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 $\omega$ 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

✓looked into phylogenetic papers on *Streptomyces* that Marie sent me

✓re-run *Streptomyces* analysis (was using the wrong tree last time)

✓tweak the selection distribution figures (fix labelling on axis)

✓gene expression paper latex template working

✓gene expression paper edits and minor formatting fixes

**Gene Expression Paper**

I sent you the latest draft of the paper which is formatted properly in the latex template that the journal provides. I spent most of the week getting this to work, editing and tweaking other minor formatting issues.

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 + Selection analysis**

I realized that I was using the wrong phylogenetic tree for the *Streptomyces* analysis, so I need to re-do this (and the selection analysis) over again. I started this late last week and this should be done by the end of this week at the latest.

I also slightly changed the selection genomic distribution figures (edited axis names). These are found in the attached supplement file.

# This Week

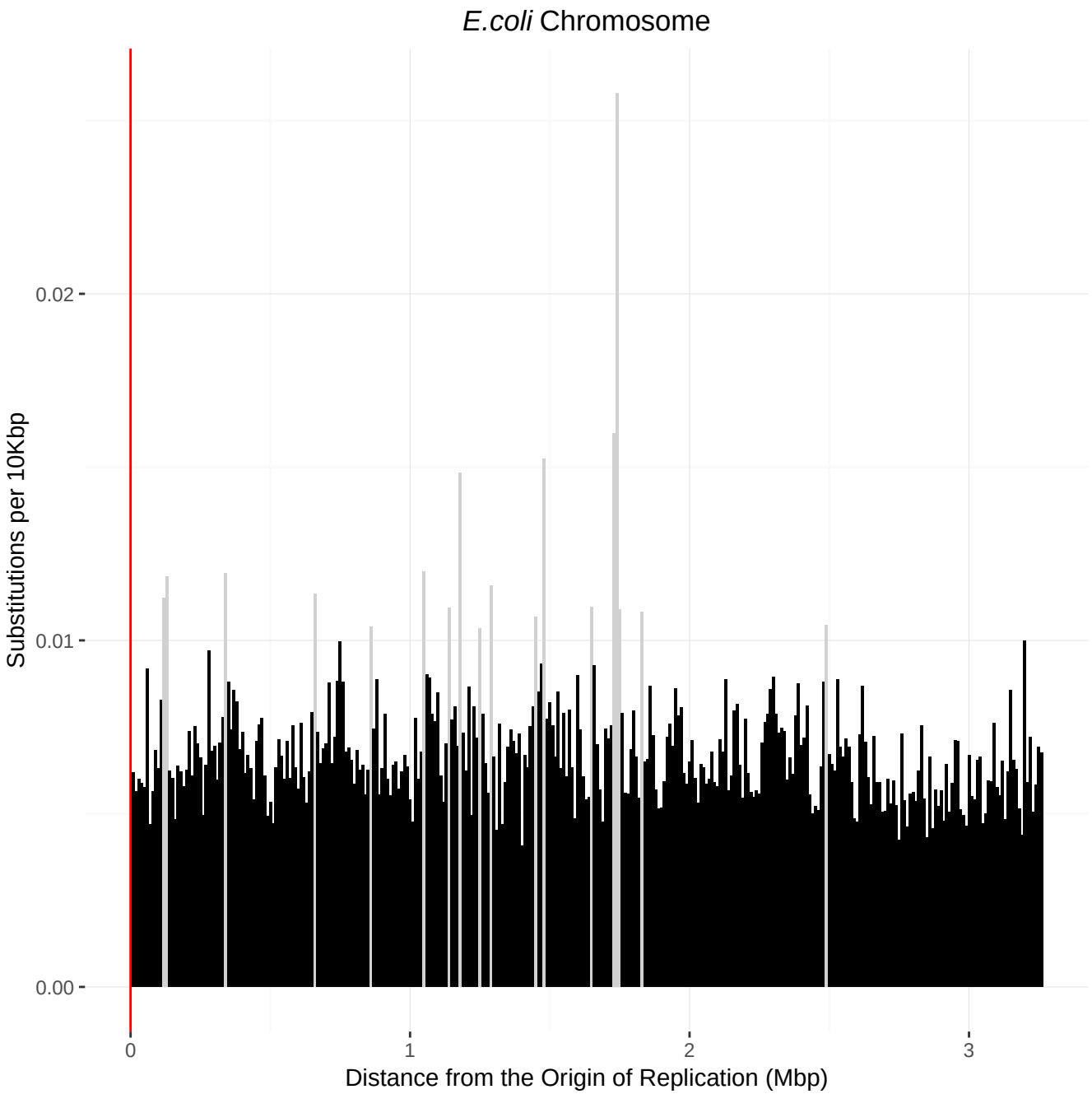
1. write cover letter for gene expression paper
2. gene expression paper data + code up on github
3. finish all final formatting for gene expression paper so all I have to do is click submit!

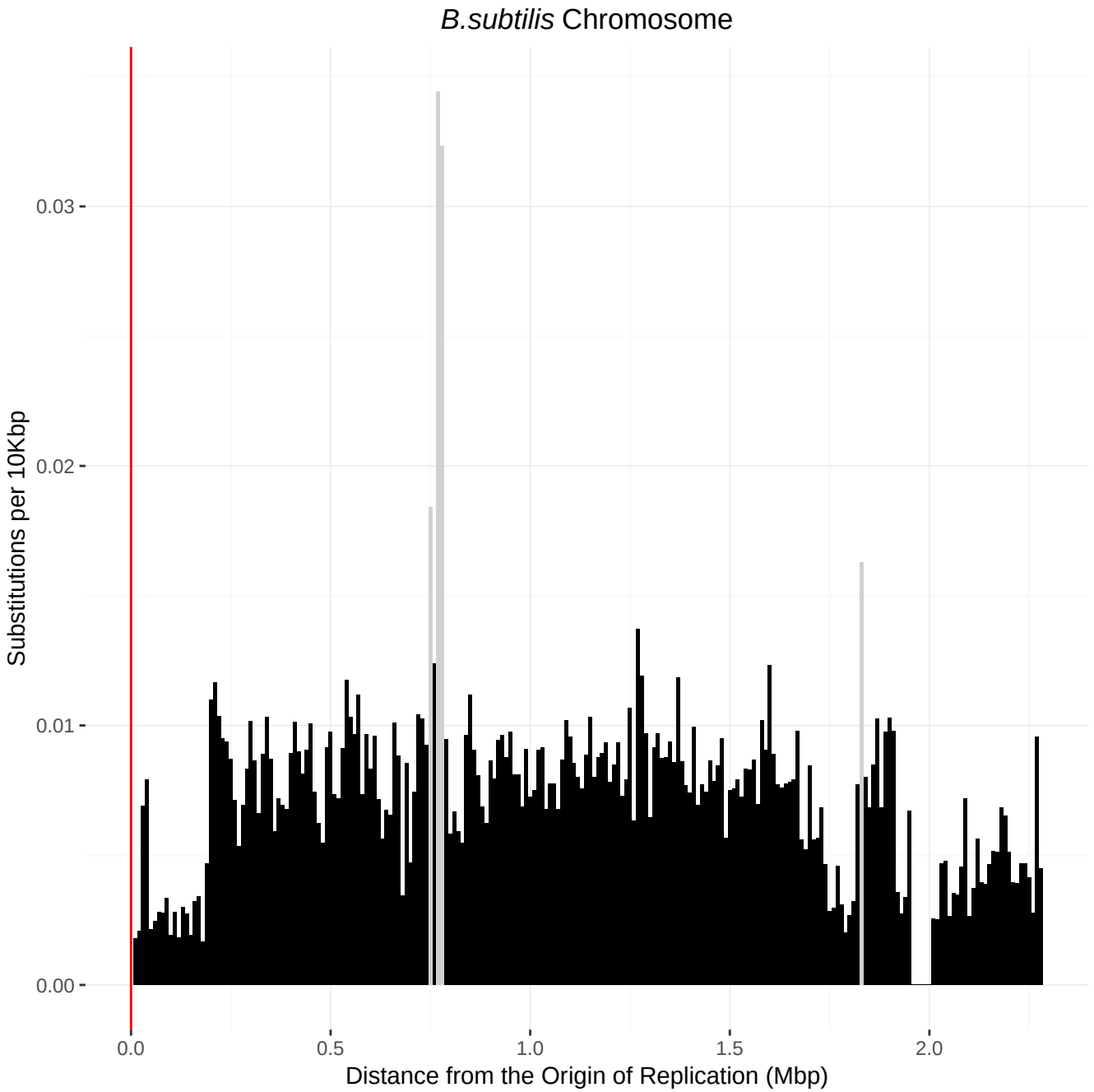
# Next Week

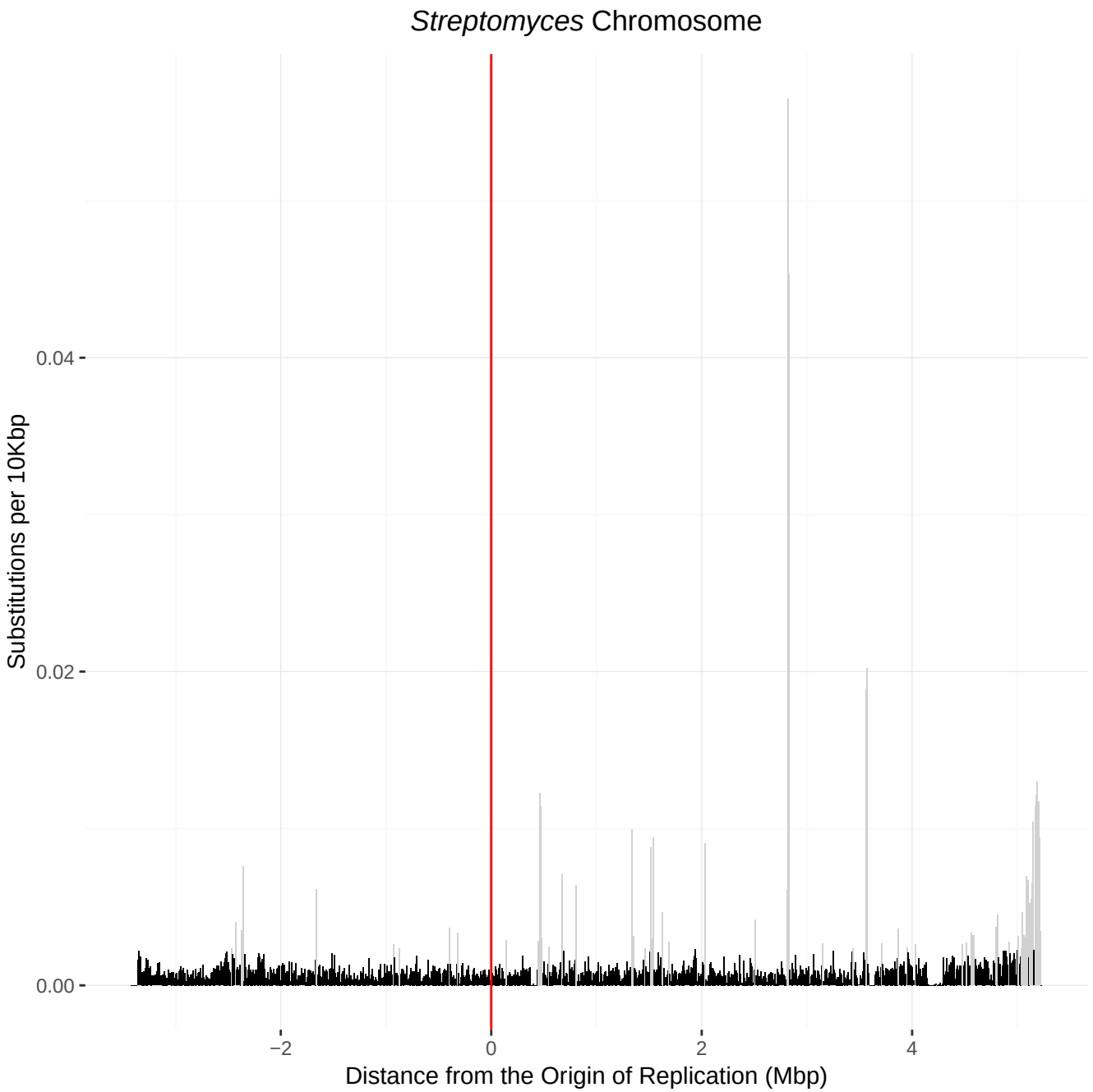
1. analize selection results
2. analize substitutions results
3. substitutions paper edits

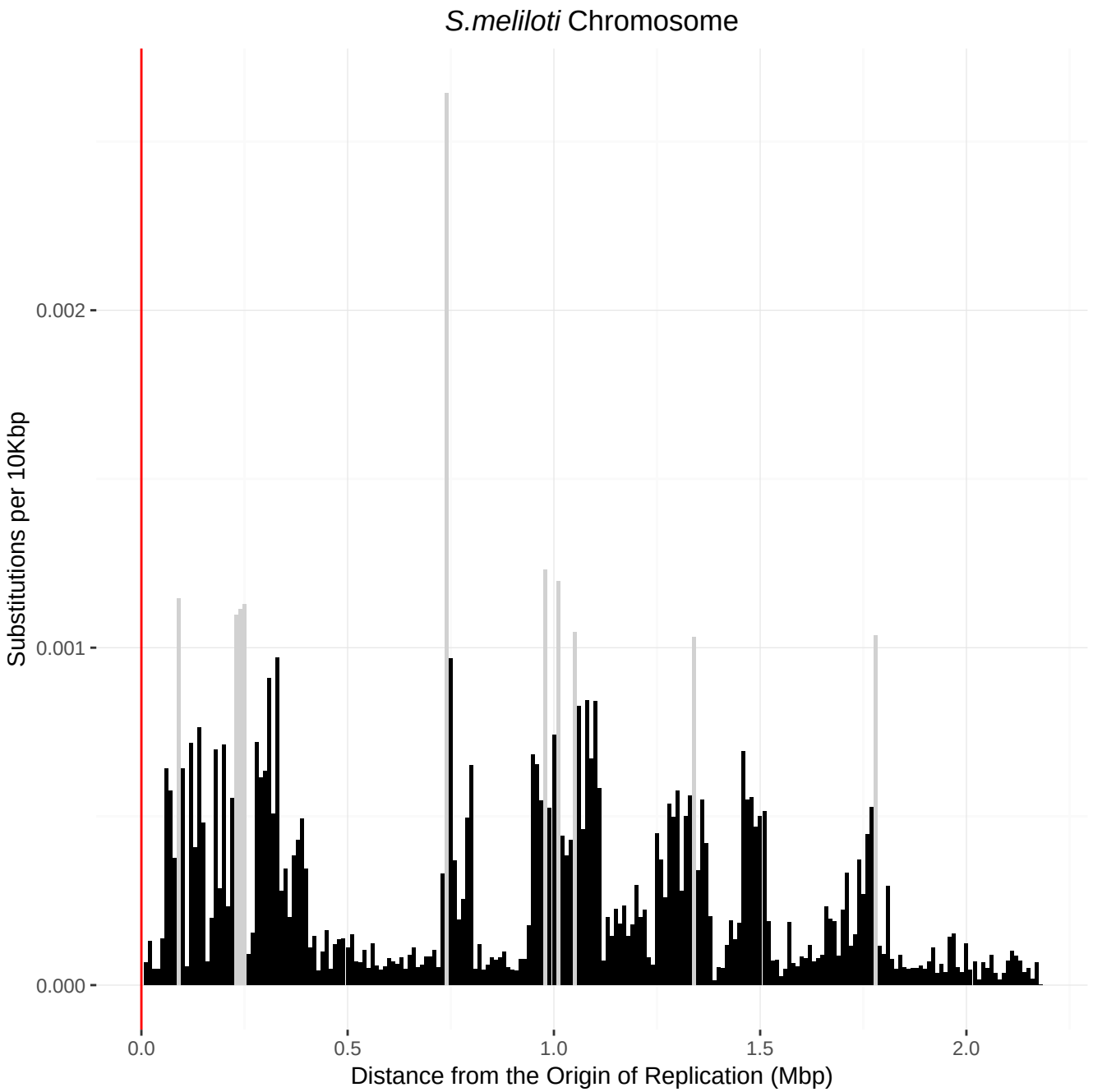
| Bacteria and Replicon | Gene Expression 10Kbp |
|---|---|
| *E. coli* Chromosome | NS |
| *B. subtilis* Chromosome | $-2.36\times10^{-5}$* |
| *Streptomyces* Chromosome | NS |
| *S. meliloti* Chromosome | NS |
| *S. meliloti* pSymA | $8.98\times10^{-4}$** |
| *S. meliloti* pSymB | NS |

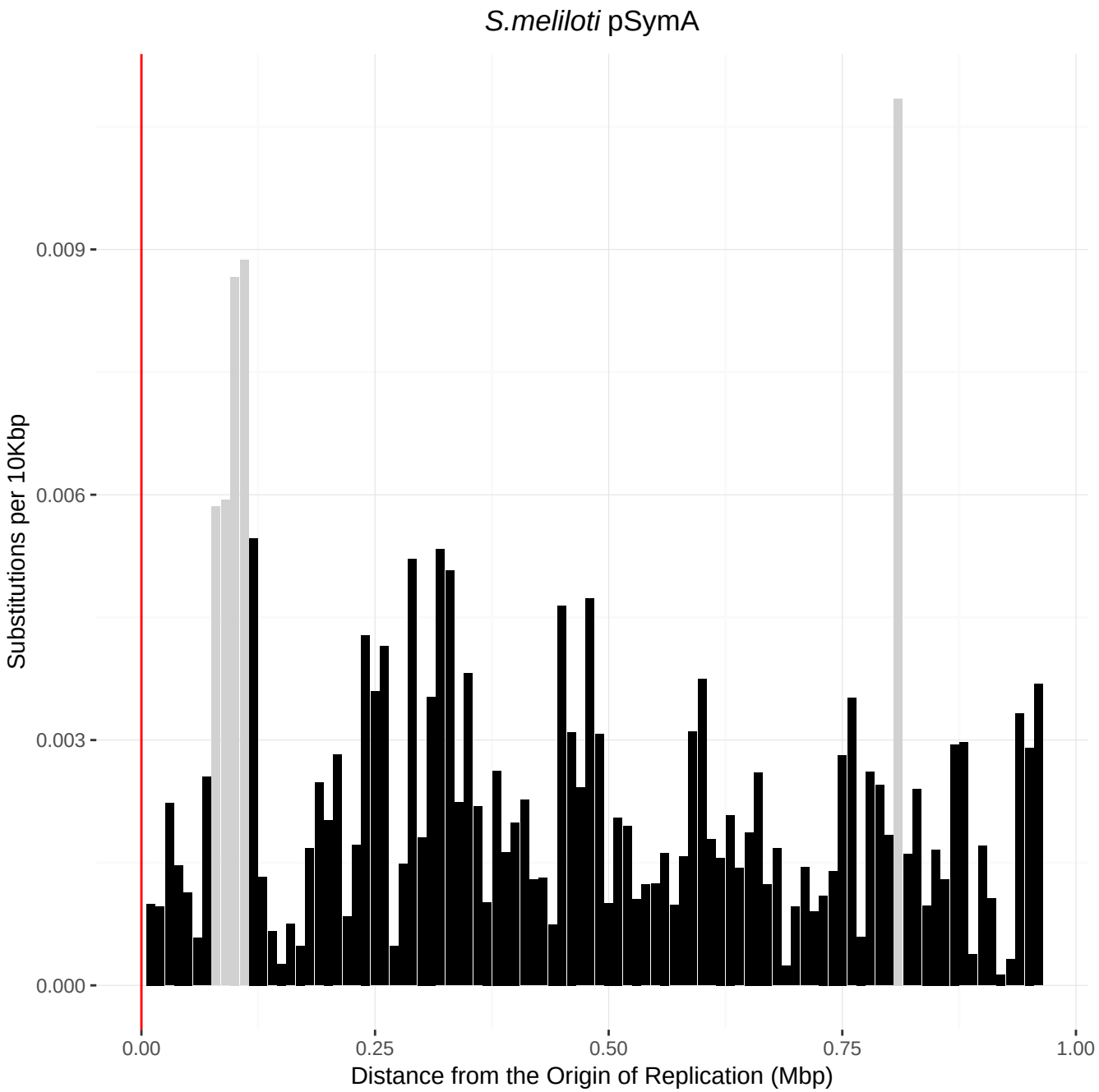
Table 1: Linear regression analysis of the median counts per million expression data for 10Kbp segments of the genome of the respective bacteria replicons. Linear regression was calculated after the origin of replication was moved to the beginning of the genome and all subsequent positions were scaled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

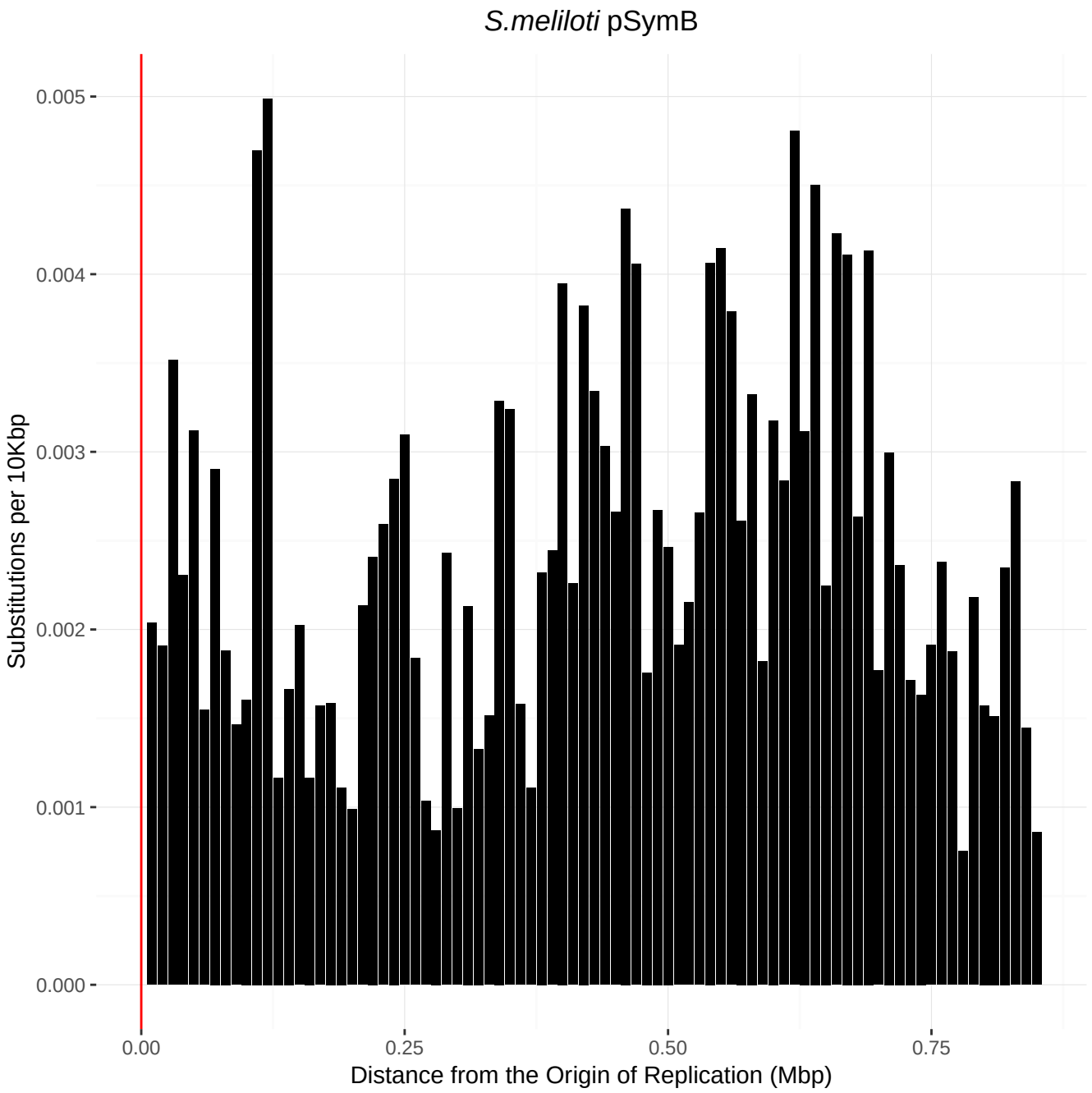

E.coli Chromosome
Substitutions per 10Kbp
0.02
0.01
0.00
0
1
2
3
Distance from the Origin of Replication (Mbp)


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
Substitutions per 10Kbp
0.04
0.02
0.00
−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
Substitutions per 10Kbp
0.009
0.006
0.003
0.000
0.00
0.25
0.50
0.75
1.00
Distance from the Origin of Replication (Mbp)


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

| Bacteria and Replicon | Protein Coding Sequences |
|---|---|
| *E. coli* Chromosome | $-2.98\times10^{-8}$*** |
| *B. subtilis* Chromosome | $-8.06\times10^{-8}$*** |
| *Streptomyces* Chromosome | $1.10\times10^{-7}$*** |
| *S. meliloti* Chromosome | $-4.32\times10^{-7}$*** |
| *S. meliloti* pSymA | $-5.18\times10^{-7}$*** |
| *S. meliloti* pSymB | $1.76\times10^{-7}$*** |

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 | $-1.17\times10^{-5}$** | NS | $5.91\times10^{-3}$ | $6.92\times10^{-3}$ |
| *B. subtilis* Chromosome | NS | $-8.96\times10^{-5}$*** | $1.97\times10^{-3}$ | $9.04\times10^{-3}$ |
| *Streptomyces* Chromosome | $7.45\times10^{-5}$*** | $-1.32\times10^{-4}$*** | $6.48\times10^{-4}$ | $6.73\times10^{-3}$ |
| *S. meliloti* Chromosome | $8.26\times10^{-5}$* | NS | $9.79\times10^{-5}$ | $5.07\times10^{-5}$ |
| *S. meliloti* pSymA | NS | NS | $9.80\times10^{-4}$ | $3.24\times10^{-3}$ |
| *S. meliloti* pSymB | $-1.42\times10^{-5}$* | $-6.32\times10^{-5}$*** | $1.97\times10^{-3}$ | $1.24\times10^{-3}$ |

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 | $-2.25\times10^{-10}$** | $-1.89\times10^{-4}$*** |
| *B. subtilis* Chromosome | $-8.35\times10^{-10}$** | $-1.91\times10^{-4}$** |
| *Streptomyces* Chromosome | $5.45\times10^{-11}$*** | NS |
| *S. meliloti* Chromosome | $-1.11\times10^{-10}$*** | $-1.28\times10^{-5}$*** |
| *S. meliloti* pSymA | NS | NS |
| *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 | $-2.51\times10^{-2}$*** |
| *B. subtilis* Chromosome | $-2.00\times10^{-2}$** |
| *Streptomyces* Chromosome | $-1.74\times10^{-3}$*** |
| *S. meliloti* Chromosome | $-1.88\times10^{-2}$*** |
| *S. meliloti* pSymA | $-2.50\times10^{-2}$** |
| *S. meliloti* pSymB | NS |

Table 5: Linear regression analysis of the total number of protein coding sites 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'.

| | Gene Average | | | Genome Average | | |
|---|---|---|---|---|---|---|
| Bacteria and Replicon | dS | dN | $\omega$ | dS | dN | $\omega$ |
| *E. coli* Chromosome | 0.2843 | 0.0145 | 0.0574 | 0.3072 | 0.0181 | 0.0663 |
| *B. subtilis* Chromosome | 0.6241 | 0.0353 | 0.0922 | 0.5708 | 0.0335 | 0.0841 |
| *Streptomyces* Chromosome | 0.0694 | 0.0062 | 0.6071 | 0.0615 | 0.0042 | 0.5877 |
| *S. meliloti* Chromosome | 0.0162 | 0.0011 | 0.1465 | 0.0171 | 0.0012 | 0.1158 |
| *S. meliloti* pSymA | 0.0865 | 0.0122 | 0.2250 | 0.0839 | 0.0106 | 0.2125 |
| *S. meliloti* pSymB | 3.2602 | 0.0256 | 0.3878 | 0.1436 | 0.0100 | 0.1943 |

Table 6: Weighted averages calculated for each bacterial replicon on a per genome basis using the gene length as the weight. Arithmetic mean calculated for the per gene averages for each bacterial replicon.