Building Tools for Modeling Disease States of the Human Gut Microbiome

Kelly L. Sovacool
Nov. 2021

Specific Aims

Changes in the taxonomic composition and metabolic activity of human microbiomes have been observed in several diseases. In the case of colorectal cancer (CRC), evidence of toxigenic activity by gut microbes implies that these changes are not only a response to disease, but may also play a role in disease etiology. Taxonomic composition is commonly defined by amplicon sequencing of the 16S rRNA gene and clustering sequences into Operational Taxonomic Units (OTUs). Previous studies have built OTU-based machine learning models to classify stool samples as normal or cancerous, to serve as a less invasive diagnostic tool for CRC than colonoscopy. Efforts to find consistent changes in taxonomic composition of microbiomes between normal and dysbiotic states have found mixed success, in part because interpersonal variability in taxonomic composition sometimes exceeds the variability between disease states. Variability of microbiome composition between individuals with the same disease status may be explained by functional redundancy, where different microbial species carry out the same functions and thus can replace each other with little effect on the overall function of the community.

16S tells us the taxonomic composition, i.e. which microbes are present in a sample. Can then compare taxonomic composition between individuals with different disease states (e.g. colorectal cancer). There is high interpersonal variability in taxonomic composition of the human gut microbiome. Ordination plots (e.g. NMDS) do not show any separation between disease states. Possibly because of functional redundancy: different microbes can perform the same functions. ML models built on taxonomic composition do an okay job of discriminating stool samples from healthy and cancerous colons (AUROC 0.7 for RF model on 490 samples). Can use metagenomics to build profiles of functional (potential) composition, i.e. what the microbes are capable of doing. omics integration: can we identify which microbes are capable of performing which functions? with metagenomics, only a small fraction of microbial genes are annotated with known functions.

Aim 1: Determine the performance of OptiFit in clustering reference-based OTUs against *de novo* OTUs.

Hypothesis: OptiFit clusters reference-based OTUs at nearly the same quality as OptiClust and with faster execution speeds.

A. *De novo* cluster reference databases into OTUs with OptiClust, then cluster datasets to the database OTUs with OptiFit.
B. *De novo* cluster half of the sequences in each dataset with OptiClust, then cluster the remaining sequences to the *de novo* OTUs with OptiFit.
C. Compare the OTU quality of the two OptiFit strategies above versus that of *de novo* clustering the datasets with OptiClust.

Aim 2. Build machine learning models with OTUs and clinical features to predict the severity of CDI and identify important micriobial features.

Hypothesis: OTU- and clinical-based machine learning models can be used to predict the severity of CDI.

A. Cluster 16S rRNA amplicon sequences from the R21 cohort dataset into *de novo* OTUs.
B. Build models with OTUs and clinical features using the R21 dataset to predict the IDSA severity score of CDI cases and identify the most important features in model performance.
C. Use OptiFit to fit the ERIN cohort dataset into the R21 OTUs, then use it as a validation test set for the best R21-based model.

Aim 3. Explore the impact of functional redundancy on the perforamnce of models to predict CRC and CDI.

Hypothesis: Incorporating functional potential along with OTUs as features improves the performance of disease prediction models.

A. Assemble metagenomes and identify known gene functions in CRC and CDI datasets.
B. Characterize the presence of functional redundancy within samples.
C. Build ML models using OTUs, known metagenomic functions, or both as features and compare performance.