

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~~
- ~~grab genes from each gbk file~~
- ~~align ↑ with a codon-aware aligner~~
- ~~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

- ✓re-run alignment classification code on re-annotated PROKKA *ecoli* genomes
- ✓applied for SMBE (Manchester) and O3EC (McMaster) conferences
- ✓figured out the “missing bars” in the gene expression data
- ✓finished corrections/comments on substitution paper

Gene Expression: The reason that there appear to be “missing bars” in the histograms for the expression data is because for some of the genes, one sample did not have an expression value for that particular gene. So, all data from that gene (for all samples) was tossed. I thought about just putting a zero expression value for these missing genes, but as we discussed this is really not proper because it is like faking data.

Alignment Column Classification: The re-annotation with PROKKA did not fix the issue with *E. coli*, where most of the sites in the alignment are NOT the same codon classification between all taxa. I thought maybe this had something to do with progressiveMauve picking the first genome to be the “reference” genome, and for *E. coli* it is not K-12 that is the reference, but as we discussed, this should not matter because the program should be picking up rearrangements no matter what the reference is.

I have also finished the corrections that you gave me for my substitutions paper.

## This Week

Look VERY closely at the *Escherichia coli* alignment and translate the proteins that it says there is a gene there and see what the heck is going on and if it makes sense.

I plan on having a detailed outline for the gene expression paper finished so I can start writing it soon.

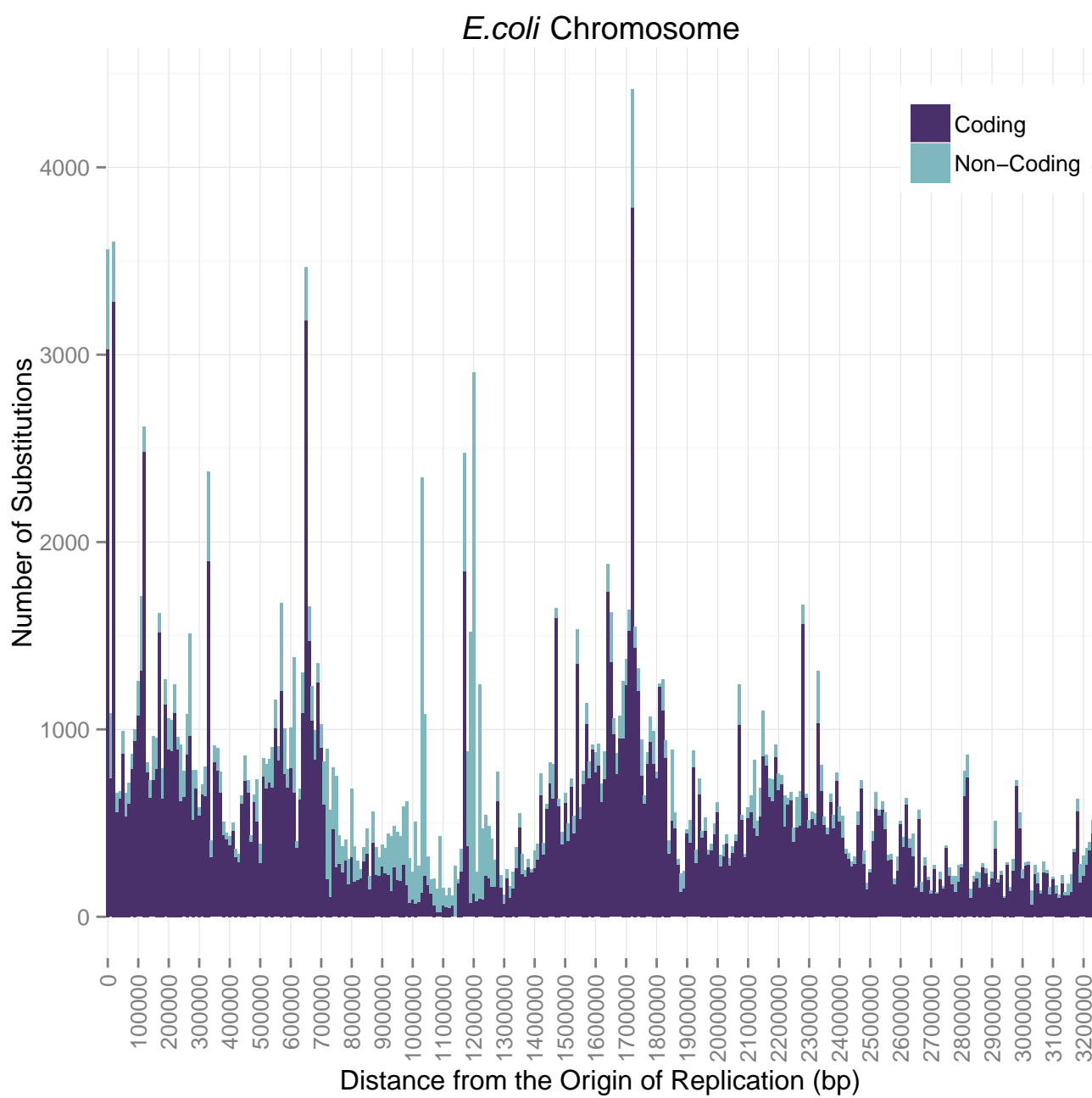
## Next Week

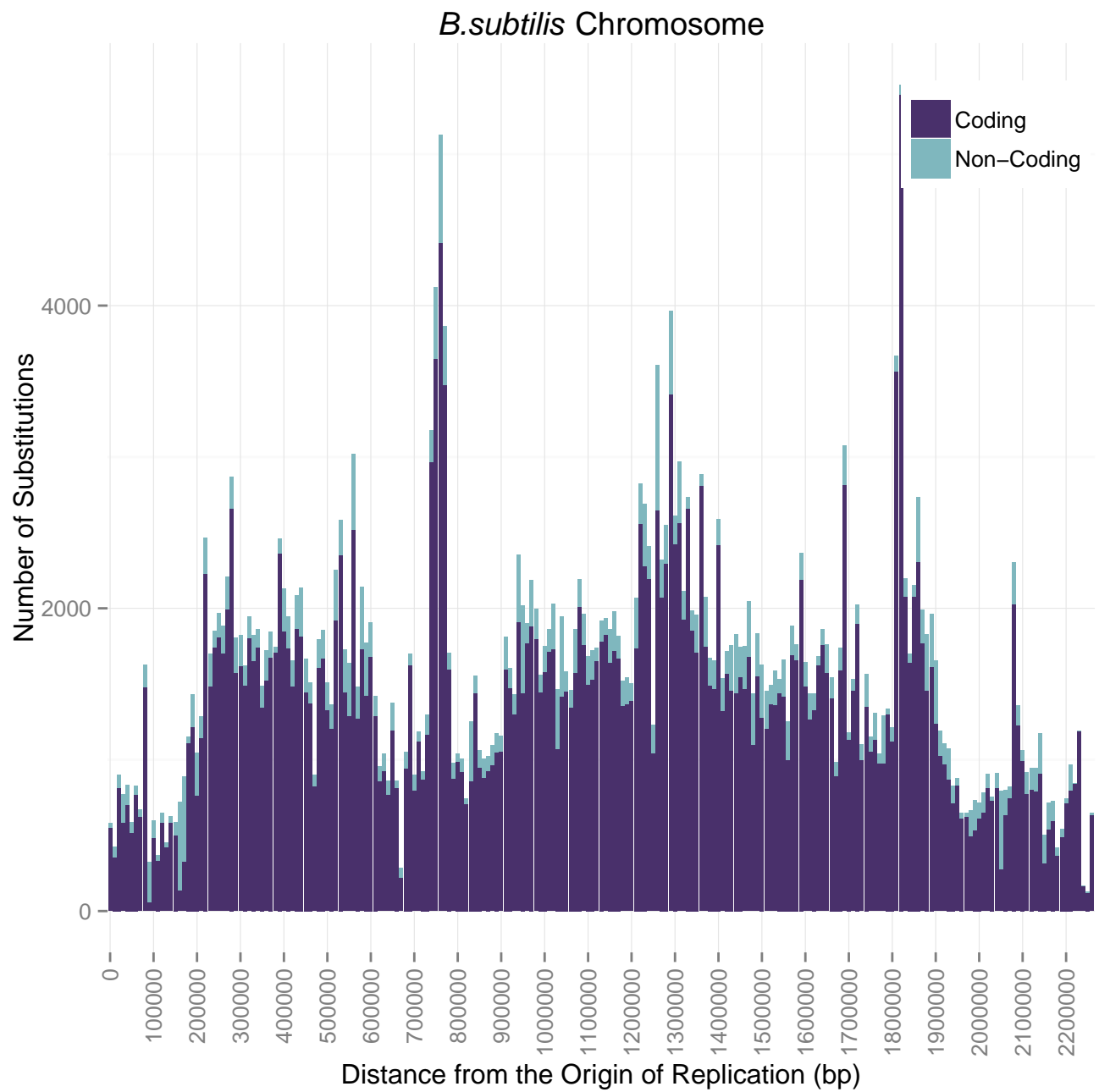
I would like to implement some sort of plan for how to deal with sites in the alignment that do not compare the same nucleotide classifications. Then I can hopefully re-run the coding and non-coding analysis and any PRANK alignments that have to happen.

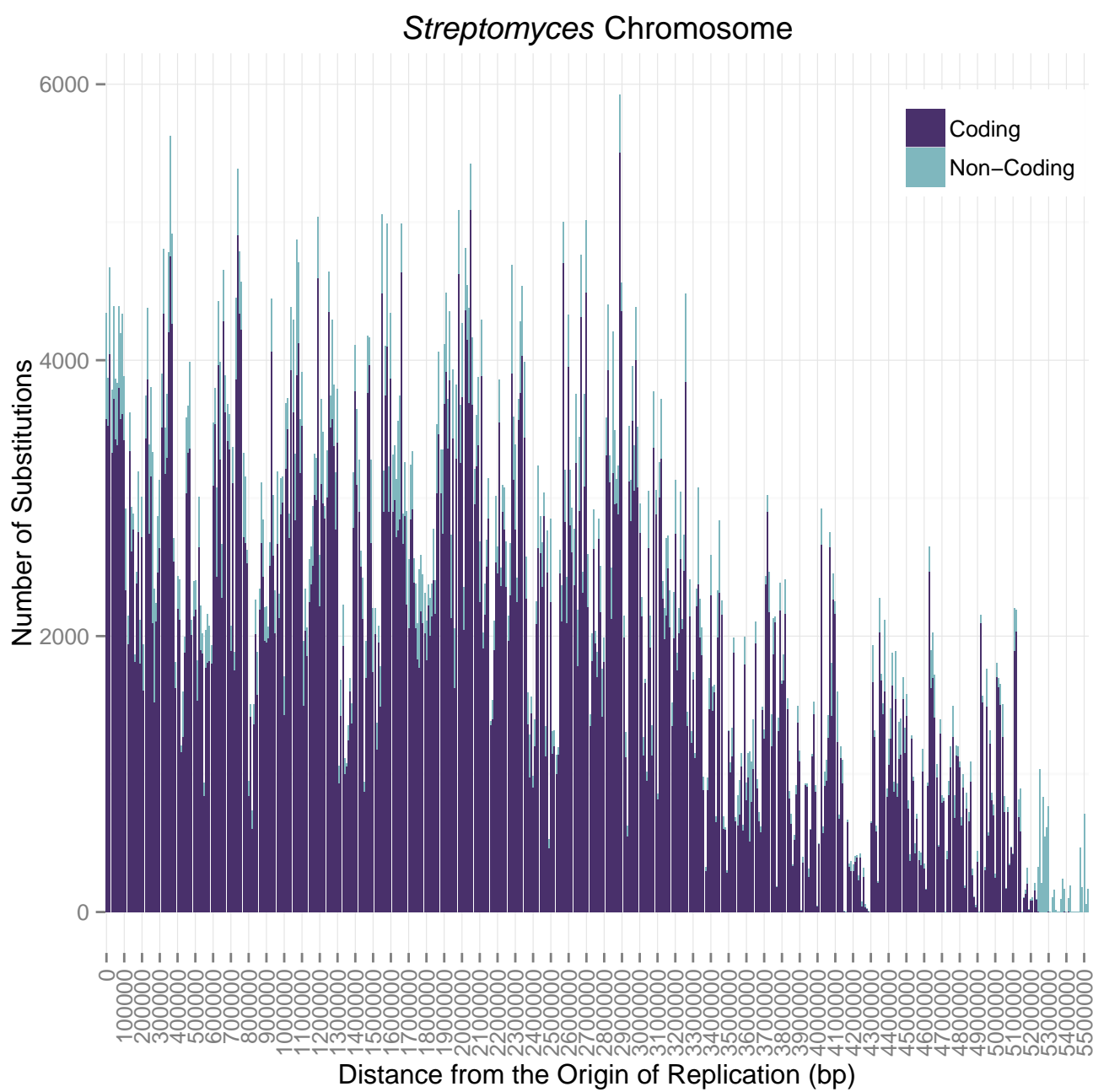
I want to begin figuring out how to obtain all inversions from the Mauve or PARSNP alignment for the inversions analysis.

Classification	Number of Nucleotides in Alignment	% of Total Alignment
Gapped	78450	31%
Not same class	138002	55%
Same class	35852	15%
Codon 1	113	<0.05%
Codon 2	113	<0.05%
Codon 3	113	<0.05%
Misc 4	0	0
Non-coding	35513	15%

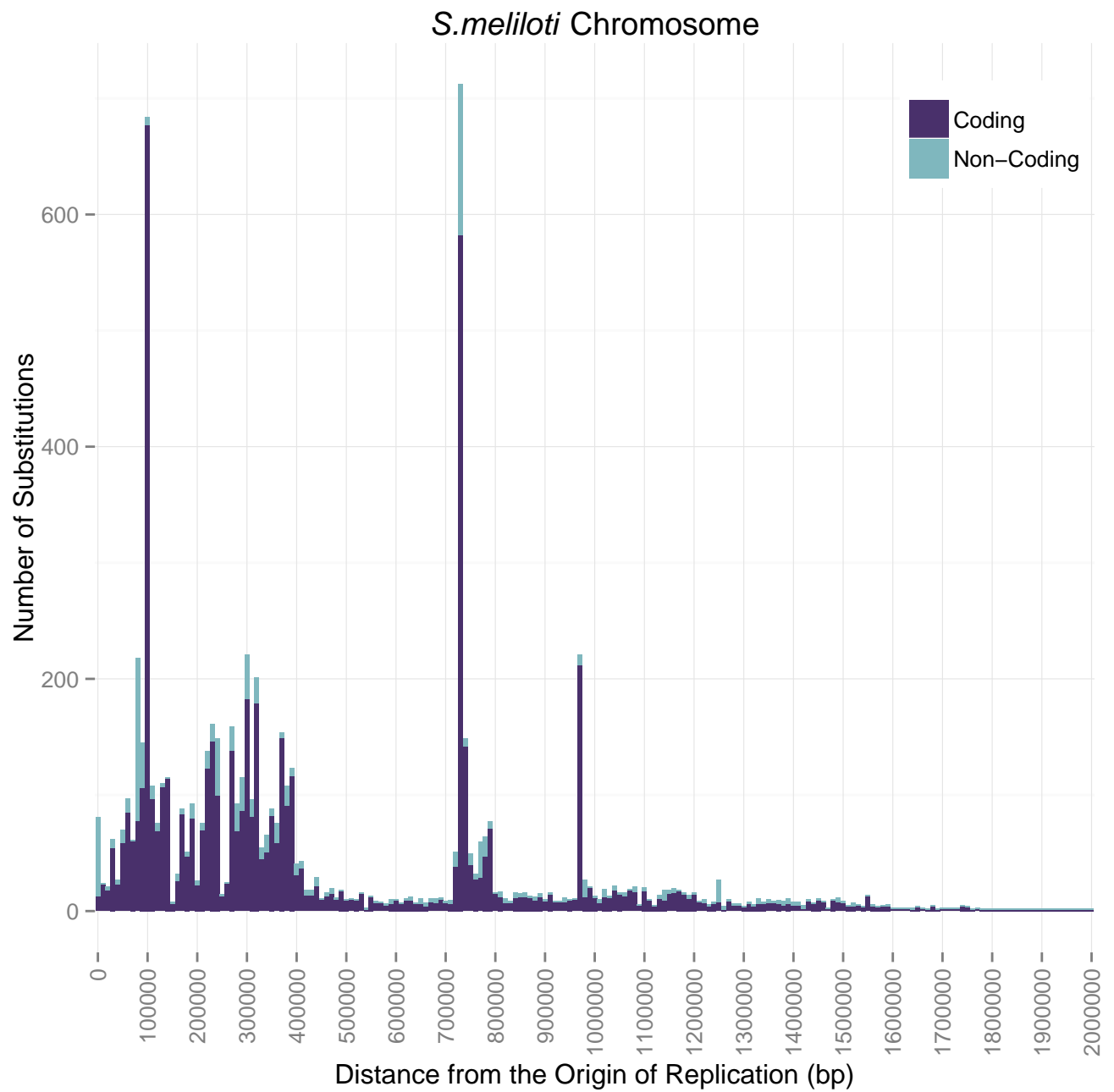
Table 1: Classification of each column in the alignment of ONE sample *E. coli* block with a total alignment length of 252304. The percentages are calculated based on the WHOLE alignment length. The "same class" classification is any column where all taxa had the same classification. "Not same class" is any column where at least one taxa did not have the same classification as the rest of the taxa in that column. The "Same class" category can be further broken down into the different nucleotide classifications to show how often those were found to be the same between all taxa in a column. The "Gapped" category denotes a site where at least one taxa had a gap present.

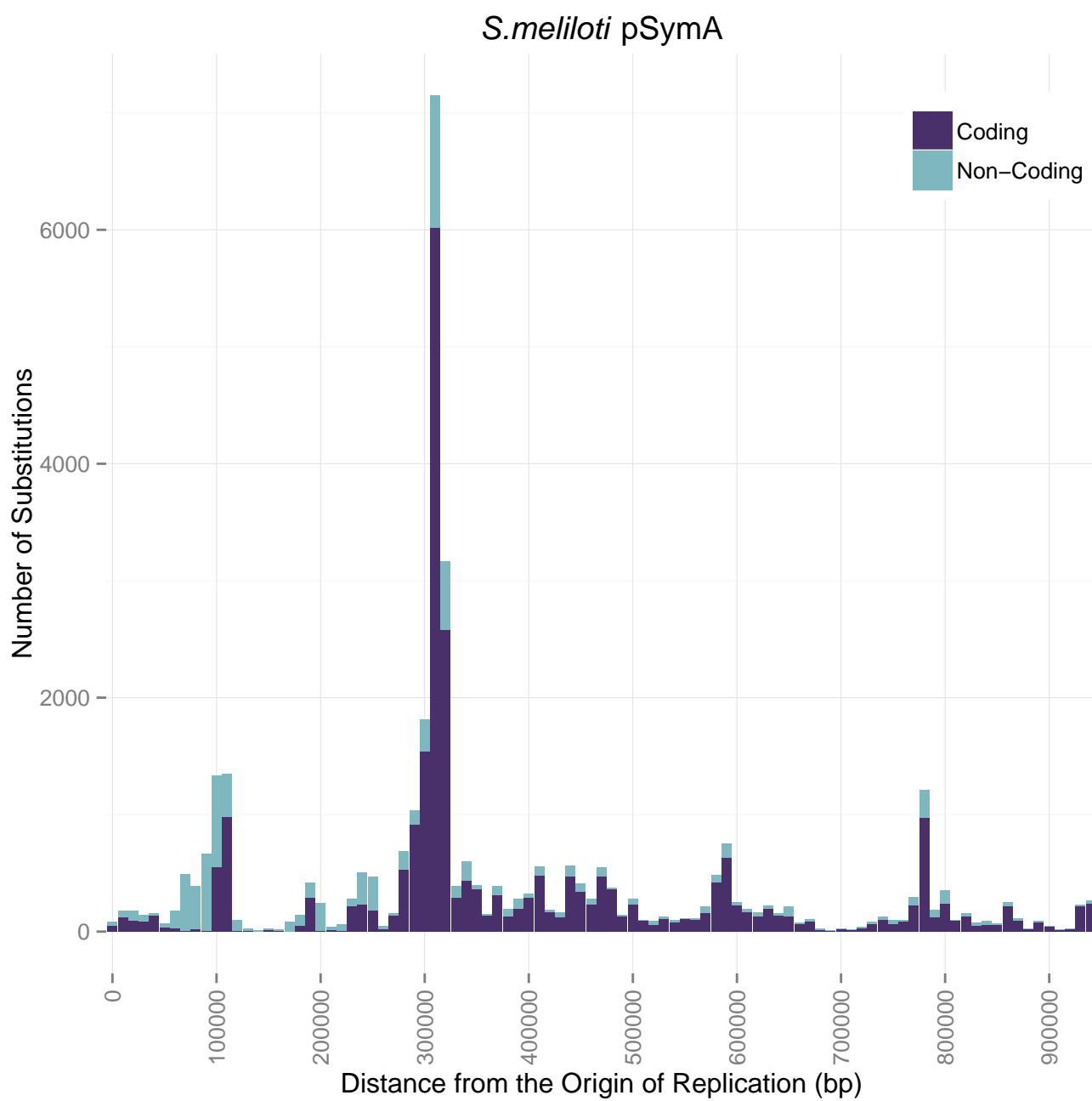


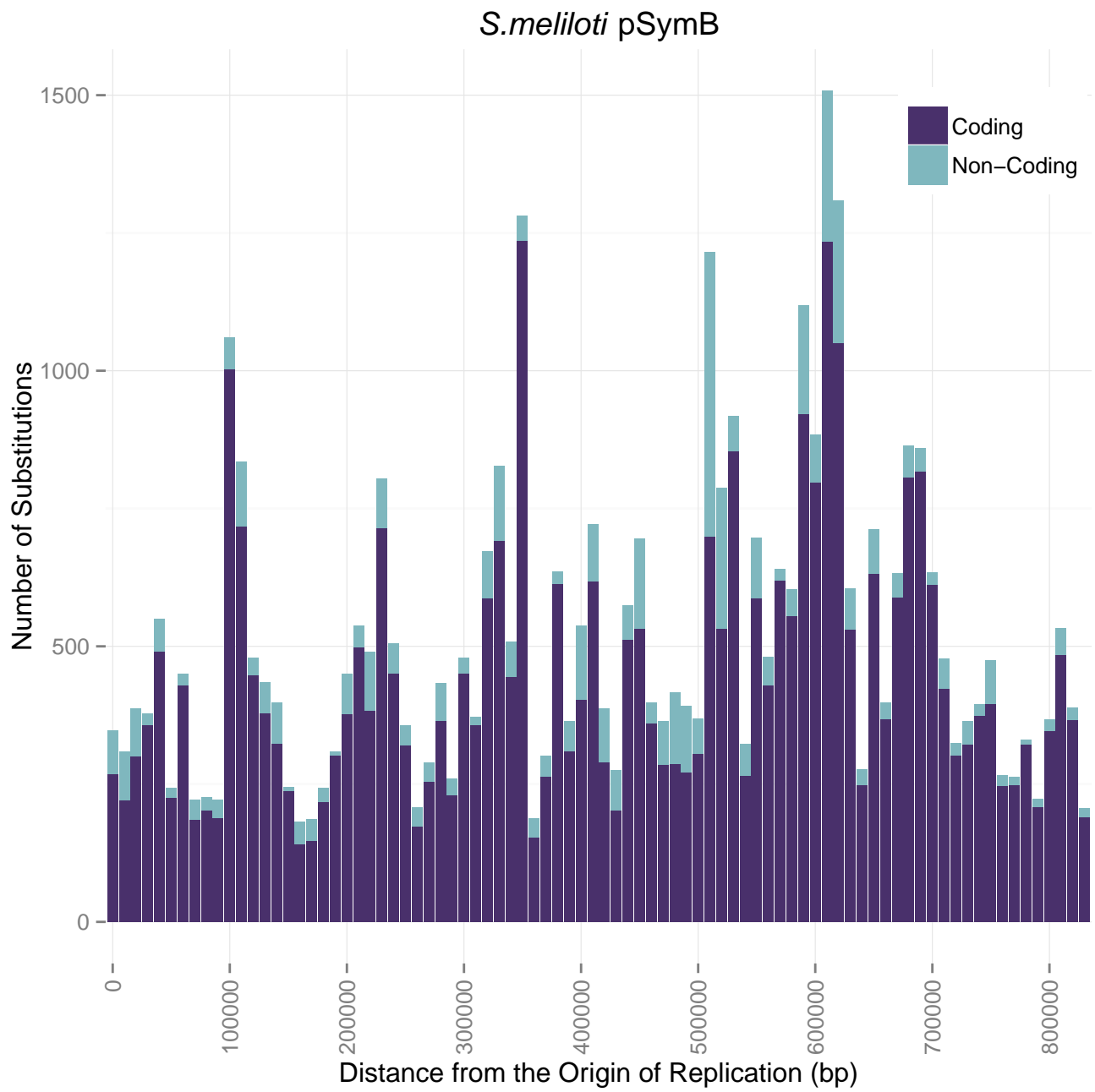


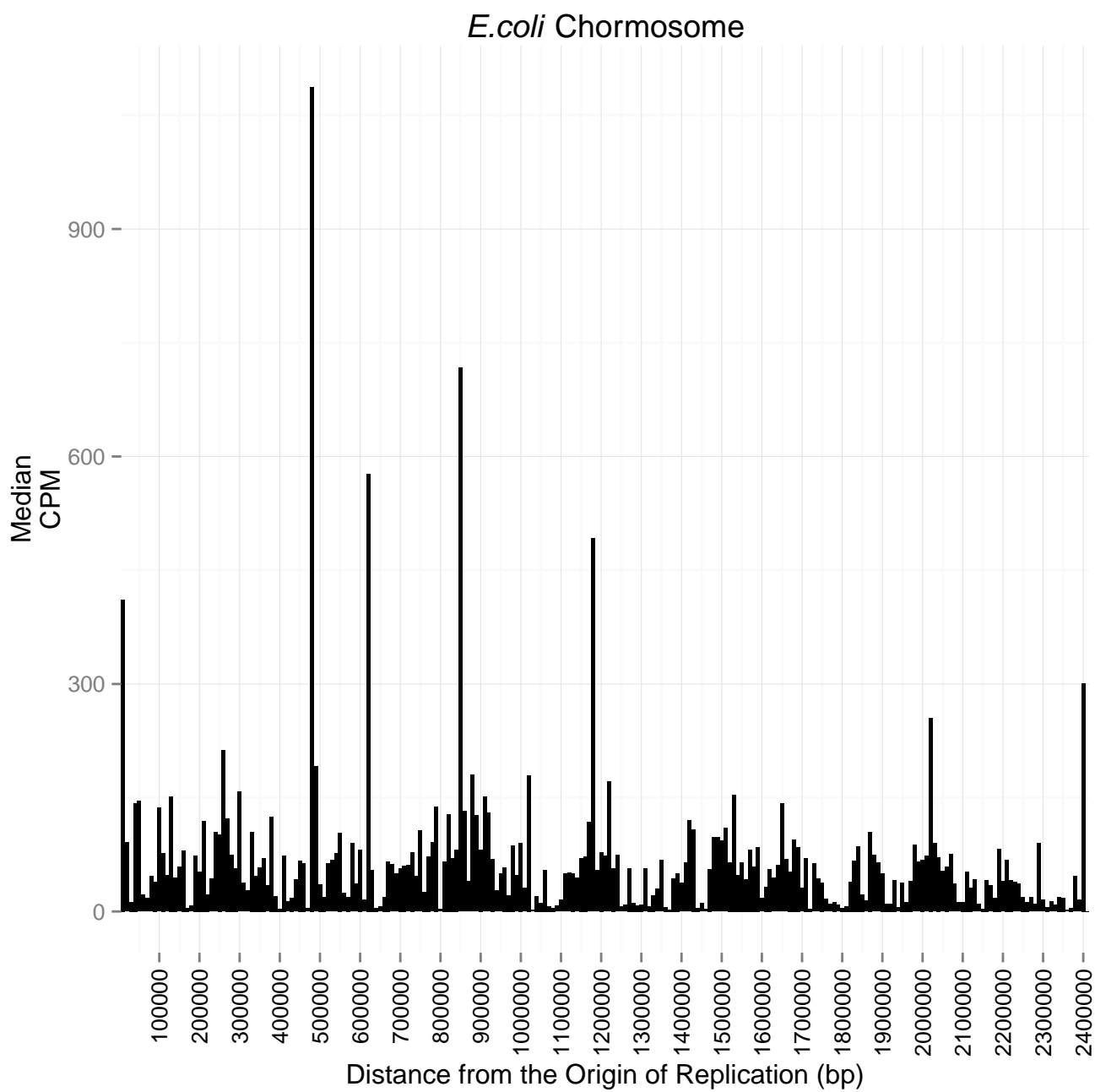


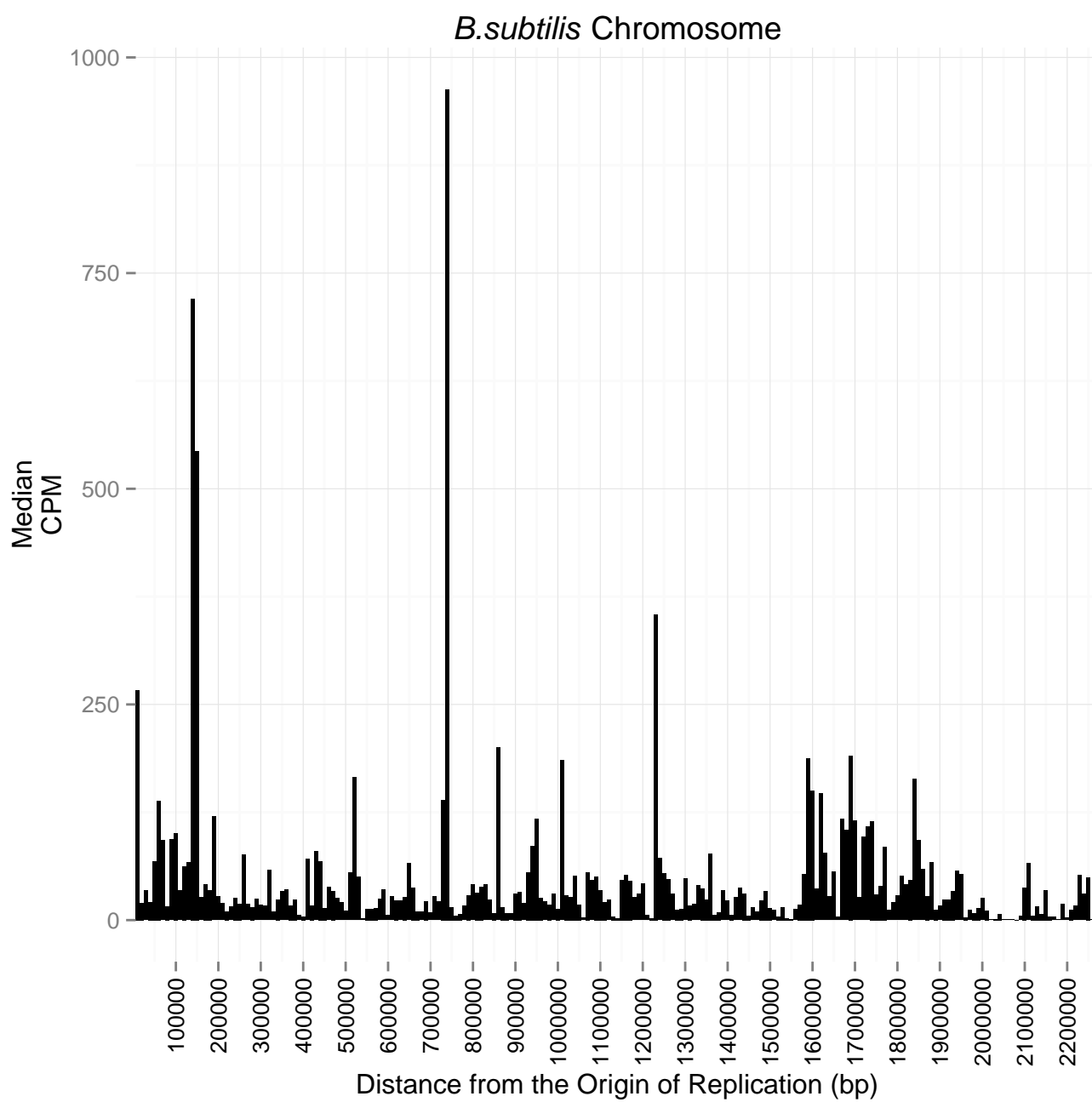


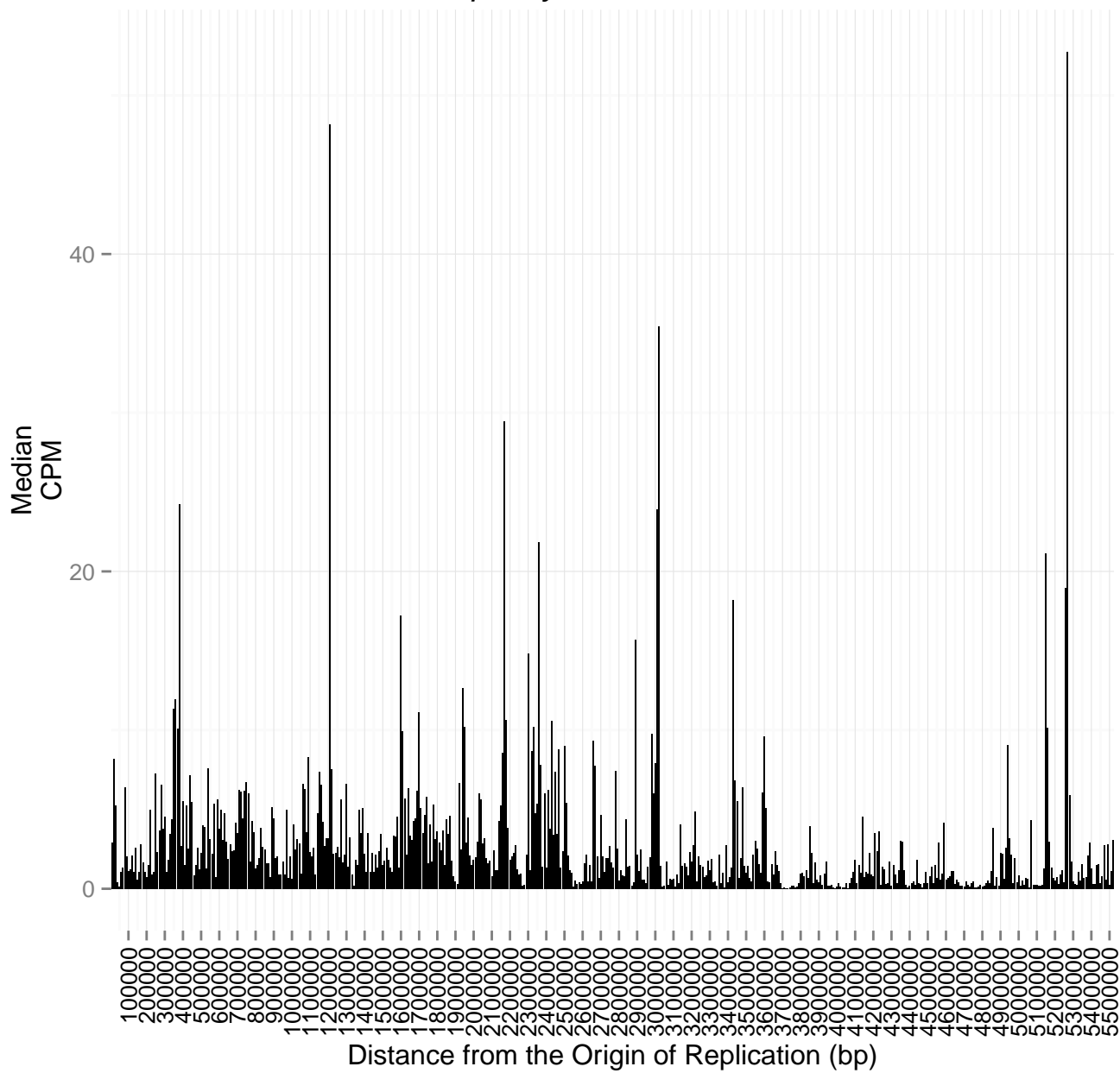


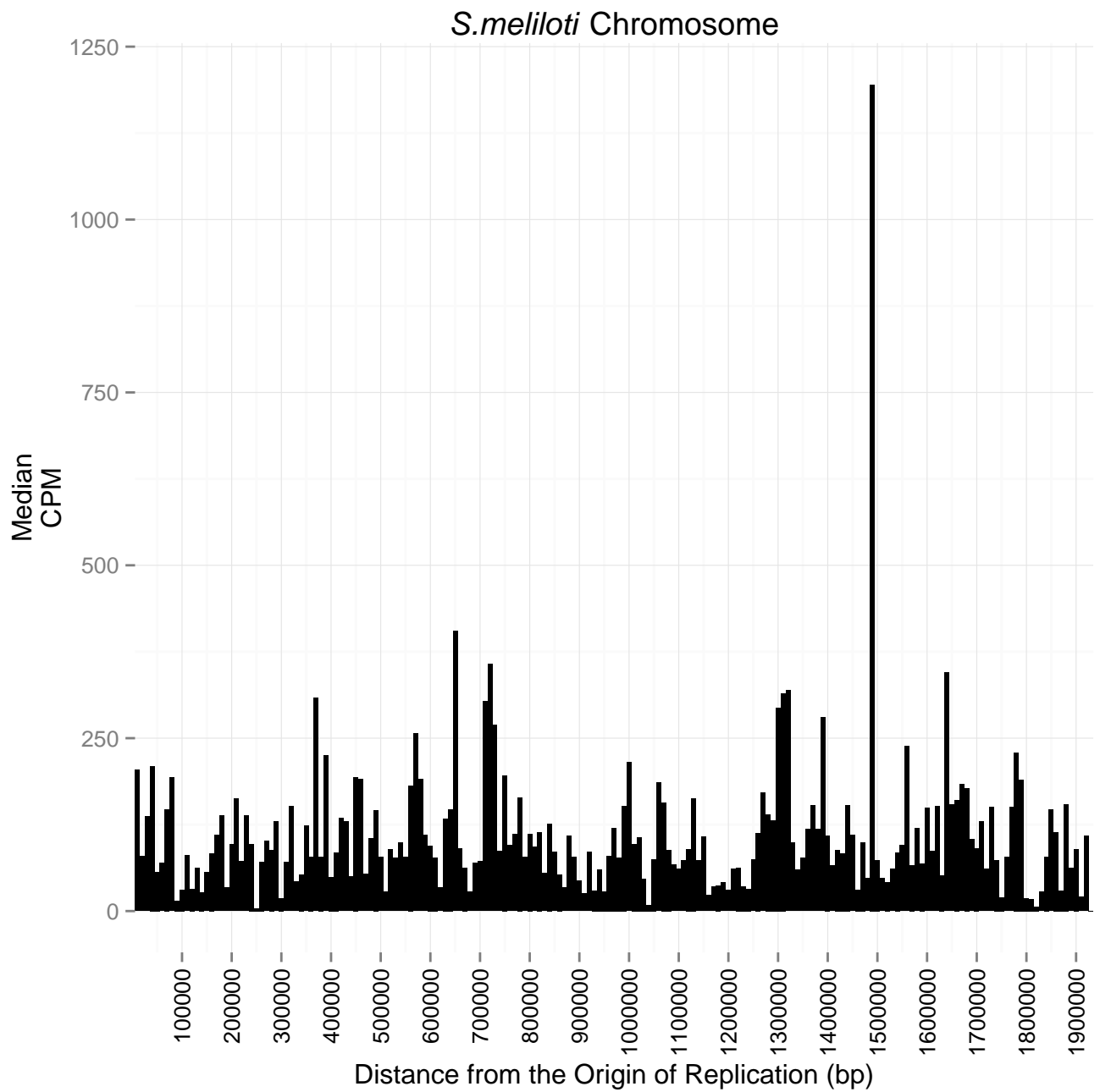


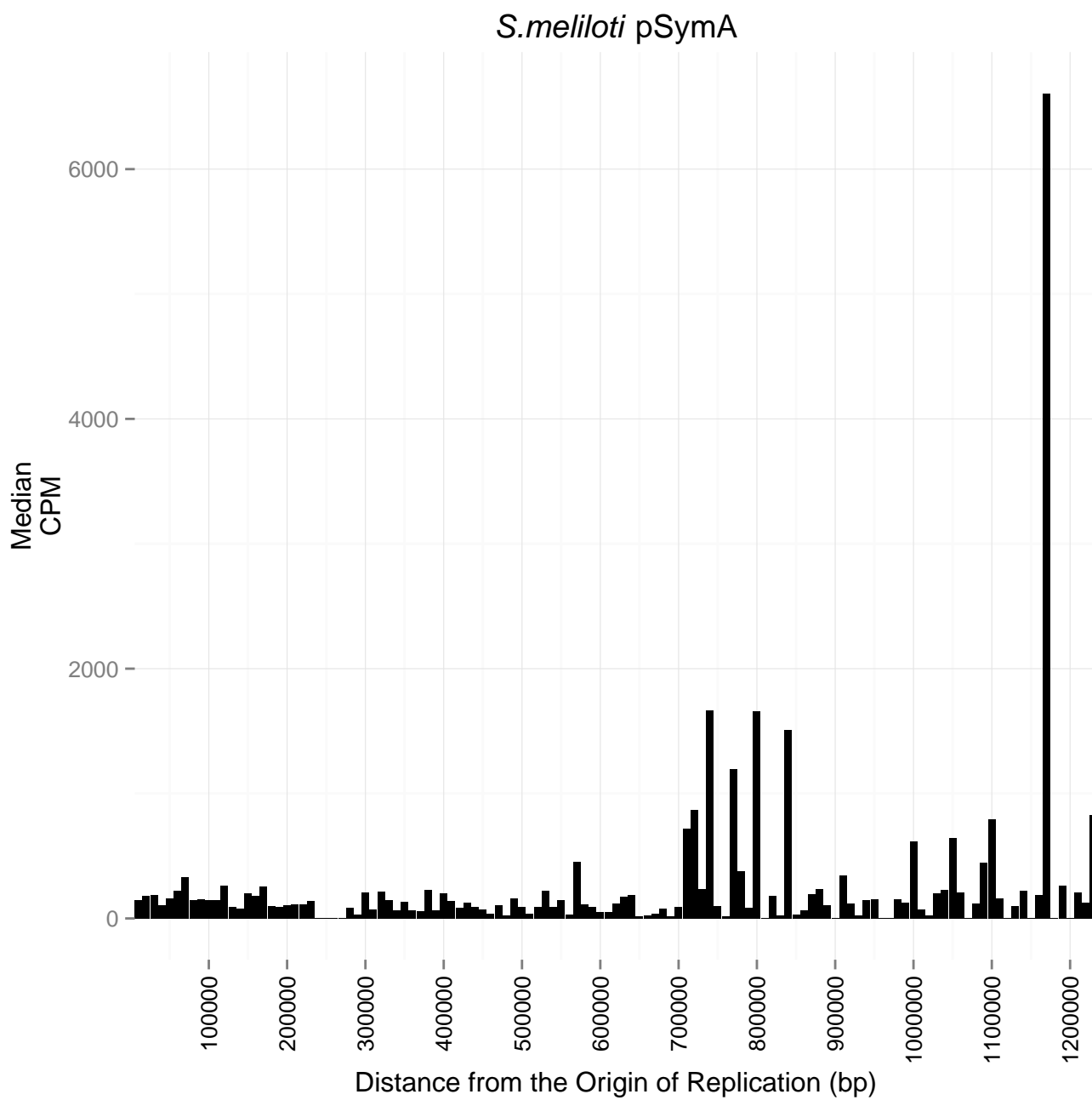




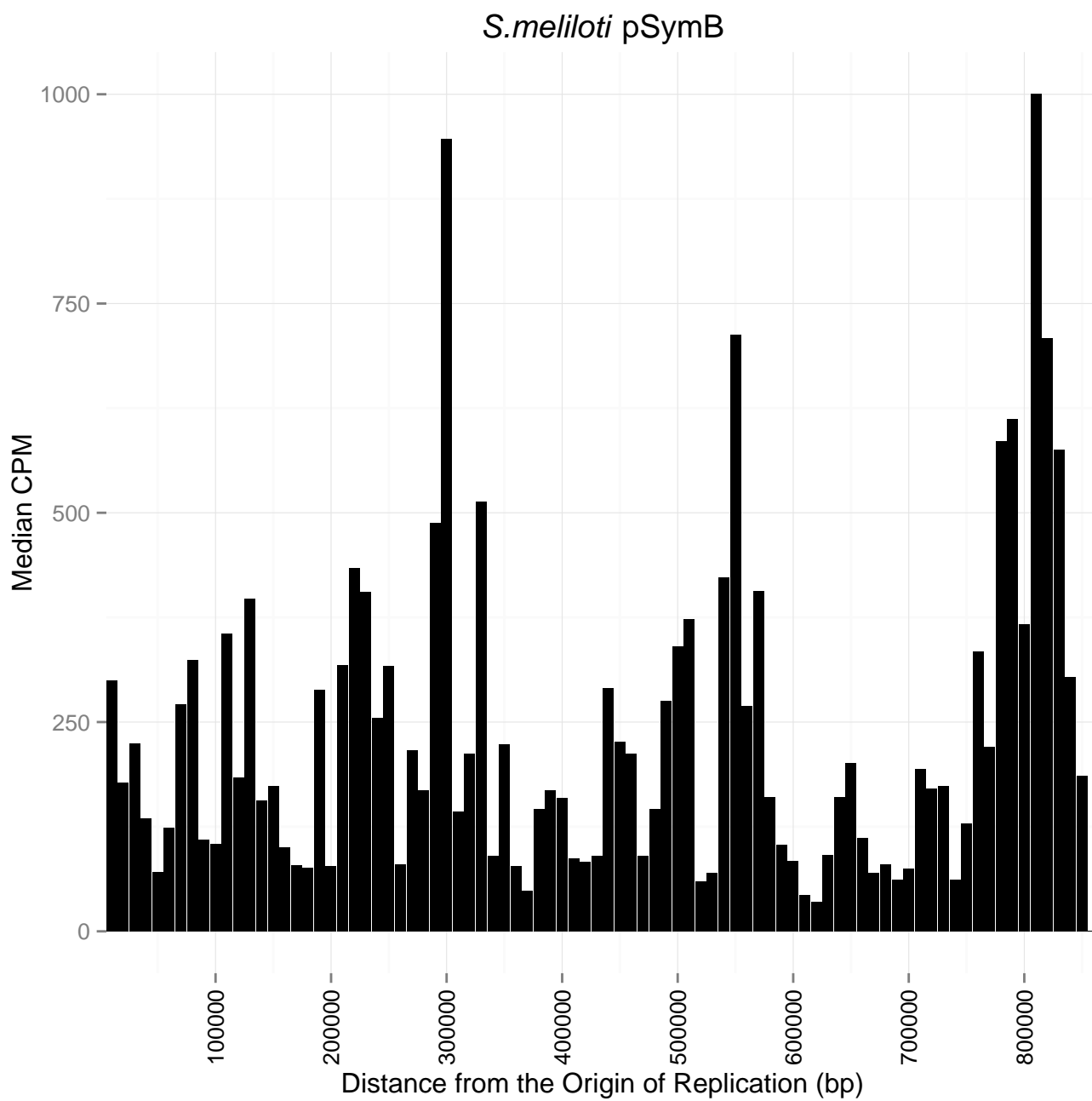


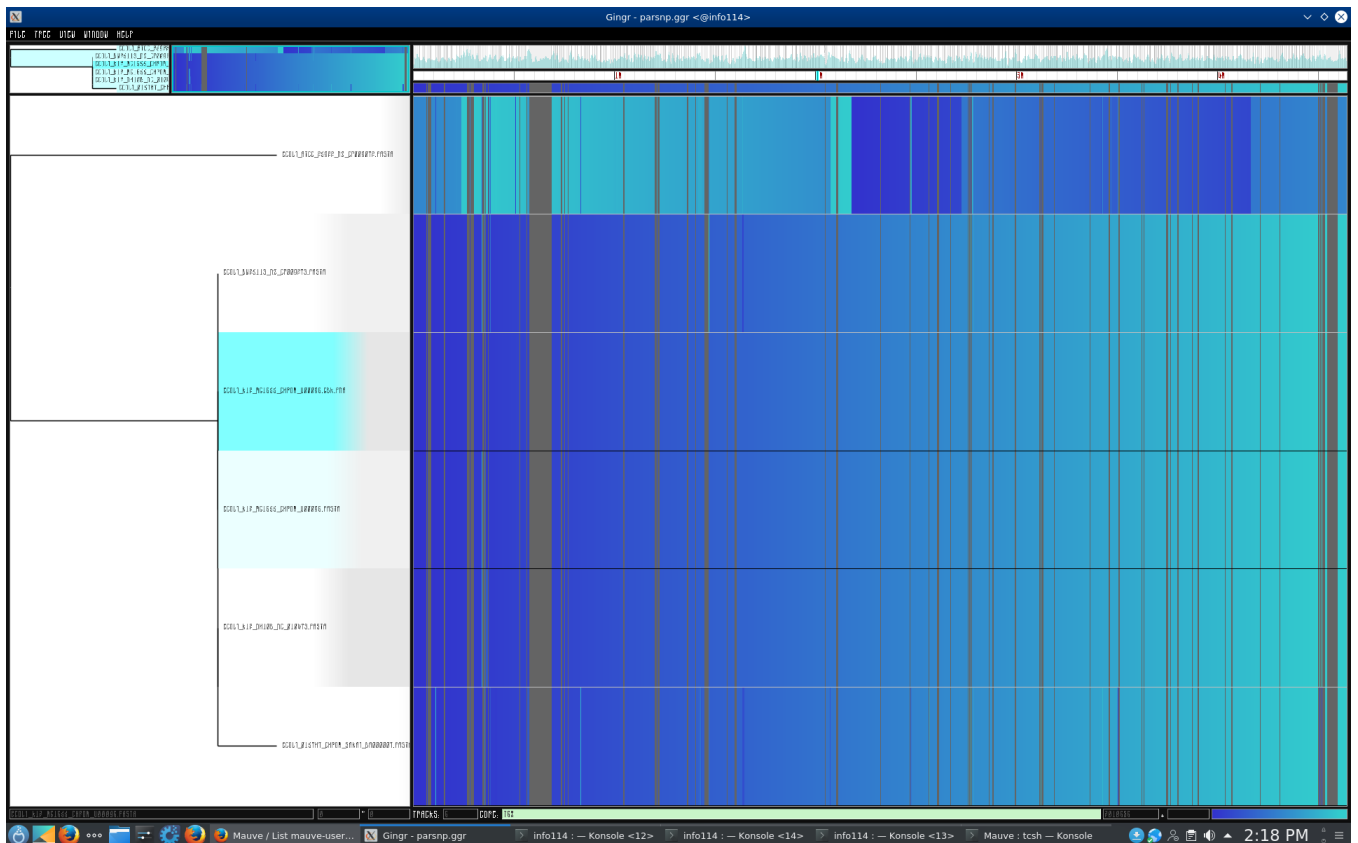
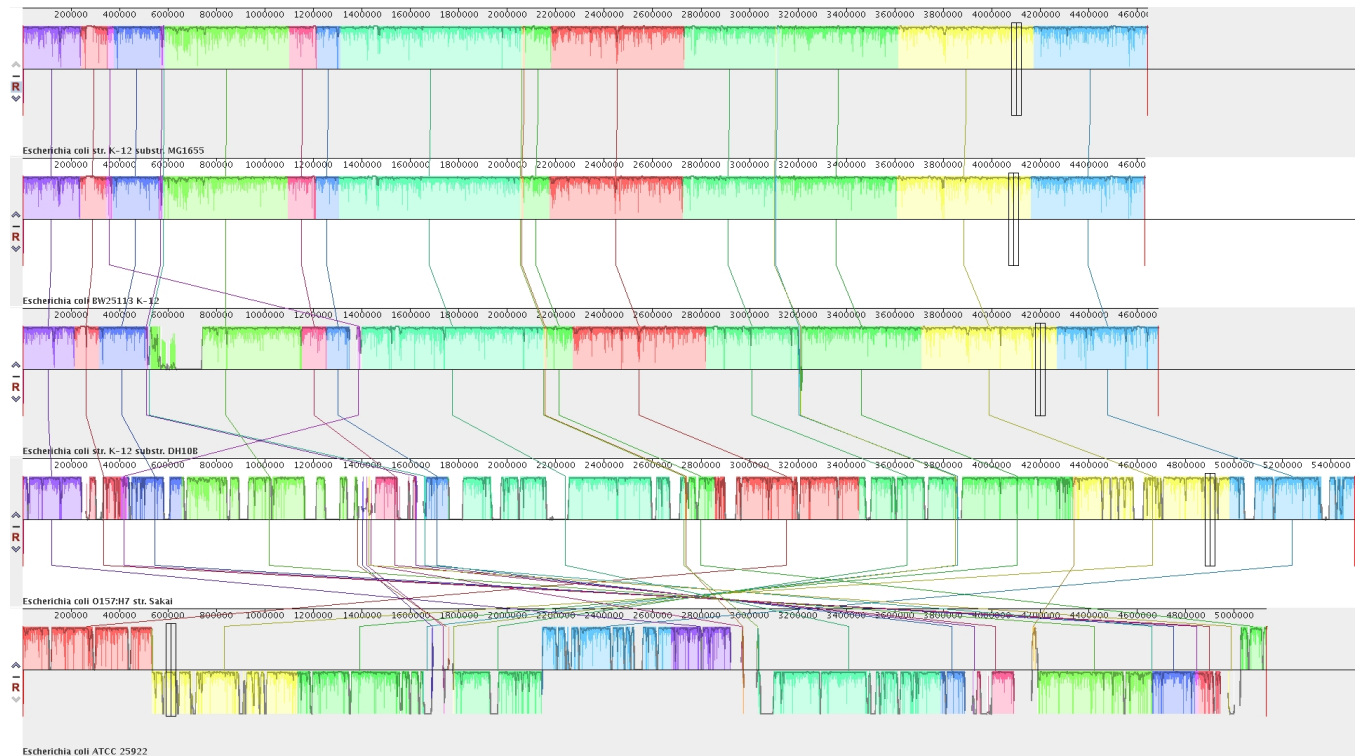
*Streptomyces* Chromosome











Bacteria and Replicon	% of Coding Sequences	% of Non-Coding Sequences	% of Subs Coding	% of Subs Non-Coding
<i>E. coli</i> Chromosome	86.47%	13.53%	5.00%	8.96%
<i>B. subtilis</i> Chromosome	87.49%	12.51%	7.31%	6.42%
<i>Streptomyces</i> Chromosome	89.03%	10.97%	13.74%	14.91%
<i>S. meliloti</i> Chromosome	86.27%	13.73%	0.19%	0.22%
<i>S. meliloti</i> pSymA	83.34%	16.66%	2.84%	4.58%
<i>S. meliloti</i> pSymB	88.81%	11.19%	2.78%	3.44%

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).

Bacteria and Replicon	Coding Sequences	Non-Coding Sequences
<i>E. coli</i> Chromosome	$-5.938 \times 10^{-8***}$	$-9.237 \times 10^{-8***}$
<i>B. subtilis</i> Chromosome	$-7.584 \times 10^{-8***}$	NS
<i>Streptomyces</i> Chromosome	$5.483 \times 10^{-7***}$	$9.182 \times 10^{-9***}$
<i>S. meliloti</i> Chromosome	$-1.448 \times 10^{-6***}$	$-7.037 \times 10^{-7***}$
<i>S. meliloti</i> pSymA	$-9.704 \times 10^{-7***}$	$-1.464 \times 10^{-7***}$
<i>S. meliloti</i> pSymB	$5.007 \times 10^{-7***}$	NS

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:  $< 0.001 = '***'$ ,  $0.001 < 0.01 = '**'$ ,  $0.01 < 0.05 = '*'$ ,  $> 0.05 = 'NS'$ .

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	$-6.03 \times 10^{-5}$	$1.28 \times 10^{-5}$	$2.8 \times 10^{-6}$
<i>B. subtilis</i> Chromosome	$-9.7 \times 10^{-5}$	$2.0 \times 10^{-5}$	$1.2 \times 10^{-6}$
<i>Streptomyces</i> Chromosome	$-1.17 \times 10^{-6}$	$1.04 \times 10^{-7}$	$< 2 \times 10^{-16}$
<i>S. meliloti</i> Chromosome	$3.97 \times 10^{-5}$	$4.25 \times 10^{-5}$	NS ( $3.5 \times 10^{-1}$ )
<i>S. meliloti</i> pSymA	$1.39 \times 10^{-3}$	$2.53 \times 10^{-4}$	$4.9 \times 10^{-8}$
<i>S. meliloti</i> pSymB	$1.46 \times 10^{-4}$	$2.03 \times 10^{-4}$	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.