

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

Last week I was mostly busy with my competition but I got a few things done. Before the break we determined that if you have one gene overlapping with 2 genes from another taxa there could be a situation where the nucleotides in one column of an alignment are not the same class. So they may not be the 1st codon position, or 2nd...etc. Since then, I have been working on determining how often this happens in the alignment of a sample block from *Escherichia coli*. I have gotten this information but, it becomes more complicated when there are genes within genes in some of the taxa. This means that some genes overlap WITHIN a taxa and so one particular site in the alignment may be classified as more than one thing (1st, 2nd, 3rd, or non coding). I am still working on how to classify/deal with this issue to get accurate information about each column in the alignment.

Once that is figured out then I can re-do the coding/non-coding part of the analysis.

I also sent you a draft of my substitutions paper before reading week which I would really appreciate your feedback on.

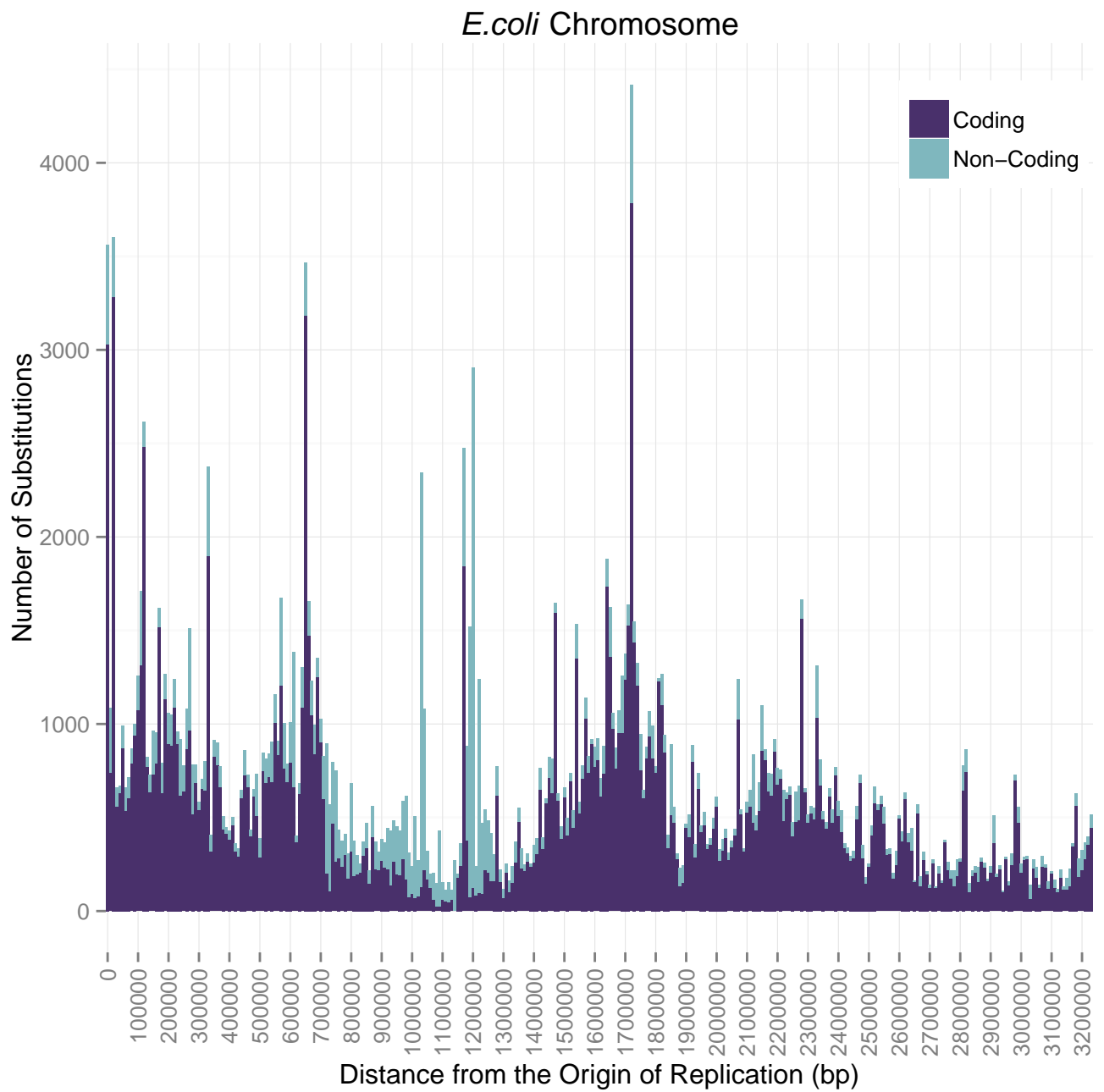
I also read and made notes on quite a few papers last week, gathering more information about gene expression and inversions.

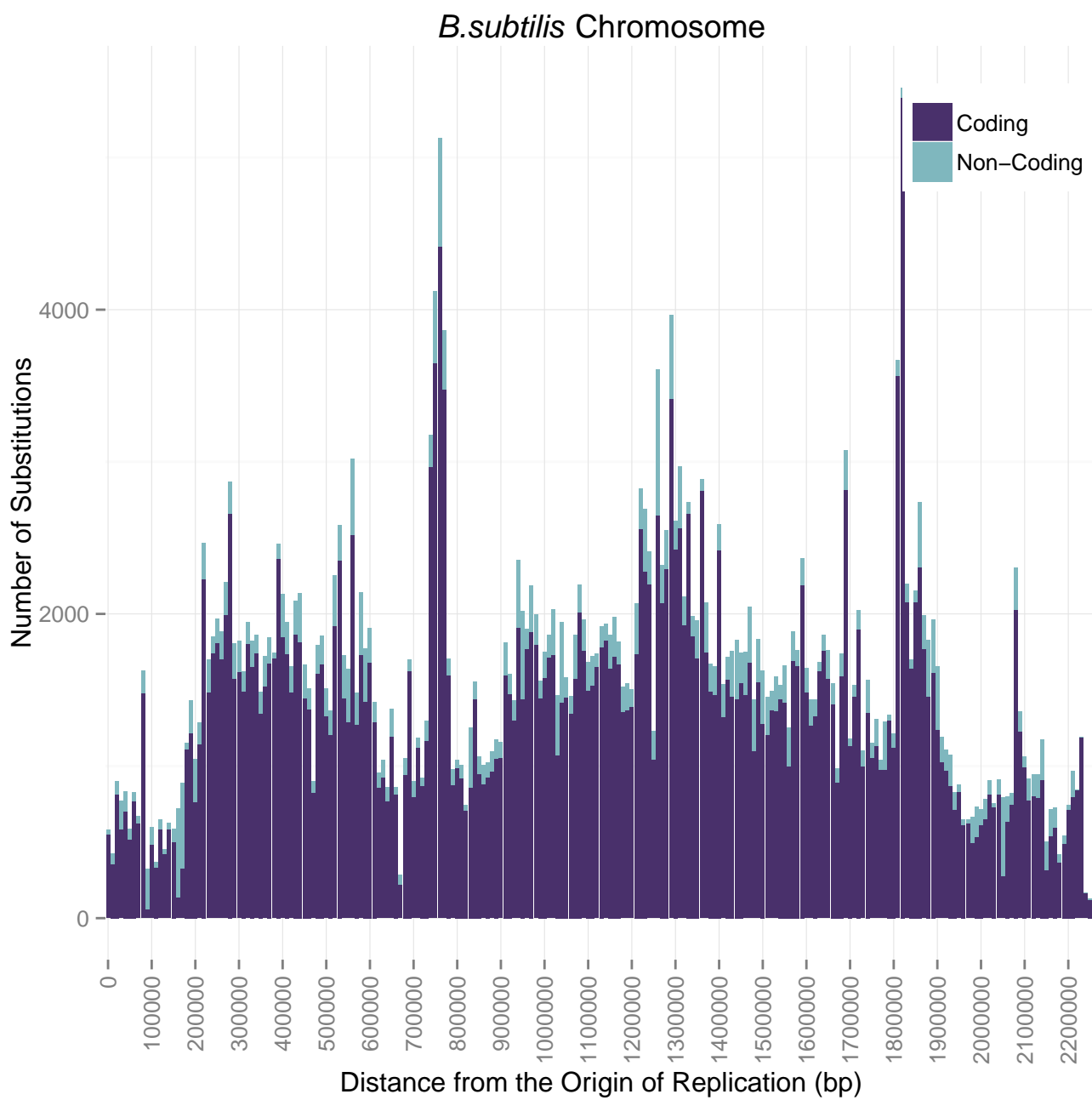
## This Week

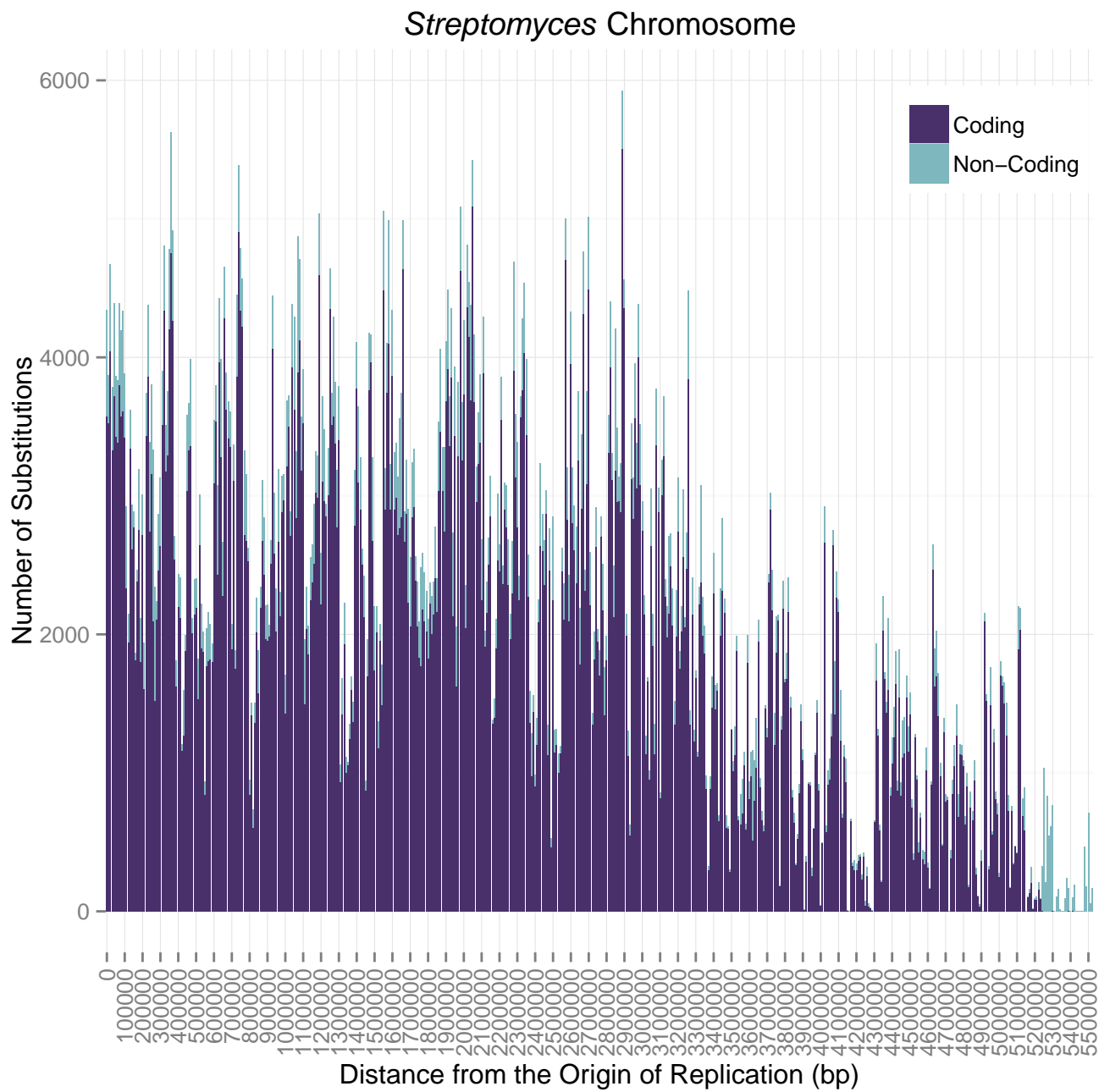
I would like to have the data on how many columns of the alignments are actually comparing the same nucleotide classification (1st, 2nd, 3rd, non-coding) and then make a decision on how to deal with the results from this.

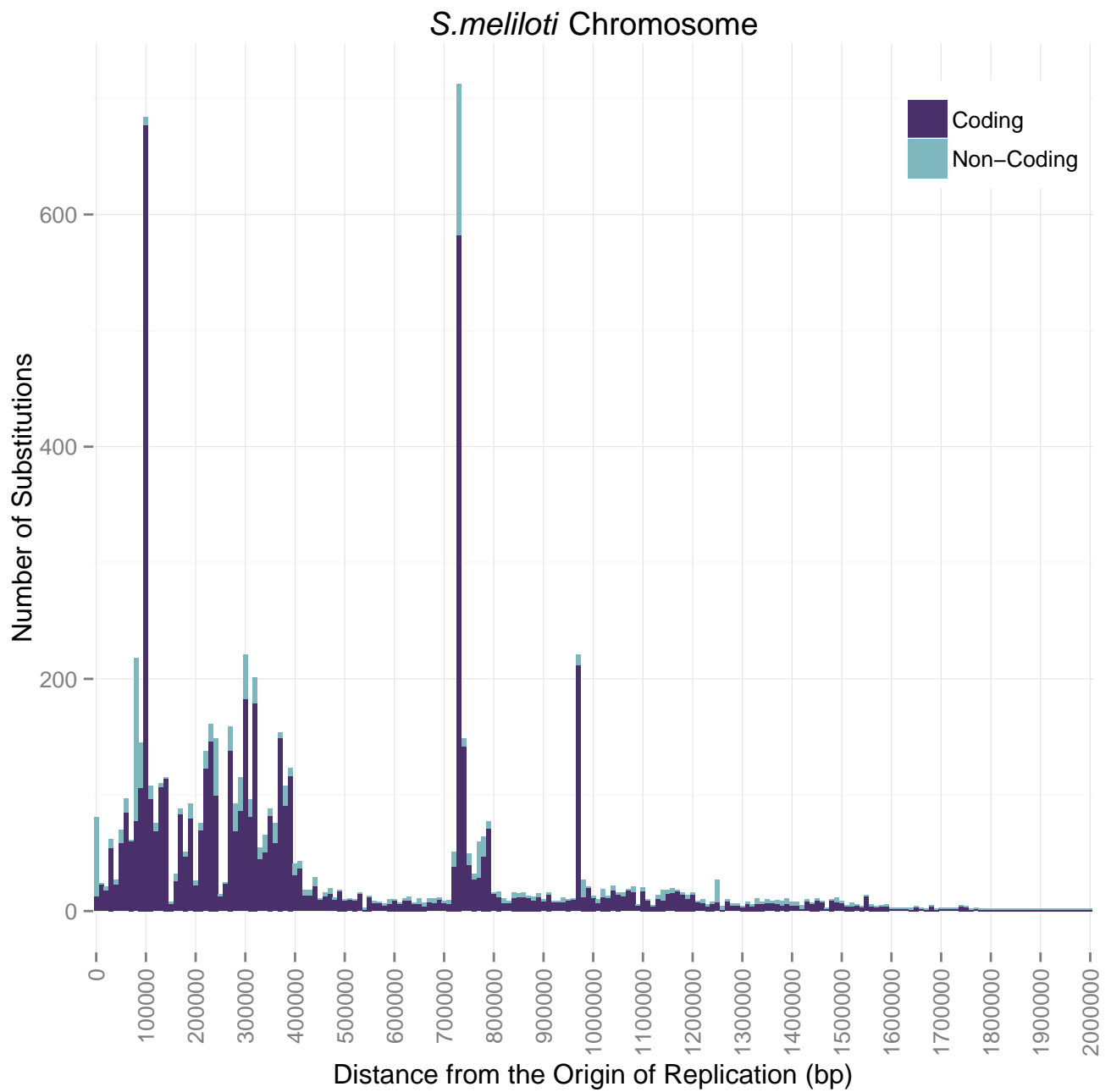
## Next Week

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.

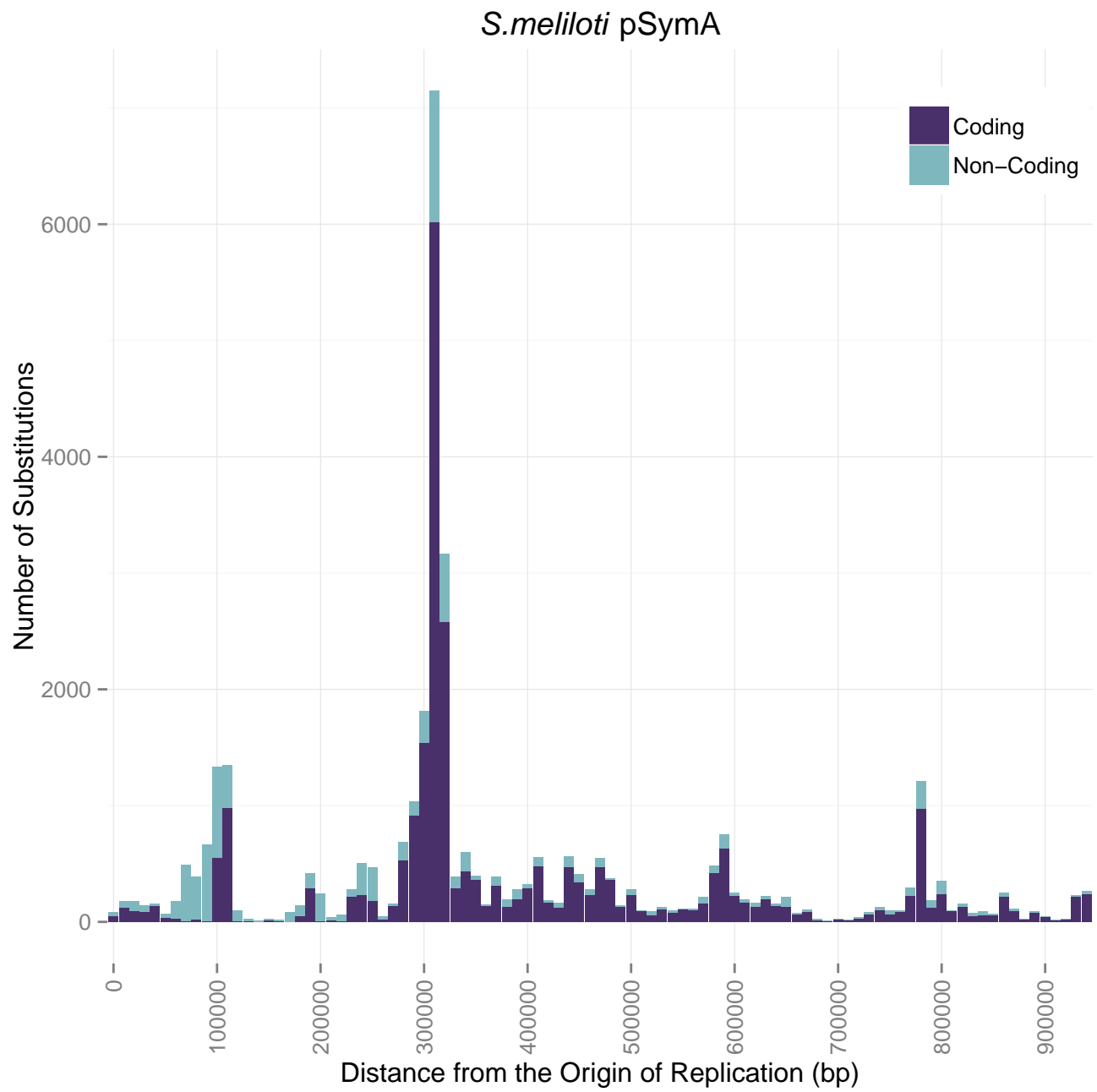


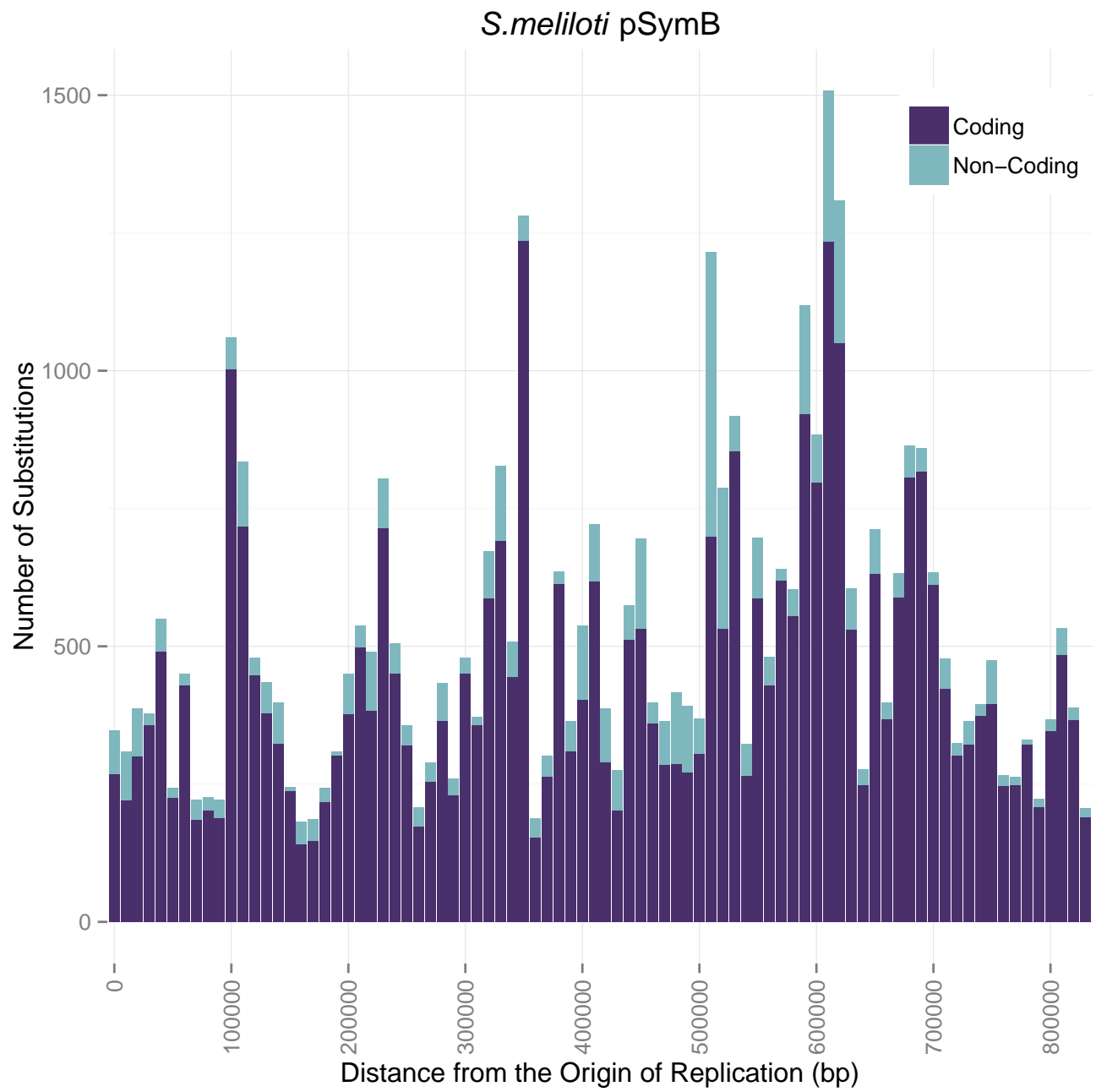


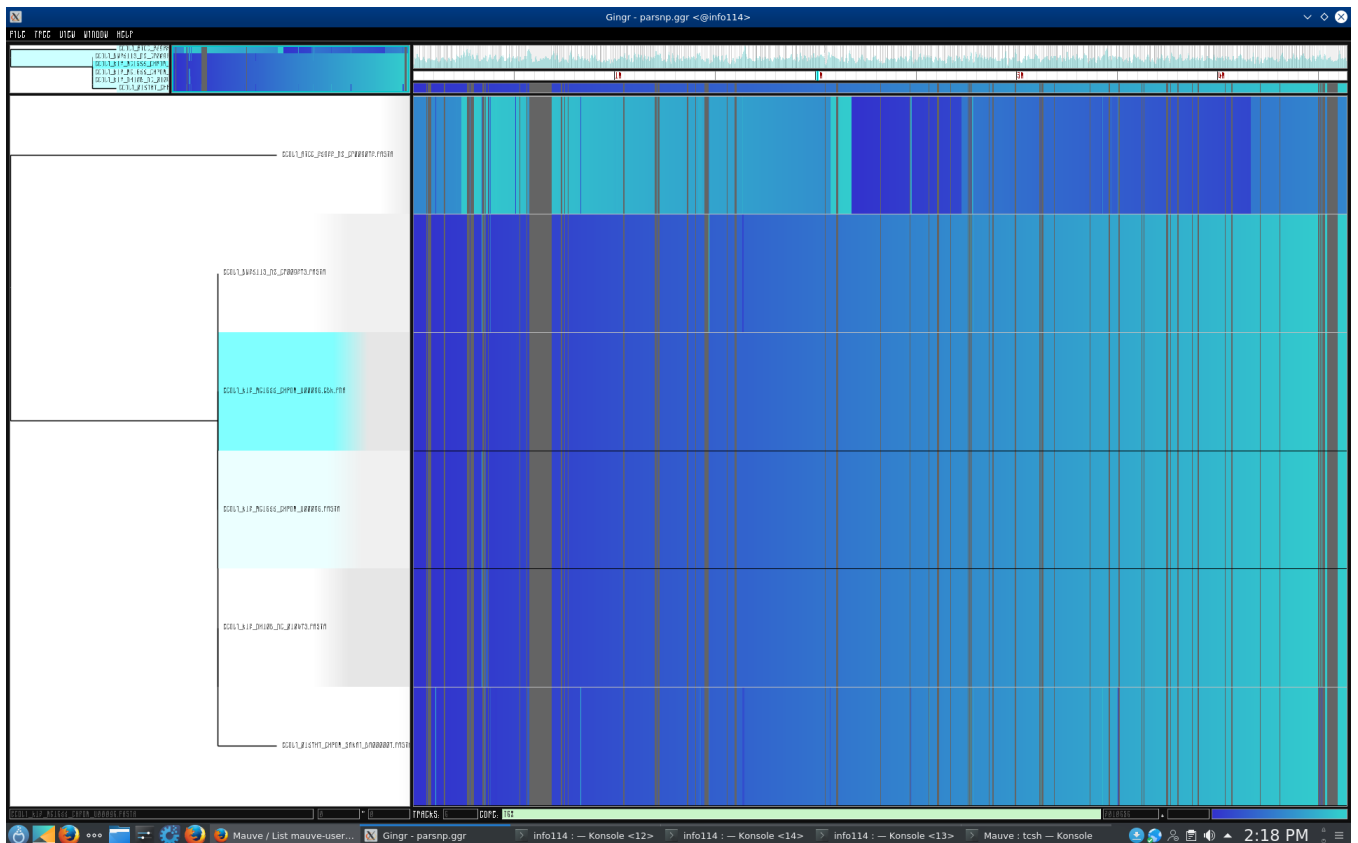
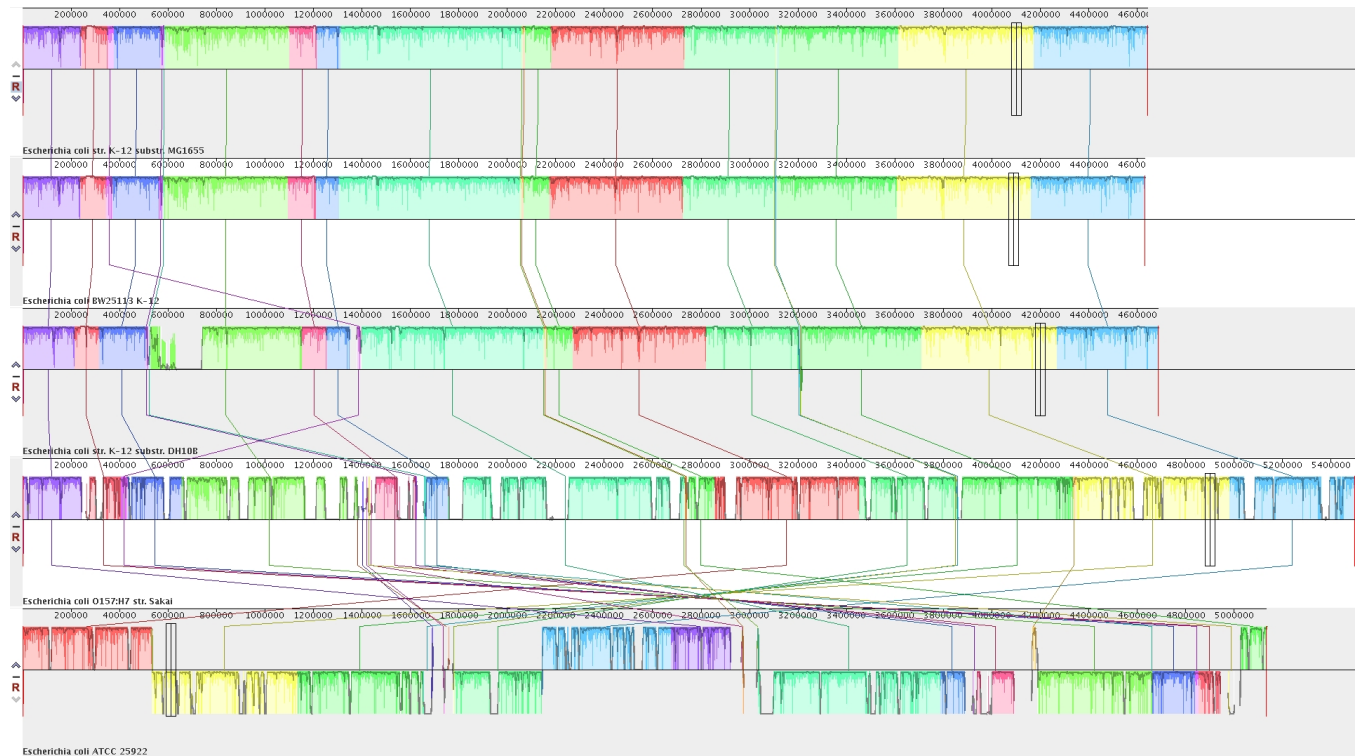












Bacteria and Replicon	% of Coding Sequences	% of Non-Coding Sequences	% of Subs Coding	% of Subs Non-Coding
<i>E. coli</i> Chromosome	86.47%	13.53%	5.00%	8.96%
<i>B. subtilis</i> Chromosome	87.49%	12.51%	7.31%	6.42%
<i>Streptomyces</i> Chromosome	89.03%	10.97%	13.74%	14.91%
<i>S. meliloti</i> Chromosome	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 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
<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 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:  $< 0.001 = '***'$ ,  $0.001 < 0.01 = '**'$ ,  $0.01 < 0.05 = '*'$ ,  $> 0.05 = 'NS'$ .

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