Subs Paper Things to Do:

- causes for weird selection and subs results in *Streptomyces*
  - see how often class 4 arises in strep to see what is going on in later portion of the genome (to see if annotation is really a problem)
  - split up the strep data into core and non core and see if results are the same
- ~~make graphs proportional to length of respective cod/non-cod regions~~
- ~~test examples for genes near and far from terminus (robust log reg/results)~~
- ~~linear regression on 10kb regions for weighted and non-weighted substitutions~~
- ~~average number of substitutions in 20kb regions near and far from the origin~~
- ~~figure out why the data is weird for number of cod/non-cod sites~~
- why are the lin reg of dN, dS and ω NS but the subs graphs are...explain!
- grey out outliers in subs graphs?
- mol clock for my analysis?
- GC content? COG? where do these fit?

Gene Expression Paper Things to Do:

- ~~linear regression on 10kb regions~~
- put new 10kb lin reg and # of genes over 10kb lin reg into paper
- write about ↑ in methods and discussion
- put expression lin reg and # coding sites log reg into supplement
- write about ↑ in paper and how results are the same
- update supplementary figures/file
- ~~correlation of gene expression across strains~~
  - ~~make graphs pretty and more informative with label names~~
  - ~~add them to supplement with a mini write up of what we did and why~~
  - ~~mention this in the actual paper~~
- if necessary add a phylogenetic component to the analysis
- potentially remove genes that have been recently translocated from the analysis
- model gene exp + position + number of genes

- split up the strep data into core and non core and see if results are the same
- what is going on with *Streptomyces* number of genes changing drastically from core to non-core
- codon bias?
- what is going on with really high gene expression bars
- edit paper
- submit paper

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
- 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)
- read and make notes on papers I found for dissertation intro

Last Week

✓edits for gene expression intro

proportional coding subs > proportional non-coding subs So I thought that I fixed this, which I did in pSymB, but when I re-ran the pipeline on the other bacteria the number of 10kb sections with proportionally more coding subs than non-coding subs still ranged from 14% - 50% of the 10kb sections.

Last week I did a lot of editing on the gene expression paper. I put the correlation of gene expression across samples into the supplement with an explanation in the main paper. I also moved the linear regression done using each gene as a data point to the supplement (with an explanation in the main paper). I added in the results from the linear regression on the number of genes over 10kb regions and expression over 10kb regions to the main section of the paper and updated the methods section to account for this.

I have also decided to try out the graduate course (Stats 744). So far it is really really cool and I am learning a lot and I think will be really beneficial.

I looked into if in each 10kb section of the substitution graph there was an increased substitution rate in the protein coding sections versus the non-protein coding section. I discussed this with you, but there were cases where one site in my data can be coding in one taxa but non-coding in another taxa. I am for sure comparing homologous sequences so this has to do with the beginning of one genome being coding but the rest being non-coding. When looking into this further I found that the reverse complement of each gene was not being printed out properly in the alignment files (when it splits up into protein coding and non-protein coding sections). I fixed this. I also found that for each site, the code was not printing out information for all the nodes in the tree unless there was a substitution when I made a little test example dataset. So I re-wrote my perl script in python and fixed this issue. So now for each site it will print null information if there was no substitution at a particular site, so I have information for all nodes in the tree for each site. I thought that this would solve the issue of having some 10kb sections with proportionally more coding substitutions than non-coding. So I re-ran the pipeline on pSymB to see if it fixed the issue and there are about 23% of 10kb sections that have the proportion of coding subs > the proportion of non-coding subs. However, the difference between these rates is 0.001 or less. I checked if it was a sample size issue, and it was not. These 10kb sections have the same amount of information as the other 10kb sections. So I am not sure what to do next, or if this is really an issue. I would appreciate any help.

I wrote the code for grabbing the dN, dS, and ω for 20 genes near and far from the origin and doing a linear regression on those genes. I am not going to run this on all the replicons until I figure out what is going on with *Streptomyces*.

I also started to look into why *Streptomyces* is weird with the selection results and I looked into *Escherichia coli* genes that had an ω value > 1. this was only about 4% of genes and they were gene fragments that were really short (average of 33bp). This was because there were gaps in the genes so the rest of the gene got "cut" so it makes sense why these genes have such high ω. I am not sure if we should be discarding these gene fragments from the analysis? Thoughts? When

I checked in with *Streptomyces*, about 70% of the genes had $\omega > 1$, which is a problem. But this is why the overall selection results for *Streptomyces* are weird. The average length of these gene fragments is 254bp. So I first checked to make sure that I was grabbing the correct sequences from the progressiveMauve annotation, and I am. All the positions in the headers of my MAFFT files match the actual genomic sequence at that position. So that is good. When I look at the alignments of these segments, they are pretty jumbled and there are a lot of substitutions happening that are changing the AA sequence. So I am a bit lost and do not know what else to check to make sure that these differences are real. Any help would be appreciated. I also see that all the venezuele strains have very similar sequences and the coelicolor and lividans strains have similar sequences to eachother. Not sure if this has anything to do with the substitutions.

This Week

I would like to figure out why *Streptomyces* has $\omega > 1$ for 70% of its genes. Not sure how I should go about looking into this.

There is a portion of my substitutions code where I remove duplicates in my data, however, I think this is wrong and I need all the information across all nodes. So I would like to re run this for pSymB and see if this is what is causing some of the 10kb sections to have proportionally more coding subs than non coding.

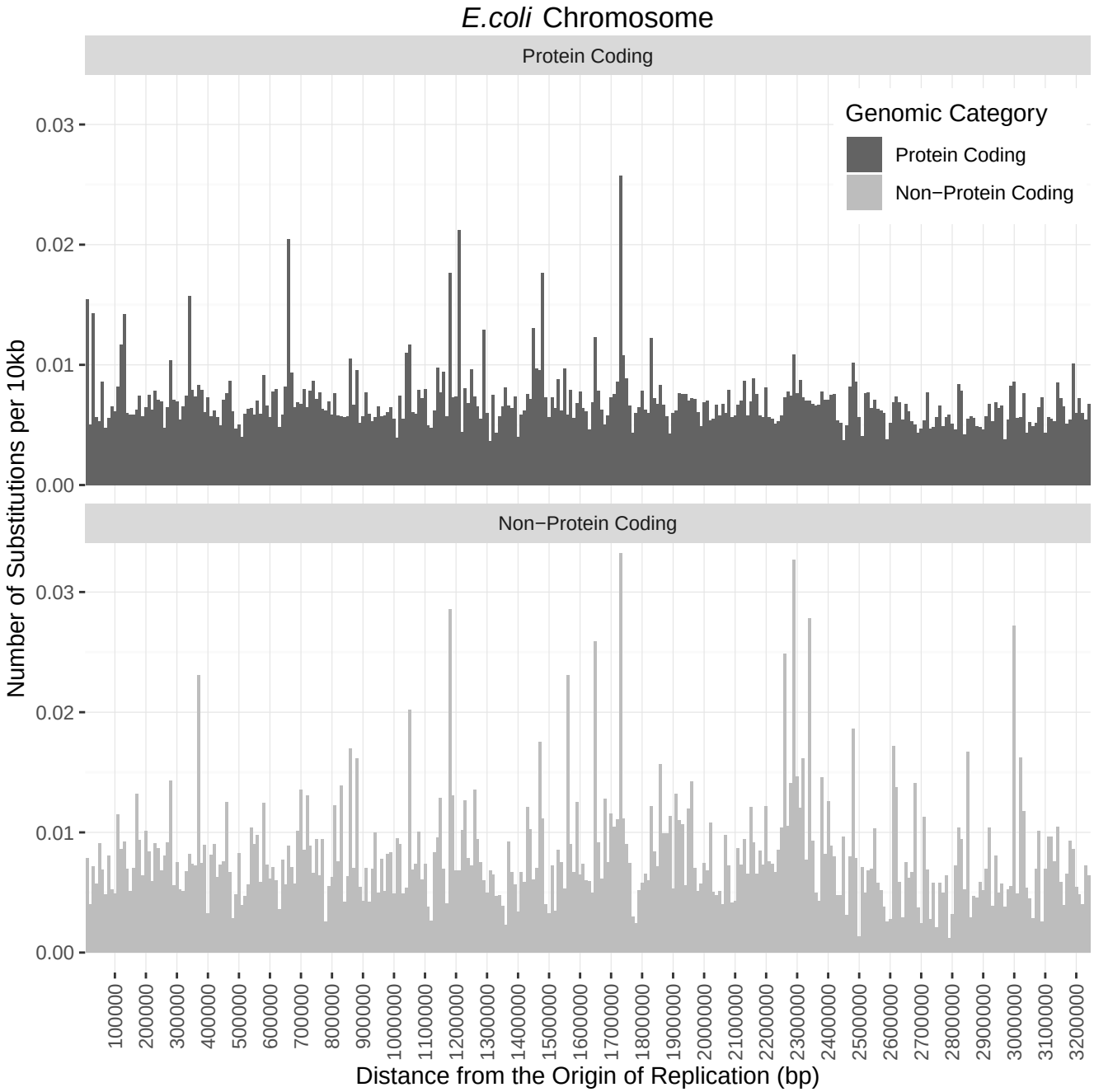
Next Week

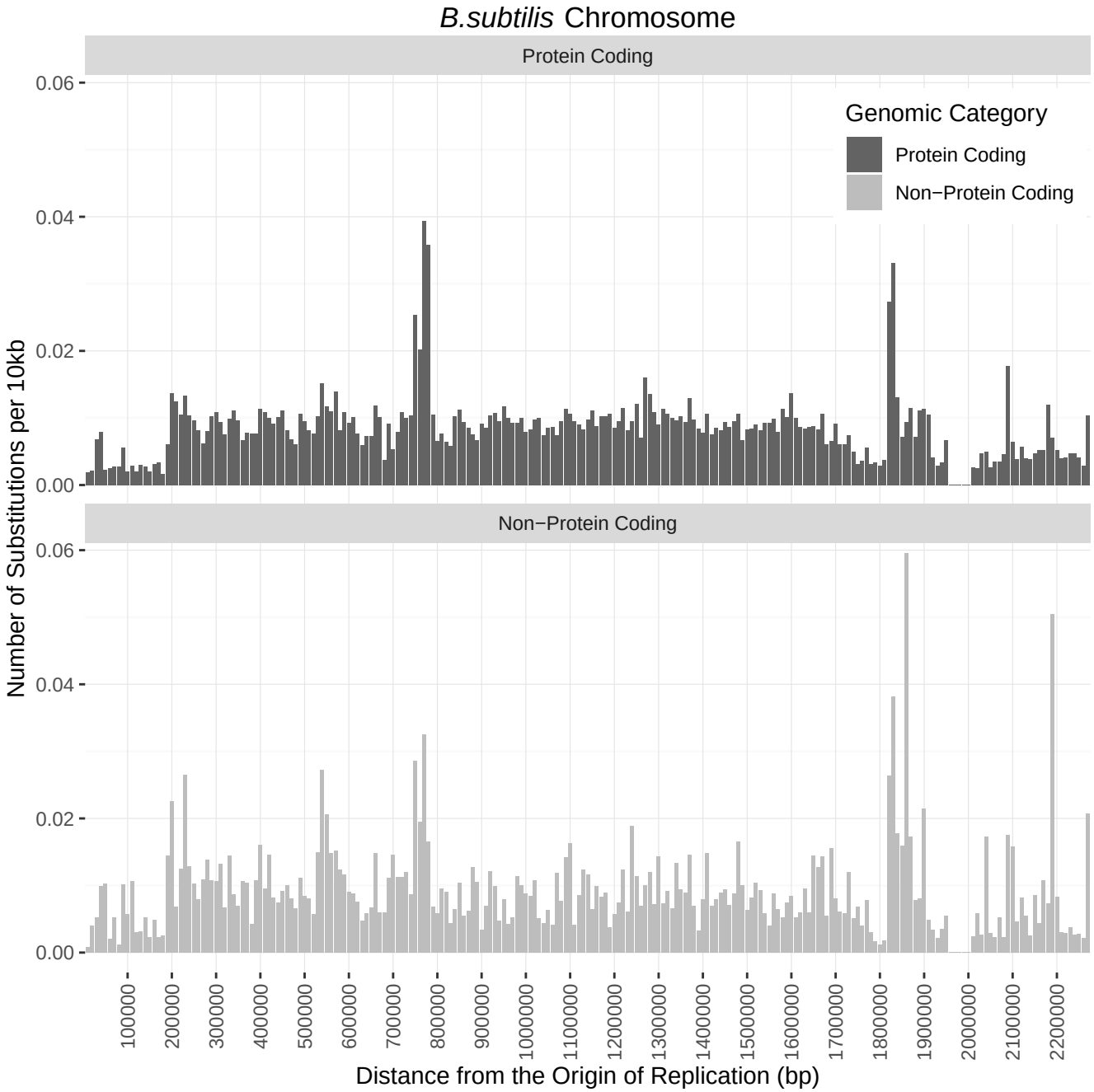
I would like to switch and work on the gene expression paper again.

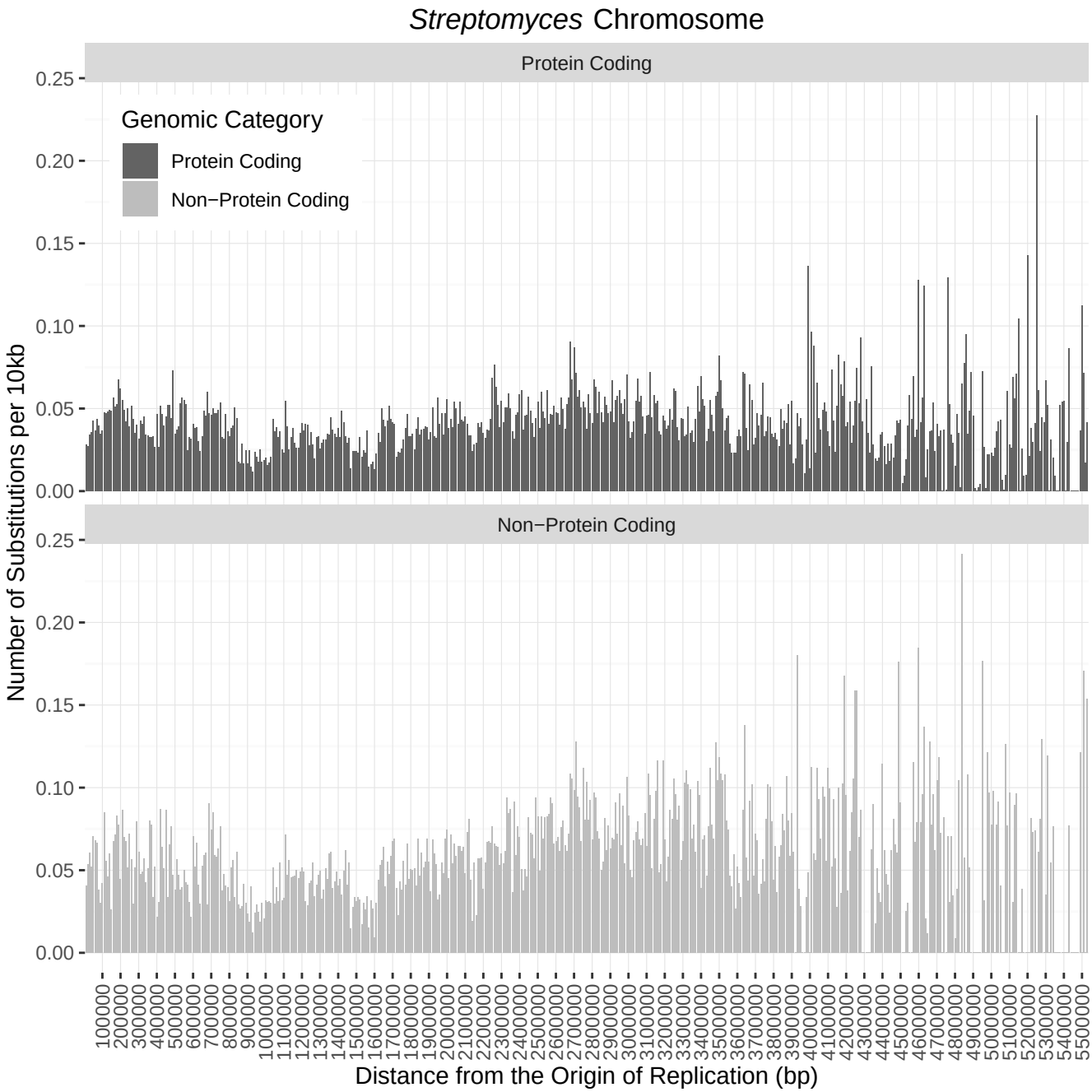
1. phylogenetic analysis with gene expression in *E. coli*?
2. remove genes that have been recently translocated from analysis?
3. model gene expression + position + number of genes
4. split up *Streptomyces* data into core and non-core and see if the results are the same (do same for number of genes)

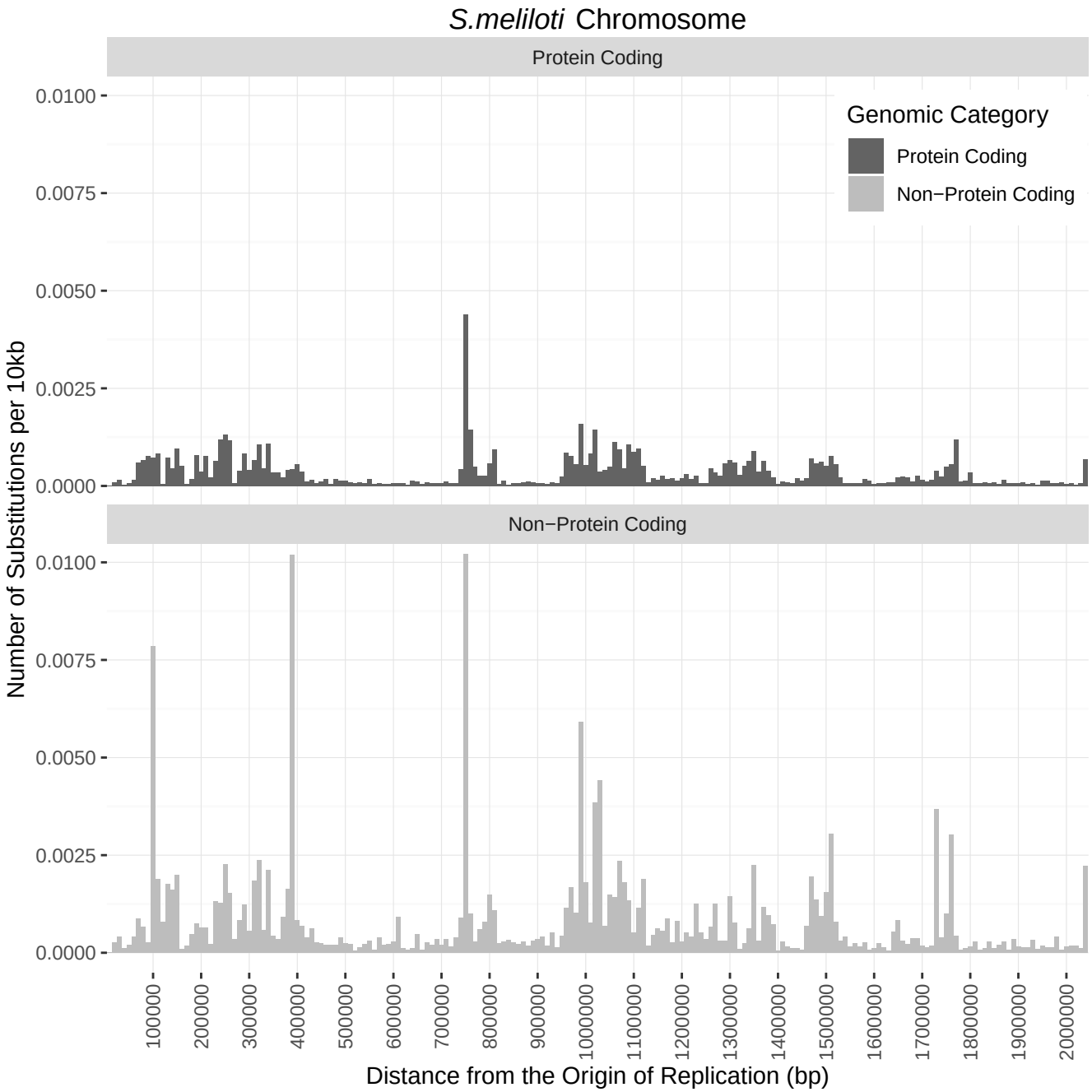
	Near Origin			Near Terminus		
Bacteria and Replicon	*dN*	*dS*	ω	*dN*	*dS*	ω
E. coli Chromosome	NS	NS	NS	NS	NS	NS
B. subtilis Chromosome	NS	NS	NS	NS	NS	NS
Streptomyces Chromosome	NS	NS	NS	NS	NS	NS
S. meliloti Chromosome	2.79×10^{-8}*	NS	NS	NS	NS	NS
S. meliloti pSymA	NS	NS	3.42×10^{-5}*	NS	NS	NS
S. meliloti pSymB	NS	NS	NS	-3.24×10^{-7}**	8.33×10^{-6}***	NS

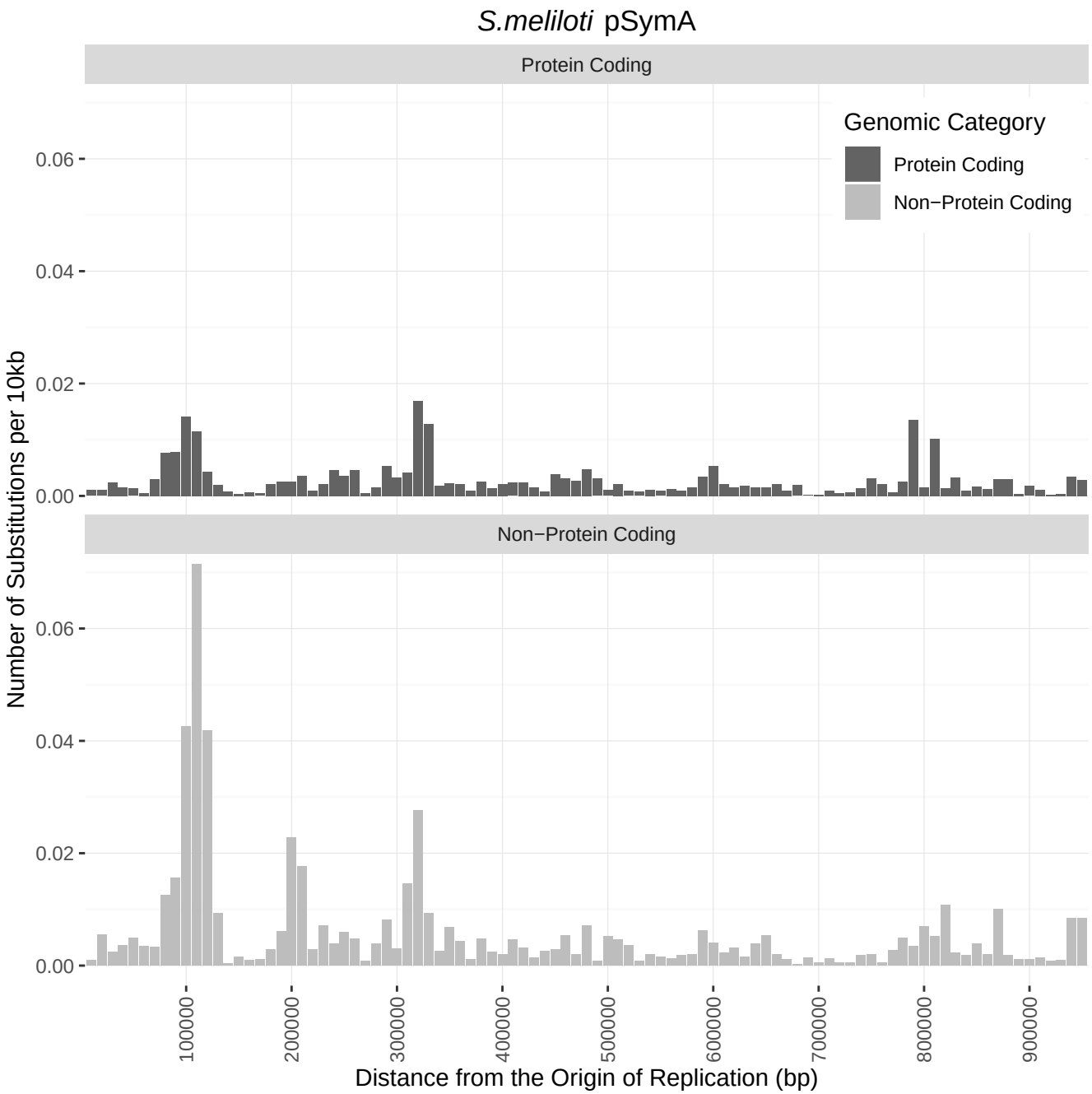
Table 1: Linear regression for dN, dS, and ω calculated for each bacterial replicon for the 20 genes closest and 20 genes farthest from the origin of replication. All results are marked with significance codes as followed: p: < 0.001 = '***', $0.001 < 0.01$ = '**', $0.01 < 0.05$ = '*', > 0.05 = 'NS'.

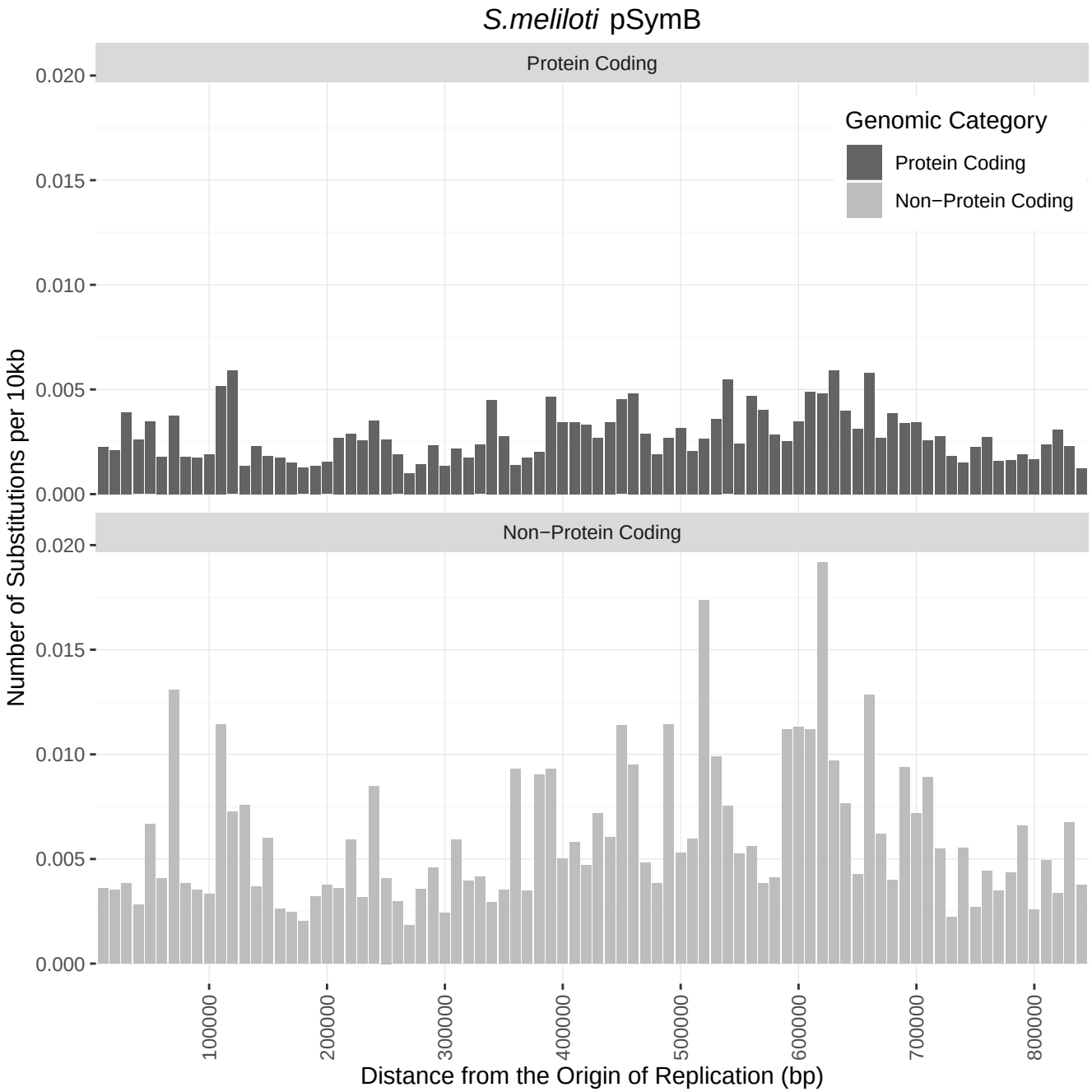

E.coli Chromosome
Protein Coding
Non−Protein Coding
Genomic Category
Protein Coding
Non−Protein Coding
Number of Substitutions per 10kb
0.00
0.01
0.02
0.03
100000
200000
300000
400000
500000
600000
700000
800000
900000
1000000
1100000
1200000
1300000
1400000
1500000
1600000
1700000
1800000
1900000
2000000
2100000
2200000
2300000
2400000
2500000
2600000
2700000
2800000
2900000
3000000
3100000
3200000
Distance from the Origin of Replication (bp)


B.subtilis Chromosome
Protein Coding
Non−Protein Coding
Genomic Category
Protein Coding
Non−Protein Coding
Number of Substitutions per 10kb
0.06
0.04
0.02
0.00
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
Protein Coding
Non−Protein Coding
Genomic Category
Protein Coding
Non−Protein Coding
Number of Substitutions per 10kb
0.25
0.20
0.15
0.10
0.05
0.00
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
Protein Coding
Non−Protein Coding
Genomic Category
Protein Coding
Non−Protein Coding
Number of Substitutions per 10kb
0.0100
0.0075
0.0050
0.0025
0.0000
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
Protein Coding
Non−Protein Coding
Genomic Category
Protein Coding
Non−Protein Coding
Number of Substitutions per 10kb
0.06
0.04
0.02
0.00
100000
200000
300000
400000
500000
600000
700000
800000
900000
Distance from the Origin of Replication (bp)


S.meliloti pSymB
Protein Coding
Non−Protein Coding
Genomic Category
Protein Coding
Non−Protein Coding
Number of Substitutions per 10kb
0.020
0.015
0.010
0.005
0.000
100000
200000
300000
400000
500000
600000
700000
800000
Distance from the Origin of Replication (bp)

Bacteria and Replicon	Protein Coding Sequences	Non-Protein Coding Sequences
E. coli Chromosome		NS
B. subtilis Chromosome	-4.971×10^{-8}***	-1.055×10^{-7}***
Streptomyces Chromosome		
S. meliloti Chromosome	-1.903×10^{-7}***	-2.900×10^{-7}***
S. meliloti pSymA	-6.642×10^{-7}***	-1.263×10^{-6}***
S. meliloti pSymB	1.769×10^{-7}***	4.771×10^{-7}***

Table 2: 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'.

	Protein Coding				Non-Protein Coding			
	Correlation Coefficient 20kb Near		Number of Substitutions per 20kb Near		Correlation Coefficient 20kb Near		Number of Substitutions per 20kb Near	
Bacteria and Replicon	Origin	Terminus	Origin	Terminus	Origin	Terminus	Origin	Terminus
E. coli Chromosome	-2.889×10^{-5}*	NS	2.87×10^{-2}	4.24×10^{-2}	-4.316×10^{-5}**	-8.209×10^{-5}*	1.095×10^{-2}	4.45×10^{-3}
B. subtilis Chromosome	NS	1.863×10^{-5}*			1.017×10^{-4}*	5.823×10^{-5}***	8×10^{-4}	6.75×10^{-3}
Streptomyces Chromosome								
S. meliloti Chromosome	NS	NS	4.05×10^{-3}	2×10^{-4}	NS	NS	9×10^{-4}	1.5×10^{-4}
S. meliloti pSymA	NS	NS	6.15×10^{-3}	1.9×10^{-3}	1.403×10^{-4}***	-2.220×10^{-4}**	2.8×10^{-3}	5.5×10^{-4}
S. meliloti pSymB					NS	-4.557×10^{-5}**	5.1×10^{-3}	5.4×10^{-3}

Table 3: Logistic regression on 20kb closest and farthest from the origin of replication after accounting for bidirectional replication and outliers. All results are marked with significance codes as followed: < 0.001 = '***', $0.001 < 0.01$ = '**', $0.01 < 0.05$ = '*', > 0.05 = 'NS'.

	Protein Coding		Non-Protein Coding	
Bacteria and Replicon	Weighted	Non-Weighted	Weighted	Non-Weighted
E. coli Chromosome	-4.87×10^{-10}**	-1.839×10^{-4}***	NS	-2.244×10^{-5}***
B. subtilis Chromosome	NS	-2.031×10^{-4}**	NS	-2.885×10^{-5}**
Streptomyces Chromosome				
S. meliloti Chromosome	-1.341×10^{-10}**	-1.461×10^{-5}**	-3.490×10^{-10}*	NS
S. meliloti pSymA	NS	NS	-1.144×10^{-8}**	-6.74×10^{-5}**
S. meliloti pSymB	NS	NS	NS	NS

Table 4: Linear regression on 10kb sections of the genome with increasing distance from the origin of replication after accounting for bidirectional replication. Weighted columns have the total number of substitutions in each 10kb section of the genome divided by the total number of protein coding and non-protein coding sites in the genome. Non-weighted columns are performing a linear regression on the total number of substitutions in each 10kb section of the genome. 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	Gene Expression 10kb
E. coli Chromosome	-2.742×10^{-5}**
B. subtilis Chromosome	-2.198×10^{-5}*
Streptomyces Chromosome	-5.230×10^{-7}***
S. meliloti Chromosome	NS
S. meliloti pSymA	NS
S. meliloti pSymB	NS

Table 5: Linear regression analysis of the median counts per million expression data for 10kb segments of 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. 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	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.

Bacteria and Replicon	Coefficient Estimate
E. coli Chromosome	NS
B. subtilis Chromosome	-2.682×10^{-6}***
Streptomyces Chromosome	-2.360×10^{-6}***
S. meliloti Chromosome	-2.074×10^{-6}***
S. meliloti pSymA	NS
S. meliloti pSymB	-4.19×10^{-6}*

Table 7: Linear regression analysis of the total number of protein coding genes per 10kb 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. 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	NS
B. subtilis Chromosome	NS	NS	-9.08×10^{-6}*
Streptomyces Chromosome	NS	NS	NS
S. meliloti Chromoeom	NS	NS	NS
S. meliloti pSymA	NS	NS	NS
S. meliloti pSymB	NS	NS	1.163×10^{-5}*

Table 8: 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 9: 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	1.0468	0.1330	1.3183	0.6491	0.0364	0.2432
B. subtilis Chromosome	4.652	0.2333	2.4200	1.0879	0.0703	0.3852
Streptomyces Chromosome	13.4950	2.0973	21.0423	5.1256	0.8911	8.9146
S. meliloti Chromosome	0.0184	0.0012	0.1069	0.0187	0.0013	0.0962
S. meliloti pSymA	1.0602	0.7451	5.1290	0.4100	0.0863	0.8311
S. meliloti pSymB	3.2602	0.0256	0.3878	0.1436	0.0100	0.1943

Table 10: 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
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 11: Summary of strains and species found for each gene expression analysis. Gene Expression Omnibus accession numbers and date accessed are provided.