Published online DD MM YYYY

Nucleic Acids Research, YYYY, Vol. xx, No. xx **1–8**
doi:10.1093/nar/gkn000

METHODS

coronaSPAdes: from biosynthetic gene clusters to RNA viral assemblies

Dmitry Meleshko [1,2,3], Iman Hajirasouliha [3,4], and Anton Korobeynikov [2,5]*

[1]Tri-Institutional PhD Program in Computational Biology and Medicine, Weill Cornell Medical College, 10021, New York, USA [2]Center for Algorithmic Biotechnology, St. Petersburg State University, 199004, St. Peterburg, Russia and [3]Institute for Computational Biomedicine, Department of Physiology and Biophysics, Weill Cornell Medicine of Cornell University, NY, 10021, USA and [4]Englander Institute for Precision Medicine, The Meyer Cancer Center, Weill Cornell Medicine, NY, 10021, USA and [5]Department of Statistical Modelling, St. Petersburg State University, 198504, St. Peterburg, Russia.

Received YYYY-MM-DD; Revised YYYY-MM-DD; Accepted YYYY-MM-DD

ABSTRACT

The COVID-19 pandemic has ignited a broad scientific interest in coronavirus research. The identification of coronaviral species in natural reservoirs typically involves de novo assembly. However, existing genome, metagenome and transcriptome assemblers often are not able to assemble coronaviruses into a single contig. Coverage variation between datasets and within dataset, presence of close strains, and contamination set a high bar for assemblers to process datasets with diverse properties. We developed coronaSPAdes, a new module of the SPAdes assembler for RNA viral species recovery in general and coronaviruses in particular. coronaSPAdes leverages the knowledge about viral genome structures to improve assembly. We have shown that coronaSPAdes outperforms existing SPAdes modes and other popular short-read and viral assemblers in the recovery of full-length RNA viral genomes.

INTRODUCTION

The COVID-19 pandemic has increased a scientific interest in coronavirus research. The analysis of the coronavirus dataset starts with obtaining full-length virus genome sequence that can be performed using read alignment (1, 2) or de novo assembly (1, 3).

The assembly pipeline based on read alignment is a tool of choice for the same strains of the close species, e.g. for SARS-CoV-2 SNP profiling of confirmed COVID-19 patients. De novo assembly is better suited for novel species recovery since read alignment for distant species is unreliable. Recently, there were multiple studies that used MEGAHIT (4) assembler to recover full-length sequence of the SARS-CoV-2 genome, also previous studies show that different SPAdes (5) modes perform well in virus recovery (6, 7) from complex metagenomes and metaviromes. Nevertheless, none of these assemblers was initially designed for viral assemblies: MEGAHIT and METASPAdes (8) are metagenomic assemblers, SPAdes (9) is designed to assemble single-cell and isolate bacterial datasets, RNASPAdes (10) is intended for accurate isoform separation from eukaryotic data. For RNA viral samples (metaviromes and metatranscriptomes) these assemblers can produce fragmented assemblies due to specific sequencing artifacts, coverage variations, host contamination, multiple strains and quasispecies, and splice events (11). Fast and correct assembly and characterization of viral species is a key step in predicting and preventing the future outbreaks.

Many RNA viruses including coronaviruses have a conserved gene structure (12, 13, 14, 15) that can help to better assemble full-length genomes. In this study, we present CORONASPAdes — a novel assembler designed for RNA viral data. While CORONASPAdes was initially developed having coronaviral species in mind, we demonstrate that overall approach is generic and applicable to assembly of other broad viral families.

We show that CORONASPAdes is able to recover full-length genomes from publicly available datasets where other popular assemblers produce fragmented assembly.

MATERIALS AND METHODS

RNA virus assemblers (16, 17, 18) have to face a number of challenges in order to assemble the sequence data into a consensus sequence. These challenges stem from the nature of the sequencing data due to the biases in the reverse transcription and polymerase chain reaction amplification process that current sequencing methods rely on. These biases are further aggravated by enormous viral population diversity causing lots of SNPs as well as structural variations. Such population diversity is explained by high error rates during the replication process of RNA viruses, essentially they occur as quasispecies (i.e, groups of related genotypes) (19).

These properties of the data cause assembly fragmentation or, even worse, make certain regions disappear from

*To whom correspondence should be addressed. Email: a.korobeynikov@spbu.ru the assembly. Additionally, some RNA viruses (including the species from the order *Nidovirales* that includes coronaviruses) are known to use discontinuous extension of negative strands to produce multiple mRNAs (20). Thus, even in case of a single virus species in the sample, assemblers should be able to deal with multiple produced "isoforms".

Over the years the SPAdes team produced several assembly pipelines aimed for a wide range of sequencing data and tasks. This includes METASPADES (8) for the assembly of consensus bacterial genomes from metagenomes, RNASPADES (10) centering around the reconstruction of multiple isoforms from eukaryotic data as well as more specialized versions such as METAVIRALSPADES (21).

None of these pipelines, however, can handle all the challenges that appear during the assembly of RNA viral data:

1. METASPADES cannot cope with the properties and sequencing artifacts typical for RNA-Seq data (outlined below). While it expects multiple species in the input data, overall it assumes relatively uniform coverage across a single bacterial genome. When applied to RNA viral data with uneven overage and extensive variation, the graph simplification procedures of METASPADES (namely, the rare strain disconnector, see (8) for more details) could confuse the main genome with high-covered variation and therefore fragment the assembly. This phenomenon is especially severe for complex metaviromes (see Results section, assembly of Inluenza and HIV data).

2. While RNASPADES certainly can cope with the specifics of RNA-Seq data, its aim is quite the opposite as required for RNA viral assembly. For RNA viral assembly the main task is to remove possible variation due to quasispecies, strain variation and sequencing artifacts. For RNA transcriptome assemble the aim is to preserve as much variation due to multiple isoforms as possible. This is why the graph simplification procedure of RNASPADES is quite "gentle" (see (10) for more details on the graph simplification procedures) which is further compensated by the isoform restoration procedure. However, the assembly graph of a typical RNA viral dataset (especially a metatranscriptoic / metaviromic one) is much more complex as compared to RNA transcriptome one. As a result, many sequencing artifacts, variation and chimeric connections are still left there which might result in fragmented assemblies, if some part is lost by an accident, or mosaic ones if isoform restoration algorithm would incorrectly resolve variation. Also, we could expect that RNASPADES as other RNA assemblers would certainly inflate the genome duplication ratio as multiple possible arrangements of variation might appear in the output.

3. METAVIRALSPADES pipeline is based on METASPADES and uses coverage-based heuristics in order to detect putative DNA virus sequences (cyclic and linear) from assembly graphs, which fails to work on RNA viral data due to uneven coverage. It does not handle RNA-Seq sequencing artifacts as well.

The outlined issues required us to develop a new assembly graph simplification pipeline that is specifically aimed to take into account the specifics of RNA viral data.

Our new CORONASPADES pipeline consists of two main steps: RNAVIRALSPADES and HMMPathExtension.

RNAVIRALSPADES

RNAVIRALSPADES is a standalone assembler on its own that takes takes a transcriptome, meta-transcriptome, virome or meta-virome dataset on input. RNAVIRALSPADES modifies approaches of METASPADES and RNASPADES in order to assemble RNA viruses on species level.

Removal of low-complexity (poly-A / poly-T) tips and edges and RNA-seq specific chimeric connections. Analysis in (10) shows that the majority of the chimeric connections in RNA-Seq data are either single-strand chimeric loops or double-strand hairpins. They are detected by analyzing the graph topology rather than nucleotide sequences or coverage.

Another characteristic of RNA-Seq datasets is the large number of low-complexity regions that originate from poly-A tails resulting from polyadenylation at the ends of mRNAs. To avoid chimeric connections and non-informative sequences low-complexity edges are removed from the de Bruijn graph.

Transcriptome and metatranscriptome datasets could be quite large and input reads often contain billions of distinct k-mers. Therefore RNAVIRALSPADES implements removal of low-complexity tips and length 1 edges (by default tips shorter than 200 k-mers and having A/T content more than 80% are removed) on both uncondensed and condensed de Bruijn graph. Early removal of large portion of sequencing artifacts before condensing the edges of de Bruijn graph helps to keep the memory consumption low and reduces the running time of further graph cleaning steps as well.

Preventing gap closure by low-complexity overlaps. There are several approaches that helps to assemble the regions of low coverage. Using multiple k-mers in iterative manner is one of them. Another one is the *gap closure* process: paired-end reads are aligned to the tips (and their neighborhoods) of the graph. And if there are enough paired-end reads that span the gap, then the ends of tips are analyzed for a possible overlap that is shorter than k-mer. If there exists an exact overlap that is longer than 10 bp, then the tips are joined into a single edge at this overlap.

It turned out that the majority of such overlaps for RNA data are again low-complexity sequences containing long stretches of "A"s or "T" with few mismatches. Almost all these overlaps are spurious and therefore the produced connection would be chimeric. We modified gap closure algorithm was to ignore such overlaps.

Collapsing of quasispecies. RNAVIRALSPADES aims for species-level assemblies (as opposed to strain-level assemblies that are certainly infeasible due to high level of variation), therefore the bulge removal procedure was refined to collapse the variation due to quasispecies. Specifically, RNAVIRALSPADES collapses long and similar (with respect to the edit distance) parallel edges in the assembly graph. By default, it does so for edges shorter than 1000 and similar to each other by more than 90% IDY.

Low-abundant strains disconnector. Unfortunately, strain differences are not only manifested as single nucleotide variations and small insertions or deletions. Such variations (especially for complex datasets containing many species) are caused by highly diverged genome regions, rearrangements, large deletions, parallel gene transfer, etc. Therefore the topology of the de Bruijn graph in the neighborhood of such variations is more complex than a few dozens of bulges complicating the strain variation masking procedure.

metaSPAdes includes a dedicated *edge disconnector algorithm* that uses the coverage ratios between adjacent edges in the assembly graph to identify edges with low coverage ratios as those that most likely originate from rare strains. The algorithm then disconnects such edges from high-covered paths. However, there are important exceptions to this approach, for example, cases when low covered edges are connected to repeats with a high copy number. In order to identify such cases, for each edge e a high-covered subgraph connected to e is constructed. In the case of repetitive region, we expect to find a component with a total sum of edge length being small.

This repeat-preserving heuristics does not work well for RNA viral datasets due to drastically increased coverage variation compared with metagenomic samples leaving many connections intact. Also, we certainly assume viral genomes to have small genome size and not include high-covered repeats (so all repeats must be intra-species).

In rnaviralSPAdes we use the disconnector algorithm without repeat-preserving heuristics and more conservative thresholds with respect to edge coverage ratio to take into account coverage variation.

Generic assembly graph simplification pipeline. Besides the important changes outlined above, rnaviralSPAdes implements the graph simplification approach similar to consensus assembly graph construction pipeline of metaSPAdes.

As a result, the rnaviralSPAdes pipeline alone allows for removal of the majority of RNA sequencing artifacts and collapsing the variation. However, still there might be very ambiguous cases when an assembler could not remove the errors neither using the coverage-based heuristics nor the graph topology. The Ginger dataset outlined above is a good example. Certainly, it might be possible to tune various assembler heuristics to deal with that particular case, however, the solution will unlikely work on other datasets as it will be overly-aggressive in error elimination. As a result, some different approach is necessary. RNA transcriptome assemblers solve the problem with isoform reconstruction step, potentially inflating the genome duplication ratio. coronaSPAdes instead uses a HMMPathExtension algorithm.

HMMPathExtension

The second step of the coronaSPAdes pipeline, HMMPathExtension, utilizes the information about viral genome organization to distinguish between putative genomic sequences from uncleaned artifacts.

HMMPathExtension is inspired by HMM-based algorithms of biosyntheticSPAdes (22). It takes a set of HMMs and the assembly graph as input. First, HMMPathExtension aligns HMMs to assembly graph and constructs a domain graph (for details of domain graph construction refer to (22)). In order to construct domain graph for an arbitrary set of HMMs, we had to add improvements to domain graph construction algorithm. Previous algorithm assumed that profile HMMs are longer than k used during de Bruijn graph construction and can not match to the same interval of the de Bruijn graph. In the current version, if profile HMM is k or shorter, it matches to the single kmer on the de Bruijn graph with prefix that have the best-scored match.

biosyntheticSPAdes guarantees that there are no positions on the de Bruijn graph that are matched more than once. This fact arouses from a distinct nature of biosynthetic gene cluster domains. However, the unprecedented diversity of viral data and building blocks of viral genomes, HMM hits can overlap, causing initial domain graph construction algorithm to fail. In order to overcome this problem, we greedily select an arbitrary HMM hit and remove all other hits that overlap with it until there are no overlapping hits. This procedure guarantees that strong and weak edges of the domain graph will be correctly added to the graph.

Similar to biosyntheticSPAdes, HMMPathExtension aims to find paths through the domain graph that traverse significant HMM matches in order and translate them to the assembly graph paths. This way the extracted genomic sequence is supported both by the graph topology and the structure of the genome. The only major difference is that biosyntheticSPAdes assumes that different gene clusters do not overlap in the assembly graph, but in case of viruses, multiple virus strains can be easily presented. Unlike biosyntheticSPAdes, HMMPathExtension does not require to thread through all close matches in a connected component of the domain graph. It allows to reconstruct multiple virus sequences of the same family. All paths produced by HMMPathExtension algorithm are maximal by inclusion, it allows to ignore non-viral sequencing artifacts since they should not have any viral matches on them.

For coronavirus assemblies coronaSPAdes is bundled with the set of HMMs obtained from Pfam SARS-CoV-2 (23) (despite the name, these HMMs are quite general and represent the profiles of various proteins that belong to coronaviruses as well as more conserved ones like RNA dependent RNA polymerase (RdRp) that is conserved across all RNA viruses (24)) and HMMs for coronaviral protein families studied in (25).

We explicitly note that the approach of HMMPathExtension is not limited to coronaviral genomes. The HMMPathExtension step allows for a custom HMM database specification effectively enabling HMM-guided assemblies of other genomes using their internal structure. In the result section, we demonstrate how coronaSPAdes can be used to assemble HIV, influenza and CoV genome sequences from diverse datasets. HMM sets used for these assemblies are also available to use.

HMMPathExtension takes advantage from hits that uniformly cover the genome of interest, allowing to reconstruct strains mixtures and splice variants.

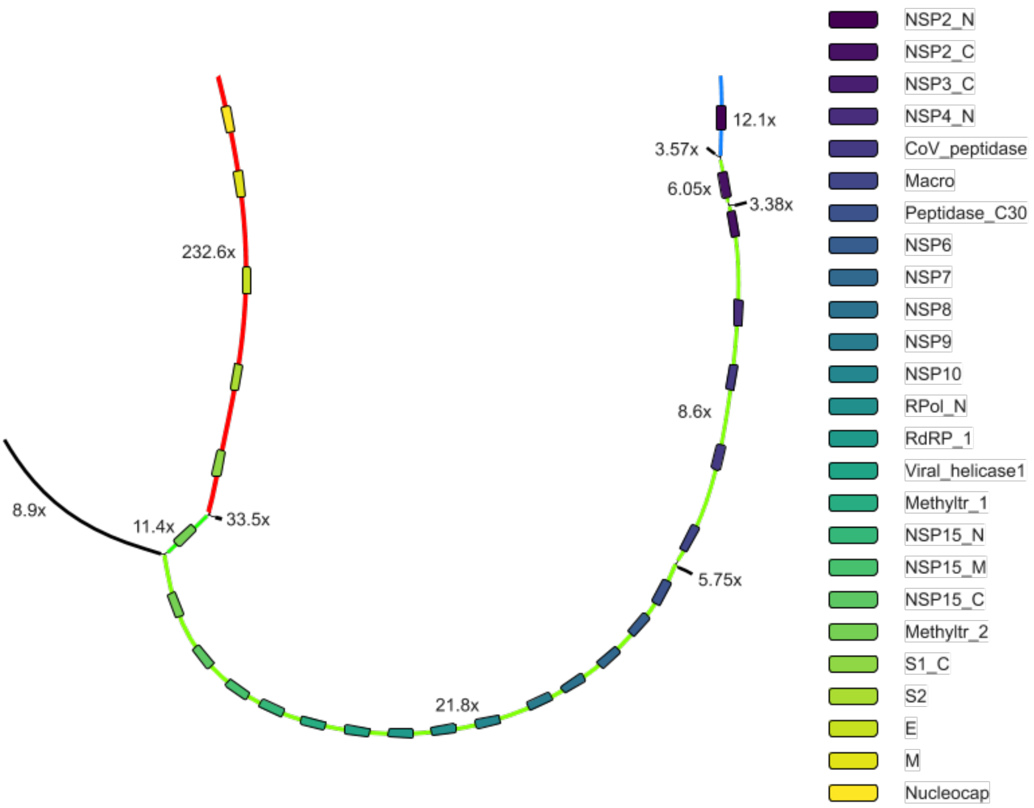

Figure 1. Part of GINGER assembly graph produced with CORONASPADES. RNAVIRALSPADES produced 8 contigs from this subgraph (red, green and blue paths on the graph and 5 black edges), therefore splitting the coronavirus genome into three parts. CORONASPADES matched viral edges of the graph with domain (rectangles of different color). Path along these matches spells a full-length viral genome

RESULTS

We highlight the features of CORONASPADES using a wide range of publicly available transcriptome and metatranscriptome datasets that include novel and known coronaviral species. We also show how CORONASPADES could be used for not only coronaviral assemblies by reproducing Influenza and HIV assembly benchmark from (16).

Coronaviral assemblies

FR4NK is a putative novel Alphacoronavirus detected in a metatranscriptome sequencing library from a Peruvian vampire bat (*Desmodus rotundus*, SRA:ERR2756788). The average coverage is 110x with variation from 2x to 1500x in different regions of the genome. This is a new species of Coronavirus based on RdRP, nucleoprotein, membrane protein and replicase 1a, which all classify this virus an Alphacoronavirus outside of all named sub-genera and most similar to a Pedacovirus.

GINGER is a putative novel Alphacoronavirus detected in a transcriptome sequencing library from a Wildcat (*Felis silvestris*, SRA:SRR72871109). The average coverage is 20x with variation from 230x to 6x in different parts of the genome.

PEDV is a known Alphacoronavirus that causes porcine epidemic diarrhea. It was assembled from a transcriptome sequencing library of epithelial cells of pig intestine (*Sus scrofa*, SRA:SRR10829957). The average coverage is 470x with variation from 30x to 8000x in different parts of the genome.

We assemble FR4NK, GINGER and PEDV using several specialized virus assemblers (IVA, PRICE), generic metagenome and transcriptome assemblers (MEGAHIT, METASPADES, RNASPADES, TRINITY (26)) and CORONASPADES.

The overview of the results could be found in Table 1. Conventional RNA assemblers (IVA and PRICE) are using a seed-and-extend approach, therefore a seed sequence was required. This property greatly reduces their applicability for novel species search. In addition, it seems they were unable to deal with the specifics of large transcriptome and metatranscriptome datasets. Other assemblers (MEGAHIT, TRINITY, METASPADES, RNASPADES) overall have shown acceptable results, however their performance was not uniform, as none of them was able to assemble complete virus genomes out of all 3 datasets. CORONASPADES was able to produce whole genomes in all cases.

HIV and Influenza assemblies

As it was mentioned previously, the HMMPathExtend approach is generic and could be applied to other viral families should the desired set of viral proteins is provided. To showcase this feature the re-create the

Table 1. Benchmarking of assemblers on several CoV datasets

		MEGAHIT	Trinity	metaSPAdes	rnaSPAdes	rnaviralSPAdes	coronaSPAdes	IVA	PRICE
Fr4nk	Longest	**29219**	**29168**	20435	20606	**29117**	**29117**	N/A**,ˆ	8504*
	CheckV completeness	105.2	105.39	73.62	72.63	101.51	101.51	N/A	31.1
	CheckV AA avg ID%	55.71	55.79	58.82	63.39	54.84	54.84	N/A	75.8
Ginger	Longest	26905	**29301**	19453	**29301**	18897	**29109**	3186**	15691*
	CheckV completeness	95.1	103.58	69.0	103.66	67.20	103.5	N/A	55.9
	CheckV AA avg ID%	85.3	85.85	93.93	85.58	93.92	85.58	N/A	95.73
PEDV	Longest	26185	**31054**; 25634 ***	26316	24682	**27973**	**27973**	N/A**,ˆ	23559*
	CheckV completeness	93.1	110.46; 91.142	93.6	82.93	99.5	99.5	N/A	83.02
	CheckV AA avg ID%	97.65	97.55; 97.94	97.96	97.92	99.5	97.5	N/A	97.9

Shown are: longest viral contig assembled, its completeness as estimated by CheckV (27) and the average amino acid identity to the closest reference in CheckV database. The best results are shown in bold.
*Seed was required, 1 kbp of coronaSPAdes assembly was used
**IVA failed to select seed automatically. 1 kbp of coronaSPAdes assembly was provided as a seed
*** Two isoforms were reported, one seems too long and likely a misassembly, second one is shorter than expected genome size.
ˆFailed to extend the seed

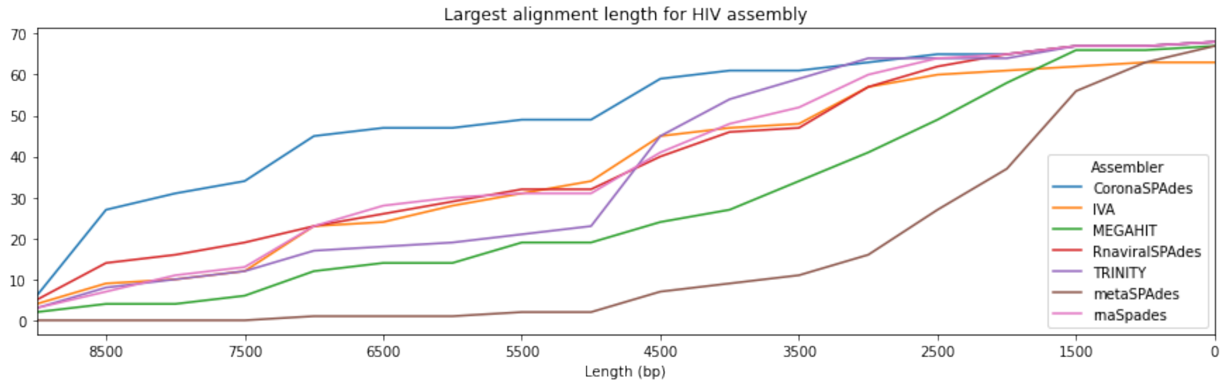

Figure 2. Performance of different assemblers on HIV datasets. Y-axis represents number of datasets which have alignment of such length or greater, similarly to a widely adopted NAx plot.

benchmarking analysis from IVA paper (16): we evaluated IVA, Trinity, MEGAHIT, rnaSPAdes, metaSPAdes, rnaviralSPAdes and coronaSPAdes on Illumina paired reads from 68 human immunodeficiency virus 1 (HIV-1) and 172 Influenza A and B virus samples. For this benchmark, coronaSPAdes used the set of HIV and Influenza HMMs extracted from U-RVDB-prot v20 database (28).

Figure 2 shows that coronaSPAdes significantly outperforms other tools in terms of assembly contiguity with 30 complete and near-complete assemblies ($>$ 8500 bp). All other assemblers have no more than 14 complete or near-complete assemblies. The number of misassemblies is another important metric, that shows assembly quality. coronaSPAdes keeps misassemblies at a relatively low level (see Figure 4), providing a good contiguity-quality trade-off. Also, these results clearly show that metagenomic assemblers might produce suboptimal results and therefore are not suitable for RNA viral assembly from metaviromes.

Influenza assembly is more complicated because influenza type A, B genomes consist of eight segments, that can have highly similar regions at the segment's ends. As a metric for the assembly contiguity, we sum the longest alignment length across all segments. Figure reffig:inf shows that rnaSPAdes has the best contiguity performance, with coronaSPAdes and Trinity at the second place. However, Figure 4 shows that MEGAHIT, rnaSPAdes, and Trinity have the worst misassembly statistics (73.0, 19.82, and 7.162 misassemblies per dataset correspondingly), while coronaSPAdes has 2.895 misassemblies per dataset. Therefore coronaSPAdes has a reasonable contiguity-correctness trade-off.

Raw assembly results are available in Supplementary Tables 1 (for Influenza datasets) and 2 (for HIV datasets).

Serratus

coronaSPAdes was used in the Serratus project (29) for a widespread search of novel CoV and CoV-like species from public sequencing libraries. From a screen of 3.8 million libraries comprising 5.6 petabases of sequencing reads, there were reported 11,120 assemblies, including sequences from 13 previously uncharacterized or unavailable CoV or CoV-like operational taxonomic units (OTUs), defined by clustering amino sequences of the RdRp gene at 97% identity. 8 of these OTUs were designated to a putative novel genus of

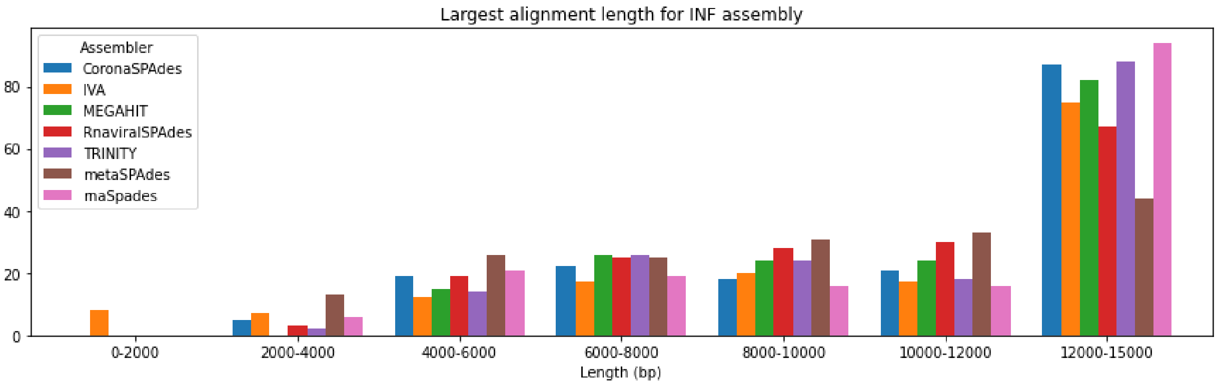

Figure 3. Performance of different assemblers on influenza datasets. Y-axis represents number of datasets which have alignment of such length.

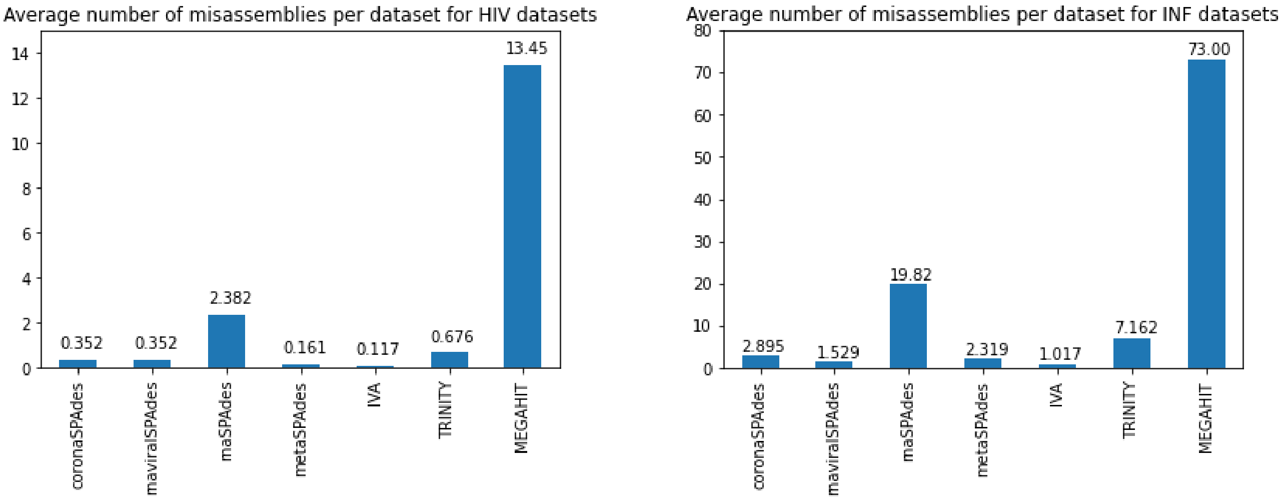

Figure 4. Misassembly statistics for the HIV and influenza datasets.
Left: Average number of misassemblies per dataset for HIV datasets. **Right**: Average number of misassemblies per dataset for influenza datasets.

coronaviruses, noting that all were found in samples from non-mammal aquatic vertebrates falling outside deltacoronaviruses genus.

DISCUSSION

GINGER dataset represents the typical case when long sequencing artifacts could influence the assembly results. All metagenome assembles lost some parts of the genome likely being unable to remove the long artifacts having coverage similar to the virus genome (see Figure 1 for assembly graph). RNA assemblers (RNASPADES and TRINITY) solved this problem via isoform restoration steps: different “isoform” paths across this subgraph were produced and one of them was a full-length viral genome. The downside of this approach is increased genome duplication ratio as multiple paths through the subgraph are produced and in more complex cases some of them might be mosaic. CORONASPADES traversed all domain matches in order and also produced a full-length viral genome.

The influenza A and B virus genomes each comprise eight negative-sense, single-stranded viral RNA segments that code 10–14 proteins, depending on the strain. Each segment possesses noncoding regions, of varying lengths, at both 3’and 5’ ends. However, the extreme ends of all segments are highly conserved among all influenza virus segments (30).

As a result, depending on the particular strain, the segments could appear glued in the assembly graph (see Figure 5), also the Influenza genomes are highly variable with many rearrangements in at least 2 segments due to the antigenic drift and shift processes (31). The challenge for an assembler here is to correctly recover the sequences of all eight segments taking into account possible variation.

Surprisingly, MEGAHIT for some unknown reason often produced the contigs of 4-8 Kbp that contained parts of multiple segments. This results in a highly elevated rate of misassemblies seen there. The results of other assemblers are overall expected as well: both TRINITY and RNASPADES shown good results in recovery of full-length segments using the isoform reconstruction procedures. However, some segments were clearly mosaic as could be seen from the Figure 4. Seed-and-extend approach of IVA resulted in the most accurate in terms of the average number of misassemblies assemblies, albeit at the expense of the recovery of the fuller segments. CORONASPADES was able

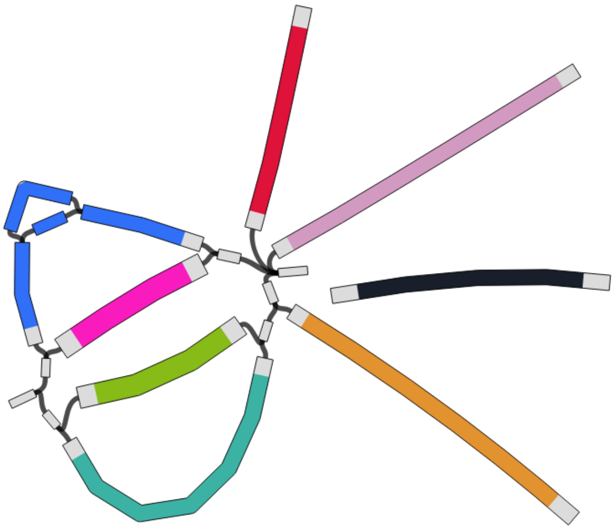

Figure 5. Typical assembly graph of Influenza A dataset. Different CDS in segments are color-coded. There is a variation in HA (blue) segment.

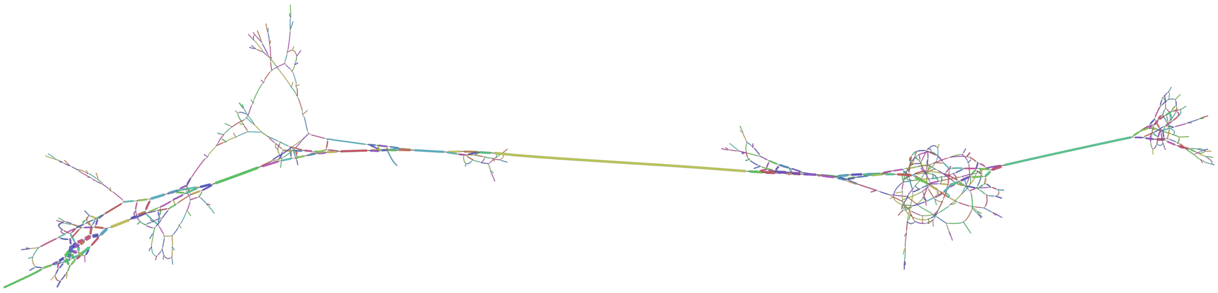

Figure 6. Typical assembly graph of a HIV dataset. Graph consists of 856 edges of 130 kbp total length with 122 dead-end edges.

to recover much more still having an acceptable misassembly rate.

HIV genome represents a true nightmare from the assembly standpoint as it employs a very complex system of differential RNA splicing to obtain more than 30 mRNA species from a less than 10 kb genome (32). This results in a very complex and tangled assembly graph (see Figure 6) that an assembler must traverse in order to recover the complete virus genome. Here the HMM-guided approach of CORONASPADES clearly allows to extend the contigs and recover significantly fuller genomes as compared to other tools.

CONCLUSION

Clearly, assembling RNA viral genomes is very challenging (16). The variety of possible kinds of input data and the overall diversity of the species multiplies these challenges even more. We demonstrated that additional information about the genome structure could significantly improve the viral genome recovery even from very complex datasets and therefore catalyze the new viral discoveries.

DATA AVAILABILITY

The datasets supporting the conclusions of this article are available in the NCBI SRA repository, under accessions ERR2756788 (FR4NK), SRR72871109 (GINGER) and SRR10829957 (PEDV). Access instructions for Serratus data can be found at https://github.com/ababaian/serratus/wiki/Access-Data-Release. Influenza and HIV-1 data were taken from IVA paper (16) (lists of accessions are available from Supplementary Tables 1 and 2). CORONASPADES version used in this article is a part of SPAdes 3.15 release and is available at http://cab.spbu.ru/software/spades and also deposited on Zenodo under doi:10.5281/zenodo.4438269. HIV and Influenza HMMs were extracted from U-RVDB-prot v20 database and are available from http://cab.spbu.ru/software/coronaspades.

ACKNOWLEDGEMENTS

This work was supported by the Russian Science Foundation (grant 19-14-00172). The research was carried out in part by computational resources provided by the Resource Center "Computer Center of SPbU". The authors are grateful to Saint Petersburg State University for the overall support of this work. DM is grateful to T32 Weill Cornell Tri-Institutional Training Program in Computational Biology and Medicine. IH is supported by a Maximizing Investigators' Research Award (MIRA) for Early Stage Investigators (R35 GM138152-01).

Conflict of interest statement. None declared.

REFERENCES

1. Sah, R., Rodriguez-Morales, A. J., Jha, R., Chu, D. K. W., Gu, H., Peiris, M., Bastola, A., Lal, B. K., Ojha, H. C., Rabaan, A. A., Zambrano, L. I., Costello, A., Morita, K., Pandey, B. D., and Poon, L. L. M. (2020) Complete Genome Sequence of a 2019 Novel Coronavirus (SARS-CoV-2) Strain Isolated in Nepal. *Microbiology Resource Announcements,* **9**(11).
2. Yin, C. (2020) Genotyping coronavirus SARS-CoV-2: methods and implications. *Genomics,* **112**(5), 3588 – 3596.
3. Kim, J.-M., Chung, Y.-S., Jo, H. J., Lee, N.-J., Kim, M. S., Woo, S. H., Park, S., Kim, J. W., Kim, H. M., and Han, M.-G. (February, 2020) Identification of Coronavirus Isolated from a Patient in Korea with COVID-19. *Osong Public Health and Research Perspectives,* **11**(1), 3–7.
4. Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (01, 2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics,* **31**(10), 1674–1676.
5. Nurk, S., Bankevich, A., Antipov, D., Gurevich, A. A., Korobeynikov, A., Lapidus, A., Prjibelski, A. D., Pyshkin, A., Sirotkin, A., Sirotkin, Y., Stepanauskas, R., Clingenpeel, S. R., Woyke, T., Mclean, J. S., Lasken, R., Tesler, G., Alekseyev, M. A., and Pevzner, P. A. (2013) Assembling Single-Cell Genomes and Mini-Metagenomes From Chimeric MDA Products. *Journal of Computational Biology,* **20**(10), 714–737 PMID: 24093227.
6. Sutton, T. D. S., Clooney, A. G., Ryan, F. J., Ross, R. P., and Hill, C. (Jan, 2019) Choice of assembly software has a critical impact on virome characterisation. *Microbiome,* **7**(1), 12.
7. Roux, S., Emerson, J. B., Eloe-Fadrosh, E. A., and Sullivan, M. B. (September, 2017) Benchmarking viromics: an in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ,* **5**, e3817.
8. Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P. A. (2017) metaSPAdes: a new versatile metagenomic assembler. *Genome Research,* **27**(5), 824–834.
9. Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., and Korobeynikov, A. (June, 2020) Using SPAdes De Novo Assembler. *Current Protocols in Bioinformatics,* **70**(1).
10. Bushmanova, E., Antipov, D., Lapidus, A., and Prjibelski, A. D. (09, 2019) rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. *GigaScience,* **8**(9) giz100.
11. Viehweger, A., Krautwurst, S., Lamkiewicz, K., Madhugiri, R., Ziebuhr, J., Hölzer, M., and Marz, M. (August, 2019) Direct RNA nanopore sequencing of full-length coronavirus genomes provides novel insights into structural variants and enables modification analysis. *Genome Research,* **29**(9), 1545–1554.
12. Masters, P. S. (2006) The Molecular Biology of Coronaviruses. In *Advances in Virus Research* pp. 193–292 Elsevier.
13. Harrach, B. (2014) Adenoviruses: General Features. In *Reference Module in Biomedical Sciences* Elsevier.
14. Dadonaite, B., Gilbertson, B., Knight, M. L., Trifkovic, S., Rockman, S., Laederach, A., Brown, L. E., Fodor, E., and Bauer, D. L. V. (July, 2019) The structure of the influenza A virus genome. *Nature Microbiology,* **4**(11), 1781–1789.
15. Watts, J. M., Dang, K. K., Gorelick, R. J., Leonard, C. W., Jr, J. W. B., Swanstrom, R., Burch, C. L., and Weeks, K. M. (August, 2009) Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature,* **460**(7256), 711–716.
16. Hunt, M., Gall, A., Ong, S. H., Brener, J., Ferns, B., Goulder, P., Nastouli, E., Keane, J. A., Kellam, P., and Otto, T. D. (02, 2015) IVA: accurate de novo assembly of RNA virus genomes. *Bioinformatics,* **31**(14), 2374–2376.
17. Ruby, J. G., Bellare, P., and DeRisi, J. L. (March, 2013) PRICE: Software for the Targeted Assembly of Components of (Meta) Genomic Sequence Data. *G3: Genes, Genomes, Genetics,* **3**(5), 865–880.
18. Yang, X., Charlebois, P., Gnerre, S., Coole, M. G., Lennon, N. J., Levin, J. Z., Qu, J., Ryan, E. M., Zody, M. C., and Henn, M. R. (2012) De novo assembly of highly diverse viral populations. *BMC Genomics,* **13**(1), 475.
19. Denison, M. R., Graham, R. L., Donaldson, E. F., Eckerle, L. D., and Baric, R. S. (March, 2011) Coronaviruses. *RNA Biology,* **8**(2), 270–279.
20. Sawicki, S. G. and Sawicki, D. L. Coronaviruses use Discontinuous Extension for Synthesis of Subgenome-Length Negative Strands pp. 499–506 Springer US Boston, MA (1995).
21. Antipov, D., Raiko, M., Lapidus, A., and Pevzner, P. A. (05, 2020) MetaviralSPAdes: assembly of viruses from metagenomic data. *Bioinformatics,* btaa490.
22. Meleshko, D., Mohimani, H., Tracanna, V., Hajirasouliha, I., Medema, M. H., Korobeynikov, A., and Pevzner, P. A. (2019) BiosyntheticSPAdes: reconstructing biosynthetic gene clusters from assembly graphs. *Genome Research,* **29**(8), 1352–1362.
23. El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C., and Finn, R. D. (10, 2018) The Pfam protein families database in 2019. *Nucleic Acids Research,* **47**(D1), D427–D432.
24. Venkataraman, S., Prasad, B., and Selvarajan, R. (February, 2018) RNA Dependent RNA Polymerases: Insights from Structure, Function and Evolution. *Viruses,* **10**(2), 76.
25. Phan, M. V. T., Tri, T. N., Anh, P. H., Baker, S., Kellam, P., and Cotten, M. (July, 2018) Identification and characterization of Coronaviridae genomes from Vietnamese bats and rats based on conserved protein domains. *Virus Evolution,* **4**(2).
26. Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., and Regev, A. (May, 2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology,* **29**(7), 644–652.
27. Nayfach, S., Camargo, A. P., Eloe-Fadrosh, E., Roux, S., and Kyrpides, N. (May, 2020) CheckV: assessing the quality of metagenome-assembled viral genomes. *bioRxiv,*.
28. Bigot, T., Temmam, S., Pérot, P., and Eloit, M. (September, 2020) RVDB-prot, a reference viral protein database and its HMM profiles [version 2; peer review: 2 approved].. *F1000Research,* **8**, 530.
29. Edgar, R. C., Taylor, J., Altman, T., Barbera, P., Meleshko, D., Lin, V., Lohr, D., Novakovsky, G., Al-Shayeb, B., Banfield, J. F., Korobeynikov, A., Chikhi, R., and Babaian, A. (2020) Petabase-scale sequence alignment catalyses viral discovery. *bioRxiv,*.
30. Bouvier, N. M. and Palese, P. (September, 2008) The biology of influenza viruses. *Vaccine,* **26**, D49–D53.
31. Webster, R. G. and Govorkova, E. A. (May, 2014) Continuing challenges in influenza. *Annals of the New York Academy of Sciences,* **1323**(1), 115–139.
32. Schwartz, S., Felber, B. K., Benko, D. M., Fenyö, E. M., and Pavlakis, G. N. (1990) Cloning and functional analysis of multiply spliced mRNA species of human immunodeficiency virus type 1.. *Journal of Virology,* **64**(6), 2519–2529.