

Subs Paper Things to Do:

- Or get 1st, 2nd, 3rd codon pos log regs
- ~~write dN/dS methods~~
- ~~write dN/dS results~~
- ~~write dN/dS discussion~~
- ~~write dN/dS into conclusion~~
- mol clock for my analysis?
- GC content? COG? where do these fit?

Gene Expression Paper Things to Do:

- ~~write abstract~~
- ~~write intro~~
- ~~add stuff from outline to Data section~~
- ~~create graphs for expression distribution (no sub data)~~
- ~~add # of genes to expression graphs (top)~~
- ~~average gene expression~~
- ~~write discussion~~
- ~~write conclusion~~
- ~~add into methods: filters for Hiseq, RT-PCR and growth phases for data collection~~
- ~~update supplementary figures/file~~

Inversions and Gene Expression Letter Things to Do:

- ~~check if opposite strand in progressiveMauve means an inversions (check visual matches the xmfa)~~
- ~~check if PARSNP and progressiveMauve both identify the same inversions (check xmfa file)~~
- ~~create latex template for paper~~
- ~~put notes from papers into doc~~
- ~~use large PARSNP alignment to identify inversions~~

- confirm inversions with dot plot
- 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

Last Week

✓sent you a draft of my poster for SMBE

As I mentioned before, the GUI for Dotter is really bad and it constantly freezes. So I am still working through how to save plots where the contrast is good enough to define the inversions. This is really frustrating and long but oh well!

As I am re-running the selection and substitutions stuff I am finding more and more issues (most of them small things, which causes my previous code to not work so I have to tweak file formats) which is very very frustrating but I am slowly working through it.

I also calculated the average gene expression per replicon for fun, this is found in Table 1. *Streptomyces* is like 2 orders of magnitude lower than everything else..which is weird so I am not sure what is going on there. Do you think this is something that needs to be put into the gene expression paper?

I have also been working to put the dN , dS , and ω values for each gene into a supplementary table on github. This is slowly getting done.

I was also wondering if I should be fitting a regression to the dN , dS , and ω data to see how those three values change (if at all) with genomic position? although to me the graphs look pretty non-linear. Thoughts?

This Week

I hope to have the re-running of the selection and substitutions analysis done by the end of the week (hopefully!!! but there are always issues that are slowing down my progress).

I would also like to have a dot plot for the inversions and gene expression analysis.

I also need to work on your comments on my poster for the conference.

Next Week

Continue working on poster for conference + print it!

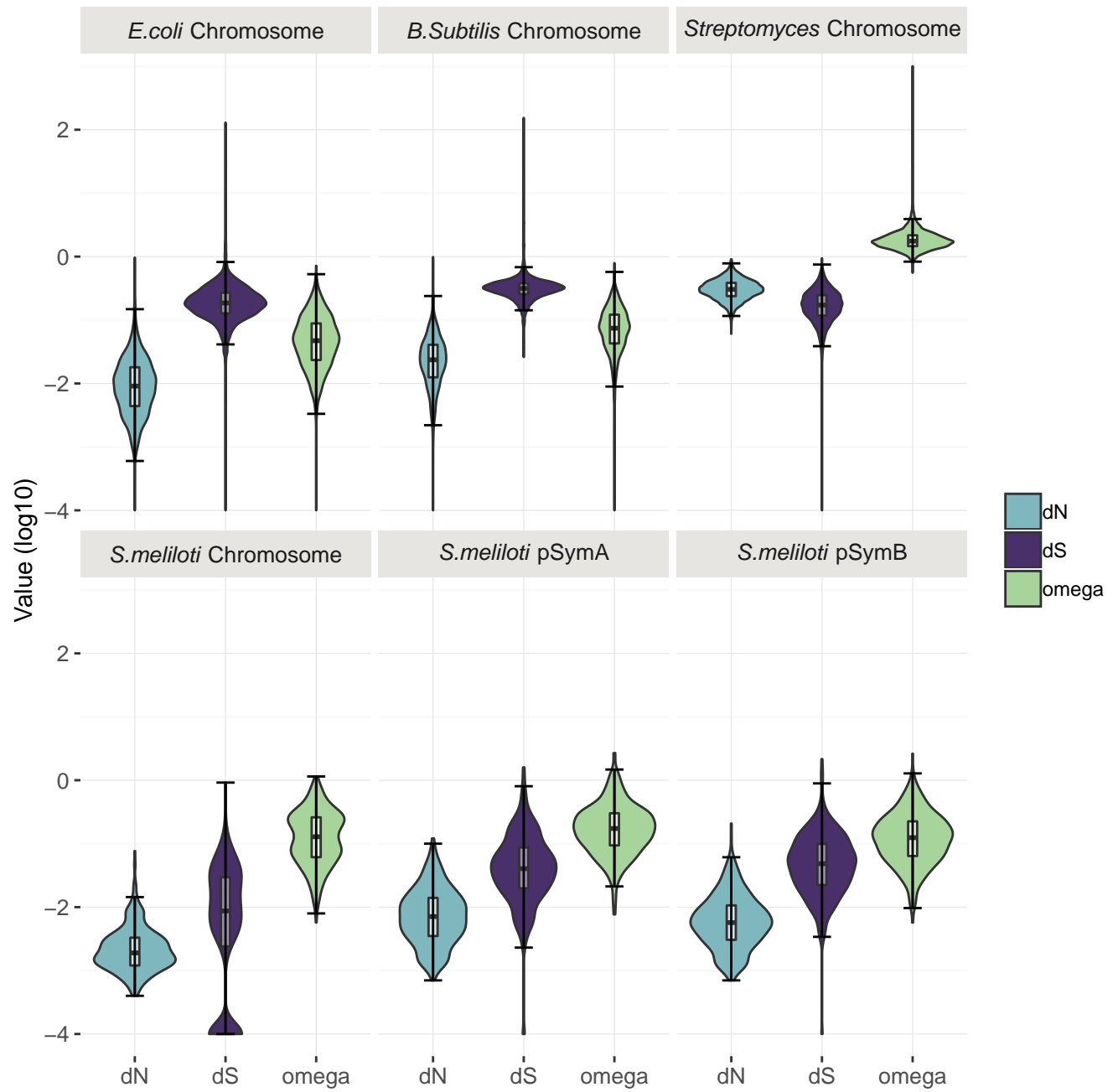
Make edits to the papers.

Move on to next steps for the inversions and gene expression analysis

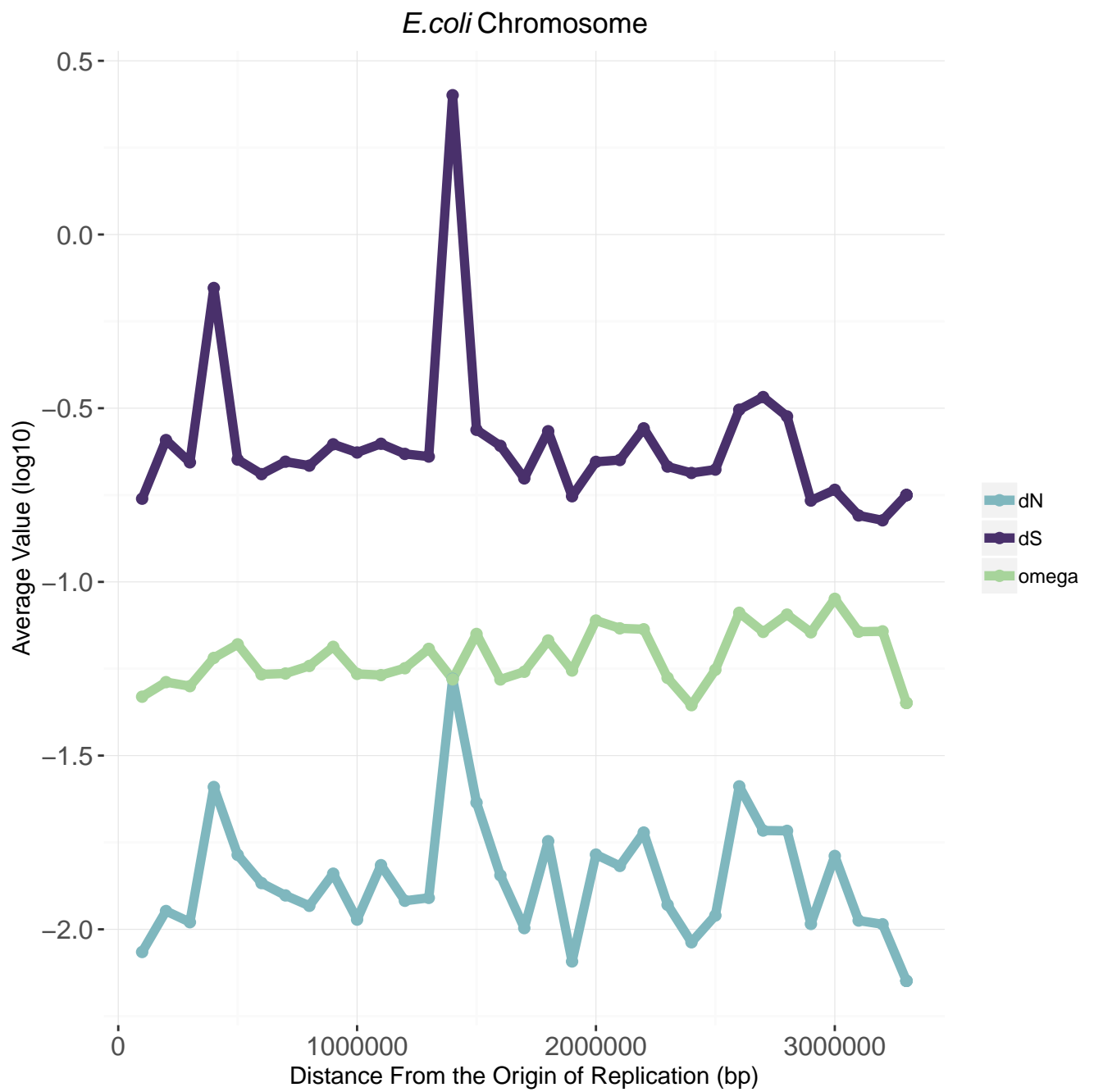
Bacteria and Replicon	Average Expression Value (CPM)
<i>E. coli</i> Chromosome	160.500
<i>B. subtilis</i> Chromosome	176.400
<i>Streptomyces</i> Chromosome	6.084
<i>S. meliloti</i> Chromosome	271.400
<i>S. meliloti</i> pSymA	690.100
<i>S. meliloti</i> pSymB	595.700

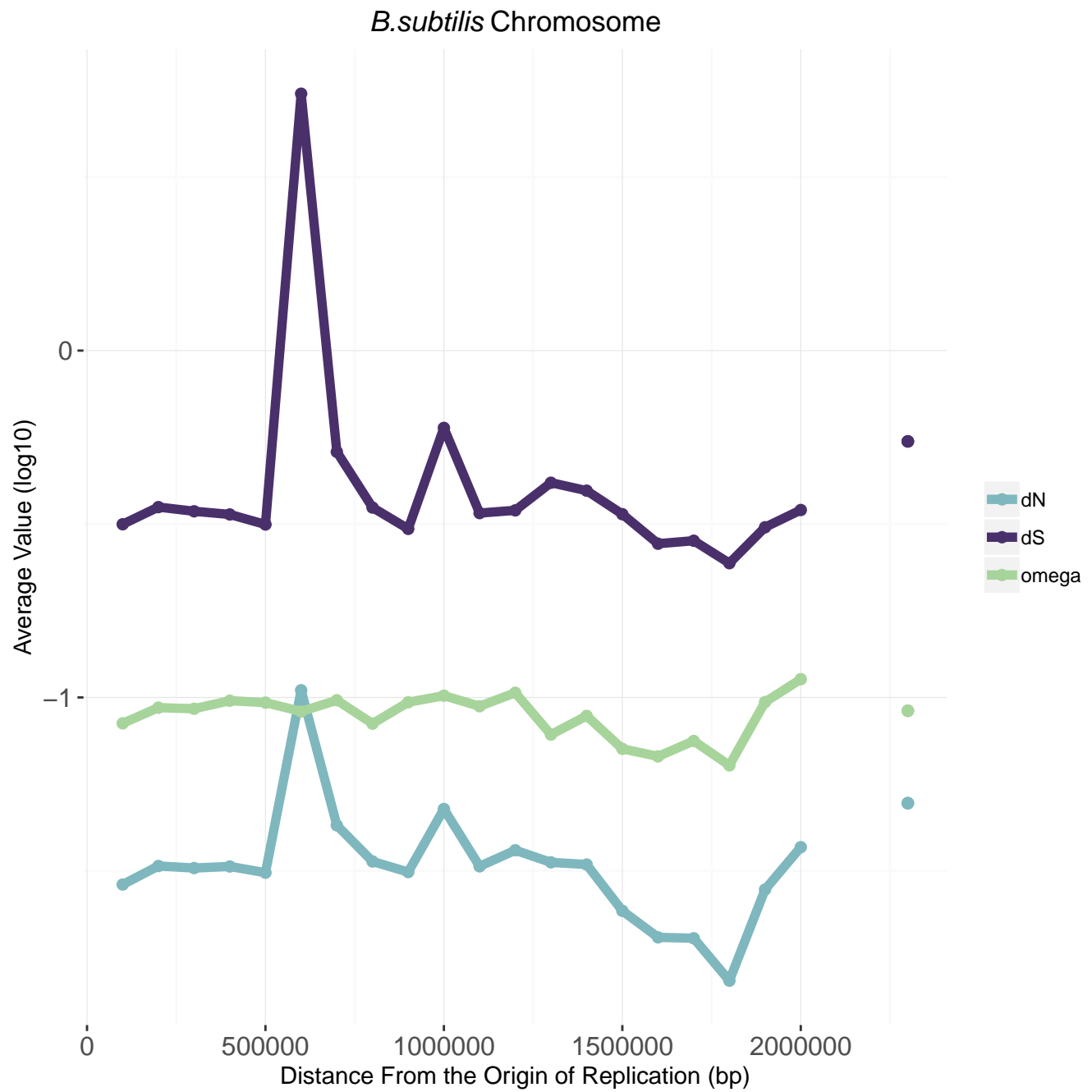
Table 1: Arithmetic gene expression calculated across all genes in each replicon. Expression values are represented in Counts Per Million.

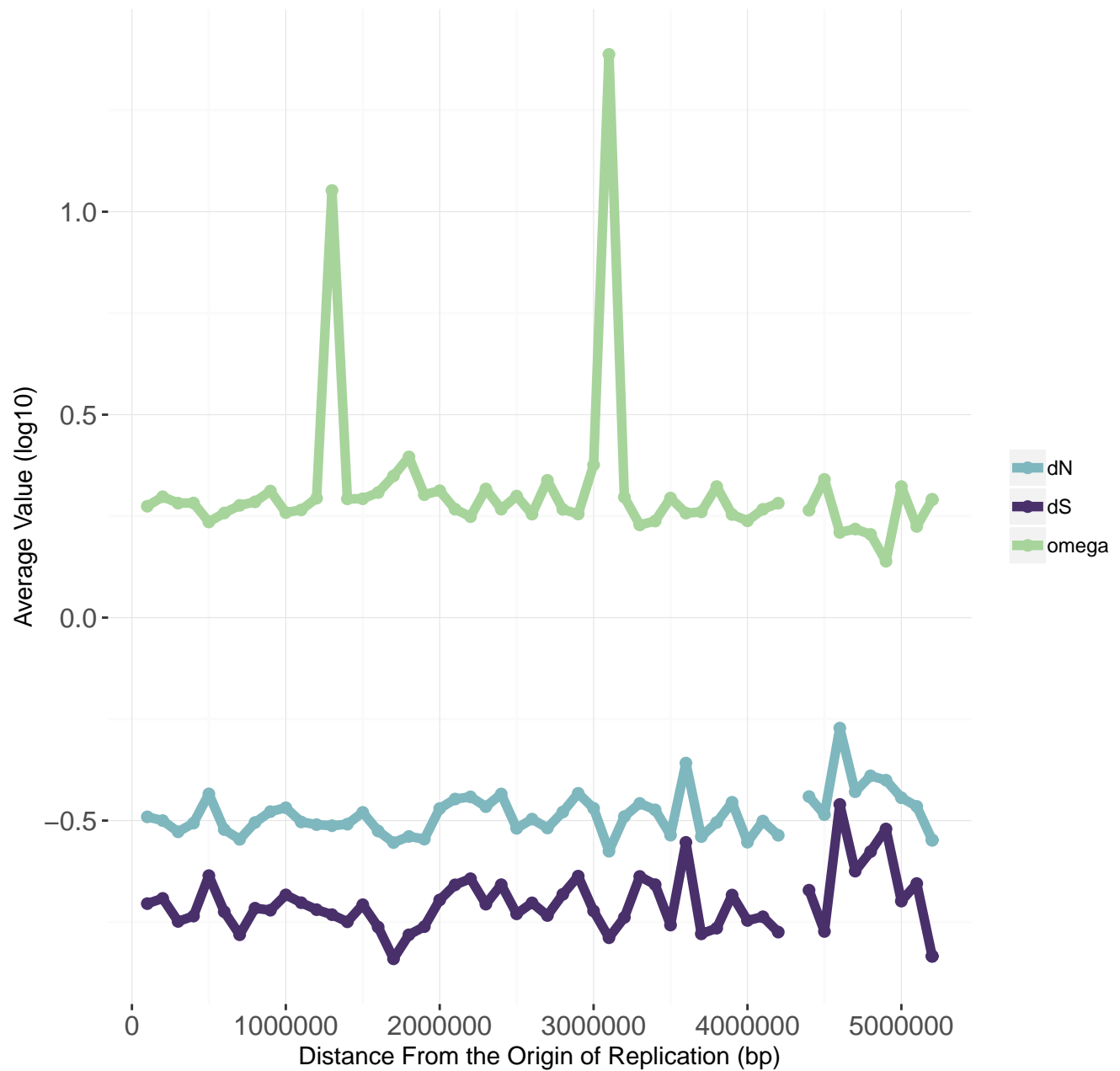
Violin plots for per gene dN, dS, and ω :

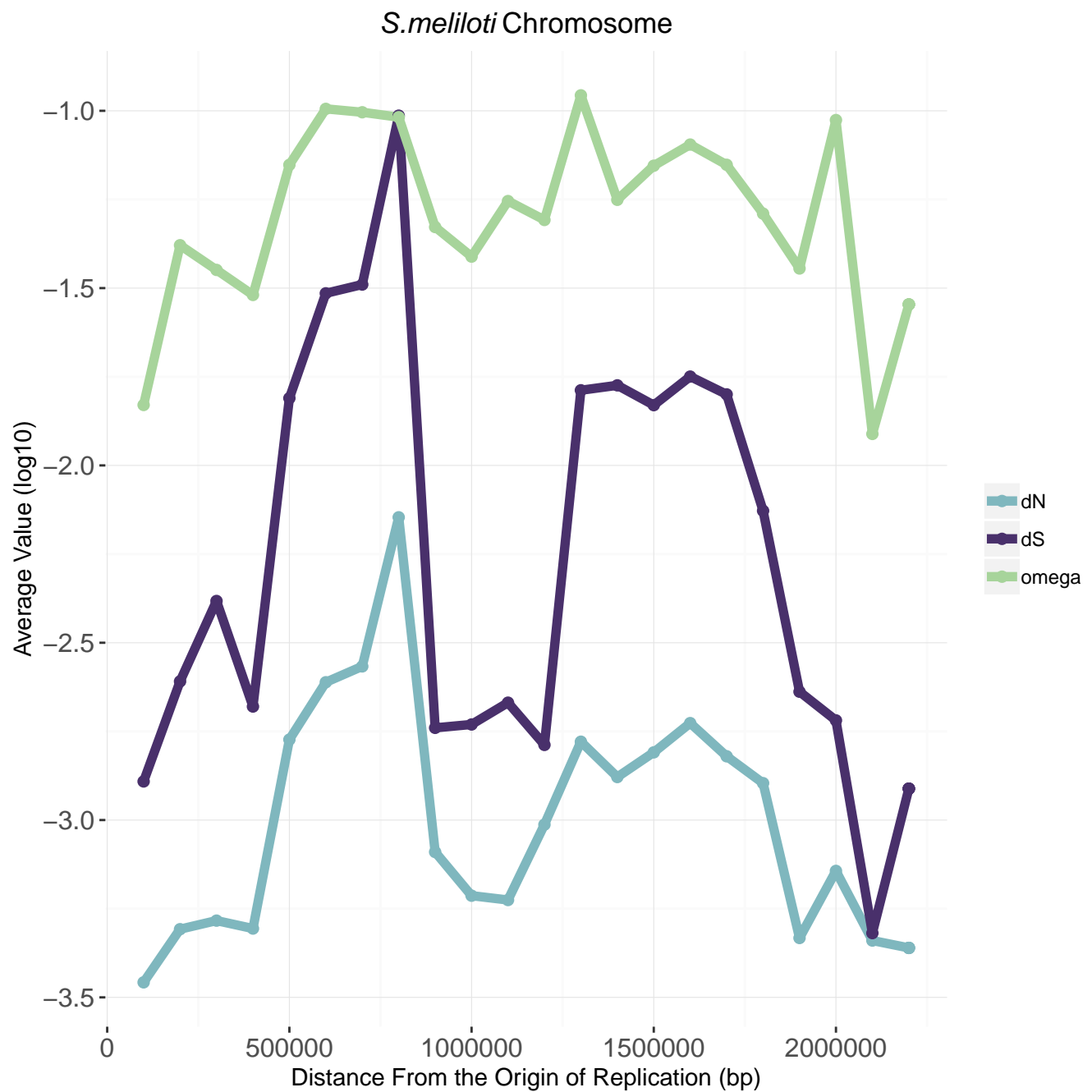


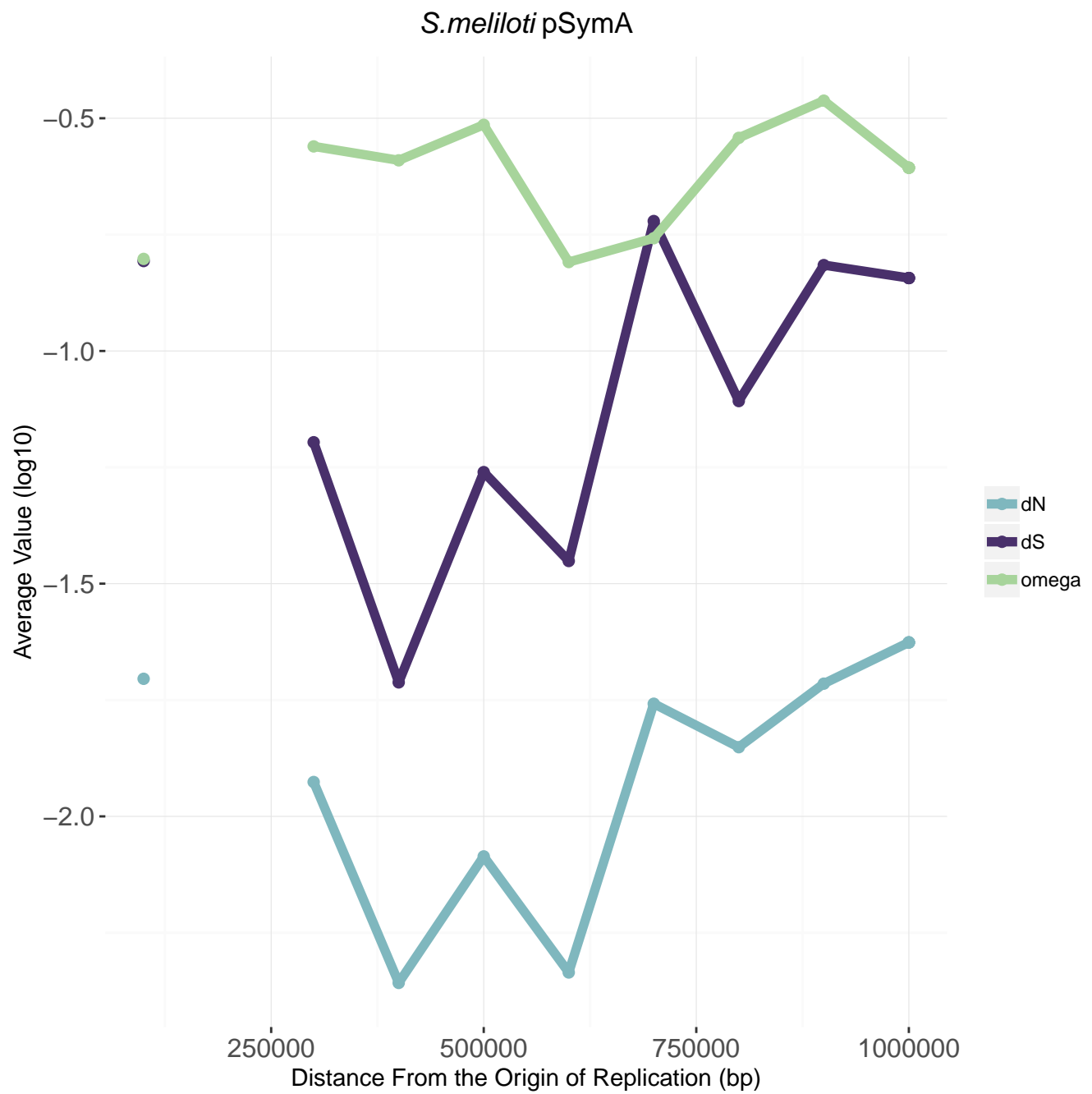
Genome Distribution for per 10kb dN, dS, and ω averages:

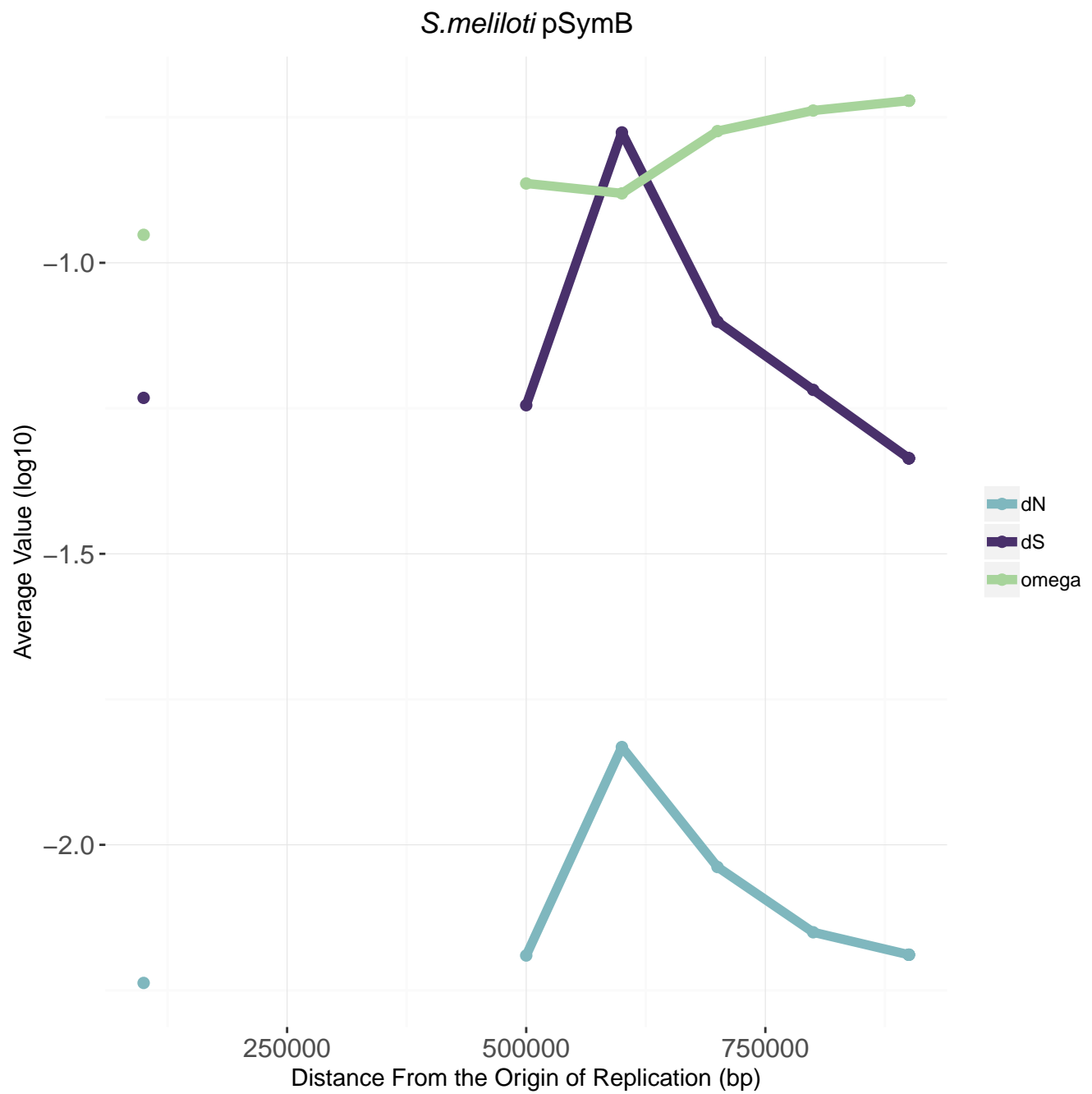




Streptomyces Chromosome







Bacteria and Replicon	Gene Average			Genome Average		
	dS	dN	ω	dS	dN	ω
<i>E. coli</i> Chromosome	0.2924	0.0144	0.0604	0.2600	0.0133	0.0556
<i>B. subtilis</i> Chromosome	0.6526	0.0358	0.0891	0.5267	0.0321	0.0828
<i>Streptomyces</i> Chromosome	0.1924	0.3201	2.6404	0.1775	0.3017	2.4358
<i>S. meliloti</i> Chromosome	0.0134	0.0014	0.0844	0.0134	0.0013	0.0930
<i>S. meliloti</i> pSymA	0.0798	0.0109	0.2320	0.0800	0.0103	0.2218
<i>S. meliloti</i> pSymB	0.0814	0.0086	0.1639	0.0782	0.0082	0.1590

Table 2: 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.

Bacteria and Replicon	Average Replicon Length	# of Coding Sites	# of Non-Coding Sites	# of Subs Coding	# of Subs Non-Coding
<i>E. coli</i> Chromosome	5082529	2960007	191748	207199	9534
<i>B. subtilis</i> Chromosome	4077077	2074653	102906	205150	6187
<i>Streptomyces</i> Chromosome	8497577	2422980	21581	551530	3670
<i>S. meliloti</i> Chromosome	3426881	1931139	199425	6684	842
<i>S. meliloti</i> pSymA	1455940	419223	34213	9832	943
<i>S. meliloti</i> pSymB	1664597	552816	22098	11699	645

Table 3: Total proportion of coding and non-coding sites in each replicon and the percentage of those sites that have a substitution (multiple substitutions at one site are counted as two substitutions).

Bacteria and Replicon	Coding Sequences	Non-Coding Sequences
<i>E. coli</i> Chromosome	$-9.983 \times 10^{-8***}$	$6.994 \times 10^{-8***}$
<i>B. subtilis</i> Chromosome	$-1.071 \times 10^{-7***}$	$-9.861 \times 10^{-8***}$
<i>Streptomyces</i> Chromosome	$-2.626 \times 10^{-8***}$	$3.615 \times 10^{-7***}$
<i>S. meliloti</i> Chromosome	$-1.367 \times 10^{-7***}$	$-1.510 \times 10^{-7*}$
<i>S. meliloti</i> pSymA	$-1.075 \times 10^{-7*}$	NS
<i>S. meliloti</i> pSymB	$2.878 \times 10^{-7***}$	$8.595 \times 10^{-7***}$

Table 4: 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 Strain/Species	GEO Accession Number	Date Accessed
<i>E. coli</i> K12 MG1655	GSE60522	December 20, 2017
<i>E. coli</i> K12 MG1655	GSE73673	December 19, 2017
<i>E. coli</i> K12 MG1655	GSE85914	December 19, 2017
<i>E. coli</i> K12 MG1655	GSE40313	November 21, 2018
<i>E. coli</i> K12 MG1655	GSE114917	November 22, 2018
<i>E. coli</i> K12 MG1655	GSE54199	November 26, 2018
<i>E. coli</i> K12 DH10B	GSE98890	December 19, 2017
<i>E. coli</i> BW25113	GSE73673	December 19, 2017
<i>E. coli</i> BW25113	GSE85914	December 19, 2017
<i>E. coli</i> O157:H7	GSE46120	August 28, 2018
<i>E. coli</i> ATCC 25922	GSE94978	November 23, 2018
<i>B. subtilis</i> 168	GSE104816	December 14, 2017
<i>B. subtilis</i> 168	GSE67058	December 16, 2017
<i>B. subtilis</i> 168	GSE93894	December 15, 2017
<i>B. subtilis</i> 168	GSE80786	November 16, 2018
<i>S. coelicolor</i> A3	GSE57268	March 16, 2018
<i>S. natalensis</i> HW-2	GSE112559	November 15, 2018
<i>S. meliloti</i> 1021 Chromosome	GSE69880	December 12, 2017
<i>S. meliloti</i> 2011 pSymA	NC_020527 (Dr. Finan)	April 4, 2018
<i>S. meliloti</i> 1021 pSymA	GSE69880	November 15, 18
<i>S. meliloti</i> 2011 pSymB	NC_020560 (Dr. Finan)	April 4, 2018
<i>S. meliloti</i> 1021 pSymB	GSE69880	November 15, 18

Table 5: Summary of strains and species found for each gene expression analysis. Gene Expression Omnibus accession numbers and date accessed are provided.

Bacteria and Replicon	Coefficient Estimate	Standard Error	P-value
<i>E. coli</i> Chromosome	-6.03×10^{-5}	1.28×10^{-5}	2.8×10^{-6}
<i>B. subtilis</i> Chromosome	-9.7×10^{-5}	2.0×10^{-5}	1.2×10^{-6}
<i>Streptomyces</i> Chromosome	-1.17×10^{-6}	1.04×10^{-7}	$< 2 \times 10^{-16}$
<i>S. meliloti</i> Chromosome	3.97×10^{-5}	4.25×10^{-5}	NS (3.5×10^{-1})
<i>S. meliloti</i> pSymA	1.39×10^{-3}	2.53×10^{-4}	4.9×10^{-8}
<i>S. meliloti</i> pSymB	1.46×10^{-4}	2.03×10^{-4}	NS ($5.34.7 \times 10^{-1}$)

Table 6: Linear regression analysis of the median counts per million expression data 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.