

## **Phylogenetic and recombination analysis of coronavirus HKU1, a novel coronavirus from patients with pneumonia**

**P. C. Y. Woo<sup>1,2,3,\*</sup>, S. K. P. Lau<sup>1,2,3,\*</sup>, Y. Huang<sup>1,\*</sup>, H.-W. Tsoi<sup>1</sup>,  
K.-H. Chan<sup>1</sup>, and K.-Y. Yuen<sup>1,2,3</sup>**

<sup>1</sup>Department of Microbiology, Faculty of Medicine,  
The University of Hong Kong, Hong Kong

<sup>2</sup>Research Centre of Infection and Immunology, Faculty of Medicine,  
The University of Hong Kong, Hong Kong

<sup>3</sup>State Key Laboratory of Emerging Infectious Diseases,  
The University of Hong Kong, Hong Kong

Received February 25, 2005; accepted April 27, 2005  
Published online June 28, 2005 © Springer-Verlag 2005

**Summary.** Phylogenetic trees constructed using predicted amino acid sequences of putative proteins of coronavirus HKU1 (CoV-HKU1) revealed that CoV-HKU1 formed a distinct branch among group 2 coronaviruses. Of the 14 trees from p65 to nsp10, nine showed that CoV-HKU1 was clustered with murine hepatitis virus. From nsp11, the topologies of the trees changed dramatically. For the eight trees from nsp11 to N, seven showed that the CoV-HKU1 branch was the first branch. The codon usage patterns of CoV-HKU1 differed significantly from those in other group 2 coronaviruses. Split decomposition analysis revealed that recombination events had occurred between CoV-HKU1 and other coronaviruses.

### **Introduction**

It has been estimated that coronaviruses [human coronaviruses 229E (HCoV-229E) and OC43 (HCoV-OC43)] cause about 5–30% of respiratory tract infections. In late 2002 and 2003, Severe Acute Respiratory Syndrome (SARS), caused by SARS coronavirus (SARS-CoV), has resulted in more than 750 deaths [12, 15, 16, 17, 22–24]. In early 2004, a novel coronavirus associated with respiratory tract infections, human coronavirus NL63 (HCoV-NL63), was discovered [3, 20]. As a result of a unique mechanism of viral replication, coronaviruses have a high frequency of recombination [9, 10, 13, 14].

\*These authors contributed equally to the manuscript.

Coronaviruses were divided into three groups, with HCoV-229E and HCoV-NL63 being group 1 coronaviruses and HCoV-OC43 a group 2 coronavirus respectively [11]. For SARS-CoV, it was initially proposed that SARS-CoV constituted a distinct group of coronavirus [15, 17]. However, after more extensive phylogenetic analysis, it was discovered that SARS-CoV probably represents a distant relative of group 2 coronaviruses [2, 18]. Further *in silico* analysis also predicted that SARS-CoV could be a product of recombination between mammalian and avian coronaviruses [19].

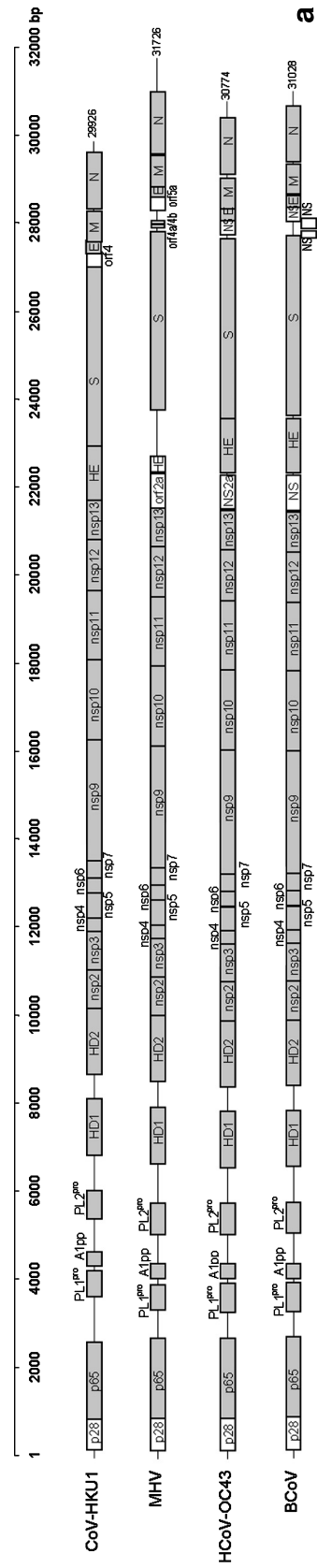
Recently, we have described the discovery of a novel coronavirus associated with pneumonia, coronavirus HKU1 (CoV-HKU1) [21]. Based on analysis of the putative chymotrypsin-like protease (3CL<sup>pro</sup>), RNA-dependent RNA polymerase (Pol), helicase, hemagglutinin-esterase (HE), spike (S), envelope (E), membrane (M) and nucleocapsid (N), CoV-HKU1 is a member of group 2 coronaviruses. However, the origin of CoV-HKU1 is still unknown. In this study, we performed a detailed phylogenetic analysis of CoV-HKU1. Possible recombination events were predicted and the origin of CoV-HKU1 discussed.

### Materials and methods

The predicted amino acid (a.a.) sequences of p65, conserved portions of nsp1 [papain-like protease 1 (PL1<sup>pro</sup>), Appr-1-p processing enzyme family (A1pp), papain-like protease 2 (PL2<sup>pro</sup>), hydrophobic domain 1 (HD1), and hydrophobic domain 2 (HD2)], nsp2–7, nsp9–13, HE, S, E, M and N were extracted from the CoV-HKU1 genome sequence (GenBank accession no. AY597011) [21]. The corresponding a.a. sequences of murine hepatitis virus (MHV), HCoV-OC43, bovine coronavirus (BCoV), porcine hemagglutinating encephalomyelitis virus (PHEV), rat sialodacryoadenitis coronavirus (SDAV) and puffinosis virus (PV) were extracted from complete genome sequences of MHV (GenBank accession no. AF201929), HCoV-OC43 (GenBank accession no. AY585229) and BCoV (GenBank accession no. NC\_003045), and sequences of PHEV, SDAV and PV available in GenBank. The a.a. sequence of HE of MHV was extracted from MHV strain JHM (GenBank accession no. BAA00661) because the HE gene in MHV (GenBank accession no. AF201929) stopped prematurely after the 97th a.a. Phylogenetic tree construction was performed using neighbour joining method with ClustalX 1.83. The corresponding a.a. sequences of HCoV-229E were used as outgroups, except for p65 and HE because these were not available in the genome of HCoV-229E. For p65 and HE, the corresponding a.a. sequences in SARS-CoV and influenza C virus were used as the outgroups respectively. Phylogenetic trees were not constructed for p28 and the predicted hypothetical protein of ORF4 and ORF8 in CoV-HKU1 because no a.a. sequences that can be used as the appropriate outgroups can be found.

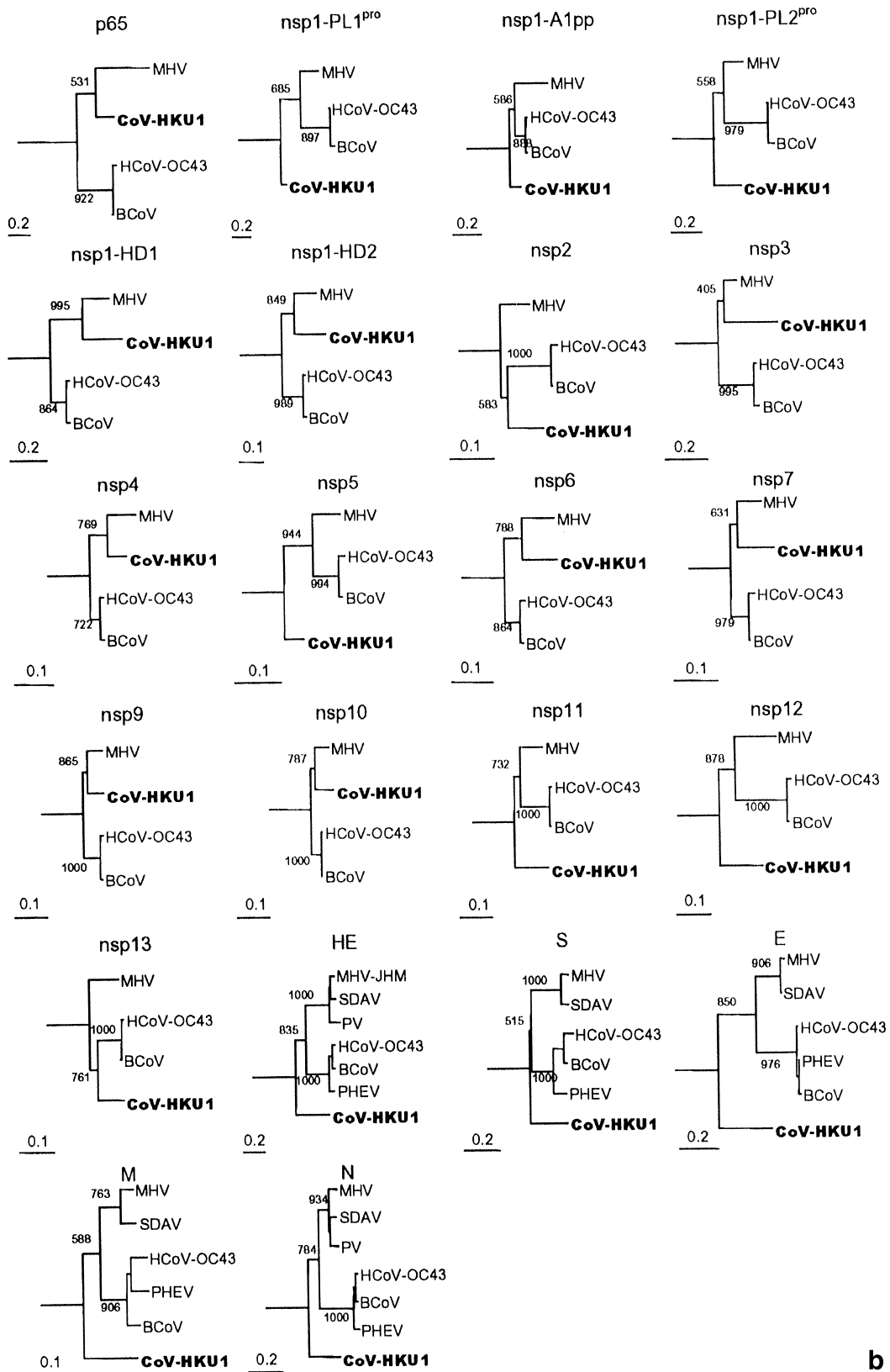
The amino-terminal 800 a.a. residues of the S proteins in various group 1 coronaviruses [porcine transmissible gastroenteritis virus (TGEV), HCoV-NL63 and HCoV-229E], various group 2 coronaviruses (PHEV, SDAV, MHV, HCoV-OC43 and BCoV), infectious bronchitis virus (IBV) (a group 3 coronavirus), SARS-CoV and CoV-HKU1 were aligned using ClustalX 1.83. The presence and positions of conserved cysteine residues in the various peptides were compared.

Correspondence analysis was used to compare the codon usage pattern variation in the different genes among group 2 coronaviruses in a multidimensional space [5]. All available sequences of ORF 1ab, HE, S, M and N of MHV, HCoV-OC43, BCoV, PHEV, SDAV, PV and SARS-CoV were downloaded from the GenBank (Table 1). Analysis of codon usage in these



**a**

**Fig. 1** (continued)



**b**

sequences and the corresponding ones in CoV-HKU1 was performed using CodonW (<http://www.molbiol.ox.ac.uk/cu/>), with each gene represented as a 59 dimensional vector, representing the 59 possible sense codons. AUG, the only codon for methionine, UGG, the only codon for tryptophan, and the three stop codons were excluded. The ORF for E was excluded because the length of the gene was too short.

To delineate the importance of recombination on the evolution of CoV-HKU1, split decomposition analysis was performed. Deduced a.a. sequences of group 1, 2 and 3 coronaviruses and SARS-CoV available in GenBank, that were homologous to 3CL<sup>pro</sup>, Pol, helicase, HE, S, ORF4, E, M and N in CoV-HKU1 [21], were retrieved. Split decomposition analysis was performed with SplitsTree version 3.2 [7] using Hamming correction and is presented with the same edge length.

## Results

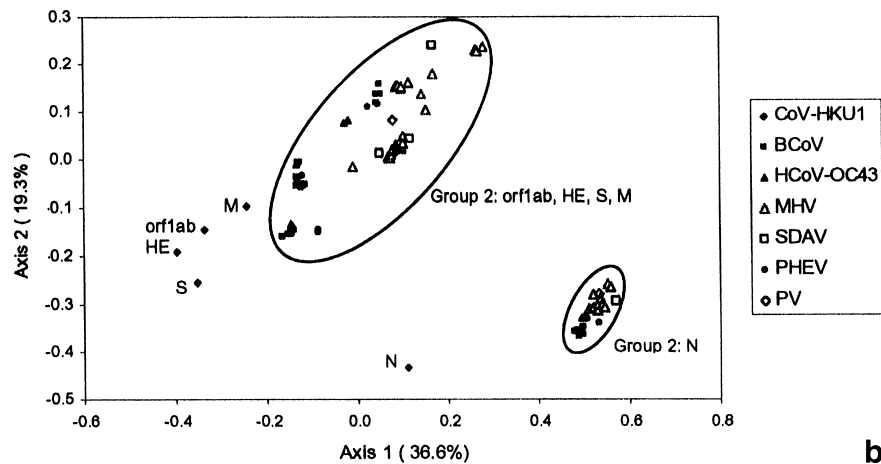
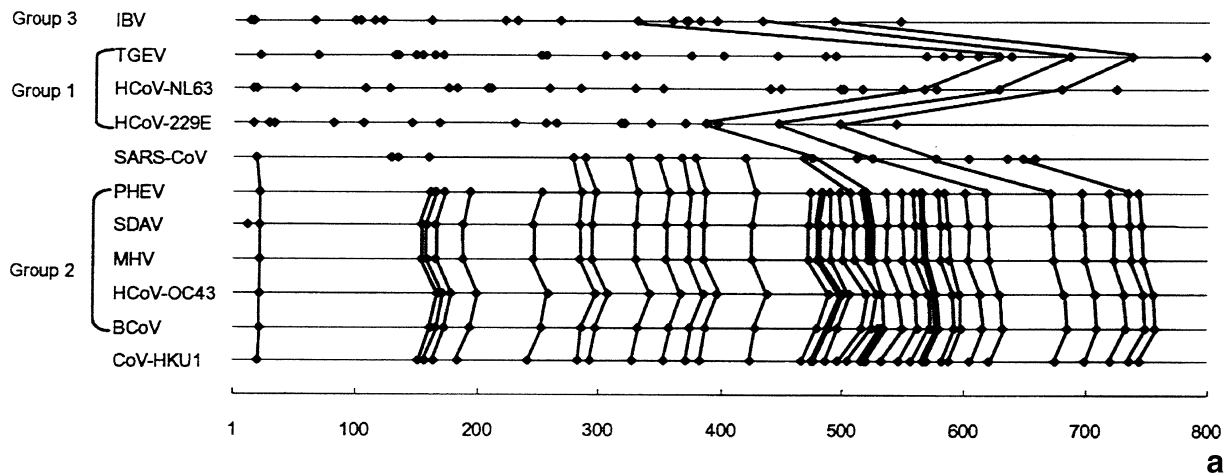
The genome organizations of CoV-HKU1 and other group 2 coronaviruses were shown in Fig. 1a. Phylogenetic trees using predicted a.a. sequences of putative proteins and polypeptides of CoV-HKU1 and other group 2 coronaviruses were constructed (Fig. 1b). The putative proteins and polypeptides included p65, conserved portions of nsp1 (PL1<sup>pro</sup>, A1pp, PL2<sup>pro</sup>, HD1 and HD2), nsp2-7, nsp9-13, HE, S, E, M and N. All trees revealed that CoV-HKU1 formed a distinct branch among group 2 coronaviruses. Interestingly, of the 14 trees of p65 to nsp10, nine (64%) (p65, HD1, HD2, nsp3, nsp4, nsp6, nsp7, nsp9 and nsp10) showed that CoV-HKU1 was clustered with MHV (Fig. 1b). However, for the eight trees of nsp11 to N, seven (88%) showed that the CoV-HKU1 branch appeared as the first branch among group 2 coronaviruses (Fig. 1b).

Comparison of the cysteine residues in the N-terminal 800 a.a. residues of S in CoV-HKU1 and those in the different groups of coronaviruses revealed that almost all the conserved cysteine residues in group 2 coronaviruses were present in CoV-HKU1 (Fig. 2a), supporting that CoV-HKU1 is a member of group 2 coronaviruses.

The number of ORF 1ab, HE, S, M and N sequences in the group 2 coronaviruses used for correspondence analysis is shown in Table 1. The results of the

←

**Fig. 1.** Genome organization and phylogenetic analysis of CoV-HKU1. **a** Genome organization of CoV-HKU1 (GenBank accession no. AY597011), MHV (GenBank accession no. AF201929), HCoV-OC43 (GenBank accession no. AY585229) and BCoV (GenBank accession no. NC\_003045). The homologous regions used for phylogenetic analysis were shaded. **b** Phylogenetic analysis of p65, conserved portions of nsp1 (PL1<sup>pro</sup>, A1pp, PL2<sup>pro</sup>, HD1 and HD2), nsp2-7, nsp9-13, HE, S, E, M and N in group 2 coronaviruses. The trees were constructed by neighbour joining method using Jukes-Cantor correction and bootstrap values calculated from 1000 trees. 578, 204, 107, 212, 421, 496, 303, 287, 89, 197, 110, 137, 928, 595, 521, 374, 299, 424, 1287, 84, 226 and 445 a.a. positions in p65, PL1<sup>pro</sup>, A1pp, PL2<sup>pro</sup>, HD1, HD2, nsp2, nsp3, nsp4, nsp5, nsp6, nsp7, nsp9, nsp10, nsp11, nsp12, nsp13, HE, S, E, M and N respectively were included in the analysis. The scale bar indicates the estimated number of substitutions per 5 or 10 a.a. as indicated. The corresponding a.a. sequences of HCoV-229E were used as the outgroups, except for p65 and HE, for which the corresponding a.a. sequences in SARS-CoV and influenza C virus were used as the outgroups respectively



**Fig. 2.** Analysis of cysteine positions in the N-terminal 800 a.a. residues of S and codon usage patterns of CoV-HKU1. **a** Schematic representation of cysteine positions (◆) in the N-terminal domain of S in CoV-HKU1 in comparison with those in other coronaviruses. Conserved cysteine residues of S in different coronaviruses are joined by solid lines. The bar indicates the a.a. residue positions on S. **b** A scattered plot of the scores for the codon usage patterns of ORF 1ab, HE, S, M and N in MHV, HCoV-OC43, BCoV, PHEV, SDAV, PV and CoV-HKU1 on the first and second axis

correspondence analysis with respect to axis 1 and 2 are shown in Fig. 2b. Axis 1 and 2 explained 36.6% and 19.3% of the variations in codon usage respectively. For ORF 1ab, HE, S and M, the scores on axis 1 in group 2 coronaviruses other than CoV-HKU1 were clustered between  $-0.16$  and  $0.28$  and those in CoV-HKU1 were clustered between  $-0.40$  and  $-0.24$  (Fig. 2b). For N, the scores on axis 1 in group 2 coronaviruses other than CoV-HKU1 were clustered between  $0.48$  and  $0.57$  and that in CoV-HKU1 was at  $0.11$  (Fig. 2b). These indicated that the codon usage patterns in the genes in CoV-HKU1 differed significantly from those in other group 2 coronaviruses.

**Table 1.** Number of ORF 1ab, hemagglutinin-esterase (HE), spike (S), membrane (M) and nucleocapsid (N) sequences in the various groups of coronaviruses used for correspondence analysis

ORF	No. of sequences used <sup>a</sup>							
	MHV	HCoV-OC43	BCoV	PHEV	SDAV	PV	SARS-CoV	CoV-HKU1
ORF 1ab	7	3	4	0	0	0	2	1
HE	3	3	8	2	1	1	0	1
S	12	3	9	2	1	0	2	1
M	7	3	6	2	1	0	2	1
N	11	3	7	2	1	1	2	1

<sup>a</sup>HCoV-OC43, human coronavirus OC43; MHV, murine hepatitis virus; BCoV, bovine coronavirus; SDAV, rat sialodacryoadenitis coronavirus; PHEV, porcine hemagglutinating encephalomyelitis virus; PV, puffinosis virus; SARS-CoV, SARS coronavirus; CoV-HKU1, human coronavirus HKU1

Split decomposition analysis revealed that recombination events had occurred between CoV-HKU1 and other group 2 coronaviruses in 3CL<sup>pro</sup>, Pol, helicase, HE, S, ORF4, E and M (Fig. 3). No evidence of recombination was shown between the N of CoV-HKU1 and those of other group 2 coronaviruses.

### Discussion

CoV-HKU1 is a distinct member of group 2 coronaviruses. It was confirmed by both phylogenetic analysis of 22 protein coding regions (Fig. 1b) and analysis of the conserved cysteine residues in the amino-terminal of the S proteins (Fig. 2a) that CoV-HKU1 is a group 2 coronavirus. Furthermore, phylogenetic analysis of the 22 protein coding regions revealed that there were 10–54% a.a. differences between a particular protein coding region in CoV-HKU1 and the corresponding region in the most closely related sequence, indicating that CoV-HKU1 is distinct from the other group 2 coronaviruses. This fact was further supported by results of correspondence analysis of codon usage (Fig. 2b).

Recombination events were common among CoV-HKU1 and other group 2 coronaviruses. Coronaviruses have high frequency of homologous RNA recombination, which has been observed in both tissue culture [10, 14] and experimentally infected animals [8]. In split tree analysis, recombination events would result in reticulations instead of simple branching structures. As shown in Fig. 3, recombination was particularly frequent in CoV-HKU1 and MHV as compared to other group 2 coronaviruses such as BCoV and HCoV-OC43. The particular high recombination frequency in MHV [1] is in line with evidence of a lot of inter-strain recombination, as shown by the high number of reticulations in various ORFs of the different MHV strains (Fig. 3). Complete genome sequencing of additional CoV-HKU1 and further split tree analysis would shed light on whether CoV-HKU1 behaves more like MHV or BCoV and HCoV-OC43.

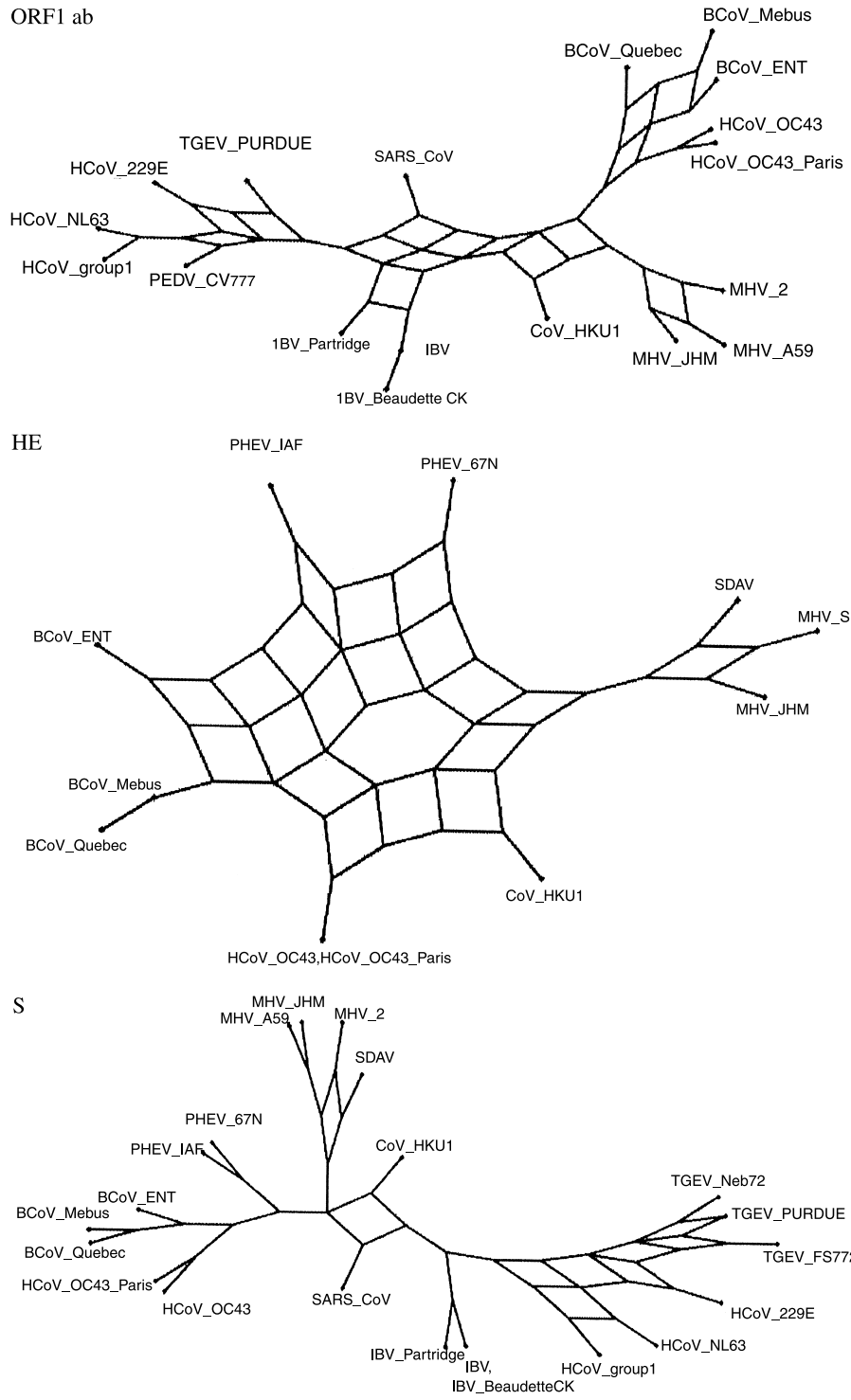


Fig. 3 (continued)



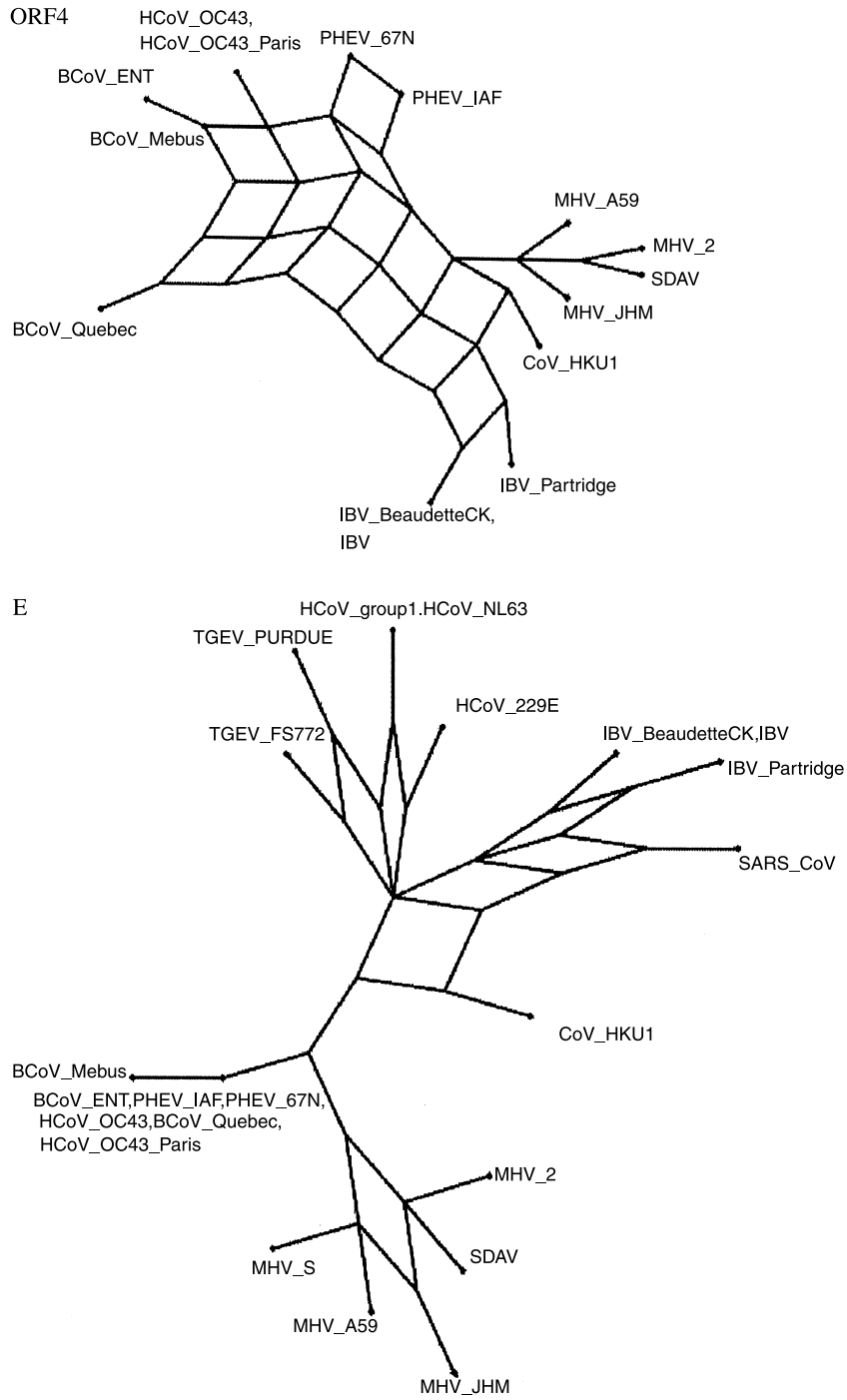
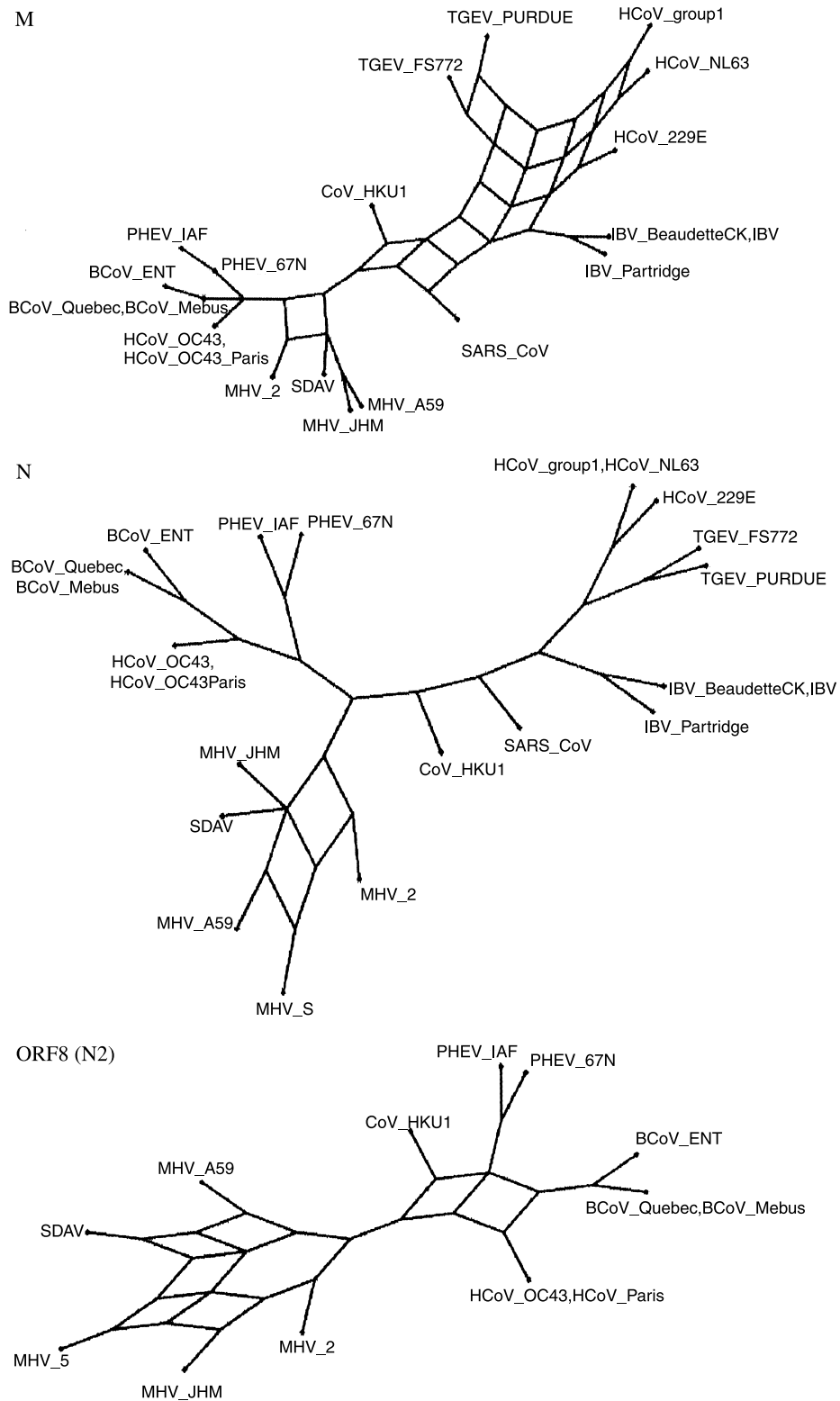


Fig. 3 (continued)



**Fig. 3.** Split decomposition graph of 3CL<sup>pro</sup>, Pol, helicase, HE, S, ORF4, E, M and N in the CoV-HKU1 genome

CoV-HKU1 may have originated from a major recombination event and numerous minor recombination events among group 2 coronaviruses. In feline coronavirus, the site of recombination has been pinpointed to a region of about 50 nucleotides in the M gene by multiple alignment [6]. As for recombination between different strains of MHV, *in vitro* studies have shown both variable sites and rates of recombination, with the S gene have a frequency three fold that of the polymerase gene [4, 14]. In CoV-HKU1, nine of the 14 phylogenetic trees constructed using deduced a.a. sequences of p65 to nsp10 showed that CoV-HKU1 was clustered with MHV (Fig. 1b). Interestingly, the topologies of the phylogenetic trees changed dramatically from nsp11. For the eight trees from nsp11 to N, seven revealed that the CoV-HKU1 branch appeared as the first branch among the group 2 coronaviruses (Fig. 1b) ( $P < 0.01$  by chi-square test). A logical explanation was that a major recombination event has taken place in the region between nsp10 and nsp11 when CoV-HKU1 first appeared. However, this recombination event was not evident in multiple alignment performed at the junction between nsp10 and nsp11 (data not shown). This is because although CoV-HKU1 is more clustered with MHV from p65 to nsp10, the difference in phylogenetic distances between CoV-HKU1 and MHV and those between CoV-HKU1 and BCoV/HCoV-OC43 is not marked (Fig. 1b), in contrast to what was observed in feline coronavirus [6]. Furthermore, bootscanning analysis in the whole genome did not reveal any putative recombination break point (data not shown). We speculate that this could be due to numerous minor recombination events between p65 and nsp10, such as between p65 and nsp1-PL1<sup>Pro</sup>, between nsp1-PL2<sup>Pro</sup> and nsp1-HD1, between nsp4 and nsp5, and between nsp5 and nsp6. This has resulted in CoV-HKU1 being clustered with MHV in only nine of the 14 phylogenetic trees constructed using deduced a.a. from p65 to nsp10, but four of the 14 trees with the CoV-HKU1 branch being the first branch among the group 2 coronaviruses.

### Acknowledgements

This work is partly supported by the Research Grant Council Grant and Research Fund for the Control of Infectious Diseases of the Health, Welfare and Food Bureau of the Hong Kong SAR Government.

### References

1. Baric RS, Fu K, Schaad MC, Stohlman SA (1990) Establishing a genetic recombination map for murine coronavirus strain A59 complementation groups. *Virology* 177: 646–656
2. Eickmann M, Becker S, Klenk HD, Doerr HW, Stadler K, Censini S, Guidotti S, Masignani V, Scarselli M, Mora M, Donati C, Han JH, Song HC, Abrignani S, Covacci A, Rappuoli R (2003) Phylogeny of the SARS coronavirus. *Science* 302: 1504–1505
3. Fouchier RA, Hartwig NG, Bestebroer TM, Niemeyer B, de Jong JC, Simon JH, Osterhaus AD (2004) A previously undescribed coronavirus associated with respiratory disease in humans. *Proc Natl Acad Sci USA* 101: 6212–6216
4. Fu K, Baric RS (1992) Evidence for variable rates of recombination in the MHV genome. *Virology* 189: 88–102

5. Grantham R, Gautier C, Gouy M (1980) Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. *Nucleic Acids Res* 8: 1893–1912
6. Herrewegh AA, Smeenk I, Horzinek MC, Rottier PJ, de Groot RJ (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
7. Huson DH (1998) SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* 14: 68–73
8. Keck JG, Matsushima GK, Makino S, Fleming JO, Vannier DM, Stohlman SA, Lai MM (1988) In vivo RNA-RNA recombination of coronavirus in mouse brain. *J Virol* 62: 1810–1813
9. Kusters JG, Jager EJ, Niesters HG, van der Zeijst BA (1990) Sequence evidence for RNA recombination in field isolates of avian coronavirus infectious bronchitis virus. *Vaccine* 8: 605–608
10. Lai MM, Baric RS, Makino S, Keck JG, Egbert J, Leibowitz JL, Stohlman SA (1985) Recombination between nonsegmented RNA genomes of murine coronaviruses. *J Virol* 56: 449–456
11. Lai MM, Cavanagh D (1997) The molecular biology of coronaviruses. *Adv Virus Res* 48: 1–100
12. Lau SK, Woo PC, Wong BH, Woo GK, Poon RW, Tsoi HW, Chan KH, Wei WI, Peiris JS, Yuen KY (2004) Detection of severe acute respiratory syndrome (SARS) coronavirus nucleocapsid protein in SARS patients by enzyme-linked immunosorbent assay. *J Clin Microbiol* 42: 2884–2889
13. Lee CW, Jackwood MW (2000) Evidence of genetic diversity generated by recombination among avian coronavirus IBV. *Arch Virol* 145: 2135–2148
14. Makino S, Keck JG, Stohlman SA, Lai MM (1986) High-frequency RNA recombination of murine coronaviruses. *J Virol* 57: 729–737
15. Marra MA, Jones SJ, Astell CR, Holt RA, Brooks-Wilson A, Butterfield YS, Khattra J, Asano JK, Barber SA, Chan SY, Cloutier A, Coughlin SM, Freeman D, Girm N, Griffith OL, Leach SR, Mayo M, McDonald H, Montgomery SB, Pandoh PK, Petrescu AS, Robertson AG, Schein JE, Siddiqui A, Smailus DE, Stott JM, Yang GS, Plummer F, Andonov A, Artsob H, Bastien N, Bernard K, Booth TF, Bowness D, Czub M, Drebot M, Fernando L, Flick R, Garbutt M, Gray M, Grolla A, Jones S, Feldmann H, Meyers A, Kabani A, Li Y, Normand S, Stroher U, Tipples GA, Tyler S, Vogrig R, Ward D, Watson B, Brunham RC, Kraiden M, Petric M, Skowronski DM, Upton C, Roper RL (2003) The Genome sequence of the SARS-associated coronavirus. *Science* 300: 1399–1404
16. Peiris JS, Lai ST, Poon LL, Guan Y, Yam LY, Lim W, Nicholls J, Yee WK, Yan WW, Cheung MT, Cheng VC, Chan KH, Tsang DN, Yung RW, Ng TK, Yuen KY, SARS study group (2003) Coronavirus as a possible cause of severe acute respiratory syndrome. *Lancet* 361: 1319–1325
17. Rota PA, Oberste MS, Monroe SS, Nix WA, Campagnoli R, Icenogle JP, Penaranda S, Bankamp B, Maher K, Chen MH, Tong S, Tamin A, Lowe L, Frace M, DeRisi JL, Chen Q, Wang D, Erdman DD, Peret TC, Burns C, Ksiazek TG, Rollin PE, Sanchez A, Liffick S, Holloway B, Limor J, McCaustland K, Olsen-Rasmussen M, Fouchier R, Gunther S, Osterhaus AD, Drost C, Pallansch MA, Anderson LJ, Bellini WJ (2003) Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science* 300: 1394–1399
18. Snijder EJ, Bredenbeek PJ, Dobbe JC, Thiel V, Ziebuhr J, Poon LL, Guan Y, Rozanov M, Spaan WJ, Gorbalenya AE (2003) Unique and conserved features of genome and

- proteome of SARS-coronavirus, an early split-off from the coronavirus group 2 lineage. *J Mol Biol* 331: 991–1004
19. Stavriniades J, Guttman DS (2004) Mosaic evolution of the severe acute respiratory syndrome coronavirus. *J Virol* 78: 76–82
  20. van der Hoek L, Pyrc K, Jebbink MF, Vermeulen-Oost W, Berkhout RJ, Wolthers KC, Wertheim-Van Dillen PM, Kaandorp J, Spaargaren J, Berkhout B (2004) Identification of a new human coronavirus. *Nat Med* 10: 368–373
  21. Woo PC, Lau SK, Chu CM, Chan KH, Tsoi HW, Huang Y, Wong BH, Poon RW, Cai JJ, Luk WK, Poon LL, Wong SS, Guan Y, Malik JS, Yuen KY (2005) Characterization and complete genome sequence of a novel coronavirus, coronavirus HKU1, from patients with pneumonia. *J Virol* 79: 884–895
  22. Woo PC, Lau SK, Tsoi HW, Chan KH, Wong BH, Che XY, Tam VK, Tam SC, Cheng VC, Hung IF, Wong SS, Zheng BJ, Guan Y, Yuen KY (2004) Relative rates of non-pneumonic SARS coronavirus infection and SARS coronavirus pneumonia. *Lancet* 363: 841–845
  23. Woo PC, Lau SK, Wong BH, Chan KH, Chu CM, Tsoi HW, Huang Y, Peiris JS, Yuen KY (2004) Longitudinal profile of immunoglobulin G (IgG), IgM, and IgA antibodies against the severe acute respiratory syndrome (SARS) coronavirus nucleocapsid protein in patients with pneumonia due to the SARS coronavirus. *Clin Diagn Lab Immunol* 11: 665–668
  24. Woo PC, Lau SK, Wong BH, Tsoi HW, Fung AM, Chan KH, Tam VK, Peiris JS, Yuen KY (2004) Detection of specific antibodies to severe acute respiratory syndrome (SARS) coronavirus nucleocapsid protein for serodiagnosis of SARS coronavirus pneumonia. *J Clin Microbiol* 42: 2306–2309

Author's address: Kwok-yung Yuen, Department of Microbiology, The University of Hong Kong, University Pathology Building, Queen Mary Hospital, Pokfulam, Hong Kong; e-mail: hkumicro@hkucc.hku.hk