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

✓finished code to get codon classification for each column in the alignment (to see if we are comparing codon position 1 with codon position 1)

✓figure out what R is doing with the stacked bar graphs

✓put gene expression notes from papers into electronic doc

✓put inversions notes from papers into electronic doc

✓re-do gene expression analysis with data from same strain of *S. meliloti* for pSymA and pSymB so it matches the chromosome

✓get only gene expression graphs (not including substitution data)

Last week I was mostly working on the code to properly classify each column in the alignment as codon position 1, 2, 3, non-coding (0), or overlapping with another gene (4). This took a very long time because I kept running into errors and realizing that I had to account for more situations (like different kinds of overlap) than I thought. A summary of what I found for ONE sample alignment block from *Escherichia coli* is below in Table 1. So overall, this finding is not good. It means that either the annotation is very wrong (which we really can not back up because we are not sequencing anything ourselves), or that Mafft and/or Mauve are doing a bad job at identifying what is homologous or not. When I look at the Mafft alignment it seems to be really good and having most columns the same nucleotide. So I need to look into this more.

I also looked into how R creates the stacked bar graphs (which are below) for the coding and non-coding separation of substitutions. It does what I thought it did, which is just put the totals of coding and non-coding stacked on top of each other in one bar. So the total height of the bar is the total number of substitutions in that 10kb section, and the total coding subs in that section is the bottom portion, plus the total non-coding subs in that section on top of that. This still confused me as to why there were more coding substitutions than non-coding BUT then I remembered that most of the genome is coding. So even though in the table below (Table 2) the proportion of coding and non-coding substitutions is relatively even, the actual NUMBER of coding substitutions will be more than non-coding because there are more coding regions of the genome. So I think my graphs make sense.

I also read some papers and organized the notes I took from these papers into an electronic doc to make it easier for me when I write out the gene expression papers.

When I was re-doing the gene expression analysis for pSymA and pSymB I noticed that the calculation for the midpoint of each gene was not an integer value and therefore might have been causing problems in later bidirectional calculations. So I am working on re-doing the gene expression analysis for all of the replicons (but I do not expect the results to change much). I also re-created the expression graphs on their own (without the substitution data) (graphs not shown) and some points it looks like there are complete bars/10kb sections "missing". I looked into the briefly and found that these bars are not actually missing, they just have only a few genes present in that section. The reason for this was potentially partially what I mentioned above about the non-integer

midpoint calculation, I also found some weird gene names in the *E. coli* that might be messing things up. But I think that the main reason is that in some of the samples there are some genes missing for whatever reason. So when I merge the data, these genes are not in the data frame and it therefore appears like some 10kb sections have no genes. I have decided that it would not be right to only use some samples to account for gene expression for some genes and all samples for other genes. If you can think of a better thing to do than to toss everything that does not have expression values for all samples for one gene, please let me know.

Another issue that we talked about on Friday was that some of my raw expression values are VERY high for some samples (100,000) and others are not (100's). I was concerned that I was not grabbing the correct raw count values from the data frames, or that the authors of the data did not actually provide the raw count data they said they did. However, when I normalize the samples this issue is fixed and all of the values "even out". I need to do one more check before I can feel at ease with this.

I have also been working on the corrections that you gave me for my substitutions paper.

This Week

As mentioned above, the results from the codon classification stuff is bad and I need to look into this more. We decided that I should be looking at a gene start and end from one of the taxa and look at the Mafft alignment to see what is going on in that section of the alignment.

I would also like to add up the values from the gene expression data to make sure that I do have raw count values. I would also like to re-do the expression analysis checking for any weird gene names and with the new midpoint calculation. I will then also have gene expression graphs.

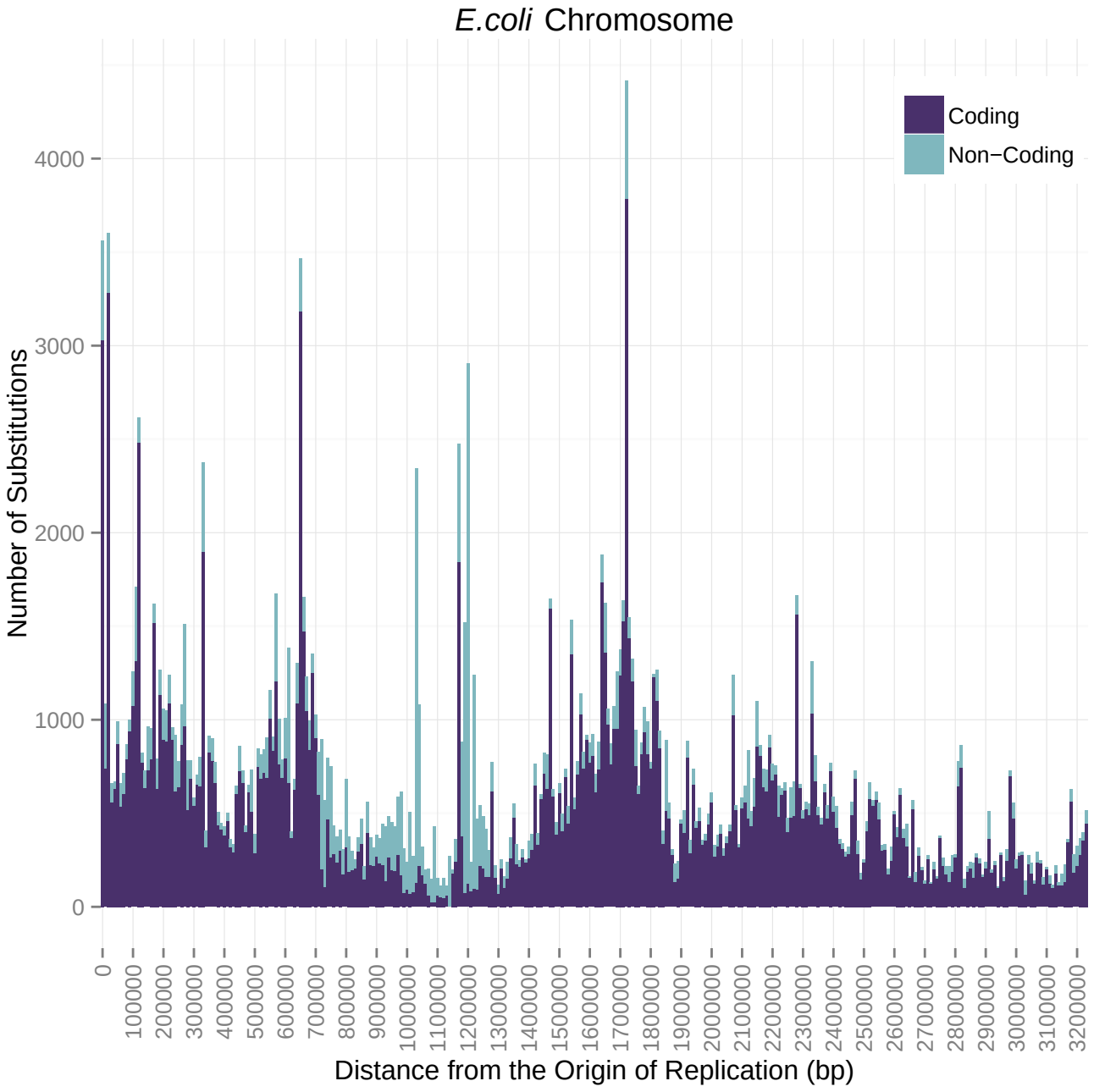
I also plan on finishing the edits for the substitutions paper this week.

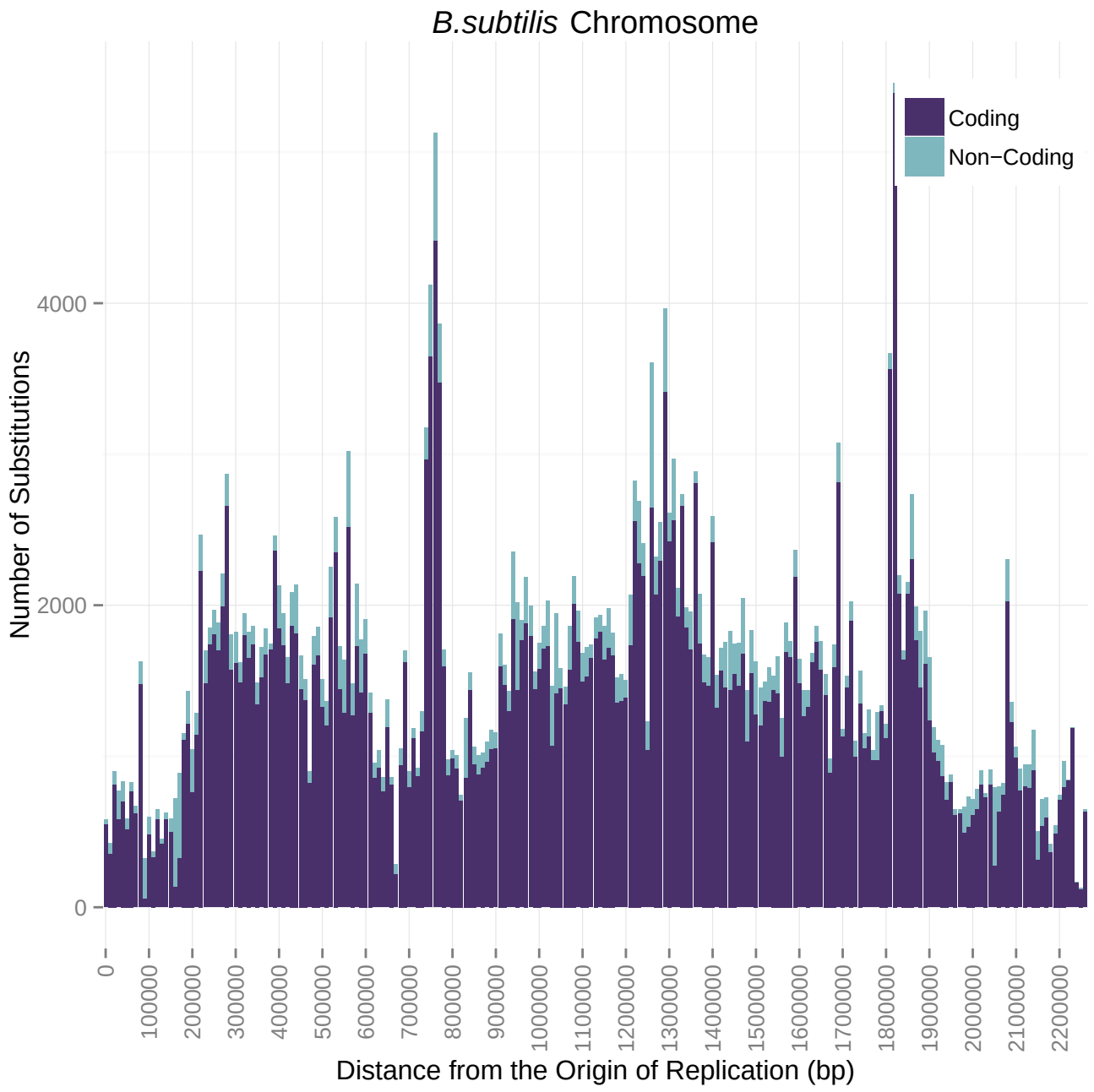
Next Week

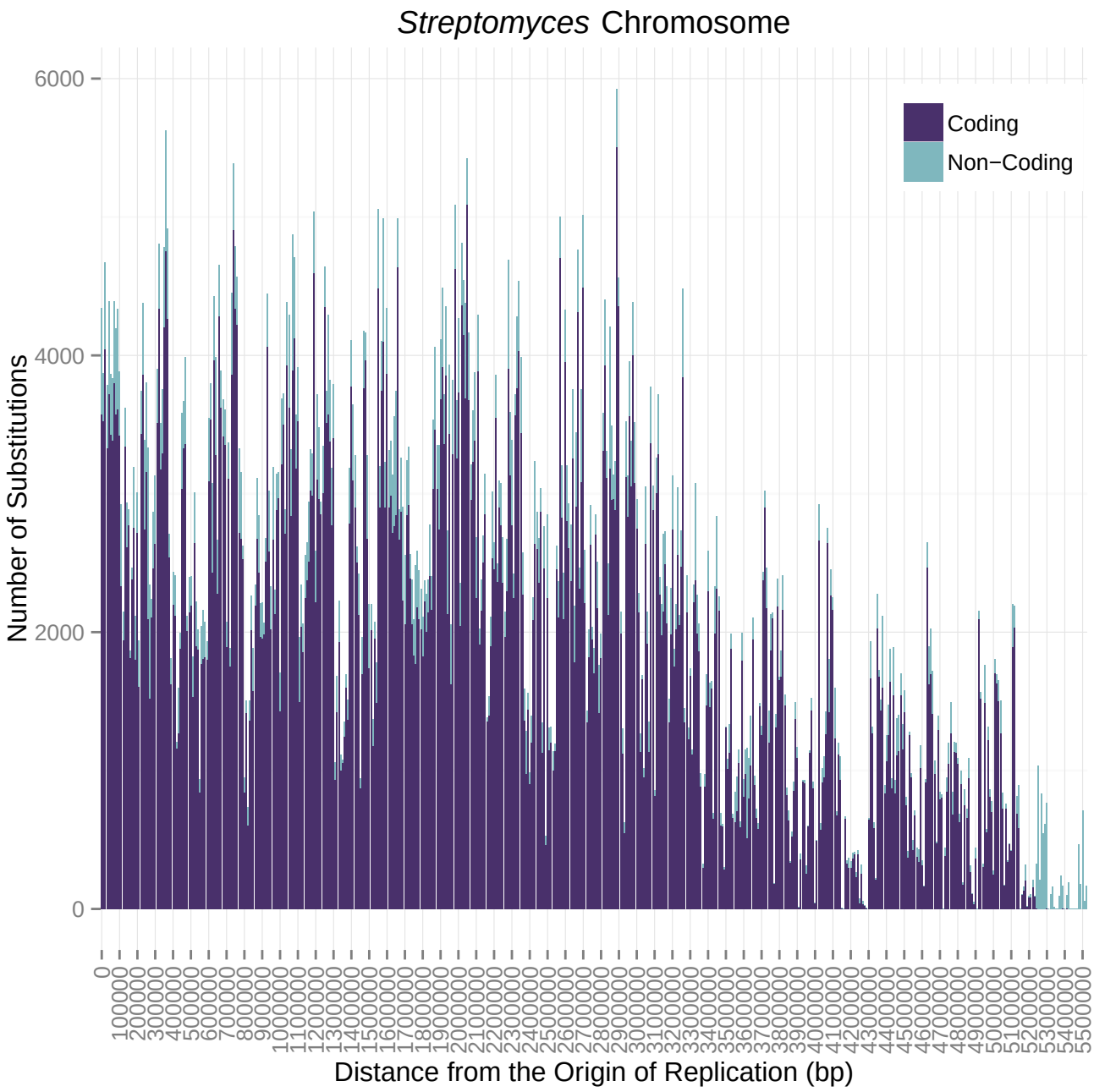
I would like to implement some sort of plan for how to deal with sites in the alignment that do not compare the same nucleotide classifications. Then I can hopefully re-run the coding and non-coding analysis and any PRANK alignments that have to happen.

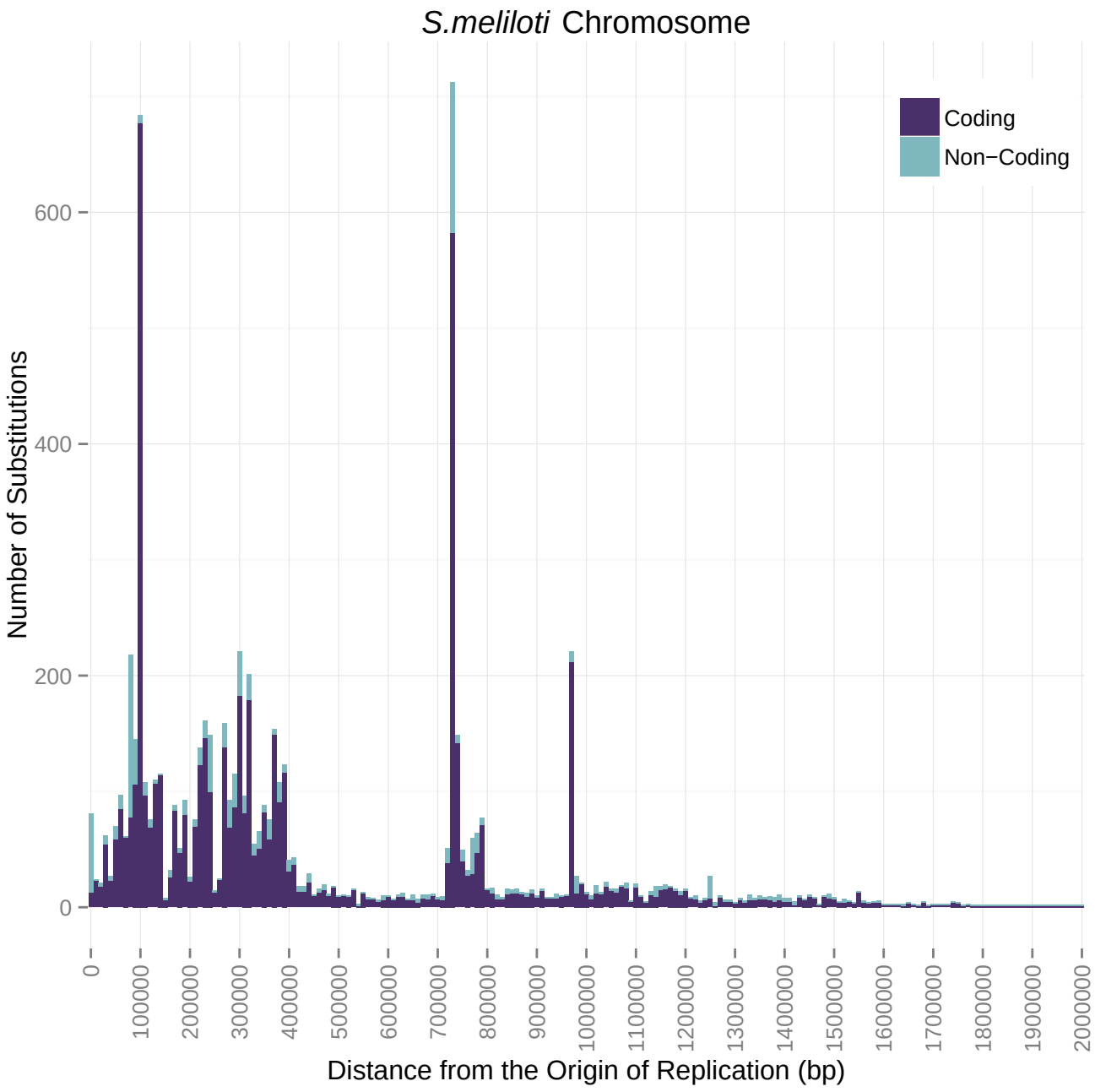
Classification	Number of Nucleotides in Alignment	% of Total Alignment
Gapped	78450	31%
Not same class	138002	55%
Same class	35852	15%
Codon 1	113	<0.05%
Codon 2	113	<0.05%
Codon 3	113	<0.05%
Misc 4	0	0
Non-coding	35513	15%

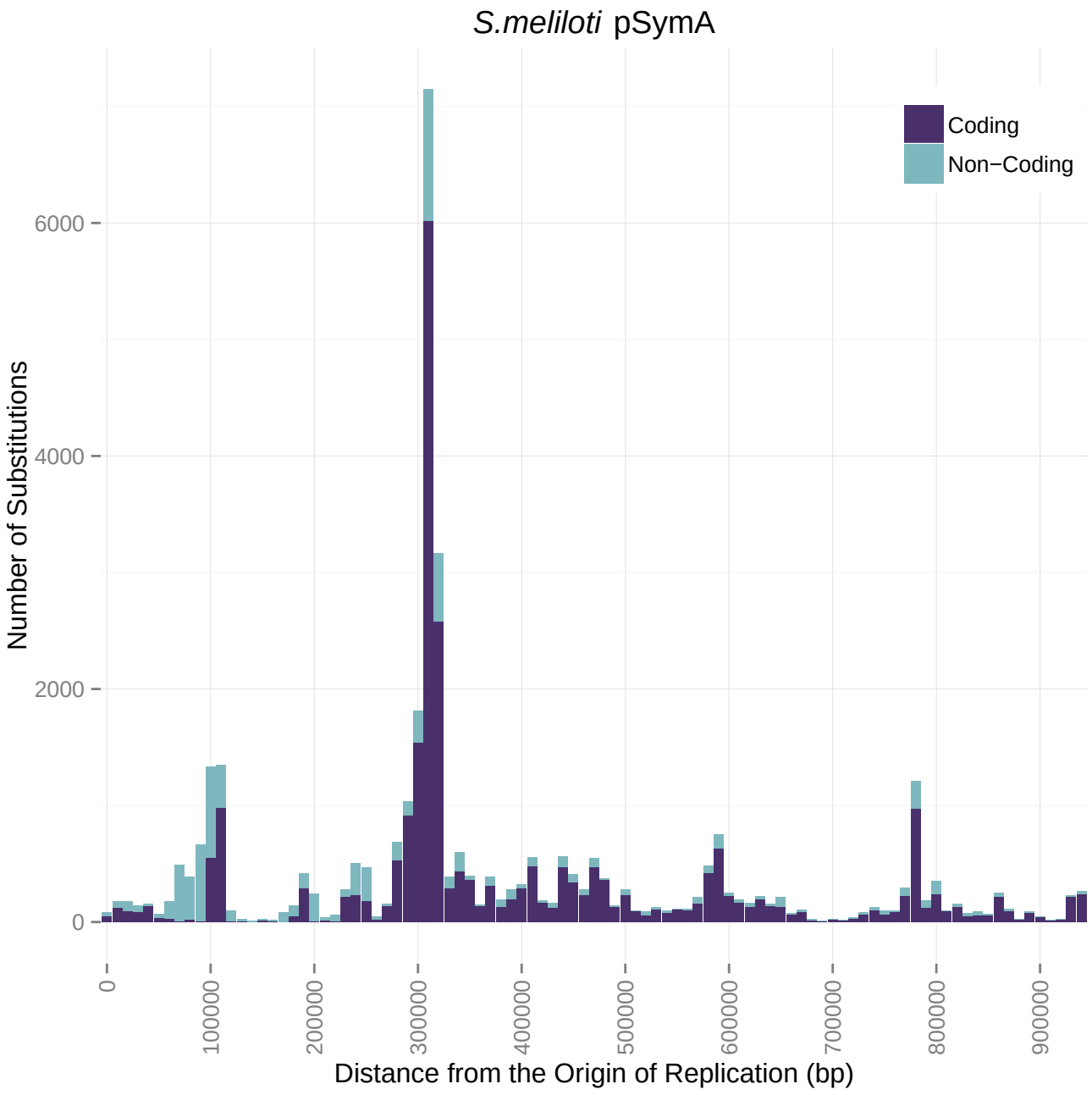
Table 1: Classification of each column in the alignment of ONE sample *E. coli* block with a total alignment length of 252304. The percentages are calculated based on the WHOLE alignment length. The "same class" classification is any column where all taxa had the same classification. "Not same class" 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.

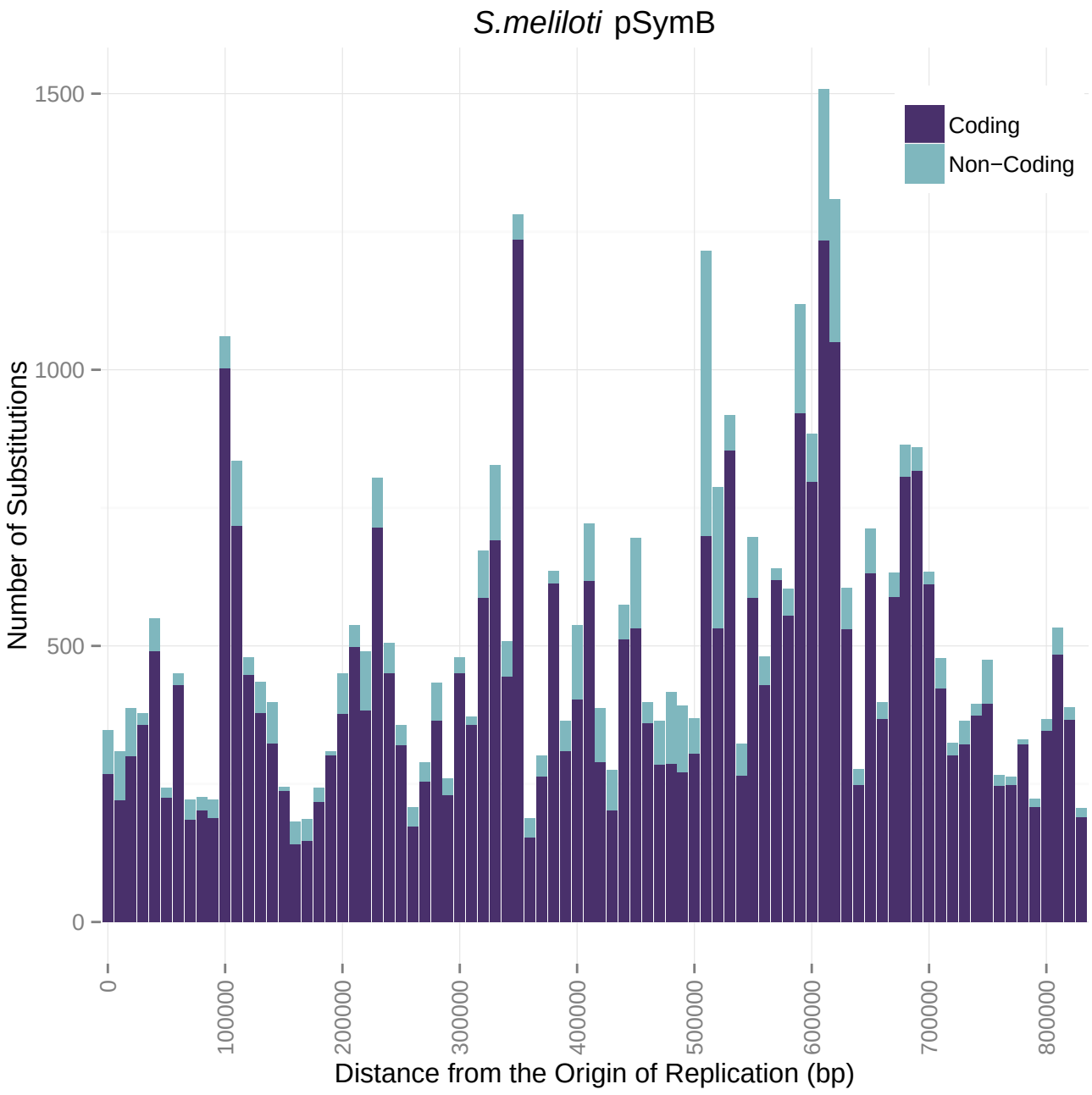

E.coli Chromosome
Coding
Non−Coding
Number of Substitutions
0
1000
2000
3000
4000
0
100000
200000
300000
400000
500000
600000
700000
800000
900000
1000000
1100000
1200000
1300000
1400000
1500000
1600000
1700000
1800000
1900000
2000000
2100000
2200000
2300000
2400000
2500000
2600000
2700000
2800000
2900000
3000000
3100000
3200000
Distance from the Origin of Replication (bp)

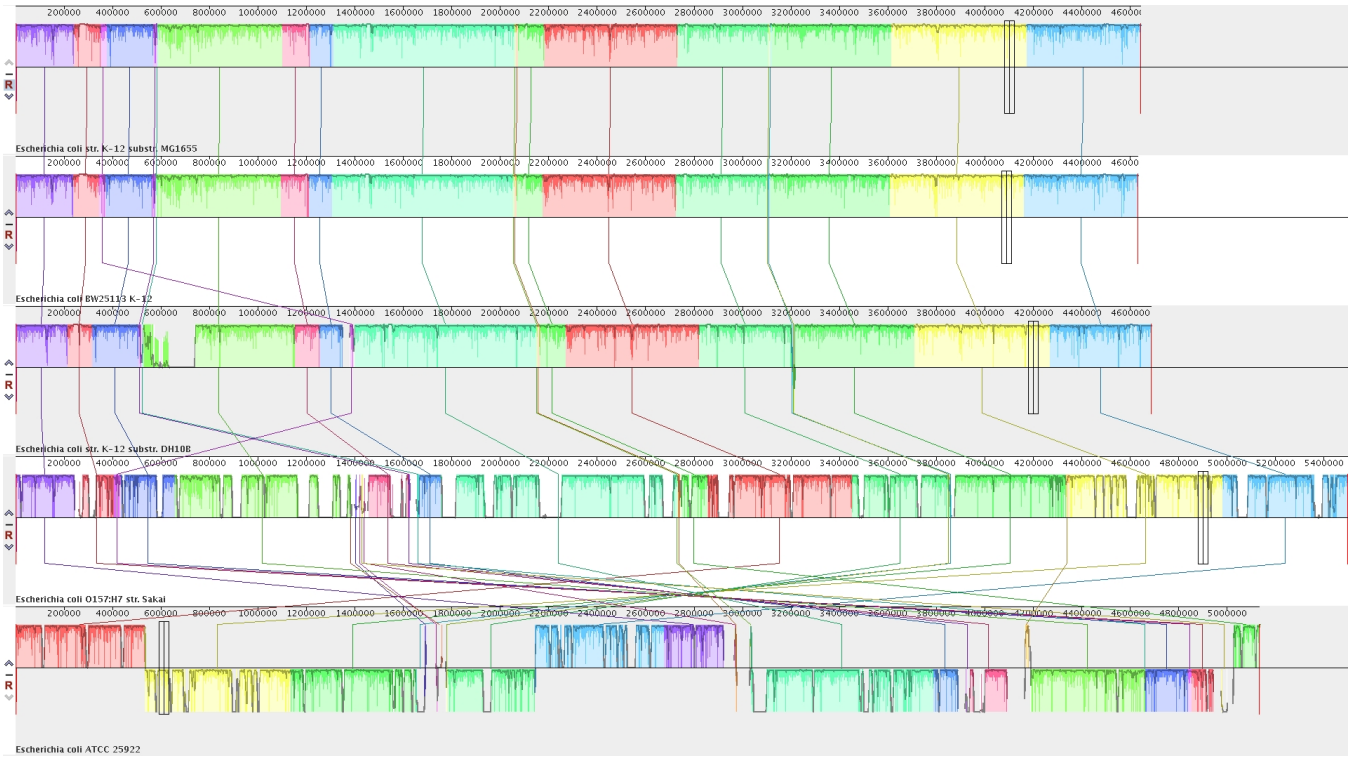

B.subtilis Chromosome
Coding
Non−Coding
Number of Substitutions
4000
2000
0
0
100000
200000
300000
400000
500000
600000
700000
800000
900000
1000000
1100000
1200000
1300000
1400000
1500000
1600000
1700000
1800000
1900000
2000000
2100000
2200000
Distance from the Origin of Replication (bp)

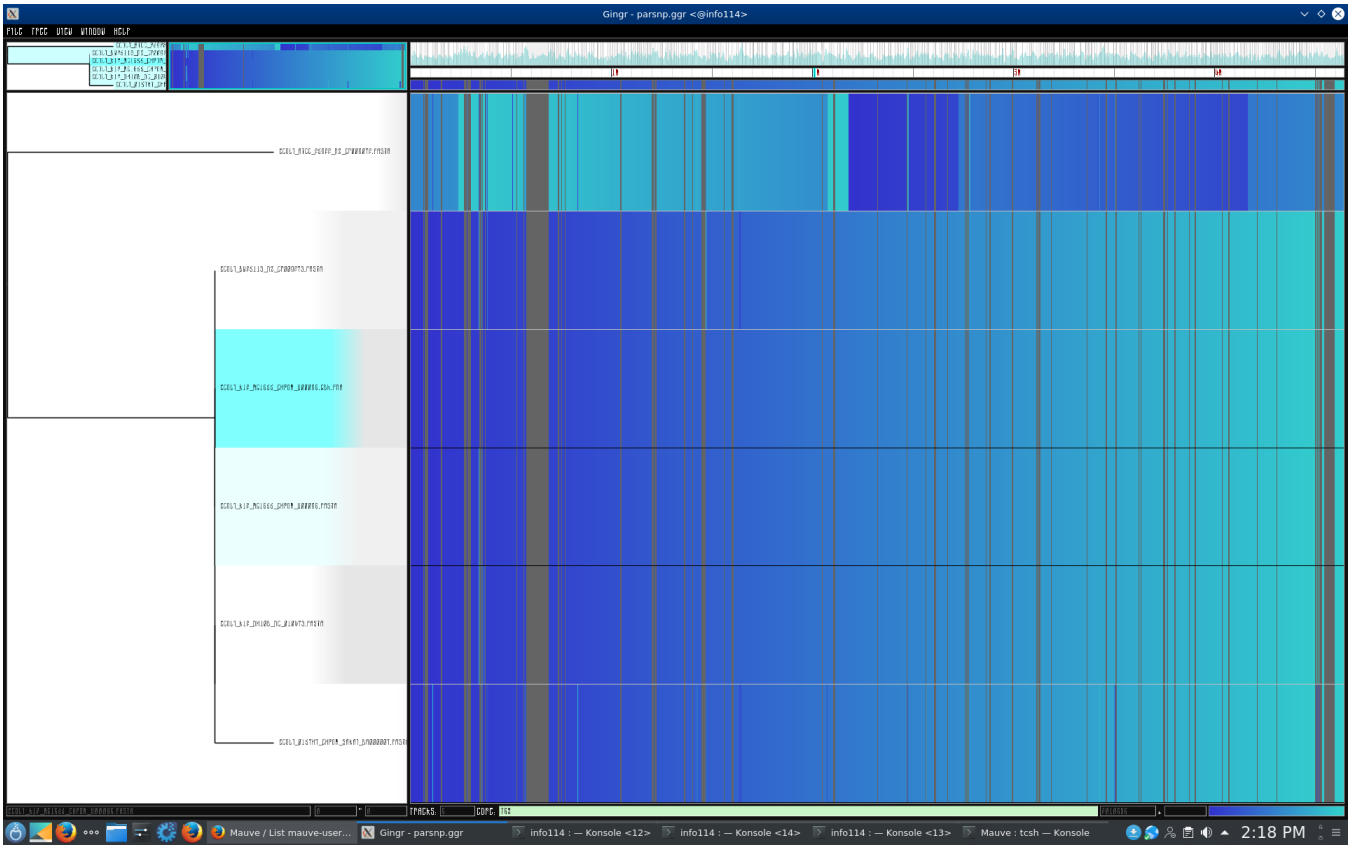

Streptomyces Chromosome
Coding
Non−Coding
Number of Substitutions
6000
4000
2000
0
0
100000
200000
300000
400000
500000
600000
700000
800000
900000
1000000
1100000
1200000
1300000
1400000
1500000
1600000
1700000
1800000
1900000
2000000
2100000
2200000
2300000
2400000
2500000
2600000
2700000
2800000
2900000
3000000
3100000
3200000
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
200
400
600
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
6000
4000
2000
0
0
100000
200000
300000
400000
500000
600000
700000
800000
900000
Distance from the Origin of Replication (bp)


S.meliloti pSymB
Coding
Non−Coding
Number of Substitutions
1500
1000
500
0
0
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 and Replicon	% of Coding Sequences	% of Non-Coding Sequences	% of Subs Coding	% of Subs Non-Coding
E. coli Chromosome	86.47%	13.53%	5.00%	8.96%
B. subtilis Chromosome	87.49%	12.51%	7.31%	6.42%
Streptomyces Chromosome	89.03%	10.97%	13.74%	14.91%
S. meliloti Chromosome	86.27%	13.73%	0.19%	0.22%
S. meliloti pSymA	83.34%	16.66%	2.84%	4.58%
S. meliloti 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
E. coli Chromosome	-5.938×10^{-8}***	-9.237×10^{-8}***
B. subtilis Chromosome	-7.584×10^{-8}***	NS
Streptomyces Chromosome	5.483×10^{-7}***	9.182×10^{-9}***
S. meliloti Chromosome	-1.448×10^{-6}***	-7.037×10^{-7}***
S. meliloti pSymA	-9.704×10^{-7}***	-1.464×10^{-7}***
S. meliloti pSymB	5.007×10^{-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
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	-1.5×10^{-6}	1.4×10^{-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.