APPLICATION OF NEXT GENERATION SEQUENCING IN MICROBIAL ECOLOGY

Ph.D. Student: Daniele Pietrucci / *Supervisor*: Prof. Antonella Canini

Cycle: XXXI *A.Y.*: 2017/2018

The advent of Next Generation Sequencing (NGS) has revolutionized the study of microbiology. The main benefits of these technologies concern the high speed sequencing of DNA and RNA, the low manual work and the lower cost when compared to the Sanger sequencing method. However, the large amount of data produced by these technologies requires the use of sophisticated bioinformatic and biostatistical analysis.

The aim of this PhD project is the optimization of bioinformatic pipelines for microbiological applications. In detail my project is focused on: 1) the qualitative and quantitative investigation of microbiota in environmental samples and food; 2) investigation of the genetic variation of *Salmonella enterica*.

During my first year I have focused my attention on a metagenomic project, aimed at describing the microbial taxonomic composition of different foodstuffs and their evolution through space and time. I described the bacterial microbiota of three different Italian cheese (“Caciotta”, “Stracchinato” and “Nobile”), trough 16S rRNA amplicon gene sequencing. In order to characterize the evolution of the microbiota during the ripening process, different samples were analysed: milk, curd, middle and end of fermentation. The data have been preprocessed and analysed trough a bioinformatic workflow. The analysis has provided an insight into the microbial composition of these products, identifying the main bacteria genus for each cheese and describing α and β diversity. I have implemented statistical test in the R software, such as repeated measures anova and nonparametric test, in order to correlate the microbial composition with the physical-chemical process used during the manufacturing procedure.

During the second year I have optimized the bioinformatic pipeline, in order to characterize and asses the best procedure to be used for targeted metagenomics. I have evaluated the pipeline using a virtual dataset, created from a reference-database, and testing different specific metrics, such as F-measure, clustering indices and diversity indices. The pipeline has been validated on a real dataset, consisting in samples extracted for rumen.

During the second year I have also studied the genomic variation of a monophasic variant of *Salmonella enterica* (Kauffman-White classification: 1,4,[5],12:i:-). This serovar is the emerging one, being the most frequently isolated over the last decades and is enlisted among the emerging foodborne pathogens worldwide. I have carried out a comparative analysis of 50 epidemiologically unrelated *S.* 1,4,[5],12:i:- genomes, belonging to different sources. Genomes were assembled and annotated. The analysis permitted the identification of specific clones, having as the most interesting genetic the presence of heavy metals tolerance gene cassettes (resistance to copper and silver) and the distribution of type II toxin-antitoxin families, that likely contribute to their persistence.