Subs Paper Things to Do:

- ~~# of coding and non-coding sites~~
- ~~# of subs in each of ↑~~
- ~~Look into *Streptomyces* non-coding issue~~
- Look into *E. coli* coding issue
- Look into pSymB coding/non-coding trend weirdness
- get dN/dS for coding/non-coding stuff
- Or get 1st, 2nd, 3rd codon pos log regs
- write up coding/non-coding results
- write up methods for coding/non-coding
- write methods and results for clustering
- take out gene expression from this paper
- write better intro/methods for distribution of subs graphs
- mol clock for my analysis?
- write discussion for coding/non-coding
- GC content? COG? where do these fit?
- write coding/non-coding into conclusion

Gene Expression Paper Things to Do:

- ~~look for more GEO expression data for *S. meliloti*~~
- ~~look for more GEO expression data for *Streptomyces*~~
- ~~look for more GEO expression data for *B. subtilis*~~
- ~~format paper and put in stuff that is already written~~
- find papers about what has been done with gene expression
- read papers ↑
- put notes from ↑ papers into word doc
- do same ancestral/phylogenetic analysis that I did in the subs paper
- Get numbers for how many different strains and multiples of each strain I have for gene expression

- write abstract
- write intro
- add stuff from outline to Data section
- create graphs for expression distribution (no sub data)
- add # of genes to expression graphs (top)
- average gene expression
- write discussion
- write conclusion
- look for more GEO expression data for *E. coli*
- add into methods: filters for Hiseq, RT PCR and growth phases for data collection
- update supplementary figures/file

Inversions and Gene Expression Letter Things to Do:

- create latex template for paper
- find papers about inversions and expression
- read papers ↑
- put notes from papers ↑ into doc
- use large PARSNP alignment to identify inversions
- confirm inversions with dot plot
- get as much GEO data as possible
- write outline for letter
- write Abstract
- write intro
- write methods
- compile tables (supplementary)
- write results
- write discussion
- write conclusion

Last Week

✓ look for more GEO expression data for *B. subtilis*

✓ format paper and put in stuff that is already written

Last week I was checking in on the results from the coding/non-coding stuff and I noticed that *Streptomyces* coding has only about 7600bp of data (including both substitutions and non-substitutions), which seems very wrong. I think that this might be because for th other bacteria we are dealing with the same sub-strains, so choosing a single sub-strain to identify coding and non-coding regions for all sub-strains. Where as for *Streptomyces* we are dealing with strains, not sub-strains. So potentially the coding and non-coding regions of one strain may not line up nicely with the regions of another strain? Should I make it so that if a base in the alignment falls within ANY coding or non-coding region of ANY strain then it should count? If I should do this, then should I change my analysis for all the bacteria to also do this?

Last week I finished up running all the coding and non-coding analysis and the results are summarized in the tables below. You can see that for most replicons, majority of the substitutions are coming from the non-coding regions of the genomes. The replicons that do not follow this trend is pSymB, pSymA (which are not chromosomes so this may still be able to be worked into a story) and *Streptomyces*. I am not sure why *Streptomyces* is showing both coding and non-coding sequences to have higher number of substitutions when moving away from the origin. Maybe because these taxa are from different strains of *Streptomyces*, so I am only capturing sequences where all taxa are present, which would likely be important genes, which would be expected to have a positive substitution trend? Thoughts?

The number of substitutions are higher in the non-coding regions than the coding regions for every replicon except pSymB, which is also really weird and interesting. Thoughts?

Escherichia coli looked like the code was doing something weird and I think it may have to do with my origin and bidirectionality scaling. I am looking into this.

I also made a table to summarize the proportion of coding and non-coding sections in each of the genomes (below).

I have also looked at the GEO datasets and attempted to find more for *Streptomyces*, and *S. meliloti*. These are summarized also in a table below. I have only just begun looking for more data in *B. subtilis*.

This Week

I will finish going through the *B. subtilis* and *E. coli* GEO data sets to see if there is any more expression data I can grab.

I would like to fix the bidirectionality issue that seems to be happening only with the *E. coli* coding analysis.

Find papers for the various gene expression papers to see what has already been done in the field and have solid background knowladge.

I would like to create a template in latex for both gene expression papers and add in information that I already have written up.

Next Week

I would like to start figuring out how to get dN/dS for coding and non-coding stuff and/or codon position logistic regression information.

Write out my methods for the coding/non-coding stuff.

Read some of the gene expression papers I will find.

Determine next steps for inversions and gene expression paper.

Bacteria and Replicon	% of Coding Sequences	% of Non-Coding Sequences	# of Subs Coding	# of Subs Non-Coding
E. coli Chromosome	87.22%	12.78%	702	256423
B. subtilis Chromosome	87.58%	12.42%	15547	287781
Streptomyces Chromosome	88.02%	11.98%	1357	1200749
S. meliloti Chromosome	85.68%	14.32%	1530	5581
S. meliloti pSymA	83.34%	16.66%	3230	10343
S. meliloti pSymB	88.70%	11.30%	37419	10596

Table 1: Total proportion of coding and non-coding sites in each replicon and the percentage of those sites that have a substitution (multiple substitutions at one site are counted as two substitutions).

Bacteria and Replicon	Coding Sequences	Non-Coding Sequences
E. coli Chromosome	2.496×10^{-5}*	-1.397×10^{-7}***
B. subtilis Chromosome	1.812×10^{-6}***	-1.439×10^{-8}***
Streptomyces Chromosome	2.984×10^{-5}***	1.689×10^{-8}***
S. meliloti Chromosome	6.993×10^{-6}***	-1.311×10^{-6}***
S. meliloti pSymA	-9.713×10^{-7}***	-1.413×10^{-7}***
S. meliloti pSymB	-4.406×10^{-7}***	5.916×10^{-7}***

Table 2: Logistic regression analysis of the number of substitutions along all positions of the genome of the respective bacteria replicons. These genomic positions were split up into the coding and non-coding regions of the genome. 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. All results are marked with significance codes as followed: $< 0.001 =$ '***', $0.001 < 0.01 =$ '**', $0.01 < 0.05 =$ '*', $0.05 < 0.1 =$ '.', $> 0.1 =$ ' '.

Bacteria and Replicon	Coefficient Estimate	Standard Error	P-value
E. coli Chromosome	2.496×10^{-5}	8.695×10^{-6}	0.0041
B. subtilis Chromosome	1.812×10^{-6}	8.913×10^{-8}	$<2\times10^{-16}$
Streptomyces Chromosome	2.984×10^{-5}	1.858×10^{-6}	$<2\times10^{-16}$
S. meliloti Chromosome	6.993×10^{-6}	6.205×10^{-7}	$<2\times10^{-16}$
S. meliloti pSymA	-9.713×10^{-7}	3.212×10^{-8}	$<2\times10^{-16}$
S. meliloti pSymB	-4.406×10^{-7}	2.317×10^{-8}	$<2\times10^{-16}$

Table 3: Logistic regression analysis of the number of substitutions along all coding portions of 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.

Bacteria and Replicon	Coefficient Estimate	Standard Error	P-value
E. coli Chromosome	-1.397×10^{-7}	2.427×10^{-9}	$< 2\times10^{-16}$
B. subtilis Chromosome	-1.439×10^{-8}	1.569×10^{-9}	$<2\times10^{-16}$
Streptomyces Chromosome	1.689×10^{-8}	7.235×10^{-10}	$<2\times10^{-16}$
S. meliloti Chromosome	-1.311×10^{-6}	3.393×10^{-8}	$<2\times10^{-16}$
S. meliloti pSymA	-1.413×10^{-7}	3.762×10^{-8}	1.73×10^{-4}
S. meliloti pSymB	5.196×10^{-7}	4.769×10^{-8}	$<2\times10^{-16}$

Table 4: Logistic regression analysis of the number of substitutions along all non-coding portions of 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.

Bacteria Strain/Species	GEO Accession Number	Date Accessed
E. coli K12 MG1655	GSE60522	December 20, 2017
E. coli K12 MG1655	GSE73673	December 19, 2017
E. coli K12 MG1655	GSE85914	December 19, 2017
E. coli K12 MG1655	GSE40313	November 21, 2018
E. coli K12 DH10B	GSE98890	December 19, 2017
E. coli BW25113	GSE73673	December 19, 2017
E. coli BW25113	GSE85914	December 19, 2017
E. coli O157:H7	GSE46120	August 28, 2018
B. subtilis 168	GSE104816	December 14, 2017
B. subtilis 168	GSE67058	December 16, 2017
B. subtilis 168	GSE93894	December 15, 2017
B. subtilis 168	GSE80786	November 16, 2018
S. coelicolor A3	GSE57268	March 16, 2018
S. natalensis A3	GSE112559	November 15, 2018
S. meliloti 1021 Chromosome	GSE69880	December 12, 2017
S. meliloti 2011 pSymA	NC_020527 (Dr. Finan)	April 4, 2018
S. meliloti 1021 pSymA	GSE69880	November 15, 18
S. meliloti 2011 pSymB	NC_020560 (Dr. Finan)	April 4, 2018
S. meliloti 1021 pSymB	GSE69880	November 15, 18

Table 5: Summary of strains and species found for each gene expression analysis. Gene Expression Omnibus accession numbers and date accessed are provided.