Subs Paper Things to Do:

- more genomes
- new outgroups? (too distant)
- explain high dS values in *B. subtilis*
- potentially poor alignment and non-orthologous genes (core genome, change methods?)
- non-parametirc analysis for subs
- gap in *Escherichia coli* fig 5
- new methods for trees
- concerned about repeated genes (TEs) and not analyzing core genome
- check if trimming respects coding frame
- clear distinction between mutations and substitutions in intro (separate sections)
- datasets from previous papers (repeat my analysis on them?)
- why would uncharacterized proteins have higher subs rates?
- R^2 values in regression analysis
- update gene exp paper ref
- why are the lin reg of dN, dS and ω NS but the subs graphs are...explain!
- mol clock for my analysis?
- GC content? COG? where do these fit?

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

Inversions + Gene Expression:

✓fix stats for Wilcox and t test (on actual df)

✓finish trimming of PARSNP aln (re-do mapping + block info file)

✓decided to be more strict with the inversions and apply the same alignment trimming method to the PARSNP blocks

✓check why genes are "missing" from mapping file (BW and K12)

✓look at duplicated rows for all genes/species mapping

✓think about if we can use/compare block with less than all taxa for inversions analysis

Subst Paper:

✓started re-doing the analysis with 23 *S. meliloti* genomes (there are not any more than this)

- figured out new RAxML pipeline for trees
- bootstrapped tree with branchlengths
- SH test on each block tree v.s. ↑ tree
- started trimming alignment to find only the good sections

✓updated gene expression paper ref with proper volume and page numbers

Inversions + Gene Expression:

We decided to apply the same strict alignment trimming method to the PARSNP alignments to ensure that we are comparing homologous genes. This will reduce the amount of data that we have, but I think it is better to be able to compare gene expression between things we are sure are similar. I am working on applying this to the data now, and then Queenie will use these new segments to produce a dataframe with gene expression, genomic position, and inversion information.

We decided to do the analysis multiple times for each instance of taxa present in each block. So for only blocks where all taxa are present, blocks where 3 or 4 taxa are present, and blocks where 3, 4, or 2 taxa are present. I will then compare the results to see if they differ. Queenie is working on producing those three dataframes for me.

I tried to run the R stats that I created using the test dataframe on the actual data and of course it did not work. So I need to figure out what happened.

Substitution Paper

The one block (435,873bp) from the 23 *S. meliloti* genome analysis finally finished running and it took 1 week to align. I looked at how large the blocks from the original analysis were for comparison, and there is a block from the *Streptomyces* analysis that was 2Mbp long. I can not remember how long that took to run but probably just over a week. So I am not sure if this can be used as justification to say that this analysis takes to long. I also ran one block (136,217bp) through my code to identify if the codon positions line up and it took 3h. Previously this would take 49min (with 4 genomes). Each block can be run in parallel so I suppose this will not take too much time. But I am not sure how this will scale with say 100 genomes. **What are your thoughts on these preliminary time constraints and scaling up the analysis?**

Finally figured out how to

This Week

- Queenie: working on getting summary statistics/graphics for inversions (how many, how big, in what taxa, where are they located, expression averages..etc)
- Queenie: looking into which block do not have all taxa present (probably due to genes being split between block and therefore removed)
- Queenie to complete visualizations on the summary statistics for inversions
- Queenie to create dataframe with raw expression data and inversions info
- Queenie to create dataframes with various taxa present
- keep working on scaling up the subst data (filtering/trimming mafft alignments and call ancestor)
- think about how to incorporate DESeq into analysis (requires raw counts)
- comment on repeated genes (TEs) in subst analysis and not using core genome

Next Week

- Queenie to create a plot of the inversions
- Queenie to compare blast results and gene mapping
- think about (and execute) how to incorporate distance from the origin into the inversion analysis
- keep working on scaling up the subst data (paml analysis and selection alignment trimming)
- work on DESeq pipeline