Subs Paper Things to Do:

- ~~# of coding and non-coding sites~~
- ~~# of subs in each of †~~
- ~~Look into *Streptomyces* non-coding issue~~
- ~~Look into *Streptomyces* 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~~
- ~~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~~
- 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
- 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 *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*

✓get as much GEO data as possible for inversions and gene exp paper

✓re-do gene expression analysis for *E. coli*

✓Look into *E. coli* coding issue for the sub paper

I finished going through all the datasets on GEO and found at least one more I could include for each of the bacteria. So I think that it would be wise to re-do the gene expression analysis with these new data sets to have the most amount of data possible for each replicon.

The summary of all the strains that I have found are in a table below. I think that *E. coli* is the only bacteria that has enough different strain information (maybe?) to do an ancestral reconstruction analysis for gene expression and to investigate inversions and their impact on gene expression. The only question I have about this is that I have 7 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?

Re-did gene expression analysis with the extra datasets I found for *B. subtilis* and *E. coli*.

I looked into the *E. coli* coding issue and I fixed it! I accidentally put in the wrong genome length and it was messing things up and making the positions negative. But it is all fixed now and the regression lines info is in the tables below.

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? It could also be the fact that we are using blocks that have ALL taxa present. So this limits the data. In addition, we are removing any column in the alignment that has at least one gap in it and treating this column as missing data. This may also account for why there is so little *Streptomyces* coding data?

Started looking into reasons why pSymB has the opposite trend than what we expect. Have not found an explanation yet, will continue to look into this.

I am still looking into the issue with *E. coli* coding and it's bidirectionality scaling.

I am almost finished going through the gene expression data. There are just about 300 more *E. coli* experiments for me to parse through. The table below summarizes what I have found so far. Let me know which taxa you think there is enough diversity to do a phylogenetic analysis on.

Some of the bacteria also have time series data (samples taken at different times), or data sampled from different growth conditions. This is *E. coli*, *Bacillus subtilis*, *Streptomyces*, and limited data for *S. meliloti* chrom. However, these are sampled at variable times ranging from minutes to days, or different growth periods and conditions. The growth phases I think may be too subjective, and the time series data may not have enough datasets sampled at the same to be comparable.

This Week

I will finish going through the *E. coli* GEO data sets to see if there is any more expression data I can grab, and focus on this to break from trying to figure out the weirdness that is happening with the coding/non-coding stuff.

I would like to fix the bidirectionality issue that seems to be happening only with the *E. coli* coding analysis, and figure out what is happening with *Streptomyces* coding and pSymB.

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 the inversions and gene expression paper.

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	1.912×10^{-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	1.912×10^{-6}	8.753×10^{-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 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 5: Summary of strains and species found for each gene expression analysis. Gene Expression Omnibus accession numbers and date accessed are provided.

Bacteria and Replicon	Coefficient Estimate	Standard Error	P-value
E. coli Chromosome	-6.03×10^{-5}	1.28×10^{-5}	2.8×10^{-6}
B. subtilis Chromosome	-9.7×10^{-5}	2.0×10^{-5}	1.2×10^{-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 6: 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.