X Jan 6: Write up methods for COG paper

✓ May 7: Revise summer goals if accepted for Chicago Conference

✓ May 11: Find gene expression papers for bacteria and specific bacteria, printed

✓ May 25: Read above papers and make notes (one a day?)

June 8: Have process down for testing position clustering

June 22: Have all clustering testing complete for all bacteria

June 12-29: Have first draft of ISMB presentation done (and present for the lab)/ prepare for conference questions

July 5: Have final edits for ISMB presentation finished

July 6-13: ISB Chicago Conference

July 20: Have date booked for Comps

July 20: Think about/compile list of inversions in *E. coli* for new paper

July 31: Gather gene expression data for the above mentioned *E. coli* strains

July 16 - August 31: Prepare for Comps

Last Week

Last week I was testing the ancestral position reconstruction code you developed. We determined that PAML does an acceptable job at reconstructing the ancestral positions the only thing that it does not know how to handle is clustering the positions, a.k.a how many base pairs apart do two positions have to be in order for them to be considered the same? So, you wrote another code that only clusters positions based on specified base pair differences. I was testing this to ensure that it worked as it was supposed to and it does! I also began to look at the Parsnp *Escherichia coli* alignment and started thinking about how to identify inversions in an automated and efficient manner. You suggested that the first thing to do is to gather gene expression data for all the *Escherichia coli* strains used and get as much as possible and then worry about identifying inversions.

This Week

I am working on choosing a date for my Comprehensive exam so I can begin preparing for that. I plan on having an automated process complete for testing the different position clustering that way I can just run this a million times on all the bacteria. If there is time I would like to begin to gather gene expression data for this inversion paper.

Next Week

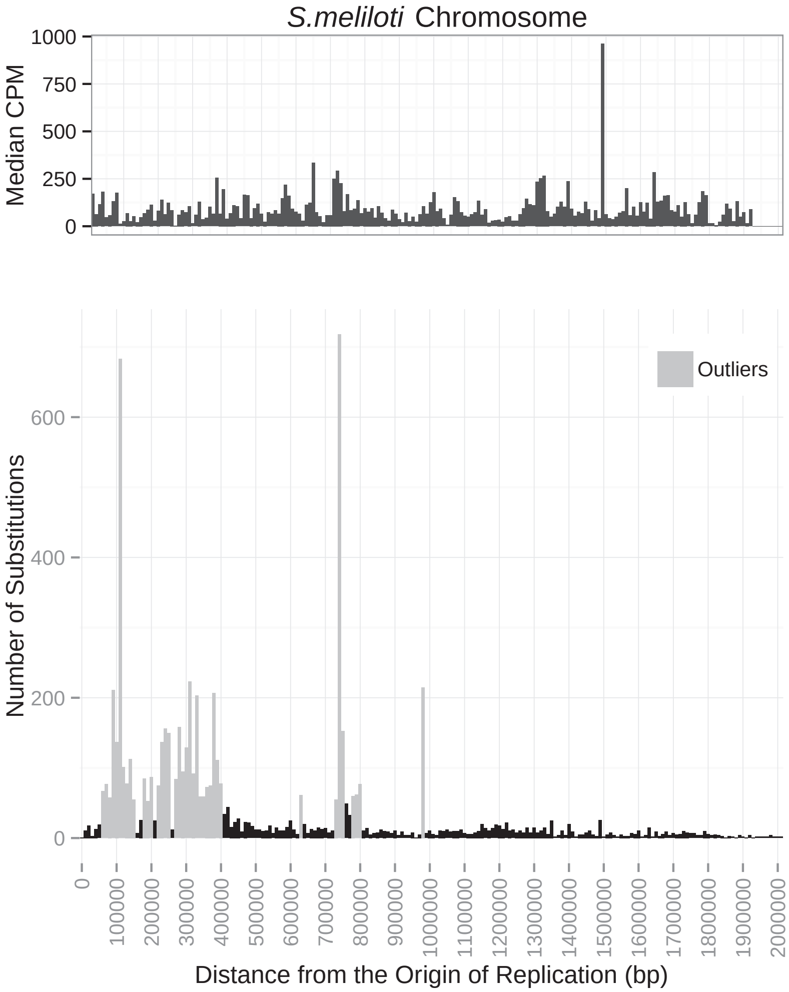
I would like to run the position clustering testing on all the bacteria and begin compiling this information. I would like to start gathering gene expression data for the inversion paper.

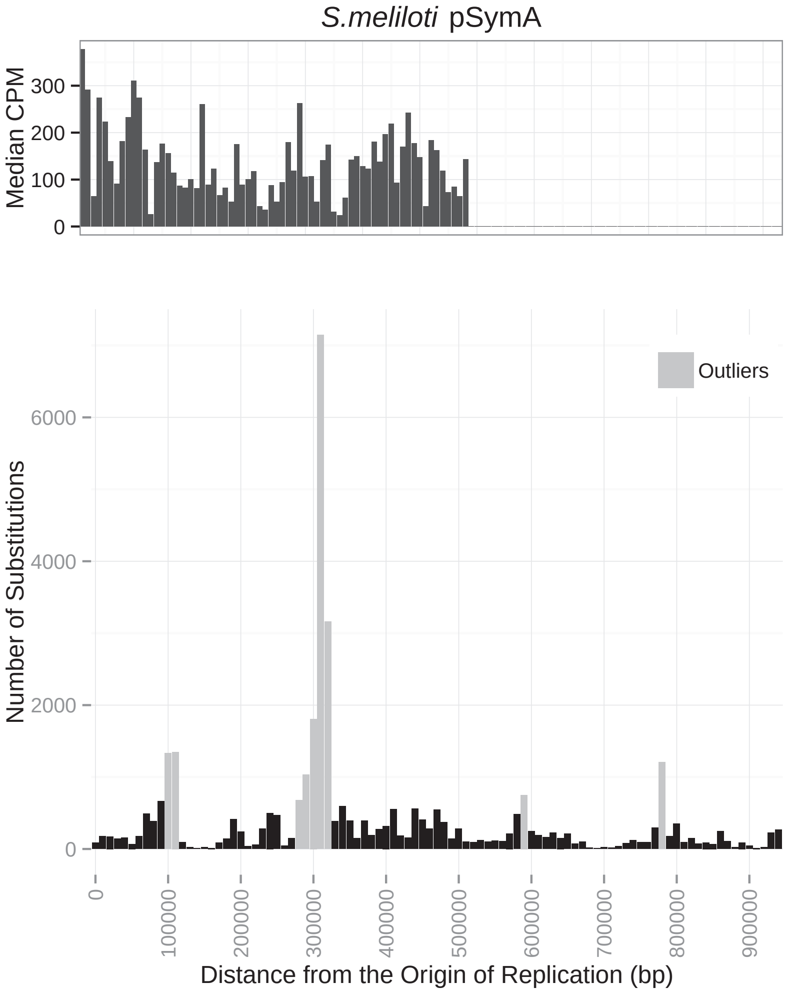
Bacteria and Replicon	Coefficient Estimate	Standard Error	P-value
E. coli Chromosome	-6.41×10^{-5}	1.65×10^{-5}	1.1×10^{-4}
B. subtilis Chromosome	-9.9×10^{-5}	2.18×10^{-5}	6×10^{-6}
Streptomyces Chromosome	-1.5×10^{-6}	1.4×10^{-7}	$<2\times10^{-16}$
S. meliloti Chromosome	3.19×10^{-5}	3.57×10^{-5}	3.7×10^{-1}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

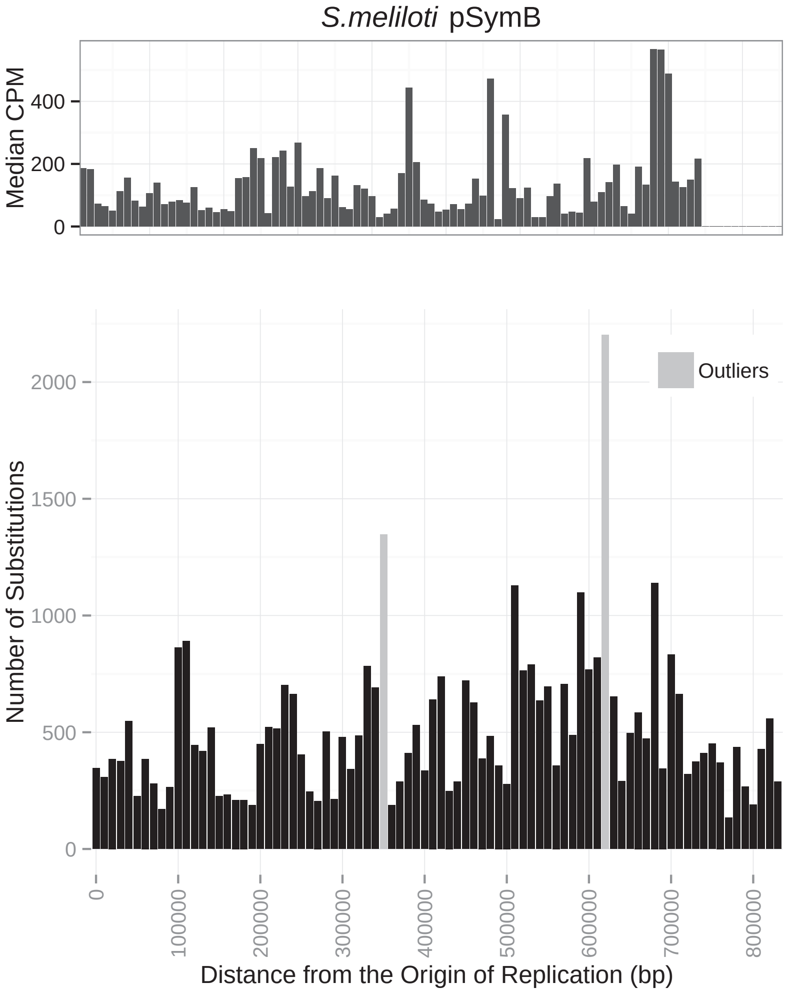
Table 1: Linear regression analysis of the median counts per million expression data along the genome of the respective bacteria replicons. Grey coloured boxes indicate statistically significant results at the 0.5 significance level. Linear regression was calculated after the origin of replication was moved to the beginning of the genome and all subsequent positions were scaled around the origin accounting for bidirectionality of replication.

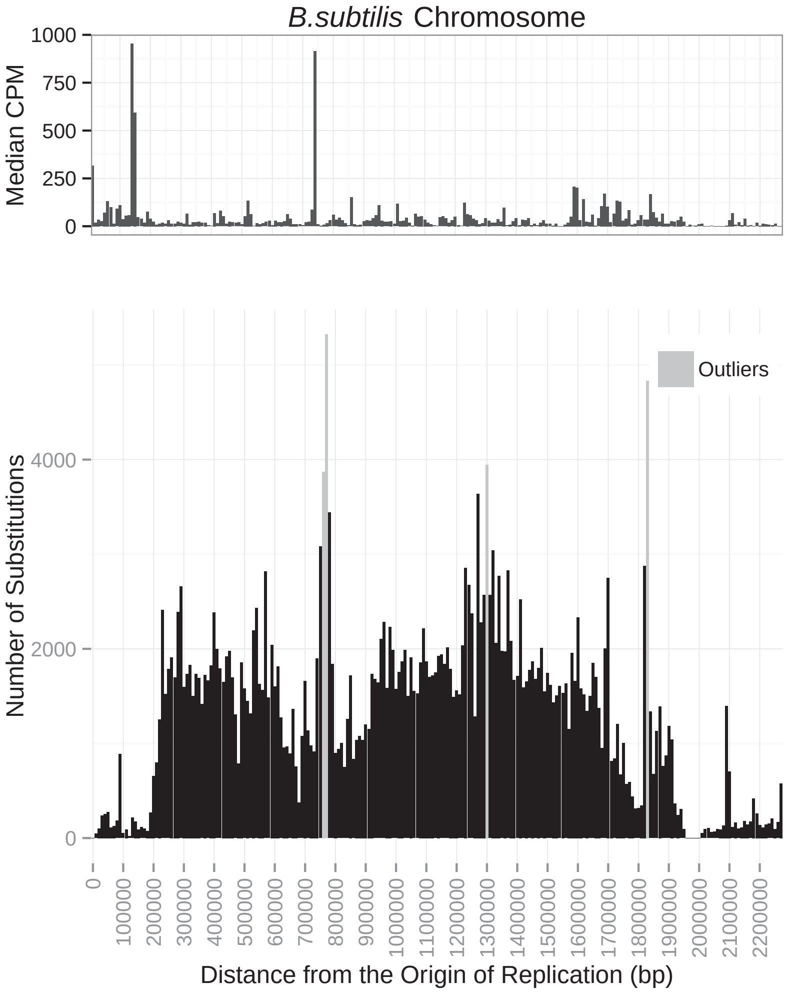
Bacteria and Replicon	Coefficient Estimate	Standard Error	P-value
E. coli Chromosome	-1.394×10^{-7}	2.425×10^{-9}	$<2\times10^{-16}$
B. subtilis Chromosome	-2.538×10^{-8}	1.58×10^{-9}	$<2\times10^{-16}$
Streptomyces Chromosome	1.736×10^{-8}	7.231×10^{-10}	$<2\times10^{-16}$
S. meliloti Chromosome	-1.541×10^{-6}	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}	$<2\times10^{-16}$

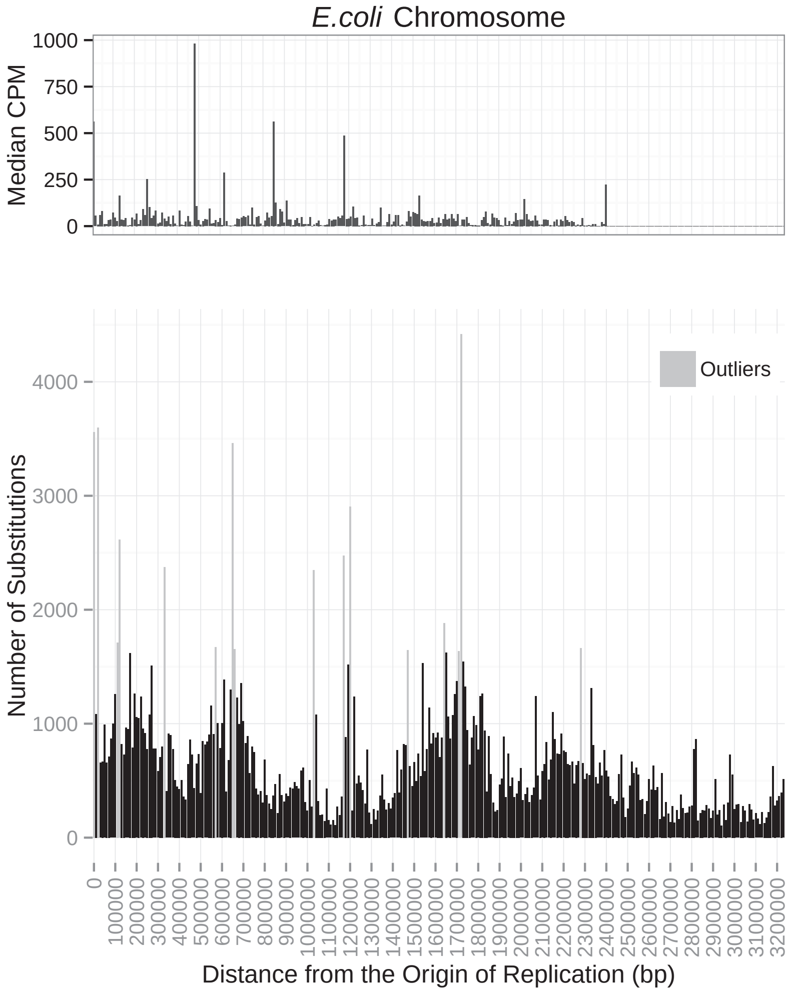
Table 2: Logistic regression analysis of the number of substitutions along the genome of the respective bacteria replicons. Grey coloured boxes indicate a negative logistic regression coefficient estimate. All results are statistically significant. Logistic regression was calculated after the origin of replication was moved to the beginning of the genome and all subsequent positions were scaled around the origin accounting for bidirectionality of replication.

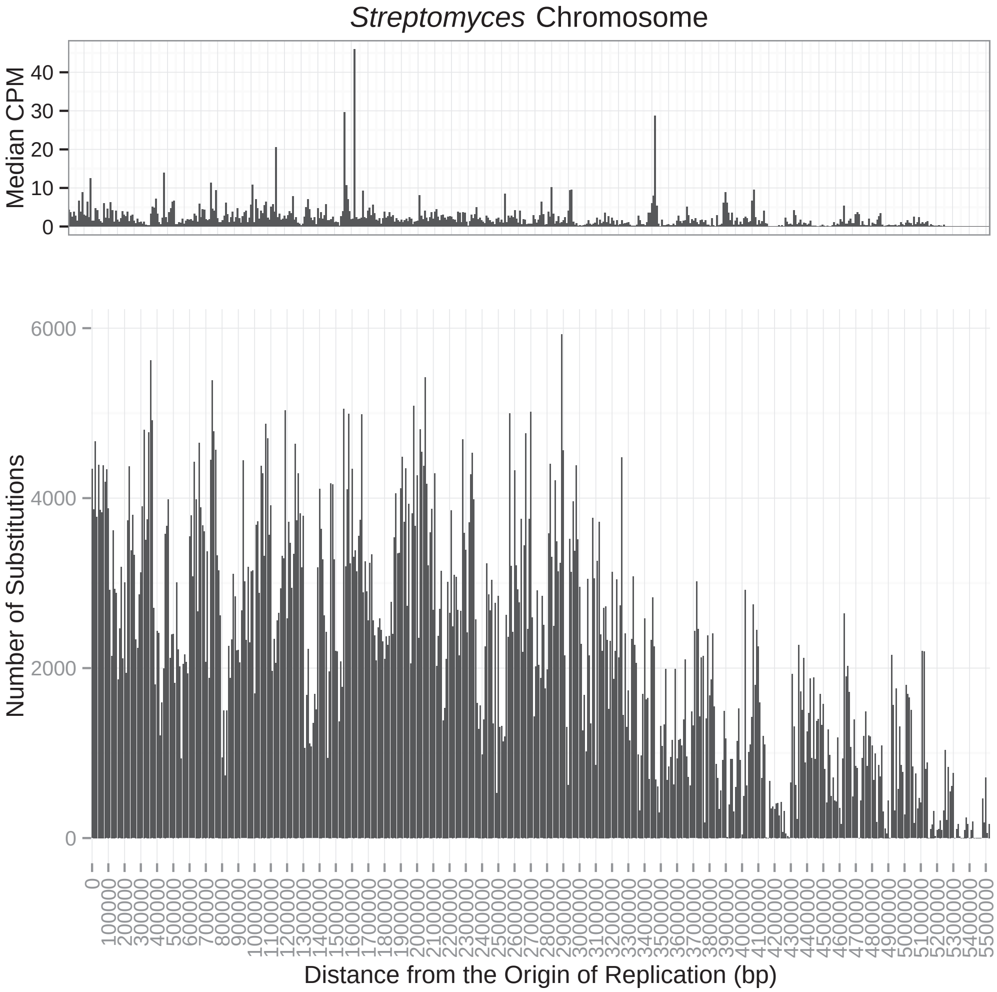

S.meliloti Chromosome
Median CPM
1000
750
500
250
0
Outliers
Number of Substitutions
600
400
200
0
0
100000
200000
300000
400000
500000
600000
700000
800000
900000
1000000
1100000
1200000
1300000
1400000
1500000
1600000
1700000
1800000
1900000
2000000
Distance from the Origin of Replication (bp)


S.meliloti pSymA
Median CPM
300
200
100
0
Outliers
6000
4000
2000
0
Number of Substitutions
0
100000
200000
300000
400000
500000
600000
700000
800000
900000
Distance from the Origin of Replication (bp)


S.meliloti pSymB
Median CPM
0
200
400
Number of Substitutions
0
500
1000
1500
2000
Outliers
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
Number of Substitutions
4000
2000
0
Outliers
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 3: 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.