

TITLE: VARSMETAGEN: AN ONLINE PLATFORM FOR ANALYSES OF MICROBIAL NGS DATA ENABLING PRECISION MEDICINE IN THE DIAGNOSIS OF INFECTIOUS DISEASES

AUTHORS: DORLASS, E. G.¹; AMGARTEN, D.¹; RIYUZO, R.¹; MANGOLINI, T. G.¹; SOARES, A. C.¹; SILVA JUNIOR, W. J.¹; CERVATO, M.¹

INSTITUTION: 1. HOSPITAL ISRAELITA ALBERT EINSTEIN (AV. ALBERT EINSTEIN, 627/701 - MORUMBI, SÃO PAULO - SP, 05652-900).

ABSTRACT

Sequencing technologies have evolved at astonishing rates over the past 20 years. As a consequence, prices to generate one-million base pairs (Mbp) have decreased to the point where it became accessible for the main public with important applications to disease diagnostics. In this work, we present the Varsmetagen platform for analyses and interpretation of microbial NGS data, which will allow users to execute bioinformatics pipelines on cloud infrastructure as well as interpret results and easily generate reports. An user-friendly online interface allows users to execute the steps of: 1. Create runs, upload files (FASTQ, BAM or CRAM) and assign patient information to individual samples. 2. Select a bioinformatics pipeline to be executed at the samples. 3. Inspect metrics for quality control and approve or disapprove samples. 4. Inspect charts and tables with information of diversity and abundance. 5. Interpret results to select and validate potential pathogens linked with patient symptoms. 6. Generate reports, which may be used for clinical or publication purposes. Currently, we have implemented two pipelines: Virome and Metagenomics pipeline. Briefly, it takes raw reads from FASTQ or BAM files, perform quality control to remove poor quality sequences, remove host contamination by mapping to the Human reference genome, perform taxonomic identification of remaining sequences using Kraken2 with a customized database, and normalize kraken2 counts by applying a bayesian method as implemented in Bracken. In parallel, short sequences are assembled in contigs using Spades for a second taxonomic identification round and several metrics are generated to help the analyst to interpret and evaluate findings. Additionally, in metagenomics pipeline we added the recovery of genomes from metagenomes step using MetaWRAP binning modules and GTDB-tk for taxonomic classification of genomes. It is worthy mentioning that other pipelines for microbial data analysis will be implemented in the platform shortly, as for example pipelines for analysis of Microbiome 16S/ITS data, genotyping of SARS-CoV-2, analysis of whole genome of bacteria and others.

Varsmetagen is currently in a beta version available to users in an early-access program, and its release for the general public is planned to happen in December 2022. Varsmetagen is free of charge for academic and non-profit users.

Keywords: virome, clinical metagenomics, clinical microbiome, innovation, pathogen