Genomic analysis of 16 Colorado human NL63 coronaviruses identifies a new genotype, high sequence diversity in the N-terminal domain of the spike gene and evidence of recombination

Samuel R. Dominguez,1,2 Gregory E. Sims,3† David E. Wentworth,3 Rebecca A. Halpin,3 Christine C. Robinson,4 Christopher D. Town3 and Kathryn V. Holmes2

1Department of Pediatrics, University of Colorado School of Medicine, Anschutz Medical Campus, 12800 E 19th Ave, Room P18-9403B, Aurora, CO 80045, USA
2Department of Microbiology, University of Colorado School of Medicine, Anschutz Medical Campus, 12800 E 19th Ave, Room P18-9403B, Aurora, CO 80045, USA
3J. Craig Venter Institute, 9704 Medical Center Drive, Rockville, MD 20850, USA
4Department of Pathology and Clinical Medicine, Children’s Hospital Colorado, 13123 E 16th Ave, Aurora, CO 80045, USA

This study compared the complete genome sequences of 16 NL63 strain human coronaviruses (hCoVs) from respiratory specimens of paediatric patients with respiratory disease in Colorado, USA, and characterized the epidemiology and clinical characteristics associated with circulating NL63 viruses over a 3-year period. From 1 January 2009 to 31 December 2011, 92 of 9380 respiratory specimens were found to be positive for NL63 RNA by PCR, an overall prevalence of 1%. NL63 viruses were circulating during all 3 years, but there was considerable yearly variation in prevalence and the month of peak incidence. Phylogenetic analysis comparing the genome sequences of the 16 Colorado NL63 viruses with those of the prototypical hCoV-NL63 and three other NL63 viruses from the Netherlands demonstrated that there were three genotypes (A, B and C) circulating in Colorado from 2005 to 2010, and evidence of recombination between virus strains was found. Genotypes B and C co-circulated in Colorado in 2005, 2009 and 2010, but genotype A circulated only in 2005 when it was the predominant NL63 strain. Genotype C represents a new lineage that has not been described previously. The greatest variability in the NL63 virus genomes was found in the N-terminal domain (NTD) of the spike gene (nt 1–600, aa 1–200). Ten different amino acid sequences were found in the NTD of the spike protein among these NL63 strains and the 75 partial published sequences of NTDs from strains found at different times throughout the world.

INTRODUCTION

Respiratory tract infections cause the greatest burden of disease worldwide, surpassing that of human immunodeficiency virus infection, malaria, cancer and heart disease. The World Health Organization estimates that respiratory tract infections are the leading cause of death in low-income countries (WHO, 2003, 2008). Even in the USA, respiratory tract infections cause more disease and death than any other type of infection, and over the past 50 years there has been little decrease in mortality associated with lung infections (Mizgerd, 2006, 2008). Over the past decade, eight novel respiratory viral pathogens have been discovered (Allander et al., 2005, 2007; Dominguez et al., 2008; Gaynor et al., 2007; Lamson et al., 2006; Lau et al., 2007; McErlean et al., 2007; van den Hoogen et al., 2001), including three new human coronaviruses (hCoVs) (Drosten et al., 2003; van der Hoek et al., 2004; Woo et al., 2005).

There are now five known hCoV species, exemplified by the prototypical strains hCoV-229E, hCoV-OC43, hCoV-HKU1,
severe acute respiratory syndrome (SARS)-CoV and hCoV-NL63. Strains within each of these hCoV species, except for SARS-CoV, cause worldwide seasonal epidemics of respiratory disease every year. With the possible exception of hCoV-OC43, humans are the only known host for these CoVs. The antigenicity, clinical presentations and disease severity of CoVs of many mammalian and avian species can be affected by mutations that cause amino acid substitutions, deletions or insertions, and by recombination between viral RNA genomes during co-infections. The prototype hCoV-NL63 strain (NL63/AMS/2004/1) was identified in 2004, and NL63 viruses have subsequently been shown to have a worldwide distribution (Fouchier et al., 2004; van der Hoek et al., 2004). Retrospective epidemiological studies by our laboratory and others suggest that the prevalence of NL63 viruses in children with respiratory illness ranges from 2 to 9%, and that these NL63 viruses are associated with asymptomatic infections and mild upper respiratory tract infections, as well as with pneumonia, bronchiolitis and croup (Arden et al., 2005; Bastien et al., 2005a, b; Chiu et al., 2005; Dominguez et al., 2009; Ebihara et al., 2005; Esper et al., 2005; Kaiser et al., 2005; Kuypers et al., 2007; Prill et al., 2012; Suzuki et al., 2005; Vabret et al., 2005; van der Hoek et al., 2005). Until recently, it has been difficult or impossible to isolate these NL63 viruses from clinical specimens in continuous cell lines, although the prototype NL63 strain has been propagated on LLC-MK2 cells (van der Hoek et al., 2004) and in primary, differentiated human bronchial–tracheal respiratory epithelial cells cultured at the air–liquid interface (Banach et al., 2009).

To date, only four full-genome sequences of NL63 viruses have been deposited in GenBank, and all of these sequences are from clinical specimens collected in the Netherlands (Fouchier et al., 2004; Pyrc et al., 2006; van der Hoek et al., 2004). Analysis of these genome sequences identified two distinct genotypes, A and B, of NL63 viruses with an overall genome identity of 99%, and one recombinant between these strains (Pyrc et al., 2006). Here, we report the epidemiology and disease association of NL63 viruses over a 3-year period (2009–2011) at the Children’s Hospital Colorado (CO, USA). Sequencing the NL63 genomes from 23 of our clinical specimens yielded 16 full-length genomes (>99% coverage) and identified genomic signatures of at least three distinct circulating genotypes of NL63 and one recombinant. These genomes are the first NL63 genomes from the western hemisphere and broaden our understanding of the complexity of NL63 hCoVs circulating globally.

RESULTS

Clinical epidemiology

In order to determine the yearly and seasonal variations in the prevalence of NL63 virus infection in our paediatric patient population, specimens were analysed over a 3-year period. From 1 January 2009 to 31 December 2011, 92 of 9380 respiratory specimens were found to be positive for NL63 viral RNA by multiplex PCR (see Methods), an overall prevalence of 1%. NL63 viruses were circulating in the paediatric population during all three of the study years (Fig. 1), primarily during the winter months from December to March. However, there was considerable yearly variation in prevalence and the month of peak incidence. During December 2010 and January 2011, there was a much higher incidence of NL63, with 29 and 21 positive specimens, accounting for 10 and 6% of all specimens tested, respectively. Of the 92 respiratory specimens positive for NL63 viral RNA during this study period, 33% were also positive for an additional virus. The most common co-infections were human rhinoviruses (13%), respiratory syncytial virus (8%), adenoviruses (5%) and hCoV-229E (5%).

Of the children positive for NL63 viruses, 42% were under 12 months of age and 56% were under 2 years of age. Seventy-one per cent of the NL63-positive patients were admitted to the hospital and 51% had an underlying medical condition. To characterize the clinical syndromes associated with NL63 viruses, we analysed the primary discharge diagnoses given by the care providers to the patients with NL63-positive respiratory specimens. The most common diagnosis was viral syndrome or upper respiratory tract infection (n=27), fever and neutropenia (n=9), croup (n=9), bronchiolitis (n=8), rule out sepsis/fever in the neonate (n=6) and pneumonia (n=5). The most common symptoms in the NL63-positive patients were fever (73%), cough (41%) and hypoxia (34%). The clinical characteristics of the patients associated with the NL63 genomes that were fully sequenced are presented in Table 1.

Two patients had multiple specimens positive for NL63 viral RNA over an extended period of time. Both of these patients were immunocompromised with underlying oncological diagnoses. The first patient had six NL63-positive respiratory specimens over a 5-month period and...
the second had nine NL63-positive specimens over a 10-month period.

**Genome sequences, phylogenetic analysis and genotypes**

Complete or nearly complete genome sequences were obtained for 19 of the 23 samples submitted for sequencing. All samples submitted for sequencing were positive only for NL63 and no other viruses were detected in these samples. The sequences ranged from 27,456 to 27,495 nt, lacking only short sequences at the 3' and 5' termini (the original NL63 isolate was 27,553 nt) and with some having one or more small gaps within the final pseudomolecule. Poor template quality precluded further sequence analysis of four samples. The nearly complete genome sequences

### Table 1. Clinical characteristics of sequenced Denver NL63 viruses

The genotypes of the 2010 strains and NL63/DEN/2005/291 are based on the S sequence only. ALTE, Acute life-threatening event; CLD, chronic lung disease; CF, cystic fibrosis; F&N, fever and neutropenia; KD, Kawasaki disease; PCKD, polycystic kidney disease; Unk, unknown; VSD, ventricular septal defect.

<table>
<thead>
<tr>
<th>NL63 isolate</th>
<th>NL63 genotype</th>
<th>Sample date</th>
<th>Sex</th>
<th>Admitted</th>
<th>Symptom(s)</th>
<th>Underlying medical condition(s)</th>
<th>Discharge diagnosis</th>
<th>Age (months)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NL63/DEN/2005/193</td>
<td>A</td>
<td>1/11/2005</td>
<td>M</td>
<td>Y</td>
<td>Apnoea and hypoxia</td>
<td>None</td>
<td>Rule out sepsis</td>
<td>1</td>
</tr>
<tr>
<td>NL63/DEN/2005/347</td>
<td>A</td>
<td>2/1/2005</td>
<td>M</td>
<td>N</td>
<td>Fever</td>
<td>None</td>
<td>Viral syndrome</td>
<td>53</td>
</tr>
<tr>
<td>NL63/DEN/2005/449</td>
<td>A</td>
<td>2/9/2005</td>
<td>F</td>
<td>N</td>
<td>Fever, sore throat</td>
<td>None</td>
<td>PCKD Viral syndrome</td>
<td>90</td>
</tr>
<tr>
<td>NL63/DEN/2005/1062</td>
<td>A</td>
<td>4/12/2005</td>
<td>M</td>
<td>Y</td>
<td>Fever, apnoea, sore throat</td>
<td>None</td>
<td>ALTE</td>
<td>7</td>
</tr>
<tr>
<td>NL63/DEN/2005/1876</td>
<td>A</td>
<td>11/21/2005</td>
<td>F</td>
<td>Y</td>
<td>Apnoea, congestion</td>
<td>None</td>
<td>ALTE</td>
<td>5</td>
</tr>
<tr>
<td>NL63/DEN/2005/232</td>
<td>B</td>
<td>1/18/2005</td>
<td>M</td>
<td>N</td>
<td>Fever, seizures, cough, congestion</td>
<td>None</td>
<td>Seizures and viral syndrome</td>
<td>221</td>
</tr>
<tr>
<td>NL63/DEN/2009/14</td>
<td>B</td>
<td>3/1/2009</td>
<td>M</td>
<td>Y</td>
<td>Fever, cough, congestion</td>
<td>None</td>
<td>KD</td>
<td>36</td>
</tr>
<tr>
<td>NL63/DEN/2010/36</td>
<td>B</td>
<td>12/30/2010</td>
<td>M</td>
<td>N</td>
<td>Fever, cough, congestion</td>
<td>None</td>
<td>Bronchiolitis</td>
<td>3</td>
</tr>
<tr>
<td>NL63/DEN/2006/16</td>
<td>C</td>
<td>1/8/2008</td>
<td>F</td>
<td>N</td>
<td>Difficulty breathing</td>
<td>Laryngomalacia</td>
<td>Mucous plugging</td>
<td>97</td>
</tr>
<tr>
<td>NL63/DEN/2009/6</td>
<td>C</td>
<td>2/25/2009</td>
<td>M</td>
<td>Y</td>
<td>Hypoxia, tachypnea</td>
<td>None</td>
<td>Viral syndrome</td>
<td>8</td>
</tr>
<tr>
<td>NL63/DEN/2009/20</td>
<td>C</td>
<td>3/12/2009</td>
<td>F</td>
<td>Y</td>
<td>Fever, cough, congestion</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>NL63/DEN/2010/20</td>
<td>C</td>
<td>12/15/2010</td>
<td>M</td>
<td>Y</td>
<td>Fever, cough, congestion</td>
<td>None</td>
<td>Rule out sepsis</td>
<td>1</td>
</tr>
<tr>
<td>NL63/DEN/2010/35</td>
<td>C</td>
<td>12/14/2010</td>
<td>F</td>
<td>Y</td>
<td>Fever, cough, anorexia</td>
<td>Hypotonia</td>
<td>Viral syndrome</td>
<td>7</td>
</tr>
<tr>
<td>NL63/DEN/2009/31</td>
<td>R</td>
<td>2/21/2009</td>
<td>M</td>
<td>Y</td>
<td>Respiratory distress</td>
<td>Tonsillectomy</td>
<td>Viral syndrome</td>
<td>33</td>
</tr>
</tbody>
</table>
consisted of 2–18 contigs after two rounds of directed closure, and these were listed as ‘draft’. However, only those in two or fewer contigs (n=16) were included in the whole-genome analysis. All full-length genomes assembled into single contigs, except for NL63/DEN/2009/6 and NL63/DEN/2009/31, which remained in two contigs with small intervening gaps. The absence of ambiguous bases in these contigs strongly suggested (but does not prove) that each of the samples sequenced contained one predominant NL63 genome and thus that the two contigs arose from different parts of the same genome. The G+C content of the 16 complete genomes was 34.5 mol%, considerably lower than that of other alpha-CoVs, which range from 39 to 42 mol% G+C (Woo et al., 2010).

The overall genome organization was identical for all 16 NL63 virus specimens and identical to the four known NL63 sequences from the Netherlands, and was similar to that of other alpha-CoVs. The putative core sequence (CS) of the transcription regulatory sequence (TRS) motif 5′-CUAAAC-3′ (Sola et al., 2005; Woo et al., 2009, 2010; Zuniga et al., 2004) was found at the 3′ end of the leader sequence and preceded each of the translated ORFs. The CS upstream of the spike (S) and envelope (E) coding regions, however, was slightly modified to 5′-CUAAUC-3′ and 5′-CUAAUC-3′, respectively. The CSs of the membrane (M) and nucleocapsid (N) genes were found in duplicate, with the terminal C of the first sequence serving as the starting C of the second sequence (5′-CUAAACUAAAC-3′). The CS and TRS sequences in all of the genes were conserved across all the full-length NL63 genomes.

A phylogenetic tree comparing the genome sequences of the 16 Colorado NL63 strains combined with the four full-length NL63 genomes found in the Netherlands is shown in Fig. 2. These 20 NL63 strains fell into three genotypes, designated genotypes A (seven strains), B (six strains) and C (four strains). Genotypes A and B were identified previously in European clinical specimens (Fouchier et al., 2004; Pyrc et al., 2006; van der Hoek et al., 2004). Two of the strains previously sequenced from the Netherlands and one of the Colorado strains formed a fourth group, which probably represents recombinants of the other genotypes (Fig. 2). VISTA analysis using the prototypical NL63 isolate from 2004 as the reference showed that the greatest areas of dissimilarity between the genotypes were from 1.0 to 3.5 kb [1a and non-structural protein 3 (nsp3) genes] and from 21.5 to 23.5 kb (S gene) (Fig. 3 and Figs S1 and S2, available in JGV Online). Fig. 3 clearly illustrates regions of the NL63 genome that had large areas of divergent sequences, shows how these contribute to the genotype clusters (A, B and C) of the NL63 full-genome sequences and illustrates how these markers can be useful for identifying potential recombinants between these genotypes. (The single thin peak near the 5′ end of the NL63/DEN/2009/31 genome and the peak near the beginning of the S gene of NL63/DEN/2009/6 represent areas of poor sequence coverage due to the presence of two contigs for these sequences.) The degree of phylogenetic differences between each ORF of the 20 NL63 genomes is pictured schematically in a heat map (Fig. 4). This provides another way of visualizing recombination along the length of the genomes. It shows a pairwise comparison of the shapes of the phylogenetic trees constructed separately for each coding sequence. Where the pairwise comparisons were highly similar (blue), these genes had the same phylogenetic relationship, and where different (yellow), the corresponding regions of the genomes had different phylogenetic relationships. The data suggested that, during the evolution of NL63 viruses, recombination events have contributed to the diversity of circulating virus strains.

As reported previously for the Netherlands NL63 sequences (Pyrc et al., 2006), we found two signature deletions in our multiple sequence alignments, one in-frame deletion of 15 nt located between nt 3310 and 3350 in gene 1a (nsp3) and a 3 nt deletion located between nt 20798 and 20800 within the S gene (S1 subunit). These signature deletions segregated with the different clades. In Fig. 2, the strains coloured in purple differed from this overall pattern (some had both the 1a and S gene deletions, whilst others had neither), indicating that these markers may have swapped between genomes after strain divergence.

Based on the VISTA analysis, we examined the nucleotide sequences of the nsp3 and S genes for potential recombination.
sites. Looking at representative sequences from each of the genotypes, genotype A (NL63/DEN/2005/1876) and genotype C (NL63/DEN/2009/20) were similar in sequence, with genotype B (NL63/DEN/2009/14) exhibiting a slightly divergent sequence until around nt 1030 of nsp3 (nt 4014 in the whole genome of reference strain NL63/AMS/2004/1). After nt 1030, genotypes B and C had similar sequences, with genotype A having a more divergent sequence. Similarly, in the S gene, genotypes A and C had similar sequences until around nt 920 (nt 21392 in the whole genome of reference strain NL63/AMS/2004/1) where their sequences diverged. From nt 920 to 1710, the three genotypes had relatively divergent sequences. Following nt 1710, genotypes B and C had relatively similar sequences. This segregation of nucleotide sequences was maintained at the amino acid level in nsp3 but only in the first part of the S gene.

Associations between the date of detection, age, sex, type and severity of clinical disease, and the presence of underlying medical conditions among the NL63 genotypes were analysed (Table 1). Included in this analysis were six additional respiratory specimens from the 2010–2011 season for which the entire S gene was sequenced. Genotypes B and C co-circulated in Colorado in 2005, 2009 and 2010, but genotype A was detected only in 2005 when it was the predominant NL63 strain. Unexpectedly, all eight of the patients with genotype B were male. Whether this represents a true association or a statistical anomaly will require sequencing of additional NL63 specimens. Genotype A was more commonly associated with a diagnosis of apnoea or acute life-threatening events than the other NL63 genotypes.

**Bootscanning analysis**

From the clustering identified in Fig. 2, we created four groups, labelled A–C and R. Fig. 5 shows the bootscans for these four groups using the unique Denver genotype C group as reference (actually called query by the Simplot program). Deviations from the mean (spikes up or down) showed regions of the genome where the query genomes were more or less similar to the reference genome.

Using all of the available NL63 genome sequences for analysis, the ratios of synonymous ($K_a$) to non-synonymous ($K_s$) amino acids for the various coding regions were calculated (Table 2). The highest $K_a:K_s$ ratio in the NL63 genomes was observed in the S gene (0.272), followed by the nsp3 (0.250) and nsp2 (0.194) genes, strongly indicating that these gene products are under the strongest selection pressure.

**Comparative sequence analysis of the S gene among NL63 viruses**

Because the S gene was identified as the most variable region of the genome, we analysed more closely the S gene from our NL63-positive samples and sequenced this gene
20 NL63 viruses. Gene tree topologies were compared in a Journal of General Virology 2392 circulating in Colorado, USA. A few 'unusual strains' three distinct genotypes of NL63 viruses (A, B and C) genome sequences from clinical respiratory specimens in This is the first study to describe 16 complete NL63 were eight potential acid differences (15 %) (Fig. 6). Within the NTD, there for the majority of these specimens, five of the ten different NTD sequences were most similar to genotype A. Among these ten different NTD sequences, there were 29 amino acid differences (15 %) (Fig. 6). Within the NTD, there were eight potential N-linked glycosylation sites (NXS or NXT), five of which were conserved in all NL63 strains. NL63 viruses containing each of the ten NTD sequence variants have been detected at different times throughout the world in the Netherlands, Belgium, Sweden, Hong Kong and the USA (Denver) (Chiu et al., 2005; Fouchier et al., 2004; Koetz et al., 2006; Leung et al., 2009; Moës et al., 2005; Pyrc et al., 2006; van der Hoek et al., 2004).

from six additional Colorado clinical specimens for further comparison. As reported previously, the greatest variability in the S gene was found in the N-terminal domain (NTD) of the S gene (nt 1–600, aa 1–200). We found ten different amino acid sequences in the NTD of our NL63 sequences and the 75 NL63 NTD S sequences in GenBank (Fig. S1). Although sequences outside the S gene were not available for the majority of these specimens, five of the ten different NTD sequences were most similar to genotype A. Among these ten different NTD sequences, there were 29 amino acid differences (15 %) (Fig. 6). Within the NTD, there were eight potential N-linked glycosylation sites (NXS or NXT), five of which were conserved in all NL63 strains. NL63 viruses containing each of the ten NTD sequence variants have been detected at different times throughout the world in the Netherlands, Belgium, Sweden, Hong Kong and the USA (Denver) (Chiu et al., 2005; Fouchier et al., 2004; Koetz et al., 2006; Leung et al., 2009; Moës et al., 2005; Pyrc et al., 2006; van der Hoek et al., 2004).

**DISCUSSION**

This is the first study to describe 16 complete NL63 genome sequences from clinical respiratory specimens in North America. We found that there have been at least three distinct genotypes of NL63 viruses (A, B and C) circulating in Colorado, USA. A few 'unusual strains' (NL63/DEN/2009/31, NL63/ROT/2004/1 and NL63/AMS/2004/057) displayed incongruent phylogenetic positions, with some regions more like genotype B and others more like genotype C. These strains probably represent recombinitants between genotypes.

In agreement with other studies (Arden et al., 2005; Bastien et al., 2005a, b; Chiu et al., 2005; Dominguez et al., 2009; Ebihara et al., 2005; Esper et al., 2005; Kuypers et al., 2007; Prill et al., 2012; Suzuki et al., 2005; Vabret et al., 2005; van der Hoek et al., 2005), only a small percentage (1 %) of all of the samples submitted for respiratory testing to the clinical virology laboratory during our 3-year study period were positive for NL63 RNA. The prevalence of NL63 infection, however, varied markedly from year to year, with a peak/mini-outbreak during the winter of 2010–2011 when 10 % of respiratory specimens submitted during December and January were positive for NL63 RNA. Our previous study also detected a high prevalence of NL63 viruses circulating during the 2004–2005 respiratory season, with 4 % of all samples positive for NL63 viruses during the months of January–March (Dominguez et al., 2009). These data are in agreement with other multi-year studies and support the notion that the prevalence of NL63, like other human CoVs, varies markedly from year to year (Dare et al., 2007; Gaunt et al., 2010; Gerna et al., 2007; Lau et al., 2006; Talbot et al., 2009; Vabret et al., 2008). Similarly, serological studies have shown yearly variations in the prevalence of hCoV-299E and hCoV-OC43 infection (Monto & Lim, 1974). Interestingly, it was shown recently that infection with one hCoV may elicit cross-protective immunity from subsequent infection with a different CoV in the same antigenic group (Dijkman et al., 2012).

We examined the epidemiology and clinical presentations associated with the different genotypes of NL63 viruses in Colorado. Genotype A was identified only in clinical samples from 2005 and was the predominant NL63 strain circulating in Colorado in that year. Although detected in all of the years studied, genotypes B and C were the predominant NL63 strains circulating in 2009 and 2010. These data suggest that there is annual variation in the prevalence of NL63 and suggest antigenic variation similar to that seen for other respiratory viruses, most notably influenza. We did not find a clear association of clinical presentation with particular genotypes of NL63, although genotype A was more commonly detected in patients with apnoea or acute life-threatening events.

It is unclear how long CoV shedding can persist in the human respiratory tract following natural infection. We previously detected HCoV-229E RNA in respiratory specimens from an immunocompromised child for 11 weeks (Dominguez et al., 2007). Here, we have reported prolonged shedding of NL63 for at least 5 and 10 months in two immunocompromised patients, suggesting that persistent infection of immunocompromised patients may provide a reservoir for adaptation and/or further spread of NL63.
CoV transcription proceeds via a discontinuous RNA synthesis step guided by the TRS. The TRS contains a CS within a hairpin structure essential for efficient production of subgenomic mRNAs (Dufour et al., 2011; Sola et al., 2005; Zúñiga et al., 2004). The CSs are conserved among all unique sequences of NL63 genomes and within all subgenomic RNAs, except for the E and S genes. The CSs of the E and S genes may alter the RNA secondary structures, affecting the production of subgenomic mRNAs. This might explain why the levels of expression of subgenomic mRNAs encoding S and E are lower than expression of ORF3 (Pyrc et al., 2004). In contrast, the M and N genes are preceded by tandem duplicates of CSs, which could affect the production of subgenomic mRNAs. As the M protein is essential for virus budding and the N protein is necessary for packaging viral RNA, this most likely represents a mechanism for increased production.

Due to their discontinuous RNA replication and their very long (~30 kb) RNA genomes, in vivo recombination occurs in many CoVs and can lead to the generation of new recombinant strains or genotypes, and may alter their virulence, immunogenicity and viral host range. For example, feline coronavirus (FCoV) type II apparently originated from recombination between FCoV type I strains and canine CoVs (Herrewegh et al., 1998). Similarly, recombination between genomes of different bat CoVs has resulted in the emergence of novel SARS-related bat and civet CoVs (Lau et al., 2010). Through whole-genome sequencing of 22 hCoV-HKU1 specimens,
Researchers in Hong Kong detected a third, novel genotype that had probably arisen from recombination of the two other known strains (Woo et al., 2006). Analysis of hCoV-OC43 genomes has suggested that a novel genotype of hCoV-OC43 has arisen through natural recombination and has become the dominant circulating strain (Lau et al., 2011). A study of NL63 genomes in the Netherlands provided the first evidence of recombination for NL63 (Pyrc et al., 2006). The present study provides an additional example of natural recombination between NL63 genomes and illustrates that recombination utilized by CoVs occurs in humans and could result in the emergence of novel strains and/or rapid dissemination of antigenic variants that escape immune recognition or have other selective advantages.

The most variable region in the NL63 genome is the S gene, particularly the NTD (nt 1–600). Comparison of the nucleotide and amino acid sequences of the NL63 S NTDs in our study with the NL63 sequences in GenBank confirmed the extensive amino acid sequence variability in NL63 strains circulating throughout the world and at different times. The functions of the NTD and the reasons for its variability are not yet known. One possibility is that the NTDs may exhibit antigenic variation, providing a mechanism to evade neutralization by host immune responses whilst maintaining the more highly conserved, receptor-binding C-domain of the S protein. This mechanism is utilized by other respiratory viruses including respiratory syncytial virus and influenza virus (Plotkin et al., 2002; Smith et al., 2004; Sullender, 2000). The presence of multiple potential glycosylation sites in the NTD, with the deletion of up to three of them in certain NL63 samples, supports this possibility. The NL63 NTD may also play a role in virus infection in several ways: it may aid in virus attachment by recognizing cell-surface carbohydrates or sialic acids, it may facilitate binding of the C-domain to the protein receptor human angiotensin-converting enzyme 2 (hACE2) or it may play a critical role in entry into cells by influencing a conformation change required for fusion of the viral envelope with the cell membrane. These possibilities are supported by studies showing that, although the NTD was not essential for binding to hACE2 or for hACE2-dependent membrane fusion of lentiviruses pseudotyped with NL63 S protein, deletion of the NTD markedly reduced both binding and membrane fusion (Hofmann et al., 2006). The core of the NTD of the CoV mouse hepatitis virus S protein has the same β-sandwich fold as human galectins (S-lectins), which are sugar-

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>L. PLAS CFCTN</td>
<td>L. SPANCDLGL VALUESQSTIV</td>
<td>I. I. I.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>30</td>
<td>40</td>
<td>50</td>
<td>60</td>
<td>70</td>
<td>80</td>
<td>90</td>
<td>100</td>
</tr>
<tr>
<td>N. RRAFALHST</td>
<td>YIDANQIVIV</td>
<td>VTRNEIGLHAS</td>
<td>VTLKICFNSR</td>
<td>NTTFDFLSNA</td>
<td>SSSYDCVYV</td>
<td>N. I. I.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>130</td>
<td>140</td>
<td>150</td>
<td>160</td>
<td>170</td>
<td>180</td>
<td>190</td>
<td>200</td>
<td>210</td>
<td>220</td>
</tr>
</tbody>
</table>

**Fig. 6.** Amino acid sequence alignment of the NTDs of the S proteins (aa 10–177 shown) of representatives of all published NL63 S genes. In addition to the NTDs from the full genomes shown in Figs 2 and 3, NTDs from partial NL63 sequences from GenBank are shown (black). Potential N-glycosylation sites are underlined.
binding proteins that modulate immune responses. Most alpha-, beta- and gamma-CoV S glycoproteins contain NTDs, and their biological significance is an important topic for future investigation.

The results presented here highlight the utility of genome sequencing of viruses in clinical specimens from different parts of the world. We found that the same NL63 genotypes (A and B) circulate in different parts of the world. We also detected a new NL63 genotype (C) and have provided further evidence for circulation of naturally occurring NL63 recombinants, which has important evolutionary implications. In our paediatric population, the predominant genotype of NL63 varied from year to year, suggesting yearly antigenic variation. Sequence analysis demonstrated high sequence diversity in the NTD of the S gene. Future structural and functional studies are needed to elucidate the role this domain may play in infection, immunity, virus entry and pathogenesis of NL63 viruses.

**METHODS**

**Clinical specimens.** Beginning in January 2009, the Clinical Virology Laboratory of the Children’s Hospital Colorado began using a multiplex PCR [Tag Respiratory Virus Panel (RVP); Lumex Molecular Diagnostics] in combination with direct fluorescent antibody (DFA) assays to identify pathogens in respiratory specimens (nasopharyngeal washes, tracheal aspirates and bronchoalveolar lavages) from children with respiratory symptoms. Clinicians had the option of testing specimens by DFA or RVP only, DFA followed by RVP if the DFA testing was negative, or DFA and RVP concurrently. Nucleic acids were extracted from specimens submitted for RVP testing using Virus Minikits version 2.0 on BioRobot EZ1 extractors (Qiagen), following the manufacturer’s instructions, and tested by RVP. This assay detects 12 types of respiratory viruses (including influenza A and B viruses, parainfluenza viruses 1–4, adenoviruses, respiratory syncytial virus A and B, human metapneumovirus, rhinoviruses/enteroviruses and hCoV-229E, hCoV-OC43, hCoV-HKU1 and hCoV-NL63). RVP is not cleared by the US Food and Drug Administration for routine reporting of hCoV results, but the ability of this test to identify hCoVs has been reported (Mahony et al., 2007; Wong et al., 2008) and was confirmed by our clinical laboratory prior to this study. As part of an ongoing investigation of the aetiology of paediatric viral respiratory infections, all clinical respiratory specimens positive for any hCoV were archived at −70°C for further analysis. Medical chart review of NL63-positive specimens was performed using a standardized form. Use of the banked specimens and clinical data for this study was approved by the Colorado Multiple Institutional Review Board.

Subsets of NL63-positive clinical specimens collected from 2009 to 2011, as well as samples collected in 2005 from our previously reported study (Dominguez et al., 2009), were selected for genome sequencing as follows. First, specimens identified initially as positive for NL63 RNA by RVP were confirmed using a modified quantitative real-time RT-PCR (qRT-PCR) assay that detected RNAs from all four non-SARS hCoVs (Dominguez et al., 2007; Kuypers et al., 2007). Specimens positive in the consensus CoV qRT-PCR assay were then analysed using NL63-specific conventional RT-PCR assays using SuperScript III reverse transcriptase (Invitrogen) with random primers, followed by PCR. Primer sets were used that bound regions unique to the NL63 CoV S and N genes (Bastien et al., 2005b; Moes et al., 2005),

**RT-PCR and full-genome sequencing.** A 96-well plate of degenerate NL63-specific primers was designed from a consensus sequence of the four NL63 reference genomes from GenBank (accession nos DQ445912, DQ445911, NC_005831 and AY567487) using a PCR primer design pipeline developed at the J. Craig Venter Institute (JCVI) (Li et al., 2008). This produced tiled amplicons with an optimal length of 550 bp, with a 100 bp overlap and at least twofold amplicon coverage at every base. An M13 sequence tag added to the 5′ end of each primer was used for sequencing.

RNA extracted from the clinical samples was subjected to RT-PCR using a Qiagen One-step kit to produce cDNA amplicons from each of the primer pairs in the 96-well plate. For each clinical specimen, two independent sets of reactions were performed to produce sequence coverage consistent with JCVI standards for finished Sanger sequences. PCR products were treated with a mixture of exonuclease and shrimp alkaline phosphatase to remove primers and dNTPs, and then subjected to Sanger sequencing using the common M13 primers at the 5′ end of each primer.

Sequencing reads were trimmed to remove ampiclon primer-linker sequences as well as low-quality sequences, and assembled using minimus, part of the open-source AMOS project (http://sourceforge.net/apps/mediawiki/amos/index.php?title=AMOS). Initial assemblies were inspected using an in-house application, Cloe (Closure Editor, http://cloe.sourceforge.net), and directed PCR-based sequencing reactions were conducted to improve the sequence. Assembled sequences for each viral sample were run through the VIGOR annotation pipeline. VIGOR (Viral Genome ORF Reader, http://www.jcvi.org/vigor; Wang et al., 2010) was developed at JCVI to decode sequences from many classes of virus by taking into account virus-specific features such as alternative splicing, internal ORFs and ribosomal slippage, and was used to validate the newly assembled NL63 genomes during the finishing process. After sequences were annotated and validated by VIGOR, the gene predictions were subjected to manual inspection and quality control before loading into the JCVI annotation database and submission to GenBank. To predict mature peptides, annotated polypeptides and ORFs were searched by BLAST against a reference set of known CoV mature peptides. The edges of each BLAST hit were scored by proximity to other BLAST hits and the quality of the hits. The hits with the highest scoring edges were used to cover the polypeptide. A final step resolved gaps and overlaps by using known cleavage motifs in the polypeptide whilst maximizing the homology of each predicted mature peptide to the reference mature peptide (Djikeng et al., 2008; Li et al., 2008; Wang et al., 2010).

**Sequencing methods for the 2010 amplicons.** For the samples collected in 2010, only the region spanning the S gene was sequenced. Extracted RNA was reverse transcribed in an oligo(dT)-primed reaction using SuperScript III (Invitrogen). An ampiclon of ~5.6 kb covering the S region was generated by PCR using Accuprime (Invitrogen) and the following primer pair: forward, 5′-TTGCA-GTCTGTGATATGGAAGTG3′, and reverse, 5′-ACCAAAACTATGGTCCAAA-3′. Each ampiclon was gel purified and then simultaneously amplified and bar-coded in two separate reactions using a modified sequence-independent single-primer amplification (SISPA) approach (Djikeng et al., 2008). A pool of these and other viral samples were used to construct an Illumina library and sequenced on a HiSeq platform. After sequencing, reads from each sample were deconvoluted by bar code and trimmed for quality and to remove the SISPA hexamer primer and barcode sequences. The Illumina reads were then assembled de novo using the clc_novo_assemble program (CLC Bio). The resulting de novo contigs were then used to pick a best reference, which was then used for a mapping assembly using the clc_ref_assemble_long program (CLC Bio). The assembled amplicons were annotated using VIGOR as described above.
Sequence analysis. The viral genome sequences were multiply aligned with MUSCLE version 3.7 (Edgar, 2004), using the sum-of-pairs profile score (VTML240) option over the course of five iterations. From this alignment, a distance-based neighbour-joining phylogenetic tree was constructed using the Kimura two-parameter distance (Kimura, 1980), with gamma-distributed rates across sites ($\gamma$=0.04) with the PHYLIP package version 3.69 (Felsenstein, 1989). Bootstrap analysis of the genome alignments was performed using SEQBOOT and CONSENSE with 1000 pseudo-replicates. To analyse possible recombination events, bootstrap analyses were undertaken on the alignment using Simplot version 3.5.1 (Lole et al., 1999). Bootstrapped phylogenetic trees were built for overlapping window segments of 200 bp. The mean bootstrap value across all members of the query group segment versus the remaining three groups was plotted versus window position. For comparison of ORFs across the genomes, a heat map was constructed. Coding sequences of each viral gene were multiply aligned with MUSCLE using a maximum of 30 iterations and the diagonal expansion option. Maximum-parsimony trees for each coding sequence were constructed with the PHYLIP utility DNAPARS. The collection of most parsimonious trees for a given coding sequence was then consolidated into a single consensus tree using PHYLIP CONSENSE. The resulting gene tree topologies were compared in a pairwise fashion using the Robinson–Foulds topological distance, $D_{RF}(a,b)$, with PHYLIP DNADIST. During calculation of $D_{RF}(a,b)$, in cases where one or more samples were missing a coding sequence in topology a, the missing genes were temporarily pruned from the tree topology b to achieve the same coding sequence in both topologies (using RETREE). The distance matrix $D_{RF}$ was visualized as a heat map using the $R$ statistical package.

Nucleotide sequence accession numbers. The clinical specimens positive for NL63 viruses were named according to the following nomenclature: virus/location/year of collection/specimen number (e.g. NL63/DEN/2009/6).

ACKNOWLEDGEMENTS

This project was funded in part with Federal funds from the National Institute of Allergy and Infectious Disease, the National Institutes of Health, and the Department of Health and Human Services under contract number HHSN272200900007C and by an NIH grant to S. R. D (1-K08AI073525). We thank Jonathan Motley, Christina Osborne, Andrew Berglund, the staff of the Children’s Hospital Colorado Clinical Virology Laboratory and members of the JCVI viral sequencing and informatics groups for their assistance with these studies, and Zhaohui Qian for critical review of this manuscript.

REFERENCES


Genome analysis of 16 Colorado NL63 coronaviruses


http://vir.sgmjournals.org


