The bioinformatics of viral metagenomics

Martin Norling[1], Oskar Karlsson[1,2,3], Erik Bongcam-Rudloff[1] ✉
[1]SLU Global Bioinformatics Centre, Department of Animal Breeding and Genetics (HGEN), Swedish University of Agricultural Sciences (SLU), Uppsala, Sweden
[2]Department of Biomedical Sciences and Veterinary Public Health (BVF), Swedish University of Agricultural Sciences (SLU), Uppsala, Sweden
[3]The Joint Research and Development Division of SLU and SVA, OIE Collaborating Centre for the Biotechnology-based Diagnosis of Infectious Diseases in Veterinary Medicine (OIE CC), Uppsala, Sweden

Motivation and Objectives

Metagenomic methods provide the veterinary and public health sciences with the promise of new and improved diagnostic tools with unprecedented ability to detect a plethora of known and unknown viromes in clinical samples. Successful metagenomics is based on three main activities where each one must be consistent and reliable for the method to be useful, these are (1) wet-lab preparation, (2) sequencing, and (3) bioinformatics analysis of the results.

We are a collaborative group from the OIE Collaborating Centre for the Biotechnology-based Diagnosis of Infectious Diseases in Veterinary Medicine, Uppsala, Sweden and the SLU Global Bioinformatics Centre, Uppsala, Sweden who are working with the development and evaluation of platforms and methods for viral metagenomics. Together with the National Veterinary Institute (SVA), we develop and test methods for extraction of viromes, feasibility of sequencing platforms to deliver metagenomic data sets within constraints of money and time as well as evaluate bioinformatics tools to do the final analysis. We also combine the tools that evaluate well into software packages for separation, classification, assembly and visualization of genomic data in metagenomic samples.

The aim of the work is to provide insight into the feasibility of using the metagenomics approach for detection of emerging viruses, monitoring wildlife for known pathogens as well as providing a tool for rapid characterization of viral pathogens in outbreak situations from a veterinary standpoint.

Methods

The bioinformatical challenge separates itself from the preparation and sequencing steps in that methodologies in bioinformatics evolve comparatively fast. A stable wet-lab and sequencing platform will still require constant updates of bioinformatical databases and tools, requiring that any system is build in a modular way where each part can easily be exchanged for an updated variety.

The current bioinformatical metagenomics pipeline is implemented with two separate front-ends. One is a classic command-line interface suitable for server automation and implementation into further pipelines and the other is a combined HTML5 and jQuery interface intended to give the power of the command-line interface to everyday users. This two-fold approach is to allow as many users as possible to use the same tools, making results easier to reproduce in different settings.

The back-end pipeline is based on a simple plug-and-play configuration where any program or script can easily be replaced to customize or update the system. The current default configuration is based on FastQC (unpublished) for quality control, Prinseq-Lite (Schmieder and Edwards 2011), MetaVelvet (Namiki, *et al.,* 2012) and BLAST (Altschul, *et al.,* 1990), but tests are being run with several other programs as well. The system can be configured to use scheduling systems in the back-end to distribute load and processing. The default scheduler is SLURM, but the system could easily be configured to use a different system. This modular approach is evident throughout the entire project – allowing the system to be used in a wide range of situations.

Results and Discussion

The system is still in BETA, and every part of the pipeline and interface is still being tested and evaluated, but a few common themes are sure to live throughout the project.

First of all, the web version of the system uses the high level of user interaction allowed by HTML5 and jQuery. The system allows for upload of large data files by fragmenting and resuming, allowing arbitrary size data sets to be uploaded

without changing server settings, as well as allowing a broken download to resume from where it was cut-off. The web system is built around responsive, intuitive interfaces giving feedback and process information in real time, sane defaults giving new users an easy start with quick guides for common tasks and clear documentation for server implementations. The entire system will also be released as open source to contribute to the public as much as possible.

Acknowledgements

This work is funded by SIDA grant SWE-2011-154: "Using next generation sequencing to support the development of infection and treatment vaccination methods in East Africa", Sweden. The bioinformatics work at SLU Global Bioinformatics Centre was also supported by grants from SLU and BILS.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search tool. *Journal of molecular biology*, **215**(3), 403-410.

Namiki T, Hachiya H, Tanaka, Sakakibara Y (2012) MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Research* **40**(20), e155. doi:10.1093/nar/gks678

Schmieder R, Edwards R (2011) Quality cotrol and preprocessing of metagenomic datasets. *Bioinformatics*, **27**(6), 863-864. doi:10.1093/bioinformatics/btr026