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

✓linear regression on 10kb regions for substitutions (weighted and non-weighted) (Table 2)

✓average number of substitutions in 20kb regions near and far from the origin (Table 1)

✓add code to check protein cod subs > non-protein coding subs

✓figured out weird *Streptomyces* number of cod and non-cod sites (Figure below)

✓correlation of gene expression across strains (Figures 1 and 2)

Last week I did linear regression on the total number of substitutions in 10kb sections of the genome as well as a proportional number of substitutions for protein coding and non-protein coding sites (Table 2). Mostly everything is not significant or negative which is good! The only weird thing is that *Streptomyces* is positive in non-protein coding when looking at proportional number of substitutions, but is negative when looking at raw counts of substitutions. I did a linear regression on the number of protein coding and non-protein coding sites and there is a significant negative correlation between the number of sites and distance from the origin of replication. So I think that there are just less sites here so that is why the proportion of substitutions per site increases near the end, but the raw number of substitutions is low. I will come discuss this with you later.

I also calculated an average number of substitutions in 20kb regions near and far from the origin, these are found in Table 1. The origin does have a higher number of substitutions near the origin than compared to the terminus, which is great! However, the number of substitutions is higher in protein coding regions than in non-protein coding regions. I also wrote a little script to check if the proportional number of substitutions was higher in each 10kb region for non-protein coding v.s. protein coding, and I found the opposite. That the proportional number of substitutions (number of subs divided by number of sites) is higher in protein coding than non-protein coding in 60% of the *E. coli* 10kb regions. I am still looking into this, I think that it may be something wrong with how I split up the alignments into protein coding and non-protein coding. So I need to look at this code carefully.

I realized that for *Streptomyces*, the weird number of protein coding and non-protein coding sites was because of zero substitutions in some sections messing up the proportional calculation. The new graph is below!

As I mentioned to you I looked into the correlation of gene expression across strains for that paper and it looks like all the datastes are roughly the same, which is great! *Streptomyces*, and all the replicons of *S. meliloti* only had one dataset, so I could not look into this with them. I thought that *Streptomyces* had more than one dataset but I was wrong and accidentally counted one of the replicates as it's own dataset! The weird genes that had super high expression were all hypothetical proteins or unknown proteins. The graphs for *E. coli* and *B. subtilis* are in Figures 1 and 2.

This Week

I would like to really investigate why *Streptomyces* has such weird values for dN, dS, and ω. I plan on first checking that for all the bacteria ω is not > 1 for any gene. Then calculating dN, dS, and ω by hand for a few *Streptomyces* genes to see what happens.

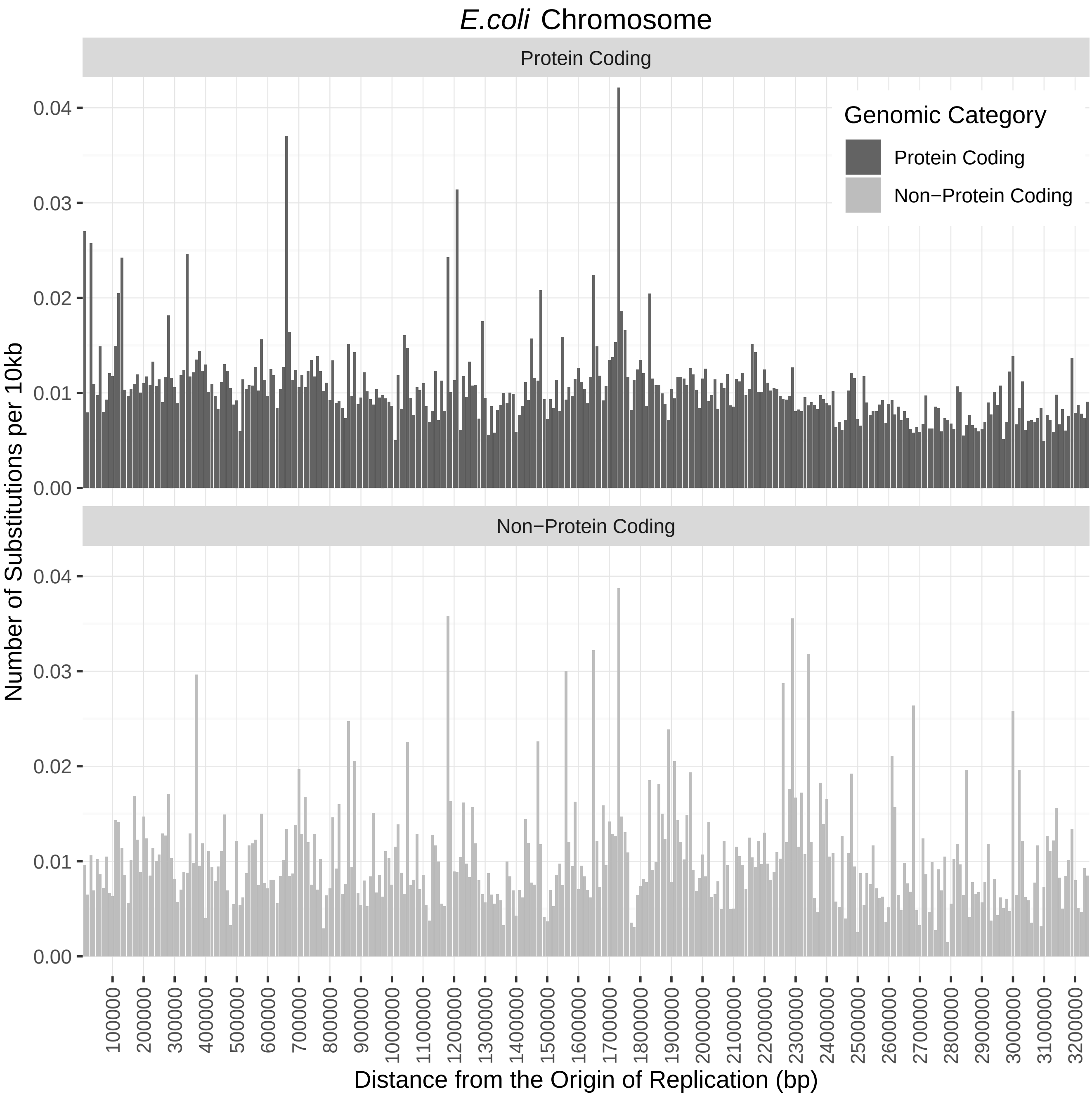
As mentioned above I need to re-look at why the proportionate number of substitutions is higher in the protein coding regions than non-protein coding sections. I think this has to do with how I split up the alignment into cod and non-cod sections so I will be looking into this further this week.

I would like to have the code ready for after I figure out the above stuff looking at 20 genes near and far from the origin to see what their dN, dS, and ω values look like.

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)


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
0.04
Distance from the Origin of Replication (bp)
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

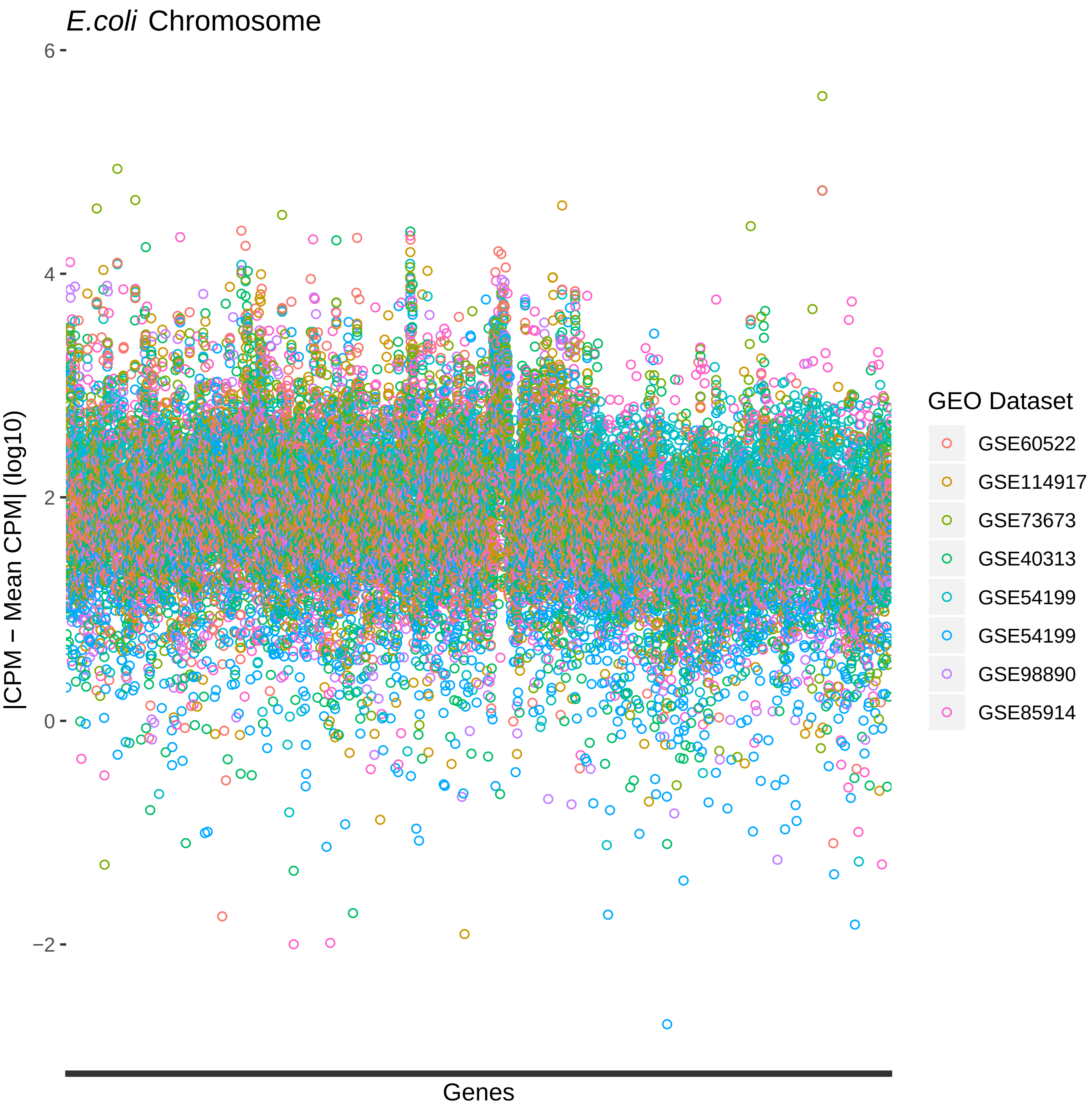


Figure 1: Distribution of the median expression value for each *E. coli* dataset minus the mean expression value for that gene across all datasets. Each gene is shown on the x-axis and the log base 10 values are on the y-axis.

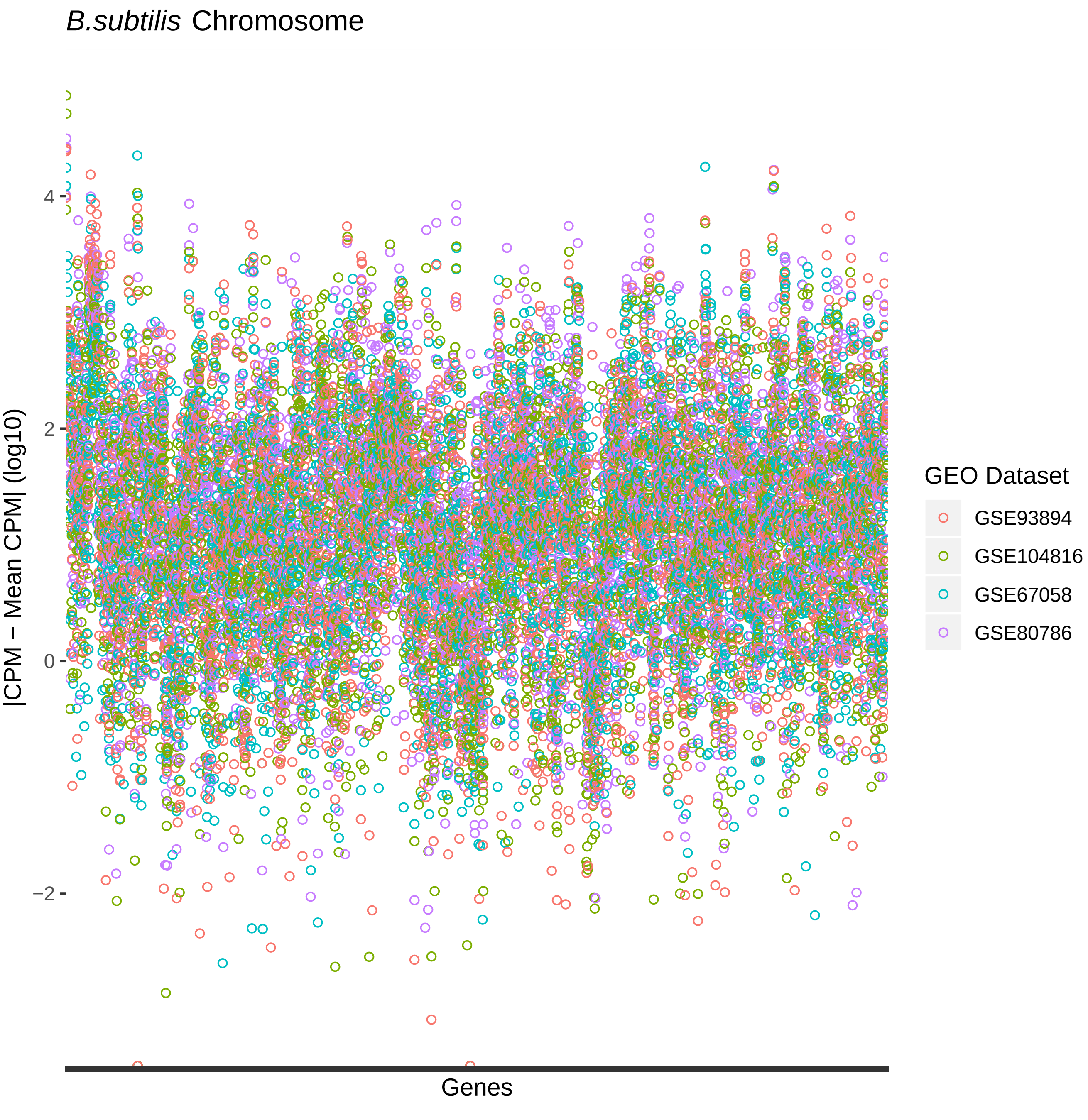
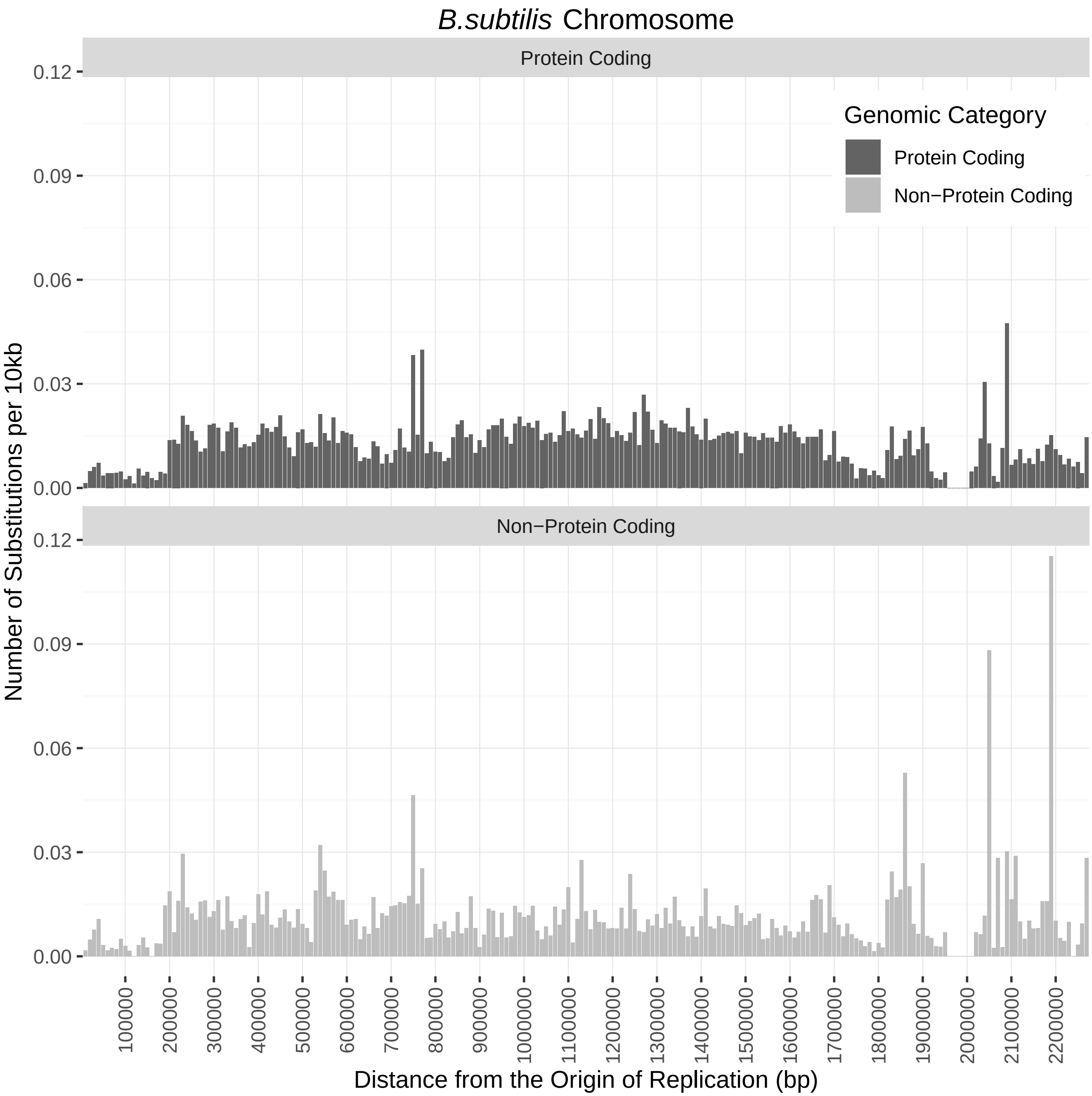
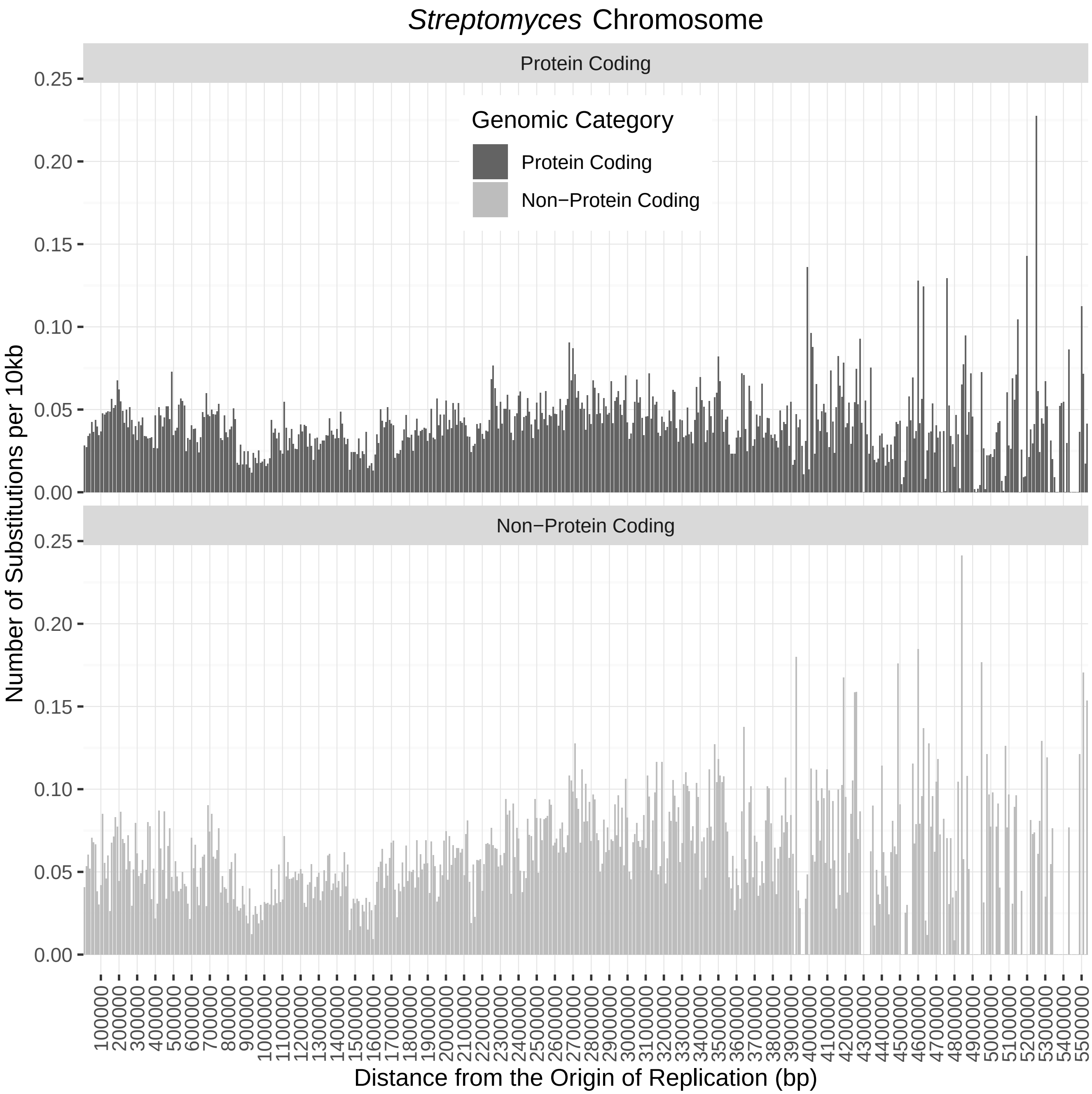
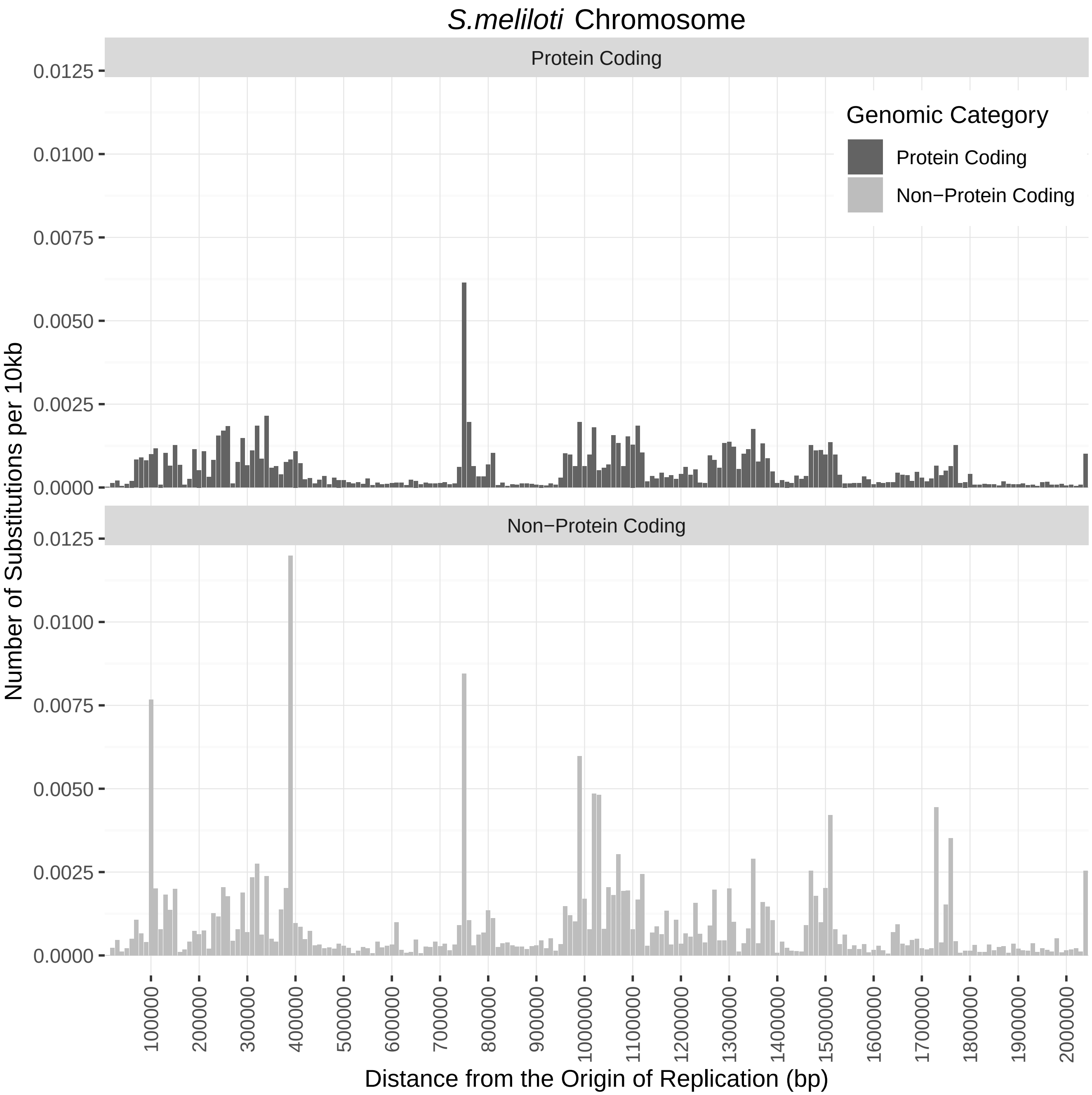
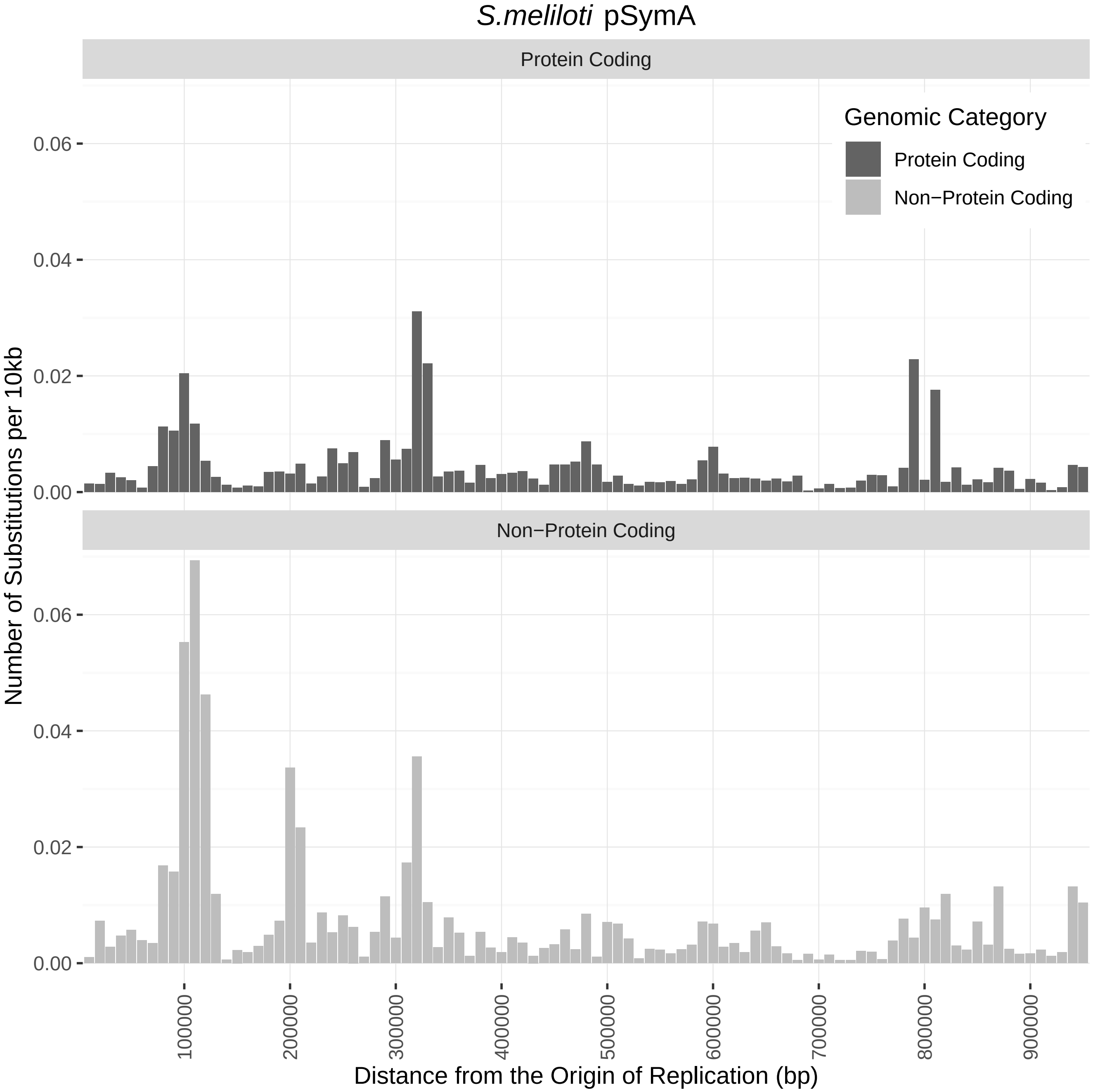


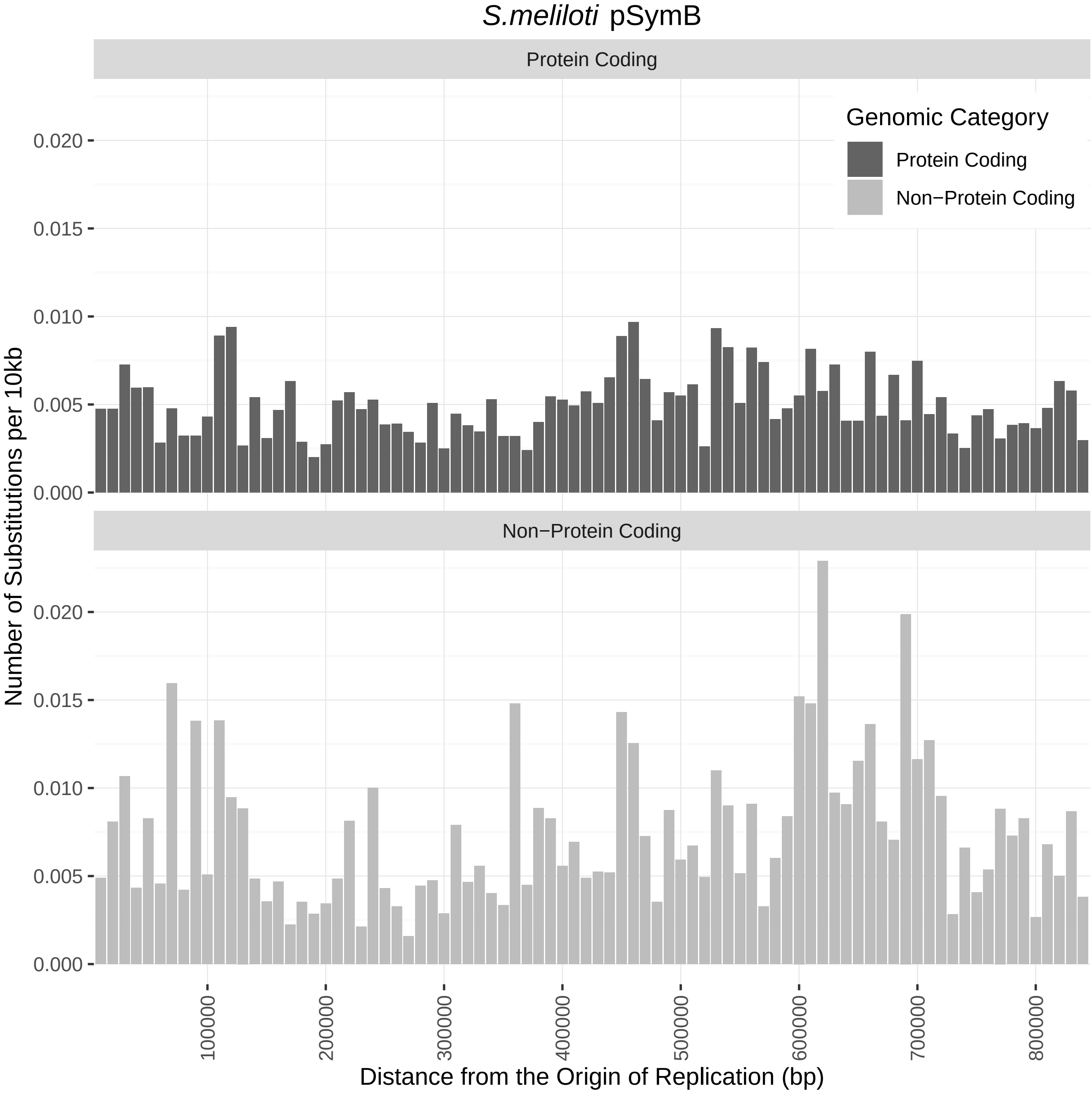
Figure 2: Distribution of the median expression value for each *B. subtilis* dataset minus the mean expression value for that gene across all datasets. Each gene is shown on the x-axis and the log base 10 values are on the y-axis.


B.subtilis Chromosome
Protein Coding
Non−Protein Coding
Genomic Category
Protein Coding
Non−Protein Coding
Number of Substitutions per 10kb
0.00
0.03
0.06
0.09
0.12
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
0.00
0.05
0.10
0.15
0.20
0.25
Distance from the Origin of Replication (bp)
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


S.meliloti Chromosome
Protein Coding
Non−Protein Coding
Genomic Category
Protein Coding
Non−Protein Coding
Number of Substitutions per 10kb
0.0000
0.0025
0.0050
0.0075
0.0100
0.0125
Distance from the Origin of Replication (bp)
100000
200000
300000
400000
500000
600000
700000
800000
900000
1000000
1100000
1200000
1300000
1400000
1500000
1600000
1700000
1800000
1900000
2000000


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
0.06
0.04
0.02
0.00
100000
200000
300000
400000
500000
600000
700000
800000
900000
Distance from the Origin of Replication (bp)



	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	-3.018×10^{-5}*	NS	2.87×10^{-2}	1.235×10^{-2}	-2.884×10^{-5}**	-5.276×10^{-5}*	1.085×10^{-2}	4.6×10^{-3}
B. subtilis Chromosome	-3.018×10^{-5}*	NS	2.87×10^{-2}	4.235×10^{-2}	NS	5.960×10^{-5}**	4×10^{-4}	6.25×10^{-3}
Streptomyces Chromosome	-1.988×10^{-5}***	-5.986×10^{-5}***	1.265×10^{-1}	2.58×10^{-2}	3.154×10^{-5}***	NS	2.23×10^{-2}	2.4×10^{-3}
S. meliloti Chromosome	5.109×10^{-6}**	NS	4.1×10^{-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}	1.425×10^{-4}***	-1.867×10^{-4}*	2.8×10^{-3}	5.5×10^{-4}
S. meliloti pSymB	NS	-4.411×10^{-5}***	3.37×10^{-2}	2.845×10^{-2}	NS	-4.669×10^{-5}**	7×10^{-3}	6.05×10^{-3}

Table 1: 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	$-1.622\times10^{-9***}$	$-1.852\times10^{-4***}$	NS	$-2.238\times10^{-5***}$
B. subtilis Chromosome	NS	$-1.761\times10^{-4**}$	NS	$-2.47\times10^{-5**}$
Streptomyces Chromosome	NS	$-4.382\times10^{-4***}$	$1.87\times10^{-9*}$	$-6.08\times10^{-5***}$
S. meliloti Chromosome	$-1.836\times10^{-10*}$	$-1.467\times10^{-5**}$	NS	NS
S. meliloti pSymA	NS	NS	$-1.266\times10^{-8**}$	$-6.74\times10^{-5**}$
S. meliloti pSymB	NS	NS	NS	NS

Table 2: 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	$-2.742\times10^{-5**}$
B. subtilis Chromosome	$-2.198\times10^{-5*}$
Streptomyces Chromosome	$-5.230\times10^{-7***}$
S. meliloti Chromosome	NS
S. meliloti pSymA	NS
S. meliloti pSymB	NS

Table 3: 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 4: 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	$-2.682\times10^{-6***}$
Streptomyces Chromosome	$-2.360\times10^{-6***}$
S. meliloti Chromosome	$-2.074\times10^{-6***}$
S. meliloti pSymA	NS
S. meliloti pSymB	$-4.19\times10^{-6*}$

Table 5: 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	Protein Coding Sequences	Non-Protein Coding Sequences
E. coli Chromosome	-1.354×10^{-7}***	NS
B. subtilis Chromosome	-6.735×10^{-8}***	NS
Streptomyces Chromosome	4.105×10^{-7}***	1.635×10^{-7}***
S. meliloti Chromosome	-9.185×10^{-8}***	-1.749×10^{-7}***
S. meliloti pSymA	-8.121×10^{-7}***	-1.247×10^{-6}***
S. meliloti pSymB	1.655×10^{-7}***	4.105×10^{-7}***

Table 6: 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	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 7: 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 8: 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 9: 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 10: Summary of strains and species found for each gene expression analysis. Gene Expression Omnibus accession numbers and date accessed are provided.