


# VIGA: a one-stop tool for eukaryotic virus identification and genome assembly from next-generation-sequencing data

Ping Fu, Yifan Wu, Zhiyuan Zhang, Ye Qiu, Yirong Wang and Yousong Peng 

Corresponding author. Yousong Peng, Bioinformatics Center, College of Biology, Hunan Provincial Key Laboratory of Medical Virology, Hunan University, Tianma Road, Yuelu District, Changsha 410082, China. Tel.: 0731-88822606; Fax: 0731-88821720; E-mail: [pys2013@hnu.edu.cn](mailto:pys2013@hnu.edu.cn)

## Abstract

Identification of viruses and further assembly of viral genomes from the next-generation-sequencing data are essential steps in virome studies. This study presented a one-stop tool named VIGA (available at <https://github.com/viralinformatics/VIGA>) for eukaryotic virus identification and genome assembly from NGS data. It was composed of four modules, namely, identification, taxonomic annotation, assembly and novel virus discovery, which integrated several third-party tools such as BLAST, Trinity, MetaCompass and RagTag. Evaluation on multiple simulated and real virome datasets showed that VIGA assembled more complete virus genomes than its competitors on both the metatranscriptomic and metagenomic data and performed well in assembling virus genomes at the strain level. Finally, VIGA was used to investigate the virome in metatranscriptomic data from the Human Microbiome Project and revealed different composition and positive rate of viromes in diseases of prediabetes, Crohn's disease and ulcerative colitis. Overall, VIGA would help much in identification and characterization of viromes, especially the known viruses, in future studies.

**Keywords:** Virus identification genome assembly; NGS; metatranscriptomic; metagenomic

## INTRODUCTION

Viruses are ubiquitous in nature and have profound impacts on the health and diversity of all living organisms [1]. However, most viruses remain unknown and unculturable due to the limitations of conventional isolation methods [2]. In recent years, the rapid development of next-generation-sequencing (NGS) technologies has revolutionized the field of metagenomics and transcriptomics [3], enabling the analysis of nucleic acid sequences obtained directly from environmental or clinical samples. For example, the Human Microbiome Project (HMP) [4] has facilitated the generation of hundreds of reference microbial genomes, along with thousands of metagenomic and metatranscriptomic datasets from multiple parts of the human body. This has greatly increased the number of viral genomes and provided valuable information for studying viral evolution, diversity and epidemiology [5].

Identification of viruses from NGS data is the first step in virome studies. There are currently two kinds of methods for virus identification. The first is the homology-based methods, such as using tools of BLAST [6] or HMMER [7] to search for viruses that have sequence similarity to known viruses. This kind of methods is commonly used in identifying eukaryotic viruses that have been studied much. For example, Moustafa *et al.* [8] identified 19 viruses by whole genome sequencing of blood from 8240 individuals using BLAST. The advantage of these methods is that they are generally accurate with a low rate of false positives [9], but they

may miss identification of novel viruses that have remote or little homology with known viruses [10]. This problem becomes serious in discovery of phages that have huge genetic diversity. To address the problem, the other kind of methods has been developed for virus identification based on machine-learning methods, such as Virtifier [11], Seeker [12] or VirFinder [13]. They can detect viruses with higher sensitivity than the homology-based methods and are generally used in identifying phages. For example, the project of Global Ocean Viromes 2.0 (GOV 2.0) [14] have identified 195 728 viral populations from the global ocean DNA virome dataset based on this kind of methods. Unfortunately, a high false-positive rate was observed for these methods [15] and there was also inconsistency between the prediction results by different methods [16].

Assembling virus genomes is also an essential step for further studies of the virome. Unfortunately, assembling large genomic fragments from short-read sequencing data is a formidable computational challenge [17]. In the case of viruses, the presence of repetitive region [18], strain heterogeneity [19] and low-abundance viral populations [20] can pose significant difficulties for accurate assembly of virus genomes. Besides, the quality of the assembly may be compromised by errors in the sequencing data. To address these challenges, researchers have developed various methods for improving the accuracy and quality of the virus genome assembly that can be mainly grouped into two categories [21]: one is the reference-based assembly methods,

Ping Fu is PhD candidate in Hunan University.

Yifan Wu is PhD candidate in Hunan University.

Zhiyuan Zhang is a master student in Hunan University.

Ye Qiu is professor in Hunan University.

Yirong Wang is professor in Hunan University.

Yousong Peng is professor in Hunan University.

Received: April 27, 2023. Revised: October 26, 2023. Accepted: November 11, 2023

© The Author(s) 2023. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

such as MetaCompass [22] and VirGenA [23], and the other is the *de novo* assembly methods, such as Trinity [24] and Haploflow [25]. The reference-based methods assembling genomes by taking a known genome as a guide. The advantage of this kind of methods is that they are generally more accurate than the *de novo* methods [22], but they are not suitable for genome assembly of viruses without reference genomes. On the contrary, the *de novo* methods can be applied to all viruses including novel viruses, although they generally require a deep sequencing depth and may not assemble complete genomes [26].

In this study, a one-stop tool named VIGA was developed for eukaryotic virus identification and genome assembly from NGS data. It integrated multiple assembly tools including both the reference-based and *de novo* ones and combines both functions of virus identification and genome assembly. Evaluation on multiple simulated and real virome datasets showed that VIGA could be used for assembling virus genomes and separating mixtures of virus strains from metagenomic and transcriptomic data.

## MATERIALS AND METHODS

### Data collection

To evaluate the performance of computational tools in virus identification and genome assembly, three datasets were used: the first is a dataset of mock viral community (accession in NCBI BioProject database [27]: PRJNA431646 [17]) that consisted of 3 090 013 paired-end sequencing reads from seven viruses with varying proportions: human poliovirus 1 (47.04%), human mastadenovirus C (30.90%), coxsackievirus B4 (13.30%), murid gammaherpesvirus 4 (7.43%), echovirus E13 (0.78%), human mastadenovirus B (0.55%) and rotavirus A (0.001%).

The second is an RNA-Seq dataset of sweet potato viromes that included 10 samples (accession in NCBI BioProject database: PRJNA517178 [28]). Previous studies have identified 10 virus species, namely, the sweet potato symptomless virus 1 (SPSMV), sweet potato latent virus (SPLV), sweet potato virus G (SPVG), sweet potato virus 2 (SPV2), sweet potato virus C (SPVC), sweet potato leaf curl virus (SPLCV), sweet potato feathery mottle virus (SPFMV), sweet potato virus E (SPVE), sweet potato virus F (SPVF) and sweet potato chlorotic fleck virus (SPCFV) from the dataset using the homology-based method and obtained their genomes (accession numbers in NCBI GenBank database: NC\_034630, NC\_020896, NC\_018093, NC\_017970, NC\_014742, NC\_004650, NC\_001841, NC\_006550, MH388501 and MH388502) by the RT-PCR method.

The third is a viral metagenomic dataset of the bird feces that included 16 samples (accessions in NCBI SRA database: SRR10873766, SRR10874069, SRR10873474, SRR10875205, SRR10874841, SRR10874829, SRR10874826, SRR10874825, SRR10874819, SRR10874379, SRR10873966, SRR10873876, SRR10873785, SRR10873770, SRR10873756, SRR10873746 [29]). Previous studies have identified 16 viral strains of three virus species including *Riboviria* sp. (accessions of viral genome sequences in NCBI GenBank database: MN933892.1, MT138199.1, MT138205.1, MT138191.1, MN933887.1, MT138420.1 and MT138407.1), CRESS virus sp. (accessions: MN928923.1, MN928933.1, MN928938.1, MN928948.1, MN928929.1, MT138070.1 and MT138040.1) and *Picornavirales* sp. (accessions: MT138390.1 and MT138137.1) from the dataset using the homology-based method and obtained virus genomes by the RT-PCR method.

To evaluate the performance of computational tools in assembling viral genomes at the strain level, two datasets were used. The first is the HIV dataset, which was created in Fritz's study

by mixing three HIV strains with about 95% sequence identity in proportions 10:5:2 [25]. The dataset contains simulated metagenomic sequencing reads with the length of 150 bp and the depth of 20 000 (available for download at <https://frrl.publisso.de/data/frrl:6424451/>). The other is the hepatitis B virus (HBV) dataset [30], which is the metagenomic sequencing of two clinical samples that were co-infected by two HBV strains with 89% sequence identity (accessions in NCBI SRA database: ERR3253398 and ERR3253399, accessions of viral genomic sequences in NCBI GenBank database: MK720628.1 and MK720631.1).

The metatranscriptomic datasets derived from the HMP (<https://portal.hmpdacc.org/>) [4] were used to illustrate the usage of VIGA. Only the paired-end sequencing samples were used as VIGA can only use the paired-end data. A total of 1321 samples were finally used. The meta information including the disease state, age and gender of samples was also obtained (available at [https://github.com/viralInformatics/VIGA/blob/master/HMP\\_datasets.xlsx](https://github.com/viralInformatics/VIGA/blob/master/HMP_datasets.xlsx)).

### The workflow of VIGA

VIGA included four modules: identification, taxonomic annotation, assembly and novel virus discovery, which were described as follows. The identification module aimed to detect eukaryotic virus sequences from NGS data. Firstly, fastp (version 0.21.0) [31] was used to trim the universal adapter sequences from raw reads and filter reads with average base quality less than Q20. Secondly, the remaining clean reads were assembled into contigs using the Trinity program (version 2.1.1) [24] with default parameter settings. Thirdly, the contigs were queried against a library of virus protein sequences retrieved from the NCBI RefSeq database on 10 June 2020 using Diamond BLASTX (version 0.9.25.126) [32]. The contigs with an E-value of less than 1E-5 to the best hit were labeled as hypothetical viral contigs. Fourthly, the hypothetical viral contigs were queried against the NR database (downloaded on 17 November 2020) with Diamond BLASTX to remove false positives. Those with the BLASTX best hit belonging to viruses were kept and considered as viral contigs. Finally, they were filtered based on host of the BLASTX best hit, and only those with the BLASTX best hit belonging to eukaryotic viruses were kept for further analysis.

The taxonomic annotation module was designed for taxonomic annotation of viral contigs identified above. The taxonomy information of viral contigs was obtained according to the amino acid identity (AAI) and coverage between the contig and the BLASTX best hit against the NR database. If the AAI and coverage were no less than 90% and 80%, respectively, the viral contig was considered to be the same virus species with the BLASTX best hit according to previous studies [33–35]; if they were no less than 70% and 60%, respectively, the viral contig was considered to be the same genus with the BLASTX best hit according to previous studies [36–38]. For viral contigs that were annotated at the species level, the assembly module would be conducted; for those that were annotated at the genus level, the novel virus discovery module would be conducted.

The Assembly module assembled and quantified the viral genome for the virus species after the taxonomy annotation. Firstly, a library of virus reference genomes was built as follows: (i) genome sequences of all eukaryotic viruses were downloaded from the NCBI Genome database on 17 August 2022. (ii) Genome sequences of the same virus species were clustered using the MMseqs2 (version 13.45111) [39] at 100% level. The representative sequence in each cluster was added to the library of virus reference genomes. Secondly, all representative sequences of the

virus species were taken as reference genomes and were used to assemble the virus genome using MetaCompass (version 2.0.0, parameter: '-l 100 -t 30') [22]. Thirdly, RagTag (version 2.1.0, parameter: '-u -C') [40, 41] was used to correct potential assembly errors in contigs based on the reference genomes. Then, MetaQUAST (version 5.0.2) [42] was used to evaluate the quality of assembled viral genomes and output the genome fraction (GF) of the assembled virus genome, which measures the genome completeness. Then, the assembled viral genomes were quantified using the FPKM (fragments per virtual kilobase per million sequenced reads) method according to Lee's study [43], which was listed as follows:

$$\text{FPKM} = \frac{M * 10^3 * 10^6}{N * L_v}$$

where  $L_v$  is the length of the assembled viral genome (bp),  $M$  is the number of reads assigned to the virus genome and  $N$  is the total number of clean reads. The genome coverage [44] and the depth coverage [45] of the assembled genome that measure the sequencing broadness and depth of the genome, respectively, would also be outputted. Finally, the depth coverage of the assembled genome would be visualized with a line chart.

The novel virus discovery module was designed for novel virus discovery. Viral contigs that were annotated as the same genus were pooled together and defined as a new virus class (NVC). The read coverage of the NVC, which measures the sequencing depth of the NVC, the AAI and coverage between viral contigs and the BLASTX best hits against the NR database, and the detailed information of the BLASTX best hits would be provided.

## Comparison of VIGA and other assembly tools

Since VIGA integrated both *de novo* and reference-based assemblers for virus identification and genome assembly, four commonly used assembly tools, namely, two reference-based ones, i.e. MetaCompass [22] and VirGenA [23], and two *de novo* ones, i.e. Haploflow [25] and Trinity [24], were used for virus genome assembly and for comparison to VIGA. For the *de novo* assemblers, Trinity is a widely used tool for *de novo* contig assembly [46]. Previous studies have shown that it had excellent performances in virus genome assembly [47–49]. Thus, it is used for comparison to VIGA in assembling virus genomes. Haploflow was designed for assembly of virus genomes, especially the haplotype reconstruction [25]. It has been demonstrated to outperform other tools in haplotype reconstruction [25]. Thus, it is used to compare its performance to VIGA in assembling virus genomes at the strain level. For the reference-based assemblers, VirGenA is a tool designed for separating mixtures of viral strains of different intraspecies genetic groups such as genotypes, subtypes and clades and for assembling a separate consensus sequence for each group in a mixture [23]. It has been reported to produce long assemblies for mixture components of extremely low frequencies (<1%) [23]. Thus, it is used for comparison to VIGA in assembling virus genomes at the strain level. MetaCompass utilizes reference genomes as a guide to construct consensus sequences and employs *de novo* assembly to address regions dissimilar to known reference genomes, thereby enhancing the overall assembly quality [22]. It is designed for metagenomic assembly of low-abundance bacterial genomes and should be particularly useful for assembling virus genomes since most viruses have low abundances in either metagenomic or metatranscriptomic sequencing data [22]. Besides, it is part of VIGA and can be used to investigate the effectiveness of integrating multiple tools in VIGA.

Besides, all competitor tools used in the study have undergone extensive validation and have been widely used in the field. They are freely open to users and easy to install. The local version of MetaCompass (version 2.0.0, available at <https://github.com/marbl/MetaCompass>) was used by specifying the reference genome and setting the read length of 100 and other parameters as default. The local version of VirGenA (version 1.4, available at <https://github.com/gFedonin/VirGenA/releases>) was used with the insertion length of 800 and other parameters as default. The local version of Haploflow (version 1.0, available at <https://github.com/hzi-bifo/Haploflow>) was used with default parameters. The local version of Trinity (version 2.1.1, available at <https://github.com/trinityrnaseq/trinityrnaseq>) was used with default parameters.

The GF [42] was used to measure the completeness of the assembled genome. It was calculated as the ratio of the virus genome covered by the bases to which at least one contig has alignment. Two additional metrics were used to measure the accuracy and precision in assembling virus genomes at the strain level. One is the mismatches per 100 kb (M100K) [42], which referred to the average number of mismatches per 100 000 aligned bases and was used to measure the accuracy of assembled virus genomes. The other is the strain precision (SP) [25], which was calculated as the ratio of correctly assembled contigs among all contigs of the virus and was used to assess the precision of assembled virus genomes.

## Virus identification and genome assembly on the HMP

VIGA was used to identify viruses and assemble virus genomes from the HMP. The host kingdom (plant, fungi and animal) of the identified viruses was inferred based on the virus family as viruses of the same family generally infect host of the same kingdom. When analyzing the virome in diseases, only animal viruses with abundance of greater than 2 FPKM were kept. Besides, viruses of the Retroviridae family were removed due to possible contamination from endogenous retroviral sequences, and viruses of the Baculoviridae family were also removed, which are commonly used in the laboratory [50].

## Statistical analysis

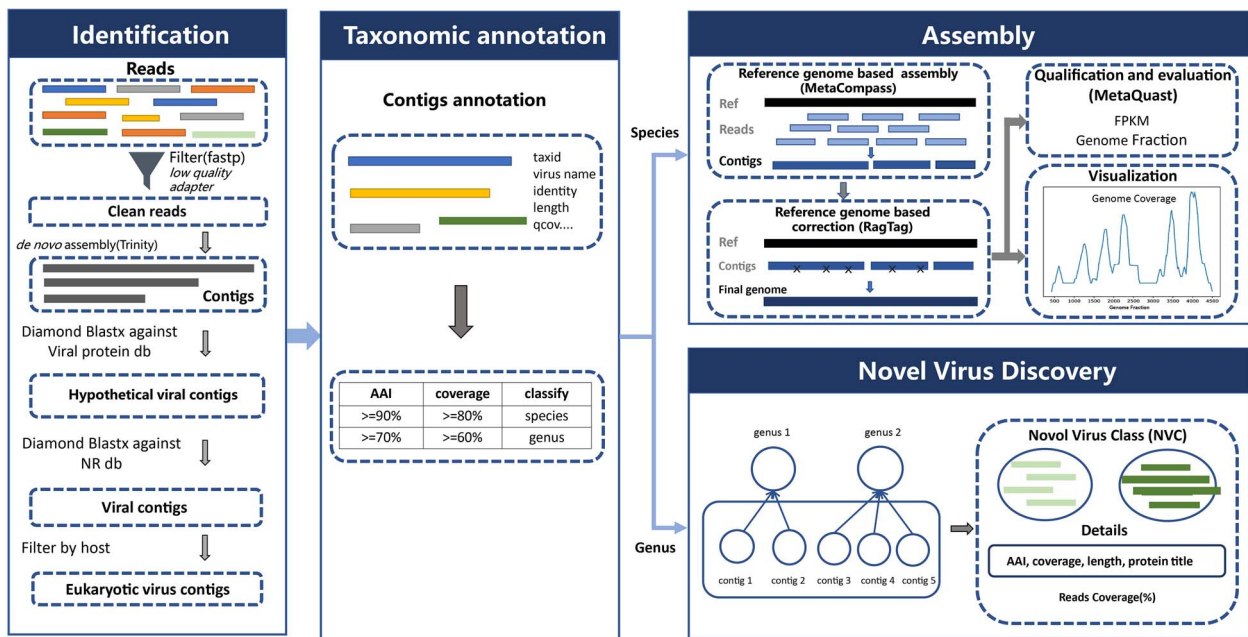
The Pearson correlation coefficient (PCC) was used to measure the correlation between virus abundance and virus percentage in the mock virus community. Python (version 3.8) was used for statistical analysis.

# RESULTS

## Overview of the VIGA pipeline

The VIGA included four modules: identification, taxonomic annotation, assembly and novel virus discovery (Figure 1), which were described briefly as follows. The identification module aimed to detect eukaryotic virus sequences from NGS data. Firstly, the raw reads were pre-processed and then were assembled into contigs. Then, these contigs were queried against virus protein sequences using BLASTX, and candidate virus contigs were obtained. Subsequently, to remove false positives, the candidate virus contigs were further queried against the NR database using BLASTX. Finally, the virus contigs were filtered based on host, and only eukaryotic virus sequences were kept for further analysis.

The Taxonomic annotation module was designed for taxonomic annotation of viral contigs identified above. The taxonomy information of viral contigs was obtained according to the AAI and



**Figure 1.** The workflow of VIGA. It includes four modules: identification, taxonomic annotation, assembly and novel virus discovery.

coverage between the contig and the BLASTX best hit against the NR database (see [Materials and methods](#)). For viral contigs that were annotated at the species level, the assembly module would be conducted; for those that were annotated at the genus level, the novel virus discovery module would be conducted.

The assembly module assembled and quantified the viral genome for the virus species after the taxonomic annotation. Firstly, the reference genomes were obtained for the virus species; then, MetaCompass was used to assemble the virus genome based on reference genomes of the virus species. Next, RagTag was used to assemble the virus genome after correcting potential assembly errors in contigs based on the reference genomes; finally, the assembled viral genomes were evaluated with MetaQUAST and quantified using FPKM.

The novel virus discovery module was designed for novel virus discovery. Viral contigs that were annotated as the same genus were pooled together and defined as a new virus class (NVC). Then, the read coverage of the NVC, the AAI and coverage between viral contigs and BLASTX best hit against the NR database, and the detailed information of the BLASTX best hits was provided.

### Evaluating the performance of VIGA on a mock viral community

The performance of VIGA was firstly evaluated on a mock viral community that consisted of seven viruses with different proportions ([Materials and methods](#)). For comparison, we also evaluated four commonly used tools, namely, two reference-based (MetaCompass and VirGenA) and two *de novo* tools (Trinity and Haploflow). As illustrated in [Figure 2A](#), VIGA successfully identified six viruses except the rotavirus A. For viruses of PV-1, EV-13, HAdVC and CV-B4, the GFs of viral genomes assembled by VIGA exceeded 0.94, which were much higher than those of its competitors. For example, the GFs of viral genomes assembled by MetaCompass and Trinity ranged from 0.74 to 0.94 and from 0.50 to 0.84, respectively. For viruses of HAdV-5 and HAdV-11, VIGA performed poorly, with the GFs of 0.636 and 0.064, respectively; however, it still outperformed its competitors much.

Previous studies have shown that the transcript abundance that was calculated based on the effective length (defined as

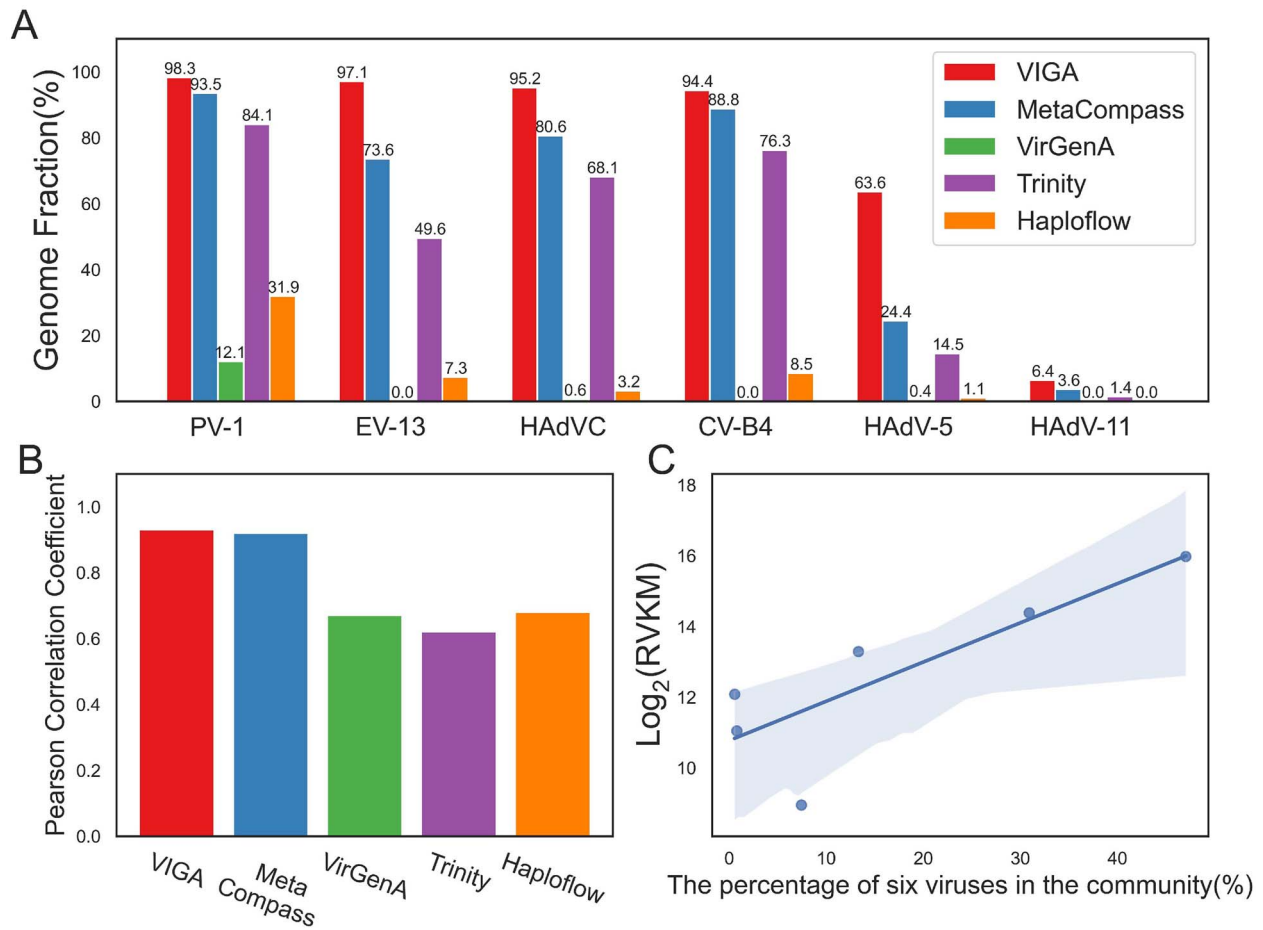
the length of transcript regions that were covered by uniquely mapped reads) can reflect the gene expression more accurately than the abundance based on the full length of the transcript. Hence, the length of the assembled virus genome was taken as the effective length and was used to calculate the virus abundance using the FPKM method ([Materials and methods](#)). The PCC was used to measure the correlation between the calculated virus abundances and the percentages of viruses in the mock virus community that reflect the real virus abundance. As shown in [Figure 2B and C](#), VIGA had a PCC of 0.93, which was similar to that of MetaCompass (0.92) and much higher than those of other software tools (0.62–0.68), suggesting that the virus quantification based on the virus genome assembled by VIGA could accurately capture the real virus abundance in complex virus communities.

### Performance of VIGA on metatranscriptomic and metagenomic datasets

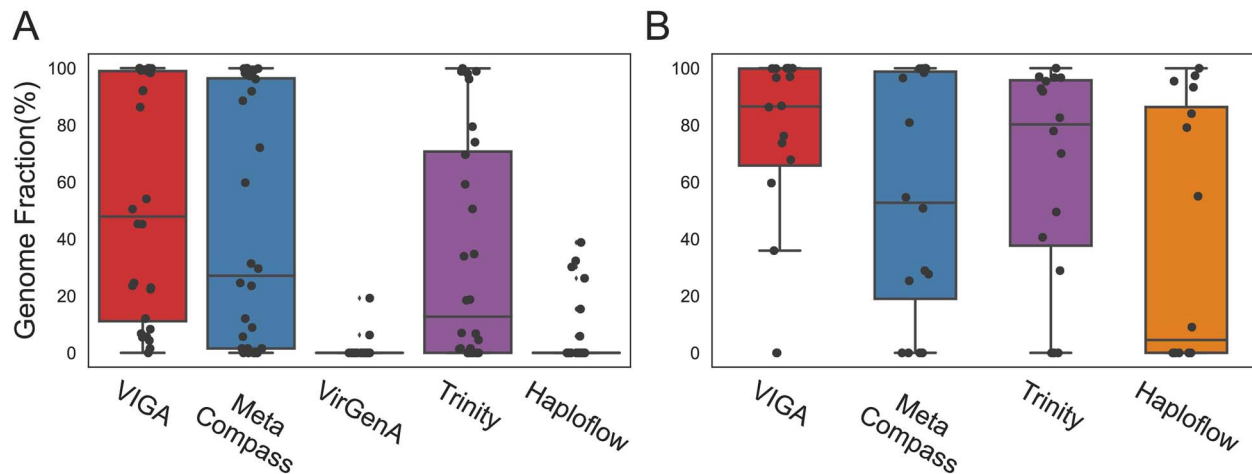
Then, VIGA was evaluated on a metatranscriptomic dataset of the sweet potato virome from which previous studies have identified 10 viruses and obtained their genomes by the RT-PCR method. VIGA and other four assemblers (MetaCompass, VirGenA, Trinity and Haploflow) were used to assemble virus genomes in the dataset. Except for SPCFV, all viruses were assembled by at least one tool. The GF of virus genomes assembled by VIGA ranged from 1.5% to 100% for these nine viruses ([Figure 3A](#)), with a median of 47.9%, which was greater than those of other software tools (median ratio ranging from 0% to 27.1%). Notably, both VIGA and MetaCompass assembled near complete genomes for five viruses, namely, sweet potato virus G (SPVG, 99.17%), sweet potato leaf curl virus (SPLCV, 100%), sweet potato feathery mottle virus (SPFMV, 98.90%), sweet potato virus E (SPVE, 99.94%) and sweet potato virus F (SPVF, 99.45%), three of which (SPLCV, SPFMV and SPVE) were also assembled well by Trinity. However, both VirGenA and Haploflow performed poorly and assembled only a small proportion of genomes for all nine viruses.

VIGA was also evaluated on a metagenomic dataset of bird droppings from which 16 viral strains of three virus species





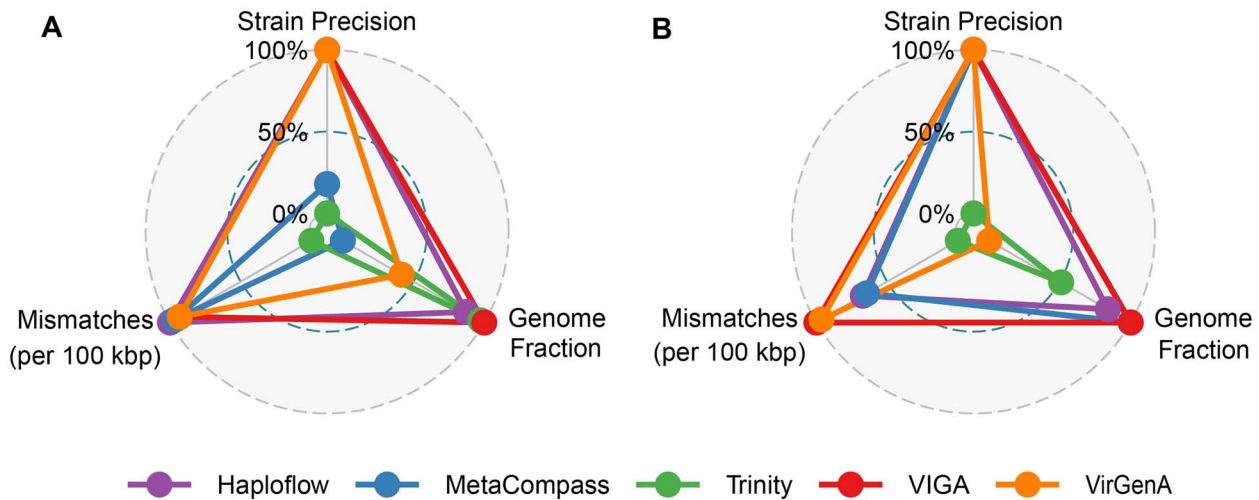
**Figure 2.** The performances of VIGA and its competitors on a mock virus community. (A) The ratios of viral genomes assembled by VIGA and its competitors including MetaCompass, VirGenA, Trinity and Haploflow. (B) The PCC between the calculated virus abundance based on the viral genomes assembled by different software tools and the viral percentages in the virus community that reflect the real virus abundance. (C) The scatter plot of the calculated virus abundance based on VIGA and the real virus abundance. The blue line referred to the linear regression and the gray area referred to 95% confidence interval. PV-1, human poliovirus 1; EV-13, human echovirus 13; HAdVC, human adenovirus C; CV-B4, coxsackievirus B4; HAdV-5, human adenovirus 5; HAdV-11, human adenovirus type 11.



**Figure 3.** Performance comparison of different software tools on metatranscriptomic (A) and metagenomic (B) datasets. Boxplots show the statistical distribution of GF of assembled virus genomes by different tools. Each dot in the boxplots referred to the GF of the virus genome assembled in a sample. The results of VirGenA on the metagenomic data were not shown as no viral genomes were assembled successfully by the assembler.

were identified, and their genomes were obtained by the RT-PCR method. VIGA stood out as the most effective tool for assembling viral genomes, with the GF ranging from 0% to 100% and a median of 86.54% (Figure 3B), while other three assemblers

(MetaCompass, Haploflow and Trinity) had the median GFs ranging from 4.5% to 80.3%. Notably, VIGA assembled high ratios of virus genomes in most samples, with only two exceptions resulting in no genome assembly.



**Figure 4.** Performance of five assemblers (Haploflow, MetaCompass, Trinity, VIGA and VirGenA) on the HIV (A) and HBV (B) datasets. Three indexes, namely, GF, SP and M100K were used to measure the performances of assemblers. Each index was normalized with the min–max normalization method before it was used in the radar plot.

### Performance of VIGA in assembling virus genomes at the strain level

We then evaluated the performance of VIGA and its competitors in assembling virus genomes at the strain level. Two datasets were used in the evaluation. The first is the HIV dataset, which was the simulated metagenomic sequencing of three HIV strains with a 95% sequence identity. VIGA assembled genomes of all three strains successfully. Three indexes, namely, GF, SP and M100K, were used to measure the performance of assemblers (see [Materials and methods](#)). As shown in [Figure 4A](#) and [Table S1](#), VIGA achieved near-perfect performance on indexes of GF and SP, with an average GF of 98.20% and an SP of 100%, while in terms of M100K, VIGA had 2787 mismatches per 100 kbp, which was much smaller than those of other assemblers except Haploflow that had the lowest mismatches per 100 kbp (33.7) among all assemblers. Compared to VIGA, Haploflow also achieved excellent performances on these indexes, with the SP of 100% and the average GF of 93.36%; other three assemblers had medium or poor performances on at least one index. For example, VirGenA had poor performances on the GF (76.14%).

The second is the HBV dataset, which was the metagenomic sequencing data of two samples infected by two HBV strains with 89% sequence identity. As shown in [Figure 4B](#) and [Table S2](#), VIGA performed best on three indexes among all assemblers, with the average GF of 99.91%, SP of 100% and 1890.7 mismatches per 100 kbp; Haploflow performed secondly among all assemblers, with the average GF of 91.19%, SP of 100% and 2355 mismatches per 100 kbp; other assemblers had median or poor performances on at least one index. For example, Trinity had an average GF of 73.48% and 3306.4 mismatches per 100 kbp.

### Performance of VIGA in assembling virus genomes with extreme genome size or GC content

To assess the robustness of VIGA, we evaluated its performance alongside that of its competitors when applied to viruses with very small or large genomes, as well as those with very low or high GC content. We initially selected HBV and Human Betaherpesvirus 5 (HHV5), whose genomes are approximately 3 and 229 kb in size, respectively. As indicated in [Table S3](#), VIGA demonstrated the highest GF and the fewest mismatches and achieved 100%

precision for both viruses. Furthermore, VIGA and MetaCompass produced only a small number of contigs for virus genome assembly, whereas other tools yielded a greater number of contigs.

Next, we examined the performance of VIGA and its competitors on viruses with either low or high GC content. Suid Herpesvirus 1 (SuHV-1) and Diachasmimorpha Longicaudata Entomopoxvirus (DIEPV) were chosen, with GC content of 73.59% and 30.07%, respectively. As shown in [Table S4](#), VIGA consistently assembled genomes with the highest GF, achieved 100% precision and produced the fewest contigs for both viruses, although it had a little more mismatches than MetaCompass. Taken together, the biased conditions had a negligible impact on the performance of VIGA in virus genome assembly.

### Configuration, installation, time and memory consumption of VIGA

We further compared the installation, configuration, time and memory consumption of VIGA and its competitors ([Table 1](#)). The installation and configuration of VIGA are more complicated compared to other tools, such as Haploflow and Trinity, because it integrated a lot of third-party tools for both virus identification and genome assembly, while other tools are only used for genome assembly. A step-by-step document about the installation, configuration and usage of VIGA was provided along with the tool to facilitate its usage. A medium-sized RNA-Seq dataset (2.8G) of porcine reproductive and respiratory syndrome virus (PRRSV) infection was used to test the time and memory consumption of VIGA and its competitors. MetaCompass ran the fastest among all tools. Both VIGA and MetaCompass completed the task in less than 100 min, which was much faster than Trinity and VirGenA. They also had much smaller peak memory consumption compared to the other tools.

### Application of VIGA in identifying and assembling virus genomes from the HMP

In order to illustrate the practical usage of VIGA on large datasets, we reanalyzed 1321 metatranscriptomic samples from the HMP project (<https://portal.hmpdacc.org/>) using VIGA, with the total data volume of 1.14T. A total of 125 known eukaryotic viruses were identified from 467 samples, with 72 plant viruses, 31 animal viruses, 4 fungal viruses and 18 viruses infecting both

**Table 1:** The level of installation, setup and usage, as well as time and memory consumption of VIGA, MetaCompass, Haploflow, Trinity and VirGenA

Tools	Installation	Configuration	Usage	Time consumption (min)	Peak memory consumption (MB)
VIGA <sup>a</sup>	Manually	Complicated	Easy	95.3	7359.1
MetaCompass	Manually	Complicated	Easy	67.2	7359.1
Haploflow	Conda	Easy	Easy	136.1	62 757.5
Trinity	Conda	Complicated	Easy	622.9	32 783.5
VirGenA <sup>b</sup>	Manually	Complicated	Complicated	846.3	36 166.5

The time and memory consumption of assemblers were tested using an RNA-Seq dataset of PRRSV infection (accession in NCBI SRA database: SRR2123928, accession of virus genome sequence in NCBI GenBank database: AF325691.1). The test was conducted on a server with the following configuration: OS, Ubuntu 16.04.7 LTS; CPU, Intel(R) Xeon(R) Silver 4114, 2.20GHz, 20 cores and 40 threads; RAM, 128GB). Memory consumption was recorded using the 'memory\_profiler' module in Python. <sup>a</sup>Only the Assembly module was run for a fair comparison. <sup>b</sup>No results returned when the task was completed as VirGenA cannot deal with large datasets.

plants and fungi (Figure 5A) (available at [https://github.com/viralinformatics/VIGA/blob/master/Virus\\_recovery\\_from\\_HMP\\_using\\_VIGA.xlsx](https://github.com/viralinformatics/VIGA/blob/master/Virus_recovery_from_HMP_using_VIGA.xlsx)). The GFs of the assembled virus genomes ranged from 1.4% to 100% with a median of 33.8% (Figure 5B). A total of 44 viruses were assembled with high completeness (GF >80%). We next focused on analysis of animal viruses since the samples were obtained from humans. After removing potential contaminations (see Materials and methods), a total of 28 animal viruses were kept for further analysis. The abundance and positive rate of 28 viruses in samples of three human diseases, namely, Crohn's disease, prediabetes and ulcerative colitis, and healthy people were analyzed. As shown in Figure 5C, a total of 11 viruses were observed in the healthy people, with the Mongoose picobornavirus and Human picobornavirus having the highest positive rate. The virome composition in disease samples was different from that in healthy people. For example, only 9 of 16 viruses in the Crohn's disease were also observed in the healthy people. For each disease group, there were one or more disease-specific viruses. For example, the Rotavirus A, which has been previously implicated in the development of intestinal diseases such as diarrhea, had high abundances in patients of the Crohn's disease with a median of 4837 FPKM. The virome composition between Crohn's disease and ulcerative colitis was similar, with nine virus species coexisting in both diseases, while the prediabetes had different virome composition to the Crohn's disease and ulcerative colitis.

## DISCUSSION

Numerous methods have been developed to identify virus or assemble virus genome from the NGS data. However, there is still a lack of an effective tool for integrating both the functions of virus identification and genome assembly. This study presented a one-stop tool named VIGA for virus identification and genome assembly from the NGS data. It has been shown to assemble more complete virus genomes than its competitors in both the metatranscriptomic and metagenomic data and perform well in assembling virus genomes at the strain level. It has also been used to characterize the virome in metatranscriptomic data from the HMP. Overall, it should be an effective tool for virus identification and genome assembly in virome studies.

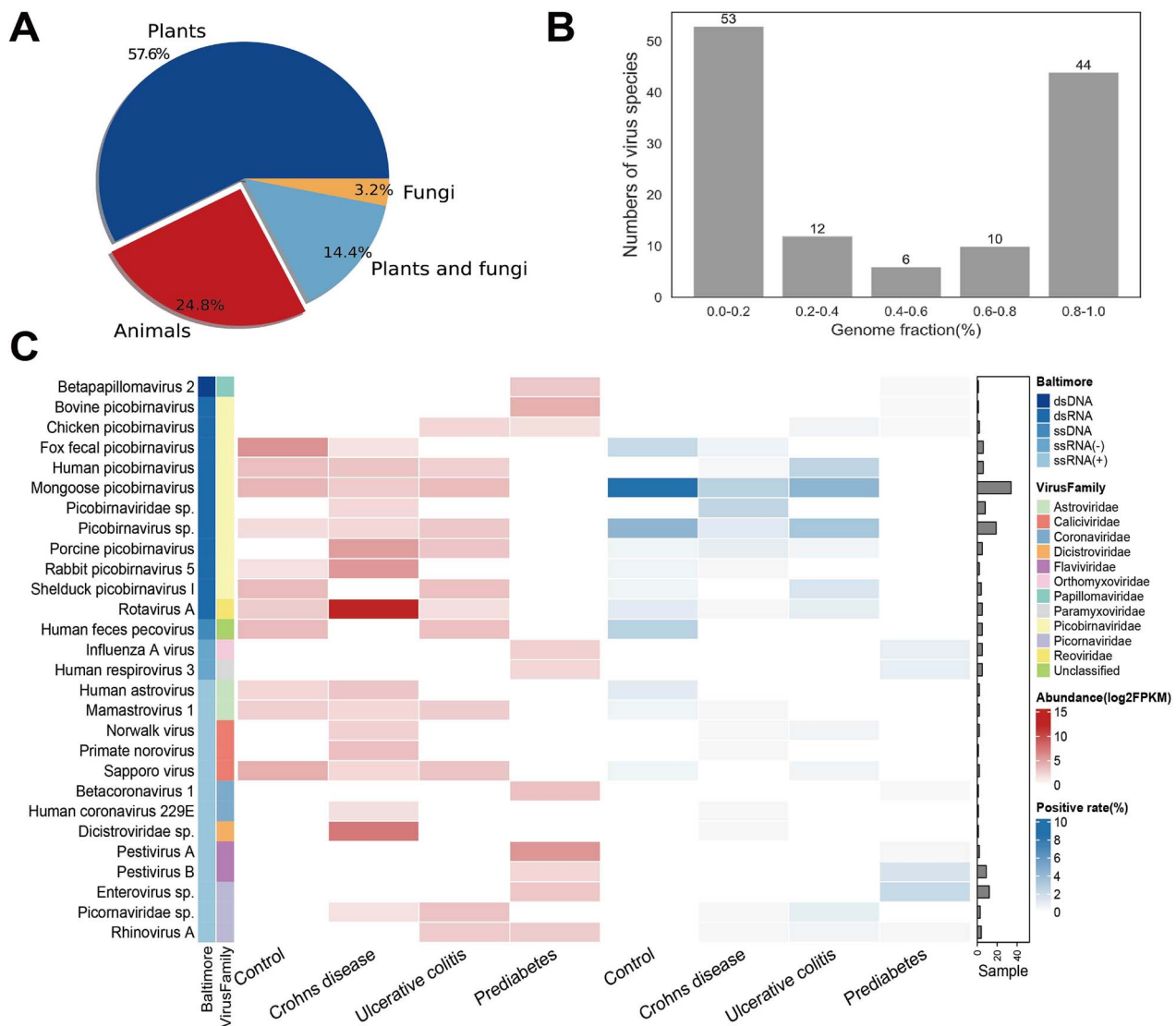
Compared to previous assemblers, VIGA assembled more complete virus genomes in both metatranscriptomic and metagenomic data. This may be because VIGA effectively integrated two reference-based tools, i.e. MetaCompass and RagTag. The reference-based tools have been reported to assemble more accurate and complete genomes than *de novo* ones [51] given high-quality reference genomes. When compared to reference-based tools such as MetaCompass, VIGA automatically

selected reference genomes from the library of virus reference genomes; more importantly, it used RagTag to further assemble virus genome obtained by MetaCompass and to correct potential errors in assembled genomes, which may lead to more complete and accurate virus genomes than those assembled by MetaCompass alone. Even for viruses with extreme genome size or GC content, VIGA also outperformed its competitors in both accuracy and completeness, suggesting its robustness in assembling virus genomes.

Both Haploflow and VirGenA have been designed for strain-resolved assembly of viral genomes. Unexpectedly, VIGA can also be used to assemble virus genomes at the strain level with excellent performances. Compared to Haploflow, VIGA assembled viral genomes with similar precision, higher completeness yet a little lower accuracy on the HIV and HBV datasets. While compared to VirGenA, VIGA assembled more complete viral genomes with higher accuracy. The ability of VIGA on separating mixtures of virus strains may be due to the comprehensive library of virus reference genomes as genome sequences of each virus species were clustered at 100% level. Thus, the library should be updated in time to capture the newly submitted virus reference genomes. However, for novel viruses without reference genomes in the virus reference genome library, it may be difficult to separate and assemble virus genomes at the strain level.

Lots of third-party tools were integrated into VIGA. There were many challenges in integrating such tools effectively. Firstly, tools used in VIGA should be selected seriously in each module. For example, in contig assembly, there were lots of candidate tools such as Trans-ABYSS [21], SPAdes-rna [52] and IVA [53]. Trinity was selected due to its excellent performance, wide usage and ease of use [46]; in virus genome assembly, the reference-based method was selected because a library of virus reference genomes was built. Secondly, the tools should be effectively integrated to make full use of their advantages. For example, in virus genome assembly, two reference-based tools, MetaCompass [22] and RagTag [40, 41], were used in VIGA. MetaCompass was firstly used to assemble virus genomes based on the virus reference genome, which usually resulted in the generation of many virus contigs. Then, RagTag was employed to correct misassemblies, scaffold genome assemblies and fill gaps, leading to a chromosome-scale virus genome and a significant improvement in the integrity of the assembled virus genomes. Thirdly, all tools used in VIGA should be effectively connected to form a unified workflow, which makes it easy for users to use.

There were some limitations to the study. Firstly, VIGA can only assemble genomes for known viruses, making it particularly suitable for identification and assembly of known virus genomes in viral genome re-sequencing. For novel viruses without reference



**Figure 5.** Identification and characterization of viromes in metatranscriptomic data from the HMP dataset. **(A)** Distribution of assembled viruses based on host type. **(B)** Number of viruses with a given level of genome fraction were assembled. For viruses that were identified in multiple samples, only the most complete genome was used in the analysis. The GF was evenly divided into five levels. **(C)** Abundance and positivity rate of animal viruses in three diseases, namely, Crohn's disease, ulcerative colitis and prediabetes and in the control group (healthy people). Viruses were organized by groups in the Baltimore classification system and by the viral family. The virus abundance and the positive rate of viruses are shown in the left and central heatmaps, respectively, according to the figure legend in the top right. The bar plot on the right indicates the number of samples in which the virus was identified.

genomes, it is difficult to assemble accurate genomes by VIGA. Secondly, the accuracy of VIGA in assembling genomes needs further improvements, especially in regions with low abundance or no transcription such as the 3' and 5' ends. Thirdly, VIGA can only be used for eukaryotic viruses that have been studied much and for which a large number of reference genomes are available. Fourthly, VIGA can only be used on paired-end NGS data as MetaCompass only accepts such data as input.

## CONCLUSION

Overall, this study developed a one-stop tool called VIGA for eukaryotic virus identification and genome assembly from the NGS data. It can be used in assembling virus genomes from both metatranscriptomic and metagenomic data and be used in separating virus strains from mixtures. It would help much

in identification and characterization of viromes in future studies.

### Key Points

- VIGA is a one-stop tool for eukaryotic virus identification and genome assembly from next-generation sequencing data.
- VIGA outperformed its competitors in assembling more complete virus genomes, even at the strain level, when evaluated on multiple simulated and real virome datasets.
- VIGA can be used to analyze the metatranscriptomic and metagenomic data.
- VIGA is freely available for public use on the GitHub platform at <https://github.com/viralInformatics/VIGA>.



## SUPPLEMENTARY DATA

Supplementary data are available online at <https://academic.oup.com/bib>.

## ACKNOWLEDGEMENTS

We thank members in PengLab for helpful discussions on the manuscript.

## FUNDING

National Key Plan for Scientific Research and Development of China (2022YFC2303802) and National Natural Science Foundation of China (32170651 and 32370700).

## DATA AVAILABILITY

All data used in the study are available in public databases. The codes for VIGA are publicly available at <https://github.com/viralInformatics/VIGA>.

## REFERENCES

- Jian H, Yi Y, Wang J, et al. Diversity and distribution of viruses inhabiting the deepest ocean on earth. *ISME J* 2021;**15**:3094–110.
- Roux S, Hallam SJ, Woyke T, Sullivan MB. Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. *Elife* 2015;**4**:e08490. <https://doi.org/10.7554/eLife.08490>.
- Cantalupo PG, Pipas JM. Detecting viral sequences in NGS data. *Curr Opin Virol* 2019;**39**:41–8.
- Proctor LM, Creasy HH, Fettweis JM, et al. The Integrative Human Microbiome Project. *Nature* 2019;**569**:641–8.
- Pybus OG, Rambaut A. Evolutionary analysis of the dynamics of viral infectious disease. *Nat Rev Genet* 2009;**10**:540–50.
- Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. *J Mol Biol* 1990;**215**:403–10.
- Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* 2011;**39**:W29–37.
- Moustafa A, Xie C, Kirkness E, et al. The blood DNA virome in 8,000 humans. *PLoS Pathog* 2017;**13**:e1006292.
- Boratyn GM, Thierry-Mieg J, Thierry-Mieg D, et al. Magic-BLAST, an accurate RNA-seq aligner for long and short reads. *BMC Bioinformatics* 2019;**20**:1–19.
- Kuchibhatla DB, Sherman WA, Chung BYW, et al. Powerful sequence similarity search methods and in-depth manual analyses can identify remote homologs in many apparently “orphan” viral proteins. *J Virol* 2014;**88**:10–20.
- Miao Y, Liu F, Hou T, Liu Y. Virtifier: a deep learning-based identifier for viral sequences from metagenomes. *Bioinformatics* 2021;**38**:1216–22.
- Auslander N, Gussow AB, Benler S, et al. Seeker: alignment-free identification of bacteriophage genomes by deep learning. *Nucleic Acids Res* 2020;**48**:e121–1.
- Ren J, Ahlgren NA, Lu YY, et al. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* 2017;**5**:1–20.
- Gregory AC, Zayed AA, Conceição-Neto N, et al. Marine DNA viral macro- and microdiversity from pole to pole. *Cell* 2019;**177**:1109–1123.e14.
- Schackart KE, III, Graham JB, Ponsoero AJ, Hurwitz BL. Evaluation of computational phage detection tools for metagenomic datasets. *Front Microbiol* 2023;**14**:1078760.
- Lu C, Zhang Z, Cai Z, et al. Prokaryotic virus host predictor: a Gaussian model for host prediction of prokaryotic viruses in metagenomics. *BMC Biol* 2021;**19**:5.
- Ajami NJ, Wong MC, Ross MC, et al. Maximal viral information recovery from sequence data using VirMAP. *Nat Commun* 2018;**9**:3205.
- Beaulaurier J, Luo E, Eppley JM, et al. Assembly-free single-molecule sequencing recovers complete virus genomes from natural microbial communities. *Genome Res* 2020;**30**:437–46.
- Warwick-Dugdale J, Solonenko N, Moore K, et al. Long-read viral metagenomics captures abundant and microdiverse viral populations and their niche-defining genomic islands. *PeerJ* 2019;**7**:e6800. <https://doi.org/10.7717/peerj.6800>.
- van Boheemen S, van Rijn AL, Pappas N, et al. Retrospective validation of a metagenomic sequencing protocol for combined detection of RNA and DNA viruses using respiratory samples from pediatric patients. *J Mol Diagn* 2020;**22**:196–207.
- Robertson G, Schein J, Chiu R, et al. De novo assembly and analysis of RNA-seq data. *Nat Methods* 2010;**7**:909–12.
- Victoria C, Liu B, Almeida M, et al. MetaCompass: reference-guided assembly of Metagenomes. *bioRxiv*. 2017;**212506**:1–34. <https://doi.org/10.1101/212506>.
- Fedonin GG, Fantin YS, Favorov AV, et al. VirGenA: a reference-based assembler for variable viral genomes. *Brief Bioinform* 2017;**20**:15–25.
- Grabherr MG, Haas BJ, Yassour M, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 2011;**29**:644–52.
- Fritz A, Bremges A, Deng ZL, et al. Haploflow: strain-resolved de novo assembly of viral genomes. *Genome Biol* 2021;**22**:212.
- Maljkovic Berry I, Melendrez MC, Bishop-Lilly KA, et al. Next generation sequencing and bioinformatics methodologies for infectious disease research and public health: approaches, applications, and considerations for development of laboratory capacity. *J Infect Dis* 2020;**221**:S292–307.
- Kodama Y, Shumway M, Leinonen R, on behalf of the International Nucleotide Sequence Database Collaboration. The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res* 2012;**40**:D54–6.
- Jo Y, Kim SM, Choi H, et al. Sweet potato viromes in eight different geographical regions in Korea and two different cultivars. *Sci Rep* 2020;**10**:2588.
- Shan T, Yang S, Wang H, et al. Virome in the cloaca of wild and breeding birds revealed a diversity of significant viruses. *Microbiome* 2022;**10**:60.
- McNaughton AL, Roberts HE, Bonsall D, et al. Illumina and Nanopore methods for whole genome sequencing of hepatitis B virus (HBV). *Sci Rep* 2019;**9**:7081.
- Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 2018;**34**:i884–90.
- Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015;**12**:59–60.
- Wang J, Pan YF, Yang LF, et al. Individual bat virome analysis reveals co-infection and spillover among bats and virus zoonotic potential. *Nat Commun* 2023;**14**:4079.
- Zhang W, Wang R, Zou X, et al. Comparative genomic analysis of alloherpesviruses: exploring an available genus/species demarcation proposal and method. *Virus Res* 2023;**334**:199163.

35. Luo C, Rodriguez-r LM, Konstantinidis KT. MyTaxa: an advanced taxonomic classifier for genomic and metagenomic sequences. *Nucleic Acids Res* 2014;**42**:e73–3.
36. Zhao R, Gu C, Zou X, et al. Comparative genomic analysis reveals new evidence of genus boundary for family Iridoviridae and explores qualified hallmark genes. *Comput Struct Biotechnol J* 2022;**20**:3493–502.
37. Ji C, Zhang Y, Sun R, et al. Isolation and identification of two clinical strains of the novel genotype enterovirus E5 in China. *Microbiology Spectrum* 2022;**10**:e02662–21.
38. Péntzes JJ, Söderlund-Venermo M, Canuti M, et al. Reorganizing the family Parvoviridae: a revised taxonomy independent of the canonical approach based on host association. *Arch Virol* 2020;**165**:2133–46.
39. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 2017;**35**:1026–8.
40. Alonge M, Lebeigle L, Kirsche M, et al. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol* 2022;**23**:258.
41. Alonge M, Soyk S, Ramakrishnan S, et al. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol* 2019;**20**:1–17.
42. Mikheenko A, Saveliev V, Gurevich A. MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics* 2015;**32**:1088–90.
43. Lee S, Seo CH, Lim B, et al. Accurate quantification of transcriptome from RNA-Seq data by effective length normalization. *Nucleic Acids Res* 2011;**39**:e9–9.
44. Hillier LW, Marth GT, Quinlan AR, et al. Whole-genome sequencing and variant discovery in *C. elegans*. *Nat Methods* 2008;**5**:183–8.
45. Sims D, Sudbery I, Ilott NE, et al. Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet* 2014;**15**:121–32.
46. Bankar KG, Todur VN, Shukla RN, Vasudevan M. Ameliorated de novo transcriptome assembly using Illumina paired end sequence data with Trinity Assembler. *Genom Data* 2015;**5**:352–9.
47. Hölzer M, Marz M. De novo transcriptome assembly: a comprehensive cross-species comparison of short-read RNA-Seq assemblers. *Gigascience* 2019;**8**:giz039.
48. Carroll MW, Haldenby S, Rickett NY, et al. Deep sequencing of RNA from blood and oral swab samples reveals the presence of nucleic acid from a number of pathogens in patients with acute Ebola virus disease and is consistent with bacterial translocation across the gut. *MSphere* 2017;**2**(4): e00325-17 <https://doi.org/10.1128/mspheredirect.00325-00317>.
49. Li D, Li Z, Zhou Z, et al. Direct next-generation sequencing of virus-human mixed samples without pretreatment is favorable to recover virus genome. *Biol Direct* 2016;**11**:1–10.
50. Ye S, Lu C, Qiu Y, et al. An atlas of human viruses provides new insights into diversity and tissue tropism of human viruses. *Bioinformatics* 2022;**38**:3087–93.
51. Ayling M, Clark MD, Leggett RM. New approaches for metagenome assembly with short reads. *Brief Bioinform* 2019;**21**:584–94.
52. Bushmanova E, Antipov D, Lapidus A, Prjibelski AD. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. *GigaScience* 2019;**8**:giz100.
53. Hunt M, Gall A, Ong SH, et al. IVA: accurate de novo assembly of RNA virus genomes. *Bioinformatics* 2015;**31**:2374–6.