Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Application Note

OXFORD

Sequence analysis

MALVIRUS: an integrated web application for viral variant calling

Simone Ciccolella [1,†], Luca Denti [1,†], Paola Bonizzoni [1], Gianluca Della Vedova [1], Yuri Pirola [1,‡,*], and Marco Previtali [1,‡]

[1] Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milano, 20126, Italy

[†] Joint First Authors. [‡] Joint Last Authors.
[*] To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Being able to efficiently call variants from the increasing amount of sequencing data daily produced from multiple viral strains is of the utmost importance, as demonstrated during the COVID-19 pandemic, in order to track the spread of the viral strains across the globe.
Results: We present MALVIRUS, an easy-to-install and easy-to-use web application that assists users in (1) computing a variant catalog consisting in a set of population SNP loci from the population sequences and (2) efficiently calling variants of the catalog that are in the sample. Tests on Illumina and Nanopore samples prove the efficiency and the effectiveness of MALVIRUS in genotyping a SARS-CoV-2 strain sample with respect to GISAID data.
Availability: MALVIRUS is freely available at `https://algolab.github.io/MALVIRUS/`.
Contact: yuri.pirola@unimib.it
Supplementary information: Supplementary data are available at *bioRxiv* online.

1 Introduction

The SARS-CoV-2 pandemic has put the global health care services to the test and many researchers are racing to face its swift and rapid spread. Since the outbreak of the virus in China and in other European countries, several studies are using sequencing technologies to track the geographical origin of SARS-Cov-2 and to analyze the evolution of sequence variants (Gudbjartsson *et al.*, 2020; Böhmer *et al.*, 2020). In this context, the availability of efficient approaches to analyze variations from the growing amount of sequencing data daily produced is of the utmost importance.

The typical pipelines for the analysis of variations within viral samples consists in aligning reads against a reference genome (Li, 2018), then analyzing the alignments to discover the variants (McKenna *et al.*, 2010; Wilm *et al.*, 2012). However, the increasing number of viral assemblies available in public databases such as GISAID (Shu and McCauley, 2017), GenBank (Sayers *et al.*, 2019), and the COVID-19 Data Portal allows to build a complete catalog of variants of a viral population. Such a catalog can be used to reduce the complexity of comparative analysis of genetic variants of sequencing samples. Clearly, to this aim, it is crucial that users are assisted by an efficient and easy-to-use method for building and updating the catalog and for calling variants that are in this catalog. In this paper, we introduce MALVIRUS, a web application for quickly genotype newly sequenced viral strains, including but not limited to the SARS-CoV-2 strains. The application is distributed as a multi-platform Docker container and it can be easily accessed using any modern Internet browser. As use case, we show that MALVIRUS is accurate at genotyping a newly-sequenced SARS-CoV-2 strain on both short and long read data.

2 Methods

To efficiently genotype a viral sample from an individual with respect to the current knowledge, we propose MALVIRUS a web application based on four state-of-the-art tools. The application is divided into two logically distinct modules: the creation of the catalog containing the SNP loci of the viral species under investigation and the variant calling from the read sample. Fig. 1 shows the MALVIRUS pipeline.

The first module requires as input the reference genome of the species under investigation, the assemblies of a set of strains of that species, and, if available, the annotation of the genes. The output of this module is the set of population SNP loci in VCF format. MALVIRUS first builds the full-length sequence alignment of the input sequences to the input reference genome using MAFFT (Katoh and Standley, 2013), then extracts

© The Author 2020.

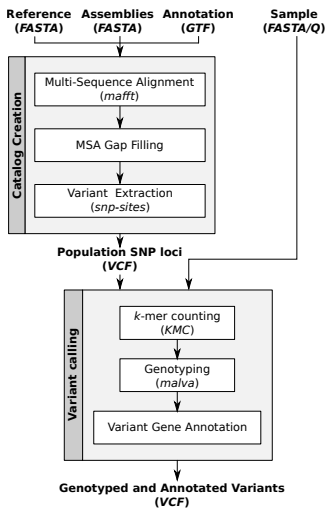

Fig. 1. Schematic representation of the pipeline integrated in MALVIRUS.

the set of population SNP loci from the multiple alignment using snp-sites (Page *et al.*, 2016). Since snp-sites is not able to output variants in positions with gaps, MALVIRUS fills the gaps in the alignment with the corresponding portions of the reference. Although this step might induce some artificial variants, it allows to preserve real ones that might be lost due to incomplete assemblies. If the population under investigation is well characterized and/or the user wants a finer control over the variant catalog, it is possible to upload a custom catalog of SNP loci in VCF format instead of relying on the automatic computation from a set of assemblies.

The second module requires as input a sample of reads in FASTA/Q format and a catalog of population SNP loci along with the corresponding reference genome chosen among the ones computed or uploaded in the first module. The output of second module is a VCF containing the genotype information of the sample and their qualities. Additionally, if gene annotation is available, this module also annotates each variant with the corresponding gene. To call the genotype of each variant, this module counts the k-mers in the sample using KMC3 (Kokot *et al.*, 2017), then it genotypes the variants using MALVA (Denti *et al.*, 2019): an efficient and accurate mapping-free approach for genotyping a set of known SNPs and indels initially developed for human individuals. We extended MALVA to support haploid organisms and high-coverage samples. Finally, the results of each analysis can be visualized as a table (see Suppl. Fig. 1 for an example) or downloaded in VCF format or as a spreadsheet for further analysis.

MALVIRUS is available as a self-hosted web application distributed as a Docker container image that can be installed and run on multiple platforms, from personal laptops to large cloud infrastructures. For user convenience, the application is distributed with a set of precomputed catalog of variants for SARS-CoV-2 based on the assemblies available on GenBank (Sayers *et al.*, 2019), therefore the user can immediately run MALVIRUS on a locally available (*e.g.,* private) viral sample. The precomputed catalogs can be easily updated from the application itself with a single click.

3 Results

As a main test case of the effectiveness of MALVIRUS, we report here the results of an analysis performed on two different runs considering an Illumina sample (SRA accession number `SRR11397721`) and a Nanopore sample (SRA accession number `SRR11397725`) sequenced from an Australian SARS-CoV-2 strain (GISAID Accession ID: `EPI_ISL_416412`). We used as input the `NC_045512.2` reference genome, its annotation, and the set of complete SARS-CoV-2 strains available on GISAID (Shu and McCauley, 2017), accessed on April 28, 2020. More details on the assessment are in the supplementary material.

To validate MALVIRUS we compared its genotype calling against the set of SNPs present in the selected strain obtained by aligning the assembly to the reference sequence using MAFFT and then extracting the SNPs with snp-sites from it, as result we found 9 variants that are also in this strain (Suppl. Sect. 2). Simulating a real case scenario, where the goal is to genotype a newly-sequenced strain, we removed the selected Australian strain from the list of assemblies downloaded from GISAID. From the remaining 9065 strains, MALVIRUS produced a VCF containing 7105 variants, where the same 8 variants, out of the 9 previously established, were called as non-reference for the two Australian samples. MALVIRUS could not genotype the last SNP since it is unique to this specific strain, making it impossible to call from the available knowledge base we created in our experimental settings.

The analysis of each sample completed in 5 minutes requiring less than 3GB of RAM. Such amount of resources is nowadays available on any computer, making MALVIRUS easily usable even on lower end machines. Furthermore the first module of our application (catalog creation) needs to be run only when new strains are available and then reused multiple times.

4 Conclusion

In this article, we presented MALVIRUS, a web application for quickly genotyping viral strains. As shown by our tests, MALVIRUS is able to efficiently and accurately genotype a newly-sequenced SARS-CoV-2 strain both from short (Illumina) and long (Oxford Nanopore) reads. Since MALVIRUS benefits from comprehensive variant catalogs, the constantly increasing number of available strains will broaden the completeness of the current variant knowledge, thus boosting the overall accuracy of our pipeline. Future steps will be devoted to replace snp-sites in order to allow MALVIRUS to genotype not only SNPs but also indels.

Funding

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 872539.

References

Böhmer, M. M. *et al.* (2020). Outbreak of COVID-19 in Germany resulting from a single travel-associated primary case.

Denti, L., Previtali, M., Bernardini, G., Schönhuth, A., and Bonizzoni, P. (2019). MALVA: genotyping by Mapping-free ALlele detection of known VAriants. *iScience*, **18**, 20–27.

Gudbjartsson, D. F. *et al.* (2020). Spread of SARS-CoV-2 in the icelandic population. *New England Journal of Medicine*.

Katoh, K. and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, **30**(4), 772–780.

Kokot, M., Długosz, M., and Deorowicz, S. (2017). KMC 3: counting and manipulating k-mer statistics. *Bioinformatics*, **33**(17), 2759–2761.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**(18), 3094–3100.

McKenna, A. *et al.* (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, **20**(9), 1297–1303.

Page, A. J. *et al.* (2016). SNP-sites: rapid efficient extraction of snps from multi-FASTA alignments. *Microbial genomics*, **2**(4).

Sayers, E. W. *et al.* (2019). GenBank. *Nucleic Acids Research*, **48**(D1), D84–D86.

Shu, Y. and McCauley, J. (2017). GISAID: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance*, **22**(13).

Wilm, A. *et al.* (2012). LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic acids research*, **40**(22), 11189–11201.