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 third round of edits on the whole paper

✓finished edits to cover letter

✓test for how big the substitutions slope can get

Inversions + Gene Expression:

✓get GitHub set up on the cluster

✓get Queenie set up on GitHub on cluster

✓started combining gene expression and PARSNP info

✓continuing to run BLAST with various parameters

✓decided to keep BW1255 strain, but use K-12 gene annotation

Substitution Paper: I made all the necessary edits to the paper and cover letter (attached in the email). If you are good with it, I think it is ready for submission!

I also did the test for looking at the substitutions slope and how large it can be. I did this by taking the largest non-outlier bar (weighted total substitutions/10Kbp) and set the position to 0, then making another fake point with position at the terminus and a substitutions value of 0, then computing a regression. The results can be found in Table 1 and the actual slope values can be found in Table 2.

Bacteria and Replicon	Test Regression slope
E. coli Chromosome	-2.94×10^{-9}
B. subtilis Chromosome	-5.08×10^{-9}
Streptomyces Chromosome	-3.92×10^{-10}
S. meliloti Chromosome	-3.32×10^{-10}
S. meliloti pSymA	-5.66×10^{-9}
S. meliloti pSymB	-5.65×10^{-9}

Table 1: Values of regression slope for each replicon using two points: 1) Highest weighted value of the number of substitutions / 10Kbp at position zero and 2) weighted value of the number of substitutions / 10Kbp of zero at the terminus. Simple linear regression was calculated. All results have no residuals (no residual degrees of freedom) because there are only two points on the line.

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 been working on the blast portion of this project and running blast with multiple parameters to see what we get.

For the BW1255 strain, since one of the data sets is mapped to the K-12 genome and the other is unclear of where it is mapped, I think that using the K-12 genome annotation for this strain is best. Additionally, since the proteomes are redundant, I think we should have minimal issues using the K-12 annotation. I have asked Queenie to see how often the genome positions for the same gene (in the gbk file) differ between BW and K-12, so we can get a sense/justify if this is the correct decision.

Bacteria and Replicon	Protein Coding Sequences
E. coli Chromosome	-2.19×10^{-8}***
B. subtilis Chromosome	-6.02×10^{-8}***
Streptomyces Chromosome	5.34×10^{-8}***
S. meliloti Chromosome	-3.80×10^{-7}***
S. meliloti pSymA	-3.15×10^{-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'.

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
- continue with blast (extract info from blast results)

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

Bacteria and Replicon	Genome Average 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 3: 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	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 4: Average number of protein coding substitutions calculated per base across all bacterial replicons. Outliers and missing data was not included in the calculation.

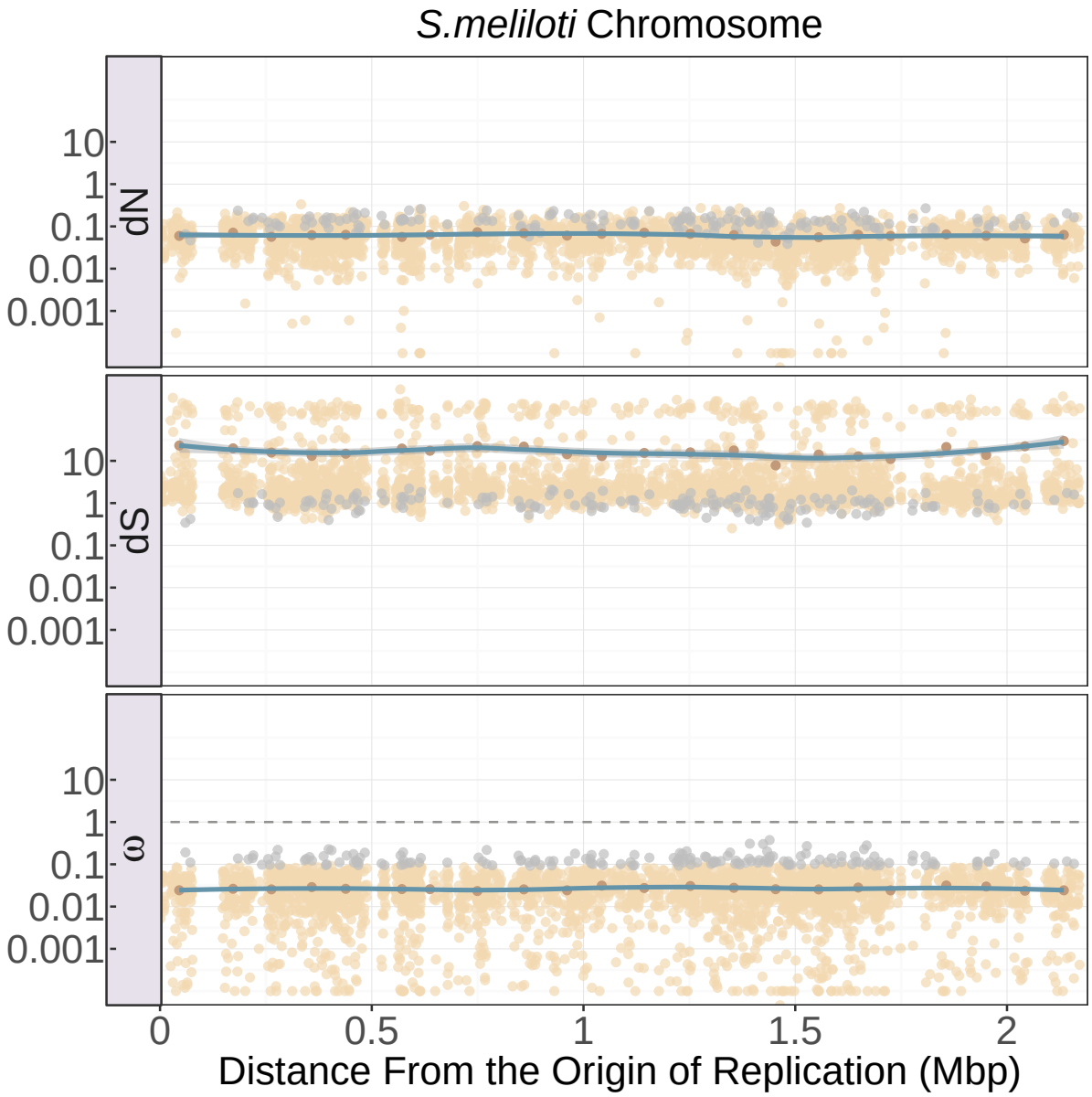
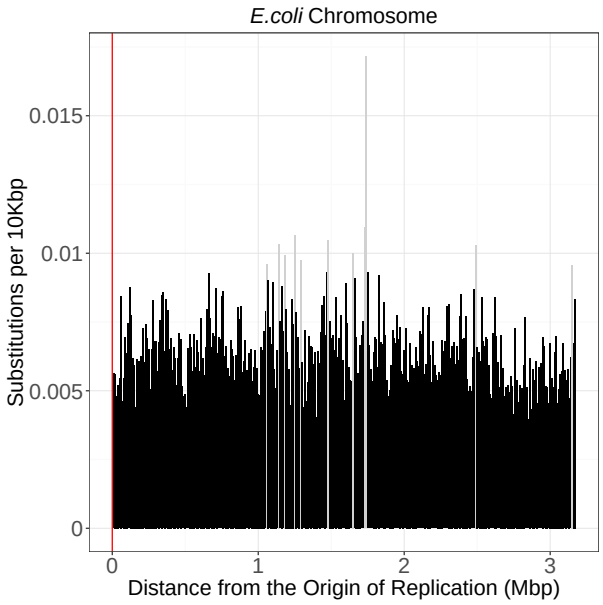
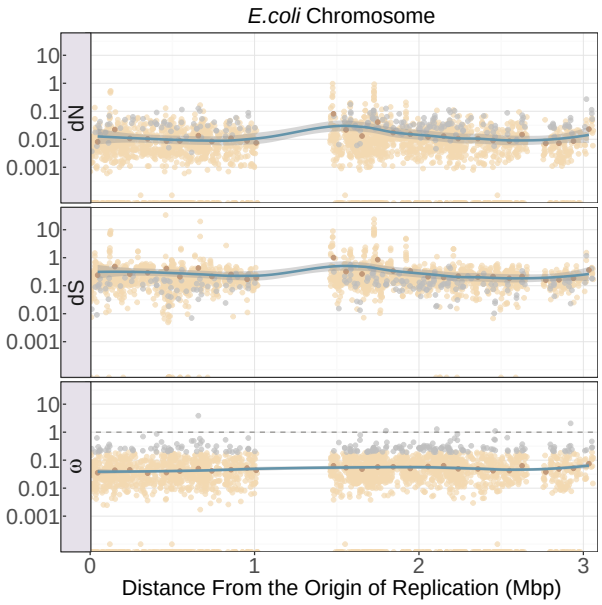


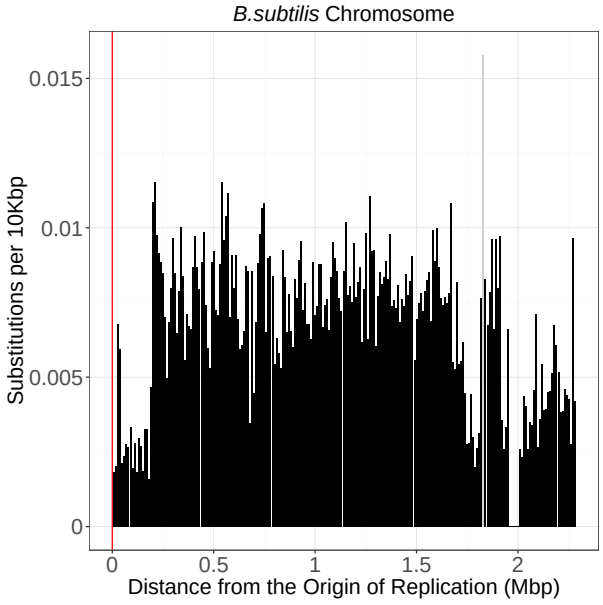
Figure 1: dN, dS, and ω values for *S. meliloti* chromosomes and ***A. tumefaciens***.


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

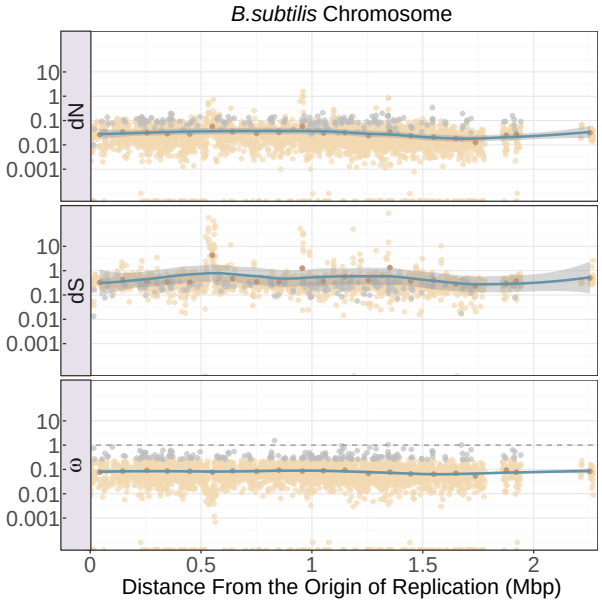
(a)


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)

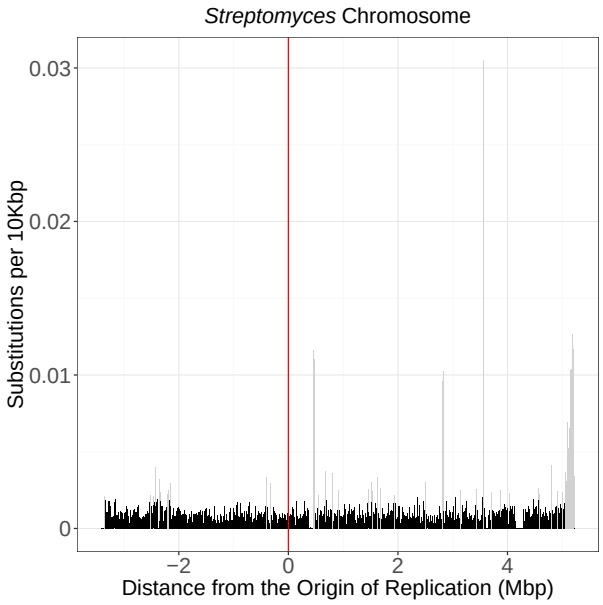
(b)


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)

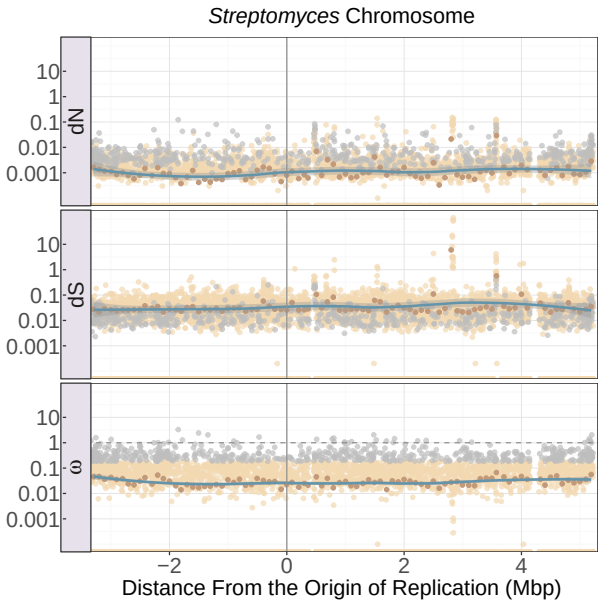
(a)


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)

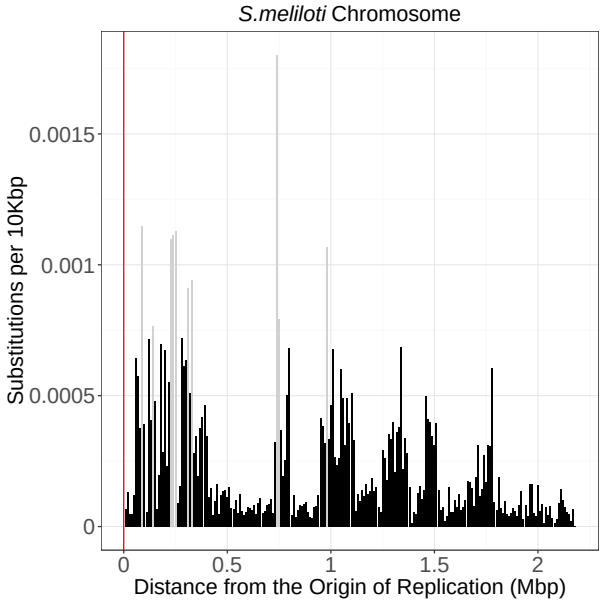
(b)


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

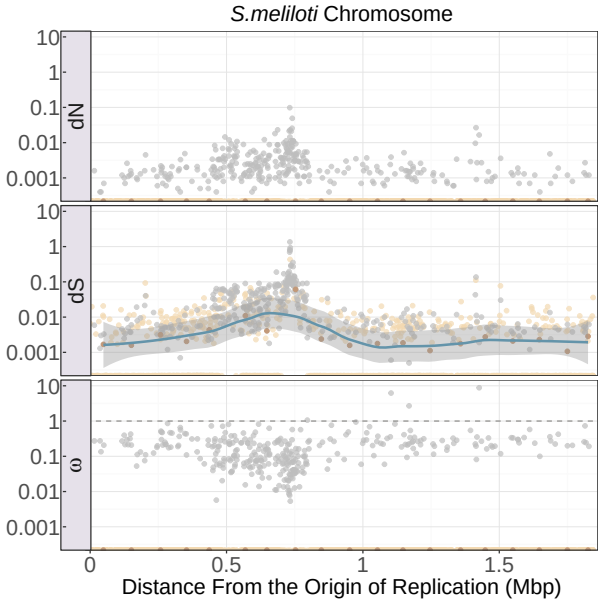
(a)


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

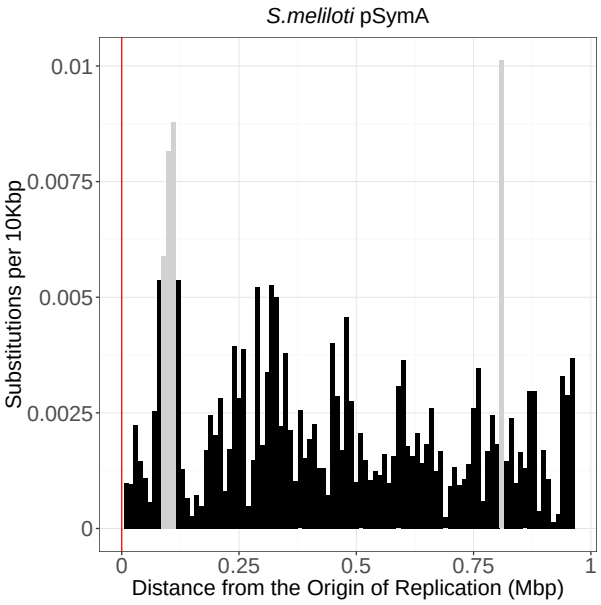
(b)


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)

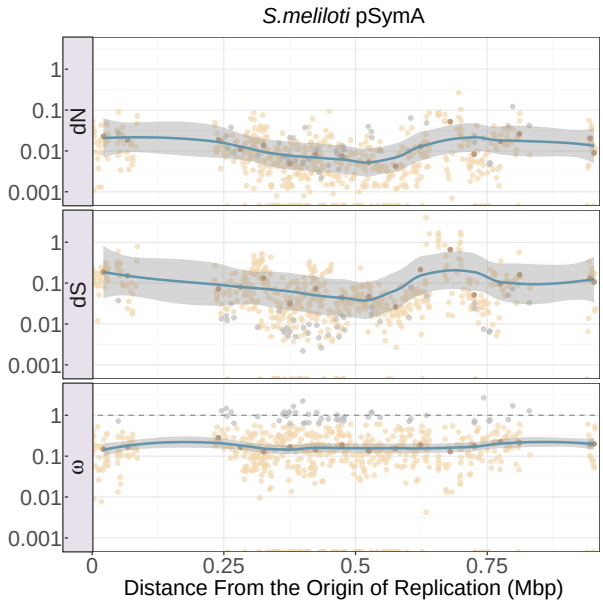
(a)


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)

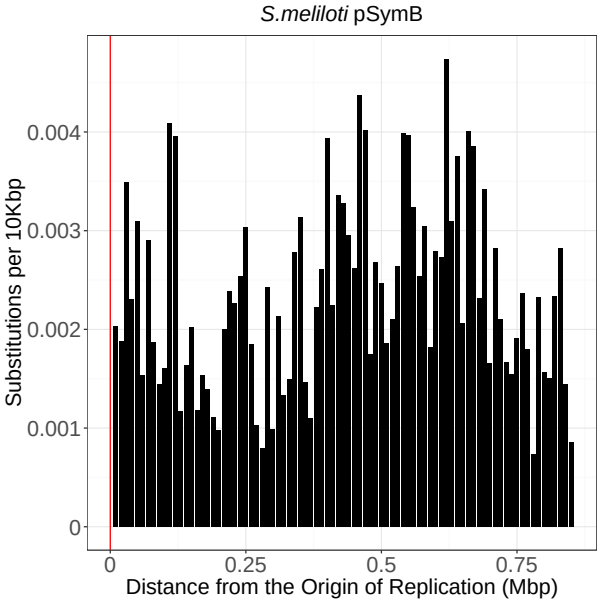
(b)


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)

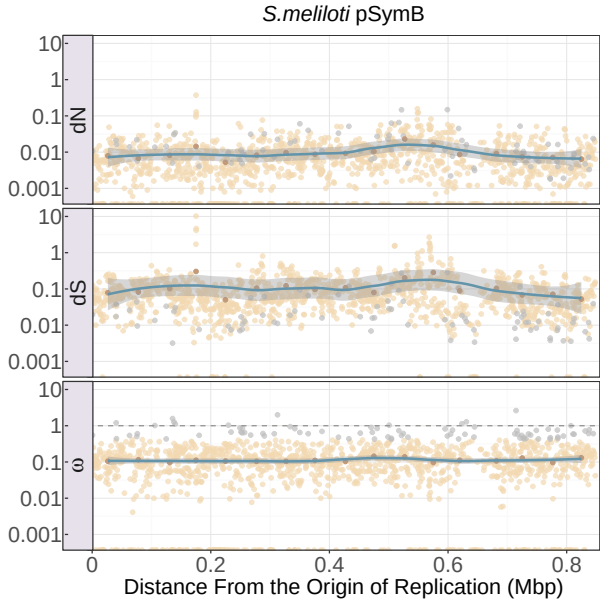
(a)


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)

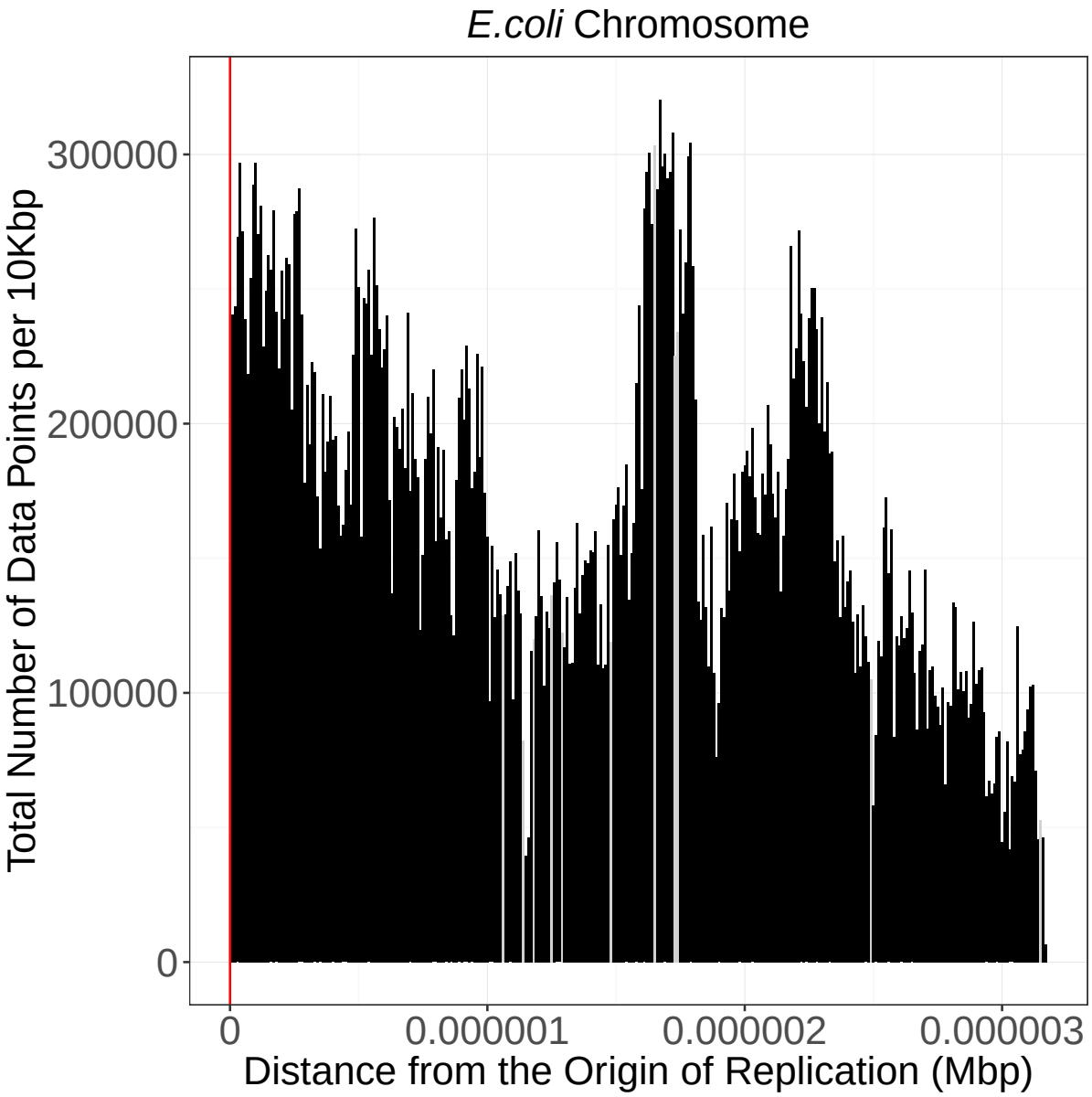


Figure 8: Distribution of total number of substitution data points per 10Kbp in genome.

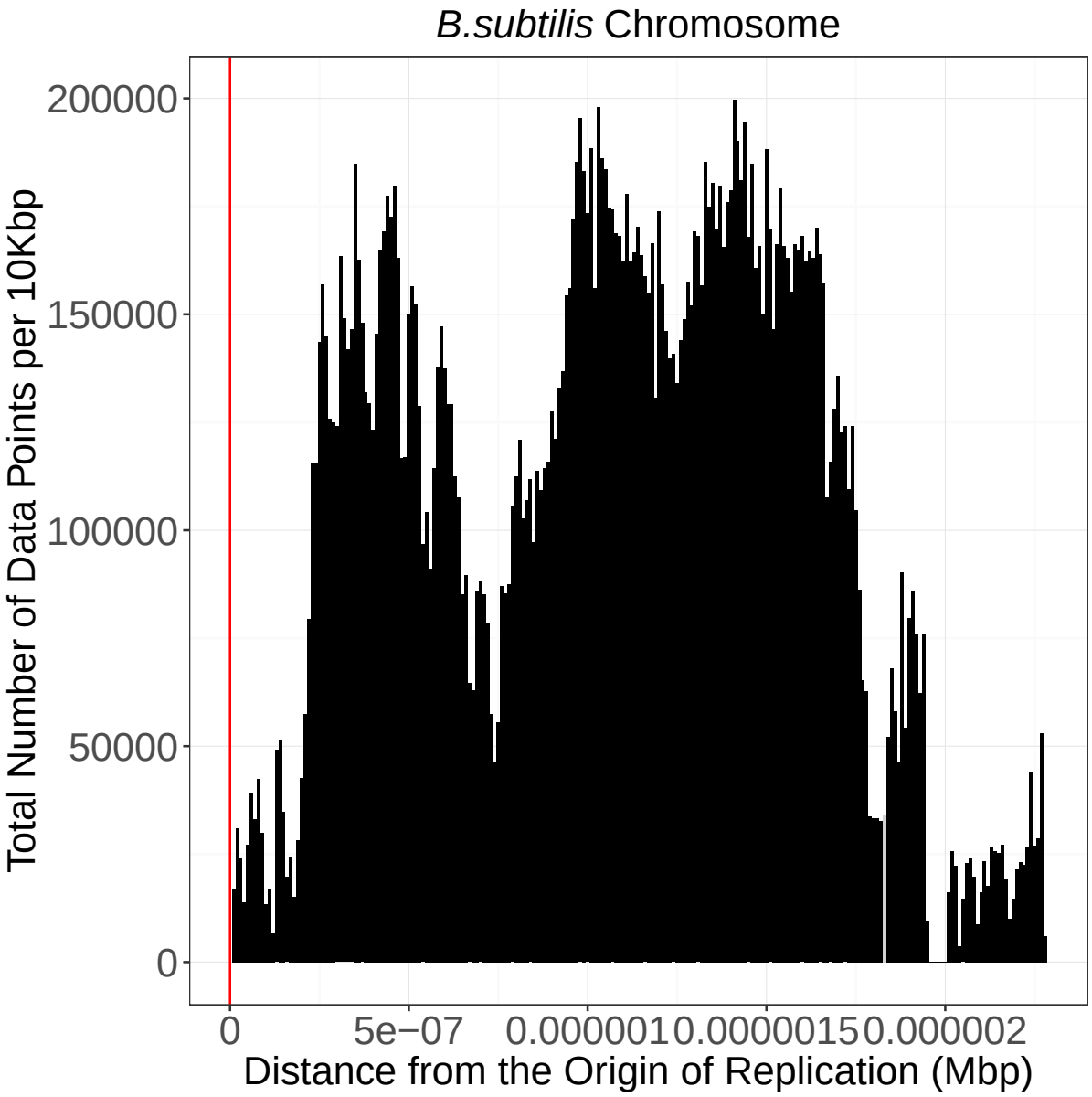


Figure 9: Distribution of total number of substitution data points per 10Kbp in genome.

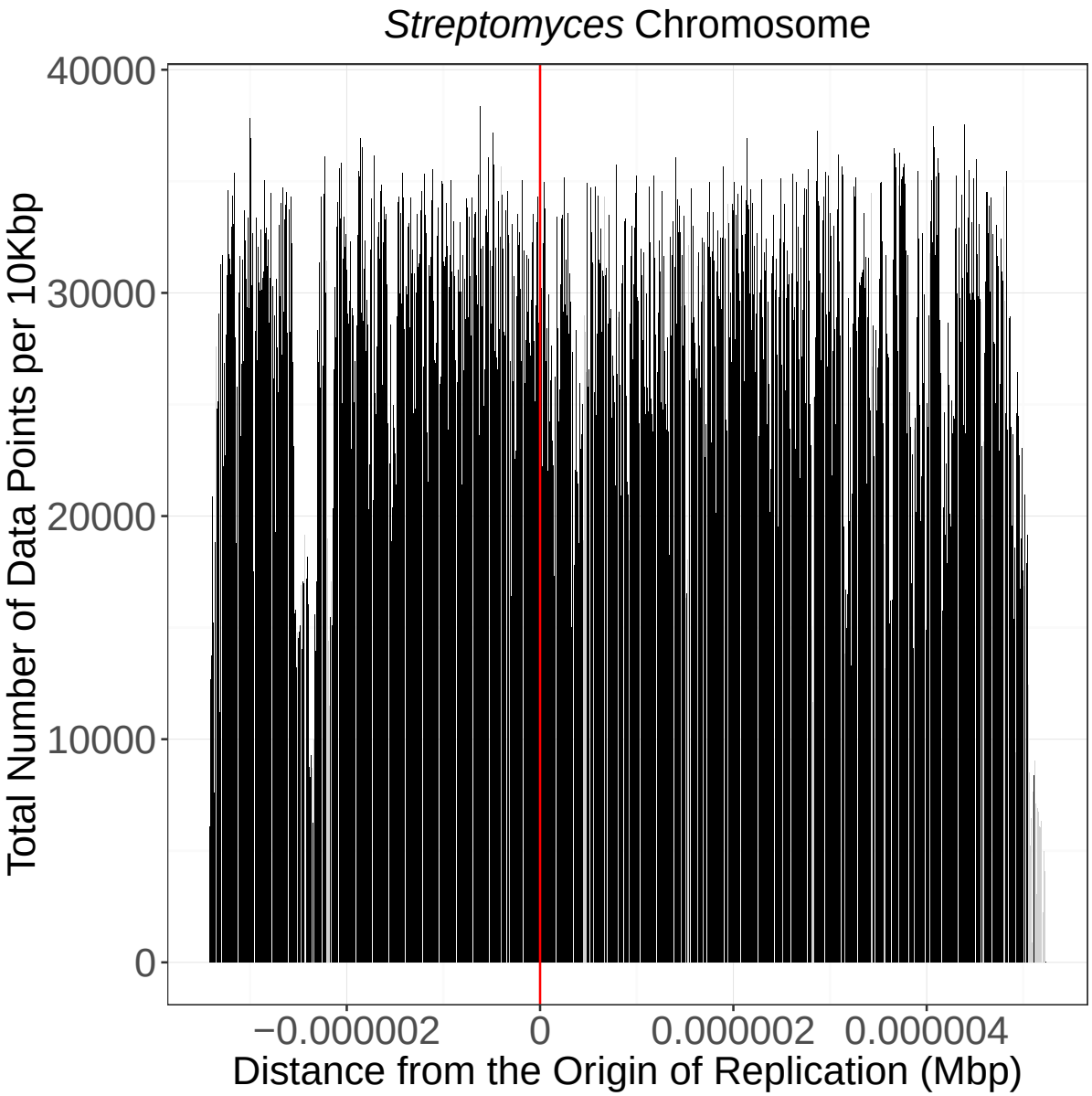


Figure 10: Distribution of total number of substitution data points per 10Kbp in genome.

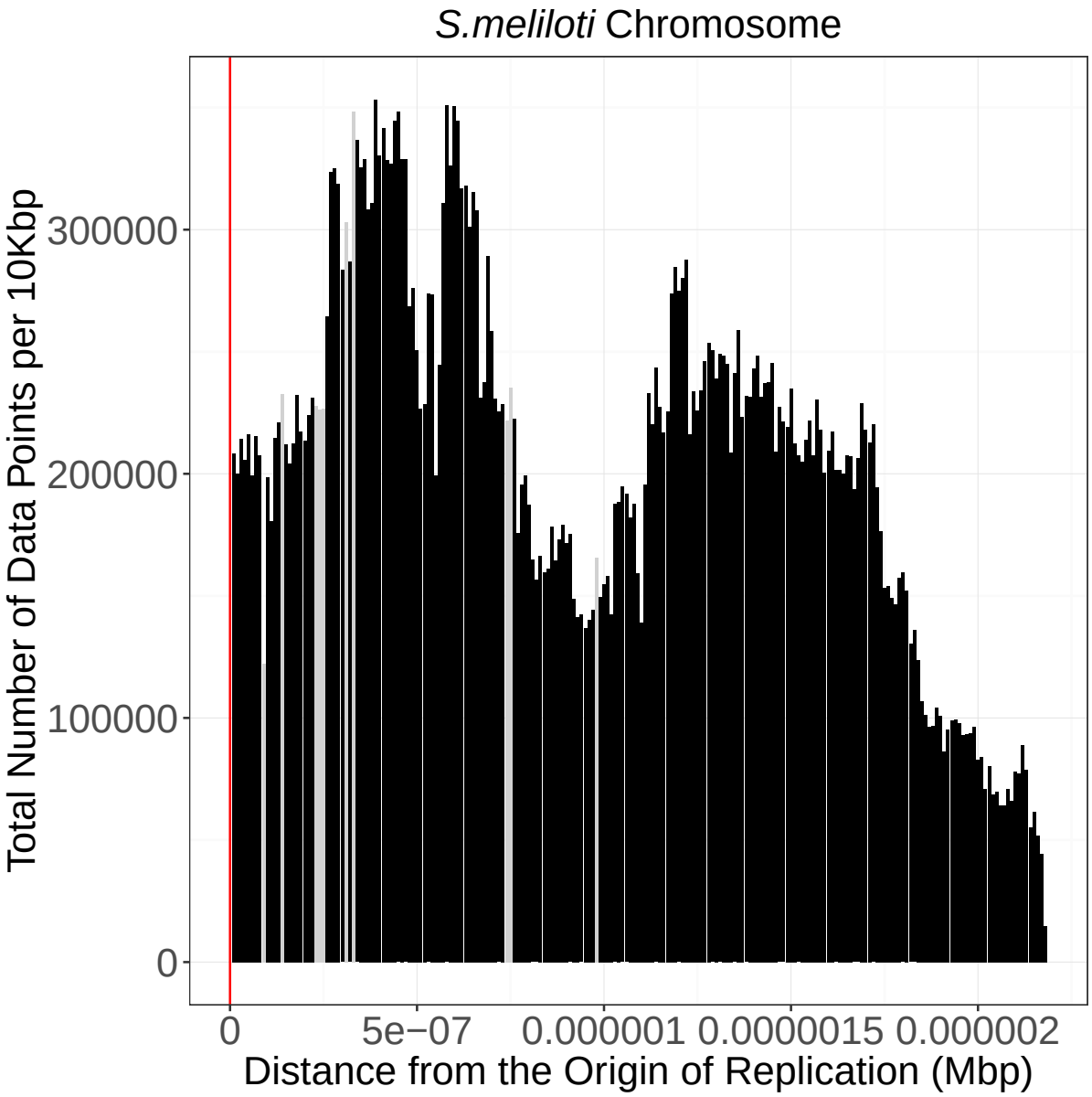


Figure 11: Distribution of total number of substitution data points per 10Kbp in genome.

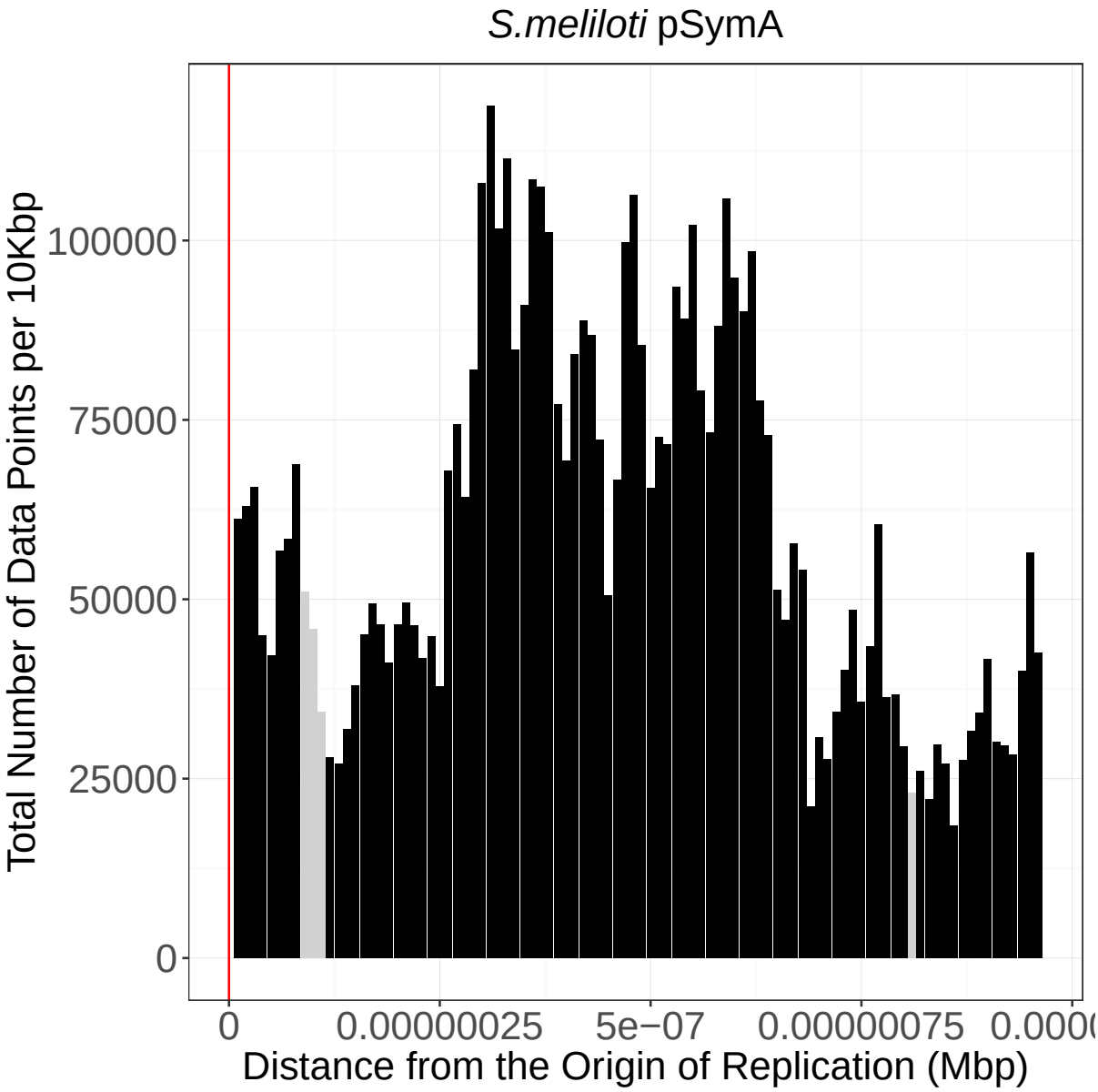


Figure 12: Distribution of total number of substitution data points per 10Kbp in genome.

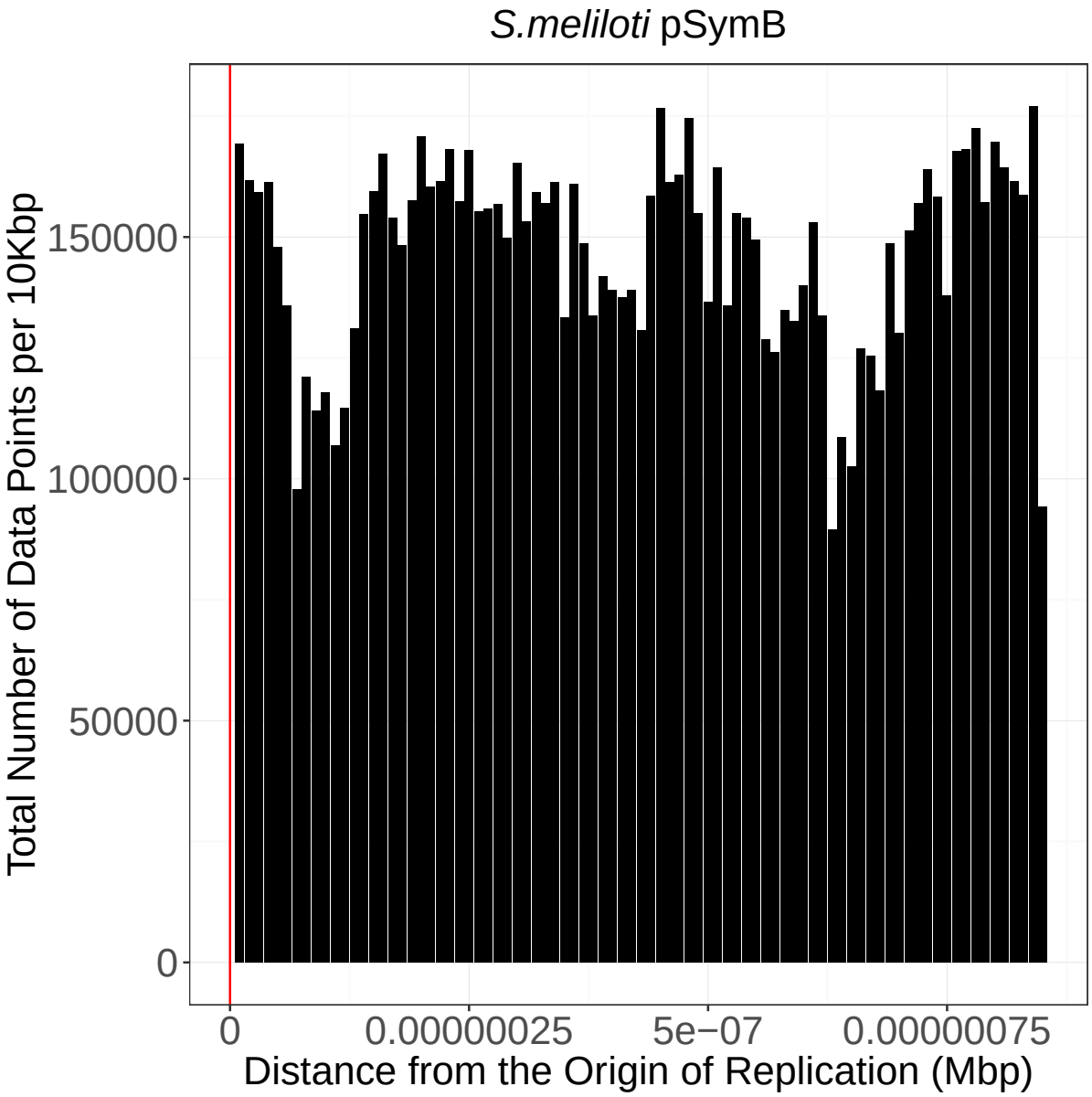


Figure 13: Distribution of total number of substitution data points per 10Kbp in genome.

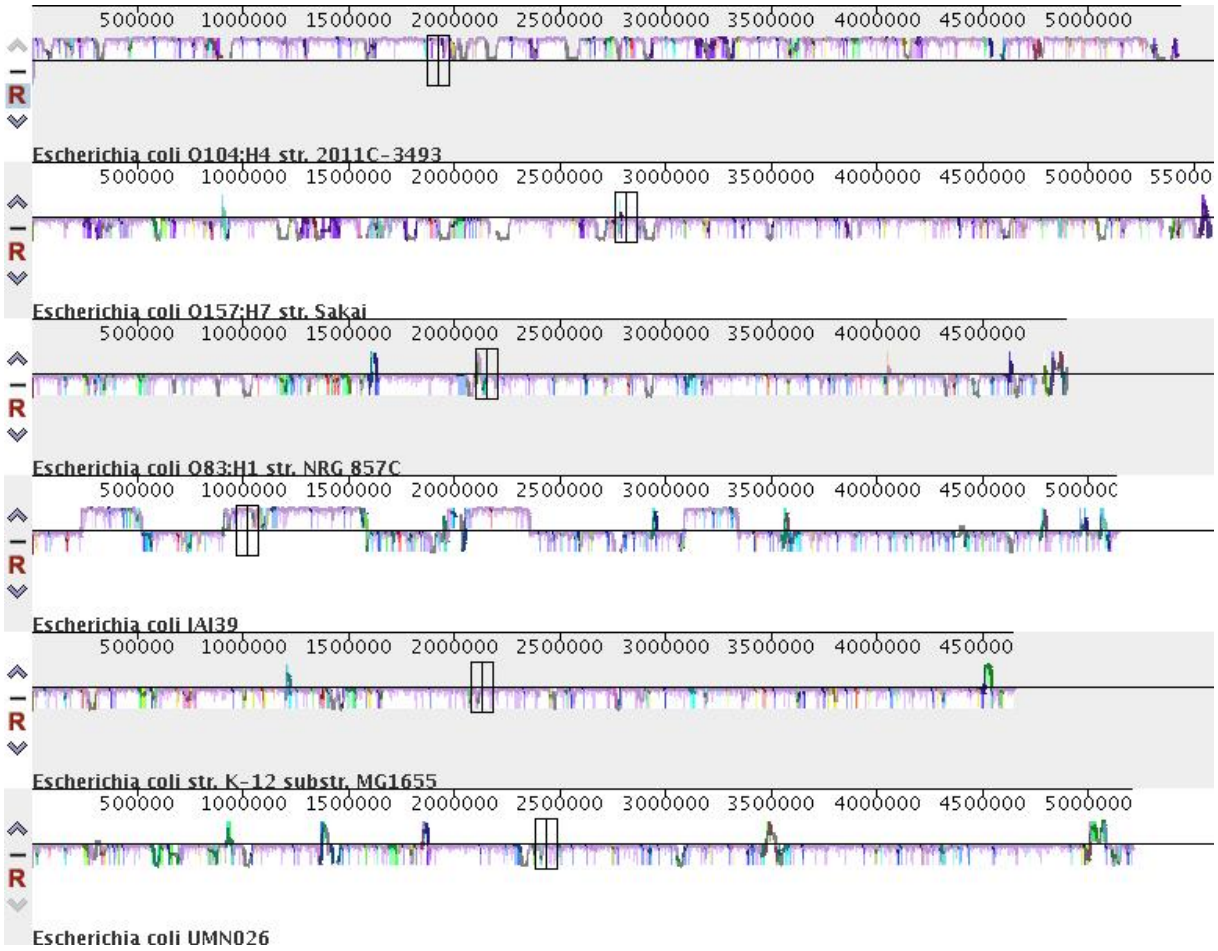


Figure 14: progressiveMauve alignment of *Escherichia coli* genomes highlighting the "backbone" of the alignment (matching regions).

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

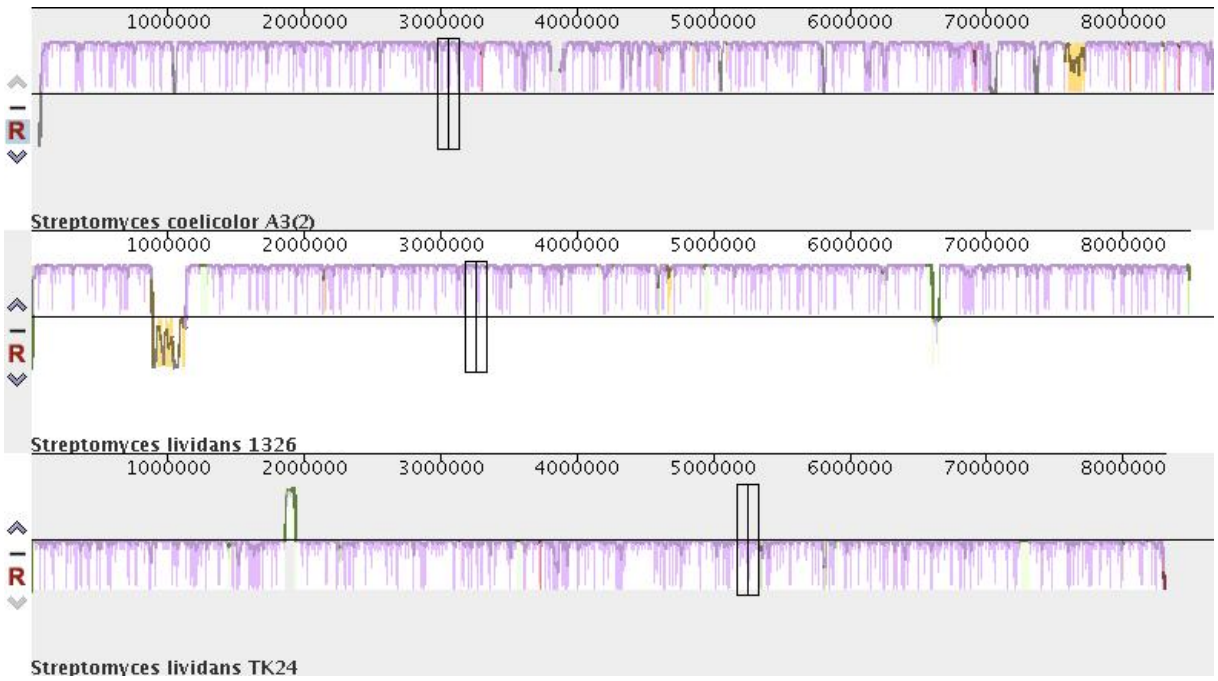


Figure 16: progressiveMauve alignment of *Streptomyces* genomes highlighting the "backbone" of the alignment (matching regions).