

Subs Paper Things to Do:

- 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

Substitutions:

✓ Submitted! Woo!

✓ double check github is up to date

Inversions + Gene Expression:

- ✓ link blast ID to gbk ids
- ✓ find reciprocal best blast hits (minus dealing with multiple RBBHs)
- ✓ Queenie: looked into overlapping blocks occurrences

General:

- ✓ edited one section of dissertation intro

Inversions + Gene Expression:

Queenie looked into how often blocks overlap but are not completely within another block (5.4%). From these blocks, 26% - 37.3% of genes are split between two blocks. I think that this is likely PARSNP not always getting the end/start position of a block completely correct. I am having Queenie check what kinds of genes are being split (pseudo genes..etc) and if these genes are within an inversion or not.

I have been working on the blast portion of this analysis (verifying the PARSNP output). I have a code that will retrieve the reciprocal best blast hits and print out the gbk gene name (to help with comparison to the parsnp output). However, it appears as though the gene names from blast do not always match what is in the gbk file. This might be a version thing so I will keep looking into it.

We also discussed creating a map between BW and K-12 to show which genes are similar. I will be working on that this week.

This Week

- create map between BW and K-12 genes
- figure out how to assess synteny when the gene names do not match
- get Queenie started on creating lists of which genes match (from all strains) in the parsnp alignment

Next Week

- edit another section of dissertation intro
- get Queenie to create a plot of the inversions
- continue to work on blast synteny and BW/K-12 map

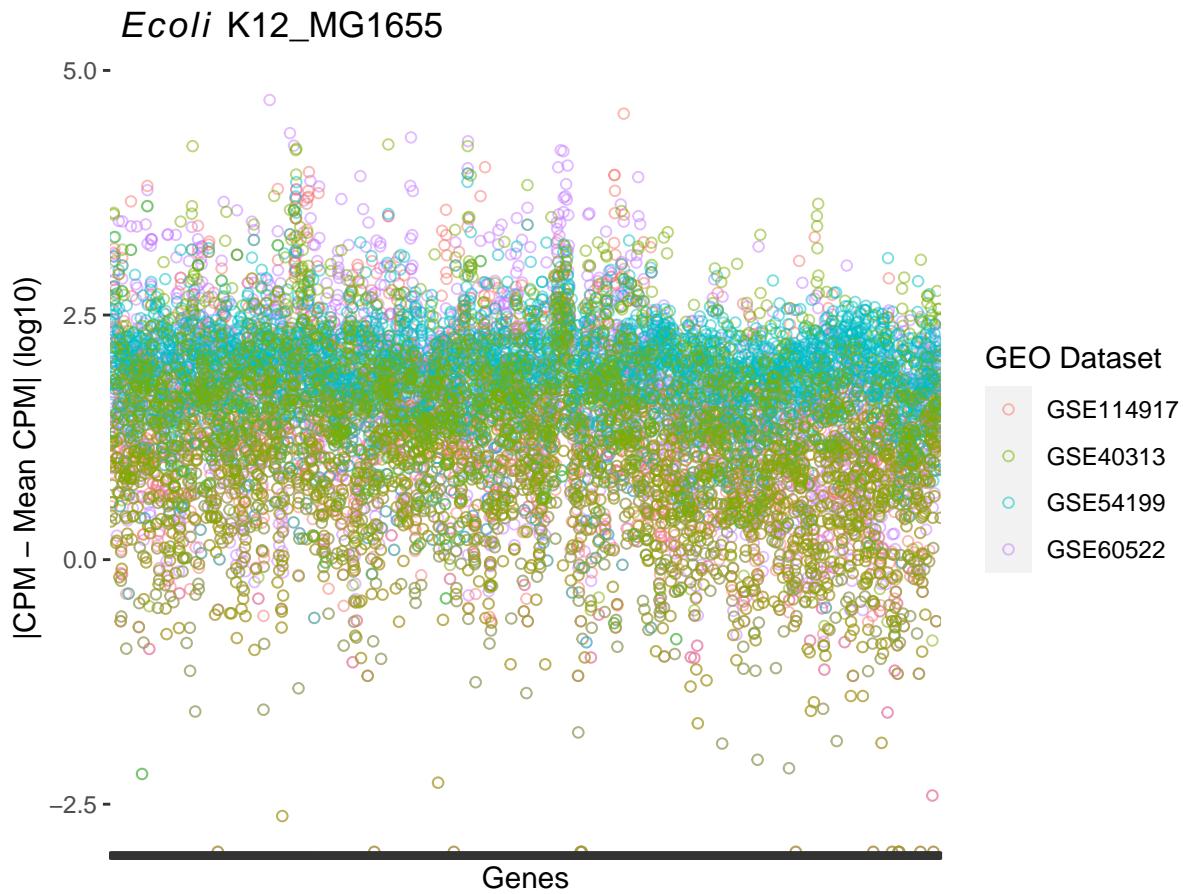


Figure 1

Bacteria and Replicon	Genome Average		
	dS	dN	ω
<i>S. meliloti</i> Chrom + <i>A. tumefaciens</i>	12.5529	0.0553	0.0265
<i>E. coli</i> Chromosome	0.2387	0.0101	0.0441
<i>B. subtilis</i> Chromosome	0.4201	0.0243	0.0714
<i>Streptomyces</i> Chromosome	0.0458	0.0011	0.0335
<i>S. meliloti</i> Chromosome	0.0029	0	0
<i>S. meliloti</i> pSymA	0.0835	0.0099	0.1645
<i>S. meliloti</i> pSymB	0.0940	0.0084	0.1142

Table 1: 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.

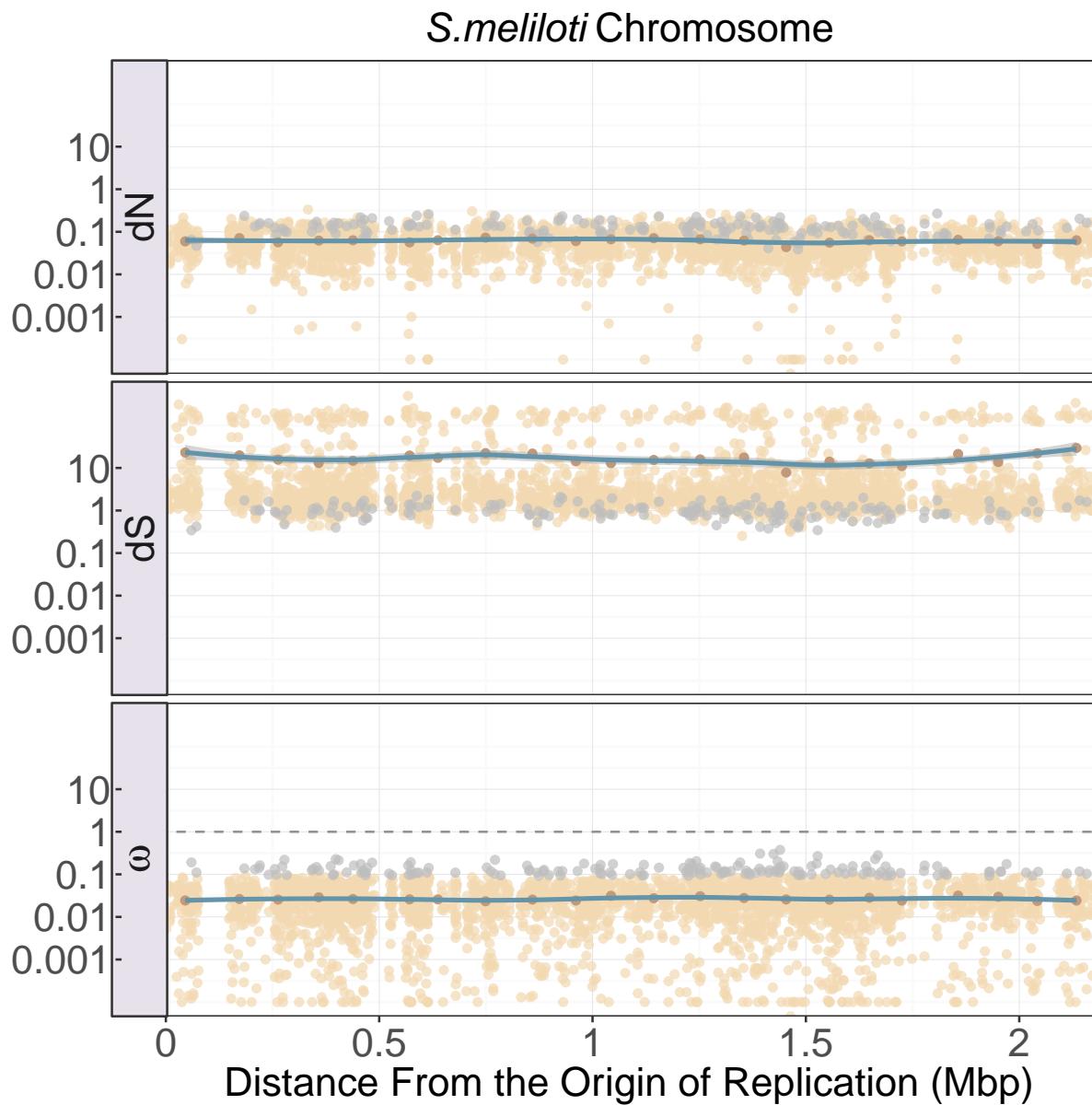
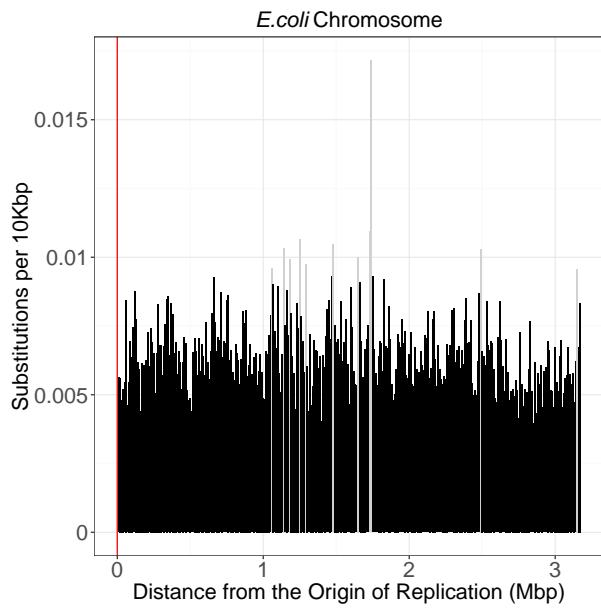


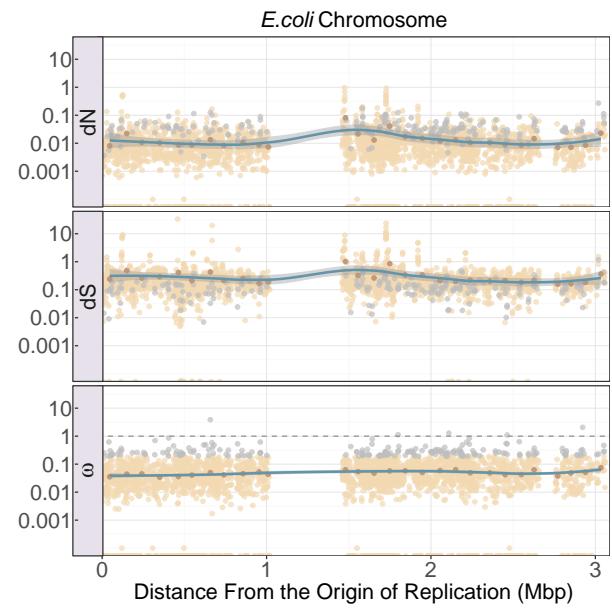
Figure 2: dN , dS , and ω values for *S. meliloti* chromosomes and *A. tumefaciens*.

Bacteria and Replicon	Average Number of Substitutions per bp
<i>E. coli</i> Chromosome	1.97×10^{-4}
<i>B. subtilis</i> Chromosome	1.93×10^{-4}
<i>Streptomyces</i> Chromosome	2.74×10^{-6}
<i>S. meliloti</i> Chromosome	9.72×10^{-5}
<i>S. meliloti</i> pSymA	6.54×10^{-5}
<i>S. meliloti</i> pSymB	1.99×10^{-4}

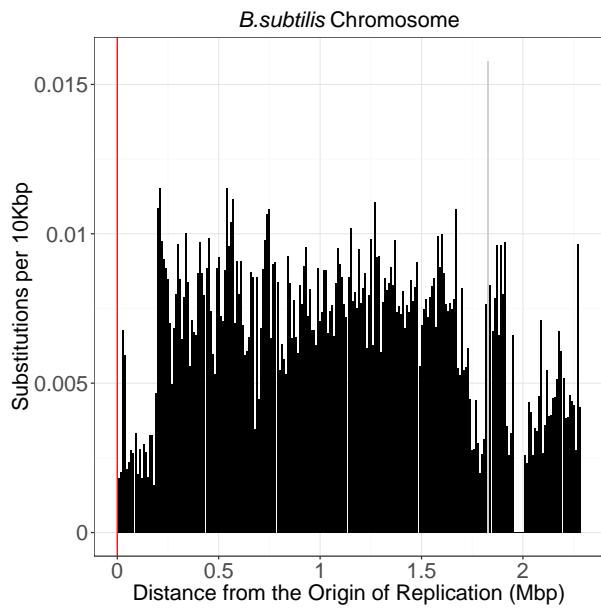
Table 2: Average number of protein coding substitutions calculated per base across all bacterial replicons. Outliers and missing data was not included in the calculation.



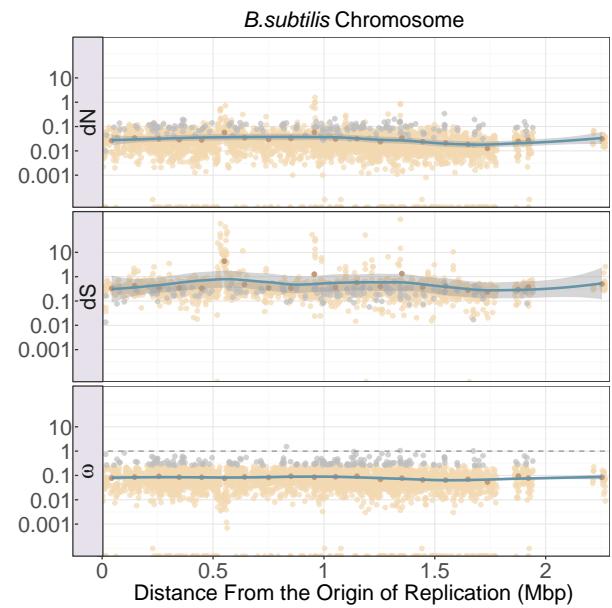
(a)



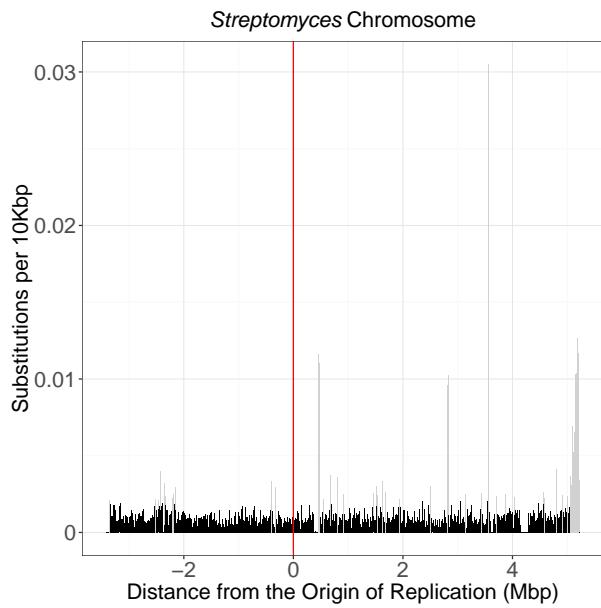
(b)



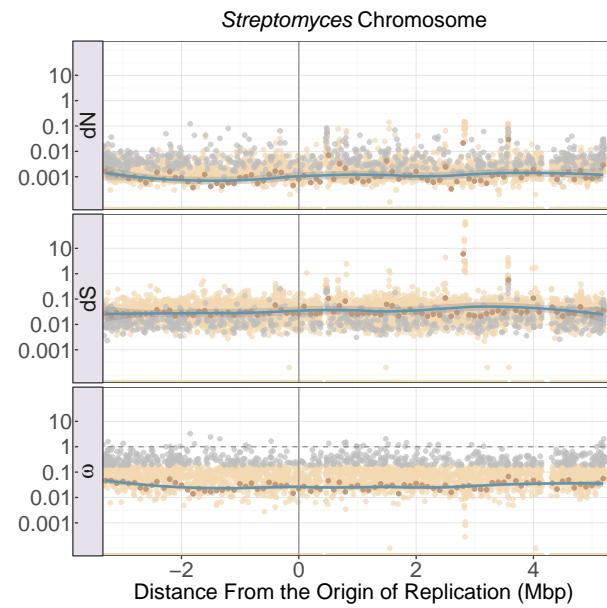
(a)



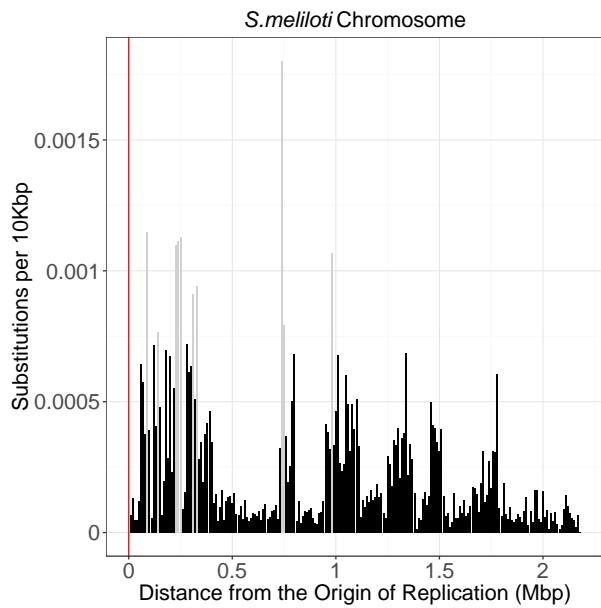
(b)



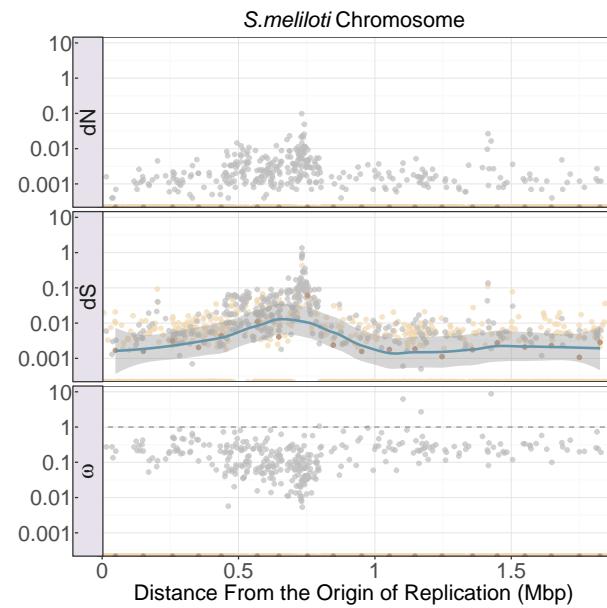
(a)



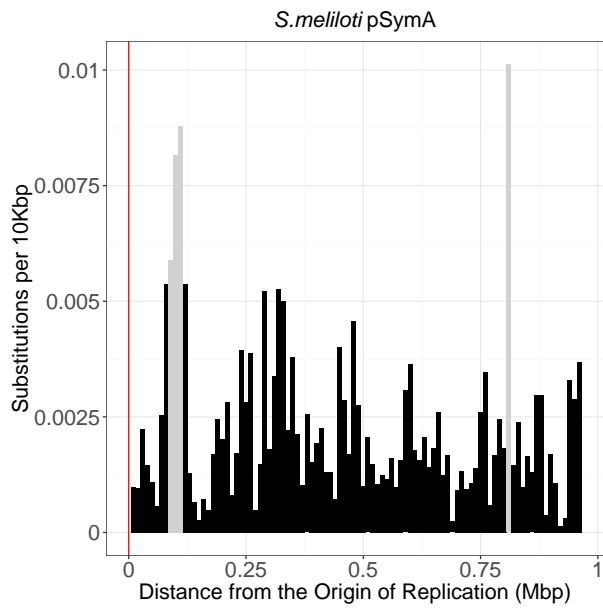
(b)



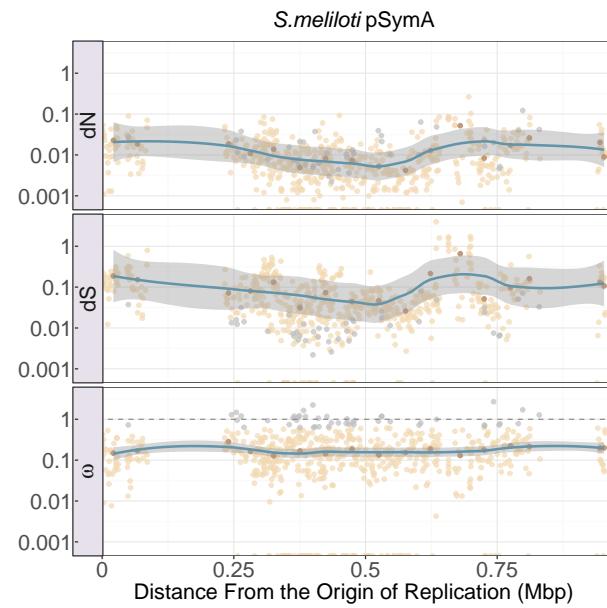
(a)



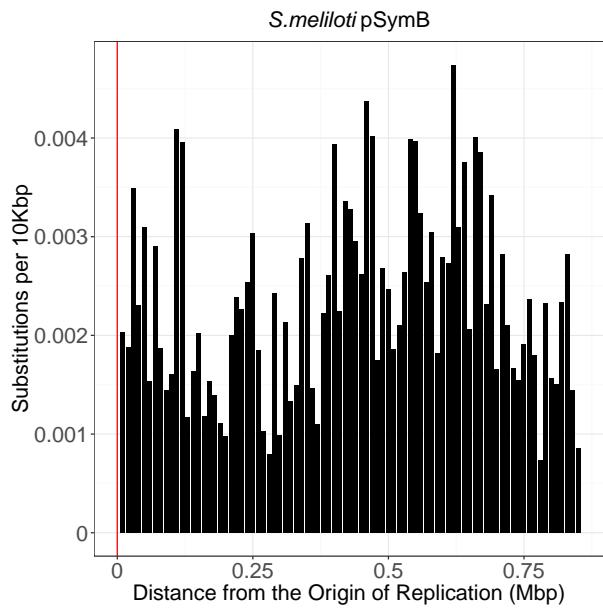
(b)



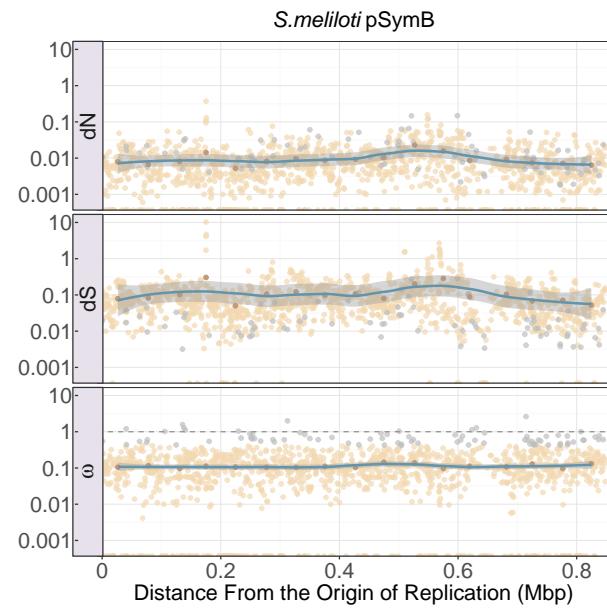
(a)



(b)



(a)



(b)

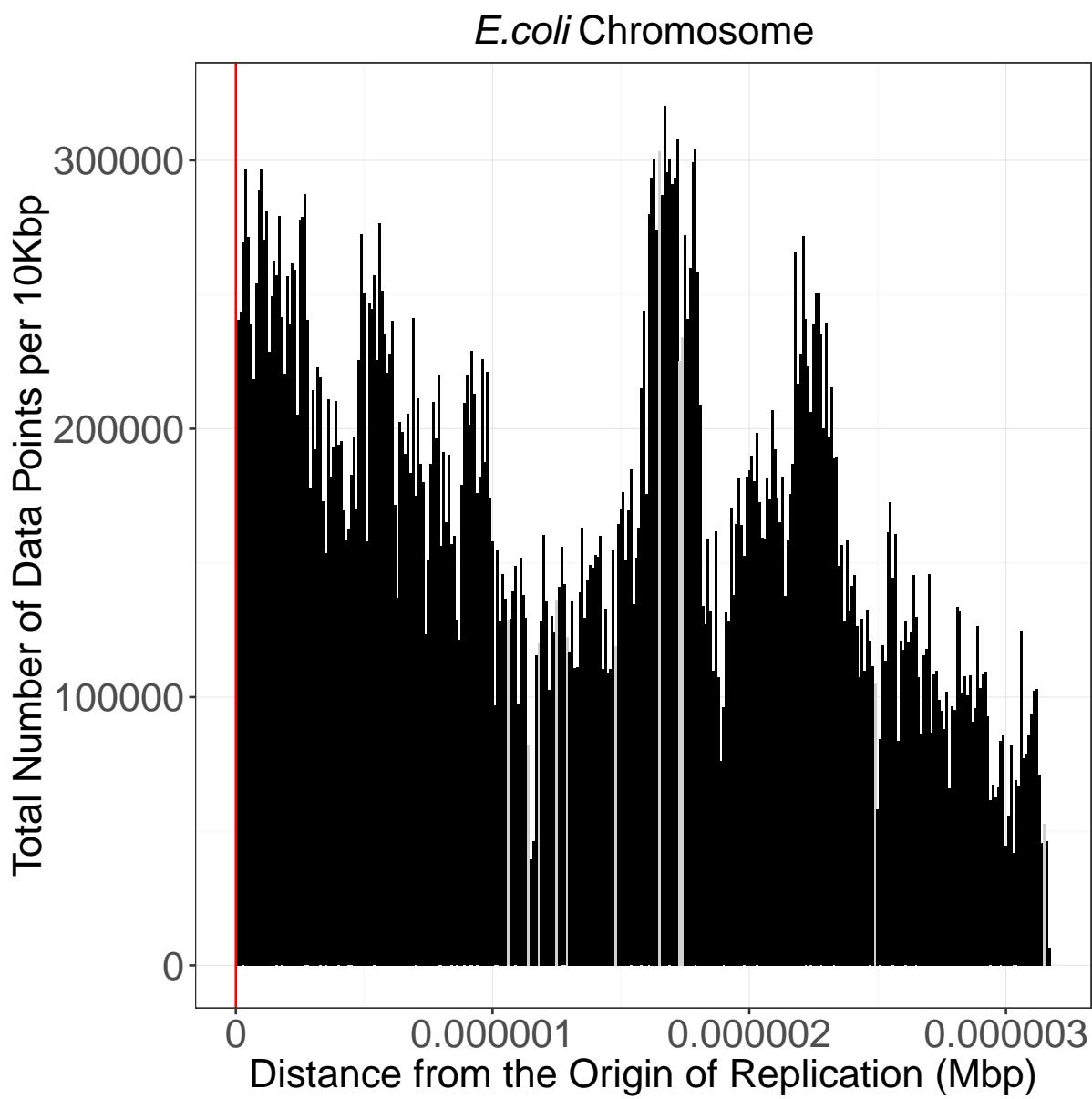


Figure 9: Distribution of total number of substitution data points per 10Kbp in genome.

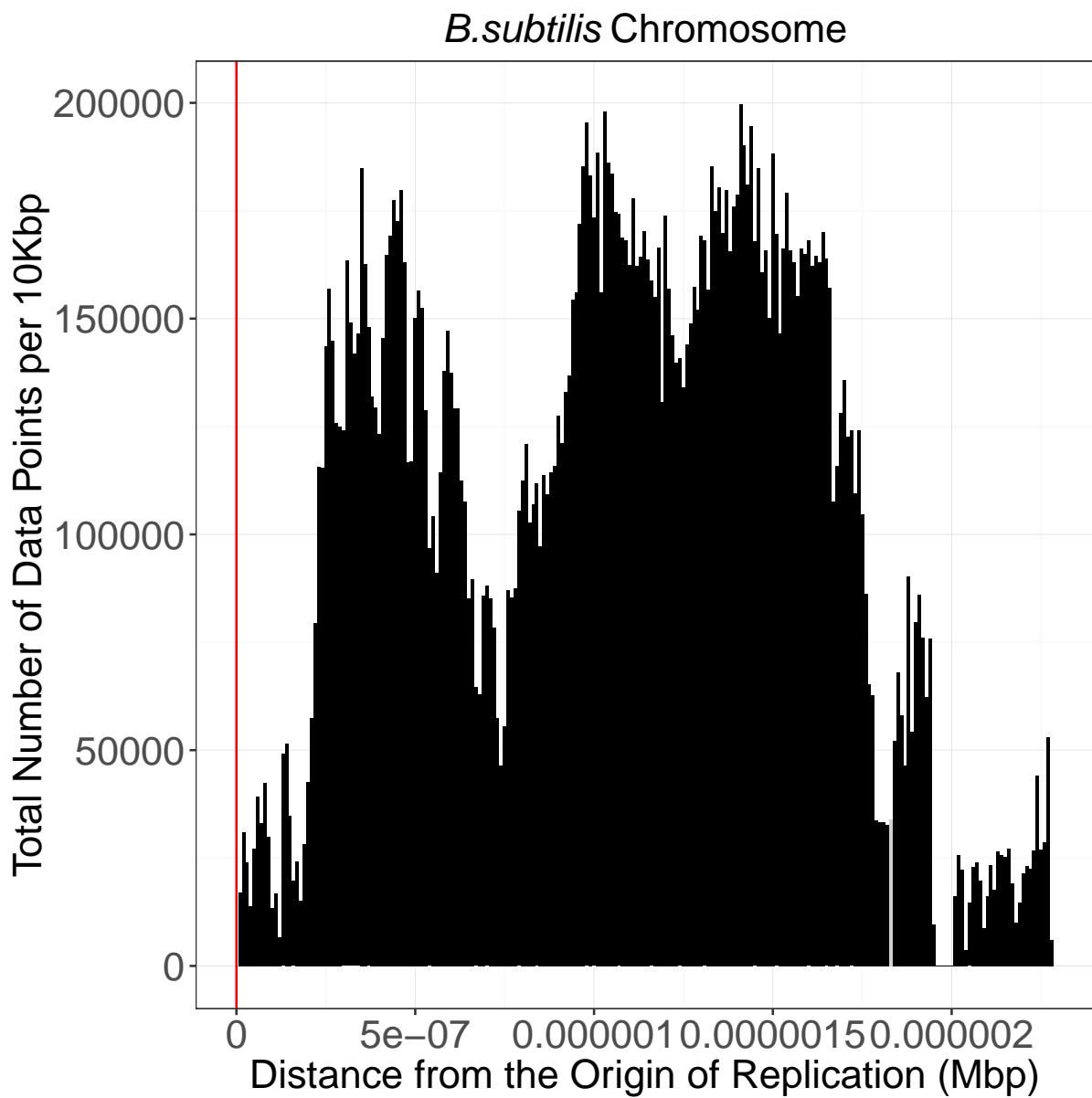


Figure 10: Distribution of total number of substitution data points per 10Kbp in genome.

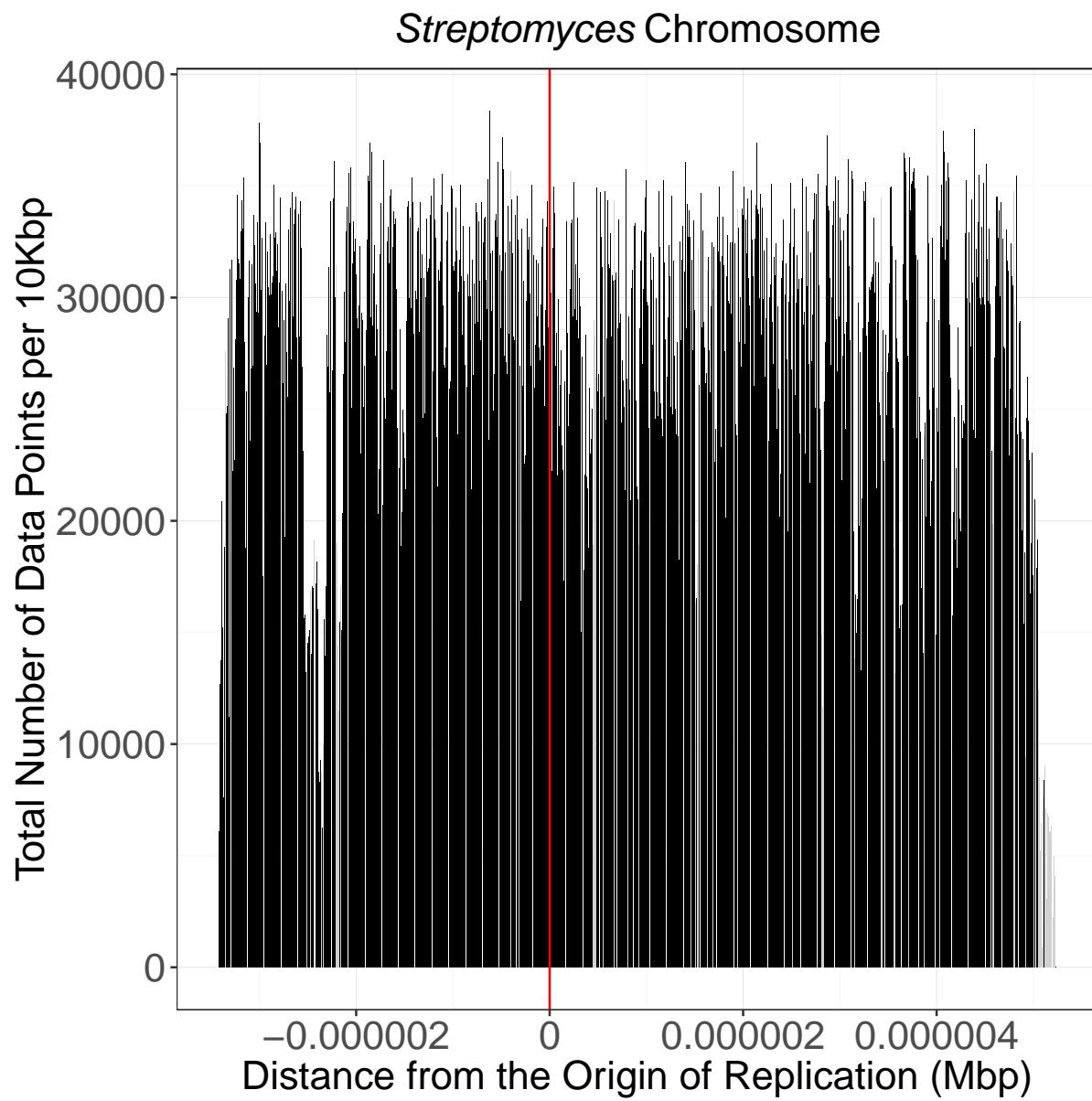


Figure 11: Distribution of total number of substitution data points per 10Kbp in genome.

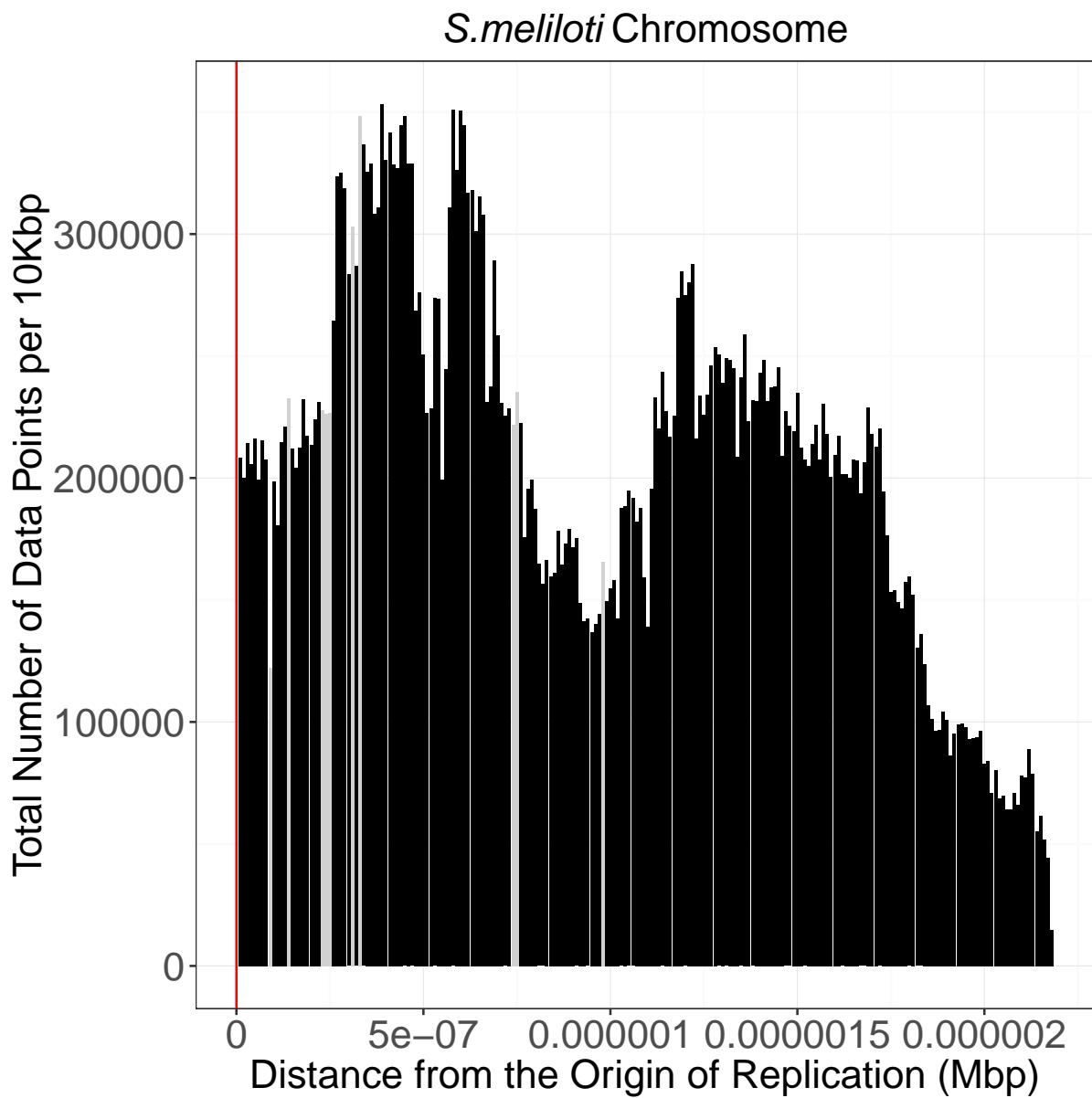


Figure 12: Distribution of total number of substitution data points per 10Kbp in genome.

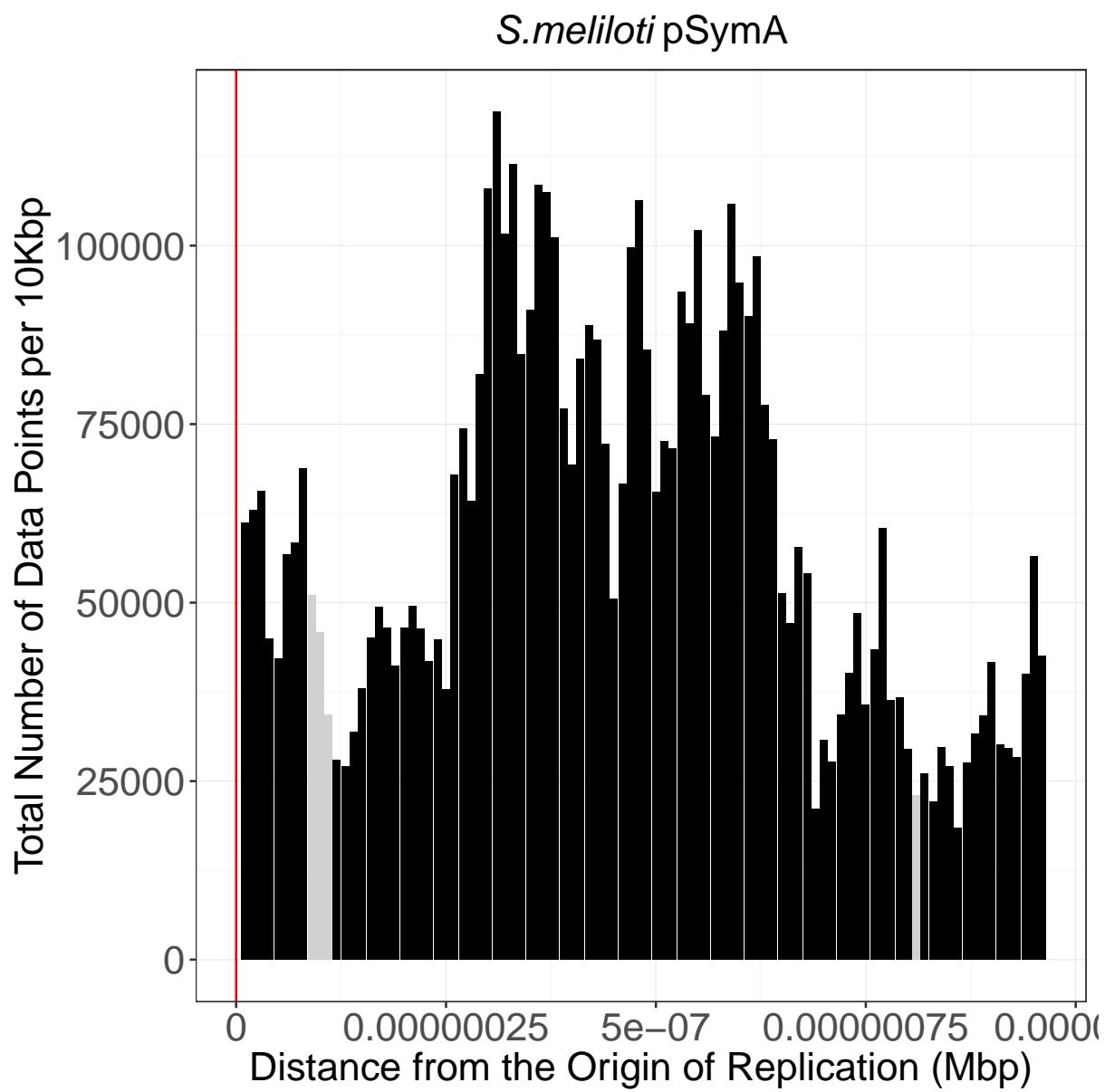


Figure 13: Distribution of total number of substitution data points per 10Kbp in genome.

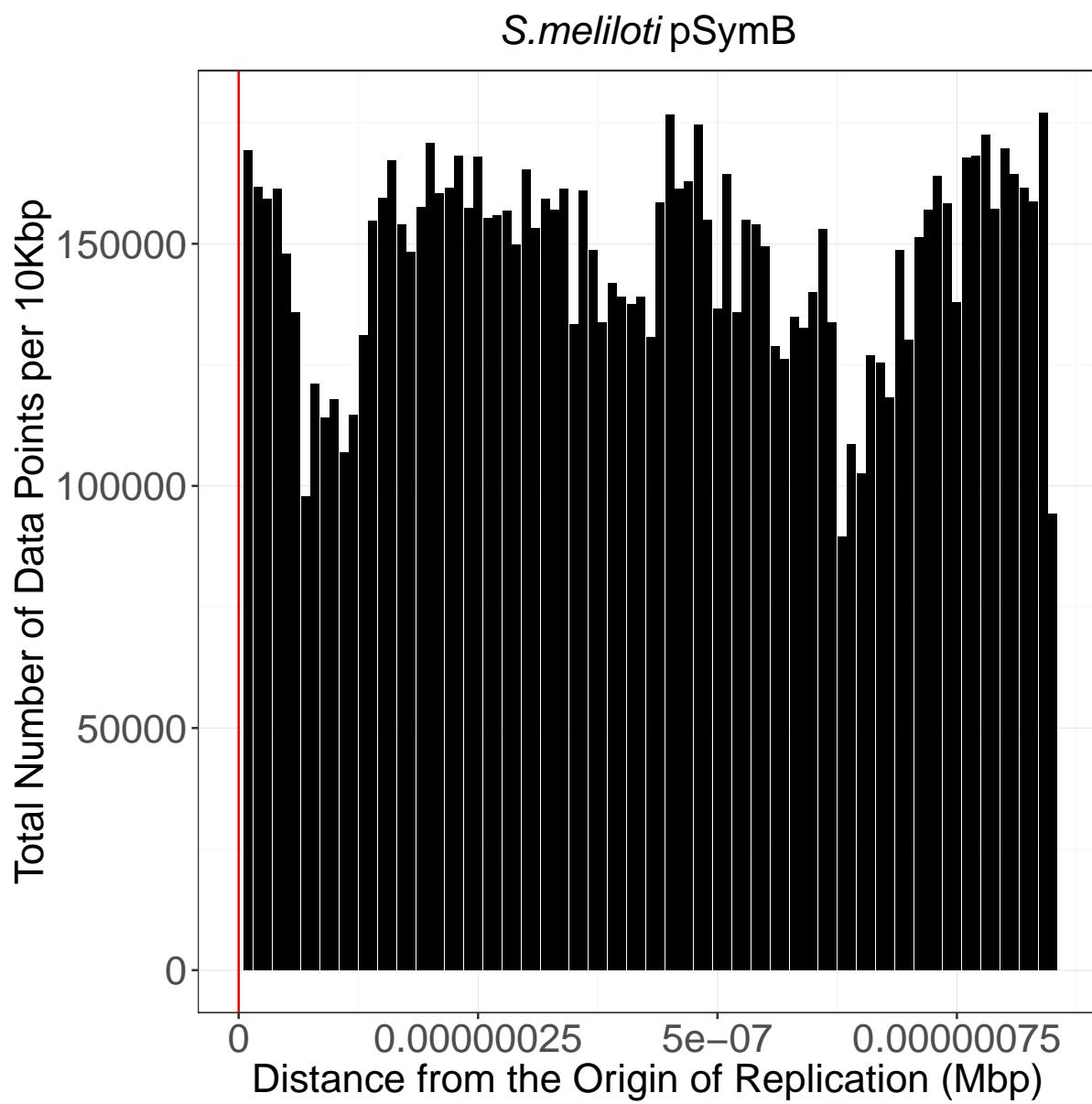


Figure 14: Distribution of total number of substitution data points per 10Kbp in genome.

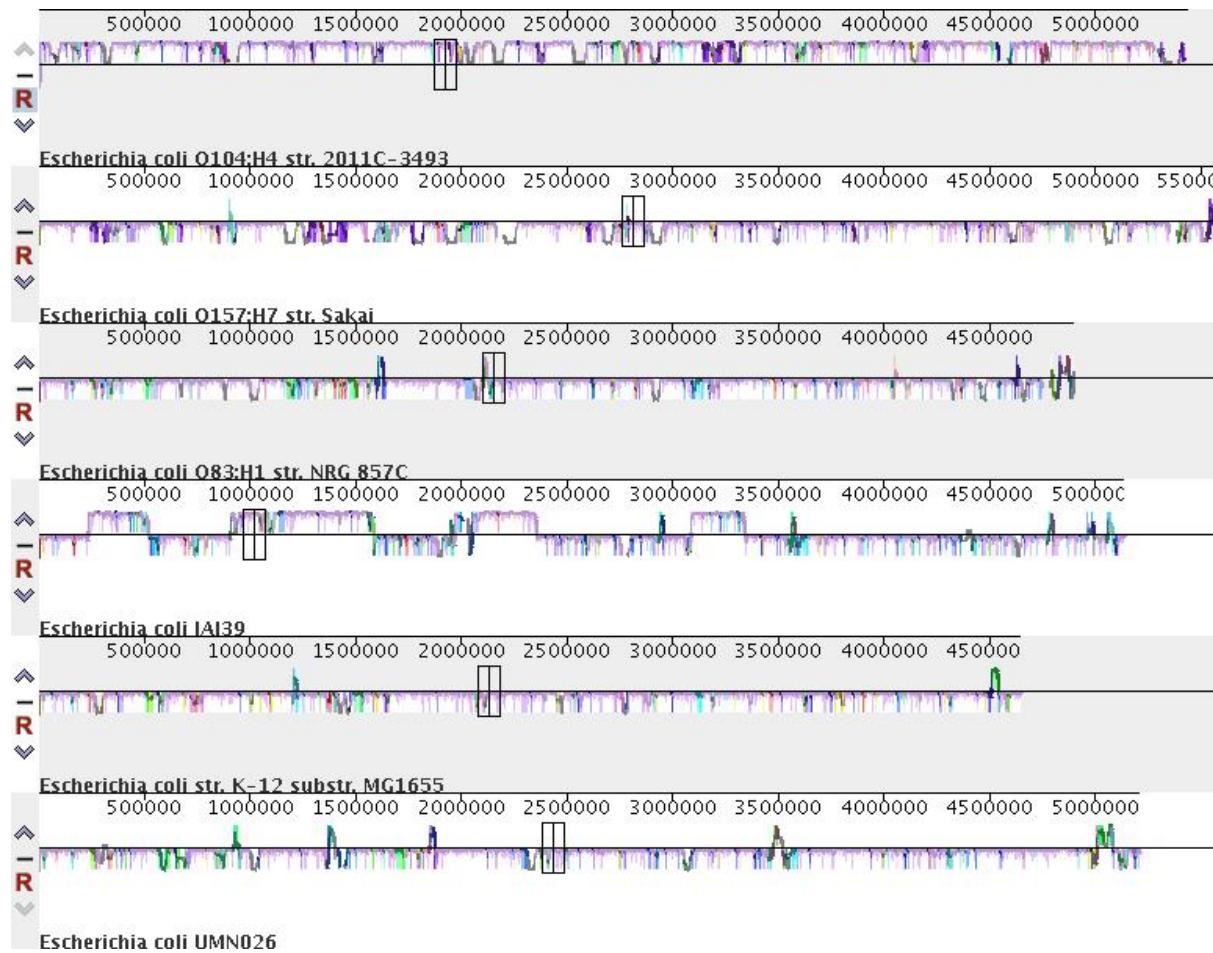


Figure 15: progressiveMauve alignment of *Escherichia coli* genomes highlighting the “backbone” of the alignment (matching regions).

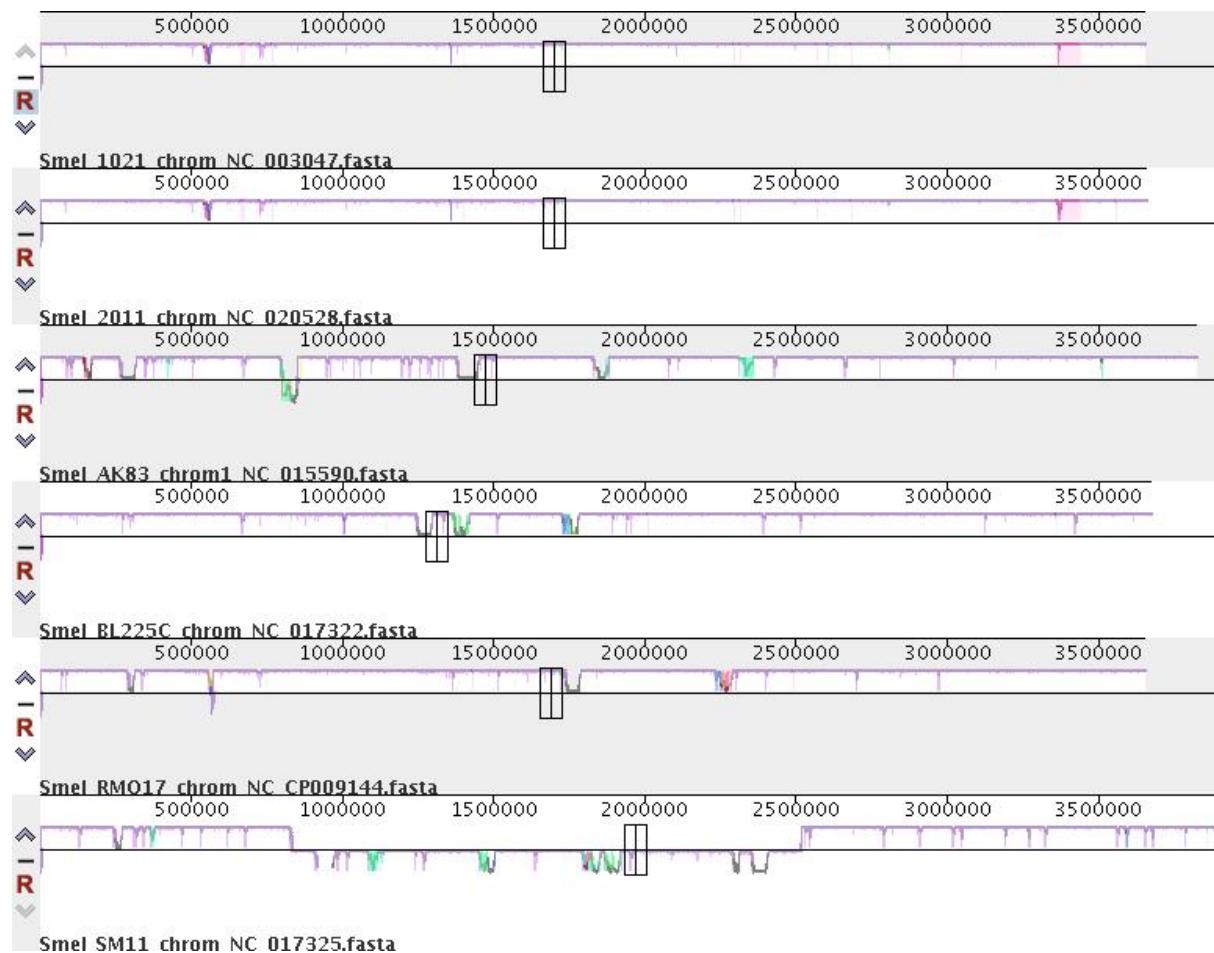


Figure 16: progressiveMauve alignment of *S. meliloti* Chromosomes highlighting the “backbone” of the alignment (matching regions).

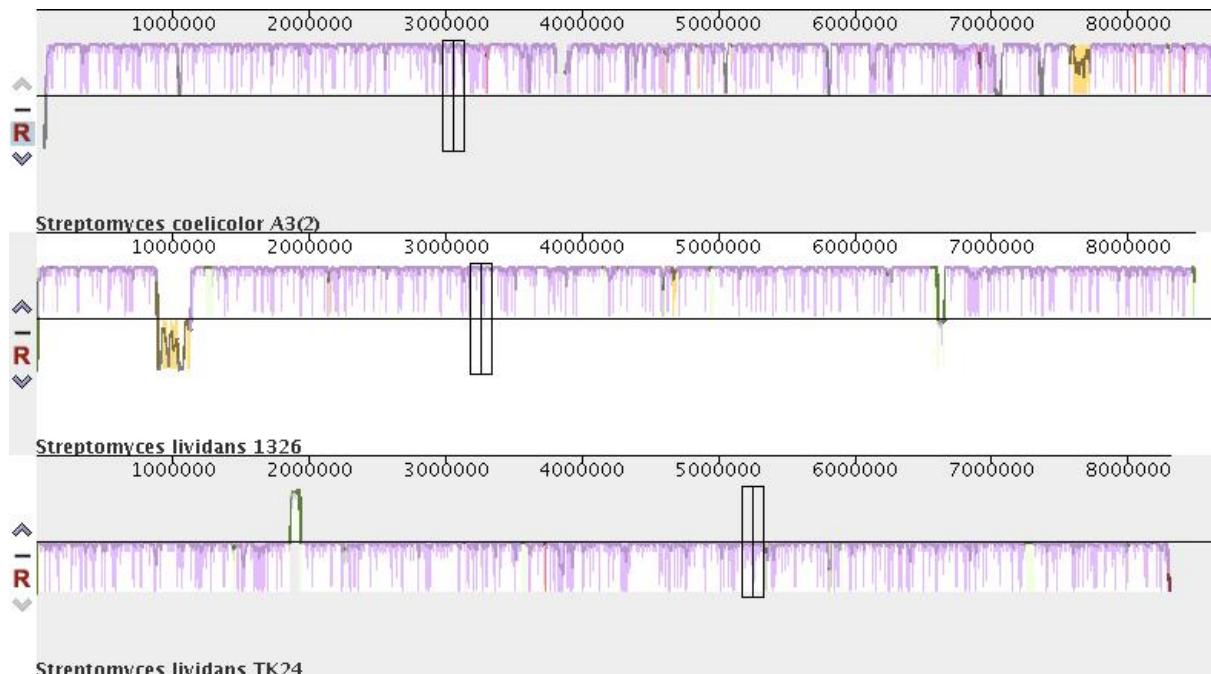


Figure 17: progressiveMauve alignment of *Streptomyces* genomes highlighting the “backbone” of the alignment (matching regions).