AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Virology

Molecular Evolutionary Dynamics of Respiratory Syncytial Virus Group A in Recurrent Epidemics in Coastal Kenya

James R. Otieno,[a] Charles N. Agoti,[a,b] Caroline W. Gitahi,[a] Ann Bett,[a] Mwanajuma Ngama,[a] Graham F. Medley,[c] Patricia A. Cane,[d] D. James Nokes[a,e]

Epidemiology and Demography Department, Kenya Medical Research Institute–Wellcome Trust Research Programme, Kilifi, Kenya[a]; Department of Biomedical Sciences, Pwani University, Kilifi, Kenya[b]; Department of Global Health and Development, London School of Hygiene and Tropical Medicine, London, United Kingdom[c]; Public Health England, Salisbury, United Kingdom[d]; School of Life Sciences and WIDER, University of Warwick, Coventry, United Kingdom[e]

ABSTRACT

The characteristic recurrent epidemics of human respiratory syncytial virus (RSV) within communities may result from the genetic variability of the virus and associated evolutionary adaptation, reducing the efficiency of preexisting immune responses. We analyzed the molecular evolutionary changes in the attachment (G) glycoprotein of RSV-A viruses collected over 13 epidemic seasons (2000 to 2012) in Kilifi ($n = 649$), Kenya, and contemporaneous sequences ($n = 1{,}131$) collected elsewhere within Kenya and 28 other countries. Genetic diversity in the G gene in Kilifi was dynamic both within and between epidemics, characterized by frequent new variant introductions and limited variant persistence between consecutive epidemics. Four RSV-A genotypes were detected in Kilifi: ON1 (11.9%), GA2 (75.5%), GA5 (12.3%), and GA3 (0.3%), with predominant genotype replacement of GA5 by GA2 and then GA2 by ON1. Within these genotypes, there was considerable variation in potential *N*-glycosylation sites, with GA2 and ON1 viruses showing up to 15 different patterns involving eight possible sites. Further, we identified 15 positively selected and 34 genotype-distinguishing codon sites, with six of these sites exhibiting both characteristics. The mean substitution rate of the G ectodomain for the Kilifi data set was estimated at 3.58×10^{-3} (95% highest posterior density interval = 3.04 to 4.16) nucleotide substitutions/site/year. Kilifi viruses were interspersed in the global phylogenetic tree, clustering mostly with Kenyan and European sequences. Our findings highlight ongoing genetic evolution and high diversity of circulating RSV-A strains, locally and globally, with potential antigenic differences. Taken together, these provide a possible explanation on the nature of recurrent local RSV epidemics.

IMPORTANCE

The mechanisms underlying recurrent epidemics of RSV are poorly understood. We observe high genetic diversity in circulating strains within and between epidemics in both local and global settings. On longer time scales (~7 years) there is sequential replacement of genotypes, whereas on shorter time scales (one epidemic to the next or within epidemics) there is a high turnover of variants within genotypes. Further, this genetic diversity is predicted to be associated with variation in antigenic profiles. These observations provide an explanation for recurrent RSV epidemics and have potential implications on the long-term effectiveness of vaccines.

Human respiratory syncytial virus (RSV) is a leading worldwide viral cause of lower respiratory tract infection (LRTI) in both young children and the elderly and is a frequent cause of hospitalization (1–3). Despite considerable effort to develop an RSV vaccine, none has been licensed. The virus is characterized into two groups (A and B) (4, 5). Strains within the groups can be further classified into genotypes and variants based on genetic divergence (6–9). The RSV attachment (G) glycoprotein is responsible for virus binding to the host cell surface receptor and is a target of human neutralizing antibodies, together with the fusion (F) glycoprotein (10, 11). The G protein shows the highest genetic and antigenic variability among all the RSV structural proteins, with evidence of accumulation of amino acid changes in its hypervariable regions over time (6, 12, 13), making its study relevant in vaccine development strategies.

The dynamics of RSV infections have been followed globally, in communities, in families, and in individuals over time (7, 14, 15). RSV epidemics often occur at regular intervals in communities, usually annually (16, 17). In Kilifi, RSV epidemics usually start between September/November of one year, peaking between January and March, with infections continuing to be detected until June/August of the following year. Individuals are found to be repeatedly reinfected with RSV throughout their lifetimes (18). The driving factors that facilitate these patterns are poorly understood. The two RSV groups have been shown to cocirculate in epidemics, with RSV-A generally occurring more frequently than RSV-B (19, 20). Additionally, sequence data have shown that a local RSV epidemic may be characterized by multiple genotypes

Received 10 December 2015 Accepted 25 February 2016

Accepted manuscript posted online 2 March 2016

Citation Otieno JR, Agoti CN, Gitahi CW, Bett A, Ngama M, Medley GF, Cane PA, Nokes DJ. 2016. Molecular evolutionary dynamics of respiratory syncytial virus group A in recurrent epidemics in coastal Kenya. J Virol 90:4990–5002. doi:10.1128/JVI.03105-15.

Editor: A. García-Sastre

Address correspondence to James R. Otieno, jotieno@kemri-wellcome.org.

Supplemental material for this article may be found at http://dx.doi.org/10.1128/JVI.03105-15.

and variants and that new variants may replace older ones in subsequent epidemics (7–9). These patterns indicate that the circulation of strains varying from those that have recently caused community epidemics is favored. A better understanding of the interplay between the virus epidemiology, the genetic and antigenic variability, and host immune dynamics in producing new RSV community epidemics may contribute to better design of control measures.

It has been suggested that genetic variability and antigenic variation within the G protein contributes to the propensity of the virus to cause repeated individual reinfections and recurrent epidemics in communities by aiding evasion of host preexisting immune responses (21). Particular amino acid substitutions and changes in glycosylation profile of the G protein have been shown to profoundly affect the antigenic profile of the virus (22, 23). The present report analyzes the genetic variability and molecular evolution of RSV group A viruses identified over 13 RSV epidemic seasons (2000 to 2012) at the rural Kenyan Coast and contemporaneous sequences collected elsewhere within Kenya and 28 other countries. Close to half of the total group A detections during the period were sequenced in their G ectodomain region. We explored the patterns of RSV-A genetic diversity in Kilifi, the influence of the observed genetic changes on the resultant protein molecule and RSV epidemiology, and the placement of the locally detected viruses on the global phylogeny.

MATERIALS AND METHODS

Study population and samples. The study was undertaken in Kilifi County, which is located at the Kenyan Coastal region with a predominantly rural catchment population of around one million. This was part of a surveillance study that aimed at understanding the epidemiology and disease burden of RSV-associated pneumonia cases (24). In this study, we used three sources of clinical samples: (i) LRTI cases for children <5 years admitted to Kilifi County Hospital (KCH) collected over the period from 2000 to 2012 (3, 24), (ii) KCH outpatient presentations for children <5 years with acute respiratory infection (ARI) from April to August 2002 (25, 26), and (iii) the Kilifi RSV birth cohort (KBC) ARI cases identified between January 2002 to August 2003 (25, 27). Nasopharyngeal aspirates, nasal washes, or nasopharyngeal swabs were collected for inpatient samples, while nasal washes were collected for the outpatient and community sampling of the birth cohort. All sample sets arise from the same catchment population and were processed similarly in the laboratory and in downstream analyses.

Sample processing and sequencing. RSV detection was conducted by antigen (immunofluorescence antibody test [IFAT]) and nucleic acid (reverse transcriptase real-time PCR) tests as previously described (24, 27, 28). For the hospital inpatient RSV-A samples collected during the surveillance period from 2002 to 2012, we randomly sampled ≥50% of the archived RSV IFAT-positive samples from each epidemic for processing. The additional hospital inpatient (2000 to 2001), outpatient (2002), and Kilifi RSV birth cohort (2002 to 2003) samples included in this analysis had previously been sequenced and preliminary results reported elsewhere (25, 27).

Total viral RNA was extracted using QIAamp viral RNA minikit (Qiagen) according to the manufacturer's instructions. Reverse transcription was performed using random primers and the OmniScript RT kit (Qiagen). The cDNA was amplified with previously described primers targeting the G ectodomain region (21, 25). BigDye v3.1 chemistry was used on an ABI 3130xl instrument to sequence the amplicons. The sequence reads were assembled into contigs using Sequencher v5.0.1 (Gene Codes Corporation, USA) and Gap4 release 2.0.0b9 (29).

Global comparison data set. All RSV-A G-gene sequences deposited in GenBank as of 20 November 2015 of the same length or longer to the Kilifi sequences and collected between 2000 and 2012 were collated and phylogenetically compared with the Kilifi sequences. This analysis aimed to determine the relatedness of the Kilifi viruses to those circulating around the world and thereby understand their global context. A total of 1,415 sequences from 29 countries were used in this analysis; 1,075 sequences from 28 countries and 340 sequences from Kenya. These sequences had been filtered from a larger data set of 1,792 sequences to include only the unique sequences per country except for Kenya where sequences from Kilifi ($n = 284$) and the rest of the country were filtered separately. In addition, the unique sequences from Kilifi were subsampled per epidemic since we had precise dates of collection (day/month/year), as opposed to the global data set, where we subsampled each country collectively from 2000 to 2012 since most of the global sequences either only had the year of collection or the samples were too sparse to sample by year. Unique sequences were identified as sequences that differ by at least one nucleotide from any other sequence over the sequenced region.

Sequence alignments and diversity analysis. All the sequences were collated and aligned using MAFFT alignment software v7.220 (30) with manual editing in Se-Al software v2.0a11 (31). Nucleotide and amino acid variability within genotypes and for all Kilifi sequences were calculated for individual epidemics and for the entire period (2000 to 2012). The analyses used MEGA v6.06 (32) for amino acid variability (p-distance) and in DnaSP v5.10 (33) for nucleotide variability (Pi [π]).

Phylogenetic analyses. The best-fit nucleotide substitution model was determined using jModelTest v2.1.7 (34). Maximum-likelihood (ML) phylogenetic trees were inferred by MEGA 6.06 under the general-time-reversible (GTR) model with site heterogeneity gamma (G) model, four gamma categories, and a proportion of the sites invariable (I) (32). Bootstrapping with 1,000 iterations was implemented to evaluate branch support of the phylogenetic clusters generated.

RSV-A genotypes were designated as previously described by Peret et al. (8). A cluster was defined as a virus or group of viruses within the same genotype that fall(s) into a phylogenetic subset away from the rest with >70% bootstrap statistical support.

Variant introduction and persistence analysis. Using a definition recently developed to characterize variants within epidemics for RSV group B viruses (9), we determined the number of variants circulating in Kilifi between 2000 and 2012 for the group A viruses. We used the same definition considering that previous estimates of the rates of nucleotide substitution for the G-gene in RSV-A and RSV-B are similar, i.e., 1.83×10^{-3} and 1.95×10^{-3} nucleotide substitutions/site/year, respectively (35, 36). The variant designation was given to a group of viruses (where the group includes a singleton) within a genotype that possesses ≥4 nucleotide differences in the G ectodomain region compared to other viruses. This analysis was done using usearch v8.1.1756 (37).

Evolutionary analysis. Evolutionary analysis used the Bayesian Markov Chain Monte Carlo-based approach implemented in BEAST v1.8.2 (38), assuming an uncorrelated lognormal relaxed molecular clock model prior to accommodate variation in molecular evolutionary rate among lineages. Input sequences were tip-dated with day, month and year of collection. We implemented the GTR (G) nucleotide substitution model of evolution and gamma (Γ) plus invariant (I) site heterogeneity model. The demographic history of the viral populations was modeled using the nonparametric Gaussian Markov random-fields (GMRF) Bayesian Skyride model (39). The analysis was run through 50 million steps with sampling after every 2,500 steps. BEAST run convergence confirmation (effective sample size [ESS] > 200) and analysis of parameter estimates were done using Tracer v1.6 (http://tree.bio.ed.ac.uk/software/tracer/). The BEAST trees were summarized using TreeAnnotator v1.8.2 (http://beast.bio.ed.ac.uk/treeannotator), and the maximum clade credibility tree was visualized by FigTree v1.4.0 (http://tree.bio.ed.ac.uk/software/figtree/).

Diversifying selection analyses. RSV-A genotype nucleotide alignments were analyzed for positive selection using PAML v4.8 (40). Data for each genotype were analyzed using the site models M0 (one-ratio), M1a (neutral), M2a (selection), and M3 (discrete). The M0 model calculates one nonsynonymous/synonymous *dN/dS* ratio (ω) for all sites, i.e., averaged over all codon sites and over the entire evolutionary time that separates the sequences. The remaining three models (M1a, M2a, and M3) allow for heterogeneous ratios among sites. M1a assumes two categories of sites: conserved sites with $\omega_0 = 0$ and neutral sites with $\omega_1 = 1$. M2a adds a third category of sites, positively selected sites ($\omega_2 > 1$), which is estimated from the data. M3 models the three heterogeneous ratios among sites as M2a but using an unconstrained discrete distribution (41).

The models M0 and M1a were used to test for selection over the whole sequenced region, while the remaining models tested for the presence of positively selected sites ($\omega > 1$). Both the Naive Empirical Bayes and Bayes Empirical Bayes methods were used to identify the sites under adaptive evolution (42).

Protein substitution analysis. The gain and loss of *N*-glycosylation and *O*-glycosylation sites were predicted using the NetNGlyc 1.0 and NetOGlyc 4.0 servers (43, 44). Only the default Asn-X-Ser/Thr sequon (where "X" is not proline) was considered in *N*-glycosylation prediction. Patterns of change in amino acids at all positions of the sequences were analyzed using python scripts.

Temporal epidemic patterns analysis. Epidemic seasons were designated to begin 1 September of one year and end 31 August of the following year. The dominant RSV group within a given epidemic was associated with ≥65% of the cases; otherwise, the groups were codominant.

Ethics statement. The samples obtained in Kilifi were collected following informed written consent from each child's guardian or parent. KEMRI Ethical Review Board, Kenya, and the Coventry Research Ethics Committee of the UK approved the study protocols (3, 24).

GenBank accession numbers. The newly generated sequences analyzed here have been deposited in GenBank under accession numbers KT765213 to KT765836. The previously sequenced and reported RSV-A sequences from Kilifi added to this analysis had been archived in GenBank under accession numbers AY524573 to AY524663 and AY660667 to AY660684.

RESULTS

A total of 2,135 RSV positive clinical samples from hospital surveillance over 13 epidemic seasons (2000 to 2012) were examined. Overall, 1,246 (58.4%) of the samples were RSV-A, 684 (32%) RSV-B, 28 (1.3%) mixed infections (RSV-A/B), and 177 (8.3%) untyped/unclassified. RSV-A viruses were detected in all the 13 epidemics with the proportion of samples being designated group A in each epidemic ranging from 12.9 to 93.7% (see Table S1 in the supplemental material).

A total of 649 RSV-A G gene nucleotide sequences were available for further analysis: 564 from hospital inpatient surveillance from 2002 to 2012, 60 sequences from an earlier collection of in- and outpatient samples collected in Kilifi from 2000 to 2002, and 25 sequences from the Kilifi RSV birth cohort (27). The sequences were 618 to 690 nucleotides in length corresponding to nucleotides 295 to 912 of the reference strain A2 (M74568). Of the 649 sequences, 284 (43.8%) were unique. The number of duplicates per sequence ranged between 2 and 37 sequences. From the 284 unique sequences that had been subsampled per epidemic, we further collectively analyzed this sequence set for uniqueness. We found only 20 sequences of the 284 that were identical across epidemics, i.e., 264 sequences were unique across all the 13 epidemics differing by at least one nucleotide.

Diversity within and between epidemics and genotypes. The genetic relatedness of the 284 unique Kilifi RSV-A sequences is shown in an ML phylogenetic tree in Fig. 1A, with the taxa color coded according to the epidemic season. On this phylogeny, the analyzed sequences fell within three major clusters corresponding to genotypes GA5, GA3, and GA2. Within these genotypes, the sequences formed distinct clusters that frequently comprised sequences that had been sampled within the same epidemic. The recently emerged genotype ON1 that has a 72-nucleotide duplication within the G ectodomain region was most closely related to genotype GA2. The genotype prevalence pattern over the period is shown in Fig. 2, with overall detection as follows: GA2 ($n = 490$, 75.5%), GA3 ($n = 2$, 0.3%), GA5 ($n = 80$, 12.3%), and ON1 ($n = 77$, 11.9%). Genotype-specific phylogenies derived from the unique GA2, GA5, and ON1 sequences are shown in Fig. 1B to D.

Genotype GA2 was observed in 11 of 13 of the epidemics, including 4 where it was the only RSV-A genotype detected. The nucleotide and amino acid variability within the genotypes in each epidemic is summarized in Table 1. For GA2, the nucleotide sequences varied by a maximum of 2.04% within an epidemic and 2.27% over the entire period with corresponding amino acid variability of up to 3.75 and 3.66%, respectively. GA5 sequences were less frequently observed and were not seen after the 2007-2008 epidemic (Fig. 2). Sequences belonging to genotype ON1 were first seen globally in 2010 to 2011 with initial detection in Kilifi in the 2011-2012 epidemic, and 77 genotype ON1 sequences from Kilifi were found in 2012 (45, 46). Genotype GA3 was represented by just 2 unique sequences of the 58 sequences obtained between 2000 and 2002 and so will not be considered further in detail.

Figure 3 illustrates the accumulation of diversity over time in the RSV-A sequences from Kilifi. It can be seen that although there was much accumulation of variation in the sequences, the sequences generally clustered closely with the sequences from previous years, with clear replacement of the previously predominant clusters with time.

Dynamics of RSV-A variants in Kilifi. A total of 89 distinct variants were determined from 649 sequences (Fig. 4) on the basis of a single sequence or cluster of sequences with ≥4 nucleotide differences from other sequences in the data set. The variants comprised of between 1 and 119 sequences, 49.4% (44/89) of which were singletons. By genotypes, there were 75, 10, 3, and 1 variants for genotypes GA2, GA5, ON1, and GA3, respectively. All the epidemics were characterized by presence of multiple variants. However, only 22.5% (20/89) of the variants were observed in several (2 to 5) epidemics, with the remainder each observed in only one epidemic. Two of the variants observed in several epidemics were detected in nonconsecutive epidemics. As well, there were two variants that were observed over four consecutive epidemics.

Rate of substitution and most recent common ancestor (MRCA). Using Bayesian analysis, the rate of substitution for the Kilifi data set was estimated at 3.58×10^{-3} nucleotide substitutions/site/year (95% highest posterior density [HPD] interval = 3.04×10^{-3} to 4.16×10^{-3}). By genotype, the rate of substitution was estimated at 2.61×10^{-3} (95% HPD = 2.22×10^{-3} to 3.02×10^{-3}) for GA2, 2.89×10^{-3} (95% HPD = 1.29 to 4.49) for ON1, and 2.28×10^{-3} (95% HPD = 1.51 to 3.05) for GA5, with no significant difference in the rate between genotypes as the 95% HPD for each overlap. For the global data set, the substitution rate was estimated at 3.06×10^{-3} nucleotide substitutions/site/year (95% HPD = 2.78 to 3.35), while the MRCA was estimated to date back 43.9 years to 1968 (95% HPD = 31.3 to 58.3). The predicted demographic history of RSV-A in Kilifi (data not shown) was characterized by fluctuating effective population size, a measure of relative genetic diversity, a finding consistent with the continual RSV case detection across the years sampled.

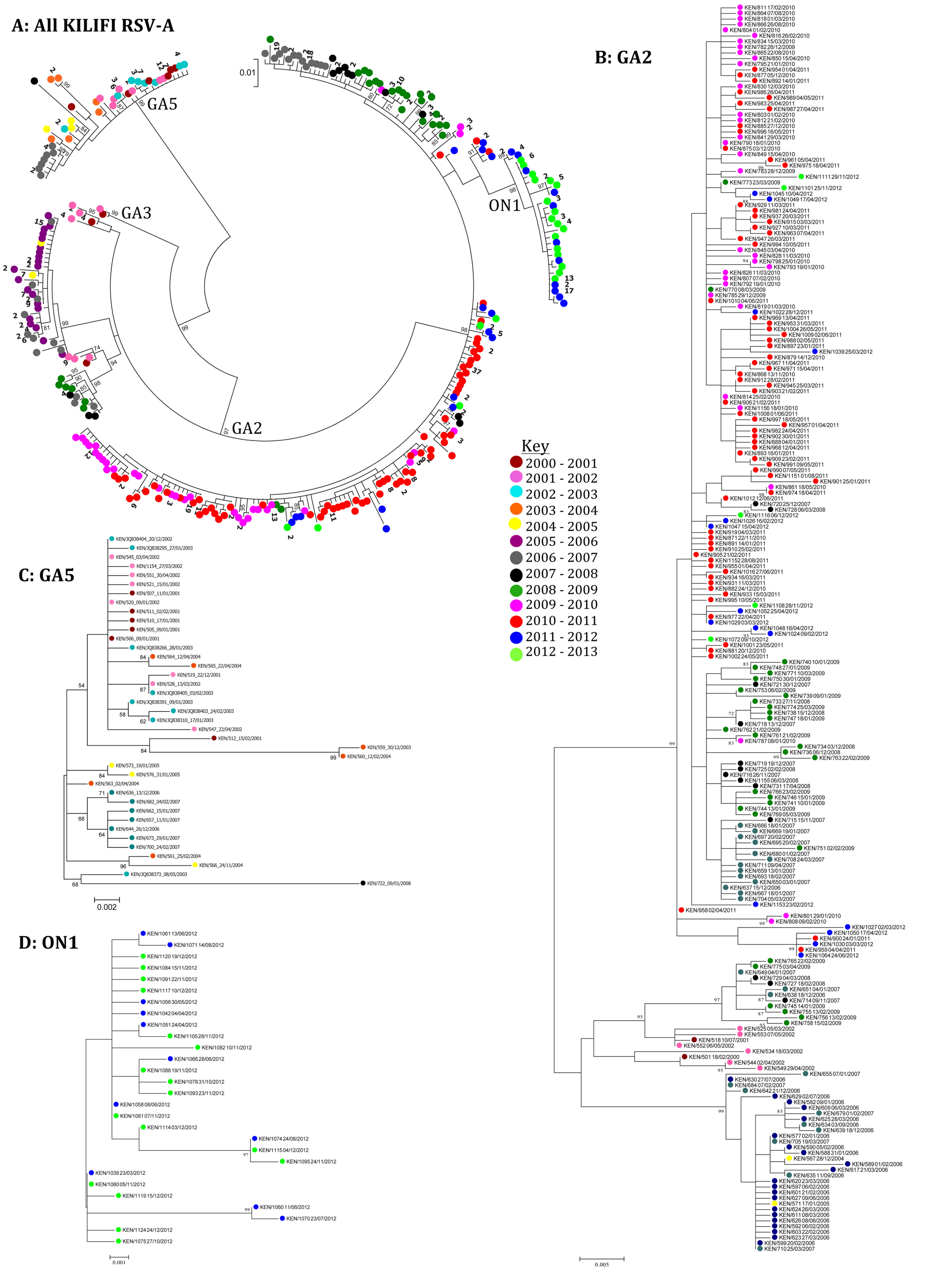


FIG 1 ML phylogenetic tree of 284 unique RSV group A G-gene, ectodomain sequences obtained in Kilifi, Kenya, from 2000 to 2012. The taxa are color coded by epidemic season of detection, as shown by the key. Panel A represents all the 284 sequences, with the numbers in boldface next to the colored circles indicating the number of identical sequences to the one shown. Genotypes, as assigned by Peret et al. (8), are shown for the main branches and only bootstrap values >70% are shown. Panels B, C, and D are subtrees from panel A representing genotypes GA2, GA5, and ON1, respectively. The taxon names denote country/serial number_date. The scale bars indicate nucleotide substitutions.

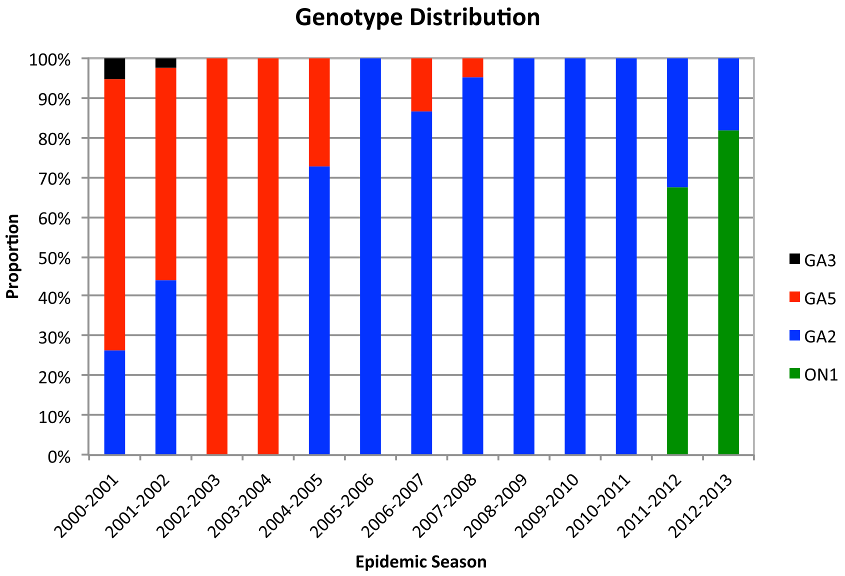


FIG 2 RSV-A genotype distribution over 13 successive epidemic seasons in Kilifi, 2000-2001 to 2012-2013. The genotypes are shown in different colors: GA2 (blue), GA5 (red), GA3 (black), and ON1 (green).

Amino acid variability and glycosylation. The alignment of unique Kilifi RSV-A amino acid sequences, ordered by date of sampling, to show amino acid variability per site is given in Fig. 5. We identified 34 codon sites with amino acid substitutions that were distinct between the genotype GA2/ON1, GA5, and GA3 viruses identified in Kilifi. Although most (31/34) of these sites were shared between two genotypes and different in the other genotype, three of these sites had a unique amino acid for each genotype. Overall, 29 (36.3%) unique amino acid sequences were observed for genotype GA5 viruses and 183 (32.2%) for genotypes GA2 and ON1. The genotype GA5 amino acid sequences were all the same length, i.e., a predicted overall G protein of 299 amino acids. The genotype GA2 sequences were 298 or 299 amino acids in length due to usage of alternative stop codons. Sequences from ON1 genotype viruses had a 24-amino-acid insertion as previously described (45).

Twelve codon positions were predicted to be potentially *N*-glycosylated (103, 105, 135, 197, 237, 238, 242, 246, 250, 251, 273, and 294). Considerable variation was seen in the individual potential *N*-glycosylation sites and the resultant combined patterns. Genotypes GA2 and ON1 viruses together showed 15 different patterns (designated 2A to 2O) involving eight different possible sites. The minimum number of possible *N*-glycosylation sites per sequence was two, and the maximum was six. The site patterns

TABLE 1 Nucleotide and amino acid variability in RSV-A genotypes identified in Kilifi, Kenya, 2000-2001 to 2012-2013

	% Variability (no. of sequences)[a]					
	GA2 and ON1[b]		GA5		Overall	
Epidemic season	Nucleotides	Amino acids	Nucleotides	Amino acids	Nucleotides	Amino acids
2000-2001	0.78 (5)	1.17 (5)	0.30 (12)	0.81 (12)	4.26 (18)	8.56 (18)
2001-2002	1.52 (18)	2.66 (18)	0.25 (21)	0.22 (21)	4.99 (40)	9.49 (40)
2002-2003	– (0)	– (0)	0.45 (25)	0.67 (25)	0.45 (25)	0.67 (25)
2003-2004	– (0)	– (0)	1.76 (7)	3.11 (7)	1.76 (7)	3.11 (7)
2004-2005	0.12 (8)	0.24 (8)	0.97 (3)	1.63 (3)	4.72 (11)	8.30 (11)
2005-2006	0.38 (56)	0.83 (56)	– (0)	– (0)	0.38 (56)	0.83 (56)
2006-2007	1.90 (70)	2.93 (70)	0.32 (11)	0.36 (11)	3.75 (81)	6.26 (81)
2007-2008	1.71 (18)	3.31 (18)	NA (1)	NA (1)	2.62 (19)	4.97 (19)
2008-2009	2.04 (47)	3.75 (47)	–	–	2.04 (47)	3.75 (47)
2009-2010	0.50 (87)	0.70 (87)	–	–	0.50 (87)	0.70 (87)
2010-2011	0.66 (154)	0.99 (154)	–	–	0.66 (154)	0.99 (154)
2011-2012	1.39 (20)	2.76 (20)	–	–	1.59 (51)	3.37 (51)
	0.43 (31)	**1.02 (31)**		–		
2012-2013	1.16 (7)	0.98 (7)	–	–	0.99 (53)	1.59 (53)
	0.37 (46)	**0.46 (46)**				
Entire period	2.27 (567)	3.66 (567)	0.77 (80)	1.23 (80)	3.87 (649)	6.44 (649)

[a] Nucleotide diversity, Pi (π), was calculated using DnaSP v5.10 (33). Amino acid diversity, p-distance, was calculated using MEGA v6.06 (32). NA, not applicable. −, no sample.
[b] ON1 values are indicated in boldface.

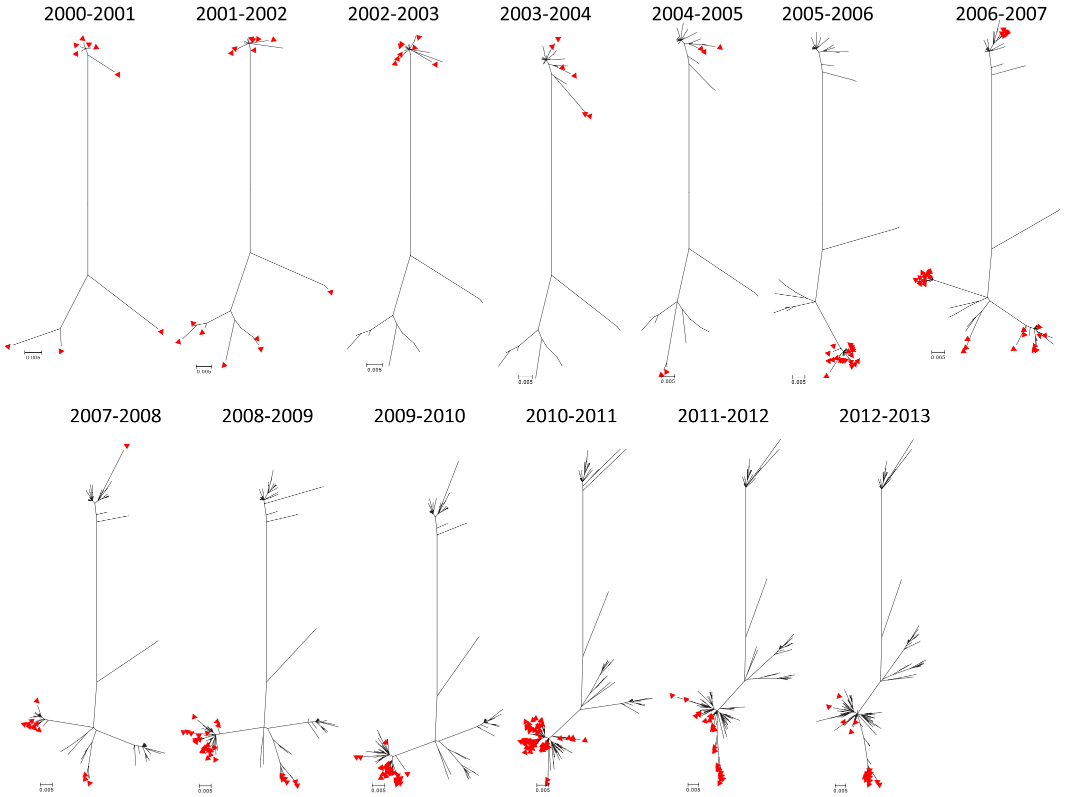


FIG 3 Accumulation of diversity in RSV-A strains over time as shown by ML radiation phylogenetic trees from Kilifi Kenya 2000-2001 to 2012-2013. Each panel shows RSV-A sequences observed since the start of the study with the sequences seen in each new epidemic highlighted with a red triangle. The clusters on the upper portion of the trees include genotype GA5 sequences, while those on the lower parts are GA2 and ON1 sequences. The lone branch in the middle includes genotype GA3 sequences.

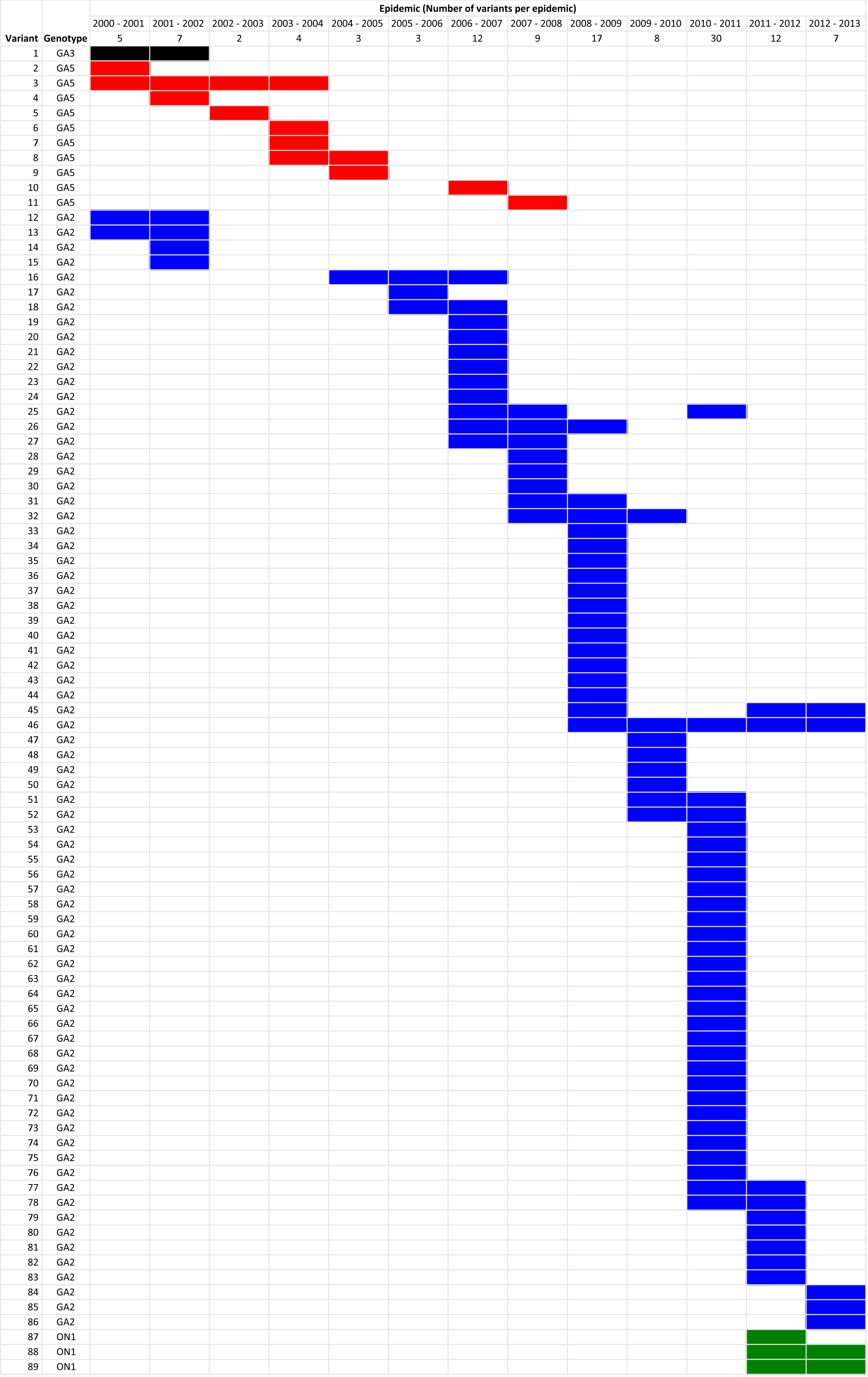
and their occurrence in the epidemics are shown in Tables 2 and 3. It can be seen that new patterns are seen in every epidemic after 2004 until 2011, when the GA2 viruses start to decline in prevalence being replaced by the emerging ON1 genotype. Five patterns were seen in multiple epidemics of between two to five before disappearance, one pattern was observed in earlier epidemics before disappearance then recurrence, but most of the patterns (nine) were only seen in single epidemics. With respect to GA5, only three *N*-glycosylation patterns were seen (5A to 5C), with one pattern being predominant from 2000 to 2001 until 2004 to 2005, occurring again in 2006 to 2007, along with a second pattern, and the third pattern only being seen in 2007 and 2008. From 2008 to 2009 until from 2012 to 2013, genotype GA5 was not detected in Kilifi.

Due to the G protein being characterized by a high prevalence of serine and threonine residues, and without the requirement of the Asn-X-Ser/Thr sequon, there were between 81 and 101 amino acid sites predicted to be *O*-glycosylated. As expected, the 24 additional amino acids in the genotype ON1 viruses offered more sites for potential *O*-glycosylation. The was also variation in the number and sites predicted to be *O*-glycosylated in successive epidemics as *N*-glycosylated, but the patterns were less clear and compounded by their ubiquity.

Positive selection analysis. Fifteen codon positions within genotypes GA2 and ON1 Kilifi sequences were observed to be under positive selection ($\omega > 1$), selection that favors change in amino acids, with a posterior probability of >0.95; Table 4. These positively selected codon sites were 101, 104, 106, 115, 147, 225, 237, 244, 248, 250, 273, 274, 275, 286 (ON1: 310), and 289 (ON1: 313). This is despite the mean nonsynonymous/synonymous (*dN*/*dS*) substitution rate ratio (ω) for the analyzed region being <1, which indicates that generally the G glycoprotein is under purifying/negative selection (selection that disfavors change in amino acids). However, the M2a site model that accommodates positive selection among sites gave the highest likelihood score against the neutral and negative selection models (M0 and M1a, respectively). No sites were identified as positively selected among the GA5 sequences, probably due to the small numbers analyzed.

Relatedness of Kilifi and global RSV-A viruses. The ML phylogenetic clustering of 1,415 global sequences from 29 countries is shown in Fig. 6A (enlarged version in Fig. S1 in the supplemental material). Kilifi viruses (blue colored taxa) were not placed into a single monophyletic group but in multiple clusters of various sizes along the tree. The composition of the clusters on the tree varied with most clusters having variants detected abroad but not in Kilifi (Fig. 6B), some clusters having variants predominantly observed in Kilifi with a few representatives from abroad (mainly Europe) (Fig. 6C), and some clusters with variants only detected in Kilifi (Fig. 6D).

All the clusters of global sequences fell within the genotypes detected in Kilifi except one cluster comprising of sequences from the United States only from 2003 to 2007 that seemed to be off the



		Epidemic (Number of variants per epidemic)												
		2000 - 2001	2001 - 2002	2002 - 2003	2003 - 2004	2004 - 2005	2005 - 2006	2006 - 2007	2007 - 2008	2008 - 2009	2009 - 2010	2010 - 2011	2011 - 2012	2012 - 2013
Variant	Genotype	5	7	2	4	3	3	12	9	17	8	30	12	7
1	GA3	⬛	⬛											
2	GA5	🟥												
3	GA5	🟥	🟥	🟥	🟥									
4	GA5		🟥											
5	GA5			🟥										
6	GA5				🟥									
7	GA5				🟥									
8	GA5				🟥	🟥								
9	GA5					🟥								
10	GA5							🟥						
11	GA5								🟥					
12	GA2	🟦	🟦											
13	GA2	🟦	🟦											
14	GA2		🟦											
15	GA2		🟦											
16	GA2					🟦	🟦	🟦						
17	GA2						🟦							
18	GA2						🟦	🟦						
19	GA2							🟦						
20	GA2							🟦						
21	GA2							🟦						
22	GA2							🟦						
23	GA2							🟦						
24	GA2							🟦						
25	GA2							🟦	🟦			🟦		
26	GA2							🟦	🟦	🟦				
27	GA2							🟦	🟦					
28	GA2								🟦					
29	GA2								🟦					
30	GA2								🟦					
31	GA2								🟦	🟦				
32	GA2								🟦	🟦	🟦			
33	GA2									🟦				
34	GA2									🟦				
35	GA2									🟦				
36	GA2									🟦				
37	GA2									🟦				
38	GA2									🟦				
39	GA2									🟦				
40	GA2									🟦				
41	GA2									🟦				
42	GA2									🟦				
43	GA2									🟦				
44	GA2									🟦				
45	GA2									🟦			🟦	🟦
46	GA2									🟦	🟦	🟦	🟦	🟦
47	GA2										🟦			
48	GA2										🟦			
49	GA2										🟦			
50	GA2										🟦			
51	GA2										🟦	🟦		
52	GA2										🟦	🟦		
53	GA2											🟦		
54	GA2											🟦		
55	GA2											🟦		
56	GA2											🟦		
57	GA2											🟦		
58	GA2											🟦		
59	GA2											🟦		
60	GA2											🟦		
61	GA2											🟦		
62	GA2											🟦		
63	GA2											🟦		
64	GA2											🟦		
65	GA2											🟦		
66	GA2											🟦		
67	GA2											🟦		
68	GA2											🟦		
69	GA2											🟦		
70	GA2											🟦		
71	GA2											🟦		
72	GA2											🟦		
73	GA2											🟦		
74	GA2											🟦		
75	GA2											🟦		
76	GA2											🟦		
77	GA2											🟦	🟦	
78	GA2											🟦	🟦	
79	GA2												🟦	
80	GA2												🟦	
81	GA2												🟦	
82	GA2												🟦	
83	GA2												🟦	
84	GA2													🟦
85	GA2													🟦
86	GA2													🟦
87	ON1												🟩	
88	ON1												🟩	🟩
89	ON1												🟩	🟩

FIG 4 Temporal occurrence patterns of the 89 RSV-A variants (rows) detected within RSV epidemic seasons in Kilifi Kenya, 2000-2001 to 2012-2013. A variant was defined as a group of viruses (where group includes a singleton) within a genotype that possesses ≥4 nucleotide differences in the G ectodomain region compared to other viruses (see Materials and Methods). The genotypes are shown in different colors: GA2 (blue), GA5 (red), GA3 (black), and ON1 (green).

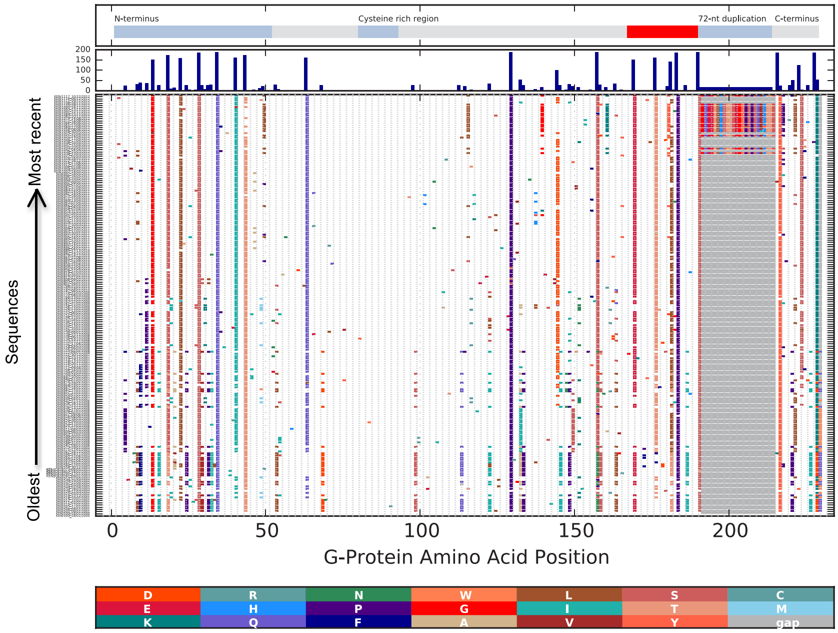


FIG 5 Changes in partial RSV-A G-protein for sequences isolated in Kilifi Kenya from 2000-2001 to 2012-2013. All unique protein sequences per epidemic were collated, aligned and the amino acid differences from the earliest sequence determined and marked with vertical colored bars, with the new amino acid residue color coded, as shown at the bottom. A gray bar indicates a gap in the query sequence. Sequences are ordered by sample collection date, with the earliest samples at the bottom of the graph. Indicated at the top of the graph are the functional domains of the G protein and the 72-nucleotide duplication of genotype ON1 (original sequence in "red" and duplicate in "light-blue steel"). The histogram in the second panel from the top indicates the total number of changes at each position.

main GA5 branch. Most of the global viruses, as was the case in Kilifi, were of genotype GA2.

DISCUSSION

The mechanisms underlying recurrent RSV epidemics have long been studied but are still poorly understood. Comparative analysis of time-stamped molecular sequence data is essential for reconstructing the genetic history of RSV and inferring the nature and extent of forces shaping the evolution of the virus that may play a role in the recurrent epidemics. This report details the genetic diversity and molecular evolution of RSV group A viruses through a long-term surveillance of RSV at the Kenyan coast, a tropical, largely rural developing country setting.

As reported elsewhere (47, 48), we observed a change from multiple group A genotype circulation in Kilifi to predominance of genotype GA2 over the 12 years studied. Nonetheless, there was

TABLE 2 Patterns of potential *N*-glycosylation sites for RSV-A sequence data from Kilifi, Kenya: genotypes GA2 and ON1

	Codon position[a]								
Designation code	103	135	197	237	246	251	273	294	Epidemic(s) observed
2A	+	+	–	+	–	+	–	+	2000-1, 2001-2, 2005-6, 2006-7, 2010-11, 2011-12
2B	+	+	–	+	–	+	+	+	2004-5, 2005-6, 2006-7
2C	–	+	–	+	–	+	+	+	2005-6
2D	+	+	–	+	+	+	–	+	2006-7, 2007-8, 2008-9
2E	+	+	–	–	–	+	+	+	2008-9, 2009-10, 2010-11, 2011-12, 2012-13
2F	–	+	–	–	–	+	+	+	2006-7, 2007-8, 2008-9
2G	–	+	–	–	–	+	–	+	2008-9
2H	–	+	–	–	–	–	–	+	2007-8
2I	+	+	–	–	+	+	–	+	2008-9
2J	+	+	–	–	–	+	–	+	2009-10, 2010-11
2K	+	+	–	–	–	–	+	+	2009-10
2L	+	–	–	–	–	+	+	+	2010-11
2 M	+	–	–	+	–	+	+	+	2010-11
2N	+	+	+	–	–	+	+	+	2010-11
2O	+	+	–	–	–	+	+	–	2011-12

[a] +, Position was detected as potentially *N*-glycosylated; –, position was not detected as *N*-glycosylated.

TABLE 3 Patterns of potential N-glycosylation sites for RSV-A sequence data from Kilifi, Kenya: genotype GA5

	Codon position[a]							
Designation code	103	105	238	242	250	273	294	Epidemic(s) observed
5A	+	+	+	–	+	+	+	2000-01, 2001-02, 2002-03-04, 2004-05, 2006-07
5B	+	+	+	–	–	+	+	2006-07
5C	+	+	+	+	+	+	+	2007-08

[a] +, Position was detected as potentially *N*-glycosylated; –, position was not detected as *N*-glycosylated.

continued diversification within the dominant genotype GA2 evidenced via (i) increased number of distinct variants over time within this genotype and (ii) proposed designation of some groups within GA2 into new genotypes, such as NA1 and NA2 (49). There appears to be a replacement period of around 7 years for some genotypes to appear, predominate, and then become undetectable—in this case GA5, then GA2, and then ON1. It could be the period it takes to build up sufficient herd immunity to a genotype through infection across all ages and thus enable replacement (at the population level) by a less-well-recognized genotype.

Phylogenetic analysis revealed ML trees with clusters of viruses derived from single and multiple epidemics, i.e., some variants were observed within a single epidemic while others were detected across several epidemics (2 to 5 years). However, persistent variants were very few comprising only 22.5% of the total number of variants assigned. It was also clear from the ML trees that individual epidemics were driven by multiple variants characterized by strains from the same epidemic placed in multiple clusters. Only two variants were seen to reappear after nondetection in consecutive epidemic seasons. Similar observations of multiple variants seeding epidemics, with little persistence between consecutive epidemics and limited variant reemergence have been reported for group B RSV from Kilifi (9). Although it remains to be understood whether multiple variants is an absolute necessity for or an occurrence of community epidemics, it is also characteristic for other seasonal respiratory viruses such as influenza (50, 51).

Similarly, the global phylogeny showed that most countries had multiple cocirculating sequence clusters from a single year. The RSV-A viruses from Kilifi often clustered closest to viruses from other regions around Kenya, which points to within-country transmission playing a pivotal role in local RSV persistence. At the global scale, however, Kilifi viruses clustered closest with viruses from Europe. Since Kilifi lies on the Kenyan coast and is a popular tourism destination especially for visitors from Europe, these findings are consistent with an argument that Europe could be a direct source of new RSV variants into Kilifi and for Kenya in general. However, more extensive sampling regionally around Kenya will point to the most probable source(s) of variants into Kilifi as there are limited sequences from Africa (only sequences from Kenya and South Africa were available in this analysis). Lastly, with many clusters on the tree not having representatives from Kilifi or Kenya, it is evident that there are many RSV-A variants circulating globally but not observed locally.

While the evolutionary rate of a virus provides a clue on its adaptive potential in a new host, the demographic history highlights changes in its relative genetic diversity over time (52). The changes in relative dynamic diversity were characterized by expansions and contractions coinciding with RSV cases peaks and troughs, respectively. Further, the major peaks in genetic diversity coincided with the epidemics that had the highest number of RSV A variants. The fact that the demographic history could capture changes in population size at fine temporal resolution implies sufficient sampling density. The mean substitution rate for the Kilifi viruses together with contemporaneous global data set was

TABLE 4 Parameter estimates, *dN/dS*, log likelihood (ℓ), and positive selection sites for RSV-A sequence data from Kilifi, Kenya

Genotype and model	*dN/dS*[a]	Parameter estimates	Positively selected sites[b]	Log likelihood (ℓ)
GA2 and ON1				
M0	0.655	$\omega = 0.655$	None	−3,681.07
M1a	0.603	$p_0 = 0.502, p_1 = 0.498; \omega_0 = 0.21, \omega_1 = 1.00$	Not allowed	−3,661.87
M2a	0.691	$p_0 = 0.496, p_1 = 0.48, p_2 = 0.03; \omega_0 = ,0.23$ $\omega_1 = 1.00, \omega_2 = 3.67$	**274L**	−3,656.22
M3	0.712	$p_0 = 0.57, p_1 = 0.41, p_2 = 0.02; \omega_0 = 0.27,$ $\omega_1 = 1.15, \omega_2 = 4.02$	101F, 104L, 106G, 115L, 147T, **225V**, 237D, 244R, 248L, 250S, 273N, **274L**, 275S, **286L (ON1: 310)**, **289P (ON1: 313)**	−3,661.99
GA5				
M0	0.547	$\omega = 0.547$	None	−1,261.92
M1a	0.547	$p_0 = 0.99, p_1 = 0.01; \omega_0 = 0, \omega_1 = 1$	Not allowed	−1,261.92
M2a	0.547	$p_0 = 1.00, p_1 = 0.00, p_2 = 0.00; \omega_0 = 0.55,$ $\omega_1 = 1.00, \omega_2 = 1.00$	None	−1,261.92
M3	0.547	$p_0 = 0.08, p_1 = 0.92, p_2 = 0.00; \omega_0 = 0.00,$ $\omega_1 = 0.55, \omega_2 = 19.09$	None	−1,261.92

[a] This *dN/dS* ratio is an average over all sites in the RSV-A alignment.

[b] The positively selected sites indicated in boldface are posterior probabilities >0.99. The positions are relative to the RSV-A reference strain (M74568).

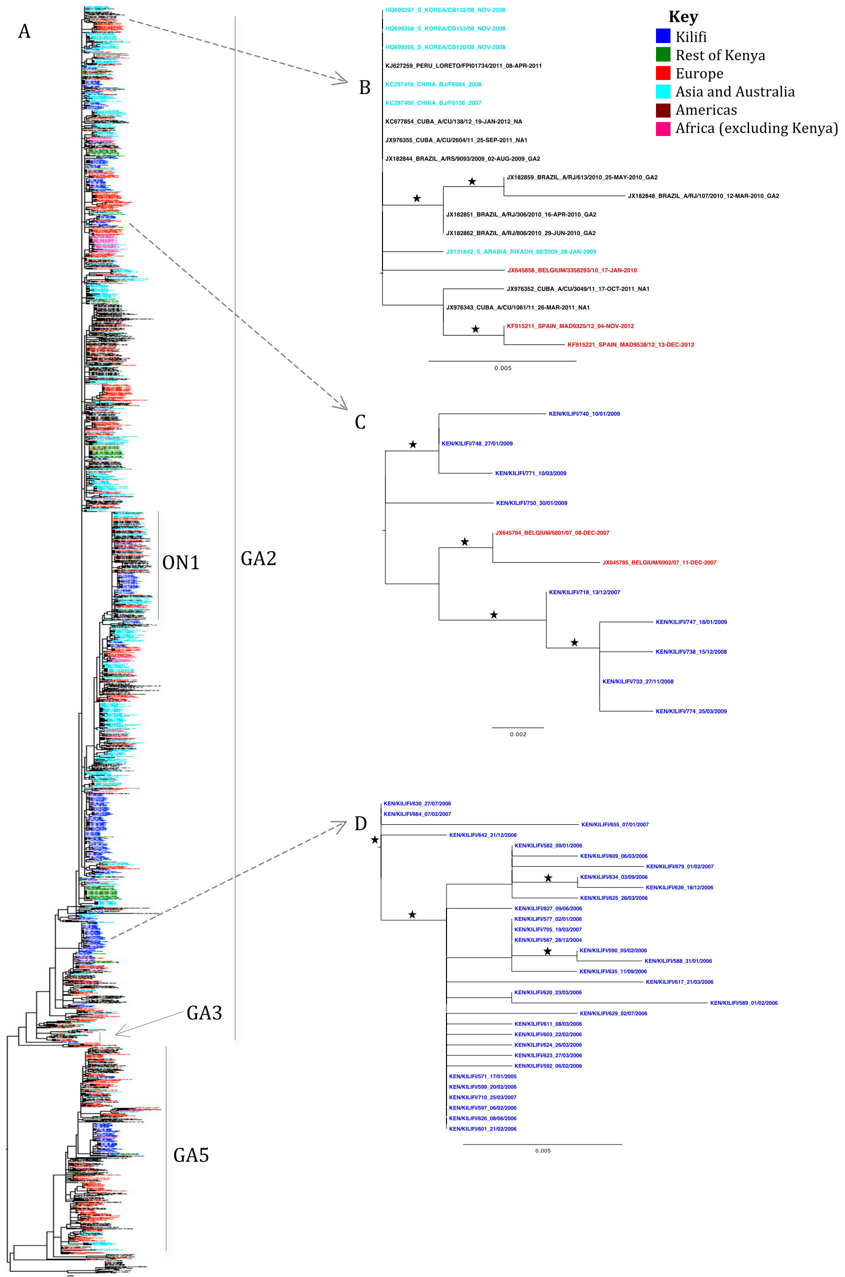

A
B
C
D
ON1
GA2
GA3
GA5
Key
🟦 Kilifi
🟩 Rest of Kenya
🟥 Europe
🟦 Asia and Australia
🟫 Americas
🟪 Africa (excluding Kenya)
HQ699267_S_KOREA/CB132/08_NOV-2008
HQ699268_S_KOREA/CB153/08_NOV-2008
HQ699266_S_KOREA/CB120/08_NOV-2008
KJ627259_PERU_LORETO/FPI01734/2011_08-APR-2011
KC297416_CHINA_BJ/F6984_2008
KC297400_CHINA_BJ/F6136_2007
KC677854_CUBA_A/CU/138/12_19-JAN-2012_NA
JX976355_CUBA_A/CU/2604/11_25-SEP-2011_NA1
JX182844_BRAZIL_A/RS/9093/2009_02-AUG-2009_GA2
JX182859_BRAZIL_A/RJ/613/2010_25-MAY-2010_GA2
JX182848_BRAZIL_A/RJ/107/2010_12-MAR-2010_GA2
JX182851_BRAZIL_A/RJ/306/2010_16-APR-2010_GA2
JX182862_BRAZIL_A/RJ/806/2010_29-JUN-2010_GA2
JX131642_S_ARABIA_RIYADH_88/2009_28-JAN-2009
JX645858_BELGIUM/3358293/10_17-JAN-2010
JX976352_CUBA_A/CU/3049/11_17-OCT-2011_NA1
JX976343_CUBA_A/CU/1061/11_26-MAR-2011_NA1
KF915211_SPAIN_MAD9325/12_04-NOV-2012
KF915221_SPAIN_MAD9538/12_13-DEC-2012
0.005
KEN/KILIFI/740_10/01/2009
KEN/KILIFI/748_27/01/2009
KEN/KILIFI/771_10/03/2009
KEN/KILIFI/750_30/01/2009
JX645784_BELGIUM/6801/07_08-DEC-2007
JX645785_BELGIUM/6902/07_11-DEC-2007
KEN/KILIFI/718_13/12/2007
KEN/KILIFI/747_18/01/2009
KEN/KILIFI/738_15/12/2008
KEN/KILIFI/733_27/11/2008
KEN/KILIFI/774_25/03/2009
0.002
KEN/KILIFI/630_27/07/2006
KEN/KILIFI/684_07/02/2007
KEN/KILIFI/655_07/01/2007
KEN/KILIFI/642_21/12/2006
KEN/KILIFI/582_09/01/2006
KEN/KILIFI/609_06/03/2006
KEN/KILIFI/679_01/02/2007
KEN/KILIFI/634_03/09/2006
KEN/KILIFI/639_18/12/2006
KEN/KILIFI/625_28/03/2006
KEN/KILIFI/627_09/06/2006
KEN/KILIFI/577_02/01/2006
KEN/KILIFI/705_19/03/2007
KEN/KILIFI/567_28/12/2004
KEN/KILIFI/590_05/02/2006
KEN/KILIFI/588_31/01/2006
KEN/KILIFI/635_11/09/2006
KEN/KILIFI/617_21/03/2006
KEN/KILIFI/620_23/03/2006
KEN/KILIFI/589_01/02/2006
KEN/KILIFI/629_02/07/2006
KEN/KILIFI/611_08/03/2006
KEN/KILIFI/603_22/02/2006
KEN/KILIFI/624_26/03/2006
KEN/KILIFI/623_27/03/2006
KEN/KILIFI/592_06/02/2006
KEN/KILIFI/571_17/01/2005
KEN/KILIFI/599_20/02/2006
KEN/KILIFI/710_25/03/2007
KEN/KILIFI/597_06/02/2006
KEN/KILIFI/626_08/06/2006
KEN/KILIFI/601_21/02/2006
0.005

FIG 6 Phylogenetic placement of RSV-A viruses from Kilifi Kenya in the global context. (A) Global ML phylogenetic tree of 1,415 unique RSV-A G-gene sequences from 29 countries collected between 2000 and 2012. The taxon names for Kilifi sequences ($n = 284$) are in blue, while the rest of Kenya ($n = 56$) are in green. Sequences from other countries are color-coded as follows: red (Europe), cyan (Asia and Australia), black (Americas), and pink (Africa excluding Kenya). (B) Example of a cluster with variants detected abroad but not in Kilifi. (C) Example of a cluster with variants predominantly observed in Kilifi, with a few representatives from abroad. (D) Example of a cluster with variants only detected in Kilifi. Branches with bootstrap values of >60% are marked with a star (★).

estimated at 3.06×10^{-3} nucleotide substitutions/site/year, which was slightly higher than previous global estimates (1.83×10^{-3} [35]), and these differences in estimates could result from differences in the source of isolates, the length of the G sequence used and the estimation method (53). The TMRCA analysis gives an upper bound to the date of the emergence of a pathogen assuming a single initial case, and from this analysis we have estimated that the MRCA of the global RSV A strains collected between 2000 and 2012 from 29 countries existed at or before 1968, similar to a previous estimate for RSV A whole genomes collected between 2001 and 2011 (54).

In this analysis, we identified 34 codon sites with amino acid substitutions that were distinct between genotypes for the group A viruses identified within Kilifi. However, a large proportion of these sites had an amino acid shared between genotypes with only three sites (250, 297, and 298) having a unique amino acid for each genotype. The sharing pattern indicates the genetic relatedness between these genotypes; GA2 and ON1 were most closely genetically related to GA3, whereas GA5 and GA3 were the most distantly related. It is interesting that previous studies have described amino acid changes at the other detected genotype defining sites 226 and 274 for RSV isolates that have lost strain-specific and group-specific epitopes (6, 55, 56).

Specific positions in the G protein have been shown to be under adaptive evolution (35, 57). We identified 15 positively selected codon sites within the two hypervariable mucin-like regions of the G protein. All of these sites have similarly been reported as positively selected, as summarized by Trento et al. (58). Amino acid replacements in sites 225, 274, and 275 have been previously described in escape mutants, selected with specific monoclonal antibodies and in natural isolates (59–61). In addition, sites 225, 274, and 286 have been described to have codon/amino acid reversal tendencies, otherwise known as the "flip-flop" pattern. Notably, only 6.5% (15/230) of the codon sites analyzed were detected as adaptive, in agreement with an average $dN/dS<1$, despite 44.6% (308/690) of the nucleotide sites being variable. This could imply that most of the substitutions are either neutral or responsible for sustained transmission of the virus through other mechanisms.

N-glycosylation is important for viral cell-cell and cell-extracellular matrix attachment (62, 63), protein folding (64), immunological properties (65–67), and stability (68). We detected eight and seven potential *N*-glycosylation sites within genotypes GA2/ON1 and GA5 sequences, respectively, with varied patterns between different epidemics. Sites 103, 135, 251, and 294 have been similarly detected as *N*-glycosylated in previous studies (53). Interestingly, no single pattern was observed continuously for the 12 years of the study period for the GA2 and ON1 sequences, with the majority of the patterns only seen in single epidemics. Similarly, none of GA5 *N*-glycosylation patterns was observed for the entire period of GA5 detection in Kilifi. There were >8-fold more potential *O*-glycosylated sites as *N*-glycosylated sites, but with similar variation in the number and sites over successive epidemics. The variations in patterns of *O*- and *N*-glycosylation were concurrent with progressive amino acid variation within the sequences analyzed. It is thought that such a high degree of variability in amino acid composition is important for molecular adaptation and is characteristic of relaxed selective constraints within these regions (69). Changes in *N*-glycosylation sites in influenza have been reported to result in marked decrease in reactivity with human sera (70) or increased virulence in seasonal influenza (71). Previous studies in RSV have indicated that epitopes within the C terminus of G-protein fragments can be masked by glycans from reactivity to certain sera, with corresponding reactivity to their nonglycosylated forms (22). Three of the 12 *N*-glycosylated sites were detected as positively selected. It is therefore plausible that these changing patterns of *N*-glycosylation sites are suggestive of a mechanism by which the virus evades preexisting immune responses.

There are conflicting views as to the impact of the variability of the ectodomain of the G protein on its reaction with antibodies and as to whether antibody selection is involved in virus evolution. Trento et al. (58) reported a high level of conservation of G protein epitopes over time, as determined by the reactions of "strain-specific" monoclonal antibodies with recent RSV isolates, and concluded that antibody-driven selection was not significant. However, an alternative explanation to their observations is that the "strain-specific" antibodies were mostly specific to genotypes rather than strains or variants. It has previously been shown that single amino acid changes can abrogate the reaction of human convalescent-phase sera with peptides whose sequences are based on naturally occurring amino acid sequences (59). For example, the change of codon 287 from proline to leucine could completely abrogate a baby's convalescent-phase serum reaction with a peptide encompassing codons 283 to 291. In addition, all the peptides found to react with babies' convalescent-phase sera comprised regions of the G protein that had potential *N*-glycosylation sites in at least some isolates. In view of the considerable variability of potential *N*-glycosylation observed within GA2, it may be that single nucleotide changes could have profound effects on the antigenicity of the G protein.

In conclusion, our analysis shows that RSV is a continually evolving virus whose persistence in the community is driven by evolutionary factors (nucleotide and amino acid variability, variation in *N*-glycosylation site patterns and selection) and epidemiological factors (multiple strain introductions and recurrence). Similar analysis with whole-genome data will illuminate further on the genetic and epidemiological characteristics of the virus and vaccine development strategies.

ACKNOWLEDGMENTS

This study was supported by the Wellcome Trust, United Kingdom (grants 102975, 100542, 084633, and 077092).

We also thank Matthew Cotten (Wellcome Trust Sanger Institute, United Kingdom) for some of the python scripts used in these analyses. This study is published with permission from the Director of the Kenya Medical Research Institute.

FUNDING INFORMATION

This work, including the efforts of James R. Otieno, was funded by Wellcome Trust (100542). This work, including the efforts of Graham Francis Medley, Patricia A. Cane, and David James Nokes, was funded by Wellcome Trust (102975). This work, including the efforts of Graham Francis Medley, Patricia A. Cane, and David James Nokes, was funded by Wellcome Trust (084633). This work, including the efforts of David James Nokes, was funded by Wellcome Trust (077092).

All funding was received from Wellcome Trust UK.

REFERENCES

1. **Nair H, Nokes DJ, Gessner BD, Dherani M, Madhi SA, Singleton RJ, O'Brien KL, Roca A, Wright PF, Bruce N, Chandran A, Theodoratou E, Sutanto A, Sedyaningsih ER, Ngama M, Munywoki PK, Kartasasmita C, Simões EA, Rudan I, Weber MW, Campbell H.** 2010. Global burden of acute lower respiratory infections due to respiratory syncytial virus in young children: a systematic review and meta-analysis. Lancet **375:**1545–1555. http://dx.doi.org/10.1016/S0140-6736(10)60206-1.
2. **Greenough A.** 2008. Long-term pulmonary outcome in the preterm infant. Neonatology **93:**324–327. http://dx.doi.org/10.1159/000121459.
3. **Nokes DJ, Okiro EA, Ngama M, Ochola R, White LJ, Scott PD, English M, Cane PA, Medley GF.** 2008. Respiratory syncytial virus infection and disease in infants and young children observed from birth in Kilifi District, Kenya. Clin Infect Dis **46:**50–57. http://dx.doi.org/10.1086/524019.
4. **Anderson LJ, Hierholzer JC, Tsou C, Hendry RM, Fernie BF, Stone Y, McIntosh K.** 1985. Antigenic characterization of respiratory syncytial virus strains with monoclonal antibodies. J Infect Dis **151:**626–633. http://dx.doi.org/10.1093/infdis/151.4.626.
5. **Mufson MA, Orvell C, Rafnar B, Norrby E.** 1985. Two distinct subtypes of human respiratory syncytial virus. J Venereal Virol **66**(Pt 10)**:**2111–2124.
6. **Cane PA, Pringle CR.** 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.
7. **Cane PA, Matthews DA, Pringle CR.** 1994. Analysis of respiratory syn-

cytial virus strain variation in successive epidemics in one city. J Clin Microbiol **32:**1–4.

8. **Peret TC, Hall CB, Schnabel KC, Golub JA, Anderson LJ.** 1998. Circulation patterns of genetically distinct group A and B strains of human respiratory syncytial virus in a community. J Gen Virol **79**(Pt 9):2221–2229. http://dx.doi.org/10.1099/0022-1317-79-9-2221.
9. **Agoti CN, Otieno JR, Ngama M, Mwihuri AG, Medley GF, Cane PA, Nokes DJ.** 2015. Successive respiratory syncytial virus epidemics in local populations arise from multiple variant introductions providing insights into virus persistence. J Virol **89:**11630–11642. http://dx.doi.org/10.1128/JVI.01972-15.
10. **Levine S, Klaiber-Franco R, Paradiso PR.** 1987. Demonstration that glycoprotein G is the attachment protein of respiratory syncytial virus. J Gen Virol **68**(Pt 9):2521–2524. http://dx.doi.org/10.1099/0022-1317-68-9-2521.
11. **Ngwuta JO, Chen M, Modjarrad K, Joyce MG, Kanekiyo M, Kumar A, Yassine HM, Moin SM, Killikelly AM, Chuang G-Y, Druz A, Georgiev IS, Rundlet EJ, Sastry M, Stewart-Jones GBE, Yang Y, Zhang B, Nason MC, Capella C, Peeples ME, Ledgerwood JE, McLellan JS, Kwong PD, Graham BS.** 2015. Prefusion F-specific antibodies determine the magnitude of RSV neutralizing activity in human sera. Sci Transl Med **7:**309ra162. http://dx.doi.org/10.1126/scitranslmed.aac4241.
12. **Cristina J, López JA, Albó C, García-Barreno B, García J, Melero JA, Portela A.** 1990. Analysis of genetic variability in human respiratory syncytial virus by the RNase A mismatch cleavage method: subtype divergence and heterogeneity. Virology **174:**126–134. http://dx.doi.org/10.1016/0042-6822(90)90061-U.
13. **Johnson PR, Spriggs MK, Olmsted RA, Collins PL.** 1987. 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. http://dx.doi.org/10.1073/pnas.84.16.5625.
14. **Hall CB, Walsh EE, Schnabel KC, Long CE, McConnochie KM, Hildreth SW, Anderson LJ.** 1990. Occurrence of groups A and B of respiratory syncytial virus over 15 years: associated epidemiologic and clinical characteristics in hospitalized and ambulatory children. J Infect Dis **162:**1283–1290. http://dx.doi.org/10.1093/infdis/162.6.1283.
15. **Munywoki PK, Koech DC, Agoti CN, Lewa C, Cane PA, Medley GF, Nokes DJ.** 2014. The source of respiratory syncytial virus infection in infants: a household cohort study in rural Kenya. J Infect Dis **209:**1685–1692. http://dx.doi.org/10.1093/infdis/jit828.
16. **Weber MW, Mulholland EK, Greenwood BM.** 1998. Respiratory syncytial virus infection in tropical and developing countries. Trop Med Int Heal TM IH **3:**268–280. http://dx.doi.org/10.1046/j.1365-3156.1998.00213.x.
17. **Felton KJ, Pandya-Smith I, Curns AG, Fry AM, Anderson LJ KN.** 2004. Respiratory syncytial virus activity-United States, 2003-2004. MMWR Morb Mortal Wkly Rep **53:**1159–1160.
18. **Henderson FW, Collier AM, Clyde WA, Denny FW.** 1979. Respiratory-syncytial-virus infections, reinfections and immunity: a prospective, longitudinal study in young children. N Engl J Med **300:**530–534. http://dx.doi.org/10.1056/NEJM197903083001004.
19. **Hendry RM, Pierik LT, McIntosh K.** 1989. Prevalence of respiratory syncytial virus subgroups over six consecutive outbreaks: 1981-1987. J Infect Dis **160:**185–190. http://dx.doi.org/10.1093/infdis/160.2.185.
20. **Mufson MA, Belshe RB, Orvell C, Norrby E.** 1988. Respiratory syncytial virus epidemics: variable dominance of subgroups A and B strains among children, 1981-1986. J Infect Dis **157:**143–148. http://dx.doi.org/10.1093/infdis/157.1.143.
21. **Agoti CN, Mwihuri AG, Sande CJ, Onyango CO, Medley GF, Cane PA, Nokes DJ.** 2012. Genetic relatedness of infecting and reinfecting respiratory syncytial virus strains identified in a birth cohort from rural Kenya. J Infect Dis **206:**1532–1541. http://dx.doi.org/10.1093/infdis/jis570.
22. **Palomo C, Cane PA, Melero JA.** 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. http://dx.doi.org/10.1002/(SICI)1096-9071(200004)60:4<468::AID-JMV16>3.3.CO;2-5.
23. **Melero JA, García-Barreno B, Martínez I, Pringle CR, Cane PA.** 1997. Antigenic structure, evolution and immunobiology of human respiratory syncytial virus attachment (G) protein. J Gen Virol **78**(Pt 10):2411–2418. http://dx.doi.org/10.1099/0022-1317-78-10-2411.
24. **Nokes DJ, Ngama M, Bett A, Abwao J, Munywoki P, English M, Scott JAG, Cane PA, Medley GF.** 2009. Incidence and severity of respiratory syncytial virus pneumonia in rural Kenyan children identified through hospital surveillance. Clin Infect Dis **49:**1341–1349. http://dx.doi.org/10.1086/606055.
25. **Scott PD, Ochola R, Ngama M, Okiro EA, Nokes DJ, Medley GF, Cane PA.** 2004. Molecular epidemiology of respiratory syncytial virus in Kilifi District, Kenya. J Med Virol **354:**344–354.
26. **Okiro EA, Ngama M, Bett A, Nokes DJ.** 2012. The incidence and clinical burden of respiratory syncytial virus disease identified through hospital outpatient presentations in Kenyan children. PLoS One **7:**e52520. http://dx.doi.org/10.1371/journal.pone.0052520.
27. **Nokes DJ, Okiro EA, Ngama M, White LJ, Ochola R, Scott PD, Cane PA, Medley GF.** 2004. Respiratory syncytial virus epidemiology in a birth cohort from Kilifi District, Kenya: infection during the first year of life. J Infect Dis **190:**1828–1832. http://dx.doi.org/10.1086/425040.
28. **Hammitt LL, Kazungu S, Welch S, Bett A, Onyango CO, Gunson RN, Scott JAG, Nokes DJ.** 2011. Added value of an oropharyngeal swab in detection of viruses in children hospitalized with lower respiratory tract infection. J Clin Microbiol **49:**2318–2320. http://dx.doi.org/10.1128/JCM.02605-10.
29. **Bonfield JK, Smith Kf, Staden R.** 1995. A new DNA sequence assembly program. Nucleic Acids Res **23:**4992–4999. http://dx.doi.org/10.1093/nar/23.24.4992.
30. **Katoh K, Standley DM.** 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol **30:**772–780. http://dx.doi.org/10.1093/molbev/mst010.
31. **Rambaut A, Charleston M.** 2001. Molecular evolution library. University of Oxford, Oxford, United Kingdom.
32. **Tamura K, Stecher G, Peterson D, Filipski A, Kumar S.** 2013. MEGA6: molecular evolutionary genetics analysis, version 6.0. Mol Biol Evol **30:**2725–2729. http://dx.doi.org/10.1093/molbev/mst197.
33. **Librado P, Rozas J.** 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics **25:**1451–1452. http://dx.doi.org/10.1093/bioinformatics/btp187.
34. **Darriba D, Taboada GL, Doallo R, Posada D.** 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods **9:**772. http://dx.doi.org/10.1038/nmeth.2109.
35. **Zlateva KT, Lemey P, Vandamme A, Ranst Van M.** 2004. Molecular evolution and circulation patterns of human respiratory syncytial virus subgroup A: positively selected sites in the attachment G glycoprotein. J Virol **78:**4675–4683. http://dx.doi.org/10.1128/JVI.78.9.4675-4683.2004.
36. **Zlateva KT, Lemey P, Moe E, Vandamme A, Ranst Van M.** 2005. Genetic variability and molecular evolution of the human respiratory syncytial virus subgroup B attachment G protein. J Virol **79:**9157–9167. http://dx.doi.org/10.1128/JVI.79.14.9157-9167.2005.
37. **Edgar RC.** 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics **26:**2460–2461. http://dx.doi.org/10.1093/bioinformatics/btq461.
38. **Drummond AJ, Suchard MA, Xie D, Rambaut A.** 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol **29:**1969–1973. http://dx.doi.org/10.1093/molbev/mss075.
39. **Minin VN, Bloomquist EW, Suchard MA.** 2008. Smooth Skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. Mol Biol Evol **25:**1459–1471. http://dx.doi.org/10.1093/molbev/msn090.
40. **Yang Z.** 2007. PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol **24:**1586–1591. http://dx.doi.org/10.1093/molbev/msm088.
41. **Yang Z, Nielsen R, Goldman N, Pedersen AM.** 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155:**431–449.
42. **Yang Z, Wong WSW, Nielsen R.** 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. Mol Biol Evol **22:**1107–1118. http://dx.doi.org/10.1093/molbev/msi097.
43. **Gupta R, Jung E, Brunak S.** 2004. Prediction of N-glycosylation sites in human proteins. Int J Cancer **46:**203–206.
44. **Steentoft C, Vakhrushev SY, Joshi HJ, Kong Y, Vester-Christensen MB, Schjoldager KT-BG, Lavrsen K, Dabelsteen S, Pedersen NB, Marcos-Silva L, Gupta R, Paul Bennett E, Mandel U, Brunak S, Wandall HH, Levery SB, Clausen H.** 2013. Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology. EMBO J **32:**1478–1488. http://dx.doi.org/10.1038/emboj.2013.79.
45. **Eshaghi A, Duvvuri VR, Lai R, Nadarajah JT, Li A, Patel SN, Low DE, Gubbay JB.** 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. http://dx.doi.org/10.1371/journal.pone.0032807.

46. **Agoti C, Otieno J, Gitahi C, Cane P, Nokes D.** 2014. Rapid spread and diversification of respiratory syncytial virus genotype ON1, Kenya. Emerg Infect Dis **20:**950–959. http://dx.doi.org/10.3201/eid2006.131438.
47. **Trento A, Casas I, Calderon A, Garcia-Garcia ML, Calvo C, Perez-Brena P.** 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. http://dx.doi.org/10.1128/JVI.00345-10.
48. **Houspie L, Lemey P, Keyaerts E, Reijmen E, Vergote V, Vankeerberghen A, Vaeyens F, De Beenhouwer H, Van Ranst M.** 2013. Circulation of HRSV in Belgium: from multiple genotype circulation to prolonged circulation of predominant genotypes. PLoS One **8:**e60416. http://dx.doi.org/10.1371/journal.pone.0060416.
49. **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. http://dx.doi.org/10.1128/JCM.00115-09.
50. **Cheng X, Tan Y, He M, Lam TT-Y, Lu X, Viboud C, He J, Zhang S, Lu J, Wu C, Fang S, Wang X, Xie X, Ma H, Nelson MI, Kung H, Holmes EC, Cheng J.** 2013. Epidemiological dynamics and phylogeography of influenza virus in southern China. J Infect Dis **207:**106–114. http://dx.doi.org/10.1093/infdis/jis526.
51. **Baillie GJ, Galiano M, Agapow P-M, Myers R, Chiam R, Gall A, Palser AL, Watson SJ, Hedge J, Underwood A, Platt S, McLean E, Pebody RG, Rambaut A, Green J, Daniels R, Pybus OG, Kellam P, Zambon M.** 2012. Evolutionary dynamics of local pandemic H1N1/2009 influenza virus lineages revealed by whole-genome analysis. J Virol **86:**11–18. http://dx.doi.org/10.1128/JVI.05347-11.
52. **Hedge J, Lycett SJ, Rambaut A.** 2013. Real-time characterization of the molecular epidemiology of an influenza pandemic. Biol Lett **9:**20130331. http://dx.doi.org/10.1098/rsbl.2013.0331.
53. **Tan L, Coenjaerts FEJ, Houspie L, Viveen MC, van Bleek GM, Wiertz EJHJ, Martin DP, Lemey P.** 2013. The comparative genomics of human respiratory syncytial virus subgroups A and B: genetic variability and molecular evolutionary dynamics. J Virol **87:**8213–8226. http://dx.doi.org/10.1128/JVI.03278-12.
54. **Tan L, Lemey P, Houspie L, Viveen MC, Jansen NJG, van Loon AM, Wiertz E, van Bleek GM, Martin DP, Coenjaerts FE.** 2012. Genetic variability among complete human respiratory syncytial virus subgroup A genomes: bridging molecular evolutionary dynamics and epidemiology. PLoS One **7:**e51439. http://dx.doi.org/10.1371/journal.pone.0051439.
55. **Rueda P, García-Barreno B, Melero JA.** 1994. Loss of conserved cysteine residues in the attachment (G) glycoprotein of two human respiratory syncytial virus escape mutants that contain multiple A-G substitutions (hypermutations). Virology **198:**653–662. http://dx.doi.org/10.1006/viro.1994.1077.
56. **Martínez I, Dopazo J, Melero JA.** 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**(Pt 10): 2419–2429. http://dx.doi.org/10.1099/0022-1317-78-10-2419.
57. **Botosso VF, Zanotto PMDA, Ueda M, Arruda E, Gilio AE, Vieira SE, Stewien KE, Peret TCT, Jamal LF, Pardini MIDMC, Pinho JRR, Massad E, Sant'anna OA, Holmes EC, Durigon EL.** 2009. Positive selection results in frequent reversible amino acid replacements in the G protein gene of human respiratory syncytial virus. PLoS Pathog **5:**e1000254. http://dx.doi.org/10.1371/journal.ppat.1000254.
58. **Trento A, Ábrego L, Rodriguez-Fernandez R, González-Sánchez MI, González-Martínez F, Delfraro A, Pascale JM, Arbiza J, Melero JA.** 2015. Conservation of g protein epitopes in respiratory syncytial virus (group A) despite broad genetic diversity: is antibody selection involved in virus evolution? J Virol **89:**7776–7785. http://dx.doi.org/10.1128/JVI.00467-15.
59. **Cane PA.** 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. http://dx.doi.org/10.1002/(SICI)1096-9071(199704)51:4<297::AID-JMV7>3.0.CO;2-0.
60. **Martinez I, Dopazo J, Melero J.** 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. http://dx.doi.org/10.1099/0022-1317-78-10-2419.
61. **Walsh EE, Falsey AR, Sullender WM.** 1998. Monoclonal antibody neutralization escape mutants of respiratory syncytial virus with unique alterations in the attachment (G) protein. J Gen Virol **79**(Pt 3):479–487.
62. **Wagner R, Wolff T, Herwig A, Pleschka S, Klenk H-D.** 2000. Interdependence of hemagglutinin glycosylation and neuraminidase as regulators of influenza virus growth: a study by reverse genetics. J Virol **74:**6316–6323. http://dx.doi.org/10.1128/JVI.74.14.6316-6323.2000.
63. **Ohuchi M, Ohuchi R, Feldmann A, Klenk H.** 1997. Regulation of receptor binding affinity of influenza virus hemagglutinin by its carbohydrate moiety. J Virol **71:**8377–8384.
64. **Doms RW, Lamb RA, Rose JK, Helenius A.** 1993. Folding and assembly of viral membrane proteins. Virology **193:**545–562. http://dx.doi.org/10.1006/viro.1993.1164.
65. **Deng R, Wang Z, Glickman RL, Iorio RM.** 1994. Glycosylation within an antigenic site on the HN glycoprotein of Newcastle disease virus interferes with its role in the promotion of membrane fusion. Virology **204:**17–26. http://dx.doi.org/10.1006/viro.1994.1506.
66. **Garcia-Beato R, Melero JA.** 2000. The C-terminal third of human respiratory syncytial virus attachment (G) protein is partially resistant to protease digestion and is glycosylated in a cell-type-specific manner. J Gen Virol **81:**919–927. http://dx.doi.org/10.1099/0022-1317-81-4-919.
67. **García-Beato R, Martínez I, Francí C, Real FX, García-Barreno B, Melero JA.** 1996. Host cell effect upon glycosylation and antigenicity of human respiratory syncytial virus G glycoprotein. Virology **221:**301–309. http://dx.doi.org/10.1006/viro.1996.0379.
68. **Papandreou MJ, Fenouillet E.** 1997. Effect of various glycosidase treatments on the resistance of the HIV-1 envelope to degradation. FEBS Lett **406:**191–195. http://dx.doi.org/10.1016/S0014-5793(97)00273-1.
69. **Wertheim JO, Worobey M.** 2009. Relaxed selection and the evolution of RNA virus mucin-like pathogenicity factors. J Virol **83:**4690–4694. http://dx.doi.org/10.1128/JVI.02358-08.
70. **Abe Y, Takashita E, Sugawara K, Matsuzaki Y, Muraki Y, Hongo S.** 2004. Effect of the addition of oligosaccharides on the biological activities and antigenicity of influenza A/H3N2 virus hemagglutinin. J Virol **78:**9605–9611. http://dx.doi.org/10.1128/JVI.78.18.9605-9611.2004.
71. **Sun X, Jayaraman A, Maniprasad P, Raman R, Houser KV, Pappas C, Zeng H, Sasisekharan R, Katz JM, Tumpey TM.** 2013. N-linked glycosylation of the hemagglutinin protein influences virulence and antigenicity of the 1918 pandemic and seasonal H1N1 influenza A viruses. J Virol **87:**8756–8766. http://dx.doi.org/10.1128/JVI.00593-13.