SARMA: a web resource for species assignment of high-throughput sequencing reads from metagenomics analysis

M. D'Antonio[1] ✉, D. Paoletti[2], M. Santamaria[3], T. Castrignanò[2], G. Pesole[1,4]

[1]Dipartimento di Bioscienze, Biotecnologie e Scienze Farmacologiche, Università degli Studi di Bari, Bari, Italy
[2]CASPUR, Consorzio interuniversitario per le Applicazioni di Supercalcolo per Università e Ricerca, Rome, Italy
[3]Istituto di Biomembrane e Bioenergetica, Consiglio Nazionale delle Ricerche, Bari, Italy

Motivations

The exceptional development of next generation sequencing platforms (NGS) has opened unprecedented possibilities for the comprehensive investigation of entire microbial and viruses communities at taxonomic and functional level in environmental and clinical samples. In the latter case, the characterization of human microbiome is a crucial achievement to fully understand the regulation of gene expression in physiological and pathological conditions. The computational approach to viral and microbial discovery is based on the premise that the assayed tissues contain a small viral and microbial component that can be detected and analyzed after the subtraction of the large excess of human sequences. Rough subtractive processes can however lead to problems due to regions in the host genome derived from previous or ancient infections (a typical example is the presence of endogenous retrovirus).

Methods

To address this issue we developed a user-friendly web resource, named SARMA (Species Assignment of Reads from Metagenomic Analysis), using as input one or more metagenomic datasets. The SARMA workflow serially subtracts input reads mapping to human sequences using both ultra fast (i.e. BWA) and sensitive (i.e. BLASTN) read aligners. To avoid false positives in this phase we utilize a custom built human genome obtained filtering out all regions derived from viruses and bacterias. After the subtractive process, any remaining unmapped reads may represent candidate non-human microbial-derived species. Only top alignments fulfilling appropriate thresholds are considered with reads inheriting the taxonomy attributes of the corresponding aligned sequence. A statistical methodology, taking into account the alignment quality, is then adopted for species assignment, allowing also for weighted partitioned assignments. Unassigned reads, possibly deriving from novel organisms, can be then assigned to higher taxonomic ranks using other publicly available tools such as MEGAN. SARMA provides a WEB interface allowing job submission and results browsing.The SARMA output consists of several tabular and tree-based visualizations of the results allowing users to browse their data from high-level summaries down to the more detailed views. Reads shared between host and other organisms are highlighted by crosslinks. Results deriving from two or more datasets can be differentially compared to identify over- and under-represented species.

Results

SARMA has been tested using a public dataset of Clonal Integration of a Polyomavirus in Human Merkel Cell Carcinoma [1]. A sample manually contaminated with a known family of virus should be a perfect training set for a resource like SARMA. The aim of this test were, of course, determining the presence of expected viruses.

References

1. Feng, H, Shuda, M, Chang, Y, and Moore, PS. Clonal Integration of a Polyomavirus in Human Merkel Cell Carcinoma. Science 2008, 319 (5866): 1096-1100. doi: 10.1126/science.1152586