✓ Dec 23: Obtain gene expression data for each bacteria

X Jan 6: Write up methods for COG and sub paper

✓ Jan 6: Read papers on gene expression

✓ Jan 6: Apply for McMaster Bursaries and Grants

✓ Jan 6: Conference Grants Completed

✓ Feb 16: Have pipeline in R for normalizing raw counts

✓ Mar 2: Have code for plotting gene expression and substitution graphs

✓ Mar 17: Have all *S. meliloti* chrom, *E. coli* and *B. subtilis* data sets combined and into one graph

✓ Mar 31: Have all *Streptomyces* data sets combined and into one graph (will take more time because it is not all the same strain)

✓ April 5: ISMB Chicago Conference Abstract Due

✓ April 7: Have something figured out for the pSymB and pSymA gene expression datasets

✓ April 27: Create regression lines for gene expression

April 27: Write up gene expression stuff

April 27: Manuscript for Substitution and Gene expression paper finished

~~April 27: Make 2nd, 3rd, 4th, order regression lines for substitution data~~

May 31: Have data for other molecular trends (GC content, number of genes, essential gene lists..etc.) combined with graphs (or in supplement) for sub analysis

May 31: Complete COG analysis

Jun 30: Gene Expression analysis write up

Jun 30: COG analysis Paper draft completed

Jul 31: Updated Sub Paper methods and results

Jul 31: Add other mol trends to Sub Paper

Last Week

I realized that for pSymB I miscalculated the bidirectionality transformation, so I had to fix this and re-run everything. It did not change the logistic regression results (seen below). However, when I re-did the origin shuffling to see if the placement of the origin changed anything, moving the origin 100kb, 90kb and 80kb to the left made the logistic regression negative. I have been trying to figure out why this is happening but I am having no luck. I thought maybe it was because these shufflings are now 700kb away from the terminus, but the actual origin is about the same distance. I am still trying to figure this out but I am not sure what to do or what it means about the robustness of the origin shuffling. I also created simple linear regressions for the gene expression data and the results are summarized in the table below.

I looked at the gene expression data for *S. meliloti* in detail and it appears to look ok. When I graph the raw data and plot the regression line it looks like there is no trend for *S. meliloti*, the points are evenly distributed throughout the genome and there appears to be no increasing or decreasing trend. When looking at the other bacteria's raw data, there is clearly a decreasing trend when moving away from the origin. Additionally the number of genes and number of replicates are all comparable between *S. meliloti* and the other bacteria. So I think that the reason *S. meliloti* does not have significant gene expression regressions is because there is simply no trend.

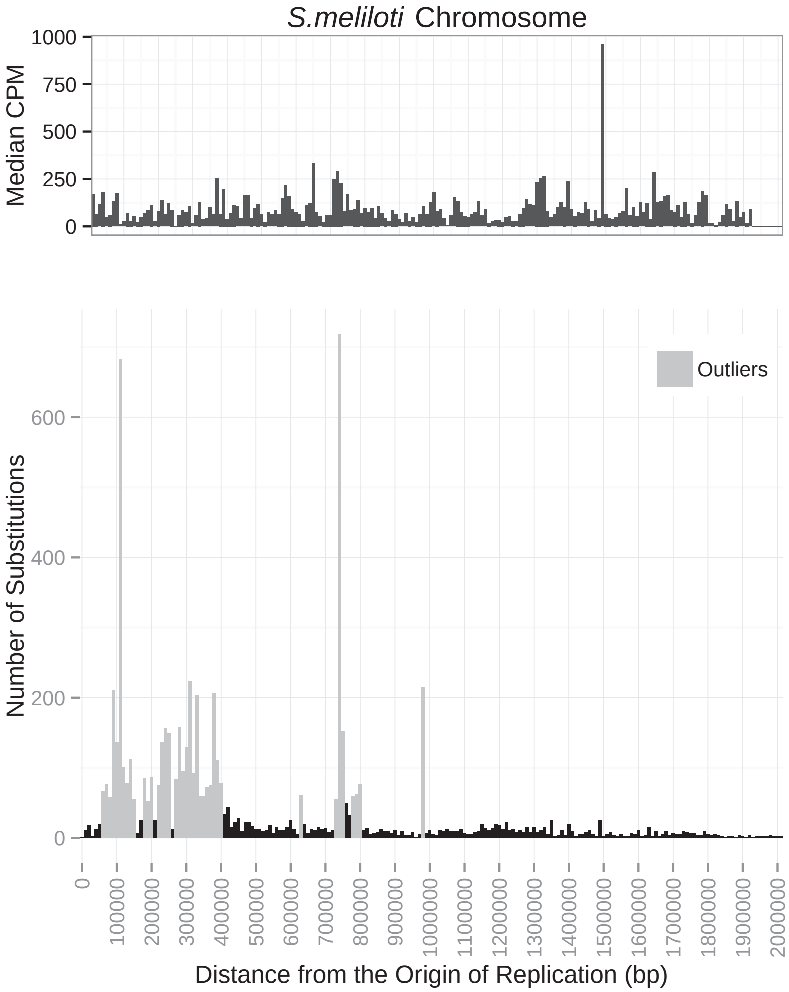
I spent a few days last week writing up the manuscript for the gene expression and substitution stuff.

This Week

I need to figure out the weird thing happening with the pSymB robust origin shuffling test. I also plan on having my manuscript finished before I leave for camping on Friday.

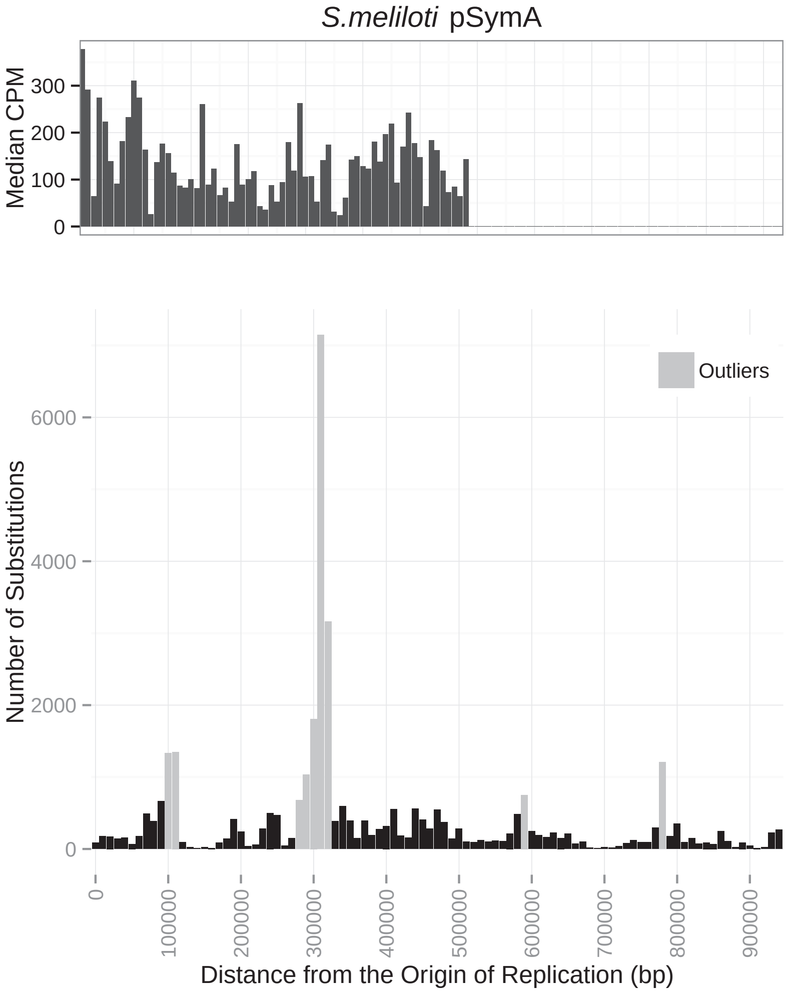
Bacteria and Replicon	Coefficient Estimate	Standard Error	P-value
E. coli Chromosome	-6.41×10^{-5}	1.65×10^{-5}	1.1×10^{-4}
B. subtilis Chromosome	-9.9×10^{-5}	2.18×10^{-5}	6×10^{-6}
Streptomyces Chromosome	-1.5×10^{-6}	1.4×10^{-7}	$<2\times10^{-16}$
S. meliloti Chromosome	3.19×10^{-5}	3.57×10^{-5}	3.7×10^{-1}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

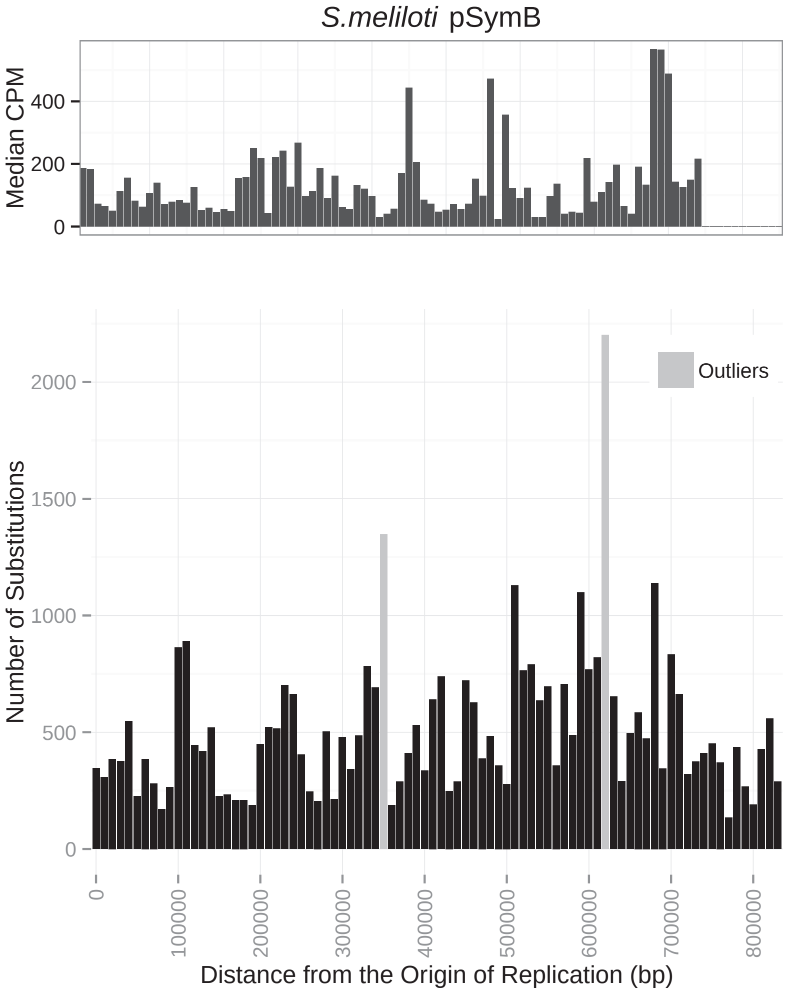
Table 1: Linear regression analysis of the median counts per million expression data along the genome of the respective bacteria replicons. Grey coloured boxes indicate statistically significant results at the 0.5 significance level. Linear regression was calculated after the origin of replication was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

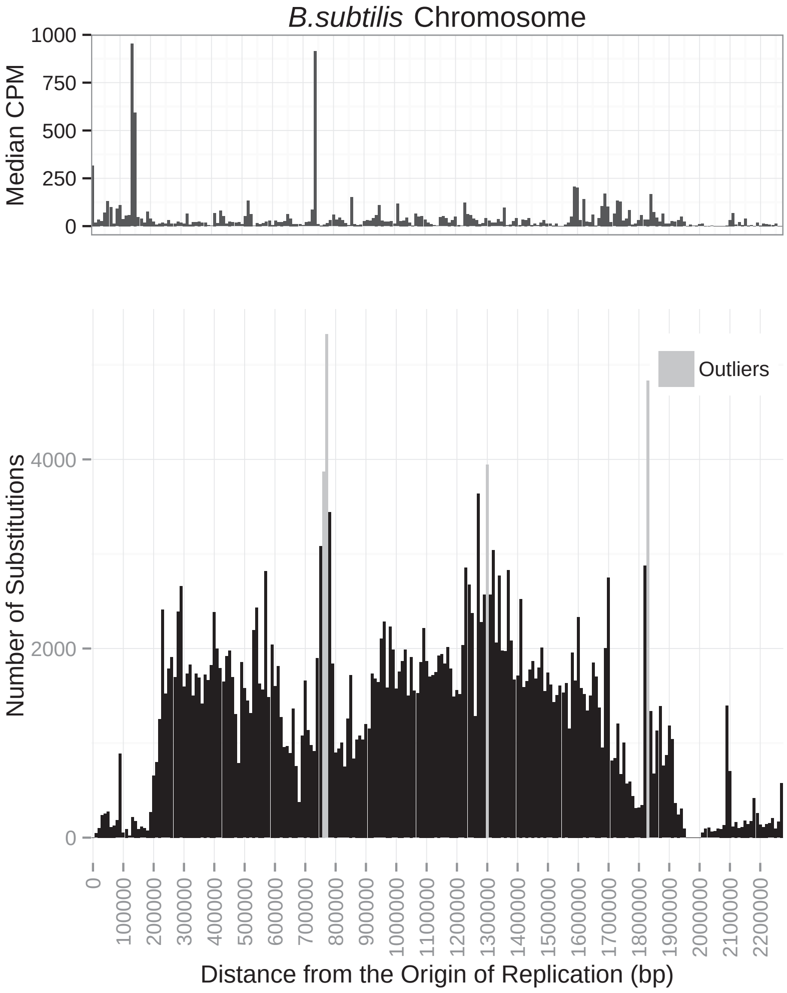


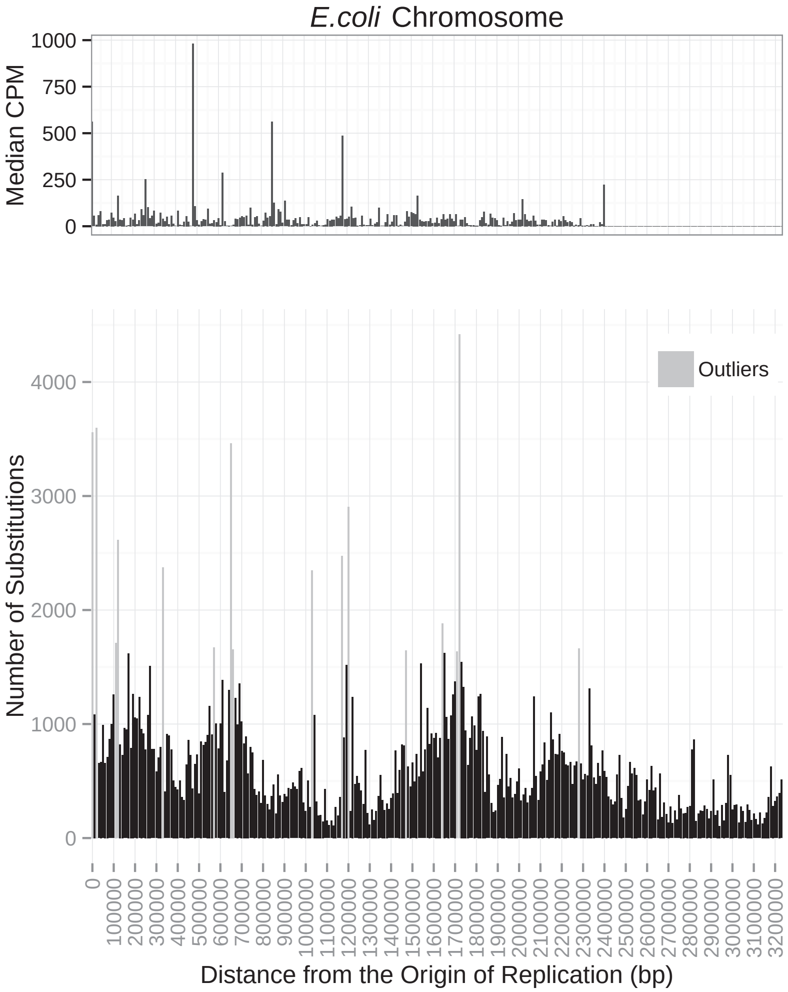
Bacteria and Replicon	Coefficient Estimate	Standard Error	P-value
E. coli Chromosome	-1.394×10^{-7}	2.425×10^{-9}	$<2\times10^{-16}$
B. subtilis Chromosome	-2.538×10^{-8}	1.58×10^{-9}	$<2\times10^{-16}$
Streptomyces Chromosome	1.736×10^{-8}	7.231×10^{-10}	$<2\times10^{-16}$
S. meliloti Chromosome	-1.541×10^{-6}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

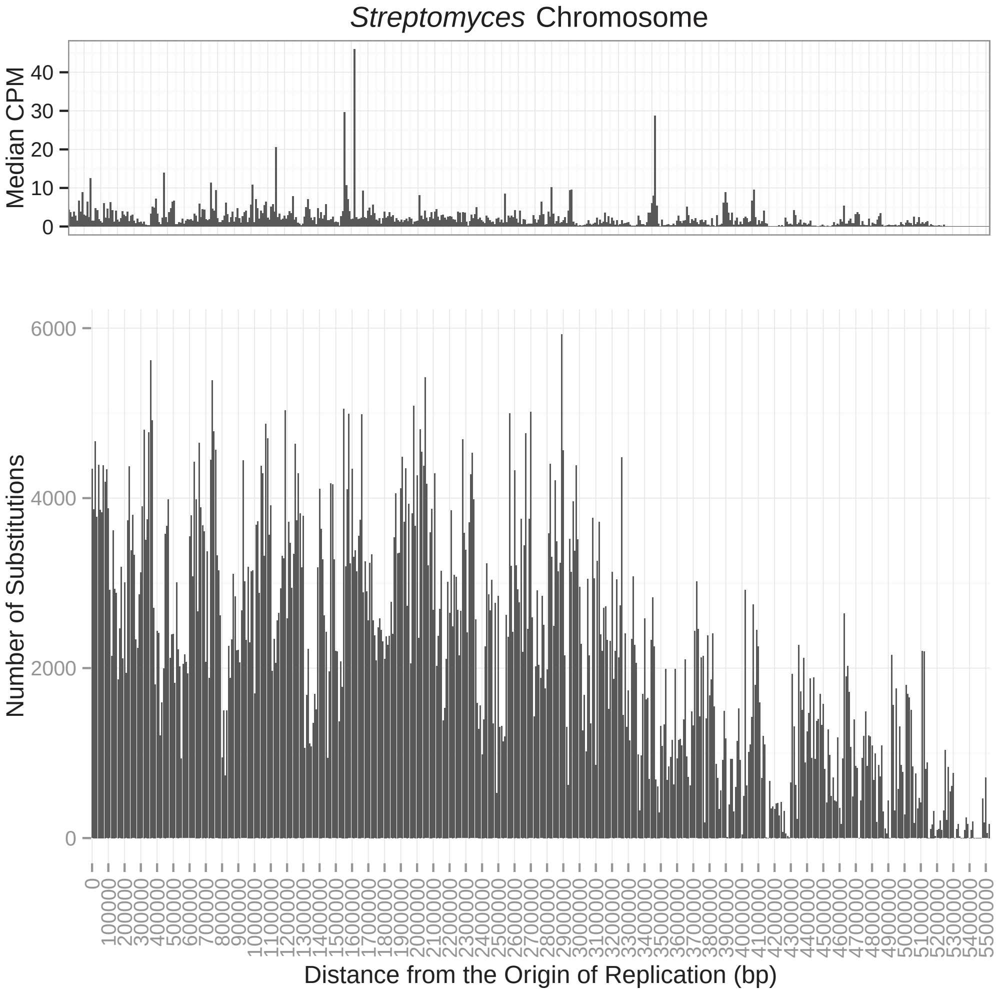
Table 2: Logistic regression analysis of the number of substitutions along the genome of the respective bacteria replicons. Grey coloured boxes indicate a negative logistic regression coefficient estimate. All results are statistically significant. Logistic regression was calculated after the origin 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 pSymA
Median CPM
0
100
200
300
Number of Substitutions
0
2000
4000
6000
Outliers
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
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)