

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-parametric 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 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 branchlengths 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
- write about TEs and repeated elements in cover letter
- re-run inversions mapping based on correct subst code
- re-run subst analysis for 25genome bac based on correct subst code

- non-parametric analysis for subst analysis
- quantify trimming loss (more subst genomes)
- selection analysis for subst analysis (6 genomes per bac)

Next Week

- Queenie: new dataframe for `limma`
- previous subst papers datasets (can I re-do?)
- 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

Authors	Species Name	Genomic Structure
Galardini et al. 2013	<i>Sinorhizobium</i> (14 genomes)	1 chromosome, 2 plasmids
Morrow et al. 2012	<i>Burkholderia</i> (4 genomes)	3 chromosomes
Cooper et al. 2010	<i>Burkholderia</i> (6 genomes)	3 chromosomes
	<i>Bordetella</i> (4 genomes)	single circular chromosome
	<i>Vibrio</i> (4 genomes)	2 chromosomes
	<i>Xanthomonas</i> (5 genomes)	single circular chromosome
Couturier et al. 2006	126 species of bacteria (only chromosome)	only primary chromosome for multi-repliconic bacteria
Flynn et al. 2010	<i>Sulfolobus</i> (6 genomes)	single circular chromosome
Sharp et al. 2005	80 species of bacteria (genomic)	only primary chromosome for multi-repliconic bacteria

Table 1: Previous studies looking at substitutions/mutation rate.

Datasets: Taxa Per Block	% of Blocks that are		
	Inverted	Inverted and Differentially Expressed	Increased in Gene Expression in the Inverted Sequences
All 4	68.15	8.66	60.61
At least 3	68.23	8.91	57.14
At least 2	68.02	8.96	58.33

Table 2: Percent of blocks in categories for various datasets (blocks with all 4 taxa, at least 3 taxa, or at least 2 taxa). The second column is any block that had at least one sequences that was inverted. The last column only deals with blocks that had at least one inverted sequence and had a significant difference in gene expression (column 3).

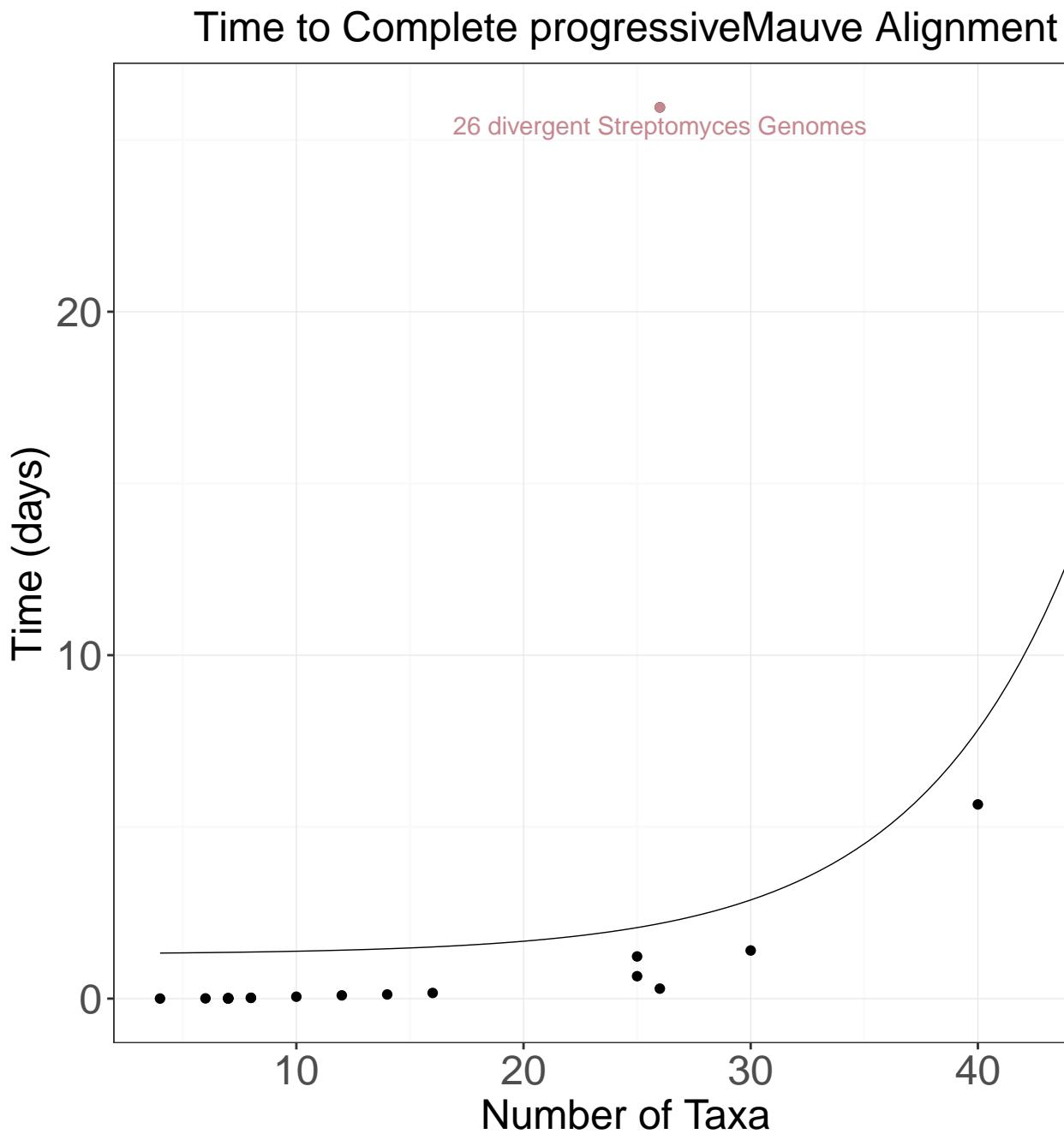


Figure 1

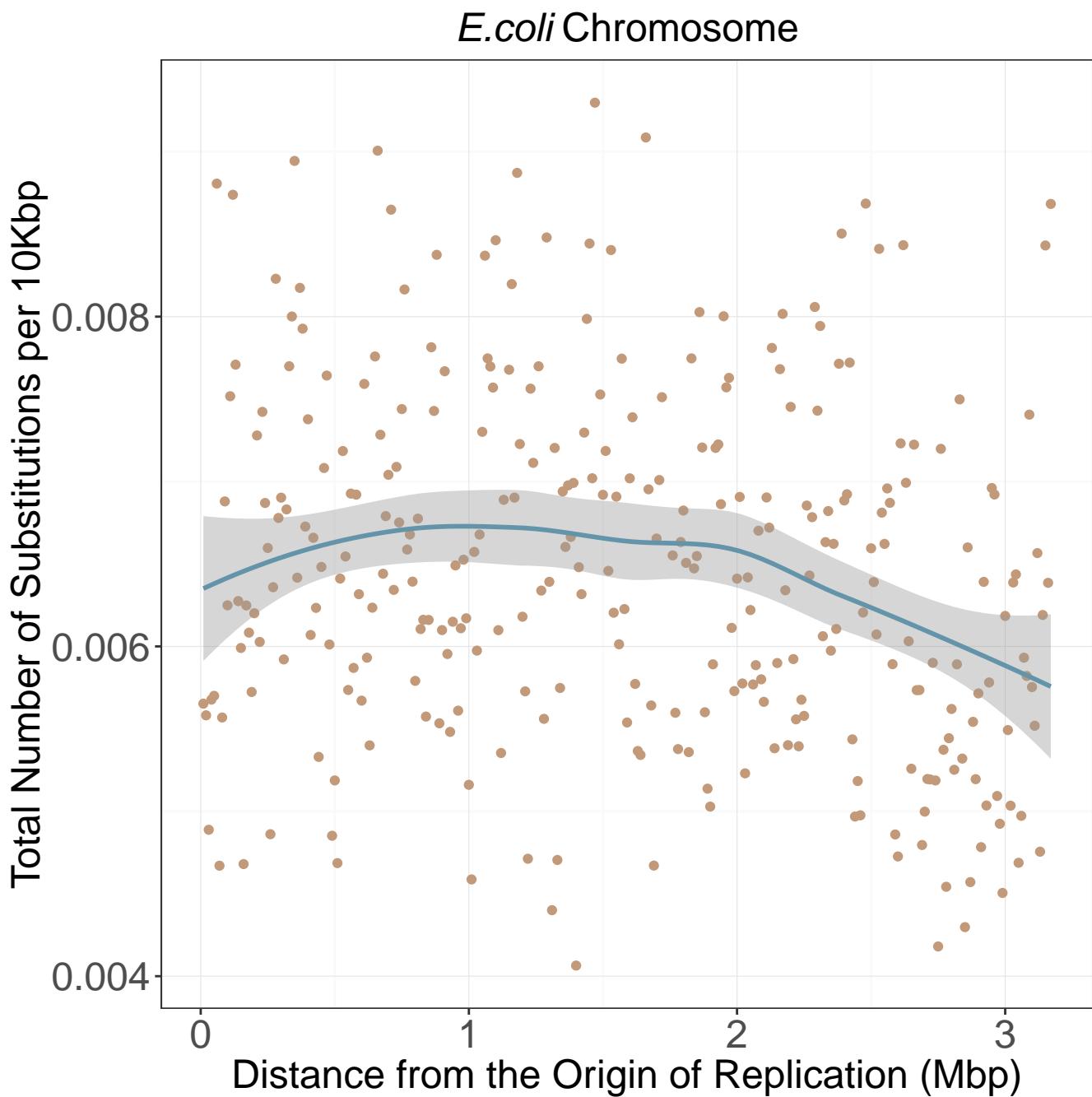


Figure 2: The graph shows the total number of substitutions weighted by the total number of protein-coding sites per 10Kbp segments of the genome. Each of these individual values are represented by beige coloured circles. A non-linear trend line (using the `geom_smooth()` function in R), was fit to these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

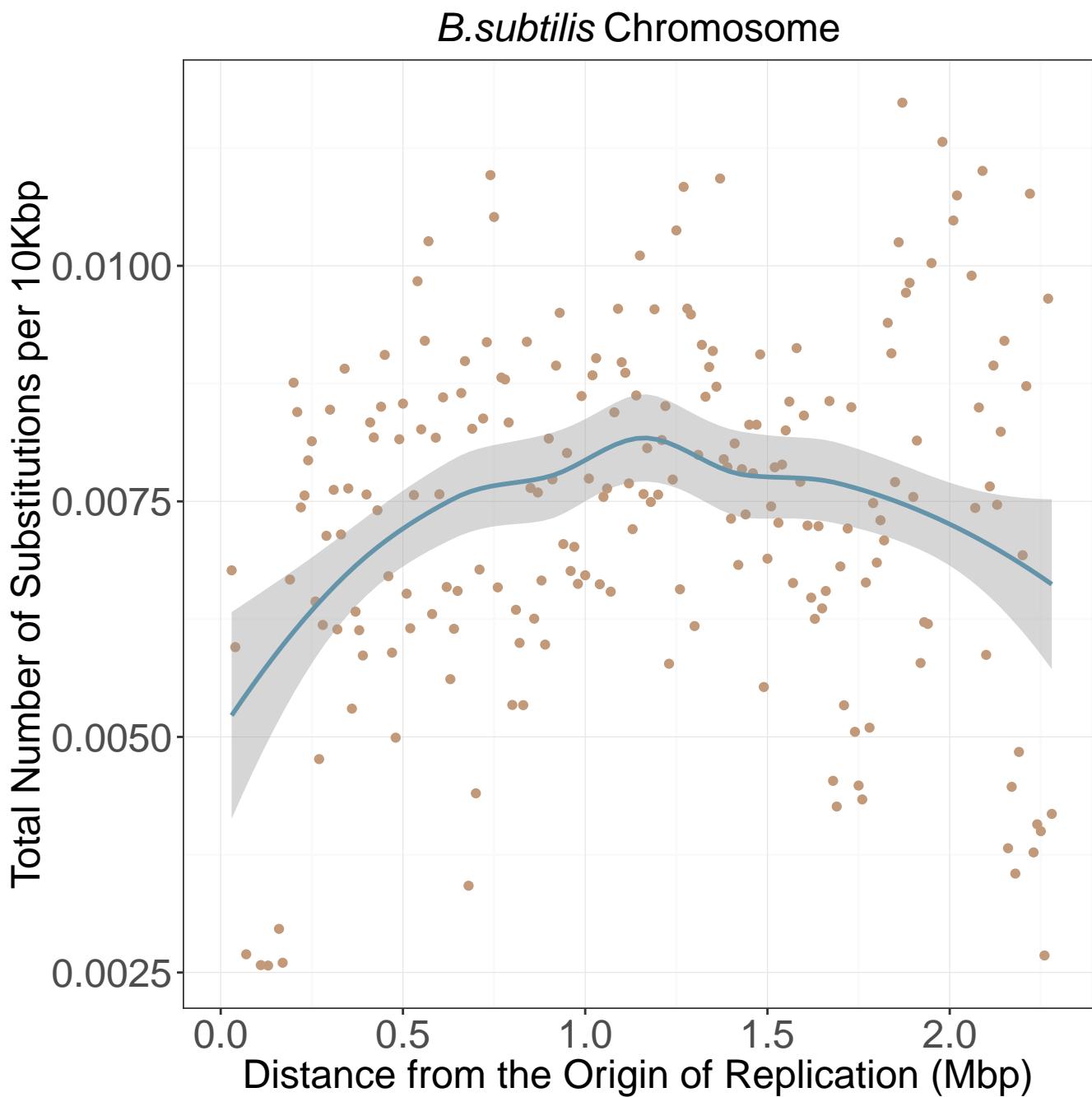


Figure 3: The graph shows the total number of substitutions weighted by the total number of protein-coding sites per 10Kbp segments of the genome. Each of these individual values are represented by beige coloured circles. A non-linear trend line (using the `geom_smooth()` function in R), was fit to these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

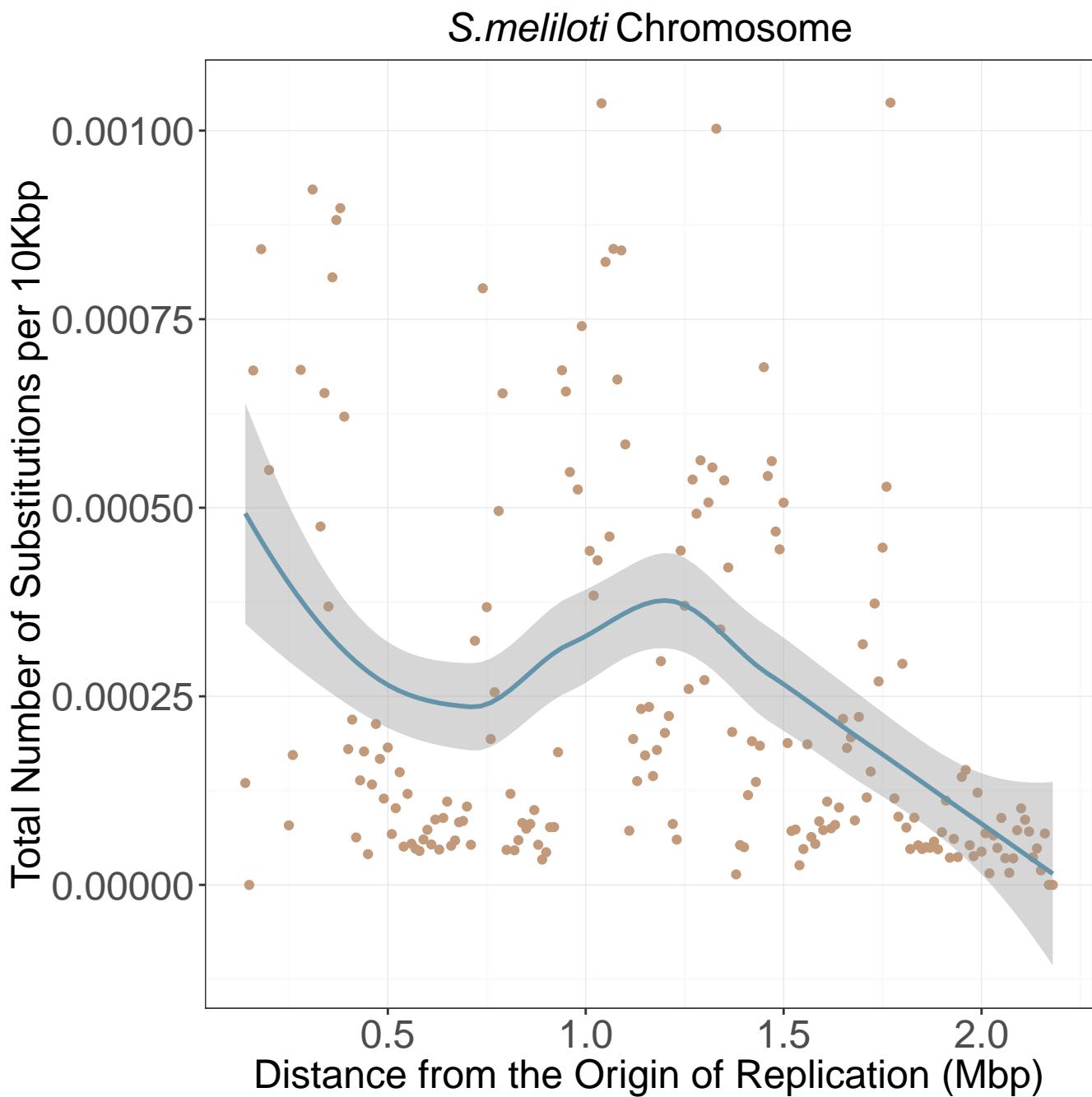


Figure 4: The graph shows the total number of substitutions weighted by the total number of protein-coding sites per 10Kbp segments of the genome. Each of these individual values are represented by beige coloured circles. A non-linear trend line (using the `geom_smooth()` function in R), was fit to these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

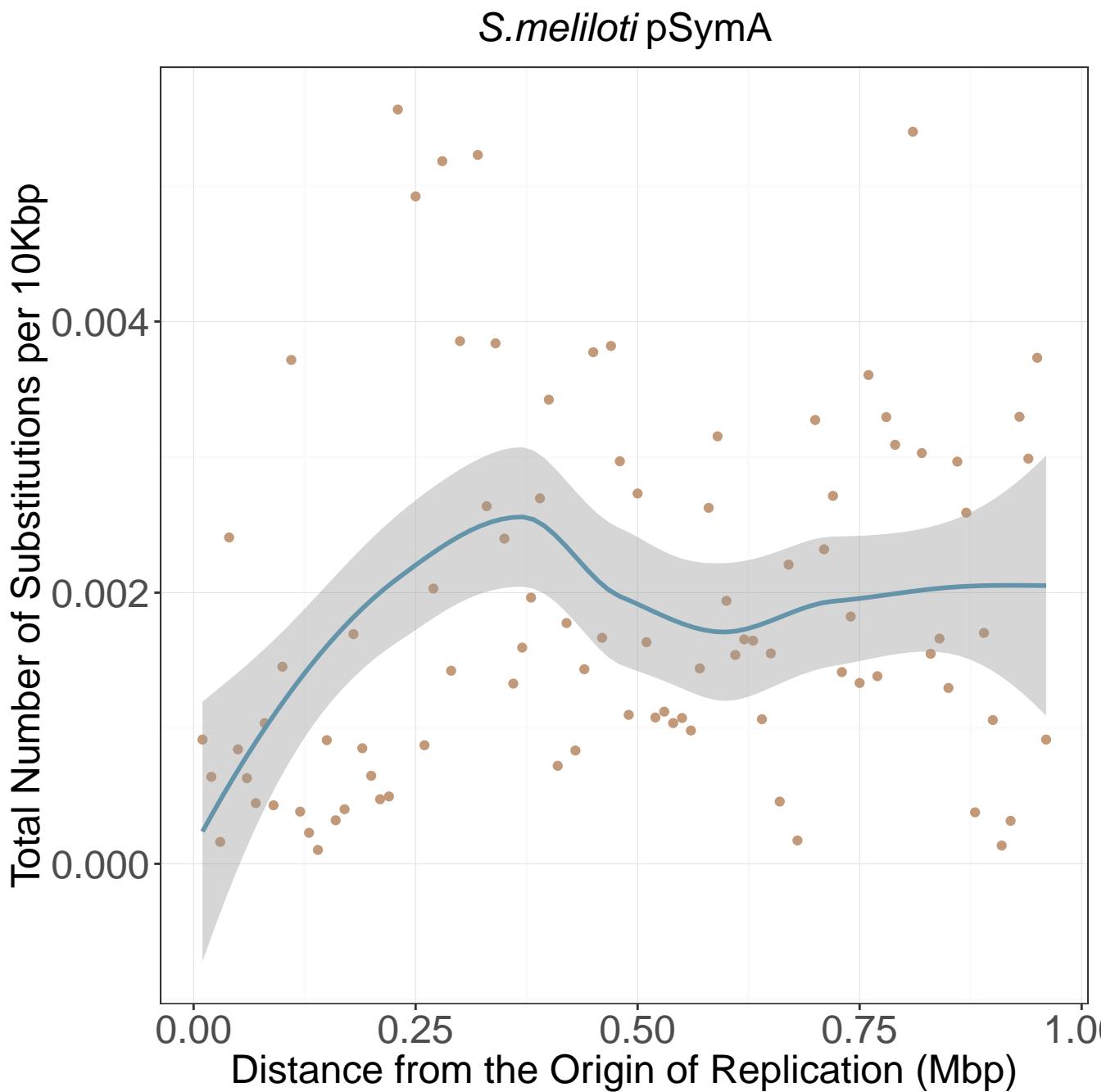


Figure 5: The graph shows the total number of substitutions weighted by the total number of protein-coding sites per 10Kbp segments of the genome. Each of these individual values are represented by beige coloured circles. A non-linear trend line (using the `geom_smooth()` function in R), was fit to these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

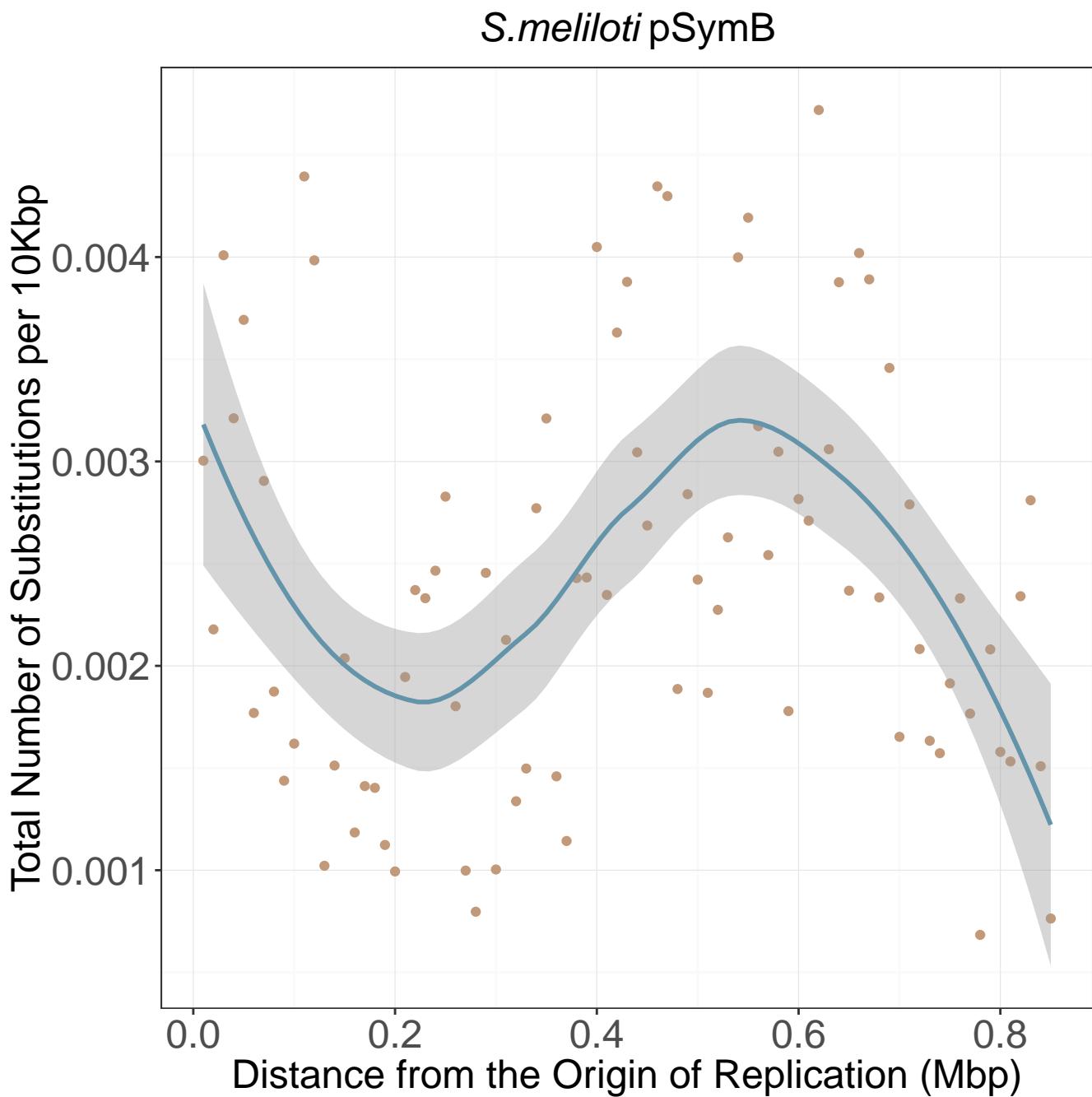
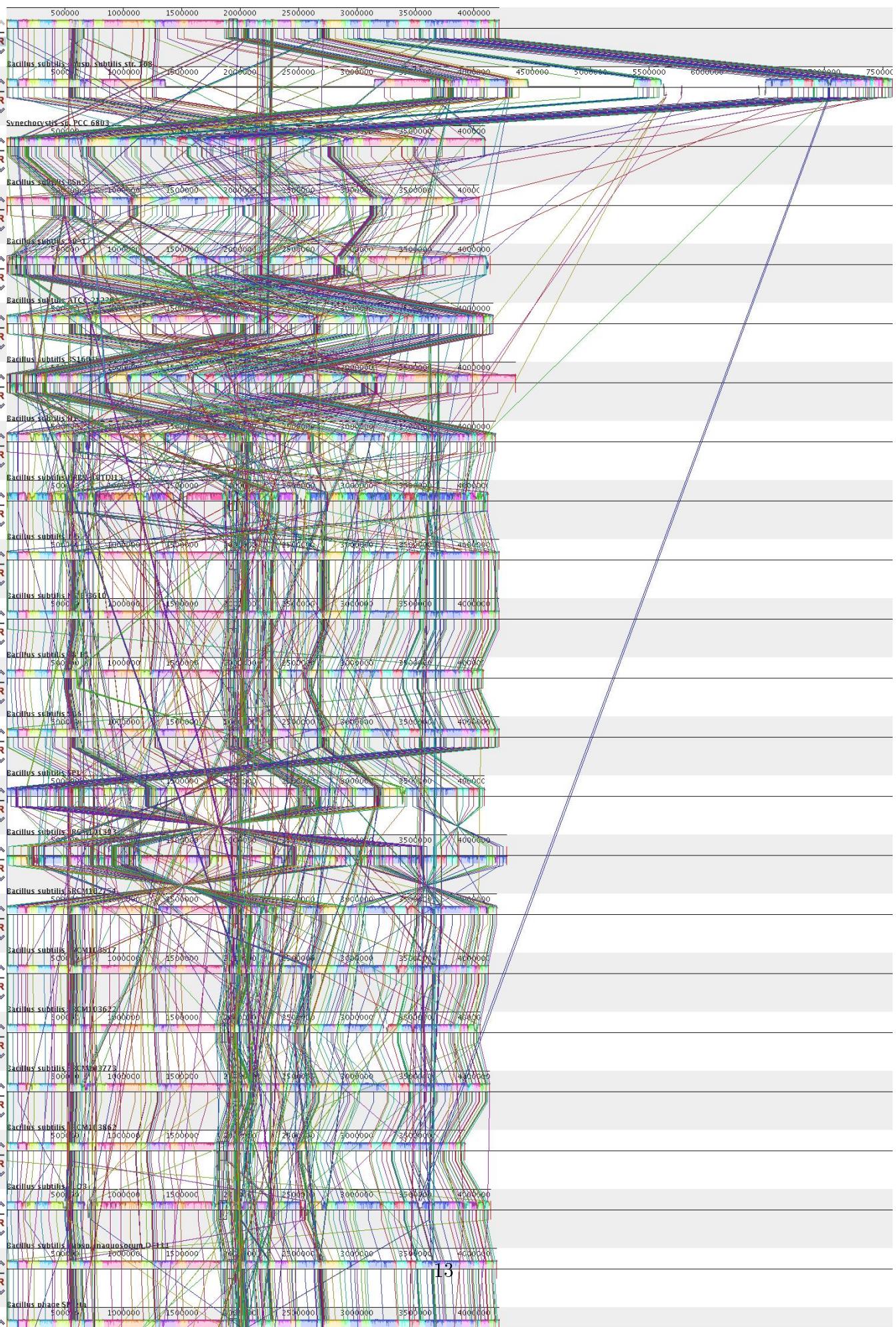


Figure 6: The graph shows the total number of substitutions weighted by the total number of protein-coding sites per 10Kbp segments of the genome. Each of these individual values are represented by beige coloured circles. A non-linear trend line (using the `geom_smooth()` function in R), was fit to these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





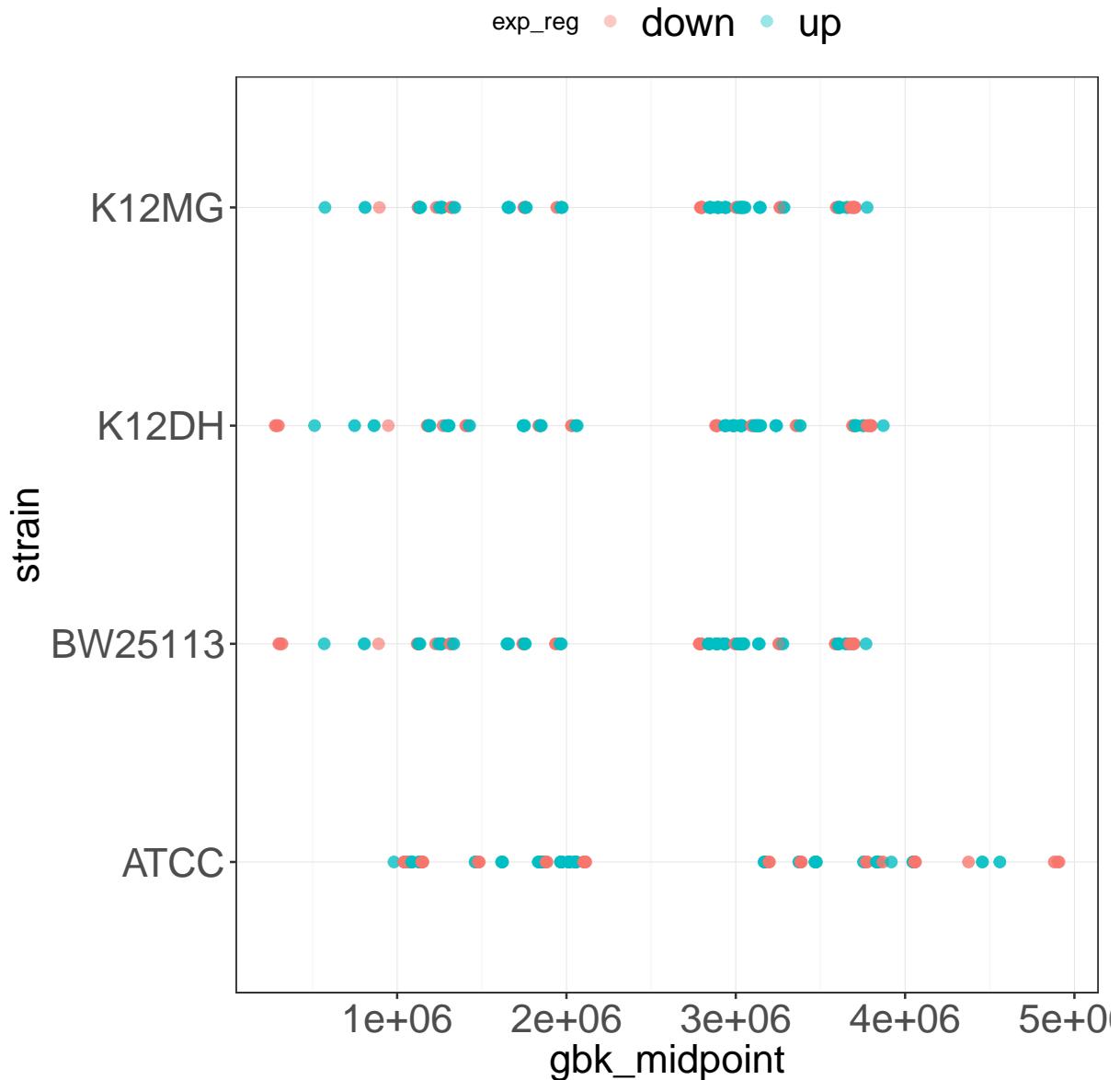


Figure 9: Test graphic for the visualization of inversions and distance from the origin of replication. Each dot represents a gene in a block where there is a significant difference in gene expression between inverted and non-inverted sequences within that block. The points are coloured based on if the inverted sequences have higher expression (“up”) or lower expression (“down”) compared to the non-inverted sequences. Genomic position is on the x-axis with NO bidirectional replication accounted for.