



## Limited Inter- and Intra-patient Sequence Diversity of the Genetic Lineage A human metapneumovirus fusion gene

THILDE NORDMANN WINTHER,<sup>1,2</sup> CHRIS D. MADSEN,<sup>2,4</sup> ANDERS G. PEDERSEN,<sup>3</sup>  
MARIE-LOUISE VON LINSTOW,<sup>1</sup> JESPER EUGEN-OLSEN<sup>2</sup> & BIRTHE HOGH<sup>1,\*</sup>

<sup>1</sup>Department of Paediatrics, Hvidovre Hospital, University of Copenhagen, Denmark

<sup>2</sup>Clinical Research Unit, Hvidovre Hospital, University of Copenhagen, Denmark

<sup>3</sup>Center for Biological Sequence Analysis, Technical University of Denmark, Denmark

<sup>4</sup>IFOM, The FIRC Institute of Molecular Oncology, Milan, Italy

Received December 6, 2004; Accepted February 6, 2005

**Abstract.** *Human metapneumovirus* (hMPV) is associated with respiratory tract illness especially in young children. Two hMPV genetic lineages, A and B, and four sublineages A1, A2 and B1, B2 have been defined. Infection with hMPV occurs through membrane fusion mediated by the hMPV fusion (F) protein. In this study, the inter- and intra-patient genetic diversity of the lineage A hMPV F gene was investigated. Ten isolates were collected from 10 hMPV infected children. Viral RNA was isolated and amplified, and approximately 10 clones from each isolate were sequenced. Altogether 108 clones were successfully sequenced. The average interpatient sequence diversity was 1.68% and 1.64% at nucleotide and amino acid levels, respectively. The samples were divisible into two groups on the basis of intrapatient sequence diversity. In group 1 (4 children) the intra-patient sequence diversity was low (nt: 0.26–0.39%, aa: 0.51–0.94%) whereas group 2 (6 children) had a higher intra-patient sequence diversity (nt: 0.85–1.98%, aa: 1.08–2.22%). Phylogenetic analyses showed that the group 1 children harboured sublineage A1 only, but interestingly group 2 children harboured both sublineages A1 and A2, indicating they had been infected with at least two viruses. Several independent viruses contained premature stop codons in exactly identical positions resulting in truncated fusion proteins. Possibly this is a mechanism for immune system evasion. The F protein is a major antigenic determinant, and the limited sequence diversity observed lay emphasis on the hMPV F gene as a putative target for future vaccine development.

**Key words:** fusion gene, hMPV, phylogenetic analysis, sequencing

### Introduction

*Human metapneumovirus* (hMPV) is a newly discovered virus associated with respiratory tract illness primarily in young children. hMPV is a member of the family Paramyxoviridae, subfamily *Pneumovirinae*, and genus *Metapneumovirus* [1,2]. Two main hMPV genetic lineages, A and B, have been defined, each of which has been divided into two sublineages, A1, A2, and B1, B2 [3,4,5,6]. The

two genetic lineages are highly related antigenically and are not distinct serotypically [5,7].

At present, it is unknown whether a patient is infected with only one hMPV sublineage or a mixed virus population, an intrinsic feature of RNA viruses [8].

Paramyxoviridae initiates infection by attaching to cell surface receptors allowing fusion of the viral membrane with the host cell plasma membrane. Two transmembrane glycoproteins, the attachment (G) protein and the fusion (F) protein, direct the infection process. The G protein serves to dock the target and attach membranes whereas the

\*Author for all correspondence:  
E-mail: birthe.hoegh@hh.hosp.dk

F protein directly mediates membrane fusion [9]. The F protein is a major antigenic determinant that mediates extensive cross-lineage neutralization and protection [5, 10]. The G protein also serves as an important antigenic determinant although it is shown to be highly variable and therefore not interesting as a target of hMPV vaccine development [4,10,11].

The hMPV fusion gene comprises 1620 base pairs, encoding a protein of 539 amino acid (aa) residues. The F protein is synthesized as a biologically inactive precursor, F0, which has to be cleaved by host cell proteases into F1 and F2 for fusion activity.

The F protein consists of an ectodomain including an amino-terminal signal sequence, a hydrophobic transmembrane domain and a cytoplasmic domain. The hMPV fusion peptide is located at the N-terminus of the F1 subunit; it is 23 aa residues in length, highly hydrophobic and thought to be inserted into the target membrane to initiate fusion. Several heptad repeats (HR) domains are also present in the protein. HR1 is 42 aa residues in length and located just carboxyl terminal to the fusion peptide. HR2 is located in the ectodomain adjacent to the transmembrane region. HR1 and HR2 are important for fusion in the infection process [6,9,11].

The aim of this study was to characterize the inter- and intra-patient sequence diversity and relatedness of the genetic lineage A hMPV F gene in isolates from infected children hospitalized in Copenhagen. We focused on the fusion peptide and the HR1, as these domains are highly conserved in the *Paramyxoviridae* family, and putative targets for vaccine development.

## Materials

Specimens comprised 11 hMPV positive nasopharyngeal aspirates (numbered 1–11) collected from children, admitted to the Department of Paediatrics, Hvidovre Hospital and Amager Hospital, Copenhagen, Denmark, with acute respiratory illnesses during November to May 1999–2000 and 2001–2002. The specimens were obtained at admission. The 11 specimens were identified as hMPV positive by LightCycler reverse transcription-polymerase chain reaction (LC-RT-PCR)

based on the nucleocapsid (N) protein gene [12]. All the identified hMPV isolates were recovered from children less than 6 months of age. None of the study children were hospitalized simultaneously at the same ward. The history of the individual virus isolates and the epidemiological and clinical data are described in detail previously [12]. The Ethics Committee of Frederiksberg, Copenhagen, Denmark approved the study (jr. nr. 01–028/03).

## Methods

### *hMPV viral RNA extraction and cDNA synthesis*

RNA extraction was carried out using the Roche Magnapure (Roche) as previously described [12]. Viral RNA was reverse transcribed using random hexamer primers (TAG Copenhagen, Denmark) and SuperScript II enzyme (Invitrogen) following the manufacturer's instructions. Briefly, 5 µl of purified RNA was incubated with 5 µM random hexamer, 0.5 mM dNTP, 1× First-Strand buffer, 1 µl RNAGuard, 10 mM DTT and brought to 20 µl with distilled water.

### *Primers*

Primers were designed based on the published hMPV sequence AY145294 (AY145294 is 100% identical with AF371337, the prototype virus for lineage A1 [6]). We used the program, primer3 input (<http://frodo.wi.mit.edu/cgi-bin/primer3/primer3.cgi/>) to design a primer set amplifying a part of the F gene, including the fusion peptide and the HR1. Sense primer: 5'-CCG ATG GAC CCA GCT TAA TA-3' (nucleotide 180–201), and antisense primer: 5'-GCT GGT GTT ATT CCA GCG TT-3' (nucleotide 627–645). Only the hMPV genetic lineage A was amplified with these primers, as alignment of the 4 predicted hMPV prototype sequences A1, A2, B1, and B2 (Genbank accession number AF371337, AY304360, AY304361 and AY304362) revealed a mismatch between the prototype B1/B2 and the 3' end of the sense primer [6].

### *PCR*

After cDNA synthesis PCR was performed using 5 µl cDNA, 0.4 µM sense and antisense primer,

0.2 mM dNTP, 1 U Easy-A High-Fidelity PCR Cloning Enzyme (Stratagene) and 1 Easy-A Reaction buffer brought to 50  $\mu$ l with distilled water. The amplification conditions comprised one initial denaturation step at 95°C for 2 min, followed by 45 cycles of 95°C for 30 s (denaturation), 55°C for 30 s (annealing) and 72°C for 40 s (extension) and finished with one final extension step at 72°C for 8 min. Negative controls were included in each run.

#### *Purification and TA cloning*

PCR amplification products were subjected to 2% agarose gel electrophoresis. Bands were excised using a clean scalpel, and DNA was purified with "Quantum Prep Freeze N Squeeze DNA Gel Extraction Spin Columns" (BIORAD) according to the manufacturer's instructions. The purified PCR-products were further amplified to ensure addition of 3'A-overhangs. The PCR conditions were identical to those applied in the first round except for only 20 cycles were used. Negative controls were included in each run.

#### *Cloning*

Purified PCR products were cloned into pCR 4-TOPO vectors by using TOPO TA cloning Kit for Sequencing (Invitrogen) and transformed into One Shot TOP-10 F'competent *E. coli* (Invitrogen). Plasmid DNA from approximately 10 clones per patient isolate was purified according to QIAprep Miniprep Kit Protocol (QIAGEN) (numbered with a patient number from 1 to 11 followed by a clone number from 1 to 28).

#### *Sequencing*

The purified clones were sequenced in both directions using M13 primers and BigDye terminator cycle sequencing Kit version 3.1 (Applied Biosystems) and run on an ABIprism 3100 Genetic Analyser (Applied Biosystems).

#### *Alignments*

Sequence alignments were made using the Clustal software package available in the software pack-

age of Bioedit version 5.0.6 (<http://jwbrown.mbio.ncsu.edu/Bioedit/bioedit.html>).

#### *Sequence diversity in a population of viral sequences*

The sequence diversity across the entire length of the partial hMPV F gene sequences was quantified using Nei's nucleotide diversity  $\Pi$  [13] (the standard error was estimated according to equation 10.7 therein). This measure gives the average percentage of different sites between all possible pairwise comparisons of the viral sequences. For instance, a nucleotide diversity of 0.02 in a population of DNA sequences means that on average any pair of sequences will be different at 2% of their sites.

#### *Sequence diversity in a single position of a gene or protein alignment*

The sequence diversity at single positions in an alignment (DNA or protein) was quantified using the Shannon entropy measure [14,15]:

$$H = - \sum_i p_i \log(p_i)$$

This measure is zero in cases where the same nucleotide or as is present in all the aligned sequences (no variability = a perfectly conserved base/aa). Conversely, Shannon entropy reaches its maximal value when all possible symbols are equally frequent (e.g., a base/aa alignment with 25% of each of the four nucleotides at a given position).

#### *Construction of phylogenetic trees*

Phylogenetic trees were constructed using a Bayesian approach as implemented in the program MrBayes (version 3.0B4) [16]. We constructed a tree based on 101 partial hMPV F gene sequences obtained from 10 patient isolates and 4 predicted prototype sequences A1, A2, B1 and B2 [6]. Another phylogenetic analysis was carried out on the sequences from patients 1 and 6, plus the prototype sequences A1 and A2. For both trees we used the general time reversible model of nucleotide substitution combined with individual rates for each of the three codon positions (the

“GTR + ss” model). Specifically, we ran the Metropolis-coupled Markov chain Monte Carlo (MCMCMC) process for 10,000,000 generations (with 4 chains) while saving a tree once every 100 generations. Proper mixing and sufficient run length were ensured using the estimated sample size-feature available in the program Tracer (Rambaut, A. & Drummond, A., [evolve.zoo.ox.ac.uk/software.html](http://evolve.zoo.ox.ac.uk/software.html)), which was also used to determine burn-in. From each set of selected trees a majority rule consensus tree was constructed and used for further analysis.

#### *Analyses of positive and negative selection*

To determine whether any part of the investigated hMPV F gene sequences had been under positive selective pressure, the MO and M3 codon models [17, 18] implemented in the program codeml from the package PALM [19], were employed. Analysis of positive and negative selection was performed using the hMPV F gene sequences and the trees constructed above.

The ratio between the non-synonymous (amino-acid replacing) and synonymous (silent) nucleotide substitution rates in a piece of coding DNA (the “dN/dS ratio”), provides useful information about the selective pressure that has been acting on the gene. Thus, dN/dS ratios < 1 indicate that amino acid changing mutations are selected against, presumably because of their disruptive effect on protein function. This is often referred to as

“negative selection”. Conversely, dN/dS ratios > 1 indicate that amino acid changing mutations are selected for, and therefore appear to be beneficial for the organism. This is called “positive selection”. It is often found that while a gene is under negative selection on average, isolated parts or even single codons of it are experiencing positive selection.

## Results

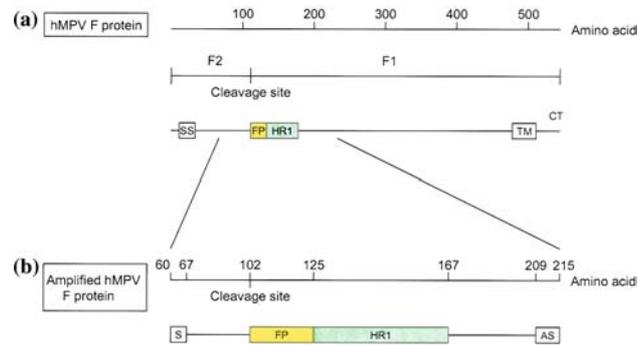
### *hMPV F gene nucleotide sequence analysis*

We investigated 11 nasopharyngeal aspirates collected from 11 children infected with hMPV. One sample (patient number 8) was excluded from the analyses, since we could not detect hMPV by the PCR system applied in this study. Data of the included patients are summarized in Table 1. Viral RNA from 10 children was reverse transcribed, amplified and cloned. A total of 108 hMPV clones were successfully sequenced (median 11 clones per child, range 8–14). The sequences were named by a patient number (1–11, excluding number 8) followed by a clone number (1–28). The amplified hMPV F gene segments were 465 bp in length (nucleotide 180–645 relative to the hMPV F gene translation start site [11]). The primer sequences were expelled, leaving fragments of 426 bp (142 aa) for analysis (nucleotide 201–627, amino acid 67–209) (Fig. 1).

*Table 1.* Epidemiological and clinical data of four children infected with hMPV sublineage A1 only, and six children infected with hMPV sublineage A1 and A2

Patient number	hMPV sublineage	Age Sex	Age (months)	Duration of hospitalization (days)	Reported fever	Nasal discharge	Cough	Tachypnoea	Chest indrawing	Rhonchi
1	A1	M	6,0	2	+	+	+	+	+	+
2	A1,A2	M	1,8	2	-	+	+	+	+	-
3	A1,A2	F	3,4	3	+	+	+	+	+	+
4	A1	F	3,5	6	+	+	+	+	+	-
5	A1,A2	F	2,3	3	+	+	+	+	-	-
6	A1,A2	M	2,4	9	+	-	+	+	+	+
7	A1,A2	F	3,5	11	+	-	+	-	-	+
9	A1	M	4,8	5	+	-	-	+	+	+
10	A1	M	6,0	3	+	+	+	+	-	+
11	A1,A2	M	4,6	2	+	+	+	+	+	+

Age: Age at admission.



*Fig. 1.* (a) Diagram of the primary sequence of the hMPV F protein. Important domains in the primary sequence of the hMPV F protein are diagrammed. The top line indicates the amino acid positions while the second line identifies the polypeptides, F1 and F2, generated by proteolytic cleavage. The yellow domain is the fusion peptide (FP) and the turquoise domain is heptad repeat 1 (HR1). Other domains are SS, signal sequence; TM, transmembrane domain and CT, cytoplasmic domain. (b) Diagram of the amplified hMPV F gene segment, amino acid 60–215 relative to the hMPV F gene translation start site [11]. Important domains in the studied hMPV F gene fragment are diagrammed. The top line indicates amino acid positions. The cleavage site is marked. The yellow domain is the fusion peptide (FP) and the turquoise domain is heptad repeat 1 (HR1). Other domains are S, sense primer: 5'-CCG ATG GAC CCA GCT TAA TA-3' and AS, antisense primer: 5'-GCT GGT GTT ATT CCA GCG TT-3'.

#### *Inter- and intra-patient nucleotide sequence diversity*

One hundred and eight hMPV nucleotide sequences obtained from 10 children were compared and revealed 60 haplotypes. The nucleotide diversity across the entire length of the hMPV F gene sequences was measured. The inter-patient nucleotide sequence diversity was 0.0168, meaning that, on average, any pair of sequences was different at 1.68% (SD  $\pm$  0.02) of their sites. The intra-patient nucleotide sequence diversity was ranging from 0.26% to 1.98%. All samples were then divided into two groups; (1) low intra-patient sequence diversity and (2) high intra-patient sequence diversity. Thus, isolates from child number 1, 4, 9 and 10 (group 1) revealed an intra-patient nucleotide sequence diversity ranging from 0.26% (SD  $\pm$  0.086) to 0.39% (SD  $\pm$  0.067), whereas the isolates from child number 2, 3, 5, 6, 7 and 11 (group 2) revealed an intra-patient nucleotide sequence diversity ranging from 0.85% (SD  $\pm$  0.31) to 1.98% (SD  $\pm$  0.23).

#### *Predicted hMPV F proteins*

The hMPV F gene nucleotide sequences were translated into amino acid sequences. The predicted polypeptide was 142 residues in length. It was observed, that seven of the obtained sequences

only encoded a fragment of 105 residues; one due to a single nucleotide substitution at position 518 (clone 6–7, G at position 518), that resulted in a premature termination codon, and six, all from the same patient, due to a frame shift mutation at position 504 (clones 3–1, 3–2, 3–4, 3–6, 3–7 and 3–9: G at position 504) (Table 2). Among the seven clones with stop mutations, five haplotypes were present. The seven sequences were excluded from further analysis.

The remaining 101 sequences contained 44 haplotypes and the average inter-patient amino acid sequence diversity was 1.64% (SD  $\pm$  0.14). The intra-patient amino acid sequence diversity ranged from 0.51% to 2.22%. The isolates were divided into the same two groups as for the nucleotide sequences. Group 1 isolates revealed an intra-patient amino acid sequence diversity ranging from 0.51% (SD  $\pm$  0.17) to 0.94% (SD  $\pm$  0.15), whereas isolates from group 2 revealed an intra-patient amino acid sequence diversity ranging from 1.08% (SD  $\pm$  0.45) to 2.22% (SD  $\pm$  0.37).

#### *Phylogenetic analyses of hMPV F gene nucleotide sequence*

Phylogenetic analyses of our 101 partial hMPV F gene sequences and the 4 predicted prototype sequences A1, A2, B1 and B2 [6] (data not shown),



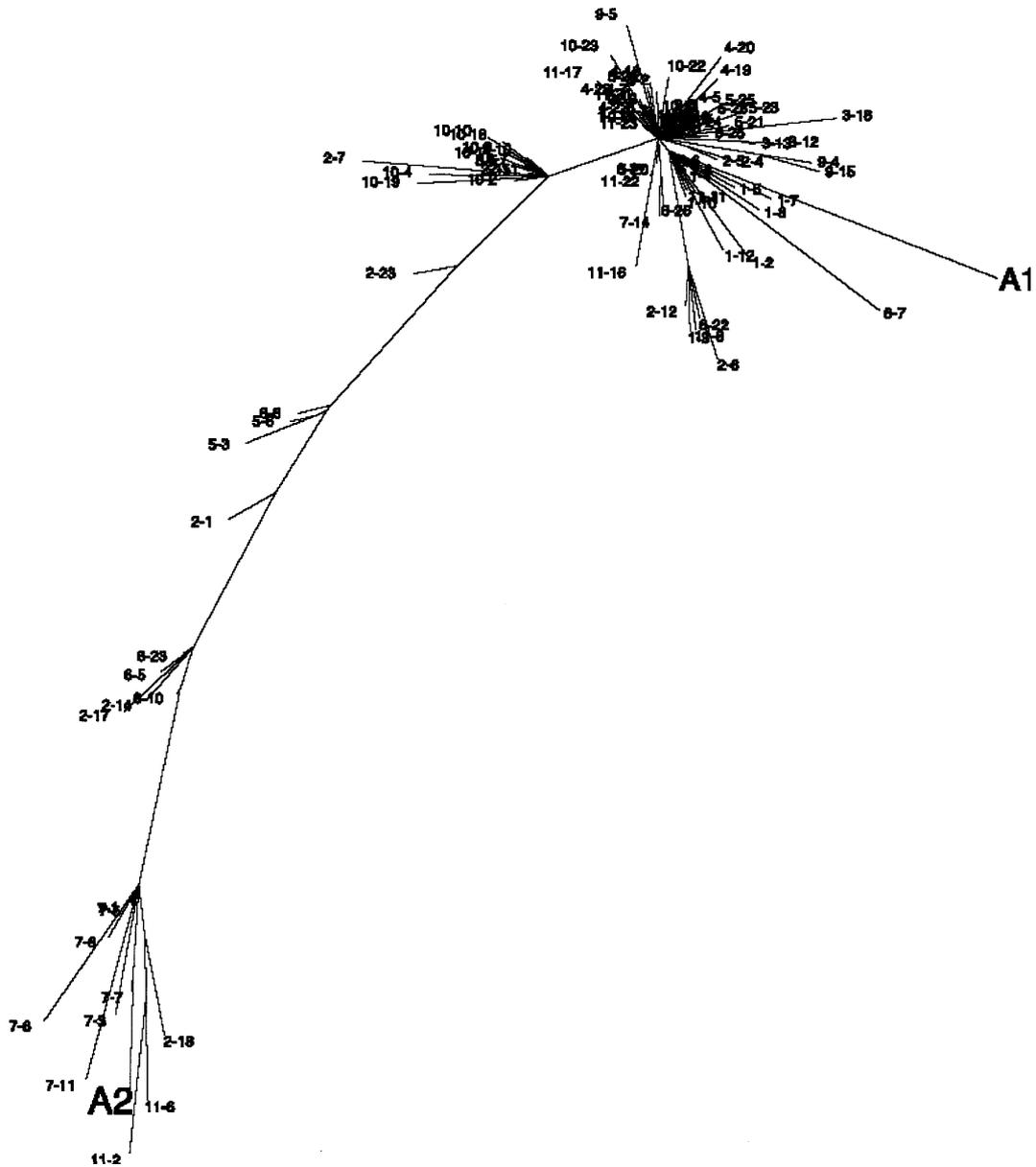


Fig. 2. Unrooted phylogenetic tree constructed on the basis of 101 partial hMPV F gene sequences and the predicted prototype sequences A1 and A2. There was a major cluster containing the predicted prototype A1 and a minor cluster consisting the predicted prototype A2. Predicted prototype sequence A1 has GenBank accession number AF371337; Predicted prototype sequence A2 has GenBank accession number AY304360. The phylogenetic tree was constructed using a Bayesian approach as implemented in the program MrBayes (version 3.0B4) [16].

and B1, B2. Members of the two different sublineages within each main lineage were 4–6% different at nt level and 1–3% at aa level. Comparison of sequences from the same sublineage revealed even less diversity; nucleotide: 0–3% and amino acid: 0–1% [3,6,11]. The results of our study are in agree-

ment with previous published results, although we report a lower diversity between members of sublineage A1 and A2 (1.68% and 1.64% at nt and aa level, respectively compared to 4–6% and 1–3% at nt and aa level, respectively). This difference could be explained by the preponderance of A1 sequences

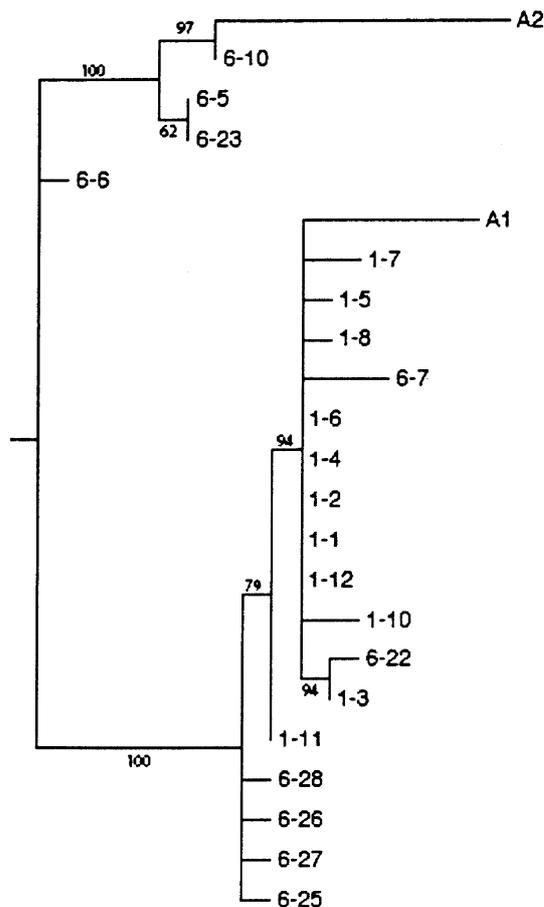


Fig. 3. Phylogenetic analysis of the hMPV F gene sequences obtained from child number 1 and 6, they represented group 1 (low intra-patience sequence diversity) and 2 (high intra-patient sequence diversity), respectively. The A1 sequences obtained from patient number 6 were more closely related to the A1 sequences obtained from patient number 1, than the A2 sequences obtained from patient number 6. The phylogenetic tree was constructed using a Bayesian approach as implemented in the program MrBayes (version 3.0B4) [16]. Bootstrap values are based on the consensus tree and relevant numbers are shown in the tree.

(82 A1 sequences, 16 A2 sequences) in the present study.

The analyzed hMPV fragment implied two striking domains; the fusion peptide (nucleotide 106–174) and HR1 (nucleotide 175–300). However, these two domains did not reveal any specific distribution of mutations. Overall our findings are consistent with the F gene of *Paramyxoviridae* having structural and functional constraints for mutations [10].

The hMPV F gene segments amplified and studied generally encoded polypeptides predicted to be 142 aa residues long, but 7 sequences encoded truncated fusion proteins of 105 aa residues. All 7 sequences had stop codons at exactly identical positions resulting in truncated forms lacking both transmembrane and cytoplasmic domains. Intriguingly, the stop codons were caused by two independent mechanisms: six sequences had single-nucleotide deletions in codon 101 that resulted in frame-shifting and premature termination at codon 106, but in one sequence the stop codon was due to a nucleotide substitution directly in codon number 106.

The use of a PCR polymerase with proof reading activity (error rate  $1.3 \times 10^{-6}$ ), the application of bi-directional sequencing, and the identification of identically positioned stop codons in 7 different sequences (five different haplotypes) obtained from two independent children, strongly indicates that the introduction of the stop codon is not due to experimental artefacts. It has previously been described that frame shift mutations in human respiratory syncytial virus generates variant attachment proteins that are resistant to monoclonal antibody neutralization [20, 21]. It is difficult to envisage the biological significance of a F protein fragment including only the first 172 a residues and it is likely to be very unstable in the infected cell. Anyway, our results raise the intriguing possibility that, rather than significantly varying the amino acid composition of the F protein, certain sublineages of hMPV may alter F protein expression as a means of partially evading immune system detection.

A phylogenetic analysis was made to determine, whether each child was infected with one hMPV sublineage or a mixed virus population. It turned out, that four of the 10 investigated children were infected with sublineage A1 only, whereas the remaining six children were infected with both sublineage A1 and A2. Nosocomial coinfection is not an opportunity since the study children were not hospitalized simultaneously at the same ward. These data show, that individuals within the same geographic area can be infected with more than one hMPV sublineage. This is in agreement with previous findings of nucleotide heterogeneity of the SH and G genes detected within one sample [10].

In conclusion, the inter- and intra-patient sequence diversity of the genetic lineage A hMPV F gene was found to be very limited in this population. However, further detailed studies are necessary for determination of the real sequence diversity of this particular gene of hMPV in general.

### Acknowledgments

Thilde Nordmann Winther was recipient of a scholarship from University of Copenhagen, Copenhagen Hospital Corporation. The Danish Lung Association and Dagmar Marshalls Foundation supported this study. No conflicts of interest.

### References

1. van den Hoogen B.G., de Jong J.C., Groen J., Kuiken T., de Groot R., Fouchier R.A., and Osterhaus A.D., *Nat Med* 7, 719–724, 2001.
2. van den Hoogen B.G., Bestebroer T.M., Osterhaus A.D., and Fouchier R.A., *Virology* 295, 119–132, 2002.
3. Boivin G., Mackay I., Sloots T.P., Madhi S., Freymuth F., Wolf D., Shemer-Avni Y., Ludewick H., Gray G.C., and LeBlanc E., *Emerg Infect Dis* 10, 1154–1157, 2004.
4. Peret T.C., Abed Y., Anderson L.J., Erdman D.D., and Boivin G., *J Gen Virol* 85, 679–686, 2004.
5. Skiadopoulos M.H., Biacchesi S., Buchholz U.J., Riggs J.M., Surman S.R., Amaro-Carambot E., McAuliffe J.M., Elkins W.R., St. Claire M., Collins P.L., and Murphy B.R., *J Virol* 78, 6927–6937, 2004.
6. van den Hoogen B.G., Herfst S., Sprong L., Cane P.A., Forleo-Neto E., de Swart R.L., Osterhaus A.D., and Fouchier R.A., *Emerg Infect Dis* 10, 658–666, 2004.
7. MacPhail M., Schickli J.H., Tang R.S., Kaur J., Robinson C., Fouchier R.A., Osterhaus A.D., Spaete R.R., and Haller A.A., *J Gen Virol* 85, 1655–1663, 2004.
8. Elena S.F., Miralles R., Cuevas J.M., Turner P.E., and Moya A., *ICBMB Life* 49, 5–9, 2000.
9. Morrison T.G., *Biochim Biophys Acta* 1614, 73–84, 2003.
10. Biacchesi S., Skiadopoulos M.H., Boivin G., Hanson C.T., Murphy B.R., Collins P.L., and Buchholz U.J., *Virology* 315, 1–9, 2003.
11. Bastien N., Normand S., Taylor T., Ward D., Peret T.C., Boivin G., Anderson L.J., and Li Y., *Virus Res* 93, 51–62, 2003.
12. von Linstow M-L., Larsen H.H., Eugen-Olsen J., Koch A., Winther T.N., Meyer A.M., Westh H., Lundgren B., Melbye M., and Heigh B., *Scand J Infect Dis* 36, 578–584, 2004.
13. Nei M., Columbia University Press, New York, 1987.
14. Schneider T.D., Stormo G.D., Gold L., and Ehrenfeucht A., *J Mol Biol* 188, 415–431, 1986.
15. Shannon C.E., *Bell System Tech J* 27, 379–423, 1948.
16. Ronquist F. and Huelsenbeck J.P., *Bioinformatics* 19, 1572–1574, 2003.
17. Nielsen R. and Yang Z., *Genetics* 148, 929–936, 1998.
18. Yang Z., Nielsen R., Goldman N., and Pedersen A.M., *Genetics* 155, 431–449, 2000.
19. Yang Z., *Comput Appl Biosci* 13, 555–556, 1997.
20. Garcia-Barreno B., Portela A., Delgado T., Lopez J.A., and Melero J.A., *EMBO J* 9, 4181–4187, 1990.
21. Sullender W.M., Mufson M.A., Anderson L.J., and Wertz G.W., *J Virol* 65, 5425–2534, 1991.