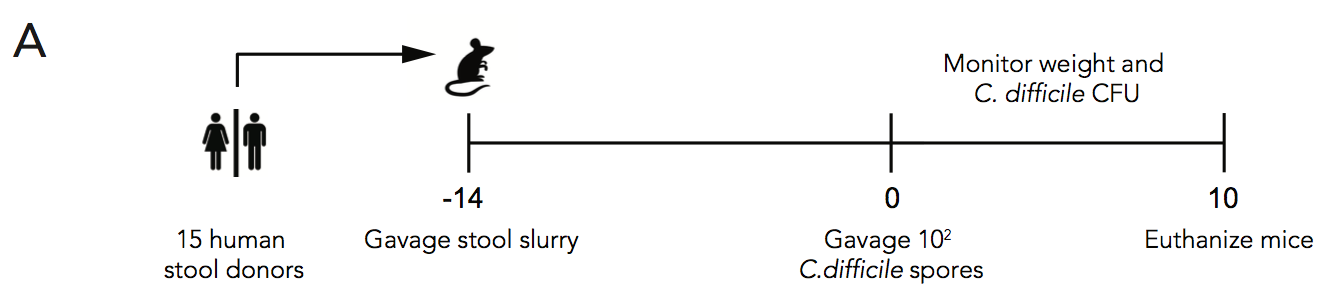
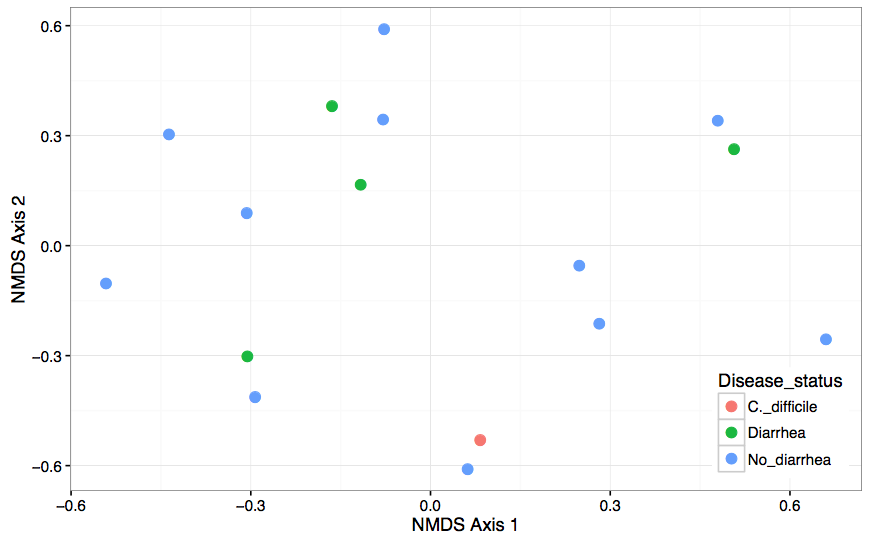
**Microbiota predict *Clostridium difficile* severity in germ-free mice colonized with human feces**

Kaitlin J Flynn\*, Nicholas Lesniak\*, Alyxandria M. Schubert, Hamide Sinani, anyone-else-ERIN-people?, Patrick D. Schloss

**Figures**

**Figure 1: Germ-free mice inoculated with human feces as a model for *C. difficile* infection**

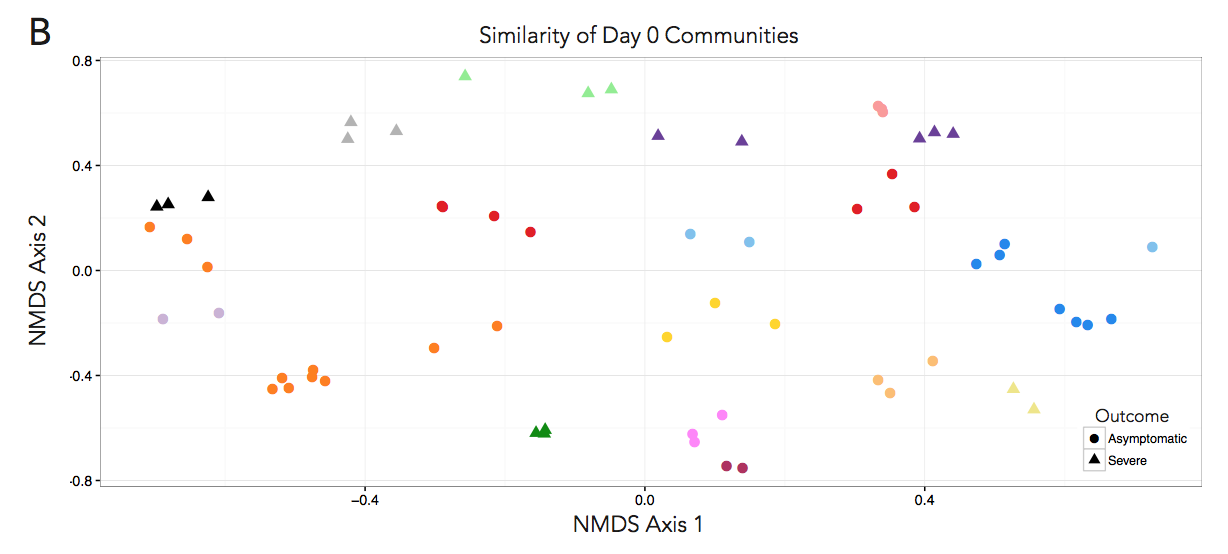
****

Figure 1. Germ-free mice inoculated with human feces as a model for *C. difficile* infection. A) Stool was collected from 16 healthy, diarrheal and CDI patients and inoculated into 3-4 germ-free mice per donor by oral gavage. After allowing the community to stabilize for 14 days, mice were orally gavaged with 100 spores of *C. difficile* strain 431. Weight and stool CFU was monitored for up to 10 days post infection. B) NMDS ordination of donor stool communities prior to inoculating mice. Each point represents one donor and donors are colored by clinical diagnosis. C) NDMS ordination of the stool communities on day 0. Each symbol represents one mouse and is colored by donor. Circles represent mice that survived the 10 days of infection and triangles represent those who suffered severe disease.

Figure to dos:  
A) Consider making this timeline a bit more streamlined/using R like Jenior’s timelines look

B) Do ADONIS to show no differences or mantel test, report value on figure or in text? Make sure these are done with 2D NMDS files and not 3D\*\*\*

C) Do mantel test for correlation of distances between severe/mild. Change legend to be mild/severe. Decide if colors are distinguishable enough (I think they are)

**Figure 2. *C. difficile* infection dynamics**

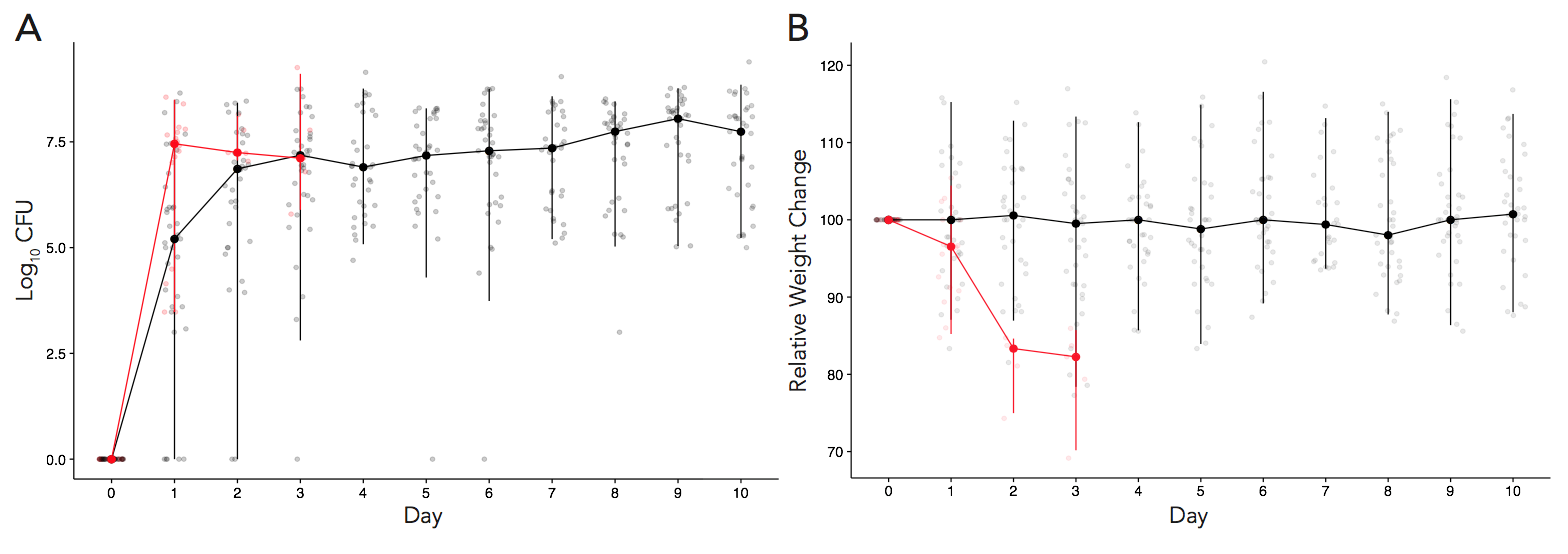
****

Figure 2. *C. difficile* infection dynamics. A) *C. difficile* CFU was enumerated by plating of mouse stool pellets daily. Each point represents a mouse and the lines represent the median (?) of all cages and error bars are interquartile ranges. Red lines and points correspond to mice that succumbed to severe disease, black lines and points correspond to mice that had mild or no disease. B) Mouse weights were recorded daily percent weight loss calculated for each mouse. Data presented as the median (?) of mice in each cage. Mice that succumbed to severe infection typically lost a significant amount of weight by day 1 or 2 post infection. Red lines correspond to severely ill mice, black to mice with mild disease.

Figure to dos:

A + B) Decide if these are the final values/error we want to represent, add back mild/severe legend

**Figure 3. Microbial community on day 0 predicts future *C. difficile* CFU**

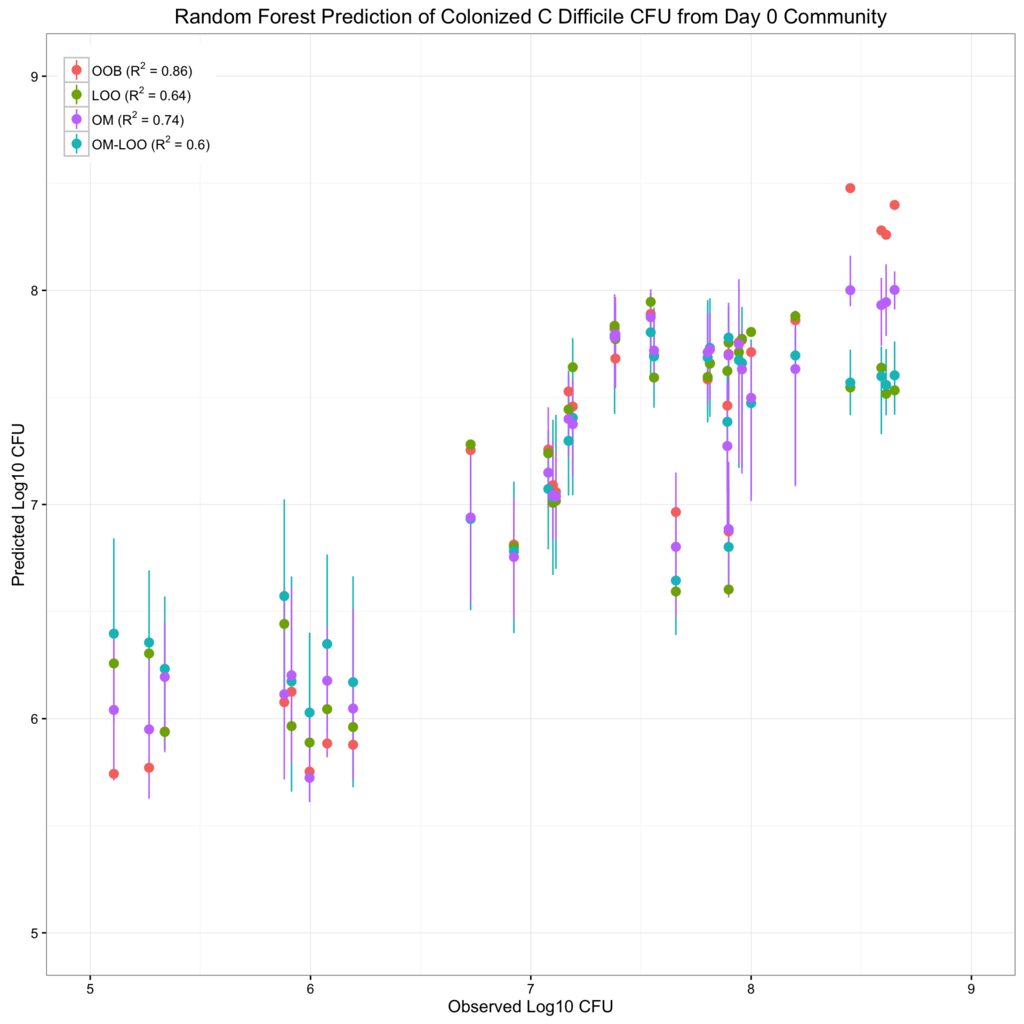


Figure 3. Random Forest prediction of *C. difficile* colonization level. A) Day 0 microbial community members above 1% relative abundance were used to predict median log10 CFU of *C. difficile* after colonization. OTUs were chosen such that they were not predictive of cage or donor. Each point is a mouse colored by cage. B) Partial dependency plots of the top six predictive OTUs. Line displats the partial dependence of log10 CFU on the relative abundance of each predictive OUT. Each median log10 CFU is plotted against its relative abundance for each predictive OTU.

Figure to-dos

A) make points bigger, probably entire plot can be smaller, B)New partial dependency representation/graph style? Nick

Meeting notes 12/5:

* don’t use term steady state
* model optimization- ntrees, etc
* if only show one model, show OM-LOO (most conservative)
* Partial dependency plots: for 4 of them they are correlated features
* Entire model in A is built on 6 OTUs, Nick builds correlation network first, then builds model on 6, could be hurting the model though by being too stringent (picked via importance in caret package)
* But by picking 6 does that affect model, is model different depending upon correlated OTU that we use?
* Build model on **all** data all mice, then test if overfit with OM-LOO.
* M1- build on all OTUs (300 OTUs)
* M2- all OTUS +donor +cage—do we get donor and cage out as being predictive? If yes, can affect overfit. If not, throw it out and don’t worry about that.
* M3 – uncorrelated
* M4 – uncorrelated + cage + donor
* Do 1 and 3 differ in output? If no, use M1, easier to interpret. Then compare vs cage/donor models
* If do M1, get list of features F1. Then can create M5, use all OTUs to predict cage. M6, use all to predict donor. Is there overlap between F5, F6?
* Nick will build models, perhaps we should double check with Jenna

**Figure 4. Random forest predicts CDI severity from day 0 microbiome.**

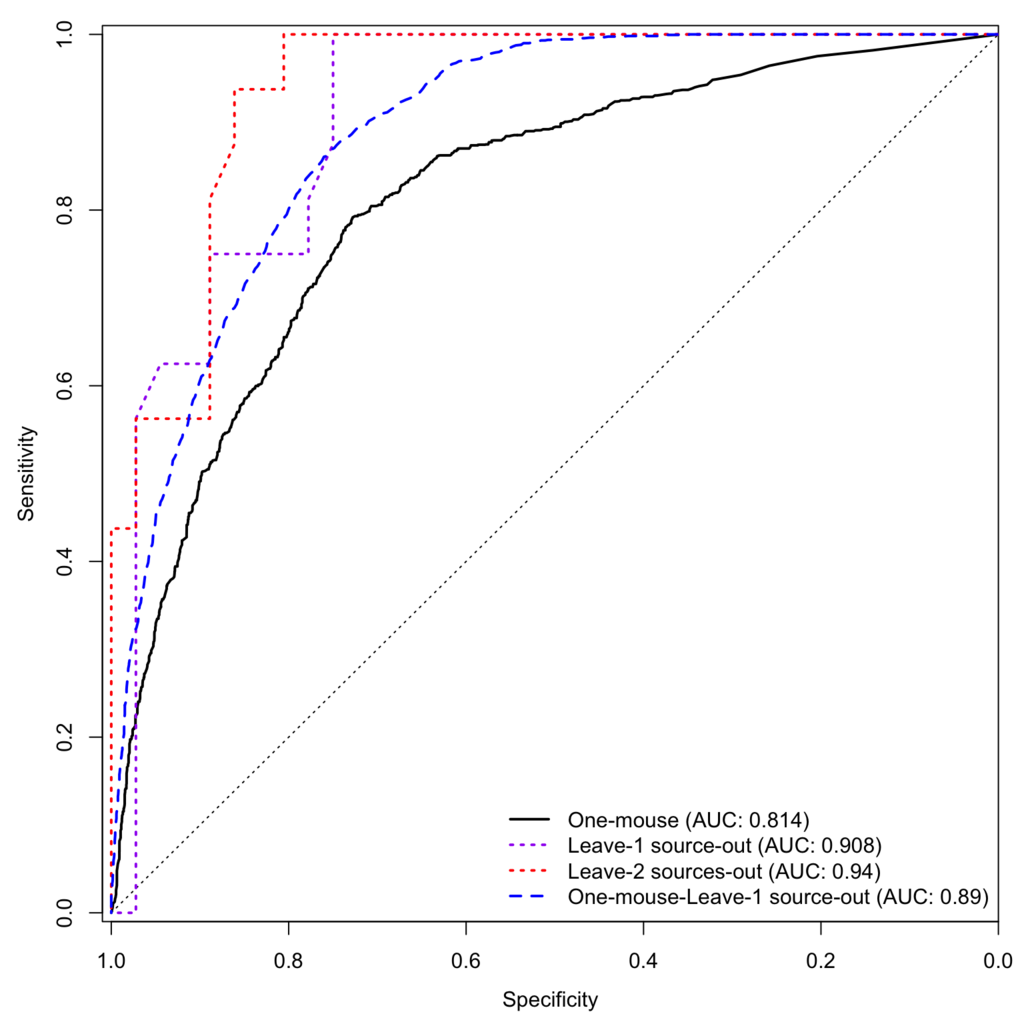


Figure 4. Random Forest prediction of CDI severity. Day 0 microbial community members above 1% relative abundance were used to predict disease severity. OTUs were chosen such that they were not predictive of cage or donor. Predictive classification tested via 10-fold (gray), leave-one-cage-out (purple dashed) or leave-one-mouse-out (blue dashed) models are displayed in A). B) Partial dependency plots of most predictive OTUs. Line displays the partial dependence of log10 CFU on OTU relative abundance. Points are the OTU relative abundance of each mouse colored by outcome (red, severe, black, mild).

Figure to –dos

A) decide on final models to present

B) new partial dependency plots/format, Nick?

Do all OTUs into this predictive model, but then cage and donor are going to predict severity.

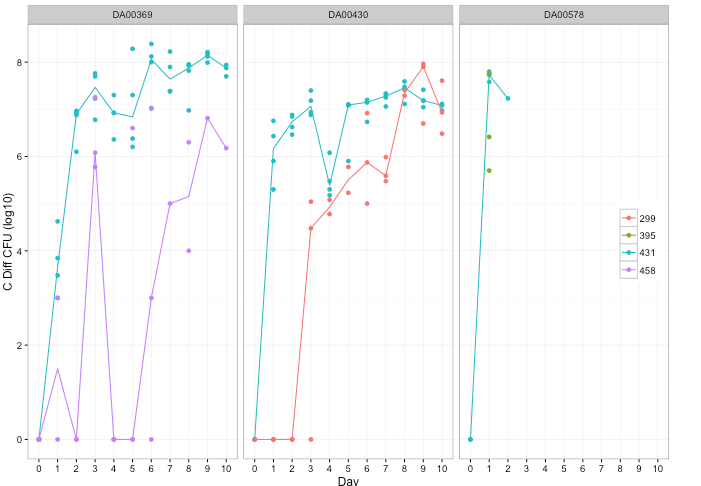
Or what happens when you do 1 mouse LOO with all OTUs?

**One mouse- training on single mouse per cage**

**Leave one source out- leave one donor out, train with only 15 sources**

**Leave 2 out- leave out 2 donors**

**Figure 5. Propensity for severe CDI is community-dependent**

****

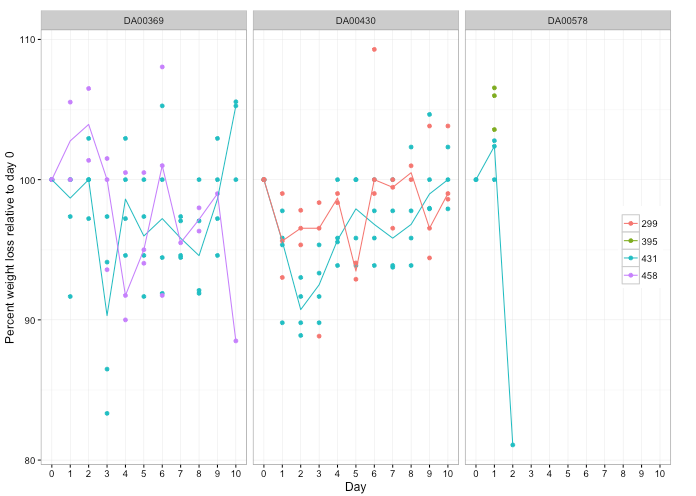
****

Figure 5. Infection of mice with different *C. difficile* strains. 3 strains of *C. difficile* were used to infect mice colonized with susceptible (DA00578) or resistant (DA00369, DA00430) human donor stool. A) *C. difficile* stool CFU was enumerated over 10 days. B) Percent weight loss was calculated each day for each mouse. In both plots, each mouse is a point and lines represent the mean of each cage.

Figure to-dos:

* + 1. decide how we want to label donors (numbers aren’t really used elsewhere)
    2. Need to add lines for green points in DA00578 plots

Take this out!

**Supplement**

**Table S1: Mouse day 0 communities by donor genera (avg + stdev of cage)**