DOI 10.1099/vir.0.81662-0

Short Communication

Analysis of human coronavirus 229E spike and nucleoprotein genes demonstrates genetic drift between chronologically distinct strains

Doris Chibo and Chris Birch

Victorian Infectious Diseases Reference Laboratory, 10 Wreckyn Street, North Melbourne, Victoria 3051, Australia

Correspondence
Doris Chibo
Doris.chibo@mh.org.au

Received 7 November 2005
Accepted 16 January 2006

Historically, coronaviruses have been recognized as a cause of minor respiratory infections in humans. However, the recent identification of three novel human coronaviruses, one causing severe acute respiratory syndrome (SARS), has prompted further examination of these viruses. Previous studies of geographically and chronologically distinct *Human coronavirus 229E* (HCoV-229E) isolates have found only limited variation within S gene nucleotide sequences. In contrast, analysis of the S genes of contemporary *Human coronavirus OC43* variants identified in Belgium revealed two distinct viruses circulating during 2003 and 2004. Here, the S and N gene sequences of 25 HCoV-229E variants identified in Victoria, Australia, between 1979 and 2004 in patients with symptomatic infections were determined. Phylogenetic analysis showed clustering of the isolates into four groups, with evidence of increasing divergence with time. Evidence of positive selection in the S gene was also established.

Coronaviruses are enveloped, positive-stranded RNA viruses with a genome of approximately 30 kb (Lai, 1990). In animals they are significant veterinary pathogens, often causing severe disease (Holmes, 2001). Until recently, only two human coronaviruses were recognized, *Human coronavirus 229E* (HCoV-229E) and *Human coronavirus OC43* (HCoV-OC43). These viruses have generally been associated with symptoms of the common cold (McIntosh, 1996) but there is growing evidence that they occasionally cause more severe infections (Pene *et al.*, 2003; Vabret *et al.*, 2003; Birch *et al.*, 2005). In late 2002, a third human coronavirus (SARS-CoV) was implicated as the aetiological agent of severe acute respiratory syndrome (SARS) (Drosten *et al.*, 2003). Since then, two more human coronaviruses have been identified, HCoV-NL63 associated with upper and lower respiratory tract infections (van der Hoek *et al.*, 2004) and HCoV-HKU1 in patients with pneumonia (Woo *et al.*, 2005). Based on their serological and genetic properties, the human coronaviruses are assigned to two of three existing antigenic serogroups: HCoV-229E and HCoV-NL63 in serogroup 1 (Siddell, 1995; van der Hoek *et al.*, 2004); HCoV-OC43, HCoV-HKU1 and, most likely, SARS-CoV in serogroup 2 (Siddell, 1995; Snijder *et al.*, 2003; Woo *et al.*, 2005).

Coronaviruses contain four structural proteins, spike (S), membrane (M), small envelope (E) and nucleocapsid (N) (Brian & Baric, 2005). A fifth, the haemagglutinin-esterase (HE) protein, exists in some serogroup 2 coronaviruses (Holmes & Lai, 1996). Limited functional and structural information exists for the HCoV-229E S protein, although receptor-binding activity is associated with the S1 subunit (Bonavia *et al.*, 2003). In the related non-human *Infectious bronchitis virus*, S1 is involved in the induction of neutralizing, serotype-specific and haemagglutination inhibiting antibodies (Cavanagh *et al.*, 1988).

Both recombination and point mutations have the capacity to drive coronavirus evolution (Navas-Martin & Weiss, 2003). Although recombination has not been detected for HCoV-229E or HCoV-OC43, a recombination breakpoint in the SARS-CoV polymerase has been described (Rest & Mindell, 2003). Few studies on variation in the human coronavirus S protein have been reported. One study of three geographically and chronologically distinct HCoV-229E isolates found limited variation within S gene nucleotide sequences (Hays & Myint, 1998). In contrast, two distinct variants of HCoV-OC43 circulated in Belgium between 2003 and 2004 (Vijgen *et al.*, 2005).

Prior to 2001, in our laboratory coronaviruses were identified using a method involving isolation from human fibroblasts, followed by confirmatory immune electron microscopy where these agents were suspected. Virus isolation attempts

The GenBank/EMBL/DDBJ accession numbers for the sequences reported in this paper are DQ243940–DQ243962 and DQ243964–DQ243987.

A figure showing the phylogenetic analysis using PHYLIP on HCoV-229E isolates and strains identified between 1972 and 2004, and tables showing primers used and their nucleotide position relative to the HCoV-229E complete genome and the summary of amino acid changes in the S and N proteins of HCoV-229E variants studied are available as supplementary material in JGV Online.

0008-1662 © 2006 SGM *Printed in Great Britain*

on respiratory specimens were discontinued in 2001 and replaced by a RT-PCR capable of differentiating between HCoV-229E and HCoV-OC43 (Birch *et al.*, 2005). RNA was extracted from virus isolates and clinical material by using a High Pure Viral Nucleic Acid kit (Roche Diagnostics) and reverse transcribed using avian myeloblastosis virus reverse transcriptase (Promega) at 42 °C for 1 h using random hexamers. The full S and N gene sequences of HCoV-229E were amplified in a nested PCR (Supplementary Table S1 available in JGV Online) using a Qiagen *Taq* DNA polymerase kit (Qiagen). The cycling programmes consisted of an initial denaturation of 4 min at 94 °C, followed by 40 cycles (first round) or 25 cycles (second round) of 30 s at 94 °C, 30 s at 60 °C and 130 s for the S gene or 90 s for the N gene at 72 °C, with a final extension of 7 min at 72 °C. RT-PCR products were purified, sequenced in both directions using a cycle sequencing reaction (ABI Prism Big Dye Terminator Cycle Sequencing Ready Reaction kit; Perkin-Elmer) and analysed by using an ABI 3730S capillary sequencer.

Nucleotide sequences were analysed and amino acid sequences determined using BIOEDIT sequence alignment editor version 7.0.1 (Hall, 1999). Alignments of nucleotide sequences were made with Multalin (Corpet, 1988) and manual editing of alignments was performed using Genedoc version 2.5 (Nicholas & Nicholas, 1997). Expected transition/transversion ratios and gamma distribution parameter alpha were estimated using TREEPUZZLE version 5.2 (Schmidt *et al.*, 2002). Phylogenetic trees based on the optimum alignment were constructed using DNAdist and neighbour-joining method with PHYLIP version 3.63 (Felsenstein, 1993) and parameters estimated from TREEPUZZLE. Unrooted phylograms were drawn with TREEVIEW version 1.5 (Page, 1996). Tests of selection were conducted using MEGA version 3.1 (Kumar *et al.*, 2004). The codon-based model of Nei–Gojobori was used to compare synonymous (dS) and non-synonymous (dN) distances. Assumptions tested were purifying ($dN < dS$), neutral ($dN = dS$) and positive ($dN > dS$) selection. The probability computed was $<0{\cdot}05$ for hypothesis rejection at the 5 % level. Potential *N*-glycosylation sites in the HCoV-229E S protein were predicted using the NetNGlyc 1.0 Server at the Centre for Biological Sequence Analysis (http://www.cbs.dtu.dk/services/NetNGlyc/).

Between 1979 and 2004, HCoV-229E was isolated or detected by RT-PCR in 25 patients aged from 0·3 to 58 years of age (Table 1). Patients suffered from symptoms varying from mild upper respiratory tract infection to pneumonia, and three (patients 12, 18 and 19) were hospitalized. Several presented with fever, fatigue and cough, consistent with a case definition of influenza-like illness (Clothier *et al.*, 2005). A total of 25 HCoV-229E sequences were analysed. Twenty-one were full-length S gene sequences, 13 from virus isolates and eight directly from clinical material. Partial S gene sequences were obtained from two samples of clinical material (patients 19 and 25 in Table 1). S gene sequences could not be obtained from the viruses detected in patients 18 and 24. All S gene sequences available were compared to an ATCC prototype strain (HCoV-229E 1973; GenBank accession no. DQ243963, containing 3522 nt, 1173 aa).

The HCoV-229E S gene encodes a 15 aa signal sequence, an N-terminal S1 domain (codons 16–560), an S2 domain (codons 561–1173) containing several heptad repeats, a transmembrane domain (codons 1117–1138) and a cytoplasmic tail (codons 1139–1173) (Bonavia *et al.*, 2003). The 21 complete S gene sequences each contained 3513 nt (1170 aa). Deletion of 9 nt occurred in two discrete positions in each of these sequences, at 681–683 and 1057–1062. Without exception this resulted in 3 aa deletions compared to the prototype strain, aspartic acid at codon 228 and valine and tyrosine at codons 353 and 354, respectively. The partial sequences comprised nt 1–957 and 1806–3513 (patient 19) and 1–1788 (patient 25). The deletion pattern described above for full-length sequences also applied to these variants within the constraints of the sequence information available. Comparison of the full or partial S amino acid sequences with the prototype sequence identified 67 individual codon differences and the three deletions described above. The deletions and most dN changes occurred in the S1 domain, between aa 16 and 560. Thirty of 54 aa substitutions detected in this region were non-conservative compared with 7 of 13 substitutions identified outside the S1 domain (Table 2 and Supplementary Table S2 available in JGV Online). The putative *N*-glycosylation pattern (Asp–X–Ser/Thr) of the S protein of all variants differed slightly compared with the prototype. The 24 *N*-glycosylation sites in the prototype strain were also present in the variants, but two additional sites were generated in variants as a result of mutations to asparagine at codons 111 and 488. Binding of HCoV-229E to its cellular receptor human aminopeptidase N occurs through interaction with aa 417–517 of the S1 domain (Bonavia *et al.*, 2003). Seven amino acid substitutions occurred in this region (Table 2). Three of them (Q430K, D444N and K488N) were non-conservative and included the generation of an *N*-glycosylation site at aa 488.

Compared with the S protein of the HCoV-229E 1973 reference isolate, many substitutions and each of the deletions were present in both clinical isolates and directly sequenced specimens. This suggests that passage in cell culture has not substantially biased the analysis of the sequences studied. Additionally, we were able to compare the sequence of a virus isolate from 1999 (strain 13–18; GenBank accession no. AY386392) with variants we obtained from that time onwards. Of 11 aa changes between our pre-1999 isolates and directly sequenced variants, eight were also present in the 1999 isolate. Overall, this indicated that the influence of cell passage on the sequences studied was limited.

The entire sequence of the N gene was obtained for 23 of 25 HCoV-229E variants; compared with the S gene, there was less heterogeneity in this gene and the most divergent variant (from patient 22 in Table 1) differed by only 25 nt (9 aa) from the 1973 prototype strain (HCoV-229E 1973; GenBank accession no. DQ243939). Seven amino acid

Table 1. Details of the patients involved and specimens from which HCoV-229E was either isolated (patients 1–13) or detected by RT-PCR directly from supplied clinical material (patients 14–25)

Patient no.	Age (years)	Gender	Strain name	Specimen type*	Symptoms	GenBank accession no. S gene	GenBank accession no. N gene
1	NA†	NA	HCoV-229E-11/6/79	NA	NA	DQ243964	DQ243940
2	14	Female	HCoV-229E-16/6/82	NTS	URTI‡	DQ243965	DQ243941
3	5	Male	HCoV-229E-22/9/82	NTS	URTI	DQ243966	DQ243942
4	41	Female	HCoV-229E-6/10/82	NPA	URTI	DQ243967	DQ243943
5	28	Female	HCoV-229E-21/10/82	NTS	Immune deficiency, fever	DQ243968	DQ243944
6	9	Male	HCoV-229E-8/11/82a	NTS	Fever, URTI	DQ243969	DQ243945
7	4	Female	HCoV-229E-8/11/82b	NTS	URTI, blocked ears	DQ243970	DQ243946
8	35	Female	HCoV-229E-29/7/84	NTS	URTI	DQ243971	DQ243947
9	48	Male	HCoV-229E-5/9/84	NPA	Pneumonia	DQ243972	DQ243948
10	12	Female	HCoV-229E-24/6/90	NPA	Pertussis	DQ243973	DQ243949
11	29·5	Male	HCoV-229E-12/5/92	NPA	Viral, URTI	DQ243975	DQ243950
12	0·3	Female	HCoV-229E-17/6/92	NPA	Bronchiolitis	DQ243976	DQ243951
13	28	Female	HCoV-229E-25/6/92	NTS	Sore throat, fever, runny nose	DQ243974	DQ243952
14	34	Female	HCoV-229E-8/8/01	NTS	Fever, cough, myalgia	DQ243977	DQ243953
15	25	Male	HCoV-229E-27/8/01	NTS	Influenza-like	DQ243978	DQ243954
16	41	Male	HCoV-229E-21/6/02	NTS§	Influenza-like	DQ243979	–
17	23	Male	HCoV-229E-6/1/03	BAL	NA	DQ243980	DQ243955
18	24	Female	HCoV-229E-28/2/03	TS‖	NA	–	DQ243960
19	0·9	Female	HCoV-229E-24/4/03	NPA¶	SARS	DQ243981/2	DQ243056
20	17	Male	HCoV-229E-30/7/03	NTS	Influenza-like	DQ243983	DQ243957
21	18	Male	HCoV-229E-14/8/03	NTS#	Fever, cough, myalgia	DQ243984	DQ243959
22	44	Male	HCoV-229E-19/8/03	NTS	Fever, cough, malaise, myalgia	DQ243985	DQ243958
23	42	Female	HCoV-229E-25/8/03	NTS	Fever, cough, headache	DQ243986	DQ243961
24	26	Female	HCoV-229E-20/1/04	NS‖	NA	–	DQ243962
25	58	Female	HCoV-229E-27/1/04	Sputum§¶	Severe pneumonia	DQ243987	–

*NTS, Nose and throat swab; NPA, nasopharyngeal aspirate; BAL, bronchio-alveolar lavage; TS, throat swab; NS, nose swab.

†NA, Not available.

‡URTI, Upper respiratory tract infection.

§Only S gene sequences were obtained from patients 16 and 25.

‖Only N gene sequences were obtained from patients 18 and 24.

¶Partial S gene sequences were obtained from patients 19 and 25.

#Influenza virus also detected in this patient.

changes, four of which were non-conservative, were shared by each variant. Four changes clustered in a hot spot between aa 224 and 229 (Supplementary Table S3 available in JGV Online). Deletions were detected in only two N gene sequences. The first was a three base (GGT) in-frame deletion at nt position 480 in the virus from patient 3 resulting in the deletion of glycine at aa 160. The second was an A deletion from a string of five As starting at nt 611 in the variant from patient 14, producing a frameshift mutation and generation of a stop codon 52 aa downstream. This sequence was generated by direct RT-PCR from a clinical specimen and may be the result of amplification of non-infectious viral RNA.

Phylogenetic analysis of the S gene assigned the 23 variants to four distinct groups each containing temporally associated viruses (Fig. 1). The groups comprised viruses circulating during the periods 1979–1982, 1982–1984, 1990–1992 and 2001–2004. The expected transition/transversion ratio and gamma distribution parameter alpha were estimated as 1·62 and 0·16, respectively, from the dataset using TREEPUZZLE (Schmidt *et al.*, 2002). These values were used in DNAdist (Kimura) to produce a more accurate tree topology. A strong correlation between bootstrap values and these groupings was found. The expected transition/transversion ratio and gamma distribution parameter alpha values estimated using the N gene dataset were 1·81 and 0·19, respectively. Phylogenetic analysis of the N gene showed clustering of variants similar to that obtained with the S gene. However, less sequence variation in the N gene made it difficult to resolve the viruses into four groups. Rather, two major clusters comprising groups 1 and 2, and groups 3 and 4, respectively, were identified (Supplementary Fig. S1 available in JGV Online).

Table 2. Summary of amino acid changes in the S protein of selected HCoV-229E variants studied. The variants chosen represent the groups shown in Fig. 1

Strain name	Amino acid position																	
	22	26	35	51	88	90	104	109	111	112	113	121	140	159	181	196	210	223
HCoV-229E1973	L	Y	Y	D	L	F	D	S	D	V	L	L	V	F	A	S	T	T
HCoV-229E-11/6/79	–*	H	–	–	S	–	–	–	N	–	S	I	–	–	S	–	–	N
HCoV-229E-6/10/82	–	H	H	–	S	–	–	Y	N	A	S	I	–	–	–	–	–	N
HCoV-229E-8/11/82a	–	H	H	–	S	–	–	–	N	-	S	I	–	–	S	A	–	N
HCoV-229E-29/7/84	M	H	H	–	S	–	–	Y	N	A	S	I	–	–	–	–	–	N
HCoV-229E-24/6/90	M	H	H	–	S	–	–	Y	N	A	S	I	A	S	–	–	–	N
HCoV-229E-8/8/01	M	Q	H	N	S	L	–	Y	N	A	S	I	A	S	–	–	S	N
HCoV-229E-14/8/03	M	H	H	N	S	L	A	Y	N	A	S	I	A	S	–	–	S	N
Strain name	**225**	**228**	**229**	**248**	**307**	**309**	**310**	**311**	**312**	**314**	**316**	**318**	**321**	**324**	**342**	**349**	**350**	**352**
HCoV-229E1973.E	D	F	S	D	K	P	Q	S	G	K	F	Y	G	D	K	Y	A	
HCoV-229E-11/6/79	–	*†	V	A	–	–	–	–	–	–	–	Y	R	V	E	–	–	–
HCoV-229E-6/10/82	–	*	V	A	–	–	L	–	–	V	R	–	R	V	–	–	–	G
HCoV-229E-8/11/82a	–	*	V	A	–	–	–	–	–	–	R	Y	R	V	E	–	–	–
HCoV-229E-29/7/84	–	*	V	A	–	–	L	–	–	V	R	–	R	V	–	–	F	G
HCoV-229E-24/6/90	–	*	V	A	N	–	L	R	–	V	R	Y	R	V	–	Q	F	G
HCoV-229E-8/8/01	A	*	V	A	N	E	L	R	R	P	R	Y	R	V	–	Q	F	G
HCoV-229E-14/8/03	A	*	V	A	N	E	L	R	R	P	R	Y	R	V	–	Q	F	G
Strain name	**353**	**354**	**355**	**356**	**357**	**358**	**401**	**404**	**406**	**407**	**408**	**409**	**410**	**411**	**424**	**430**	**444**	**462**
HCoV-229E1973.V	Y	A	N	V	G	V	W	Y	S	K	Y	Y	T	G	Q	D	V	
HCoV-229E-11/6/79	*	*	–	–	–	–	–	L	–	I	N	S	–	–	–	K	N	I
HCoV-229E-6/10/82	*	*	–	–	F	–	–	L	–	L	N	S	–	–	V	K	N	I
HCoV-229E-8/11/82a	*	*	–	–	–	–	–	L	–	I	N	S	–	–	–	K	N	I
HCoV-229E-29/7/84	*	*	–	–	F	–	–	L	–	L	N	S	–	–	V	K	N	I
HCoV-229E-24/6/90	*	*	–	–	F	–	M	L	N	L	N	S	H	–	V	K	N	I
HCoV-229E-8/8/01	*	*	V	K	F	D	M	L	N	L	N	S	H	N	V	K	N	I
HCoV-229E-14/8/03	*	*	V	K	F	D	M	L	N	L	N	S	H	N	V	K	N	I
Strain name	**465**	**488**	**530**	**642**	**669**	**676**	**681**	**714**	**765**	**775**	**849**	**871**	**937**	**971**	**1002**	**1005**		
HCoV-229E1973	D	K	L	R	D	I	T	N	V	A	S	T	I	T	I	M		
HCoV-229E-11/6/79	–	N	–	M	–	–	–	K	A	–	–	I	L	R	–	–		
HCoV-229E-6/10/82	–	N	M	M	–	–	–	K	A	S	–	I	–	R	–	–		
HCoV-229E-8/11/82a	–	N	–	M	E	V	–	K	A	–	–	I	L	R	V	–		
HCoV-229E-29/7/84	–	N	M	M	–	–	–	K	A	S	–	I	L	R	–	–		
HCoV-229E-24/6/90	–	N	M	M	–	–	R	K	A	S	–	I	L	R	–	I		
HCoV-229E-8/8/01	–	N	M	M	–	–	R	K	A	S	F	I	L	R	–	I		
HCoV-229E-14/8/03	E	N	M	M	–	–	R	K	A	S	–	I	L	R	–	I		

*–, Denotes same amino acid as first sequence.

†*, Denotes amino acid deletion.

Within the clinical variants there was substantial evidence for temporal mutations. Many of these occurred in 1982 and persisted, while others, for example D51N, F90L and T210S, first occurred in 2001 and were present subsequently in all variants. Some mutations (A181S, S196A, D342E, S407I, D669E, I676V and I1002V) were selected only briefly and then replaced by variants possessing the codon existing before that time. At the time of generation of a new codon it was often possible to detect the prevailing variant co-circulating with the strain containing the new mutation; for example, L22 with M22 mutants in 1984 and D104 with A104 mutants in 2003 (Table 2 and Supplementary Table S2 available in JGV Online).

Alignment of S protein sequences and phylogenetic analysis of the S gene revealed several discrete time points where amino acid substitutions occurred and were then retained during subsequent years. These mutations may represent

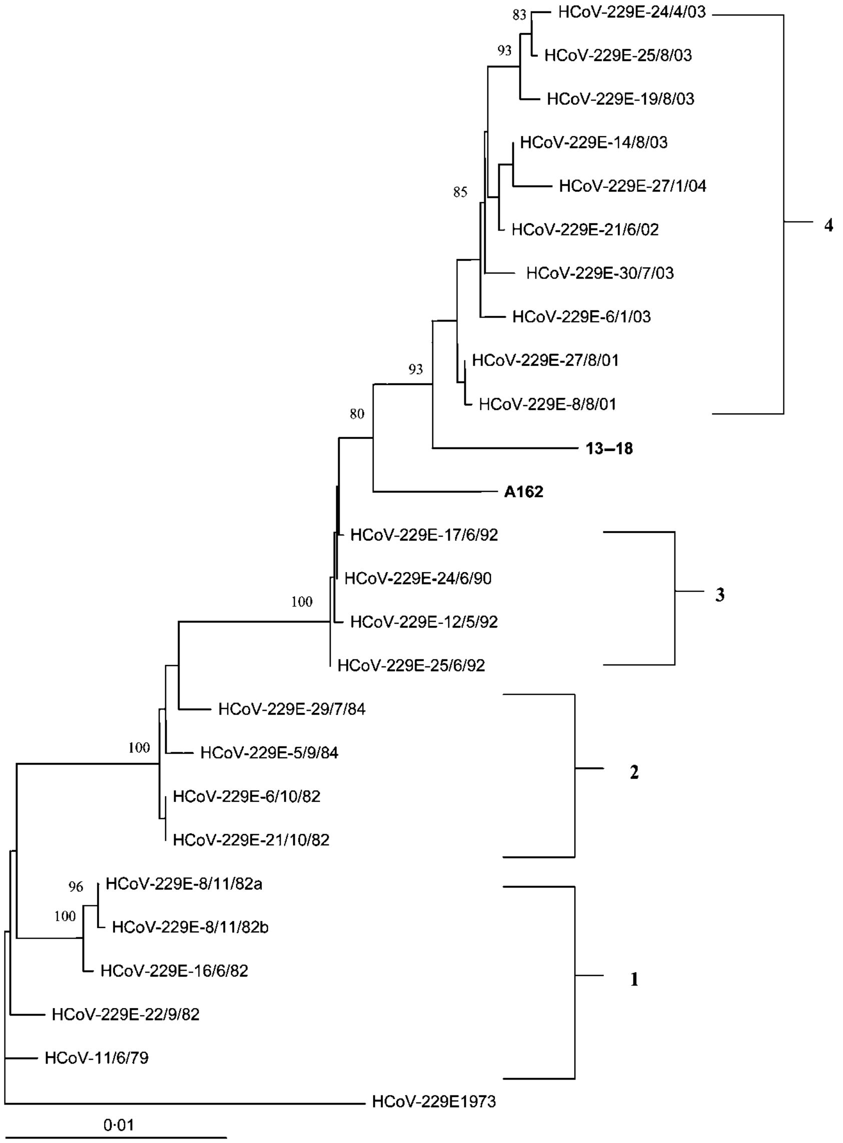

Fig. 1. Phylogenetic analysis using PHYLIP (DNAdist/neighbour-joining, 1000 bootstrap replicates) on HCoV-229E isolates and strains identified between 1979 and 2004. The DNA sequence of the entire coding region of the HCoV-229E S gene was analysed. Sequences in bold were obtained through GenBank and are strains detected in geographical locations other than Australia. Significant bootstrap values (>80 %) and the clustering of viruses analysed into groups (1–4) are indicated.

the outcome of natural evolution of the virus in the context of antibody pressure. Comparing rates of dS and dN nucleotide substitutions can test three forms of evolutionary selection on a gene. Equal dS and dN substitution rates demonstrate an absence of selection, whereas an excess of dS substitutions indicates purifying selection, operating to preserve protein structure and function. Excess dN substitution rates indicate positive selection and occur mainly in genes with adaptive functions (Bush, 2001). Most positively selected genes previously identified have been pathogen surface proteins (Yang & Bielawski, 2000). To avoid recognition by antibodies generated in response to prior antigen exposure, surface proteins of pathogens must change their structure. Hence, it is likely that evasion of the host immune system drives repeated amino acid replacements in surface proteins (Bush, 2001). Using the Nei–Gojobori method for estimating statistical differences between dS and dN substitution rates (Kumar *et al.*, 2004), we confirmed that the HCoV-229E S gene sequences had undergone positive selection over time. Probabilities of 0·00 were obtained for purifying and neutral selection and both hypotheses were rejected. A probability of 1·0 was obtained when testing for positive selection and this hypothesis was therefore accepted.

We could not show evidence for the emergence of distinct subtypes of HCoV-229E as recently reported for HCoV-OC43, where strains circulating in Belgium between 2003 and 2004 showed sufficient unrelatedness at the amino acid level (3·1 %) to suggest the existence of two genetically distinct strains (Vijgen *et al.*, 2005). For the HCoV-229E variants we studied, the overall amino acid difference between the 1979 strain and the 2004 strain was 3·3 %. When apparent antigenic drift occurred, the extent of amino acid difference between the old variant and its replacement was less than 1 %.

When the HCoV-229E S gene sequences from this study were compared to 26 sequences available through GenBank, only two variants were similar. The first, strain A162 (accession no. Y10051), was isolated in 1995 from an adult in Ghana (Hays & Myint, 1998). While the submitted sequence was 483 nt short of full-length, it grouped phylogenetically with coronaviruses circulating in Australia between 1990 and 1992 (Fig. 1). Also similar was strain 13–18, first isolated in 1999 in the USA. Although the sequence available was less than half that of the complete S gene, this virus grouped with variants circulating between 2001 and 2004 in Australia (Fig. 1). Hence, the time of circulation of distinct HCoV-229E variants in south-eastern Australia generally coincided with that of variants circulating elsewhere.

Overall, sequencing and analysis of the S and N gene products of HCoV-229E strains circulating in Victoria, Australia, between 1979 and 2004 has provided the first evidence for genetic drift and positive selection as part of the evolution of this virus. The similarity between those strains circulating in Victoria and a small number of strains identified in other geographical locations at similar times indicates that HCoV-229E, despite having the potential, has not undergone major recombination events since it was first isolated in 1967.

References

Birch, C. J., Clothier, H. J., Seccull, A., Tran, T., Catton, M. C., Lambert, S. B. & Druce, J. D. (2005). Human coronavirus OC43 causes influenza-like illness in residents and staff of aged-care facilities in Melbourne, Australia. *Epidemiol Infect* **133**, 273–277.

Bonavia, A., Zelus, B. D., Wentworth, D. E., Talbot, P. J. & Holmes, K. V. (2003). Identification of a receptor-binding domain of the spike glycoprotein of human coronavirus HCoV-229E. *J Virol* **77**, 2530–2538.

Brian, D. A. & Baric, R. S. (2005). Coronavirus genome structure and replication. *Curr Top Microbiol Immunol* **287**, 1–30.

Bush, R. M. (2001). Predicting adaptive evolution. *Nat Rev Genet* **2**, 387–392.

Cavanagh, D., Davis, P. J. & Mockett, A. P. (1988). Amino acids within hypervariable region 1 of avian coronavirus IBV (Massachusetts serotype) spike glycoprotein are associated with neutralization epitopes. *Virus Res* **11**, 141–150.

Clothier, H. J., Fielding, J. E. & Kelly, H. A. (2005). An evaluation of the Australian Sentinel Practice Research Network (ASPREN) surveillance for influenza-like illness. *Commun Dis Intell* **29**, 231–247.

Corpet, F. (1988). Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res* **16**, 10881–10890.

Drosten, C., Gunther, S., Preiser, W. & 23 other authors (2003). Identification of a novel coronavirus in patients with severe acute respiratory syndrome. *N Engl J Med* **348**, 1967–1976.

Felsenstein, J. (1993). PHYLIP: phylogeny inference package (version 3.5c). Department of Genetics, University of Washington, Seattle, WA, USA.

Hall, T. A. (1999). BIOEDIT: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp Ser* **41**, 95–98.

Hays, J. P. & Myint, S. H. (1998). PCR sequencing of the spike genes of geographically and chronologically distinct human coronaviruses 229E. *J Virol Methods* **75**, 179–193.

Holmes, K. V. (2001). Coronaviruses. In *Fields Virology*, 3rd edn, pp. 1187–1203. Edited by B. N. Fields, D. M. Knipe & P. M. Howley. Philadelphia: Lippincott Williams & Wilkins.

Holmes, K. V. & Lai, M. M. C. (1996). *Coronaviridae*: the viruses and their replication. In *Fields Virology*, 3rd edn, vol. 1, pp. 1075–1093. Edited by B. N. Fields, D. M. Knipe & P. M. Howley. Philadelphia: Lippincott Raven.

Kumar, S., Tamura, K. & Nei, M. (2004). MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief Bioinform* **5**, 150–163.

Lai, M. M. (1990). Coronavirus: organization, replication and expression of genome. *Annu Rev Microbiol* **44**, 303–333.

McIntosh, K. (1996). Coronaviruses. In *Fields Virology*, 3rd edn, vol. 1, pp. 1095–1103. Edited by B. N. Fields, D. M. Knipe & P. M. Howley. Philadelphia: Lippincott Raven.

Navas-Martin, S. & Weiss, S. R. (2003). SARS: lessons learned from other coronaviruses. *Viral Immunol* **16**, 461–474.

Nicholas, K. B. & Nicholas, H. B., Jr (1997). Genedoc: a tool for editing and annotating multiple sequence alignments. Distributed by author, 2·6·002 edn. http://www.psc.edu/biomed/genedoc

Page, R. D. M. (1996). TREEVIEW: an application to display phylogenetic trees on personal computers. *Comput Appl Biosci* **12**, 357–358.

Pene, F., Merlat, A., Vabret, A., Rozenberg, F., Buzyn, A., Dreyfus, F., Cariou, A., Freymuth, F. & Lebon, P. (2003). Coronavirus 229E-related pneumonia in immunocompromised patients. *Clin Infect Dis* **37**, 929–932.

Rest, J. S. & Mindell, D. P. (2003). SARS associated coronavirus has a recombinant polymerase and coronaviruses have a history of host-shifting. *Infect Genet Evol* **3**, 219–225.

Schmidt, H. A., Strimmer, K., Vingron, M. & von Haeseler, A. (2002). TREEPUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18**, 502–504.

Siddell, S. G. (1995). The *Coronaviridae*: an introduction. In The *Coronaviridae*. Edited by S. G. Siddell. New York: Plenum Press.

Snijder, E. J., Bredenbeek, P. J., Dobbe, J. C. & 7 other authors (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.

Vabret, A., Mourez, T., Gouarin, S., Petitjean, J. & Freymuth, F. (2003). An outbreak of coronavirus OC43 respiratory infection in Normandy, France. *Clin Infect Dis* **36**, 985–989.

van der Hoek, L., Pyrc, K., Jebbink, M. F. & 7 other authors (2004). Identification of a new human coronavirus. *Nat Med* **10**, 368–373.

Vijgen, L., Keyaerts, E., Lemey, P., Moes, E., Li, S., Vandamme, A. M. & Van Ranst, M. (2005). Circulation of genetically distinct contemporary human coronavirus OC43 strains. *Virology* **337**, 85–92.

Woo, P. C., Lau, S. K., Chu, C. M. & 12 other authors (2005). Characterization and complete genome sequence of a novel coronavirus, coronavirus HKU1, from patients with pneumonia. *J Virol* **79**, 884–895.

Yang, Z. & Bielawski, J. P. (2000). Statistical methods for detecting molecular adaptation. *Trends Ecol Evol* **15**, 496–503.