Virus Evolution, 2015, 1(1): vev016

doi: 10.1093/ve/vev016
Research article

Host ecology determines the dispersal patterns of a plant virus

Nídia Sequeira Trovão,[1,*,†] Guy Baele,[1] Bram Vrancken,[1] Filip Bielejec,[1] Marc A. Suchard,[2,3] Denis Fargette,[4] and Philippe Lemey[1,‡]

[1]Department of Microbiology and Immunology, Rega Institute, KU Leuven, 3000 Leuven, Belgium, [2]Departments of Biomathematics and Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA 0095-1766, USA, [3]Department of Biostatistics, UCLA Fielding School of Public Health, University of California, Los Angeles, CA 0095-1766, USA and [4]Institut de Recherches pour le Développement (IRD), UMR IPME (IRD, CIRAD, Université de Montpellier), 34394 Montpellier, France

*Corresponding author: E-mail: Nidia.SequeiraTrovao@rega.kuleuven.be
†http://orcid.org/0000-0002-2106-1166
‡http://orcid.org/0000-0003-2826-5353

Abstract

Since its isolation in 1966 in Kenya, rice yellow mottle virus (RYMV) has been reported throughout Africa resulting in one of the economically most important tropical plant emerging diseases. A thorough understanding of RYMV evolution and dispersal is critical to manage viral spread in tropical areas that heavily rely on agriculture for subsistence. Phylogenetic analyses have suggested a relatively recent expansion, perhaps driven by the intensification of agricultural practices, but this has not yet been examined in a coherent statistical framework. To gain insight into the historical spread of RYMV within Africa rice cultivations, we analyse a dataset of 300 coat protein gene sequences, sampled from East to West Africa over a 46-year period, using Bayesian evolutionary inference. Spatiotemporal reconstructions date the origin of RMYV back to 1852 (1791–1903) and confirm Tanzania as the most likely geographic origin. Following a single long-distance transmission event from East to West Africa, separate viral populations have been maintained for about a century. To identify the factors that shaped the RYMV distribution, we apply a generalised linear model (GLM) extension of discrete phylogenetic diffusion and provide strong support for distances measured on a rice connectivity landscape as the major determinant of RYMV spread. Phylogeographic estimates in continuous space further complement this by demonstrating more pronounced expansion dynamics in West Africa that are consistent with agricultural intensification and extensification. Taken together, our principled phylogeographic inference approach shows for the first time that host ecology dynamics have shaped the historical spread of a plant virus.

Key words: plant virus; RYMV; phylogeography; Bayesian inference; viral evolution; disease ecology.

1. Introduction

Although phylodynamics have become a burgeoning area of research focused on many human and animal viruses, comparatively fewer analyses have targeted the interaction between evolutionary and ecological dynamics in plant viruses. On the one hand, this may be explained by a biased interest in viruses that directly impact human health or that may emerge as zoonotic pathogens. On the other hand, it is unclear to what extent phylodynamic concepts apply to plant viruses because their evolutionary and ecological dynamics may not necessarily occur on the same time scale. Lower rates of plant virus evolution have been inferred based on co-divergence assumptions,

but also from sequence analysis of old samples (Rodríguez-Cerezo *et al.*, 1991; Fraile *et al.*, 1997; Gibbs *et al.*, 2010). In recent years however, evidence has accumulated for a rapid evolutionary rate in specific plant viruses, as first demonstrated for rice yellow mottle virus (RYMV) (Fargette *et al.*, 2008b) and zucchini yellow mosaic virus (Simmons, Holmes, and Stephenson, 2008). Nucleotide substitution rates falling within the range of animal RNA virus rates have also been reported for particular *Geminiviridae* (Duffy and Holmes, 2008, 2009; Monjane *et al.*, 2011) and *Luteoviridae* (Pagán and Holmes, 2010).

In addition to clarifying the tempo and time scale of plant virus evolution, molecular sequence analyses may also probe spatial population structure and shed light on the transmission dynamics that gave rise to the current spatial distribution of plant viral lineages. It is therefore not surprising that the field of plant virus epidemiology has started to adopt recent statistical inference methodology that integrates temporal and spatial dynamics in a phylogenetic context (Lemey *et al.*, 2009, 2010; Drummond *et al.*, 2012). As an example of this, the ongoing global spread of tomato yellow leaf curl virus (TYLCV) has attracted significant interest as a potential threat to tomato production in all temperate parts of the world. Motivated by the need to unravel the ecological and economic risks associated with such viral invasions (Lefeuvre *et al.*, 2010) applied Bayesian phylogeographic methods to reconstruct the spatiotemporal history of TYLCV spread and diversification. This revealed that, while the virus likely originated in the Middle East during the first half of the 20th century, this area remained epidemiologically relatively isolated. Instead, many global movements of TYLCV appear to have been seeded from the Mediterranean basin. As another example of a tropical plant virus that poses a threat to African food security, maize streak virus (MSV) has caused severe epidemics throughout the maize growing regions of Africa. Recent insights gained from Bayesian spatiotemporal reconstructions point at southern Africa as the most probable location from which MSV emerged at the beginning of the 20th century, and subsequently spread transcontinentally at an average rate of 32.5 km/year (Monjane *et al.*, 2011). As the etiological agent of the most damaging plant virus disease in the world, cassava mosaic-like virus (CMV) has caused devastating crop losses across sub-Saharan Africa. This epidemic was estimated to have originated in the late 1930s in mainland Africa with subsequent introductions to the southwest Indian ocean islands between 1988 and 2009 (De Bruyn *et al.*, 2012).

Among the fast evolving plant viruses, RYMV is also of particular interest because it circulates in most rice growing countries on the African continent (Bakker *et al.*, 1974; Abubakar *et al.*, 2003), impacting the lives of millions of impoverished Africans that rely on rice agriculture for subsistence (Abo, Sy, and Alegbejo, 1998). Symptoms of RYMV infection range from discolouration, stunting and ultimately sterilisation of the plant, resulting in devastating epidemics with yield losses that vary from 10 to 100% depending on how early the infection sets in, the type of rice cultivation and the rice cultivars used (Allarangaye *et al.*, 2006). RYMV is a member of the *Sobemovirus* genus with a genome composed of a single-stranded positive RNA molecule encompassing about 4450 nucleotides, organised into five open reading frames (ORFs) that overlap (except for ORF1) (Ling *et al.*, 2013; Sõmera, Sarmiento, and Truve, 2015). The virus is transmitted by chrysomelid beetles (Bakker *et al.*, 1974), by mammals (Sarra and Peters, 2003), and by contact during cultural practices (Traoré *et al.*, 2006), but no evidence of seed transmission has been found (Konate *et al.*, 2001). The known natural host range of RYMV is limited to the two species of cultivated rice *Oryza sativa L.* and *Oryza glaberrima Steud*, and a few related wild grasses (Bakker *et al.*, 1974).

Although the history of rice agriculture in Africa dates back many centuries, RYMV was only first reported in 1966 in Kenya (Bakker *et al.*, 1974). With nucleotide substitution rates ranging from 4×10^{-4} to 1.2×10^{-3} nucleotides/site/year, evolutionary studies have characterised the virus as a measurable evolving population with a most recent common ancestor (MRCA) dating back to around 1811 (Fargette *et al.*, 2008a; Pinel-Galzi *et al.*, 2009). In addition to rapid evolutionary rates, specific RYMV gene sequences also show evidence for recombination between particular ORFs, but not within individual ORFs (Pinel-Galzi *et al.*, 2009). Early spatial genetic analyses have suggested a fairly regular pattern of spread with a correlation between genetic and geographic distances and no evidence of long-range dispersal. Based on comparisons of genetic diversity, these analyses have also implicated East Africa as the area of early diversification (Abubakar *et al.*, 2003). Specifically, more recent surveys confirm a large concentration of RYMV diversity in eastern Tanzania (Pinel-Galzi *et al.*, 2009), a region that is isolated by the Indian Ocean to the east and by the Eastern Arc Mountains to the west. A relatively long history of co-existence of RYMV strains in conditions that support habitat fragmentation indeed point at this region as a putative origin for the virus (Fargette *et al.*, 2004).

RYMV diversity shows a pronounced and characteristic geographic structure, and has been classified into S1–S6 strains based on serological typing and phylogenetics. Five serological profiles have been identified: three in West and Central Africa (Ser1, Ser2, and Ser3) and two in East Africa (Ser4 and Ser5). Apart from Ser5, which is divided into the S5 and S6 strains, these serotypes also correspond to the S1–S6 strain types (Pinel *et al.*, 2000; Traore *et al.*, 2005). The spatial structure of the epidemic, with different strains circulating in different countries, suggests a relatively recent expansion perhaps driven by the intensification of agricultural practices (Konaté and Fargette, 2001; Abubakar *et al.*, 2003). Evolutionary and ecological hypotheses about the origin and spread of RYMV have however not been examined in a coherent statistical framework. Recent extensions of Bayesian phylogenetic diffusion models for discrete traits now offer the opportunity to formally evaluate predictors of spatial spread. In particular, the recently developed GLM approach parameterises rates of diffusion as a function of potential predictors (Lemey *et al.*, 2014). This approach has for example identified human and animal transportation measures as the drivers of spatial spread for different influenza viruses (Lemey *et al.*, 2014; Nelson *et al.*, 2015), and it may also be useful for identifying the factors responsible for plant virus spread.

Here, we demonstrate the value of Bayesian phylodynamic inference methodologies in plant molecular epidemiology by focusing on the patterns of RYMV spread across Africa and reconstructing its phylogeographic history. We test spatiotemporal hypotheses about the origins of RYMV using state-of-the-art Bayesian statistical inference, quantify the dynamics of spatial spread in both East and West Africa, and formally assess the relationship between RYMV spread and the history of rice cultivation in Africa.

2. Methodology

2.1 Dataset compilation

We have assembled a RYMV sequence dataset by retrieving all publicly available ORF4 CP gene sequences from GenBank (on 4 September 2012) and combining these with additional samples made available by collaborators, which have now been published in recent studies (Hubert *et al.*, 2013; Longué *et al.*, 2013).

The sequences were aligned using MAFFT version 6.864b (Katoh and Toh, 2008) and manually edited in Se-Al (tree.bio.ed.ac.uk/software/seal). The final dataset consists of 300 sequences that were sampled between 1966 and 2012 in 20 countries across East and West Africa (Supplementary Fig. S1) and covers all countries in which RYMV has been reported. Although we used countries as locations in the discrete analyses, more specific location coordinates for all the 300 isolates were available and these were used in the continuous phylogeographic reconstructions.

To evaluate the impact of sampling bias on the root location estimate, we applied two different subsampling procedures to the sequence data. Specifically, we subsampled Tanzanian samples down to the next highest sampled country (Côte d'Ivoire) by: (1) randomly selecting 51 Tanzanian isolates and (2) selecting the same number of sequences that best represent the Tanzanian RYMV diversity using the Phylogenetic Diversity Analyzer tool (www.cibiv.at/software/pda) Minh, Klaere, and von Haeseler (2006, 2009). Both down-sampling procedures resulted in datasets of 266 sequences.

To contextualise particular continuous phylogeographic estimates, we also assembled two datasets for MSV and East African CMV (EACMV), both imposing an enormous burden to crops worldwide and in Africa specifically. For MSV, we resorted to the 333 full-genome recombinant-free dataset of Monjane *et al.* (2011). For EACMV, we obtained a dataset comprising 65 full-genomes from De Bruyn *et al.* (2012) by focusing on non-recombinant sequences originating from mainland Africa by and excluding outliers in a linear regression analysis of root-to-tip divergence (see 'Temporal signal' section below).

2.2 Temporal signal

In order to visually examine the degree of temporal signal—or signal for divergence accumulation over the sampling time interval—in the RYMV CP sequence data, we employed an exploratory linear regression approach. We first followed a standard approach of estimating a maximum likelihood (ML) tree under a non-clock (unconstrained) generalised time-reversible (GTR) substitution model with discrete Γ-distributed rate variation among sites using PhyML (Guindon *et al.*, 2010) and plotted the root-to-tip divergences as a function of sampling time according to a rooting that maximises the Pearson product-moment correlation coefficient using Path-O-Gen (tree.bio.ed.ac.uk/software/pathogen). For comparison, we only plot the root-to-tip divergences for the same subset of taxa that is used in the procedure discussed below.

We also explored an alternative approach that attempts to avoid rate heterogeneity imposed on the deep branches connecting different RYMV clusters, and only considers the overall divergence accumulation within these specific clusters. To identify rooted clusters, we used the maximum clade credibility (MCC) tree from the Bayesian analyses (see below) and re-estimated branch lengths under a non-clock GTR + Γ substitution model using PAUP* v4.0b10. We then selected distinct phylogenetic clusters that contain taxa with a minimum sampling time interval of 15 years and that were associated with a posterior probability support >0.75 (Supplementary Fig. S2). For each of these clusters, we obtained cluster-specific MRCA-to-tip divergences as a function of sampling time based on the branch lengths estimated under the non-clock model. To explicitly model a cluster effect in the MRCA-to-tip divergence data d_{ij} for taxon i assigned to cluster j, we fit the following regression model:

$$d_{ij} = \beta_j + \delta x_{ij} + \epsilon_{ij}, \quad (1)$$

where β_j is the intercept for cluster j, δ is the phylogenetically unadjusted rate of substitution, x_{ij} is the sampling time and ϵ_{ij} an independent 0-mean error term. To visually plot a single regression line through the MRCA-to-tip divergence data with δ as slope, we subtract the estimated cluster effect from the divergence measurements and plot the resulting values, $d_{ij} - E[\beta_j]$, as a function of sampling time. We acknowledge that linear regression techniques are not appropriate estimators of divergence through time as sequences do not represent independent data, but we merely employ these approaches to tentatively examine temporal signal in our data.

In addition to the visual exploration, we also conducted a date-randomisation test to evaluate to what extent Bayesian evolutionary rate estimates (using the Bayesian Evolutionary Analysis Sampling Trees (BEAST) package, see below) from the time-stamped data deviate significantly from estimates based on randomised tip dates, for which no particular relationship between sampling time and root-to-tip divergence is expected (Firth *et al.*, 2010). For this purpose, we here propose a novel implementation of this test that avoids having to analyse multiple date-randomised datasets. Because of the computationally expensive nature of Bayesian phylogenetic analyses, the number of these randomisations is generally limited (e.g. 20) in the standard test procedure (Duchêne *et al.*, 2015). Our BEAST implementation makes use of novel transition kernels that effectively randomise dates during the Markov chain Monte Carlo (MCMC) sampling procedure. Therefore, we do not have to rely on a number of specific randomisations, but conveniently, we average over all possible randomisations in a single analysis. We follow Duchêne *et al.* (2015) and use as our criterion for a significant temporal signal that the 95% credible interval (CI) for the rate estimate obtained from correct sampling times should not overlap with the CI for the estimate obtained while randomising sampling times.

2.3 Bayesian evolutionary inference

We reconstructed time-calibrated phylogenetic and phylogeographic histories using a Bayesian statistical framework implemented in the software package BEAST v1.8 (Drummond *et al.*, 2012). BEAST uses MCMC integration to average over tree space, so that each tree is weighted proportional to its posterior probability. All analyses were performed using the Broad-platform Evolutionary Analysis General Likelihood Evaluator (BEAGLE) library to enhance computation speed (Suchard and Rambaut, 2009; Ayres *et al.*, 2012).

2.4 Sequence evolution

To model the nucleotide substitution process, we partitioned the codon positions into first+second and third positions (Shapiro, Rambaut, and Drummond, 2006) and applied a separate Hasegawa–Kishino–Yano 85 (HKY85) substitution model (Hasegawa, Kishino, and Yano, 1985) to the two partitions, each with a discretised Γ distribution (HKY + Γ) to model rate heterogeneity across sites. To accommodate among-lineage rate variation we applied an uncorrelated relaxed molecular clock that models branch rate variation according to a lognormal distribution (Drummond *et al.*, 2006). To investigate the sensitivity of the time to MRCA (TMRCA) estimates with respect to the coalescent prior, we tested all currently available flexible, non-parametric demographic priors: the skyride (Minin *et al.*, 2008) (using uniform smoothing over all inter-coalescent intervals), skyline (Drummond *et al.*, 2005), and skygrid (Gill *et al.*, 2013)

(using a cut-off to 200 years with 100 grid points) model. Whereas it is generally recommended to employ 'time-aware' smoothing for the skyride model, which weighs the smoothing such that the effective population size changes between small, consecutive inter-coalescent intervals are penalised more than changes between intervals of larger size (Minin *et al.*, 2008), this appeared problematic for our dataset without strong temporal signal and resulted in MRCA estimates that were close to the oldest sample. We ran three independent runs for 100 million generations, sampling every 10 000th and discarded 10% as the chain burn-in. Stationarity and mixing (e.g. based on effective sample sizes $\geq$ 200 for the continuous parameters) were examined using Tracer version 1.6 (tree.bio.ed.ac.uk/software/tracer), and MCC trees were summarised using TreeAnnotator.

2.6 Discrete phylogeography

We modelled discrete location transitioning of RYMV between the 20 African countries throughout the phylogenetic history using both a reversible and non-reversible continuous-time Markov chain (CTMC) process (Lemey *et al.*, 2009; Edwards *et al.*, 2011) and performed the analysis with and without a Bayesian stochastic search variable selection (BSSVS) procedure to identify a sparse migration graph, which includes a restricted number of non-zero rates in the CTMC matrix. These analyses were performed both on the full dataset and the two subsampled datasets. We evaluated model fit using (log) marginal likelihood estimates obtained through path sampling (Lartillot and Philippe, 2006) and stepping-stone sampling (Xie *et al.*, 2011) procedures as implemented in BEAST (Baele *et al.*, 2012, 2013; Baele and Lemey, 2013). We ran various computational settings to assess convergence of the (log) marginal likelihood estimates. The number of location transitions ('Markov jumps') and the time spent in each location state ('Markov rewards') were estimated using stochastic mapping techniques (Minin and Suchard, 2008a,b).

In order to quantify the spatial structure, we measured the phylogenetic association in the location trait data by applying the association index (AI) to our posterior set of trees (Wang *et al.*, 2001; Lemey *et al.*, 2009). This metric quantifies the degree to which the same traits tend to cluster together relative to the expectation for randomised trait assignments. AI values close to 0 reflect strong phylogeny-location correlation whereas AI values close to 1 reflect the absence of phylogenetic structure for the trait (Wang *et al.*, 2001; Lemey *et al.*, 2009).

For both the discrete as well as the continuous phylogeographic analysis (*cfr*. below), we use TreeAnnotator (Drummond *et al.*, 2012) to summarise the location estimates on a MCC tree and visualise the tree with annotations using FigTree (tree.bio.ed.ac.uk/software/figtree). We converted the location-annotated trees to keyhole markup language format using the Spatial Phylogenetic Reconstruction of Evolutionary Dynamics software package (Bielejec *et al.*, 2011) and visualise the spatial projections using Cartographica (www.macgis.com). We also used GenGIS (Parks *et al.*, 2013) to visualise the MCC tree as a tanglegram in a map adapted from Natural Earth (www.naturalearthdata.com).

In order to test the contribution of various predictors to the patterns of RYMV spread, we adopted a recent GLM extension of discrete phylogeographic diffusion (Lemey *et al.*, 2014). This approach models diffusion rates as a log linear function of a number of explanatory variables, and performs Bayesian model averaging to identify the combination of variables that is predictive of spatial spread while simultaneously reconstructing the phylogeographic history. The support and effect size for each predictor is estimated using inclusion probabilities and GLM coefficients, respectively (Lemey *et al.*, 2014). We considered the following predictors in our GLM-diffusion model: (1) great-circle distances between the centroids of each pair of countries; (2) intensities of rice cultivation by country (area of cultivated rice divided by the total country area (hectares per year)) at two different time points (1960 and 1990, obtained from faostat3.fao.org); (3) spatially disaggregated rice production statistics (area harvested in hectares) around the year 2000, obtained using the Spatial Production Allocation Model (HarvestedChoice, 2011) (Supplementary Fig. S3; since this is expressed in hectares of harvested rice for a 5-arc minute grid cell, we consider this as a measure of host connectivity); (4) precipitation by country in millimetres per year (www.climatemps.com); and (5) sample sizes (number of sequences included per country).

Because we model predictors of diffusion rates between pairs of locations, we include both an 'origin' and 'destination' predictor for location-specific measures such as intensity of rice cultivation, precipitation and sample size. In order to derive pairwise predictor values from the spatially mapped rice production statistics, we employ circuit theory to measure distances on a heterogeneous landscape, with rice area harvested as the heterogeneity factor. Specifically, we use Circuitscape version 3.5 (Shah and McRae, 2008) to compute the distances among pairs of locations in the rice production landscape based on a map of about 320 000 cells that encompasses all 20 sampling regions from East to West Africa. Cells with lower area of rice harvested provide higher resistance than cells with higher rice production. The landscape therefore represents a resistance surface that models small distances between nearby locations that are separated by high rice production and large distances between distant locations that are separated by low rice production. We estimated distances between all pairs of sampling regions and chose to connect cells to their eight neighbouring cells, not only to connect cells to their four cardinal neighbours but also to connect diagonally adjacent cells (Shah and McRae, 2008). All predictors were log transformed and standardised prior to their inclusion in the GLM analyses. We follow Lemey *et al.* (2014) and specify prior inclusion probabilities that put 50% prior probability on no predictor being included, and a normal prior with a mean of 0 and a standard deviation of 2 on the coefficients in log space. Bayes factor (BF) support for predictors was calculated based on the ratio of posterior to prior odds for predictor inclusion.

2.6 Continuous phylogeography

To study the geographic spread of RYMV in continuous space and quantify its tempo of dispersal, we used a phylogenetic Brownian diffusion approach that models the change in coordinates (latitude and longitude) along each branch in the evolutionary history as a bivariate normal random deviate (Lemey *et al.*, 2010). As an alternative to homogeneous Brownian motion, we adopt a relaxed random walk (RRW) extension that models branch-specific variation in dispersal rates similar to uncorrelated relaxed clock approaches (Drummond *et al.*, 2006; Lemey *et al.*, 2010). Specifically, we independently draw branch-specific scalers of the RRW precision from a log-normal distribution to relax the assumption of a constant precision (=1/variance) among branches (Lemey *et al.*, 2010). The original implementation of multivariate diffusion models in BEAST (Lemey *et al.*, 2010) resorted to data augmentation of the

unobserved locations of ancestral nodes in the phylogeny to compute multivariate trait likelihoods. Here, we employ a more recent dynamic-programming approach that integrates over all possible realisations of the unobserved traits (Pybus et al., 2012), and provides a more tractable, efficient, and stable inference for large datasets with considerable diffusion rate heterogeneity.

Bayesian estimates under continuous diffusion models yield a posterior distribution of phylogenetic trees, each having ancestral nodes annotated with location estimates. To quantify the spatial epidemic dynamics, we summarise several statistics from the posterior estimates of the continuous phylogenetic diffusion process, as previously introduced by Pybus et al .(2012). Specifically, we provide mean posterior estimates and 95% highest posterior density (HPD) intervals for: (1) dispersal rate (km/year), summarised as the total great-circle distance traveled across the phylogenetic branches divided by the total time elapsed on the branches; (2) wavefront rate (km/year), summarised as the largest great-circle distance traveled from the root location estimate divided by the time since the MRCA; and (3) diffusion coefficient (km^2/year), which reflects the diffusivity or the area that an infected host explores per time unit. Here, we use a 'weighted average' alternative of the diffusion coefficient ($\hat{D}$) introduced by Pybus et al. (2012) because this has recently been shown to provide estimates with considerably lower variances (Trovão et al., 2015). This statistic is defined as follows (Trovão et al., 2015):

$$\hat{D} = \frac{\sum_{k=1}^{n} g_k^2}{\sum_{i=k}^{n} 4t_k}, \quad (2)$$

where g_k and t_k represent the great-circle distance and time, respectively, along branch $k = 1, \ldots 2N-2$ of the random phylogeny.

3. Results

3.1 Evolutionary rate and divergence time estimation

As a standard check prior to fitting dated-tip molecular clock models, we first explored to what extent our dataset contained visually-detectable signal for sequence divergence throughout the sampling time interval. Despite the fact that previous evolutionary rate estimates for RYMV are fairly consistent (Fargette et al., 2008a), our standard linear regression exploration of root-to-tip distances as a function of sampling time did not reveal clear evidence for temporal signal in the complete dataset (Fig. 1A).

We therefore hypothesised that clusters of more closely related variants may still contain temporal information, but extensive rate heterogeneity along the deeper branches connecting these clusters may confound visual detection of such signal. Indeed, particular clusters in the rooted tree with branch lengths estimated using an unconstrained (non-clock) model (Fig. 1A and Supplementary Fig. S2), have tips that are systematically more divergent from the root than other clusters. To examine the impact of this on root-to-tip divergence as a function of sampling time, we perform a similar analysis based on MRCA-to-tip divergences as a function of sampling time for specific clusters and level-out differences in cluster heights prior to plotting all the divergence data (*cfr*. 'Methods' section and Supplementary Fig. S2). This effectively ignores the rate heterogeneity on the deeper branching and removes the cluster effects on the root-tip-regression (Fig. 1B), resulting in a somewhat more discernible divergence accumulation through time. Together with an improved fit (adjusted R^2 increase from 0 to 0.12), this suggests that the rate variation among the deeper branches can indeed affect the temporal signal estimate for the complete dataset.

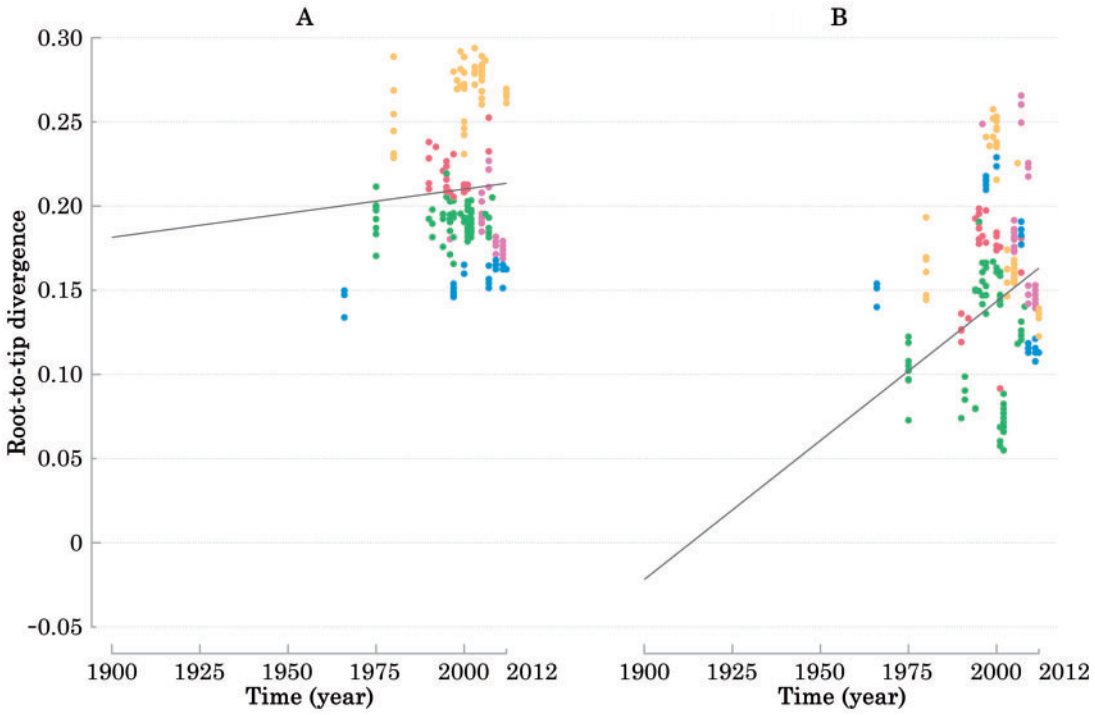

Figure 1. Root-to-tip divergence as a function of sampling time for ML tree clusters (A) and for MCC tree clusters after removing the deep branch effects (B). Colour-coding identifies the 5 different clusters included.

The presence of temporal signal remains however questionable, which urged us to complement this exploration with a date-randomisation test implemented in BEAST (*cfr*. 'Methods' section). Although the rate was estimated to be 1.01×10^{-3} (95%HPD : $7.51 \times 10^{-4} - 1.32 \times 10^{-3}$) nucleotide substitutions per year per site by BEAST using the correct sampling dates, averaging over all possible date randomisations resulted in a far lower estimate of 2.48×10^{-5} ($1.12 \times 10^{-5} - 3.76 \times 10^{-5}$). Given that the 95% HPDs do not overlap for these estimates, we follow Duchêne *et al.* (2015) in considering this as evidence for significant temporal signal in the time-stamped data. We estimated these rates using a relaxed molecular clock model (Drummond *et al.*, 2006), which is better supported by the data than a strict clock model (see Supplementary Table S1). This is perhaps not surprising given the lack of a clear divergence accumulation over the sampling time interval in the exploratory linear regression analyses (Fig. 1A), and the substantial variation of the rate about its mean (coefficient of variation = 0.75; Table 1).

Table 1. Impact of coalescent model on the TMRCA estimate

	Date for the MRCA[a] (year)	Evolutionary rate (μ) (substitutions/site/year)	Coefficient of variation for μ
Skyline	1818	9.71×10^{-4}	0.76
	[1727–1889]	[7.18×10^{-4}–1.25×10^{-3}]	[0.58–0.96]
Skyride	1818	9.39×10^{-4}	0.75
	[1736–1891]	[7.01×10^{-4}–1.21×10^{-3}]	[0.58–0.95]
Skygrid	1852	1.01×10^{-3}	0.78
	[1791–1903]	[7.51×10^{-4}–1.32×10^{-3}]	[0.59–0.99]

Values in between brackets represent 95% HPD intervals.
[a]MRCA, most recent common ancestor.

Because the absence of strong temporal signal may lead to a more pronounced impact of tree priors on divergence time estimates, we estimated TMRCAs using three different flexible nonparametric approaches (Table 1). Whereas the rate and TMRCA estimates under the skyline (Drummond *et al.*, 2005) and skyride (Minin *et al.*, 2008) models are very similar, the skygrid model results in a somewhat higher rate and younger TMRCA estimate (Table 1), but HPDs remain widely overlapping for estimates under the different models. We note that the estimates under the skyride model were sensitive to the way population sizes estimates are smoothed across the evolutionary time scale (*cfr*. 'Methods' section), which is likely due to the lack of strong temporal signal. As previously shown through simulations (Gill *et al.*, 2013), the skygrid model performs the best for divergent time estimates and the substitution rate under this model is also more consistent with previous studies (Fargette *et al.*, 2008a). We therefore use this coalescent tree prior in all further analyses.

3.2 Discrete geography

By mapping the tip locations of a cladogram representation of the MCC tree in geographic space (Fig. 2), we highlight a clear separation of East and West African RYMV diversity. The modal

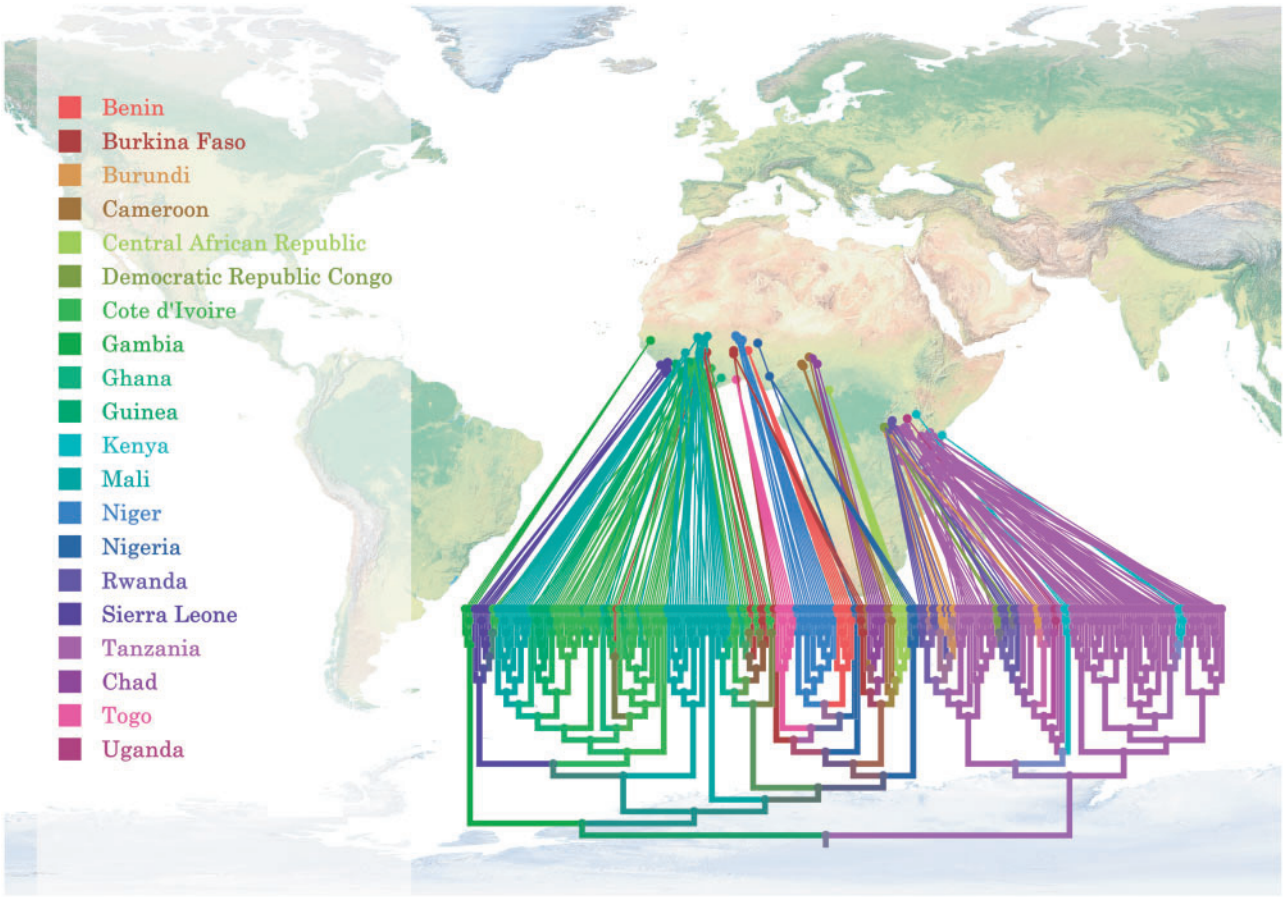

Figure 2. Tanglegram representation of the RYMV history by mapping the tip locations of the MCC cladogram to the geographic location of sampling. Branches are coloured according to the modal location state estimates obtained by a discrete phylogeographic reconstruction under a reversible CTMC model with BSSVS.

location state estimates obtained by discrete phylogeographic reconstruction (represented by branch colours in the tanglegram and in the equivalent time-measured tree in Fig. 3), also reveal a strongly spatially structured viral population. We quantified this through the degree of phylogenetic clustering by location as summarised using the AI (Wang *et al.*, 2001), and found a low AI of 0.109 (0.08–0.14) indicating that the degree of spatial structuring is not so far from absolute (AI = 0).

To infer the discrete ancestral location states, we applied both reversible and non-reversible discrete diffusion models with and without a BSSVS procedure (Lemey *et al.*, 2009; Edwards *et al.*, 2011) and compared model fit for the four combinations using (log) marginal likelihood estimation (Baele *et al.*, 2012, 2013; Baele and Lemey, 2013). Although we report the results for the best fitting model (reversible with BSSVS, see Supplementary Table S2), we note that the ancestral reconstructions are robust with respect to diffusion model specification. For example, all four model combinations find support for Tanzania as the geographical origin of RYMV (Fig. 3B), and this remained the best supported root location when the Tanzanian strains were downsampled to the same number as for Côte d'Ivoire (see Supplementary Table S3). The support for Tanzania emerges from a relatively high RYMV diversity in this country, encompassing most of the diversity in the East Africa clade (lineage S4, S5, and S6 in Fig. 3A), and hence a strong support for this location state at ancestral nodes up to the root node. West Africa (lineage S1–S3) was seeded relatively early in the RYMV history as its MRCA dates back to 1887 (1840–1919) and was estimated to have originated from Côte d'Ivoire.

Using BSSVS (Lemey *et al.*, 2009), we quantify the support for different diffusion pathways under the form of BF support for non-zero rates. Not surprising, we find support for a separate East and West African diffusion network (Fig. 4). We complement the support for the rates by estimating the number of transitions that occurred between the states involved using Markov jump counting (Minin and Suchard, 2008a) (Fig. 4). In East Africa, we find support for diffusion out of Tanzania to Kenya, Uganda, and Rwanda, and from the latter country also to Burundi and the Democratic Republic of Congo (DRC). The western diffusion network involves more countries and is characterised by a high degree of seeding from Côte d'Ivoire with diffusion pathways that extend eventually to Central Africa (the Central African Republic). Taken together, the diffusion pathways in East and West Africa we display in Figure 4 account for 84% of the location state transitions recovered in the RYMV evolutionary history.

To investigate what process of RYMV spread has led to the spatial genetic patterns we describe here, we apply a recent

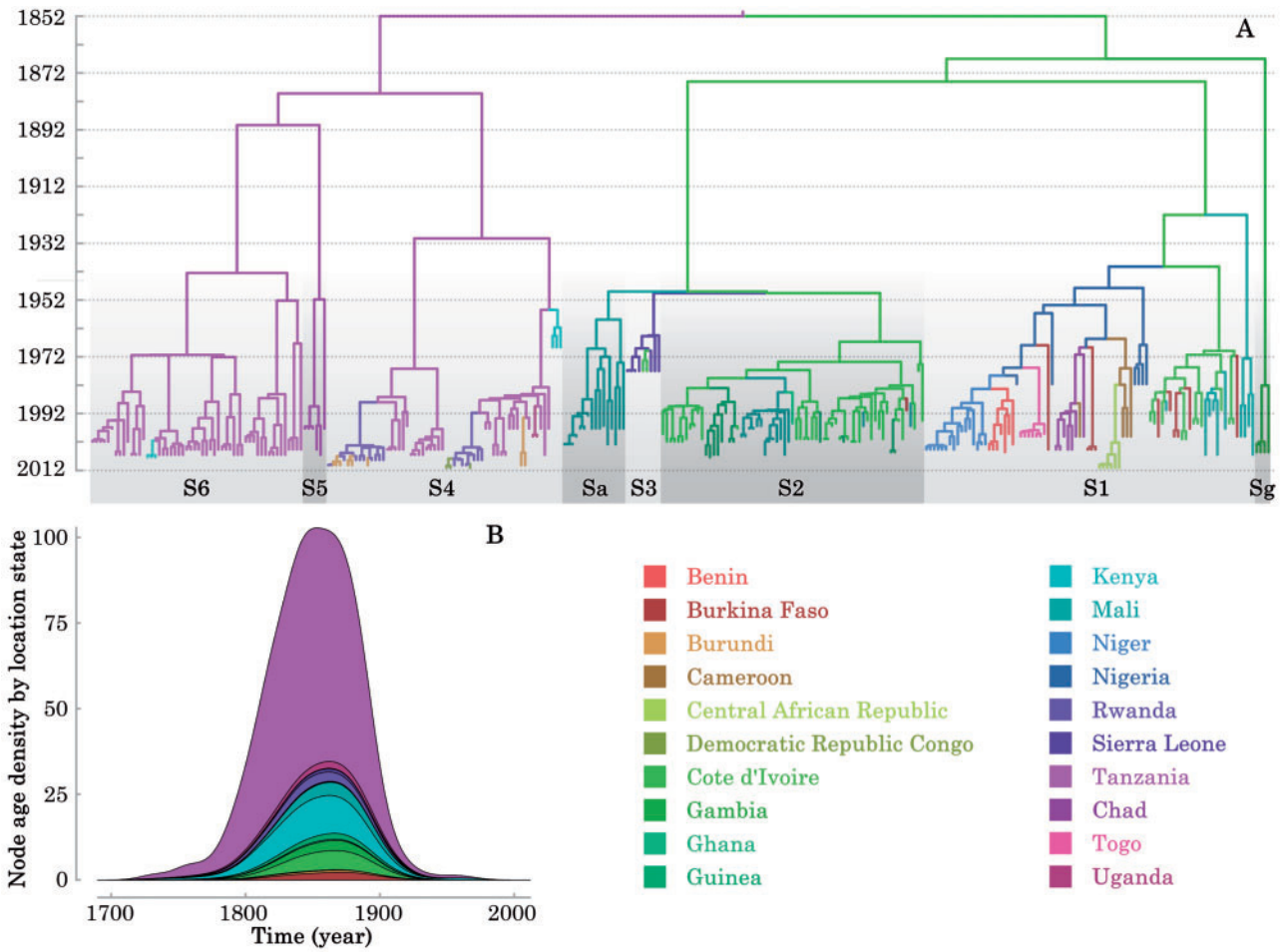

Figure 3. Time-calibrated MCC tree inferred for 300 CP sequences of RYMV. Branches are coloured according to the most probable location state, indicated in the coloured legend (A). Posterior probability densities of the root location state for discrete reversible model with BSSVS; 66.87% of the posterior mass for the root location supports Tanzania as the origin location of RYMV epidemics (B). Early separation of the East-West epidemics in RYMV history, with 87 and 43% of posterior mass for the West (green) and East (violet) lineage clades, respectively.

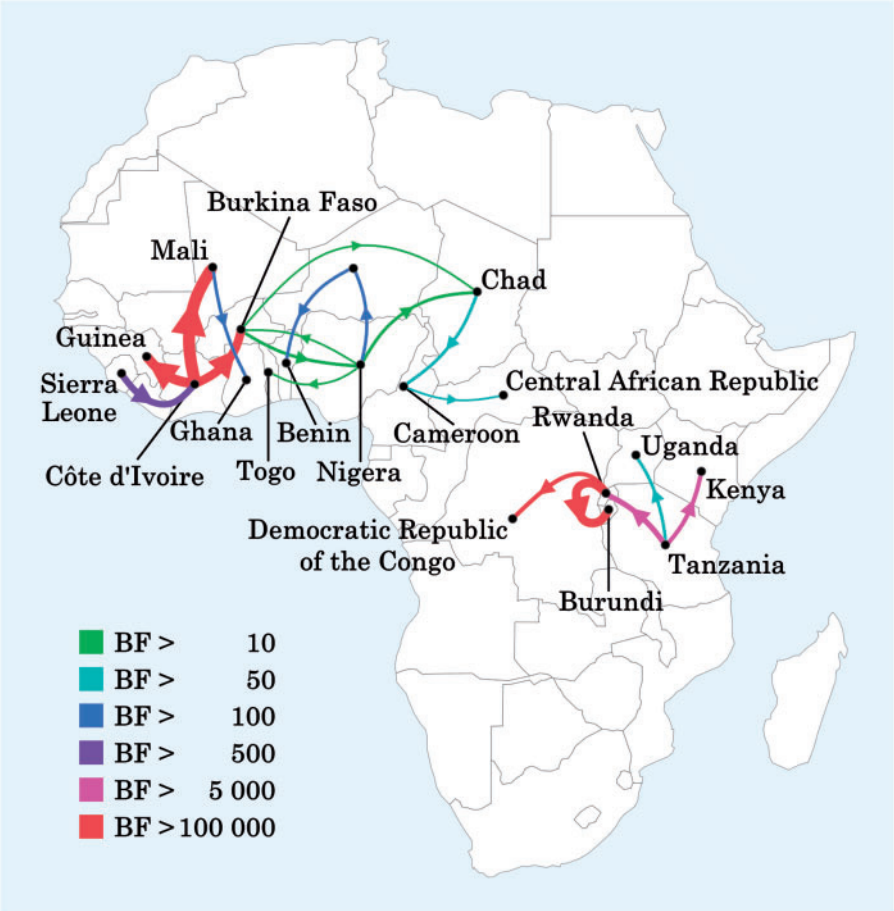

Figure 4. BF test support for discrete diffusion rates. Rates supported by a BF > 4 are indicated. The line colour represents the relative strength by which the rates are supported: green lines and red lines suggest relatively weak and strong support, respectively. The thickness of the arrows indicates increasing number of Markov jumps between locations.

GLM extension of the discrete phylogeographic diffusion model that allows to test different potential predictors of viral dispersal (Faria *et al.*, 2013; Lemey *et al.*, 2014). We consider geographic distances, distances measured on a resistance landscape of harvested rice area (*cfr.* 'Methods' section and Supplementary Fig. S3), location-specific rice intensities at two different time points and precipitation as potential explanatory variables of the patterns of spread (Fig. 5). To examine whether the predictor support is robust to sample size heterogeneity we also include sample sizes as an explanatory variable. The GLM procedure, which attempts to identify the linear combination of predictors of spatial diffusion while reconstructing the phylogeographic history, finds maximal support for distances measured on a landscape of harvested rice area as a predictor of RYMV dispersal (Fig. 5). This predictor has a negative log effect size implying an inverse relationship with transition rates. That is, a high distance in the resistance landscape, as reflected by a large geographic distance and/or low harvested rice area between these locations, correlates with less intense viral dispersal. The importance of harvested rice intensity is also reinforced by a relative modest additional support for origin rice intensity in 1990 (BF =14.2), which is accompanied by a positive log effect size, suggesting higher viral dispersal out of locations with higher rice intensity. In addition to a host connectivity component, the distances in the harvested rice area landscape also incorporate geography. The fact that both are important is demonstrated by a GLM analysis that excludes the landscape distance predictor, which results in clear support for distance as well as origin rice intensity in 1990 (Supplementary Table S4). No other predictor yielded noticeable support in our analyses, and remarkably, also sample sizes did not help to explain viral diffusion intensities (Fig. 5). By repeating the analysis separately on the East and West african clade, we demonstrate that the signal for predictor support can be entirely attributed to the more pronounced and dynamic West African spread. Whereas highly similar predictor support and effect sizes are obtained for West Africa, none of the predictors yield noticeable support in East Africa (Supplementary Fig. S4).

3.3 Continuous phylogeography

In order to quantify the dynamics of RYMV spread in continuous space, we also applied multivariate phylogenetic diffusion models to the CP sequences and their geographic coordinates (Lemey *et al.*, 2010). Because of the clear separation between the spread of East and West African RYMV lineages (Figs. 2 and 4), we perform separate analyses on the data for both regions. We tested a model of strict Brownian diffusion against several versions of RRW models using marginal likelihood estimation (Baele *et al.*, 2012, 2013), and found that a lognormal-RRW provided the best fit to the dispersal dynamics (Supplementary Table S5).

The spatiotemporal patterns of spread under this model are summarised in Figure 6. In agreement with the discrete

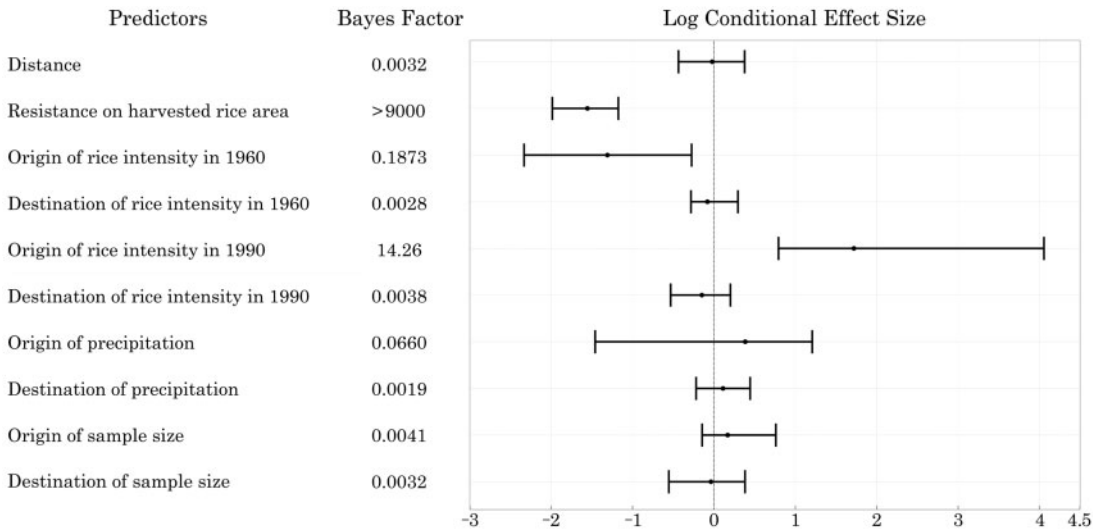

Figure 5. Predictors of RYMV dispersal across Africa. For each potential predictor, the BF support and the conditional effect size obtained using the GLM diffusion approach implemented in BEAST are shown (posterior mean and 95% Bayesian CI). Note that the credibility intervals for the cES of the predictors with BF > 14 exclude zero, which can be considered as additional evidence for its importance.

phylogeographic results, the eastern root location is estimated in Tanzania, and from the early 1930s, the virus spreads in the direction of Kenya and Uganda. The viral expansion continues within these countries and finally, by 2012, the spread of RYMV also includes Burundi, Rwanda and the DRC (Ndikumana *et al.*, 2011; Hubert *et al.*, 2013). In West Africa, the credible contour for the origin location overlaps mostly with Côte d'Ivoire, Mali and Senegal from where the virus spreads to the south, west and east. By 1932, the eastward expansion includes Nigeria and continues towards Chad. Extensive diffusion dynamics further develop within and between Côte d'Ivoire and Mali, and also western locations including Sierra Leone and Guinea appear to be seeded from these countries, respectively. Recent years mark the arrival in the most eastern location (Central African Republic) (Longué *et al.*, 2013).

Based on quantitative summaries of the eastern and western dynamics (listed in Table 2), both datasets are characterised by similar dispersal rates and diffusion coefficients with overlapping HPDs. In terms of wavefront dynamics; however, we observe a slower invasion rate in the East as compared with its western counterpart. In line with different invasion rates on similar time-scales, our phylogeographic reconstruction estimated East and West wavefront distances of ~1025 (95% HPD: 743–1,239) km and 2,869 (2,370–3,356) km, respectively. By plotting how these wavefront distances evolved over time (Fig. 7), we show that from the mid-1800s viral expansion begins in both geographical regions, but whereas it levels off at around 1960 in the east, RYMV continued to expand its spread in West Africa. Similar continuous diffusion statistics for the East and West Africa dynamics (Table 2) indicate that these account for a considerable degree of heterogeneity in the spatial spread dynamics. For instance, the dispersal rate and diffusivity are similar in both East and West Africa whereas the wavefront rate is three times lower in East than in West Africa, possibly because of the numerous barriers to spread in East Africa, whereas the Niger-Bénoué river axis in West Africa may have been an efficient means of viral propagation. We note that RYMV, MSV, and EACMV are characterised by dispersal statistics that are generally within the same order of magnitude, although some statistics suggest more pronounced dynamics for the latter two. This might be explained by transmission through leafhoppers for MSV and whiteflies for EACMV, but also human-mediated dispersal through infected cuttings.

4 Discussion and conclusion

As the main viral disease of rice, RYMV has been reported in all major rice producing countries in sub-Saharan Africa. Yielding losses up to 100%, it represents one of the tropical plant emergent diseases with the highest socio-economical impact (Fargette *et al.*, 2006). Evolutionary studies have only relatively recently characterised RYMV as a rapidly evolving plant virus based on heterochronous sequence data. In this study, we expand on the work of Fargette *et al.* (2008a,b) and Abubakar *et al.* (2003) by reconstructing the RYMV phylogeographic history in both discrete (Lemey *et al.*, 2009) and continuous (Lemey *et al.*, 2010) space using Bayesian inference, and specifically test and quantify a range of potential predictors of spatial spread (Lemey *et al.*, 2014).

Although our RYMV evolutionary rate estimate is consistent with previous studies (Fargette *et al.*, 2008a), it remains difficult to clearly detect accumulation of sequence divergence over the sampling time interval in the currently available data. We note that such temporal signal depends on both the evolutionary process and how we are able to sample from this process. A high overall tempo of evolution and a constant pattern of substitution accumulation will both increase the probability of measurable evolution over a particular time interval. A large temporal spread in sampling dates and more homogenous sampling throughout this interval will further contribute to the temporal signal (Seo *et al.*, 2002). An average RYMV substitution rate of about 0.001 substitutions per site per year and a sampling time interval of 44 years—even if sampling is more dense towards the present as is generally the case—may provide a relatively good opportunity to detect temporal signal in the CP gene. Substitution rate variability, however, appears to be a major confounding factor for RYMV, in particular because this may have acted on a relatively long evolutionary time scale of about 160 years. This is apparent through the consistently higher or lower tip divergences from particular clusters in the RYMV

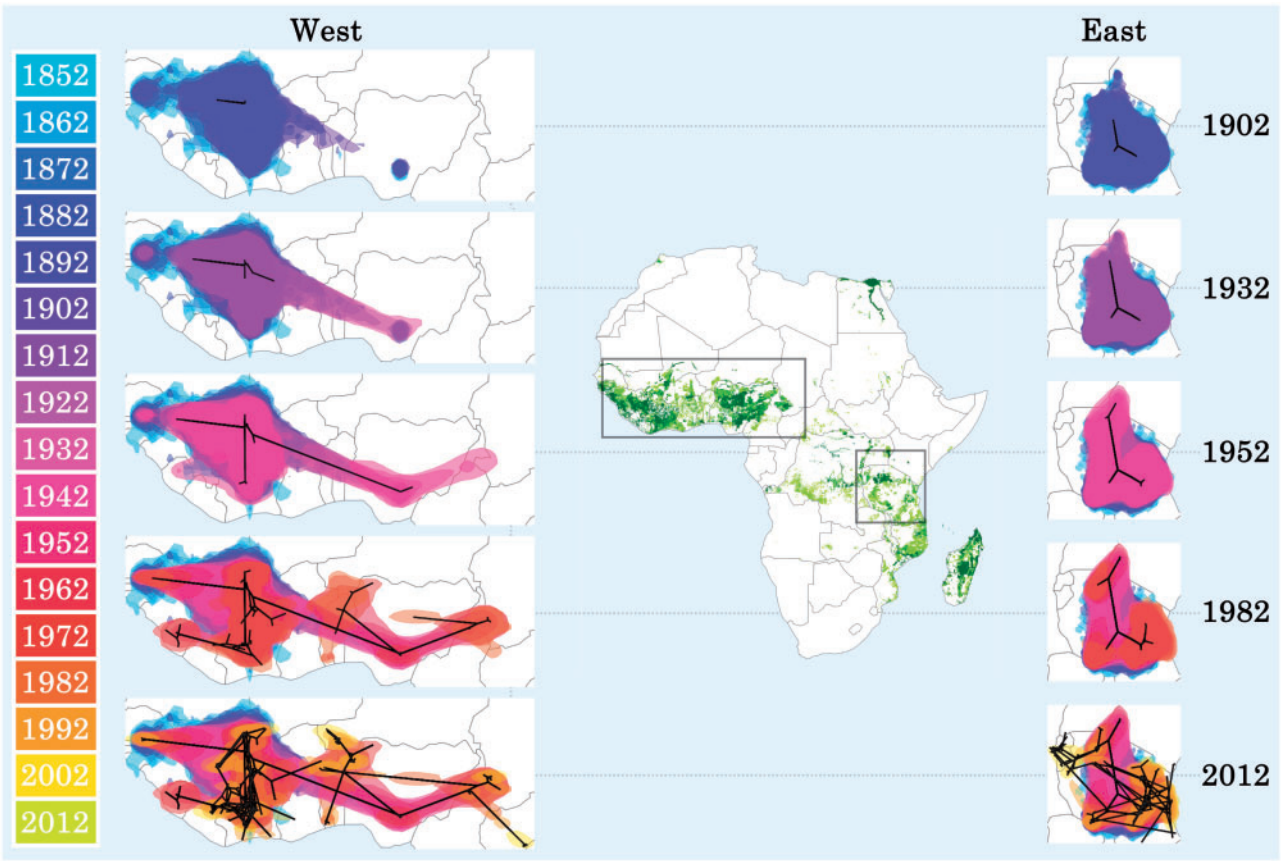

Figure 6. Reconstruction of the continuous spatiotemporal dispersal of RYMV in West and East Africa, shown from 1852 to 2012 at intervals that capture the major dispersal events. Black lines show a spatial projection of the representative phylogeny. Coloured clouds represent statistical uncertainty in the estimated locations of RYMV internal nodes (95% HPD intervals).

Table 2. Dispersal and ecological parameters

	Dispersal rate (km/year)	Diffusivity (km^2/year)	Wavefront rate (km/year)
West	13.08 [10.25–16.03]	1559.25 [1186.13–1976.88]	23.13 [14.33–31.85]
East	16.16 [11.55–21.19]	1595.53 [1106.99–2151.71]	7.51 [3.80–12.32]
MSV	33.17 [28.70–7.82]	11665.18 [9133.97–14979.72]	74.79 [45.25–109.36]
EACMV	13.30 [8.07–0.65]	3748.74 [2196.90–5890.32]	32.45 [15.67–56.12]

Values in between brackets represent 95% HPD intervals.

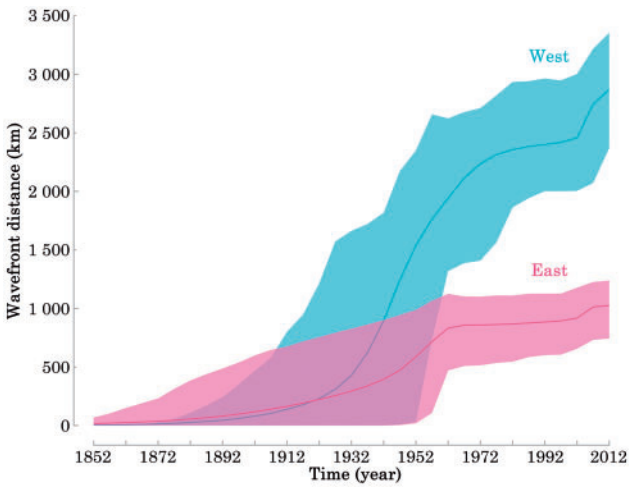

Figure 7. Mean wavefront distances for the West (blue) and East (pink) epidemics. Mean values indicated by darker lines and 95% HPD intervals indicated by coloured shadows.

phylogeny (Fig. 1). Several factors may be responsible for rate heterogeneity in the phylogenetic history, including variation in mutation rate and replication rate as well as variation in selective pressure and host population sizes, but it remains difficult to disentangle these factors for RYMV. In general, intrinsic differences in mutation rate and replication dynamics are more likely to act between more distantly related viruses, such as different viral families. For more closely related viruses, host factors have been shown to impact viral evolutionary rates (Streicker *et al.*, 2012; Worobey, Han, and Rambaut, 2014). This together with the dynamics that impact the fixation of substitutions, represent interesting subjects for further RYMV research.

Even if exploratory linear regression plots do not suggest clear temporal signal, tip calibration may still prove useful if rate variation among branches is satisfactorily accommodated (Firth *et al.*, 2010). Relaxed molecular clocks may indeed perform reasonably well in modelling rate along the relatively long branches that separate distinct RYMV clusters. However, a test is needed to assure that tip calibrations will lead to

meaningful estimates. For this purpose, a date-randomisation procedure has been proposed that tests whether the real rate estimate deviates from rates obtained in the absence of temporal structure in the tip-calibrations (Ramsden *et al.*, 2008). Here, we provide a convenient BEAST implementation of this test that does not require multiple independent randomisations, but averages over all plausible randomisations during the rate estimation process. Based on a relatively stringent test criterion (Duchêne *et al.*, 2015), we still find a significant association between substitutions and time in our RYMV sequence data.

Although significant, the relative weakness of the temporal signal explains the sensitivity to the coalescent prior we observed in our analyses. Our TMRCA estimate using the skygrid model was somewhat more recent than that obtained under the skyline and skyride model, but more importantly, different settings in the skyride model strongly impacted the evolutionary time-scale. This provides a general warning that weak temporal signal may not simply be reflected in uncertainty of date and/or rate estimation in a Bayesian coalescent framework, but coalescent priors may also affect mean TMRCA estimates. In our study, we highlight the skygrid TMRCA estimate of 1852 [1791–1903] because this model has been shown to outperform the other flexible coalescent priors for divergence time estimation (Gill *et al.*, 2013). We acknowledge that there is a limit to TMRCA estimation for rapidly evolving viruses in general because saturation and strong purifying selection can lead to an underestimation of old viral origins (Wertheim and Kosakovsky Pond, 2011). However, a TMRCA of about 160 years may be well below the time-scale on which saturation becomes truly important (Bielejec *et al.*, 2014).

Our study represents a natural extension of earlier descriptions of RYMV spatial genetics (e.g. Abubakar *et al.*, 2003). Using statistical reconstructions of discrete phylogeographic diffusion, we confirm a clear east-west separation, a strong spatial structure in general, and a likely origin in Tanzania. In this region, and maybe in the Eastern Arc Mountains biodiversity hotspot in particular, RYMV may have emerged in cultivated rice from an ancestor infecting wild graminaceous (e.g. perennial wild rice species such as *O. longistaminata*) before spreading to other parts of Africa. However, currently identified RYMV isolates in perennial wild hosts appear to be spill-over events from cultivated rice (Traore *et al.*, 2005), and no related viruses have been identified remote from rice crop areas. In line with an estimated origin in Tanzania, recent analyses of complete genome data have shown that the West and Central African RYMV diversity is nested as a monophyletic clade within the Tanzania diversity (Ochola *et al.*, 2015). Although we did not consider this in our analysis, we note that a Tanzanian origin also appears to be supported by three coding insertion-deletion polymorphisms at two positions of the CP gene (amino acid position 18 and 60). The three forms are distributed over different clades, but they are only found together in Eastern Tanzania. In the remaining parts of East Africa and in West Africa, only one of the three forms has been identified, with S5 strain (Eastern Arc Mountain in Eastern Tanzania) with both K19 and R60, S6 strain (Eastern Arc Mountain in Eastern Tanzania) with R60 and a deletion of codon K19, and strains S1, S2, S3, Sa, Sg (West Africa), and S4 (Eastern Arc Mountain in Eastern Tanzania) have K19 and a deletion of codon R60. Following its emergence in East Africa, we can only speculate on how the virus was introduced relatively early in West Africa. The long-distance movement event appears to be unique to the natural history of RYMV and could have occurred through human trading practices (Carpenter, 1978).

Phylogeographic analyses allow the description of the spatiotemporal patterns of viral spread, which in turn may lead to the formulation of hypotheses about the underlying processes that shape the dynamics of spread. Agricultural intensification and extensification are considered to strongly facilitate the establishment and epidemic spread of emerging viruses (Thresh, 1982; Elena, 2011), and this has also been invoked as a potential driver of RYMV expansion by molecular epidemiology (Konaté and Fargette, 2001; Abubakar *et al.*, 2003). Specifically, the increasing adoption of new production modes such as water-fed rice farming, annual double cropping and high-yielding Asian varieties highly susceptible to RYMV are likely to have contributed to its spread (Fargette *et al.*, 2006; Konaté and Fargette, 2001). By identifying the well supported rates of diffusion, we delineated the major RYMV pathways of spread, but until very recently it has remained challenging to formally test the drivers of spatial spread. We address this here using a GLM extension of the discrete phylogeographic model (Faria *et al.*, 2013; Lemey *et al.*, 2014), which aims at determining which subset of explanatory variables helps to explain the relative intensities of viral dispersal among pairs of locations. Agricultural intensification leads to higher rice densities and harvest, which we incorporated as a predictor in our analyses. To this purpose, we used circuit theory to build a resistance landscape with the harvested area of rice in 2000 as resistance factor. This strongly predicted the patterns of RYMV spread, and both the geographic and host ecology component of the distances measured on the resistance landscape appeared to be important. The inclusion of origin rice intensity in 1990 as an additional predictor points at a degree of asymmetry in spread facilitated by rice connectivity, with stronger effect on the dispersal out of areas with high rice intensity, which seems in line with spread facilitated by agricultural extensification. The fact that a more recent measure of rice intensity (1990 vs. 1960) provides better explanatory power may be related to the higher branch density in the recent evolutionary history covering more dispersal events around that time. To our knowledge, our study is the first to formally demonstrate the role of host ecology in plant virus spread using genetic data. It is interesting to note that a historical approach based on a historical map of rice distribution Portères (1950, 1957, 1962) (Pinel-Galzi *et al.*, 2015), and our statistical approach using a spatiotemporal reconstruction incorporating present day rice statistics, converged towards the same conclusion that harvested rice intensity or connectivity is the main determinant of RYMV emergence and spread.

Earlier RYMV phylogeographic analyses have established an isolation-by-distance pattern for RYMV (Abubakar *et al.*, 2003), and spread as a function of geographic distance was also evident from our phylogeographic test approach. This motivated the complementary application of phylogeographic reconstruction in continuous space (Lemey *et al.*, 2010) allowing us to quantify the tempo and mode of spread using several spatial summary statistics (Pybus *et al.*, 2012). Given the clear distinction of East and West African RYMV lineages, we compared separate estimates from both regions. Considerable differences exist in terms of climate, ecology and host range between these regions, with East Africa growing only the Asiatic rice *O. sativa* whereas both the African rice *O. glaberrima*, which is genetically quite different from *O. sativa*, and the Asiatic rice are cultivated in West Africa. Despite such differences, we found that the overall rate of RYMV spread and diffusivity was highly similar. These measures do however not take into account the directionality of spread, and when the directionality is considered to be the distance from the estimated origin in both regions (the wavefront distance), we

find higher wavefront velocities in the West. So, it seems that rice densities, which are more pronounced in the West, do not necessarily increase the overall rate of spread, but they facilitate more extensive expansion dynamics. By summarising these expansion dynamics over time (Fig. 7), we revealed that agricultural intensification and extensification had a more prominent impact in the West, which is in accordance with the fact that our GLM-diffusion estimates were essentially informed by the data from West Africa. The major expansion dynamics in the west are characterised by spread from Côte d'Ivoire or Mali in an eastern direction, towards countries that have relatively lower rice production like Niger, Chad, or Central African Republic, in line with the asymmetry suggested by the origin rice intensity predictor in the GLM analysis. In Mali, various lineages have been identified in the Inner Niger Delta, suggesting that this specific area may have been the West African centre of diversification (Traore *et al.*, 2005; Fargette *et al.*, 2006). The propagation towards Central Africa is likely to have followed the more accessible routes of transmission, in particular along the Niger-Bénoué rivers. In East Africa, the comparatively sparser and less intense rice production, along with physical (mountains) and ecological (tropical forests) boundaries have restricted viral expansion. The continuous phylogeographic reconstruction shows only recent spread out of its likely area of origin. However, further spread may continue in the future and molecular surveillance will be needed to track these dynamics.

Several aspects of our phylogeographic analyses may be further improved or fine-tuned in the future. Longer genome regions offer more phylogenetic resolution and are likely to increase the temporal signal, but they may also require taking into account recombination (Pinel-Galzi *et al.*, 2009). Sampling biases represent an important challenge for ancestral reconstructions, and we acknowledge that such biases may also burden the sample we analysed, even though the GLM analysis did not associate sampling numbers with diffusion intensities. Structured coalescent approaches are expected to be less sensitive to sampling biases and represent interesting alternatives for discrete phylogeographic reconstructions (De Maio *et al.*, 2015). Furthermore, it may prove interesting to expand on the predictors of RYMV dispersal if systematic data would be available, including for example, on vector demographics, rice cultivar resistance to RYMV, mode of watering and other agricultural practices. Despite these areas of potential improvement, our current analysis takes an important step towards hypothesis testing in plant virus epidemiology and ecology. Over the last decade, RYMV has become the main threat to rice cultivation in Africa and Madagascar (Konaté and Fargette, 2001). The finding that host ecology has shaped RYMV spread suggests predictable patterns of spread that may help to inform predictive models for RYMV control and public health policies. More generally, our results reinforce the concept that host population ecology is crucial for the onward transmission and epidemic potential of any emerging virus (Woolhouse and Gowtage-Sequeria, 2005).

Supplementary data

Supplementary data are available at *Virus Evolution* online.

Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme [FP7/2007-2013] under Grant Agreement no. 278433-PREDEMICS and ERC Grant Agreement no. 260864, the National Institutes of Health under R01 AI107034 and R01 HG006139, the National Science Foundation under DMS 1264153, and the Agropolis Foundation project 'Bioagressors and Invasive Species: from Individual to Population to Species (BIOFIS)'.

Conflict of interest: None declared.

References

Abo, M., Sy, A., and Alegbejo, M. (1998) 'Rice Yellow Mottle Virus in Africa: Evolution, Distribution, Economic Significance and Sustainable Rice Production and Management Strategies', *Journal of Sustainable Agriculture*, 11: 85–111.

Abubakar, Z. et al. (2003) 'Phylogeography of Rice Yellow Mottle Virus in Africa', *Journal of General Virology*, 84: 733–43.

Allarangaye, M. et al. (2006) 'Evidence of Non-Transmission of Rice Yellow Mottle Virus Through Seeds of Wild Host Species', *Journal of Plant Pathology*, 83: 309–15

Ayres, D. L., et al. (2012) 'Beagle: An Application Programming Interface and High-Performance Computing Library for Statistical Phylogenetics', *Systems Biology*, 61: 170–3.

Baele, G., and Lemey, P. (2013) 'Bayesian Evolutionary Model Testing in the Phylogenomics Era: Matching Model Complexity with Computational Efficiency', *Bioinformatics*, 29: 1970–9.

—— et al. (2012) 'Improving the Accuracy of Demographic and Molecular Clock Model Comparison While Accommodating Phylogenetic Uncertainty', *Molecular Biology and Evolution*, 29: 2157–67.

——, Lemey, P., and Vansteelandt, S. (2013) 'Make the Most of Your Samples: Bayes Factor Estimators for High-Dimensional Models of Sequence Evolution', *BMC Bioinformatics*, 14: 85.

Bakker, W. et al. (1974) *Characterization and Ecological Aspects of Rice Yellow Mottle Virus in Kenya*. Wageningen: Centre for Agricultural Publishing and Documentation.

Bielejec, F. et al. (2011) 'Spread: Spatial Phylogenetic Reconstruction Of Evolutionary Dynamics', *Bioinformatics*, 27: 2910–2.

—— et al. (2014) 'piBUSS: A Parallel BEAST/BEAGLE Utility for Sequence Simulation Under Complex Evolutionary Scenarios', *BMC Bioinformatics*, 15: 133.

Carpenter, A. J. (1978) *'The History of Rice in Africa'*, in I. W., Buddenhagen, and G., Persley (eds.) *Rice in Africa*, pp. 5–10. London: Academic Press.

De Bruyn, A. et al. (2012) 'East African Cassava Mosaic-Like Viruses from Africa to Indian Ocean Islands: Molecular Diversity, Evolutionary History and Geographical Dissemination of a Bipartite Begomovirus', *BMC Evolutionary Biology*, 12: 228.

De Maio, N., et al. (2015) 'New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation', *PLoS Genetics*, 11: e1005421.

Drummond, A. J. et al. (2005) 'Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences', *Molecular Biology and Evolution*, 22: 1185–92.

—— et al. (2006) 'Relaxed Phylogenetics and Dating with Confidence', *PLoS Biology*, 4: e88.

—— et al. (2012) 'Bayesian Phylogenetics with BEAUti and the BEAST 1.7', *Molecular Biology and Evolution*, 29: 1969–73.

Duchêne, S. et al. (2015) 'The Performance of the Date-Randomization Test in Phylogenetic Analyses of Time-Structured Virus Data', *Molecular Biology and Evolution*, 32: 1895–906.

Duffy, S., and Holmes, E. C. (2008) 'Phylogenetic Evidence for Rapid Rates of Molecular Evolution in the Single-Stranded DNA Begomovirus Tomato Yellow Leaf Curl Virus', *Journal of Virology*, 82: 957–65.

——, and —— (2009) 'Validation of High Rates of Nucleotide Substitution in Geminiviruses: Phylogenetic Evidence from East African Cassava Mosaic Viruses. *Journal of General Virology*, 90: 1539–47.

Edwards, C. J. et al. (2011) 'Ancient Hybridization and An Irish Origin for the Modern Polar Bear Matriline', *Current Biology*, 21: 1251–8.

Elena, S. F. (2011) Evolutionary Constraints on Emergence of Plant RNA Viruses', in C., Caranta, M. A., Aranda, M., Tepfer, and J. J., Lopez-Moya (eds.) *Recent Advances in Plant Virology*. Norfolk, UK: Caister Academic Press.

Fargette, D. et al. (2004) 'Inferring the Evolutionary History of Rice Yellow Mottle Virus from Genomic, Phylogenetic, and Phylogeographic Studies', *Journal of virology*, 78: 3252–61.

—— et al. (2006) 'Molecular Ecology and Emergence of Tropical Plant Viruses', *Annual Review of Phytopathology*, 44: 235–60.

—— et al. (2008a) 'Diversification of Rice Yellow Mottle Virus and Related Viruses Spans the History of Agriculture From the Neolithic to the Present', *PLoS Pathogens*, 4: e1000125.

—— et al. (2008b) 'Rice Yellow Mottle Virus, An RNA Plant Virus, Evolves as Rapidly as most RNA Animal Viruses', *Journal of Virology*, 82: 3584–9.

Faria, N. R. et al. (2013) 'Simultaneously Reconstructing Viral Cross-Species Transmission History and Identifying the Underlying Constraints', *Philosphical Transactions of the Royal Society of London B Biological Sciences*, 368: 20120196.

Firth, C. et al. (2010) 'Using Time-Structured Data to Estimate Evolutionary Rates of Double-Stranded DNA Viruses', *Molecular Biology and Evolution*, 27: 2038–51.

Fraile, A. et al. (1997) 'A Century of Tobamovirus Evolution in an Australian Population of Nicotiana Glauca', *Journal of Virology*, 71: 8316–20.

Gibbs, A. J. et al. (2010) 'Time–The Emerging Dimension of Plant Virus Studies', *Journal of General Virology*, 91: 13–22.

Gill, M. S. et al. (2013) 'Improving Bayesian Population Dynamics Inference: A Coalescent-Based Model for Multiple Loci', *Molecular Biology and Evolution*, 30: 713–24.

Guindon, S. et al. (2010) 'New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0', *Systems Biology*, 59: 307–21.

HarvestedChoice. (2011) 'Rice Area Harvested (ha) (2000)', International Food Policy Research Institute, Washington, DC., and University of Minnesota, St. Paul, MN. <http://harvestchoice.org/node/4799>.

Hasegawa, M., Kishino, H., and Yano, T. (1985) 'Dating of the Human-Ape Splitting by a Molecular Clock of Mitochondrial DNA', *Journal of Molecular Evolution*, 22: 160–74.

Hubert, J. G. et al. (2013) 'First Report of Rice Yellow Mottle Virus on Rice in the Democratic Republic of Congo', *Plant Disease*, 97: 162.

Katoh, K., and Toh, H. (2008) 'Recent Developments in the MAFFT Multiple Sequence Alignment Program', *Brief Bioinformatics*, 9: 286–98.

Konaté, G., and Fargette, D. (2001) 'Overview of Rice Yellow Mottle Virus', in Hughes, J. d'A., and Odu, B. O. (eds.) *Plant Virology in Sub-Saharan Africa, Proceedings of a Conference Organized by IITA*, pp. 1–17, http://www.iita.org/plant-virology.

——, Sarra, S., and Traore, O. (2001) 'Rice Yellow Mottle Virus is Seed-Borne but not Seed Transmitted in Rice Seeds', *European Journal of Plant Pathology*, 107: 361–64.

Lartillot, N., and Philippe, H. (2006) 'Computing Bayes Factors Using Thermodynamic Integration', *Systems Biology*, 55: 195–207.

Lefeuvre, P. et al. (2010) 'The Spread of Tomato Yellow Leaf Curl Virus from the Middle East to the World', *PLoS Pathogens*, 6: e1001164.

Lemey, P. et al. (2009) 'Bayesian Phylogeography Finds Its Roots', *PLoS Computational Biology*, 5: e1000520.

—— et al. (2010) 'Phylogeography Takes a Relaxed Random Walk in Continuous Space and Time', *Molecular Biology and Evolution*, 27: 1877–85.

—— et al. (2014) 'Unifying Viral Genetics and Human Transportation Data to Predict the Global Transmission Dynamics of Human Influenza H3N2', *PLoS Pathogens*, 10: e1003932.

Ling, R. et al. (2013) An essential fifth coding ORF in the sobemoviruses. *Virology*, 446: 397–408.

Longué, D. R. S. et al. (2013) 'First Report of Rice Yellow Mottle Virus in Rice in the Central African Republic', *Plant Disease*, 98: 162.

Minh, B. Q., Klaere, S., and von Haeseler, A. (2006) 'Phylogenetic Diversity Within Seconds', *Systems Biology*, 55: 769–73.

——, ——, and —— (2009) 'Taxon Selection Under Split Diversity', *Systems Biology*, 58: 586–94.

Minin, V. N., and Suchard, M. A. (2008a) 'Counting Labeled Transitions in Continuous-Time Markov Models of Evolution', *Journal of Mathematical Biology*, 56: 391–412.

——, and —— (2008b) 'Fast, Accurate and Simulation-Free Stochastic Mapping', *Philosphical Transactions of the Royal Society of London B Biological Sciences*, 363: 3985–95.

——, Bloomquist, E. W., and Suchard, M. A. (2008) 'Smooth Skyride Through a Rough Skyline: Bayesian Coalescent-Based Inference of Population Dynamics', *Molecular Biology and Evolution*, 25: 1459–71.

Monjane, A. L. et al. (2011) 'Reconstructing the History of Maize Streak Virus Strain A Dispersal to Reveal Diversification Hot Spots and Its Origin in Southern Africa', *Journal of Virology*, 85: 9623–36.

Ndikumana, I. et al. (2011) 'Rice Yellow Mottle Virus in Rice in Rwanda: First Report and Evidence of Strain Circulation', *New Disease Reports*, 23.

Nelson, M. I. et al. (2015) 'Global Migration of Influenza a Viruses in Swine', *Nature Communications*, 6: 6696.

Ochola, D. et al. (2015) 'Emergence of Rice Yellow Mottle Virus in Eastern Uganda: Recent and Singular Interplay Between Strains in East Africa and in Madagascar', *Virus Research*, 195: 64–72.

Pagán, I., and Holmes, E. C. (2010) 'Long-Term Evolution of the Luteoviridae: Time Scale and Mode of Virus Speciation', *Journal of Virology*, 84: 6177–87.

Parks, D. H. et al. (2013) 'Gengis 2: Geospatial Analysis of Traditional and Genetic Biodiversity, with New Gradient Algorithms and An Extensible Plugin Framework', *PLoS One*, 8: e69885.

Pinel, A. et al. (2000) 'Molecular Variability of Geographically Distinct Isolates of Rice Yellow Mottle Virus in Africa', *Archives of Virology*, 145: 1621–38.

Pinel-Galzi, A. et al. (2009) 'Recombination, Selection and Clock-Like Evolution of Rice Yellow Mottle Virus', *Virology*, 394: 164–72.

—— et al. (2015) The biogeography of viral emergence: rice yellow mottle virus as a case study. *Current Opinion in Virology*, 10: 7–13.

Portères, R. (1950) 'Vieilles Agricultures de l'afrique Intertropicale', *Agronomie Tropicale*, 5: 489–507.

—— (1957) 'Compagnonnage Agraire et Génétique Biogéographique Chez les Riz Cultivés', *CR Social de Biogéographie*, 34: 68–99.

—— (1962) 'Berceaux Agricoles Primaires Sur Le Continent Africain', *The Journal of African History*, 3: 195–210.

Pybus, O. G. et al. (2012) 'Unifying the Spatial Epidemiology and Molecular Evolution of Emerging Epidemics'. *Proceedings of the*

National Academy of Sciences ƒ the United States of America, 109: 15066–71.

Ramsden, C. et al.; VGDN Consortium. (2008) 'High Rates of Molecular Evolution in Hantaviruses', *Molecular Biology and Evolution*, 25: 1488–92.

Rodríguez-Cerezo, E. et al. (1991) 'High Genetic Stability in Natural Populations of the Plant RNA Virus Tobacco Mild Green Mosaic Virus', *Journal of Molecular Evolution*, 32: 328–32.

Sarra, S., and Peters, D. (2003) 'Rice Yellow Mottle Virus is Transmitted by Cows, Donkeys, and Grass Rats in Irrigated Rice Crop Plants', *Plant Disease*, 87: 804–8.

Seo, T.-K. et al. (2002) 'A Viral Sampling Design for Testing the Molecular Clock and for Estimating Evolutionary Rates and Divergence Times', *Bioinformatics*, 18: 115–23.

Shah, V. B., and McRae, B. H. (2008) 'Circuitscape: A Tool for Landscape Ecology', in G., Varoquaux, T., Vaught, and J., Millman (eds.) *Proceedings of the 7th Python in Science Conference* (*SciPy 2008*), pp. 62–66, https://28d97f1d-a-56512a53-s-sites.googlegroups.com/a/circuitscape.org/circuitscape/downloads/files/Shah_McRae_Circuitscape_Python_Scipy08.pdf?attachauth=ANoY7co02PPMzKhFUEcVPTdHJH9PWt2TbATfQAJG4lZa-ooDuY_Sgvg667Cx-Y2SDZ4cIi_2-jeybaLEJjYWyzrwyhjNu6OtGnjjCZQJmBTK7c8DpBNy1HSeM_YCxUd_0OPETtejWxVHuE_mU_A8YK51PYIzpXxJsUd35t-aU__fHxlACEhVQqIt5eYoRE2Tvk1KRGhhzS4hIRqe-1zPa19Lj2BteRfZHlCBMpJLQkfDukKMmmSW8pQSU4iYgTSiLlmdqS26p71RLBjc6XsgyyQeV9gf88XJCQ%3D%3D&attredirects=0.

Shapiro, B., Rambaut, A., and Drummond, A. J. (2006) 'Choosing Appropriate Substitution Models for the Phylogenetic Analysis of Protein-Coding Sequences. *Molecular Biology and Evolution*, 23: 7–9.

Simmons, H. E., Holmes, E. C., and Stephenson, A. G. (2008) 'Rapid Evolutionary Dynamics of Zucchini Yellow Mosaic Virus', *Journal of General Virology*, 89: 1081–5.

Sõmera, M., Sarmiento, C., and Truve, E. (2015) 'Overview on Sobemoviruses and a Proposal for the Creation of the Family Sobemoviridae', *Viruses*, 7: 3076–115.

Streicker, D. G. et al. (2012) 'Rates of Viral Evolution are Linked to Host Geography in Bat Rabies', *PLoS Pathogens*, 8: e1002720.

Suchard, M. A., and Rambaut, A. (2009) 'Many-Core Algorithms for Statistical Phylogenetics', *Bioinformatics*, 25: 1370–6.

Thresh, J. M. (1982) 'Cropping Practices and Virus Spread', *Annual Review of Phytopathology*, 20: 193–216.

Traore, O. et al. (2005) 'Processes of Diversification and Dispersion of Rice Yellow Mottle Virus Inferred from Large-Scale and High-Resolution Phylogeographical Studies', *Molecular Ecology*, 14: 2097–110.

—— et al. (2006) 'Rice Seedbed as a Source of Primary Infection by Rice Yellow Mottle Virus', *European Journal of Plant Pathology*, 115: 181–6.

Trovão, N. S. et al. (2015) 'Bayesian Inference Reveals Host-Specific Contributions to the Epidemic Expansion of Influenza A H5N1', *Molecular Biology and Evolution*, pii: msv185.

Wang, T. H. et al. (2001) Identification of shared populations of human immunodeficiency virus type 1 infecting microglia and tissue macrophages outside the central nervous system. *Journal of Virology*, 75: 11686–99.

Wertheim, J. O., and Kosakovsky Pond, S. L. (2011) 'Purifying Selection can Obscure the Ancient Age of Viral Lineages', *Molecular Biology and Evolution*, 28: 3355–65.

Woolhouse, M. E. J., and Gowtage-Sequeria, S. (2005) 'Host Range and Emerging and Reemerging Pathogens', *Emerging Infectious Diseases*, 11: 1842–7.

Worobey, M., Han, G.-Z., and Rambaut, A. (2014) 'A Synchronized Global Sweep of the Internal Genes of Modern Avian Influenza Virus', *Nature*, 508: 254–7.

Xie, W. et al. (2011) 'Improving Marginal Likelihood Estimation for Bayesian Phylogenetic Model Selection', *Systems Biology*, 60: 150–60.