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:

✓created mapping file to show which genes match in all the strains in each block (based on subst paper "good" alignments)

✓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)

Subst Paper:

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

- automated downloading genomes
- ran progressiveMauve
- started running mafft
- started getting tree

✓completed supplemental analysis with multiple window sizes (10Kbp, 25Kbp, 50Kbp, 100Kbp, 200Kbp and 400Kbp)

✓started writing re-submission cover letter and added explanations for changes I have made so far

✓add in new paper notes to subst intro (more recent papers)

Inversions + Gene Expression:

I did not spend a lot of time on this project this week (I was preoccupied with the subst paper). Queenie is slowly working through the summary graphics/stats about the inversions. She also found some blocks that did not have all taxa present, likely because there was only one gene in that block and it was removed because it was only partially found within a block. She is looking into this further and making sure that they are removed for legitimate reasons. **Do you think we need blocks were all taxa are present? How important do you think that is in this analysis?**

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?**

I have spent most of the week working on the up scaling of the *S. meliloti* analysis. There are a total of 23 complete *S. meliloti* genomes. I am trying to include them all. The progressiveMauve ran smoothly, but there is one really large block (435,873bp) that is taking days to run (3days and counting). My plan is to just ignore this block for now and keep working through the analysis (so I can get the code and steps figured out) and then re-do everything once it is done. **should I be concerned with how long this block is taking to re-align using mafft?**

I had to re-write a lot of my code to make some steps more automated (which did not take me that long! I am improving!). I already discussed this with you, but phylip does not handle these large and numerous sequences well so I decided to switch to RAxML for building the trees. This seems to be going ok so far!

This Week

- check why genes are "missing" from mapping file (BW and K12)
- think about how to incorporate DESeq into analysis (requires raw counts)
- get Queenie to start creating visualizations on the summary statistics for inversions
- keep working on scaling up the subst data (complete trees (including SH test) and filtering/trimming mafft alignments)
- get Queenie to create dataframe with raw expression data and inversions info
- 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

Next Week

- keep working on scaling up subst data (complete until call ancestor step)
- comment on repeated genes (TEs) in subst analysis and not using core genome
- get Queenie to create a plot of the inversions
- think about (and execute) how to incorporate distance from the origin into the inversion analysis