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

✓write results and discussion for gene expression paper

✓write methods for gene expression paper

✓get coding and non-coding substitution graphs for each replicon

✓get proportions of coding and non-coding sections for each genome

✓results and outline of slides for OE3C

I obtained a break down of the coding and non-coding proportions of the genome for each replicon. These can be found in Table 1. The proportion of coding sites that have substitutions, and non-coding sites that have substitutions are about the same (5% - 10% of sites). The only concerning thing is that because we are now only looking at sites in the alignment where all taxa are the same classification (codon position 1, 2, 3, non-coding...etc), we are now only looking at less than half of the total sites in the genome. I am slightly concerned that some of the regions in the genome may be underrepresented. What are your thoughts on this?

I realized when working through the dN/dS pipeline that there was still "stop" codons present and so I then realized that I forgot to account for the reverse complement of the sequence. So I had to add in this extra step into the pipeline, and re-run the coding/non-coding substitution analysis.

I worked on writing the methods, results and discussion for the gene expression paper/letter. I also have started working on my presentation for the OE3C, I have most of the intro, and results.

This Week

I realized that I made a slight mistake in the genome positions that each block prints out, so I have fixed this and need to re-do the substitution analysis again. I began this process this past weekend but it takes some time to run so I am still working on this.

I would like to run the dN/dS pipeline on all of the replicons.

I would like to have a presentation complete for the OE3C so I can practice it in front of everyone next week.

Next Week

I want to begin figuring out how to obtain all inversions from the Mauve or PARSNP alignment for the inversions analysis.

I will mostly be at the OE3C conference (May 2-4).

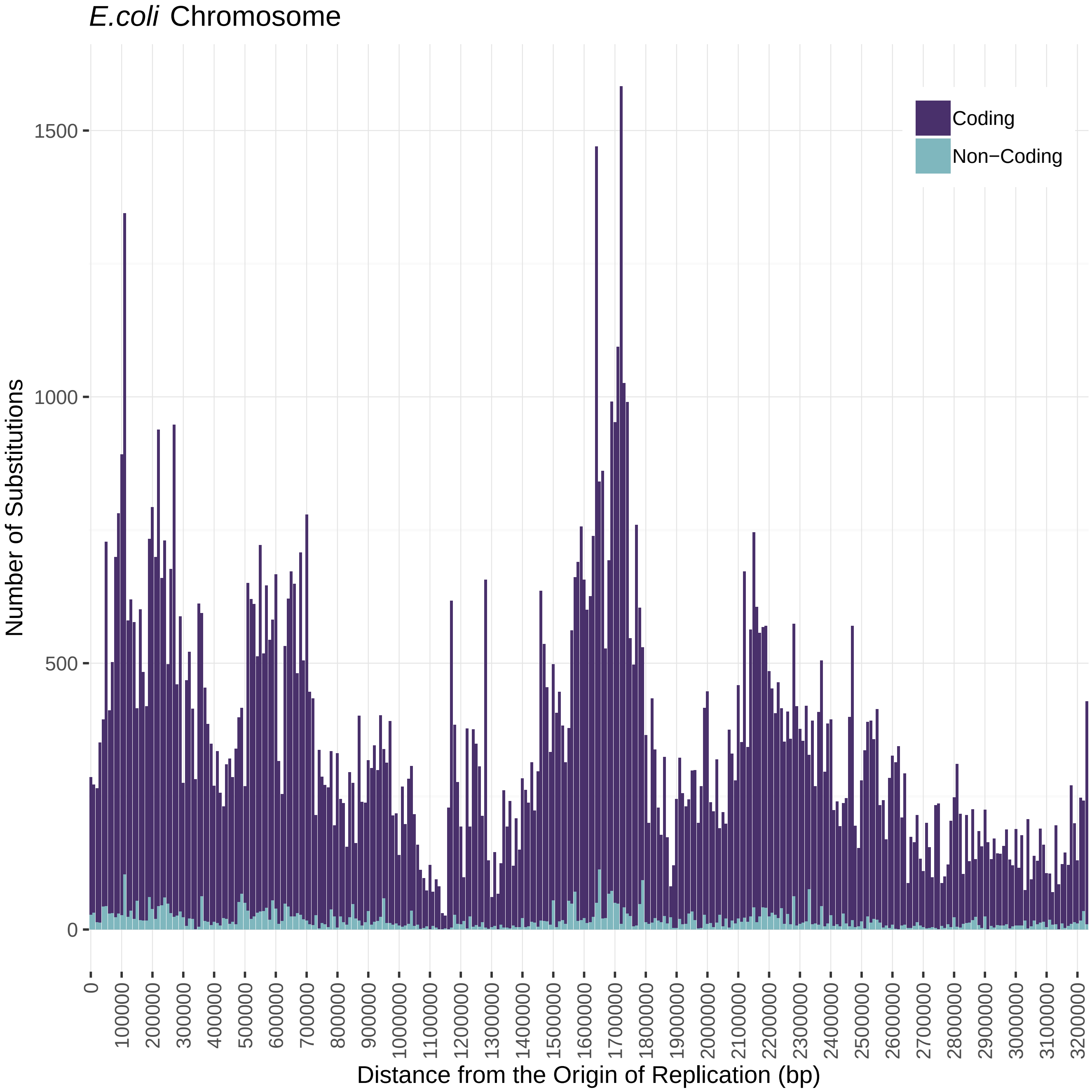
Bacteria and Replicon	# of Coding Sites	# of Non-Coding Sites	# of Subs Coding	# of Subs Non-Coding
E. coli Chromosome	2243082	141885	129595	6721
B. subtilis Chromosome	2073824	102906	171510	6187
Streptomyces Chromosome	1996354	18584	246447	2808
S. meliloti Chromosome	1928634	199425	5153	842
S. meliloti pSymA	419531	68426	7709	943
S. meliloti pSymB	559311	22098	10039	645

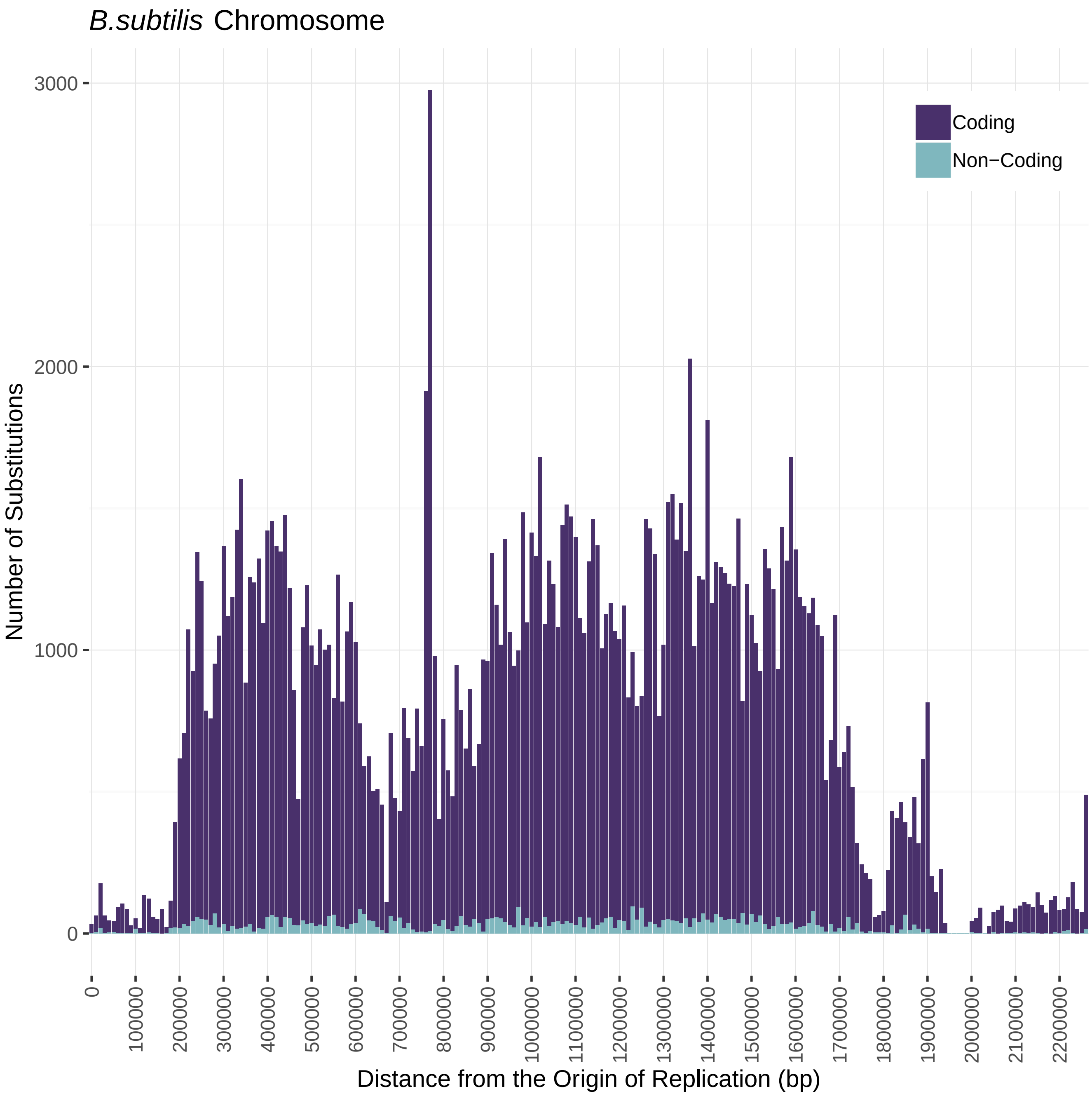
Table 1: Total proportion of coding and non-coding sites in each replicon and the percentage of those sites that have a substitution (multiple substitutions at one site are counted as two substitutions).

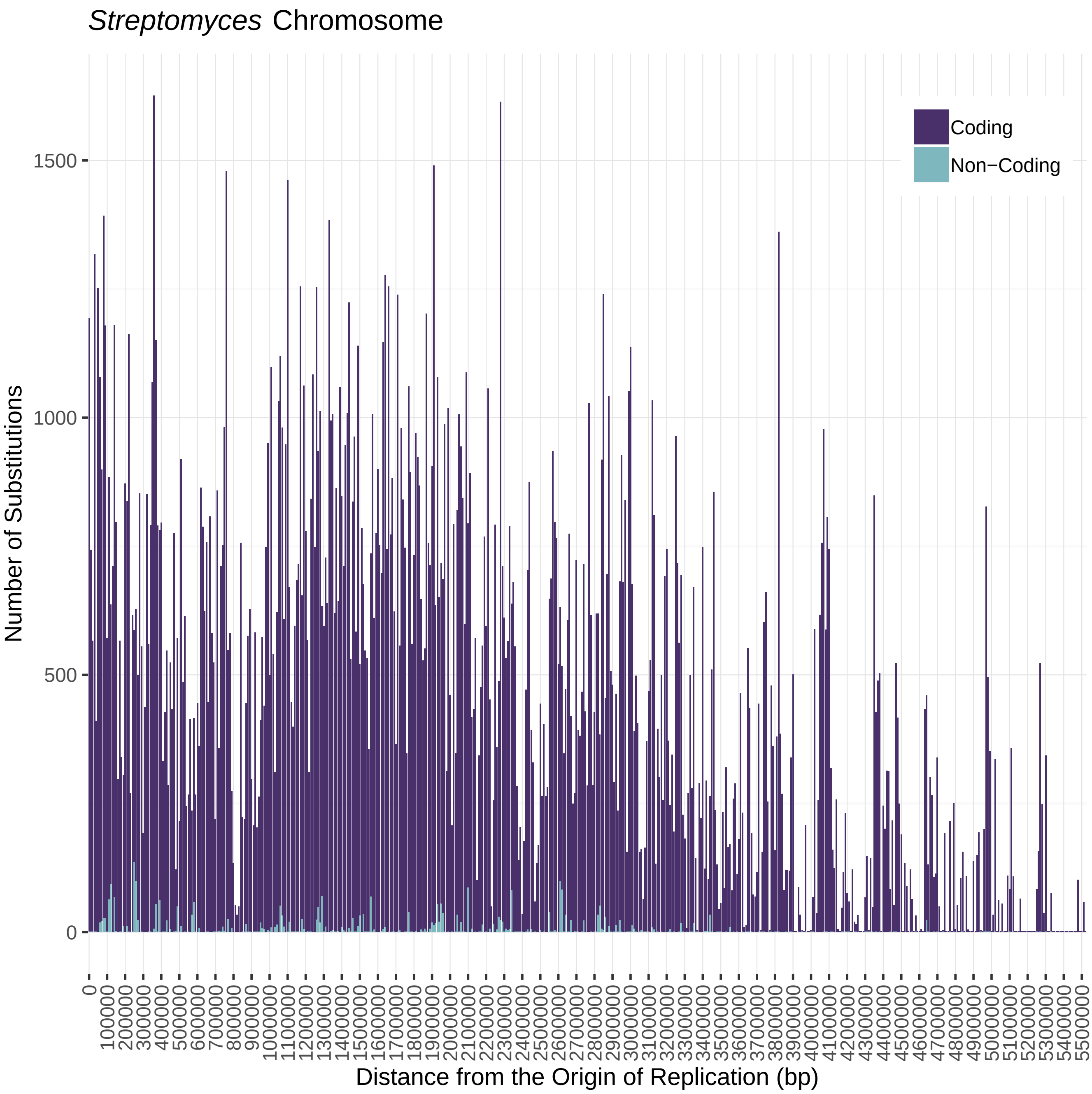
Bacteria and Replicon	Coding Sequences	Non-Coding Sequences
E. coli Chromosome	$-6.454\times10^{-8***}$	NS
B. subtilis Chromosome	$-1.159\times10^{-7***}$	$-9.861\times10^{-8***}$
Streptomyces Chromosome	$-8.464\times10^{-9***}$	$3.572\times10^{-7***}$
S. meliloti Chromosome	$-1.269\times10^{-7***}$	$-1.51\times10^{-7*}$
S. meliloti pSymA	$-2.02\times10^{-7***}$	NS
S. meliloti pSymB	$2.618\times10^{-7***}$	$8.591\times10^{-7***}$

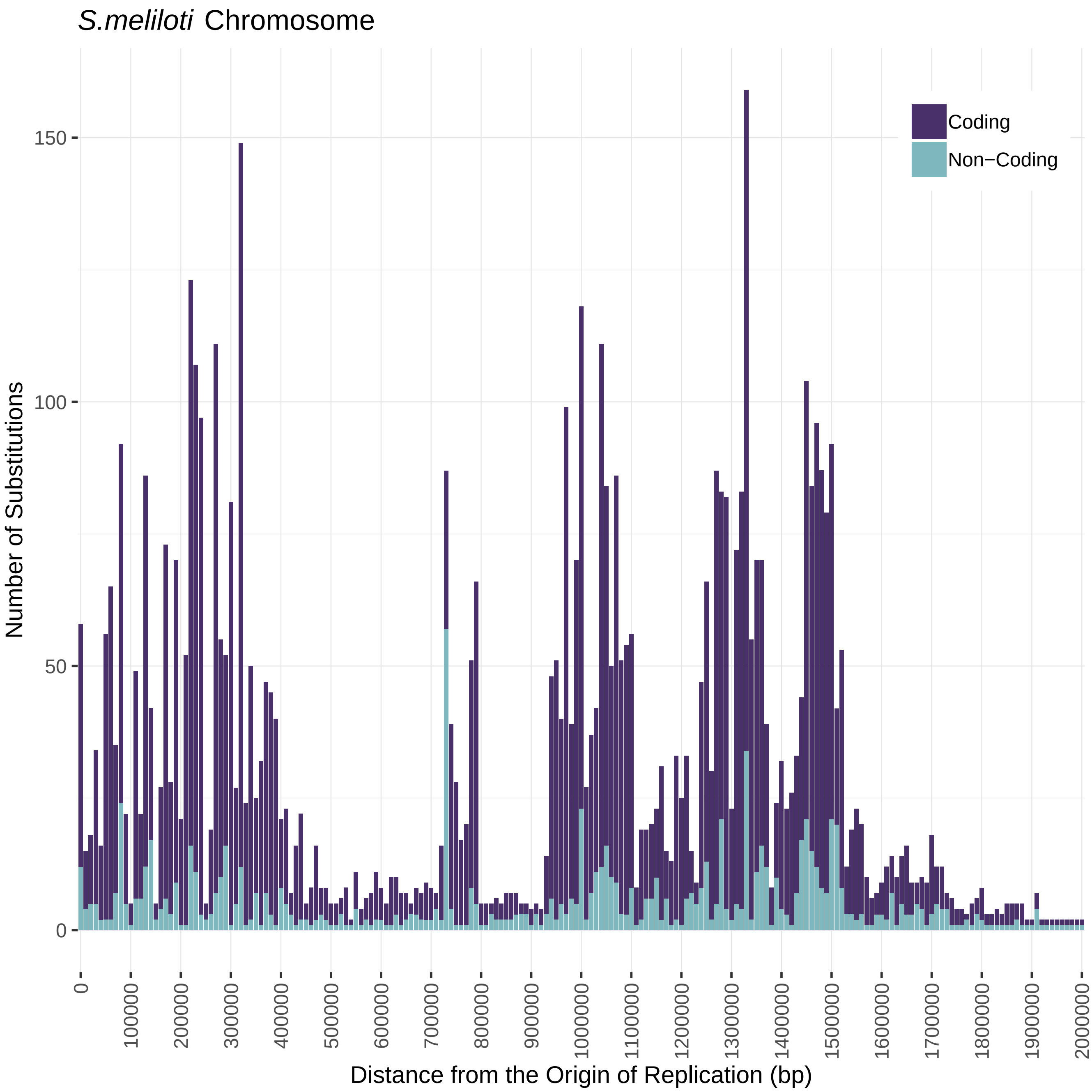
Table 2: Logistic regression analysis of the number of substitutions along all positions of the genome of the respective bacteria replicons. These genomic positions were split up into the coding and non-coding regions of the genome. Grey coloured boxes indicate a negative logistic regression coefficient 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'.

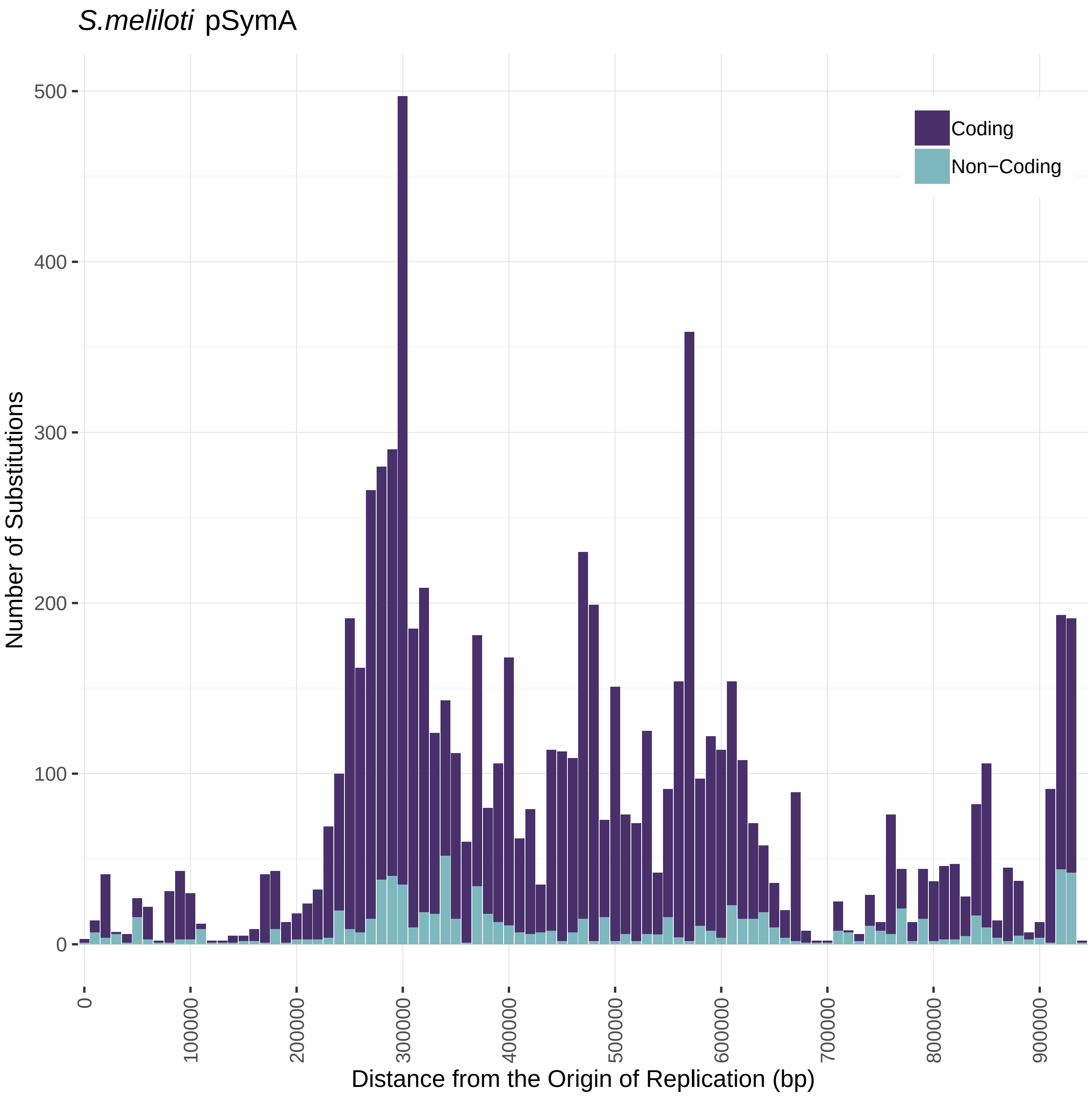
Sub density graphs with coding and non-coding information

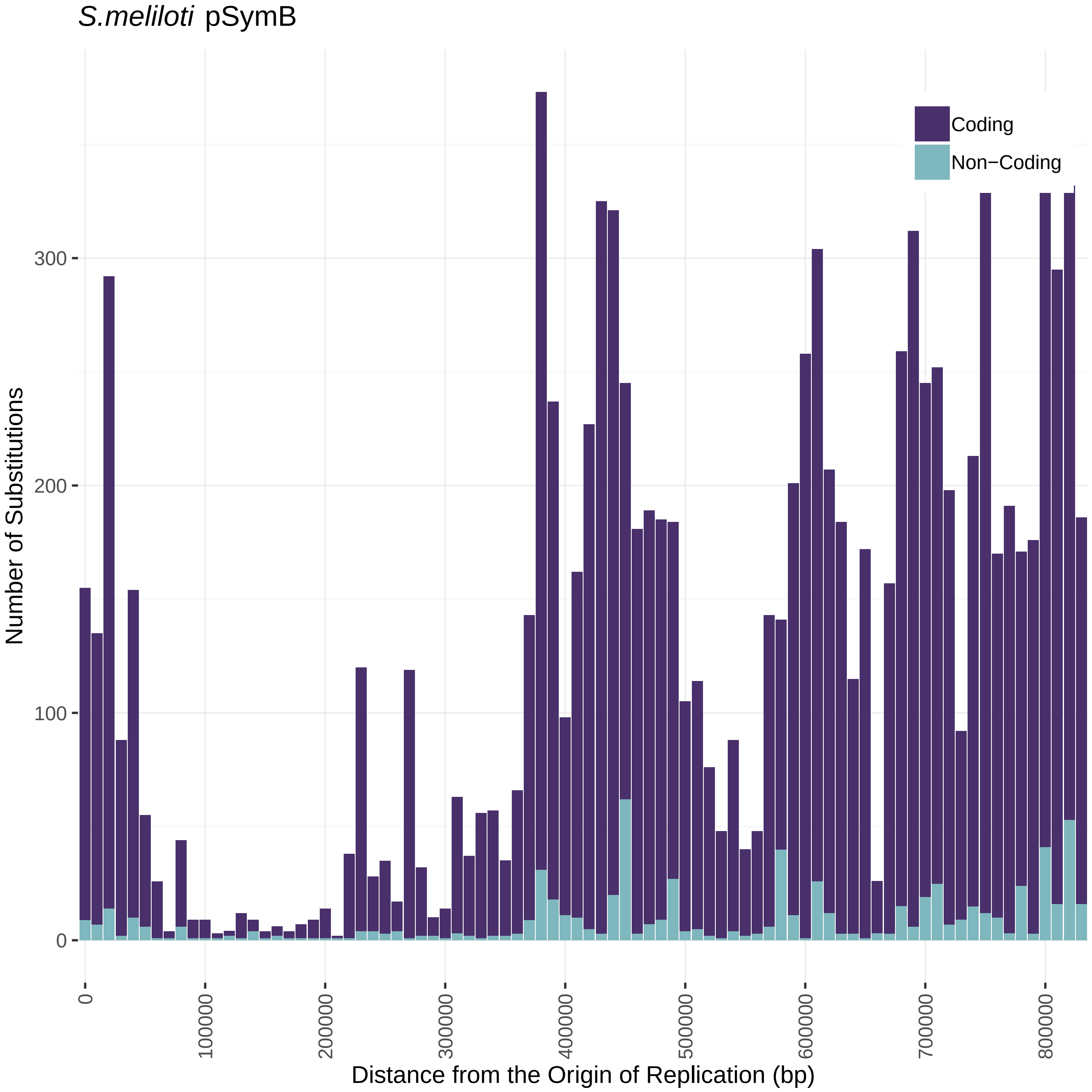

E.coli Chromosome
Coding
Non−Coding
Number of Substitutions
1500
1000
500
0
0
100000
200000
300000
400000
500000
600000
700000
800000
900000
1000000
1100000
1200000
1300000
1400000
1500000
1600000
1700000
1800000
1900000
2000000
2100000
2200000
2300000
2400000
2500000
2600000
2700000
2800000
2900000
3000000
3100000
3200000
Distance from the Origin of Replication (bp)


B.subtilis Chromosome
Number of Substitutions
3000
2000
1000
0
Coding
Non−Coding
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)

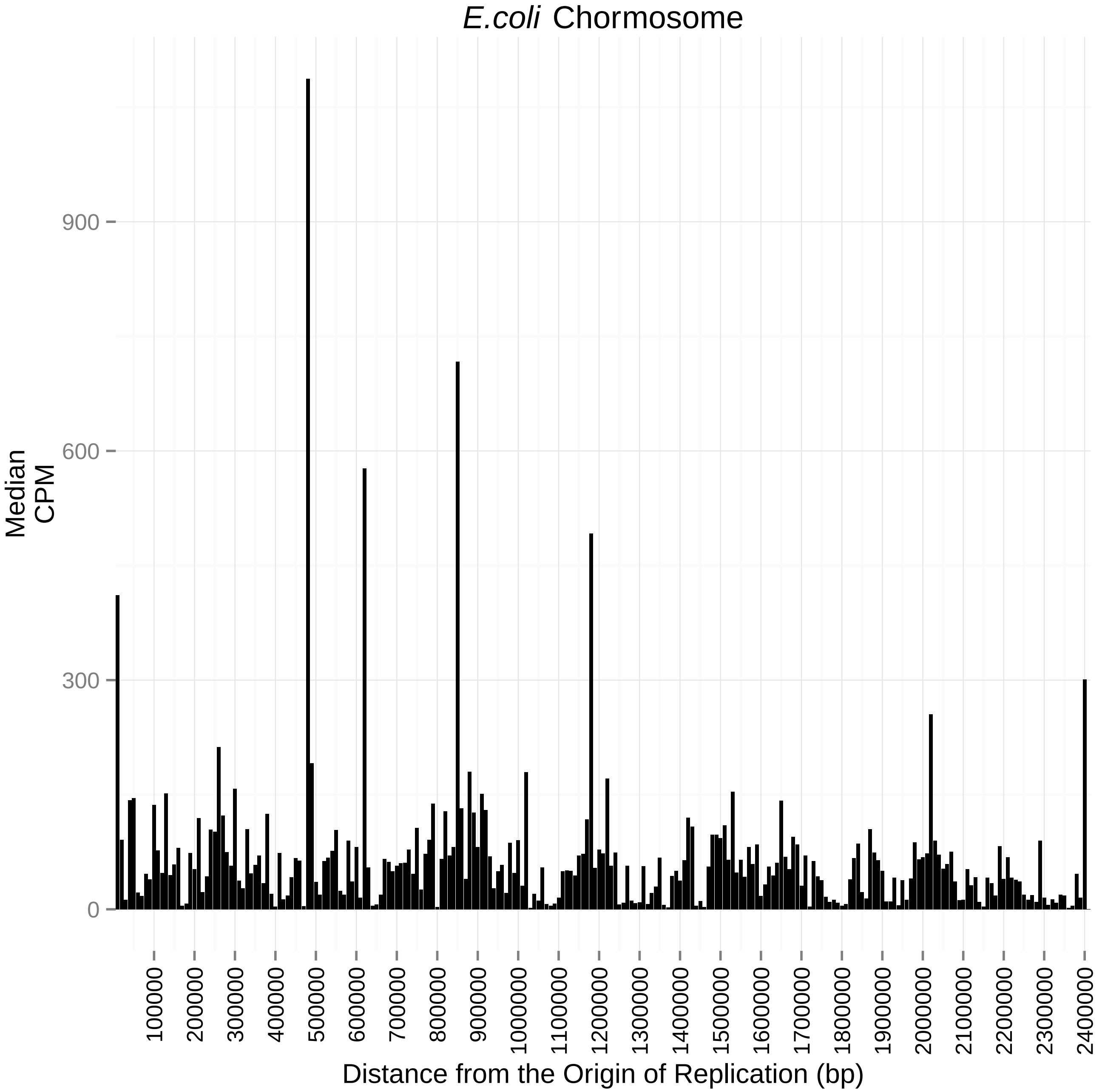

Streptomyces Chromosome
Number of Substitutions
1500
1000
500
0
Coding
Non−Coding
0
100000
200000
300000
400000
500000
600000
700000
800000
900000
1000000
1100000
1200000
1300000
1400000
1500000
1600000
1700000
1800000
1900000
2000000
2100000
2200000
2300000
2400000
2500000
2600000
2700000
2800000
2900000
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)

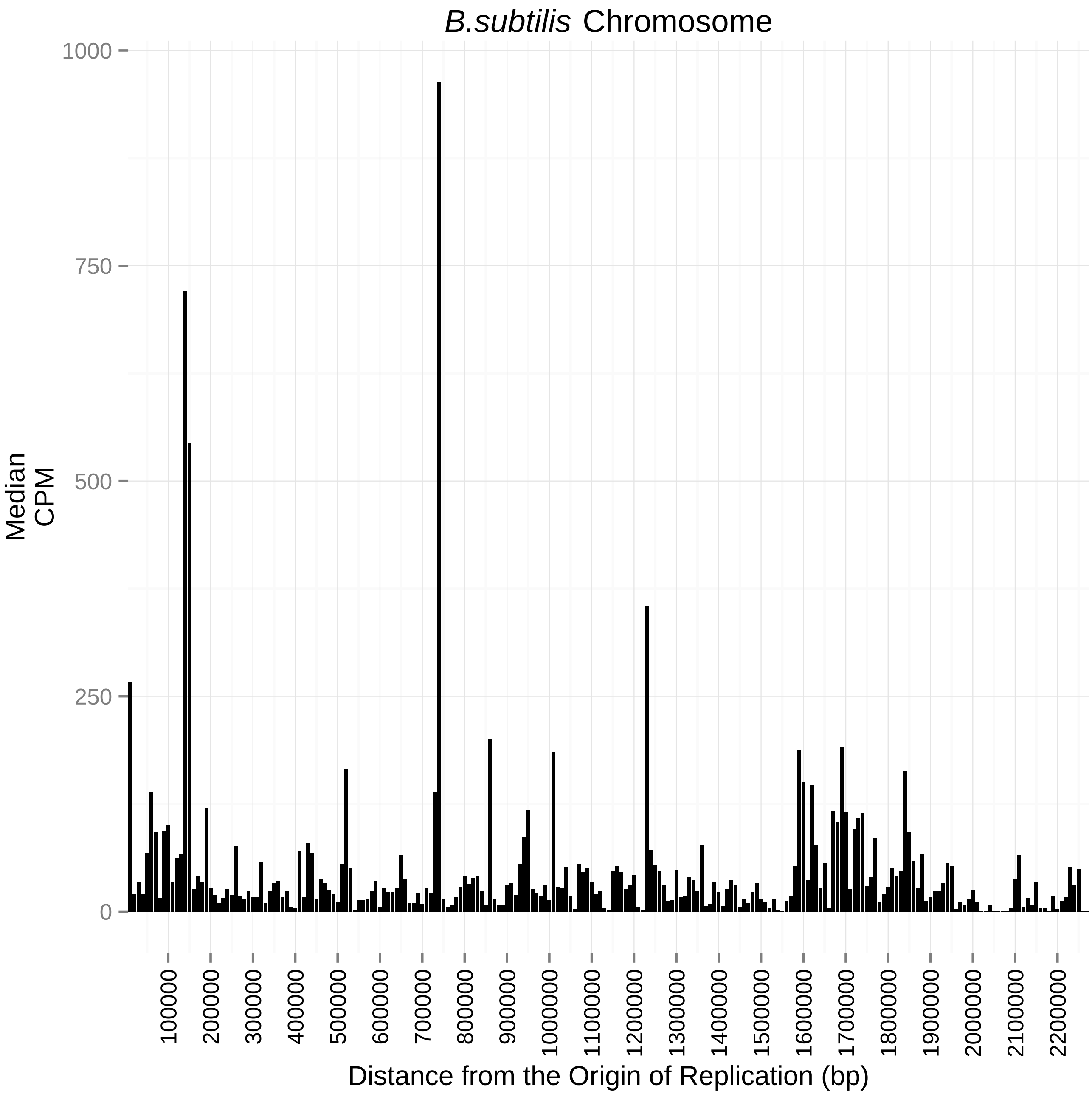

S.meliloti Chromosome
Number of Substitutions
0
50
100
150
Coding
Non−Coding
0
100000
200000
300000
400000
500000
600000
700000
800000
900000
1000000
1100000
1200000
1300000
1400000
1500000
1600000
1700000
1800000
1900000
2000000
Distance from the Origin of Replication (bp)

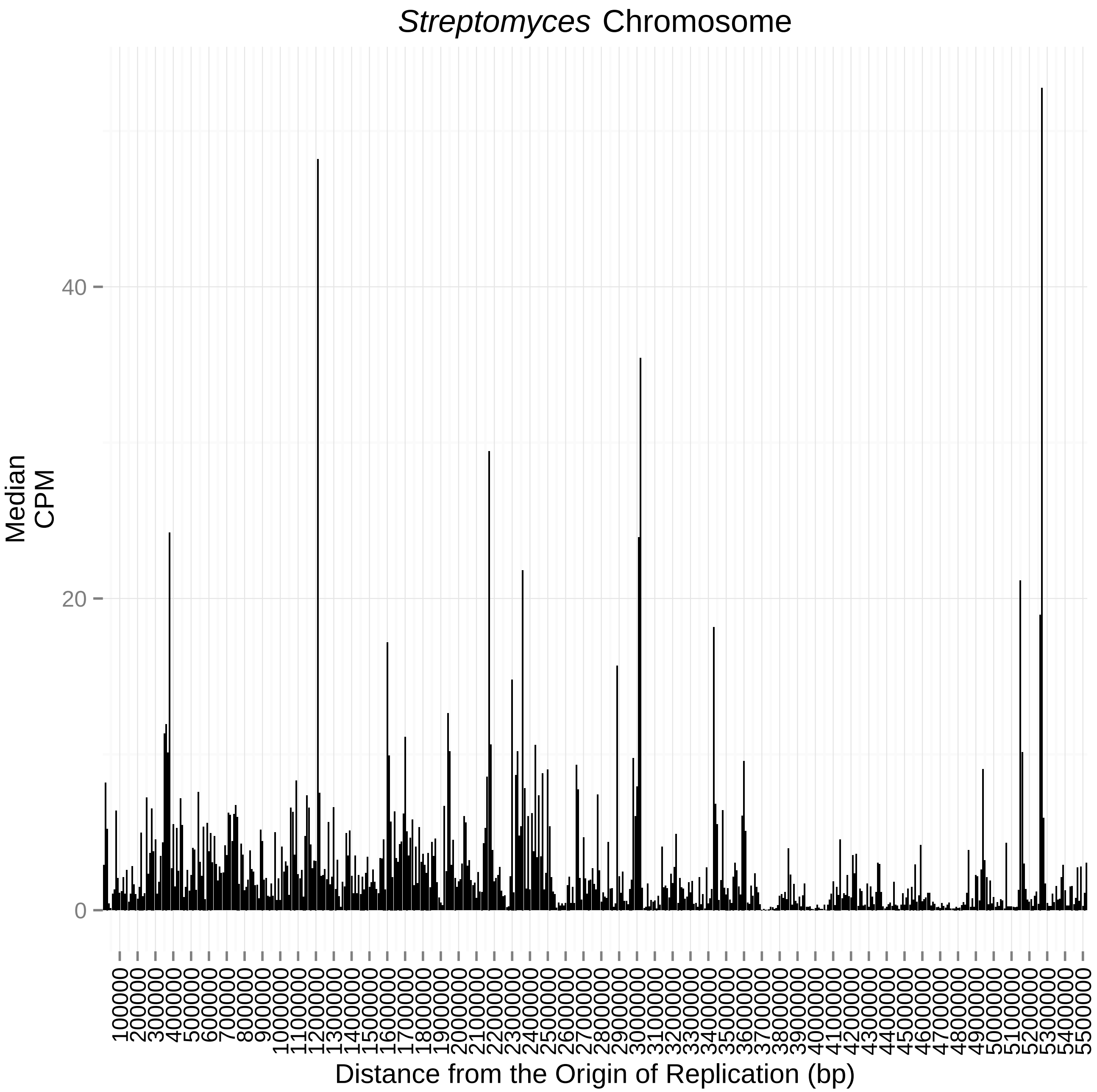

S.meliloti pSymA
Number of Substitutions
500
400
300
200
100
0
Coding
Non−Coding
0
100000
200000
300000
400000
500000
600000
700000
800000
900000
Distance from the Origin of Replication (bp)

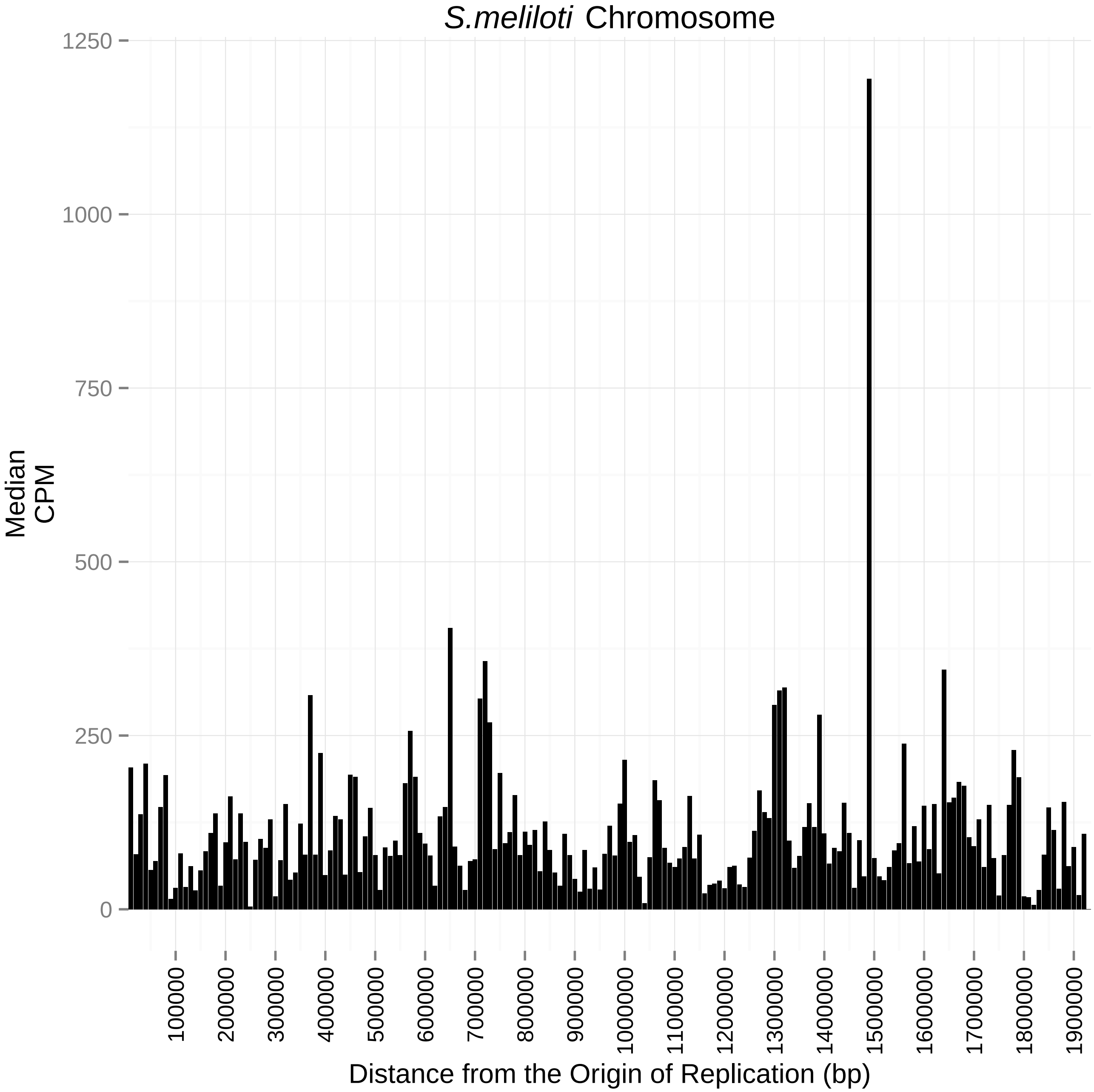


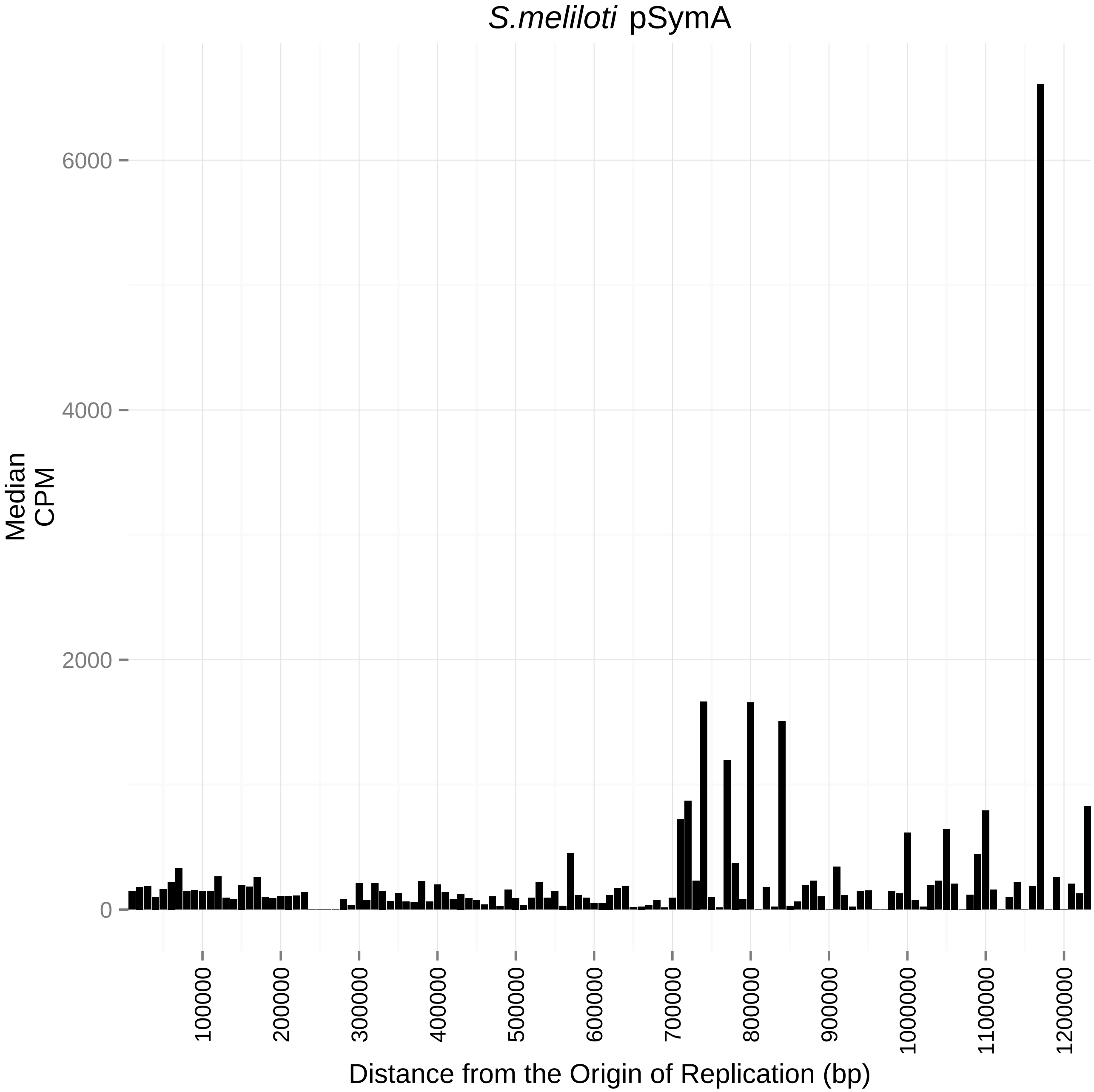
Gene expression graphs

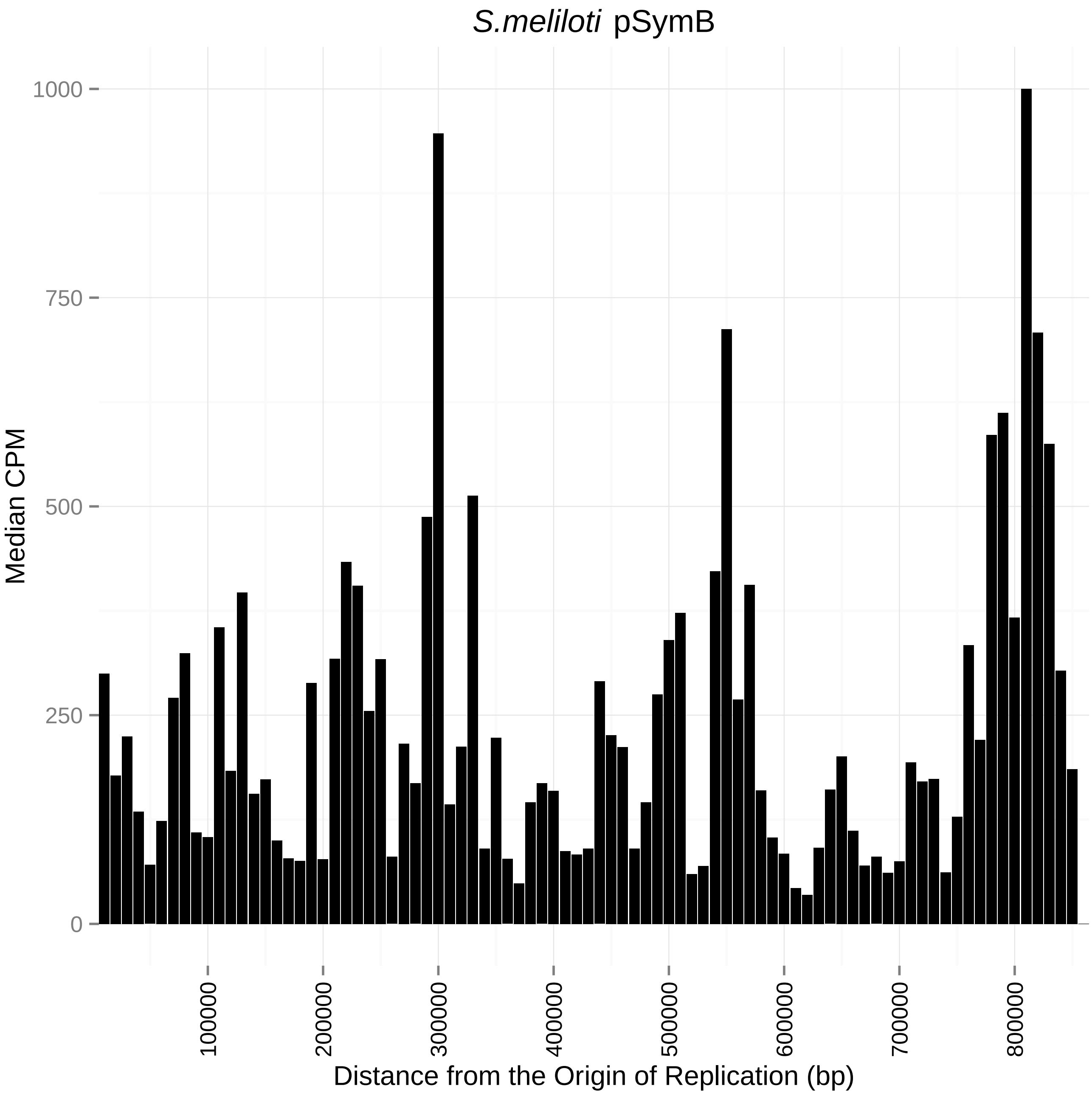

E.coli Chormosome
Median CPM
900
600
300
0
100000
200000
300000
400000
500000
600000
700000
800000
900000
1000000
1100000
1200000
1300000
1400000
1500000
1600000
1700000
1800000
1900000
2000000
2100000
2200000
2300000
2400000
Distance from the Origin of Replication (bp)

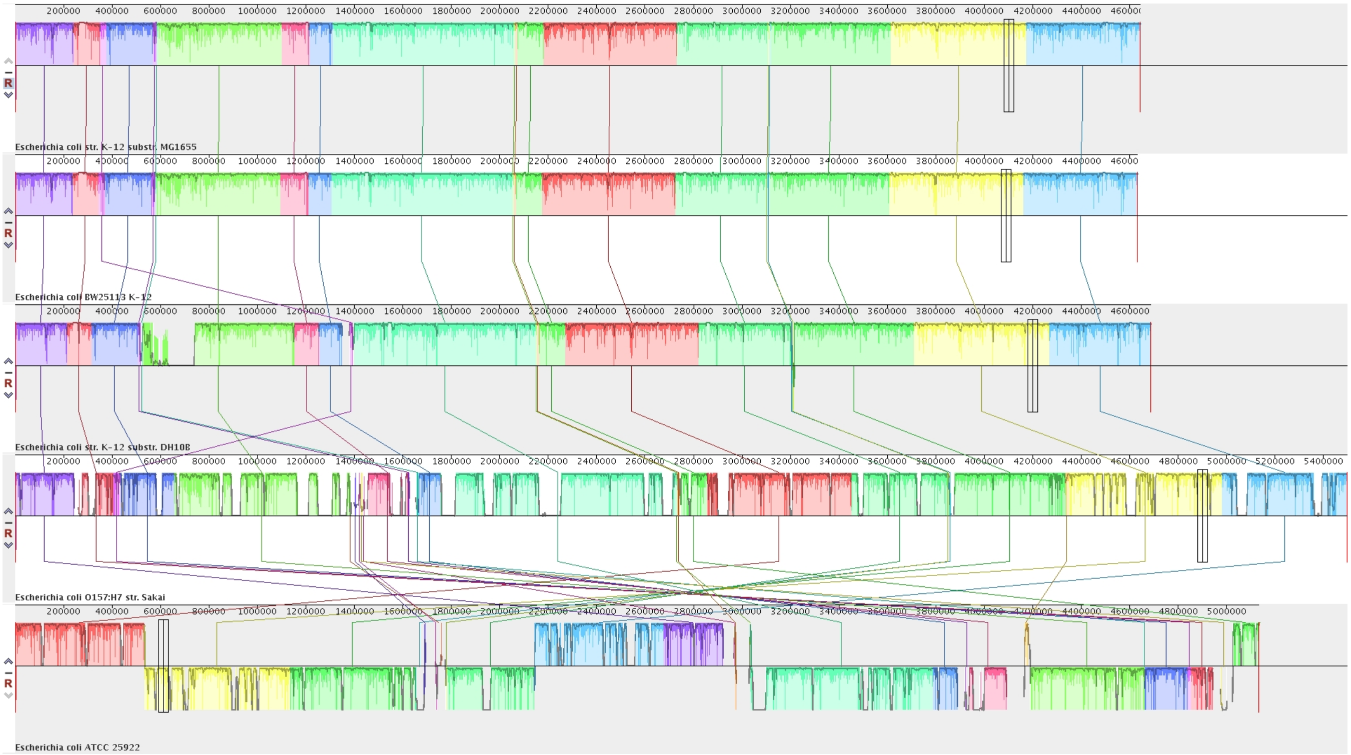

B.subtilis Chromosome
Median CPM
1000
750
500
250
0
100000
200000
300000
400000
500000
600000
700000
800000
900000
1000000
1100000
1200000
1300000
1400000
1500000
1600000
1700000
1800000
1900000
2000000
2100000
2200000
Distance from the Origin of Replication (bp)

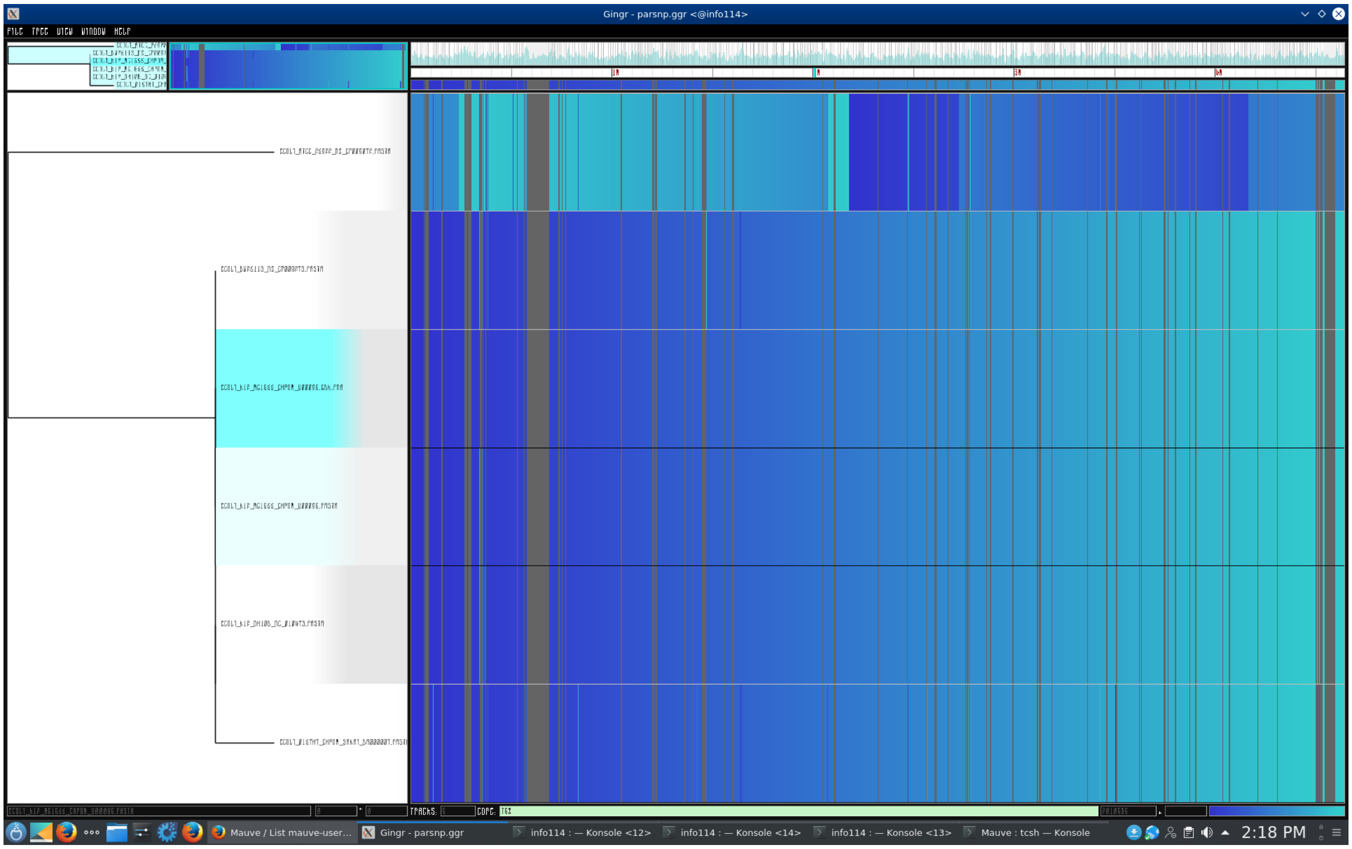

Streptomyces Chromosome
Median CPM
0
20
40
100000
200000
300000
400000
500000
600000
700000
800000
900000
1000000
1100000
1200000
1300000
1400000
1500000
1600000
1700000
1800000
1900000
2000000
2100000
2200000
2300000
2400000
2500000
2600000
2700000
2800000
2900000
3000000
3100000
3200000
3300000
3400000
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
Median CPM
6000
4000
2000
0
100000
200000
300000
400000
500000
600000
700000
800000
900000
1000000
1100000
1200000
Distance from the Origin of Replication (bp)


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


200000 400000 600000 800000 1000000 1200000 1400000 1600000 1800000 2000000 2200000 2400000 2600000 2800000 3000000 3200000 3400000 3600000 3800000 4000000 4200000 4400000 4600000
R
Escherichia coli str. K-12 substr. MG1655
R
Escherichia coli BW25113 K-12
R
Escherichia coli str. K-12 substr. DH10B
4800000 5000000 5200000 5400000
R
Escherichia coli O157:H7 str. Sakai
R
Escherichia coli ATCC 25922


Gingr - parsnp.ggr <@info114>
FILE TREE VIEW WINDOW HELP
ECOLI_ATCC_25922_NZ_CP009072.FASTA
ECOLI_BW25113_NZ_CP009273.FASTA
ECOLI_K12_MG1655_CHROM_U00096.GBK.FNA
ECOLI_K12_MG1655_CHROM_U00096.FASTA
ECOLI_K12_DH10B_NC_010473.FASTA
ECOLI_O157H7_CHROM_SAKAI_BA000007.FASTA
TRACKS:
CORE:
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 3: 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}	$<2 \times 10^{-16}$
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 4: 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.