Subs Paper Things to Do:

- more genomes
- ~~new outgroups? (too distant)~~
- explain high dS values in *B. subtilis*
- potentially poor alignment and non-orthologous genes (core genome, change methods?)
- ~~non-parametirc analysis for subs~~
- gap in *Escherichia coli* fig 5
- ~~new methods for trees~~
- ~~concerned about repeated genes (TEs) and not analyzing core genome~~
- ~~check if trimming respects coding frame~~
- clear distinction between mutations and substitutions in intro (separate sections)
- ~~datasets from previous papers (repeat my analysis on them?)~~
- why would uncharacterized proteins have higher subs rates?
- ~~R^2 values in regression analysis~~
- ~~update gene exp paper ref~~
- why are the lin reg of dN, dS and ω NS but the subs graphs are...explain!
- mol clock for my analysis?
- GC content? COG? where do these fit?

Inversions and Gene Expression Letter Things to Do:

- ~~create latex template for paper~~
- confirm inversions with dot plot
- make dot plot of just gene presence and absence matrix (instead of each site) to see if this will go better
- look up inversions and small RNA's paper Marie was talking about at Committee meeting
- write outline for letter
- write Abstract
- ~~write intro~~

- write methods
- compile tables (supplementary)
- write results
- write discussion
- write conclusion
- do same ancestral/phylogenetic analysis that I did in the subs paper

General Things to Do:

- summarize references 40 and 56 from Committee meeting report (Brian was asking)

Last Week

Inversions + Gene Expression:

✓Queenie: comparing blast and gene alignment homologs

✓Queenie: start creating dataframe that is compatible with `limma`

Subst Paper:

✓Commented on using previous papers datasets (Cooper, Morrow, Sharp, Flynn ...etc)

✓completed non-linear (non-parametric?) analysis for subst

✓new subst analysis with new `RAxML` trees and *Streptomyces* genomes is complete

✓Quantify 25 genomes alignment loss due to trimming

✓*Streptomyces* 25 genomes progressiveMauve finished running (30 days)

Inversions + Gene Expression:

Queenie is just finishing up creating the final `limma` dataframe I asked for and creating the final list for differences in homologs between the `BLAST` output and my alignment code. However, due to missing gene names, it is sometimes not possible to compare my alignment code to the blast output. I was therefore considering only excluding homologs where blast and my alignment code were different. The rest are either matching, or there is a missing gene name/identifier so it can not be compared. **What do you think about only excluding mismatches between the blast output and my alignment code?**

Substitution Paper

The outliers that were determined for *B. subtilis* subst analysis seem a bit off. The short bars near the origin were considered outliers because they fall within the lower extreme end of the distribution (Figure 7). This loss of data I suspect is what has caused the overall sign for the *B. subtilis* number of subs and distance form the origin of replication to change (Table 1). **Do you think I should count these bars as outliers? Is it "wrong" to remove them when the same code was used to determine outliers for all the other replicons?**

Looking at the selection values, *Streptomyces* and *S. meliloti* Chromosome have a lot of zero values (like last time) (Table 2 and Figures 8 and 9). However, previously ALL of the non-zero ω values for *S. meliloti* chromosome were considered outliers because of the large number of zero ω values. We therefore decided to re-do the selection analysis for *S. meliloti* chromosome without removing any outliers. Now with the new results (from the new RAxML trees) we have non-zero ω values that are not considered outliers, therefore we do not have the same problem as before. **Do you think it is necessary for me to re-do the *S. meliloti* chromosome (and maybe the *Streptomyces*) selection analysis without removing outliers?** The only reason we did it before was because there were no non-zero ω values that were not outliers.

The 26 *Streptomyces* genomes (of unknown strain, therefore very divergent) progressiveMauve has finished running! It took about a month. I have added this to my progressiveMauve computational time projections, but I think because this point is so drastically different from the others, it pushes the estimated exponential line to infinity almost immediately, so my projected timeline does not include this point (Figure 1). To me, this further shows that the more divergent the taxa are, the more infeasible progressiveMauve computation time is, and therefore that it would take too long to align more genomes. **What are your thoughts on this?**

I have finished quantifying how much alignment we lose from our current conservative alignment trimming methods (to ensure homologous genes) when we increase the number of genomes. For 26 *B. subtilis* genomes, we go from a total of 3849474bp before trimming to 4972bp after trimming. This resulted a 99.87% loss of sites. For the 25 *E. coli* genomes, we go from a total of 4378570bp before trimming to 26409bp after trimming. This resulted a 99.4% loss of sites. I think the reason for the poor retention of sites is because progressiveMauve has a really hard time coming up with appropriately aligned blocks when the genomes are divergent. These blocks are therefore not comparing homologous genes, and so most of the sites get thrown out based on our trimming methods. For the 23 *S. meliloti* Chromosomes, I saw a loss of only 35% of sites, which is comparable to what I found with only 6 genomes of *S. meliloti*, where about 25% of sites were lost. I think that this is fairly convincing as to why we can not do this particular analysis (the current pipeline) with more genomes. We would have to re-structure the pipeline completely if we wanted to include more genomes (potentially using just the core genome?). **What are you thoughts on all this? Do you think this will satisfy the reviewer? They did suggest to do a rigorous core genome analysis, which we said we did not do because we wanted to include as much info as possible. Should I be doing a core analysis with more genomes?**

I did the non-linear subst analysis and to me it still looks variable and inconsistent between bacterial replicons (Figures 2 - 6). They do for the most part mirror the linear results (higher/lower subs near the origin and lower/higher subs near the terminus), but with some peaks and valleys in between. **What are your thoughts on this non-linear analysis?**

I finished re-doing the subst analysis with the new phylogenetic trees and *Streptomyces*

genomes. The results (based on the graphs) seem the same. I have not re-entered the regression results into the paper yet, but I suspect they will be the same. The only thing that changed in this analysis were the branchlengts of most trees, the topology of pSymA and pSymB trees, and an increased amount of data used for all replicons (because all block trees matched the overall tree so nothing was thrown out, unlike before.). I will get these results to you as soon as I have them.

This Week

- Queenie: compare blast results and alignments
- Queenie: new dataframe for `limma`
- continue to re-run the selection analysis with the new RAxML trees
- re-run the supplementary subst window analysis (with new trees)
- add new subst results (new tree) to main and supp of paper
- check into gap at beginning of *S. meliloti* chrom subst graph

Next Week

- why do uncharacterized proteins have higher sub rates?
- gap in *E. coli* fig 5
- *B. subtilis* high dS values should not be present
- blast to confirm homologs in subst analysis
- distinction between mutations and substitutions in subst paper intro
- update new code on git (subst paper)

Bacteria and Replicon	Protein Coding Sequences	
	Coefficient Estimate	R^2
E. coli Chromosome	-2.66×10^{-8}***	
B. subtilis Chromosome	2.76×10^{-8}***	
Streptomyces Chromosome	7.19×10^{-8}***	
S. meliloti Chromosome	-6.57×10^{-7}***	
S. meliloti pSymA	2.74×10^{-7}***	
S. meliloti pSymB	1.09×10^{-7}***	

Table 1: one reviewer requested R^2 values for the regressions. For a logistic regression, the R^2 value is not explicitly calculated by the glm() function. Should I calculate this myself? Or do you think the reviewer only wanted the R^2 value on the linear regressions? Logistic regression analysis of the number of substitutions along all protein coding positions of the genome of the respective bacteria replicons. 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 bidirectional 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	8.84	50.70	28.21	50.70
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	6.16	13.17	5.22	13.17

Table 2: Percent of data that was calculated to be an outlier or had a selection variable (dN, dS, and ω) value of zero.

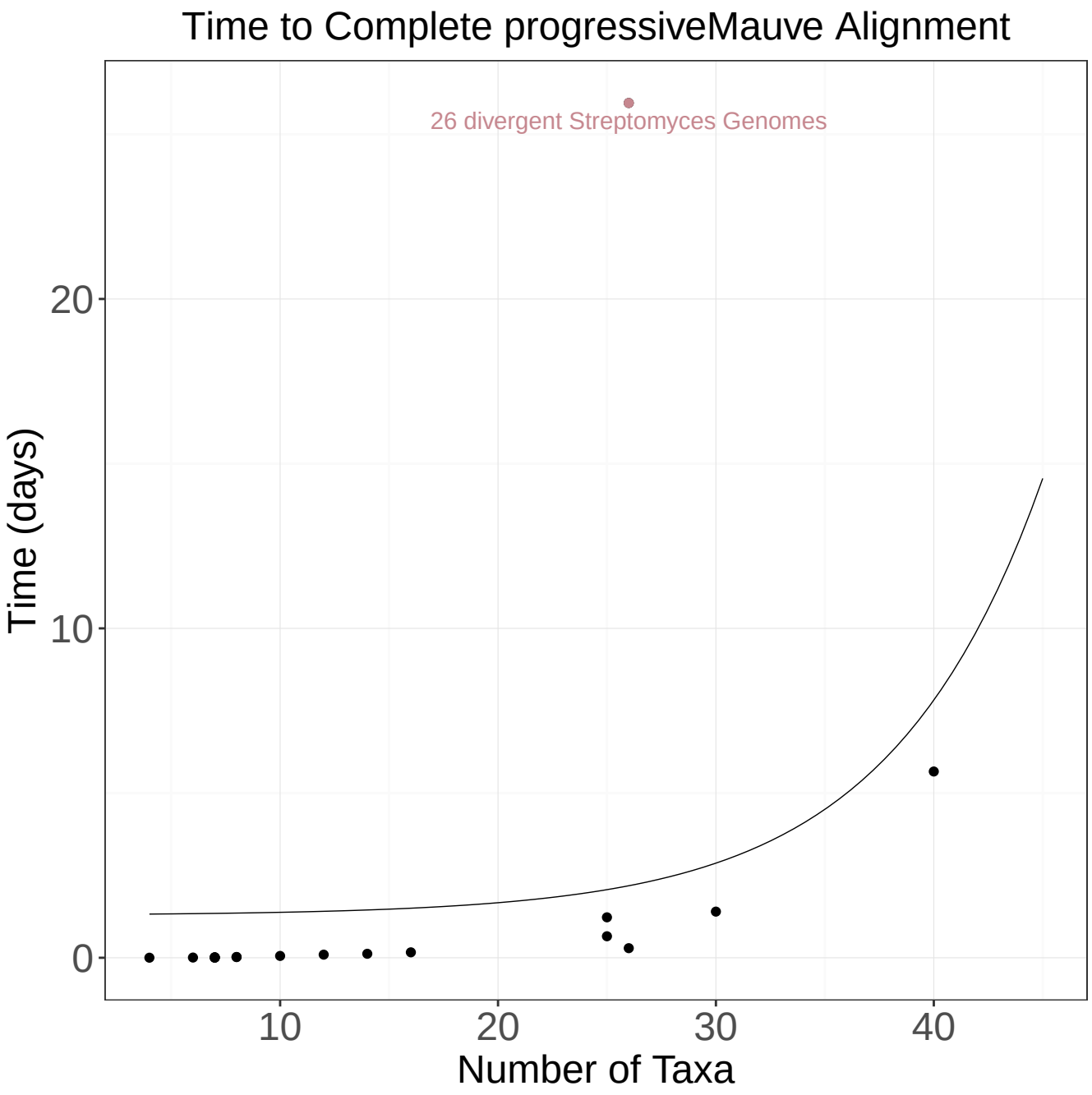
Time to Complete progressiveMauve Alignment
26 divergent Streptomyces Genomes
Time (days)
0
10
20
10
20
30
40
Number of Taxa

Figure 1

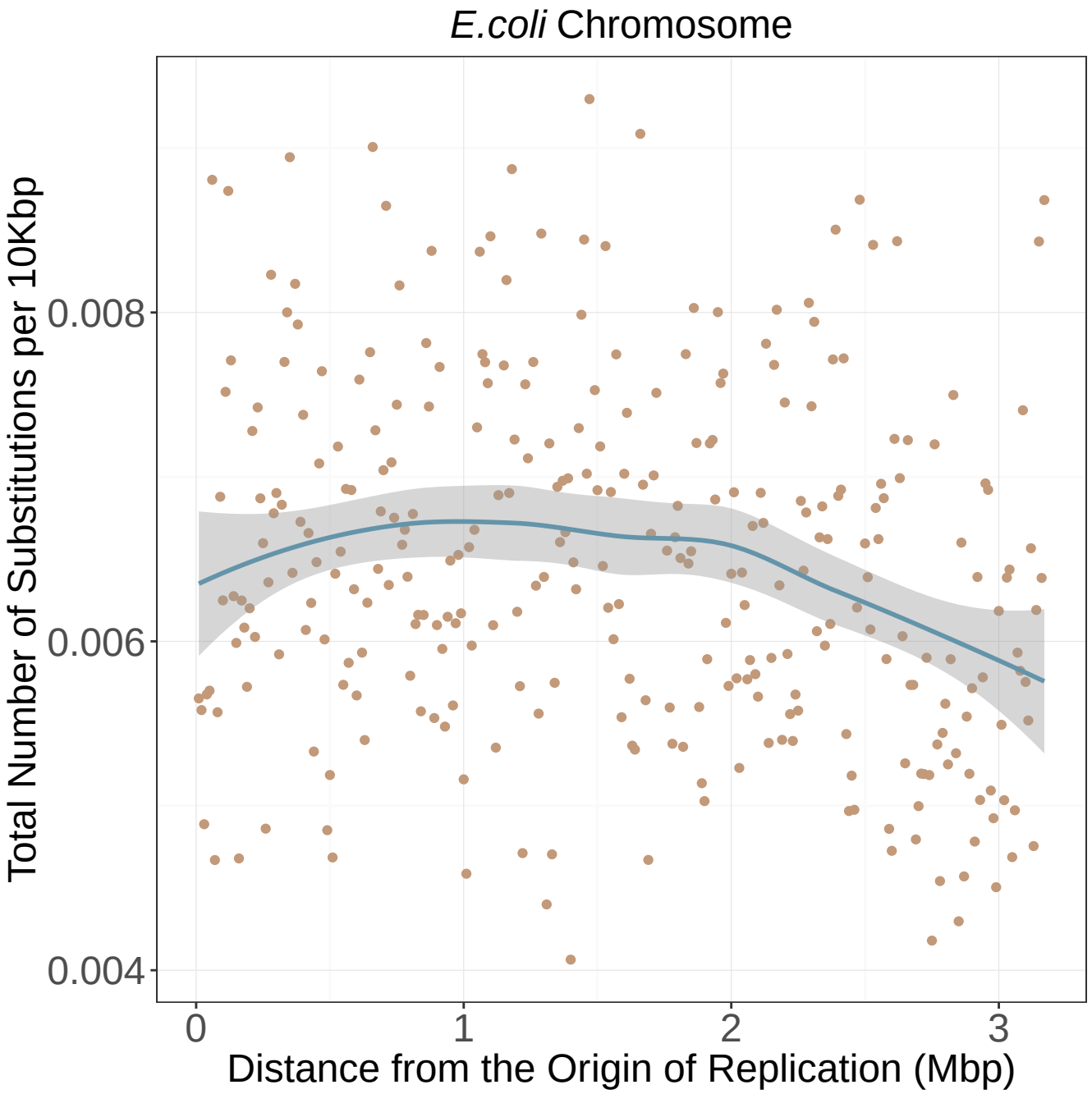

Figure 2: The graph shows the total number of substitutions weighted by the total number of protein-coding sites per 10Kbp segments of the genome. Each of these individual values are represented by beige coloured circles. A non-linear trend line (using the `geom_smooth()` function in `R`), was fit to 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.

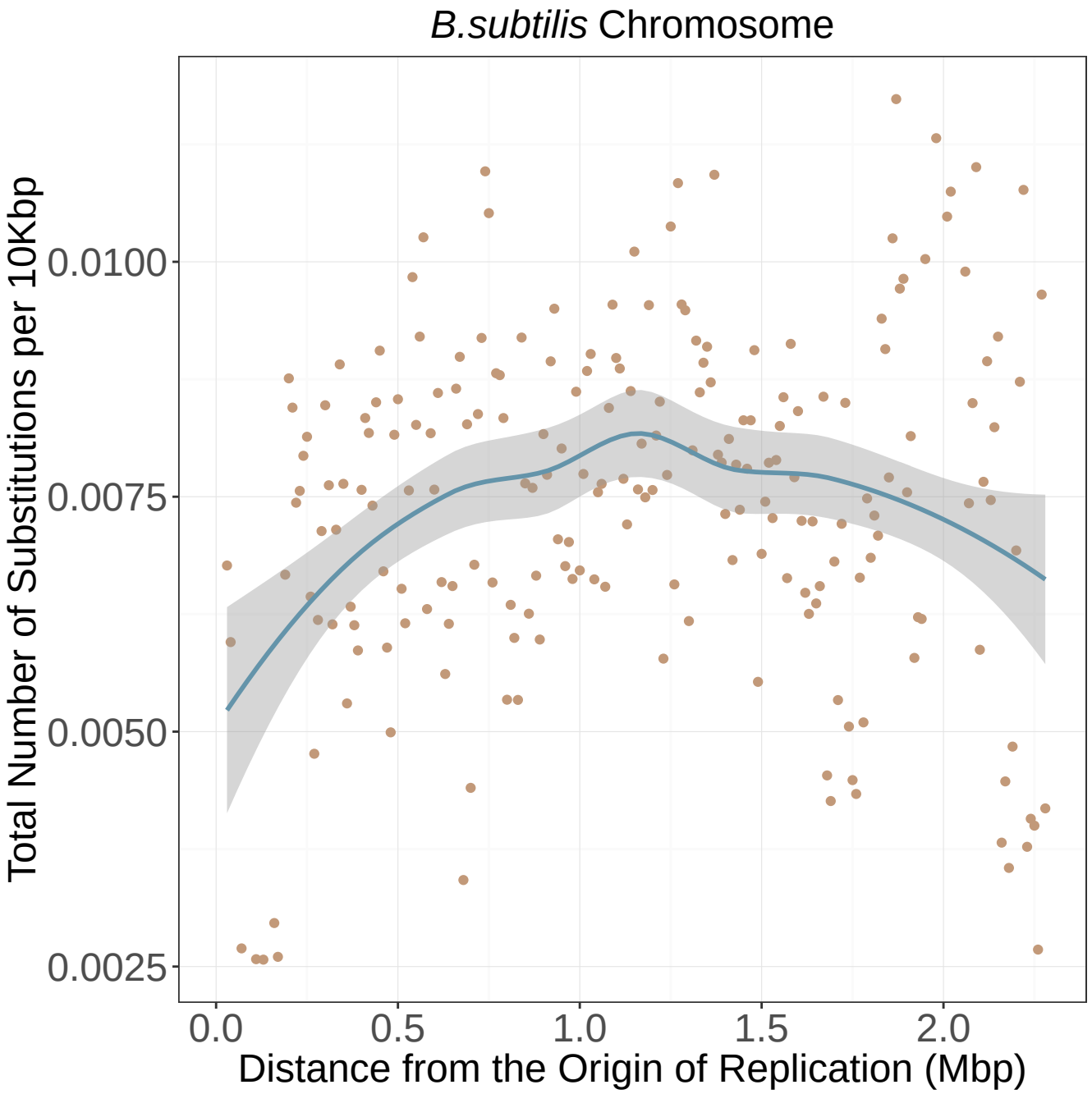

Figure 3: The graph shows the total number of substitutions weighted by the total number of protein-coding sites per 10Kbp segments of the genome. Each of these individual values are represented by beige coloured circles. A non-linear trend line (using the `geom_smooth()` function in `R`), was fit to 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.

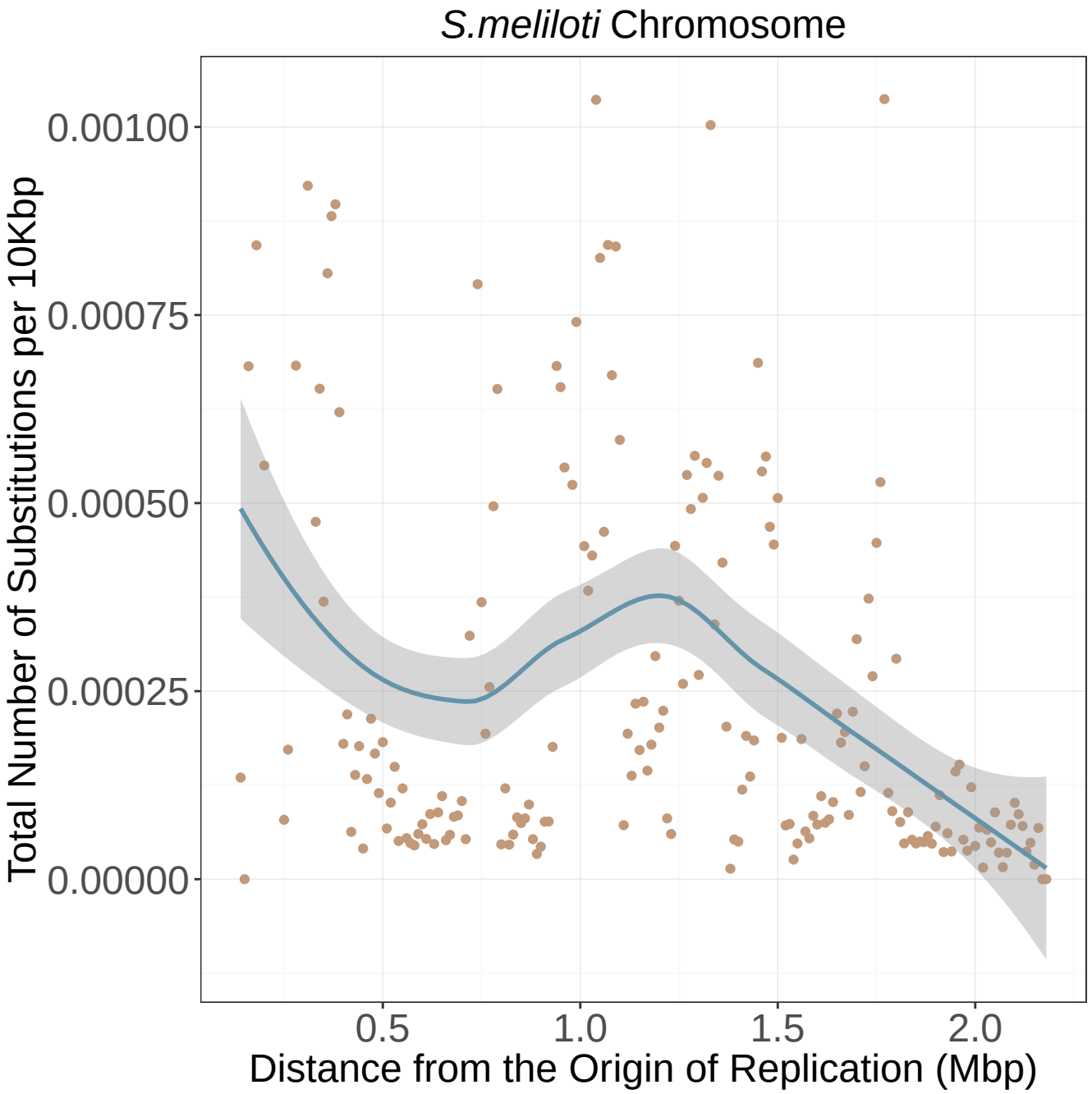

Figure 4: The graph shows the total number of substitutions weighted by the total number of protein-coding sites per 10Kbp segments of the genome. Each of these individual values are represented by beige coloured circles. A non-linear trend line (using the `geom_smooth()` function in `R`), was fit to 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.

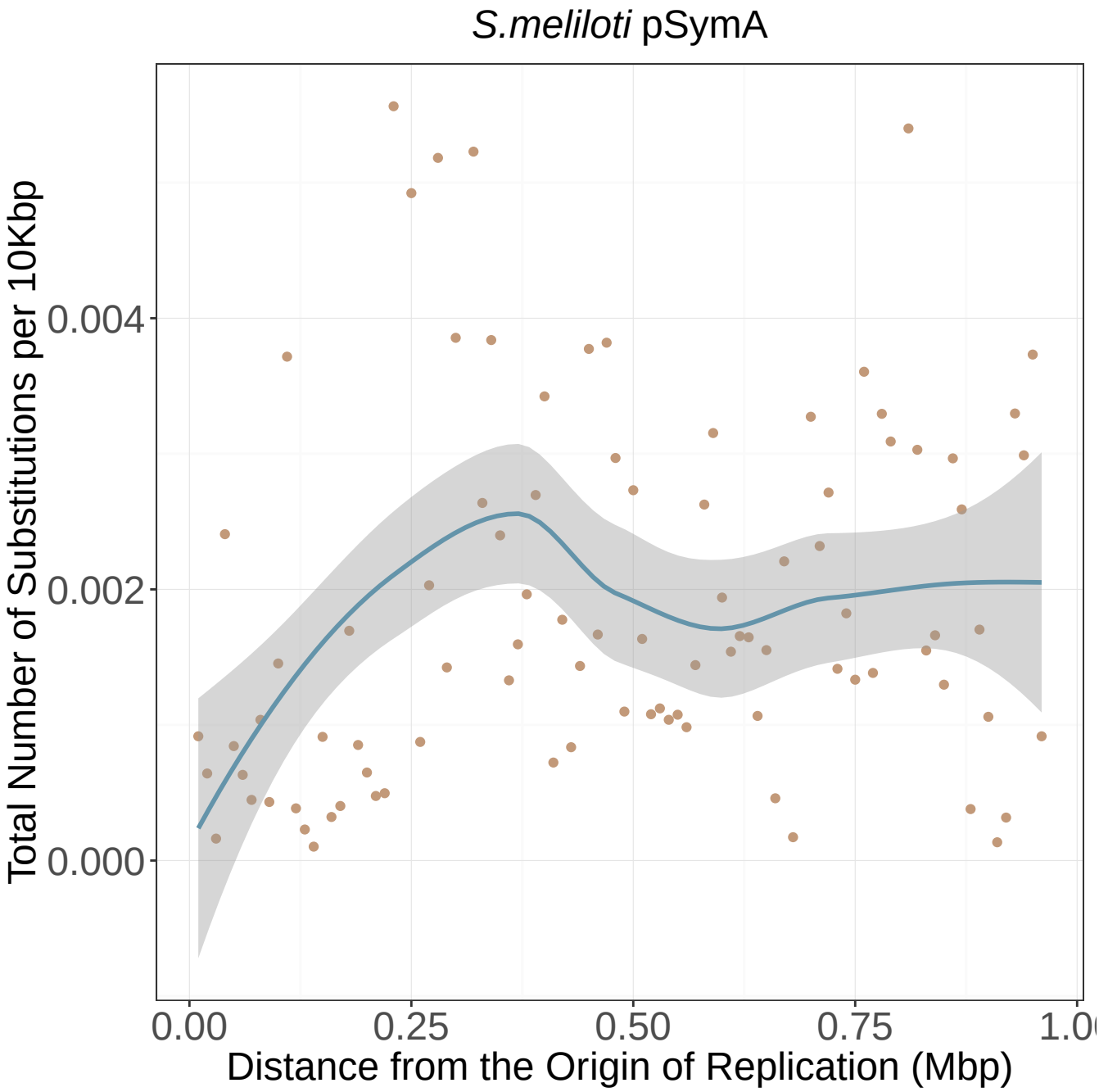

Figure 5: The graph shows the total number of substitutions weighted by the total number of protein-coding sites per 10Kbp segments of the genome. Each of these individual values are represented by beige coloured circles. A non-linear trend line (using the `geom_smooth()` function in `R`), was fit to 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.

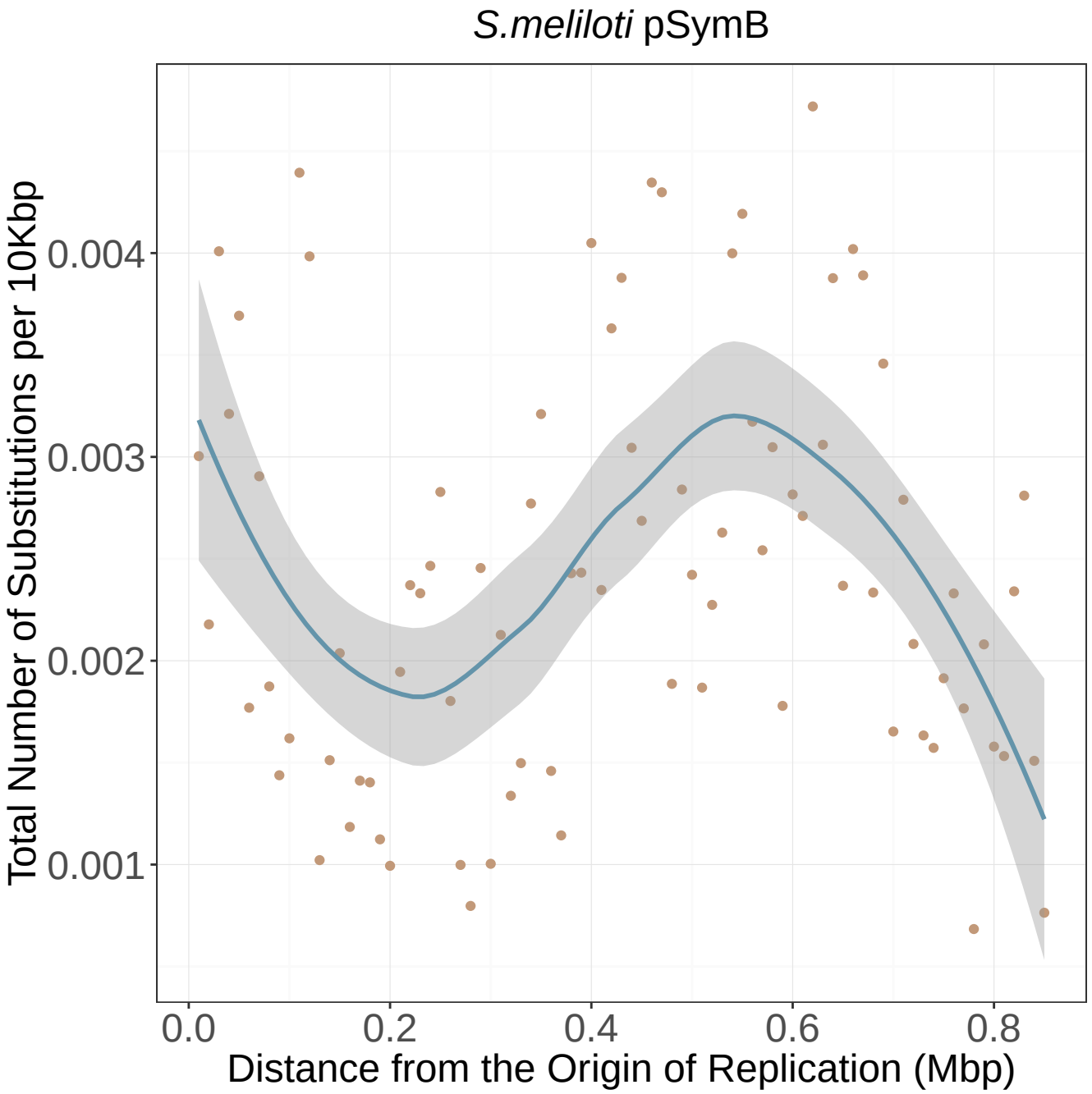

Figure 6: The graph shows the total number of substitutions weighted by the total number of protein-coding sites per 10Kbp segments of the genome. Each of these individual values are represented by beige coloured circles. A non-linear trend line (using the `geom_smooth()` function in `R`), was fit to 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.

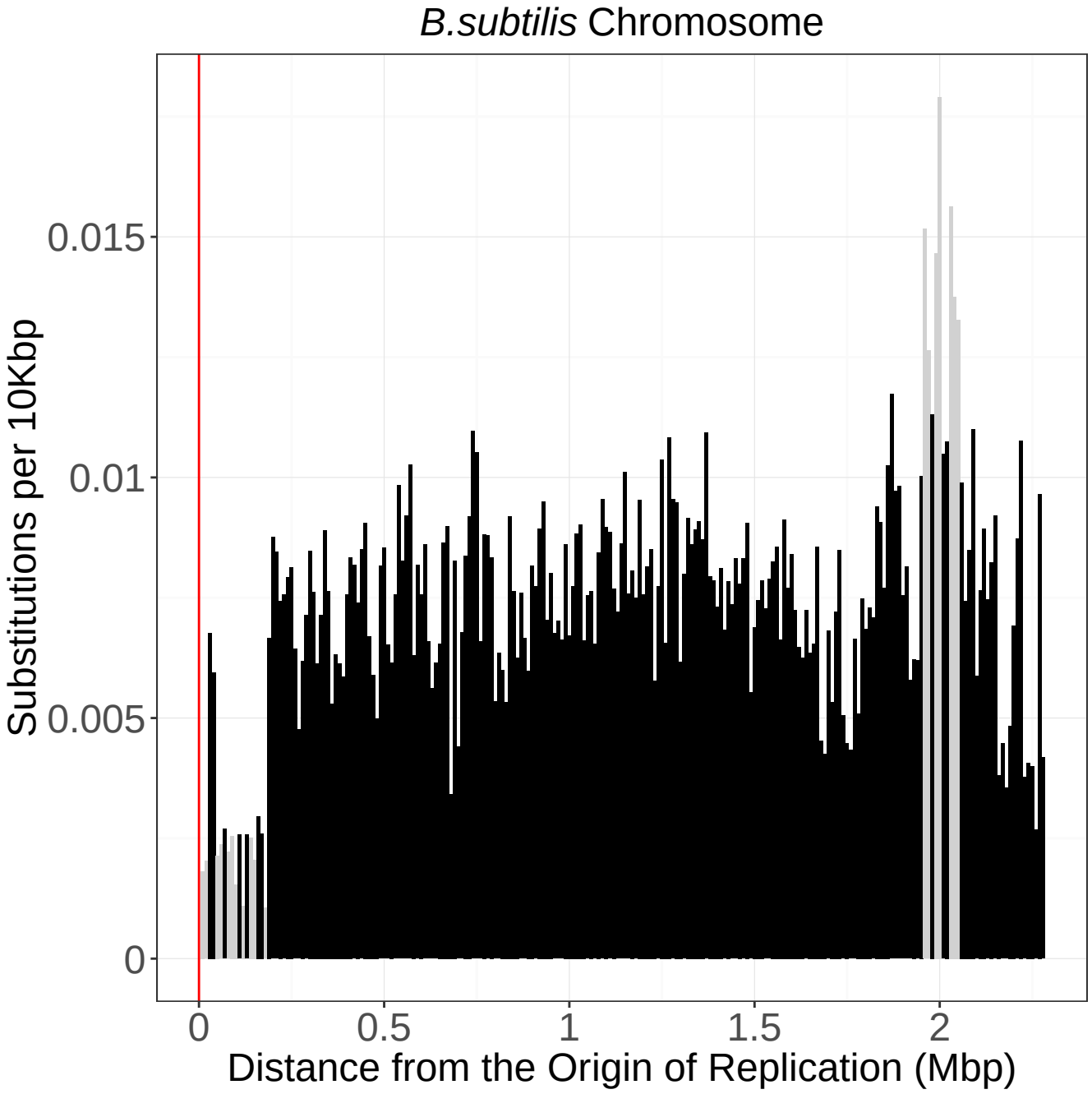

Figure 7: *B. subtilis* subs graph

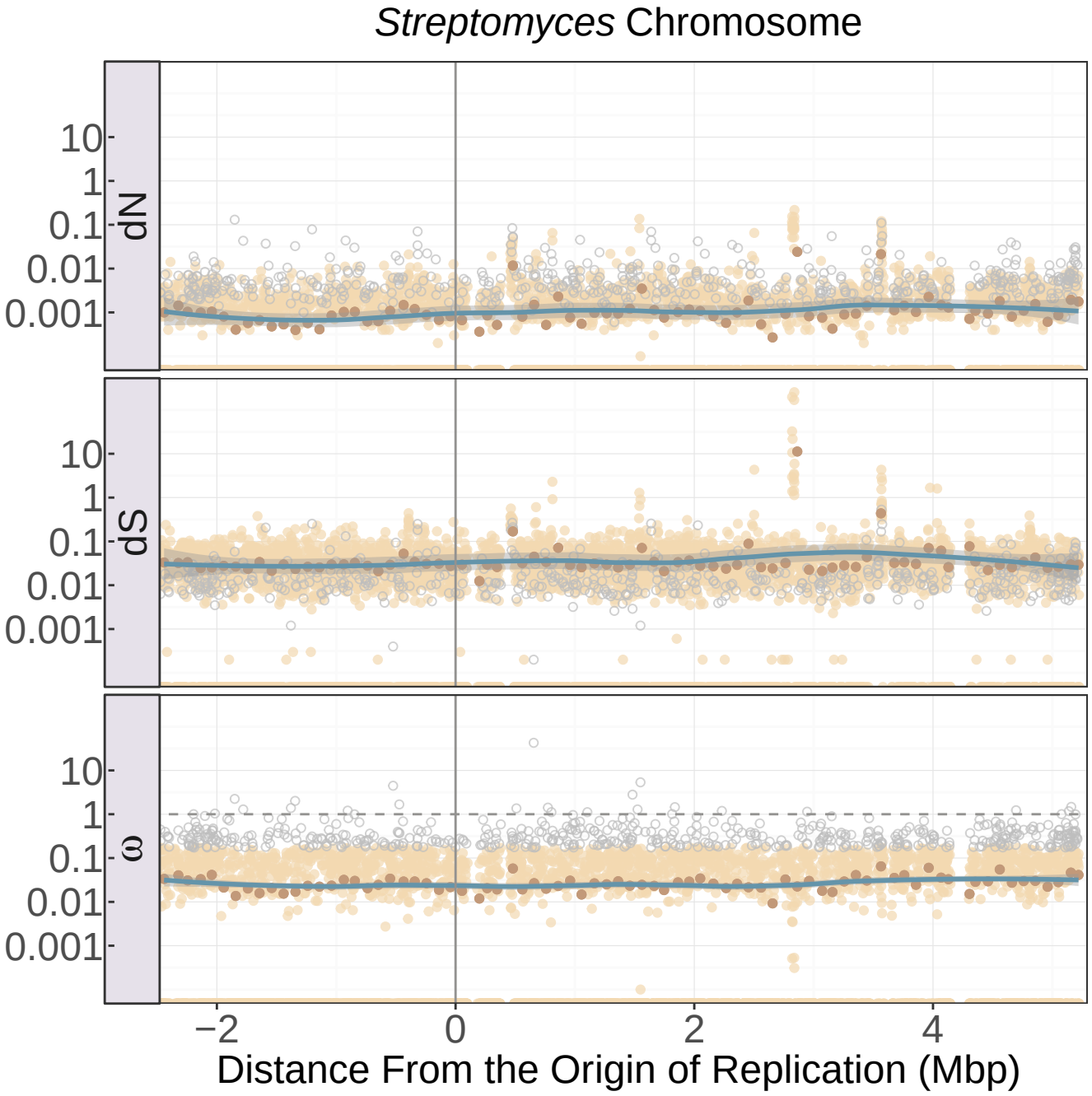
Streptomyces Chromosome
dN
dS
ω
10
1
0.1
0.01
0.001
−2
0
2
4
Distance From the Origin of Replication (Mbp)

Figure 8

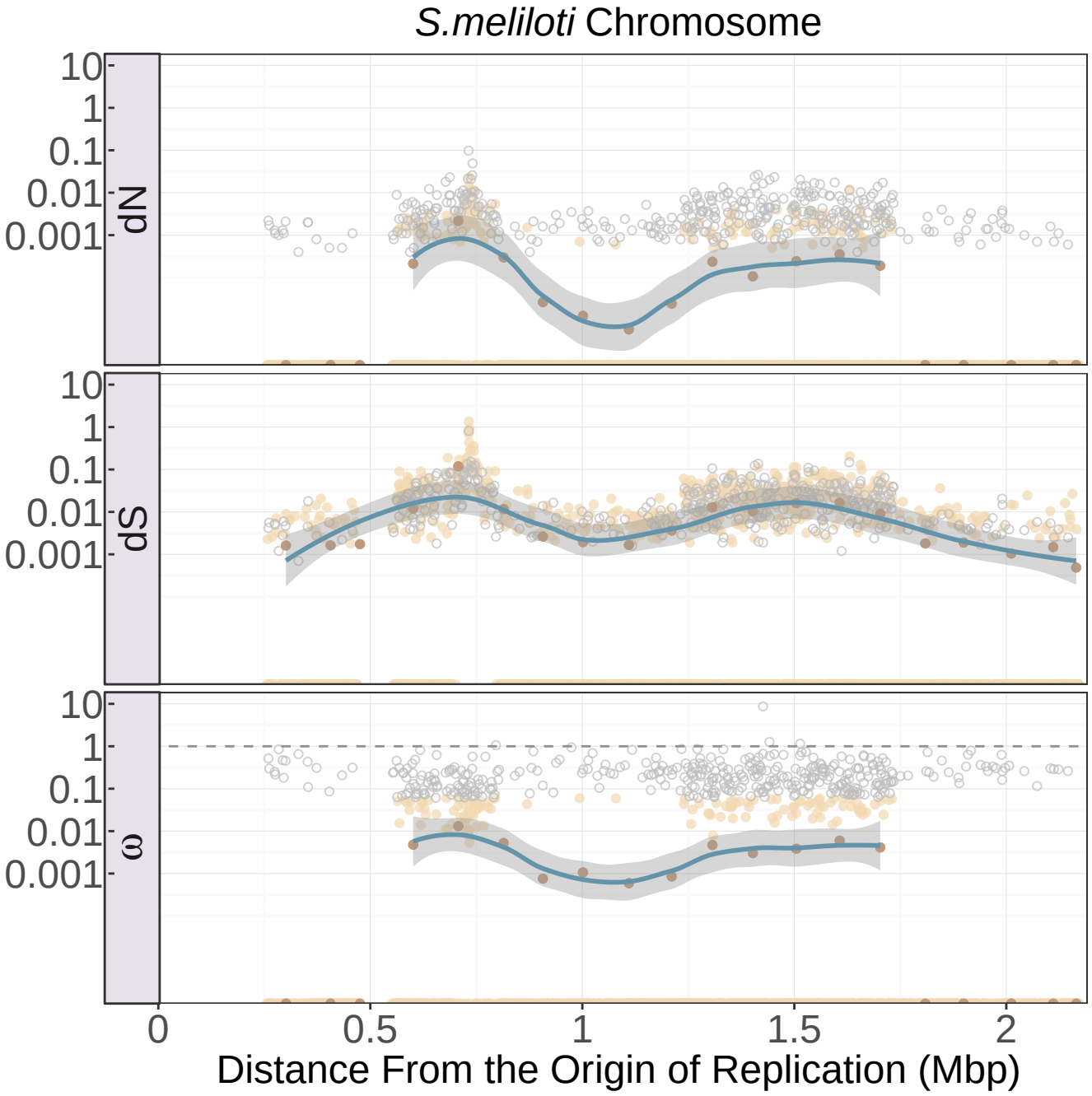
S.meliloti Chromosome
dN
dS
ω
10
1
0.1
0.01
0.001
0
0.5
1
1.5
2
Distance From the Origin of Replication (Mbp)

Figure 9