Elsevier Editorial System(tm) for Infection, Genetics and Evolution

Manuscript Draft

Manuscript Number: MEEGID-D-16-00266R1

Title: MARBURG VIRUS DISTRIBUTION IN AFRICA: AN EVOLUTIONARY APPROACH

Article Type: Research paper

Keywords: Marburg virus, Ravn virus, geographic dispersion, phylodynamic

Corresponding Author: Prof. Gianguglielmo Zehender, PhD

Corresponding Author's Institution: University of Milan

First Author: Gianguglielmo Zehender, PhD

Order of Authors: Gianguglielmo Zehender, PhD; Chiara Sorrentino; Carla Veo; Lisa Fiaschi; Sonia Gioffrè; Erika Ebranati; Elisabetta Tanzi; Massimo Ciccozzi; Alessia Lai; Massimo Galli

Abstract: The aim of this study was to investigate the origin and geographical dispersion of Marburg virus, the first member of the Filoviridae family to be discovered. Seventy-three complete genome sequences of Marburg virus isolated from animals and humans were retrieved from public databases and analysed using a Bayesian phylogeographical framework. The phylogenetic tree of the Marburg virus data set showed two significant evolutionary lineages: Ravn virus (RAVV) and Marburg virus (MARV). MARV divided into two main clades; clade A included isolates from Uganda (five from the European epidemic in 1967), Kenya (1980) and Angola (from the epidemic of 2004-2005); clade B included most of the isolates obtained during the 1999-2000 epidemic in the Democratic Republic of the Congo (DRC) and a group of Ugandan isolates obtained in 2007-2009. The estimated mean evolutionary rate of the whole genome was 3.3x10-4 substitutions/site/year (credibility interval 2.0-4.8). The MARV strain had a mean root time of the most recent common ancestor of 177.9 years ago (YA) (95% highest posterior density 87-284), thus indicating that it probably originated in the mid-XIX century, whereas the RAVV strain had a later origin dating back to a mean 33.8 YA. The most probable location of the MARV ancestor was Uganda (state posterior probability, spp = 0.41), whereas that of the RAVV ancestor was Kenya (spp = 0.71). There were significant migration rates from Uganda to the DRC (Bayes Factor, BF=42.0) and in the opposite direction (BF=5.7). Our data suggest that Uganda may have been the cradle of Marburg virus in Africa.

Dear Editor:

We would like to thank you for the competence of your reviewers and for your helpful suggestions.

Following your encouragement, we have reviewed the manuscript on the basis of the referees' comments and done our outmost to satisfy all their requests.

In particular, as you suggested we added more complete information about the methodology employed and about the strains included in the analysis.

Moreover the manuscript was revised by a professional English mother-tongue.

We hope that you will now find the paper suitable for publication in Infection Genetic and Evolution.

Yours sincerely,

Gianguglielmo Zehender

***Reply to Reviewers**

Reviewer#1

We would like to thank the Reviewer for his/her helpful suggestions to improve our manuscript.

1. The manuscript has been revised by a professional English mother-tongue.
2. According to the request of the referee, supplementary data describing in more detail the methodology employed has been added to the manuscript.

Reviewer#2

We would like to thank the Reviewer for his/her helpful suggestions to improve our manuscript.

1. All the parameter of the model used to study the selection pressure have been estimated by Datamonkey. This information has now been added to the manuscript.
2. We agree with the referee that the discussion section is a little bit long and sometime redundant. Now we have shortened at our best the discussion section, by erasing the repetitions.
3. The figure legends now report the meaning of the scale bars.
4. More detailed information about the sequences used (acc.numbers, sampling dates and locations) have been now reported in supplementary table 1.

HIGHLIGHTS

• Evolutionary rate estimation of $3.3x10^{-4}$ substitutions/site/year for the whole Marburg virus genome

• Bayesian analysis suggests that MARV strain originated about 177.9 years ago most probably in Uganda

• Bayesian analysis suggests that RAVV strain originated about 33.8 years ago in an area between Uganda and Kenya

• Our data highlight a central role of Uganda as the cradle of Marburg virus in Africa

DISTRIBUTION OF MARBURG VIRUS IN AFRICA: AN EVOLUTIONARY APPROACH

Gianguglielmo Zehender[1#], Chiara Sorrentino[1], Carla Veo[1], Lisa Fiaschi[1], Sonia Gioffrè[1], Erika Ebranati[1], Elisabetta Tanzi[2], Massimo Ciccozzi[3], Alessia Lai[1], Massimo Galli[1]

[1]Department of Biomedical and Clinical Sciences "Luigi Sacco", University of Milan, Milan, Italy.

[2]Department of Biomedical Sciences for Health, University of Milan, Milan, Italy.

[3]Department of Infectious Parasitic and Immunomediated Diseases, National Institute of Health, Rome, Italy.

[#]Corresponding author:

Prof. Gianguglielmo Zehender,

Luigi Sacco Department of Biomedical and Clinical Sciences,

Infectious Diseases and Immunopathology Section,

University of Milan,

Via G.B. Grassi 74,

20157 Milano,

Italy.

Phone (+39) 0250319770, Fax: (+39) 0250319768; e-mail: gianguglielmo.zehender@unimi.it

ABSTRACT

The aim of this study was to investigate the origin and geographical dispersion of Marburg virus, the first member of the *Filoviridae* family to be discovered. Seventy-three complete genome sequences of Marburg virus isolated from animals and humans were retrieved from public databases and analysed using a Bayesian phylogeographical framework. The phylogenetic tree of the Marburg virus data set showed two significant evolutionary lineages: Ravn virus (RAVV) and Marburg virus (MARV). MARV divided into two main clades; clade A included isolates from Uganda (five from the European epidemic in 1967), Kenya (1980) and Angola (from the epidemic of 2004-2005); clade B included most of the isolates obtained during the 1999-2000 epidemic in the Democratic Republic of the Congo (DRC) and a group of Ugandan isolates obtained in 2007-2009. The estimated mean evolutionary rate of the whole genome was $3.3x10^{-4}$ substitutions/site/year (credibility interval 2.0-4.8). The MARV strain had a mean root time of the most recent common ancestor of 177.9 years ago (YA) (95% highest posterior density 87-284), thus indicating that it probably originated in the mid-XIX century, whereas the RAVV strain had a later origin dating back to a mean 33.8 YA. The most probable location of the MARV ancestor was Uganda (state posterior probability, spp = 0.41), whereas that of the RAVV ancestor was Kenya (spp = 0.71). There were significant migration rates from Uganda to the DRC (Bayes Factor, BF=42.0) and in the opposite direction (BF=5.7). Our data suggest that Uganda may have been the cradle of Marburg virus in Africa.

HIGHLIGHTS

• The estimated evolutionary rate of the whole Marburg virus genome is $3.3x10^{-4}$ substitutions/site/year

• Bayesian analysis suggests that the MARV strain most probably originated in Uganda about 177.9 years ago

• It also suggests that the RAVV strain originated in an area between Uganda and Kenya about 33.8 years ago

• Our data indicates Uganda as the cradle of the Marburg virus in Africa

KEY WORDS

Marburg virus; Ravn virus; geographic dispersion; phylodynamics

1. INTRODUCTION

Marburg virus was the first discovered member of the *Filoviridae* (Siegert et al., 1968), a family of enveloped, non-segmented, negative-stranded RNA viruses with a characteristic filamentous structure (Rougeron et al., 2015). The *Marburgvirus* genus is one of three known *Filoviridae* genera, and accounts for a single viral species, *Marburg marburgvirus*, that includes the two distinct and quite divergent strains of Marburg virus (MARV) and Ravn virus (RAVV) (Kuhn et al., 2010). These tow strains are the causative agents of Marburg virus disease (MVD), a highly lethal condition that clinically resembles Ebola virus disease (EVD) (Brauburger et al., 2012), and has so far been reported in Uganda, the north-eastern area of the Democratic Republic of the Congo (DRC), Kenya, Zimbabwe and northern Angola. These areas are separated by distances of sometimes even thousands of kilometres, and lie in ecologically diverse regions that include rain forests and arid woodlands (Brauburger et al., 2012; Peterson et al., 2006). Unlike EVD, MVD has been more frequently observed as sporadic cases or small epidemics (Adjemian et al., 2011; Conrad et al., 1978; Johnson et al., 1996; Smith et al., 1982), with the significant exceptions of two large-scale epidemics: one in Durba, the DRC, in 1998-99 (Bausch et al., 2006; Kuhn, 2008), and the other in Uige, Angola, in 2004-2005 (Feldmann, 2006; Towner et al., 2006). Moreover, at least eight of the 13 MVD cases published so far may have been epidemiologically related to human entry into caves or mines (Adjemian et al., 2011; Amman et al., 2014; Brauburger et al., 2012).

The isolation of MARV from specimens of the common Egyptian fruit bat (*Rousettus aegyptiacus*) captured in Gabon in 2007 (Towner et al., 2007), and healthy infected *R. aegyptiacus* bats caught in the Kitaka mine in Uganda (Towner et al., 2009) and the Goroumbwa mine near Durba in the DRC, makes it likely that this species is a natural reservoir of the disease. Moreover, other fruit and insectivorous bat species have also been found to be positive for MARV DNA or antibodies in different parts of Africa (Pourrut et al., 2009; Swanepoel et al., 2007). Interestingly, both Marburg virus lineages can co-circulate in humans and bats in the same area (Towner et al., 2009), and

closely similar isolates can be found in widely separated locations (Towner et al., 2009). The origin, host range and path of dispersion of Marburg virus have not yet been defined, and the same is true of the ecological factors that influence their spread and transmission to humans.

The aim of this study was to investigate the origin and mechanisms of dispersion of Marburg virus within Africa using a Bayesian phylogeographical approach.

2. MATERIALS AND METHODS

2.1. Sequence data sets

This analysis is based on a total of 73 complete genome sequences of Marburg virus isolated in various African countries and retrieved from public databases (GenBank at http://www.ncbi.nlm.nih. gov/genbank/).

The sequences were selected using the following inclusion criteria: i) they had to have previously published in peer-reviewed journals; ii) their non-recombinant subtype assignment had to be certain; iii) the city/state of origin and year of sampling had to be known and clearly established in the original publication.

The sampling dates ranged from 1967 to 2012, and the sampling locations were Uganda (UG, n=22), Zimbabwe (ZW, n=1), Kenya (KE, n=11), Angola (AO, n=9), and the Democratic Republic of the Congo, DRC (CD, n=30). More information regarding the used dataset are shown in Supplementary Table 1. The 1967 Marburg epidemic sequences were considered Ugandan.

All of the sequences were aligned using ClustalX software (Thompson et al., 1994) and then manually edited using Bioedit software v. 7.2.5 (freely available at http://www.mbio.ncsu.edu/bioedit/bioedit.html). Only the coding regions of the whole genome were considered.

2.2. Likelihood mapping

In order to obtain an overall impression of the phylogenetic signals in the analysed sequences, we made a likelihood-mapping analysis of 10,000 random quartets generated using TreePuzzle (Schmidt et al., 2002). A likelihood map consists of an equilateral triangle: each dot within the triangle represents the likelihoods of the three possible unrooted trees for a set of four sequences (a quartet) randomly selected from the data set. The dots close to the corners or at the sides respectively represent tree-like (fully resolved phylogenies in which one tree is clearly better than the others) or network-like phylogenetic signals (three regions in which it is not possible to decide

between two topologies); the central area of the map represents a star-like signal (the region in which the star tree is the optimal tree).

2.3. Root-to-tip regression analysis

In order to investigate the temporal structure and "clock-likeness" of the Marburg virus data set, we made a regression analysis of the root-to-tip genetic distance against sampling years using Path-O-Gen (Rambaut, 2000), and a maximum-likelihood tree without a molecular clock assumption (Guindon and Gascuel, 2003).

2.4. Phylogenetic reconstruction

The evolutionary model that best fitted the data was selected using an information criterion implemented in JmodelTest v. 2.1.7 (Posada, 2008) (freely available at http://darwin.uvigo.es/software/jmodeltest.html), which selected the general time reversible + gamma distribution (GTR+Γ) model of nucleotide substitution.

Phylogeny was initially analysed using the selected model and MrBayes v. 3.7 program (Huelsenbeck and Ronquist, 2001). A Markov Chain Monte Carlo (MCMC) search was made for $10x10^6$ generations until reaching convergence using tree sampling every 100^{th} generation with a burn-in fraction of 50%. Statistical support for specific clades was obtained by calculating the posterior probability (pp) of each monophyletic clade, and a posterior consensus tree was generated after a 50% burn-in.

The evolutionary rates were estimated using a Bayesian MCMC method implemented in BEAST 1.8.0 (Drummond et al., 2012) under strict and relaxed clock conditions, and an uncorrelated log normal rate distribution model. Four demographic models of population growth were compared as coalescent priors: constant size, exponential growth, logistic growth, and a piecewise-constant Bayesian skyline plot (BSP) (Drummond et al., 2005).

The chains were run for 100 million generations until reaching convergence, and sampled every 10,000 steps. Convergence was assessed on the basis of an effective sampling size (ESS) of >200

after a 10% burn-in using Tracer software v. 1.5 (http://tree.bio.ed.ac.uk/software/tracer/). The analysis was made after partitioning the entire coding sequences into codon positions. Uncertainty in the estimates was indicated by the 95% highest posterior density (95% HPD) intervals, and the best fitting models were selected using a Bayes factor (BF) with marginal likelihoods implemented in BEAST. In accordance with (Kass and Raftery, 1995), the strength of the evidence against H_0 was evaluated as follows: 2lnBF <2 no evidence; 2–6 weak evidence; 6–10 strong evidence, and >10 very strong evidence. A negative 2lnBF indicates evidence in favour of H_0. Only values of $\geq$6 were considered significant. The trees were summarised as a maximum clade credibility (MCC) tree (the tree with the largest product of posterior clade probabilities) after a 10% burn-in using the Tree Annotator program included in the BEAST package. The time of the most recent common ancestor (tMRCA) estimates were expressed as mean values and 95% HPD years before the most recent sampling dates, corresponding to 2012 for the analysed dataset. The final trees were visualised and manipulated using FigTree v.1.4.0 (available at http://tree.bio.ed.ac.uk/software/figtree/).

2.5. Bayesian phylogeography

The geographical analysis was made using the continuous time Markov chain (CTMC) process over discrete sampling locations implemented in BEAST, and the Bayesian Stochastic Search Variable Selection (BSSVS) model, which allows diffusion rates to be zero with a positive prior probability (Lemey et al., 2009). Comparison of the posterior and prior probabilities that the individual rates were zero provided a formal BF for testing the significance of the linkage between locations. Rates with a BF of >3 were considered well supported and formed the migration pathway.

The MCC tree was selected from the posterior tree distribution after a 10% burn-in using the TreeAnnotator program, version 1.8.0, and the final tree was visualised using FigTree, version 1.4. The significant migration rates were analysed and visualised using SPREAD (Bielejec et al., 2011), which is available at http://www.kuleuven.be/aidslab/phylogeography/SPREAD.html.

The 73 isolates were assigned to five distinct geographical groups corresponding to the sampled locations (Uganda, Zimbabwe, Kenya, Angola, and the DRC). In order to provide a spatial projection, the migration routes indicated by the tree were visualised using Google Earth (http://earth.google.com).

2.6. Viral gene flow analysis

The MacClade program, version 4 (Sinauer Associates, Sunderland, MA) was used to test viral gene out/in using a modified version of the Slatkin and Maddison test (Slatkin and Maddison, 1989). A one-character data matrix was obtained from the original data set by assigning a one-letter code indicating geographical origin to each taxon in the tree. The putative origin of each ancestral sequence (i.e. internal node) in the tree was then inferred by finding the most parsimonious reconstruction (MPR) of the ancestral character. No ACCTRAN or DELTRAN resolving options were used because there were no uncertainties in the tree. The final tree length (i.e. the number of observed viral gene flow events in the genealogy) can be computed and compared with the tree-length distribution of 10,000 trees obtained by random joining-splitting (null distribution) whose observed genealogies are significantly shorter than random trees, thus indicating the presence of subdivided populations with a restricted gene flow. Viral gene flow (migration) was traced using MacClade state changes and stasis software, which counts the number of changes in a tree for each pair-wise character state. Gene flow was also calculated for a null distribution in order to assess whether the gene flow events observed in the actual tree were significantly higher (>95%) or lower (<95%) than the values in the null distribution at a p level of 0.05.

The isolates were divided into five distinct groups based on their sampling locations (Uganda, Zimbabwe, Kenya, Angola and the DRC).

2.7. Selection pressure analysis

The non-synonymous/synonymous rate ratio dN/dS (ω) was estimated using the maximum likelihood (ML) approach under a global single-ratio model implemented in the HyPhy program

(Kosakovsky Pond and Frost, 2005). In particular, the global model (which assumes a single selective pressure for all branches) was compared with the local model (which allows selective pressure to change along every branch) using the likelihood ratio test (LRT). Site-specific positive or negative selection was estimated using three different algorithms: single likelihood ancestor counting (SLAC) derived from the Suzuki-Gojobori approach (Suzuki and Gojobori, 1999); fixed-effects likelihood (FEL), which fits an ω ratio to every site and uses the likelihood ratio to test whether $dN \neq dS$; and random effect likelihood (REL), a variant of the Nielsen-Yang approach (Yang and Nielsen, 2000) that assumes the existence of a discrete distribution of rates across sites and allows both dS and dN to vary independently site-by-site, together with a mixed effects model of evolution (MEME) in order to detect episodic diversifying selection. The three methods are described in more detail elsewhere (Kosakovsky Pond and Frost, 2005) (Murrell et al., 2012).

Finally, in order to investigate whether the sampled sequences had been subject to selective pressure at population level (i.e. along internal branches), an internal fixed effects likelihood (IFEL) method (Pond et al., 2006) was also used.

The PRIME model was used to take into account the biochemical properties of the amino acids on the basis of their chemical composition, polarity, volume, iso-electric point, and hydropathy (Conant et al., 2007).

In order to select the sites under selective pressure, we assumed a p value of ≤ 0.1 or a posterior probability of ≥ 0.9. The analyses were made using the Web-based Datamonkey interface (http://www.datamonkey.org/)(Pond et al., 2005).

3. RESULTS

3.1. Likelihood mapping

The entire data set of 73 complete MARV genomes underwent likelihood mapping. The evaluation of 10,000 randomly chosen quartets showed that 0.5% fell into the central area of the likelihood map and 99% were at the corners of the triangle, thus suggesting that the alignment contained sufficient phylogenetic information (Fig. 1).

3.2. Phylogenetic analysis

Bayesian analysis of the entire Marburg virus data set showed two highly significant evolutionary lineages (pp=1) corresponding to MARV and RAVV (highlighted in Fig. 2). The MARV lineage included two main clades, the first of which (A) branched into three highly significant subclades (pp from 1 to 0.99): A1 consisted of 10 Ugandan isolates (five obtained during the first documented Marburg epidemic in 1967, and five sporadic cases from 2008-2012) and a single isolate from the 2000 epidemic in the DRC; A2 consisted of a group of Kenyan isolates obtained in 1980; and A3 consisted of isolates obtained during the 2004-2005 epidemic in Angola. The second major clade (B), which had an outgroup consisting of a single 1975 isolate from Zimbabwe, branched into two subclades: B1 (pp=1) included the majority of the isolates obtained during the 1999-2000 epidemic in the DRC; and B2 consisted of a number of Ugandan isolates obtained in 2007-2009. Interestingly, the DRC isolates in B1 formed at least three highly significant monophyletic groups (pp=1). The RAVV isolates segregated into two highly significant subclades (pp=1): RAVV1 included Kenyan strains isolated in 1987, and RAVV2 a number of Ugandan isolates from 2007-2009 and a single sequence obtained in the DRC in 1999. The analysis of genetic distances showed a median nucleotide difference of 16.9 substitutions/100 sites (range 16.7-17.0) between the MARV and RAVV lineages, whereas the median difference between the MARV subclades (A1-3 and B1-2) was 6.0 substitutions/100 sites (range 0.9-6.5).

Root-to-tip regression analysis showed that the root of the MARV strain was between clade A and

B, and the correlation coefficient of 0.83 suggests a significant relationship between genetic divergence and time.

3.3. Evolutionary rate estimates and dated tree reconstruction

Comparison of the different coalescent models using the BF test on the entire dataset showed that the model best fitting the data was a coalescent prior BSP (2lnBF constant *vs* BSP = 54.4, and exponential *vs* BSP = 1693.8) under a log-normal relaxed clock (2lnBF strict *vs* relaxed clock = 23.8). The estimated mean evolutionary rate was $3.3x10^{-4}$ subs/site/year (95%HPD 2.0-$4.8x10^{-4}$). Analysis of the MARV strain using the tree root estimated by the Path O Gen program confirmed the value obtained using the entire dataset (evolutionary rate $3.0x10^{-4}$; 95% HPD 1.8-$4.2x10^{-4}$).The analysis was also made after partitioning the entire coding sequences into codon positions: on average, the rate in codons 1 + 2 was six times lower than that in codon 3 (0.388 *vs* 2.22).

Figure 3 shows the phylogeographical maximum clade credibility (MCC) tree, the topology of which is identical to that described above, with high posterior probability values for the nodes corresponding to the main clades and subclades.

The tMRCAs of the internal nodes were estimated on the basis of the mean evolutionary rate values and intervals (Tab. 1). The root tMRCA was 353.8 (between 150 and 635) years ago (YA). The MARV strain had an estimated mean root tMRCA of 177.9 YA (95%HPD 87-284 YA), which indicates that particularly clade A originated in the 19th century, whereas clade B (including the majority of the DRC isolates) originated between the 1920s and 1970s (on average, in 1952). The mean tMRCA of subclade A1 (including the sequences of the first European outbreak) dates back to the 1950s (between 1937 and 1962), whereas those of subclades A2 (Kenya) and A3 (Angola) preceded the beginning of their respective epidemics by only a few years (1-6 years). The 1999-2000 DRC isolates in subclade B1 coalesced in 1987 (credibility interval 1981-1993), and the Ugandan isolates in subclade B2 originated in 1999 (95%HPD 9.32-18.19 YA). Finally, the RAVV strain had a mean tMRCA of 33.8 YA (95%HPD 27-42 YA) and therefore dated back to 1978.

3.4. Phylogeographical analysis

The location of the tree root was between Kenya (state posterior probability, spp=0.35) and Uganda (spp=0.29); this region had a 0.6 combined spp of being the original location of circulating viruses. As shown in Figure 3, the most probable origin of the MARV strain was Uganda (spp=0.41 *vs* 0.28 for the second most probable location), and that of RAVV was Kenya (spp=0.71 *vs* 0.18 for the second most probable location).

Table 1 shows the most probable locations and the associated state posterior probabilities of the various clades and subclades. The MARV clade A had an MRCA that was most probably located in Uganda (spp=0.45), whereas that of clade B was located in the DRC (spp=0.5). The MRCA locations of the subclades always corresponded with the sampling locations.

The significance of the spatio-temporal linkage between the different geographical locations was calculated using a BF test, assuming that the posterior and prior probabilities of the individual rates were zero. There was a mean of 5.2 non-zero rates (95%HPD=4-7) out of 10, and the rates with a BF of >3 were between Uganda and the DRC (BF=3569), between Kenya and Uganda (BF=3.4), and between Zimbabwe and the DRC (BF=3.5). The asymmetrical substitution model for discrete traits, which allows for different rates of migration between locations, showed that the significant pathways were from Uganda to the DRC (BF=42.0) and in the opposite direction (BF=5.7), and from Kenya to Uganda (BF=6.8). Figure 4 shows the migration routes and estimated times.

The gene flows (migrations) between the different geographical areas were also investigated using a modified version of the Slatkin and Maddison method. The null hypothesis of panmixia was rejected for all of the countries in the bubblegram by the randomisation test ($p = 0.0001$). The gene flows were highly asymmetrical, with the viral infections expanding from Uganda to various countries: almost 42.9% were from Uganda to the DRC, whereas those from Uganda to Zimbabwe, Kenya and Angola accounted for respectively 14.3%, 28.6% and 14.3% of the migrations (Fig. 5).

3.5. Selection pressure analysis

The ML estimate of the dN/dS ratio (ω) gave a mean value of 0.126 (95% CI: 0.116-0.136), with significant differences between the lineages (the LRT from the global and local model indicated a 2Δ likelihood of 132, p=0.0176). The analysis of site-by-site selection pressure showed that 24 sites were under positive selection (supported by at least two methods at a significance level of 90%; Table 2), 14 of which were significantly supported by the IFEL method (which detects sites selected along the internal branches) and mainly corresponded to subclade-specific substitutions (Fig. 6). The majority of codons under positive pressure were in the L protein (14/2231 amino acids, 0.62%) and glycoprotein (5/681 amino acids, 0.73%). There were also codons under positive pressure in the NP (2/695 amino acids, 0.28%), and VP35, VP40 and VP30 proteins, each of which had a single residue under positive selection (0.3-0.35% of residues).

Finally, the PRIME method was used to estimate the changes in biochemical properties induced by the evolutionary process. As shown in Table 2, four residues had modified properties (three in the polymerase and one in the VP30 protein).

4. DISCUSSION

Marburg virus is still classified as a single viral species despite the considerable genetic differences between the MARV and RAVV strains (16% at nucleotide level in this study). Our phylogenetic analysis showed that the MARV isolates grouped into two main clades (A and B) that had a mean genetic difference of 6% of nucleotides, and branched into a number of subclades that largely corresponded to the geographical sources of the isolates. The isolates of the initial European outbreak in 1967 (a virus probably originating in Uganda) formed a significant group that fell into a single subclade (A1) together with the other Ugandan human and bat isolates sampled between 2008 and 2012, and a sequence obtained in Durba (DRC, 1999). Other isolates obtained in Kenya in 1987, and during the 2004-2005 outbreak in Angola, formed separate subclades (A2 and A3). Clade B included the large majority of MARV isolates from the 1999-2000 epidemic in Durba (B1), and Ugandan isolates obtained in 2007-2009 grouped together in the closely related but separate subclade B2.

The main aim of this study was to assess the spatio-temporal phylogeny of Marburg virus. The genetic heterogeneity of the MARV and RAVV strains made it difficult to obtain a single estimate of the substitution rate. The global substitution rate for the entire viral genome ($3.3x10^{-4}$; CI 2.0-$4.8x10^{-4}$ subs/site/year) was confirmed by the independent analysis of the MARV strain alone, which gave a high correlation coefficient in the regression analysis of the genetic distance against the sampling years. This value is similar to that previously published by (Carroll et al., 2013) and is included within his confidence interval.

Our Bayesian phylogeographical analysis assigned similar posterior probabilities to Kenya and Uganda at the tree-root, with the combined posterior probability of 0.6 indicating the origin of the viruses in this area. The most probable location of the MARV lineage was Uganda (spp=0.41), and that of the RAVV lineage was Kenya (spp=0.71). The use of a different approach based on maximum parsimony indicated that Uganda was the most probable location of the tree root, and this

was supported by the highly significant flows from Uganda to all of the other sites of Marburg virus infections.

The uncertainty of the root of the tree including both MARV and RAVV was also confirmed by the tMRCA estimates dating back to an average of 354 YA, but with a broad credibility interval (from 150 to 635 years) that has also been found by other authors (Carroll et al., 2013). This suggests the absence of one or more steps in the reconstruction of the evolutionary history of the two viruses. More precise estimates have been obtained for the tMRCAs of MARV (dating back to about the mid-19th century) and RAVV, which did not emerge until the 1970s.

Interestingly, the oldest MARV subclade (subclade A1), which includes the virus that caused the first known outbreak in Europe in 1967, had an estimated tMRCA in the early 1950s, thus indicating the existence of MARV in an enzootic cycle long before it was first documented in humans.

On the basis of our phylogeographic reconstruction, Marburg virus spread westward to an area of the DRC that is relatively close to the Ugandan border, and southward, where it was identified in a cave in Zimbabwe between the 1950s and the 1980s (Fig. 3). We estimated that subclade B1 reached the DRC in the late 1980s and, in line with this, a disease similar to Marburg hemorrhagic fever has been documented in the area of Durba at least since it was first recognised in a miner in 1987, and it is known that outbreaks occurred in the same area in 1990, 1992, 1994 and 1997, before the large-scale epidemic of 1999-2000 (Bausch et al., 2006). The epidemic in the DRC was singular insofar as it was caused by at least nine lineages, possibly due to multiple spillovers from the animal reservoir (Bausch et al., 2006). The different lineages circulating in the DRC did not become extinct, but re-emerged in the Kitaka mine and Python Cave in Uganda seven years after the DRC outbreak. These observations suggest that the strains involved in the 1999-2000 epidemic were characterised by ecological/epidemiological dynamics that were very different from those of all of the other strains (in Kenya, Angola and the Gabon), which remained confined to a single

outbreak and became extinct immediately afterwards. The significant bi-directional genetic flow between Uganda and the DRC found by our phylogeographical analysis suggests the possibility of repeated exchanges of strains between these two areas.

These results indicate that new viral variants may emerge periodically in reservoir animals, and become predominant in different places as a result of continuous viral evolution prevalently driven by genetic drift. This is supported by our analysis of positive selection pressure, which showed that only a minority of codons in the entire viral genome were subject to significant positive selection. The protein with the highest percentage of positively selected amino acid sites (0.7%) was the polymerase, followed by the glycoprotein (0.3%): however, the rate of selected sites was always <1%, frequently episodic, and did not cause any significant and long-lasting changes at population level. Only four sites showed a significant change in the biochemical properties of the mutated amino acids.

In conclusion, Marburg viruses seem to have two different patterns of viral dispersion. In the first, a single viral variant arises a median of four years before an outbreak, remains temporally and spatially confined to a specific episode, and apparently disappears immediately afterwards: this pattern includes the large-scale epidemic in Angola and the two episodes in Kenya. The second is characterised by long-term persistence and frequent re-circulation, as seen in the case of many Ugandan and some DRC strains, which may have persisted for years (from the early 1950s to 2012). The presence of all the major viral clades in Python Cave and the Kitaka mine supports a central role of Uganda as the cradle of the Marburg virus in Africa, although further studies of a larger number of isolates are needed to support this hypothesis.

FIGURE LEGENDS

Fig. 1. Likelihood mapping of Marburg virus sequences. Each dot represents the likelihoods of the three possible unrooted trees for each quartet randomly selected from the data set: the dots near the corners or sides respectively represent tree-like (fully resolved phylogenies in which one tree is clearly better than the others) or network-like phylogenetic signals (three regions in which it is not possible to decide between two topologies). The central area of the map represents a star-like signal (the region in which the star tree is the optimal tree). The numbers indicate the percentage of dots in the centre of the triangle.

Fig. 2. MCC tree of the 73 Marburg virus genomes reconstructed using MrBayes. The numbers on the branches represent posterior probabilities (see Materials and Methods for details). The main significant clades/subclades are highlighted. The scale axis below the tree shows the number of expected changes per site.

Fig. 3. The maximum clade credibility (MCC) tree of the Marburg virus whole genome sequences. The branches are coloured on the basis of the most probable location of the descendent nodes (AO = Angola, ZW = Zimbabwe, CD = Democratic Republic of the Congo, UG = Uganda, KE = Kenya). The numbers on the internal nodes indicate significant posterior probabilities (pp >0.8), and the scale at the bottom of the tree represents calendar years. The main geographical clades are highlighted. The histograms on the left show the state posterior probabilities (spp) of the different locations of the roots of the MARV and RAVV lineages.

Fig. 4. Significant non-zero migration rates of Marburg virus worldwide (i.e. those supported by a BF of >3). The relative strength of the support is indicated by the colour of the lines (from light

red = strong to dark red = weak). The map was reconstructed using SPREAD (see Materials and Methods). The numbers indicate the mean estimated year in which the virus entered the area.

Fig. 5. Phylogeographical mapping of the MARV genome sequences. The bubblegrams show the frequency of gene flows (migrations) to/from the African countries shown in Figure 2. The surface of each circle is proportional to the percentage of observed migrations in the ML genealogy. The migrations were inferred using a modified version of the Slatkin and Maddison algorithm.

Fig. 6. Phylogenetic tree based on the Marburg virus nucleotide sequences (the same as those shown in Figure 2.). The most represented amino acid substitutions are indicated on the basis of their position in the tree. The table on the left shows a total of 24 sites under positive selection pressure.

WEB REFERENCES

The National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/genbank/) 18 March 2016.

Biological sequence alignment editor (http://www.mbio.ncsu.edu/bioedit/bioedit.html) 18 March 2016.

jModelTest 2: HPC selection of models of nucleotide substitution (https://code.google.com/p/jmodeltest2/) 18 March 2016.

Tracer (http://tree.bio.ed.ac.uk/software/tracer/) 18 March 2016.

Tree Figure Production – FigTree (http://tree.bio.ed.ac.uk/software/figtree/) 18 March 2016.

Bielejec F., Rambaut A., Suchard M.A & Lemey P. SPREAD: Spatial Phylogenetic Reconstruction of Evolutionary Dynamics. Bioinformatics, 2011, 27:2910-2912. (http://www.kuleuven.be/aidslab/phylogeography/SPREAD.html) 18 March 2016.

Google Earth (https://www.google.it/intl/it/earth/) 18 March 2016.

Adaptive Evolution Service – Datamonkey (http://www.datamonkey.org/) 18 March 2016

Red List Maps (http://maps.iucnredlist.org/map.html?id=29730)

REFERENCES

Adjemian, J., et al., 2011. Outbreak of Marburg hemorrhagic fever among miners in Kamwenge and Ibanda Districts, Uganda, 2007. J Infect Dis 204 Suppl 3, S796-799.

Amman, B.R., et al., 2014. Marburgvirus resurgence in Kitaka Mine bat population after extermination attempts, Uganda. Emerg Infect Dis 20, 1761-1764.

Bausch, D.G., et al., 2006. Marburg hemorrhagic fever associated with multiple genetic lineages of virus. N Engl J Med 355, 909-919.

Bielejec, F., et al., 2011. SPREAD: spatial phylogenetic reconstruction of evolutionary dynamics. Bioinformatics 27, 2910-2912.

Brauburger, K., et al., 2012. Forty-five years of Marburg virus research. Viruses 4, 1878-1927.

Carroll, S.A., et al., 2013. Molecular evolution of viruses of the family Filoviridae based on 97 whole-genome sequences. J Virol 87, 2608-2616.

Conant, G.C., et al., 2007. Modeling amino acid substitution patterns in orthologous and paralogous genes. Mol Phylogenet Evol 42, 298-307.

Conrad, J.L., et al., 1978. Epidemiologic investigation of Marburg virus disease, Southern Africa, 1975. Am J Trop Med Hyg 27, 1210-1215.

Drummond, A.J., et al., 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol 22, 1185-1192.

Drummond, A.J., et al., 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 29, 1969-1973.

Feldmann, H., 2006. Marburg hemorrhagic fever--the forgotten cousin strikes. N Engl J Med 355, 866-869.

Guindon, S., Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52, 696-704.

Huelsenbeck, J.P., Ronquist, F., 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17, 754-755.

Johnson, E.D., et al., 1996. Characterization of a new Marburg virus isolated from a 1987 fatal case in Kenya. Arch Virol Suppl 11, 101-114.

Kass, R.E., Raftery, A.E., 1995. Bayes factors. Journal of American Statistical Association 90, 773-795.

Kosakovsky Pond, S.L., Frost, S.D., 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol 22, 1208-1222.

Kuhn, J.H., 2008. Filoviruses. A compendium of 40 years of epidemiological, clinical, and laboratory studies. Arch Virol Suppl 20, 13-360.

Kuhn, J.H., et al., 2010. Proposal for a revised taxonomy of the family Filoviridae: classification, names of taxa and viruses, and virus abbreviations. Arch Virol 155, 2083-2103.

Lemey, P., et al., 2009. Bayesian phylogeography finds its roots. PLoS Comput Biol 5, e1000520.

Murrell, B., et al., 2012. Detecting individual sites subject to episodic diversifying selection. PLoS Genet 8, e1002764.

Peterson, A.T., et al., 2006. Geographic potential for outbreaks of Marburg hemorrhagic fever. Am J Trop Med Hyg 75, 9-15.

Pond, S.L., et al., 2006. Adaptation to different human populations by HIV-1 revealed by codon-based analyses. PLoS Comput Biol 2, e62.

Pond, S.L., et al., 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21, 676-679.

Posada, D., 2008. jModelTest: phylogenetic model averaging. Mol Biol Evol 25, 1253-1256.

Pourrut, X., et al., 2009. Large serological survey showing cocirculation of Ebola and Marburg viruses in Gabonese bat populations, and a high seroprevalence of both viruses in Rousettus aegyptiacus. BMC Infect Dis 9, 159.

Rambaut, A., 2000. Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. Bioinformatics 16, 395-399.

Rougeron, V., et al., 2015. Ebola and Marburg haemorrhagic fever. J Clin Virol 64, 111-119.

Schmidt, H.A., et al., 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18, 502-504.

Siegert, R., et al., 1968. Detection of the "Marburg Virus" in patients. Ger Med Mon 13, 521-524.

Slatkin, M., Maddison, W.P., 1989. A cladistic measure of gene flow inferred from the phylogenies of alleles. Genetics 123, 603-613.

Smith, D.H., et al., 1982. Marburg-virus disease in Kenya. Lancet 1, 816-820.

Suzuki, Y., Gojobori, T., 1999. A method for detecting positive selection at single amino acid sites. Mol Biol Evol 16, 1315-1328.

Swanepoel, R., et al., 2007. Studies of reservoir hosts for Marburg virus. Emerg Infect Dis 13, 1847-1851.

Thompson, J.D., et al., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22, 4673-4680.

Towner, J.S., et al., 2009. Isolation of genetically diverse Marburg viruses from Egyptian fruit bats. PLoS Pathog 5, e1000536.

Towner, J.S., et al., 2006. Marburgvirus genomics and association with a large hemorrhagic fever outbreak in Angola. J Virol 80, 6497-6516.

Towner, J.S., et al., 2007. Marburg virus infection detected in a common African bat. PLoS One 2, e764.

Yang, Z., Nielsen, R., 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol 17, 32-43.

Table(s)

Table 1. Estimated times of the most recent common ancestors (tMRCAs) of the main clades and credibility intervals (95%HPD), with calendar years, most probable locations, and state posterior probabilities (spp) of the 73 Marburg virus complete genomes.

		tMRCA			Years				
CLADE	**Subclade**	**Mean**	**[1]LCI**	**[2]UCI**	**Mean**	**LCI**	**UCI**	**Location**	**[3]spp**
TREE ROOT		353.8	150	635	1658	1862	1377	KE	0.35
ROOT MARV		177.9	87	284	1834	1925	1728	UG	0.41
A		142.65	67.66	221.45	1869	1944	1790	UG	0.45
	A1	61.01	50.08	74.66	1951	1962	1937	UG	0.98
	A2 (KE80)	35.07	33.04	38.05	1977	1979	1974	KE	1
	A3 (AO)	9.07	7.62	10.92	2003	2004	2001	AO	1
B		59.73	39.76	91.43	1952	1972	1920	CD	0.5
	B1 (CD)	25.1	18.74	31.11	1987	1993	1981	CD	1
	B2 (UG)	13.01	9.32	18.19	1999	2003	1994	UG	1
ROOT RAVN		33.8	27	42	1978	1985	1970	KE	0.71
	RAVN1	28.61	26.27	31.41	1983	1986	1981	KE	1
	RAVN2	20	13.7	26.92	1992	1998	1985	UG	0.58

[1]LCI: Lower limit of the Credibility Interval; [2]UCI: Upper limit of the Credibility Interval; [3]spp: State posterior probability;

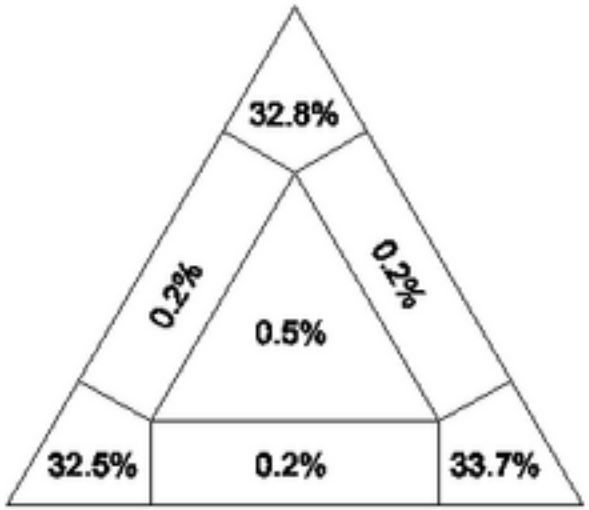
Table(s)

Table 2. Selection pressure analysis. SLAC, FEL, IFEL, REL, MEME and PRIME analyses of the 73 Marburg virus genomes (see Methods).

		METHODS					
Proteins	**CODON POSITION**	**p - VALUE FEL**	**p - VALUE IFEL**	**PP REL**	**p - VALUE MEME**	**p - VALUE PRIME**	**Modified biochemical properties**
NP	440	0.06	0.03	0.99	-	-	
NP	502	-	0.08	0.95	0.05	-	
VP35	725	0.07	-	0.96	0.04	-	
VP40	1208	0.06	-	0.99	0.0009	-	
GP	1333	-	0.09	0.94	-	-	
GP	1543	-	-	0.96	0.08	-	
GP	1614	0.07	-	0.99	0.04	-	
GP	1676	0.03	0.04	0.99	0.01	-	
GP	1697	0.07	0.04	0.99	0.03	-	
VP30	2132	-	-	-	2.8 e-5	0.03 - 0.000	Isoelectric Point - Hydrophobicity
L	2627	-	-	0.95	0.09	-	
L	2878	0.09	0.05	0.99	-	-	
L	2884	0.06	0.08	0.99	0.09	-	
L	2976	-	-	0.98	3.2 e-7	0.03	Hydrophobicity
L	3031	-	-	-	8.6 e-6	0.001	Isoelectric Point
L	3897	-	-	0.99	0.01	-	
L	4251	-	0.1	0.99	0.004	0.024	Volume
L	4293	-	-	0.99	0.06	-	
L	4314	-	0.07	0.96	-	-	
L	4323	0.07	0.04	0.99	0.1	-	
L	4333	-	0.07	0.95	-	-	
L	4347	-	0.06	0.99	-	-	
L	4378	-	0.08	0.95	-	-	
L	4408	0.06	0.05	0.96	0.07	-	

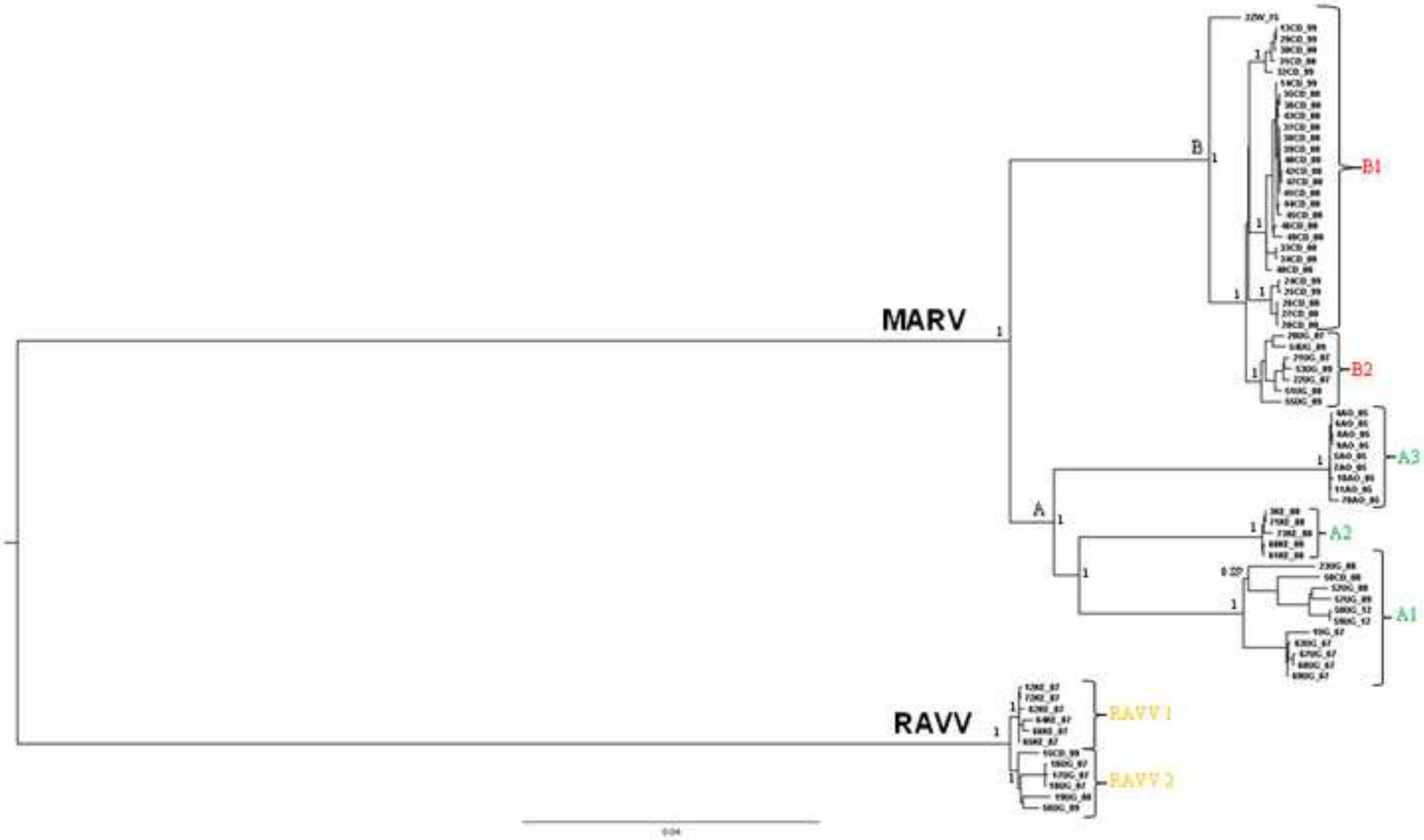
Figure(s)

Click here to download high resolution image



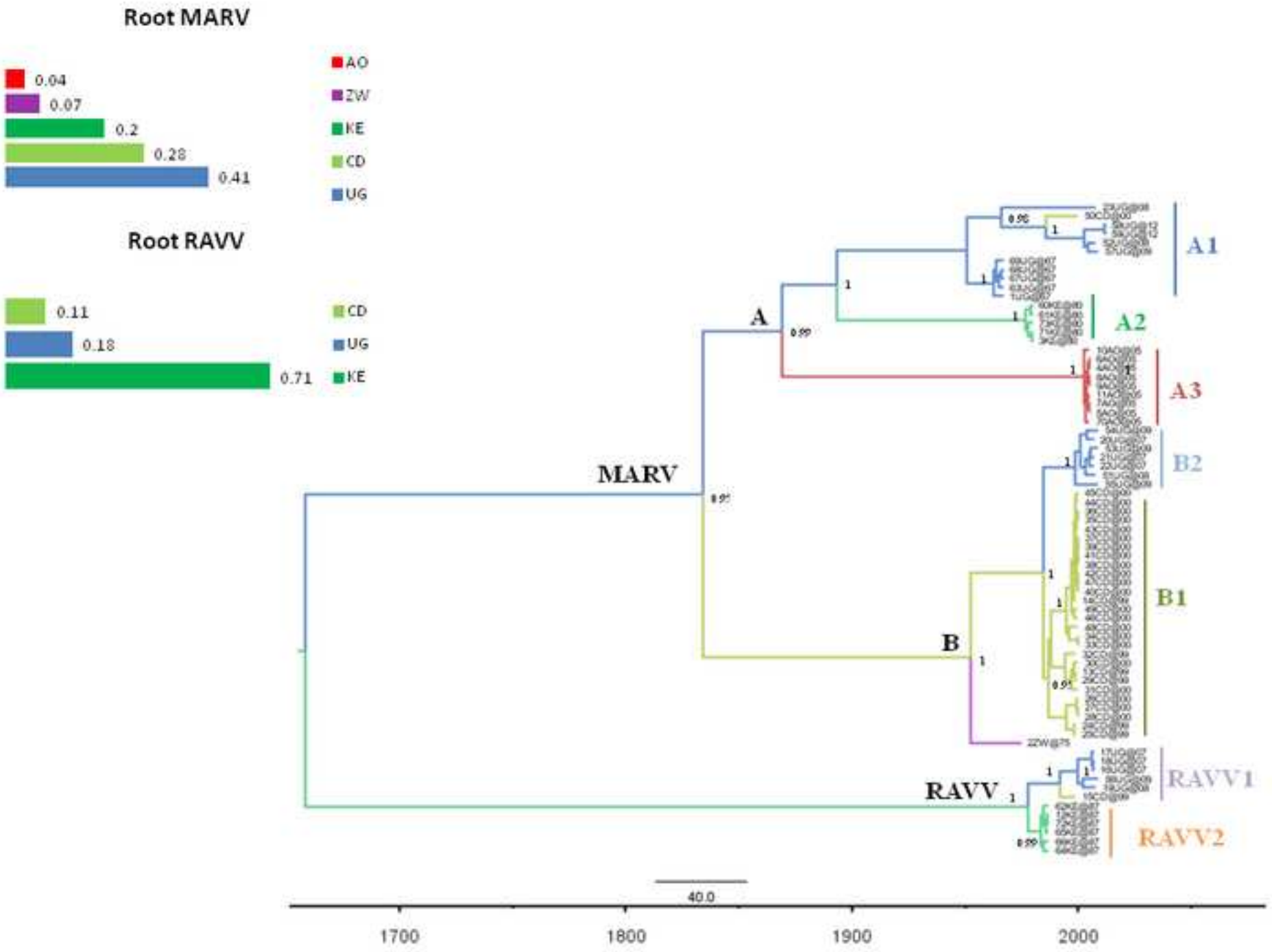
Figure(s)

Click here to download high resolution image



Figure(s)

Click here to download high resolution image

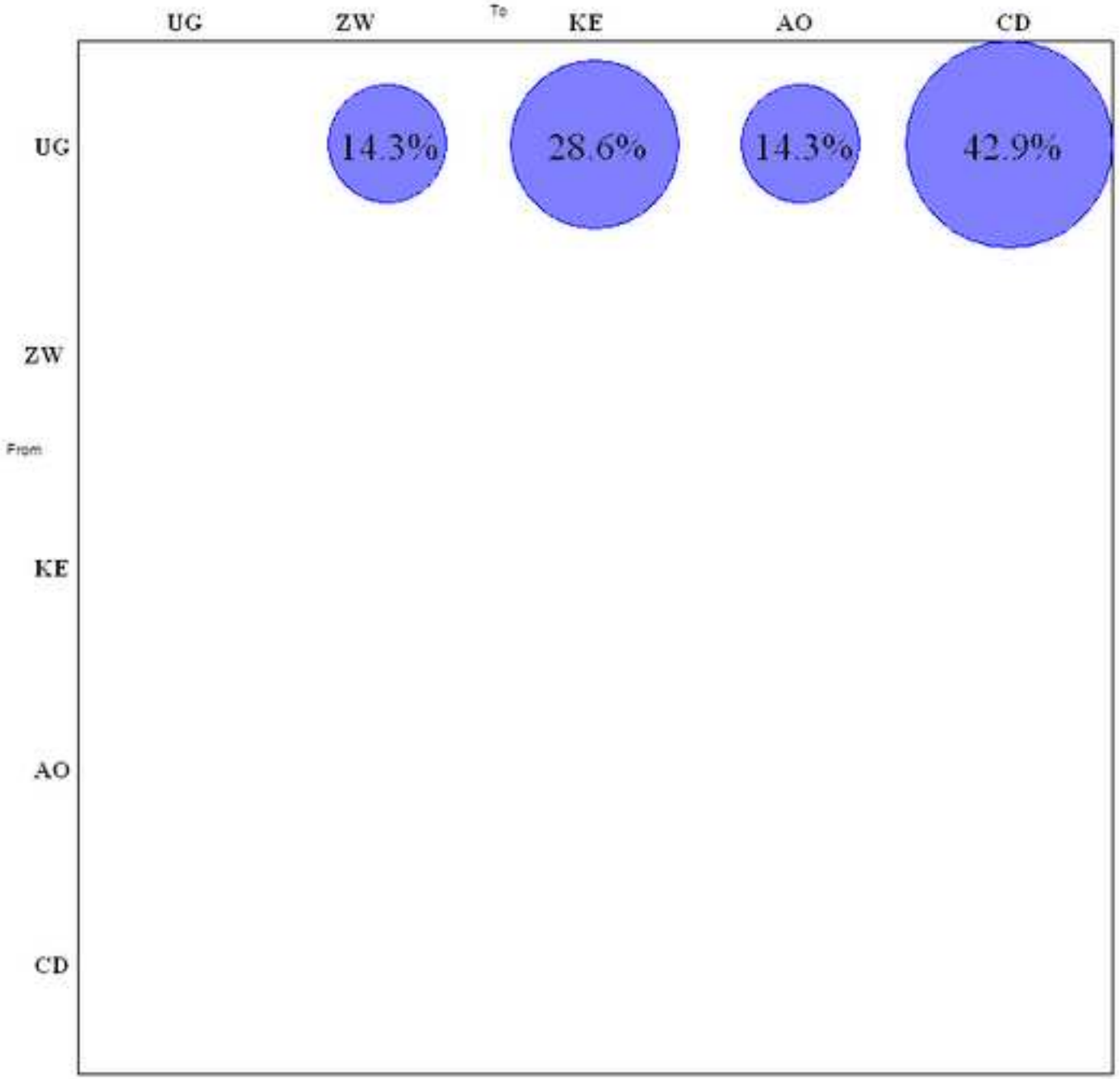


Figure(s)

Click here to download high resolution image

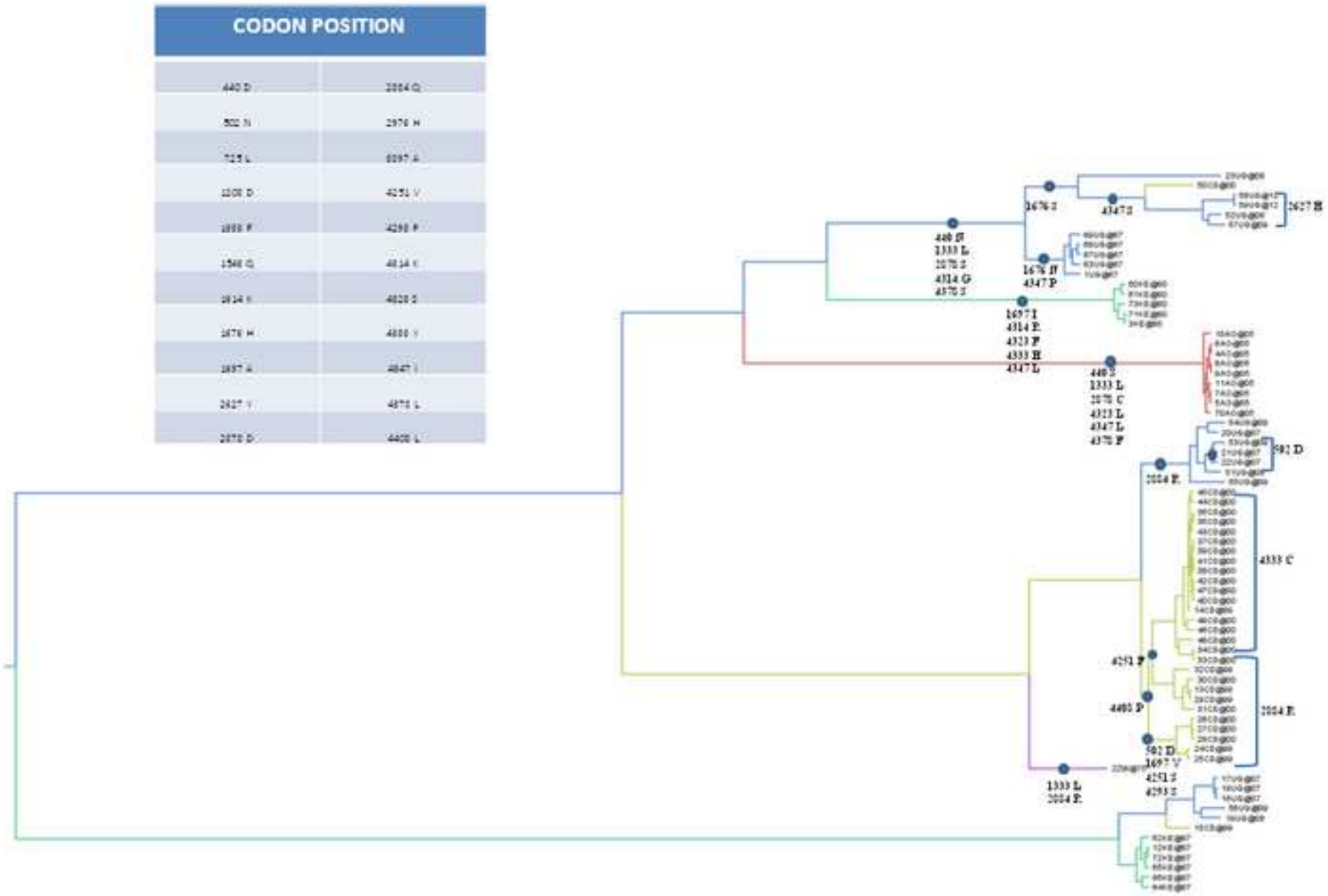
Figure(s)

Click here to download high resolution image



Figure(s)

Click here to download high resolution image



Supplementary Material

Click here to download Supplementary Material: Supplementary Table 1.docx