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 ω~~
- 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 was working on addressing some of the weird things about the substitution analysis. I was mostly looking into why *Escherichia coli* looks like it is missing a whole bunch of data in the dN, dS, ω genome distribution graphs. I already talked to you about this but it looks like it is because some genes have a “join” in the gene location and because of pseudo genes. So I had to re-code how the gene start and stop positions are being gathered. I am still making sure that this does not mess up any analysis down the line and that this will fix the issue with pSymA also missing a big chunk of data.

This Week

This week I overall want to fix/figure out why there are a bunch of weird things happening with the substitution and selection results. The goal is to get this all fixed so this weekend and next week I can re-run the analysis.

I want to continue to work out why *Escherichia coli* is missing so much data.

I would also like to investigate why *S. meliloti* Chromosome has a bunch of zero values for dS and ω.

I would also like to figure out why *Streptomyces* has $dN > dS$.

Next 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).

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

Bacteria and Replicon	Coding Sequences	Non-Coding Sequences
E. coli Chromosome	$3.557\times10^{-7***}$	NS
B. subtilis Chromosome	$-7.804\times10^{-8***}$	$-3.170\times10^{8*}$
Streptomyces Chromosome	$4.545\times10^{-8***}$	$1.584\times10^{-7***}$
S. meliloti Chromosome	$-9.183\times10^{-8***}$	$-1.718\times10^{-7***}$
S. meliloti pSymA	$-8.121\times10^{-7***}$	$-1.247\times10^{-7***}$
S. meliloti pSymB	$1.655\times10^{-7***}$	$4.105\times10^{-7***}$

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	ω
E. coli Chromosome	NS	NS	1.416×10^{-4}
B. subtilis Chromosome	$-1.383\times10^{-6***}$	NS	$-1.309\times10^{-5**}$
Streptomyces Chromosome	NS	NS	NS
S. meliloti Chromosome	NS	NS	NS
S. meliloti pSymA	NS	NS	NS
S. meliloti pSymB	NS	NS	$1.167\times10^{-5*}$

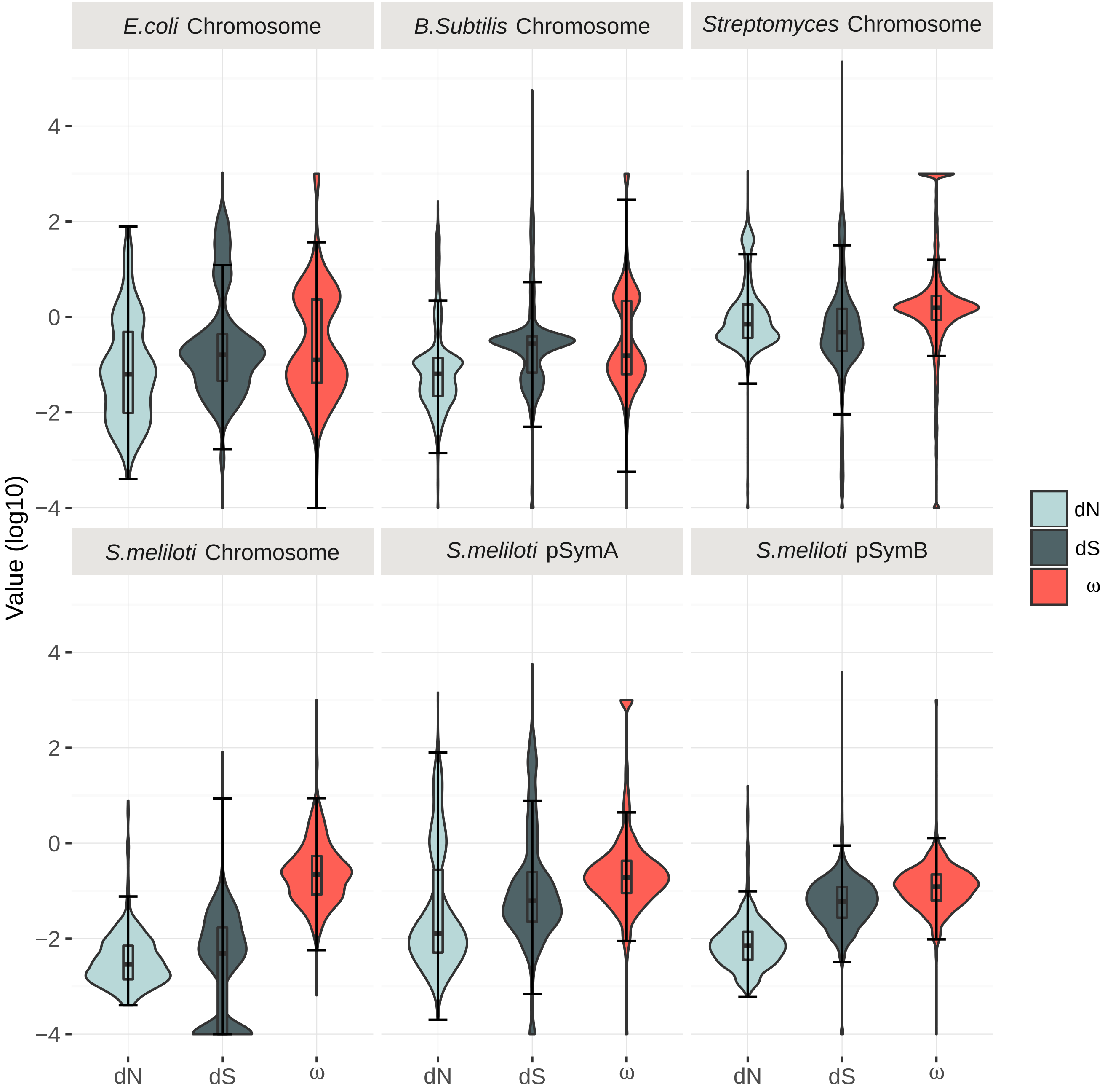
Table 2: Linear regression for dN, dS, and ω calculated for each bacterial replicon on a per genome basis. All results are marked with significance codes as followed: p: < 0.001 = ‘***’, $0.001 < 0.01$ = ‘**’, $0.01 < 0.05$ = ‘*’, > 0.05 = ‘NS’.

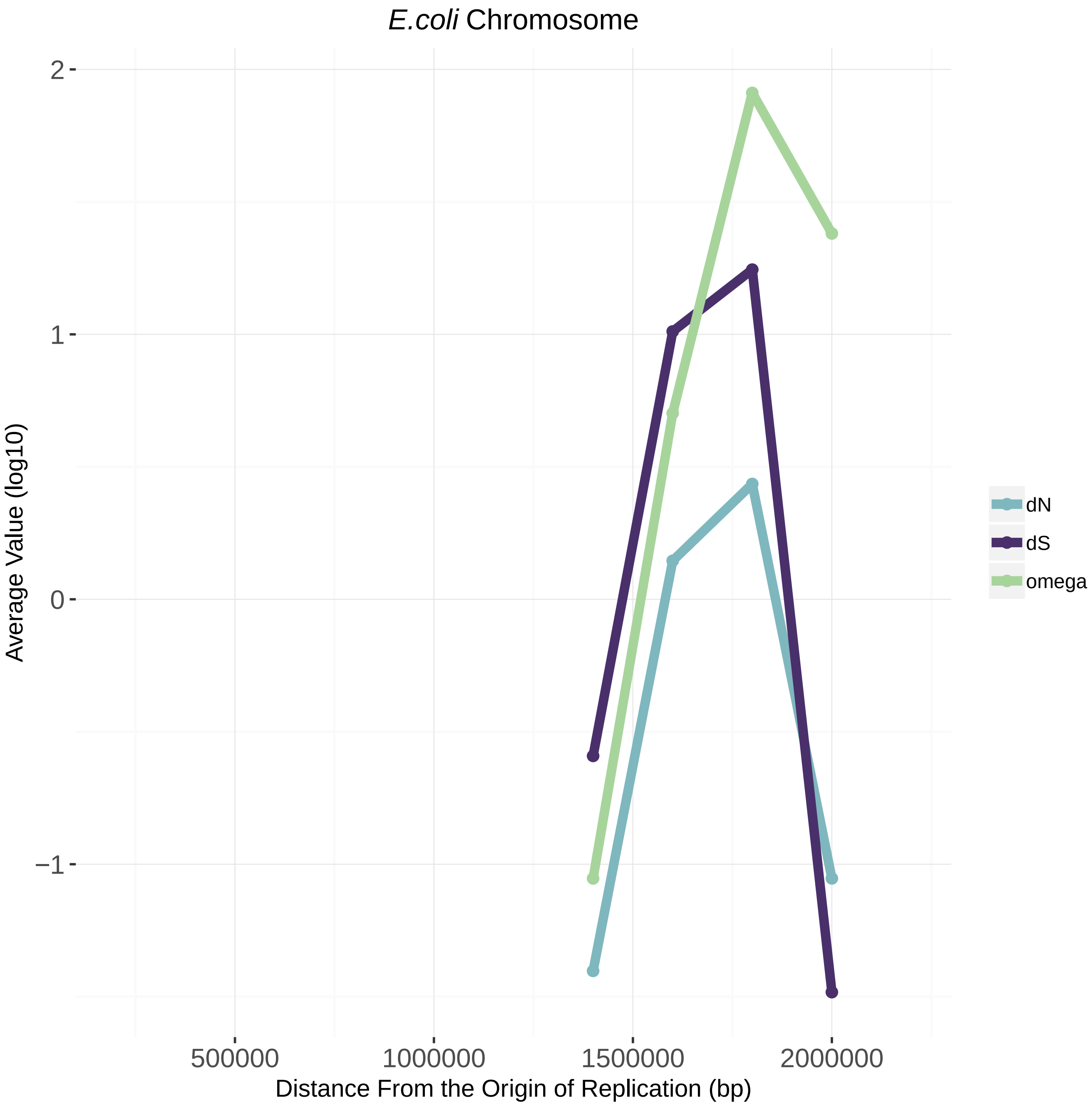
Bacteria and Replicon	Average Expression Value (CPM)
E. coli Chromosome	160.500
B. subtilis Chromosome	176.400
Streptomyces Chromosome	6.084
S. meliloti Chromosome	271.400
S. meliloti pSymA	690.100
S. meliloti pSymB	595.700

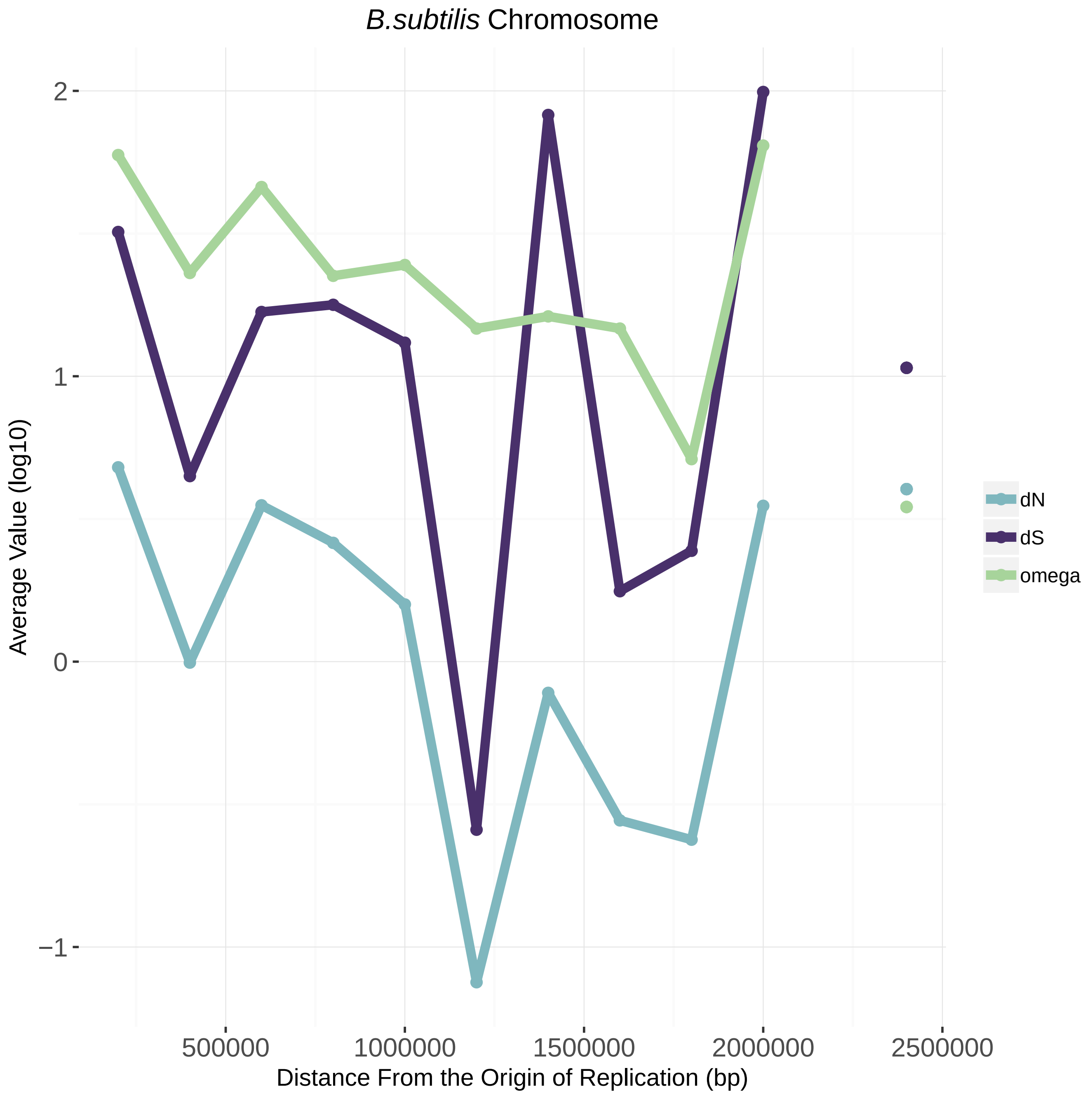
Table 3: Arithmetic gene expression calculated across all genes in each replicon. Expression values are represented in Counts Per Million.

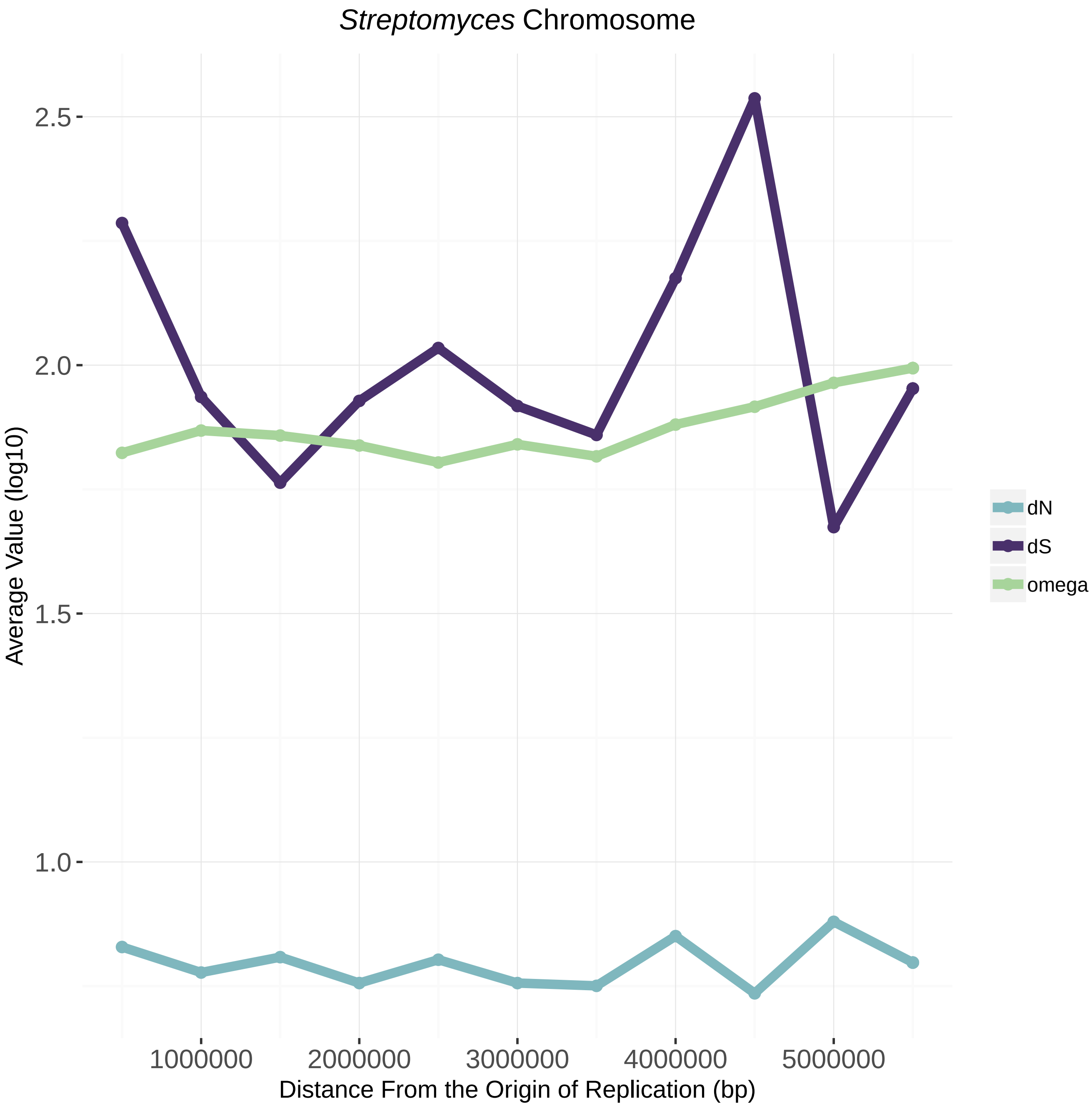
	Gene Average			Genome Average		
Bacteria and Replicon	dS	dN	ω	dS	dN	ω
E. coli Chromosome	1.3694	0.1433	3.2789	0.7445	0.0752	1.5618
B. subtilis Chromosome	4.4557	0.2755	8.9584	0.7947	0.1140	4.8677
Streptomyces Chromosome	0.1924	0.3201	2.6404	0.1775	0.3017	2.4358
S. meliloti Chromosome	0.0134	0.0014	0.0844	0.0134	0.0013	0.0930
S. meliloti pSymA	1.0602	0.7451	5.1290	0.4100	0.0863	0.8311
S. meliloti pSymB	3.3091	0.0260	0.3878	0.1436	0.0100	0.1950

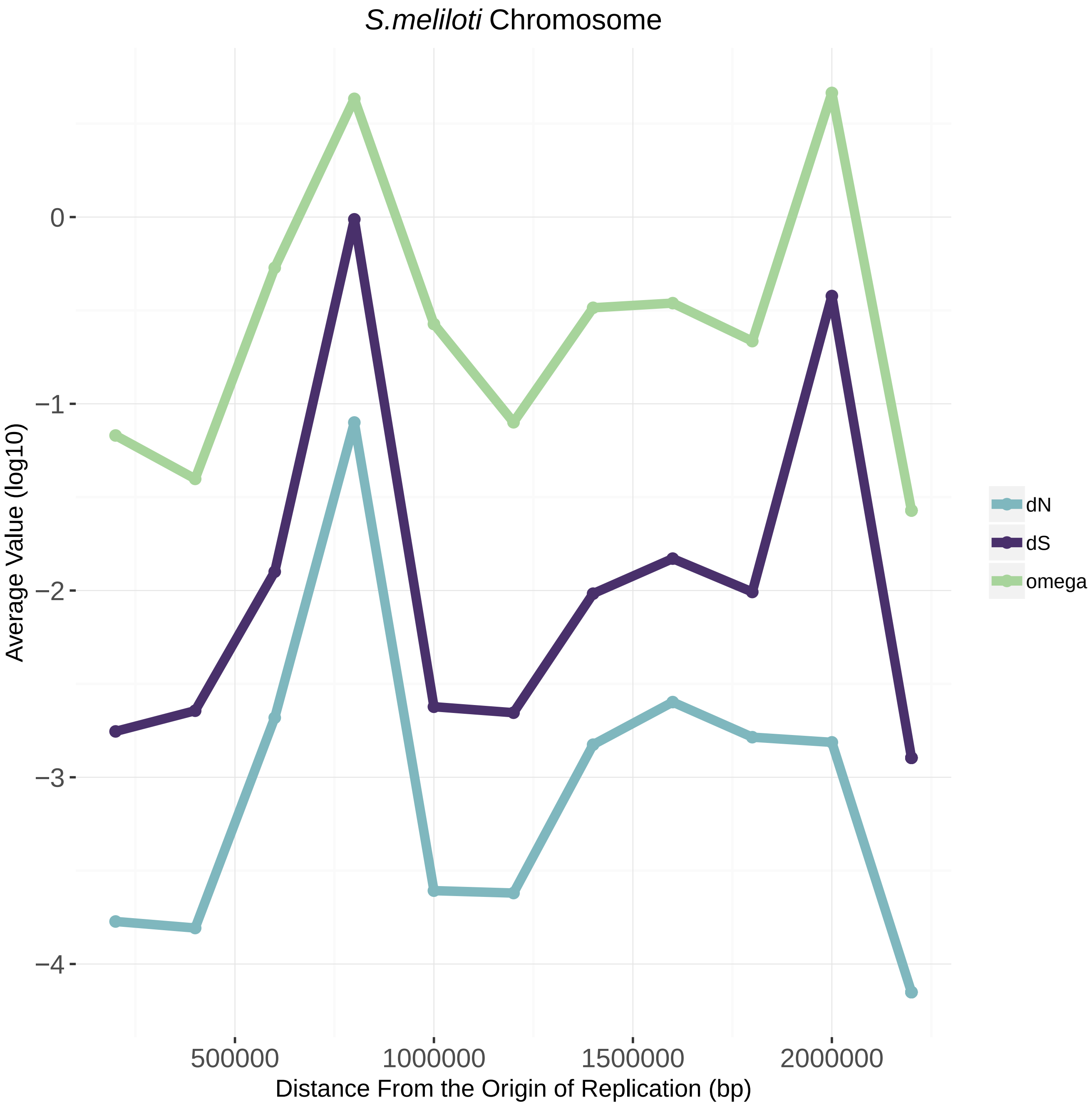
Table 4: Weighted averages calculated for each bacterial replicon on a per genome basis using the gene length as the weight. Arithmetic mean calculated for the per gene averages for each bacterial replicon.

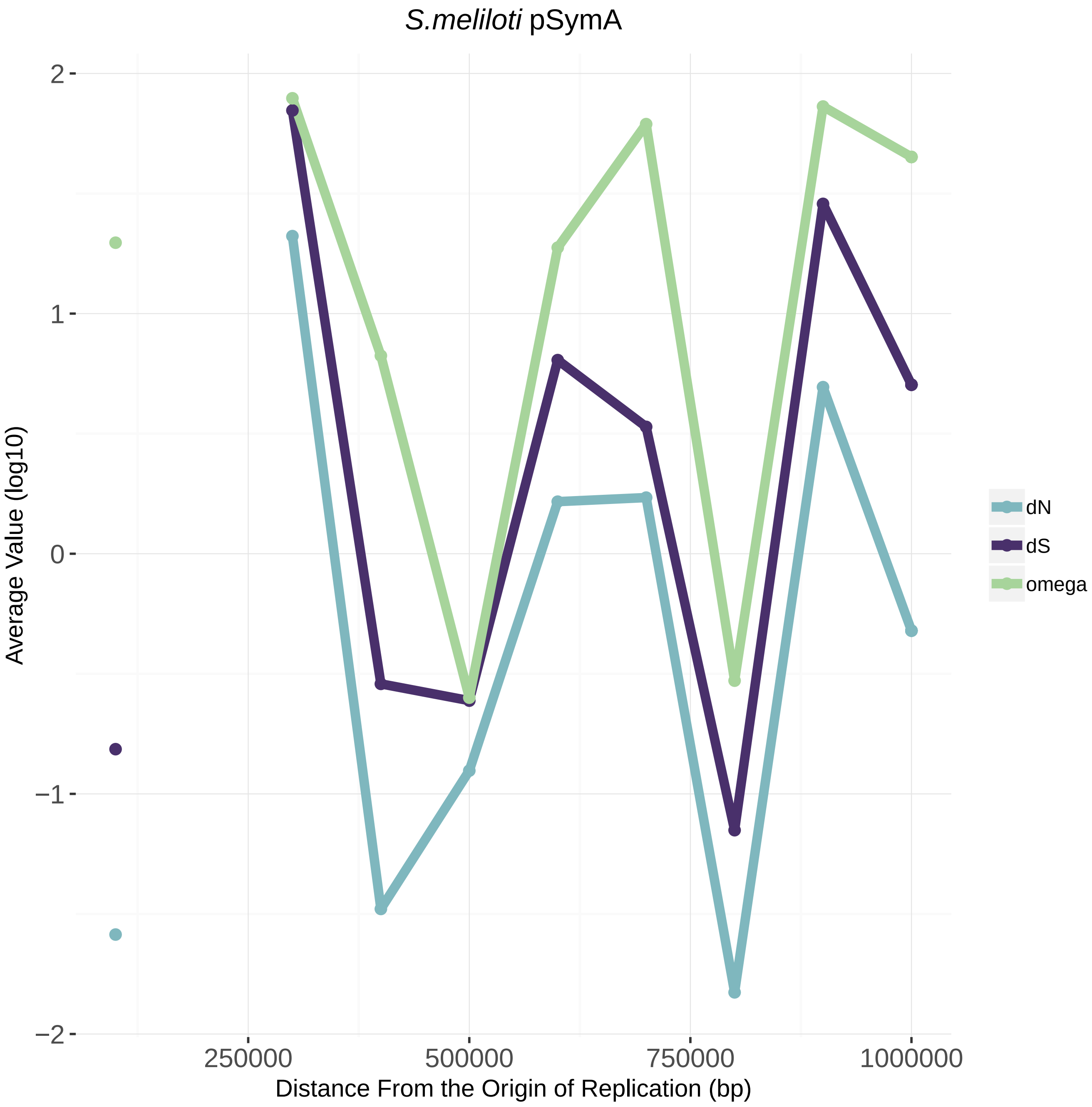

E.coli Chromosome
B.Subtilis Chromosome
Streptomyces Chromosome
S.meliloti Chromosome
S.meliloti pSymA
S.meliloti pSymB
Value (log10)
4
2
0
−2
−4
dN
dS
ω
dN
dS
ω

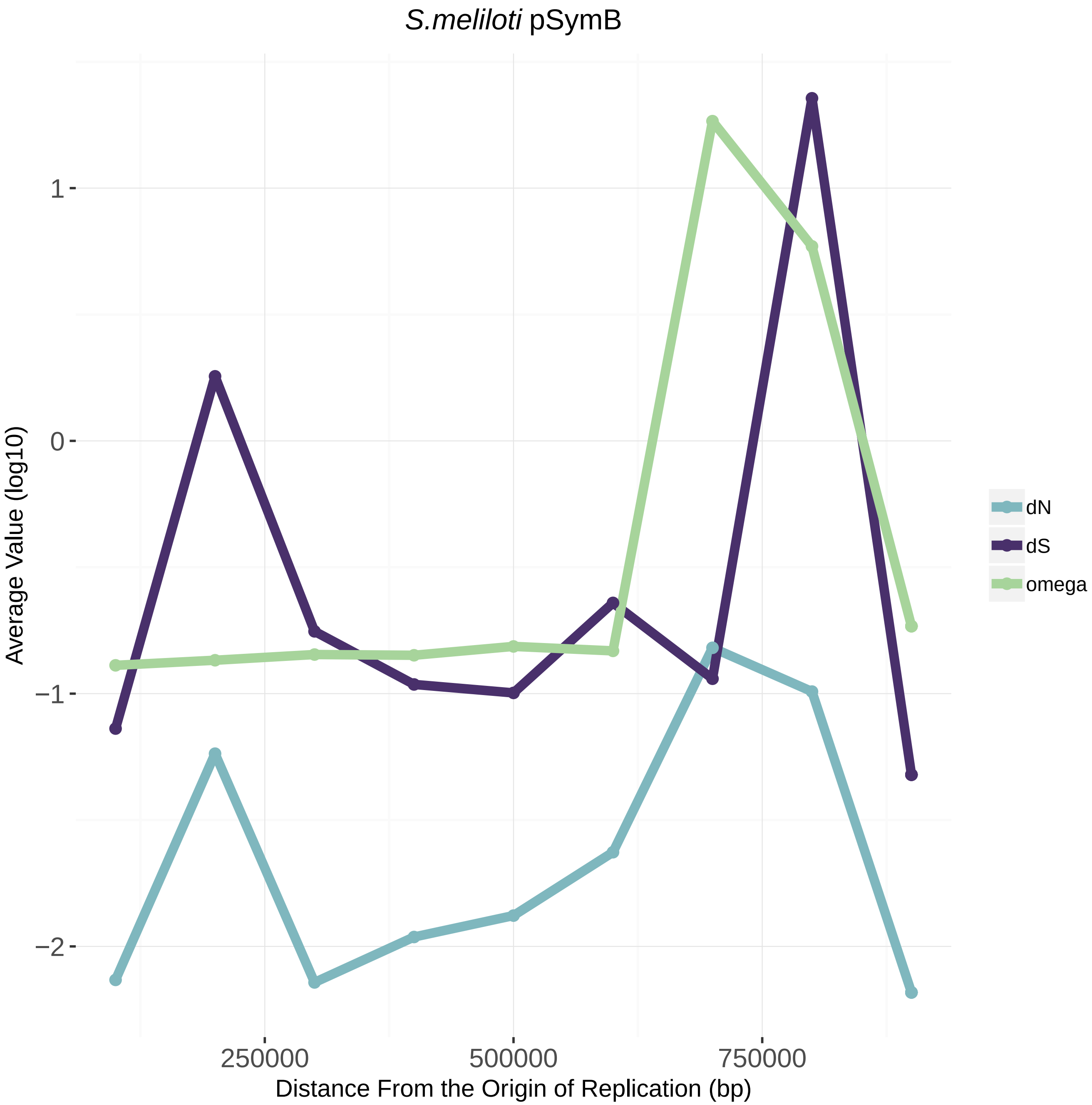

E.coli Chromosome
2
1
0
−1
Average Value (log10)
500000
1000000
1500000
2000000
Distance From the Origin of Replication (bp)
dN
dS
omega

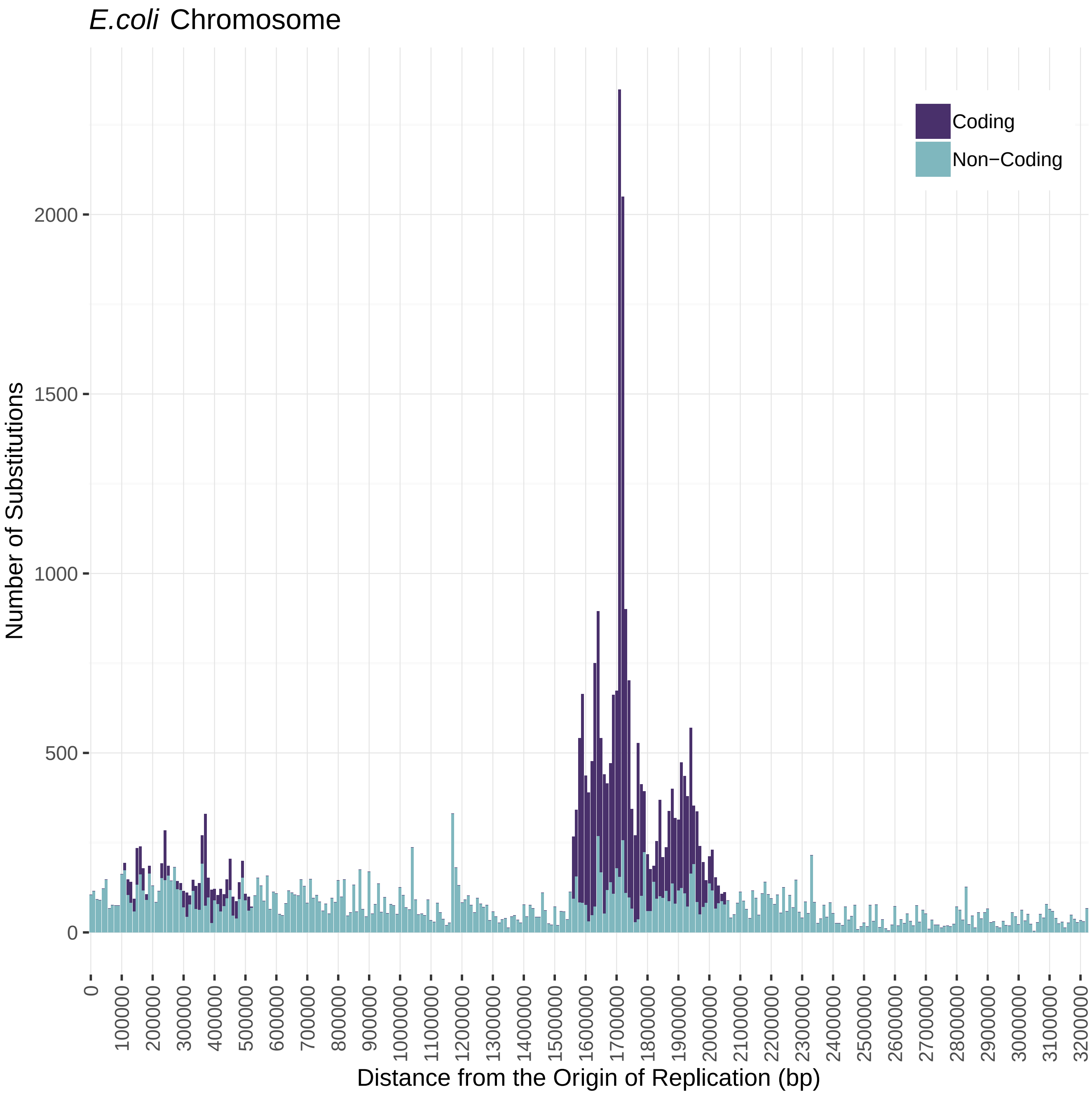

B.subtilis Chromosome
Average Value (log10)
2
1
0
−1
500000
1000000
1500000
2000000
2500000
Distance From the Origin of Replication (bp)
dN
dS
omega

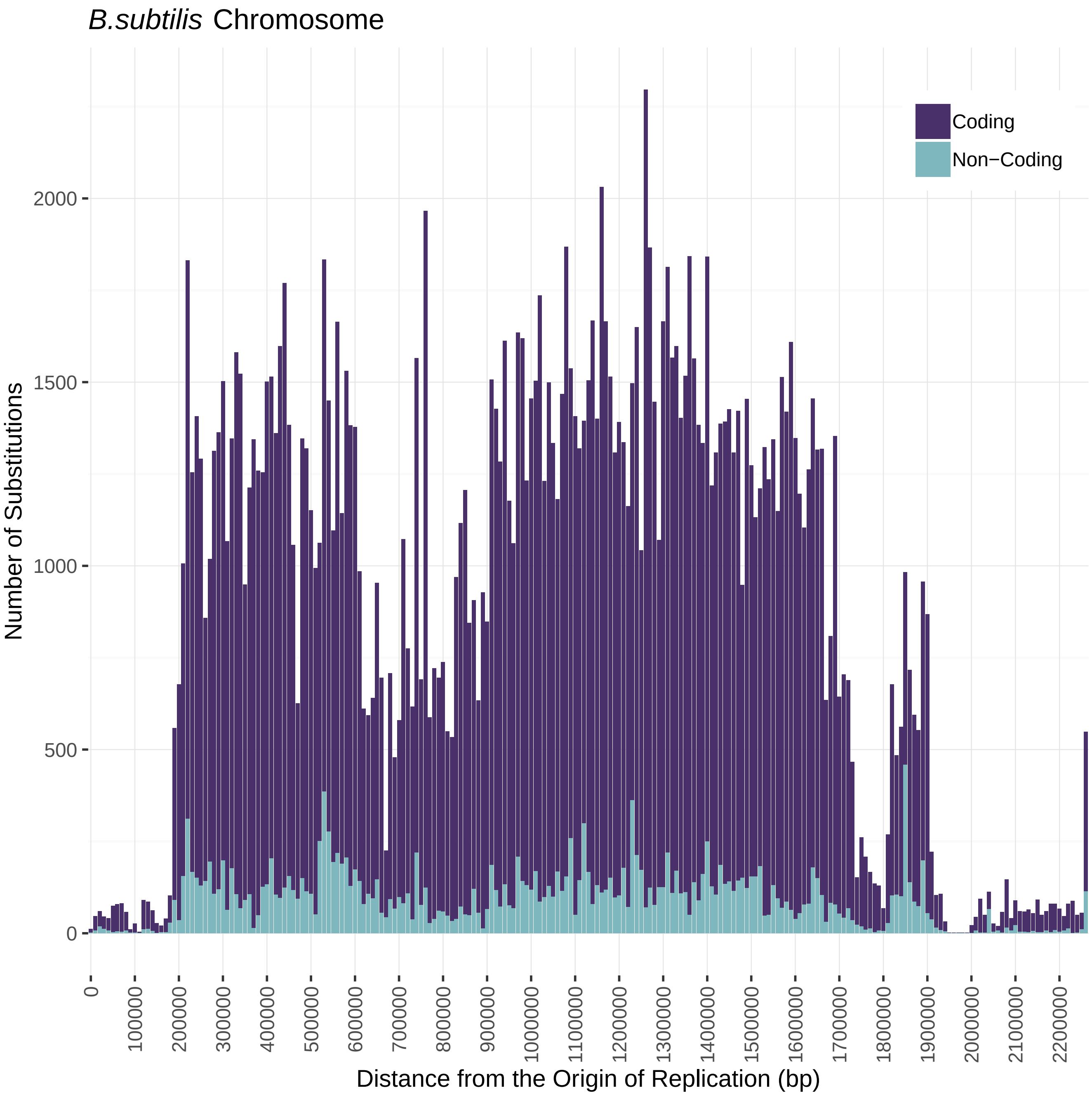

Streptomyces Chromosome
Average Value (log10)
2.5
2.0
1.5
1.0
1000000
2000000
3000000
4000000
5000000
Distance From the Origin of Replication (bp)
dN
dS
omega

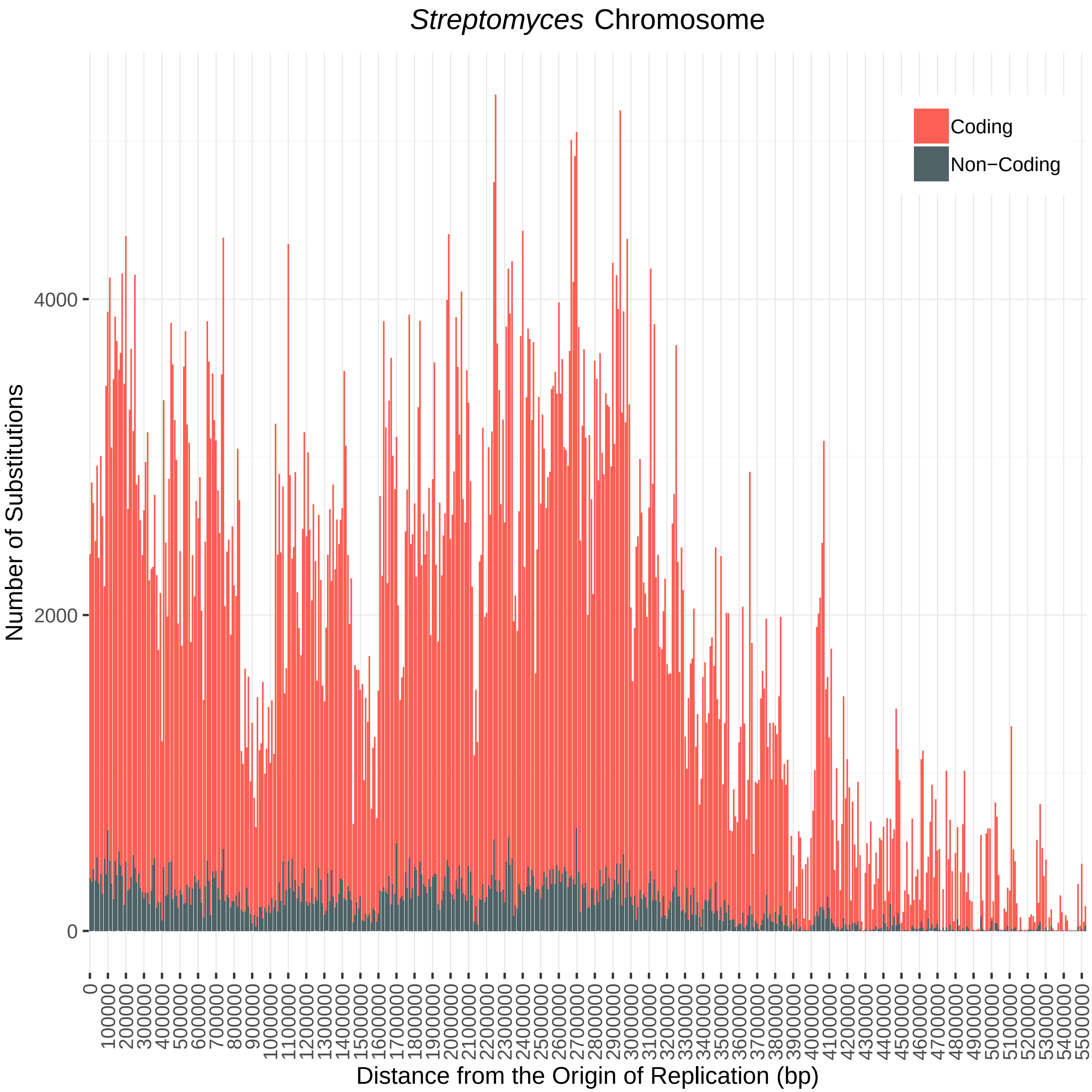

S.meliloti Chromosome
Average Value (log10)
0
−1
−2
−3
−4
500000
1000000
1500000
2000000
Distance From the Origin of Replication (bp)
dN
dS
omega

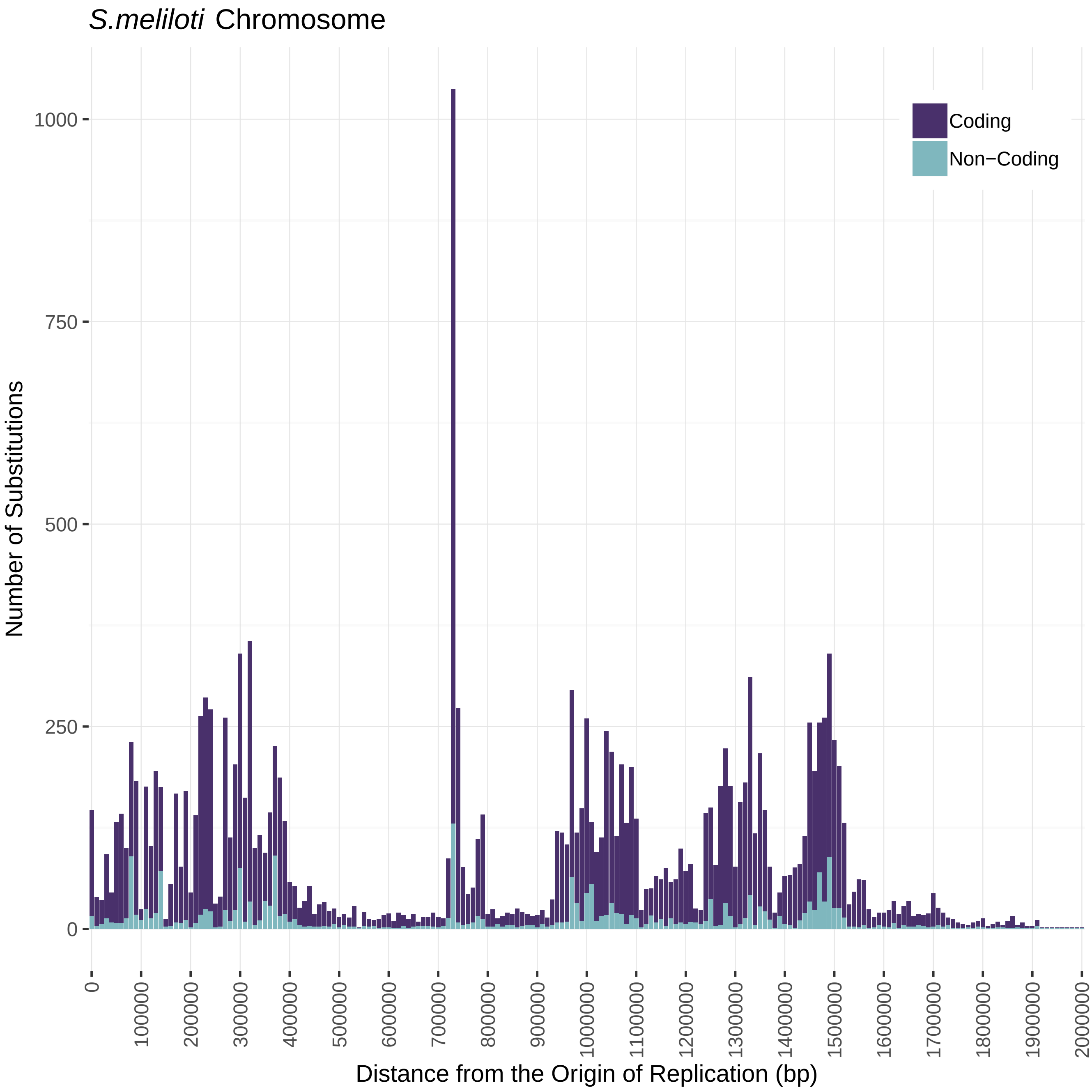

S.meliloti pSymA
Average Value (log10)
2
1
0
−1
−2
250000
500000
750000
1000000
Distance From the Origin of Replication (bp)
dN
dS
omega

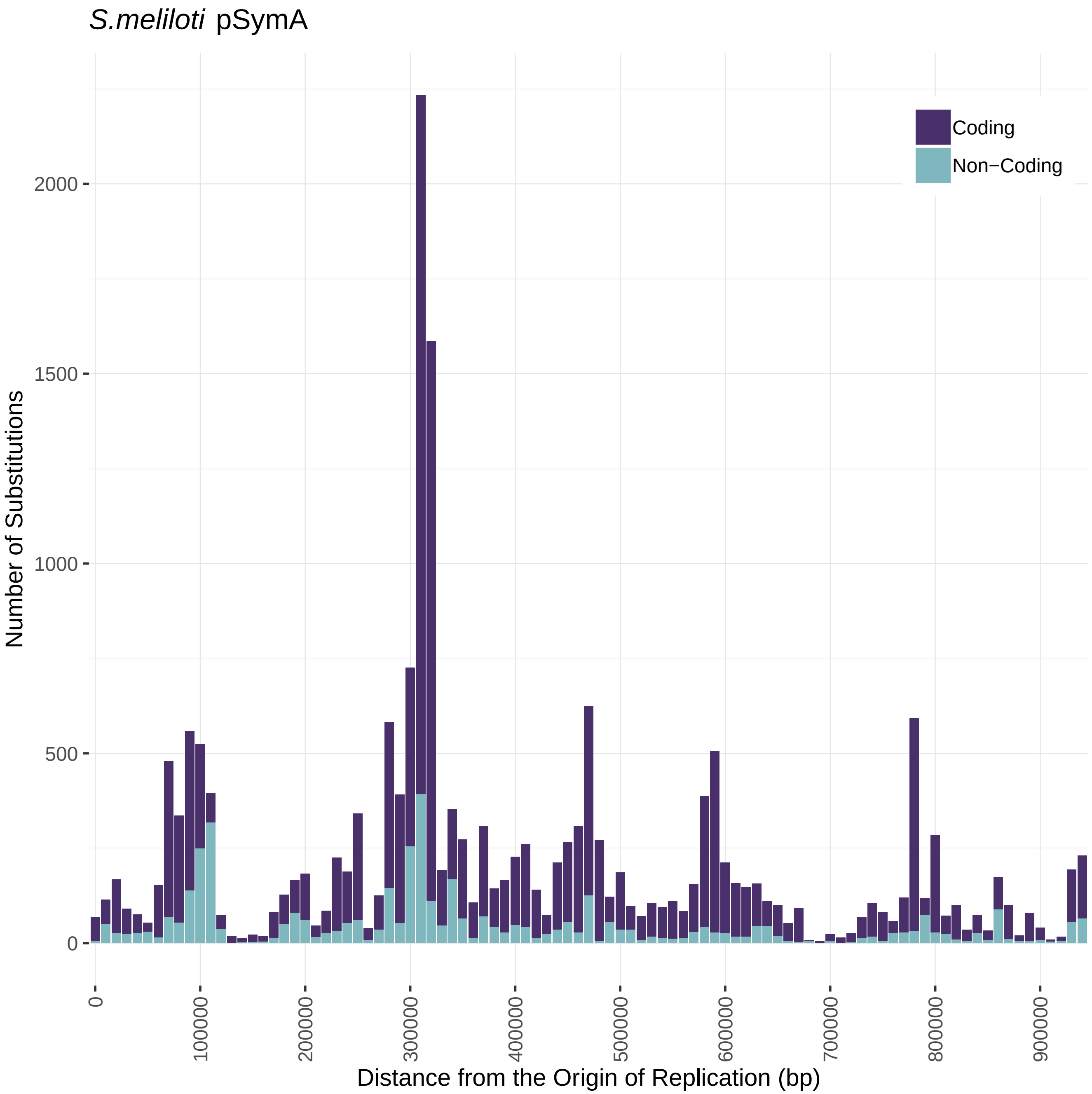

S.meliloti pSymB
Average Value (log10)
1
0
−1
−2
250000
500000
750000
Distance From the Origin of Replication (bp)
dN
dS
omega

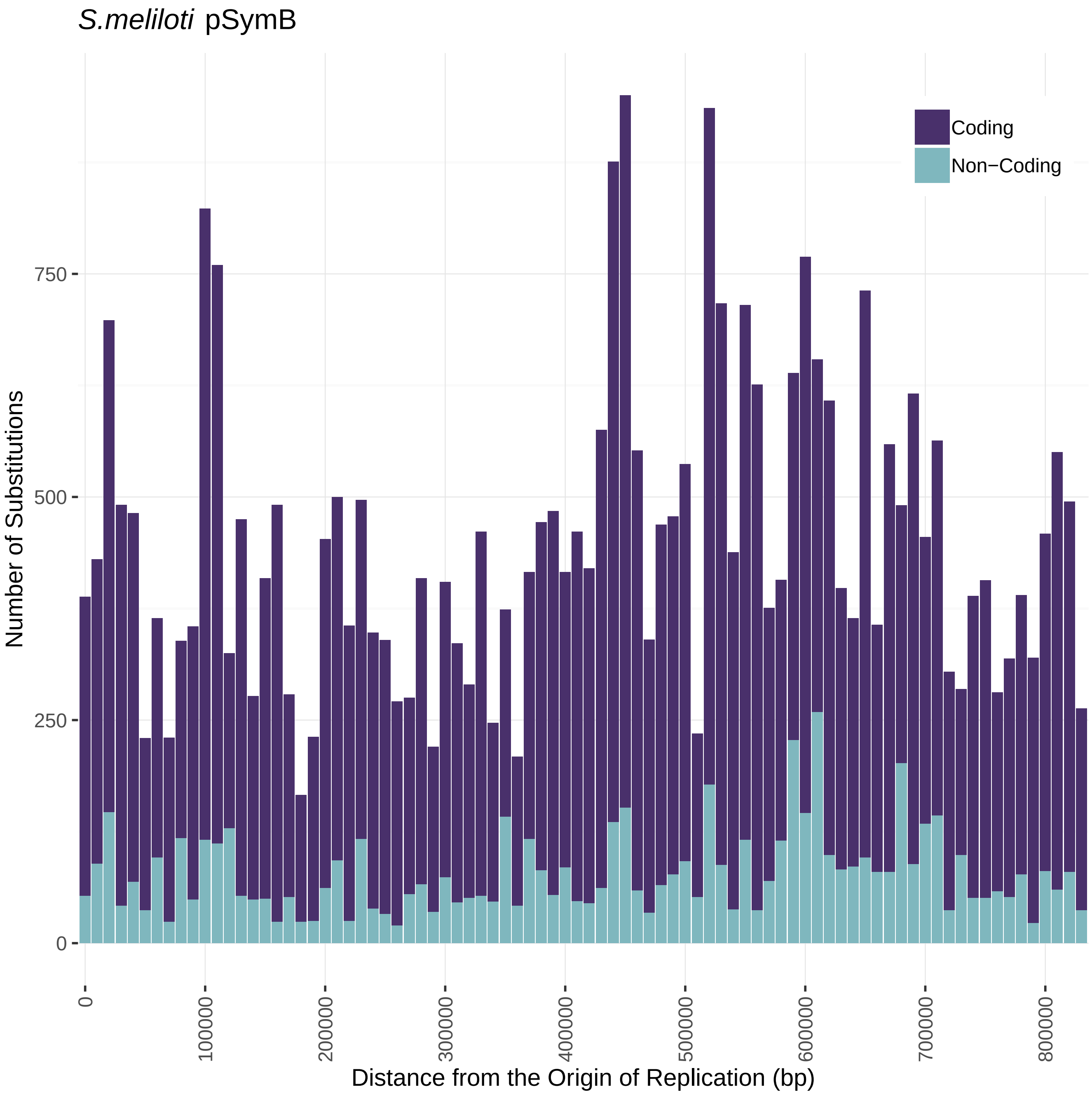

E.coli Chromosome
Coding
Non−Coding
Number of Substitutions
0
500
1000
1500
2000
0
100000
200000
300000
400000
500000
600000
700000
800000
900000
1000000
1100000
1200000
1300000
1400000
1500000
1600000
1700000
1800000
1900000
2000000
2100000
2200000
2300000
2400000
2500000
2600000
2700000
2800000
2900000
3000000
3100000
3200000
Distance from the Origin of Replication (bp)


B.subtilis Chromosome
Coding
Non−Coding
Number of Substitutions
0
500
1000
1500
2000
0
100000
200000
300000
400000
500000
600000
700000
800000
900000
1000000
1100000
1200000
1300000
1400000
1500000
1600000
1700000
1800000
1900000
2000000
2100000
2200000
Distance from the Origin of Replication (bp)


Streptomyces Chromosome
Coding
Non−Coding
Number of Substitutions
0
2000
4000
0
100000
200000
300000
400000
500000
600000
700000
800000
900000
1000000
1100000
1200000
1300000
1400000
1500000
1600000
1700000
1800000
1900000
2000000
2100000
2200000
2300000
2400000
2500000
2600000
2700000
2800000
2900000
3000000
3100000
3200000
3300000
3400000
3500000
3600000
3700000
3800000
3900000
4000000
4100000
4200000
4300000
4400000
4500000
4600000
4700000
4800000
4900000
5000000
5100000
5200000
5300000
5400000
5500000
Distance from the Origin of Replication (bp)


S.meliloti Chromosome
Coding
Non−Coding
Number of Substitutions
0
250
500
750
1000
0
100000
200000
300000
400000
500000
600000
700000
800000
900000
1000000
1100000
1200000
1300000
1400000
1500000
1600000
1700000
1800000
1900000
2000000
Distance from the Origin of Replication (bp)


S.meliloti pSymA
Coding
Non−Coding
Number of Substitutions
0
500
1000
1500
2000
0
100000
200000
300000
400000
500000
600000
700000
800000
900000
Distance from the Origin of Replication (bp)


S.meliloti pSymB
Coding
Non−Coding
Number of Substitutions
0
250
500
750
Distance from the Origin of Replication (bp)
0
100000
200000
300000
400000
500000
600000
700000
800000

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.