A priori estimation of sequencing effort in complex microbial metatranscriptomes

Antonio Monleon-Getino[1] and Jorge Frias-Lopez[2]

[1]University of Barcelona
[2]University of Florida

May 6, 2020

Abstract

1. Accurate differential expression of microbial metatranscriptomes based on Next Generation Sequencing depends partly on the depth of the libraries used to perform the analysis. Therefore, estimating the sequencing depth required to sample the metatranscriptome of interest using RNA-seq effectively is an essential first step to both obtain robust results in further analysis and avoiding over-expending once the information contained in the library reaches saturation. 2. Here we present a method to calculate the effort in saturation curves and a priori genes prediction using a simulated series of metatranscriptomic/metagenomic matrices. This method is based on the extrapolation rarefaction curve using a Weibull growth model to estimate the maximum number of genes/OTUs as a function of sequencing depth using a machine learning approach. This approach allows us to compute the effort at different confidence intervals and to obtain an approximate a priori effort using based on an initial fraction of sequences. 3. The accuracy of the results obtained with simulations and real samples (15 datasets of metatranscriptomes from the oral cavity, RNA sequences consist of vectors of 105-1.5x107 reads depth with a 10000 and 600000 genes size) allows one to use an initial shallowly sequenced sample (in this case 20% of the total amount of reads sampled; accuracy R2>0.99 simulated samples and 60-93% for real samples) to estimate the expected sequencing effort needed to cover the whole metatranscriptome/ metagenome from the same sample, so can be used to estimate the estimate the sample size. The algorithm containing the proposed method was saved as a function for R. 4. This proposed method of estimation of the maximum number of gene/OTUs, reads to reach 90, 95 and 99% of maximum number of gene/OTUs, is efficient to help researchers to know if the sampling is sufficient or otherwise need to be increased.

Hosted file

`ver_10_01_2020.pdf` available at https://authorea.com/users/30461/articles/447390-a-priori-estimation-of-sequencing-effort-in-complex-microbial-metatranscriptomes