Specific Aims

background and motivation, testable hypothes(is/es), two specific aims, and a summary of the proposed research. [1]

bacteria that aren't taxonomically closely related can be capable of producing some of the same gene products. just like different charismatic megafauna can have the same niche in an ecosystem (predator X and Y both eat prey Z), different bacteria can cover the same niche. looking only at taxonomy may not be enough to understand the role and response of the gut microbiome in health and disease processes. some studies couldn't find strong signal in 16S OTUs to differentiate CRC from normal, but found different metabolites [Weir 2013]

some studies use PiCrust in an ill-conceived attempt to predict the functional potential of communities from 16S sequence [Perera 2018]. That's a dumb idea. Why use a prediction when you can get the real thing? Let's use metagenomics and metabolomics instead, since metagenomes tells us the actual genes capable of producing products, and metabolomics tells us the functions that the microbiome are performing at the time of sampling.

- "Functional redundancy is very important because it introduces the idea that, up to a point, species are interchangeable in a given microbiota in terms of function" (Moya and Ferrer 2016) - "Another issue is that there is an overlap between metabolites that are externally derived, produced by host metabolism, and produced through microbial cometabolism, which adds to the difficulty in setting up these types of databases; DAS is one such metabolite." - Johnson 2016 - interpersonal variation of taxonomic composition can outweigh the variation observed between disease states. - "Community-level function is often more conserved than community composition" [humann2 paper] - get taxonomic community composition and functional community structure - should get fcnl comm struct from metagenomes and/or metabolomes? - find one paper that classifies fcnl potential from metagenomes, another from metabolomes. use both techniques and compare? - Royalty and Steen 2020 preprint: quantify functional redundancy from metagenomes. "We define functional redundancy as how evenly taxa in a community contribute to the total quantity of trait exhibited by the community" - "definition of functional redundancy indicating the mere ability of multiple distinct organisms to perform a specific function... is practical" [louca 2018]. - but if we have metabolomes, we can verify - are these more or less correlated in CRC compared to healthy? - wrt carcinogenic functions / genes / metabolites - comparisons of taxonomic composition, functional gene composition, and metabolomes: - within disease states - between disease states - taxonomic vs functional gene composition - metabolomic vs functional gene composition - limitations: - temporal changes in metabolites sampled. this study is not longitudinal.

- TODO: look at notes from talk by Robert (lastname?) of MSU from last year. used GNPS and QIIME in gut microbiome study.

Hypotheses: - There is functional redundancy in CRC communities. - When CRC communities have different taxonomic composition, they have similar functional composition. - Functional redundancy explains the taxonomic variability between individual hosts with the same disease status. - Clustering communities based on functional potential improves classification into disease states compared to clustering based on taxonomic composition. - When there is not a significant difference in taxonomic composition of CRC and normal communities, there is a significant difference in their functional potential and in their metabolomic profiles. - Variation in functional profiles within disease states is less than the variation in taxonomic profiles within disease states. - Assess how well the putative microbial metabolites match functional gene profiles from metagenomes and the difference in CRC vs non-cancerous.

Questions: - is there correlation between taxonomic composition and functional potential (metagenome composition) of microbiome? - does the idea of functional redundancy hold up during CRC? - is concordance stronger/weaker in CRC vs normal? - where OTU composition doesn't show a difference between CRC and normal [Weir 2013], does functional potential show a significant difference? - does taking fcnl redundancy into account improve classification of samples into CRC vs normal? even in studies that did find a difference? - how many genes (functional potential) have corresponding

metabolites actually shown? - some toxigenic gene products produced by bacteria are known. can we use GNPS to identify unknown features as potentially toxigenic products based on their similarity to known toxigenic ones? - does functional redundancy from metagenomes correlate with actual metabolites? - find intersection of biosynthetic genes (metagenomes -> antiSMASH) and known metabolites that are present. how well do they match?

Aim 1. Assess the importance of functional gene redundancy of the gut microbiome in CRC.

Hypothesis: Using functional gene profiles instead of only taxonomic profiles improves the classification modeling of samples as CRC or non-cancerous because of functional redundancy in the gut microbiome.

A. Build taxonomic profiles with OTUs from 16S rRNA gene sequences and build profiles of functional gene potential from metagenomes.
B. Compare taxonomic composition to functional gene potential of microbiomes within and between disease states to determine presence and degree of functional redundancy.
C. Build machine learning models to classify samples as CRC or non-cancerous with taxonomic composition, functional gene potential profiles, or both as model features and compare performance.

Aim 2. Validate functional gene potential with active metabolites to improve classification of CRC microbiome samples.

Hypothesis: Using active metabolic pathways confirmed with mass spectrometry instead of all potential metabolic pathways from metagenomes 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.
C. Build machine learning models to classify samples as CRC or non-cancerous with all potential metabolic pathways or only confirmed active metabolic pathways as model features and compare performance.

References

[1] S. Bikel, A. Valdez-Lara, F. Cornejo-Granados, K. Rico, S. Canizales-Quinteros, X. Soberón, L. Del Pozo-Yauner, and A. Ochoa-Leyva. Combining metagenomics, metatranscriptomics and viromics to explore novel microbial interactions: towards a systems-level understanding of human microbiome. *Computational and Structural Biotechnology Journal*, 13:390–401, Jan. 2015.