

- ✓ 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
- X Sep 14: Have all clustering testing complete for all bacteria
- X 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

I was running the position clustering and I am almost done. The updated results are below. This is taking a bit longer than I expected it to.

I spent majority of last week figuring out how to accurately obtain the coding and non-coding sequences. I wanted to check how often the following happened: within the same column of the alignment, is there at least one taxa that is coding while the rest are non-coding? Or vice versa? Initially when I checked this I was seeing HUGE discrepancies. There would be 2000bp where one taxa was non-coding and the rest were coding. However, after looking into this more carefully all weekend, I found that this “discrepancy” is mostly consisting of gaps and therefore gets thrown out in my analysis. Where there are no gaps, and there is still one taxa that is non-coding while the rest are coding, the sequence alignment is nearly identical between all taxa. These same results are seen when I look at other blocks with “discrepancies”. So, this makes me think that it may be mostly annotation error and that the sections that are “non-coding” are actually coding but have not been annotated properly. Another explanation could be that the promoter for some reason got mutated and is not transcribed any more, but because I see near perfect sequence identity between all taxa in pretty much all the blocks, I suspect it is poor annotation and not the promoter.

I also spent quite a few hours last week doing administrative stuff like figuring out and testing the assignment submission for 3S03, and helping Jennifer from Bio 720.

This Week

I would like to continue to work on the clustering testing, and I think that I can get it all done by the end of next week.

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

I would also like to start writing code to gather info about the coding and non-coding sections of the alignments for the substitution analysis.

Next Week

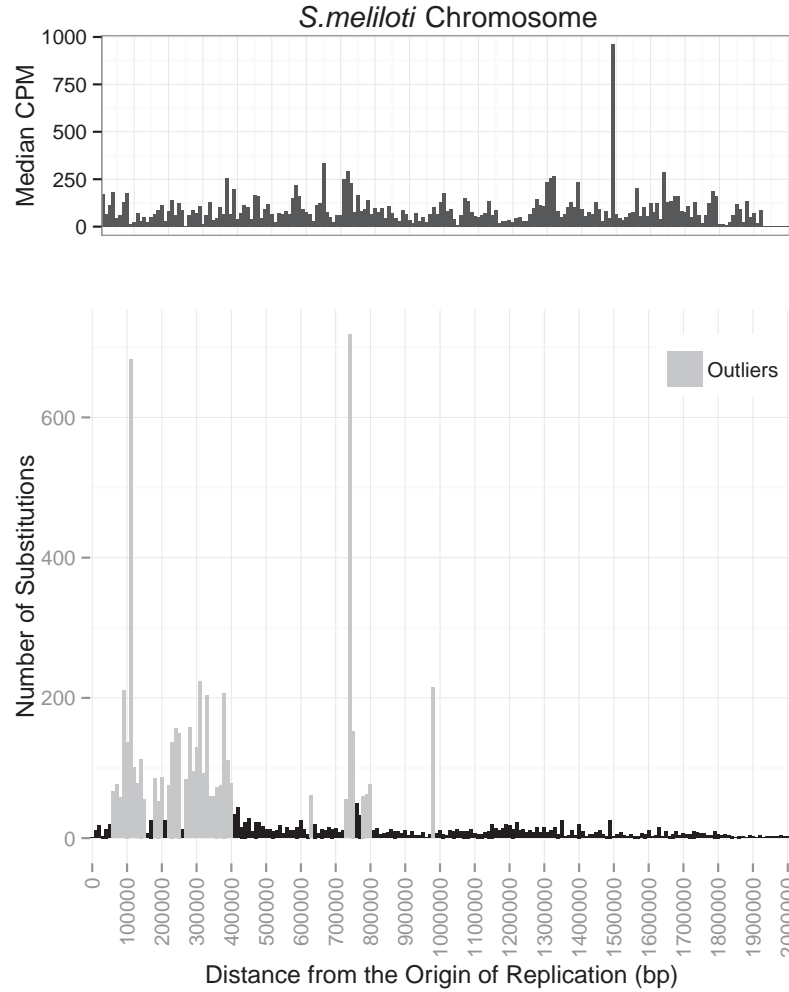
I would like to compile all my notes from comps into one file so that I can use this to update my substitutions paper intro. I would like to keep working on the coding v.s. non-coding stuff. I would like to complete the remainder of the clustering testing.

Position Difference	<i>E. coli</i> Chromosome	<i>B. subtilis</i> Chromosome	<i>Streptomyces</i> Chromosome	<i>S. meliloti</i> Chromosome	<i>S. meliloti</i> pSymA	<i>S. meliloti</i> pSymB
1bp	-1.394 $\times 10^{-7}$ ***	-2.538 $\times 10^{-8}$ ***	1.736 $\times 10^{-8}$ ***	-1.541 $\times 10^{-6}$ ***	-9.130 $\times 10^{-7}$ ***	2.488 $\times 10^{-7}$ ***
10bp	-1.394 $\times 10^{-7}$ ***	-2.518 $\times 10^{-8}$ ***	-4.484 $\times 10^{-9}$ ***	-1.627 $\times 10^{-6}$ ***	-9.13 $\times 10^{-7}$ ***	3.487 $\times 10^{-7}$ ***
100bp	-1.764 $\times 10^{-7}$ ***	-1.417 $\times 10^{-8}$ ***	1.448 $\times 10^{-8}$ ***		-1.166 $\times 10^{-6}$ ***	4.021 $\times 10^{-7}$ ***
1000bp	-1.784 $\times 10^{-7}$ ***	-1.417 $\times 10^{-8}$ ***	1.505 $\times 10^{-8}$ ***	-1.605 $\times 10^{-6}$ ***	-1.153 $\times 10^{-6}$ ***	4.021 $\times 10^{-7}$ ***
10000bp	-1.712 $\times 10^{-7}$ ***	-3.496 $\times 10^{-8}$ ***			-3.570 $\times 10^{-8}$ *	3.784 $\times 10^{-7}$ ***
100000bp						3.784 $\times 10^{-7}$ ***

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.

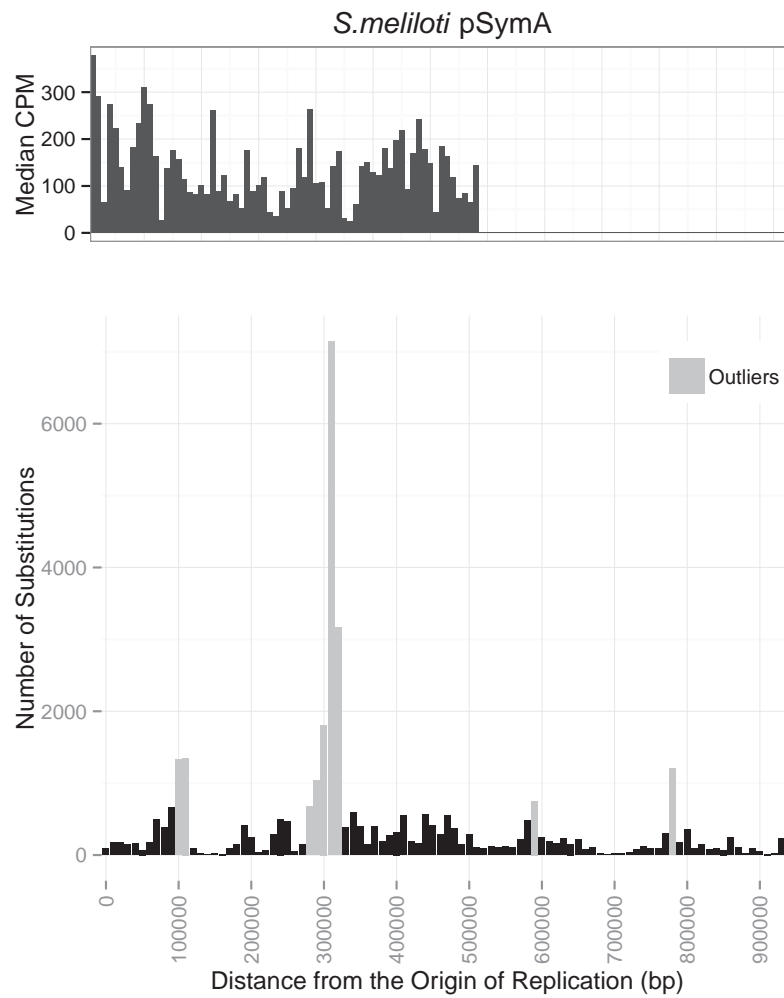
Bacteria and Replicon	Coefficient Estimate	Standard Error	P-value
<i>E. coli</i> Chromosome	-6.41 $\times 10^{-5}$	1.65 $\times 10^{-5}$	1.1 $\times 10^{-4}$
<i>B. subtilis</i> Chromosome	-9.9 $\times 10^{-5}$	2.18 $\times 10^{-5}$	6 $\times 10^{-6}$
<i>Streptomyces</i> Chromosome	-1.5 $\times 10^{-6}$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

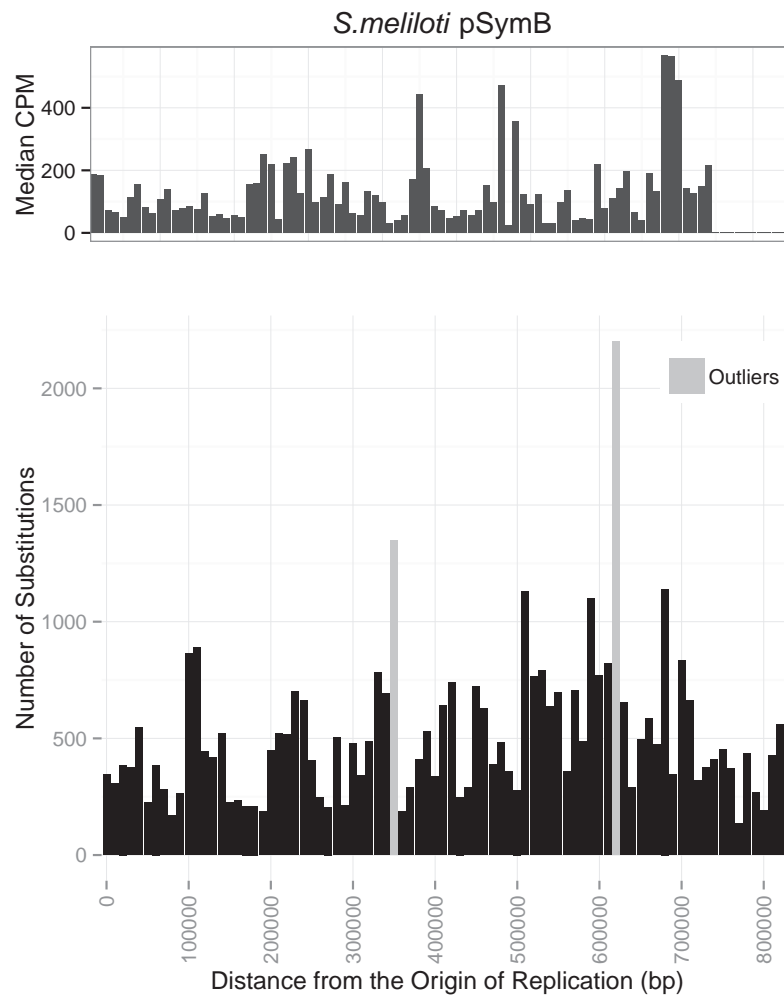
Table 2: Linear regression analysis of the median counts per million expression data along the genome of the respective bacteria replicons. Grey coloured boxes indicate statistically significant results at the 0.5 significance level. Linear regression was calculated after the origin of replication was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

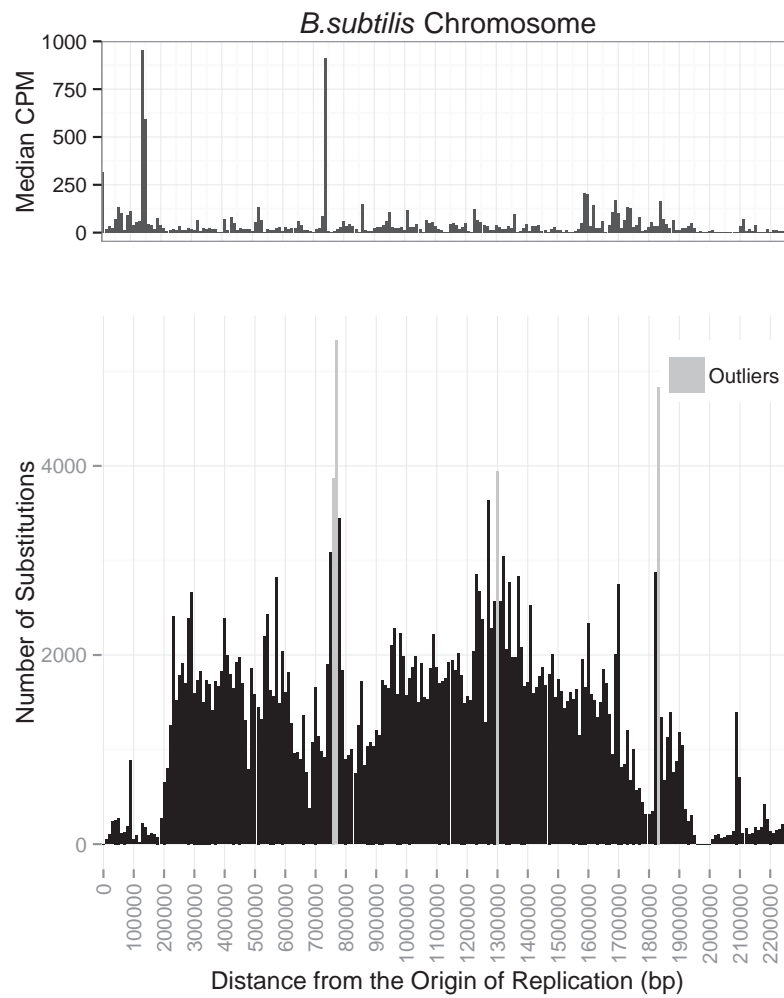


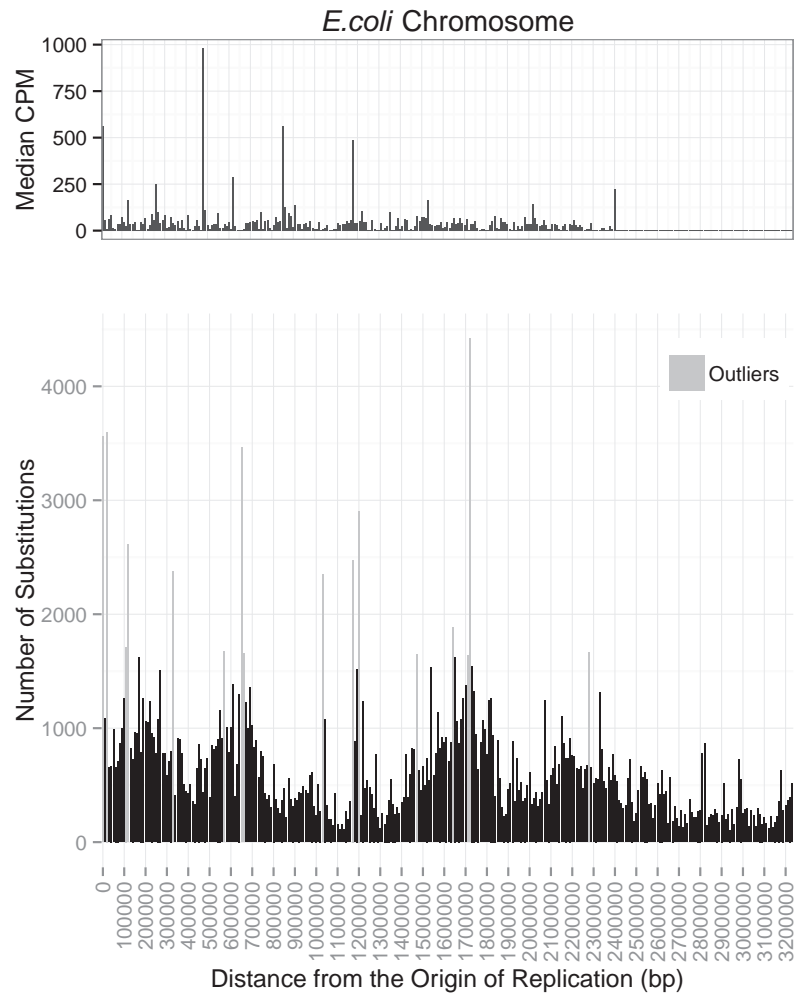
Bacteria and Replicon	Coefficient Estimate	Standard Error	P-value
<i>E. coli</i> Chromosome	-1.394×10^{-7}	2.425×10^{-9}	$< 2 \times 10^{-16}$
<i>B. subtilis</i> Chromosome	-1.265×10^{-8}	1.562×10^{-9}	5.430×10^{-16}
<i>Streptomyces</i> Chromosome	1.736×10^{-8}	7.231×10^{-10}	$< 2 \times 10^{-16}$
<i>S.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

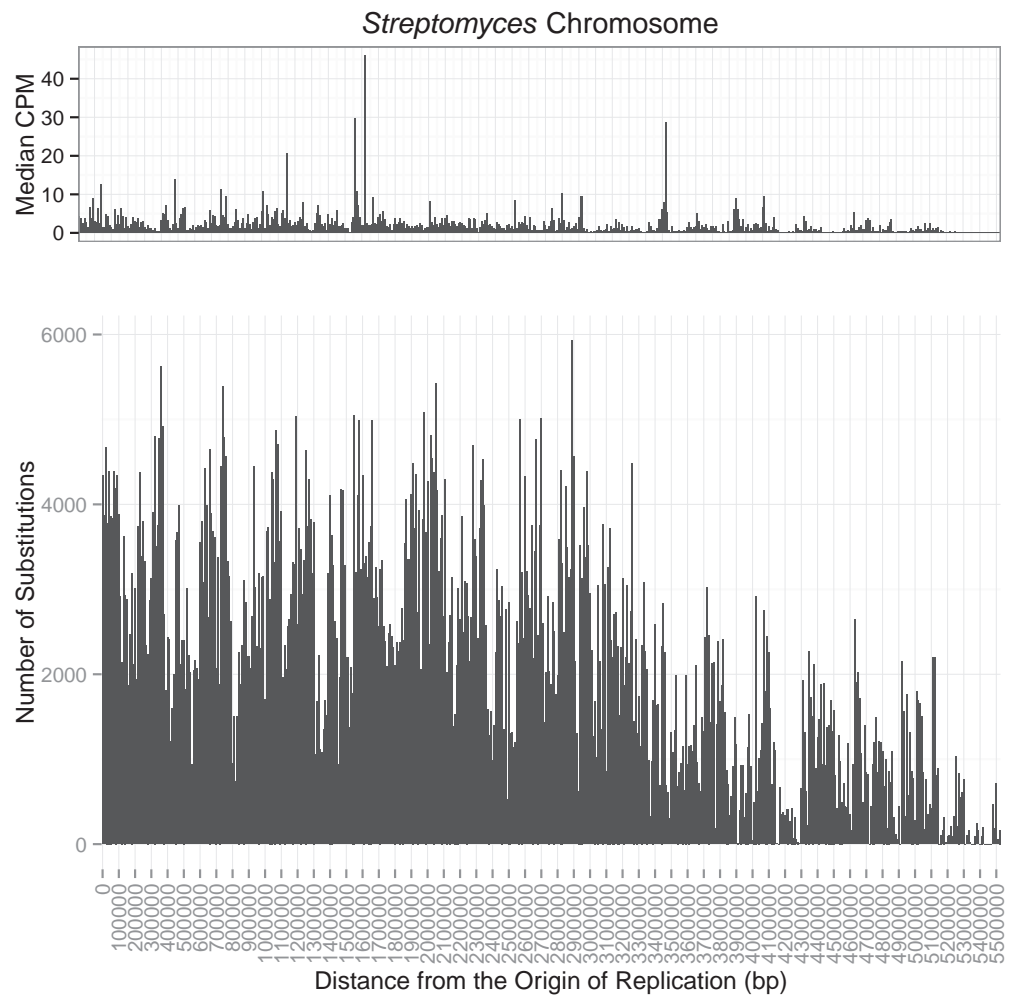
Table 3: Logistic regression analysis of the number of substitutions along the genome of the respective bacteria replicons. Grey coloured boxes indicate a negative logistic regression coefficient estimate. All results are statistically significant. Logistic regression was calculated after the origin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











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

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.