Machine learning identifies unique taxa differentiating proximal and distal human colonic microbiota
Kaitlin Flynn
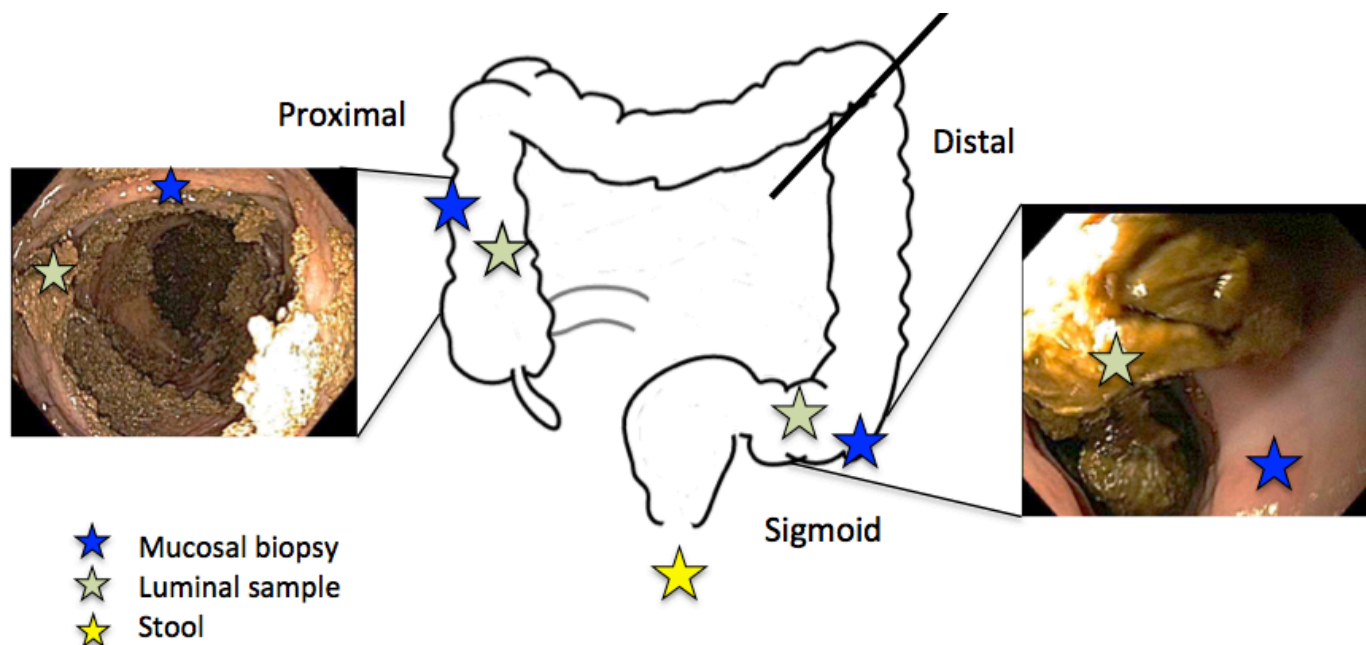Figure set as of 6/12/17

Figure 1:

Figure 1: Sampling strategy. A flexible sigmoidoscope was used to sample the distal colonic luminal contents and mucosa. The scope was inserted ~25 cm into the patient and biopsy forceps were used to sample the luminal contents (green star). A separate set of biopsy forceps was used to sample the distal mucosa (blue star). The sigmoidoscope was removed. A pediatric colonoscope was inserted and used to access the proximal colon. Biopsies were taken of the proximal luminal contents and mucosa as described. One week prior to the procedure stool was collected at home and brought in to the appointment.

To do for this figure/analysis:
- put into photoshop or something to get the proper resolution
- probably change the colors of the stars, talk to Nick.
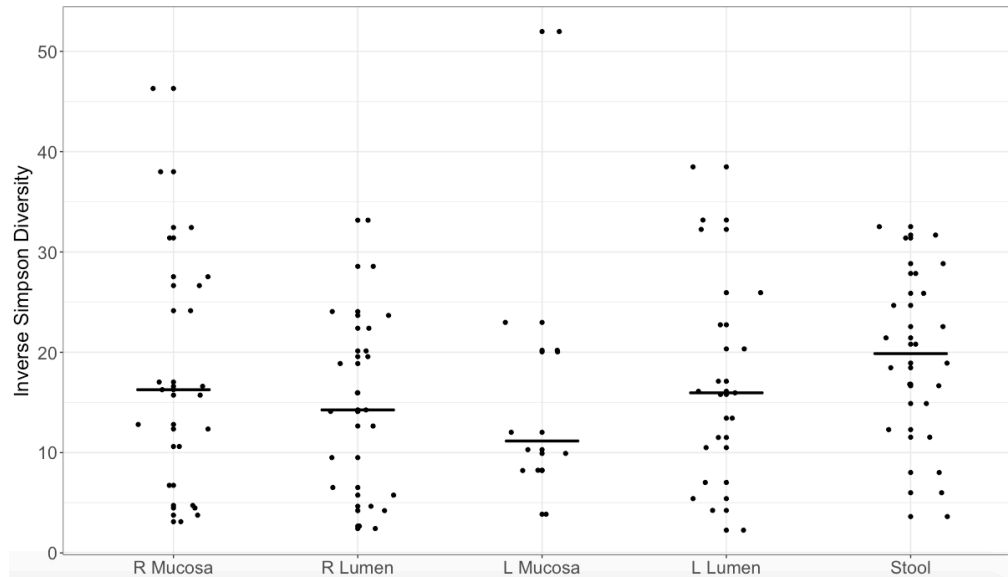
Figure 2:

A
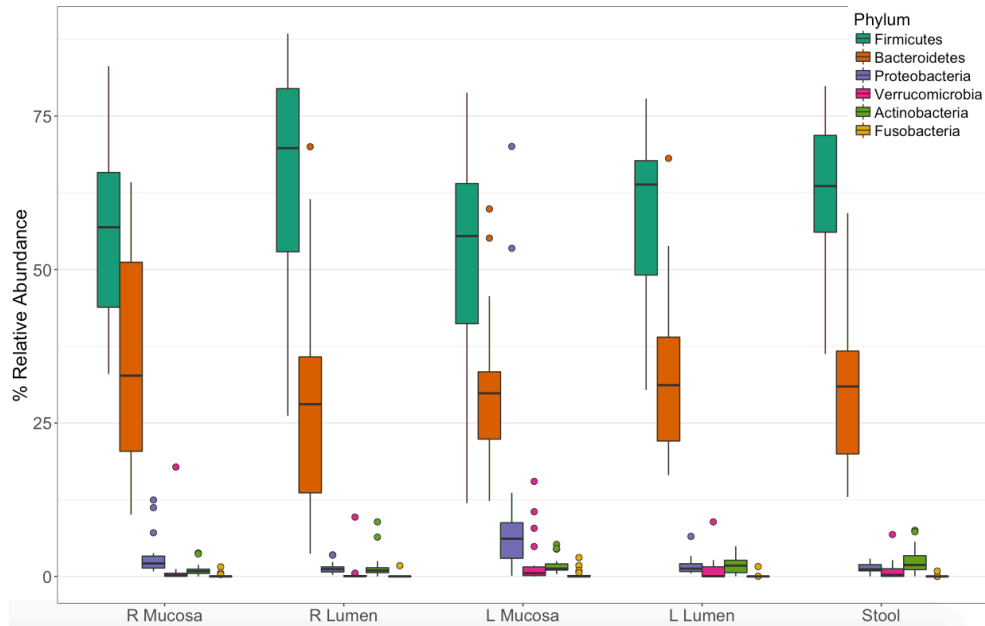
B

Figure 2: Microbial membership and diversity of the proximal and distal human colon. A) Relative abundance of the top five bacterial phyla in each sampling site. Each box represents the median and confidence intervals. B) Simpson diversity of the microbial communities at each location. The lines represent the median values.

To be done for this:
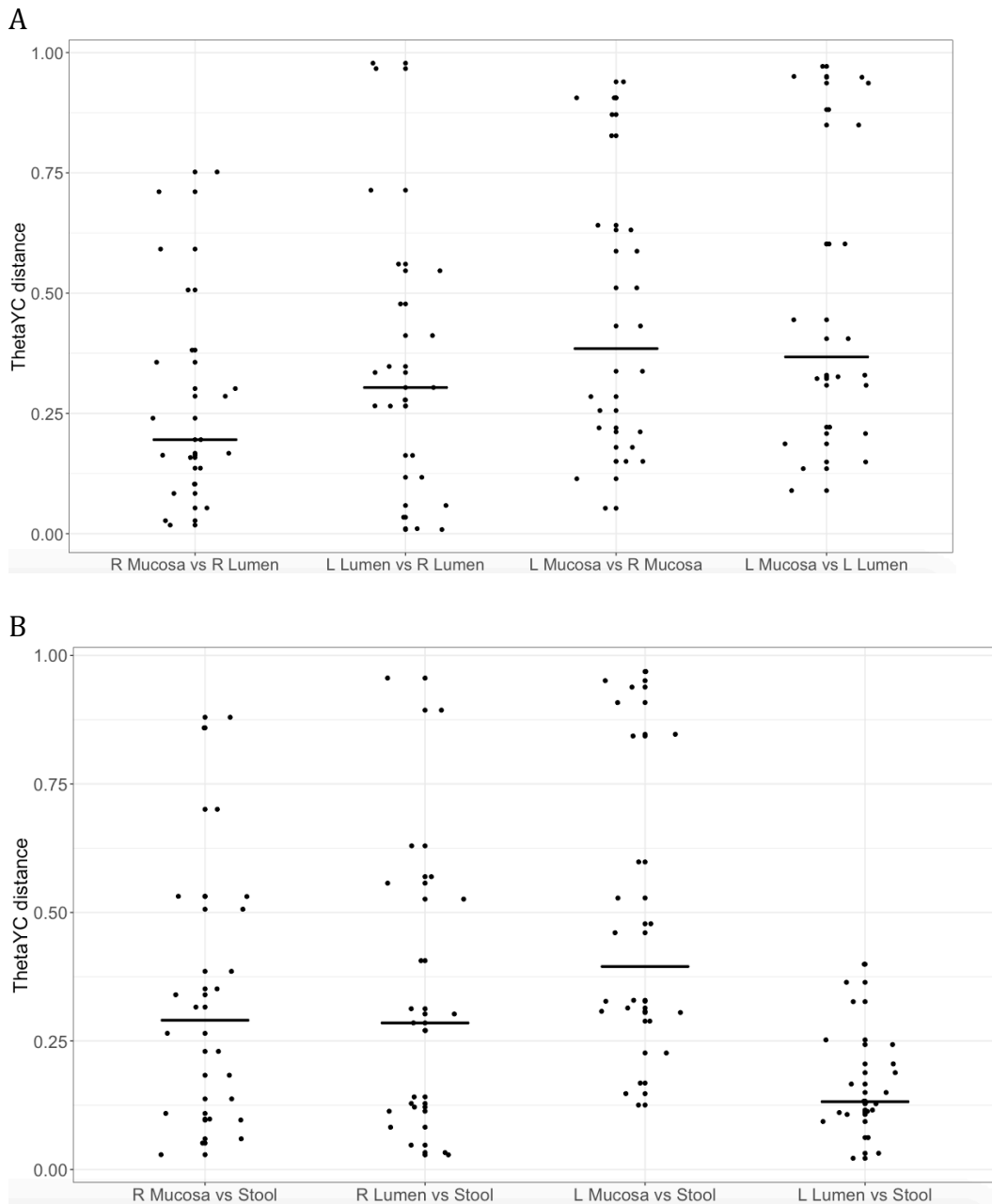- export as final sizes/ resolutions / all on one plot

Figure 3

A

B

Figure 3: Similarity of microbial community structure between sites of the gut. ThetaYC distances are shown for interpersonal similarities between two sites – each point represents one individual. In (A), comparisons of the right and left mucosal and lumen are shown. In (B), comparisons of each site to the exit stool are shown.

TBD:
- make the new plot with pat suggestions.
- Median point for each that shows the distance between pt1 and all other pts. Then should only have 20 points or so for each
- Plot as a box and whiskers plot
- add stats * or lines on here somewhere
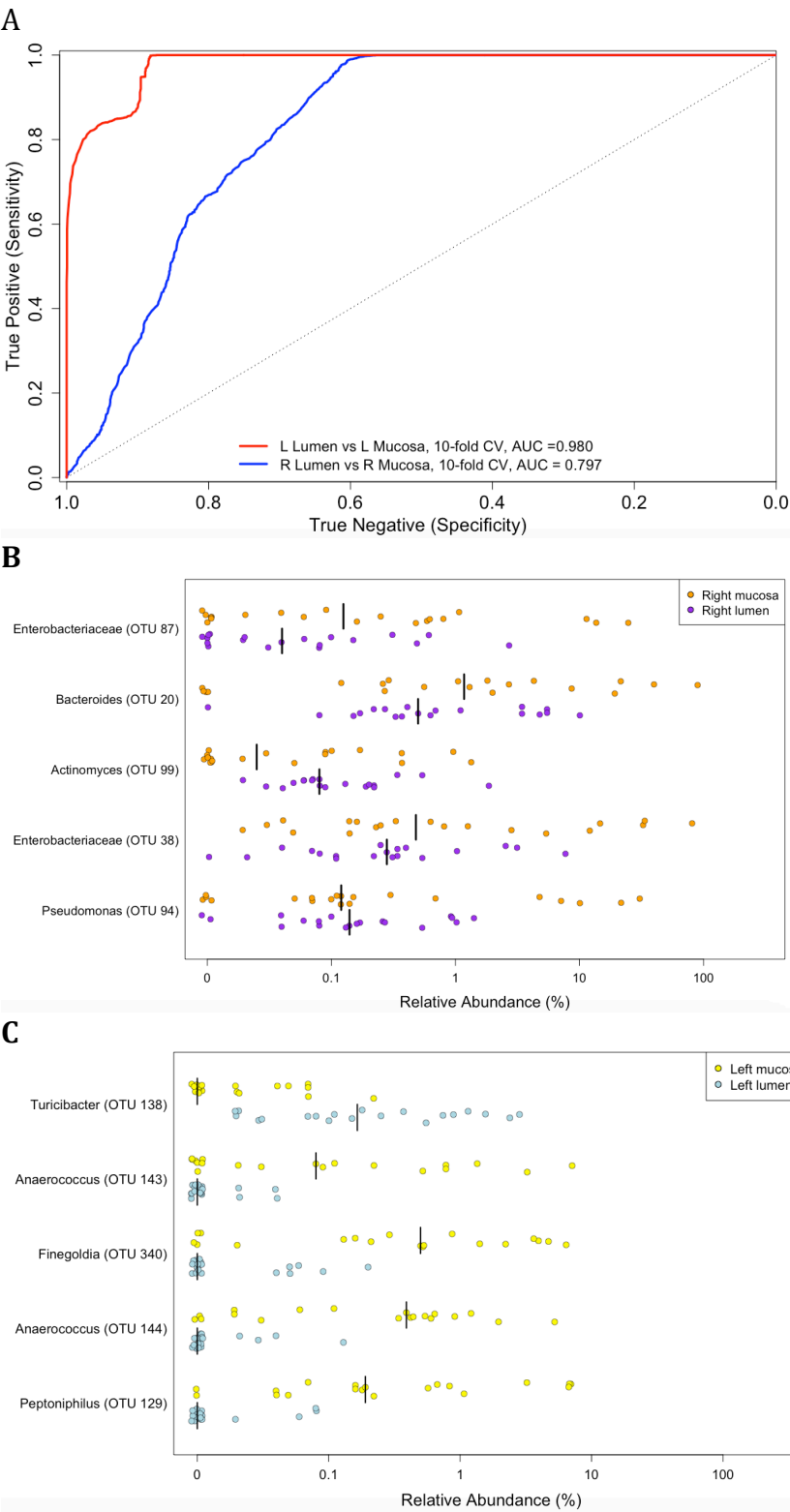
Figure 4:

A

B

C

Figure 4 legend: Random Forest classifies the mucosa and lumen of each side of the colon. A) Receiver Operator Characteristic curves are shown for the 10-fold cross validation of the Random Forest model classifying lumen and mucosal samples for the left and right sides of the colon. (B) Top five OTUs that are most important for the classification model for the left mucosa and lumen (B) and the right mucosa and lumen (C).

TBD:

- lumpy model- discuss w Geof
- discuss Xopt and number of top OTUs to display with Marc
- finish plot tweaks/export as one plot
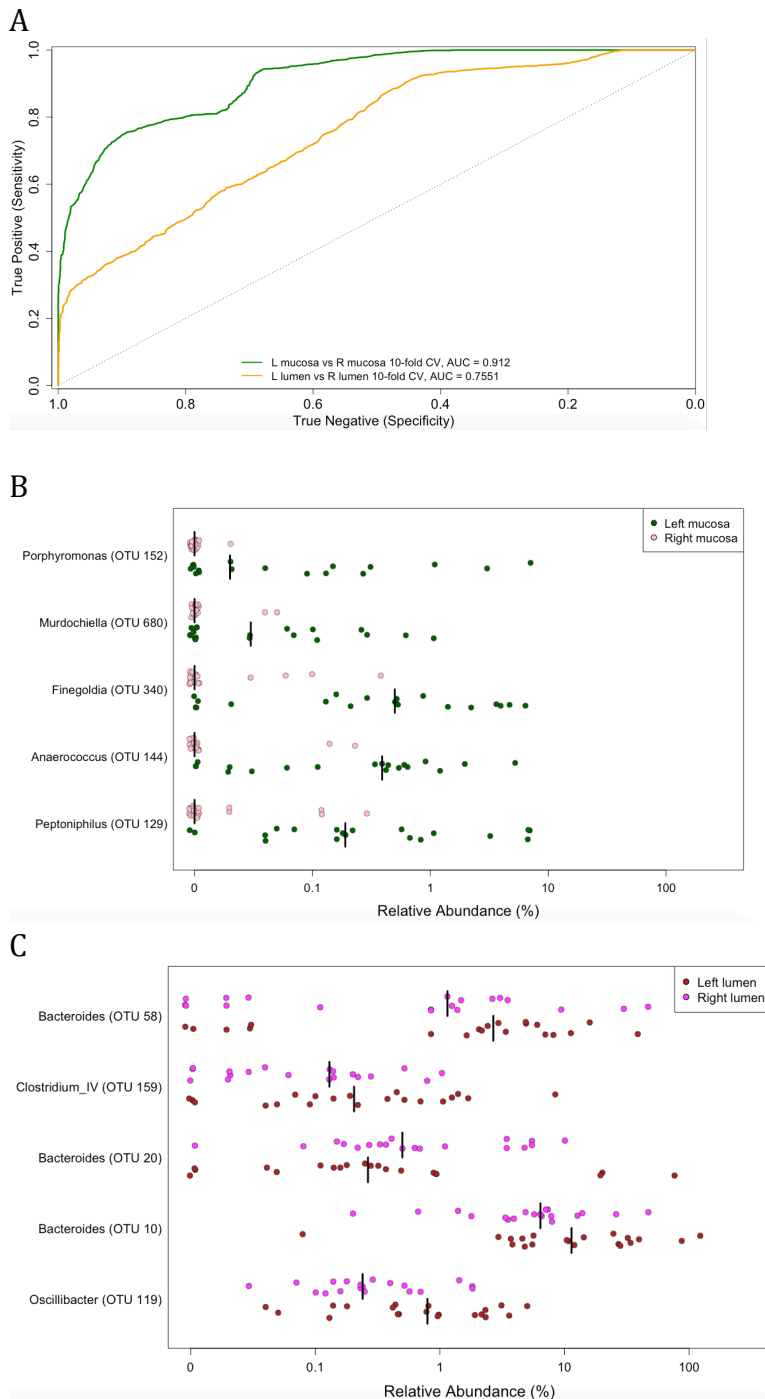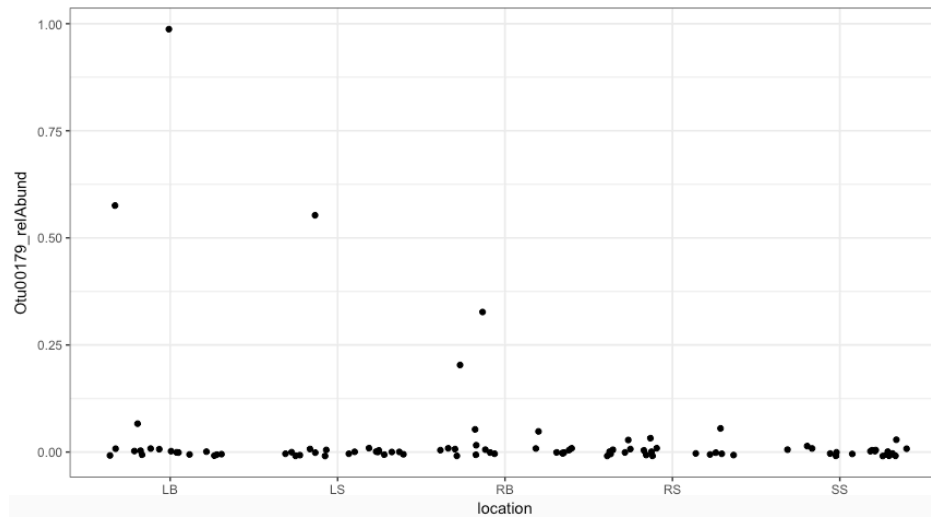
Figure 5:

A

B

C

Figure 5 legend. Random Forest classifies the left and right sides of the colon. A) Receiver Operator Characteristic curves are shown for the 10-fold cross validation of the Random Forest model classifying left lumen versus right lumen (orange) and left mucosa vs right mucosa (green). (B) Top five OTUs that are most important for the classification model for the left and right mucosa (B) and the left and right lumen (C).

TBD:
- lumpy model- discuss w Geof
- discuss Xopt and number of top OTUs to display with Marc
- finish plot tweaks/export as one plot

Figure 6

A: F nucleatum abundance
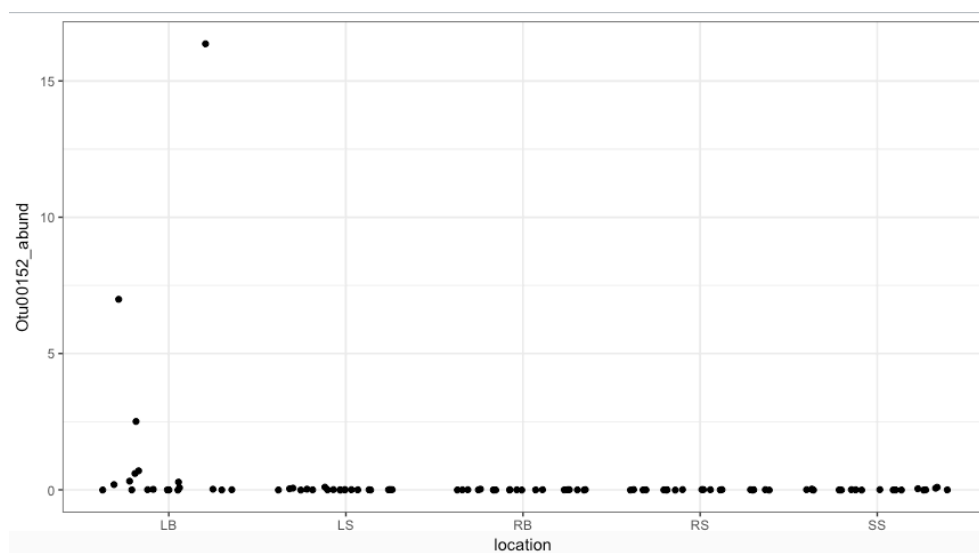

B: P. asacharolytica abundance

Figure 6 legend: Location and abundance of cancer-associated OTUs. Relative abundance was calculated and plotted by sample site for each OTU of interest: (A) *Fusobacterium nucleatum* and (B) Porphyromonas asacharolytica

TBD:
- decide on final figures or table
- include other OTUs?

Figure 7, possible model/schematic of results?