Whole genome analysis of local Kenyan and global sequences unravels the epidemiological and molecular evolutionary dynamics of RSV genotype ON1 strains

Otieno JR[1#], Kamau EM[1], Oketch JW[1], Ngoi JM[1], Gichuki AM[1], Binter Š[2,3], Otieno GP[1], Ngama M[1], Agoti CN[1, 4], Cane PA[5], Kellam P[3,6], Cotten M[2,7], Lemey P[8], and Nokes DJ[1, 9]

Affiliations:

1. *Epidemiology and Demography Department, Kenya Medical Research Institute (KEMRI) – Wellcome Trust Research Programme, Kilifi, Kenya*
2. *Wellcome Trust Sanger Institute, Hinxton UK*
3. *Kymab Ltd., Babraham Research Campus, Cambridge, UK*
4. *Department of Biomedical Sciences, Pwani University, Kilifi, Kenya*
5. *Public Health England, Salisbury, United Kingdom*
6. *Department of Medicine, Division of Infectious Diseases, Imperial College London, London, UK.*
7. *Department of Viroscience, Erasmus Medical Center, Rotterdam*
8. *Department of Microbiology and Immunology, KU Leuven- University of Leuven, Leuven, Belgium*
9. *School of Life Sciences and Zeeman Instutute (SBIDER), University of Warwick, Coventry, United Kingdom*

#Corresponding author

James R. Otieno

KEMRI– Wellcome Trust Research Programme, Kilifi, Kenya

P.O. Box 230, 80108, Kilifi, Kenya

e-mail: jotieno@kemri-wellcome.org phone: +254726387202

Word count

Abstract: 286

Author Summary: 190

Main Text: 5666

Running title: RSV genotype ON1 dynamics in Kilifi and globally

Abstract

The respiratory syncytial virus (RSV) group A variant with the 72-nucleotide duplication in the G gene, genotype ON1, was first detected in Kilifi in 2012 and has almost completely replaced previously circulating genotype GA2 strains. This replacement suggests some fitness advantage of ON1 over the GA2 viruses, and might be accompanied by important genomic substitutions in ON1 viruses. Close observation of such a new virus introduction over time provides an opportunity to better understand the transmission and evolutionary dynamics of the pathogen. We have generated and analyzed 184 RSV-A whole genome sequences (WGS) from Kilifi (Kenya) collected between 2011 and 2016, the first ON1 genomes from Africa and the largest collection globally from a single location. Phylogenetic analysis indicates that RSV-A transmission into this coastal Kenya location is characterized by multiple introductions of viral lineages from diverse origins but with varied success in local transmission. We identify signature amino acid substitutions between ON1 and GA2 viruses within genes encoding the surface proteins (G, F), polymerase (L) and matrix M2-1 proteins, some of which were identified as positively selected, and thereby provide an enhanced picture of RSV-A diversity. Furthermore, five of the eleven RSV open reading frames (ORF) (i.e. G, F, L, N and P), analyzed separately, formed distinct phylogenetic clusters for the two genotypes. This might suggest that coding regions outside of the most frequently studied G ORF play a role in the adaptation of RSV to host populations with the alternative possibility that some of the substitutions are nothing more than genetic hitchhikers. Our analysis provides insight into the epidemiological processes that define RSV spread, highlights the genetic substitutions that characterize emerging strains, and demonstrates the utility of large-scale WGS in molecular epidemiological studies.

Author summary

Respiratory syncytial virus (RSV) is the leading viral cause of severe pneumonia and bronchiolitis among infants and children globally. No vaccine exists to date. The high genetic variability of this RNA virus, characterized by group (A or B), genotype (within group) and variant (within genotype) replacement in populations, may pose a challenge to effective vaccine design by enabling immune response escape. To date most sequence data exists for the highly variable G gene encoding the RSV attachment protein, and there is little globally-sampled RSV genomic data to provide a fine resolution of the epidemiology and evolutionary dynamics of the pathogen. Here we use long-term RSV surveillance in coastal Kenya to track the introduction, spread and evolution of a new RSV genotype known as ON1 (having a 72-nucleotide duplication in the G gene). We present a set of 184 RSV-A whole genomes, including 154 of RSV ON1 (the first from Africa), describe patterns of local ON1 spread and show genome-wide changes between the two major RSV-A genotypes that may define the pathogen's adaptation to the host. These findings have implications for vaccine design and improved understanding of RSV epidemiology and evolution.

Introduction

Respiratory syncytial virus (RSV) is the leading viral cause of severe pneumonia and bronchiolitis among infants and children globally (1,2). Individuals remain susceptible to RSV upper respiratory tract reinfection throughout life even though they develop immune responses following primary and secondary RSV infections in childhood (3). No licensed RSV vaccine exists, partly due to the antigenic variability in the virus (4).

The single stranded, negative sense RSV genome encodes 11 proteins of which the attachment glycoprotein (G) is the most variable and a key player of adaptive evolution of the virus (5). The rate of nucleotide substitution for the G gene encoding the attachment protein has been estimated to be 1.83×10^{-3} and 1.95×10^{-3} nucleotide (nt) substitutions/site/year for group A and B, respectively, with some variation dependent on the timescale of observation (6,7). There is evidence of immune driven selection of the G gene (4,8). Although at a lower rate of evolution than for the G gene, there is significant ongoing accumulation of substitutions across the whole genome, again dependent upon the timescale of observation (9,10). At present, there is limited analysis of the selective forces acting on genes other than for the G gene as a result of paucity of whole genome sequences (WGS), particularly from one location over a period of time spanning multiple seasons (11,12). Therefore, it is not apparent whether there are genetic signatures across the rest of the genome that might additionally inform on the adaptive mechanisms of RSV viruses following introduction into communities.

125 RSV is classified into two Groups, RSV-A and RSV-B (13), differing antigenically
126 (14), with each group further characterized into genotypes (with genotype defined as
127 a cluster of viruses each of which has greater genetic distance from viruses of any
128 other genotype compared to that between viruses of the most diverse genotype
129 (15,16)). A genotype can be further divided into (i) imported variants which show
130 greater genetic difference than expected from *in situ* diversification (17,18), and (ii)
131 local variants arising from recent introduction which subsequently diversify *in situ*
132 (without time for purifying selection from, for example inter-epidemic bottlenecks)
133 (10). Studies from our group have shown that within RSV epidemics, there is co-
134 circulation of RSV viruses belonging to different groups, genotypes and variants both
135 imported and local, (10,17,18), with the latter not clearly distinguished through partial
136 G gene sequencing. Consequently, full genome sequencing offers the opportunity to
137 differentiate introduced from persistent RSV viruses within a given location.
138
139 Two recent RSV genotypes with large duplications within the G glycoprotein, BA and
140 ON1, have been detected globally. The RSV-B BA genotype is characterized by a 60-
141 nucleotide (nt) duplication while the RSV-A ON1 genotype is characterized by a 72-
142 nucleotide duplication. Initially detected in Buenos Aires Argentina in 1999, the BA
143 genotype subsequently spread rapidly throughout the world becoming the
144 predominant group B genotype and replacing all previous circulating RSV-B
145 genotypes in certain regions (19,20). The ON1 genotype was first detected in 2010 in
146 Ontario Canada, a decade after BA, and has also spread globally (21–29). Of interest
147 is what could be driving the apparent fitness advantage of these emergent genotypes
148 over the preceding genotypes (30), and whether such insights could be mined from
149 whole genome sequences.

In this study, we sought to gain a deeper understanding of the epidemiological and evolutionary dynamics of RSV viral populations through extensive whole genome sequencing and analysis of samples collected as part of on-going surveillance studies of respiratory viruses within Kilifi, Coastal Kenya (2011-2016). This was done by monitoring the unique genotype ON1 72-nucleotide duplication tag whose temporal progression can be directly followed from when the ON1 viruses first entered the ON1 'naïve' Kilifi population. This WGS analysis advances previous work on the patterns of introduction and persistence of the ON1 variant within this community that utilized partial G gene sequences (31,32), and provides a higher resolution of the RSV genetic structure, spread and identification of variation that may be associated with molecular adaptation and apparent fitness advantages.

Materials and Methods

Study population

This study is part of ongoing surveillance of respiratory viruses within Kilifi County, coastal Kenya, and across the country that is aimed at understanding the epidemiology and disease burden of respiratory viruses in this region (33). Two sets of samples were used in the current analysis; (i) samples collected from children (under 5 years of age) admitted to the Kilifi County Hospital (KCH) presenting with syndromically defined severe or very severe pneumonia (32,33), and (ii) samples collected from patients of all ages presenting at health facilities within the Kilifi Health and Demographic Surveillance System (KHDSS) (34) with acute respiratory illnesses (ARI).

175 RNA extraction and RT-PCR

176 All KCH specimens had previously been screened for RSV, RSV group and RSV-A
177 genotype status (32), while the KHDSS samples were screened afresh. The criterion
178 for proceeding to WGS was a sample real-time PCR cycle threshold (Ct) value < 30
179 based on the success rate from previous experience (9), with the exception of four test
180 samples that were PCR negative or had Ct>30. Viral RNA was extracted using
181 QIAamp Viral RNA Mini Kit (QIAGEN). Reverse transcription (RT) of RNA
182 molecules and polymerase chain reaction (PCR) amplification were performed with a
183 six-amplicon, six-reaction strategy (9), or using a 6 or 14-amplicon strategy
184 (unpublished) split into two reactions of three and seven amplicons, respectively for
185 each, *Figure 1A*. Amplification success was confirmed by observing the expected
186 PCR product size (1200-1500 bp) on 0.6% agarose gels. For the successfully
187 amplified samples, all the six or two reactions per sample were pooled and purified
188 for Illumina library preparation.

189

190 Illumina library construction and sequencing

191 The purified PCR products were quantified using Qubit fluorimeter 2.0 (Life
192 Technologies) and normalized to 0.2 ng/μL. The normalized DNA was tagmented (a
193 process of fragmentation and tagging) using the Nextera XT (Illumina) library prep
194 kit as per the manufacturer's instructions. Indices were ligated to the tagmented DNA
195 using the Nextera XT index kit (Illumina). The barcoded libraries were then purified
196 using 0.65X Ampure Xp beads. Library quality control was carried out using the
197 Agilent high sensitivity DNA kit on the Agilent 2100 Bioanalyzer (Agilent) to
198 confirm the expected size distributions and library quality. Each library was
199 quantified using the Qubit fluorimeter 2.0 (Life Technologies), after which the

libraries were then normalized and pooled at equimolar concentrations based on the Qubit results. The pooled libraries were sequenced on either (i) Illumina HiSeq system using 2 x 250 bp PE sequencing at the Wellcome Trust Sanger Institute (UK), or (ii) Illumina MiSeq using 2 x 250 bp PE sequencing at the KEMRI-Wellcome Trust Research Programme (Kilifi, Kenya).

To determine the proportion of RSV and non-RSV reads in the samples used here, Kraken v0.10.6 (35) was used with a pre-built Kraken database provided by viral-ngs (36,37) (downloaded in December, 2015; https://storage.googleapis.com/sabeti-public/meta_dbs/kraken_ercc_db_20160718.tar.gz). A preliminary quality check of the sequence reads was done using fastqc (38) with the output per batch aggregated and visualized by multiqc (39).

Depletion of human reads

Prior to deposition of the raw short reads into NCBI short read archive (SRA), datasets were depleted of human reads. The raw reads were mapped onto the human reference genome hg19 using bowtie2 (40) while samtools (41) was used to filter, sort and recover the unmapped (non-human) reads. The final reads are available in the NCBI BioProject database under the study accession PRJNA438443.

Genome assembly and coverage

Sequence reads were taxonomically filtered within the viral-ngs pipeline using an RSV genotype ON1 reference, KC731482. The RSV reads were then used to generate consensus genome assemblies using viral-ngs versions 1.18.0 and 1.19.0 (36,37) and/or SPAdes version 3.10.1 (42), selecting the most complete assembly from either

assemblers. In addition, the available Sanger G-gene sequences (31,32) for these same samples were used to confirm agreement with the WGS assemblies. The genomes generated in this study are available in GenBank under accession numbers MH181878 - MH182061. The genomes were aligned using MAFFT alignment software v7.305 (43) using the parameters '--localpair --maxiterate *1000*'.

To calculate and visualize depth of coverage, sample raw reads were mapped onto individual assemblies with BWA (44), samtools (41) used to sort and index the aligned bam files, and finally bedtools (45) used to generate the coverage depth statistics. Plotting of the depth of coverage was done in R (46) in the RStudio (47).

Global comparison dataset

All complete and partial genome sequences available in GenBank Nucleotide database (https://www.ncbi.nlm.nih.gov/genbank/) as on 19/09/2017 were added to a global RSV-A genotype ON1 genomic and G-gene dataset. To prepare the global ON1 dataset, we downloaded all RSV sequences from GenBank (search terms: respiratory syncytial virus), created a local blast database in Geneious (48), and performed a local blast search using the 144 nucleotide sequence region of the ON1 genotype. To remove duplicates, the sequences were binned by country of sample collection, filtered of duplicates and then re-collated into a single dataset. For the global G-gene dataset of 1,167 sequences, the sequence length ranged from 238-690bp. The final alignment of 344 ON1 genome sequences comprised the sequences reported in this study (n=154) and additional publicly available GenBank ON1 sequences (n=190). In addition to the ON1 genomes, we generated 30 genotype GA2 genome sequences from Kilifi. The alignments were inspected in AliView (49) and

edited manually removing unexpected spurious frame-shift indels (largely homopolymeric and most likely sequencing errors).

Maximum likelihood phylogenetic analyses and root-to-tip regression

Separate Maximum-Likelihood (ML) phylogenetic trees were generated using multiple sequence alignments of the three datasets, i.e. Kilifi WGS, and global G-gene and WGS datasets. The ML trees were inferred using both PhyML and RaxML, with each optimizing various parts of the tree generation process (i.e. borrowing strengths of both approaches), using the script generated and deposited by Andrew Rambaut at (https://github.com/ebov/space-time/tree/master/Data/phyml_raxml_ML.sh). The GTR+G model was used after determination as the best substitution model by IQ-TREE v.1.4.2 (50).

To determine presence of temporal signal ('clockiness') in our datasets, we used TempEst v1.5 (51) to explore the relationship between root-to-tip divergence and sample dates. The data were exported to R (46) to perform a regression with the 'lm' function.

Estimating number local variant introductions

To differentiate between local variants arising from a recent introduction and imported variants with greater genetic differences than is expected from local diversification, we used a pragmatic criterion previously described by *Agoti et al.* in (17). Briefly, a variant is a virus (or a group of viruses) within a genotype that possesses $\geq x$ nucleotide differences compared to other viruses. This x nucleotide differences is a product of the length of the genomic region analyzed, estimated

substitution rate for that region, and time. This analysis was done using usearch v8.1.1861 (29).

Protein substitution and selection analysis

Using the aligned Kilifi (ON1 and GA2) genome dataset, patterns of change in nucleotides (single nucleotide polymorphisms or SNPs) and amino acids were sought using Geneious v11.1.2 (48) and BioEdit 7.2.5 (52), respectively. Potential positively selected and co-evolving sites within the coding regions were identified using HyPhy (53) and phyphy (54). SNPs were called from both the complete dataset and from an alignment of the consensus sequences from GA2 and ON1, whereby a consensus nucleotide was determined as the majority base at a given position. For the positive selection analysis, two strategies were used; gene-wide selection detection [BUSTED (55)] and site-specific selection [SLAC, FEL (56), FUBAR (57) and MEME (58)]. Codon positions with a p-value <0.1 for either the SLAC, FEL and MEME models or with a posterior of probability >0.9 for the FUBAR method were considered to be under positive selection.

Bayesian phylogenetics

To infer time-structured phylogenies, Bayesian phylogenetic analyses were performed using BEAST v.1.8.4 (59). Because of sparse data at the 5' and 3' termini and in the non-coding regions of the genomic datasets, only the coding sequences (CDS) were used as input. The SRD06 substitution model (60) was used on the CDS and three coalescent tree priors were tested, i.e. a constant-size population, an exponential growth population, and a Bayesian Skyline (61). For each of these tree priors, combinations with the strict clock model and an uncorrelated relaxed clock model

with log-normal distribution (UCLN) (62) were tested with the molecular clock rate set to use a non-informative continuous time Markov chain rate reference prior (CTMC) (63). For each of the molecular clock and coalescent model combinations, the analyses were run for 150 million Markov Chain Monte Carlo (MCMC) steps and performed both path-sampling (PS) and stepping-stone (SS) to estimate marginal likelihood (64,65). The best fitting model was a relaxed clock with a Skyline coalescent model, *Supplementary sheet 1*.

BEAST was then run with 300-400 million MCMC steps using the SRD06 substitution model, Skyline tree prior, and relaxed clock model to estimate Bayesian phylogenies. For the time to the most recent common ancestor (TMRCA) estimates, the same substitution model and tree prior were used as above but with a strict clock model. For the global G-gene dataset, BEAST was run with 400 million MCMC steps using the HKY substitution model, Skyline tree prior, and a relaxed clock model. We used Tracer v1.6 to check for convergence of MCMC chains and to summarize substitution rates. Maximum clade credibility (MCC) trees were identified using TreeAnnotator v1.8.4 after removal of 10% burn-in and then visualized in FigTree v1.4.3.

Principal component analysis

To check on any clustering and stratification patterns, principal component analysis (PCA) was performed using the R package FactoMineR (66). The input data were a matrix of pairwise distances from genome sequence alignment using the "N" model of DNA evolution, i.e. the proportion or the number of sites that differ between each

pair of sequences. Each genome on the PCA plot was annotated by the continent of sample origin.

Results

Genome sequencing and assemblies

A total of 184 RSV-A genomes were generated in this study, comprising genotypes ON1 (*n*=154) and GA2 (*n*=30), collected between February 2012 and April 2016; *Supplementary sheet 2*. This dataset included 176 genomes from inpatients at KCH and 8 genomes from peripheral health centres within the KHDSS. Between 0.2 to 4.3 million short reads were available per sample of which RSV specific reads ranged between 0.001 to 3.9 million reads. The genome assemblies had a median length of 15,054 nucleotides (range: 13,966-15,322) and mean depth of base coverage per genome ranging from 39 to 66457.

Whereas the samples for WGS were generally of high viral content (lower Ct value), it is apparent there was reduced genome yield (proportion of genome assembled) from samples with lower viral loads (i.e. higher Ct values); *Figure 1B*. However, since the samples collected at the hospital are from children presenting with severe or very severe pneumonia cases they generally have high viral loads as shown in *Figure 1C*. The median fraction of the genome with unambiguous base calls was 98% with reference length from KC731482. Read coverage across the genomes was non-uniform, *Figure 1D,* suggesting varied PCR amplification efficiency among primer pair combinations combined with increased sequencing yield from the ends of the amplicons.

348

349 *Bayesian reconstruction of ON1 epidemiological and evolutionary history*

350 The global ON1 whole-genome MCC phylogenetic tree, *Figure 2A*, shows
351 evolutionary relationship among ON1 viruses from five sampled continents. The
352 TMRCA of the ON1 strains from the most recent tip (7 April 2016) was estimated to
353 be 11.07 years [95% HPD: 9.85-12.31], resulting in an estimated ON1 emergence
354 date of between December 2003 and June 2006. This estimated date of emergence is
355 earlier than a previous estimate (2008-2009) using the G-gene alone (29), but such a
356 difference could ideally be a reflection of the different datasets (by geography and
357 sampling dates). *Comas-Garcia et al.* have reported the earliest ON1 strain identified
358 to date in November 2009 from central Mexico (67), and from our estimates this
359 suggests a period of 3-6 years of circulation of this virus before first detection. The
360 genome-wide substitution rate for the ON1 viruses was estimated at 5.97 X 10^{-4}
361 nucleotide substitutions per site per year [95% HPD: 5.42-6.58], similar to previous
362 estimates for RSV group A full length sequences sampled over several epidemics
363 (9,12) but slower than estimates from both samples collected from a household study
364 over a single epidemic within the same location and using a global ON1 G-gene
365 dataset (10,29). Across the genome, estimates of evolutionary rates for individual
366 ON1 open reading frames (ORFs) varied, *Figure 2B*, with the mean substitution rate
367 highest in the G-gene, lowest in NS1, and moderate (with tight 95% HPD intervals)
368 for the whole genome.

369

370 The Kilifi ON1 genomes formed three distinct lineages (labelled LI-LIII) on the
371 global tree in *Figure 2A*. These three ON1 lineages were however placed into two
372 clades within the Kilifi WGS MCC tree, *Figures 2C* whereby lineages LI and LIII

were placed in clade A while clade B comprised only lineage LII sequences. The two clades were temporal with clade A mostly comprising sequences from the 2011-2013 RSV epidemic period and clade B comprising sequences from the epidemic period 2013-2016. These clade and temporal patterns are further highlighted by the PCA analysis in *Figure 2D*. Based on the phylogenetic placement of the Kilifi ON1 lineages on the global tree, we estimate that there could have been at least three separate introductions of ON1 viruses into Kilifi. For lineage LIII, we sampled only 4 cases, and this is consistent with limited local transmission. In addition, all the eight outpatient ON1 viruses collected outside KCH were placed within lineage LII and were interspersed with viruses sampled from inpatient admissions at KCH implying that our sampling at the hospital might be well representative of the KHDSS community.

Using the global whole genome ON1 substitution rate estimate above, the Kilifi ON1 genomes dataset (length 15,404 bp) and a pragmatic criterion previously described by *Agoti et al.* in (17) to differentiate between local and imported variants, we estimated that there were up to 73 ON1 introductions into Kilifi. Even when we used the higher substitution rate previously estimated from ON1 partial G-gene sequences by *Duvvuri et al* in (29), i.e. 4.10×10^{-3} substitutions/site/year which translates to a difference of at least 63 nucleotides between any two genomes to be classified as separate introductions, this resulted in an estimate of 6 separate introductions. This implies that multiple seeding introductions of viruses within lineages LI and LII may have been required to sustain their local transmission.

Global ON1 spatiotemporal dynamics

As there are far more partial G gene sequences than full genomes, we explored ON1 spatiotemporal patterns using a set of 1,167 global G gene sequences. The global G gene MCC tree is shown in *Figure 3* with the corresponding sampling locations in *Supplementary figure 1*. Although they do not correspond to well-supported monophyletic clades, we classify the clusters in C1, C2 and C3 for convenience. While there was neither a single cluster that was comprised solely of viruses from a specific continent nor a continent whose viruses were only found within a single cluster, there was predominance of African/European viruses in cluster C1, Asian in C2 and European/Asian in C3, suggesting both intra and inter-continental circulation patterns. The majority of the Kilifi ON1 lineages (LI and LII in *Figure 2A*) were found in cluster C1, suggesting perhaps a predominantly European source of RSV introductions into Kilifi, while the lineage LIII viruses were found in cluster C3. All the African viruses in cluster C1 were from Kilifi while all the Nigerian and all the South African viruses were only found in clusters C2 and C3 respectively. Further, the viruses closely related to the ON1 lineage (LIII) with limited transmission in Kilifi described above were frequently isolated in other locations (cluster C3).

Genomic diversity of Kilifi RSV-A viruses

Pairwise intra-genotypic genetic diversity analysis of the GA2 and ON1 genomes from Kilifi, *Figure 4A* and *4B*, show unimodal and bimodal distributions respectively consistent with two genetically distinct circulating strains of the ON1 viruses. *Figure 4C* shows an entropy plot with protein substitution density based on amino acid polymorphisms for a concatenated set of all RSV proteins from both Kilifi genotype ON1 and GA2 viruses. Analyzing for substitutions across the genomes, we identified a total of 746 single nucleotide polymorphisms (SNPs) with frequencies of >1% (in

the set of 184 genomes). Of these SNPs, the majority (589, 78.9%) were found within coding sequences/regions (CDS) with only 145/589 (24.6%) of these coding mutations resulting in non-synonymous changes, *Supplementary sheet 3*. The three CDSs with the most substitutions were the polymerase L (39.6%), the glycoprotein G (14.8%) and the fusion F (14.6%). However, most of the non-synonymous changes occurred within G, SH and M2-2.

Is the 72-nucleotide duplication a red-herring or masking the complete genomic identity of ON1 viruses?

The currently known or *de facto* distinguishing feature of the ON1 from GA2 strains is the 72-nucleotide duplication within the G gene. It has been shown from phylogenetic analysis of the G-gene that RSV group A genotypes form distinct clusters (16). However, it has not been investigated if the distinct clustering is replicated in the other genes especially for the closely related genotypes GA2 and ON1 viruses. Through an exploratory root-to-tip regression analysis of ORF specific ML trees, we confirmed that all but the NS1, NS2 and SH proteins had good temporal signals to proceed with this analysis, *Supplementary figure 2*. Our observations using the Kilifi GA2 and ON1 WGS dataset indicate that this phylogenetic divergence is present in the concatenated set of all the 11 RSV-A ORFs (*Figure 5A*) and five individual coding regions (G, F, L, N and P), *Supplementary figure 3*. However, node posterior support for divergence between GA2 and ON1 in the N and P proteins was quite low (50-70%) despite observation of distinct clusters. The date of the MRCA of the Kilifi ON1 strains was estimated to June 2010 [95% HPD: November 2009-November 2010], implying a lapse of at least one year before initial detection of genotype ON1 in Kilifi in February 2012.

448

449 Based on the results above indicating evolutionary divergence across the five ON1
450 and GA2 proteins, we asked the question; Is the 72-nucleotide duplication the only
451 marker of the ON1 strains or an accompanying mutation? To answer this question, we
452 analyzed a concatenated set of 10 RSV ORFs (excluding the G) whereby we observed
453 distinct and well supported ON1 and GA2 clusters indicating the presence of genetic
454 markers outside the 72-nucleotide duplication and the G ORF that differentiate
455 viruses belonging to these two genotypes. With the possibility of additional
456 substitutions between GA2 and ON1 viruses across the genome and assuming a single
457 point source of ON1 viruses, we hypothesized two likely scenarios (i) a single ON1-
458 GA2 split event in which the founder ON1 virus possessed distinctive substitutions
459 across the five RSV ORFs, or (ii) progressive but rapid accumulation of substitutions
460 between ON1 and GA2 viruses within the five ORFs.

461

462 With regard to scenario (ii) above, it would be important to know what could have
463 come first; the 72-nucleotide duplication in the G or the changes in the other ORFs?
464 However, considering that these substitutions could have happened anywhere on the
465 branch between the GA2-ON1 split time and ON1 TMRCA in *Figure 5A*, it is
466 impossible to distinguish the order of such substitutions. This dilemma is confirmed
467 by the overlapping intervals in the divergence timings of the individual ORFs in
468 *Figure 5B*.

469

470 *Identification of signature substitutions differentiating ON1 from GA2 viruses*

471 The presence of phylogenetic divergence between five ORFs of ON1 and GA2
472 viruses above indicates potential SNPs between these two strains. Through a

473 comparative genome-wide scan along the RSV-A coding genome, we sought to pick
474 out SNPs between the consensus sequences of the Kilifi ON1 and GA2 viruses. We
475 identified 66 signature nucleotide substitutions, i.e. where a signature substitution was
476 defined as an SNP differentiating ON1 from GA2 viruses, *Supplementary sheet 4*.
477 While most of these signature substitutions were synonymous, we found 14 non-
478 synonymous substitutions between the ON1 from GA2 viruses, *Table 1*, of which
479 nine substitutions were in the G protein, two each in the F and L proteins, and one in
480 the M2-1 protein. None of these signature substitutions were observed to have an
481 effect on our RSV multiplex PCR diagnostics as they occur outside the target primer
482 binding sites in the N gene. Changes at the codon sites 142 and 237 of the G protein
483 that had these signature substitutions have previously been shown to characterize
484 antibody escape mutants, and were located within strain-specific epitopes (68). The
485 two signature substitutions in the F protein (116 and 122) occur within site p27, which
486 is the most variable antigenic site in the F protein among RSV group A and B
487 genotypes (69). However, determining the effect of these signature AA substitutions
488 on potential fitness and phenotype differences between ON1 and GA2 strains will
489 require targeted functional assays.
490
491 It should be noted that the signature ON1-GA2 SNPs above were called from
492 consensus genome sequences of each genotype, whereby a consensus nucleotide was
493 a simple majority at a given position. However, five nucleotide positions with the
494 identified signature polymorphisms resulting in non-synonymous substitutions had
495 100% nucleotide consensus for GA2 and ON1 sequences and are of great interest, i.e.
496 codon positions 232, 253, 274 and 314 in the G protein and 598 in the L protein
497 (respective nucleotide positions in *Table 1*). Additionally, we observed the following

498 substitutions in ON1 genomes (data not shown) that might also be of interest; (i) ON1
499 viruses undergoing convergent evolution at particular sites by acquiring nt/AA
500 substitutions similar to GA2 viruses in recent epidemics (2014-2016), and (ii) ON1
501 viruses undergoing further diversification through acquisition of nt and/or AA that is
502 different from GA2 in recent epidemics whereas these two genotypes possessed
503 different or shared similar nt/AA in earlier epidemics.
504
505 *Signature substitutions between lineages with successful and limited local*
506 *transmission*
507 In an attempt to unravel the molecular basis of the ON1 lineages in Kilifi with varied
508 local transmission outcomes, we performed a similar genome-wide comparative scan
509 between the consensus of genomes of viruses with successful (lineages LI and LII)
510 and those with limited local transmission (lineage LIII) for characteristic signature
511 polymorphisms. We identified 33 SNPs between these two groups of lineages,
512 *supplementary sheet 3*, of which nine resulted in non-synonymous changes; five in G,
513 two in F and one each in M2-2 and L. In three of these nine SNPs with non-
514 synonymous substitutions between the two groups of lineages, lineage LIII possessed
515 similar nucleotides as the GA2 viruses (G: codons P274L and P310L, and F: codon
516 A122T). Whether these polymorphisms that characterize the two groups of lineages
517 are neutral mutations or influence local transmission of the virus warrants further
518 investigation.
519
520 *Patterns of selective pressure*
521 It is expected that different codon sites in genes would be under differential
522 evolutionary pressures. We conducted selection analysis on all 11 RSV ORFs for the

dataset. ORF-wide episodic diversifying selection was only detected in the NS1 and M proteins. A total of nine positively selected codon sites were identified within the G (73, 201, 250, 251, 273, 310), NS2 (15) and the L (2030, 2122) by at least one method. Notably, sites 273 and 310 within the G protein detected to be under positive selection also had the signature SNPs described above between ON1 and GA2 viruses. However, the number of positively selected sites could have been underestimated in the analysis that was limited to Kilifi RSV-A genomes and care should be taken while interpreting these results as some of the positively selected sites were only detected by one method and at default (less stringent) cut-offs.

Discussion

Here we report an in-depth analysis of local and global RSV genotype ON1 evolution and transmission using whole genome sequence data. We describe RSV-A genomic diversity and identify polymorphisms with the most potential in influencing RSV evolution and phenotype. Utilizing genomes from samples collected between 2010–2016, including 184 complete genomes from Kilifi alone, we obtained a finer resolution on the pattern of RSV introductions, persistence and evolution in a defined location, and the changes within the genome that might be important for the survival of the virus.

Genetic variation not only provides important insights into RSV relatedness by which to infer transmission but also highlights potential functional changes in the virus. From our analysis, we find that substitutions are widespread across the RSV genome but occur at higher frequency within the structural proteins (G and F) and in parts of the polymerase (L). The G protein has the most genetic flexibility of the RSV

548 ORFs to accommodate frequent substitutions including large duplications, and
549 previous studies have described epitope positions within this protein that characterize
550 escape mutants selected by specific monoclonal antibodies or by natural isolates
551 (5,68,70,71). Site p27 in the F protein with two signature substitutions has been
552 shown to possess greater binding activity in sera from young children (<2 years) than
553 any of the other antigenic sites in the F protein and may be responsible for group
554 specific immunity due to its great variability between RSV-A and RSV-B viruses
555 (72). However, the implications of the substitutions in the L protein of the ON1
556 viruses remain unclear, but considering both its role in genome replication and the
557 presence of the 72-nucleotide duplication in the G ORF, we posit that either (i) these
558 polymorphisms might have resulted in a sloppy polymerase, that further resulted in a
559 slip that generated the 72-nucleotide duplication in the G ORF (73), or (ii) the 72-
560 nucleotide duplication in the G presented a larger metabolic challenge in replicating a
561 larger genome and thereby facilitating adaptive polymorphisms within the polymerase
562 (74). While we also found a considerable number of SNPs in other ORFs other than
563 the G, F and L proteins, only a very minor proportion of those changes resulted in
564 amino acid substitutions implying very strong purifying selection in these portions of
565 the genome.
566
567 Based on distinct phylogenetic clustering of ON1 and GA2 viruses in five ORFs, the
568 emergence of ON1 may be characterized by additional substitutions across the
569 genome in addition to the 72-nucleotide duplication within the G gene. However,
570 assuming ON1 diverged from GA2 and through a single ancestral virus, it is unclear
571 whether the multiple signature substitutions differentiating ON1 from GA2 viruses all
572 arose from that single split event or have been acquired progressively over time. In

573 case of the latter, it is unclear the chronology of the changes across the different
574 ORFs. Understanding how and which mutations define the emergence of a new RSV
575 variant may be important in describing substitutions that are either crucial for the
576 survival of the variant and/or of some complementary structural or functional
577 integrity. It is also likely that some of these substitutions are nothing more than
578 genetic hitchhikers. Notwithstanding this lack of clarity on ON1 emergence, it has
579 been shown for influenza A viruses that linked selection amongst antigenic and non-
580 antigenic genes influences the evolutionary dynamics of novel antigenic variants (75).
581 Further, it has been demonstrated experimentally that adaptive evolution is a multi-
582 step process that occurs in waves (76). There is an initial adaptive wave that occurs
583 rapidly and is characterized by founder or gatekeeper mutations. Thereafter,
584 additional waves of evolutionary fine-tuning occur (77). Similar studies in RSV
585 would be important in determining if such dynamics do characterize their
586 evolutionary history and also might inform the design of an RSV vaccine.
587
588 Undoubtedly, ON1 is rapidly replacing GA2 in Kilifi suggesting that this variant has
589 some fitness advantage. We previously showed that genotype ON1 viruses did not
590 result in more severe disease compared to GA2 viruses in Kilifi (32). However, there
591 are conflicting reports globally with some indicating that ON1 is more virulent and
592 others reporting ON1 being less virulent than GA2 (78,79). Even with the discordant
593 results, which may also be due to differences in study populations and analysis
594 methods, there might be phenotypic differences between viruses belonging to these
595 two genotypes. Identification of such phenotypic differences and the potential drivers
596 might augment our current understanding of the pathogenesis of this virus.
597

598 Observations from this study using whole genomes reinforce previous findings based
599 on partial G-gene sequences (17,18,22,32) that RSV epidemics are characterized by
600 the introduction and circulation of multiple variants. In addition, persistence within
601 the community seems to be sustained by only a proportion of these introductions. We
602 have characterized genomic substitutions that distinguish between successful and
603 dead-end ON1 lineages in Kilifi. Nonetheless, it is evident that besides viral genetic
604 factors there could be other determinants of successful onward transmission of a virus
605 lineage considering that the non-persistent ON1 strains in Kilifi were abundant in
606 other parts of the world albeit with varied frequencies relative to other genotypes.
607 Such determinants warrant further investigations and could include the host factors
608 (e.g. births, immunity, genetics, contact patterns and mobility) and environmental
609 factors (e.g. temperature, rainfall and humidity).
610
611 We live in times of rapid global movement of people, which may influence the spread
612 of infectious diseases. The observation that_most of the Kilifi sequences clustered
613 with sequences from Europe and Asia suggests that RSV introductions into Kilifi
614 originate predominantly from these two continents. It might not be surprising that
615 Europe is a primary source of RSV introduction into Kilifi, or even a destination for
616 viruses from Kilifi, considering that it accounts for the largest single group of tourists
617 to Kenya (80). In addition, the increasing Chinese economic interests in Africa
618 (including Kenya) in becoming Africa's largest trade partner has resulted in an influx
619 of Chinese into Africa for trade, work and tourism (81). However, there are far too
620 few partial ON1 sequences from Africa (only from Kenya, South Africa and Nigeria)
621 and no ON1 genomes from outside Kilifi Kenya to help define intra-African spread
622 dynamics in detail (which we hypothesize might be more impactful on the many local

623 introductions). In fact, a recent study suggests that in the recent past domestic tourism
624 accounts for more than half of the growth in Kenya's tourism (82). As such,
625 availability of sequences from across the country would be critical in deciphering if
626 and how such tourist activities influence virus transmission patterns in Kenya.
627 Accordingly, such observations could be helpful in the design of future RSV
628 transmission intervention strategies.
629
630
631

632 **References**

633 1. Nair H, Nokes DJ, Gessner BD, Dherani M, Madhi SA, Singleton RJ, et al.
634 Global burden of acute lower respiratory infections due to respiratory syncytial
635 virus in young children: a systematic review and meta-analysis. Lancet
636 [Internet]. Lancet Publishing Group; 2010 May 1 [cited 2012 Apr
637 16];375(9725):1545–55. Available from:
638 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2864404&tool=pm
639 centrez&rendertype=abstract

640 2. Nokes DJ, Okiro EA, Ngama M, Ochola R, White LJ, Scott PD, et al.
641 Respiratory syncytial virus infection and disease in infants and young children
642 observed from birth in Kilifi District, Kenya. Clin Infect Dis [Internet]. 2008
643 Jan 1 [cited 2012 Mar 22];46(1):50–7. Available from:
644 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2358944&tool=pm
645 centrez&rendertype=abstract

646 3. Agoti CN, Mwihuri AG, Sande CJ, Onyango CO, Medley GF, Cane PA, et al.

Genetic relatedness of infecting and reinfecting respiratory syncytial virus strains identified in a birth cohort from rural Kenya. J Infect Dis [Internet]. 2012 Nov 15 [cited 2013 Aug 7];206(10):1532–41. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3475639&tool=pmcentrez&rendertype=abstract

4. Cane PA. Molecular epidemiology of respiratory syncytial virus. Rev Med Virol [Internet]. 2001 Jan [cited 2014 May 9];11(2):103–16. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11262529

5. Cane PA, Pringle CR. Evolution of subgroup A respiratory syncytial virus: evidence for progressive accumulation of amino acid changes in the attachment protein. J Virol [Internet]. 1995 May 1 [cited 2014 Apr 9];69(5):2918–25. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=188990&tool=pmcentrez&rendertype=abstract

6. Zlateva KT, Lemey P, Vandamme A, Van Ranst M. Molecular Evolution and Circulation Patterns of Human Respiratory Syncytial Virus Subgroup A : Positively Selected Sites in the Attachment G Glycoprotein Molecular Evolution and Circulation Patterns of Human Respiratory Syncytial Virus Subgroup A : Positi. J Virol. 2004;78(9):4675–83.

7. Zlateva KT, Lemey P, Moës E, Ranst M Van, Moe E. Genetic Variability and Molecular Evolution of the Human Respiratory Syncytial Virus Subgroup B Attachment G Protein Genetic Variability and Molecular Evolution of the Human Respiratory Syncytial Virus Subgroup B Attachment G Protein. J Virol. 2005;79(14):9157.

671 8. Sullender WM. Respiratory syncytial virus genetic and antigenic diversity. Clin
672 Microbiol Rev. 2000 Jan;13(1):1–15, table of contents.

673 9. Agoti CN, Otieno JR, Munywoki PK, Mwihuri AG, Cane PA, Nokes DJ, et al.
674 Local evolutionary patterns of human respiratory syncytial virus derived from
675 whole-genome sequencing. J Virol [Internet]. 2015 Jan 21 [cited 2015 Feb
676 5];89(7):3444–54. Available from:
677 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4403408&tool=pm
678 centrez&rendertype=abstract

679 10. Agoti CN, Munywoki PK, Phan MVT, Otieno JR, Kamau E, Bett A, et al.
680 Transmission patterns and evolution of respiratory syncytial virus in a
681 community outbreak identified by genomic analysis. Virus Evol [Internet].
682 2017;3(1). Available from:
683 https://academic.oup.com/ve/article/3066353/Transmission

684 11. Tan L, Lemey P, Houspie L, Viveen MC, Jansen NJG, van Loon AM, et al.
685 Genetic Variability among Complete Human Respiratory Syncytial Virus
686 Subgroup A Genomes: Bridging Molecular Evolutionary Dynamics and
687 Epidemiology. PLoS One [Internet]. 2012 Jan [cited 2014 Feb
688 13];7(12):e51439. Available from:
689 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3517519&tool=pm
690 centrez&rendertype=abstract

691 12. Tan L, Coenjaerts FEJ, Houspie L, Viveen MC, van Bleek GM, Wiertz EJHJ,
692 et al. The comparative genomics of human respiratory syncytial virus
693 subgroups A and B: genetic variability and molecular evolutionary dynamics. J
694 Virol [Internet]. 2013 Jul [cited 2013 Aug 7];87(14):8213–26. Available from:

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3700225&tool=pmcentrez&rendertype=abstract

13. Mufson MA, Orvell C, Rafnar B, Norrby E. Two Distinct Subtypes of Human Respiratory Syncytial Virus. J Veneral Virol [Internet]. 1985;66 (Pt 10(10):2111–24. Available from: http://www.ncbi.nlm.nih.gov/pubmed/2413163

14. Sande CJ, Mutunga MN, Medley GF, Cane PA, Nokes DJ. Group- and genotype-specific neutralizing antibody responses against respiratory syncytial virus in infants and young children with severe pneumonia. J Infect Dis [Internet]. 2013 Feb 1 [cited 2015 Feb 26];207(3):489–92. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3541697&tool=pmcentrez&rendertype=abstract

15. Trento A, Ábrego L, Rodriguez-Fernandez R, González-Sánchez MI, González-Martínez F, Delfraro A, et al. Conservation of G Protein Epitopes in Respiratory Syncytial Virus (Group A) Despite Broad Genetic Diversity: Is Antibody Selection Involved in Virus Evolution? J Virol [Internet]. 2015 May 20 [cited 2015 Jul 1];89(May):JVI.00467-15. Available from: http://jvi.asm.org/lookup/doi/10.1128/JVI.00467-15

16. Peret TC, Hall CB, Schnabel KC, Golub JA, Anderson LJ. Circulation patterns of genetically distinct group A and B strains of human respiratory syncytial virus in a community. J Gen Virol [Internet]. 1998 Sep;79 (Pt 9):2221–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9747732

17. Agoti CN, Otieno JR, Ngama M, Mwihuri AG, Medley GF, Cane P a., et al. Successive Respiratory Syncytial Virus Epidemics in Local Populations Arise

719 from Multiple Variant Introductions Providing Insights into Virus Persistence.
720 J Virol [Internet]. 2015;89(September):JVI.01972-15. Available from:
721 http://jvi.asm.org/lookup/doi/10.1128/JVI.01972-15

722 18. Otieno JR, Agoti CN, Gitahi CW, Bett A, Ngama M, Medley GF, et al.
723 Molecular evolutionary dynamics of respiratory syncytial virus group A in
724 recurrent epidemics in coastal Kenya. J Virol [Internet].
725 2016;90(10):JVI.03105-15. Available from:
726 http://jvi.asm.org/lookup/doi/10.1128/JVI.03105-15

727 19. Trento a. Major changes in the G protein of human respiratory syncytial virus
728 isolates introduced by a duplication of 60 nucleotides. J Gen Virol [Internet].
729 2003 Nov 1 [cited 2012 Jun 18];84(11):3115–20. Available from:
730 http://vir.sgmjournals.org/cgi/doi/10.1099/vir.0.19357-0

731 20. Trento A, Casas I, Calderón A, Garcia-Garcia ML, Calvo C, Perez-Breña P, et
732 al. Ten years of global evolution of the human respiratory syncytial virus BA
733 genotype with a 60-nucleotide duplication in the G protein gene. J Virol.
734 2010;84(15):7500–12.

735 21. Eshaghi A, Duvvuri VR, Lai R, Nadarajah JT, Li A, Patel SN, et al. Genetic
736 variability of human respiratory syncytial virus a strains circulating in Ontario:
737 A novel genotype with a 72 nucleotide G gene duplication. PLoS One
738 [Internet]. 2012 Jan [cited 2012 Apr 10];7(3):e32807. Available from:
739 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3314658&tool=pm
740 centrez&rendertype=abstract

741 22. Agoti CN, Otieno JR, Gitahi CW, Cane PA, James Nokes D. Rapid spread and
742 diversification of respiratory syncytial virus genotype ON1, Kenya. Emerg

743 Infect Dis. 2014;20(6):950–9.

744 23. Pierangeli A, Trotta D, Scagnolari C, Ferreri ML, Nicolai A, Midulla F, et al.
745 Rapid spread of the novel respiratory syncytial virus a on1 genotype, central
746 Italy, 2011 to 2013. Eurosurveillance [Internet]. European Centre for Disease
747 Prevention and Control (ECDC) - Health Comunication Unit; 2014 Jul 3 [cited
748 2016 Apr 29];19(26):20843. Available from:
749 http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=20843

750 24. Prifert C, Streng A, Krempl CD, Liese J, Weissbrich B. Novel respiratory
751 syncytial virus a genotype, Germany, 2011-2012. Emerg Infect Dis [Internet].
752 2013 Jun [cited 2016 Apr 29];19(6):1029–30. Available from:
753 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3713827&tool=pm
754 centrez&rendertype=abstract

755 25. Auksornkitti V, Kamprasert N, Thongkomplew S, Suwannakarn K,
756 Theamboonlers A, Samransamruajkij R, et al. Molecular characterization of
757 human respiratory syncytial virus, 2010-2011: Identification of genotype ON1
758 and a new subgroup B genotype in Thailand. Arch Virol [Internet]. 2014 Mar
759 [cited 2016 Apr 29];159(3):499–507. Available from:
760 http://www.ncbi.nlm.nih.gov/pubmed/24068580

761 26. Avadhanula V, Chemaly RF, Shah DP, Ghantoji SS, Azzi JM, Aideyan LO, et
762 al. Infection with novel respiratory syncytial virus genotype Ontario (ON1) in
763 adult hematopoietic cell transplant recipients, Texas, 2011-2013. J Infect Dis
764 [Internet]. 2015 Feb 15 [cited 2016 Apr 29];211(4):582–9. Available from:
765 http://www.ncbi.nlm.nih.gov/pubmed/25156562

766 27. Valley-Omar Z, Muloiwa R, Hu N-C, Eley B, Hsiao N-Y. Novel respiratory

767 syncytial virus subtype ON1 among children, Cape Town, South Africa, 2012.
768 Emerg Infect Dis [Internet]. 2013 Apr [cited 2014 Oct 22];19(4):668–70.
769 Available from:
770 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3647422&tool=pm
771 centrez&rendertype=abstract

772 28. Tsukagoshi H, Yokoi H, Kobayashi M, Kushibuchi I, Okamoto-Nakagawa R,
773 Yoshida A, et al. Genetic analysis of attachment glycoprotein (G) gene in new
774 genotype ON1 of human respiratory syncytial virus detected in Japan.
775 Microbiol Immunol [Internet]. 2013 Sep [cited 2014 Oct 22];57(9):655–9.
776 Available from: http://www.ncbi.nlm.nih.gov/pubmed/23750702

777 29. Duvvuri VR, Granados A, Rosenfeld P, Bahl J, Eshaghi A, Gubbay JB.
778 Genetic diversity and evolutionary insights of respiratory syncytial virus A
779 ON1 genotype: global and local transmission dynamics. Sci Rep [Internet].
780 Nature Publishing Group; 2015;5(April):14268. Available from:
781 http://www.nature.com/doifinder/10.1038/srep14268

782 30. Hotard AL, Laikhter E, Brooks K, Hartert T V., Moore ML. Functional
783 Analysis of the 60 Nucleotide Duplication in the Respiratory Syncytial Virus
784 Buenos Aires Strain Attachment Glycoprotein. J Virol [Internet].
785 2015;89(16):JVI.01045-15. Available from:
786 http://jvi.asm.org/lookup/doi/10.1128/JVI.01045-15

787 31. Agoti C, Otieno J, Gitahi C, Cane P, Nokes D. Rapid Spread and
788 Diversification of Respiratory Syncytial Virus Genotype ON1, Kenya. Emerg
789 Infect Dis. 2014;20(6).

790 32. Otieno JR, Kamau EM, Agoti CN, Lewa C, Otieno G, Bett A, et al. Spread and

Evolution of Respiratory Syncytial Virus A Genotype ON1, Coastal Kenya, 2010–2015. Emerg Infect Dis [Internet]. 2017 Feb;23(2). Available from: http://wwwnc.cdc.gov/eid/article/23/2/16-1149_article.htm

33. Nokes DJ, Ngama M, Bett A, Abwao J, Munywoki P, English M, et al. Incidence and severity of respiratory syncytial virus pneumonia in rural Kenyan children identified through hospital surveillance. Clin Infect Dis [Internet]. 2009 Nov 1 [cited 2012 Mar 22];49(9):1341–9. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2762474&tool=pmcentrez&rendertype=abstract

34. Scott JAG, Bauni E, Moisi JC, Ojal J, Gatakaa H, Nyundo C, et al. Profile: The Kilifi health and demographic surveillance system (KHDSS). Int J Epidemiol [Internet]. 2012 Jun [cited 2015 Jan 27];41(3):650–7. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3396317&tool=pmcentrez&rendertype=abstract

35. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol [Internet]. 2014;15(3):R46. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2014-15-3-r46

36. Park DJ, Dudas G, Wohl S, Goba A, Whitmer SLM, Andersen KG, et al. Ebola Virus Epidemiology, Transmission, and Evolution during Seven Months in Sierra Leone. Cell. 2015;161(7):1516–26.

37. Park DJ, Tomkins-Tinch C, Ye S, Jungreis I, Metsky H, Shlyakhter I, et al. Broad Institute viral-ngs [Internet]. 2016. Available from: https://zenodo.org/record/1040266#.WgL5nLaQ1Ps

38. Andrews S. FastQC: A quality control tool for high throughput sequence data [Internet]. 2010. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

39. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32(19):3047–8.

40. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol [Internet]. 2009 Jan [cited 2013 May 21];10(3):R25. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2690996&tool=pmcentrez&rendertype=abstract

41. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics [Internet]. 2009 Aug 15 [cited 2013 Feb 27];25(16):2078–9. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2723002&tool=pmcentrez&rendertype=abstract

42. Bankevich A, Nurk S, Antipov D, Gurevich A a., Dvorkin M, Kulikov AS, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol [Internet]. 2012 May [cited 2013 Nov 6];19(5):455–77. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3342519&tool=pmcentrez&rendertype=abstract

43. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol [Internet]. 2013

839 Apr 1 [cited 2014 Jul 13];30(4):772–80. Available from:
840 http://mbe.oxfordjournals.org/content/30/4/772

841 44. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler
842 transform. Bioinformatics. 2009;25(14):1754–60.

843 45. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing
844 genomic features. Bioinformatics. 2010;26(6):841–2.

845 46. R Core Team. R: A Language and Environment for Statistical Computing
846 [Internet]. R Foundation for Statistical Computing Vienna Austria. 2015. p.
847 {ISBN} 3-900051-07-0. Available from: http://www.r-project.org/

848 47. RStudio Team -. RStudio: Integrated Development for R. [Online] RStudio,
849 Inc, Boston, MA URL http//www rstudio com. 2016;RStudio, Inc., Boston,
850 MA.

851 48. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al.
852 Geneious Basic: An integrated and extendable desktop software platform for
853 the organization and analysis of sequence data. Bioinformatics [Internet]. 2012
854 Jun 15;28(12):1647–9. Available from:
855 https://academic.oup.com/bioinformatics/article-
856 lookup/doi/10.1093/bioinformatics/bts199

857 49. Larsson A. AliView: A fast and lightweight alignment viewer and editor for
858 large datasets. Bioinformatics. 2014;30(22):3276–8.

859 50. Chernomor O, von Haeseler A, Minh BQ. Terrace Aware Data Structure for
860 Phylogenomic Inference from Supermatrices. Syst Biol [Internet]. Oxford
861 University Press; 2016 Apr 26 [cited 2016 Sep 6];syw037. Available from:
862 http://sysbio.oxfordjournals.org/lookup/doi/10.1093/sysbio/syw037

863 51. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. Exploring the temporal
864 structure of heterochronous sequences using TempEst (formerly Path-O-Gen).
865 Virus Evol [Internet]. 2016;2(1):vew007. Available from:
866 https://academic.oup.com/ve/article-lookup/doi/10.1093/ve/vew007

867 52. Hall T. BioEdit: a user-friendly biological sequence alignment editor and
868 analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser [Internet].
869 1999;41:95–8. Available from:
870 http://jwbrown.mbio.ncsu.edu/JWB/papers/1999Hall1.pdf

871 53. Pond SLK, Frost SDW, Muse S V. HyPhy: hypothesis testing using
872 phylogenies. Bioinformatics [Internet]. 2005 Mar 1;21(5):676–9. Available
873 from: http://www.ncbi.nlm.nih.gov/pubmed/15509596

874 54. J. Spielman S. phyphy: Python package for facilitating the execution and
875 parsing of HyPhy standard analyses. J Open Source Softw [Internet]. 2018 Jan
876 17;3(21):514. Available from: http://joss.theoj.org/papers/10.21105/joss.00514

877 55. Murrell B, Weaver S, Smith MD, Wertheim JO, Murrell S, Aylward A, et al.
878 Gene-wide identification of episodic selection. Mol Biol Evol.
879 2015;32(5):1365–71.

880 56. Kosakovsky Pond SL, Frost SDW, Pond SLK, Frost SDW. Not So Different
881 After All: A Comparison of Methods for Detecting Amino Acid Sites Under
882 Selection. Mol Biol Evol [Internet]. 2005 May 1 [cited 2014 Jan
883 27];22(5):1208–22. Available from:
884 http://mbe.oxfordjournals.org/content/22/5/1208%5Cnhttp://mbe.oxfordjournal
885 s.org/content/22/5/1208.full.pdf%5Cnhttp://mbe.oxfordjournals.org/content/22/
886 5/1208.short%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/15703242

887 57. Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Kosakovsky Pond SL,
888 et al. FUBAR: A fast, unconstrained bayesian AppRoximation for inferring
889 selection. Mol Biol Evol. 2013;30(5):1196–205.

890 58. Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond
891 SL. Detecting individual sites subject to episodic diversifying selection. Malik
892 HS, editor. PLoS Genet [Internet]. Public Library of Science; 2012 Jan [cited
893 2014 Jan 20];8(7):e1002764. Available from:
894 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3395634&tool=pm
895 centrez&rendertype=abstract

896 59. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with
897 BEAUti and the BEAST 1.7. Mol Biol Evol [Internet]. 2012 Aug [cited 2014
898 Jan 21];29(8):1969–73. Available from:
899 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3408070&tool=pm
900 centrez&rendertype=abstract

901 60. Shapiro B, Rambaut A, Drummond AJ. Choosing appropriate substitution
902 models for the phylogenetic analysis of protein-coding sequences. Mol Biol
903 Evol [Internet]. 2006 Jan [cited 2012 Oct 27];23(1):7–9. Available from:
904 http://www.ncbi.nlm.nih.gov/pubmed/16177232

905 61. Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian coalescent
906 inference of past population dynamics from molecular sequences. Mol Biol
907 Evol [Internet]. 2005;22(5):1185–92. Available from:
908 http://mbe.oupjournals.org/cgi/doi/10.1093/molbev/msi103

909 62. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. Relaxed phylogenetics and
910 dating with confidence. PLoS Biol. 2006;4(5):699–710.

911 63. Ferreira MAR, Suchard MA. Bayesian analysis of elapsed times in continuous-
912 time Markov chains. Can J Stat. 2008;36(3):355–68.

913 64. Baele G, Lemey P, Bedford T, Rambaut A, Suchard MA, Alekseyenko A V.
914 Improving the accuracy of demographic and molecular clock model
915 comparison while accommodating phylogenetic uncertainty. Mol Biol Evol.
916 2012;29(9):2157–67.

917 65. Baele G, Li WLS, Drummond AJ, Suchard MA, Lemey P. Accurate model
918 selection of relaxed molecular clocks in Bayesian phylogenetics. Mol Biol
919 Evol. 2013;30(2):239–43.

920 66. Lê S, Josse J, Husson F. FactoMineR : An R Package for Multivariate
921 Analysis. J Stat Softw [Internet]. 2008;25(1):253–8. Available from:
922 http://linkinghub.elsevier.com/retrieve/pii/S016041200800113X%5Cnhttp://w
923 ww.jstatsoft.org/v25/i01/

924 67. Comas-García A, Noyola DE, Cadena-Mota S, Rico-Hernández M, Bernal-
925 Silva S. Respiratory Syncytial Virus-A ON1 Genotype Emergence in Central
926 Mexico in 2009 and Evidence of Multiple Duplication Events. J Infect Dis
927 [Internet]. 2018 Jan 24; Available from: https://academic.oup.com/jid/advance-
928 article/doi/10.1093/infdis/jiy025/4823205

929 68. Martínez I, Dopazo J, Melero JA. Antigenic structure of the human respiratory
930 syncytial virus G glycoprotein and relevance of hypermutation events for the
931 generation of antigenic variants. J Gen Virol [Internet]. 1997 Oct 1 [cited 2014
932 Apr 9];78(10):2419–29. Available from:
933 http://www.ncbi.nlm.nih.gov/pubmed/9349460

934 69. Hause AM, Henke DM, Avadhanula V, Shaw CA, Tapia LI, Piedra PA.

935 Sequence variability of the respiratory syncytial virus (RSV) fusion gene
936 among contemporary and historical genotypes of RSV/A and RSV/B. PLoS
937 One. 2017;12(4).

938 70. García O, Martín M, Dopazo J, Arbiza J, Frabasile S, Russi J, et al.
939 Evolutionary pattern of human respiratory syncytial virus (subgroup A):
940 cocirculating lineages and correlation of genetic and antigenic changes in the G
941 glycoprotein. J Virol [Internet]. 1994 Sep [cited 2015 Feb 3];68(9):5448–59.
942 Available from:
943 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=236945&tool=pmc
944 entrez&rendertype=abstract

945 71. Cane P a. Analysis of linear epitopes recognised by the primary human
946 antibody response to a variable region of the attachment (G) protein of
947 respiratory syncytial virus. J Med Virol [Internet]. 1997 Apr;51(4):297–304.
948 Available from: http://www.ncbi.nlm.nih.gov/pubmed/9093944

949 72. Fuentes S, Coyle EM, Beeler J, Golding H, Khurana S, Wilson PC. Antigenic
950 Fingerprinting following Primary RSV Infection in Young Children Identifies
951 Novel Antigenic Sites and Reveals Unlinked Evolution of Human Antibody
952 Repertoires to Fusion and Attachment Glycoproteins. PLOS Pathog [Internet].
953 1005;1–24. Available from: http://dx.doi.org/10.1371/journal.ppat.1005554

954 73. Komissarova N, Kashlev M. Transcriptional arrest: Escherichia coli RNA
955 polymerase translocates backward, leaving the 3’ end of the RNA intact and
956 extruded. Proc Natl Acad Sci U S A. 1997;94(March):1755–60.

957 74. Canchaya C, Proux C, Fournous G, Bruttin A, Brüssow H. Prophage genomics.
958 Microbiol Mol Biol Rev [Internet]. Am Soc Microbiol; 2003;67(2):238–276,

table of contents. Available from: http://mmbr.asm.org/cgi/content/abstract/67/2/238

75. Raghwani J, Thompson RN, Koelle K. Selection on non-antigenic gene segments of seasonal influenza A virus and its impact on adaptive evolution. Virus Evol [Internet]. 2017;3(2). Available from: http://academic.oup.com/ve/article/doi/10.1093/vex034/4614565

76. Stern A, Yeh M Te, Zinger T, Smith M, Wright C, Ling G, et al. The Evolutionary Pathway to Virulence of an RNA Virus. Cell. 2017;169(1):35–46.e19.

77. Grubaugh ND, Andersen KG. Experimental Evolution to Study Virus Emergence. Cell. 2017. p. 1–3.

78. Yoshihara K, Le MN, Okamoto M, Wadagni ACA, Nguyen HA, Toizumi M, et al. Association of RSV-A ON1 genotype with Increased Pediatric Acute Lower Respiratory Tract Infection in Vietnam. Sci Rep [Internet]. 2016 Jun 16;6:27856. Available from: http://www.nature.com/articles/srep27856

79. Panayiotou C, Richter J, Koliou M, Kalogirou N, Georgiou E, Christodoulou C. Epidemiology of respiratory syncytial virus in children in Cyprus during three consecutive winter seasons (2010-2013): age distribution, seasonality and association between prevalent genotypes and disease severity. Epidemiol Infect [Internet]. 2014 Nov [cited 2016 Apr 29];142(11):2406–11. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24476750

80. The Report: Kenya 2017 [Internet]. Oxford Business Group; 2017 [cited 2017 Dec 5]. Available from: https://www.oxfordbusinessgroup.com/kenya-2017/tourism

81. More than minerals [Internet]. The Economist; [cited 2017 Dec 5]. Available from: https://www.economist.com/news/middle-east-and-africa/21574012-chinese-trade-africa-keeps-growing-fears-neocolonialism-are-overdone-more

82. Sunday F. US not Kenya's largest tourism market. Standard Group Limited [Internet]. Nairobi; 2018 Feb 20; Available from: https://www.standardmedia.co.ke/business/article/2001270358/us-not-kenya-s-largest-tourism-market

1004 **Figure Legends**

1005

1006 **Figure 1: Sample sequencing and genome details**

1007 The two RSV-A whole genome amplification strategies used in this study are shown
1008 in (A), i.e. six and fourteen amplicons. For each panel the positions of primer targets
1009 for each amplicon are indicated. The locations of the 11 RSV ORFs are indicated on
1010 top of panel 1. (B) The proportion of RSV genome length sequence recovered (using
1011 KC731482 as the reference) was plotted as a function of sample's diagnostic real-time
1012 PCR Ct value. (C) The distribution of the diagnostic real-time PCR Ct values for the
1013 samples reported here. (D) Shows the log values of the sequencing depth (see
1014 Methods) at each position of the genome assemblies along the concatenated RSV
1015 ORFs (i.e. excluding the intergenic regions).

1016

1017 **Figure 2: Global and local ON1 MCC trees and PCA**

1018 (A) Maximum clade credibility tree inferred from 344 global full genome sequences
1019 (see Methods), annotated with the identified Kilifi lineages (I-III) and the tips colour
1020 coded with the continent of sample collection. Node labels are posterior probabilities
1021 indicating support for the selected nodes. (B) shows the evolutionary rate estimates
1022 for the different genotype ON1 ORFs. (C) is an MCC tree inferred from 154 ON1
1023 genomes from Kilifi annotated with identified clades A and B, and the tips colour
1024 coded with the epidemic season. (D) is a PCA analysis (see Methods) of the same
1025 dataset as (C) and similarly annotated. Percentage of variance explained by each
1026 component is indicated on the axis.

1027

1028

Figure 3: Global ON1 G-gene MCC phylogenetic tree

A maximum clade credibility tree inferred from 1,167 partial ON1 G gene global sequences with the tips colour coded with the source continent.

Figure 4: Pairwise genomic distances and genome-wide amino acid variation

The distribution of pairwise genetic distances between genotype GA2 and ON1 genome sequences are shown in (A) and (B), respectively. (C) is an entropy plot showing amino acid variation along the concatenated Kilifi RSV-A genomes.

Figure 5: Estimated TMRCA for Kilifi RSV-A viruses and ORFs

(A) Maximum clade credibility tree inferred from 184 RSV-A complete genome sequences (coding regions only) from Kilifi with the tips colour coded by genotype, i.e. ON1 (cyan) and GA2 (red). The two node bars indicate the 95% HPD interval for the TMRCA for the Kilifi GA2 and ON1 viruses (grey), and Kilifi ON1 strains (blue). Node labels are posterior probabilities indicating support for the selected nodes. (B) shows the TMRCA (with 95% HPD interval) of the different ORFs of the RSV-A genotype GA2 and ON1 viruses. The (*) indicates node posterior support of <0.9 for the split between GA2 and ON1 in the N and P ORFs.

Supporting Information Legends

S1 Fig: Sampling locations of the global ON1 G-gene dataset

A map showing source country locations of the global ON1 G-gene sequences dataset analyzed here with circles representative of relative proportion of contributing sequences by country

S2 Fig: Root-to-tip regression analysis of Kilifi RSV-A ORFs

A root-to-tip regression analysis of ML trees from whole genomes and 11 separate coding regions, with the points colour coded by genotype; GA2 (red) and ON1 (cyan).

S3 Fig: BEAST MCC ON1-GA2 divergence trees for different ORFs

MCC trees inferred from different ORFs of 184 RSV-A complete genome sequences from Kilifi with the tips colour coded by genotype, i.e. ON1 (cyan) and GA2 (red).

S1 Table: Model selection to infer time-structured phylogenies

S2 Table: Study samples and genomes details

S3 Table: SNPs identified from dataset of all Kilifi genomes

S4 Table: Signature SNPs between ON1 and GA2 viruses

S5 Table: Signature SNPs between successful and limited transmission ON1 lineages

Figure 1: Sample sequencing and genome details

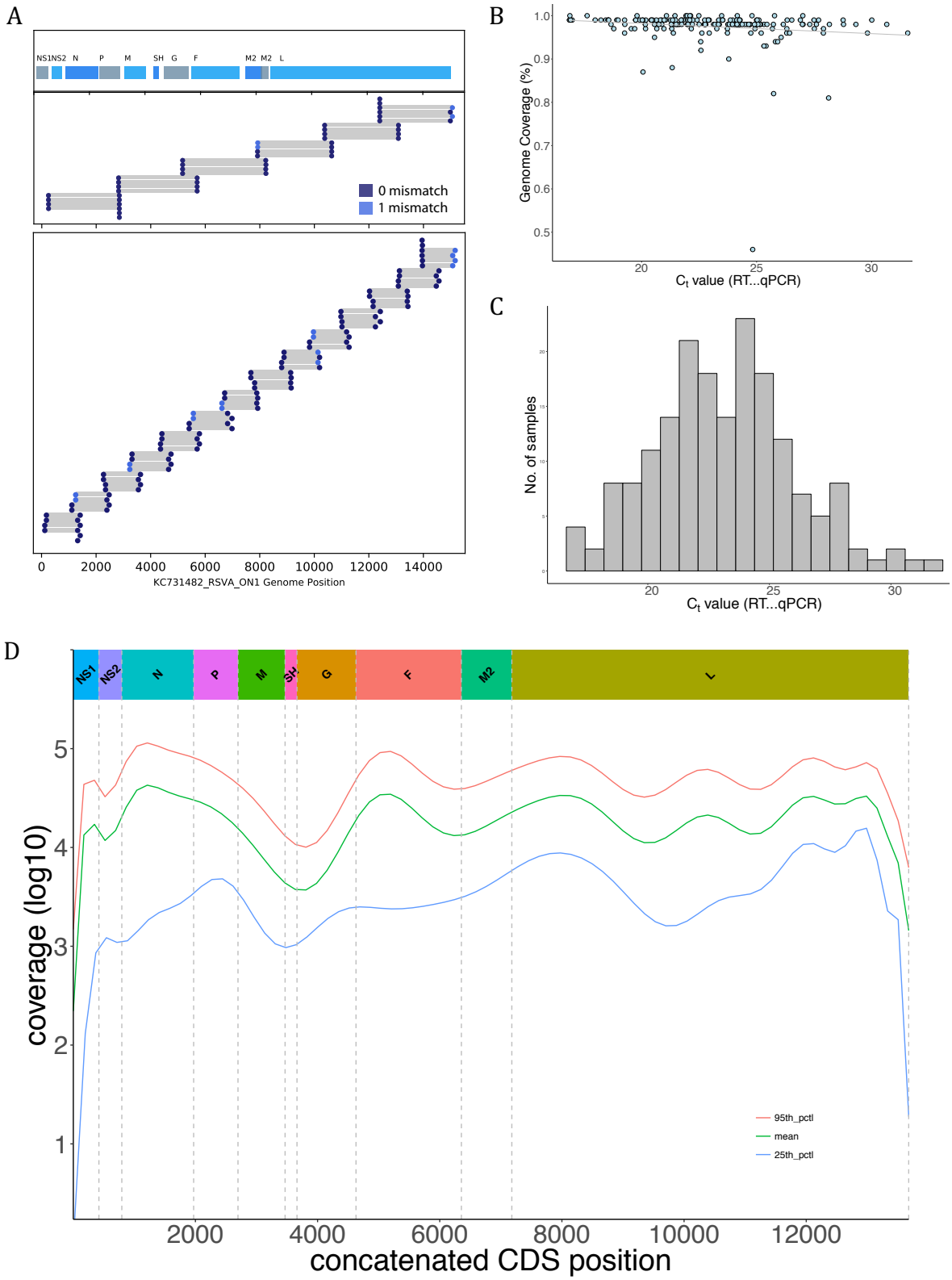


Figure 2: Global and local (Kilifi) MCC time-resolved trees and evolutionary rate estimates

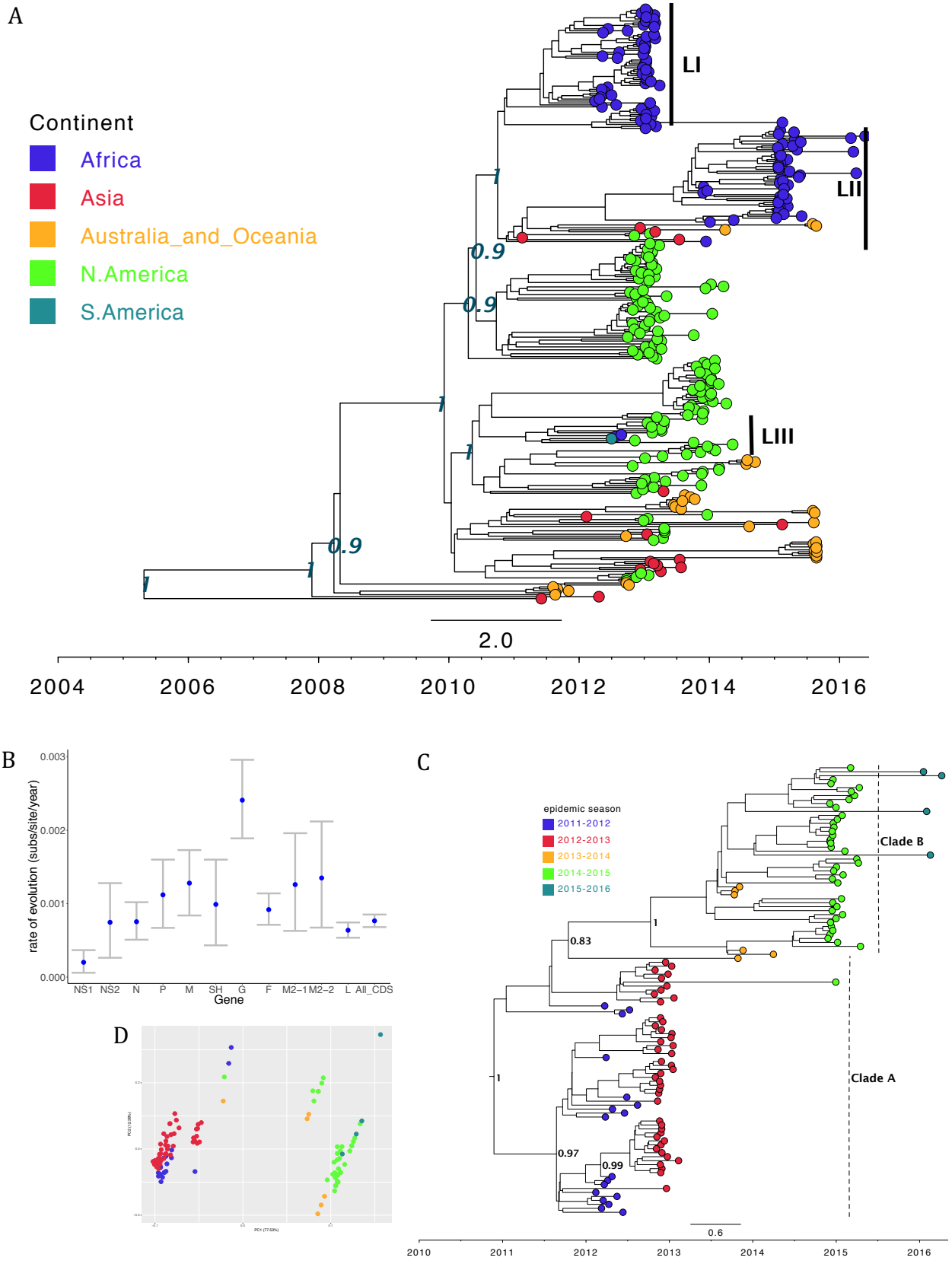


Figure 3: Global ON1 Bayesian G-gene phylogenetic tree

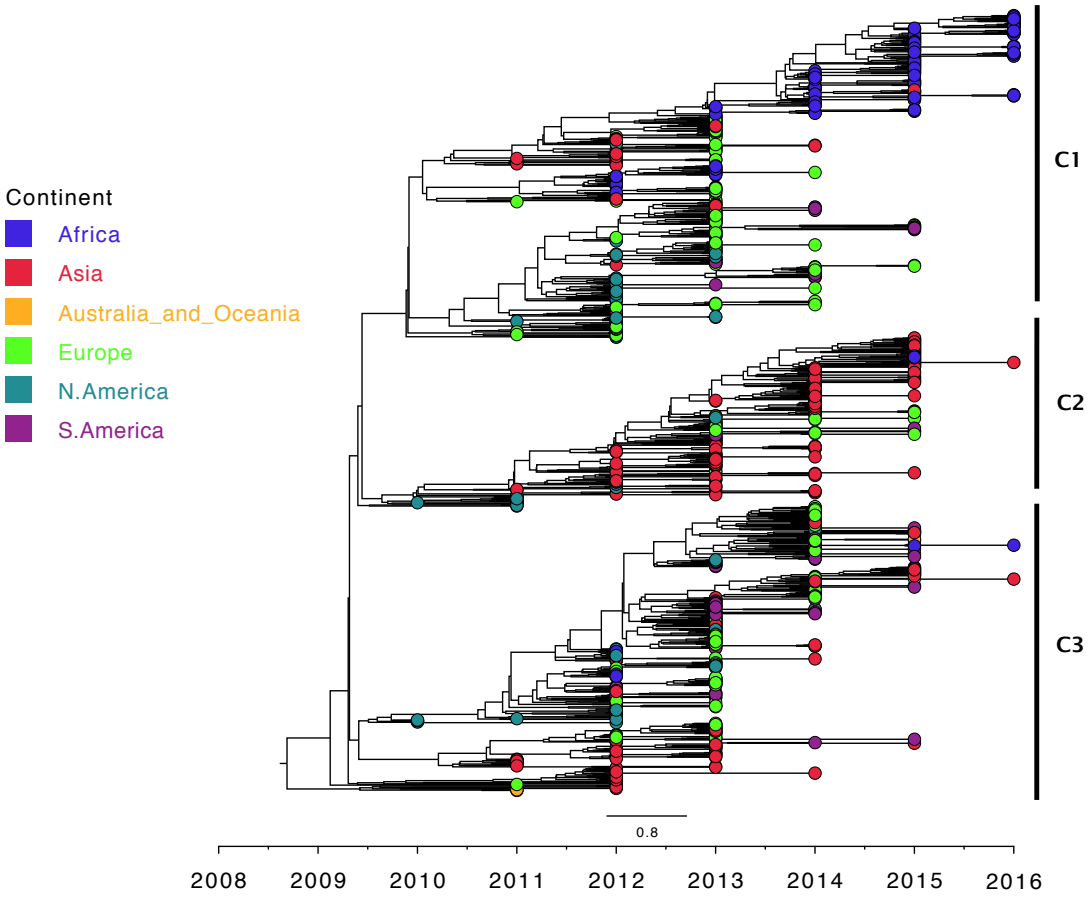


Figure 4: Pairwise genomic distances and genome-wide amino acid variation

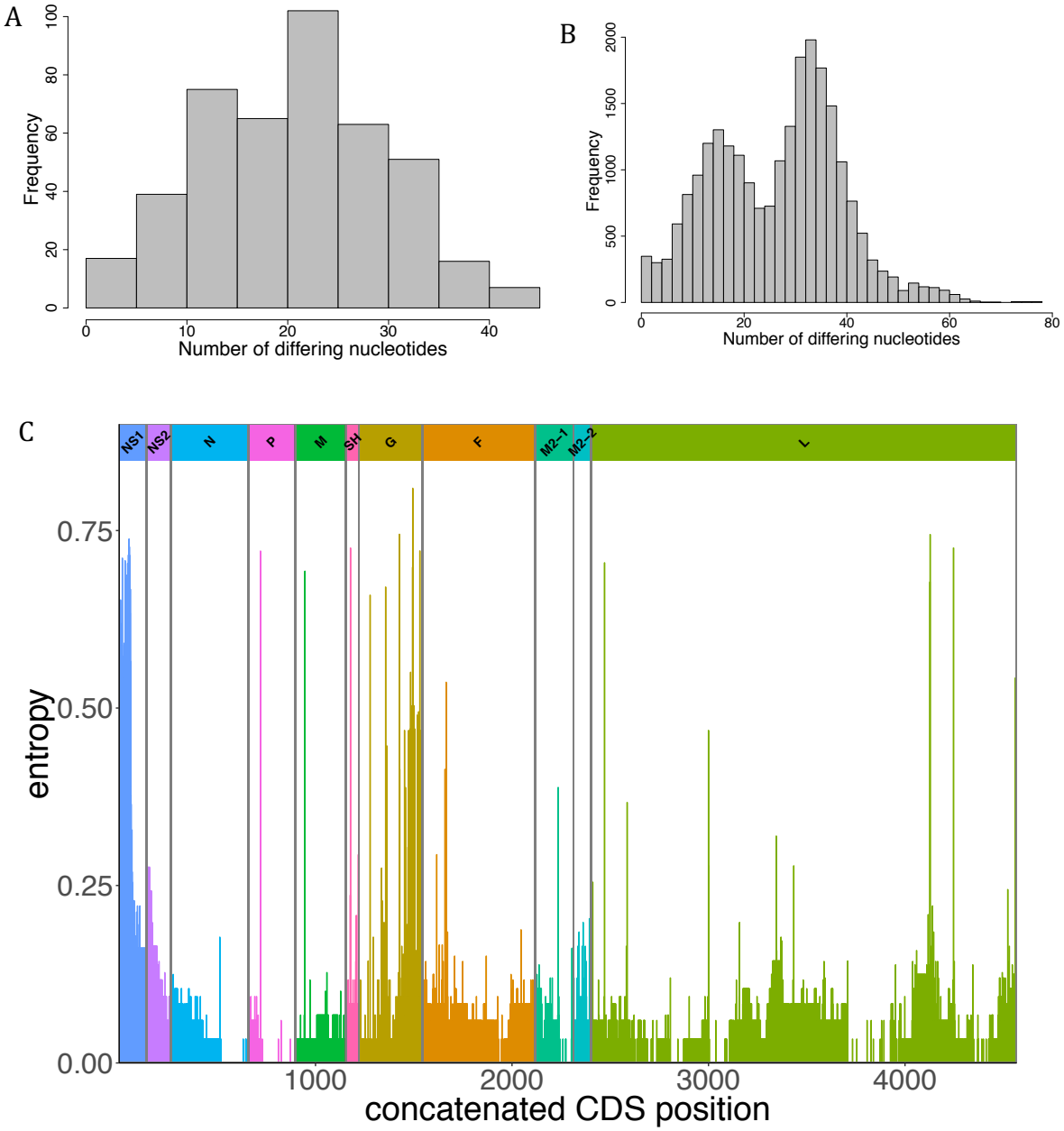


Figure 5: Estimated TMRCA for Kilifi RSV-A viruses and ORFs

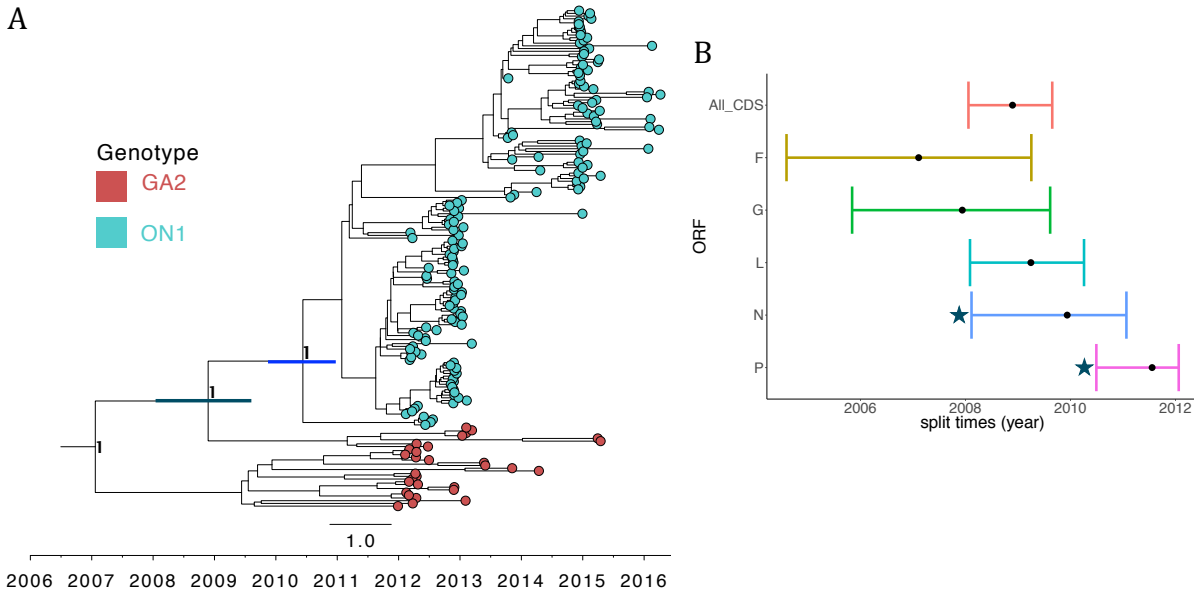
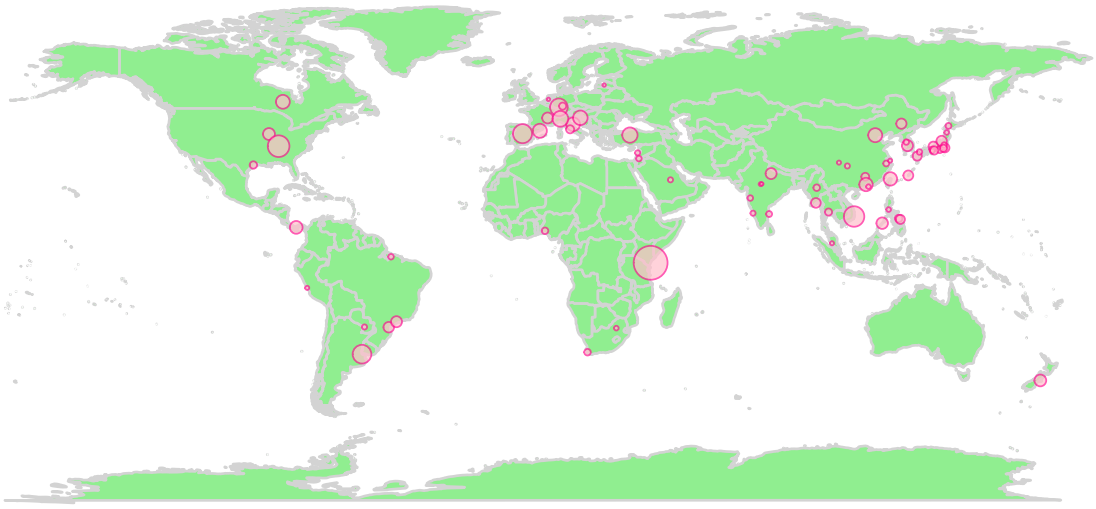


Table 1: Signature nucleotide and amino acid polymorphisms between genotype ON1 and GA2 viruses

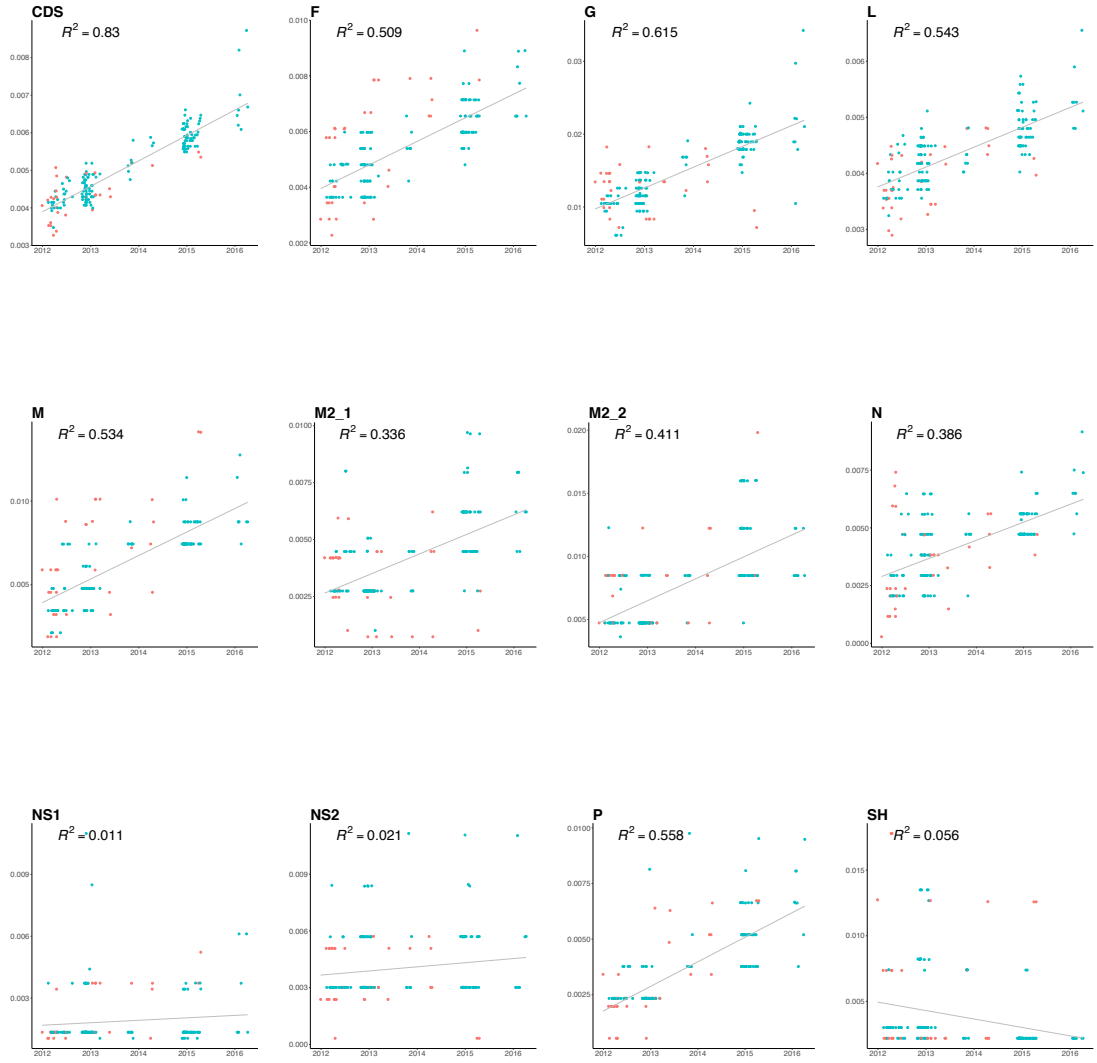
ORF	[α]ORF Nt Pos.	ORF AA Pos.	Change	AA Change	SNP Type
G	424	142	TT -> CA	L -> Q	Substitution
G	622	208	C -> A	L -> I	Transversion
G	695	232	G -> A	G -> E	Transition
G	709	237	A -> G	N -> D	Transition
G	758	253	A -> C	K -> T	Transversion
G	817	273	T -> A	Y -> N	Transversion
G	821	274	C -> T	P -> L	Transition
G	851	284	72 nt duplication	24 AA insertion	Deletion
G	929 (GA2: 857)	310	C -> T	P -> L	Transition
G	941 (GA2: 869)	314	T -> C	L -> P	Transition
F	346	116	A -> G	N -> D	Transition
F	364	122	G -> A	A -> T	Transition
M2-1	349	117	A -> C	N -> H	Transversion
L	1792	598	C -> T	H -> Y	Transition
L	5175	1725	A -> T	E -> D	Transversion

ORF=Open Reading Frame, Nt=Nucleotide, AA=Amino Acid, Pos.=Position
[α]Positions are relative to ON1 strains, in which complementary positions in GA2 (without the duplication) within the G protein are shown in brackets.

Supplementary figure 1: *Sampling locations of the global ON1 G-gene dataset with circles representative of relative proportion of contributing sequences by country*



Supplementary figure 2: *Root-to-tip regression analysis of Kilifi RSV-A ORFs*



Supplementary figure 3: *BEAST MCC trees showing divergence between ON1(cyan) and GA2 (red) ORFs*