doi:10.1016/j.jmb.2006.09.074 J. Mol. Biol. (2006) 364, 964–973

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Mosaic Structure of Human Coronavirus NL63, One Thousand Years of Evolution

Krzysztof Pyrc[1]*, Ronald Dijkman[1], Lea Deng[2], Maarten F. Jebbink[1] Howard A. Ross[2], Ben Berkhout[1] and Lia van der Hoek[1]*

[1]*Laboratory of Experimental Virology, Department of Medical Microbiology, Center for Infection and Immunity Amsterdam (CINIMA) Academic Medical Center University of Amsterdam Meibergdreef 15, 1105 AZ Amsterdam, The Netherlands*

[2]*School of Biological Sciences and Bioinformatics Institute University of Auckland Private Bag 92019, Auckland New Zealand*

**Corresponding authors*

Before the SARS outbreak only two human coronaviruses (HCoV) were known: HCoV-OC43 and HCoV-229E. With the discovery of SARS-CoV in 2003, a third family member was identified. Soon thereafter, we described the fourth human coronavirus (HCoV-NL63), a virus that has spread worldwide and is associated with croup in children. We report here the complete genome sequence of two HCoV-NL63 clinical isolates, designated Amsterdam 57 and Amsterdam 496. The genomes are 27,538 and 27,550 nucleotides long, respectively, and share the same genome organization. We identified two variable regions, one within the 1a and one within the S gene, whereas the 1b and N genes were most conserved. Phylogenetic analysis revealed that HCoV-NL63 genomes have a mosaic structure with multiple recombination sites. Additionally, employing three different algorithms, we assessed the evolutionary rate for the S gene of group Ib coronaviruses to be $\sim 3 \times 10^{-4}$ substitutions per site per year. Using this evolutionary rate we determined that HCoV-NL63 diverged in the 11th century from its closest relative HCoV-229E.

Keywords: coronavirus; HCoV-NL63; recombination; evolution; molecular clock

Introduction

Coronaviruses, a genus of the *Coronaviridae* family, are enveloped viruses with a large plus-strand RNA genome. The genomic RNA is 27–32 kb in size, capped and polyadenylated. Coronaviruses have been identified in bats, mice, rats, chickens, turkeys, swine, dogs, cats, rabbits, horses, cattle and humans and cause highly prevalent diseases such as respiratory, enteric, cardiovascular and neurological disorders.[1,2] Originally, coronaviruses were classified on the basis of antigenic cross-reactivity in three antigenic groups.[3] When coronavirus genome sequence data began to accumulate, the original antigenic groups were converted into genetic groups based on similarity of the nucleotide sequences.

The coronaviruses possess a characteristic genome composition. The 5′ two-thirds of a coronavirus genome encodes two polyproteins (1a and 1ab) that contain all proteins necessary for RNA replication. The 3′ one-third of a coronavirus genome encodes several structural proteins such as spike (S), envelope (E), membrane (M) and nucleocapsid (N) proteins that, among other functions, participate in the budding process and are incorporated into the virus particle. Additional accessory protein genes are located in the 3′ part of the genome in a coronavirus species-specific position.

HCoV-NL63, a recently discovered member of the *Coronaviridae* family,[4–6] has spread worldwide, is observed most frequently in the winter season and is associated with acute respiratory illness and croup in young children, elderly and immunocompromised patients.[7–10] A recent report suggested that HCoV-NL63 is one of the pathogens underlying Kawasaki disease,[11] although other studies could not confirm this association.[12–14]

HCoV-NL63 belongs to the group I coronaviruses according to phylogenetic analyses. The highest similarity is observed with HCoV-229E and porcine epidemic diarrhoea virus (PEDV), 65% and 61%, respectively. Phylogenetic analysis based on gene 1a sequences indicates the presence of diverse

Abbreviations used: aa, amino acid residue(s); RdRp, RNA-dependent RNA polymerase; ACE, angiotensin converting enzyme.

E-mail addresses of the corresponding authors: k.a.pyrc@amc.uva.nl; c.m.vanderhoek@amc.uva.nl

HCoV-NL63 strains with distinct molecular markers.[4] The increasing number of HCoV-NL63 sequences from several locations provides further evidence for this genetic diversity and confirms the presence of two main genetic clusters.[10,15,16] However, drawing conclusions based on phylogenetic analysis of a single gene sequence and sometimes even a partial gene sequence requires caution as the true phylogeny can only be demonstrated by analyzing complete genome sequences. Full genome sequences of field isolates were, however, not available. Therefore, we sequenced the complete genomes of two HCoV-NL63 field isolates (Amsterdam 57 and Amsterdam 496) and genome fragments of 21 additional field isolates. Here, we present evidence for the in-field recombination of HCoV-NL63. Furthermore, we characterized the molecular variability of HCoV-NL63 isolates to have insight into the evolution of the virus. We observed high variability at certain genome regions and a molecular clock analysis revealed that the virus has been present in the human population for centuries.

Results

HCoV-NL63 isolates

We sequenced the complete genomes of two HCoV-NL63 isolates (Amsterdam 57 and Amsterdam 496) directly from patient material. Isolate 496 was obtained from the throat swab of an eight-month old boy (February 2003) and isolate 57 was amplified from the bronchoalveolar lavage of a 57-year old woman (December 2002), both suffering from acute respiratory illness. Both isolates displayed the same basic genome structure as previously described for HCoV-NL63. Furthermore, we partially sequenced additional isolates directly from patient material to analyze the variability of HCoV-NL63 (Table 1). The regions were: 1a gene nt 3004–3888 (3K) and nt 5815–6280 (6K), S gene nt 20,497–21,003 (21K), ORF3 gene nt 24,521–25,206 (25K), and N gene nt 26,136–27,166 (26K). In some cases only a few regions were amplified because of the low virus load in some patient samples.

Genetic variability along the genome

Pair-wise sequence alignments of isolates Amsterdam 1,[4] NL,[17] 57 and 496 demonstrate an overall genome similarity of 99.0% between the HCoV-NL63 strains. We plotted the frequency of polymorphic nucleotides along the genome to visualize variable sites (Figure 1(a)) and identified two hypervariable regions, one in the 5′ part of the 1a gene encoding nsp1-nsp3 (nt 170–5000) and in the 5′ part of the spike gene (nt 20,300–22,000). The latter region encompasses the S1 region that contains a unique insert in HCoV-NL63 when compared to its closest relative HCoV-229E.

Table 1. Clinical isolates of HCoV-NL63

Isolate name	Sampling date	Viral load in patient sample (copies/ml)	Fragments sequenced
3	n.k.	5.66×10^5	3K
27	n.k.	2.23×10^6	3K; 21K
42	n.k.	n.d.	26K
57	08.01.2004	2.00×10^8	Full genome
63c	n.k.	n.d.	26K
72	31.12.2002	196305	3K; 6K; 21K; 25K; 26K
111	12.01.2004	n.d.	21K
120	13.01.2004	n.d.	21K
173	n.k.	n.d.	3K
202	31.03.2003	n.d.	21K
212	01.02.2003	n.d.	21K; 25K
223	08.01.2003	n.d.	21K; 25K; 26K
242	10.02.2003	n.d.	25K; 26K
246	13.01.2003	1.80×10^4	6K; 21K; 25K; 26K
248	16.01.2003	1.98×10^6	3K; 6K; 25K; 26K
251	07.01.2003	8.34×10^7	25K
466	04.02.2003	5.36×10^5	3K; 6K; 21K; 25K; 26K
496	25.02.2003	1.70×10^6	Full genome
705	n.k.	3.30×10^6	3K; 21K
744	n.k.	7.20×10^6	3K
791	n.k.	3.70×10^6	21K
857	10.12.2003	3.50×10^7	3K; 6K; 21K; 25K; 26K
890	n.k.	7.39×10^5	3K

n.k., not known. n.d., not determined.

Within the variable region 1–5000 nt in the 1a gene, we identified 126 variable sites among the four full genome isolates (55 non-synonymous substitutions), which resulted in 51 variable amino acid (aa) positions. In this region, we also identified an in-frame deletion of 15 nt in isolate 496 and NL (corresponding to nt 3321–3335 of Amsterdam 1). To determine the prevalence of this deletion in the virus population, we analyzed partial sequences of the 1a gene (region 3K) of additional HCoV-NL63 isolates (Table 1) and found it in three more patients (003, 890 and 248) while in eight patients no deletion was observed. The second variable region encompasses the S1 part of the S gene. Out of 175 polymorphic nucleotides (56 non-synonymous substitutions) leading to 51 aa substitutions, 119 are located in the first 1200 nt (46 non-synonymous substitutions) of the spike gene leading to 41 aa changes. Furthermore, we observed a 3 nt deletion within the S gene (corresponding to nt 20798 and 20800 of Amsterdam 1) in isolates 496, 57 and NL. Sequencing of 12 additional patient samples identified no additional variants with this 3 nt deletion.

The analysis of synonymous/non-synonymous substitutions along the genome indicates that synonymous substitutions are generally in excess over non-synonymous substitutions (Figure 1(b)). To determine if the high variability in the 3K and 21K regions is driven by positive selection we analyzed these regions with PAML software, which provides maximum likelihood estimates of the extent of positive selection. Likelihood ratio tests

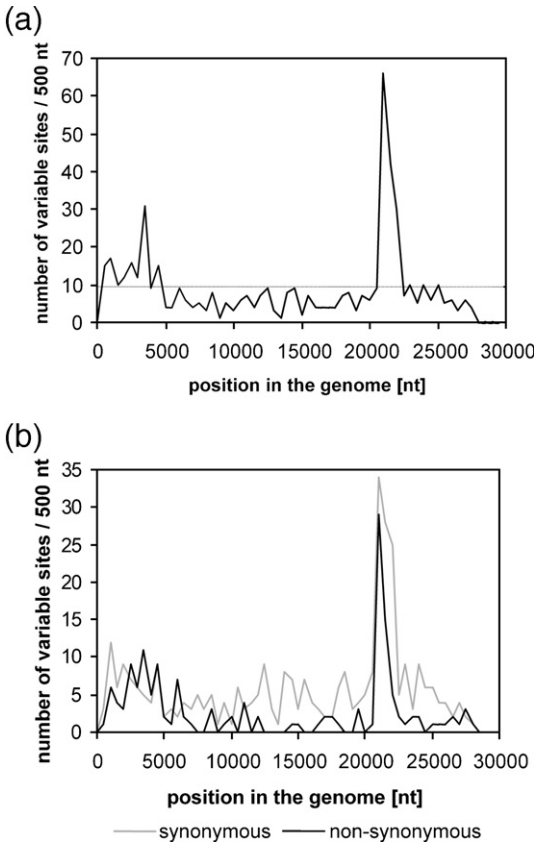

Figure 1. Molecular variability of HCoV-NL63 along the genome. (a) Frequency of polymorphic sites at the nucleotide level among four isolates of HCoV-NL63. (b) Frequency of polymorphic sites on the synonymous and non-synonymous positions among four isolates of HCoV-NL63. The analysis was done with a 500 nt non-overlapping window.

were used to assess whether a model, which included positive selection, was significantly better than one that did not. When positive selection was indicated, empirical Bayes' methods were used to identify which individual sites were under positive selection. According to the PAML analysis the 3K and 21K regions showed no significant sign of positive selection.

We analyzed the most conserved genome regions to identify suitable targets for the development of a PCR-based diagnostic assay that can detect all HCoV-NL63 isolates. The 1b polyprotein gene is highly conserved, with 25 variable nucleotides and only one aa substitution in the region nt 12401–15195 among the four isolates. This region encodes the RNA-dependent RNA polymerase (RdRp). The second most conserved region is the N protein and we confirmed the homogeneity in nine additional patient isolates. Of the 24 variable positions scored in a 1031 nt region, only five were non-silent and resulted in 4 aa changes. Although it was previously mentioned that the ORF3 gene is highly variable in strain NL,[17] we observed a low heterogeneity in Amsterdam 1, 57 and 496. Also among 11 patient isolates, only ten polymorphic nucleotides were observed (one non-synonymous substitution), resulting in only 1 aa change in the patient isolates. A 3 nt insertion and an additional 1 aa change were observed only in the cultured NL isolate.[17]

HCoV-NL63 recombination

We analyzed the full genome sequence of the four HCoV-NL63 isolates for possible recombination events. As the explorative bootscan analysis was not suggestive due to the low number of highly similar sequences and a stochastic noise that could not be distinguished from the real signal, we decided to analyze only the regions showing high number of informative sites. We analyzed the partial sequences of five regions (3K, 6K, 21K, 25K, and 26K) for 57, 496, NL and Amsterdam 1 isolates and nine additional patient isolates. The phylogenetic analysis confirms that in the 3K region the isolates Amsterdam 1 and 57 do cluster together in one subgroup while NL and 496 are located in the second (Figures 2 and 3). Phylogenetic analysis of the 6K region of several isolates reveals that the clustering pattern changes, with isolates NL, 496 and Amsterdam 1 in one subgroup and isolate 57 being an outgroup (Figures 2 and 3). In the 21K region the analysis shows that Amsterdam 1 is a single representative of one cluster, while NL, 496 and 57 are tightly clustering in the second group (Figures 2 and 3). Therefore, there is clear evidence that HCoV-NL63 isolates are mosaics with multiple recombinations along the genome.

Because in several regions the sequence was highly conserved and the analysis did not show signs of the presence of two genetic clusters it was very difficult to identify the exact location of the recombination spots. The S gene does contain enough informative sites, and we identified two spots of recombination: one between positions nt 21,072 and 21,161 and the second between positions nt 21,662 and 21,884 in the Amsterdam 1 genome (Figure 4).

Interspecies recombination

We also analyzed whether recombination within the coronavirus family can be identified. The analysis was performed with the SimPlot software by plotting the similarity between different members of the *Coronaviridae* family as well as by scanning the genome with the bootscan tool. Along the genome the similarity between HCoV-NL63 and HCoV-229E is the highest, except for one part of the M gene. The similarity graph shows that the 3′ region of the M gene has a higher nucleotide similarity to PEDV than to HCoV-229E (Figure 5(a)). Additionally, the bootscan analysis suggests recombination between an ancestral HCoV-NL63 strain and PEDV in that region (Figure 5(b)). To rule out that the observed effect is the result of convergent evolution we analyzed the synonymous substitution pattern between HCoV-NL63 and HCoV-229E. It has been described that the synonymous substitution

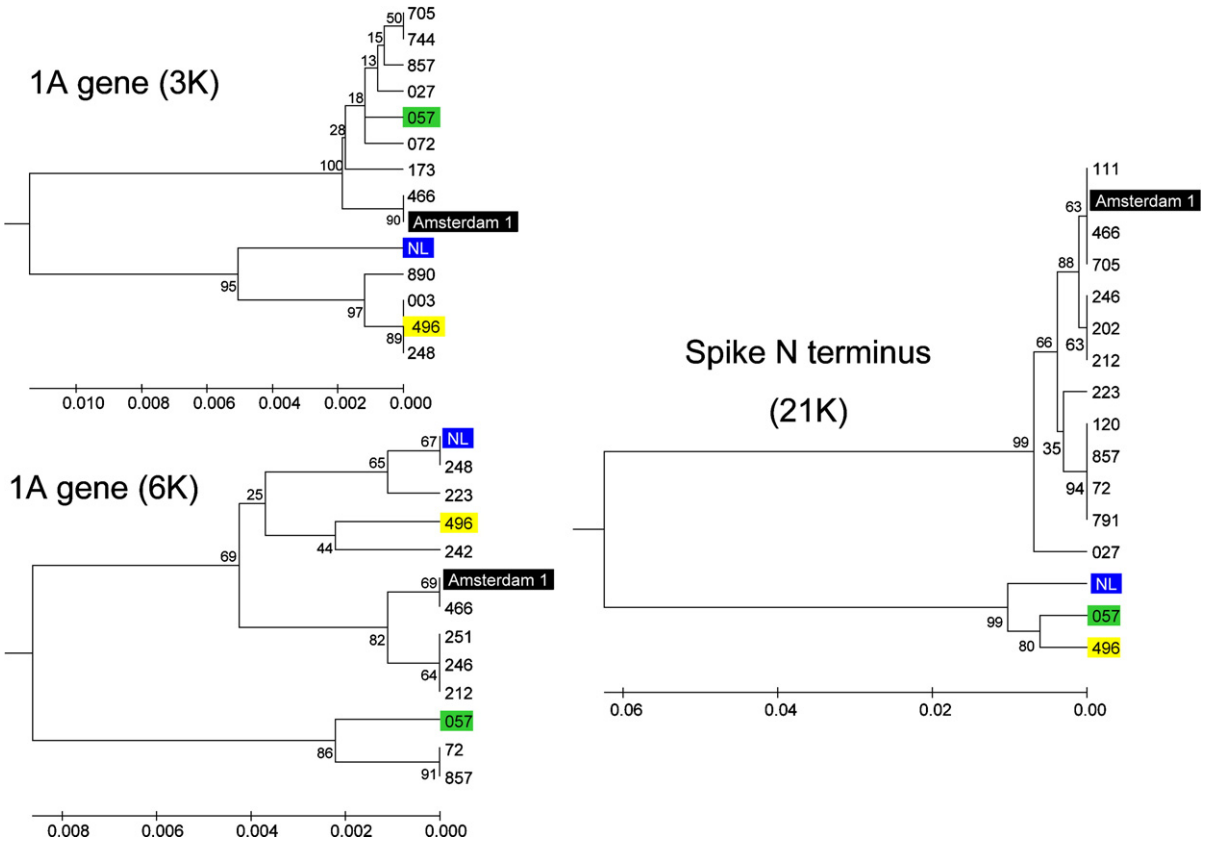

Figure 2. Discordance in phylogenetic clustering of different isolates of HCoV-NL63 at regions 3K, 6K and 21K. Phylogenetic trees were constructed as described in Materials and Methods using an UPGMA algorithm. The scale bar unit represents a 0.002 substitution per site. The trees were rooted with the sequences of it's closest relative: for 3K and 6K the HCoV-229E and for the 21K region the PEDV. Four completely sequenced isolates are marked with colored boxes,

rate is increased at genome regions, which originated from another species and may be used as a marker of gene transfer from another species.[18,19] Indeed a rise in the synonymous substitution rate in the 3′ region of the M gene was observed (Figure 5(c)).

Molecular clock analysis

Based on the nucleotide sequence coding for the S protein (nt 20,649–22,269), a maximum-likelihood phylogenetic tree was constructed for HCoV-NL63 and several HCoV-229E strains for which the date of isolation was known (three isolates from year 1967, five isolates from year 1999 and six isolates from year 2000). Based on these sequence data, the evolutionary rate of HCoV-229E was calculated by Bayesian coalescent approach,[20] serial ML estimate[21,22] and sUPGMA[23] approaches. The evolutionary rates estimated with these three approaches were of very similar magnitude and were 3.28×10^{-4} (95% confidence interval, 1.72×10^{-4} to 5.00×10^{-4}), 6.17×10^{-4} and 2.82×10^{-4} (95% confidence interval, 1.36×10^{-4}–4.42×10^{-4}) substitutions per site per year, respectively. Assuming a constant evolutionary rate in time and between the branches for HCoV-NL63 and HCoV229E, the time to the most recent common ancestor (TMRCA) of HCoV-NL63 and HCoV-229E was dated by the Bayesian coalescent approach around the year 1053 (95% highest posterior density interval, year 966 to 1142). This estimate was highly consistent under different demographic models, including an exponential growth (TMRCA around year 1105 (95% confidence interval, 1017 to 1188)), and expansion growth (TMRCA around year 1124 (95% confidence interval, 1038 to 1206)). A likelihood ratio test indicated that the molecular clock hypothesis could not be rejected ($P=0.05$). We also attempted to date back the split of two HCoV-NL63 lineages, but due to several recombination spots in the spike gene region that we sequenced (see also Figure 4) this analysis was not possible. The 3K and 6K fragments could not be used because we only know the substitution rate for the 20,649–22,269 region in the HCoV-NL63 genome.

Discussion

Homologous recombination is well known for several RNA viruses, including coronaviruses.[24,25] A "copy-choice" mechanism has been proposed, in which the RdRp, together with the nascent RNA strand, dissociates from the original template and re-associates at the same position on another template

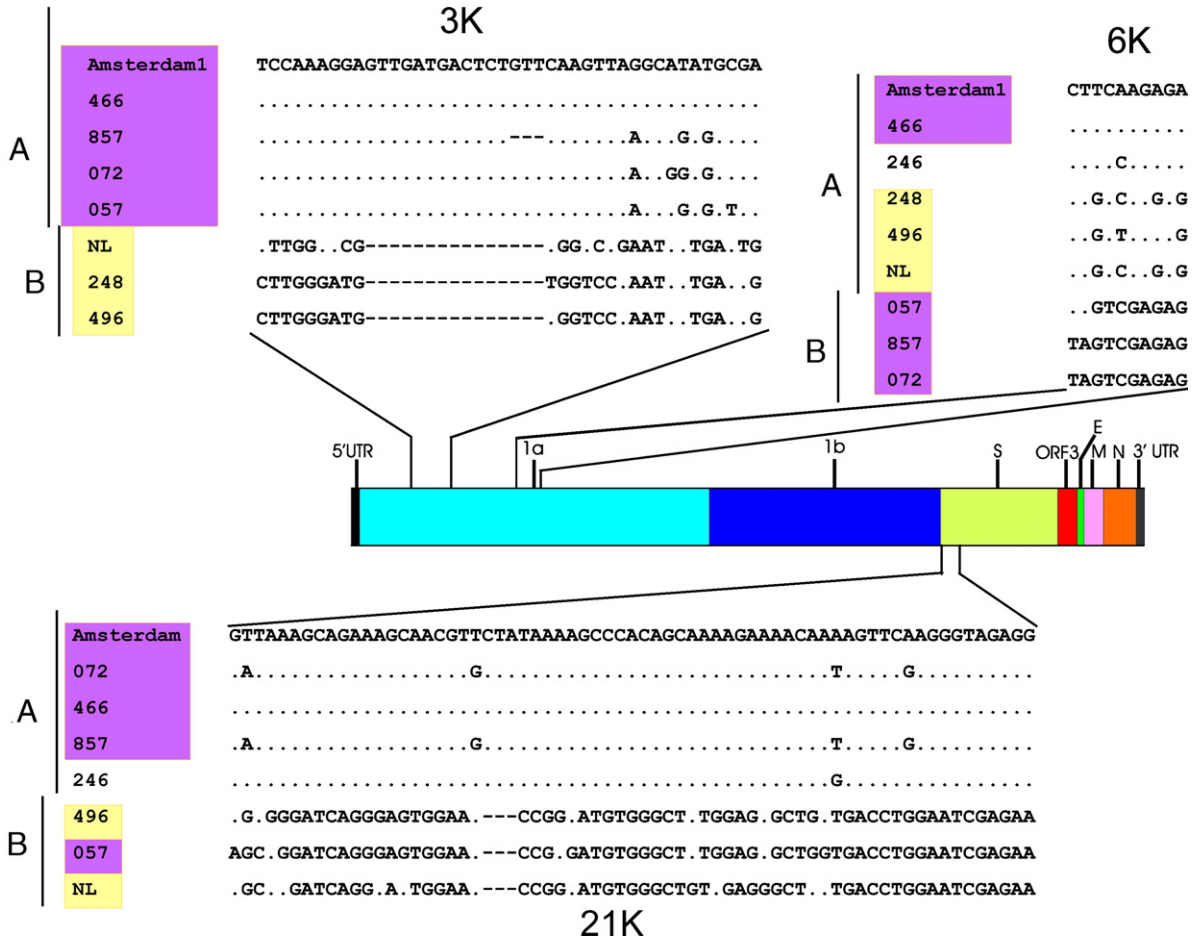

Figure 3. Discordance in clustering of different isolates of HCoV-NL63 at regions 3K, 6K and 21K. Three alignments of only variable sites subtracted from the original sequence with DnaSP 4.0 software, shows the discordance in clustering at different regions of the genome. Groups A and B were created arbitrarily to show the discordance. Group A was defined

subsequently recommencing RNA synthesis.[25] Recombination of coronavirus genomes has been observed *in vitro* in cell culture,[24–26] in experimentally infected animals,[27] and in embryonated eggs.[28] In the case of infectious bronchitis virus, there is evidence for homologous recombination occurring in the field.[29–32] We present evidence that recombination has occurred during the evolution of HCoV-NL63 and that viral isolates possess a mosaic genome structure. Recombination was discovered by full

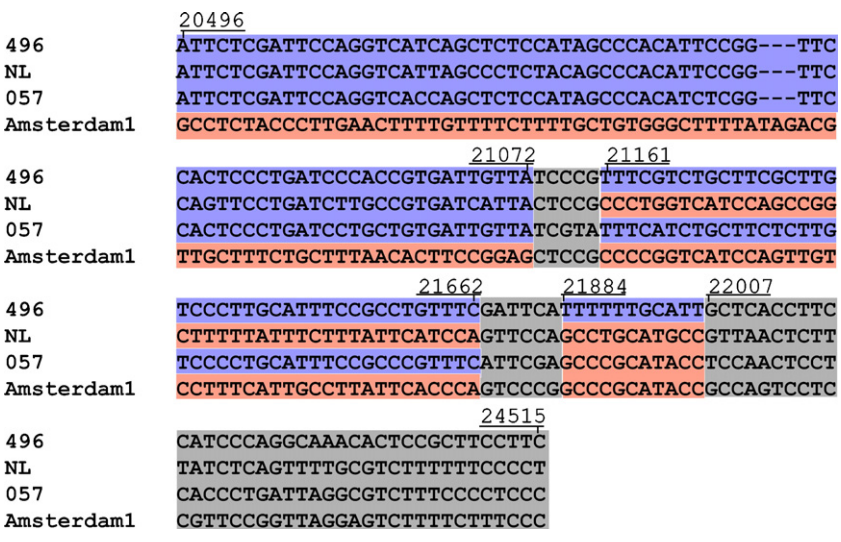

Figure 4. Identification of recombination sites in the S gene. The alignment includes only subtracted variable sites. These variable sites were subtracted with DnaSP 4.0 software. The change of color represents the alternation of genetic clustering between isolates. The numbers at the top represent the beginning of the S gene (nt 20,472 in the Amsterdam 1 isolate), coordinates of recombination spots inside the S gene (nt 21,061–21,072 and 21,575–21,576 in the Amsterdam 1 isolate) and the 3′ terminus of the S gene (nt 24,542 in the Amsterdam 1 isolate).

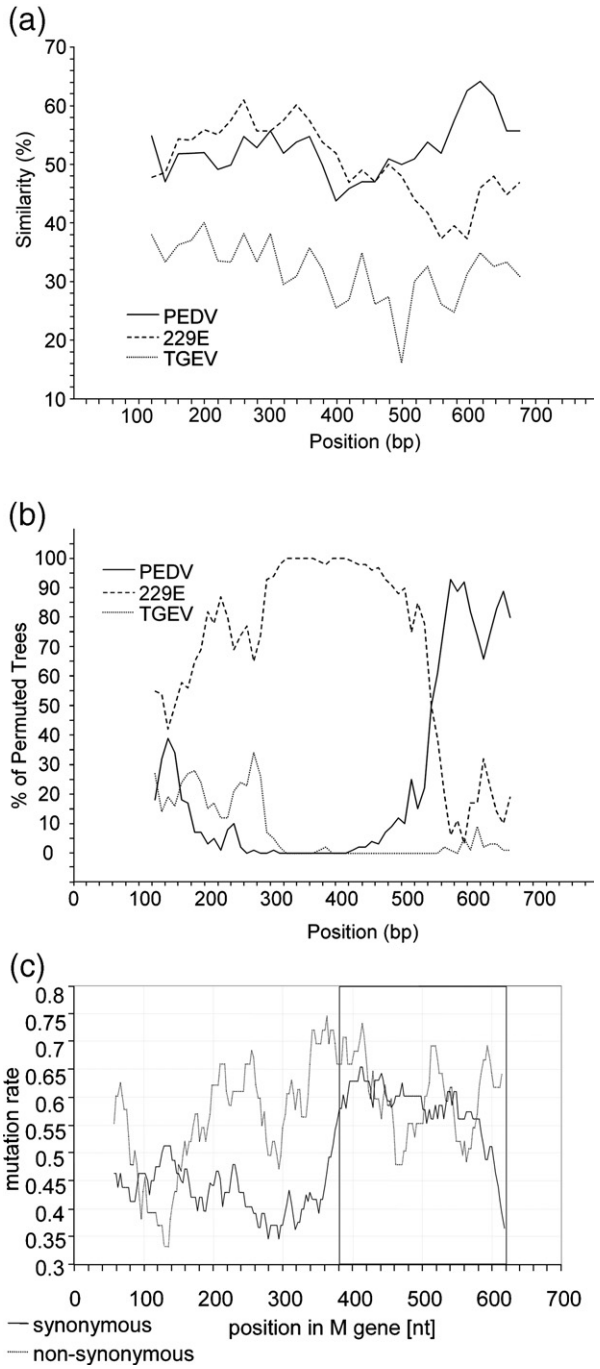

Figure 5. Signs of interspecies recombination in the M gene. Similarity plot (a) and bootscan analysis with the Kimura (two-parameter) distance model, neighbor-joining tree model and 500 bootstrap replicates (b) of the HCoV-NL63 M protein. (c) Analysis of the substitution pattern on the synonymous and non-synonymous level between the M gene of HCoV-NL63 and HCoV-229E. The window used was 40 nucleotides with a step of one nucleotide.

genome sequence analysis of HCoV-NL63 variants from clinical samples. Analysis of several genome regions showed discordance in the phylogenetic clustering along the genome, a clear sign of recombination. Because the majority of informative sites are located at non-coding positions one can exclude that related genome sequences were the result of convergent evolution due to positive selection.

HCoV-NL63 is the causative agent of up to 10% of all respiratory illnesses.[4,7–10,15,33–36] This high prevalence obviously increases the possibility of a recombination event through co-infection with another human or zoonotically transmitted animal coronavirus. Thus, recombination might enable highly pathogenic recombinant virus variants to arise. There is some evidence for recombination between PEDV and an ancestral HCoV-NL63 strain. Whereas HCoV-NL63 is mostly similar to HCoV-229E, a part of the HCoV-NL63 M gene shows the highest similarity to PEDV, suggesting a possible interspecies recombination event. The M protein is a key player in virus assembly and budding, thereby interacting with other structural viral proteins. The M protein of coronaviruses has been shown to span the virion membrane three or four times. The N_{exo}-C_{endo} topology is adopted by most coronaviral M proteins, but it has been proposed that the M protein of TGEV is present in the viral envelope in two topological states, N_{exo}-C_{endo} and N_{exo}-C_{exo}.[37] The sequence analysis with the TMHMM v 2.0 prediction software† suggests a similar scenario for HCoV-NL63 (data not shown). Also the M protein of PEDV, but notably not HCoV-229E, shares these characteristics, which suggests that this domain of the M gene is obtained by interspecies recombination.

Coronaviruses are well equipped to adapt rapidly to changing ecological niches by the high substitution rate of their RNA genome. The average substitution rate for this family was estimated to be 10^{-4} substitutions per year per site.[38,39] HCoV-NL63 is a member of the group I coronaviruses, with highest similarity to HCoV-229E and PEDV. These three species cluster with a recently described bat coronavirus (BatCoV, strain 61)[40] in subgroup Ib. Our efforts to establish the substitution rate of HCoV-NL63 failed, as there is not enough sequence data available from isolates of past years. For this reason we decided to calculate the substitution rate for HCoV-229E, using partial sequences of the S gene from different dates. Based on this substitution rate, we dated the divergence time of HCoV-229E and HCoV-NL63 to the 11th century. The reliability of molecular dating is dependent on the validity of the molecular clock hypothesis, which assumes that the substitution rate is roughly constant. A maximum likelihood test confirmed that the molecular clock hypothesis is suitable for the coronavirus data set investigated here.

We propose that around 900 years ago the HCoV-229E and HCoV-NL63 viruses started to evolve from a common ancestor into the direction of separate species. For HCoV-OC43 it has been estimated that it emerged at the end of the 19th century or the beginning of the 20th century, implying that it has only been around for 100 years.[41,42] The SARS-CoV was introduced in humans in 2002, and for HCoV-HKU1, this has not been investigated. The

† http://www.cbs.dtu.dk/services/TMHMM/

heterogeneity of HCoV-HKU1, which appears in two separate genotypes and a third genotype that is a recombinant of these two genotypes,[43,44] suggests that this virus was not recently introduced into the human population, similar to the situation of HCoV-NL63. The divergence of HCoV-NL63 and HCoV-229E was followed by a separation of HCoV-NL63 into two lineages, of which we suspect that this occurred at geographically distinct locations. Subsequently the two lineages recombined during co-infection, illustrated by the fact that all four full genome sequences available thus far display a mosaic genome organization.

The variability of coronaviruses has not been studied thoroughly. Although there are several reports concerning SARS-CoV, the relatively short time that this virus was present in the human population makes a long term study impossible. The first variable region of HCoV-NL63 encodes the three proteins nsp1–nsp3. The biological function of nsp1 and nsp2 proteins is thought to be linked to virus replication.[45,46] The nsp3 protein is a coronaviral papain-like proteinase (PL^{Pro}) that is expected to be a multifunctional protein with several domains that mediate various enzymatic activities.[47,48] Thus, the high variability of this region might influence the viral replication and interaction with cellular proteins. The second variable region is located in the 5′ part of the S gene. This region contains 24% of all polymorphic nucleotides, whereas it encompasses only ~4% of the genome. The coronavirus S protein is an important determinant for the host cell specificity and tissue tropism, which is largely determined by the distribution of its receptor. Recently, Hofmann *et al*. reported that HCoV-NL63 uses the angiotensin converting enzyme 2 (ACE2) molecule as a receptor. The interaction of NL63-S with ACE2 is surprising, as HCoV-NL63 is closely related to HCoV-229E, which uses CD13 as a receptor. Furthermore, NL63-S shares no appreciable amino acid identity with SARS-CoV-S, which does use ACE2. The amino acid sequence of the CD13-binding site in 229E-S is 57% conserved in NL63-S, whereas the alignment of the ACE2-binding site of SARS-CoV-S with NL63-S reveals only 14% aa identity. These data suggest that NL63-S and SARS-CoV-S might have evolved different strategies to interact with ACE2. The receptor binding domain of NL63-S protein resides in the S1 region.[49] Thus, the variability that we observed may alter the ACE2-binding properties of the S protein or, alternatively, the binding to a co-receptor.

Besides the two hypervariable regions, we also identified regions with a remarkably low substitution rate. The 1b gene is extremely conserved, most prominently in the region encoding RdRp (nt 12416–15195). Analysis of the ORF3 gene shows high conservation of this gene, unlike what has been reported for HCoV-NL63.[17] This suggests a vital function of the ORF3 protein during natural infection. Further investigations on the ORF3 protein function is needed to determine its real biological relevance.

The observation of recombination within the HCoV-NL63 group indicates that two lineages, identified in previous reports in the 1a gene, cannot be treated as separate lineages. Characterization and typing of currently circulating strains should be performed with at least two assays, based on sequences derived from hypervariable regions 1–6000 nt and 20,000–21,000. On the contrary, a sensitive diagnostic assay for detection of HCoV-NL63 should be designed in the regions with highest stability such as the 1b or N gene.

Materials and Methods

Patient isolates

HCoV-NL63 positive patients were identified by a diagnostic nested RT-PCR as described[4] or by a real time PCR that was performed with primers NF (5′ GCGTGTTCCTACCAGAGAGGA 3′) and NR (5′ GCTGTGGAAAACCTTTGGCA 3′), and HCoV-NL63 was detected with probe NP (5′ FAM–ATGTTATTCAGTGCTTTG GTCCTCGTGAT–TAMRA 3′) as described.[13] A total of 23 NL63-positive patients were identified within the Academic Medical Center, Amsterdam. Two HCoV-NL63 isolates were selected for full genome sequencing (57 and 496) and several genome fragments were sequenced for the other 21 isolates (Table 1). Sampling dates and patient characteristics are summarized in Table 1. Sequencing was performed on an ABI 3700 machine (Perkin-Elmer Applied Biosystems) using the BigDye terminator cycle sequencing kit (version 1.1). Chromatogram sequence files were inspected and assembled with CodonCode 1.4 and further corrected manually. Several genomic regions were amplified using the following primers. For the 1a gene: sense 5′ GGTCACTATGTAGTTTATGATG 3′ and sense 5′ GGATTTTTCA TAACCACTTAC 3′; antisense 5′ CTTTTGATAACGGTCACTATG 3′ and antisense 5′ CTCA TTACATAAAACATCAAACGG 3′. For the S gene: sense 5′ GGTTGTTGTTACGCAATAAT GGTCGT 3′; antisense 5′ ACACGGCCATTATGTGTGGT 3′. For ORF3: sense 5′ ATTGTT TAACTTCATCAATGC 3′; antisense 5′ CCATAAAATGGAATTGAGGACAATAC 3′. For N: sense 5′ CTCTCAGGAGGGTGTTTTGTCAGAAAG 3′; antisense 5′ ATAATAAACATTCA ACTGGAATTA C 3′.

RT-PCR and full genome sequencing

The cDNA used for sequencing was generated with MMLV-RT, 1 μg of random hexamer DNA primers, in 10 mM Tris (pH 8.3), 50 mM KCl, 0.1% (v/v) Triton-X100, 6 mM $MgCl_2$ and 50 μM dNTPs at 37 °C for 90 min. The cDNA was converted into double-stranded DNA in a standard PCR reaction with 1.25 units of Taq polymerase (Perkin-Elmer) per reaction and appropriate primers. Full genome sequencing of the two field isolates of HCoV-NL63 was performed with single round RT-PCR as described above, with a set of overlapping PCR products (average size 700 bp) encompassing the entire genome. Primers were designed in regions that are conserved between the Amsterdam 1 and NL isolates of HCoV-NL63. Primer sequences used for full genome sequencing are available on request. The 5′ and 3′-terminal sequence were determined by 5′ RACE (Invitrogen) and 3′ RACE as described.[4] Each PCR fragment was sequenced on both

strands and the virus isolates were amplified and sequenced on separate dates to prevent sample contamination. Each experiment contained negative extraction controls. Sequencing was performed as described above.

Sequence analysis

Multiple sequence alignments were prepared with ClustalX 1.83 and manually edited in BioEdit. Phylogenetic analyses were conducted using MEGA, version 3.1. Bootscan and similarity graphs were prepared with SimPlot 2.5 software.[50] Analysis of HCoV-NL63 variability and synonymous and non-synonymous substitutions was done with DnaSP 4.0 software. Positive selection was analyzed with PAML 3.14 software.[51] The analysis is according to the codon-based evolution models (one-ratio, neutral and selection models) developed by Nielsen and Yang,[52] which allows the ratio of synonymous and non-synonymous substitution rates to vary among amino acid positions. This method uses d_S and d_N to denote the rates of synonymous and non-synonymous substitutions, respectively. Their ratio reflects the selection intensity at the amino acid level.

Molecular clock analysis

Evolutionary rates were estimated using three approaches: Bayesian inference in BEAST, version 1.2 (kindly made available by A. J. Drummond and A. Rambaut, University of Oxford‡) and serial ML estimate and sUPGMA with PEBBLE 1.0.[21] Divergence times were estimated using Bayesian inference in BEAST, version 1.2‡. The Markov chain Monte Carlo (MCMC) length was 10^8 with a sample frequency of 10^3 and effective sample size of 9×10^4. The burn-in was 10^7. Convergence to stationarity was investigated using the Tracer 1.3 MCMC trace analysis tool. The molecular clock hypothesis was tested by the likelihood ratio test.

Nucleotide sequence accession numbers

The sequence of HCoV-NL63 isolate Amsterdam 57 and Amsterdam 496 described here were deposited in GenBank under accession numbers: DQ445911 and DQ445912, respectively. The partial sequences of patient isolates were deposited in GenBank under accession numbers DQ462752-DQ462792. The GenBank accession number of HCoV-NL63 isolate Amsterdam-1 is NC_005831, isolate NL is AY518894, HCoV-229E is AF304460, and PEDV (strain CV777) is AF353511. The GenBank accession numbers of the 6K region of isolates 072, 246, 248 and 466 are AY567494, AY567489, AY567493 and AY567488, respectively.

Acknowledgement

We thank A. de Ronde for critical reading of the manuscript. Lia van der Hoek is supported by VIDI grant 016.066.318 from Netherlands Organization for Scientific Research (NWO).

‡ http://www.evolve.zoo.ox.ac.uk/beast/

References

1. Guy, J. S., Breslin, J. J., Breuhaus, B., Vivrette, S. & Smith, L. G. (2000). Characterization of a coronavirus isolated from a diarrheic foal. *J. Clin. Microbiol.* **38**, 4523–4526.
2. Lai, M. M. C. & Holmes, K. V. (2005). Coronaviridae and their replication. In *Fields-Virology* (Howley, P., Griffin, D., Lamb, R., Martin, M., Roizman, B., Straus, S. & Knipe, D., eds), pp. 1163–1185, Lippincott Williams and Wilkins, London.
3. Cavanagh, D., Brian, D. A., Brinton, M. A., Enjuanes, L., Holmes, K. V., Horzinek, M. C. *et al.* (1993). The Coronaviridae now comprises two genera, coronavirus and torovirus: report of the Coronaviridae Study Group. *Adv. Exp. Med. Biol.* **342**, 255–257.
4. van der Hoek, L., Pyrc, K., Jebbink, M. F., Vermeulen-Oost, W., Berkhout, R. J., Wolthers, K. C. *et al.* (2004). Identification of a new human coronavirus. *Nature Med.* **10**, 368–373.
5. Pyrc, K., Jebbink, M. F., Berkhout, B. & van der Hoek, L. (2004). Genome structure and transcriptional regulation of human coronavirus NL63. *Virol. J.* **1**, 7.
6. Pyrc, K., Bosch, B. J., Berkhout, B., Jebbink, M. F., Dijkman, R., Rottier, P. & van der, H. L. (2006). Inhibition of human coronavirus NL63 infection at early stages of the replication cycle. *Antimicrob. Agents Chemother.* **50**, 2000–2008.
7. van der Hoek, L., Sure, K., Ihorst, G., Stang, A., Pyrc, K., Jebbink, M. F. *et al.* (2005). Croup is associated with the novel coronavirus NL63. *PLoS. Med.* **2**, e240.
8. Ebihara, T., Endo, R., Ma, X., Ishiguro, N. & Kikuta, H. (2005). Detection of human coronavirus NL63 in young children with bronchiolitis. *J. Med. Virol.* **75**, 463–465.
9. Chiu, S. S., Chan, K. H., Chu, K. W., Kwan, S. W., Guan, Y., Poon, L. L. & Peiris, J. S. (2005). Human coronavirus NL63 infection and other coronavirus infections in children hospitalized with acute respiratory disease in Hong Kong. *China. Clin. Infect. Dis.* **40**, 1721–1729.
10. Arden, K. E., Nissen, M. D., Sloots, T. P. & Mackay, I. M. (2005). New human coronavirus, HCoV-NL63, associated with severe lower respiratory tract disease in Australia. *J. Med. Virol.* **75**, 455–462.
11. Esper, F., Shapiro, E. D., Weibel, C., Ferguson, D., Landry, M. L. & Kahn, J. S. (2005). Association between a novel human coronavirus and Kawasaki disease. *J. Infect. Dis.* **191**, 499–502.
12. Ebihara, T., Endo, R., Ma, X., Ishiguro, N. & Kikuta, H. (2005). Lack of association between New Haven coronavirus and Kawasaki disease. *J. Infect. Dis.* **192**, 351–352.
13. Shimizu, C., Shike, H., Baker, S. C., Garcia, F., van der Hoek, L., Kuijpers, T. W. *et al.* (2005). Human coronavirus NL63 is not detected in the respiratory tracts of children with acute Kawasaki disease. *J. Infect. Dis.* **192**, 1767–1771.
14. Chang, L. Y., Chiang, B. L., Kao, C. L., Wu, M. H., Chen, P. J., Berkhout, B. *et al.* (2006). Lack of association between infection with a novel human coronavirus (HCoV), HCoV-NH, and Kawasaki disease in Taiwan. *J. Infect. Dis.* **193**, 283–286.
15. Bastien, N., Anderson, K., Hart, L., Van Caeseele, P., Brandt, K., Milley, D., Hatchette, T., Weiss, E. C. & Li, Y. (2005). Human coronavirus NL63 infection in Canada. *J. Infect. Dis.* **191**, 503–506.
16. Moes, E., Vijgen, L., Keyaerts, E., Zlateva, K., Li, S., Maes, P. *et al.* (2005). A novel pancoronavirus RT-PCR

assay: frequent detection of human coronavirus NL63 in children hospitalized with respiratory tract infections in Belgium. *BMC Infect. Dis.* **5**, 6.
17. Fouchier, R. A., Hartwig, N. G., Bestebroer, T. M., Niemeyer, B., de Jong, J. C., Simon, J. H. & Osterhaus, A. D. (2004). A previously undescribed coronavirus associated with respiratory disease in humans. *Proc. Natl Acad. Sci. USA*, **101**, 6212–6216.
18. Sharp, P. M., Rogers, M. S. & McConnell, D. J. (1984). Selection pressures on codon usage in the complete genome of bacteriophage T7. *J. Mol. Evol.* **21**, 150–160.
19. Smits, S. L., Lavazza, A., Matiz, K., Horzinek, M. C., Koopmans, M. P. & de Groot, R. J. (2003). Phylogenetic and evolutionary relationships among torovirus field variants: evidence for multiple intertypic recombination events. *J. Virol.* **77**, 9567–9577.
20. Drummond, A. J., Nicholls, G. K., Rodrigo, A. G. & Solomon, W. (2002). Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, **161**, 1307–1320.
21. Goode, M. & Rodrigo, A. G. (2004). Using PEBBLE for the evolutionary analysis of serially sampled molecular sequences. In *Current Protocols in Bioinformatics* (Baxevanis, A. D., Davison, D. B., Page, R. D. M., Petsko, G. A., Stein, L. D. & Stormo, G. D., eds), pp. 6.8.1–6.8.26, John Wiley and Sons, Inc., New York.
22. Drummond, A., Forsberg, R. & Rodrigo, A. G. (2001). The inference of stepwise changes in substitution rates using serial sequence samples. *Mol. Biol. Evol.* **18**, 1365–1371.
23. Drummond, A. & Rodrigo, A. G. (2000). Reconstructing genealogies of serial samples under the assumption of a molecular clock using serial-sample UPGMA. *Mol. Biol.Evol.* **17**, 1807–1815.
24. Lai, M. M., Baric, R. S., Makino, S., Keck, J. G., Egbert, J., Leibowitz, J. L. & Stohlman, S. A. (1985). Recombination between non-segmented RNA genomes of murine coronaviruses. *J. Virol.* **56**, 449–456.
25. Makino, S., Keck, J. G., Stohlman, S. A. & Lai, M. M. (1986). High-frequency RNA recombination of murine coronaviruses. *J. Virol.* **57**, 729–737.
26. Baric, R. S., Stohlman, S. A., Razavi, M. K. & Lai, M. M. (1985). Characterization of leader-related small RNAs in coronavirus-infected cells: further evidence for leader-primed mechanism of transcription. *Virus Res.* **3**, 19–33.
27. Keck, J. G., Matsushima, G. K., Makino, S., Fleming, J. O., Vannier, D. M., Stohlman, S. A. & Lai, M. M. (1988). In vivo RNA-RNA recombination of coronavirus in mouse brain. *J. Virol.* **62**, 1810–1813.
28. Kottier, S. A., Cavanagh, D. & Britton, P. (1995). Experimental evidence of recombination in coronavirus infectious bronchitis virus. *Virology*, **213**, 569–580.
29. Jia, W., Karaca, K., Parrish, C. R. & Naqi, S. A. (1995). A novel variant of avian infectious bronchitis virus resulting from recombination among three different strains. *Arch. Virol.* **140**, 259–271.
30. Kusters, J. G., Jager, E. J., Niesters, H. G. & van der Zeijst, B. A. (1990). Sequence evidence for RNA recombination in field isolates of avian coronavirus infectious bronchitis virus. *Vaccine*, **8**, 605–608.
31. Wang, L., Junker, D. & Collisson, E. W. (1993). Evidence of natural recombination within the S1 gene of infectious bronchitis virus. *Virology*, **192**, 710–716.
32. Herrewegh, A. A., Smeenk, I., Horzinek, M. C., Rottier, P. J. & de Groot, R. J. (1998). Feline coronavirus type II strains 79–1683 and 79–1146 originate from a double recombination between feline coronavirus type I and canine coronavirus. *J .Virol.* **72**, 4508–4514.
33. Kaiser, L., Regamey, N., Roiha, H., Deffernez, C. & Frey, U. (2005). Human coronavirus NL63 associated with lower respiratory tract symptoms in early life. *Pediatr. Infect. Dis. J.* **24**, 1015–1017.
34. Bastien, N., Robinson, J. L., Tse, A., Lee, B. E., Hart, L. & Li, Y. (2005). Human coronavirus NL-63 infections in children: a 1-year study. *J. Clin. Microbiol.* **43**, 4567–4573.
35. Vabret, A., Mourez, T., Dina, J., van der Hoek, L., Gouarin, S., Petitjean, J. *et al.* (2005). Human coronavirus NL63, France. *Emerg. Infect. Dis.* **11**, 1225–1229.
36. Suzuki, A., Okamoto, M., Ohmi, A., Watanabe, O., Miyabayashi, S. & Nishimura, H. (2005). Detection of human coronavirus-NL63 in children in Japan. *Pediatr. Infect. Dis. J.* **24**, 645–646.
37. Risco, C., Anton, I. M., Sune, C., Pedregosa, A. M., Martin-Alonso, J. M., Parra, F. *et al.* (1995). Membrane protein molecules of transmissible gastroenteritis coronavirus also expose the carboxy-terminal region on the external surface of the virion. *J. Virol.* **69**, 5269–5277.
38. Sanchez, C. M., Gebauer, F., Sune, C., Mendez, A., Dopazo, J. & Enjuanes, L. (1992). Genetic evolution and tropism of transmissible gastroenteritis coronaviruses. *Virology*, **190**, 92–105.
39. Vijgen, L., Lemey, P., Keyaerts, E. & Van Ranst, M. (2005). Genetic variability of human respiratory coronavirus OC43. *J .Virol.* **79**, 3223–3224.
40. Poon, L. L., Chu, D. K., Chan, K. H., Wong, O. K., Ellis, T. M., Leung, Y. H. *et al.* (2005). Identification of a novel coronavirus in bats. *J. Virol.* **79**, 2001–2009.
41. Vijgen, L., Keyaerts, E., Moes, E., Thoelen, I., Wollants, E., Lemey, P. *et al.* (2005). Complete genomic sequence of human coronavirus OC43: molecular clock analysis suggests a relatively recent zoonotic coronavirus transmission event. *J .Virol.* **79**, 1595–1604.
42. Vijgen, L., Keyaerts, E., Lemey, P., Maes, P., Van Reeth, K., Nauwynck, H. *et al.* (2006). Evolutionary history of the closely related group 2 coronaviruses: porcine hemagglutinating encephalomyelitis virus, bovine coronavirus, and human coronavirus OC43. *J. Virol.* **80**, 7270–7274.
43. Woo, P. C., Lau, S. K., Chu, C. M., Chan, K. H., Tsoi, H. W., Huang, Y. *et al.* (2005). Characterization and complete genome sequence of a novel coronavirus, coronavirus HKU1, from patients with pneumonia. *J. Virol.* **79**, 884–895.
44. Woo, P. C., Lau, S. K., Yip, C. C., Huang, Y., Tsoi, H. W., Chan, K. H. & Yuen, K. Y. (2006). Comparative analysis of 22 coronavirus HKU1 genomes reveals a novel genotype and evidence of natural recombination in coronavirus HKU1. *J. Virol.* **80**, 7136–7145.
45. Brockway, S. M. & Denison, M. R. (2005). Mutagenesis of the murine hepatitis virus nsp1-coding region identifies residues important for protein processing, viral RNA synthesis, and viral replication. *Virology*, **340**, 209–223.
46. Graham, R. L., Sims, A. C., Brockway, S. M., Baric, R. S. & Denison, M. R. (2005). The nsp2 replicase proteins of murine hepatitis virus and severe acute respiratory syndrome coronavirus are dispensable for viral replication. *J. Virol.* **79**, 13399–13411.
47. Shi, S. T., Schiller, J. J., Kanjanahaluethai, A., Baker, S. C., Oh, J. W. & Lai, M. M. (1999). Colocalization and membrane association of murine hepatitis virus

gene 1 products and de novo-synthesized viral RNA in infected cells. *J. Virol.* **73**, 5957–5969.
48. Tijms, M. A., van Dinten, L. C., Gorbalenya, A. E. & Snijder, E. J. (2001). A zinc finger-containing papain-like protease couples sub-genomic mRNA synthesis to genome translation in a positive-stranded RNA virus. *Proc. Natl Acad. Sci. USA*, **98**, 1889–1894.
49. Hofmann, H., Pyrc, K., van der Hoek, L., Geier, M., Berkhout, B. & Pohlmann, S. (2005). Human coronavirus NL63 employs the severe acute respiratory syndrome coronavirus receptor for cellular entry. *Proc. Natl Acad. Sci. USA*, **102**, 7988–7993.
50. Lole, K. S., Bollinger, R. C., Paranjape, R. S., Gadkari, D., Kulkarni, S. S., Novak, N. G. *et al.* (1999). Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *J. Virol.* **73**, 152–160.
51. Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. BioSci.* **13**, 555–556.
52. Nielsen, R. & Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, **148**, 929–936.

Edited by J. Karn

(Received 25 August 2006; received in revised form 24 September 2006; accepted 25 September 2006)
Available online 3 October 2006