Intratransient Variation of the Respiratory Syncytial Virus Attachment Protein Gene

Charles N. Agoti,1 Jean L. Mbisa,2 Ann Bett,1 Graham F. Medley,3 D. James Nokes,1,3 and Patricia A. Cane2*

Kenya Medical Research Institute, Centre for Geographic Medicine Research-Coast, Kilifi, Kenya1; Health Protection Agency, Centre for Infections, Colindale, London, United Kingdom2; and Department of Biological Sciences, University of Warwick, Coventry, United Kingdom3

Received 2 June 2010/Accepted 12 July 2010

Intratransient variability of the attachment (G) protein gene of respiratory syncytial virus (RSV) was examined using both population and single-genome sequencing. Samples from three patients infected with a group B virus variant which has a 60-nucleotide duplication in the G protein gene were examined. These samples were chosen because occasional mixed sequence bases were observed. In a minority of RSV genomes from these patients considerable variability was found, including point mutations, insertions, and deletions. Of particular note, the deletion of the exact portion of the gene which had been duplicated in some isolates was observed in viral RNAs from two patients.

Human respiratory syncytial virus (RSV) is a major cause of lower respiratory tract infection in infants and vulnerable adults (3, 9) and is unusual in that it can repeatedly reinfect individuals (5, 6). RSV isolates are classified into two groups, A and B, and the attachment (G) protein, a target for neutralizing antibodies, is the most variable of the viral proteins, showing considerable genetic and antigenic variability both within and between the groups (7, 8). The G protein is able to accommodate drastic changes, which have been observed both in culture during the selection of monoclonal antibody escape variants (4, 12, 13) and in vivo with the emergence of new variants, including a group B strain with a duplication of 60 nucleotides (17). This strain with a 60-nucleotide duplication was first reported from Buenos Aires in 1999 (17) and then was subsequently detected in samples from 1998 in Madrid (16). The strain then became the dominant group B strain worldwide, indicating a selective advantage for this variant (16, 18). Thus, major genetic changes can be introduced into the G gene sequence while the virus replicates in its natural host, which can then be selected under favorable epidemiological conditions.

Previous investigations of the genetic diversity of RSV exploited direct sequencing of PCR-amplified products (2), which represent the population average of the in vivo variants. Such sequences are derived from multiple copies of cDNA and represent the dominant sequence, and they thus do not allow detection of minority populations below about 20% prevalence (11). Information on intratransient viral diversity during infections may therefore be missed, knowledge of which could be important in the overall understanding of the genetic diversity of this virus. We report here the analysis of individual RSV RNA molecules derived by single-genome amplification (SGA) and sequencing from clinical samples using a methodology developed for the analysis of HIV genomes (11, 14).

RSV-positive samples were collected from infants admitted to Kilifi District Hospital, Kenya (10). Viral RNA extraction and cDNA synthesis were carried out as previously described (15). For population sequencing, a nested PCR was carried out on the cDNA using primers that amplified the ectodomain-coding part of the G protein gene, with the PCR product being directly sequenced. In the 2007-2008 RSV epidemic in Kilifi, group B viruses were predominant. By population sequencing of ~100 group B samples, all were found to have the 60-nucleotide duplication observed in the Buenos Aires variant (data not shown). However, in some specimens there were some mixed bases at some positions, so the variability at the level of the single cDNA molecule was further investigated.

Three samples that gave occasional mixed signals in the sequence chromatograms were further analyzed by SGA and sequencing. For SGA the cDNAs were serially diluted 3-fold up to 1:6,361. Ten nested PCRs were carried out on each dilution using Platinum high-fidelity PCR SuperMix (Invitrogen) (containing Taq polymerase together with the proofreading enzyme Pyrococcus species GB-D polymerase). Based on the Poisson distribution, it has been shown that for a sample dilution yielding approximately 30% positive PCRs there is an 80% likelihood that each PCR is derived from a single cDNA molecule (11). For each of the identified endpoint dilutions, the cDNA was amplified in 80 separate nested PCRs using the high-fidelity enzyme and the positive reaction products sequenced. The nomenclature for the sequences reported in this paper is place of isolation (Kenya [Ken])/year of isolation/strain number. For SGA sequences an additional Roman number is given.

The predicted length derived by population sequencing of the G proteins of the three samples examined by SGA was 310 amino acids, showing a 6-nucleotide deletion and a changed stop codon relative to the Buenos Aires strain (Fig. 1). The dominant sequences represented 60 to 88% of the sequences.
derived by SGA. The differences were due to point mutations, duplications, and deletions, as summarized in Table 1; the consequences of these changes for the predicted length of the G protein are shown in Table 2.

For sample 1 (Ken/07/80395), 17 separate PCR products from SGA were sequenced. Sequence Ken/07/80395/i was the dominant sequence, occurring in 15/17 (88%) of the products. A synonymous mutation (T-C at 264) was seen in Ken/07/80395/ii, while Ken/07/80395/iii had an 8-nucleotide duplication starting at position 599, which resulted in a frameshift with a subsequent premature stop codon.

For sample 2 (Ken/08/80900), 12 separate SGA PCR products were sequenced, with the dominant sequence (Ken/08/80900) and a minority sequence (Ken/08/80900/ii), with the sequences of prototype group B strain CH18537 (accession number M17213) and Buenos Aires strain BA/3833/99B (accession number AY333362). This shows the duplication of 60 nucleotides in the Kenyan and Buenos Aires viruses relative to CH18537 and the loss of the same 60 nucleotides in the Kenyan minority sequence. Termination codons are underlined.

FIG. 1. Nucleotide sequence alignment of part of the G protein gene (from nucleotide 400) of the sample 2 population sequence (Ken/08/80900) and a minority sequence (Ken/08/80900/ii), with the sequences of prototype group B strain CH18537 (accession number M17213) and Buenos Aires strain BA/3833/99B (accession number AY333362). This shows the duplication of 60 nucleotides in the Kenyan and Buenos Aires viruses relative to CH18537 and the loss of the same 60 nucleotides in the Kenyan minority sequence. Termination codons are underlined.

<table>
<thead>
<tr>
<th>Strain</th>
<th>Total no. of SGA-derived sequences</th>
<th>No. of:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Variants</td>
</tr>
<tr>
<td>Ken/07/80395</td>
<td>17</td>
<td>3</td>
</tr>
<tr>
<td>Ken/08/80900</td>
<td>12</td>
<td>3</td>
</tr>
<tr>
<td>Ken/08/80767</td>
<td>15</td>
<td>6</td>
</tr>
</tbody>
</table>

* The insertion involved an 8-nucleotide duplication.
* The deletion was of 60 nucleotides.
* The deletions ranged from 1 to 104 nucleotides, including one of 60 nucleotides.
tested during the period in question had the duplication, and duplication can be excluded because first, all the samples 60-nucleotide deletion in the same position as the original low (0.011%) and without major deletions and insertions (11). The error rate in SGA analysis of HIV-1 has been found to be high-fidelity enzyme mix was used in these experiments, and the possibility of artifacts should be considered. However, a nonsynonymous mutation was seen at position 539 (T-C), causing amino acid change 1175T.

The samples examined by SGA in this study were originally selected because of mixed bases observed by population sequencing, so the results reported here may not be applicable to the community immunity. Such strains have been largely replaced in recent years, may be length (i.e., without the 20-amino-acid insertion), which, as described by Trento et al. (17). Such deletions provide the potential for emergence of viruses with the “normal” protein structure of the viral RNA sequence that is duplicated, as described by Henderson, F. W., A. M. Collier, W. A. Clyde, Jr., and F. W. Denny. (6).

The results reported here confirm and extend the observations that the RSV G gene is highly prone to errors during replication, including within the patient, and so provides a pool of variants that can be subject to selection during transmission. The detection of minority genomes that showed a 60-nucleotide deletion at the exact position as the original duplication observed in the Buenos Aires strains was a surprising finding. The mechanisms for such a deletion may lie in the stem-loop structure of the viral RNA sequence that is duplicated, as described by Trento et al. (17). Such deletions provide the potential for emergence of viruses with the “normal” protein length (i.e., without the 20-amino-acid insertion), which, as such strains have been largely replaced in recent years, may be able to overcome the community immunity.

### Nucleotide sequence accession numbers

The GenBank accession numbers for the sequences determined in this study are GU811691 to GU811706.

We thank the Kilifi RSV project staff for collecting the samples analyzed here, and we also thank the study participants.

This study is published with permission of the Director of KEMRI. This work was supported by a Wellcome Trust Masters Training Fellowship to C.N.A. (grant no. 085882) and program grant no. 084663.

### REFERENCES


<table>
<thead>
<tr>
<th>Strain</th>
<th>Sequence no.</th>
<th>Mutation type (nucleotide position)</th>
<th>Effect on predicted polypeptide</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ken/07/80395</td>
<td>Ken/07/80395/ii</td>
<td>Substitution (264)</td>
<td>None</td>
</tr>
<tr>
<td>Ken/07/80395</td>
<td>Ken/07/80395/iii</td>
<td>Duplication (599–606)</td>
<td>Frameshift with a subsequent premature stop codon</td>
</tr>
<tr>
<td>Ken/08/80900</td>
<td>Ken/08/80900/ii</td>
<td>Deletion (792–851)</td>
<td>Reduction of length by 20 amino acids</td>
</tr>
<tr>
<td>Ken/08/80767</td>
<td>Ken/08/80767/ii</td>
<td>Substitution (592)</td>
<td>Isoleucine-to-threonine change</td>
</tr>
<tr>
<td>Ken/08/80767</td>
<td>Ken/08/80767/iii</td>
<td>Substitution (233)</td>
<td>Threonine-to-serine change</td>
</tr>
<tr>
<td>Ken/08/80767</td>
<td>Ken/08/80767/iv</td>
<td>Substitution (664)</td>
<td>Immediate premature stop codon</td>
</tr>
<tr>
<td>Ken/08/80767</td>
<td>Ken/08/80767/v</td>
<td>Deletion (599)</td>
<td>Frameshift with a subsequent premature stop codon</td>
</tr>
<tr>
<td>Ken/08/80767</td>
<td>Ken/08/80767/vi</td>
<td>Deletion (599)</td>
<td>No effect due to earlier frameshift; otherwise this would cause a 20-amino-acid deletion</td>
</tr>
<tr>
<td>Ken/08/80767</td>
<td>Ken/08/80767/vii</td>
<td>Deletion (836–939)</td>
<td>Frameshift, with no subsequent stop codon within the region sequenced</td>
</tr>
<tr>
<td>Ken/07/80767</td>
<td>Ken/07/80767/vi</td>
<td>Substitution (664)</td>
<td>Immediate premature stop codon</td>
</tr>
<tr>
<td>Ken/08/80767</td>
<td>Ken/08/80767/vii</td>
<td>Deletion (836–939)</td>
<td>Frameshift, with no subsequent stop codon within the region sequenced</td>
</tr>
</tbody>
</table>


