Subs Paper Things to Do:

- causes for weird selection and subs results in *Streptomyces*
  - see how often class 4 arises in strep to see what is going on in later portion of the genome (to see if annotation is really a problem)
  - split up the strep data into core and non core and see if results are the same
- ~~make graphs proportional to length of respective cod/non-cod regions~~
- ~~test examples for genes near and far from terminus (robust log reg/results)~~
- ~~linear regression on 10kb regions for weighted and non-weighted substitutions~~
- ~~average number of substitutions in 20kb regions near and far from the origin~~
- ~~figure out why the data is weird for number of cod/non-cod sites~~
- why are the lin reg of dN, dS and ω NS but the subs graphs are...explain!
- grey out outliers in subs graphs?
- mol clock for my analysis?
- GC content? COG? where do these fit?

Gene Expression Paper Things to Do:

- ~~linear regression on 10kb regions~~
- ~~put new 10kb lin reg and # of genes over 10kb lin reg into paper~~
- ~~write about ↑ in methods and discussion~~
- ~~put expression lin reg and # coding sites log reg into supplement~~
- ~~write about ↑ in paper and how results are the same~~
- ~~update supplementary figures/file~~
- ~~correlation of gene expression across strains~~
  - ~~make graphs pretty and more informative with label names~~
  - ~~add them to supplement with a mini write up of what we did and why~~
  - ~~mention this in the actual paper~~
- if necessary add a phylogenetic component to the analysis
- ~~potentially remove genes that have been recently translocated from the analysis~~
- ~~model gene exp + position + number of genes~~

- ~~split up the strep data into core and non core and see if results are the same~~
- ~~what is going on with *Streptomyces* number of genes changing drastically from core to non-core~~
- codon bias?
- ~~what is going on with really high gene expression bars~~
- edit paper
- submit paper

Inversions and Gene Expression Letter Things to Do:

- ~~check if opposite strand in progressiveMauve means an inversions (check visual matches the xmfa)~~
- ~~check if PARSNP and progressiveMauve both identify the same inversions (check xmfa file)~~
- create latex template for paper
- ~~put notes from papers into doc~~
- ~~use large PARSNP alignment to identify inversions~~
- confirm inversions with dot plot
- make dot plot of just gene presence and absence matrix (instead of each site) to see if this will go better
- look up inversions and small RNA's paper Marie was talking about at Committee meeting
- write outline for letter
- write Abstract
- write intro
- write methods
- compile tables (supplementary)
- write results
- write discussion
- write conclusion
- do same ancestral/phylogenetic analysis that I did in the subs paper

General Things to Do:

- summarize references 40 and 56 from Committee meeting report (Brian was asking)
- read and make notes on papers I found for dissertation intro

Last Week

✓length cutoff for non-coding sections

✓re-run subs analysis

✓new graphs for gene expression data

✓talked to Marie about *Streptomyces*

✓started selection analysis

Subs analysis I re-did the analysis for the substitutions part of my thesis. I had the results all laid out but my new laptop died so I didn't have time to put it all together here. But as soon as I have the results together I will let you know! From what I remember it looks roughly the same as before. We decided to use the same cutoff for the non-coding sections (100bp) as for the coding sections.

I also made new gene expression graphs for all the bacteria which has a more efficient code, but also has *Streptomyces* having negative numbers to make the graph clearer. Again, once I get my laptop sorted I will show you these figures. The only graph that was not working was the *S. meliloti* chrom.

Talk with Marie I spoke with Marie about the *Streptomyces* and she suggested as an alternative solution to look at just the core genome of *Streptomyces*. There are a lot of papers that state that the core is similar but the terminal ends are just really out of wack. So she thinks it would not be too hard to justify. She sent me some papers with other phylogenies of *Streptomyces* so I am going to have a look at the to see if I can find any appropriate groupings with more genomes.

Selection analysis I started to re-do this analysis with the incorporation of Gblocks, but have hit some snags. Because these sequences need to begin with the start of a codon, and end with the end of a codon, I am having trouble making this happen. Hopefully I can figure this out this week.

This Week

Substitutions project:

1. get new tree picture for ne *Streptomyces* genomes
2. figure out issue with alignment length for selection analysis
3. re-run selection analysis
4. figure out why *S. meliloti* chromosome gene expression graph is not working

Next Week

Gene Expression:

1. new graphs for substitutions? (negative positions), write new (and better) code for this
2. look into journal requirements for submission
3. write cover letter for gene expression paper
4. substitutions paper edits

Bacteria and Replicon	Protein Coding Sequences	Non-Protein Coding Sequences
E. coli Chromosome	-1.887×10^{-8}***	6.462×10^{-8}***
B. subtilis Chromosome	-7.200×10^{-8}***	-1.296×10^{-7}***
Streptomyces Chromosome	3.703×10^{-8}***	1.775×10^{-7}***
S. meliloti Chromosome	-2.024×10^{-7}***	-1.594×10^{-7}**
S. meliloti pSymA	-5.894×10^{-7}***	-6.904×10^{-7}***
S. meliloti pSymB	1.361×10^{-7}***	4.475×10^{-7}***

Table 1: 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'.

	Protein Coding				Non-Protein Coding			
	Correlation Coefficient 20kb Near		Number of Substitutions per 20kb Near		Correlation Coefficient 20kb Near		Number of Substitutions per 20kb Near	
Bacteria and Replicon	Origin	Terminus	Origin	Terminus	Origin	Terminus	Origin	Terminus
E. coli Chromosome	-1.343×10^{-5}**	NS	5.45×10^{-3}	6.60×10^{-3}	NS	-9.231×10^{-5}**	2.95×10^{-3}	4.67×10^{-3}
B. subtilis Chromosome	NS	3.114×10^{-5}***	1.24×10^{-3}	8.00×10^{-3}	NS	7.906×10^{-5}*	1.35×10^{-3}	1.74×10^{-2}
Streptomyces Chromosome	9.938×10^{-5}***	-8.242×10^{-5}***	1.22×10^{-3}	7.77×10^{-3}	NS	-1.023×10^{-4}***	1.66×10^{-3}	2.14×10^{-2}
S. meliloti Chromosome	NS	NS	9.45×10^{-5}	4.25×10^{-5}	NS	NS	1.61×10^{-4}	1.41×10^{-4}
S. meliloti pSymA	NS	NS	9.78×10^{-4}	6.89×10^{-4}	1.431×10^{-4}**	NS	2.51×10^{-3}	9.44×10^{-4}
S. meliloti pSymB	-2.375×10^{-5}***	-6.976×10^{-5}***	2.18×10^{-3}	1.61×10^{-3}	NS	-6.134×10^{-5}**	2.64×10^{-3}	5.23×10^{-3}

Table 2: Logistic regression on 20kb closest and farthest from the origin of replication after accounting for bidirectional replication and outliers. All results are marked with significance codes as followed: < 0.001 = '***', $0.001 < 0.01$ = '**', $0.01 < 0.05$ = '*', > 0.05 = 'NS'.

	Protein Coding		Non-Protein Coding	
Bacteria and Replicon	Weighted	Non-Weighted	Weighted	Non-Weighted
E. coli Chromosome	-2.91×10^{-10}*	-1.57×10^{-4}***	NS	-9.29×10^{-6}***
B. subtilis Chromosome	-1.150×10^{-9}**	-1.993×10^{-4}**	NS	-8.24×10^{-6}**
Streptomyces Chromosome	NS	-8.98×10^{-6}***	3.87×10^{-10}***	-1.065×10^{-6}***
S. meliloti Chromosome	-1.389×10^{-10}**	-1.425×10^{-5}**	NS	NS
S. meliloti pSymA	-2.01×10^{-9}*	-1.06×10^{-4}*	-3.95×10^{-9}**	NS
S. meliloti pSymB	NS	NS	NS	NS

Table 3: Linear regression on 10kb sections of the genome with increasing distance from the origin of replication after accounting for bidirectional replication. Weighted columns have the total number of substitutions in each 10kb section of the genome divided by the total number of protein coding and non-protein coding sites in the genome. Non-weighted columns are performing a linear regression on the total number of substitutions in each 10kb section of the genome. All results are marked with significance codes as followed: < 0.001 = '***', $0.001 < 0.01$ = '**', $0.01 < 0.05$ = '*', > 0.05 = 'NS'.

Bacteria and Replicon	Coefficient Estimate
E. coli Chromosome	-9.89×10^{-8}***
B. subtilis Chromosome	-2.239×10^{-8}***
Streptomyces Chromosome	-8.23×10^{-8}***
S. meliloti Chromosome	1.265×10^{-7}***
S. meliloti pSymA	3.084×10^{-7}888
S. meliloti pSymB	-2.172×10^{-7}***

Table 4: Linear regression analysis of the total number of protein coding genes per 10kb along the genome of the respective bacteria replicons. 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. All results are marked with significance codes as followed: < 0.001 = '***', $0.001 < 0.01$ = '**', $0.01 < 0.05$ = '*', > 0.05 = 'NS'.