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:

✓finished Brian's second round of edits on the whole paper

Inversions + Gene Expression:

✓get GitHub set up on the cluster

✓make Queenie instructions on how to use GitHub on her laptop (Rstudio and cluster)

✓updated methods with removed O7 strain (discussed months ago but I forgot to actually do it)

✓continued to look into blast

General:

✓complete formatting for gene expression paper in dissertation

✓got abbreviations to work in dissertation file!

Substitution Paper: I made some more minor edits to the substitutions paper based on your comments: minor edits, changed the title, (attempt at) shortened the discussion. **If you could look over this again (mostly the discussion and title) that would be great! And the cover letter.**

Inversions + Gene Expression: Queenie is slowly working away at the tasks I gave her: verifying that gene expression is consistent across datasets for each strain and combining the parsnp inversions information with the gene expression + genome position data frame. I have also tried to encourage her to use GitHub so that I can keep track of her code (right now it is just on her laptop so I have no idea what she is doing). So I have gotten that set up on the machines and have instructed her on how to make it work with Rstudio (her preferred method for using R) so that I can keep track of what she is doing! I also continued to look into blast for confirming inversions. The only thing I am struggling with is what parameters to choose for the blast. **Which parameters do you think would be the best for this analysis?**

Still not sure what to do with the BW25113 data (some is mapped to K-12, not BW25113) and if we should include it or not. There are a total of 6 inversions between K-12 MG1655 and BW25113 (out of a total of 715). They vary in length from 22bp-272bp (all but one are between 22bp-38bp). I suppose we do not NEED this strain since it is so similar to K-12. K-12 MG1655, BW25113 and K-12 DH10B are all very similar. Most inversions happen between these three and the ATCC strain. I think the original thought was that we wanted to include as many strains as possible, even if they were similar because it could potentially help later on with future analysis. If we do get rid of the BW25113 strain, then we are only dealing with 3 strains total for this analysis, which is not necessarily bad since the similarity between the K-12 strains makes them essentially the same, which would happen regardless of if BW25113 is included or not. **What do you think?**

This Week

- check on Queenie's progress and double check her normalization code
- Queenie should be done graphs to check that expression between samples is comparable
- Get Queenie started on combining information about Parsnp inversions and gene expression

data

- figure out what I need to blast for the reciprocal part of the blast
- continue with blast (extract info from blast results)
- edit cover letter for substitution paper

Next Week

- help queenie with anything else she might need
- continue to work on blast (reciprocal part of blast hit and extract info from blast)
- edit dissertation intro
- submit substitutions paper

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

Bacteria and Replicon	Protein Coding Sequences
E. coli Chromosome	-1.43×10^{-8}***
B. subtilis Chromosome	-5.55×10^{-8}***
Streptomyces Chromosome	7.49×10^{-8}***
S. meliloti Chromosome	-5.99×10^{-7}***
S. meliloti pSymA	-5.18×10^{-7}***
S. meliloti pSymB	1.67×10^{-7}***

Table 2: 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'.

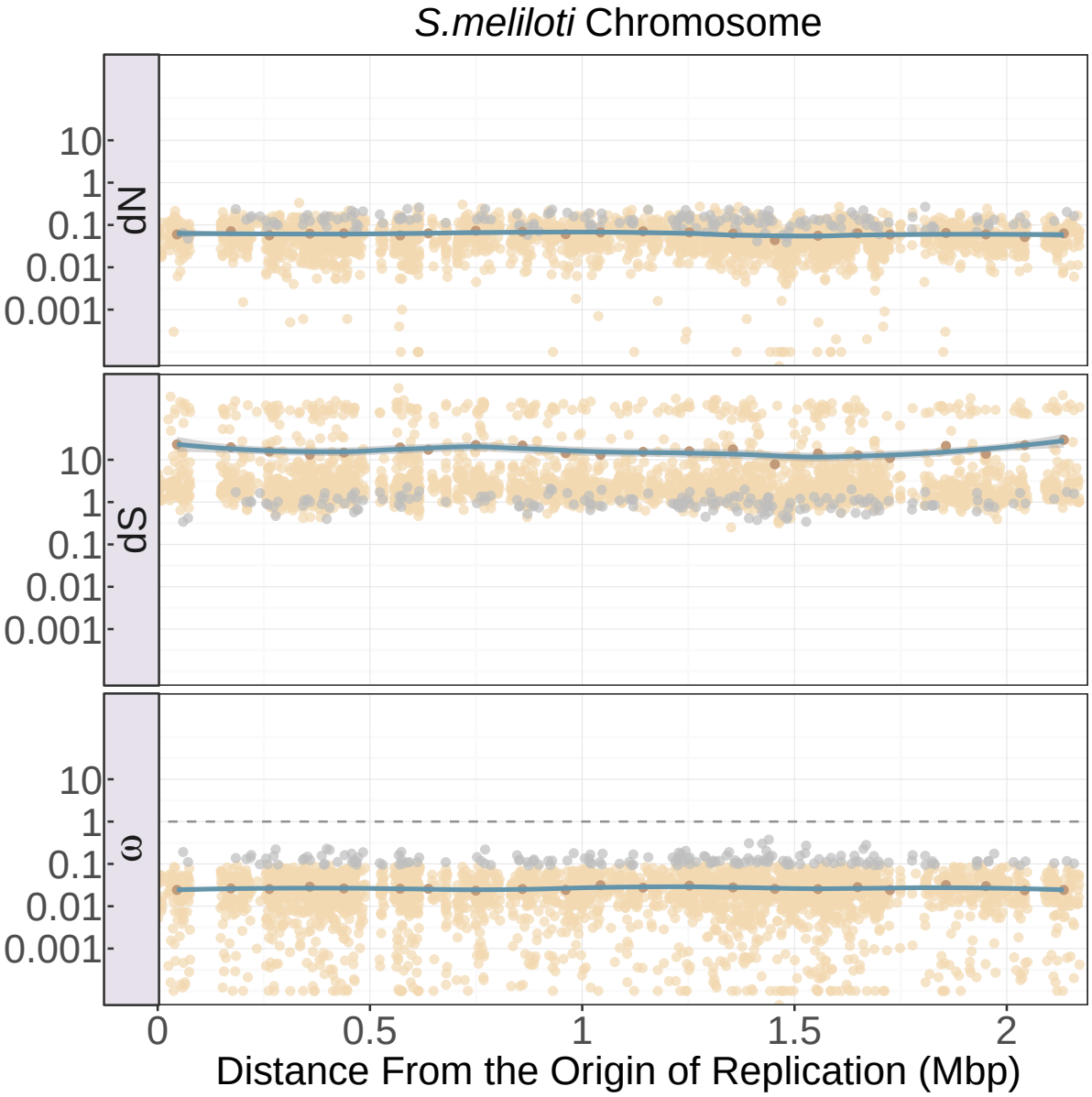
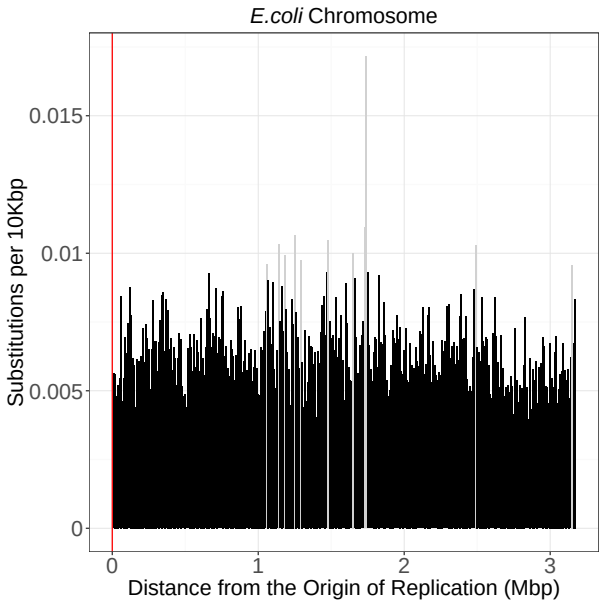
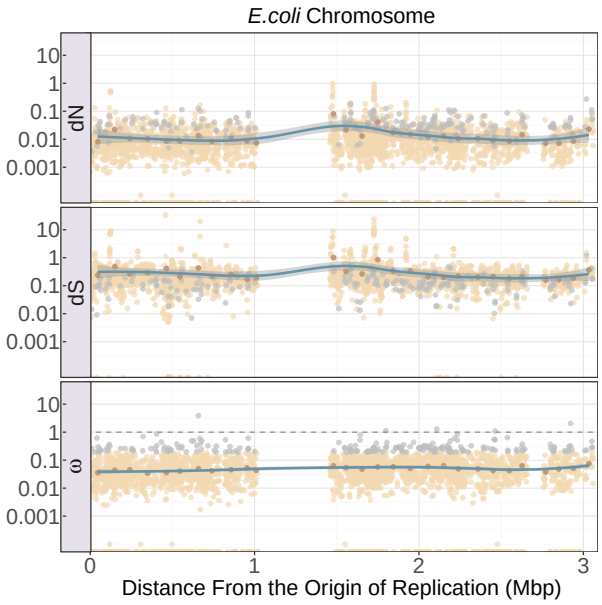


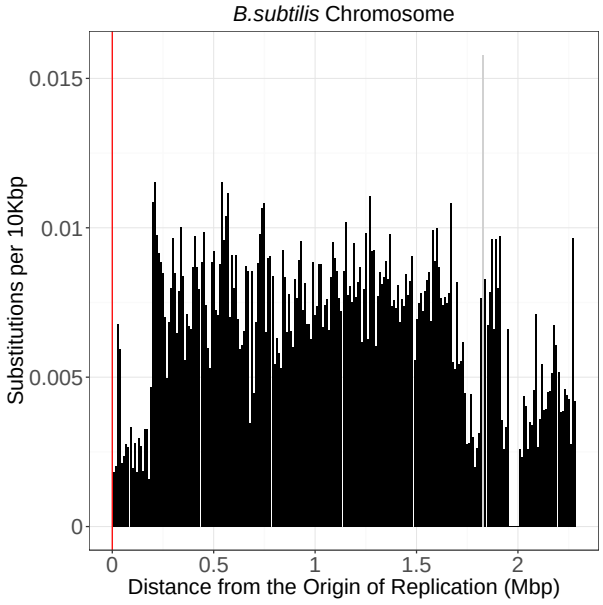
Figure 1: dN, dS, and ω values for *S. meliloti* chromosomes and ***A. tumefaciens***.


E.coli Chromosome
Substitutions per 10Kbp
0.015
0.01
0.005
0
0
1
2
3
Distance from the Origin of Replication (Mbp)

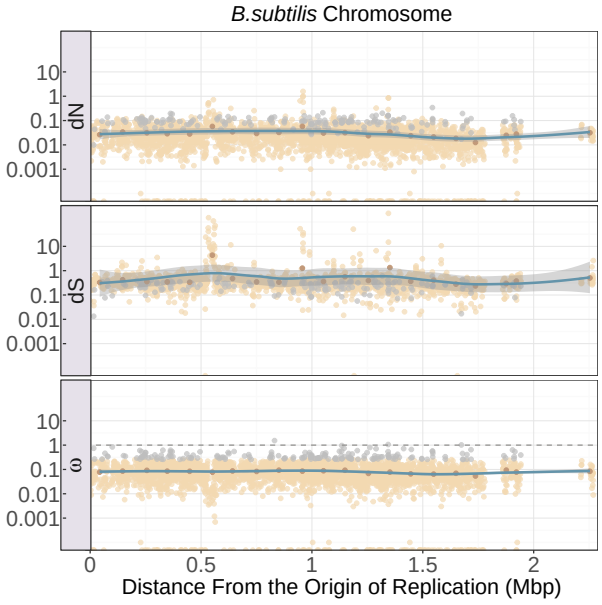
(a)


E.coli Chromosome
dN
dS
ω
10
1
0.1
0.01
0.001
0
1
2
3
Distance From the Origin of Replication (Mbp)

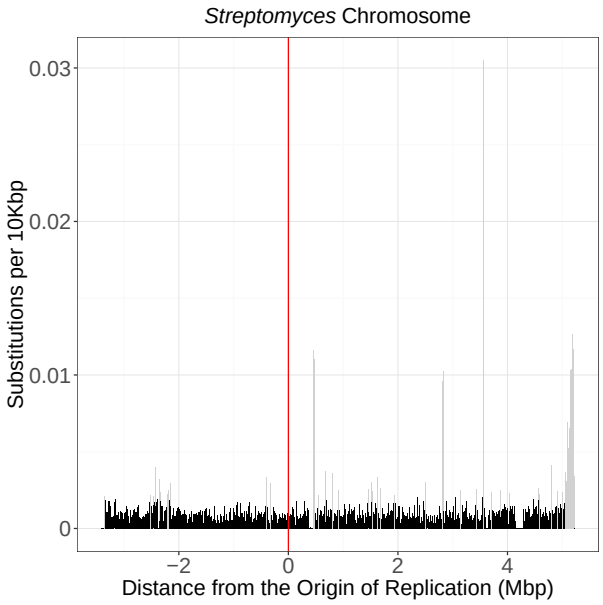
(b)


B.subtilis Chromosome
Substitutions per 10Kbp
0.015
0.01
0.005
0
0
0.5
1
1.5
2
Distance from the Origin of Replication (Mbp)

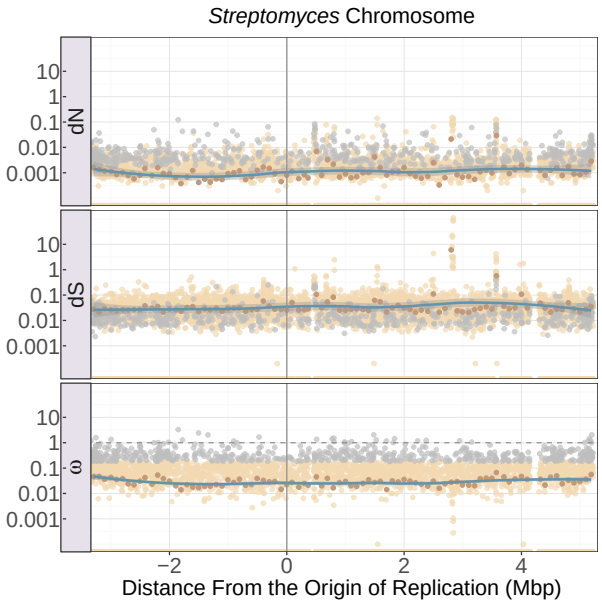
(a)


B.subtilis 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)

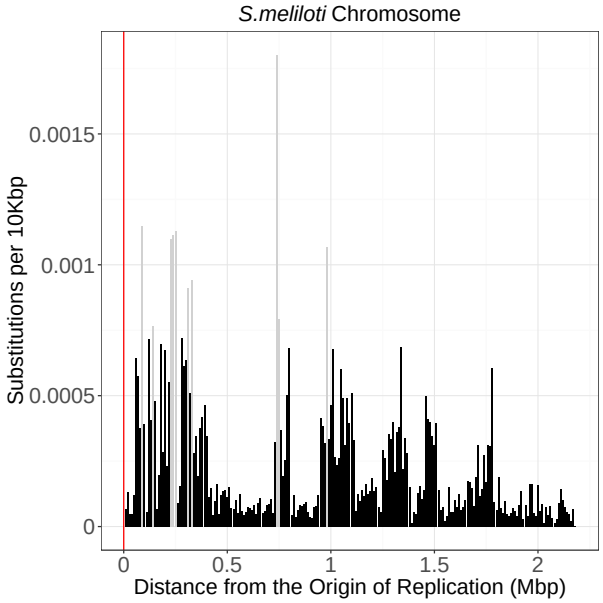
(b)


Streptomyces Chromosome
Substitutions per 10Kbp
0.03
0.02
0.01
0
−2
0
2
4
Distance from the Origin of Replication (Mbp)

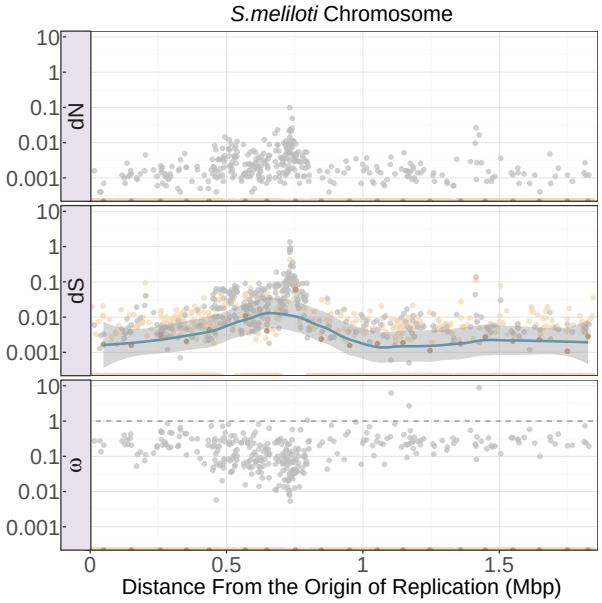
(a)


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

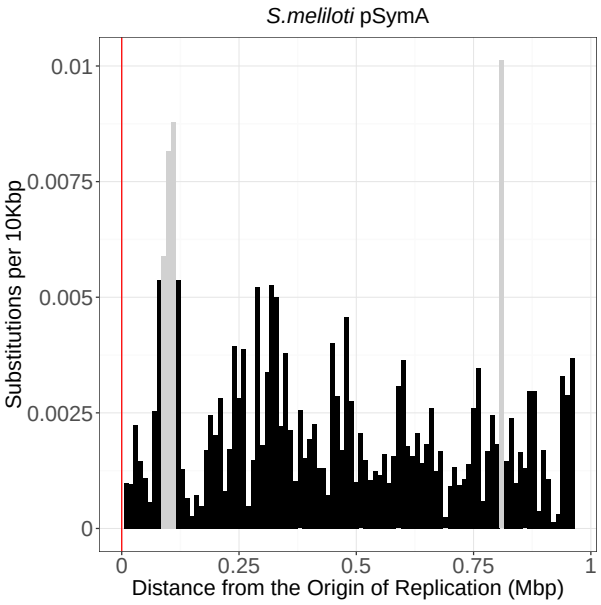
(b)


S.meliloti Chromosome
Substitutions per 10Kbp
0.0015
0.001
0.0005
0
0
0.5
1
1.5
2
Distance from the Origin of Replication (Mbp)

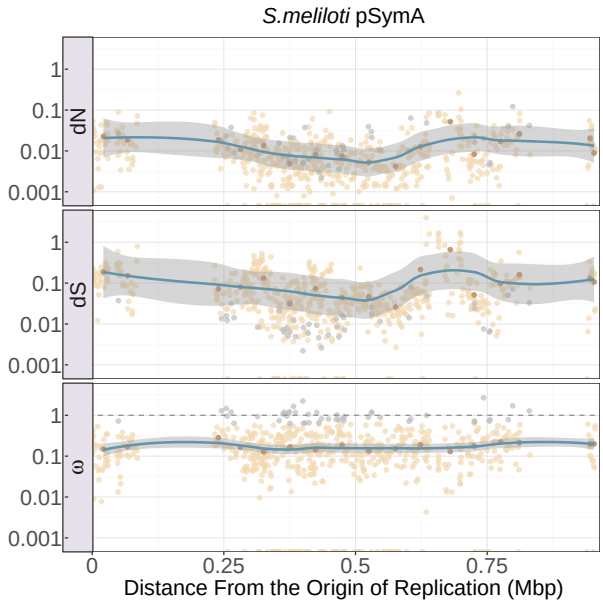
(a)


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

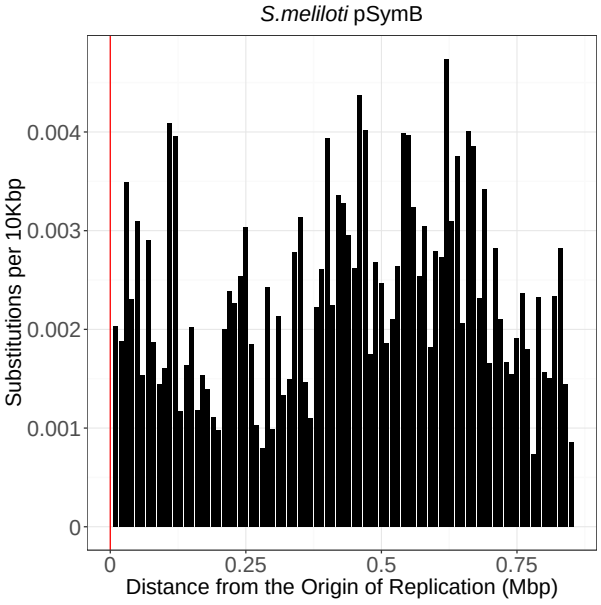
(b)


S.meliloti pSymA
Substitutions per 10Kbp
0.01
0.0075
0.005
0.0025
0
0
0.25
0.5
0.75
1
Distance from the Origin of Replication (Mbp)

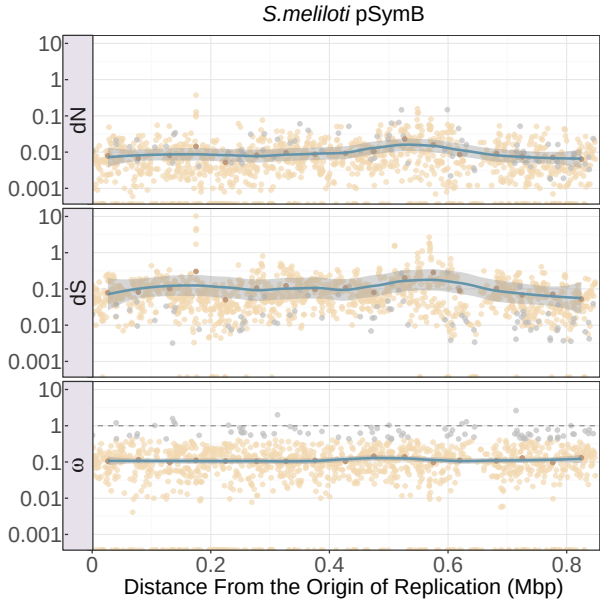
(a)


S.meliloti pSymA
dN
dS
ω
1
0.1
0.01
0.001
0
0.25
0.5
0.75
Distance From the Origin of Replication (Mbp)

(b)


S.meliloti pSymB
Substitutions per 10Kbp
0.004
0.003
0.002
0.001
0
0
0.25
0.5
0.75
Distance from the Origin of Replication (Mbp)

(a)


S.meliloti pSymB
dN
dS
ω
10
1
0.1
0.01
0.001
0
0.2
0.4
0.6
0.8
Distance From the Origin of Replication (Mbp)

(b)

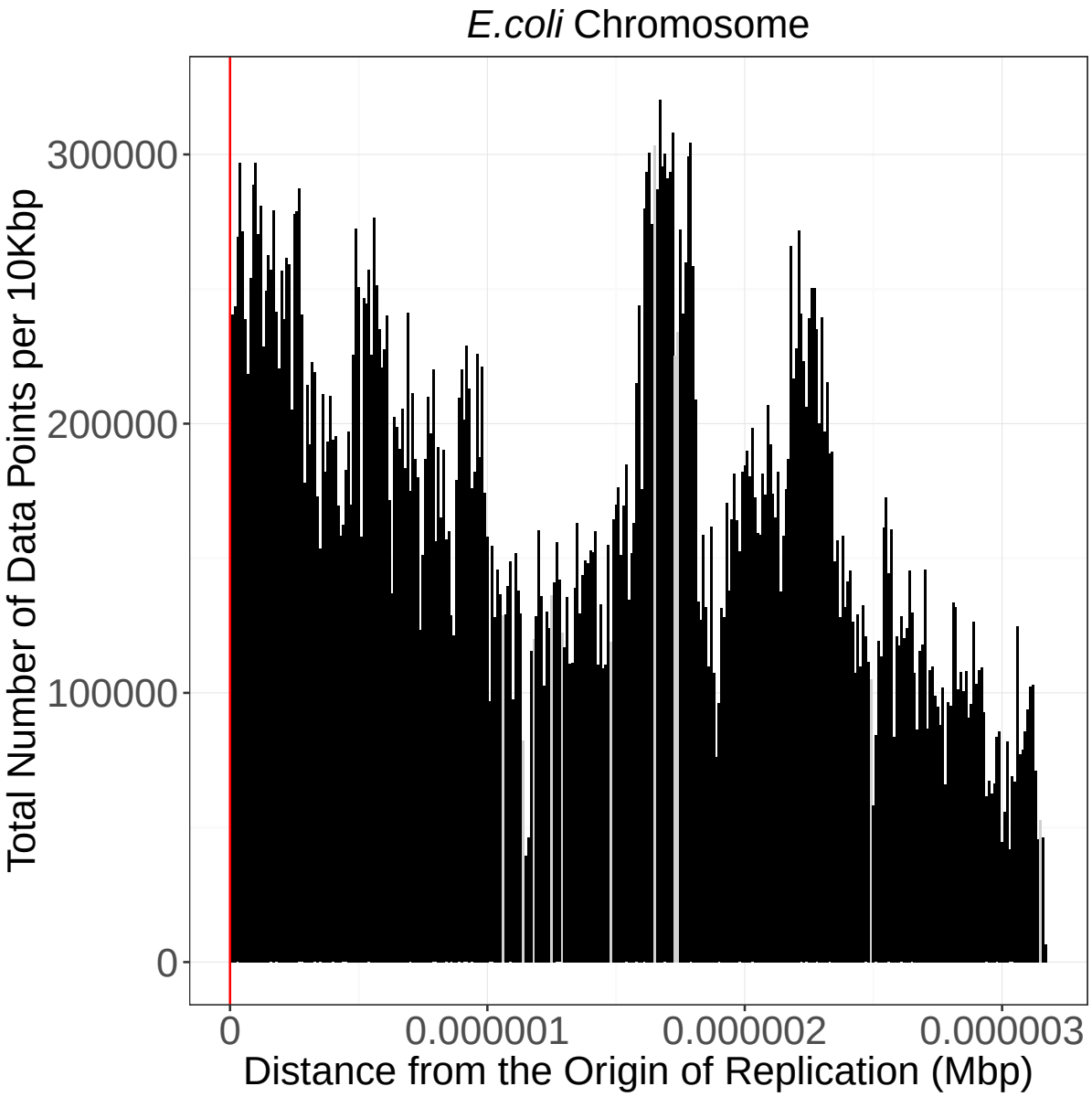


Figure 8: Distribution of total number of substitution data points per 10Kbp in genome.

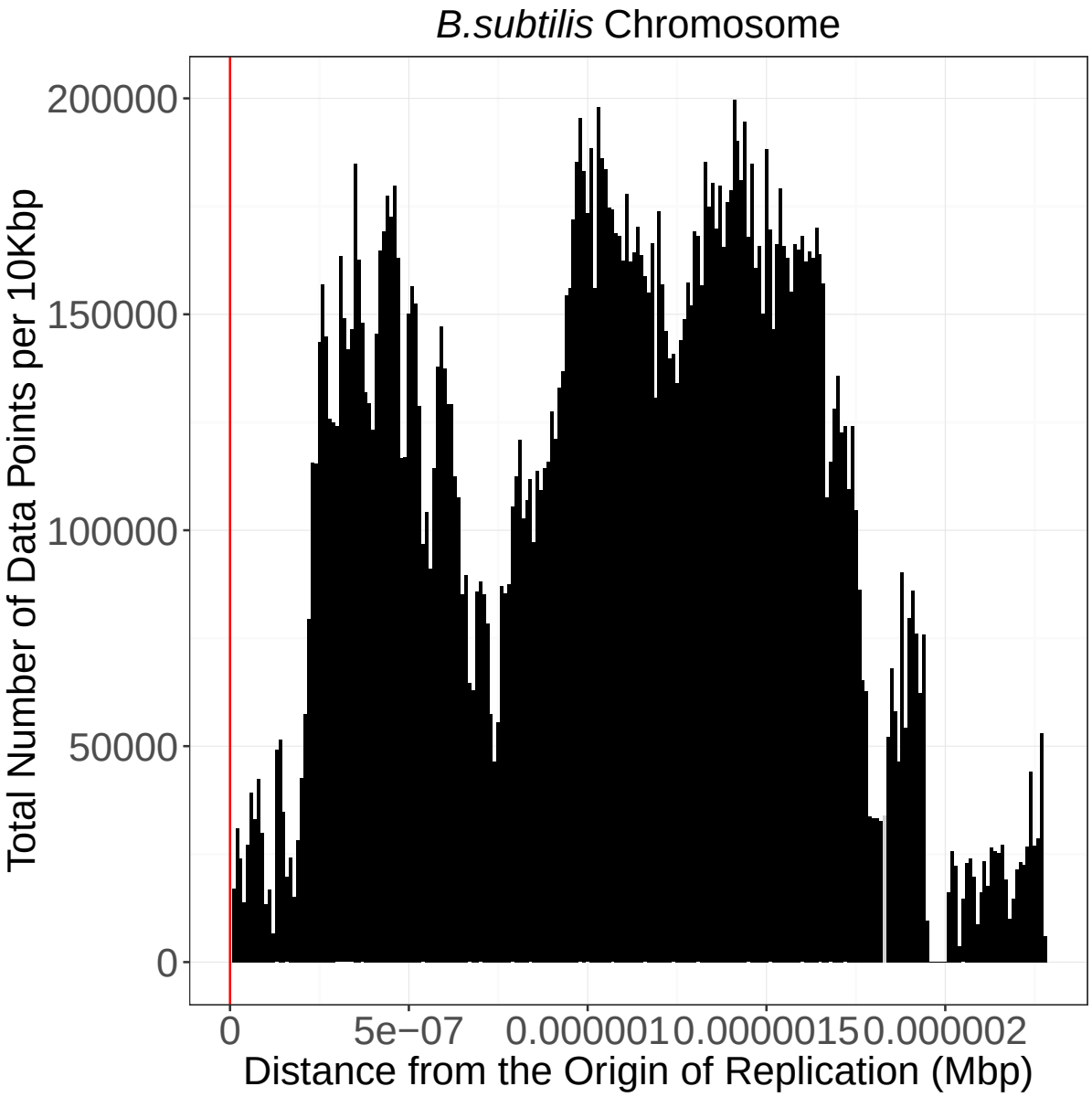


Figure 9: Distribution of total number of substitution data points per 10Kbp in genome.

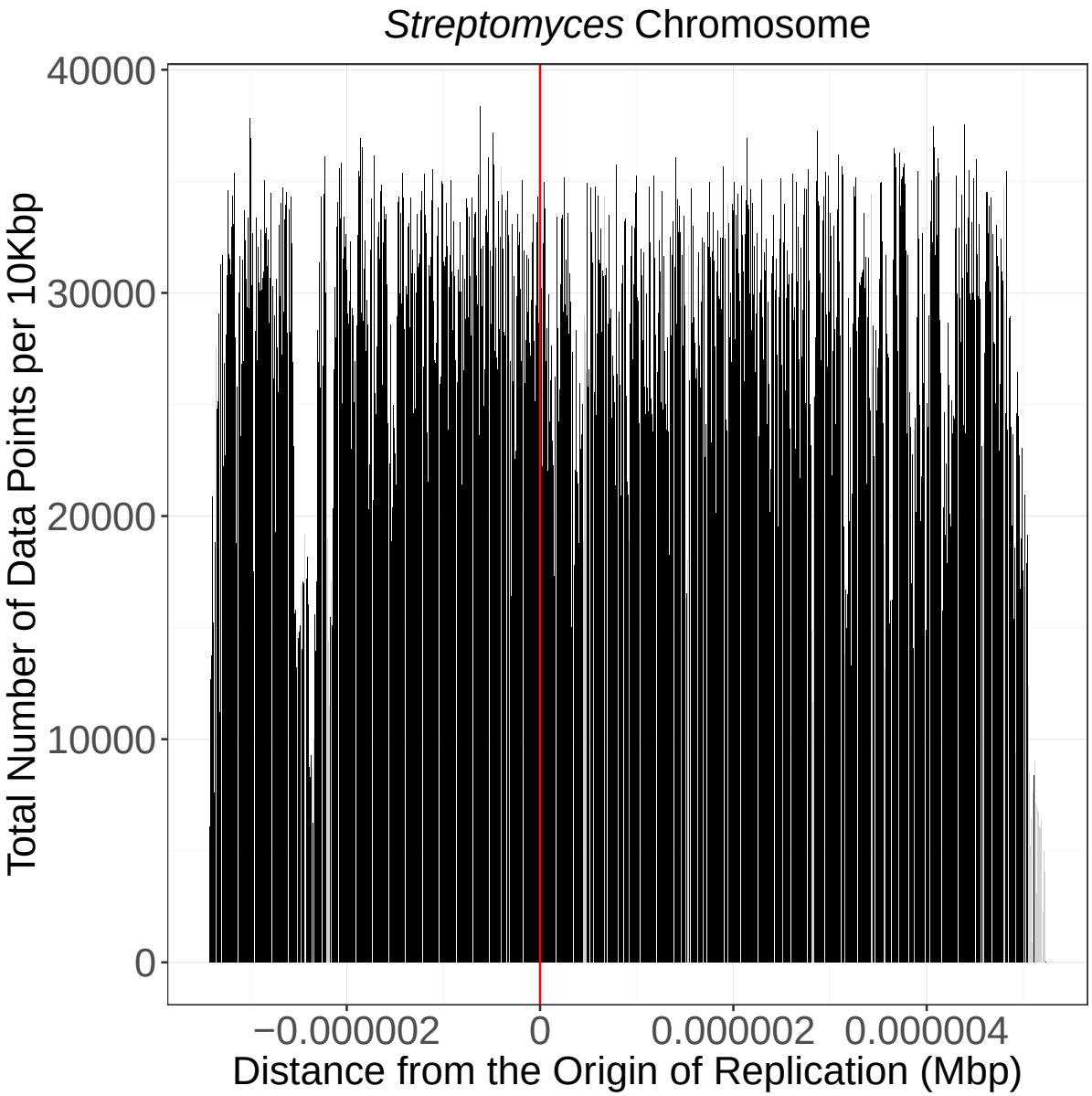


Figure 10: Distribution of total number of substitution data points per 10Kbp in genome.

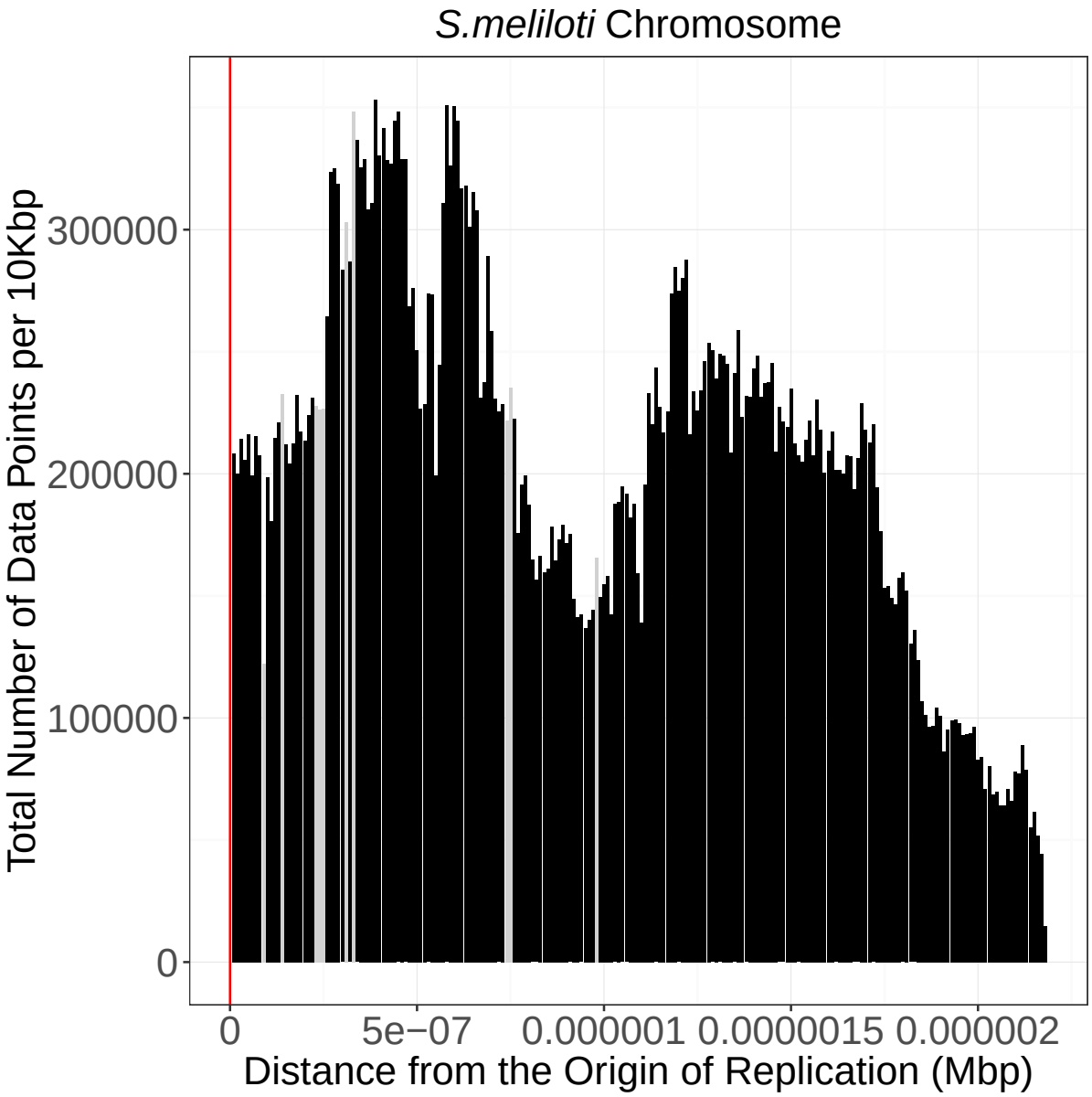


Figure 11: Distribution of total number of substitution data points per 10Kbp in genome.

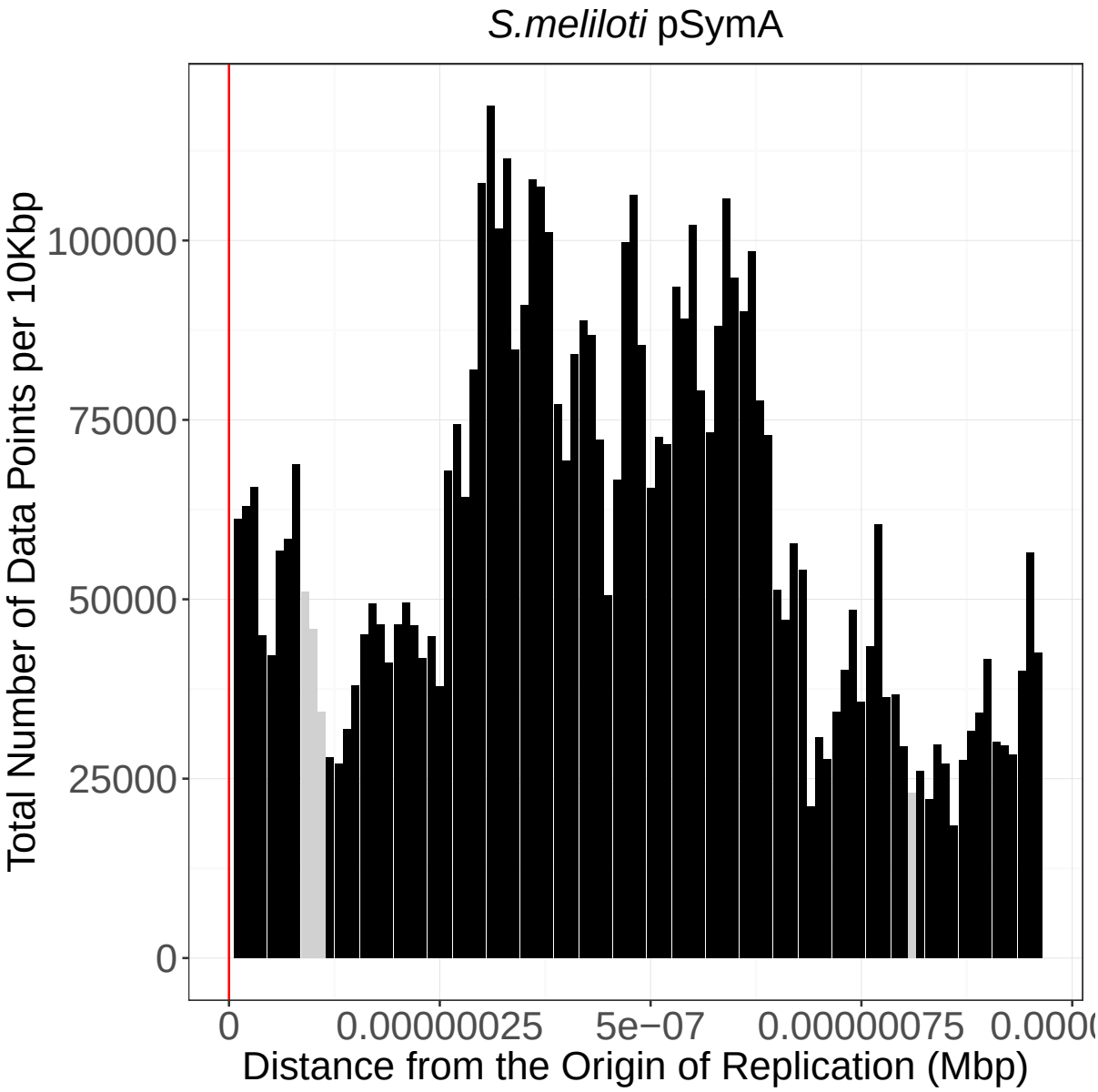


Figure 12: Distribution of total number of substitution data points per 10Kbp in genome.

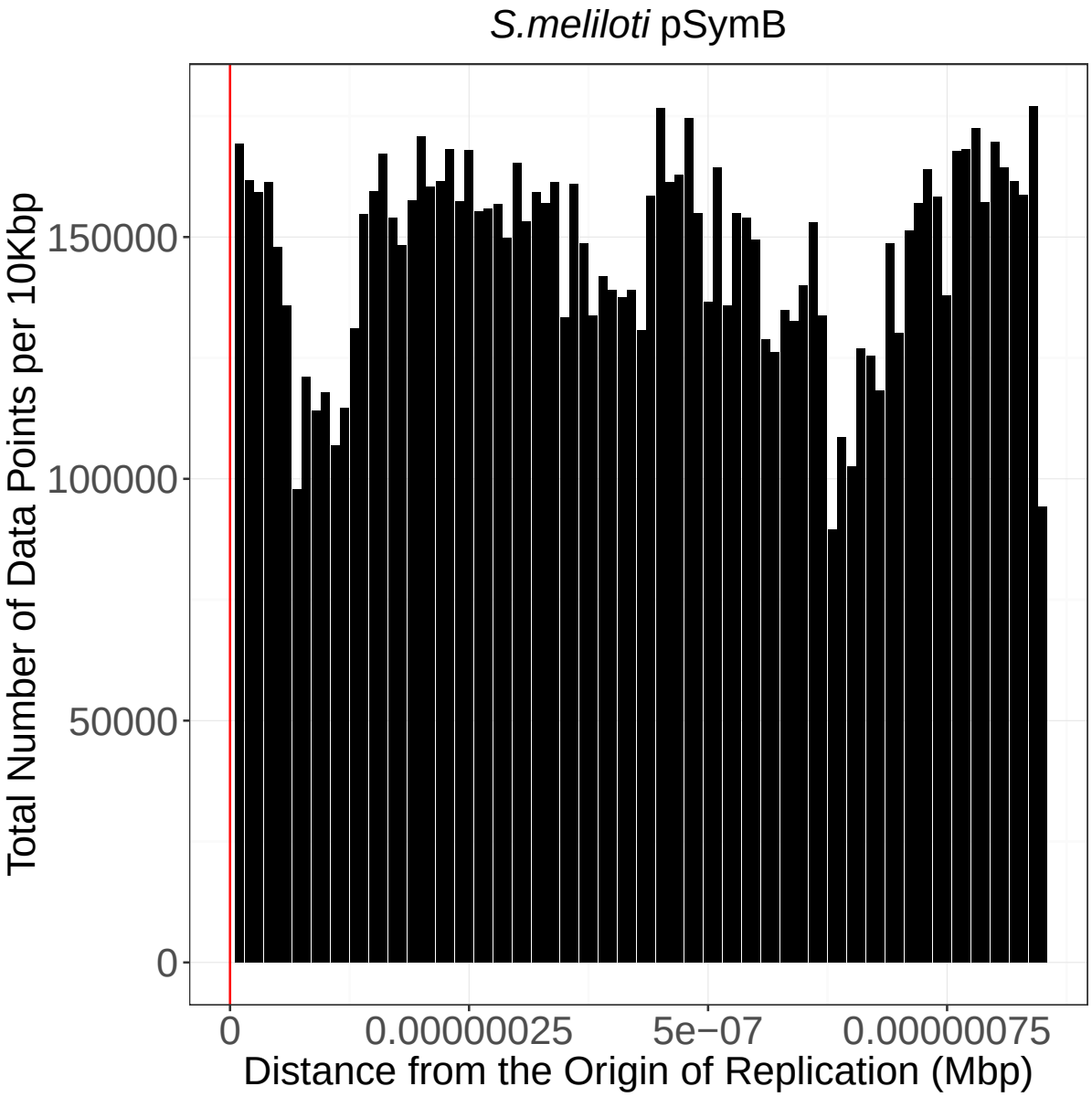


Figure 13: Distribution of total number of substitution data points per 10Kbp in genome.

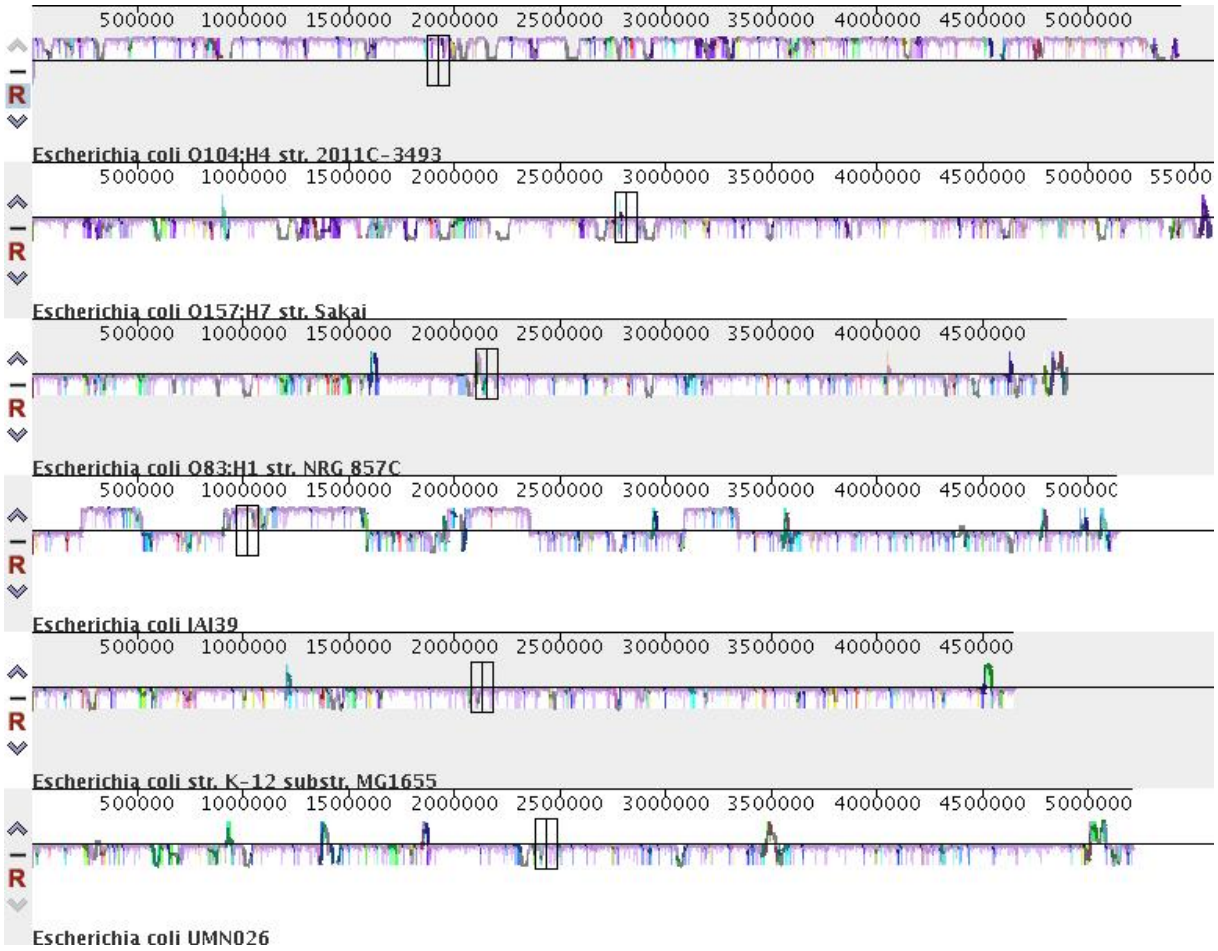


Figure 14: progressiveMauve alignment of *Escherichia coli* genomes highlighting the "backbone" of the alignment (matching regions).

Figure 15: progressiveMauve alignment of *S. meliloti* Chromosomes highlighting the "backbone" of the alignment (matching regions).

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

Table 3: Average number of protein coding substitutions calculated per base across all bacterial replicons. Outliers and missing data was not included in the calculation.

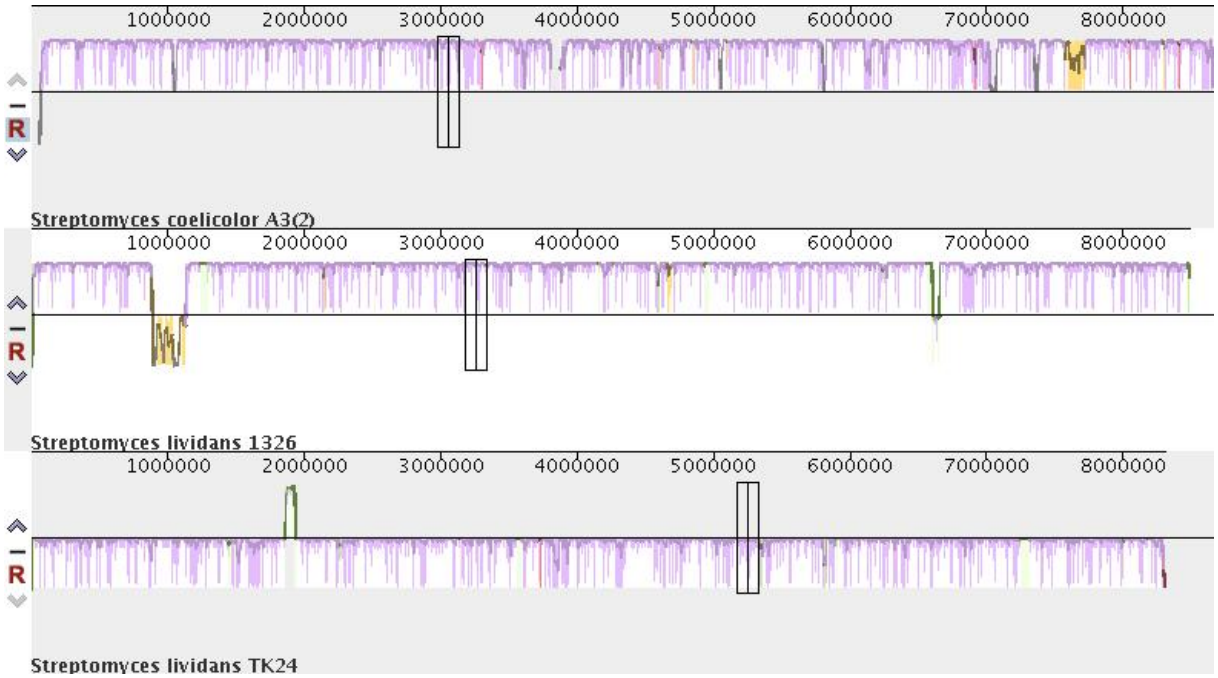


Figure 16: progressiveMauve alignment of *Streptomyces* genomes highlighting the "backbone" of the alignment (matching regions).