✓ 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

X 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 14: Have Lab Meeting Presentation done

Sep 18: Present in Lab Meeting

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

The clustering testing has been very frustrating. A few weeks ago I combined all of my steps so that I would have to do less "checking in" on the processes. But in doing that, something has gone wrong and some of the data files are coming up empty. I have been trying to work this out all week and I still can not figure it out. This issue appears to only be happening for *S. meliloti* Chrom and pSymA. This is what I spent majority of my time on last week, trying to fix this issue, which is setting me back for the rest of the clustering analysis.

I tried to run as much as I could for the rest of the replicons, the current results are below.

I was also going through the list of bacteria you used in the PARSNP *Escherichia coli* alignment and seeing how many of those strains have usable gene expression data. I have about 30 more strains to look through.

This Week

I would like to continue to work on the clustering testing.

I will mostly be working on my lab meeting presentation which will be almost exactly the same as my comprehensive exam presentation.

I would also like to finish going though the *Escherichia coli* alignment to find gene expression data to use for the inversions and expression analysis.

I would also like to think about how to gather information on coding and non-coding sections of the genomes to incorporate this into the substitution analysis to help tease apart selection.

Next Week

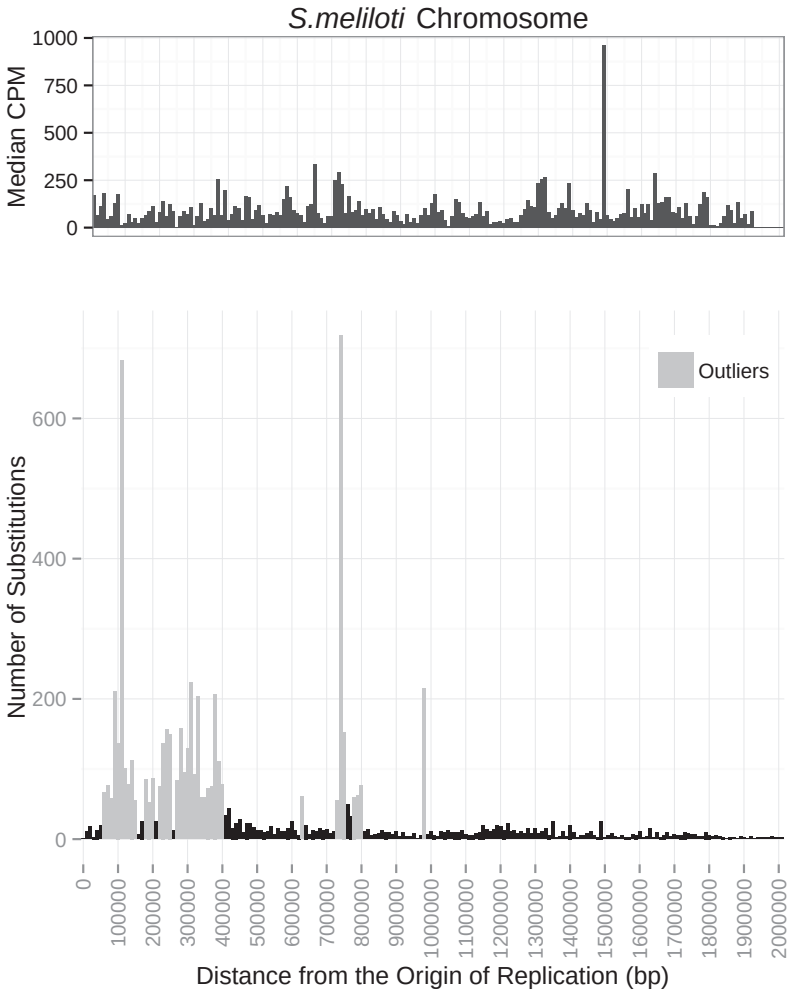
I would like to compile all my notes from comps into one file so that I can use this to update my substitutions paper intro. I would like to complete the remainder of the clustering testing.

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}***	1.448×10^{-8}***			
1000bp		-1.417×10^{-8}***			-1.153×10^{-6}***	4.021×10^{-7}***

Table 1: Position clustering analysis. Logistic regression analysis of the number of substitutions along the genome of the respective bacteria replicons to test position differences. Each row denotes different base pair distances that the positions were clustered together as. 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.

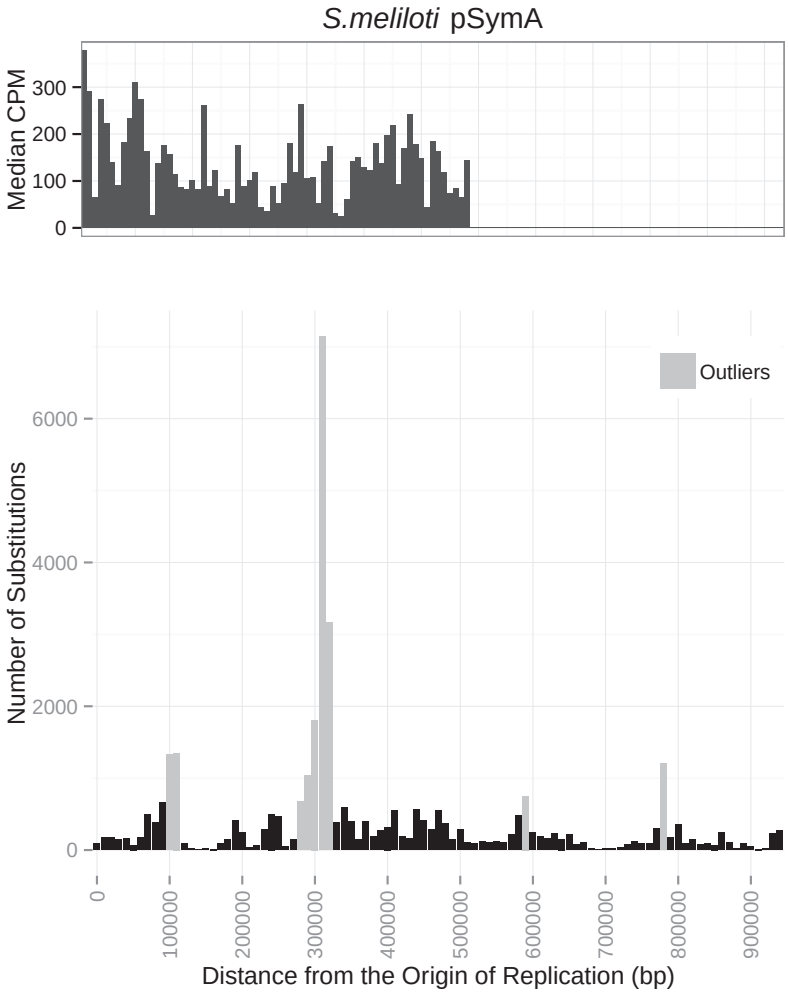
Bacteria and Replicon	Coefficient Estimate	Standard Error	P-value
E. coli Chromosome	-6.41×10^{-5}	1.65×10^{-5}	1.1×10^{-4}
B. subtilis Chromosome	-9.9×10^{-5}	2.18×10^{-5}	6×10^{-6}
Streptomyces Chromosome	-1.5×10^{-6}	1.4×10^{-7}	$<2\times10^{-16}$
S. meliloti Chromosome	3.19×10^{-5}	3.57×10^{-5}	3.7×10^{-1}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

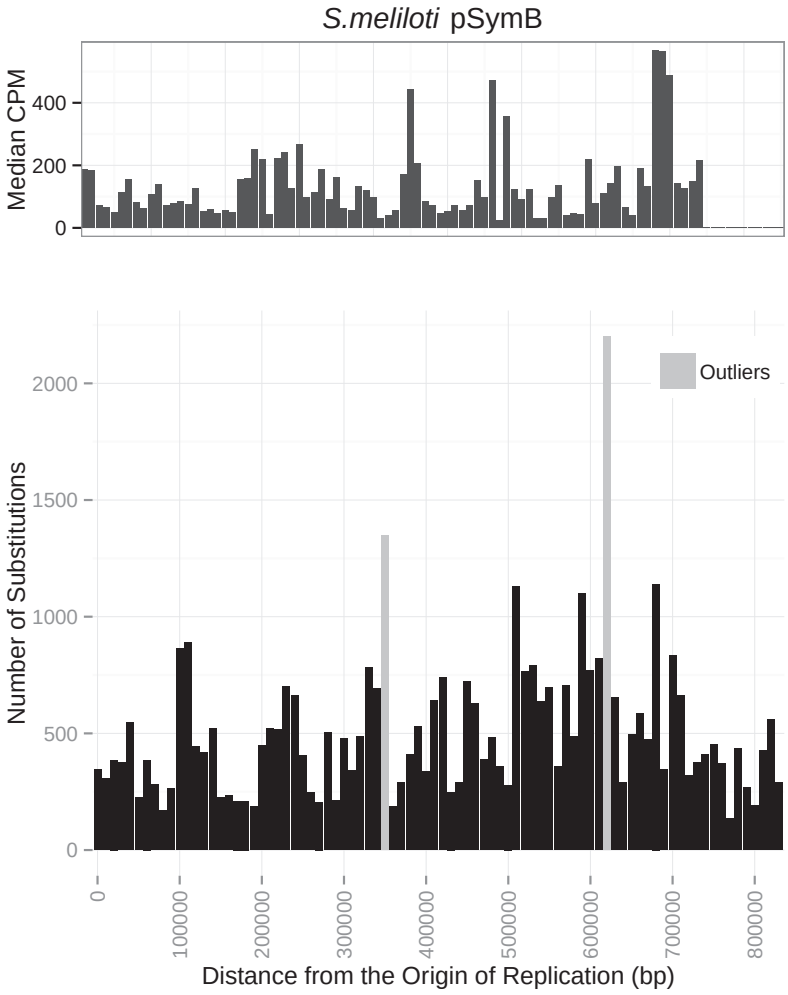
Table 2: Linear regression analysis of the median counts per million expression data along the genome of the respective bacteria replicons. Grey coloured boxes indicate statistically significant results at the 0.5 significance level. Linear regression was calculated after the origin of replication was moved to the beginning of the genome and all subsequent positions were scaled around the origin accounting for bidirectionality of replication.

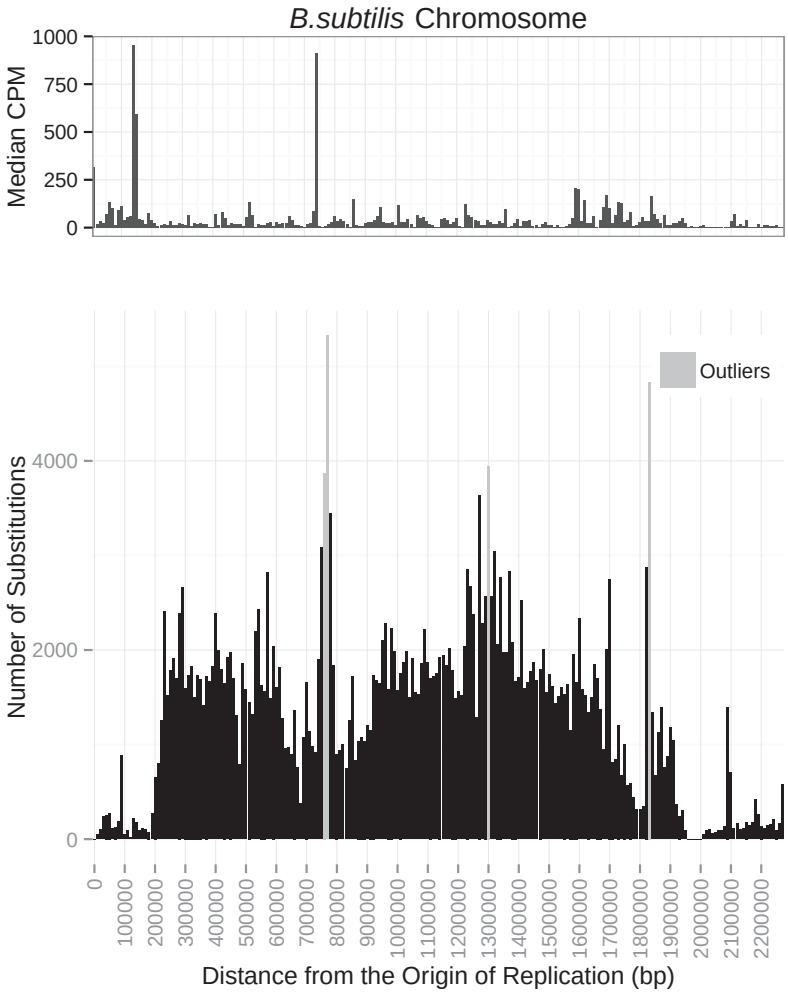


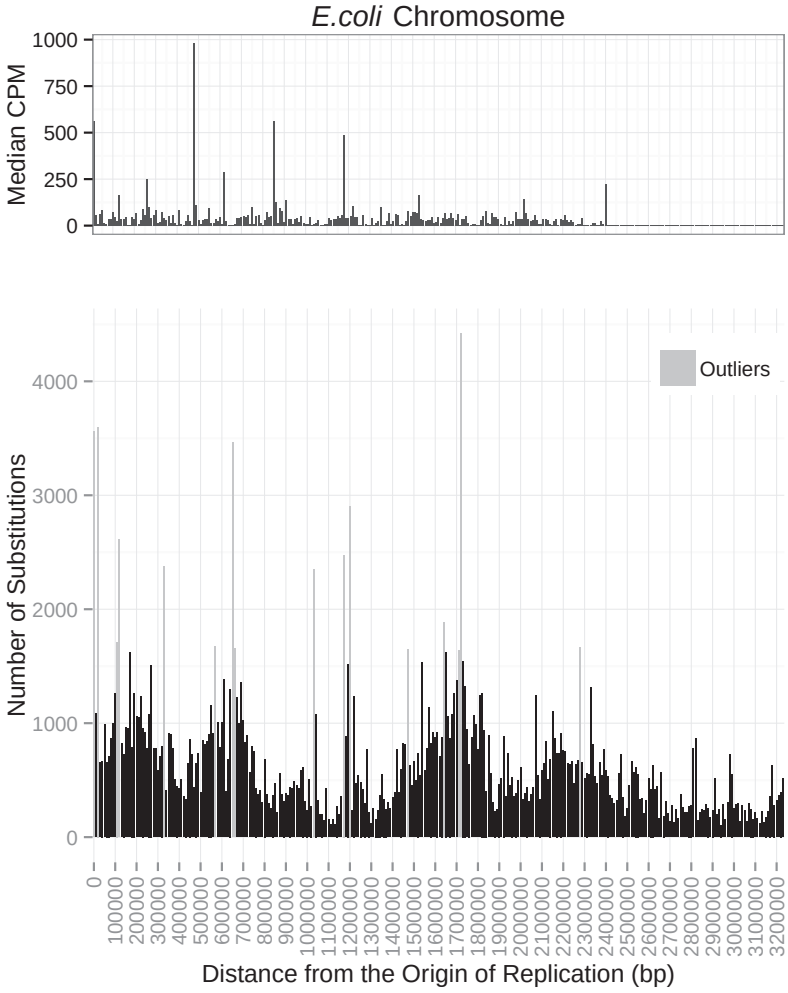
Bacteria and Replicon	Coefficient Estimate	Standard Error	P-value
E. coli Chromosome	-1.394×10^{-7}	2.425×10^{-9}	$<2\times10^{-16}$
B. subtilis Chromosome	-1.265×10^{-8}	1.562×10^{-9}	5.430×10^{-16}
Streptomyces Chromosome	1.736×10^{-8}	7.231×10^{-10}	$<2\times10^{-16}$
S. meliloti Chromosome	-1.541×10^{-6}	3.042×10^{-8}	$<2\times10^{-16}$
S. meliloti pSymA	-9.130×10^{-7}	1.975×10^{-8}	$<2\times10^{-16}$
S. meliloti pSymB	2.488×10^{-7}	1.964×10^{-8}	$<2\times10^{-16}$

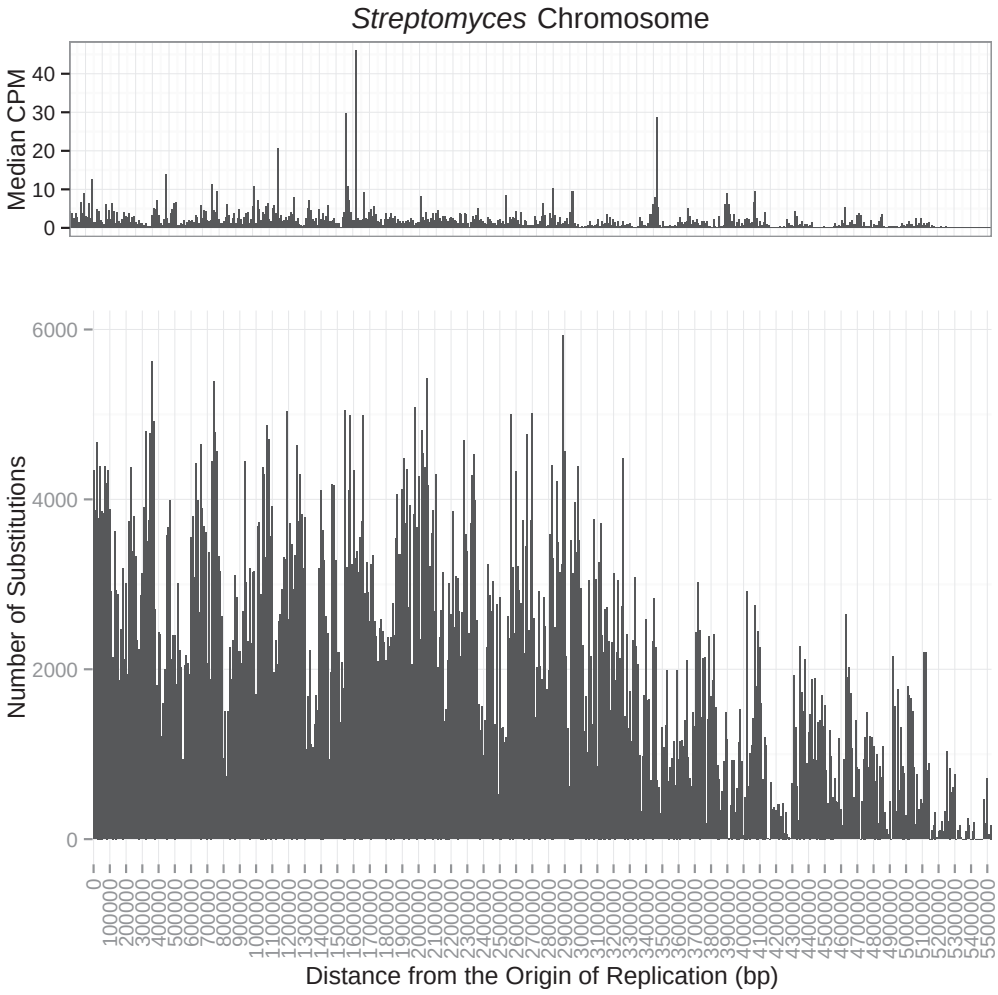
Table 3: Logistic regression analysis of the number of substitutions along the genome of the respective bacteria replicons. Grey coloured boxes indicate a negative logistic regression coefficient estimate. All results are statistically significant. Logistic regression was calculated after the origin of replication was moved to the beginning of the genome and all subsequent positions were scaled around the origin accounting for bidirectionality of replication.


S.meliloti pSymA
Median CPM
300
200
100
0
Outliers
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
Median CPM
400
200
0
Outliers
2000
1500
1000
500
0
Number of Substitutions
0
100000
200000
300000
400000
500000
600000
700000
800000
Distance from the Origin of Replication (bp)


B.subtilis Chromosome
Median CPM
1000
750
500
250
0
Outliers
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)


E.coli Chromosome
Median CPM
1000
750
500
250
0
Outliers
Number of Substitutions
4000
3000
2000
1000
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
Distance from the Origin of Replication (bp)


Streptomyces Chromosome
Median CPM
0
10
20
30
40
Number of Substitutions
0
2000
4000
6000
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)

Origin Location	*E. coli* Chromosome	*B. subtilis* Chromosome	*Streptomyces* Chromosome	*S. meliloti* Chromosome	*S. meliloti* pSymA	*S. meliloti* pSymB
Moved 100kb Left	$-1.445\times10^{-7***}$	$4.374\times10^{-9*}$	$6.909\times10^{-9***}$	$-1.316\times10^{-6***}$	$-1.058\times10^{-6***}$	$-2.009\times10^{-7***}$
Moved 90kb Left	$-1.544\times10^{-7***}$	$-1.036\times10^{-7***}$	$5.677\times10^{-9***}$	$-1.32\times10^{-6***}$	$-1.246\times10^{-6***}$	$-1.357\times10^{-7***}$
Moved 80kb Left	$-1.65\times10^{-7***}$	$-1.072\times10^{-7***}$	$8.11\times10^{-9***}$	$-1.338\times10^{-6***}$	$-1.398\times10^{-6***}$	$-6.57\times10^{-8***}$
Moved 70kb Left	$-1.667\times10^{-7***}$	$-1.102\times10^{-7***}$	$6.716\times10^{-9***}$	$-1.363\times10^{-6***}$	$-1.405\times10^{-6***}$	9.83×10^{-8}
Moved 60kb Left	$-1.64\times10^{-7***}$	$-1.19\times10^{-7***}$	$8.7\times10^{-9***}$	$-1.324\times10^{-6***}$	$-1.394\times10^{-6***}$	$1.129\times10^{-7***}$
Moved 50kb Left	$-1.446\times10^{-7***}$	$-1.211\times10^{-7***}$	$1.045\times10^{-8***}$	$-1.36\times10^{-6***}$	$-1.403\times10^{-6***}$	$1.521\times10^{-7***}$
Moved 40kb Left	$-1.4\times10^{-7***}$	$-1.299\times10^{-7***}$	$1.214\times10^{-8***}$	$-1.255\times10^{-6***}$	$-1.422\times10^{-6***}$	$1.543\times10^{-7***}$
Moved 30kb Left	$-1.498\times10^{-7***}$	$-1.292\times10^{-7***}$	$1.24\times10^{-8***}$	$-1.26\times10^{-6***}$	$-1.392\times10^{-6***}$	$1.63\times10^{-7***}$
Moved 20kb Left	$-1.51\times10^{-7***}$	$-1.1\times10^{-7***}$	$1.395\times10^{-8***}$	$-1.525\times10^{-6***}$	$-1.412\times10^{-6***}$	$1.603\times10^{-7***}$
Moved 10kb Left	$-1.262\times10^{-7***}$	-2.602×10^{-9}	$1.563\times10^{-8***}$	$-1.599\times10^{-6***}$	$-9.499\times10^{-7***}$	$2.973\times10^{-7***}$
Moved 10kb Right	$-1.305\times10^{-7***}$	$-2.045\times10^{-8***}$	$1.578\times10^{-8***}$	$1.614\times10^{-6***}$	$-1.026\times10^{-6***}$	$3.505\times10^{-7***}$
Moved 20kb Right	$-1.454\times10^{-7***}$	$-1.006\times10^{-7***}$	$1.903\times10^{-8***}$	$-1.634\times10^{-6***}$	$-1.475\times10^{-6***}$	$1.649\times10^{-7***}$
Moved 30kb Right	$-1.548\times10^{-7***}$	$-8.596\times10^{-8***}$	$2.046\times10^{-8***}$	$-1.698\times10^{-6***}$	$-1.417\times10^{-6***}$	$1.526\times10^{-7***}$
Moved 40kb Right	$-1.632\times10^{-7***}$	$-8.378\times10^{-8***}$	$2.125\times10^{-8***}$	$-1.719\times10^{-6***}$	$-1.367\times10^{-6***}$	$1.589\times10^{-7***}$
Moved 50kb Right	$-1.856\times10^{-7***}$	$-7.879\times10^{-8***}$	$1.957\times10^{-8***}$	$-1.735\times10^{-6***}$	$-1.277\times10^{-6***}$	$1.654\times10^{-7***}$
Moved 60kb Right	$-1.91\times10^{-7***}$	$-6.98\times10^{-8***}$	$1.974\times10^{-8***}$	$-1.788\times10^{-6***}$	$-1.169\times10^{-6***}$	$1.645\times10^{-7***}$
Moved 70kb Right	$-1.892\times10^{-7***}$	$-6.634\times10^{-8***}$	$1.934\times10^{-8***}$	$-1.854\times10^{-6***}$	$-1.059\times10^{-6***}$	$1.843\times10^{-7***}$
Moved 80kb Right	$-1.879\times10^{-7**}$	$-5.814\times10^{-8***}$	$2.313\times10^{-8***}$	$-1.891\times10^{-6***}$	$-9.07\times10^{-7***}$	$1.90\times10^{-7***}$
Moved 90kb Right	$-1.862\times10^{-7***}$	$-4.314\times10^{-8***}$	$2.304\times10^{-8***}$	$-1.865\times10^{-6***}$	$-7.171\times10^{-7***}$	$2.415\times10^{-7***}$
Moved 100kb Right	$-1.799\times10^{-7***}$	$-2.597\times10^{-8***}$	$1.945\times10^{-8***}$	$-1.525\times10^{-6***}$	$-6.572\times10^{-7***}$	$3.095\times10^{-7***}$

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.