Title: Assembly and protein annotation of viral metagenomes from aquatic, terrestrial and aerial habitats.

Authors: Oscar Ortega Recalde, Elkin Ramón Alvis Narváez

Abstract

Virus are widespread globally and, most likely, are the most abundant microbes in the entire biosphere. In spite of its abundance and ecological relevance, the function of many of their proteins and its significance in specific habitats is widely unknown. Currently, next generation sequencing (NGS) methods enable us to sequence large viral communities (viromes), providing a large amount of data which could be used to resolve this question. Given that every study uses different bioinformatic pipelines for genome assembly and protein annotation is difficult to establish a proper comparison intra and inter habitat. This proposal aims to standardize a pipeline which allow us to compare viromes sampled from three habitats: marine, soil and human gut and which could be subsequently extended. Raw data will be downloaded from public repositories and processed with several bioinfomatic tools for data analysis, assembly and annotation. All of this bioinformatic work will be done in the computer cluster in the Universidad de los Andes and will use several tools currently available through the server such as Galaxy, SPADes and Prodigal. Results will be used to describe specific blueprints for every ecosystem and potentially identify new viral domains and proteins.

Introduction

Recently, it has been identifying that the earth's biosphere is plenty of viruses and they have a variety of important roles on it. Although, include viruses into a new feature is controversial, their activity is not exclusively parasitism but also they have an active participation on the ecosystems modelling, the probational dynamics and different ecological, geochemical and biomedical processes.

Though experimentally, the study of the viruses are still considered very difficult, the progress of molecular biology techniques and particularly next-generation sequencing has enabled broaden and deepen our knowledge about these. Currently, there are several

projects whose objective is to describe the totality of particular habitats virus (viromes). Examples of these projects include attempts to identify the human viroma and the the Tara Oceans expedition, in which one of the objectives was to identify an oceanic viroma. The information obtained from these projects is available in public repositories and it is an important source of data largely untapped for several reasons. First of all, viruses can be considered one of the most diverse groups on earth given the population size and it wide distribution. In addition, it has been found that the vast majority of sequences obtained (63% to 93%) by expedition Tara have yet unknown functions, making them one of the largest global gene reservoirs.

In view of the above, in particular the complex interaction of the viruses with the environment and the lack of large amount of information regarding their genome the research question arises: is it possible to identify, on a metagenomic level, differences between viromes of diverse environments? The overall objective of this project is to characterize viral genomes of particular habitats at a population scale. This characterization seeks to identify new domains and/or specific predicted proteins of each environment and it also pretend to contribute to the future comprehension of the interaction between the viruses with biotic and abiotic factors present on a ecosystem.

Methodology

This study will use several computational tools in order to analyze metagenomes from three different environments: human gut, an aquatic environment and one land or air environment.

Data Collection

The data will be downloaded from public databases such as iVirus, Metavir and NCBI database for metagenomes (SRA). According to the data available three studies for each of the selected environments will be chosen. A review of the literature will be conduct to determine the characteristics of the data if they are not present in the repository.

Assembly

Although some of the raw data are also assembled, the aim of this step is to ensure the homogeneity between the samples to evaluate. The data will be processed with the FASTQC tool to assess the quality of the readings; then it will be processed with Trimmomatic to optimize the quality of the pre-assembly. The assembly algorithm to use will be SPAdes or IDBA-UD according to the preliminary analysis of the data quality obtained for each of these.

References

1. Kristensen, D. M., Mushegian, A. R., Dolja, V. V. & Koonin, E. V. New dimensions of the virus world discovered through metagenomics. *Trends Microbiol.* **18,** 11–19 (2010).
2. Cesar Ignacio-Espinoza, J., Solonenko, S. A. & Sullivan, M. B. The global virome: Not as big as we thought? *Curr. Opin. Virol.* **3,** 566–571 (2013).
3. Koonin, E. V., Dolja, V. V. & Krupovic, M. Origins and evolution of viruses of eukaryotes: The ultimate modularity. *Virology* **479-480,** 2–25 (2015).
4. Nasir, A. & Caetano-Anollés, G. A phylogenomic data-driven exploration of viral origins and evolution. *Sci. Adv.* **1,** e1500527 (2015).
5. Kristensen, D. M. *et al.* Orthologous gene clusters and taxon signature genes for viruses of prokaryotes. *J. Bacteriol.* **195,** 941–950 (2013).
6. Reyes, A., Semenkovich, N. P., Whiteson, K., Rohwer, F. & Gordon, J. I. Going viral: next-generation sequencing applied to phage populations in the human gut. *Nat. Rev. Microbiol.* **10,** 607–17 (2012).
7. Rasmussen, A. L. Probing the viromic frontiers. *MBio* **6,** 1–3 (2015).
8. Lecuit, M. & Eloit, M. The human virome: New tools and concepts. *Trends Microbiol.* **21,** 510–515 (2013).
9. Virgin, H. W. The virome in mammalian physiology and disease. *Cell* **157,** 142–150 (2014).
10. Brum, J. R. *et al.* Patterns and ecological drivers of ocean viral communities. *Science (80-.).* **348,** 1261498–1261498 (2015).
11. Hulo, C. *et al.* ViralZone: a knowledge resource to understand virus diversity. *Nucleic Acids Res.* **39,** D576–82 (2011).
12. Merchant, N. *et al.* The iPlant Collaborative: Cyberinfrastructure for Enabling Data to Discovery for the Life Sciences. *PLoS Biol.* **14,** e1002342 (2016).
13. Roux, S., Tournayre, J., Mahul, A., Debroas, D. & Enault, F. Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. *BMC Bioinformatics* **15,** 76 (2014).
14. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **4,** 311–23 (2015).
15. Oulas, A. *et al.* Metagenomics: tools and insights for analyzing next-generation

sequencing data derived from biodiversity studies. *Bioinform. Biol. Insights* **9,** 75–88 (2015).
16. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–20 (2014).
17. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19,** 455–77 (2012).
18. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28,** 1420–8 (2012).
19. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11,** 119 (2010).
20. Delcher, A. L., Bratke, K. A., Powers, E. C. & Salzberg, S. L. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23,** 673–9 (2007).
21. Keegan, K. P., Glass, E. M. & Meyer, F. MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function. *Methods Mol. Biol.* **1399,** 207–33 (2016).
22. Reyes, A. *et al.* Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* **466,** 334–338 (2010).
23. Martínez Martínez, J., Swan, B. K. & Wilson, W. H. Marine viruses, a genetic reservoir revealed by targeted viromics. *ISME J.* **8,** 1079–88 (2014).
24. Aguirre de Cárcer, D., López-Bueno, A., Pearce, D. A. & Alcamí, A. Biodiversity and distribution of polar freshwater DNA viruses. *Sci. Adv.* **1,** e1400127 (2015).
25. Reavy, B. *et al.* Distinct circular single-stranded DNA viruses exist in different soil types. *Appl. Environ. Microbiol.* **81,** 3934–45 (2015).