✓ 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 finished the gene expression and substitution graphs for all the bacteria. The graphs are below. 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 (also seen below). I also created simple linear regressions for the gene expression data and the results are summarized in the table below. The significant linear regressions were all negative as predicted by other authors. So gene expression decreases as you move further from the origin of replication. All of *S. meliloti* did not have a significant linear regression which is interesting. It might be because there was only one gene expression dataset I could use, therefore less samples than the other bacteria. I started to write up these gene expression results and methods into my manuscript.

I still need to talk to you about the gene expression inversion stuff.

This Week

I would like to write code for the 2nd, 3rd, and 4th order regression lines for the substitution data and possibly the gene expression data. I would also like to start writing the code for the gene expression inversion simulation.

Bacteria and Replicon	Coefficient Estimate	Standard Error	P-value
E. coli Chromosome	-6.41×10^{-5}	1.44×10^{-5}	8.1×10^{-6}
B. subtilis Chromosome	-9.51×10^{-5}	2.05×10^{-5}	3.7×10^{-6}
Streptomyces Chromosome	-1.46×10^{-6}	1.3×10^{-7}	<2×10^{-16}
S. meliloti Chromosome	-5.68×10^{-6}	2.98×10^{-5}	8.5×10^{-1}
S. meliloti pSymA	-2.74×10^{-4}	4.85×10^{-4}	5.73×10^{-1}
S. meliloti pSymB	4.5×10^{-4}	2.46×10^{-4}	6.8×10^{-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.

Next Week

I would like to finish up the above mentioned stuff and begin writing and editing my manuscript.

Bacteria and Replicon	Coefficient Estimate	Standard Error	P-value
E. coli Chromosome	-1.394×10^{-7}	2.425×10^{-9}	<2×10^{-16}
B. subtilis Chromosome	-2.538×10^{-8}	1.58×10^{-9}	<2×10^{-16}
Streptomyces Chromosome	1.736×10^{-8}	7.231×10^{-10}	<2×10^{-16}
S. meliloti Chromosome	-1.541×10^{-6}	3.042×10^{-8}	<2×10^{-16}
S. meliloti pSymA	-9.130×10^{-7}	1.975×10^{-8}	<2×10^{-16}
S. meliloti pSymB	2.488×10^{-7}	1.964×10^{-8}	<2×10^{-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 pSymB

E.coli Chromosome

Streptomyces Chromosome