**Notes to Jon about his rotation project**

**Goal**: Infer the demographic history of common commensal gut bacteria

**ToDo:** Create a hoffman2 account under ngarud

**Papers to read:**

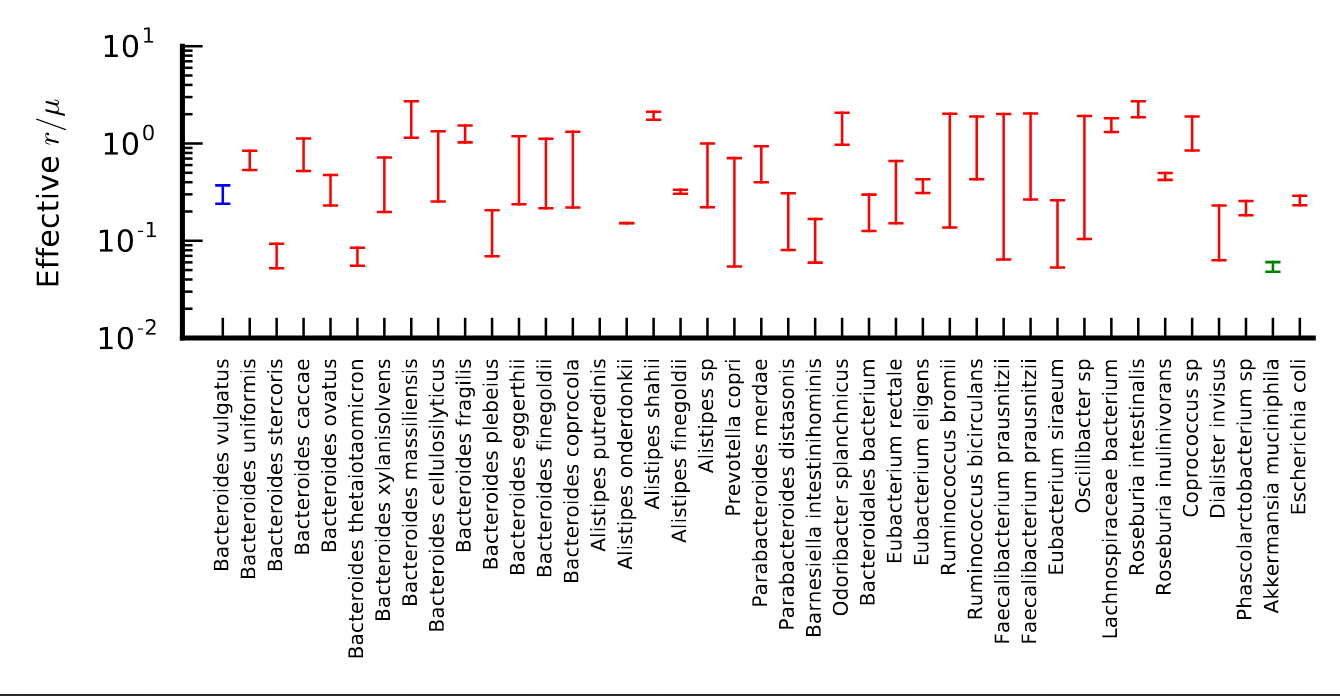
[https://academic.oup.com/mbe/article/30/4/881/1068944](https://academic.oup.com/mbe/article/30/4/881/1068944" \t "_blank) -- I think we will follow this paper rather closely. I like the way they did their demographic inference and we can use their models as a starting point.

<https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1003543>

(also see other papers by Caitlin Pepperell) – solid popgen on a human-associated bacterial species.

Garud, Good 2019 PLoS Biology – the original source of the data you will be working with.

As you read, think about the role of recombination in these analyses. Compare rho/mu estimate in Omar Cornejo’s paper with this supplemental figure from my paper. What do you make of it?



Step 1

* Create SFSs for syn and nonsyn sites for top 20 prevalent bacteria.
* Start with the merged data set from Garud, Good et al.
* We would want to subset samples into clades. The clades are already defined in the github.
* Will has code for the SFS – point Jon to that.
* Overlay on top an expected SFS (1/f)
* Goals: learn a bit about the code base, the data, the type of bacteria in the dataset, and look at SFS patterns.

Step 2:

* Using this same data, make a PCA plot. Color dots by geographic origin. There are data sets from Europe, North America, and China. Return to this plot later once we add on African and Peruvian data (see below).

Step 3:

* Run DaDi! As a starting point, test the models in Omar Cornejo’s model. In general, we can ask if a model of human demography is a good fit – this is a good starting point/null.

Step X:

* Once the above starts making sense, we can start to download new metagenomic data and process it. I’m thinking African data? This data could also be helpful for Will’s project, and so we can work with him on this.

Step 4:

* So at this point, the rest is all kind of future directions if we have time for it:
  + Run Fit DaDi to infer a DFE?
  + Once you have a demographic model, simulate it, plot LD, and see if it matches estimates in the data?
  + Try running DaDi on a 2D SFS using data from Africa?