

Subs Paper Things to Do:

- ~~write dN/dS methods~~
- ~~write dN/dS results~~
- ~~write dN/dS discussion~~
- ~~write dN/dS into conclusion~~
- ~~spatial analysis of  $dN$ ,  $dS$ , and  $\omega$~~
- ~~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

Last week I finally fixed the previous issue where pseudogenes and genes with “joins” in them were messing up my code. With this, I decided to more accurately describe the coding and non-coding sections of the genome. All coding regions will now be referred to as protein coding regions, and non-coding regions will be referred to as non-protein coding regions. So all intergenic, RNA, and pseudogenes are classified as non-coding regions. I started to re-run the selection and substitution analysis on Thursday. *Streptomyces* takes a really really long time to run because it has the most blocks (and the biggest genome) so it is still running through the both analysis. All the other bacteria are done with the substitutions analysis and the graphs and results can be found below.

I hit a small snag in the selection analysis, it looks like the alignments that are outputted from my program do not all have a multiple of 3 as its length. I am currently looking into this. Hopefully I can solve this issue by today and have the results for the selection analysis by the end of the week.

I also wanted your opinion about something that was mentioned at the conference. Someone asked me if I was including RNA in my “coding” sections? Because RNA often is under different selective pressures than coding or other “non-coding” sections, they suggested that I do my analysis on just all RNA to see if there is any sort of trend with respect to distance from the origin. I was wondering what you think about this and if you think it is worth it for me to do?

I have begun compiling my committee report for Aug 26th. I am hoping to send it out to everyone by Monday Aug 19th, so there is time to read it before the meeting.

## This Week

This week I need to figure out why there are non-multiple of 3 genes being outputted by my analysis (which should not be happening). Once this is complete, I can continue to re-run the selection analysis.

I would like to finish my committee report by the end of the weekend.

While all that is running, I need to keep working on the inversions and gene expression stuff.

## Next Week

Assuming the re-running of my analysis is complete, I would like to work intensely on the inversions and gene expression stuff and get a good chunk of that started.

Bacteria and Replicon	Coding Sequences	Non-Coding Sequences
<i>E. coli</i> Chromosome	$3.557 \times 10^{-7***}$	NS
<i>B. subtilis</i> Chromosome	$-7.804 \times 10^{-8***}$	$-3.170 \times 10^{8*}$
<i>Streptomyces</i> Chromosome	$4.545 \times 10^{-8***}$	$1.584 \times 10^{-7***}$
<i>S. meliloti</i> Chromosome	$-9.183 \times 10^{-8***}$	$-1.718 \times 10^{-7***}$
<i>S. meliloti</i> pSymA	$-8.121 \times 10^{-7***}$	$-1.247 \times 10^{-7***}$
<i>S. meliloti</i> pSymB	$1.655 \times 10^{-7***}$	$4.105 \times 10^{-7***}$

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	$dN$	$dS$	$\omega$
<i>E. coli</i> Chromosome	NS	NS	$1.416 \times 10^{-4}$
<i>B. subtilis</i> Chromosome	$-1.383 \times 10^{-6***}$	NS	$-1.309 \times 10^{-5**}$
<i>Streptomyces</i> Chromosome	NS	NS	NS
<i>S. meliloti</i> Chromosome	NS	NS	NS
<i>S. meliloti</i> pSymA	NS	NS	NS
<i>S. meliloti</i> pSymB	NS	NS	$1.167 \times 10^{-5*}$

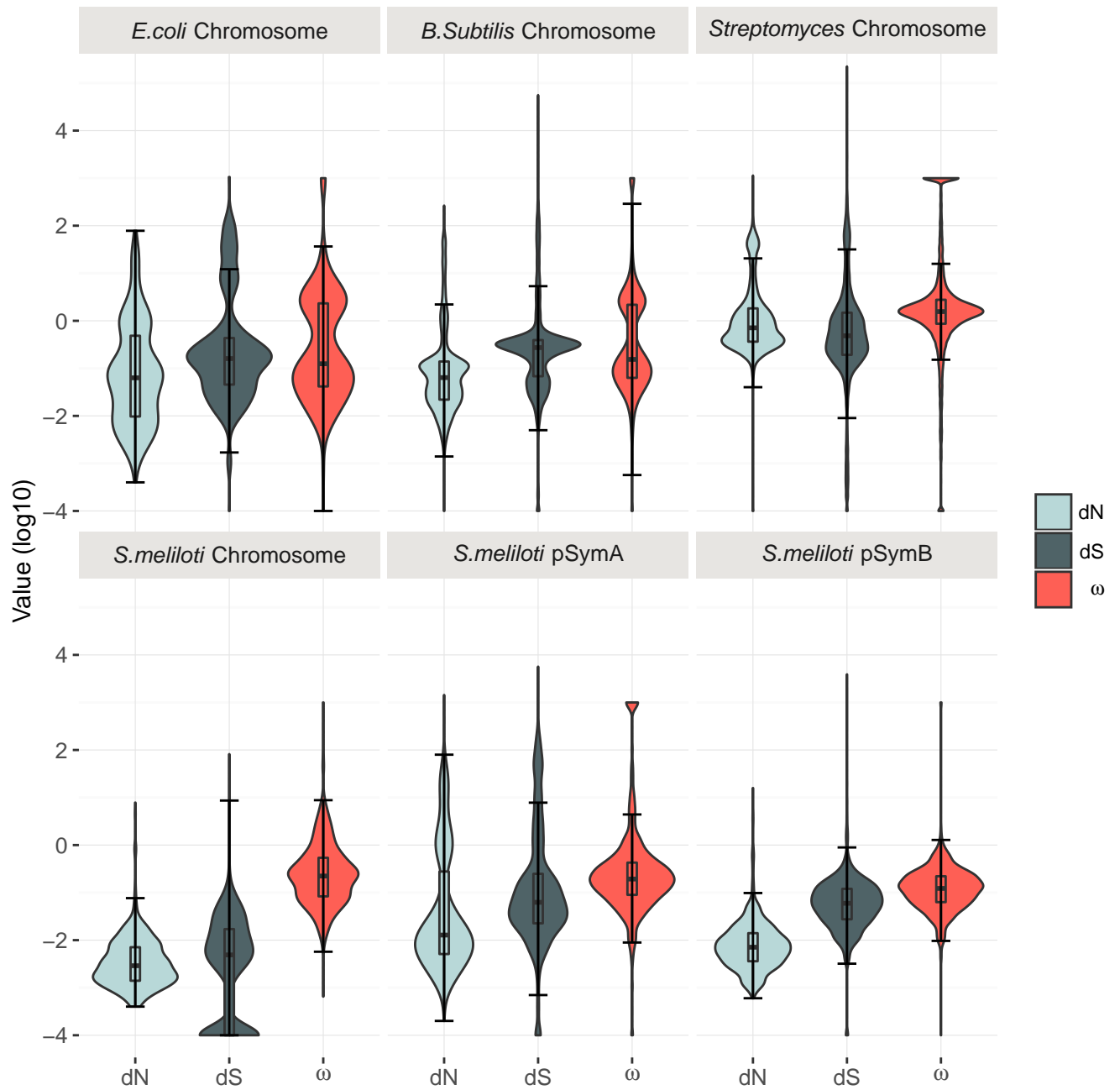
Table 2: Linear regression for  $dN$ ,  $dS$ , and  $\omega$  calculated for each bacterial replicon on a per genome basis. All results are marked with significance codes as followed:  $p: < 0.001 = '***'$ ,  $0.001 < 0.01 = '**'$ ,  $0.01 < 0.05 = '*'$ ,  $> 0.05 = 'NS'$ .

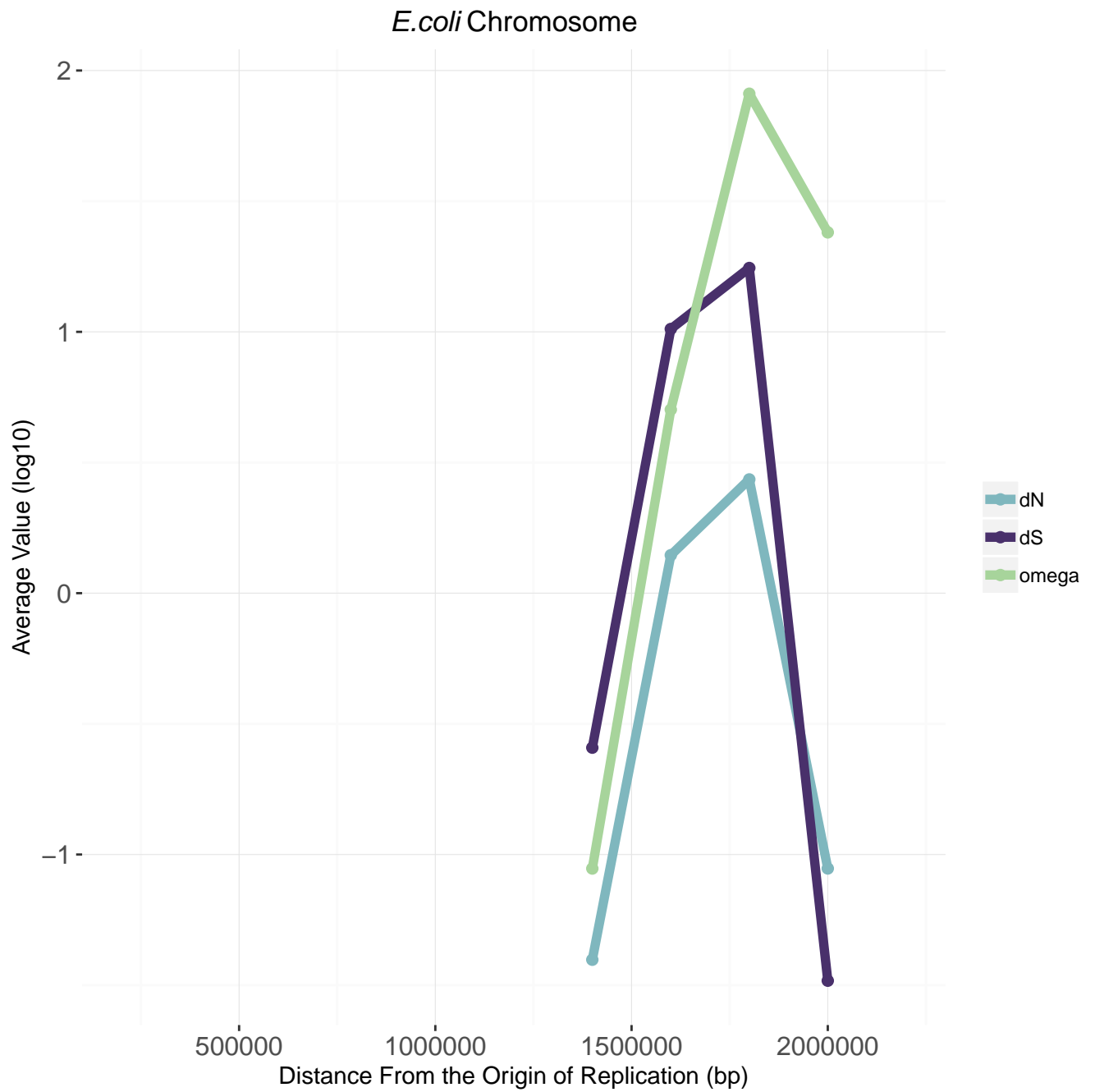
Bacteria and Replicon	Average Expression Value (CPM)
<i>E. coli</i> Chromosome	160.500
<i>B. subtilis</i> Chromosome	176.400
<i>Streptomyces</i> Chromosome	6.084
<i>S. meliloti</i> Chromosome	271.400
<i>S. meliloti</i> pSymA	690.100
<i>S. meliloti</i> pSymB	595.700

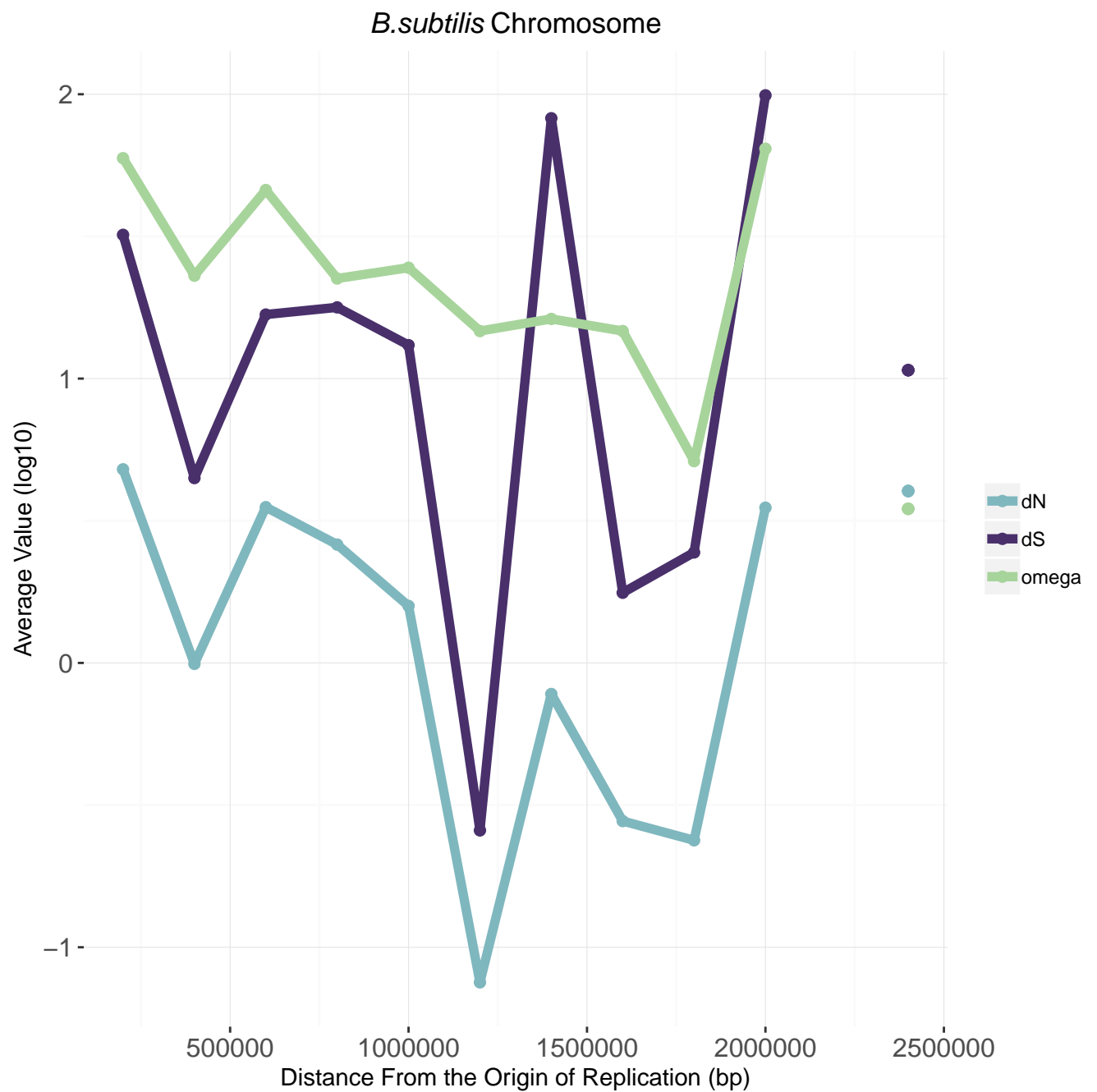
Table 3: Arithmetic gene expression calculated across all genes in each replicon. Expression values are represented in Counts Per Million.

Bacteria and Replicon	Gene Average			Genome Average		
	dS	dN	$\omega$	dS	dN	$\omega$
<i>E. coli</i> Chromosome	1.3694	0.1433	3.2789	0.7445	0.0752	1.5618
<i>B. subtilis</i> Chromosome	4.4557	0.2755	8.9584	0.7947	0.1140	4.8677
<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	1.0602	0.7451	5.1290	0.4100	0.0863	0.8311
<i>S. meliloti</i> pSymB	3.3091	0.0260	0.3878	0.1436	0.0100	0.1950

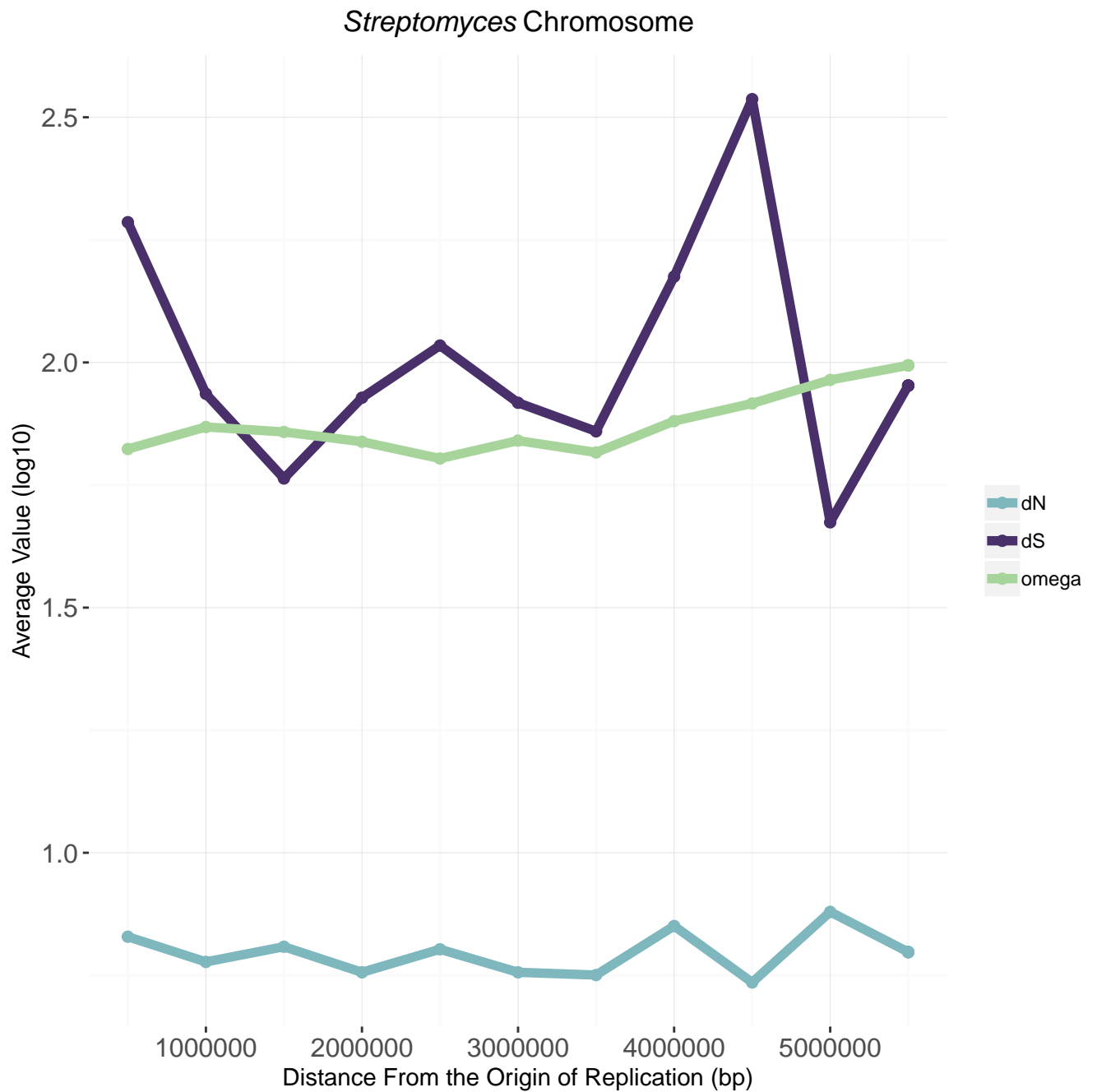
Table 4: Weighted averages calculated for each bacterial replicon on a per genome basis using the gene length as the weight. Arithmetic mean calculated for the per gene averages for each bacterial replicon.

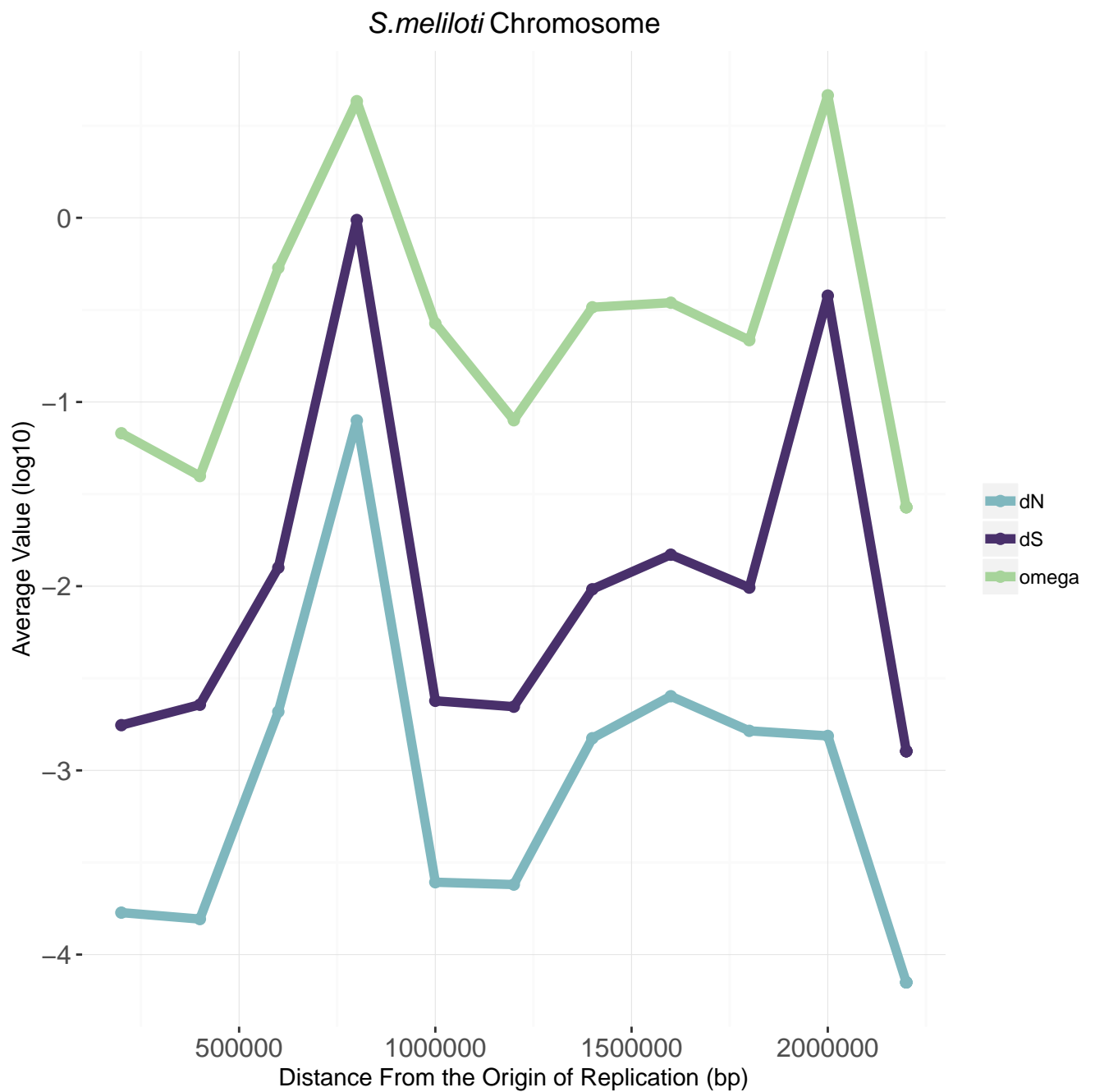


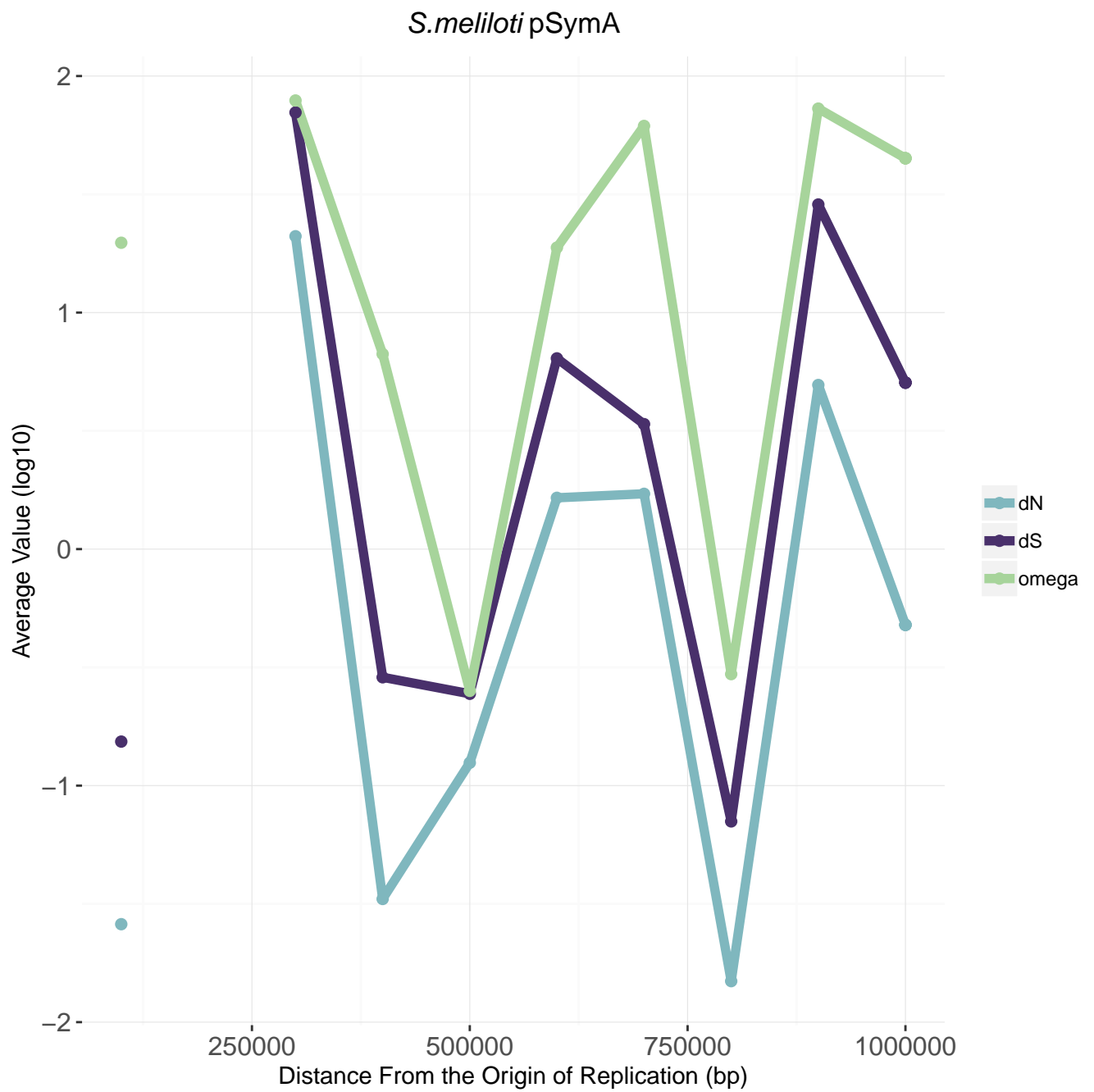


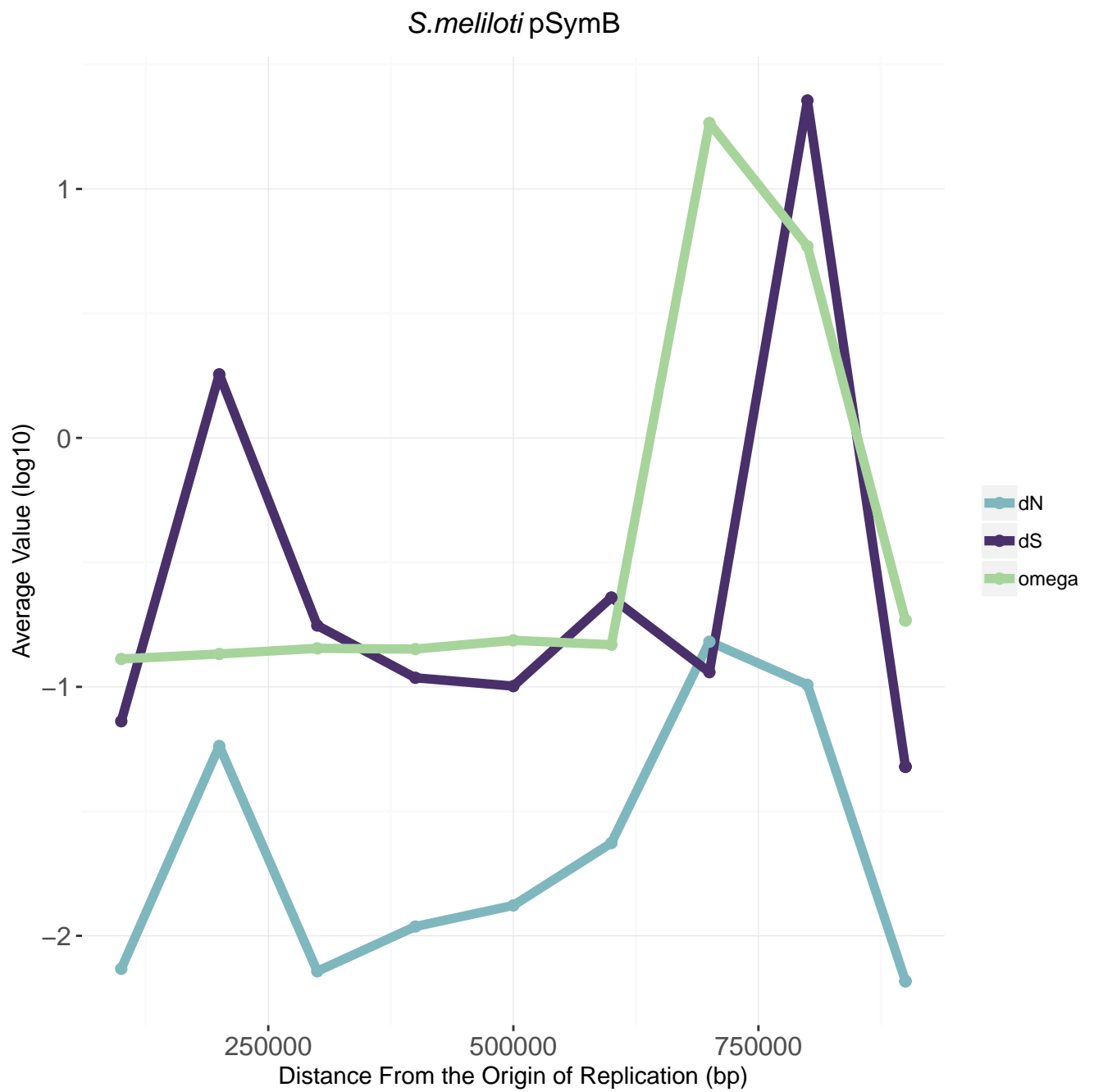


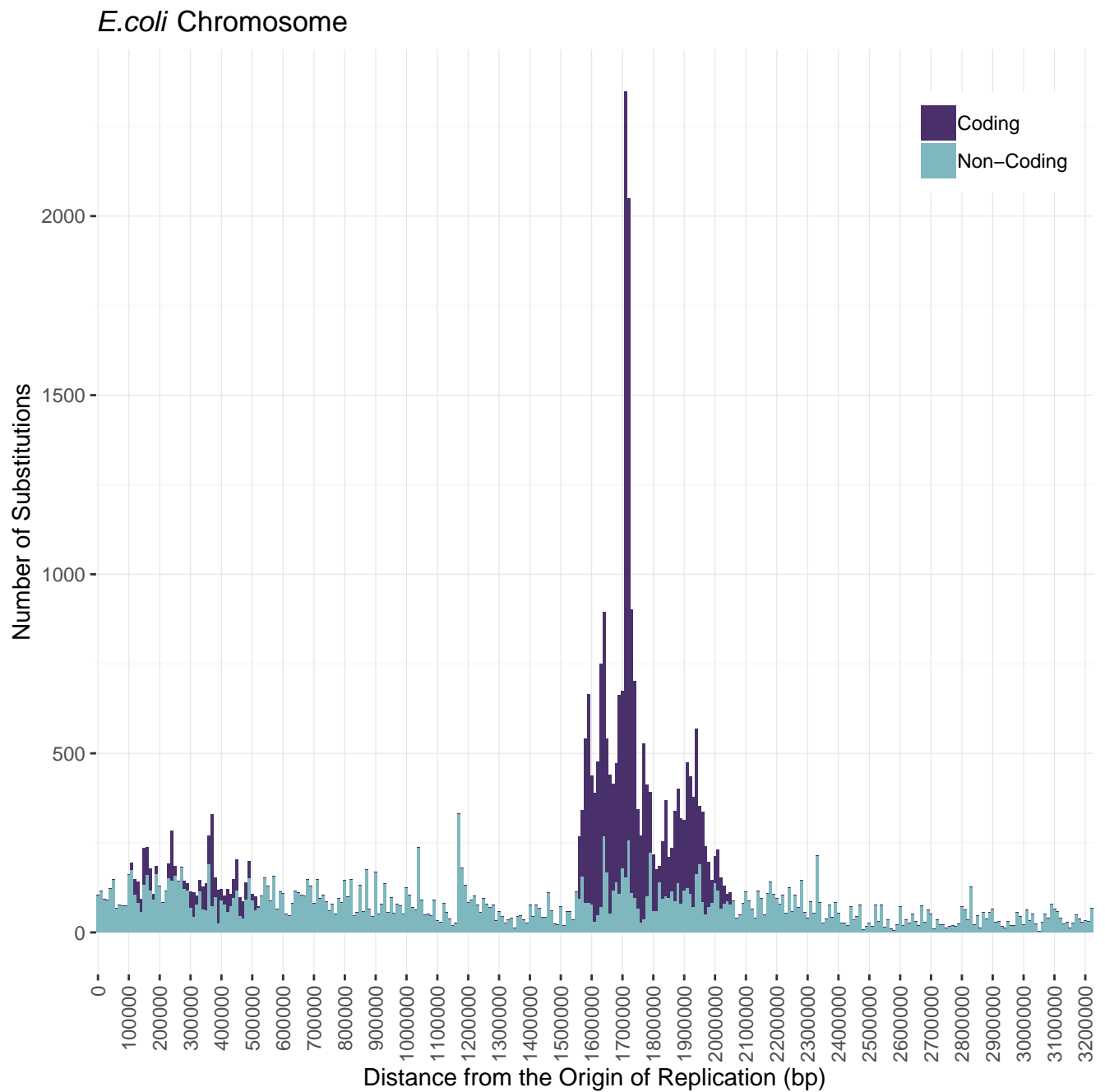


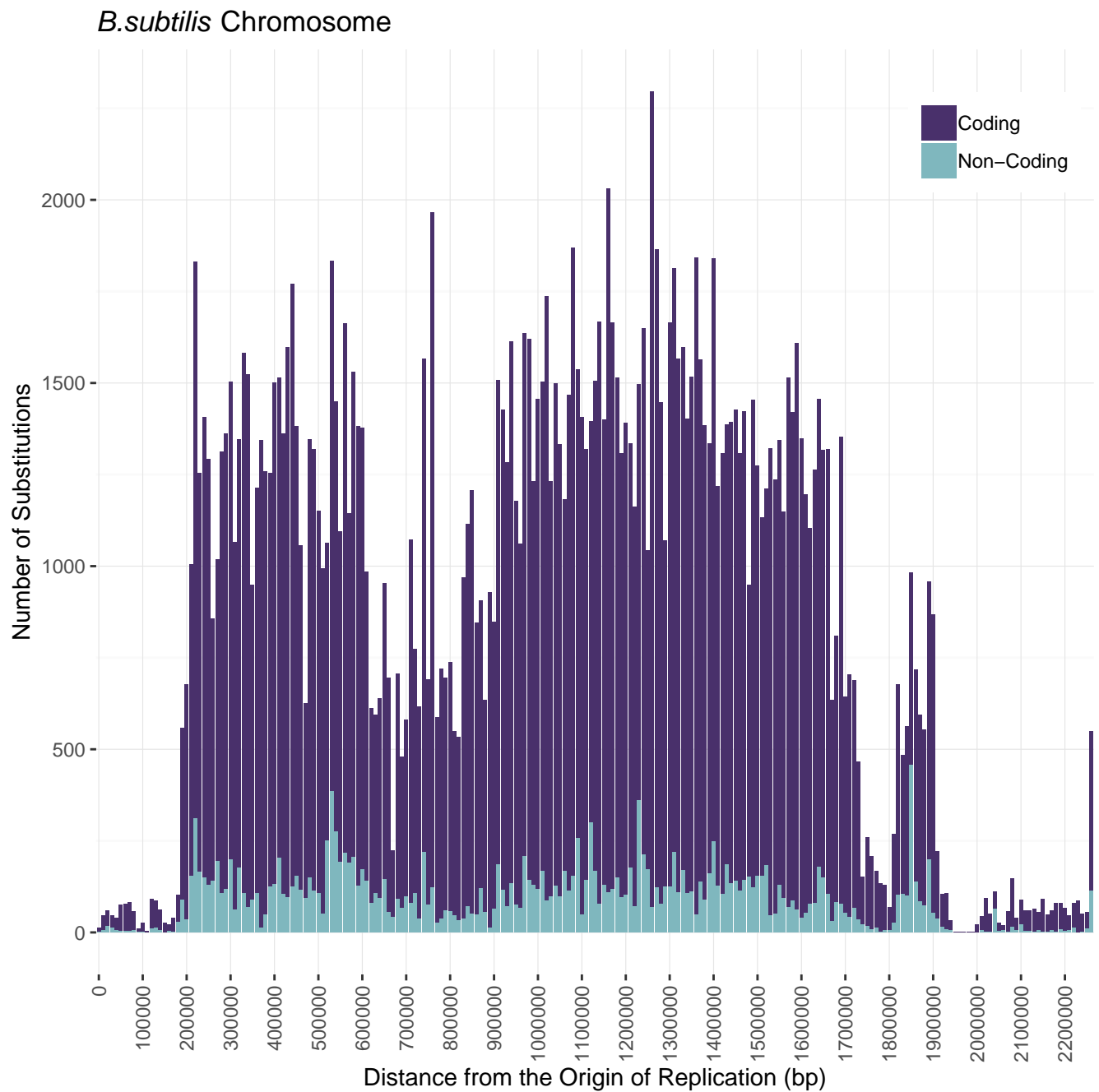


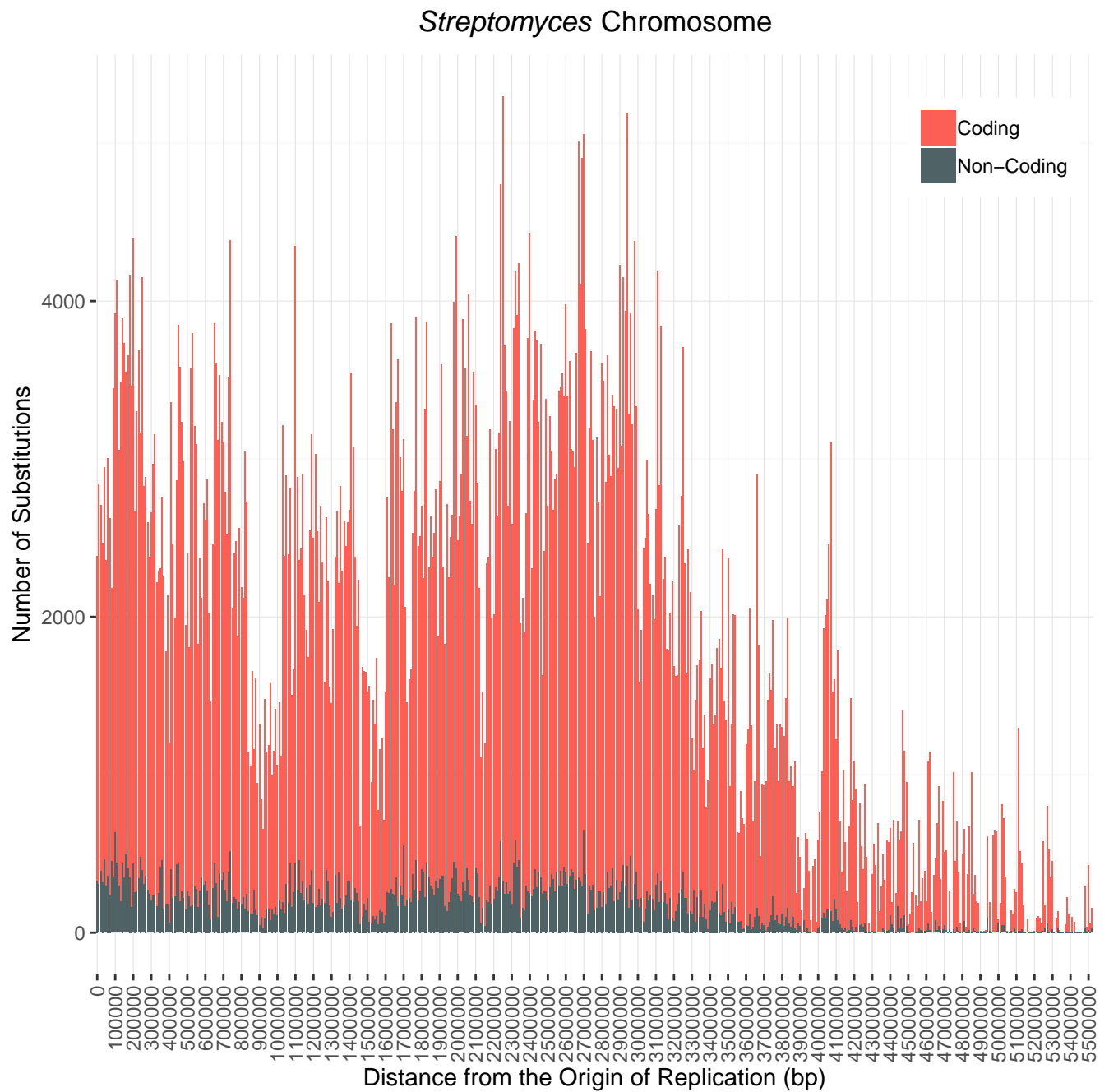


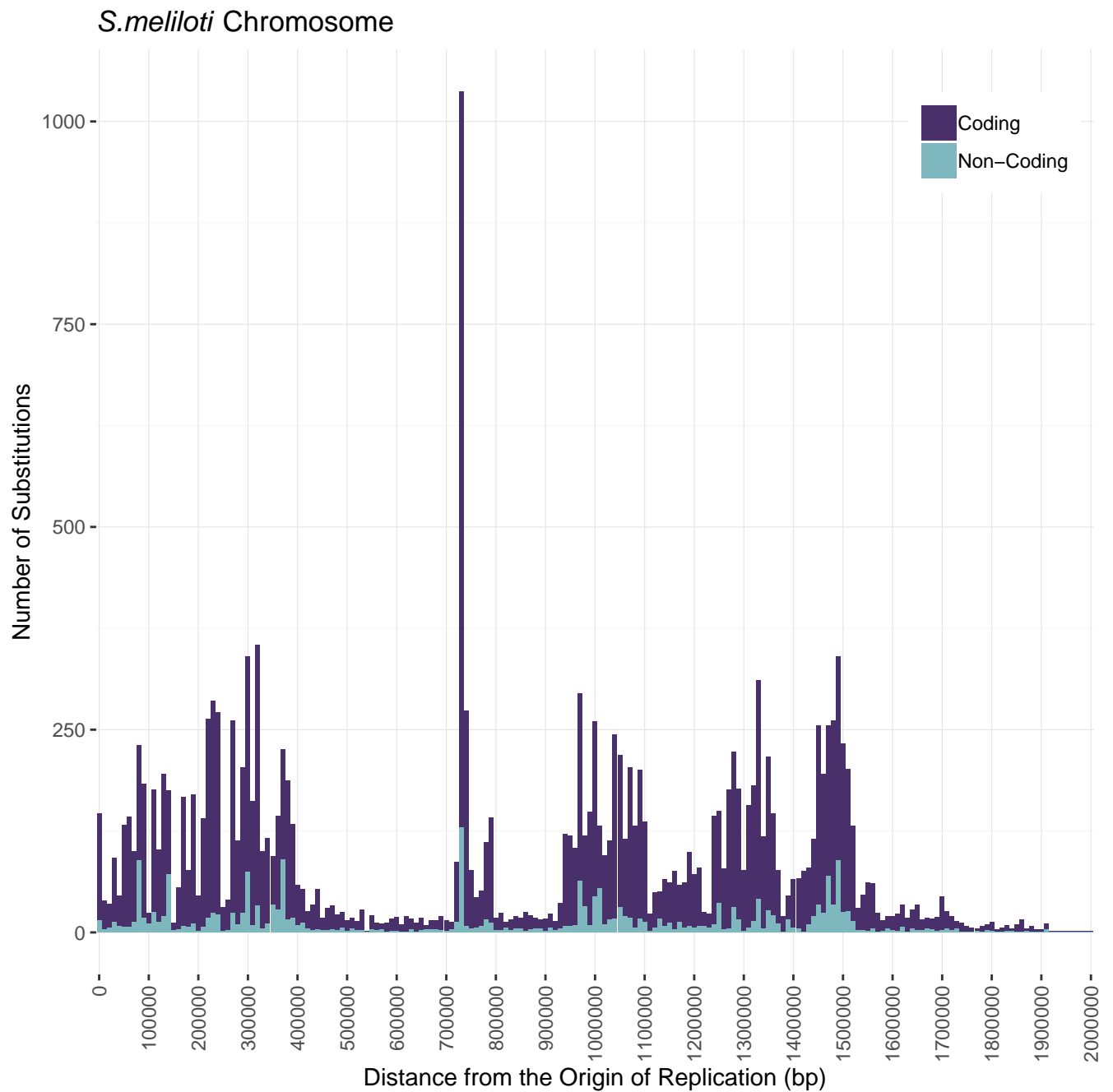




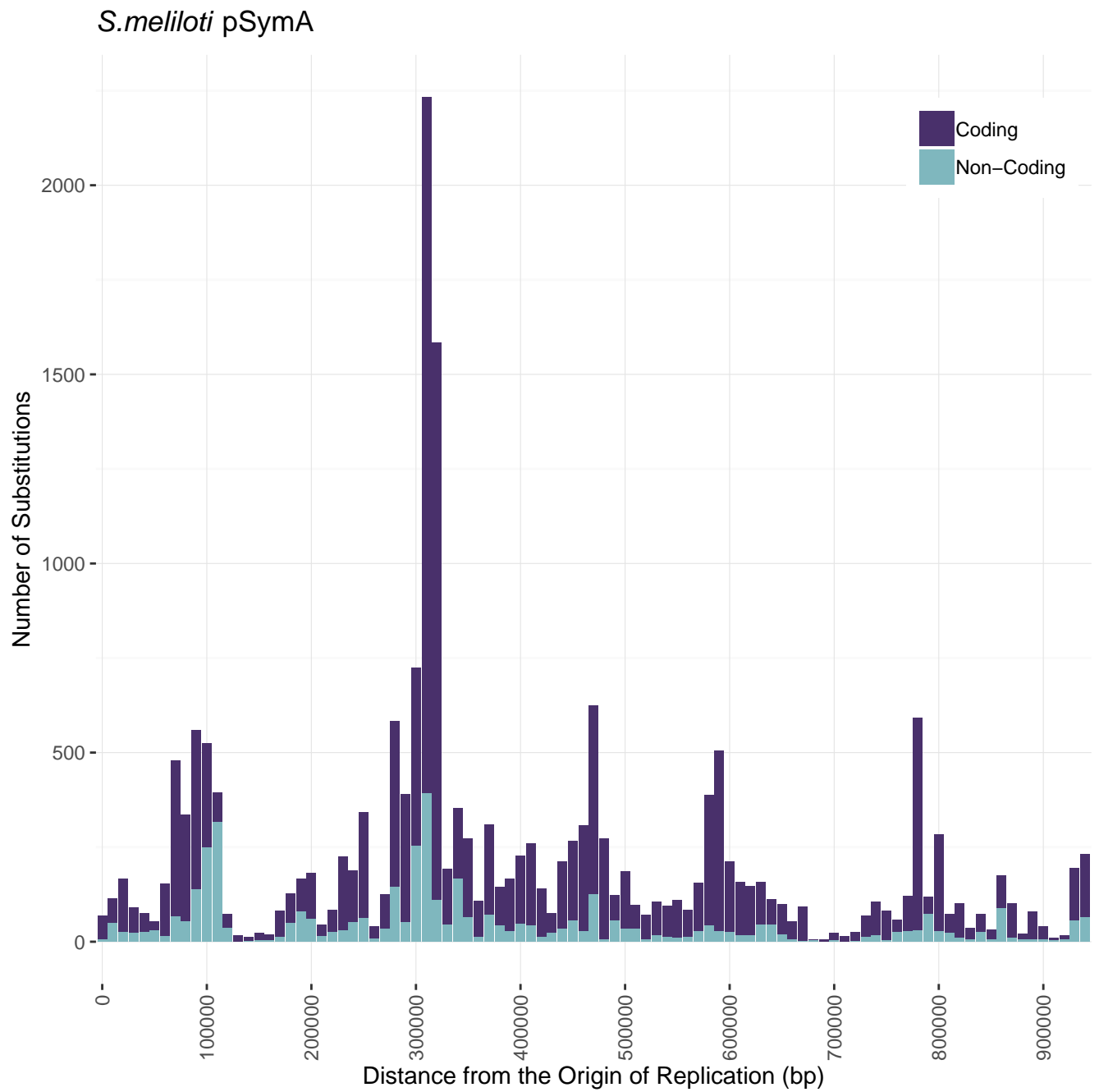


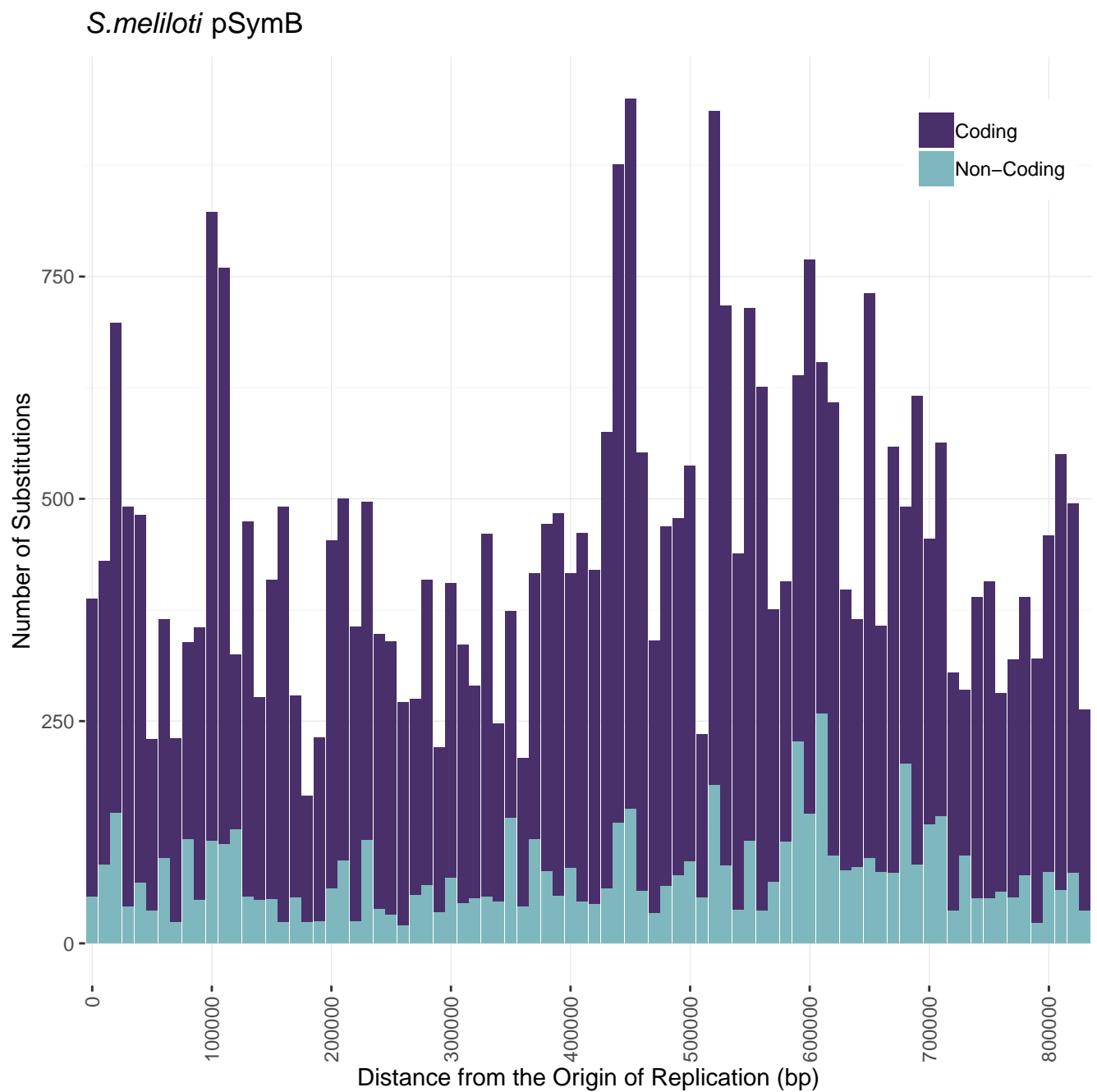












Bacteria Strain/Species	GEO Accession Number	Date Accessed
<i>E. coli</i> K12 MG1655	GSE60522	December 20, 2017
<i>E. coli</i> K12 MG1655	GSE73673	December 19, 2017
<i>E. coli</i> K12 MG1655	GSE85914	December 19, 2017
<i>E. coli</i> K12 MG1655	GSE40313	November 21, 2018
<i>E. coli</i> K12 MG1655	GSE114917	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	$-6.03 \times 10^{-5}$	$1.28 \times 10^{-5}$	$2.8 \times 10^{-6}$
<i>B. subtilis</i> Chromosome	$-9.7 \times 10^{-5}$	$2.0 \times 10^{-5}$	$1.2 \times 10^{-6}$
<i>Streptomyces</i> Chromosome	$-1.17 \times 10^{-6}$	$1.04 \times 10^{-7}$	$< 2 \times 10^{-16}$
<i>S. meliloti</i> Chromosome	$3.97 \times 10^{-5}$	$4.25 \times 10^{-5}$	NS ( $3.5 \times 10^{-1}$ )
<i>S. meliloti</i> pSymA	$1.39 \times 10^{-3}$	$2.53 \times 10^{-4}$	$4.9 \times 10^{-8}$
<i>S. meliloti</i> pSymB	$1.46 \times 10^{-4}$	$2.03 \times 10^{-4}$	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.