OXFORD

G3 Genes | Genomes | Genetics

G3, 2021, 11(1), jkaa054
DOI: 10.1093/g3journal/jkaa054
Advance Access Publication Date: 5 January 2021
Investigation

Evolution of Chi motifs in Proteobacteria

Angélique Buton and Louis-Marie Bobay *

Department of Biology, University of North Carolina Greensboro, Greensboro, NC 27402, USA

*Corresponding author: Department of Biology, University of North Carolina Greensboro, 321 McIver Street, Greensboro, NC 27402, USA. ljbobay@uncg.edu

Abstract

Homologous recombination is a key pathway found in nearly all bacterial taxa. The recombination complex not only allows bacteria to repair DNA double-strand breaks but also promotes adaption through the exchange of DNA between cells. In Proteobacteria, this process is mediated by the RecBCD complex, which relies on the recognition of a DNA motif named Chi to initiate recombination. The Chi motif has been characterized in *Escherichia coli* and analogous sequences have been found in several other species from diverse families, suggesting that this mode of action is widespread across bacteria. However, the sequences of Chi-like motifs are known for only five bacterial species: *E. coli*, *Haemophilus influenzae*, *Bacillus subtilis*, *Lactococcus lactis*, and *Staphylococcus aureus*. In this study, we detected putative Chi motifs in a large dataset of Proteobacteria and identified four additional motifs sharing high sequence similarity and similar properties to the Chi motif of *E. coli* in 85 species of Proteobacteria. Most Chi motifs were detected in *Enterobacteriaceae* and this motif appears well conserved in this family. However, we did not detect Chi motifs for the majority of Proteobacteria, suggesting that different motifs are used in these species. Altogether these results substantially expand our knowledge on the evolution of Chi motifs and on the recombination process in bacteria.

Keywords: homologous recombination; bacteria; genome evolution; DNA motifs

Introduction

Bacteria frequently suffer DNA double-strand breaks; and these damages need to be efficiently repaired to avoid cell death. One of the primary repair mechanisms is mediated by homologous recombination which allows the exchange of homologous sequences and thus the repair of damaged DNA (Resnick 1976). Homologous recombination is also a key mechanism contributing to the rapid evolution of bacteria and their viruses (Bobay *et al.* 2013). In Proteobacteria, the main enzymatic complex involved in this process is the RecBCD helicase–nuclease complex (McKittrick and Smith 1989; Rocha *et al.* 2005). Following a double-strand break, RecBCD binds to the broken end of the DNA and unwinds it (Taylor and Smith 1980; Wilkinson *et al.* 2016) upon encountering a Chi (Crossover Hotspot Instigator) site. The recognition of the Chi motif triggers an endonucleolytic nick near the 3′ end of the Chi motif by RecBCD (Smith 2012) and initiates the loading of RecA proteins on the DNA strand (Anderson and Kowalczykowski 1998). The DNA–RecA complex can then initiate the search and exchange of genetic material with a homologous sequence (Smith 2012). The recognition of Chi motifs, mediated by the RecC subunit (Handa *et al.* 2012), thus plays a key function in the initiation of the homologous recombination.

Chi sites have been well defined in *Escherichia coli*, where the sequence of this 8 nt-long motif is 5′-GCTGGTGG-3′ (Smith *et al.* 1981a, 1981b). This motif is overrepresented on the genome with one motif every 5 kb on average (Touzain *et al.* 2011). Chi sites are also more frequent in the conserved regions of the bacterial genome (*i.e.*, the core genome) than in the genes recently acquired (Halpern *et al.* 2007; Cockram *et al.* 2015) and in genomic repeats (Li *et al.* 2019), suggesting the importance of the repair of the regions that encode the most essential cellular functions. This motif is also polarized on the chromosome, which is an important feature for the recognition of Chi sites by RecBCD (Taylor *et al.* 1985): in *E. coli*, Chi motifs are found almost exclusively on the leading strand of replication. Chi sites have an important role in the RecBCD pathway as its recognition by this enzymatic complex is a key step in initiating this mechanism. However, most of the knowledge on Chi sites has been derived from *E. coli* and little is known about this mechanism in other species.

Chi motifs and functionally related motifs (*i.e.*, Chi-like motifs) have only been identified in four other bacteria from diverse taxonomic groups: *Haemophilus influenzae* (5′-GNTGGTGG-3′/5′G(G/C)TGGAGG-3′), *Bacillus subtilis* (5′-AGCGG-3′), *Lactococcus lactis* (5′-GCGCGT-3′) (El Karoui *et al.* 1999), and *Staphylococcus aureus* (5′-GAAGCGG-3′) (Halpern *et al.* 2007). Note that it remains unclear whether the 5′-AGCGG-3′ motif stimulates homologous recombination in *B. subtilis* and that it cannot therefore be considered a true Chi motif. For this reason, we will refer to “Chi motifs” to designate the motifs found in *E. coli* and related bacteria with a similar sequence. In contrast, we will use “Chi-like motifs” to designate all the motifs that are thought to be associated with homologous recombination in prokaryotes. For most of the species with Chi-like sequences identified, Chi-like sequences are statistically overrepresented in the core genome (El Karoui *et al.* 1999). The presence of Chi-like motifs across highly divergent bacteria suggests that this mechanism of action is widespread in bacteria

Received: August 11, 2020. **Accepted:** November 1, 2020

and possibly universal. Moreover, *H. influenzae* and *E. coli* are both Gammaproteobacteria and present related Chi sequences. Although Chi-like sequences have likely evolved toward different sequences in most species, this last case suggests that they are evolving relatively slowly and indicates that it might be possible to identify them with computational approaches.

In this study, we searched for the presence of Chi motifs related to *E. coli*'s Chi motif across hundreds of species of Proteobacteria. By using sequence similarity, oligonucleotide frequency statistics and the known properties of Chi motifs (*i.e.*, polarization on the leading strand and overrepresentation on the core genome), we detected the presence of candidate Chi motifs across 87 species of Proteobacteria. The identified motifs appear restricted to five sequences: (GCTGGTGG, GCTGGCGG, GCTGCTGG, GGTGGTGG, and GCTGGAGG). We further used phylogenomic approaches to reconstruct the evolution of these sequences in Proteobacteria. Our results underline that these sequences are well conserved in *Enterobactericeae*, suggesting a common mode of action of homologous recombination in these species. However, we did not find related sequences across most Proteobacteria, indicating that Chi motifs are either absent in these lineages or possess highly divergent sequences.

Materials and methods

Data collection

Genomic sequences were downloaded from GenBank in May 2018. One genome per species was selected. If multiple genomes were available, we selected the one with the most complete assembly (the one with the smallest number of contigs), and if a species presented multiple assemblies with the same quality, we randomly selected one of the genomes. Following this procedure, we obtained 8924 genomes from distinct species. From this list, we extracted the species belonging to the Proteobacteria and the Terrabacteria, which yielded a total of 2603 species, composed of 1071 and 992 species of Proteobacteria and Terrabacteria, respectively. We then selected the genomes with complete assemblies (*i.e.*, assemblies at the scaffold or the contig level were discarded), and this yielded a total or 495 Proteobacteria and 363 Terrabacteria. Dataset information such as species name, GenBank accession number, taxonomy, genome size, and nucleotide composition (GC-content) are detailed in Supplementary Dataset S1.

Definition of the core genome

It has been previously observed that Chi sites are only statistically overrepresented on the core genome (El Karoui *et al.* 1999), likely because these regions need to be repaired efficiently. In order to improve the sensitivity of our approach, we restricted our analyses of Chi motifs to the core genome of each species. The core genome of a bacterial species is typically defined by all the genes present in every or almost every strain of a species. However, it is not possible to build a core genome for most species, since most species present a single or a few sequenced genome(s). To circumvent this issue, we built a pseudo core genome for each species by using the set of orthologous genes shared among closely related species. Indeed, the core genes of a given species typically encode for the most essential cellular functions (typically housekeeping genes) and those are expected to be shared by closely related species. For each species, we used HMMer v3 (Eddy 2011) and the set of 43 universal protein profiles identified in a previous study (Raymann *et al.* 2015). Each of the universal proteins identified were extracted and aligned with MAFFT v7 (Katoh and Standley 2013). For each alignment, poorly aligned regions were removed using Gblocks v0.91b (Castresana 2000). The 43 universal protein alignments were then concatenated into a single alignment. All pairwise distances were computed on the concatenate and the distances were used to identify all the pairs of most closely related species. For each pair of most related species, the set of shared orthologs was defined as best reciprocal hits using USEARCH v11 (global alignment) with a minimum of 70% sequence identity and 80% length conservation (Edgar 2010). For each species, the pseudo core genome was defined as the set of genes shared by the two most closely related species.

Detection of Chi sites

Candidate Chi motifs were inferred by using multiple criteria. First, we searched for the presence of candidates Chi sites with an identical or highly similar sequence of the motif found in *E. coli* (5′-GCTGGTGG-3′). All the motifs with one nucleotide difference compared with the motif of *E. coli* have been analyzed, namely a total of 25 motifs: GCTGGTGG, ACTGGTGG, CCTGGTGG, TCTGGTGG, GATGGTGG, GGTGGTGG, GTTGGTGG, GCAGGTGG, GCGGGTGG, GCCGGTGG, GCTAGTGG, GCTTGTGG, GCTCGTGG, GCTGATGG, GCTGTTGG, GCTGCTGG, GCTGGAGG, GCTGGCGG, GCTGGGGG, GCTGGTAG, GCTGGTCG, GCTGGTTG, GCTGGTGA, GCTGGTGC, and GCTGGTGT. The pseudo core genomes of the 2063 species were analyzed by the software R'MES v3.1.0 (Schbath and Hoebeke 2011). The statistics of all the possible octamers were performed using a Gaussian law (Prum *et al.* 1995; Schbath *et al.* 1995; Schbath 1997) in R'MES v3.1.0, allowing the calculation of the expected frequency of the motifs relative to the oligonucleotide composition of the genomes. The overrepresentation of each motif was then evaluated by a statistical comparison between the observed and expected numbers using different Markov models to take into account the different levels of oligonucleotide composition of the sequences (different composition bias: mono-nucleotides, di-nucleotides, tri-nucleotides, tetra-nucleotides, penta-nucleotides, and hexa-nucleotides). For each model, a motif was considered as significantly overrepresented based on the statistics observed for the Chi site frequencies of *E. coli* (thresholds used are presented Supplementary Table S1). The detected motifs were considered as candidate Chi motifs when found significantly overrepresented in the core genome with at least three different Markov models and when found polarized on the genome. The polarization statistics were computed with the Wilcoxon signed-rank test implemented in R version 3.5.1. Since we could not confidently identify the origin and terminus of replication for all species, we reasoned that, in the absence of polarization on the genome, Chi sites should be randomly distributed on both DNA strands. If Chi sites are polarized on the chromosome, we should observe a biased distribution where consecutive Chi sites should be found on the same strand more frequently than expected by chance. We compared the distribution of the positions of Chi sites on the positive strand and on the negative strand. We considered that the candidate Chi motifs were polarized on the genome when the distribution was significantly biased ($P < 0.001$, Wilcoxon test) from the null expectation.

Phylogenetic analysis

The concatenated alignment of 43 protein sequences described above was used to build the phylogenetic tree of the Proteobacteria (Raymann *et al.* 2015). We built the phylogenetic tree with a maximum likelihood approach using RAxML v8

(Stamatakis 2014) with the LG model, which was selected as best-scoring model by RAxML. The tree was rooted by including the sequences of three Firmicutes: *Bacillus cereus*, *Lactobacillus casei* and *Staphylococcus saprophyticus*. The resulting tree was edited with colors and legends with iTOL (Letunic and Bork 2007). The Newick format of the tree is available in Supplementary Dataset S2.

Data availability

All genomes analyzed in this study were downloaded from GenBank in May 2018 (https://www.ncbi.nlm.nih.gov/genome/). The list of genome IDs and other information such as species name, taxonomy, genome size and nucleotide composition (GC-content) are provided in Supplementary Dataset S1. The Newick file of the bacterial phylogenetic tree is available in Supplementary Dataset S2. The list of species with polarized candidate Chi motifs (25 polarized motifs searched) is provided in Supplementary Dataset S3. The list of species with non-polarized candidate Chi motifs (25 non-polarized motifs searched) is available in Supplementary Dataset S4. The extended list of species with non-polarized candidate Chi motifs (830 non-polarized motifs searched) is available in Supplementary Dataset S5. The list of all statistically overrepresented octamers is presented in Supplementary Dataset S6 (*de novo* motif identification).

Supplementary material is available at figshare DOI: https://doi.org/10.25387/g3.13202909.

Results

Detection of five Chi motifs in Proteobacteria

We used multiple criteria to detect putative Chi sites in our dataset of 495 Proteobacteria: (1) sequence similarity to the Chi motif of *E. coli*, (2) polarization with the leading strand of replication, and (3) statistical overrepresentation on the core genome relative to the oligonucleotide composition of these genomes (see Methods for details). Because multiple oligonucleotide biases can impact the inference of Chi motifs, we computed the statistical overrepresentation of these motifs in the core genomes with R'MES using different Markov models: from model 0, which accounts for the mononucleotide composition of the core genome, to model 5, which accounts for the hexanucleotide composition of the core genome. A Chi motif was inferred as a candidate when statistically overrepresented in the core genome with three Markov models (see Supplementary Tables S2–S3 for the detailed results across all Markov models). Using these different criteria, we identified 85 species of Proteobacteria with candidate Chi sites (see Supplementary Dataset S3 for details). Across these species, five main motifs were identified as potential Chi motifs in Proteobacteria: GCTGGTGG, GCTGGCGG, GCTGCTGG, GGTGGTGG, and GCTGGAGG (Figure 1). The most frequent motif, GCTGGTGG, is the Chi motif shown to be functional in *E. coli*.

Estimation of false positives

To assess the quality of our detection approach, we used a large sample of Terrabacteria ($n = 363$) to estimate our rate of false positive detection (see Supplementary Tables S1–S3 for the detailed results across all Markov models). Terrabacteria are highly divergent from Proteobacteria but they also exhibit a large diversity of genome size and nucleotide composition. We reasoned that these divergent bacteria are not expected to present similar motifs to Proteobacteria. Indeed, the Chi motifs previously identified in *B. subtilis*, *L. lactis*, and *S. aureus* are not related to the motifs inferred in Proteobacteria. Using the same detection procedure for the 25 putative motifs, we estimated our rate of false positives to be 3.9%. Interestingly, the putative Chi motifs are GC-rich and the false positives were mostly found in species of Terrabacteria with high GC-content, indicating that a weak detection bias due to sequence composition remains even though our framework should compensate for mononucleotide and oligonucleotide compositions.

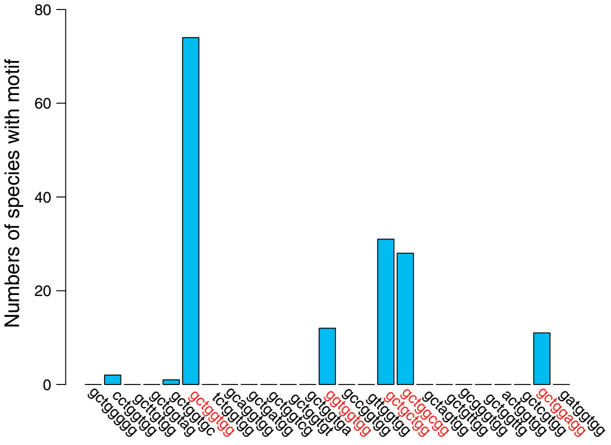

Figure 1 Frequencies and sequences of candidate Chi motifs identified in Proteobacteria. Five main Chi motifs have been detected in a total of 87 species of Proteobacteria based on our criteria.

GC-content and Chi sites

The reference Chi motif of *E. coli* is a GC-rich sequence and all the related Chi motifs that we searched in these genomes present high GC-content (76.7% on average). We observed that the species with inferred candidate Chi motifs present higher genomic GC-content relative to the species without Chi motifs (respectively, 60.8% and 51.5%; $P < 10^{-5}$; Wilcoxon test; Supplementary Figure S1A). This difference was also observed when limiting the comparison to Gammaproteobacteria ($P < 10^{-5}$; Wilcoxon test; Supplementary Figure S1B). However, the genomic GC-content of Proteobacteria is much lower than the GC-content of Chi motifs (respectively, 60.8% and 76.6%; $P < 10^{-5}$; Wilcoxon test; Supplementary Figure S1A). These results suggest that species harboring these Chi motifs tend to be relatively GC-rich, but their GC-content remains much lower than the GC-content of the motif itself.

Evolution of Chi sites in Proteobacteria

We further analyzed the evolution of Chi motifs in Proteobacteria. To do so, we reconstructed the phylogenetic tree of Proteobacteria using a set 43 universal proteins previously identified (Raymann *et al.* 2015). The distribution of species with inferred Chi motifs in the Proteobacteria tree confirms that the vast majority of these species are close relatives of *E. coli*: mostly *Enterobacteriaceae* (Gammaproteobacteria) (Figure 2). Several species of Gammaproteobacteria—other than the *Enterobacteriaceae*—were inferred to possess candidate Chi motifs (*Aeromonas*, *Zobellella*, *Pseudomonas* et *Kushneria*) but their genomes contain a GC-content of about 60%, which suggests that these could represent false positives. In the other groups of Proteobacteria, we detected Chi motifs among Betaproteobacteria (*Paraburkholderia* genus), as well as in Epsilon/Deltaproteobacteria (*Myxococcus* genus) and Alphaproteobacteria (*Azorhizobium caulinodans* and *Komagataeibacter europaeus*). Those species all present a relatively high GC-content (GC% > 60%), also suggesting that the Chi motifs

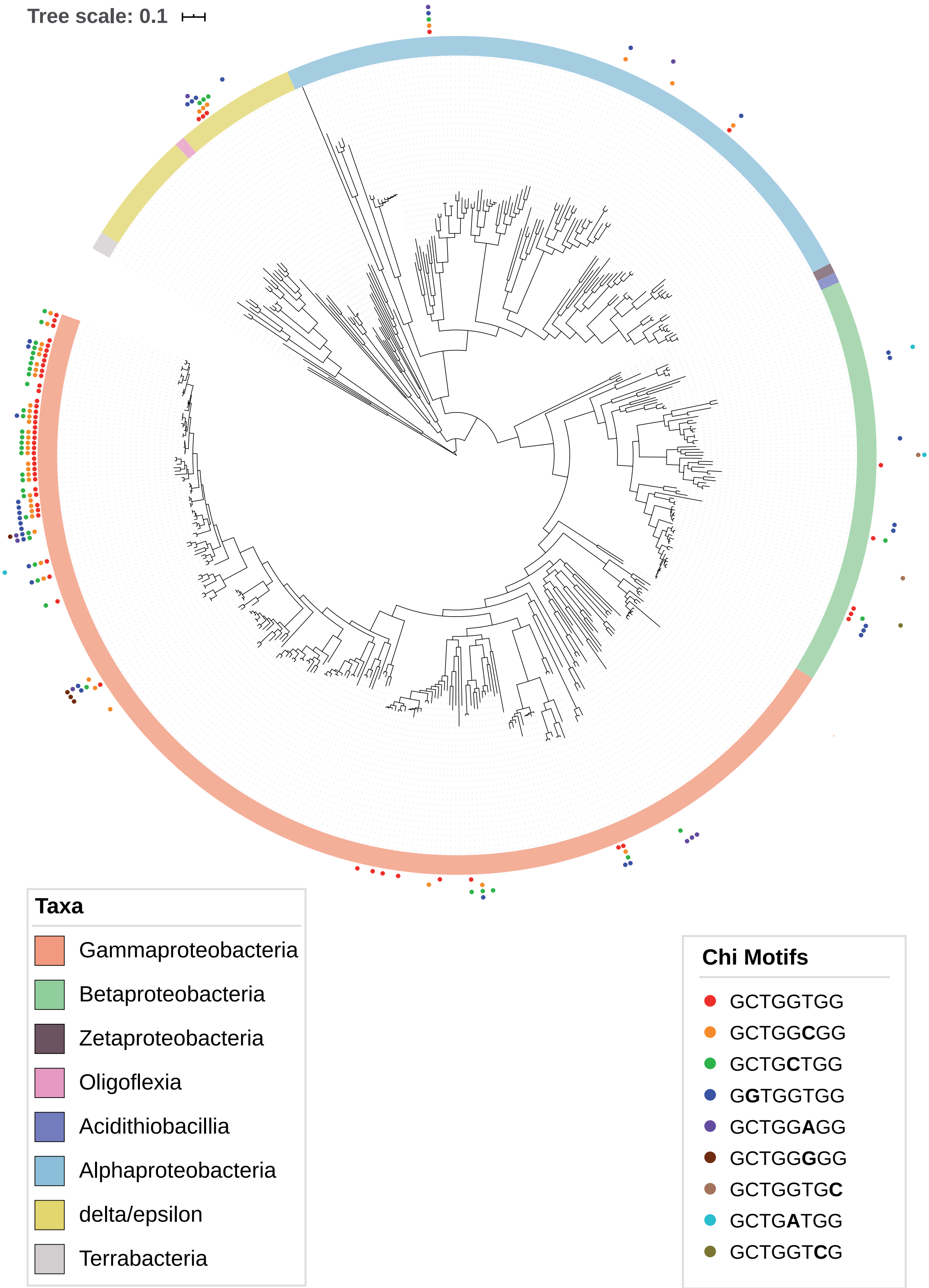

Figure 2 Phylogenetic tree of Proteobacteria and distribution of species with inferred Chi motifs. The tree was inferred using a set of 43 universal proteins and with RaxML (see *Materials and methods*) using the LG substitution model and rooted using three Firmicutes species.

inferred in these species likely represent false positive. For comparison, the GC-content of *E. coli* is 52.3%.

In contrast, virtually all *Enterobacteriaceae* presented Chi motifs. One notable exception is the genus *Yersinia* for which we did not detect the presence of Chi sites for any of the available species (Figure 3). The phylogenetic distribution of species with inferred Chi sites suggests that this motif was present in the ancestor of all *Enterobacteriaceae* and was subsequently lost in

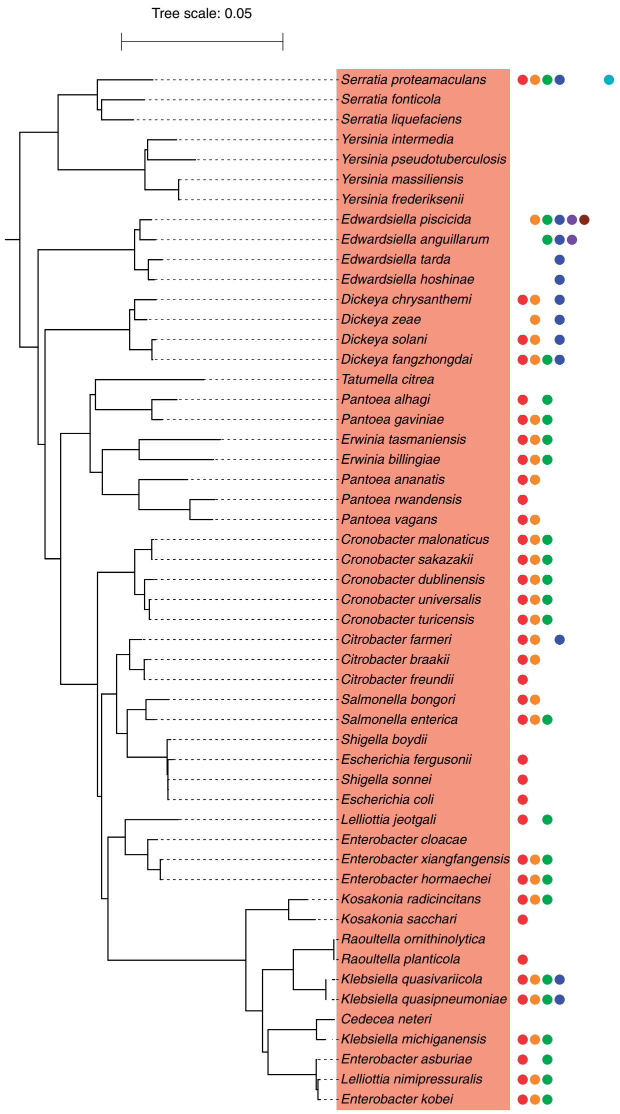

Figure 3 Distribution of candidate Chi motifs in *Enterobacteriacaea*. The subtree was extracted from the tree of Proteobacteria presented in Figure 2 and rooted with non-*Enterobacteriacaea* species.

Yersinia and several *Serratia* species. Alternatively, it is possible that Chi sites were acquired after the divergence of the Yersinia/Serratia lineage and subsequently transferred to *Serratia proteamaculans*. Interestingly, many species of *Enterobacteriaceae* possess multiple Chi motifs. Indeed, three motifs are frequently identified in the same species: GCTGGTGG, GCTGGCGG, and GCTGCTGG. This suggests that several species might recognize the degenerated Chi motif GCTG[G/C][T/C]GG. Finally, some of the more basally branching species of *Enterobacteriaceae* such as members of the genera *Dickeya* and *Edwardsiella* were predominantly found to present the motif GGTGGTGG. These results suggest that Chi motifs are relatively well conserved across *Enterobacteriaceae* but that these sequences have experienced some minor sequence changes over time.

Extending the search of Chi motifs

Our initial analysis was based on stringent parameters to identify candidate Chi sites with high confidence using the same properties observed for *E. coli*'s Chi sites. It is very likely, however, that some Chi sites do not possess similar properties in other species. We conducted the same detection strategy without imposing any constraints on sequence polarity. This extended our list to a total of 234 species of Proteobacteria with candidate Chi motifs (Supplementary Dataset S4) and these motifs were found in 64 Terrabacteria (21% of the species). Although this likely increases our rate of false positives, it is possible that many of these candidate motifs are true Chi motifs that need not be polarized to be functional.

We then extended the search of putative Chi motifs to more divergent sequences. Starting from the five candidate motifs that we detected, we generated 830 additional non-redundant motifs, each containing two different nucleotides relative to one of these five motifs. The same procedure was then used to search for statistically overrepresented motifs based on oligonucleotide compositions on the core genome and sequence polarity. We found very few polarized motifs statically overrepresented relative to oligonucleotide compositions of the genomes (Supplementary Table S4). Using our previous criteria, we did not identify any of these motifs to be polarized and statistically overrepresented using at least three Markov models (Supplementary Table S5). Finally, we searched for these 830 Chi-like motifs without imposing that these motifs should be polarized (*i.e.*, by lifting the polarity criterion). This yielded a list of only 11 additional species of Proteobacteria with unpolarized candidate Chi motifs and 16 species of Terrabacteria (Supplementary Dataset S5), which strongly suggests that these candidate motifs are likely false positives. This result further indicates that reducing the stringency of our parameters does not substantially increase the list of species with candidate Chi motifs.

Inferring *de novo* candidate motifs

Our results suggest that Chi-like motifs in most Proteobacteria are too divergent—or not related—to the candidate Chi motifs found in *Enterobacteriaceae*. This indicates that *de novo* detection of DNA motifs is needed to define new candidate motifs that might be involved in homologous recombination. We used our motif statistics to generate a list of new 8-nt long motifs that likely represent functional sequences based on their overrepresentation in the genome and their frequency across species of Proteobacteria. The list was generated using the same criteria of motif frequency from mononucleoide composition to hexanucleotide composition, as used above to infer Chi motifs similar to *E. coli*'s motif (criteria defined in Supplementary Table S1). Interestingly, we identified 75 motifs that were found more frequently that the Chi motif of *E. coli* (Supplementary Dataset S6) and the vast majority of these motifs presented little similarity to Chi-like motifs defined above. The most commonly identified motif (AGGCGCTG) was found statistically overrepresented in 39% of all Proteobacteria (average determined across the six Markov models). We aligned the 75 most common motifs using Muscle (no internal gaps were allowed) and we identified that the 4-nt palindrome GCGC was found overrepresented across these motifs. Although they might not be involved in homologous recombination, these motifs constitute preferential targets for future experimental analyses aiming at characterizing new functional motifs in Proteobacteria.

Discussion

This study suggests that a relatively small proportion of Proteobacteria use Chi motifs. We found that a minority of Gammaproteobacteria (12.5% of the total Proteobacteria) is likely using a similar Chi motif to the one of *E. coli*. However, this study was limited to a list of motifs closely related to the one found in *E. coli*, *i.e.*, a list of motifs with a one nucleotide difference (25 motifs). Among the 25 sequences, only five were found frequently in the Proteobacteria: 5′-GCTGGTGG-3′, 5′-GCTGGCGG-3′, 5′-GCTGCTGG-3′, 5′-GGTGGTGG-3′, and 5′-GCTGGAGG-3′. Some bacterial species are known to completely lack (*e.g.*, *Buchnera aphidicola*) or show a strong reduction in homologous recombination (*e.g., Yersinia pestis*) (Cui and Song 2016) and is it theoretically possible that Chi sites might be completely absent in additional species. This last scenario remains highly unlikely given the high prevalence of homologous recombination in bacteria (Vos and Didelot 2009; Bobay and Ochman 2017). In addition, some species of Proteobacteria, such as Alphaproteobacteria, rely on the AddBC recombination system instead of RecBCD (Rocha *et al.* 2005). However, given the focus of our procedure on Chi sites related to *E. coli*'s motif, it is very likely that more divergent—or unrelated—Chi-like sequences exist in Proteobacteria. Searching for additional motifs could enable the identification of Chi sites in these species. This endeavor remains difficult due to the sensitivity of *ab initio* detections and that no Chi-like motifs unrelated to *E. coli*'s motif have been identified in Proteobacteria so far. In addition, it may not be possible at all to detect such motifs using computational approaches if these sequences are simply not statistically overrepresented in the genome (see later section).

Given the large dataset used for this study, it is likely that some Chi motifs identified correspond to detection errors (*i.e.*, false positives). By using a large set of Terrabacteria species, which are thought to use different motifs (El Karoui *et al.* 1999), we identified our rate of false positives to reach about 4% of detected motifs. The detection of false positives seems to be closely linked to the richness in GC-content since most false positives were found in GC-rich genomes (*e.g.*, Actinobacteria). Although the nucleotide composition biases were taken into account by the statistics generated by R'MES, it seems that these filters did not completely eliminate all the false positives due to nucleotide composition. This suggests that the candidate Chi motifs inferred in GC-rich species should be taken with care.

The evolution of Chi sites and other DNA motifs is still poorly understood; however, we can conclude that species related to *E. coli* likely use very similar motifs, indicating that these motifs have been conserved for millions of years. Based on our phylogenetic analysis we can infer that this motif appeared at least since the divergence between the *Enterobacteriaceae* and the rest of the Gammaproteobacteria and that the sequence of the motif evolved in some of these taxa. In agreement with our results, it was previously observed that several species of *Enterobacteriaceae* were capable of Chi-dependent cleavage of DNA constructs containing *E. coli*'s Chi motif (Schultz and Smith 1986; Smith *et al.* 1986). Our results extend these findings by showing that most *Enterobacteriaceae* contain polarized and statistically overrepresented Chi motifs in their core genomes and that some of these species likely use slightly different motifs than *E. coli*'s motif (or a degenerated version of this motif). However, we found that the majority—but not all—species of *Enterobacteriaceae* use Chi motifs similar to *E. coli*'s motif. Indeed, we did not detect the presence of polarized Chi motifs in *Vibrionaceae*, although several of these species were shown capable of Chi-dependent cleavage of DNA constructs containing *E. coli*'s Chi motif (Schultz and Smith 1986). However, we identified a non-polarized candidate Chi motif (GCTGGTGG) in 23 species of *Vibrio* (Supplementary Dataset S4), indicating that polarity might not be required for the proper function of Chi motifs in these species. We did not detect candidate Chi motifs in *H. influenzae* (*Pasteurellaceae*), although Chi motifs were identified for this species in a previous study (El Karoui *et al.* 1999). In fact, the original study that identified Chi sites in *H. influenzae* (El Karoui *et al.* 1999) showed that these motifs are degenerated, less represented statistically and less polarized than the ones found in *E. coli*, even though their function has been experimentally validated (Sourice *et al.* 1998). We used stringent criteria to avoid the detection of false positives, but it is possible that some species present Chi motifs too degenerated to be confidently identified using *in silico* approaches. Furthermore, it is likely that other species of Proteobacteria contain Chi sites with other sequences than those we searched for in this study. It is possible that the stringency of our approach prevented us from identifying additional motifs but relaxing our search criteria directly leads to a much higher rate of false positives. Our results do not constitute a proof that most Proteobacteria are devoid of Chi or Chi-like motifs, but our analyses rather suggest that their putative motifs must substantially differ from *E. coli*'s motif. Considering that some experimentallycharacterized motifs, such as the ones of *H. influenzae*, virtually escape to all bioinformatic detection procedure, it is possible that only an experimental approach would allow to confidently detect new Chi-like motifs in Proteobacteria. Together, these results contribute to better characterizing the signals and pathways involved in homologous recombination in bacteria, as well as the evolution of genomic motifs.

Overall, our results indicate that Chi motifs similar to *E. coli*'s motif evolved at least since the ancestor of *Enterobacteriaceae* and were lost in several lineages. Some of these species likely use motifs whose sequence is slightly different from *E. coli*'s motif. Using *in silico* approaches we did not confidently detect such motifs in other species of Proteobacteria, suggesting that these lineages use unrelated sequences or that these motifs do not present all the characteristics of *E. coli*'s motif (*i.e.*, polarization on the leading strand). Experimental analyses will be needed to confirm that these motifs can efficiently trigger RecBCD-mediated homologous recombination. Overall this study revealed new candidate Chi motifs and extends our understanding of the mechanisms controlling homologous recombination in Proteobacteria.

Acknowledgments

We thank Shealynn Burgess for assistance with the analyses, Kasie Raymann for criticisms on the manuscript, and Caroline Stott for assistance with the figures.

L.-M.B. and A.B. designed the study, analyzed the data, and wrote the manuscript. Both authors read and approved the final manuscript.

Funding

This work was supported by the National Science Foundation under Grant No. DEB-1831730 by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R01GM132137.

Conflicts of interest: None declared.

Literature cited

Anderson DG, Kowalczykowski SC. 1998. Reconstitution of an SOS response pathway: derepression of transcription in response to DNA breaks. Cell. **95**:975–979.

Bobay LM, Ochman H. 2017. Biological species are universal across life's domains. Genome Biol Evol. **9**:491–501.

Bobay LM, Rocha EP, Touchon M. 2013. The adaptation of temperate bacteriophages to their host genomes. Mol Biol Evol. **30**:737–751.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. **17**:540–552.

Cockram CA, Filatenkova M, Danos V, Karoui ME, Leach DR. 2015. Quantitative genomic analysis of RecA protein binding during DNA double-strand break repair reveals RecBCD action in vivo. Proc Natl Acad Sci USA. **112**:E4735–E4742.

Cui Y, Song Y. 2016. Genome and evolution of *Yersinia pestis*. Adv Exp Med Biol. **918**:171–192.

Eddy SR. 2011. Accelerated profile HMM searches. PLoS Comput Biol. **7**:e1002195.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. **26**:2460–2461.

El Karoui M, Biaudet V, Schbath S, Gruss A. 1999. Characteristics of Chi distribution on different bacterial genomes. Res Microbiol. **150**:579–587.

Halpern D, Chiapello H, Schbath S, Robin S, Hennequet-Antier C, *et al.* 2007. Identification of DNA motifs implicated in maintenance of bacterial core genomes by predictive modeling. PLoS Genet. **3**: 1614–1621.

Handa N, Yang L, Dillingham MS, Kobayashi I, Wigley DB, *et al.* 2012. Molecular determinants responsible for recognition of the single-stranded DNA regulatory sequence, Chi, by RecBCD enzyme. Proc Natl Acad Sci USA. **109**:8901–8906.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. **30**:772–780.

Letunic I, Bork P. 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics. **23**: 127–128.

Li C, Danilowicz C, Tashjian TF, Godoy VG, Prevost C, *et al.* 2019. The positioning of Chi sites allows the RecBCD pathway to suppress some genomic rearrangements. Nucleic Acids Res. **47**:1836–1846.

McKittrick NH, Smith GR. 1989. Activation of Chi recombinational hotspots by RecBCD-like enzymes from enteric bacteria. J Mol Biol. **210**:485–495.

Prum B, Rodolphe F, de Turckheim E. 1995. Finding words with unexpected frequencies in deoxyribonucleic acid sequences. J Royal Stat Soc. **57**:205–220.

Raymann K, Brochier-Armanet C, Gribaldo S. 2015. The two-domain tree of life is linked to a new root for the Archaea. Proc Natl Acad Sci USA. **112**:6670–6675.

Resnick MA. 1976. The repair of double-strand breaks in DNA; a model involving recombination. J Theor Biol. **59**:97–106.

Rocha EPC, Cornet E, Michel B. 2005. Comparative and evolutionary analysis of the bacterial homologous recombination systems. PLoS Genet. **1**:e15.

Schbath S. 1997. An efficient statistic to detect over- and under-represented words in DNA sequences. J Comput Biol. **4**: 189–192.

Schbath S, Hoebeke M. 2011. R'MES: A Tool to Find Motifs with a Significantly Unexpected Frequency in Biological Sequences. Singapore: World Scientific.

Schbath S, Prum B, Turckheim E. D. 1995. Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences. J Comput Biol. **2**:417–437.

Schultz DW, Smith GR. 1986. Conservation of Chi cutting activity in terrestrial and marine enteric bacteria. J Mol Biol. **189**:585–595.

Smith GR. 2012. How RecBCD enzyme and Chi promote DNA break repair and recombination: a molecular biologist's view. Microbiol Mol Biol Rev. **76**:217–228.

Smith GR, Comb M, Schultz DW, Daniels DL, Blattner FR. 1981a. Nucleotide sequence of the Chi recombinational hot spot Chi +D in bacteriophage lambda. J Virol. **37**:336–342.

Smith GR, Kunes SM, Schultz DW, Taylor A, Triman KL. 1981b. Structure of Chi hotspots of generalized recombination. Cell. **24**: 429–436.

Smith GR, Roberts CM, Schultz DW. 1986. Activity of Chi recombinational hotspots in *Salmonella typhimurium*. Genetics. **112**:429–439.

Sourice S, Biaudet V, Karoui ME, Ehrlich SD, Gruss A. 1998. Identification of the Chi site of *Haemophilus influenzae* as several sequences related to the *Escherichia coli* Chi site. Mol Microbiol. **27**: 1021–1029.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. **30**: 1312–1313.

Taylor A, Smith GR. 1980. Unwinding and rewinding of DNA by the RecBC enzyme. Cell. **22**:447–457.

Taylor AF, Schultz DW, Ponticelli AS, Smith GR. 1985. RecBC enzyme nicking at Chi sites during DNA unwinding: location and orientation-dependence of the cutting. Cell. **41**:153–163.

Touzain F, Petit MA, Schbath S, Karoui ME. 2011. DNA motifs that sculpt the bacterial chromosome. Nat Rev Microbiol. **9**:15–26.

Vos M, Didelot X. 2009. A comparison of homologous recombination rates in bacteria and archaea. ISME J. **3**:199–208.

Wilkinson M, Chaban Y, Wigley DB. 2016. Mechanism for nuclease regulation in RecBCD. Elife. **5**:e18227.

Communicating editor: M. Zetka