The Prevalence and Evolutionary Conservation of Inverted Repeats in Proteobacteria

Bar Lavi[1,2], Eli Levy Karin[1,3], Tal Pupko[1,*], and Einat Hazkani-Covo[2,*]

[1]Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Israel

[2]Department of Natural and Life Sciences, The Open University of Israel, Ra'anana, Israel

[3]Department of Molecular Biology & Ecology of Plants, George S. Wise Faculty of Life Sciences, Tel Aviv University, Israel

*Corresponding authors: E-mails: talp@post.tau.ac.il; einatco@openu.ac.il.

Accepted: February 21, 2018

Abstract

Perfect short inverted repeats (IRs) are known to be enriched in a variety of bacterial and eukaryotic genomes. Currently, it is unclear whether perfect IRs are conserved over evolutionary time scales. In this study, we aimed to characterize the prevalence and evolutionary conservation of IRs across 20 proteobacterial strains. We first identified IRs in *Escherichia coli* K-12 substr MG1655 and showed that they are overabundant. We next aimed to test whether this overabundance is reflected in the conservation of IRs over evolutionary time scales. To this end, for each perfect IR identified in *E. coli* MG1655, we collected orthologous sequences from related proteobacterial genomes. We next quantified the evolutionary conservation of these IRs, that is, the presence of the exact same IR across orthologous regions. We observed high conservation of perfect IRs: out of the 234 examined orthologous regions, 145 were more conserved than expected, which is statistically significant even after correcting for multiple testing. Our results together with previous experimental findings support a model in which imperfect IRs are corrected to perfect IRs in a preferential manner via a template switching mechanism.

Key words: inverted repeats, palindrome evolution, palindrome conservation, template switching.

Introduction

A perfect inverted repeat (IR) is a stretch of DNA consisting of two DNA sequences (arms) in reverse complement orientation (e.g., 5′ *AGAACAxxxxTGTTCT* 3′). When the arms are directly adjacent to each other, the sequence is called a palindrome (e.g., 5′ *AGAACATGTTCT* 3′). When the arms do not perfectly match, the segment would be termed an imperfect IR, or a quasi-palindrome. The reverse symmetry between the arms allows them to form base pairs with each other, resulting in DNA secondary structures, such as hairpins and cruciforms (Lilley 1980; Panayotatos and Wells 1981).

IRs are important for a wide range of biological processes, including replication, transcription, and DNA repair. IRs are found to be abundant in viral origins of replication (Pearson et al. 1996) and in bacterial plasmids (del Solar et al. 1998). In addition, in *Escherichia coli*, imperfections in the IR sequence of the *lac* repressor gene regulatory region were shown to decrease the binding affinity of the transcription machinery to that gene (Sadler et al. 1983). Moreover, IRs can be involved in alternative termination of bacterial genes (Li et al. 1997) and participate in the process of immunoglobulin V(D)J gene rearrangement (Cuomo et al. 1996). IRs are also important for RNA functionality. Endogenous transcripts bearing IRs could be processed to produce small RNAs that are involved in gene silencing (Okamura et al. 2008; Wroblewski et al. 2014).

Numerous studies of the DNA repair mechanism have demonstrated the role of long IRs in promoting genome instability both in prokaryotes and eukaryotes. For example, the bacterial transposon Tn5 that includes a 1.5-kb long IR was found to be unstable when inserted into *Saccharomyces cerevisiae* (Gordenin et al. 1993). Following DNA replication, IRs can undergo a process in which their length increases, enhancing genome instability (Butler et al. 1996; Tanaka et al. 2002; Brewer et al. 2011). Instability of long IRs was shown to be associated with higher probability for translocation and deletion events (Richard et al. 2008; Mizuno et al. 2009; Branzei and Foiani 2010; Lu et al. 2015). It was additionally shown that nearby IRs in yeast fuse to form unstable dicentric chromosomes, resulting in translocations and other gross chromosomal changes (Paek et al. 2009). In addition,

Alus inserted as IRs in the yeast genome (*Alu*-IR) induce double strand breaks (DSBs) that are terminated by hairpins and are mitotic recombination hotspots. Failure to process the hairpins leads to generation of chromosome inverted duplications (Lobachev et al. 2002). Furthermore, secondary structures of IRs can stall the DNA replication process, leading to DSBs and deletions, for example, *Alu*-IRs were shown to be highly unstable and to stall the replication fork (Voineagu et al. 2008). In addition, *Alu*-IRs are very rare in the human genome, suggesting that long IRs are selected against (Lobachev et al. 2000). Secondary structures formed by IRs were also found to be related to gene amplification (Pearson et al. 1996; Tanaka et al. 2002; Narayanan et al. 2006). For example, in breast cancer it was shown that palindrome formation proceeds gene amplification (Tanaka et al. 2005; Tanaka and Yao 2009).

In contrast to long IRs, several reports showed that perfect short IRs are abundant in genomes. Cox and Mirkin (1997) showed that short perfect IRs are enriched in human, *S. cerevisiae, E. coli*, and *Haemophilus influenzae.* This enrichment of short perfect IRs in bacterial genomes was further reported by others (Lillo et al. 2002; van Noort, Snel, et al. 2003; Ladoukakis and Eyre-Walker 2008; Strawbridge et al. 2010).

A striking characteristic of IRs is their tendency to undergo concerted evolution, which leads to the removal of variation between the two arms of an IR. For example, it was reported that in chloroplasts of land plants, long IRs undergo frequent recombination to produce identical IR forms (Kolodner and Tewari 1979; Grant et al. 1980; Turmel et al. 2017). Concerted evolution of IRs can be mediated by two mechanisms: homologous recombination and template switching (see below). Homologous recombination is known to occur in long IRs, while template switching can occur in both long and short IRs. While concerted evolution of long IRs through homologous recombination is relatively well studied (Warburton et al. 2004; Chen et al. 2007; Darmon and Leach 2014), the impact of template switching of short IRs on genome evolution received little to no attention.

DNA template switching occurs when the DNA polymerase hops between templates. Template switching between IR arms was first identified in bacteriophages by Ripley (1982). The mechanism requires two template switching steps and can occur through an intra or inter-molecular template switching (Ripley 1982). An example of the template switching mechanism is shown in figure 1. Starting with a template represented by an imperfect IR of arm size 5 bp (lower sequence in fig. 1*A*), the first switch occurs after the newly synthesized strand replicates the template of the 3′ arm. This switch is to the complementary arm, either in the nascent DNA strand (intramolecular, fig. 1*B*) or across the replication fork (intermolecular, fig. 1*C*). The second switch (fig. 1*D*) occurs when the replication fork realigns back to the native template. At the end of this process, one strand has an imperfect IR and the other strand has a perfect IR. At the next round of replication, one of the daughter cells harbors a perfect IR and the other one—does not. The resulting perfect IR is shown as the upper sequence in figure 1*A*.

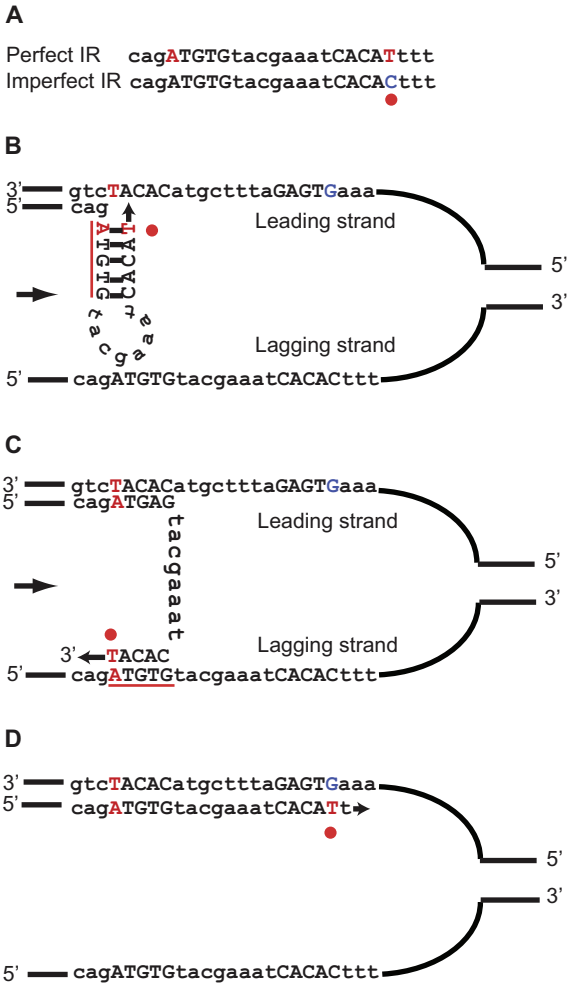

Fig. 1.—Template switching converts an imperfect IR to a perfect one. (*A*) The upper and lower sequences represent a perfect and an imperfect IR, respectively, located in an orthologous locus in two genomes. (*B*) The first switch under intramolecular template switching. Here the nascent strand is used as template. (*C*) The first switch under intermolecular template switching. Here the strand across the fork is used as template. (*D*) The second switch returns the nascent strand into the original template, resulting in a perfect IR as represented by the upper sequence in A. Upper case letters represent the IR arms while red dots represent mismatches between the arms. The noncanonical template is marked with a red line. The direction of the replication fork is indicated with an arrow.

Template switching is known to occur in numerous organisms including bacteriophage T4 (Ripley 1982), *E. coli* (Rosche et al. 1997; Viswanathan et al. 2000), *S. cerevisiae* (Hampsey et al. 1988), as well as human (Greenblatt et al. 1996; Bissler 1998). Template switching between IR arms was reported to be associated with mutation hotspots in the *thyA* and *rpsL* genes in bacteria (Mo et al. 1991; Viswanathan et al. 2000;

Yoshiyama et al. 2001) and in numerous genetic diseases, such as thromboembolism, osteogenesis-related disease, familial hypercholesteremia, and Duchenne muscular dystrophy (Bissler 1998).

Factors affecting template switching mutations include the directionality of the replication fork (Yoshiyama et al. 2001; Kim et al. 2013), the DNA strand (Seier et al. 2011), and the local sequence context (Schultz and Drake 2008). In yeast, quasi-palindrome to palindrome correction is enriched in highly transcribed regions (Kim et al. 2013). Fork replication problems as well as DNA repair proteins affect the rate of IR template switching. For example, the loss of the yeast *RAD27* gene, a key player in Okazaki fragment maturation, increases the rate of template switching (Omer et al. 2017). Moreover, template switching of IRs is enriched in the absence of nucleotide excision repair genes and is dependent on translesion synthesis DNA polymerases (Kim et al. 2013).

Previous studies analyzed short IRs in a genomic context (Cox and Mirkin 1997; van Noort, Worning, et al. 2003; Lisnić et al. 2005; Ladoukakis and Eyre-Walker 2008). These studies suggested that overrepresentation of short-perfect IRs in genomes stems from template switching. The studies mentioned above analyzed a single genome at a time, and hence the importance of template switching in the context of evolution was not directly evaluated. The analysis of orthologous IRs in closely related genomes, through a comparative evolutionary approach, is currently lacking.

Here, we aimed to characterize IR dynamics from an evolutionary perspective in order to understand their long-term impact on genomes. To this end, we examined the abundance of perfect IRs and tested their evolutionary conservation in the context of 20 proteobacteria genomes.

Materials and Methods

Genome Sequences

All proteobacterial genomes reported in this study were obtained from GenBank (*Actinobacillus equuli* NZ_CP007715.1; *Citrobacter amalonaticus* NZ_CP011132.1; *Citrobacter freundii* NZ_CP012554.1; *Enterobacter aerogenes* NZ_CP011574.1; *E. aerogenes* NZ_CP011539.1; *E. aerogenes* NZ_CP011539.1; *Enterobacter lignolyticus* NZ_CP012871.1; *Escherichia albertii* NZ_CP007025; *E. coli* K-12 substr MG1655 U00096; *Escherichia fergusonii* NC_011740.1; *Klebsiella variicola* NZ_CP012871.1; *K. variicola* NC_013850.1; *Kluyvera intermedia* NZ_CP011602.1; *Kosakonia radicincitans* NZ_CP015113.1; *Kosakonia sacchari* NZ_CP016337.1; *Leclercia adecarboxylata* NZ_CP013990.1; *Lelliottia amnigena* NZ_CP015774.1; *Raoultella ornithinolytica* NC_021066.1; *Salmonella enterica* NZ_LN868943.1; *Serratia fonticola* NZ_CP011254.1; *Shigella sonnei* NC_007384).

Genome Partition

Noncoding (NC) regions of the *E. coli* K-12 substr MG1655 genome were determined according to the GenBank annotation. NC regions shorter than 10 bp were filtered out. This resulted in 3,436 NC sequence regions.

Single Genome Analyses

IRs Detection

IRs were detected in NC sequence regions of *E. coli* K-12 substr MG1655 as well as in simulated data sets (see below). We searched for IRs with an arm length of at least 5 bp using the EMBOSS palindrome package (Rice et al. 2000). In our search, we allowed a spacer of up to 70 bp between the two IR arms. This spacer was chosen since the probability of forming a cruciform with a longer spacer is small (Schroth and Ho 1995).

Sequence Simulation from the Null Model

Simulated data sets were generated according to two different algorithms. The first algorithm is based on the Fisher–Yates method (Fisher and Yates 1948), which takes the content of the input sequence and shuffles it to obtain a set of permutated sequences. The second algorithm is based on a second-order Markov chain. The parameters of the chain (base frequencies and transition probabilities, e.g., the probability to observe 'A' after the dinucleotide 'AC') were learnt from the concatenated NC sequence regions of *E. coli* K-12 substr MG1655. These parameters were used to generate random sequences. The second algorithm preserves the trinucleotide composition of the input sequence. For each algorithm, a data set of 100 simulated NC genomes (each simulation contained 3,436 NC sequence regions) were generated using a Perl script.

Statistical Test of Enrichment

In order to test whether the IR lengths observed in the real data come from the same distribution of IR lengths of the null data sets, we performed a likelihood ratio test. Let N (M) denote the total number of IRs across all real (simulated) NC regions. Let x_i (y_i) denote the total number of IRs of length i across all NC regions in the real data set (simulated data sets). Due to small sample sizes, all IRs of length ≥ 13 were counted in x_{13} (y_{13}). We computed the proportion of IRs of length i in the real (simulated) collection as $p_i = \frac{x_i}{N}$ $\left(q_i = \frac{y_i}{M}\right)$. Thus, a total of $k = 9$ length categories were considered ($i = 5, \ldots, 13$). The set of p_i values represents the maximum likelihood estimate for the proportions of each length-category in an alternative model (i.e., the real data set comes from its own distribution). The set of q_i values represents the parameters for the proportions of each length-category under the null model. Either set of parameters can be used to compute the probability of observing the real data $D = (x_5, \ldots, x_{13})$ using a

multinomial distribution. For computational reasons, the probability of observing the data under either model was first approximated using a multivariate normal distribution and then a likelihood ratio test was performed to select between the models.

Evolutionary Analysis

Sequence Segment Definitions and Orthologs Detection

The MG1655 genome contains 4,288 coding sequences (CDSs). We denote each CDS and the NC region downstream of it as "CDS–NC unit." For each such CDS–NC unit, we added the CDS–NC unit immediately upstream and downstream of it. The resulting segment contained three CDSs and three NCs. We consider the middle CDS–NC unit as the focus. Such a three-unit segment was constructed for each identified CDS–NC unit, thus each CDS–NC unit served as the focus in its turn. Hereon, we denote each such resulting segment of three units as "reference query." We have chosen to work with such a reference query to decrease the potential of erroneous classification of regions as orthologous due to horizontal gene transfer. Of note, each reference query has a middle NC region. We retained only reference queries for which the middle NC region contained at least one IR (2,320 out of 3,436). We then used each of the retained reference queries as input for a BLASTN search (BLAST+, Camacho et al. 2009) against the 20 proteobacterial genomes. The BLAST search was limited to hits with an *E*-value ≤ 0.001. For each such reference query, we chose the best hit in each of the 20 genomes, conditioning that this hit was similar in length to the reference query (at least 70% but no more than 130% in length of the reference query). If no hit met these conditions—no ortholog from that species was collected. Finally, we retained only orthologous sets with at least 10 orthologs. A total of 605 such orthologous sets were retained.

Aligning Orthologous Sequences and Tree Reconstruction

Each of the 605 orthologous sets was aligned using MAFFT V3.705 (Katoh et al. 2009) with default parameters for nucleotide alignment. The three CDS blocks were removed from the resulting multiple sequence alignment (MSA). Each resulting MSA thus contained only NC regions. Only MSAs of length >150 bp were retained for further analysis (234 out of 605). Each such MSA was given as input to PhyML V3.1 (Guindon et al . 2010 broken ref.) to reconstruct the maximum likelihood tree. The model of nucleotide substitution for the PhyML reconstruction was HKY (Hasegawa et al. 1985). Rate variation among sites was modeled by a discrete gamma distribution with four rate categories (Yang 1994).

Rooting Phylogenetic Trees

We generated rooted versions for each phylogenetic tree by first turning MG1655 into an internal node by adding a leaf below it with a branch of length 0. Next, the modified tree was rooted on MG1655 using the R package ape (Paradis et al. 2004).

Control Sequences

In order to compare conservation levels between an aligned IR segment and an aligned segment without an IR, we searched for IR-less alignment segments of the same length as the IR in the same orthologous set described above. We could identify such control regions for 207 NC regions.

Conservation Score

We defined an IR conservation score as the ratio between the number of species that displayed the exact IR sequence as the root (including the root) and the total number of species in the phylogenetic tree. The conservation score is defined as the average IR conservation score in a given NC focus region. The same computation was repeated with the control sequences replacing the IRs. This resulted in 234 conservation scores for IR regions and 207 conservation scores for control regions.

Sequence Simulations

Sequences were simulated along rooted phylogenetic trees using the module Simseq of the R package phangorn V2.2.0 (Schliep 2011). Each IR and each control sequence were simulated according to the phylogenetic tree using the inferred evolutionary model parameters in the region in which they reside. In each simulation, the starting root sequence was set to be the MG1655 sequence. For each IR and for each control sequence, we performed 1,000 simulations. IRs and control sequences in the same region were grouped together, thus obtaining a collection of 1,000 artificial copies for each region. In total, we computed 234,000 simulated conservation scores for IR regions and 207,000 simulated conservation scores for control regions.

Results

Characterization of IRs within the *E. coli* str. K-12 substr. MG1655 Genome

We first examined the prevalence and features of IRs in the *E. coli* str. K-12 substr. MG1655 genome (Blattner et al. 1997). The annotated genome is 4,639,221 bp in length and contains 4,288 CDSs. We identified 233,730 short IRs in coding regions and 27,678 short IRs in Noncoding (NC) regions (from here on "the NC collection"). All these short IRs had an arm length shorter than 20 bp and a spacer of up to 70 bp (see section Materials and Methods). Below, we focused on NC IRs as the selection and mutation processes on these IRs are not masked by other selective forces to preserve protein function.

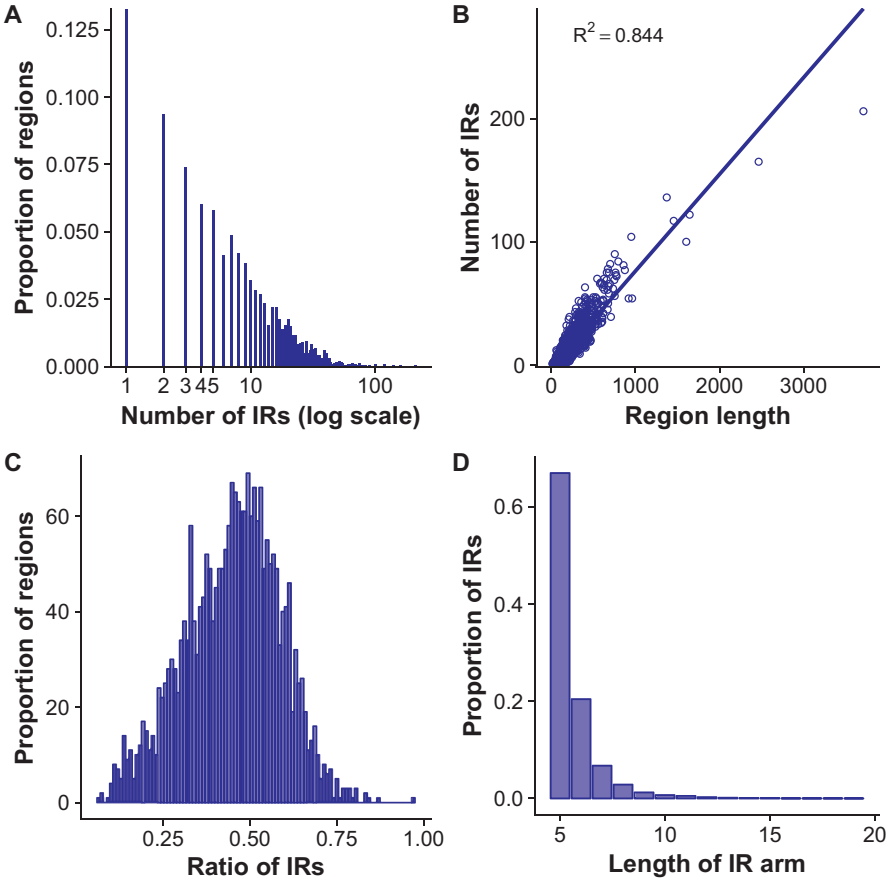

Fig. 2.—Characteristics of detected IRs in the MG1655 genome. (*A*) The proportion of regions in the NC collection for which a certain number of IRs was detected. (*B*) The number of detected IRs in a region as a function of the length of the region. (*C*) The proportion of regions in a collection for which a certain ratio of IR base-pairs to total number of region base-pairs was detected. (*D*) The proportion of IRs for which a certain arm length was detected.

For the total of 27,678 IRs detected in the MG1655 NC collection, we found an average of nine IRs per NC region. As shown in figure 2*A*, most NC regions include few IRs and only a small fraction of NC regions includes many IRs. We next accounted for the variation in the length of the NC regions by dividing either the number of detected IRs or the number of IR base-pairs by the total number of base-pairs in a given region. We found that the number of detected IRs increases linearly with the length of the region ($R^2 = 0.84$, $P < 2e{-}15$, fig. 2*B*) and an average ratio of 0.44 ± 0.13 of IR bases per NC region (fig. 2*C*). We then examined the distribution of arm lengths across all detected IRs. We found an average arm length of 5.75 ± 1.16 bp per IR and an inverse relationship between the arm length and its prevalence across IRs (fig. 2*D*).

Next, we aimed to determine whether the prevalence of IRs and their length distribution in the MG1655 genome are different from what one could expect when no special mechanism for IR creation or repair exists. Without such a special mechanism, IR regions are expected to evolve under the same mutation-selection regime as other NC regions (our null model). To this end, we first simulated sequence regions under these null conditions (see section Materials and Methods). Specifically, for each NC region in our data, we simulated 100 corresponding null regions. This yielded a total of 100 simulated null collections, each of which corresponds to the entire MG1655 collection. We then measured the prevalence and lengths of IRs in each simulated null collection in an identical manner to the characterization of the real MG1655 data set.

When simulating the null collections, we examined two models: the Fisher and Yates (Fisher and Yates 1948) algorithm and a second-order Markov chain (see full details in the Materials and Methods section). The first algorithm offers a simple shuffle of the nucleotide base-pairs. The second model takes into account the trinucleotide composition of the region. The total number of IRs in the MG1655 NC collection was significantly greater than the total number of IRs seen in each null collection. This was true for both simulation schemes (Fisher–Yates and second-order Markov chain; empirical $P < 0.01$, fig. 3). Furthermore, we evaluated each region in the MG1655 collection individually by computing the proportion of null regions that contained an equal to or greater than

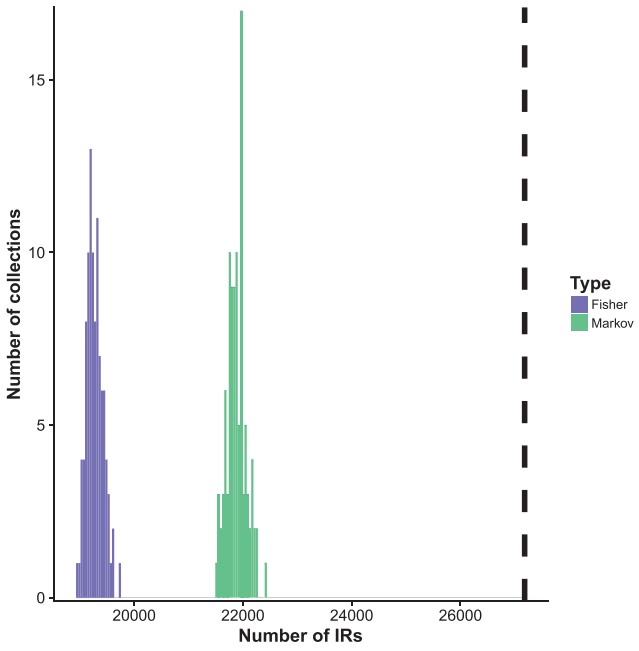

Fig. 3.—Detected IRs in the entire MG1655 genome compared to simulations. The total number of IRs in the MG1655 NC regions (bold dashed line) compared to the total number of IRs in each of its corresponding null collections.

number of IRs as detected for the real region. This proportion served as an empirical *P*-value for that region. Specifically, under null conditions one could expect a uniform distribution of *P*-values. However, when examining the distribution of the obtained empirical *P*-values, we noticed a strong inflation of small *P*-values (see supplementary fig. S1*A* and *B* for the two models, Supplementary Material online). This deviation is statistically significant (Kolmogorov–Smirnov test against a uniform distribution, $P < 2e{-}7$ in all cases). This suggests that null conditions, in which no mechanism exists to create or repair IRs are unlikely to explain the abundance of IRs observed in the MG1655 genome.

We have noticed that the length of the IR arm is an important factor in determining its deviation from the null expectation. Specifically, we first computed the proportion of IRs with arm length *i* in the MG1655 collection. We then computed these proportions in the null simulated collections. Next, a likelihood ratio test was performed to test whether the counts observed for the IR arm length categories in the real collection were likely to originate from the same distribution of IR arm lengths computed from the null simulations. We found that the IR arm length distribution of the MG1655 NC collection was significantly different from that computed from the null simulations using either the Fisher or the second-order Markov algorithm (likelihood ratio test with eight degrees of freedom; $P < 1e{-}40$ in both cases). We have further analyzed the deviation from expectation as a function of the IR arm length. We found that

Table 1

The Observed/Expected Ratio for all IR Arm Length Categories

	IR Arm Length Category								
	5	6	7	8	9	10	11	12	≥13
Fisher	0.89	1.07	1.39	2.3	3.88	8.26	24.64	49.08	152.86
Markov	0.91	1.04	1.29	2.04	3.36	7.12	20.44	38.2	148.86

Note.—Expected values were inferred using the Fisher and second-order Markov chain algorithms.

longer IRs are enriched in the MG1655 genome for all arms longer than 5 bp (table 1).

Evolutionary Conservation of IRs in Proteobacteria

The above analysis suggests that perfect IRs are more abundant in a given genome than randomly expected. An important question follows this finding: are IRs conserved throughout evolution? To test the conservation of IRs in a broader evolutionary context, we collected orthologous sequence units for the MG1655 genome across 20 additional proteobacteria (see section Materials and Methods). Out of 27,678 IRs in MG1655 we were able to identify 914 IRs that appear in at least 10 species. These orthologs reside in 234 NC regions and were further analyzed. For each of these 234 regions, we computed a conservation score (see fig. 4*A* and *B* and Materials and Methods for full details).

Next, we aimed to compare the 234 conservation scores computed from the real NC regions to the expectation under null conditions, in which IRs evolve similarly to any other NC sequence. To this end, we generated 1,000 corresponding simulated data sets for each of the 234 real NC regions. For each such NC region, the empirical distribution of the 1,000 conservation scores served as a null distribution to which we compared the conservation score of the real region. An example of one NC region, which is more conserved than the null distribution, is shown in figure 4*C*.

Out of the 234 NC regions, 145 were found to have a significantly higher conservation score than the expectation under null conditions (empirical $P < 0.05$, fig. 5). Generally, the distribution of the 234 empirical *P*-values deviates from a uniform distribution with a strong inflation of small *P*-values (fig. 5). The high IR conservation level reported above could potentially result from a purifying selection regime acting on the NC region where the IRs are located. In order to test whether the elevated sequence conservation is specific to IRs or general to their location on the genome, we analyzed control sequences adjacent to the IRs. A lower conservation level of control sequences versus the IRs would suggest that IRs are more conserved than their surrounding sequences. We computed 234 conservation scores for IR regions and 207 control region conservation scores. Next, using simulations (see Materials and Methods), we computed empirical *P*-values

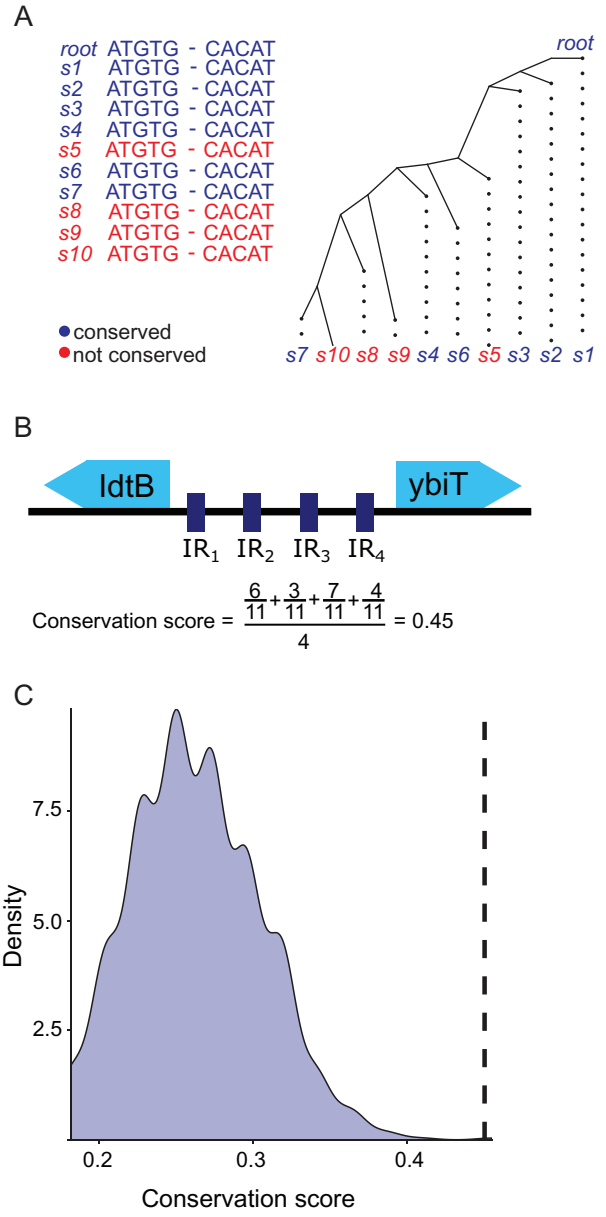

Fig. 4.—Conservation analysis. (*A*) An example alignment of an IR and its mapping onto its corresponding phylogenetic tree, with the IR of the MG1655 as the root sequence. The IR conservation score is 7/11, since 7 out of 11 sequences are identical to the root IR. (*B*) Conservation score computation for the entire NC region, located between the *ldtB* and the *yblT* genes in the MG1655 genome, which contains three additional IRs. (*C*) Analysis of conservation of IRs in the region located between the *ldtB* and the *yblT* genes. The distribution of 1,000 conservation scores computed using simulated data is shown in blue and the conservation score value computed from real data is shown as a bold dashed line. The detailed template switching mechanism that can explain this example is shown in fig 1.

for each such region, obtaining a total of 234 *P*-values for IR regions and 207 *P*-values for control regions. We found these distributions to be significantly different, with the IR conservation *P*-value distribution having a heavier left tail

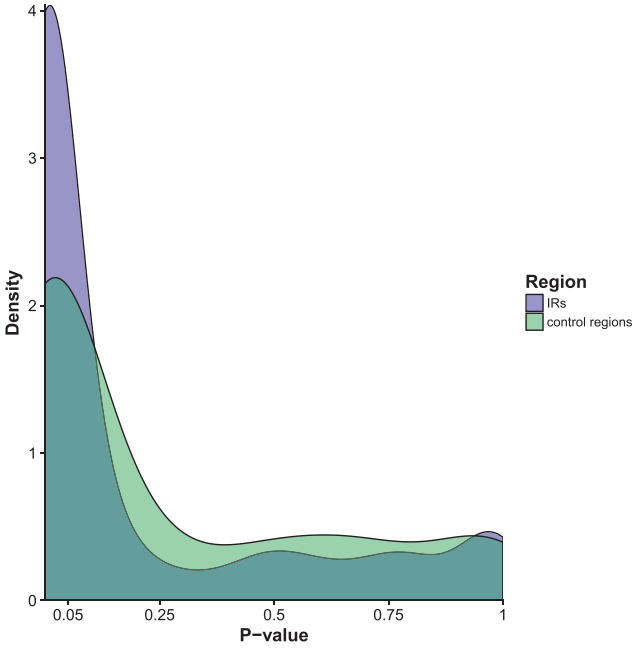

Fig. 5.—Comparison of conservation significance between IR regions and control regions. An empirical *P*-value was computed for each region based on its 1,000 corresponding simulated data sets. Shown in blue are the *P*-values for the IR regions and in green for the control regions.

(Kolmogorov–Smirnov test, $P < 0.028$, fig. 5). These results indicate that the selective forces acting on IRs differ from their adjacent regions, suggesting the operation of a special mechanism to conserve IRs.

Discussion

Theoretical studies suggested that genomes are enriched with short perfect IRs. This has been observed in individual genomes across the tree of life (Smith et al. 1995; Lisnić et al. 2005; Ladoukakis and Eyre-Walker 2008). The abundance of IRs in genomes implies a mechanism, acting to maintain perfect IRs in genomes. The mechanism of template switching that was reported in experimental systems to homogenize IR arms during replication is the best known mechanism to explain the enrichment of perfect IRs in genomes (Lovett 2017).

Here, we chose to focus on *E. coli*, in which the template switching mechanism was studied the most. Considering the single *E. coli* MG1655 genome, we first validated that perfect IRs are more abundant in *E. coli* than expected by chance. The added value of our analysis of a single genome is the findings concerning the arm length. Our results clearly indicate that the deviation from the null expectation increases as a function of the arm length (table 1). This deviation can be explained as longer IRs are more likely to form secondary structures, which promote template switching (Strawbridge et al. 2010).

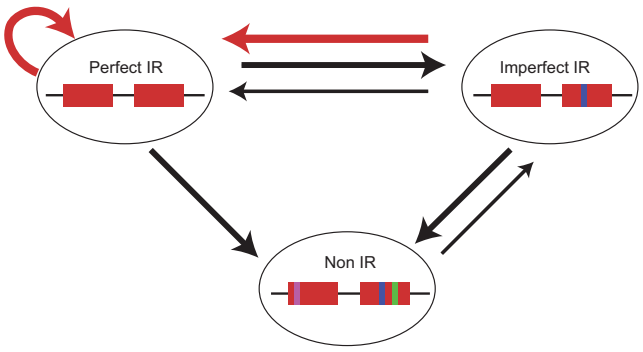

Fig. 6.—Two main processes (template switching in red and the formation of substitutions and indels in black) are dictating the dynamics of short IRs. In this scheme there are three sequence states and the transition between them occurs due to the two processes. Black arrows indicate substitutions and indels, red arrows indicate template switching. The width of the arrows indicates the rate of each process.

The main contribution of this study is the analysis of IRs using an evolutionary genomics approach. If template switching occurs in genomes then we expect it to contribute to the correction of imperfect IRs to perfect ones, resulting in the conservation of IRs through evolution. This hypothesis is in line with the observed overabundance of perfect IRs when analyzing single genomes. However, the conservation of IRs was hitherto never demonstrated in a comparative evolutionary manner and here we have shown it for proteobacterial data. Specifically, the rejection of the null hypothesis, according to which there is no special mechanism to retain and preserve IRs, supports the model of IR arm corrections by template switching through evolution.

We propose an illustrative model of IRs evolutionary dynamics (fig. 6). The model consists of two major evolutionary forces: substitution and small insertion–deletion (indel) events and a correction mechanism (template switching). The contribution of the correction mechanism likely depends on the arm and spacer lengths. Changes in these lengths should affect secondary structures of DNA, and hence the rate of template switching. Although we have shown the evolutionary conservation of IRs only in proteobacteria, given the prevalence of template switching in other organisms (Dutra and Lovett 2006; Löytynoja and Goldman 2017) we speculate that the model holds across large sections of the tree of life.

Much research effort was previously invested in accurate modeling of base pair substitutions and recently, also in indel events (Aris-Brosou et al. 2012; Ezawa 2016; Levy Karin et al. 2017). In contrast, very little effort was invested in modeling the evolutionary dynamics of IRs. Of note, a model that accounts for IR evolutionary dynamics and template switching in particular between pairs of closely related genomes was recently presented by Löytynoja and Goldman (2017). Their work and ours should ideally be extended to a generative model for sequence evolution, in which IR corrections and generations are explicitly accounted for. In the same vein, IRs are currently ignored in most evolutionary analyses, including alignment algorithms, phylogenetic tree inference, ancestral sequence reconstruction, and molecular dating. In such cases, differences between orthologous IRs are considered as independent events, which is clearly not the case. The impact of ignoring IRs in such evolutionary analyses remains to be studied.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

E.L.K. is a fellow of the Edmond J. Safra Center for Bioinformatics at Tel Aviv University. T.P. is supported by the Israel Science Foundation grant 802/16. E.H.C. was supported by the Israel Cancer Association grant 20150038 and by the Open University of Israel Research fund.

Literature Cited

Aris-Brosou S, Rodrigue N, Anisimova M. 2012. The essentials of computational molecular evolution. Methods Mol Biol. 855:111–152.
Bissler JJ. 1998. DNA inverted repeats and human disease. Front Biosci. 3(4):d408–d418.
Blattner FR, et al. 1997. The complete genome sequence of *Escherichia coli* K-12. Science 277(5331):1453–1462.
Branzei D, Foiani M. 2010. Leaping forks at inverted repeats. Genes Dev. 24(1):5–9.
Brewer BJ, Payen C, Raghuraman MK, Dunham MJ. 2011. Origin-dependent inverted-repeat amplification: a replication-based model for generating palindromic amplicons. PLoS Genet. 7(3):e1002016.
Butler DK, Yasuda LE, Yao M. 1996. Induction of large DNA palindrome formation in yeast: implications for gene amplification and genome stability in eukaryotes. Cell 87(6):1115–1122.
Camacho C, et al. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.
Chen J-M, Cooper DN, Chuzhanova N, Férec C, Patrinos GP. 2007. Gene conversion: mechanisms, evolution and human disease. Nat Rev Genet. 8(10):762–775.
Cox R, Mirkin SM. 1997. Characteristic enrichment of DNA repeats in different genomes. Genetics 94(10):5237–5242.
Cuomo CA, Mundy CL, Oettinger MA. 1996. DNA sequence and structure requirements for cleavage of V(D)J recombination signal sequences. Mol Cell Biol. 16(10):5683–5690.
Darmon E, Leach DRF. 2014. Bacterial genome instability. Microbiol Mol Biol Rev. 78(1):1–39.
Dutra BE, Lovett ST. 2006. Cis and trans-acting effects on a mutational hotspot involving a replication template switch. J Mol Biol. 356(2):300–311.
Ezawa K. 2016. General continuous-time Markov model of sequence evolution via insertions/deletions: are alignment probabilities factorable?. BMC Bioinformatics 17(1):397.
Fisher RA, Yates F. 1948. Statistical tables for biological, agricultural and medical research. London: Oliver and Boyd. p. 37–39.
Gordenin DA, et al. 1993. Inverted DNA repeat: a source of eukaryotic genomic instability. Mol Cell Biol. 13(9):5315–5322.

Grant DM, Gillham NW, Boynton JE. 1980. Inheritance of chloroplast DNA in *Chlamydomonas reinhardtii*. Proc Natl Acad Sci U S A 77(10):6067–6071.
Greenblatt MS, Grollman AP, Harris CC. 1996. Deletions and insertions in the p53 tumor suppressor gene in human cancers: confirmation of the DNA polymerase slippage/misalignment model. Cancer Res. 56(9):2130–2136.
Guindon S, et al. 2010. New algorithms and methods in PHYML 3.0. Syst Biol. 59:307–321.
Hampsey DM, Ernst JF, Stewart JW, Sherman F. 1988. Multiple base-pair mutations in yeast. J Mol Biol. 201(3):471–486.
Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol. 22(2):160–174.
del Solar G, Giraldo R, Ruiz-Echevarría MJ, Espinosa M, Díaz-Orejas R. 1998. Replication and control of circular bacterial plasmids. Microbiol Mol Biol Rev. 62(2):434–464.
Katoh K, Asimenos G, Toh H. 2009. Multiple alignment of DNA sequences with MAFFT. Methods Mol Biol. 537:39–64.
Kim N, Cho J, Li YC, Jinks-robertson S. 2013. RNA: DNA hybrids initiate quasi-palindrome-associated mutations in highly transcribed yeast DNA. PLoS Genet. 9:e1003924.
Kolodner R, Tewari KK. 1979. Inverted repeats in higher plants. Proc Natl Acad Sci U S A 76(1):41–45.
Ladoukakis ED, Eyre-Walker A. 2008. The excess of small inverted repeats in prokaryotes. J Mol Evol. 67(3):291–300.
Levy Karin E, Shkedy D, Ashkenazy H, Cartwright RA, Pupko T. 2017. Inferring rates and length-distributions of indels using approximate Bayesian computation. Genome Biol Evol. 9(5):1280–1294.
Li X, Lindahl L, Sha Y, Zengel JM. 1997. Analysis of the bacillus subtilis S10 ribosomal protein gene cluster identifies two promoters that may be responsible for transcription of the entire 15-kilobase s10-spc-alpha cluster. J Bacteriol. 179(22):7046–7054.
Lilley DM. 1980. The inverted repeat as a recognizable structural feature in supercoiled DNA molecules. Proc Natl Acad Sci U S A 77(11):6468–6472.
Lillo F, Basile S, Mantegna RN. 2002. Comparative genomics study of inverted repeats in bacteria. Bioinformatics 18(7):971–979.
Lisnić B, Svetec I-K, Sarić H, Nikolić I, Zgaga Z. 2005. Palindrome content of the yeast *Saccharomyces cerevisiae* genome. Curr Genet. 47(5):289–297.
Lobachev KS, Gordenin DA, Resnick MA. 2002. The Mre11 complex is required for repair of hairpin-capped double-strand breaks and prevention of chromosome rearrangements. Cell 108(2):183–193.
Lobachev KS, et al. 2000. Inverted Alu repeats unstable in yeast are excluded from the human genome. EMBO J. 19(14):3822–3830.,
Lovett ST. 2017. Template-switching during replication fork repair in bacteria. DNA Repair (Amst) 56:118–128.
Löytynoja A, Goldman N. 2017. Short template switch events explain mutation clusters in the human genome. Genome Res. 27(6):1039–1049.
Lu S, et al. 2015. Short inverted repeats are hotspots for genetic instability: relevance to cancer genomes. Cell Rep. 10(10):1674–1680.
Mizuno KI, Lambert S, Baldacci G, Murray JM, Carr AM. 2009. Nearby inverted repeats fuse to generate acentric and dicentric palindromic chromosomes by a replication template exchange mechanism. Genes Dev. 23(24):2876–2886.
Mo JY, Maki H, Sekiguchi M. 1991. Mutational specificity of the dnaE173 mutator associated with a defect in the catalytic subunit of DNA polymerase III of *Escherichia coli*. J Mol Biol. 222(4):925–936.
Narayanan V, Mieczkowski PA, Kim H, Petes TD, Lobachev KS. 2006. The pattern of gene amplification is determined by the chromosomal location of hairpin-capped breaks. Cell 125(7):1283–1296.
van Noort V, Snel B, Huynen MA. 2003. Predicting gene function by conserved co-expression. Trends Genet. 19(5):238–242.
van Noort V, Worning P, Ussery DW, Rosche WA, Sinden RR. 2003. Strand misalignments lead to quasipalindrome correction. Trends Genet. 19(7):365–369.
Okamura K, Chung W-J, Lai EC. 2008. The long and short of inverted repeat genes in animals: microRNAs, mirtrons and hairpin RNAs. Cell Cycle 7(18):2840–2845.
Omer S, Lavi B, Mieczkowski PA, Covo S, Hazkani-Covo E. 2017. Whole-genome sequence analysis of mutations accumulated in *rad27* Δ yeast strains with defects in the processing of okazaki fragments indicates template-switching events. G3 7:3775–3787.
Paek AL, et al. 2009. Fusion of nearby inverted repeats by a replication-based mechanism leads to formation of dicentric and acentric chromosomes that cause genome instability in budding yeast. Genes Dev. 23(24):2861–2875.
Panayotatos N, Wells RD. 1981. Cruciform structures in supercoiled DNA. Nature 289(5797):466–470.
Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20(2):289–290.
Pearson CE, Zorbas H, Price GB, Zannis-Hadjopoulos M. 1996. Inverted repeats, stem-loops, and cruciforms: significance for initiation of DNA replication. J Cell Biochem. 63(1):1–22.
Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European molecular biology open software suite. Trends Genet. 16(6):276–277.
Richard G-F, Kerrest A, Dujon B. 2008. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol Mol Biol Rev. 72(4):686–727.
Ripley LS. 1982. Model for the participation of quasi-palindromic DNA sequences in frameshift mutation. Proc Natl Acad Sci U S A 79(13):4128–4132.
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.
Sadler JR, Sasmor H, Betzt JL. 1983. A perfectly symmetric lac operator binds the lac repressor very tightly. Biochemistry 80(22):6785–6789.
Schliep KP. 2011. phangorn: phylogenetic analysis in R. Bioinformatics 27(4):592–593.
Schroth GP, Ho PS. 1995. Occurrence of potential cruciform and forming sequences in genomic DNA H-DNA. Nucleic Acids Res. 23(11):1977–1983.
Schultz GE, Drake JW. 2008. Templated mutagenesis in bacteriophage T4 involving imperfect direct or indirect sequence repeats. Genetics 178(2):661–673.
Seier T, et al. 2011. Insights into mutagenesis using *Escherichia coli* chromosomal lacZ strains that enable detection of a wide spectrum of mutational events. Genetics 188(2):247–262.
Smith H, Tomb J, Dougherty B, Fleischmann R, Venter J. 1995. Frequency and distribution of DNA uptake signal sequences in the *Haemophilus influenzae* Rd genome. Science 269:538–540.
Strawbridge EM, Benson G, Gelfand Y, Benham CJ. 2010. The distribution of inverted repeat sequences in the *Saccharomyces cerevisiae* genome. Curr Genet. 56(4):321–340.
Tanaka H, Bergstrom DA, Yao M-C, Tapscott SJ. 2005. Widespread and nonrandom distribution of DNA palindromes in cancer cells provides a structural platform for subsequent gene amplification. Nat Genet. 37(3):320–327.
Tanaka H, Tapscott SJ, Trask BJ, Yao M-C. 2002. Short inverted repeats initiate gene amplification through the formation of a large DNA palindrome in mammalian cells. Proc Natl Acad Sci U S A 99(13):8772–8777.
Tanaka H, Yao M-C. 2009. Palindromic gene amplification—an evolutionarily conserved role for DNA inverted repeats in the genome. Nat Rev Cancer 9(3):216–3126.

Turmel M, Otis C, Lemieux C. 2017. Divergent copies of the large inverted repeat in the chloroplast genomes of ulvophycean green algae. Sci Rep. 7(1):1–15.
Viswanathan M, Lacirignola JJ, Hurley RL, Lovett ST. 2000. A novel mutational hotspot in a natural quasipalindrome in *Escherichia coli*. J Mol Biol. 302(3):553–564.
Voineagu I, Narayanan V, Lobachev KS, Mirkin SM. 2008. Replication stalling at unstable inverted repeats: interplay between DNA hairpins and fork stabilizing proteins. Proc Natl Acad Sci U S A 105(29):9936–9941.
Warburton PE, Giordano J, Cheung F, Gelfand Y, Benson G. 2004. Inverted repeat structure of the human genome: the X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes. Genome Res. 14(10A):1861–1869.
Wroblewski T, et al. 2014. Distinctive profiles of small RNA couple inverted repeat-induced post-transcriptional gene silencing with endogenous RNA silencing pathways in *Arabidopsis*. RNA 20(12):1987–1999.
Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol. 39(3):306–314.
Yoshiyama K, Higuchi K, Matsumura H, Maki H. 2001. Directionality of DNA replication fork movement strongly affects the generation of spontaneous mutations in *Escherichia coli*. J Mol Biol. 307(5):1195.

Associate editor: Ruth Hershberg