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 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
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
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;
et al. (2021). 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 R&D Program 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-Mendoza A, Tomori O, Mazet J (2018) The global virome project. Science 359:872–874
Cloney 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-Fadrosh 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
