BMC Bioinformatics

SOFTWARE

Open Access

HAVoC, a bioinformatic pipeline for reference-based consensus assembly and lineage assignment for SARS-CoV-2 sequences

Phuoc Thien Truong Nguyen[1*], Ilya Plyusnin[2,3], Tarja Sironen[1,3], Olli Vapalahti[1,3,4], Ravi Kant[1,3†] and Teemu Smura[1,4†]

*Correspondence:
phuoc.truong@helsinki.fi
†Ravi Kant and Teemu Smura have contributed equally to this work.
[1] Department of Virology, Faculty of Medicine, University of Helsinki, Helsinki, Finland
Full list of author information is available at the end of the article

Abstract

Background: SARS-CoV-2 related research has increased in importance worldwide since December 2019. Several new variants of SARS-CoV-2 have emerged globally, of which the most notable and concerning currently are the UK variant B.1.1.7, the South African variant B1.351 and the Brazilian variant P.1. Detecting and monitoring novel variants is essential in SARS-CoV-2 surveillance. While there are several tools for assembling virus genomes and performing lineage analyses to investigate SARS-CoV-2, each is limited to performing singular or a few functions separately.

Results: Due to the lack of publicly available pipelines, which could perform fast reference-based assemblies on raw SARS-CoV-2 sequences in addition to identifying lineages to detect variants of concern, we have developed an open source bioinformatic pipeline called *HAVoC* (Helsinki university Analyzer for Variants of Concern). *HAVoC* can reference assemble raw sequence reads and assign the corresponding lineages to SARS-CoV-2 sequences.

Conclusions: *HAVoC* is a pipeline utilizing several bioinformatic tools to perform multiple necessary analyses for investigating genetic variance among SARS-CoV-2 samples. The pipeline is particularly useful for those who need a more accessible and fast tool to detect and monitor the spread of SARS-CoV-2 variants of concern during local outbreaks. *HAVoC* is currently being used in Finland for monitoring the spread of SARS-CoV-2 variants. *HAVoC* user manual and source code are available at https://www.helsinki.fi/en/projects/havoc and https://bitbucket.org/auto_cov_pipeline/havoc, respectively.

Keywords: SARS-CoV2, Variant detection, Reference assembly, Lineage identification, Coronavirus, Sequence analysis

Background

Emerging pathogens pose a continuous threat to mankind, as exemplified by the Ebola virus epidemic in West Africa in 2014 [1], Zika virus pandemic in 2015 [2], and the ongoing Coronavirus disease 2019 (COVID-19) pandemic. These viruses are zoonotic, i.e. have crossed species barriers from animals to humans, alike the majority of emerging human pathogens [3, 4]. The likelihood of this host switching is enhanced by several factors, e.g. global movement of people and animals, environmental changes, increased proximity of humans, wildlife and livestock, and population expansion into new environments [5].

The mutation and evolution rate of RNA viruses is considerably higher than their hosts, which is advantageous for viral adaptation. Mutations in the viral genome are most of the time silent or, if affecting phenotype, related to attenuation, although mutations can also lead to more pathogenic strains. A new virus variant may have one or more mutations that separate it from the wild-type virus already circulating among the general population.

Coronaviruses (family *Coronaviridae*) are enveloped single-stranded RNA viruses, which cause respiratory, enteric, hepatic, and neurological diseases of a broad spectrum of severity among different animals and humans. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a novel evolutionary divergent virus responsible for the present pandemic, has devastated societies and economies globally. The SARS-CoV-2 pandemic has already infected more than 100 million people in 221 countries, causing over 2.2 million global deaths as of 3rd February 2021 [6]. In autumn 2020, a new variant of SARS-CoV-2 known as 20B/501Y.V1 (B.1.1.7) was detected in south-eastern England, Wales, and Scotland [7]. This variant has since spread globally to more than 80 countries. The variant has undergone 23 mutations with 13-nonsynonymous mutations, four amino acid deletions, and six synonymous mutations making the virus more transmissible [8]. Another variant 20C/501Y.V2 (B.1.351) was detected in South Africa which was genetically distant from the UK 20B/501Y.V1 variant [9]. This South African variant with its two mutations in the receptor-binding motif that mainly forms the interface with the human ACE2 receptor has also been widely spreading to circulate globally. It has been noticed that some existing vaccines against SARS-CoV-2 are less effective against the 20C/501Y.V2 variant [10–12]. A third variant being closely monitored is P.1 detected first in Brazil [13]. Interestingly, all these three variants have a mutation in the receptor binding domain (RBD) of the spike protein at position 501, where the amino acid asparagine (N) has been replaced with tyrosine (Y) enabling specific PCR to detect the N501Y mutation [14].

Sequencing and computing consensus sequences or genomes from RNA viruses present certain challenges. The mutation and evolution rate of RNA viruses is considerably higher than their hosts, which is advantageous for viral adaptation. Mutations in the viral genome are most of the time silent or, if affecting phenotype, related to attenuation, although mutations can also lead to more pathogenic strains. A new virus variant may have one or more mutations that separate it from the wild-type virus already circulating among the general population. To detect these minor changes in viral genomes requires bioinformatic tools with high accuracy and sensitivity, which tend to come at the expense of computing speed. Currently, the most common sequencing method employed with SARS-CoV-2 is a combination of PCR-based amplification followed by Illumina sequencing. The resulting reads are then pre-processed and quality-controlled for either a de novo or a reference-based assembly. To identify to what variant a sequenced SARS-CoV-2 belongs to, it is usually analyzed and classified with pangolin [15] or Nextclade, which is a part of Nextstrain [16]. Each tool utilizes different nomenclature for classification [17, 18]. To date, most of the resulting sequences are submitted to the GISAID database [19, 20] (contains over 1 million sequences as of May 1, 2020), which has another classification system for the virus.

As more transmissible coronavirus variants are circulating worldwide, the role of researchers and technology specialists in controlling the pandemic has received more emphasis. The surveillance of virus variants by sequencing the SARS-CoV-2 genomes would provide a fast way to monitor variants and their spread, however, there are only few publicly available methods for quick reference-based consensus assembly and lineage assignment for SARS-CoV-2 samples. For this purpose, we have developed a simple pipeline, called *HAVoC* (Helsinki university Analyzer for Variants of Concern), for quick reference-based consensus assembly from short reads sequenced with Illumina and lineage assignment for SARS-CoV-2 samples. This will provide the end user a quick and accessible method of variant identification and monitoring. The pipeline was developed to be run on Unix/Linux operating systems, and thus can also be used in remote servers, e.g. CSC—IT Center for Science, Finland.

Implementation

HAVoC consists of a single shell script, which performs reference-based consensus assemblies to query SARS-CoV-2 FASTQ sequence libraries and assigns lineages to them individually in succession. It does this using several bioinformatic tools publicly available in Bioconda on Unix/Linux platforms. For *HAVoC* to be utilized, the user is required to install these dependencies. This can be done for example via Biocanda with the following command:

```
conda install fastp trimmomatic bowtie2 bwa sambamba samtools bedtools lofreq bcftools pangolin
```

The script can be started by typing the following line into your command line terminal:

```
sh HAVoC.sh [FASTQ directory]
```

The computing of consensus sequences starts with the tool detecting FASTQ files generated via paired end sequencing in a given input directory and checking that each query FASTQ file has its corresponding counterpart, i.e. mates file. The names of the files are modified to be more concise, e.g. Query-Seq:1_X123_Y000_R1_000.fastq.gz to Query-Seq:1_R1.fastq.gz. The pipeline accepts FASTQ files both in gzipped and uncompressed format.

For the analyses, the user can choose one of two bioinformatic tools to utilize in each phase of the assembly. This can be done by typing the tool wanted (*tools_prepro, tools_aligner* and *tools_sam*) within the options section in the beginning of the script file. For example, if the user wants to deploy Trimmomatic to pre-process FASTQ files, the following line can be changed as follows:

From

```
tools_prepro="fastp"
```

To

```
tools_prepro="trimmomatic"
```

Other options include the number of threads, minimum coverage below which a region is masked (*min_coverage*), and whether to run pangolin to assign lineages to the consensus genome (*run_pangolin*). An additional option allows *HAVoC* to be run in the CSC servers (*run_in_csc*).

The pre-alignment quality control, e.g. removing and trimming low quality reads and bases, removing adapter sequences, can be done with either fastp [21] or Trimmomatic [22]. The former is used by default due to its faster processing speed. Also unlike Trimmomatic, fastp does not require a separate file with adapters (NexteraPE-PE.fa or other provided by the user) making it more usable for different sequencing protocols. In addition to increasing the quality of the reads, this step reduces the time of the following alignment process in which the reads are then aligned to a reference genome of SARS-CoV-2 isolate Wuhan-Hu-1 (Genbank accession code: NC_045512.2) provided in the ref.fa file with BWA-MEM [23] or Bowtie 2 [24]. The end-user may also provide their own reference genome in fasta format by copying it directly to the ref.fa file. BWA-MEM is chosen by default due to it being faster compared to other contemporary tools utilizing different alignment techniques [25]. For higher coverage contigs, the user may want to use Bowtie 2 instead. The resulting SAM and BAM files are processed (includes sorting, filling in mate coordinates, marking duplicate alignments, and indexing reads) with Sambamba [26] or Samtools [27] and the low coverage regions are masked with BEDtools [28]. After masking a variant call is done with LoFreq [29] before computing the consensus sequence via BCFtools of Samtools [27]. Finally, the consensus sequence is analyzed with pangolin [15] to assign a lineage. The whole process is depicted in Fig. 1.

Usage example

We are going to demonstrate a common use case for *HAVoC* with FASTQ files containing reads for SARS-CoV-2 sequences, provided by the Viral zoonoses research unit at University of Helsinki, Finland. The reads were produced via Illumina sequencing. The test files within the Example_FASTQs folder contain paired-end FASTQ files for the UK variant (UK-variant-1) and the South African variant (S-Africa-variant-1). To analyse these example files, the aforementioned command needs to be deployed as follows:

```
sh HAVoC.sh Example_FASTQs
```

Results

The FASTQ files are processed and analyzed with the default options utilizing faster bioinformatic tools (fastp, BWA-MEM and Sambamba) in ca. 2–5 min, depending on the performance of the platform (local or server) and the size of the input FASTQ files. After *HaVoc* has finished the analyses, each FASTQ file is moved to their respective result folders within the FASTQ directory. Each result folder contains a FASTA file for the consensus sequence (e.g. UK-variant-1_consensus.fa) and a CSV file with the lineage information produced by pangolin (e.g. UK-variant-1_pangolin_lineage.csv). In addition to these main result files, each directory contains the original FASTQ files, BAM files

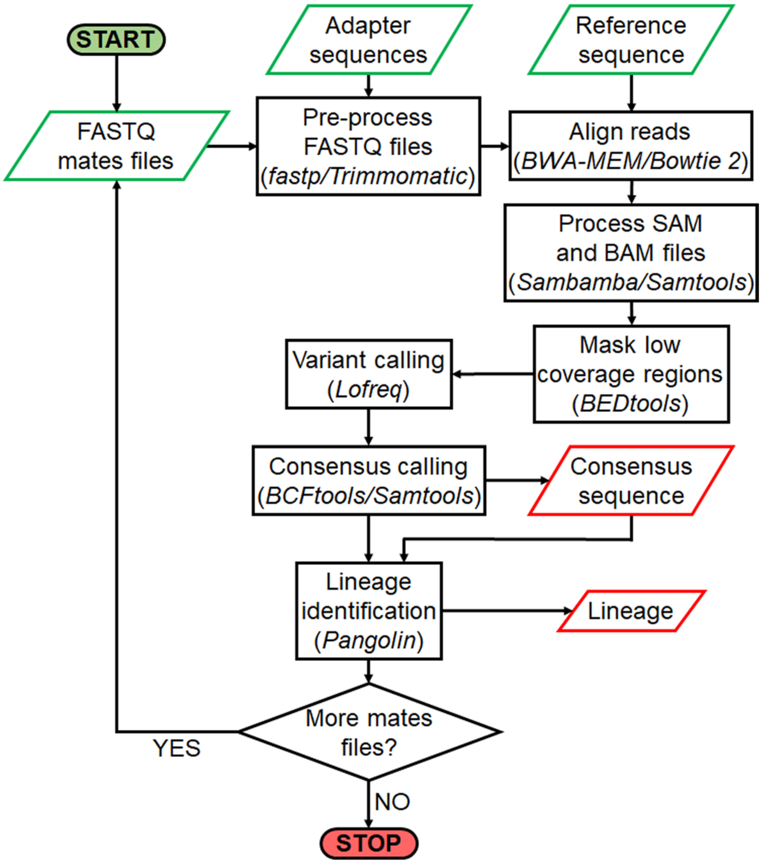
START
Adapter sequences
Reference sequence
FASTQ mates files
Pre-process FASTQ files (fastp/Trimmomatic)
Align reads (BWA-MEM/Bowtie 2)
Process SAM and BAM files (Sambamba/Samtools)
Variant calling (Lofreq)
Mask low coverage regions (BEDtools)
Consensus calling (BCFtools/Samtools)
Consensus sequence
Lineage identification (Pangolin)
Lineage
More mates files?
YES
NO
STOP

Fig. 1 Flowchart describing processes and steps performed by *HAVoC* pipeline. The pipeline constructs consensus sequences from all FASTQ files in an input directory and then compares the resulting sequences to other established SARS-CoV-2 genomes to assign them the most likely lineages. The pipeline requires a FASTA file of adapter sequences for FASTQ pre-processing and a reference genome of SARS-CoV-2 in a separate FASTA file. The adapter file is not required when running the pipeline with fastp option. Input files are highlighted in green and the outputs in red

(original, indexed and sorted), variant call files (VCF) with mutation data, BED file used for masking regions, and fastp report files with the results of FASTQ processing. The resulting directory and file structure with the example files will look as follows:

Example_FASTQs/

UK-variant-1/

UK-variant-1.bam
UK-variant-1_R1.fastq.gz
UK-variant-1_R2.fastq.gz
UK-variant-1_consensus.fa
UK-variant-1_fixmate.bam
UK-variant-1_indel.bam
UK-variant-1_indel.vcf

```
        UK-variant-1_indel_flt.vcf
        UK-variant-1_lowcovmask.bed
        UK-variant-1_markdup.bam
        UK-variant-1_namesort.bam
        UK-variant-1_pangolin_lineage.csv
        UK-variant-1_sorted.bam
        fastp.html
        fastp.json
        S-Africa-variant-1.bam

    S-Africa-variant-1/

        S-Africa-variant-1_R1.fastq.gz
        S-Africa-variant-1_R2.fastq.gz
        S-Africa-variant-1_consensus.fa
        S-Africa-variant-1_fixmate.bam
        S-Africa-variant-1_indel.bam
        S-Africa-variant-1_indel.vcf
        S-Africa-variant-1_indel_flt.vcf
        S-Africa-variant-1_lowcovmask.bed
        S-Africa-variant-1_markdup.bam
        S-Africa-variant-1_namesort.bam
        S-Africa-variant-1_pangolin_lineage.csv
        S-Africa-variant-1_sorted.bam
        fastp.html
        fastp.json
```

Each of the example UK variants should have been categorized as B.1.1.7 and the South African variants as B.1.351 (with pangoLEARN release 2021-02-06). It is important to note however, that as more sequences are uploaded and the pangolin lineage nomenclature updated, the assigned lineages may differ from the expected ones described in this paper.

Regions with low coverages (with default setting under 30) are marked with the letter N during masking and represent gaps in the final consensus sequences.

HAVoC is comparable to alternative combinations of tools, e.g. Jovian and pangolin, in both speed and accuracy. These tools however operate separately, and as of publishing, there are no single public tools that can both perform a reference-based consensus assembly and a lineage identification in an easily accessible manner.

Conclusions

Early detection and understanding of the potential impact of emerging variants of SARS-CoV-2 is of primary importance and can assist in more efficient surveillance and control of the disease. The likelihood of emergence of novel SARS-CoV-2 variants of concern is increased and accelerated by the high mutation rates typical in RNA viruses and the growing number of transmissions and infections both locally and globally.

With the rising number of variants detected worldwide and with many of them associated with increased transmissibility and lower vaccine efficacy, there is an emerging need for fast, efficient and reliable pipelines to help detect, identify and trace SARS-CoV-2 lineages. These pipelines should in addition be accessible to researchers who may not be familiar with utilizing complex bioinformatic tools or scripting pipelines.

Due to these challenges, we have developed *HAVoC*, a simple, reliable and user-friendly pipeline, which can be simply downloaded from our repository and run without being installed. The pipeline performs reference-based assemblies and lineage assignment from SARS-CoV-2 samples sequenced with Illumina utilizing a combination of multiple well-established third-party bioinformatic tools in current use. All its dependencies can be installed via existing package managers, of which we recommend Bioconda. *HAVoC* is currently being developed and updated in addition to being utilized for detecting and tracing SARS-CoV-2 variants of concern, mainly B.1.1.7, B1.351 and P.1, in Finland.

Abbreviations
SARS-CoV-2: Severe acute respiratory syndrome coronavirus 2; COVID-19: Coronavirus disease 2019; *HAVoC*: Helsinki university Analyzer for Variants of Concern.

Acknowledgements
None.

Authors' contributions
Conceptualization: PTN IP RK TS TSi OV. Development: PTN IP RK TS. Testing/Formal analysis: PTN IP RK TS. Funding acquisition: TSi OV. Investigation: PTN IP RK TS. Methodology: PTN IP RK TS. Project administration: RK TS OV. Resources: PTN RK IP TS TSi OV. Validation: PTN IP RK TS. Writing—original draft: PTN RK. Writing—review and editing: IP TS TSi OV. All authors read and approved the final manuscript.

Funding
This study was supported by the Academy of Finland (Grant Number 336490), VEO—European Union's Horizon 2020 (Grant Number 874735) and the Jane and Aatos Erkko Foundation.

Availability of data and materials
Project name: HAVoC (Helsinki university Analyzer for Variants of Concern). Project home page: https://www.helsinki.fi/en/projects/havoc and https://bitbucket.org/auto_cov_pipeline/havoc. Operating system(s): Linux, Mac. Programming language: Shell script. Other requirements: Trimmomatic or fastp, BWA-MEM or Bowtie 2, Samtools, BEDtools, BCFtools, LoFreq and pangolin. License: GNU GPL. Any restrictions to use by non-academics: license needed. Publicly available at https://bitbucket.org/auto_cov_pipeline/havoc.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Virology, Faculty of Medicine, University of Helsinki, Helsinki, Finland. [2]Institute of Biotechnology, University of Helsinki, Helsinki, Finland. [3]Department of Veterinary Biosciences, University of Helsinki, Helsinki, Finland. [4]Department of Virology, University of Helsinki and Helsinki University Hospital, Helsinki, Finland.

Received: 12 February 2021 Accepted: 8 July 2021
Published online: 17 July 2021

References

1. Dixon MG, Schafer IJ, Centers for Disease Control and Prevention (CDC). Ebola viral disease outbreak–West Africa, 2014. MMWR Morb Mortal Wkly Rep. 2014;63:548–51.
2. Kindhauser MK, Allen T, Frank V, Santhana RS, Dye C. Zika: the origin and spread of a mosquito-borne virus. Bull World Health Organ. 2016;94:675-686C. https://doi.org/10.2471/BLT.16.171082.
3. Taylor LH, Latham SM, Woolhouse ME. Risk factors for human disease emergence. Philos Trans R Soc Lond B Biol Sci. 2001;356:983–9. https://doi.org/10.1098/rstb.2001.0888.
4. Woolhouse MEJ, Gowtage-Sequeria S. Host range and emerging and reemerging pathogens. Emerg Infect Dis. 2005;11:1842–7. https://doi.org/10.3201/eid1112.050997.
5. Morens DM, Fauci AS. Emerging pandemic diseases: how we got to COVID-19. Cell. 2020;182:1077–92. https://doi.org/10.1016/j.cell.2020.08.021.
6. Worldometer. COVID-19 Virus Pandemic. Worldometer. https://www.worldometers.info/coronavirus/. Accessed 3 Feb 2021.
7. Rambaut A, Loman N, Pybus O, Barclay W, Barrett J, Carabelli A, et al. Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. Virological. 2020. https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563. Accessed 2 Feb 2021.
8. Leung K, Shum MH, Leung GM, Lam TT, Wu JT. Early transmissibility assessment of the N501Y mutant strains of SARS-CoV-2 in the United Kingdom. Euro Surveill. 2020. https://doi.org/10.2807/1560-7917.ES.2020.26.1.2002106.
9. Tegally H, Wilkinson E, Giovanetti M, Iranzadeh A, Fonseca V, Giandhari J, et al. Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. medRxiv. 2020. https://doi.org/10.1101/2020.12.21.20248640.
10. Mahase E. Covid-19: Novavax vaccine efficacy is 86% against UK variant and 60% against South African variant. BMJ. 2021;372:n296. https://doi.org/10.1136/bmj.n296.
11. Kupferschmidt K. Vaccine 2.0: Moderna and other companies plan tweaks that would protect against new coronavirus mutations. Science. 2021. https://doi.org/10.1126/science.abg7691.
12. Edwards E. J&J says vaccine effective against Covid, though weaker against South Africa variant. NBC News. 2021. https://www.nbcnews.com/health/health-news/j-j-vaccine-effective-against-covid-though-weaker-against-south-n1255400. Accessed 10 Feb 2021.
13. Faria NR, Claro IM, Candido D, Franco LAM, Andrade PS, Coletti TM, et al. Genomic characterisation of an emergent SARS-CoV-2 lineage in Manaus: preliminary findings. Virological. 2021. https://virological.org/t/genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-manaus-preliminary-findings/586. Accessed 3 Feb 2021.
14. Centers for Disease Control and Prevention (CDC). Emerging SARS-CoV-2 Variants. https://www.cdc.gov/coronavirus/2019-ncov/more/science-and-research/scientific-brief-emerging-variants.html. Accessed 12 Feb 2021.
15. O'Toole Á, Scher E, Underwood A, Jackson B, Hill V, McCrone JT, et al. pangolin: lineage assignment in an emerging pandemic as an epidemiological tool. https://github.com/cov-lineages/pangolin. Accessed 12 Feb 2021.
16. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 2018;34:4121–3. https://doi.org/10.1093/bioinformatics/bty407.
17. Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol. 2020;5:1403–7. https://doi.org/10.1038/s41564-020-0770-5.
18. Bedford T, Hodcroft EB, Neher RA. Updated Nextstrain SARS-CoV-2 clade naming strategy. Nextstrain. 2021. https://nextstrain.org/blog/2021-01-06-updated-SARS-CoV-2-clade-naming/. Accessed 7 May 2021.
19. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. Glob Challenges. 2017;1:33–46. https://doi.org/10.1002/gch2.1018.
20. Shu Y, McCauley J. GISAID: global initiative on sharing all influenza data—from vision to reality. Euro Surveill. 2017;22:30494. https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494.
21. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34:i884–90. https://doi.org/10.1093/bioinformatics/bty560.
22. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20. https://doi.org/10.1093/bioinformatics/btu170.
23. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 2013.
24. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9. https://doi.org/10.1038/nmeth.1923.
25. Borozan I, Watt SN, Ferretti V. Evaluation of alignment algorithms for discovery and identification of pathogens using RNA-Seq. PLoS ONE. 2013;8:e76935. https://doi.org/10.1371/journal.pone.0076935.
26. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. Bioinformatics. 2015;31:2032–4. https://doi.org/10.1093/bioinformatics/btv098.
27. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9. https://doi.org/10.1093/bioinformatics/btp352.
28. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2. https://doi.org/10.1093/bioinformatics/btq033.
29. Wilm A, Aw PPK, Bertrand D, Yeo GHT, Ong SH, Wong CH, et al. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res. 2012;40:11189–201. https://doi.org/10.1093/nar/gks918.

Publisher's Note