BRIGHAM AND WOMEN'S HOSPITAL

HARVARD MEDICAL SCHOOL

181 Longwood Avenue
Boston, Massachusetts 02115-5804

Department of Medicine
Channing Division of Network Medicine

Channing Microbiome Seminar

April 26 (Friday), 2019, 11am @ 5th-floor conference room

Isabel Fernández Escapa, PhD

Forsyth Institute

Achieving species-level resolution from 16S rRNA gene short-read data using a high-resolution training set with the RDP naïve Bayesian Classifier

Microbiome studies must achieve species-level resolution for clinical relevance, since both harmless and pathogenic species of the same genus are often native to a body site. Moreover, there is a need for large-scale molecular epidemiological studies of the microbiota of thousands of humans to gain clinically useful insights. The cost of which is feasible with 16S rRNA gene-based short-read sequencing. Therefore, we developed a microbiota analysis pipeline that achieves species-level resolution from 16S rRNA gene short-read sequences. We focused on bacterial microbiota of the human aerodigestive tract (nasal passages, sinuses, throat, esophagus, and mouth) because it has the potential to reveal new insights for promoting human health. We first overcame technical limitations and successfully Illumina sequenced the 16S rRNA gene V1-V3 region, the most informative for classifying bacteria native to the human aerodigestive tract. We parsed sequences into high-resolution Amplicon Sequence Variants (ASVs) using Minimum Entropy Decomposition (MED) or Divisive Amplicon Denoising Algorithm (DADA2). To accomplish accurate and optimally informative taxonomic assignment to these ASVs, we generated a high-resolution V1-V3 region training set from our actively curated, and comprehensive, expanded Human Oral Microbiome Database (eHOMD) for use with the Ribosomal Database Project naïve Bayesian Classifier. We also generated a full-length eHOMD 16S rRNA gene training set to analyze PacBio-sequenced data, which we used to validate the representation of species in our training sets. Our approach facilitates species/supraspecies taxonomic assignment to ASVs derived from both short-read and full-length 16S rRNA gene sequences, enhancing the utility of 16S rRNA gene sequencing.

Bio: As Director of Forsyth's Sequencing and Bioinformatics Core, Isabel Fernández Escapa assists Forsyth investigators in facing the technological challenges of this field. Escapa's most recent research focused on the expansion of the Human Oral Microbiome Database (eHOMD) to include bacteria resident to the entire human aerodigestive tract. This project, in collaboration with Drs. Katherine Lemon and Floyd Dewhirst, has allowed Forsyth bioinformaticians to develop and validate a comprehensive workflow for analysis of 16S rRNA gene tag sequencing data that uses a training set derived from the expanded database (eHOMD). This eHOMD custom training set, in conjunction with high-resolution algorithms, allows species-level identification of most bacterial taxa present in human aerodigestive microbiome samples. Thanks to this new pipeline the Forsyth Institute positions itself at the frontier of oral and nasal microbiome research, offering an outstanding quality 16S rRNA gene microbiome workflow. Escapa aims to develop the Sequencing and Bioinformatics Core into a key asset that could support Forsyth basic, translational, and clinical research. "Our goal is to leverage our knowledge by generating high quality proprietary microbiome data that could be mined for the discovery of new therapeutic/diagnostic tools."

Hosted by Yang-Yu Liu