✓ Aug 21: Comprehensive Exam 10:30am

✓ Aug 26: make new list of dates for goals

✓ Sep 7: Write Up methods for clustering testing and add to substitutions paper

Sep 9: Gene expression data for the inversions project

Sep 14: Have all clustering testing complete for all bacteria

Sep 14: Compile notes from comps papers into one document

Sep 30: New intro for Substitution paper

Sep 15-28: Apply for NSERC (if applicable)

Oct 3: NSERC Due

Oct 5-12: Apply for Mac Scholarships and Awards

Oct 31: Write out methods for gene expression paper

Sep 9: Think about/compile list of inversions in *E. coli* for new paper

Nov 15: Think about how to better look at the COG data

Nov 25: Complete any extra analysis needed for Substitution paper

Dec 4: Mac Scholarships and Awards Due

Dec 1: Write out COG methods

Dec 15: Gather papers for COG paper intro

Dec 15: Implement COG stuff

Last Week

I was working on getting a good chunk of the clustering testing done. There is still a bit more to complete but it should be done by the end of next week!

I also wrote out detailed methods for the clustering testing and added it to the substitutions paper write-up. I also working on finding more gene expression data for the inversions and gene expression project.

This Week

I would like to continue to work on the clustering testing. In between having this finished I would like to keep looking for gene expression data for *Escherichia coli* and have this finished by the end of the week.

Next Week

I would like to compile all my notes from comps into one file so that I can use this to update my substitutions paper intro. I would like to complete the remainder of the clustering testing.

Position Difference	*E. coli* Chromosome	*B. subtilis* Chromosome	*Streptomyces* Chromosome	*S. meliloti* Chromosome	*S. meliloti* pSymA	*S. meliloti* pSymB
1bp	-1.394×10^{-7}**	-2.538×10^{-8}**	1.736×10^{-8}**	-1.541×10^{-6}**	-9.130×10^{-7}**	2.488×10^{-7}***
10bp	-1.394×10^{-7}***	-2.518×10^{-8}***	-4.484×10^{-9}***	-1.627×10^{-6}***	-9.13×10^{-7}***	3.487×10^{-7}***
100bp	-1.764×10^{-7}***	-1.417×10^{8}***				

Table 1: Logistic regression analysis of the number of substitutions along the genome of the respective bacteria replicons to test position differences. All results are marked with significance codes as followed: $< 0.001 = $ '***', $0.001 < 0.01 = $ '**', $0.01 < 0.05 = $ '*', $0.05 < 0.1 = $ '.', $> 0.1 = $ ' '. Logistic regression was calculated after the positions in the genome were determined to be the same at each position difference listed in the first column.

| Bacteria and Replicon	Coefficient Estimate	Standard Error	P-value
E. coli Chromosome	-6.41×10^{-5}	1.65×10^{-5}	1.1×10^{-4}
B. subtilis Chromosome	-9.9×10^{-5}	2.18×10^{-5}	6×10^{-6}
Streptomyces Chromosome	-1.5×10^{-6}	1.4×10^{-7}	$<2\times10^{-16}$
S. meliloti Chromosome	3.19×10^{-5}	3.57×10^{-5}	3.7×10^{-1}
S. meliloti pSymA	-5.36×10^{-5}	6.34×10^{-4}	9.33×10^{-1}
S. meliloti pSymB	5.05×10^{-4}	2.6×10^{-4}	5.3×10^{-2}

Table 2: 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.

3
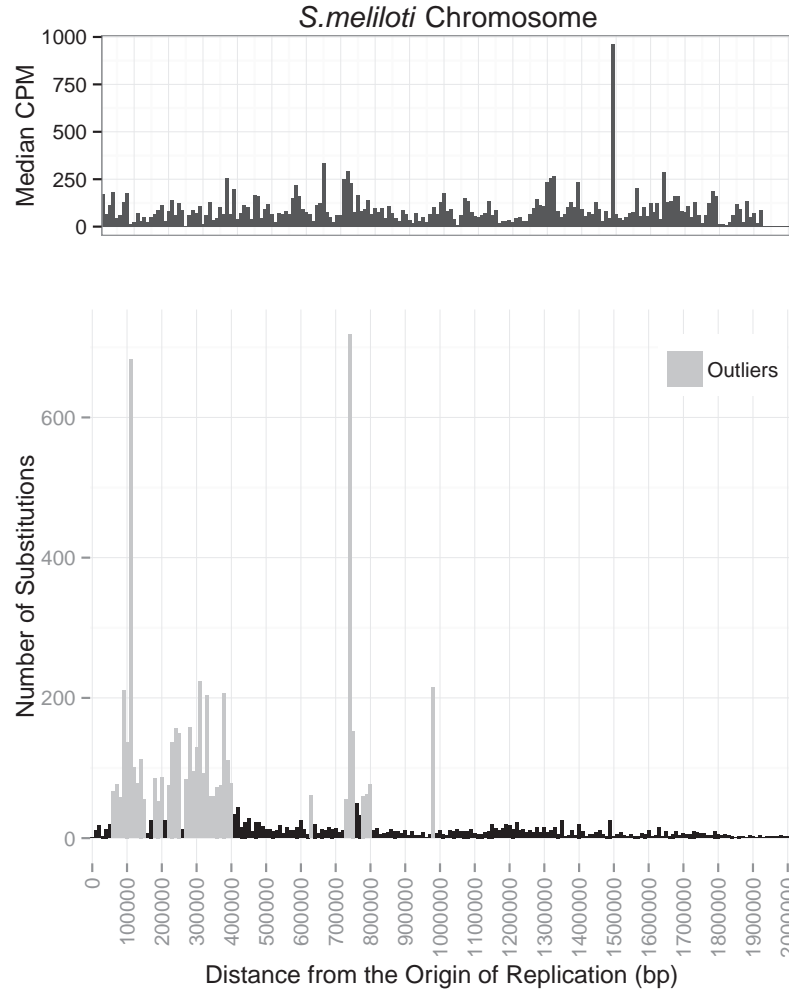
Bacteria and Replicon	Coefficient Estimate	Standard Error	P-value
E. coli Chromosome	-1.394×10^{-7}	2.425×10^{-9}	$<2 \times 10^{-16}$
B. subtilis Chromosome	-1.265×10^{-8}	1.562×10^{-9}	5.430×10^{-16}
Streptomyces Chromosome	1.736×10^{-8}	7.231×10^{-10}	$<2 \times 10^{-16}$
S. meliloti Chromosome	-1.541×10^{-6}	3.042×10^{-8}	$<2 \times 10^{-16}$
S. meliloti pSymA	-9.130×10^{-7}	1.975×10^{-8}	$<2 \times 10^{-16}$
S. meliloti pSymB	2.488×10^{-7}	1.964×10^{-8}	$<2 \times 10^{-16}$

Table 3: Logistic regression analysis of the number of substitutions along the genome of the respective bacteria replicons. 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.
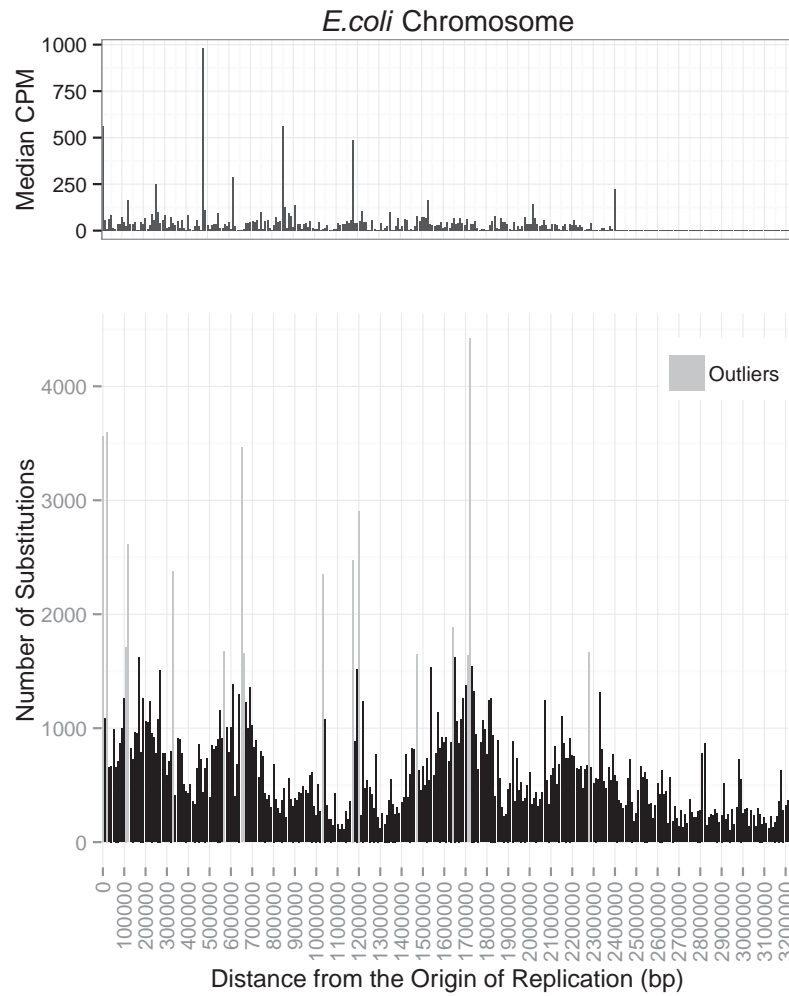
4

S.meliloti pSymA

S.meliloti pSymB

B.subtilis Chromosome

Streptomyces Chromosome
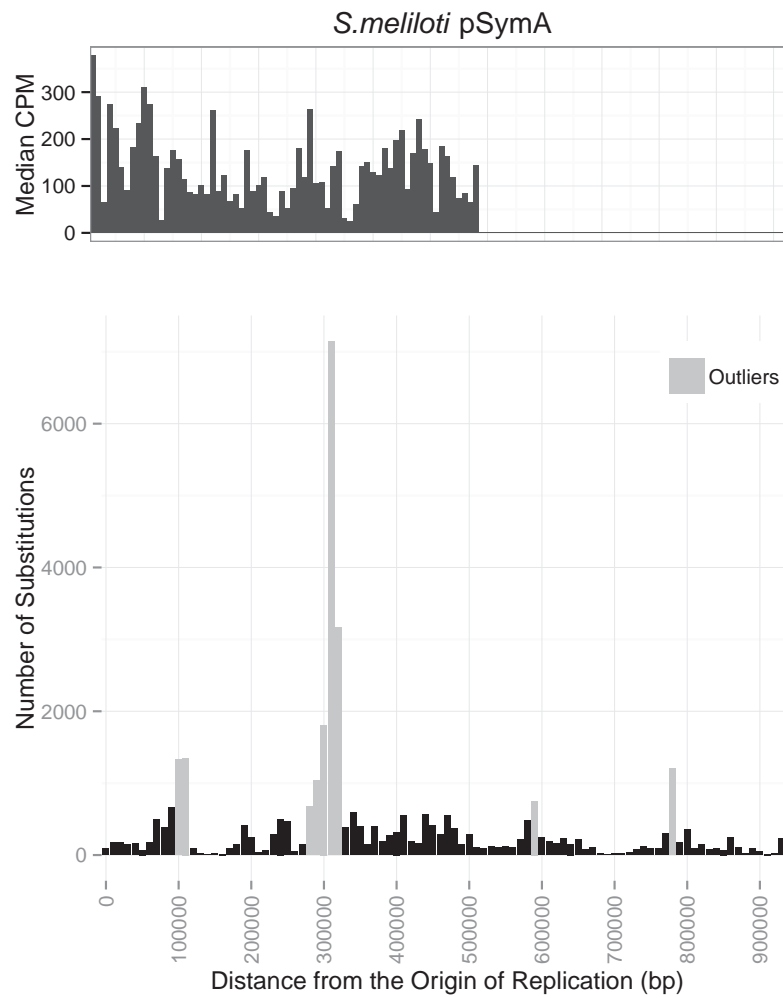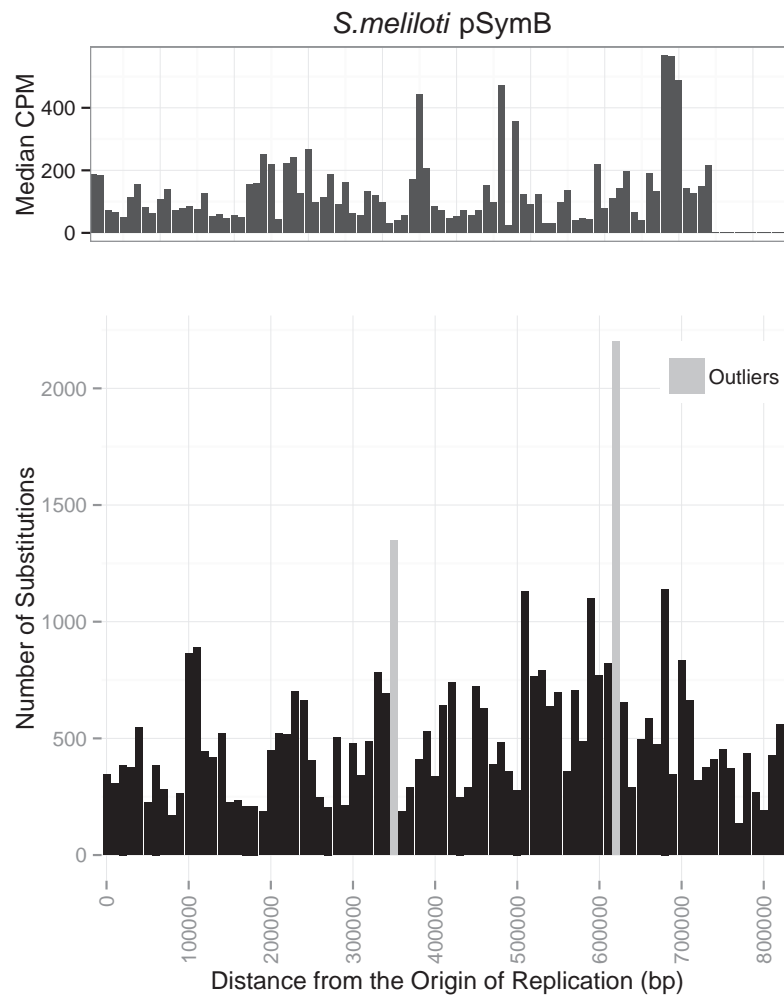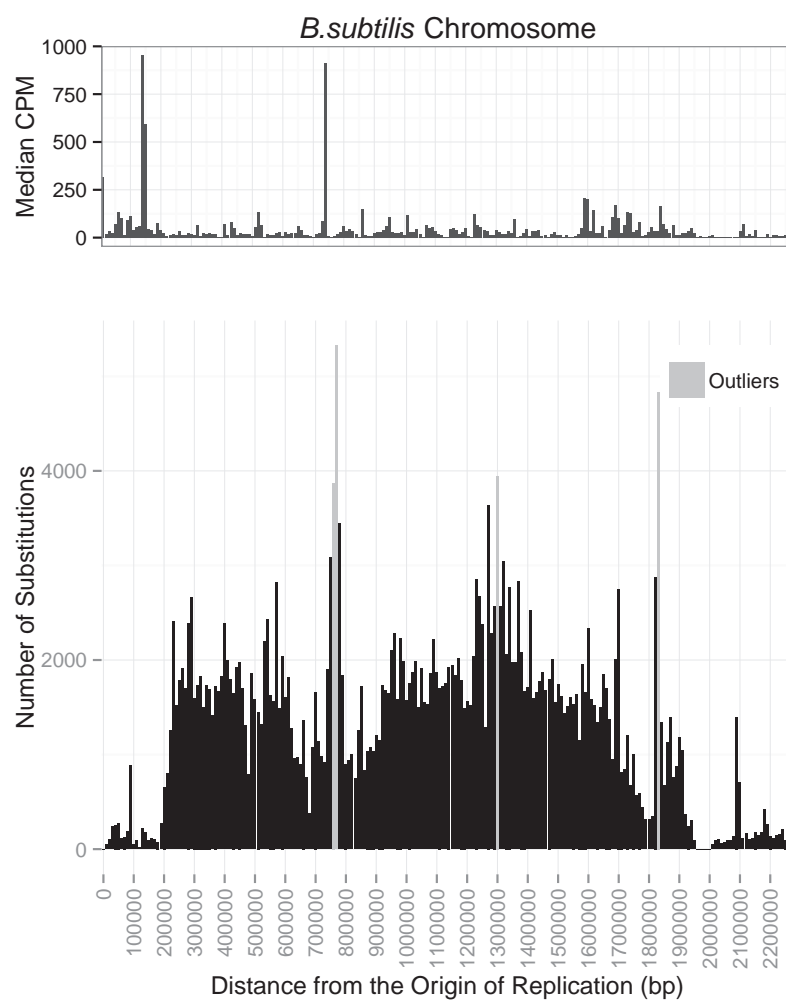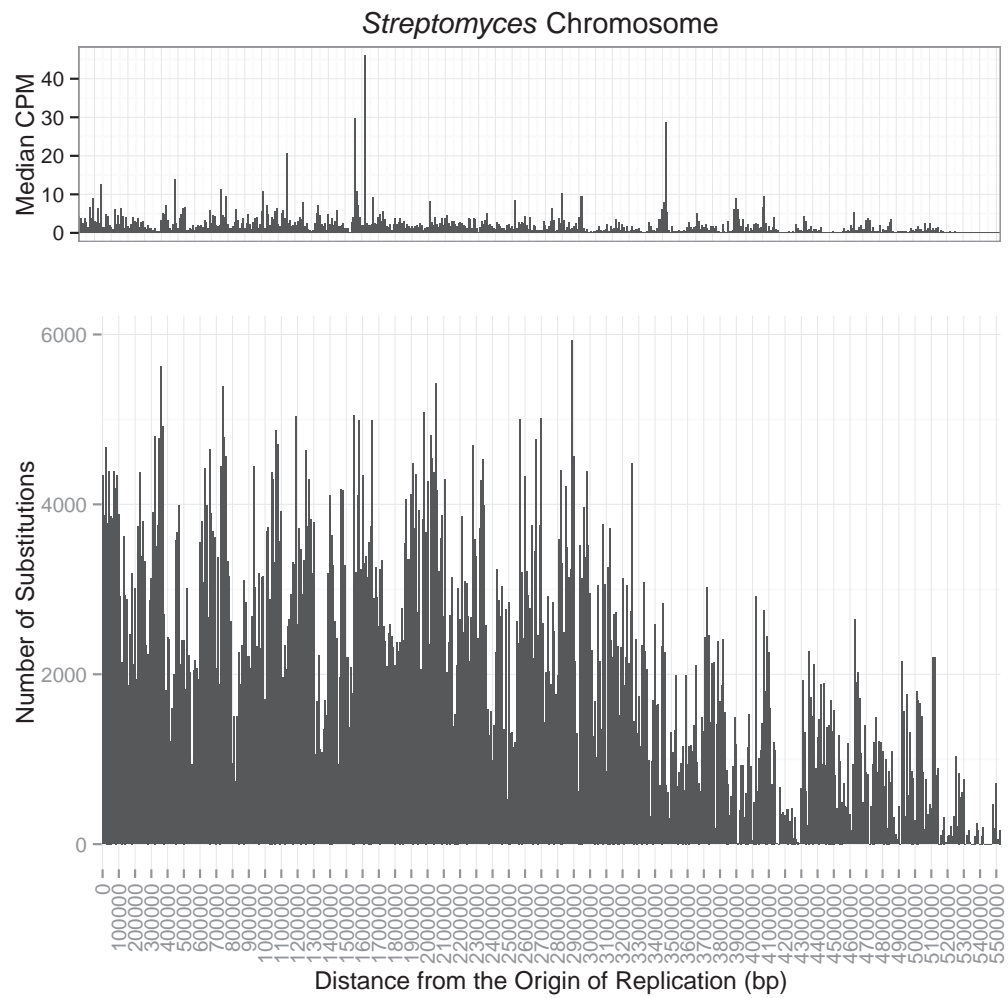
Origin Location	*E. coli* Chromosome	*B. subtilis* Chromosome	*Streptomyces* Chromosome	*S. meliloti* Chromosome	*S. meliloti* pSymA	*S. meliloti* pSymB
Moved 100kb Left	-1.445×10^{-7}***	4.374×10^{-9}*	6.909×10^{-9}***	-1.316×10^{-6}***	-1.058×10^{-6}***	-2.009×10^{-7}***
Moved 90kb Left	-1.544×10^{-7}***	-1.036×10^{-7}***	5.677×10^{-9}***	-1.32×10^{-6}***	-1.246×10^{-6}***	-1.357×10^{-7}***
Moved 80kb Left	-1.65×10^{-7}***	-1.072×10^{-7}***	8.11×10^{-9}***	-1.338×10^{-6}***	-1.398×10^{-6}***	-6.57×10^{-8}***
Moved 70kb Left	-1.667×10^{-7}***	-1.102×10^{-7}***	6.716×10^{-9}***	-1.363×10^{-6}***	-1.405×10^{-6}***	9.83×10^{-8}
Moved 60kb Left	-1.64×10^{-7}***	-1.19×10^{-7}***	8.7×10^{-9}***	-1.324×10^{-6}***	-1.394×10^{-6}***	1.129×10^{-7}***
Moved 50kb Left	-1.446×10^{-7}***	-1.211×10^{-7}***	1.045×10^{-8}***	-1.36×10^{-6}***	-1.403×10^{-6}***	1.521×10^{-7}***
Moved 40kb Left	-1.4×10^{-7}***	-1.299×10^{-7}***	1.214×10^{-8}***	-1.255×10^{-6}***	-1.422×10^{-6}***	1.543×10^{-7}***
Moved 30kb Left	-1.498×10^{-7}***	-1.292×10^{-7}***	1.24×10^{-8}***	-1.26×10^{-6}***	-1.392×10^{-6}***	1.63×10^{-7}***
Moved 20kb Left	-1.51×10^{-7}***	-1.1×10^{-7}***	1.395×10^{-8}***	-1.525×10^{-6}***	-1.412×10^{-6}***	1.603×10^{-7}***
Moved 10kb Left	-1.262×10^{-7}***	-2.602×10^{-9}	1.563×10^{-8}***	-1.599×10^{-6}***	-9.499×10^{-7}***	2.973×10^{-7}***
Moved 10kb Right	-1.305×10^{-7}***	-2.045×10^{-8}***	1.578×10^{-8}***	1.614×10^{-6}***	-1.026×10^{-6}***	3.505×10^{-7}***
Moved 20kb Right	-1.454×10^{-7}***	-1.006×10^{-7}***	1.903×10^{-8}***	-1.634×10^{-6}***	-1.475×10^{-6}***	1.649×10^{-7}***
Moved 30kb Right	-1.548×10^{-7}***	-8.596×10^{-8}***	2.046×10^{-8}***	-1.698×10^{-6}***	-1.417×10^{-6}***	1.526×10^{-7}***
Moved 40kb Right	-1.632×10^{-7}***	-8.378×10^{-8}***	2.125×10^{-8}***	-1.719×10^{-6}***	-1.367×10^{-6}***	1.589×10^{-7}***
Moved 50kb Right	-1.856×10^{-7}***	-7.879×10^{-8}***	1.957×10^{-8}***	-1.735×10^{-6}***	-1.277×10^{-6}***	1.654×10^{-7}***
Moved 60kb Right	-1.91×10^{-7}***	-6.98×10^{-8}***	1.974×10^{-8}***	-1.788×10^{-6}***	-1.169×10^{-6}***	1.645×10^{-7}***
Moved 70kb Right	-1.892×10^{-7}***	-6.634×10^{-8}***	1.934×10^{-8}***	-1.854×10^{-6}***	-1.059×10^{-6}***	1.843×10^{-7}***
Moved 80kb Right	-1.879×10^{-7}**	-5.814×10^{-8}***	2.313×10^{-8}***	-1.891×10^{-6}***	-9.07×10^{-7}***	1.90×10^{-7}***
Moved 90kb Right	-1.862×10^{-7}***	-4.314×10^{-8}***	2.304×10^{-8}***	-1.865×10^{-6}***	-7.171×10^{-7}***	2.415×10^{-7}***
Moved 100kb Right	-1.799×10^{-7}***	-2.597×10^{-8}***	1.945×10^{-8}***	-1.525×10^{-6}***	-6.572×10^{-7}***	3.095×10^{-7}***

Table 4: Logistic regression analysis of the number of substitutions along the genome of the respective bacteria replicons. All results are marked with significance codes as followed: $< 0.001 = $ '***', $0.001 < 0.01 = $ '**', $0.01 < 0.05 = $ '*', $0.05 < 0.1 = $ '.', $> 0.1 = $ ' '. 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.