J Mol Evol (2008) 67:291–300
DOI 10.1007/s00239-008-9151-z

The Excess of Small Inverted Repeats in Prokaryotes

Emmanuel D. Ladoukakis · Adam Eyre-Walker

Received: 6 May 2008 / Accepted: 11 July 2008 / Published online: 12 August 2008

Abstract Recent analyses have shown that there is a large excess of perfect inverted repeats in many prokaryotic genomes but not in eukaryotic ones. This difference could be due to a genuine difference between prokaryotes and eukaryotes or to differences in the methods and types of data analyzed – full genome versus protein coding sequences. We used simulations to show that the method used previously tends to underestimate the expected number of inverted repeats. However, this bias is not large and cannot explain the excess of inverted repeats observed in real data. In contrast, our method is unbiased. When both methods are applied to bacterial protein coding sequences they both detect an excess of inverted repeats, which is much lower than previously reported in whole prokaryotic genomes. This suggests that the reported large excess of inverted repeats is due to repeats found in intergenic regions. These repeats could be due to transcription factor binding sites, or other types of repetitive DNA, on opposite strands of the DNA sequence. In contrast, the smaller, but significant, excess of inverted repeats that we report in protein coding sequences may be due to sequence-directed mutagenesis (SDM). SDM is a process where one copy of a small, imperfect, inverted repeat corrects the other copy via strand misalignment, resulting in a perfect repeat and a series of mutations. We show by simulation that even very low levels of SDM, relative to the rate of point mutation, can generate a substantial excess of inverted repeats.

Keywords Sequence-directed mutagenesis · Genomic pattern · Prokaryotic genomes

E. D. Ladoukakis · A. Eyre-Walker
Centre for the Study of Evolution and School of Life Sciences, University of Sussex, Brighton, UK

E. D. Ladoukakis (✉)
Department of Biology, University of Crete, 71409 Iraklion, Greece
e-mail: ladoukakis@biology.uoc.gr

A. Eyre-Walker
National Evolutionary Synthesis Center, Durham, NC 27705, USA

Introduction

DNA repeats are known to be involved in several mutagenic processes. For example, the variation in copy number of short sequence DNA repeats, which is related to the virulence of many pathogenic bacteria (Hood et al. 1996), is caused by slippage of the enzymatic machinery during DNA replication (van Belkum et al. 1998). Direct and inverted DNA repeats can also form secondary structures when DNA is single stranded, which can lead to insertions, deletions, rearrangements, or sequence changes of the DNA. For example, in human cultured cells deletion of a 122-bp inverted repeat appears to be induced by the presence of 6-bp direct repeats which flank the sequence (Kramer et al. 1996). Recombination might also play an important role in repeat-related mutagenesis. Either homologous recombination (gene conversion) or illegitimate recombination (Chuzhanova et al. 2003) can cause sequence altering or DNA rearrangements. A specific repeat-related mutagenic process which does not involve recombination but is facilitated by the secondary structure of single-strand DNA is sequence-directed mutagenesis.

This is a process by which two imperfect inverted repeats correct each other to form two perfect repeats and a set of mutations. This process was originally discovered in the T4 lysozyme gene (Streisinger et al. 1966), in which spontaneous frameshift mutations appeared to be associated with DNA repeats (Okada et al. 1972). This led to the suggestion that repeated sequences mediated mutations by allowing local misalignments of the complementary strands of DNA. Sequence-directed mutagenesis has been observed in a variety of organisms such as *E. coli* (Halliday and Glickman 1991; Rosche et al. 1997) and yeast (Ripley and Shoemaker 1982). Blisser (1998) has also suggested that several human diseases might be caused by this process including fragile X, osteogenesis imperfecta, muscular dystrophy, and familial hypertension. However, the evidence for this is not strong, as we have shown elsewhere (Ladoukakis and Eyre-Walker 2007); there is very little evidence that sequence-directed mutagenesis occurs in eukaryotes generally.

In contrast to the situation in eukaryotes, there does appear to be widespread evidence of sequence-directed mutagenesis in prokaryotes. In a survey of 106 prokaryotic genomes, van Noort et al. (2003) showed that perfect inverted repeats were significantly more common than one would expect by chance; in some bacteria, repeats were nearly twice as common as one might expect. This led to the suggestion that sequence-directed mutagenesis was frequent and important.

Here we set out to investigate, first, whether the high excess of perfect inverted repeats previously reported in prokaryotes is due to a problem with the approach used or whether the excess is real. We show that the method is mildly biased, but we also confirm that there is a genuine excess of inverted repeats, though the excess is not as great as van Noort et al. (2003) report. Second, we quantify the rate of sequence-directed mutagenesis that is required to generate the excess of repeats found in prokaryotic genomes using computer simulations. We find that very low rates of sequence-directed mutagenesis can generate the observed levels of repeats.

Materials and Methods

Data Used

We downloaded the genes from 183 fully sequenced and annotated prokaryotic genomes from the NCBI database (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). We had downloaded all genomes of the database in October 2005. We used one genome per bacterial "species" to avoid redundancy. We used genes longer than 2000 bp and we restricted the analysis to the coding part of the genes because we wanted to avoid the inverted repeats which might have been due to transposable elements or generated by spontaneous duplications of small DNA sequences. We restricted our analysis to long protein coding sequences to ensure that there was sufficient sequence to allow proper randomization. Of the 183 prokaryotic genomes, we excluded all genomes that had fewer than 10 genes >2000 bp, which left 159 prokaryotic genomes (see Appendix). The number of large genes per genome ranged from 11 (*Nostoc* sp. plasmid) to 655 (*Rhodopirellula baltica*) genes, with a median of 132.

In every gene we counted the number of inverted repeats, setting two criteria: (i) the length of the repeats should be 6, 7, 8, or 9 bp; and (ii) the length of DNA between the two copies should be ≤50 bp. Every repeat was assigned to a length class (i.e., 6mers, 7mers, 8mers, and 9mers) when its length fit the specific length class exactly. For example, the 7mers included only repeats of 7 bp, not repeats of 8 bp, which would, of course, include two overlapping 7mers. Length classes and the length of the DNA between the copies of the repeats were based on previous studies (Fieldhouse and Golding 1991; van Noort et al. 2003; Ladoukakis and Eyre-Walker 2007).

We define an imperfect repeat as an inverted repeat that differs by a single nucleotide excluding the terminal base pairs; e.g., for the 7mers a difference could occur at only one of the five internal nucleotides of the repeat, and not at the first or the last nucleotide. Deletions and insertions were not taken into account.

Perfect and imperfect repeats were counted in actual and randomized sequences. We calculated 95% confidence intervals of the observed/expected ratio using bootstrapping by gene within a genome (Ladoukakis and Eyre-Walker 2007).

Randomization Methods

We used three different methods of randomization to estimate the number of inverted repeats expected by chance alone. First, we randomized the sequences preserving the dinucleotide composition of the sequence. This method is widely used for analysis of prokaryotic genomes (Lillo et al. 2002; van Noort et al. 2003) and is based on the observation that nucleotide composition (Fleischmann et al. 1995) and dinucleotide composition is relatively homogeneous over the entire bacterial chromosome (Karlin et al. 1997) (but see Kerr et al. [1997] for a notable exception). Second, we randomized the sequences preserving the amino acid order and codon usage by swapping synonymous codons. Finally, we randomized the sequences by only swapping the synonymous codons that were followed by the same nucleotide (e.g., CAC.G could be

swapped with CAT.G but not with CAT.C). This randomization pattern, as we have explained elsewhere (Ladoukakis and Eyre-Walker 2007), takes into account neighboring nucleotide effects, which can increase the probability of observing perfect repeats. With the last two randomization methods, we take into account the fact that some inverted repeats might be overrepresented because specific amino acid combinations are favored by natural selection.

Simulations

We ran two sets of simulations. The first set was run to evaluate the three methods of randomization. The simulated sequences were 3000-bp-long protein coding sequences. The amino acid composition of each sequence was random but there was codon bias toward a specific G+C content. We simulated 11 sets of sequences. Each set had 1000 sequences and its sequences had the same G+C content at synonymous sites, ranging between 0% and 100% in increments of 10% (0, 10, 20,…, 100%).

The second simulation analysis was run to estimate how the excess of perfect repeats depended on the relative rates of sequence-directed mutagenesis and point mutation. We simulated a single 3000-bp protein coding sequence with random amino acid composition. Each generation this sequence was subjected to point mutation and sequence-directed mutagenesis. In all simulations the rate of point mutation, u, was set at 0.001 per nucleotide per generation and the rate of sequence-directed mutagenesis was varied. If the mutation introduced by either process changed an amino acid, it was rejected with a probability of 0.97. This probability was chosen to reflect the very high level of constraint which is typically observed in bacteria; e.g., the average ratio of nonsynonymous-to-synonymous substitutions (dN/dS) between *Escherichia coli* and *Salmonella enterica* is about 0.05 (Clark et al. 1999; Charlseworth and Eyre-Walker 2006). Sequence-directed mutagenesis was applied in the following manner. We first determined the expected number of sequence-directed mutagenesis mutations which would occur in our sequence given the mutation rate; for example, if the rate of sequence-directed mutagenesis mutation was 0.0001, we would on average introduce 0.3 of a sequence-directed mutagenesis mutation each generation (i.e., 3000 bp $\times$ 0.0001 = 0.3). We generated a random Poisson deviate with a mean of 0.3 to give us the number of sequence-directed mutagenesis mutations in that generation. If this was $\geq$1, we randomly chose a 7-bp oligo from within the 3000-bp sequence. We then searched 50 bp upstream and downstream of the oligo for an imperfect inverted repeat which differed from the oligo by 1 bp. If this imperfect repeat was found, we corrected this oligo to be identical to the target oligo, thus inducing a sequence-directed mutagenesis mutation. If an imperfect repeat was not found, another target oligo was randomly chosen from the protein-coding sequence and the process repeated until the required number of sequence-directed mutagenesis events had been introduced in that generation. The sequence was allowed to equilibrate for $1/u$ generations before being sampled every $1/u$ generations until 1000 samples had been taken. At each sampling point we calculated the number of 7-bp perfect and imperfect repeats within 50 bp. We also applied the three randomization methods.

Results

Investigation of the Randomization Methods

To test whether there is an excess of inverted repeats we need to know how many inverted repeats we expect by chance alone; this expectation is estimated by randomizing the sequences but there are several ways in which this may be done. van Noort et al. (2003) randomized the entire genome sequence preserving the frequencies of dinucleotides, but this can be biased if the base composition varies between sites in a systematic fashion. We have proposed, alternatively, that one might randomize protein coding sequences by shuffling synonymous codons either with or without reference to the next nucleotide in the sequence. To investigate the relative merits of these three methods we simulated the evolution of a protein coding sequences that was subject to point mutation, but no sequence-directed mutagenesis. We allow there to be base composition bias at the third codon position.

As expected, randomization of the sequence keeping the dinucleotide composition constant is biased; the method tends to underestimate the expected number of inverted repeats (Fig. 1). However, the effect is small compared to the excess of repeats found in real sequences (Fig. 2). In contrast randomizing synonymous codons is unbiased.

Excess of Inverted Repeats in Prokaryotic Genomes

We used our unbiased method of randomization to investigate the excess of perfect inverted repeats in a large set of prokaryotic genomes. In total, 159 genomes were analyzed, which included genomes analyzed in previous studies (van Noort et al. 2003). We tested for an excess of four length classes of repeats; 6mers, 7mers, 8mers, and 9mers. Here we present the summation of the four classes because each individual class did not give different results. We found

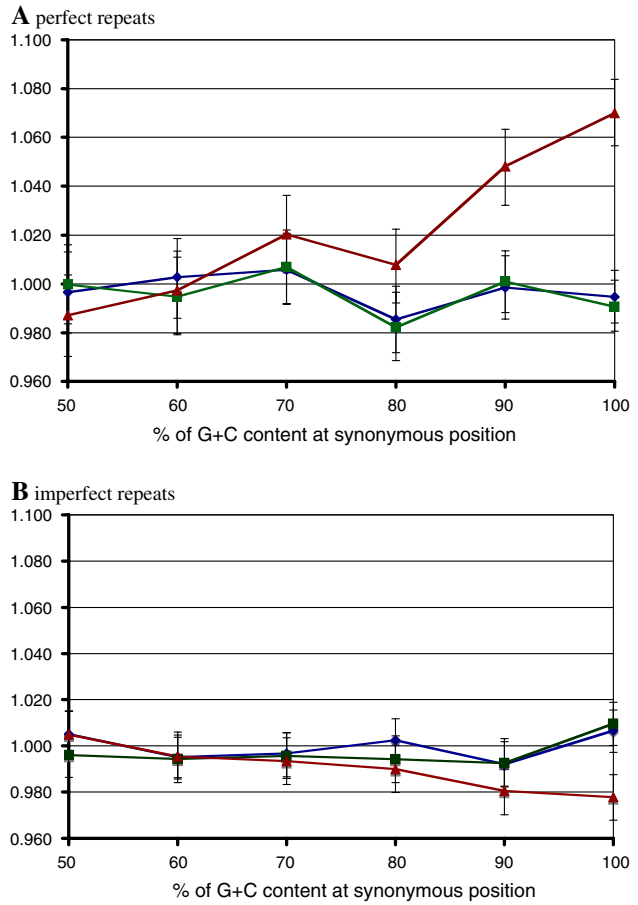

Fig. 1 Ratio of observed (obs)-to-expected (exp) numbers of (**a**) perfect and (**b**) imperfect repeats in simulated sequences using three randomization methods: the dinucleotide method (triangles), shuffling synonymous sites (diamonds), and shuffling synonymous sites and preserving the neighbor nucleotide (squares)

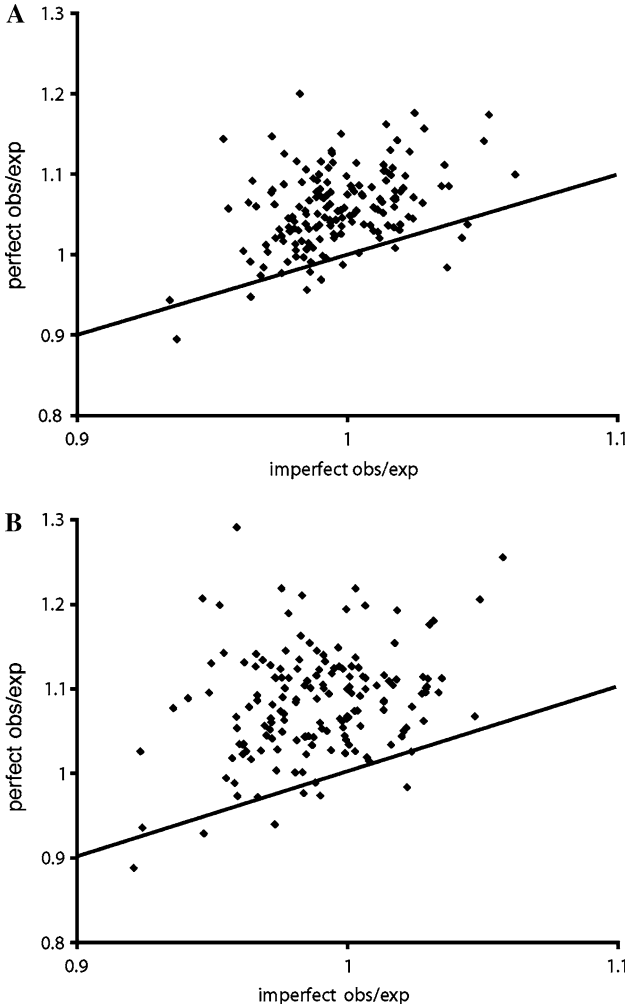

Fig. 2 The relative abundance of perfect and imperfect inverted repeats in 159 prokaryotic genomes. Each point is a genome. The black line in each graph shows the case where the ratio of observed to expected is equal for perfect and imperfect repeats. Each dot represents the summation over all repeat lengths considered: 6-, 7-, 8-, and 9mers. (**a**) Our randomization method. (**b**) Dinucleotide randomization method

that most prokaryotic genomes show an excess of perfect repeats, although in most cases the excess is quite small, much smaller than previously reported (van Noort et al. 2003) (Fig. 2a). To compare different randomization methods in real data, we also used dinucleotide randomization to estimate the expected number of inverted repeats in our protein coding sequence collection. We observed an increase in the ratio of observed to expected inverted repeats (Fig. 2b) in accordance with the simulation results (see previous paragraph).

Based on our unbiased randomization method, the biggest excess of inverted repeats we observed when repeats of all lengths were considered together was 20%, which was in a plasmid of *Nostoc* sp. In contrast, the excess that van Noort et al. (2003) observed in some prokaryotic genomes was close to 100%. In all, 102 of the 159 prokaryotic genomes that we analyzed had a significant excess of perfect repeats and 8 genomes had a significant deficit, when we considered all repeats together. The biggest deficit of repeats was 10.5% in (*Synechocystis* sp.); in the other cases it was lower than 5.6 %.

Simulations of Sequence-Directed Mutagenesis vs. Point Mutations

Although there is a significant excess of inverted repeats, it is not immediately apparent whether this is due to a high or a low rate of sequence-directed mutagenesis. To investigate this we ran another set of simulations in which sequences were subjected to both sequence-directed mutagenesis and point mutations; point mutations will tend to eliminate perfect inverted repeats and sequence-directed mutagenesis will tend to generate them. We kept the rate of point mutation constant (mutation rate, 0.001) and we applied

Table 1 Estimating the strength of sequence-directed mutagenesis relative to random mutagenesis: ratio of point mutation rate to sequence-directed mutation rate and ratio of observed (obs)-to-expected (exp) perfect inverted repeats N

Mutation rate/sequence-directed mutagenesis rate	Obs/exp
10	5.342 (5.297–5.392)
100	1.610 (1.592–1.629)
1000	1.027 (1.014–1.040)

Note: Numbers in parentheses are 95% confidence intervals

three different sequence-directed mutagenesis rates, 0.0001, 0.00001, and 0.000001, which are 10, 100, and 1000 times lower than the rate of point mutation, respectively. Surprisingly, even when the level of sequence-directed mutagenesis was 100 times lower than the rate of point mutation, the number of imperfect repeats was much higher than expected in randomized sequences (Table 1 and Appendix). Thus, very low levels of sequence-directed mutagenesis can induce a substantial excess of inverted repeats.

Discussion

Theoretical studies suggest that sequence-directed mutagenesis could be an important source of perfect inverted repeats in genomes (Fieldhouse and Golding 1991). This process has been observed in individual loci in a variety of organisms, ranging from T4 virus (Okada et al. 1972; Streisinger et al. 1966) to bacteria (Halliday and Glickman 1991). Recently, two studies investigated the impact of sequence-directed mutagenesis in the formation of short inverted repeats on a genomic scale. The first study looked at 106 prokaryotic genomes (bacteria and archaea) and reported a high excess of perfect inverted repeats, which was attributed to sequence-directed mutagenesis (van Noort et al. 2003). On the other hand, Ladoukakis and Eyre-Walker (2007) examined protein coding genes from five eukaryotic genomes and found very little excess of inverted repeats. This contradictory result could be due to three reasons. First, one of the two methodologies might be biased. Second, there might be other sources of small inverted repeats that were not taken into account because the two studies used different genomic sequences (the first used whole genomes and the second used protein coding sequences). Third, there might be a real difference between prokaryotic and eukaryotic genomes. To investigate these issues we have performed here an analysis of prokaryotic coding sequences. We find that there is an excess of perfect inverted repeats in prokaryotic DNA using methods identical to those used in eukaryotes. This suggests that there is a fundamental difference between prokaryotes and eukaryotes in the frequency of inverted repeats.

However, van Noort et al. (2003) found a much larger excess of repeats than we found in this study. This could be due to differences in methodology or differences in the data used. We have shown that the method employed by van Noort et al. (2003) is biased; it tends to overestimate the number of perfect repeats relative to random expectations when there is strong base composition bias, as we find in genomes such as *Mycoplasma* (G+C content of 25%) or *Micrococcus* (G+C content of 75%) (Lawrence and Ochman 1997). However, the bias in their method is not large, not nearly large enough to explain why van Noort et al. (2003) found a much larger excess of repeats than we did. Furthermore, we find a rather similar excess of inverted repeats when we use their methodology on our protein coding sequences. It therefore seems likely that the reason for the difference between the two studies is due to the data used. van Noort et al. (2003) used complete genomes, whereas we only used protein coding sequences; it therefore seems that many of the perfect repeats they found were in intergenic DNA. This is not unexpected since bacterial genomes are known to contain transposable elements which are rich in inverted repeats (Kleckner 1981). Furthermore, transcription factor binding sites on opposite strands of the DNA could also generate an excess of inverted repeats, and it has been shown that short inverted repeats are preferentially located in noncoding DNA, close to the 3′ end of the nearest coding region (Lillo et al. 2002) and are involved in important biological processes, such as regulation of transcription and translation (Blatt et al. 1993; Raghunathan et al. 1991), genome rearrangements (Bi and Liu 1996), and genome instability (Achaz et al. 2002).

Most interestingly we show that the excess of repeats, which we found in the coding sequences we surveyed, could be due to very low levels of sequence-directed mutagenesis. Sequence-directed mutagenesis can be several orders of magnitude lower than the rate of point mutation and still produce an excess of perfect inverted repeats. It therefore seems that while sequence-directed mutagenesis may have altered the pattern of DNA sequences, it does not generally generate many mutations.

Acknowledgments E.L. was funded by a Marie Curie fellowship, A.E.-W. was funded by the National Evolutionary Synthesis Center and the Biotechnology and Biological Sciences Research Council.

Appendix

List of prokaryotic genomes

	GenBank accession no.	No. of genes (≥2000 bp)	Name	Obs/exp: perfect repeats	Obs/exp: imperfect repeats
1	NC_005966	204	*Acinetobacter* sp.	1.099	1.0211
2	NC_000854	71	*Aeropyrum pernix*	1.0324	0.9745
3	NC_003062	148	*Agrobacterium tumefaciens*	1.0031	1.0038
4	NC_004842	91	*Anaplasma marginale*	1.0248	0.995
5	NC_000918	79	*Aquifex aeolicus*	1.0605	0.9818
6	NC_000917	83	*Archaeoglobus fulgidus*	1.0927	0.9645
7	NC_006513	283	*Azoarcus* sp.	1.0177	0.9759
8	NC_003997	170	*Bacillus anthracis*	1.0793	1.0192
9	NC_006274	192	*Bacillus cereus*	1.0604	1.017
10	NC_006582	149	*Bacillus clausii*	1.0757	0.9996
11	NC_002570	181	*Bacillus halodurans*	1.0711	0.9908
12	NC_006270	165	*Bacillus licheniformis*	1.102	1.015
13	NC_000964	167	*Bacillus subtilis*	1.0933	0.9884
14	NC_000117	79	*Chlamydia trachomatis*	1.1631	1.014
15	NC_000853	94	*Thermotoga maritima*	1.0956	0.9871
16	NC_000868	73	*Pyrococcus abyssi*	1.0478	0.9794
17	NC_000911	231	*Synechocystis* sp.	0.8959	0.9365
18	NC_000913	257	*Escherichia coli* K12	0.9975	0.9834
19	NC_000915	113	*Helicobacter pylori*	1.1481	0.9717
20	NC_000916	76	*Methanothermobacter thermautotrophicus*	1.1272	0.994
21	NC_000958	17	*Deinococcus radiodurans*	0.9923	0.9637
22	NC_000961	69	*Pyrococcus horikoshii*	1.0788	0.9935
23	NC_000963	63	*Rickettsia prowazekii*	1.0471	0.9913
24	NC_001988	12	*Clostridium acetobutylicum*	0.9848	1.0366
25	NC_002163	85	*Campylobacter jejuni*	1.1303	0.9938
26	NC_002488	170	*Xylella fastidiosa*	1.0871	1.0039
27	NC_002516	370	*Pseudomonas aeruginosa*	1.0594	0.9985
28	NC_002578	59	*Thermoplasma acidophilum*	1.0929	1.014
29	NC_002662	120	*Lactococcus lactis*	1.0907	0.9922
30	NC_002663	137	*Pasteurella multocida*	1.0303	1.0093
31	NC_002677	117	*Mycobacterium leprae*	1.0815	0.9719
32	NC_002678	307	*Mesorhizobium loti*	1.0489	1.0005
33	NC_002689	60	*Thermoplasma volcanium*	1.0386	1.0441
34	NC_002696	277	*Caulobacter crescentus*	0.9995	0.9905
35	NC_002737	97	*Streptococcus pyogenes*	1.0519	0.9886
36	NC_002754	96	*Sulfolobus solfataricus*	0.9916	0.9862
37	NC_002927	271	*Bordetella bronchiseptica*	1.038	0.9893
38	NC_002928	247	*Bordetella parapertussis*	1.0351	0.978
39	NC_002929	213	*Bordetella pertussis*	1.0423	0.98
40	NC_002935	149	*Corynebacterium diphtheriae*	1.0529	1.0129
41	NC_002936	74	*Dehalococcoides ethenogenes*	1.1007	0.9891
42	NC_002937	230	*Desulfovibrio vulgaris*	1.0363	0.9975
43	NC_002939	270	*Geobacter sulfurreducens*	1.0215	0.9947
44	NC_002940	88	*Haemophilus ducreyi*	1.0315	1.0189
45	NC_002944	232	*Mycobacterium avium*	1.0657	0.963

continued

	GenBank accession no.	No. of genes (≥2000 bp)	Name	Obs/exp: perfect repeats	Obs/exp: imperfect repeats
46	NC_002945	256	*Mycobacterium bovis*	1.0328	0.9853
47	NC_002946	105	*Neisseria gonorrhoeae*	1.0884	0.9761
48	NC_002947	377	*Pseudomonas putida*	1.0336	0.9801
49	NC_002950	175	*Porphyromonas gingivalis*	1.0795	1.0021
50	NC_002951	118	*Staphylococcus aureus*	1.0652	1.0275
51	NC_002967	180	*Treponema denticola*	1.1125	1.013
52	NC_002971	93	*Coxiella burnetii*	1.0698	1.0175
53	NC_002977	245	*Methylococcus capsulatus*	1.0512	0.9844
54	NC_003037	50	*Sinorhizobium meliloti*	1.0201	0.9894
55	NC_003098	109	*Streptococcus pneumoniae*	1.0702	1.0118
56	NC_003103	64	*Rickettsia conorii*	1.074	1.0053
57	NC_003106	91	*Sulfolobus tokodaii*	0.9884	0.998
58	NC_003112	115	*Neisseria meningitidis*	1.0829	0.9896
59	NC_003155	519	*Streptomyces avermitilis*	0.9749	0.9676
60	NC_003197	270	*Salmonella typhimurium*	1.0549	0.996
61	NC_003198	232	*Salmonella enterica*	1.0443	0.9954
62	NC_003210	150	*Listeria monocytogenes*	1.0154	0.9853
63	NC_003212	152	*Listeria innocua*	0.9485	0.9638
64	NC_003228	487	*Bacteroides fragilis*	1.0638	0.9932
65	NC_003240	11	*Nostoc* sp.	1.2011	0.982
66	NC_003295	210	*Ralstonia solanacearum*	0.9779	0.9754
67	NC_003317	105	*Brucella melitensis*	1.0457	1.0239
68	NC_003361	92	*Chlamydophila caviae*	1.1576	1.028
69	NC_003364	91	*Pyrobaculum aerophilum*	1.1513	0.9973
70	NC_003413	69	*Pyrococcus furiosus*	1.0595	0.9919
71	NC_003454	111	*Fusobacterium nucleatum*	1.0863	1.0372
72	NC_003551	66	*Methanopyrus kandleri*	1.0848	1.0086
73	NC_003552	297	*Methanosarcina acetivorans*	1.109	0.9927
74	NC_003869	114	*Thermoanaerobacter tengcongensis*	1.0732	0.99
75	NC_003888	505	*Streptomyces coelicolor*	0.9697	0.9899
76	NC_003901	212	*Methanosarcina mazei*	1.0986	0.9993
77	NC_003902	384	*Xanthomonas campestris*	0.9985	0.9806
78	NC_003910	387	*Colwellia psychrerythraea*	1.0747	0.992
79	NC_003911	205	*Silicibacter pomeroyi*	1.0064	0.9806
80	NC_003912	87	*Campylobacter jejuni*	1.1066	0.9843
81	NC_003919	401	*Xanthomonas axonopodis*	1.0043	0.9701
82	NC_004061	35	*Buchnera aphidicola*	1.048	1.0223
83	NC_004113	158	*Thermosynechococcus elongatus*	0.9444	0.9339
84	NC_004116	118	*Streptococcus agalactiae*	1.0558	1.003
85	NC_004129	407	*Pseudomonas fluorescens*	1.0313	0.9815
86	NC_004193	130	*Oceanobacillus iheyensis*	1.0573	1.0109
87	NC_004307	165	*Bifidobacterium longum*	0.9573	0.9847
88	NC_004310	99	*Brucella suis*	1.0219	1.042
89	NC_004337	205	*Shigella flexneri*	1.0268	0.9946
90	NC_004347	332	*Shewanella oneidensis*	1.1053	1.0136
91	NC_004350	113	*Streptococcus mutans*	1.0681	0.9855
92	NC_004369	178	*Corynebacterium efficiens*	1.0375	1.0081
93	NC_004459	180	*Vibrio vulnificus*	1.0704	1.0168

continued

	GenBank accession no.	No. of genes (≥2000 bp)	Name	Obs/exp: perfect repeats	Obs/exp: imperfect repeats
94	NC_004461	111	*Staphylococcus epidermidis*	1.0597	1.0092
95	NC_004463	457	*Bradyrhizobium japonicum*	1.0178	0.9827
96	NC_004545	34	*Buchnera aphidicola*	0.9967	0.9918
97	NC_004547	264	*Erwinia carotovora*	1.0554	0.9841
98	NC_004551	61	*Tropheryma whipplei*	1.1171	0.9809
99	NC_004552	92	*Chlamydophila abortus*	1.131	1.0155
100	NC_004557	135	*Clostridium tetani*	1.0215	1.0115
101	NC_004567	158	*Lactobacillus plantarum*	1.0722	1.0174
102	NC_004578	400	*Pseudomonas syringae*	1.0216	0.973
103	NC_004603	190	*Vibrio parahaemolyticus*	1.0418	1.0013
104	NC_004668	187	*Enterococcus faecalis*	1.0412	0.9934
105	NC_004757	203	*Nitrosomonas europaea*	1.1157	0.9943
106	NC_004917	112	*Helicobacter hepaticus*	1.078	0.9715
107	NC_005027	655	*Rhodopirellula baltica*	1.076	1.005
108	NC_005042	80	*Prochlorococcus marinus*	1.0267	0.9972
109	NC_005061	36	Candidatus *Blochmannia floridanus*	1.0139	0.9804
110	NC_005070	118	*Synechococcus* sp.	1.1152	1.0029
111	NC_005085	267	*Chromobacterium violaceum*	1.0634	0.9727
112	NC_005090	132	*Wolinella succinogenes*	1.1263	0.9763
113	NC_005125	320	*Gloeobacter violaceus*	1.046	0.9778
114	NC_005126	316	*Photorhabdus luminescens*	1.091	0.9828
115	NC_005213	25	*Nanoarchaeum equitans*	1.1124	1.0356
116	NC_005296	325	*Rhodopseudomonas palustris*	1.061	0.9659
117	NC_005362	109	*Lactobacillus johnsonii*	1.129	1.0226
118	NC_005363	250	*Bdellovibrio bacteriovorus*	1.0346	1.0168
119	NC_005791	57	*Methanococcus maripaludis*	1.177	1.0245
120	NC_005823	219	*Leptospira interrogans*	1.0666	1.0125
121	NC_005861	191	Candidatus *Protochlamydia amoebophila*	1.0344	0.988
122	NC_005877	71	*Picrophilus torridus*	1.0826	1.0015
123	NC_005955	94	*Bartonella quintana*	1.0863	1.0344
124	NC_005956	108	*Bartonella henselae*	1.0582	1.0099
125	NC_005957	175	*Bacillus thuringiensis*	1.0724	1.0241
126	NC_006085	163	*Propionibacterium acnes*	1.0244	0.9754
127	NC_006087	90	*Leifsonia xyli*	0.9918	0.9775
128	NC_006138	215	*Desulfotalea psychrophila*	1.0562	0.9972
129	NC_006142	65	*Rickettsia typhi*	1.0373	1.0147
130	NC_006177	155	*Symbiobacterium thermophilum*	1.0391	0.9843
131	NC_006300	120	*Mannheimia succiniciproducens*	1.0723	0.988
132	NC_006348	159	*Burkholderia mallei*	0.9796	0.9858
133	NC_006350	217	*Burkholderia pseudomallei*	1.0096	0.9869
134	NC_006449	83	*Streptococcus thermophilus*	1.08	1.0172
135	NC_006512	212	*Idiomarina loihiensis*	1.1049	1.013
136	NC_006526	121	*Zymomonas mobilis*	1.1749	1.052
137	NC_006570	91	*Francisella tularensis*	1.1005	1.0617
138	NC_006576	151	*Synechococcus elongatus*	1.1168	0.9898
139	NC_006624	87	*Thermococcus kodakarensis*	1.013	0.9695
140	NC_006672	12	*Gluconobacter oxydans*	1.0583	0.9556
141	NC_006814	111	*Lactobacillus acidophilus*	1.1431	1.0181

continued

	GenBank accession no.	No. of genes (≥2000 bp)	Name	Obs/exp: perfect repeats	Obs/exp: imperfect repeats
142	NC_006831	77	*Ehrlichia ruminantium*	1.0299	0.9783
143	NC_006833	54	*Wolbachia* endosymbiont strain TRS of *Brugia malayi*	1.0092	1.0173
144	NC_006834	376	*Xanthomonas oryzae*	1.0075	0.9845
145	NC_006840	168	*Vibrio fischeri*	1.1001	1.0158
146	NC_006932	99	*Brucella abortus*	1.0385	1.0192
147	NC_006958	171	*Corynebacterium glutamicum*	1.0441	0.9932
148	NC_007109	79	*Rickettsia felis*	1.0838	1.0204
149	NC_007164	155	*Corynebacterium jeikeium*	0.9852	0.9686
150	NC_007168	112	*Staphylococcus haemolyticus*	1.1422	1.0502
151	NC_007181	80	*Sulfolobus acidocaldarius*	1.0518	1.0021
152	NC_007204	147	*Psychrobacter arcticus*	1.0385	1.0057
153	NC_007205	68	Candidatus *Pelagibacter ubique*	1.0369	0.9934
154	NC_007292	38	Candidatus *Blochmannia pennsylvanicus*	1.0865	1.0009
155	NC_007298	332	*Dechloromonas aromatica*	1.0469	0.9977
156	NC_007333	216	*Thermobifida fusca*	1.109	1.0164
157	NC_007336	21	*Ralstonia eutropha*	1.0055	0.9611
158	NC_007354	80	*Ehrlichia canis*	1.0292	1.0109
159	NC_007356	66	*Dehalococcoides* sp.	1.1447	0.9537

Note: The ratios of observed-to-expected (obs/exp) repeats contain all length classes (i.e., 6-, 7-, 8-, and 9mers). Expected numbers of repeats were estimated using our unbiased method (see Materials and Methods)

References

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

Bi X, Liu LF (1996) DNA rearrangement mediated by inverted repeats. Proc Natl Acad Sci USA 93:819–823

Blatt NB, Osborne SE, Cain RJ, Glick GD (1993) Conformational studies of hairpin sequences from the ColE1 cruciform. Biochimie 75:433–441

Blisser JJ (1998) DNA inverted repeats and human diseases. Front Biosci 3:d408–d418

Charlseworth J, Eyre-Walker A (2006) The rate of adaptive evolution in enteric bacteria. Mol Biol Evol 23:1348–1356

Chuzhanova N, Abeysinghe SS, Krawczak M, Cooper DN (2003) Translocation and gross deletion breakpoints in human inherited disease and cancer II: potential involvement of repetitive sequence elements in secondary structure formation between DNA ends. Hum Mutat 22:245–251

Clark MA, Moran NA, Baumann P (1999) Sequence evolution in bacterial endosymbionts having extreme base compositions. Mol Biol Evol 16:1586–1598

Fieldhouse D, Golding B (1991) A source of small repeats in genomic DNA. Genetics 129:563–572

Fleischmann RD, Adams MD, White O et al (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269:496–512

Halliday JA, Glickman BW (1991) Mechanisms of spontaneous mutation in DNA repair-proficient Escherichia coli. Mutat Res 250:55–71

Hood DW, Deadman ME, Jennings MP, Bisercic M, Fleischmann RD, Venter JC, Moxon ER (1996) DNA repeats identify novel virulence genes in Haemophilus influenzae. Proc Natl Acad Sci USA 93:11121–11125

Karlin S, Mrazek J, Campbell AM (1997) Compositional biases of bacterial genomes and evolutionary implications. J Bacteriol 179:3899–3913

Kerr AR, Peden JF, Sharp PM (1997) Systematic base composition variation around the genome of Mycoplasma genitalium, but not Mycoplasma pneumoniae. Mol Microbiol 25:1177–1179

Kleckner N (1981) Transposable elements in prokaryotes. Annu Rev Genet 15:341–404

Kramer PR, Stringer JR, Sinden RR (1996) Stability of an inverted repeat in a human fibrosarcoma cell. Nucleic Acids Res 24:4234–4241

Ladoukakis ED, Eyre-Walker A (2007) Searching for sequence directed mutagenesis in eukaryotes. J Mol Evol 64:1–3

Lawrence JG, Ochman H (1997) Amelioration of bacterial genomes: rates of change and exchange. J Mol Evol 44:383–397

Lillo F, Basile S, Mantegna RN (2002) Comparative genomics study of inverted repeats in bacteria. Bioinformatics 18:971–979

Okada Y, Newton J, Owen J, Tsugita A, Streisin G, Inouye M (1972) Molecular basis of a mutational hot spot in lysozyme gene of bacteriophage-T4. Nature 236:338

Raghunathan G, Jernigan RL, Miles HT, Sasisekharan V (1991) Conformational feasibility of a hairpin with two purines in the loop. 5′-d-GGTACIAGTACC-3′. Biochemistry 30:782–788

Ripley LS, Shoemaker NB (1982) Polymerase infidelity and frameshift mutation. Basic Life Sci 20:161–178

Rosche WA, Trinh TQ, Sinden RR (1997) Leading strand specific spontaneous mutation corrects a quasipalindrome by an intermolecular strand switch mechanism. J Mol Biol 269: 176–187

Streisinger G, Okada Y, Emrich J, Newton J, Tsugita A, Terzaghi E, Inouye M (1966) Frameshift mutations and the genetic code. Cold Spring Harb Symp Quant Biol 31:77–84

van Belkum A, Scherer S, van Alphen L, Verbrugh H (1998) Short-sequence DNA repeats in prokaryotic genomes. Microbiol Mol Biol Rev 62:275–293

van Noort V, Worning P, Ussery DW, Rosche WA, Sinden RR (2003) Strand misalignments lead to quasipalindrome correction. Trends Genet 19:365–369