Genomic epidemiology and evolutionary dynamics of respiratory syncytial virus group B in Kilifi, Kenya, 2015-17

Everlyn Kamau [1,*,#], James R. Otieno [1], Nickson Murunga [1], John W. Oketch [1], Joyce M. Ngoi [1], Zaydah R. de Laurent [1], Anthony Mwema [1], Joyce U. Nyiro [1], Charles N. Agoti [1,2], D. James Nokes [1,3]

[1] Epidemiology and Demography Department, KEMRI–Wellcome Trust Research Programme, Kilifi, Kenya

[2] Pwani University, School of Health and Human Sciences, Kilifi, Kenya.

[3] School of Life Sciences and Zeeman Institute (SBIDER), University of Warwick, Coventry, UK

[*] Correspondence: everlyn.kamau@ndm.ox.ac.uk

[#] Current address: Nuffield Department of Medicine, University of Oxford, Oxford, UK

Abstract

Respiratory syncytial virus (RSV) circulates worldwide and is a leading cause of acute respiratory illness in young children. There is paucity of genomic data from purposively sampled populations by which to investigate evolutionary dynamics and transmission patterns of RSV. Here we present an analysis of 295 RSV group B genomes from Kilifi, coastal Kenya, sampled from individuals seeking outpatient care in 9 health facilities across a defined geographical area (890 km^2), over 2 RSV epidemics between 2015 and 2017. RSVB diversity was characterized by multiple viral introductions into the area and co-circulation of distinct genetic groups or clusters, which transmitted and diversified locally but with varying frequency. Bayesian analyses indicated a strong spatially and temporally structured viral population suggesting extensive within-epidemic virus transmission. Phylogeographic analysis provided a strong support for epidemiological linkage from one central health facility to other facilities. Increase in relative diversity paralleled increase in seasonal viral incidence. Importantly, we identified a cluster of viruses (n=91) that emerged in the 2016/17 epidemic, carrying distinct amino-acid signatures including a novel non-synonymous change (K68Q) in antigenic site ∅ in the Fusion gene. A different non-synonymous change K68N was recently associated with escape from a potent neutralizing monoclonal antibody (MEDI8897). RSVB diversity was additionally marked by signature non-synonymous substitutions that were unique to particular genomic clusters, some of which were under diversifying selection. Our findings provide insights into recent evolutionary and epidemiologic behaviors of RSV group B, and highlight possible emergence of a novel antigenic variant, which has implications on current prophylactic development strategies.

Introduction

Respiratory syncytial virus (RSV) is the most common cause of acute lower respiratory tract infection in children <5 years of age worldwide, with an estimated associated mortality of up to 199,000 deaths per year mostly in developing countries (1,2). RSV is also an important cause of community-acquired pneumonia among hospitalized adults of all ages (3). RSV is a member of the family *Pneumoviridae*, subfamily *Orthopneumovirus,* has an enveloped, non-segmented, single-stranded, negative sense RNA genome of approximately 15,000 nucleotides encoding 11 proteins: NS1, NS2, N, P, M, SH, G, F, M2-1, M2-2, and L (4). RSV clinical isolates are classified into two groups (RSVA and RSVB) based on antigenic and genetic variability (5). Distinct genotypes of RSV have been shown to circulate locally and globally (6-9) and strains may vary from location to location in any given season with viruses identified in one location being similar to those from vastly different geographic locations identified in different years (10), suggestive of rapid global transmission. The available therapeutic modalities for RSV are chiefly supportive, and prophylactic treatment with RSV-specific neutralizing antibody is effective in reducing RSV morbidity in infants, but its use is currently limited to high-risk populations in high-resource settings (11). There is no licensed RSV vaccine for routine use in immunization, however, prophylactic vaccine candidates and monoclonal antibodies (mAbs) are now in advanced clinical trials (12).

We have previously characterised the community dynamics of RSV in the coastal region of Kenya, using sequences of the highly variable G glycoprotein gene (encoding the attachment protein) for RSV groups A and B and using whole genomes of RSVA genotype ON1, almost exclusively from samples from pneumonia patients admitted to the Kilifi County Hospital (13-15). From these studies, RSV displays high genetic diversity of locally

61 circulating strains, within and between consecutive epidemics. Furthermore, recurrent RSV
62 epidemics in Kilifi are depicted by sequential replacement of genotypes, over the long term,
63 and high turnover of variants within genotypes in the short term (13,14). In the study
64 reported here, samples arise from a design aimed to limit temporal, age-related, illness
65 severity, geographical, and health care access bias. Recruitment was carried throughout a
66 study location, from representative health facilities, simultaneously, and of patients of any
67 age with mild acute respiratory symptoms (16).

68 Phylodynamic and phylogeographic methods have been used increasingly to study
69 molecular epidemiology and evolutionary dynamics of virus populations e.g. Ebola, Zika, and
70 Influenza (17-21). However, despite the importance of RSV to pneumonia hospitalisation
71 and mortality among children (1), equivalent genome-scale studies to examine RSV
72 transmission patterns and evolutionary dynamics within a community setting, are still few
73 and remain a major gap in our understanding of what viral factors shape epidemiologic and
74 evolutionary processes RSV at the local level. While most molecular studies on RSV have
75 focused on the G glycoprotein gene because of its high genetic diversity and utility as a
76 phylogenetic marker, genome-wide genetic signatures in RSV genome additionally inform
77 on diversity and the adaptive mechanisms of RSV viruses following introduction into
78 communities (15).

79 Here, we applied a combination of molecular clock, coalescent and discrete diffusion
80 phylogeographic models to whole genome sequences to measure genomic diversity and
81 deduce spatial and temporal circulation of RSVB in rural Kilifi, coastal Kenya. We sought to
82 characterize introductions, transmission and spread of RSVB from samples collected
83 through outpatient surveillance of respiratory viruses, analogous to studying community
84 epidemics. We present estimates of evolutionary parameters such as the genomic rate of

evolution and relative genetic diversity and describe spatial and temporal clustering patterns of viruses within each epidemic as demonstrated by phylogenetically distinct clades. In particular, we determine possible selection-driven emergence of a novel RSVB variant carrying distinct amino acid signatures. Our study significantly increases the number of publicly available complete RSVB genomes, which will enable further studies of RSV evolution.

Materials and methods

Study site and patient recruitment. The annual RSV epidemics in Kilifi, coastal Kenya, are seasonal starting from November through May, with a peak around January and an average duration of 18 weeks (22). The study from which the samples arise and used in the present report was done from December 2015 to March 2018, a period covering three RSV seasonal epidemics (2015/16, 2016/17 and 2017/18), and was carried out within the Kilifi Health and Demographic Surveillance System (KHDSS) (23). The study was conducted to document the community-wide burden of respiratory virus infections. Nine out of the 24 public outpatient health facilities in KHDSS were purposively selected (**Figure 1(A)**) – Matsangoni (MAT), Ngerenya (NGE), Mtondia (MTO), Sokoke (SOK), Mavueni (MAV), Jaribuni (JAR), Chasimba (CHA), Pingilikani (PIN) and Junju (JUN) - to provide a broad representation across the geographical region, covering major road networks into the location and variation in population density (16). Participant recruitment and specimen collection was integrated within the routine patient care and led by a resident clinician or nurse as detailed in (16). Patients of any age presenting with one or more ARI symptoms of cough, sneezing, nasal congestion, difficulty breathing, or increased respiratory rate for age

were eligible. Written individual informed consent was sought from adult patients and parents/guardians of patients below 18 years.

The study was approved by the KEMRI-Scientific and Ethical Review Unit (SERU# 3103) and the University of Warwick Biomedical and Scientific Research Ethics Committee (BSREC# REGO-2015-6102).

Sample collection and laboratory testing. A nasopharyngeal swab (NPS) was collected from each participant and stored in universal virus transport media (Copan Diagnostics, USA). RNA was extracted from samples by Qiacube HT using an RNeasy extraction kit (Qiagen, Germany) and screened for a range of respiratory viruses including RSVA and RSVB using a multiplex real-time PCR assay system (24,25). Samples with real time RT-PCR cycle threshold (Ct) of <35.0 were defined as positive for the target virus.

RSVB whole genome amplification and sequencing. Whole genome amplification and sequencing was attempted for all RSVB positive samples with PCR cycle threshold (Ct) value < 35.0 collected between mid-December 2015 to end of May 2017. No RSVB was detected from June 2017 to the end of the study in March 2018 (i.e. 2017/18 epidemic was RSVA only). Reverse transcription of RNA molecules and PCR amplification were performed with a six-amplicon, six-reaction strategy presented in detail in (26). Briefly, extracted RNA was converted to cDNA using Superscript III First Strand Synthesis kit (Invitrogen) and forward primers, followed by PCR using Phusion High-Fidelity DNA polymerase (NEB). Amplification success was confirmed by observing the expected PCR product size (2300–2500 bp) on 0.6% agarose gels. Amplicons were fragmented, tagged to adapters, and indexed using the Nextera XT (Illumina, San Diego, CA, USA) library prep kit as per

manufacturer's instructions. Size distribution of the barcoded libraries was assessed by Agilent's 2100 Bioanalyzer. Pooled and normalized libraries in batches of seventy or eighty were sequenced on Illumina MiSeq system using 2 x 250 bp or 2 x 300 bp paired-end (PE) sequencing at the KEMRI-Wellcome Trust Research Program laboratories.

Short read data assembly. Methods for quality check of the sequence reads, depletion of human reads, generation of consensus genome assemblies and calculation of coverage depth, were as described in (15). Briefly, quality check of the sequence reads was done using FastQC. Consensus genome assemblies were generated using viral-ngs pipeline v1.19.0 (Broad Institute). The raw reads were depleted of human reads by mapping onto the human reference genome hg19 using bowtie2 (27), and samtools (28) used to filter, sort, and recover the unmapped reads. The raw reads were mapped onto individual consensus assemblies with bowtie2, then samtools used to sort and index the aligned .bam files, and BEDtools (29) used to generate the coverage depth statistics.

Sequence data compilation. To contextualize the diversity of RSVB in Kilifi in 2015 to 2017, the new Kilifi genomes were appended to a global dataset of other RSVB genomes available in Genbank (>14000 nucleotides (nt)), retrieved on 29 September 2019 (**Supplementary Table 1**). Only global sequences belonging to the RSVB BA genotype were considered, since the Kilifi sequences were of the BA genotype. The global dataset comprised sequences of viruses collected between 2012 and 2016 from UK, US, Russia, Philippines, Nicaragua, Jordan and India. Sequences of viruses sampled before 2012 and published without date or location of sampling were excluded.

Phylogenetic and phylogeographic analyses. Sequence alignment was done using MAFFT v.7.221 (30) using the parameters '–localpair –maxiterate 1000'. To determine the degree of temporal signal of divergence in the Kilifi RSVB genomes, a maximum likelihood (ML) tree was estimated using RAxML (31) under general time-reversible (GTR) nucleotide substitution model with gamma-distributed among-site rate heterogeneity, and examined by exploratory linear regression analysis in Tempest v1.5.1 (32). jModelTest (33) was used to determine best-fit substitution model. Root-to-tip divergence was plotted as a function of sampling date (day-month-year). Branch support for ML trees was evaluated using 1000 bootstrap replicates.

Bayesian molecular clock phylogenies and discrete trait phylogeographic analyses were done using BEAST v1.10 (34) for 300 million MCMC steps, sampling parameters and trees every 5000 steps. HKY nucleotide substitution model with a gamma-distributed rate variation among sites was used. Path-sampling and stepping-stone marginal likelihood estimator (MLE) models (35) were used to estimate the most probable combination of uncorrelated lognormal relaxed molecular clock and coalescent models. The best fitting combination was the uncorrelated log-normal distributed relaxed molecular clock and Skygrid's Gaussian Markov random field (GMRF) tree prior. A non-informative continuous time Markov chain reference prior was used on the molecular clock. Mixing and convergence of the MCMC sampler in the posterior target distribution was evaluated using Tracer v1.6 (http://beast.bio.ed.ac.uk/Tracer), and a maximum clade credibility (MCC) trees summarized with TreeAnnotator after removal of 10% burn-in. MCC trees were visualized with FigTree 1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/).

To describe and quantify viral movement in Kilifi, we performed a discrete trait diffusion analysis, using sampling location as a discrete trait. We used a Bayesian stochastic

search variable selection (BSSVS) procedure (36) and compared reversible and non-reversible (asymmetric) discrete diffusion phylogeographic models to estimate the most relevant transition pathways within the study locations. The standard implementation of the discrete asymmetrical phylogeographic approach assigns the same prior probability distribution to all of the transition rates in the continuous time markov chain (CTMC) (37). On the other hand, the reversible CTMC model allows more balanced transitions in the phylogeny or free lineage exchange between location states without preconditioned directionality but may have a poor fit for locations that could have unidirectional links in reality e.g. spatially expanding epidemics (36). Statistical support for migratory events was measured using Bayes factors (BF) and summarized using SpreaD3 software (38) after discarding 10% burn-in. We assumed that a BF > 6 is strong evidence for a well-supported viral pathway between two locations.

Molecular evolution and adaptation. Gene-specific ratio of nonsynonymous to synonymous substitutions per site (dN/dS) were estimated for codon-based alignments using the single likelihood ancestor counting (SLAC) method available at the Datamonkey webserver (39). We also investigated episodic positive or diversifying selection using mixed-effects model of evolution (MEME) and Fast, Unconstrained Bayesian AppRoximation (FUBAR) approaches also available in Datamonkey. MEME aims to detect sites evolving under positive selection in a proportion of branches, while FUBAR uses a Bayesian approach and assumes that the selection pressure is constant along the entire phylogeny.

Phylogenomic clusters. We defined a phylogenomic cluster as genetically linked viral sequences (≥2) that were more closely related than any randomly selected sequences

within genetic distance threshold of <*d*. Maximum likelihood evolutionary distances, an estimate of the average number of changes per base pair, were calculated for each pair of consensus genomes using IQ-TREE (40) to measure genetic similarity or dissimilarity between pairs of taxa. ML distances, depicting genetic distance between sequences in a phylogenetic tree measured in nucleotide substitutions per site, were selected over the conventional nucleotide *p*-distances in order to correct for multiple nucleotide substitutions.

The distribution histogram plots of pairwise genetic distances were generated in R, and a distance cutoff value was determined as the least frequent value (*d*) between the first and second peaks in the histogram. Median values of pairwise ML evolutionary distance distributions of the identified phylogenomic cluster were checked if they were below *t*-percentile of the overall distance distribution. The *t*-percentile corresponded to the *d* distance threshold. In addition, a phylogenomic cluster was only considered eligible if it had reliability of $\geq$ 90% phylogenetic bootstrap support or Bayesian posterior probability. Each inferred phylogenomic cluster was evaluated for size, temporal and genetic characteristics.

Phylogeny-trait association analysis. We used the Bayesian Tip-association Significance (BaTS) software (41), which uses the posterior sets of trees from Bayesian MCMC analysis, to measure the degree of association between sampling location and the phylogeny and are inversely correlated with the degree of association. Each sequence was assigned a location trait reflecting its origin of sampling and the first 10% of tree states were removed as burn-in. The overall statistical significance was determined by estimating the parsimony score (PS) and association index (AI) metrics, where the null hypothesis was that clustering by location was not more than expected by chance alone. In addition, the

maximum clade size (MC) metric was used to compare the strength of clustering at each location by calculating the expected (null) and the observed mean clade size from each study location. A significance level of 0.05 was used in all cases. The PS, AI and MC statistics were computed for a null distribution with 1000 replicates.

Sequence data availability. The final sequencing reads are available in the NCBI BioProject database under the study accession PRJNA562116 and the genomes generated in this study are available in GenBank under accession numbers MN365302 to MN365600.

Results

RSV occurrence in Kilifi, 2015 to 2017. Between December 2015 and June 2017, a total of 8127 nasopharyngeal swab samples were tested for RSV. RSV was detected in 503 (6.19%) samples (Ct <35): 95 (18.9%) were RSV group A (RSVA) and 408 (81.1%) were RSV group B (RSVB). The frequency and pattern of occurrence of RSVB for each health facility are shown in **Figure 1(B)**. The mean (SD) Ct value was 26.9 (4.16) and 25.0 (4.16) for RSVA and RSVB, respectively. The proportion of RSV positive individuals differed by age (*p* value <0.001), study location (*p* value = 0.003) but not by gender (*p* value = 0.078) (**Table 1**). The distribution of RSV viral load (equated to rRT-PCR Ct value) did not differ by outpatient health facility (**Supplementary Figure 1(A)**). The median age of RSV positive individuals was 20 months (interquartile range (IQR), 8-43 months), 81.7% (411/503) were aged 5 years or younger, and 272 (54.1%) of the cases were female (**Table 1**). RSV prevalence across different age groups by gender is shown in **Supplementary Figure 2**.

RSV co-infection with other respiratory viruses was observed in 41/503 (8.2%) samples: rhinovirus (n =18) and adenovirus (n=11) were co-detected most frequently with

253 RSV. Comparative RSV incidence data was obtained from a contemporaneous respiratory
254 virus surveillance study that recruits patients aged <59 months admitted with acute lower
255 respiratory tract infection at the pediatric ward of the Kilifi County Hospital (KCH, see
256 Supplementary Methods). The frequency of RSV occurrence among children <5 years of age
257 differed significantly (p-value <0.001) between the outpatient and inpatient setting during
258 the period of December 2015 to June 2017. The RSV viral load was similar between the
259 inpatient and outpatient settings (**Supplementary Figure 1(B)**). Across both inpatient and
260 outpatient care settings, the peak time for RSV case detections occurred from November to
261 May the following year (**Supplementary Figure 3**). The distribution of clinical symptoms
262 among patients attending the nine outpatient health facilities, and by age and gender are
263 shown in **Supplementary Table 2.**
264
265 **RSVB genome sequences from Kilifi**. WGS and data assembly was successful for
266 299/408 (73.3%) RSVB positive samples collected between December 2015 and June 2017
267 from the 9 selected health facilities in Kilifi county, Kenya. The remaining 26% (109/408)
268 sampling fraction were either of low viral load and we could not obtain more than 4 PCR
269 amplicons (81.6%) or were sequenced at insufficient depth or quality for genome assembly
270 (18.4%). 295/299 genomes were coding-complete (all the 11 RSV coding genes were
271 assembled) and were used in subsequent analysis. The median genome length was 15205
272 (range 11519 to 15257 nt). All the sequenced viruses belonged to the RSVB BA genotype,
273 characterized by the presence of 60-nt duplication in the C-terminal region of the G
274 glycoprotein gene. Genome coverage did not vary by rRT-PCR Ct value (**Figure 2(A)**).
275

Genome-wide sequence diversity and evolution. The sequence data showed linear relationship between genetic change and time (root-to-tip correlation coefficient of 0.82, **Figure 2(B)**); a temporal signal that supported the use of molecular clock models. Bayesian analyses estimated an evolutionary rate of 1.063×10^{-3} (95% HPD: 9.2422×10^{-4} – 1.2063×10^{-3}) nucleotide substitutions/site/year, and the genetic diversity traced back to a common ancestor dated in 2014 (date in decimal format: 2014.89; 95% HPD: 2014.63 – 2015.14). Demographic reconstruction (**Figure 2(C)**) showed seasonal periodicity in relative genetic diversity that broadly mirrored RSVB incidence in the two epidemics. The peak genetic diversity occurred in January followed by an inter-epidemic bottleneck that indicated lineage or variant displacement events between epidemics.

The Kilifi genome sequences contained 838 consensus level non-gap single nucleotide polymorphisms (SNPs), 554 (66%) of which were parsimony informative. 503/554 (91%) SNPs were located within coding regions, of which 332/503 (66%) were non-synonymous and non-uniformly distributed across the RSV genome. Non-synonymous changes peaked in density at the mucin-like domains of G gene; in the N-terminal of fusion (F) gene; as well as in the N- and C-terminals of RNA-dependent RNA polymerase (L) gene (**Figure 2(D)**).

Selection pressure analyses showed that G and SH glycoproteins had higher global non-synonymous (dN)/synonymous (dS) substitution rate ratio estimates than all other genes (**Table 2**). SLAC analyses identified three amino-acid sites (135, 217 and 285) in G gene under significant positive selection (P <0.1). In addition, we used MEME and FUBAR methods to identify codons under pervasive and episodic (diversifying) positive selection. MEME analyses detected 3 diversifying codons in the F gene, and 11 in the G gene (P < 0.1)

(**Table 2**). The FUBAR method identified 2 codon sites in F gene and 7 in G gene, under episodic positive selection with significant support (posterior probability >0.9) (**Table 2**).

Viral introductions and spread within Kilifi. Genomic phylogenetic and phylogeographic analyses afforded an in-depth look into the introduction and spread of RSVB in the study population between December 2015 and June 2017. To determine viral introductions, we analyzed the Kilifi RSVB genomes (*n=295*) in the context of other global RSVB genomes (*n=500*). Sequences from Kilifi fell into eight major clades (numbered I to VIII, **Figure 3(A)**) in which they clustered more closely among themselves and less within the diversity of contemporaneous strains, indicating at least eight separate introductions into the study population. The absence of external/global strains within the eight clades might suggest local diversification of viruses, although we cannot exclude additional importation from unsampled locations. Still, the low availability of RSVB sequence data from other regions within Kenya or neighboring countries makes it difficult to estimate the precise number of viral introductions.

The first four introductions (I-IV) comprised viruses confined to a single RSV epidemic (**Figure 3(B)**). Clades I, II and IV comprised 25 viruses in total, from the 2015/16 epidemic, while clade III comprised 73 viruses uniquely from the 2016/17 epidemic. The other four clades (V-VIII) comprised 76, 16, 36 and 69 viruses, respectively, a mix from both RSV epidemics. All the eight introductions circulated within more than one location (table inset, **Figure 3(A)**).Viruses in clade III in Figure 3(A) were closely related to RSVB strains collected in January and August 2016 in UK and Australia, respectively. Clades I, IV, V and VIII, shared common ancestors with sequences from Australia. Viruses in Clade VI were closely related to a sequence collected in USA in 2015, while those in clade II were closely

related to a Dutch strain. There is a strong potential for bias in these results due to heterogeneous sampling globally. We used TreeTime (42) for maximum likelihood dating of the inferred clades, and the most likely dates of introduction were placed between July 2007 and September 2014 (**Supplementary Table 3**).

Analyzed separately, the time measured MCC tree of Kilifi genomes showed phylogenetic clustering by epidemic, which depicted a temporally structured viral population (**Figure 3(B)**). Our analysis of the geographical signal revealed high AI and PS values (**Table 3**), suggesting relatively extensive viral migration dynamics among study locations. The association between phylogenetic clustering and study location was significant (p value < 0.001) in at least 8/9 study locations as revealed by the maximum clade size values (**Table 3**). Differences in the observed and expected MC values (**Table 3**) suggested that Mavueni exhibited the most spatial structure (difference of 8.7) and Mtondia had the least (difference of 0.3).

We compared the model fit for symmetrical (reversible) and asymmetrical (non-reversible) discrete diffusion models with BSSVS procedure. The asymmetrical CTMC model gave a better fit (marginal likelihood, path sampling = -264.14 and stepping-stone sampling = -316.3) compared to the reversible's marginal likelihood path sampling of -287.44 and stepping-stone sampling of -351.9. The non-reversible model is a more realistic description of the diffusion process that allows location exchange rates to vary according to directionality (37). Mavueni may have played a central role in virus dissemination to other study locations: it was the most probable location for most ancestral nodes with location posterior support >0.9, (**Figure 4(A)**), and viral lineage movements from Mavueni were statistically supported with Bayes Factor >1000 (**Table 4**). In addition, there was a relatively high genomic diversity in Mavueni (red-colored taxa in **Supplementary Figure 4**). The

inferred viral movements between the nine study locations are visualized as geographic links in **Figure 4(B)**. There were very few statistically supported viral movements between Sokoke and other study locations.

Phylogenomic clusters. The RSVB viruses in Kilifi were further summarized as phylogenomic clusters based on genetic similarity of consensus genomes. For this, we examined the distribution of maximum likelihood pairwise evolutionary distances among the 295 samples (43365 pairs). The genetic distance distribution was multi-modal (**Figure 5 (A)**), and we defined a phylogenomic cluster as a group of viruses (≥2) within an epidemic that contained evolutionary pairwise distances of <0.0024 nucleotide substitutions per site (red dashed line **Figure 5 (A)**) or 37 pairwise nucleotide differences. The distance threshold of <0.0024 nucleotide substitutions per site demarcated the 17th percentile of the entire distance distribution and was additionally supported by clear separation of viruses into closely related phylogenetic clusters with >0.9 posterior probability support.

Overall, at least 20 phylogenomic clusters were identified, all of which were supported by strong posterior probability support (>0.9); nine in the 2015/16 epidemic (**Figure 6(A)**) and eleven (**Figure 6(B)**) in the 2016/17 epidemic. The lowest within-cluster genetic diversity was found in cluster 12 in the 2016/17 epidemic and the highest in cluster four in the 2015/16 epidemic (**Figure 5(B)**). Genetically, cluster five in the 2015/16 epidemic fell basal to cluster 17 in the 2016/17 epidemic and both made the 6th RSVB introduction (clade VI in **Figure 3(A)**), a phylogenetic pattern compatible with in-situ virus evolution and persistence. The clusters ranged in size, from 2 to 72 viruses. Overall, there was abundant viral diversity in Pingilikani in both epidemics (13 clusters), while Matsangoni and Ngerenya

had the lesser diversity (six clusters). In particular, only one cluster or viral lineage was circulating in Matsangoni in the 2015/16 epidemic.

Sample collection dates were used to estimate the duration of each phylogenomic cluster. Clusters 20 (172 days), 10 (173 days) and six (211 days) circulated over a wide temporal scale up to several months. Geographically, clusters one and 20 were the largest in size and most pervasive circulating widely in all the nine study locations (**Figure 6**). Two viruses in the 2016/17 epidemic (marked with '*' in **Figure 6**) were singletons and could not be assigned to any phylogenomic cluster, likely an artifact of limited sampling. The number of clusters identified in each study location was weakly associated with the number of genomes from that location (Spearman correlation *rho* = 0.2, *P* = 0.6)

Amino acid diversity in F and G genes. Given their role in the initial phases of infection and as the major antigens for eliciting neutralizing antibody responses, we examined amino-acid variation in the two major glycoproteins on the surface of the RSV virion, the attachment (G) glycoprotein and the fusion (F) glycoprotein. The F protein is also the major target for antiviral immuno-prophylaxis. In all the sequenced viruses, there were 37 and 93 amino-acid (codon) sites containing non-synonymous changes in F and G gene, respectively, and the frequency of the observed non-synonymous variants was up to 24.7% and 90.4% in F and G gene, respectively. While most of these variable sites contained a single non-synonymous change, one codon in F contained two changes (K419N/R), and three codons in G contained multiple non-synonymous changes (T107A/D, P289L/S and E290A/G/V). The K419N and K419R substitutions were seen in a subset (n=6) of sequences in cluster 13 and in cluster 17, respectively. The T107D substitution was observed in a subset (n=7) of sequences in cluster 13, while the T107A was seen in all other clusters

except for seven, eight and nine. The P289L and P289S substitutions were observed in cluster seven and eight, and six, respectively. The E290A, E290G and E290V substitutions were observed in clusters 13, 6 and 4 and 14, respectively. These amino-acid variants were not limited to a study location and were relative to the earliest dated sequence in the dataset (14 December 2015).

In F gene, viruses in clusters 10 and 11 (**Figure 6(B)**) in the 2016/17 epidemic contained two unique amino-acid changes (N99T and T129I). Additionally, three non-synonymous substitutions (V103A, L172Q and S173L) located in the F protein antigenic site V were unique to clusters seven, eight and nine in the 2015/16 epidemic. In G gene, 90% of the genomes contained T107A non-synonymous substitution, 64% had P217L substitution, and 45% contained a H285Y substitution, compared to the earliest sequence. A number of amino-acid changes occurred at the node between cluster nine and other viruses (marked A in **Figure 3(A)**) and these included N213D, A227I, P255S, and N303E in G gene; and T19A and P125L in F gene. A non-synonymous change from CAA(Q) resulted in early Stop codon (TAA) at amino acid position 311 in 135/295 (45%) sequenced viruses.

Several amino-acid sites appeared to undergo reversal mutation, for instance in G gene, T/S75S/T occurred within the 2015/16 epidemic, while T/I252I/T, L/P270P/L and G/S135S/G occurred in both epidemics. The I/L542L/I reversal mutation in F gene was seen in the 2015/16 epidemic. The T/S75S/T (G gene) and I/L542L/I (F gene) reversal substitutions were both observed in phylogenomic clusters seven and eight that comprised clade II (**Figure 3(A)**). The L/P270P/L reversal mutation occurred in clusters 5, 7 and 17. Reversible evolution may contribute to the escape from the human population immune response, thereby facilitating viral transmission (43).

Emerging genetically distinct variant in 2016/17 epidemic. Several non-synonymous changes were unique to phylogenomic cluster 20 in the 2016/17 epidemic (**Figure 6(B)**), including K68Q in F gene; Y90H, L91F, T225N, T273I, and A301T in G gene. These six non-synonymous changes were co-occurring in all the sequences in cluster 20 and were detected at an intermediate frequency (41% of samples in 2016/17) in the study population. Importantly, the amino-acid change K68Q is at the antigenic site ∅, the binding site of monoclonal antibody (mAb) MEDI8897. A variant with the mutation K68N was previously detected in 2% of sampled viruses in the US (44). Other amino-acid changes distinctive to the cluster 20 are listed in **Supplementary Table 4**. This phylogenomic cluster, which was also unequivocally an independent introduction (clade III, **Figure 3(A)**), was the most prevalent and its detection through genomic analysis rules out any sample contamination and sequencing errors.

We went further and checked whether this emerging variant was also present in the wider Kilifi county based on G gene sequence data of samples collected from pediatric ward admissions (<59 months of age) at the Kilifi County Hospital (KCH, formerly Kilifi District Hospital) during the 2016/17 RSV epidemic. KCH provides primary care and inpatient referral services to a larger catchment area. We found that 32% of RSVB positive hospital samples in the 2016/17 epidemic clustered with the emerging variant and likewise contained the non-synonymous changes Y90H, L91F, T225N, T273I, and A301T in G gene, which further increased the overall prevalence of clade III in the 2016/17 epidemic.

We also investigated whether the emerging variant persisted in Kilifi after the 2016/17 RSV epidemic, based on G gene sequences of RSVB positive samples collected from pediatric admissions at KCH during the 2018/19 epidemic. To note, there were no RSVB cases in the 2017/18 epidemic in the KHDSS study (outpatient surveillance in the nine health

facilities) (**Supplementary Figure 5**) and only 2 RSVB cases were detected in the in-patient pediatric admissions at KCH (<0.01% of RSV positive samples). We found that viruses in 2018/19 epidemic clustered closely with clade III or cluster 20 and the branch leading to the 2018/19 RSVB epidemic shared a common ancestor with cluster 20 (**Supplementary Figure 6**). The 2018/19 viruses likely descended from the same introduction as clade III, suggesting that the mutations defining clade III were epidemiologically important. It is possible that RSVB underwent a significant population (or genetic) bottleneck after the 2016/17 epidemic and older viruses were eliminated given that no RSVB viruses or persistent variants were transmissible in the 2017/18 epidemic.

Amino-acid diversity in other regions of the genome. Additional variable amino acid sites observed in the other genomic regions are listed in (**Supplementary Table 4**). Prominent amino acid changes occurring in more than one phylogenomic cluster are underlined in **Supplementary Table 4**. The emerging variant (cluster 20) discussed above had two other unique amino acid differences in NS2 (non-structural protein 2) and L (polymerase) genes. Even in other genomic regions, phylogenomic clusters 5 (2015/16 epidemic) and 17 (2016/17 epidemic), which are nested within the same virus introduction (clade VI, **Figure 3(A)**), shared amino-acid differences. Clusters 10 to 14 had the same amino-acid differences in F and L genes.

Discussion

Here we report whole genome molecular epidemiology and phylodynamics of respiratory syncytial virus group B providing a detailed view of the introduction and spread the virus in Kilifi county, coastal Kenya. The results were obtained from genomic analyses of

295 samples originating from representative sampling across the KHDSS area, over two consecutive RSV epidemics. Phylogenetic analyses revealed multiple virus introductions, each introduction commonly circulating in all the study locations, suggesting substantial spatial spread and transmission between locations in a relatively short time. Although RSV surveillance has improved in many regions across the world, publicly available RSVB genomic data from recent years is quite insufficient and may have limited our inference of spatial and temporal placement of virus introductions in Kilifi.

Analysis of RSVB transmission dynamics in Kilifi suggested extensive viral migrations among the study locations associated with the nine health facilities as well as strong spatial substructures within each location. The substructures might represent predominant local transmission and diversification processes. Using the BSSVS approach and an asymmetrical diffusion model to reconstruct RSVB dispersal, we observed significant and strong support for epidemiological links between one central location (Mavueni), located at the intersection of the main roads through the KHDSS and other study locations (BF > 1000, Table 4). However, we note that Mavueni had a larger proportion of RSVB genomes compared to other locations, which might bias the ancestral location estimates. Improved road infrastructure and transportation within KHDSS has facilitated mobility thus increasingly connecting the local population and expanding virus transmission networks.

Temporal sequence divergence and accumulation of nucleotide substitutions was detectable over the sampling timeframe and varied by epidemic intervals as shown by two separate groups of tips in the linear regression plot (Figure 2B). Time-scaled phylogenies also exhibited a marked epidemic behavior indicative of chronological generation of new variants. The two epidemics were characterized by multiple clades for viruses sampled within the same epidemic, indicating continued transmission generated and sustained by

490 increasing spatial connectivity in the wider Kilifi county. Changes in the relative genetic
491 diversity coincided with RSVB case detection and captured fine temporal resolution of
492 changes in the viral population size, which also implied sufficient sampling density (14).
493 While estimating nucleotide substitution rate is useful in revealing the dynamics and
494 processes of viral evolution (45), we could not directly compare our estimate of the mean
495 nucleotide substitution rate with previous studies due to varying sampling timescales,
496 different molecular clock (fixed vs. relaxed) and coalescent models, as well as epidemiologic
497 variations.

498 Phylogenetic clusters have been used to investigate epidemiologically significant HIV
499 hotspots and characterize groups burdened by a high rate of HIV transmission (46). We
500 inferred 20 phylogenomic clusters of closely related viruses based on genetic relatedness,
501 which we put forward as potential transmission units. Stochastic difference in
502 transmissibility or circulation, or infection rates could explain the differences in prevalence
503 of the different phylogenomic clusters.

504 Using SLAC, MEME, and FUBAR methods, there was limited detectable diversifying
505 selection for the amino acid substitutions that characterized or defined the different
506 phylogenomic clusters. Only three cluster-defining codons in G (144, 294 and 303) and three
507 in F (125, 172 and 173) glycoproteins were subject to episodic positive selection. It is likely
508 that the observed codon replacements follow non-selective epidemiological processes and
509 these substitutions are compensatory mutations to retain function, or hitchhikers carried
510 along by chance (47), or that immune driven positive selection could not be identified by the
511 three methods used in this study. Nonetheless, the implications of these mutations on
512 protein function, viral evolution and fitness are uncertain and warrant further functional
513 investigation.

514 A previous study showed that viruses carrying the K68N substitution in the US,
515 affected the binding of MEDI8897 (48). MEDI8897 is an RSV pre-F-specific human mAb with
516 an extended serum half-life, under clinical evaluation as a passive immunization of all
517 infants entering their first RSV season (49). It is probable the K68Q substitution identified in
518 Kilifi promoted evasion of pre-existing immune responses. The unequivocal support for the
519 monophyly of the cluster 20 as an introduction into Kilifi (clade III, **Figure 3(A)**), further
520 supports that this phylogenetic clade was a single (epidemiologically successful)
521 introduction event. Surprisingly, the K68Q F gene amino-acid substitution observed in
522 phylogenomic cluster 20 was not detected as under any selection pressure even though this
523 residue is located at structurally determined mAb binding epitopes (50). Perhaps an
524 explanation for this is the low rate of nonsynonymous evolution (conversely, high sequence
525 conservation) at codon 68 in our dataset. In any case conventional approaches for
526 measuring selection pressure consistently detect positive selection only at codon sites with
527 high rates of nonsynonymous evolution (51).
528 A high sequence conservation was reported in the MEDI8897 binding site among
529 naturally occurring RSV isolates collected from 1965 to 2014 (50). Our study provides a
530 novel sequence polymorphism (K68Q) within the MEDI8897 binding site with a frequency of
531 nearly 50% in the 2016/17 RSV epidemic in our study population. Functional
532 characterization is required to determine MEDI8897 neutralization and/or binding to viruses
533 containing the K68Q mutation. Additionally, the viruses with the K68Q mutation in 2016/17
534 epidemic possessed five distinctive amino-acid mutations in G gene, including Y90H and
535 L91F in two consecutive codons. We cannot exclude the possibility that these are relevant
536 antigenic epitopes. Our study underscores the need for continued genomic surveillance of

contemporary clinical strains particularly at F and G protein antigenic sites as this has implications on RSV therapeutic and vaccine development.

RSVB viruses containing A103V/L172Q/S173L amino acid-changes in the F protein were also detected during the 2015/16 RSV season in USA and estimated to have likely emerged around 2014 (44) and in China (52), suggesting global circulation of this variant. However, unlike in the US, none of the samples from 2016/17 in Kilifi had these three substitutions probably due to removal by purifying selection.

In conclusion, we present the utility of genomic analyses to investigate virus transmission and genetic diversity including detection of a novel antigenically distinct variant. Further studies are required to determine whether the K68Q mutation is adaptive and/or a result of escape from antibody-mediated selection and constitutes a naturally acquired antiviral resistance-associated mutation that potentially disrupts neutralizing antibody recognition and binding. An important future surveillance effort for us is to assess if the K68Q mutation has become more prevalent and gradually fixed since the 2016/17 epidemic. Additional sequencing of RSVB from other regions in Kenya and neighboring countries is essential to refine evolutionary dynamics and draw better conclusions about geographic origins of viral introductions in the study population in Kilifi. The present study makes publicly available a large number of newly acquired coding-complete RSVB genomes useful for further molecular evolution studies.

Data availability

The replication data and analysis scripts for this manuscript are available from the Harvard Dataverse: ???. Some of the clinical dataset contains potentially identifying information on

560 participants and is stored under restricted access. Requests for access to the restricted
561 dataset should be made to the Data Governance Committee (dgc@kemri-wellcome.org).
562
563 **Competing interests**
564 The authors declare no competing interests.
565
566 **Acknowledgements**
567 We thank all the study participants for their contribution of samples and data. We also
568 thank the Dispensary / Health Centre management committees for allowing us to conduct
569 the study within their health facilities. We are grateful to the field study team for participant
570 recruitment and the laboratory staff of the KEMRI-Wellcome Trust Research Programme /
571 Virus Epidemiology and Control research group. We would also like to thank D. Collins
572 Owuor for his assistance with the BaTS software and Mark Otiende for his assistance with
573 demographics data. This paper is published with the permission of the Director of KEMRI.
574
575 **Funding**
576 This work was supported by the Wellcome Trust [grant 102975, 203077]. CNA is supported
577 by the Initiative to Develop African Research Leaders (IDeAL) through the DELTAS Africa
578 Initiative [DEL-15-003]. The DELTAS Africa Initiative is an independent funding scheme of the
579 African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa
580 (AESA) and supported by the New Partnership for Africa's Development Planning and
581 Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust
582 [107769/Z/10/Z] and the UK government. The views expressed in this publication are those

583 of the authors and not necessarily those of AAS, NEPAD Agency, Wellcome Trust or the UK
584 government.

585

586 **Figure legends**

587 **Figure 1**(**A**) A map showing the geographical area covered in the Kilifi Health Demographic
588 Surveillance System (KHDSS), expanded from a map of Kenya. The nine participating public
589 health facilities are indicated in the map. The dark lines within the polygons indicate the
590 road structure within KHDSS. The maps were rendered using QGIS 2.18.17
591 (https://www.qgis.org/) (**B**) Monthly RSVB occurrence by study location: temporal and
592 spatial distribution of RSVB positive cases (left Y axis) and number of clinical samples tested
593 (right Y axis) from each participating health facilities. Abbreviations: CHA = Chasimba, JAR =
594 Jaribuni, JUN = Junju, MAV = Mavueni, MAT = Matsangoni, PIN = Pingilikani, NGE =
595 Ngerenya, SOK = Sokoke, MTO = Mtondia

596

597 **Figure 2**(**A**) Genome coverage for each virus isolate versus the viral load (rRT-PCR Cycle
598 threshold (Ct) value). (**B**) Root-to-tip divergence as a function of sampling time colored by
599 study location. (**C**) Relative genetic diversity through time estimated using the Gaussian
600 Markov Random Field (GMRF) Skygrid model. Solid line represents mean relative genetic
601 diversity while the corresponding dashed lines indicate the 95% HPD intervals. (**D**) Relative
602 frequencies of potential non-synonymous changes across codon-aligned RSV genome
603 sequences. The frequencies for each codon position are calculated as the number of non-
604 synonymous nucleotide substitutions for all pairwise comparisons in a sequence alignment,
605 while excluding ambiguous bases or insertions. Abbreviations: CT = cytoplasmic, TM =
606 transmembrane, CCD = central conserved domain; SP = signal peptide; RdRp = RNA

607 dependent RNA polymerase, Cap = capping, and MT = methyltransferase, CD = connector
608 domain, CTD = C-terminal domain.
609
610 **Figure 3**(**A**) Phylogenetic placement of RSVB genomes from Kilifi relative to sequences from
611 other geographical regions. The clade or introduction assignments are indicated as I to VIII
612 on the trees and the circles at the branch tips are colored by location or country of origin.
613 Only the contemporaneous (global) sequences that are closely related to or share common
614 ancestors with RSVB genomes from Kilifi, are colored. The black circles represent other
615 global sequences used in the phylogenetic analysis. Geographic distribution of the RSVB
616 introductions in Kilifi is shown in the inset table. Abbreviations: CHA = Chasimba, JAR =
617 Jaribuni, JUN = Junju, MAV = Mavueni, MAT = Matsangoni, PIN = Pingilikani, NGE =
618 Ngerenya, SOK = Sokoke, MTO = Mtondia. (**B**) Maximum likelihood phylogeny (1000
619 bootstrap resampling) showing the temporal distribution of the RSVB introductions in Kilifi.
620 The individual sequences from Kilifi are colored by respective RSV epidemic. The black
621 circles represent other global sequences used in the phylogenetic analysis. Bootstrap
622 support values are shown for the most basal nodes of the inferred introductions.
623
624 **Figure 4**(**A**) Temporal-scaled phylogeographic maximum clade credibility tree of Kilifi RSVB
625 genomes. Branch colors indicate the modal location (most probable reconstructed location
626 at each node) inferred under the asymmetrical discrete phylogeographic model. Location-
627 state posterior probabilities are shown next to relevant nodes along with the posterior
628 support. (**B**) Spatial diffusion pathways of RSVB in Kilifi. Well supported (Bayes Factor >6)
629 discrete diffusion asymmetric rates of viral movement between the study locations are

630 indicated. The type and thickness of the arrows represents the relative strength of the
631 diffusion rate.
632
633 **Figure 5**(**A**) Histogram of the whole genome sequences patristic distance frequency
634 distribution. The vertical red dashed line corresponds to the 17th percentile distance
635 threshold (0.0024 expected nucleotide substitutions per site) for which phylogenomic
636 clusters were identified. The distances were measured in units of nucleotide substitutions
637 per site and extracted from a maximum likelihood phylogeny (1000 bootstrap resampling).
638 (**B**) Maximum pairwise patristic distances for sequences belonging to the different
639 phylogenomic clusters.
640
641 **Figure 6** The assigned phylogenomic clusters are shown on maximum likelihood phylogenies
642 (1000 bootstrap resampling): 9 clusters in the 2015/16 epidemic (**A**) and 11 clusters in the
643 2016/17 epidemic (**B**). Geographical composition and number of viruses/sequences per
644 phylogenomic cluster are shown (inset tables). The '*' in (B) indicates sequences that were
645 not assigned into a phylogenomic cluster. All sequences had epidemiological and
646 geographical information. Amino-acid changes in F and G glycoproteins unique to the
647 emergent variant (cluster 21) in the 2016/17 are labeled on the major branch leading to the
648 cluster. Scale bar represents the number of substitutions per site.
649
650 **Supplementary Figure 1** (**A**) Violin plots show the distribution (median, IQR) of RSVB qRT-
651 PCR Cycle threshold (C_t) values (<=40) for outpatient participants. (**B**) Comparison of C_t
652 distribution between inpatient (KCH) and the outpatient facilities for children <5 years for

653 the period of December 2015 to June 2017. The threshold used for determining positive and
654 negative samples is shown by dashed line (C_t=35.0).
655
656 **Supplementary Figure 2** Frequency (primary Y axis) and proportion (secondary Y axis) of
657 RSV across different age categories in both genders (dashed and dotted lines) from
658 participants presenting to the nine outpatient health facilities over the period December
659 2015 to June 2017.
660
661 **Supplementary Figure 3** Distribution by month of RSVB positive samples collected from the
662 inpatient (severe pneumonia admissions to KCH) and outpatient facilities over the period of
663 December 2015 to June 2017. Secondary Y axis records sample tested (dotted line) and
664 proportion (dashed line) that is RSVB positive per month.
665
666 **Supplementary Figure 4** Temporal-scaled phylogeographic maximum clade credibility tree
667 showing sequences of samples collected in Mavueni location (red taxa) in Kilifi County,
668 coastal Kenya.
669
670 **Supplementary Figure 5** The number of RSVB cases through time (2015 to 2019) for severe
671 pneumonia pediatric admissions to Kilifi County Hospital. RSV epidemics (usually from
672 October to July of the following year) are indicated.
673
674 **Supplementary Figure 6** Maximum likelihood phylogeny of the attachment G glycoprotein
675 gene. The orange colored branches represent the recently emergent variant (phylogenomic

676 cluster 20). The purple colored branches represent the 2018/19 RSVB viruses collected from

677 severe pneumonia pediatric admissions to Kilifi County Hospital.

678

679

680 **Tables**

681 **Table 1** Characteristics of RSV positive and negative study participants by age, gender, and

682 the 9 outpatient health facilities in Kilifi County, coastal Kenya.

Characteristic	NPS virus positive(n)	%	NPS virus negative (n)	%	Total(n)	P value
	(n=503)	6.19	7624	93.8	8127	
Age in years						
Mean			149.4		144.0	
Median (IQR*)	20 (8-43)		54(18-194)		49 (17-188)	
Sex						
Male	231	6.7	3195	93.3	3426	0.078
Female	272	5.8	4428	94.2	4700	
Age Category						
0–5 mo	85	11.3	665	88.7	750	<0.001
6–11 mo	84	11.0	681	89.0	765	
12–23 mo	102	9.1	1021	90.9	1123	
24–35 mo	76	9.6	712	90.4	788	
3–4 y	64	6.7	898	93.4	962	
5–9 y	37	3.4	1046	96.6	1083	
10–19 y	21	2.0	1024	98.0	1045	
20–49 y	23	2.1	1098	98.0	1121	
50–100 y	11	2.3	479	97.8	490	
Health Facility						
Matsangoni	58	6.1	895	93.9	953	0.003
Ngerenya	58	6.5	840	93.5	898	
Sokoke	50	5.7	826	94.3	876	
Mtondia	66	6.9	898	93.2	964	
Mavueni	79	8.6	838	91.4	917	
Jaribuni	43	5.1	803	94.9	846	
Chasimba	67	7.5	826	92.5	893	
Junju	40	4.3	889	95.7	929	
Pingilikani	42	4.9	809	95.1	851	

683

684 **Table 2** The predicted nature of selection pressures acting on each genomic region: 1st

685 column shows the computed mean dN/dS rate ratio using SLAC and the 2nd column shows

686 amino-acid sites in F and G gene under episodic selection as identified by MEME analyses.

687 Sites also detected using the FUBAR method, in addition to MEME, are underlined.

Non-synonymous (dN)/synonymous (dS) substitution rate ratio per site		**Sites subject to episodic positive/diversifying selection**
NS1	0.12	*G gene*
NS2	0.236	135, 144, 154, 172, 208, 217,
N	0.0832	285, 291, 294, 298, 303
M	0.0525	252*
P	0.0642	
F	0.179	*F gene*
G	0.487	125, 172, 173
SH	0.426	
M2-1	0.264	
M2-2	0.267	
L	0.122	

688 * This site was detected by the FUBAR method only.

689

690

691 **Table 3** Results of Bayesian analysis of phylogeographic structure of RSVB viruses in Kilifi,
692 coastal Kenya, 2015-2017. *P* values correspond to the proportion of trees from the expected
693 (null) distribution equal to, or more extreme than, the median posterior of the statistic.
694 Abbreviations: CHA = Chasimba, JAR = Jaribuni, JUN = Junju, MAV = Mavueni, MAT =
695 Matsangoni, PIN = Pingilikani, NGE = Ngerenya, SOK = Sokoke, MTO = Mtondia.

	Association Index (AI) (95% CI) †			Parsimony Score (PS) (95% CI) †			Maximum Clade size (95% CI) ‡			
Location	**Observed**	**Expected**	***P* value**	**Observed**	**Expected**	***P* value**	**Observed**	**Expected**	***P* value**	**Difference §**
ALL	14.8 (13.8-15.8)	31 (29.5-32.3)	0.0	132.6 (129-136)	208.4 (202-214.2)	0.0	-	-	-	-
CHA	-	-	-	-	-	-	4.6 (4-6)	1.84 (1.1-2.7)	10E-4	2.62
JAR	-	-	-	-	-	-	5.27 (5-6)	1.56 (1-2.08)	10E-4	3.71
JUN	-	-	-	-	-	-	6 (6-6)	1.55 (1-2.03)	10E-4	4.45
MAT	-	-	-	-	-	-	6.32 (4-10)	1.71 (1-2.2)	10E-4	4.61
MAV	-	-	-	-	-	-	11 (11-11)	2.3 (1.76-3)	10E-4	8.7
MTO	-	-	-	-	-	-	2 (2-2)	1.7 (1-2.3)	0.21	0.3
NGE	-	-	-	-	-	-	3.65 (2-4)	1.64 (1-2.2)	10E-4	2.01
PIN	-	-	-	-	-	-	3.1 (3-4)	1.73 (1-2.2)	0.0084	1.37
SOK	-	-	-	-	-	-	4 (3-5)	1.8 (1.1-2.3)	10E-4	2.2

696 †AI and PS metrics were determined for all locations combined.
697 ‡ Maximum clade size was determined for each specific location.
698 § Difference between observed and expected (null) clade size.

699

700 **Table 4** Statistically supported state transitions indicating viral migration events. Bayes

701 factor >6 was considered significant.

Transition between		Distance (Km)$^{\Phi}$	Bayes factor	Mean indicator $^{\delta}$	Mean rate
Chasimba	Mtondia	21.5	6.2	0.46	1
Jaribuni	Junju	17.5	6.7	0.48	1
Chasimba	Ngerenya	25.5	8.3	0.53	0.8
Sokoke	Matsangoni	20.6	9.8	0.57	0.9
Mtondia	Junju	24	12.2	0.63	0.76
Chasimba	Pingilikani	6.2	13.7	0.65	0.9
Pingilikani	Mtondia	21.6	18.2	0.71	1
Junju	Jaribuni	17.5	20.8	0.74	1
Mavueni	Mtondia	12.5	28.5	0.80	1
Matsangoni	Ngerenya	12.3	35.0	0.83	0.75
Mavueni	Matsangoni	29.3	179.9	0.96	0.87
Mavueni	Junju	11.5	434.6	0.98	0.87
Junju	Pingilikani	5.4	1166.6	0.99	1.4
Ngerenya	Mtondia	4.6	1624.5	1.00	1.3
Mavueni	Ngerenya	17	2062.5	1.00	1.2
Sokoke	Mtondia	10.9	12678.4	1.00	1.7
Mavueni	Jaribuni	12.8	24571.2	1.00	1.09
Mavueni	Pingilikani	9.8	30243.2	1.00	1.5
Mavueni	Chasimba	12.1	196620.8	1.00	1.75
Mavueni	Sokoke	16.5	393248.8	1.00	1.78

702 $^{\Phi}$ Great circle distance estimates between centroid longitude and latitude of the respective locations.
703 $^{\delta}$ The posterior probability of observing non-zero migration rates in the Bayesian sampled trees.
704

705

706 **Supplementary Table 1** Information (accession number, country of sampling and collection

707 date) on RSVB genome sequences of >14,000 nucleotide length retrieved from NCBI

708 GenBank.

709

710

711 **Supplementary Table 2** Frequency of distribution of symptoms among study participants by

712 9 outpatient health facilities, age, and gender, between December 2015 and June 2017.

Characteristic	Fever, %	Chest indrawing, %	Crackles, %	Wheeze, %	Cough, %	Nasal discharge, %	Nasal flare, %	Difficulty breathing, %	Total participants, n
Health Facility									
Matsangoni	82.8	0.0	3.5	6.9	98.3	72.4	3.45	10.3	953
Ngerenya	84.5	1.7	1.7	0.0	98.3	65.5	0.0	10.3	898
Sokoke	84.0	2.0	2.0	6.0	100	80.0	4.0	18.0	876
Mtondia	81.8	6.1	3.0	1.5	97.0	78.8	4.6	7.6	964
Mavueni	74.7	3.8	2.5	1.3	97.5	77.2	5.1	15.2	917
Jaribuni	79.1	2.3	4.7	9.3	97.7	83.7	0.0	4.7	846
Chasimba	65.7	3.0	1.5	6.0	97.0	71.6	3.0	9.0	893
Junju	80.0	0.0	5.0	7.5	95.0	90.0	2.5	20.0	929
Pingilikani	85.7	2.4	2.4	9.5	95.2	76.2	4.8	40.5	851
***P* value**	**0.156**	**0.588**	**0.974**	**0.142**	**0.900**	**0.192**	**0.741**	**0.001**	
Age Category									
0–5 mths	77.7	3.5	2.4	5.9	98.8	69.4	10.6	23.5	750
6–11 mths	83.3	8.3	4.8	6.0	100	82.1	3.6	22.6	765
12–23 mths	84.3	2.0	3.9	5.9	97.1	73.5	1.0	13.7	1123
24–35 mths	82.9	1.3	2.6	4.0	96.1	81.6	2.6	9.2	788
3–4 yrs	79.7	0.0	1.6	6.3	98.4	76.6	1.6	9.4	962
5–9 yrs	83.8	0.0	2.7	2.7	91.9	83.8	0.0	8.1	1083
10–19 yrs	66.7	0.0	0.0	0.0	100	66.7	0.0	4.8	1045
20–49 yrs	52.2	0.0	0.0	0.0	91.3	87.0	0.0	4.4	1121
50–100 yrs	45.5	0.0	0.0	0.0	100	54.6	0.0	0.0	490
***P* value**	**0.003**	**0.041**	**0.883**	**0.830**	**0.129**	**0.150**	**0.009**	**0.010**	
Sex									
Male	80.5	3.5	3.0	5.6	98.7	73.2	2.2	16.5	3426
Female	77.9	1.8	2.6	4.0	96.3	79.4	4.0	12.1	4700
***P* value**	**0.478**	**0.252**	**0.756**	**0.406**	**0.094**	**0.099**	**0.231**	**0.166**	

713

714 **Supplementary Table 3** Rate of nucleotide substitution and the estimated date of

715 introduction or emergence of each clade.

		Rate of evolution (10^{-3} subs/site/y)		
Clade	**Estimated date**	**Mean**	**Lower 95% HPD**	**Upper 95% HPD**
I	September 2008	1.2	0.27	2.3
II	July 2007	1.18	0.33	2.32
III	September 2014	1.06	0.35	1.95
IV	August 2012	0.43	0.26	0.54
V	March 2013	1.22	0.87	2.02
VI	August 2012	0.67	0.58	1.1
VII	May 2013	1.64	1.03	2.17
VIII	March 2014	0.92	0.41	1.87

716

717

718 **Supplementary Table 4** Amino acid substitutions identified in RSV genomic regions, which

719 were distinctive to the various phylogenomic clusters.

Cluster (2015/16)	Amino acid substitutions in genomic regions	Cluster (2016/17)	Amino acid substitutions in genomic regions
1	*P gene* (Ile_{60}); *L gene* (Ile_{105}, Asn_{183})	10	*F gene* (Val_{5}, Thr_{99}, Thr_{129}); *G gene* (Ser_{101}, Phe_{267}); *NS2 gene* (Asn_{53}); *L gene* (Asp_{176})
2	*N gene* (Val_{97}); *G gene* (His_{144})	11	*F gene* (Val_{5}, Thr_{99}, Thr_{129}); *L gene* (Asp_{176}, Leu_{1708}); *G gene* (Phe_{267})
3	*N gene* (Val_{97})	12	*F gene* (Ser_{104}, Val_{5},); *G gene* (Phe_{267}); *NS2 gene* (Arg_{101}); *L gene* (Asp_{176})
4	G gene (Val_{290})	13	*F gene* (Val_{5},); *L gene* (Asp_{176}); *G gene* (Thr_{268}, Ala_{290}, Leu_{304})
5	*NS2 gene* (Thr_{5}); *L gene* (Ile_{1166}, Arg_{2066}); *G gene* (Val_{269}, Pro_{270})	14	*F gene* (Val_{5},); *L gene* (Asp_{176}, Phe_{1980}); *G gene* (Thr_{268}, Val_{290})
6	*F gene* (Thr_{16}); *L gene* (Ala_{1744}, Gly_{1787})	15	*N gene* (Val_{97}); *G gene* (Thr_{205}, Leu_{214}, Ala_{249}, Leu_{293}, Pro_{305})
7	*F gene* (Leu_{172}, Ser_{173}, Leu_{542}); *G gene* (Ser_{75}, Asn_{103}, Thr_{107}, Arg_{136}, Gly_{260}, Pro_{270}, Ile_{279}, Leu_{289}, Lys_{312}); *M2-1 gene* (Val_{181});	16	*F gene* (Cys_{25}); *G gene* (His_{144}); *L gene* (Glu_{304}, Ile_{1956})

	M2-2 gene (Asn_{26}, Lys_{49}); *N gene* (His_{216}); *NS1 gene* (Asp_{69}, Ile_{72}); *NS2 gene* (Thr_{26}); *SH gene* (Thr_{49}); *L gene* (Leu_{56}, Val_{715}, Ala_{1712}, Asn_{1736}, Ile_{2019}, Ile_{2069})		
8	*G gene* (Ser_{75}, Ile_{76}, Asn_{103}, Thr_{107}, Leu_{110}, Arg_{136}, Ser_{176}, Leu_{191}, Ser_{235}, Gly_{260}, Ile_{279}, Leu_{289}, Lys_{312}); *F gene* (Ala_{103}, Thr_{115}, Leu_{172}, Ser_{173}, Ile_{303}, Leu_{542}); *M2-1 gene* (Val_{181}); *M2-2 gene* (Asn_{26}); *N gene* (His_{216}); *SH gene* (Thr_{32}, Thr_{37}, Thr_{49}); *L gene* (Leu_{56}, Ala_{173}, Arg_{260}, Val_{715}, Ala_{1712}, Asn_{1736}, Ile_{2069})	17	*F gene* (Arg_{419}); *NS2 gene* (Thr_5); *L gene* (Ile_{1166}, Arg_{2066}); *G gene* (Val_{269}, Pro_{270}, Asp_{294})
9	*G gene* (Thr_{107}, Arg_{136}, Asn_{213}, Leu_{221}, Ala_{227}, Pro_{255}, Asn_{303}, Ile_{279}, Lys_{312}); *F gene* (Ala_{103}, Pro_{125}, Leu_{172}, Ser_{173}, Thr_{19}); *M2-1 gene* (Val_{181}); *M2-2 gene* (Asn_{26}); *M gene* (Met_{73}); *N gene* (His_{216}); *SH gene* (Ile_{26}, Thr_{49});	18	*P gene* (Ile_{60}); *L gene* (Ile_{105}, His_{141}, Asn_{183})

	L gene (Leu_{56}, Val_{715}, Ala_{1712}, $Asn_{1736,}$ $Tyr_{1897,}$ Val_{514})		
		19	*P gene* (Ile_{60}); *L gene* ($Ile_{105,}$ Asn_{183}); *G gene* (Pro_{315})
		20	*G gene* (His_{90}, Phe_{91}, Asn_{225}, Ile_{273}, Thr_{301}); *F gene* (Gln_{68}); *NS2 gene* (Arg_{80}); *L gene* (Ser_{184})

720
721

References

1. Pneumonia Etiology Research for Child Health Study, G. (2019) 'Causes of severe pneumonia requiring hospital admission in children without HIV infection from Africa and Asia: the PERCH multi-country case-control study', *Lancet*, 394/10200: 757-779.
2. Scheltema, N. M., Gentile, A., Lucion, F., Nokes, D. J., Munywoki, P. K., Madhi, S. A., Groome, M. J., Cohen, C., Moyes, J., Thorburn, K., Thamthitiwat, S., Oshitani, H., Lupisan, S. P., Gordon, A., Sanchez, J. F., O'Brien, K. L., Gessner, B. D., Sutanto, A., Mejias, A., Ramilo, O., Khuri-Bulos, N., Halasa, N., de-Paris, F., Pires, M. R., Spaeder, M. C., Paes, B. A., Simoes, E. A. F., Leung, T. F., da Costa Oliveira, M. T., de Freitas Lazaro Emediato, C. C., Bassat, Q., Butt, W., Chi, H., Aamir, U. B., Ali, A., Lucero, M. G., Fasce, R. A., Lopez, O., Rath, B. A., Polack, F. P., Papenburg, J., Roglic, S., Ito, H., Goka, E. A., Grobbee, D. E., Nair, H. and Bont, L. J. (2017) 'Global respiratory syncytial virus-associated mortality in young children (RSV GOLD): a retrospective case series', *Lancet Glob Health*, 5/10: e984-e991.
3. Dowell, S. F., Anderson, L. J., Gary, H. E., Jr., Erdman, D. D., Plouffe, J. F., File, T. M., Jr., Marston, B. J. and Breiman, R. F. (1996) 'Respiratory syncytial virus is an important cause of community-acquired lower respiratory infection among hospitalized adults', *J Infect Dis*, 174/3: 456-462.
4. Sullender, W. M. (2000) 'Respiratory syncytial virus genetic and antigenic diversity', *Clin Microbiol Rev*, 13/1: 1-15, table of contents.
5. Melero, J. A., Garcia-Barreno, B., Martinez, I., Pringle, C. R. and Cane, P. A. (1997) 'Antigenic structure, evolution and immunobiology of human respiratory syncytial virus attachment (G) protein', *J Gen Virol*, 78 (Pt 10): 2411-2418.
6. Trento, A., Casas, I., Calderon, A., Garcia-Garcia, M. L., Calvo, C., Perez-Brena, P. and Melero, J. A. (2010) 'Ten years of global evolution of the human respiratory syncytial virus BA genotype with a 60-nucleotide duplication in the G protein gene', *J Virol*, 84/15: 7500-7512.
7. Zlateva, K. T., Vijgen, L., Dekeersmaeker, N., Naranjo, C. and Van Ranst, M. (2007) 'Subgroup Prevalence and Genotype Circulation Patterns of Human Respiratory Syncytial Virus in Belgium during Ten Successive Epidemic Seasons', *J Clin Microbiol*, 45/9: 3022.
8. Agoti, C. N., Otieno, J. R., Gitahi, C. W., Cane, P. A. and Nokes, D. J. (2014) 'Rapid spread and diversification of respiratory syncytial virus genotype ON1, Kenya', *Emerg Infect Dis*, 20/6: 950-959.
9. Otieno, J. R., Kamau, E. M., Agoti, C. N., Lewa, C., Otieno, G., Bett, A., Ngama, M., Cane, P. A. and Nokes, D. J. (2017) 'Spread and Evolution of Respiratory Syncytial Virus A Genotype ON1, Coastal Kenya, 2010-2015', *Emerg Infect Dis*, 23/2: 264-271.
10. Bose, M. E., He, J., Shrivastava, S., Nelson, M. I., Bera, J., Halpin, R. A., Town, C. D., Lorenzi, H. A., Noyola, D. E., Falcone, V., Gerna, G., De Beenhouwer, H., Videla, C., Kok, T., Venter, M., Williams, J. V. and Henrickson, K. J. (2015) 'Sequencing and analysis of globally obtained human respiratory syncytial virus A and B genomes', *PLoS One*, 10/3: e0120098.
11. Neuzil, K. M. (2016) 'Progress toward a Respiratory Syncytial Virus Vaccine', *Clinical and Vaccine Immunology*, 23/3: 186-188.
12. Gerretsen, H. E. and Sande, C. J. (2017) 'Development of respiratory syncytial virus (RSV) vaccines for infants', *J Infect*, 74 Suppl 1: S143-s146.

13. Agoti, C. N., Otieno, J. R., Ngama, M., Mwihuri, A. G., Medley, G. F., Cane, P. A. and Nokes, D. J. (2015) 'Successive Respiratory Syncytial Virus Epidemics in Local Populations Arise from Multiple Variant Introductions, Providing Insights into Virus Persistence', *J Virol*, 89/22: 11630-11642.
14. Otieno, J. R., Agoti, C. N., Gitahi, C. W., Bett, A., Ngama, M., Medley, G. F., Cane, P. A. and Nokes, D. J. (2016) 'Molecular Evolutionary Dynamics of Respiratory Syncytial Virus Group A in Recurrent Epidemics in Coastal Kenya', *J Virol*, 90/10: 4990-5002.
15. Otieno, J. R., Kamau, E. M., Oketch, J. W., Ngoi, J. M., Gichuki, A. M., Binter, S., Otieno, G. P., Ngama, M., Agoti, C. N., Cane, P. A., Kellam, P., Cotten, M., Lemey, P. and Nokes, D. J. (2018) 'Erratum: Whole genome analysis of local Kenyan and global sequences unravels the epidemiological and molecular evolutionary dynamics of RSV genotype ON1 strains', *Virus Evol*, 4/2: vey036.
16. Nyiro, J. U., Munywoki, P., Kamau, E., Agoti, C., Gichuki, A., Etyang, T., Otieno, G. and Nokes, D. J. (2018) 'Surveillance of respiratory viruses in the outpatient setting in rural coastal Kenya: baseline epidemiological observations', *Wellcome Open Res*, 3: 89.
17. Dudas, G., Carvalho, L. M., Bedford, T., Tatem, A. J., Baele, G., Faria, N. R., Park, D. J., Ladner, J. T., Arias, A., Asogun, D., Bielejec, F., Caddy, S. L., Cotten, M., D'Ambrozio, J., Dellicour, S., Di Caro, A., Diclaro, J. W., Duraffour, S., Elmore, M. J., Fakoli, L. S., Faye, O., Gilbert, M. L., Gevao, S. M., Gire, S., Gladden-Young, A., Gnirke, A., Goba, A., Grant, D. S., Haagmans, B. L., Hiscox, J. A., Jah, U., Kugelman, J. R., Liu, D., Lu, J., Malboeuf, C. M., Mate, S., Matthews, D. A., Matranga, C. B., Meredith, L. W., Qu, J., Quick, J., Pas, S. D., Phan, M. V. T., Pollakis, G., Reusken, C. B., Sanchez-Lockhart, M., Schaffner, S. F., Schieffelin, J. S., Sealfon, R. S., Simon-Loriere, E., Smits, S. L., Stoecker, K., Thorne, L., Tobin, E. A., Vandi, M. A., Watson, S. J., West, K., Whitmer, S., Wiley, M. R., Winnicki, S. M., Wohl, S., Wolfel, R., Yozwiak, N. L., Andersen, K. G., Blyden, S. O., Bolay, F., Carroll, M. W., Dahn, B., Diallo, B., Formenty, P., Fraser, C., Gao, G. F., Garry, R. F., Goodfellow, I., Gunther, S., Happi, C. T., Holmes, E. C., Kargbo, B., Keita, S., Kellam, P., Koopmans, M. P. G., Kuhn, J. H., Loman, N. J., Magassouba, N., Naidoo, D., Nichol, S. T., Nyenswah, T., Palacios, G., Pybus, O. G., Sabeti, P. C., Sall, A., Stroher, U., Wurie, I., Suchard, M. A., Lemey, P. and Rambaut, A. (2017) 'Virus genomes reveal factors that spread and sustained the Ebola epidemic', *Nature*, 544/7650: 309-315.
18. Faria, N. R., Quick, J., Claro, I. M., Theze, J., de Jesus, J. G., Giovanetti, M., Kraemer, M. U. G., Hill, S. C., Black, A., da Costa, A. C., Franco, L. C., Silva, S. P., Wu, C. H., Raghwani, J., Cauchemez, S., du Plessis, L., Verotti, M. P., de Oliveira, W. K., Carmo, E. H., Coelho, G. E., Santelli, A., Vinhal, L. C., Henriques, C. M., Simpson, J. T., Loose, M., Andersen, K. G., Grubaugh, N. D., Somasekar, S., Chiu, C. Y., Munoz-Medina, J. E., Gonzalez-Bonilla, C. R., Arias, C. F., Lewis-Ximenez, L. L., Baylis, S. A., Chieppe, A. O., Aguiar, S. F., Fernandes, C. A., Lemos, P. S., Nascimento, B. L. S., Monteiro, H. A. O., Siqueira, I. C., de Queiroz, M. G., de Souza, T. R., Bezerra, J. F., Lemos, M. R., Pereira, G. F., Loudal, D., Moura, L. C., Dhalia, R., Franca, R. F., Magalhaes, T., Marques, E. T., Jr., Jaenisch, T., Wallau, G. L., de Lima, M. C., Nascimento, V., de Cerqueira, E. M., de Lima, M. M., Mascarenhas, D. L., Neto, J. P. M., Levin, A. S., Tozetto-Mendoza, T. R., Fonseca, S. N., Mendes-Correa, M. C., Milagres, F. P., Segurado, A., Holmes, E. C., Rambaut, A., Bedford, T., Nunes, M. R. T., Sabino, E. C., Alcantara, L. C. J., Loman, N.

J. and Pybus, O. G. (2017) 'Establishment and cryptic transmission of Zika virus in Brazil and the Americas', *Nature*, 546/7658: 406-410.

19. Lemey, P., Rambaut, A., Bedford, T., Faria, N., Bielejec, F., Baele, G., Russell, C. A., Smith, D. J., Pybus, O. G., Brockmann, D. and Suchard, M. A. (2014) 'Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2', *PLoS Pathog*, 10/2: e1003932.
20. Faria, N. R., Suchard, M. A., Rambaut, A., Streicker, D. G. and Lemey, P. (2013) 'Simultaneously reconstructing viral cross-species transmission history and identifying the underlying constraints', *Philos Trans R Soc Lond B Biol Sci*, 368/1614: 20120196.
21. Zehender, G., Veo, C., Ebranati, E., Carta, V., Rovida, F., Percivalle, E., Moreno, A., Lelli, D., Calzolari, M., Lavazza, A., Chiapponi, C., Baioni, L., Capelli, G., Ravagnan, S., Da Rold, G., Lavezzo, E., Palu, G., Baldanti, F., Barzon, L. and Galli, M. (2017) 'Reconstructing the recent West Nile virus lineage 2 epidemic in Europe and Italy using discrete and continuous phylogeography', *PLoS One*, 12/7: e0179679.
22. Nokes, D. J., Ngama, M., Bett, A., Abwao, J., Munywoki, P., English, M., Scott, J. A., Cane, P. A. and Medley, G. F. (2009) 'Incidence and severity of respiratory syncytial virus pneumonia in rural Kenyan children identified through hospital surveillance', *Clin Infect Dis*, 49/9: 1341-1349.
23. Scott, J. A. G., Bauni, E., Moisi, J. C., Ojal, J., Gatakaa, H., Nyundo, C., Molyneux, C. S., Kombe, F., Tsofa, B., Marsh, K., Peshu, N. and Williams, T. N. (2012) 'Profile: The Kilifi Health and Demographic Surveillance System (KHDSS)', *International Journal of Epidemiology*, 41/3: 650-657.
24. Hammitt, L. L., Kazungu, S., Welch, S., Bett, A., Onyango, C. O., Gunson, R. N., Scott, J. A. and Nokes, D. J. (2011) 'Added value of an oropharyngeal swab in detection of viruses in children hospitalized with lower respiratory tract infection', *J Clin Microbiol*, 49/6: 2318-2320.
25. Kamau, E., Agoti, C. N., Lewa, C. S., Oketch, J., Owor, B. E., Otieno, G. P., Bett, A., Cane, P. A. and Nokes, D. J. (2017) 'Recent sequence variation in probe binding site affected detection of respiratory syncytial virus group B by real-time RT-PCR', *J Clin Virol*, 88: 21-25.
26. Agoti, C. N., Otieno, J. R., Munywoki, P. K., Mwihuri, A. G., Cane, P. A., Nokes, D. J., Kellam, P. and Cotten, M. (2015) 'Local evolutionary patterns of human respiratory syncytial virus derived from whole-genome sequencing', *J Virol*, 89/7: 3444-3454.
27. Langmead, B. and Salzberg, S. L. (2012) 'Fast gapped-read alignment with Bowtie 2', *Nature Methods*, 9: 357.
28. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, 25/16: 2078-2079.
29. Quinlan, A. R. and Hall, I. M. (2010) 'BEDTools: a flexible suite of utilities for comparing genomic features', *Bioinformatics*, 26/6: 841-842.
30. Katoh, K. and Standley, D. M. (2013) 'MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability', *Molecular Biology and Evolution*, 30/4: 772-780.
31. Stamatakis, A. (2014) 'RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies', *Bioinformatics*, 30/9: 1312-1313.

32. Rambaut, A., Lam, T. T., Max Carvalho, L. and Pybus, O. G. (2016) 'Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen)', *Virus Evolution*, 2/1: vew007-vew007.
33. Posada, D. (2008) 'jModelTest: phylogenetic model averaging', *Mol Biol Evol*, 25/7: 1253-1256.
34. Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J. and Rambaut, A. (2018) 'Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10', *Virus Evolution*, 4/1.
35. Baele, G., Li, W. L., Drummond, A. J., Suchard, M. A. and Lemey, P. (2013) 'Accurate model selection of relaxed molecular clocks in bayesian phylogenetics', *Mol Biol Evol*, 30/2: 239-243.
36. Lemey, P., Rambaut, A., Drummond, A. J. and Suchard, M. A. (2009) 'Bayesian phylogeography finds its roots', *PLoS Comput Biol*, 5/9: e1000520.
37. Faria, N. R., Hodges-Mameletzis, I., Silva, J. C., Rodes, B., Erasmus, S., Paolucci, S., Ruelle, J., Pieniazek, D., Taveira, N., Trevino, A., Goncalves, M. F., Jallow, S., Xu, L., Camacho, R. J., Soriano, V., Goubau, P., de Sousa, J. D., Vandamme, A. M., Suchard, M. A. and Lemey, P. (2012) 'Phylogeographical footprint of colonial history in the global dispersal of human immunodeficiency virus type 2 group A', *J Gen Virol*, 93/Pt 4: 889-899.
38. Bielejec, F., Baele, G., Vrancken, B., Suchard, M. A., Rambaut, A. and Lemey, P. (2016) 'SpreaD3: Interactive Visualization of Spatiotemporal History and Trait Evolutionary Processes', *Mol Biol Evol*, 33/8: 2167-2169.
39. Weaver, S., Shank, S. D., Spielman, S. J., Li, M., Muse, S. V. and Kosakovsky Pond, S. L. (2018) 'Datamonkey 2.0: A Modern Web Application for Characterizing Selective and Other Evolutionary Processes', *Mol Biol Evol*, 35/3: 773-777.
40. Nguyen, L. T., Schmidt, H. A., von Haeseler, A. and Minh, B. Q. (2015) 'IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies', *Mol Biol Evol*, 32/1: 268-274.
41. Parker, J., Rambaut, A. and Pybus, O. G. (2008) 'Correlating viral phenotypes with phylogeny: accounting for phylogenetic uncertainty', *Infect Genet Evol*, 8/3: 239-246.
42. Sagulenko, P., Puller, V. and Neher, R. A. (2018) 'TreeTime: Maximum-likelihood phylodynamic analysis', *Virus Evol*, 4/1: vex042.
43. Botosso, V. F., Zanotto, P. M., Ueda, M., Arruda, E., Gilio, A. E., Vieira, S. E., Stewien, K. E., Peret, T. C., Jamal, L. F., Pardini, M. I., Pinho, J. R., Massad, E., Sant'anna, O. A., Holmes, E. C., Durigon, E. L. and Consortium, V. (2009) 'Positive selection results in frequent reversible amino acid replacements in the G protein gene of human respiratory syncytial virus', *PLoS Pathog*, 5/1: e1000254.
44. Bin, L., Liu, H., Tabor, D. E., Tovchigrechko, A., Qi, Y., Ruzin, A., Esser, M. T. and Jin, H. (2019) 'Emergence of new antigenic epitopes in the glycoproteins of human respiratory syncytial virus collected from a US surveillance study, 2015-17', *Sci Rep*, 9/1: 3898.
45. Duffy, S., Shackelton, L. A. and Holmes, E. C. (2008) 'Rates of evolutionary change in viruses: patterns and determinants', *Nat Rev Genet*, 9/4: 267-276.
46. Aldous, J. L., Pond, S. K., Poon, A., Jain, S., Qin, H., Kahn, J. S., Kitahata, M., Rodriguez, B., Dennis, A. M., Boswell, S. L., Haubrich, R. and Smith, D. M. (2012) 'Characterizing HIV transmission networks across the United States', *Clin Infect Dis*, 55/8: 1135-1143.

47. Smith, D. J., Lapedes, A. S., de Jong, J. C., Bestebroer, T. M., Rimmelzwaan, G. F., Osterhaus, A. D. and Fouchier, R. A. (2004) 'Mapping the antigenic and genetic evolution of influenza virus', *Science*, 305/5682: 371-376.
48. Zhu, Q., Lu, B., McTamney, P., Palaszynski, S., Diallo, S., Ren, K., Ulbrandt, N. D., Kallewaard, N., Wang, W., Fernandes, F., Wong, S., Svabek, C., Moldt, B., Esser, M. T., Jing, H. and Suzich, J. A. (2018) 'Prevalence and Significance of Substitutions in the Fusion Protein of Respiratory Syncytial Virus Resulting in Neutralization Escape From Antibody MEDI8897', *J Infect Dis*, 218/4: 572-580.
49. Domachowske, J. B., Khan, A. A., Esser, M. T., Jensen, K., Takas, T., Villafana, T., Dubovsky, F. and Griffin, M. P. (2018) 'Safety, Tolerability and Pharmacokinetics of MEDI8897, an Extended Half-life Single-dose Respiratory Syncytial Virus Prefusion F-targeting Monoclonal Antibody Administered as a Single Dose to Healthy Preterm Infants', *Pediatr Infect Dis J*, 37/9: 886-892.
50. Zhu, Q., McLellan, J. S., Kallewaard, N. L., Ulbrandt, N. D., Palaszynski, S., Zhang, J., Moldt, B., Khan, A., Svabek, C., McAuliffe, J. M., Wrapp, D., Patel, N. K., Cook, K. E., Richter, B. W. M., Ryan, P. C., Yuan, A. Q. and Suzich, J. A. (2017) 'A highly potent extended half-life antibody as a potential RSV vaccine surrogate for all infants', *Sci Transl Med*, 9/388.
51. Kosakovsky Pond, S. L. and Frost, S. D. (2005) 'Not so different after all: a comparison of methods for detecting amino acid sites under selection', *Mol Biol Evol*, 22/5: 1208-1222.
52. Chen, X., Xu, B., Guo, J., Li, C., An, S., Zhou, Y., Chen, A., Deng, L., Fu, Z., Zhu, Y., Liu, C., Xu, L., Wang, W., Shen, K. and Xie, Z. (2018) 'Genetic variations in the fusion protein of respiratory syncytial virus isolated from children hospitalized with community-acquired pneumonia in China', *Sci Rep*, 8/1: 4491.

Figure 1 (A)

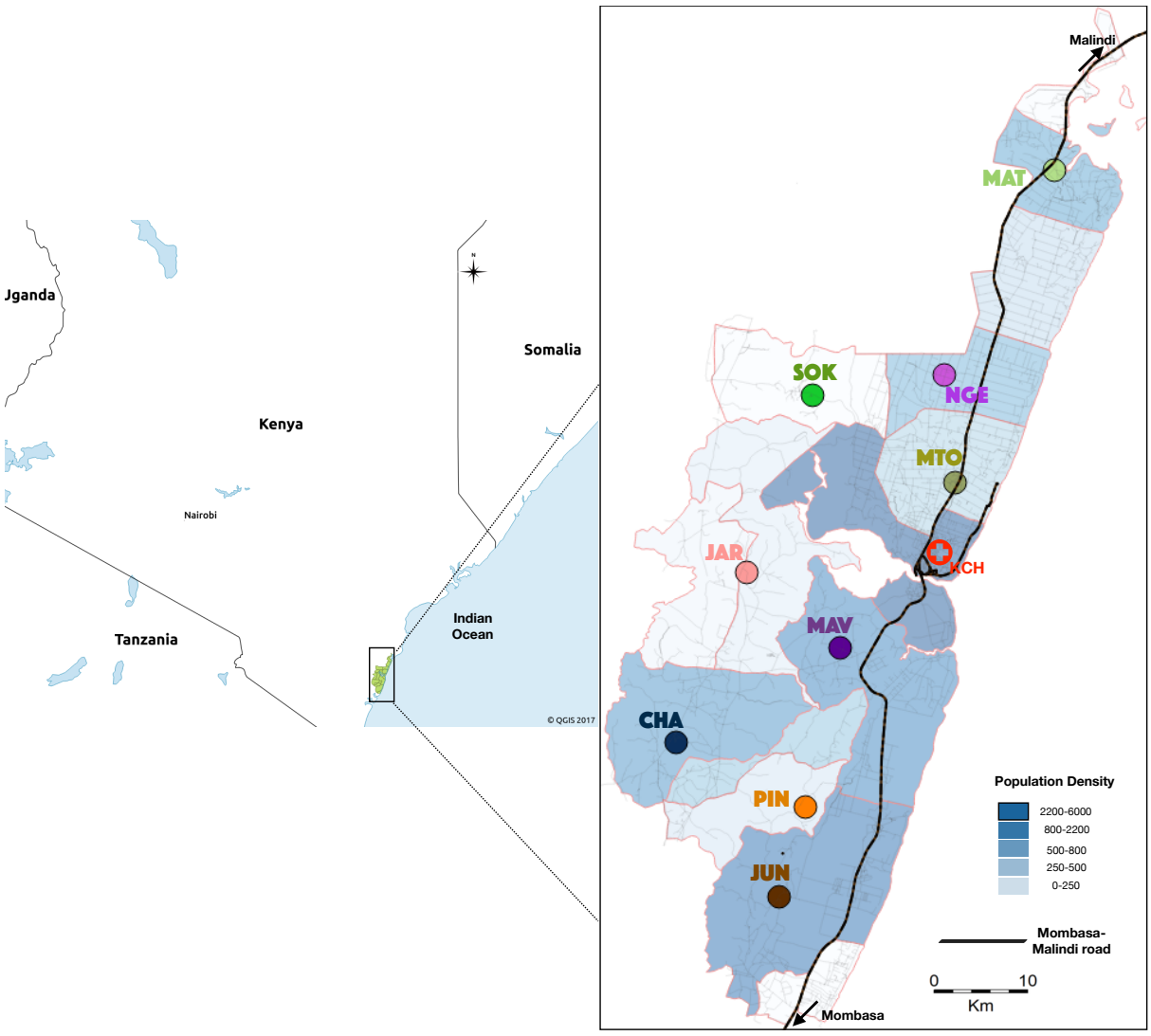

Jganda
Kenya
Nairobi
Tanzania
Somalia
Indian
Ocean
© QGIS 2017
Malindi
MAT
SOK
NGE
MTO
JAR
KCH
MAV
CHA
PIN
JUN
Mombasa
Population Density
2200-6000
800-2200
500-800
250-500
0-250
Mombasa-
Malindi road
0
10
Km

Figure 1 (B)

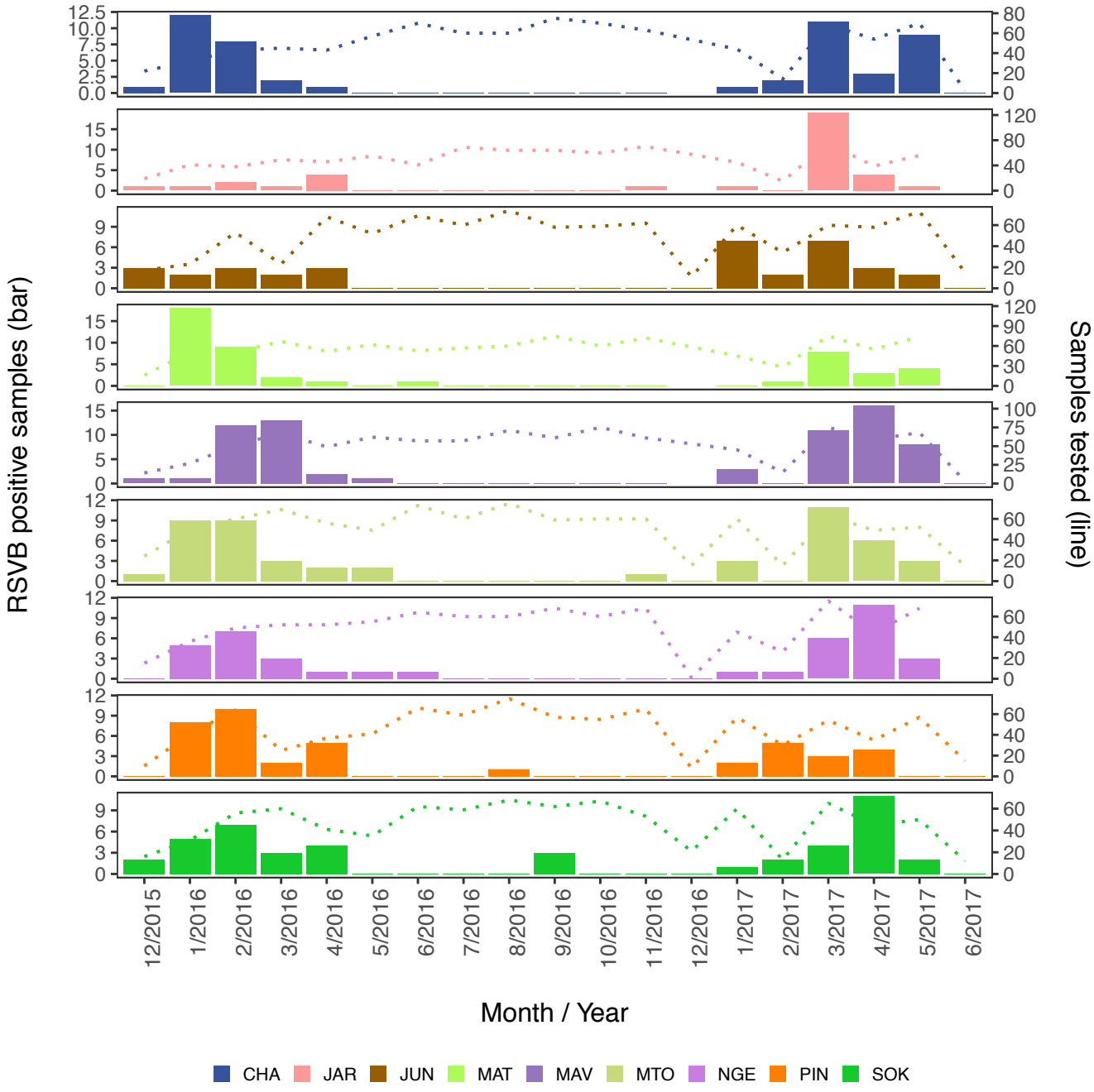

RSVB positive samples (bar)
Samples tested (line)
Month / Year
12/2015
1/2016
2/2016
3/2016
4/2016
5/2016
6/2016
7/2016
8/2016
9/2016
10/2016
11/2016
12/2016
1/2017
2/2017
3/2017
4/2017
5/2017
6/2017
CHA
JAR
JUN
MAT
MAV
MTO
NGE
PIN
SOK

Figure 2 (A)

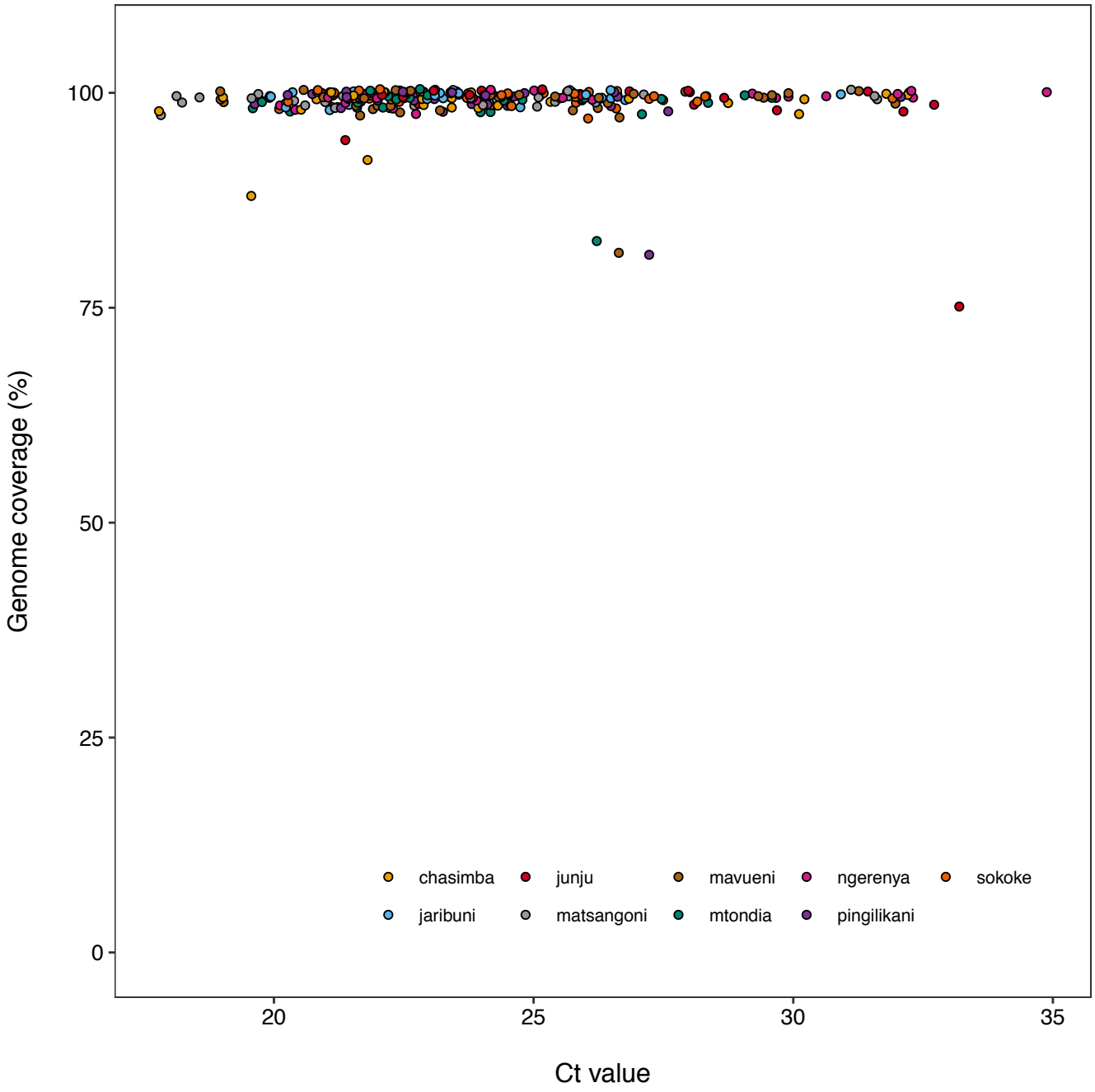

Genome coverage (%)
100
75
50
25
0
20
25
30
35
Ct value
chasimba
junju
mavueni
ngerenya
sokoke
jaribuni
matsangoni
mtondia
pingilikani

Figure 2 (B)

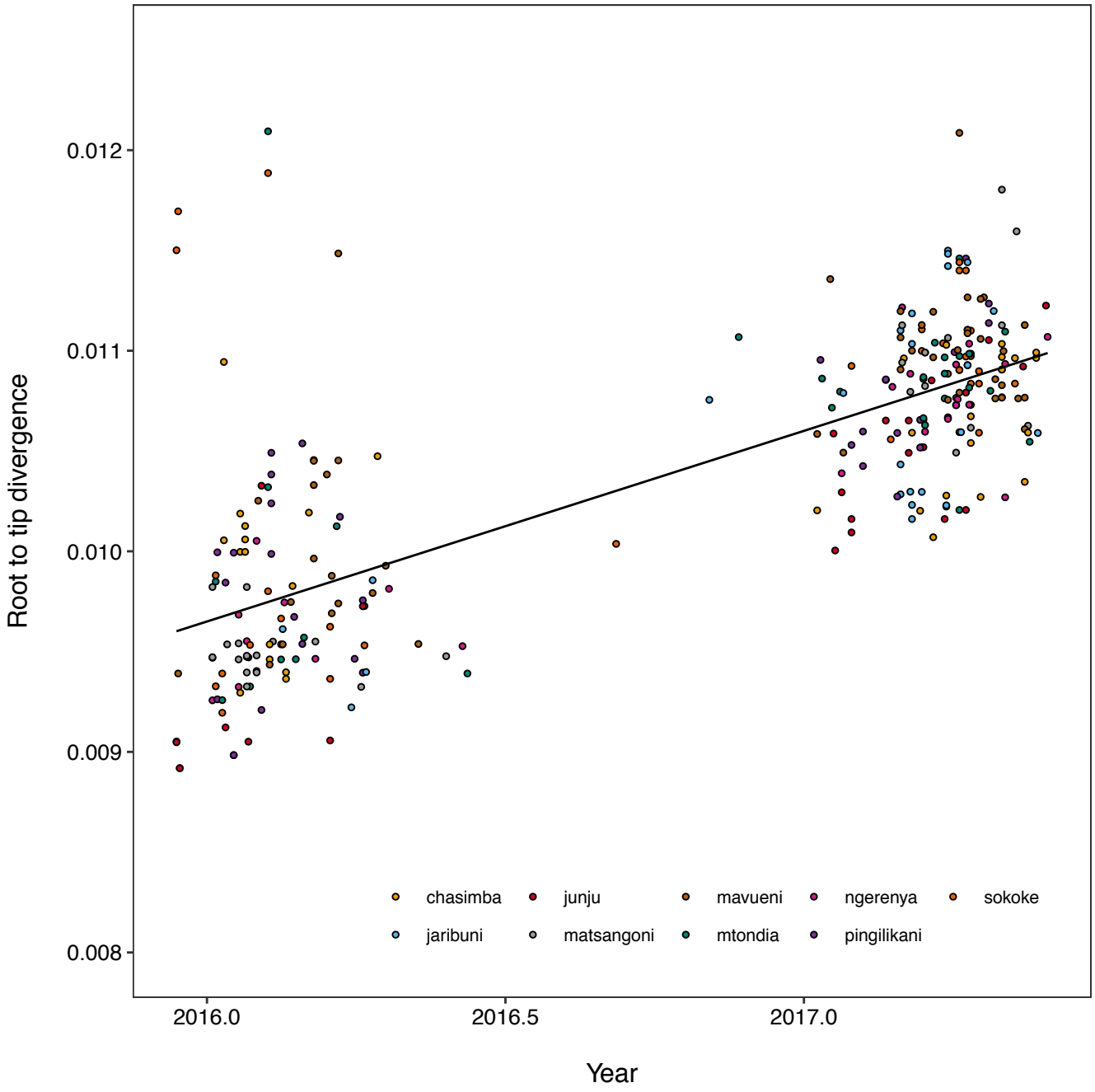

Root to tip divergence
0.012
0.011
0.010
0.009
0.008
2016.0
2016.5
2017.0
Year
chasimba
junju
mavueni
ngerenya
sokoke
jaribuni
matsangoni
mtondia
pingilikani

Figure 2(C)

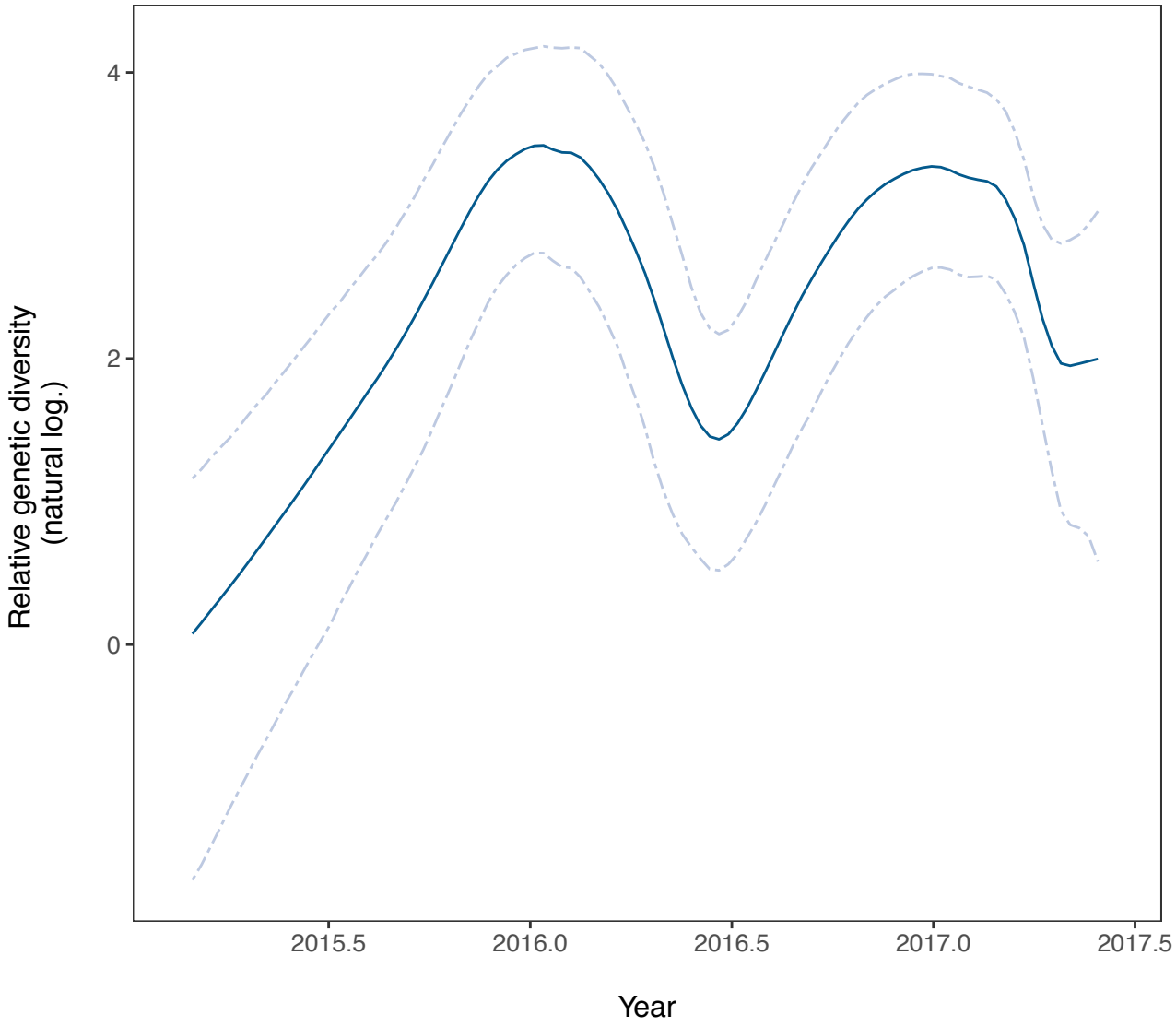

Relative genetic diversity
(natural log.)
4
2
0
2015.5
2016.0
2016.5
2017.0
2017.5
Year

Figure 2 (D)

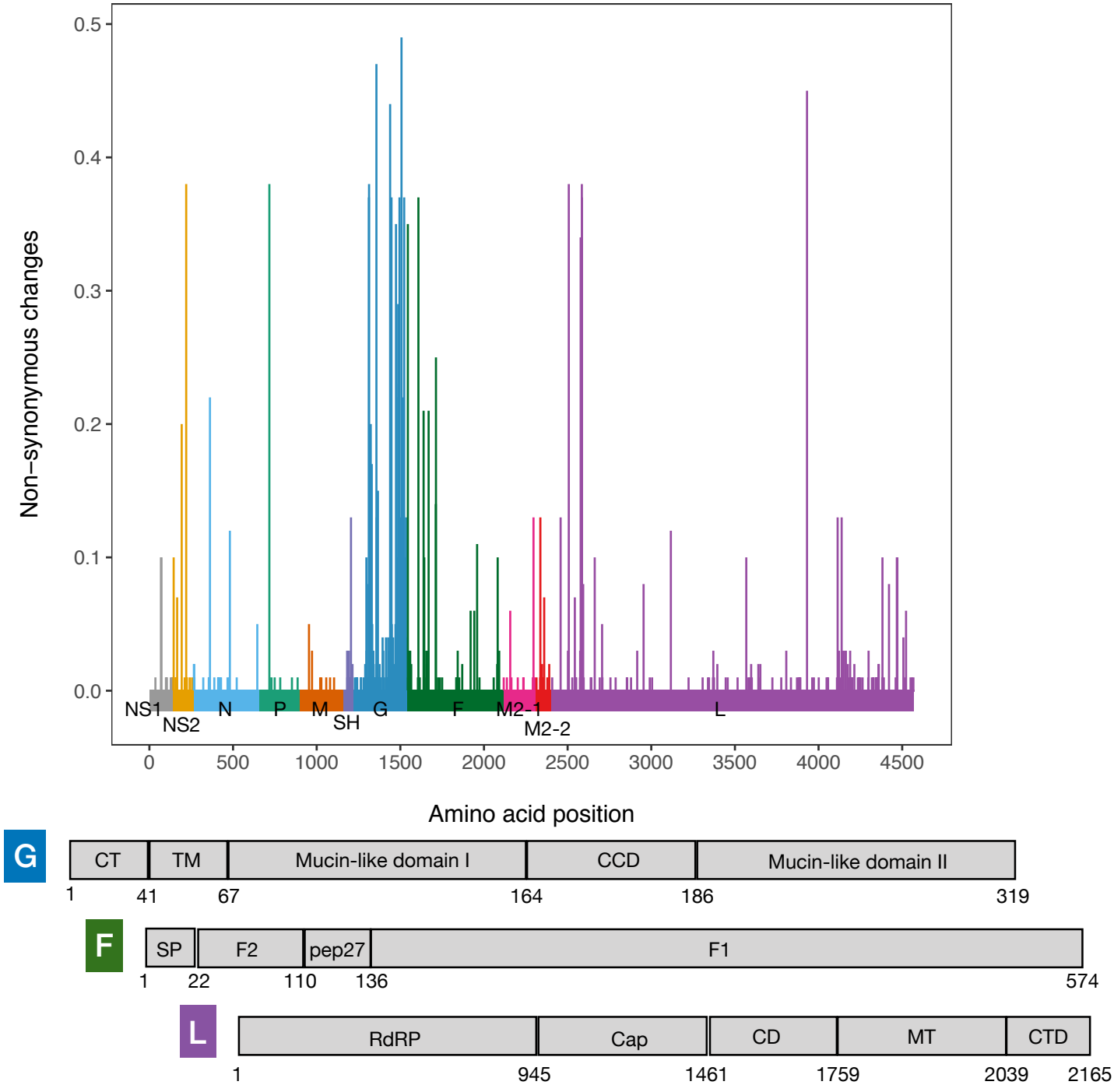

Non−synonymous changes
0.5
0.4
0.3
0.2
0.1
0.0
NS1
NS2
N
P
M
SH
G
F
M2-1
M2-2
L
0
500
1000
1500
2000
2500
3000
3500
4000
4500
Amino acid position
G
CT
TM
Mucin-like domain I
CCD
Mucin-like domain II
1
41
67
164
186
319
F
SP
F2
pep27
F1
1
22
110
136
574
L
RdRP
Cap
CD
MT
CTD
1
945
1461
1759
2039
2165

Figure 3(A)

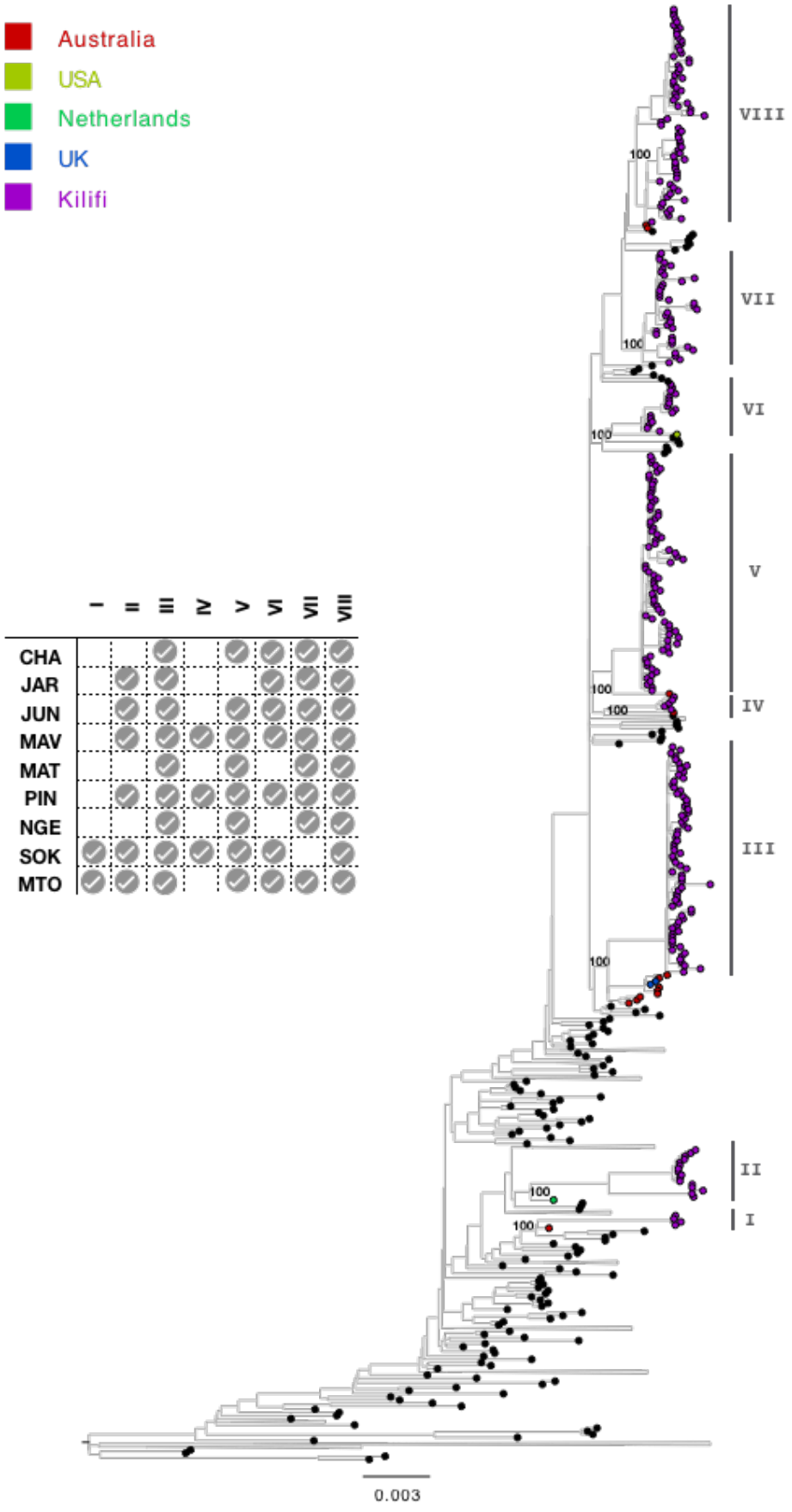

Australia
USA
Netherlands
UK
Kilifi
I
II
III
IV
V
VI
VII
VIII
CHA
JAR
JUN
MAV
MAT
PIN
NGE
SOK
MTO
100
0.003

Figure 3(B)

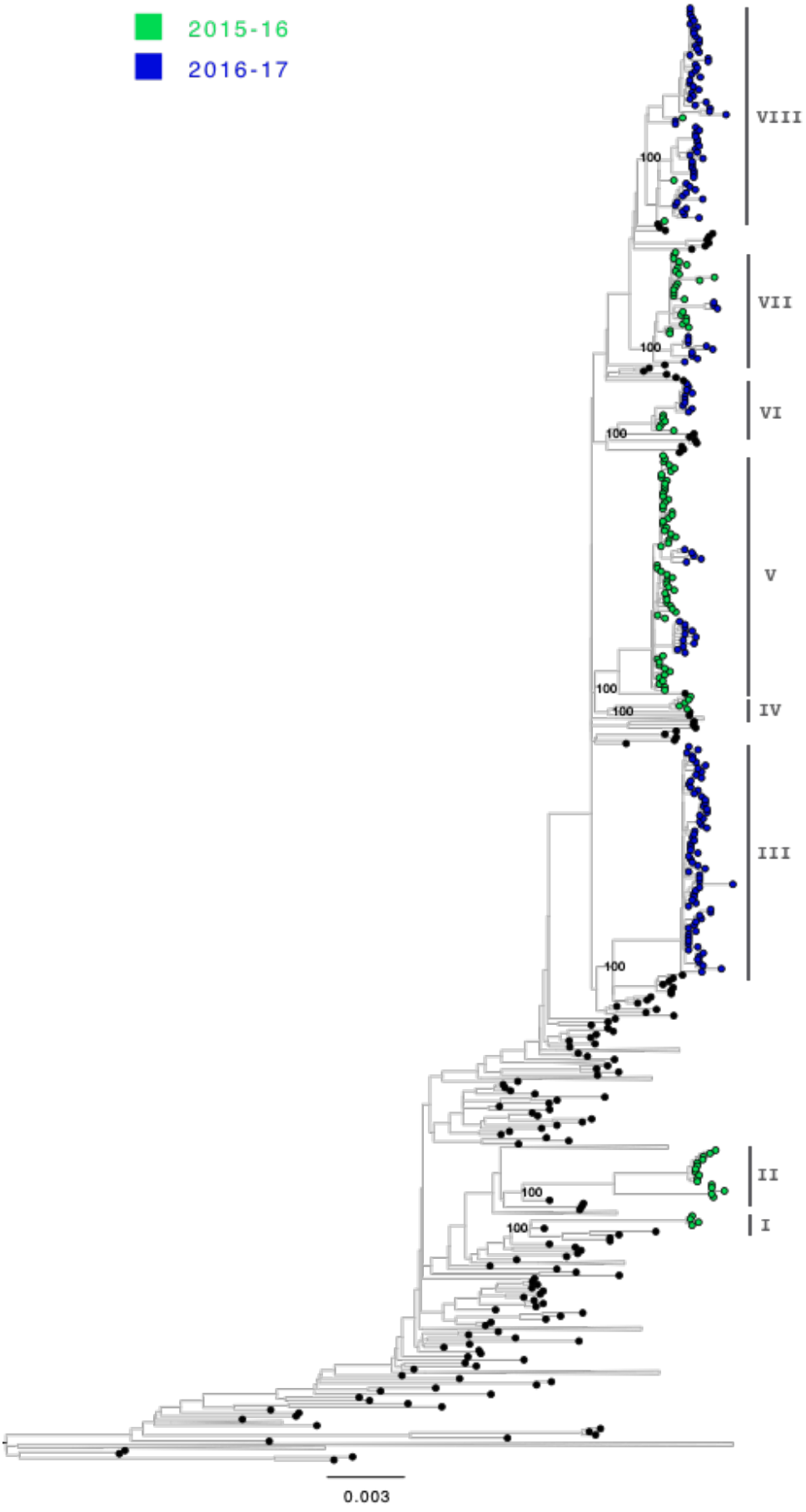

2015-16
2016-17
VIII
VII
VI
V
IV
III
II
I
100
100
100
100
100
100
100
0.003

Figure 4 (A)

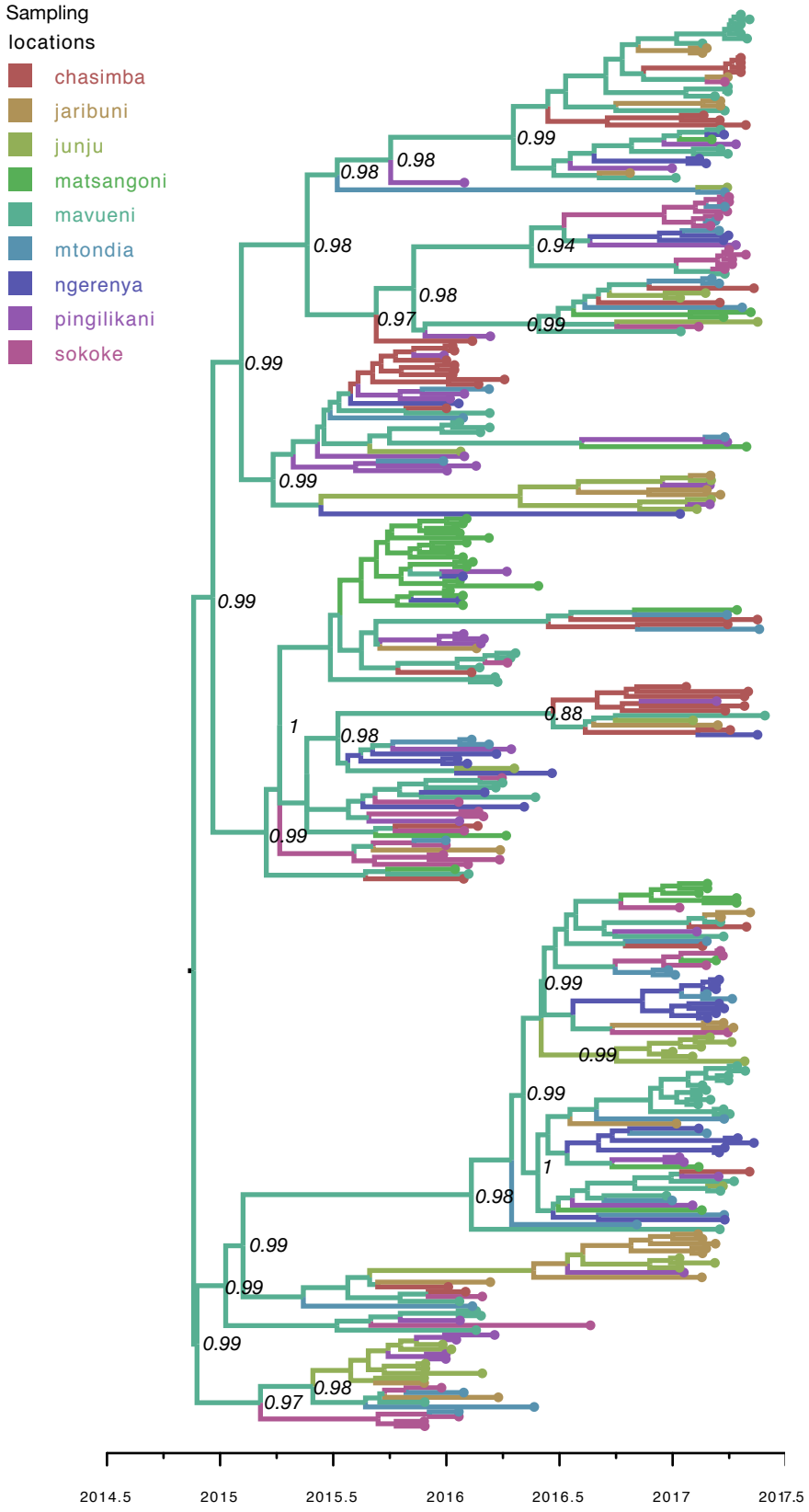

Sampling locations
chasimba
jaribuni
junju
matsangoni
mavueni
mtondia
ngerenya
pingilikani
sokoke
0.99
0.98
0.98
0.98
0.94
0.98
0.97
0.99
0.99
0.99
0.99
1
0.98
0.88
0.99
0.99
0.99
1
0.98
0.99
0.99
0.99
0.98
0.97
2014.5
2015
2015.5
2016
2016.5
2017
2017.5

Figure 4(B)

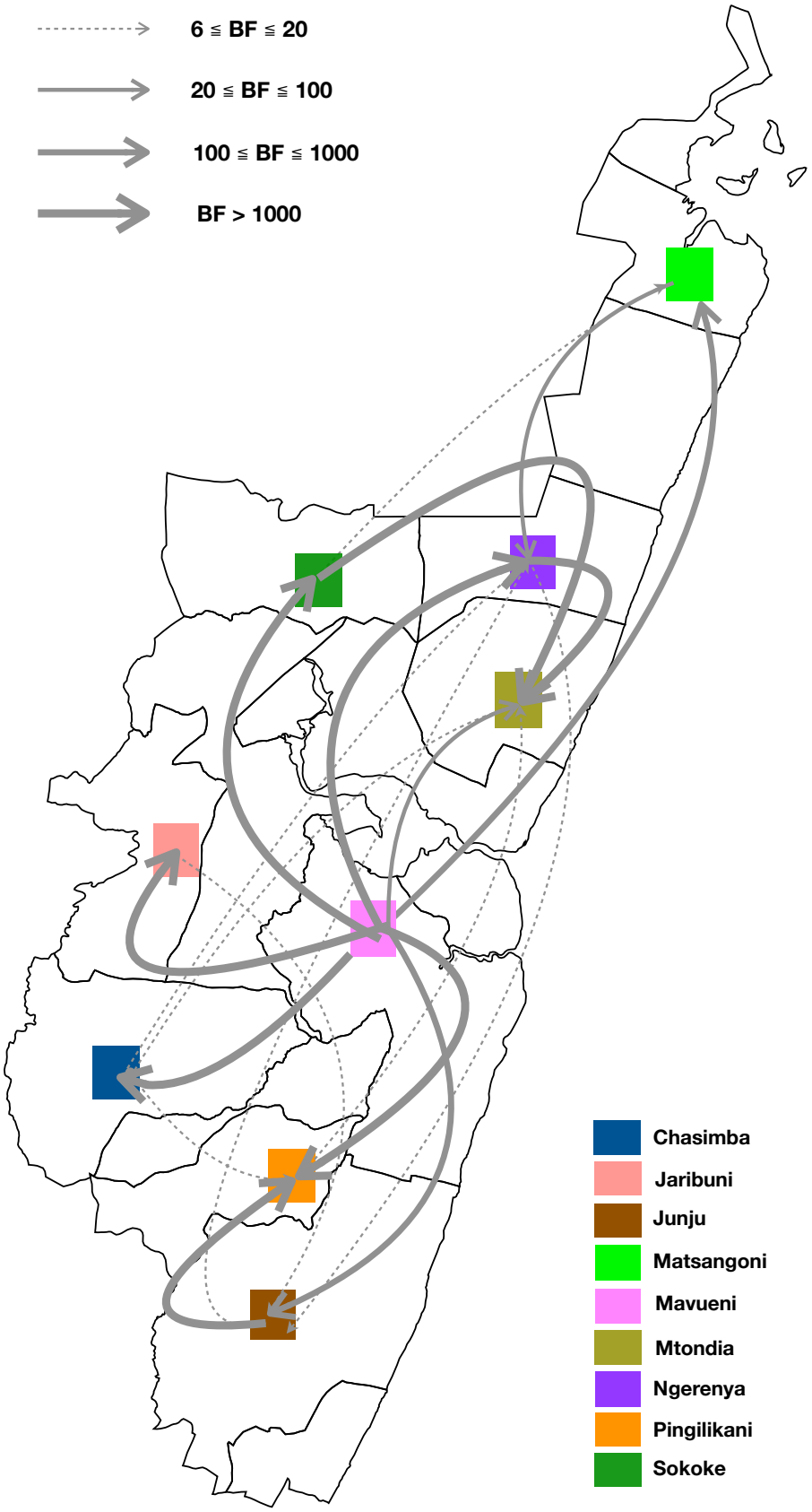

6 ≦ BF ≦ 20
20 ≦ BF ≦ 100
100 ≦ BF ≦ 1000
BF > 1000
Chasimba
Jaribuni
Junju
Matsangoni
Mavueni
Mtondia
Ngerenya
Pingilikani
Sokoke

Figure 5(A)

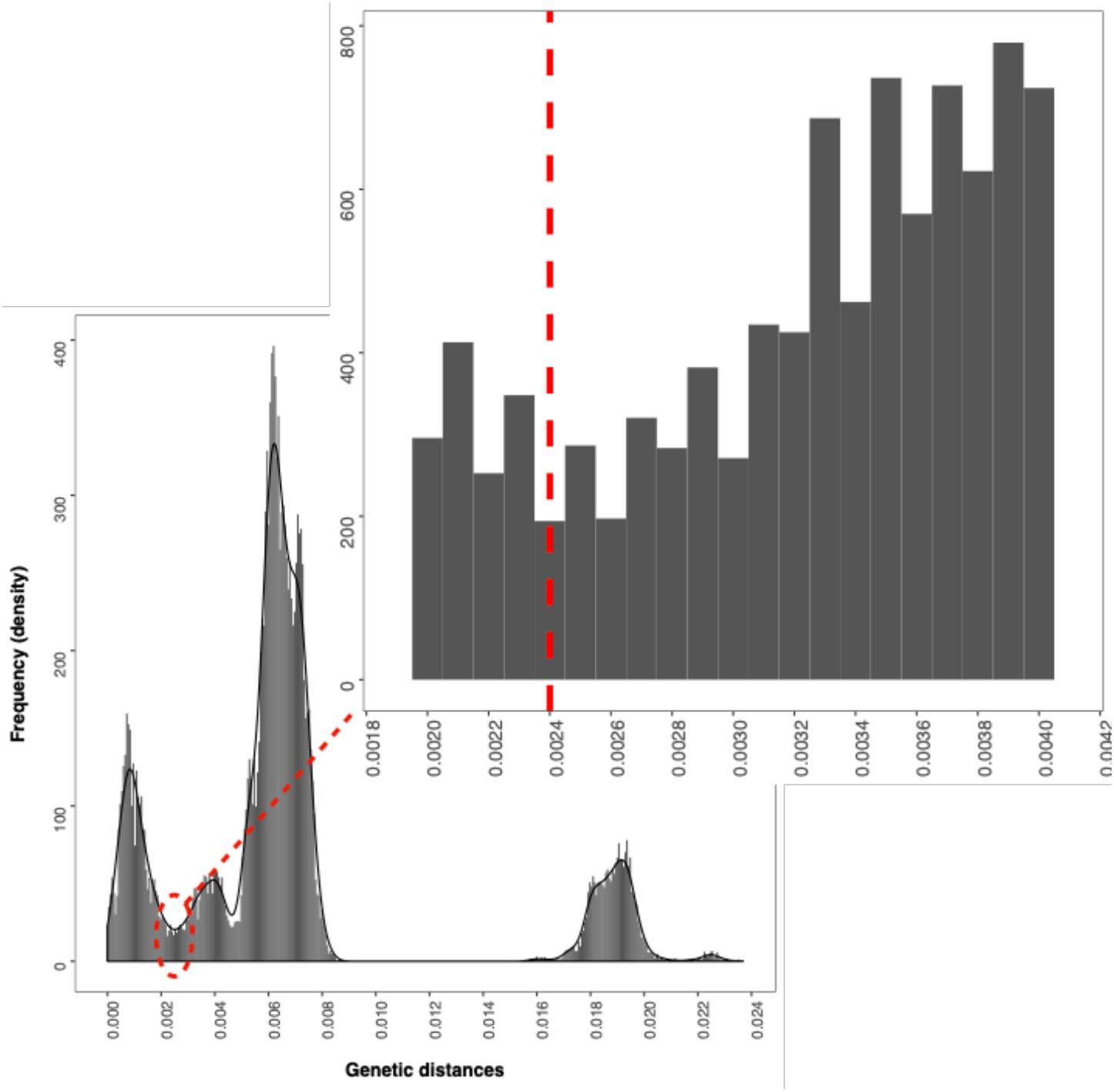

Frequency (density)
0
100
200
300
400
0.000
0.002
0.004
0.006
0.008
0.010
0.012
0.014
0.016
0.018
0.020
0.022
0.024
Genetic distances
0
200
400
600
800
0.0018
0.0020
0.0022
0.0024
0.0026
0.0028
0.0030
0.0032
0.0034
0.0036
0.0038
0.0040
0.0042

Figure 5(B)

2015/16		2016/17	
1	0.0013362	10	0.0018424
2	0.0021504	11	0.0003305
3	0.0002829	12	1E-06
4	0.0022764	13	0.0019125
5	0.0012214	14	0.0018402
6	0.0012025	15	0.0021738
7	0.0009352	16	0.000208
8	0.0006647	17	0.0010525
9	0.0005339	18	0.0015105
		19	0.0017768
		20	0.0019813

Figure 6(A)

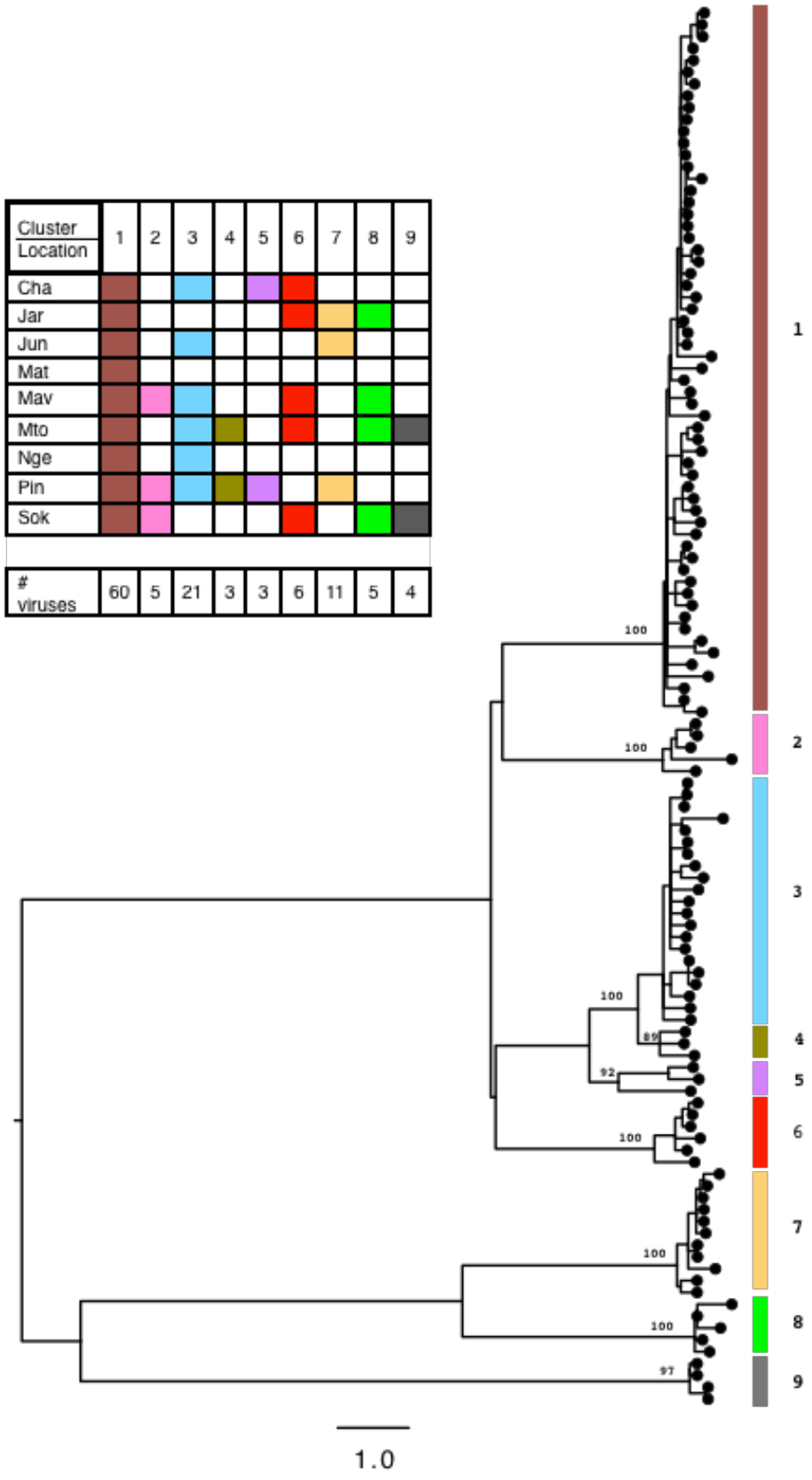
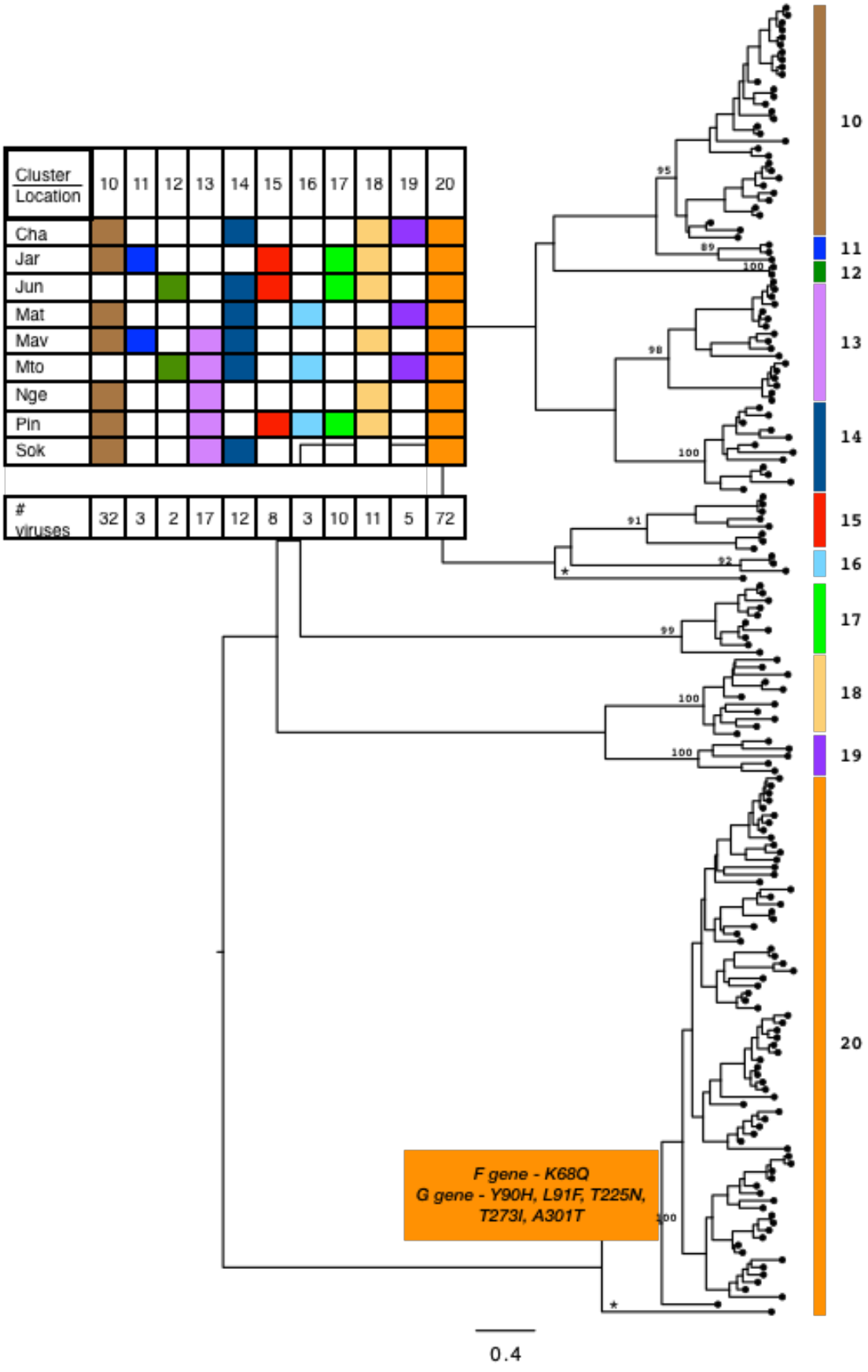

Cluster
Location
1 2 3 4 5 6 7 8 9
Cha
Jar
Jun
Mat
Mav
Mto
Nge
Pin
Sok
viruses
60 5 21 3 3 6 11 5 4
100
100
100
89
92
100
100
100
97
1
2
3
4
5
6
7
8
9
1.0

Figure 6(B)


Cluster / Location
10
11
12
13
14
15
16
17
18
19
20
Cha
Jar
Jun
Mat
Mav
Mto
Nge
Pin
Sok
viruses
32
3
2
17
12
8
3
10
11
5
72
95
89
100
98
100
91
92
99
100
100
100
F gene - K68Q
G gene - Y90H, L91F, T225N, T273I, A301T
0.4