Virology 570 (2022) 123–133

Contents lists available at ScienceDirect

Virology

journal homepage: www.elsevier.com/locate/virology

Genomic evolution of the *Coronaviridae* family

Christian M. Zmasek [a], Elliot J. Lefkowitz [b], Anna Niewiadomska [a], Richard H. Scheuermann [a,c,d,e,*]

[a] Department of Informatics, J. Craig Venter Institute, La Jolla, CA, 92037, USA
[b] Department of Microbiology, UAB School of Medicine, Birmingham, AL, 35294, USA
[c] Department of Pathology, University of California, San Diego, CA, 92093, USA
[d] Division of Vaccine Discovery, La Jolla Institute for Immunology, La Jolla, CA, 92037, USA
[e] Global Virus Network, Baltimore MD, 21201, USA

ARTICLE INFO

Keywords:
Nidovirales
Coronaviridae
Orthocoronavirinae
Evolution
Phylogenetics
Phylogenomics
Protein domains
Genome
Hidden Markov models

ABSTRACT

The current outbreak of coronavirus disease-2019 (COVID-19) caused by SARS-CoV-2 poses unparalleled challenges to global public health. SARS-CoV-2 is a Betacoronavirus, one of four genera belonging to the *Coronaviridae* subfamily *Orthocoronavirinae*. *Coronaviridae*, in turn, are members of the order *Nidovirales*, a group of enveloped, positive-stranded RNA viruses. Here we present a systematic phylogenetic and evolutionary study based on protein domain architecture, encompassing the entire proteomes of all *Orthocoronavirinae*, as well as other *Nidovirales*. This analysis has revealed that the genomic evolution of *Nidovirales* is associated with extensive gains and losses of protein domains. In *Orthocoronavirinae*, the sections of the genomes that show the largest divergence in protein domains are found in the proteins encoded in the amino-terminal end of the polyprotein (PP1ab), the spike protein (S), and many of the accessory proteins. The diversity among the accessory proteins is particularly striking, as each subgenus possesses a set of accessory proteins that is almost entirely specific to that subgenus. The only notable exception to this is ORF3b, which is present and orthologous over all Alphacoronaviruses. In contrast, the membrane protein (M), envelope small membrane protein (E), nucleoprotein (N), as well as proteins encoded in the central and carboxy-terminal end of PP1ab (such as the 3C-like protease, RNA-dependent RNA polymerase, and Helicase) show stable domain architectures across all *Orthocoronavirinae*. This comprehensive analysis of the *Coronaviridae* domain architecture has important implication for efforts to develop broadly cross-protective coronavirus vaccines.

1. Introduction

Coronaviridae is a family of enveloped, positive-strand RNA viruses that infect a wide variety of animals. The *Coronaviridae* family belongs to the suborder *Cornidovirineae,* which, together with *Tornidovirineae* belong to the order *Nidovirales* (enveloped, positive-strand RNA viruses) (Fig. 1). Recent phylogenetic studies based on RNA-directed RNA polymerases indicate that *Nidovirales*, together with *Picornavirales, Caliciviridae, Astroviridae*, and their relatives form a distinct supergroup of RNA viruses (Picornavirus supergroup) (Koonin et al., 2020; Wolf et al., 2018). *Nidovirales* can infect a wide range of animal hosts, including insects, mollusks, crustaceans, and vertebrates, suggesting horizontal virus transfer across metazoan species (Dolja and Koonin, 2020). *Coronaviridae* are divided into two subfamilies *Letovirinae* and *Orthocoronavirinae*, the latter of which are the main focus of this work. *Orthocoronavirinae* in turn are divided into four genera, Alpha-, Beta-, Gamma, and Deltacoronaviruses. Currently, there are seven *Orthocoronavirinae* species or sub-species, which have been found to infect humans, two members of the Alphacoronavirus genus: Human coronavirus 229E and Human coronavirus NL63, and five members of the Betacoronavirus genus: Human coronavirus OC43, Human coronavirus HKU1, Middle East respiratory syndrome-related coronavirus (MERS-CoV), Severe acute respiratory syndrome coronavirus (SARS-CoV), and Severe acute respiratory syndrome coronavirus 2 (2019-nCoV, SARS-CoV-2 (Andersen et al., 2020; Drosten et al., 2003; Fan et al., 2019; Fehr and Perlman, 2015).

All *Orthocoronavirinae* viruses possess four shared structural proteins, the spike (S), envelope (E), membrane (M) and nucleocapsid (N) proteins. The genome is packed inside a helical capsid formed by the nucleoprotein N. This in turn is surrounded by an envelope containing

* Corresponding author. Department of Informatics, J. Craig Venter Institute, La Jolla, CA, 92037, USA.
E-mail address: RScheuermann@jcvi.org (R.H. Scheuermann).

https://doi.org/10.1016/j.virol.2022.03.005
Received 12 November 2021; Received in revised form 11 March 2022; Accepted 18 March 2022
Available online 30 March 2022

the E and M proteins, which are involved in virus assembly, and the spike glycoprotein protein S, which mediates virus entry into host cells (McBride and Fielding, 2012). *Orthocoronavirinae* have relatively large viral genomes in comparison to other RNA viruses, with sizes ranging from 26 to 32 kilobases. The first two open reading frames, ORF1a and ORF1b, code for two overlapping large replicase-containing polyproteins, pp1a and pp1ab, with the larger pp1ab translated as a result of a -1 ribosomal frameshifting (Fig. 2A). These large polyproteins are subsequently (self) cleaved into 15 or 16 mature proteins referred to as non-structural proteins (nsps). And while the PP1ab, S, E, M, and N proteins are found in all *Coronaviridae* family genomes, the individual protein domains show surprising diversity. In addition, depending on the specific strain, many coronaviruses contain additional ORFs coding for accessory proteins, many of which remain poorly characterized (Fig. 2B).

In this work, we performed a protein domain-centric evolutionary comparative genomics analysis of *Coronaviridae* genomes, revealing the complex domain architectures that have resulted from recombination and a complicated evolutionary history.

Homologs are genes that are related by shared ancestry. Orthologs were defined by Fitch in 1970 as homologous genes in different species that diverged by speciation. Genes that diverged by gene duplication, either in the same or different species, have been termed paralogs (Fitch, 1970, 2000). While the terms ortholog and paralog have no functional implications (Jensen, 2001), orthologs are often thought of as more functionally similar than paralogs at the same level of sequence divergence (Altenhoff et al., 2012; Eisen, 1998).

Protein domains are distinct functional and/or structural units of a protein. Domains tend to form stable compact three-dimensional structures that can often be independently folded. Many proteins are composed of multiple domains, with each domain having its own evolutionary history and biochemical function. Thus, the architecture of a protein is a product of the ordered arrangement of its constituent domains and their overall tertiary structure. During evolution, multiple domains can combine, creating a vast number of distinct domain combinations, even within the same species (Moore et al., 2008). Assembling multiple domains into a single protein creates an entity whose function can be more than the sum of its constituent parts. The generation of proteins with novel combinations of duplicated and then diverged domains is a major mechanism for rapid evolution of new functionality in genomes (Itoh et al., 2007; Peisajovich et al., 2010). This modular structure of proteins enables rapid emergence of a multitude of novel protein functions from an initially limited array of functional domains. Proteins can gain or lose domains via genome rearrangements; the domains themselves can be modified by small-scale mutations (Christian M. Zmasek and Godzik, 2012).

Here we use the Domain-architecture Aware Inference of Orthologs (DAIO) approach described in (Zmasek et al., 2019) to compare the arrangement of protein domains (and by extension, proteins) in polyproteins and ORFs from different *Orthocoronavirinae* sub-genera, updating and expanding our knowledge of *Nidovirales* genome evolution at the domain level, which, for example, has been review previously in (Gorbalenya et al., 2006). This approach places proteins into groups in which all members are not only orthologous to each other but also have the exact same domain architecture. This analysis resulted in the classification of *Coronaviridae* proteins into "Strict Ortholog Groups" (SOGs), in which all proteins are orthologous to each other (related by speciation events) and exhibit the same domain architecture. The SOG classification also enabled the development of an informative naming convention for each SOG that includes information about the protein's function (if known) and a suffix indicating the taxonomic group (such as Betacoronavirus) where a particular SOG is present. The SOG classification results are publicly available through the Virus Pathogen Resource (ViPR) (Pickett et al., 2012) at https://www.viprbrc.org.

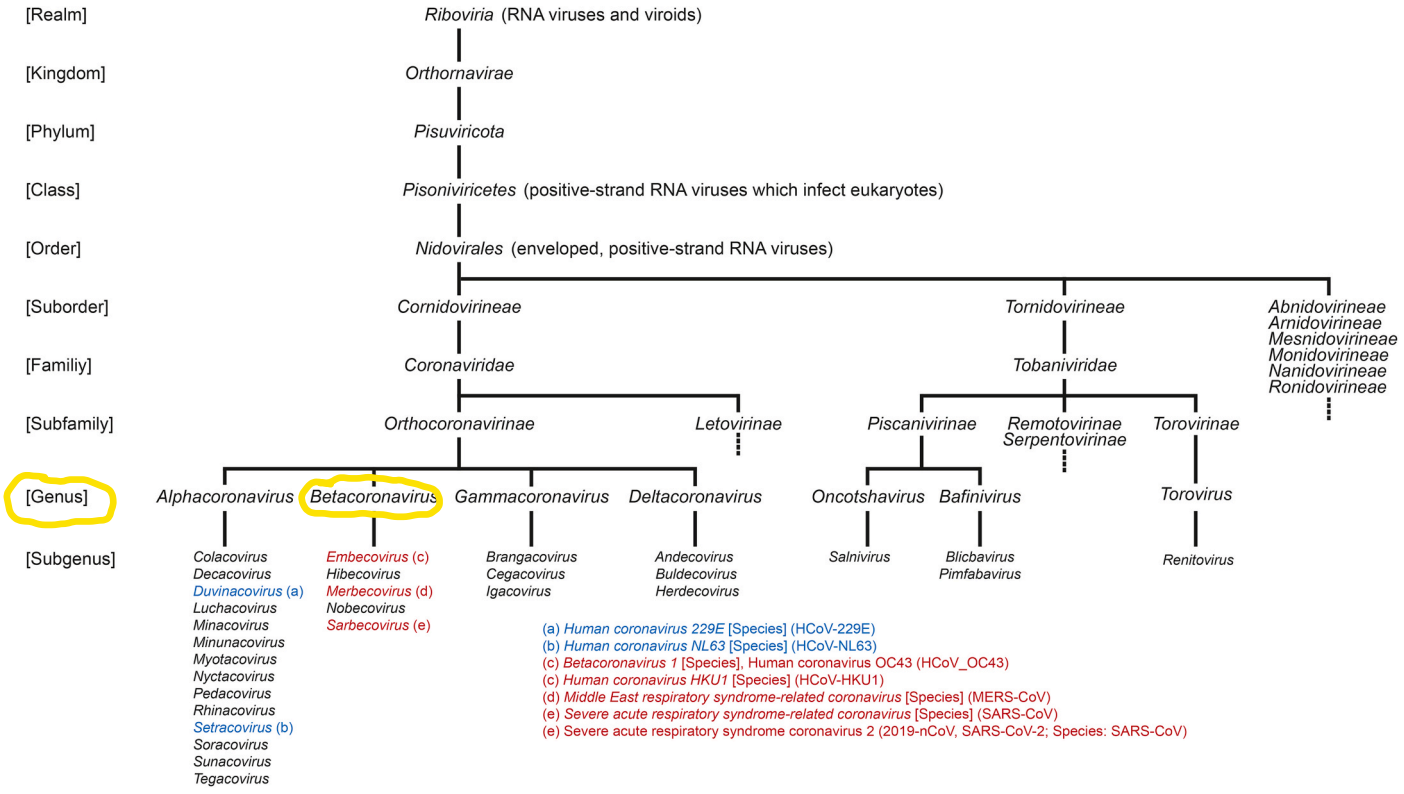

Fig. 1. *Nidovirales* taxonomy. This figure is based on the taxonomy established by the International Committee on Taxonomy of Viruses (ICTV) and currently used by the U.S. National Center for Biotechnology Information (NCBI) and the Universal Protein Resource (UniProt) databases. Viruses which infect humans are listed in blue (Alphacoronaviruses) and red (Betacoronaviruses). Their taxonomic level is indicated in square brackets. For some viruses, no taxonomic level has been established as of this writing. An example of this is Human coronavirus OC43.