Subs Paper Things to Do:

- why are the lin reg of $dN$, $dS$ and $\omega$ NS but the subs graphs are...explain!
- mol clock for my analysis?
- GC content? COG? where do these fit?

Gene Expression Paper Things to Do:

- if necessary add a phylogenetic component to the analysis
- codon bias?

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

✓added a table of high substitutions bars and what genes were in them to supplement

✓deal with properly removing outliers in subs analysis

✓updated outlier removal in methods

✓accounted for your comments in my methods

✓re-did the average substitution calculation

✓added some supporting info about other opposing molecular trends that people found to my intro and discussion

When I was making the table for the contents of the higher substitutions bars for each replicon, I noticed that for all the replicons of *S. meliloti*, there were a LOT of misc features listed in the gbk file. I need to look into this more but this might be why things are so weird with the *S. meliloti* chromosome, I did not include misc features in my analysis, and if everything is listed as a misc feature, then that leaves very little data. I need to look into this more.

I properly removed all outliers for the substitutions analysis. This is what I spent most of the week on because my brain was really slow this week so it took me a while to conceptually figure it all out. All the outliers are properly removed now and the results/methods are updated in the paper draft.

I made corrections based on your comments on my methods section and sent you a revised version. During this I realized that the way I was calculating the average number of substitutions per site for each genome was weird and wrong (as you pointed out). So I re-did this calculation and updated the results and methods.

I am still really confused about why the selection graph for *S. meliloti* Chromosome looks so odd and the only explanation I can come up with is that the sequences are just really really similar. **Do you have any thoughts on this or suggestions for other things I could investigate to figure out what is going on?**

# This Week

- continue look into whats up with *S. meliloti* chrom bc it does not look right at all
- look into the numerous misc features in *S. meliloti*
- change caption for selection distribution figures so that they match how many bp the averages were calculated over (PA and PB are different)
- fix the results to properly talk about the selection and substitution figs

- make a comment about why there are two lines in the box plot (one at 0 and one at about 0.0001), in caption? or discussion?
- add that high dS values are also real and due to real changes where most of the gene is syn changes with very few non-syn changes and therefore it skews the whole calculation, creating a very high dS value. mention supplemental high subs bar

# Next Week

- think about if the selection distribution figs or the summary selection fig should be in the main paper
- re-word captions for all figures so they make sense with current figures
- fix discussion

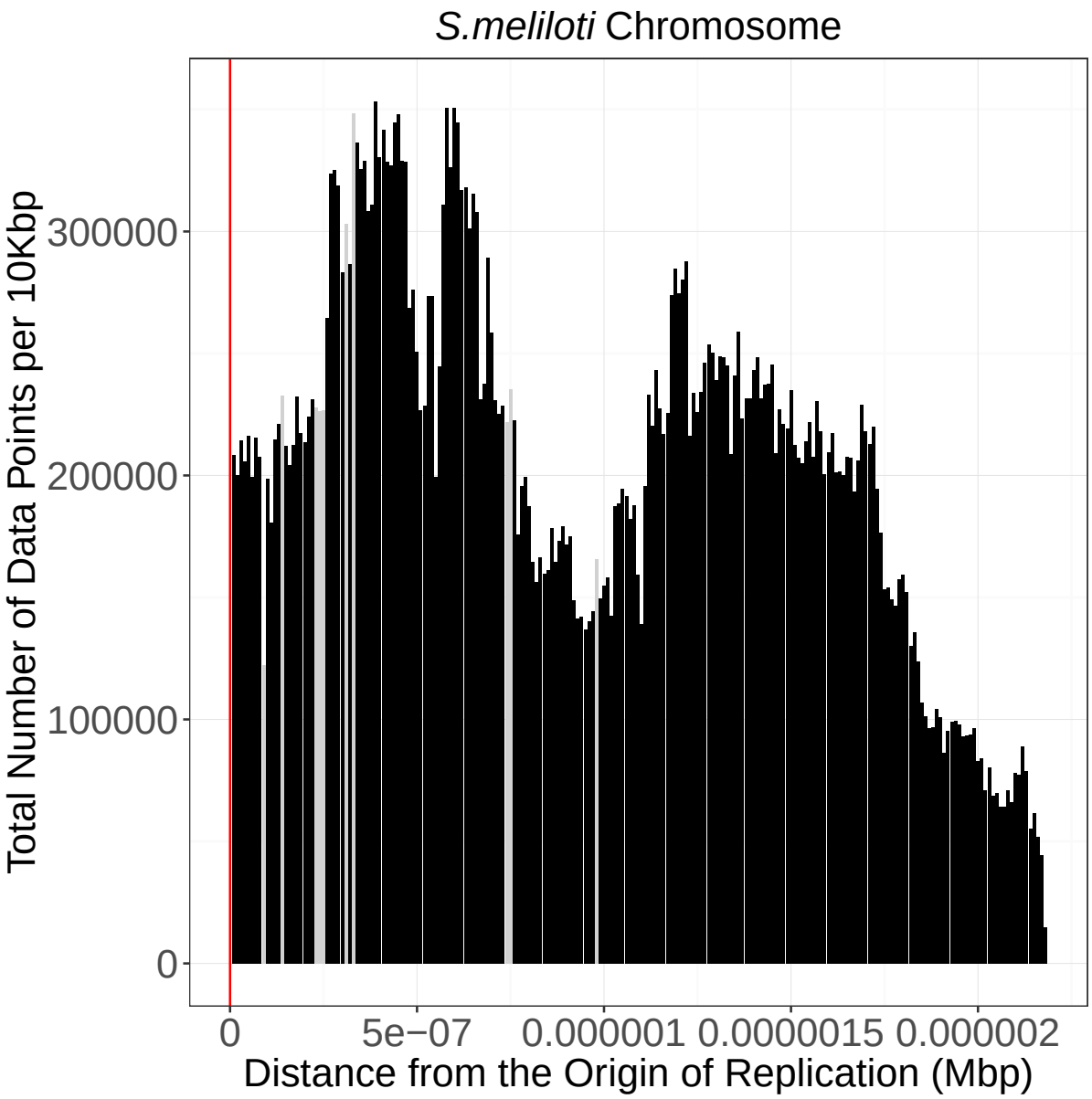



Figure 1: Distribution of total number of substitution data points per 10Kbp in genome.

| | Genome Average | | |
|---|---|---|---|
| Bacteria and Replicon | dS | dN | $\omega$ |
| ***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.

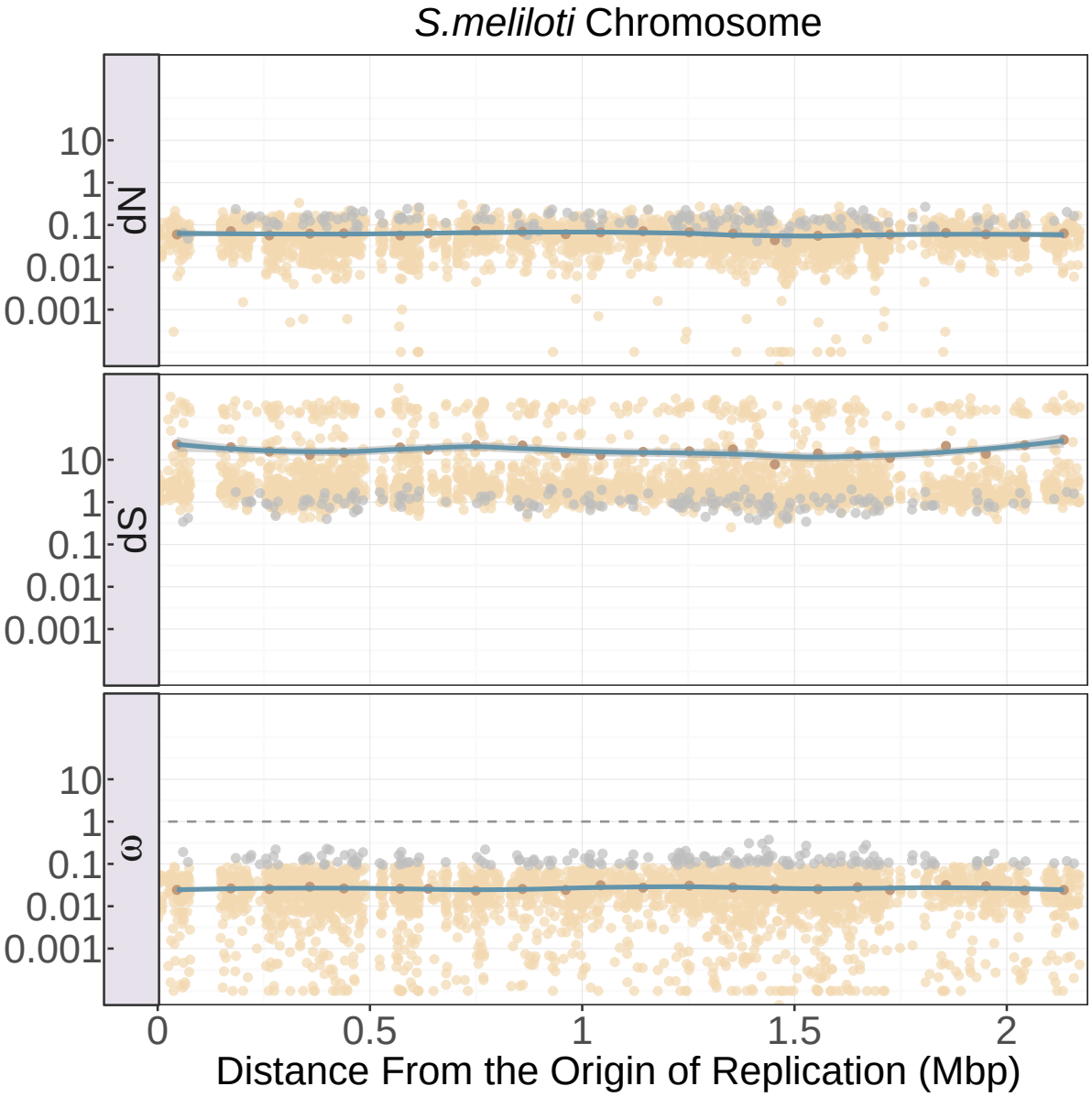
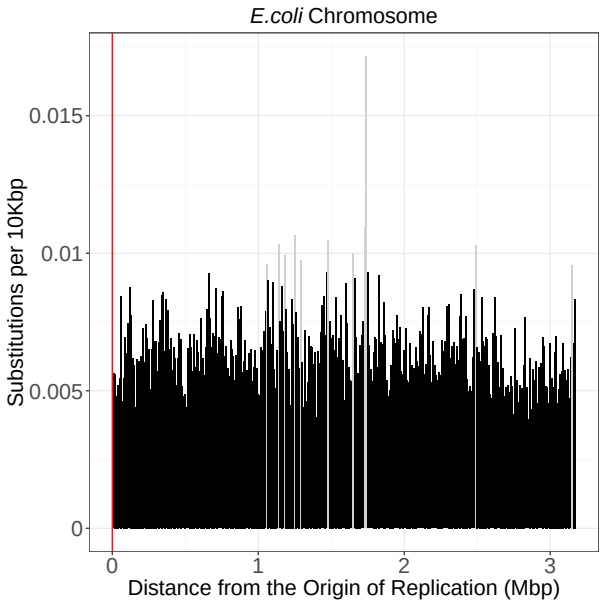
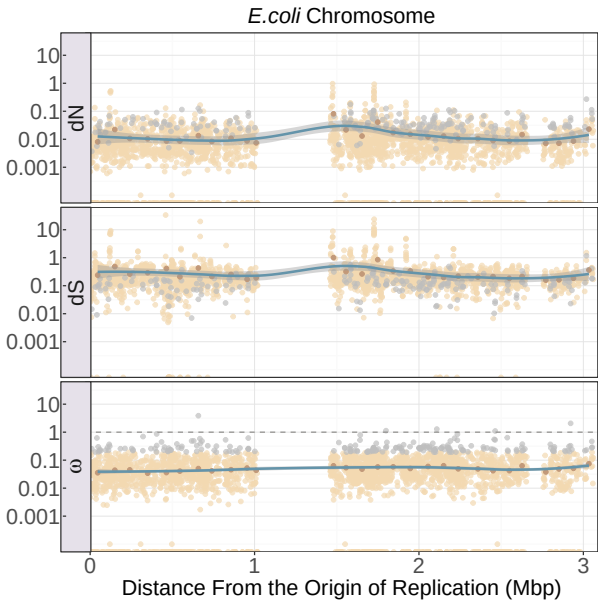


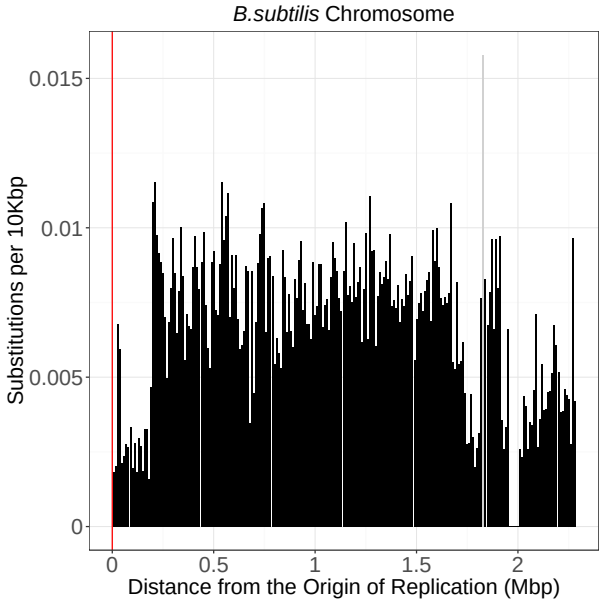

Figure 2: $dN$, $dS$, and $\omega$ values for *S. meliloti* chromosomes and ***A. tumefaciens***.


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

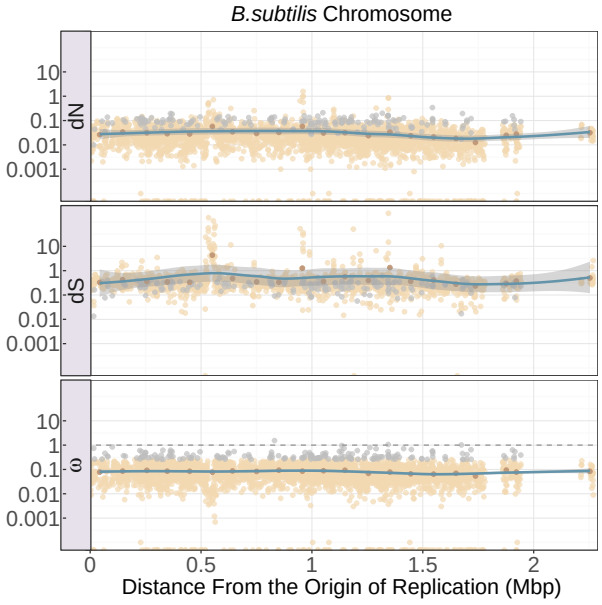
(a)


*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)

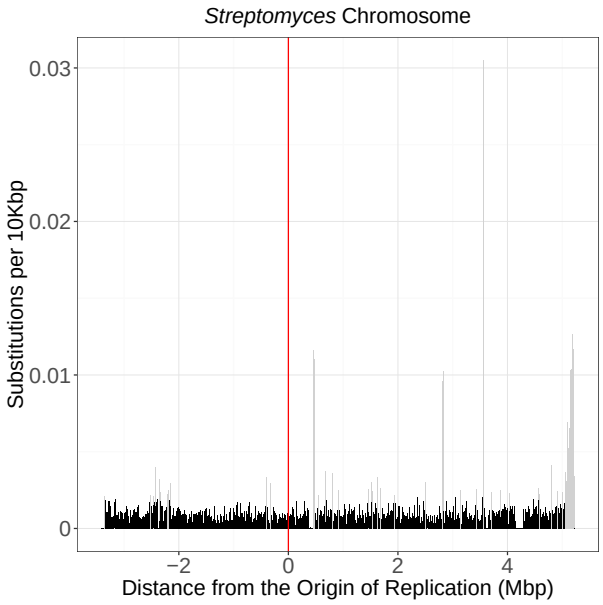
(b)


*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)

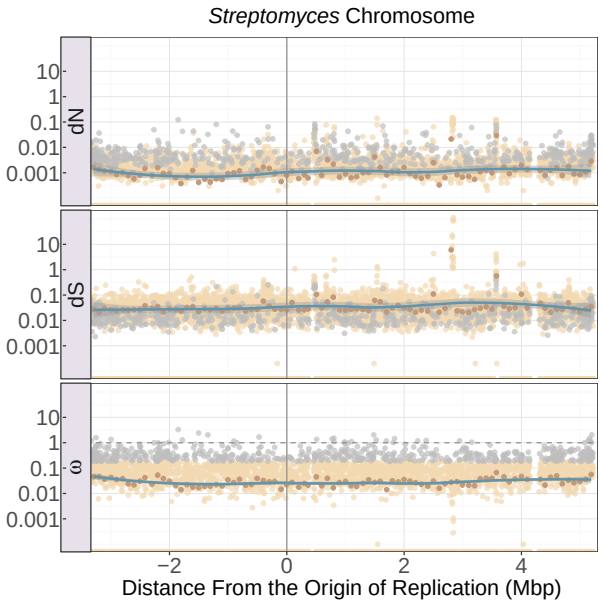
(a)


*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)

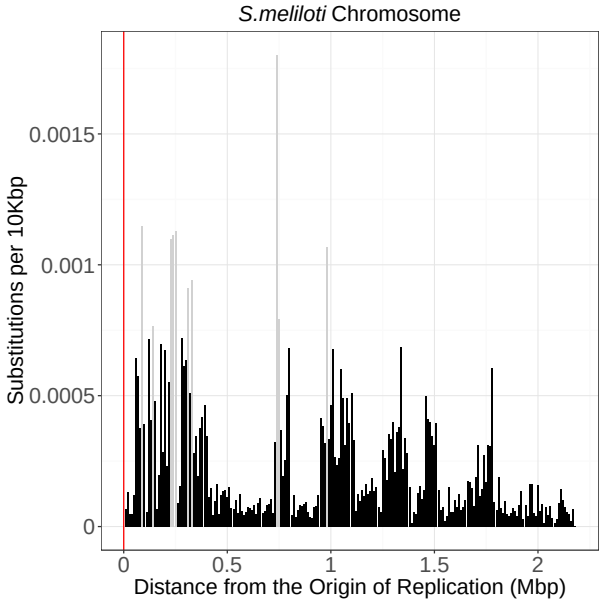
(b)


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

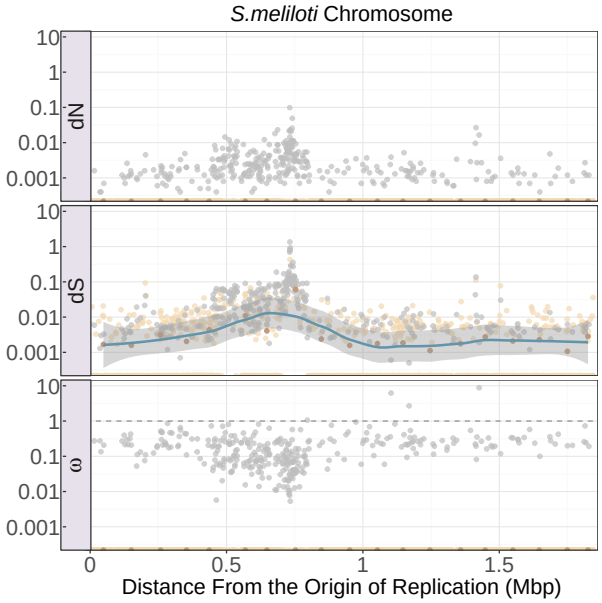
(a)


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

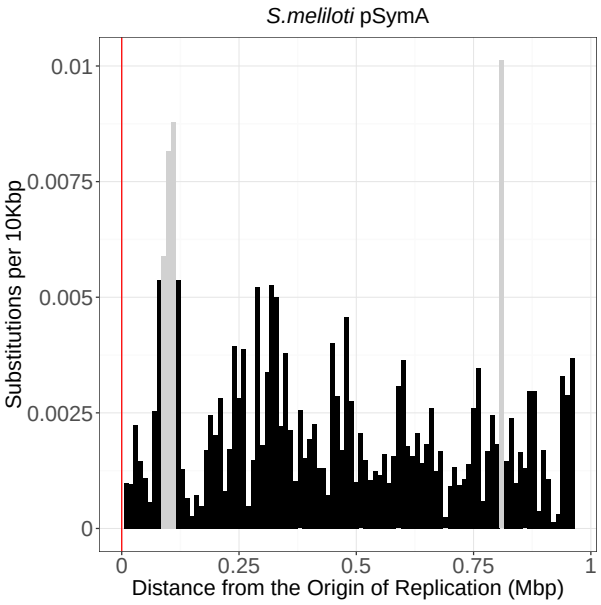
(b)


*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)

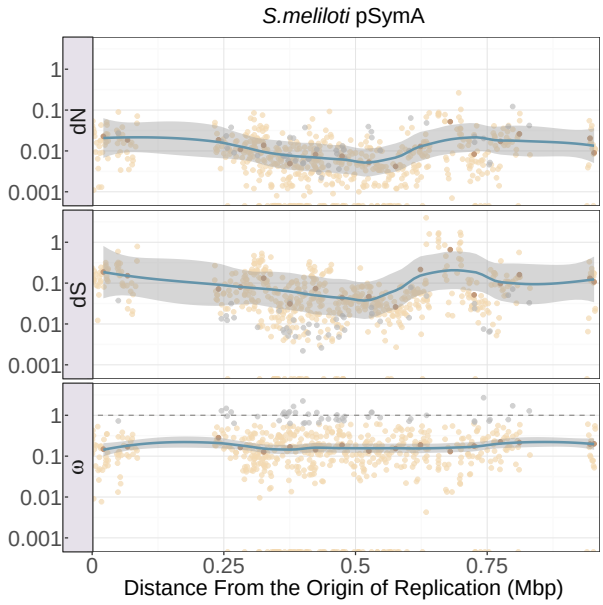
(a)


*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)

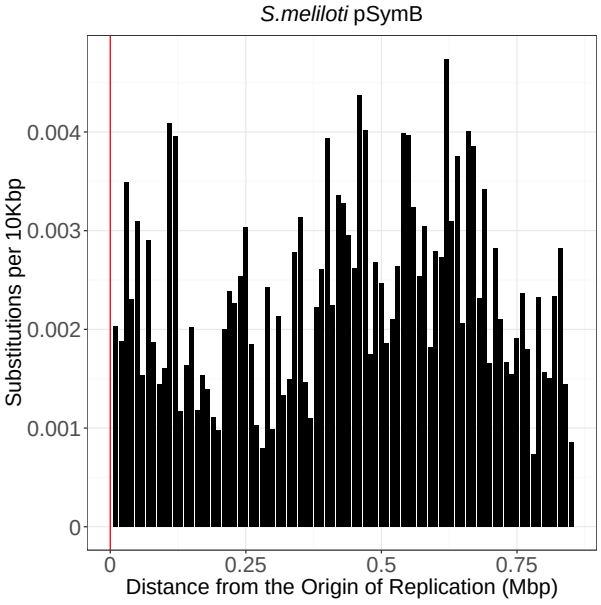
(b)


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)

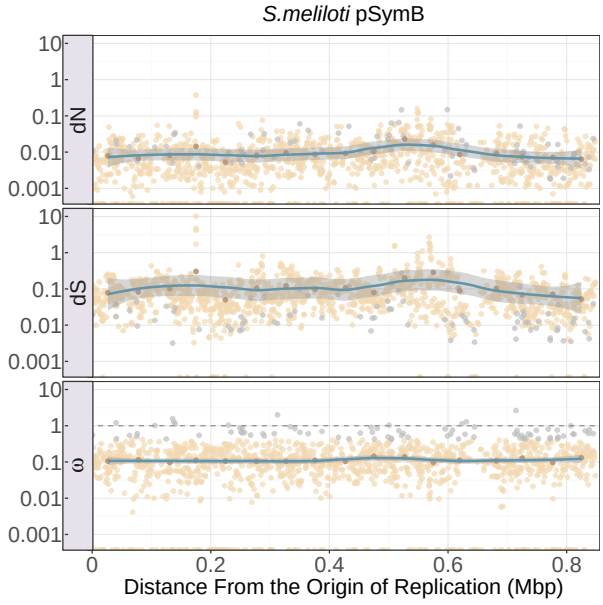
(a)


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)

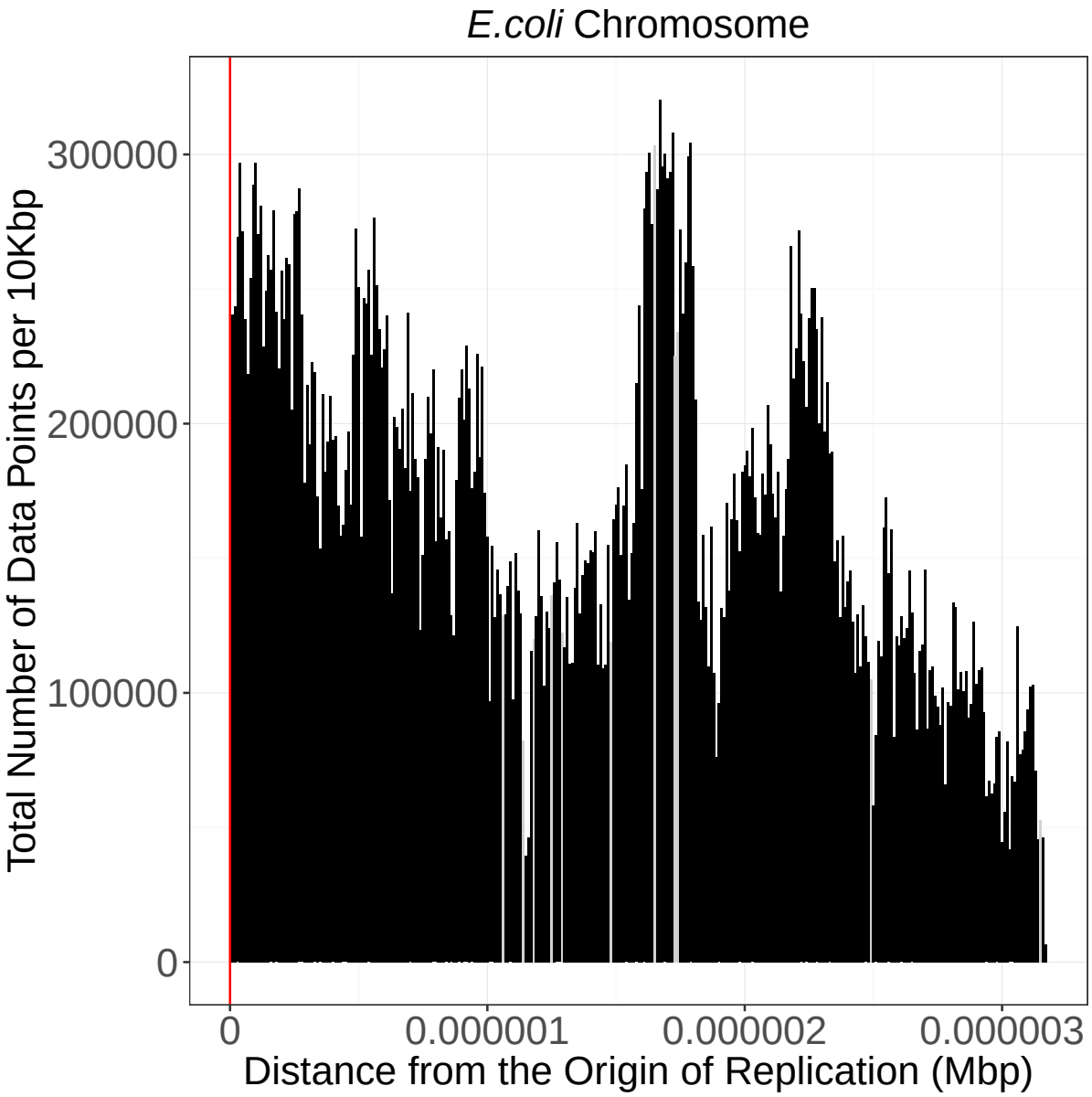



Figure 9: Distribution of total number of substitution data points per 10Kbp in genome.

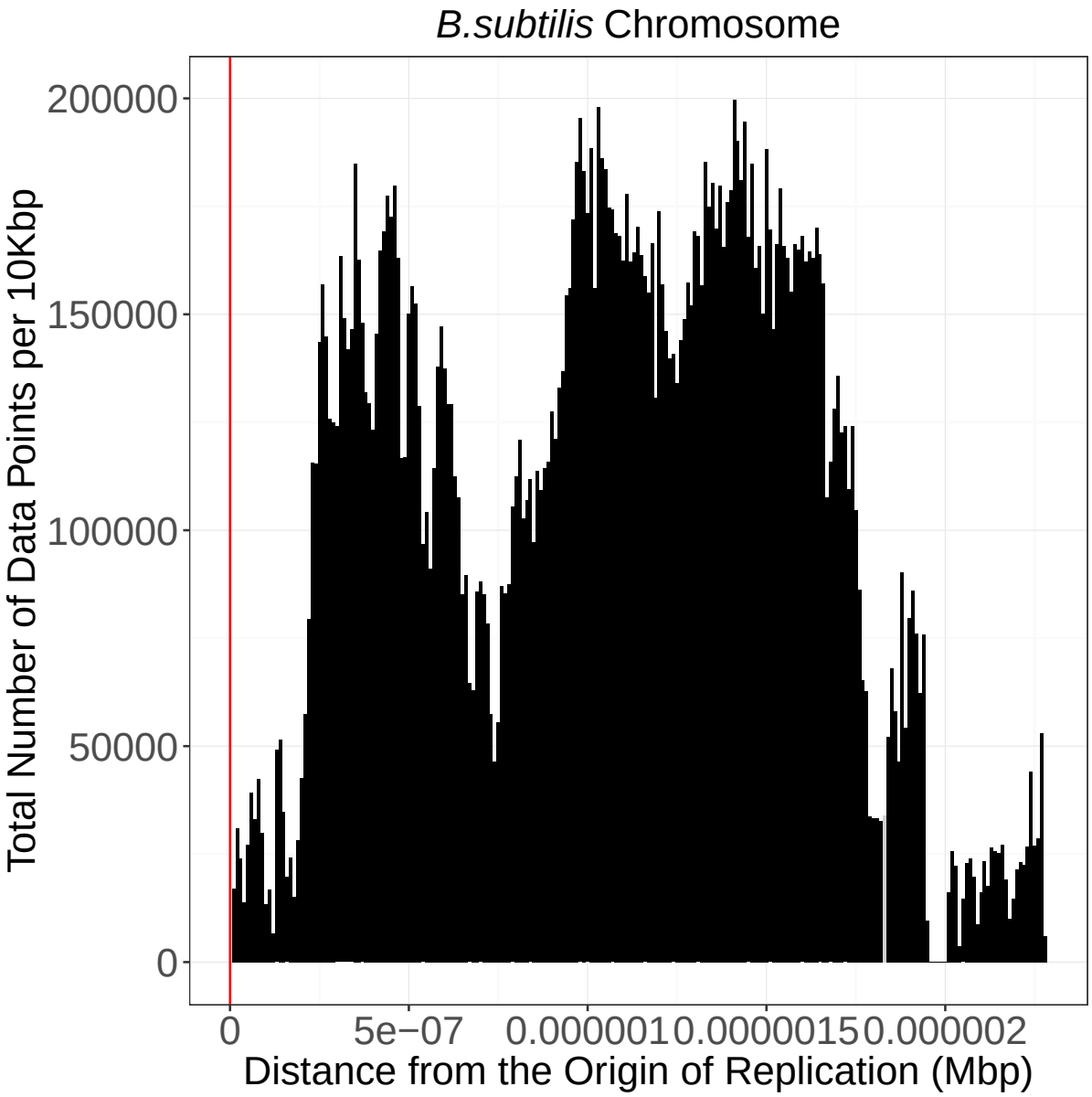



Figure 10: Distribution of total number of substitution data points per 10Kbp in genome.

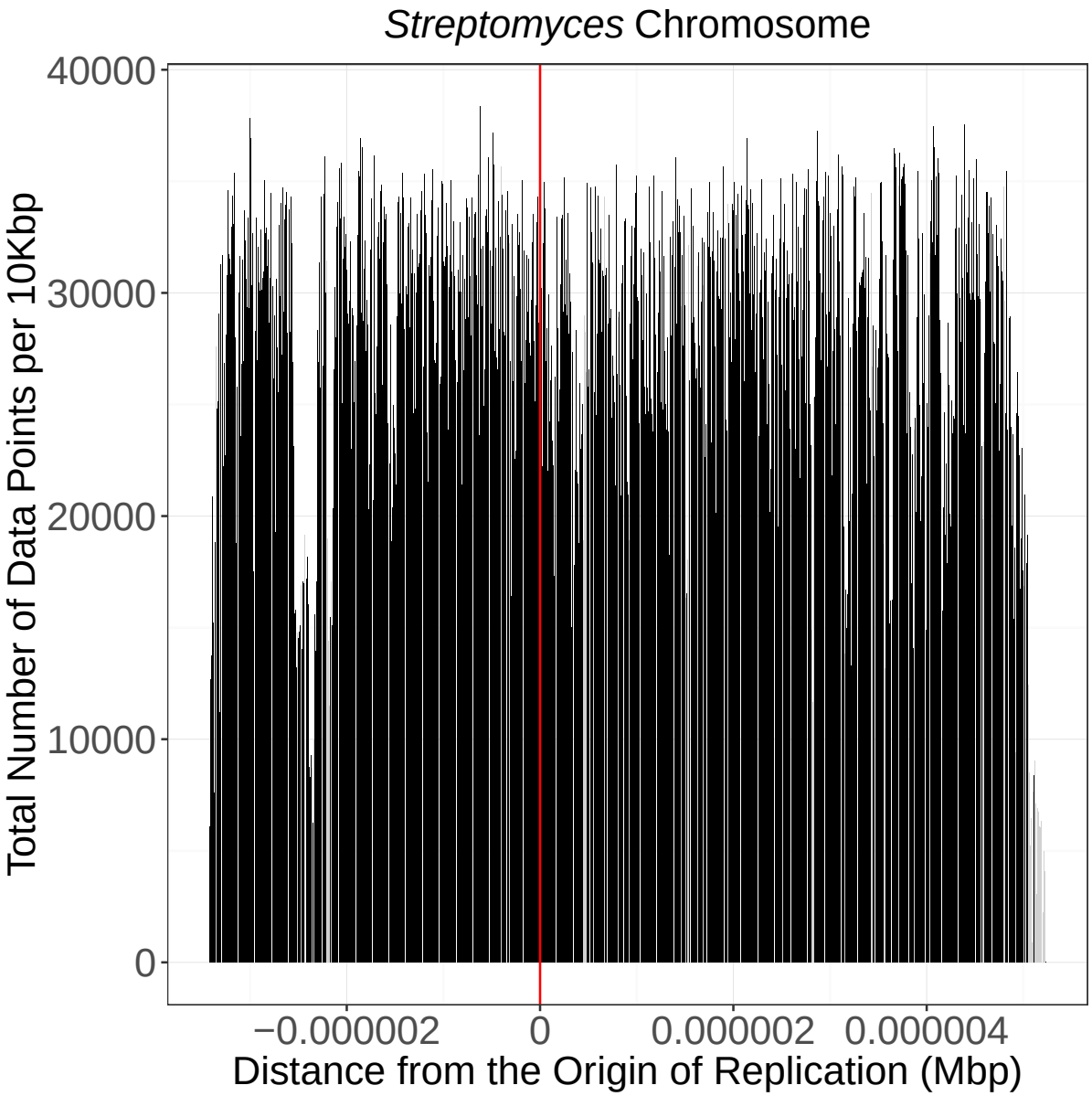



Figure 11: Distribution of total number of substitution data points per 10Kbp in genome.

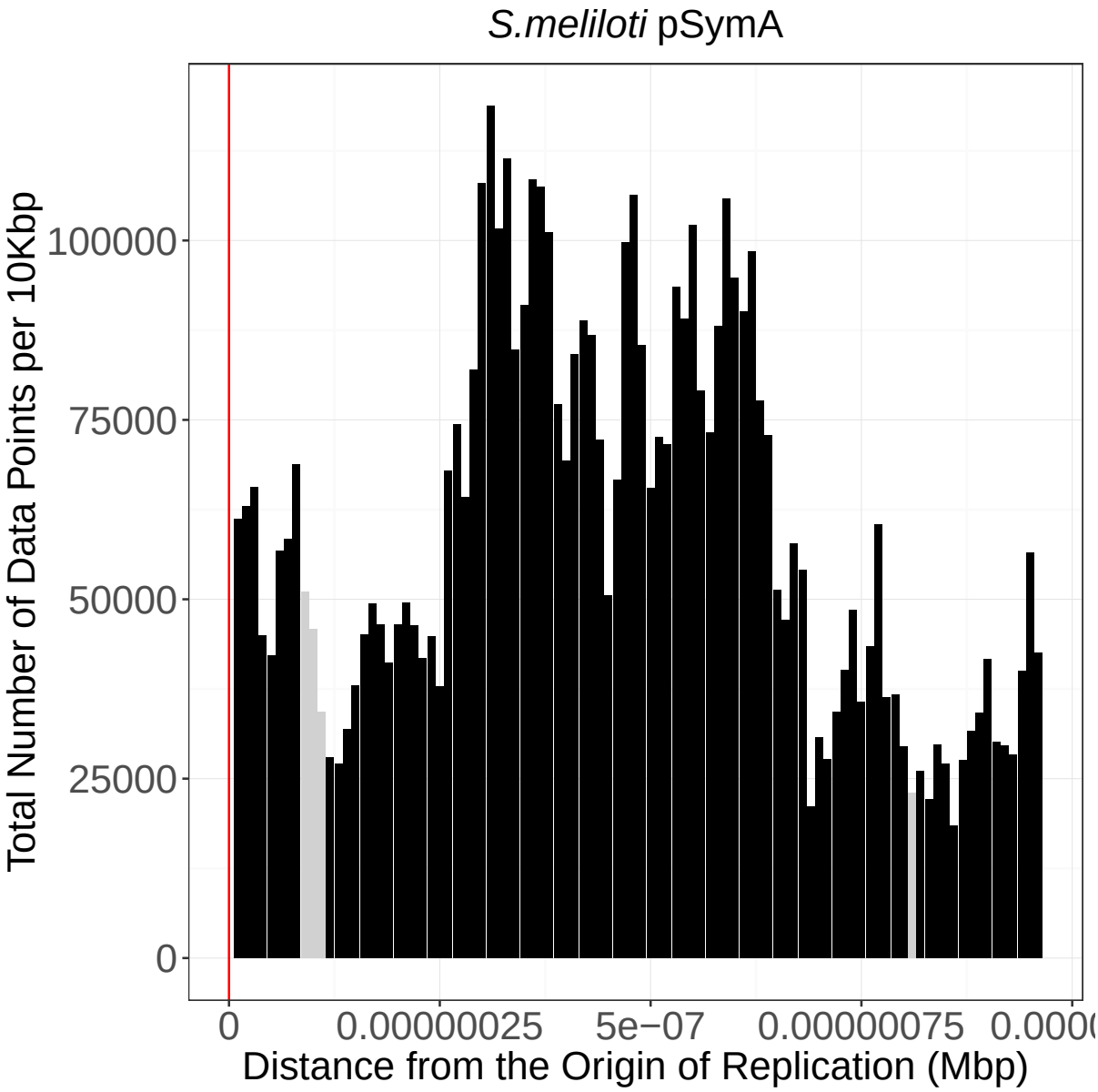



Figure 12: Distribution of total number of substitution data points per 10Kbp in genome.

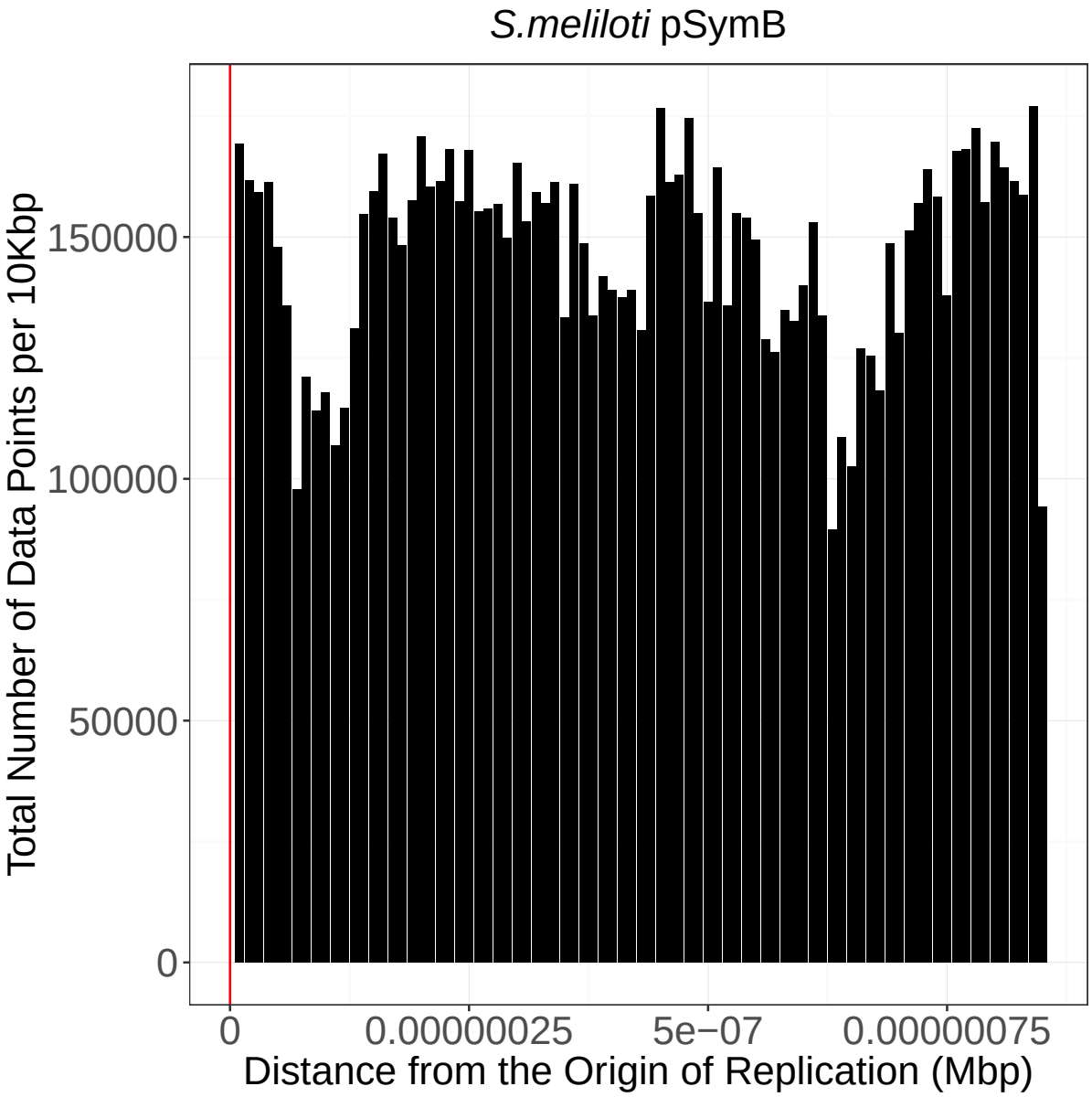



Figure 13: Distribution of total number of substitution data points per 10Kbp in genome.

| Bacteria and Replicon | Protein Coding Sequences |
|---|---|
| *E. coli* Chromosome | $-1.43\times10^{-8}$*** |
| *B. subtilis* Chromosome | $-5.55\times10^{-8}$*** |
| *Streptomyces* Chromosome | $7.49\times10^{-8}$*** |
| *S. meliloti* Chromosome | $-5.99\times10^{-7}$*** |
| *S. meliloti* pSymA | $-5.18\times10^{-7}$*** |
| *S. meliloti* pSymB | $1.67\times10^{-7}$*** |

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'.

| Bacteria and Replicon | Average Number of Substitutions per bp |
|---|---|
| *E. coli* Chromosome | $1.97\times10^{-4}$ |
| *B. subtilis* Chromosome | $1.93\times10^{-4}$ |
| *Streptomyces* Chromosome | $2.74\times10^{-6}$ |
| *S. meliloti* Chromosome | $9.72\times10^{-5}$ |
| *S. meliloti* pSymA | $6.54\times10^{-5}$ |
| *S. meliloti* pSymB | $1.99\times10^{-4}$ |

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.

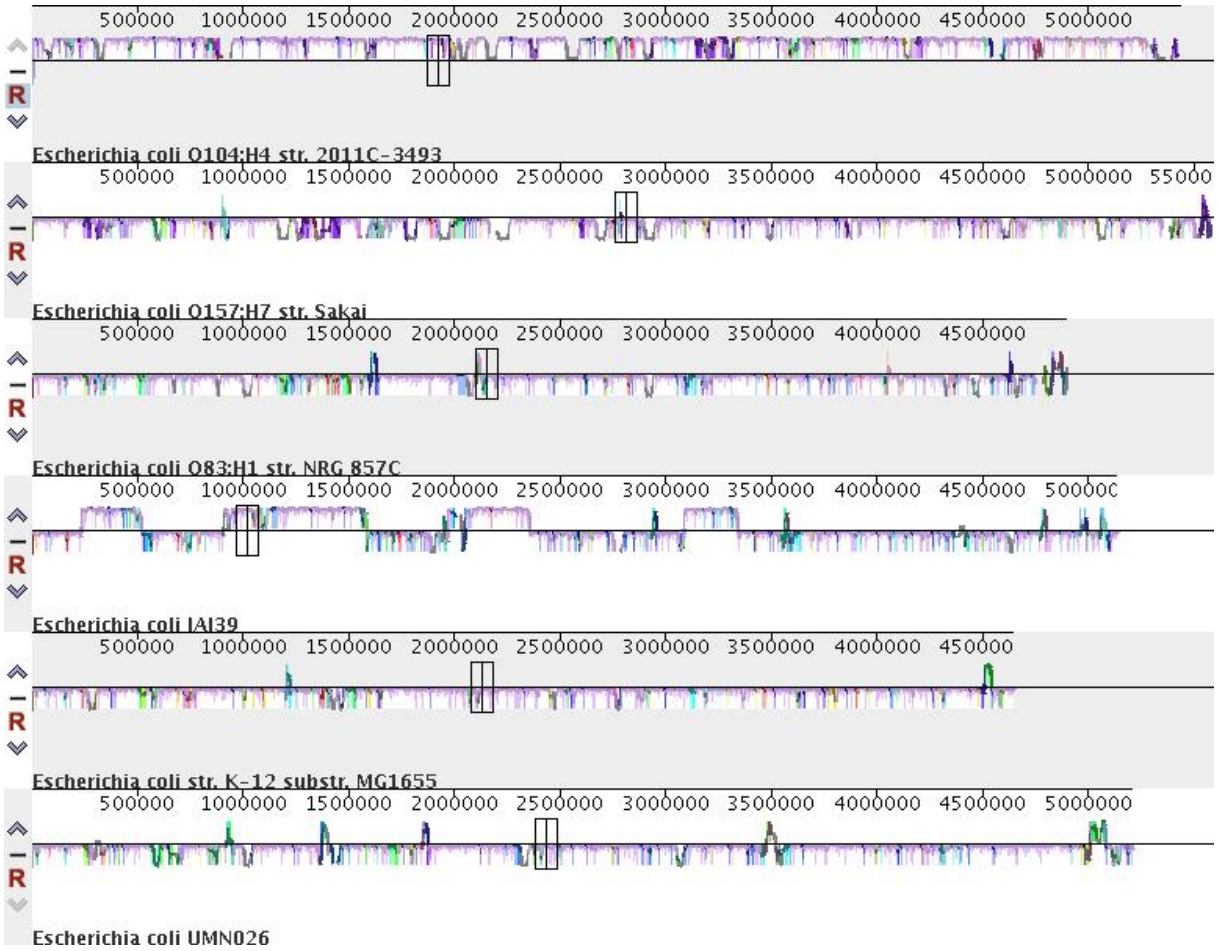



Figure 14: progressiveMauve alignment of *Escherichia coli* genomes highlighting the "backbone" of the alignment (matching regions).


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

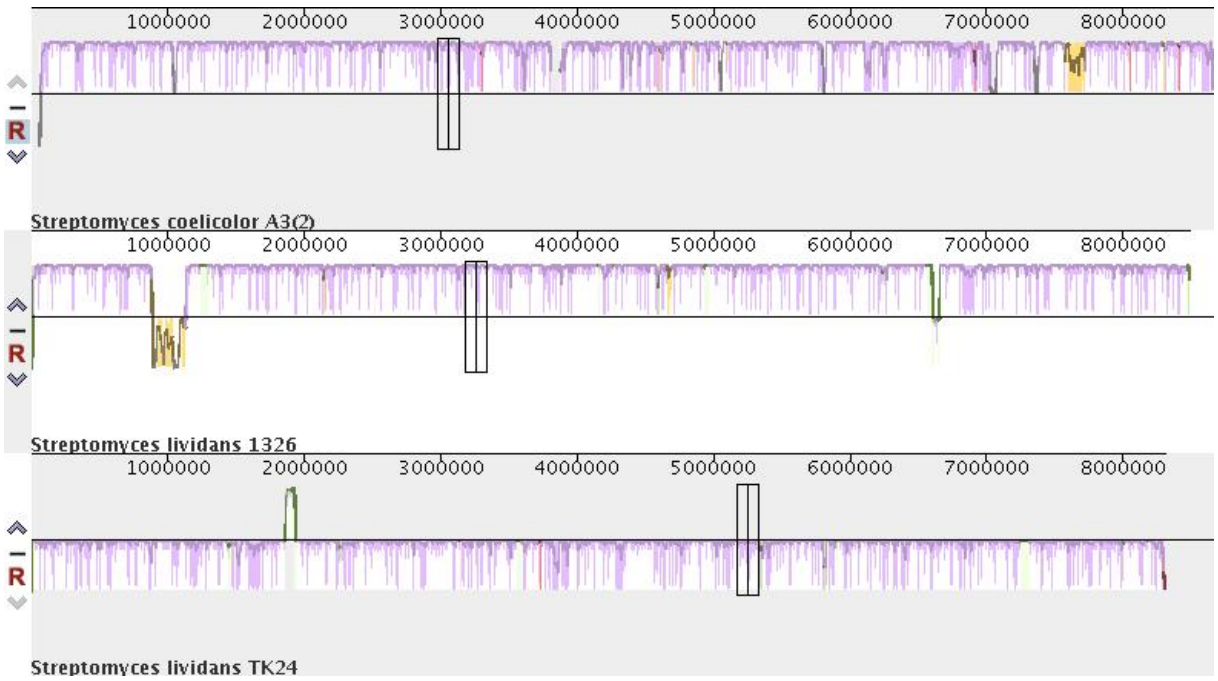



Figure 16: progressiveMauve alignment of *Streptomyces* genomes highlighting the "backbone" of the alignment (matching regions).