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~~
- ~~Figure out why *Streptomyces* appears to have tons of coding data missing~~
- Figure out what is going on with cod/non-cod code and why it is still not working!
- 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~~
- ~~look for more GEO expression data for *E. coli*~~
- ~~Get numbers for how many different strains and multiples of each strain I have for gene expression~~

- ~~re-do gene expression analysis for *B. subtilis*~~
- ~~re-do gene expression analysis for *E. coli*~~
- ~~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
- 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
- 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:

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

- write methods
- compile tables (supplementary)
- write results
- write discussion
- write conclusion

Last Week

✓find papers about inversions and expression

✓find papers about what has been done with gene expression

As I was going through the coding and non-coding code I kept finding that in some bacteria the code was doing what it was properly supposed to be doing, and sometimes it was not. I realized that this was because there are some genes that overlap or there are situations where one gene completely resides within the region of another gene. This is what was messing up my code, and obviously when trying to fix this a bunch of other new errors arose. I am still working on fixing the problem and HOPEFULLY this will be done by today and I will have the coding/non-coding results by tomorrow!

I have found about 30 articles on both gene expression and/or inversions in bacteria. My plan is to read and make notes on all of these over the break and apply for a few more scholarships

I made a detailed plan of what I want to do over the break, which consists of reading all the articles I found on gene expression and inversions and gene expression. As well as writing out the methods and results of the coding/non-coding stuff in the substitutions paper.

Summary of GEO data I found in a table below. Some questions for you:

1. Do you think there is enough there to do the same ancestral reconstruction analysis for *E. coli*?
2. Follow up to 1. : I have 6 datasets for *E. coli* K-12 MG1655, would these all be combined to obtain one gene expression value? Or would they all be considered separate taxa on the tree? The issue with that is that they were all mapped to the reference K-12 MG1655 genome. Thoughts?
3. Do you think there is enough *E. coli* data to look at inversions and gene expression?

This Week

I NEED to have this coding and non-coding business all finished for good! It is taking way too long to figure this out. This week I would like to figure out how to calculate dN/dS for the

substitution data.

Next Week/Holiday Break

Next week I want to get started on performing the ancestral reconstruction with the *E. coli* genomes for the gene expression data.

I have found about 30 articles on both gene expression and/or inversions in bacteria. My plan is to read and make notes on all of these over the break and apply for a few more scholarships

Bacteria and Replicon	% of Coding Sequences	% of Non-Coding Sequences	% of Subs Coding	% of Subs Non-Coding
E. coli Chromosome	86.47%	13.53%	5.00%	8.96%
B. subtilis Chromosome	87.49%	12.51%	7.31%	6.42%
Streptomyces Chromosome	89.03%	10.97%	13.74%	14.91%
S. meliloti Chromosome	86.27%	13.73%	0.19%	0.22%
S. meliloti pSymA	83.34%	16.66%	2.84%	4.58%
S. meliloti pSymB	88.81%	11.19%	2.78%	3.44%

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	-5.938×10^{-8}***	-9.237×10^{-8}***
B. subtilis Chromosome	-9.791×10^{-8}***	NS
Streptomyces Chromosome		9.182×10^{-9}***
S. meliloti Chromosome	-1.498×10^{-6}***	-7.037×10^{-7}***
S. meliloti pSymA	-1.230×10^{-6}***	-1.464×10^{-7}***
S. meliloti pSymB	3.295×10^{-7}***	NS

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 = 'NS'.

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 MG1655	GSE114917	November 22, 2018
E. coli K12 MG1655	GSE54199	November 26, 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
E. coli ATCC 25922	GSE94978	November 23, 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 HW-2	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 3: Summary of strains and species found for each gene expression analysis. Gene Expression Omnibus accession numbers and date accessed are provided.