Replication-Dependent Organization Constrains Positioning of Long DNA Repeats in Bacterial Genomes

Nitish Malhotra * and Aswin Sai Narain Seshasayee*

National Centre for Biological Sciences, Tata Institute of Fundamental Research, Bellary Road, Bangalore 560065, India

*Corresponding authors: E-mails: nitishmalhotra1411@gmail.com (N.M.); aswin@ncbs.res.in (A.S.N.S.).

Accepted: 27 June 2022

Abstract

Bacterial genome organization is primarily driven by chromosomal replication from a single origin of replication. However, chromosomal rearrangements, which can disrupt such organization, are inevitable in nature. Long DNA repeats are major players mediating rearrangements, large and small, via homologous recombination. Since changes to genome organization affect bacterial fitness—and more so in fast-growing than slow-growing bacteria—and are under selection, it is reasonable to expect that genomic positioning of long DNA repeats is also under selection. To test this, we identified identical DNA repeats of at least 100 base pairs across ~6,000 bacterial genomes and compared their distribution in fast- and slow-growing bacteria. We found that long identical DNA repeats are distributed in a non-random manner across bacterial genomes. Their distribution differs in the overall number, orientation, and proximity to the origin of replication, between fast- and slow-growing bacteria. We show that their positioning—which might arise from a combination of the processes that produce repeats and selection on rearrangements that recombination between repeat elements might cause—permits less disruption to the replication-dependent genome organization of bacteria compared with random suggesting it as a major constraint to positioning of long DNA repeats.

Key words: bacterial genomics, genome evolution, bioinformatics, genome organization, comparative genomics, chromosomal rearrangements.

Significance

Long DNA repeats cause chromosomal rearrangements leading to disruption of bacterial genome organization and thus may be under selection. The questions we address include the following: 1) How are such repeats distributed in the bacterial genome? 2) What are the constraints that define the positions of these repeats in the genome? Our study shows that long DNA repeats are present in a non-random manner across bacterial genomes with their genomic distribution being different in fast- and slow-growing bacteria. We highlight replication-dependent genome organization as one of the major constraints in genomic positioning of repeats in bacteria.

Introduction

Most bacterial genomes consist of a single circular chromosome. Similar to the genomes of all living organisms, the bacterial genome is condensed and organized inside the cell. Despite sharing different strategies of gene organization with eukaryotes (Lawrence 2002), unlike eukaryotes, bacterial genome organization is primarily driven by chromosome replication (Rocha 2004).

Chromosome replication in bacteria begins at a single locus called the origin of replication (*oriC*) and terminates diametrically opposite at the terminus of replication, *ter* (Duggin and Bell 2009). The movement of the replisome

from *oriC* to *ter* is bidirectional, creating replichores on both sides of the *oriC*. The average doubling time of fast-growing bacteria such as *Escherichia coli* is less than the time required to replicate its chromosome (Cooper and Helmstetter 1968). To compensate for the lag between chromosome replication time and division time, a new replication cycle begins at *oriC* even before the previous one ends at *ter*. Consequently, at any point of time during replication, the copy number of regions near *oriC* is higher than those near *ter*, resulting in an *oriC*-*ter* dosage gradient. This dosage gradient can be as high as 8:1 in *E. coli*. Even in slower-growing bacteria in which DNA replication occupies a substantial portion of the cell cycle, a 2:1 dosage gradient between *oriC* and *ter* will be common. As a possible consequence of selection arising from this *oriC*–*ter* gene dosage gradient, highly expressed genes—primarily those coding for the translation machinery—are found proximal to *oriC* whereas stress response genes and horizontally acquired genes are localized proximal to *ter* (Rocha 2004; Couturier and Rocha 2006).

In addition to the dosage gradient observed, the genomic content in bacteria is differentially distributed across strands and replichores formed during replication (Rocha 2004). At a very local scale, nucleotide composition in the leading strand is skewed towards G and T. Additionally there is often a gradient of decreasing G + C content from *oriC* to *ter* (Lobry 1996; Frank and Lobry 1999). The leading strand is also abundant in DNA motifs involved in recombination-mediated repair, namely Chi (Lam et al. 1974; Uno et al. 2000) and chromosomal segregation, namely KOPS (Bigot et al. 2005). Furthermore, the leading strand encodes more genes than the lagging strand and is in particular enriched for essential and highly expressed genes. This is often attributed to the detrimental effect of head-on collisions between the DNA polymerase and the RNA polymerase, which are more likely to happen during the transcription of highly expressed genes encoded on the lagging strand (Rocha 2003, 2004, 2008).

Despite organizational features of the genome being shared across bacteria (Tamames 2001; Couturier and Rocha 2006; Khedkar and Seshasayee 2016), the events that generate chromosome variations are inevitable in nature and large rearrangements are one of the major contributors to these variations. Chromosome rearrangements like deletions, duplications, and inversions are caused by cellular processes like homologous recombination and transpositions under different genetic and environmental conditions (Sonti and Roth 1989; Srinivasan et al. 2015; Bishop and Schiestl 2000; Tillier and Collins 2000; Veetil et al. 2020). DNA repeats are one of the major players mediating such events (Bi and Liu 1996; Lambert et al. 1999; Achaz et al. 2003; Treangen et al. 2009). DNA repeats or duplicated stretches of DNA range from dinucleotides to thousands of nucleotides and are abundant in bacterial genomes (Treangen et al. 2009). DNA recombination between repeats results in structural variations and its rate is linearly dependent on substrate length (Shen and Huang 1986; Vulić et al. 1997). "Long" DNA repeats, that is, DNA repeats of the length of at least 100 nucleotides can lead to intrachromosomal rearrangements by acting as substrates to the bacterial recombination machinery (Shen and Huang 1986; Treangen et al. 2009).

The type of rearrangements mediated by long DNA repeats depend on the relative orientation of the repeat pairs. Direct repeats—that is, repeat pairs present in the same orientation—result in duplication or deletion of the genomic region flanked by them. On the contrary, inverted repeat pairs—that is, repeat pairs that are positioned in the genome in opposite orientation—lead to inversions. Deletions lead to the removal of the chromosomal region thus resulting in gene loss whereas duplications lead to doubling of the genomic segment consequently increasing the copy number and presumably the expression levels of the affected genes (Straus and Hoffmann 1975; Sonti and Roth 1989; Skovgaard et al. 2011; Srinivasan et al. 2015). On the contrary, inversions caused by inverted repeat pairs flip the repeat-flanked region thus reversing its orientation and can cause detrimental head-on collisions between the two polymerases (DNA and RNA polymerase), especially at any highly expressed gene in the inverted segment. Large inversions also result in a significant disruption of gene dosage gradient, affecting fitness particularly in conditions supporting fast growth (Srivatsan et al. 2010). Taken together, repeat mediated rearrangements disrupt the bacterial genome organization by altering the dosage and orientation of genes.

Previous studies have reported events of bacterial rearrangements under different stresses. The rearrangements observed were associated with repeat elements like insertion sequences, and were in turn advantageous or disadvantageous in different environments (Sonti and Roth 1989; Maharjan et al. 2013; Adler et al. 2014; Srinivasan et al. 2015; Repar et al. 2017; Veetil et al. 2020). Since these repeat-associated changes in genome organization play a role in affecting fitness, there might be selection on the positioning of such repeats on the chromosome. Studies indicating non-random genomic distribution of repeats or landscapes of chromosomal rearrangements suggest chromosomal composition, relative position with respect to origin of replication, or pathogenicity as constraints on genomic presence of such repeats (Rocha et al. 1999; Repar and Warnecke 2017).

In this study, we used comparative genomics to investigate the association between replication-dependent genome organization and long DNA repeats. Through this work, we asked the following questions on the genomic distribution of long DNA repeats: 1) Are long repeats present randomly across genomes? 2) Does their distribution

reflect selection imposed by the nature of structural variation they mediate? 3) How does their genomic arrangement vary across bacteria with different growth rates? Using ~6,000 bacterial genomes across different genera and classes of bacteria representing fast- and slow-growing bacteria, we found that long identical DNA repeats are distributed non-randomly across bacterial genomes. The genomic distribution of these repeats differs in number, orientation and *oriC*-proximity in fast- and slow-growing bacteria. Repeat pairs are present in such a way that repeat-mediated rearrangements result in less disruptions to overall genome organization than that by random chance. Such an observation is stronger in fast-growing bacteria than in slower-growing ones. Taken together, our study identifies bacterial growth as one of the constraints to genomic organization of long DNA repeats in bacterial genomes.

Results

Repeat Density is Not Correlated with Bacterial Genome Size

The premise of this study is that long DNA repeats, that is, DNA repeats of at least 100 bp (base pairs) length are capable of intrachromosomal recombination leading to a variety of structural variations (Shen and Huang 1986). These variations have the potential to alter bacterial genome organization and can be beneficial or deleterious in different conditions. Consequently, this might impose selection on where such repeats are positioned on the chromosome.

We identified intrachromosomal identical repeats whose repeating units are at least 100 bp in length using the MUMmer (Delcher et al. 2002; Kurtz et al. 2004) package across 6387 bacterial genomes obtained from the NCBI RefSeq database (O'Leary et al. 2016). These genomes represent 795 bacterial species from 462 genera and 27 phyla.

~78% of DNA repeats overlapped coding sequences (CDS) and ~11% of repeats had an overlap with rDNA regions. Additionally, an average of ~60% of the repeats reside in horizontally acquired regions. Insertion Sequence (IS) elements and prophages constitute ~23% and ~2% of the repeats, respectively (fig. 1*a*). Note that these regions can be overlapping with each other and therefore the percentages may not sum up to 100.

rDNA sequences are present in multiple copies and are an essential part of the bacterial translation machinery. These sequences are proximal to *oriC* and are known to show gene dosage in a replication-dependent manner. For these reasons, we specifically removed repeat pairs that had an overlap with these regions (i.e., an average ~11% of repeats) *oriC* to avoid any unwanted bias in our study due to their nonuniform organization. This left us with 6340 bacteria with atleast one repeat and a median of 120 distinct repeats ranging from a minimum of 1 repeat to 9,217 repeats per genomes.

The repeat density, defined by the proportion of the repetitive genome, ranged from ~0.005% in *Advenella* to ~49% in *Orientia* with a median of ~1.5%. This density varied across different classes of bacteria (fig. 1*b*) and did not correlate (Pearson's correlation coefficient; $r = -0.011$, $P = 0.36$, $N = 6{,}340$) with genome size (fig. 1*c*). We observed a similar lack of correlation with genome size even after removing redundancy at the species level (Pearson's correlation coefficient; $r = -0.05$, $P = 0.13$, $N = 795$) (supplementary fig. S1, Supplementary Material online). This is contrary to an earlier study done on a smaller set ($N = 53$) of genomes by (Rocha et al. 1999) which found a negative correlation between repeat density and bacterial genome size.

Bacterial cells spend a substantial portion of their cell cycles replicating the chromosome. In fast-growing organisms in which chromosome replication takes longer than the average population doubling time, there are multiple replication cycles ongoing at any time. This establishes an *oriC–ter* dosage gradient. Even in slow-growing bacteria, there will be a gradient for the proportion of the cell cycle during which replication is going on. However, the *oriC–ter* gradient will be steeper in fast-growing bacteria than in slower-growing ones. To incorporate this effect of growth rate on the replicative structure in our analysis we classified bacteria into fast- or slow-growing organisms and compared the genomic distribution of repeats between them. This classification was based on a quantity called R-factor (Rf), a rough estimate of the *oriC–ter* gradient determined by (Couturier and Rocha 2006), defined as the number of replication initiations per cell cycle. Fast-growing bacteria are defined as those with $Rf > 1$; where the chromosome replication time is greater than the average doubling time, indicating a steep replication-dependent gene dosage gradient. In contrast, slow-growing bacteria have $Rf \leq 1$. It is to be noted that this categorization is a prediction based on 16S rRNA count, its relationship to reported minimum doubling times of a subset of experimentally characterized bacteria (Couturier and Rocha 2006; Khedkar and Seshasayee 2016) and an assumption of uniform replication rate across bacteria. This becomes a necessary simplification considering the lack of availability of data on replication rates across bacteria.

In this study, we identified 2,337 fast-growing and 4,003 slow-growing bacteria with at least one repeat pair in their genomes. In this categorization, fast-growing bacteria span 6 phyla and 16 classes while slow-growing bacteria span 27 phyla and 75 classes of bacteria (Supplementary File available at https://doi.org/10.6084/m9.figshare.19367048).

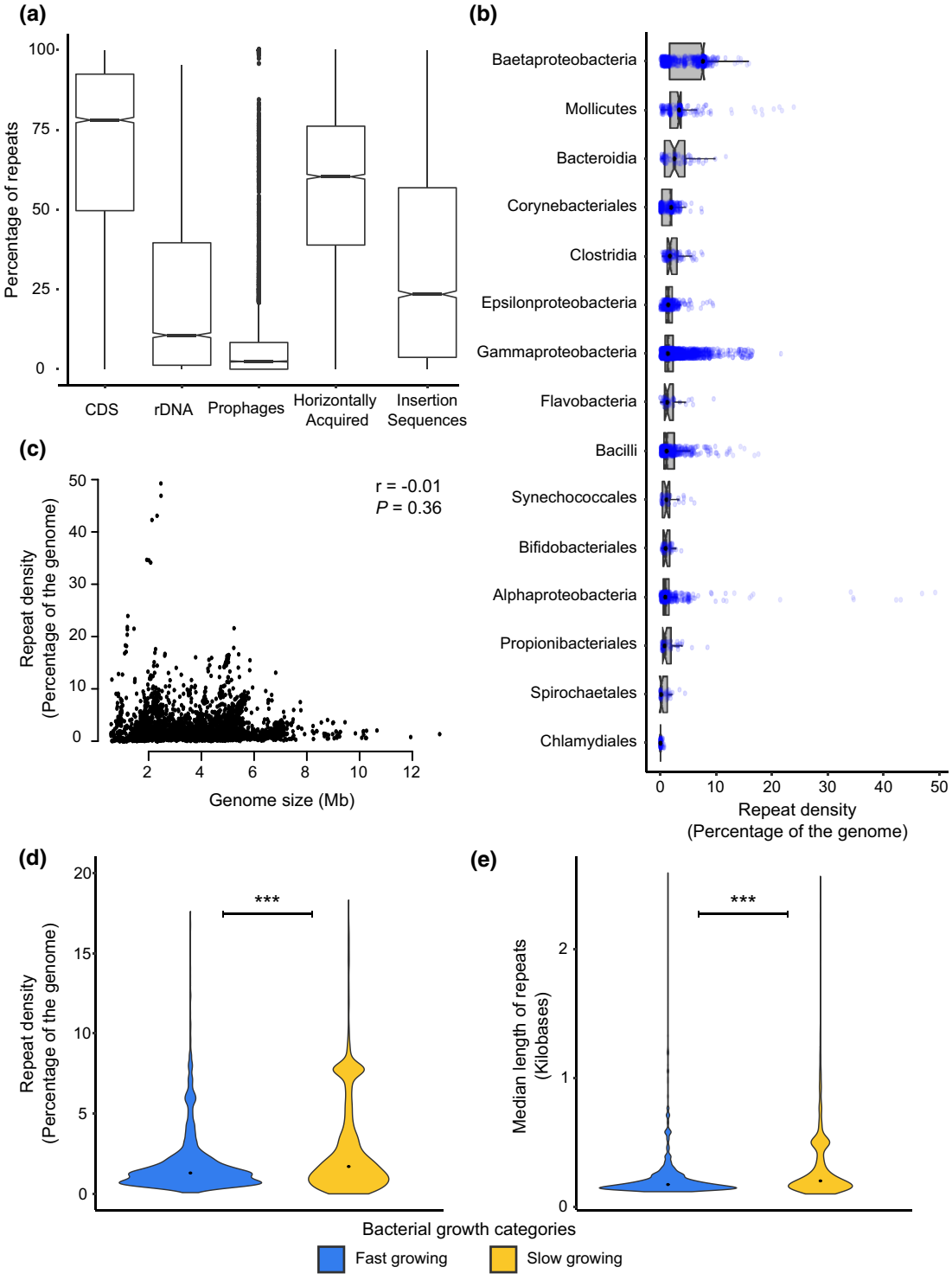

Fig. 1.—Repeat density is not correlated with bacterial genome size or growth. (*a*) Boxplots of percentage of repeats overlapping with annotated CDS, rDNA regions, and predicted prophages (using Phaster), horizontally acquired regions (using Alien Hunter) and insertions sequences (ISEScan). (*b*) Boxplots of repeat densities across classes of bacteria with jitter showing the distribution. Only classes with at least 30 genomes are shown here. (*c*) The scatter plot showing relation between repeat density and genome size (Megabases; Mb). The correlation is calculated by Pearson's correlation using cor.test() function in *R*. *N* (total number of genomes) = 6,340. (*d*) Violin plots showing repeat density distribution in fast-growing and slow-growing bacteria. Wilcoxon's rank-sum test is used to calculate the significance in the differences of two medians. (*e*) Violin plots showing distribution of median length of repeats in fast-growing and slow-growing bacteria. Wilcoxon's rank-sum test is used to calculate the significance in the differences of two medians (n.s.:$P \geq 10^{-2}$, *$P < 10^{-2}$, **$P < 10^{-4}$, ***$P < 10^{-6}$).

We note that the repeat density of fast-growing bacterial genomes is significantly lower than that of slow-growing bacteria (Wilcoxon rank-sum test; $P < 10^{-10}$) (fig. 1*d*). We further observed that the median length of repeats in fast-growing bacteria is also lower than in slow-growing bacteria (Wilcoxon rank-sum test; $P < 10^{-10}$) (fig. 1*e*). These suggest that repeat-mediated recombination events may be selected against, particularly in fast-growing bacteria.

Inverted Repeats are Less Common than Direct Repeats in Bacterial Genomes

Following the identification of long DNA repeats in bacterial genomes, we studied the distribution of repeats responsible for specific kinds of structural variation. The nature of structural variation caused by repeat pairs depends on their relative orientation. Repeat pairs are classified as *direct* if both repeats in a pair are oriented in the same direction or *inverted* if they are oriented opposite to each other. Direct repeats can lead to duplications and deletions whereas inverted repeats can cause inversions of the region flanked by them.

In our dataset, we found that ~79% (4,984/6,340) of genomes have a lower proportion of inverted repeats than direct repeats (i.e., number of inverted repeat pairs/number of direct repeat pairs <1). This is consistent with the earlier report by (Achaz et al. 2003). Similar to their report, we observed that this lower proportion of inverted repeats is more prominent in genomes with a lower number of repeat pairs. This proportion approaches 1 with the increase in repeat pairs (fig. 2*a*). To understand if this observed proportion is within the range expected by random chance, we generated a null distribution by allocating an orientation to each repeat element in a genome randomly. While creating the null distribution, we keep the number of pairs the same as that in the observed set. We further calculated the count of direct and inverted repeat pairs for that genome for 1,000 such random allocations. For every genome, we applied a z-statistic to the observed count of direct and inverted repeats and calculated the *P*-value associated with the count. We found that ~46% (2,903/6,340) of the genomes had a significantly lower number of inverted repeat pairs ($z < 0$, $P \leq 0.01$) while just ~2% (106/6,340) of the genomes had a significantly higher proportion inverted repeat pairs ($z > 0$, $P \leq 0.01$). Slow-growing bacteria (median inverted repeat proportion; 0.62) encode a relatively lower proportion of inverted repeat pairs than fast-growing organisms (median inverted repeat proportion; 0.74) (Wilcoxon rank-sum test; $P < 10^{-10}$) (fig. 2*b*).

The presence of significantly lower inverted repeats suggests that the process by which repeats are generated is more likely to produce direct repeats (see Discussion), especially at lower total repeat counts, and/or selection against inversions in bacterial genomes across genera, the two explanations not being mutually exclusive.

Repeats are Associated with Large Genomic Rearrangements

To test if the role of repeats in causing chromosomal rearrangements can be seen in extant genomes, we analysed inverted repeats and their association with large (>1 kb) chromosomal inversions. We specifically selected inversions for this analysis because large inversions are relatively more stable than large deletions and duplications (Straus and Hoffmann 1975; Adler et al. 2014), and are relatively simpler to detect in the assembled genomes. For each species in our dataset, we randomly chose a reference genome and compared all genomes in that species with the reference using MUMmer (Delcher et al. 2002; Kurtz et al. 2004). We extracted genomic positions of large inversions (≥1 kb) with respect to the reference and then calculated the distance of the nearest repeat element flanking that inverted region in both the reference and the compared genome. We further selected the shortest distance amongst the distances obtained in either the reference or the compared genome and then tested the null hypothesis that the shortest distances between inversions and the nearest inverted repeat element are not significantly different from that obtained if repeats were randomly distributed. The apriori condition for a repeat to be considered was that it should form an inverted pair in the same genome, that is, another instance of the repeat should be present in the same genome but with opposite orientation. The underlying assumption for this analysis is that repeats responsible for the rearrangement in that genome are found in close proximity to the site of inversion. We generated a null model for every bacterial genome by assigning inverted repeat elements to random positions on the genome. To test our hypothesis, we compared distributions of median shortest distances observed in the genomes with the distribution of median distances obtained from 1,000 iterations using a Wilcoxon signed-rank test (paired). Using the null model thus generated and the statistical tests performed, we rejected the null hypothesis ($P < 10^{-10}$) and found that the average distances in the observed dataset are significantly lower than expected had the null hypothesis been true (fig. 2*c*).

Long DNA Repeats are Distributed Non-randomly Across Bacterial Genomes

Having established that repeats are indeed associated with structural variations, we set out to test whether there are any constraints on the genomic positions of such repeats. To test the null hypothesis that repeats are positioned randomly on the chromosome, we considered genomes with at least 30 distinct repeat-containing loci (5,970 genomes).

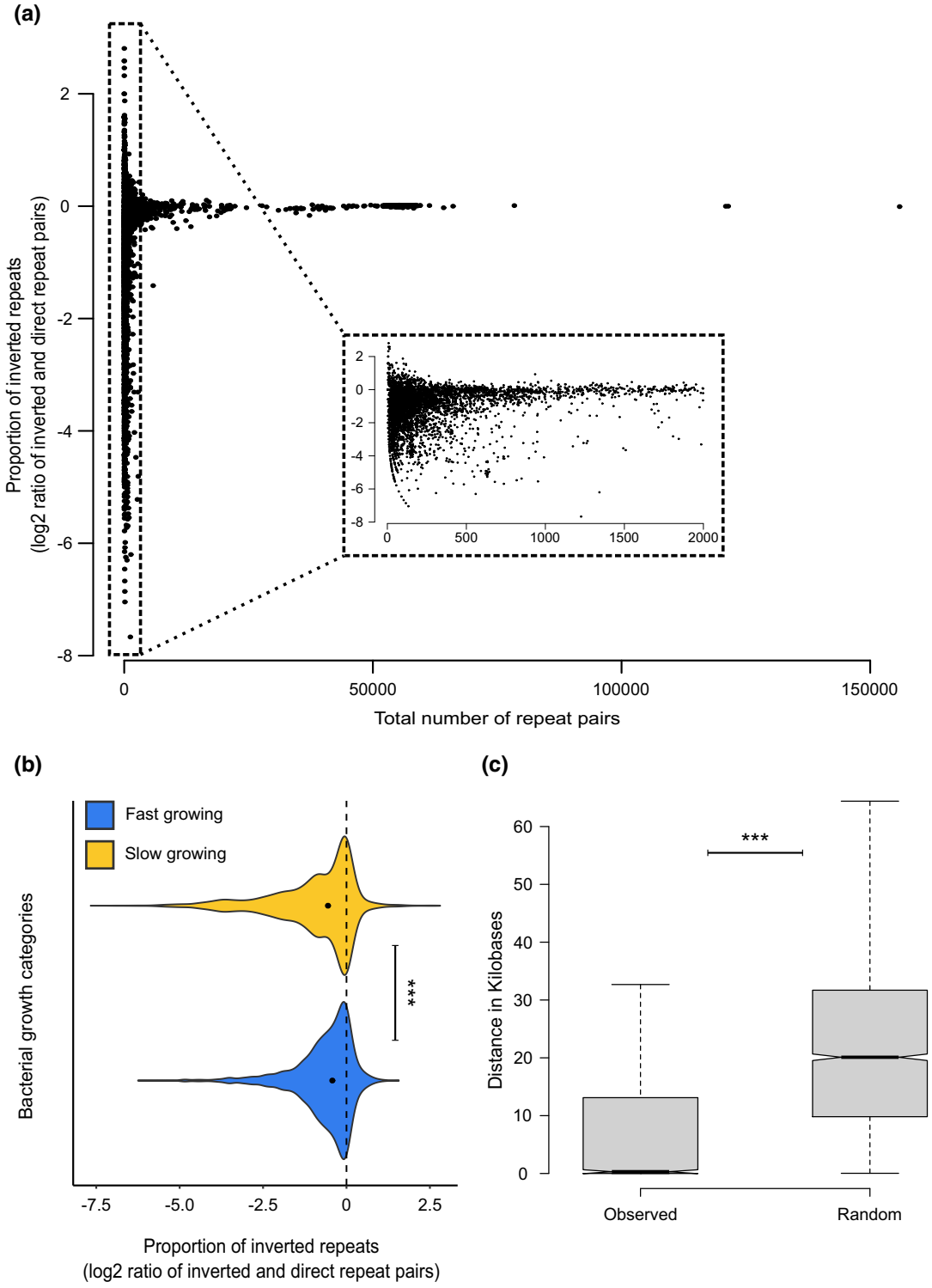

Fig. 2.—Inverted repeats are less common than direct repeats in bacterial genomes. (*a*) The scatter plot showing relation between proportion of inverted repeat pairs shown as log2 (number of inverted repeat pairs/number of direct repeat pairs). The inset shows a zoomed in version from 0 to 2,000 repeat pairs where inverted repeat pairs are under-represented and have not reached closer to that of direct repeat pairs. (*b*) Violin plots showing distribution of log2 proportion of inverted repeat pairs in fast-growing and slow-growing bacteria. Wilcoxon's rank-sum test is used to calculate the significance in the differences of two medians. (*c*) Boxplots of the median of minimum distances (in Kilobases; Kb) of the inverted repeats with the genomic inversion in the query sequence with respect to the reference. For every genome, there is one point in each ofobserved and random set. Wilcoxon's signed-rank test is used to calculate the significance in the differences of two medians (n.s.:$P \geq 10^{-2}$, *$P < 10^{-2}$, **$P < 10^{-4}$, ***$P < 10^{-6}$).

For every genome, we calculated the distances between adjacent repeat elements and compared its distribution with the distribution obtained by randomizing their genomic positions. The comparisons were done using Kolmogorov–Smirnov (KS) two-sample test for 1,000 random iterations followed by the Bonferroni test for multiple corrections resulting in 1,000 *P* values for every genome. A genome was defined as having a non-random distribution of repeats if the distance distribution of repeats showed significant deviation from the random distribution, that is, at least 90% of the iterations of the comparison mentioned have a $P \leq 0.01$. It is to be noted that the stringent *P*-value thresholds used here might compromise the sensitivity in rejecting the null hypothesis. We observed that in ~83% (4,977/5,970) of the bacteria, the genomic distribution of long DNA repeats is non-random. To reduce the effect of the strict identity threshold (100%) to predict repeats and overrepresentation of repeats in our analysis due to possible overlaps between proximal repeats, we merged repeats that were separated by ≤100 bases to one repeat. This produced a set of 5,520 genomes with at least 30 repeats. Following merging, we found that only ~38% (2,086/5,520) of the bacteria show non-random genomic distribution of repeats. This deviation of the genomic distribution of repeats from random can be seen across classes of bacteria with varying degrees (fig. 3*a* and *b*).

We next investigated the constraints imposed by the growth rate on the distribution of repeats. We found that ~90% (2,113/2,336) of fast-growing bacteria and ~78% (2,864/3,634) of slow-growing bacteria have non-random genomic distribution of repeats (≥90% iterations with $P \leq 0.01$). On merging proximal repeats, the number changes to ~50% (1,169/2,326) fast-growing bacteria and ~29% (917/3,194) slow-growing bacteria (fig. 3*c* and *d*).

We next investigated whether there are regions in the chromosome that are enriched with repeats. We divided each genome into 20 equally-sized bins taking *oriC* and *ter* as reference positions and then calculated the proportion of repeats in each bin as the percentage of the bin covered with repeats (fig. 4*a*). We observed that the repeat distribution across bins varied in fast- and slow-growing bacteria (fig. 4*b*). On comparing the distributions of the median proportion of repeats across bins, we found that the variability in the distribution of repeats across bins (statistically calculated as variance) was significantly less (*F*-test; $P = 0.0005$) in slow-growing bacterial genomes than in fast-growing bacterial genomes (fig. 4*c*).

Taken together, a larger fraction fast-growing bacteria show non-randomness in repeat distribution than slow-growing ones. This consistent difference between fast- and slow-growing bacteria strengthens the association between replication derived parameters and genomic distribution of repeats.

oriC Proximal Repeats are Involved in Lower Long-range Interactions

To identify patterns in repeat proximity to *oriC* and ter, we divided the genome into four quadrants as described earlier in (Khedkar and Seshasayee 2016). The Ori quadrant is centred around *oriC* and the Ter quadrant is around *ter*. The quadrants between *oriC* and *ter* on either replichores are called Right and Left quadrants, respectively (fig. 5*a*). For each quadrant pair, we calculated the proportion of repeat pairs as the count of repeat pairs in the pair of quadrants to the total number of repeat pairs in that genome. We observed that repeat pairs across genomes are present closer to each other, that is, as intraquadrant pairs (both repeats of a pair in the same quadrant) as compared with interquadrant pairs (repeats of a pair in different quadrants). This is clearly evident from the "X"-shaped pattern in the heatmap of median repeat pairs proportion across bacteria (fig. 5*b*).

On comparing repeat pairs proportion across quadrants, we observed that both ori and ter quadrants show the highest proportion of intraquadrant repeat pairs (fig. 5*c*). In contrast, the proportion of all interquadrant repeat pairs involving the ori quadrant (one repeat of a pair in ori quadrant and the other repeat in any of the left, right or *ter* quadrant) is less than that for all other quadrants (fig. 5*d*). Furthermore, we observed that this difference in the proportion of interquadrant repeat pairs and intraquadrant pairs involving ori is more significant in fast-growing bacteria as compared with the slow-growing bacteria (fig. 5*e* and *f*).

Thus, despite having more repeats near the origin of replication, fast-growing bacteria have a lower number of repeats potentially involved in highly disruptive, long-range recombination events involving *oriC*-proximal regions.

Distribution of Repeats Across Replichores is Different Depending on their Orientations

The impact of repeat pairs on fitness would depend on the repeat type and on whether the two members of the pair are present on the same replichore (i.e., same side of the *oriC–ter* axis) or on alternative replichores (i.e., on either side of the *oriC–ter* axis). Towards studying this effect, we categorized repeat pairs into two categories; inter-replichore repeats, that is, repeat pairs across two halves or replichores divided by the *oriC–ter* axis, and intra-replichore repeats as repeat pairs present on the same side of the axis (fig. 6*a*). This is of particular interest to inverted repeats: for example, inversions caused by recombination between repeats in the same replichore switch affected genes from leading to lagging strand and vice-versa whereas that between repeats across replichores do not. Additionally, inversion of the gene also leads to change in the gene dosage due to the relative change in *oriC* proximity in both inter- and intra-replichore pairs; however, if the two recombining

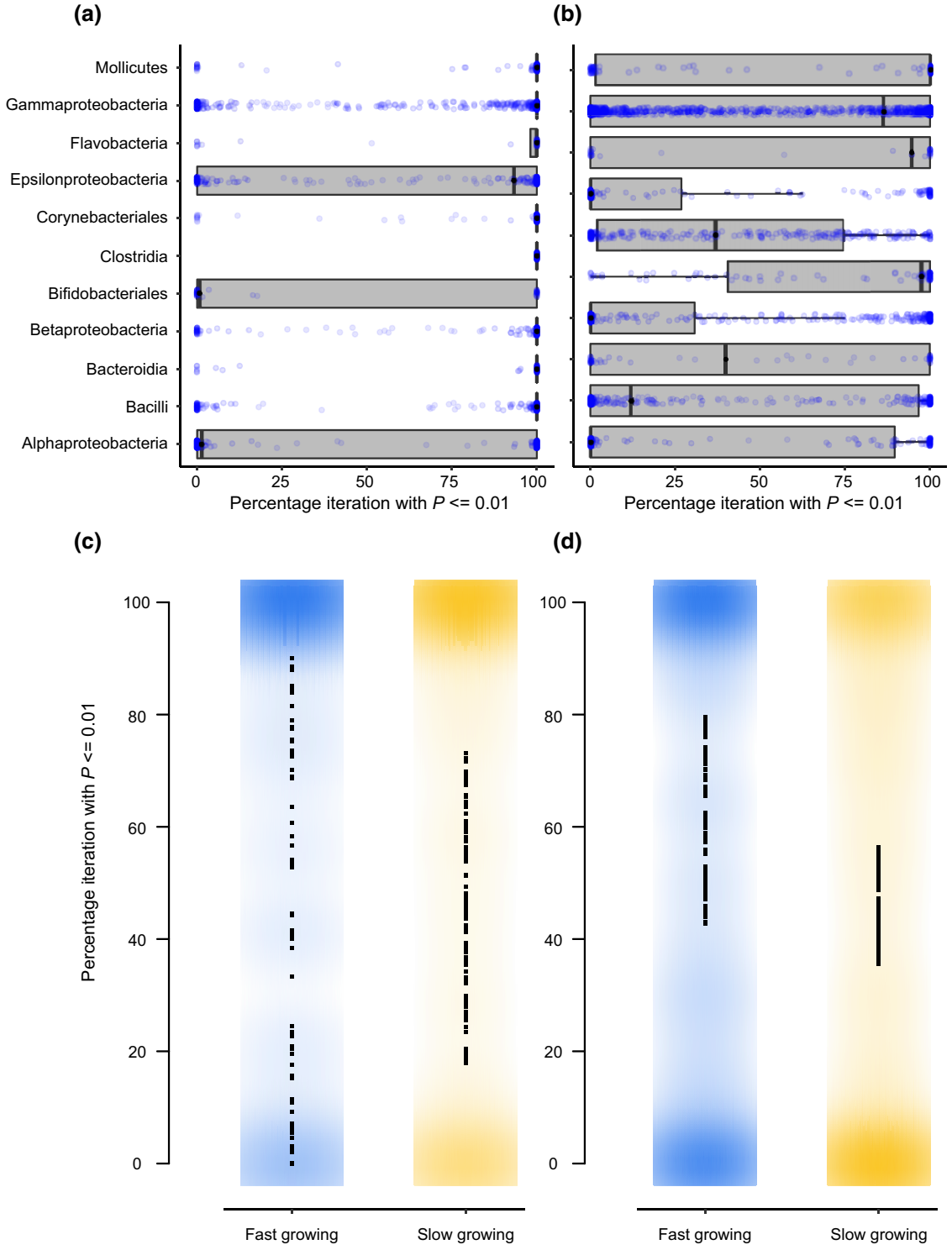

Fig. 3.—Long DNA repeats are distributed non-randomly across bacterial genomes. Boxplots of percentage iterations with $P \leq 0.01$ obtained to check randomness in repeat distribution of repeats (*a*) without merging the overlapping repeats and (*b*) after merging overlapping repeats with blue jitter showing the distribution. Only classes with at least 30 genomes are shown here. A smooth scatter plot of percentage iterations with $P \leq 0.01$ obtained to check randomness in repeat distribution in fast-growing and slow-growing bacteria (*c*) without merging the overlapping repeats and (*d*) after merging overlapping repeats. Higher transparency in the plot represents lower density of data and vice-versa. The *P* values in all the plots were calculated using KS test between the observed distribution of distances between repeats to that of random for 1,000 iterations. The *P* values were adjusted using Bonferroni correction for multiple correction.

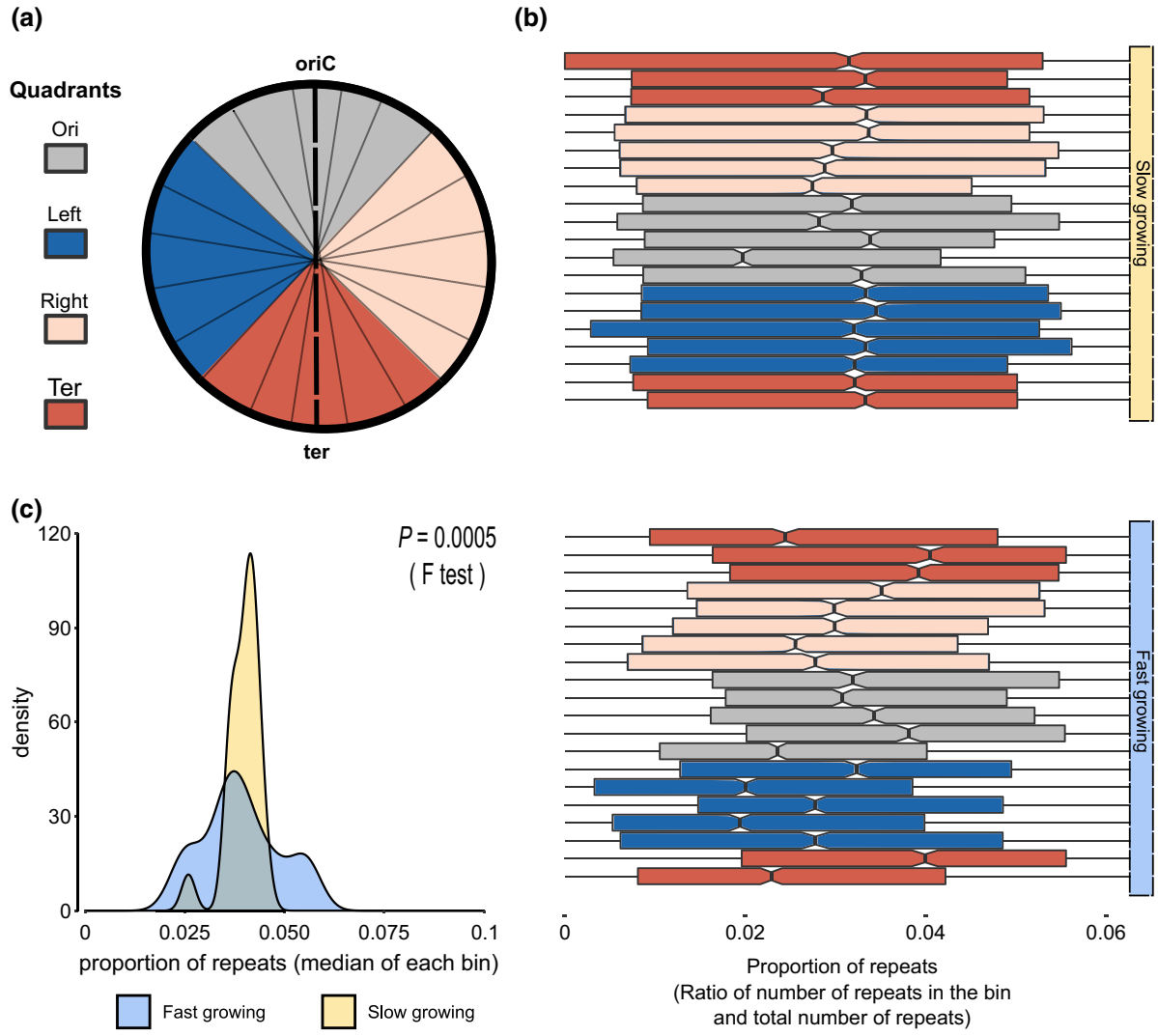

Fig. 4.—Differential genomic distribution of long DNA repeats in fast- and slow-growing bacteria. (*a*) Cartoon picture representing bacterial chromosome with the *oriC-ter* axis and 20 bins. The colors represent different quadrants made using *oriC* and *ter* as references. (*b*) Boxplots of repeat proportions in bins across bacterial genomes. The colors represent different quadrants made using *oriC* and *ter* as references. The upper plot represents the distribution in slow-growing bacteria while the lower set of the plot represent distribution in fast-growing bacteria. (*c*) Kernel density plot of proportion of repeats (median of each bin) obtained from (*b*) in fast- and slow-growing bacteria. The F test using var.test() function in R is used to calculate the difference in variance in the two distributions and obtain the *P* value.

inter-replichore inverted repeat elements are positioned symmetrically around the *oriC*, the inversions would affect gene dosage minimally. In case of deletions or duplications mediated by direct repeat pairs, the gene dosage of the affected region will be altered similarly in both intra-replichore and inter-replichore arrangements. However, deletion of the *oriC* itself would be immediately lethal and large duplication across replichores mediated by direct repeats would be unstable.

To perform statistical inference on the proportion of intra/inter-replichore repeats, we randomly allotted a replichore to all the repeat elements present in the genome. We then calculated the proportion of intra-replichore repeats in that randomization event. For every genome, a median of the proportions obtained from 1,000 randomization events was taken. On comparing the distribution of observed genomic proportions of intra-replichore repeats with the median proportions obtained by randomization, we found that repeat pairs are significantly higher (Wilcoxon signed-rank test [paired]; $P < 10^{-10}$) in the same replichore (intra-replichore) as compared with that by random event, in both fast- and slow-growing bacteria

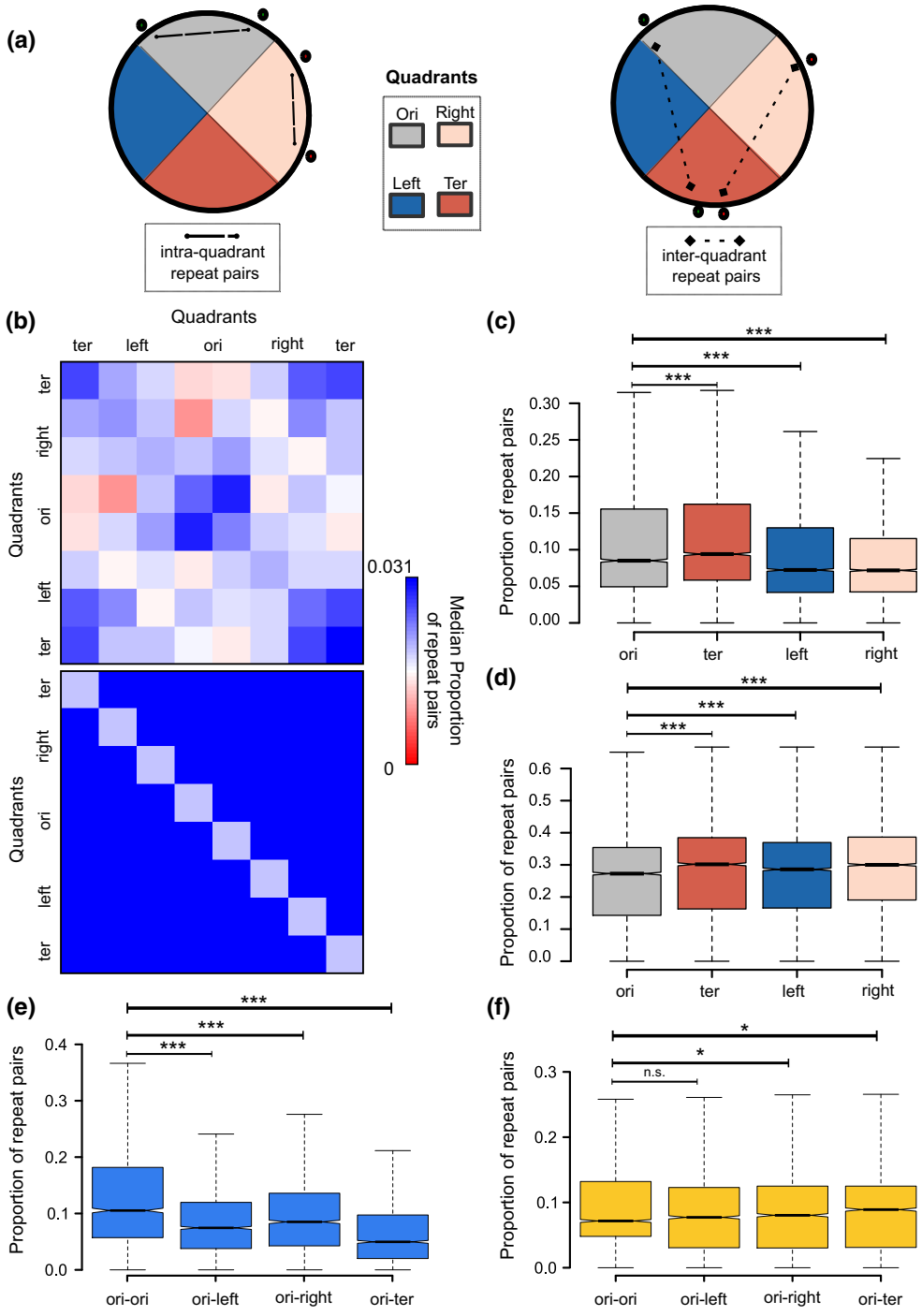

Fig. 5.—*oriC* proximal repeats are involved in lower long-range interactions. (*a*) Cartoon picture representing bacterial chromosome divided into four quadrants. The upper picture shows intraquadrant repeat pairs (dashed line with dots at the end); that is, both repeats of a pair are in the same quadrant while the lower picture shows interquadrant repeat pairs (dotted line with diamonds at the end); that is, repeats of a pair are in different quadrants. The colors represent different quadrants made using *oriC* and *ter* as references. (*b*) The heatmap of median (across genomes) proportion of repeat pairs across bin pairs. "X"-shaped pattern is visible indicating a higher proportion of intraquadrant repeat pairs. Boxplots of proportion of repeat pairs present as (*c*) intraquadrants pairs and (*d*) interquadrants pairs across quadrants. Boxplots of proportion of repeat pairs present as interquadrant pairs with ori as one of the quadrants in (*e*) fast-growing bacteria and (*f*) slow-growing bacteria Wilcoxon's rank-sum test is used to calculate the significance in the differences of two medians in each pair of comparison (n.s.:$P \geq 10^{-2}$, *$P < 10^{-2}$, **$P < 10^{-4}$, ***$P < 10^{-6}$).

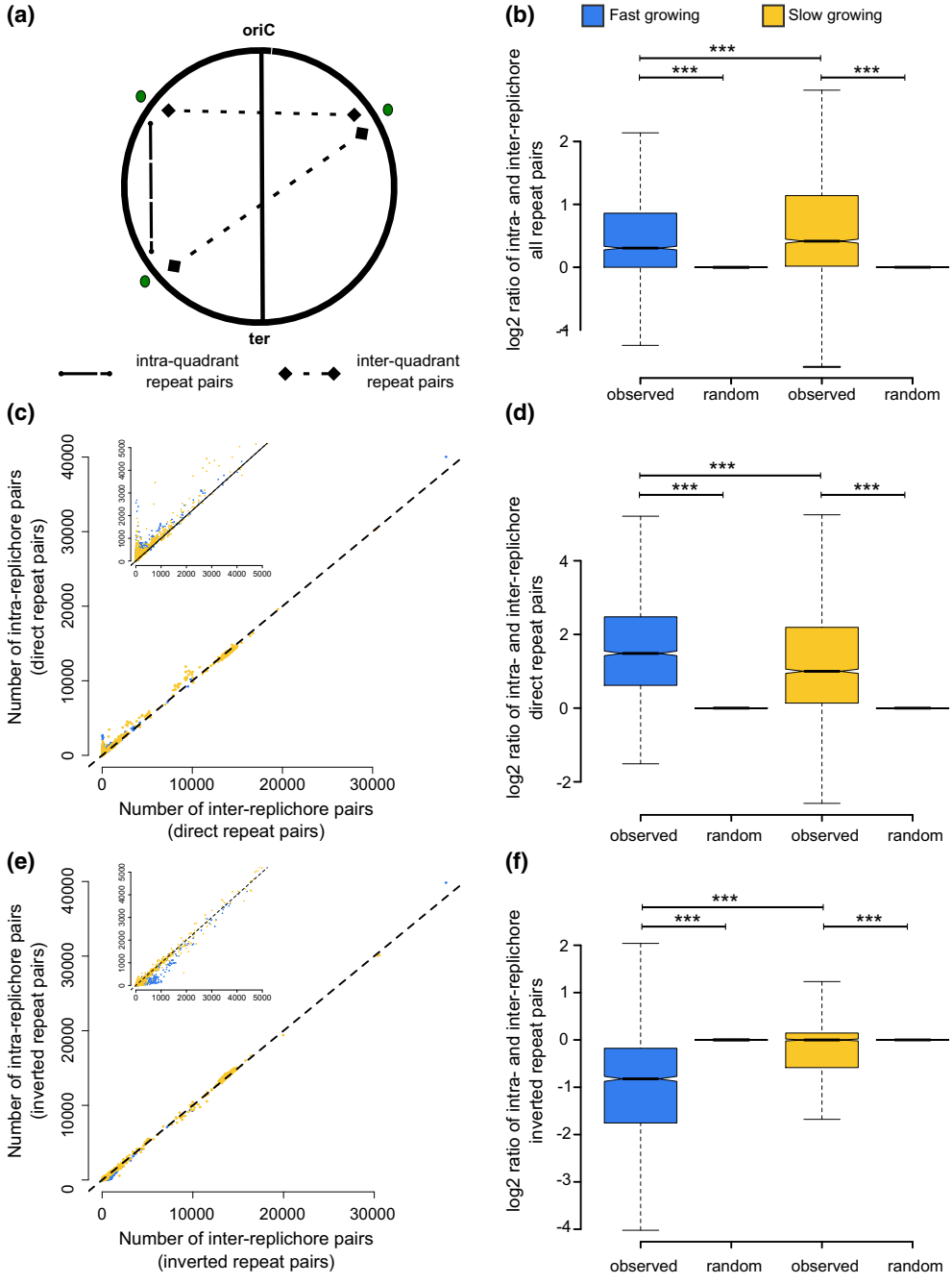

Fig. 6.—Distribution of repeats across replichores is different on the basis of their orientation. (*a*) Cartoon picture representing bacterial chromosome divided into two replichores across the *oriC-ter* axis. Oval dots on the chromosome represent repeats present as intra-replichore pair (dashed line with dots at the end) and as inter-replichore pair (dotted line with diamonds at the end). (*b*) Boxplots of proportion of intra-replichore repeat pairs shown as log2 (number of intra-replichore repeat pairs/number of inter-replichore repeat pairs) in all repeat pairs (both direct and inverted repeats). (*c*) The scatter plot of number of inter- and intra-replichore direct repeat pairs with colored dots indicating bacterial growth categories. (*d*) Boxplots of proportion of intra-replichore repeat pairs shown as log2 (number of intra-replichore repeat pairs/number of inter-replichore repeat pairs) for only direct repeat pairs in fast-growing and slow-growing bacteria. (*e*) The scatter plot of number of inter- and intra-replichore inverted repeat pairs with colored dots indicating bacterial growth categories. (*f*) Boxplots of proportion of intra-replichore repeat pairs shown as log2 (number of intra-replichore repeat pairs/number of inter-replichore repeat pairs) for only inverted repeat pairs in fast-growing and slow-growing bacteria. Wilcoxon's rank-sum test is used to calculate the significance in the differences of two medians in each comparison between fast- and slow-growing bacteria whereas Wilcoxon's signed-rank test (paired) is used to calculate the significance in the differences of medians of the observed and random populations in each comparison (n.s.:$P \geq 10^{-2}$, *$P < 10^{-2}$, **$P < 10^{-4}$, ***$P < 10^{-6}$).

(fig. 6*b*). This proportion is significantly less in fast-growing bacteria as compared with slow-growing bacteria (Wilcoxon rank-sum test; $P < 10^{-10}$) (fig. 6*b*).

Direct repeats were higher than by random chance intra-replichore (fig. 6*c* and *d*) (Wilcoxon signed-rank test [paired]; $P < 10^{-10}$) (fig. 6*e* and *f*), whereas inverted repeat pairs were present more in inter-replichore arrangement (Wilcoxon signed-rank test [paired]; $P < 10^{-10}$) (fig. 6*e* and *f*). When compared across bacterial growth categories, we observed that the number of intra-replichore direct repeats was significantly higher in fast-growing as compared with slow-growing bacteria (Wilcoxon rank-sum test; P-value $= 10^{-10}$) whereas intra-replichore inverted repeat pairs were found significantly less frequently in fast-growing organisms (Wilcoxon rank-sum test; $P < 10^{-10}$) (fig. 6*d* and *f*).

Genomic Distribution of Repeats Enables Less Disruption to the Replication-dependent Genome Organization

If repeat pairs are present in the same replichore (intra replichore repeats), a higher spacing between members of a repeat pair would lead to the rearrangement of a larger genomic region and thus disrupt the gene dosage gradient to a greater extent. On the contrary, if inter-replichore repeats are present symmetrical about the *oriC-ter* axis, the extent of disruption to gene dosage would be minimal.

In the case of direct repeat pairs, we observed that the average distance between repeat pairs in the same replichore is significantly (Wilcoxon signed-rank test [paired]; $P < 10^{-10}$) less than the average distance obtained from the randomly shuffled positions of the repeats in both fast- and slow-growing bacteria. This can be seen from the darker diagonal in the heatmap of the median proportion of direct repeats (fig. 7*a*). This distance was significantly lower (Wilcoxon rank-sum test; $P = 7.7 \times 10^{-6}$) in fast-growing bacteria when compared with that in slow-growing bacteria (fig. 7*b*). Surprisingly, in the case of inverted repeat pairs, the diagonal in the heat map of the median proportion of inverted repeats is not visible (fig. 7*c*). The distance between repeats is significantly higher than random (Wilcoxon signed-rank test [paired]; $P < 10^{-10}$) with no difference in fast- and slow-growing bacteria (fig. 7*d*).

To find out if repeat pairs are present symmetrical around *oriC*, we calculated the symmetry of inter-replichore repeat pairs as 1-absolute difference in the normalized relative position of the repeat units with respect to *oriC*. These values theoretically range from 0 to 1 with a value of "1" implying that the repeat pair is symmetrical around *oriC* while a repeat pair, that is, most asymmetric around the *oriC–ter* axis, that is, present on the axis itself, has a value of 0. While direct repeats are slightly more symmetric than random in terms of statistical significance, the effect size is small (fig. 7*e*–*h*). This significance holds only for fast-growing, but not in slow-growing bacteria (Wilcoxon signed-rank test [paired]; $P < 3.17 \times 10^{-5}$ for fast-growing bacteria; fig. 7*f*). However, inverted repeat pairs are symmetrical in both fast- and slow-growing bacteria with a clear difference from random expectation (Wilcoxon signed-rank test [paired]; $P < 10^{-10}$), and these repeats were significantly more symmetrical in fast-growing bacteria (Wilcoxon rank-sum test; $P < 10^{-10}$) (fig. 7*h*).

These observations of close intra-replichore direct repeats and symmetric inter-replichore inverted repeats remain unchanged when bacteria were classified based on growth estimates using the dataset by (Vieira-Silva and Rocha 2010) (supplementary fig. S2, Supplementary Material online). These results hold true even after removing species redundancy (supplementary fig. S3, Supplementary Material online) and in most of the comparisons in major phyla (supplementary fig. S4, Supplementary Material online) with varying degrees of significance. To account for the effect of phylogenetic relatedness in the patterns observed in our study, we also performed phylogenetic analysis of variance (ANOVA) using "phylANOVA" function in "phytools" package of R with a phylogenetic tree of 524 nonredundant species. We observed that fast- and slow-growing genomes were significantly ($P = 0.001$) different in the closeness of intra-replichore direct repeats and the symmetry of inter-replichore inverted repeats.

Our observations of the symmetric presence of long DNA repeats are consistent with previous reports on symmetric inversions ("X" shaped patterns on pairwise genome alignment) in *Azotobacter vinelandii* and *Vibrio chloreae* genomes (Repar and Warnecke 2017) and symmetric translocations in *Caulobacter crescentus* (Khedkar and Seshasayee 2016). In a scatter plot of normalized positions of inter-replichore repeat pairs, we found that most of the repeats in *Vibrio* and *Caulobacter* are closer to the diagonal which indicates their symmetry around *oriC* as seen in fig. 8*a* and *b*. This can be clearly seen in the distribution of absolute difference in the relative positions of the repeats, where the mode is closer to 0 (closer/symmetric the repeat pairs, lower the difference in their relative positions). Though we do not see such clear symmetry in the position of repeats with respect to the diagonal in *Azotobacter*, the distribution of the absolute difference in positions has a mode closer to 0 (fig. 8*c*). These examples also show proximity of repeat-pairs with respect to each other.

Discussion

Changes in genome organization in terms of orientation and positions relative to *oriC* affect bacterial fitness and thus are under selection. Likewise, anything mediating these changes will also be under selection. Since these changes are primarily caused by repeat-mediated

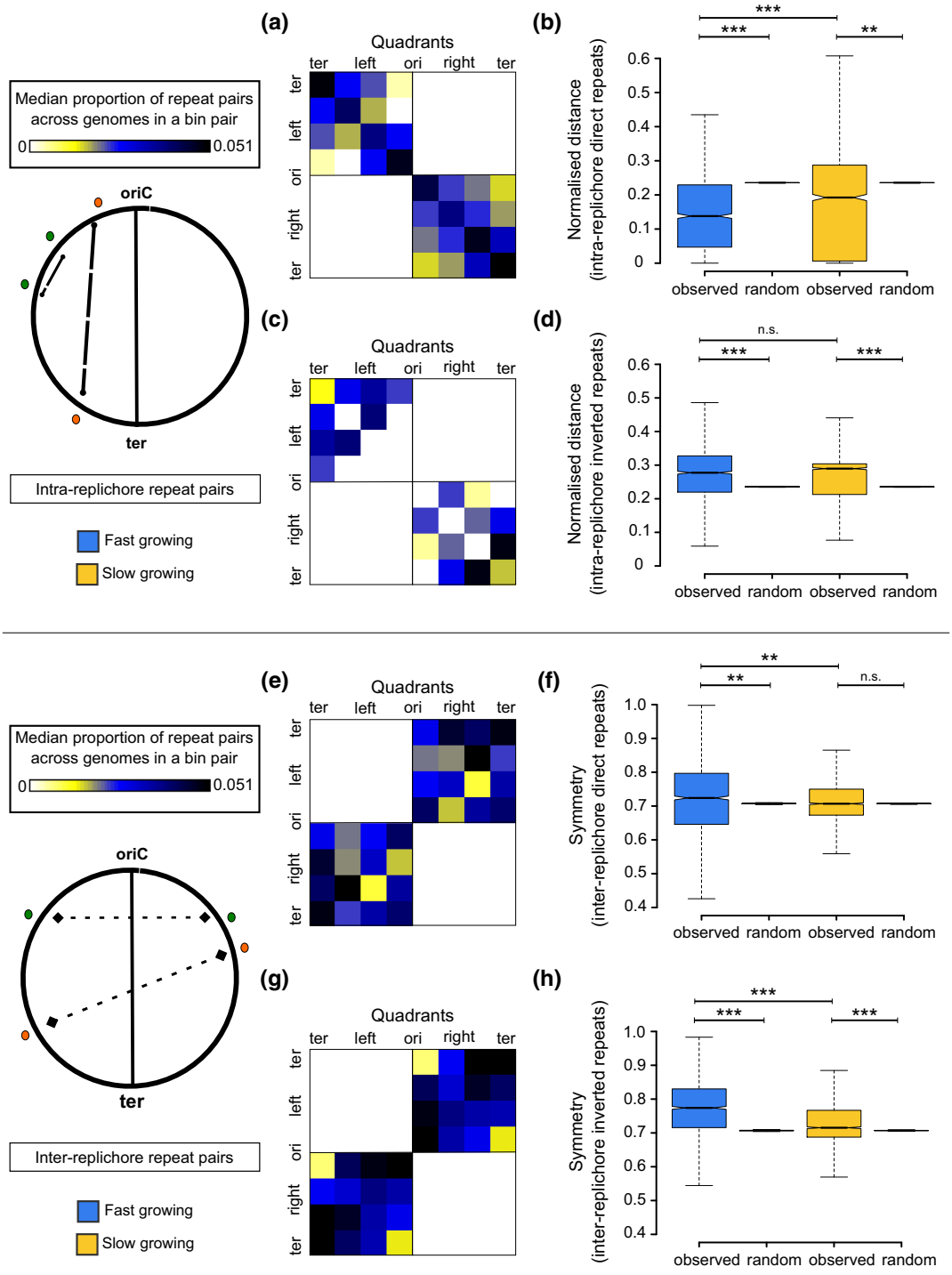

Fig. 7.—Genomic distribution of repeats enables less disruption to the replication-dependent genome organization. (*a*) The heatmap of median (across genomes) proportion of intra-replichore direct repeat pairs across bin pairs. (*b*) Boxplots of normalized distances between intra-replichore direct repeat pairs in fast- and slow-growing bacteria. (*c*) The heatmap of median (across genomes) proportion of intra-replichore inverted repeat pairs across bin pairs. (*d*) Boxplots of normalized distances between intra-replichore inverted repeat pairs in fast- and slow-growing bacteria. (*e*) The heatmap of median (across genomes) proportion of inter-replichore direct repeat pairs across bin pairs. (*f*) Boxplots of symmetry between inter-replichore direct repeat pairs in fast- and slow-growing bacteria. (*g*) The heatmap of median (across genomes) proportion of inter-replichore inverted repeat pairs across bin pairs. (*h*) Boxplots of symmetry between inter-replichore inverted repeat pairs in fast- and slow-growing bacteria. Wilcoxon's rank-sum test is used to calculate the significance in the differences of two medians in each comparison between fast- and slow-growing bacteria whereas Wilcoxon's signed-rank test (paired) is used to calculate the significance in the differences of medians of the observed and random populations in each comparison (n.s.:$P \geq 10^{-2}$, *$P < 10^{-2}$, **$P < 10^{-4}$, ***$P < 10^{-6}$).

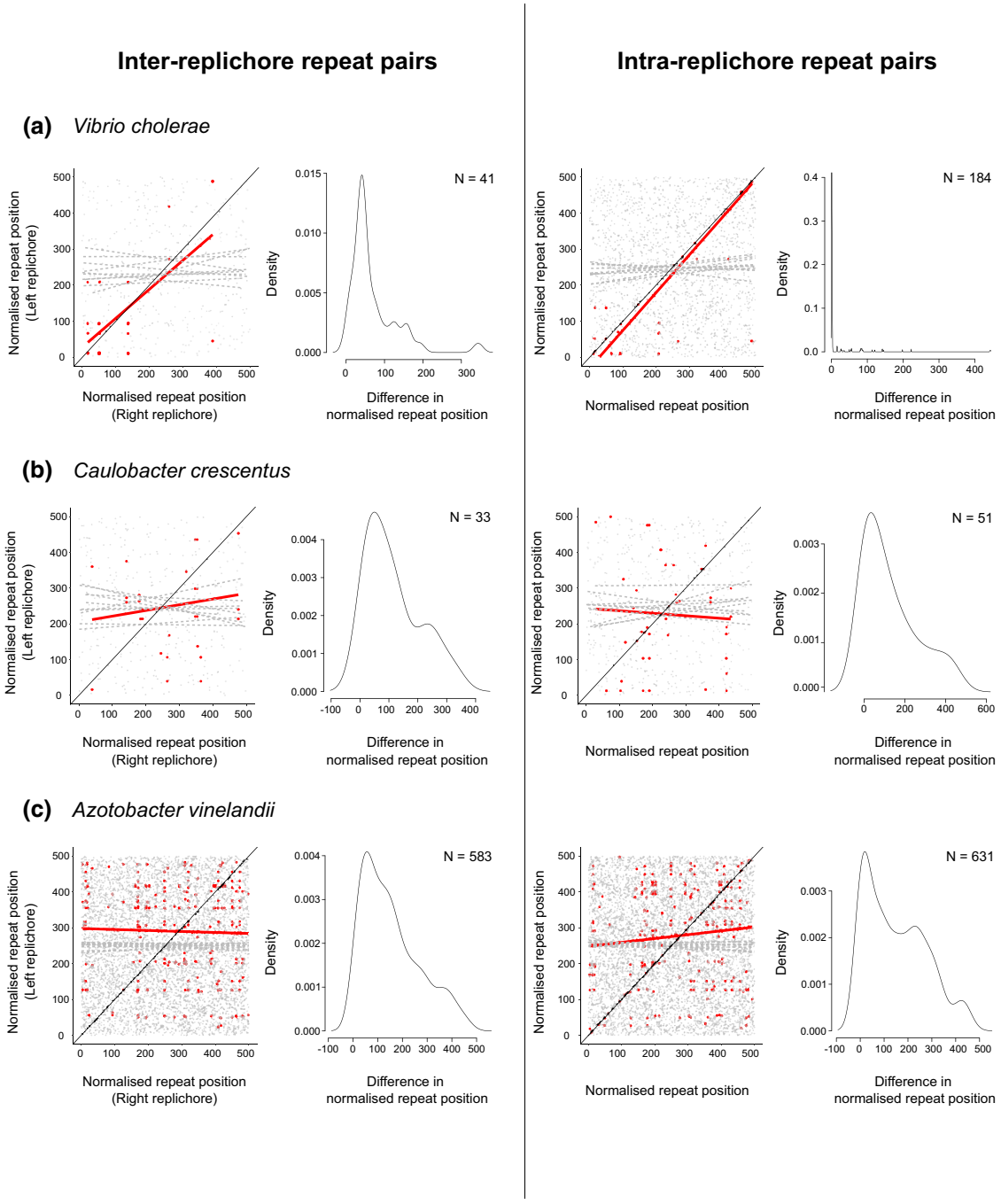

Fig. 8.—Genomic positions of long DNA repeats show patterns similar to observed symmetric inversions and translocations. The scatter plot of normalized repeat positions of the repeats in a pair in case of inter-replichore and intra-replichore repeat pairs with kernel density distribution of absolute differences in the normalized relative positions of the repeats in (*a*) *Vibrio cholerae*, (*b*) *Caulobacter crescentus,* and (*c*) *Azotobacter vinelandii*. The red colored points in the scatter plot represents the positions of observed dataset with red line indicating the linear model fit whereas grey points and dotted lines represent the positions of randomly shuffled set for 10 iterations.

rearrangements, this begs the question of how these repeats are distributed in the genome. Despite a large increase in the number of fully-sequenced bacterial genomes in the recent past, this question has attracted limited attention since (Achaz et al. 2003). In this study, we used a large dataset of ~6,000 genomes and show that long repeats are distributed in a non-random manner across bacteria. Though regions near *oriC* and *ter* are repeat-rich in fast-growing bacteria, these are positioned in a manner that they can cause short-range

rearrangements and thereby disrupt genome organization minimally. This was supported by the study by (Valens et al. 2016) demonstrating lower inversion frequency between Ori and Right macrodomains relative to intra-macrodomain frequency. In line with previous reports that found symmetry in inversions across bacterial genomes (Helm et al. 2003; Kong et al. 2009; Repar and Warnecke 2017; Eisen et al.), we find that inter-replichore inverted repeat pairs tend to be more symmetric around the *oriC–ter* axis than expected by random chance.

Rearrangements and Genome Organization

Since selection would act on rearrangements and these are a function of not only repeat pair presence and distance/symmetry but also their propensity to interact physically, variations in 3D conformations in the chromosome can enhance, limit or even abolish selection on repeat positions. We took 3C or Hi-C data available for *C. crescentus*, *E. coli*, *Bacillus subtilis*, and *Mycoplasma pneumoniae* from previous studies (Le et al. 2013; Wang et al. 2014; Trussart et al. 2017; Lioy et al. 2018) and compared the normalized interaction scores (as reported in the study) of the genomic regions (bin pairs) containing repeats with the regions devoid of repeats in the WT condition. Though we see a significant difference in interaction scores between the groups in *E. coli*, we do not see any difference in the average interaction scores of *C. crescentus* and *B. subtilis.* The average interaction scores of regions with repeats were significantly lower than the rest of the regions in *M. pneumoniae* (supplementary fig. S5, Supplementary Material online). The inconsistency in patterns observed across these few limited bacterial genomes suggests that the product of repeat position and repeat interaction is likely to vary considerably across organisms. As more Hi-C and related data for bacteria accumulate, a more comprehensive analysis of how chromosome shape might affect selection on gene and repeat positioning could be achieved.

Repeat Distribution Patterns: Mechanistic Processes and Selection

How do repeats originate? Achaz et al. (2002) proposed that repeats arise at the first instance by tandem duplication and such repeat pairs are direct repeats. Additional direct repeats and inverted repeats are created by other events causing structural rearrangements such as inversions, translocations, inversions, etc. We performed a toy simulation in which a single repeat element placed in one of 100 random genomic slots is allowed to undergo several tandem duplications and rearrangements at arbitrary rates; in this simulation, we assumed that rearrangements maintain or reverse the orientation of the affected repeat at equal probability. By definition, this would produce more direct than inverted repeats as observed in real data—more so at lower total repeat counts before enough translocation events have accumulated—unless the rate of rearrangements is high enough relative to that of duplications and the probability of rearrangements generating inverted repeats is much higher than those producing more direct repeats.

As expected, even without invoking selection, this model of repeat generation would place direct repeats closer to each other than expected by pure random chance. If selection were imposed over this such that the probability of the loss of an intra-replichore direct repeat would be higher the farther the members of the pair are from each other, the distribution of the distance between the repeat elements would decrease (supplementary fig. S6a, Supplementary Material online, for direct repeats). Our observation from genome data in which intra-replichore direct repeats are closer together more in fast-growing than in slow-growing bacteria suggests a role for selection unless this difference could be explained purely by more frequent replication events causing more tandem duplications. Inter-replichore direct repeat pairs would almost always be generated by translocation-related processes and no additional origin of selection, beyond distance between them, can be envisaged for them.

Inverted repeats present a more interesting situation. They are not generated by tandem duplication and assuming that translocation events place them at random positions on the genome, there is no reason to assume that without selection their relative position should be any different from random. Curiously, we observe in bacterial genomes that intra-replichore inverted repeats are placed further apart from each other than random expectation (though this trend is lost when we picked one random representative genome per species). Assuming this pattern is biologically meaningful, at least in a subset of genomes considered here, there are two possibilities that can cause such a pattern to arise: 1) there could be a minimum distance over which translocations can occur in some genomes, and this would likely push inverted repeat pairs further apart than expected by random chance; 2) selection operates *against* inverted repeat pairs that are relatively close to each other (supplementary fig. S6a, Supplementary Material online, for inverted repeats), presumably because chromosome conformations make recombination events between closely positioned repeats more likely; this would be less effective for direct repeat pairs for which the process of generation by tandem duplication would be a strong counter-force. Inter-replichore inverted repeat pairs would be positioned randomly in the absence of selection. It would be reasonable to assume that selection would operate against inter-replichore inverted repeat pairs that are positioned asymmetrically about the *oriC-ter* axis. Therefore, imposing a probability of repeat loss under selection, that is, inversely proportional to symmetry would tend to place inter-replichore inverted repeat pairs in more symmetric positions than random

(supplementary fig. S6b, Supplementary Material online) producing a distribution of repeat pair symmetry that is shifted to the right, similar to the results obtained here as well as for the translocation data observed by (Khedkar and Seshasayee 2016).

Further exploration of this toy model with a deep sampling of the parameter space might provide additional insight into the processes underlying the repeat distributions observed in this study.

Conclusion

Maintenance of genome organization and chromosomal variations are interdependent and go hand-in-hand in the course of genome evolution. On one hand, structural variations play a major role in the adaptation and evolution of the organism (Straus and Hoffmann 1975; Sonti and Roth 1989; Rocha et al. 1999; Rocha 2002; Srinivasan et al. 2015; Veetil et al. 2020; Gorkovskiy and Verstrepen 2021), and on the other hand, they can disrupt the genomic organization and thus affect fitness (Hill and Harnish 1981; Hill and Gray 1988; Rocha 2004; Adler et al. 2014). Taken together, by analyzing ~6,000 genomes, our study finds an association between intrachromosomal long DNA repeats and replication-dependent bacterial genome organization. However, we considered only homologous recombination as the mechanism of repeat-mediated disruptions and do not take account of disruptions mediated by transposition events (e.g., by IS). We suggest maintenance of replication-dependent genome organization as a selection pressure in the positioning of long DNA repeats, involved in causing chromosomal rearrangements. There exists a balance between the interplay of genome variability caused by repeats and genome maintenance, and the presence of repeats at certain positions enables a reduction in disruption to the replication-dependent genome organization.

Materials and Methods

Data

DNA sequence files (.fna), RNA files (.fna), and genomic feature files (.gff) of 6,387 completely sequenced (as of April, 2019) bacterial genomes were downloaded from RefSeq database (O'Leary et al. 2016) at the NCBI FTP website. Only the main chromosome of a genome was used for this study. The classification of these genomes into phyla, class, genera, and species was done based on the information available on sequence headers and KEGG classification of NCBI genomes (https://www.genome.jp/kegg/docs/cmp_prok.html). These genomes had at least one of their species representatives with a single predicted origin of replication in the DoriC database (as of April, 2019) (Gao and Zhang 2007) (DoriC 6.5, http://tubic.tju.edu.cn/doric/public/index.php).

Classification of Bacteria Based on Growth

Bacterial genomes were classified into two categories: fast-growing and slow-growing by calculating a parameter called Rf factor (Couturier and Rocha 2006; Khedkar and Seshasayee 2016). It was defined as the ratio of estimated time taken for a full round of chromosome replication (Tr) and minimum doubling time (Td). Tr was calculated by taking the ratio of half the genome size to 600 nt/s which is an average speed of DNA replication (Reyes-Lamothe et al. 2008; Milo et al. 2009). For Tr, genome sizes were defined as string length of the sequence in .fna files using in house python script, while Td was estimated by rDNA copy number and doubling time of the bacteria as mentioned in Couturier and Rocha (2006); Freilich et al. (2009); Khedkar and Seshasayee (2016).

Bacteria with Rf>1 are expected to initiate replication more than once per cell cycle on average which is true for fast-growing bacteria. On the contrary, slow-growing bacteria have Rf $\leq$ 1. In our data, 2,337 bacteria belong to the fast-growing and 4,003 bacteria to the slow-growing category. These categories share 91.7% similarity (5,817/6,340 bacteria) (correlation between Rf values: 0.98) with an independent categorization done using a larger dataset by Vieira-Silva and Rocha (2010).

Identification of Long Intrachromosomal DNA Repeats and Associated Genomic Regions

Identical long repeats of length at least 100 bp were identified by the repeat-match algorithm of the MUMmer (version 3.23) software using -n 100 option (Delcher et al. 2002; Kurtz et al. 2004). The positions and orientation of the repeat pairs were extracted using an in-house python script. Repeat pairs were classified as direct pairs when both the repeat partners are in the same orientation, that is, the positions of the exact matches in the mummer output are on the same strand and inverted pairs when it was otherwise.

Genomic coordinates of CDS and rDNA regions were extracted from the .gff files. IS elements, prophages, and horizontally acquired regions were identified using ISEScan (version 1.7.2) (Xie and Tang 2017), Phaster web server (https://phaster.ca/) (Arndt et al. 2016), and AlienHunter (version 1.7) (Vernikos and Parkhill 2006), respectively. To identify the genomic regions associated with repeats, the overlaps between the identified repeats and these regions were calculated by comparing the respective genomic coordinates. A repeat was reported to be associated with these regions if there was an overlap of even a single nucleotide base with the respective region. Repeat pairs in which any of the repeat partners had an overlap with rDNA regions were removed from further analysis. This led us to 6,340 genomes with at least one repeat pair.

Association of Inversions with Inverted Repeat Elements

For every species in our dataset, a reference genome was chosen randomly and other genomes were used as a query for predicting intergenomic inversions. To predict large chromosomal inversions of at least 1 kb size, nucmer -l 1000 function of MUMmer (version 3.23) (Delcher et al. 2002; Kurtz et al. 2004) package was used. Using the orientation information in the output file, large intergenomic inversions were found. Their positions were parsed using in-house python scripts and compared with the positions of intrachromosomal long inverted repeats. Only those genomes were chosen for the analysis which had at least 30 distinct genomic positions with repeats in the genome and at least one predicted inversion with respect to the reference. This left us with 3,730 genomes for this analysis.

For every inversion detected in the genome with respect to the reference, distance (in bases) was calculated between the nearest repeat (as part of inverted repeat pairs identified) and the inversion observed. This gave us four different distance values; from 5′ and 3′ sides in the query and its reference; and among which the shortest distance was used to compare the distances for this particular analysis.

Calculation of Relative Position of Repeats with Respect to *oriC*

For every bacterial species, the *oriC* of a representative genome in the D*oriC* database was used as a reference and blastn (version 2.2.31+) (Altschul et al. 1990) was performed to predict *oriC* sequences in the rest of the genomes of that species. Blast hits with the highest score and with at least 90% identity were chosen as the *oriC* sequence of that genome. For every genome with predicted *oriC*, the shortest distance between the repeat and the *oriC* was calculated thus providing the relative position of repeats with respect to the origin of replication. The relative position was then normalized to 1,000 bp, that is, divided by genome size and then multiplied by 1,000 to make it comparable across genomes.

Calculation of Genomic Distribution of Repeats

To understand the genomic distribution of repeats, every genome was divided into equal parts of 2, 4, and 20 bins centred around the *oriC*-*ter* axis. Ori bin was centred at *oriC* while *ter* bin was referred to as the region diametrically opposite to the Ori bin. Left and right bin(s) were classified as regions which were on the left replichore or upstream of *oriC* and right replichore or downstream of *oriC*, respectively. This is similar to the approach taken in Khedkar and Seshasayee (2016).

To calculate bin level enrichment, repeats were counted for every bin using their normalized positions and were divided by the total number of repeats in the genome. These genomes (5,970 genomes) had at least 30 distinct genomic positions with repeats. While comparing the enrichment between fast-growing and slow-growing bacteria, the distributions of medians of proportions across bins were made across two groups and compared using F-test. A similar analysis was performed for calculating intraquadrant repeat enrichment.

Calculation of Distances and Symmetry Between Repeat Pairs

For intra-replichore repeat pairs, the distance between repeats was calculated as the absolute difference in the normalized position of repeat pairs. And symmetry in inter-replichore repeat pairs was calculated as 1—absolute difference in the normalized relative position of the repeat units with respect to *oriC*, similar to Khedkar and Seshasayee (2016). These values ranged from 0 to 1 with 0 being most symmetric around the ori-*ter* axis while 1 being least symmetric, that is, repeat pair is on the axis itself. For this analysis, the genomes that had at least 30 distinct positions of repeats involved in direct and inverted repeat pairs were taken. A total of 4,165 genomes were used in this analysis.

Phylogenetic ANOVA

16S rDNA sequences were extracted from RNA files using their description. One rDNA sequence with minimum Ns and maximum length was chosen for each genome. For each species, one representative genome was chosen randomly to make a phylogenetic tree. Multiple sequence alignment of the sequences followed by making phylogenetic tree was done by using "GTR" as evolutionary model and "Gamma" rate model for likelihoods in "FastTree" program of SILVA ACT browser (https://www.arb-silva.de/aligner/) (Pruesse et al. 2012). The phylogenetic tree was visualized and pruned in iTOL (https://itol.embl.de/) (Letunic and Bork 2019) web browser. After pruning, a phylogenetic tree with 524 species were obtained. With tree and the median values for distance/symmetry for each species as input, phylANOVA function in phytools package (Revell 2012) was used to calculate the statistical significance in the difference between fast- and slow-growing bacterial categories for distance/symmetry.

Null Models and Comparisons Across Parameters

To test the statistical significance of the observed values of the parameters in all our analyses, we compared the observed dataset relevant to that analysis with that of the null model generated through randomization. The underlying assumptions of all comparisons were that individual repeat elements are independent of each other and there is no other constraint than bacterial growth. To consider phylogenetic dependency of the genomes, we show major

observations at the phyla level and also after removing redundancy in species. All the comparisons and their corresponding null models are explained in the relevant section.

Supplementary Material

Supplementary materials are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This work was supported by a DBT/Wellcome Trust India Alliance Intermediate Fellowship [IA/I/16/2/502711 to A.S.N.S] and core funding from the Department of Atomic Energy, Government of India [Project No. 12-R&D-TFR-5.04-0800]. We thank Anjana Badrinarayanan, Dasaradhi Palakodeti, and Dimple Notani for helpful discussions as part of the thesis advisory committee. We also thank Akshara Dubey, Meghna Nandy, Mohak Sharda, and other members of the ASN lab for valuable inputs and discussions throughout the study. We appreciate contributions by Akshaya Seshadri, Anurag Kumar Singh, Kushi Anand, and Neha Sontakke for proofreading the manuscript.

Data Availability

The data and scripts underlying this article are available at Figshare repository which can be accessed via https://tinyurl.com/yh3y84un.

Literature Cited

Achaz G, Coissac E, Netter P, Rocha EPC. 2003. Associations between inverted repeats and the structural evolution of bacterial genomes. Genetics 164:1279–1289. doi:10.1093/genetics/164.4.1279

Achaz G, Rocha EPC, Netter P, Coissac E. 2002. Origin and fate of repeats in bacteria. Nucleic Acids Res. 30:2987–2994. doi:10.1093/nar/gkf391

Adler M, Anjum M, Berg OG, Andersson DI, Sandegren L. 2014. High fitness costs and instability of gene duplications reduce rates of evolution of new genes by duplication-divergence mechanisms. Mol Biol Evol. 31:1526–1535. doi:10.1093/molbev/msu111

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410. doi:10.1016/S0022-2836(05)80360-2

Arndt D, et al. 2016. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res. 44:W16–W21. doi:10.1093/nar/gkw387

Bi X, Liu LF. 1996. DNA rearrangement mediated by inverted repeats. Proc Natl Acad Sci USA 93:819–823. doi:10.1073/pnas.93.2.819

Bigot S, et al. 2005. KOPS: DNA motifs that control E. coli chromosome segregation by orienting the FtsK translocase. EMBO J 24:3770–3780. doi:10.1038/sj.emboj.7600835

Bishop AJR, Schiestl RH. 2000. Homologous recombination as a mechanism for genome rearrangements: environmental and genetic effects. Hum Mol Genet. 9:2427–2334. doi:10.1093/hmg/9.16.2427

Cooper S, Helmstetter CE. 1968. Chromosome replication and the division cycle of Escherichia coli B/r. J Mol Biol. 31:519–540. doi:10.1016/0022-2836(68)90425-7

Couturier E, Rocha EPC. 2006. Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes: Gene dosage effects and genome organisation. Mol Biol. 59:1506–1518.

Delcher AL, Phillippy A, Carlton J, Salzberg SL. 2002. Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res. 30:2478–2483. doi:10.1093/nar/30.11.2478

Duggin IG, Bell SD. 2009. Termination structures in the Escherichia coli chromosome replication fork trap. J Mol Biol. 387:532–539. doi:10.1016/j.jmb.2009.02.027

Frank AC, Lobry JR. 1999. Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. Gene 238:65–77. doi:10.1016/S0378-1119(99)00297-8

Freilich S, et al. 2009. Metabolic-network-driven analysis of bacterial ecological strategies. Genome Biol. 10:R61. doi:10.1186/gb-2009-10-6-r61

Gao F, Zhang C-T. 2007. DoriC: a database of oriC regions in bacterial genomes. Bioinformatics 23:1866–1867. doi:10.1093/bioinformatics/btm255

Gorkovskiy A, Verstrepen KJ. 2021. The role of structural variation in adaptation and evolution of yeast and other fungi. Genes 12:699. doi:10.3390/genes12050699

Helm RA, Lee AG, Christman HD, Maloy S. 2003. Genomic rearrangements at rrn Operons in Salmonella. Genetics 165:951–959. doi:10.1093/genetics/165.3.951

Hill CW, Gray JA. 1988. Effects of chromosomal inversion on cell fitness in Escherichia coli K-12. Genetics 119:771–778. doi:10.1093/genetics/119.4.771

Hill CW, Harnish BW. 1981. Inversions between ribosomal RNA genes of Escherichia coli. Proc Natl Acad Sci. 78:7069–7072. doi:10.1073/pnas.78.11.7069

Khedkar S, Seshasayee ASN. 2016. Comparative genomics of interreplichore translocations in bacteria: A measure of chromosome topology? G3 Genes|Genomes|Genetics 6:1597–1606.

Kong S-G, et al. 2009. Inverse symmetry in complete genomes and whole-genome inverse duplication Aziz, RK, editor. PLoS ONE 4:e7553. doi:10.1371/journal.pone.0007553

Kurtz S, et al. 2004. Versatile and open software for comparing large genomes. Genome Biol. 5:R12. doi:10.1186/gb-2004-5-2-r12

Lam ST, Stahl MM, McMilin KD, Stahl FW. 1974. Rec mediated recombinational hot spot activity in bacteriophage lambda. II. A mutation which causes hot spot activity. Genetics 77:425–433. doi:10.1093/genetics/77.3.425

Lambert S, et al. 1999. Analysis of intrachromosomal homologous recombination in mammalian cell, using tandem repeat sequences. Mutat Res. 433:159–168. doi:10.1016/S0921-8777(99)00004-X

Lawrence JG. 2002. Shared strategies in gene organization among prokaryotes and eukaryotes. Cell 110:407–413. doi:10.1016/S0092-8674(02)00900-5

Le TBK, Imakaev MV, Mirny LA, Laub MT. 2013. High-resolution mapping of the spatial organization of a bacterial chromosome. Science 342:731–734. doi:10.1126/science.1242059

Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 47:W256–W259. doi:10.1093/nar/gkz239

Lioy VS, et al. 2018. Multiscale structuring of the E. coli chromosome by nucleoid-associated and condensin proteins. Cell 172:771–-783.e18. doi:10.1016/j.cell.2017.12.027

Lobry JR. 1996. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol Biol Evol. 13:660–665. doi:10.1093/oxfordjournals.molbev.a025626

Maharjan RP, et al. 2013. A case of adaptation through a mutation in a tandem duplication during experimental evolution in Escherichia coli. BMC Genomics 14:1. doi:10.1186/1471-2164-14-441
Milo R, Jorgensen P, Moran U, Weber G, Springer M. 2009. BioNumbers—the database of key numbers in molecular and cell biology. Nucleic Acids Res. 38:D750–D753. doi:10.1093/nar/gkp889
O'Leary NA, et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 44:D733–D745. doi:10.1093/nar/gkv1189
Pruesse E, Peplies J, Glöckner FO. 2012. SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics 28:1823–1829. doi:10.1093/bioinformatics/bts252
Repar J, et al. 2017. Elevated rate of genome rearrangements in radiation-resistant bacteria. Genetics 205:1677–1689. doi:10.1534/genetics.116.196154
Repar J, Warnecke T. 2017. Non-random inversion landscapes in prokaryotic genomes are shaped by heterogeneous selection pressures 34:1902–1911.
Revell LJ. 2012. phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol Evol. 3:217–223. doi:10.1111/j.2041-210X.2011.00169.x
Reyes-Lamothe R, Possoz C, Danilova O, Sherratt DJ. 2008. Independent positioning and action of Escherichia coli replisomes in live cells. Cell 133:90–102. doi:10.1016/j.cell.2008.01.044
Rocha EPC. 2002. Genomic repeats, genome plasticity and the dynamics of Mycoplasma evolution. Nucleic Acids Res. 30:2031–2042. doi:10.1093/nar/30.9.2031
Rocha EPC. 2003. Gene essentiality determines chromosome organisation in bacteria. Nucleic Acids Res. 31:6570–6577. doi:10.1093/nar/gkg859
Rocha EPC. 2004. The replication-related organization of bacterial genomes. Microbiology 150:1609–1627. doi:10.1099/mic.0.26974-0
Rocha EPC. 2008. The organization of the bacterial genome. Annu Rev Genet. 42:211–233. doi:10.1146/annurev.genet.42.110807.091653
Rocha EP, Danchin A, Viari A. 1999. Analysis of long repeats in bacterial genomes reveals alternative evolutionary mechanisms in Bacillus subtilis and other competent prokaryotes. Mol Biol Evol. 16:1219–1230. doi:10.1093/oxfordjournals.molbev.a026212
Shen P, Huang HV. 1986. Homologous recombination in Escherichia coli: dependence on substrate length and homology. Genetics 112:441–457. doi:10.1093/genetics/112.3.441
Skovgaard O, Bak M, Løbner-Olesen A, Tommerup N. 2011. Genome-wide detection of chromosomal rearrangements, indels, and mutations in circular chromosomes by short read sequencing. Genome Res. 21:1388–1393. doi:10.1101/gr.117416.110
Sonti RV, Roth JR. 1989. Role of gene duplications in the adaptation of Salmonella typhimurium to growth on limiting carbon sources. Genetics 123:19–28. doi:10.1093/genetics/123.1.19
Srinivasan R, Scolari VF, Lagomarsino MC, Seshasayee ASN. 2015. The genome-scale interplay amongst xenogene silencing, stress response and chromosome architecture in Escherichia coli. Nucleic Acids Res. 43:295–308. doi:10.1093/nar/gku1229
Srivatsan A, Tehranchi A, MacAlpine DM, Wang JD. 2010. Co-orientation of replication and transcription preserves genome integrity moran, NA, editor. PLoS Genet 6:e1000810. doi:10.1371/journal.pgen.1000810
Straus DS, Hoffmann GR. 1975. Selection for a large genetic duplication in salmonella typhimurium. Genetics 80:227–237. doi:10.1093/genetics/80.2.227
Tamames J. 2001. Evolution of gene order conservation in prokaryotes. Genome Biol. 2:research0020.1. doi:10.1186/gb-2001-2-6-research0020
Tillier ERM, Collins RA. 2000. Genome rearrangement by replication-directed translocation. Nat Genet. 26:195–197. doi:10.1038/79918
Treangen TJ, Abraham AL, Touchon M, Rocha EPC. 2009. Genesis, effects and fates of repeats in prokaryotic genomes. FEMS Microbiol Rev. 33:539–571. doi:10.1111/j.1574-6976.2009.00169.x
Trussart M, et al. 2017. Defined chromosome structure in the genome-reduced bacterium Mycoplasma pneumoniae. Nat Commun. 8:14665. doi:10.1038/ncomms14665
Uno R, Nakayama Y, Arakawa K, Tomita M. 2000. The orientation bias of Chi sequences is a general tendency of G-rich oligomers. Gene 259:207–215. doi:10.1016/S0378-1119(00)00430-3
Valens M, Thiel A, Boccard F. 2016. The MaoP/maoS site-specific system organizes the ori region of the E. coli chromosome into a Macrodomain Casadesús, J, editor. PLoS Genet. 12:e1006309. doi:10.1371/journal.pgen.1006309
Veetil RT, Malhotra N, Dubey A, Seshasayee ASN. 2020. Laboratory evolution experiments help identify a predominant region of constitutive stable DNA replication initiation Bowman, GR, editor. mSphere 5. doi:10.1128/mSphere.00939-19
Vernikos GS, Parkhill J. 2006. Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the Salmonella pathogenicity islands. Bioinformatics 22:2196–2203. doi:10.1093/bioinformatics/btl369
Vieira-Silva S, Rocha EPC. 2010. The systemic imprint of growth and its uses in ecological (Meta)genomics Moran, NA, editor. PLoS Genet. 6:e1000808. doi:10.1371/journal.pgen.1000808
Vulić M, Dionisio F, Taddei F, Radman M. 1997. Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. Proc Natl Acad Sci USA 94:9763–9767. doi:10.1073/pnas.94.18.9763
Wang X, Montero Llopis P, Rudner DZ. 2014. Bacillus subtilis chromosome organization oscillates between two distinct patterns. Proc Natl Acad Sci USA 111:12877–12882. doi:10.1073/pnas.1407461111
Xie Z, Tang H. 2017. ISEScan: automated identification of insertion sequence elements in prokaryotic genomes Hancock, J, editor. Bioinformatics 33:3340–3347. doi:10.1093/bioinformatics/btx433

Associate editor: Ruth Hershberg