Last Week

- Started to look at deriving gene families in order to create presence-absence matrices to ultimately solve for rates of lateral gene transfer
- Did blastp reciprocal hits for salmonella
- Retrieved output from blastp run and started running the R code.

This Week

I finally finished the "rarefaction" curves for the Roary runs i.e. seeing the steady increase of pangenome size with each addition of one genome of the species each to the run (running Roary with two genomes from the sample, then three, then four and so on the way up to 20 for 20 samples (point) at each respective number of genomes tested). As expected, a steady increase was shown, and all of them indicated similar patterns in order to further reassure for the consistency of Roary. I have attached the summary file separately with this report (once again I apologize for the formatting and I'm sure there's a better to do it).

Summary of the "Rarefaction" Data: As mentioned, it shows the expected steady positive linear correlation that we expect it to. I do want to figure out the rationality as to why most of the figures start off mostly linear, but then hit that sudden "dent" around fifteen genomes. For now, I assume it is some sort of equilibrium, though I kind of wish to test this theory should time permits for numbers greater than 20.

Out of all of these, the figure most worth noticing is probably *E. coli*, as we suspected, because its range of pangenome size by far exceeds all the other samples. It goes all the way up to 12 000 (range = 8000) and still shows signs of increasing. The next largest range belongs to *K. pneumoniae* is only 4000, and goes up to only 9000. This probably has something to do with *E. coli*'s supposed open genome, but of course we can't conclude that for certain from that alone.

I also spent this week messing around with regular expressions and rewriting Athena's code to correlate to the files that I obtained. I managed to generate my own Python code which retrieves accession numbers and protein ids from Genbank files using regular expressions. This should've generated the similar column output that Athena had for her "listofallgenes.txt" file. Although it was a bit longwinded and probably avoidable, it was a good chance for me to learn regex and how to use it in Python to generate the output that I would need in the future.

Upon concatenation of the weeded-out files, I had put the entire file through Athena's "genefamily11.R" which should indicate the gene families for my salmonella files. I made sure to run this on the entire concatenated file instead of the separated files, because it wouldn't make sense to find genefamilies amongst different files. Since this is a big file, I've left in running for awhile. In order to hopefuly speed up the process from now on, I ran blastp on the other bacteria on the background so that they'll immediately be ready when I need them next week.

Next Week

- Start figuring out how to reformat figures to make it look nice and insert into the current writing document
- Hopefully get some results from the R run and then find a more efficient method often thoroughly understand what kind of output files are required, because as it remains this is not the most productive way to do things and requires a lot of waiting time still (rewrite code? use parallel?)
- Continue writing, since figure are now available and have been created, so I will continue with my results (appendix?) section based on them.