bioRxiv preprint doi: https://doi.org/10.1101/2020.07.28.224584; this version posted July 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

CoronaSPAdes: from biosynthetic gene clusters to coronaviral assemblies

Dmitry Meleshko (d.meleshko@spbu.ru)[1,2], Anton Korobeynikov (a.korobeynikov@spbu.ru)[1,3*]

[1]Center for Algorithmic Biotechnology, Institute for Translational Biomedicine, St. Petersburg State University, St. Petersburg, Russia 199004;

[2]Tri-Institutional PhD Program in Computational Biology and Medicine, Weill Cornell Medical College, New York, New York 10021, USA;

[3]Department of Statistical Modelling, St. Petersburg State University, St. Petersburg, Russia 198504;

*Corresponding author

Abstract

COVID-19 pandemy has ignited a broad scientific interest in coronavirus research. The identification of coronaviridae species in natural reservoirs often requires de novo assembly. However, existing transcriptome assemblers often are not able to assemble coronaviruses into a single contig. We developed coronaSPAdes, a new module for SPAdes assembler for coronavirus species recovery. coronaSPAdes uses the knowledge about coronaviridae genome structure to improve assembly. We have shown that coronaSPAdes outperforms existing SPAdes modes and other popular short-read assemblers in the recovery of full-length coronavirus genomes. This should allow to better understand the coronaviridae spread and diversity.

Keywords

Genome assembly, transcriptome assembly, COVID-19, coronaviridae, hidden markov models

Background

COVID-19 pandemy has increased a scientific interest in coronavirus research. The analysis of 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. Though previous studies show that different SPAdes[5] modes also perform well in virus recovery[6]. Nevertheless, none of these assemblers was initially designed for viral assemblies in general and for coronaviridae species recovery in particular. MEGAHIT and metaSPAdes[7] are metagenomic assemblers, SPAdes[8] is designed to assemble single-cell and isolate bacterial datasets. All these assemblers can produce fragmented assemblies due to sequencing artifacts, coverage variations, host contamination, multiple strains presence, and coronavirus splice events[9]. Fast and correct characterization of virus datasets might be a key step in predicting and preventing the future outbreaks. Coronaviruses have a conserved gene structure[10] that can help to better assemble full-length genomes. In this study, we present coronaSPAdes - a new mode for SPAdes assembler designed to assemble coronaviridae species.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.28.224584; this version posted July 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

coronaSPAdes borrows its algorithmic ideas from the existing SPAdes[11] modes (metaSPAdes[7], rnaSPAdes[12], metaviralSPAdes[13]), and includes HMM-guided assembly inspired by biosyntheticSPAdes[14]. We show that coronaSPAdes is able to recover novel full-length genomes from publicly available datasets where other popular assemblers produce fragmented assembly.

Results and Discussion

coronaSPAdes was used by the Serratus project [15] to assemble more than 10,000 putative coronaviral genomes out of 3.6 million SRA datasets. The Serratus benchmarking and results will be provided elsewhere. Here we highlight the features of coronaSPAdes using the range of publicly available transcriptome and metatranscriptome datasets that include novel and known species.

Fr4nk is a putative novel Alphocoronavirus detected in a metatranscriptome sequencing library from a Peruvian vampire bat (*Desmodus rotundus*, SRA:ERR2756788). Assembly of this sequencing library with coronaSPAdes yielded a 29264 nt viral genome. 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). Assembly of this sequencing library with coronaSPAdes yielded a 29277 nt viral genome (see Fig. 1). The average coverage is 20x with variation from 230x to 6x in different parts of the genome.

[insert Figure 1 around here]

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:SRR10829953). Assembly of this sequencing library with coronaSPAdes resulted in a 27973 nt viral genome. The average coverage is 470x with variation from 30x to 8000x in different parts of the genome.

We benchmark **Fr4nk**, **Ginger** and **PEDV** using several specialized virus assemblers (IVA, PRICE), generic metagenome and transcriptome assemblers (MEGAHIT, metaSPAdes, rnaSPAdes) 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 nover species search. In addition, it seems they were unable to deal with the specifics of large transcriptome and metatranscriptome datasets. Other assemblers (MEGAHIT, 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.

[insert Table 1 around here]

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.28.224584; this version posted July 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Conclusions

In summary, we developed coronaSPAdes, a specialized assembler for coronaviridae genomes. coronaSPAdes allows to recover coronaviridae genomes with lower fragmentation and higher contiguity from the datasets of different nature. Moreover, our results prove the utility of a HMM-guided assembly approach that can be adapted to other genome types.

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. These properties of the data cause assembly fragmentation or, even worse, make certain regions disappear from the assembly. Additionally, coronaviruses are known to use discontinuous extension of negative strands to produce multiple mRNAs [19]. So, even in case of a single virus in the sample, assemblers should be able to deal with multiple produced “isoforms”.

Over the years SPAdes team produced several assembly pipelines aimed for different kinds of sequencing data and tasks. This includes metaSPAdes for assembly of consensus genomes from metagenomes, rnaSPAdes centered around reconstruction of multiple isoforms from eukaryotic data as well as more specialized versions such as metaplasmidSPAdes and metaviralSPAdes. The latter pipeline uses coverage-based heuristics in order to detect putative DNA virus sequences (cyclic and linear) from assembly graphs.

It turned out that none of these pipelines could cope with all the challenges connected with RNA viral data, however, each of it contains some useful algorithms and approaches. The modular structure of SPAdes allowed deep reuse and extension of the existing algorithms to be combined into a coronaSPAdes pipeline.

Essentially, the coronaSPAdes pipeline consists of two main steps: rnaviralSPAdes and HMMPathExtension.

rnaviralSPAdes constructs assembly graph from input RNA viral dataset (transcriptome, meta-transcriptome, virome and meta-virome datasets are expected and supported). Briefly, it is based on metaSPAdes pipeline with several important additions and changes adopted from other SPAdes pipelines:

1. Removal of low-complexity (poly-A / poly-T) tips and edges and RNA-seq specific chimeric connections. rnaSPAdes introduced extensive procedures to remove from the de Bruijn graph artifacts that are specific to RNA data. As stated in [12] 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.
2. Removal of subspecies-bases variation. As rnaviralSPAdes like metaviralSPAdes aims for species-level assemblies (as opposed to strain-level assemblies that are certainly infeasible due to high level of variation), the bulge removal procedure was refined to collapse this variation. Specifically, it collapses long and similar (with respect to the edit distance) parallel edges in the assembly graph.
3. Additionally, we relaxed the metagenomic edge disconnector condition (see [7] for more info) due to coronavirus genome size and drastically increased coverage variation compared with metagenomic samples.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.28.224584; this version posted July 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

rnaviralSPAdes pipeline allows for removal of the majority of RNA sequencing artifacts and collapsing the variation. However, still there are cases when an assembler could not remove the errors neither using the coverage-based heuristic 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 this particular case, however, the solution will unlikely work on other datasets and would require extensive benchmarking in order not to be overly-aggressive in error elimination.

The second step of coronaSPAdes pipeline, HMMtraversal, deals with outlined problems in a completely different way and tries to use the information about coronaviral genome organization to distinguish between putative genomic sequences from uncleaned artifacts.

HMMPathExtension extends HMM-based algorithms of biosyntheticSPAdes. Similarly with biosyntheticSPAdes, HMMtraversal tries to find paths on the assembly graph that go through all significant HMM matches in order and agree with rnaviralSPAdes contigs. On contig breakpoints (e.g. in the case of remaining erroneous connections or variations), HMMtraversal is able to do a search for the nearest feasible match. This way the extracted genomic sequence is supported both by the graph topology and the structure of the genome.

coronaSPAdes is bundled with the Pfam SARS-CoV-2 set of HMMs [20]. Note that 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 RdRp that is conserved across all RNA viruses [21]. Hits from these HMMs uniformly cover the genome of coronaviruses, allowing to reconstruct strains mixtures and splice variants.

We explicitly note that the approach of HMMPathExtension is not limited to coronaviral genomes. HMMPathExtension step allows for custom HMM database specification effectively enabling HMM-guided assemblies of other genomes using their internal structure.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

Serratus data, including Frank and Ginger datasets are publically available. Access instructions can be found at https://github.com/ababaian/serratus/wiki/Access-Data-Release. coronaSPAdes version used in this article is available at http://cab.spbu.ru/software/coronaspades/. coronaSPAdes will be included in the future SPAdes releases.

Competing interests

The authors declare that they have no competing interests

Funding

This work was supported by the Russian Science Foundation (grant 19-14-00172). Research was carried out in part by computational resources provided by Resource Center “Computer Center of SPbU”. The authors are grateful to Saint Petersburg State University for the overall support of this work (project id: 51555639).

Authors contribution

DM came up with an algorithmic solution. DM and AK developed the software, analyzed the data, wrote and revised the manuscript. The authors read and approved the final manuscript.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.28.224584; this version posted July 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Acknowledgements

DM is grateful to T32 Weill Cornell Tri-Institutional Training Program in Computational Biology and Medicine and Iman Hajirasouliha for support.

References

Sah R, Rodriguez-Morales AJ, Jha R, Chu DK, Gu H, Peiris M, Bastola A, Lal BK, Ojha HC, Rabaan AA, Zambrano LI. Complete genome sequence of a 2019 novel coronavirus (SARS-CoV-2) strain isolated in Nepal. Microbiology Resource Announcements. 2020 Mar 12;9(11).

Yin C. Genotyping coronavirus SARS-CoV-2: methods and implications. Genomics. 2020 Apr 27.

Kim JM, Chung YS, Jo HJ, Lee NJ, Kim MS, Woo SH, Park S, Kim JW, Kim HM, Han MG. Identification of Coronavirus Isolated from a Patient in Korea with COVID-19. Osong public health and research perspectives. 2020 Feb;11(1):3.

Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015 May 15;31(10):1674-6.

Sutton TD, Clooney AG, Ryan FJ, Ross RP, Hill C. Choice of assembly software has a critical impact on virome characterisation. Microbiome. 2019 Dec;7(1):12.

Nurk S, Bankevich A, Antipov D, Gurevich A, Korobeynikov A, Lapidus A, Prjibelski A, Pyshkin A, Sirotkin A, Sirotkin Y, Stepanauskas R, Clingenpeel SR, Woyke T, Mclean JS, Lasken R, Tesler G, Alekseyev MA, and Pevzner PA. Assembling Single-Cell Genomes and Mini-Metagenomes From Chimeric MDA Products. Journal of Computational Biology.Oct 2013. 20(10):714-737

Roux S, Emerson JB, Eloe-Fadrosh EA, Sullivan MB. Benchmarking viromics: an in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. PeerJ. 2017 Sep 21;5:e3817.

Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome research. 2017 May 1;27(5):824-34.

Viehweger A, Krautwurst S, Lamkiewicz K, Madhugiri R, Ziebuhr J, Hölzer M, Marz M. Direct RNA nanopore sequencing of full-length coronavirus genomes provides novel insights into structural variants and enables modification analysis. Genome research. 2019 Sep 1;29(9):1545-54.

Masters PS. The molecular biology of coronaviruses. Advances in virus research. 2006 Jan 1;66:193-292.

Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. Using SPAdes De Novo Assembler. Current Protocols in Bioinformatics. 2020 Jun;70(1):e102.

Bushmanova E, Antipov D, Lapidus A, Prjibelski AD. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. GigaScience. 2019 Sep;8(9):giz100.

Antipov D, Raiko M, Lapidus A, Pevzner PA. metaviralSPAdes: assembly of viruses from metagenomic data. Bioinformatics. 2020 May 15.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.28.224584; this version posted July 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Meleshko D, Mohimani H, Tracanna V, Hajirasouliha I, Medema MH, Korobeynikov A, Pevzner PA. BiosyntheticSPAdes: reconstructing biosynthetic gene clusters from assembly graphs. Genome research. 2019 Aug 1;29(8):1352-62.

Babaian A., et al. Petabase-scale sequence search uncovers novel Coronaviruses. Submitted. 2020

Hunt M, Gall A, Ong SH, Brener J, Ferns B, Goulder P, Nastouli E, Keane JA, Kellam P, Otto TD. IVA: accurate de novo assembly of RNA virus genomes. Bioinformatics. 2015 Jul 15;31(14):2374-6.

Ruby JG, Bellare P, DeRisi JL. PRICE: software for the targeted assembly of components of (Meta) genomic sequence data. G3: Genes, Genomes, Genetics. 2013 May 1;3(5):865-80.

Yang X, Charlebois P, Gnerre S, Coole MG, Lennon NJ, Levin JZ, Qu J, Ryan EM, Zody MC, Henn MR. De novo assembly of highly diverse viral populations. BMC genomics. 2012 Dec 1;13(1):475.

Sawicki SG, Sawicki DL. Coronaviruses use discontinuous extension for synthesis of subgenome-length negative strands. InCorona-and Related Viruses 1995 (pp. 499-506). Springer, Boston, MA.

Sara El-Gebali, Jaina Mistry, Alex Bateman, Sean R Eddy, Aurélien Luciani, Simon C Potter, Matloob Qureshi, Lorna J Richardson, Gustavo A Salazar, Alfredo Smart, Erik L L Sonnhammer, Layla Hirsh, Lisanna Paladin, Damiano Piovesan, Silvio C E Tosatto, Robert D Finn, The Pfam protein families database in 2019, Nucleic Acids Research. 2019; 47(D1):D427–D432.

Venkataraman S, Prasad BV, Selvarajan R. RNA dependent RNA polymerases: insights from structure, function and evolution. Viruses. 2018 Feb;10(2):76.

Nayfach S, Camargo AP, Eloe-Fadrosh E, Roux S, Kyrpides N. CheckV: assessing the quality of metagenome-assembled viral genomes. BioRxiv. 2020 Jan 1.

(Vancouver style)

		MEGAHIT	metaSPAdes	rnaSPAdes	coronaSPAdes	IVA	PRICE
Fr4nk	Longest	**29219**	26321	20607	**29164**	N/A**,^	8504*
	CheckV completeness	105.2	94.85	72.48	101.7	N/A	31.1
	CheckV AA avg ID%	55.71	56.13	63.34	54.84	N/A	75.8
Ginger	Longest	23905	19453	**29301**	**29277**	3186**	15691*
	CheckV completeness	91.1	69.0	103.66	103.5	N/A	55.9
	CheckV AA avg ID%	85.3	93.93	85.58	85.58	N/A	95.73
PEDV	Longest	**28530**	26316	23312	**27973**	N/A**,^	23559*

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.28.224584; this version posted July 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

	CheckV completeness	101.4	93.6	82.93	99.5	N/A	83.02
	CheckV AA avg ID%	97.55	97.96	97.92	97.5	N/A	97.9

Table 1: Benchmarking of assemblers on several datasets. Shown are: longest viral contig assembled, its completeness as estimated by CheckV[22] 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
^Failed to extend the seed

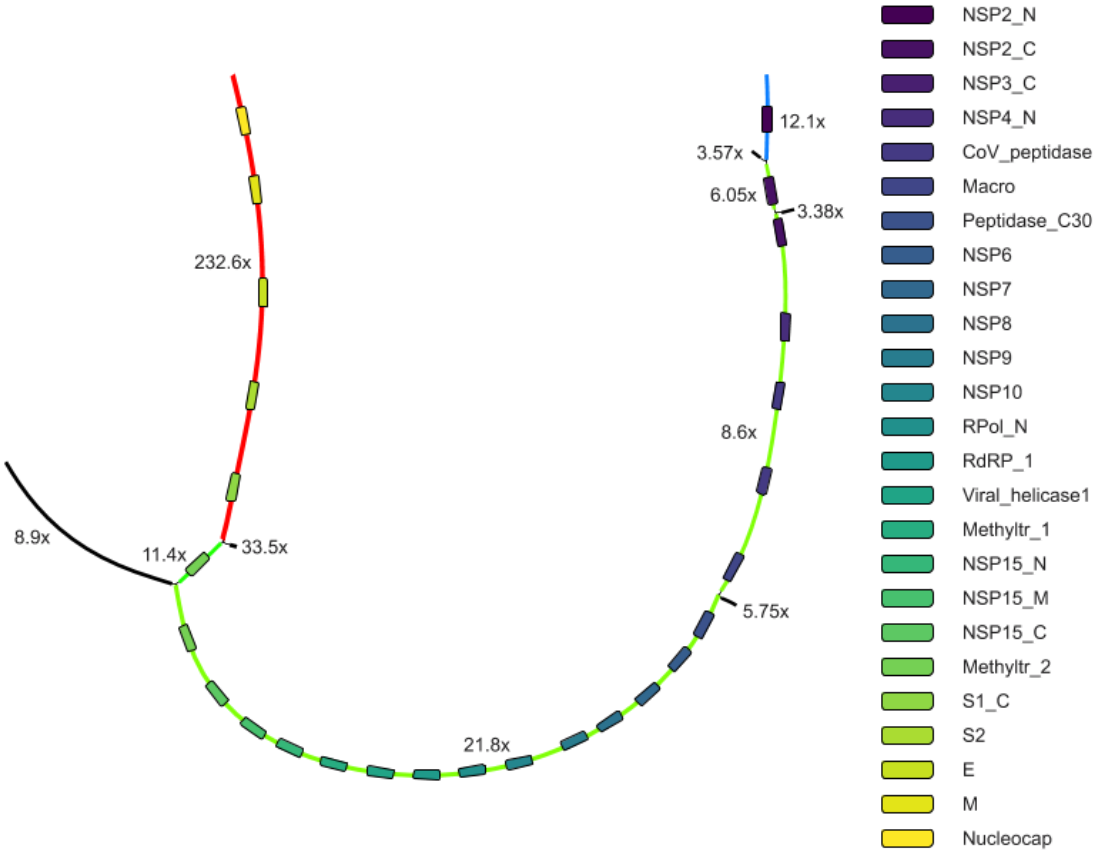
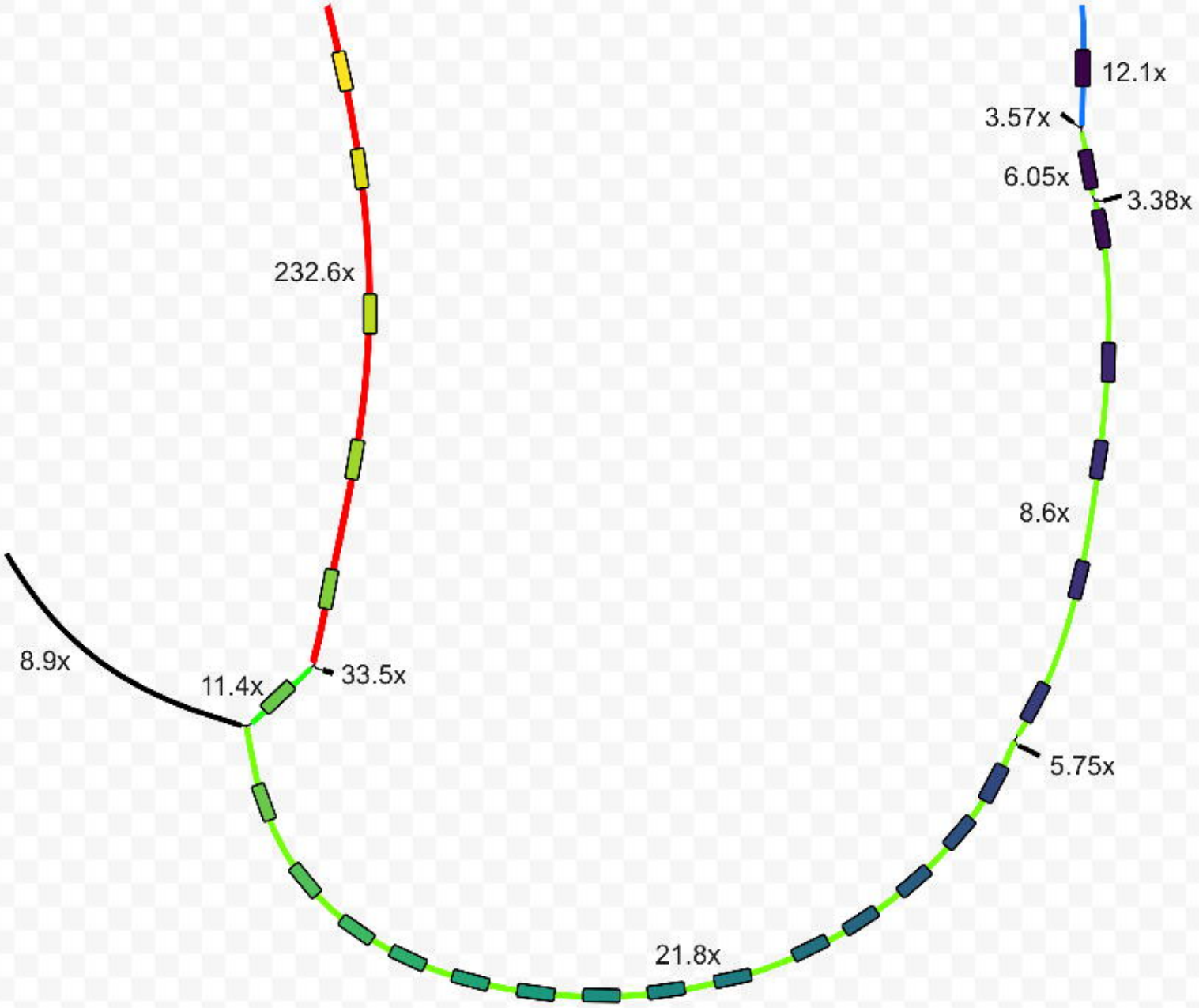


Figure 1: Part of Ginger assembly graph produced with coronaSPAdes. rnaviralSPAdes produced 8 contigs from this subgraph (red, green and blue paths on the graph and 5 black edges), therefore splitting the coronavirus genome into three parts. coronaSPAdes matched viral edges of the graph with domain (rectangles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



NSP2_N
NSP2_C
NSP3_C
NSP4_N
CoV_peptidase
Macro
Peptidase_C30
NSP6
NSP7
NSP8
NSP9
NSP10
RPol_N
RdRP_1
Viral_helicase1
Methyltr_1
NSP15_N
NSP15_M
NSP15_C
Methyltr_2
S1_C
S2
E
M
Nucleocap