Subs Paper Things to Do:

- causes for weird selection and subs results in *Streptomyces*
  - see how often class 4 arises in strep to see what is going on in later portion of the genome (to see if annotation is really a problem)
  - split up the strep data into core and non core and see if results are the same
- ~~make graphs proportional to length of respective cod/non-cod regions~~
- ~~test examples for genes near and far from terminus (robust log reg/results)~~
- ~~linear regression on 10kb regions for weighted and non-weighted substitutions~~
- ~~average number of substitutions in 20kb regions near and far from the origin~~
- ~~figure out why the data is weird for number of cod/non-cod sites~~
- why are the lin reg of dN, dS and ω NS but the subs graphs are...explain!
- grey out outliers in subs graphs?
- mol clock for my analysis?
- GC content? COG? where do these fit?

Gene Expression Paper Things to Do:

- ~~linear regression on 10kb regions~~
- ~~put new 10kb lin reg and # of genes over 10kb lin reg into paper~~
- ~~write about ↑ in methods and discussion~~
- ~~put expression lin reg and # coding sites log reg into supplement~~
- ~~write about ↑ in paper and how results are the same~~
- ~~update supplementary figures/file~~
- ~~correlation of gene expression across strains~~
  - ~~make graphs pretty and more informative with label names~~
  - ~~add them to supplement with a mini write up of what we did and why~~
  - ~~mention this in the actual paper~~
- if necessary add a phylogenetic component to the analysis
- ~~potentially remove genes that have been recently translocated from the analysis~~
- ~~model gene exp + position + number of genes~~

- ~~split up the strep data into core and non core and see if results are the same~~
- ~~what is going on with *Streptomyces* number of genes changing drastically from core to non-core~~
- codon bias?
- ~~what is going on with really high gene expression bars~~
- edit paper
- submit paper

Inversions and Gene Expression Letter Things to Do:

- ~~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
- make dot plot of just gene presence and absence matrix (instead of each site) to see if this will go better
- look up inversions and small RNA's paper Marie was talking about at Committee meeting
- 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

General Things to Do:

- summarize references 40 and 56 from Committee meeting report (Brian was asking)
- read and make notes on papers I found for dissertation intro

Last Week

✓new phylo tree picture for new *Streptomyces* analysis (see below fig)

✓edited gene expression paper and updated results/figs

✓protein coding subs > non-protein coding subs

✓wrote new script for subs analysis

Gene Expression Paper I made a `GitHub` link for the gene expression paper that has the supplementary information on it. I also edited the intro and discussion. I made a plot of the coefficient estimates for the gene expression linear regression to see if you like these better for the paper compared to a table of values. Thoughts? (see fig below)

Last week when I re-did the gene expression graphs, I noticed that my code for doing the bidirectionality of replication was wrong so I fixed this and re-ran everything. However, the results changed ever so slightly and most of the replicons are not significant (see Table 1). When looking at the graphs I think that this is because some of the high gene expression bars (like in *E. coli*) are driving the linear regression. So I am not sure if I should be removing outliers? Thoughts?

I also noticed that *Streptomyces* does not really have any weirdly high expression bars like the other replicons. So I am not sure if I still need to give examples of what genes are in those sections?

protein coding subs > non-protein coding subs I did some research about what proportion of a bacterial genome is protein coding and it looks like between 40-90% with an average of 88%. These numbers often include pseudogenes as protein coding, which I am not. My data for *Escherichia coli* is estimating 85% coding and 15% non-coding. So really I am not that far off so I think things are fine!

Substitution analysis As I was re-running this analysis I decided to re-write the R portion of my analysis into one concise script. As I was doing this I also changed the *Streptomyces* figures to have negative genomic positions and overall my code runs faster. I also realised I was removing outliers in places where I should not have, so I have fixed that. I am not sure if I should be removing outliers in the various 10kb linear regressions. Thoughts?

This Week

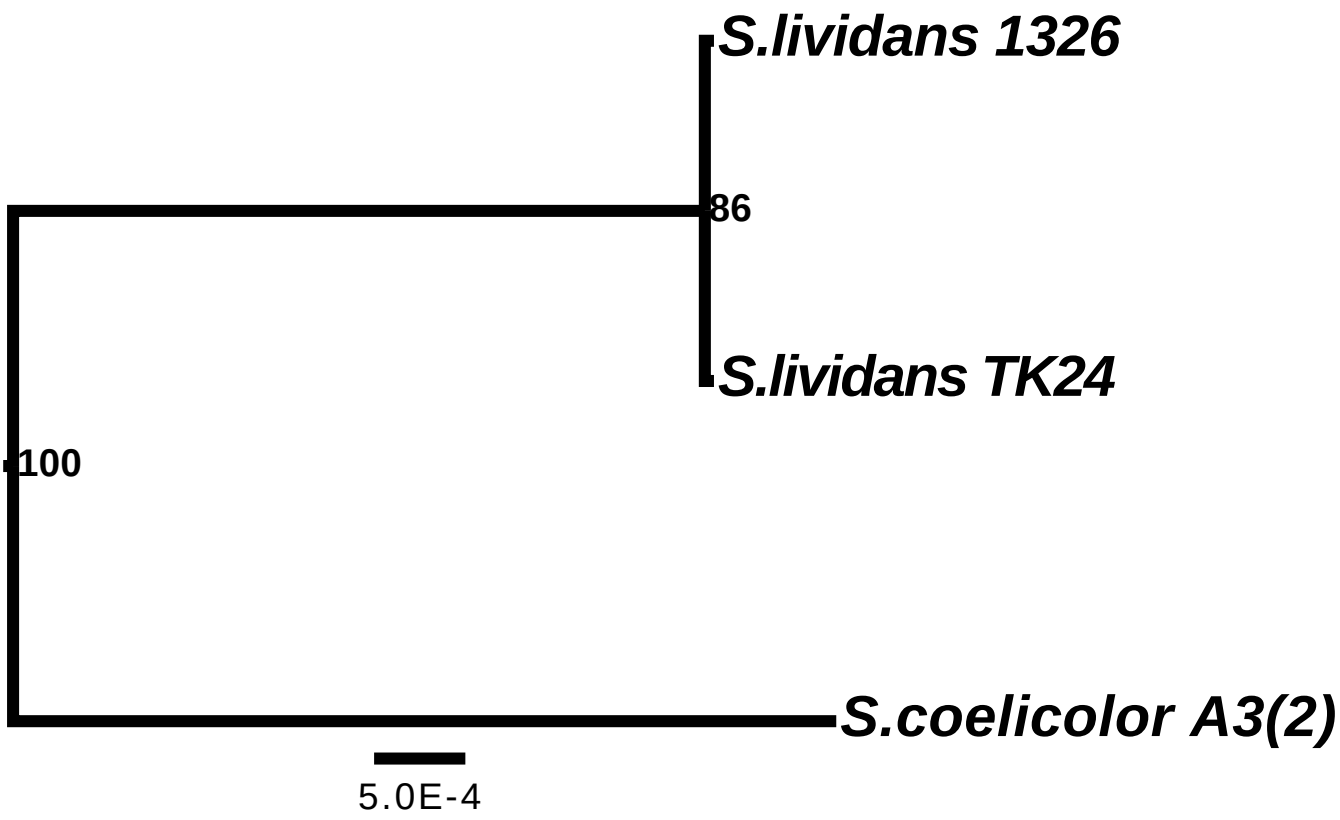
Substitutions project:

1. figure out issue with alignment length for selection analysis
2. re-run selection analysis

Next Week

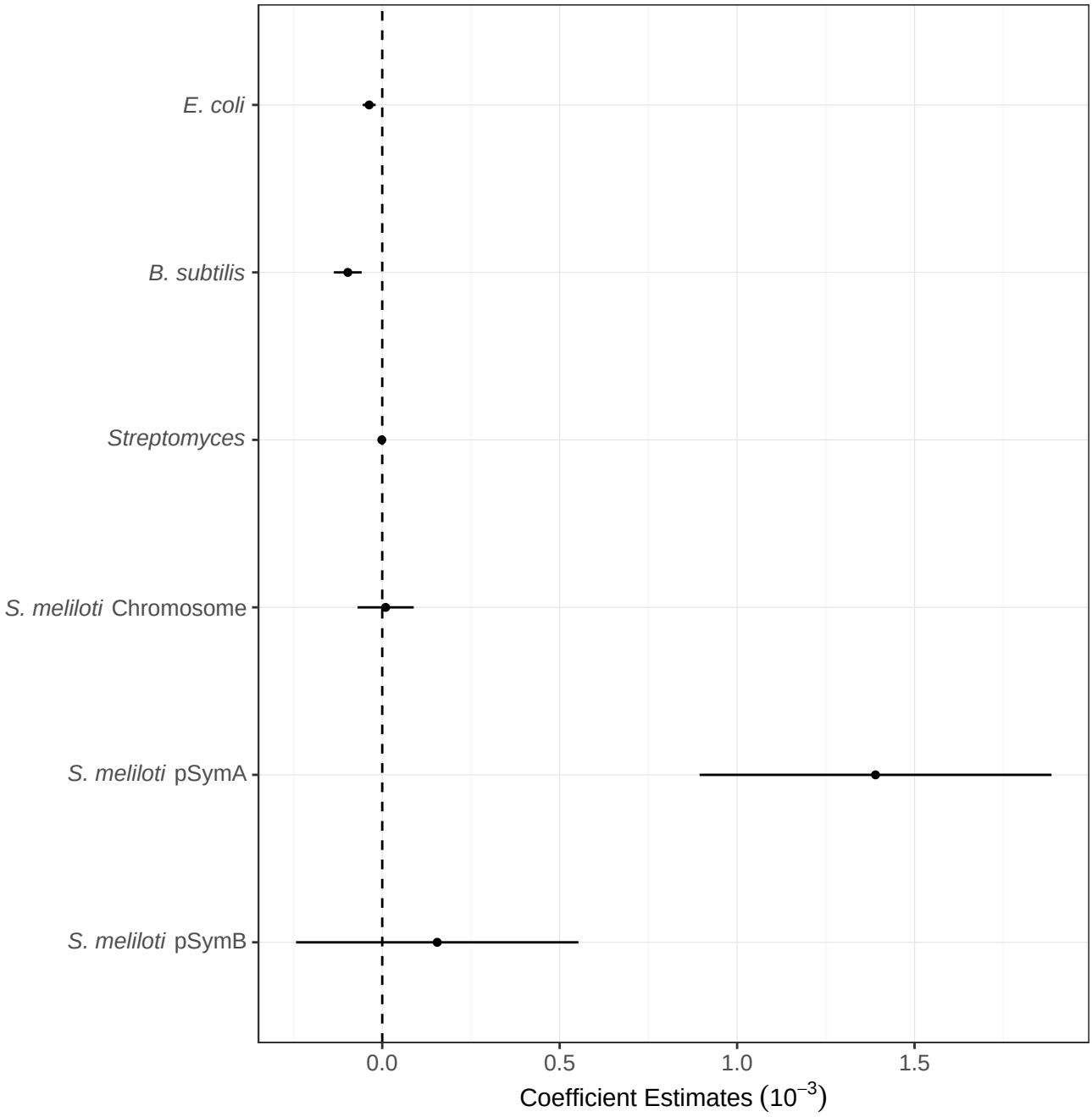
Gene Expression:

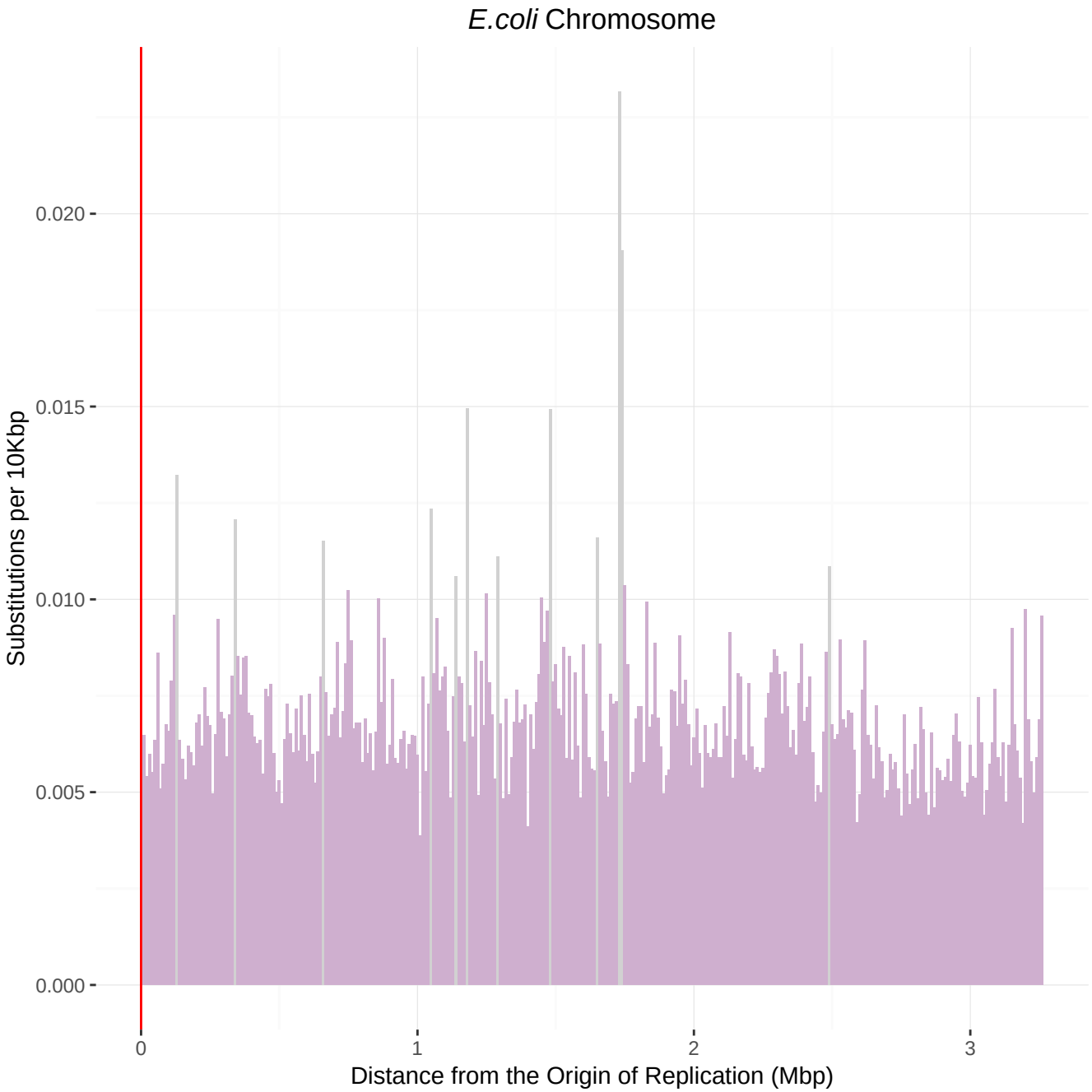
1. look into journal requirements for submission
2. write cover letter for gene expression paper
3. substitutions paper edits

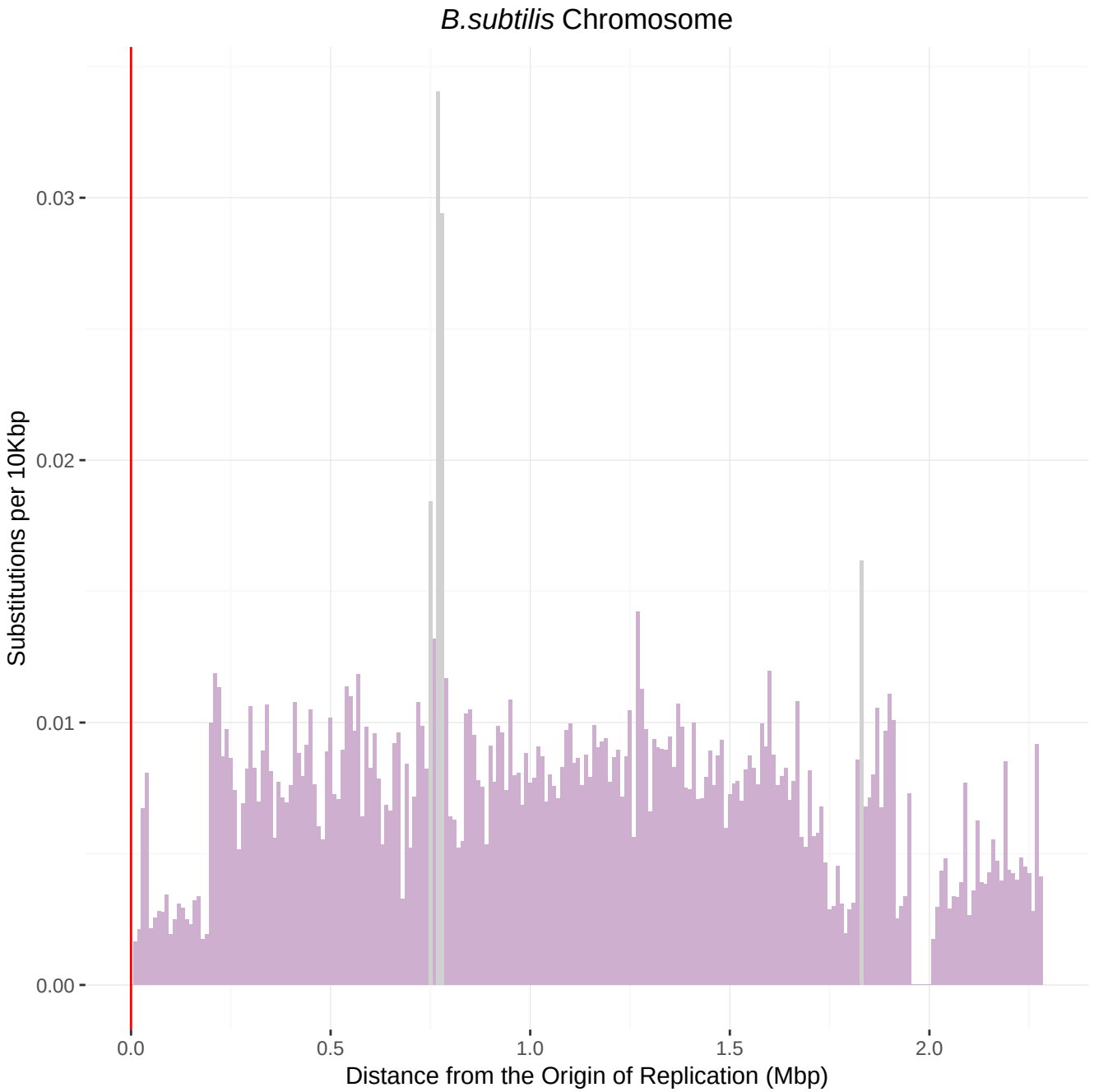


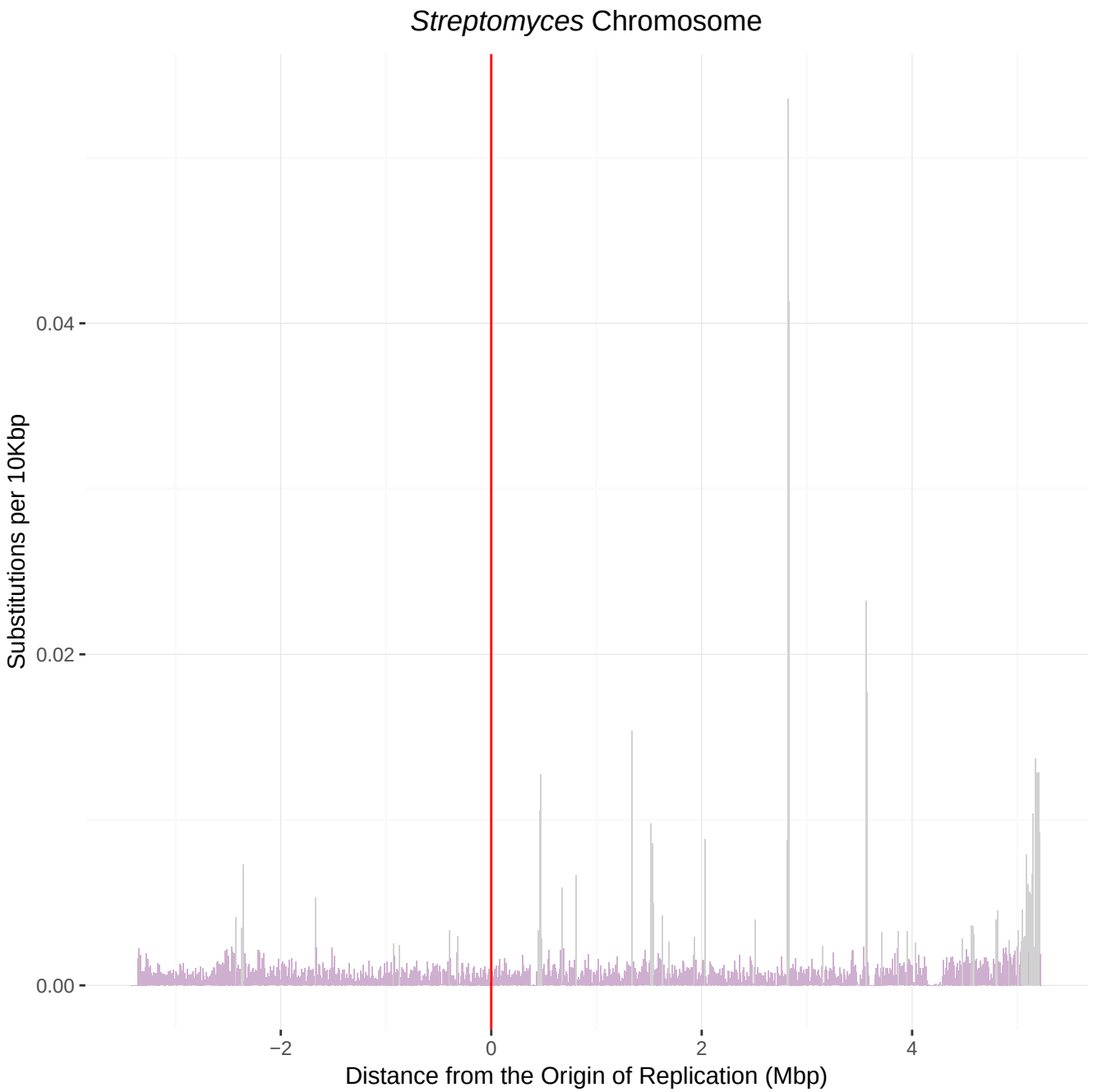
Bacteria and Replicon	Gene Expression 10Kbp
E. coli Chromosome	NS
B. subtilis Chromosome	-2.36×10^{-5}*
Streptomyces Chromosome	NS
S. meliloti Chromosome	NS
S. meliloti pSymA	8.98×10^{-4}**
S. meliloti pSymB	NS

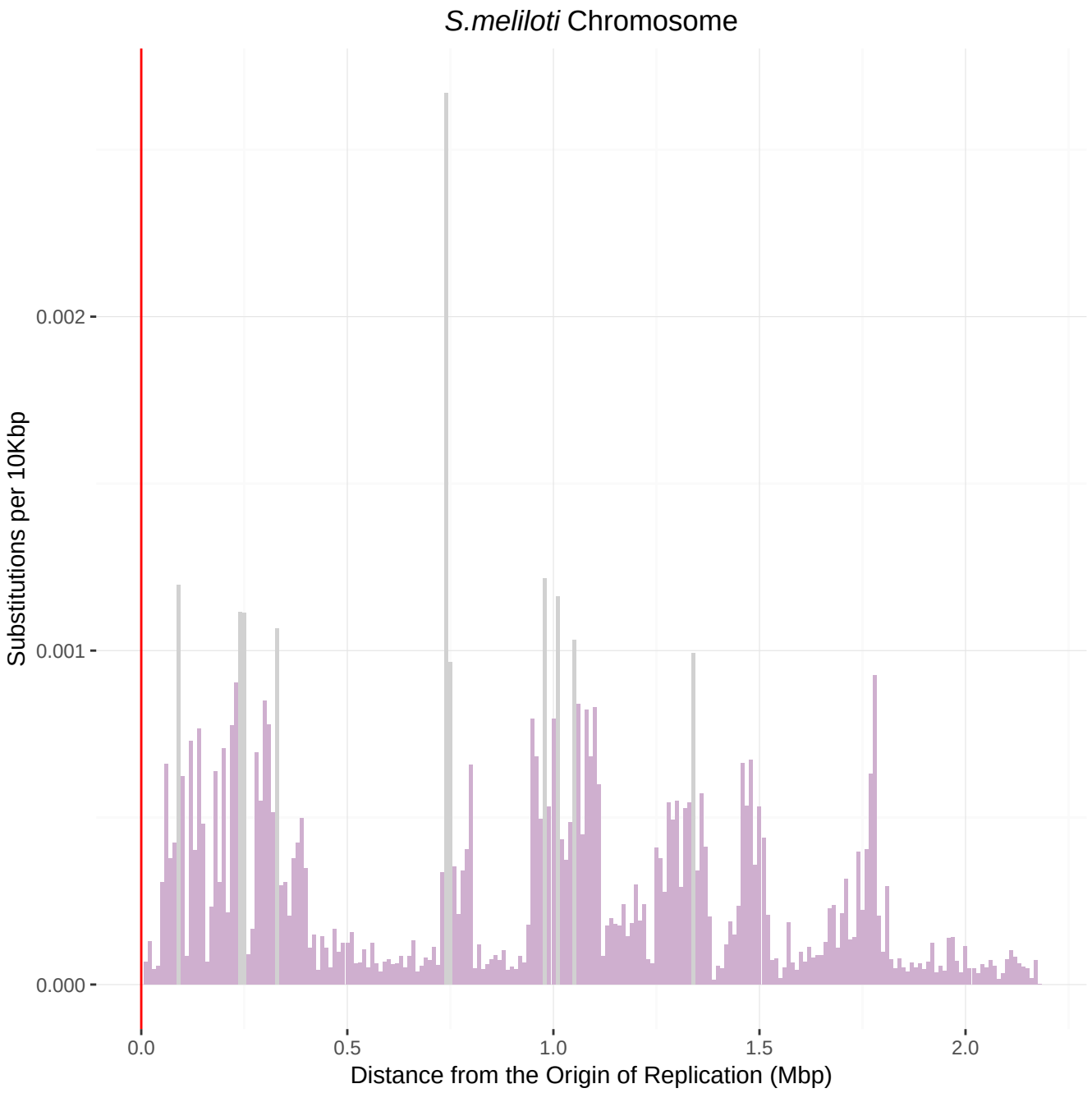
Table 1: Linear regression analysis of the median counts per million expression data for 10Kbp segments of the genome of the respective bacteria replicons. Linear regression was calculated after the origin of replication was moved to the beginning of the genome and all subsequent positions were scaled around the origin accounting for bidirectionality of replication. All results are marked with significance codes as followed: < 0.001 = '***', $0.001 < 0.01$ = '**', $0.01 < 0.05$ = '*', > 0.05 = 'NS'.

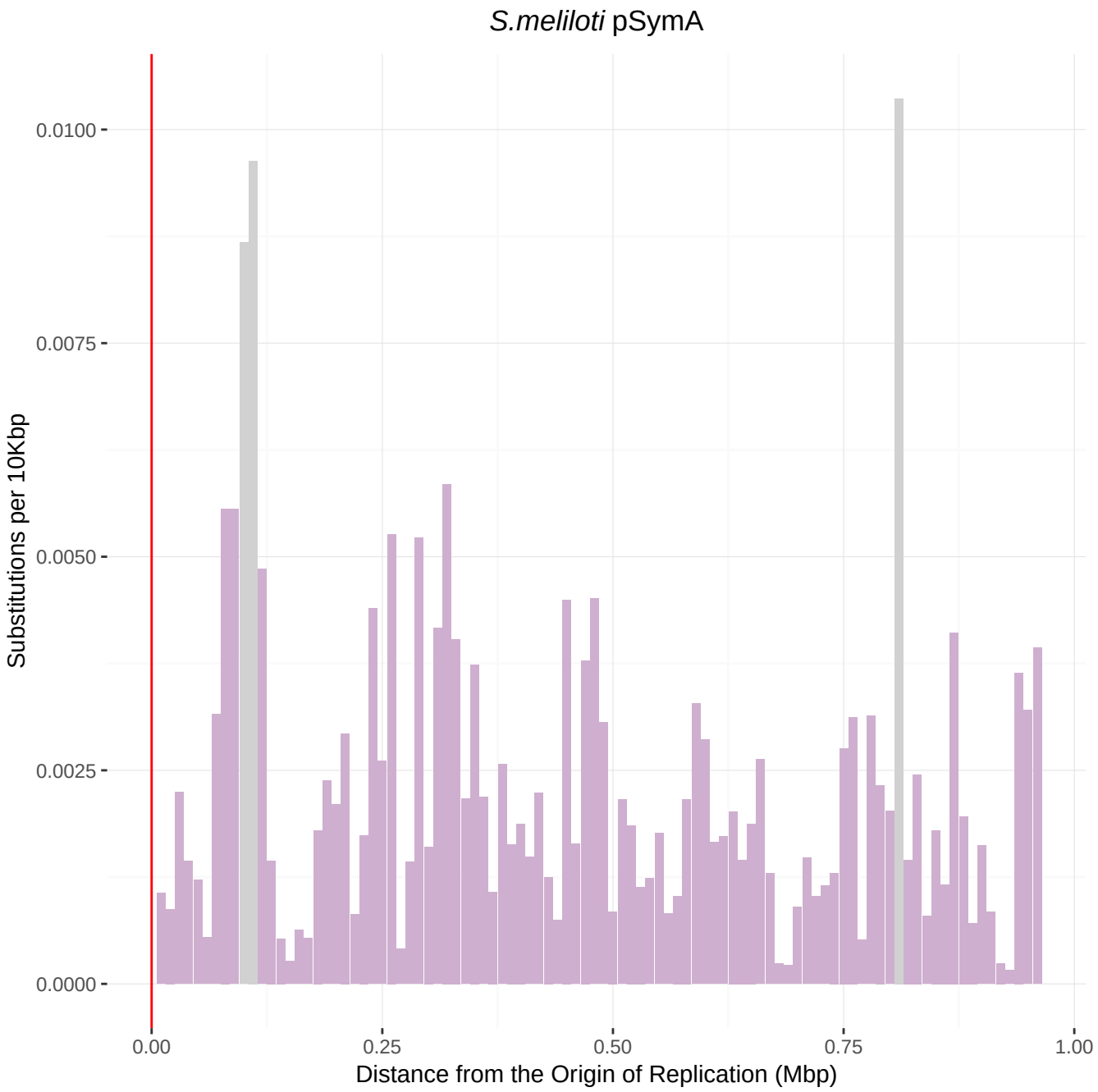

E. coli
B. subtilis
Streptomyces
S. meliloti Chromosome
S. meliloti pSymA
S. meliloti pSymB
0.0
0.5
1.0
1.5
Coefficient Estimates (10−3)

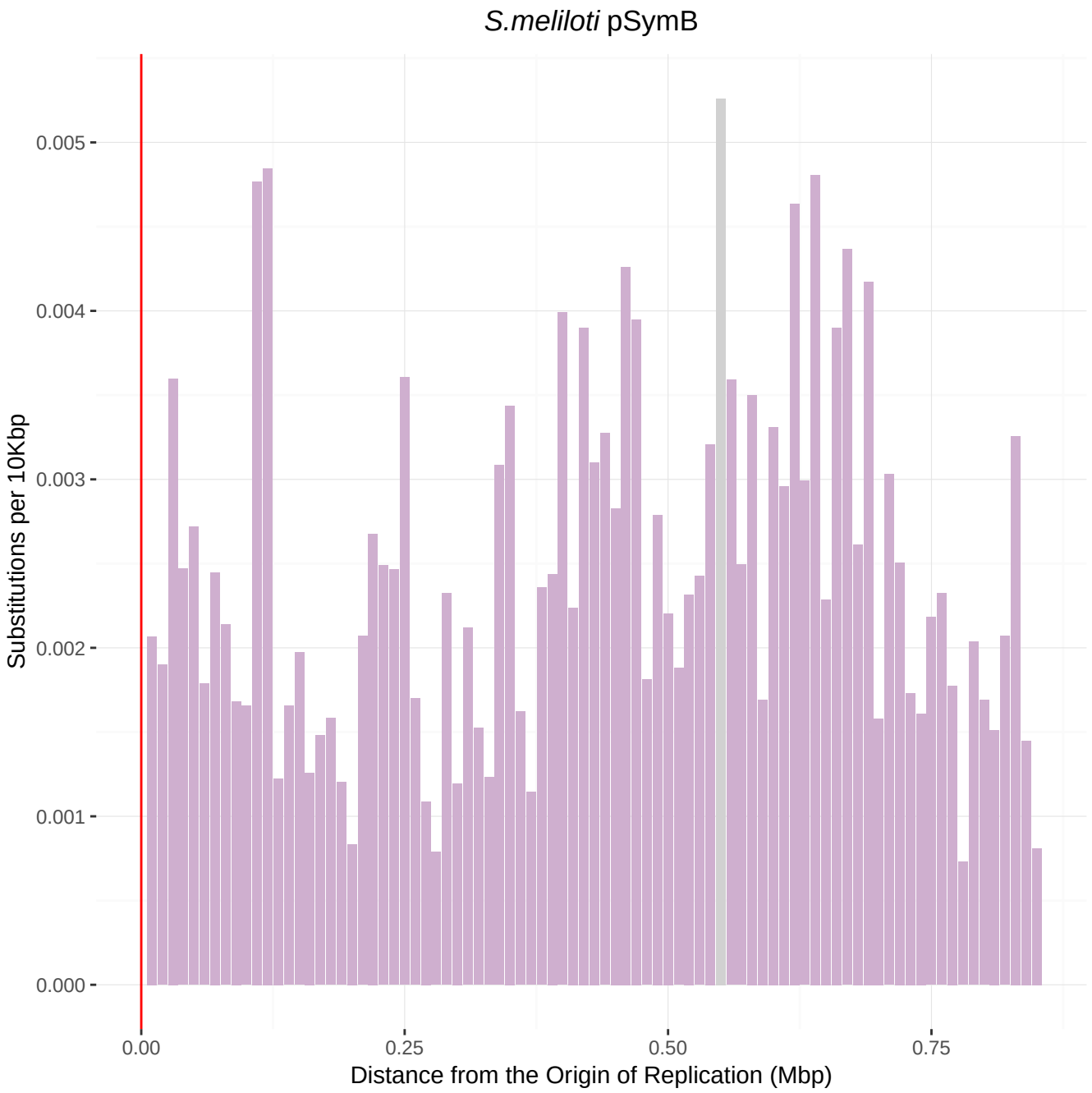

E.coli Chromosome
Substitutions per 10Kbp
0.020
0.015
0.010
0.005
0.000
0
1
2
3
Distance from the Origin of Replication (Mbp)


B.subtilis Chromosome
Substitutions per 10Kbp
0.03
0.02
0.01
0.00
0.0
0.5
1.0
1.5
2.0
Distance from the Origin of Replication (Mbp)


Streptomyces Chromosome
0.04
0.02
0.00
Substitutions per 10Kbp
−2
0
2
4
Distance from the Origin of Replication (Mbp)


S.meliloti Chromosome
Substitutions per 10Kbp
0.002
0.001
0.000
0.0
0.5
1.0
1.5
2.0
Distance from the Origin of Replication (Mbp)


S.meliloti pSymA
0.0100
0.0075
0.0050
0.0025
0.0000
Substitutions per 10Kbp
0.00
0.25
0.50
0.75
1.00
Distance from the Origin of Replication (Mbp)


S.meliloti pSymB
Substitutions per 10Kbp
0.005
0.004
0.003
0.002
0.001
0.000
0.00
0.25
0.50
0.75
Distance from the Origin of Replication (Mbp)

Bacteria and Replicon	Protein Coding Sequences
E. coli Chromosome	-2.99×10^{-8}***
B. subtilis Chromosome	-8.05×10^{-8}***
Streptomyces Chromosome	1.10×10^{-7}***
S. meliloti Chromosome	-4.32×10^{-7}***
S. meliloti pSymA	-5.00×10^{-7}***
S. meliloti pSymB	1.77×10^{-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'.

	Protein Coding			
	Correlation Coefficient 20kb Near		Number of Substitutions per 20kb Near	
Bacteria and Replicon	Origin	Terminus	Origin	Terminus
E. coli Chromosome	-1.21×10^{-5}**	-2.68×10^{-5}*	5.95×10^{-3}	7.03×10^{-3}
B. subtilis Chromosome	NS	-4.25×10^{-5}**	1.97×10^{-3}	8.75×10^{-3}
Streptomyces Chromosome	9.05×10^{-5}***	-1.36×10^{-4}***	6.65×10^{-4}	6.05×10^{-3}
S. meliloti Chromosome	NS	NS	9.85×10^{-5}	5.18×10^{-5}
S. meliloti pSymA	NS	NS	9.61×10^{-4}	3.53×10^{-3}
S. meliloti pSymB	-2.29×10^{-5}***	-5.60×10^{-5}***	1.98×10^{-3}	1.22×10^{-3}

Table 3: Logistic regression on 20kb closest and farthest from the origin of replication after accounting for bidirectional replication and outliers. All results are marked with significance codes as followed: < 0.001 = '***', $0.001 < 0.01$ = '**', $0.01 < 0.05$ = '*', > 0.05 = 'NS'.

	Protein Coding	
Bacteria and Replicon	Weighted	Non-Weighted
E. coli Chromosome	-3.31×10^{-10}**	-1.93×10^{-4}***
B. subtilis Chromosome	1.01×10^{-9}***	-2.08×10^{-4}***
Streptomyces Chromosome	2.64×10^{-10}***	NS
S. meliloti Chromosome	-1.49×10^{-10}***	-1.82×10^{-5}***
S. meliloti pSymA	NS	-1.17×10^{-4}*
S. meliloti pSymB	NS	NS

Table 4: Linear regression on 10kb sections of the genome with increasing distance from the origin of replication after accounting for bidirectional replication. Weighted columns have the total number of substitutions in each 10kb section of the genome divided by the total number of protein coding and non-protein coding sites in the genome. Non-weighted columns are performing a linear regression on the total number of substitutions in each 10kb section of the genome. All results are marked with significance codes as followed: < 0.001 = '***', $0.001 < 0.01$ = '**', $0.01 < 0.05$ = '*', > 0.05 = 'NS'.

Bacteria and Replicon	Coefficient Estimate
E. coli Chromosome	-2.50×10^{-2}***
B. subtilis Chromosome	-1.99×10^{-2}**
Streptomyces Chromosome	-1.74×10^{-3}***
S. meliloti Chromosome	-1.90×10^{-2}***
S. meliloti pSymA	-2.44×10^{-2}*
S. meliloti pSymB	NS

Table 5: Linear regression analysis of the total number of protein coding genes per 10kb along the genome of the respective bacteria replicons. 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. All results are marked with significance codes as followed: < 0.001 = '***', $0.001 < 0.01$ = '**', $0.01 < 0.05$ = '*', > 0.05 = 'NS'.