

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

- ✓make an R theme for the graphs so they all look the same
- ✓updated the *Streptomyces* selection figures to have negative positions (to match the substitution graph)
- ✓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
- ✓add in reviewers comments from gene expression paper about why we chose these bacteria to substitution paper

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  $\omega$  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* alignment (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 similar)? **I am 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?**

Another thing I notices while looking into the selection analysis for *S. meliloti* chromosome is that sometiems PAML puts a  $dS$  value of 0.0001 even though there are no substitutions at all. I am not sure how often this happens and if these values (0.0001) are only associated with no substitutions. **Thoughts on what to do with these values?**

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

- continue look into whats up with *S. meliloti* chrom bc it does not look right at all
- make small changes based on Brian's comments from questions above
- update methods in paper draft (make sure they are the same as what I actually did bc things have changed a lot)
- create a script that will match up homologous genes from different *Escherichia coli* genomes for Queenie to use for the inversions project

## Next Week

- add to intro/discussion about how lots of people have found opposing substitutions trends that do not match previous expectations
- create table explaining what genes are present in high substitutions bars (like in gene exp paper) for supplement
- continue to update methods for substitutions paper

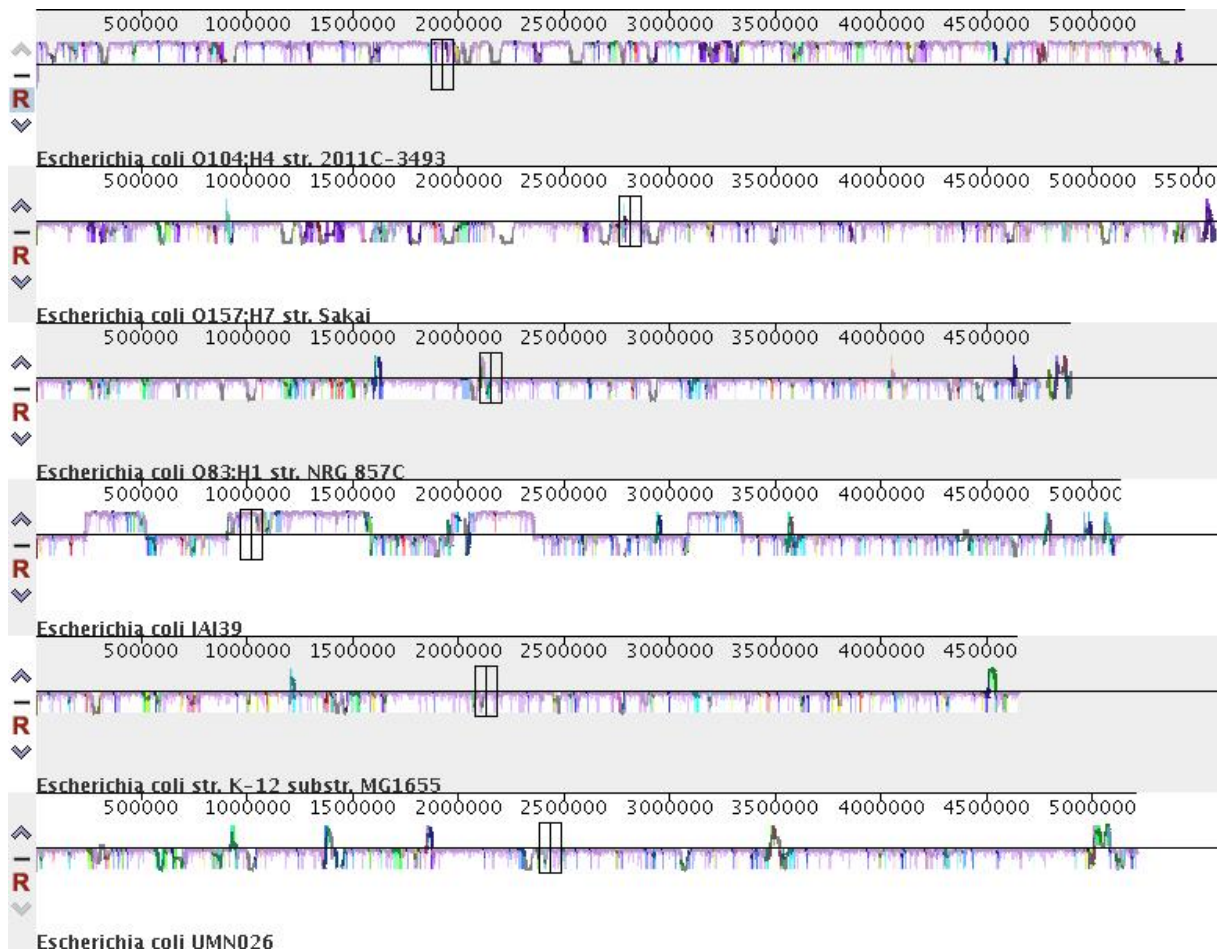


Figure 1: progressiveMauve alignment of *Escherichia coli* genomes highlighting the “backbone” of the alignment (matching regions).

Bacteria and Replicon	Protein Coding Sequences
<i>E. coli</i> Chromosome	$-1.43 \times 10^{-8}***$
<i>B. subtilis</i> Chromosome	$-5.55 \times 10^{-8}***$
<i>Streptomyces</i> Chromosome	$7.49 \times 10^{-8}***$
<i>S. meliloti</i> Chromosome	$-5.99 \times 10^{-7}***$
<i>S. meliloti</i> pSymA	$-5.18 \times 10^{-7}***$
<i>S. meliloti</i> pSymB	$1.67 \times 10^{-7}***$

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:  $< 0.001 = ***$ ,  $0.001 < 0.01 = **$ ,  $0.01 < 0.05 = *$ ,  $> 0.05 = \text{'NS'}$ .



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
<i>E. coli</i> Chromosome	$1.97 \times 10^{-4}$
<i>B. subtilis</i> Chromosome	$1.93 \times 10^{-4}$
<i>Streptomyces</i> Chromosome	$2.74 \times 10^{-6}$
<i>S. meliloti</i> Chromosome	$9.72 \times 10^{-5}$
<i>S. meliloti</i> pSymA	$6.54 \times 10^{-5}$
<i>S. meliloti</i> pSymB	$1.99 \times 10^{-4}$

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.

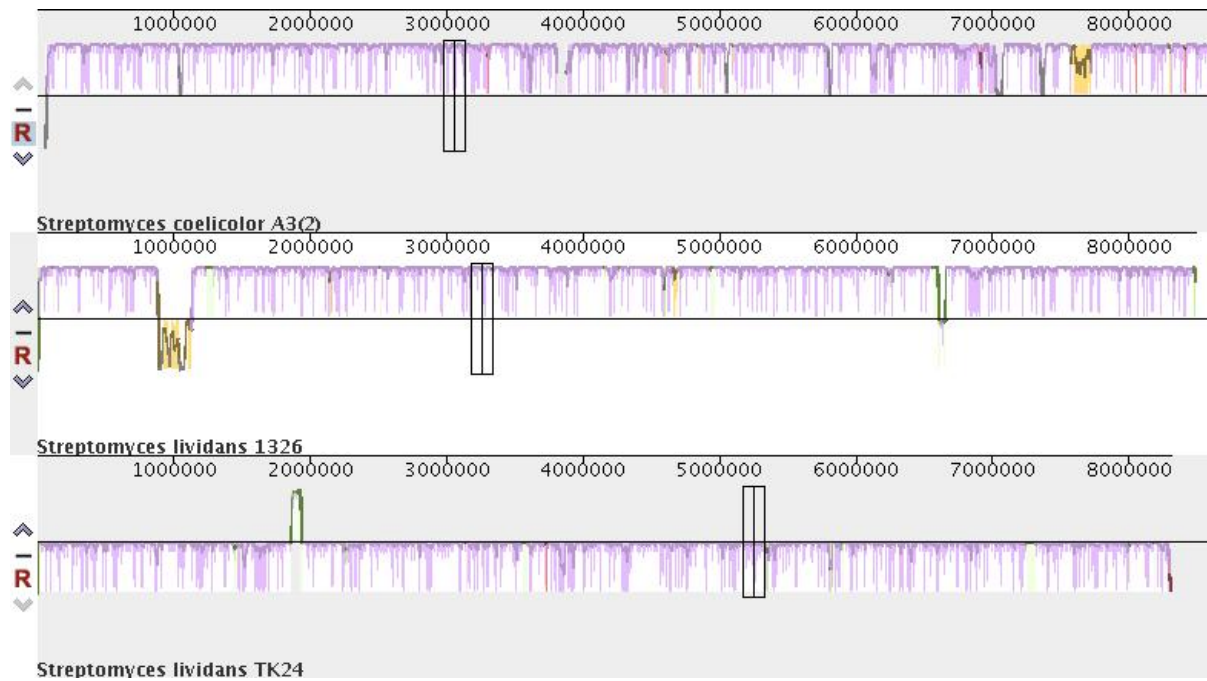
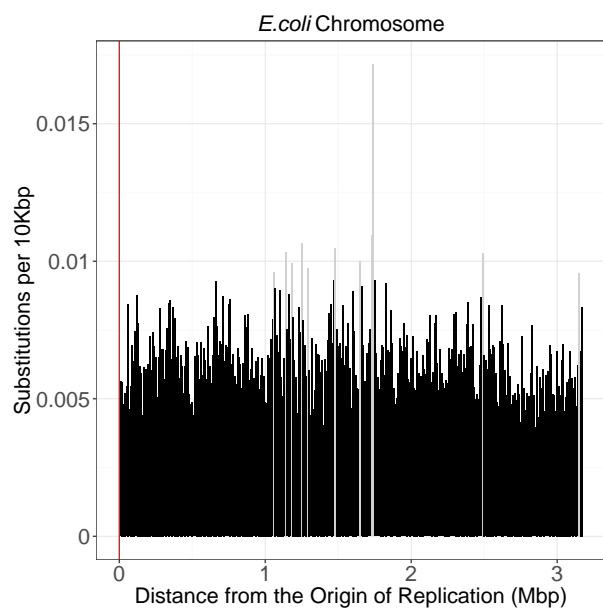


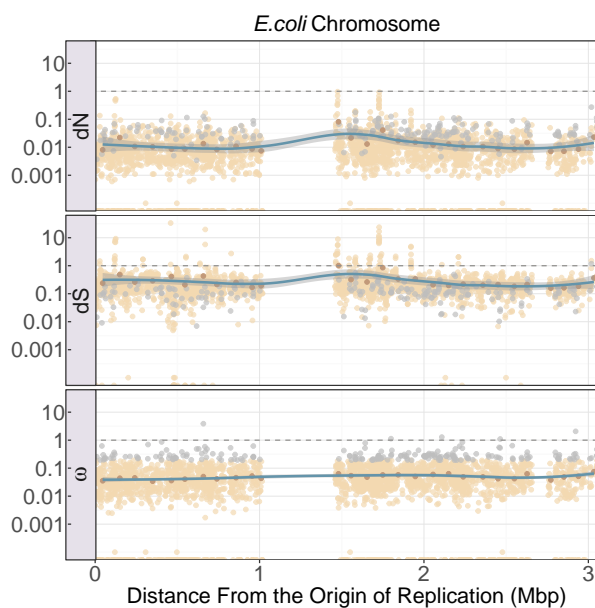
Figure 3: progressiveMauve alignment of *Streptomyces* genomes highlighting the “backbone” of the alignment (matching regions).

Bacteria and Replicon	Genome Average		
	dS	dN	$\omega$
<i>E. coli</i> Chromosome	0.2387	0.0101	0.0441
<i>B. subtilis</i> Chromosome	0.4201	0.0243	0.0714
<i>Streptomyces</i> Chromosome	0.0458	0.0011	0.0335
<i>S. meliloti</i> Chromosome	0.0029	0	0
<i>S. meliloti</i> pSymA	0.0874	0.0099	0.1645
<i>S. meliloti</i> pSymB	0.0940	0.0084	0.1142

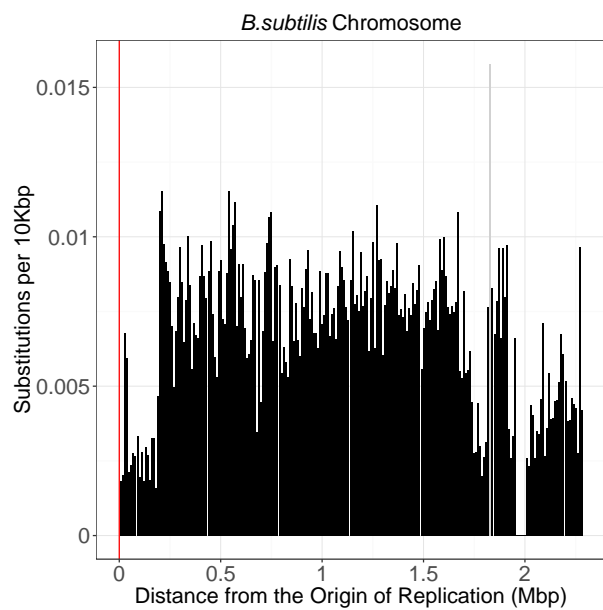
Table 3: Weighted averages calculated for each bacterial replicon on a per genome basis using the gene length as the weight. Arithmetic mean calculated for the per gene averages for each bacterial replicon.



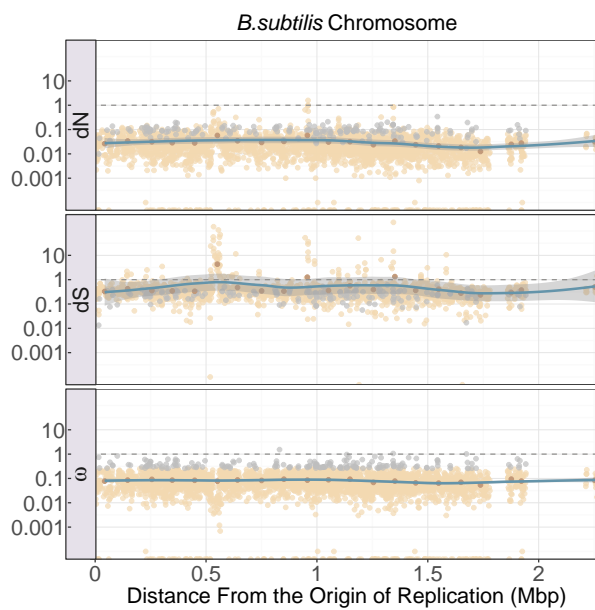
(a)



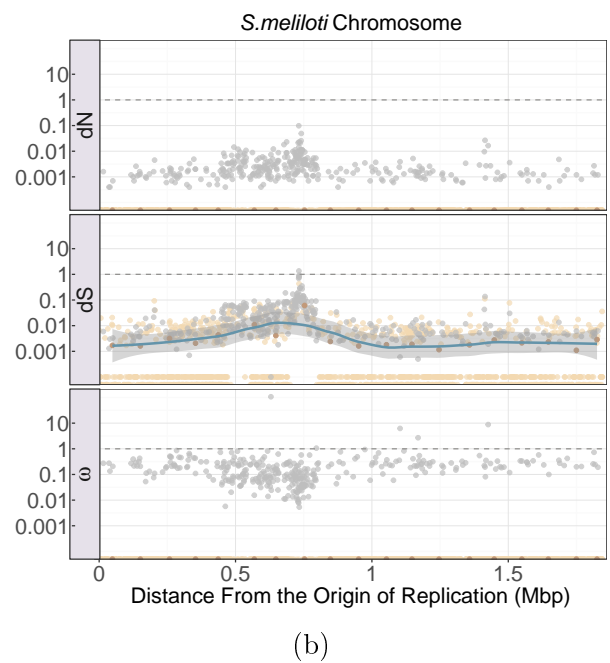
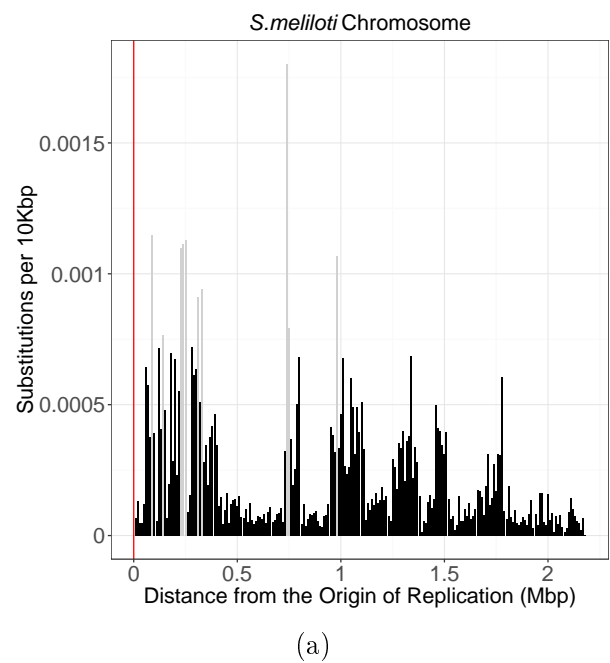
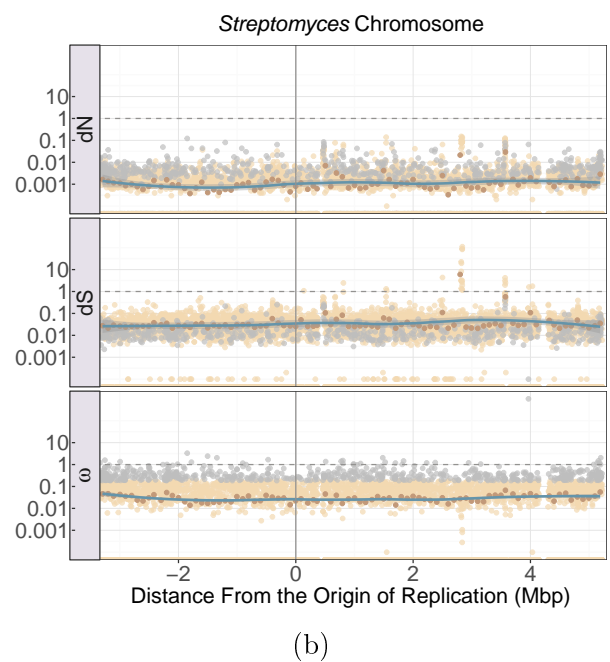
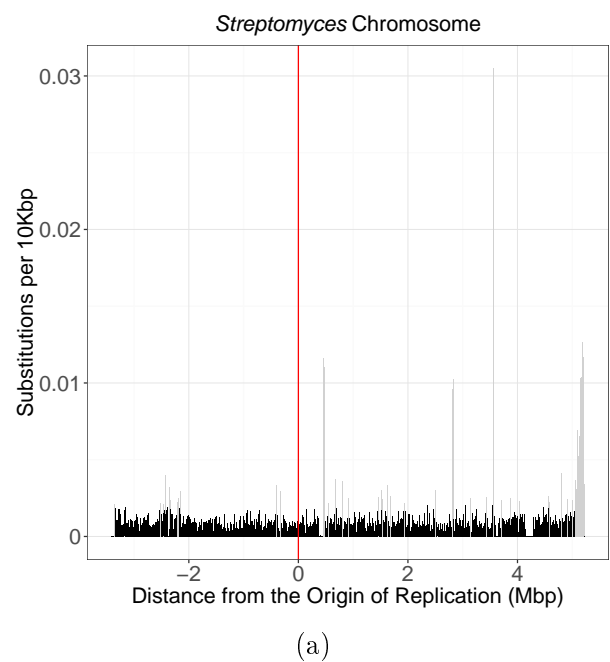
(b)



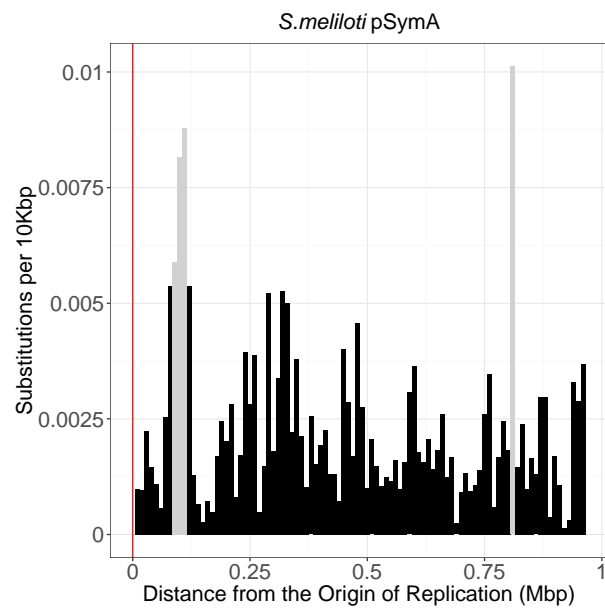
(a)



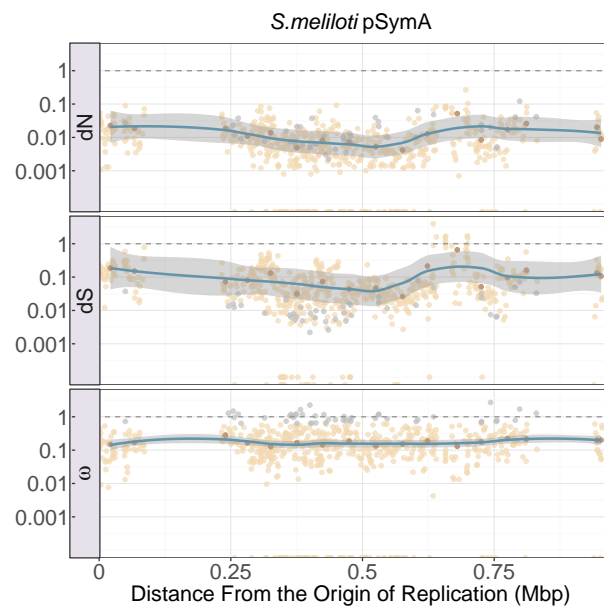
(b)



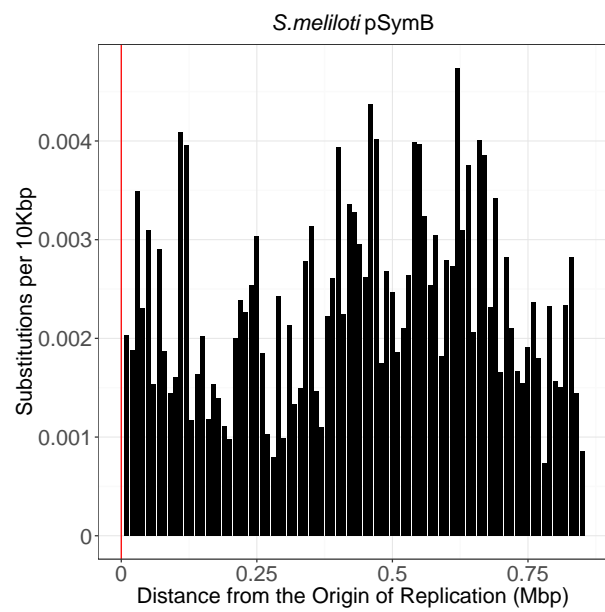




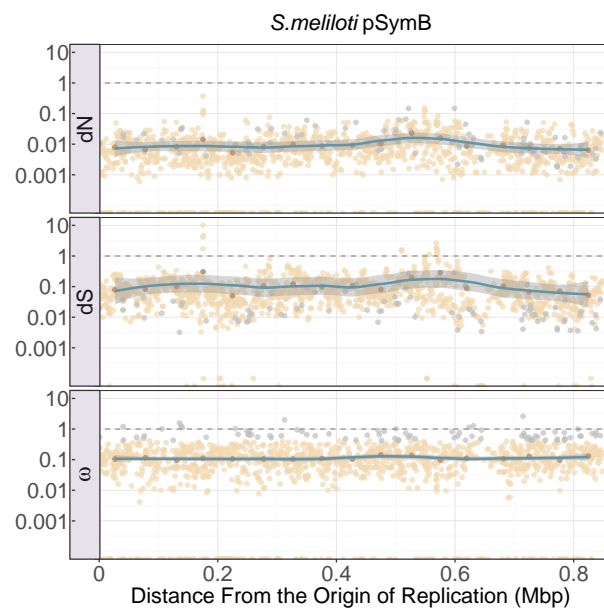
(a)



(b)



(a)



(b)