Two Weeks Ago

- Mostly file organization and getting the github ready and started. Bacterial files were being tracked and the remaining blastp runs should be completed over the time that I was away.

This Week

Obtained first indelmiss data on sample of five salmonella after reading through the documentation and trying to figure out how to work it.

This is the results that were shown:

```
Call:
indelrates(usertree = tree, userphyl = userphyl)

 5 taxa with 4818 gene families and 32 different phyletic gene pat
```

```
Groups of nodes with the same rates:
[[1]]
[1] 1 2 3 4 5 6 7 8 9
```

```
M1
$rates
          [,1]
mu 0.8269211
nu 0.8269211

$se
$se$rates
            [,1]
mu 0.02280058
nu 0.02280058
```

```
Loglikelihood for model M1 : -15138
AIC           for model M1 : -30278
BIC           for model M1 : -30277.61
```

```
M2
$rates
        [,1]
mu 0.6838036
nu 0.6838036

$p
[1] 0.1164538

$se
$se$rates
          [,1]
mu 0.01819685
nu 0.01819685

$se$p
[1] 0.005380595


Number of genes estimated as missing corresponding to the missing
 [1] 419

Loglikelihood for model M2 : -13700.91
AIC           for model M2 : -27405.82
BIC           for model M2 : -27405.04
```

```
M3
$rates
        [,1]
mu 0.8277939
nu 0.8337843
```

```
$se
$se$rates
          [,1]
mu 0.02429750
nu 0.06850801



Loglikelihood for model M3 : -15137.99
AIC           for model M3 : -30279.99
BIC           for model M3 : -30279.21
```

```
M4
$rates
         [,1]
mu 0.6716713
nu 0.5186157

$p
[1] 0.1164697

$se
$se$rates
          [,1]
mu 0.01766361
nu 0.05162335

$se$p
[1] 0.005380365


Number of genes estimated as missing corresponding to the missing
 [1] 419

Loglikelihood for model M4 : -13695.94
AIC           for model M4 : -27397.88
BIC           for model M4 : -27396.71
```

```
Time taken: 1.977 seconds.
```

Next Week

Tentative steps to plan the final R data frames output:

Columns: Bacteria Name -> Pangenome Sizes (Pan, Core etc.), Distance, Indelmiss (M1, M2, M3, M4) run1 run2 . . . run100

What we want to end up with: a presence-absence matrix with only the 20 species and their respective gene families.

1. Consult roaryinput lists: 100 group runs of 20 species each. Each run group contains 20 individuals which we want to obtain thelateral gene transfer rates from indelmiss.
2. Use the reference files to extract the proper faa files
3. Run reciprocal blasts on each of the 100 runs for 20 species
4. Weed out and run on genefamily11.pl (use TaxaNamesandprots.bash to generate directory of prot files, and then run the perl code, finally, this output should be able to be run on indelmiss)
5. After the results are obtained, I will have to write a code to read them all into R data frames