Distributed gene quantification in large metagenomes

Fredrik Boulund[1], **Anders Sjgren**[1], **Erik Kristiansson**[1]
[1] Mathematical Sciences, Chalmers University of Technology, Sweden

Metagenomics is the study of microorganisms by sequencing random DNA fragments from microbial communities. Since no cultivation of individual organisms is required, metagenomics can be used to analyze the large proportion of microorganisms that are hard or impossible to grow in laboratories. Metagenomics therefore holds great promise for understanding complex microbial communities and their interactions with their respective environments. The analysis of metagenomes in the human gut, oral cavities and on our skin can, for example, provide information about what microorganisms are present and their effects on human health.

High-throughput DNA sequencing has significantly increased the size of the metagenomes which today can contain trillions of nucleotides (terabases) from one single study. Current bioinformatics methods are however not designed with these large amounts of data in mind and they are consequently forced to significantly reduce the sensitivity to achieve acceptable computational times. We have therefore developed a new method for distributed gene quantification in metagenomes. Through efficient data dissemination and state-of-the-art sequence alignment algorithms the framework can rapidly and accurately estimate gene abundances in very large metagenomes while still maintaining high sensitivity. The output of our framework is adapted for further statistical analysis, such as comparative metagenomics to identify both quantitative and qualitative differences between metagenomes.

Our method is shown to scale well with the number of available worker nodes and provides a flexible way to optimally utilize the available computer resources. It provides fast and sensitive analysis of terabase size metagenomes and thus enables analysis of studies of microbial communities at a resolution and sensitivity previously not feasible.

Keywords: metagenomics, next generation DNA sequencing, high performance computing