Application Note

Genome Detective Coronavirus Typing Tool for rapid identification and characterization of novel coronavirus genomes

Sara Cleemput, Wim Dumon, Vagner Fonseca, Wasim Abdool Karim, Marta Giovanetti, Luiz Carlos Alcantara, Koen Deforce, Tulio de Oliveira

1Emweb bv, Herent, Belgium; 2KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), School of Laboratory Medicine and Medical Sciences, College of Health Sciences, University of KwaZulu-Natal, Durban, South Africa; 3Laboratório de Genética Celular e Molecular, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil; 4Coordenação Geral dos Laboratórios de Saúde Pública/Secretaria de Vigilância em Saúde, Ministério da Saúde, Brasília, Brazil, 5Laboratório de Flavivírus, Instituto Oswaldo Cruz, Fundação Oswaldo Cruz, Rio de Janeiro, RJ, Brazil, 6Centre for the AIDS Programme of Research in South Africa (CAPRISA), Durban, South Africa, 7Department of Global Health, University of Washington, Seattle, USA

*To whom correspondence should be addressed.

Associate Editor: Pier Luigi Martelli
Received on 31 Jan 2020; revised on 17 Feb 2020; accepted on 18 Feb 2020

Abstract

Summary: Genome Detective is a web-based, user-friendly software application to quickly and accurately assemble all known virus genomes from next generation sequencing datasets. This application allows the identification of phylogenetic clusters and genotypes from assembled genomes in FASTA format. Since its release in 2019, we have produced a number of typing tools for emergent viruses that have caused large outbreaks, such as Zika and Yellow Fever Virus in Brazil. Here, we present the Genome Detective Coronavirus Typing Tool that can accurately identify the novel severe acute respiratory syndrome (SARS) related coronavirus (SARS-CoV-2) sequences isolated in China and around the world. The tool can accept up to 2,000 sequences per submission and the analysis of a new whole genome sequence will take approximately one minute. The tool has been tested and validated with hundreds of whole genomes from ten coronavirus species, and correctly classified all of the SARS-related coronavirus (SARSr-CoV) and all of the available public data for SARS-CoV-2. The tool also allows tracking of new viral mutations as the outbreak expands globally, which may help to accelerate the development of novel diagnostics, drugs and vaccines to stop the COVID-19 disease.

Availability: https://www.genomedetective.com/app/typingtool/cov
Contact: koen@emweb.be and deoliveira@ukzn.ac.za
Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction
We are currently faced with a potential global epidemic of a new coronavirus that has infected thousands of people in China and is spreading rapidly around the world. In the end of January 2020, the WHO has declared it a global emergency (WHO, 2020). The novel coronavirus (SARS-CoV-2), first isolated in Wuhan China, has already caused more infections than the previous severe acute respiratory syndrome (SARS) outbreak of 2002 and 2003. The virus is a SARS related coronavirus (SARSr-CoV), and it is genetically associated with SARSr-CoV strains that infect bats in China (Zhu et al., 2020, Lu et al., 2020). It causes severe respiratory illness, which the WHO recently named COVID-19.
The phylogenetic reference dataset was used to create an automated Coronavirus Typing Tool using the Genome Detective framework (Vilsker et al., 2019, Fonseca et al., 2019). To determine the accuracy of this tool, each of the 431 test WGS was considered for evaluation (i.e. 384 reference sequences from VIPR and 47 public SARS-CoV-2 sequences). The sensitivity, specificity and accuracy of our method was calculated for both species assignment and phylogenetic clustering of SARS-CoV. Sensitivity was computed by the formula $\frac{TP}{TP+FN}$, specificity by $\frac{TN}{TN+FP}$ and accuracy by $\frac{TP+TN}{TP+TN+FP+FN}$, where TP = True Positives, FP = False Positives, TN = True Negatives and FN = False Negatives.

Classifying query sequences in an automated fashion involves two steps. The first step enables virus species assignments and the second, which is restricted to SARSr-CoV, includes phylogenetic analysis. The first classification analysis subjects a query sequence to BLAST and AGA analysis. AGA is a novel alignment method for nucleic acid sequences against annotated genomes from NCBI RefSeq Virus Database. AGA (Deforce 2017) expands the optimal alignment algorithms of Smith-Waterman (Smith & Waterman 1981) and Gotoh (Gotoh 1982) based on an induction state with additional parameters. The result is a more accurate aligner, as it takes into account both nucleotide and protein scores and identifies all of the polymorphisms at nucleotide and amino acid levels. In the second step, a query sequence is aligned against the phylogenetic reference dataset using -add alignment option in the MAFFT software (Kato & Standley 2013). In addition, a Neighbor Joining phylogenetic tree is constructed using the HKY distance metric with gamma among-site rate variation with 1,000 bootstrap replicates using PAUP* (Swofford). The query sequence is assigned to a particular phylogenetic cluster if it clusters monophyletically with that clade or a subset of it with bootstrap support >70%. If the bootstrap support is <70%, the genotype is reported to be unassigned.

The result of the phylogenetic and mutational analysis performed by AGA is available in a detailed report. This report contains an interactive phylogenetic tree and genome mapper (Supplementary Figure 1). It also presents the virus species and cluster assignments and a detailed table that provides information about open reading frames (ORFs), CDS and proteins. This table can be expanded to show nucleotide and amino acid mutations that differentiate a query sequence from their species RefSeq or from a sequence in the phylogenetic reference dataset. All results can be exported to a variety of file formats (XML, CSV, Excel, Nexus or FASTA).

3 Testing and Validation
The Genome Detective Coronavirus Typing Tool correctly classified all of the 175 SARSr-CoV sequences at species level, i.e. specificity, sensitivity and accuracy of 100%. Furthermore, all of the 47 SARS-CoV-2 WGS that were isolated in China (n=36), USA (n=5), France (n=2), Thailand (n=3), Japan (n=1) and Taiwan (n=1) were correctly classified at phylogenetic cluster level as SARS-CoV-2, which may be renamed as SARS-B. In addition, we classified with very high specificity, sensitivity and accuracy (i.e. 100%) all of the 112 SARS outbreak WGS of 2002 and 2003. We also achieved perfect classification (i.e. specificity, sensitivity and accuracy of 100%) for all of Beta coronavirus, Human Coronavirus HKU1, MERS-CoV, Rousettus Bat coronavirus HKU9 and Tylonycteris bat coronavirus HKU4 at species level. For a detailed overview of assignment performance, please refer to the Supplementary Table 3.

Our tool also allows detailed analysis of coding regions and proteins for each of the coronavirus species. For example, the analysis of the first released SARS-CoV-2 sequence, the WH_Humani1 China 2019Dec (GenBank: MN908847) demonstrated at genome level, the nucleotide (NT) identity was 79.0% to the reference strain SARS-CoV (ACCESSION: NC_004718.3) and that the Envelop Small Membrane Protein (protein E) is the most similar protein. In total, 94.8% (73/77) of the amino acids were identical; the four amino acid differences were located at positions 55 (T55S), 56 (V56F), 69 (69deletion) and 70 (G70R). The spike protein (protein S), which can be associated with virulence, was 76.2% identical to the reference strain of SARS-CoV (Supplementary Table 4A). Interestingly there were four amino acid insertions at position 237 (A237 F238insHRSY, genome NT position 22202_22203insCATAGAAGTTAT)), which is just upstream from a cleavage site. There is also a four amino acid insertion PRRA at the spike protein at positions 681 to 684 This is at the junction of S1 and S2 and creates a new polybase cleavage site. Our tool also allows us to compare mutations with other related sequences, such as the Pangolin, Bat RaTG13, the Bat SARS-CoV and SARS Sm890 (Figure 2 and supplementary table 2). The most diverse coding regions were the CD8 Sars8a and Sars8b. In these two regions, only 30% of the amino acids were identical. Sars8b protein was truncated early and its CDS had four stop codons (Supplementary Table 4A).

Figure 2: Output from Genome Detective Coronavirus Typing Tool showing: A) SARS-CoV-2 complete genome map. Top bar represents the genome (nucleotide positions 1 to 29,903). Bottom segments represent the open reading frames (ORFs). B) Amino-acid alignment of the spike protein highlighting a four amino acid insertion (PRRA), which creates a new polybase cleavage site (RRAR) for SARS-CoV-2. Amino acid (aa) alignment is compared with four related coronaviruses. The tool also calculates the percentage aa identities with reference to SARS-CoV-2 as shown here for the complete (1,274 aa) spike protein.

Our Coronavirus Typing Tool also allows a query sequence to be analysed against a sequence in the phylogenetic reference dataset. For example, the WH_Human1 China 2019Dec strain (GenBank: MG772934) was one of the Bat-CoV sequences that were most related to n2019-CoV (Lu et al. 2020). The Envelop Small Membrane Protein (protein E) was 100% identical (Supplementary Table 4B). When the SARS-CoV-2 isolated from France (BetaCoV/France/IDF0373/2020) was analysed with our tool and compared with the WH_Human1 China 2019Dec strain (Accession: MN908847), this sequence was 99.9% identical and had only two NT mutations (Supplementary Table 4C). These two differences were located on positions: 22551G>T & 26016G>T), which caused three amino acid mutations (E2 glycoprotein Protein mutation: V354F (22551G>T), sars3a protein mutations: G250V (26016G>T) and sars3b protein mutation: V110F (26016G>T) (Table 3). The analysis of a WGS in FASTA format takes approximately 60 seconds.

DISCUSSION

We developed and released the Genome Detective Coronavirus Typing tool as a free-of-charge resource in the third week of January 2020 in order to help the rapid characterization of COVID-19 infections. This tool allows the analysis of whole or partial viral genomes within minutes. It accepts assembled genomes in FASTA format or raw NGS data in FASTQ format from Illumina, Ion Torrent, PACBIO or Oxford Nanopore Technologies (ONT) can be submitted to the Genome Detective Virus Tool (Vilsker et al., 2019) to automatically assemble the consensus genome prior to executing the Coronavirus Typing Tool. User effort is minimal, and a user can submit multiple FASTA sequences at once.

The tool uses a novel and dynamic aligner, AGA, to allow submitted sequences to be queried against reference genomes, using both nucleotide and amino acid similarity scores. This allows accurate identification of other coronavirus species and the tracking of new viral mutations as the outbreak expands globally. It also performs detailed analysis of the coding regions and proteins. Moreover, it can easily be updated to add new phylogenetic clusters if new outbreaks arise or if the classification nomenclature changes. The tool has been able to correctly classify all the recently released SARS-CoV-2 genomes, as well as all the 2002-2003 SARS outbreak sequences.

In conclusion, the Genome Detective Coronavirus Typing Tool is a web-based and user-friendly software application that allows the identification and characterization of novel coronavirus genomes.

Acknowledgements

Genome data for SARS-CoV-2 made kindly available by National Institute for Communicable Disease Control and Prevention (NICDC) Chinese Center for Disease Control and Prevention (China CDC), Institute of Pathogen Biology, Chinese Academy of Medical Sciences & Peking Union Medical College, Hubei Provincial Center for Disease Control and Prevention, Wuhan Institute of Virology, Chinese Academy of Sciences, National Institute for Viral Disease Control and Prevention, China CDC, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Department of Medical Sciences, Ministry of Public Health, Thailand | That Red Cross Emerging Infectious Diseases - Health Science Centre | Department of Disease Control, Ministry of Public Health, Thailand, Department of Microbiology, Guangdong Provincial Center for Diseases Control and Prevention, Department of Microbiology, Zhejiang Provincial Center for Disease Control and Prevention, Division of Viral Diseases, Centers for Disease Control and Prevention, Centers for Disease Control, R.O.C. (Taiwan) | Centers for Disease Control, R.O.C. (Taiwan), California Department of Public Health | Pathogen Discovery, Respiratory Viruses Branch, Division of Viral Diseases, Centers for Disease Control and Prevention, Arizona Department of Health Services | Pathogen Discovery, Respiratory Viruses Branch, Division of Viral Diseases, Centers for Disease Control and Prevention, Guangdong Provincial Centers for Diseases Control and Prevention | Guangdong Provincial Institute of Public Health, University of Hong Kong-Shenzhen Hospital, Shenzhen, Guangdong, State Key Laboratory of Virology, Wuhan University, Department of Infectious and Tropical Diseases, Bichat Claude Bernard Hospital, Paris | National Reference Center for Viruses of Respiratory Infections, Institut Pasteur, Paris. We would like to acknowledge all of the data contributors.

Funding

Research reported in this publication was supported by a research Flagship grant from the South African Medical Research Council (MRC-RFA-UFSF-01-2013/JUKRI HIVEPI), the VIROGENESIS project, which received funding from the European Union’s Horizon 2020 Research and Innovation Programme (under Grant Agreement no. 634650) and the National Human Genome Research Institute of the National Institutes of Health under Award Number 2U4HG006941. H3ABioNet is an initiative of the Human Health and Heredity in Africa Consortium (H3Africa). The content is solely the responsibility of the
authors and does not necessarily represent the official views of the National Institutes of Health

Conflict of Interest: Dr. Koen Deforche and Wim Dumon are some of the owners of the commercial company, EMWEB. This company has allowed the coronavirus typing tool to openly available on the web.

References


Rambaut A. FigTree v1.4.4. Institute of Evolutionary Biology, University of Edinburgh, Edinburgh. 2018, http://tree.bio.ed.ac.uk/software/figtree/


