Subs Paper Things to Do:

- why does sinoC have omega lin reg = 0 near and far from the origin?
- create new graphs for selection analysis
- find and example of high substitution bar in *Streptomyces* and put this into supplement as an example of really diverged taxa (and that subs are real!)
- discuss removing omega outliers in methods
- double check that the ter and ori and max genome pos are correct
- ~~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:

- if necessary add a phylogenetic component to the analysis
- codon bias?
- make corrections based on Brian's edits
- create a clean copy of the paper (no strikeout) for re-submission

Inversions and Gene Expression Letter Things to Do:

- create latex template for paper
- 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)

Last Week

✓finished writing first draft of thesis intro!

✓first draft for lab meeting presentation

✓first version of new selection graphs

✓implemented Brian's comments/concerns on reviewers comments

✓figure out why sinoC has so many omega values = 0

I have not written one of these in a while, but I am trying to get back into the swing of things. I finished a first draft of my intro for my dissertation! It still needs to be edited a lot and probably shortened but at least all the ideas are out there!

I have finished the first draft of my lab meeting presentation. I need to practice it a lot and will probably have to change/add some slides.

I showed you the first version of the selection graphs that I made last week, and based on our discussion will be altering them this week.

As we discussed, I made all the changes you asked me to for the Gene Expression paper. I will be correcting the new comments you gave me today and hopefully we can re-submit this week!

For the selection analysis in the Substitution paper, it seems that *S. meliloti* Chromosome genuienly has a lot of values where $dN = 0$ and therefore $\omega = 0$. These are making all the graphs

look weird (with respect to the outliers and the box plots) so I need to figure out a way to deal with this.

This Week

- make new corrections for Gene Expression paper
- get proper copy of gene expression paper ready (+ update stuff on `GitHub` supplementary folder)
- re-submit gene expression paper
- new selection graphs
- practice/perfect lab meeting presentation

Next Week

- why *S. meliloti* chromosome has an omega linear regression of 0 near and far from the origin
- find high substitution bar in *Streptomyces* as a supplementary example for the paper (to show that high subs that are left are real!)
- double check that I am using the correct ori, ter and max genome pos

Bacteria and Replicon	Protein Coding Sequences
E. coli Chromosome	-1.98×10^{-8}***
B. subtilis Chromosome	-5.55×10^{-8}***
Streptomyces Chromosome	7.49×10^{-8}***
S. meliloti Chromosome	-4.19×10^{-7}***
S. meliloti pSymA	-5.18×10^{-7}***
S. meliloti pSymB	1.67×10^{-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			
	Correlation Coefficient 20kb Near		Number of Substitutions per 20kb Near	
Bacteria and Replicon	Origin	Terminus	Origin	Terminus
E. coli Chromosome	NS	NS	5.62×10^{-3}	6.66×10^{-3}
B. subtilis Chromosome	NS	-8.37×10^{-5}***	1.95×10^{-3}	9.10×10^{-3}
Streptomyces Chromosome	7.91×10^{-5}***	-1.32×10^{-4}***	6.74×10^{-4}	6.73×10^{-3}
S. meliloti Chromosome	8.26×10^{-5}*	NS	9.79×10^{-5}	5.07×10^{-5}
S. meliloti pSymA	NS	NS	9.75×10^{-4}	3.23×10^{-3}
S. meliloti pSymB	-1.44×10^{-5}*	-6.32×10^{-5}***	1.96×10^{-3}	1.24×10^{-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	
Bacteria and Replicon	Weighted	Non-Weighted
E. coli Chromosome	-2.28×10^{-10}***	-1.65×10^{-4}***
B. subtilis Chromosome	-7.96×10^{-10}**	-1.73×10^{-4}**
Streptomyces Chromosome	2.38×10^{-11}*	NS
S. meliloti Chromosome	-1.05×10^{-10}***	-1.24×10^{-5}***
S. meliloti pSymA	NS	NS
S. meliloti pSymB	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	-2.32×10^{-2}***
B. subtilis Chromosome	-1.93×10^{-2}**
Streptomyces Chromosome	-1.24×10^{-3}***
S. meliloti Chromosome	-1.88×10^{-2}***
S. meliloti pSymA	-2.50×10^{-2}*
S. meliloti pSymB	NS

Table 4: Linear regression analysis of the total number of protein coding sites 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'.

	Gene/Genome Average		
Bacteria and Replicon	dS	dN	ω
E. coli Chromosome	0.2351	0.0101	0.0444
B. subtilis Chromosome	0.4201	0.0243	0.0714
Streptomyces Chromosome	0.0458	0.0011	0.0335
S. meliloti Chromosome	9.0094	0.0001	0.0015
S. meliloti pSymA	0.0872	0.0099	0.1642
S. meliloti pSymB	0.0940	0.0084	0.1142

Table 5: Weighted averages calculated for each bacterial replicon on a per genome basis using the gene length as the weight. Arithmetic mean calculated for the per gene averages for each bacterial replicon.