Developing tools for characterizing the taxonomic and functional composition of host-associated microbiomes

Kelly L. Sovacool
Feb. 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).

LC-MS/MS tells us the metabolites present – the inputs and outputs of chemical reactions. Can use metabolomics to build profiles of functional composition, i.e. what the microbes (and the host) are doing. omics integration: can we identify which microbes are performing which functions?

with metabolomics, only 2 percent of untargeted lc-ms/ms features can be annotated with known metabolites. why should we throw out that data? microbial ecology gets around this type of problem by using database-independing approaches for clustering sequences into OTUs. let's take this approach and apply it to metabolomics, cluster ms features into Operational Metabolomic Units. with metagenomics, only a small fraction of microbial genes are annotated with known functions.

Aim 1: Build a tool to cluster untargeted LC-MS/MS features into Operational Metabolomic Units.

Hypothesis: LC-MS/MS features can be clustered into chemically similar OMUs.

A. Calculate cosine similarity scores for all pairs of features and cluster features into Operational Metabolomic Units (OMUs) using OptiClust.
B. Label OMUs based on the known compounds they contain using GNPS.
C. Test the tool using LC-MS/MS data from known compounds to determine the optimal cosine threshold for OMU clustering.

Aim 2: Assess the impact of integrating metabolic profiles with taxonomic profiles on CRC classification.

Hypothesis: Incorporating both OMUs and OTUs improves ML model performance for CRC classification over using only OTUs.

A. Build taxonomic profiles with OTUs from 16S rRNA gene sequences and metabolic profiles with OMUs from LC-MS/MS features.
B. Compare profiles of taxonomic and metabolomic composition within and between disease states.
C. Build machine learning models to classify samples as CRC or non-cancerous with OTUs, OMUs, or both and compare model performance.

Aim 3. Assess the impact of integrating active metabolites with functional gene potential on CRC classification.

Hypothesis: Using all measured metabolites instead of only known bacterial metabolites improves the classification modeling of samples as CRC or non-cancerous.

A. Annotate compounds from untargeted mass spectrometry with the GNPS database and select those known to be products of bacterial metabolic pathways with the MetaCyc database.
B. Calculate the intersection of pathways associated with active metabolites and the pathways from functional potential profiles from whole metagenomes.
C. Build machine learning models to classify samples as CRC or non-cancerous using all OMUs or only confirmed active bacterial metabolites and compare performance.