BMC Genomics

SOFTWARE Open Access

UnCoVar: a reproducible and scalable workflow for transparent and robust virus variant calling and lineage assignment using SARS-CoV-2 as an example

Alexander Thomas[1†], Thomas Battenfeld[1†], Ivana Kraiselburd[1], Olympia Anastasiou[2], Ulf Dittmer[2], Ann-Kathrin Dörr[1], Adrian Dörr[1], Carina Elsner[2], Jule Gosch[1], Vu Thuy Khanh Le-Trilling[2,3], Simon Magin[1], René Scholtysik[3], Pelin Yilmaz[1], Mirko Trilling[2,3], Lara Schöler[2,3,4], Johannes Köster[5,6†] and Folker Meyer[1*†]

Abstract

Background At a global scale, the SARS-CoV-2 virus did not remain in its initial genotype for a long period of time, with the first global reports of variants of concern (VOCs) in late 2020. Subsequently, genome sequencing has become an indispensable tool for characterizing the ongoing pandemic, particularly for typing SARS-CoV-2 samples obtained from patients or environmental surveillance. For such SARS-CoV-2 typing, various in vitro and in silico workflows exist, yet to date, no systematic cross-platform validation has been reported.

Results In this work, we present the first comprehensive cross-platform evaluation and validation of in silico SARS-CoV-2 typing workflows. The evaluation relies on a dataset of 54 patient-derived samples sequenced with several different in vitro approaches on all relevant state-of-the-art sequencing platforms. Moreover, we present UnCoVar, a robust, production-grade reproducible SARS-CoV-2 typing workflow that outperforms all other tested approaches in terms of precision and recall.

Conclusions In many ways, the SARS-CoV-2 pandemic has accelerated the development of techniques and analytical approaches. We believe that this can serve as a blueprint for dealing with future pandemics. Accordingly, UnCoVar is easily generalizable towards other viral pathogens and future pandemics. The fully automated workflow assembles virus genomes from patient samples, identifies existing lineages, and provides high-resolution insights into individual mutations. UnCoVar includes extensive quality control and automatically generates interactive visual reports. UnCoVar is implemented as a Snakemake workflow. The open-source code is available under a BSD 2-clause license at *github.com/IKIM-Essen/uncovar*.

Keywords SARS-CoV-2, Workflow, Variant calling, Lineage assignment, Next generation sequencing

[†]Alexander Thomas and Thomas Battenfeld Shared first authors.

[†]Johannes Köster and Folker Meyer Shared last authors

*Correspondence:
Folker Meyer
folker.meyer@uk-essen.de

Full list of author information is available at the end of the article

Introduction

Since its first identification, more than 760 million cases of coronavirus disease 2019 (COVID-19) have been reported[1] since December 2019. The causative pathogen SARS-CoV-2 has affected the lives of billions of people, and researchers found infection or vaccination-induced antibodies in 96% of their subjects in a longitudinal study [1]. High infection rates and continuing uncontrolled transmission led to the emergence and spread of viral lineages carrying fitness-enhancing mutations [2–11], while controlling transmission and vaccination promoted the evolution of immune-evasive alterations in the viral genome [12]. Due to their relatively high transmissibility [3, 5, 10, 13, 14], such variants of concern (VOCs) carrying mutations beneficial for the virus have replaced the wild type [3, 15, 16], making whole-genome sequencing with next-generation sequencing (NGS) approaches instrumental for assessing the genomic diversity of the virus in patients.

Previous work has focused on the reconstruction of virus genomes [17–22] and the surveillance of SARS-CoV-2 genomes [23–25]; however, limited attention has been given to reproducibility and portability. In addition, no comprehensive multiplatform benchmark dataset from various protocols and sequencing instruments, including Sanger sequences as ground truth for assessing SARS-CoV-2-related workflows, has been devised thus far. In this work, we present both benchmark dataset and UnCoVar, a reproducible, transparent, and scalable analysis workflow that accepts sequencing products from various protocols. UnCoVar has been extensively optimized for SARS-CoV-2 in routine clinical application and environmental surveillance during the pandemic while being straightforwardly adaptable to future outbreaks of other viruses.

Methods

The UnCoVar workflow

UnCoVar is a Snakemake [26] workflow for virus analysis that provides full in silico reproducibility, diagnostic transparency, uncertainty awareness, and extensive interactive graphical exploration of results.

UnCoVar is publicly available at https://github.com/IKIM-Essen/uncovar under the BSD-2-clause license. Detailed information on the software tools and libraries used in UnCoVar and its usage are available at https://ikim-essen.github.io/uncovar.

UnCoVar consists of four main modules: (i) preprocessing and quality control of raw sequence data, (ii) de novo assembly, reference-guided scaffolding, variant calling and consensus building, (iii) lineage detection, and (iv) aggregation of results and report generation (Fig. 1). Instructions on installation, deployment, configuration and execution as well as a detailed description of the tool chain can be found in the online documentation.

The approaches used for quality control, preprocessing, assembly and lineage detection are described in Table 1. We highlight that only the lineage detection tool Pangolin is SARS-CoV-2 specific (specificity tested against Non-SARS-CoV-2 Corona viruses; Appendix Chap. 5), and the pipeline can easily be adapted to other viral pathogens by registering the respective reference genomes and using Kallisto [27] instead of Pangolin [28] for lineage detection. Moreover, we expect Pangolin (or a successor) to be adapted in the case of future non-SARS-CoV-2 pandemics. With this amount of flexibility, UnCoVar serves the concept for a Disease X [29] analysis tool, a yet unknown pathogen with the potential for an endemic or pandemic outbreak.

UnCoVar is adjustable via a thoroughly documented configuration file. It supports whole-genome shotgun and amplicon-based sequencing from Illumina and Nanopore sequencing and has been extensively tested with data from both sequencing methods from a clinical dataset. In the following, we provide methodological details of the major functionalities of UnCoVar.

Variant calling

UnCoVar employs technology-specific variant callers (short reads: freeBayes [40] for small variants and DELLY [41] for structural variants; long reads: Medaka variant [35] for small variants and Longshot [42] for structural variants) to obtain a list of candidate variants for each investigated sample. The candidate variants are subsequently given to the generic variant classification functionality of Varlociraptor[2] [43].

Complementary genome reconstruction methods

A variety of assemblers have been compared and two default assembly options have been selected according to each library preparation method (MEGAHIT [36] for shotgun, metaSPAdes [37] for amplicon derived samples) based on the best performance (Appendix Chap. 4 and Figures A2+A3). All alternative assembly tools tested (Trinity [44], Velvet [45], MEGAHIT-meta large/sensitive and corona- [46] /rnaviral-/ standard SPAdes [47]), are available and can be used via selecting them in UnCoVar's config file. The pipeline offers two approaches for determining the genomic sequence of a given virus sample. The first (and preferred) approach uses de novo assembly, followed by reference-guided scaffolding. Then, reads are mapped against the obtained assembly using BWA-MEM

[1] https://www.who.int/news-room/fact-sheets/detail/coronavirus-disease-(covid-19).

[2] Using Varlociraptor's variant calling grammar: https://varlociraptor.github.io/docs/calling/#generic-variant-calling.

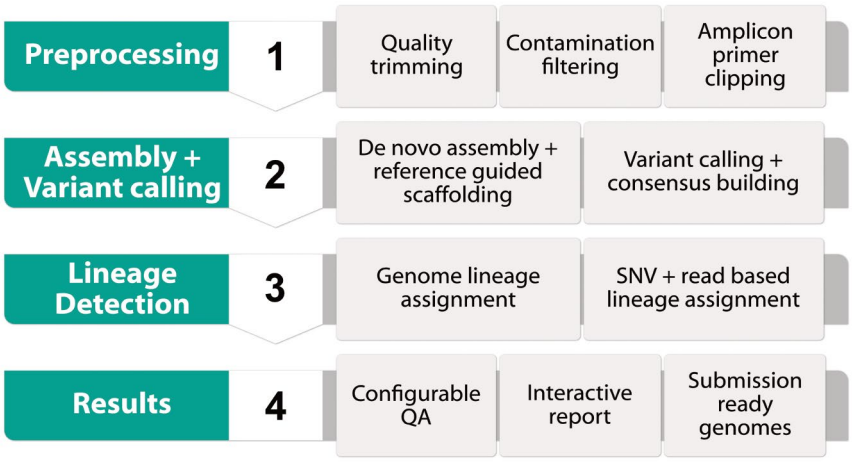

Fig. 1 Outline of UnCoVar. The individual steps of the workflow can be summarized as follows: preprocessing, assembly and variant calling, lineage detection and result generation

Table 1 Tools used in UnCoVar depending on the type of input data (Illumina short reads or Nanopore long reads)

Stage	Step	Tool Illumina	Tool Nanopore	SARS-CoV-2 specific
Preprocessing	primer clipping	BAMClipper [30]	NoTrAmp [31]	no
	quality clipping	fastp [32]		
	contamination removal	Kraken2 [33]		
	Denoising	***[a]	Canu [34], Medaka [35]	
Assembly	Assembly[b]	MEGAHIT [36], metaSPAdes [37]		
	scaffolding	RaGOO [38]		
	polishing	BCFtools consensus [39]	Medaka consensus	
Variant calling	SNV calling	freeBayes [40], DELLY [41]	Medaka variant, Longshot [42]	
	SNV validation	Varlociraptor [43]		
Lineage detection	read based lineage assignment	Kallisto [27]		
	lineage call	Pangolin [28]		yes

[a]No denoising is performed for Illumina reads. Instead, assembly products are polished with uncertainty-aware variant calls from Varlociraptor

[b]Besides the default assembly options, Trinity, Velvet, MEGAHIT-meta large/sensitive and corona-/rnaviral-/ standard SPAdes are available for use

[48], and variants are called with Varlociraptor (see section Variant calling). Variants for which the subclonal-major, subclonal-high and clonal events summed to at least a probability of 0.95 were used to polish the assembly. Low-quality loci (those with low read depth and ambiguous basecalls) are masked by N or IUPAC codes[3]

The second approach maps reads against the primary reference genome of the investigated virus and applies the above polishing approach to the reference genome, including the masking of uncovered regions.

Lineage assignment

Lineage calling approaches fall into two distinct classes: those requiring an almost fully reconstructed genome sequence [49] and those using raw sequencing products in the form of reads, without the necessity of sequence assembly [50–55].

UnCoVar offers three approaches for assigning lineages to samples. First, based on the obtained genome reconstruction, in the case of SARS-CoV-2, UnCoVar utilizes the machine learning driven method Pangolin [28] to assign a lineage.

Second, it employs Kallisto [27] to quantify the numbers of reads originating from given lineage reference sequences, and subsequently calculates their fraction among the total amount of mappable reads. This approach has the advantage of being able to detect lineage mixtures within a single sample, which can allow the detection of mixed infections or the assessment of wastewater samples.

To account for the rapid evolution of SARS-CoV-2, UnCoVar offers a comparison between the investigated sample and the most similar lineages at the level of individual variants. The pipeline obtains the catalog of all known amino acid and noncoding alterations of variants/

[3] In line with criteria defined by the Robert-Koch-Institute, Germany: https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/DESH/Qualitaetskriterien.pdf?__blob=publicationFile

lineages of concern (VOCs) available on covariants.org. Amino acid alterations are back-translated into all potential causing multiple nucleotide variants (MNVs). The resulting set of candidate variants is called using Varlociraptor (see section Variant calling, leveraging Varlociraptor's functionality to classify any set of candidate variants). To determine the degree of similarity between the sample and the VOCs, we performed the following scoring. Let n be the total number of variants and m be the number of VOCs. Let X be a binary matrix that relates variants with VOCs, namely, $X_{i,j} = 1$ if and only if variant i is in VOC_j, with $i = 1,\ldots,n$ and $j = 1,\ldots,m$. Let θi be the latent allele frequency of variant i in the given sample and $\hat{\theta}_i$ be the maximum a posteriori estimate of θi as provided by Varlociraptor. Let $p_i = Pr(\theta_i > 0 \mid D)$ be the posterior probability that the variant i is present in the sample (i.e., the probability that its latent allele frequency is greater than zero, given the data D). Then, the similarity of a given sample to VOC_j can be calculated as the Jaccard-like similarity score:

$$\frac{\sum_{i=1}^{n} p_i \cdot \hat{\theta}_i \cdot X_{i,j} + (1 - p_i) \cdot (1 - X_{i,i})}{n}$$

The better the variant pattern of a VOC is matched (present true positive plus absent true negative mutations), the closer the similarity score is to one. In contrast, if the variant pattern tends toward being the opposite of a VOC, the corresponding score tends toward zero. A low maximum a posteriori VAF estimate or a weaker probability lowers the summand for a specific variant i, such that the overall score decreases.

In this way, UnCoVar is capable of assigning lineage similarities without a fully reconstructed genome, which is especially useful when the analyzed samples are derived from patients with low-level viremia or from environmental sewage water samples, where the abundance of viral RNAs is low and potentially contains several different lineages. UnCoVar reports the top 10 VOC lineages found based on the calculated Jaccard-like similarity of all present and absent mutations included in the VOC database.

Graphical report

UnCoVar's high-level interactive graphical reporting interface allows noncomputational scientists to navigate the details of the analysis and results. The user interface provides an accurate picture of uncertainties in the data. The Snakemake-generated report is portable and maintenance-free since it does not require a running and constantly maintained database or web service. It can be easily archived, distributed via email, a static web server, or any file-sharing platform and solely requires an HTML5[4] compliant web browser to be viewed (see Fig. 2). A detailed overview of all included results can be found on the GitHub pages of UnCoVar (https://ikim-essen.github.io/uncovar/).

The benchmark dataset

We present a benchmark dataset of 54 clinical SARS-CoV-2 patient samples. These 54 samples were obtained from SARS-CoV-2 qPCR-positive patients at the University Hospital Essen, Germany, covering the period from February to September 2021[5]. For all the samples, cDNA synthesis was conducted using the LunaScript® RT SuperMix kit (New England Biolabs, USA). Subsequently, all samples underwent Sanger sequencing for a portion of the genome region encoding the S protein as well as two separate procedures for the targeted sequencing of complete SARS-CoV-2 genomes by tiled amplicons. First, a sequencing library was prepared according to the LoCost nCoV-2019 sequencing protocol (Quick) using the ARTICv3 primer panel (Integrated DNA Technologies, Coralville, Iowa) for loading onto an R9.4.1 Flow Cell and subsequent sequencing on a GridION device (Oxford Nanopore Technologies, Oxford, United Kingdom). Second, library preparation was performed using the EasySeq™ SARS-CoV-2 WGS kit (NimaGen, Nijmegen, The Netherlands), followed by sequencing on an Illumina MiSeq sequencing platform (Illumina Inc., San Diego, CA, USA). For a subset of 32 samples, library preparation was additionally performed using the Illumina COVIDSeq Assay Kit (Illumina Inc., San Diego, CA, USA) followed by sequencing on an Illumina MiSeq platform. The approaches of the Illumina and NimaGen library preparation kits are similar, except that the EasySeq kit combines cDNA amplification and index PCR in a single reaction. Since the dataset was collected in the middle of the SARS-CoV-2 pandemic, we carefully curated the matching between the different technologies to rule out human errors. As a result, we discarded 22 Illumina samples due to potential sample swaps. The data generated by all different approaches were then analyzed and used to benchmark precision and recall across different sequencing platforms in terms of calling individual mutations, virus lineages, and sequence assemblies. The raw NGS benchmarking data is available for download at https://www.ebi.ac.uk/ena/browser/view/PRJEB73579.

Results

Using the presented benchmark dataset, we analyzed the performance of UnCoVar and other state-of-the-art analysis pipelines in terms of precision, recall and runtime. Furthermore, we investigated the required number and

[4] http://www.w3.org/TR/html5.

[5] ethics vote #20-9512-BO.

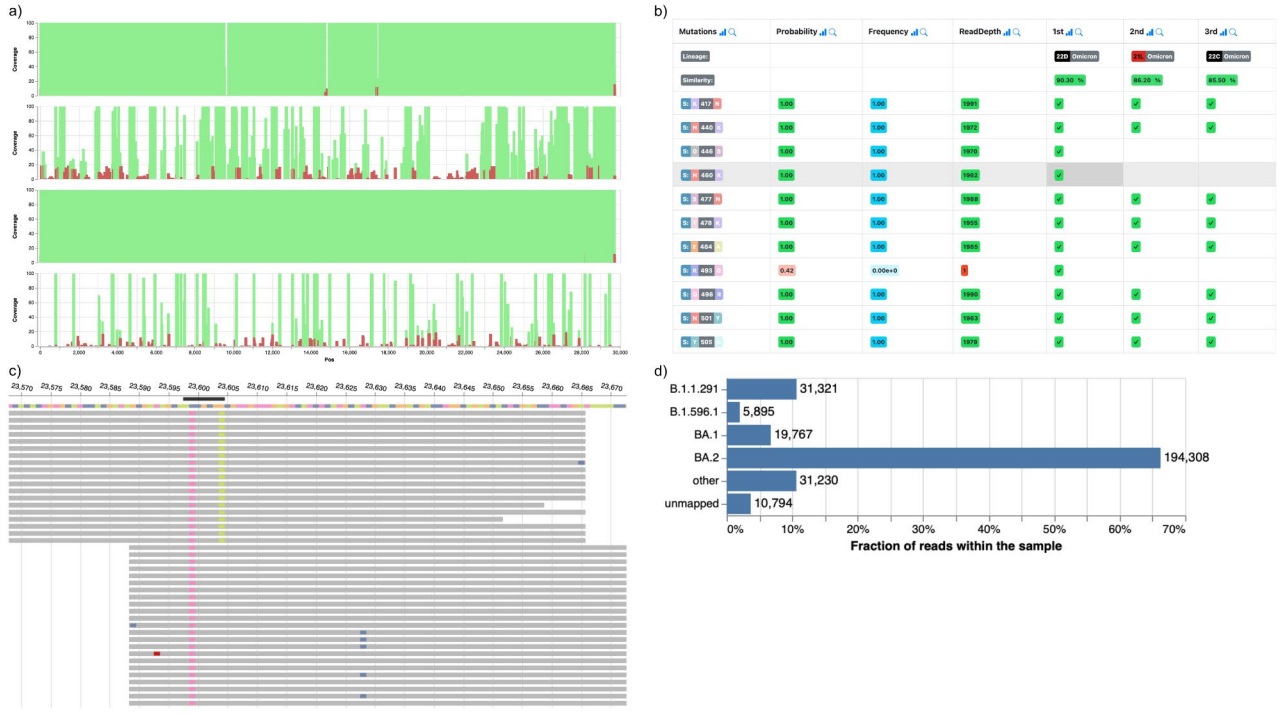

Fig. 2 Four different example elements of the results generated by UnCoVar: **(a)** The genome coverage of the aligned reads, visualized for multiple samples, **(b)** evaluation of known protein alterations from VOCs for one sample, **(c)** a pileup of reads at the position of one protein alteration. The mutations observed for multiple reads (gray bars) for a single sample, here in the S gene, **(d)** The lineage assignments inferred for single reads for one sample. Unmapped reads can be attributed to low sequence quality and variation beyond the considered lineages

length of reads needed for accurate lineage assignment in UnCoVar and examined the time required to execute the workflow. Finally, we compared UnCoVar against other available pipelines using the benchmark dataset described above.

Benchmarking

To assess the different sequencing protocols, the preprocessed reads (see Methods) were aligned to the primary SARS-CoV-2 reference genome from Wuhan (NC_045512.2 or MN908947), and variants were called using UnCoVar. Only variants with a posterior probability ≥ 0.95 for presence according to Varlociraptor were considered, thereby controlling the local false discovery rate in a Bayesian sense at 0.05. The observed variants were compared with those found in the Sanger sequence in the corresponding region. Variants outside of the Sanger sequenced region were omitted. If a variant also occurred in the Sanger sequencing, it was considered a true positive; if not, it was considered a false positive (assuming that Sanger sequencing has the highest possible accuracy). Sanger-based variants that did not occur in the investigated sample were considered false negatives. Let TP, FP, and FN be the respective numbers of true positives, false positives, and false negatives across all samples. We defined precision as the fraction TP/(TP+FP) of true positives among all predicted variants and recall as the fraction TP/(TP+FN) of true positives among all variants in the Sanger sequences. Obviously, the recall can drop with decreasing sequencing depth. More details on the sequencing depth necessary for proper lineage assignment can be found in Appendix Chap. 6. In general, we can conclude, that Pangolin and Kallisto perform equally well in case of nearly complete assemblies while the lineage prediction accuracy of Pangolin decreases with the of completeness of the assembly (Appendix Chap. 6, Figure A4).

Table 2 UnCoVars precision/recall of observed variants per sequencing kit vs. observed variants of Sanger sequencing. *Reduced number of samples for the Illumina kit due to potential sample swaps; see methods

	Sanger	Artic/ONT	Illumina	NimaGen
Variants	160/85*	160	83*	161
Precision	-	1.0	1.0	0.97
Recall	-	1.0	0.98	0.98

As shown in Table 2, the UnCoVar pipeline achieves outstanding precision and recall when using Sanger sequencing data as a true positive gold standard across three different in vitro technologies. Importantly, the precision always stays above the expectation of 0.95 defined by the controlled FDR of 0.05, indicating that the

statistical model of Varlociraptor used in UnCoVar manages to properly assess the uncertainty in the data.

Comparison with other pipelines

We compared UnCoVar against other available state-of-the-art pipelines using the above-mentioned benchmark dataset (Table 3).

All but two of the pipelines achieved a precision and recall above 0.90. This represents a satisfactory result both for the software developers and for clinicians using the results from these pipelines. UnCoVar was the only pipeline that consistently achieved at least 0.97 for both precision and recall across the different in vitro platforms used. When measuring and comparing the execution time to produce variant callings between all considered workflows (average of the individual processing time for all 54 samples), UnCoVar was the fastest for one and close to the compared other pipelines in the other Illumina in vitro approach. A more detailed view of the run times of UnCoVar with differing sequencing depths can be found in Appendix Chap. 2. For Oxford Nanopore data, UnCoVar is one of two pipelines capable of processing such samples without errors at the time of writing, with perfect precision and recall rates and the quickest average execution time measured for all 54 samples. We note that we cannot guarantee overall correct usage of the other software products and executed the compared workflows based on the available documentation. Any bugs that occurred were reported to the original authors. An overview of the tested pipelines and reasons for exclusion can be found in the appendix (Appendix Chap. 7, Table A1).

We posit that the presence of a vendor- and platform-agnostic gold standard for NGS data supported by non-NGS data will enable other groups to use the data for benchmarking their approaches.

Discussion

We present UnCoVar, a fully automated, reproducible workflow for analyzing viral pathogen sequencing data. In addition, we present a thoroughly investigated gold standard benchmark dataset of 54 SARS-CoV-2 samples sequenced with multiple technologies. Using this dataset, we show that UnCoVar outperforms all other available analysis pipelines in terms of recall and precision. UnCoVar thereby manages to accurately control the false discovery rate using Varlociraptor [43].

By using a combination of Snakemake [26], Conda/Mamba, and Snakedeploy, the workflow is portable, reproducible, transparent, and adaptable to any viral pathogen. A combination of different state-of-the-art tools delivers a robust analysis that accepts sequencing products from a range of different instruments and protocols as input.

During the SARS-CoV-2 pandemic, rapid viral mutations played a major role in increasing infection rates [12, 59–61]. While other approaches [56, 57] commonly use only one strategy for crucial steps in the analysis (e.g., de novo assembly or SNV-based consensus building), UnCoVar provides complementary functions for assembly, variant calling, genome reconstruction, and lineage identification. With the strength of using Varlociraptor and its powerful features for the probabilistic re-evaluation of identified mutations, we integrated a unique addition to conventional variant calling methods, as confident identification of SNVs and other mutations played a crucial role in pandemic surveillance. The widely used tool Pangolin for SARS-CoV-2 lineage assignment depends on accurate genome assembly, which UnCoVar achieves by automated SNV-based consensus building, integrated quality assurance and postprocessing of reconstructed genomes. While this is commonly achieved when sequencing patient samples, a lack of full-genome amplification and sequencing and therefore, incomplete genome assembly often occurs in the case of analyzing

Table 3 Computing time and precision/recall comparison of the identified variants between UnCoVar and three other state-of-the-art pipelines. Computing time is given as the median computing time per sample when running all considered benchmark samples and performing only the variant callings. *Reduced number of samples for Illumina kit due to potential sample swaps (Nimagen/ONT = 54 samples; Illumina = 32 samples)

Kit + Sequencer	Pipeline	Precision	Recall	Computing time (h: mm: ss)
Illumina/Illumina*	UnCoVar	1.0	0.98	0:03:44
	Nf-core-viralrecon [53]	0.98	0.99	0:03:46
	V-Pipe [56]	0.66	0.95	1:15:41
	CoVPipe [57]	0.92	0.99	0:06:41
NimaGen/Illumina	UnCoVar	0.97	0.98	0:06:48
	Nf-core-viralrecon	0.99	0.76	0:05:21
	V-Pipe	0.9	0.33	1:12:46
	CoVPipe	0.78	0.87	0:04:18
Artic/ONT	UnCoVar	1.0	1.0	0:02:02
	Artic-medaka [58]	0.98	1.0	0:06:03

environmental – for example, wastewater – samples. Furthermore, evaluating known SARS-CoV-2 protein alterations and not being dependent on a fully reconstructed genome allows us to identify the occurrence of new virus variants through the exclusivity of specific mutations. By providing all these "belts and suspenders", UnCoVar is a versatile all-in-one pipeline with considerable potential, not only for analyzing SARS-CoV-2 samples.

Future work will entail the potential addition of BUSCO [62] for assembly quality assessment. Moreover, we will investigate the use of pangenome references [63] for further improving contamination detection and reducing reference bias in read alignment.

We will work on updating the benchmark dataset with additional SARS-CoV-2 variants and attempt to include other sequencing platforms. As we continue to analyze patient-derived samples from our institution, we will maintain the SARS-CoV-2 analysis and include additional viral pathogens (RSV, influenza A and B) for analysis with UnCoVar. UnCoVar was efficiently employed for the characterization of SARS-CoV-2 variants from wastewater samples [64], and a prototypical module of UnCoVar was employed in a SARS-CoV-2 surveillance project at neighborhoods and city scales in the metropolitan Ruhr area of Germany (Thomas et al., in preparation).

Availability and requirements

Project name: UnCoVar.
Project home page: *github.com/IKIM-Essen/uncovar*.
Operating system(s): platform independent.
Programming language: Python.
Other requirements: Conda, Snakemake 6.9. or higher.
License: BSD-2-Clause License.
Any restrictions to use by non-academics: None.

Abbreviations

COVID-19	Coronavirus Disease 2019
NGS	Next-Generation Sequencing
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
VOCs	Variants of Concern
SNVs	Single Nucleotide Variants

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-024-10539-0.

Supplementary Material 1

Acknowledgements

We thank Felix Wiegand for his dedicated work on the visualization of the UnCoVar report. We would also like to thank Lena Kinzel and Cara-Lila van Bassewitz for their ongoing support in the development and improvement of UnCoVar.

Author contributions

AT, TB, PY, JK, and FM conceived the study. AT, TB, JK, and SM implemented the workflow. AT, TB, IK, PY, MT, JK, and FM wrote the manuscript. AT and TB conducted the associated analyses. IK, SM, JG, FM, AD, OA, CE, LS, VL, RS, MT and UD generated the benchmark datasets in vitro. JK and FM supervised the work. All the authors have read and approved the final manuscript.

Funding

This work was partially supported by the WBEready consortium (grant no. ZMII2-2523COR10A-E), funded by the German Federal Ministry for Health (Bundesministerium für Gesundheit, BMG), and by the SMITH-Medizininformatik-Konsortium-Nachwuchsgruppe Vorhersage von Sepsis auf Basis von Mikrobiomsequenzdaten (MicrobiomSepsisPred, grant no. 01ZZ2013) , funded by the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung, BMBF).
Open Access funding enabled and organized by Projekt DEAL.

Data availability

Sequence data that support the findings of this study have been deposited in the European Nucleotide Archive with the primary accession code PRJEB73579.

Declarations

Ethics approval and consent to participate

The studies involving human participants were reviewed and approved by the Ethics Committee of the Faculty of Medicine at University Duisburg-Essen (Approval number 20-9512-BO). Informed written consent for participation and publication was given by all participants.

Competing interests

The authors declare no competing interests.

Author details

[1]Data Science Research Group, Institute for Artificial Intelligence in Medicine (IKIM), University Hospital of Essen, University of Duisburg-Essen, Essen, Germany
[2]Institute for Virology, University Hospital of Essen, University of Duisburg-Essen, Essen, Germany
[3]Institute for the Research on HIV & AIDS-associated Diseases, University Hospital of Essen, University of Duisburg-Essen, Essen, Germany
[4]Institute of Cell Biology (Cancer Research), University Hospital of Essen, University of Duisburg-Essen, Essen, Germany
[5]Bioinformatics and Computational Oncology, Institute for Artificial Intelligence in Medicine (IKIM), University Hospital of Essen, University of Duisburg-Essen, Essen, Germany
[6]Division of Molecular and Cellular Oncology, Department of Medical Oncology, Harvard Medical School, Boston, MA, USA

Received: 4 April 2024 / Accepted: 18 June 2024
Published online: 28 June 2024

References

1. Jones JM, Manrique IM, Stone MS, Grebe E, Saa P, Germanio CD, Spencer BR, Notari E, Bravo M, Lanteri MC, et al. Estimates of SARS-CoV-2 seroprevalence and incidence of primary SARS-CoV-2 infections among blood donors, by COVID-19 Vaccination Status - United States, April 2021-September 2022. MMWR Morb Mortal Wkly Rep. 2023;72(22):601–5.
2. Bloom JD, Neher RA. Fitness effects of mutations to SARS-CoV-2 proteins. Virus Evol 2023, 9(2).
3. Harvey WT, Carabelli AM, Jackson B, Gupta RK, Thomson EC, Harrison EM, Ludden C, Reeve R, Rambaut A, Peacock SJ, et al. SARS-CoV-2 variants, spike mutations and immune escape. Nat Reviews Microbiol 2021. 2021;19(7):7.
4. Kemp SA, Collier DA, Datir RP, Ferreira I, Gayed S, Jahun A, Hosmillo M, Rees-Spear C, Mlcochova P, Lumb IU, et al. Author correction: SARS-CoV-2 evolution during treatment of chronic infection. Nature. 2022;608(7922):E23.
5. Markov PV, Ghafari M, Beer M, Lythgoe K, Simmonds P, Stilianakis NI, Katzourakis A, Markov PV, Ghafari M, Beer M, et al. The evolution of SARS-CoV-2. Nat Reviews Microbiol 2023. 2023;21(6):6.
6. Meng B, Kemp SA, Papa G, Datir R, Ferreira I, Marelli S, Harvey WT, Lytras S, Mohamed A, Gallo G, et al. Recurrent emergence of SARS-CoV-2

spike deletion H69/V70 and its role in the alpha variant B.1.1.7. Cell Rep. 2021;35(13):109292.
7. Munnink BBO, Sikkema RS, Nieuwenhuijse DF, Molenaar RJ, Munger E, Molenkamp R, van der Spek A, Tolsma P, Rietveld A, Brouwer M, et al. Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans. Science. 2021;371(6525):172–7.
8. Obermeyer F, Jankowiak M, Barkas N, Schaffner SF, Pyle JD, Yurkovetskiy L, Bosso M, Park DJ, Babadi M, MacInnis BL et al. Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness. *Science* 2022, 376(6599).
9. Schröder S, Richter A, Veith T, Emanuel J, Gudermann L, Friedmann K, Jeworowski LM, Mühlemann B, Jones TC, Müller MA et al. Characterization of intrinsic and effective fitness changes caused by temporarily fixed mutations in the SARS-CoV-2 spike E484 epitope and identification of an epistatic precondition for the evolution of E484A in variant Omicron. *Virology Journal 2023 20:1* 2023, 20(1).
10. Wang X, Hu M, Liu B, Xu H, Jin Y, Wang B, Zhao Y, Wu J, Yue J, Ren H. Evaluating the effect of SARS-CoV-2 spike mutations with a linear doubly robust learner. Front Cell Infect Microbiol 2023, 13.
11. Yurkovetskiy L, Wang X, Pascal KE, Tomkins-Tinch C, Nyalile TP, Wang Y, Baum A, Diehl WE, Dauphin A, Carbone C, et al. Structural and functional analysis of the D614G SARS-CoV-2 spike protein variant. Cell. 2020;183(3):739–e751738.
12. Carabelli AM, Peacock TP, Thorne LG, Harvey WT, Hughes J, Peacock SJ, Barclay WS, de Silva TI, Towers GJ, Robertson DL et al. SARS-CoV-2 variant biology: immune escape, transmission and fitness. *Nature Reviews Microbiology* 2023 21:3 2023-01-18, 21(3).
13. Cheng MH, Krieger JM, Kaynak B, Arditi M, Bahar I. Impact of South African 501.V2 variant on SARS-CoV-2 spike infectivity and neutralization: a structure-based Computational Assessment. *bioRxiv* 2021:2021.2001.2010.426143.
14. Petersen E, Koopmans M, Go U, Hamer DH, Petrosillo N, Castelli F, Storgaard M, Al Khalili S, Simonsen L. Comparing SARS-CoV-2 with SARS-CoV and influenza pandemics. Lancet Infect Dis. 2020;20(9):e238–44.
15. Kirca F, Aydoğan S, Gözalan A, Kayipmaz AE, Özdemir FAE, Tekçe YT, Beşer İO, Gün P, Ökten RS, Dinç B. Comparison of clinical characteristics of wild-type SARS-CoV-2 and Omicron. Revista Da Associação Médica Brasileira 2022, 68(10).
16. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, Zhao X, Huang B, Shi W, Lu R, et al. A novel coronavirus from patients with Pneumonia in China, 2019. N Engl J Med. 2020;382(8):727–33.
17. Chen J, Huang J, Sun Y. TAR-VIR: a pipeline for TARgeted VIRal strain reconstruction from metagenomic data. BMC Bioinformatics. 2019;20(1):305.
18. Libin PJK, Deforche K, Abecasis AB, Theys K. VIRULIGN: fast codon-correct alignment and annotation of viral genomes. Bioinformatics. 2019;35(10):1763–5.
19. Pickett BE, Greer DS, Zhang Y, Stewart L, Zhou L, Sun G, Gu Z, Kumar S, Zaremba S, Larsen CN, et al. Virus pathogen database and analysis resource (ViPR): a comprehensive bioinformatics database and analysis resource for the coronavirus research community. Viruses. 2012;4(11):3209–26.
20. Schäffer AA, Hatcher EL, Yankie L, Shonkwiler L, Brister JR, Karsch-Mizrachi I, Nawrocki EP. VADR: validation and annotation of virus sequence submissions to GenBank. BMC Bioinformatics. 2020;21(1):211.
21. Vilsker M, Moosa Y, Nooij S, Fonseca V, Ghysens Y, Dumon K, Pauwels R, Alcantara LC, Vanden Eynden E, Vandamme AM, et al. Genome detective: an automated system for virus identification from high-throughput sequencing data. Bioinformatics. 2019;35(5):871–3.
22. Wang S, Sundaram JP, Spiro D. VIGOR, an annotation program for small viral genomes. BMC Bioinformatics. 2010;11:451.
23. Mercatelli D, Triboli L, Fornasari E, Ray F, Giorgi FM. Coronapp: a web application to annotate and monitor SARS-CoV-2 mutations. J Med Virol. 2021;93(5):3238–45.
24. Singer J, Gifford R, Cotten M, Robertson D. CoV-GLUE: a web application for Tracking SARS-CoV-2 genomic variation. In.: Preprints.org; 2020.
25. Wittig A, Miranda F, Hölzer M, Altenburg T, Bartoszewicz JM, Beyvers S, Dieckmann MA, Genske U, Giese SH, Nowicka M, et al. CovRadar: continuously tracking and filtering SARS-CoV-2 mutations for genomic surveillance. Bioinformatics. 2022;38(17):4223–5.
26. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, Forster J, Lee S, Twardziok SO, Kanitz A et al. Sustainable data analysis with Snakemake. *F1000Research* 2021, 10:33.
27. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34(5):525–7.
28. O'Toole Á, Scher E, Underwood A, Jackson B, Hill V, McCrone JT, Colquhoun R, Ruis C, Abu-Dahab K, Taylor B, et al. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. Virus Evol. 2021;7(2):veab064.
29. Jiang S, Shi Z-L. The First Disease X is caused by a highly transmissible Acute Respiratory Syndrome Coronavirus. Virol Sin. 2020;35(3):263–5.
30. Au CH, Ho DN, Kwong A, Chan TL, Ma ESK. BAMClipper: removing primers from alignments to minimize false-negative mutations in amplicon next-generation sequencing. Sci Rep. 2017;7(1):1567.
31. NoTrAmp. Normalization and Trimming of long-read (ONT, PB) amplicon sequencing data [https://github.com/simakro/NoTrAmp]].
32. Chen S, Zhou Y, Chen Y, Gu J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884–90.
33. Davis MP, van Dongen S, Abreu-Goodger C, Bartonicek N, Enright AJ. Kraken: a set of tools for quality control and analysis of high-throughput sequence data. Methods. 2013;63(1):41–9.
34. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27(5):722–36.
35. sequence correction provided by, Research ONT. [https://github.com/nanoporetech/medaka]].
36. 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;31(10):1674–6.
37. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 2017;27(5):824–34.
38. Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, Lippman ZB, Schatz MC. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. Genome Biol. 2019;20(1):224.
39. Twelve years of. SAMtools and BCFtools - PubMed. *GigaScience* 02/16/2021, 10(2).
40. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv Preprint arXiv:12073907 2012.
41. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012;28(18):i333–9.
42. Edge P, Bansal V, Edge P, Bansal V. Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing. Nat Commun 2019. 2019;10(1):1.
43. Köster J, Dijkstra LJ, Marschall T, Schönhuth A. Varlociraptor: enhancing sensitivity and controlling false discovery rate in somatic indel discovery. Genome Biol. 2020;21(1):98.
44. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29(7):2011–05.
45. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de bruijn graphs. Genome Res 2008/05, 18(5).
46. Meleshko D, Hajirasouliha I, Korobeynikov A. coronaSPAdes: from biosynthetic gene clusters to RNA viral assemblies. Bioinf 2021/12/22, 38(1).
47. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455–77.
48. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:13033997* 2013.
49. Hu T, Li J, Zhou H, Li C, Holmes EC, Shi W. Bioinformatics resources for SARS-CoV-2 discovery and surveillance. Brief Bioinform. 2021;22(2):631–41.
50. Brandt C, Krautwurst S, Spott R, Lohde M, Jundzill M, Marquet M, Hölzer M. poreCov-An Easy to use, fast, and robust workflow for SARS-CoV-2 Genome Reconstruction via Nanopore Sequencing. Front Genet. 2021;12:711437.
51. Desai S, Rashmi S, Rane A, Dharavath B, Sawant A, Dutt A. An integrated approach to determine the abundance, mutation rate and phylogeny of the SARS-CoV-2 genome. Brief Bioinform. 2021;22(2):1065–75.
52. Desai S, Rane A, Joshi A, Dutt A. IPD 2.0: to derive insights from an evolving SARS-CoV-2 genome. BMC Bioinformatics. 2021;22(1):247.
53. Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, Garcia MU, Di Tommaso P, Nahnsen S. The nf-core framework for community-curated bioinformatics pipelines. Nat Biotechnol. 2020;38(3):276–8.
54. Nasir JA, Kozak RA, Aftanas P, Raphenya AR, Smith KM, Maguire F, Maan H, Alruwaili M, Banerjee A, Mbareche H et al. A comparison of whole genome

sequencing of SARS-CoV-2 using amplicon-based sequencing, Random Hexamers, and bait capture. Viruses 2020, 12(8).
55. Sabato LD, Vaccari G, Knijn A, Ianiro G, Bartolo ID, Morabito S. SARS-CoV-2 RECoVERY: a multi-platform open-source bioinformatic pipeline for the automatic construction and analysis of SARS-CoV-2 genomes from NGS sequencing data. *bioRxiv* 2021:2021.2001.2016.425365.
56. Posada-Céspedes S, Seifert D, Topolsky I, Jablonski KP, Metzner KJ, Beerenwinkel N. V-pipe: a computational pipeline for assessing viral genetic diversity from high-throughput data. Bioinformatics. 2021;37(12):1673–80.
57. Lataretu M, Drechsel O, Kmiecinski R, Trappe K, Hölzer M, Fuchs S. Lessons learned: overcoming common challenges in reconstructing the SARS-CoV-2 genome from short-read sequencing data via CoVpipe2. *F1000Research* 2023, 12:1091.
58. Tyson JR, James P, Stoddart D, Sparks N, Wickenhagen A, Hall G, Choi JH, Lapointe H, Kamelian K, Smith AD et al. Improvements to the ARTIC multiplex PCR method for SARS-CoV-2 genome sequencing using nanopore. *bioRxiv* 2020.
59. Kistler KE, Huddleston J, Bedford T. Rapid and parallel adaptive mutations in spike S1 drive clade success in SARS-CoV-2. Cell Host Microbe 2022/04/04, 30(4).
60. Sonnleitner ST, Prelog M, Sonnleitner S, Hinterbichler E, Halbfurter H, Kopecky DBC, Almanzar G, Koblmüller S, Sturmbauer C, Feist L et al. Cumulative SARS-CoV-2 mutations and corresponding changes in immunity in an immunocompromised patient indicate viral evolution within the host. Nat Commun 2022, 13(1).
61. Weber S, Ramirez CM, Weiser B, Burger H, Doerfler W. SARS-CoV-2 worldwide replication drives rapid rise and selection of mutations across the viral genome: a time-course study – potential challenge for vaccines and therapies. EMBO Mol Med 2021-05-31, 13(6).
62. Manni M, Berkeley MR, Seppey M, Zdobnov EM. BUSCO: assessing genomic data Quality and Beyond. Curr Protoc. 2021;1(12):e323.
63. Computational Pan-Genomics C. Computational pan-genomics: status, promises and challenges. Brief Bioinform. 2018;19(1):118–35.
64. Schmiege D, Kraiselburd I, Haselhoff T, Thomas A, Doerr A, Gosch J, Schoth J, Teichgräber B, Moebus S, Meyer F. Analyzing community wastewater in subsewersheds for the small-scale detection of SARS-CoV-2 variants in a German metropolitan area. Sci Total Environ 2023/11/10, 898.

Publisher's Note