

Subs Paper Things to Do:

- ~~# of coding and non-coding sites~~
- ~~# of subs in each of \uparrow~~
- 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

Preparing for the OE3C conference and the conference itself took up majority of my time last week.

Other than that, I noticed that I was still getting some random “stop” codons in the codeml analysis so I have been looking into this further and found that since we are only using one taxa to specify the gene starts and stops, there are some rare cases where in the reference taxa the codon is TGG but this has been changed in one of the other taxa to TGA (stop codon) so this whole column is therefore omitted from the rest of the dN/dS analysis. I think this situation is unavoidable and since the column is removed anyways it should not impact the results. This also happens rarely.

I also realized that for some reason some of the blocks did not run for the coding/non-coding substitution analysis... so I am currently re-running those. I do not expect the results to change that much.

I talked to you about the weirdly high dN/dS ratios. These are because in some of the blocks/sections there are ONLY non-synonymous substitutions, and so the ratio is undefined and codeml really freaks out and gives high ratios like 999.99. Codeml also freaks out and gives wonky ratios when there are no substitutions. However, we discussed this and you suggested just taking the total counts of non-synonymous and synonymous substitutions and dividing this by the total number of sites and then get ratios based on that.

This Week

I would like to figure out how to combine the dN/dS results into one summary table for each of the bacteria and calculate the actual number of synonymous and non-synonymous substitutions.

I also want to finish re-running the coding/non-coding analysis.

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

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.

Bacteria and Replicon	dN/dS Ratio	
	Gene Average	Genome Average
<i>E. coli</i> Chromosome		0.0612
<i>B. subtilis</i> Chromosome		0.0911
<i>Streptomyces</i> Chromosome		
<i>S. meliloti</i> Chromosome		157.0089
<i>S. meliloti</i> pSymA		30.1216
<i>S. meliloti</i> pSymB		21.0320

Table 1: dN/dS ratio calculated for each bacteria on a per gene and per genome basis. The per gene average was calculated by taking the average dN/dS of each gene, and then averaging these. The per genome average was calculated by taking the average of all dN/dS ratios.

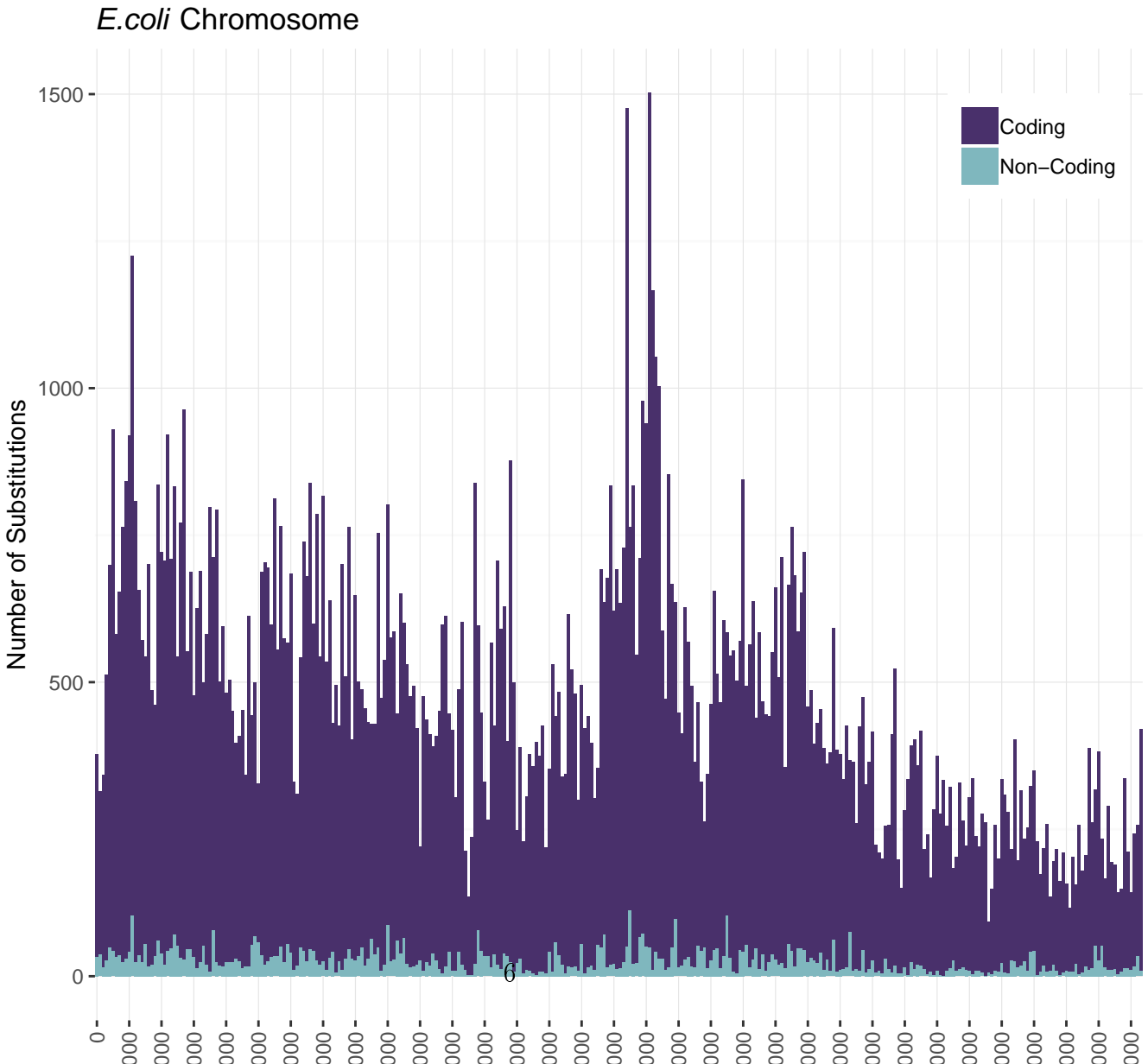
Bacteria and Replicon	Average Replicon Length	# of Coding Sites	# of Non-Coding Sites	# of Subs Coding	# of Subs Non-Coding
<i>E. coli</i> Chromosome	5082529	5203089	191748	169620	9534
<i>B. subtilis</i> Chromosome	4077077	2074653	102906	205150	6187
<i>Streptomyces</i> Chromosome	8497577	2422980	21581	551530	3670
<i>S. meliloti</i> Chromosome	3426881	1931139	199425	6684	842
<i>S. meliloti</i> pSymA	1455940	419223	34213	9832	943
<i>S. meliloti</i> pSymB	1664597	552816	22098	11699	645

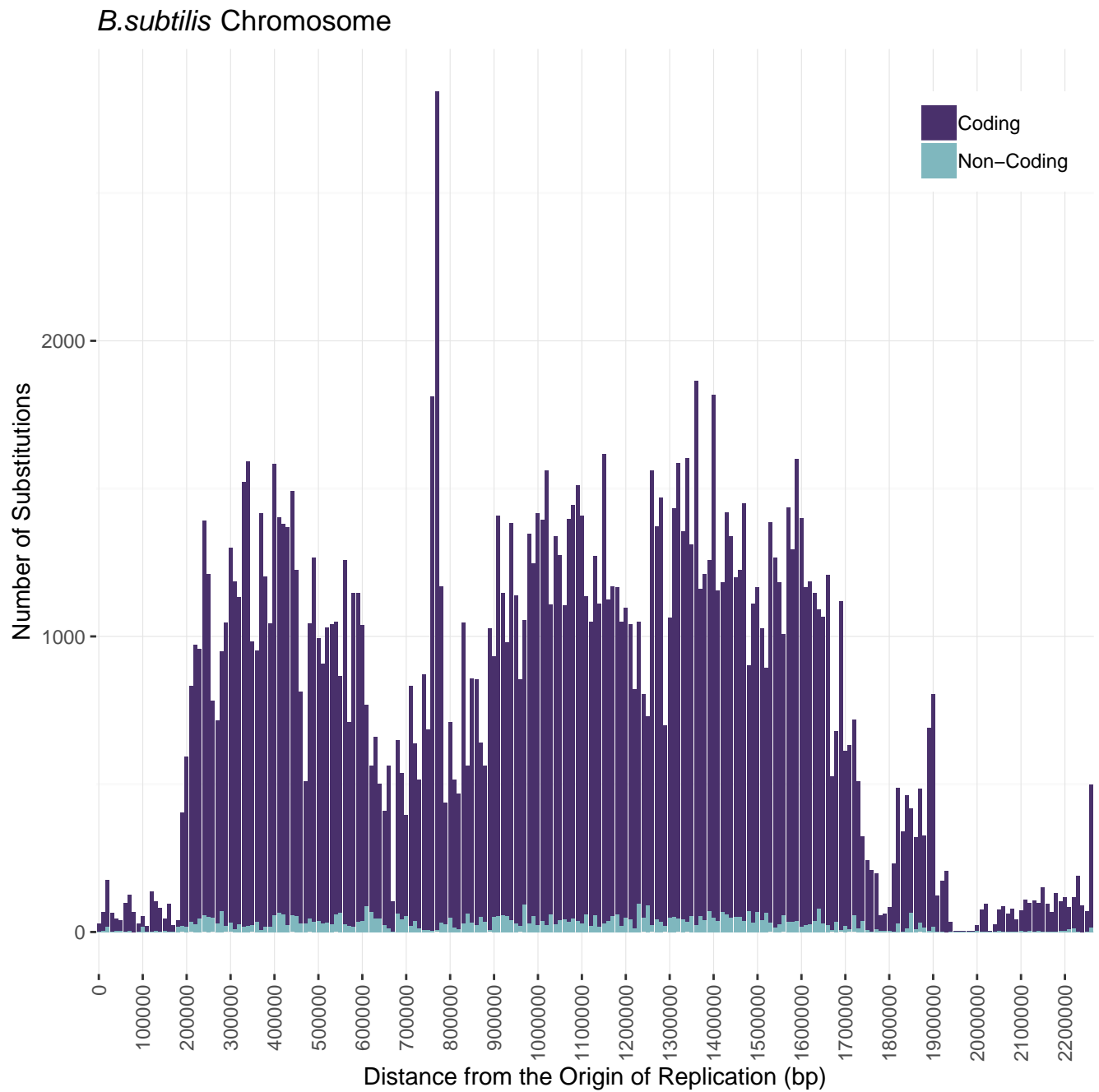
Table 2: Total proportion of coding and non-coding sites in each replicon and the percentage of those sites that have a substitution (multiple substitutions at one site are counted as two substitutions).

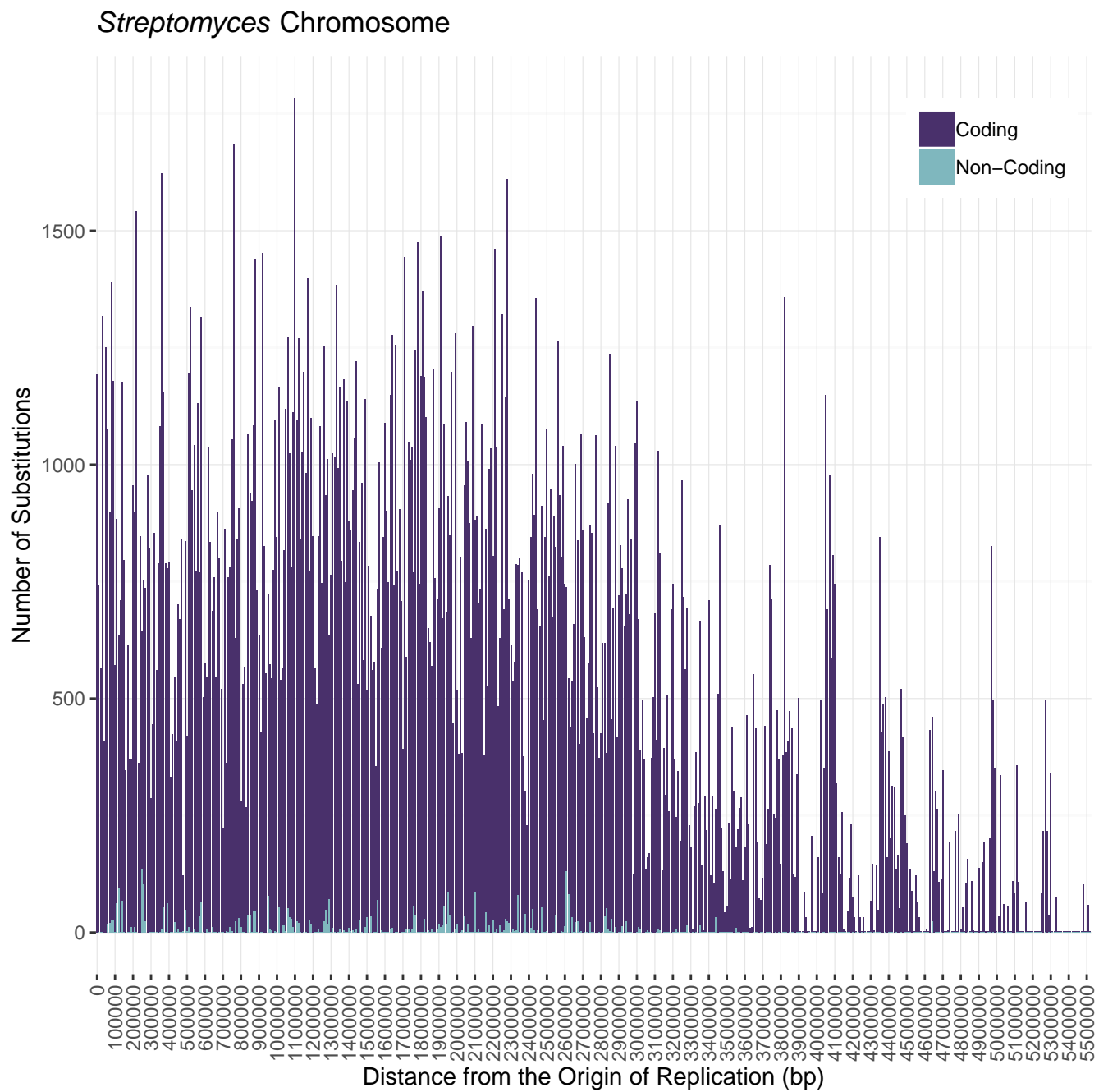
Sub density graphs with coding and non-coding information

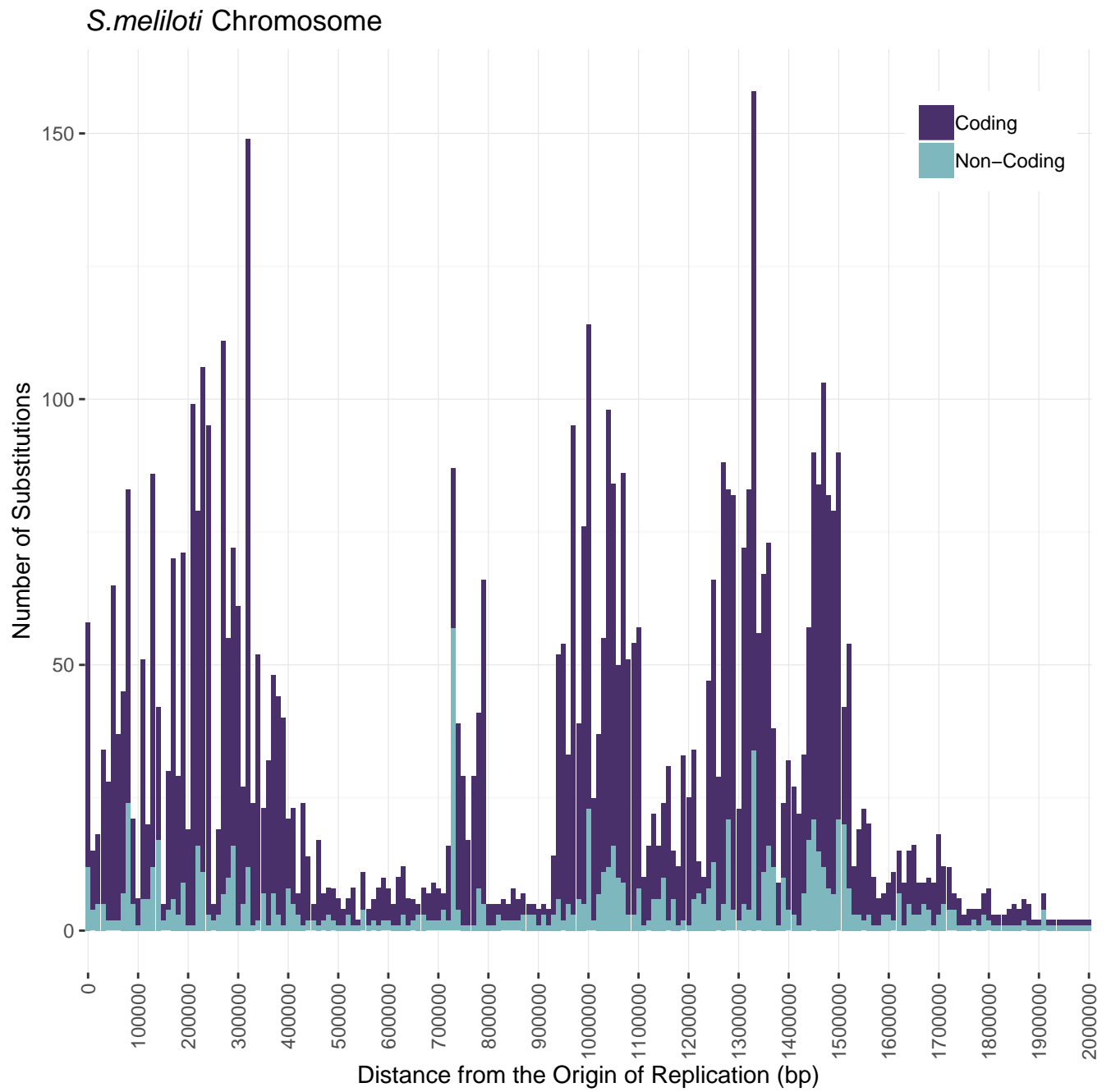
Bacteria and Replicon	Coding Sequences	Non-Coding Sequences
<i>E. coli</i> Chromosome	$-9.119 \times 10^{-8***}$	$7.022 \times 10^{-8***}$
<i>B. subtilis</i> Chromosome	$-1.071 \times 10^{-7***}$	$-9.861 \times 10^{-8***}$
<i>Streptomyces</i> Chromosome	$-7.945 \times 10^{-9***}$	$3.637 \times 10^{-7***}$
<i>S. meliloti</i> Chromosome	$-1.367 \times 10^{-7***}$	$-1.510 \times 10^{-7*}$
<i>S. meliloti</i> pSymA	$-1.075 \times 10^{-7*}$	NS
<i>S. meliloti</i> pSymB	$2.878 \times 10^{-7***}$	$8.595 \times 10^{-7***}$

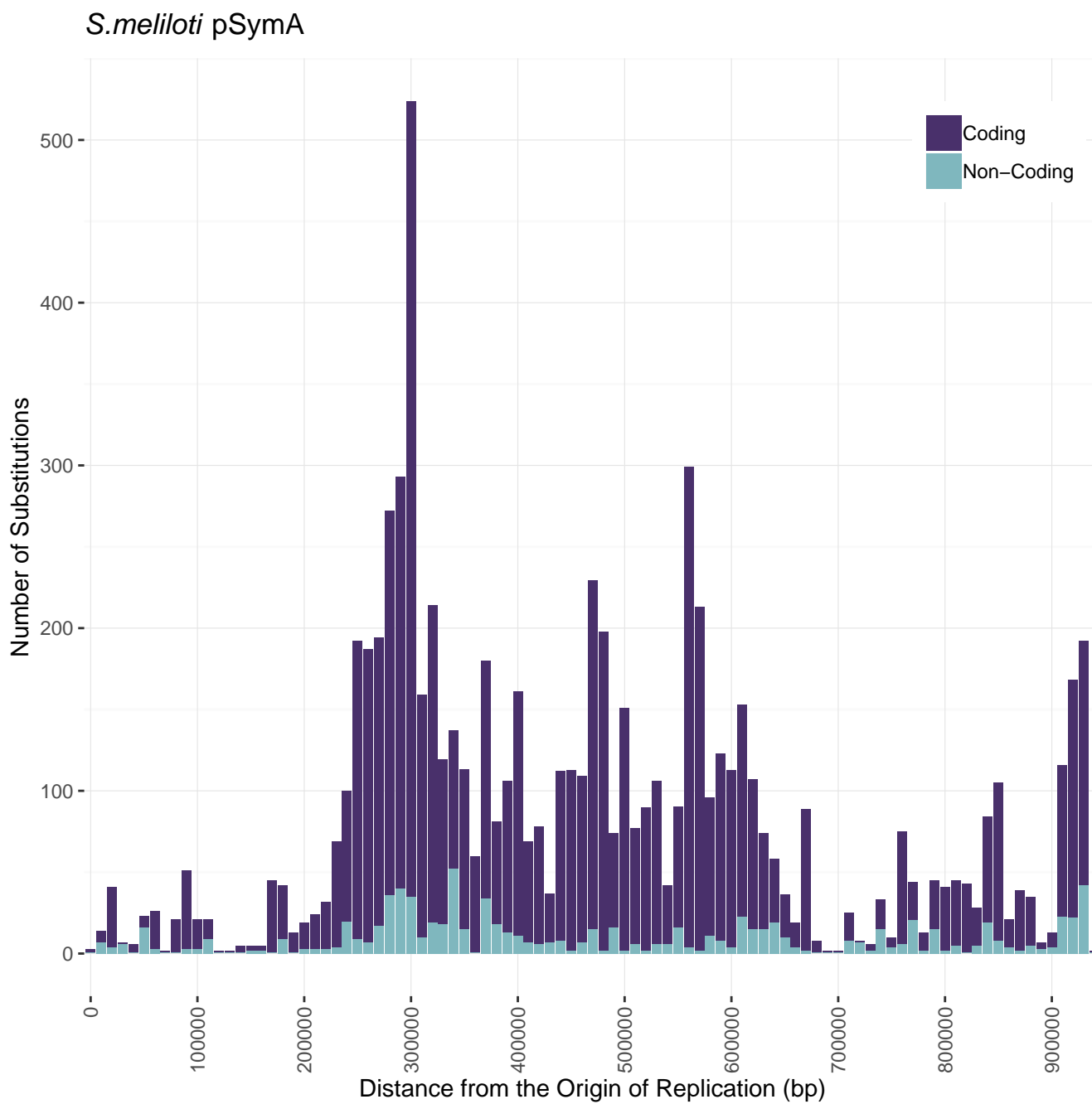
Table 3: Logistic regression analysis of the number of substitutions along all positions of the genome of the respective bacteria replicons. These genomic positions were split up into the coding and non-coding regions of the genome. Grey coloured boxes indicate a negative logistic regression coefficient 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: $< 0.001 = '***'$, $0.001 < 0.01 = '**'$, $0.01 < 0.05 = '*'$, $> 0.05 = 'NS'$.

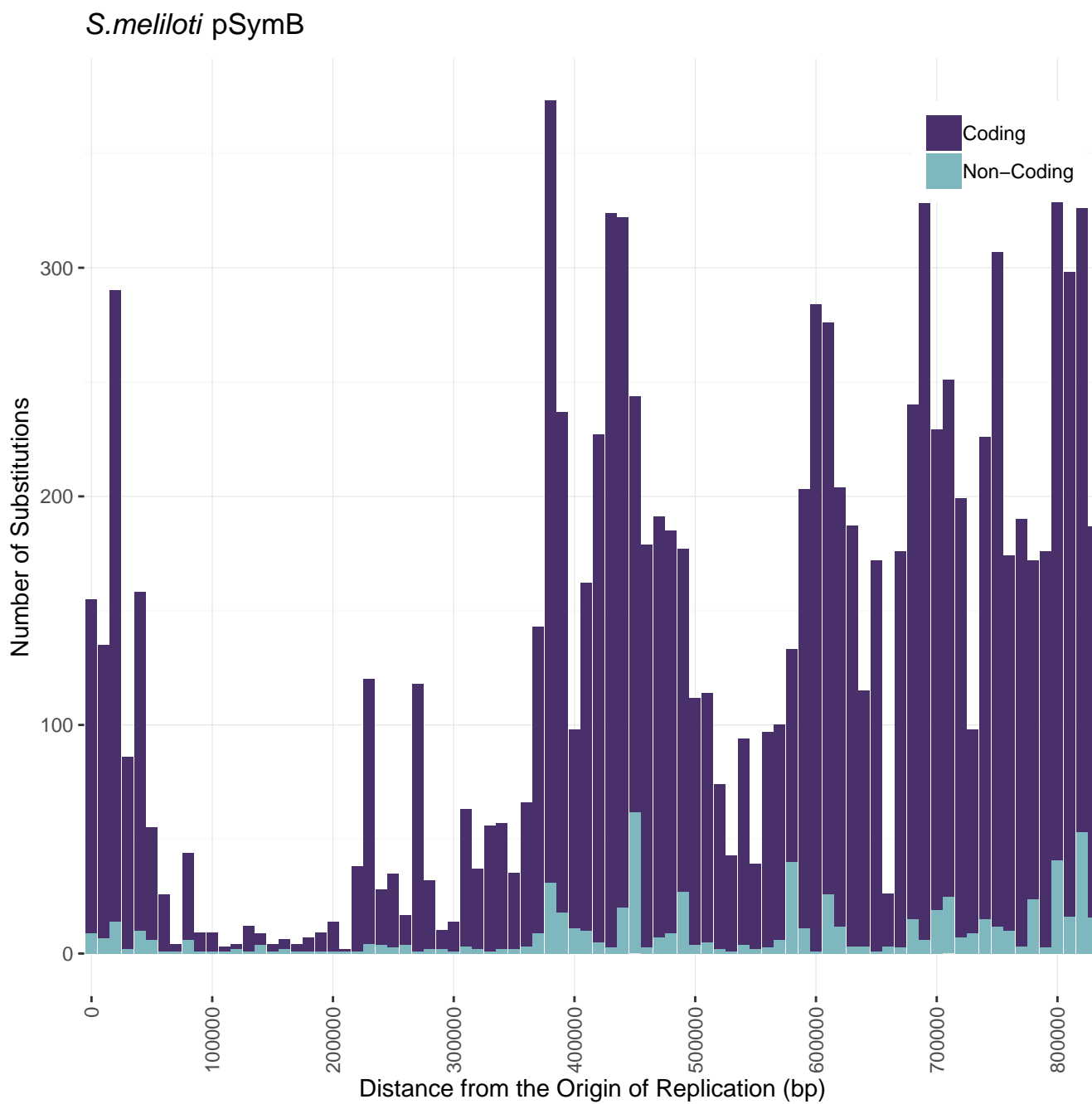




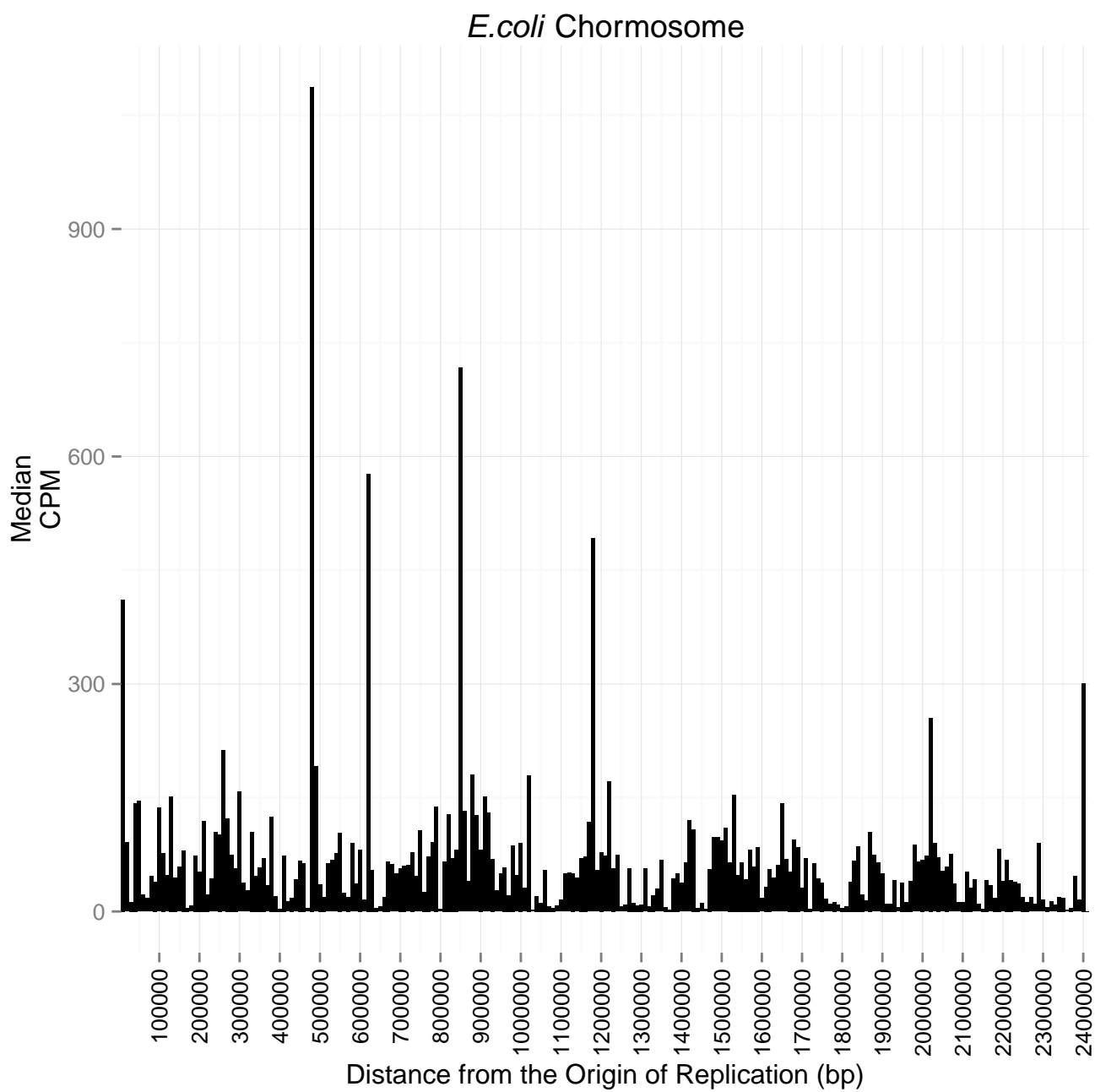


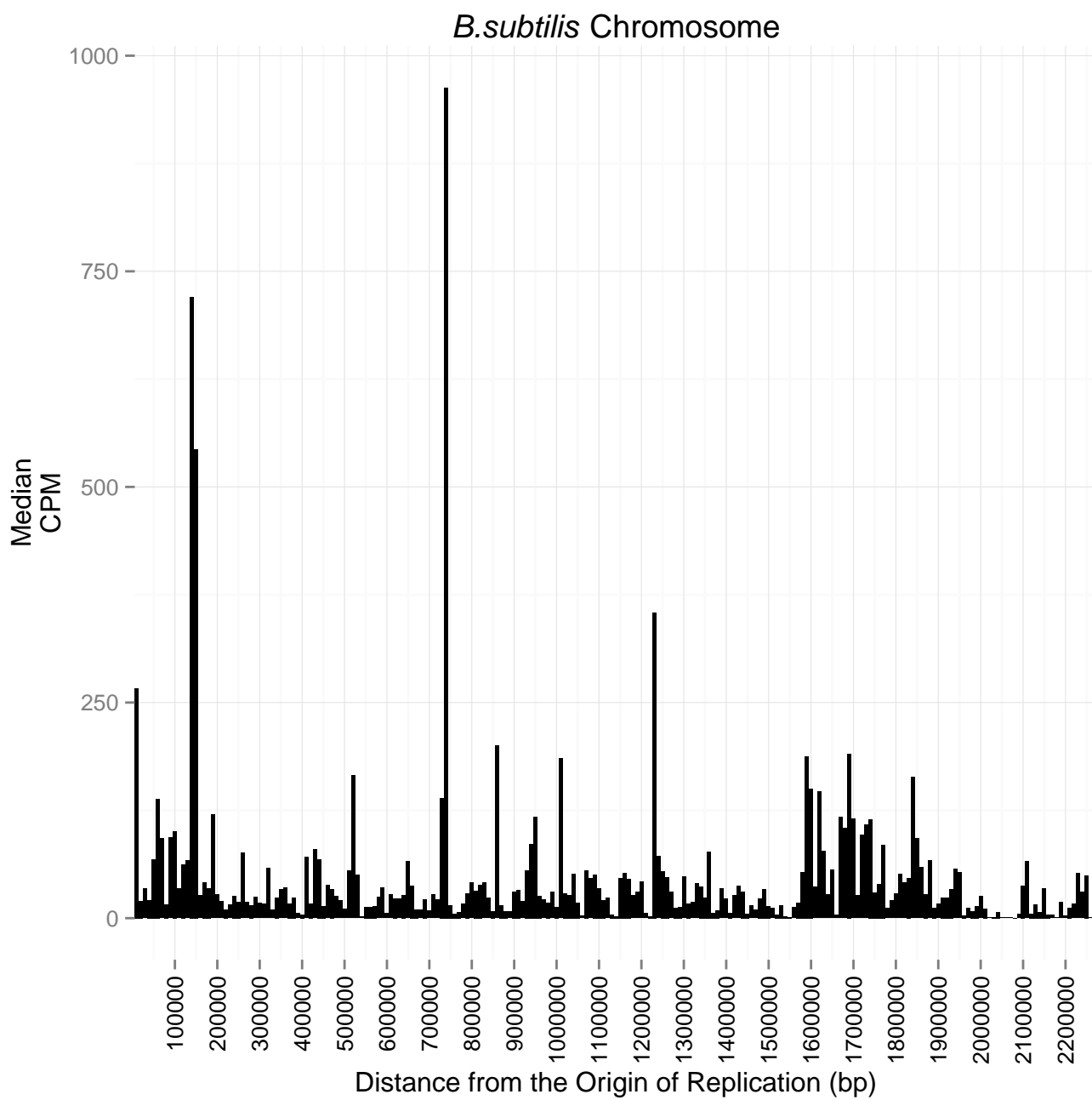


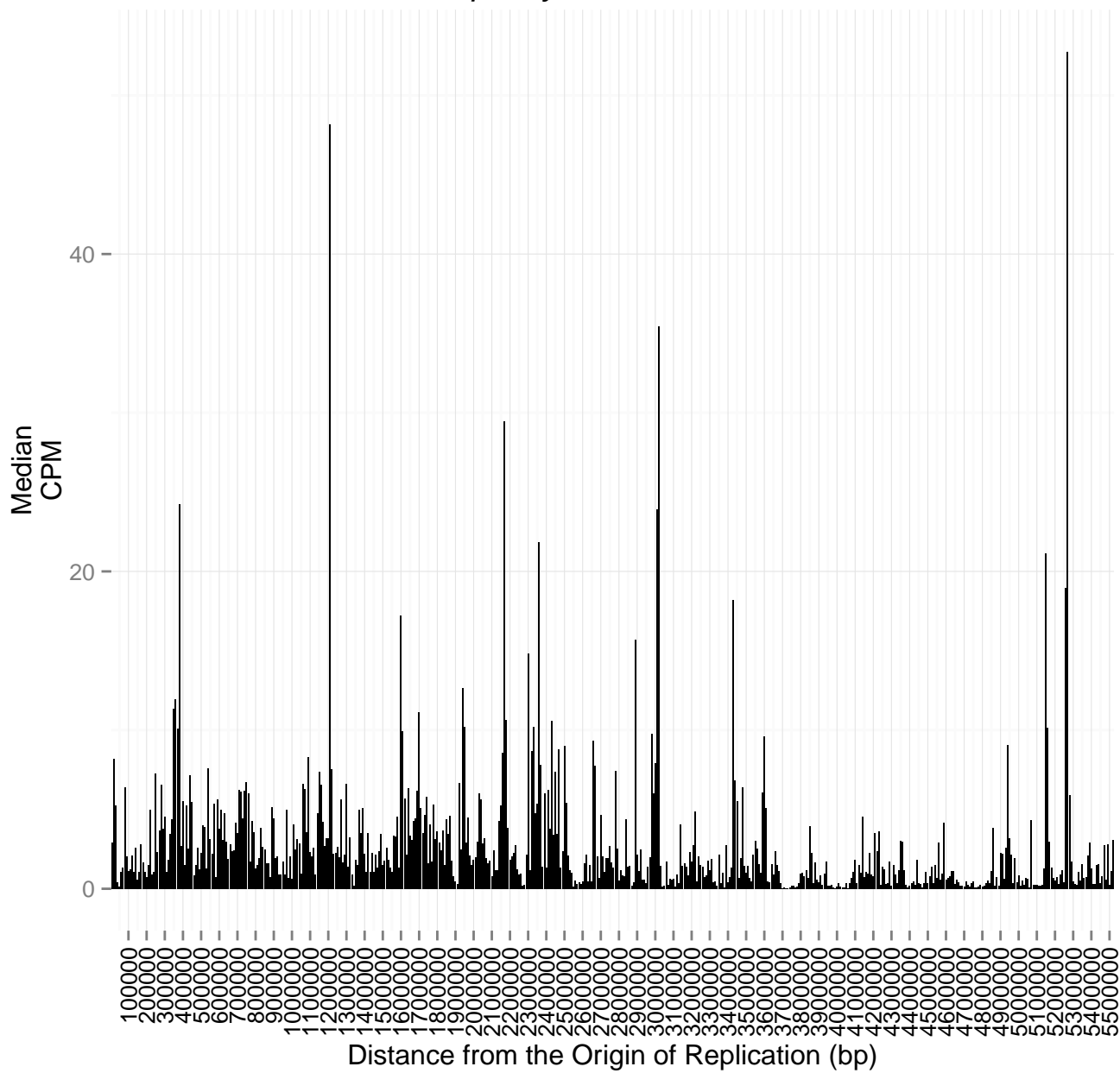


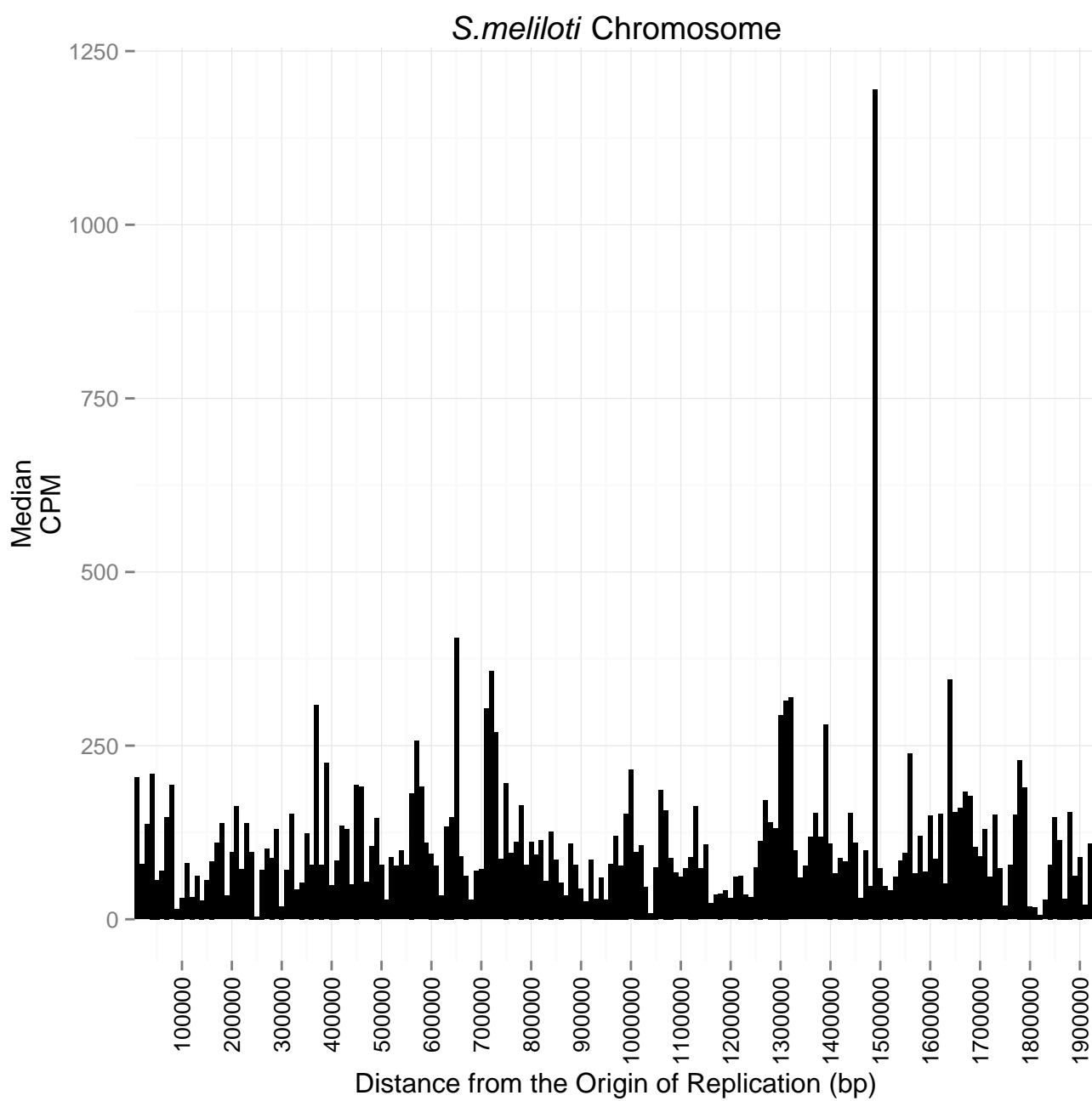


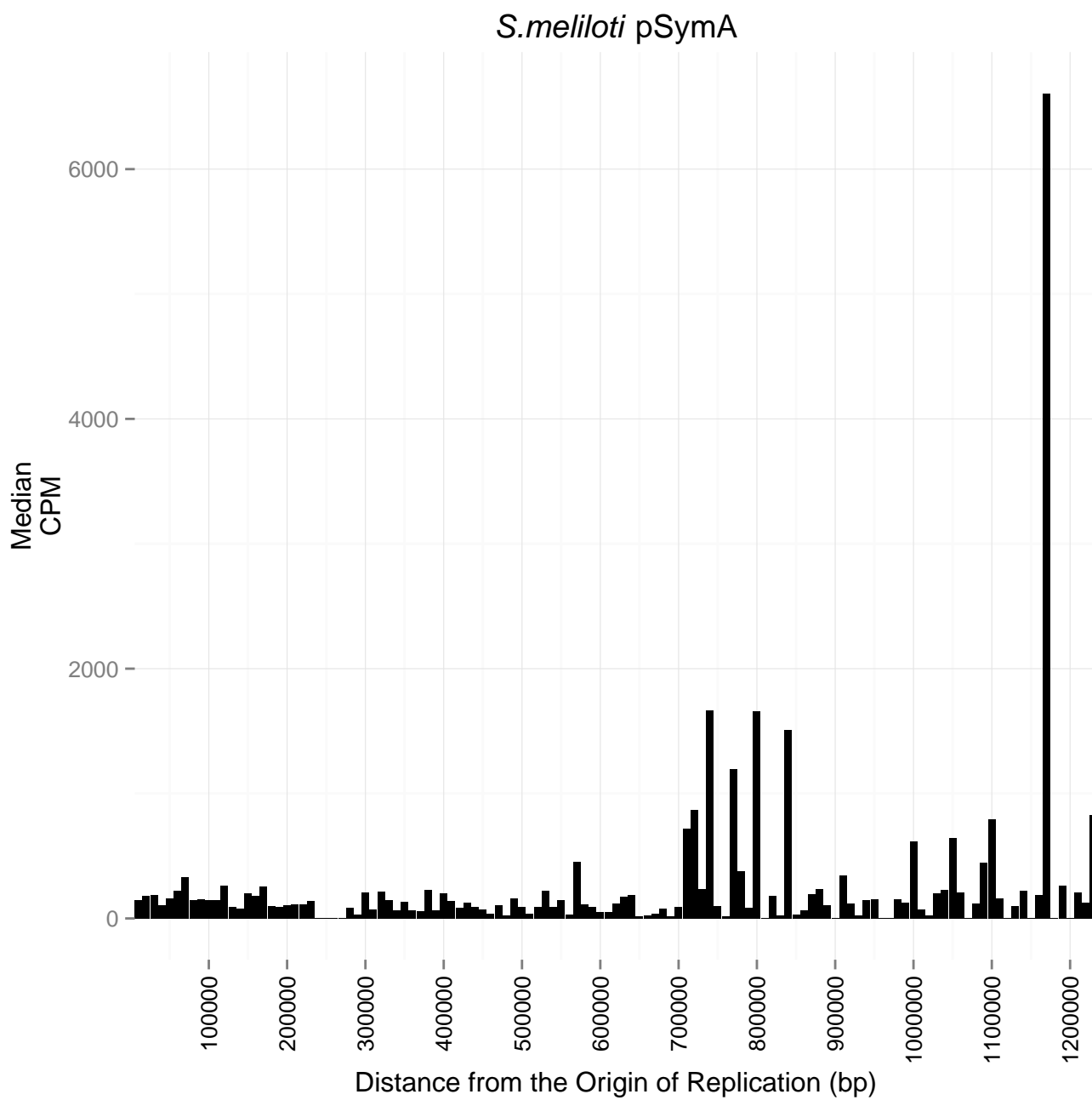
Gene expression graphs

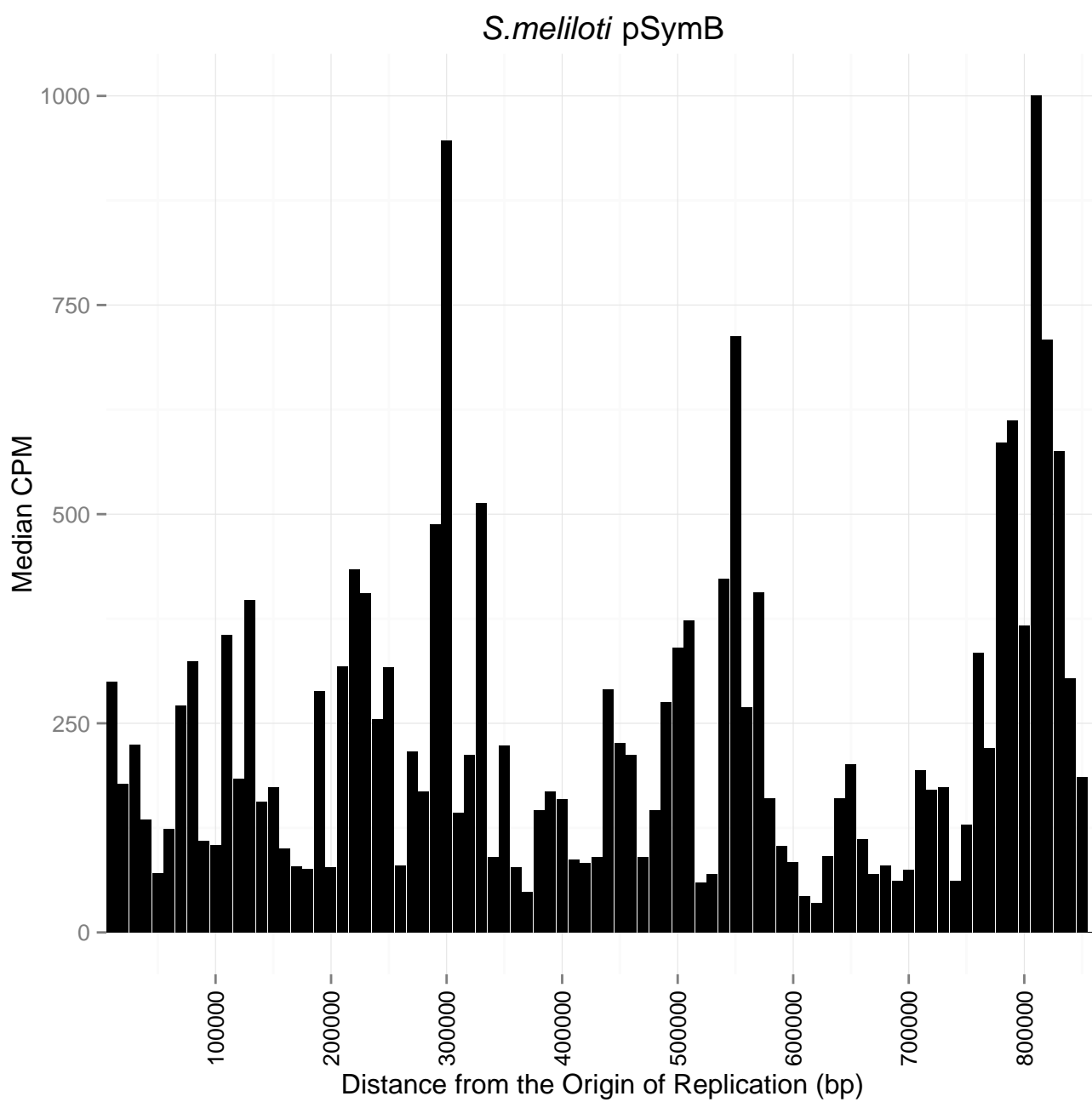


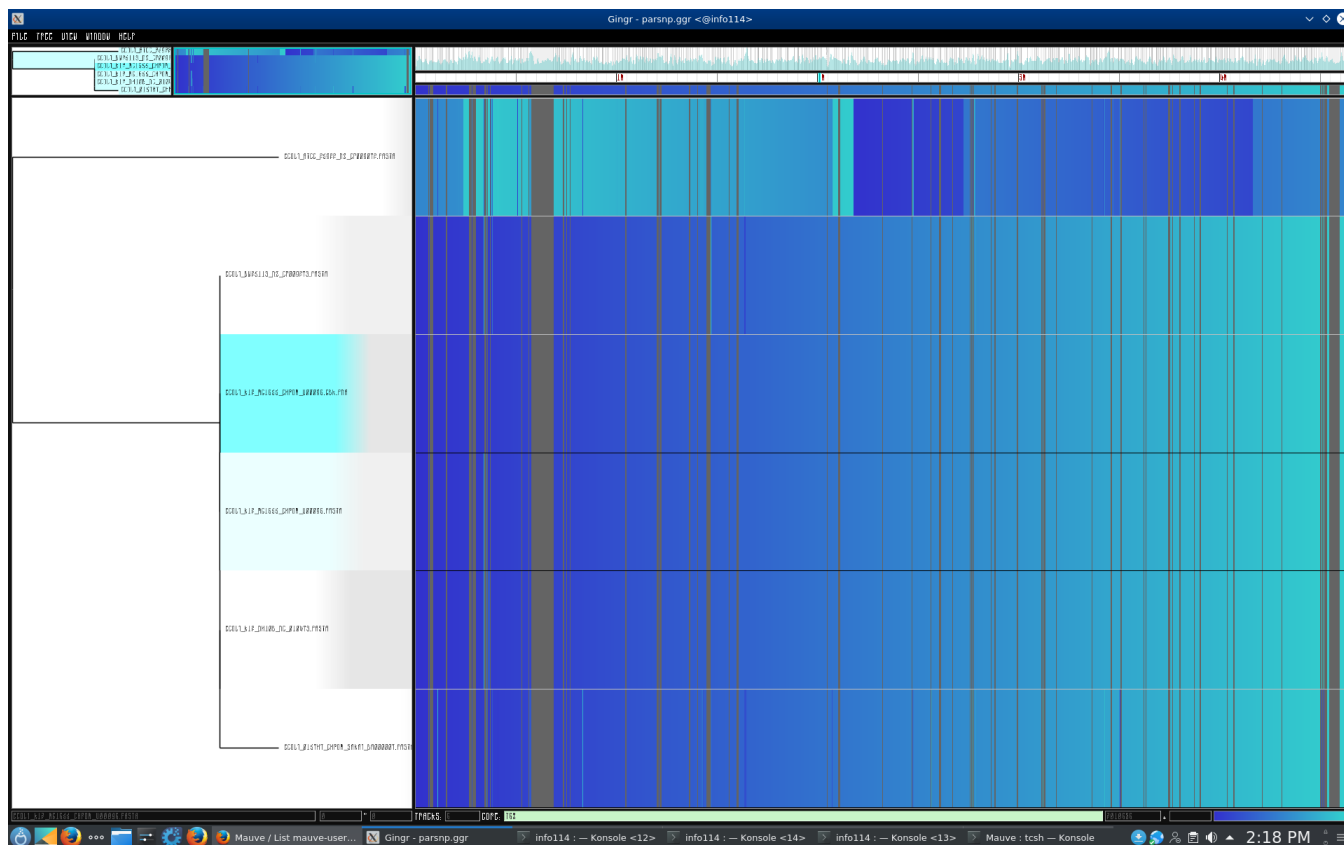
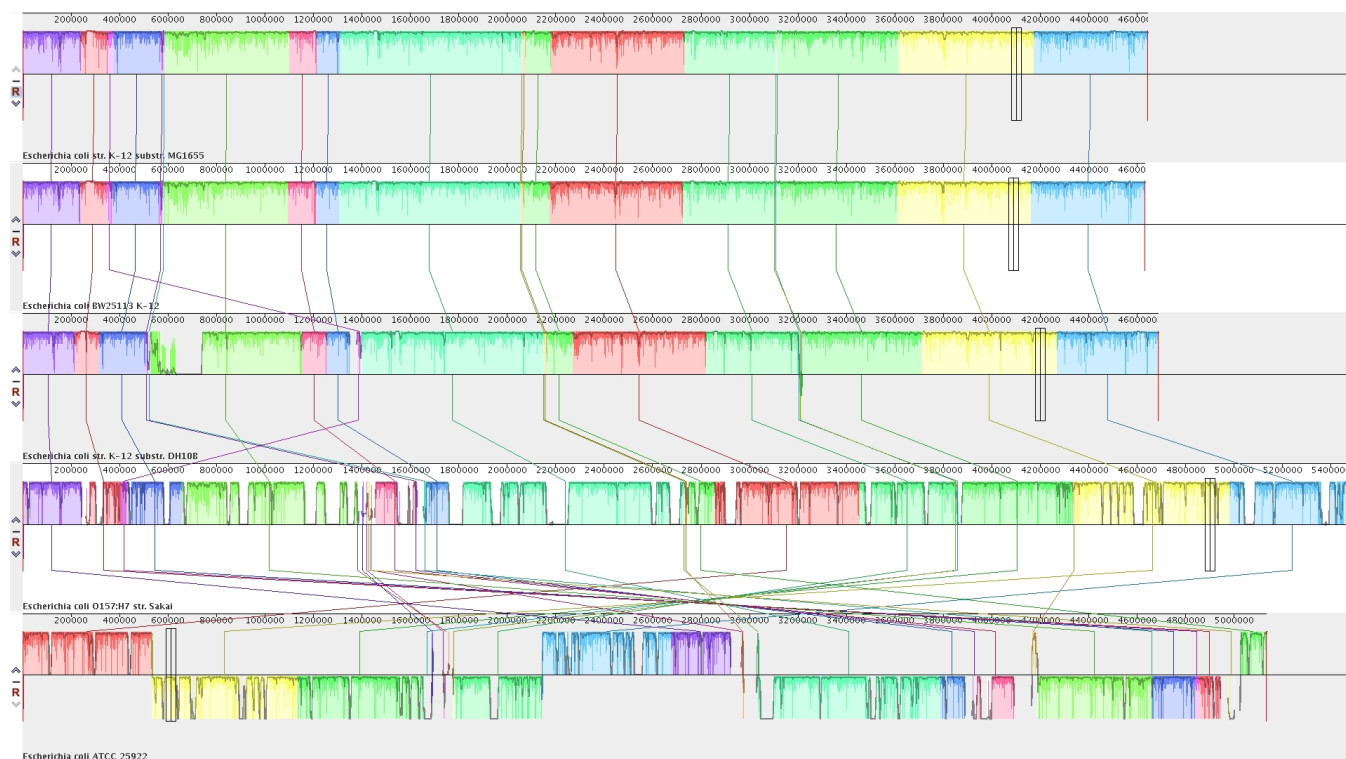


Streptomyces Chromosome









Bacteria Strain/Species	GEO Accession Number	Date Accessed
<i>E. coli</i> K12 MG1655	GSE60522	December 20, 2017
<i>E. coli</i> K12 MG1655	GSE73673	December 19, 2017
<i>E. coli</i> K12 MG1655	GSE85914	December 19, 2017
<i>E. coli</i> K12 MG1655	GSE40313	November 21, 2018
<i>E. coli</i> K12 MG1655	GSE114917	November 22, 2018
<i>E. coli</i> K12 MG1655	GSE54199	November 26, 2018
<i>E. coli</i> K12 DH10B	GSE98890	December 19, 2017
<i>E. coli</i> BW25113	GSE73673	December 19, 2017
<i>E. coli</i> BW25113	GSE85914	December 19, 2017
<i>E. coli</i> O157:H7	GSE46120	August 28, 2018
<i>E. coli</i> ATCC 25922	GSE94978	November 23, 2018
<i>B. subtilis</i> 168	GSE104816	December 14, 2017
<i>B. subtilis</i> 168	GSE67058	December 16, 2017
<i>B. subtilis</i> 168	GSE93894	December 15, 2017
<i>B. subtilis</i> 168	GSE80786	November 16, 2018
<i>S. coelicolor</i> A3	GSE57268	March 16, 2018
<i>S. natalensis</i> HW-2	GSE112559	November 15, 2018
<i>S. meliloti</i> 1021 Chromosome	GSE69880	December 12, 2017
<i>S. meliloti</i> 2011 pSymA	NC_020527 (Dr. Finan)	April 4, 2018
<i>S. meliloti</i> 1021 pSymA	GSE69880	November 15, 18
<i>S. meliloti</i> 2011 pSymB	NC_020560 (Dr. Finan)	April 4, 2018
<i>S. meliloti</i> 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
<i>E. coli</i> Chromosome	-6.03×10^{-5}	1.28×10^{-5}	2.8×10^{-6}
<i>B. subtilis</i> Chromosome	-9.7×10^{-5}	2.0×10^{-5}	1.2×10^{-6}
<i>Streptomyces</i> Chromosome	-1.17×10^{-6}	1.04×10^{-7}	$< 2 \times 10^{-16}$
<i>S. meliloti</i> Chromosome	3.97×10^{-5}	4.25×10^{-5}	NS (3.5×10^{-1})
<i>S. meliloti</i> pSymA	1.39×10^{-3}	2.53×10^{-4}	4.9×10^{-8}
<i>S. meliloti</i> pSymB	1.46×10^{-4}	2.03×10^{-4}	NS ($5.34.7 \times 10^{-1}$)

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.