A global pandemic of the coronavirus disease COVID-19, a severe respiratory illness caused by a novel virus from the family Coronaviridae (SARS-CoV-2), has infected millions and caused hundreds of thousands of deaths (World Health Organization 2020a). COVID-19 (SARS-CoV-2), has infected millions and caused hundreds of respiratory illness caused by a novel virus from the family

A Comprehensive, Flexible Collection of SARS-CoV-2 Coding Regions

Dae-Kyum Kim,*††† Jennifer J. Knapp,*††† Da Kuang,*††† Aditya Chawla,*††† Patricia Cassonnet,†*‡‡‡ Hunsang Lee,*† Dayag Sheyhkarimili,*†,††† Payman Samavarchi-Tehrani,†† Hala Abdouni,††† Ashyad Rayhan,*††† Roujia Li,*††† Oxana Pogoutse,*††† Étienne Coyaud,†† Sylvie van der Werf,††† Dae-Kyum Kim,*††† Caroline Demeret,††† Anne-Claude Gingras,†‡‡ Brian Raught,*** Yves Jacob,§§§††† and Frederick P. Roth†††,‡‡‡‡

*Donnelly Centre, 1Department of Molecular Genetics, 2Department of Medical Biophysics, Princess Margaret Cancer Centre, 1111Department of Computer Science, University of Toronto, Ontario, Canada M5S 2E4, 4Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Ontario, Canada M5G 1X5, 5Unité de Génétique Moléculaire des Virus à ARN, Département Virologie, Institut Pasteur, Paris, France 75015, **UMR3569, Centre National de la Recherche Scientifique, Paris, France, 11Université de Paris, France, 75016 12Univ. Lille, Inserm, CHU Lille, U1192 - Protéomique Réponse Inflammatoire Spectrométrie de Masse - PRISM, F-59000 Lille, France, and 58Molecular Architecture of Life Program, Canadian Institute for Advanced Research, Toronto, Ontario, Canada M5G 1M1

ORCID ID: 0000-0003-3554-0464 (D.K.)

ABSTRACT The world is facing a global pandemic of COVID-19 caused by the SARS-CoV-2 coronavirus. Here we describe a collection of codon-optimized coding sequences for SARS-CoV-2 cloned into Gateway-compatible entry vectors, which enable rapid transfer into a variety of expression and tagging vectors. The collection is freely available. We hope that widespread availability of this SARS-CoV-2 resource will enable many subsequent molecular studies to better understand the viral life cycle and how to block it.

KEYWORDS SARS-CoV-2 coding sequence collection Gateway-compatible TEV (tobacco etch virus) sequence

Copyright © 2020 Kim et al. doi: https://doi.org/10.1534/g3.120.401554

Manuscript received July 10, 2020; accepted for publication July 22, 2020; published Early Online August 6, 2020.

This is an open-access article published under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Supplemental material available at figshare: https://doi.org/10.25387/g3.12725096.

†These authors contributed equally to this work.

‡Corresponding authors: Yves Jacob. Dept de Virologie- Institut Pasteur-28 rue du Dr Roux –75724. Phone: (33) 1 45 68 87 53. E-mail: yves.jacob@pasteur.fr. Frederick P. Roth. Donnelly Centre, 160 College Street, Toronto, Ontario, Canada M5S 3E1. Phone: (1) 416-946-5130. E-mail: fritz.roth@utoronto.ca.
ORF1AB, a large polyprotein which is post-translationally processed into 16 proteins (Chan et al. 2020). More recently, Wu et al. discovered two additional viral ORFs (ORF9Bwu and ORF10wu) with unclear functions (Wu et al. 2020). Progress on molecular characterization has been made on several viral proteins (Walls et al. 2020; Zhang et al. 2020), providing valuable insights into host-virus interaction, but more research is necessary. The Gateway system offers efficient and high-throughput transfer of the viral coding sequences (CDSs) into a large selection of Gateway-compatible destination vectors used for protein expression in many biological systems, e.g., Escherichia coli, Saccharomyces cerevisiae, insect, or mammalian cells (Walhout et al. 2000). Broad availability of a collection of SARS-CoV-2 CDSs has the potential to enable many downstream biochemical and structural studies and thus a better understanding of processes within the viral life cycle, including scalable assays for screening drug candidates that could potentially disrupt these processes.

### MATERIALS AND METHODS

#### Synthesis of viral coding sequences

Based on the published annotation of the genome sequence of the HKU-SZ-005b (GenBank MN975262; Chan et al. 2020) and Wuhan-Hu-1 (GenBank MN908947; Wu et al. 2020) isolates of SARS-CoV-2, we requested the synthesis of viral coding sequences (GenScript and Integrated DNA Technologies), including termination codons and

| Gene Symbol | CDS Name                  | Putative Function/Domain                                      | AA Length | Clone Status |
|-------------|---------------------------|--------------------------------------------------------------|-----------|--------------|
| ORF1AB      | NSP1                      | Suppress antiviral host response                             | 180       | ✓ ✓ ✓        |
|             | NSP2                      | Unknown                                                      | 639       | ✓ ✓ ✓        |
|             | NSP3                      | Putative PL-pro domain                                       | 1,946     | ✓ ✓ ✓        |
|             | NSP3-Cys857Ala            | Putative PL-pro domain (with Cys857Ala variant)              | 1,946     | ✓ ✓ NA       |
|             | NSP4                      | Complex with NSP3 & 6 for DMV (double-membrane vesicle)      | 501       | ✓ ✓ ✓        |
|             | NSP5                      | 3CL-pro domain                                               | 307       | ✓ ✓ ✓        |
|             | NSP5-Cys146Ala            | 3CL-pro domain (with Cys146Ala variant)                      | 307       | ✓ ✓ NA       |
|             | NSP6                      | Complex with NSP 3 & 4 for DMV formation                      | 291       | ✓ ✓ ✓        |
|             | NSP7                      | DNA primase subunit                                          | 84        | ✓ ✓ ✓        |
|             | NSP8                      | DNA primase subunit                                          | 199       | ✓ ✓ ✓        |
|             | NSP9                      | RNA/DNA binding activity                                     | 114       | ✓ ✓ ✓        |
|             | NSP10                     | Complex with NSP14: Replication fidelity                     | 140       | ✓ ✓ ✓        |
|             | NSP12                     | RNA-dependent RNA polymerase                                 | 919       | ✓ ✓ ✓        |
|             | NSP13                     | Helicase                                                     | 602       | ✓ ✓ ✓        |
|             | NSP14                     | ExoN: 3′-5′ exonuclease                                       | 528       | ✓ ✓ ✓        |
|             | NSP15                     | XendoU: poly(U)-specific endoribonuclease                    | 347       | ✓ ✓ ✓        |
|             | NSP16                     | 2′-O′-MT: 2′-O-ribo methyltransferase                        | 299       | ✓ ✓ ✓        |
| S           | S                         | Spike glycoprotein trimer that binds to host cell receptors (e.g., ACE2) | 1,273    | ✓ ✓ ✓        |
|             | S-24nt                    | Spike glycoprotein trimer (minus 8 amino acids)              | 1,265     | ✓ ✓ NA       |
|             | S-frag1                   | Entire Ectodomain                                            | 1,213     | NA NA NA     |
|             | S-frag2                   | Entire Ectodomain without the signal peptide                | 1,199     | NA NA NA     |
|             | S-frag3                   | N-term fragment after the furin cleavage                     | 686       | NA NA NA     |
|             | S-frag4                   | N-term fragment after the furin cleavage without the signal peptide | 672 | NA NA NA |
|             | S-frag5                   | C-terminal Ectodomain from the furin cleavage site           | 528       | NA NA NA     |
|             | S-frag6                   | C-terminal Ectodomain from the Tmpress 2 priming site        | 399       | NA NA NA     |
| ORF3A       | 3A                        | Induce inflammatory response and apoptosis                   | 275       | ✓ ✓ ✓        |
| ORF3B       | 3B                        | Induce inflammatory response and inhibit the expression of IFNβ | 58        | ✓ ✓ ✓        |
| E           | E                         | Envelope protein pentamer                                    | 75        | ✓ ✓ ✓        |
|             | E-27nt                    | Envelope protein pentamer (minus 9 amino acids)              | 66        | ✓ ✓ NA       |
| M           | M                         | Membrane protein                                             | 222       | ✓ ✓ ✓        |
| ORF6        | 6                         | Antagonize STAT1 function and IFN signaling, and induce DNA synthesis | 61 | ✓ ✓ ✓ |
| ORF7A       | 7A                        | Induce inflammatory response and apoptosis                   | 121       | ✓ ✓ ✓        |
| ORF7B       | 7B                        | Induce inflammatory response                                 | 43        | ✓ ✓ ✓        |
| ORF7B       | 7B-trunc                  | Induce inflammatory response (with N terminus truncated)     | 20        | ✓ ✓ NA       |
| ORF8        | 8                         | Induce apoptosis and DNA synthesis                           | 121       | ✓ ✓ ✓        |
| N           | N                         | Facilitate viral RNA packaging                               | 419       | ✓ ✓ ✓        |
| ORF9B       | 9B                        | Induce apoptosis                                             | 98        | ✓ ✓ ✓        |
| ORF9Bwu     | 9Bwu                      | Unknown                                                      | 73        | ✓ ✓ NA       |
| ORF10wu     | 10wu                      | Unknown                                                      | 38        | ✓ ✓ NA       |

✓ indicates that clone is available; NA indicates that the clone was not available at the time of this writing.
**RESULTS AND DISCUSSION**

A total of 98 clones (Table 1) are currently included in the Gateway-compatible collection, covering 28 out of 29 total an-notated CDSSs in the SARS-CoV-2 genome. **NSP11** was omitted due to the incompatibility of its 36 base pair length with the Gateway cloning system (Cheo et al. 2004). All 28 of these CDSS regions are available as clones with and without termination codons. The ‘no-stop’ collection was further extended to include six clones encoding different cleaved products of the spike (S) protein — “S-fragment” 1–6. We also included two CDSS variants with in-frame deletions (“S-24nt” and “E-27nt”), one truncated CDSS variant (“ORF8B-truncated”), that were each detected by recent viral transcriptome mapping efforts (Davidson et al. 2020, Kim et al. 2020) and two missense catalytic variants (NSP3 C857A and NSP5 C146A; Gordon et al. 2020).

Although our collection facilitates tagging of SARS-CoV-2 pro-teins for various functional studies, certain applications require re-moveal of tags at some stage, for example, after protein purification. Fusion proteins can potentially interfere with the yield, structure, and function of purified proteins, such as during large scale production and crystallography studies (Booth et al. 2018). To address this we expanded our collection to include clones containing an N-terminal recognition sequence for the nuclear inclusion protease from tobacco etch virus (TEV; Carrington and Dougherty 1987; Carrington and Dougherty 1988). The TEV sequence is one of the best characterized and widely used endoproteolytic reagents due to its stringent se-quence specificity, ease of production, and ability to tolerate a variety of residues at the P1’ position of its recognition site (Waugh 2011). We note that our clones are not expression vectors in and of themselves, and we have not yet assessed the expression of any of our clones after moving to a Gateway Destination expression vector. However, we note that our Gateway-compatible collection allows users the flexibility to conveniently move any of the SARS-CoV-2 ORFs into any Gateway Destination expression vector with any preferred N-terminal or C-terminal fusion.

To promote open-access dissemination of the collection, all clones have been deposited to the non-profit organization Addgene (Kamens 2015), and are freely available from the authors under circumstances where Addgene cannot be used. Table S2 summarizes all CDSSs in the collection, together with their nucleotide sequences, nucleotide and amino acid lengths and links for ordering clones. We hope that this SARS-CoV-2 CDS-clone collection will be a valuable resource for many applications, including study of how coronaviruses can exploit cellular processes for the viral replication cycle (de Wilde et al. 2018), understanding virus-host protein-protein interactions (Gordon et al. 2020; Lasso et al. 2019), production of recombinant virus proteins for structural studies (Edavellet et al. 2012), mapping of protein subcellular localization using N-terminal fluorescent reporters (Tanz et al. 2013), or development of vaccines or other therapeutics (Jing et al. 2012; McDonald et al. 2007).

**ACKNOWLEDGMENTS**

This work was supported by a Canadian Institutes for Health Re-search Foundation Grant (F.P.R.); the Canada Excellence Research Chairs Program (F.P.R.); the Thistledown Foundation (B.R., F.P.R. and A.-C.G.); the LabEx Integrative Biology of Emerging Infectious Diseases (10-LABX-0062; Y.J.) and Platform for European Preparedness Against (Re-)emerging Epidemics, EU (602525; Y.J.).

**LITERATURE CITED**

Booth, W. T., C. R. Schlachter, S. Pote, N. Ussin, N. J. Mank et al., 2018 Impact of an N-terminal polyhistidine tag on protein thermal stability. ACS Omega 3: 760–768. https://doi.org/10.1021/acsomega.7b01598

Carrington, J. C., and W. G. Dougherty, 1987 Small nuclear inclusion protein encoded by a plant potyvirus genome is a protease. J. Virol. 61: 2540–2548. https://doi.org/10.1128/JVI.61.8.2540-2548.1987

Carrington, J. C., and W. G. Dougherty, 1988 A viral cleavage site cassette: Identification of amino acid sequences required for tobacco etch virus polyprotein processing. Proc. Natl. Acad. Sci. USA 85: 3391–3395. https://doi.org/10.1073/pnas.85.10.3391

Chan, J. F. W., K. H. Kok, Z. Zhu, H. Chu, K. K. W. To et al., 2020 Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. Emerg. Microbes Infect. 9: 221–236. https://doi.org/10.1080/22221751.2020.1719902
