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
- Working on the reciprocal blastp portion for salmonella after using the the getfeatures program on the salmonella genomes
- Code originally ran:

  ```
  blastp -db parsed10mycogenomeswotransposase_seqs -query
  10mycogenomeswotransposase_seqs.fna -out
  allvsall_wotransposase_softmask_sw -evalue 5e-2 -num_threads 10
  -outfmt '7 qseqid qlen qstart qend length sseqid slen qcovs score
  bitscore evalue' -soft_masking true -use_sw_tback &
  ```

  Note that no -num-threads argument was used.

This Week

Evidently the method with taking out the -num_threads argument was not at all efficient for such a large input file–I was running since Thursday and it still had not completed by Tuesday, so I took an alternatively approach by splitting up the giant 3000000+ line output file from getfeatures into 16 smaller files and ran without the num_threads argument in a for & loop to avoid the segmentation fault error that would always occur when I ran the whole large file with the num_threads argument.

After fragmentation, the files worked but then I bumped into the issue of sequence identifiers: the ones generated by the "modified_get_features.pl" code seems to not correlate with the the sequence ids that Athena had orginally used to extract certain features from the code. Since I'm not entirely sure what the output files are supposed to look like at this point, I will have to go in and either a) rewrite the modifed_get_features code in the language that I'm not too familiar with or b) create all the necessary files I need using my own codes.

I started with generating the "weeded-out" versions of the blast-p data: i.e. I picked out all the data that had less than an 85% match and hit to itself. I wrote this R, and although it is probably not the most efficient way to generate these, it makes the most sense to me as of now. I will leave this code running. I believe the code needs some adjustment so I will work on it next week.

In the meantime, I ran the "rarefaction" plots in the background and generated those. I will attach them in a separate file for reference next week (it should've been done this week but off-by-one error is absolutely no fun. I have two bacteria that are still running but other than that the data has been plotted for each species.

Next Week

- Concatenate the rarefaction figures into one organized document.
- Sort out the whole gene families and hopefully get results for at least salmonella (My mistake was trying to do it with the giant file so I'll probably start with the smaller fragents first, and if that ends up working I'll continue that and concatenate it all at the end once again.) Hopefully I should get to the end by Friday next week.