Subs Paper Things to Do:

- # of coding and non-coding sites
- # of subs in each of ↑
- Look into *E. coli* coding issue
- 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:

- 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

- format paper and put in stuff that is already written
- 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

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

✓ Make check lists for each of the 3 papers ✓ Read paper on detecting selection in bacterial genomes

Last week was spent analyzing the coding and non-coding data. The results are summarized in the tables below. They are exciting! For the chromosomes of all the bacteria (so far) it looks like the coding sections have a positive trend and the non-coding sections have a negative trend! Which makes sense biologically! *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. It appears to be only pSymA and pSymB that do not follow these trends but they are not chromosomes so I think we can still make a convincing argument as to why they are not following the trends of the other replicons. I think that maybe why we were seeing only negative trends before was because there were more substitutions in the non-coding regions than coding and these non-coding subs were driving the logistic regression to be negative.

I have been sticking to my goal of reading one paper a week during my off time while TA-ing.

This Week

Finish up the *Streptomyces* non-coding analysis. Put numbers to the proportion of each genome that is coding/non-coding and how many substitutions are in each so I know if it truly was more substitutions in the non-coding regions that were driving the previous whole genome substitution trends. I

need to keep looking at coding *E. coli* and double checking it is doing what it is supposed to be doing.

I would like to make 3 check-lists for the 3 papers that we talked about today, so that I can start getting things done for them.

Next Week

I will have a more solid list of tasks for next week based on the lists I will be making this week for the papers. So I will be starting these this week/next week.

Bacteria and Replicon	Proportion of Coding Sequences	Proportion of Non-Coding Sequences
E. coli Chromosome		
B. subtilis Chromosome		
Streptomyces Chromosome		
S. meliloti Chromosome		
S. meliloti pSymA	83.34%	16.66%
S. meliloti pSymB		

Table 1: Total proportion of coding and non-coding sites in each replicon.

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}***	1.3×10^{-8}***
S. meliloti Chromosome	4.425×10^{-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	4.425×10^{-6}	5.155×10^{-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	1.3×10^{-8}	3.393×10^{-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.