INVESTIGATION

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

Dae-Kyum Kim,[*,†,‡,1] Jennifer J. Knapp,[*,†,‡,1] Da Kuang,[*,†,‡,1] Aditya Chawla,[*,†,‡] Patricia Cassonnet,[§,,††] Hunsang Lee,[*,†] Dayag Sheykhkarimli,[*,†,‡] Payman Samavarchi-Tehrani,[‡] Hala Abdouni,[‡] Ashyad Rayhan,[*,†,‡] Roujia Li,[*,†,‡] Oxana Pogoutse,[*,†,‡] Étienne Coyaud,[‡‡] Sylvie van der Werf,[§,**,††] Caroline Demeret,[§,**,††] Anne-Claude Gingras,[†,‡] Mikko Taipale,[*,†,§§] Brian Raught,[***] Yves Jacob,[§,**,††,2] and Frederick P. Roth[*,†,‡,†††,2]**

[*]Donnelly Centre, [†]Department of Molecular Genetics, [***]Department of Medical Biophysics, Princess Margaret Cancer Centre, [†††]Department of Computer Science, University of Toronto, Ontario, Canada M5S 2E4, [‡]Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Ontario, Canada M5G 1X5, [§]Unité de Génétique Moléculaire des Virus à ARN, Département Virologie, Institut Pasteur, Paris, France 75015, [**]UMR3569, Centre National de la Recherche Scientifique, Paris, France, [††]Université de Paris, France, 75016 [‡‡]Univ. Lille, Inserm, CHU Lille, U1192 - Protéomique Réponse Inflammatoire Spectrométrie de Masse - PRISM, F-59000 Lille, France, and [§§]Molecular 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

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 manifestation in patients can range from a lack of symptoms to severe pneumonia and death (Huang *et al.* 2020). Person-to-person spread through respiratory droplets has been identified as a major source of transmission of the virus (Yu *et al.* 2020). Various measures, from social distancing to nationwide lockdowns, have been imposed to contain and control the transmission of SARS-CoV-2 (Cohen and Kupferschmidt 2020). Despite these measures, the number of confirmed COVID-19 cases has continued to rise (World Health Organization 2020a), highlighting the need for an effective vaccine and antiviral agents. Furthermore, the extrapolations concerning the evolution of the pandemic are particularly alarming (Ferguson *et al.* 2020). It is therefore of intense and pressing interest to better understand this virus and its interaction with host cells on a molecular level.

Shortly after the outbreak, the complete genome of two SARS-CoV-2 strains were published (Chan *et al.* 2020; Wu *et al.* 2020). Using the genome sequence as a reference, Chan *et al.* identified 12 viral open reading frames (ORFs), including one encoding

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 distributed 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.

[1] These authors contributed equally to this work.

[2] 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

Table 1 The genome-scale SARS-CoV-2 coding sequence clone collection

				Clone Status		
Gene Symbol	CDS Name	Putative Function/Domain	AA Length	STOP	NO STOP	TEV
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) formation	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	S-24nt	Spike glycoprotein trimer (minus 8 amino acids)	1,265	✓	✓	NA
S	S-frag1	Entire Ectodomain	1,213	NA	✓	NA
S	S-frag2	Entire Ectodomain without the signal peptide	1,199	NA	✓	NA
S	S-frag3	N-term fragment after the furin cleavage	686	NA	✓	NA
S	S-frag4	N-term fragment after the furin cleavage without the signal peptide	672	NA	✓	NA
S	S-frag5	C-terminal Ectodomain from the furin cleavage site	528	NA	✓	NA
S	S-frag6	C-terminal Ectodomain from the Tmpress 2 priming site	399	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	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 the time of this writing.

attB recombination sequences, with optimization of codon usage to reduce GC content and optimize expression in human and insect cells. A start codon was added to *NSP2–16* to allow independent transcription and translation, as the endogenous products are derived from *ORF1AB* by post-translational processing. *ORF9Bwu*, an alternative ORF within the *N* gene from SARS-COV-2 (Wu *et al.* 2020), was subsequently amplified by polymerase chain reaction (PCR) from the viral *N* gene with primers listed in Table S1.

Generation of Gateway-compatible viral coding sequence clone collections

Synthesized viral coding sequences were incorporated into Gateway Entry plasmids: either pDONR207 (Invitrogen Cat #12213013) or pDONR223 (Rual *et al.* 2004). To enable C-terminal fusion constructs, we also generated an equivalent set of Gateway-compatible clones without termination codons. These clones were made by either PCR-amplifying the whole plasmid with primers that eliminated the stop codon, or by amplifying CDS regions from the first collection, using downstream primers with complementary regions that were internal to each stop codon, and which simultaneously incorporated the flanking sequences necessary for incorporation into a Gateway Entry plasmid [pDONR207, pDONR221 (Invitrogen Cat #12536017) or pDONR223].

Expression clones with N-terminal fusion tags (*e.g.*, for purification) can be produced simply by preparing the appropriate Gateway-compatible Destination vector. However, to enable the subsequent removal of such N-terminal fusion tags, we generated an additional set of clones containing, at the N-terminus of the ORF, a recognition sequence for nuclear inclusion protease from tobacco etch virus (TEV). TEV sequences were incorporated by amplifying CDS regions from the first collection using forward primers that also provide TEV sequences with the original reverse primers.

Each SARS-CoV-2 CDS bacterial clone (DH5α *E. coli* strain, NEB Cat# C2987) was isolated from a single colony, and its inserted CDS was confirmed by full-length Sanger sequencing (The Centre for Applied Genomics, Toronto, Canada). All clones with a pDONR221 or pDONR223 backbone were sequenced with M13F and M13R primers. Clones with a pDONR207 backbone were sequenced with customized forward and reverse primers. All primer sequences are available in Table S1.

Data availability

Clones are available through Addgene. Table S1 contains all primers used. Table S2 contains detailed descriptions of clones in the collection and links to the clone resource available from Addgene. Supplemental material available at figshare: https://doi.org/10.25387/g3.12725096.

RESULTS AND DISCUSSION

A total of 98 clones (Table 1) are currently included in the Gateway-compatible collection, covering 28 out of 29 total annotated CDSs 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 CDS 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 CDS variants with in-frame deletions ("S-24nt" and "E-27nt"), one truncated CDS 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 proteins for various functional studies, certain applications require removal 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 sequence 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 CDSs 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 host 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 (Edavettal *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 Research 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

Cheo, D. L., S. A. Titus, D. R. Byrd, J. L. Hartley, G. F. Temple *et al.*, 2004 Concerted assembly and cloning of multiple DNA segments using in vitro site-specific recombination: functional analysis of multi-segment expression clones. Genome Res. 14: 2111–2120. https://doi.org/10.1101/gr.2512204

Cohen, J., and K. Kupferschmidt, 2020 Countries test tactics in "war" against COVID-19. Science 367: 1287–1288. https://doi.org/10.1126/science.367.6484.1287

Davidson, A. D., M. K. Williamson, S. Lewis, D. Shoemark, M. W. Carroll *et al.*, 2020 Characterisation of the transcriptome and proteome of SARS-CoV-2 reveals a cell passage induced in-frame deletion of the furin-like cleavage site from the spike glycoprotein. Genome Med. 12: 68. https://doi.org/10.1186/s13073-020-00763-0

de Wilde, A. H., E. J. Snijder, M. Kikkert, and M. J. van Hemert, 2018 Host Factors in Coronavirus Replication, pp. 1–42 in *Roles of host gene and non-coding RNA expression in virus Infection*, edited by Tripp, E., and S. Tompkins. Springer, Cham.

Edavettal, S. C., M. J. Hunter, and R. V. Swanson, 2012 Genetic construct design and recombinant protein expression for structural biology. Methods Mol. Biol. 841: 29–47. https://doi.org/10.1007/978-1-61779-520-6_2

Ferguson, N., D. Laydon, G. G. Nedjati, N. Imai, K. Ainslie *et al.*, 2020 Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand. doi: 10.25561/77482 (Posted March 16, 2020).https://doi.org/10.25561/77482

Gordon, D. E., G. M. Jang, M. Bouhaddou, J. Xu, K. Obernier *et al.*, 2020 A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature 583: 459–468. https://doi.org/10.1038/s41586-020-2286-9

Huang, C., Y. Wang, X. Li, L. Ren, J. Zhao *et al.*, 2020 Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 395: 497–506. https://doi.org/10.1016/S0140-6736(20)30183-5

Jing, L., J. Haas, T. M. Chong, J. J. Bruckner, G. C. Dann *et al.*, 2012 Cross-presentation and genome-wide screening reveal candidate T cells antigens for a herpes simplex virus type 1 vaccine. J. Clin. Invest. 122: 654–673. https://doi.org/10.1172/JCI60556

Kamens, J., 2015 The Addgene repository: an international nonprofit plasmid and data resource. Nucleic Acids Res. 43: 1152–1157. https://doi.org/10.1093/nar/gku893

Kim, D., J. Y. Lee, J. S. Yang, J. W. Kim, V. N. Kim *et al.*, 2020 The architecture of SARS-CoV-2 transcriptome. Cell 181: 914–921.e10. https://doi.org/10.1016/j.cell.2020.04.011

Lasso, G., S. V. Mayer, E. R. Winkelmann, T. Chu, O. Elliot *et al.*, 2019 A structure-informed atlas of human-virus interactions. Cell 178: 1526–1541. https://doi.org/10.1016/j.cell.2019.08.005

McDonald, W. F., J. W. Huleatt, H. G. Foellmer, D. Hewitt, J. Tang *et al.*, 2007 A West Nile virus recombinant protein vaccine that coactivates innate and adaptive immunity. J. Infect. Dis. 195: 1607–1617. https://doi.org/10.1086/517613

Rual, J. F., T. Hirozane-Kishikawa, T. Hao, N. Bertin, S. Li *et al.*, 2004 Human ORFeome version 1.1: a platform for reverse proteomics. Genome Res. 14: 2128–2135. https://doi.org/10.1101/gr.2973604

Tanz, S. K., I. Castleden, I. D. Small, and A. H. Millar, 2013 Fluorescent protein tagging as a tool to define the subcellular distribution of proteins in plants. Front. Plant Sci. 4: 214. https://doi.org/10.3389/fpls.2013.00214

Walhout, A. J. M., G. F. Temple, M. A. Brasch, J. L. Hartley, M. A. Lorson *et al.*, 2000 Gateway recombinational cloning: Application to the cloning of large numbers of open reading frames or ORFeomes. Methods Enzymol. 328: 575–592. https://doi.org/10.1016/S0076-6879(00)28419-X

Walls, A. C., Y. J. Park, M. A. Tortorici, A. Wall, A. T. McGuire *et al.*, 2020 Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. Cell 181: 281–292.e6. https://doi.org/10.1016/j.cell.2020.02.058

Waugh, S. D., 2011 An overview of enzymatic reagents for the removal of affinity tags. Protein Expr. Purif. 80: 283–293. https://doi.org/10.1016/j.pep.2011.08.005

World Health Organization, 2020a COVID-19 situation reports.

Wu, F., S. Zhao, B. Yu, Y. M. Chen, W. Wang *et al.*, 2020 A new coronavirus associated with human respiratory disease in China. Nature 579: 265–269. https://doi.org/10.1038/s41586-020-2008-3

Yu, P., J. Zhu, Z. Zhang, and Y. Han, 2020 A familial cluster of infection associated with the 2019 novel coronavirus indicating possible person-to-person transmission during the incubation period. J. Infect. Dis. 221: 1757–1761. https://doi.org/10.1093/infdis/jiaa077

Zhang, L., D. Lin, X. Sun, U. Curth, C. Drosten *et al.*, 2020 Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. Science 368: 409–412.

Communicating editor: J. Hesselberth