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

✓fixed/looked at weird points in the genome distribution of dN, dS, and ω graphs

✓write abstract for gene expression paper

✓write code to get the gene name associated with each dN, dS, and ω value

✓calculate average gene expression value for each replicon and gene expression paper

✓look into dot plot for inversions paper

I looked into the weird points of the distribution of dN, dS, and ω across the genome and the points where dS is higher than $dN = \omega$ are real, and the high number of $dS = 0$ points in *S. meliloti* chrom is also real. We discussed this and you said to just leave everything the way it is.

I looked into a number of programs that create a dot plot from an alignment or two genomes. There are lots of different programs like MUMerplot (which needs its own alignment), D-GENIES (only AA seqs), Dotlet (no command line), Dotter, Dotplot (need to install, makes its own alignment), and LAST (uses its own alignment, good for big data but talks about mapping reads..). Dotter is already on the machines so I was playing with that. To align 2 whole *Escherichia coli* genomes took 7 days in Dotter. I am still working out how to save the resulting dot plot and make it look nice. The GUI seems to freeze sometimes so this is difficult. But I am working on it!

I also started looking into why pSymB is missing so much data, and why *Streptomyces* had $dN > dS$ for the whole genome. I think these two may be related to a small issue in my code. I am still testing this now.

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 almost 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?

I have also pretty much finished paper drafts for the Substitutions paper and the gene expression paper.

This Week

Reminder that I will be away from June 14-19th camping in Calgary

I need to continue to look into the weird things about the selection distribution plots.

Continue working on the inversions and gene expression analysis, by confirming inversions with a dot plot.

Send you paper drafts for the substitutions paper and the gene expression paper.

Next Week

Start thinking about poster for SMBE.

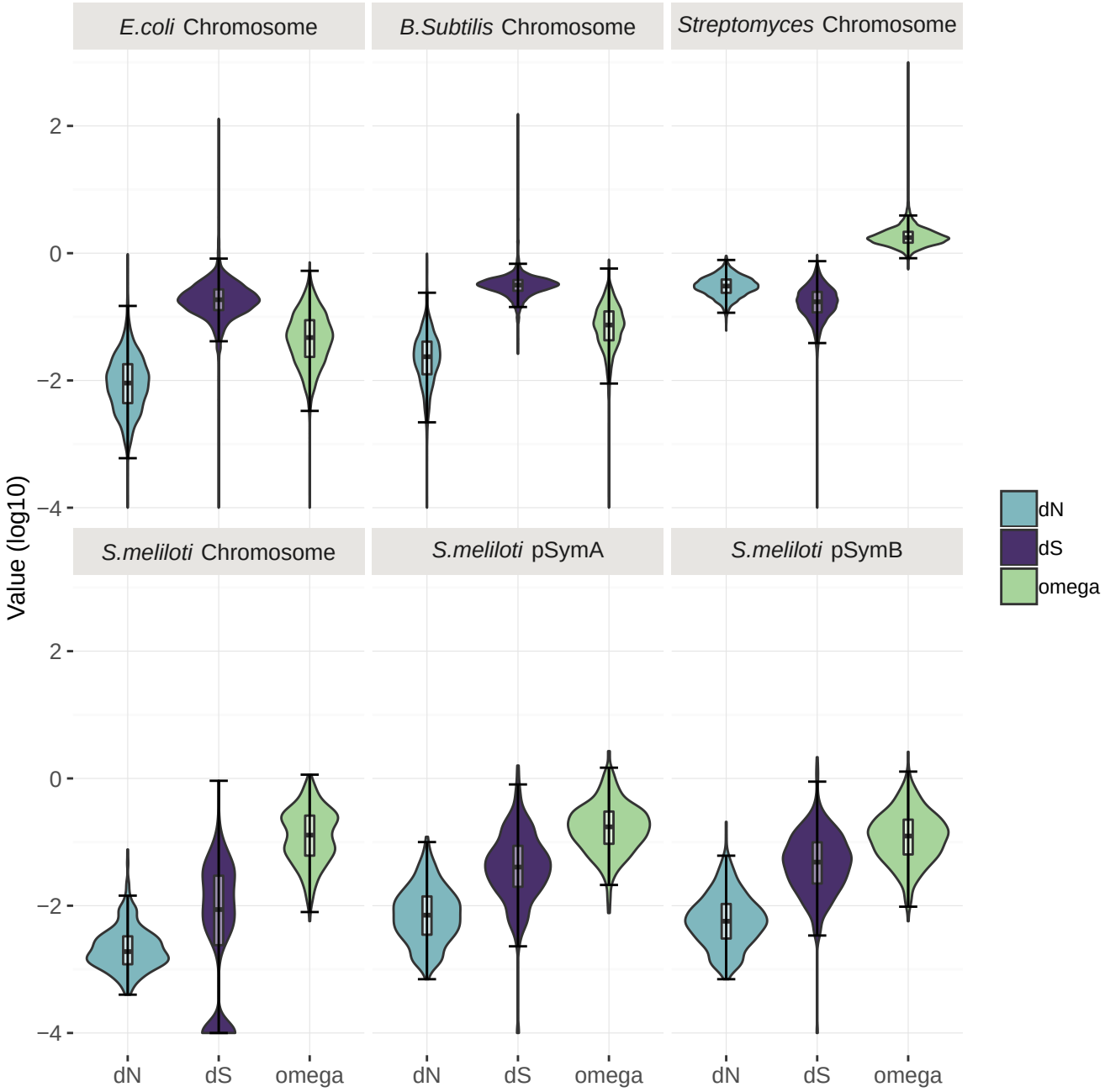
Continue working on inversions and gene expression

Make any necessary edits to the papers.

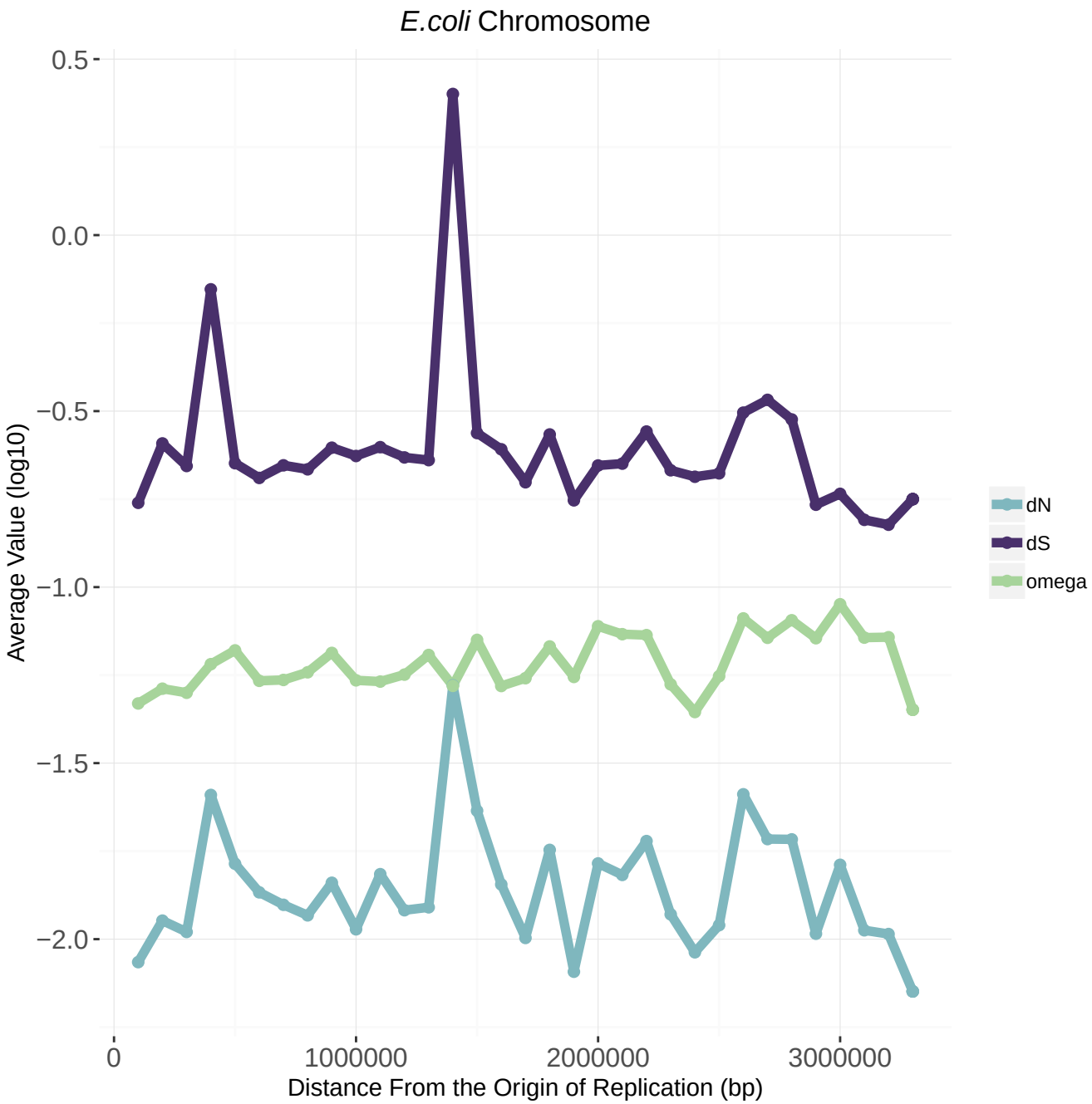
Bacteria and Replicon	Average Expression Value (CPM)
E. coli Chromosome	160.500
B. subtilis Chromosome	176.400
Streptomyces Chromosome	6.084
S. meliloti Chromosome	271.400
S. meliloti pSymA	690.100
S. meliloti pSymB	595.700

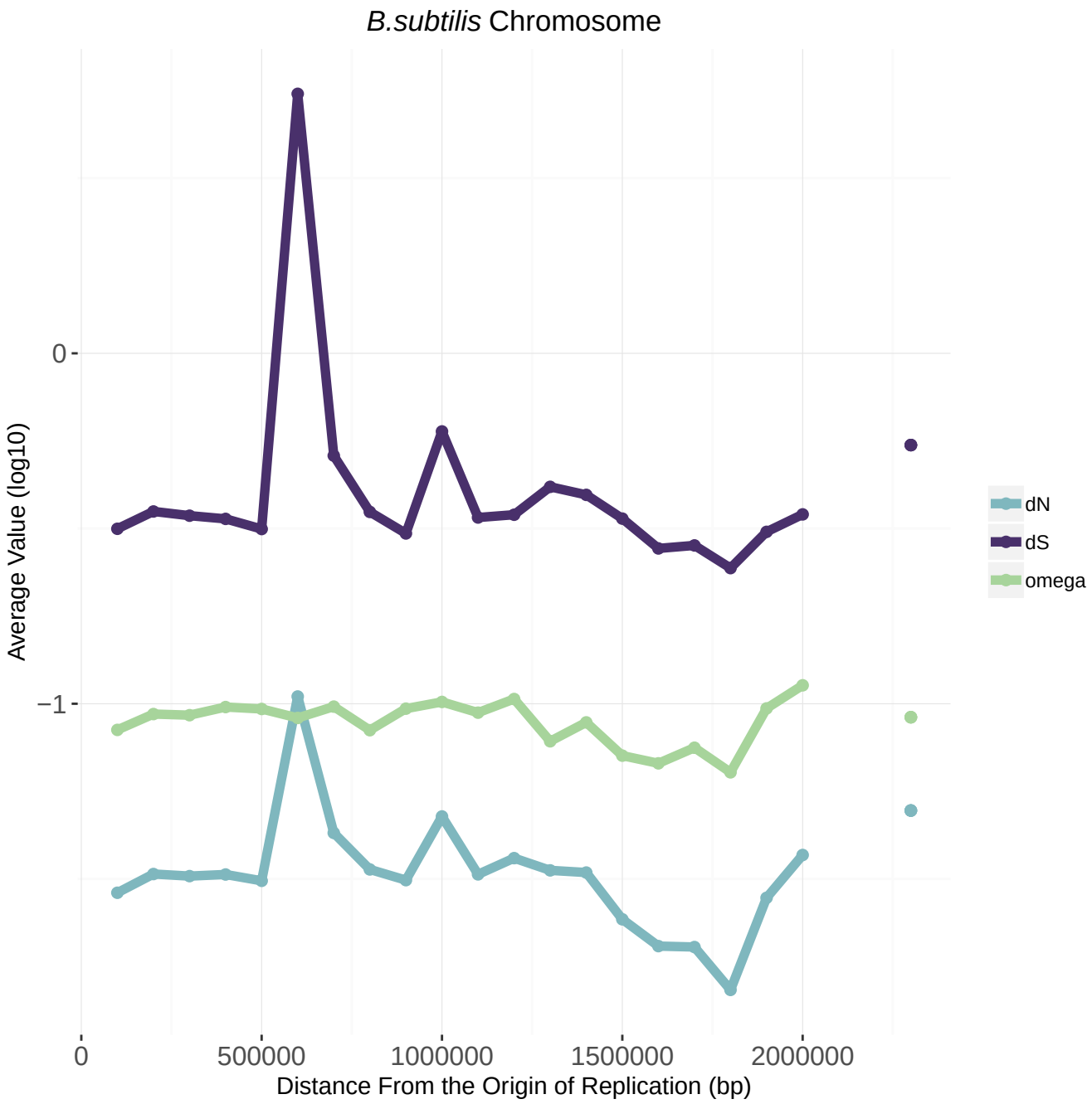
Table 1: Arithmetic gene expression calculated across all genes in each replicon. Expression values are represented in Counts Per Million.

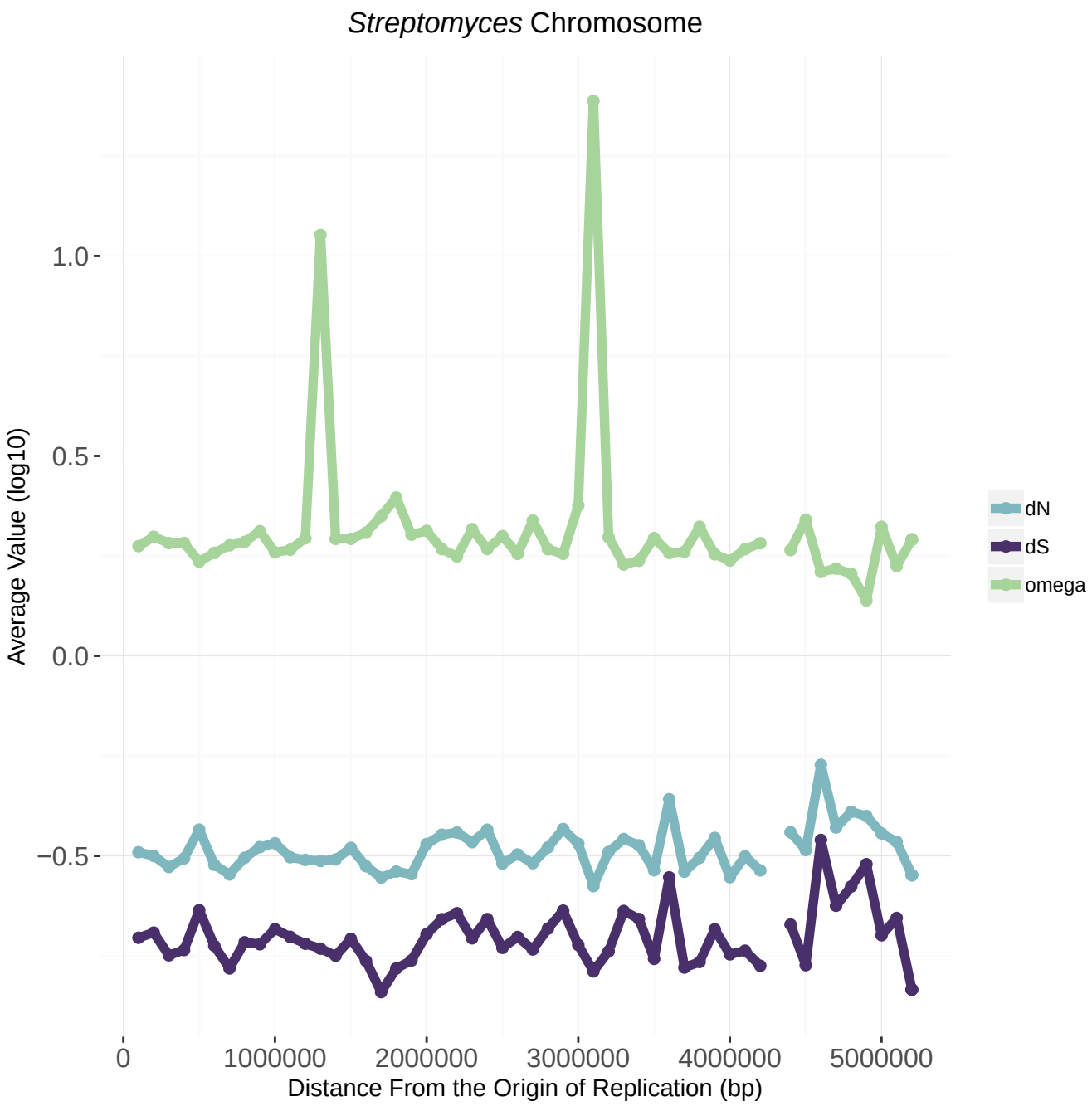
Violin plots for per gene dN, dS, and ω:

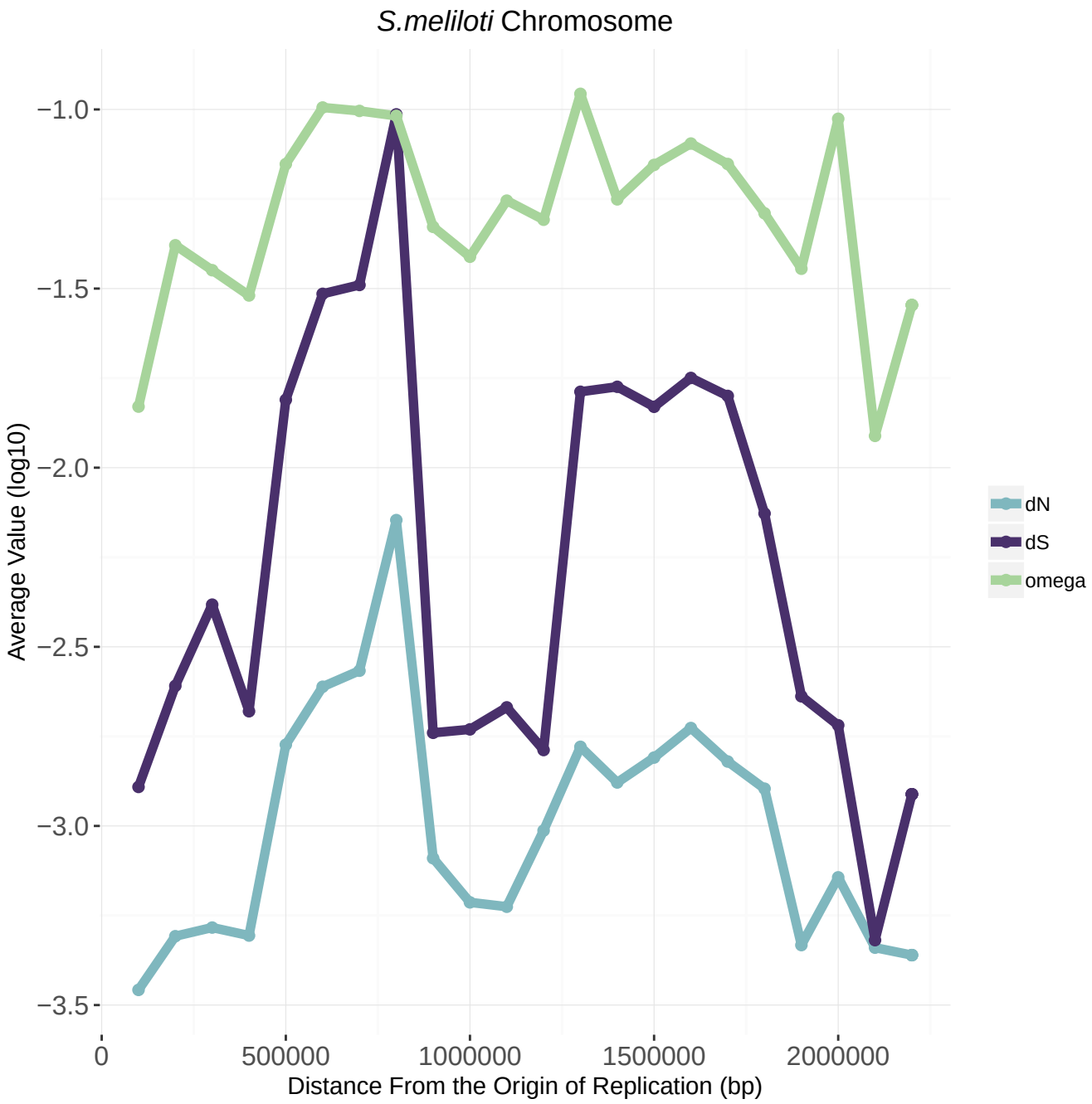


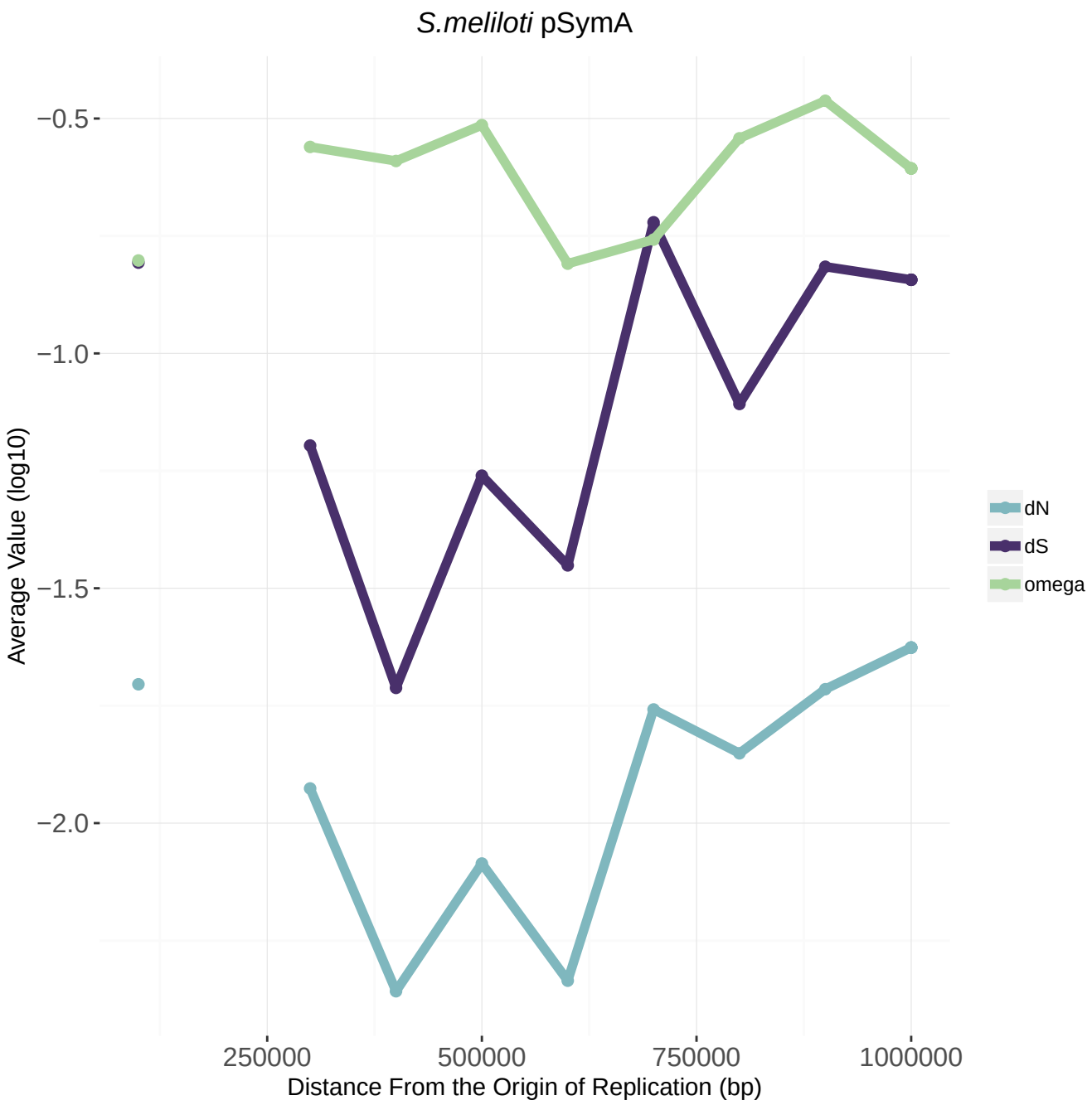
Genome Distribution for per 10kb dN, dS, and ω averages:

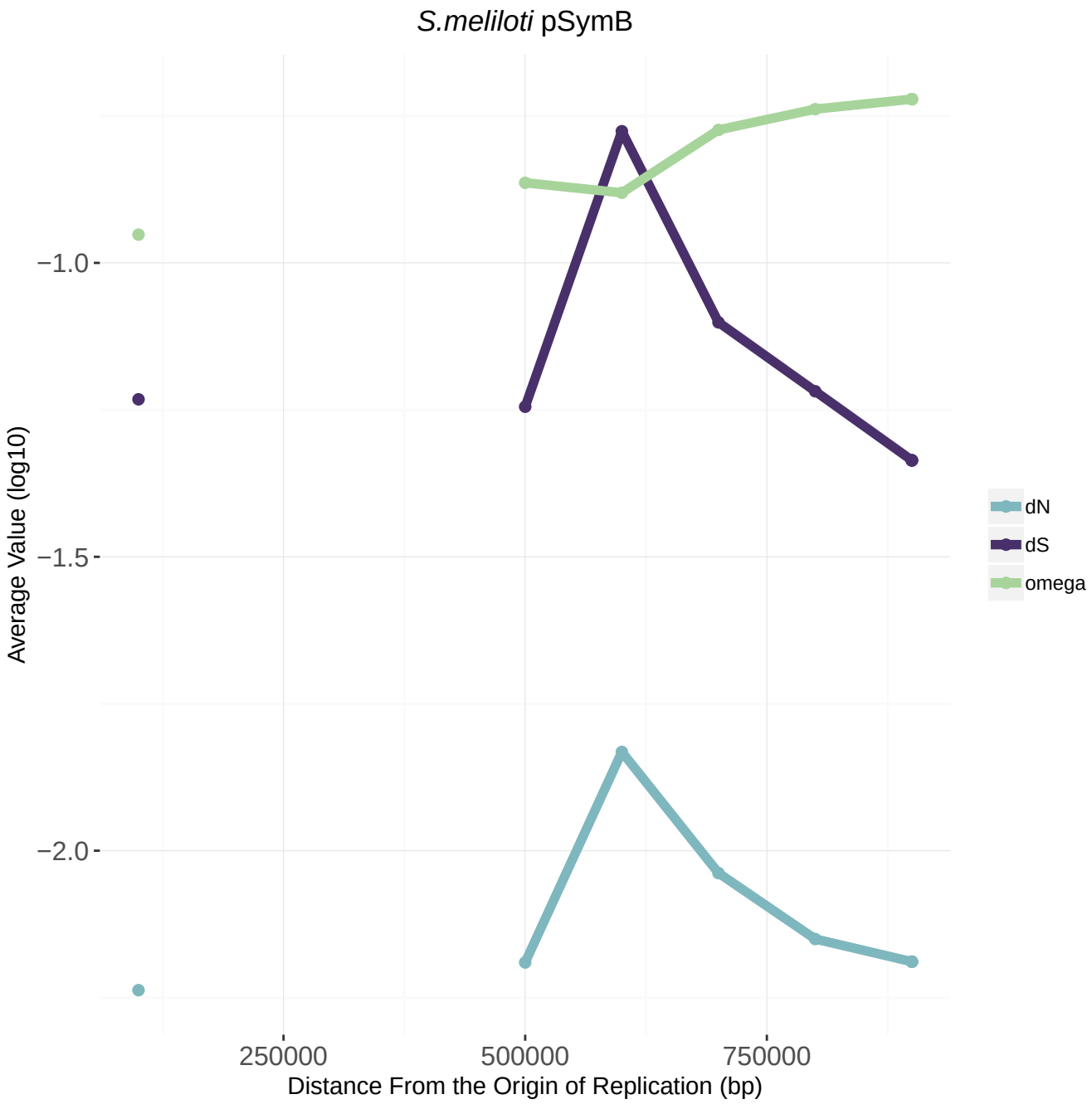

E.coli Chromosome
0.5
0.0
−0.5
−1.0
−1.5
−2.0
Average Value (log10)
0
1000000
2000000
3000000
Distance From the Origin of Replication (bp)
dN
dS
omega


B.subtilis Chromosome
Average Value (log10)
0
−1
0
500000
1000000
1500000
2000000
Distance From the Origin of Replication (bp)
dN
dS
omega


Streptomyces Chromosome
Average Value (log10)
1.0
0.5
0.0
−0.5
0
1000000
2000000
3000000
4000000
5000000
Distance From the Origin of Replication (bp)
dN
dS
omega


S.meliloti Chromosome
−1.0
−1.5
−2.0
−2.5
−3.0
−3.5
Average Value (log10)
0
500000
1000000
1500000
2000000
Distance From the Origin of Replication (bp)
dN
dS
omega


S.meliloti pSymA
−0.5
−1.0
−1.5
−2.0
Average Value (log10)
250000
500000
750000
1000000
Distance From the Origin of Replication (bp)
dN
dS
omega


S.meliloti pSymB
−1.0
−1.5
−2.0
Average Value (log10)
250000
500000
750000
Distance From the Origin of Replication (bp)
dN
dS
omega

Bacteria and Replicon	Gene Average			Genome Average		
	dS	dN	ω	dS	dN	ω
E. coli Chromosome	0.2924	0.0144	0.0604	0.2600	0.0133	0.0556
B. subtilis Chromosome	0.6526	0.0358	0.0891	0.5267	0.0321	0.0828
Streptomyces Chromosome	0.1924	0.3201	2.6404	0.1775	0.3017	2.4358
S. meliloti Chromosome	0.0134	0.0014	0.0844	0.0134	0.0013	0.0930
S. meliloti pSymA	0.0798	0.0109	0.2320	0.0800	0.0103	0.2218
S. meliloti 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
E. coli Chromosome	5082529	2960007	191748	207199	9534
B. subtilis Chromosome	4077077	2074653	102906	205150	6187
Streptomyces Chromosome	8497577	2422980	21581	551530	3670
S. meliloti Chromosome	3426881	1931139	199425	6684	842
S. meliloti pSymA	1455940	419223	34213	9832	943
S. meliloti 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
E. coli Chromosome	-9.983×10^{-8}***	6.994×10^{-8}***
B. subtilis Chromosome	-1.071×10^{-7}***	-9.861×10^{8}***
Streptomyces Chromosome	-2.626×10^{-8}***	3.615×10^{-7}***
S. meliloti Chromosome	-1.367×10^{-7}***	-1.510×10^{-7}*
S. meliloti pSymA	-1.075×10^{-7}*	NS
S. meliloti pSymB	2.878×10^{-7}***	8.595×10^{-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
E. coli K12 MG1655	GSE60522	December 20, 2017
E. coli K12 MG1655	GSE73673	December 19, 2017
E. coli K12 MG1655	GSE85914	December 19, 2017
E. coli K12 MG1655	GSE40313	November 21, 2018
E. coli K12 MG1655	GSE114917	November 22, 2018
E. coli K12 MG1655	GSE54199	November 26, 2018
E. coli K12 DH10B	GSE98890	December 19, 2017
E. coli BW25113	GSE73673	December 19, 2017
E. coli BW25113	GSE85914	December 19, 2017
E. coli O157:H7	GSE46120	August 28, 2018
E. coli ATCC 25922	GSE94978	November 23, 2018
B. subtilis 168	GSE104816	December 14, 2017
B. subtilis 168	GSE67058	December 16, 2017
B. subtilis 168	GSE93894	December 15, 2017
B. subtilis 168	GSE80786	November 16, 2018
S. coelicolor A3	GSE57268	March 16, 2018
S. natalensis HW-2	GSE112559	November 15, 2018
S. meliloti 1021 Chromosome	GSE69880	December 12, 2017
S. meliloti 2011 pSymA	NC_020527 (Dr. Finan)	April 4, 2018
S. meliloti 1021 pSymA	GSE69880	November 15, 18
S. meliloti 2011 pSymB	NC_020560 (Dr. Finan)	April 4, 2018
S. meliloti 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
E. coli Chromosome	-6.03×10^{-5}	1.28×10^{-5}	2.8×10^{-6}
B. subtilis Chromosome	-9.7×10^{-5}	2.0×10^{-5}	1.2×10^{-6}
Streptomyces Chromosome	-1.17×10^{-6}	1.04×10^{-7}	$<2\times10^{-16}$
S. meliloti Chromosome	3.97×10^{-5}	4.25×10^{-5}	NS (3.5×10^{-1})
S. meliloti pSymA	1.39×10^{-3}	2.53×10^{-4}	4.9×10^{-8}
S. meliloti pSymB	1.46×10^{-4}	2.03×10^{-4}	NS ($5.34.7\times10^{-1}$)

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.