

Subs Paper Things to Do:

- ~~# of coding and non-coding sites~~
- ~~# of subs in each of  $\uparrow$~~
- ~~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~~
- ~~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

✓Finish re-doing the coding/non-coding analysis (after fixing the code)

✓re-calculated the number of subs in the coding and non-coding regions

Finished re-doing the coding/non-coding analysis (again) with the proper code. The results are summarized in the table below. I also re-calculated the number of subs in the coding and non-coding regions (see table below). But two things are weird: 1) pSymA and pSymB both have more substitutions in the coding regions than the non-coding regions, and 2) *E. coli* and *S. meliloti* Chrom have VERY few subs in the coding regions. For 2) I am not sure how I can be getting a significant log reg for *E. coli* when there are not that many data points?

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. This also fixed the pSymB coding and non-coding issue! Everything seems to make sense again!

However, when I was looking into why *Streptomyces* appeared to have only 7600bp of “coding” data, I realized that my code was not doing what I thought it was doing. It was missing identifying a majority of the coding sites because the data file was not sorted in ascending order and therefore the coding regions at the beginning of the genome were getting skipped over. I fixed this by simply ordering the datafile. I am just waiting for the rest of the replicons to finish running. The logistic regression results that I do have are below.

## This Week

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

I would like to create a template in latex for the inversions and gene expression paper.

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.

## Next Week

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 |
|--------------------------------|-----------------------|---------------------------|------------------|----------------------|
| <i>E. coli</i> Chromosome      | 87.22%                | 12.78%                    | 700              | 256425               |
| <i>B. subtilis</i> Chromosome  | 87.58%                | 12.42%                    | 15546            | 287782               |
| <i>Streptomyces</i> Chromosome | 88.02%                | 11.98%                    | 210592           | 991514               |
| <i>S. meliloti</i> Chromosome  | 85.68%                | 14.32%                    | 1529             | 5582                 |
| <i>S. meliloti</i> pSymA       | 83.34%                | 16.66%                    | 32040            | 10333                |
| <i>S. meliloti</i> pSymB       | 88.70%                | 11.30%                    | 37420            | 10595                |

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       |
|--------------------------------|----------------------------|----------------------------|
| <i>E. coli</i> Chromosome      | $2.509 \times 10^{-5**}$   | $-1.397 \times 10^{-7***}$ |
| <i>B. subtilis</i> Chromosome  | $1.909 \times 10^{-6***}$  | $-1.439 \times 10^{-8***}$ |
| <i>Streptomyces</i> Chromosome | $5.483 \times 10^{-7***}$  | NS                         |
| <i>S. meliloti</i> Chromosome  | $6.993 \times 10^{-6***}$  | $-1.546 \times 10^{-6***}$ |
| <i>S. meliloti</i> pSymA       | $-9.704 \times 10^{-7***}$ | $-1.464 \times 10^{-7***}$ |
| <i>S. meliloti</i> pSymB       | $6.182 \times 10^{-7***}$  | $5.916 \times 10^{-7***}$  |

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 Strain/Species            | GEO Accession Number  | Date Accessed     |
|------------------------------------|-----------------------|-------------------|
| <i>E. coli</i> K12 MG1655          | GSE60522              | December 20, 2017 |
| <i>E. coli</i> K12 MG1655          | GSE73673              | December 19, 2017 |
| <i>E. coli</i> K12 MG1655          | GSE85914              | December 19, 2017 |
| <i>E. coli</i> K12 MG1655          | GSE40313              | November 21, 2018 |
| <i>E. coli</i> K12 MG1655          | GSE114917             | November 22, 2018 |
| <i>E. coli</i> K12 MG1655          | GSE54199              | November 26, 2018 |
| <i>E. coli</i> K12 DH10B           | GSE98890              | December 19, 2017 |
| <i>E. coli</i> BW25113             | GSE73673              | December 19, 2017 |
| <i>E. coli</i> BW25113             | GSE85914              | December 19, 2017 |
| <i>E. coli</i> O157:H7             | GSE46120              | August 28, 2018   |
| <i>E. coli</i> ATCC 25922          | GSE94978              | November 23, 2018 |
| <i>B. subtilis</i> 168             | GSE104816             | December 14, 2017 |
| <i>B. subtilis</i> 168             | GSE67058              | December 16, 2017 |
| <i>B. subtilis</i> 168             | GSE93894              | December 15, 2017 |
| <i>B. subtilis</i> 168             | GSE80786              | November 16, 2018 |
| <i>S. coelicolor</i> A3            | GSE57268              | March 16, 2018    |
| <i>S. natalensis</i> HW-2          | GSE112559             | November 15, 2018 |
| <i>S. meliloti</i> 1021 Chromosome | GSE69880              | December 12, 2017 |
| <i>S. meliloti</i> 2011 pSymA      | NC_020527 (Dr. Finan) | April 4, 2018     |
| <i>S. meliloti</i> 1021 pSymA      | GSE69880              | November 15, 18   |
| <i>S. meliloti</i> 2011 pSymB      | NC_020560 (Dr. Finan) | April 4, 2018     |
| <i>S. meliloti</i> 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.

| Bacteria and Replicon          | Coefficient Estimate   | Standard Error        | P-value               |
|--------------------------------|------------------------|-----------------------|-----------------------|
| <i>E. coli</i> Chromosome      | $-6.03 \times 10^{-5}$ | $1.28 \times 10^{-5}$ | $2.8 \times 10^{-6}$  |
| <i>B. subtilis</i> Chromosome  | $-9.7 \times 10^{-5}$  | $2.0 \times 10^{-5}$  | $1.2 \times 10^{-6}$  |
| <i>Streptomyces</i> Chromosome | $-1.5 \times 10^{-6}$  | $1.4 \times 10^{-7}$  | $< 2 \times 10^{-16}$ |
| <i>S. meliloti</i> Chromosome  | $3.19 \times 10^{-5}$  | $3.57 \times 10^{-5}$ | $3.7 \times 10^{-1}$  |
| <i>S. meliloti</i> pSymA       | $-5.36 \times 10^{-5}$ | $6.34 \times 10^{-4}$ | $9.33 \times 10^{-1}$ |
| <i>S. meliloti</i> pSymB       | $5.05 \times 10^{-4}$  | $2.6 \times 10^{-4}$  | $5.3 \times 10^{-2}$  |

Table 4: 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.