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:

✓Read one H-NS paper

✓diagram for inversions visualization

✓reverse complement ATCC for ↑ picture

✓preliminary code for correlation btwn H-NS binding and inversions/inverted blocks with significant gene expression differences

General:

✓Wrote 3pages for conclusion of Dissertation

Inversions + Gene Expression:

H-NS Preliminary Analysis: I made some preliminary code to see if there is a correlation between H-NS binding and inverted blocks as well as inverted blocks with significant differences in gene expression. I did this using a simple Pearson correlation. **Do you think I need to do something more complicated than this?** At the moment, there appears to be no correlation between inverted blocks and H-NS binding. However, I realized that there was a small formatting error in my data frame so this needs to be re-done and I would take these results with a grain of salt.

Inversion Visualization: I have been looking at programs to best depict the inversions we are seeing but most take too long to run (dot plots) or are quite complicated in the input files. I found a way to use parallel sets in `R` to get around this and I think they are pretty good (Figure 1)! In Figure 1, it appears as though the whole ATCC genome (CP009072) is inverted. So I decided to reverse complement the entire genome of ATCC and then re-create the diagram to see if it would make a difference (Figure 2). Figure 2 does clean things up a bit and still shows some of the inversions. Additionally, I am a bit confused because theoretically just reversing one sequence should still identify the same blocks in `PARSNP` when the sequence is not reversed. However, the blocks are all slightly different in their positions but the overall `PARSNP` visualization looks identical. So I am a bit confused about which image I should use in the paper. My main concern with this is if it is ok to use one image for the visualization (the reverse complemented image Figure 2), and the data from the other figure (non-reverse complemented Figure 1) for the data analysis. I mention this because when ATCC is reverse complemented it complicates the genomic position annotation. So I would have to ensure that all my annotation is also reverse complemented and that I am grabbing the correct sequence (reverse of the reverse complement?) when I re-align with `MAFFT`. **What are your thoughts on this?**

I also noticed from Figures 1 and 2 that a lot of the inversions have been rearranged in addition to being inverted. I am not sure how best to account for this in my calculations. **Do you have any thoughts on this?**

This Week

- Queenie: new dataframe for DESeq (combining blast results and raw expression data)
- Keep working on position and inversion visualization (finalize + make it pretty)
- determine what H-NS datasets we will be using (combine, intersect, separate..etc)

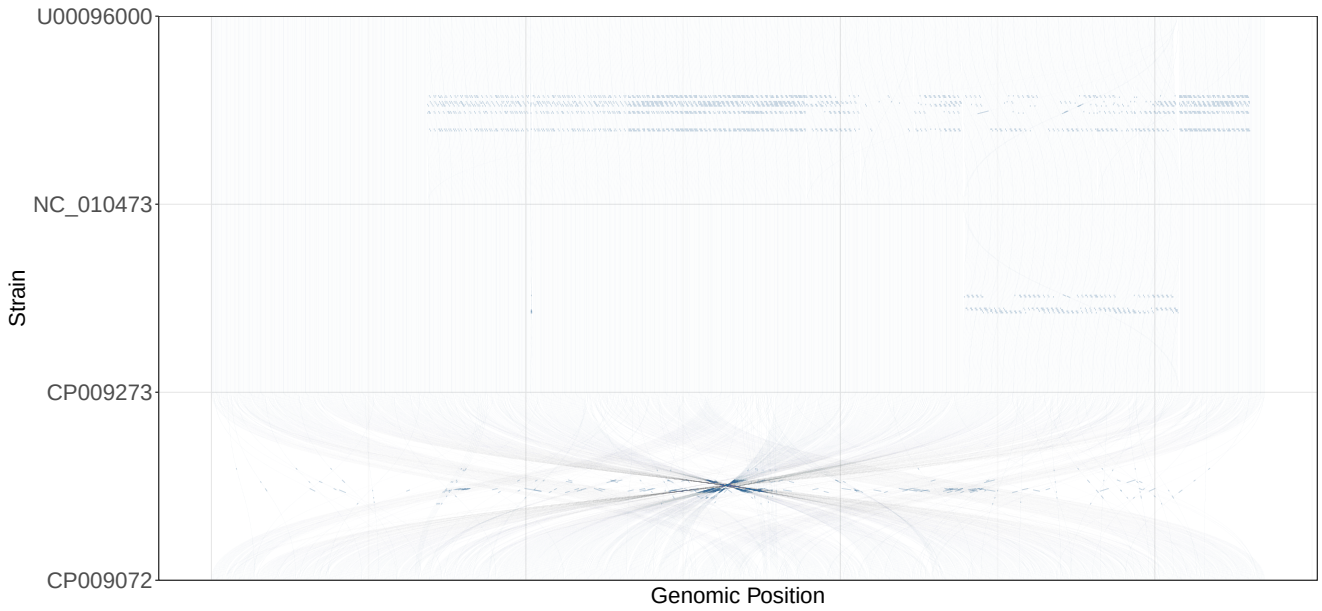

Figure 1: Visualization of rearrangements and inversions in all *E. coli* strains. ATCC is in the GenBank listed orientation.

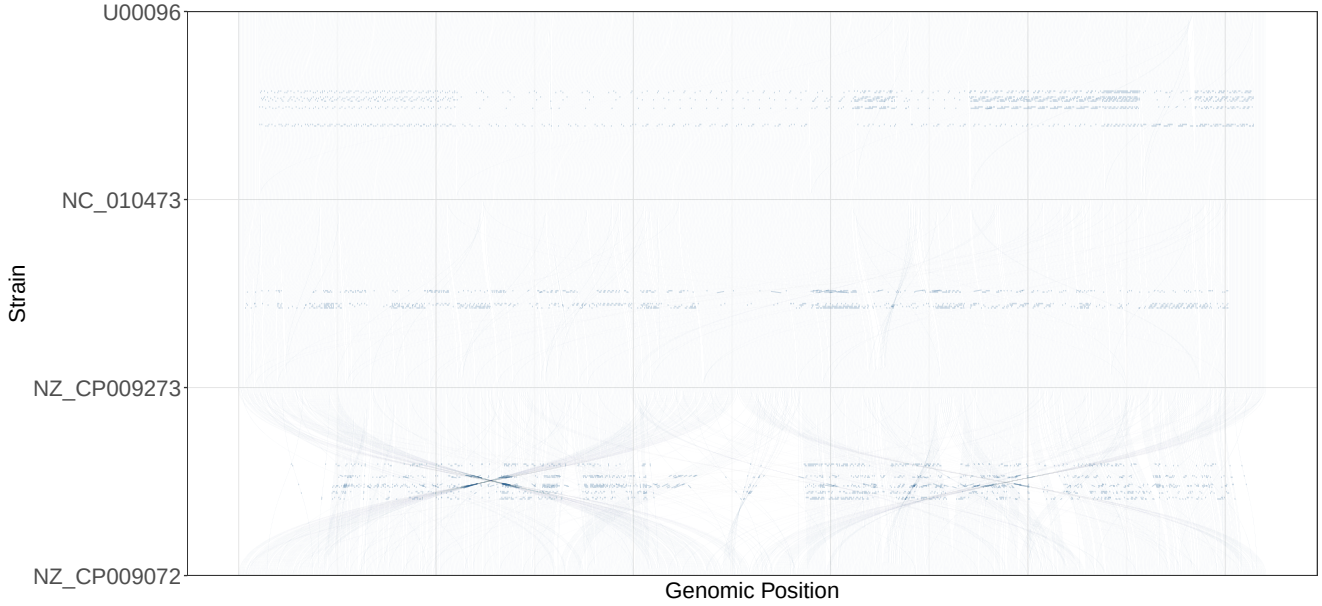

Figure 2: Visualization of rearrangements and inversions in all *E. coli* strains. ATCC is reverse complemented from the GenBank listed orientation.

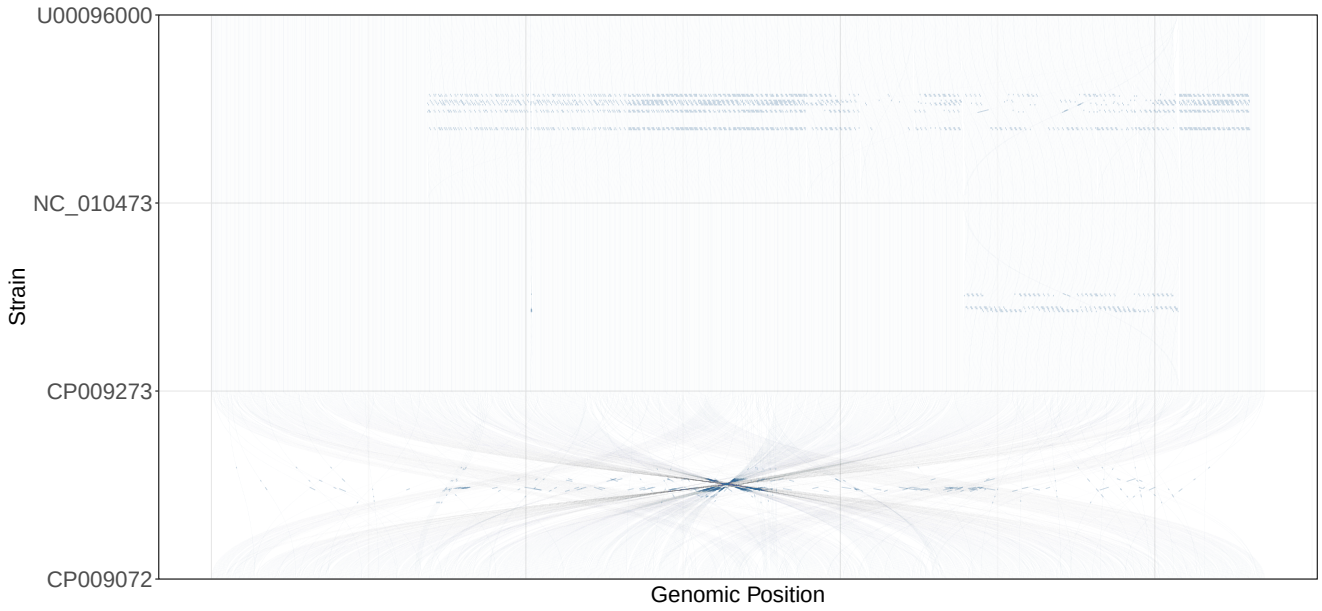

Figure 3: Visualization of the difference in gene expression between inverted and non-inverted sequences within alignment blocks. Each alignment block represents homologous sequences between the *Escherichia coli* strains insert table ref here. Each alignment block has one point on the graph to represent the average expression value in **C**ounts **P**er **M**illion (**CPM**) for all inverted (circles) and non-inverted (triangles) sequences within the block. Blocks that had a significant difference in gene expression (using a Wilcoxon sign-ranked test, see Materials and Methods) have the inverted and non-inverted gene expression averages highlighted in pink circles and purple triangles respectively. A smoothing line (`loewss`) was added to link the average gene expression values for the inverted (pink solid) and non-inverted (purple dashed) sequences within block that had a significant difference in gene expression (using a Wilcoxon sign-ranked test, see Materials and Methods). All blocks that did not have a significant difference in average gene expression between inverted and non-inverted sequences within alignment blocks have the average inversion (circles) and non-inversion (triangles) gene expression values coloured in light grey.

- write code to compare all H-NS datasets to inversions based on what is decided ↑

Next Week

- actual analysis on DESeq data
- visualizations/results for ↑
- read papers on H-NS proteins
- think about how to visualize H-NS and inversions info