The VirAnnot Pipeline: A Resource for Automated Viral Diversity Estimation and Operational Taxonomy Units Assignment for Virome Sequencing Data

Marie Lefebvre, Sébastien Theil, Yuxin Ma, and Thierry Candresse[†]

UMR 1332 BFP, INRA, University of Bordeaux, CS20032, 33882 Villenave d'Ornon cedex, France

Abstract

Viral metagenomics relies on high-throughput sequencing and on bioinformatic analyses to access the genetic content and diversity of entire viral communities. No universally accepted strategy or tool currently exists to define operational taxonomy units (OTUs) and evaluate viral alpha or beta diversity from virome data. Here we present a new bioinformatic resource, the VirAnnot (automated viral diversity estimation) pipeline, which performs the automated identification of OTUs. Reverse-position-specific BLAST (RPS-Blastn) is used to detect conserved viral protein motifs. The corresponding contigs are then aligned and a clustering approach is used to group in the same OTU contigs sharing more than a set identity threshold. A 10% threshold has been validated as producing OTUs that reasonably approach, in many families, the International Committee for the Taxonomy of Viruses taxonomy and can therefore be used as a proxy to viral species.

Metagenomic approaches rely on high-throughput sequencing (HTS) and on bioinformatic analyses of HTS sequence data to access the genetic content and diversity of entire communities in an unbiased way. The use of metagenomic data has a wide variety of applications for plant pathogen studies, including the identification of potential pathogens for better food security (MacDiarmid et al. 2013) or further our knowledge on the effects of microbiomes and viromes on plants (Bulgarelli et al. 2013; Roossinck 2015).

Whereas metagenomics face identification, storage, and computational challenges, viral metagenomics is confronted by specific taxonomic assignment difficulties. Unlike cellular organisms such as fungi or bacteria for which universally conserved genes (internal transcribed spacers and 16S ribosomal RNA) can be used to define operational taxonomic units (OTUs) through a clustering approach (Caporaso et al. 2010), no such universally shared pattern exists for viruses. A consequence is that no universally accepted strategy or tool currently exists to define OTUs and evaluate viral alpha or beta diversity from virome data (Nooij et al. 2018; Simmonds 2015).

We have developed an automated routine addressing this OTU definition problem and integrated it in VirAnnot, a bioinformatic virome sequence analysis pipeline that already performs the assembly of reads, the identification of viral contigs, and their Blast-based taxonomic assignment, steps considered standard for the analysis of HTS data (Nooij et al. 2018). These steps are then followed by a new clustering strategy that allows for grouping together in the same OTU contigs that share more than a set identity threshold.

Sequence datasets. In order to analyze the reproducibility of the VirAnnot OTU clustering output and the closeness of the identified OTUs to taxonomic species recognized in the current

Funding

Financial support for the early development phase was supported by ANR NET-BIOME 2010 SafePGR project. Y. Ma was supported by a Chinese Science Council (CSC) Scholarship Ph.D. grant.

Keywords

bioinformatics, metagenomics, operational taxonomy units, viral diversity, virology, virome

Current address for S. Theil: INRA UMRP, 20, côte de Reyne, 15000 Aurillac, France.

First and second authors contributed equally to this work.

[†]Corresponding author: T. Candresse; Thierry.candresse@inra.fr

The author(s) declare no conflict of interest.

Accepted for publication 25 August 2019.

International Committee for the Taxonomy of Viruses (ICTV), several datasets generated from complex pools of plants (ca. 40 species and 200 individual plants, Ma et al., *unpublished data*) were used. These datasets have been sequenced on an Illumina HiSeq 3000 system at the GeT-PlaGe platform (INRA, Toulouse, France) and deposited in the INRA National Data Portal under the identifier <https://doi.org/10.15454/TVWBCQ>.

VirAnnot pipeline workflow. The VirAnnot pipeline integrates standard bioinformatic tools in three main steps: (i) reads cleaning, (ii) contigs assembly, and (iii) taxonomic classification. Briefly, the first step consists of raw reads quality trimming with a minimum score of 20 and a minimum read length of 70. The reads are then demultiplexed followed by adapters, polyA, and if necessary, multiplex identifier (MID) tag removal using cutadapt (Martin 2011). In order to limit intersample cross talk associated with index-hopping (Illumina 2017; van der Valk et al. 2019), a subroutine can be implemented to retain only reads having identical MID tags on both pairs. The second step performs the assembly of the reads from all selected libraries into contigs using either IDBA-UD (Peng et al. 2012) or SPAdes (Bankevich et al. 2012). There is however no specific step implemented to identify terminal redundancies to identify and assemble circular viral genomes. Through the third step, the contigs are annotated using BlastN or BlastX (Altschul et al. 1990) against the NCBI GenBank nr or nt sequence databases with a user defined significance threshold (default e-value of 10^{-4}). In addition to similarity searches against protein databases, a search against the PFAM database (Punta et al. 2012) is carried out in all six reading frames using RPS-Blast (reverse-position-specific BLAST) (Marchler-Bauer et al. 2002) with again a user defined threshold (default e-value of 10^{-4}).

Clustering approach of OTU identification. After the RPS-Blast and BlastX searches, the VirAnnot pipeline identifies OTUs based on a clustering approach performed on sequences encoding conserved viral protein domains (Itzhaki 2011; Koo et al. 2009). All contigs encoding a given virus-specific conserved protein motif, identified by the RPS-Blast annotation, are aligned with reference sequences using the ETE3 Toolkit, taking into account the motif-coding information so that all sequences are aligned in the correct orientation (Huerta-Cepas et al. 2016). Centering each alignment on a motif-encoding region guarantees that comparable regions are aligned, irrespective of the overall size of the assembled contigs. For each viral motif, a distance matrix is then computed using pairwise distances between the locally aligned sequences corresponding to the region encoding the conserved motif and 50 amino acids upstream and downstream of it. User-defined variables at this stage include the minimum contigs overlap (default of 20 amino acids) and a distance threshold between OTUs. We routinely use a 10% divergence value but this can be adjusted at will. The matrix is used to generate two identical phylogenetic trees: one using the ETE3 Toolkit for visualization purposes and one using the hclust clustering method (Müllner 2013). This second tree is cut according to the distance threshold, determining the OTUs and the contigs integrated into each OTU.

Validation of the OTU definition threshold in relation to ICTV taxonomic species. Various datasets (see above) were analyzed using a 10% clustering cut-off value for OTUs defined on the basis of the RNA-dependent RNA polymerase (RdRp) conserved motifs. For several single-stranded RNA (ssRNA) or double-stranded RNA (dsRNA) viral families, the average pairwise distance between OTUs (amino acid divergence in a 100 amino acid small region extending on both sides of the GDD conserved triplet) was then compared with that between viral species recognized by the ICTV using the MEGA software (Kumar et al. 2016).

Pipeline availability. VirAnnot is freely available at <https://github.com/marieBvr/virAnnot>. All documentation about the implementation, installation, and use of the pipeline is available at <https://virannot-docs.readthedocs.io/en/latest/>.

Repeatability. To validate the clustering routine implemented in VirAnnot, its repeatability was evaluated by running the whole pipeline analysis five times on two datasets. A comparison of the OTUs defined on the basis of the well-known RdRp signature sequences of RNA viruses (RdRp1, RdRp2, RdRp3, and RdRp4) in these five independent analyses was then performed. For RdRp1, $26 \leq \text{OTUs} \leq 27$ were obtained; for RdRp2, $45 \leq \text{OTUs} \leq 46$; for RdRp3, 40 OTUs; and for RdRp4, $111 \leq \text{OTUs} \leq 117$. Comparison of the OTUs identified between the different clustering repetitions showed that the same OTUs were repeatedly defined (not shown). These comparisons therefore show a good stability of the VirAnnot output and the limited variations observed are likely due to the known variability of the assembly and clustering processes.

ICTV group/threshold validation. Given the variability of taxonomic criteria used to define species in different viral families (Simmonds 2015), the use of a unique simple rule cannot allow to define OTUs closely mimicking species under all circumstances. Similarly, since the extent of conservation varies between viral conserved motifs, the OTUs cut-off distance parameter may

Table 1. Statistics on amino acid pairwise distances for operational taxonomy units (OTUs) defined using the RNA-dependent RNA polymerase conserved motif and for formal species recognized by the International Committee for the Taxonomy of Viruses (ICTV)

Viral family	Category	Total number of comparisons	Mean	Standard deviation	Min	Median	Max
Single-stranded RNA	<i>Potyviridae</i> OTU	36	0.542	0.076	0.377	0.538	0.635
	<i>Potyviridae</i> ICTV	16,471	0.531	0.104	0.108	0.512	0.823
Double-stranded RNA	<i>Totiviridae</i> OTU	1,225	0.764	0.122	0.25	0.788	0.942
	<i>Totiviridae</i> ICTV	378	0.646	0.13	0.273	0.651	0.969
	<i>Partitiviridae</i> OTU	169	0.727	0.191	0.237	0.808	0.942
	<i>Partitiviridae</i> ICTV	990	0.661	0.158	0.057 ^a	0.723	0.861

^a A few pairwise distance values were less than 10%.

need to be adjusted depending on the motif used. VirAnnot takes this aspect into consideration and the OTU divergence threshold is a user definable parameter. We have however found that a 10% cut-off value defines in many viral families RdRp OTUs that appear to be a reasonable proxy to viral species. We compared the amino acid pairwise distances in the short conserved region around the RdRp motifs between VirAnnot OTUs and between valid ICTV species (Table 1). For comparisons, in the ssRNA virus family *Potyviridae*, the average distance between OTUs closely matched that between ICTV species. For the dsRNA families *Totiviridae* and *Partitiviridae*, the average distance between OTUs was slightly higher than that between ICTV species but with a large overlap in values (Table 1), supporting the meaningfulness of the 10% threshold.

Perspectives. This Resource Announcement describes a new HTS virome analysis pipeline, VirAnnot, which allows the automated evaluation of viral OTU richness (alpha diversity) in metavirome data or the comparison of diversity between viromes (beta diversity). In addition, because of the underlying phylogenetic approach, VirAnnot is also compatible with the UniFrac strategy of comparison of microbial communities (Chen et al. 2012; Lozupone and Knight 2005). Given the constraints imposed by the strategy used (existence of a contig encoding a conserved protein motif, minimum overlap between contigs for a given motif...), the VirAnnot OTU output represents a lower bound estimate of the total virome complexity. On the other hand, the conserved protein motifs that are at the core of the strategy implemented in VirAnnot are the most conserved elements in viral genomes. They are therefore among the most efficient tools we have to identify viral sequences. Given that the search performed by VirAnnot can target all known viral conserved motifs, only viruses that do not share any such motif would not be identified. The stability and repeatability of the VirAnnot OTU assignment and OTU richness estimation have been confirmed, therefore allowing an easy and direct comparison of viral richness between samples. Although VirAnnot is not able to reconstruct circular genomes, it performs well with viruses with circular genomes, including pararetroviruses (*Caulimoviridae*, *Badnaviridae*) and viruses with ssDNA genomes such as *Geminiviridae*, *Genomoviridae*, and *Nanoviridae* and OTUs can be efficiently defined for these agents provided they are based on alignments of regions encoding protein motifs conserved in the corresponding viral families. There are for example 11 conserved Pfam motifs typical of the family *Geminiviridae*, all of which could be used to define *Geminiviridae* OTUs. The setting at 10% user-defined divergence threshold between OTUs provides results that approximate, in different families, the ICTV species level, allowing to use such OTUs as a proxy to taxonomic species. The use of higher or lower threshold values may allow the definition of OTUs representing different taxonomic levels.

Acknowledgments

We thank Sandy Contreras for her contribution to the early phases of the VirAnnot pipeline development.

Literature Cited

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403-410.
- Bankovich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Pribelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., and Pevzner, P. A. 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19:455-477.
- Bulgarelli, D., Schlaeppi, K., Spaepen, S., van Themaat, E. V. L., and Schulze-Lefert, P. 2013. Structure and functions of the bacterial microbiota of plants. *Annu. Rev. Plant Biol.* 64:807-838.
- Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., Reeder, J., Sevinsky, J. R., Turnbaugh, P. J., Walters,

- W. A., Widmann, J., Yatsunenko, T., Zaneveld, J., and Knight, R. 2010. QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7: 335-336.
- Chen, J., Bittinger, K., Charlson, E. S., Hoffmann, C., Lewis, J., Wu, G. D., Collman, R. G., Bushman, F. D., and Li, H. 2012. Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics* 28: 2106-2113.
- Huerta-Cepas, J., Serra, F., and Bork, P. 2016. ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* 33:1635-1638.
- Illumina. 2017. Effects of index misassignment on multiplexing and downstream analysis. <https://www.illumina.com/content/dam/illumina-marketing/documents/products/whitepapers/index-hopping-white-paper-770-2017-004.pdf>
- Itzhaki, Z. 2011. Domain-domain interactions underlying Herpesvirus-human protein-protein interaction networks. *PLoS One* 6:e21724.
- Koo, Q. Y., Khan, A. M., Jung, K. O., Ramdas, S., Miotto, O., Tan, T. W., Brusic, W., Salmon, J., and August, J. T. 2009. Conservation and variability of West Nile virus proteins. *PLoS One* 4:e5352.
- Kumar, S., Stecher, G., and Tamura, K. 2016. MEGA7: Molecular evolutionary genetics analysis Version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33:1870-1874.
- Lozupone, C., and Knight, R. 2005. UniFrac: A new phylogenetic method for comparing microbial communities. *Appl. Env. Microbiol.* 71:8228-8235.
- MacDiarmid, R., Rodoni, B., Melcher, U., Ochoa-Corona, F., and Roossinck, M. 2013. Biosecurity implications of new technology and discovery in plant virus research. *PLoS Pathog.* 9:e1003337.
- Marchler-Bauer, A., Panchenko, A. R., Shoemaker, B. A., Thiessen, P. A., Geer, L. Y., and Bryant, S. H. 2002. cdd: A database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res.* 30:281-283.
- Martin, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 17:10-12.
- Müllner, D. 2013. Fastcluster: Fast hierarchical, agglomerative clustering routines for R and python. *J. Stat. Softw.* 53:1-18.
- Nooij, S., Schmitz, D., Vennema, H., Kroneman, A., and Koopmans, M. P. G. 2018. Overview of virus metagenomic classification methods and their biological applications. *Front. Microbiol.* 9:749.
- Peng, Y., Leung, H. C. M., Yiu, S. M., and Chin, F. Y. L. 2012. IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28:1420-1428.
- Punta, M., Coghill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E. L., Eddy, S. R., Bateman, A., and Finn, R. D. 2012. The Pfam protein families database. *Nucleic Acids Res.* 40:D290-D301.
- Roossinck, M. J. 2015. Plants, viruses and the environment: Ecology and mutualism. *Virology* 479-480:271-277.
- Simmonds, P. 2015. Methods for virus classification and the challenge of incorporating metagenomic sequence data. *J. Gen. Virol.* 96:1193-1206.
- van der Valk, T., Vezzi, F., Ormestad, M., Dalén, L., and Guschanski, K. 2019. Index hopping on the Illumina HiSeqX platform and its consequences for ancient DNA studies. *Mol. Ecol. Resour.* doi.org/10.1111/1755-0998.13009