Subs Paper Things to Do:

- ~~# of coding and non-coding sites~~
- ~~# of subs in each of †~~
- ~~Look into *Streptomyces* non-coding issue~~
- ~~Look into *E. coli* coding issue~~
- ~~Look into pSymB coding/non-coding trend weirdness~~
- ~~Figure out why *Streptomyces* appears to have tons of coding data missing~~
- ~~Figure out what is going on with cod/non-cod code and why it is still not working!~~
- ~~write up methods for coding/non-coding~~
- ~~write methods and results for clustering~~
- ~~start code to split alignment into multiple alignments of each gene~~
- ~~figure out how to deal with overlapping genes~~
- ~~figure out how to deal with gaps in gene of ref taxa~~
- ~~split up the alignment into multiple alignments of each gene~~
- ~~check if each gene alignment is a multiple of 3 (proper codon alignment)~~
- ~~get dN/dS for coding/non-coding stuff per gene~~
- Or get 1st, 2nd, 3rd codon pos log regs
- ~~write up coding/non-coding results~~
- ~~take out gene expression from this paper~~
- ~~write better intro/methods for distribution of subs graphs~~
- ~~write discussion for coding/non-coding~~
- ~~write coding/non-coding into conclusion~~
- ~~figured out pipeline for CODEML to calculate dN/dS for each gene~~
- ~~make a list of what should be in supplementary files for subs paper~~
- ~~put everything in list into supplementary file for subs paper~~
- write dN/dS methods
- write dN/dS results
- write dN/dS discussion

- write dN/dS into conclusion
- ~~new bar graph with coding and non-coding sites separated~~
- mol clock for my analysis?
- GC content? COG? where do these fit?

Gene Expression Paper Things to Do:

- ~~look for more GEO expression data for *S. meliloti*~~
- ~~look for more GEO expression data for *Streptomyces*~~
- ~~look for more GEO expression data for *B. subtilis*~~
- ~~format paper and put in stuff that is already written~~
- ~~look for more GEO expression data for *E. coli*~~
- ~~Get numbers for how many different strains and multiples of each strain I have for gene expression~~
- ~~re-do gene expression analysis for *B. subtilis*~~
- ~~re-do gene expression analysis for *E. coli*~~
- ~~find papers about what has been done with gene expression~~
- ~~read papers †~~
- ~~put notes from † papers into word doc~~
- 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:

- ~~get as much GEO data as possible~~
- ~~find papers about inversions and expression~~
- ~~see how many inversions I can identify in these strains of *Escherichia coli* with gene expression data~~
- ~~read papers about inversions~~
- ~~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

✓Finished re-running the coding and non-coding substitution analysis (Table 3)

✓Get per gene and per genome dN, dS, and ω results (Table 1)

✓started thinking more about inversions and gene expression

✓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)

I re-ran the coding and non-coding substitution analysis because I realized that my genome position numbering was off for the blocks. The results are still the same and can be found in Table 3. The updated distribution of substitutions histograms are also below.

dN, dS, ω Calculated the per gene and per genome dN, dS, and ω results for each bacteria. For the per gen rates I calculated both a weighted average (weighted by the length of the gene) and the non-weighted average to see if there was a big difference. In doing this, I noticed that my calculations for the per genome and weighted per gene averages were the same... So I need to look into this more to see what is happening.

I was also thinking about maybe associating each of the dN, dS, and ω averages per gene with the midpoint of that gene and getting a distribution of these rates across the genome. But I am not sure if this is useful or conventional. Thoughts?

Inversions and Gene Expression:

I have been thinking about the inversions and gene expression project a lot this week and the only bacteria that has enough gene expression data is *E. coli*. So all analysis will have to be done on only that. I have 3 strains of *E. coli* gene expression data: k-12, ATCC, and Saki. Within K-12, I have 3 substrains: MG1655, DH108 and BW25113. Which makes for a total of 5 genomes. I was thinking about maybe doing the same ancestral reconstruction process with these 5 strains but I am not sure of a few things:

- Is 5 genomes going to provide enough data/information?
- would I use a midpoint for each gene and have one data point per node in the tree per gene?
- or should I have all positions in the gene have the same expression value?

Re-doing this analysis should not be too bad but would require me to re-code a few things to make the new data work.

I was also wondering if an inversion and reverse complement of a sequence is the same thing? That is, does an inverted piece of sequence have to be the reverse complement to be re-inserted into the genome? Another thing that I realized is that progressiveMauve and PARSNP both allow for the blocks to be present in different parts of the genome for the different taxa. So this would be blending the rearrangements information and inversions information. Which then I am not sure if we can say that the gene expression differences are due to inversions alone unless we correct for this some how. But overall, PARSNP and progressiveMauve are identifying the same large inversions, the only difference is that PARSNP can identify very small inversions better. I think this is because progressiveMauve is trying to make the biggest blocks possible and is aligning the whole genome, where as PARSNP is just doing the core genome. Another difference is that PARSNP has very very few gaps in its alignment, I think this a result from aligning just the core? Not sure how this will impact future results.

This Week

Continue to check in with the dN/dS stuff and make sure my per gene and genome rates are correct and try to interpret these!

I would like to obtain all inversions from Mauve or PARSNP alignment for the inversions and gene expression analysis.

I want to work on finishing up some scholarships for travel to SMBE.

Next Week

Create histograms with the total number of genes in each 10kb section of the genome to supplement the gene expression analysis.

Continue working on the inversions and gene expression analysis.

Write up interpretation of dN/dS results.

	Gene Average			Genome Average		
Bacteria and Replicon	dS	dN	ω	dS	dN	ω
E. coli Chromosome	0.2924	0.0144	0.0604	0.2600	0.0133	0.0556
B. subtilis Chromosome	0.6526	0.0358	0.0891	0.5267	0.0321	0.0828
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	0.0798	0.0109	0.2320	0.0800	0.0103	0.2218
S. meliloti pSymB	0.0814	0.0086	0.1639	0.0782	0.0082	0.1590

Table 1: 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.

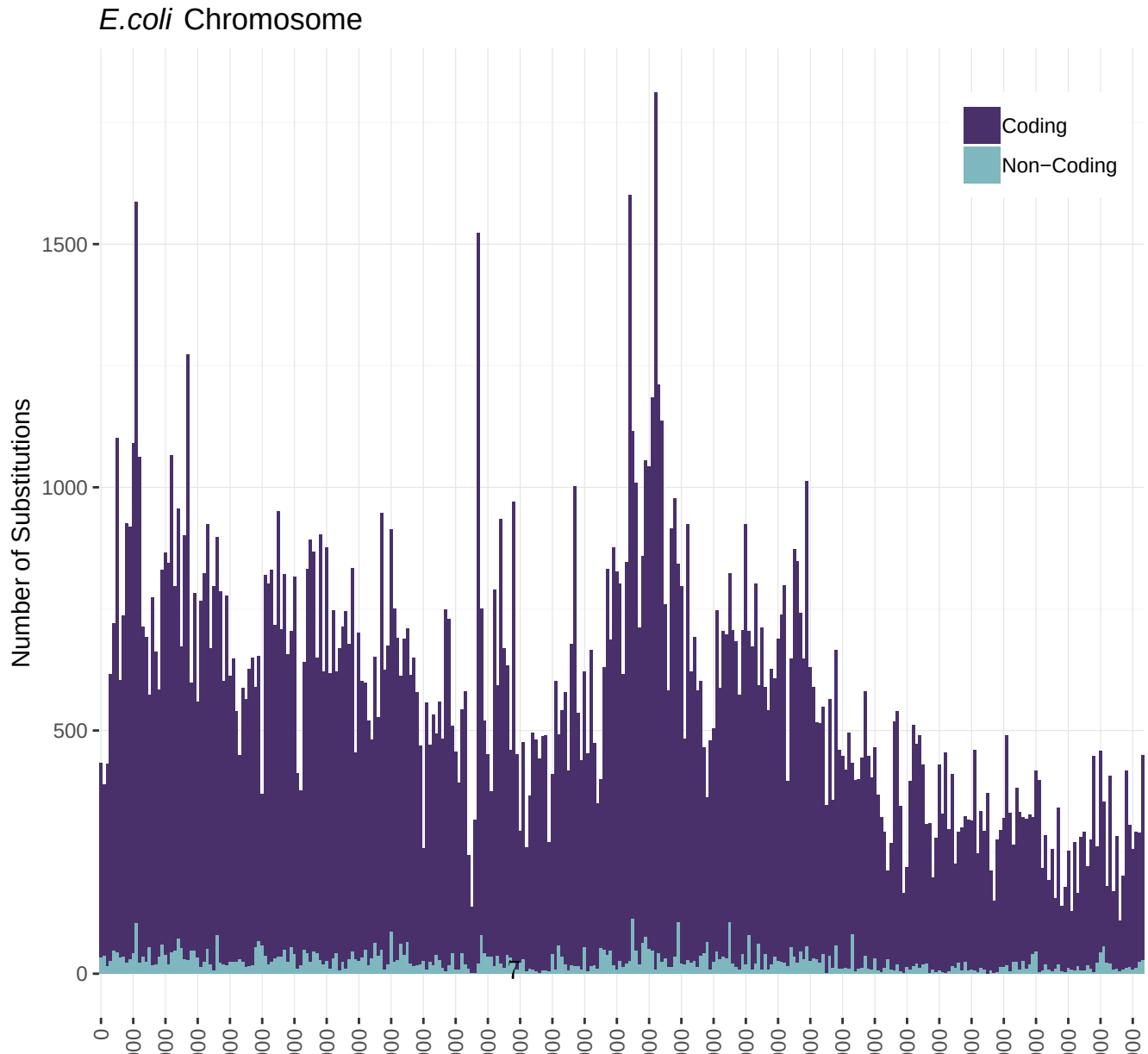
Bacteria and Replicon	Average Replicon Length	# of Coding Sites	# of Non-Coding Sites	# of Subs Coding	# of Subs Non-Coding
E. coli Chromosome	5082529	2960007	191748	207199	9534
B. subtilis Chromosome	4077077	2074653	102906	205150	6187
Streptomyces Chromosome	8497577	2422980	21581	551530	3670
S. meliloti Chromosome	3426881	1931139	199425	6684	842
S. meliloti pSymA	1455940	419223	34213	9832	943
S. meliloti pSymB	1664597	552816	22098	11699	645

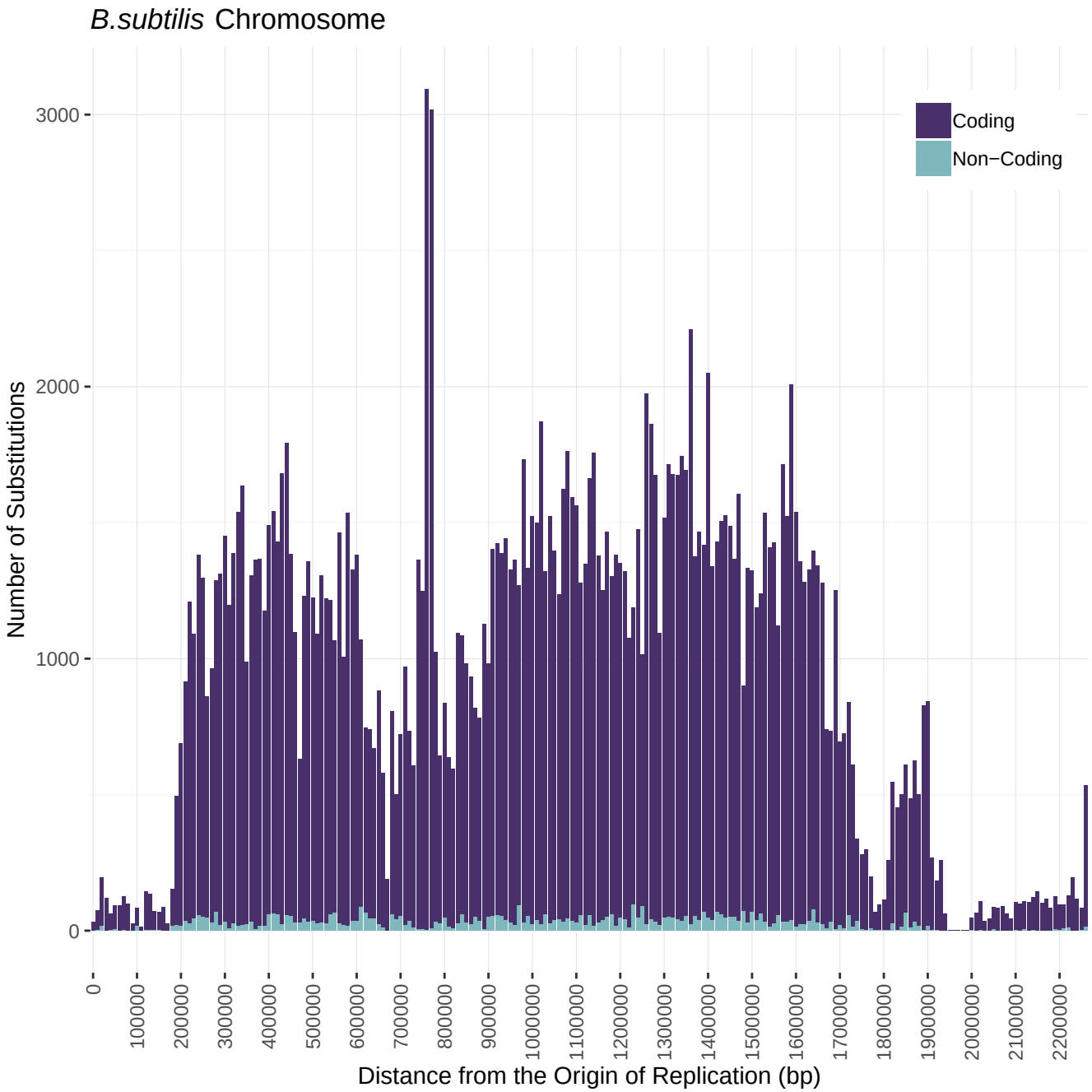
Table 2: Total proportion of coding and non-coding sites in each replicon and the percentage of those sites that have a substitution (multiple substitutions at one site are counted as two substitutions).

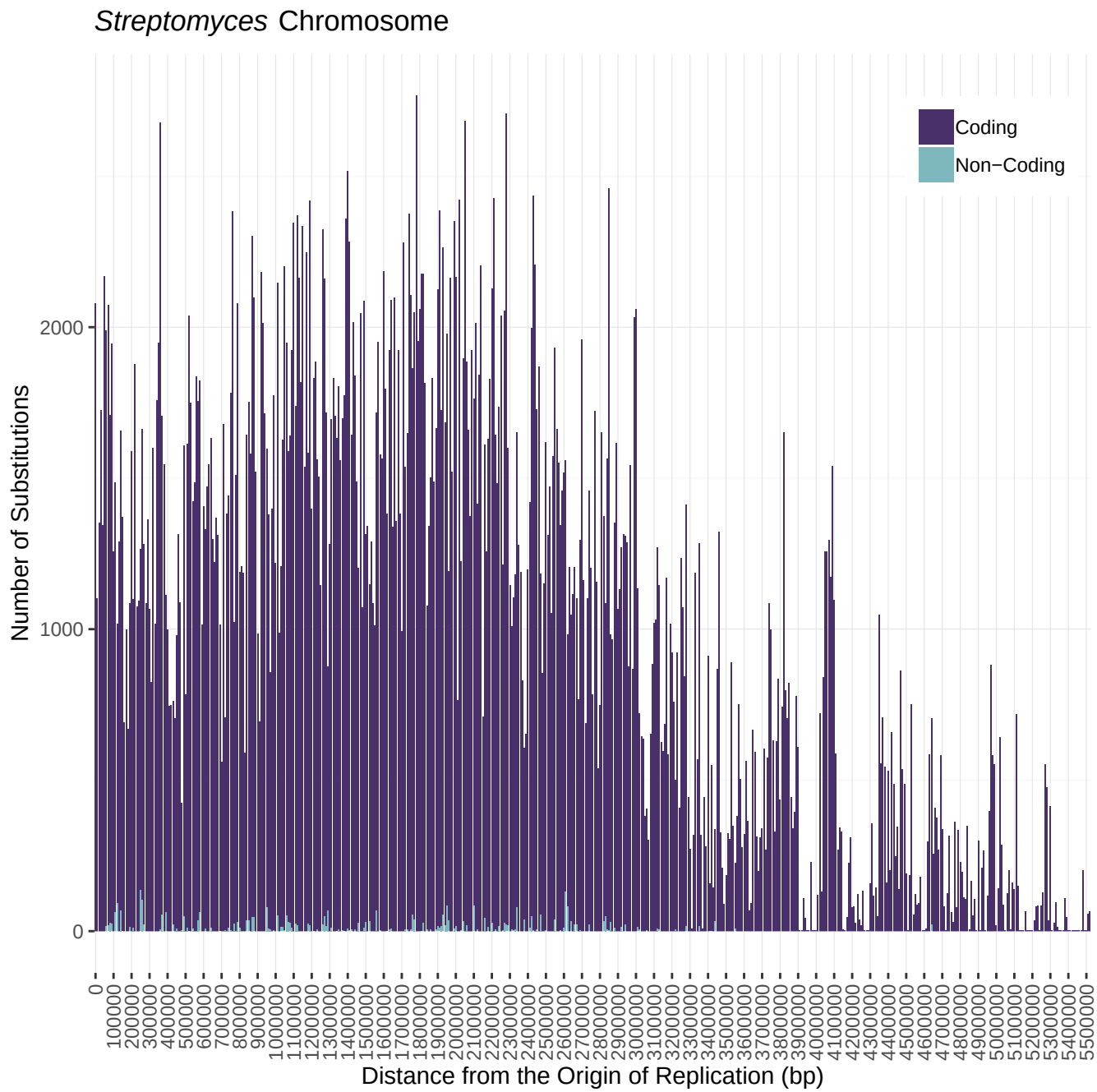
Sub density graphs with coding and non-coding information

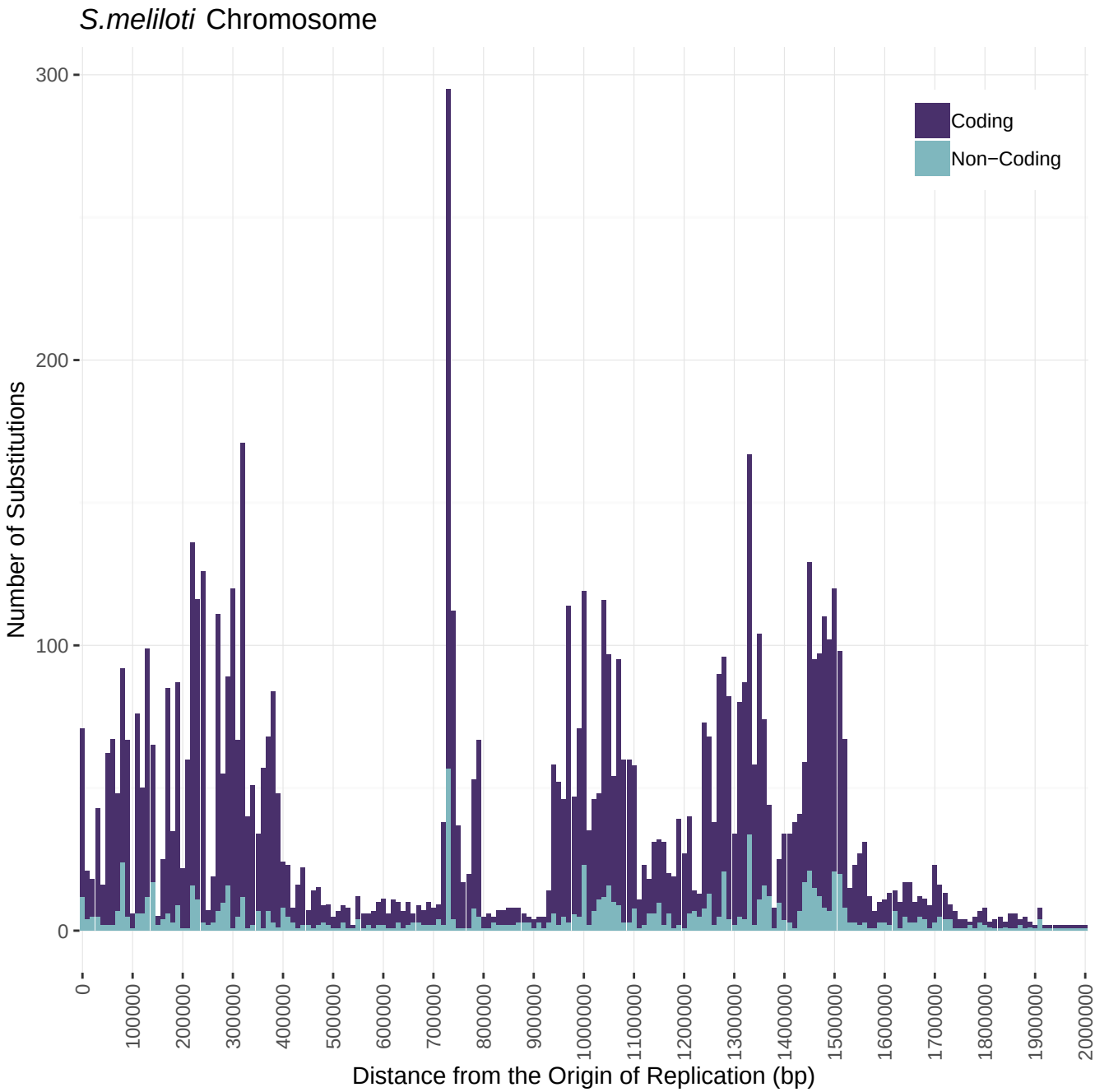
Bacteria and Replicon	Coding Sequences	Non-Coding Sequences
E. coli Chromosome	-9.983×10^{-8}***	6.994×10^{-8}***
B. subtilis Chromosome	-1.071×10^{-7}***	-9.861×10^{8}***
Streptomyces Chromosome	-2.626×10^{-8}***	3.615×10^{-7}***
S. meliloti Chromosome	-1.367×10^{-7}***	-1.510×10^{-7}*
S. meliloti pSymA	-1.075×10^{-7}*	NS
S. meliloti pSymB	2.878×10^{-7}***	8.595×10^{-7}***

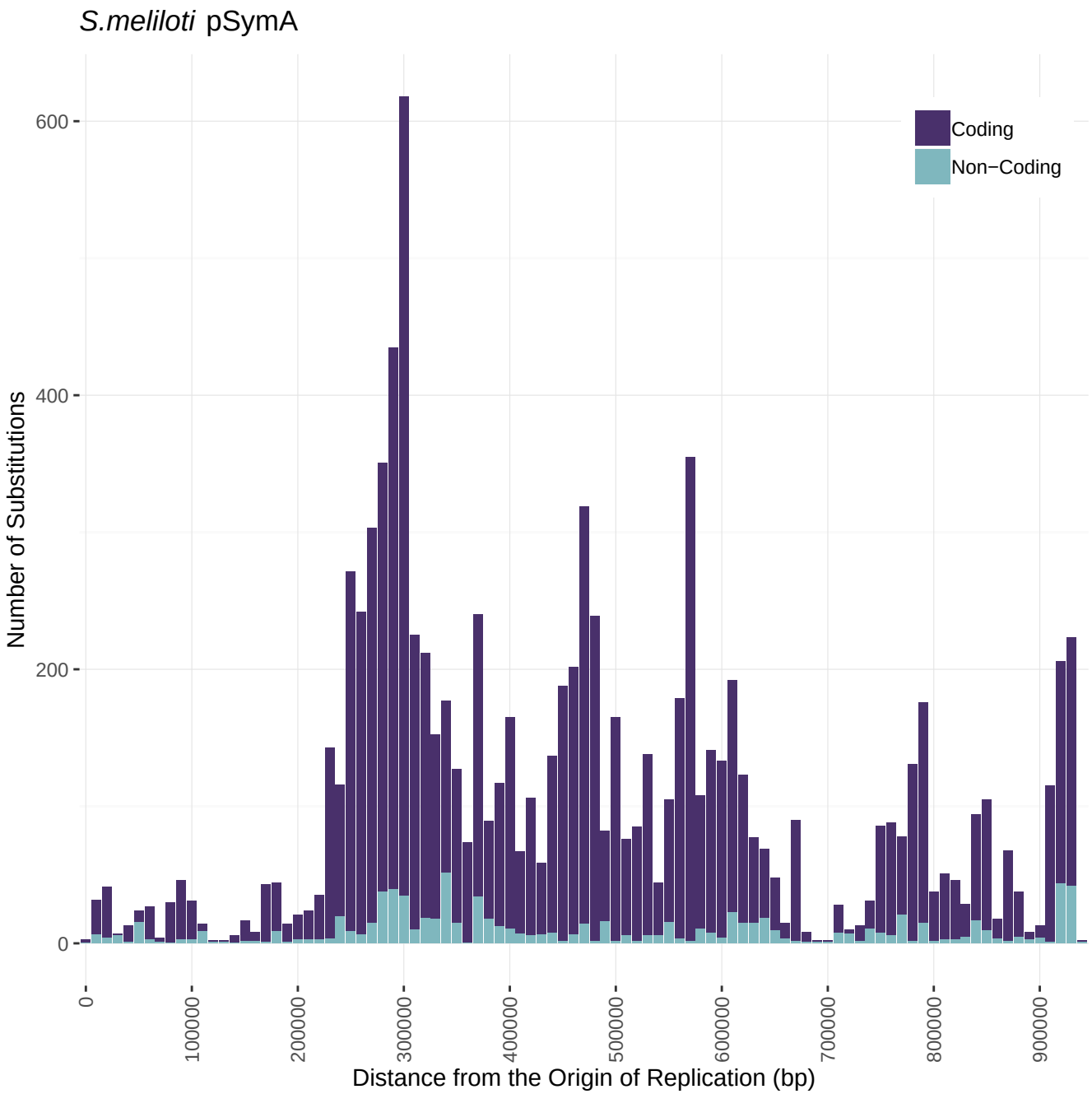
Table 3: Logistic regression analysis of the number of substitutions along all positions of the genome of the respective bacteria replicons. These genomic positions were split up into the coding and non-coding regions of the genome. Grey coloured boxes indicate a negative logistic regression coefficient 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'.

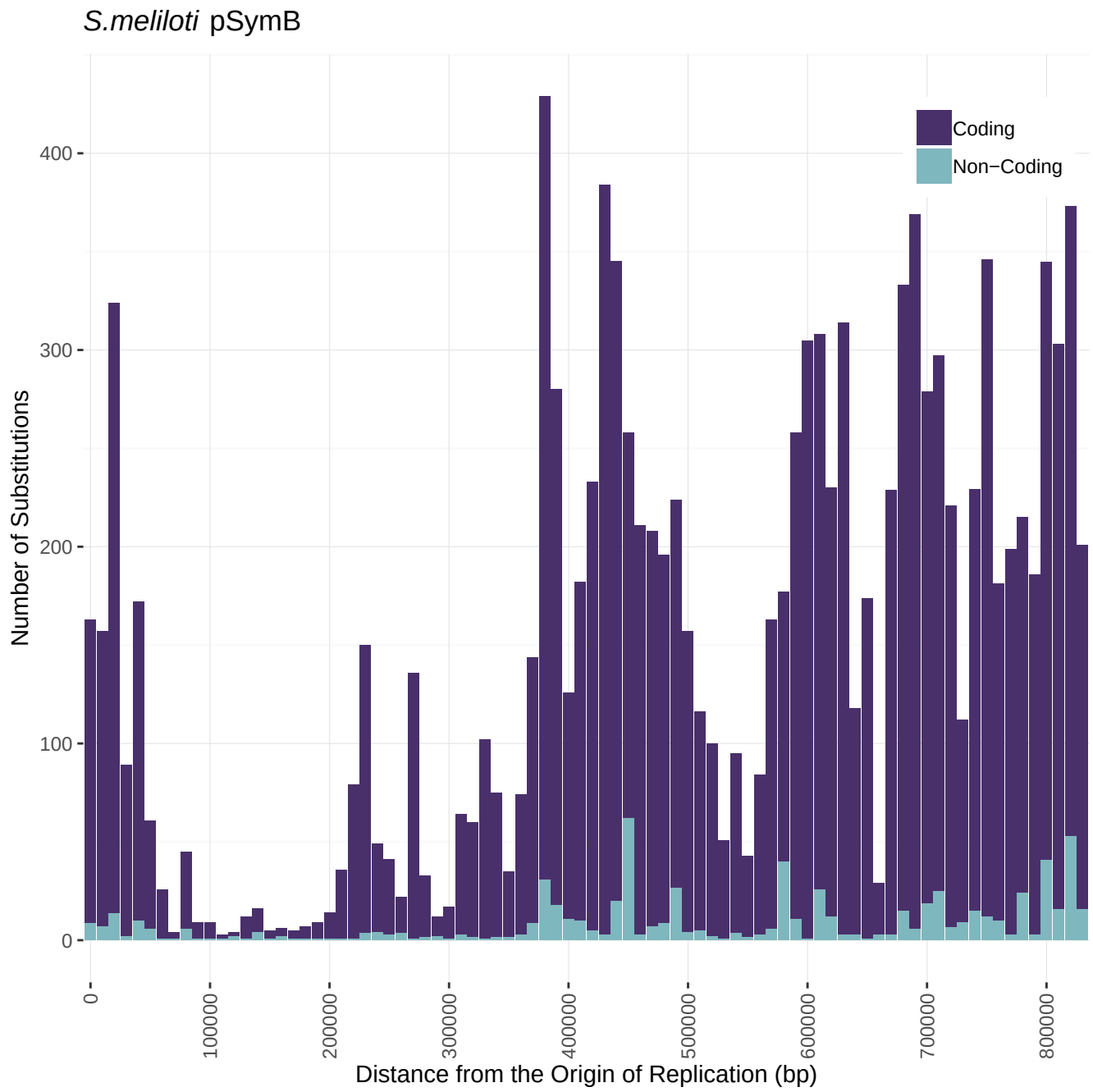



B.subtilis Chromosome
Coding
Non−Coding
Number of Substitutions
3000
2000
1000
0
0
100000
200000
300000
400000
500000
600000
700000
800000
900000
1000000
1100000
1200000
1300000
1400000
1500000
1600000
1700000
1800000
1900000
2000000
2100000
2200000
Distance from the Origin of Replication (bp)

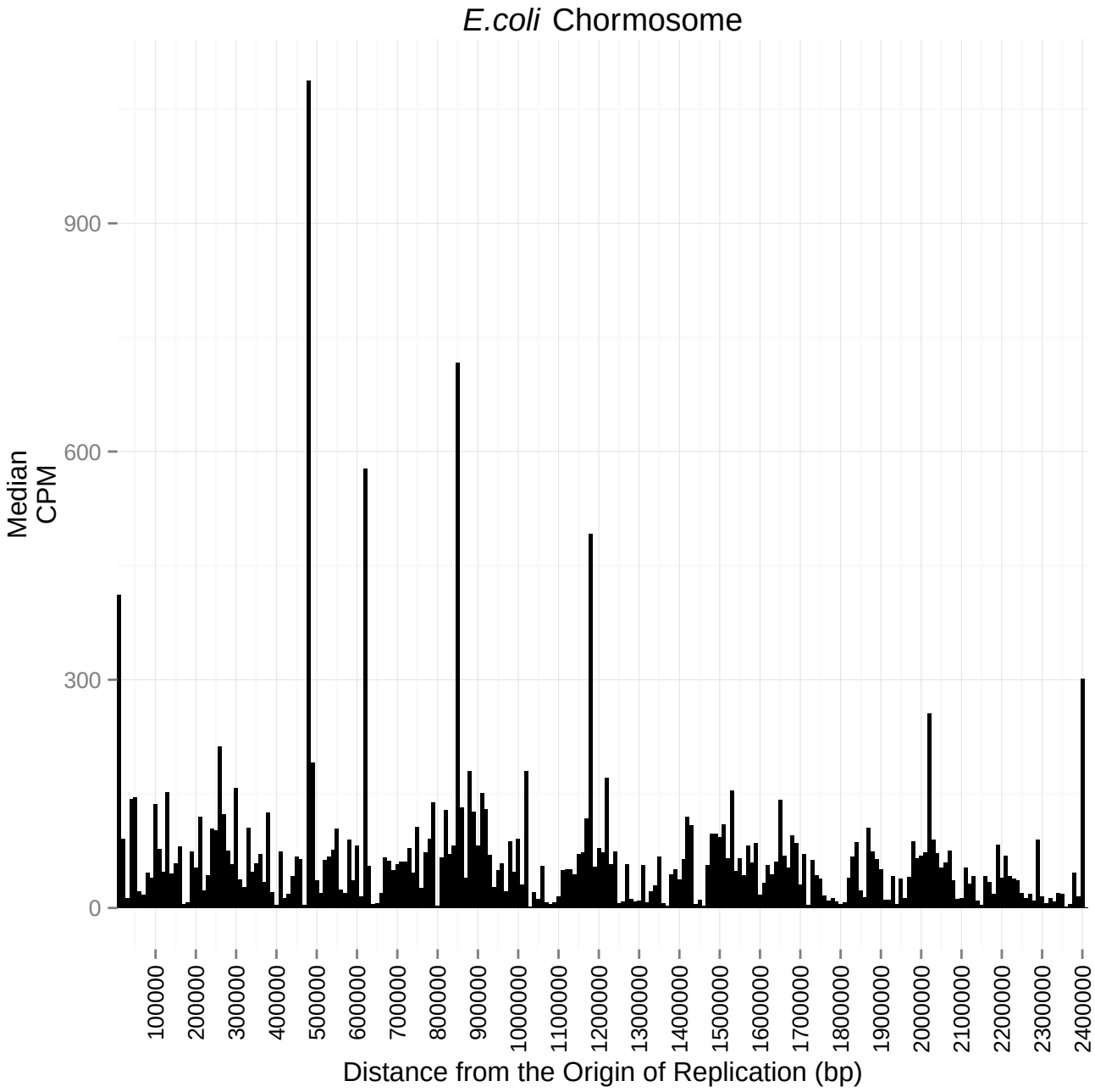

Streptomyces Chromosome
Coding
Non−Coding
Number of Substitutions
0
1000
2000
0
100000
200000
300000
400000
500000
600000
700000
800000
900000
1000000
1100000
1200000
1300000
1400000
1500000
1600000
1700000
1800000
1900000
2000000
2100000
2200000
2300000
2400000
2500000
2600000
2700000
2800000
2900000
3000000
3100000
3200000
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)

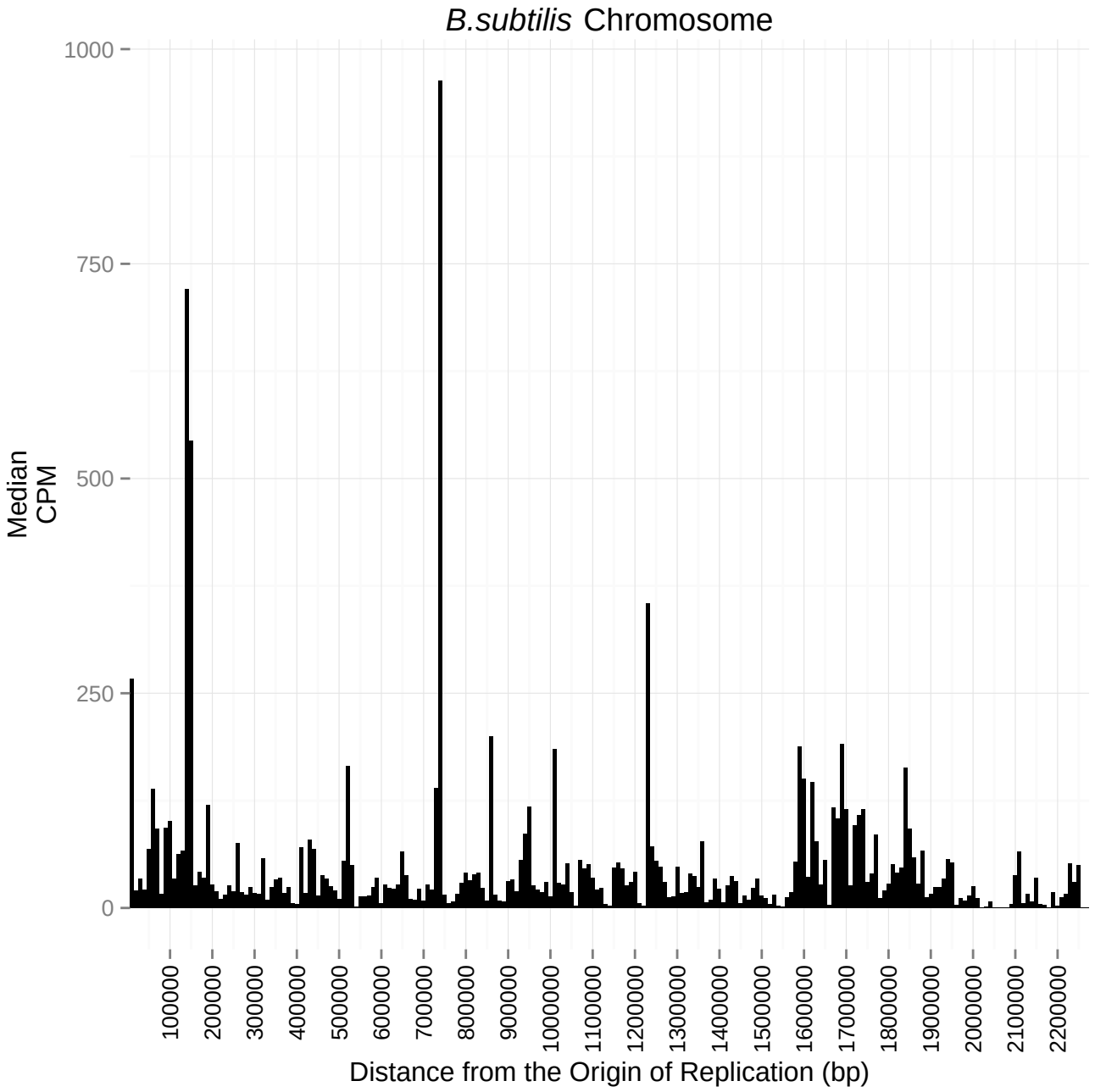

S.meliloti Chromosome
Coding
Non−Coding
Number of Substitutions
0
100
200
300
0
100000
200000
300000
400000
500000
600000
700000
800000
900000
1000000
1100000
1200000
1300000
1400000
1500000
1600000
1700000
1800000
1900000
2000000
Distance from the Origin of Replication (bp)

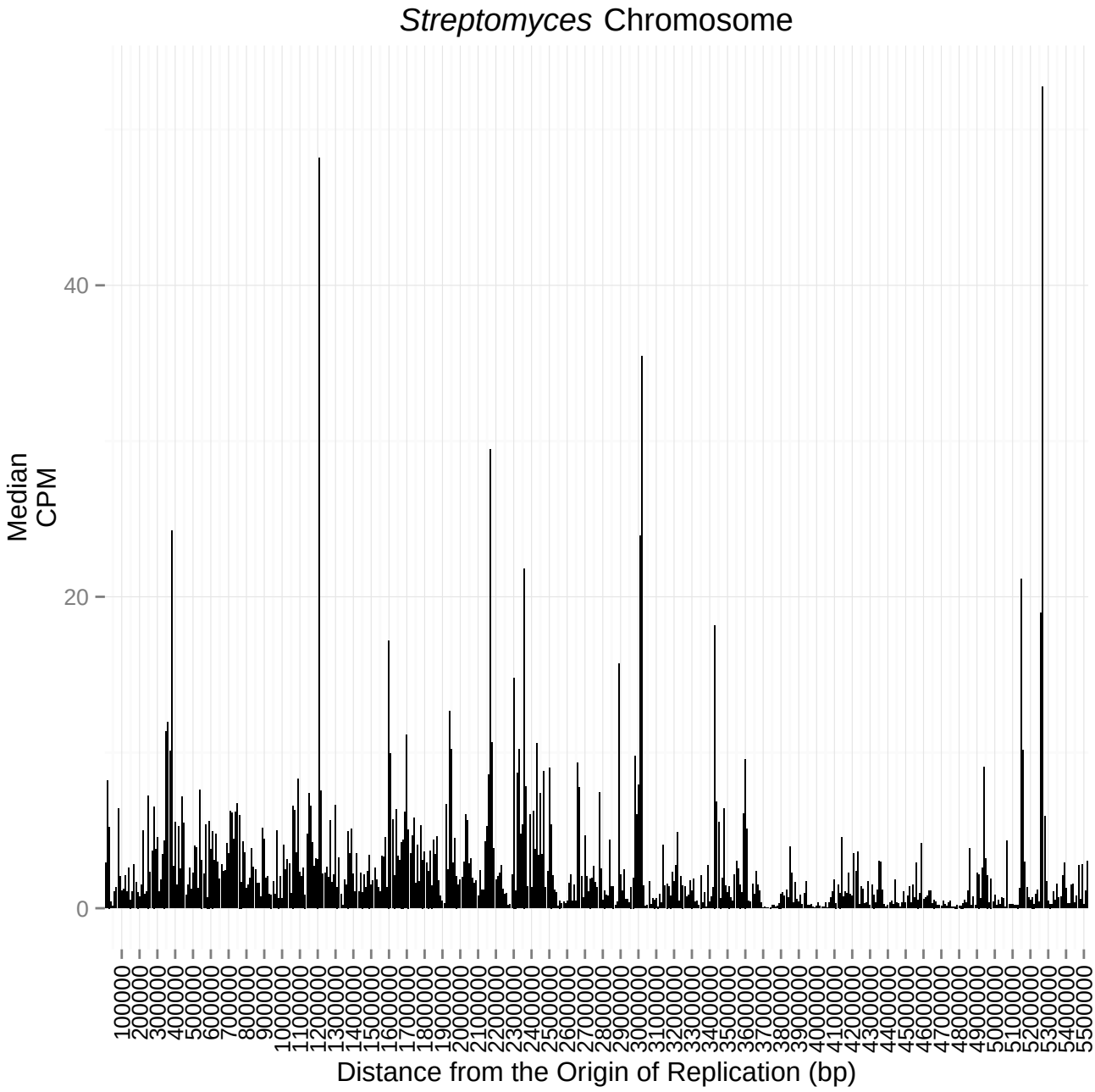

S.meliloti pSymA
Coding
Non−Coding
Number of Substitutions
0
200
400
600
0
100000
200000
300000
400000
500000
600000
700000
800000
900000
Distance from the Origin of Replication (bp)

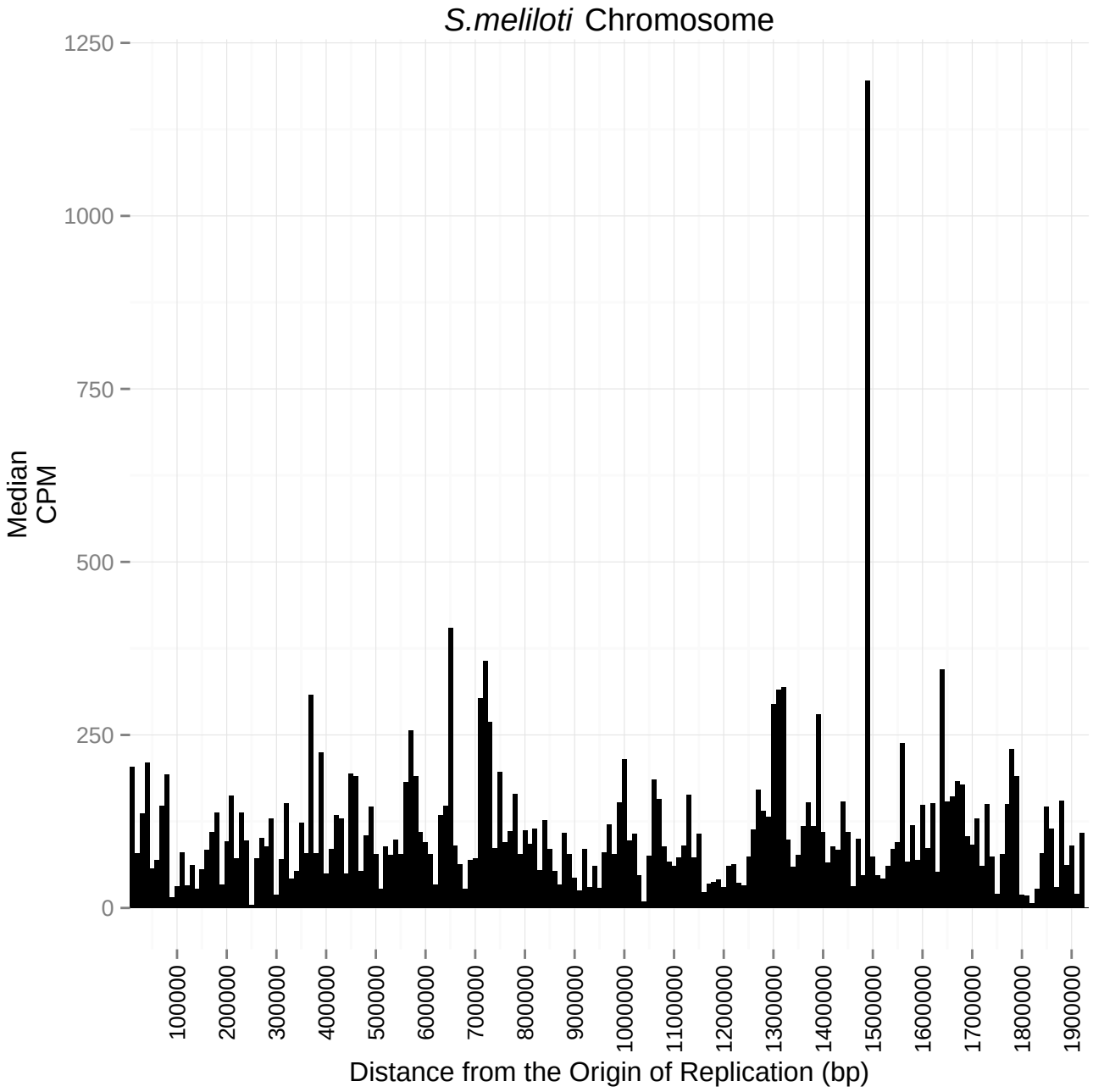


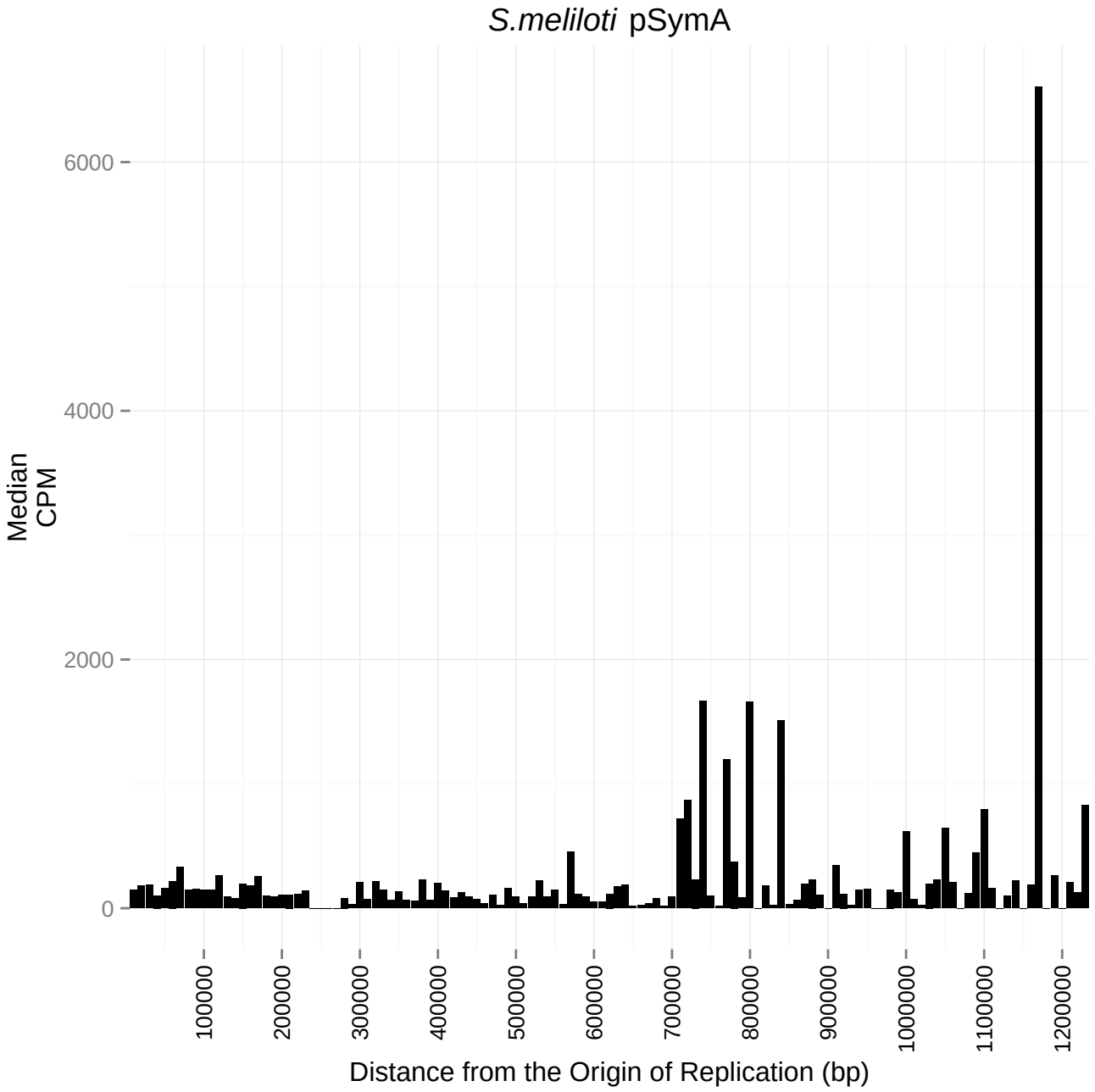
Gene expression graphs

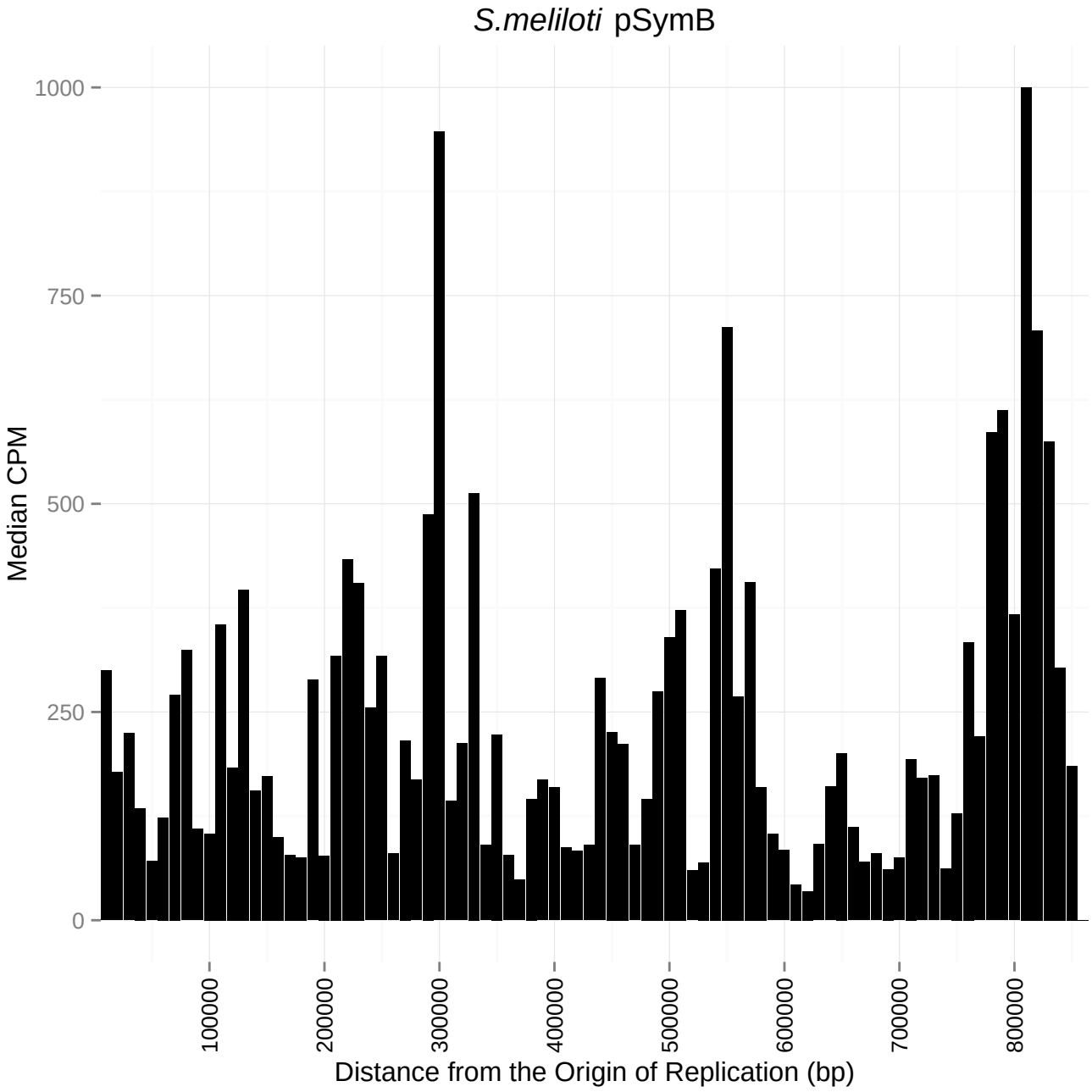

E.coli Chormosome
Median CPM
900
600
300
0
100000
200000
300000
400000
500000
600000
700000
800000
900000
1000000
1100000
1200000
1300000
1400000
1500000
1600000
1700000
1800000
1900000
2000000
2100000
2200000
2300000
2400000
Distance from the Origin of Replication (bp)

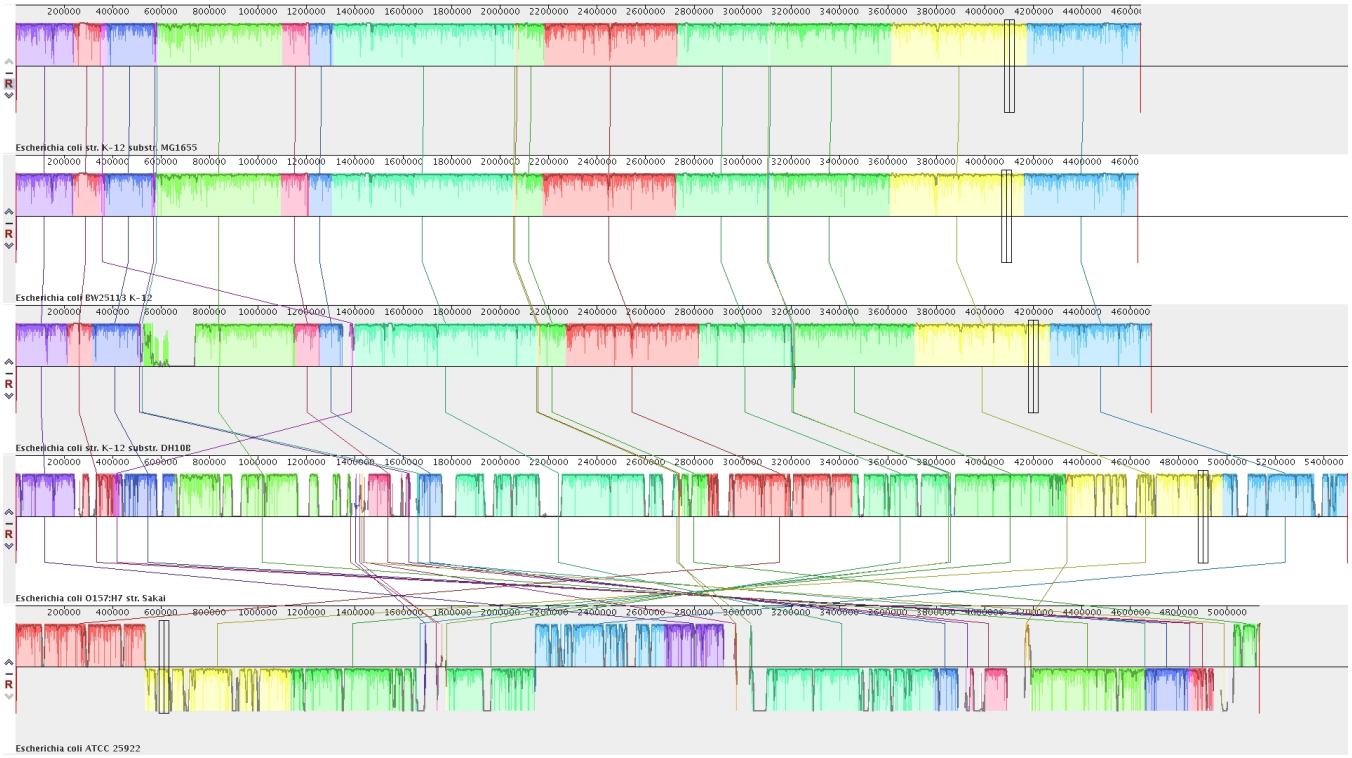

B.subtilis Chromosome
1000
750
500
250
0
Median CPM
100000
200000
300000
400000
500000
600000
700000
800000
900000
1000000
1100000
1200000
1300000
1400000
1500000
1600000
1700000
1800000
1900000
2000000
2100000
2200000
Distance from the Origin of Replication (bp)

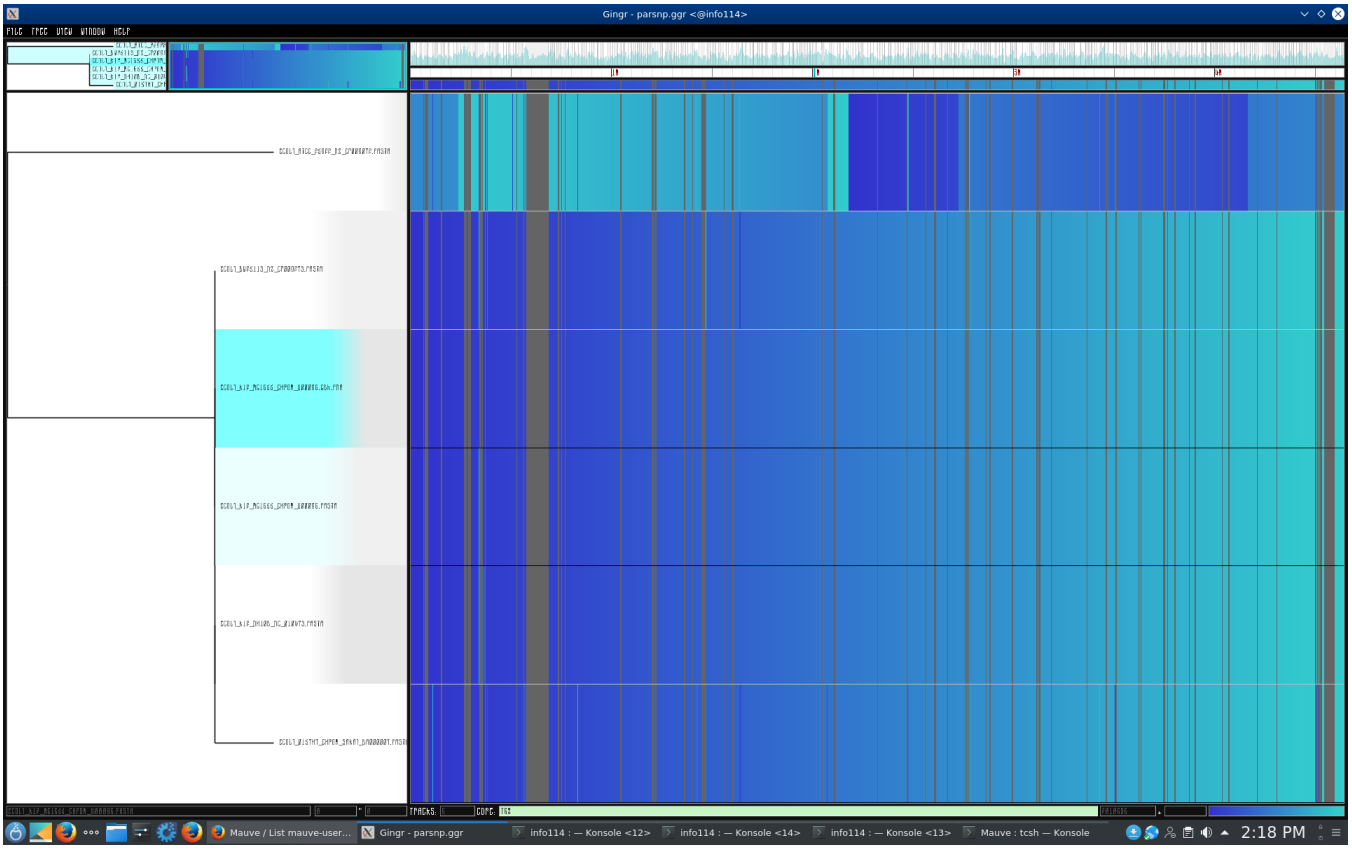

Streptomyces Chromosome
Median CPM
40
20
0
100000
200000
300000
400000
500000
600000
700000
800000
900000
1000000
1100000
1200000
1300000
1400000
1500000
1600000
1700000
1800000
1900000
2000000
2100000
2200000
2300000
2400000
2500000
2600000
2700000
2800000
2900000
3000000
3100000
3200000
3300000
3400000
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
Median CPM
1250
1000
750
500
250
0
100000
200000
300000
400000
500000
600000
700000
800000
900000
1000000
1100000
1200000
1300000
1400000
1500000
1600000
1700000
1800000
1900000
Distance from the Origin of Replication (bp)


S.meliloti pSymA
6000
4000
2000
0
Median CPM
100000
200000
300000
400000
500000
600000
700000
800000
900000
1000000
1100000
1200000
Distance from the Origin of Replication (bp)


S.meliloti pSymB
1000
750
500
250
0
Median CPM
100000
200000
300000
400000
500000
600000
700000
800000
Distance from the Origin of Replication (bp)


Escherichia coli str. K-12 substr. MG1655
Escherichia coli BW25113 K-12
Escherichia coli str. K-12 substr. DH10B
Escherichia coli O157:H7 str. Sakai
Escherichia coli ATCC 25922


Gingr - parsnp.ggr <@info114>
FILE TREE VIEW WINDOW HELP
Mauve / List mauve-user...
Gingr - parsnp.ggr
info114 : — Konsole <12>
info114 : — Konsole <14>
info114 : — Konsole <13>
Mauve : tcsh — Konsole
2:18 PM

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 4: 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 5: Linear regression analysis of the median counts per million expression data along the genome of the respective bacteria replicons. Grey coloured boxes indicate statistically significant results at the 0.5 significance level. 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.