Subs Paper Things to Do:

- 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:

✓done with mapping BW genes to K-12MG genes

✓Queenie to create a toy dataframe that I can begin working on the analysis of comparing inverted and non-inverted sections of the genome

✓Queenie: minimal genes found in overlapping regions (0.5%)

General:

✓Edited some of dissertation intro

Inversions + Gene Expression:

I finished mapping the BW genes to the K-12MG genes. So I now have a list of all the genes that are similar between those two strains. I have also repeated this for the other proteomes that were labelled as "redundant" so that I can ensure we are only looking at genes that actually exist in the strains we are looking at. I first aligned the two genomes for mapping with PARSNP. I think split the alignment into it's blocks, re-aligned with mafft and ran this through my substitutions code (which is fairly conservative) to find sections of the alignment which are well aligned (no gaps, start at the same codon position, min 100bp length) and then printed out which gene was in each of these sections from each genome. **does this sound like an ok plan? It is essentially using the same pipeline we used in the substitutions paper.**

Queenie quantified that there is only 0.05% of genes that are found in overlapping regions which I think is fine to just chuck and not use in the analysis. There was not a lot of information about these genes (i.e. not peseudo genes) so I assume that the overlap is just PARSNP being unsure about the definitive borders of the blocks.

Queenie has set up a small "toy" dataframe with (some of) the gene expression and inversion info so I can start to play with this and develop some sort of analysis pipeline so that when the final dataframe is ready, we can run it through the analysis quickly!

This Week

- think about how I am defining inversions and if I should classify inversions based on each strain vs if there is an inversion present in the block
- start playing with "toy" dataframe to determine analysis directions
- get Queenie started on creating lists of which genes match (from all strains) in the parsnp alignment

Next Week

- edit another section of dissertation intro
- get Queenie to create a plot of the inversions

- continue playing with "toy" dataframe to determine analysis directions