Associations Between Inverted Repeats and the Structural Evolution of Bacterial Genomes

Guillaume Achaz,*,[1] Eric Coissac,*,[2] Pierre Netter* and Eduardo P. C. Rocha†,‡,[3]

**Structure et Dynamique des Génomes, Institut Jacques Monod, 75251 Paris, France, ‡Unité GGB, CNRS URA 2171, Institut Pasteur, 75015 Paris, France and †Atelier de BioInformatique, Université Pierre et Marie Curie, 75005 Paris, France*

Manuscript received December 17, 2002
Accepted for publication April 14, 2003

ABSTRACT

The stability of the structure of bacterial genomes is challenged by recombination events. Since major rearrangements (*i.e.*, inversions) are thought to frequently operate by homologous recombination between inverted repeats, we analyzed the presence and distribution of such repeats in bacterial genomes and their relation to the conservation of chromosomal structure. First, we show that there is a strong underrepresentation of inverted repeats, relative to direct repeats, in most chromosomes, especially among the ones regarded as most stable. Second, we show that the avoidance of repeats is frequently associated with the stability of the genomes. Closely related genomes reported to differ in terms of stability are also found to differ in the number of inverted repeats. Third, when using replication strand bias as a proxy for genome stability, we find a significant negative correlation between this strand bias and the abundance of inverted repeats. Fourth, when measuring the recombining potential of inverted repeats and their eventual impact on different features of the chromosomal structure, we observe a tendency of repeats to be located in the chromosome in such a way that rearrangements produce a smaller strand switch and smaller asymmetries than expected by chance. Finally, we discuss the limitations of our analysis and the influence of factors such as the nature of repeats, *e.g.*, transposases, or the differences in the recombination machinery among bacteria. These results shed light on the challenges imposed on the genome structure by the presence of inverted repeats.

THE advances of the last decade on genome sequencing and pulsed field gel electrophoresis provide a puzzling image concerning the organization and stability of bacterial genomes. On one hand, many features of genome organization have been found or further unraveled, such as the impact of replication in imposing compositional strand biases (Lobry 1996) and constraining gene distribution (McLean *et al.* 1998). Coding sequences cover 90% of most bacterial genomes and transcriptional regulation can be very complex, suggesting selection for structural stability. On the other hand, the genome structure is extremely fluid. Operons are not well conserved between distant species (Itoh *et al.* 1999) and gene content varies at a very high rate in some bacterial lineages (Casjens 1998), partly because of frequent horizontal transfers (Ochman *et al.* 2000). Distinctive features of the organization of bacterial genomes, notably in relation to replication, have different importance in different species (Rocha and Danchin 2001). Among groups such as Firmicutes, one observes very different compositional bias, as well as different gene positional biases (Rocha 2002). Thus, the understanding of the trade-offs between genome stability and the requirements of genotypic diversity is becoming a major issue in the study of genome evolution.

Intrachromosomal homologous recombination can lead to deletions, duplications, translocations (for direct repeats), and inversions (for inverted repeats; Smith 1988; Roth *et al.* 1996; Romero and Palacios 1997). All these events change the genome composition, but most of them do not induce very important shifts in its structure. Indeed, large deletions are counterselected, large insertions are rare, and large tandem duplications are not observed in currently sequenced bacterial genomes, probably because they are too unstable. Therefore, inversions have been regarded as one of the main motors of chromosome structural change (Liu and Sanderson 1996; Roth *et al.* 1996; Hughes 2000). Pairwise comparisons among completely sequenced genomes show that the first large chromosome rearrangements are caused by inversions (Eisen *et al.* 2000; Tillier and Collins 2000a; Zivanovic *et al.* 2002). Currently, the sole exception to this rule is provided by the comparison of *Mycoplasma pneumoniae* and *M. genitalium*, which reveals several translocations and no inversion (Himmelreich *et al.* 1997). However, inverted repeats capable of mediating chromosomal inversions are strongly underrepresented in these genomes, being

[1] *Present address:* Wakeley Lab, BioLabs, Harvard University, 16 Divinity Ave., Cambridge, MA 02138.

[2] *Present address:* Equipe Hélix, INRIA Rhône-Alpes, Zirst, 655 Avenue de l'Europe, Montbonnot, 38334 Saint Ismier Cedex, France.

[3] *Corresponding author:* Atelier de BioInformatique, Université Pierre et Marie Curie, 12 Rue Cuvier, 75005 Paris, France.
E-mail: erocha@pasteur.fr

present up to 60 times less frequently than direct repeats (Rocha and Blanchard 2002).

It is difficult to define genome stability without experimental support or a large number of very close genomes. Thus, we use replication compositional bias as a proxy of genome stability. DNA replication is asymmetric; one strand is replicated continuously (leading strand) whereas the other is replicated in discrete steps through the use of Okasaki fragments (lagging strand; Marians 1992). Since the origins of replication in bacteria, when they are known, seem to be unique, the asymmetry in replication creates a durable asymmetry in the structure of the chromosome (Frank and Lobry 1999). This leads to different nucleotide compositions in each replicating strand, which seem to result from an essentially neutral mutational bias (Frank and Lobry 1999; Tillier and Collins 2000b; Rocha and Danchin 2001). Thus, the intensity of the bias is shaped by the strength of the mutational mechanism and by the rates of genome rearrangement. Assuming that the strength of the mechanism has small variations between genomes, strand bias should be highly correlated to the stability of the chromosome.

Chromosomal inversions seem to be rare in nature but very frequent in the laboratory (Louarn *et al.* 1985; Roth *et al.* 1996). This suggests the existence of selection pressure for maintaining chromosomal structure. As a result of this, the large inversions observed in bacterial genomes are symmetrical in relation to the origin of replication (Eisen *et al.* 2000; Tillier and Collins 2000a). It is then important to understand how chromosomes face recombination events and especially inversions. Here, we tackle this question by accounting for the distribution of repeats capable of producing rearrangements (inverted repeats). We also take into consideration the effects of such potential rearrangements on different elements of the chromosomal structure, in particular chromosome asymmetry and replication strand bias.

METHODS AND DATA

Data: Data on the complete bacterial genomes were taken from Entrez Genomes (http://www.ncbi.nlm.nih.gov), and the annotations were taken from the GenBank files. Except when noted otherwise, we used only one strain for each species to avoid any bias in favor of species represented several times in GenBank. This resulted in a data set of 63 chromosomes, representing 58 bacterial genomes.

Identification of large strict repeats: To compute the threshold minimal length of large repeats, we used a statistic of extremes that takes into account the nucleotide composition and the length of the genome (Karlin and Ost 1985). Among bacteria, the minimal length for which the probability of finding one exact repeat in the genome is <1‰ is in the range 21–26 nucleotides (nt) ($P < 0.001$; Rocha *et al.* 1999a). The search for such large, strictly identical repeats was done using *Reputer* (Kurtz and Schleiermacher 1999).

Deriving large nonstrict repeats: To investigate the influence of genome structure on repeats, we identified nonstrict repeats from strict repeats using an extension process previously described (Achaz *et al.* 2000, 2002). The method identifies nonstrict repeats by extending both sides of strict repeats when they share significant similarity in sequence. This is based on a local alignment procedure (Smith and Waterman 1981). Nucleotide frequencies differ widely between bacterial species and identity matrix scores produce artificially longer repeats in highly biased genomes. To avoid this effect, we used an empirical scoring matrix for each chromosome, which takes into account the frequencies of nucleotides (Achaz *et al.* 2002). After comparing different methods to build such a matrix, we used the one providing closer average lengths for repeats detected in random genomes with different nucleotide compositions (from very low to very high values of G + C content). This scoring matrix is the following:

$$\text{match}_{i/i} = -10 \times \ln(p_i^2), \quad \text{mismatch}_{i/j} = 10 \times \ln(p_i \times p_j)$$
$$\text{match}_{N/i} = -10 \times \frac{1}{4} \times \frac{1}{4} \times \sum_{i=A}^{T} \ln(p_i^2),$$
$$\text{Gap}_{\text{ext}} = -4 \times \text{match}_{N/i}, \quad \text{Gap}_{\text{open}} = 4 \times \text{Gap}_{\text{ext}},$$

where p_i is the frequency of the nucleotide i in the genome. This matrix provides scores of matches ranging from 20 to 41 and scores of mismatches ranging from −41 to −20. The score of $\text{match}_{N/i}$ is either 7 or 8, depending on the genome bias.

Strand compositional bias: Linear discriminant analyses followed by skew analyses were used to identify genomes with significant strand bias, as in Rocha and Danchin (2001). Once origin and terminus were identified, compositional strand bias was quantified in terms of ΔGC skews. These are defined as the average difference in GC skews between the genes in the leading and the lagging strand. $\Delta GC = (G_{\text{lead}} - C_{\text{lead}})/(G_{\text{lead}} + C_{\text{lead}}) - (G_{\text{lag}} - C_{\text{lag}})/(G_{\text{lag}} + C_{\text{lag}})$, where X_i is the nucleotide frequency of the nucleotide X (*i.e.*, G or C) in the genes of strand i (*i.e.*, lead or lag). This normalizes the replication biases in terms of the genome average bias in nucleotide composition.

MODELS OF GENOME REARRANGEMENT

Before proceeding, we must model the potential outcome of recombination between repeats. We consider a random model where each copy of a repeat can recombine with another copy of the repeat in a random way. We further suppose that couples of repeats of identical size recombine at identical frequency. Yet, two factors are taken into account. First, since one expects larger repeats to recombine more often than smaller ones (in a linear fashion according to Shen and Huang 1986;

Vulic *et al.* 1997), we weight each repeat by its length when computing indices of potential rearrangement. Second, we incorporate the fact that recombination always proceeds between two copies of a repeat. A repeat present in two copies recombines only in a single way between the two copies. However, a repeat with three copies (*e.g.*, A, B, and C) can recombine in three different ways (A with B, A with C, and B with C), each resulting in different outcomes. Thus, because recombination takes place between pairs of repeats we count the latter repeat as three couples of repeats. Therefore, counting pairs of potentially recombining repeats is equivalent to counting couples of repeats.

Assumptions: To accomplish this analysis we had to proceed to several assumptions:

1. We implicitly assume that homologous recombination may proceed with all statistically significant large repeats. Such minimal length varies from 21 to 26 nt depending on genome size and composition. In fact, it corresponds to the minimal length requirements for the start of homologous recombination by the RecBCD system in *Escherichia coli* and its functional homolog AddAB in *Bacillus subtilis*, the only bacterial species for which such studies have been conducted (Shen and Huang 1986; Roberts and Cohan 1993).
2. We analyze the distribution of repeats as they occur in the published genomes. Thus, we do not take into account the changes of that distribution if rearrangements do occur. Naturally, more refined models should be developed in the future to tackle this question. Such models should take into account the results of rearrangements on the relative positioning of the other repeats (even though that is a very hard computational problem), and the rate of repeat creation and loss.
3. Given the lack of experimental comparative studies of recombination mechanisms and frequencies in most bacteria, we implicitly assume that the frequency of intrachromosomal recombination is the same in different genomes. All bacteria here analyzed, except Buchnera (Shigenobu *et al.* 2000), have RecA, the major protein in homologous recombination pathways. However, the different elements of the homologous recombination pathways vary significantly between genomes (Eisen and Hanawalt 1999).
4. We consider that all repeats are involved in the dynamics of the chromosome in the same way. Since self-replicating repeated elements, such as IS, have special dynamics, we analyze their influence separately. We discuss the impact of violations to these assumptions in the interpretation of the results.

Measures of global rearrangement: The inversion produced by a recombination event between two occurrences of a repeat implicates the inversion of the region between the repeats—the *spacer*. This element contains less than half of the chromosome, by definition. A simple way of analyzing the potential for genome rearrangement is simply to divide the total number of pairs of inverted repeats by the length of the genome, thereby computing a *density of pairs of repeats.* However, the analysis of direct repeats has shown that the average spacer length is different between genomes (Rocha *et al.* 1999a). Also, the frequency of recombination between copies of a repeat is expected to be proportional to the repeat's length. Therefore, a more precise measure for the average rearrangement length potentially induced by the inverted repeats in a genome is given by

$$R_{\mathrm{L}} = \frac{1}{G_{\mathrm{L}}} \times \frac{\sum \mathrm{Lr}_i \times \mathrm{Lsp}_i}{\mathrm{Lr}_{\mathrm{T}}},$$

where R_{L} is the potential rearrangement length associated with the repeats in the genome; Lr_i, the length of the repeat i; Lsp_i, its spacer length; G_{L}, the genome length; and Lr_{T}, the sum of the repeats' lengths.

Inversions and replication structure: Compositional strand bias and chromosomal symmetry are differently affected by recombination between inverted repeats (Figure 1). By definition, copies of inverted repeats occur in different DNA strands. However, they can be in the same type of replicating strand (*i.e.*, both copies in the same chirochore—either leading or lagging strand) or in the same replichore (same replicating half of the chromosome). If they are in the same replichore (I_{R}), then an inversion will produce a shift of the spacer from one replicating strand to the other, so that the sequence of the spacer that was on the leading strand switches to the lagging strand and vice versa. However, because in this case the spacer does not include the origin or the terminus of replication, the symmetry of the chromosome (*i.e.*, the opposite placement of origin and terminus of replication) will not be affected. Naturally, close occurrences will induce small changes, whereas distant occurrences induce large changes. One can then define a measure of average strand switch (SS) potentially induced by all I_{R} repeats in a genome as

$$\mathrm{SS} = \frac{1}{G_{\mathrm{L}}} \times \frac{\sum \mathrm{Lr}_i \times \mathrm{Lsp}_i}{\mathrm{Lr}_{\mathrm{T}}} = R_{\mathrm{L}} \quad (\text{for } I_{\mathrm{r}} \text{ repeats only}).$$

Conversely, the spacer of a repeat with occurrences in the same chirochore (I_{C}) encompasses the origin or the terminus of replication. In this case, an inversion will not change the leading/lagging character of the spacer, but may induce changes in the relative positions of the origin and terminus of replication. The average asymmetry switch (AS) induced by the inversion will be proportional to the distance of the position of the center of the spacer (P_i) to the closer origin/terminus of replication ($P_{\mathrm{ori/ter}}$):

$$\mathrm{AS} = \frac{1}{G_{\mathrm{L}}} \times \frac{\sum |P_i - P_{\mathrm{ori/ter}}| \times 2 \times \mathrm{Lr}_i}{\mathrm{Lr}_{\mathrm{T}}} \quad (\text{for } I_{\mathrm{c}} \text{ repeats only}).$$

Expected values: We determined the expected values

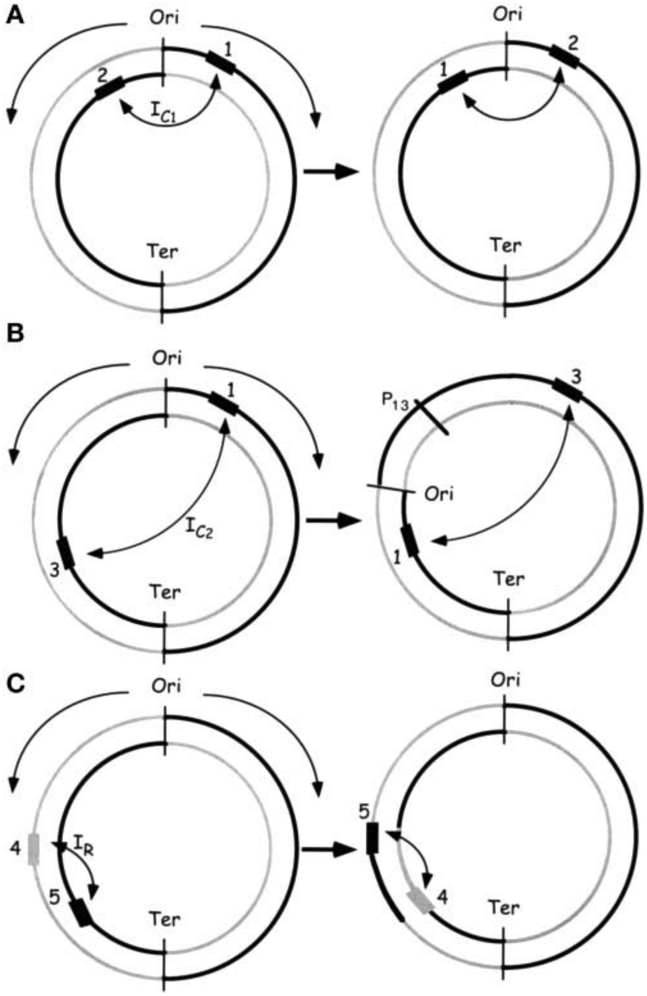

FIGURE 1.—Examples of intrachromosomal recombination between inverted repeats. Solid and shaded semicircles represent different replicating strands. (A) Repeats in the same chirochore (I_C) and symmetrical around the origin of replication. The inversion through recombination between 1 and 2 results in $R_L = \text{Lsp}_{12}/G_L$, AS ~ 0, SS $= 0$. (B) Repeats in the same chirochore (I_C) and asymmetrical around the origin of replication. The inversion through recombination between 1 and 3 results in $R_L = \text{Lsp}_{13}/G_L$, AS $= 2 \times |P_{13}, P_{\text{Ori}}|/G_L$, SS $=$ 0. (C) Repeats in the same replichore (I_R). The inversion through recombination between 4 and 5 results in $R_L = \text{Lsp}_{45}/G_L$, AS $= 0$, SS $= R_L$. Lsp_{ij} is the shortest distance between i and j in the circle (*i.e.*, the length of the spacer), G_L is the length of the genome, P_{ij} is the geometric center of the spacer between the copies i and j, P_{Ori} is the position of the origin of replication.

of R_L, SS, and AS under a model where pairs of copies of repeats engage into recombination randomly. The null model corresponds to a random placement of repeats in the chromosomes. Thus, approximate values for the expectations of R_L, SS, and AS can be easily determined by simulation. Here, we detail the derivation of the exact expressions. Under the model of random placement of repeats in the chromosome, the distance between two copies of a repeat is distributed uniformly in the interval $]0, G_L/2]$. Therefore, the expected value of R_L is ¼ $(1/G_L \times G_L/4)$.

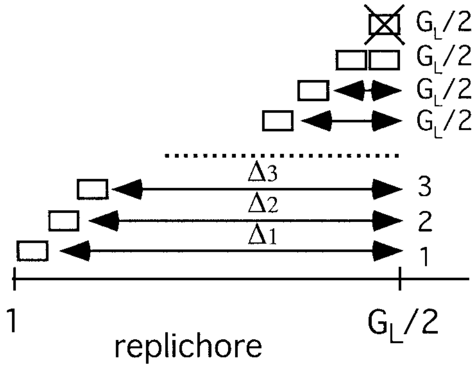

FIGURE 2.—Diagram depicting the rationale of the method to determine the probability density function associated with SS. Let us suppose the above replichore. A repeat will be counted only if both copies occur in the same replichore. Let us fixate the first copy of the repeat. If this copy is at position 1 (repeat 1), then the other copy has a uniform probability of occurrence in the replichore except at position 1; *i.e.*, it has a uniform probability density in Δ_1 (which has a length of $G_L/2 - 1$). Now suppose the repeat whose first occurrence is at position 2. The second occurrence of this repeat has a uniform probability of occurrence in Δ_2 (which has a length of $G_L/2 - 2$). A similar reasoning applies to all repeats whose first copy is at positions 3 to $G_L/2 - 1$. For these repeats, the second copy is distributed with a uniform probability of occurrence in Δ_3 to $\Delta_{G_L/2-1}$. A copy at $G_L/2$ cannot have a second occurrence in the replichore. Thus, the distribution of the length of the spacer (Δ) is a linear function of the length of the replichore (*i.e.*, it is a linear function of $G_L/2$).

For the determination of the expected values of SS and AS we assume, as previously, uniform distribution for the distance between copies. For simplicity, but this does not affect generality, we assume that all repeats have the same length. Under these conditions, we call SSi the strand switch associated with a repeat and allow it to take one of two values: either the length of the spacer (both copies in the same replichore) or 0 (both copies in the same chirochore). Given the symmetry of the system, the value SSi $= 0$ has a probability 0.5. Thus, one has to determine only the expression for the probability density function of SSi when repeats are in the same replichore (which sums to 0.5). This results in a function that depends linearly on the spacer length (see Figure 2) and is constrained by two conditions: (i) the cumulated probability is 0.5 and (ii) the function evaluates to zero at $G_L/2$. Thus, the probability density function is given by

$$f(x) = \alpha + \beta x,$$

$$\text{constrained by (i)} \int_0^{G_L/2} f(x)\,dx = 0.5 \text{ and (ii) } f(G_L/2) = 0,$$

which results in a function

$$f(x) = \frac{2}{G_L} - \frac{4}{G_L^2}\,x,$$

whose expected value is given by

$$E(x) = \int xf(x)\,dx = G_L/12.$$

Since SS is the sum of each partial SSi, divided by the genome length, its expected value is $\frac{1}{12}$. Excluding from the analysis the repeats in the same chirochore, for which SS = 0, the expected value becomes $\frac{1}{6}$. A similar reasoning applies to the determination of the expected value of AS.

RESULTS AND DISCUSSION

Relative distribution of inverted repeats in bacterial genomes: *Absolute numbers of repeats:* The distribution of direct and inverted repeats in bacterial genomes has recently been analyzed in the context of horizontal transfer (Rocha *et al.* 1999a) and of repeat generation (Achaz *et al.* 2002). These works have shown that bacterial genomes contain a considerable amount of large repeats. Furthermore, the abundance of such repeats is highly variable among species. In our data set, one finds a maximum of 66,860 pairs of inverted repeats in *Neisseria meningitidis* and no inverted repeats in *Chlamydia trachomatis* (Table 1). Interestingly, both bacteria are human pathogens and seem to have a functional RecBCD system. However, *C. trachomatis* is an obligatory and intracellular parasite, whereas Neisseria is neither. A low level of repeats is typical of obligatory intracellular bacteria (see below). On the other hand, Neisseria is a very extreme case of repeat abundance, mostly for effects of antigenic variation (Saunders *et al.* 2000). For clarity it is removed from most graphs where the number of repeats is taken explicitly into account. The average length of the repeats in the different genomes is nearly always well above the lower threshold of statistical significance. Indeed, the average length of strict repeats in the genomes is 207 nucleotides.

Inverted repeats are underrepresented compared to direct ones: One would expect to find more direct than inverted repeats if selection acts toward minimizing inversions. On one hand, inverted repeats may induce inversions. On the other hand, if repeats originate mainly from close direct repeats (Achaz *et al.* 2002), inversions are required to create inverted repeats from direct repeats. In any case, our analysis indicates that inverted repeats are usually underrepresented compared to direct repeats: the ratio of inverted/direct repeats is almost always <1 (Figure 3). This still holds if one excludes close direct repeats, which are thought to be the result of an active process of duplication (Achaz *et al.* 2002). The strongest underrepresentation of inverted repeats tends to occur when the total number of repeats is smaller. This suggests that when repeats are avoided (*e.g.*, by structural reasons), inverted repeats are even more strongly avoided, possibly because of their major role in chromosomal inversions. Genomes saturated with repeats, *e.g.*, Neisseria, show no difference between inverted and direct repeats, possibly because selection for a direct positioning *vs.* the inverted becomes inefficient at such a high level of repeat density (or possibly because their recombination apparatus is less sensitive to repeats).

Rearrangement length: Although relative avoidance of inverted repeats may suggest counterselection of sequences capable of producing inversions, many different causes can underlie such avoidance. In particular, if the magnitude of the rearrangements' counterselection were simply proportional to their length, one would expect a selection for close repeats that could induce small rearrangements. However, the average observed/expected (O/E) R_L is 0.963, which is not significantly different from 1 ($P > 0.4$, signed-rank test; Figure 4). One is then inclined to think that although selective pressure against rearrangements may cause the avoidance of inverted repeats, relative to direct ones, there is no systematic tendency toward the minimization of the length of the potential rearrangement.

Support for the hypothesis that inverted repeats challenge the chromosomal stability: *Analyses of close genomes:* The genomes presenting the lowest values of observed/expected rearrangement length are the ones containing fewer repeats, notably Chlamydia, some Mycoplasma, Rickettsia, and Buchnera. These are also the genomes with smaller inverted/direct ratios. Interestingly, recent works have shown that many obligatory intracellular bacterial genomes keep a remarkable synteny (Suyama and Bork 2001; Wolf *et al.* 2001). In light of their small populations one would expect less efficient purifying selection and therefore larger differences in gene order. The observations that such genomes contain a reduced recombination potential, especially when it involves inversions, may thus explain their stability. Recently, a second genome of Buchnera has been published (Tamas *et al.* 2002), which indicates that for 50 million years these genomes remained strictly colinear, showing no inversion. This is not surprising since these genomes have both <10 large inverted repeats in their genomes and a deficient homologous recombination machinery. However, the other small stable genomes presenting few repeats do code for both RecA and RecBCD or RecF-like systems.

Closely related bacteria with very different repeat abundance show increased levels of synteny loss. For example, the strains KIM and CO92 of *Yersinia pestis* are very closely related (average 99.9% of protein similarity) but show a considerable amount of rearrangement in their genomes (Deng *et al.* 2002). The closely related *Salmonella enterica typhi* and *typhimurium* (mean protein similarity of 98.6%) show only two large rearrangements. This can be put into relation with their different numbers of repeats: ∼5000 in *Y. pestis*, many of them insertion sequences, and <1000 in *S. enterica typhimurium* (for genomes of similar lengths). The correlation between abundance of repeats and genome stability seems to be valid also in Archaea. A recent comparative study of three Pyrococcus (*Pyrococcus abyssi*, *P. horikoshii*,

TABLE 1

General results

Chromosome	Accession no.	Length (kb)	ΔGC skew	N_{I}	N_{D}	$N_{\mathrm{I}}/N_{\mathrm{D}}$	$O/E\ R_{\mathrm{L}}$	O/E SS	O/E AS
Aeropyrum pernix	NC_000854	1670	—	572	1437	0.40	1.34	—	—
Agrobacterium tumefaciens C58 chr1	AE007869	2842	0.050	904	880	1.03	0.83	0.56	0.72
A. tumefaciens C58 chr2	AE007870	2075	0.039	241	295	0.82	1.00	0.42	0.76
A. tumefaciens C58 pl AT	AE007872	543	—	6	43	0.14	1.32	—	—
Aquifex aeolicus	AE000657	1551	—	204	253	0.81	1.23	—	—
Archaeoglobus fulgidus	AE000782	2178	—	3092	3970	0.78	0.73	—	—
Bacillus halodurans C-125	BA000004	4202	0.095	2245	4551	0.49	1.06	0.90	0.91
B. subtilis 168	AL009126	4215	0.084	407	755	0.54	1.14	0.68	1.04
Borrelia burgdorferii B31	AE000783	911	0.284	1	52	0.02	1.74	—	—
Brucella melitensis 16M chr 1	AE008917	2117	0.067	786	1009	0.78	1.00	0.94	0.97
B. melitensis 16M chr 2	AE008918	1178	0.067	217	221	0.98	0.99	0.75	0.86
Buchnera APS	AP000398	641	0.046	1	1	1.00	1.13	—	—
Caulobacter crescentus CB15	AE005673	4017	0.035	2175	2213	0.98	0.97	0.84	0.84
Campylobacter jejuni NCTC11168	AL111168	1641	0.142	11	123	0.09	0.75	0.70	0.52
Chlamydia muridarum MoPn	AE002160	1069	0.269	16	47	0.34	0.14	0.21	0.05
C. pneumoniae CWL029	AE001363	1230	0.183	3	133	0.02	0.01	—	—
C. trachomatis D	AE001273	1046	0.251	0	12	0.00	—	—	—
Clostridium acetobutylicum ATCC824	AE001437	3941	0.228	208	966	0.22	0.96	0.36	1.24
C. perfringens 13	BA000016	3031	0.205	732	1011	0.72	0.74	0.92	0.64
Deinococcus radiodurans R1 chr 1	AE000513	2649	—	1637	1677	0.98	0.99	—	—
D. radiodurans R1 chr 2	AE001825	412	—	14	51	0.27	1.09	—	—
Escherichia coli MG1655	U00096	4639	0.051	3688	4394	0.84	0.98	0.93	0.83
Haemophilus influenzae Rd	L42023	1830	0.064	381	646	0.59	1.69	0.73	1.95
Halobacterium sp. NRC-1	AE004437	2014	0.080	111	237	0.47	0.41	0.86	0.91
Helicobacter pylori 26695	AE000511	1668	0.064	143	595	0.24	1.22	0.96	1.22
Lactococcus lactis IL1403	AE005176	2366	0.111	728	725	1.00	0.89	1.07	0.83
Listeria innocua Clip11262	AL592022	3011	0.091	246	383	0.64	0.87	0.33	0.84
L. monocytogenes EGD	NC_003210	2945	0.098	94	335	0.28	1.02	0.61	1.05
Methanocaldococcus jannaschii	L77117	1665	—	1372	4216	0.33	0.77	0.80	0.68
Mesorhizobium loti MAFF303099	NC_002678	7036	—	1116	1891	0.59	0.87	—	—
Methanobacterium thermoautotrophicum ΔH	AE000666	1751	0.023	5917	9000	0.66	1.11	1.20	0.57
Mycoplasma genitalium G37	L43967	580	—	22	699	0.03	1.40	—	—
Mycobacterium leprae TN	AL450380	3268	0.110	1134	1104	1.03	0.91	1.02	0.97
Mycoplasma pneumoniae M129	NC_000912	816	—	303	2539	0.12	1.21	0.65	0.94
M. pulmonis UABCTIP	AL445566	964	—	114	1592	0.07	1.69	0.48	1.77
Mycobacterium tuberculosis H37Rv	AL123456	4412	0.049	1198	2888	0.41	1.04	0.80	0.99
Neisseria meningitides Z2491	AL162759	2184	0.115	66860	68354	0.98	1.02	0.94	0.98
Nostoc sp. PCC 7120	BA000019	6414	—	5868	8033	0.73	0.97	—	—
Pasteurella multocida PM70	AE004439	2257	—	101	1436	0.07	1.10	—	—
Pseudomonas aeruginosa PA01	AE004091	6264	0.067	488	1081	0.45	1.30	1.11	1.20
Pyrococcus abyssi	AL096836	1765	0.021	708	705	1.00	0.35	2.36	0.27
Pyrobaculum aerophilum IM2	AE009441	2222	—	923	4244	0.22	0.50	—	—
Pyrococcus horikoshii OT3	NC_000961	1739	0.060	499	3395	0.15	1.80	1.99	0.24
Ralstonia solanacearum GMI1000	AL646052	3716	0.051	345	657	0.53	1.17	0.81	0.96
Rickettsia conorii Malish 7	AE006914	1269	0.084	1178	1348	0.87	0.98	0.91	0.98
R. prowazekii Madrid E	AJ235269	1112	0.096	4	19	0.21	0.22	0.22	0.51
Salmonella typhimurium LT2	AE006468	4857	0.068	812	2120	0.38	1.28	0.96	0.79
Sinorhizobium meliloti 1021	AL591688	3654	0.035	5861	5963	0.98	0.96	0.98	0.96
Staphylococcus aureus N315	BA000018	2815	0.130	911	1419	0.64	1.01	0.92	0.92
Streptococcus pneumoniae TIGR4	AE005672	2161	0.174	8152	10954	0.74	1.04	0.89	0.88
S. pyogenes M1	AE004092	1852	0.135	237	241	0.98	1.16	0.75	0.79
Sulfolobus solfataricus P2	AE006641	2992	—	16671	21531	0.77	0.24	—	—
S. tokodaii strain7	BA000023	2695	—	4260	8503	0.50	0.01	—	—
Synechocystis PCC6803	AB001339	3573	—	1624	1799	0.90	1.02	—	—
Thermoplasma acidophilum	AL139299	1565	—	40	1083	0.04	1.40	—	—
Thermotoga maritima MSB8	AE000512	1861	0.036	1530	1732	0.88	0.87	0.68	0.47
Thermoplasma volcanium GSS1	NC_002689	1585	0.011	166	742	0.22	1.38	—	—

(*continued*)

TABLE 1

(Continued)

Chromosome	Accession no.	Length (kb)	ΔGC skew	N_I	N_D	N_I/N_D	$O/E\ R_L$	O/E SS	O/E AS
Treponema pallidum pallidum	AE000520	1138	0.176	34	170	0.20	0.68	0.54	0.45
Ureaplasma urealyticum S3	AF222894	752	—	11	116	0.09	0.17	—	—
Vibrio cholerae N16961 chr 1	AE003852	2961	0.094	1022	1117	0.91	1.07	0.96	0.85
V. cholerae N16961 chr 2	AE003853	1072	0.106	167	21545	0.01	0.95	0.93	0.92
Xylella fastidiosa 9a5c	AE003849	2679	0.207	576	1952	0.30	1.40	0.59	0.91
Yersinia pestis CO92	AL590842	4654	0.058	4915	5389	0.91	0.99	0.96	0.90

The GenBank accession numbers of the chromosomes, their length, their ΔGC skews, the number of repeats in inverse (N_I) and direct (N_D) sense, and their ratio (N_I/N_D), and the observed/expected (O/E) ratios of R_L, SS, and AS are shown. The chromosomes for which there are no values of ΔGC skews, SS and AS, are the ones for which the origin and terminus of replication could not be determined or for which there is no significant compositional strand bias (for ΔGC skews).

and *P. furiosus*) has indicated that *P. furiosus* is much more subject to genome rearrangements (Zivanovic *et al.* 2002). The close comparison between these genomes seemed to implicate repeats in these rearrangements. Indeed, the comparison of the number of pairs of inverted repeats in these genomes (of nearly identical genome length) is in good agreement with these observations: 503 for *P. horikoshii*, 711 for *P. abyssi*, and 2004 for *P. furiosus*.

The case of Rickettsia conorii: One major exception to this trend concerns the comparison of *Rickettsia conorii* with *R. prowazekii*. *R. conorii* is 14% larger than *R. prowazekii*, but the genomes are colinear, thus supposedly stable, even though *R. conorii* contains 1180 inverted repeats that have been proposed to replicate in a selfish manner (Ogata *et al.* 2000) for only 6 inverted repeats in *R. prowazekii*. A closer analysis of the former genome indicates that its repeats are all small, since 70% of repeats have between 25 and 30 bp, and only 2 repeats are >85 bp. Since the genome does not contain a homolog of the RecBCD system, homologous recombination is expected to follow the RecF pathway (Shen and Huang 1986). The RecF pathway is thought to be involved in restarting replication after replication fork disassociation, and the importance of its role in homologous recombination has been disputed (Courcelle *et al.* 2001; Amundsen and Smith 2003). For this pathway, the minimal length of strict homology required to start homologous recombination in *E. coli* is much larger than that required for the RecBCD pathway. It might be as large as 90 bp, whereas it is 20–30 bp for the RecBCD pathway (Shen and Huang 1986). It is thus quite possible that these repeats are not targeted by homologous recombination in Rickettsia, because of the peculiarities of its recombination machinery. This would explain the stability of these genomes in spite of the large number of small repeats.

Support to use replication composition bias as a proxy of genome stability: We have previously suggested a link between the number of repeats in a genome and the replication compositional strand bias (Rocha *et al.* 1999b). Compositional replication strand bias seems to result from a fast asymmetric mutational bias causing inverted genes to adapt fast to the new strand (Tillier and Collins 2000b). The magnitude of strand bias can vary either by the intensity of the mutational bias or by processes counteracting its establishment, such as genome

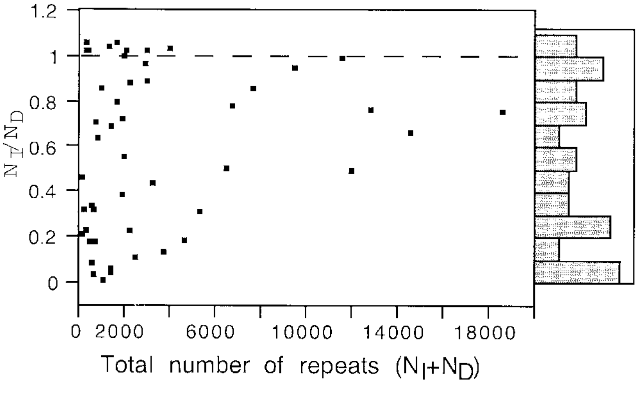

Figure 3.—Distribution of the ratio of inverted (N_I) over direct (N_D) repeats in the function of the total number of repeats and corresponding histogram.

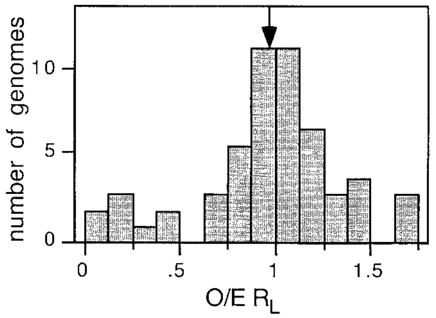

Figure 4.—Observed/expected median rearrangement length of repeats in the bacterial genomes. The arrow indicates the median (0.96), which is not significantly different from 1 ($P > 0.4$, signed-rank test).

rearrangements. In this sense, important levels of strand bias can be established only if the genomes are stable. Among the 63 chromosomes, 44 exhibit a significant replication bias. Genomes with significant strand bias have a median of 394 repeats/genome, whereas the remaining genomes have a median of 708 repeats/genome (even though the former genomes are 23% larger). Moreover, genomes with significant strand bias show a negative correlation ($\rho = -0.30$, $P < 0.05$, Spearman's rank test) between the number of inverted repeats and the intensity of the bias (measured as ΔGC skew). These results suggest that the chromosomal stability is highly challenged by inverted repeats. As a consequence, in very stable chromosomes, the number of inverted repeats might tend to be minimized.

How do inverted repeats challenge the chromosomal stability? To tackle this question, we divided inverted repeats into two categories: repeats in the same chirochore (further labeled as I_C) and repeats in the same replichore (I_R; see MODELS OF GENOME REARRANGEMENT and Figure 1). We also developed simple measures of the impact of these repeats on AS and SS. AS measures the consequences of potential rearrangements between I_C. SS measures the consequences of potential rearrangements between I_R. Therefore the ratio of observed/expected of these indices indicates the association between the positioning of repeats and the instabilities they might induce on genomes.

Differences between I_C and I_R suggest selection for chromosomal stability: Repeats are causes of change in chromosomal structure, but the distribution and maintenance of repeats is also constrained by the characteristics of that structure. In genomes containing strong compositional strand biases, the mutation pattern is similar for both copies of I_C, but different for both copies of I_R (Rocha and Danchin 2001). Thus, faster divergence between copies of I_R repeats, relative to I_C, could lead to differences in number and length between I_C and I_R (Table 2). If this is so, we should expect higher similarity between copies of I_C than between copies of I_R. Naturally this hypothesis cannot be tested with the data on strict repeats, which are identical (by definition). Therefore, we extended by dynamic programming the exact repeats into larger nonstrict repeats by searching for significant similarity at the edge of the strict repeats (as described in METHODS AND DATA). The comparison of nonstrict repeats confirms that I_C are more numerous and longer than I_R (Table 2). However, the average identity percentage does not differ between I_C and I_R repeats. This suggests that the different abundance of each type of repeats is not due to larger rates of divergence among I_R repeats. The avoidance of I_R could then be a consequence of negative selection on the distribution of repeats. Such selection pressure may have different origins. First, inversions change the relative distance of the genes to the origin of replication. This is expected to be counterselected in genomes selecting for highly expressed genes near the origin of replication. Second, genes on the leading strand will be transferred to the lagging strand and vice versa. This is also expected to be counterselected for highly expressed genes and for genomes containing two dedicated DNA polymerases (Rocha 2002). Finally, it has been proposed that higher levels of substitutions in inverted genes may lead to gene loss (Mackiewicz *et al.* 2001).

TABLE 2

Comparison of I_C and I_R within strict repeats and nonstrict repeats

Repeats	$I_C > I_R$	$I_C < I_R$	P value (χ^2)	Total
Strict				
N_I	29	6	$P < 0.001$	35
Length	27	8	$P < 0.01$	35
Nonstrict				
N_I[a]	25	8	$P < 0.01$	34
Length	27	7	$P < 0.001$	34
Identity %	18	16	NS	34

Here we compared the number of inverted repeats (N_I), the average length, and, for nonstrict repeats, the average percentage of identity. We analyzed genomes for which the origin and terminus of replication could be well defined and that had at least 10 repeats I_C and 10 repeats I_R. There are 35 such genomes by considering strict repeats and 34 by considering nonstrict repeats (several strict repeats can be part of the same larger nonstrict repeat). NS, not significant.

[a] For one genome (*Helicobacter pylori* 26695) $I_C = I_R$.

Chromosomes tend to keep their symmetry: Using the positions of the origins and termini of replication, one can determine the relative lengths of the two replichores. We analyzed the 48 genomes for which the origin and the terminus can be reliably predicted. In these genomes the length of the two replichores never differed by >20%. Further, the ratio of the lengths of the smallest over the largest replichores of each genome shows a median of 0.95 (data not shown). Such similarity between replichore lengths is in good agreement with the existence of a selective pressure against inversions increasing the asymmetry of the chromosome. A similar selection pressure has been observed in horizontal transfer between strains of *E. coli* and Salmonella, since genomic variation tends to occur in equal amounts on both replichores, thus keeping chromosomal symmetry (Bergthorsson and Ochman 1998). Further, inversions between the rRNA operons of *E. coli* that strongly change the symmetry of the chromosome have been found to be severely detrimental (Hill and Gray 1988). This is also in good agreement with data indicating preference for symmetrical rearrangements around the origin and terminus of replication (Eisen *et al.* 2000; Tillier and Collins 2000a). It has been proposed that such inversions could result from illegitimate recombination between the two newly replicated chromosomes

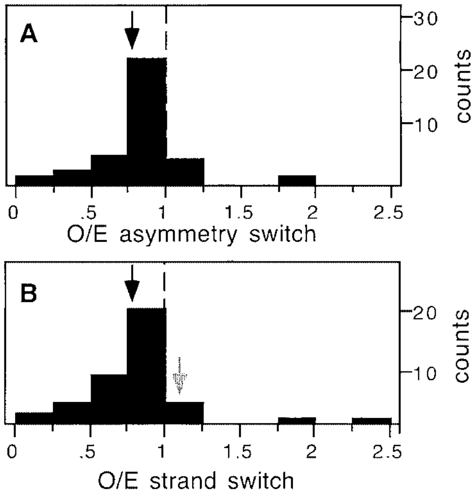

Figure 5.—(A) Observed/expected values for the asymmetry switch. The solid arrow indicates the median value (0.86). (B) Observed/expected values for the strand switch. The solid arrow indicates the median of the genomes with significant compositional strand bias (0.80), whereas the shaded arrow indicates the median for the other genomes (1.08).

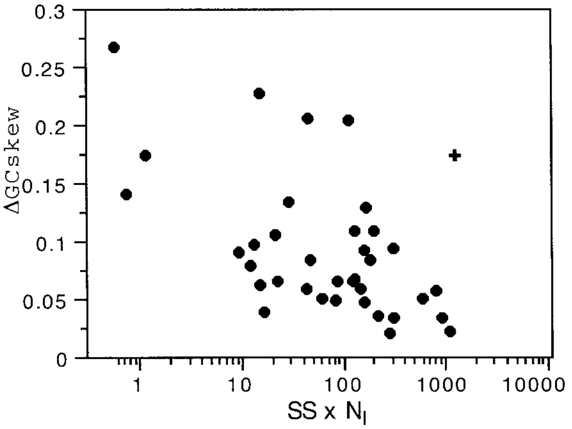

Figure 6.—Strand switch (SS × N_I) *vs.* the strand composition bias (Δ*GC* skew). Spearman's rank correlation is −0.553 ($P < 0.001$). The cross corresponds to the genome of *S. pneumoniae* (excluded as an outlier, along with *N. meningitidis*; see text).

at the moment of replication (Tillier and Collins 2000a), but there is still no experimental evidence of such a mechanism. The analyses of AS indicate *O/E* ratios systematically smaller than one (average AS = 0.86, $P < 0.001$, signed-rank test; Figure 5). This indicates that potential rearrangements caused by homologous recombination between I_C tend to be symmetrical and that such I_C repeats may be less negatively selected.

Strand switch and replication compositional bias: An inversion between two I_R switches the strands of the spacer and thus switches the compositional biases in each strand. The comparison of genomes with and without significant compositional strand biases shows a different median observed/expected SS (respectively, 0.80 and 1.08, $P < 0.01$, Wilcoxon test). Genomes lacking strand compositional bias have a median observed/expected SS not significantly different from 1 (median 1.08, not significant), whereas the others show a ratio systematically smaller than one (median 0.80, $P < 0.001$, signed-rank test). Further, among these genomes there is a significant negative correlation between the potential of repeats to induce strand switch and their genome Δ*GC* skew (−0.553, $P < 0.001$, Spearman ρ; Figure 6). Although the correlation is highly significant, the analysis of its residuals shows a considerable dispersion and two outliers, *Streptococcus pneumoniae* and *N. meningitidis* ($P < 0.01$). This is an indication that other factors affect strand bias and/or that some of our basic assumptions are oversimplified (*e.g.*, the assumption of similar recombination mechanisms and frequencies in different bacteria).

General picture: Both AS and SS indicate observed/expected ratios systematically smaller than 1 (Figure 5), and the differences between AS and SS are not statistically significant. One should note that avoiding simultaneously AS and SS can be done it two different ways. First, it can be done if the occurrences of repeats are close. However, the analysis of R_L for all inverted repeats and the relative abundance of I_R and I_C indicates that is not the case. Second, it can be done by selecting the placement of the two copies of repeats in the same chirochore and in a symmetrical way around the origin or the terminus of replication (see Figure 1). Our results point toward the latter hypothesis.

The special role of transposases: Among the simplifications we have made at the beginning of this work, we assumed that repeats induced rearrangements through homologous recombination. This is an oversimplification for some types of sequences and especially when transposases are concerned. We have thus tried to further analyze the impact of these elements in the induction of genome rearrangements. We have identified 40 bacterial genomes containing genes coding for putative transposases, using the annotation files. As expected, these genomes contain a much larger density of repeats (4.5 times larger, $P < 0.002$, Wilcoxon test). Further, the density of repeats correlates well with the number of transposases (ρ = +0.45, $P < 0.005$, Spearman rank test) with two clear outliers (*S. solfataricus* and *S. pneumoniae*). However, only 19% of the repeats directly concern sequences coding for transposases. Part of the difference may be explained by the difficulty in identifying unknown families of transposases or by the existence of insertion sequence (IS) remnants that no longer contain intact transposases. Only in three genomes (*Bacillus halodurans*, Synechocystis C125, and *Y. pestis*) do transposase-coding sequences include >55% of the genome's inverted repeats (respectively, 76%, 74%, and 72%).

Genomes lacking IS have smaller ratios of inverted/direct repeats (median 0.22) than genomes containing

IS (median 0.69, $P < 0.01$), although both values are significantly <1 ($P < 0.01$). There is also a positive and similar effect of transposases on the O/E values for AS and SS, which tend to get closer to 1, with the existence and with the number of transposases in the genome ($P < 0.01$). Thus, the presence of transposases in shuffling the genome seems to exceed the one of simple repeats targeted by homologous recombination. It is likely that their self-replicative behavior further shuffles the chromosome.

CONCLUSION

The availability of complete genomes of close species, or strains within a species, has brought to light the importance of genome rearrangements in fashioning the bacterial genome (Hughes 1999). Almost without exception, the first major rearrangements observed in recently divergent bacterial strains or species concern inversions that are symmetrical around the origin and terminus of replication. Here, we have tried to understand the relation between such analyses and the potential for intrachromosomal rearrangements mediated by the long repeats present in bacterial chromosomes. Selective processes are probably at the basis of the different abundance and characteristics of inverted repeats. These repeats have important consequences for genome stability, as we have seen, but they can also be under positive selection for antigenic variation or gene dosage effects. This seems to be a particular case of the trade-off between the necessity of generating genotypic diversity and the problems that are derived from that need.

To be able to compare different genomes we were forced to make several simplifying assumptions. Some, *e.g.*, the role of transposases, could be tackled in this work, but most will have to be tested as more experimental works on homologous recombination in other bacteria become available. In particular, it is of outmost importance to determine the relative levels of homologous recombination between repeats in different genomes as well as the minimal lengths required for homologous recombination. The results of this work suggest that these requirements are likely to be different, since some genomes, such as Neisseria, contain an astonishingly high level of repeats. The genome of *S. pneumoniae* shows particularly striking features, since it contains very high numbers of repeats for its size and large numbers of transposases (46 genes), but exhibits strong ΔGC skews and 80% of the genes in the leading strand. Such a well-ordered genome structure contrasts with the quantity of elements capable of disrupting it. It remains an open question if this is due to differences in the recombination machinery or to other processes.

Most of the results we have presented are compatible with the hypothesis that repeats challenge the structure of bacterial chromosomes. We found low values of AS and SS, a frequent association of repeat density with differential stability of close genomes, and a systematic underrepresentation of inverted repeats relative to direct ones. However, one would have also expected to find O/E R_L values significantly <1, which was not the case. However, considering only I_R, O/E R_L are <1, resulting in O/E SS < 1 (the underrepresentation of I_R as compared to I_C leads to that apparent randomness). On the other hand, the lack of a global bias in R_L shows that mechanisms creating repeats at short distances are not biasing our results. O/E R_L values close to 1 could result if the other elements contributing to the selection of a stable chromosomal structure are not sensitive to the length of the rearrangement. For example, selection of operon structures should be equally effective on small and on large rearrangements, since in both cases only the two operons at the breakpoints of rearrangements are disrupted (and this if repeats are inside different operons). Considering that many large repeats in bacteria are inside coding sequences (Rocha *et al.* 1999a), selection for minimization of operon disruption would be effective only through the avoidance of inverted repeats relative to direct ones (as observed). Thus, the distribution of repeats in genomes would be constrained by the structure of the chromosome in terms of replication, which is dependent on the length and the type of rearrangement, and of some other factors, which are possibly independent of the length of the inverted segments (*i.e.*, the rearrangement length). The relation between the distribution of repeats in bacterial chromosomes and other genomic features is still a largely unexplored field. For example, several works have suggested that nonpermissive intervals of rearrangement exist in *E. coli* (Segall *et al.* 1988; Guijo *et al.* 2001) and that some regions of the chromosome are particularly prone to recombination events (Louarn *et al.* 1991). Further work will be required to tackle these questions.

We are very grateful to Isabelle Gonçalves for carefully reading the manuscript. Guillaume Achaz was funded by "La Société de Secours des Amis des Sciences." Eric Coissac and Pierre Netter are at the Université Pierre et Marie Curie and Eduardo Rocha at the Centre National de la Recherche Scientifique. This work was partially funded by the Association pour la Recherche sur le Cancer, contract 4672.

LITERATURE CITED

Achaz, G., E. Coissac, A. Viari and P. Netter, 2000 Analysis of intrachromosomal duplications in yeast *Saccharomyces cerevisiae*: a possible model for their origin. Mol. Biol. Evol. **17:** 1268–1275.

Achaz, G., E. P. C. Rocha, P. Netter and E. Coissac, 2002 Origin and fate of repeats in bacteria. Nucleic Acids Res. **30:** 2987–2994.

Amundsen, S. K., and G. R. Smith, 2003 Interchangeable parts of the *Escherichia coli* recombination machinery. Cell **112:** 741–744.

Bergthorsson, U., and H. Ochman, 1998 Distribution of chromosome length variation in natural isolates of *Escherichia coli.* Mol. Biol. Evol. **15:** 6–16.

Casjens, S., 1998 The diverse and dynamic structure of bacterial genomes. Annu. Rev. Genet. **32:** 339–377.

Courcelle, J., A. K. Ganesan and P. C. Hanawalt, 2001 Therefore, what are recombination proteins there for? Bioessays **23:** 463–470.

Deng, W., V. Burland, G. Plunkett, III, A. Boutin, G. F. Mayhew *et al.*, 2002 Genome sequence of *Yersinia pestis* KIM. J. Bacteriol. **184:** 4601–4611.

Eisen, J. A., and P. C. Hanawalt, 1999 A phylogenomic study of DNA repair genes, proteins, and processes. Mutat. Res. **435:** 171–213.

Eisen, J. A., J. F. Heidelberg, O. White and S. L. Salzberg, 2000 Evidence for symmetric chromosomal inversions around the replication origin in bacteria. Genome Biol. **1:** 11.1–11.9.

Frank, A. C., and J. R. Lobry, 1999 Asymmetric patterns: a review of possible underlying mutational or selective mechanisms. Gene **238:** 65–77.

Guijo, M. I., J. Patte, M. Del Mar Campos, J. M. Louarn and J. E. Rebollo, 2001 Localized remodeling of the *Escherichia coli* chromosome: the patchwork of segments refractory and tolerant to inversion near the replication terminus. Genetics **157:** 1413–1423.

Hill, C. W., and J. A. Gray, 1988 Effects of chromosomal inversion on cell fitness in *Escherichia coli* K-12. Genetics **119:** 771–778.

Himmelreich, R., H. Plagens, H. Hilbert, B. Reiner and R. Herrmann, 1997 Comparative analysis of the genomes of the bacteria *Mycoplasma pneumoniae* and *Mycoplasma genitalium.* Nucleic Acids Res. **25:** 701–712.

Hughes, D., 1999 Impact of homologous recombination on genome organization and stability, pp. 109–128 in *Organization of the Prokaryotic Genome*, edited by R. L. Charlebois. ASM Press, Washington, DC.

Hughes, D., 2000 Co-evolution of the *tuf* genes links gene conversion with the generation of chromosomal inversions. J. Mol. Biol. **297:** 355–364.

Itoh, T., K. Takemoto, H. Mori and T. Gojobori, 1999 Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. Mol. Biol. Evol. **16:** 332–346.

Karlin, S., and F. Ost, 1985 Maximal segmental match length among random sequences from a finite alphabet, pp. 225–243 in *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, edited by L. M. L. Cam and R. A. Olshen. Wadsworth, New York.

Kurtz, S., and C. Schleiermacher, 1999 REPuter: fast computation of maximal repeats in complete genomes. BioInformatics **15:** 426–427.

Liu, S. L., and K. E. Sanderson, 1996 Highly plastic chromosomal organization in *Salmonella typhi.* Proc. Natl. Acad. Sci. USA **93:** 10303–10308.

Lobry, J. R., 1996 Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol. **13:** 660–665.

Louarn, J. M., J. P. Bouche, F. Legendre, J. Louarn and J. Patte, 1985 Characterization and properties of very large inversions of the *E. coli* chromosome along the origin-to-terminus axis. Mol. Gen. Genet. **201:** 467–476.

Louarn, J.-M., J. Louarn, V. François and J. Patte, 1991 Analysis and possible role of hyperrecombination in the termination region of the *E. coli* chromosome. J. Bacteriol. **173:** 5097–5104.

Mackiewicz, P., D. Mackiewicz, A. Gierlik, M. Kowalczuk, A. Nowicka *et al.*, 2001 The differential killing of genes by inversions in prokaryotic genomes. J. Mol. Evol. **53:** 615–621.

Marians, K. J., 1992 Prokaryotic DNA replication. Annu. Rev. Biochem. **61:** 673–719.

McLean, M. J., K. H. Wolfe and K. M. Devine, 1998 Base composition skews, replication orientation and gene orientation in 12 prokaryote genomes. J. Mol. Evol. **47:** 691–696.

Ochman, H., J. G. Lawrence and E. A. Groisman, 2000 Lateral gene transfer and the nature of bacterial innovation. Nature **405:** 299–304.

Ogata, H., S. Audic, V. Barbe, F. Artiguenave, P. E. Fournier *et al.*, 2000 Selfish DNA in protein-coding genes of *Rickettsia.* Science **290:** 347–350.

Roberts, M. S., and F. M. Cohan, 1993 The effect of DNA sequence divergence on sexual isolation in Bacillus. Genetics **134:** 401–408.

Rocha, E. P. C., 2002 Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes? Trends Microbiol. **10:** 393–396.

Rocha, E. P. C., and A. Blanchard, 2002 Genomic repeats, genome plasticity and the dynamics of *Mycoplasma* evolution. Nucleic Acids Res. **30:** 2031–2042.

Rocha, E. P. C., and A. Danchin, 2001 Ongoing evolution of strand composition in bacterial genomes. Mol. Biol. Evol. **18:** 1789–1799.

Rocha, E. P. C., A. Danchin and A. Viari, 1999a Analysis of long repeats in bacterial genomes reveals alternative evolutionary mechanisms in *Bacillus subtilis* and other competent prokaryotes. Mol. Biol. Evol. **16:** 1219–1230.

Rocha, E. P. C., A. Danchin and A. Viari, 1999b Functional and evolutionary roles of long repeats in prokaryotes. Res. Microbiol. **150:** 725–733.

Romero, D., and R. Palacios, 1997 Gene amplification and genomic plasticity in prokaryotes. Annu. Rev. Genet. **31:** 91–111.

Roth, J. R., N. Benson, T. Galitski, K. Haack, J. G. Lawrence *et al.*, 1996 Rearrangements of the bacterial chromosome: formation and applications, pp. 2256–2276 in *Escherichia coli and Salmonella: Cellular and Molecular Biology*, edited by R. C. H. Neinhardt, J. L. Ingraham, E. C. C. Lin, K. Brooks Low, B. Magasanik *et al.* ASM Press, Washington, DC.

Saunders, N. J., A. C. Jeffries, J. F. Peden, D. W. Hood, H. Tettelin *et al.*, 2000 Repeat-associated phase variable genes in the complete genome sequence of *Neisseria meningitidis* strain MC58. Mol. Microbiol. **37:** 207–215.

Segall, A., M. J. Mahan and J. R. Roth, 1988 Rearrangement of the bacterial chromosome: forbidden inversions. Science **241:** 1314–1318.

Shen, P., and H. V. Huang, 1986 Homologous recombination in *Escherichia coli*: dependence on substrate length and homology. Genetics **112:** 441–457.

Shigenobu, S., H. Watanabe, M. Hattori, Y. Sakaki and H. Ishikawa, 2000 Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. Nature **407:** 81–86.

Smith, G. R., 1988 Homologous recombination in procaryotes. Microbiol. Rev. **52:** 1–28.

Smith, T. F., and M. S. Waterman, 1981 Comparison of bio-sequences. Adv. Appl. Math. **2:** 482–489.

Suyama, M., and P. Bork, 2001 Evolution of prokaryotic gene order: genome rearrangements in closely related species. Trends Genet. **17:** 10–13.

Tamas, I., L. Klasson, B. Canback, A. K. Naslund, A. S. Eriksson *et al.*, 2002 50 million years of genomic stasis in endosymbiotic bacteria. Science **296:** 2376–2379.

Tillier, E. R., and R. A. Collins, 2000a Genome rearrangement by replication-directed translocation. Nat. Genet. **26:** 195–197.

Tillier, E. R., and R. A. Collins, 2000b Replication orientation affects the rate and direction of bacterial gene evolution. J. Mol. Evol. **51:** 459–463.

Vulic, M., F. Dionisio, F. Taddei and M. Radman, 1997 Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. Proc. Natl. Acad. Sci. USA **94:** 9763–9767.

Wolf, Y. I., I. B. Rogozin, A. S. Kondrashov and E. V. Koonin, 2001 Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. Genome Res. **11:** 356–372.

Zivanovic, Y., P. Lopez, H. Philippe and P. Forterre, 2002 *Pyrococcus* genome comparison evidences chromosome shuffling-driven evolution. Nucleic Acids Res. **30:** 1902–1910.

Communicating editor: M. A. F. Noor