

ORIGINAL ARTICLE

Identification and evolutionary dynamics of two novel human coronavirus OC43 genotypes associated with acute respiratory infections: phylogenetic, spatiotemporal and transmission network analyses

Xiang Yong Oong^{1,*}, Kim Tien Ng^{2,*}, Yutaka Takebe^{1,3}, Liang Jie Ng⁴, Kok Gan Chan⁵, Jack Bee Chook², Adeeba Kamarulzaman² and Kok Keng Tee¹

Human coronavirus OC43 (HCoV-OC43) is commonly associated with respiratory tract infections in humans, with five genetically distinct genotypes (A to E) described so far. In this study, we obtained the full-length genomes of HCoV-OC43 strains from two previously unrecognized lineages identified among patients presenting with severe upper respiratory tract symptoms in a cross-sectional molecular surveillance study in Kuala Lumpur, Malaysia, between 2012 and 2013. Phylogenetic, recombination and comparative genomic analyses revealed two distinct clusters diverging from a genotype D-like common ancestor through recombination with a putative genotype A-like lineage in the non-structural protein (nsp) 10 gene. Signature amino acid substitutions and a glycine residue insertion at the N-terminal domain of the S1 subunit of the spike gene, among others, exhibited further distinction in a recombination pattern, to which these clusters were classified as genotypes F and G. The phylogeographic mapping of the global spike gene indicated that the genetically similar HCoV-OC43 genotypes F and G strains were potentially circulating in China, Japan, Thailand and Europe as early as the late 2000s. The transmission network construction based on the TN93 pairwise genetic distance revealed the emergence and persistence of multiple sub-epidemic clusters of the highly prevalent genotype D and its descendant genotypes F and G, which contributed to the spread of HCoV-OC43 in the region. Finally, a more consistent nomenclature system for non-recombinant and recombinant HCoV-OC43 lineages is proposed, taking into account genetic recombination as an important feature in HCoV evolution and classification.

Emerging Microbes & Infections (2017) 6, e3; doi:10.1038/emi.2016.132; published online 4 January 2017

Keywords: comparative genomic analyses; evolutionary dynamics; human coronavirus OC43; recombination; transmission network

INTRODUCTION

Human coronavirus OC43 (HCoV-OC43), belonging to the *Betacoronavirus* genus of the Coronaviridae family,¹ continues to cause respiratory tract infections in children and adult populations worldwide.^{2,3} HCoV-OC43 and other human coronaviruses (HKU1, NL63, 229E, SARS-CoV and MERs-COV) contain a large positive-sense single-stranded RNA with a genome size from ~27 to 31 kb.⁴ Previous studies have focused on investigating the molecular epidemiology of HCoV-OC43 to understand its evolution and pathogenicity.^{5–11} HCoVs continue to evolve through homologous RNA recombination and exhibit high nucleotide substitution rates across the genome,^{12,13} resulting in the emergence of novel variants that can adapt to new hosts or ecological niches.^{14–18}

Since the first description of HCoV-OC43 in the 1960s, five genetically distinct genotypes (A through E) have been identified

based on phylogenetic analysis of main genes, such as the spike (S), RNA-dependent RNA polymerase (RdRP) and nucleocapsid (N) genes and complete viral genome.^{7,9} Genotypes A and B were estimated to have emerged around the 1950s and 1990s, respectively, whereas genotypes C, D and E were detected more recently in the 2000s.^{7,9} Genotype D arose from recombination between genotypes B and C and was dominant in parts of Asia and Europe.^{7–9,19} Likewise, genotype E was generated from recombination among genotypes B, C and D in Asia,⁹ underlining the importance of recombination in driving the evolution of HCoV-OC43.

A cross-sectional molecular surveillance of HCoV-OC43 and HCoV-HKU1 was conducted among patients presented with acute upper respiratory tract infection (URTI) in Kuala Lumpur, Malaysia.²⁰ Both HCoV-OC43 and HCoV-HKU1 were co-circulating throughout the year, but the lowest detection rates were reported between October

¹Department of Medical Microbiology, Faculty of Medicine, University of Malaya, 50603 Kuala Lumpur, Malaysia; ²Department of Medicine, Faculty of Medicine, University of Malaya, 50603 Kuala Lumpur, Malaysia; ³AIDS Research Center, National Institute of Infectious Diseases, 162-8640 Tokyo, Japan; ⁴Faculty of Information Science & Technology, Multimedia University, 75450 Melaka, Malaysia and ⁵Division of Genetics and Molecular Biology, Institute of Biological Sciences, Faculty of Science, University of Malaya, 50603 Kuala Lumpur, Malaysia

*These authors contributed equally to this work.

Correspondence: KK Tee

E-mail: k2tee@um.edu.my

Received 9 September 2016; revised 18 November 2016; accepted 21 November 2016

and January,²⁰ a period that coincides with the Northeast Monsoon season (November to March), which brings in more rainfall compared with the Southwest Monsoon.²¹ Interestingly, phylogenetic analysis of the partial S gene (S1 domain) revealed that a majority of the HCoV-OC43 strains shared a genotype D-like common ancestor but diverged into two unique clusters. In this study, we obtained the full-length genome sequences of these unique strains and performed phylogenetic and recombination analyses, suggesting a possible emergence of two novel recombinant genotypes descended from genotype D, which were designated as genotypes F and G. Through a database search of global S gene sequences, Bayesian coalescent phylogenetic and amino acid sequence analyses implied that these two novel genotypes were likely to have emerged around the late 2000s to early 2010s with a wide geographical dispersion. Their origins were probably mapped to Asia where the putative parent genotype D was circulating at high prevalence, driven in part by the emergence and persistence of multiple sub-epidemic transmission networks of respiratory tract infections.

MATERIALS AND METHODS

Clinical specimens

This study was approved by the University of Malaya Medical Centre (UMMC) Medical Ethics Committee (MEC890.1). Standard, multi-lingual consent forms from the Medical Ethics Committee were used, and written consent was obtained from all study participants. A total of 2060 consenting outpatients presented with symptoms of acute URTI were recruited at the primary care clinics of University Malaya Medical Centre in Kuala Lumpur, Malaysia between March 2012 and February 2013. The nasopharyngeal swabs collected from the patients were transferred to the laboratory in universal transport media (Copan Diagnostics, Inc., Murrieta, CA, USA) and stored at -80°C . The xTAG Respiratory Virus Panel (RVP) FAST multiplex RT-PCR assay (Luminex Molecular, Toronto, ON, Canada) and Luminex's proprietary Universal Tag sorting system on Luminex 200 IS platform (Luminex, Austin, TX, USA) were used to detect HCoV-OC43 in the samples according to the manufacturer's protocol.²² As reported previously, through phylogenetic analysis of the partial S gene (S1 domain), 21 out of 2060 nasopharyngeal samples (1.02%), which were positive for HCoV-OC43, formed two distinct clades

provisionally designated as lineages 1 and 2 that shared a genotype D-like common ancestor.²⁰

Full-length genome sequencing

To characterize and evaluate the novelty of the two distinct HCoV-OC43 lineages, 16/21 strains (nine from lineage 1 and seven from lineage 2) from 16 infected patients were prepared for further whole-genome analysis. The demographic and clinical profile of patients infected with HCoV-OC43 lineages 1 and 2 are summarized in Table 1. To obtain the full-length genome of these unique strains, viral RNA was extracted by the NucliSENS easyMAG automated nucleic acid extraction system (bioMérieux, Marcy l'Etoile, France)²³ and reversely transcribed into cDNA using SensiFAST cDNA Synthesis Kit (Bioline, London, UK), which contains anchored oligo(dT) and random hexamer primers. The full-length genome cDNA of ~ 30 kb in size (which flanks from the 5' end of *ORF1a* gene to 3' end of the poly-A tail) was amplified by a genome walking method that involved a total of 44 overlapping fragments using a set of previously published primers with minor modifications for improved sequence coverage (Supplementary Table S1).⁹ PCR thermocycling conditions were set as follows: initial denaturation at 95°C for 1 min, 35 cycles of amplification at 95°C for 15 s, 50°C for 15 s and 72°C for 30 s using the MyTaq HS Red Mix (Bioline, London, UK) kit. PCR products were purified, and sequencing reactions were performed in ABI PRISM 3730XL Genetic Analyzer using the BigDye Terminator v3.1 cycle sequencing kit chemistry (Applied Biosystems, Foster City, CA, USA). Finally, sequence reads were assembled into a contig and manually edited using BioEdit 7.2 (Ibis Therapeutics, Carlsbad, CA, USA) to produce a final sequence of full-length HCoV-OC43 genomes. All sequences generated in this study are available from GenBank under accession numbers KX538964–KX538979.

Phylogenetic, recombination and amino acid sequence analyses

To determine the evolutionary relationship among the unique and global HCoV-OC43 strains, phylogenetic analysis was conducted using full-length genome sequences. All 16 unique sequences were first aligned with published global reference sequences (genotypes A to E) retrieved from GenBank (accessed on 31 March, 2016) (Supplementary

Table 1 Demographic and clinical profile of patients infected with HCoV-OC43 lineages 1 and 2

Lineage	Strain ID	Collection date	Demographic profile		Symptoms reported
			Age	Sex	
Lineage 1	MY-U208/12	28 March 2012	61	F	Sneezing, nasal discharge, sore throat, hoarseness of voice, cough
	MY-U413/12	2 May 2012	72	F	Nasal discharge
	MY-U464/12	9 May 2012	38	M	Nasal congestion, headache, sore throat, hoarseness of voice, cough
	MY-U523/12	18 May 2012	74	M	Sneezing, cough
	MY-U732/12	25 June 2012	53	F	Sneezing, nasal discharge, headache, cough
	MY-U868/12	16 July 2012	59	M	Nasal congestion, cough
	MY-U945/12	1 August 2012	11	M	Sneezing, nasal discharge, nasal congestion, headache
	MY-U1024/12	24 August 2012	61	F	Nasal congestion
	MY-U1140/12	10 September 2012	21	M	Sneezing, nasal discharge, sore throat, hoarseness of voice
Lineage 2	MY-U002/12	22 February 2012	71	F	Nasal congestion, headache, cough
	MY-U236/12	2 April 2012	19	M	Nasal discharge, nasal congestion, headache
	MY-U710/12	20 June 2012	50	F	Sneezing, nasal discharge, nasal congestion, cough
	MY-U774/12	2 July 2012	32	F	Sneezing, nasal congestion, headache, sore throat, hoarseness of voice, cough
	MY-U1057/12	27 August 2012	58	F	Sore throat, hoarseness of voice, cough
	MY-U1758/13	2 January 2013	56	M	Sneezing, nasal discharge, nasal congestion, headache, sore throat, cough
	MY-U1975/13	15 February 2013	52	F	Sneezing, headache, hoarseness of voice, cough

Abbreviations: Female, F; Male, M.

Table S2) using a web-based multiple sequence alignment program MAFFT.²⁴ Phylogenetic tree reconstruction using the neighbor-joining (NJ) method and inter-genotype pairwise genetic distance calculation for sequence divergence comparison were performed using MEGA 6.0.²⁵ The maximum-likelihood (ML) method was also performed for reconstruction of a phylogenetic tree, which was heuristically inferred using subtree pruning and regrafting and nearest neighbor interchange algorithms with a general time-reversible (GTR) nucleotide substitution model, a proportion of invariant sites (+I) and four categories of gamma rate heterogeneity (+ Γ_4), which were implemented in PAUP version 4.0.²⁶ Kimura's two-parameter model with a reliability of branching order analyzed by bootstrap replicates of 1000 was used. Subsequently, bootscanning was performed using SimPlot version 3.5.1 to determine possible recombination events and location of breakpoints in the viral genome of unique strains. This approach has been previously reported.^{7,9,15,27} Sub-genomic regions located between recombination breakpoints were subjected to additional phylogenetic analysis using the neighbor-joining method to infer the recombination structure and the parental genotype of each region. Signature nucleotide and amino acid substitutions of the unique strains were determined by Sequence Data Explorer in MEGA.

Estimation of divergence times

The Bayesian Evolutionary Analysis by Sampling Trees (BEAST) program has been widely used to investigate the spatiotemporal and evolutionary dynamics of viral pathogens using time-stamped nucleotide sequence data sets.²⁸ Previously, estimations of divergence times of HCoV-OC43 strains relied mainly on the S gene sequence data^{9–11} given that the S protein is the major antigenic protein with high selection pressure and genetic diversity compared with other viral proteins.⁴ In this study, using the query ($n = 16$) and global reference full-length genome sequences ($n = 13$), the divergence times of all HCoV-OC43 genotypes and lineages 1 and 2²⁰ were estimated to determine when these strains emerged. The divergence times were also re-estimated using all S gene sequences available in the public database (S1 domain: 23 644–25 125 nt). The estimation was performed by molecular clock dating analysis using the Bayesian Markov chain Monte Carlo (MCMC) coalescence method implemented in BEAST 1.7.²⁸ Two parametric demographic models (constant and exponential population sizes) and one non-parametric model (Bayesian Skyline Plot (BSP)) coalescent tree priors were used to infer the viral phylogenies, nucleotide substitution rates and time of most recent common ancestor (tMRCA). The uncorrelated exponential relaxed, uncorrelated lognormal relaxed and strict molecular clock models were tested. Analyses were performed under the general time-reversible nucleotide substitution model with a proportion of invariant sites (GTR+I). MCMC runs for the full-length genome and S gene were 50 million steps long, with sampling every 50 000 states. Using Tracer version 1.6 (<http://tree.bio.ed.ac.uk/software/tracer/>), the output was assessed for convergence by means of effective sampling size greater than 200 after a 10% burn-in. Bayesian maximum clade credibility (MCC) trees were annotated using the Tree Annotator program included in the BEAST package by choosing the tree with the maximum sum of posterior probabilities after a 10% burn-in. The final MCC trees were visualized in FigTree (<http://tree.bio.ed.ac.uk/software/figtree/>).

Transmission network analysis of HCoV-OC43 genotype D and its related recombinants

As HCoV-OC43 genotype D has been the most prevalent and persistent genotype circulating in East Asia in recent years,^{7,9,11} an

estimation of the transmission network of genotype D and its related recombinants²⁰ could be a useful strategy to elucidate the degree of spread and dynamics of infection attributed to these genotypes within and between countries.^{29,30} To deduce the transmission pattern of HCoV-OC43 genotype D and its related recombinants in recent years, a transmission cluster was deduced from new and published S gene sequences based on the Tamura-Nei 93 (TN93) pairwise distance estimates performed using a custom script in Python (release 3.2.6) with a bootstrap analysis of 1000 replicates.^{29,30} In the present study, a transmission cluster is defined as a cluster consisting at least two individuals (nodes) whose viral sequences are genetically linked (edges) at a given genetic distance threshold supported by bootstrap value of >90%.²⁹ The genetic distance threshold was determined between the highest and lowest values of the intra- and inter-patient patristic distances, respectively, measured in nucleotide substitutions per site.^{31,32} Given that HCoV-OC43 causes acute respiratory tract infection and hinders the estimation of intra-patient viral genetic distance, the most probable threshold value was determined from the 95% confidence interval of the lower 0.025 percentile of the inter-patient genetic distances²⁹ as calculated from globally available and published S gene reference sequences ($n = 27$)

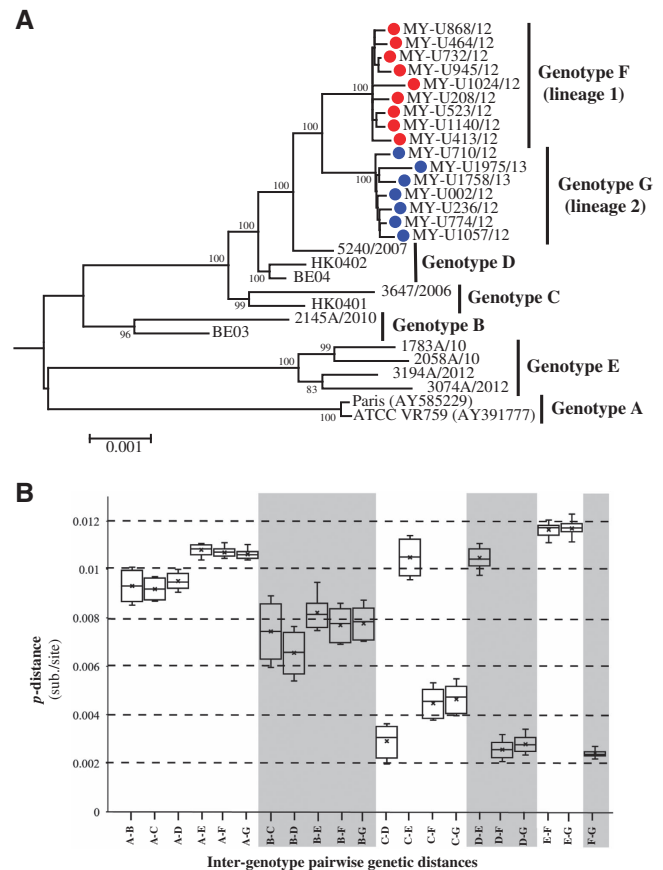


Figure 1 (A) Phylogenetic analysis of the HCoV-OC43 strains based on the full-length genome. Trees were reconstructed using the neighbor-joining method and Kimura 2-parameter model in MEGA 6.0. Bootstrap values were calculated from 1000 trees. Bootstrap values > 70% were indicated on the branch nodes. The scale bar of an individual tree indicates the substitutions per site. (B) Estimation of pairwise genetic distances between HCoV-OC43 genotypes based on the full-length genome sequences. Genotypes F and G were previously classified as lineages 1 and 2, respectively.²⁰

(Supplementary Tables S3 and S4). HCoV-OC43 sequences from different patients with a patristic distance less than the estimated threshold were identified either as transmission dyads (consists of two nodes) or networks (more than two nodes),³³ reflecting the transmission linkages and genetic relatedness of the infecting HCoV-OC43 strains.

RESULTS

Phylogenetic analysis of unique HCoV-OC43 strains using full-length genome sequences

The phylogenetic tree reconstructed by the NJ method for the full-length genome is illustrated in Figure 1A, which consists of the unique Malaysian HCoV-OC43 strains and all available global reference sequences from 2001 to 2013 (except for prototype strain ATCC

VR759, which was isolated in the 1960s) (Supplementary Table S2). These reference sequences were classified previously as genotypes A to E, and these reference viruses were isolated from patients with acute respiratory tract infection (ARTI) in Paris,³⁴ Belgium,⁵ China⁹ and Hong Kong.⁷ Phylogenetic trees were also reconstructed by the NJ and ML methods for full-length genome, which include genotyped (published, $n = 13$) and un-genotyped (unpublished, $n = 76$) reference sequences as well as 2 cell-adapted/neurovirulent strain sequences, as shown in Supplementary Figure S1. Two distinct clusters (lineages 1 and 2) appeared to branch out from a genotype D-like common ancestor with high bootstrap support (100%) (Figure 1A; Supplementary Figure S1). The topology of this phylogenetic tree based on a full-length genome was similar to the tree topology based on a partial S gene reported in a previous study.²⁰ On the basis of the

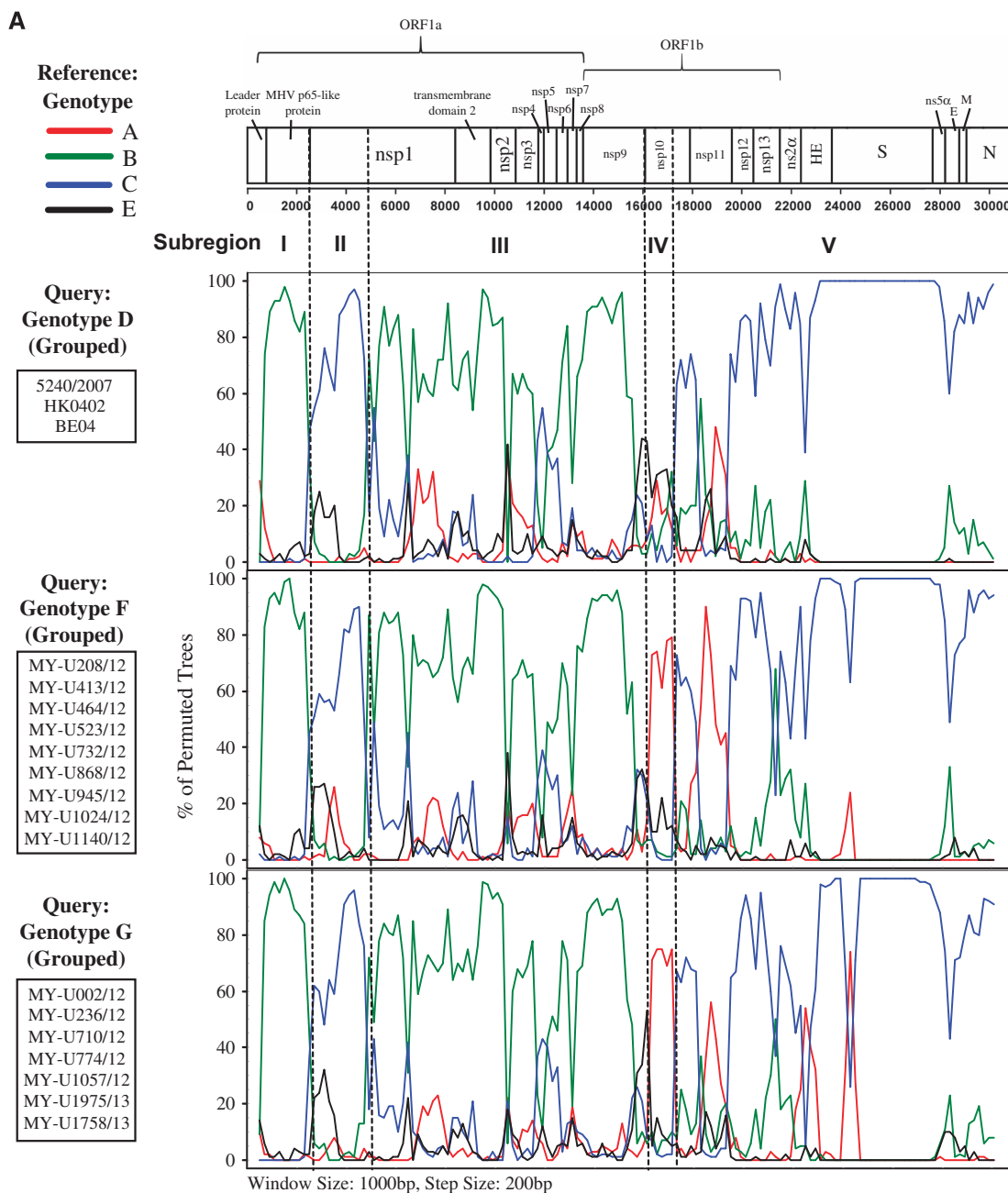


Figure 2 Continued.

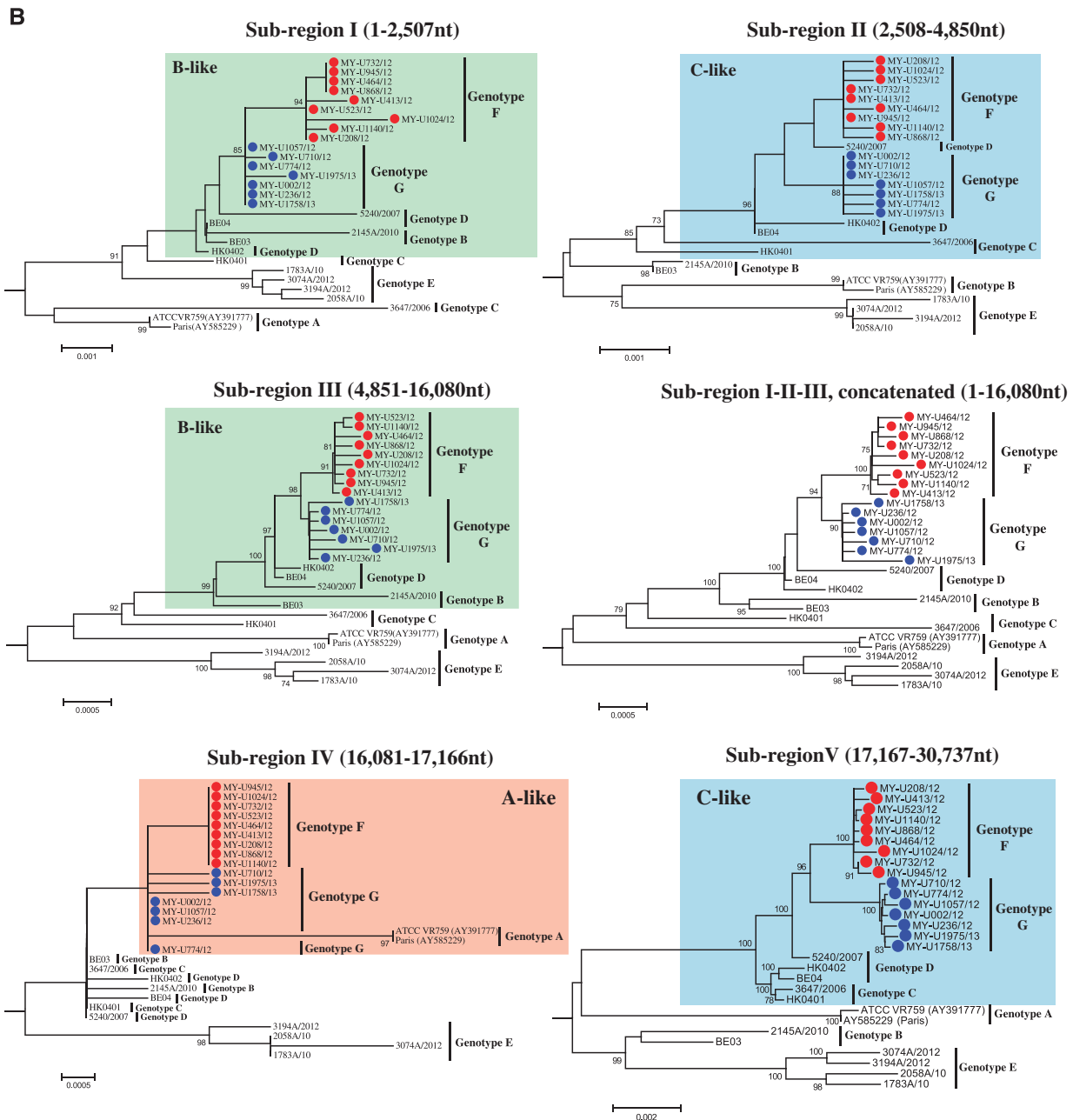


Figure 2 (A) Comparison of the mosaic recombination structure of the full-length genome between genotypes D, F and G. Bootscan analysis was performed using published and genotyped reference genomes for genotypes A, B, C and E as putative parental genotypes. (B) Putative parental genotype determination and confirmation in sub-genomic regions (sub-regions I–V) of genotypes D, F and F recombinants using the neighbor-joining method. Breakpoints determined by informative site analysis. Red, green and blue shades indicate genotypes A, B and C putative parental genotypes, respectively. Numbering of nucleotide (nt) positions is based on prototype ATCC VR759 reference strain.

estimation of inter-genotype pairwise genetic distances (Figure 1B), the distances of lineages 1 and 2 compared with genotypes A, B and E were $> 0.7\%$ (0.007 substitutions/site), whereas distances were $< 0.5\%$ when compared with genotypes C and D. This finding indicates that lineages 1 and 2 were more similar to genotypes C and D compared with genotypes A, B and E. Although the low mean genetic distance of $0.29 \pm 0.03\%$ between genotypes C and D is probably attributed to the recombinant nature of genotype D (genotype D was generated from recombination between genotypes B and C),⁷ it is the lowest reported full-length genetic distance that separates HCoV-OC43 genotypes.

Using this benchmark, with a genetic distance of $0.26 \pm 0.02\%$ between genotype D and lineage 1 and $0.27 \pm 0.02\%$ between genotype D and lineage 2, it is suggestive that the two lineages, which are designated as genotypes F and G hereafter, may have arisen and diverged from genotype D.

Mosaic recombination structures of HCoV-OC43 genotypes F and G

The mosaic recombination structures of genotypes F and G were determined and compared with genotype D by performing bootscan

analysis (sliding window size: 1000 bp, step size: 200 bp).^{7,9} Published reference full-length genomes for genotypes A (prototype strain ATCC VR759, Paris strain—AY585229), B (Belgium 2003, 2145/2010), C (HK0401, 3647/2006) and E (1783A/10, 2058A/10, 3194A/12, 3074A/12) were used as putative parental genotypes. When the genomes of grouped genotypes F and G strains were used as query sequences, several potential recombination sites in the viral genomes were observed, separating the genome into at least five sub-regions (Figure 2A). From the 5' end of the genome to position 16 080 nt, bootscan analysis showed that sub-regions I (positions 1 to 2507 nt) and III (4851 to 16 080 nt) of genotypes F and G were closely related to genotype B, whereas sub-region II (2508 to 4850 nt) was closely related to genotype C (supported by sub-region NJ trees in Figure 2B). Bootscan and sub-region tree analyses also revealed that these regions (sub-regions I–II–III) shared high homology with genotype D. From positions 16 081 to 17 166 nt (sub-region IV), genotypes F and G were closely related to genotype A, whereas positions 17 167 to 30 737 nt (sub-region V) were grouped with genotypes C and D (Figures 2A and 2B). When the mosaic recombination structures of genotypes F and G were compared with genotype D, it is noticeable that all three genotypes shared similar recombination breakpoints between 2500–3000 nt, 4500–5000 nt and 16 000–17 000 nt. However, both genotypes F and G had an additional recombination breakpoint between 17 000–17 500 nt, which was not observed in genotype D; thus, parts of the nsp10 gene (sub-region IV, 16 081–17 166 nt) were genotyped as A-like. This finding indicates that recombination events led to the emergence of novel genotypes F and G with a putative genotype A-like parental strain in the nsp10 region despite sharing similar recombination structure in most parts of the genome with genotype D strains.

Nucleotide and amino acid sequence analysis

The whole genome of genotypes F and G strains was further subjected to nucleotide and amino acid sequence analysis to detect signature substitutions in their respective genomes given that the mosaic recombination pattern between genotypes F and G could not be clearly distinguished. However, both were evidently distinct from genotype D (Figure 2). As observed in Figure 3, using the prototype ATCC VR759 as the reference strain for nucleotide and amino acid positions, 34 and 32 nucleotide substitutions unique to genotypes F and G, respectively, were mapped across the whole genome. Corresponding to these nucleotide substitutions, 15 and 10 non-synonymous amino acid substitutions were observed in genotypes F and G, respectively. Of note, all genotype F strains had a unique 3-nucleotide insertion (GGC) between 23 988 and 23 989 nt that was not observed in genotype G as well as other genotypes, resulting in a glycine insertion at position 119 in the S protein. Likewise, genotype A-like nucleotide substitutions were observed in the S gene at positions 23 707, 24 186, 24 430 and 24 434 nt in genotype G (but not in genotype F) (Figure 3), indicating a plausible recombination event in genotype G that involved genotype A. Altogether, these findings indicated that genotypes F and G had their respective distinctive genotypic features at the nucleotide and amino acid levels. Collectively, with the phylogenetic clustering, pairwise genetic distance, recombination and comparative genomic analyses between genotypes indicate that both genotypes F and G represent two distinct HCoV-OC43 genetic lineages that have descended from the genotype D parental lineage through recombination with a genotype A-like lineage.

Global circulation and divergence times of genotypes F and G

Phylogenetic analysis of the partial S gene (S1 domain) in a previous study demonstrated that several HCoV-OC43 strains from China ($n=2$), Thailand ($n=3$) and Japan ($n=5$) were clustered together with genotype F (previously known as lineage 1), whereas another two strains from China were clustered with genotype G (lineage 2).²⁰ When amino acid sequence analysis was performed on the global partial S gene (S1 domain: 23 644 to 25 125 nt) in this study (using HCoV-OC43 genotype D as the reference strain), these genotype F strains from China, Thailand, Japan, and three newly deposited sequences from France (clustered together with genotype F strains in the MCC tree in Figure 5A) shared a signature amino acid substitution Y176H with the newly sequenced genotype F strains from Malaysia (Figure 4A). In addition, amino acid substitutions R26K, T93K and I181S, which were observed in genotype F strains from Malaysia, were also present in strains reported in China and/or France. In addition, it is interesting to note that the glycine insertion at position 119 was only present in the Malaysian strains. Though, despite this unique insertion, the bootscan analysis of the partial S gene (using a narrow sliding window of 300 bp and step size of 10 bp for improved resolution) revealed that all genotype F strains in these countries shared similar mosaic recombination structure (Figure 4B). The same could be observed for genotype G strains from China ($n=2$), which shared four signature amino acid substitutions P22T, D267K, I268D and T271S and similar mosaic recombination pattern in the partial S gene with the Malaysian strains (Figures 4A and 4B). In general, the amino acid and bootscan analyses of the partial S gene indicated that genetically similar HCoV-OC43 genotypes F and G strains can be found circulating in a number of Asian countries and Europe (Figure 5B).

To investigate the spatiotemporal and evolutionary dynamics of all HCoV-OC43 genotypes, divergence times were estimated by performing molecular clock dating analysis on 29 full-length genome sequences (including 16 new full-length genomes generated in this study) (Supplementary Table S2). Given increased accessibility to the S gene sequences in the public domain, similar dating analysis was estimated using 114 complete and partial S gene sequences (Supplementary Table S3). To infer the mean tMRCA and the 95% highest posterior density (HPD), the exponential population size under a relaxed-clock model with BSP distribution and uncorrelated exponential distribution were adopted for the S gene and full-length genome, respectively. Both models were the best data-fitting coalescent models and were selected by means of marginal likelihoods (specifically Bayes factor), as estimated using the smoothed harmonic mean estimator³⁵ and by means of Akaike's Information Criterion for MCMC samples estimated using the method-of-moments estimator implemented in Tracer (data not shown).³⁶ The molecular clock dating analysis estimated the mean evolutionary rate (and 95% HPD) for the S gene and full-length genome of all HCoV-OC43 strains based on their respective coalescent models at 5.8×10^{-4} (4.4×10^{-4} to 7.1×10^{-4}) and 1.8×10^{-4} (1.2×10^{-4} to 2.4×10^{-4}) nucleotide substitutions per site per year, respectively. The estimate of the mean evolutionary rate for the S gene is comparable to previous findings of 6.1×10^{-4} – 6.7×10^{-4} nucleotide substitutions per site per year,^{7,20,37} whereas the rate for a full-length genome, to our knowledge, is newly estimated in this study. As shown in Table 2, the estimated tMRCA based on the S gene and full-length genome data for genotype A was in the 1960s, genotypes B to E were in the late 1990s to mid-2000s, and genotypes F and G were in the late 2000s to early 2010s. The estimates from both sets of data were comparable, indicating that the S gene or the full-length genome could be used for tMRCA estimation.

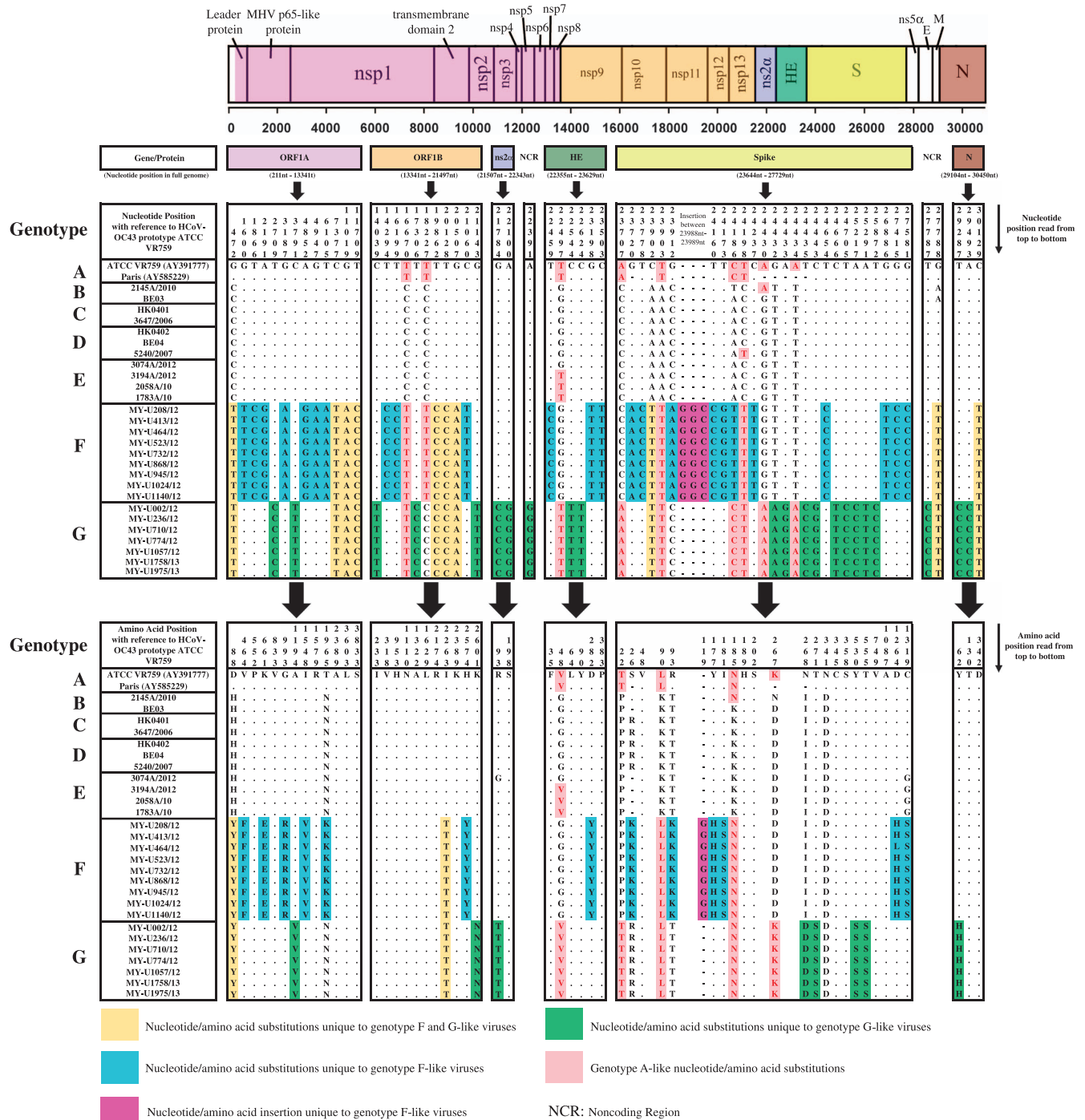


Figure 3 Signature nucleotide and amino acid substitution differences across the whole genome between genotypes F and G strains. Nucleotide and amino acid positions are numbered with a reference to HCoV-OC43 prototype strain ATCC VR759.

HCOV-OC43 transmission network

To investigate the transmission pattern of HCoV-OC43 genotypes D, F and G, transmission clusters were constructed based on the pairwise distances using the TN93 model estimated from 86S gene sequences, which included sequences from China, Japan, Thailand, France and Malaysia collected between 2002 and 2013 (Supplementary Table S3). The 95% confidence interval of the lower 0.025 percentile of the inter-person patristic distance was calculated at 0.001 substitutions per site, which represented the distance threshold for estimating HCoV-OC43

transmission cluster (Supplementary Table S4). Forty-eight sequences (55.8%, 48/86) formed a total of ten transmission clusters with strong spatial structure, of which four dyads and six networks of different sizes ranging between 3 and 13 nodes per network were estimated (Figure 6). For genotype D, five transmission clusters involved sequences that were isolated from China between 2007 and 2010, whereas one transmission network was shared among sequences from China and Thailand sampled within a 2-year period (2008–2010). However, three genotype F and one genotype G clusters were

circulating exclusively within their particular countries of origin: Japan in 2011, China and Malaysia in 2012 (Figure 6).

DISCUSSION

HCoV-OC43 strains are associated with respiratory diseases and have caused outbreaks worldwide.^{8,19,38,39} Despite its first discovery in 1967,³⁴ full-length genomes of known and published HCoV-OC43 genotypes were limited. More recently, genotyping studies have become more common, with the first description of a complete genome from a laboratory ATCC strain and a clinical isolate from France in 2004 (genotype A)³⁴ followed by two Belgium strains in 2005 (genotypes B and C).⁵ In the early 2010s, genotyping studies further highlighted the epidemiological impact of recombination in driving the emergence of two more novel HCoV-OC43 genotypes

(genotypes D and E). Genotype D was a result of recombination between genotypes B and C,⁷ whereas genotype E was a recombinant among genotypes B, C and D.⁹

The global emergence and re-emergence of viral respiratory disease outbreaks have prompted more active pathogen surveillance initiatives in major healthcare settings worldwide, including in Southeast Asia. In Malaysia, a cross-sectional molecular surveillance of HCoV-OC43 was conducted among patients with acute URTI in a major teaching hospital in Kuala Lumpur.²⁰ On the basis of the phylogenetic analysis of the S gene, two unique lineages (lineages 1 and 2) appeared to diverge from genotype D. HCoV-OC43 strains found in other geographical regions were also grouped within the two lineages. Phylogenetic incongruence found in the partial genes of these two unique lineages indicated possible recombination between genotypes,

A

Genotypes	Year	Country	Subunit/Domain Amino Acid Positons	NTD	NTD	NTD	NTD	NTD	NTD	NTD	NTD	NTD	NTD	NTD
				22	26	90	93	119	177	181	185	267	268	271
			Amino Acid Positons (as according to Ren et al.)	22	26	90*	93*		176*	180	184*	266*	267*	270
Genotype D	2002	Hong Kong, China	HK04-02	P	R	K	T	-	Y	I	K	D	I	T
	2004	Belgium	BE04	-
	2007	Beijing, China	5240/2007	-
	2008	Beijing, China	892A/08	-
Genotype F	2009	Thailand	CU-H967/2009#	.	.	L	.	-	H	.	N	.	.	.
	2009	France	HCoV-OC43/FRA_EPI/Caen/2009/11#	.	.	L	.	-	H	.	N	.	.	.
	2010	Thailand	CU-H1444/2010#	.	.	L	.	-	H	.	N	.	.	.
	2010	Thailand	CU-H1367/2010#	.	.	L	.	-	H	.	N	.	.	.
	2011	Japan	HCoV-OC43/Niigata.JPN/11-769#	.	.	L	.	-	H	.	N	.	.	.
	2011	Japan	HCoV-OC43/Niigata.JPN/11-981#	.	.	L	.	-	H	.	N	.	.	.
	2011	Japan	HCoV-OC43/Niigata.JPN/11-768#	.	.	L	.	-	H	.	N	.	.	.
	2011	Japan	HCoV-OC43/Niigata.JPN/11-764#	.	.	L	.	-	H	.	N	.	.	.
	2011	Japan	HCoV-OC43/Niigata.JPN/11-833#	.	.	L	.	-	H	.	N	.	.	.
	2012	France	HCoV-OC43/FRA_EPI/Caen/2012/14#	.	K	L	.	-	H	.	N	.	.	.
	2013	France	HCoV-OC43/FRA_EPI/Caen/2013/15#	.	K	L	.	-	H	.	N	.	.	.
	2012	Beijing, China	12691/12#	.	K	L	K	-	H	S	N	.	.	.
	2012	Beijing, China	3269A/12#	.	K	L	K	-	H	S	N	.	.	.
	2012	Malaysia	MY-U868/12	.	K	L	K	G	H	S	N	.	.	.
	2012	Malaysia	MY-U208/12	.	K	L	K	G	H	S	N	.	.	.
	2012	Malaysia	MY-U413/12	.	K	L	K	G	H	S	N	.	.	.
	2012	Malaysia	MY-U464/12	.	K	L	K	G	H	S	N	.	.	.
	2012	Malaysia	MY-U523/12	.	K	L	K	G	H	S	N	.	.	.
	2012	Malaysia	MY-U732/12	.	K	L	K	G	H	S	N	.	.	.
2012	Malaysia	MY-U945/12	.	K	L	K	G	H	S	N	.	.	.	
2012	Malaysia	MY-U1024/12	.	K	L	K	G	H	S	N	.	.	.	
2012	Malaysia	MY-U1140/12	.	K	L	K	G	H	S	N	.	.	.	
Genotype G	2012	Beijing, China	12694/12#	T	.	L	.	-	.	.	N	K	D	S
	2012	Beijing, China	12689/12#	T	.	L	.	-	.	.	N	K	D	S
	2012	Malaysia	MY-U002/12	T	.	L	.	-	.	.	N	K	D	S
	2012	Malaysia	MY-U710/12	T	.	L	.	-	.	.	N	K	D	S
	2012	Malaysia	MY-U774/12	T	.	L	.	-	.	.	N	K	D	S
	2012	Malaysia	MY-U1057/12	T	.	L	.	-	.	.	N	K	D	S
	2012	Malaysia	MY-U236/12	T	.	L	.	-	.	.	N	K	D	S
	2013	Malaysia	MY-U1758/12	T	.	L	.	-	.	.	N	K	D	S
	2013	Malaysia	MY-U1975/12	T	.	L	.	-	.	.	N	K	D	S

Amino acid substitutions shared by Genotype F and G-like viruses
 Amino acid substitutions shared by Genotype G-like viruses

Amino acid substitutions shared by Genotype F-like viruses
 Amino acid substitutions shared by Malaysian/China/France Genotype F-like viruses

* Positively selected sites
 # Previously classified as Genotype D-like viruses
 NTD Non-terminal domain

Figure 4 Continued.

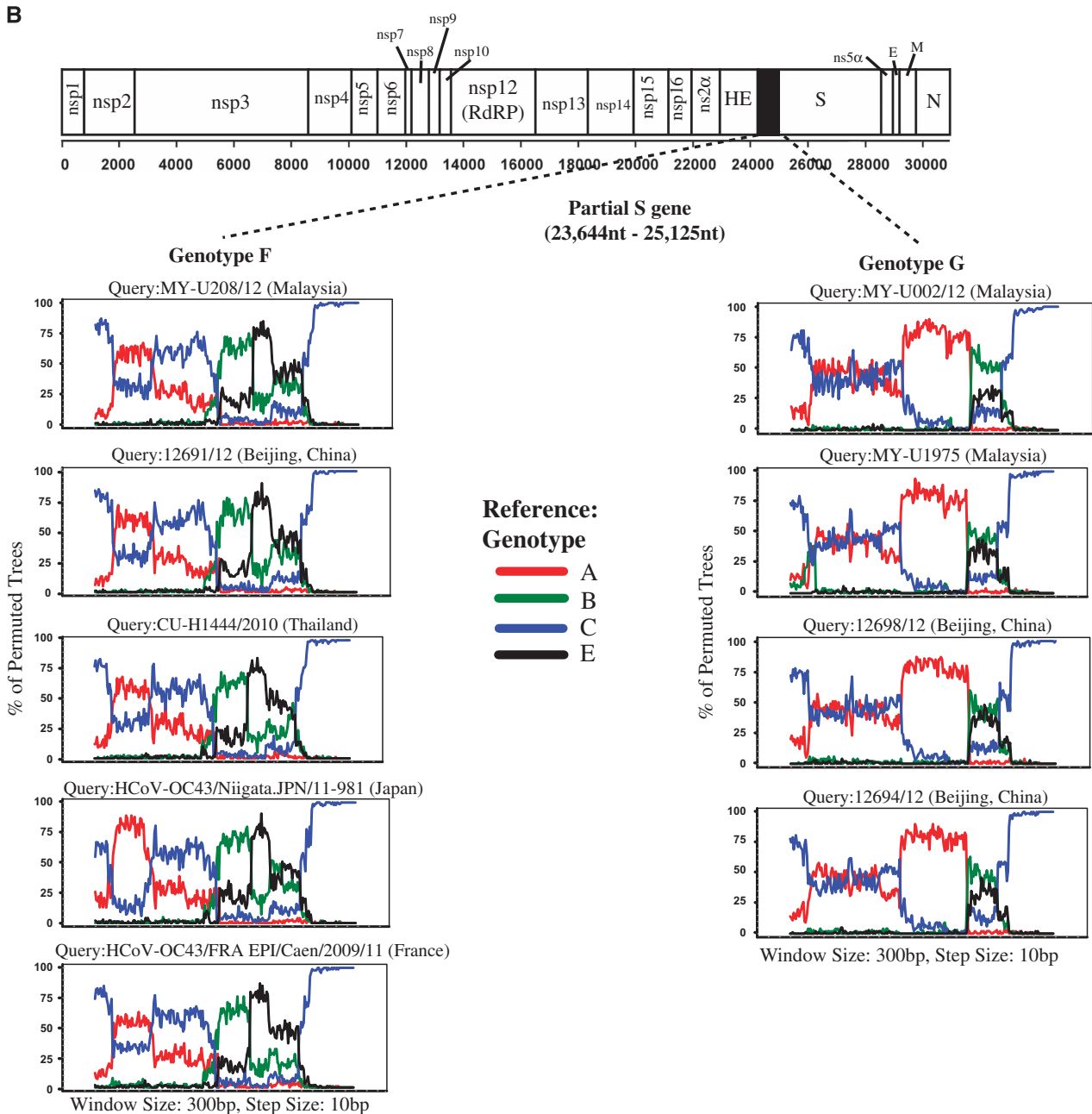


Figure 4 Evidence from (A) amino acid substitutions and (B) bootscan analysis on the partial S gene data to confirm the presence of genotypes F and G strains. The amino acid sequence and bootscan analyses are performed on the partial S gene region (23 644–25 125 nt). Bootscan analysis is performed with a window and step size of 300 and 10 bp, respectively.

which prompted the sequencing of their complete genomes presented in this study.

Our analyses on the full-length genome sequences of the unique HCoV-OC43 strains from novel lineages 1 and 2²⁰ confirm the identification of two novel genotypes, which are designated as genotypes F and G. These two novel genotypes were descendants of a previously reported recombinant genotype D,⁷ which contained genotypes B and C as the putative parental genotypes and a genotype A-like genetic signal in the ORF1b gene. The recombination breakpoints were located at ~16 000–17 000 nt and 17 000–17 500 nt, which

corresponds to the nsp9/nsp10 junctions (Figure 2A). Previous studies on HCoV-OC43 genotypes D and E genomes have reported potential recombination sites at the nsp2/nsp3, nsp6/nsp7, nsp9/nsp10, nsp12/nsp13, ns2 α /HE, ns5 α /E and M/N junctions.^{7,9} In addition, recombination breakpoints in the nsp5/nsp6, nsp16/S and nsp14/nsp15 junctions were also identified in other HCoV genomes, such as HCoV-HKU1,¹⁵ SARS-CoV⁴⁰ and MERS-CoV,⁴¹ respectively. It is notable that the ORF1ab region (a region that encodes for non-structural proteins) is probably more recombination-prone compared with other regions in the HCoV genome. Recombination in this

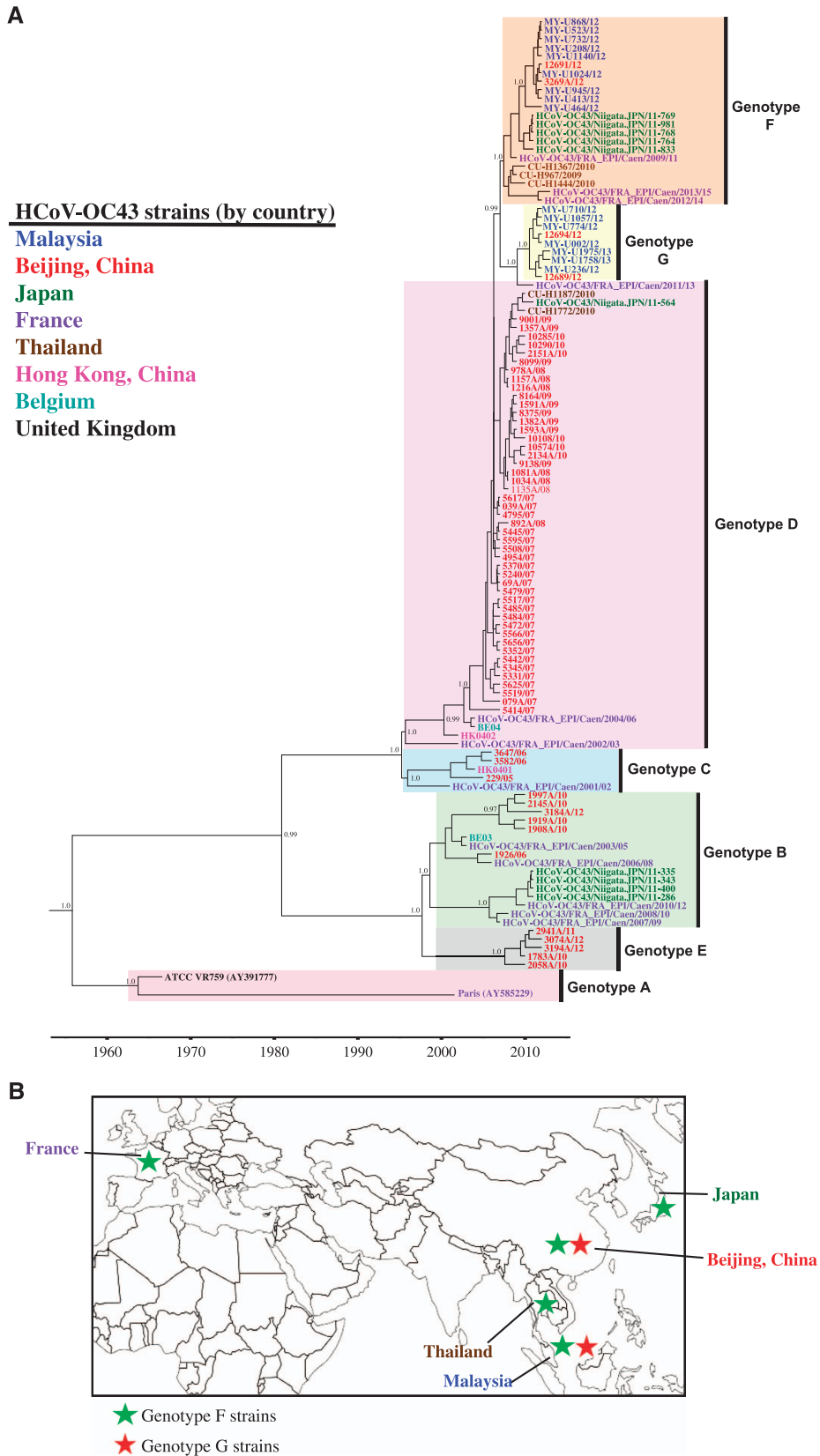


Figure 5 (A) Maximum clade credibility (MCC) tree of HCov-OC43 strains based on the 114 complete and partial global S gene data. MCC posterior probability values were indicated on the nodes of each genotype. (B) Global distribution of genotypes F and G strains.

Table 2 Time of most recent common ancestor (tMRCA) for HCoV-OC43 genotypes A to G estimated based on the spike (S) gene and full-length genome

Genotype	tMRCA (95% HPD)	
	Spike (S) gene	Full-length genome
A	1963.8 (1959.4–1966.7)	1964.7 (1960.3–1966.8)
B	1999.1 (1996.6–2001.5)	1999.0 (1993.5–2002.6)
C	1997.2 (1994.0–1999.7)	1998.0 (1994.3–2000.5)
D	1997.3 (1994.2–2000.4)	1999.3 (1996.5–2001.5)
E	2007.9 (2005.9–2009.4)	2003.3 (1999.1–2006.9)
F	2010.8 (2010.0–2011.4)	2009.8 (2008.1–2011.2)
G	2010.7 (2010.0–2011.4)	2010.1 (2008.5–2011.4)

Abbreviation: highest posterior density, HPD.

pressure exerted by the host immune response.⁴ Three signature amino acid mutations at the spike protein, H177Y (genotype F), D267K and I268D (genotype G) (or H176Y, D266K and I267D according to Ren *et al.*,¹¹ were among the positively selected sites identified at the N-terminal domain (NTD) of the S1 subunit. More interestingly, a three-nucleotide insertion (GGC), which resulted in the introduction of a glycine residue in the NTD, was unique only to the Malaysian genotype F strains. Whether the introduction of glycine has a role in enhancing the binding of NTD to the sugar receptors of the host cells, which subsequently enhances the pathogenicity of HCoV-OC43,⁴ requires further investigation.

In this study, evidence from phylogenetic, amino acid and recombination analyses was used to demonstrate the global distribution of HCoV-OC43 genotypes F and G. Due to limited published and genotyped full-length genome data in the database, a larger amount of S gene sequence data was utilized. Evidence from these analyses revealed that several strains that were previously classified as genotype D from China, Thailand, Japan and France belonged to genotypes F or G. Such misclassification highlights the potential weaknesses in the current classification system that relies primarily on the phylogenetic analysis of three different parts of the genome: S, RdRP (nsp12), and N genes. The use of these parts promotes bias towards misclassification of strains with recombination occurring outside these regions.^{7,9,10,19,20,42} In addition, the highly conserved RdRP and N genes typically result in poorly resolved phylogenetic trees,^{9,10,20} which is not ideal for precise genotype classification. Therefore, to minimize the underestimation of recombinant strains, full-length genomes should be characterized for new genotype designation and shared in public databases as reference genotypes. As shown in Figure 1, the pairwise genetic distances of full-length genome sequences estimated between non-recombinant HCoV-OC43 genotypes (A vs. B, A vs. C, and B vs. C) were >0.6%, which is suggestive of a minimal genetic distance for a genotype assignment. Such a cutoff, however, is not applicable for recombinant lineages that contain genetic information from multiple parental genotypes with varying evolutionary histories. Similar to the more established nomenclature systems used for the human immunodeficiency virus type 1 and hepatitis C virus,^{43,44} we propose that inter-genotype recombinant strains of HCoV-OC43 identified among multiple individuals can be assigned as ‘recombinant form’ (RF) to reflect the recombination origin of the lineage. The RF candidates, which descended from the same parental genotypes via recombination event(s), must share an identical mosaic recombinant structure based on the full-length genome sequences. RF candidates are suffixed with an identifying number in the order in which they were first described (01, 02, 03 and so on) followed by letters (listed alphabetically), indicating the parental genotypes involved in the recombination. For example, HCoV-OC43 genotype D, which was the first described inter-genotype recombinant involving genotypes B and C, can be designated as RF01_BC. For clarity, if more than two parental genotypes are involved in the recombination, a complex (cpx) recombinant lineage can be assigned. Therefore, in the case of genotype E and the newly described genotypes F and G, they can be (re)designated as RF02_cpx, RF03_cpx and RF04_cpx, respectively. The newly proposed nomenclature system for non-recombinant and recombinant HCoV-OC43 herein will establish a platform for a more consistent and reliable naming system that considers genetic recombination as an important and common feature in HCoV evolution. The application of a naming system that reflects the evolutionary histories of HCoV-OC43 strains will also enable better mapping and tracking of newly emerging recombinant lineages, which have become increasingly prevalent in large parts of Asia and elsewhere.^{7,9,20}

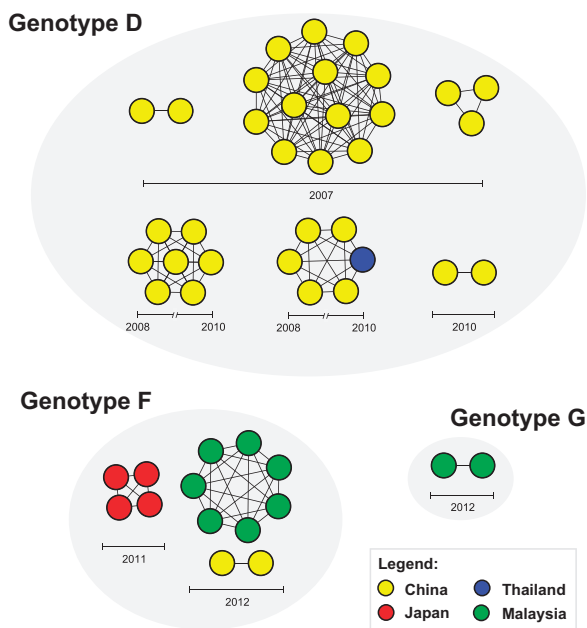


Figure 6 Transmission network of HCoV-OC43 genotypes D, F and G. Transmission clusters were inferred from 86 S gene sequences based on the Tamura-Nei 93 (TN93)²⁹ pairwise distance estimates performed using a custom script in Python (release 3.2.6) with 1000 bootstrap replicates.

region typically contributes to the generation of new HCoV genotypes,^{7,9,15,40,41} suggesting the importance to target this region for molecular evolutionary and epidemiological investigations.

Apart from recombination analysis, genotypes F and G strains also shared four and eight nucleotide substitutions, respectively, with genotype A within the ORF1b, HE and S gene region (Figure 3). This finding implies that minute genotype A-like genetic signals found within these regions differentiate genotypes F and G; however, the breakpoints could not be clearly resolved through bootscan analysis. In addition to minor differences in the recombination pattern, the distinction of the genomes between genotypes F and G could be attributed to their possession of unique signature nucleotide and amino acid mutations across the full-length genomes. Most of the signature mutations found only in genotypes F or G occurred in the S gene (Figure 3), which is not uncommon given that the spike protein is a major antigenic surface protein that undergoes high selection

Communicable diseases, such as virus-associated respiratory infections, are transmitted through close contacts between individuals within networks. The role of networks in fueling and sustaining the onward transmission of pathogen is profound,⁴⁵ unless effective intervention measures are being introduced. The inclusion of transmission clusters information coupled with epidemiological surveillance data could help to better understand the spread and dynamics of viral diseases attributed to specific genotypes circulating in the population. To the best of our knowledge, such analysis represents an inventive approach of which the spread and dynamics of HCoV-OC43 transmission was mapped. Using the TN93 pairwise distance estimates of the S gene to establish genetically-related HCoV-OC43 strains with apparent transmission linkage, we inferred the transmission clusters of genotypes D, F and G sampled across Asia and Europe between 2002 and 2013. Greater than 55% of the global spread of genotype D were linked to transmission clusters, in which 90% of the transmissions were restricted within China, except for one genotype D network that involved sequences from China and Thailand, suggesting cross-border transmission and persistence of a transmission network for up to a 2-year period (Figure 6). In addition, the emergence and persistence of multiple sub-epidemic clusters of genotype D corroborates with the recent predominance and continual transmission of this genotype in China.^{9,11}

In summary, herein we report two novel HCoV-OC43 recombinant genotypes, designated as genotypes F and G, which were identified among patients presenting with acute respiratory tract symptoms. Using observations from phylogenetic, recombination, and comparative genomic analyses on full-length genome sequences, HCoV-OC43 genotypes F and G were likely to co-circulate worldwide. Bayesian coalescent dating analysis implied that both genotypes probably diverged concurrently around the late 2000s to early 2010s from a genotype D-like common ancestor through natural recombination a few years after genotype D was first identified in Asia. The wide distribution of HCoV-OC43 genotype D and its recombinant lineages were driven in part by the emergence and persistence of transmission networks in this region. Overall, our results highlight the importance of recombination in HCoV-OC43 evolution, which warrants a more consistent nomenclature system for classification and better tracking of newly emerged recombinant lineages. More comprehensive molecular surveillance studies on HCoV-OC43 are essential to better understand the evolutionary dynamics, pathogenesis, and disease burden of HCoV-OC43 infections.

ACKNOWLEDGEMENTS

This work was supported by grants from the Ministry of Education, Malaysia: High Impact Research UM.C/625/1/HIR/MOE/CHAN/02/02 to KKT and the Postgraduate Research Fund (PG084-2015A) to XYO. The funders were not involved in study design or data collection and analysis. We would like to thank Nyoke Pin Wong (Department of Medical Microbiology, Faculty of Medicine, University of Malaya, 50603 Kuala Lumpur, Malaysia), Joon Ling Tan (Department of Medical Microbiology, Faculty of Medicine, University of Malaya, 50603 Kuala Lumpur, Malaysia), Maryam Nabel Al-Khannaq (Department of Medicine, Faculty of Medicine, University of Malaya, 50603 Kuala Lumpur, Malaysia), Yong Kek Pang (Department of Medicine, Faculty of Medicine, University of Malaya, 50603 Kuala Lumpur, Malaysia) and Nik Sherina Hanafi (Department of Primary Care Medicine, Faculty of Medicine, University of Malaya, 50603 Kuala Lumpur, Malaysia) for assistance and support.

- 1 Büchen-Osmond C, Dallwitz MJ. *ICTVdB: The Universal Virus Database, Version 3*. Columbia University: New York, NY, USA; 2006.
- 2 Lee J, Storch GA. Characterization of human coronavirus OC43 and human coronavirus NL63 infections among hospitalized children <5 years of age. *Pediatr Infect Dis J* 2014; **33**: 814–820.
- 3 Jean A, Quach C, Yung A *et al*. Severity and outcome associated with human coronavirus OC43 infections among children. *Pediatr Infect Dis J* 2013; **32**: 325–329.
- 4 Masters PS, Perlman S. Coronaviridae. *Fields Virol* 2013; **1**: 826–859.
- 5 Vijgen L, Keyaerts E, Lemey P *et al*. Circulation of genetically distinct contemporary human coronavirus OC43 strains. *Virology* 2005; **337**: 85–92.
- 6 Vabret A, Dina J, Mourez T *et al*. Inter- and intra-variant genetic heterogeneity of human coronavirus OC43 strains in France. *J Gen Virol* 2006; **87**: 3349–3353.
- 7 Lau SK, Lee P, Tsang AK *et al*. Molecular epidemiology of human coronavirus OC43 reveals evolution of different genotypes over time and recent emergence of a novel genotype due to natural recombination. *J Virol* 2011; **85**: 11325–11337.
- 8 Hu Q, Lu R, Peng K *et al*. Prevalence and genetic diversity analysis of human coronavirus OC43 among adult patients with acute respiratory infections in Beijing, 2012. *PLoS One* 2014; **9**: e100781.
- 9 Zhang Y, Li J, Xiao Y *et al*. Genotype shift in human coronavirus OC43 and emergence of a novel genotype by natural recombination. *J Infect* 2015; **70**: 641–650.
- 10 Kin N, Miszczak F, Lin W *et al*. Genomic analysis of 15 human coronavirus OC43 (HCoV-OC43s) circulating in France from 2001 to 2013 reveals a high intra-specific diversity with new recombinant genotypes. *Viruses* 2015; **7**: 2358–2377.
- 11 Ren L, Zhang Y, Li J *et al*. Genetic drift of human coronavirus OC43 spike gene during adaptive evolution. *Sci Rep* 2015; **5**: 11451.
- 12 Makino S, Keck JG, Stohlman SA *et al*. High-frequency RNA recombination of murine coronaviruses. *J Virol* 1986; **57**: 729–737.
- 13 van der Most RG, Heijnen L, Spaan WJ *et al*. Homologous RNA recombination allows efficient introduction of site-specific mutations into the genome of coronavirus MHV-A59 via synthetic co-replicating RNAs. *Nucleic Acids Res* 1992; **20**: 3375–3381.
- 14 Woo PC, Wang M, Lau SK *et al*. Comparative analysis of twelve genomes of three novel group 2c and group 2d coronaviruses reveals unique group and subgroup features. *J Virol* 2007; **81**: 1574–1585.
- 15 Woo PC, Lau SK, Yip CC *et al*. Comparative analysis of 22 coronavirus HKU1 genomes reveals a novel genotype and evidence of natural recombination in coronavirus HKU1. *J Virol* 2006; **80**: 7136–7145.
- 16 Menachery VD, Yount Jr BL, Debbink K *et al*. A SARS-like cluster of circulating bat coronaviruses shows potential for human emergence. *Nat Med* 2015; **21**: 1508–1513.
- 17 Wang Y, Liu D, Shi W *et al*. Origin and possible genetic recombination of the Middle East Respiratory Syndrome Coronavirus from the first imported case in China: phylogenetics and coalescence analysis. *MBio* 2015; **6**: e01280–01215.
- 18 Gerna G, Campanini G, Rovida F *et al*. Genetic variability of human coronavirus OC43-, 229E-, and NL63- like strains and their association with lower respiratory tract infections of hospitalized infants and immunocompromised patients. *J Med Virol* 2006; **78**: 938–949.
- 19 Kon M, Watanabe K, Tazawa T *et al*. Detection of human coronavirus NL63 and OC43 in children with acute respiratory infections in Niigata, Japan, between 2010 and 2011. *Jpn J Infect Dis* 2012; **65**: 270–272.
- 20 Al-Khannaq MN, Ng KT, Oong XY *et al*. Molecular epidemiology and evolutionary histories of human coronavirus OC43 and HKU1 among patients with upper respiratory tract infections in Kuala Lumpur, Malaysia. *Viral J* 2016; **13**: 33.
- 21 Malaysian Meteorological Department. General climate of Malaysia. Petaling Jaya: MMD; 2014. Available at: <http://www.met.gov.my/web/metmalaysia/climate/general-information/malaysia>. Accessed on 1 March 2015.
- 22 Pabbaraju K, Wong S, Tokaryk KL *et al*. Comparison of the Luminex xTAG respiratory viral panel with xTAG respiratory viral panel fast for diagnosis of respiratory virus infections. *J Clin Microbiol* 2011; **49**: 1738–1744.
- 23 Loens K, Bergs K, Ursi D *et al*. Evaluation of NucliSens easyMAG for automated nucleic acid extraction from various clinical specimens. *J Clin Microbiol* 2007; **45**: 421–425.
- 24 Katoh K, Asimenos G, Toh H. Multiple alignment of DNA sequences with MAFFT. *Bioinform DNA Seq Anal* 2009; **537**: 39–64.
- 25 Tamura K, Stecher G, Peterson D *et al*. MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol* 2013; **30**: 2725–2729.
- 26 Swofford DL. *PAUP*: Phylogenetic Analysis Using Parsimony (* and other methods)*, Version 4. Sinauer Associates: Sunderland, MA, USA, 2003.
- 27 Lole KS, Bollinger RC, Paranjape RS *et al*. Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *J Virol* 1999; **73**: 152–160.
- 28 Drummond AJ, Suchard MA, Xie D *et al*. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* 2012; **29**: 1969–1973.
- 29 Wertheim JO, Brown AJL, Hepler NL *et al*. The global transmission network of HIV-1. *J Infect Dis* 2013; **209**: 304–313.
- 30 Chow WZ, Chan YF, Oong XY *et al*. Genetic diversity, seasonality and transmission network of human metapneumovirus: identification of a unique sub-lineage of the fusion and attachment genes. *Sci Rep* 2016; **6**: 27730.
- 31 Poon AF, Joy JB, Woods CK *et al*. The impact of clinical, demographic and risk factors on rates of HIV transmission: a population-based phylogenetic analysis in British Columbia, Canada. *J Infect Dis* 2014; **211**: 926–935.

- 32 Olmstead AD, Joy JB, Montoya V *et al*. A molecular phylogenetics-based approach for identifying recent hepatitis C virus transmission events. *Infect Genet Evol* 2015; **33**: 101–109.
- 33 Little SJ, Pond SLK, Anderson CM *et al*. Using HIV networks to inform real time prevention interventions. *PLoS One* 2014; **9**: e98443.
- 34 St-Jean JR, Jacomy H, Desforges M *et al*. Human respiratory coronavirus OC43: genetic stability and neuroinvasion. *J Virol* 2004; **78**: 8824–8834.
- 35 Suchard MA, Weiss RE, Sinsheimer JS. Bayesian selection of continuous-time Markov chain evolutionary models. *Mol Biol Evol* 2001; **18**: 1001–1013.
- 36 Baele G, Lemey P, Bedford T *et al*. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Mol Biol Evol* 2012; **29**: 2157–2167.
- 37 Vijgen L, Keyaerts E, Lemey P *et al*. Evolutionary history of the closely related group 2 coronaviruses: porcine hemagglutinating encephalomyelitis virus, bovine coronavirus, and human coronavirus OC43. *J Virol* 2006; **80**: 7270–7274.
- 38 Vabret A, Mourez T, Gouarin S *et al*. An outbreak of coronavirus OC43 respiratory infection in Normandy, France. *Clin Infect Dis* 2003; **36**: 985–989.
- 39 Walsh EE, Shin JH, Falsey AR. Clinical impact of human coronaviruses 229E and OC43 infection in diverse adult populations. *J Infect Dis* 2013; **208**: 1634–1642.
- 40 Lau SK, Li KS, Huang Y *et al*. Ecoepidemiology and complete genome comparison of different strains of severe acute respiratory syndrome-related Rhinolophus bat coronavirus in China reveal bats as a reservoir for acute, self-limiting infection that allows recombination events. *J Virol* 2010; **84**: 2808–2819.
- 41 Sabir JS, Lam TT, Ahmed MM *et al*. Co-circulation of three camel coronavirus species and recombination of MERS-CoVs in Saudi Arabia. *Science* 2016; **351**: 81–84.
- 42 Suwannakarn K, Chieochansin T, Vichiwattana P *et al*. Prevalence and genetic characterization of human coronaviruses in Southern Thailand from July 2009 to January 2011. *Southeast Asian J Trop Med Public Health* 2014; **45**: 326–336.
- 43 Robertson DL, Anderson JP, Bradac JA *et al*. HIV-1 nomenclature proposal. *Science* 2000; **288**: 55–56.
- 44 Simmonds P, Bukh J, Combet C *et al*. Consensus proposals for a unified system of nomenclature of hepatitis C virus genotypes. *Hepatology* 2005; **42**: 962–973.
- 45 Craft ME. Infectious disease transmission and contact networks in wildlife and livestock. *Philos Trans R Soc B* 2015; **370**: 20140107.



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

© The Author(s) 2017

Supplementary Information for this article can be found on the *Emerging Microbes & Infections* website (<http://www.nature.com/emi>)