Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Research paper

Virus detection in high-throughput sequencing data without a reference genome of the host

Jochen Kruppa[a,*], Wendy K. Jo[b], Erhard van der Vries[c], Martin Ludlow[b], Albert Osterhaus[b], Wolfgang Baumgaertner[d], Klaus Jung[a]

[a] *Institute for Animal Breeding and Genetics, University of Veterinary Medicine Hannover, Foundation, Bünteweg 17p, Hannover 30559, Germany*
[b] *Research Center for Emerging Infections and Zoonoses, University of Veterinary Medicine Hannover, Foundation, Bünteweg 17, Hannover 30559, Germany*
[c] *Department of Immunity and Infection, University of Utrecht, Yalelaan 1, Utrecht, CN 3584, the Netherlands*
[d] *Department of Pathology, University of Veterinary Medicine Hannover, Foundation, Bünteweg 17, Hannover 30559, Germany*

ARTICLE INFO

Keywords:
Decoy database
Metagenomics
Read mapping
Reference genome
Virus discovery

ABSTRACT

Discovery of novel viruses in host samples is a multidisciplinary process which relies increasingly on next-generation sequencing (NGS) followed by computational analysis. A crucial step in this analysis is to separate host sequence reads from the sequence reads of the virus to be discovered. This becomes especially difficult if no reference genome of the host is available. Furthermore, if the total number of viral reads in a sample is low, de novo assembly of a virus which is a requirement for most existing pipelines is hard to realize.

We present a new modular, computational pipeline for discovery of novel viruses in host samples. While existing pipelines rely on the availability of the hosts reference genome for filtering sequence reads, our new pipeline can also cope with cases for which no reference genome is available. As a further novelty of our method a decoy module is used to assess false classification rates in the discovery process. Additionally, viruses with a low read coverage can be identified and visually reviewed. We validate our pipeline on simulated data as well as two experimental samples with known virus content. For the experimental samples, we were able to reproduce the laboratory findings.

Our newly developed pipeline is applicable for virus detection in a wide range of host species. The three modules we present can either be incorporated individually in other pipelines or be used as a stand-alone pipeline. We are the first to present a decoy approach within a virus detection pipeline that can be used to assess error rates so that the quality of the final result can be judged. We provide an implementation of our modules via Github. However, the principle of the modules can easily be re-implemented by other researchers.

1. Introduction

Samples from humans and animals suspected of a virus infection on clinical grounds, are usually analyzed by classical and modern molecular virological assays, when applicable supported by histo-pathology data. Meanwhile, the advent of next-generation sequencing (NGS) has provided us with the opportunity of reading all sequence information in a biological sample, therefore becoming an important tool for virus discovery which will undoubtedly find its way into routine virus diagnostic practice. However, virus detectionusing NGS data is by no means a straightforward task, but should involve close communication between the clinician, the virologist, the pathologist and the bioinformatician (Smits et al., 2015; Smits and Osterhaus, 2013).

The overall problem of virus identification in NGS data from a host sample is to identify all sequences that don't originate from the host itself. While most sequencing reads will usually belong to the host or other non-relevant microorganisms only a small proportion of reads will belong to the virus to be discovered. The assignment of sequencing reads to the host and other non-relevant organisms and viruses relies on reference genomes available in databases. Here, we present a new bioinformatics pipeline for virus metagenomics that is also applicable if no reference genome data from the host is available.

Currently available bioinformatics pipelines or software solutions for virus sequence detection in NGS data rely on approaches that can be divided into two categories. Category I involve approaches, that first remove all host reads from the sample and map or align the remaining reads to a viral database. In this case, the host's reference genome sequence must be available in a sufficient quality to make sure that all

[*] Corresponding author.
E-mail address: jochen.kruppa@charite.de (J. Kruppa).

https://doi.org/10.1016/j.meegid.2018.09.026
Received 2 August 2018; Received in revised form 25 September 2018; Accepted 27 September 2018
Available online 03 October 2018

host reads are removed and only non-host sequences are among the unmapped reads. Such approaches work for example well with samples from humans or mice and other species for which reference genome data is available at the Ensemble database (Ensemble Database, n.d.). Approaches from the category II first assemble the raw sequencing reads (or only the unmapped reads) to larger contigs, which are further used in the analysis pipeline. Larger contigs allow for a higher mapping accuracy than short reads, and including an assembly step into a detection pipeline is therefore advantageous. Nevertheless, to achieve large contigs - that are longer than the single reads - the coverage of the single viral strains of the sequencing reads must be high. If there are not enough viral reads of a single strain in the sample, the gaps between the reads are too large and contigs cannot be built preventing virus identification. A common element of the approaches in both categories is that reads or contigs are aligned to a given virus sequence database, and a sorted list of detected viruses (or at least taxonomic groups) is returned. The approach we present here belongs to category I, i.e. raw reads instead of contigs are mapped against reference genomes. In the following, we provide a brief summary of other existing pipelines and their usability in the case of samples generated with low sequencing coverage and a non-availability of a host reference.

Among category II pipelines, Iterative Virus Assembler (IVA) (Hunt et al., 2015) uses its own de novo assembler to generate contigs from the raw or host-free reads. Generated contigs can afterwards be mapped to a virus database using SMALT (SMALT, n.d.) and Kraken (Wood and Salzberg, 2014) to determine the viral strain. IVA reports only the virus strain that appears most frequently with quality information, whereas the report produced by Kraken gives the user more information on the identified taxa. Thus, the limitaion of IVA consists on de novo assembly of contigs, which is not possible when the overall sequence coverage is low and the generation of only a single viral strain. Another pipeline, called RIEMS (Scheuch et al., 2015), also first assembles the raw reads to contigs, which are afterwards mapped to a virus database using the NCBI BLAST software suite (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+). Assigned reads are then classified taxonomicaly. As an additional feature, the RIEMS pipeline can also translate the assembled sequences to amino acid sequences and use these sequences for further detection on the protein level.

Among category I pipelines (i.e., direct mapping of sequencing reads), the approach by Petty et al. (Petty et al., 2014) describes the standard procedure in a human pilot study. The raw reads are first mapped to the human reference genome, and the unmapped reads are then mapped to a virus database. The removal of the host reads is a crucial step and has been implemented in VirusFinder (Wang et al., 2013), VirusHunter (Zhao et al., 2013), VirusSeq (Chen et al., 2012), and Vy-PER (Forster et al., 2015). Mostly, these pipelines have been demonstrated on example of human samples. In general, these pipelines demand for a known host reference genome of sufficient quality to remove all non-viral reads from the downstream analysis. First, host reads are removed to obtain data cleaned from host sequences. If host reads are kept in the data, false positive mapping to virus reference sequences may occur. Second, further noise is removed from the data to improve the mapping accuracy. By removing the host reads it is assumed that only viral reads remain. Nielsen et al. (Nielsen et al., 2014) describe a reference free identification and assembly without the implementation as a software solution or ready to use pipeline. In the following the mentioned pipelines of category I are described in more detail.

VirusFinder (Wang et al., 2013) performs first a preprocessing step, in which the raw sequencing reads are mapped to the human reference genome using Bowtie2 (Langmead and Salzberg, 2012; Langmead et al., 2009). Then, unmapped reads are extracted and aligned to a viral database using BLAT (Kent, 2002). Finally, the reads are assembled using Trinity (Grabherr et al., 2011). VirusFinder assumes a high sequencing coverage so that good assembly results can be obtained, and is mainly developed to detect virus integrations sites in the human genome. The examples presented in the VirusFinder article have a sequencing coverage ranging from 31.7× to 121.2×. The assembled contigs are then used for the generation of phylogenetic trees and the estimation of relationship to each other. VirusHunter (Zhao et al., 2013) uses BLASTn to filter first the reads belonging to the host after some quality assessment. The host-free reads are then classified using BLASTn and BLASTx into taxonomic groups. Therefore, VirusHunter needs a good host reference genome to filter the reads into host-free and host reads. In addition, the repeated BLAST runs to process the reads are also time consumable. VirusSeq (Chen et al., 2012) focuses on the identification of viral strains in human cancer tissue. First all human sequence reads are removed by mapping to the human reference genome. The remaining human-free reads are then aligned to a viral reference database using the MOSAIK aligner in both steps (Lee et al., 2014). VirusSeq uses the overall count number of matched reads to identify the viral strain. Nevertheless, the threshold is set to 1.000 reads per virus regarding an overall 30× coverage of the whole-genome sequencing data. This threshold can be modified, but VirusSeq is developed for sample with a high read coverage. Therefore, it cannot be used to analyze low coverage datasets. Vy-PER (Forster et al., 2015) uses in the first step the human reference genome to remove all host sequence reads. Reads, which are not mapped to the human reference are then filtered and aligned to the NCBI viral genome database using BLAT (Kent, 2002). The described example on leukemia samples is done with a very high coverage (80× cases and 40× controls), which is not a requirement, but precondition for the elimination of false positives.

All mentioned pipelines of category II make use of an alignment or a mapping software such as Blast or Bowtie2. A broader and more comprehensive overview on available mappers is given by Fonseca et al. (Fonseca et al., 2012), in which the authors discuss the mapper characteristics and the problems of comparing different mappers. In the case of virus detection some specific issues play a role: 1) high heterogeneity of the genomes, 2) mutation rates, 3) insertions of whole genomic areas, and 4) infection of new hosts with adaptions of the viral genome. Hence, problems occur when dealing with samples from a high variety of potential virus infected species. First, a fully assembled reference genome is only available for a small number of animals. Furthermore, the quality of the reference genomes can differ as only the human and mouse genome are of sufficient quality, whereas there is no good reference genome available for many animals. Second, the number of viral sequence reads in a biological sample depends on the production circle of the virus, the time point of infection, and the selection of the correct tissue type to get most of the virus out of the sample. Therefore, the number of possible detectable viral sequences might be low. For building contigs by an assembly process many viral sequence reads must be available, i.e. a good coverage of viral reads must be given, and these reads should not be contaminated with sequences from the host organism. Both issues are not applicable to studies which don't focus on human or mice samples with a small area of possible viral infection.

In this study, three bioinformatics modules for virus detection and two example data samples are described. Module I allows to evaluate the false positive findings by a decoy database approach, module II shows the host-free mapping of DNA sequencing reads to an artificial viral reference genome, and module III describes the mapping of the translated DNA sequence reads to a artificial amino acid viral reference genome. In the results section we demonstrate the results of a simulation study using the decoy database (comparing different mapping softwares within our pipeline) and present the analysis of two example data sets. We close this article with some conclusions. Moreover, we describe the combination of all three modules into a virus detection pipeline in the supplementary material.

2. Methods

In this section, we describe in detail modules that form our new virus detection pipeline. We chose a modular composition of our