A flexible genome-scale SARS-CoV-2 clone resource

Dae-Kyum Kim[1,2,3,11], Jennifer J. Knapp[1,2,3,11], Da Kuang[1,2,3,11], Aditya Chawla[1,2,3], Patricia Cassonnet[4,5,6], Hunsang Lee[1,2], Dayag Sheykhkarimli[1,2,3], Payman Samavarchi-Tehrani[3], Hala Abdouni[3], Ashyad Rayhan[1,2,3], Oxana Pogoutse[1,2,3], Étienne Coyaud[7], Sylvie van der Werf[4,5,6], Caroline Demeret[4,5,6], Anne-Claude Gingras[2,3], Mikko Taipale[1,2,8], Brian Raught[9], Yves Jacob[4,5,6,*], Frederick P. Roth[1,2,3,10,*]

[1] Donnelly Centre, University of Toronto, Toronto, Ontario, M5S 3E1, Canada

[2] Department of Molecular Genetics, University of Toronto, Toronto, Ontario, M5S 1A8, Canada

[3] Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Ontario, M5G 1X5, Canada

[4] Unité de Génétique Moléculaire des Virus à ARN, Département Virologie, Institut Pasteur, Paris, 75724, France

[5] UMR3569, Centre National de la Recherche Scientifique, Paris, 75015, France

[6] Université de Paris, Paris, 75015, France

[7] Univ. Lille, Inserm, CHU Lille, U1192 - Protéomique Réponse Inflammatoire Spectrométrie de Masse - PRISM, F-59000 Lille, 59655, France

[8] Molecular Architecture of Life Program, Canadian Institute for Advanced Research, Toronto, Ontario, M5G 1M1, Canada

[9] Department of Medical Biophysics, Princess Margaret Cancer Centre, University of Toronto, Toronto, Ontario, M5G 2C1, Canada

[10] Department of Computer Science, University of Toronto, Toronto, Ontario, M5T 3A1, Canada

[11] These authors contributed equally.

* Correspondence: yves.jacob@pasteur.fr (Y.J), fritz.roth@utoronto.ca (F.P.R.)

Keywords: SARS-CoV-2, coding sequence collection, Gateway-compatible

Summary

The world is facing a major health crisis, the global pandemic of COVID-19 caused by the SARS-CoV-2 coronavirus, for which no approved antiviral agents or vaccines are currently available. 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 via Addgene. 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.

Introduction

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 asymptomatic (no symptoms) to severe pneumonia and death (Huang et al., 2020). Early analysis of the outbreak in China outlines symptoms that commonly include fever, dry cough, shortness of breath and myalgia (World Health Organization, 2020b). Person-to-person spread through respiratory droplets has been identified as a major source of transmission of the virus (Yu et al., 2020). To limit contagion, 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*. (Chan et al., 2020) identified 12 viral open reading frames (ORFs), including ORF1ab, a large polyprotein which is post-translationally processed into 16 proteins. 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. However, 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, possibly yielding scalable assays for screening drug candidates that could disrupt these processes.

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

Gene Symbol	**CDS Name**	**Putative Function/Domain**	**AA Length**	**Clone Status**		
				STOP	NO STOP	TEV
ORF1AB	NSP1	Suppress antiviral host response	180	✓	✓	✓
	NSP2	Unknown	639	✓	✓	✓
	NSP3	Putative PL-pro domain	1,946	✓	✓	✓
	NSP4	Complex with NSP3 & 6 for DMV (double-membrane vesicle) formation	501	✓	✓	✓
	NSP5	3CL-pro domain	307	✓	✓	✓
	NSP6	Complex with NSP 3 & 4 for DMV formation	291	✓	✓	✓
	NSP7	DNA primase subunits	84	✓	✓	✓
	NSP8		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 signalling, 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 currently at Addgene; NA indicates that clones is not available in the collection.

Results and Discussion

A total of 94 clones 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 its 36 base pair length, which makes it incompatible with the Gateway cloning system (ThermoFisher). All 28 of these CDS regions are available in 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) and one truncated CDS variant (ORF8B-truncated) that were detected by recent viral transcriptome mapping efforts (Davidson et al., 2020; Kim 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. To address this we have expanded our collection to include clones containing an N-terminal recognition sequence for 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).

To promote open-access dissemination of the collection, all clones have been deposited to Addgene (Kamens, 2015). Table S2 summarizes all CDSs in the collection, together with their nucleotide sequences, nucleotide and amino acid lengths and direct links to Addgene.

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).

Acknowledgements

This work was supported by Canadian Institutes for Health Research Foundation Grant (F.P.R.); the Canada Excellence Research Chairs Program (F.P.R.); 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.). D.-K.K. was supported by a Banting Postdoctoral Fellowship through the Natural Sciences and Engineering Research Council of Canada and by the Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Education (2017R1A6A3A03004385).

Authors Contribution

Conceptualization, D.-K.K., B.R., Y.J., and F.P.R.; Methodology, D.-K.K., J.J.K., H.L., D.S., M.T., and F.P.R.; Investigation, D.-K.K., J.J.K., A.C., P.C., H.L., D.S., P.S.-T., H.A., A.R., and O.P.; Writing – original draft, D.-K.K., D.K. and D.S.; Writing – review & editing, D.-K.K., J.J.K., D.K., H.L., A.R., Y.J., and F.P.R.; Supervision, É.C., S.V.D.W., C.D., A.-C.G., M.T., B.R., Y.J., and F.P.R.; Funding acquisition, D.-K.K., A.-C.G., B.R., Y.J., and F.P.R.

Declaration of Interests: The authors declare no competing interests.

STAR METHODS

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Frederick P. Roth (fritz.roth@utoronto.ca).

Materials Availability

Plasmids generated in this study have been deposited to Addgene (see Table S2 for links), available with a completed Materials Transfer Agreement.

Data and Code Availability

This study did not generate any unique datasets or code.

METHOD DETAILS

Synthesis of viral coding sequences

Based on the published annotation of the genome sequence of the HKU-SZ-005b (Chan et al., 2020) and Wuhan-Hu-1 (Wu et al., 2020) isolates of SARS-CoV-2, we requested the synthesis of viral coding sequences (GenScript, IDT), including termination codons and *attB* recombination sequences, with optimization of codon usage to reduce GC content and optimize expression in human and insect cells. A start codon is added to NSP2–16 to allow independent transcription, as they are prophetically cleaved from ORF1 post translation in host cells. ORF9Bwu, an alternative ORF within the N gene from the 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 then incorporated into Gateway Entry plasmids: either pDONR207 (Invitrogen) 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 and pDONR223).We further expanded our collection to include clones containing N-terminal recognition sequence for nuclear inclusion protease from tobacco etch virus (TEV) to enable the *ad-hoc* removal of fusion tags. TEV sequences were incorporated by amplifying CDS regions from the first collection using forward primers containing TEV sequences and original reverse primers.

Each SARS-CoV-2 CDS bacterial clone was isolated from a single colony, and its inserted CDS was confirmed by full-length Sanger sequencing (TCAG DNA sequencing facility, 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.

Supplementary information

Table S1. Primers used for amplifying and sanger sequencing viral coding sequences.

Table S2. Clones in the genome-scale SARS-CoV-2 coding sequence collection, together with their nucleotide and amino acid lengths, coding sequence and direct links to Addgene.

References

Carrington, J.C. and Dougherty, W.G. (1987). Small nuclear inclusion protein encoded by a plant potyvirus genome is a protease. J. Virol. *61*, 2540–2548.

Carrington, J.C. and Dougherty, W.G. (1988). A Viral Cleavage Site Cassette: Identification of Amino Acid Sequences Required for Tobacco Etch Virus Polyprotein Processing. Proc. Natl. Acad. Sci. U.S.A. *85*, 3391–3395.

Chan, J.F.-W., Kok, K.-H., Zhu, Z., Chu, H., To, K.K.-W., Yuan, S., and Yuen, K.-Y. (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.

Cohen, J., and Kupferschmidt, K. (2020). Countries test tactics in “war” against COVID-19. Science *367*, 1287–1288.

Davidson, A.D., Williamson, M.K., Lewis, S., Shoemark, D., Carroll, M.W., Heesom, K., Zambon, M., Ellis, J., Lewis, P.A., Hiscox, J.A., et al. (2020). Characterisation of the transcriptome and proteome of SARS-CoV-2 using direct RNA sequencing and tandem mass spectrometry reveals evidence for a cell passage induced in-frame deletion in the spike glycoprotein that removes the furin-like cleavage site. https://doi.org/10.1101/2020.03.22.002204.

Edavettal, S.C., Hunter, M.J., and Swanson, R.V. (2012). Genetic construct design and recombinant protein expression for structural biology. Methods Mol. Biol. *841*, 29–47.

Ferguson, N., Laydon, D., Nedjati Gilani, G., Imai, N., Ainslie, K., Baguelin, M., Bhatia, S., Boonyasiri, A., Cucunuba Perez, Z., Cuomo-Dannenburg, G., et al. (2020). Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand. https://doi.org/10.25561/77482.

Gordon, D.E., Jang, G.M., Bouhaddou, M., Xu, J., Obernier, K., White, K.M., O’Meara, M.J., Rezelj, V.V., Guo, J.Z., Swaney, D.L., et al. (2020). A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature. https://doi.org/10.1038/s41586-020-2286-9

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., et al. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet *395*, 497–506.

Jing, L., Haas, J., Chong, T.M., Bruckner, J.J., Dann, G.C., Dong, L., Marshak, J.O., McClurkan, C.L., Yamamoto, T.N., Bailer, S.M., 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.

Kamens, J. (2015). The Addgene repository: an international nonprofit plasmid and data resource. Nucleic Acids Res. *43*, D1152–D1157.

Kim, D., Lee, J.Y., Yang, J.S., Kim, J.W., Kim, V.N. and Chang, H. (2020). The architecture of SARS-CoV-2 transcriptome. Cell. https://doi.org/10.1016/j.cell.2020.04.011

Lasso, G., Mayer, S.V., Winkelmann, E.R., Chu, T., Elliot, O., Patino-Galindo, J.A., Park, K., Rabadan, R., Honig, B., and Shapira, S.D. (2019). A Structure-Informed Atlas of Human-Virus Interactions. Cell *178*, 1526–1541.

McDonald, W.F., Huleatt, J.W., Foellmer, H.G., Hewitt, D., Tang, J., Desai, P., Price, A., Jacobs, A., Takahashi, V.N., Huang, Y., et al. (2007). A West Nile virus recombinant protein vaccine that coactivates innate and adaptive immunity. J. Infect. Dis. *195*, 1607–1617.

Rual, J.-F., Hirozane-Kishikawa, T., Hao, T., Bertin, N., Li, S., Dricot, A., Li, N., Rosenberg, J., Lamesch, P., Vidalain, P.-O., et al. (2004). Human ORFeome version 1.1: a platform for reverse proteomics. Genome Res. *14*, 2128–2135.

Tanz, S.K., Castleden, I., Small, I.D., and Millar, A.H. (2013). Fluorescent protein tagging as a tool to define the subcellular distribution of proteins in plants. Front. Plant Sci. *4*, 214.

ThermoFisher Gateway Recombination and Seamless Cloning Support.

Walhout, A.J.M., Temple, G.F., Brasch, M.A., Hartley, J.L., Lorson, M.A., van den Heuvel, S., and Vidal, M. (2000). GATEWAY recombinational cloning: Application to the cloning of large numbers of open reading frames or ORFeomes. In Methods in Enzymology, J. Thorner, S.D. Emr, and J.N. Abelson, eds. (Academic Press), pp. 575–592.

Walls, A.C., Park, Y.-J., Tortorici, M.A., Wall, A., McGuire, A.T., and Veesler, D. (2020). Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell *181*, 281–292.

Waugh S.D. (2011). An overview of enzymatic reagents for the removal of affinity tags. Protein

Expr. Purif. *80*, 283–293.

de Wilde, A.H., Snijder, E.J., Kikkert, M., and van Hemert, M.J. (2018). Host Factors in Coronavirus Replication. In Roles of Host Gene and Non-Coding RNA Expression in Virus Infection, R.A. Tripp, and S.M. Tompkins, eds. (Cham: Springer International Publishing), pp. 1–42.

World Health Organization (2020a). COVID-19 Situation Reports.

World Health Organization (2020b). Report of the who-china joint mission on coronavirus disease 2019 (COVID-19).

Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., Hu, Y., Tao, Z.-W., Tian, J.-H., Pei, Y.-Y., et al. (2020). A new coronavirus associated with human respiratory disease in China. Nature *579*, 265–269.

Yu, P., Zhu, J., Zhang, Z., and Han, Y. (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.

Zhang, L., Lin, D., Sun, X., Curth, U., Drosten, C., Sauerhering, L., Becker, S., Rox, K., and Hilgenfeld, R. (2020). Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. Science *368*, 409–412.