Spatial Patterns of Substitutions in Bacterial Genomes

Daniella F Lato[1] and G Brian Golding[1]*

Paper Draft

November 9, 2019

Abstract

Increasing evidence supports the notion that different regions of a genome have different rates of molecular change. This variation is particularly evident in bacterial genomes where gene expression, essentiality, and mutation rate tend to decrease with increasing distance from the origin of replication. In multi-repliconic bacteria, smaller replicons are often comprised of less conserved genes with higher substitution rates compared to the larger replicons. Other genomic arrangements, for instance linear chromosomes, may therefore have varying substitution and other molecular trends. Here, we explore this phenomena by mapping substitutions to the genomes of *Escherichia coli*, *Bacillus subtilis*, *Streptomyces*, and *Sinorhizobium meliloti*, quantifying how many substitutions have occurred at each position in the genome. Preceding work has indicated that the number of substitutions should be higher as distance from the origin increases. Using a larger sample size and accounting for genome rearrangements, our analysis demonstrates that the number of substitutions instead decreased when moving away from the origin of replication in most of the bacteria analyzed. The exceptions to this were *Streptomyces* and a replicon of *S. meliloti* which had the number of substitutions increase as distance from the origin increased. dN, dS and ω were examined across all genes and spatially there was no significant trend between dN, dS, or ω and distance from the origin of replication. This study sheds light on the impact genomic structure and distance from the origin has on molecular trends in bacterial replicons and illustrates that these spatial trends are important to consider in any molecular evolutionary analysis.

[1] Department of Biology, McMaster University, Hamilton, ON, Canada
* Author for correspondence: G. Brian Golding, Department of Biology, Life Science Building, McMaster University, Hamilton, ON, Canada, L8S 4K1. Email: golding@mcmaster.ca.

Key Words: genome location, substitution, genomic structure, origin of replication

Introduction

Bacterial genomes are subject to the introduction and reorganization of genetic information through processes such as horizontal gene transfer (HGT), rearrangements, duplications, and inversions. These processes happen frequently and are important sources of genomic variation (Ochman et al. 2000; Epstein et al. 2014). It has been estimated that 10-15% of the *E. coli* genome has been horizontally introduced (Koski et al. 2001). DNA that is acquired through HGT can come from the same and/or different species of bacteria, allowing useful genes to be integrated into new genomes (Ochman et al. 2000). Genomic reorganization such as rearrangements, duplications, and inversions provide bacteria with the opportunity to fine tune existing gene expression, dosage, and replication. Bacteria can not escape genome reorganizations, and therefore incorporating reorganization potential is a crucial component of bacterial evolutionary analysis and can be done through multi-genome alignment programs such as `progressiveMauve` (Darling et al. 2010).

The genomic structure of a bacterial genome may provide new genomic landscapes capable of altering gene regulation. Here we will consider three main types of bacterial genomic structures: circular chromosomes, linear chromosomes, and multi-repliconic genomes. Secondary replicons of multi-repliconic bacteria are hypothesized to predominantly contain niche specific genes (Heidelberg et al. 2000; Egan et al. 2005). These replicons generally contain genes that have distinctive rates of evolution and selection acting upon them (Heidelberg et al. 2000). This allows the bacteria to thrive in rapidly changing environments, with varying molecular traits associated with each replicon (Heidelberg et al. 2000; Cooper et al. 2010; Morrow and Cooper 2012; Galardini et al. 2013).

Although thousands of bacterial genomes have been sequenced for bacteria with different genomic structures, majority of these genomes are incomplete and composed of scaffolds or contigs. For this analysis, a complete genome, free of gaps or contigs, was necessary to accurately track substitutions and their genomic locations. Incomplete genomes would have gaps in genome positions, leaving missing information about substitutions for these segments of sequence. Therefore, we wished to consider only complete genomes.

A previous multipartite genome investigation with four genomes of *Burkholderia* has shown that the primary chromosome is highly conserved and has higher gene expression compared to the secondary replicons which are less conserved (Morrow and Cooper 2012). A similar study using a minimum of four genomes from *Burkholderia*, *Vibrio*, *Xanthomonas*, and *Bordetella* also discovered that the primary chromosomes are conserved, with higher gene expression compared to the secondary replicons (Cooper et al. 2010). Additionally, primary chromosomes have lower substitution (Morrow and Cooper 2012) and evolutionary (Cooper et al. 2010) rates compared to the secondary replicons. Housekeeping genes usually reside on the primary chromosome, and the secondary replicons usually contain parts of the accessory genome, which could account for the substitution and evolutionary rate differences between primary and secondary replicons (Cooper et al. 2010; Flynn et al. 2010; Morrow and Cooper 2012). It has been suggested that the differences in gene content between replicons of multi-repliconic bacteria could be due to delays in replication (Flynn et al. 2010; Morrow and Cooper 2012). To maintain synchronization, due to the offset of different sequence lengths between primary and secondary replicons, the secondary replicons begin replication after the primary chromosome (Flynn et al. 2010; Morrow and Cooper 2012).

Prior research on molecular trends when moving from the origin of replication to the terminus have determined that gene expression is increased near the origin (Sharp et al. 2005; Couturier and Rocha 2006) and that genes become less conserved with increasing distance from the origin (Couturier and Rocha 2006). Analyses with fewer bacterial species have replicated these results and found that gene expression decreases with increasing distance from the origin (Morrow and Cooper 2012)(one species) and substitution rates (non-synonymous (dN), synonymous (dS), and dN/dS) increase with distance from the origin of replication (Cooper et al. 2010) (4 species) (Morrow and Cooper 2012). However, a few studies found no correlation between distance from the origin of replication and the frequencies of mutations, but they did find mutation rate to vary with position along the *E. coli* chromosome Juurik et al. (2012) or intermediate positions had a higher non-synonymous mutation rate than positions farther from the origin (Ochman 2003). It is speculated that the recombination rate moving away from the origin of replication is less regulated, genes located near the terminus have more variation and are less conserved as those near the origin of replication (Flynn et al. 2010; Sharp et al. 1989). Additionally, genes found withing the core genome are typically located near the origin of replication, while genes associated with the accessory genome are found near the terminus (Couturier and Rocha 2006; Sharp et al. 2005; Flynn et al. 2010). The placement of these two gene categories may explain why near the origin, gene expression and essentiality are high and mutation rate is low (Couturier and Rocha 2006; Flynn et al. 2010; Sharp et al. 2005). However, there are cases in *Rhodobacteraceae* where core genes in some species were concentrated near the terminus, not the origin of replication (Kopejtka et al. 2019). Other species had a mosaic patter of core genes dispersed throughout the genome (Kopejtka et al. 2019). It is speculated that other factors such as HGT, phage insertion, and replication may be responsible for the conflicting placement of core genes in various *Rhodobacteraceae* species (Kopejtka et al. 2019).

There are a number of additional factors that are dependent on distance from the origin such as transposon insertion events (Gerdes et al. 2003), gene order (Mackiewicz et al. 2001), number of replication forks (Couturier and Rocha 2006), and nucleotide composition (Mackiewicz et al. 1999; Karlin 2001). These phenomena are important to consider when analyzing molecular trends with respect to distance from the origin of replication.

The majority of these studies used an average of 3 genomes per bacteria analyzed (Sharp et al. 2005; Couturier and Rocha 2006; Flynn et al. 2010; Cooper et al. 2010; Morrow and Cooper 2012) and failed to analyze secondary replicons of multipartite genomes (Sharp et al. 2005; Couturier and Rocha 2006; Flynn et al. 2010). This work examines the spatial substitution trends in *Escherichia coli* (6 genomes), *Bacillus subtilis* (7 genomes), *Streptomyces* (6 genomes), and *Sinorhizobium meliloti* (6 genomes). These bacteria contain genomic structures that range from a single circular chromosome (*E. coli* and *B. subtilis*), linear chromosome (*Streptomyces*), and a multi-repliconic genome (*S. meliloti*). For each, the effects of genomic rearrangements and number of replicons were taken into account. We show here that for the majority of the replicons investigated, the number of substitutions decrease when moving away from the origin of replication towards the terminus. The exceptions were the second largest replicon of *S. meliloti* (pSymB) and the non-protein coding regions of *Streptomyces* and *E. coli*, where the number of substitutions increase with increasing distance from the origin. Possible causes and consequences of this pattern are discussed.

Materials and Methods

Sequence Data

Whole genomes of different strains of *E. coli*, *B. subtilis*, and *S. meliloti*, as well as various species of *Streptomyces* were downloaded from NCBI. Access date and accession numbers are given in Supplementary Table S1. These bacteria inhabit a variety of different living environments and have contrasting genomic structures, providing a well rounded sample for this analysis. Although *E. coli*, *B. subtilis*, and *Streptomyces* contain small plasmids, they are not considered multi-repliconic bacteria and therefore their plasmids were not included in this analysis. *S. meliloti* is a multi-repliconic bacteria and its two large secondary replicons were included in the analysis (pSymA and pSymB). The replicons of *S. meliloti* are known to differ in genetic content, and therefore, all analyses were performed on each individual replicon of *S. meliloti*. The genomes used in each bacterial analysis attempted to consist of as many reference genomes as possible at the time of the data collection (Supplementary Table: S1).

Sequence Alignment

Alignments were performed using `progressiveMauve` (Darling et al. 2010) to group the alignment into locally colinear blocks (LCBs). This method allows for rearrangements, duplications and inversions to be taken into account. A LCB is frequently found at different genomic positions in each of the taxa analyzed. These segments of sequence must be similar between at least two of the taxa, but not necessarily between all of them. To obtain accurate information for subsequent analysis, only the subset of LCBs that were present in all taxa were considered. Each locally colinear block was then re-aligned with MAFFT (Katoh et al. 2014) to obtain a more accurate local alignment. Instances in the alignment where a gap was present in minimally one taxa were removed from the remainder of the analysis.

Phylogenetic Trees

Rearrangements, duplications, and inversions that happen frequently throughout bacterial genomes must be considered when analyzing spatial genomic trends. Phylogenetic trees were created to assist in mapping the evolutionary history of large scale and local DNA rearrangements onto the phylogeny. These trees were used to determine the number of substitutions and record the genomic location of the substitution for each respective replicon. Each of the LCBs specified by `progressiveMauve` were combined to create a single large "super sequence" and a series of programs (`seqboot`, `dnadist`, `neighbor`, `consense`, and `dnaml`) from the `PHYLIP` (Felsenstein 2008) package were used to create a phylogenetic tree for each bacterial replicon. Phylogenetic consensus trees with bootstrap support values can be found in Supplementary Figures:S2-S7.

An SH-test (Shimodaira and Hasegawa 1999; Goldman et al. 2000) was performed to determine if there was a significant difference between the "super sequence" tree topology for each bacterial replicon using all LCBs, and the tree topology of each LCB individually. The number of LCBs that were determined to have a topology that were statistically similar to the "super sequence" tree is summarized in Supplementary Table S2. Any LCBs that had a topology that was significantly different (at the 5% significance level) from the "super sequence" topology was removed from the remainder of the analysis.

Origin of Replication and Bidirectionality

For each bacteria demarked, the beginning of the origin of replication was denoted as the beginning of the *oriC* region for the chromosomal replicons, and the beginning of the *repC* (Pinto et al. 2011) region for the secondary replicons of *S. meliloti* (Supplementary Table S2). This origin of replication position was calibrated to be the beginning of the genome, position 1, and remaining positions in the genome were all scaled around this origin of replication taking into account the bidirectional nature of bacterial replication (Figure 1).

The terminus of replication was determined using the Database of Bacterial Replication Terminus (Kono et al. 2011), which uses the prediction of *dif* sequences, and attempts to predict the *dif* sequences normally found at the termini (Clerget 1991; Blakely et al. 1993). These were therefore used as a proxy for the termini location. For pSymA and pSymB of *S. meliloti* the terminus is not listed in the database, so the midpoint between the origin of replication and the end of the genome was used as the terminus location. Replication in the linear chromosome of *Streptomyces* begins at the origin of replication, located to the right of the middle of the replicon (Heidelberg et al. 2000), and terminates at each end of the chromosome arms (Heidelberg et al. 2000)(Supplementary Table S3).

We have chosen a single base to represent the origin of replication. In reality, the origin of replication is often multiple base pairs long. To determine the effect of the exact location of the origin, we performed permutation tests shuffling the *oriC* position by 10,000bp increments in each direction from the original origin (Supplementary Table: S3) to a maximum of 100,000bp in each direction.

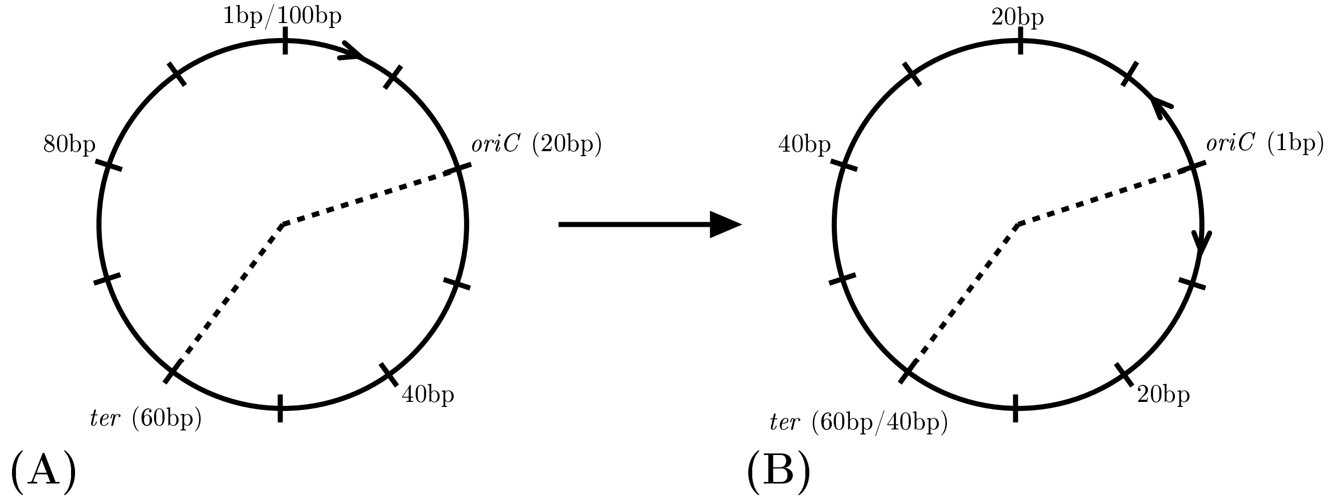


Figure 1: Schematic of the transformation used to scale the positions in the genome to the origin of replication and account for bidirectionality of replication. Circle (A) represents the original replicon genome without any transformation. Circle (B) represents the same replicon genome after the transformation. The origin of replication is denoted by "*oriC*" and the terminus of replication is denoted by "*ter*". The dashed line represents the two halves of the replicon separate by replication. The replicon genome in this example is 100 base pairs in length. Every 10 base pairs is denoted by a tick on the genome. The origin in (A) is at position 20 in the genome and is transformed in (B) to become position 1. The terminus is at position 60 in (A) and position 60 and 40 in (B). The terminus has two positions in (B) depending on which replicon half is being accounted for. If the replication half to the right of the origin is considered, the terminus will be at position 40. If the replication half to the left of the origin is considered, the terminus will be at position 60. Position 40 in (A) becomes position 20 in (B). Position 80 in (A) becomes position 40 in (B), because of the bidirectional nature of bacterial replication.

These results showed that moving the origin of replication does not affect the results of the analysis (Supplementary Table: S4).

Ancestral Reconstruction

To track genome rearrangements, nucleotide substitutions and genomic positions were reconstructed in extinct ancestors on the given phylogenies. We have used the PAML (Yang 1997) package of programs, with slight modification, to reconstruct genome location and substitutions in extinct ancestors (Figure 2).

Nucleotide Substitutions

The baseml program in the `PAML` package (Yang 1997) was used to determine single nucleotide substitutions within each of the alignments. This program determined the ancestral state of each nucleotide in the alignment at each node in the replicon's respective phylogenetic tree (Figure 2). Multiple substitutions at one site were allowed and accounted for as two separate substitutions. Any nucleotides, or columns, in the alignment that had at least one gap present were not used in the analysis because the baseml program inaccurately classifies substitutions when a gap is involved. These gaped positions were categorized as missing data.

Coding and Non-Coding Substitutions

To further specify the ancestral and extant substitutions present, we have classified the substitutions as part of protein coding or non-protein coding regions of the genome. The protein coding regions of the genome were determined by the known and predicted transcribed protein coding annotation found in the GenBank files of each of the bacteria, all other regions of the genome were denoted as non-protein coding (Supplementary Table S5). A custom python script was used to associate each of these regions with their genomic positions and determine if the ancestral and extant substitutions were found in each region. This classification was performed on each replicon in the alignment and the remainder of the methods were performed on the protein coding or non-protein coding regions of each replicon separately.

Genomic Position

Genomic rearrangements were accounted for using the genome locations specified by `progressiveMauve` to determine the ancestral genome positions of each taxa (Figure 2). These locations were inferred for each nu-

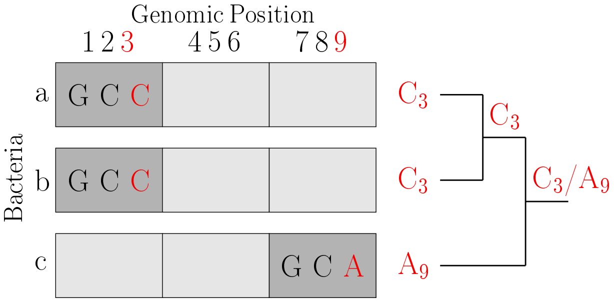


Figure 2: Schematic of the ancestral reconstruction of both the nucleotide and genomic position. Each horizontal row of rectangles represents three hypothetical bacterial genomes (a, b, c). The genomic position is indicated at the top of the diagram. The phylogenetic tree showing the relationship between all three bacteria is pictured on the right of the diagram. The dark grey rectangle denotes the genomic segment of interest. In bacteria (a) and (b), this segment is located at genomic positions 1-3. In bacteria (c), this segment is located at genomic positions 7-9. Within this genomic region of interest there is a substitution where the nucleotides changed from C → A, this is highlighted in red. This substitution is at position 3 in bacteria (a) and (b), and in position 9 in bacteria (c). This is depicted by the values (C_3) and (A_9). The ancestral reconstruction process in this analysis can be seen at the inner nodes of the phylogenetic tree by the values (C_3). The most parsimonious reconstruction of the sequence and associated genomic position is having the value (C_3) present at the ancestor of bacteria (a) and (b). The ancestral node of all three bacteria would have a reconstruction of the sequence and associated genomic position of (C_3 / A_9). In this situation where there is a "tie" for two most parsimonious options, the option with the highest likelihood estimate would be chosen using maximum-likelihood methods (see (Yang 1997) for more details). This would mean that in bacteria (c) there was a substitution from C → A which is also associated with a genomic position of 9.

cleotide in the alignment. We modified the codeml (Yang 1997) program from the PAML package to reconstruct the ancestral genome positions at each node within the phylogenetic tree (Supplementary Trees: S2- S7) of each respective replicon for each position in the alignment (Figure 2).

Each branch in the tree possesses information on how each nucleotide in the alignment has moved through the genome to its current position in each of the taxa (Figure 2). Therefore, each segment of sequence in the genome has the opportunity to be present in one position in the genome of one taxa, and a completely different position in another taxa (Figure 2). The density of ancestral substitutions in both protein coding and non-protein regions across each bacterial replicon can be seen in Figures 3 and 4. These Figures provide information on the frequency of substitutions in relation to the distance from the origin of replication while also taking into account the bidirectionality of bacterial replication (See Methods: Origin of Replication and Bidirectionality).

For this portion of the analysis each genomic position was considered unique and distinct, including positions that were separated by one base pair. We performed a supplementary analysis to determine if clustering genomic positions based on how many base pairs separate them, would significantly alter the overall spatial substitution results (See Supplemental for more details). We determined that considering each genomic position to be unique and distinct or clustering the positions did not alter the results.

Logistic Regression

The binary nature of the data is ideal for logistic regression to determine the statistical significance of substitution and position trends at both protein coding and non-protein coding regions of the genome in each bacterial replicon (Table 1). Multiple substitutions at any given genomic location were allowed. The substitution data had to be within the first quartile minus 1.5 times the interquartile range, and the third quartile plus 1.5 times the interquartile range. Any points outside this range were classified as outliers and were subsequently removed.

Average Number of Substitutions

The average number of substitutions per base of each bacterial replicon was calculated as the average number of substitutions per 10kb region divided by the farthest distance from the origin of replication, while accounting for bidirectionality of replication. The results are summarized in Table 2.

Selection

Separating the substitutions data into substitutions that occur within protein coding and non-protein coding regions allows us to determine how selection may be acting differently on each of these sections. Within the protein coding regions of the genome, we wanted to observe how selection may be acting on each of the genes in the various bacterial replicons. Calculating the synonymous (dS) and non-synonymous (dN) substitution rates for each gene allows for an in depth analysis of the selective pressures throughout the genome while accounting for rearrangements between the bacterial taxa. We can then relate this information to the location of the genes in the genome and determine if there are any trends between selection and distance from the origin. It has been found previously that genes closest to the origin of replication are conserved (Couturier and Rocha 2006) and tend to be a part of the core genome (Sharp et al. 2005; Couturier and Rocha 2006; Flynn et al. 2010). We therefore expect genes closer to the origin to have fewer substitutions and therefore lower values for dS and dN.

Calculating dN and dS

The CODEML program in the PAML package (Yang 1997) was used to calculate the synonymous (dS) and non-synonymous (dN) substitution rates as well as estimate a value for ω. This program estimates the dN and dS values for each gene of each replicon as mentioned in the previous section. No molecular clock was used and the basic substitution model where each nucleotide is able to be substituted for another at the same rate was used. The varying nucleotide models have minimal impact on the dN and dS calculations because the overall number of synonymous and non-synonymous substitutions per site were small. We then used the per gene dN, dS, and ω values to calculate an arithmetic average of dN, dS, and ω for each replicon weighted by the length of each gene. Any genes where both dN and dS or dS were equal to zero were removed from the weighted ω calculation. A summary of the average dN and dS results are found in Table 3.

Results

Logistic Regression

The number of substitutions decreased when moving away from the origin of replication for the protein coding regions of *E. coli*, *B. subtilis*, the chromosome of *S. meliloti* and pSymA of *S. meliloti*. Additionally, the non-protein coding regions of pSymA and the chromosome of *S. meliloti* also had the number of substitutions decrease when moving away from the origin of replication. This implies that the area near the terminus of replication in these replicon sections had less substitutions than the area near the origin of replication. We could not detect a significant logistic regression coefficient estimate for the non-protein coding sections of the *E. coli* and *B. subtilis*. pSymB of *S. meliloti* and *Streptomyces* showed the opposite trend from the other bacterial replicons. pSymB and *Streptomyces* had a positive logistic regression coefficient for both the protein coding and non-protein coding segments of the genomes. The positive coefficient indicates that in pSymB and *Streptomyces* there is an decreased number of substitutions present near the origin of replication compared to the terminus. All logistic regression and supporting statistical information for the substitution trends are found in Table 1. Additionally, majority of the bacteria in this study proportionally had similar number of substitutions within non-protein coding regions than protein coding regions (Supplementary Table: S8).

Table 1: Logistic regression analysis of the number of substitutions along all positions of the genome of the respective bacteria replicons. These genomic positions were split up into the coding and non-coding regions of the genome. Grey coloured boxes indicate a negative logistic regression coefficient estimate. All results are statistically significant. Logistic regression was calculated after the origin of replication was moved to the beginning of the genome and all subsequent positions were scaled around the origin accounting for bidirectionality of replication. All results are marked with significance codes as followed: < 0.001 = '***', $0.001 < 0.01$ = '**', $0.01 < 0.05$ = '*', > 0.05 = 'NS'.

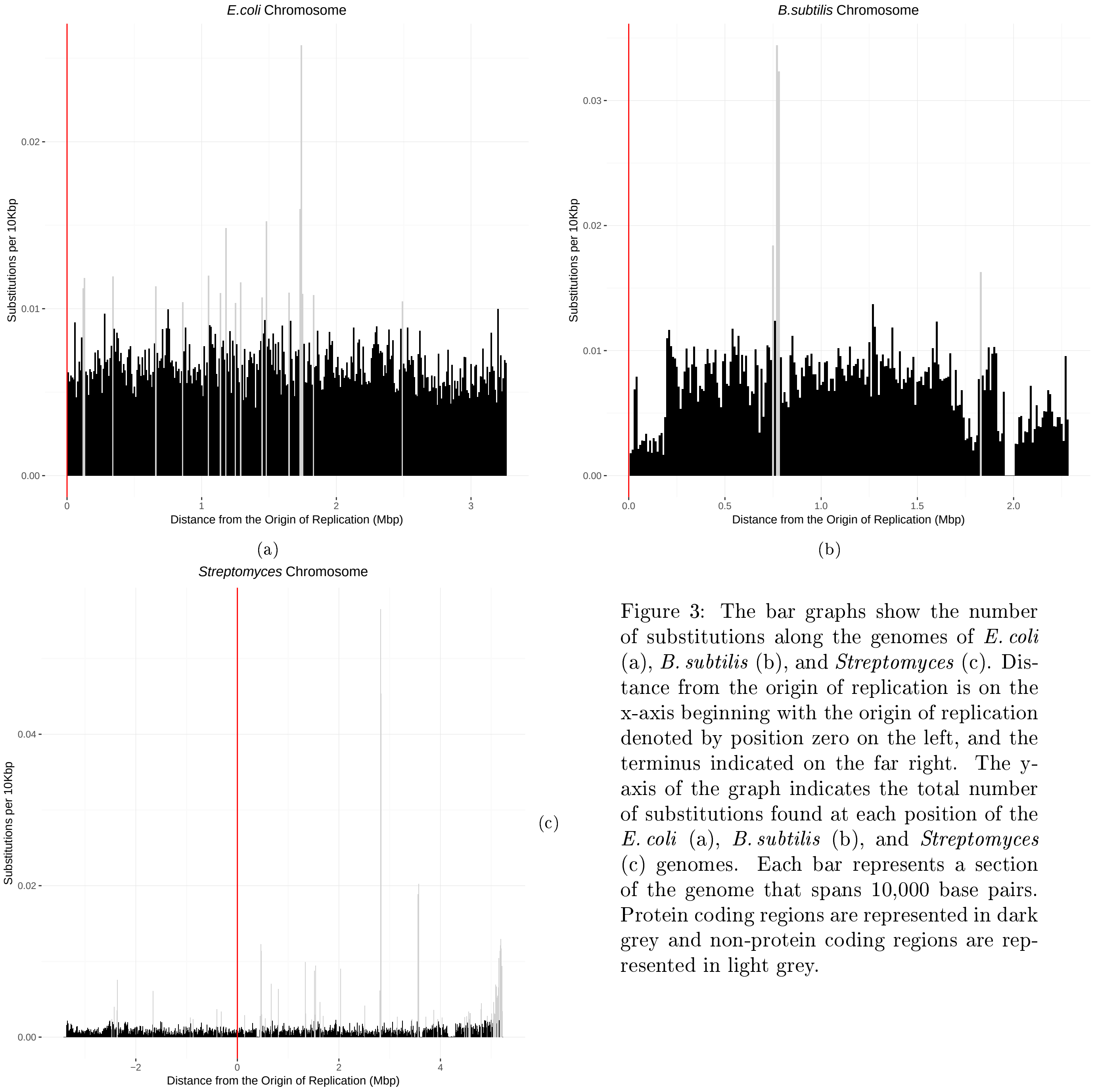
Bacteria and Replicon	Protein Coding Sequences
E. coli Chromosome	-2.98×10^{-8}***
B. subtilis Chromosome	-8.06×10^{-8}***
Streptomyces Chromosome	1.10×10^{-7}***
S. meliloti Chromosome	-4.32×10^{-7}***
S. meliloti pSymA	-5.18×10^{-7}***
S. meliloti pSymB	1.76×10^{-7}***

Ancestral Reconstruction

The number of substitutions (both ancestral and extant) in protein coding and non-protein coding regions of the genomes are reflected in the graphs for each of the replicons analyzed (Figures: 3 and 4). For all of the bacterial replicons there is a higher proportion of substitutions in the non-protein coding region compared to the proportion of substitutions in the protein coding regions (Supplementary Table S8).

Average Number of Substitutions

Overall there appears to be no relationship between replicon length and average number of substitutions. The largest and third largest genomes respectively, *B. subtilis* and *Streptomyces*, have the largest average number of substitutions (Table 2). Although *E. coli* is the second largest genome, it has the third lowest average number of substitutions. The smaller replicons of *S. meliloti* - pSymA and pSymB - have faster substitution rates compared to the larger chromosomal replicon of the same bacteria. pSymB has a slightly faster substitution rate compared to pSymA. These results are consistent with the general knowledge of the gene types between the smaller replicons of *S. meliloti* and the chromosome. The smaller replicons are expected to evolve more quickly and therefore should have a faster average substituion rate. It is curious that pSymB has a slightly higher average substitution rate compared to pSymA because pSymA has been shown to be more variable than pSymB Galardini et al. 2013.

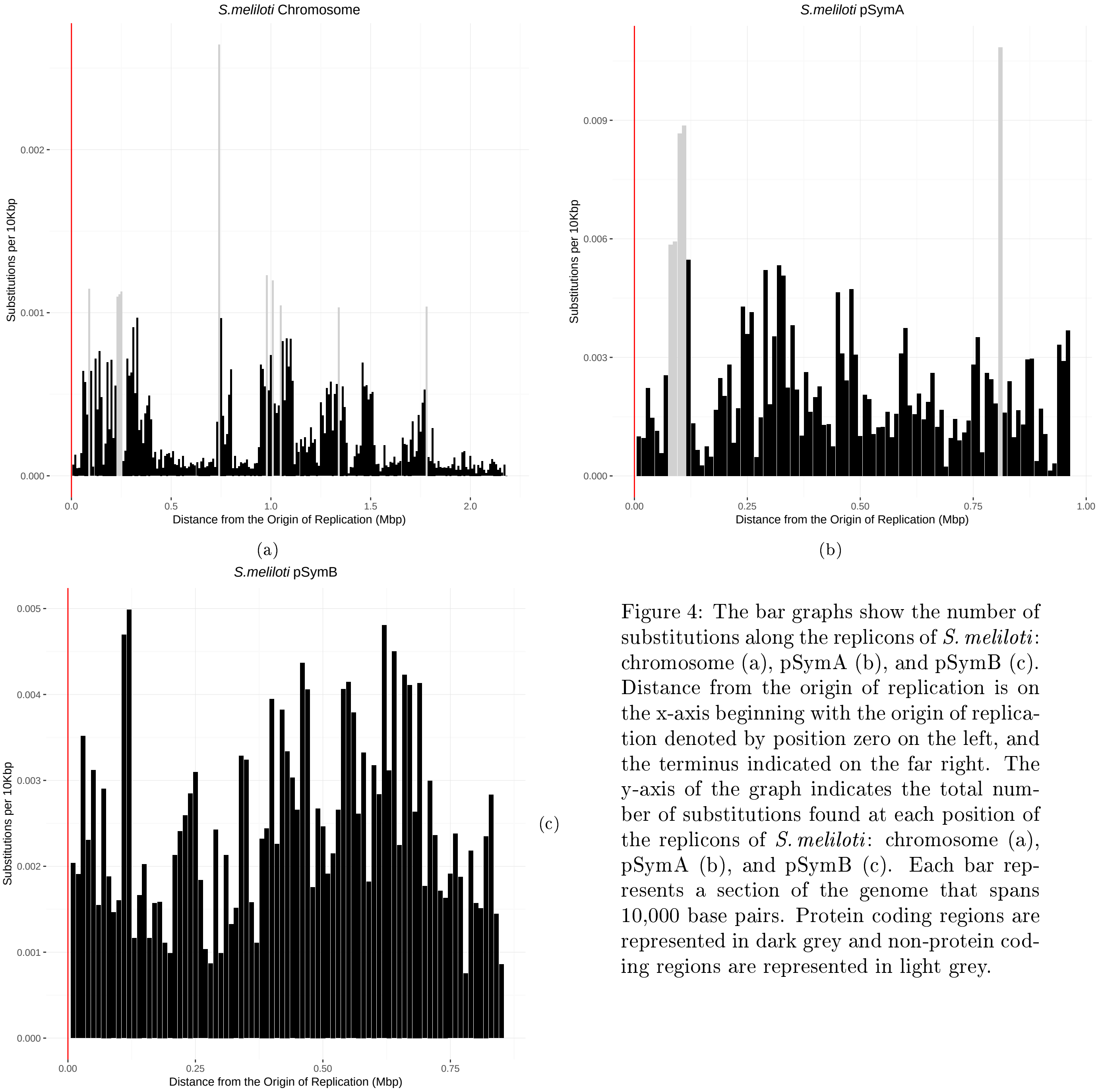


(a)

(b)

(c)

Figure 3: The bar graphs show the number of substitutions along the genomes of *E. coli* (a), *B. subtilis* (b), and *Streptomyces* (c). Distance from the origin of replication is on the x-axis beginning with the origin of replication denoted by position zero on the left, and the terminus indicated on the far right. The y-axis of the graph indicates the total number of substitutions found at each position of the *E. coli* (a), *B. subtilis* (b), and *Streptomyces* (c) genomes. Each bar represents a section of the genome that spans 10,000 base pairs. Protein coding regions are represented in dark grey and non-protein coding regions are represented in light grey.



(a)

(b)

(c)

Figure 4: The bar graphs show the number of substitutions along the replicons of *S. meliloti*: chromosome (a), pSymA (b), and pSymB (c). Distance from the origin of replication is on the x-axis beginning with the origin of replication denoted by position zero on the left, and the terminus indicated on the far right. The y-axis of the graph indicates the total number of substitutions found at each position of the replicons of *S. meliloti*: chromosome (a), pSymA (b), and pSymB (c). Each bar represents a section of the genome that spans 10,000 base pairs. Protein coding regions are represented in dark grey and non-protein coding regions are represented in light grey.

Table 2: Average number of substitutions calculated per base across all bacterial replicons. Outliers and missing data was not included in the calculation.

Bacteria and Replicon	Average Number of Substitutions per bp
E. coli Chromosome	1.81×10^{-4}
B. subtilis Chromosome	7.73×10^{-4}
Streptomyces Chromosome	7.84×10^{-4}
S. meliloti Chromosome	2.42×10^{-5}
S. meliloti pSymA	1.58×10^{-4}
S. meliloti pSymB	5.63×10^{-4}

Selection

All bacterial replicons had average per gene and per genome *dS* values that were higher than the respective *dN* values. The only exception to this was *Streptomyces* which had a per gene and per genome *dN* value that was higher than the *dS* value. Complete selection results can be found in Table 3 with distribution box plots found in Figure **??**.

Table 3: Weighted averages calculated for each bacterial replicon on a per genome basis using the gene length as the weight. Arithmetic mean calculated for the per gene averages for each bacterial replicon.

	Gene Average			Genome Average		
Bacteria and Replicon	dS	dN	ω	dS	dN	ω
E. coli Chromosome	0.2843	0.0145	0.0574	0.3072	0.0181	0.0663
B. subtilis Chromosome	0.6241	0.0353	0.0922	0.5708	0.0335	0.0841
Streptomyces Chromosome	0.0854	0.0042	0.7384	0.0945	0.0062	0.7676
S. meliloti Chromosome	0.0162	0.0011	0.1465	0.0171	0.0012	0.1158
S. meliloti pSymA	0.0865	0.0122	0.2250	0.0839	0.0106	0.2125
S. meliloti pSymB	3.2602	0.0256	0.3878	0.1436	0.0100	0.1943

Discussion

The logistic regression results for both the proetin coding and non-protein coding regions of *E. coli*, *B. subtilis*, the chromosome of *S. meliloti* and pSymA of *S. meliloti* indicate that the number of substitutions decreases when moving away from the origin of replication. In both the non-protein coding and protein coding regions of pSymB of *S. meliloti* and *Streptomyces*, the logistic regression results imply that the number of substitutions increases when moving away from the origin of replication. The exception to this was the non-protein coding regions of *E. coli* and *B. subtilis* where we did not observe a significant correlation between distance from the origin and number of substitutions. The non-protein coding regions of all bacteria are smaller than their protein coding regions and usually with a proportionally lower number of substitutions (Supplementary Table S8). It has been observed that pSymB of *S. meliloti* has many similar properties to the chromosome of *S. meliloti* (Charles and Finan 1991; Finan et al. 2001; Wong and Golding 2003). A number of previous studies have complementary results regarding increasing substitution trends of bacterial replicons. These studies observed gene expression, gene essentiality and mutation rate decrease as the positions become further from the origin of replication (Prescott and Kuempel 1972; Sharp et al. 2005; Morrow and Cooper 2012; Galardini et al. 2013). Genes that are less essential and often expressed less tend to evolve quickly compared to more conserved genes with higher expression levels (Sharp et al. 1989). pSymB of *S. meliloti* has been known to house essential genes (Cooper et al. 2010; Morrow and Cooper 2012), which may explain the increased number of substitutions near the terminus.

The chromosome of *Streptomyces* contains the majority of it's essential genes near the origin of replication (Bentley et al. 2002; Kirby 2011). These genes should have decreased number of substitutions and therefore coincide with the increasing substitution rate when moving away from the origin of replication.

When considering selective pressures acting upon these replicons, *Streptomyces* is the only replicon that has an average ω value larger than 1, indicating sites in the protein coding regions of this replicon may be under positive selection. The other replicons have an average ω value less than 1, indicating sites in the protein coding regions of these replicons may be under purifying selection. For the *Streptomyces* analysis the lack of available strain data may be contributing to why substitutions in these species are undergoing positive selections because the species in the *Streptomyces* analysis are less closely related than the strains from the other bacterial analysis, therefore they would benefit from allowing beneficial alleles to propagate in the population. The other taxa in the *E. coli*, *B. subtilis*, and *S. meliloti* analysis are all from the same strain and therefore more closely related.

The selection results illustrate that majority of the bacteria are undergoing purifying selection within the protein coding regions of the replicons as expected. This therefore does not give us any further insight into the interesting decreasing substitution trend found in majority of the replicons in this analysis.

Not sure what to say about the violin plots or if I need to say anything at all about them? thoughts?

Molecular composition, gene content, and replication may all be factors contributing to the curious decreasing number of substitutions with increasing genomic distance found in this study.

Molecular architecture such as GC content or nucleotide composition varies along a bacterial genome.

Both of these trends have been found to significantly change around the origin of replication and terminus (Mackiewicz et al. 1999; Ikeda et al. 2003). This disparity in composition between the origin and terminus may be a contributing factor in explaining why in this study we have found a higher number of substitutions near the origin of replication and warrants further investigation.

These differences in molecular composition can also impact where genes are located along a replicon. Core genes are typically found near the origin of replication (Couturier and Rocha 2006; Cooper et al. 2010; Morrow and Cooper 2012; Flynn et al. 2010) and accessory genes are usually found near the terminus (Couturier and Rocha 2006; Cooper et al. 2010; Flynn et al. 2010; Morrow and Cooper 2012). Determining the distribution and placement of the core and accessory genes in *E. coli*, *B. subtilis*, pSymA and the chromosome of *S. meliloti* could elucidate why these replicons appear to have a higher number of substitutions near the origin of replication. Additionally, potential genomic and pathogenicity islands have been found near the origin of replication in *Mycobacterium tuberculosis* and *Haloquadratum walsbyi* (Karlin 2001; Mira et al. 2010). These islands were found to have genomic signatures such as codon bias, that deviated from the rest of the genome (Karlin 2001). Deviations in these genomic signatures may extend to substitution rates and provide another potential explanation as to why most of the replicons in this study has an increased number of substitutions near the origin of replication.

Part of the reason for distinct placement of the core accessory genes across the genome is speculated to be in part due to the nature of replication. Many translocations happen at replication forks as they advance along the chromosome(Tillier and Collins 2000; Mackiewicz et al. 2001). If these replication forks were concentrated near the origin of replication, this could increase the number of translocations present in that area, which would provide an opportunity for new genomic signatures such as substitution rate to arise. This could be a reason why we are seeing an increased number of substitutions near the origin of replication.

Rearrangements, inversions, duplications, and HGT all play a major role in shaping bacterial replicons. These affect all aspects of the genome such as gene content, expression, order and substitutions. Some studies have found that the density of transposon insertion events peaks at the origin of replication and is at a minimum at the terminus in *E. coli* (Gerdes et al. 2003). Once again, the differences in various genomic signatures caused by genome reorganization such as transposon insertion and duplication events, may be a justification for the high number of substitutions seen in this analysis.

The multi-repliconic nature of *S. meliloti* appears to have a small effect on the overall spatial substitution trends of each replicon. For the smaller replicon pSymA and the chromosome, the number of substitutions decreased as distance from the origin increases, while pSymB of *S. meliloti* showed the opposite trend. The *S. meliloti* chromosome has a slower substitution rate than pSymA and pSymB. This may be due to an over representation of highly expressed/essential genes located on the chromosome. Smaller replicons, like pSymA, have many genes that are less conserved (Cooper et al. 2010; Morrow and Cooper 2012), and used for local environmental adaptation (Medini et al. 2008). Genes that are not crucial to the survival of an organism may be subject to increased substitutions. Overall, determining the placement of essential and non-essential genes in each of the bacteria analyzed might assist in explaining the substitution trends and guide understanding of bacterial genome evolution.

The circularity or linearity of a bacterial chromosome appears to have no effect on the spatial substitution trends. However, in the *Streptomyces* substitution analysis there appears to be a bell shaped distribution (Figure 3). This implies that most substitutions are happening near the origin of replication, located in the middle portion of the linear chromosome. This central region of the *Streptomyces* genome contains many conserved essential genes (Bentley et al. 2002). It is curious that these core genes would be undergoing more substitutions than the less conserved genes at the chromosome termini. Bentley et al. (2002) identified laterally transferred genes that were potentially acquired recently. Some of these genes were found near the *oriC* region. If the genes or regions near the origin were recently laterally acquired, they may have increased substitutions because of their new location. Fitting a non-linear model might improve classification of *Streptomyces*'s spatial substitution trends.

Determining the spatial trends of genome descriptors such as gene essentiality and number of substitutions gives insight into bacterial genome evolution. Further in-depth analysis of other molecular trends in each segment of bacterial genomes is necessary. Determining how the number of substitutions are distributed spatially throughout bacterial genomes broadens our understanding of their evolution.

Conclusions

For most replicons considered in this study, the number of substitutions decrease with increasing distance from the origin of replication. The exceptions to this are pSymB of *S. meliloti* and *Streptomyces* where the number of substitutions increase when moving away from the origin of replication. These spatial substitution results combined with the selection results can be used to determine if all bacteria possess the same evolutionary patterns. Spatial location of essential genes and functional classification of those genes will assist in answering questions about the evolution of bacteria.

Supplementary Material

Supplementary Figures S1- S13 and Tables S1- S8 are available at Genome Biology and Evolution online (http://www.oxfordjournals.org/our_journals/gbe/). Further supplemental code and information are available on `GitHub` at `www.github.com/dlato/Spatial_Patterns_of_Substitutions`.

Acknowledgments

We thank Caitlin Simopoulos for comments on the manuscript. We thank NSERC for funding for this project (grant XXXXXX to GBG).

References

Bentley S D, Chater K F, Cerdeno-Tarraga A M, Challis G L, Thomson N R, James K D, Harris D E, Quail M A, Kieser H, Harper, et al. (2002). Complete genome sequence of the model actinomycete <i>Streptomyces coelicolor A3(2)</i>. Nature 417(6885), 141–147.

Blakely G, May G, McCulloch R, Arciszewska L K, Burke M, Lovett S T, and Sherratt D J (1993). Two related recombinases are required for site-specific recombination at dif and cer in <i>E. coli K12</i>. Cell 75(2), 351–361.

Charles T C and Finan T M (1991). Analysis of a 1600-kilobase <i>Rhizobium meliloti</i> megaplasmid using defined deletions generated in vivo. Genetics 127(1), 5–20.

Clerget M (1991). Site-specific recombination promoted by a short DNA segment of plasmid R1 and by a homologous segment in the terminus region of the <i>Escherichia coli</i> chromosome. New Biol 3(8), 780–788.

Cooper V S, Vhor S H, Wrocklage S C, and Hatcher P J (2010). Why genes evolve faster on secondary chromosomes in bacteria. PLoS Comp Biol 6(4), e1000732.

Couturier E and Rocha E P (2006). Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes. Mol Microbiol 59(5), 1506–1518.

Darling A E, Mau B, and Perna N T (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS one 5(6), e11147.

Egan E S, Fogel M A, and Waldor M K (2005). Divided genomes: negotiating the cell cycle in prokaryotes with multiple chromosomes. Mol Microbiol 56(5), 1129–1138.

Epstein B, Sadowsky M J, and Tiffin P (2014). Selection on horizontally transferred and duplicated genes in Sinorizobium (Ensifer), the root-nodule symbionts of Medicae. Genome Biol Evol 6(5), 1199–1209.

Felsenstein J (2008). Comparative methods with sampling error and within-species variation: contrasts revisited and revised. Am Nat 171(6), 713–725.

Finan T M, Weidner S, Wong K, Buhrmester J, Chain P, Vorhölter F J, Hernandez-Lucas I, Becker A, Cowie A, and Gouzy J (2001). The complete sequence of the 1,683-kb pSymB megaplasmid from the N2-fixing endosymbiont Sinorhizobium meliloti. Proc Natl Acad Sci 98(17), 9889–9894.

Flynn K M, Vohr S H, Hatcher P J, and Cooper V S (2010). Evolutionary rates and gene dispensability associate with replication timing in the archaeon <i>Sulfolobus islandicus</i>. Genom Biol Evol 2, 859–869.

Galardini M, Pini F, Bazzicalupo M, Biondi E G, and Mengoni A (2013). Replicon-dependent bacterial genome evolution:the case of Sinorhizobium meliloti. Genome Biol Evol 5(3), 542–558.

Gerdes S Y, Scholle M D, Campbell J W, Balazsi G, Ravasz E, Daugherty M D, Somera A L, Kyrpides N C, Anderson I, Gelfand M S, et al. (2003). Experimental determination and system level analysis of essential genes in <i>Escherichia coli MG1655</i>. J Bacteriol 185(19), 5673–5684.

Goldman N, Anderson J P, and Rodrigo A G (2000). Likelihood-based tests of topologies in phylogenetics. System Biol 49(4), 652–670.

Heidelberg J F, Eisen J A, Nelson W C, Clayton R A, Gwinn M L, Dodson R J, Haft D H, Hickey E K, Peterson J D, Umayam L, et al. (2000). DNA sequence of both chromosomes of the cholera pathogen <i>Vibrio cholerae</i>. Nature 406(6795), 477–483.

Ikeda H, Ishikawa J, Hanamoto A, Shinose M, Kikuchi H, Shiba T, Sakaki Y, Hattori M, and Omura S (2003). Complete genome sequence and comparative analysis of the industrial microorganism Streptomyces avermitilis. Nat Biotechnol 21(5), 526–531.

Juurik T, Ilves H, Yeras R, Ilmjarv T, Tavita K, Ukkivi K, Teppo A, Mikkel K, and Kivisaar M (2012). Mutation frequency and spectrum of mutations vary at different chromosomal positions of <i>Pseudomonas putida</i>. PLOS 7, e48511.

Karlin S (2001). Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. Trends in Microbiol 9(7), 335–343.

Katoh K, Misawa K, Kuma K, and Miyata T (2014). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30(14), 3059–3066.

Kirby R (2011). Chromosome diversity and similarity within the Actinomycetales. FEMS Microbiol Lett 319(1), 1–10.

Kono N, Arakawa K, and Tomita M (2011). Comprehensive prediction of chromosome dimer resolution sites in bacterial genomes. BMC Genomics 12, 19.

Kopejtka K, Lin Y, Jakubovičová M, Kobl\'\ižek M, and Tomasch J (2019). Clustered core-and pan-genome content on Rhodobacteraceae chromosomes. Genome Biol Evol 11(8), 2208–2217.

Koski L B, Morton R A, and Golding G B (2001). Codon bias and base composition are poor indicators of horizontally transferred genes. Mol Biol Evol 18(3), 404–412.

Mackiewicz P, Gierlik A, Kowalczuk M, Dudek M R, and Cebrat S (1999). How does replication-associated mutational pressure influence amino acid composition of proteins? Genome Res 9(5), 409–416.

Mackiewicz P, Mackiewicz D, Kowalczuk M, and Cebrat S (2001). Flip-flop around the origin and terminus of replication in prokaryotic genomes. Genome Biol 2(12), interactions1004–1.

Medini D, Serruto D, Parkhill J, Relman D A, Donati C, Moxon R, Falkow S, and Rappuoli R (2008). Microbiology in the post-genomic era. Nat Rev Microbiol 6(6), 419–430.

Mira A, Martin-Cuadrado A B, D'Auria G, and Rodriguez-Valera F (2010). The bacterial pangenome:a new paradigm in microbiology. Intl Microbiol 13(2), 45–57.

Morrow J D and Cooper V S (2012). Evolutionary effects of translocations in bacterial genomes. Genom Biol Evol 4(12), 1256–1262.

Ochman H (2003). Neutral mutations and neutral substitutions in bacterial genomes. Mol Biol Evol 20(12), 2091–2096.

Ochman H, Lawrence J G, and Groisman E A (2000). Lateral gene transfer and the nature of bacterial innovation. Nature 405(6784), 299.

Pinto U M, Flores-Mireles A L, Costa E D, and Winans S C (2011). RepC protein of the octopine-type Ti plasmid binds to the probable origin of replication within repC and functions only in cis. Mol Microbiol 81(6), 1593–1606.

Prescott D M and Kuempel P L (1972). Bidirectional replication of the chromosome in <i>Escherichia coli</i>. Proc Natl Acad Sci 69(10), 2842–2845.

Sharp P M, Bailes E, Grocock R J, Peden J F, and Sockett R E (2005). Variation in the strength of selected codon usage bias among bacteria. Nucleic Acids Res 33(4), 1141–1153.

Sharp P M, Shields D C, Wolfe K H, and Li W.-H (1989). Chromosomal location and evolutionary rate variation in <i>Enterobacterial</i> genes. Science 246, 808–810.

Shimodaira H and Hasegawa M (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol 16(8), 1114.

Tillier E R and Collins R A (2000). Genome rearrangement by replication-directed translocation. Nat Genet 26(2), 195–197.

Wong K and Golding G B (2003). A phylogenetic analysis of the pSymB replicon from the Sinorhizobium meliloti genome reveals a complex evolutionary history. Can J Microbiol 49(4), 269–280.

Yang Z (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. Bioinfor 13(5), 555–556.