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?

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

✓look into what is up with the chromosome of *S. meliloti*

Edits to the Substitution Paper:

✓change selection and substitution figure captions

✓comment about 0.0001 lines in selection plots

✓add that high values of dS are real

✓edits to methods based on Brian's comments

✓finish personal edits to results section to send to Brian

✓edits to results based on Brian's edits

✓think about and write/edit future directions in discussion

✓think about and edit broad questions in intro and discussion

✓think about and write/edit weakness in discussion

Continuing to look into what is going on with the selection stuff in *S. meliloti* chromosome and basically the only thing I can see is that *S. meliloti* chromosome is really really similar between these taxa. I looked at the progressiveMauve alignment plots of *S. meliloti* replicons, and it basically tells the same story. the two plasmids have more variation in the backbone (which parts are similar between the taxa) than the chromosome.

I also looked at the misc_features in each of the bacterial replicons because these are NOT included as part of my analysis. *S. meliloti* as a whole (all replicons) has about 75-76% of its genes as misc_features, where as all the others are much less (*E. coli* = 0.5%, *B. subtilis* = 1.6% and *Streptomyces* = 46%). The two bacteria that have lower substitutions (*Streptomyces* and *S. meliloti*) are the ones with more misc_features, but since the secondary replicons of *S. meliloti* have about the same percentage of misc_features as the chromosome, I do not think that the misc_features are influencing why the chromosome of *S. meliloti* is so weird.

I looked at dN, dS, and ω values between the chromosome of *S. meliloti* and the secondary replicons (looking at sections of the alignment by hand) and again, it all is real. The massive amounts of zero values for dN and ω in the chromosome are real. So it again looks like the chromosome just has less variation. However, the issue is that because there are so many zero values for ω in the chromosome, anything where ω is $>$ zero is considered an outlier. Which makes all the calculations for linear regressions and such very off. But we can't really do anything about this because we would then have to change the way outliers are calculated for all the bacteria. Additionally, in the genes where ω is > 0 in the chromosome, there are still VERY few substitutions compared to genes where $\omega > 0$ in the other bacteria/replicons. So again, it looks like the chromosome of *S. meliloti* just has less variation.

I created graphs showing the number of sites per 10Kbp for the substitutions analysis to see if maybe the chromosome of *S. meliloti* was under represented (Figures 9 - 14). From these graphs, we can see that all the bacteria have some areas where there are an under representation of sites in different places in the genomes. So I do not think that lack of data is the cause for the weirdness in the chromosome of *S. meliloti*.

Even when looking at the average number of substitutions per genome, we see that the chromosome of *S. meliloti* has the fewest substitutions. *Streptomyces* is on the same order of magnitude, which to a degree makes sense because it is also very similar (based on the progressiveMauve alignment), but is also puzzling because *Streptomyces* appears to not have these issues.

I added *A. tumefaciens* as a taxa to the *S. meliloti* chromosome analysis to test if this changes the values of dN, dS, and ω to be similar in magnitude to the other bacteria (at your request), see Table 4. we see that dS is 3 orders of magnitude bigger than dS, but all the ω values are < 1 (even outliers), which I suppose sort of makes sense because these taxa are more distantly related than the taxa in the other bacterial analysis. **Should I instead use a different *Sinorhizobium* species to add to the analysis? What are your thoughts on all of this extra digging I have done regarding why the chromosome of *S. meliloti* looks so weird?**

I made lots of little edits and wrote more for the substitution paper (see above checklist).

The section below are the results from INCLUDING outliers (non-zero dN and ω values) in the *S. meliloti* chromosome selection analysis. The results are basically the same which is good! Because it means that even with these values, our point still stands that there is no evidence of a correlation between dN, dS and ω and distance from the origin of replication. **I would really appreciate it if you could read it over and let me know of any edits I should make.**

0.1 *S. meliloti* Chromosome Selection Analysis Without Outliers

Due to the extremely high sequence similarity of the *S. meliloti* chromosomes in this analysis, there are a relatively low number of substitutions and therefore many dN and ω values that are equal to zero (see Table **??**). The high number of zero values were included in the original calculation of outliers (see Main Paper for more details) causing all of the non-zero dN and ω values to be classified as outliers (see Figure 6 in the Main Paper). We decided to perform the same calculations on dN, dS, and ω but including the outliers to see what the results would have been. A visualization of the distribution of dN, dS, and ω along the chromosome of *S. meliloti* is seen in Figure 1. The average values for dN, dS, and ω are found in Table 1 and the linear regression to determine if there is a correlation between distance from the origin of replication and dN, dS, and ω values for the chromosome of *S. meliloti* is found in Table 2. We also looked at the values of dN, dS, and ω in the 20Kbp regions near and far from the origin of replication (including outliers) in the *S. meliloti* chromosome, these results are summarized in Table 3. The methods for these calculations are the same as in the Main Paper and in section **??**, however, outliers were not removed from these calculations.

The results in Tables 1 - 3 closely reflect the results of the *S. meliloti* analysis when outliers were included (see Main Paper). The significant correlation between dS and distance from the origin

of replication is small and the lack of significant correlation between dN and ω and distance from the origin of replication suggest that the results are inconclusive. Even when the outliers (non-zero values of dN and ω) are included in the selection analysis, we still can not conclude that there is an overall trend between distance from the origin of replication and dN, dS, and ω values.

	Genome Average		
Bacteria and Replicon	dS	dN	ω
S. meliloti Chromosome	0.0100	0.0007	0.0677

Table 1: Weighted averages of dN, dS, and ω values calculated for *S. meliloti* chromosome using the gene length as the weight. Arithmetic mean was calculated for the per gene averages for *S. meliloti* chromosome. Outliers were included in the calculation.

Bacteria and Replicon	dN	dS	ω
S. meliloti Chromosome	NS	-2.29×10^{-9}*	NS

Table 2: Linear regression to determine the correlations between dN, dS, and ω values and distance from the origin of replication. Outliers were included in the calculation. All results are marked with significance codes as followed: p: < 0.001 = '***', $0.001 < 0.01$ = '**', $0.01 < 0.05$ = '*', > 0.05 = 'NS'.

	Near Origin			Near Terminus		
Bacteria and Replicon	dN	dS	ω	dN	dS	ω
S. meliloti Chromosome	NS	NS	NS	NS	NS	NS

Table 3: Linear regression for dN, dS, and ω calculated for the 20 genes closest and 20 genes farthest from the origin of replication in the *S. meliloti* chromosome. Outliers were included in the calculation. All results are marked with significance codes as followed: p: < 0.001 = '***', $0.001 < 0.01$ = '**', $0.01 < 0.05$ = '*', > 0.05 = 'NS'.

This Week

- incorperate any feedback Brian has about why *S. meliloti* chromosome looks so weird
- discuss selection linear regression results
- alter main focus of paper in discussion
- interpret all other results
- edit discussion for flow
- send Brian discussion to edit

Next Week

- do Brian's edits on discussion
- alter main focus of paper in conclusion
- edit/work on conclusion
- send Brian conclusion to edit
- edit/read through whole paper (in prep to send to Brian)

	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 4: 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 5: 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	1.97×10^{-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 6: Average number of protein coding substitutions calculated per base across all bacterial replicons. Outliers and missing data was not included in the calculation.

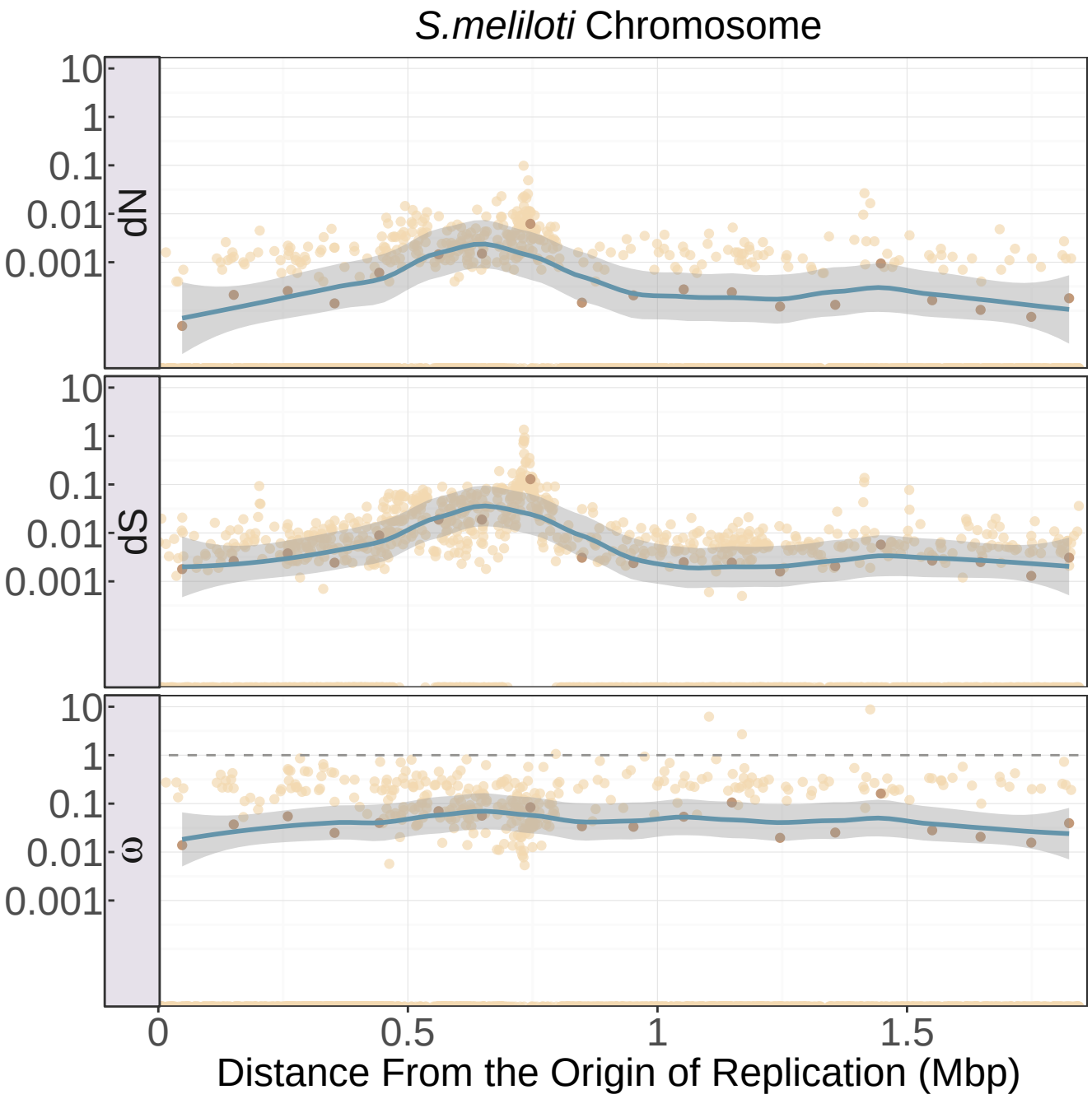


Figure 1: The graph show the values of dN, dS, and ω along the chromosome of *S. meliloti*. Distance from the origin of replication is on the x-axis beginning with the origin of replication denoted by position zero on the left, and the terminus indicated on the far right. The y-axis of the graph indicates the value of dN, dS, and ω found at each gene segment position of the chromosome. Outliers are included in this graph. The average dN, dS, and ω values for each 10,000bp regions of the replicon were calculated and represented by the dark brown points. A trend line represented in blue (using the `loess` method), was fit to these average values and the associated 95% confidence intervals for this line is represented by the grey ribbon around the blue trend line. For a complete list zero value information, please see Supplementary Material.

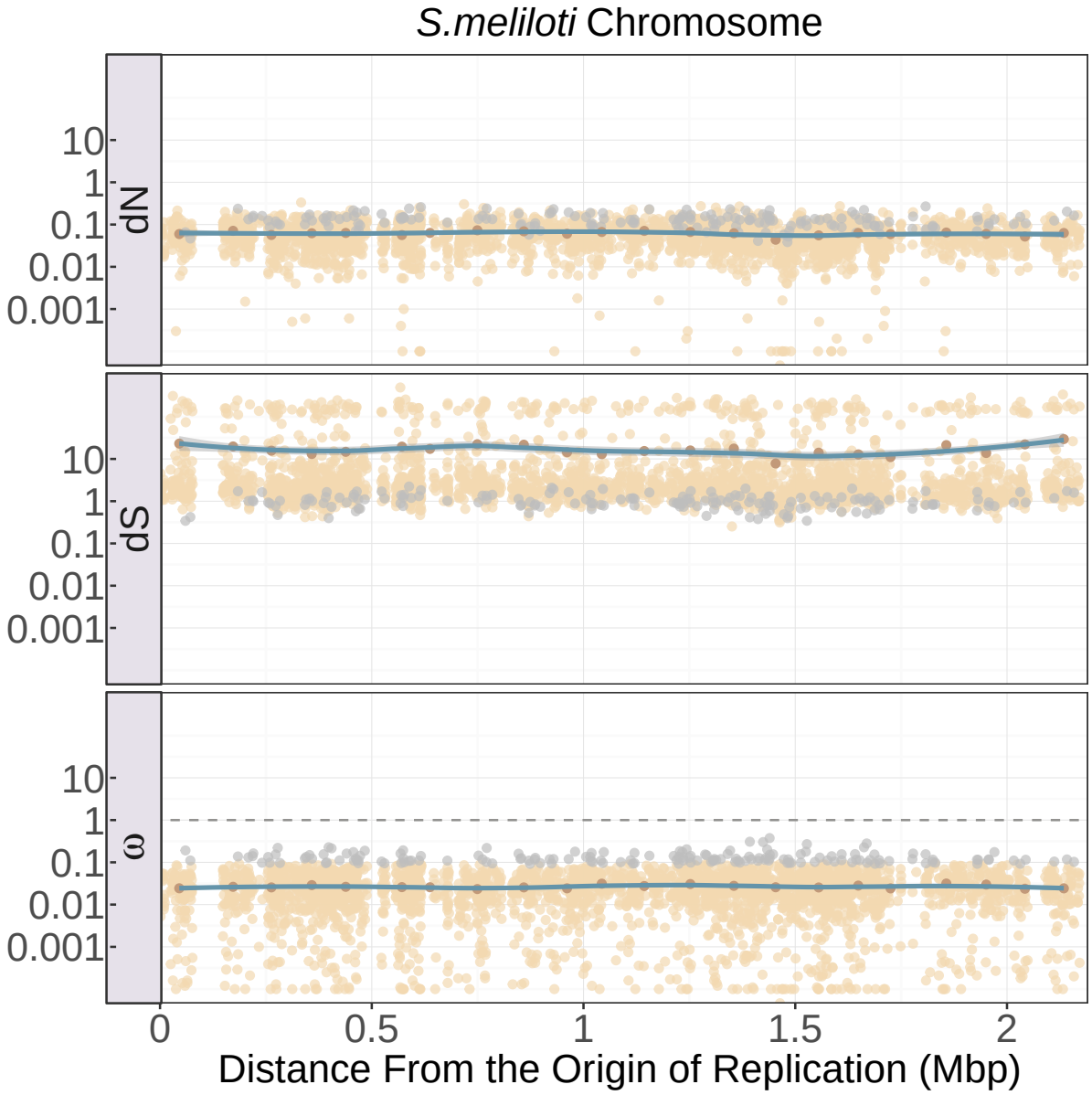
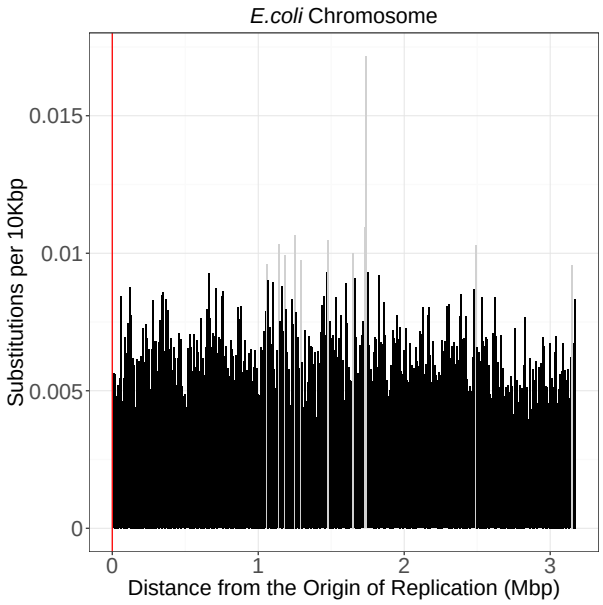
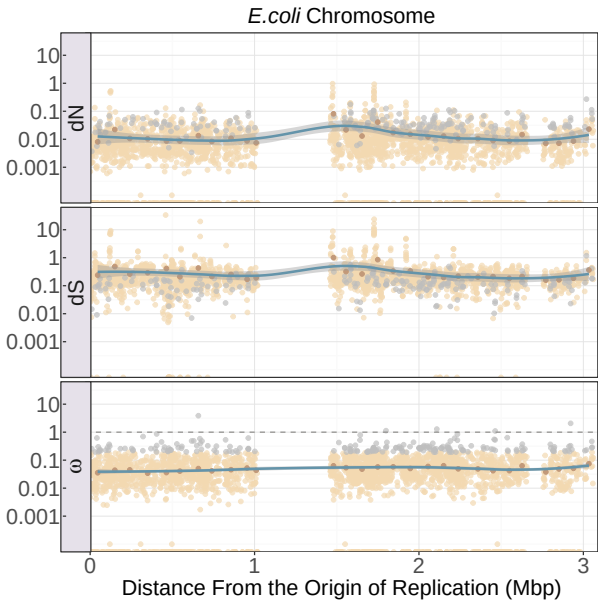


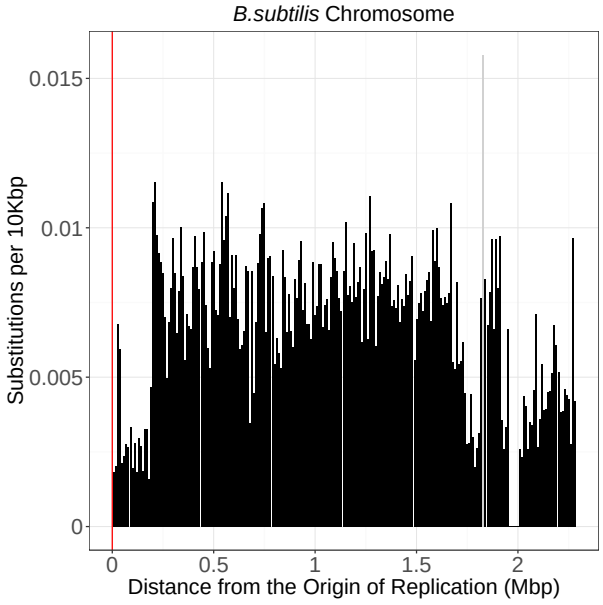
Figure 2: dN, dS, and ω values for *S. meliloti* chromosomes and ***A. tumefaciens***.


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

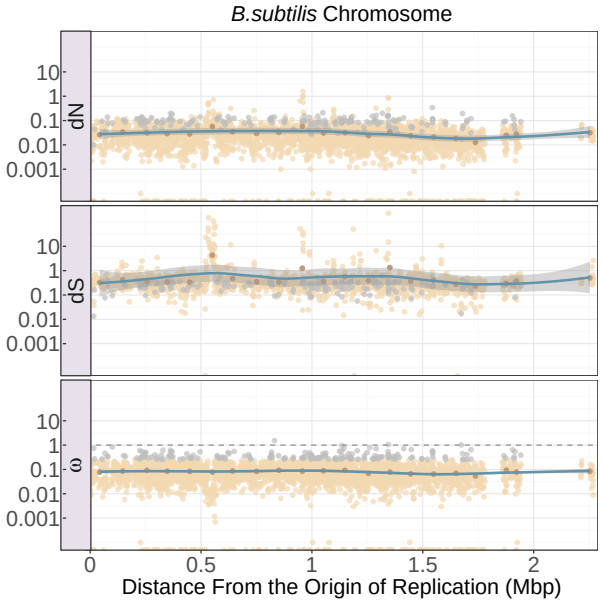
(a)


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)

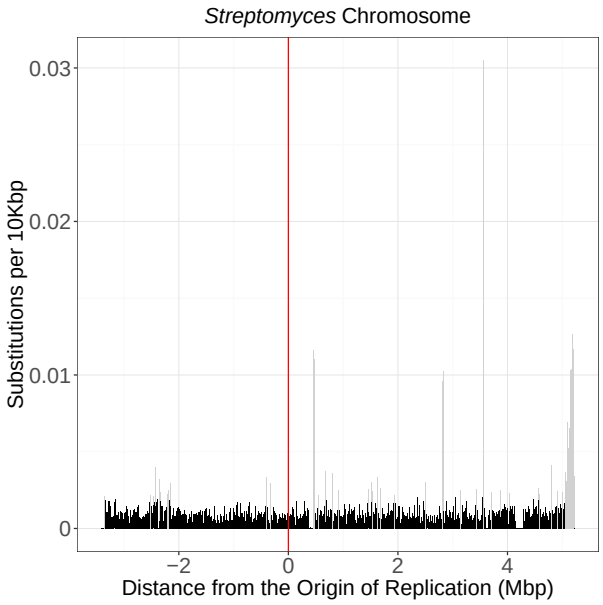
(b)


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)

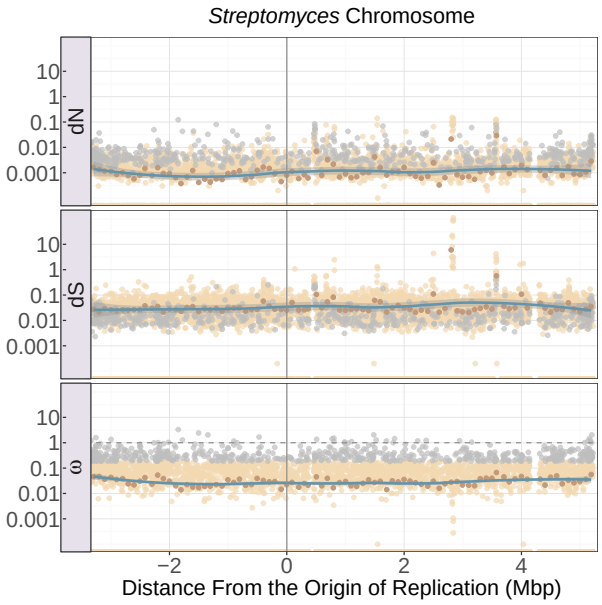
(a)


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)

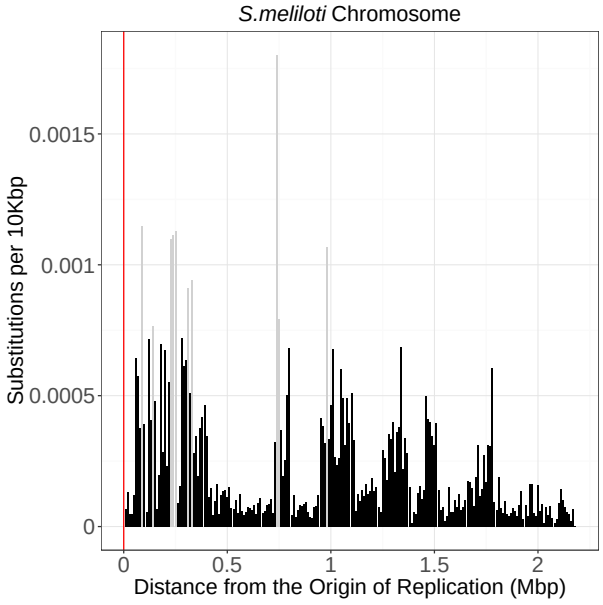
(b)


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

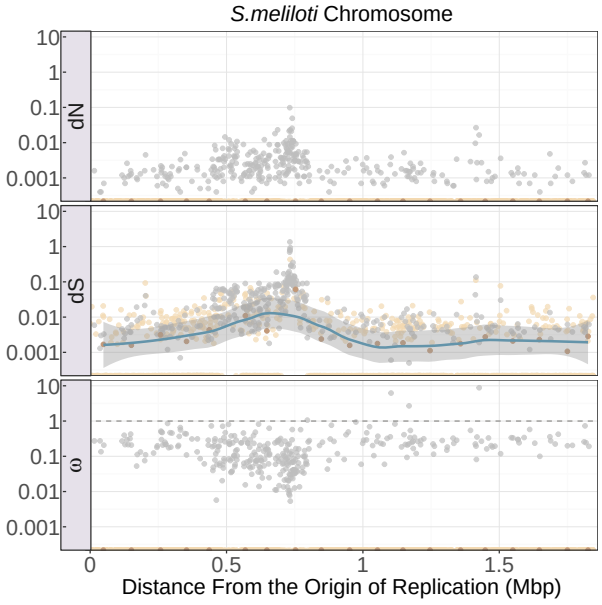
(a)


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

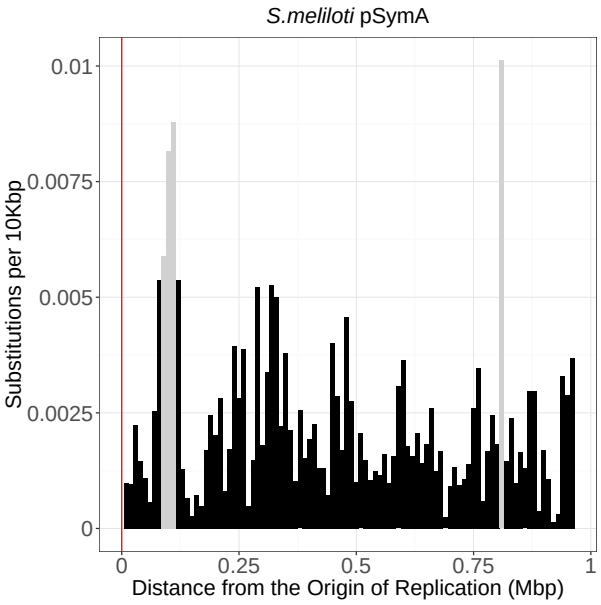
(b)


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)

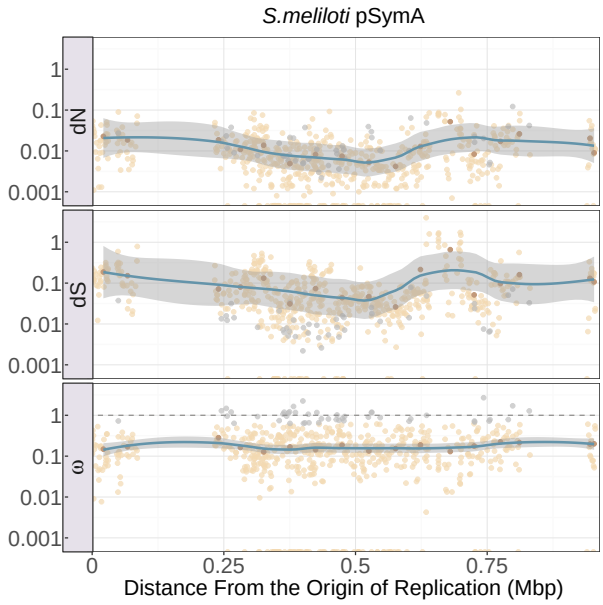
(a)


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)

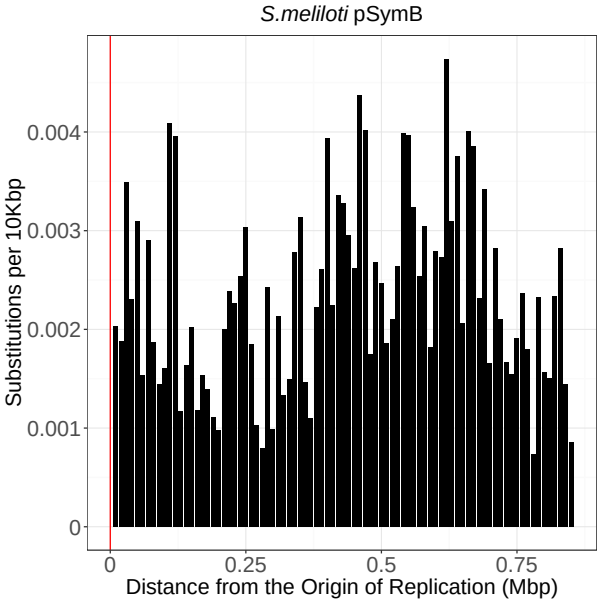
(b)


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)

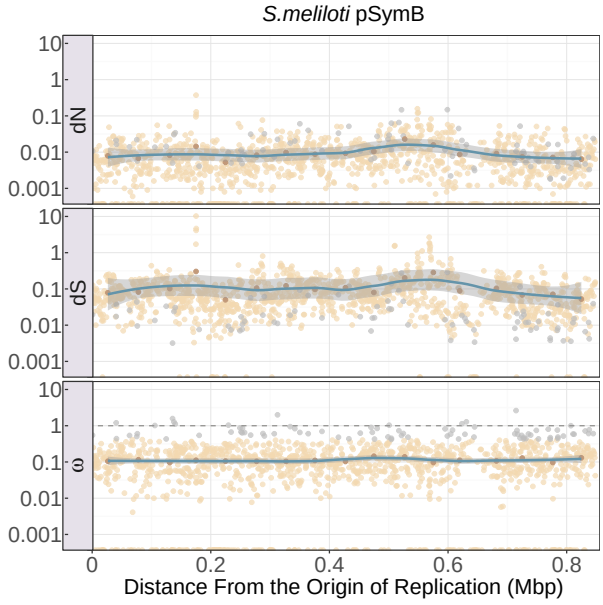
(a)


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)

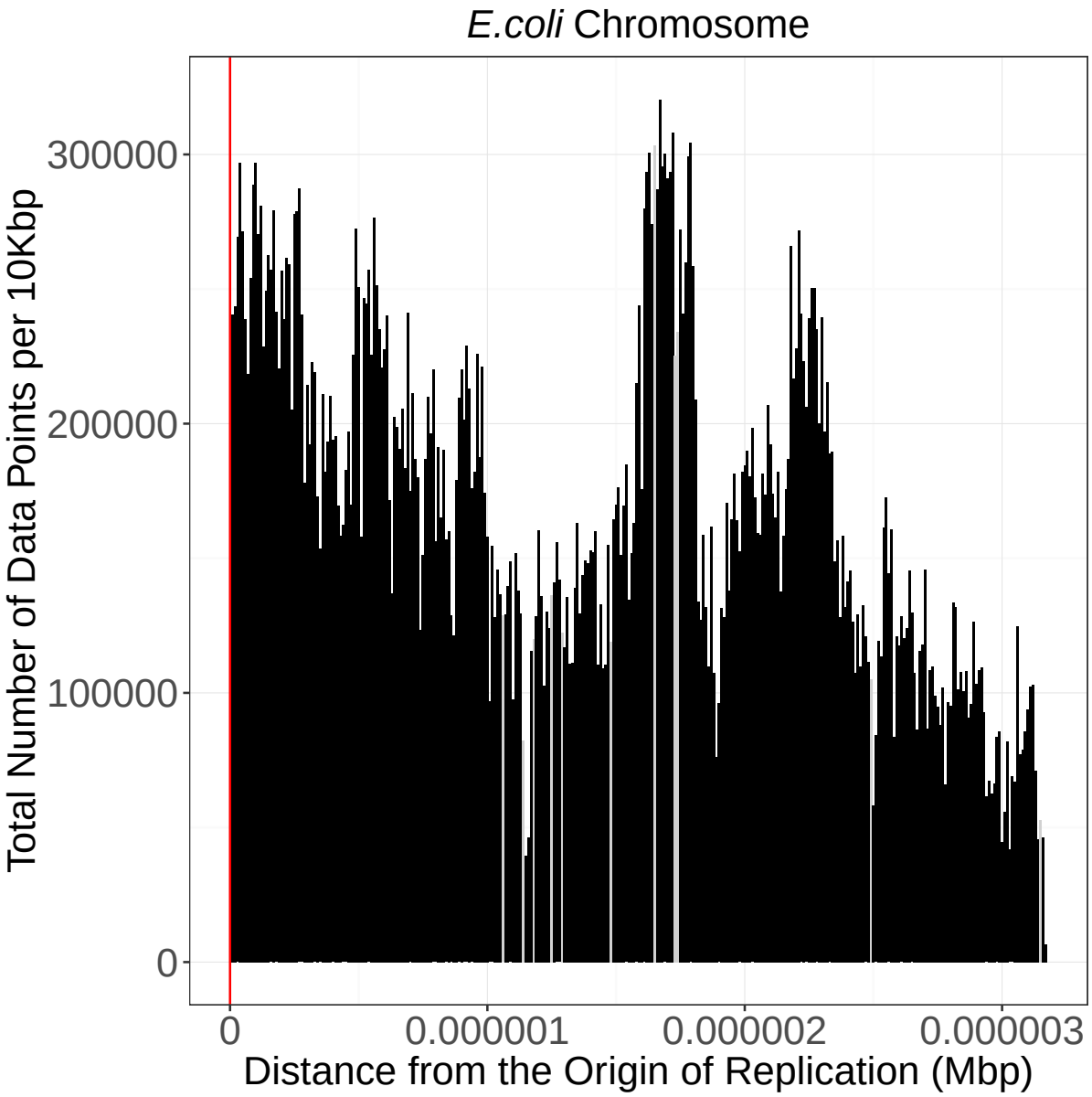


Figure 9: Distribution of total number of substitution data points per 10Kbp in genome.

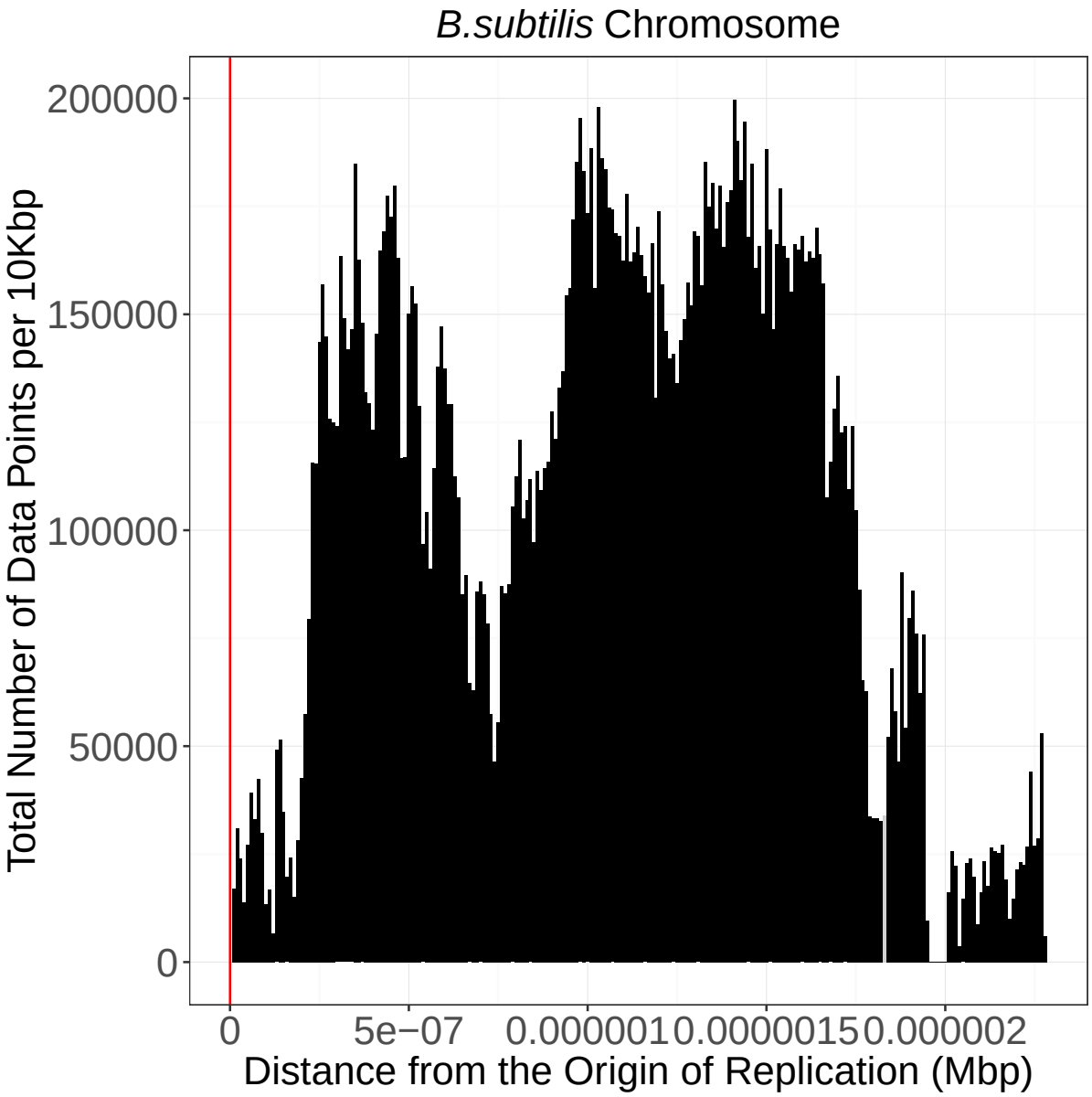


Figure 10: Distribution of total number of substitution data points per 10Kbp in genome.

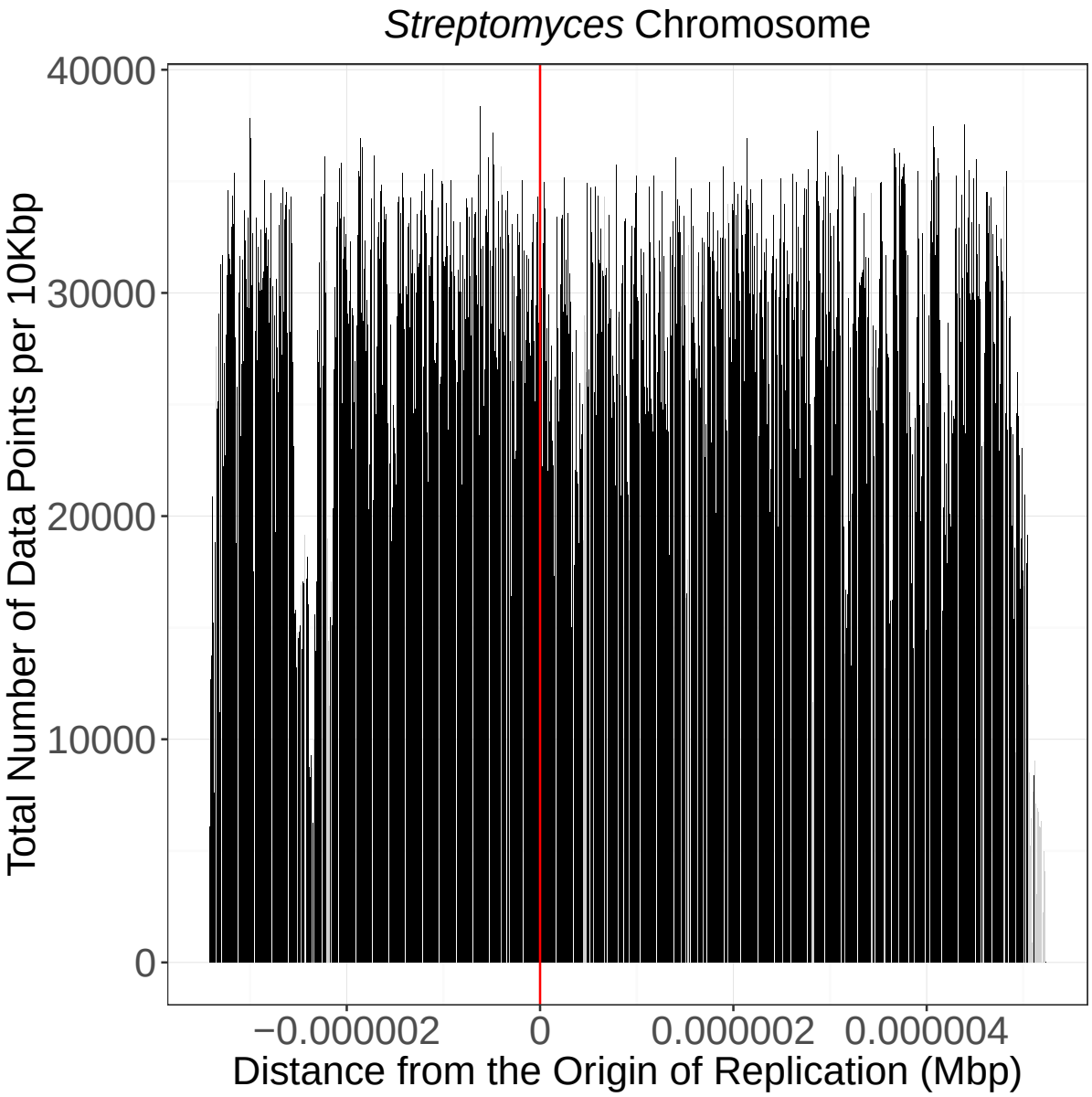


Figure 11: Distribution of total number of substitution data points per 10Kbp in genome.

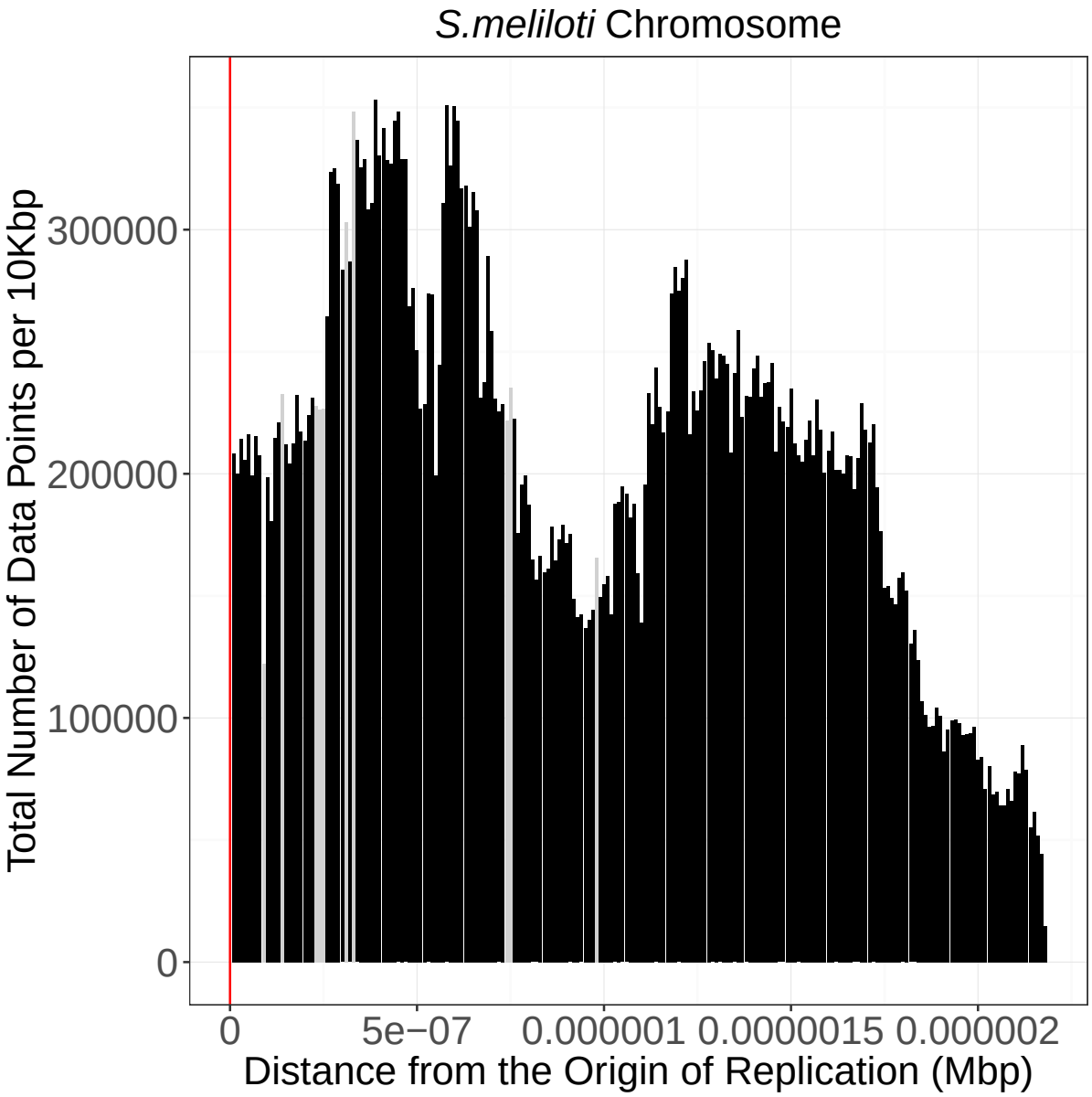


Figure 12: Distribution of total number of substitution data points per 10Kbp in genome.

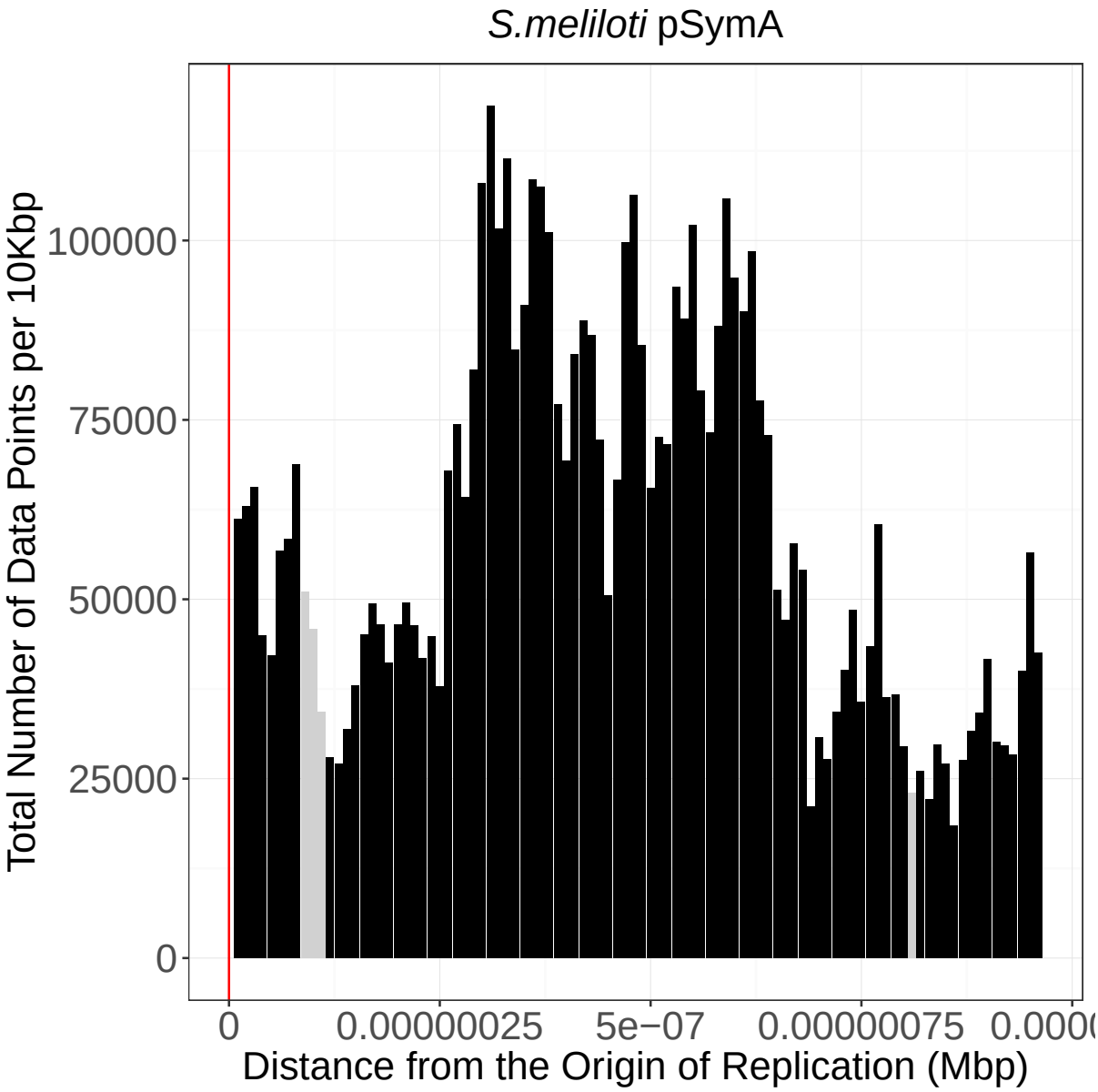


Figure 13: Distribution of total number of substitution data points per 10Kbp in genome.

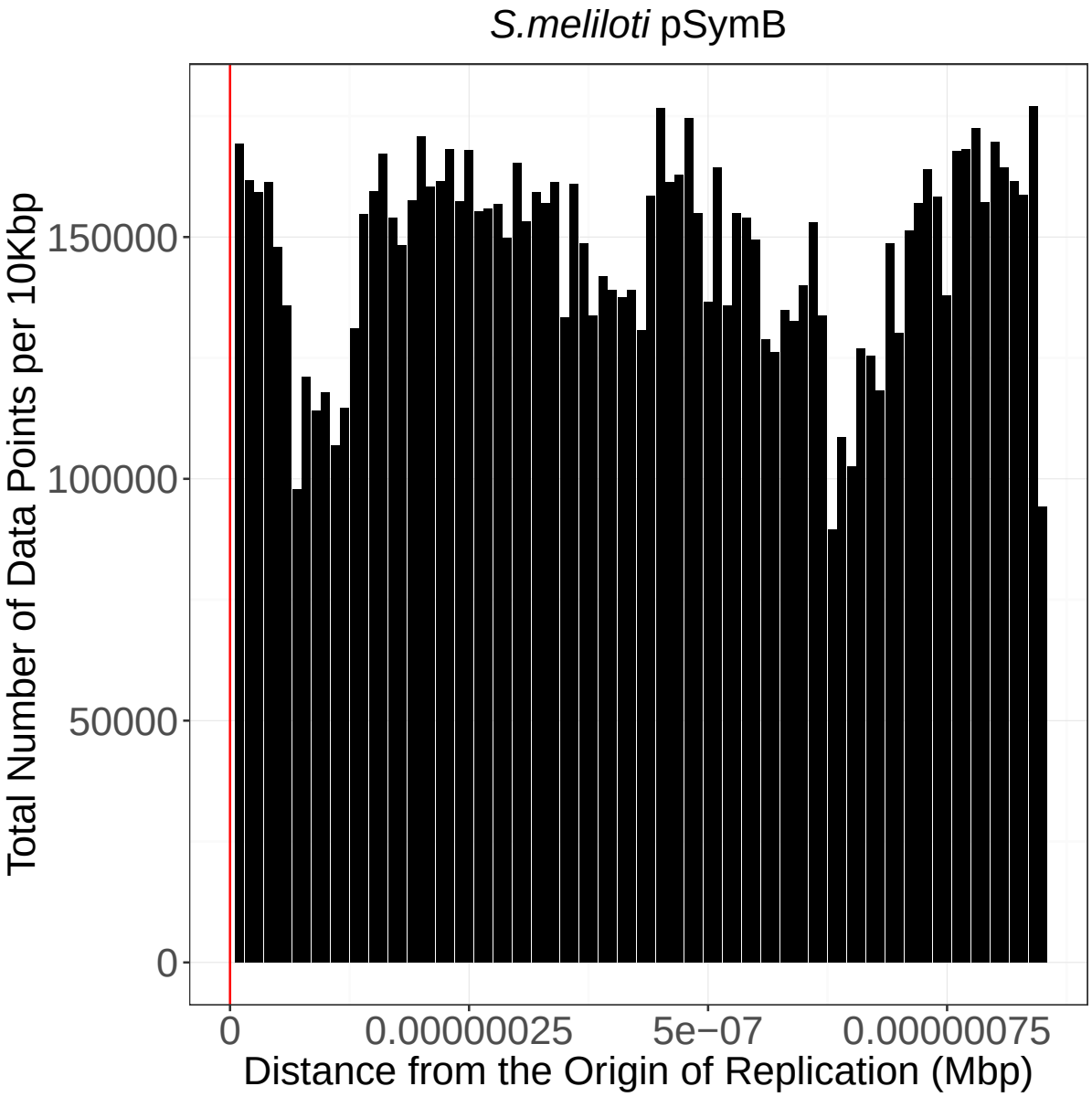


Figure 14: Distribution of total number of substitution data points per 10Kbp in genome.

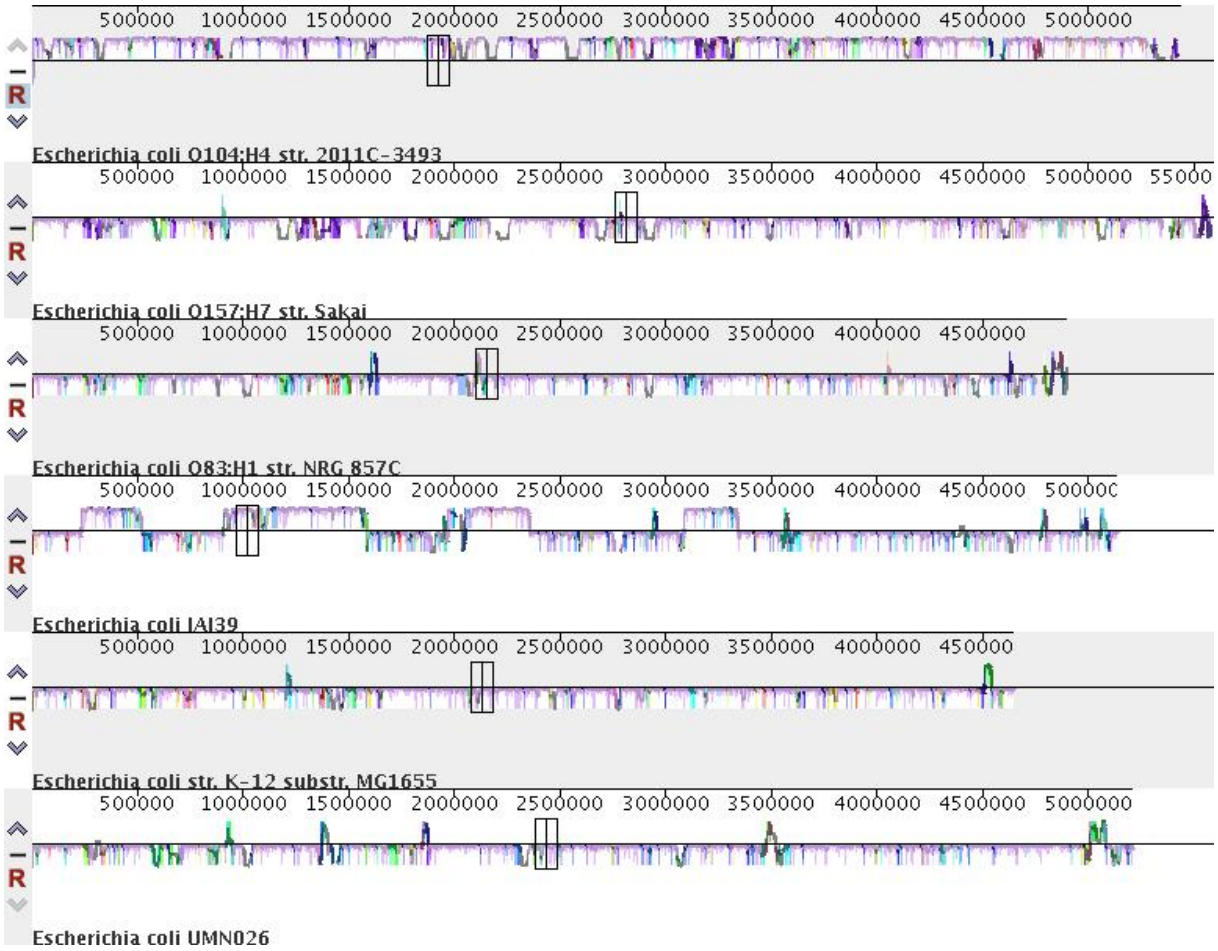


Figure 15: progressiveMauve alignment of *Escherichia coli* genomes highlighting the "backbone" of the alignment (matching regions).

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

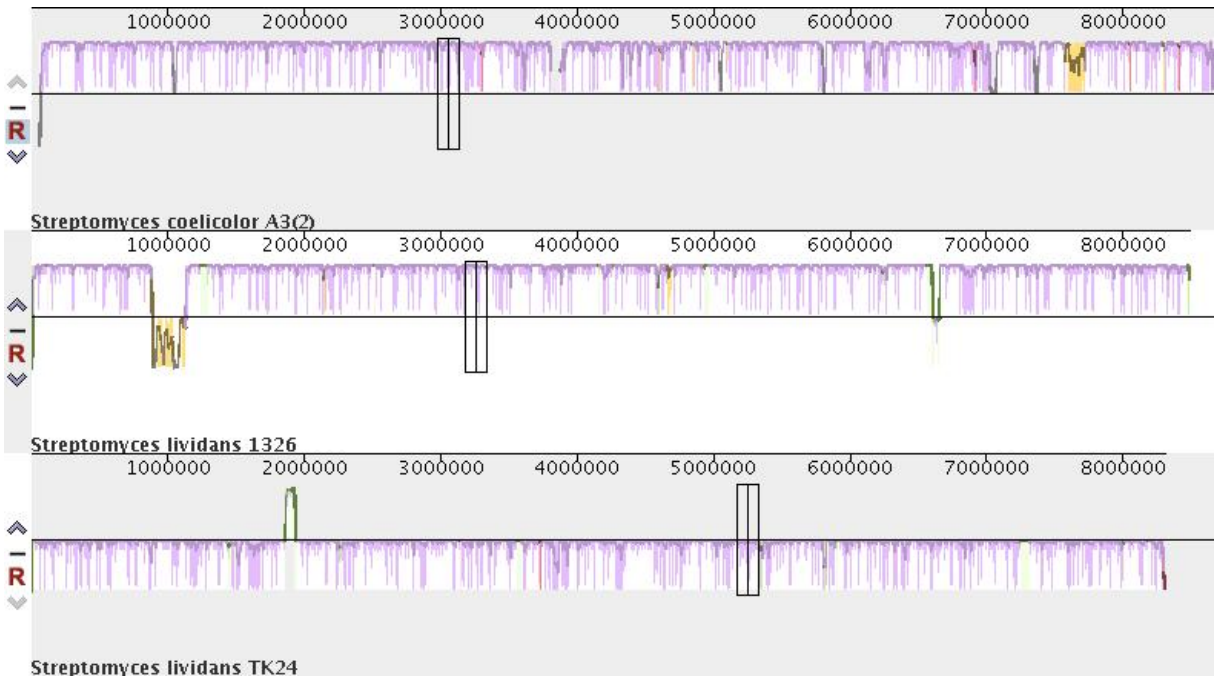


Figure 17: progressiveMauve alignment of *Streptomyces* genomes highlighting the "backbone" of the alignment (matching regions).