Bioinformatics, 2020, 1–2
doi: 10.1093/bioinformatics/btaa1047
Advance Access Publication Date: 21 December 2020
Applications Note

Genome analysis

CorGAT: a tool for the functional annotation of SARS-CoV-2 genomes

Matteo Chiara [1,2,*], Federico Zambelli[1,2], Marco Antonio Tangaro[2], Pietro Mandreoli[1,2], David S. Horner[1,2,] and Graziano Pesole [2,3,*]

[1]Department of Biosciences, University of Milan, 20133 Milan, Italy[2]Institute of Biomembranes, Bioenergetics and Molecular Biotechnology, National Research Council, 70126 Bari, Italy and [3]Department of Biosciences, Biotechnology and Biopharmaceutics, University of Bari "A. Moro", 70126 Bari, Italy

*To whom correspondence should be addressed.
Associate Editor: Alfonso Valencia

Received on August 3, 2020; revised on October 30, 2020; editorial decision on December 6, 2020; accepted on December 9, 2020

Abstract

Summary: While over 200 000 genomic sequences are currently available through dedicated repositories, *ad hoc* methods for the functional annotation of SARS-CoV-2 genomes do not harness all currently available resources for the annotation of functionally relevant genomic sites. Here, we present CorGAT, a novel tool for the functional annotation of SARS-CoV-2 genomic variants. By comparisons with other state of the art methods we demonstrate that, by providing a more comprehensive and rich annotation, our method can facilitate the identification of evolutionary patterns in the genome of SARS-CoV-2.

Availabilityand implementation: Galaxy http://corgat.cloud.ba.infn.it/galaxy; software: https://github.com/matteo14c/CorGAT/tree/Revision_V1; docker: https://hub.docker.com/r/laniakeacloud/galaxy_corgat.

Contact: matteo.chiara@unimi.it or graziano.pesole@uniba.it

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The recent outbreak of COVID-19 has underlined the importance of rapid and effective sharing of molecular data for combating the spread of human pathogens and tracing possible routes of infection. At present, more than 200 000 genomic SARS-CoV-2 sequences have been deposited in dedicated repositories (Shu and McCauley, 2017) along with associated metadata. Harnessing this wealth of information to identify functionally relevant genomic changes and/or recognizing the emergence of novel viral strains is of pivotal importance in the fight against COVID-19. Currently tools for functional annotation of genomic sequences have not been specifically devised for the analysis of SARS-CoV-2, e.g. do not take into account the unusual mechanisms of transcription and post-translational processing of coronaviruses gene products (Sawicki *et al.*, 2007). Additionally, while a wealth of resources and datasets for the fine-grained annotation of functional genomic elements are currently available, including: detailed studies of transcriptional mechanisms (Kim et al., 2020), conserved regulatory sequences (Sawicki *et al.*, 2007), sites under evolutionary selection (http://hyphy.org/covid/), predicted epitopes (Kiyotani et al., 2020) and non-coding secondary structure elements, are available, these are not normally incorporated in the functional annotation of SARS-CoV-2 genomic variants. To overcome these limitations, we propose a novel, highly effective and user friendly approach for the functional annotation of SARS-CoV-2 genomes: CorGAT - the Coronavirus Genome Analysis Tool. By integrating a curated selection of datasets and resources, CorGAT provides a richer and more detailed annotation of SARS-CoV-2 variants when compared with other state of the art methods. To illustrate its advantages, we apply CorGAT to the complete collection of 57 558 currently available SARS-CoV-2 genomic sequences, and derive relevant insights concerning the evolution of this novel pathogen.

2 Implementation

CorGAT has been made available as a collection of Perl script and annotation files at https://github.com/matteo14c/CorGAT/tree/Revision_V1. A user friendly version of the software is available in the form of a standalone Galaxy (Afgan *et al.*, 2018) implementation, based on the Laniakea@ReCaS Galaxy on-demand service (Tangaro et al., 2020) at http://corgat.cloud.ba.infn.it/galaxy. A Docker container image can be obtained from https://hub.docker.

com/r/laniakeacloud/galaxy_corgat. A complete account of the resources used for the annotation of the SARS-CoV-2 genomes, and of their integration in CorGAT is presented in the Supplementary Materials. A detailed user manual is available at https://corgat.readthedocs.io/. Functional annotation files incorporated in CorGAT are updated on a monthly basis, to cope with the constant increase in publicly available data and genomic sequences of SARS-CoV-2. CorGAT has a modular architecture (see Supplementary Materials), allowing the rapid inclusion of novel or even custom types of annotations, simply by editing plain text files.

3 Results

To demonstrate the application of CorGAT, we compared the functional annotation of the complete collection of 20 045 genetic variants derived from 57 558 genomic sequences of SARS-CoV-2 (see Supplementary Materials) by CorGAT, with the annotations by SNPeff (Cingolani *et al.*, 2012) and by the Variant Annotation Integrator (Hinrichs *et al.*, 2016). Simple statistics concerning the number and types of variants are reported in Supplementary Table S1. As outlined in Supplementary Table S2, all the tools herein considered provided highly consistent annotations of functional effects of variants associated with protein coding genes, thus confirming that CorGAT attains the same level of sensitivity as the other methods. However, as illustrated in Supplementary Table S3, CorGAT provides additional layers of annotation that are not provided by other methods, for a total of 14753 single distinct annotations. These include 33 variants associated with regulatory elements (transcription regulatory sequences, TRS), 69 variants associated to consensus cleavage sites (Kiemer *et al.*, 2004) in the ORF1a and ORF1ab polyproteins, 1164 variants associated with sites under selection according to Hyphy (Kosakovsky Pond *et al.*, 2020) and 161 variants in conserved secondary structure elements (Supplementary Table S3). According to our analyses, a highly significant reduction of missense substitutions is observed at sites predicted to be under negative selection (Fisher P-value $< 2.2e\text{-}16$), compared to the background of all the substitutions in protein coding genes. Nevertheless, 229 missense substitutions alter highly conserved amino acid residues that are predicted to be under negative selection. Furthermore, analysis of genetic variants associated with functional non-coding elements in the genome of SARS-CoV-2 highlight some potentially interesting patterns. While the 5′ and 3′ UTRs are the most variable regions of the genome, TRS and secondary structure elements in general are considerably less variable, and show levels of conservation comparable to protein coding genes (Supplementary Table S4). This is well exemplified by the TRS-L element, which is the single most conserved region in the 5′ UTR (Supplementary Fig. S1). Strikingly, the s2m element in the 3′ UTR (Tengs et al., 2013) exhibits more of variability and recurrent indels than other annotated functional elements (Supplementary Tables S4 and S5). Interestingly, our functional annotation (see Supplementary Materials), indicates that several observed substitutions might result in substantial changes to s2m structure consistent with change or loss of s2m function in SARS-CoV-2 (Chiara *et al.*, 2020). We conclude, that CorGAT constitutes a useful addition to the collection of tools for the functional characterization of SARS-CoV-2 genomes.

Acknowledgements

The authors thank ELIXIR-Italy for providing the computing and bioinformatics facilities. We gratefully acknowledge the authors, originating and submitting laboratories of the sequences from GISAID's EpiFlu™ Database on which this research is based.

Funding

This work was supported by the Italian Ministero dell'Università e Ricerca: PRIN 2017, Consiglio Nazionale delle Ricerche, H2020 projects EOSC-Life (GA: 824087), EOSC-Pillar (GA: 857650) and ELIXIR Converge (GA: 871075), and Elixir-IIB.

Conflict of Interest

none declared.

References

Afgan,E. *et al.* (2018) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.*, **46**, W537–W544.

Chiara,C. *et al.* (2020) Comparative genomics provides an operational classification system and reveals early emergence and biased spatio-temporal distribution of SARS-CoV-2. Unpublished data, bioRxiv. doi: 10.1101/2020.06.26.172924. Preprint.

Cingolani,P. *et al.* (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*, **6**, 80–92.

Hinrichs,A.S. *et al.* (2016) UCSC data integrator and variant annotation integrator. *Bioinformatics*, **32**, 1430–1432.

Kiemer,L. *et al.* (2004) Coronavirus 3CLpro proteinase cleavage sites: possible relevance to SARS virus pathology. *BMC Bioinformatics*, **5**, 72.

Kim,D. *et al.* (2020) The Architecture of SARS-CoV-2 Transcriptome. *Cell*, **181**, 914–921.e10.

Kiyotani,K. *et al.* (2020) Bioinformatic prediction of potential T cell epitopes for SARS-Cov-2. *J. Hum. Genet.*, **65**, 569–575.

Kosakovsky Pond,S.L. *et al.* (2020) HyPhy 2.5—a customizable platform for evolutionary hypothesis testing using phylogenies. *Mol. Biol. Evol.*, **37**, 295–299.

Sawicki,S.G. *et al.* (2007) A contemporary view of coronavirus transcription. *J. Virol.*, **81**, 20–29.

Shu,Y. and McCauley,J. (2017) GISAID: global initiative on sharing all influenza data – from vision to reality. *Euro Surveill.*, **22**, 30494.

Tangaro,M.A. *et al.* (2020) Laniakea: an open solution to provide Galaxy "on-demand" instances over heterogeneous cloud infrastructures. *Gigascience*, **9**, giaa033.

Tengs,T. *et al.* (2013) A mobile genetic element with unknown function found in distantly related viruses. *Virol. J.*, **10**, 132.