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)
- 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
- 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)

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

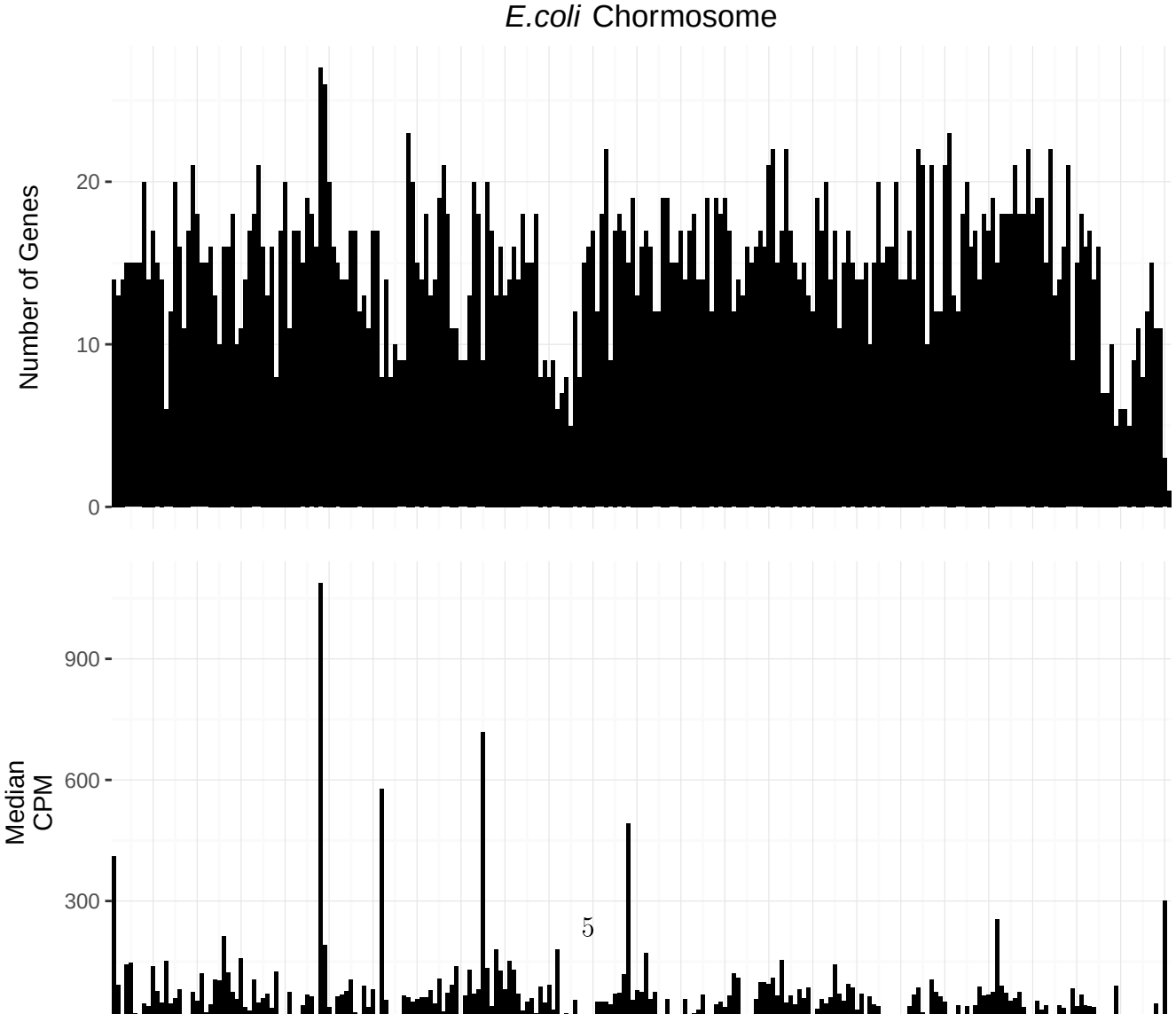
Assuming the re-running of my analysis is complete, I would like to work intensely on the inversions and gene expression stuff and get a good chunk of that started.

	20kb Near	
Bacteria and Replicon	Origin	Terminus
E. coli Chromosome		
B. subtilis Chromosome		
Streptomyces Chromosome		
S. meliloti Chromosome		
S. meliloti pSymA		
S. meliloti pSymB		

Table 1: Logistic regression on 20kb closest and farthest from the origin of replication after accounting for bidirectional replication and outliers.

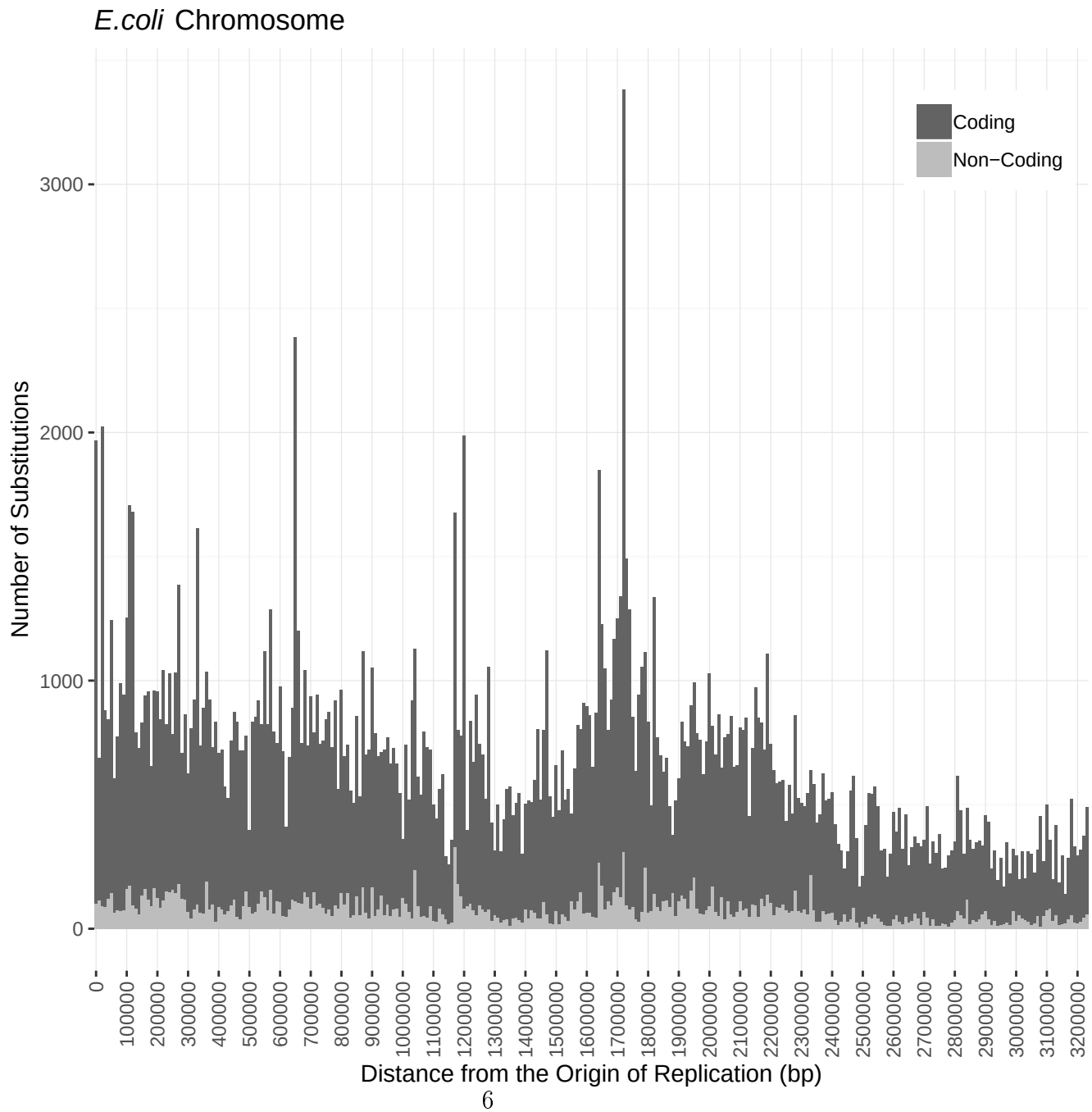
Bacteria and Replicon	Gene Expression 10kb
E. coli Chromosome	-2.742×10^{-5}**
B. subtilis Chromosome	-2.198×10^{-5}*
Streptomyces Chromosome	-5.230×10^{-7}***
S. meliloti Chromosome	NS
S. meliloti pSymA	NS
S. meliloti pSymB	NS

Table 2: 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'.



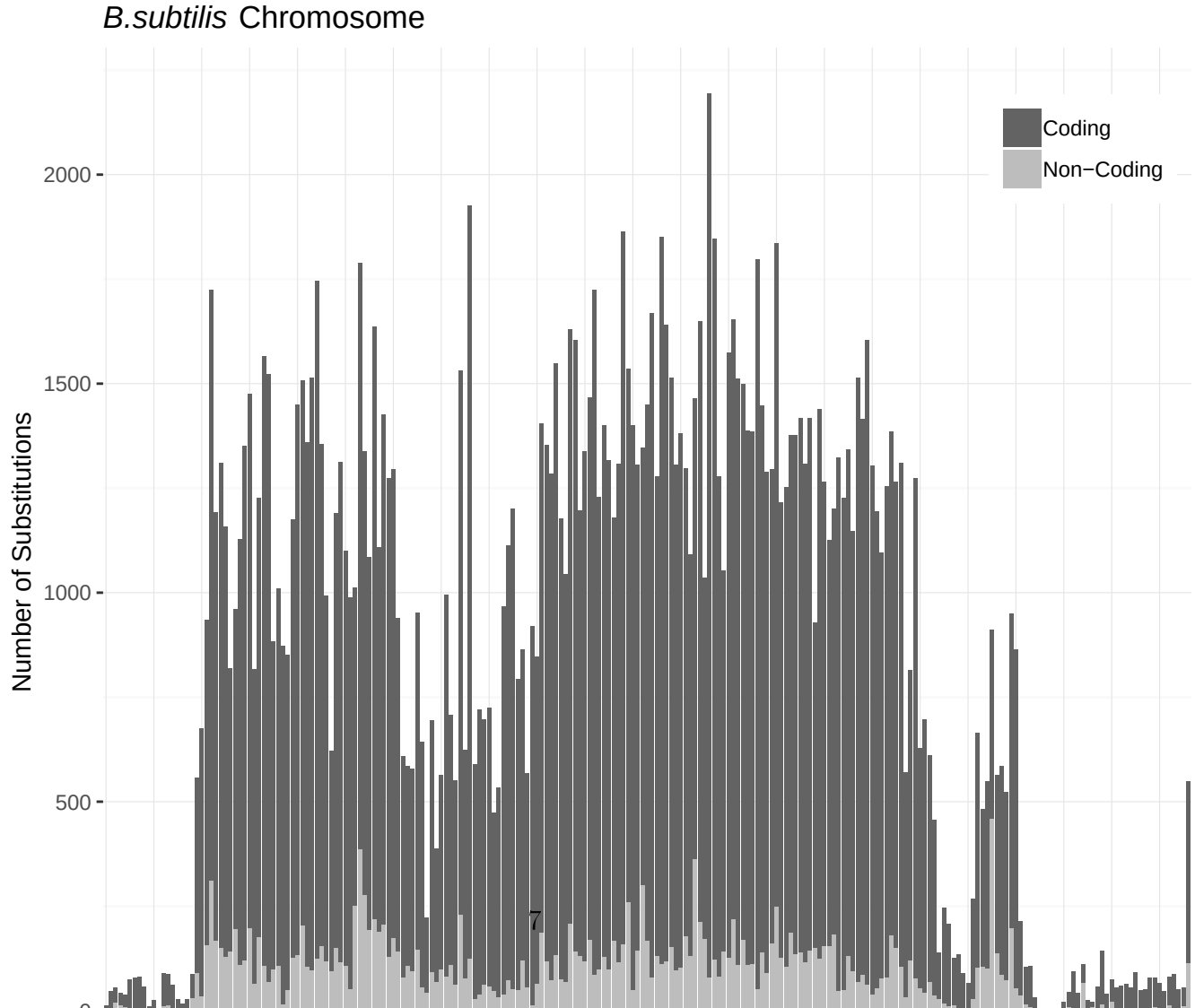
Bacteria and Replicon	Coefficient Estimate	Standard Error	P-value
E. coli Chromosome	-6.03×10^{-5}	1.28×10^{-5}	2.8×10^{-6}
B. subtilis Chromosome	-9.7×10^{-5}	2.0×10^{-5}	1.2×10^{-6}
Streptomyces Chromosome	-1.17×10^{-6}	1.04×10^{-7}	$<2\times10^{-16}$
S. meliloti Chromosome	3.97×10^{-5}	4.25×10^{-5}	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 3: 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.



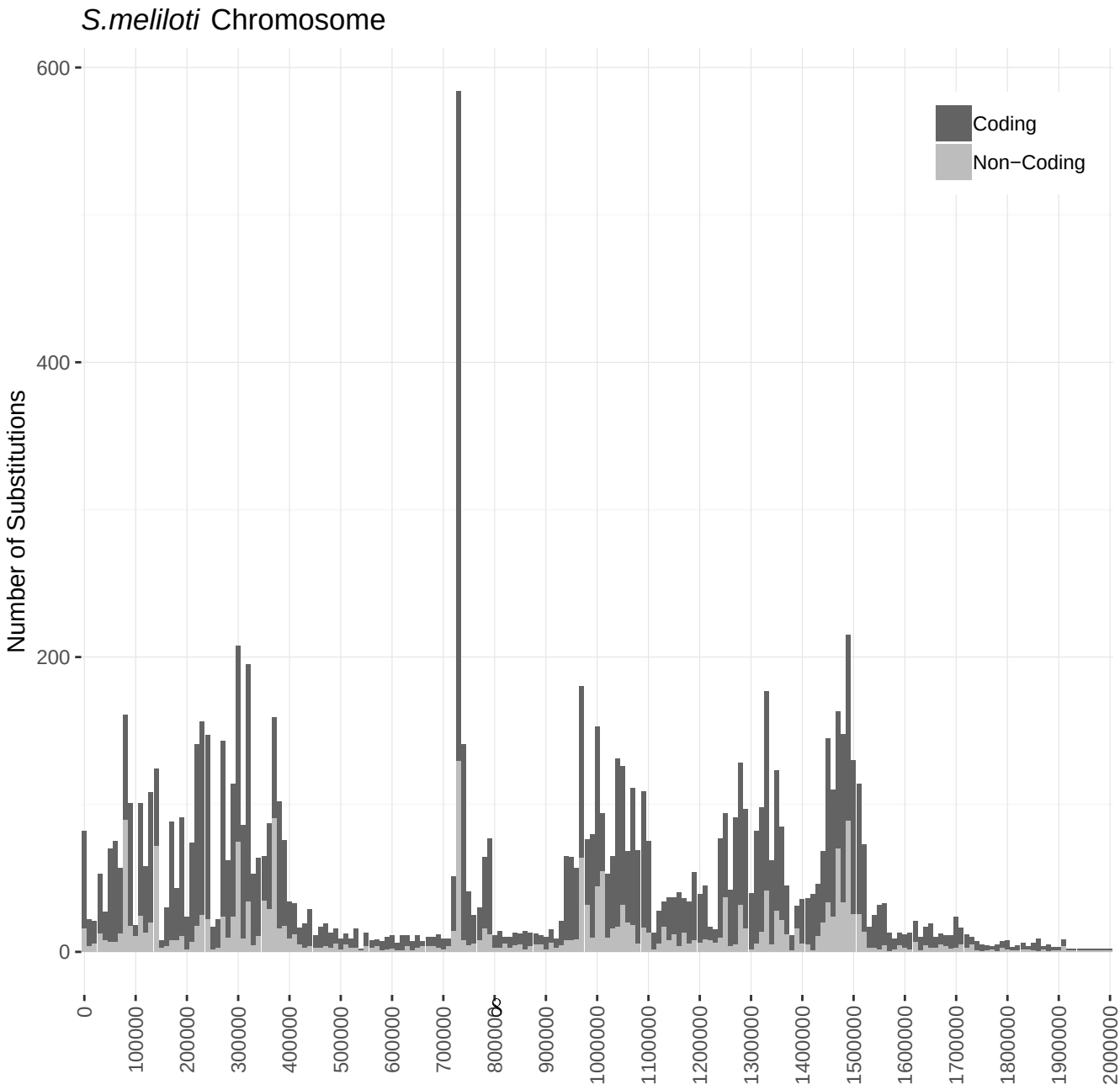
Bacteria and Replicon	Coefficient Estimate
E. coli Chromosome	NS
B. subtilis Chromosome	-2.682×10^{-6}***
Streptomyces Chromosome	-2.360×10^{-6}***
S. meliloti Chromosome	-2.074×10^{-6}***
S. meliloti pSymA	NS
S. meliloti pSymB	-4.19×10^{-6}*

Table 4: 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'.



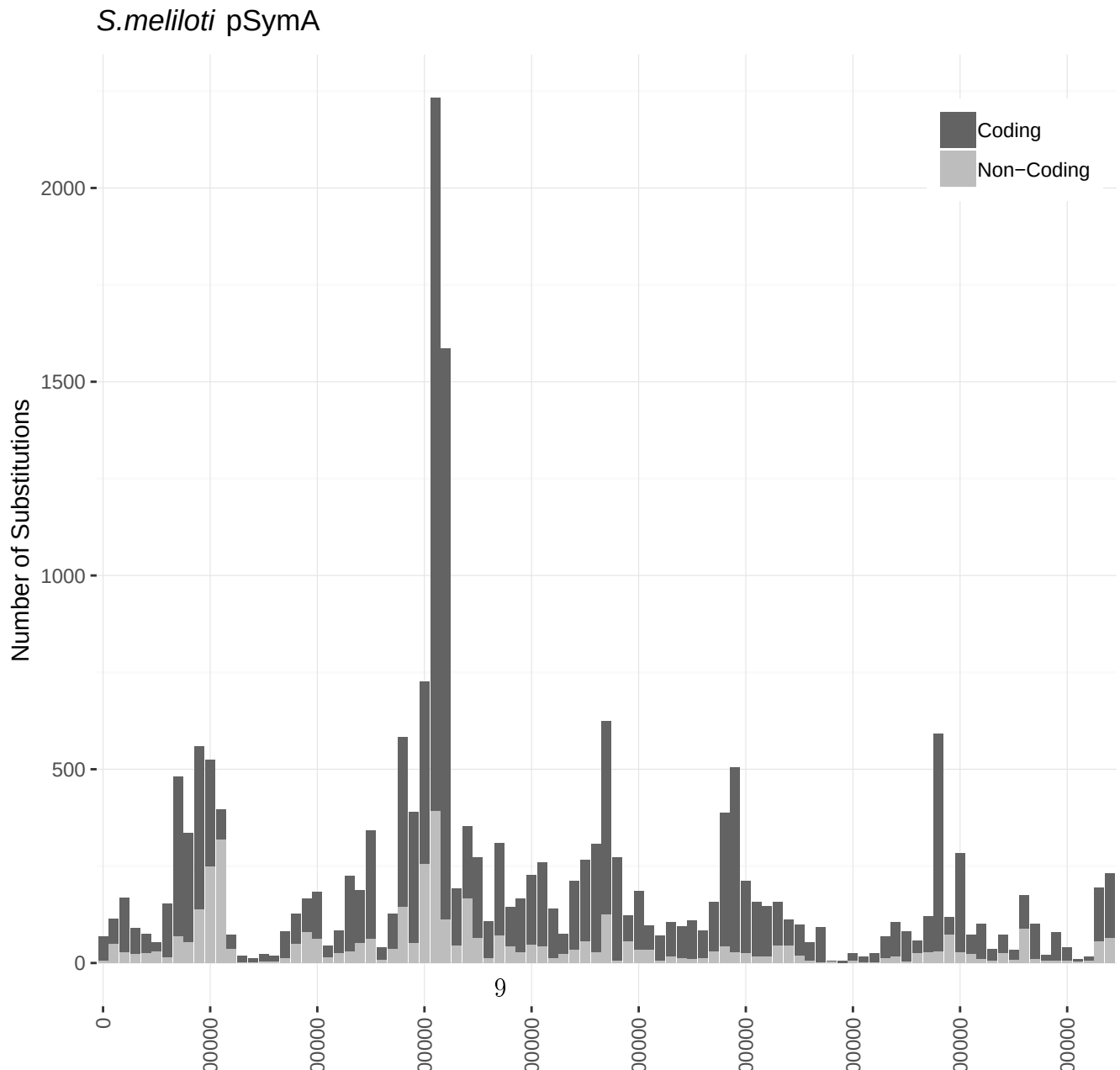
Bacteria and Replicon	Protein Coding Sequences	Non-Protein Coding Sequences
E. coli Chromosome	-1.354×10^{-7}***	NS
B. subtilis Chromosome	-6.735×10^{-8}***	NS
Streptomyces Chromosome	4.105×10^{-7}***	1.635×10^{-7}***
S. meliloti Chromosome	-9.185×10^{-8}***	-1.749×10^{-7}***
S. meliloti pSymA	-8.121×10^{-7}***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

Table 5: 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	$-9.08\times10^{-6}*$
Streptomyces Chromosome	NS	NS	NS
S. meliloti Chromoeom	NS	NS	NS
S. meliloti pSymA	NS	NS	NS
S. meliloti pSymB	NS	NS	$1.163\times10^{-5}*$

Table 6: 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'.

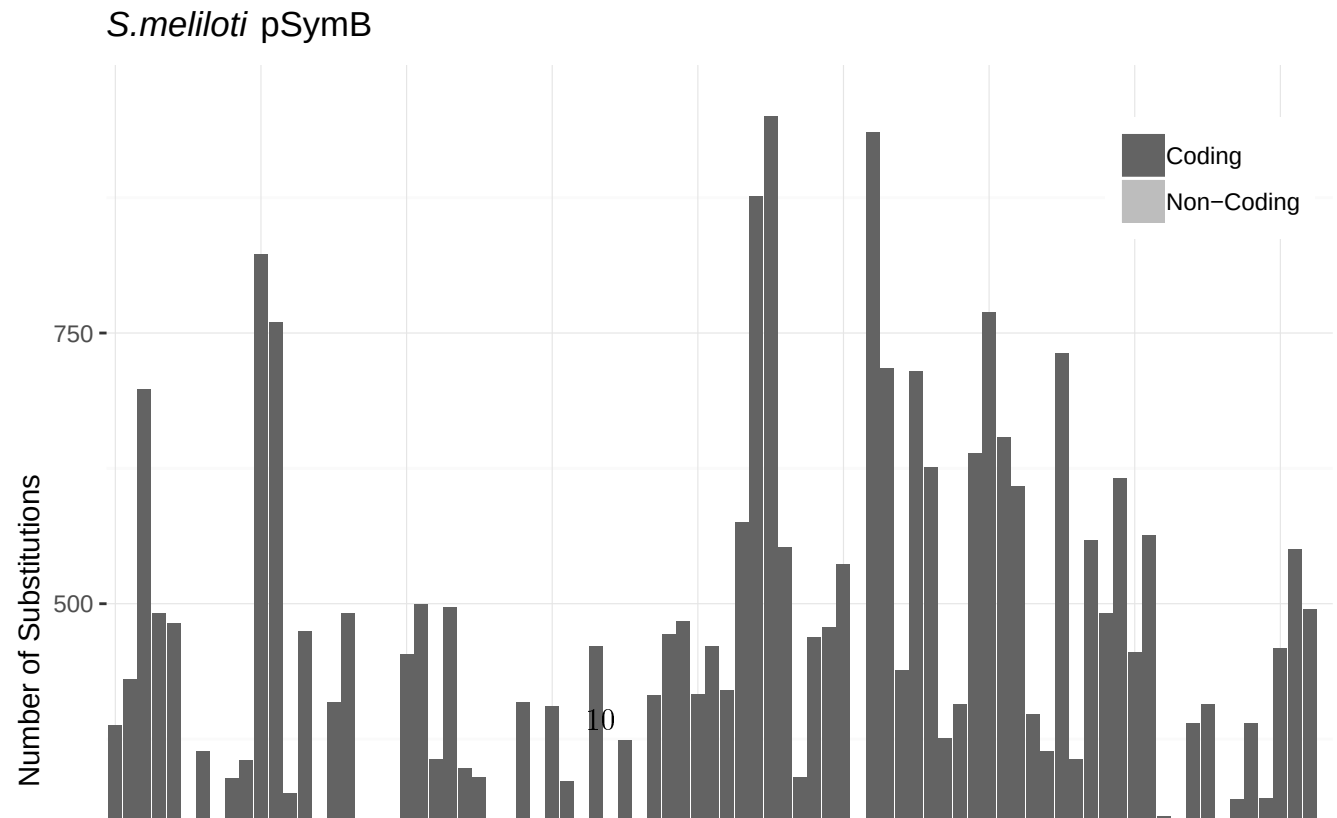


Bacteria and Replicon	Average Expression Value (CPM)
E. coli Chromosome	160.500
B. subtilis Chromosome	176.400
Streptomyces Chromosome	6.084
S. meliloti Chromosome	271.400
S. meliloti pSymA	690.100
S. meliloti pSymB	595.700

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