

Subs Paper Things to Do:

- why does ~~sineC~~ 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  $\omega$  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

✓ re-submit gene expression paper! Woo!

✓ change colours for outlier and actual data in selection distribution graphs

✓ why *S. meliloti* chromosome has an omega linear regression of 0 near and far from the origin

✓ find high substitution bar in *Streptomyces* as a supplementary example for the paper (to show that high subs that are left are real!)

✓ figure out how to deal with significant number of duplicates in one GEO gene expression dataset (for inversions analysis)

✓ begin making changes to first year presentation

✓ minor edits to selection graphs

We resubmitted the gene expression paper! Woo!

**add in graphs and ask brian which he likes best**

I changed the colours for the selection graphs to make it a bit easier to see the outliers. I also made minor edits to the graphs like making sure the trendline means are in the middle of the genomic region they are averaging (instead of at the beginning). I also attempted to see if R will plot a trendline if one of the points is zero, and it will not, at least not on a log scale. So we are

stuck with some segments of the graph “missing” a line. But, the mean points are still at zero so the information that the average selection value is zero in these genomic regions is still visible. I also tried to plot the selection values as  $\log(x + 1)$  to try and include the zeros, but it did not work, everything was still squished to the x-axis. So I will have to make a note in my captions about the number of zero values in each graph.

I also looked into why the *S. meliloti* chromosome had linear regression of zero near and far from the origin, and it is because all the values in these regions are zero. Any non-zero values were labelled as outliers. Not sure how we should be explaining this in the paper or if we should be altering some part of the analysis? **Thoughts?**

I also chose a particularly high substitutions bar from the substitutions graphs to add into the supplement as an example showing that whatever remaining high numbers of substitutions are left, are actually real! I also added info about this in the main portion of the paper.

I started to make changes to the first year bioinformatics lecture, however, in light of the school cancelling all in-person classes, I assume this will be postponed until the fall.

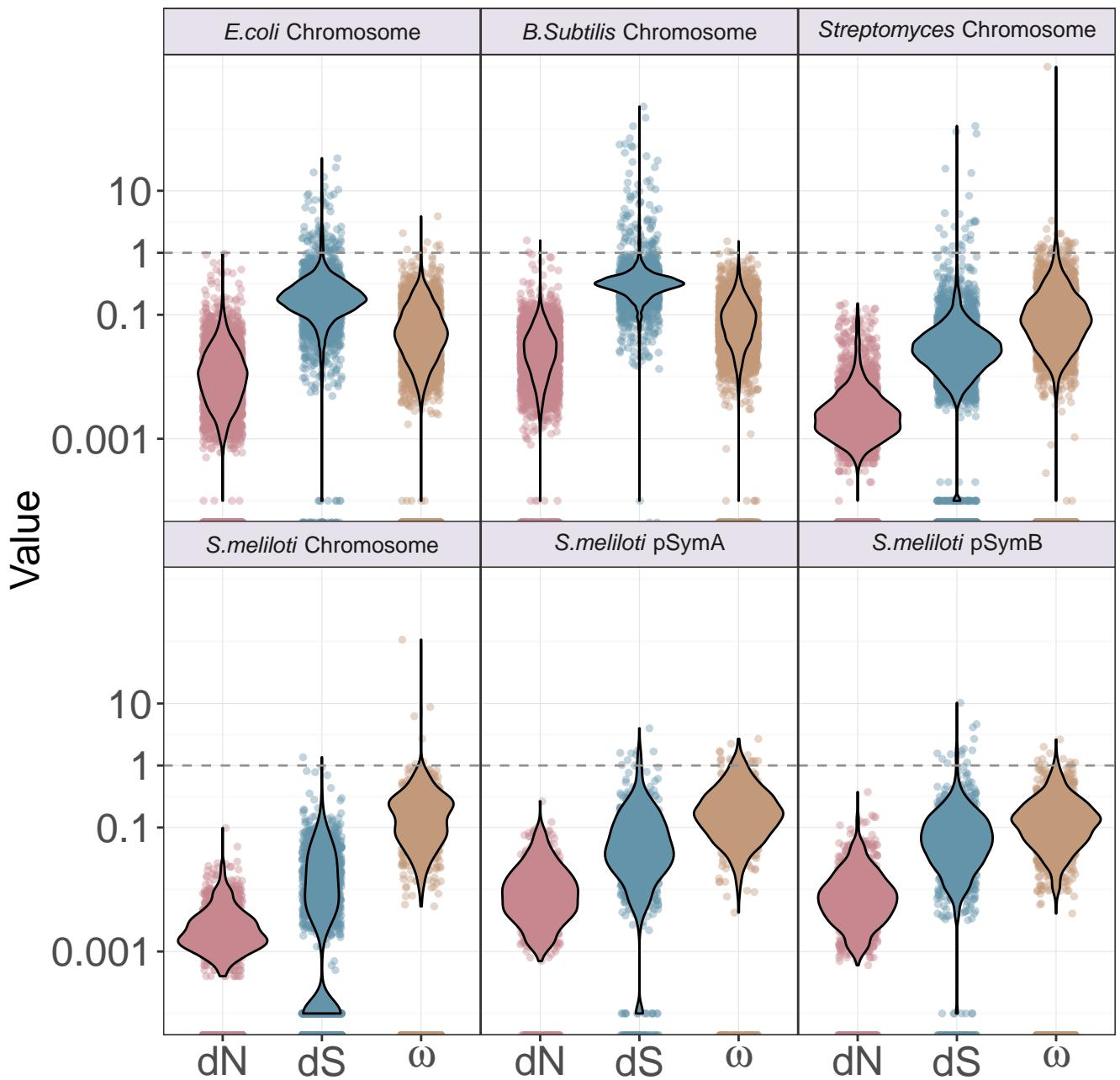
I also spoke with you about the duplicated genes in one of the GEO datasets for the inversion analysis. I contacted the author and they said that I can just add up all the raw read values to get the total for that gene. So this is what Queenie is working on as well as normalizing all the gene expression data.

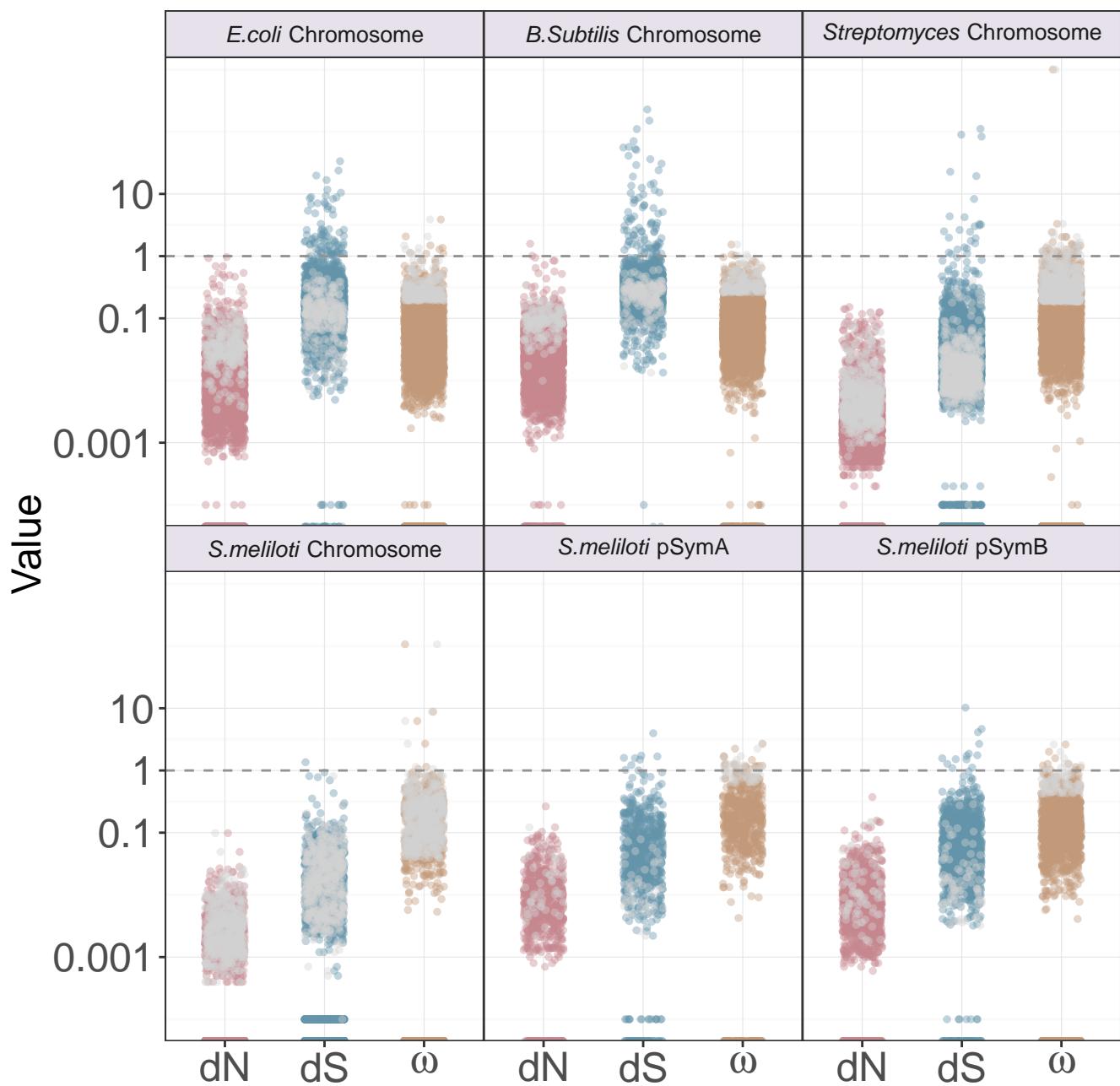
## This Week

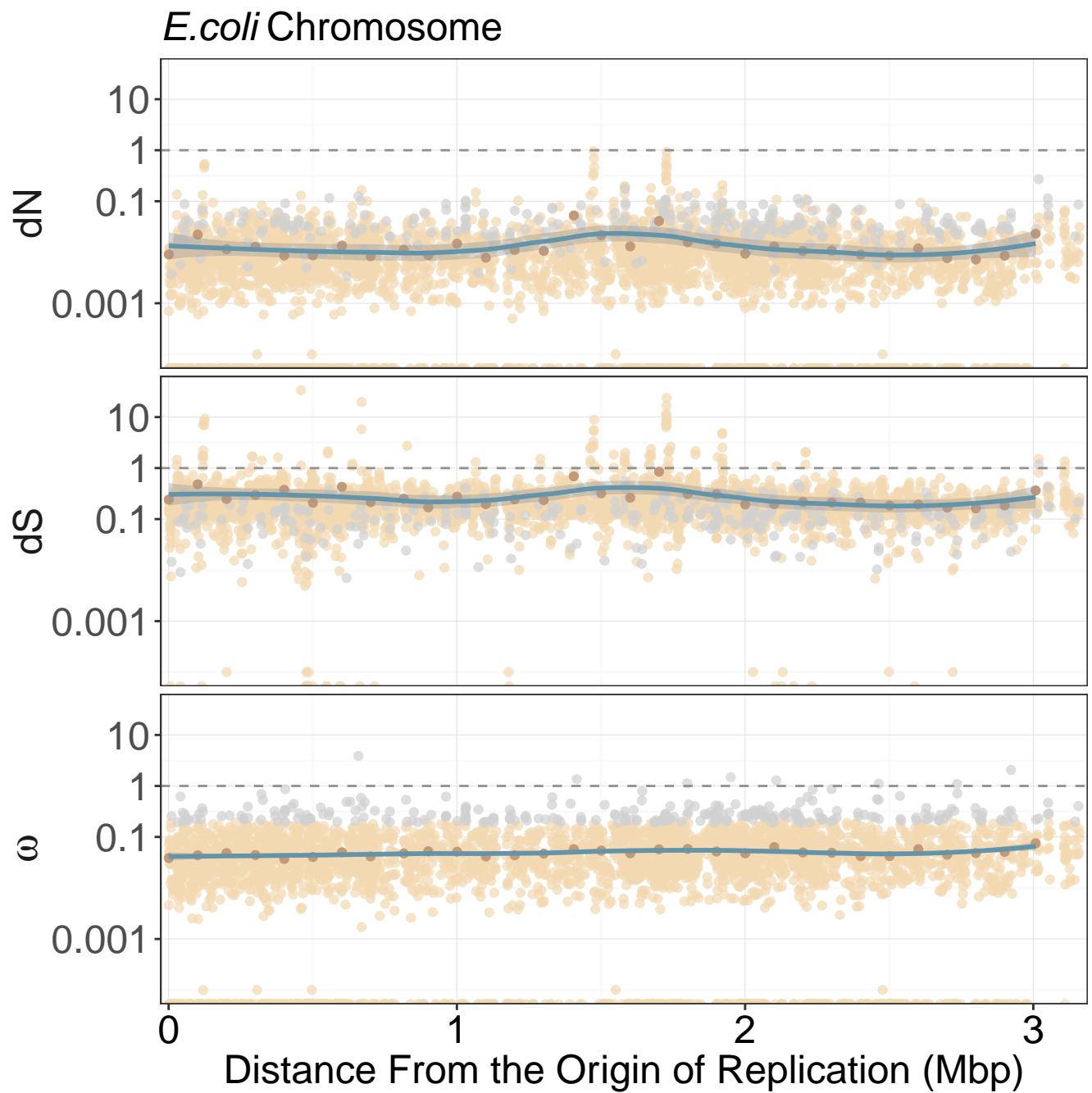
- make changes to first year presentation

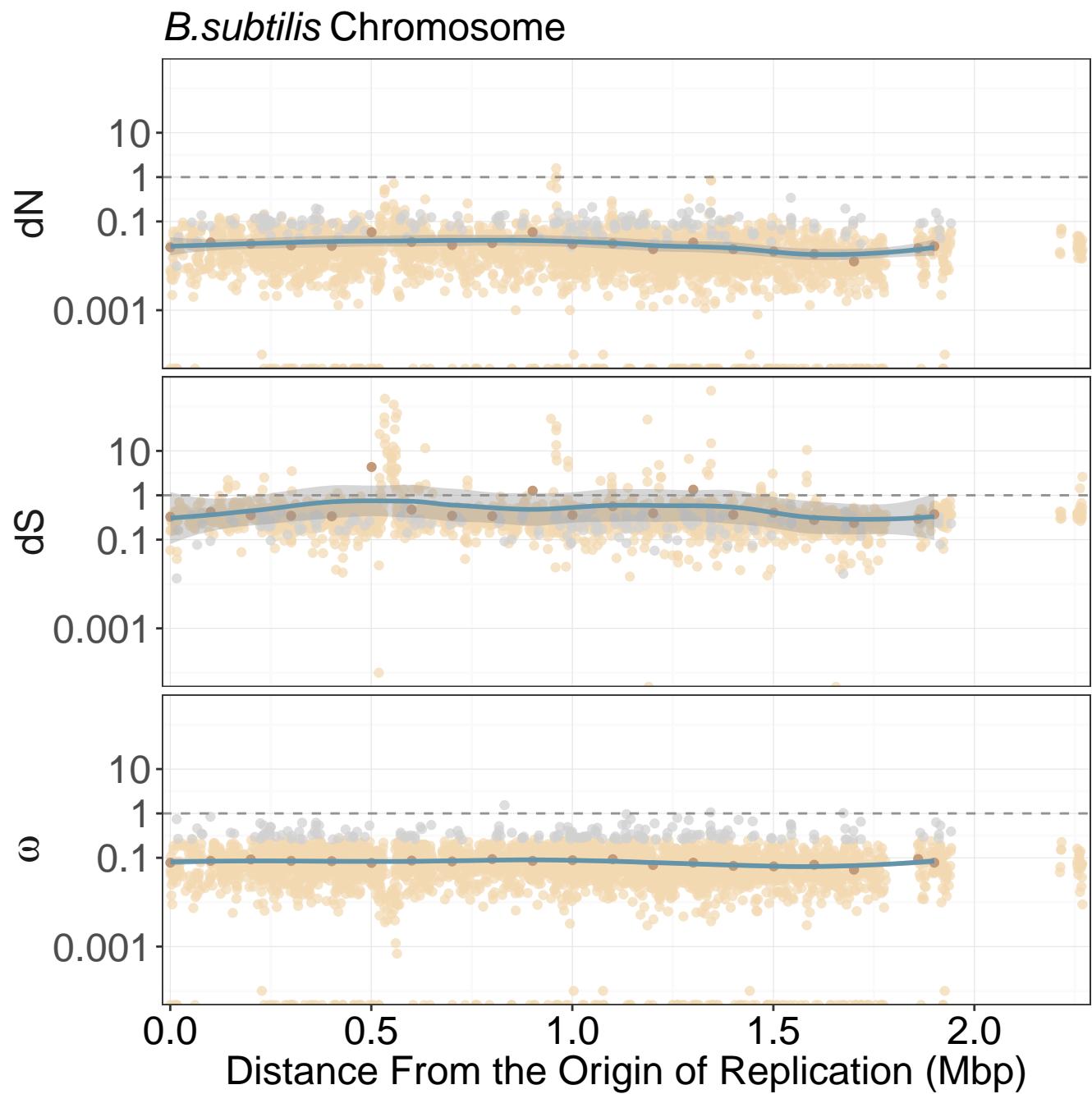
## Next Week

- find a block where mauve aligns non-homologous regions and put into supplement
- define a theme for the substitutions graphs (and selection graphs) and re-do these and put in paper
- Why does *S. meliloti* (and *B. subtilis*) have “missing data” in selection distribution graphs

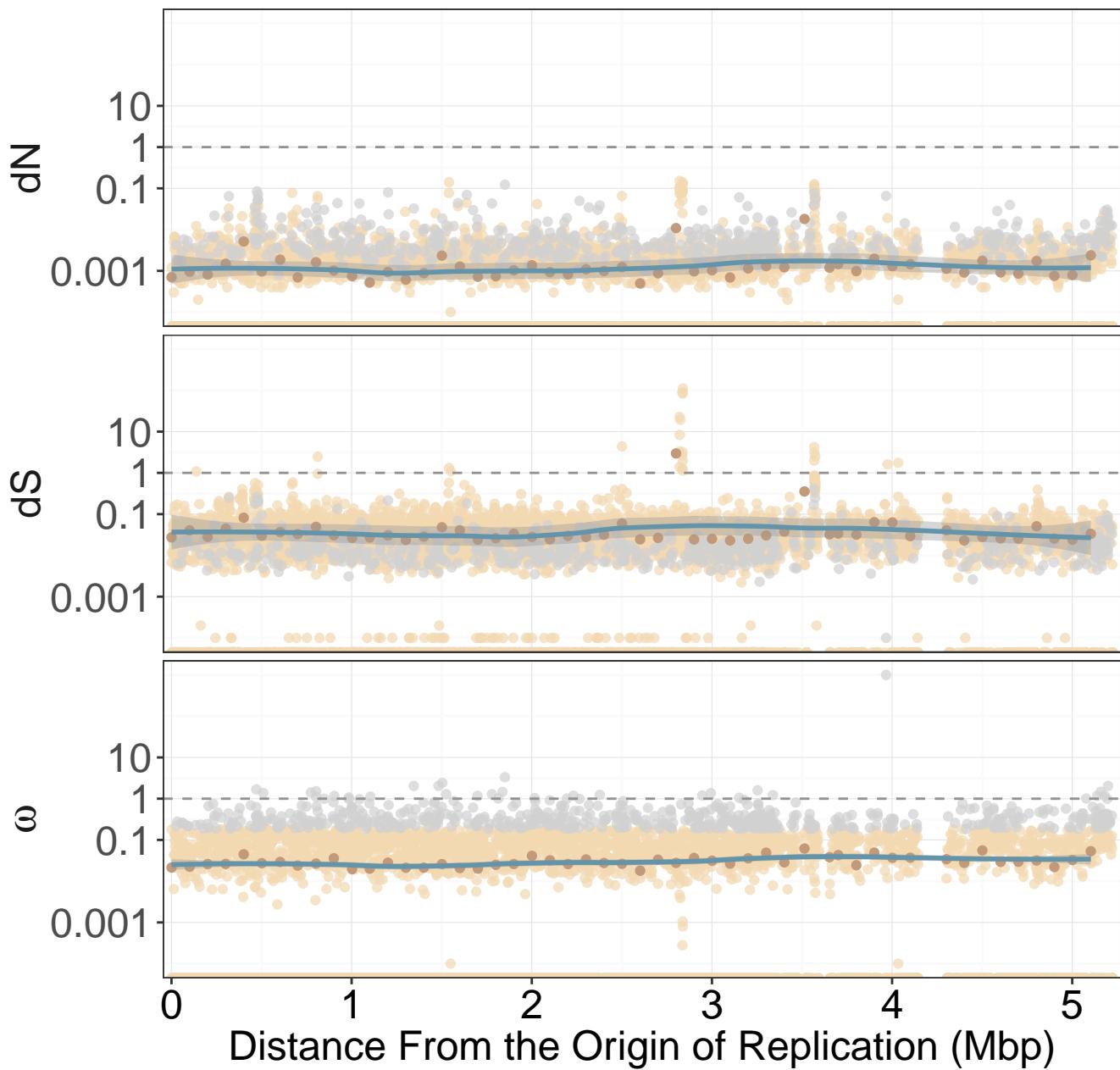


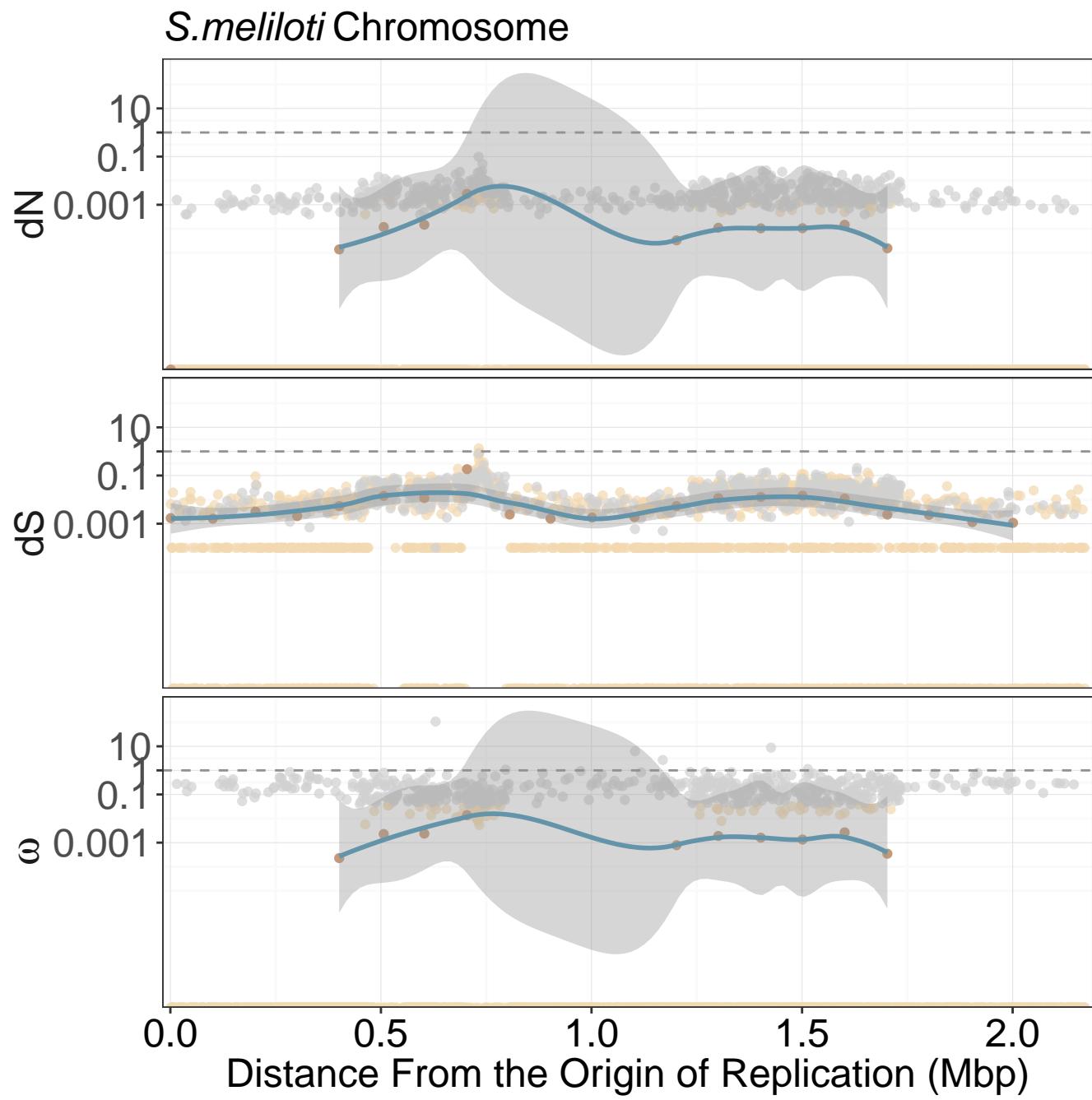


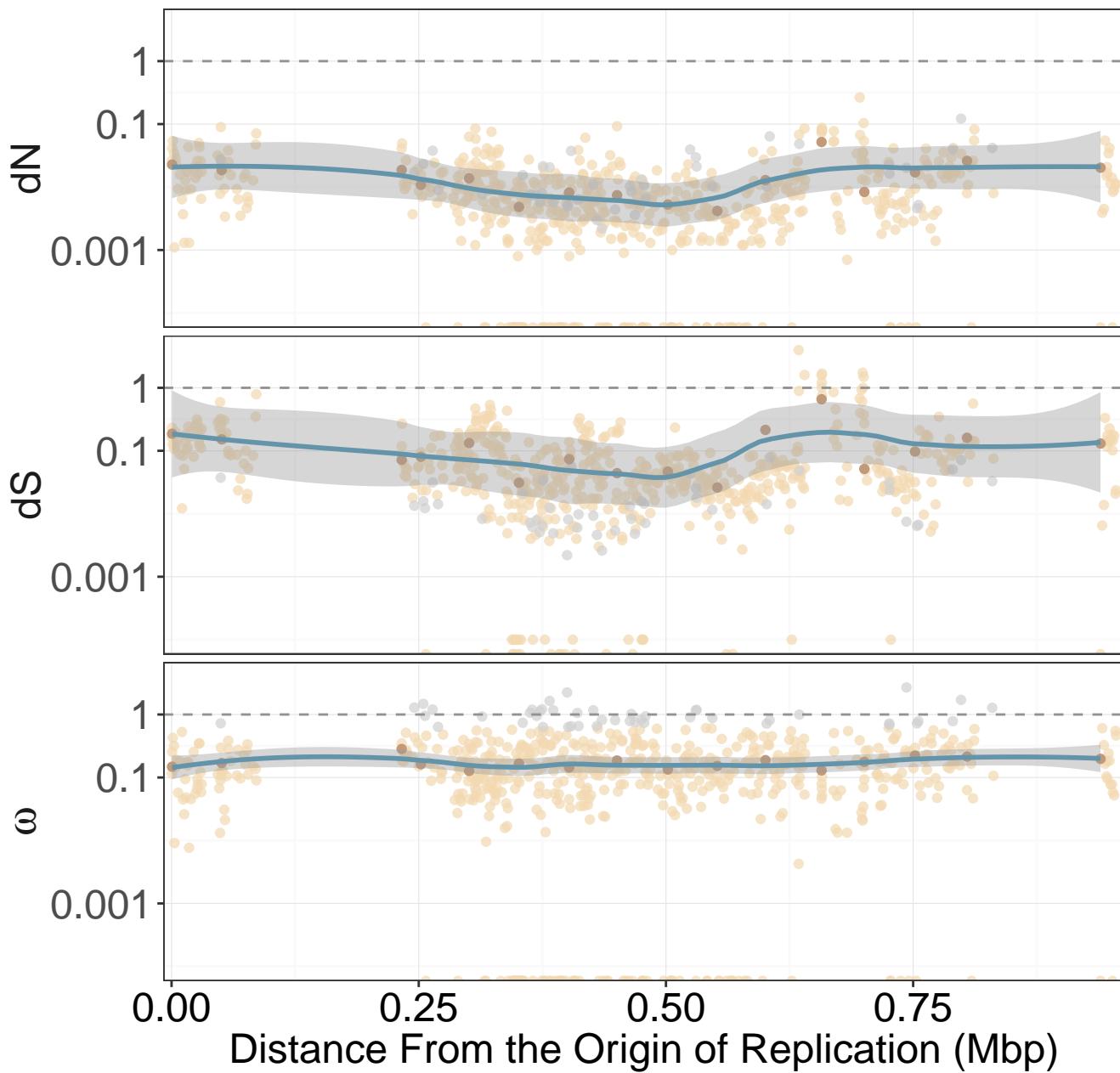


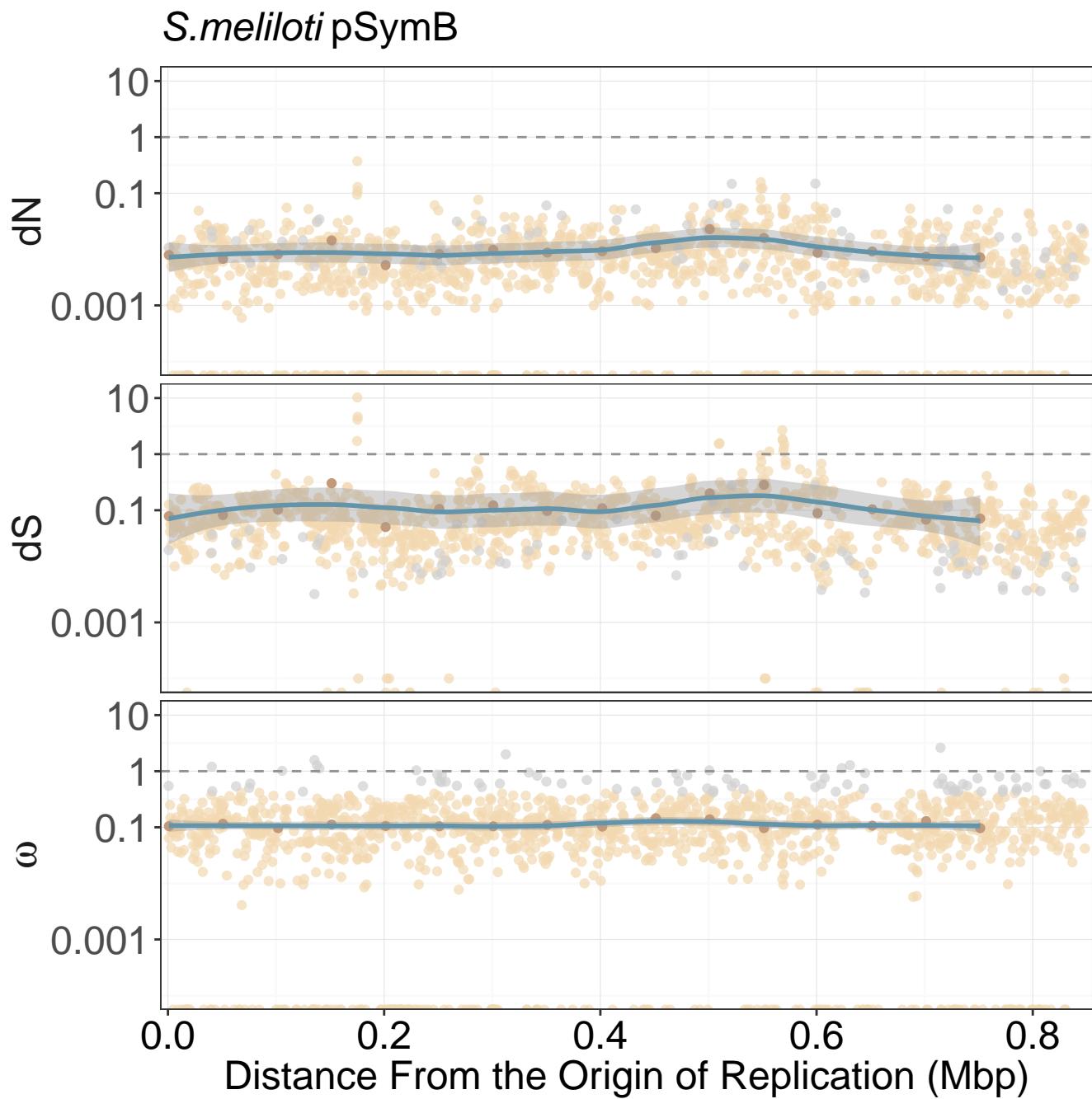


## *Streptomyces* Chromosome





*S.meliloti* pSymA



| Bacteria and Replicon          | Protein Coding Sequences  |
|--------------------------------|---------------------------|
| <i>E. coli</i> Chromosome      | $-1.98 \times 10^{-8}***$ |
| <i>B. subtilis</i> Chromosome  | $-5.55 \times 10^{-8}***$ |
| <i>Streptomyces</i> Chromosome | $7.49 \times 10^{-8}***$  |
| <i>S. meliloti</i> Chromosome  | $-4.19 \times 10^{-7}***$ |
| <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 positions of the genome of the respective bacteria replicons. These genomic positions were split up into the coding and non-coding regions of the genome. 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 bidirectionality of 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          | Protein Coding                       |                           |  |                       |
|--------------------------------|--------------------------------------|---------------------------|--|-----------------------|
|                                | Correlation Coefficient<br>20kb Near |                           | Number of Substitutions<br>per 20kb Near |                       |
|                                | Origin                               | Terminus                  | Origin                                   | Terminus              |
| <i>E. coli</i> Chromosome      | NS                                   | NS                        | $5.62 \times 10^{-3}$                    | $6.66 \times 10^{-3}$ |
| <i>B. subtilis</i> Chromosome  | NS                                   | $-8.37 \times 10^{-5}***$ | $1.95 \times 10^{-3}$                    | $9.10 \times 10^{-3}$ |
| <i>Streptomyces</i> Chromosome | $7.91 \times 10^{-5}***$             | $-1.32 \times 10^{-4}***$ | $6.74 \times 10^{-4}$                    | $6.73 \times 10^{-3}$ |
| <i>S. meliloti</i> Chromosome  | $8.26 \times 10^{-5}*$               | NS                        | $9.79 \times 10^{-5}$                    | $5.07 \times 10^{-5}$ |
| <i>S. meliloti</i> pSymA       | NS                                   | NS                        | $9.75 \times 10^{-4}$                    | $3.23 \times 10^{-3}$ |
| <i>S. meliloti</i> pSymB       | $-1.44 \times 10^{-5}*$              | $-6.32 \times 10^{-5}***$ | $1.96 \times 10^{-3}$                    | $1.24 \times 10^{-3}$ |

Table 2: Logistic regression on 20kb closest and farthest from the origin of replication after accounting for bidirectional replication and outliers. 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          | Protein Coding             |                           |
|--------------------------------|----------------------------|---------------------------|
|                                | Weighted                   | Non-Weighted              |
| <i>E. coli</i> Chromosome      | $-2.28 \times 10^{-10}***$ | $-1.65 \times 10^{-4}***$ |
| <i>B. subtilis</i> Chromosome  | $-7.96 \times 10^{-10}**$  | $-1.73 \times 10^{-4}**$  |
| <i>Streptomyces</i> Chromosome | $2.38 \times 10^{-11}*$    | NS                        |
| <i>S. meliloti</i> Chromosome  | $-1.05 \times 10^{-10}***$ | $-1.24 \times 10^{-5}***$ |
| <i>S. meliloti</i> pSymA       | NS                         | NS                        |
| <i>S. meliloti</i> pSymB       | NS                         | NS                        |

Table 3: Linear regression on 10kb sections of the genome with increasing distance from the origin of replication after accounting for bidirectional replication. Weighted columns have the total number of substitutions in each 10kb section of the genome divided by the total number of protein coding and non-protein coding sites in the genome. Non-weighted columns are performing a linear regression on the total number of substitutions in each 10kb section of the genome. 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          | Coefficient Estimate      |
|--------------------------------|---------------------------|
| <i>E. coli</i> Chromosome      | $-2.32 \times 10^{-2}***$ |
| <i>B. subtilis</i> Chromosome  | $-1.93 \times 10^{-2}**$  |
| <i>Streptomyces</i> Chromosome | $-1.24 \times 10^{-3}***$ |
| <i>S. meliloti</i> Chromosome  | $-1.88 \times 10^{-2}***$ |
| <i>S. meliloti</i> pSymA       | $-2.50 \times 10^{-2}*$   |
| <i>S. meliloti</i> pSymB       | NS                        |

Table 4: Linear regression analysis of the total number of protein coding sites per 10kb along the genome of the respective bacteria replicons. Linear 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 bidirectionality of 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          | Gene/Genome Average |        |          |
|--------------------------------|---------------------|--------|----------|
|                                | dS                  | dN     | $\omega$ |
| <i>E. coli</i> Chromosome      | 0.2351              | 0.0101 | 0.0444   |
| <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  | 9.0094              | 0.0001 | 0.0015   |
| <i>S. meliloti</i> pSymA       | 0.0872              | 0.0099 | 0.1642   |
| <i>S. meliloti</i> pSymB       | 0.0940              | 0.0084 | 0.1142   |

Table 5: 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.