Title: The Location of Substitutions and Bacterial Genome Arrangements

Authors: Daniella F. Lato and G. Brian Golding

Journal: Genome Biology and Evolution

DOI:

Corresponding Author Information:
G. Brian Golding
McMaster Univeristy
Department of Biology
1280 Main St. West
Hamilton, ON
Canada
L8S 4K1
Email: golding@mcmaster.ca

Supplementary Material

Further supplemental information and code are available on `GitHub` at `www.github.com/dlato/Spatial_Patterns_of_Substitutions`.

Software Version Numbers

Program	Version Number	Build Date
`baseml`	4.9	March 2015
`codeml`	4.9	March 2015
`consense`	3.6b	NA
`dnadist`	3.6b	NA
`dnaml`	3.6b	NA
`MAFFT`	v7.045b	June 5, 2013
`neighbor`	3.6b	NA
`progressiveMauve`	Snap Shot	June 7, 2012
`RAxML`	8.0.25	June 16, 2014
`seqboot`	3.6b	NA
`trimAl`	v1.4.rev15	December 17, 2013

Table S1: Version numbers and build dates for each of the programs used.

Sequences

Bacteria Strain/Species	Accession Number	Date Accessed
Escherichia coli		
E. coli 0104H4	CP003289	September 29, 2016
E. coli 0157H7	BA000007	September 29, 2016
E. coli 083H1	CP001855	September 29, 2016
E. coli IAI39	CU928164	September 26, 2016
E. coli K12 *	U00096	September 26, 2016
E. coli UMN026	CU928163	September 26, 2016
Outgroup: *E. fergusonii* ATCC 35469T	NC_011740	August 26, 2020
Bacillus subtilis		
B. subtilis 168 *	NC_000964	November 10, 2016
B. subtilis BS38	NZ_CP017314	November 11, 2016
B. subtilis BSn5	NC_014976	November 11, 2016
B. subtilis PY79	NC_022898	November 11, 2016
B. subtilis QB928	NC_018520	November 11, 2016
B. subtilis RONN1	NC_017195	November 11, 2016
B. subtilis W23	NC_014479	November 11, 2016
Outgroup: *B. cereus* FDAARGOS_797	NZ_CP053931	August 26, 2020
Streptomyces		
S. lividans TK24	NZ_GG657756	August 26, 2020
S. lividans 1362	NZ_CM001889	August 26, 2020
S. coelicolor A3 *	AL645882	November 30, 2016
S. coelicolor A32 CFB NCB	NZ_CP042324	August 26, 2020
S. coelicolor M1154/pAMX4/pGP1416	NZ_CP050522	August 26, 2020
Outgroup: *S. aureofaciens* DM1	NZ_CP020567	August 26, 2020
S. meliloti Chromosome		
S. meliloti 2011	NC_020528	April 24, 2017
S. meliloti 1021 *	NC_003047	June 3, 2014
S. meliloti AK83	NC_015590	June 3, 2014
S. meliloti BL225C	NC_017322	June 3, 2014
S. meliloti SM11	NC_017325	June 3, 2014
S. meliloti RMO17	NC_CP009144	April 24, 2017
Outgroup: *Rhizobium leguminosarum* trifolii WSM1689 chromosome	NZ_CP007045	August 26, 2020
S. meliloti pSymA		
S. meliloti 2011	NC_020527	April 24, 2017
S. meliloti 1021 *	NC_003037	June 3, 2014
S. meliloti AK83	NC_015591	June 3, 2014
S. meliloti BL225C	NC_017324	June 3, 2014
S. meliloti SM11	NC_017327	June 3, 2014
S. meliloti RMO17	NC_CP009145	April 24, 2017
Outgroup: *Rhizobium leguminosarum* trifolii WSM1689 plasmid pRLG202	NC_0113665	August 26, 2020
S. meliloti pSymB		
S. meliloti 2011	NC_020560	April 24, 2017
S. meliloti 1021 *	NC_003078	June 3, 2014
S. meliloti AK83	NC_015596	June 3, 2014
S. meliloti BL225C	NC_017323	June 3, 2014
S. meliloti SM11	NC_017326	June 3, 2014
S. meliloti RMO17	NC_CP009146	April 24, 2017
Outgroup: *Rhizobium leguminosarum* trifolii WSM1689 plasmid pRLG201	NC_011368	August 26, 2020

Table S2: Strains and species used for each replicon analysis. Accession numbers, date accessed, and outgroups for each replicon are provided. An astrix (*) insicates the strain that was used as the representative strain.

Constraints to Number of Sequence Chosen

Computational time constraints and the nature of the data were limiting factors for the number of strains that were chosen for each bacterial species. `progressiveMauve` is a multiple sequence alignment program which is useful for accounting for local and large scale genomic rearrangements. Some of the bacterial strains are very similar and therefore there was no issue finding a sufficient number of locally co-linear blocks (LCBs) without having the genomes broken into an overwhelming number of blocks. *E. coli*, *Bacillus subtilis*, and *S. meliloti* were among the bacteria where this was the case. However, the *Streptomyces* strains were slightly too distantly related so when we tried to use a comparable number of strains to the other bacteria (six genomes), `progressiveMauve` split the genomes into 521 LCBs (Supplementary Figure S1). These blocks were therefore very small in length and resulted in many blocks that were comparing sequences with poor homology. Consequently, we had to reduce the number of genomes used for the *Streptomyces* analysis and after many iterations of genome combinations, we settled on three *Streptomyces* genomes with a total of 6 LCBs (Supplementary Figure S2). This allowed for the correct comparison of homologous sequences, while also accounting for recombination.

The computational time required to run `progressiveMauve` was an additional constraint that needed to be considered. `progressiveMauve` can align multiple whole genomes an identify regions that have been rearranged within the taxa provided. This process happens in relatively quick computational time, however, like most other programs, the addition of more data increased the amount of time required to complete the process. We ran multiple instances of `progressiveMauve` with varying numbers of *E. coli* genomes (Supplementary Figure S3). These data points were connected using a locally estimated scatterplot smoothing method and confidence intervals (Supplementary Figure S3). From this data, we can see that increasing the number of genomes exponentially increases the run time of `progressiveMauve`. It becomes impractical to align more than 27 genomes with `progressiveMauve`, as anything over that would take more than 24h to run. This information combined with `progressiveMauve`'s inability to pair of homologous sequences in LCBs of distantly related taxa, has limited the total number of genomes we can use per taxa to a maximum of 7. This provides the most accurate data and the most reasonable analysis duration.

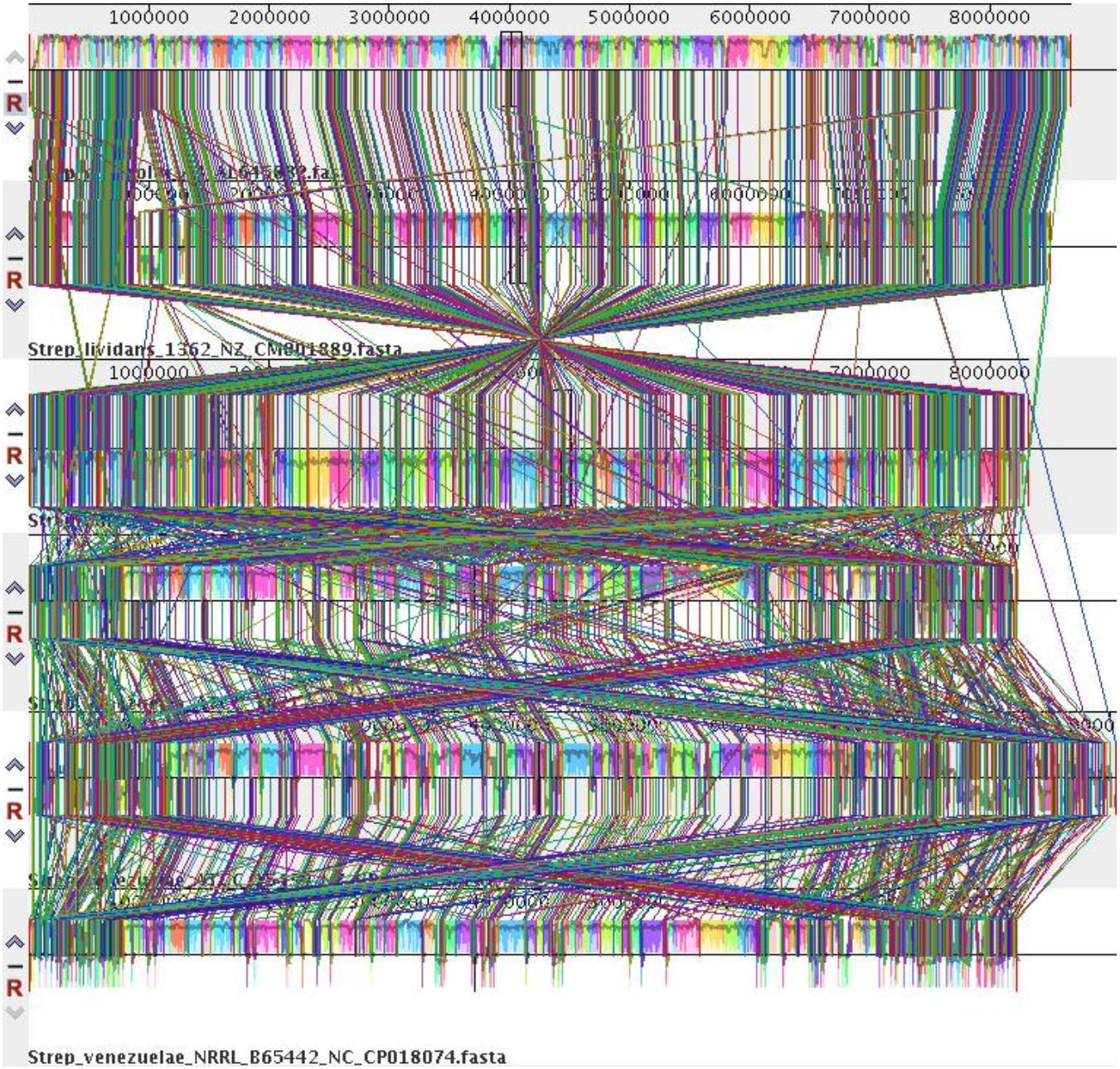

Figure S1: Visualization of the `progressiveMauve` alignment of 6 *Streptomyces* genomes (from top to bottom): *S. coelicolor* AL645882, *S. lividans* NZ_CM001889, *S. lividans* NZ_GG657756, *S. venezuelae* NC_018750, *S. venezuelae* NZ_CP013129, and *S. venezuelae* NC_CP018074. Each coloured block represents a different locally co-linear block (LCB). Coloured lines connect LCBs that are similar btween taxa. The black lines underneath each LCB represent the whole genome sequence of each of the *Streptomyces* taxa. Each LCB can be treated as a rearrangement, there have therefore been 521 rearrangements between these *Streptomyces* genomes.

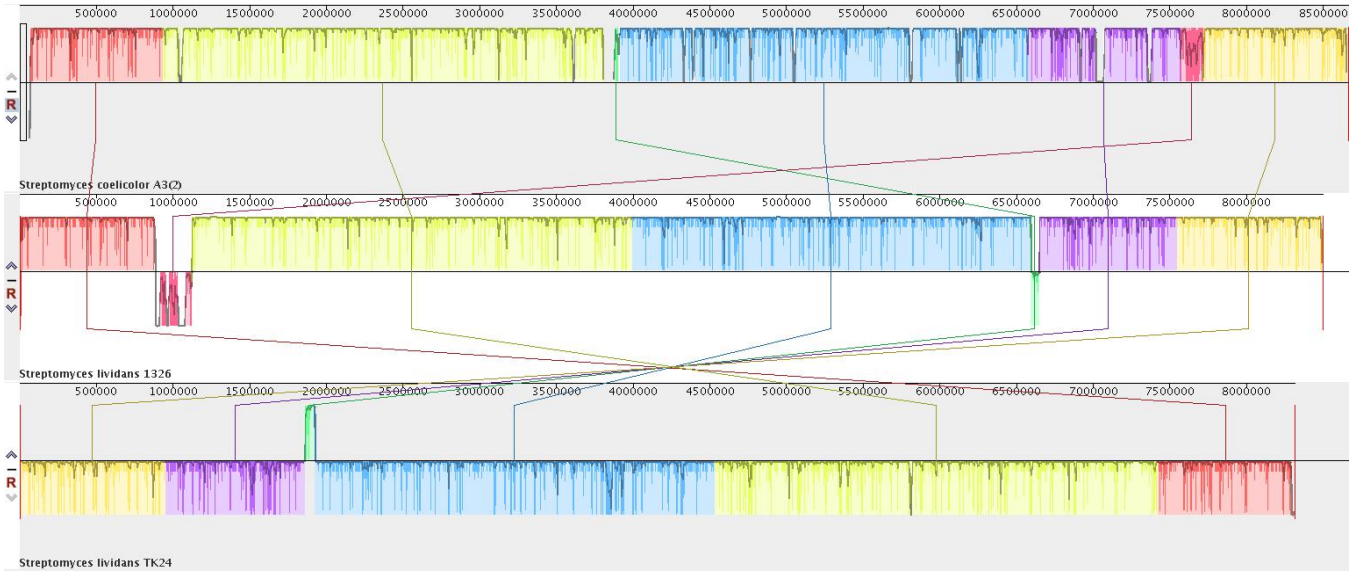

Figure S2: Visualization of the `progressiveMauve` alignment of the 3 *Streptomyces* genomes chosen for this analysis (from top to bottom): *S. coelicolor* AL645882, *S. lividans* NZ_CM001889, and *S. lividans* NZ_GG657756. Each coloured block represents a different locally co-linear block (LCB). Coloured lines connect LCBs that are similar btween taxa. The black lines underneath each LCB represent the whole genome sequence of each of the *Streptomyces* taxa. Each LCB can be treated as a rearrangement, there have therefore been 6 rearrangements between these *Streptomyces* genomes.

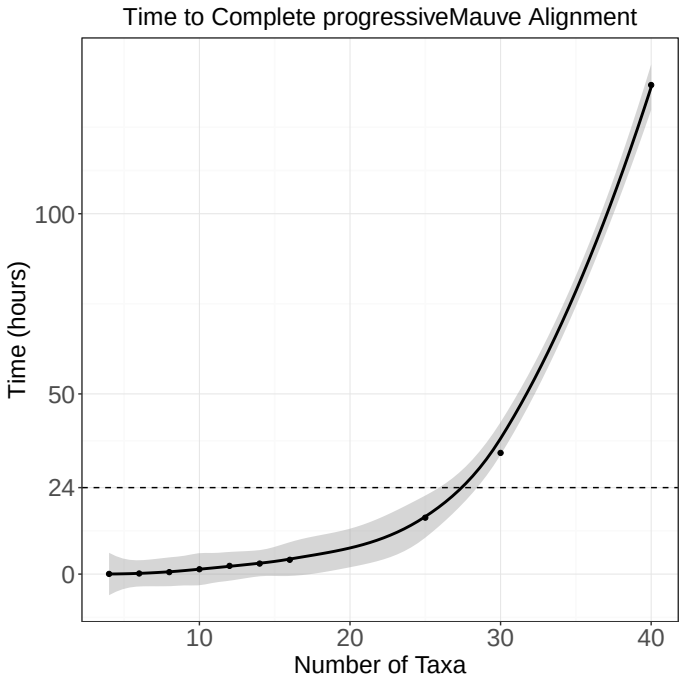

Figure S3: This graph shows the time to complete a `progressiveMauve` alignment with varying numbers of *E. coli* genomes. The total number of genomes or taxa is along the x-axis and the total time in hours is along the right axis. Each black point represents data from one `progressiveMauve` alignment. All data points are connected by calculating locally estimated scatterplot smoothing (black line) with confidence intervals (grey band).

`progressiveMauve` Alignment

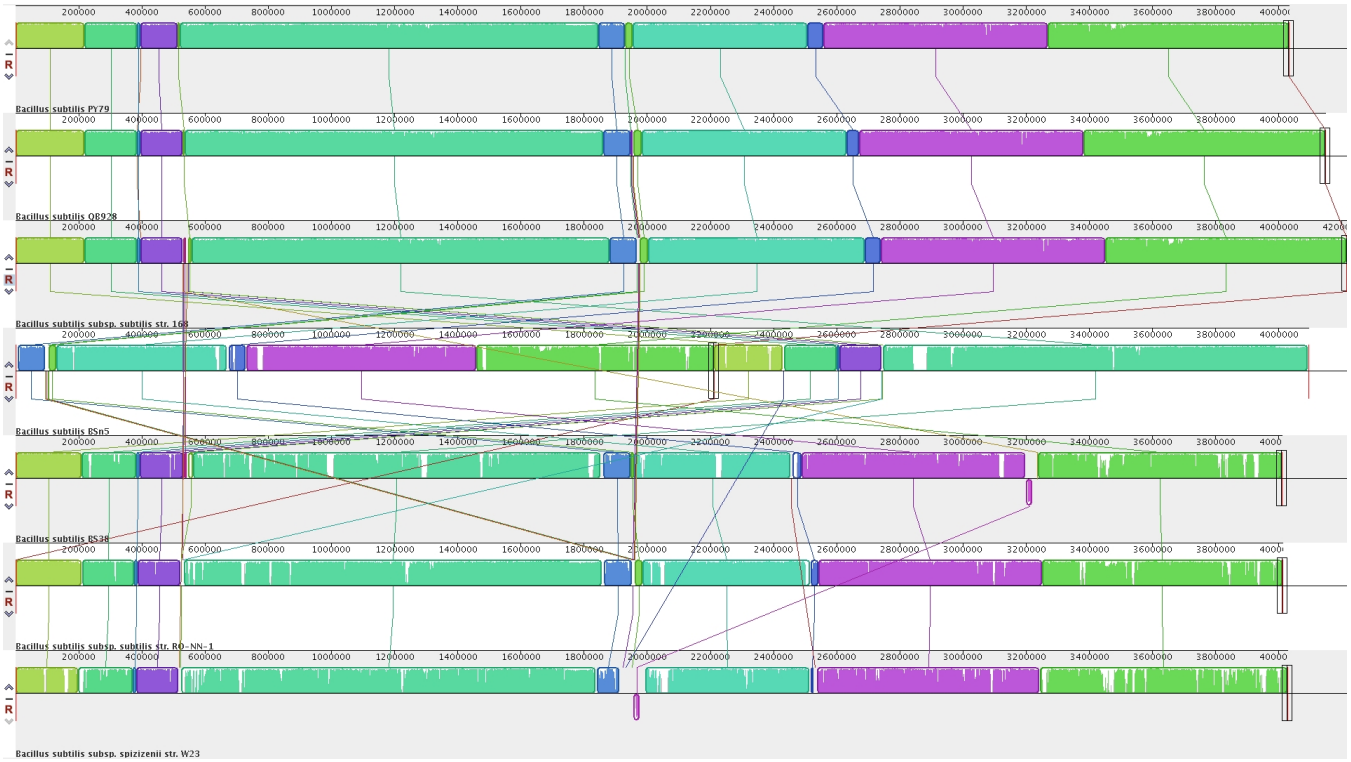

Figure S4: Visualization of the `progressiveMauve` alignment of the *B. subtilis* genomes. Each coloured block represents a different locally colinear block (LCB). Coloured lines connect LCBs that are similar btween taxa. The black lines underneath each LCB represent the whole genome sequence of each of the *B. subtilis* taxa. From top to bottom the taxa are: *B. subtilis* PY79, *B. subtilis* QB928, *B. subtilis* 168, *B. subtilis* BSn5, *B. subtilis* BS38, *B. subtilis* RONN1, *B. subtilis* W23. Each LCB can be treated as a rearrangement, there have therefore been 12 rearrangements between these *B. subtilis* genomes.

Poor Sequence Alignment

After a re-alignment of `progressiveMauve` LCBs with `MAFFT` there were still regions of the alignment that were visibly poor. This prompted the additional alignment quality trimming using a custom `Python` script and `trimAl` (Capella-Gutiérrez et al. 2009). An example of what a "poor" alignment would look like can be found in Figure S5. The `FASTA` format of this segment of the alignment can be found on `GitHub` labelled as file "`poor_ecoli_alignment_example.fna`".

This segment of `MAFFT` alignment (Figure S5) appears to have completely misaligned the second sequence (*E. coli* O157H7). When we look at the genes that these regions of DNA are found within (Table S3), we see that the second sequence (*E. coli* O157H7) does not have the same protein sequence as the other bacteria genes. Poor sequence alignments like this, as well as other non-homologous alignment regions were removed from the analysis. Please see the main paper for more detailed methods.

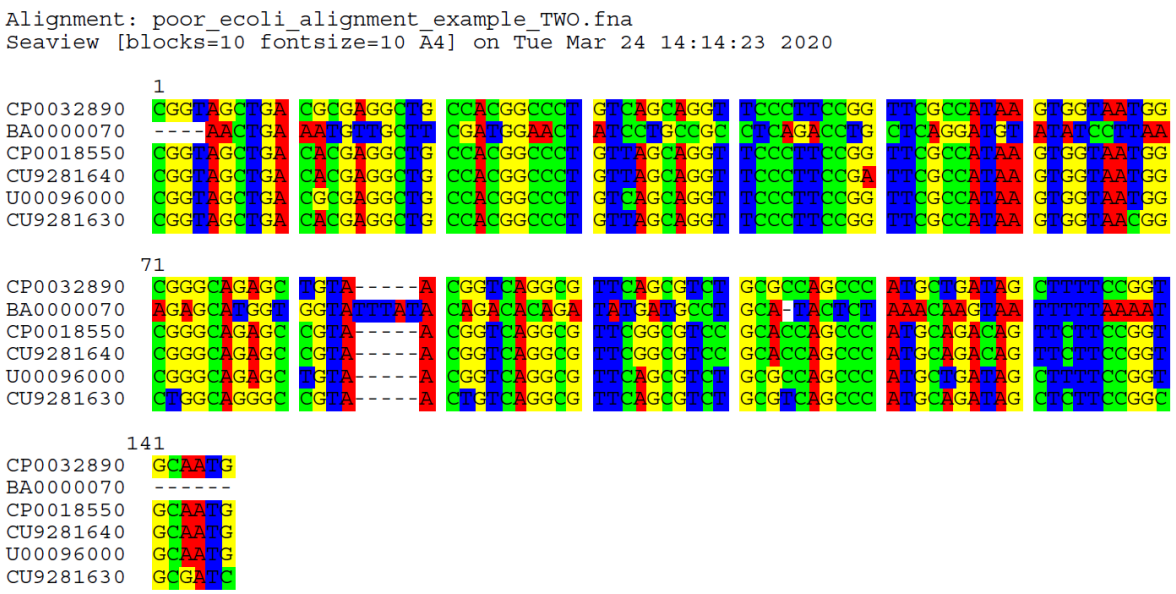

Figure S5: Visualization of a section of `MAFFT` alignment between the six strains of *E. coli*. This alignment was visualized with the `SeaView` graphical interface (Gouy et al. 2010).

E. coli Strain	NCBI Accession Number	Alignment Gene Id
0104H4	CP003289	O3K_04155
O157H7	BA000007	ECs3861
083H1	CP001855	NRG857_18350
IAI39	CU928164	yghE
K12	U00096	yghE
UMN026	CU928163	yghE

Table S3: *E. coli* strain, NCBI accession number, and Gene Id for the genes in the poor alignment example (Figure S5).

Phylogenetic Trees

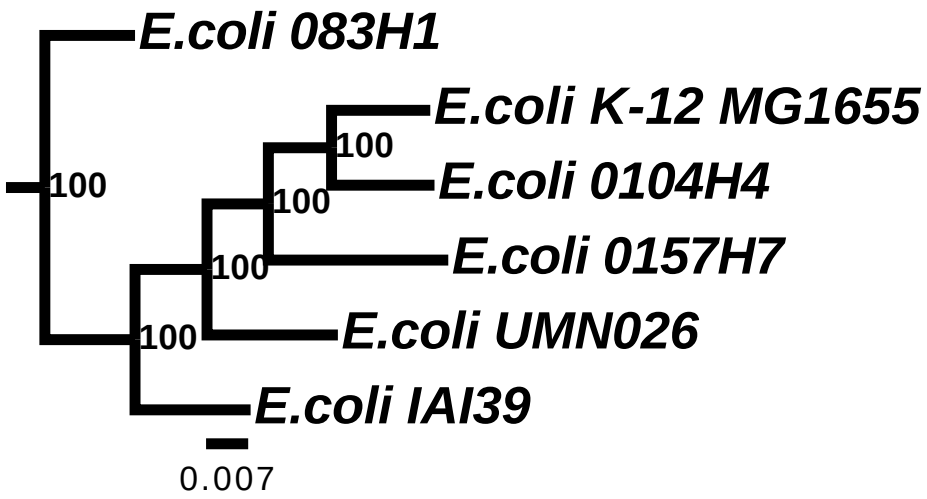

Figure S6: Phylogenetic tree of *E. coli* genomes. *Salmonella enterica* was used as an outgroup to root the tree. Branch lengths are to scale. The numbers at each node indicate the bootstrap value as a percentage. The number of bootstrapped trees was 100.

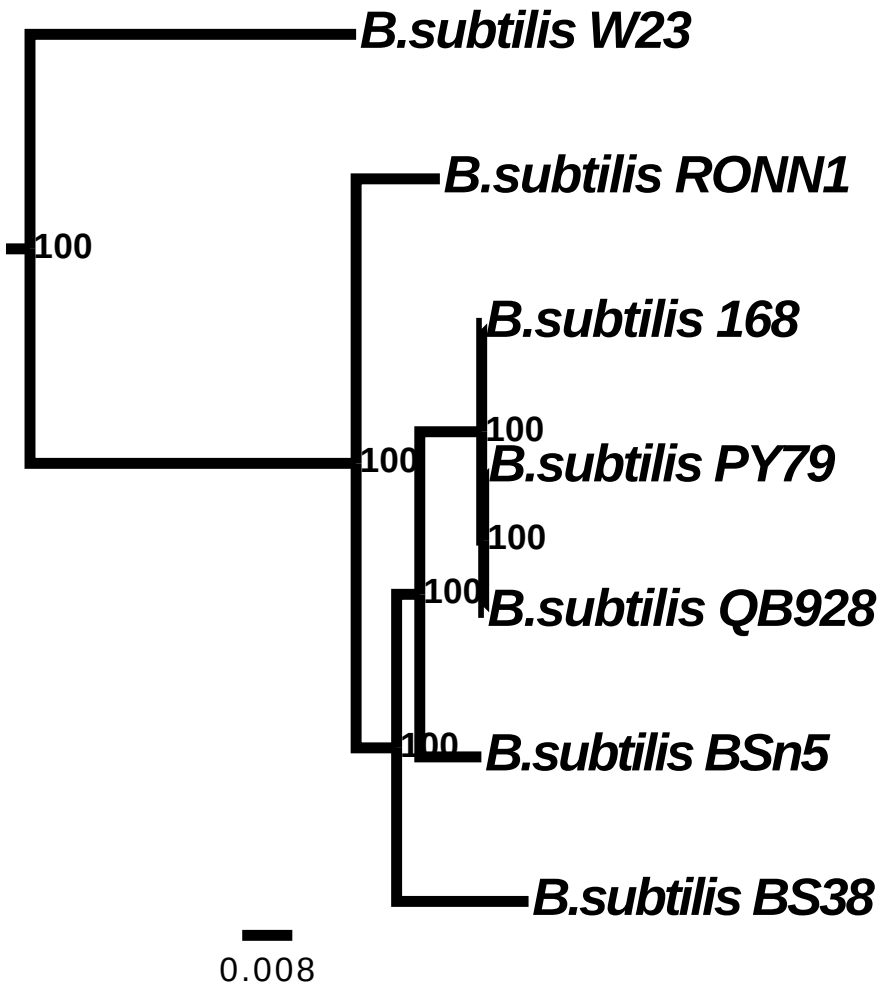

Figure S7: Phylogenetic tree of *B. subtilis* genomes. *Listeria monocytogenes* was used as an outgroup to root the tree. Branch lengths are to scale. The numbers at each node indicate the bootstrap value as a percentage. The number of bootstrapped trees was 100.

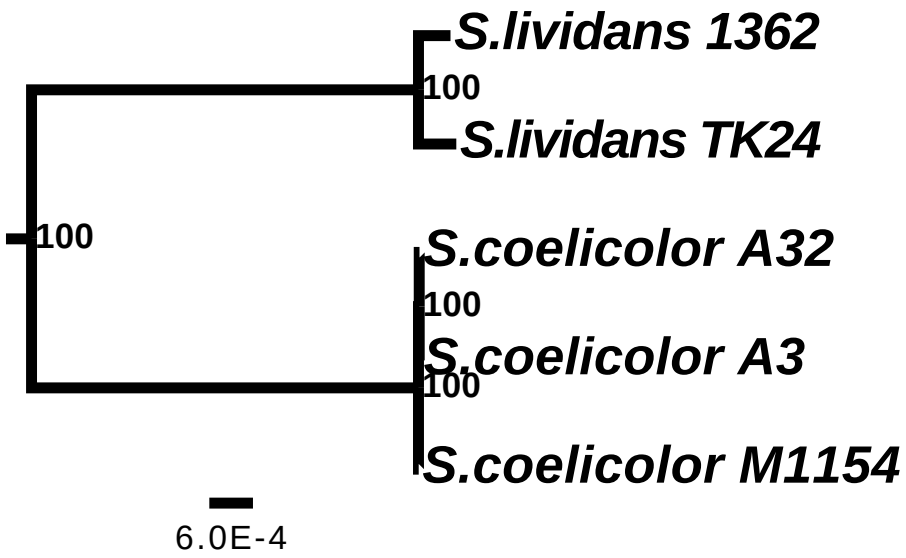

Figure S8: Phylogenetic tree of *Streptomyces* genomes. *Mycobacterium tuberculosis* was used as an outgroup to root the tree. Branch lengths are to scale. The numbers at each node indicate the bootstrap value as a percentage. The number of bootstrapped trees was 100.

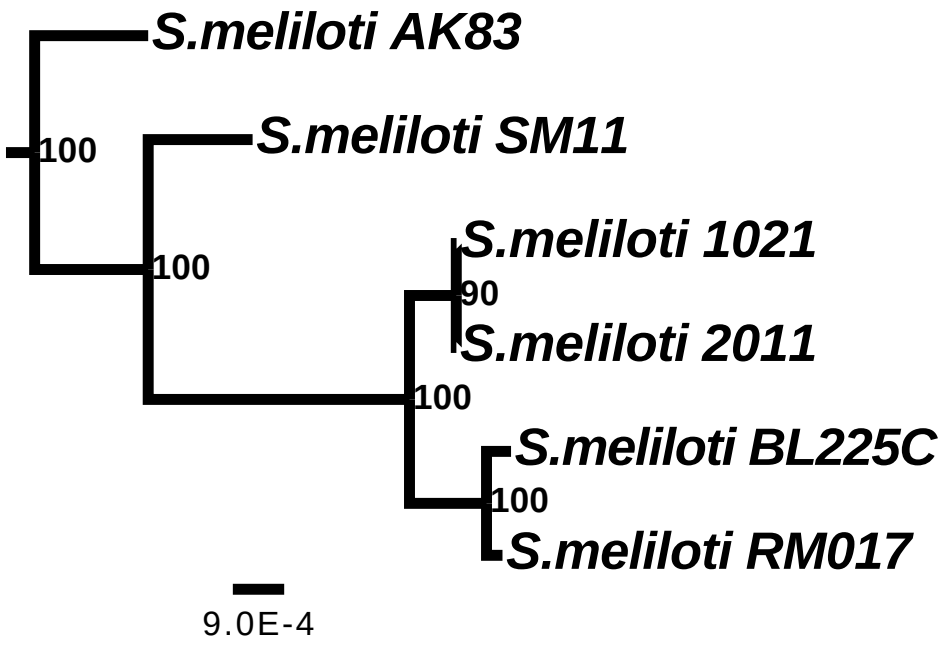

Figure S9: Phylogenetic tree using only the chromosomes of *S. meliloti*. *A. tumefaciens* circular chromosome was used as an outgroup to root the tree. Branch lengths are to scale. The numbers at each node indicate the bootstrap value as a percentage. The number of bootstrapped trees was 100.

Origin and Terminus Locations

Each of the bacterial strains used in this analysis vary in total genomic length, in some cases this difference is up to 856Kbp like in *E. coli* (Table S4). This will cause the farthest point from the origin of replication to appear larger

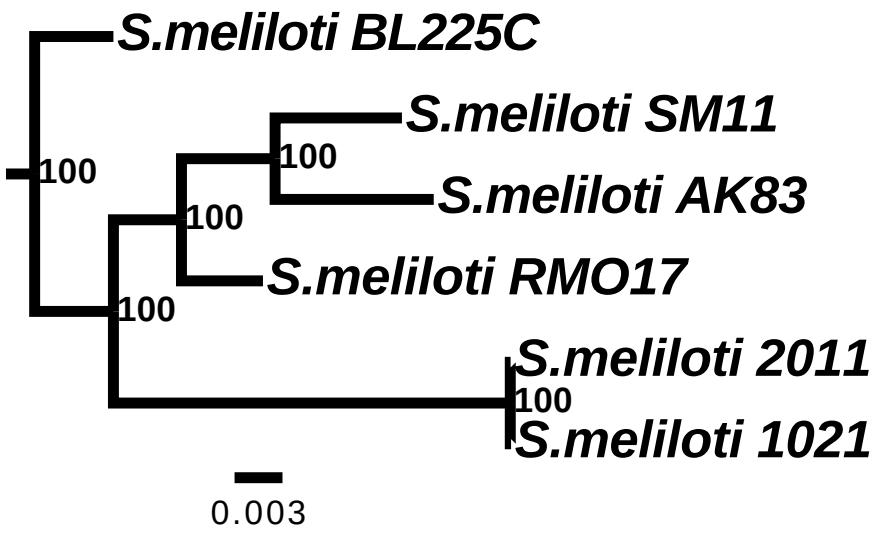

Figure S10: Phylogenetic tree using only pSymA of *S. meliloti*. *A. tumefaciens* circular plasmid was used as an outgroup to root the tree. Branch lengths are to scale. The numbers at each node indicate the bootstrap value as a percentage. The number of bootstrapped trees was 100.

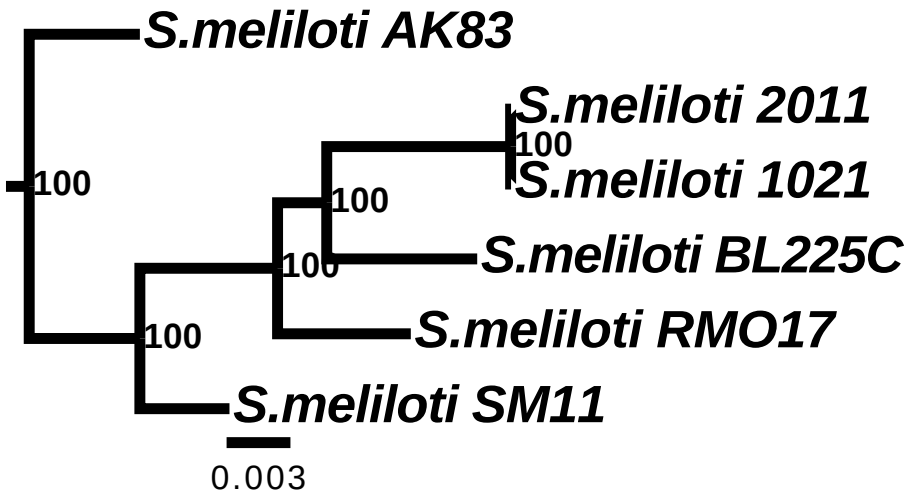

Figure S11: Phylogenetic tree using only pSymB of *S. meliloti*. *A. tumefaciens* circular chromid was used as an outgroup to root the tree. Branch lengths are to scale. The numbers at each node indicate the bootstrap value as a percentage. The number of bootstrapped trees was 100.

because of the increased genome size of some strains.

Bacteria	Origin of Replication	Terminus of Replication	Length of Longest Genome (bp)
E. coli	3925744	1588773	5498450
B. subtilis	1	1942542	4215606
Streptomyces	3419363	1 & 8667664	8667664
S. meliloti Chromosome	1	1735626	3908022
S. meliloti pSymA	1350001	672888	1633319
S. meliloti pSymB	55090	896756	1690594

Table S4: Origin of replication and terminus of replication positions in replicons of *E. coli*, *B. subtilis*, *Streptomyces*, and *S. meliloti*. The origin and terminus of replication are values from the representative strain of each bacteria, which can be found in Supplementary Table S2. The linear nature of *Streptomyces* chromosome gives it two termini, one at each end of the chromosome. The lenght of the longest genome is the longest genome length from all strains/species of each bactiera. This is not necessarally the same as the genome length of the representative strain.

Origin Location	*E. coli* Chromosome	*B. subtilis* Chromosome	*Streptomyces* Chromosome	*S. meliloti* Chromosome	*S. meliloti* pSymA	*S. meliloti* pSymB
Moved 100kb Left	-1.445×10^{-7}***	4.374×10^{-9}*	6.909×10^{-9}***	-1.316×10^{-6}***	-1.058×10^{-6}***	-2.009×10^{-7}***
Moved 90kb Left	-1.544×10^{-7}***	-1.036×10^{-7}***	5.677×10^{-9}***	-1.32×10^{-6}***	-1.246×10^{-6}***	-1.357×10^{-7}***
Moved 80kb Left	-1.65×10^{-7}***	-1.072×10^{-7}***	8.11×10^{-9}***	-1.338×10^{-6}***	-1.398×10^{-6}***	-6.57×10^{-8}***
Moved 70kb Left	-1.667×10^{-7}***	-1.102×10^{-7}***	6.716×10^{-9}***	-1.363×10^{-6}***	-1.405×10^{-6}***	9.83×10^{-8}
Moved 60kb Left	-1.64×10^{-7}***	-1.19×10^{-7}***	8.7×10^{-9}***	-1.324×10^{-6}***	-1.394×10^{-6}***	1.129×10^{-7}***
Moved 50kb Left	-1.446×10^{-7}***	-1.211×10^{-7}***	1.045×10^{-8}***	-1.36×10^{-6}***	-1.403×10^{-6}***	1.521×10^{-7}***
Moved 40kb Left	-1.4×10^{-7}***	-1.299×10^{-7}***	1.214×10^{-8}***	-1.255×10^{-6}***	-1.422×10^{-6}***	1.543×10^{-7}***
Moved 30kb Left	-1.498×10^{-7}***	-1.292×10^{-7}***	1.24×10^{-8}***	-1.26×10^{-6}***	-1.392×10^{-6}***	1.63×10^{-7}***
Moved 20kb Left	-1.51×10^{-7}***	-1.1×10^{-7}***	1.395×10^{-8}***	-1.525×10^{-6}***	-1.412×10^{-6}***	1.603×10^{-7}***
Moved 10kb Left	-1.262×10^{-7}***	-2.602×10^{-9}	1.563×10^{-8}***	-1.599×10^{-6}***	-9.499×10^{-7}***	2.973×10^{-7}***
Moved 10kb Right	-1.305×10^{-7}***	-2.045×10^{-8}***	1.578×10^{-8}***	1.614×10^{-6}***	-1.026×10^{-6}***	3.505×10^{-7}***
Moved 20kb Right	-1.454×10^{-7}***	-1.006×10^{-7}***	1.903×10^{-8}***	-1.634×10^{-6}***	-1.475×10^{-6}***	1.649×10^{-7}***
Moved 30kb Right	-1.548×10^{-7}***	-8.596×10^{-8}***	2.046×10^{-8}***	-1.698×10^{-6}***	-1.417×10^{-6}***	1.526×10^{-7}***
Moved 40kb Right	-1.632×10^{-7}***	-8.378×10^{-8}***	2.125×10^{-8}***	-1.719×10^{-6}***	-1.367×10^{-6}***	1.589×10^{-7}***
Moved 50kb Right	-1.856×10^{-7}***	-7.879×10^{-8}***	1.957×10^{-8}***	-1.735×10^{-6}***	-1.277×10^{-6}***	1.654×10^{-7}***
Moved 60kb Right	-1.91×10^{-7}***	-6.98×10^{-8}***	1.974×10^{-8}***	-1.788×10^{-6}***	-1.169×10^{-6}***	1.645×10^{-7}***
Moved 70kb Right	-1.892×10^{-7}***	-6.634×10^{-8}***	1.934×10^{-8}***	-1.854×10^{-6}***	-1.059×10^{-6}***	1.843×10^{-7}***
Moved 80kb Right	-1.879×10^{-7}**	-5.814×10^{-8}***	2.313×10^{-8}***	-1.891×10^{-6}***	-9.07×10^{-7}***	1.90×10^{-7}***
Moved 90kb Right	-1.862×10^{-7}***	-4.314×10^{-8}***	2.304×10^{-8}***	-1.865×10^{-6}***	-7.171×10^{-7}***	2.415×10^{-7}***
Moved 100kb Right	-1.799×10^{-7}***	-2.597×10^{-8}***	1.945×10^{-8}***	-1.525×10^{-6}***	-6.572×10^{-7}***	3.095×10^{-7}***

Table S5: Logistic regression analysis of the number of substitutions along the genome of the respective bacterial replicons after the origin location was moved by the specified increments from the original origin of replication position (listed in Table S4). All results are marked with significance codes as followed: $<$ 0.001 = '***', 0.001 $<$ 0.01 = '**', 0.01 $<$ 0.05 = '*', 0.05 $<$ 0.1 = '.', $>$ 0.1 = ' '. Logistic regression was calculated after the origin of replication was moved to the new location in the genome and all subsequent positions were scaled around the origin accounting for bidirectionality of replication.

Bacteria Strain	Accession Number	Date Accessed
E. coli K12 Chromosome	U00096	September 26, 2016
B. subtilis 168 Chromosome	NC_000964	November 10, 2016
S. coelicolor A3 Chromosome	AL645882	November 30, 2016
S. meliloti Chromosome 1021	NC_003047	June 3, 2014
S. meliloti pSymA 1021	NC_003037	June 3, 2014
S. meliloti pSymB 1021	NC_003078	June 3, 2014

Table S6: Strains and species used for determining the protein coding regions of each bacterial replicon. GenBank reference annotation was used to determine all protein coding sections of the replicons. NCBI accession numbers and date accessed are provided.

Genomic Position Clustering

A custom `R` script was used to cluster genomic positions together based on a user specified genetic distance using single-link clustering. An illustration of the clustering method used in this supplemental test can be found in Figure S12. This clustering was done for genomic distances beginning at 1bp and increasing by one order of magnitude until 1,000,000bp difference exists between the taxa genomic positions. These newly clustered genomic positions were then put into the same substitution analysis as mentioned previously to determine the impact of this position clustering on the spatial substitution trends through a linear regression. A complete table of the statistical results from the clustering assessment are found in Table S7. The results from this analysis indicate that genomic positions up to 1,000,000bp apart can be considered a singular genomic position without altering the overall spatial substitution analysis.

Figure S12: Visualization of the genomic position clustering method. In this example, the user specified the genetic distance to be 2, all genomic positions within 2 base pairs would be clustered together. In this example we are looking at 6 taxa with genomic positions 10, 14, 12, 25, 22, and 20. Based on the clustering algorithm, positions 10, 14 and 12 would be grouped into a cluster (outlined in green), position 25 would be its own cluster (outlined in pink), and positions 22 and 20 would be grouped into another cluster (outlined in blue). Once the clusters are determined, a new genomic position for each of the clusters is calculated using the average of all positions within that cluster. In this example, the green cluster would have a new genomic position of 12 (the average between those three positions), the pink cluster would have the sane genomic position of 25, and the blue cluster would have a new genomic position of 21. The new list of genomic positions for the 4 taxa would be: 12, 12, 12, 25, 21 and 21.

Position Difference	*E. coli* Chromosome	*B. subtilis* Chromosome	*Streptomyces* Chromosome	*S. meliloti* Chromosome	*S. meliloti* pSymA	*S. meliloti* pSymB
1bp	-1.394×10^{-7}**	-2.538×10^{-8}**	1.736×10^{-8}**	-1.541×10^{-6}**	-9.130×10^{-7}**	2.488×10^{-7}***
10bp	-1.394×10^{-7}***	-2.518×10^{-8}***	-4.484×10^{-9}***	-1.627×10^{-6}***	-9.13×10^{-7}***	3.487×10^{-7}***
100bp	-1.764×10^{-7}***	-1.417×10^{-8}***	1.448×10^{-8}***	-1.605×10^{-6}***	-1.166×10^{-6}***	4.021×10^{-7}***
1000bp	-1.784×10^{-7}***	-1.417×10^{-8}***	1.505×10^{-8}***	-1.605×10^{-6}***	-1.153×10^{-6}***	4.021×10^{-7}***
10000bp	-1.712×10^{-7}***	-3.496×10^{-8}***	4.790×10^{-8}***	-1.605×10^{-6}***	-3.570×10^{-8}*	3.784×10^{-7}***
100000bp	-2.061×10^{-7}***	-3.561×10^{-8}***	4.167×10^{-9}***	-1.605×10^{-6}***	-4.676×10^{-7}***	3.784×10^{-7}***
1000000bp	4.229×10^{-8}***	-7.710×10^{-9}***	6.083×10^{-8}***	-1.605×10^{-6}***	4.285×10^{-6}***	-8.888×10^{-7}***

Table S7: Results from the position clustering analysis. Logistic regression analysis of the number of substitutions along the genome of the respective bacteria replicons to test position differences. The "Position Difference" column denotes different base pair distances that the positions in the genome were clustered together as. All results are marked with significance codes as followed: < 0.001 = '***', $0.001 < 0.01$ = '**', $0.01 < 0.05$ = '*', $0.05 < 0.1$ = '.', > 0.1 = ' '. Logistic regression was calculated after the positions in the genome were determined to be the same at each position difference listed in the first column.

Bacteria and Replicon	Average Replicon Length	Number of Sites	Number of Substitutions
E. coli Chromosome	5082529	2318259	353740
B. subtilis Chromosome	4077077	2032176	185060
Streptomyces Chromosome	8494093	6057063	24046
S. meliloti Chromosome	3426881	1892874	11210
S. meliloti pSymA	1455940	571278	13132
S. meliloti pSymB	1664597	1248879	28941

Table S8: Total number of protein coding sites in each replicon for this analysis and the number of those sites that have a substitution (multiple substitutions at one site are counted as two substitutions).

High Substitutions Gene Example

Throughout this analysis there are a few genes/gene segments in all the bacterial replicons that have relatively high numbers of substitutions when compared to other genes or gene segments. These high numbers of substitutions are indeed real changes seen in homologous genes. To illustrate this, we have chosen a segment of alignment from *Streptomyces*. Information about the genes involved in this segment can be found in Table S9. A protein alignment for these genes can be found on `GitHub` (`https://github.com/dlato/Spatial_Patterns_of_Substitutions`) under the file name "`Streptomyces_high_substitutions_gene_example.txt`".

Both *S. lividans* strains have 100% sequence identity at the DNA level, while the *S. coelicolor* species has 87.2% sequence identity with the *S. lividans* strains for this particular alignment. Dispite this high sequence identity and almost identical protein alignment (Figures S13 and S14), there are a total of 31 substitutions (across all nodes of the phylogenetic tree, Figure S8) within this short stretch of sequence. It is segments like these that are resulting in the appearance of extremely high numbers of substitutions in sections of all the bacterial repliconic genomes.

Species	NCBI Acession Number	Gene Id
S. coelicolor A3	AL645882	SCO6334
S. lividans 1362	NZ_CM001889	SLI_RS32020
S. lividans TK24	NZ_GG657756	SSPG_RS06405

Table S9: Information about the example gene segment with high number of substitutions.

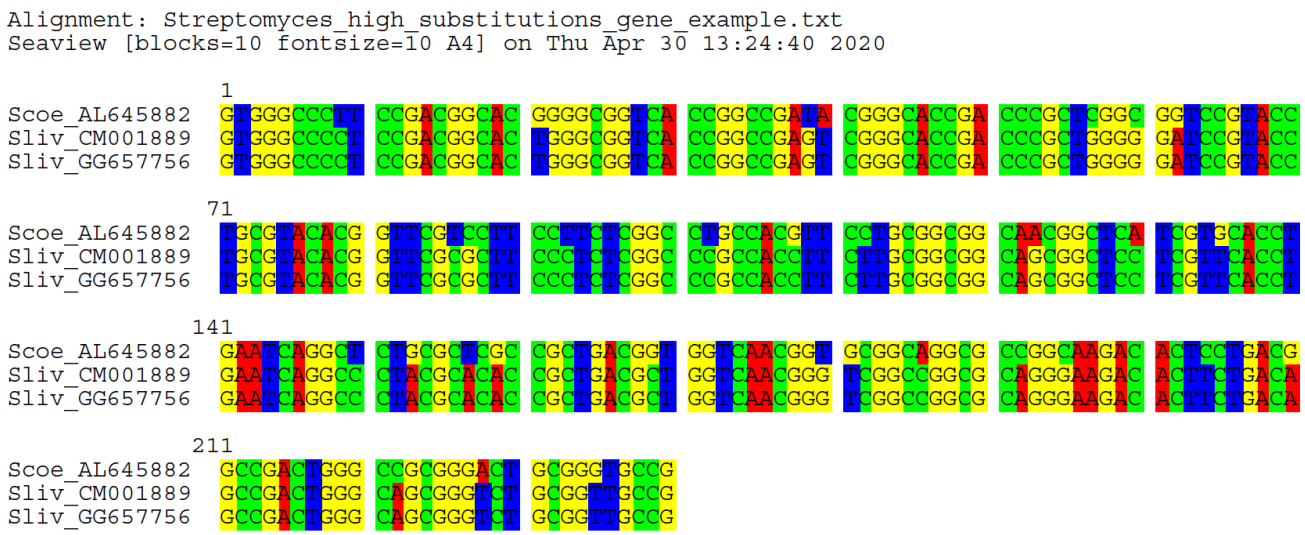

Figure S13: Visualization of the nucleotide alignment of *Streptomyces* genes with high numbers of substitutions. Alignment visualization was performed with `SeaView` (Gouy et al. 2010)

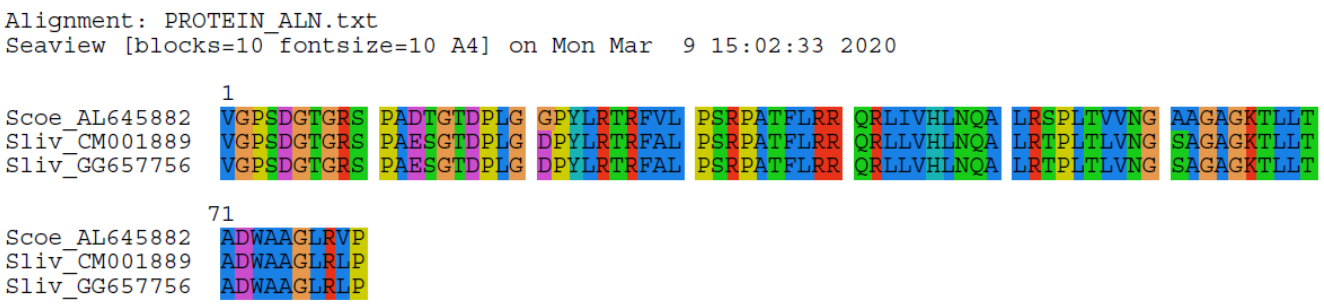

Figure S14: Visualization of the protein alignment of *Streptomyces* genes with high numbers of substitutions. Alignment visualization was performed with `SeaView` (Gouy et al. 2010)

High Substitution Distribution

Bacteria and Replicon	Bidirectional Genomic Position (bp)	Protein/Gene Examples
E. coli Chromosome	1130000 - 1140000	Uncharacterized protein
		Hypothetical protein
		Lipoprotein
		Transcriptional activator
	1720000 - 1740000	Hypothetical protein
		Predicted protein
		Small toxic polypeptide
B. subtilis Chromosome	560000 - 570000	Hypothetical protein
		Derived by automated computational analysis
		Membrane protein
	1820000 - 1380000	Derived by automated computational analysis
Streptomyces Chromosome	3550000 - 3570000	Hypothetical protein
		Derived by automated computational analysis
		Putative integral membrane protein
		Reductase
S. meliloti Chromosome	80000 - 90000	Hypothetical proteins
	730000 - 740000	Hypothetical proteins
		Putative proteins
S. meliloti pSymA	100000 - 110000	Hypothetical proteins
	800000 - 810000	Hypothetical protein
		Transporter protein
S. meliloti pSymB	450000 - 460000	Hypothetical protein
		Putative oxidoreductase
		Hypothetical proteins
	610000 - 620000	Hypothetical protein
		Putative transport regulator
		Predicted membrane protein

Table S10: Table of high number of substitutions per 10Kbp genomic regions for each bacterial replicon and examples of the associated proteins/gene functions found in that region. The genomic position begins at the origin of replication and continues in both directions until the terminus of replication (bidirectional replication).

Weighted, Non-weighted, and 20Kbp Near and Far From the Origin Substitution Linear Regression Analysis

	Protein Coding Window Size					
Bacteria and Replicon	10Kbp	25Kbp	50Kbp	100Kbp	200Kbp	400Kbp
E. coli Chromosome	-2.27×10^{-10}*** (0.038)	-2.54×10^{-10}** (0.078)	-2.32×10^{-10}** (0.112)	-2.36×10^{-10}* (0.133)	NS (0.200)	NS (0.362)
B. subtilis Chromosome	NS (0.009)	NS (0.001)	NS (0.0002)	NS (0.002)	NS (0.019)	NS (0.484)
Streptomyces Chromosome	NS (2.49×10^{-5})	NS (4.46×10^{-5})	NS (0.009)	NS (0.0002)	3.72×10^{-11}* (0.132)	4.34×10^{-11}* (0.197)
S. meliloti Chromosome	-1.21×10^{-10}** (0.076)	-1.71×10^{-10}*** (0.137)	-1.86×10^{-10}** (0.126)	-2.78×10^{-10}** (0.350)	NS (0.150)	NS (0.397)
S. meliloti pSymA	NS (0.032)	NS (0019)	NS (0.135)	NS (0.0124)	NS (0.034)	NS (1.42×10^{-30})
S. meliloti pSymB	NS (0.001)	NS (0.003)	NS (0.008)	NS (0.006)	NS (2.12×10^{-8})	NS (0.043)

Table S11: Linear regression on various sections of the genome (10Kbp, 25Kbp, 50Kbp, 100Kbp, 200Kbp, and 400Kbp) with increasing distance from the origin of replication after accounting for bidirectional replication. The total number of substitutions in each section of the genome was divided by the total number of protein coding sites in that genomic region. All results are marked with significance codes as followed: < 0.001 = '***', $0.001 < 0.01$ = '**', $0.01 < 0.05$ = '*', > 0.05 = 'NS'. The R^2 value for each coefficient estimate is found below the value in brackets ().

Multiple linear regressions were performed to determine if there was any correlation between number of substitutions and distance from the origin of replication. A linear regression to determine how the weighted and non-weighted total number of substitutions in various sections of the genome (10Kbp, 25Kbp, 50Kbp, 100Kbp, 200Kbp, and 400Kbp) changes with genomic position was performed (Tables S11 and S12). All additional linear regression results (Tables S11 and S12) mirror the results from the logistic regression on presence or absence of substitutions and changes in genomic position (see the Main Paper results section for more information). The results from these supplemental tests are consistent with the results from the linear regression found in the Main Paper, most bacterial replicons have a decreasing number of substitutions when moving away from the origin of replication.

To calculate the non-weighted values of the total number of substitutions per 10Kbp region of the genome, the total number of substitutions was summed up over each region of the genome (10Kbp, 25Kbp, 50Kbp, 100Kbp, 200Kbp, and 400Kbp), while accounting for bidirectional replication (see Main Paper for details). A linear regression on these total number of substitutions in each section of the genome (10Kbp, 25Kbp, 50Kbp, 100Kbp, 200Kbp, and 400Kbp) was performed to see how the number of substitutions changes with distance from the origin of replication (Table S12). The weighted values of the total number of substitutions per various region of the genome, the total number of substitutions was summed up over each region of the genome (10Kbp, 25Kbp, 50Kbp, 100Kbp, 200Kbp, and 400Kbp) while accounting for bidirectional replication (see Main Paper for details). These summed values were then divided by the total number of protein coding sites in each region to obtain the weighted value. A linear regression on these weighted total number of substitutions in each section of the genome was performed to see how the number of substitutions changes with distance from the origin of replication (Table S11).

We took a closer look at 20Kbp regions of the replicons close and far from the origin of replication. We performed a logistic regression on the presence or absence of a substitution with distance from the origin of replication. Data points from the 20Kbp regions closest to the origin of replication and data points from the 20Kbp regions closest to the terminus of replication were used for this portion of the analysis. Outliers were removed from this analysis. The number of substitutions per site was also calculated in each of these 20Kbp regions for each bacterial replicon. We were unable to determine a consistent spatial substitution trend when considering only the 20Kbp near and far from the orign of replication in all bacterial replicons. Some bacterial replicons had a positive correlation coefficient, indicating that the number of substitutions increases with increasing distance from the origin of replication (Table

	Protein Coding Window Size					
Bacteria and Replicon	10Kbp	25Kbp	50Kbp	100Kbp	200Kbp	400Kbp
E. coli Chromosome	-1.66×10^{-4}*** (0.398)	-4.12×10^{-4}*** (0.476)	-8.64×10^{-4}*** (0.563)	-1.71×10^{-3}*** (0.509)	-3.42×10^{-3}** (0.534)	-6.71×10^{-3}* (0.592)
B. subtilis Chromosome	NS (0.004)	NS (0.004)	NS (0.001)	NS (0.001)	NS (0.145)	NS (0.027)
Streptomyces Chromosome	NS (0.002)	NS (0.007)	NS (0.004)	NS (0.027)	NS (0.075)	NS (0.076)
S. meliloti Chromosome	-8.97×10^{-6}*** (0.040)	-3.72×10^{-5}** (0.098)	-7.76×10^{-5}* (0.126)	-1.64×10^{-4}* (0.188)	NS (0.082)	NS (0.427)
S. meliloti pSymA	NS (0.027)	NS (0.001)	NS (0.006)	NS (0.193)	NS (0.050)	NS (1.59×10^{-31})
S. meliloti pSymB	NS (0.035)	NS (0.053)	NS (0.010)	NS (0.002)	NS (0.495)	NS (0.491)

Table S12: Linear regression on various sections of the genome (10Kbp, 25Kbp, 50Kbp, 100Kbp, 200Kbp, and 400Kbp) with increasing distance from the origin of replication after accounting for bidirectional replication. The linear regression was performed on the total number of substitutions in each section of the genome without accounting for the number of sites in each genomic region. All results are marked with significance codes as followed: < 0.001 = '***', $0.001 < 0.01$ = '**', $0.01 < 0.05$ = '*', > 0.05 = 'NS'. The R^2 value for each coefficient estimate is found below the value in brackets ().

S13). Other replicons had a negative correlation coefficient, suggesting that the number of substitutions decreases with increasing distance from the origin of replication (Table S13). Additionally, it was unclear if the number of substitutions per site locally were higher near the origin of replication or near the terminus. Some bacteria had higher number of substitutions per site near the origin (*E. coli*, *S. meliloti* chromosome and pSymB), while other replicons has the opposite trend (*B. subtilis*, *Streptomyces* and *S. meliloti* pSymA) (Table S13). These results suggest that on a small local scale, there are varying patterns of substitutions with respect to distance from the origin of replication. This varies between bacteria, and in some cases even within the same bacteria (*E. coli*). This variation locally does not allow us to make any overarching statements about the local distribution of substitutions in bacterial genomes. It is therefore more useful to consider the global (genome wide) pattern of substitutions when making overarching statements about genomic substitution arrangements.

Non-linear Analysis of Number of Substitutions and Distance From the Origin of Replication

Using a simple smoothed conditional means method (`geom_smooth()` function in `R`), a non-linear trend analysis was performed on all bacterial replicons. The previous mentioned weighted data (see the previous subsection), was used in this analysis. The weighted data represents the total number of substitutions divided by the total number of protein-coding sites in 10Kbp segments of the genomes. Outliers were removed. The results from this non-linear analysis can be seen in Figures S15 - S20. The visual results from this analysis mirror the findings from the main paper, the total number of substitutions varies with distance from the origin of replication, but the direction of this trend is unclear and inconsistent between bacterial replicons.

Total Number of Sites Linear Regression

We performed a linear regression on the total number of protein coding sites and distance from the origin of replication. We found that the total number protein coding sites decreases with distance from the origin of replication in majority of the bacterial replicons in this analysis. We were unable to detect a significant relationship between the number of protein coding sites and distance from the origin of replication in pSymB of *S. meliloti*.

	Protein Coding			
	Correlation Coefficient 20kb Near		Number of Substitutions per 20kb Near	
Bacteria and Replicon	Origin	Terminus	Origin	Terminus
E. coli Chromosome	NS	6.16×10^{-6}**	5.85×10^{-3}	6.47×10^{-3}
B. subtilis Chromosome	1.18×10^{-6}*	1.57×10^{-5}***	4.23×10^{-3}	5.01×10^{-3}
Streptomyces Chromosome	NS	NS	2.26×10^{-4}	2.05×10^{-5}
S. meliloti Chromosome	7.11×10^{-6}***	NS	1.51×10^{-3}	3.86×10^{-5}
S. meliloti pSymA	-6.94×10^{-5}***	NS	2.03×10^{-3}	3.27×10^{-3}
S. meliloti pSymB	1.58×10^{-5}***	-7.10×10^{-5}***	3.06×10^{-3}	1.25×10^{-3}

Table S13: Logistic regression on 20kb closest and farthest from the origin of replication after accounting for bidirectional replication and outliers. Number of substitutions was calculated by taking the total number of substitutions in each of the 20Kbp regions and dividing by the total number of sites in those regions. All results are marked with significance codes as followed: < 0.001 = '***', $0.001 < 0.01$ = '**', $0.01 < 0.05$ = '*', > 0.05 = 'NS'. The R^2 values for each estimate are in brackets.

Average *dN*, *dS*, and ω per Gene Values

The average dN, dS, and ω values per gene were calculated. For genes that were split into multiple parts (due to the presence of gaps or poor homology in the alignment), the dN, dS, and ω values for each gene part were averaged to obtain a single average value per gene. A complete list of these values can be found on `GitHub` (`www.github.com/dlato/Spatial_Patterns_of_Substitutions`) under the file name "`Supplementary_table_per_gene_dN_dS_omega.pdf`".

Distribution of *dN*, *dS*, and ω

20Kbp Near and Far From Origin Selection Linear Regression Analysis

We additionally took a closer look at 20 genes close and far from the origin of replication. We performed a linear regression on the change in selection values (dN, dS, and ω) with distance from the origin of replication in these genes (Table S16). For majority of the bacterial replicons we failed to find a trend, which is not surprising since there was no evidence of an overall genomic trend when looking at these values (see Main Paper for results). Again, we are unable to conclude that there is a consistent overall trend for any of the selection values, dN, dS, and ω.

Bacteria and Replicon	Coefficient Estimate	R^2
E. coli Chromosome	-2.33×10^{-2}***	0.423
B. subtilis Chromosome	NS	0.001
Streptomyces Chromosome	-4.08×10^{-3}***	0.095
S. meliloti Chromosome	NS	0.013
S. meliloti pSymA	NS	0.002
S. meliloti pSymB	2.69×10^{-2}**	0.081

Table S14: Linear regression analysis of the total number of protein coding sites per 10kb along the genome of the respective bacteria replicons. Linear 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	Outliers (%)	Zero Value (%)		
		dN	*dS*	ω
E. coli Chromosome	7.49	13.82	1.05	13.82
B. subtilis Chromosome	5.41	4.40	0.16	4.40
Streptomyces Chromosome	4.74	25.70	14.48	25.70
S. meliloti Chromosome	17.05	61.21	59.26	61.21
S. meliloti pSymA	6.69	11.28	9.75	11.28
S. meliloti pSymB	6.13	13.20	5.20	13.20

Table S15: Percent of data that was calculated to be an outlier or had a selection variable (*dN*, *dS*, and ω) value of zero.

	Near Origin			Near Terminus		
Bacteria and Replicon	*dN*	*dS*	ω	*dN*	*dS*	ω
E. coli Chromosome	NS	NS	NS	NS	NS	NS
B. subtilis Chromosome	NS	NS	NS	NS	NS	NS
Streptomyces Chromosome	NS	NS	-9.36×10^{-7}* (0.328)	NS	NS	NS
S. meliloti Chromosome	NS	NS	NS	NS	NS	NS
S. meliloti pSymA	NS	NS	NS	-2.53×10^{-7}* (0.238)	NS	NS
S. meliloti pSymB	NS	6.19×10^{-6}** (0.372)	NS	NS	4.92×10^{-6}* (0.232)	NS

Table S16: Linear regression for *dN*, *dS*, and ω calculated for each bacterial replicon for the 20 genes closest and 20 genes farthest from the origin of replication. All results are marked with significance codes as followed: p: $<$ 0.001 = '***', 0.001 $<$ 0.01 = '**', 0.01 $<$ 0.05 = '*', $>$ 0.05 = 'NS'. The R^2 values for each estimate are in brackets.

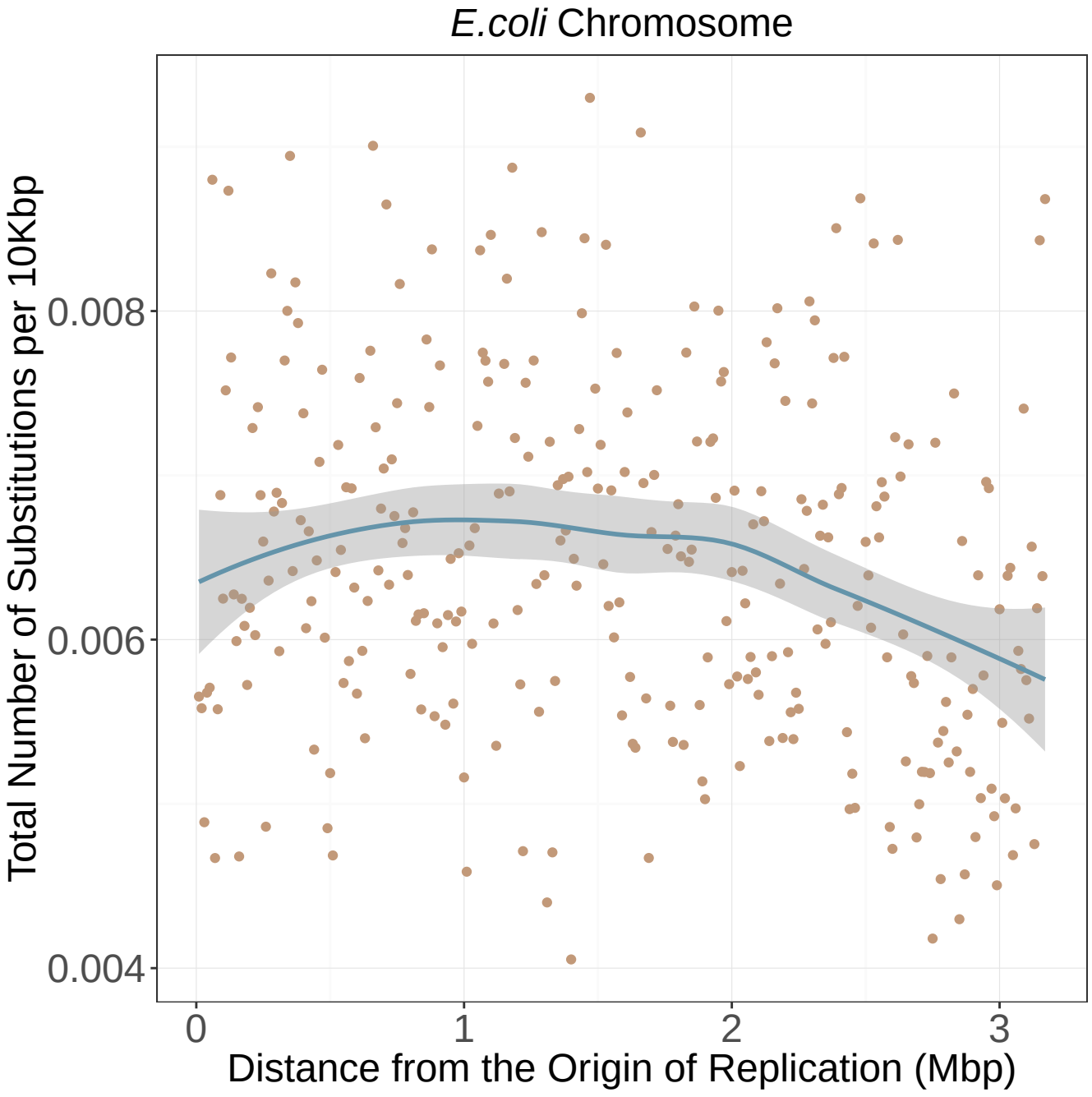

Figure S15: The graph shows the total number of substitutions weighted by the total number of protein-coding sites per 10Kbp segments of the *E. coli* genome. Each of these individual values are represented by beige coloured circles. A non-linear trend line (using the `geom_smooth()` function in `R`), was fit to these average values and the associated 95% confidence intervals for this line is represented by the grey ribbon around the blue trend line. Outliers were removed from this graph.

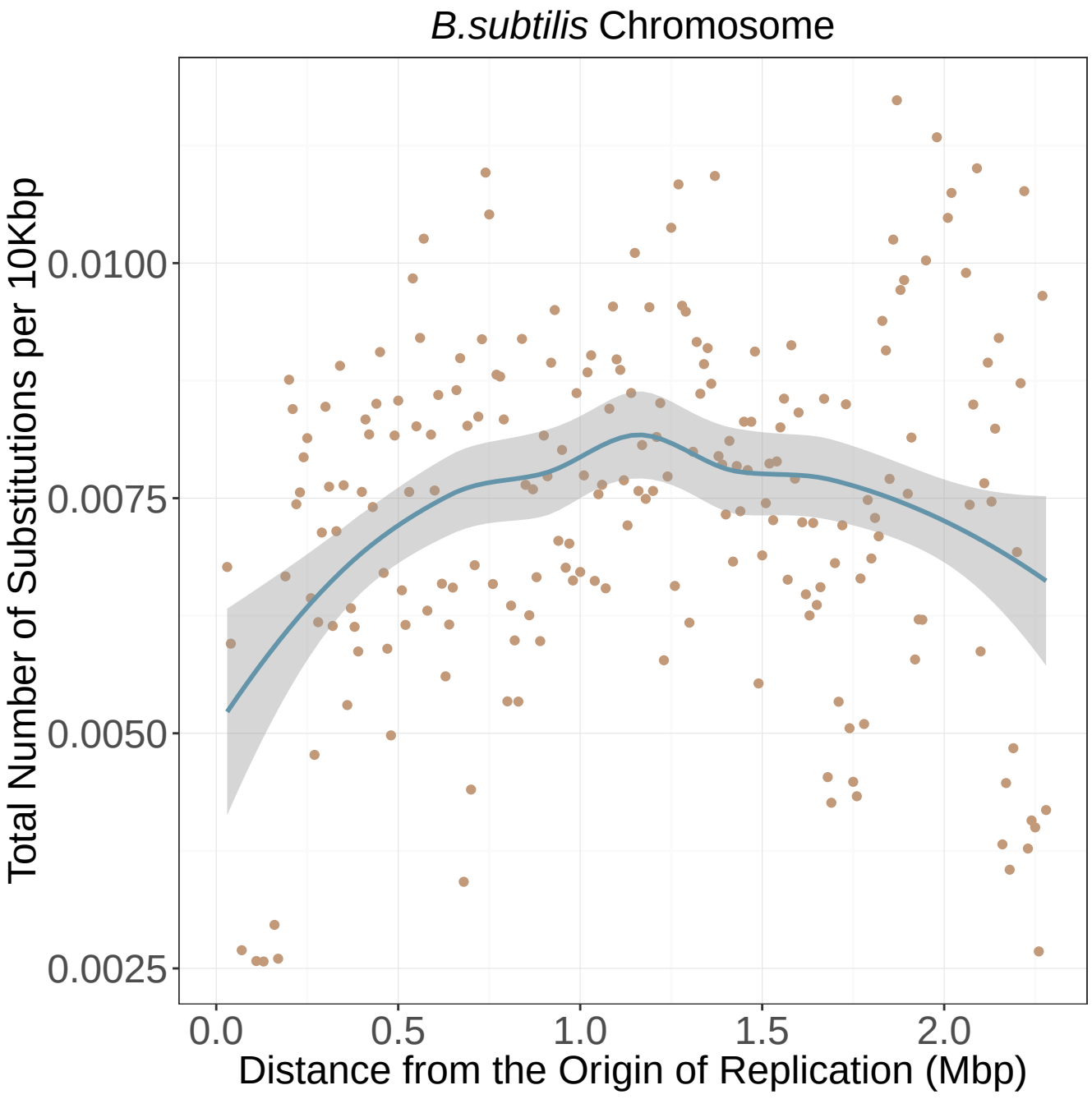

Figure S16: The graph shows the total number of substitutions weighted by the total number of protein-coding sites per 10Kbp segments of the *B. subtilis* genome. Each of these individual values are represented by beige coloured circles. A non-linear trend line (using the `geom_smooth()` function in `R`), was fit to these average values and the associated 95% confidence intervals for this line is represented by the grey ribbon around the blue trend line. Outliers were removed from this graph.

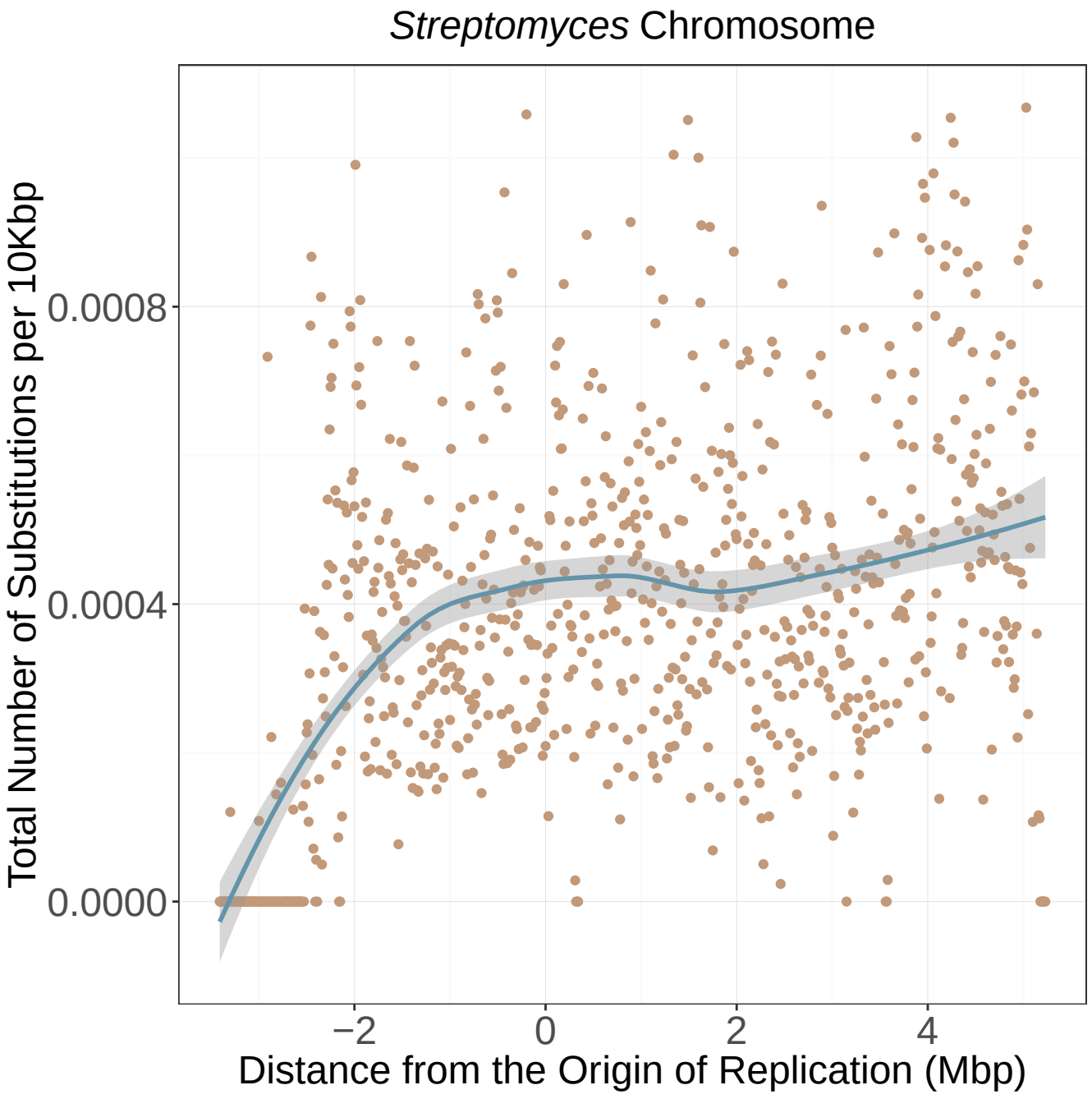

Figure S17: The graph shows the total number of substitutions weighted by the total number of protein-coding sites per 10Kbp segments of the *Streptomyces* genome. Each of these individual values are represented by beige coloured circles. A non-linear trend line (using the `geom_smooth()` function in `R`), was fit to these average values and the associated 95% confidence intervals for this line is represented by the grey ribbon around the blue trend line. Outliers were removed from this graph.

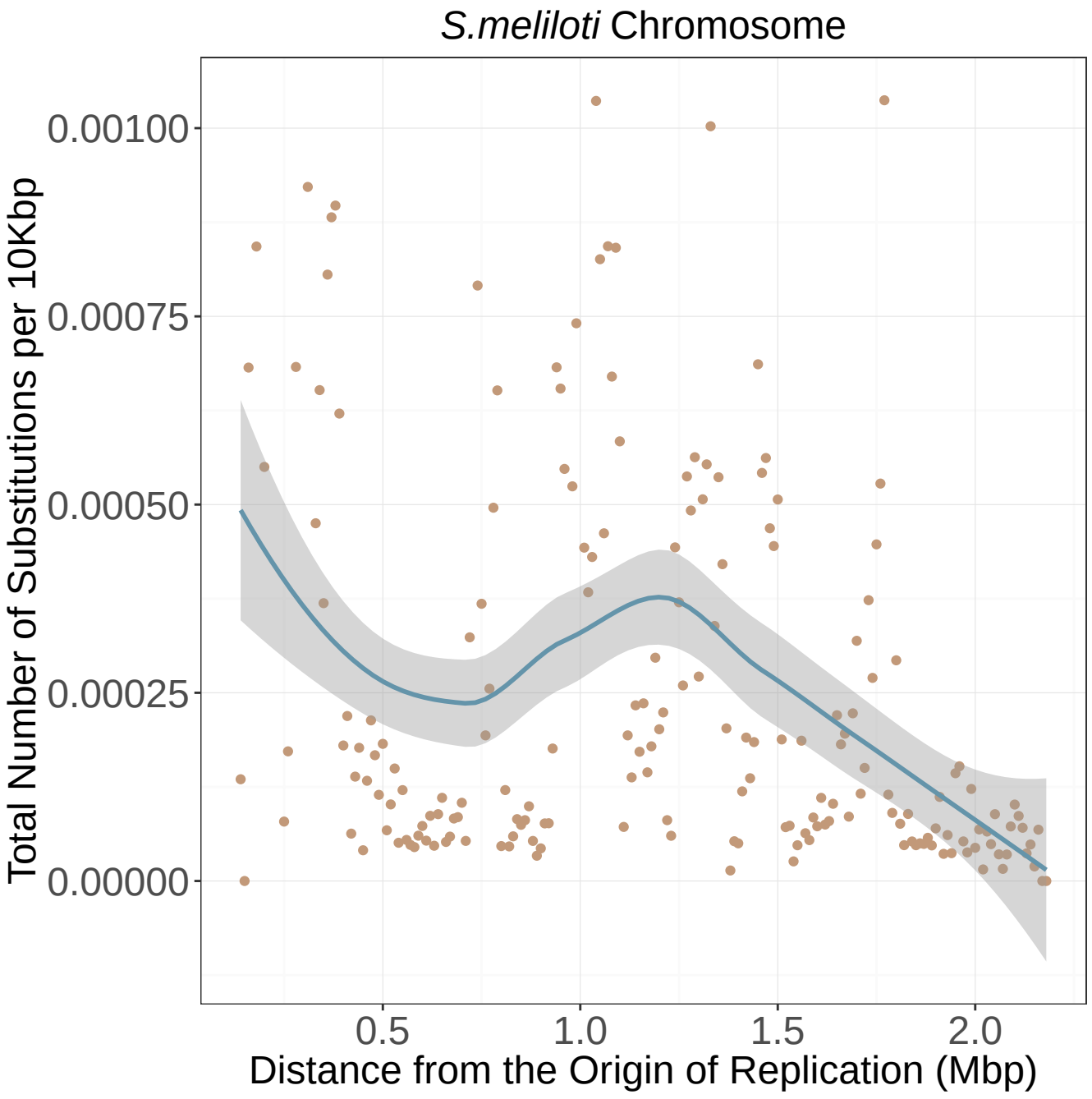

Figure S18: The graph shows the total number of substitutions weighted by the total number of protein-coding sites per 10Kbp segments of the *S. meliloti* Chromosome. Each of these individual values are represented by beige coloured circles. A non-linear trend line (using the `geom_smooth()` function in `R`), was fit to these average values and the associated 95% confidence intervals for this line is represented by the grey ribbon around the blue trend line. Outliers were removed from this graph.

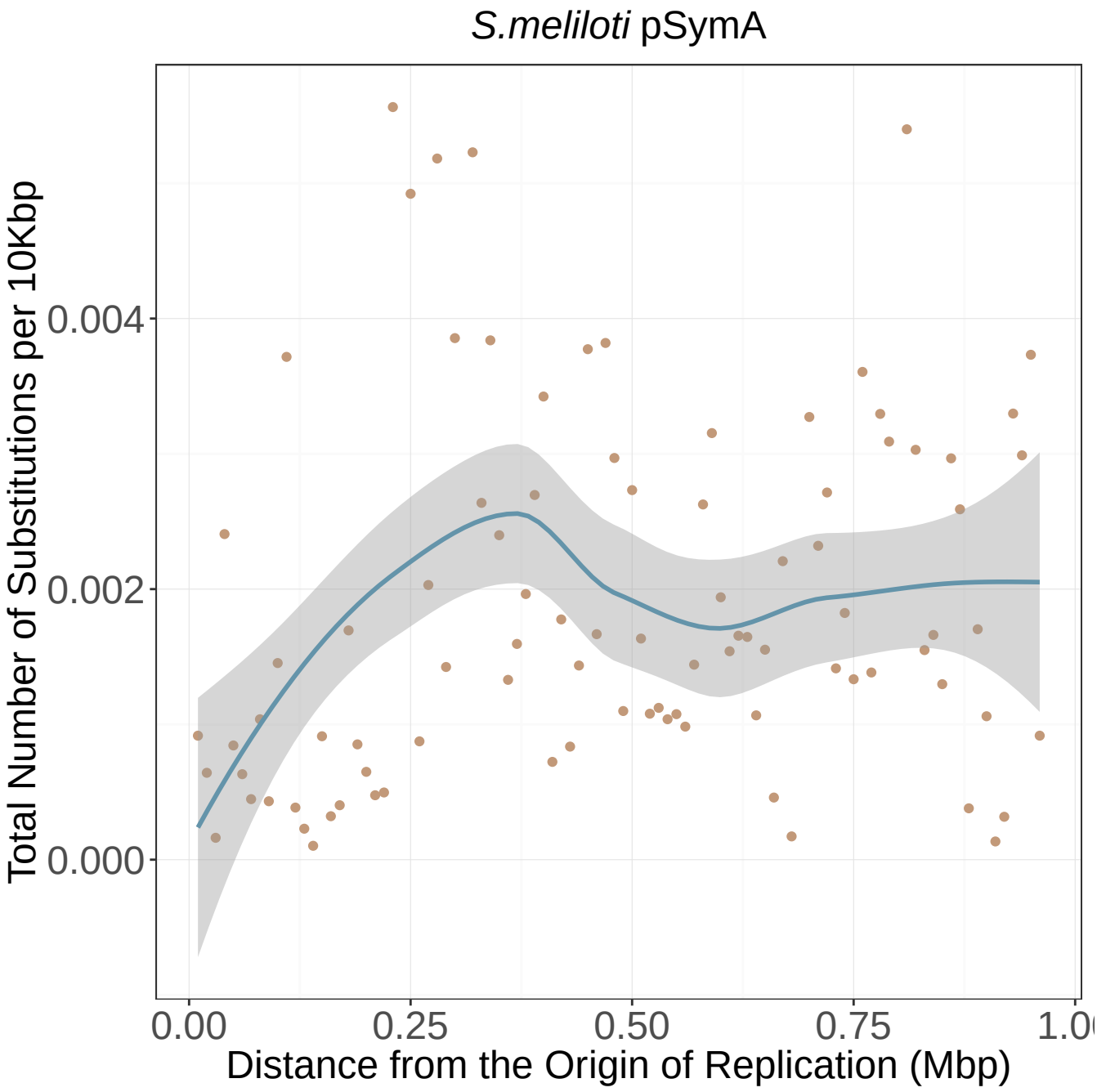

Figure S19: The graph shows the total number of substitutions weighted by the total number of protein-coding sites per 10Kbp segments of the *S. meliloti* pSymA replicon. Each of these individual values are represented by beige coloured circles. A non-linear trend line (using the `geom_smooth()` function in `R`), was fit to these average values and the associated 95% confidence intervals for this line is represented by the grey ribbon around the blue trend line. Outliers were removed from this graph.

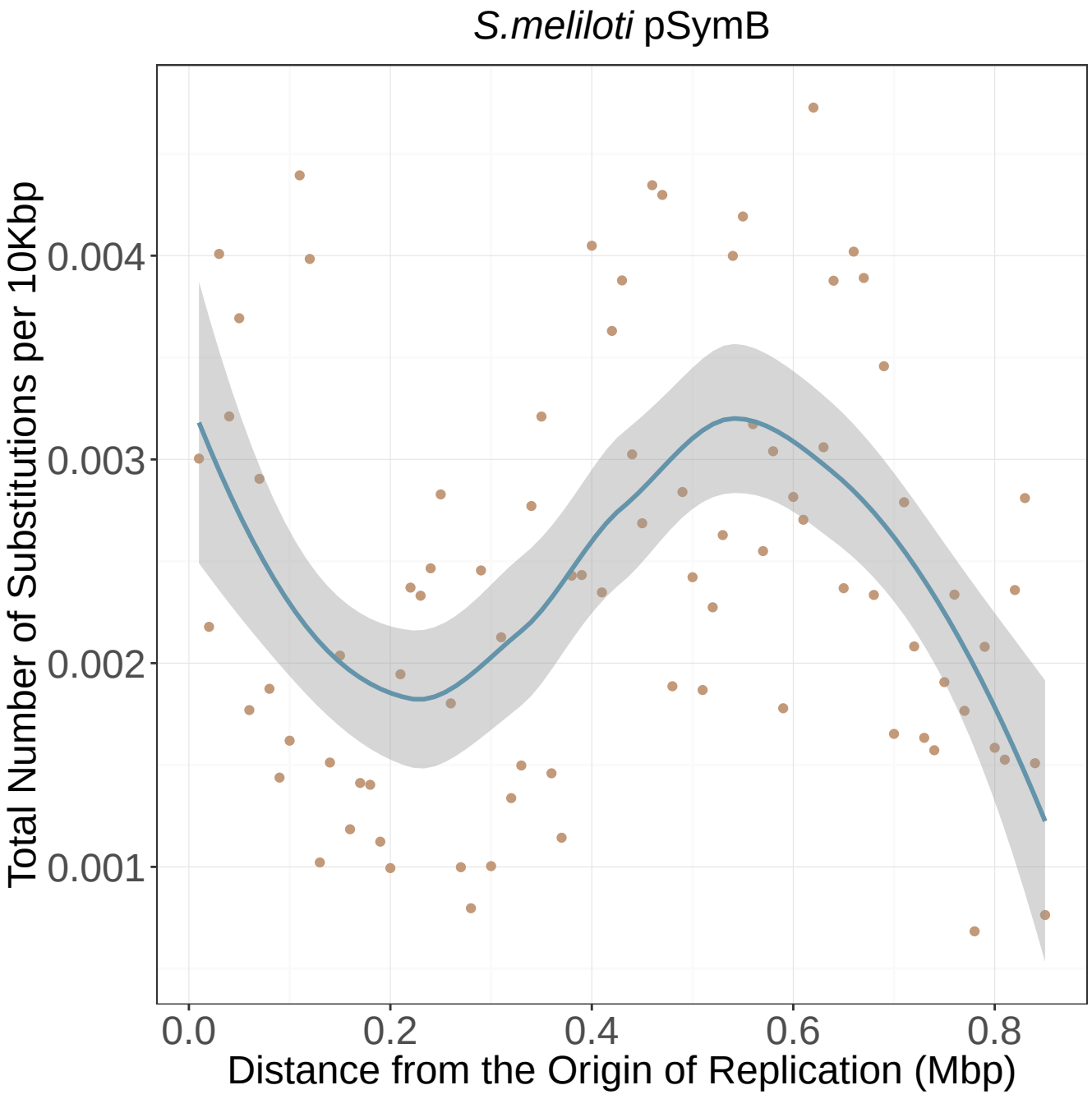

Figure S20: The graph shows the total number of substitutions weighted by the total number of protein-coding sites per 10Kbp segments of the *S. meliloti* pSymB replicon. Each of these individual values are represented by beige coloured circles. A non-linear trend line (using the `geom_smooth()` function in `R`), was fit to these average values and the associated 95% confidence intervals for this line is represented by the grey ribbon around the blue trend line. Outliers were removed from this graph.

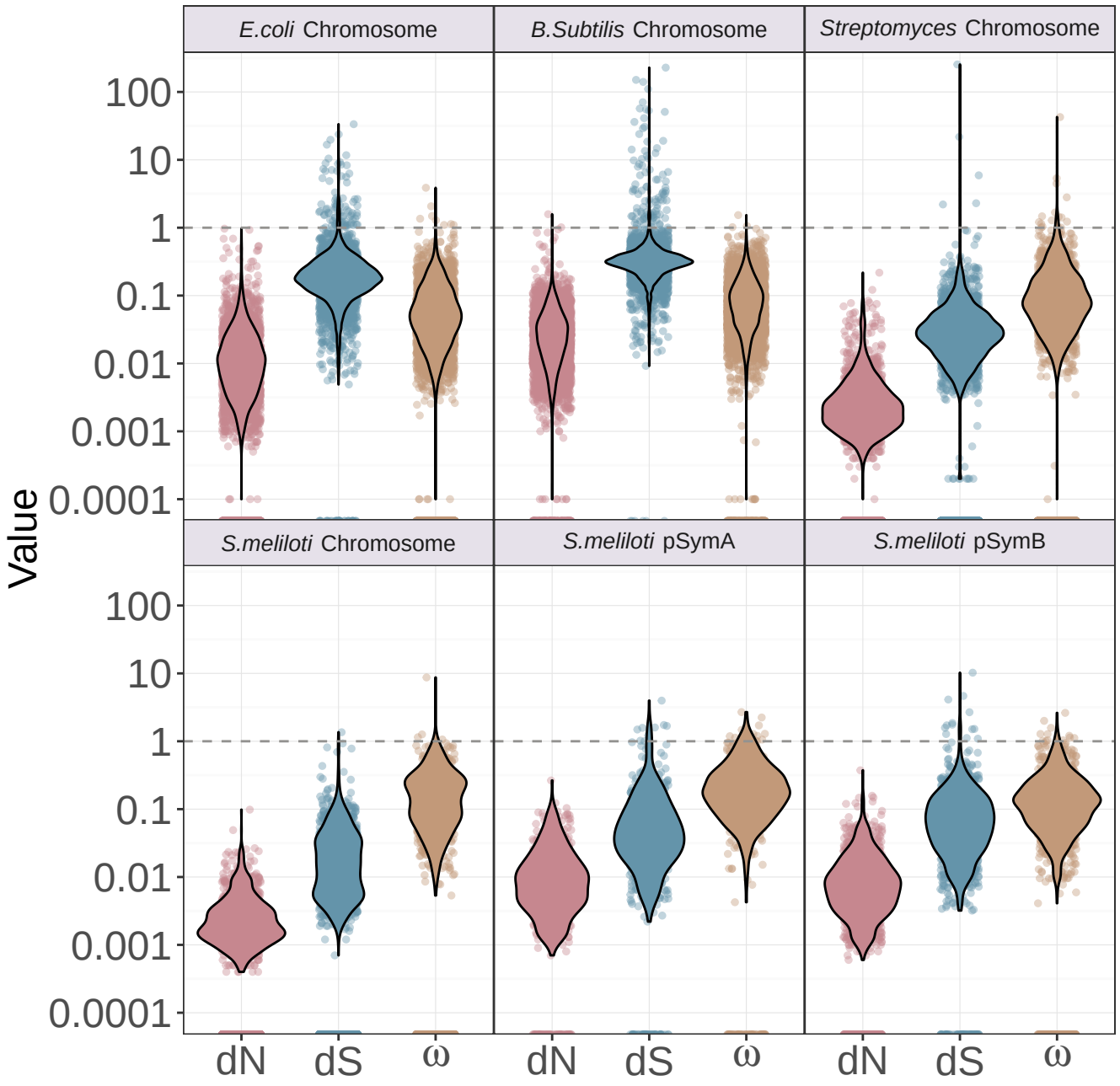

Figure S21: Distribution of all dN, dS, and ω values on a log base 10 scale for each replicon. Individual points are shown as a strip chart (which has been jittered in the x-direction in `R` (Wickham et al. 2019)), and the density of these selection values is shown in the overlaid violin plot. All points are included in this graphic including outliers. For more information on how outliers were calculated, please see the main paper. Any dN, dS, or ω values that had a value of zero is pushed to the bottom of the x-axis. Since these values will not appear on a log base 10 scale, they are not included in the violin portions of this graphic. For a complete list of zero values in each of the selection categories please refer to Table S15. In these graphs there is a horizontal line of values at 0.0001 for most of the selection coefficients in most of the bacterial replicons. This is due to rounding practices when `codeml` (Yang 1997) calculates dN, dS, and ω values.

References

Capella-Gutiérrez S, Silla-Martinez J M, and Gabaldón T (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinfor 25(15), 1972–1973.

Gouy M, Guindon S, and Gascuel O (2010). SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol 27, 221–224.

Stamatakis A (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinfor 30(9), 1312–1313.

Wickham H, Averick M, Bryan J, Chang W, McGowan L D, François R, Grolemund G, Hayes A, Henry L, Hester J, et al. (2019). Welcome to the {tidyverse}. Journal of Open Source Software 4(43), 1686.

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