Journal of General Virology (2015), **96**, 3470–3483 **DOI** 10.1099/jgv.0.000298

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,[1]† Andreas Wilm,[3]† H. Rogier van Doorn,[1,2] 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 Dac,[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,[1] Menno de Jong,[1,2,6] Swaine Chen,[3] Niranjan Nagarajan,[3] Juliet E. Bryant[1,2] and Martin L. Hibberd[3]

Correspondence
Lien Anh Ha Do
dr.anhha@gmail.com

[1]Oxford University Clinical Research Unit, Wellcome Trust Major Overseas Program, Ho Chi Minh City, Vietnam

[2]Nuffield Department of Clinical Medicine, University of Oxford, Oxford, UK

[3]Genome Institute of Singapore, Genome Building, 138672 Singapore

[4]Children's Hospital 1, Ward 10, District 10, Ho Chi Minh City, Vietnam

[5]Children's Hospital 2, Ben Nghe Ward, District 1, Ho Chi Minh City, Vietnam

[6]Department 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.

Received 2 July 2015
Accepted 23 September 2015

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

†These authors contributed equally to this paper.

The GenBank/EMBL/DDBJ accession numbers for the genome sequences for RSV A and B are KJ939919–KJ939971. Short reads for all samples were uploaded to the European Nucleotide Archive with accession ID PRJEB6237.

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).

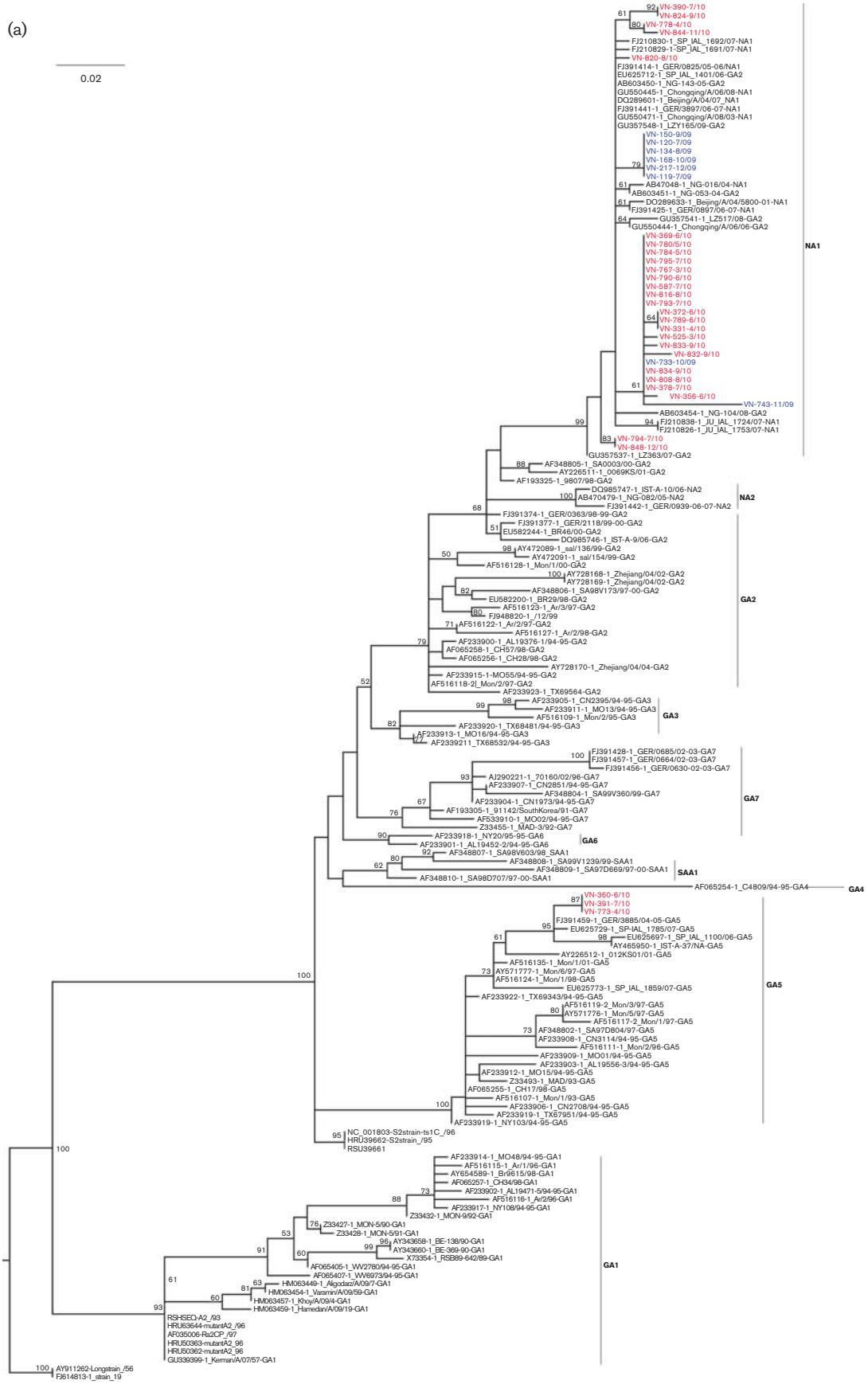
Vietnam is a high-burden country for infant respiratory infection, morbidity and mortality, with RSV as the leading cause amongst hospitalized children (Do *et al.*, 2011; Singh, 2005; Yoshida *et al.*, 2010). Here, we used whole-genome, next-generation Illumina sequencing and a rigorous variant-calling 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 Vietnam 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



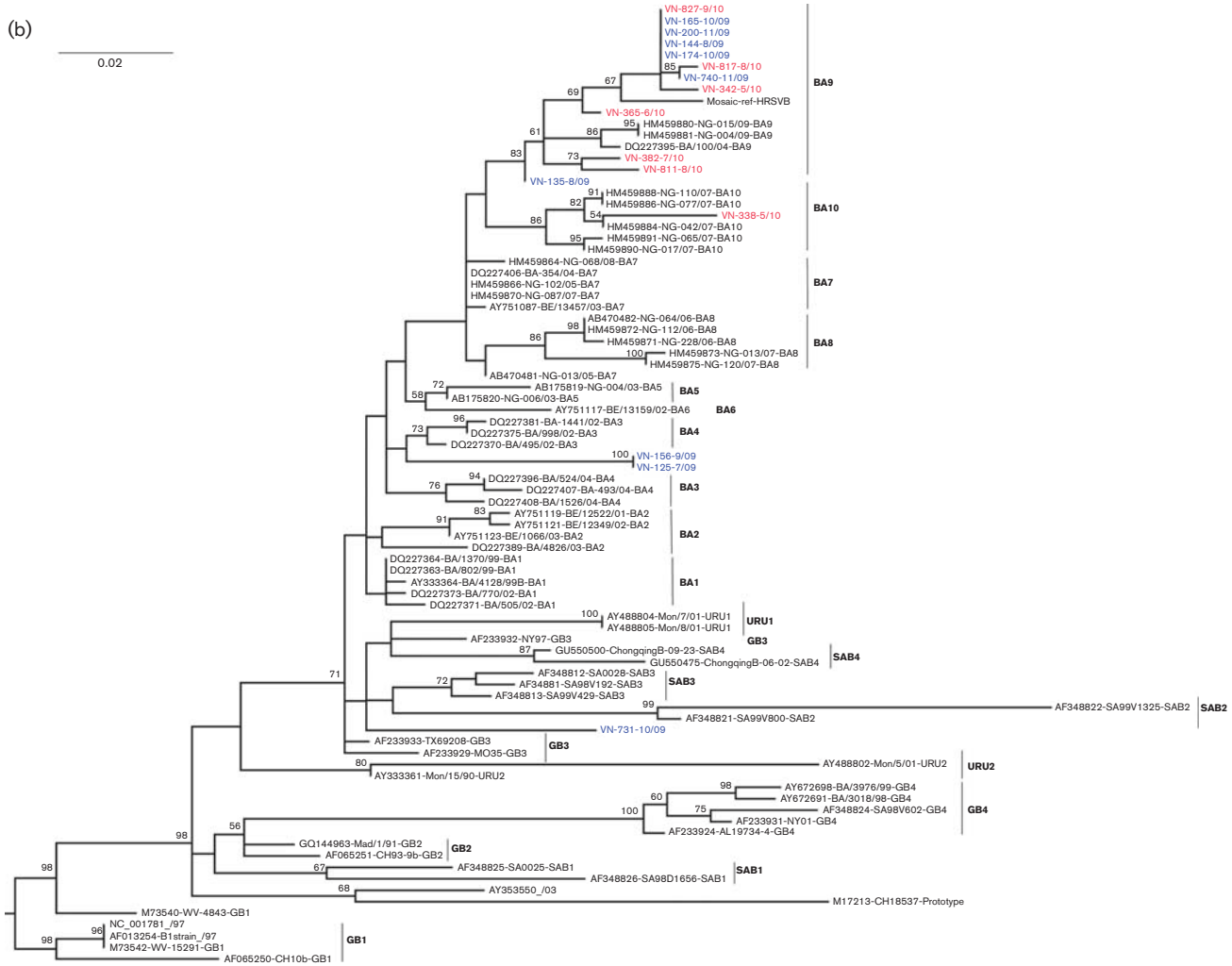


Fig. 1. Maximum-likelihood phylogenetic analysis of the C-terminal second hypervariable region (HVR2) of the G gene of Vietnamese (a) RSV A and (b) RSV B. A discrete Γ distribution was used to model evolutionary rate differences amongst sites (four categories). Branch labels indicate the stability of the branches over 1000 bootstrap replicates. Trees are drawn to scale, with branch lengths measured in the number of substitutions per site. Bar, rate of nucleotide substitutions per site. For RSV A, the cluster of two reference sequences Long and FJ614813 was used as the outgroup; for RSV B, the cluster of three reference sequences NC_001781, AF013254 and M73542 was used. Vietnamese sequences are indicated by 'VN' followed by the month and year of collection; those in blue and in red were collected in 2009 and 2010, respectively. The remaining sequences are from prototype strains and representative genotypes which are indicated with GenBank accession number, place, year of collection and their assigned genotype.

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 d*N*/d*S* 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 (d*N*/d*S*=4.42, $P>50$ %), 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

Table 1. Summary of selection pressure analysis

Gene	RSV A (79 sequences)					RSV B (25 sequences)			
	Length (bp)	No. of unique sequences	No. of RB* (position; bp)	Fragment coordinates (if RB >0)	LRT†	Length (bp)	No. of unique sequences	No. of RB* (position; bp)	LRT†
NS1	417	32	0		1	417	14	0	1
NS2	372	26	0		0.88	372	13	0	1
N	1173	53	0		0.73	1173	16	0	1
P	723	45	0		1	723	17	0	1
M	768	39	1			768	14	0	0.50
				1–168	1				
				169–768	1				
SH	192	25	0		1	195	11	0	0.65
G	891	57	1			936	21	0	0.002
				1–570	0.23693				
				571–891	2.87×10^{-10}				
F	1722	62	1		1	1722	20	0	0.29
				1–543	1				
				544–1722	1				
M2-1	582	38	0		1	585	17	0	0.73
M2-2	264	29	0		−1	270	10	0	0.74
L	6495	74	5			6498	23	0	1
				1–741	1				
				742–1599	0.63128				
				1600–2805	0.39852				
				2806–3924	1				
				3925–5574	1				
				5575–6495	1				

*Recombination breakpoints (RB) predicted by SBP and GARD.
†Likelihood ratio test (LRT) is significant if <0.05.

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 ($dN/dS=6.67$, $P>50$ %). Two of these have not been described before: aa 159 ($P=99$ %) and ($P=88$ %; 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, *G* score >0.5) were predicted to be *O*-glycosylated and three sites were predicted to be *N*-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 *O*-glycosylation motifs; in addition, aa 289 was both positively selected and predicted to be *O*-glycosylated in all NA1 sequences, and aa 250 was positively selected and predicted to be *N*-glycosylated in all GA5 sequences (Fig. 2a). In contrast, none of the RSV B potentially positively selected sites were predicted to be *O*- or *N*-glycosylated (Fig. 2b).

Amongst Vietnamese RSV A sequences, predicted *N*-glycosylation patterns in the G protein HVR2 varied by genotype (Fig. 2a). Potential *N*-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 *N*-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 *O*-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 2. Rates of synonymous (d*S*) and non-synonymous (d*N*) mutations in Vietnamese RSV A and B

Gene	RSV A					RSV B				
	Mean d*S*	Mean d*N*	d*N*/d*S*	Total sites	Substitutions per site*	Mean d*S*	Mean d*N*	d*N*/d*S*	Total sites	Substitutions per site*
NS1	0.11	0.003	0.03	420	0.006	0.04	0.003	0.08	420	0.004
NS2	0.13	0.003	0.03	375	0.005	0.06	0.003	0.04	375	0.005
N	0.11	0.002	0.02	1176	0.005	0.04	0.0008	0.02	1176	0.003
P	0.09	0.003	0.03	726	0.005	0.07	0.001	0.02	726	0.005
M	0.10	0.0005	0.01	771	0.004	0.03	0.0007	0.03	771	0.002
SH	0.13	0.004	0.03	195	0.005	0.04	0.007	0.18	198	0.002
G	0.10	0.04	0.39	897	0.020	0.07	0.02	0.33	960	0.008
F	0.11	0.006	0.05	1725	0.008	0.04	0.03	0.06	1725	0.004
M2-1	0.10	0.0061	0.05	585	0.005	0.06	0.004	0.06	588	0.005
M2-2	0.05	0.03	0.48	267	0.009	0.04	0.009	0.26	273	0.004
L	0.07	0.003	0.05	6498	0.006	0.05	0.002	0.03	6501	0.003

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

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.

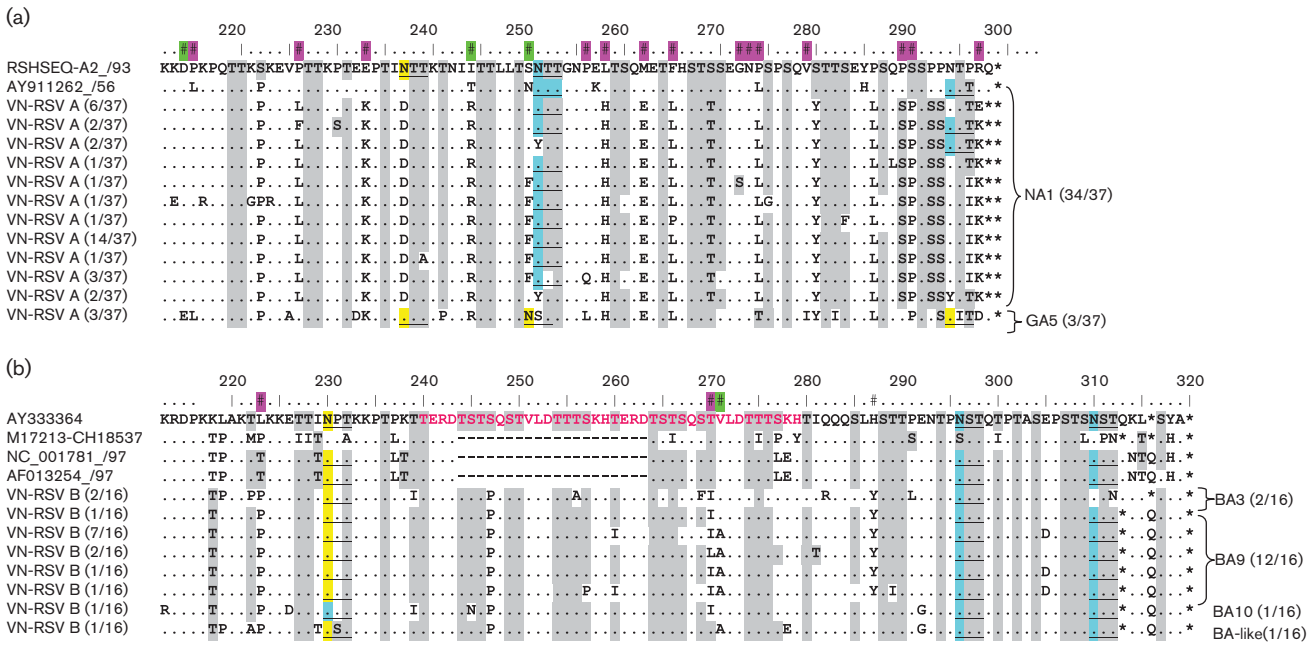


Fig. 2. Predicted positively selected and glycosylated sites in the C-terminal HVR2 of the G glycoprotein from Vietnamese (a) RSV A and (b) RSV B sequences. Hash symbols (top row) indicate positions identified as positively selected (purple, previously identified positions; green, newly detected positions). 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.

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

Dots indicate identical nucleotides and dashes indicate deleted nucleotides (compared with consensus sequences).

Gene	Gene start	No. of polymorphism sequences amongst RSV A and B	Gene end	No. of polymorphism sequences amongst RSV A and B
Consensus sequence*	**GGGGCAAATA**		**AGTTAATATAAAA**	
NS1	..........†		.....	
NS2	.....		..... T.TA..–	37/37 RSV A
			..... A.T..A..–	16/16 RSV B
N	..........†		.....A.....A	34/37 RSV A
			..... C.A.....A	3/37 RSV A
			.....C.A ....TA	16/16 RSV B
P	..........†		..... CA.A.....	36/37 RSV A
			....CA.A ...AAG	1/37 RSV A
			.... A.CA.A	15/16 RSV B
			..... A.CA.A.G..	1/16 RSV B
M	..........†		.....A.....–	23/37 RSV A
			..... C.A..	1/37 RSV A
			.....A..	13/37- RSV A
			..G ..... ATA.–	16/16 RSV B
SH	.....	37/37 RSV A	.....TA.....	34/37 RSV A
	.....T.....	16/16 RSV B	..... CTA.....	3/37 RSV A
			..... T.TA.....	16/16 RSV B
G	.....G†		..... C.T..A..–	36/37 RSV A
			..... C.T.TA.–	1/37 RSV A
			–..... T.CA...	3/16 RSV B
			–.....T.TA	13/16 RSV B
F	..........†		..... TATA.–	37/37 RSV A
			.....CATA	15/16 RSV B
			.....CATA...G	1/16 RSV B
M2	..........†		..... T.TA..–	37/37 RSV A
			.....TCTA–	16/16 -RSV B
L	... A...A–†		..... T.TA..–	37/37 RSV A
			.....T..A ....–	16/16 RSV B

*Consensus sequences of gene start and gene ends for both RSV A and B.
†Identical gene start sequences for both RSV A and B.

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 (aa 104–163) (Llorente *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.

Although hotspots in the SH and G genes were found in both RSV A and B, the specific locations of these hotspots differed by subgroup (Fig. 3a, b). Both genes are non-essential for viral replication *in vitro* (Karron *et al.*, 1997) and could therefore be more likely to acquire SNVs without negative effects. RSV A has a large number of novel hotspots in the L gene, clustered within the *S*-adenosyl-L-methionine-dependent methyl-transferase domain of the polymerase. Novel hotspots were also observed in RSV A NS1 and N genes, potentially

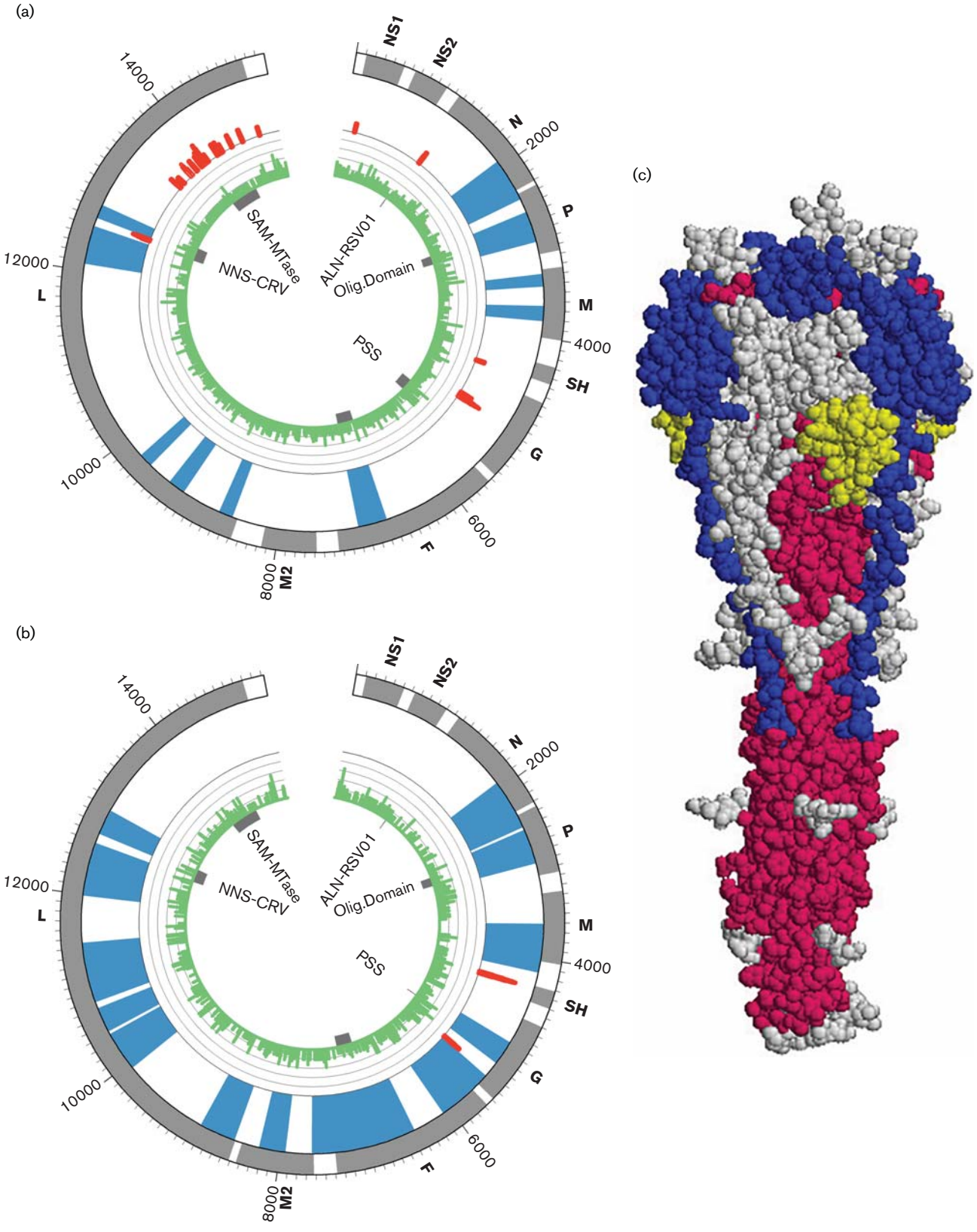


Fig. 3. Distribution of mutational hotspots and coldspots across the RSV genome and on the F glycoprotein. Circos plots (Krzywinski *et al.*, 2009) with genomic features for (a) RSV A and (b) RSV B. Blue wedges, regions depleted in intra-host variants (coldspots); red bars, regions with significant excess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

of times they were identified); green bars, sequence conservation in the corresponding multiple sequence alignment measured as Shannon entropy. ALN-RSV01, location of therapeutic small interfering RNA of the same name; Olig.Domain, oligomerization domain of P. PSS, positively selected sites; NNS-CRV, conserved region V of non-segmented negative-sense RNA viruses; SAM-MTase, the *S*-adenosyl-L-methionine-dependent methyl-transferases domain of the polymerase. (c) Three-dimensional model of the RSV-F post-fusion trimer (based on Protein Data Bank ID: 3RKI); also marked as grey box in (a) and (b). Red, α-helical regions; blue, predicted coldspots; yellow, motavizumab epitope.

indicating uncharacterized functional regions. Interestingly, the N 5′-end hotspot is just 60 bp downstream of the region targeted by the therapeutic small interfering RNA ALN-RSV01 (Alvarez *et al.*, 2009), which is currently in clinical trials.

RSV A sample numbers were sufficient to allow us to conduct a hotspot analysis separately for severe and non-severe cases. Intriguingly, we found a hotspot unique to severe cases located at the mucinoid I region of the G gene (aa 114–141), whilst hotspots uniquely found in non-severe cases were located in NS1 and NS2, the IGS between the SH and G genes, and L gene regions (Fig. 3a, b).

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 (Escribano-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) (García *et al.*, 1994; Martínez *et al.*, 1997; Rueda *et al.*, 1991) or in WT strains (aa 214, 215, 226, 265 and 272) (Cane, 1997; Cane & Pringle, 1995; García *et al.*, 1994). For example, sites P215L 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; García *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 (Martínez *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 hotspot 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 antigenome 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; Fearns *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 (Martínez *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 JN_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 AF013255.1, AF13254.1, AY353550.1 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 IGSs 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 + Γ4 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 (d*S*) and non-synonymous substitutions (d*N*) (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 SBP 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 d*N*/d*S*>1 were identified as possibly under positive selection; those with high 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/CSB5/2015-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, Huynh thi Phuong Thao and Tran thi Tuyet 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

Alvarez, R., Elbashir, S., Borland, T., Toudjarska, I., Hadwiger, P., John, M., Roehl, I., Morskaya, S. S., Martinello, R. & other authors (2009). RNA interference-mediated silencing of the respiratory syncytial virus nucleocapsid defines a potent antiviral strategy. *Antimicrob Agents Chemother* **53**, 3952–3962.

Arnott, A., Vong, S., Mardy, S., Chu, S., Naughtin, M., Sovann, L., Buecher, C., Beauté, J., Rith, S. & other authors (2011). A study of the genetic variability of human respiratory syncytial virus (HRSV) in Cambodia reveals the existence of a new HRSV group B genotype. *J Clin Microbiol* **49**, 3504–3513.

Baek, Y. H., Choi, E. H., Song, M. S., Pascua, P. N., Kwon, H. I., Park, S. J., Lee, J. H., Woo, S. I., Ahn, B. H. & other authors (2012). Prevalence and genetic characterization of respiratory syncytial virus (RSV) in hospitalized children in Korea. *Arch Virol* **157**, 1039–1050.

Blanc, A., Delfraro, A., Frabasile, S. & Arbiza, J. (2005). Genotypes of respiratory syncytial virus group B identified in Uruguay. *Arch Virol* **150**, 603–609.

Botosso, V. F., Zanotto, P. M., Ueda, M., Arruda, E., Gilio, A. E., Vieira, S. E., Stewien, K. E., Peret, T. C., Jamal, L. F. & other authors (2009). Positive selection results in frequent reversible amino acid replacements in the G protein gene of human respiratory syncytial virus. *PLoS Pathog* **5**, e1000254.

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

attachment (G) protein of respiratory syncytial virus. *J Med Virol* **51**, 297–304.

Cane, P. A. & Pringle, C. R. (1995). Evolution of subgroup A respiratory syncytial virus: evidence for progressive accumulation of amino acid changes in the attachment protein. *J Virol* **69**, 2918–2925.

Choudhary, M. L., Wadhwa, B. S., Jadhav, S. M. & Chadha, M. S. (2013). Complete genome sequences of two human respiratory syncytial virus genotype A strains from India, RSV-A/NIV1114046/11 and RSV-A/NIV1114073/11. *Genome Announc* **1**, e00165–e00113.

Coates, H. V., Alling, D. W. & Chanock, R. M. (1966). An antigenic analysis of respiratory syncytial virus isolates by a plaque reduction neutralization test. *Am J Epidemiol* **83**, 299–313.

Collins, P. L., Olmsted, R. A., Spriggs, M. K., Johnson, P. R. & Buckler-White, A. J. (1987). Gene overlap and site-specific attenuation of transcription of the viral polymerase L gene of human respiratory syncytial virus. *Proc Natl Acad Sci U S A* **84**, 5134–5138.

Collins, P. L., Mink, M. A. & Stec, D. S. (1991). Rescue of synthetic analogs of respiratory syncytial virus genomic RNA and effect of truncations and mutations on the expression of a foreign reporter gene. *Proc Natl Acad Sci U S A* **88**, 9663–9667.

Connors, M., Crowe, J. E. Jr, Firestone, C. Y., Murphy, B. R. & Collins, P. L. (1995). A cold-passaged, attenuated strain of human respiratory syncytial virus contains mutations in the F and L genes. *Virology* **208**, 478–484.

Crowe, J. E. Jr, Firestone, C. Y., Whitehead, S. S., Collins, P. L. & Murphy, B. R. (1996). Acquisition of the *ts* phenotype by a chemically mutagenized cold-passaged human respiratory syncytial virus vaccine candidate results from the acquisition of a single mutation in the polymerase (L) gene. *Virus Genes* **13**, 269–273.

Cui, G., Qian, Y., Zhu, R., Deng, J., Zhao, L., Sun, Y. & Wang, F. (2013). Emerging human respiratory syncytial virus genotype ON1 found in infants with pneumonia in Beijing, China. *Emerg Microbes Infect* **2**, e22.

Dapat, I. C., Shobugawa, Y., Sano, Y., Saito, R., Sasaki, A., Suzuki, Y., Kumaki, A., Zaraket, H., Dapat, C. & other authors (2010). New genotypes within respiratory syncytial virus group B genotype BA in Niigata, Japan. *J Clin Microbiol* **48**, 3423–3427.

Do, A. H., van Doorn, H. R., Nghiem, M. N., Bryant, J. E., Hoang, T. H., Do, Q. H., Van, T. L., Tran, T. T., Wills, B. & other authors (2011). Viral etiologies of acute respiratory infections among hospitalized Vietnamese children in Ho Chi Minh City, 2004-2008. *PLoS One* **6**, e18176.

Do, L. A., van Doorn, H. R., Bryant, J. E., Nghiem, M. N., Nguyen Van, V. C., Vo, C. K., Nguyen, M. D., Tran, T. H., Farrar, J. & de Jong, M. D. (2012). A sensitive real-time PCR for detection and subgrouping of human respiratory syncytial virus. *J Virol Methods* **179**, 250–255.

Dormitzer, P. R., Grandi, G. & Rappuoli, R. (2012). Structural vaccinology starts to deliver. *Nat Rev Microbiol* **10**, 807–813.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792–1797.

Empey, K. M., Peebles, R. S. Jr & Kolls, J. K. (2010). Pharmacologic advances in the treatment and prevention of respiratory syncytial virus. *Clin Infect Dis* **50**, 1258–1267.

Escribano-Romero, E., Rawling, J., García-Barreno, B. & Melero, J. A. (2004). The soluble form of human respiratory syncytial virus attachment protein differs from the membrane-bound form in its oligomeric state but is still capable of binding to cell surface proteoglycans. *J Virol* **78**, 3524–3532.

Eshaghi, A., Duvvuri, V. R., Lai, R., Nadarajah, J. T., Li, A., Patel, S. N., Low, D. E. & Gubbay, J. B. (2012). Genetic variability of human respiratory syncytial virus A strains circulating in Ontario: a novel genotype with a 72 nucleotide G gene duplication. *PLoS One* **7**, e32807.

Fearns, R., Collins, P. L. & Peeples, M. E. (2000). Functional analysis of the genomic and antigenomic promoters of human respiratory syncytial virus. *J Virol* **74**, 6006–6014.

Firestone, C. Y., Whitehead, S. S., Collins, P. L., Murphy, B. R. & Crowe, J. E., Jr (1996). Nucleotide sequence analysis of the respiratory syncytial virus subgroup A cold-passaged (*cp*) temperature sensitive (*ts*) *cpts*-248/404 live attenuated virus vaccine candidate. *Virology* **225**, 419–422.

García, O., Martín, M., Dopazo, J., Arbiza, J., Frabasile, S., Russi, J., Hortal, M., Perez-Breña, P., Martínez, I. & other authors (1994). Evolutionary pattern of human respiratory syncytial virus (subgroup A): cocirculating lineages and correlation of genetic and antigenic changes in the G glycoprotein. *J Virol* **68**, 5448–5459.

Gaunt, E. R., Jansen, R. R., Poovorawan, Y., Templeton, K. E., Toms, G. L. & Simmonds, P. (2011). Molecular epidemiology and evolution of human respiratory syncytial virus and human metapneumovirus. *PLoS One* **6**, e17427.

Ghildyal, R., Ho, A., Dias, M., Soegiyono, L., Bardin, P. G., Tran, K. C., Teng, M. N. & Jans, D. A. (2009). The respiratory syncytial virus matrix protein possesses a Crm1-mediated nuclear export mechanism. *J Virol* **83**, 5353–5362.

Grad, Y. H., Newman, R., Zody, M., Yang, X., Murphy, R., Qu, J., Malboeuf, C. M., Levin, J. Z., Lipsitch, M. & DeVincenzo, J. (2014). Within-host whole-genome deep sequencing and diversity analysis of human respiratory syncytial virus infection reveals dynamics of genomic diversity in the absence and presence of immune pressure. *J Virol* **88**, 7286–7293.

Hall, T. (1999). BioEdit: a user-friendly biological sequence aligment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp Ser* **41**, 95–98.

Johnson, P. R. Jr, Olmsted, R. A., Prince, G. A., Murphy, B. R., Alling, D. W., Walsh, E. E. & Collins, P. L. (1987a). Antigenic relatedness between glycoproteins of human respiratory syncytial virus subgroups A and B: evaluation of the contributions of F and G glycoproteins to immunity. *J Virol* **61**, 3163–3166.

Johnson, P. R., Spriggs, M. K., Olmsted, R. A. & Collins, P. L. (1987b). The G glycoprotein of human respiratory syncytial viruses of subgroups A and B: extensive sequence divergence between antigenically related proteins. *Proc Natl Acad Sci U S A* **84**, 5625–5629.

Karron, R. A., Buonagurio, D. A., Georgiu, A. F., Whitehead, S. S., Adamus, J. E., Clements-Mann, M. L., Harris, D. O., Randolph, V. B., Udem, S. A. & other authors (1997). Respiratory syncytial virus (RSV) SH and G proteins are not essential for viral replication *in vitro*: clinical evaluation and molecular characterization of a cold-passaged, attenuated RSV subgroup B mutant. *Proc Natl Acad Sci U S A* **94**, 13961–13966.

Khor, C. S., Sam, I. C., Hooi, P. S. & Chan, Y. F. (2013). Displacement of predominant respiratory syncytial virus genotypes in Malaysia between 1989 and 2011. *Infect Genet Evol* **14**, 357–360.

Korber, B. (2000). HIV signature and sequence variation analysis. In *Computational Analysis of HIV Molecular Sequences*, pp. 55–72. Edited by A. G. Rodrigo & G. H. Learn. Dordrecht: Kluwer.

Krusat, T. & Streckert, H. J. (1997). Heparin-dependent attachment of respiratory syncytial virus (RSV) to host cells. *Arch Virol* **142**, 1247–1254.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J. & Marra, M. A. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res* **19**, 1639–45.

Kumaria, R., Iyer, L. R., Hibberd, M. L., Simões, E. A. & Sugrue, R. J. (2011). Whole genome characterization of non-tissue culture adapted HRSV strains in severely infected children. *Virol J* **8**, 372.

Kuo, L., Grosfeld, H., Cristina, J., Hill, M. G. & Collins, P. L. (1996). Effects of mutations in the gene-start and gene-end sequence motifs

on transcription of monocistronic and dicistronic minigenomes of respiratory syncytial virus. *J Virol* **70**, 6892–6901.

Kuo, L., Fearns, R. & Collins, P. L. (1997). Analysis of the gene start and gene end signals of human respiratory syncytial virus: quasi-templated initiation at position 1 of the encoded mRNA. *J Virol* **71**, 4944–4953.

Lamb, R. A., Collins, P. L., Kolakofsky, D., Melero, J. A., Nagai, Y., Oldstone, M. B. A., Pringle, C. R. & Rima, B. K. (2005). *Paramyxoviridae.* In *Eighth Report of the International Committee on Taxonomy of Viruses*, pp. 655–668. Edited by C. M. Fauquet, M. A. Mayo, J. Maniloff, U. Desselberger & L. A. Ball. London: Elsevier Academic Press.

Lee, W. J., Kim, Y. J., Kim, D. W., Lee, H. S., Lee, H. Y. & Kim, K. (2012). Complete genome sequence of human respiratory syncytial virus genotype A with a 72-nucleotide duplication in the attachment protein G gene. *J Virol* **86**, 13810–13811.

Li, J., Rahmeh, A., Morelli, M. & Whelan, S. P. (2008). A conserved motif in region V of the large polymerase proteins of nonsegmented negative-sense RNA viruses that is essential for mRNA capping. *J Virol* **82**, 775–784.

Llorente, M. T., García-Barreno, B., Calero, M., Camafeita, E., López, J. A., Longhi, S., Ferrón, F., Varela, P. F. & Melero, J. A. (2006). Structural analysis of the human respiratory syncytial virus phosphoprotein: characterization of an alpha-helical domain involved in oligomerization. *J Gen Virol* **87**, 159–169.

Llorente, M. T., Taylor, I. A., López-Viñas, E., Gomez-Puertas, P., Calder, L. J., García-Barreno, B. & Melero, J. A. (2008). Structural properties of the human respiratory syncytial virus P protein: evidence for an elongated homotetrameric molecule that is the smallest orthologue within the family of paramyxovirus polymerase cofactors. *Proteins* **72**, 946–958.

Lo, M. S., Brazas, R. M. & Holtzman, M. J. (2005). Respiratory syncytial virus nonstructural proteins NS1 and NS2 mediate inhibition of Stat2 expression and alpha/beta interferon responsiveness. *J Virol* **79**, 9315–9319.

Martínez, I., Dopazo, J. & Melero, J. A. (1997). Antigenic structure of the human respiratory syncytial virus G glycoprotein and relevance of hypermutation events for the generation of antigenic variants. *J Gen Virol* **78**, 2419–2429.

Martínez, I., Valdés, O., Delfraro, A., Arbiza, J., Russi, J. & Melero, J. A. (1999). Evolutionary pattern of the G glycoprotein of human respiratory syncytial viruses from antigenic group B: the use of alternative termination codons and lineage diversification. *J Gen Virol* **80**, 125–130.

Matheson, J. W., Rich, F. J., Cohet, C., Grimwood, K., Huang, Q. S., Penny, D., Hendy, M. D. & Kirman, J. R. (2006). Distinct patterns of evolution between respiratory syncytial virus subgroups A and B from New Zealand isolates collected over thirty-seven years. *J Med Virol* **78**, 1354–1364.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S. & other authors (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297–1303.

Melero, J. A., García-Barreno, B., Martínez, I., Pringle, C. R. & Cane, P. A. (1997). Antigenic structure, evolution and immunobiology of human respiratory syncytial virus attachment (G) protein. *J Gen Virol* **78**, 2411–2418.

Mink, M. A., Stec, D. S. & Collins, P. L. (1991). Nucleotide sequences of the 3′ leader and 5′ trailer regions of human respiratory syncytial virus genomic RNA. *Virology* **185**, 615–624.

Moudy, R. M., Harmon, S. B., Sullender, W. M. & Wertz, G. W. (2003). Variations in transcription termination signals of human respiratory syncytial virus clinical isolates affect gene expression. *Virology* **313**, 250–260.

Nei, M. & Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* **3**, 418–426.

Nielsen, R. & Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**, 929–936.

Olmsted, R. A., Murphy, B. R., Lawrence, L. A., Elango, N., Moss, B. & Collins, P. L. (1989). Processing, surface expression, and immunogenicity of carboxy-terminally truncated mutants of G protein of human respiratory syncytial virus. *J Virol* **63**, 411–420.

Otto, T. D., Sanders, M., Berriman, M. & Newbold, C. (2010). Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology. *Bioinformatics* **26**, 1704–1707.

Palomo, C., García-Barreno, B., Peñas, C. & Melero, J. A. (1991). The G protein of human respiratory syncytial virus: significance of carbohydrate side-chains and the C-terminal end to its antigenicity. *J Gen Virol* **72**, 669–675.

Palomo, C., Cane, P. A. & Melero, J. A. (2000). Evaluation of the antibody specificities of human convalescent-phase sera against the attachment (G) protein of human respiratory syncytial virus: influence of strain variation and carbohydrate side chains. *J Med Virol* **60**, 468–474.

Parveen, S., Sullender, W. M., Fowler, K., Lefkowitz, E. J., Kapoor, S. K. & Broor, S. (2006). Genetic variability in the G protein gene of group A and B respiratory syncytial viruses from India. *J Clin Microbiol* **44**, 3055–3064.

Peret, T. C., Hall, C. B., Schnabel, K. C., Golub, J. A. & Anderson, L. J. (1998). Circulation patterns of genetically distinct group A and B strains of human respiratory syncytial virus in a community. *J Gen Virol* **79**, 2221–2229.

Peret, T. C., Hall, C. B., Hammond, G. W., Piedra, P. A., Storch, G. A., Sullender, W. M., Tsou, C. & Anderson, L. J. (2000). Circulation patterns of group A and B human respiratory syncytial virus genotypes in 5 communities in North America. *J Infect Dis* **181**, 1891–1896.

Pond, S. L., Frost, S. D. & Muse, S. V. (2005). HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21**, 676–679.

Posada, D. & Crandall, K. A. (1998). MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**, 817–818.

Power, U. F. (2008). Respiratory syncytial virus (RSV) vaccines - two steps back for one leap forward. *J Clin Virol* **41**, 38–44.

Rebuffo-Scheer, C., Bose, M., He, J., Khaja, S., Ulatowski, M., Beck, E. T., Fan, J., Kumar, S., Nelson, M. I. & Henrickson, K. J. (2011). Whole genome sequencing and evolutionary analysis of human respiratory syncytial virus A and B from Milwaukee, WI 1998-2010. *PLoS One* **6**, e25468.

Roca, A., Loscertales, M. P., Quintó, L., Pérez-Breña, P., Vaz, N., Alonso, P. L. & Saiz, J. C. (2001). Genetic variability among group A and B respiratory syncytial viruses in Mozambique: identification of a new cluster of group B isolates. *J Gen Virol* **82**, 103–111.

Rueda, P., Delgado, T., Portela, A., Melero, J. A. & García-Barreno, B. (1991). Premature stop codons in the G glycoprotein of human respiratory syncytial viruses resistant to neutralization by monoclonal antibodies. *J Virol* **65**, 3374–3378.

Rueda, P., Palomo, C., García-Barreno, B. & Melero, J. A. (1995). The three C-terminal residues of human respiratory syncytial virus G glycoprotein (Long strain) are essential for integrity of multiple epitopes distinguishable by antiidiotypic antibodies. *Viral Immunol* **8**, 37–46.

Salter, A., Laoi, B. N. & Crowley, B. (2011). Molecular epidemiology of human respiratory syncytial virus subgroups A and B identified in adults with hematological malignancy attending an Irish hospital between 2004 and 2009. *J Med Virol* **83**, 337–347.

Shobugawa, Y., Saito, R., Sano, Y., Zaraket, H., Suzuki, Y., Kumaki, A., Dapat, I., Oguma, T., Yamaguchi, M. & Suzuki, H. (2009). Emerging genotypes of human respiratory syncytial virus subgroup A among patients in Japan. *J Clin Microbiol* **47**, 2475–2482.

Singh, V. (2005). The burden of pneumonia in children: an Asian perspective. *Paediatr Respir Rev* **6**, 88–93.

Spann, K. M., Tran, K. C., Chi, B., Rabin, R. L. & Collins, P. L. (2004). Suppression of the induction of alpha, beta, and lambda interferons by the NS1 and NS2 proteins of human respiratory syncytial virus in human epithelial cells and macrophages [corrected]. *J Virol* **78**, 4363–4369.

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690.

Stec, D. S., Hill, M. G. III & Collins, P. L. (1991). Sequence analysis of the polymerase L gene of human respiratory syncytial virus and predicted phylogeny of nonsegmented negative-strand viruses. *Virology* **183**, 273–287.

Sullender, W. M. (2000). Respiratory syncytial virus genetic and antigenic diversity. *Clin Microbiol Rev* **13**, 1–15.

Swanson, K. A., Settembre, E. C., Shaw, C. A., Dey, A. K., Rappuoli, R., Mandl, C. W., Dormitzer, P. R. & Carfi, A. (2011). Structural basis for immunization with postfusion respiratory syncytial virus fusion F glycoprotein (RSV F) to elicit high neutralizing antibody titers. *Proc Natl Acad Sci U S A* **108**, 9619–9624.

Tan, L., Lemey, P., Houspie, L., Viveen, M. C., Jansen, N. J., van Loon, A. M., Wiertz, E., van Bleek, G. M., Martin, D. P. & Coenjaerts, F. E. (2012). Genetic variability among complete human respiratory syncytial virus subgroup A genomes: bridging molecular evolutionary dynamics and epidemiology. *PLoS One* **7**, e51439.

Tawar, R. G., Duquerroy, S., Vonrhein, C., Varela, P. F., Damier-Piolle, L., Castagné, N., MacLellan, K., Bedouelle, H., Bricogne, G. & other authors (2009). Crystal structure of a nucleocapsid-like nucleoprotein-RNA complex of respiratory syncytial virus. *Science* **326**, 1279–1283.

Tolley, K. P., Marriott, A. C., Simpson, A., Plows, D. J., Matthews, D. A., Longhurst, S. J., Evans, J. E., Johnson, J. L., Cane, P. A. & other authors (1996). Identification of mutations contributing to the reduced virulence of a modified strain of respiratory syncytial virus. *Vaccine* **14**, 1637–1646.

Trento, A., Galiano, M., Videla, C., Carballal, G., García-Barreno, B., Melero, J. A. & Palomo, C. (2003). Major changes in the G protein of human respiratory syncytial virus isolates introduced by a duplication of 60 nucleotides. *J Gen Virol* **84**, 3115–3120.

Trento, A., Casas, I., Calderón, A., García-García, M. L., Calvo, C., Perez-Breña, P. & Melero, J. A. (2010). Ten years of global evolution of the human respiratory syncytial virus BA genotype with a 60-nucleotide duplication in the G protein gene. *J Virol* **84**, 7500–7512.

Venter, M., Madhi, S. A., Tiemessen, C. T. & Schoub, B. D. (2001). Genetic diversity and molecular epidemiology of respiratory syncytial virus over four consecutive seasons in South Africa: identification of new subgroup A and B genotypes. *J Gen Virol* **82**, 2117–2124.

Vignuzzi, M., Stone, J. K., Arnold, J. J., Cameron, C. E. & Andino, R. (2006). Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature* **439**, 344–348.

Vignuzzi, M., Wendt, E. & Andino, R. (2008). Engineering attenuated virus vaccines by controlling replication fidelity. *Nat Med* **14**, 154–161.

Weese, D., Emde, A. K., Rausch, T., Döring, A. & Reinert, K. (2009). RazerS - fast read mapping with sensitivity control. *Genome Res* **19**, 1646–1654.

Whitehead, S. S., Juhasz, K., Firestone, C. Y., Collins, P. L. & Murphy, B. R. (1998). Recombinant respiratory syncytial virus (RSV) bearing a set of mutations from cold-passaged RSV is attenuated in chimpanzees. *J Virol* **72**, 4467–4471.

WHO (2006). Collecting, preserving and shipping specimens for the diagnosis of avian influenza A(H5N1) virus infection. WHO/CDS/EPR/ARO/2006.1. http://www.who.int/csr/resources/publications/surveillance/WHO_CDS_EPR_ARO_2006_1/en/.

Wilm, A., Aw, P. P., Bertrand, D., Yeo, G. H., Ong, S. H., Wong, C. H., Khor, C. C., Petric, R., Hibberd, M. L. & Nagarajan, N. (2012). LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res* **40**, 11189–11201.

Yang, Z., Wong, W. S. & Nielsen, R. (2005). Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol Biol Evol* **22**, 1107–1118.

Yoshida, L. M., Suzuki, M., Yamamoto, T., Nguyen, H. A., Nguyen, C. D., Nguyen, A. T., Oishi, K., Vu, T. D., Le, T. H. & other authors (2010). Viral pathogens associated with acute respiratory infections in central Vietnamese children. *Pediatr Infect Dis J* **29**, 75–77.