Do you cov me? Effect of coverage rarefaction on species identification in complex matrices

Marroni F(1)[†] , Radovic S(1), Jurman I(1,2), Cattonaro F(1)

(1) IGA Technology services s.r.l., Udine
(2) Istituto di Genomica Applicata, Udine

[†] Email: marroni@appliedgenomics.org

Motivation
Whole genome shotgun metagenomics is a powerful technique to characterize the microbial composition of several substrates, such as soil, food matrices, and the human body. Compared to target analysis, one of the major advantages of the WGS approach is the possibility of targeting prokaryotes and eukaryotes in a single experiment, without having to perform two or more library preparations, and being free from biases due to different efficiency of amplification in different target species. On the other hand, WGS is more expensive, and requires a higher number of reads. However, several investigations are aiming at the overall characterization of the microbial community of matrices, thus requiring ability of identifying species with frequencies around 1%. We set out to estimate the amount of read coverage needed to efficiently perform such studies. As an example, we show results of rarefaction studies on WGS sequencing of three complex matrices (composed of prokaryotes, fungi, plants, and vertebrates).
Methods
To assess the effect of coverage on the ability of sampling the major components of the complex matrices, the full dataset was used as a gold standard, and several subsets were compared against it. Random samples of 100, 1000, 10000, 100000 and 1000000 reads were extracted from the full data sets and analysis repeated. OTUs identification was performed using Kraken v1.0, with a custom database including the whole nt database of NCBI. This enabled the detection of prokaryotes and eukaryotes, possibly at the expenses of precision of classification of prokaryotes, compared to curated prokaryotes databases. Results in the rarefied samples were compared to those obtained with the full dataset by comparing the number and the relative abundance of OTUs identified in the different data sets. An additional comparison was performed assuming that part of the work required de-novo assembly of the metagenome. Assembly was performed using megahit. Comparison was based on the number and relative abundance of OTUs represented in the assembly of rarefied data-sets compared to the full set and on the fraction of conserved genes observed in the rarefied data-set compared to the full set.
Results
We observed that moderately rarefied samples (i.e. with 10000 reads or more) still have the ability to characterize the major components of complex matrices. Most of the species with frequency of 1% or above were consistently identified and the relative abundance in the rarefied samples was similar to that observed in the full

dataset (see Figure 1). Results based on the de-novo assembly are more sensitive to rarefaction. Only samples with 100000 reads or more showed consistency with the full dataset in terms of proportion of OTUs identified in the assembled scaffolds. Thus, the use of low-coverage shotgun sequencing is not an option if part of the scientific question is the de-novo assembly of the organisms present in the sample. In conclusion, our results suggest that low coverage whole genome shotgun sequencing could be a effective approach for identifying and quantifying species present in complex matrices by aligning sequences against databases.