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

✓remove horizontal reference lines in dN and dS portion of graphs

✓change all dS values of 0.0001 to 0

✓check into why number of subs near the ter is higher than near the origin

✓calculate proper near/far from ori substitutions linear regression

✓edit/update methods

✓write code to get block information needed for Queenie

I double checked that all dS values of 0.0001 actually do correspond to no synonymous substitutions. These have all be changed to values of 0 in all bacterial replicons. **Should I write in my methods that sometimes codeml will put 0.0001 as a value for *dS* even if there are no synonymous substitutions? and that we made these 0.0001 values zero?**

When going through the latest results from the substitutions and selection analysis I noticed that the number of subs near the terminus was higher than near the origin, even though the overall trend is that the number of subs decreases when moving away from the origin. This is found in *E. coli*, *B. subtilis*, *Streptomyces* and pSymA. So I think what is happening is that when you look locally (a small region, like 20Kbp) then we are seeing that the weighted number of subs near the origin in higher than the weighted 20Kbp number of subs near the terminus. But globally, when we look at the whole genome, we still see a decreasing trend. I think that I will need to re-work my discussion to talk about this. I also noticed that outliers were NOT removed in my near/far from the origin calculations. The way that I am calculating outliers is by considering a whole bar in my subs graphs (so weighted subs over each 10Kbp, light grey bars in my graphs) as an outlier. **Therefore, should I be removing ALL points (substitutions and no substitutions, or just subs?) in that bar and just skip to the next 10Kbp bar and consider that part of the 20Kbp "near" the origin? Should I not remove any outliers? Should I be calculating outliers another way, like using only binary data (is that even possible?)?**

When checking into why the substitutions near the terminus were higher than near the origin, I realized that for *Streptomyces* the calculations of what was "near" and "far" from the origin were wrong. They were just calculating the two ends of the chromosome not near zero. So I fixed this and re-did that portion of the analysis.

I made a lot of minor edits to the methods section, adding in new methods and generally updating. I sent you the latest draft of this and will be making revisions based on your comments this week.

I wrote a script to get all of the necessary Block information for the PARSNP alignments so Queenie can eventually use this data frame to find/confirm homologous genes between each of the *Escherichia coli* genomes in the inversions.

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
- 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
- revise methods based on Brian's edits
-

Next Week

- find and look at notes from Why genes evolve faster on secondary chromosomes in bacteria (see what I can add to intro)
- make sure weighted and non weighted calculations match up with the same sign
- 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 methods / discussion to properly talk about the selection 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
- think about if the selection distribution figs or the summary selection fig should be in the main paper

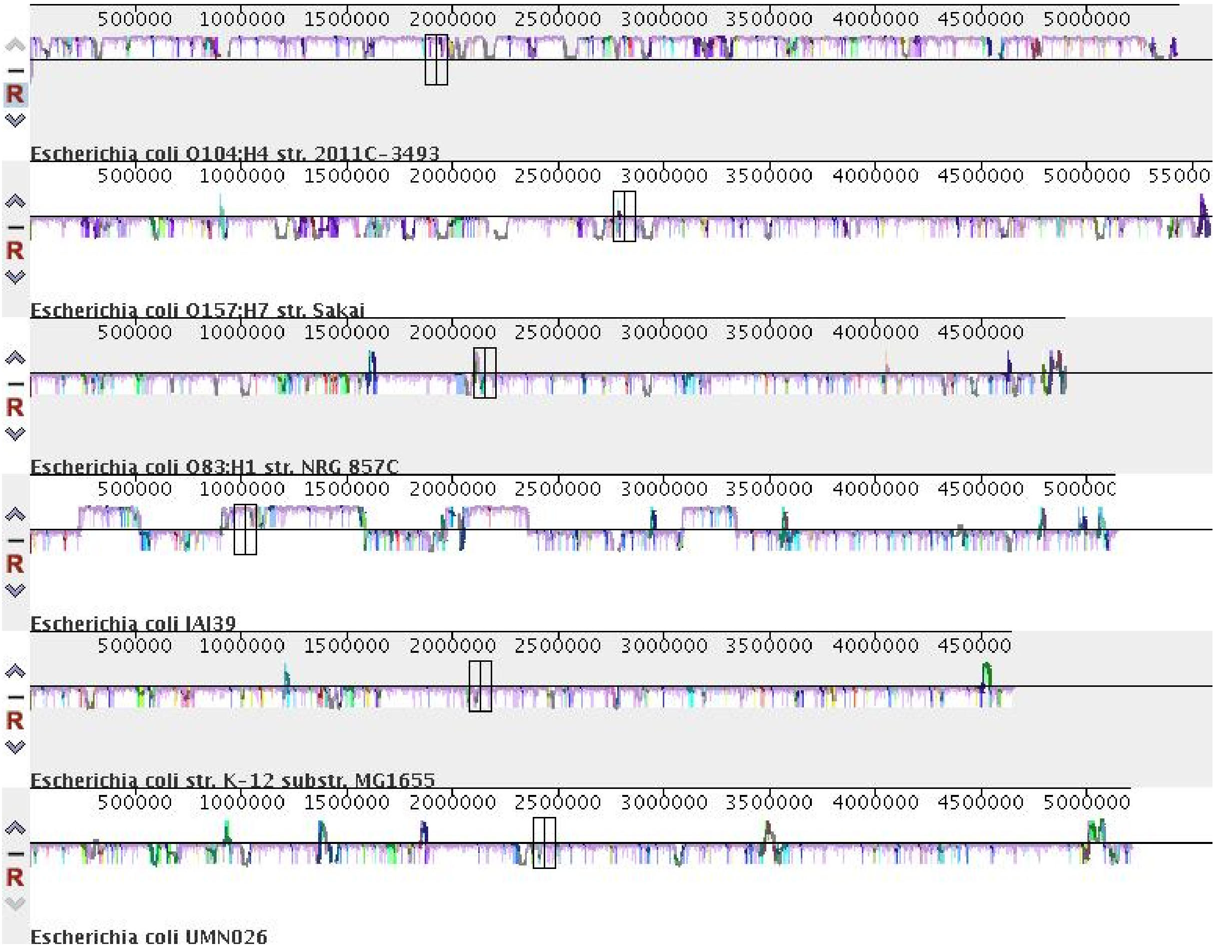


Figure 1: progressiveMauve alignment of *Escherichia coli* genomes highlighting the “backbone” of the alignment (matching regions).

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’.

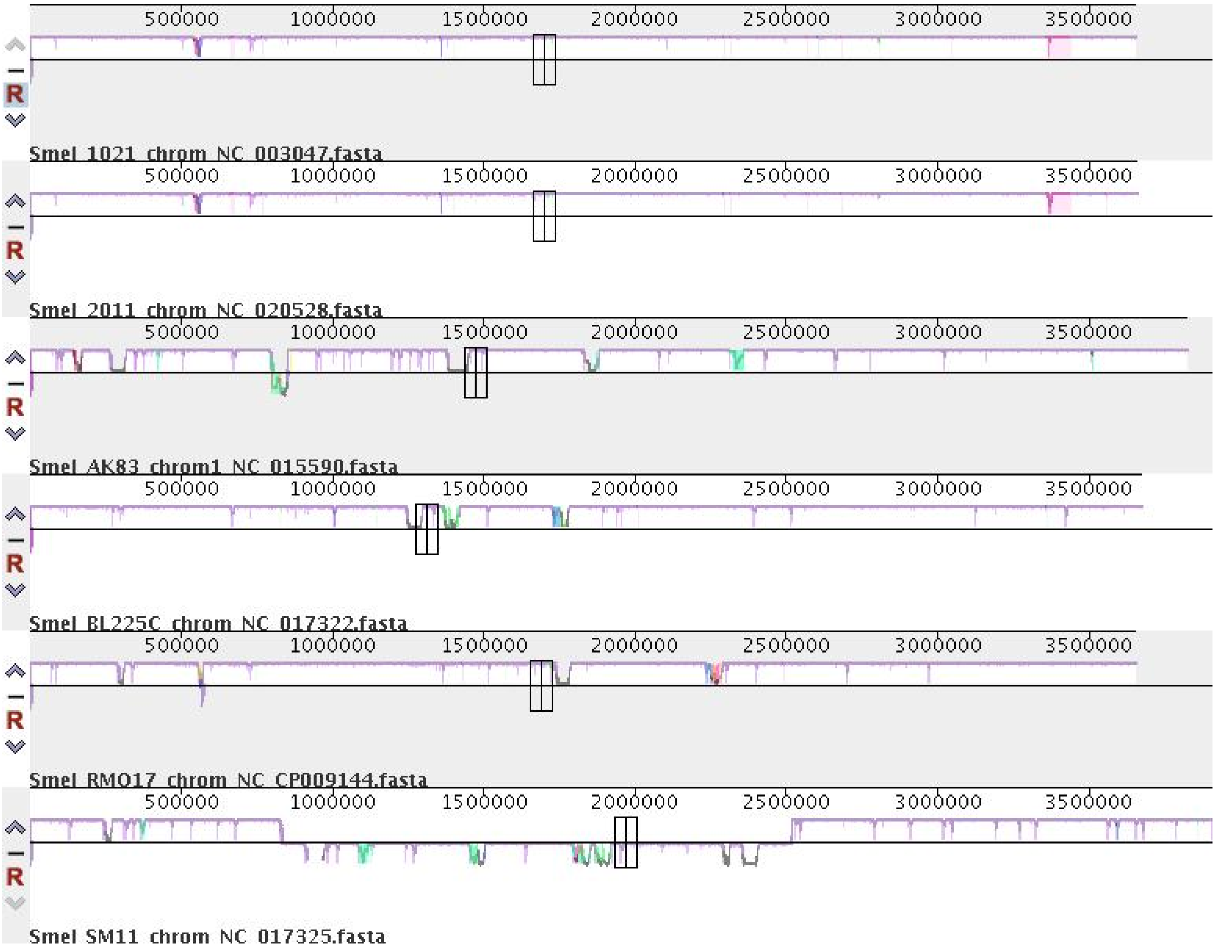

500000
1000000
1500000
2000000
2500000
3000000
3500000
R
Smel_1021_chrom_NC_003047.fasta
Smel_2011_chrom_NC_020528.fasta
Smel_AK83_chrom1_NC_015590.fasta
Smel_BL225C_chrom_NC_017322.fasta
Smel_RMO17_chrom_NC_CP009144.fasta
Smel_SM11_chrom_NC_017325.fasta

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.

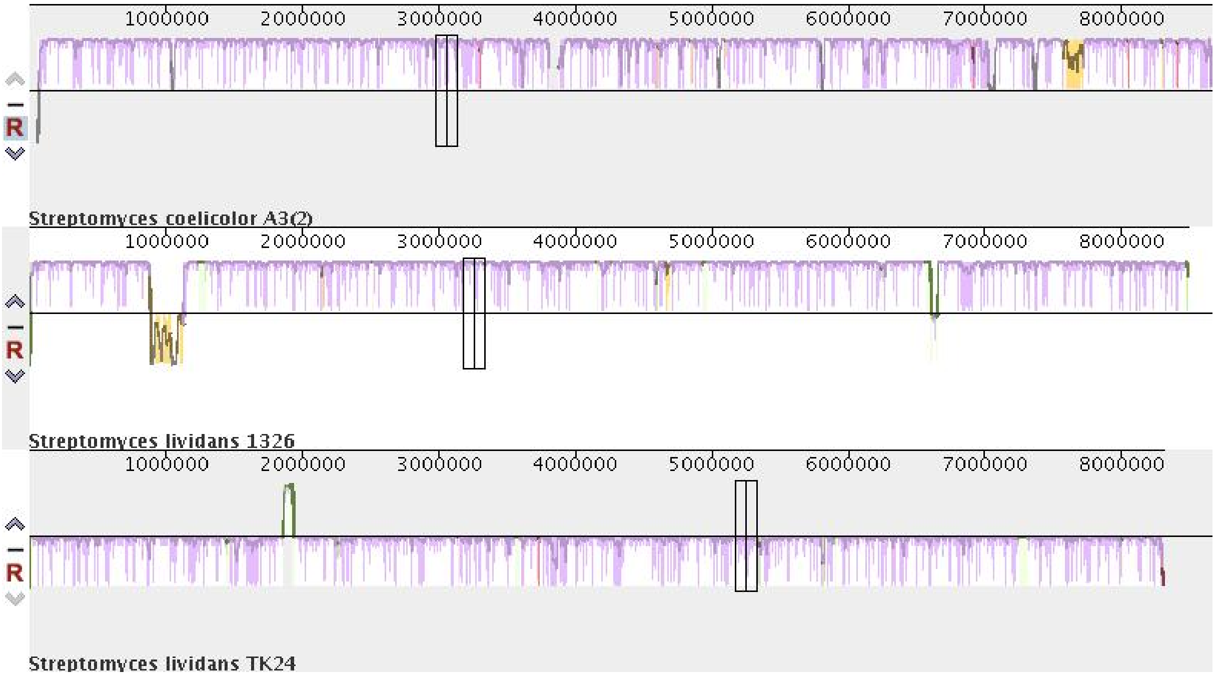
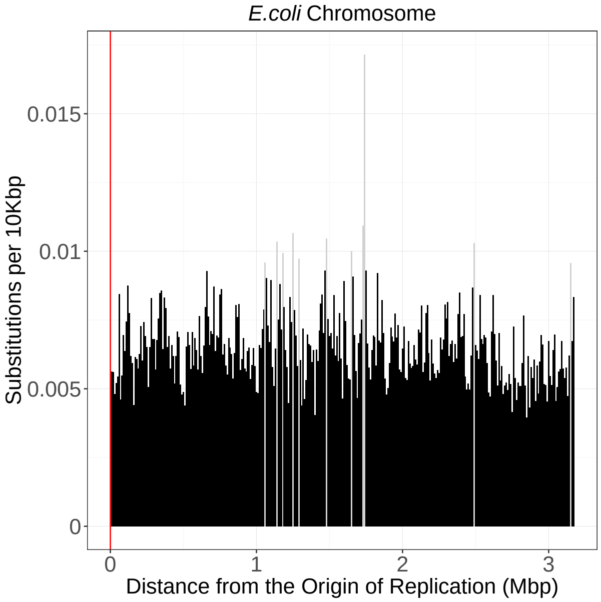

1000000
2000000
3000000
4000000
5000000
6000000
7000000
8000000
R
Streptomyces coelicolor A3(2)
Streptomyces lividans 1326
Streptomyces lividans TK24

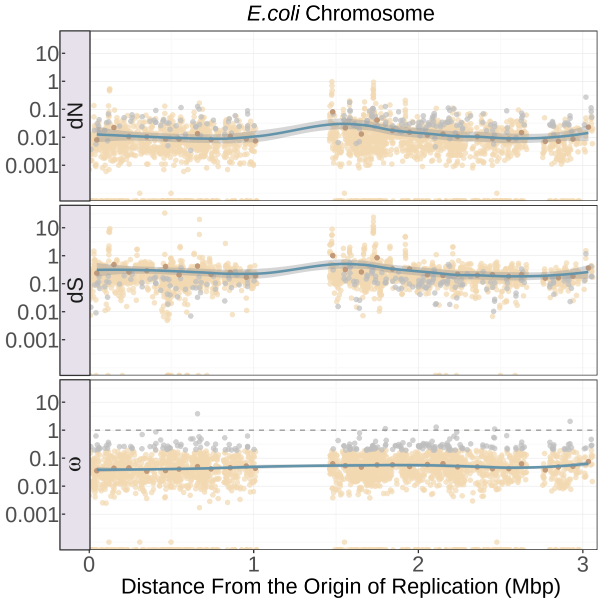
Figure 3: progressiveMauve alignment of *Streptomyces* genomes highlighting the "backbone" of the alignment (matching regions).

	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	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

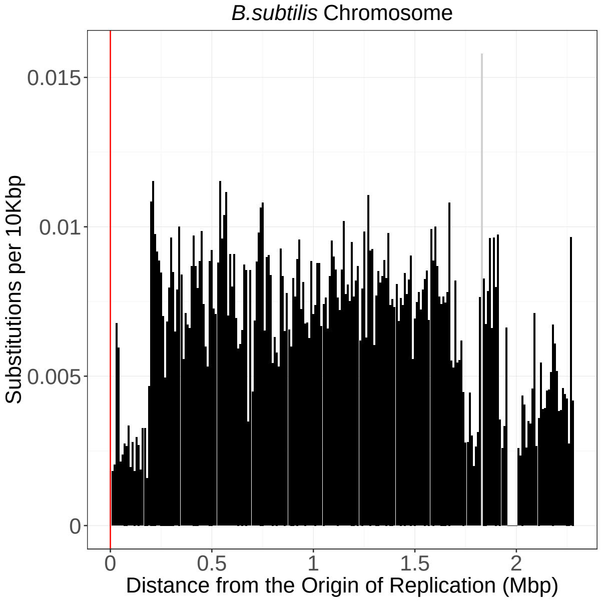
Table 3: Weighted averages calculated for each bacterial replicon on a per genome basis using the gene length as the weight. Arithmetic mean calculated for the per gene averages for each bacterial replicon.


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

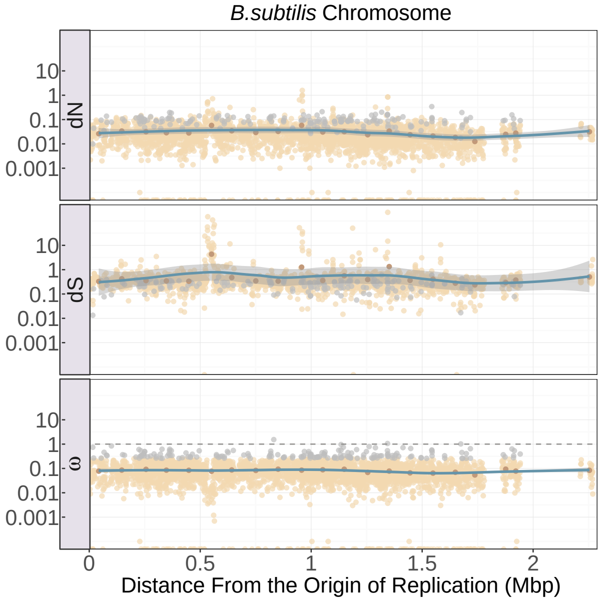
(a)


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)

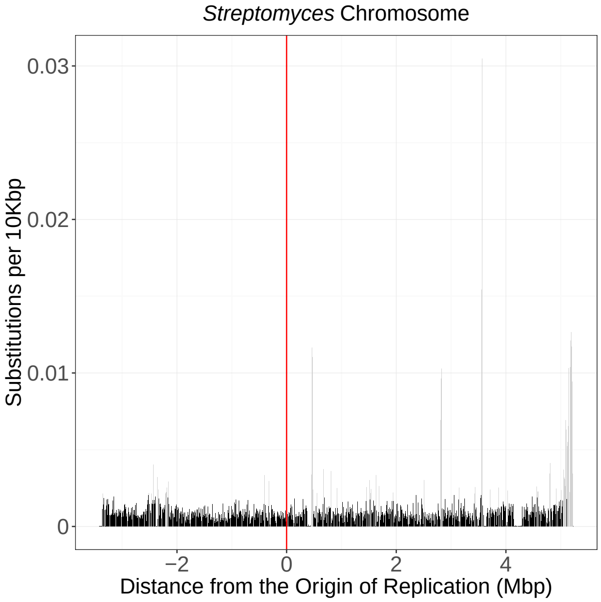
(b)


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)

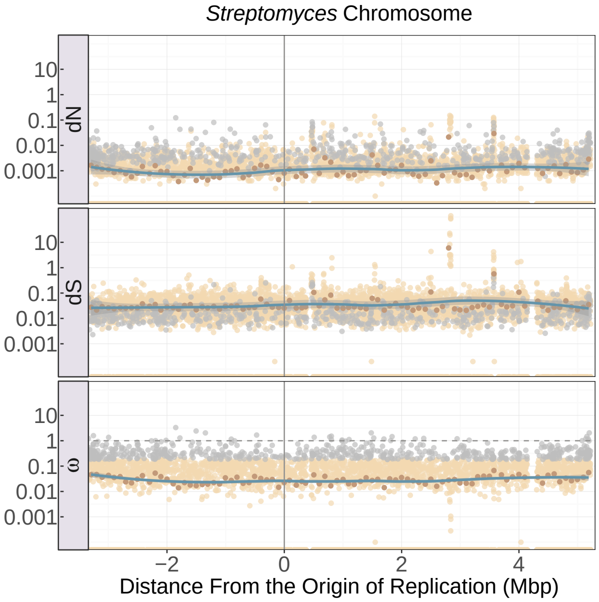
(a)


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)

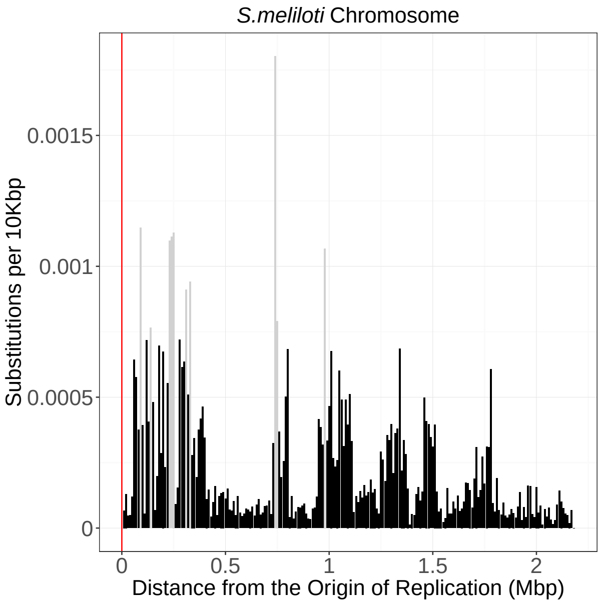
(b)


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

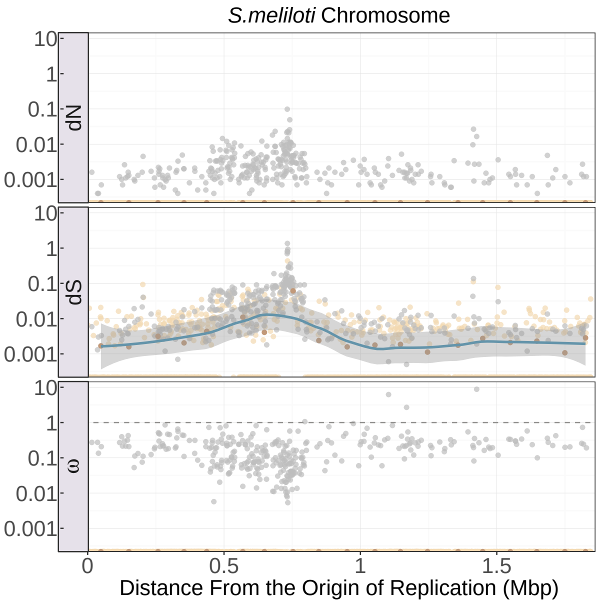
(a)


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

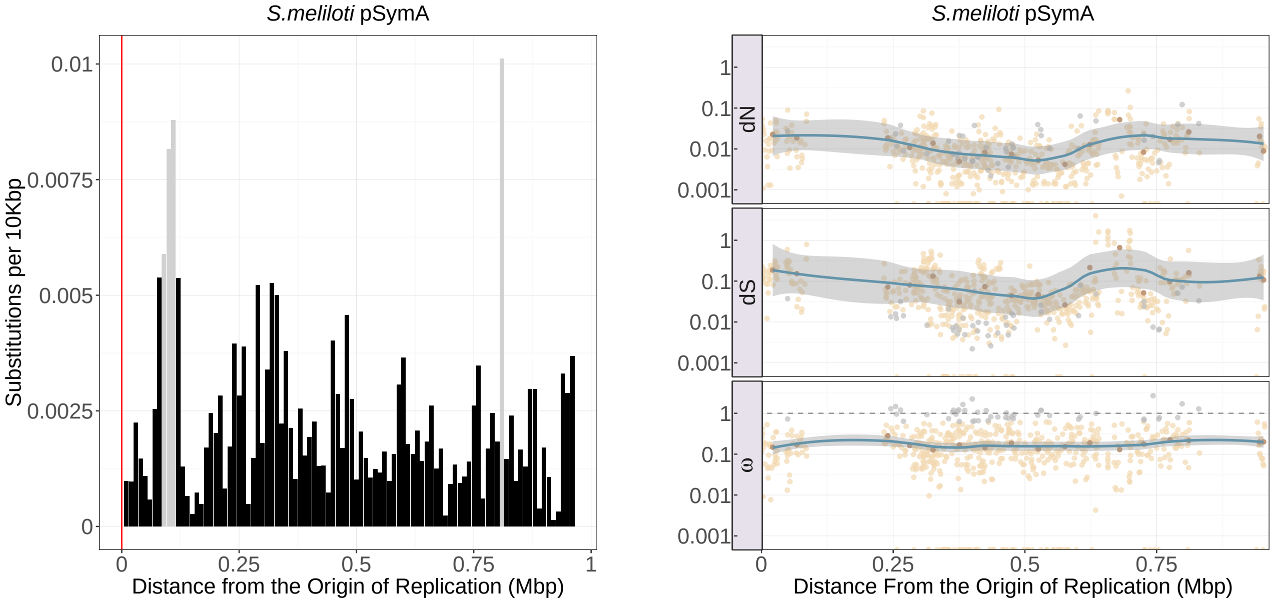
(b)


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)

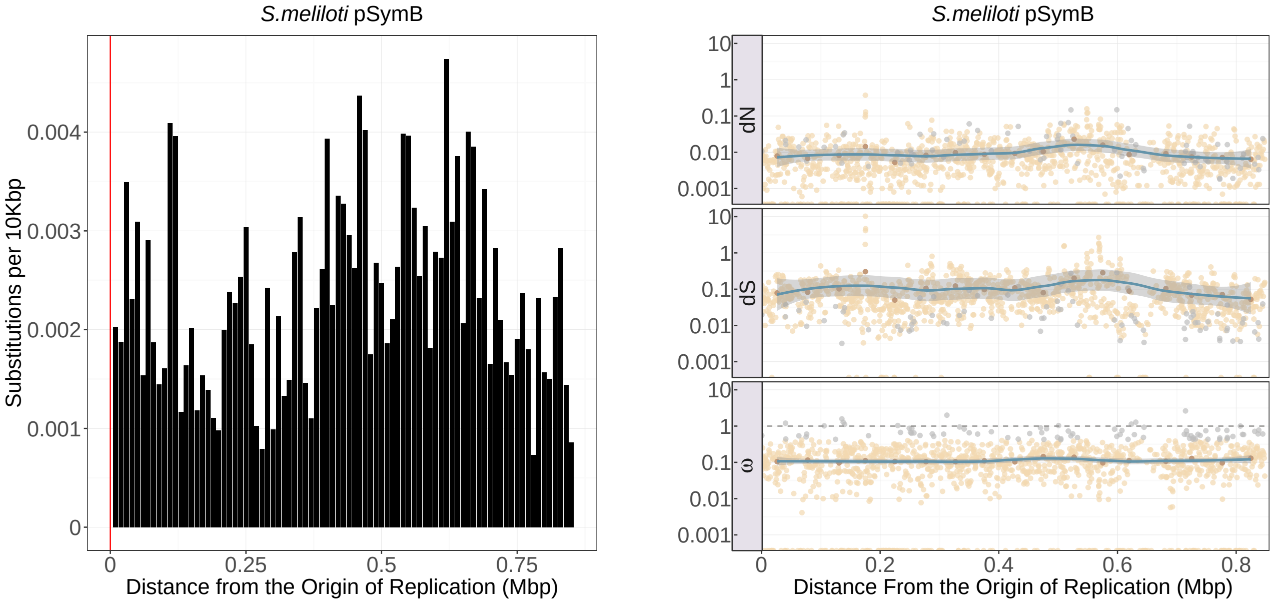
(a)


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)
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)

(a) (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)
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)

(a) (b)