Identification of DNA Motifs Implicated in Maintenance of Bacterial Core Genomes by Predictive Modeling

David Halpern[1], Hélène Chiapello[2], Sophie Schbath[2], Stéphane Robin[3], Christelle Hennequet-Antier[2], Alexandra Gruss[1], Meriem El Karoui[1*]

1 INRA, UR888, Unité des Bactéries Lactiques et pathogènes Opportunistes, Jouy en Josas, France, **2** INRA, UR1077, Unité Mathématique, Informatique, et Génome, Jouy en Josas, France, **3** AgroParisTech/INRA, UMR518, Unité Mathématiques et Informatique appliquées, Paris, France

Bacterial biodiversity at the species level, in terms of gene acquisition or loss, is so immense that it raises the question of how essential chromosomal regions are spared from uncontrolled rearrangements. Protection of the genome likely depends on specific DNA motifs that impose limits on the regions that undergo recombination. Although most such motifs remain unidentified, they are theoretically predictable based on their genomic distribution properties. We examined the distribution of the "crossover hotspot instigator," or Chi, in *Escherichia coli,* and found that its exceptional distribution is restricted to the core genome common to three strains. We then formulated a set of criteria that were incorporated in a statistical model to search core genomes for motifs potentially involved in genome stability in other species. Our strategy led us to identify and biologically validate two distinct heptamers that possess Chi properties, one in *Staphylococcus aureus,* and the other in several streptococci. This strategy paves the way for wide-scale discovery of other important functional noncoding motifs that distinguish core genomes from the strain-variable regions.

Citation: 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(9): e153. doi:10.1371/journal.pgen.0030153

Introduction

Analyses of bacterial pan genomes reveal a high level of genetic diversity, even within a short evolutionary scale [1,2]. This raises the question of how bacterial chromosomes remain organized yet allow recombination events to occur. Noncoding functional DNA motifs have been implicated in bacterial genome maintenance, although their identification is rare and limited to studies in very few organisms. Where examined, they tend to be species or genus specific. Some examples are the highly frequent DNA uptake sequences, involved in discriminating self from foreign entering DNA during competence in *Haemophilus influenzae* and *Neisseria meningitidis* [3], the crossover hotspot instigator (Chi), first identified in *E. coli,* involved in recombinational repair [4,5], and a recently characterized chromosome dimer resolution motif named KOPS (FtsK orienting polar sequences) in *E. coli* [6,7].

Where studied, motifs with a biological function appear to occur nonrandomly in the genome. However, prediction of genomic motifs with biological function based on their distribution remains a challenging question and, to our knowledge, few approaches exist to answer it. Most currently developed motif-finding methods are targeted towards discovery of sequences involved in gene expression, which presupposes a constrained position of motifs with respect to genes. The approaches employed are not suitable for identification of genome organization motifs, which tend to be scattered around the chromosome. In this work, we formulate the following hypotheses upon which we devise a strategy to predict such motifs: (i) Natural selection of such motifs guarantees their overall distribution, rather than their occurrence at specific positions [8]. Prediction criteria for such sequences should therefore be based on motif distribution properties on complete genomes. (ii) Functional motifs should be conserved across strains of the same species, which can be assessed through comparative genome analysis. As recently shown, bacterial genomes can be segmented into a core genome ("backbone") conserved among all strains of the same species, and strain-variable regions ("loops") [1,9]. Motifs related to general cellular processes would be expected to be enriched on the backbone.

A predictive approach for motif identification was undertaken, and the parameters for this approach were based on a well-documented bacterial motif, the Chi site, an essential component of double-strand break repair [4]. Chi was first identified as the sequence 5′-GCTGGTGG-3′ in *E. coli* [5], where it is a key modulator of enzymatic activities of the RecBCD complex [10,11]. Chi orientation-dependent recognition by RecBCD [12] leads to repair of genomic breaks incurred during DNA replication [13]. Most sequenced bacterial genomes encode functional RecBCD analogs [14,15]. However, Chi motifs are not conserved between species and only a few have been identified [16]. We successfully applied

Editor: Ivan Matic, Université Paris V, France

Received April 16, 2007; **Accepted** July 23, 2007; **Published** September 14, 2007

Abbreviations: Chi, crossover hotspot instigator; HMW, high molecular weight multimer; RC, rolling circle

* To whom correspondence should be addressed. E-mail: Meriem.El_Karoui@jouy.inra.fr

Author Summary

Availability of bacterial "pan genomes," based on multiple genome sequences of a given species, has revealed the existence of core genomes, but also of high levels of variable, nonconserved regions. The nature of bacterial strategies that assure genome organization while permitting biodiversity remains an intriguing question. A first clue in addressing this question comes from growing evidence for the existence of noncoding functional DNA motifs with specific distribution properties along the chromosome, which are implicated in DNA integrity. In this work, we addressed the challenging problem of predicting such motifs from pan-genome information. We analyzed characteristics of the "crossover hotspot instigator" motif, Chi, which, in *E. coli,* is an 8-base-pair sequence involved in chromosome maintenance. Our results show that Chi has specific distribution properties that are restricted to core genomes common to related *E. coli* strains. Using statistical modeling combined with comparative genomics, we then predicted the identity of Chi in core genomes of *S. aureus* and several streptococci, and confirmed them in vivo. The strategy developed in this study may be extended to reveal and characterize novel functional motifs, and should be instrumental in analyzing genome biodiversity mechanisms.

this approach to identify two sequences having Chi activity, one in *S. aureus* and another in three streptococcal species.

Results/Discussion

Chi Overrepresentation and Skew Are Characteristic of the *E. coli* Backbone

A set of criteria was generated to describe Chi distribution characteristics in the *E. coli* genome. A backbone/loop segmentation of the *E. coli* genome was performed by complete genome alignment of commensal K-12 [17], enterohemorragic O157:H7Sakai [1], and uropathogenic CFT073 [18] strains. The backbone is 3.7 Mb long and each strain carries several hundred loops (Table 1; data accessible on the MOSAIC database Web site http://genome.jouy.inra.fr/mosaic/). Chi motif frequency and overrepresentation were assessed separately on backbone and loop segments (the loop segment correspond to all loops concatenated). Chi is highly frequent on the backbone (its frequency is 1 every 6 kb, noted 1/6kb), but not on loops (Table 1). Statistical significance of these frequencies was checked by comparing observed and expected Chi frequencies based on composition in mono- to hepta-nucleotides of the backbone or of the loops. This was done by calculating an overrepresentation *p*-value based on a Gaussian approximation of short motif counts under a Markov model of order 6 [19]. On the backbone, Chi is the most overrepresented octamer among 65,536 possible sequences. In contrast, Chi occurrence on *E. coli* loops is not exceptional and is explained by loop DNA composition (Table 1). Moreover, comparison of all octamers distributed on the complete genome versus the backbone showed that other octamers appeared to be more overrepresented than Chi on the complete genome (its overrepresentation rank is 5 compared to a highest overrepresentation rank of 1). These were filtered out when analyses were restricted to the backbone (see motifs in red circle, Figure S1). This suggests that the analysis of backbone helps reduce the "noise" produced by motifs that are overrepresented due to the loops, and thus eliminates the candidates that are unlikely to have Chi activity. Interestingly, among these motifs, we identified octamers that are part of the BIME elements (a family of repeated elements that is significantly enriched on the loops [9]). This result supports the assumption that restricting the motif search to backbone facilitates identification of biologically important genomic motifs such as Chi.

Table 1. Chi Is Overrepresented on the *E. coli* Backbone but Not on the Loops

Sequence Features		Backbone	K-12 Loops	O157:H7 Sakai Loops	CFT073 Loops
Total length (Mb)		3.7	0.9	1.7	1.5
Number of loops		—	887	868	851
Chi Distribution Properties	Freq	1/6 kb	1/11 kb	1/11 kb	1/14 kb
	p_F	$2\ 10^{-24}$	ns	ns	ns
	Rank	1	260	3615	465
	Skew	0.77	0.68	0.71	0.68
	p_S	10^{-5}	ns	ns	ns

Distribution properties were assessed on the leading strand of the backbone and loop set of each strain. A p_F higher than 6×10^{-8} is reported as not significant (ns). A p_S higher than 5×10^{-2} is reported as not significant (ns).
Freq, frequency on leading strand; p_F, overrepresentation *p*-value; p_S, skew *p*-value; rank, overrepresentation rank for the candidate motif compared to all possible octamers; skew, observed skew.
doi:10.1371/journal.pgen.0030153.t001

In *E. coli*, 75% of Chi sites are skewed towards the replicative leading strand [16,20,21], in keeping with their function in stimulating double-strand break repair upon replication fork collapse [13,22]. As the *E. coli* genome shows a strong GC skew, and Chi is G-rich, this bias might simply reflect the underlying GC skew. This is not the case with Chi skew; when compared to that expected from mono-nucleotide composition, it was found to be significant on the *E. coli* backbone, beyond what might be expected due to leading-strand bias (Figure 1).

Prediction of Chi on the *S. aureus* Backbone

The above results validate our initial hypotheses concerning motif distribution for the *E. coli* Chi site. We therefore

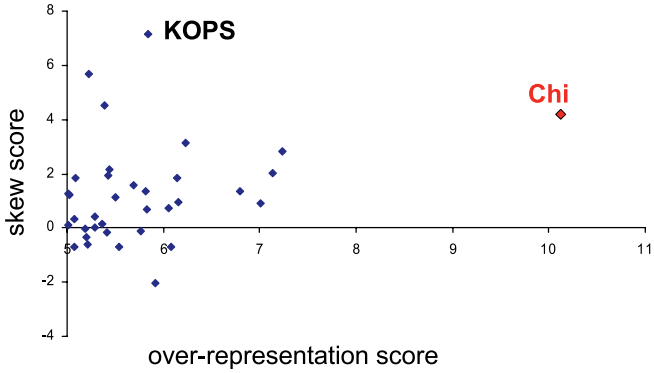

Figure 1. The *E. coli* Chi site, 5′-GCTGGTGG-3′, Is Significantly Overrepresented and Skewed on *E. coli* Backbone DNA

Statistical scores of overrepresentation and skew are plotted for all octamers whose overrepresentation scores are higher than 5 (corresponding to a *p*-value $< 10^{-7}$). Chi and one of the motifs of the KOPS family [6] are indicated.
doi:10.1371/journal.pgen.0030153.g001

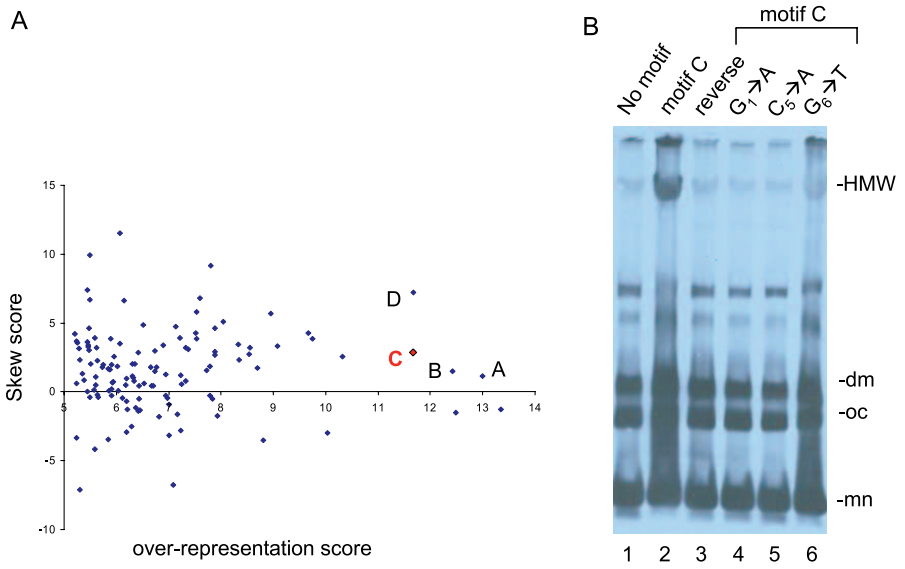

Figure 2. Motif C, 5′-GAAGCGG-3′, Is the *S. aureus* Chi Site

(A) Overrepresentation and skew of heptamers on the *S. aureus* backbone. Plots are represented as in Figure 1. Only heptamers with an overrepresentation score higher than 5 (corresponding to a *p*-value $< 10^{-7}$) are shown. The four best candidates (motifs A to D) are indicated.
(B) Experimental validation that motif C has Chi activity in *S. aureus*. Motif C was cloned in plasmid pRC and evaluated for HMW accumulation, an indicator of Chi activity [27]. Total genomic DNA was hybridized with a probe specific to the RC plasmid. Plasmid monomer (mn), dimer (dm), and open circle (oc) forms, and HMW are indicated by arrows. Motif C (lane 2, 5′-GAAGCGG-3′) is HMW-positive, but it gives a negative signal when cloned in the reverse orientation (lane 3) or when it carries a mismatch (lanes 4–6).
doi:10.1371/journal.pgen.0030153.g002

used its distribution properties as prediction criteria for Chi identification by an in silico approach in species where its sequence was unknown. We first chose the human pathogen *S. aureus*, for which numerous available genome sequences reveal a high level of intraspecies diversity [23]. Identification of Chi and characterization of its properties could provide insight into genome plasticity in this problematic pathogen. The *S. aureus* backbone was defined from an alignment of six *S. aureus* genomes (Mu50, Mw2, N315, COL, MRSA252, and MSSA476). The 2.44-Mb backbone corresponds to 86 % of the mean genome length. It was used to analyze oligomer distribution on the replicative leading strand. We looked for motifs that were (i) significantly overrepresented; (ii) frequent, with average frequencies higher than 1 in 15 kb; and (iii) skewed (≥60%) with a significant score. Chi sites previously identified in different bacteria ranged in length from 8 to 5 nucleotides [24–26]. We started by analyzing octamers, as longer motifs can be more readily reduced to find the minimal active motif than the converse. None of the overrepresented and skewed octamers were frequent enough to be retained as potential candidates. We thus focused on heptamers (Figure 2A). Four motifs were strikingly over-represented, of which motif C (5′-GAAGCGG-3′) and D (5′-GAATTAG-3′) matched the skew score criteria. As motifs A (5′-GAAAATG-3′) and B (5′-GGATTAG-3′) had skew scores only slightly lower than the threshold (1.13 and 1.54, respectively) they were also retained as potential candidates.

The Motif 5′-GAAGCGG-3′ Is the Functional Chi Site of *S. aureus*

The selected motifs were tested for *S. aureus* Chi activity using a biological screening assay previously used to identify Chi in *Lactococcus lactis*, *H. influenzae*, and *Bacillus subtilis* [24–26]. Briefly, a rolling circle (RC) plasmid may generate a double-strand DNA extremity during replication, which provides an entry point for RecBCD-like enzymes [27]. This extremity is degraded unless the plasmid contains a correctly oriented Chi site, in which case degradation is aborted, and high molecular weight multimers (HMW) accumulate. Motifs A and D were present in both orientations in the initial RC plasmid, which does not accumulate HMW in *S. aureus* (Figure 2B, lane 1), thus ruling them out as *S. aureus* Chi site candidates. Motif B was also ruled out, as it did not provoke HMW when cloned in the RC plasmid (unpublished data). However, insertion of an oligonucleotide containing motif C into the RC plasmid vector resulted in a strong HMW signal (Figure 2B, lane 2). The RC plasmid containing the inverse orientation of motif C did not generate HMW (Figure 2B, lane 3), in keeping with the orientation dependency of the *E. coli* Chi [12]. Chi sites identified in other bacteria vary in length and degeneracy [24–26]. It was thus possible that the *S. aureus* Chi motif necessitated only a subsequence of the identified seven-nucleotide motif. Substitutions at any of the seven defined nucleotide positions tested negative for HMW when present on the RC plasmid. Specifically, alterations in the first or last nucleotide of the motif were also negative (Table S3; Figure 2B, lanes 4–6). We conclude that the seven-nucleotide motif 5′-GAAGCGG-3′ is necessary and sufficient to confer Chi activity in *S. aureus*.

The Chi Sequence Is Overrepresented Only on the *S. aureus* Backbone

The *S. aureus* Chi site is very frequent and overrepresented on the backbone (1/9 kb; *p*-value 10^{-28}, rank 5), and 88% of Chi motifs are on the leading strand. As in *E. coli*, over-representation of the *S. aureus* Chi site did not apply to loops (Table 2). Chi site occurrence in *S. aureus* loops ranged from 1/15 kb to 1/20 kb, and in only Mu50 loops were they slightly

Table 2. Chi Is Overrepresented on the *S. aureus* Backbone

Sequence Features		Backbone	N315 Loops	Mu50 Loops	Mw2 Loops	MRSA252 Loops	MSSA476 Loops	Col Loops
Total length (Mb)		2.44	0.37	0.43	0.37	0.45	0.35	0.36
Number of loops		—	700	699	702	703	696	699
Chi Distribution Properties	Freq	1/9kb	1/19kb	1/15kb	1/17kb	1/18kb	1/20kb	1/18kb
	p_F	10^{-28}	ns	ns	ns	ns	ns	ns
	Rank	5	369	42	159	135	399	217
	Skew	0.88	0.68	0.83	0.9	0.78	0.9	0.91
	p_s	1.8×10^{-3}	ns	ns	ns	ns	ns	ns

Distribution properties were assessed on the leading strand of the backbone and loop set of each strain. A p_F higher than 6×10^{-8} is reported as not significant (ns). A p_S higher than 5×10^{-2} is reported as not significant (ns).

Freq, frequency on leading strand; p_F, overrepresentation *p*-value; p_S, skew *p*-value; rank, overrepresentation rank for the candidate motif compared to all possible octamers; skew, observed skew.

doi:10.1371/journal.pgen.0030153.t002

overrepresented (*p*-value 8×10^{-6}, rank 42), although levels were below our threshold values. As observed in *E. coli*, comparison of heptamer distribution on strain N315 complete genome versus backbone shows that some motifs appear overrepresented on the complete genome mostly due to their presence on loops, and that these motifs are filtered out by analysis on the backbone (motifs in red circle, Figure S2).

The staphylococcal Chi sequence comprises a nonactive submotif, which was previously identified as the five-nucleotide Chi site of *B. subtilis*, 5′-AGCGG-3′ [26] (Table S3). Analyses on the *B. subtilis* complete genome show that its Chi motif is very frequent (1/0.6 kb) and significantly overrepresented (*p*-value 6×10^{-19}, rank 54) but does not fully conform to criteria described for Chi in *E. coli* and *S. aureus*, with respect to skew significance (its skew is only 58% and is not statistically significant). We considered the possibility that distribution of the *S. aureus* Chi site on the *B. subtilis* genome would meet the criteria we described for Chi. This is not the case; the motif, although frequent (1/4.4 kb), is not overrepresented (*p*-value not significant, rank 2,959). Analysis of two other longer motifs (5′-AAGCGGC-3′ and 5′-AGCGGCGC-3′), reported in [26] as having very high Chi activity, shows that they are likewise not overrepresented (respective ranks 11,872 and 4,258, both with nonsignificant *p*-values). A more thorough analysis will require additional *B. subtilis* genome sequences to extract backbone DNA, but both the shortness of the sequence and its genome features in this species leads us to suggest that Chi properties may not be universal, and that Chi evolution in this bacterial branch may undergo different selective pressures leading to a role for Chi on both DNA strands.

Prediction and Validation of a Chi Candidate Active in Streptococci

The generality of the Chi prediction method was challenged by examination of four species of the *Streptococcus* genus: pathogens *Streptococcus pyogenes* and *Streptococcus pneumoniae*, the opportunist pathogen *Streptococcus agalactiae*, and the nonpathogen *Streptococcus thermophilus*. These species belong to the same family as the nonpathogenic food bacterium, *L. lactis*. A backbone was constructed for each of the species (resulting from an alignment of two to six genomes; see Table 3 for details) and heptamer distribution was analyzed. The best candidate fulfilling the prediction criteria is the same in each of the four species (Table 3), and corresponds exactly to the *L. lactis* Chi site [24]. This prediction was confirmed by experimental validation in *S. agalactiae*, *S. thermophilus*, and *S. pneumoniae* (Figure S3; Table S4). As observed in *E. coli* and *S. aureus*, Chi is mostly not overrepresented on loops of streptococci (Table 3). These results indicate that five species of the Streptococcaceae family seem to share the Chi site independent of their pathogenicity status, which leads us to speculate that DNA protection by Chi evolved prior to functional differentiation of these bacteria. More generally, a comparison of all known Chi motifs in the context of the phylogenetic relationships of the corresponding species (Figure 3) indicates at least partial Chi conservation among closely related bacteria, regardless of their niche or pathogenic status.

Why Is Chi Overrepresented Only on the Backbone?

Chi has been considered as a key element in preventing chromosomal fork collapse during replication [13], thus ensuring faithful DNA break repair. The observed enrichment of Chi on backbone DNA suggests that Chi plays an important role in maintaining integrity of backbone regions that presumably encode essential functions. The much lower occurrence of Chi motifs on loops (of which the largest are up to 100 kb), likely reflects their exogenous origin. This might result in less efficient repair and hence lower stability of these regions. Conversely, moderate Chi overrepresentation in certain loop sets (e.g., in *S. aureus* strain Mu50) might suggest that these loops are older and have persisted by conferring a selective advantage to the bacterium. Thus, Chi enrichment on backbone, which is common to all species we examined, probably reflects the evolutionary balance between genome stability and diversification through horizontal transfer.

Prediction of Functional Motifs with Remarkable Features on Genomes

The approach reported here to predict genome organization motifs is based on a combination of comparative genomics and statistical analysis of motif distribution properties. It may prove useful in the discovery of other genomic motifs with known or novel functions via their specific characteristics in the complete genome, or in defined genomic regions. In addition to Chi, analysis of the *S. aureus* backbone also identified that heptamer motifs B (5′-GGAT-

Table 3. A common Chi motif is conserved among pathogenic and non-pathogenic Streptococci, and is not over-represented on loops.

Species	Chi Distribution Properties	Backbone	Strain 1 Loops	Strain 2 Loops	Strain 3 Loops	Strain 4 Loops	Strain 5 Loops	Strain 6 Loops
S. agalactiae **GCGCGTG**	Freq	1/18 kb[a]	1/378 kb	1/42 kb	1/63 kb	—	—	—
	p_F	1.2×10^{-15}	ns	ns	ns	—	—	—
	Rank	5	2373	32	3540	—	—	—
	Skew	0.82	1	0.44	1	—	—	—
	p_s	7×10^{-3}	ns	ns	ns			
S. pneumoniae **GCGCGTG**	Freq	1/8 kb	1/21 kb	1/25kb	—	—	—	—
	p_F	1×10^{-34}	ns	ns	—	—	—	—
	Rank	1	217	573	—	—	—	—
	Skew	0.85	1	0.62	—	—	—	—
	p_s	3×10^{-5}	ns	Ns	—	—	—	—
S. pyogenes GCGCGTG	Freq	1/12 kb	1/25 kb	1/29 kb	1/108 kb	1/44 kb	1/27 kb	1/107 kb
	p_F	5×10^{-21}	ns	ns	ns	ns	ns	ns
	Rank	2	89	33	2606	737	480	2013
	Skew	0.77	0.65	0.65	1	0.86	0.67	3
	p_s	4×10^{-3}	ns	ns	ns	ns	ns	ns
S. thermophilus **GCGCGTG**	Freq	1/11.2 kb	1/26 kb	No motif	—	—	—	—
	p_F	3×10^{-25}	ns	—	—	—	—	—
	Rank	1	1371	—	—	—	—	—
	Skew	0.81	0.5	—	—	—	—	—
	p_s	2×10^{-4}	ns	—	—	—	—	—

Strains analyzed: *S. agalactiae* 1, NEM316; 2, A909; and 3, 2603VR. *S. thermophilus* 1, CNRZ1066 and 2, LMG18311. *S. pneumoniae* 1, R6; and 2, TIGR4. *S. pyogenes* 1, M1GAS; 2, MGAS10394; 3, MGAS315; 4, MGAS5005; 5, MGAS6180; and 6, MGAS8232. Abbreviations as in Table 1. Motifs (written 5′ to 3′) in bold were experimentally confirmed to have Chi activity in *S. agalactiae*, *S. pneumoniae,* and *S. thermophilus*.

[a]In *S. agalactiae,* the frequency of Chi is slightly lower than the selected threshold (1/15kb)

doi:10.1371/journal.pgen.0030153.t003

TAG-3′) and D (5′-GAATTAG-3′) were overrepresented on the backbone and skewed. Like Chi, these motifs are essentially not remarkable on the loops (Table S5). Interestingly, motif B is also highly overrepresented and skewed in two species related to *S. aureus: Bacillus cereus* and *Bacillus anthracis*. These results suggest that this motif (or the degenerated motif 5′-G(G/A)ATTAG-3′) might have a single biological function in several related species, which remains to be discovered. We recently used statistical analysis to characterize the KOPS family of octamers in *E. coli*, which orient the chromosome for segregation during cell division [6]. Like Chi, KOPS are overrepresented, frequent, and skewed on the backbone (Figure 1). Characteristics of the non-Chi motif identified here make it a tempting potential candidate to test as the KOPS homolog in *S. aureus*. Using the strategy presented here, KOPS and other biologically characterized repeated sequences could be modeled to discover motifs with similar roles in very distant bacteria.

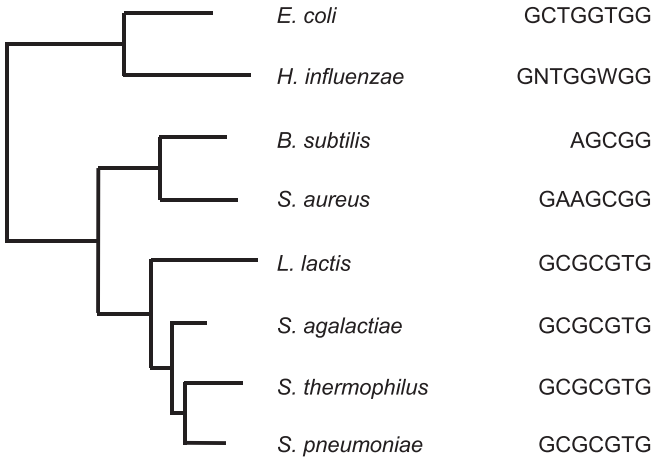

Figure 3. Chi Sequence Conservation through Evolution

The tree shows phylogenetic relationships of the eight bacterial species in which Chi is known, based on 16S RNA, and was constructed using the tree-builder command of the Ribosomal Database Project (https://rdp.cme.msu.edu/). *E. coli, H. influenzae, L. lactis,* and *B. subtilis* Chi sites were previously identified [5,25,26]. Chi sites of *S. aureus* and the three streptococci were identified by predictive modeling and confirmed experimentally (this work). Partial conservation of Chi correlates with phylogenetic conservation among bacteria.

doi:10.1371/journal.pgen.0030153.g003

Materials and Methods

Genome analysis. *Sequence data.* We recently described a method for genome segmentation into backbone and loops based on multiple genome alignments [9]. In *E. coli,* the backbone/loops segmentation resulted from comparison of strains K-12, O157:H7 Sakai, and CFT073. Note that these analyses could not be performed in other species where Chi is known, due to the absence of appropriate data that would allow construction of a reliable backbone. For *S. aureus* segmentation, we used strains Mu50, Mw2, N315, COL, MRSA252, and MSSA476. Data are available at http://genome.jouy.inra.fr/mosaic/. Leading strands were defined as the DNA strand reported in Genbank files downstream of the replication origin up to the terminus and the reverse complement strand from the replication terminus to the origin. In *E. coli,* the position of the replication origin is annotated for the K-12 strain, and was determined in the other strains by a sequence similarity search. The origin position was previously mapped and assigned position 1 in all the other species. The terminus position was chosen as the first nucleotide of the chromosome dimer resolution *dif* site as described for *E. coli* and *B. subtilis* [28,29]. The *S. aureus, B. cereus* and, *B. anthracis dif*-like sites were identified as sequences similar to the *B. subtilis dif* site 5′-ACTTCCTAGAATATATATTATGTAAACT-3′, (allowing two mismatches, one insertion, and one deletion), by performing a PATSCAN

analysis [30] on the Micado database (http://genome.jouy.inra.fr/micado). For the four streptococci, we used the same strategy using the streptococcus-specific *dif* site identified by P. Le Bourgeois [31]. The terminus positions thus determined coincided with those reported by other methods [8,32,33].

Motif overrepresentation score. Motif count analyses were performed on the leading strand of the backbone common to the aligned genomes for a given species, and on the leading strand of all loops of each strain. To assess overrepresentation of short motifs, the observed count of each motif was compared to the count expected in random sequences showing the same oligonucleotide composition. The significance of the difference between the counts was evaluated by calculating the associated *p*-value, which is the probability that the count of a given motif "**u**" in a random sequence (under a chosen stationary Markov model, see below) is greater than the observed count for this motif. The *p*-value is obtained using a Gaussian approximation of motif counts [34], which has been shown to be reliable for short motifs in comparatively long sequences [35]. For a motif "**u,**" the *p*-value denoted $p_F(\mathbf{u})$ is given by:

$$p_F(\mathbf{u}) = P(X \geq [N_{obs}(\mathbf{u}) - EN(\mathbf{u})]/\gamma)$$

where X is distributed according to the standard Gaussian distribution $N(0,1)$, $N_{obs}(\mathbf{u})$ the observed count of **u**, $EN(\mathbf{u})$ is the estimated expected count of **u** in the chosen model, and γ is an explicit normalizing factor [34].

Note that $[N_{obs}(\mathbf{u}) - EN(\mathbf{u})]/\gamma$ can be directly interpreted as an overrepresentation score: the higher its positive value, the more overrepresented is the motif **u**.

The random sequence models are Markov chain models that take into account the nucleotide monomer, dimer, trimer, etc., $(m + 1)$-mer composition of the sequence, depending on the order m that is used; When examining Chi among *E. coli* octamers we chose the maximal model *M*6 (which takes into account the monomer to heptamer nucleotide composition). Analyses on heptamers in *S. aureus* and streptococci were performed using the *M*5 model (based on monomer to hexamer composition). Parameters of those models were estimated separately on the backbone and the loops to take into account the difference in oligomer composition of loops compared to backbone. We checked that in all cases the sequences were long enough to reliably estimate the parameters of the model. Note that *p*-values obtained on the backbone and loops cannot be directly compared because of differences in the sequence lengths [36]. We therefore present the overrepresentation ranks (in which the overrepresentation score of a given motif is compared to all possible motifs of the same length), which are directly comparable.

Calculations were made using R'MES software (http://genome.jouy.inra.fr/ssb/rmes/), which provides scores for all motifs of a given length, and ranks them according to their overrepresentation score. The Bonferroni correction was used to choose the significance threshold; for a given significance level (10^{-3} in our case), the motif *p*-values have to be divided by the number of tests (16,384 for heptamers and 65,536 for octamers, respectively). Thus, a *p*-value is significant when it is $\leq 6 \times 10^{-8}$ (corresponding score 5.2) for a given heptamer, and $\leq 1.5 \times 10^{-8}$ (corresponding score 5.6) for a given octamer.

Motif skew score. The significance of the skew for a motif "**u**" was also determined by calculating a *p*-value, namely the probability that its skew in a random sequence is greater than the observed skew S_{obs}. Since the skew is the ratio between the counts of **u** and that of $\bar{\mathbf{u}}$ ($\bar{\mathbf{u}}$ is the reverse complement of **u**), the *p*-value denoted $p_S(\mathbf{u})$ was obtained using a Gaussian approximation of motif counts as above. More precisely, we have:

$$p_S(\mathbf{u}) = P(X \geq -\mu/\sigma),$$

where X is distributed according to the standard Gaussian distribution $N(0,1)$,

$$\mu = EN(\mathbf{u}) - S_{obs}EN(\bar{\mathbf{u}}), \text{ and}$$

$$\sigma^2 = \mathrm{Var}(N(\mathbf{u})) - 2S_{obs}\mathrm{cov}(N(\mathbf{u}), N(\bar{\mathbf{u}})) + S_{obs}^2\mathrm{Var}(N(\bar{\mathbf{u}})).$$

We chose a Markov model of order 0 to take into account the richness in G of the leading strand, which is due to G/C skew. Under such a model, the expectation, variance, and covariance of motif counts are easily derived [34].

As above, the quantity $-\mu/\sigma$ can be directly interpreted as a score of skew exceptionality: The higher the positive value, the higher the exceptionality of the skew. These calculations can also be obtained with the third version of the R'MES software (http://genome.jouy.inra.fr/ssb/rmes/). Because skew is only a secondary criterion for Chi activity, we used a much less stringent significance level ($p < 0.05$, corresponding score 1.6) for both heptamers and octamers.

Criteria for motif selection. We selected the most overrepresented motifs on the backbone leading strand that were also frequent (frequency higher than 1/15kb), skewed (skew higher than 60%), and with a skew *p*-value lower than 0.05. This allowed us to construct a list of potential candidates ordered by decreasing overrepresentation *p*-value.

Experimental validation of Chi activity. The strains and plasmids used in this study are listed in Table S1. *S. aureus* strain RN4220 (an avirulent strain that accepts foreign DNA by transformation of competent cells) was grown in Brain Heart Infusion medium with aeration at 37 °C. *S. pneumoniae* and *S. agalactiae* were grown in Todd Hewitt yeast extract at 37 °C. *S. thermophilus* strains were grown in M17 medium supplemented with 1% lactose at 42 °C. *E. coli* strain TG1 was used for plasmid constructions. Chloramphenicol (10 μg/ml in *S. aureus*, 20 μg/ml in *E. coli*, 3 μg/ml in *S. thermophilus*, 5 μg/ml in *S. pneumoniae*, and 7 μg/ml in *S. agalactiae*) was used for plasmid selection. Plasmid DNA preparation, PCR amplifications, and DNA modifications were carried out according to suppliers' instructions. *S. aureus* cell suspensions were lysed using 100 μg/ml lysostaphin (Sigma) for 30 min at 37 °C as described [37]. DNA transformation was performed by electroporation [38].

Oligonucleotide cloning. Oligonucleotides (Table S2) were inserted at the HincII/EcoRI sites of the multiple cloning site of pRC_2 (Table S1) using standard methods. Plasmid constructions were carried out in *E. coli* and confirmed by DNA sequencing. The pACYC184 replication origin present on pRC_2 was then deleted by NciI/PvuII digest, giving rise to a pRC plasmid carrying the designed oligonucleotide. The resulting plasmid was then transferred to *S. aureus*.

Detection of Chi activity by HMW accumulation. The RC plasmid used for Chi activity detection in *S. aureus* is based on the pVS41 (pC194) replicon [39]. It was derived from pRC_2 deleted for its pACYC184 replication origin by NciI/PvuII digest, Klenow fill-in, and self ligation, giving rise to pRC. pRC carrying oligonucleotides containing putative Chi motifs (see above for construction) was used for HMW detection. Total DNA was extracted from strains carrying different constructions and HMW detected by Southern blot hybridization as described [40].

To identify mutations leading to abolition of Chi activity, we used pRC that carried oligonucleotides differing from Chi by a single point mutation (see above for construction and Table S2). We also made use of pRC constructs containing random 1–2-kb NciI/PvuII DNA fragments derived from *Pseudomonas aeruginosa* PAO1 that were screened for HMW accumulation; DNA inserts of plasmids giving no HMW signal were sequenced and checked against the complete PAO1 genome sequence. Sequences were screened for motifs that differed by single mutations from the putative Chi site.

The RC plasmid used for Chi activity detection in *S. agalactiae* and *S. thermophilus* is pRC_2. Its derivative containing an oligonucleotide with the putative streptococcal Chi site (5′-GCGCGTG-3′) is called pRC_2Chi_2 [40]. HMW accumulation was detected by Southern blot hybridization as above.

Detection of Chi activity by measurement of transformation efficiency. Attempts to visualize HMW in *S. pneumoniae* by Southern blot hybridization were not reproducible, despite numerous efforts and modifications. We therefore used a test previously described for Chi identification in *B. subtilis* [26]. Briefly, in *B. subtilis*, multimeric plasmid molecules transform cells better than monomeric or dimeric species. Therefore, HMWs, being composed of multiple repeats of plasmid molecules, confer increased transformation ability to plasmid extracts. We monitored the amount of HMW by measuring transformation efficiencies in *B. subtilis* (strain 168) of total DNA extracts from different *S. pneumoniae* strains relative to that of a wild-type strain carrying the pRC_2 vector. Total DNA extracts were quantified by spectrophotometric dosage at 260 nm and calibration against known quantities of λ BstEII on agarose gels. Total DNA extracts were performed as described and standard transformation techniques were used [26].

Supporting Information

Figure S1. Comparison of Octamer overrepresentation on the *E. coli* K-12 Strain Complete Genome versus Backbone

Overrepresentation scores on the complete genome (*y*-axis) and the backbone (*x*-axis) are plotted. Among the most overrepresented motifs of the complete genome, a significant number are not

overrepresented on the backbone (in red circle). Note that among these motifs, we identified several motifs that are submotifs of BIME elements; BIMEs comprise a family of repeated elements specifically enriched on loops [9]. The motif shown in red is Chi.

Found at doi:10.1371/journal.pgen.0030153.sg001 (24 KB PDF).

Figure S2. Comparison of Heptamer Overrepresentation on the *S. aureus* N315 Strain Complete Genome versus Backbone

Overrepresentation scores on the complete genome (*y*-axis) and the backbone (*x*-axis) are plotted. As seen above for *E. coli,* among the most overrepresented motifs of the complete genome, a significant number are not overrepresented on the backbone (in red circle). The motif shown in red is Chi.

Found at doi:10.1371/journal.pgen.0030153.sg002 (22 KB PDF).

Figure S3. The *L. lactis* Chi Site Is Active in *S. agalactiae* and *S. thermophilus*

The candidate Chi motif cloned in plasmid pRC_2 [40] was transformed into *S. agalactiae* and *S. thermophilus,* and evaluated for HMW accumulation, an indicator of Chi activity [40]. Total genomic DNA was hybridized with a probe specific to the pRC_2 plasmid. Plasmid monomer (mn), dimer (dm), and open circle (oc) forms, and HMW are indicated. The candidate (5′-GCGCGTG-3′) is HMW positive (lanes 2 and 4). Note that Chi activity of 5′-GCGCGTG-3′ in *S. thermophilus* was not observed in our previous study [24]. In this species, cloning of the Chi motif in pRC_2 induced a strong decrease in plasmid copy number. To reveal Chi activity, sample concentrations were corrected such that the same quantity of pRC_2 and pRC_2Chi_2 plasmid monomers was loaded on gels.

Found at doi:10.1371/journal.pgen.0030153.sg003 (21 KB PDF).

Table S1. Strains and Plasmids Used in This Study

Found at doi:10.1371/journal.pgen.0030153.st001 (37 KB DOC).

Table S2. Oligonucleotides Used in This Study

Found at doi:10.1371/journal.pgen.0030153.st002 (30 KB DOC).

Table S3. Point Mutations in the *S. aureus* Chi Motif Abolish Activity

The Chi substitutions shown were present on different RC plasmids tested in this study for HMW accumulation. In all cases, single base alterations in Chi abolished HMW accumulation. Bottom line corresponds to the *B. subtilis* Chi site.
[a]HMW generated from RC plasmids containing a properly oriented Chi site [27].
+, HMW is accumulated; −, no HMW (compare to Figure 2B).

Found at doi:10.1371/journal.pgen.0030153.st003 (35 KB DOC).

Table S4. The *L. lactis* Chi site Is Active in *S. pneumoniae*

Total DNA extracts from the wild-type strain carrying a plasmid with the putative Chi site show increased transformation efficiency, indicative of HMW accumulation by this plasmid. As previously observed in *B. subtilis,* transformation efficiency of plasmid extracts from a *rexB* strain that accumulates HMW due to recombinase *rexAB* inactivation is intermediate [26].
[a]Average of three experiments; [b]transformation of the wild-type strain with pRC_2 gives an average of 3.3×10^{-2} transformants per μg of DNA.

Found at doi:10.1371/journal.pgen.0030153.st004 (22 KB DOC).

Table S5. Distribution Properties of Two Significant Non-Chi Motifs in *S. aureus*

Distribution properties of motifs (written 5′ to 3′) were assessed on the leading strand of the backbone and each loops set. A p_F higher than 6×10^{-8} is reported as not significant (ns). A p_S higher than 5×10^{-2} is reported as not significant (ns).
Freq, frequency on leading strand; p_F, overrepresentation *p*-value; rank, overrepresentation rank for the candidate motif compared to all possible octamers; skew, observed skew; p_S, skew *p*-value.

Found at doi:10.1371/journal.pgen.0030153.st005 (36 KB DOC).

Accession Numbers

The National Center for Biotechnology Information (NCBI) Genbank (http://www.ncbi.nlm.nih.gov/sites/gquery) accession number of the genomes used in this study are as follows: *E. coli* strains K-12, O157:H7 Sakai, and CFT073: U000096, BA000007, and AE14075, respectively; *S. aureus* strains Mu50, Mw2, N315, COL, MRSA252, and MSSA476: BA000017, BA000033, BA000018, CP000046, BX571856, and BX571857, respectively; *S. agalactiae* strains NEM316, A909, and 2603VR: AL732656, CP000114, and AE009948, respectively; *S. thermophilus* strains CNRZ1066 and LMG18311: CP000024 and CP000023, respectively; *S. pneumoniae* strains R6 and TIGR4: AE007317 and AE005672, respectively; *S. pyogenes* strains M1GAS, MGAS10394, MGAS315, MGAS5005, MGAS6180, and MGAS8232: AE004092, CP000003, AE014074, CP000017, CP000056, and AE009949, respectively.

Acknowledgments

We are grateful to Annie Gendrault (INRA) for help in database management and the MIGALE bioinformatics platform (INRA) for providing computational resources. We acknowledge Pascal Le Bourgeois for communicating data prior to publication. We thank Philippe Bouloc, Marie-Agnes Petit, Francois-Xavier Barre, and Simonetta Gribaldo for suggestions and critical reading of the manuscript. We thank the anonymous referees for their insightful comments.

Author contributions. AG and MEK conceived and designed the experiments. DH performed the experiments. SS and MEK analyzed the data. HC, SS, SR, and CHA contributed reagents/materials/analysis tools. SS, AG, and MEK wrote the paper.

Funding. This work was supported in part by French grants from the Programme de Recherche Fondamentale en Microbiologie et Maladies Infectieuses et Parasitaires, Action Concertée Incitative Informatique Mathématique et Physique pour la Biologie (IMPBio) of the French Ministère de l'Education Nationale, de la Recherche et de la Technologie, and the Programme inter-EPST bioinformatique.

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

References

1. Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K, et al. (2001) Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. DNA Res 8: 11–22.
2. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R (2005) The microbial pan-genome. Curr Opin Genet Dev 15: 589–594.
3. Smith HO, Gwinn ML, Salzberg SL (1999) DNA uptake signal sequences in naturally transformable bacteria. Res Microbiol 150: 603–616.
4. Lam ST, Stahl MM, McMilin KD, Stahl FW (1974) Rec-mediated recombinational hot spot activity in bacteriophage lambda. II. A mutation which causes hot spot activity. Genetics 77: 425–433.
5. Smith GR, Kunes SM, Schultz DW, Taylor A, Triman KL (1981) Structure of chi hotspots of generalized recombination. Cell 24: 429–436.
6. Bigot S, Saleh OA, Lesterlin C, Pages C, El Karoui M, et al. (2005) KOPS: DNA motifs that control *E. coli* chromosome segregation by orienting the FtsK translocase. Embo J 24: 3770–3780.
7. Levy O, Ptacin JL, Pease PJ, Gore J, Eisen MB, et al. (2005) Identification of oligonucleotide sequences that direct the movement of the *Escherichia coli* FtsK translocase. Proc Natl Acad Sci U S A 102: 17618–17623.
8. Hendrickson H, Lawrence JG (2006) Selection for chromosome architecture in bacteria. J Mol Evol 62: 615–629.
9. Chiapello H, Bourgait I, Sourivong F, Heuclin G, Gendrault-Jacquemard A, et al. (2005) Systematic determination of the mosaic structure of bacterial genomes: species backbone versus strain-specific loops. BMC Bioinformatics 6: 171.
10. Anderson DG, Kowalczykowski SC (1997) The recombination hot spot chi is a regulatory element that switches the polarity of DNA degradation by the RecBCD enzyme. Genes Dev 11: 571–581.
11. Kowalczykowski SC, Dixon DA, Eggleston AK, Lauder SD, Rehrauer WM (1994) Biochemistry of homologous recombination in *Escherichia coli.* Microbiol Rev 58: 401–465.
12. 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.
13. Kuzminov A (1995) Collapse and repair of replication forks in *Escherichia coli.* Mol Microbiol 16: 373–384.
14. Rocha EP, Cornet E, Michel B (2005) Comparative and evolutionary analysis of the bacterial homologous recombination systems. PLoS Genet 1: e15. doi:10.1371/journal.pgen.0010015
15. Petit M-A (2005) Mechanisms of homologous recombination in bacteria. In: Mullany P, editor. The dynamic bacterial genome: Cambridge: Cambridge University Press. pp. 3–32.
16. El Karoui M, Biaudet V, Schbath S, Gruss A (1999) Characteristics of Chi distribution on different bacterial genomes. Res Microbiol 150: 579–587.

17. Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V, et al. (1997) The complete genome sequence of *Escherichia coli* K-12. Science 277: 1453–1474.
18. Welch RA, Burland V, Plunkett G 3rd, Redford P, Roesch P, et al. (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. Proc Natl Acad Sci U S A 99: 17020–17024.
19. Schbath S (1997) An efficient statistic to detect over- and under-represented words in DNA sequences. J Comput Biol 4: 189–192.
20. Salzberg SL, Salzberg AJ, Kerlavage AR, Tomb JF (1998) Skewed oligomers and origins of replication. Gene 217: 57–67.
21. Uno R, Nakayama Y, Arakawa K, Tomita M (2000) The orientation bias of Chi sequences is a general tendency of G-rich oligomers. Gene 259: 207–215.
22. Bidnenko V, Ehrlich SD, Michel B (2002) Replication fork collapse at replication terminator sequences. EMBO J 21: 3898–3907.
23. Lindsay J, Holden M (2006) Understanding the rise of the superbug: Investigation of the evolution and genomic variation of *Staphylococcus aureus*. Funct Integr Genomics: 6: 186–201.
24. Biswas I, Maguin E, Ehrlich SD, Gruss A (1995) A 7-base-pair sequence protects DNA from exonucleolytic degradation in *Lactococcus lactis*. Proc Natl Acad Sci U S A 92: 2244–2248.
25. Sourice S, Biaudet V, El Karoui M, 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.
26. Chedin F, Noirot P, Biaudet V, Ehrlich SD (1998) A five-nucleotide sequence protects DNA from exonucleolytic degradation by AddAB, the RecBCD analogue of *Bacillus subtilis*. Mol Microbiol 29: 1369–1377.
27. Dabert P, Ehrlich SD, Gruss A (1992) Chi sequence protects against RecBCD degradation of DNA in vivo. Proc Natl Acad Sci U S A 89: 12073–12077.
28. Sciochetti SA, Piggot PJ, Blakely GW (2001) Identification and characterization of the dif site from *Bacillus subtilis*. J Bacteriol 183: 1058–1068.
29. Kuempel PL, Henson JM, Dircks L, Tecklenburg M, Lim DF (1991) dif, a recA-independent recombination site in the terminus region of the chromosome of *Escherichia coli*. New Biol 3: 799–811.
30. Dsouza M, Larsen N, Overbeek R (1997) Searching for patterns in genomic data. Trends Genet 13: 497–498.
31. Le Bourgeois P, Bugarel M, Campo N, Daveran-Mingot M-L, Labonté J, et al. (2007). The unconventional Xer recombination machinery of streptococci/lactococci. PLoS Genet. 3: e17. doi:10.1371/journal.pgen.0030117
32. Kuroda M, Ohta T, Uchiyama I, Baba T, Yuzawa H, et al. (2001) Whole genome sequencing of meticillin-resistant *Staphylococcus aureus*. Lancet 357: 1225–1240.
33. Worning P, Jensen LJ, Hallin PF, Staerfeldt HH, Ussery DW (2006) Origin of replication in circular prokaryotic chromosomes. Environ Microbiol 8: 353–361.
34. Robin S, Rodophe F, Schbath S (2005) DNA, words and models. Cambridge: Cambridge University Press. 158 p.
35. Robin S, Schbath S (2001) Numerical comparison of several approximations of the word count distribution in random sequences. J Comput Biol 8: 349–359.
36. Robin S, Schbath S, Vandewalle V (2007) Statistical tests to compare motif count exceptionalities. BMC Bioinformatics 8: 84.
37. Rigoulay C, Entenza JM, Halpern D, Widmer E, Moreillon P, et al. (2005) Comparative analysis of the roles of HtrA-like surface proteases in two virulent *Staphylococcus aureus* strains. Infect Immun 73: 563–572.
38. Kraemer GR, Iandolo JJ (1990) High-frequency transformation of *Staphylococcus aureus* by electroporation. Current Microbiology V21: 373–376.
39. von Wright A, Saarela M (1994) A variant of the staphylococcal chloramphenicol resistance plasmid pC194 with enhanced ability to transform *Lactococcus lactis subsp. lactis*. Plasmid 31: 106–110.
40. el Karoui M, Ehrlich D, Gruss A (1998) Identification of the lactococcal exonuclease/recombinase and its modulation by the putative Chi sequence. Proc Natl Acad Sci U S A 95: 626–631.