Direct whole-genome deep-sequencing of human respiratory syncytial virus A and B from Vietnamese children identifies distinct patterns of inter- and intra-host evolution

Lien Anh Ha Do,† Andreas Wilm,† H. Rogier van Doorn,† Ha Minh Lam,1 Shuzhen Sim,3 Rashmi Sukumaran,3 Anh Tuan Tran,4 Bach Hue Nguyen,4 Thi Thu Loan Tran,5 Quynh Huong Tran,5 Quoc Bao Vo,5 Nguyen Anh Tran Duc,5 Hong Nhien Trinh,4 Thi Thanh Hai Nguyen,4 Bao Tinh Le Binh,4 Khanh Le,4 Minh Tien Nguyen,4 Quang Tung Thai,4 Thanh Vu Vo,4 Ngoc Quang Minh Ngo,4 Thi Kim Huyen Dang,5 Ngoc Huong Cao,5 Thu Van Tran,5 Lu Viet Ho,5 Jeremy Farrar,† Menno de Jong,†,‡,§ Swaine Chen,3 Niranjan Nagarajan,3 Juliet E. Bryant†,‡ and Martin L. Hibberd3

1Oxford University Clinical Research Unit, Wellcome Trust Major Overseas Program, Ho Chi Minh City, Vietnam
2Nuffield Department of Clinical Medicine, University of Oxford, Oxford, UK
3Genome Institute of Singapore, Genome Building, 138672 Singapore
4Children’s Hospital 1, Ward 10, District 10, Ho Chi Minh City, Vietnam
5Children’s Hospital 2, Ben Nghe Ward, District 1, Ho Chi Minh City, Vietnam
6Department of Medical Microbiology, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands

Human respiratory syncytial virus (RSV) is the major cause of lower respiratory tract infections in children <2 years of age. Little is known about RSV intra-host genetic diversity over the course of infection or about the immune pressures that drive RSV molecular evolution. We performed whole-genome deep-sequencing on 53 RSV-positive samples (37 RSV subgroup A and 16 RSV subgroup B) collected from the upper airways of hospitalized children in southern Vietnam over two consecutive seasons. RSV A NA1 and RSV B BA9 were the predominant genotypes found in our samples, consistent with other reports on global RSV circulation during the same period. For both RSV A and B, the M gene was the most conserved, confirming its potential as a target for novel therapeutics. The G gene was the most variable and was the only gene under detectable positive selection. Further, positively selected sites in G were found in close proximity to and in some cases overlapped with predicted glycosylation motifs, suggesting that selection on amino acid glycosylation may drive viral genetic diversity. We further identified hotspots and coldspots of intra-host genetic diversity in the RSV genome, some of which may highlight previously unknown regions of functional importance.

INTRODUCTION

Human respiratory syncytial virus (RSV) is the most frequently detected virus amongst young children hospitalized with acute respiratory infection worldwide (Lamb et al., 2005), with no effective therapy or approved vaccine currently available. RSV (family Paramyxoviridae) is an enveloped, single-stranded, negative-sense RNA virus...
with a ~15.2 kb genome that encodes 11 viral proteins. Two genetically and serologically distinct RSV subgroups, A and B (Coates et al., 1966), co-circulate with varying frequency. Ten RSV A genotypes and 19 RSV B genotypes have been recognized worldwide (Blanc et al., 2005; Dapat et al., 2010; Peret et al., 1998, 2000; Shobugawa et al., 2009; Trento et al., 2010; Venter et al., 2001); genotyping is based on the highly variable G (glycoprotein) gene, which encodes one of two principal surface antigens.

Epidemiological studies have reported periodic shifts in the predominance of RSV A and B (Baek et al., 2012; Dapat et al., 2010). This pattern is thought to be driven by the dynamics of population immunity, where short-lived, subtype-specific herd immunity (predominantly directed against the G protein) over one or two seasons favours dissemination of the alternate subtype in a subsequent season (Botosso et al., 2009). Repeated infections with the same RSV strain within individuals and co-circulation of multiple genotypes suggest the lack of an effective long-term host immune response, and consequently the lack of strong selective pressures, such as those that induce yearly antigenic drift and sequential lineage replacement in influenza virus (Power, 2008). This suggests that RSV is able to evade or modulate the host immune response and several viral proteins have indeed been reported to be involved in this. The highly variable, extensively glycosylated G protein has multiple immune modulation functions (Johnson et al., 1987b, Power, 2008; Roca et al., 2001). The other major surface antigen, the F (fusion) protein, has been shown to block proliferation of peripheral blood lymphocytes (Power, 2008), whilst the NS1 and NS2 nucleocapsid proteins are known to suppress the IFN response (Spann et al., 2004).

Little is known about RSV intra-host genetic diversity over the course of infection or about the immune pressures that drive RSV molecular evolution. A recent case study that examined RSV intra-host genetic variation within a chronically infected infant with severe combined immune deficiency syndrome before and after bone marrow transplantation reported increased diversity, mostly within the G protein, after engraftment, suggesting that adaptive immunity plays an important role in driving viral diversity (Grad et al., 2014).

Many epidemiological questions remain that would benefit from this type of analysis, if done on a larger scale. For example, whilst similar RSV genotypes circulate simultaneously in different geographical regions (Sullender, 2000), it is unclear if specific genetic signatures for each region, and hence region-specific differences in herd immunity, exist. Currently, most published whole-genome RSV sequences are consensus sequences of passaged isolates from the USA and Europe, the majority of which are RSV A (Collins et al., 1987; Connors et al., 1995; Crowe et al., 1996; Firestone et al., 1996; Karron et al., 1997; Lo et al., 2005; Mink et al., 1991; Rebuffo-Scheer et al., 2011; Stec et al., 1991; Tan et al., 2012; Tolley et al., 1996; Whitehead et al., 1998).

Vietnam is a high-burden country for infant respiratory infection, morbidity and mortality, with RSV as the leading cause amongst hospitalized children (Do et al., 2011; Singh, 2005; Yoshida et al., 2010). Here, we used whole-genome, next-generation Illumina sequencing and a rigorous variant-calling algorithm to characterize RSV inter- and intra-host genetic diversity in clinical samples collected during two consecutive seasons from otherwise healthy children hospitalized in Ho Chi Minh City. This dataset places RSV in Vietnam into the context of global RSV epidemiology and provides important insights into the immune pressures that drive genetic diversity in RSV populations.

RESULTS

RSV in Vietnam in the context of global RSV epidemiology

The global C-terminal G gene phylogeny of 165 and 74 representative sequences indicated that the 37 Vietnamese RSV A sequences formed two distinct genotypes: NA1 (n=34) and GA5 (n=3), with bootstrap values of 99 and 85 %, respectively (Fig. 1a). All RSV A sequences from 2009 were NA1 genotype, whereas NA1 and GA5 co-circulated in 2010. Of the 16 Vietnamese RSV B sequences, 13 belonged to the BA genotypes BA9 (n=12) and BA10 (n=1). Of the remaining three sequences, two (VN-125-7/09 and VN-156-9/09) were most closely related to a BA3 reference strain, but were not classified as BA3 due to low bootstrap values, and one (VN-731-10/09) did not cluster with any reported genotype (Fig. 1b). In general, the Vietnamese RSV genomes fit into the current classification system.

We next reconstructed whole-genome phylogenies (data available upon request) to examine the relationship of Vietnamese RSV to published full-length sequences, which at the time of analysis comprised 63 (50 RSV A and 13 RSV B) sequences from the USA and Europe (Collins et al., 1987; Connors et al., 1995; Crowe et al., 1996; Firestone et al., 1996; Karron et al., 1997; Kumaria et al., 2011; Lo et al., 2005; Mink et al., 1991; Rebuffo-Scheer et al., 2011; Stec et al., 1991; Tolley et al., 1996; Whitehead et al., 1998) collected between 1956 and 2010. Of these, 27 had been directly sequenced from clinical samples (Kumaria et al., 2011; Rebuffo-Scheer et al., 2011). The 34 NA1 Vietnamese RSV A whole-genome sequences from 2009 and 2010 formed the same five distinct clusters as on the C-terminal G gene phylogeny. These five clusters grouped together with four genomes (GenBank accession numbers JF920052, JF920053, JF920054 and JF920055) collected in 2010 in Wisconsin, USA, which were previously assigned to genotype GA2 (not NA1), although they were clearly divergent from other GA2 strains (Rebuffo-Scheer et al., 2011). The remaining three GA5 Vietnamese RSV A sequences (VN-360, VN-391 and VN-773) were from the 2010 season.
and were related to a 2007 GA5 strain collected in the USA (GenBank accession numbers JF920058) (Rebuffo-Scheer et al., 2011). Of the 16 Vietnamese RSV B genomes, 13 were most closely related to BA strains collected from 2007 to 2010 in Wisconsin (Rebuffo-Scheer et al., 2011). Unlike RSV A, we did not observe clustering by year for RSV B.

**RSV inter-host (consensus) genetic diversity**

Analysis of inter-host consensus differences revealed the highest overall substitution rate in G and the lowest in M, for both RSV A and B. Using a maximum-likelihood method, evidence of positive selection was only found in the G gene (Table 1), although G and M2-2 showed the highest dN/dS ratios in both subgroups (Table 2).

We next identified individual G gene codons under selection. Amongst RSV A genomes, a total of 18 possible positively selected sites (6 % of the G gene) were observed (dN/dS 95 %), with seven of 18 of these sites reaching significance (P> 95%). Three of these 18 have not been described before as positively selected: aa 214 (P=82 %), 244 (P=70 %) and 250 (P=99 %; all highlighted in green in Fig. 2a). All potentially positively selected sites from
RSV A were located in the C-terminal second hypervariable region (HVR2) of the G gene (Fig. 2a). Amongst RSV B genomes, five potentially positively selected sites (2% of the G gene) were found ($d_N/d_S = 6.67$, $P_{\text{w}} = 50\%$). Two of these have not been described before: aa 159 ($P = 99\%$) and (aa 78, $P = 88\%$; highlighted in green in Fig. 2b). Consistent with an earlier report (Parveen et al., 2006), many serine and threonine residues in HVR2 (34 and 42 for RSV A and B, respectively) were predicted to be $\Omega$-glycosylated and three sites were predicted to be $\Omega$-glycosylated (Fig. 2). We hypothesize that host immune pressure could result in selection on amino acid glycosylation, which is well known to influence the antigenicity of the G protein. In support of this, nine of 18 positively selected RSV A sites (aa 226, 244, 250, 258, 274, 279, 289, 290 and 297) were found next to $\Omega$-glycosylation motifs; in addition, aa 289 was both positively selected and predicted to be $\Omega$-glycosylated in all NA1 sequences, and aa 250 was positively selected and predicted to be $\Omega$-glycosylated in all GA5 sequences (Fig. 2a). In contrast, none of the RSV B potentially positively selected sites were predicted to be $\Omega$- or $\Omega$-glycosylated (Fig. 2b).

Amongst Vietnamese RSV A sequences, predicted $\Omega$-glycosylation patterns in the G protein HVR2 varied by genotype (Fig. 2a). Potential $\Omega$-glycosylation sites at aa 237 and 250 were present in GA5, but not NA1 sequences, and a third site at aa 294 was observed in all three GA5 strains, but only in four of 37 NA1 strains (Fig. 2a). We similarly found three polymorphic $\Omega$-glycosylation sites amongst the RSV B sequences, two of which (aa 230 and 296) were conserved across genotypes. A substitution was found in the third site (aa 310) in BA3 strains (Fig. 2b). No specific $\Omega$-glycosylation patterns were observed amongst different genotypes for either RSV subgroup.

No distinctive patterns between RSV A and B were found in non-coding regions (NCRs) or intergenic non-coding regions (IGSs) (data available upon request). 5'-NCRs were generally shorter and more conserved than 3'-NCRs, and IGSs were more variable than either NCR. Whilst gene start sequences were well conserved and identical for both RSV A and B, gene end sequences showed much higher variation (Table 3).

### RSV intra-host genetic diversity

Low-frequency intra-host single nucleotide variants (SNVs) in the viral genome were identified using the

<table>
<thead>
<tr>
<th>Gene</th>
<th>RSV A (79 sequences)</th>
<th>RSV B (25 sequences)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Length (bp)</td>
<td>No. of unique sequences</td>
</tr>
<tr>
<td>NS1</td>
<td>417</td>
<td>32</td>
</tr>
<tr>
<td>NS2</td>
<td>372</td>
<td>26</td>
</tr>
<tr>
<td>N</td>
<td>1173</td>
<td>53</td>
</tr>
<tr>
<td>P</td>
<td>723</td>
<td>45</td>
</tr>
<tr>
<td>M</td>
<td>768</td>
<td>39</td>
</tr>
<tr>
<td>SH</td>
<td>192</td>
<td>25</td>
</tr>
<tr>
<td>G</td>
<td>891</td>
<td>57</td>
</tr>
<tr>
<td>F</td>
<td>1722</td>
<td>62</td>
</tr>
<tr>
<td>M2-1</td>
<td>582</td>
<td>38</td>
</tr>
<tr>
<td>M2-2</td>
<td>264</td>
<td>29</td>
</tr>
<tr>
<td>L</td>
<td>6495</td>
<td>74</td>
</tr>
<tr>
<td></td>
<td>1–168</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1–570</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1–543</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1–741</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1600–2805</td>
<td>0.39852</td>
</tr>
<tr>
<td></td>
<td>3925–5574</td>
<td>1</td>
</tr>
</tbody>
</table>

*Recombination breakpoints (RB) predicted by SBP and GARD.
†Likelihood ratio test (LRT) is significant if <0.05.
sensitive, quality-aware variant caller LoFreq (Wilm et al., 2012). It was found that 22% of SNVs were present as the dominant variant in other samples. SNVs were further aggregated to identify regions of the genome with an excess (hotspot) or depletion (coldspot) of intra-host diversity (Wilm et al., 2012) (Fig. 3a, b). Despite sharing some key features, RSV A and B exhibited distinct patterns of intra-host diversity.

Table 2. Rates of synonymous (dS) and non-synonymous (dN) mutations in Vietnamese RSV A and B

<table>
<thead>
<tr>
<th>Gene</th>
<th>RSV A</th>
<th>RSV B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean dS</td>
<td>Mean dN</td>
</tr>
<tr>
<td>NS1</td>
<td>0.11</td>
<td>0.003</td>
</tr>
<tr>
<td>NS2</td>
<td>0.13</td>
<td>0.003</td>
</tr>
<tr>
<td>N</td>
<td>0.11</td>
<td>0.002</td>
</tr>
<tr>
<td>P</td>
<td>0.09</td>
<td>0.003</td>
</tr>
<tr>
<td>M</td>
<td>0.10</td>
<td>0.0005</td>
</tr>
<tr>
<td>SH</td>
<td>0.13</td>
<td>0.004</td>
</tr>
<tr>
<td>G</td>
<td>0.10</td>
<td>0.04</td>
</tr>
<tr>
<td>F</td>
<td>0.11</td>
<td>0.006</td>
</tr>
<tr>
<td>M2-1</td>
<td>0.10</td>
<td>0.0061</td>
</tr>
<tr>
<td>M2-2</td>
<td>0.05</td>
<td>0.03</td>
</tr>
<tr>
<td>L</td>
<td>0.07</td>
<td>0.003</td>
</tr>
</tbody>
</table>

*Substitutions per site was based on maximum-likelihood phylogeny trees of each coding gene.

Fig. 2. Predicted positively selected and glycosylated sites in the C-terminal HVR2 of the G glycoprotein from Vietnamese (a) RSV A and (b) RSV B sequences. Hash symbols (top row) indicate positions identified as positively selected (purple, previously identified positions; green, newly detected positions). Potential N-glycosylation sites, e.g. NXT motifs, where X is not a proline, are underlined; asparagines strongly predicted to be N-glycosylated are highlighted in yellow (default threshold of >0.5 and supported by nine networks), whilst potentially predicted asparagine sites are highlighted in blue (default threshold of >0.5, but not clearly supported by networks); potential O-glycosylation sites (KPx motif) are highlighted in grey. Red text indicates residues previously reported to belong to an immunogenic region or antibody epitope. Dots indicate identical residues and asterisks indicate stop codons. Only one representative of each sequence group is shown and labelled ‘VN’; the number of times the sequence was detected is indicated in parentheses.
RSV A and B share characteristic coldspots in the region flanked by the nucleoprotein gene N and the phosphoprotein gene P, including the intergenic region, highlighting their conserved functions. N binds viral RNA by forming a groove between the N- and C-terminal domains (Tawar et al., 2009); P is known to form homo-tetramers and the corresponding α-helical oligomerization domain is covered by another coldspot, also shared between RSV A and B (Karron et al., 2006, 2008). Similarly, shared coldspots in the polymerase gene L overlap with conserved region V of the polymerase, which is known to have mRNA capping activity and is conserved across non-segmented negative-sense RNA viruses (Li et al., 2008). Finally, a shared coldspot was also detected on the F gene, which is a drug and vaccine target (Dormitzer et al., 2012; Swanson et al., 2011). This coldspot covers two of the three subunits that form the ‘head’ of the structure (Fig. 3c) and are exposed on the protein surface. It extends towards the α-helical regions that form the ‘stalk’ of the trimeric protein, but does not overlap with the α-helical regions that are mainly responsible for the complex structural rearrangements that F undergoes during cell entry, suggesting greater genomic plasticity in this region. In both RSV A and B, the P and M genes had the least intra-host diversity, as measured by the fraction of the gene covered by coldspots. Overall, a larger fraction of the RSV B genome was covered in coldspots; despite this, a subtype-specific coldspot was detected in the N terminus of the M gene in RSV A.

Table 3. Gene start and gene end sequences of RSV A and B

<table>
<thead>
<tr>
<th>Gene</th>
<th>Gene start</th>
<th>No. of polymorphism sequences amongst RSV A and B</th>
<th>Gene end</th>
<th>No. of polymorphism sequences amongst RSV A and B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Consensus sequence*</td>
<td>GGGGCAAAATA</td>
<td>AGTTAATATAAAA</td>
<td></td>
</tr>
<tr>
<td>NS1</td>
<td>.</td>
<td>37/37 RSV A</td>
<td>.</td>
<td>16/16 RSV B</td>
</tr>
<tr>
<td>NS2</td>
<td>.</td>
<td>16/16 RSV B</td>
<td>.</td>
<td>37/37 RSV A</td>
</tr>
<tr>
<td>N</td>
<td>.</td>
<td>34/37 RSV A</td>
<td>.</td>
<td>3/37 RSV A</td>
</tr>
<tr>
<td>P</td>
<td>.</td>
<td>36/37 RSV A</td>
<td>.</td>
<td>1/37 RSV A</td>
</tr>
<tr>
<td>M</td>
<td>.</td>
<td>23/37 RSV A</td>
<td>.</td>
<td>13/37 RSV A</td>
</tr>
<tr>
<td>SH</td>
<td>.</td>
<td>37/37 RSV A</td>
<td>.</td>
<td>16/16 RSV B</td>
</tr>
<tr>
<td>G</td>
<td>.</td>
<td>37/37 RSV A</td>
<td>.</td>
<td>1/37 RSV A</td>
</tr>
<tr>
<td>M2</td>
<td>.</td>
<td>37/37 RSV A</td>
<td>.</td>
<td>1/16 RSV B</td>
</tr>
</tbody>
</table>

*Consensus sequences of gene start and gene ends for both RSV A and B.
†Identical gene start sequences for both RSV A and B.
Fig. 3. Distribution of mutational hotspots and coldspots across the RSV genome and on the F glycoprotein. Circos plots (Krzewinski et al., 2009) with genomic features for (a) RSV A and (b) RSV B. Blue wedges, regions depleted in intra-host variants (coldspots); red bars, regions with significant excess of intra-host variants (hotspots; height proportional to the number
DISCUSSION

Here, we report the development and application of a high-throughput sequencing strategy to sequence Vietnamese RSV whole genomes directly from clinical samples, enabling the study of inter- and intra-host genetic diversity. Overall, the data highlight evolutionary dynamics of individual genes and their impact on RSV fitness, specifically in the context of immune evasion.

The dominant Vietnamese RSV genotypes (RSV A NA1 and RSV B BA9) were similar to those circulating during the same period in Cambodia, Canada and the USA, consistent with previous observations of global dissemination since their first description in 2008–2009 (Arnott et al., 2011; Dapat et al., 2010; Eshaghi et al., 2012). Recently, a novel genotype (ON1) containing a 72 bp duplication in the HVR2 of the G gene (aa 283) has become dominant in Canada, whilst being reported only sporadically in India, South Korea, Malaysia (2010–2011) and China (2012) (Choudhary et al., 2013; Cui et al., 2013; Eshaghi et al., 2012; Khor et al., 2013; Lee et al., 2012). We did not observe this genotype amongst our 37 RSV A sequences from 2009 to 2010, nor amongst 331 Vietnamese RSV G gene sequences from a 2010–2011 nosocomial cohort of hospitalized children with acute respiratory infection (unpublished data). This suggests that its occurrence in Asia is sporadic and that it has not yet spread widely. In contrast, the BA genotype of RSV B with a duplication of 60 bp at HVR2 of the G gene (at aa 239 using GenBank accession number AY333364 as reference) (Fig. 2b) spread globally within a short period of time and is now continuously evolving with regular detection of new BA variants (Arnott et al., 2011; Eshaghi et al., 2012; Rebuffo-Scheer et al., 2011; Salter et al., 2011; Trento et al., 2003, 2010). In our study, all Vietnamese RSV B sequences were BA-like genotypes containing this 60 bp duplication and the majority (12/16) were classified as BA9. Novel subgenotypes (BA7–BA10) have been reported from 2006 to 2010 in different regions in the world (Dapat et al., 2010). One RSV B sequence (VN-731) of the BA-like genotype appears to be a novel variant of BA10, based on 96 % nucleotide similarity in the G gene and specific amino acid substitutions observed in the prototype BA10 strain (AY333364). The data suggest that the emergence and evolution of novel subgenotypes is an ongoing phenomenon.

The G gene is the most variable in the genome and encodes a surface glycoprotein that carries host cell receptor binding sites and neutralizing antibody epitopes (Escritano-Romero et al., 2004; Johnson et al., 1987a, b; Krusat & Streckert, 1997). Thus, immune pressure on this gene is likely to play a key role in driving the evolution of RSV genotypes described above. In our RSV A dataset, positively selected sites on the G protein were strongly associated with known antibody epitopes, as described in escape mutants of specific mAbs (aa 226, 265, 274 and 290) (Garcia et al., 1994; Martinez et al., 1997; Rueda et al., 1991) or in WT strains (aa 214, 215, 226, 265 and 272) (Cane, 1997; Cane & Pringle, 1995; Garcia et al., 1994). For example, sites P213L and P226L/F, which belong to an immunogenic region of G (Olmsted et al., 1989), were identified to be under positive selection in three of 37 and 34 of 37 Vietnamese RSV A sequences, respectively. Substitutions F265L and P274L/T, located within epitope 25G (Cane & Pringle, 1995; Garcia et al., 1994, Rueda et al., 1991), were under positive selection in 36 of 37 and six of 37 Vietnamese RSV A sequences, respectively. Moreover, amino acid substitution R297E/K/D, demonstrated to influence the integrity of multiple overlapping strain-specific epitopes (Rueda et al., 1995), was positively selected in all our sequences. Further epitope mapping and site-directed mutagenesis studies are required to confirm the effect of specific substitutions. In contrast, none of the positively selected RSV B sites were associated with previously described epitopes (Martinez et al., 1997); there is, however, much less information on RSV B epitopes. Interestingly, positively selected sites in RSV B, such as aa 227, 257, 276, 291 and 293, were associated with the major division of the RSV B phylogenetic tree into two branches (Botosso et al., 2009).

Glycosylation dramatically influences the antigenicity of the G protein (Palomo et al., 1991) and can thus contribute to immune evasion by masking or creating antigenic sites, abolishing G protein recognition by carbohydrate-specific
antibodies (Melero et al., 1997; Palomo et al., 2000), or enhancing the reactivity of certain antibodies (Palomo et al., 2000). These influences might help explain the poor immune memory for RSV. In our RSV A data, predicted glycosylation sites were located next to nine of 18 positively selected sites and overlapped with two, suggesting possible selection on amino acid glycosylation. There is a need to further explore a possible correlation between glycosylation patterns and seasonal shifts in subgroups that have been observed in many previous studies, including ours (data not shown). Note that these studies (including our previous study) have shown strong seasonal peaks of RSV prevalence during the rainy season from May to October (Do et al., 2011).

Our analysis of intra-host diversity showed similarly high rates of genetic variation in the G gene of RSV A and B (Fig. 3), which overlaps with the positive selection findings from our inter-host analysis (Table 1). Intriguingly, a hot-spot region was also identified in the SH gene of both RSV A and B strains (and one in the L gene for RSV A) which was not detected to be under positive selection. This could potentially be due to the small number of sequences analysed for positive selection signatures. Alternatively, this could suggest a more direct role for the nucleic acid in viral survival that does not impact selection at the protein level. Our analysis also identified several coldspots that correlate well with known functionally important regions in RSV, suggesting that this approach could complement traditional methods (e.g. multiple sequence alignments to detect sequence conservation) for identifying new functional regions and drug targets. Overall, the large number of coldspots detected is consistent with the observation that RSV is genetically very constrained and has replicated largely unchanged over the past 50 years (Tan et al., 2012).

Our inter- and intra-host analysis indicated that the M gene was the most conserved in the genome, consistent with its biological function as a major structural protein involved in viral replication, assembly and interaction with host cells (Kumaria et al., 2011; Rebuffo-Scheer et al., 2011; Tan et al., 2012). The presence of highly conserved M gene regions in both RSV A and B clinical samples (Fig. 3a, b) makes it a promising candidate for vaccine and drug development. Coldspots in M overlap with leucine-rich nuclear export signals (aa 195–206 and 46–60); in vitro inhibition of nuclear export by leptomycin has been shown to block virus assembly and RSV virion production (Ghildyal et al., 2009). Other previously explored drug targets within the RSV genome include highly conserved regions of the N and F proteins (ALN-RSV01/RSV-604 and motavizumab, respectively) (Dormitzer et al., 2012; Empey et al., 2010). Regions containing signals that direct viral mRNA transcription or antigenic synthesis, such as the 3′ extragenic leader and the gene start and gene end sequences, were also found to be conserved in our study (Table 3) (Collins et al., 1991; Fears et al., 2000; Kuo et al., 1996, 1997; Mink et al., 1991; Moudy et al., 2003).

Some studies have suggested that intra-host variation correlates with disease severity (Vignuzzi et al., 2006, 2008); however, our phylogenetic analysis of consensus sequences did not reveal any such clustering. Nevertheless, the identification of unique hotspots in the mucinoid I region of the G protein and in the central region of the L protein in severe RSV cases is intriguing, although the limited number of severe RSV cases (n=6) in this study precludes statistical analysis. Our focus on patients with high viraemia also limits our ability to explore factors that could be involved in viral attenuation, such as hotspots that differentiate between patients with high versus low viral loads.

We were also able to compare the evolution of RSV A and B. Similar to Rebuffo-Scheer et al. (2011), rates of synonymous and non-synonymous substitution, the rate of nucleotide substitutions per site, and the number of positively selected sites per gene and in total were higher in RSV A than in RSV B. In contrast, others have shown that RSV B evolves faster than RSV A (Martinez et al., 1999; Matheson et al., 2006). This discordance could be explained by the larger number of sequences from subgroup A collected over a longer period and emphasizes the need for further studies on RSV B.

METHODS

Ethics. This study was approved by the Institutional Review Board of Children’s Hospitals 1 and 2, the Scientific and Ethical Committee of the Hospital for Tropical Diseases, Ho Chi Minh City, Vietnam and the Oxford University Tropical Research Ethical Committee, Oxford, UK. Written informed consent was obtained from parents or legal guardians of children before enrolment into the study.

Collection, preparation and deep-sequencing of clinical samples. Nasopharyngeal swabs were collected from 301 RSV-positive children enrolled in a study with 632 enrolled patients on acute lower respiratory infections between May 2009 and December 2010 at the two largest paediatric referral hospitals in southern Vietnam: Children’s Hospitals 1 and 2 in Ho Chi Minh City. Swabs were placed in 1 ml viral transport medium (WHO, 2006), kept at 4 °C for a maximum of 24 h, and then aliquoted and stored at −80 °C. A shift in the predominant subgroup from B to A was seen between the 2009 and 2010 seasons (data not shown). Severe cases were defined as patients hospitalized in the paediatric intensive care unit requiring supplementary oxygen/mechanical ventilation or having a peripheral capillary oxygen saturation (SpO$_2$) <92 %.

After viral RNA extraction with a QIAamp Viral RNA Mini kit (Qiagen), 53 out of 301 samples were selected for whole-genome sequencing. These showed high viral loads, as determined by quantitative reverse transcription-PCR (Do et al., 2012), and encompassed two consecutive transmission seasons. More details of the sequencing protocol (primer sequences, read statistics and example coverage plots) are available upon request. In all, 53 RSV-containing samples (37 RSV A and 16 RSV B) were selected for analysis as representative of the total (summary information of all sequenced samples is available upon request) and comprised 18 % (53/301) of the enrolled RSV-positive patients. All sequences have been uploaded to GenBank and next-generation sequencing data have been uploaded to the European Nucleotide Archive.

Assembly of full-length consensus sequences. For each sample, we computed consensus/master genome sequences with iCORN (Otto et al., 2010), which iteratively maps reads against a reference
sequence and extracts a new reference. As an initial reference, we used a mosaic sequence created from Sanger-sequenced fragments (from samples VN-217 and VN-144 for RSV A and B, respectively), which covered the genome only partially and filled the gaps with RefSeq NC_001803 for RSV A and the recently sequenced IN_032120 (Rebuffo-Scheer et al., 2011) for RSV B. The latter was preferred over RefSeq NC_001781 because it contains a 60 base duplication in the G gene, which we also identified in de novo assembled contigs otherwise too short to be useful (data not shown). Reads of each sample were mapped against each sample’s consensus sequence with RazerS (Weese et al., 2009). Reads overlapping PCR primer positions were removed and base-quality recalibration was performed with GATK (McKenna et al., 2010) ignoring sites with >1% variation.

**Phylogenetic analyses.** For genotyping, maximum-likelihood phylogenetic trees of the hypervariable region of the G gene of RSV A and B were reconstructed from an assembled database of 165 and 74 sequences, respectively (database available upon request). RSV genotypes were assigned based on the maximum-likelihood trees and relationship to sequences of representative genotypes if bootstrap support was >70% (Arnott et al., 2011; Dapat et al., 2010; Gaunt et al., 2011; Venter et al., 2001).

To study the relationships of Vietnamese RSV to other whole genomes available in GenBank, we used MUSCLE 3.8 (Edgar, 2004) to generate alignments with sequences from Rebuffo-Scheer et al. (2011) (n=25 for RSV A, n=9 for RSV B), Kumaria et al. (2011) (n=14 for RSV A), 11 reference RSV A sequences (GenBank accession numbers U39661.1, NC_001803.1, AF035006.1, U39662-S2, U50363.1, U50362.1, U63644.1, M74568.1, AY911262.1, FJ948820.1 and FJ614813.1) (Collins et al., 1987; Connors et al., 1995; Crowe et al., 1996; Firestone et al., 1996; Lo et al., 2005; Mink et al., 1991; Stec et al., 1991; Tolley et al., 1996; Whitehead et al., 1998) and four reference RSV B sequences (GenBank accession numbers AF103255.1, AF13254.1, YJ535550.1 and NC_001781.1) (Karron et al., 1997).

For each subgroup, separate alignments of protein-coding sequences (NS1, NS2, N, M, P, G, F, SH, M2 and L), NCRs and IGs were generated using BioEdit 7.0.9.0 (Hall, 1999). Maximum-likelihood phylogenies were reconstructed for the whole genome and protein-coding sequences using the GTR + I + M model of nucleotide substitution determined by ModelTest 3.7 (Posada & Crandall, 1998) in RAXML 7.0.4 (Stamatakis, 2006) with 1000 bootstrap replicates. Phylogenies were viewed with FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

**Substitution rates and positive selection analysis.** Protein-coding sequence alignments were analysed with SNAP (Korber, 2000) to estimate overall substitution rates, and the rates of synonymous (dS) and non-synonymous substitutions (dN) (Nei & Gojobori, 1986). Percentages of conserved nucleotides in each NCR for RSV A and B subgroups were calculated and compared with reference viruses with GenBank accession numbers NC_001803 (RSV A) and NC_001781 (RSV B). The HyPhy software package (Pond et al., 2005) was used to identify individual codons within the protein-coding sequence evolving under positive selection. Duplicate sequences were removed from multiple sequence alignments. Recombination breakpoints were predicted using a combination of SBE and GARD programs, and used to split the multiple sequence alignments into recombination-free subalignments, which were analysed for positive selection (Nielsen & Yang, 1998) using the Nielsen–Yang (NY98) method. In subalignments with evidence for positive selection, a Bayesian calculation for posterior probabilities was used to identify individual codons under selection (Yang et al., 2005); sites with a posterior probability of P>0.5 for having dN/dS>1 were identified as possibly under positive selection; those with higher posterior probabilities (P>0.95) were identified as significant.

**Variation in coding genes and predicted glycosylation sites.** The deduced amino acid sequences of the C-terminal HVR2 of the G protein were compared with the following references (GenBank accession number): the prototype A2 (M11486), BA4128/99B (AY33364), Long (AY911262), CH18537 (M17213), B1 (NC_001781) and BA1 (AF013254). Potential N- and O-glycosylation sites in the C-terminal HVR2 of the G protein were identified with the online services NetOGlyc 3.1 and NetNGlyc 1.0 (http://www.cbs.dtu.dk/services/).

**Analysis of intra-host genetic variation.** Low-frequency SNVs present in the viral population in each sample were predicted with the sensitive, quality-aware variant caller LoFreq (Wilm et al., 2012). We only considered variants passing a P-value threshold of <5% after multiple testing correction (Bonferroni). See https://github.com/CSBS/2015-do-hrsv for a list of all predicted intra-host variants. To identify mutational hotspots per sample, a scanning window approach was used (window size of 20 and overlap of 5 nt) to look for an excess of SNVs compared with the genome-wide mean (binomial test; Bonferroni-corrected P-value <0.05). For coldspots, SNVs were pooled from all samples and SNV-free windows of significantly large size were detected (binomial test; Bonferroni-corrected P-value <0.05). Variant calling and hotspot/coldspot analysis followed the recipes described and validated in Wilm et al. (2012).

**ACKNOWLEDGEMENTS**

The study would not have been possible without excellent support from clinicians at the Paediatric Intensive Care Unit, Emergency Department and the Paediatric Respiratory Wards at the Children’s Hospitals 1 and 2, Ho Chi Minh City, Vietnam. Special thanks to nurses Hoang thi Minh Tu, Nguyen thi Hong Ngoc, Nguyen Viet Truong, Nguyen thi Ngoc Ha, Nguyen thi Thanh Nha, Le thi Kim Loan, Huyh thi Phuong Thao and Tran thi Tuey Nhung at Children’s Hospitals 1 and 2 for help with sample collection. We would like to thank Pauline Poh Kim Aw for her technical advice in setting up the RSV whole-genome sequencing.

**REFERENCES**


Cane, P. A. (1997). Analysis of linear epitopes recognised by the primary human antibody response to a variable region of the


