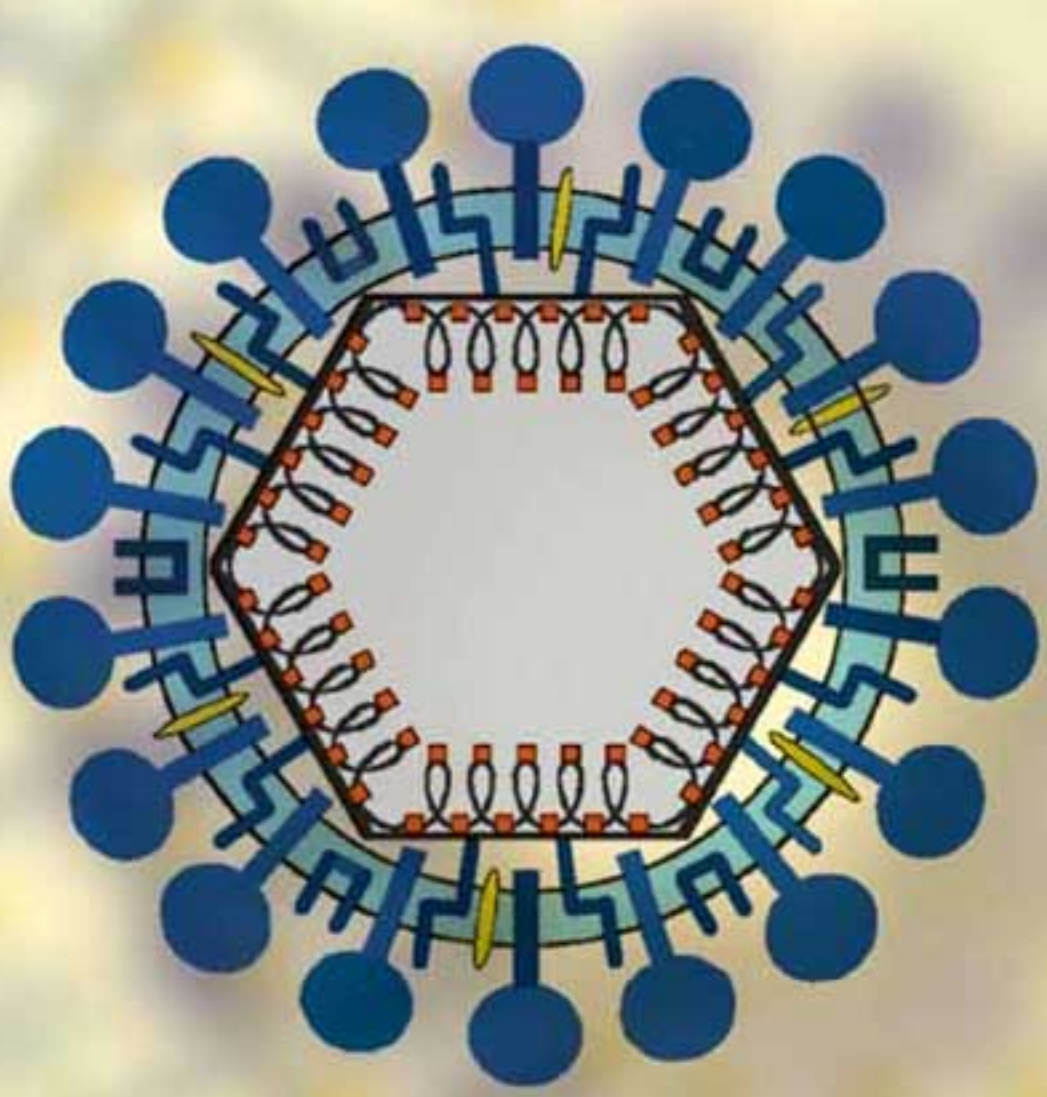
L. Enjuanes (Ed.)

Coronavirus Replication and Reverse Genetics



Springer

287
Current Topics in Microbiology and Immunology

L. Enjuanes (Ed.)

Coronavirus Replication and Reverse Genetics

With 49 Figures

Springer

Professor Dr. Luis Enjuanes
Centro Nacional de Biotecnología
Department of Molecular and Cell Biology
Campus Universidad Autónoma, Cantoblanco
38049 Madrid
Spain

e-mail: L.Enjuanes@cnb.uam.es

Cover illustration by Luis Enjuanes
Simplified structure of a coronavirus prototype (transmissible gastroenteritis virus, TGEV) showing the envelope with several structural proteins: spike (S), two topologies of the membrane (M or M') proteins, envelope (E) protein, and nucleoprotein (N). A nucleocapsid can be observed inside the envelope virus, containing the genomic RNA, the N protein, and the carboxiterminus of the M protein.

Library of Congress Catalog Card Number 72-152360

ISSN 0070-217X
ISBN 3-540-21494-1 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media
springeronline.com

Printed in Germany

Editor: Dr. Rolf Lange, Heidelberg
Desk editor: Anne Clauss, Heidelberg
Production editor: Andreas Gösling, Heidelberg
Cover design: design & production GmbH, Heidelberg
Typesetting: Stürtz AG, Würzburg
Printed on acid-free paper 27/3150/ag – 5 4 3 2 1 0

Preface

The *Coronaviridae* family is included in the *Nidovirales* order together with the *Arteriviridae* and *Roniviridae*. Possibly the first recorded coronavirus-related disease was feline infectious peritonitis in 1912. However, until the late 1960s the coronaviruses were not recognized as pathogens responsible for human diseases (common cold), and it was in 2003 when human coronaviruses (HCoVs) received worldwide attention with the emergence of the severe and acute respiratory syndrome (SARS), produced by a coronavirus (SARS-CoV), that has infected more than 8,000 people in 32 countries, killing about 10%. The increase in research on coronaviruses soon led to the discovery of another human coronavirus (HCoV-NL63), which is prevalent in 7% of hospital patients and has been associated with bronchiolitis and, possibly, conjunctivitis.

Coronaviruses have been identified in mice, rats, chickens, turkeys, pigs, dogs, cats, rabbits, horses, cows, and humans. Coronaviruses are associated mainly with respiratory, enteric, hepatic, and central nervous system diseases. In humans and fowl, coronaviruses primarily cause upper respiratory tract infections, while porcine and bovine coronaviruses establish enteric infections that result in severe economic losses. HCoVs are responsible for 10%–20% of common colds, and have been implicated in gastroenteritis, high and low respiratory tract infections, and rare cases of encephalitis. HCoVs have also been associated with infant necrotizing enterocolitis and are tentative candidates for multiple sclerosis.

In some coronavirus members, such as transmissible gastroenteritis virus (TGEV), three levels can be distinguished in the virion structure: the envelope, the core, and the nucleocapsid formed by the genome and the nucleoprotein (N). The CoV genome is a single-stranded positive-sense RNA genome of 27–32 kb that is infectious. Coronaviruses have the largest genome known for an RNA virus and probably one of the longest stable RNAs in nature. The genome of all coronaviruses contains a basic set of genes: the replicase (Rep 1a and 1b), the spike (S), envelope (E), membrane (M), and nucleoprotein (N) arranged in the order 5′-Rep1a-1b-S-E-M-N-3′ and a variable number of genes encoding nonstructural proteins intercalated between these genes in a position characteristic of each virus group. The production of coronavirus subgenomic mRNAs involves the fusion of sequences that are noncontiguous in the viral genome. Several models have been proposed to explain this discontinuous synthesis. Never-

theless, the model of Sawicki and Sawicki (1995), which proposes a discontinuous step during minus-strand RNA synthesis, is best supported by the available biochemical and genetic studies.

Coronavirus reverse genetics was performed in the two last decades using defective genomes, as cDNAs encoding full-length genomes were not available due to the large size of the coronavirus RNAs, posing relevant limitations. During the 1990s, reverse genetics in coronaviruses was possible by targeted recombination developed by Paul Master's group. This useful technology enabled the modification of the coronavirus genome by recombination between a replicating coronavirus genome and nonreplicating or replicating RNAs introduced into the same cell. This technology still remains a very useful tool for modifying the coronavirus genome. Significant progress was made in 2000 with the construction of infectious cDNAs encoding coronavirus genomes using a variety of technologies. Historically, the construction of infectious cDNA clones started with the assembly of QB phage (4.5 kb) cDNA, and was followed by the construction of cDNA of RNA viruses with increasing complexity such as brome mosaic virus (with three RNA segments with sizes between 2.1 and 3.2 kb), poliovirus (7.5 kb), closteroviruses (19 kb) and, finally, coronaviruses (27–32 kb). The two main problems associated with the construction of infectious cDNA clones—the fidelity of reverse transcriptase and the toxicity of sequences derived from eukaryotic viruses in bacteria—were aggravated for coronavirus genomes due to their extremely large size. These problems have been overcome by following different strategies. The first infectious cDNA clone was constructed for TGEV in bacterial artificial chromosomes (BACs). These plasmids presented a reduced toxicity in bacteria since only a single copy, maximum two per cell, is produced. Other approaches were the assembly of a full-length cDNA in vitro, or the use of poxviruses as cloning vectors.

The construction of infectious cDNAs for different coronaviruses [TGEV, HCoV-229E, infectious bronchitis virus (IBV), mouse hepatitis virus (MHV), and SARS-CoV] enables nowadays deep insights into coronavirus replication and transcription mechanisms, virus–host interactions, and development of strategies to protect against coronavirus-induced diseases. This volume consists of eight chapters. The first two provide a perspective on coronavirus genome structure, replication, and transcription. The third chapter concentrates on the replicase, the most complex coronavirus gene. The fourth chapter reviews the viral and cellular proteins involved in coronavirus replication. This chapter includes cellular factors involved in virus–host interaction, a new avenue that is attracting the attention of many scientists. The fifth chapter reviews the design of targeted recombination in coronaviruses and its application to the analysis of the replication and morphogenesis of this virus family. The construction of virus vectors derived from RNA viruses is a comprehensive process that requires for optimum performance the availability of an infectious cDNA clone, knowledge of virus transcription mechanisms to optimize mRNA levels, determination of the essential and nonessential genes to create room for

heterologous genes, and a strategy for vector safety. These aspects are reviewed in the sixth chapter, using TGEV as a model, and also in the last two chapters, using HCoV-229E and MHV genomes. The content of these chapters clearly shows that a new family of virus vectors based on coronavirus genomes has emerged with high potential, due to the variety of tissue and species tropism of these vectors, their large size (providing them with a large cloning capacity), and the possibility of multigenic expression using the extensive repertoire of subgenomic mRNAs that are transcribed in coronaviruses. In addition, in the last chapter the construction of the first infectious cDNA clone of SARS-CoV is described, allowing the study of the molecular biology of this virus in order to understand the molecular basis of its virulence and the development of recombinant vaccines to prevent coronavirus-induced diseases.

I certainly hope that this volume will be useful to academic researchers, scientists involved in human and animal health, and enterprises involved in the fight against coronavirus-induced diseases. I would like to thank all the authors, the staff of Springer, and my colleagues at the CNB, CSIC (Madrid), for their help in the preparation of this book.

L. Enjuanes

List of Contents

List of Contributors

(Their addresses can be found at the beginning of their respective chapters.)

CTMI (2005) 287:1–30

Coronavirus Genome Structure and Replication

D. A. Brian[1] (✉) · R. S. Baric[2, 3]

[1] Departments of Microbiology and Pathobiology, University of Tennessee, College of Veterinary Medicine, Knoxville, TN 37996-0845, USA
dbrian@utk.edu
[2] Department of Microbiology and Immunology, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NV 27599-7400, USA
[3] Department of Epidemiology, Program of Infectious Diseases, School of Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7400, USA

Abstract In addition to the SARS coronavirus (treated separately elsewhere in this volume), the complete genome sequences of six species in the coronavirus genus of the coronavirus family [avian infectious bronchitis virus-Beaudette strain (IBV-Beaudette), bovine coronavirus-ENT strain (BCoV-ENT), human coronavirus-229E strain (HCoV-229E), murine hepatitis virus-A59 strain (MHV-A59), porcine transmissible gastroenteritis-Purdue 115 strain (TGEV-Purdue 115), and porcine epidemic diarrhea virus-CV777 strain (PEDV-CV777)] have now been reported. Their lengths range from 27,317 nt for HCoV-229E to 31,357 nt for the murine hepatitis virus-A59, establishing the coronavirus genome as the largest known among RNA

viruses. The basic organization of the coronavirus genome is shared with other members of the Nidovirus order (the torovirus genus, also in the family *Coronaviridae*, and members of the family *Arteriviridae*) in that the nonstructural proteins involved in proteolytic processing, genome replication, and subgenomic mRNA synthesis (transcription) (an estimated 14–16 end products for coronaviruses) are encoded within the 5′-proximal two-thirds of the genome on gene 1 and the (mostly) structural proteins are encoded within the 3′-proximal one-third of the genome (8–9 genes for coronaviruses). Genes for the major structural proteins in all coronaviruses occur in the 5′ to 3′ order as S, E, M, and N. The precise strategy used by coronaviruses for genome replication is not yet known, but many features have been established. This chapter focuses on some of the known features and presents some current questions regarding genome replication strategy, the *cis*-acting elements necessary for genome replication [as inferred from defective interfering (DI) RNA molecules], the minimum sequence requirements for autonomous replication of an RNA replicon, and the importance of gene order in genome replication.

1 Introduction

Despite its unique property as the largest of the known plus-strand RNA genomes, the coronavirus genome shares with those of other plus-strand RNA viruses (excepting retroviruses) the properties of (1) infectiousness [and not using a packaged RNA-dependent RNA polymerase (RdRp)] (Brian et al. 1980; Schochetman et al. 1977) and (2) replication in the cytoplasm in close association with cellular membranes (Denison et al. 1999; Dennis and Brian 1982; Gosert et al. 2002; Sethna and Brian 1997; Shi et al. 1999; van der Meer et al. 1999). Many of the basic features of coronavirus genome structure and replication have been described in recent reviews (Cavanagh et al. 1997; Enjuanes et al. 2000a, 2000b; Lai and Cavanagh 1997; Lai and Holmes 2001; Luytjes 1995; van der Most and Spaan 1995). With the advent of reverse genetics enabling site-directed mutagenesis of any part of the genome (Almazan et al. 2000; Casais et al. 2001; Masters 1999; Thiel et al. 2001; Yount et al. 2000, 2002), many of the mechanistic features of coronavirus genome replication that could previously be learned only from direct manipulation of defective interfering (DI) RNA can now be examined in the context of the whole virus genome. In this chapter, we review the current knowledge of coronavirus genome structure and organization and the *cis*-acting elements in coronavirus replication and raise selected questions that we believe are important for approaching a better understanding of coronavirus genome replication.

2
Common Features of Genome Structure Among Coronaviruses

In addition to the SARS coronavirus (treated separately elsewhere in this volume), the genomes of six species of coronaviruses have now been fully sequenced and reported in GenBank (as of November 2002): IBV-Beaudette (NC 001451, Boursnell et al. 1987), BCoV-ENT (NC 003045, Chouljenko et al. 2001), MHV-A59 (NC 001846, Leparc-Goffart et al. 1997), HCoV-229E (NC 002645, Herold et al. 1993; Thiel et al. 2001), TGEV-Purdue (NC 002306, Almazan et al. 2000; Eleouet et al. 1995; Penzes et al. 2001), and PEDV-CV777 2001 (NC 003436, Kocherhans et al. 2001). These, representing all three coronavirus serogroups (Siddell 1995), are schematically depicted in Fig. 1. Additional strains of BCoV [BCoV-LUN (AF391542, Chouljenko et al. 2001)], BCoV-Mebus (U00735, Nixon and Brian, unpublished data) and BCoV-Quebec (AF220295, Yoo and Pei 2001), and MHV [MHV-2 (AF201929, Sarma et al. 1999)] have also been reported. The genome sizes range from 27,317 nt for HCoV-229E to 31,357 nt for MHV-A59, establishing them as the largest known among RNA viruses (Enjuanes et al. 2000a; Lai and Cavanagh 1997). The following similarities in genome structure among the six can be noted:

1. The 5′ UTRs ranging in length from 209 to 528 nt contain a similarly positioned short, AUG-initiated open reading frame (ORF) relative to the 5′ end [Table 1; a situation that, by current terminology, is problematic because the "untranslated region" now becomes in part potentially translatable and thus should preferably be called a "leader" (Morris and Geballe 2000). The term "leader," however, has an established meaning in the nidovirus lexicon (Lai and Cavanagh 1997; see subsequent chapters, this volume) of a 5'-terminal, genome-encoded sequence of 65–98 nt appearing on the 5′ terminus of each subgenomic mRNA species]. For purposes of this review, "5′ UTR" will refer to the sequence upstream of ORF 1 (gene 1) despite the internally positioned short ORF. The short AUG-initiated ORFs (except for HCoV-229E) begin in a suboptimal Kozak context for translation (Table 1) (Kozak 1991) and potentially encode peptides of 3–11 amino acids.
2. The 3′ UTRs range from 288 to 506 nt [although some strains of IBV have 3′ UTRs of greater length because of internal sequence duplications (Williams et al. 1993)], all possess an octameric sequence of GGAAGAGC beginning at base 73 to 80 upstream from the poly(A) tail, and all possess a 3′-terminal poly(A) tail (Table 1).

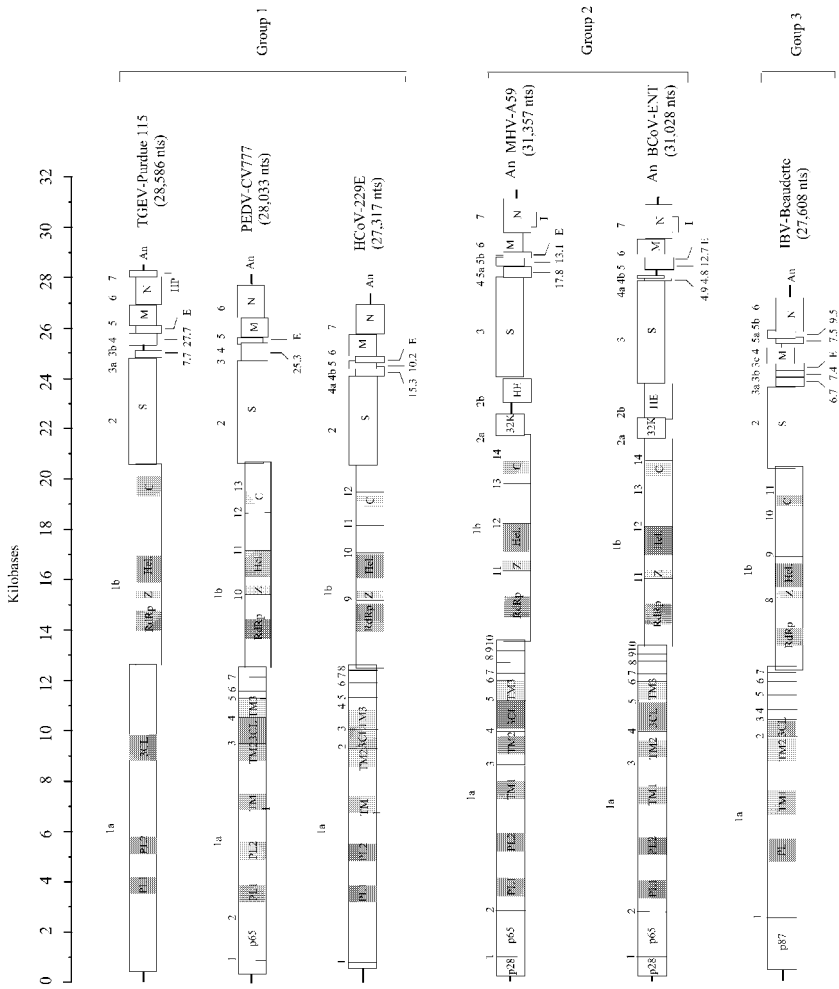


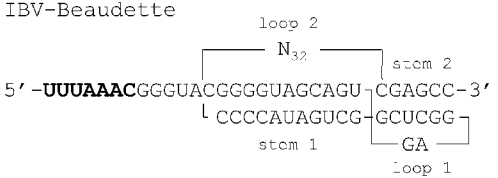
Fig. 1. Genomes of the six sequenced species of coronaviruses known prior to the discovery of the SARS coronavirus. Maps are drawn to approximate scale, and species are shown in decreasing order of size within each of the three groups. The representations are derived from data in the GenBank as of November 2002. For gene 1 (ORFs 1a and 1b) the predicted protease cleavage sites are indicated by numbers and domains of known or predicted function are shaded and identified (*PL*, papain-like protease; *3CL*, poliovirus 3C-like protease; *TM*, transmembrane domain; *RdRp*, RNA-dependent RNA polymerase; *Z*, zinc finger (metal-binding) domain; *Hel*, helicase domain; *C*, conserved sequence domain). Genes 2–8 (or 9) are identified by their transcript name (1a, 1b, etc.) or their abbreviated name of the protein product (*S*, spike; *E*, envelope; *M*, membrane; *N*, nucleocapsid; *HP*, hydrophobic protein; *HE*, hemagglutinin-esterase; *I*, internal). Literature references are described with the GenBank information (see text)

Table 1. Properties of the coronavirus 5′ UTR, intra-5′ UTR short ORF, and 3′ UTR

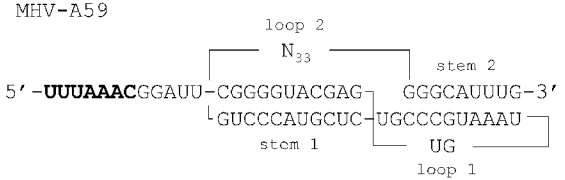
Coronavirus	Length of 5′ UTR (number of nt upstream of gene 1)	Position and Kozak context[a] of the intra-5′ UTR short ORF start codon	Number of amino acids encoded by the 5′ UTR short ORF	Amino acid sequence of the 5′ UTR short ORF product	Length of 3′ UTR (number of nt)	Position of the first nt in the octamer GGAAGAGC upstream from the 3′ poly(A) tail
TGEV-Purdue	314	117UCUaugA	3	MKS	279	76
PEDV-CV777	296	105GUUaugC	10	MLLEAGVEFH	334	73
HCoV-229E	292	86GCUaugG	11	MAGIFDAGVVV[b]	462	74
MHV-A59	209	99UCCaugC	8	MPAGLVLS	324	81
BCoV-ENT	210	100UCUaugC	8	MPVGVDFS[c]	288	78
IBV-Beaudette	528	131UGGaugG	11	MAPGHLSGFCY	506	80

[a] The optimal Kozak context for translation initiation is GCCaugG (Kozak 1991).
[b] A second ORF beginning 16 nt downstream from the first and in the plus 1 reading frame relative to the first encodes the amino acids MLES.
[c] The second amino acid in the BCoV-Mebus strain is L.

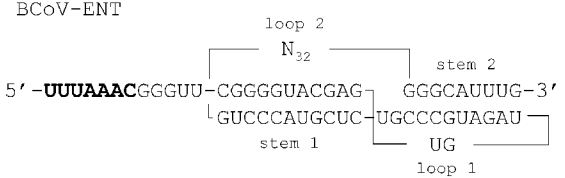
IBV-Beaudette



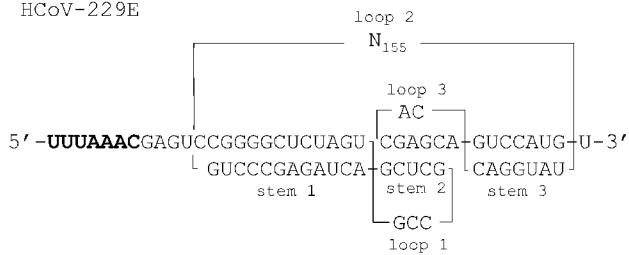
MHV-A59



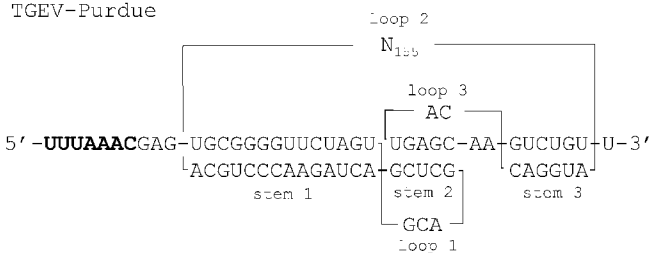
BCoV-ENT



HCoV-229E



TGEV-Purdue



PEDV-CV777