Computational Viromics: Applications of the Computational Biology in Viromics Studies

Congyu Lu¹ · Yousong Peng¹

Received: 2 June 2020 / Accepted: 14 April 2021 / Published online: 31 May 2021
© Wuhan Institute of Virology, CAS 2021

Viruses are a kind of biological entities which rely on host cells for survival. Depending on the genetic materials and replication mode, they can be grouped into double-stranded DNA (dsDNA), single-stranded DNA (ssDNA), double-stranded RNA (dsRNA), positive-sense single-stranded RNA (+ssRNA), negative-sense single-stranded RNA (−ssRNA), ssRNA reverse transcriptase viruses (ssRNA-RT) and dsDNA reverse transcriptase viruses (dsDNA-RT) (Walker *et al.* 2020). Viruses can infect most kinds of biological entities, including viruses, bacteria, archaea and eukaryote (La Scola *et al.* 2008; Fermin, 2018). They have a great impact on the earth by shaping bacterial population dynamics and balancing the global ecosystem (Suttle, 2007). For humans, viruses, on the one hand, can cause high human morbidity and mortality and serious economic loss (Baud *et al.* 2020), on the other hand, they can promote and maintain the healthy balance of the gut microbiome (Seo and Kweon, 2019). Besides, some phages can be applied as the therapy of bacterial infections, especially for the bacterial strains resistant to multiple antibiotics (Altamirano and Barr, 2019).

The viromics studies based on the high-throughput sequencing technology have become increasingly popular in recent years, and novel viruses are being discovered at an unprecedented pace (Gregory *et al.* 2019). For example, the Tara Oceans Project recently identified 195,728 viral populations which were more than 10 times as many as the known global ocean DNA virome (Gregory *et al.* 2019). However, several challenges exist in analyzing the sequencing data from viromics studies. Firstly, it is difficult to identify all viral nucleotide sequences from the nucleotide sequences that mixed with the sequences of other species and the possible pollutions (Roux *et al.*

2015a; Ren *et al.* 2017; Fang *et al.* 2019; Kieft *et al.* 2020); secondly, the annotation of viral nucleotide sequences is still challenging, especially for those with remote or no homology with the known viruses (Roux *et al.* 2015b; McNair *et al.* 2019; Zhang *et al.* 2019a); thirdly, the taxonomic assignment of novel viruses is difficult due to a lack of a unified classification system for viruses (Low *et al.* 2019); fourthly, rapid functional characterization of a large number of newly discovered viruses such as identifying the viral hosts is extremely difficult to achieve by using traditional experimental methods (Jofre and Muniesa, 2020). According to the above analysis, an emerging area of computational viromics which is defined as using the computational methods to solve the problems in viromics studies was proposed in the present study. It includes but not limited to the following aspects:

Identification of Viral Genomic Sequences

The first step of viromics studies is to identify the viral nucleotide sequences from the metagenomic sequencing data which often contain lots of DNA sequences of host cells such as bacteria and human (Edwards and Rohwer, 2005). Both the eukaryotic and prokaryotic viruses are usually identified using approaches based on the homology of marker genes or the genomic sequences with known viral nucleotide sequences, such as ViromeScan (Rampelli *et al.* 2016). Unfortunately, viruses lack the universal marker genes like the 16s ribosomal RNA (rRNA) in other species. Therefore, the marker genes used for viral sequence identification should be carefully selected to ensure the coverage of viruses. For example, the RNA-dependent RNA polymerases (RdRp) can be taken as the marker gene for the RNA viruses (Wolf *et al.* 2018). Nevertheless, the homology-based methods show a significant limitation when they are used in identifying the viruses with large diversification from the known viruses (Gregory *et al.* 2019). To overcome the limitation, the

✉ Yousong Peng
pys2013@hnu.edu.cn

¹ Bioinformatics Center, College of Biology, Hunan Provincial Key Laboratory of Medical Virology, Hunan University, Changsha 410082, China

sequence homology independent methods have been developed (Ren *et al.* 2017; Fang *et al.* 2019). For example, the VirFinder identified viral sequences based on the k -mer frequencies in the nucleotide sequences (Ren *et al.* 2017). In addition, the methods such as VIBRANT combining homology-based methods and homology-independent methods were also further developed to improve the viral sequence identification (Kieft *et al.* 2020).

Annotation of Viral Genomes

The annotation of viral genomes is important for the further characterization of viruses, and the identification of the genes in viral genomes is most crucial to genome annotation. Currently, there are a large number of methods that can be used for gene prediction (Hyatt *et al.* 2010; McNair *et al.* 2019; Zhang *et al.* 2019a). Although most of them are not designed for viruses, they should be suitable for the gene predictions of viruses since the viruses can use the machinery of their host cells for transcription and translation (Stern-Ginossar *et al.* 2019). The characteristics of viral genes including the compact gene structure, no or few introns, overlapped or co-transcribed genes, can be incorporated to optimize the gene prediction methods, such as the PHANOTATE and Vgas (McNair *et al.* 2019; Zhang *et al.* 2019a). Besides, the combination of several different tools may improve gene predictions (Zhang *et al.* 2019a). Moreover, the emerging metaproteomics which is defined as large-scale identification and quantification of proteins from microbial communities may help the identification of genes in viromics studies to a great extent (Brum *et al.* 2016).

Taxonomic Assignment of the Virome

Most of the newly-discovered viruses lack biological features and cannot be classified by the current classification system proposed by the International Committee on Taxonomy of Viruses (ICTV) (Walker *et al.* 2020). The usage of viral sequences in virus taxonomy is valid and is supported by the ICTV (Simmonds *et al.* 2017). Currently, the homology-based methods have been used for taxonomic assignment of the virome. However, they can only be suitable for a small proportion of viruses with sequence homology to those with known taxonomy (Gregory *et al.* 2020). A comprehensive classification of the whole viral sequence space remains challenging due to the lack of universal marker genes in viruses. The gene content-based methods such as vConTACT and GRAViTy proposed in recent studies can accurately classify viruses and may

provide novel frameworks for a unified classification of all viruses (Eloe-Fadrosh, 2019).

Evolution of the Virome

In the era of viromics, molecular evolution on the virome level can provide a global view on the origin, diversity and evolution of viruses (Wolf *et al.* 2018; Gregory *et al.* 2019). Due to the lack of universal marker genes and the large diversity of viral nucleotide sequences, evolutionary analysis was often conducted on a group of viruses which share one or more marker genes (Wolf *et al.* 2018; Low *et al.* 2019). For example, Wolf *et al.* analyzed the origins and evolution of the RNA virome using the RNA-dependent RNA polymerase (RdRp) (Wolf *et al.* 2018). The results obtained from the study determined the evolutionary relationship among the double-stranded RNA, positive-stranded RNA and negative-stranded RNA viruses, and revealed the extensive gene module exchange among diverse viruses and the horizontal virus transfer between the distantly related hosts (Wolf *et al.* 2018).

Host Prediction of Viruses

The identification of viral hosts is essential for characterizing viruses, as viruses must rely on host cells for survival. Host predictions of eukaryotic viruses have been usually conducted based on viral sequences alone, such as those for influenza viruses and coronaviruses (Xu *et al.* 2017; Tian, 2020); while those of prokaryotic viruses have been usually conducted based on the similarity of sequence features or sequences between viruses and hosts. At present, two kinds of computational methods have been developed to predict prokaryotic virus hosts based on genomic sequences (Edwards *et al.* 2016; Ahlgren *et al.* 2017; Galiez *et al.* 2017; Lu *et al.* 2021). The first kind of methods rely on the sequence similarity search between the query viruses and the candidate host genomes since viruses and their hosts may share the same genes and/or short nucleotide sequences such as the spacer sequences used in CRISPR systems (Edwards *et al.* 2016). This kind of method can usually predict viral hosts with high accuracy, especially the CRISPR-spacer-based method (Edwards *et al.* 2016). However, they can be only used for a small proportion of viruses since only some viruses have sequence similarities with their hosts (Edwards *et al.* 2016). Another kind of methods can predict the viral hosts based on the sequence composition similarity between viruses and their hosts, such as the Prokaryotic virus Host Predictor (PHP) (Lu *et al.* 2021), VirHostMatcher (Ahlgren *et al.* 2017) and WiSH (Galiez *et al.* 2017). Although the latter kind of

Table 1 Illustration of computational methods or case studies in computational viromics.

Fields	Methods or case studies	Summary	Advantages	Limitations
Identification of viral genomic sequences	VIBRANT (Kieft <i>et al.</i> 2020)	Recovery, annotation and curation of microbial viruses from genomic sequences	Automated; user-friendly; accurate	Only for prokaryotic virus detection
	ViromeScan (Rampelli <i>et al.</i> 2016)	Detect eukaryotic viruses based on the homologous search method	Customized virus database; taxonomic assignment	Difficult to identify novel viruses
Annotation of viral genomes	PHANOTATE (McNair <i>et al.</i> 2019)	Gene annotations in phages based on viral gene characteristics	Reference-free; predicting more genes than other gene callers	Only for phages
	Vgas (Zhang <i>et al.</i> 2019a)	Combining <i>ab initio</i> and similarity-based method for predicting viral genes	High precision and recall rate	Accuracy needs further improvement
Taxonomic assignment of the virome	vConTACT (Eloe-Fadrosh, 2019)	Classification of prokaryotic virome based on the gene sharing network	Universal, scalable and automated	Not for short fragments, the singleton or outlier sequences; only tested for phages
	GRAViTy (Eloe-Fadrosh, 2019)	Classification of eukaryotic viruses at the family level within each Baltimore group	Concise and clear	Not for short contigs; only for eukaryotic viruses
Evolution of the virome	Analyzed the origins and evolution of the RNA virome (Wolf <i>et al.</i> 2018)	Reconstruct the RNA virus evolution using the RdRp protein, and reveal extensive gene module exchange and horizontal virus transfer among diverse viruses	A far more complete reconstruction of the evolution of RNA viruses than those in previous studies	Only for RNA viruses
Host prediction of phages	PHP (Lu <i>et al.</i> 2021)	A Gaussian model for host prediction of prokaryotic viruses in metagenomics	Accurate, fast and user-friendly	Accuracy needs further improvement
Virus-host interactions	P-HIPSTer (Lasso <i>et al.</i> 2019)	Prediction of PPIs between 1,001 human-infecting viruses and human based on structure information	Comprehensive and accurate	The codes are not available
	Predict the receptorome of human viruses (Zhang <i>et al.</i> 2020b)	Predicting the receptorome of the human-infecting virome based on the unique features of mammalian virus receptors	Fast and comprehensive	Accuracy needs further improvement; unable to predict virus-receptor interactions directly
Virus culturomics	KOMODO (Oberhardt <i>et al.</i> 2015)	A platform for recommending microbial media	Data-rich, user-friendly and relatively accurate	Only suitable for bacteria and archaea
Association of virome and human health	MicroPro (Zhu <i>et al.</i> 2019)	A data analysis pipeline for analysis of the association between the microbes in human body and the complex diseases	Combining both known and unknown microbial organisms	Did not consider the complex interactions between microbes
	HumanVirusFinder (Zhang <i>et al.</i> 2019b)	A machine learning model for identification of human-infecting viruses from viral metagenomic data	Fast, easy to use, suitable for genomes, contigs and reads	Limited data used in training the models

method predicts viral hosts with lower accuracy than the former, they can be used for any prokaryotic viruses. Viruses can change their hosts or spill over into other species after genetic mutations or recombinations (Letko *et al.* 2020), which poses a great challenge for the prediction of viral hosts. For better prediction of viral hosts, it is important to understand the host specificity of viruses determined by the interactions between viruses and their hosts.

Virus–Host Interactions

The complex interactions between virome and hosts are difficult to be resolved in viromics studies (Seo and Kweon, 2019). Besides, novel viruses are being discovered at an unprecedented pace (Gregory *et al.* 2019). So, the high-throughput experimental techniques and computational methods are in urgent need to analyze the interactions between viruses and their hosts (Lasso *et al.* 2019;

Lian *et al.* 2021; Zhang *et al.* 2020a, 2020b). The protein–protein interaction (PPI) prediction methods can help the identification of the interactions between virome and hosts to a great extent (Lasso *et al.* 2019; Lian *et al.* 2021). For example, Lasso *et al.* used the structural information to develop a computational framework to predict the PPIs between 1,001 human-infecting viruses and human, and they obtained a series of new findings about human-virus interactions such as the shared and unique machinery employed across human-infecting viruses and the previously unappreciated cellular circuits that act on human-infecting viruses (Lasso *et al.* 2019).

Virus Culturomics

The virus isolation and cultivation are the basis for further studies of the virus. Culturomics is defined as a systematic method to find the optimum culture conditions such as the culture medium and the incubation temperature for microbial cultivation (Greub, 2012). Many achievements in the field of bacteria culturomics have been obtained (Lagier *et al.* 2018). For example, Oberhardt *et al.* integrated known medium databases and a novel prediction tool into a platform that predicts the culture medium given an organism's 16S rRNA sequence (Oberhardt *et al.* 2015). Moreover, this platform can also predict culture media for new organisms using a transitivity property and a phylogeny-based collaborative filtering method. Similar to the work by Oberhardt *et al.*, it is possible to predict the cell line or tissue that can be used in virus cultivation based on the similarity of genomic sequences and the predicted PPIs between viruses and hosts.

Association of Virome and Human Health

The virome has a significant impact on human health. Previous studies have shown that the virome is associated with multiple diseases. However, the detailed mechanism is still unknown due to the complex interactions between the virome and their hosts (Clooney *et al.* 2019). Computational methods are needed to identify the viruses and their roles in causing human diseases. For example, Zhu *et al.* developed a metagenomic data analysis pipeline, MicroPro, to analyze the association between the microbes in the human body and complex diseases (Zhu *et al.* 2019). The virome is also closely related to the early warnings of newly emerging viruses. The Global Virome Project (GVP) has estimated that there are 631,000–827,000 unknown viruses with the potential of infecting humans (Carroll *et al.* 2018). Recent studies have developed machine-learning methods to identify the human-infecting virome

based on sequence features (Zhang *et al.* 2019b). More efforts are needed to validate their usage in applications.

Taken together, this perspective provides an overall view of computational viromics which includes the identification, annotation and taxonomic assignment of viral genomics sequences, phenotype prediction of viruses, evolution of viromes, virus-host interactions, virus culturomics, association of the virome and human health, and so on (Table 1). The computational viromics is still in the beginning stage. Much more computational methods and experimental efforts are needed to characterize the virome and its interactions with the hosts and environments considering the huge diversity of the global virome.

Acknowledgements This work was supported by the National Key Plan for Scientific Research and Development of China (2016YFD0500300) and Hunan Provincial Natural Science Foundation of China (2020JJ3006).

Compliance with Ethical Standards

Conflict of interest The authors declare that they have no conflict of interest.

Animal and Human Rights Statement This article does not contain any studies with human or animal subjects performed by the author.

References

- Ahlgren NA, Ren J, Lu YY, Fuhrman JA, Sun F (2017) Alignment-free oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic Acids Res* 45:39–53
- Altamirano FLG, Barr JJ (2019) Phage therapy in the postantibiotic era. *Clin Microbiol Rev* 32:e00066–e118
- Baud D, Qi X, Nielsen-Saines K, Musso D, Pomar L, Favre G (2020) Real estimates of mortality following Covid-19 infection. *Lancet Infect Dis* 20:773
- Brum JR, Ignacio-Espinoza JC, Kim E-H, Trubl G, Jones RM, Roux S, VerBerkmoes NC, Rich VI, Sullivan MB (2016) Illuminating structural proteins in viral “dark matter” with metaproteomics. *Proc Natl Acad Sci USA* 113:2436–2441
- Carroll D, Daszak P, Wolfe ND, Gao GF, Morel CM, Morzaria S, Pablos-Méndez A, Tomori O, Mazet JA (2018) The global virome project. *Science* 359:872–874
- Clooney AG, Sutton TD, Shkoporov AN, Holohan RK, Daly KM, O'Regan O, Ryan FJ, Draper LA, Plevy SE, Ross RP (2019) Whole-virome analysis sheds light on viral dark matter in inflammatory bowel disease. *Cell Host Microbe* 26:764–778.e5
- Edwards RA, Rohwer F (2005) Viral metagenomics. *Nat Rev Microbiol* 3:504–510
- Edwards RA, McNair K, Faust K, Raes J, Dutilh BE (2016) Computational approaches to predict bacteriophage–host relationships. *FEMS Microbiol Rev* 40:258–272
- Eloe-Fadrosch EA (2019) Towards a genome-based virus taxonomy. *Nat Microbiol* 4:1249–1250
- Fang Z, Tan J, Wu S, Li M, Xu C, Xie Z, Zhu H (2019) Ppr-meta: A tool for identifying phages and plasmids from metagenomic fragments using deep learning. *GigaScience* 8:giz066

- Fermin G (2018) Host range, host–virus interactions, and virus transmission. In: Tennant P, Fermin G, Foster JE (eds) *Viruses: molecular biology, host interactions, and applications to biotechnology*, 1st edn. Academic Press, London, pp 101–134
- Galiez C, Siebert M, Enault F, Vincent J, Söding J (2017) WiSH: Who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. *Bioinformatics* 33:3113–3114
- Gregory AC, Zayed AA, Conceição-Neto N, Temperton B, Bolduc B, Alberti A, Ardyna M, Arkhipova K, Carmichael M, Cruaud C (2019) Marine DNA viral macro-and microdiversity from pole to pole. *Cell* 177:1109–1123.e1114
- Gregory AC, Zablocki O, Zayed AA, Howell A, Bolduc B, Sullivan MB (2020) The gut virome database reveals age-dependent patterns of virome diversity in the human gut. *Cell Host Microbe* 28:724–740.e8
- Greub G (2012) Culturomics: a new approach to study the human microbiome. *Clin Microbiol Infect* 18:1157–1159
- Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ (2010) Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinf* 11:119
- Jofre J, Muniesa M (2020) Bacteriophage isolation and characterization: phages of *Escherichia coli*. In *horizontal gene transfer*. *Methods Mol Biol* 2075:61–79
- Kieft K, Zhou Z, Anantharaman K (2020) Vibrant: Automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* 8:1–23
- La Scola B, Desnues C, Pagnier I, Robert C, Barrassi L, Fournous G, Merchat M, Suzan-Monti M, Forterre P, Koonin E (2008) The virophage as a unique parasite of the giant mimivirus. *Nature* 455:100–104
- Lagier JC, Dubourg G, Million M, Cadoret F, Bilen M, Fenollar F, Levasseur A, Rolain JM, Fournier PE, Raoult D (2018) Culturing the human microbiota and culturomics. *Nat Rev Microbiol* 16:540–550
- Lasso G, Mayer SV, Winkelmann ER, Chu T, Elliot O, Patino-Galindo JA, Park K, Rabadan R, Honig B, Shapira SD (2019) A structure-informed atlas of human–virus interactions. *Cell* 178:1526–1541.e16
- Letko M, Seifert SN, Olival KJ, Plowright RK, Munster VJ (2020) Bat-borne virus diversity, spillover and emergence. *Nat Rev Microbiol* 18:461–471
- Lian X, Yang X, Yang S, Zhang Z (2021) Current status and future perspectives of computational studies on human–virus protein–protein interactions. *Brief Bioinf*. <https://doi.org/10.1093/bib/bbab029>
- Low SJ, Džunková M, Chaumeil P-A, Parks DH, Hugenholtz P (2019) Evaluation of a concatenated protein phylogeny for classification of tailed double-stranded DNA viruses belonging to the order *Caudovirales*. *Nat Microbiol* 4:1306–1315
- Lu C, Zhang Z, Cai Z, Zhu Z, Qiu Y, Wu A, Jiang T, Zheng H, Peng Y (2021) Prokaryotic virus host predictor: a gaussian model for host prediction of prokaryotic viruses in metagenomics. *BMC Biol* 19:5
- McNair K, Zhou C, Dinsdale EA, Souza B, Edwards RA (2019) Phanotate: A novel approach to gene identification in phage genomes. *Bioinformatics* 35:4537–4542
- Oberhardt MA, Zarecki R, Gronow S, Lang E, Klenk H-P, Gophna U, Ruppin E (2015) Harnessing the landscape of microbial culture media to predict new organism–media pairings. *Nat Commun* 6:8493
- Rampelli S, Soverini M, Turrone S, Quercia S, Biagi E, Brigidi P, Candela M (2016) Viromescan: a new tool for metagenomic viral community profiling. *BMC Genom* 17:1–9
- Ren J, Ahlgren NA, Lu YY, Fuhrman JA, Sun F (2017) Virfinder: A novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* 5:69
- Roux S, Enault F, Hurwitz BL, Sullivan MB (2015a) Virsorter: mining viral signal from microbial genomic data. *PeerJ* 3: e985
- Roux S, Hallam SJ, Woyke T, Sullivan MB (2015b) Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. *Elife* 4: e08490
- Seo SU, Kweon MN (2019) Virome–host interactions in intestinal health and disease. *Curr Opin Virol* 37:63–71
- Simmonds P, Adams MJ, Benkő M, Breitbart M, Brister JR, Carstens EB, Davison AJ, Delwart E, Gorbalenya AE, Harrach B (2017) Consensus statement: Virus taxonomy in the age of metagenomics. *Nat Rev Microbiol* 15:161–168
- Stern-Ginossar N, Thompson SR, Mathews MB, Mohr I (2019) Translational control in virus-infected cells. *Cold Spring Harb Perspect Biol* 11:a033001
- Suttle CA (2007) Marine viruses—major players in the global ecosystem. *Nat Rev Microbiol* 5:801–812
- Tian BP (2020) The potential intermediate hosts for SARS-CoV-2. *Front Microbiol* 11:580137
- Walker PJ, Siddell SG, Lefkowitz EJ, Mushegian AR, Adriaenssens EM, Dempsey DM, Dutilh BE, Harrach B, Harrison RL, Hendrickson RC (2020) Changes to virus taxonomy and the statutes ratified by the international committee on taxonomy of viruses (2020). *Arch Virol* 165:2737–2748
- Wolf YI, Kazlauskas D, Iranzo J, Lucía-Sanz A, Kuhn JH, Krupovic M, Dolja VV, Koonin EV (2018) Origins and evolution of the global rna virome. *Mbio* 9:e02329-18
- Xu B, Tan Z, Li K, Jiang T, Peng Y (2017) Predicting the host of influenza viruses based on the word vector. *PeerJ* 5:e3579
- Zhang KY, Gao YZ, Du MZ, Liu S, Dong C, Guo FB (2019a) Vgas: a viral genome annotation system. *Front Microbiol* 10:184
- Zhang Z, Cai Z, Tan Z, Lu C, Peng Y (2019b) Rapid identification of human-infecting viruses. *Transbound Emerg Dis* 66:2517–2522
- Zhang Z, Yu F, Zou Y, Qiu Y, Wu A, Jiang T, Peng Y (2020a) Phage protein receptors have multiple interaction partners and high expressions. *Bioinformatics* 36:2975–2979
- Zhang Z, Ye S, Wu A, Jiang T, Peng Y (2020b) Prediction of the receptorome for the human-infecting virome. *Virol Sin* 36:133–140
- Zhu Z, Ren J, Michail S, Sun F (2019) Micropro: Using metagenomic unmapped reads to provide insights into human microbiota and disease associations. *Genome Biol* 20:154