Peer-reviewed version available at *G3* **2020**, *10*; doi:10.1534/g3.120.401554

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

Dae-Kyum Kim[1,2,3¶], Jennifer J. Knapp[1,2,3¶], Da Kuang[1,2,3¶], 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, Canada

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

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

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

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

[6] Université de Paris, Paris, France

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

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

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

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

* Corresponding authors. E-mail: yves.jacob@pasteur.fr (YJ), fritz.roth@utoronto.ca (FPR)

¶ These authors contributed equally to this work.

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.

Author Summary

The current COVID-19 pandemic is motivating widespread efforts to understand the life cycle and spread of the SARS-CoV-2 virus, as a foundation for rational development of preventive and therapeutic measures. We therefore generated a comprehensive and flexible collection of SARS-CoV-2 DNA fragments encoding viral protein coding sequences (CDSs), and have made it widely available both directly and via a non-profit distributor. We used the Gateway cloning system to allow efficient transfer of any viral CDS into a wide range of vectors enabling a wide variety of studies, such as expression, tagging, purification, or various interaction or activity assays, to better understand the virus and interaction with its host.

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 [1]. COVID-19 manifestation in patients can range from asymptomatic (no symptoms) to severe pneumonia and death [2]. Early analysis of the outbreak in China outlines symptoms that commonly include fever, dry cough, shortness of breath and myalgia [3]. Person-to-person spread through respiratory droplets has been identified as a major

Peer-reviewed version available at *G3* **2020**, *10*; doi:10.1534/g3.120.401554

source of transmission of the virus [4]. To limit contagion, various measures from social distancing to nationwide lockdowns, have been imposed to contain and control the transmission of SARS-CoV-2 [5]. Despite these measures, the number of confirmed COVID-19 cases has continued to rise [1], highlighting the need for an effective vaccine and antiviral agents. Furthermore, the extrapolations concerning the evolution of the pandemic are particularly alarming [6]. 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 [7,8]. Using the genome sequence as a reference, Chan *et al*. identified 12 viral open reading frames (ORFs), including ORF1ab, a large polyprotein which is post-translationally processed into 16 proteins [7]. More recently, Wu *et al.* discovered two additional viral ORFs (ORF9Bwu and ORF10wu) with unclear functions [8]. Progress on molecular characterization has been made on several viral proteins [9,10], 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 [11]. 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.

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 [12]. All 28

Peer-reviewed version available at *G3* **2020**, *10*; doi:10.1534/g3.120.401554

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), one truncated CDS variant (ORF8B-truncated), that were each detected by recent viral transcriptome mapping efforts [13,14] and two missense catalytic variants (NSP3 C857A and NSP5 C146A) [20].

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	✓	✓	✓
	NSP3-C857A	Putative PL-pro domain (with C857A variant)	1,946	✓	✓	NA
	NSP4	Complex with NSP3 & 6 for DMV (double-membrane vesicle) formation	501	✓	✓	✓
	NSP5	3CL-pro domain	307	✓	✓	✓
	NSP5-C146A	3CL-pro domain (with C146A variant)	307	✓	✓	NA
	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 available; NA indicates that the clone was not available the time of this writing.

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 expanded our collection to include clones containing an N-terminal recognition sequence for nuclear inclusion protease from tobacco etch virus (TEV) [15,16]. 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 [17].

To promote open-access dissemination of the collection, all clones have been deposited to the non-profit organization Addgene [18], and freely available from the authors in circumstances where Addgene cannot be used. S2 Table summarizes all CDSs in the collection, together with their nucleotide sequences, nucleotide and amino acid lengths and direct links to order 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 [19], understanding virus-host protein-protein interactions [20,21], production of recombinant virus proteins for structural studies [22], mapping of protein subcellular localization using N-terminal fluorescent reporters [23], or development of vaccines or other therapeutics [24,25].

Materials and Methods

Synthesis of viral coding sequences

Based on the published annotation of the genome sequence of the HKU-SZ-005b (GenBank MN975262) [7] and Wuhan-Hu-1 (GenBank MN908947) [8] isolates of SARS-CoV-2, we requested the synthesis of viral coding sequences (GenScript, IDT), including termination

Peer-reviewed version available at *G3* **2020**, *10*; doi:10.1534/g3.120.401554

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 was added to NSP2–16 to allow independent transcription and translation, as the endogenous product is derived from ORF1 by post-translational processing. ORF9Bwu, an alternative ORF within the N gene from the SARS-COV-2 [8], was subsequently amplified by Polymerase Chain Reaction (PCR) from the viral N gene with primers listed in S1 Table.

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 [26]. 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) and pDONR223).

To enable the removal of N-terminal fusion tags, we further expanded our collection to include clones containing N-terminal 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 containing TEV sequences and original reverse primers.

Each SARS-CoV-2 CDS bacterial clone (DH5alpha *E. coli* strain, NEB Cat #C2987) 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 S1 Table.

Supporting information

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

S2 Table. 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

1. World Health Organization. COVID-19 Situation Reports. 2020 [Cited 2020 June 24]. Available from: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports
2. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 2020;395: 497-506.
3. World Health Organization. Report of the who-china joint mission on coronavirus disease 2019 (COVID-19). 2020 [cited 2020 Jun 24]. Available from: https://www.who.int/publications/i/item/report-of-the-who-china-joint-mission-on-coronavirus-disease-2019-(covid-19)
4. Yu P, Zhu J, Zhang Z, Han Y. A Familial Cluster of Infection Associated With the 2019 Novel Coronavirus Indicating Possible Person-to-Person Transmission During the Incubation Period. J. Infect. Dis. 2020;221: 1757-1761.
5. Cohen J, Kupferschmidt K. Countries test tactics in “war” against COVID-19. Science. 2020;367: 1287-1288.
6. Ferguson N, Laydon D, Nedjati GG, Imai N, Ainslie K, Baguelin M, Bhatia S, et al. Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand. 2020 Mar:1-20. Available from: https://doi.org/10.25561/77482
7. Chan JFW, Kok KH, Zhu Z, Chu H, To KKW, Yuan S, et al. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. Emerg. Microbes Infect. 2020;9: 221-236.
8. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020;579: 265-269.
9. Walls AC, Park YJ, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell. 2020;181: 281-292.

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

11. Walhout AJM, Temple GF, Brasch MA, Hartley JL, Lorson MA, van den Heuvel S, et al. GATEWAY recombinational cloning: Application to the cloning of large numbers of open reading frames or ORFeomes. In: Thorner J, Emr SD, Abelson JN, editors. Methods in Enzymology. Cambridge: Academic Press; 2000. pp. 575–592.

12. Cheo DL, Titus SA, Byrd DR, Hartley JL, Temple GF, Brasch MA. Concerted assembly and cloning of multiple DNA segments using in vitro site-specific recombination: functional analysis of multi-segment expression clones. Genome Res. 2004;14: 2111-2120.

13. Davidson AD, Williamson MK, Lewis S, Shoemark D, Carroll MW, Heesom K, et al. 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. BioRxiv [Preprint]. 2020 bioRxiv 2020.03.22.002204 [posted 2020 Mar 24; cited 2020 Jun 24]: [39 p.]. Available from: https://www.biorxiv.org/content/10.1101/2020.03.22.002204v1 doi: 10.1101/2020.03.22.002204

14. Kim D, Lee JY, Yang JS, Kim JW, Kim VN, Chang H. The architecture of SARS-CoV-2 transcriptome. Cell. 2020;181: 914-921.

15. Carrington JC, Dougherty WG. Small nuclear inclusion protein encoded by a plant potyvirus genome is a protease. J. Virol. 1987;61: 2540-2548.

16. Carrington JC, Dougherty WG. A Viral Cleavage Site Cassette: Identification of Amino Acid Sequences Required for Tobacco Etch Virus Polyprotein Processing. Proc. Natl. Acad. Sci. U.S.A. 1988;85: 3391-3395.

17. Waugh SD. An overview of enzymatic reagents for the removal of affinity tags. Protein Expr.

Purif. 2011;80: 283-293.

18. Kamens J. The Addgene repository: an international nonprofit plasmid and data resource. Nucleic Acids Res. 2015;43: D1152-D1157.
19. de Wilde AH, Snijder EJ, Kikkert M, van Hemert MJ. Host Factors in Coronavirus Replication. In: Tripp RA, Tompkins SM, editors. Roles of Host Gene and Non-Coding RNA Expression in Virus Infection. Cham: Springer International Publishing; 2018. pp. 1-42.
20. Gordon DE, Jang GM, Bouhaddou M, Xu J, Obernier K, White KM, et al. (2020). A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature. Forthcoming 2020.
21. Lasso G, Mayer SV, Winkelmann ER, Chu T, Elliot O, Patino-Galindo JA, et al. A Structure-Informed Atlas of Human-Virus Interactions. Cell. 2019;178: 1526-1541.
22. Edavettal SC, Hunter MJ, Swanson RV. Genetic construct design and recombinant protein expression for structural biology. Methods Mol. Biol. 2012;841: 29-47.
23. Tanz SK, Castleden I, Small ID, Millar AH. Fluorescent protein tagging as a tool to define the subcellular distribution of proteins in plants. Front. Plant Sci. 2013;4: 214.
24. Jing L, Haas J, Chong TM, Bruckner JJ, Dann GC, Dong L, et al. Cross-presentation and genome-wide screening reveal candidate T cells antigens for a herpes simplex virus type 1 vaccine. J. Clin. Invest. 2012;122: 654-673.
25. McDonald WF, Huleatt JW, Foellmer HG, Hewitt D, Tang J, Desai P, et al. A West Nile virus recombinant protein vaccine that coactivates innate and adaptive immunity. J. Infect. Dis. 2007;195: 1607-1617.
26. Rual JF, Hirozane-Kishikawa T, Hao T, Bertin N, Li S, Dricot A, et al. Human ORFeome version 1.1: a platform for reverse proteomics. Genome Res. 2004;14: 2128-2135.