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:

✓done with reciprocal best blast hits (including dealing with multiple best hits)

✓Queenie: looked into overlapping blocks occurrences

General:

✓edited one section of dissertation intro

Inversions + Gene Expression:

I have finally finished my code for getting the reciprocal best blast hits and deal with multiple best hits. **Could you let me know if my method seems ok and logical?**

1. Find all best hits for each query protein (in AvsB and BvsA blast output). This is based on the lowest e-value score.
2. If there are multiple best hits (i.e. multiple subject proteins have the same lowest e-value score), calculate the difference in genomic position between the two genes.
3. The single best hit is determined by the lowest difference in genomic positions (my attempt at synteny)
4. Find all instances where the best hit for blast A (AvsB) query c is subject d and the best hit for blast B (BvsA) query d is subject c (reciprocal part of the rbbh)

I was unsure how to best assess synteny because it is supposed to be gene order, but each blast hit only tells you one gene that the query matches to. So I decided to use genomic position (from the gbk files) to tell me if these genes are in the same part of the genome. Usually this is accurate and will select the hit that is the same gene (same gene name, protein seq...etc). However, there are a few cases where a different subject gene is selected because it is closer in genomic position compared to the subject gene with the same gene (same gene name, protein seq..etc). **Thoughts on me using the genomic position difference between the query and subject genes as a good measure of synteny?**

Additionally, the proteomes of the *E. coli* strains are very redundant. So in the case of *E. coli* K-12 DH10B, the proteome used was for *E. coli* K-12 W3110 and for *E. coli* BW25113 the proteome use was *E. coli* K-12 MG1655, and finally for *E. coli* ATCC the proteome used was *E. coli* CFT073. This was because the proteomes are redundant and the second proteome had more complete information about gene names. **Does this make sense and seem like a sound thing to do?**

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 (paeudo 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 PASNP 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

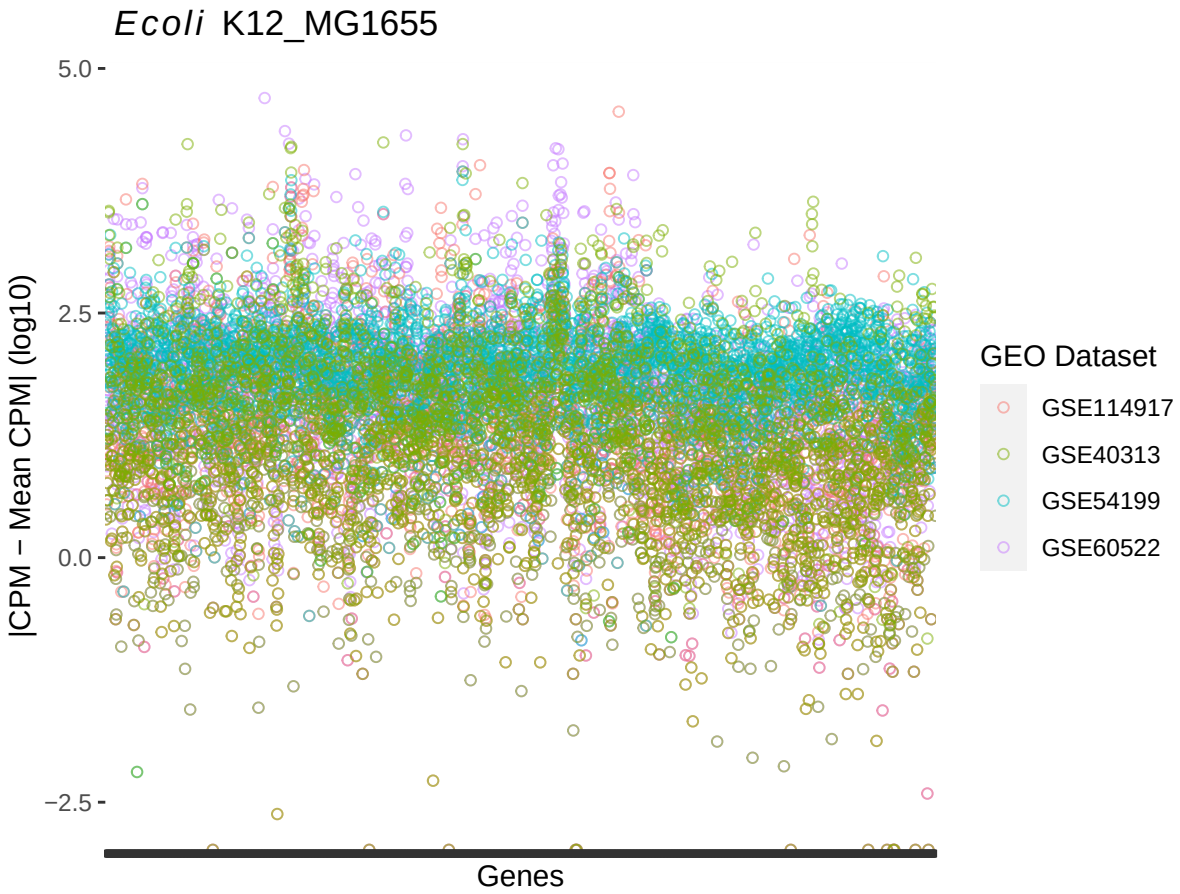


Figure 1

	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.

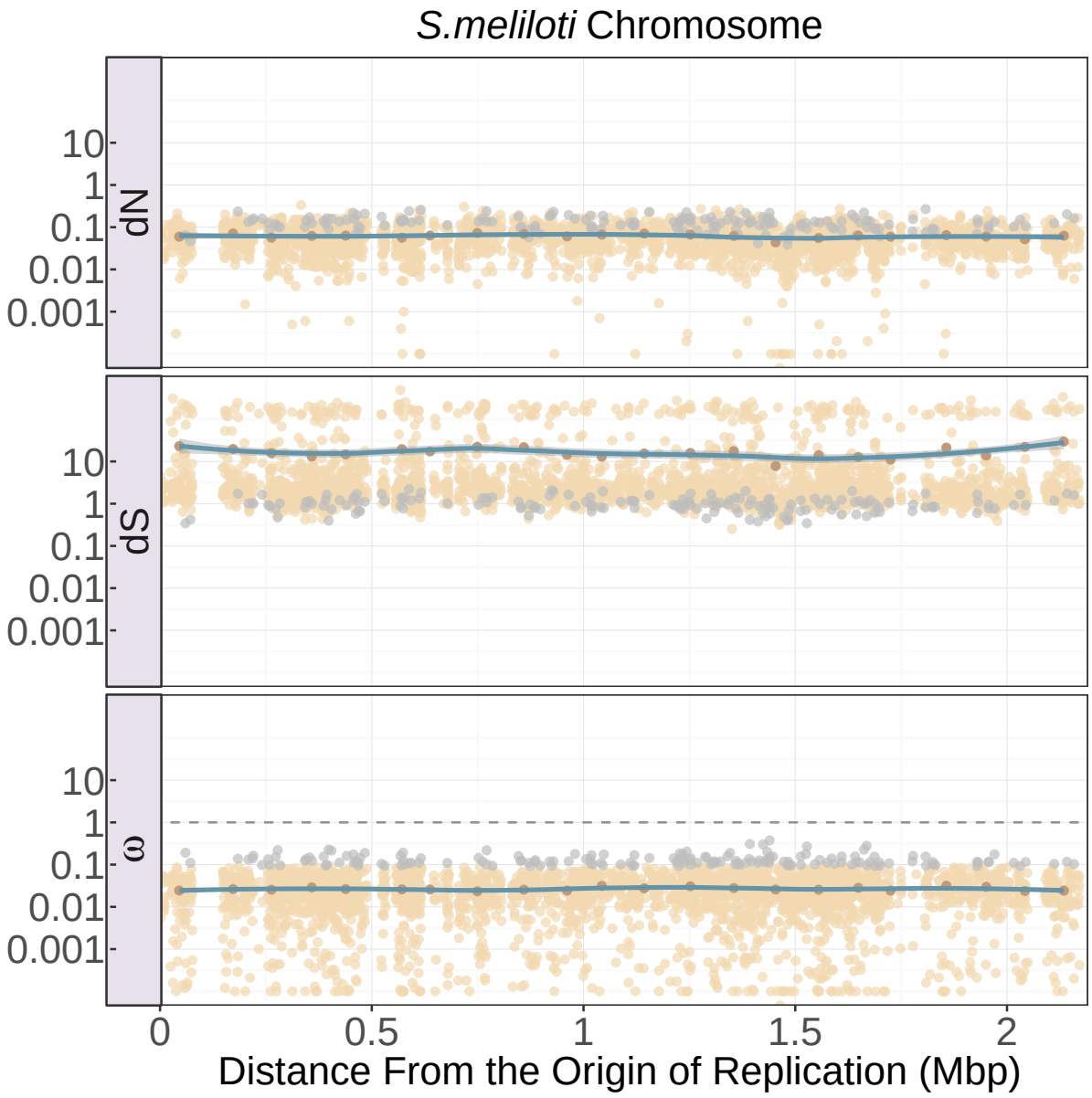
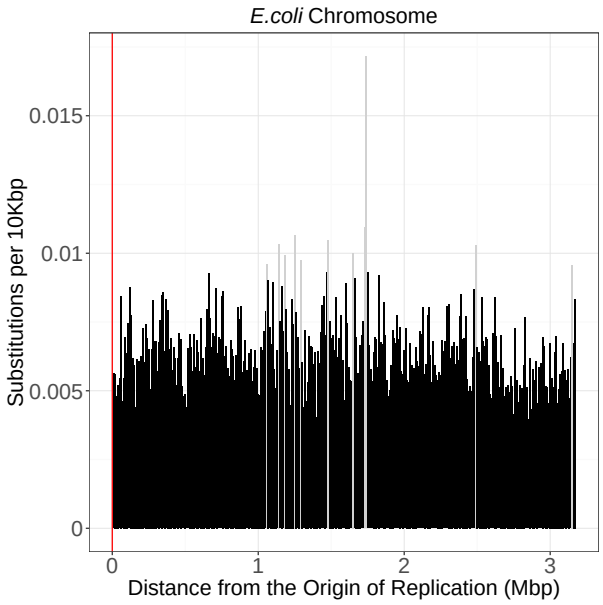


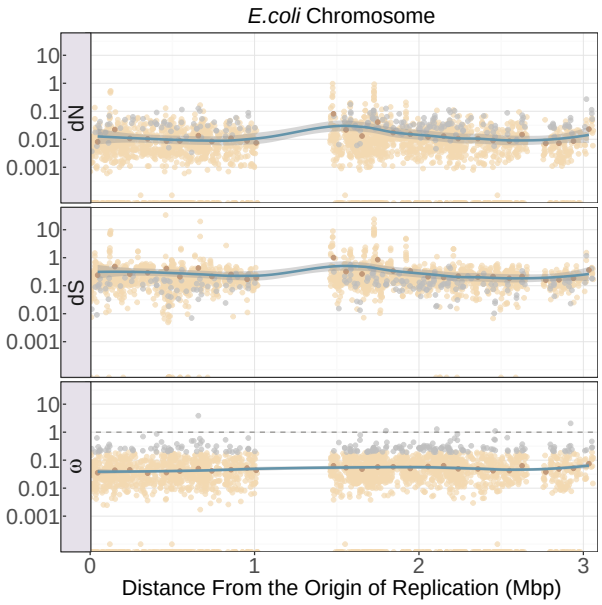
Figure 2: dN, dS, and ω values for *S. meliloti* chromosomes and ***A. tumefaciens***.

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}

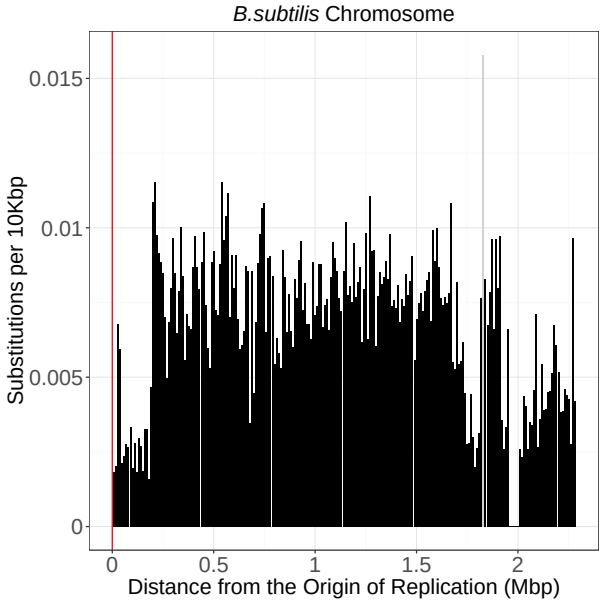
Table 2: Average number of protein coding substitutions calculated per base across all bacterial replicons. Outliers and missing data was not included in the calculation.


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

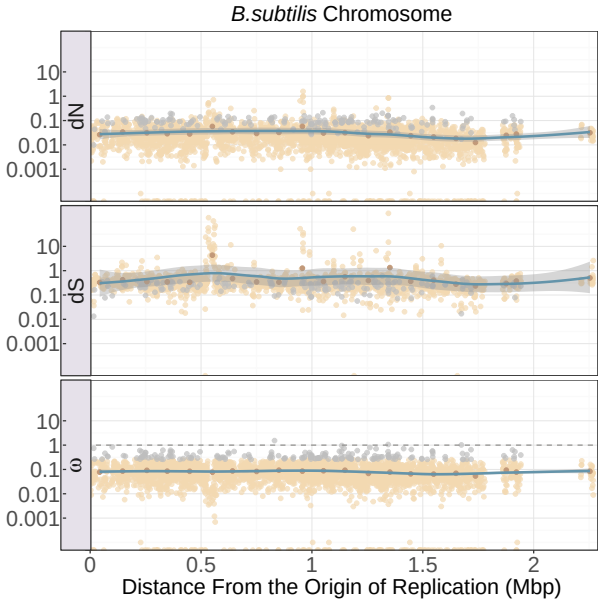
(a)


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)

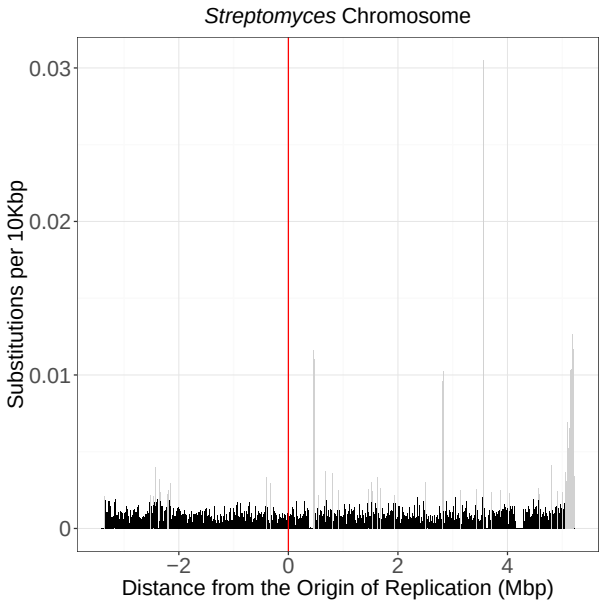
(b)


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)

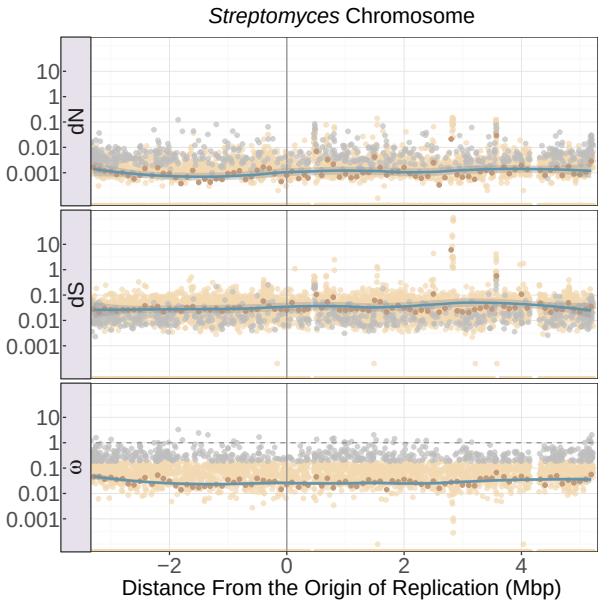
(a)


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)

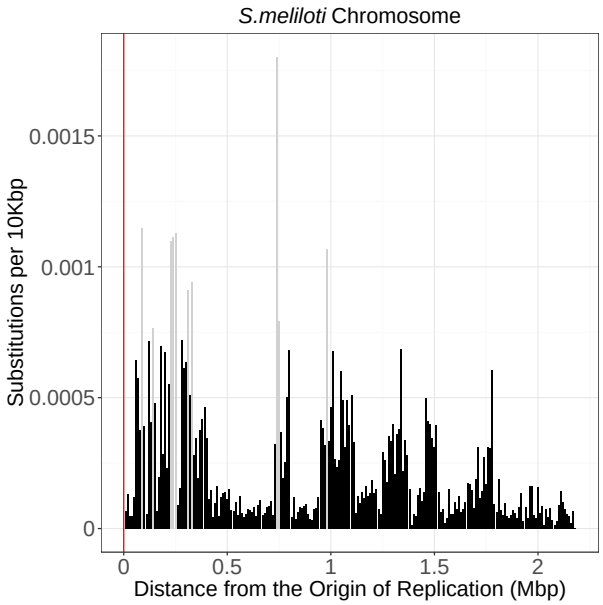
(b)


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

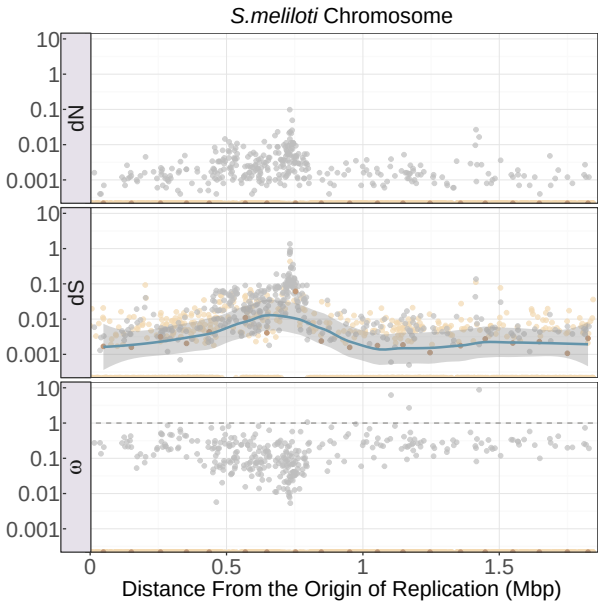
(a)


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

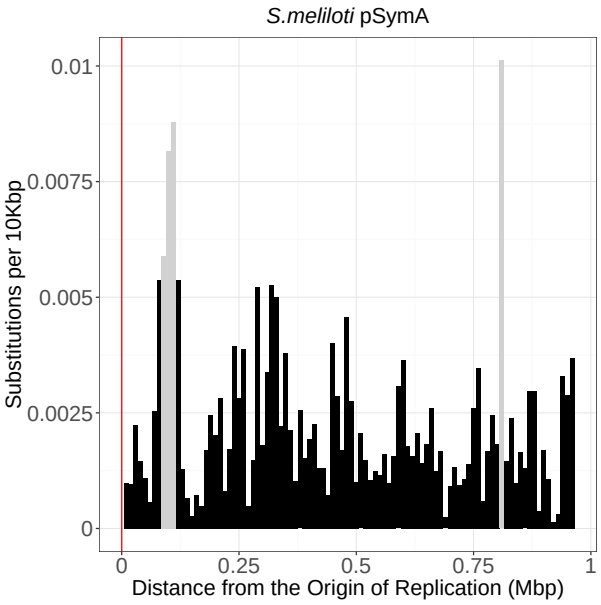
(b)


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)

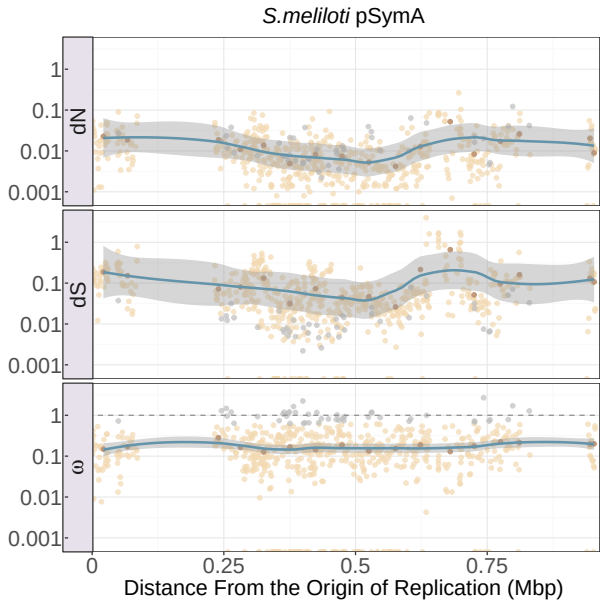
(a)


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)

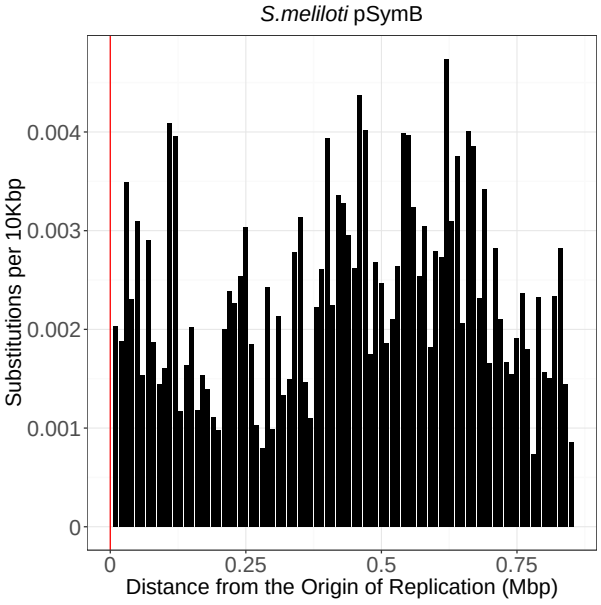
(b)


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)

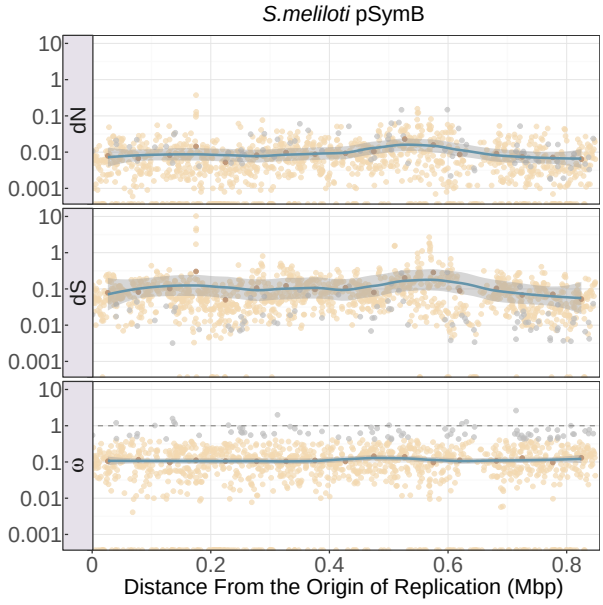
(a)


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)

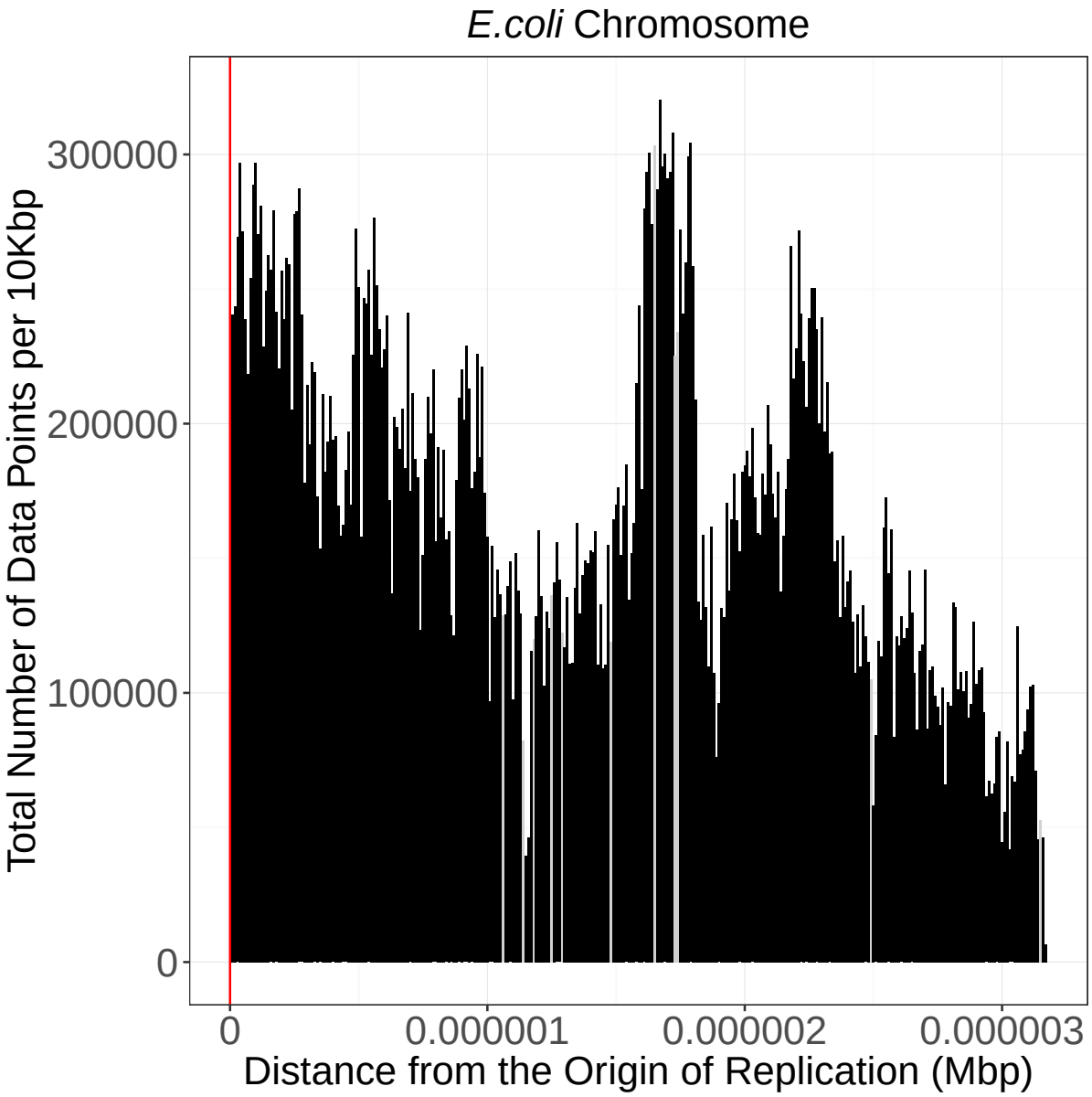


Figure 9: Distribution of total number of substitution data points per 10Kbp in genome.

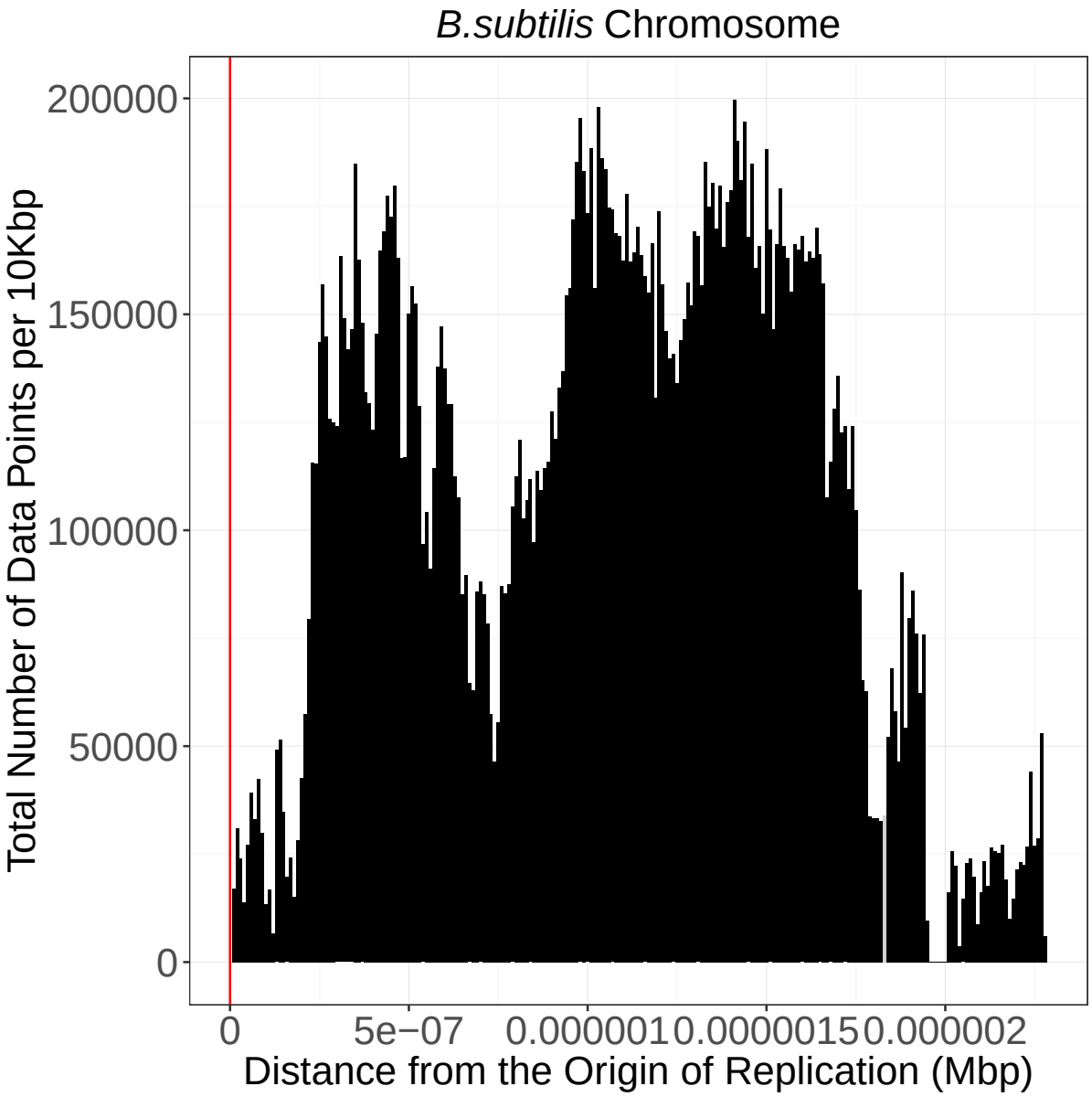


Figure 10: Distribution of total number of substitution data points per 10Kbp in genome.

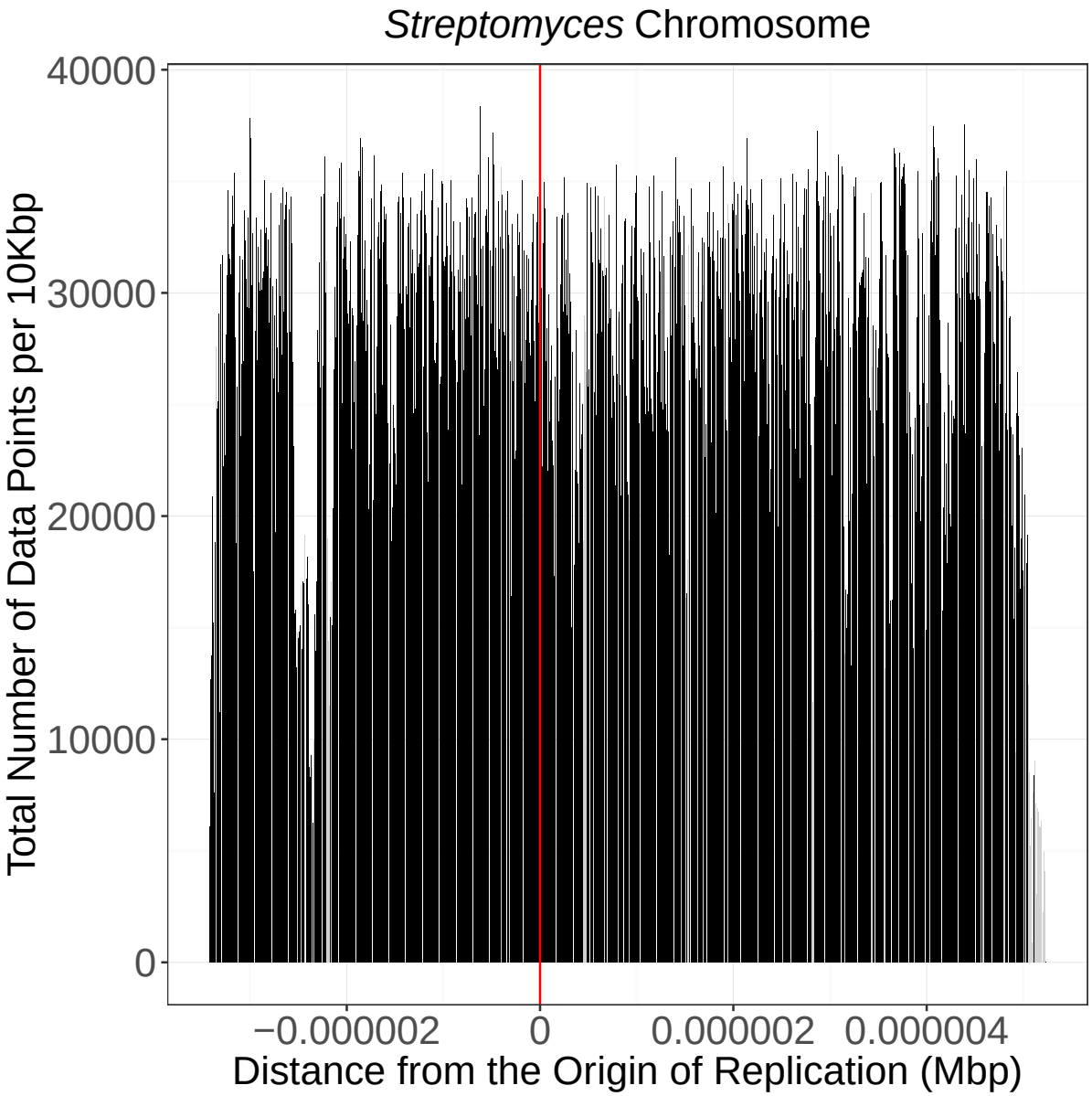


Figure 11: Distribution of total number of substitution data points per 10Kbp in genome.

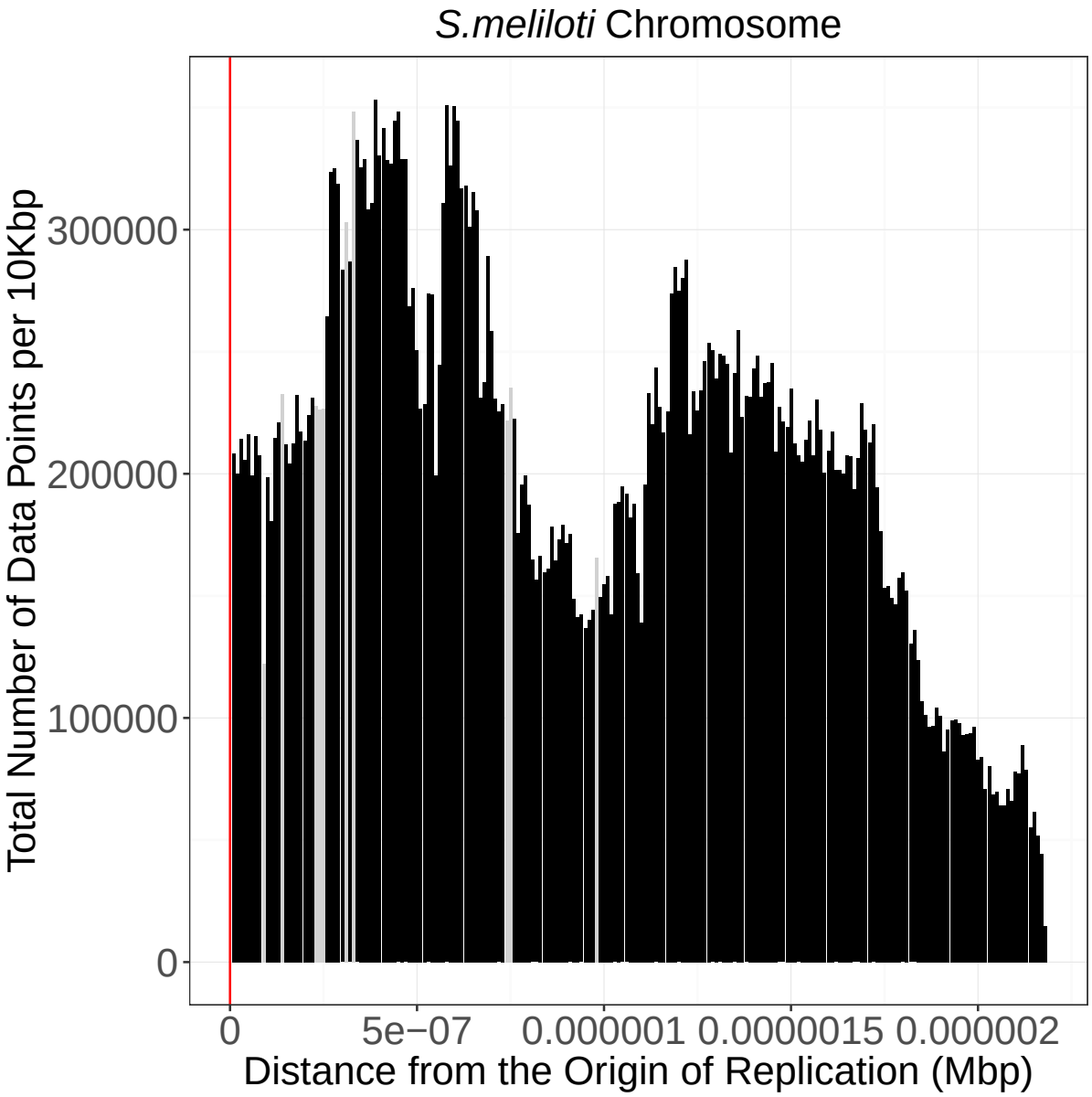


Figure 12: Distribution of total number of substitution data points per 10Kbp in genome.

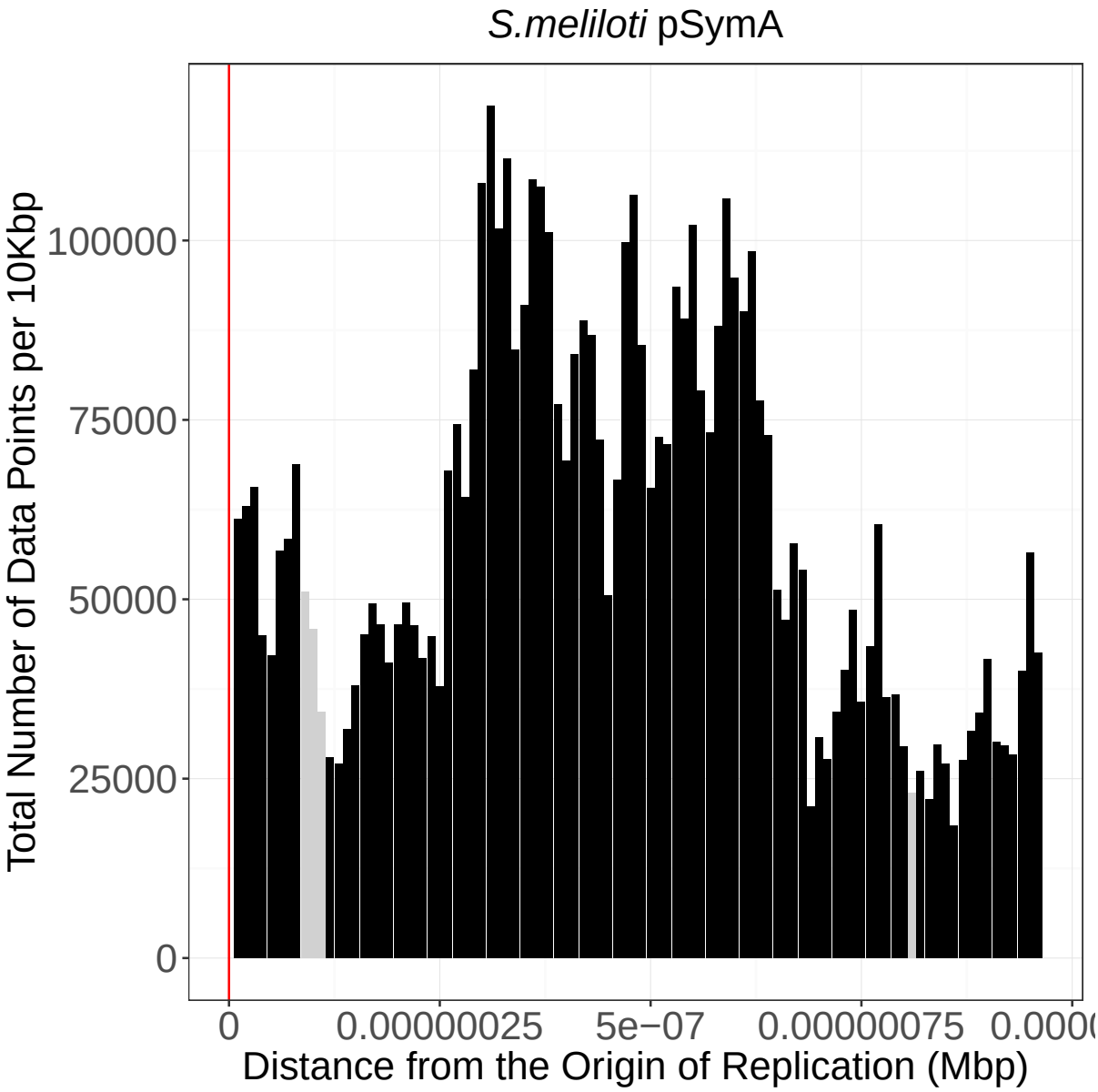


Figure 13: Distribution of total number of substitution data points per 10Kbp in genome.

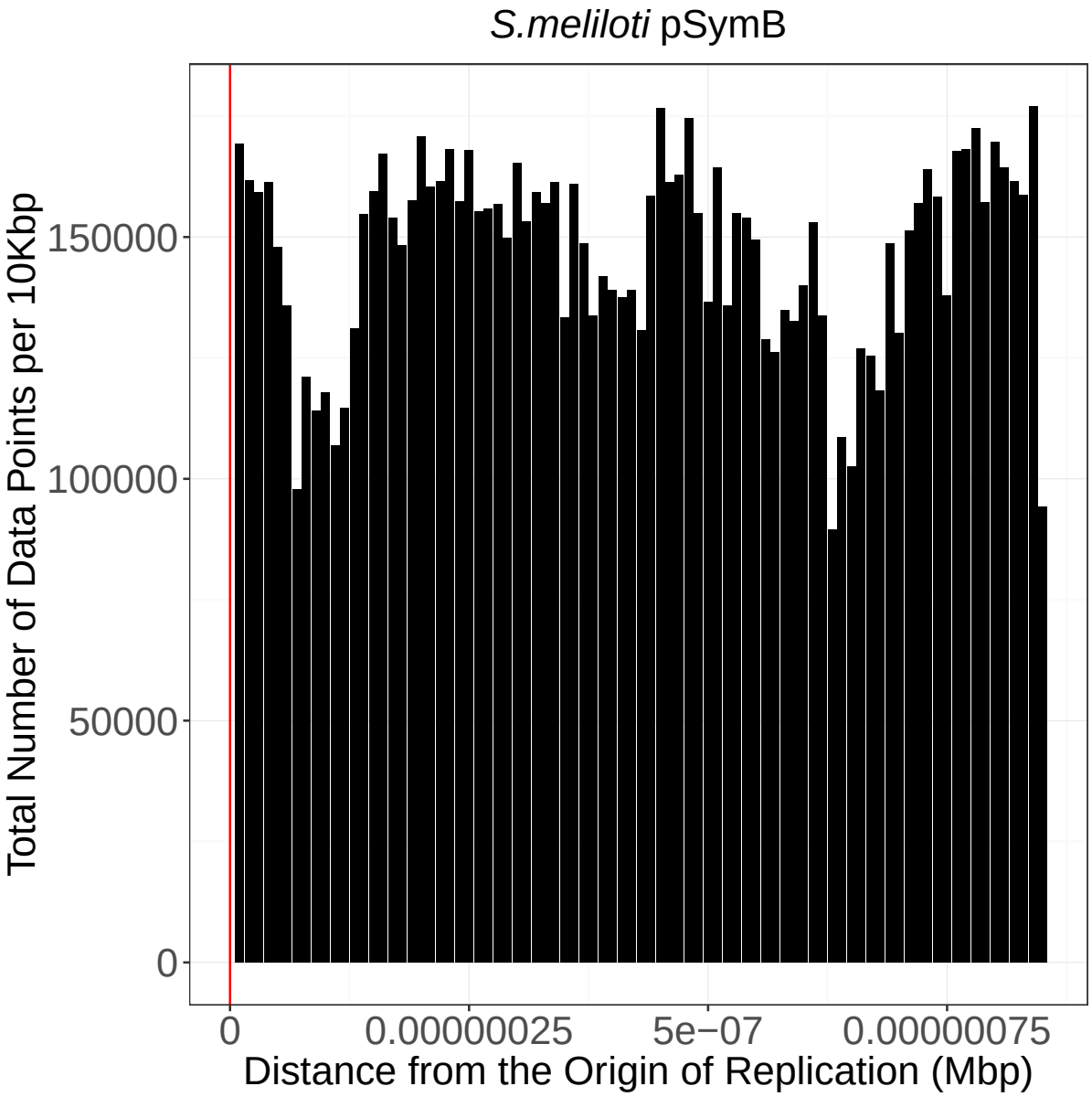


Figure 14: Distribution of total number of substitution data points per 10Kbp in genome.

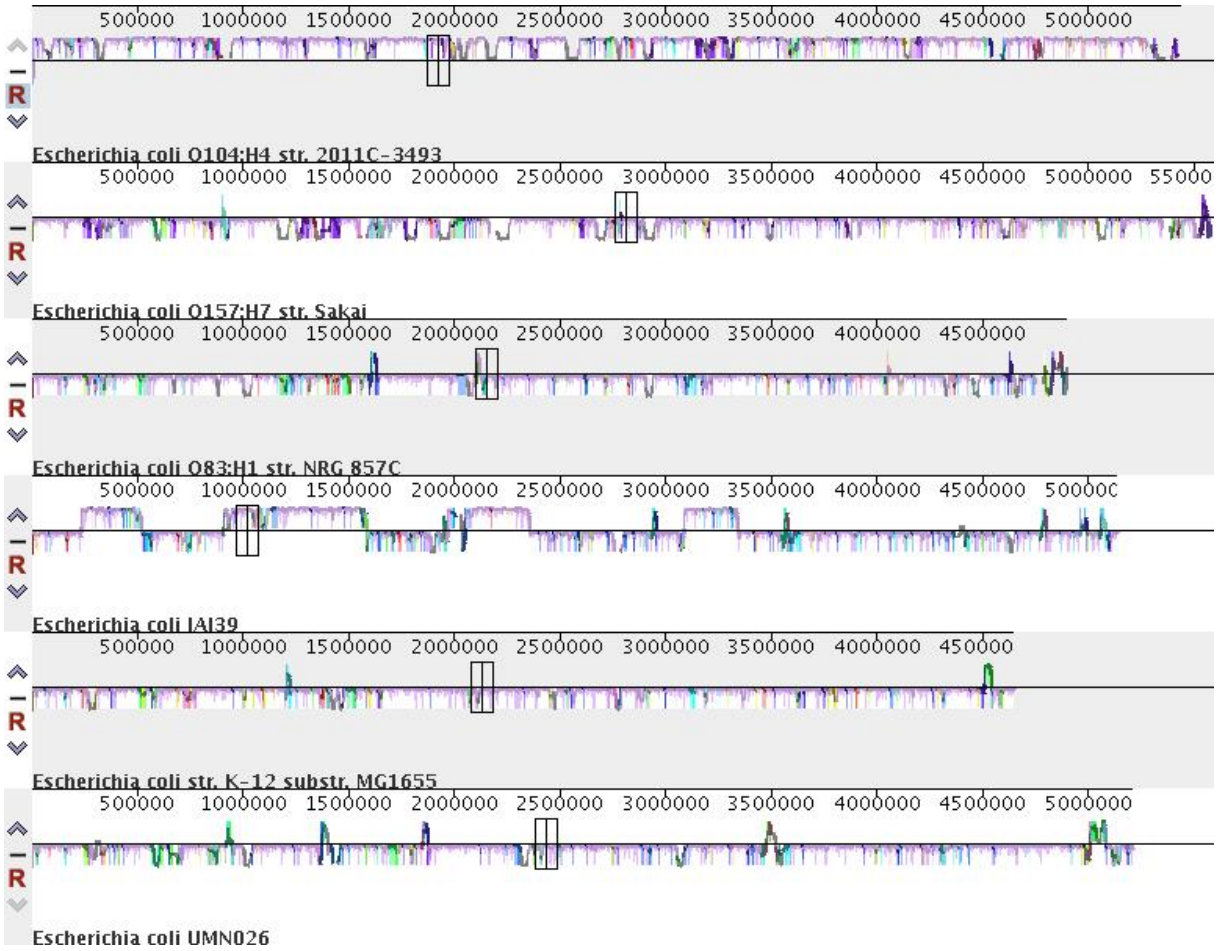


Figure 15: progressiveMauve alignment of *Escherichia coli* genomes highlighting the "backbone" of the alignment (matching regions).

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

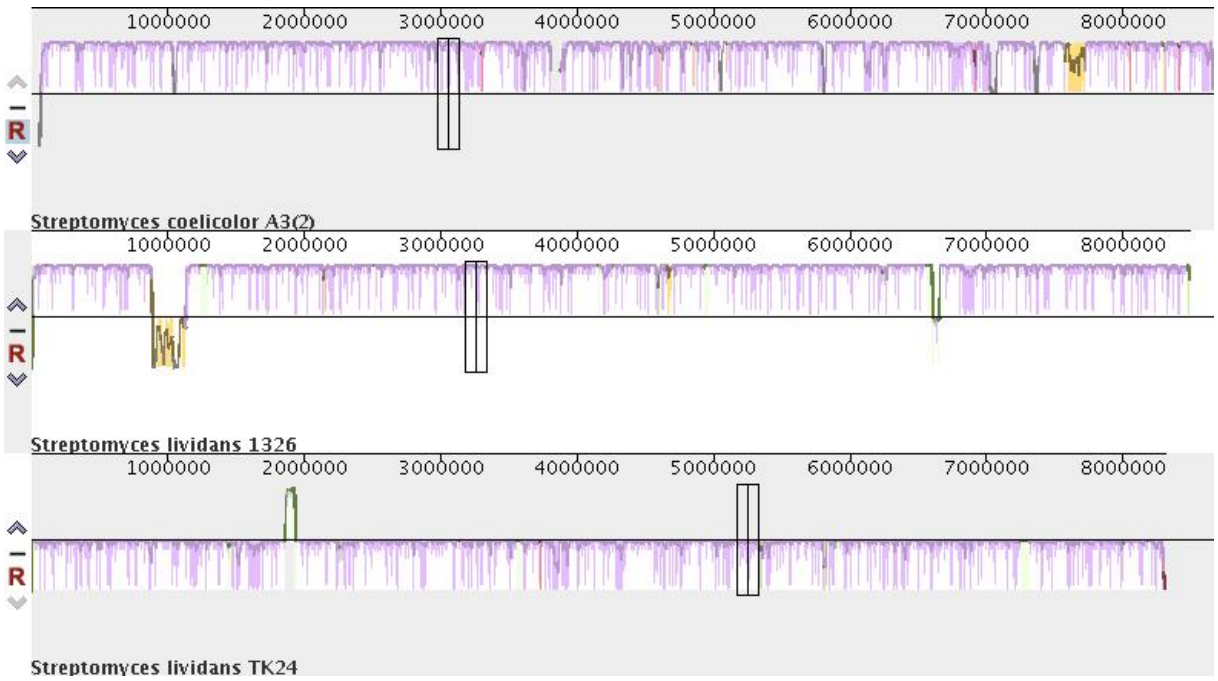


Figure 17: progressiveMauve alignment of *Streptomyces* genomes highlighting the "backbone" of the alignment (matching regions).