Trends in Microbiology

CellPress

Review

Molecular Evolution of Human Coronavirus Genomes

Diego Forni,[1] Rachele Cagliani,[1] Mario Clerici,[2,3] and Manuela Sironi[1,*]

Human coronaviruses (HCoVs), including SARS-CoV and MERS-CoV, are zoonotic pathogens that originated in wild animals. HCoVs have large genomes that encode a fixed array of structural and nonstructural components, as well as a variety of accessory proteins that differ in number and sequence even among closely related CoVs. Thus, in addition to recombination and mutation, HCoV genomes evolve through gene gains and losses. In this review we summarize recent findings on the molecular evolution of HCoV genomes, with special attention to recombination and adaptive events that generated new viral species and contributed to host shifts and to HCoV emergence.

Human Coronaviruses Are Zoonotic Pathogens

The recent emergence of severe acute respiratory syndrome-related coronavirus (SARS-CoV) and of Middle East respiratory syndrome-related Coronavirus (MERS-CoV) (order Nidovirales, family Coronaviridae, subfamily Coronavirinae) as dangerous zoonoses stirred great interest in the ecology and evolution of coronaviruses. Before the SARS-CoV epidemic only two HCoVs were known: HCoV-229E and HCoV-OC43. Two additional HCoVs, HCoV-NL63 and HCoV-HKU1, were discovered in 2004–2005 from clinical specimens [1]. These viruses originated in animals and are mainly responsible for respiratory diseases in humans (Figure 1A, Key Figure). Specifically, all HCoVs are thought to have a bat origin, with the exception of lineage A beta-CoVs, which may have reservoirs in rodents [2]. The phylogenetic relationships of HCoVs and other animal CoVs mentioned in this review are summarized in Figure 1A.

A number of field studies identified and sequenced viruses related to HCoVs in wildlife reservoirs, and phylogenetic reconstruction provided important clues on the most likely events that led to the introduction of HCoVs in human populations. Several recent excellent reviews delve into the knowns and unknowns of HCoV origin in terms of reservoir species, amplification host, and, more generally, of CoV ecology [1,3–5]. In this review we instead focus on the molecular evolution of HCoV genomes. The general concepts of evolutionary analyses in viruses are outlined in Box 1, whereas the most common approaches that were applied to the analysis of CoV sequence evolution in terms of phylogenetic reconstruction, detection of recombination, and identification of selection signatures are summarized in Boxes 1 and 2.

HCoV Genome Organization

CoVs are positive-sense, single-strand RNA viruses with a likely ancient origin, and HCoVs repeatedly emerged during the past 1000 years (Box 3). All CoVs have nonsegmented genomes that share a similar organization. About two thirds of the genome consists of two large overlapping **open reading frames** (ORF1a and ORF1b; see Glossary), that are translated into the pp1a and pp1ab polyproteins. These are processed to generate 16 nonstructural proteins (nsp1 to 16). The remaining portion of the genome includes ORFs for the structural proteins: spike (S),

Trends

Human coronaviruses (HCoVs) are zoonotic pathogens with large and complex genomes. Some HCoV accessory proteins were acquired from host genes, and some were lost or split during HCoV evolution. Most likely SARS-CoV ORF8 became dispensable during the shift to the human/civet host.

HCoV spike proteins adapted to use diverse cellular receptors. This occurred by divergence followed, in some cases, by convergent evolution to bind the same receptor.

Recombination and positive selection shaped the diversity of CoV genomes, especially the S gene. Positive selection in the S gene of MERS-CoV and related CoVs mainly acted on the heptad repeats.

In MERS-CoV and other lineage C beta-CoVs, positive selection targeted the nonstructural components, particularly ORF1a. Most adaptive events occurred in nsp3, which acts as a viral protease and contributes to suppression of interferon responses.

[1]Scientific Institute IRCCS E. MEDEA, Bioinformatics, Bosisio Parini, Italy
[2]Department of Physiopathology and Transplantation, University of Milan, Milan, Italy
[3]Don C. Gnocchi Foundation ONLUS, IRCCS, Milan, Italy

*Correspondence: manuela.sironi@bp.lnf.it (M. Sironi).

Box 1. Molecular Evolution in Viruses: General Concepts

RNA viruses are rapidly evolving pathogens that can accumulate considerable genetic diversity in relatively short time periods. This is mostly due to their high nucleotide mutation rates (but see text for CoVs). The diversity of extant viral sequences can be analyzed to construct phylogenetic relationships among species/strains and to infer the underlying evolutionary patterns. In the presence of recombination a single phylogenetic tree is unable to describe the evolution of homologous sequences. Because recombination is common in many viruses, including HCoVs, the evolution of viral genomes is best modeled by several phylogenies, one for each nonrecombinant fragment (Box 2). By reassorting mutations, recombination has the potential to generate novel viral phenotypes. Thus, not only recombination is of interest *per se*, but failure to account for its presence can distort phylogeny-based analyses, including estimates of natural selection [66]. Natural selection acts pervasively on viral sequences. When coding regions are concerned, natural selection is commonly estimated in terms of ω (also referred to as dN/dS) – that is, the observed number of nonsynonymous differences per nonsynonymous site (dN) over the observed number of synonymous differences per synonymous site (dS). Under neutral evolution, ω is expected to be equal to 1, as the rate at which amino acid substitutions accumulate is similar to the rate for synonymous changes. Due to the fact that essential protein domains can often tolerate only minor sequence changes, most amino acid replacements are eliminated by selection; this generates ω values <1, a situation referred to as negative (or purifying) selection. Nevertheless, amino acid replacements can be advantageous for a virus in terms, for example, of host adaptation or immune evasion: in this case ω values can be higher than 1 (positive selection). Thus, evaluation of how ω varies from site to site or from branch to branch in a phylogeny is commonly used to describe selective events. Some possible caveats should nevertheless be kept in mind. (i) The saturation of substitution rates (especially dS) may occur and affect evolutionary inference when fast-evolving sequences are analyzed (see Box 3 for an example, and Box 2 for methods to overcome this problem). (ii) In viral genomes synonymous substitutions are not always neutral; this may be due to the presence of overlapping reading frames, conserved RNA secondary structures, packaging signals, and other functional elements (Box 3). (iii) A relaxation in the intensity of both negative and positive selection may occasionally occur (Box 2 and Figure 2A).

Glossary

dN: the observed number of nonsynonymous substitutions per nonsynonymous site.
dS: the observed number of synonymous substitutions per synonymous site.
Hemagglutinin-esterases (HEs): a family of viral proteins that mediate binding to *O*-acetylated sialic acids.
Homology: the relationship between elements (e.g., genes, proteins) deriving from a common ancestor.
Lectins: a group of proteins with carbohydrate recognition activity. Lectins are categorized in many distinct families depending on structural and functional properties.
Maximum likelihood (ML): is a statistical method for estimating population parameters from a data sample. Given one or more unknown parameters and a sample data, the ML estimates of the parameters are the values maximizing the probability of obtaining the observed data.
Open reading frame (ORF): the part of a reading frame that contains no stop codons. An ORF is a continuous stretch of nucleotide triplets that have the potential to code for a protein or a peptide.
Phosphodiesterases (PDEs): are enzymes that break a phosphodiester bond. PDEs belonging to the 2H family are characterized by two H-Φ-[S/T]-Φ motifs (where Φ is a hydrophobic residue) separated by an average of 80 residues.
Positive selection: the accumulation of favorable amino acid-replacing substitutions, which results in more nonsynonymous changes than expected under neutrality (dN/dS>1).
Purifying selection: the elimination of deleterious amino acid-replacing substitutions, which results in fewer nonsynonymous changes than expected under neutrality (dN/dS < 1) (it is also referred to as negative selection).
Viroporins: hydrophobic viral proteins that can promote the formation of channels following insertion into the host cell membrane and oligomerization.

envelope (E), membrane (M) and nucleoprotein (N). A variable number of accessory proteins are also encoded by distinct viruses (Figure 1B).

Among RNA viruses, CoVs have exceptionally long genomes (up to 32 kb). Genome expansion in CoVs is believed to be at least partially mediated by increased replication fidelity. Although estimates of the mutation rate for CoVs differ, possibly depending on the phase of CoV adaptation to novel hosts, several studies have shown that these viruses may possess an unusually high replication fidelity [6–8]. Indeed, a major step that allowed genome expansion in CoVs and, more generally, in Nidovirales, was the acquisition of a set of RNA-processing enzymes that improved the low fidelity of RNA replication [9]. These enzymes include an RNA 3′-to-5′ exoribonuclease (ExoN) and possibly an endoribonuclease (NendoU) [9]. Additional evidence, though, suggests that features distinct from replication fidelity underlie genome expansion in Nidovirales. These include a peculiar genome organization [9] and a processive replication complex [10].

Importantly, CoV genome expansion allowed the acquisition and maintenance of genes encoding diverse accessory proteins that may promote virus adaptation to specific hosts and often contribute to the suppression of immune responses, as well as to virulence. Accessory proteins differ in number and sequence even among CoVs belonging to the same lineage (Figure 1B), raising interesting questions about their origin and evolution.

Gene Gains and Gene Losses

The acquisition (or loss) of novel protein-coding genes has the potential to drastically modify viral phenotypes. Thus, tracing these gain/loss events may identify important turning points in viral evolution.

Among SARS-CoV accessory proteins, the origin of ORF8 has remained mysterious for a while, as SARS-CoV-related (SARSr) bat viruses were isolated but found to encode divergent ORF8 proteins (amino acid identity with SARS-CoV ORF8 around 33%) [11–13]. Very recently, SARSr-BatCoVs from *Rhinolophus sinicus* (Rs) and *Rhinolophus ferrumequinum* (Rf) were isolated