In October 2008, the cost of sequencing a bacterial genome of typical length declined to $900, which was 105 times lower than the previous year’s seriously prohibitive $95 000. The reason was the arrival of a new DNA sequencing technology, one of the first among the many that followed and defined the upcoming period of DNA sequencing as the era of the next-generation sequencing (NGS).

The low cost and massive throughput of NGS techniques have revolutionised many fields of genome research. One of these fields is metagenomics, the study of the biodiversity, composition, and functional capacity of environmental microbial communities by genome sequencing and analysis. From a biomedical perspective, the most interesting of these communities is the human microbiome, the collection of microbial species that live on and inside our bodies. Revealed by metagenomics, the vast number of these microbes and their remarkable functional repertoire, as well as their relationship with our health, have put a spotlight on the human microbiome like a newly discovered organ. This thesis is a compilation of computational studies, where the common objective is to improve the analyses of the metagenomic data obtained from our microbiome, in order to make better sense of it.

Briefly, most of the computational work presented here deals with the two most popular questions in metagenomics and microbiome research: “what species are there?” and “what are they doing?”. Regarding the first question, in Chapter 2, we show that the biodiversity and composition estimates in metagenomic studies are improved by optimising the choice of computational methods used via elaborate computer simulations of metagenomic sequencing data. This is followed by Chapter 3, where we present NGS-eval; our tool for getting the most important parameter in our simulations right: the sequencing error rate. NGS-eval can be considered a data quality checker since it determines the sequencing error rate in a certain class of NGS datasets. In addition, it detects novel sequence variants, which are microbial genomic sequences that are absent in databases.

In Chapter 4, we change our focus to the “what are they doing?” question. Here, we look closely at two comparative metatranscriptomic datasets. These are NGS datasets of microbial genes, which were not only present but also active in the oral microbiomes of individuals with either good oral health or oral disease. Based on this data, in the same chapter we describe a novel computational framework called metaModules. This
framework facilitates identifying subnetworks of significantly differentially active, as well as interacting microbial functions that potentially mediate microbiome-related diseases. Finally, in Chapter 5 we present the application of a single-species flavour of metaModules on comparative transcriptomic datasets of either low or high biofilm forming *Candida albicans* isolates. We show that the subnetwork of interacting *C. albicans* genes identified by our method accurately highlights a target for the disruption of molecular mechanisms that are associated with high mortality.