Rickettsia prowazekii str. Madrid E | 1111523     | 0.290       | 628915           | TGTTCTATAATCTATTGGAATAT | 596105         | 0.179       | 32810               | yes                       |
| Sphingomonadales                 |                 |             |                  |                         |                  |             |                     |                          |
| Novosphingobium aromaticivorans DSM 12444 | 3561584 | 0.652       | 2030402          | AGATTGATATATACATTTGAAATA | 2048172       | 0.179       | 17770               | yes                       |
| **γ-proteobacteria**             |                 |             |                  |                         |                  |             |                     |                          |
| Burkholderiales                  |                 |             |                  |                         |                  |             |                     |                          |
| Bordetella pertussis Tohama I    | 4086189        | 0.677       | 2227724          | AATGCCTGAATTTTATATAATGAAAT | 2229069       | 0.214       | 1345                | yes                       |
| Burkholderia mallei ATCC 23344 chr. 1 | 3510148   | 0.681       | 1086094          | AATGCCTGAATTTTATATAATGAAAT | 1081309       | 0.214       | 4785                | hyp. prot                |
| Burkholderia mallei ATCC 23344 chr. 2 | 2325379    | 0.689       | 1077185          | AATGCCTGAATTTTATATAATGAAAT | 1075135       | 0.286       | 2050                | yes                       |
| Ralstonia solanacearum GM1000    | 3716413        | 0.670       | 2009173          | GATTGCCATAATTTTAATTGAAAT | 2031219       | 0.250       | 22046               | yes                       |
| Rhodotherix ferrireducens DSM 15236 | 4712337   | 0.598       | 2483317          | AGATTGATATATACATTTGAAATA | 2472550       | 0.250       | 12767               | yes                       |
| Hydrogenophilales                |                 |             |                  |                         |                  |             |                     |                          |
| Thiobacillus denitrificans       | 2909809        | 0.660       | 1440104          | ACTGCGATAATTTTATATGAAAT | 1430783       | 0.214       | 9321                | yes                       |
| Methylocophilales                |                 |             |                  |                         |                  |             |                     |                          |
| Methyllobacillus flagellatus KT  | 2971517        | 0.557       | 1573478          | ACTGCGATAATTTTATATGAAAT | 1564653       | 0.214       | 8825                | yes                       |
| Neisseriales                     |                 |             |                  |                         |                  |             |                     |                          |
| Neisseria meningitidis MCS8      | 2272351        | 0.515       | 1231577          | AGATTGATATATACATTTGAAAT | 1229349       | 0.214       | 2228                | hyp. prot                |
| Nitrosomonadales                 |                 |             |                  |                         |                  |             |                     |                          |
| Species                        | Genome features | putative dIF features |
|-------------------------------|-----------------|-----------------------|
|                               | Species size (bp) | Maximum GC skew | Nucleotide sequence (1) | position on genome | G+C content | distance from GCG skew | Intergenic location (2) |
| Nitrosomonas europaea ATCC 19718 | 2812094         | 0.507               | 964528                | ATTTCGTATAATTGTAATTGTGAAAT | 974219         | 0.143                       | 9691               | yes                     |
| Rhodocyclales                 |                 |                      |                       |                      |              |                            |                    |                         |
| Dechloromonas aromatica RCB   | 4501104         | 0.592               | 2186143               | ACGCGCATAATTTGCAATTGTGAAAT | 2192508         | 0.286                       | 6365               | yes                     |
| *δ*-proteobacteria            |                 |                      |                       |                      |              |                            |                    |                         |
| Bdellovibrionales             |                 |                      |                       |                      |              |                            |                    |                         |
| Bdellovibrio bacteriovorus HD100 | 3782950        | 0.506               | 1940732               | TCTTGTAATGTTATATTGTAACG | 1946858         | 0.286                       | 6126               | yes                     |
| Desulfobacterales             |                 |                      |                       |                      |              |                            |                    |                         |
| Desulfotalea psychrophila LSv54 | 3523383        | 0.468               | 2260306               | TAAGGAGATAATGATTGAAACG | 2338241         | 0.250                       | 77935              | yes                     |
| Desulfuvibionales             |                 |                      |                       |                      |              |                            |                    |                         |
| Desulfuvibrio vulgaris subsp. Vulgaris str. Hildenborough | 3570858 | 0.631 | 1735879 | ATGTCCCATATTGAATTGTGAACT | 1754277 | 0.250 | 18398 | yes |
| Myxococcales                  |                 |                      |                       |                      |              |                            |                    |                         |
| Anaeromyxobacter dehalogenans ZCP-C | 5013479 | 0.749 | 1906268 | ATGTCCGAAATGGAATTGTGAACT | 1906477 | 0.357 | 209 | yes |
| Myxococcus xanthus DK 1622    | 9139763         | 0.688               | 4547166               | ATGTCCGAAATGGAATTGTGAACT | 4489697         | 0.393                       | 57469              | yes                     |
| Syntrophobacterales           |                 |                      |                       |                      |              |                            |                    |                         |
| Syntrophus aciditrophicus SB  | 3179300         | 0.514               | 1665473               | TGGTCCATAAGTATATTGTAACC | 1665477         | 0.250                       | 388                | yes                     |
| *γ*-proteobacteria            |                 |                      |                       |                      |              |                            |                    |                         |
| Campylobacteriales            |                 |                      |                       |                      |              |                            |                    |                         |
| Sulfurimonas denitrificans DSM 1251 | 2201561 | 0.34 | 1135161 | TTCAATAGAATTACATATTGTAACC | 1122264 | 0.175 | 12897 | yes |
| *γ*-proteobacteria            |                 |                      |                       |                      |              |                            |                    |                         |
| Aeromonadales                 |                 |                      |                       |                      |              |                            |                    |                         |
| Aeromonas hydrophila ATCC7966  | 4744448         | 0.615               | 2494705               | ACCGGCCATAATGTTATATTGTAAT | 2514936         | 0.286                       | 20231              | yes                     |
| Alteromonadales               |                 |                      |                       |                      |              |                            |                    |                         |
| Idiomarina loihiensis L2TR    | 2839318         | 0.470               | 1411650               | ATGCGATAATGTTATATTGTAAT | 1387623         | 0.179                       | 24027              | yes                     |
| Shewanella oneidensis MR-1    | 4969795         | 0.459               | 2490130               | ACATCCCCGAAATGTTATATTGTAAT | 2476928         | 0.286                       | 13202              | yes                     |
| Chromatiales                  |                 |                      |                       |                      |              |                            |                    |                         |
| Nitrosococcus oceanis ATCC 19707 | 3481691     | 0.503               | 1849931               | TGGTCCGATAATTACATATTGTAAT | 1850410         | 0.214                       | 479                | yes                     |
| Enterobacterales              |                 |                      |                       |                      |              |                            |                    |                         |
| Erwinia carotovora subsp. atroseptica SCRI1043 | 5064019 | 0.509 | 2552458 | GGTGGCCATAATGTTATATTGTAAT | 2532133 | 0.250 | 20325 | yes |
| Escherichia coli K12           | 4639675         | 0.507               | 1549688               | GGTGGCCATAATGTTATATTGTAAT | 1588788         | 0.286                       | 39100              | yes                     |
| Sodalis glossinidius str. ‘morsitans’ | 4717116 | 0.546 | 2461688 | AGTGCCGATAATGTTATATTGTAAT | 2471418 | 0.250 | 9460 | yes |
| Yersinia pestis CO92           | 4653728         | 0.476               | 2562641               | GGTGGCCATAATGTTATATTGTAAT | 2562919         | 0.286                       | 278                | yes                     |
| Methyllococcales              |                 |                      |                       |                      |              |                            |                    |                         |
### Table 1. Cont.

| Species                                | Genome features | putative dif features |
|-----------------------------------------|-----------------|-----------------------|
|                                         | size (bp)       | G+C content | Maximum CGC skew | Nucleotide sequence (1) | position on genome | G+C content | distance from GCG skew | intergenic location (2) |
| *Methyllococcus capsulatus* str. Bath   | 3304553         | 0.635       | 1531625          | TATGGCATAATGTAATATGTTAAT | 1492525           | 0.214       | 39100                 | yes                    |
| Oceanospirillales                       |                 |             |                 |                           |                   |             |                      |                        |
| *Hahella chejuensis* KCTC 2396          | 7215267         | 0.538       | 3493027          | AGTGGCATAATGTAATATGTTAAT | 3437061           | 0.214       | 1966                  | yes                    |
| Pasteurellales                          |                 |             |                 |                           |                   |             |                      |                        |
| *Haemophilus influenzae* Rd KW20        | 1830023         | 0.381       | 1474989          | ATTTGCATAATATATATGTTAAT | 1473975           | 0.143       | 1014                  | yes                    |
| *Pseudomonadales*                       |                 |             |                 |                           |                   |             |                      |                        |
| *Acinetobacter* sp. ADP1                | 3598621         | 0.404       | 1847121          | GATTCGATATGTAATATGTTAAT | 1848733           | 0.179       | 1612                  | yes                    |
| *Pseudomonas aeruginosa* PAO1           | 6264403         | 0.665       | 2428120          | GATTCGATAATATATATGTTAAT | 2443082           | 0.214       | 14962                 | yes                    |
| Thiiorchiales                           |                 |             |                 |                           |                   |             |                      |                        |
| *Francisella tularensis* subsp. tularensis Schu 4 | 1892819         | 0.322       | 950050           | CATTCGATAATATATATGTTAAT | 994689            | 0.143       | 44639                 | yes                    |
| *Vibionales*                            |                 |             |                 |                           |                   |             |                      |                        |
| *Vibrio cholerae* O1 biovar eltor str. N16961 chr. I | 2961116         | 0.476       | 1564264          | GATTCGATATGTAATATGTTAAT | 1564118           | 0.250       | 146                   | yes                    |
| *Vibrio cholerae* O1 biovar eltor str. N16961 chr. II | 1072311         | 0.469       | 512448           | GATTCGATATGTAATATGTTAAT | 507996            | 0.357       | 4452                  | yes                    |
| *Xanthomonadales*                       |                 |             |                 |                           |                   |             |                      |                        |
| *Xanthomonas campestris* pv. campestris str. ATCC 33913 | 5076172         | 0.650       | 2442019          | TCCTGACATAATATATATGTTAAT | 2441762           | 0.286       | 257                   | yes                    |

(1) The central nucleotide in bold defines the position of the *dif* sequence on the chromosome. The nucleotides involved in the palindrome are underlined.

(2) hyp.prot. = *dif* inserted into a hypothetical protein-encoding gene.

(3) *Sulfurimonas* denitrificans strain DSM 1251 = *Thiomicrospira* denitrificans ATCC 33889.

(4) Maximum of the GC skew.

doi:10.1371/journal.pone.0006531.t001
in \( \text{dif}^{\text{XerC}} \) and 18 to 23 in \( \text{dif}^{\text{XerD}} \) do not vary between chromosomes (Figure 1B) and therefore these positions may well be critical for recombinase binding.

When scanning the genome of the multi-chromosome \( A. \) \textit{tumefaciens} for \( \text{dif} \)-related sequences, we found a \( \text{dif} \) sequence on the larger circular chromosome (2.84 Mb) but none on the smaller linear chromosome (2.07 Mb). This finding is not surprising, as it has been shown that the \( E. \) \textit{coli} \( \text{dif} \) sequence is dispensable after linearization of the circular chromosome [39]. The origin of the linear chromosome in \( A. \) \textit{tumefaciens} is unknown but some sequence features suggest that it derives from a plasmid [40,41]. If the plasmid origin of the linear chromosome is confirmed, one can hypothesize that the \( \text{dif} \) sequence would have been lost after the chromosome became linear. Furthermore, when analyzing the presence of Xer homologs in \( A. \) \textit{tumefaciens}, we observed that the gene coding for the XerD-like recombinase is present on the linear chromosome, whereas the \( \text{xerC} \) homolog gene was located on the circular chromosome. This distribution of the Xer recombinase genes seems to be specific to \( A. \) \textit{tumefaciens}, since both recombinases are located on the larger chromosome in the recently sequenced genomes of \( A. \) \textit{citri} and \( A. \) \textit{radiobacter}. Hence, \( A. \) \textit{tumefaciens} is the only known multi-chromosome bacterium in which the XerC and XerD-encoding genes are on different chromosomes. This example suggests a \( \text{xer} \) gene

---

**Figure 1.** Nucleotide variability within \( \text{dif} \)-related sequences. (A) Consensus sequence and \( \text{dif} \) nucleotide variability for 161 \( \text{dif} \)-related sequences from 137 proteobacterial species. Nucleotide sequence characters in bold represent the \( \text{dif} \) sequence (28-mer). If the nucleotide frequency represents more than 50%, it is written in upper case letters; if not, the nucleotide is written in lower case letters. The nucleotide variability at each position in the 28-mer was defined as \( 1-f \), where \( f \) is the frequency of the most frequent nucleotide. Nucleotide frequencies at each position are given in Table S2. Black bars represent \( \text{dif}^{\text{XerC}} \) and \( \text{dif}^{\text{xerC}} \) nucleotides, whereas grey bars correspond to the the \( \text{dif}^{\text{cent}} \) nucleotides. White bars represent nucleotides outside \( \text{dif} \). (B) Degree of variability in the \( \text{dif} \) sequence in 21 multi-strain species and in 19 multi-chromosome species. The degree of variability was calculated for each nucleotide position, as described in the Methods section.
doi:10.1371/journal.pone.0006531.g001
translocation from the larger, conserved chromosome to the smaller, less conserved one [42].

Low G+C content, palindromicity, close association with the terminus and presence in a non-coding region are conserved features of the \textit{dif}-related sequences.

Nucleotide analysis of the \textit{dif}-related sequences revealed that for a given species, the G+C content of the \textit{dif} motif was systematically lower than the G+C content of the corresponding chromosomes (a difference of between 8.3% and 58.4%, median value = 29.9%) (Table S1). Furthermore, as palindromicity seems to be essential for \textit{dif} functionality [11,43], we next searched for palindromes in the \textit{dif} sequences from the 48 selected species. Of the 28 nucleotides of the \textit{dif} sequence, 16.3±3.0 are involved in a palindrome. When the analysis was performed with 28-mers randomly generated from the initial \textit{dif} sequences, the number of nucleotides involved in palindrome was significantly lower (9.3±4.4) (p<0.001; Student’s test, n = 48) confirming that palindromicity is a key feature of the \textit{dif} motif. We then analyzed each nucleotide position in the \textit{dif} sequences for their involvement in a palindromic structure. Positions 5, 6, 23, 24 of the 28-mer are rarely involved in a palindromic structure, whereas nucleotides at position 8 to 11 and 18 to 21 (corresponding to the inner part of \textit{dif} XerC and \textit{dif} XerD) are frequently associated (Figure S1).

We then compared the \textit{dif} position on the chromosome relative to the maximum cumulative GC skew. On the 161 chromosomal...

**Figure 2. Phylogeny of proteobacterial XerC and XerD recombinases.** Representative proteobacterial species of each taxon were selected for the analysis (Table 1). β-proteobacterial species are represented in blue, with γ in red, δ in green, α in magenta and ε in black. Amino acid sequence alignments were performed using Clustal W (MEGA 4 [60]). The evolutionary history was inferred by using the Neighbor-Joining method [61] conducted in MEGA4. Similar results were obtained using the Minimum Evolution method (data not shown). Only significant bootstrap values (≥90%) obtained with 1000 runs are indicated next to the branches (white with a grey background). The tree is drawn to scale, with branch lengths (below the branches) in the same units as those of the evolutionary distances used to infer the phylogenetic tree. Branch lengths below the value 0.05 are not shown. The evolutionary distances were computed using the Poisson correction method and are given as the number of amino acid substitutions per site. doi:10.1371/journal.pone.0006531.g002
sequences, the median distance between the \textit{dif} motif and the replication terminus as defined by the cumulative GC skew method was calculated to be 7277 bp (first quartile = 2055.5; third quartile = 21,445.5), with a distance ranging from 146 bp for chromosome 1 in \textit{V. cholerae} to 199612 bp for \textit{Syntrophus aciditrophicus}. The great distance between the location of \textit{dif} and the peak of the cumulative GC skew curve for a few species can be mainly explained by a noisy GC skew signal blurring the precise location of its maximum value. However, despite this difficulty, we noted a high degree of correlation ($R^2 = 0.9978$) between \textit{dif}'s position and the peak of the cumulative GC skew curve (Figure 3), which confirms the close association previously observed in a smaller number of species [22].

Furthermore, we analyzed the gene environment of the 161 \textit{dif} motifs and found that most (96.3\%) were located in non-coding regions. This observation clearly indicates that \textit{dif} intergenicity is another key feature. However, in a few cases (6 sequences out of 161, 3.7\%), \textit{dif} was present within coding sequences, four of which correspond to hypothetical proteins, plus two associated to characterized open reading frames (ORFs). Whereas the sequence was inserted within a bacteriophage protein-coding sequence in \textit{Vibrio parahaemolyticus} (chromosome 1), the motif was located in a gene coding for a major facilitator family transporter in the third chromosome of \textit{Burkholderia ambifaria} (Table S1).

Analysis of the flanking coding sequences of the 161 representative \textit{dif}-related sequences revealed that 10.9\% were flanked by proteins of phage origin and 14.2\% were associated with insertion sequences or transposase- or integrase-encoding genes. This shows that about a quarter of the \textit{dif} sequences are associated with ORFs whose products are involved in mobility. This number might even be an underestimate, since 60\% of the \textit{dif} sequences have ORFs with unknown functions in their vicinity (upstream, downstream or both). These results emphasize the propensity of the terminus region in general and \textit{dif} in particular to facilitate DNA mobility.

### Two \textit{dif}/\textit{Xer} systems in \textit{\varepsilon}-proteobacteria

During the initial analysis of the \textit{\varepsilon}-proteobacterial chromosomes, we found that only one species (\textit{Sulfurimonas denitrificans}) out of 8 had a \textit{dif} sequence (Table 1 and Table S1). In order to understand the apparent absence of a \textit{dif}-related sequence in the genome of \textit{Helicobacter} and \textit{Campylobacter} species, we searched for the presence of the \textit{Xer}-like recombinases in this subgroup. A \textit{XerD} homolog was found in all bacteria belonging to this \textit{\varepsilon}-subgroup, although the corresponding protein had a low degree of homology with \textit{E. coli} \textit{XerD} and was longer (between 355 and 363 amino acids versus 298 amino acids for \textit{E. coli} \textit{XerD}). Surprisingly, we did not detect any other recombinases that unambiguously corresponded to a \textit{XerC} homolog. Blastp analysis with \textit{E. coli} \textit{XerC} showed the presence of \textit{XerC}-like recombinases but none was ubiquitously found in the \textit{Helicobacter} and \textit{Campylobacter} species. Some of these \textit{XerC}-like recombinases probably correspond to the transposable element-associated recombinases found in \textit{Helicobacter} and designated “\textit{XerT}” by Kersulyte et al. [44]. We thus concluded that this \textit{\varepsilon}-proteobacteria sub-group expresses only one ubiquitous \textit{Xer} recombinase that we designated here as “\textit{XerH}” because \textit{Helicobacter} is a major representative of this group. The presence of a single \textit{Xer} recombinase is not unique in the bacterial kingdom. Indeed, it was recently shown that \textit{Streptococcus} and \textit{Lactococcus} species display an unconventional \textit{dif} sequence (\textit{difSL}) which requires a single 356-amino acid recombinase, \textit{XerS} [23]. Although \textit{XerS} and \textit{XerH} exhibit a similar size, the proteins appear to be phylogenetically unrelated (Figure 4). However, when BLASTing

\begin{figure}
\centering
\includegraphics[width=\textwidth]{Fig3.png}
\caption{Correlation between the position of the \textit{dif} sequence and the terminus of replication as defined by cumulative GC skew. The analysis was performed on the 161 proteobacterial chromosomes from the 137 representative \textit{dif}+ species (Table S1). Chromosome of \textit{Wolbachia} endosymbiont of \textit{Drosophila melanogaster} and chromosome 2 of \textit{Pseudoalteromonas haloplanktis} were not included in the analysis since no terminus of replication could be located for these species by the method of the cumulative GC skew. The equation of the plot and the coefficient of determination ($R^2$) are given. doi:10.1371/journal.pone.0006531.g003}
\end{figure}
difSL on ε-proteobacteria genomes, we discovered a difSL homolog presenting all key features of a dif motif, i.e. (i) located near to the peak of the cumulative GC skew, (ii) present in non-coding regions and (iii) with a low G+C content (Table 2). Furthermore, this difSL-like sequence designated as difSL was composed of 2 highly conserved, inverted repeats separated by a central hexanucleotide variable region - another hallmark of the classical features of the difSL homolog (Table S4). It seems that genome fitting can be associated with specific phylogenetic groups. Experimental approaches are now required to test the functionality of the newly discovered ε Xer-like recombination system.

The dif/Xer system is not present in all proteobacteria

Our approach revealed that 12.2% of the studied proteobacterial species do not contain a dif motif. Most of them lack the XerC, XerD, XerH or XerS recombinases, justifying the absence of dif (Table S4). It seems that genome size should be taken into account when considering the absence of the dif/Xer system. Indeed, insect endosymbiont bacteria (Buchnera sp, Blochmannia sp, Carsonella ruddii, Rhabdiza magnifica, Baumannia cicadellinicola and Wigglesworthia glossinidiae) have a genome size ranging from 0.159 to 1.1 Mb and lack the dif/Xer system (Table S4). During their co-evolution with their host, the endosymbiotic bacteria have lost a large part of their genome and have retained only genes that are essential for survival [46]. The absence of the dif/Xer system in these bacteria indicates that this recombination system is not required for microbial symbiosis. Likewise, the marine α-proteobacteria Pelagibacter ubique has the smallest known genome of a free-living microorganism (1.3 Mb) [47] and, like the endosymbiotic bacteria, does not possess a dif/Xer system. This confirms that genome fitting can affect non-vital systems, such as the Xer machinery. However, low chromosome size is not always associated with the absence of the dif/Xer system, since the Rickettsiales (α-proteobacteria with a genome ranging from 0.85 to 1.52 Mb in size) do harbor dif/Xer recombination machinery (Table 1 and Table S1). Furthermore,
Table 2. Features of the putative dif sequences of ε-proteobacteria.

| ε-proteobacteria species (1) | chromosomal features (2) | putative dif characteristics |
|------------------------------|--------------------------|------------------------------|
|                              | size (bp) | G+C content | CGC skew | sequence (40-mer) (3) | G+C content | position on chromosome | distance from GC skew (bp) | distance from xerE (bp) | intergenic location (4) |
| Arcobacter butzleri RM4018   | 2341251  | 0.270       | 1232417  | TAATAGTAGAAGAAACTATAATTCTCAAATAAAAATA 0.100   | 1197716 | 34701 | 66                   | yes                |
| Campylobacter concisus 13826 | 2052006  | 0.394       | 971215   | ATTTTTGTTAAGAAACTTAATTTCTCAAATTTATTTT 0.125 | 999842 | 28627 | 59459                | hyp. prot            |
| Campylobacter curvus 525.92  | 1971264  | 0.445       | 1008113  | ATTTTTGTTAAGAAACTTAATTTCTCAAATTTATTTT 0.100 | 991768 | 16345 | 68030                | hyp. prot            |
| Campylobacter fetus subsp. fetus 82-40 | 1773615 | 0.333       | 908630   | TAATTTTGTTAAGAAACTTAATTTCTCAAATTTATTTT 0.150 | 886842 | 21788 | 35968                | yes                |
| Campylobacter hominis ATCC BAA-381 | 1711272 | 0.317       | 875024 | TAATTATTGGATTATTGAAAATTATTTTTCAATTTTGT 0.125 | 851021 | 24003 | 214673               | yes                |
| Campylobacter jejuni subsp. doyle 269.97 | 1845106 | 0.306       | 853771  | TAATTTTGTTAAGAAACTTAATTTCTCAAATTTTATTTT 0.125 | 892865 | 39094 | 217                  | yes                |
| Campylobacter jejuni RM1221 | 1777831  | 0.303       | 893799   | TAATTTTGTTAAGAAACTTAATTTCTCAAATTTTATTTT 0.150 | 888360 | 5439  | 215                  | yes                |
| Helicobacter acinonychis str. Sheeba | 1553927 | 0.382       | 748099   | TAGTTTAGTTAAGAAACTGACTTTCTCAATTATTAAAT 0.225 | 747275 | 824  | 282                  | yes                |
| Helicobacter hepaticus ATCC 51449 | 1799446 | 0.359       | 1794500  | TGATTATGTTAAGAAACTGACTTTCTCAATTATTAAAT 0.175 | 1765790 | 28710 | 125                  | yes                |
| Helicobacter pylori 269.95 | 1667867  | 0.388       | 813426   | TCAATTTGTTAAGAAACTGACTTTCTCAATTATTAAAT 0.225 | 723517 | 89909 | 1981                 | yes                |
| Nitratispira sp. SB155-2 | 1879311  | 0.397       | 927614   | TTTATTAGTTAAGAAACTTAATTTCTCAAATTTTATT 0.125 | 1001399 | 73785 | 52                   | yes                |
| Wolinella succinogens DSM 1740 | 2110355 | 0.484       | 1188027  | TCAATTTGTTAAGAAACTAATTTCTCAAATTTTATT 0.200 | 1170384 | 17643 | 16                   | yes                |
| consensus sequence (6)      | ---TT--T---AATAACCTTTGAAAAC---TTTTCAAA--------- | classical dif | distance from XerC / XerD (bp) | 705981/193695 | yes |

| Sulfurimonas denitrificans DSM 1251 (7) | 2201561 | 0.345 | 1135161 | AAAATTTCAATAGAATTTCATATGTATTAACTAACTATATA 0.175 | 1122264 | 12897 | 705981/139695 | yes |
| Sulfurovum sp. NBC37-1 | 2562277  | 0.439 | 1189307 | TGCATTGATTAGAATTTCATATGTATTAACTAACTATATA 0.150 | 1186929 | 2378 | 1109695/758085 | yes |
| consensus sequence (6) | ---TTT-AATAACCTTTGAAAAC---TTTTCAAA--------- | classical dif | distance from XerC / XerD (bp) | 705981/193695 | yes |

(1) One representative per species.
(2) All genomes are circular.
(3) The position of the putative dif motif on the chromosome corresponds to the nucleotide in bold type, located between the two inverted repeats.
(4) Hyp. prot. = hypothetical protein.
(5) Maximum of the GC skew.
(6) Underlined nucleotides correspond to the inverted repeats.
(7) Sulfurimonas denitrificans strain DSM 1251 = Thioregulosa denitrificans ATCC 33889.

doi:10.1371/journal.pone.0006531.t002
the absence of the dif/Xer system cannot only be explained by chromosome fitting, since bacteria with a larger chromosome (like the Legionellales, *Caldiffus pyroides* or *Sachschephas* *degrediens*: genome size ranging from 2 Mb to 3 Mb) also lack this machinery. Surprisingly, a dif/Xer system was not found in *Aronatidium aromatica* str EhN1 (also designated as *Azoarcus* sp EhN1) whereas the complete system was revealed in *Azoarcus* BH72. This difference could be attributed to the low degree of synteny seen for the genomes of these two phylogenetically similar species [49].

Chromosome dimerization is a prerequisite for dif/Xer activity and requires the presence of RecA, RecBC, and RecF pathways for homologous recombination between sister chromosomes, RecA being the most efficient for this function [49,4]. Except for *Candidatus Ruthia magnifica* which does not display RecA, RecB or RecF homologs, all other dif-deficient species encode at least one enzyme that may be responsible for chromosome dimerization (Table S4). This observation raises the question of the fate of bacterial cells in which dimerization occurs without the rescue by the dif/Xer system.

**Discussion**

In the present study, 234 chromosomes from 156 proteobacterial species were analyzed for the presence of a dif-related sequence by using a strategy mainly based on homology with experimentally-defined dif sequences and a close association with the chromosome terminus defined by the cumulative GC skew. We now have an overview of the features of the dif/Xer systems present in proteobacteria. Most species display a “classical” dif sequence composed of two undecanucleotides, a conserved dif Xor and a more variable dif Xor separated by an hexanucleotide region (dif Xor). These dif motifs (i) contain inverted repeats forming a palindrom, (ii) are located intergenically, with no apparent specific genetic environment, (iii) have a lower G+C content than the chromosomal G+C content and (iv) are located near the replication terminus as identified in GC skew analyses. These sequences are found in bacteria harboring XerD- and XerC-like recombinases. Other proteobacteria, notably a subgroup of ε-proteobacteria, display a sequence (dif Xor) which is homologous to dif Xor, from streptococci and lactococci [23]. As the canonical dif motif, dif Xor (i) exhibits a low G+C content (ii) is located intergenically, near the terminus defined by the GC skew, (iii) is not associated with specific genetic elements or open reading frames, (iv) displays a palindromic structure and, (v) like dif Xor, can be located in the immediate vicinity of its recombinase. Furthermore, as for the streptococci and lactococci, a single Xer-like recombinase (XerH) was found in species displaying a dif Xor sequence. However, no phylogenetic association between XerS and XerH could be found, which strongly suggests the existence of two unrelated dif/Xer systems. Taken as a whole, these data demonstrate that at least two types of dif/Xer systems exist in proteobacteria: the classical machinery found in most species and an atypical system present in a sub-group of ε proteobacteria. Exhaustive analysis of the dif/Xer systems in other bacterial taxa is now required to evaluate the distribution of these systems in the bacterial kingdom. The general features of dif defined in our study should facilitate this investigation.

Our analysis also demonstrated that the dif/Xer system is not as universal as initially thought. Indeed, 12.2% of the studied proteobacterial species do not harbor this recombination machinery - an absence that could be explained by genome fitting for small genome microorganisms but not for bacteria with large chromosomes (like the Legionellales, *Sachschephas* *degrediens* or *Caldiffus pyroides*). It is presently unclear whether the large chromosome in these microorganisms lost the Xer recombination system, never acquired it or developed a substitutive system to deconcatenate the chromosomes. The consequences of this absence are also intriguing, as most of these dif-deficient species seem to possess the enzymatic machinery (RecA, RecBCD and RecF) potentially responsible for chromosome dimerization by homologous recombination (Table S4) [49,4]. In the absence of dif and Xer recombinase, how do bacterial cells handle chromosome deconcatenation? Can these bacteria survive without the need to resolve chromosome dimers or does an alternative recombination system replace the dif/Xer system? It has already been shown that the loxP/Cre resolvase system (but not res/Tn3) can suppress the filamentation phenotype of a dif-deficient *E. coli* but only when loxP is located at the chromosome terminus [6]. This demonstrates that the dif/Xer machinery can be replaced by other recombination systems. However, there is presently no evidence to suggest that dif/Xer-deficient proteobacteria harbor loxP/Cre resolvase-like systems. In the case of Legionella, the absence of dif/Xer agrees with an early observation showing filamentous cells in *Legionella* cultures [50]. Experimental evidence is now required in order to establish whether the filamentous phenotype in *L. pneumophila* results from the absence of dif/Xer recombination. This question could be answered by reintroducing a functional dif/Xer system into *Legionella* and then checking for the filamentous phenotype.

Compared with other recombination targets, the dif motif harbors a particular structure in view of the presence of two recombinases. It is composed of two recombinase–specific outer regions and two inner regions with dyad symmetry, close to the central hexanucleotide. Our analysis of nucleotide variability in proteobacteria species revealed that the inner regions of dif Xor and dif Xor are highly conserved, whereas the outer regions are much more variable (Figure 1A). Nucleotides at position 23 and 24 (located in the outer part of dif Xor) are highly variable and are rarely part of the palindrom. Interestingly, these positions were experimentally defined in *E. coli* as major contributors to the XerD binding specificity [43] and analysis of the crystal structure of XerD predicted that the dif nucleotide at position 24 interacts directly with the highly conserved amino acid residue Q221 of XerD [51]. Furthermore, this position is much less variable in multi-strain species and multichromosome species (Figure 1B). This observation shows that the variability of the nucleotide at position 24 is primarily inter-species variability and could even be

---

**Figure 5. Alignment of difX and difSL.** The difX sequence corresponds to the putative dif motif of *H. pylori* 26695 (Table 2), whereas difSL was described by Le Bourgeois et al. [23]. Asterisks indicate the common nucleotides and arrows designate inverted repeats.

`doi:10.1371/journal.pone.0006531.g005`
considered as a species marker. Lastly, an adenine residue is highly conserved at position 25 within the outer part of difXerD and could represent a general feature of the dif in proteobacteria.

As for difXerD, the variable outer region of difXerC corresponds to the recombinase binding site, since positions 2 and 3 have been described as major contributors to XerC binding [43]. For the outer difXerC region, our study shows that nucleotides in position 2 of dif display the highest variability, whereas the residue located in position 5 is the least variable. It would be interesting to know whether the most conserved position in dif XerC (position 5) is associated with a conserved amino acid residue in XerC. Unfortunately, structure/function analysis of XerC is prevented by the lack of structural data.

The dif conserved motif was a hexamer in all the proteobacterial genomes that we analyzed, suggesting that the size of the central region separating the recombinase binding sites is a critical feature. In E. coli, it has been demonstrated that the 6 bp-distance between the XerC and XerD binding sites is optimal for chromosomal recombinase activity and cleavage [32,53]. A 8-bp central region is found in natural plasmids like CoIE1 but is always associated with adjacent DNA sequence and accessory proteins [54,55,56]. The presence of a 6-bp central region in proteobacteria thus suggests that chromosomal recombination at dif in these species does not require accessory elements. Furthermore, positions in the central hexamer do not appear to be equivalent. Indeed, our overall analysis suggests that the nucleotide at position 13 within dif conserved motif is highly variable (Figure 1A), whereas it is the least variable residue of the hexanucleotide in the genomes of multi-chromosome bacteria (Figure 1B). Hence, this position may represent an important feature for species discrimination. Moreover, in multi-chromosome species, dif conserved motif is more variable than the dif XerD or dif XerC regions (Figure 1B). This observation agrees with the study by Val et al. [16] and confirms that the central hexanucleotide is a key region for discriminating between chromosomes within the same bacterium and for avoiding chromosome fusion.

This study represents the first comprehensive analysis of the dif motif and its recombinases; it revealed a new dif/Xer recombination system in proteobacteria and constitutes an important step toward the characterization of the dif/Xer-like systems in bacteria with circular chromosomes.

Methods
Identification of dif-like motifs in proteobacteria

The dif-related sequences were identified by using genomic similarity search tools, such as the Basic Local Alignment Search Tool (BLAST) [http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi] [57] and the YASS DNA pairwise alignment tool [http://bioinf.pl.cnrs.fr/yass/yass.php] [58]. Sequences of the experimentally characterized dif elements from E. coli, B. subtilis, C. crescentus, X. campesiris, V. cholerae, streptococci and lactococci [13,14,15,16,23] were used as query sequences. Given that previous studies had revealed conservation of the dif sequences, we used this feature to develop an approach for characterizing dif homologs in phylogenetically related species.

Our analysis of dif-related motifs was performed on all the 234 completed proteobacterial chromosome sequences released before January 1st, 2007 (Table S1). This corresponds to 156 species and represents 53.1% (234 out of 440) of all the bacterial chromosomes sequenced as of that date. The nucleotide sequences were downloaded from the National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov). Forty-eight species were selected as being representative of the different proteobacterial taxae (Table 1). Information on the coding sequences flanking the dif-related sequence was obtained from the protein tables (.ptt file) summarizing the genome annotation at NCBI.

Skew analysis

The position of the candidate dif sequences was compared with that of the DNA replication terminus, as defined by the maximum of cumulative GC nucleotide (CGC) skew obtained by nucleotide skew analyses of the chromosome sequences [59]. When the maximum of CGC skew was undetectable, we chose the maximum of GC skew (GC) of the chromosome or at the first position of the codon (GC1).

Determination of the dif consensus sequence

To define the dif consensus sequence for proteobacteria, we aligned the dif-related sequences extracted from the available chromosomes. To avoid redundancy when several genome sequences were available for one bacterial species, only the information on the first-published chromosome (according to the NCBI release date) was used. In the end, 161 bacterial chromosomes were selected for determination of the dif consensus sequence; since some species have several chromosomes, the number of chromosomes is higher than the number of species. The degree of nucleotide variability (z) at each position of the 28-mer was defined as z = 1 - f, where f is the frequency of the most frequent nucleotide.

Measurement of the degree of variability

The intra-species nucleotide variability of the dif sequence was measured in the 21 dif species represented by at least two strains (termed “multi-strain species” in this study). Intra-species variability was measured at each position of the 28-nucleotide dif sequence. A score of 1 was attributed to the position when the nucleotides differed between strains of the same species; if not, the score at this position was 0 (i.e. conservation). For each position of the 28-nucleotide dif sequence, the scores obtained for all the species were added and normalized against the number of species (n = 21). This value obtained corresponds to the degree of variability at each position and, hence, a low value corresponds to low nucleotide variability at the position.

A similar approach was adopted for analyzing the nucleotide variability of dif in 19 out of 20 multi-chromosome species (listed in Table S3). Multi-chromosome Agrobacterium tumefaciens was not included in the analysis since only one of its two chromosomes display a dif sequence. Within the same strain, chromosomes were compared in terms of dif sequence. A score of 1 was attributed to the position if the nucleotides differed for the 2 or 3 chromosomes in the same strain; if not, the position was scored as 0. Next, for each position, the scores were added and normalized against the number of species (n = 19) to obtain a value representing the degree of variability in multi-chromosome species, a low value being associated with a low nucleotide variability at the position.

Paliindromicity

Paliindromicity was analyzed by comparing the 28-nt dif sequence with its inverted complementary counterpart in the 48 selected proteobacterial species (Table 1). The palindrome was defined as the conserved nucleotide sequence between dif and its inverted, complementary strand. When a nucleotide was found both in dif and in the reverse complementary sequence, a value of 1 was given to the position. Next, the values for the 48 dif sequences for each position were added together to give the n value. The palindromicity frequency (fpal) was then estimated as: fpal = n / 48, with 48 being the number of dif sequences analyzed. A
fpl value of 1 to a nucleotide position means that the nucleotide is always part of a palindrome.

In order to demonstrate that the presence of a palindrome is a key feature of dif motifs, we compared each dif sequence with a randomly generated 28-mer obtained by shuffling the nucleotide of the original dif. Next, the nucleotides involved in the palindrome were counted in both dif and the randomized 28-mers and the average numbers of nucleotide involved in a palindrome were calculated and compared.

BLASTp analysis and the phylogeny of the Xer recombinases

BLASTp analysis were performed using reference amino acid sequences from E.coli K12 XerC and XerD recombinases (protein reference on NCBI: NP_418256 and NP_417370, respectively), from Lactococcus lactis Il1403 XerS (NP_267388) and from E. coli K12 RecA, RecB and RecF (NP_417179, NP_417297.1 and NP_417308, respectively).

Phylogenetic analysis of the Xer recombinases was performed with MEGA version 4 [61]. Sequences were aligned with ClustalW, whereas phylogeny was build using the Neighbor-Joining method [61].

Supporting Information

Figure S1 Palindromicity of the dif-related sequences. The frequency of palindromicity was calculated from the 48 representative dif sequences (Table 1), as described in the Methods section. Black bars represent dif XerC and dif XerD nucleotides, whereas grey bars correspond to dif cent nucleotides. White bars represent nucleotides outside dif.

References

1. Blakely G, May G, McCulloch R, Arciszewska LK, Burke M, et al. (1993) Two related recombinases are required for site-specific recombination at dif and cer in E. coli K12. Cell 75: 351–361.
2. Steiner W, Liu G, Donachie WD, Kuempel P (1999) The cytoplasmic domain of FtsK protein is required for resolution of chromosome dimers. Mol Microbiol 31: 579–583.
3. Aussel I, Barre FX, Aroyo M, Stasiak A, Stasiak AZ, et al. (2002) FtsK is a DNA motor protein that activates chromosome dimer resolution by switching the catalytic state of the XerC and XerD recombinases. Cell 108: 193–205.
4. Kuempel PL, Henson JM, Dircks L, Tecklenburg M, Linn DF (1991) a recA independent recombination site in the terminus region of the chromosome of Escherichia coli. New Biol 3: 799–811.
5. Blakely G, Colloms S, May G, Burke M, Serratir D (1991) Escherichia coli XerC recombinase is required for chromosomal segregation at cell division. New Biol 3: 789–798.
6. Leslie NR, Serratir DJ (1995) Site-specific recombination in the replication terminus region of Escherichia coli: functional replacement of dif and cer. Embry J 14: 1561–1570.
7. Cornet F, Louarn J, Patte J, Louarn JM (1996) Restriction of the activity of the replication site dif to a small zone of the Escherichia coli chromosome. Gene 187: 85–94.
8. Kuempel P, Hogaard A, Nielsen M, Naqappan O, Tecklenburg M (1996) Use of a transposon (Tn651) to obtain suppressing and non-suppressing insertions of the dif resolution site of Escherichia coli. Genes Dev 10: 1162–1171.
9. Peralba K, Corbet F, Merlet Y, Delon I, Louarn JM (2000) Functional polarization of the Escherichia coli chromosome terminus: the dif site acts in chromosome dimer resolution only when located between long stretches of opposite polarity. Mol Microbiol 36: 33–43.
10. Stirling CJ, Stewar G, Serratir DJ (1980) Multicopy plasmid stability in Escherichia coli requires host-encoded functions that lead to plasmid site-specific recombination. Mol Gen Genet 184: 80–84.
11. Clerget M (1991) Site-specific recombination promoted by a short DNA segment of plasmid R1 and by a homologous segment in the terminus region of the Escherichia coli chromosome. New Biol 3: 780–788.
12. Neilson L, Blakely G, Serratir DJ (1999) Site-specific recombination at dif by Haemophilus influenzae XerC. Mol Microbiol 31: 915–926.
13. Sciscioetti SA, Pugot PJ, Blakely GW (2001) Identification and characterization of the dif Site from Bacillus subtilis. J Bacteriol 183: 1058–1068.
14. Jensen RB (2000) Analysis of the terminus region of the Caulobacter crescentus chromosome and identification of the dif site. J Bacteriol 188: 6016–6019.

Acknowledgments

We thank Claire Bertelli for her help with the Bayesian analysis and Michel Simonet, Florent Sebbane and Stephanie Herwegh for critical reading of the paper. We also thank Douglas Berg for his helpful comments and Mohammed Benghezal for his collaboration.

Author Contributions

Conceived and designed the experiments: CC. CAR. Performed the experiments: CC. Analyzed the data: CC. Contributed reagents/materials/analysis tools: CAR. Wrote the paper: CC.

Found at: doi:10.1371/journal.pone.0006531.s001 (6.24 MB TIF)

Table S1 Genome, dif, and Xer recombinase features of 234 probotacteria

Found at: doi:10.1371/journal.pone.0006531.s002 (0.17 MB XLS)

Table S2 Nucleotide frequency (%) of the dif-related sequences from 161 probotacterial chromosomes.

Found at: doi:10.1371/journal.pone.0006531.s003 (0.07 MB DOC)

Table S3 dif-related sequences in multi-chromosome proteobacteria

Found at: doi:10.1371/journal.pone.0006531.s004 (0.15 MB DOC)

Table S4 Xer recombinases and RecA, RecB and RecF homologs in dif-deficient proteobacteria

Found at: doi:10.1371/journal.pone.0006531.s005 (0.03 MB XLS)
42. Slater SC, Goldman BS, Goodher B, Setubal JC, Farrand SK, et al. (2009) Genome sequences of three agrobacterium biocars help elucidate the evolution of multi-chromosome genomes in bacteria. J Bacteriol 191: 2501–2511.

43. Hayes J, Sherratt DJ (1997) Recombinease binding specificity at the chromosome dimer resolution site of Escherichia coli. J Mol Biol 266: 525–537.

44. Kersulyte D, Lee W, Subramaniam D, Anant S, Herrero P, Cabrera L, Balqui J, Barabas O, Kalia A, Gilman RH, Berg DE (2009) Helicobacter pylori's plasticity zones are novel transposable elements. PLoS ONE, in press.

45. Campbell BJ, Engel AS, Porter ML, Takai K (2006) The versatile epsilon-proteobacteria: key players in sulphidic habitats. Nat Rev Microbiol 4: 435–468.

46. Andersson SG (2006) The bacterial world gets smaller. Science 314: 259–260.

47. Giovannoni SJ, Tripp HJ, Givan S, Podar M, Verjin KL, et al. (2005) Genome streaming in a cosmopolitan oceanic bacterium. Science 309: 1242–1245.

48. Krause A, Ramakumar A, Bartels D, Battistoni F, Bekel T, et al. (2006) Complete genome of the mutualistic, N₂-fixing grass endophyte Azanara sp. strain BH72. Nat Biotechnol 24: 1385–1391.

49. Steinr WW, Kuepfer PL (1998) Sister chromlatid exchange frequencies in Escherichia coli analyzed by recombination at the dfr resolution site. J Bacteriol 180: 6289–6275.

50. Brenner DJ, Feeley JC, Weaver RE (1984) Family VII. Legionellaceae. In: Krieg NRH, J.G, eds. Bergey's Manual of Systematic Bacteriology Williams & Wilkins Co.: Baltimore ed. pp 279–288.

51. Subramanya HS, Arisareksa IJ, Baker RA, Bird LE, Sherratt DJ, et al. (1997) Crystal structure of the site-specific recombinase, XerD. Embo J 16: 5178–5187.

52. Blakely GW, Davidson AO, Sherratt DJ (1997) Binding and cleavage of nicked substrates by site-specific recombinases XerC and XerD. J Mol Biol 265: 30–39.

53. Blakely G, Sherratt D (1996) Determinants of selectivity in Xer site-specific recombination. Genes Dev 10: 762–773.

54. Summers DK, Sherratt DJ (1988) Resolution of ColE1 dimers requires a DNA sequence implicated in the three-dimensional organization of the ori site. Embo J 7: 851–856.

55. Stirling CJ, Szatmari G, Stewart G, Smith MC, Sherratt DJ (1988) The arginine repressor is essential for plasmid-stabilizing site-specific recombination at the ColE1 ori locus. Embo J 7: 4389–4393.

56. Stirling CJ, Collins SD, Collins JF, Szatmari G, Sherratt DJ (1989) xerB, an Escherichia coli gene required for plasmid ColE1 site-specific recombination, is identical to pepA, encoding aminopeptidase A, a protein with substantial similarity to bovine lens leucine aminopeptidase. Embo J 8: 1625–1627.

57. Alsctul SF, Gish W, Miller W, Myers EW, Lipman D (1990) Basic local alignment search tool. J Mol Biol 215: 403–410.

58. Nei L, Kucherov G (2005) YASS: enhancing the sensitivity of DNA similarity search. Nucleic Acids Res 33: W580–584.

59. Lobry JR (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. Mol Biol Evol 13: 660–665.

60. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24: 1596–1599.

61. Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406–425.