OGG-CoV: Ontology Representation and Analysis of Genes and Genomes of Coronaviruses

Anthony Huffman [a,1] and Yongqun He [a]
[a] *University of Michigan Medical School Ann Arbor, MI 48109, USA*

Abstract. The current SARS-CoV-2 pandemic has brought about significant influx of coronavirus data for better disease understanding and treatment and vaccine development. As such, ontologies to help categorize the massive amount of research and information being done are required. The Ontology of Genes and Genomes (OGG) systematically represents genes and genomes for specific organisms. OGG-CoV is a branch of OGG that provides an ontological representation of genes and genomes within different coronavirus strains and species. OGG-CoV adopts a pan-genome strategy and systematically represents coronavirus genes based on their ortholog classification.

Keywords. Ontology, Coronavirus, Genes, Genomes

1. Introduction

Coronaviruses are a subfamily of RNA viruses, some of which are responsible for various human respiratory illnesses [1]. The latest, SARS-COV-2, has caused the COVID-19 pandemic that, as of 08/17/2020, has reached over 21,707,773 confirmed cases in 188 countries and has prompted a massive expansion of genomic sequencing to support a better understanding of the disease and rational development of effective and safe treatments and cures (https://coronavirus.jhu.edu/us-map). The NIH already has 16,084 sequences uploaded for this virus and there are 23,295 articles on PubMed written within the last eight months, demonstrating the need to categorize the massive data influx.

Coronavirus genes are generally divided into structural proteins and accessory proteins. The structural proteins are responsible for the formation of the virion and its transmission and have well-characterized functions and roles [2]. Accessory proteins form parts of the virion and vary more often between species, with some being linked as contributing to the coronavirus pathology [3]. RNA viruses are highly mutable, with SARS-CoV-2 having mutation rates of ~30% within some of their genes [4].

Ontologies are a tool used to classify and systematize the information of entities and relations to aid further research and inquiry. The Ontology of Genes and Genomes (OGG) is an ontology developed to classify the genes and genomes of organisms, with sub-ontologies developed for specific model species [5]. The Coronavirus Infectious Disease Ontology (CIDO) (https://github.com/CIDO-ontology/) is a community-based

[1]Anthony Huffman, Corresponding author, University of Michigan Medical School, Ann Arbor, MI, USA. E-mail: huffmaar@umich.med.edu,

ontology in the domain of coronavirus diseases, which covers various topics such as virus etiology, hosts, host-virus interactions, drugs and vaccines [6]

In this study, we present OGG-CoV, an extension of the OGG core with a focus on logically modeling and representing genes and genomes in various coronaviruses. OGG-CoV has then been imported into the CIDO to support integrative representation and analysis of coronavirus genes and their usages in different applications.

2. Methods

2.1. Coronavirus gene and genome information extraction

The genes and genetic sequences were all collected and extracted from NCBI. Sections of polyprotein genes that produce functional gene products were assigned from UniProt's InterPro.

2.2. OGG-CoV development

OGG-CoV was developed by aligning it with OGG and CIDO [7]. Orthologs between coronaviruses were determined using the NCBI gene annotation and literature annotation for genes with known protein products. Genes that do not have labeled orthologs were classified using nucleotide alignments from MUSCLE (https://www.ebi.ac.uk/Tools/msa/) and ViPR BLAST (https://www.viprbrc.org/brc/-home.spg?decorator=vipr). If both programs predicted a different set of orthologs or there was no match, the gene was placed within its own class. The basic OGG terms and structure were obtained using Ontofox [6]. Extracted genome and gene data were uploaded using Ontorat into version 5.5 of Protégé-OWL editor (https://protege.stanford.edu/) for editing.

2.3. OGG-CoV source code

The source code of OGG-CoV is available at GitHub: https://github.com/CIDO-ontology/OGG-CoV. The source code uses the license CC-BY.

3. Results

3.1. High level ontology design

Figure 1 lays out the high level design, using Basic Formal Ontology (BFO) inherited from the high level design of OGG. The OGG-CoV is designed to represent specific

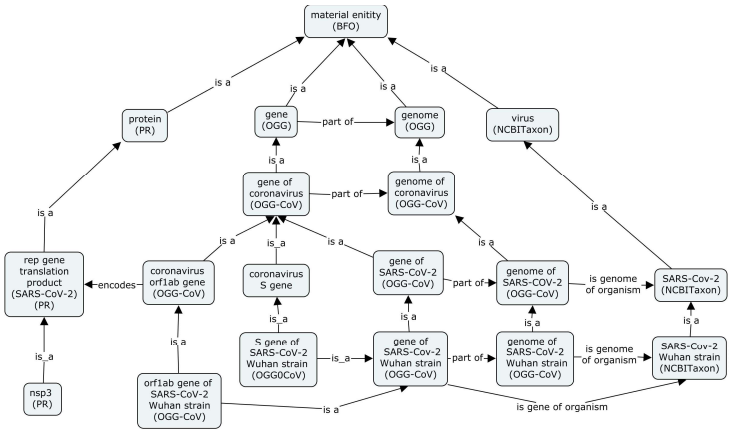

Figure 1. Screenshot of high level of design of OGG-CoV. Each gene from a specific strain is a subclass for the gene within a species and the gene within the strain. The origin of each terms is shown below each class in parenthesis: BFO: Basic Formal Ontology, PR: Protein Ontology, OGG: Ontology of Genes and Genomes, NCBITaxon: NCBITaxon.

genes (e.g., orf1ab gene of SARS-CoV-2 Wuhan strain) under the gene of its own strain (e.g., gene of SARS-CoV-2 Wuhan strain) and meanwhile under its evolution-derived ortholog hierarchy (e.g., coronavirus S gene). This design asserts S gene of SARS-CoV-2 Wuhan strain under two parent classes. However, this design allows us to compare different genes within an ortholog.

3.2. Specific OGG-CoV development

In our current OGG-CoV development, we included 13 strains from 8 different species. SARS-CoV-2 Wuhan strain, the first sequenced SARS-CoV-2 strain, has 10 genes. As seen in Figure 2, the coronavirus genes are classified based on their taxonomic classification and orthology gene type. For example, the S gene of SARS-CoV-2 Wuhan strain can be classified as a gene under different strains (e.g., Wuhan strain) (Figure 2a) or as a gene under an ortholog S gene type (Figure 2b).

The orf1ab gene is made up of two different genes, orf1a and orf1b. Orf1a gene encodes a single polypeptide which produces 11 specific nonstructural proteins (nsps) Orf1b produces a polypeptide that which produces all but the last nsp produced by orf1a in addition to 5 unique nsps responsible for viral replication. This is due to a frame shift caused by ribosomal slippage during translation of orf1a. Together they code for the coronavirus replicase polyprotein. Each nsp-encoding portion of this replicase gene is represented in ontology as a part of the polygene sequence which encodes a specific nsp protein.

4. OGG-CoV Use case: Creation of a Coronavirus Pan-genome

Pan-genome is defined as the entire set of genes for all strains within a clade (i.e., a group of organisms evolved from a common ancestor) [9]. While originally restricted to bacteria, this can be expanded to other organisms, including viruses. A pan-genome typically includes a set of core genes that are present in all strains with the clade and accessory genes that are present in some but not all strains. Our OGG-CoV is designed to represent the genes in all the coronavirus clade based on the ortholog analysis, which allows pan-genome representation.

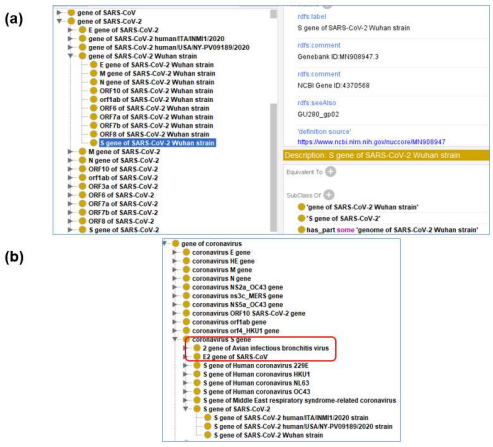

Figure 2. Classification of coronavirus genes in OGG-CoV. Each class that contains a specific s gene is sorted as a subclass for its gene taxonomy origin (A) and gene type (B). Orf1a and orf1b were labeled as a single gene, orf1ab, on the NCBI website. Each gene is purposely designed to have two explicit parents. The genes circled in red are classified as S genes due to NCBI listing a spike polypeptide product for these genes.

The selected coronaviruses genomes had a consistent pattern of the orf1ab gene, and the S (spike), E (envelope), M (membrane), and N (nucleocapsid) structure genes occurring in this specific order from the 5' to 3' direction. The one exception, AIBV NCBI reference genome, did not have a gene labeled as an M protein but did have an unlabeled gene between the E and N proteins. The orf1ab always demonstrated a -1 frameshift which coded for a single polypeptide that is cleaved to create 13-16 total nonstructural proteins, depending on the species. This seems to indicate that the coronavirus pan-genome has these structural proteins has its core genes.

The accessory genome of coronaviruses contains some structures genes that are part of the core genome for certain coronavirus sub-clades and the various accessory genes. We found that the structural gene HE (hemagglutinin-esterase) was found to be part of the core genome subgenus *Embecovirus* but not present in other coronavirus species. The various accessory genes differed in quantity between each species and were labeled by their ordinal location within the genome from the 5' to 3' direction. This nomenclature, while allowing for quick comparison of the number of accessory genes between coronaviruses, is insufficient to match accessory proteins within the ontology without supporting evidence from sequence alignment.

Interestingly, we found the name of the orf10 gene only in SARS-CoV-2 strains but not in other branches of coronaviruses. However, a unique name does not guarantee its unique presence among all coronavirus genes. To address this issue, we performed a MUSCLE analysis, and found that orf10 had no significant alignment with any other coronavirus proteins, suggesting that orf10 is uniquely present in SARS-CoV-2. We will later include more genomes from different coronavirus strains to confirm the unique presence of this gene in SARS-CoV-2. If the conclusion is true, orf10 likely plays an important role in making SARS-CoV-2 unique in terms of the viral pathogenesis and/or transmission.

5. Discussion and Conclusion

OGG-CoV is developed to serves as a knowledge base of coronavirus genes and genomes. Instead of presenting genes and genomes in a single species as done in previous OGG branches, OGG-CoV adopts a pan-genome strategy to systematically represent coronavirus genes not only in terms of its genomes, but also based on their ortholog classification. The ortholog classification method will then allow us to easily define the core and accessory genes of the pan-genome of all coronaviruses. Our use case demonstrates that many structure genes belong to the core genes. However, some genes such as the orf10 gene is unique is only SARS-CoV-2 strain.

Future work will include its integration with the other parts of CIDO and allow the other entities (e.g., vaccine and drug target) defined in CIDO to possibly link to the genes and genomes in OGG-CoV. We will also expand OGG-CoV to possibly identify and represent gene mutations that occur naturally in various coronaviruses, allowing for mechanistic analysis of host-virus interactions and fight against COVID-19 pandemic.

Acknowledgments

Edison Ong and Lauren Austin for their aid and commentary. This was funded by the ImmPort project (NIH NIAID grant 1UH2AI132931).

References

[1] Cui J, Li F, Shi ZL. Origin and evolution of pathogenic coronaviruses. Nat Rev Microbiol. 2019 Mar;17(3):181-192. doi: 10.1038/s41579-018-0118-9. PMID: 30531947; PMCID: PMC7097006.

[2] Fehr AR, Perlman S. Coronaviruses: an overview of their replication and pathogenesis. Methods Mol Biol. 2015;1282:1-23. doi: 10.1007/978-1-4939-2438-7_1. PMID: 25720466; PMCID: PMC4369385.

[3] Yue Y, Nabar NR, Shi CS, Kamenyeva O, Xiao X, Hwang IY, Wang M, Kehrl JH. SARS-Coronavirus Open Reading Frame-3a drives multimodal necrotic cell death. Cell Death Dis. 2018 Sep 5;9(9):904. doi: 10.1038/s41419-018-0917-y. PMID: 30185776; PMCID: PMC6125346.

[4] Pachetti M, Marini B, Benedetti F, Giudici F, Mauro E, Storici P, Masciovecchio C, Angeletti S, Ciccozzi M, Gallo RC, Zella D, Ippodrino R. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. J Transl Med. 2020 Apr 22;18(1):179. doi: 10.1186/s12967-020-02344-6. PMID: 32321524; PMCID: PMC7174922.

[5] He Y, Liu Y, Zhao B. (2014) OGG: a Biological Ontology for Representing Genes and Genomes in Specific Organisms. In: Hogan WR, Arabandi S, Brochhausen, M, editor. Proceedings of the 5th International conference on Biomedical Ontologies; October 8-9 2014; Houston, TX. Published on CEUR-WS; c2014 p. 13–20

[6] He Y, Yu H, Ong E, Wang Y, Liu Y, Huffman A, Huang HH, Beverley J, Hur J, Yang X, Chen L, Omenn GS, Athey B, Smith B. CIDO, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis. Sci Data. 2020 Jun 12;7(1):181. doi: 10.1038/s41597-020-0523-6. PMID: 32533075; PMCID: PMC7293349.

[7] Xiang Z, Courtot M, Brinkman RR, Ruttenberg A, He Y. OntoFox: web-based support for ontology reuse. BMC Res Notes. 2010 Jun 22;3:175. doi: 10.1186/1756-0500-3-175. PMID: 20569493; PMCID: PMC2911465.

[8] Xiang Z, Zheng J, Lin Y, He Y. Ontorat: automatic generation of new ontology terms, annotations, and axioms based on ontology design patterns. J Biomed Semantics. 2015 Jan 9;6:4. doi: 10.1186/2041-1480-6-4. PMID: 25785185; PMCID: PMC4362828.

[9] Computational Pan-Genomics Consortium. Computational pan-genomics: status, promises and challenges. Brief Bioinform. 2018 Jan 1;19(1):118-135. doi: 10.1093/bib/bbw089. PMID: 27769991; PMCID: PMC5862344.