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 to gene expression paper intro

✓ re-run selection analysis with fixed gene complement issue

✓ look into/test Gblocks alignment assessment

✓ look into other *Streptomyces* genomes available

✓ figure out why *Streptomyces* failed to split the alignment into coding and non-coding

Gblocks So I looked into Gblocks and what it does. It seems very similar to trimal, so it breaks the alignment up into good and bad segments. It spits out this big alignment into phylip interleaved format so it should be pretty easy to write a code to parse this and get the alignment into a format I can use. So for *Streptomyces*, Gblocks will throw out between 10% - 99% of the alignment for all the blocks. It also had only 30% of the blocks where 50% or more of the sequence was kept. The segments of alignment that it kept were between 10bp - 500bp. Comparing this to *E. coli*, 67% of the blocks had 50% or more of the sequence kept and the segments of alignment that were kept were much longer than in *Streptomyces* and ranged from 10bp - 30000bp. Some of these chunks (for *Streptomyces* and *E. coli*) are really small and I think these small segments are also showing up in my code and giving weird selection coefficients. So I think having some sort of alignment length cut off, plus assessing the alignment with Gblocks will help eliminate these and then the selection results will hopefully look better and be more accurate.

I then looked closely at what Gblocks does v.s. what my code does. Most of the time my code and Gblocks lines up, BUT there are cases where I kept something that Gblocks tossed, and cases where Gblocks kept something that I tossed. So I am not sure what this means but I think using both will be the most stringent.

I also looked at the smaller segments of the alignment of *E. coli* and found that the alignment was not aligning homologous sequences either, but only for the smaller segments. So I think we can't trust anything that is really short because we just can not be confident that it is correct.

$dN/dS > 1$ for some genes

As mentioned above, these really short segments are just inaccurate. So I think making a cut off and using Gblocks should hopefully fix this issue

Streptomyces Alternate Genomes

I found a paper that shows multiple phylogenetic trees for *Streptomyces* and found some consistent groupings of strains. I tried to find genomes related to *S. venezuelae* but when I did progressiveMauve on these, there was still way too many blocks. I am going to talk to Marie to see if she has any insight on this. There may be more hope for genomes related to *S. coelicolor*, so I am going to look into this more and see if I can get a progressiveMauve alignment that is less crazy.

Alignment split

The reason why one of the blocks for *Streptomyces* failed to split is that the sequences are not homologous. The code is a bit complicated to explain but the reason that it failed is that some taxa had an extra "gene" than others. But again, this is because the sequences are not homologous so this should be fixed when I implement the above.

This Week

- look for more *Streptomyces* phylo trees and associated genomes (*S. coelicolor*)

- check coding/non-coding code to see how often sites do not match up for *Streptomyces*

- check how often the tree for each block matches the overall tree for *Streptomyces*

- figure out what is going on with cod subs > non-cod subs

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)

Bacteria and Replicon	Near Origin			Near Terminus		
	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	—	—	—	—	—	—
S. meliloti Chromosome	3.77×10^{-8}**	3.54×10^{-7}**	1.23×10^{-6}**	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

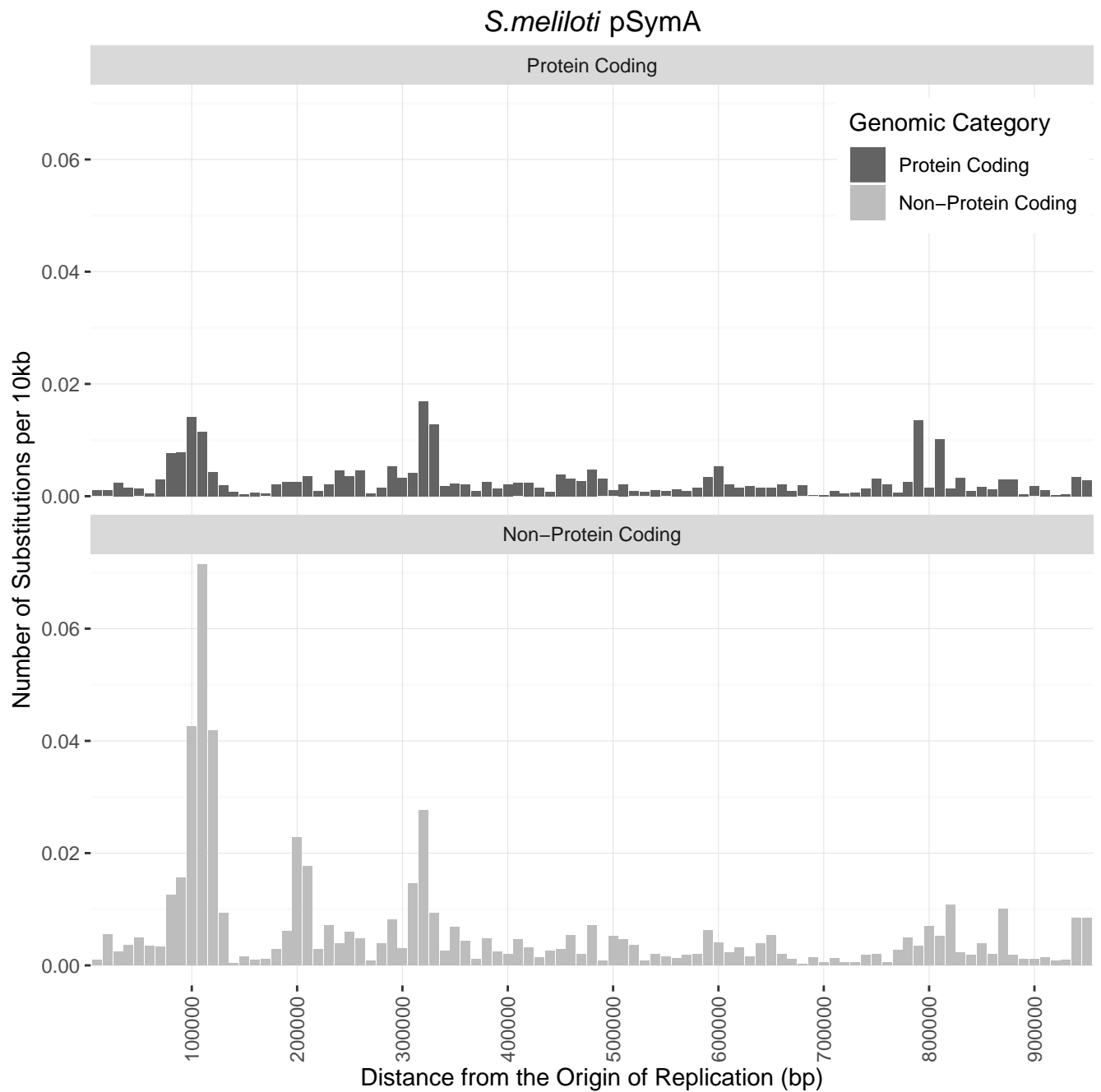Table 1: Linear regression for dN, dS, and ω calculated for each bacterial replicon for the 20 genes closest and 20 genes farthest from the origin of replication. All results are marked with significance codes as followed: p: $< 0.001 = $ '***', $0.001 < 0.01 = $ '**', $0.01 < 0.05 = $ '*', $> 0.05 =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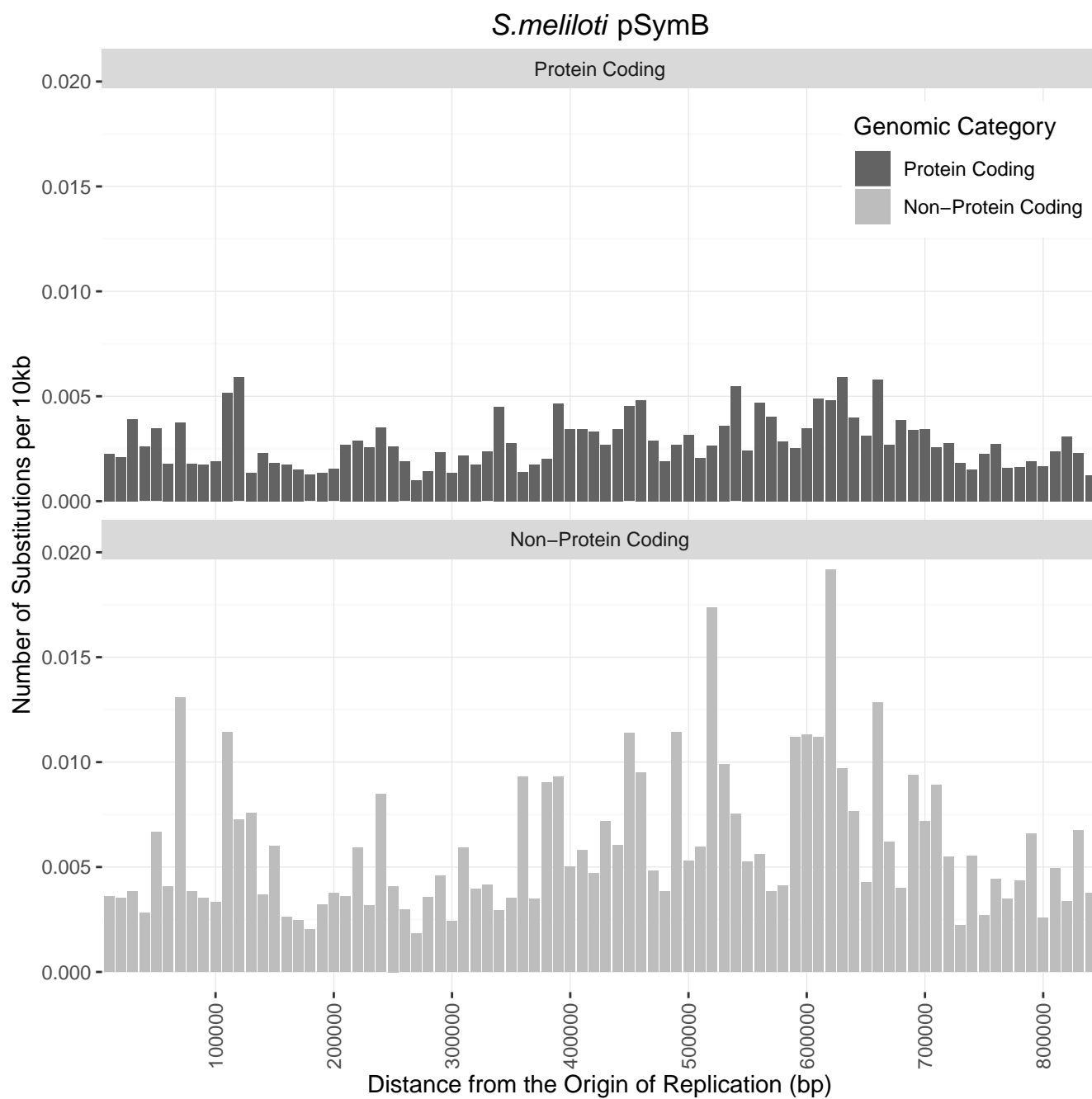
Streptomyces Chromosome

S.meliloti Chromosome

S.meliloti pSymA
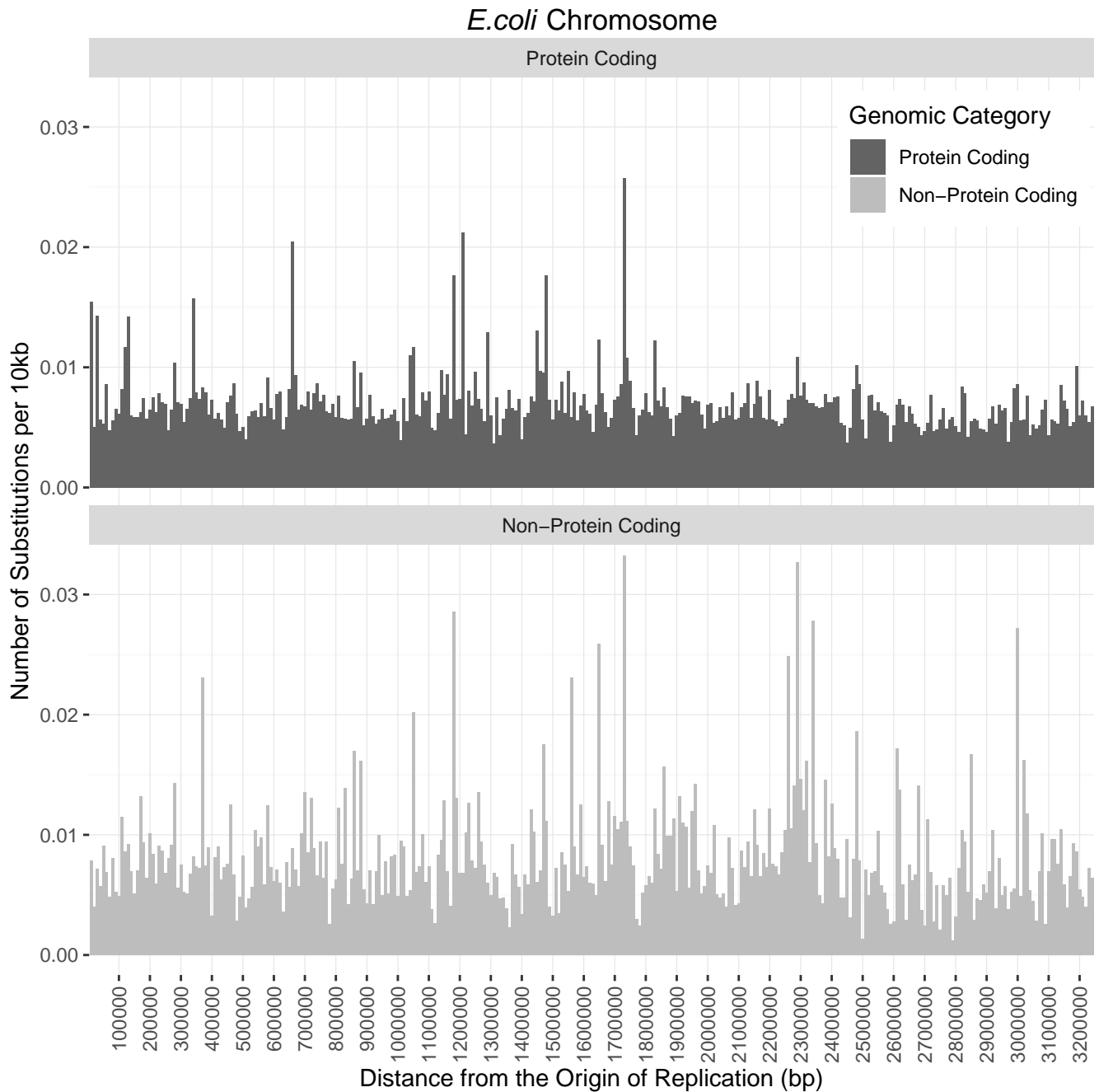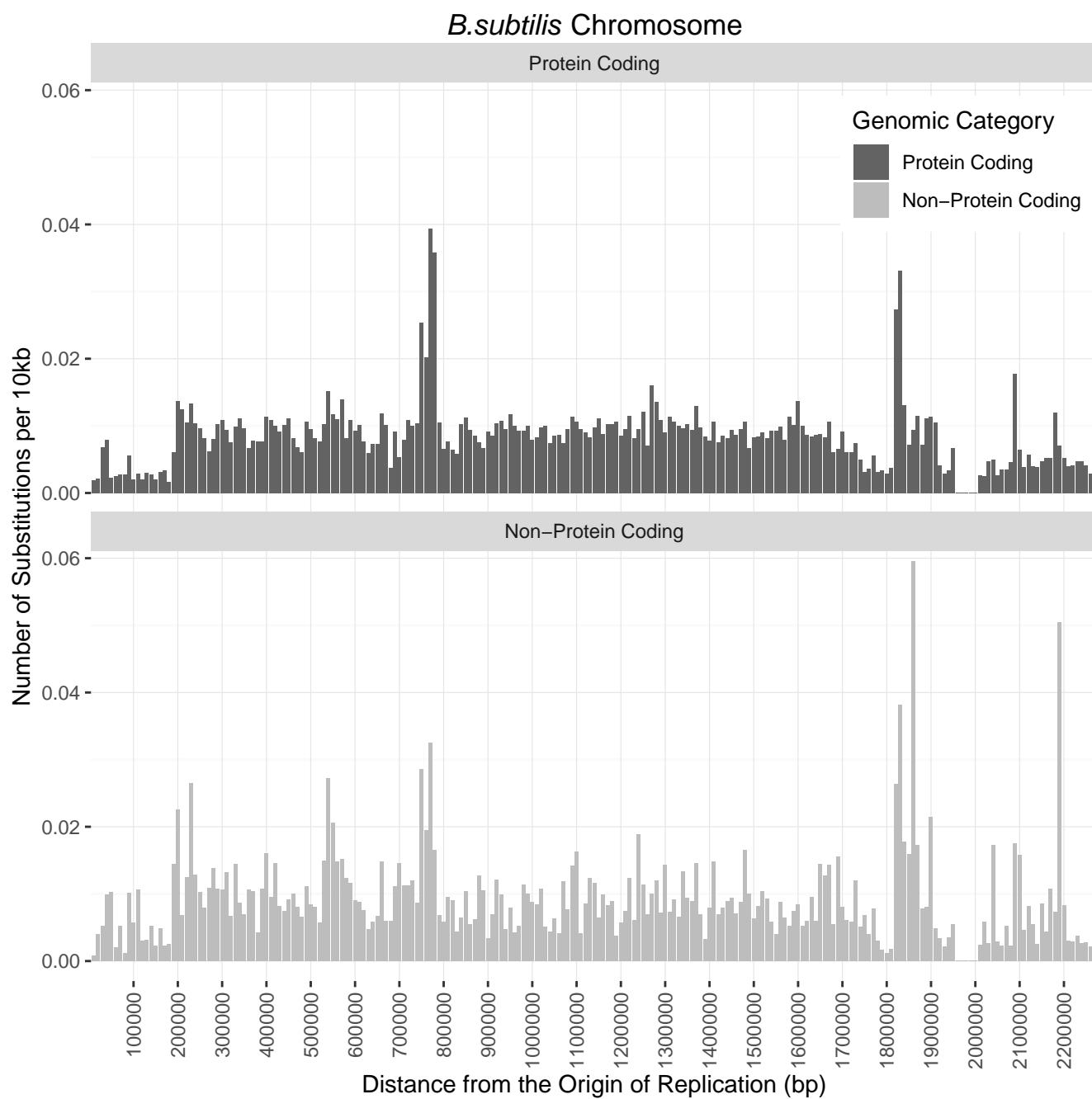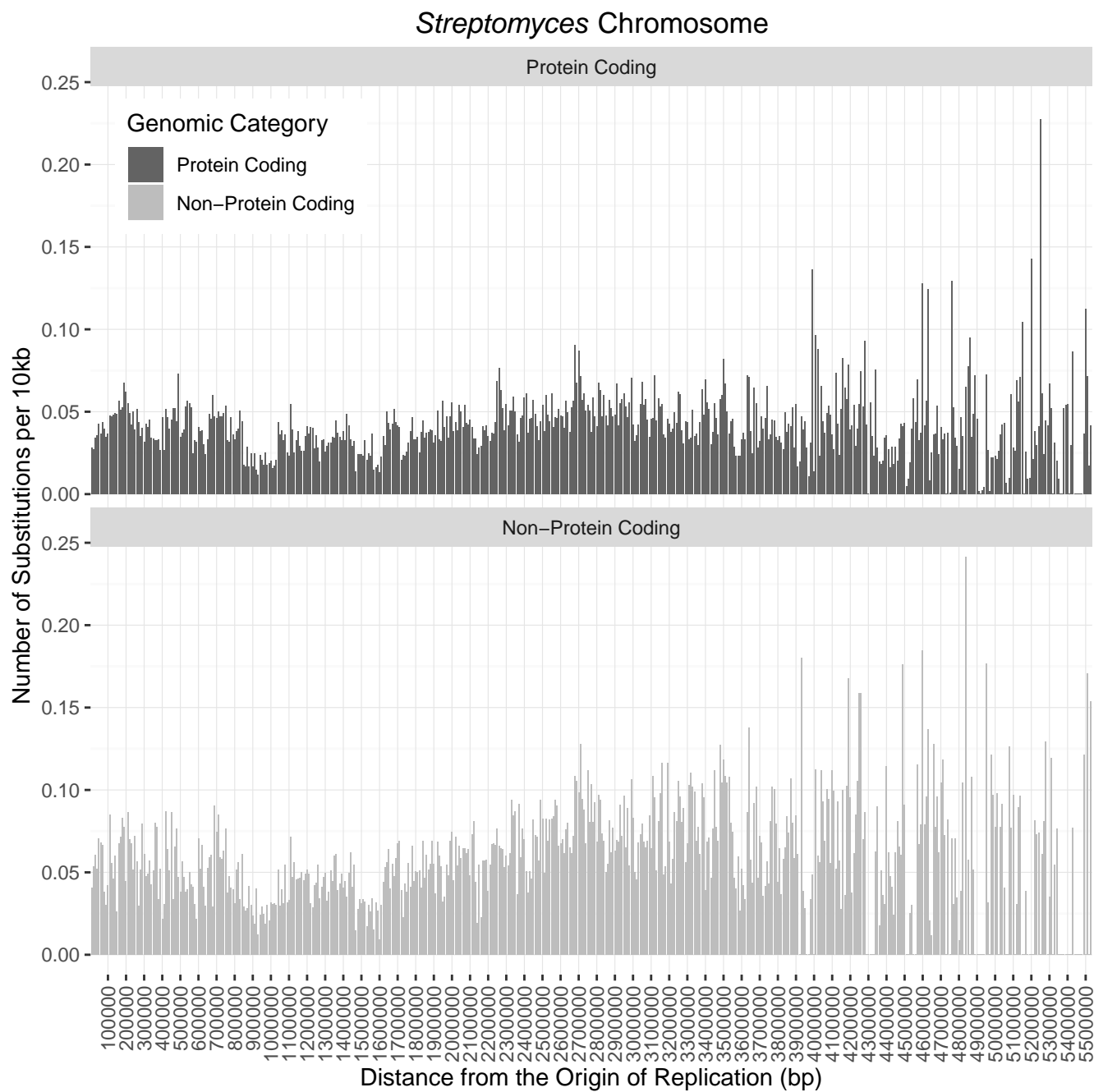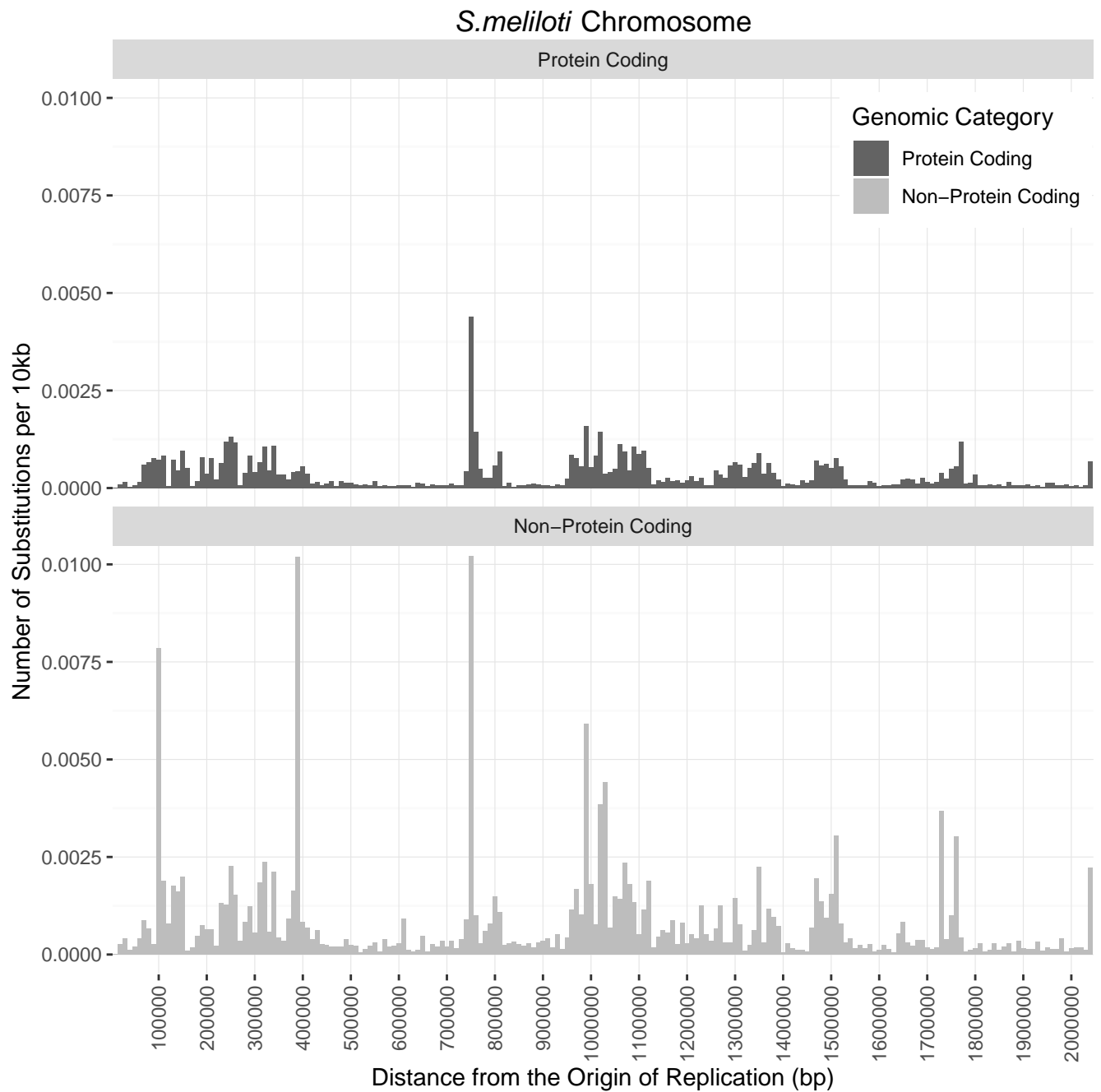
S.meliloti pSymB

Bacteria and Replicon	Protein Coding Sequences	Non-Protein Coding Sequences
E. coli Chromosome	-4.308×10^{-8}***	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}*	4.8×10^{-3}	3.06×10^{-2}	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	-1.553×10^{-5}*	-4.908×10^{-5}***	3.23×10^{-2}	2.36×10^{-2}	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'.

Bacteria and Replicon	Protein Coding		Non-Protein Coding	
	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	$-2.742 \times 10^{-5}**$
B. subtilis Chromosome	$-2.198 \times 10^{-5}*$
Streptomyces Chromosome	$-5.230 \times 10^{-7}***$
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}	$<2 \times 10^{-16}$
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 ($5.34.7 \times 10^{-1}$)

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.

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.

Bacteria and Replicon	Gene Average			Genome Average		
	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.