Subs Paper Things to Do:

- ~~why does sinoC have omega lin reg = 0 near and far from the origin?~~
- create new graphs for selection analysis
- ~~find and example of high substitution bar in *Streptomyces* and put this into supplement as an example of really diverged taxa (and that subs are real!)~~
- ~~discuss removing omega outliers in methods~~
- ~~double check that the ter and ori and max genome pos are correct~~
- ~~make graphs proportional to length of respective cod/non-cod regions~~
- ~~test examples for genes near and far from terminus (robust log reg/results)~~
- ~~linear regression on 10kb regions for weighted and non-weighted substitutions~~
- ~~average number of substitutions in 20kb regions near and far from the origin~~
- ~~figure out why the data is weird for number of cod/non-cod sites~~
- why are the lin reg of dN, dS and ω NS but the subs graphs are...explain!
- ~~grey out outliers in subs graphs~~?
- 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?
- ~~make corrections based on Brian's edits~~
- ~~create a clean copy of the paper (no strikeout) for re-submission~~

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

✓make an R theme for the graphs so they all look the same

✓re-run the necessary substitution and selection analysis

✓made some edits to the ancestral reconstruction figure in the paper to make it more apparent that it extends for the whole sequence (and not just 9 nucleotides)

✓Look into why *S. meliloti* Chromosome looks odd

✓supplementary example of non-homologous alignment

I created a new theme for the selection and substitution graphs so that they all look relatively the same (similar margins, font size..etc). Last week when I was re-doing the SH-Test (to see which block trees were different from the overall tree), I realized that some of these blocks were not removed from the analysis in *E. coli*, *S. meliloti* Chromosome and pSymA. I spent most of this week re-running these analysis with the correct number of blocks. The new figures and results can be found below and in the attached Supplementary File for the paper. Nothing has changed significantly. *S. meliloti* chromosome looks really terrible. All non-outlier points for dN and ω are zero values. Therefore a regression is pointless and makes the selection graph look really odd. I started to look into this but it looks like the *S. meliloti* chromosomes are just so similar that there are hardly any substitutions. This is particularly evident when you look at the backbone of the progressiveMauve alignment (which shows similar sequences). When looking at the *Escherichia coli* alignement (Figure 1) we see that the backbone is very "spiky" indicating regions where the

nucleotides are not similar. The same can be said for the *Streptomyces* genomes (Figure 3), even though these are more similar to each other than the *Escherichia coli* genomes. When we look at the *S. meliloti* Chromosome alignment (Figure 2), we see that the backbone is almost completely flat, meaning that there is hardly any variation in the nucleotide sequence. This is especially curious because there appears to be no different in the average number of substitutions in the *S. meliloti* chromosome (Table 2). It could be that all the variation is being considered an "outlier" because most of the values are zero (because they are so smilar)? **I am really confused about why the selection graph for *S. meliloti* Chromosome looks so odd and the only explination 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?**

I found an example of the `MAFFT` alignment where non-homologous genes were aligned to put in the supplement as an illustrative example.

This Week

- find a block where mauve aligns non-homologous regions and put into supplement
- re-do selection and substitutions analysis (if necessary, because of not throwing out blocks with different trees than the overall tree)
- re-do substitutions and selection graphs (with new theme) and these and put in paper

Next Week

- look into whats up with *S. meliloti* chrom bc it does not look right at all
- update methods in paper draft (make sure they are the same as what I actually did bc things have changed a lot)
- add reviewers comments from gene expression paper into this one (most will apply)

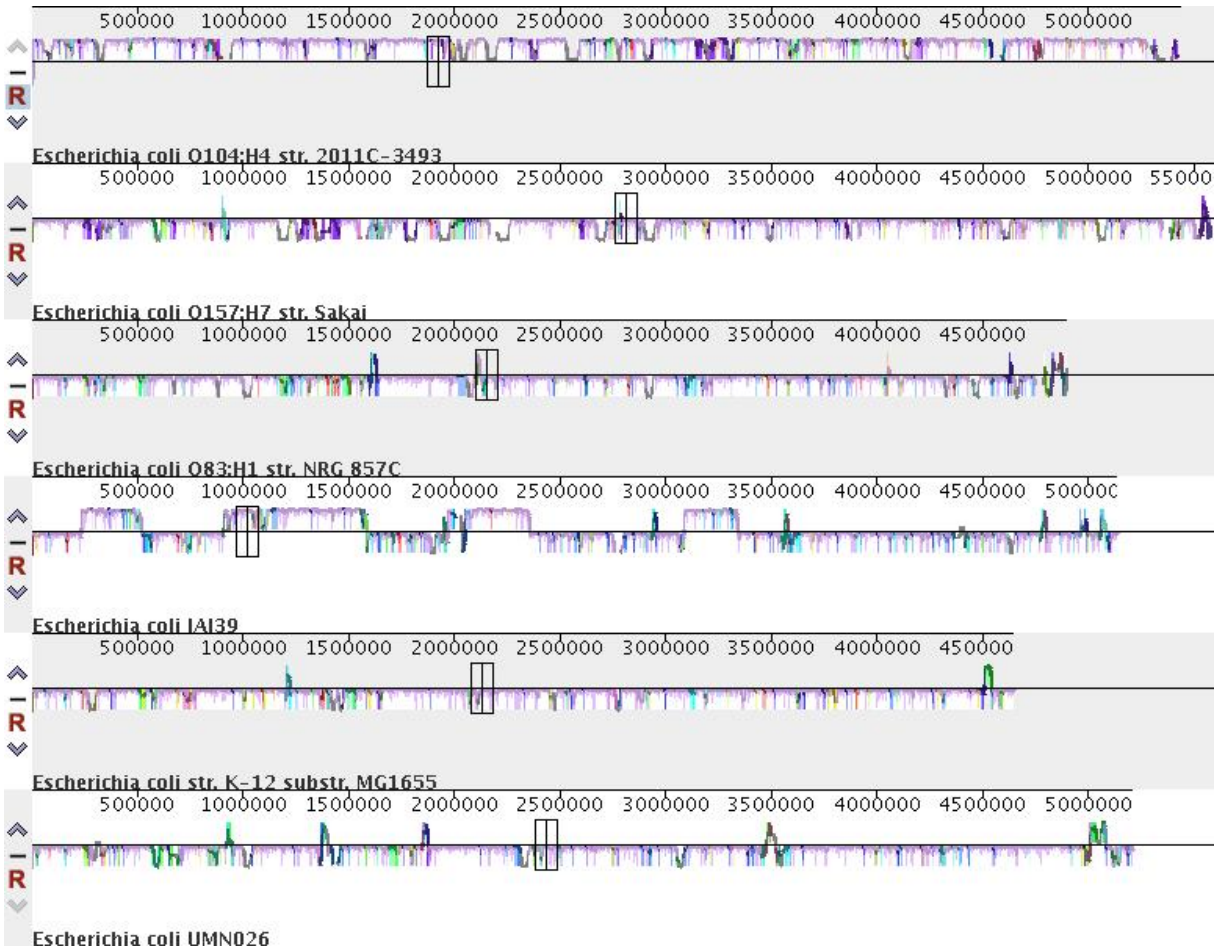


Figure 1: progressiveMauve alignment of *Escherichia coli* genomes highlighting the "backbone" of the alignment (matching regions).

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 1: 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 2: 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 2: Average number of protein coding substitutions calculated per base across all bacterial replicons. Outliers and missing data was not included in the calculation.

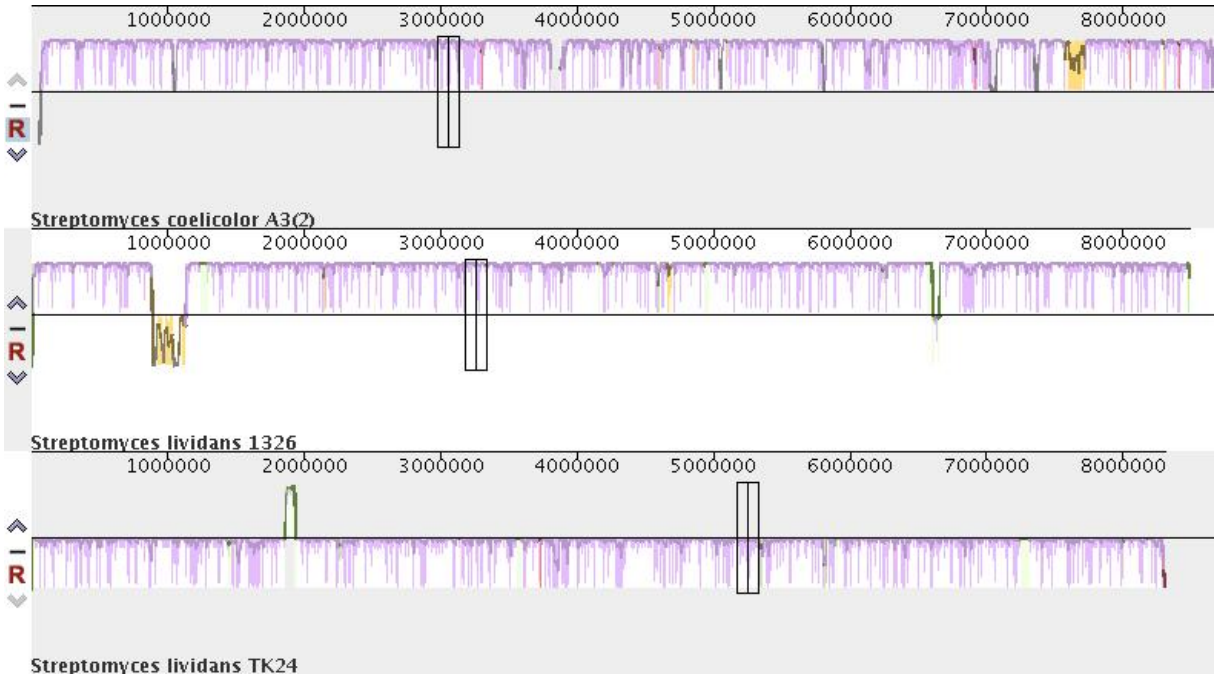
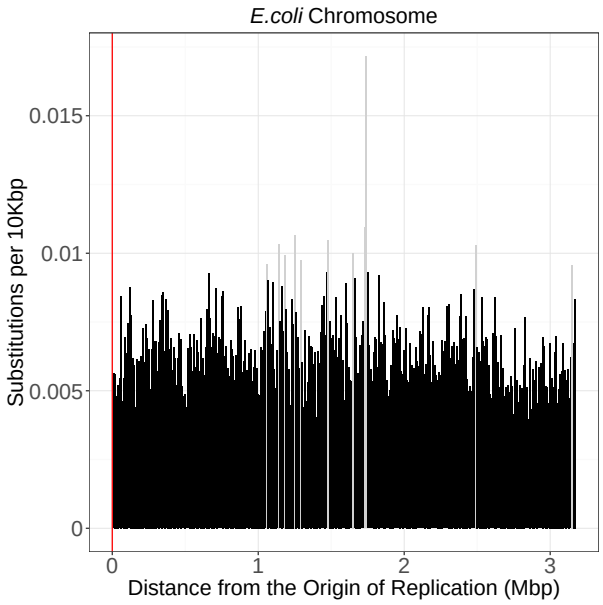


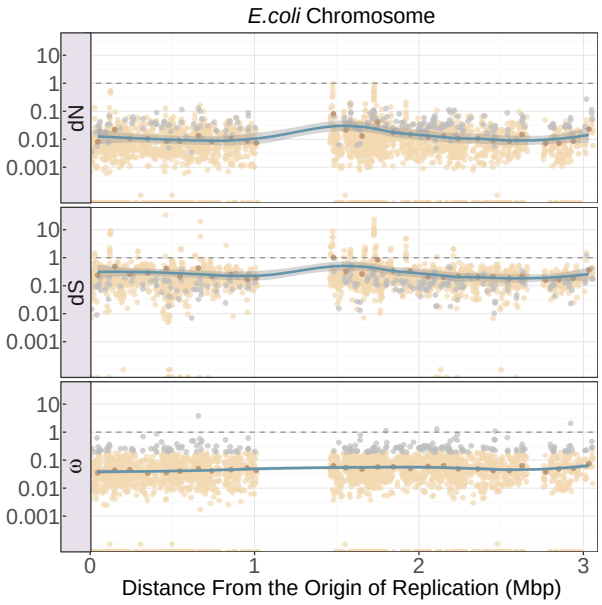
Figure 3: progressiveMauve alignment of *Streptomyces* genomes highlighting the "backbone" of the alignment (matching regions).

	Genome Average		
Bacteria and Replicon	dS	dN	ω
E. coli Chromosome	0.2387	0.0101	0.0441
B. subtilis Chromosome	0.4201	0.0243	0.0714
Streptomyces Chromosome	0.0458	0.0011	0335
S. meliloti Chromosome	0.0029	0	0
S. meliloti pSymA	0.0874	0.0099	0.1645
S. meliloti pSymB	0.0940	0.0084	0.1142

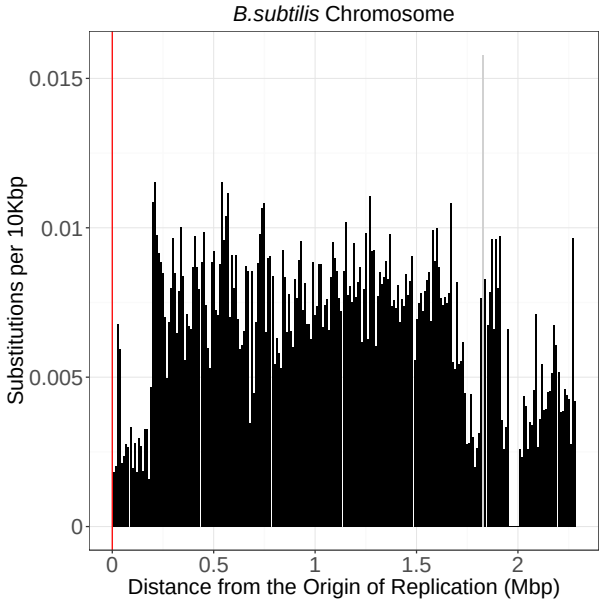
Table 3: Weighted averages calculated for each bacterial replicon on a per genome basis using the gene length as the weight. Arithmetic mean calculated for the per gene averages for each bacterial replicon.


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

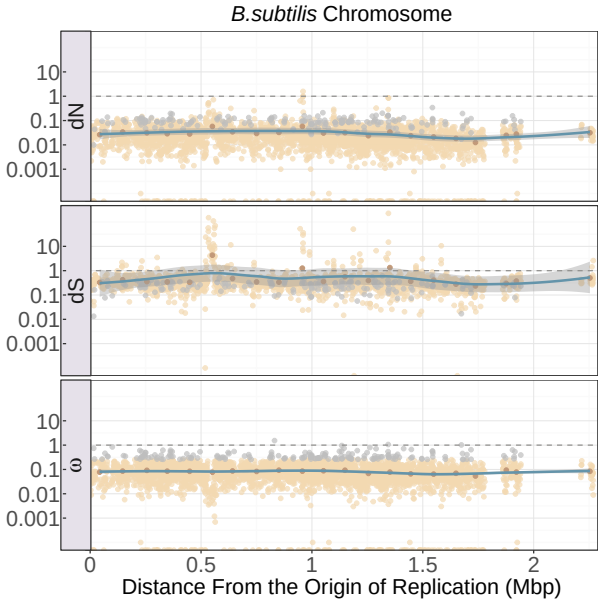
(a)


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)

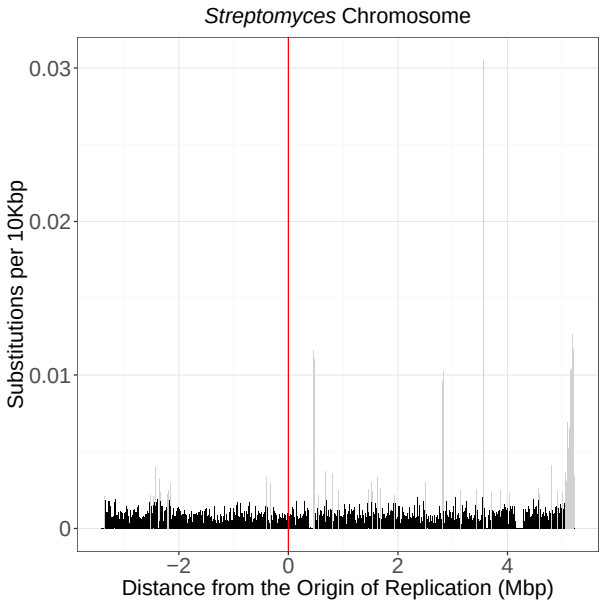
(b)


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)

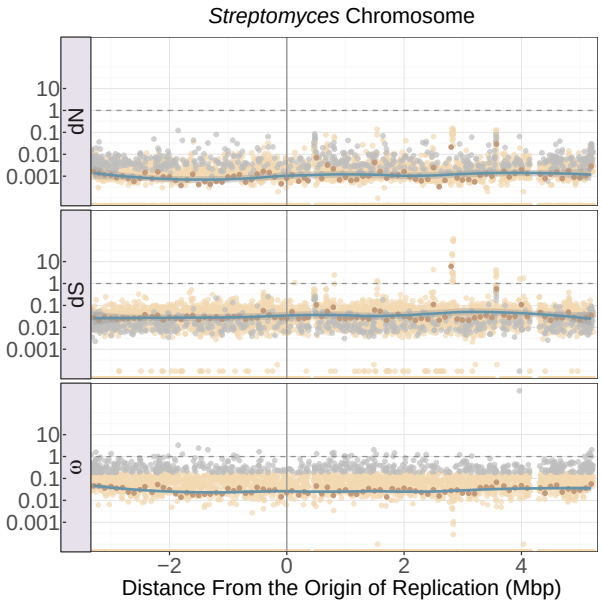
(a)


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)

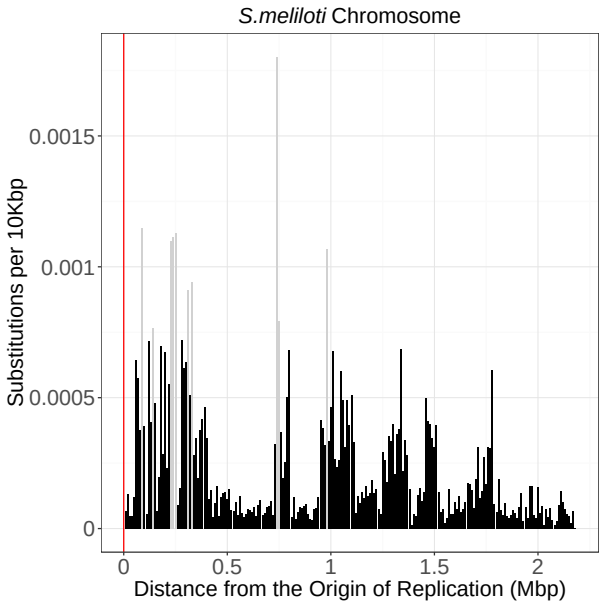
(b)


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

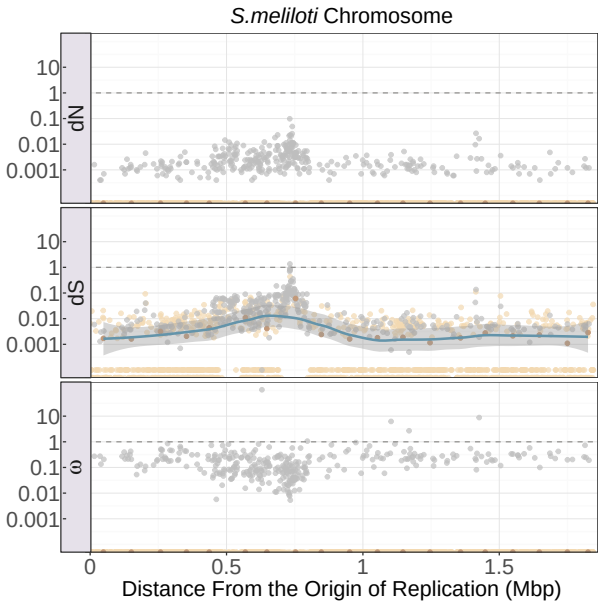
(a)


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

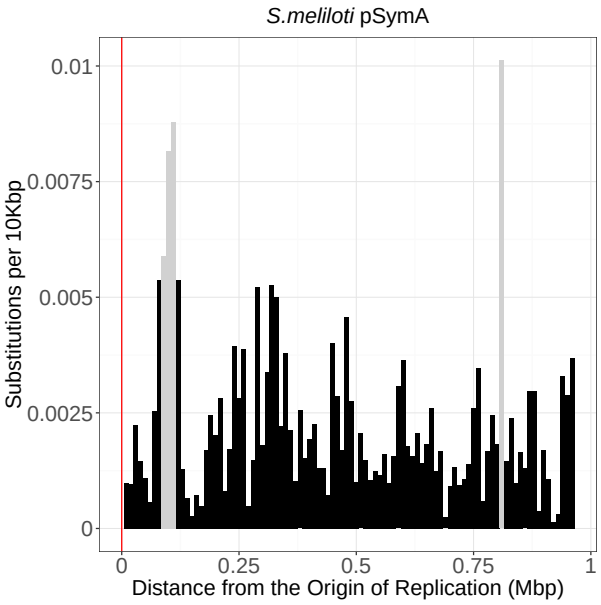
(b)


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)

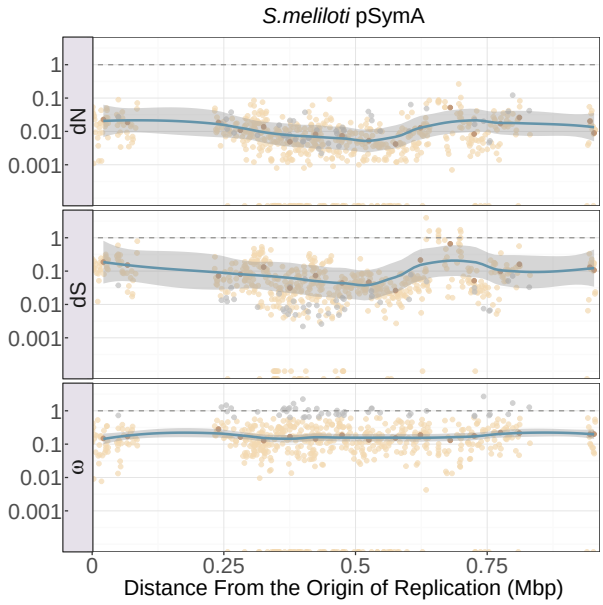
(a)


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)

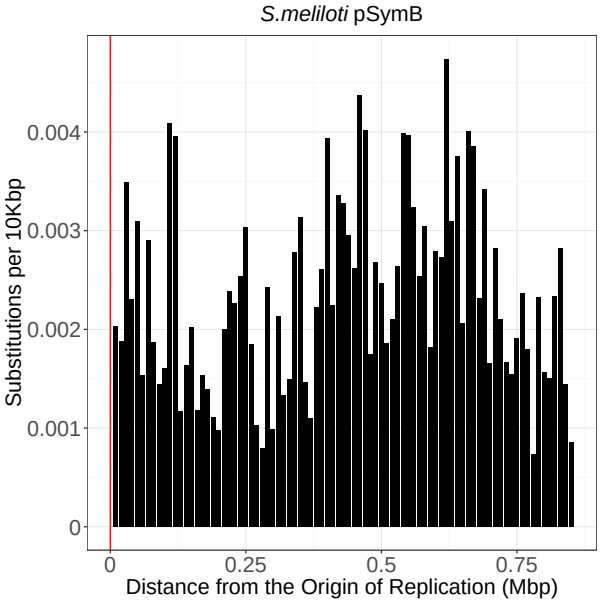
(b)


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)

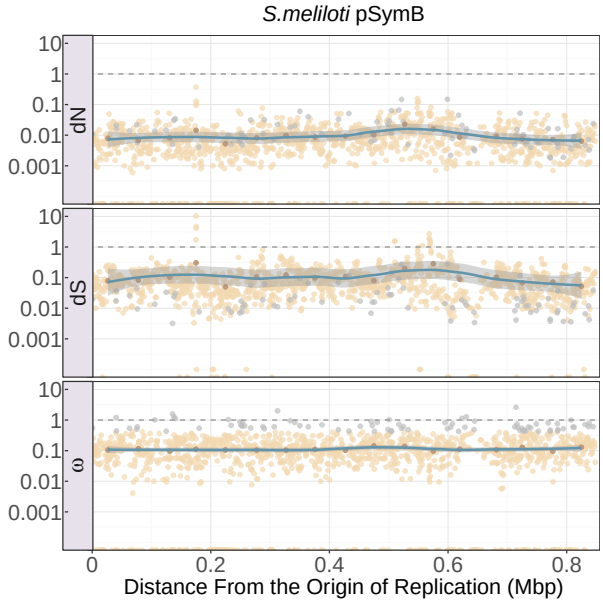
(a)


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)