Last Week

- Ran blasts on all bacterial files, as well as the old genefamily11.R code when they were finished
- Ran a few with plasmids as preliminaries, but removed plasmids on the rest (using gbktrimmer.py)
- Played with and adjusted Sid's code for generating the Presence-Absence matrix
- Initiated 'sample output' for five salmonella species, just to test the results.

This Week

I finally got around to using github so my progress as well as most relevant files will be kept there. I will also add a git log to every weekly report from now on as a more detailed look at my commits if needed.

Github: https://github.com/le-ann/2017pange_hgt

Five Species Salmonella Test Output: Five salmonella species had been chosen randomly from the full sample set and were tested against the orginal instructions within Athena's genefamily2.txt file. I finally managed to derive the sample presence-absence matrix for the five species. The steps were listed as followed, with file names corresponding to relevant files. The bolded file names are attached with the e-mail (these files can also be found on my github):

1. Obtain gbff for the five individual files
2. Use "modified_get_features" to generate faa files for each gbff file
3. Form a blast database and run the reciprocal blast on the five files (output: allvsallwotransposase -> **output.txt**

4. Put the output.txt file through 'genefamily11.**R**'(output: gfcode1.Rdata, renamed **salmtest.Rdata**

5. Run the Rdata file on Sid's 'createPresenceAbsencemtrix.R' code to get the final matrix. Export the output (**presenceabsence.Rdata**, matrix: **presenceAbsence.csv**)

   The final output, as can be inferred, is a matrix of ones and zeros corresponding to the existence of in each gene in each ascession number, each of course corresponding to a particular species.

   If possible, I will like to work with this sample matrix next week and see exactly what kind of data it gives me.

   In the meantime, the data for *S. pyogenes* has somehow finished running on the old 'genefamily11.R' file, so I will likely be using those for comparison purposes. The matrix has already been generated through Sid's code, and I am currently running the same blastp file through Dr Golding's perl code. I still need to generate a matrix from the perl code, and am running the matrix code. When it finishes, I will extract both and compare the data.

   Some large bacterial blastp files are still running on the 'genefamily11.R'. Since they've already started, I will leave them. But within the next few weeks I will try to get the perl running on them as well. All the files I am focussing on at this point are plasmid removed, but I'm keeping track of both of them for comparison purposes later since they're there.

Next Week

- Run the remaining bacterial blast files on perl
- Compare output data for the five species *Salmonella* data
- Compare output data for the 51 species *S. pyogenes*
- Look at indelmiss(?)

Git Log:

```
commit a8faa23254873915c9281beb1b4a7b78e0f9ac26
Author: le-ann <lea22@mcmaster.ca>
Date:   Fri Aug 4 15:55:33 2017 -0400

    Added sample output for five random salmonella species

up to presence-absence matrix. Blast file

commit f2b20adab53ce3d0b70c6e3f271ede06d7a217fa
Author: le-ann <lea22@mcmaster.ca>
Date:   Wed Aug 2 19:44:39 2017 -0400

    Add Dr. Golding's genefamily11.pl program

commit 46ab289c137b8b49506e5cf6745ff863e5705c20
Author: le-ann <lea22@mcmaster.ca>
Date:   Wed Aug 2 19:42:11 2017 -0400

    Add Dr Golding's perl version of genefamily11.pl

commit a30db79671301a04bb7c723d2116147ee5189d0f
Author: le-ann <lea22@mcmaster.ca>
Date:   Tue Aug 1 16:08:23 2017 -0400

    First intiation of project
```