Subs Paper Things to Do:

- more genomes
- ~~new outgroups? (too distant)~~
- explain high dS values in *B. subtilis*
- potentially poor alignment and non-orthologous genes (core genome, change methods?)
- ~~non-parametirc analysis for subs~~
- gap in *Escherichia coli* fig 5
- ~~new methods for trees~~
- ~~concerned about repeated genes (TEs) and not analyzing core genome~~
- ~~check if trimming respects coding frame~~
- clear distinction between mutations and substitutions in intro (separate sections)
- ~~datasets from previous papers (repeat my analysis on them?)~~
- why would uncharacterized proteins have higher subs rates?
- ~~R^2 values in regression analysis~~
- ~~update gene exp paper ref~~
- 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?

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

Inversions + Gene Expression:

✓Queenie: comparing blast and gene alignment homologs

✓Queenie: start creating dataframe that is compatible with `limma`

Subst Paper:

✓finished re-running subst and selection analysis (with new RAxML trees)

Inversions + Gene Expression:

Queenie is still finishing things up but she is moving at a glacial pace. Which for now is fine because I am busy with the subst paper, but soon I will have to take over from her to get everything done. I had some issues with the gene mapping code but I fixed this. It was an issue where one gene is so large that it matches to multiple smaller genes in another taxa. For now I consider both of these a match and consider it a match to the blast results if blast also aligned either of those smaller genes. **What do you think about this?**

Substitution Paper I finished re-running the selection and subst analysis with the new RAxML trees. Everything is essentially the same, except for *B. subtilis* where the logistic regression changed sign. I think this is because of the outliers. I want to send you the paper and results once all the revisions are done, **but if you want it earlier, let me know!**

The outliers that were determined for the *B. subtilis* subst analysis seem a bit off. The short bars near the origin were considered outliers because they fall within the lower extreme end of the distribution (Figure 1). This loss of data I suspect is what has caused the overall sign for the *B. subtilis* number of subs and distance form the origin of replication to change (Table 1). **Do you**

think I should count these bars as outliers? Is it "wrong" to remove them when the same code was used to determine outliers for all the other replicons?

Looking at the selection values, *Streptomyces* and *S. meliloti* Chromosome have a lot of zero values (like last time) (Table 2 and Figures 2 and 3). However, previously ALL of the non-zero ω values for *S. meliloti* chromosome were considered outliers because of the large number of zero ω values. We therefore decided to re-do the selection analysis for *S. meliloti* chromosome without removing any outliers. Now with the new results (from the new RAxML trees) we have non-zero ω values that are not considered outliers, therefore we do not have the same problem as before. **Do you think it is necessary for me to re-do the *S. meliloti* chromosome selection analysis without removing outliers?** The only reason we did it before was because there were no non-zero ω values that were not outliers. Either way, I think I need to address the large amount of zero values because it skews the trend lines for dN, dS, and ω.

One reviewers comments was on the high dS values (> 10), particularly in the *B. subtilis* genome. I did address some of this in the supplement, but I guess it was not enough. I looked into these high values and there are 23 gene segments that have a $dS > 10$ in *B. subtilis*. These segments range from 105bp-327bp, our minimum length is 100bp so these are close to that minimum. I looked into the highest dS value and the alignment is not great, but is correct in the sense that it is what mauve and mafft have said should be aligned. It seems like even with our rigorous alignment trimming, some poor alignments still slip through the cracks. When I looked into the genes that are encoded in this segment of the genome, they align really really well (see email for attached Clustal Omega protein alignment) for some regions, but very poorly for others. It looks like some taxa have extra proteins while others are missing those proteins. I suspect this is what is causing the poor alignment. Based on the product names of the genes, it is hard to tell if they are truly similar. Some are listed as terminase, hypothetical proteins, phage like elements related to protein Xkdv. When looking at the genomic position of these proteins, they are all within the same 70,000bp of the genome in all taxa (the gene is about 2000bp long). **What do you think should be done about this?**

This brings me to BLAST. I wanted to include an extra check for the alignments by verifying the genes with the reciprocal best blast hit results from the same taxa, just like what I am doing for my inversions and gene expression analysis. However, this is proving to be very complicated. Sometimes there is no proteome available for the strains that I have, gene names are often missing from the genbank file (leaving me with no way to confirm if it is a match), and Queenie has been jumping through some hoops to get this all coded for me (and honestly I am not sure how scalable her code is with other taxa). We want to get the re-submission for this subst paper out ASAP and I am not sure how long (or how much of a headache) it will take me to implement blast into my pipeline, so I am wondering **if you think I should include this extra blast check? Should I say we checked a subset of the data with blast and things line up x% of the time so we think our alignment is good? I am not sure what to do.**

This Week (aka next week)

- Queenie: compare blast results and alignments

- Queenie: new dataframe for `limma`
- why do uncharacterized proteins have higher sub rates?
- *B. subtilis* high dS values should not be present
- start blast to confirm homologs in subst analysis?
- distinction between mutations and substitutions in subst paper intro
- update new code on git (subst paper)

Next Week (aka 2 weeks)

- finish blast to confirm homologs in subst analysis
- final edits of subst paper
- send subst paper to Brian to review
- make ↑ changes to subst paper
- submit subst paper!

Bacteria and Replicon	Protein Coding Sequences	
	Coefficient Estimate	R^2
E. coli Chromosome	$-2.66\times10^{-8***}$	
B. subtilis Chromosome	$2.76\times10^{-8***}$	
Streptomyces Chromosome	$7.21\times10^{-8***}$	
S. meliloti Chromosome	$-6.57\times10^{-7***}$	
S. meliloti pSymA	$2.74\times10^{-7***}$	
S. meliloti pSymB	$1.10\times10^{-7***}$	

Table 1: one reviewer requested R^2 values for the regressions. For a logistic regression, the R^2 value is not explicitly calculated by the glm() function. Should I calculate this myself? Or do you think the reviewer only wanted the R^2 value on the linear regressions? 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	Outliers (%)	Zero Value (%)		
		dN	dS	ω
E. coli Chromosome	7.49	13.82	1.05	13.82
B. subtilis Chromosome	5.41	4.40	0.16	4.40
Streptomyces Chromosome	4.74	25.70	14.48	25.70
S. meliloti Chromosome	17.05	61.21	59.26	61.21
S. meliloti pSymA	6.69	11.28	9.75	11.28
S. meliloti pSymB	6.13	13.20	5.20	13.20

Table 2: Percent of data that was calculated to be an outlier or had a selection variable (dN, dS, and ω) value of zero.

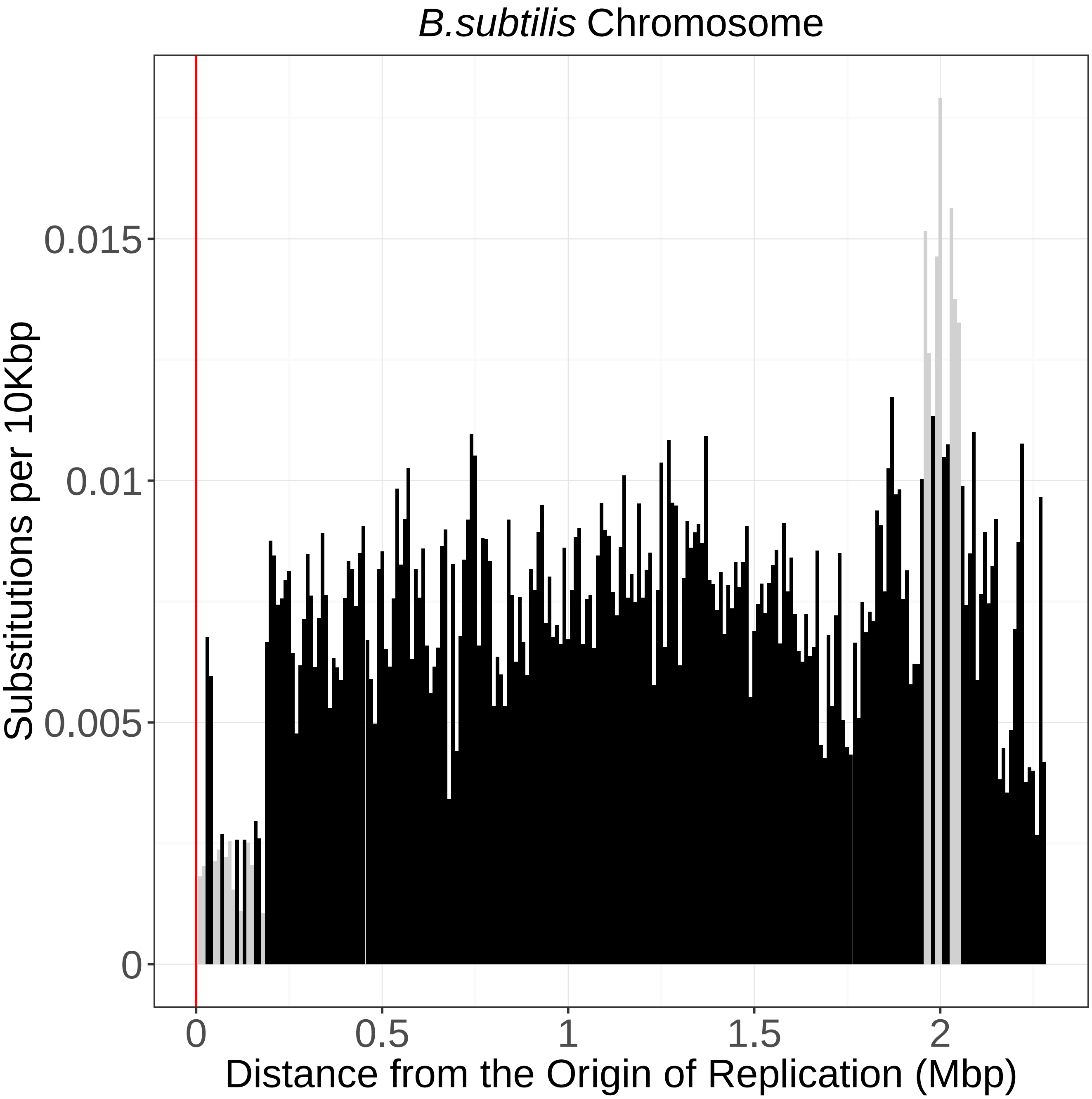

Figure 1: *B. subtilis* subs graph

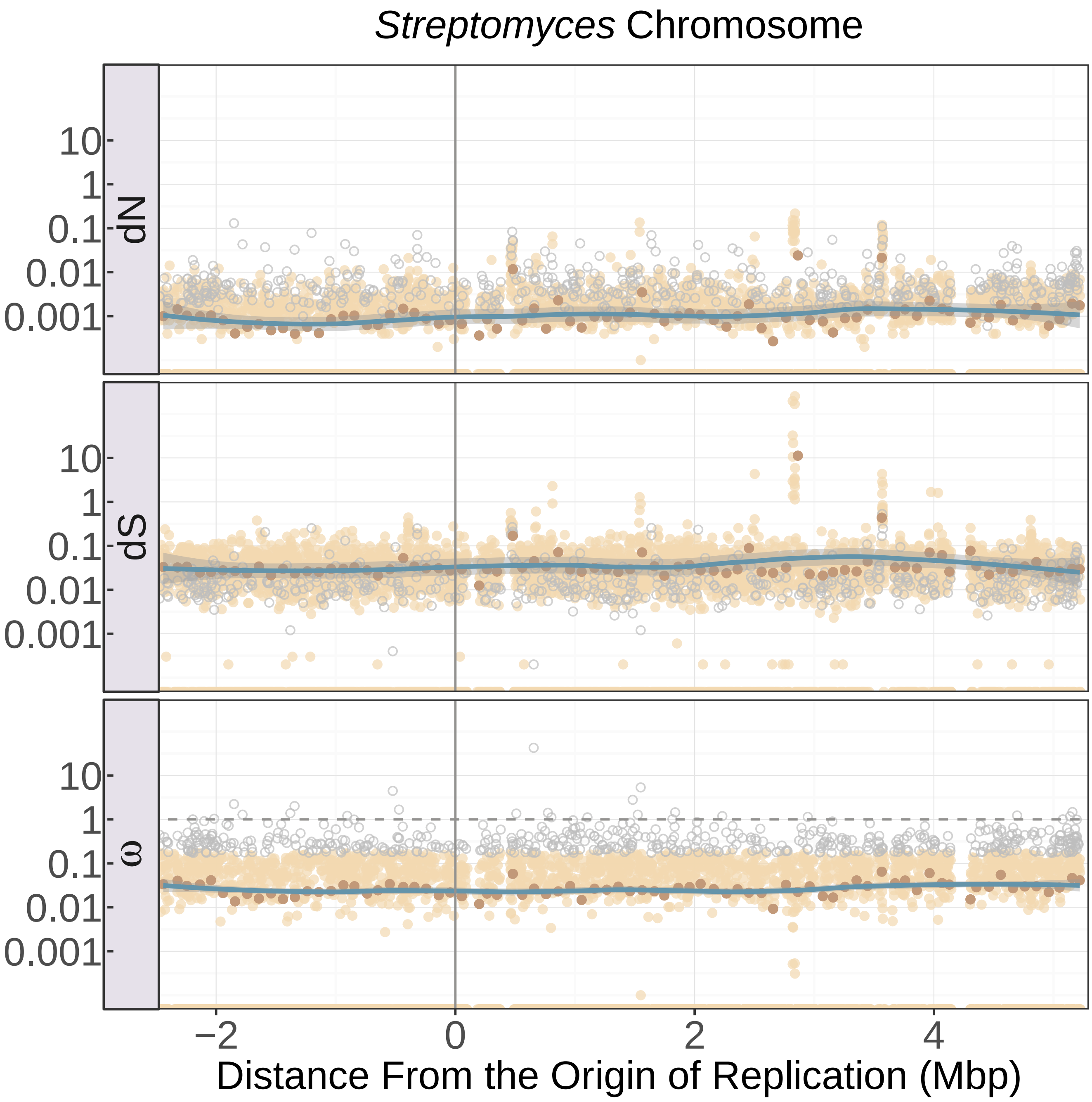
Streptomyces Chromosome
dN
dS
ω
10
1
0.1
0.01
0.001
−2
0
2
4
Distance From the Origin of Replication (Mbp)

Figure 2

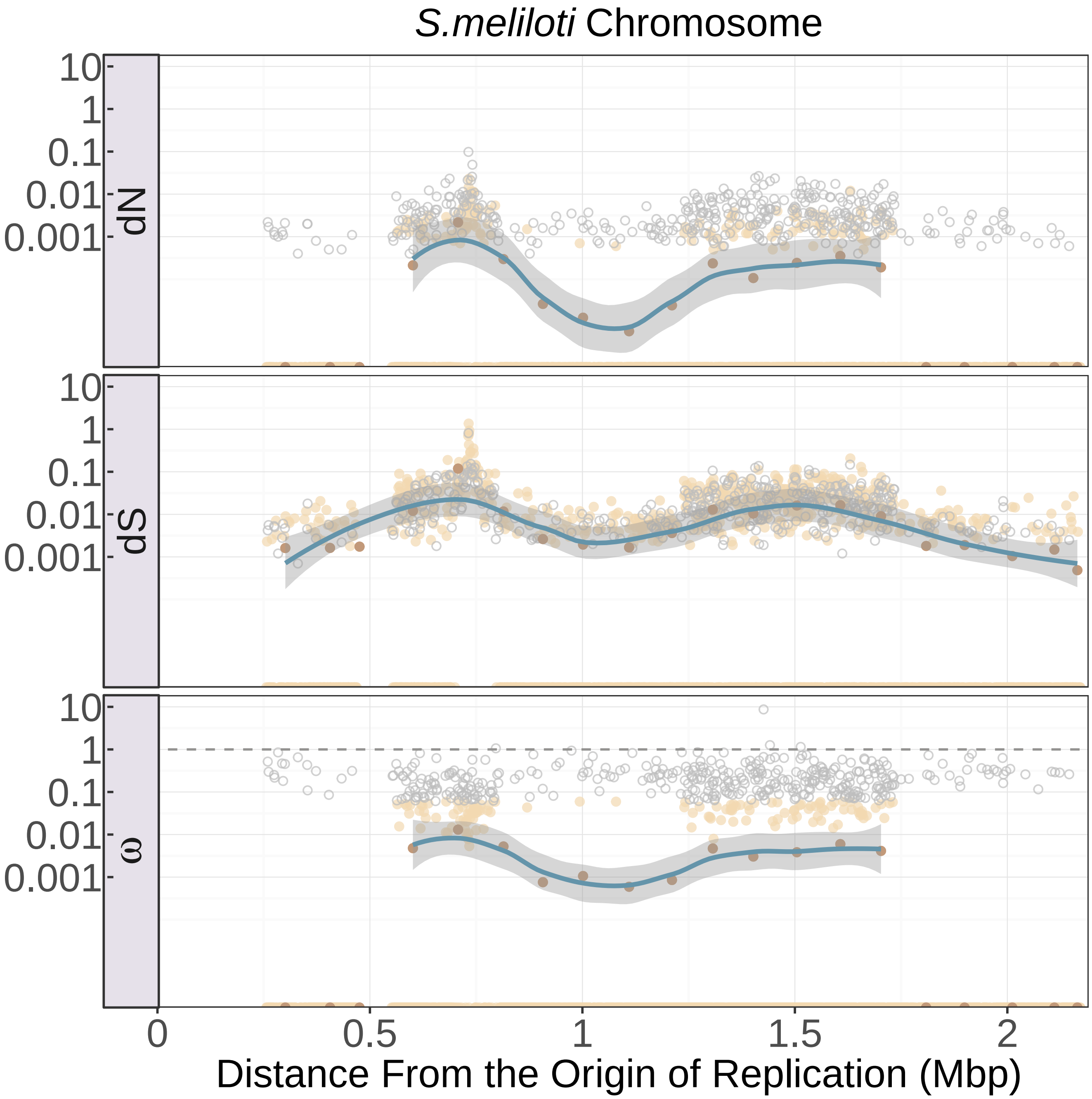
S.meliloti Chromosome
dN
dS
ω
10
1
0.1
0.01
0.001
0
0.5
1
1.5
2
Distance From the Origin of Replication (Mbp)

Figure 3