

Appendix S1 for

Ebolavirus evolution and emergence are associated with land use change

Authors: Christian E. Lange, Thomas R. Barnum, David J. McIver, Matthew LeBreton, Karen Saylor, Charles Kumakamba, Sara Lowes, Eduardo Montero, Robert L. Cohen

Ecological Monographs

This PDF file includes:

Section S1
Figures S1 to S13
Tables S1 to S4
References

Section S1

Supplementary Methods

Estimating spatiotemporal ranges of most recent common ancestors (MRCAs)

Polygons representing the potential geographic ranges of each MRCA between two given outbreaks were reconstructed as the intersection of circles drawn around each outbreak site, with the latitude and longitude for each outbreak's index case (Mylne et al. 2014) as the centroid for each circle and the radii of the circles defined by the MRCA's 95% Highest Posterior Density (HPD) (Appendix S1: Fig. S1). The radius of the circle drawn around the less recent outbreak was the distance from that outbreak's index case to the point at which the MRCA's older limit would have existed if it then diverged and traveled at constant velocity to the two subsequent outbreak sites in the observed years. The radius of the circle drawn around the more recent outbreak was the distance from that outbreak's index case to the point the MRCA's more recent limit would have existed if it then diverged and then traveled at constant velocity to the two subsequent outbreak sites in the observed years. If the more recent limit of the MRCA's 95% HPD was more recent than the latter outbreak, it was constrained to be the same year as that outbreak. The calculations were as follows:

$$radius_A = distance_{AB} * \frac{(yr_A - MRCA_{UL})}{(yr_A - MRCA_{UL}) + (yr_B - MRCA_{UL})} \quad (\text{Equation S1})$$

$$radius_B = distance_{AB} * \frac{(yr_B - MRCA_{LL})}{(yr_A - MRCA_{LL}) + (yr_B - MRCA_{LL})} \quad (\text{Equation S2})$$

where yr_i is the year in which outbreak i (A or B) was reported, $distance_{AB}$ is the Haversine distance between outbreaks A and B, A is the less recent of the two outbreaks, and $MRCA_{UL}$ and $MRCA_{LL}$ are the years of the upper and lower limit of the MRCA's 95% HPD (the less and more recent limits, respectively, where $MRCA_{LL}$ was constrained to $\leq yr_B$).

The geographic centroids and median MRCA estimates of each polygon were then used as the estimated location for that node to repeat the procedure retrospectively to reconstruct each prior divergence event until reaching the root of the phylogenetic tree. The distance traversed between each MRCA was estimated by randomly sampling 10,000 points within each polygon and measuring the distance between each pair. These path distances were summed to simulate 10,000 possible paths and distances traversed for EBOV and SUDV. Velocities reported for each outbreak represent the mean and 95% confidence interval of distance sampled divided by the difference in years between the outbreak and its MRCA. We held these methods constant for all outbreaks and nodes for EBOV and SUDV except for the Yambio 2004 SUDV outbreak, since this method would model SUDV traveling from South Sudan to Uganda (Node D) and back, which is unlikely given the phylogeographic analysis (Figs. 1 and 3). Instead, we assumed the 2004 Yambio SUDV outbreak had the same MRCA as the 1979 Maleo SUDV outbreak. We also dropped SUDV Node B from this analysis since it was likely an artifact of passaging the 1976 SUDV sequences, with its position in the tree having only 50% confidence. Finally, outbreaks of EBOV which occurred in very close spatiotemporal proximity with >99.9% phylogenetic similarity (such as outbreaks in Luebo, DRC in 2007 and again in 2008, or on the Gabon/Congo border from 2001-2003) were considered as the same outbreak for simplicity.

This method assumes that MRCAs existed between nodes or outbreaks that were adjacent in the phylogenetic tree rather than a random location. This assumption is somewhat validated by the finding that polygons representing the estimated geographic range of the origin MRCAs of EBOV and SUDV are located near the first two known outbreak sites of each virus. While Yambuku falls just outside the EBOV origin polygon, that polygon does contain the village of Bonduni, and an epidemiologic investigation of the 1977 Bonduni case suggested a prior outbreak of EVD-like disease in 1972 in the nearby town of Tandala, with filovirus seroprevalence of 10% for people over age 30 and 6% for those under 30 (Heymann et al. 1980). Nevertheless, these polygons produce smaller estimated ranges for MRCAs than SERAPHIM would, and this increased precision may reduce accuracy, an important limitation that requires caution and triangulation when drawing inferences based on these polygons.

The polygons that represent the geographic location where splits in the phylogenetic tree occurred were estimates that did not consider ecotone or climate. Sophisticated methods that consider factors such as ecotone and climate to identify a geographic location where a basal population may have emerged do exist, but because a specific host species is not known, and because most polygons for each virus occurred across similar ecotones, a method that used fewer inputs was necessary. Instead, our polygons focused on accounting for the temporal uncertainty around when the split in the phylogenetic tree occurred to identify a potential geographic range.

Regression analysis of spatial spread for EBOV and SUDV

Using the 10,000 simulated paths of spread for each virus, we hypothesized that each virus spread as an advancing wave across Africa, as suggested previously (Walsh, Biek, and Real 2005), and which would be possible through its suspected airborne reservoir. We hypothesized a positive relationship between the distance traversed by each genetic virus variant, the time since the phylogenetic origin, and phylogenetic distance (% of nucleotide dissimilarity). We used ordinary least squares regression with robust standard errors to test how distance traversed varied by time, phylogenetic distance, and virus species.

Estimating speeds of dispersal from polygons

Using the 10,000 simulated paths of spread for each virus, 10,000 estimates of dispersal speeds from polygons to their descendent outbreak sites were calculated directly, using the median estimate for the MRCAs represented by each polygon, and the known year of the outbreaks (Fig. 2b,d). In the Monte Carlo simulations, an analogous dispersal speed for each dispersal segment was calculated as a random estimate drawn from the normal distribution calculated for that path length divided by the time between the outbreak year and the random estimate drawn for the Monte Carlo simulation from the relevant MRCA 95% highest posterior density (see Methods, section on Monte Carlo simulation).

For the estimated dispersal speed between polygons, we could not use randomly sampled path distances and MRCA years, since these could sometimes yield unrealistically fast dispersal speeds (for example, when nodes were randomly sampled to be separated by one year). Therefore, we estimated a normal distribution for the diffusion speed between polygons, before and after 2000, using the results from the 10,000 simulations shown in Fig 2a and 2c. For each segment in each trial in the Monte Carlo simulation, a random value for each between-polygon dispersal speed was drawn from these normal distributions, for EBOV and SUDV separately.

Comparing measures of fragmentation at coarser and finer spatial scales

The HYDE data is suitable for exploring changes in human population densities across landscapes through time, but the effects of land use change on wildlife populations may occur both at the landscape scale and at smaller spatial scales (Johansson, Primmer, and Merilä 2007). To assess whether the HYDE data may be suitable for comparing effects on wildlife populations during decades preceding the availability of high resolution satellite data, we compared changes in the HYDE data to tree cover loss as measured by remote sensing data from Global Forest Watch (GFW). GFW measures tree cover loss at a 30 m x 30 m resolution from 2000 to 2019 (Hansen et al. 2013). Although remotely-sensed forest cover data does not go back to 1900, an important limitation in studying twentieth century virus emergence, we compared the similarity of detected forest fragmentation from the GFW data to that from the HYDE data in order to test whether HYDE provided a reasonable estimate for fragmentation at smaller spatial scales. The most recently available version of the forest cover data (V1.8) was reprocessed from 2011 to 2017, allowing for year-to-year comparisons in forest fragmentation. We used a mixed effects model with the lme function in R, with polygon as a random effect, to compare CV from 2011 to 2017 of annual HYDE data to annual GFW data for nine of ten EBOV polygons (excluding J/Lue-Gui due to its large size). The model revealed a significant, positive relationship between the two data types, at all human population densities, suggesting the HYDE data does indeed capture the broader pattern of forest fragmentation (Appendix S1: Fig. S8). The scale of fragmentation may vary at different population densities.

Global Forest Watch data query

The most recently available GFW data at the time of the analysis (V1.8) was accessed at <https://storage.googleapis.com/earthenginepartners-hansen/GFC-2020-v1.8/download.html>. Data for the Congo basin were downloaded by selecting four tiles, which produced the following queries, and which were analyzed in R version 3.6 (R Core Team 2019):

1. https://storage.googleapis.com/earthenginepartners-hansen/GFC-2020-v1.8/Hansen_GFC-2020-v1.8_lossyear_10N_010E.tif
2. https://storage.googleapis.com/earthenginepartners-hansen/GFC-2020-v1.8/Hansen_GFC-2020-v1.8_lossyear_10N_020E.tif
3. https://storage.googleapis.com/earthenginepartners-hansen/GFC-2020-v1.8/Hansen_GFC-2020-v1.8_lossyear_00N_010E.tif
4. https://storage.googleapis.com/earthenginepartners-hansen/GFC-2020-v1.8/Hansen_GFC-2020-v1.8_lossyear_00N_020E.tif

Assessing a discontinuity in the HYDE data

While researchers have spent decades refining historical maps of LUC (Klein Goldewijk et al. 2017), the estimates can be impacted by changes in political boundaries. For example, the maps for Sudan show an unexpected shift in 2011 after South Sudan became a country (Fig. 6), likely reflecting unreconciled data reported by the governments of Sudan in 2010 and South Sudan in 2011 (Klein Goldewijk et al. 2017; Food and Agriculture Organization 2020). While this should not affect inference around the time of the SUDV original phylogenetic split in the 1950s, in which we detect LUC that aligns well with historical events, the abrupt change in 2011 underscores the need for linking changes in the maps to historical events. In our study, we

accounted for these data limitations by including various types of uncertainty in different regression models.

Model selection methodology for generalized estimating equations

Model selection criteria for generalized estimating equations (GEEs) are an active area of research and optimized methods still being pursued (M. Wang 2014). Because GEEs do not use maximum likelihood estimation, selection criteria such as the Akaike Information Criterion (AIC) cannot be used (M. Wang 2014; Cui 2007). In 2001, Pan proposed a Quasi-Likelihood Information Criterion (QIC) as an analog of AIC (M. Wang 2014; Cui 2007). Subsequent empirical usage and simulation studies has shown that QIC does not often perform well in selecting the true model (M. Wang 2014; Y. Wang et al. 2015). The current recommendation is to use the Correlation Information Criterion (CIC) to select a “working” correlation structure, since GEEs produce consistent parameter estimates even when the correlation structure is misspecified (M. Wang 2014; Hin and Wang 2009), and then use the Extended Quasi-Likelihood Information Criterion (EQIC) to select covariates. We followed this recommendation, using the `qic` command in Stata to calculate CIC (Cui 2007) and QIC, and calculating EQIC as defined previously (M. Wang 2014; Y.-G. Wang and Hin 2010).

Stata code

To run the GEEs for each of the six land use change (LUC) scenarios defined in Table 1, we used the following code in Stata/MP 16.1 (“Stata Statistical Software: Release 16.” 2019):

```
foreach num of numlist 1/6 {
  xtgee substitution_rate i.sudan##c.delta_CV##c.speed i.sudan##c.percent i.sudan##c.time if
  luc_scenario == `num', corr(ar 1) vce(robust) family(poisson) link(log) i(uniqueid) t(decade)
  eform }
```

Monte Carlo simulation rejection sampling

We used an ordinary Monte Carlo simulation with rejection sampling (Brooks et al. 2011) to create 1,200 possible simulations of the phylogenetic trees. We used the 95% HPD normal or log-normal distributions from the phylogenetic analysis as the prior distributions for each split. Since these distributions could be sampled from easily, Markov chains were not necessary and would have been complex due to the inherent multidimensionality of the distribution. Instead, the samples were tuned to approximate the prior distributions using rejection sampling. The accept-reject function, which was validated empirically (Appendix S1: Table S3), had a uniform distribution g as the denominator tuned to avoid oversampling in the later part of distributions, given that each split y was constrained to occur more recently than the previous split x :

For each split s , sample $y_s \sim h(\mu_s, \sigma_s)$ where h is a normal or log-normal distribution of the split defined by the phylogenetic trees (Fig. 1) with probability density function f :
 Accept y with probability $A = a(x,y)$ where (Equation S3)
 $A = 0$ if $y < x$
 $A = f(y_s, \mu_s, \sigma_s)/g(x,y)$ if $y > x$
 where $g(x,y) = y - x$ or $\ln(y) - \ln(x)$
 and μ , x and y are all years.

Institutional Review Board Approval

These surveys had Institutional Review Board (IRB) approval from the Harvard Committee on the Use of Human Subjects (CUHS) (Harvard IRB 16-1065).

Supplementary Figures and Tables

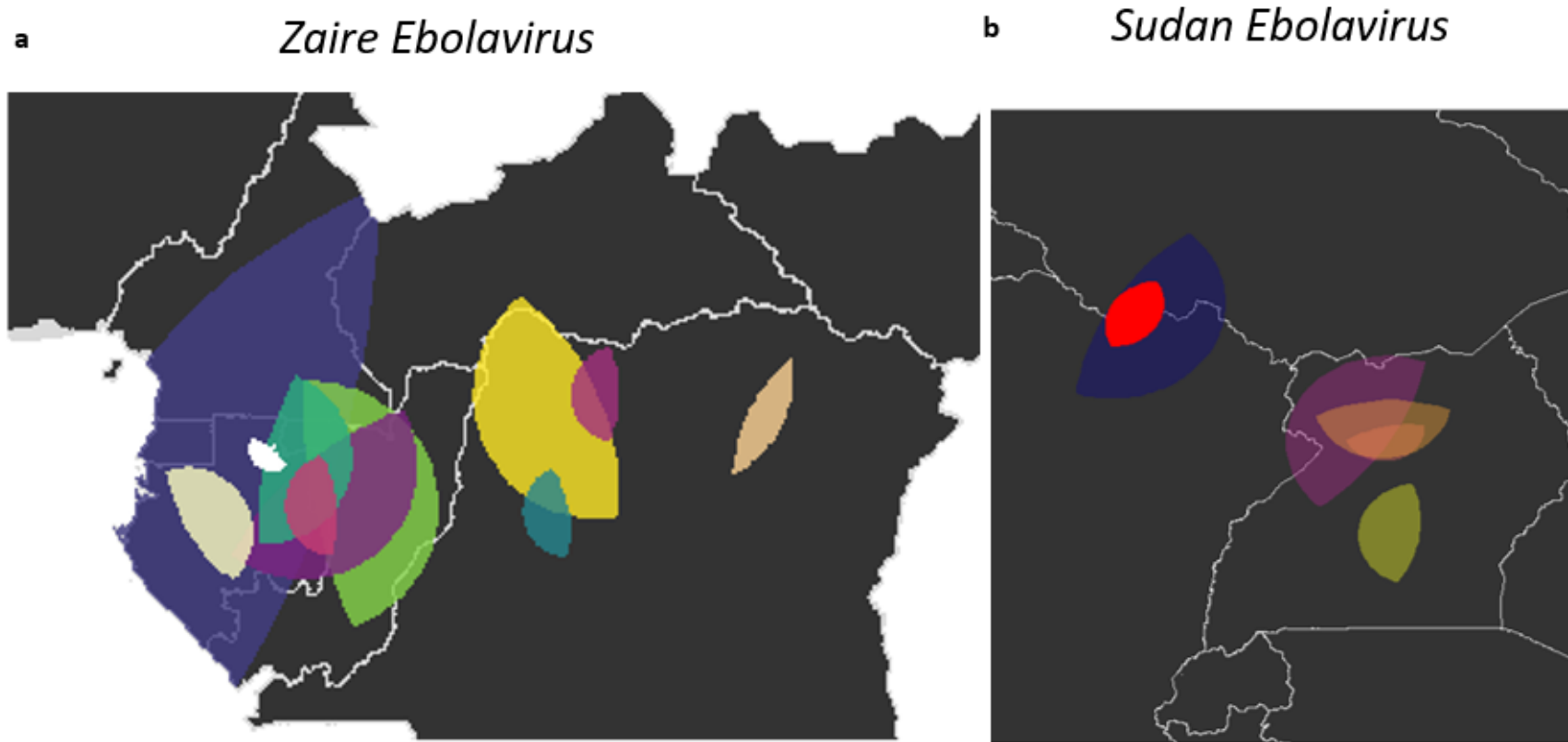


Figure S1. Estimated polygons representing ranges for major splits in phylogenetic trees. (a) EBOV. (b) SUDV. Estimated origin MRCAs are in yellow for EBOV and blue for SUDV. The estimated ranges for many adjacent EBOV MRCAs overlap on the Gabon/ROC border in the 1980s-90s, reflecting circulation and evolutionary events there. We use the large blue polygon for the SUDV origin rather than the small red polygon since the red is completely enveloped by blue, reflecting the fact that the 1976 and 1979 outbreaks started in the same town (Nzara).

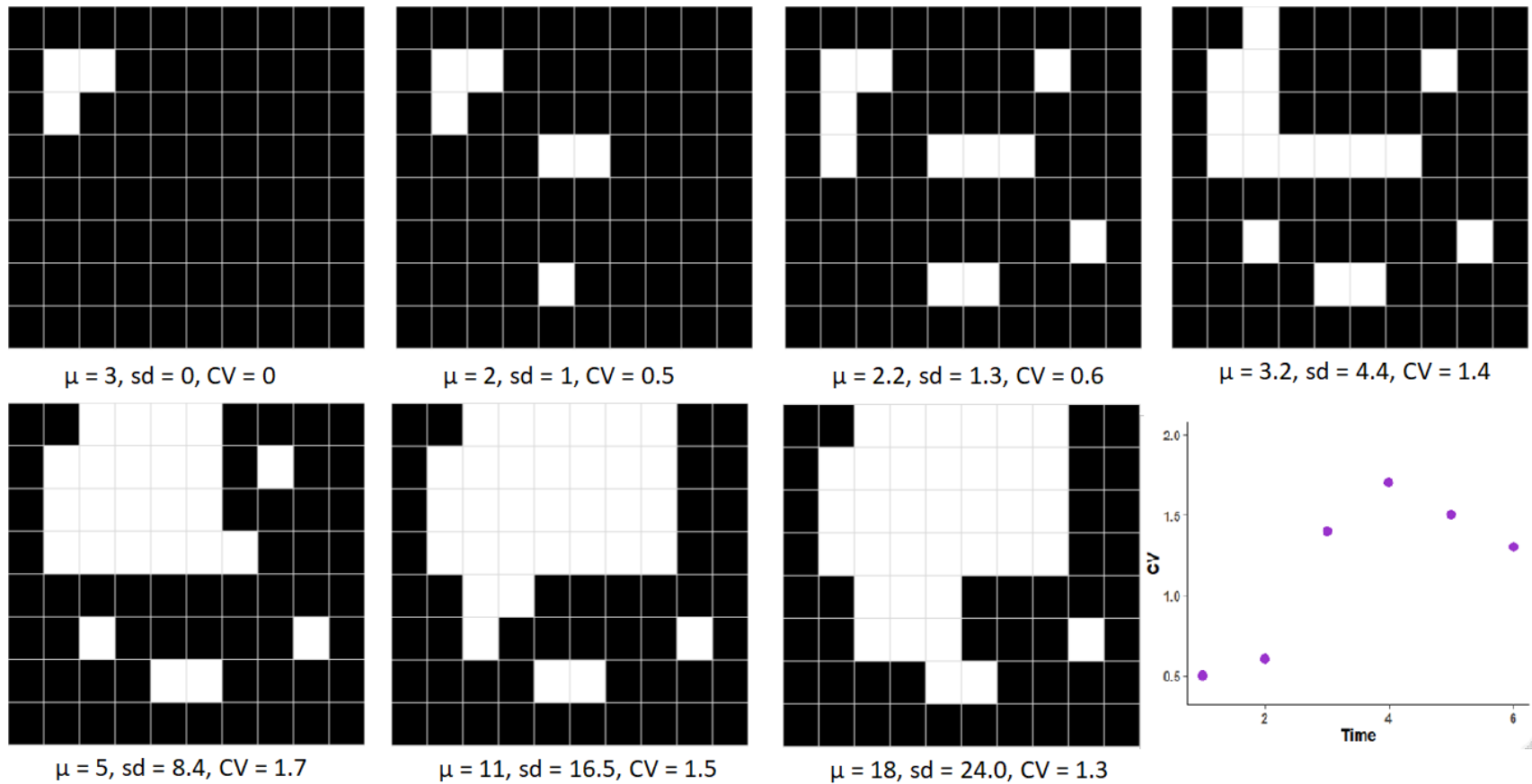


Figure S2. Coefficient of variation as a measure of disturbance. Schematic showing how coefficient of variation (CV) was calculated and can change as a polygon shifts from natural, or undisturbed habitat, to mostly anthropogenic (Tilman 1996). The top left panel represents a polygon that is predominantly undisturbed (black squares) with a small anthropogenic disturbance (white squares). CV is the standard deviation (sd) divided by the mean (μ). Moving along the top row, the anthropogenic disturbance expands in a patchy pattern, representing the clusters of anthropogenic activity that occurred in the MRCA polygons. The bottom left panel shows how the presence of patches in the landscape can increase the CV, with only 31% of the squares anthropogenic, showing that a high CV and fragmentation is not necessarily synonymous with anthropogenic land use dominance. Anthropogenic squares continue to expand in the next two panels, but as the clusters of anthropogenic land use become connected, the CV decreases. The scatter plot in the lower right plots the change in CV through time as the polygon shifted from a predominantly undisturbed landscape to a landscape nearly evenly split by anthropogenic land use.

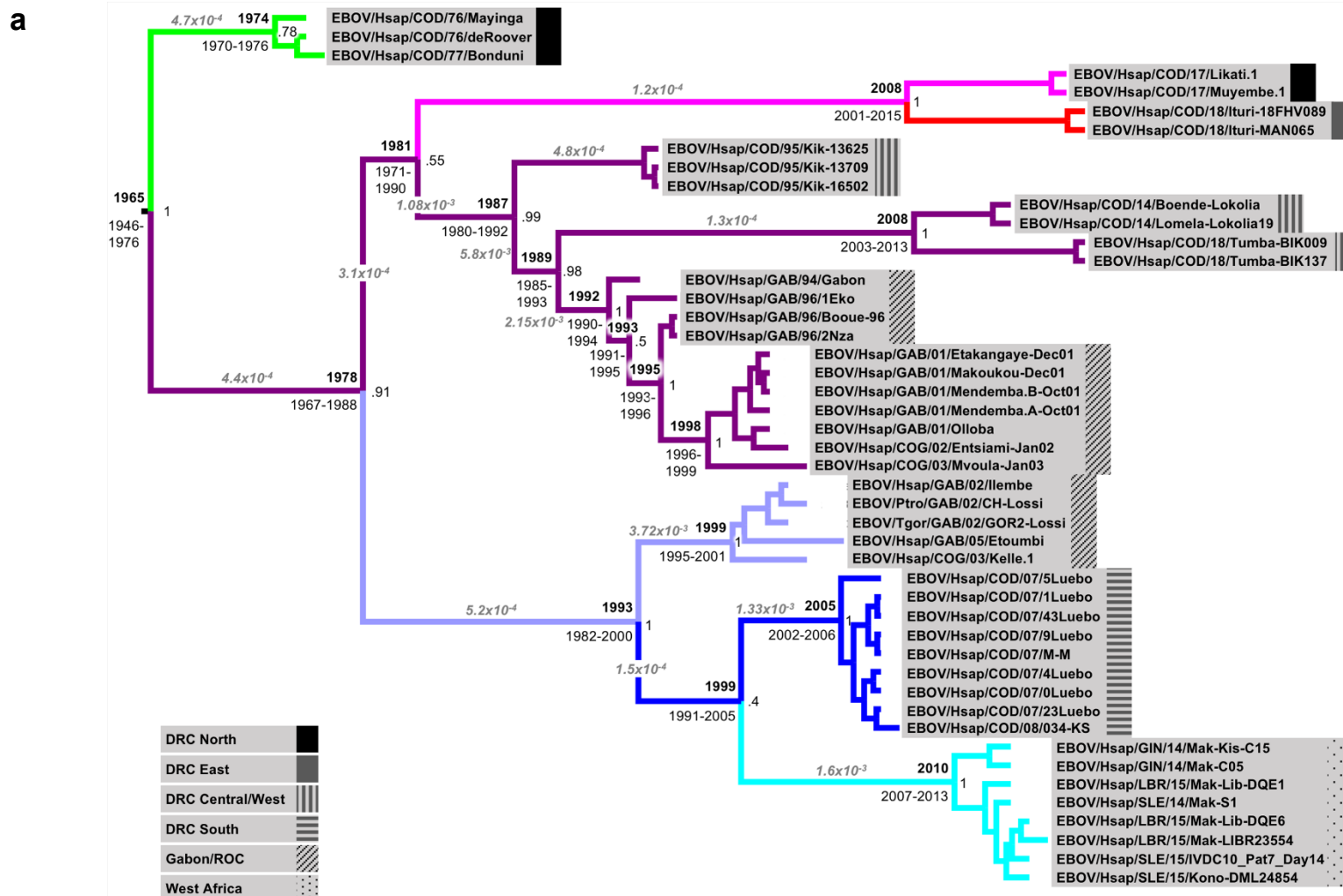


Figure S3. Additional phylogenetic analyses. (a) Phylogenetic tree based on the glycoprotein (GP) sequences of 47 EBOV isolates. Major nodes are indicated by capital letters; sequences included are labeled using “medium-length designations.” Branches of the same color indicate sequence differences of less than 1% among individual sequences in the branch. Indicated at major nodes are the MRCA year (above node), the 95% HPD for the MRCA (below node), and the bootstrap support for the node (right of node). Branches show the rate of substitution (*italics*). (continued on next page)...

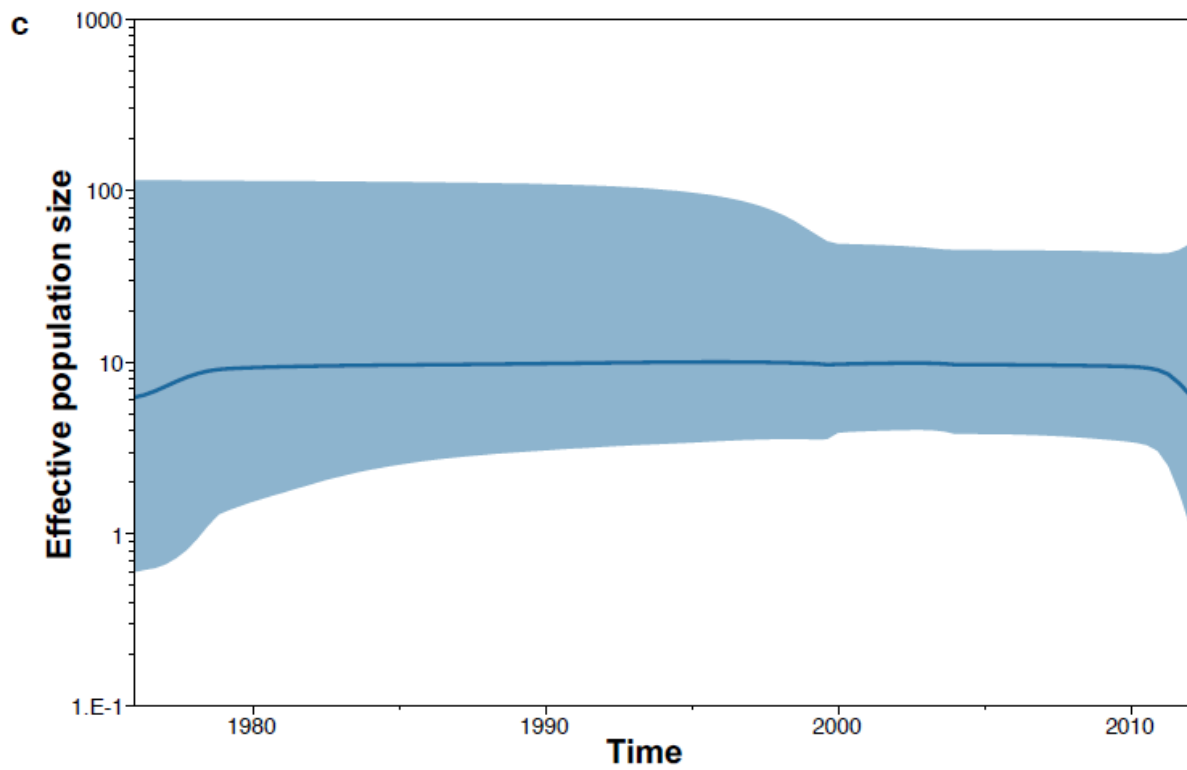
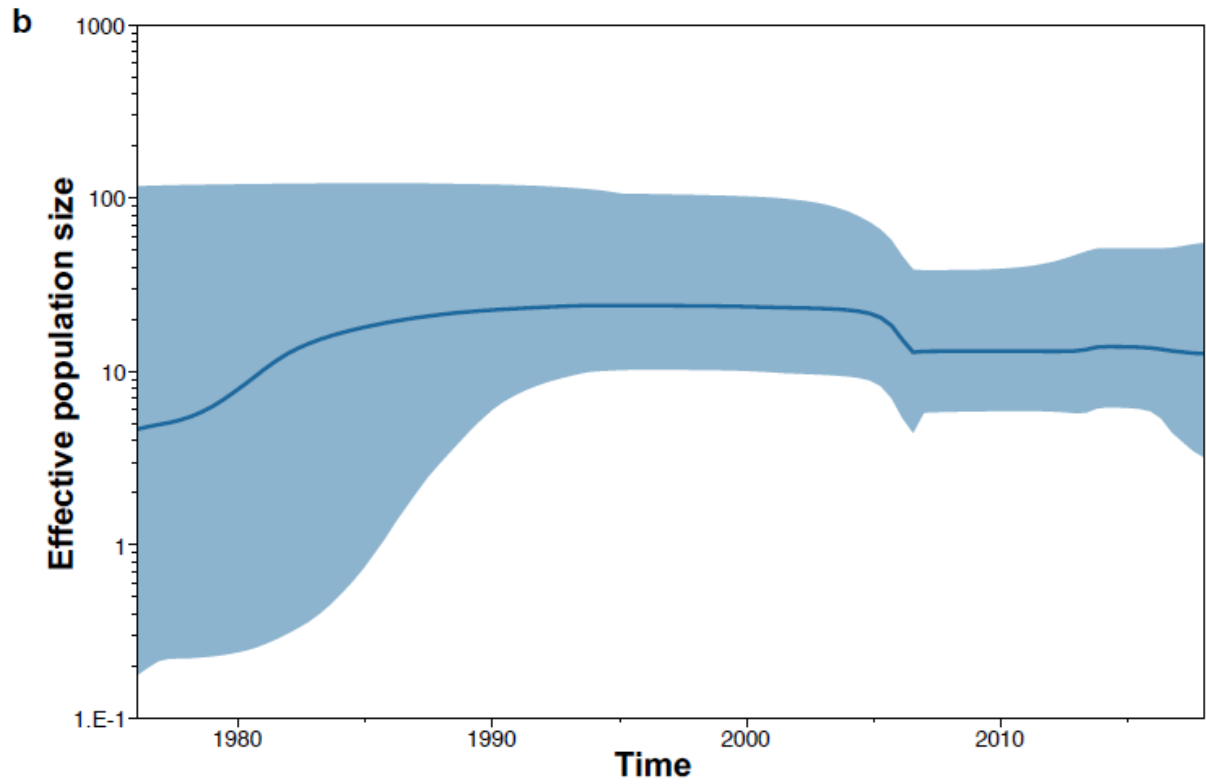


Figure S3. Additional phylogenetic analyses (continued). Bayesian skyline plots showing effective population size over time for (b) EBOV and (c) for SUDV.

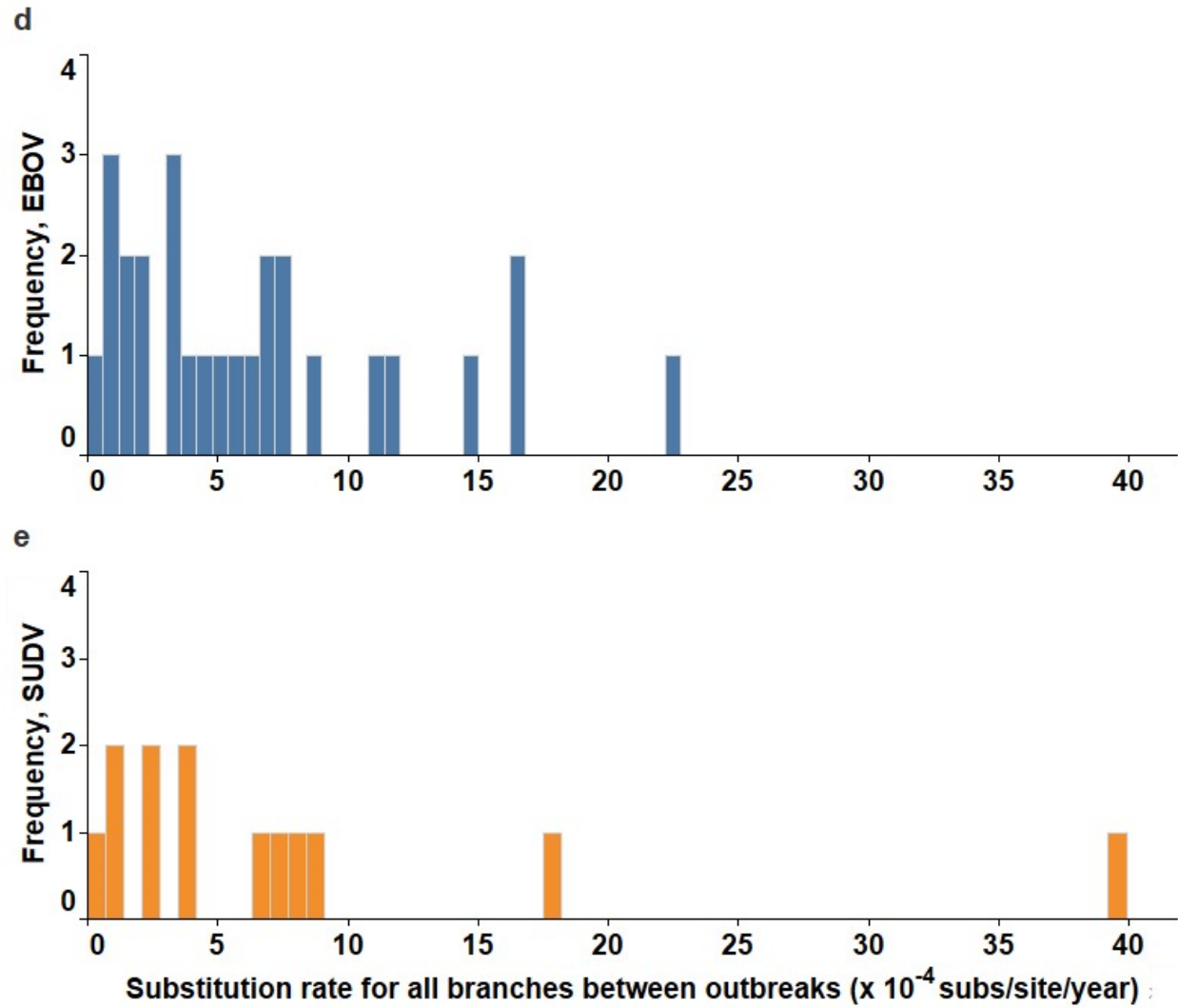


Figure S3. Additional phylogenetic analyses (continued). (d) Histogram of all substitution rates calculated from root-to-tip between major branches (between outbreaks and/or MRCAs) from the unrooted trees using complete genomes (Fig. 1) for EBOV and (e) for SUDV.

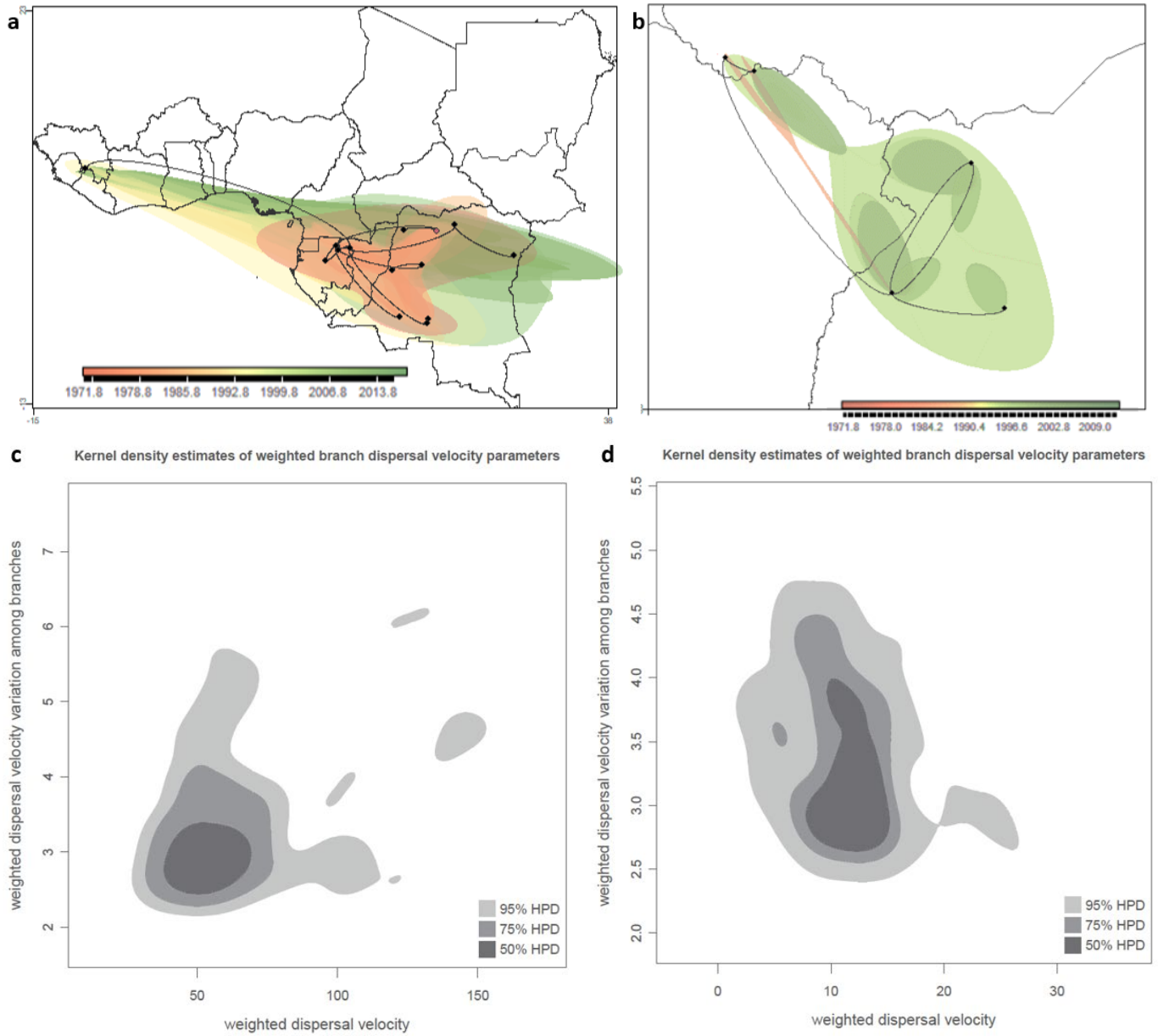


Figure S4. Continuous phylogeography results generated by SERAPHIM. (a) Reconstructed spatiotemporal diffusion of EBOV showing 95% HPD regions based on 100 trees sampled from the posterior distribution, using methods described in Dellicour et al. 2016. The model calculates that EBOV circulated across the northern Congo basin in the 1970s, and then spread as an advancing wave across equatorial Africa. First recorded outbreak is the red dot (Yambuku), and all subsequent outbreaks are black dots. (b) Reconstructed spatiotemporal diffusion of SUDV from South Sudan to Uganda. The estimated SUDV original range was much smaller than that of EBOV. Black dots are outbreak sites. (c) Kernel density estimates of weighted branch velocity parameters (coefficient of variation against mean values, with weights calculated for each branch using SERAPHIM) for EBOV, in km/yr. (d) Kernel density estimates of weighted branch velocity parameters for SUDV, in km/yr. The diffusion of SUDV was slower and less variable across different branches than was the diffusion of EBOV, reflecting the more limited geographic range and different host species of SUDV.

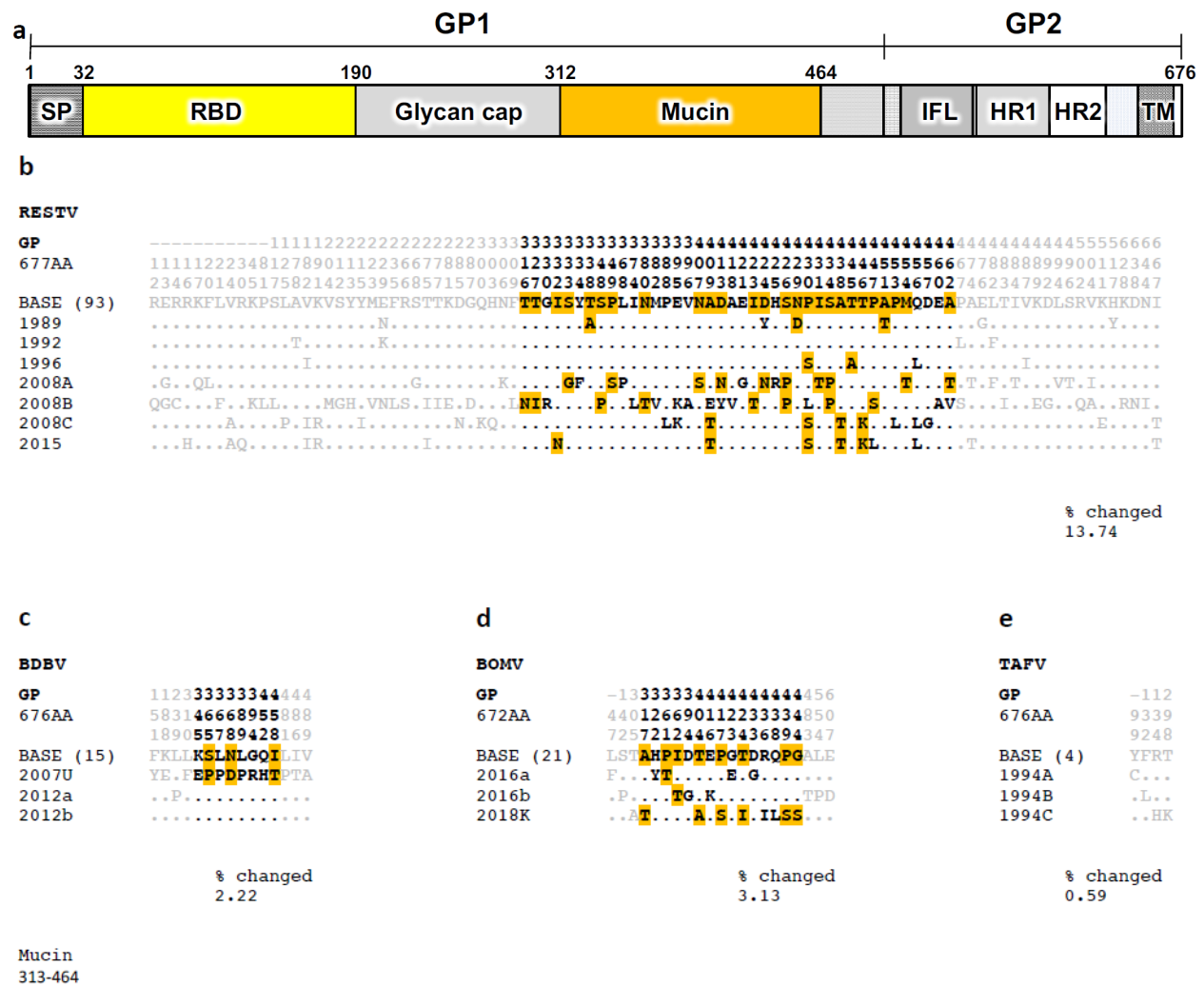
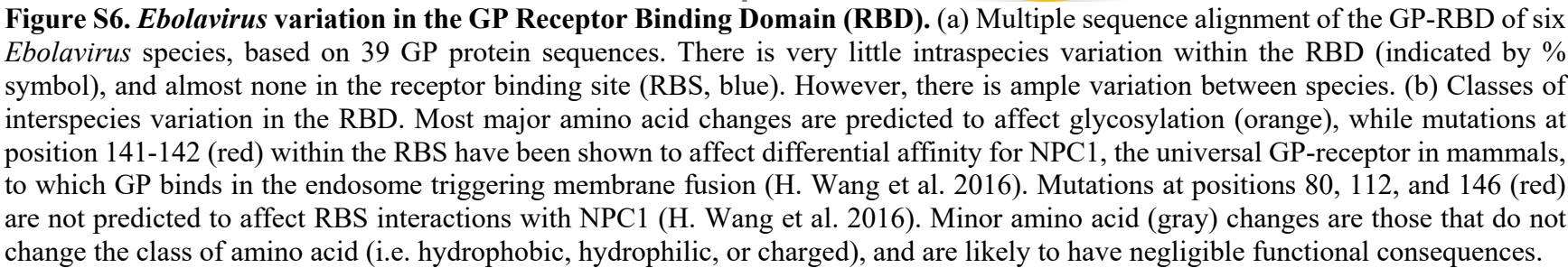


Figure S5. Multiple sequence alignment of surface glycoprotein mucin-like domain (GP-MLD) for other *Ebolavirus* species. The most common type of intraspecies variation is of potential glycosylation targets in the mucin-like domain, as for EBOV and SUDV. (a) Domains of the *Ebolavirus* surface glycoprotein (GP), which is cleaved by furin into subunits GP1 and GP2 prior to virus assembly (H. Wang et al. 2016). The MLD spans positions 313-464. SP = Signal peptide; RBD = Receptor-binding domain; IFL = Internal fusion loop; HR = Heptad repeat; TM = Transmembrane domain. (b) For RESTV, 65% of MLD variation is of potential glycosylation targets (orange). (c) For BDBV, 38%. (d) For BOMV, 53%. (e) For TAFV, 0%. Because there is only one known outbreak of TAFV, variation between sequences is likely an artifact of passaging.

ALL Ebolavirus species



%: Amino acid level variations exist within the species



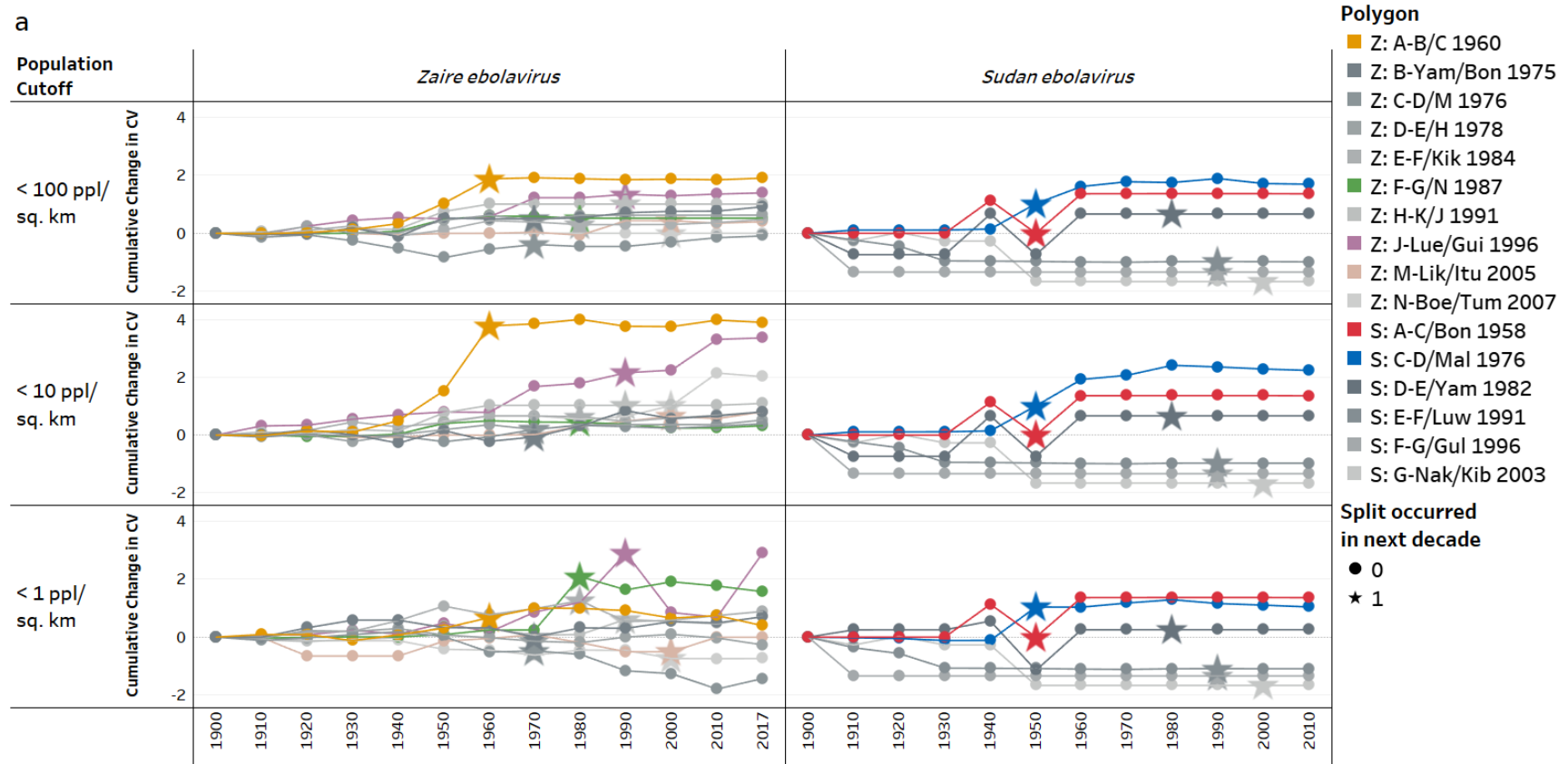


Figure S7. Cumulative change in land use change (LUC) 1900-2018 for polygons estimating the ranges of *Ebolavirus* MRCAs. (a) Coefficient of variation (CV) for each polygon as a function of time under different LUC scenarios, with different population human densities (people per sq. km) in semi-natural lands defining the anthropogenic biomes and thus reflecting different degrees of anthropogenic fragmentation of semi-natural ecosystems, as defined in Table 1. (b) (next page), Total percent anthropogenic land use in each polygon. Polygons are labeled based on nomenclature in Fig. 1, and colored based on Fig. S1 for polygons that were important drivers of statistical significance in the population averaged model. Polygons which were less important drivers of statistical significance are hues of gray. Stars indicate the decade in which a phylogenetic split occurred in the relevant polygon in the maximum clade credibility tree (Fig. 1). Z = EBOV; S = SUDV. Changes in CV in the decade prior to phylogenetic splits are largest in some polygons, notably both origin MRCAs (A-B/C for EBOV (yellow, Fig. 5) and C-D/Maleo for SUDV (blue, Fig. 6)), and those on the Gabon/ROC Border (F-G/N (green) and J-Lue/Gui (purple), Fig. S9). CV changes (a) are more pronounced than are the increases in total anthropogenic land conversion (b). C-D/Maleo (blue) is used as SUDV's origin since it completely envelops A-C/Bon (red), (Fig. S1).

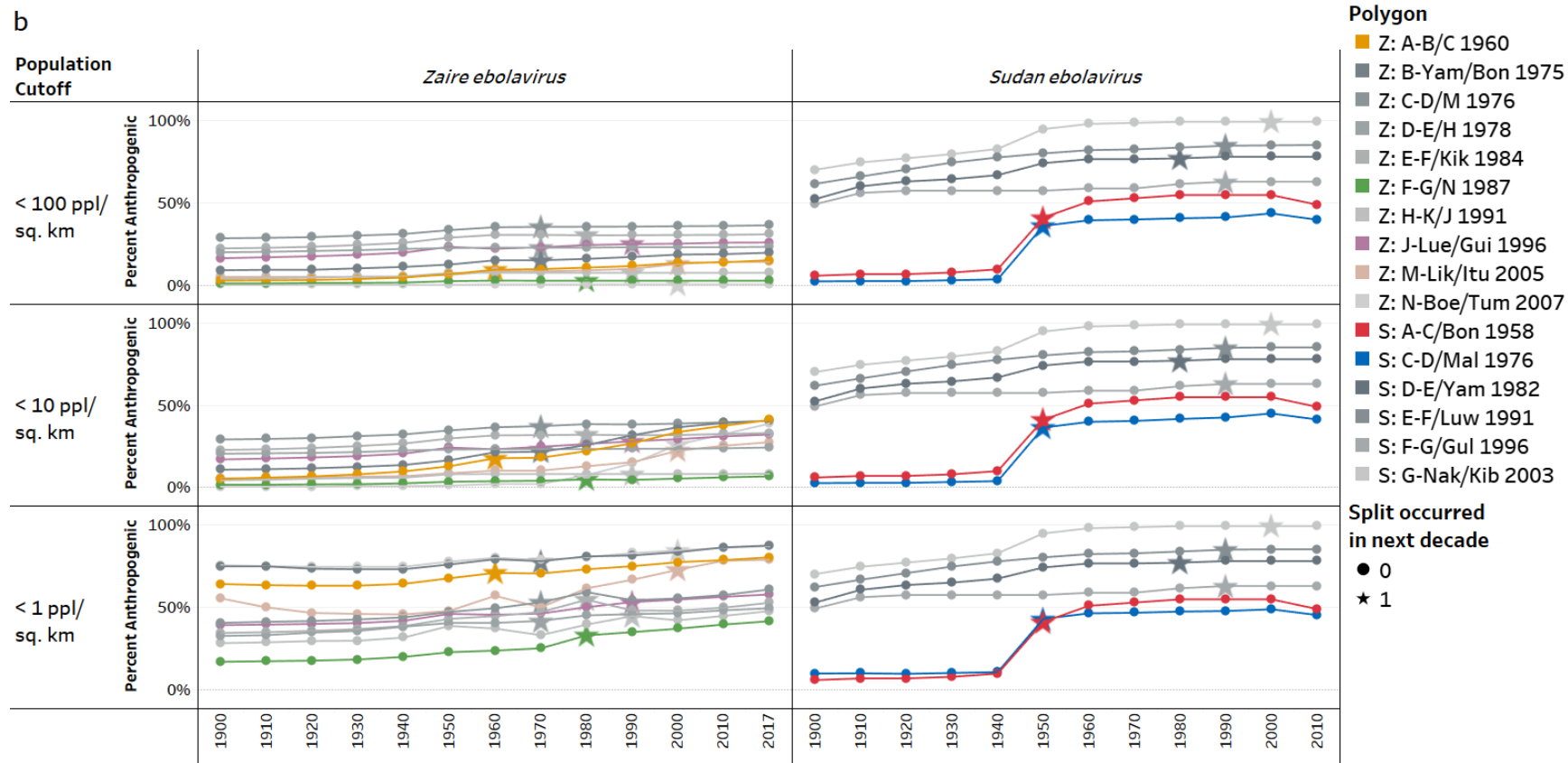


Figure S7. Cumulative change in percentage of anthropogenic land use for study polygons (continued).

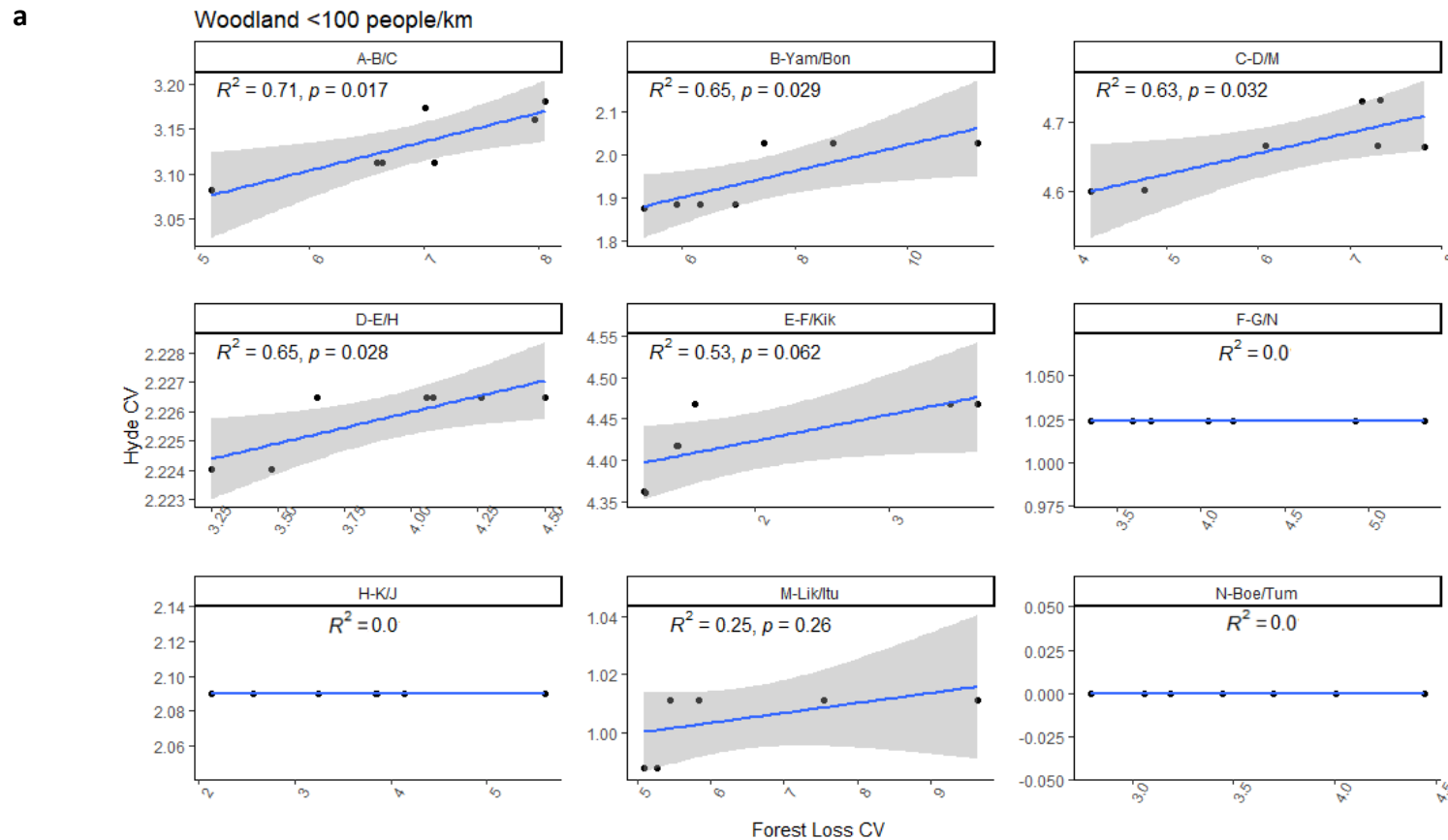


Figure S8. Comparison of fragmentation as measured by coefficient of variation (CV) at coarser and finer spatial scales. There is a positive association between CV as calculated from the HYDE 3.2 dataset (Klein Goldewijk et al. 2017), which measures landscape fragmentation at the coarser spatial scale of a 5' x 5' global grid (approximately 10 km x 10 km), and from Global Forest Watch (GFW), which measures forest loss with a 30 m x 30 m resolution. As a test case, the CV from three different HYDE LUC scenarios (Table 1), and from GFW were calculated annually for years 2011-2017 for all EBOV MRCA polygons except J-Lue/Gui, due to its very large size. The positive correlation is statistically significant on average in each LUC scenario (i.e. at all human population densities), which is acceptable since our model uses population-averaged generalized estimating equations, and suggestive of how the scale of fragmentation may vary at different population densities. (a) HYDE Residential Semi-natural Scenario vs. GFW, which compares fragmentation measures for all woodlands at the coarser and finer spatial scales (Wald test, $P < 0.01$), (continued next page)...

b

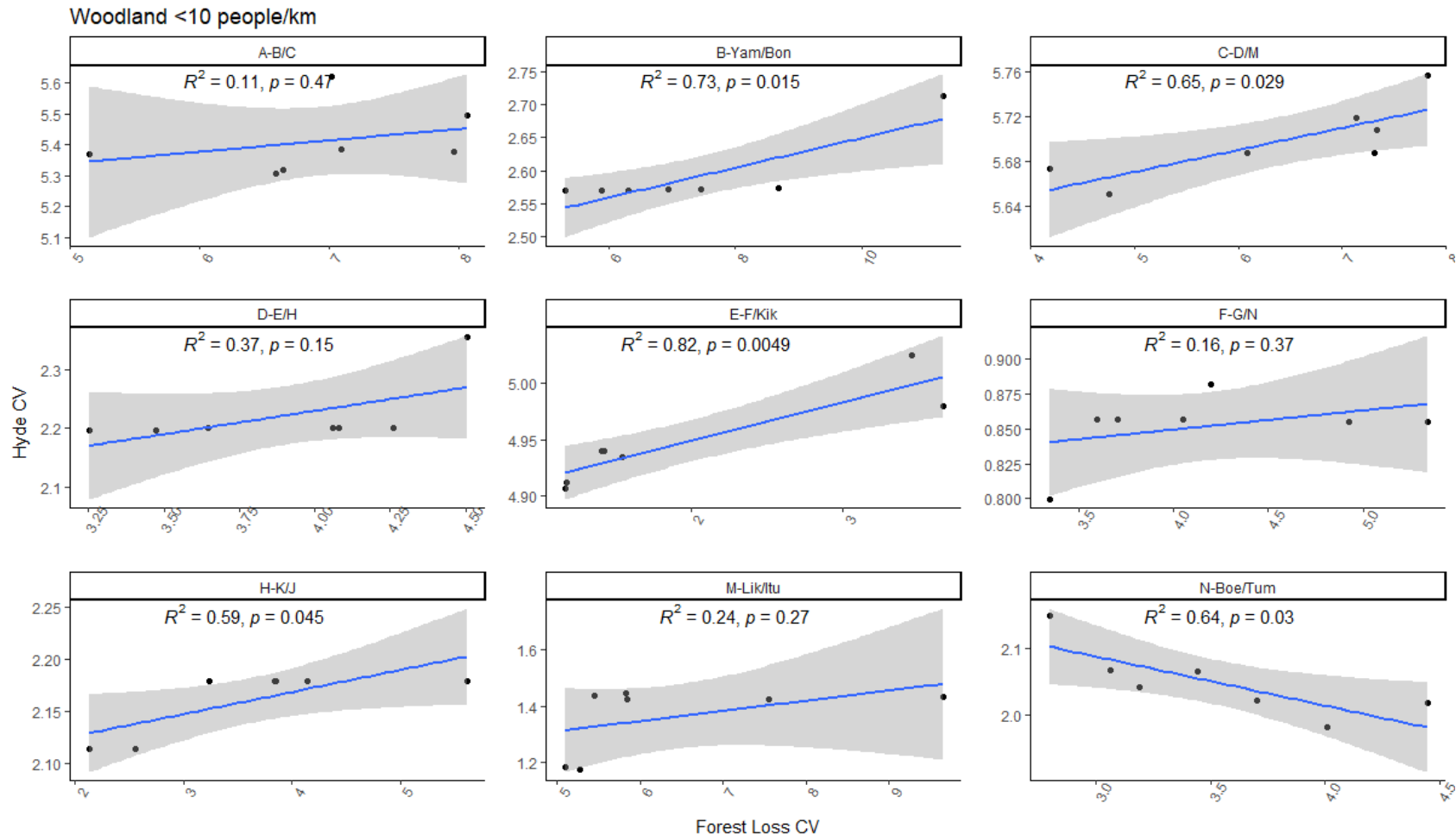


Figure S8. Comparison of fragmentation measures at coarse and fine spatial scales (continued). ... (b) HYDE Populated Semi-natural Scenario vs. GFW, (Wald test, $P < 0.01$), with woodlands with human population density $> 10/\text{km}^2$ considered anthropogenic, thus potentially comparing more granular fragmentation projected by the HYDE data with GFW (Wald test, $P < 0.01$), and (continued next page)...

c

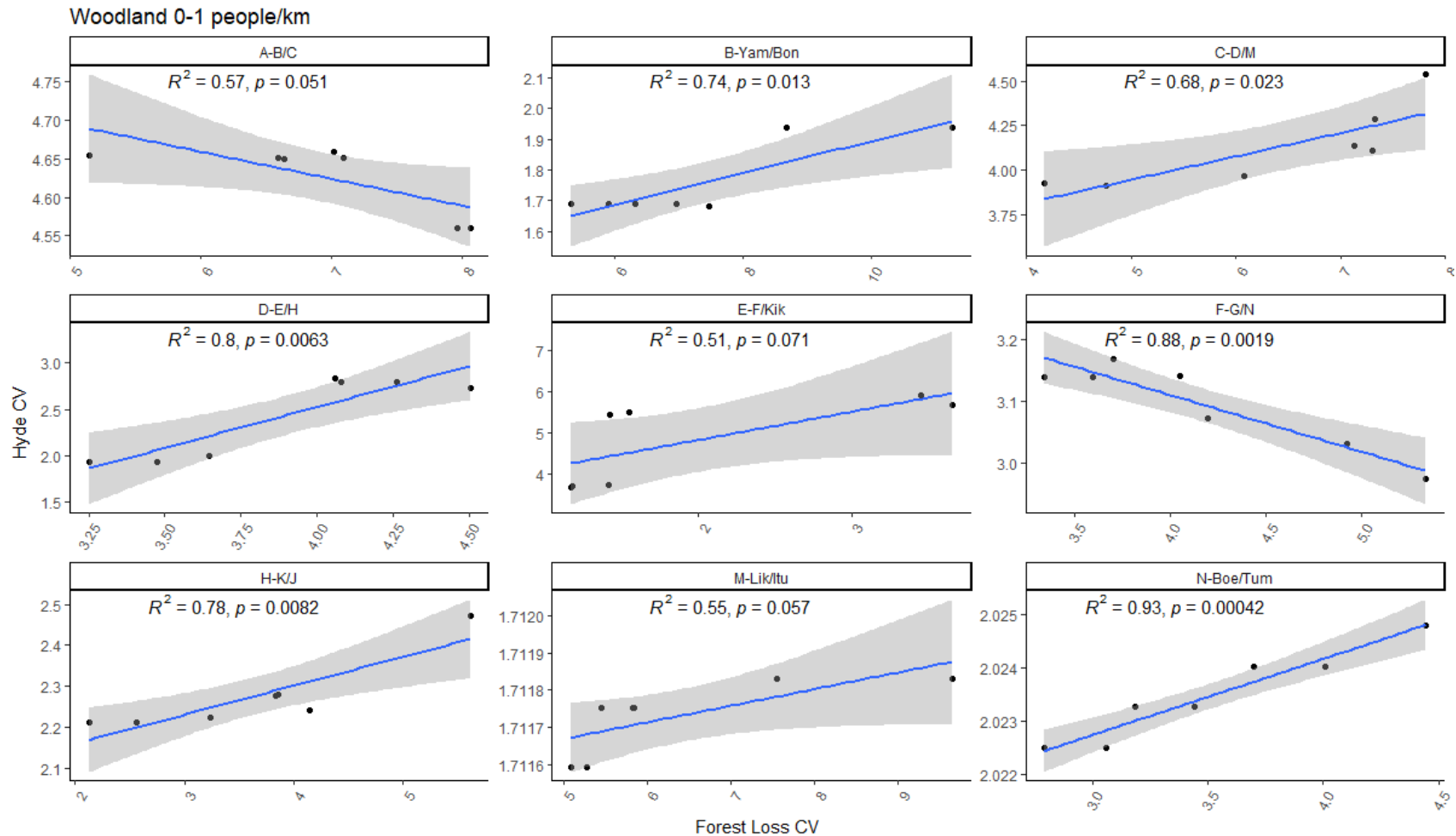
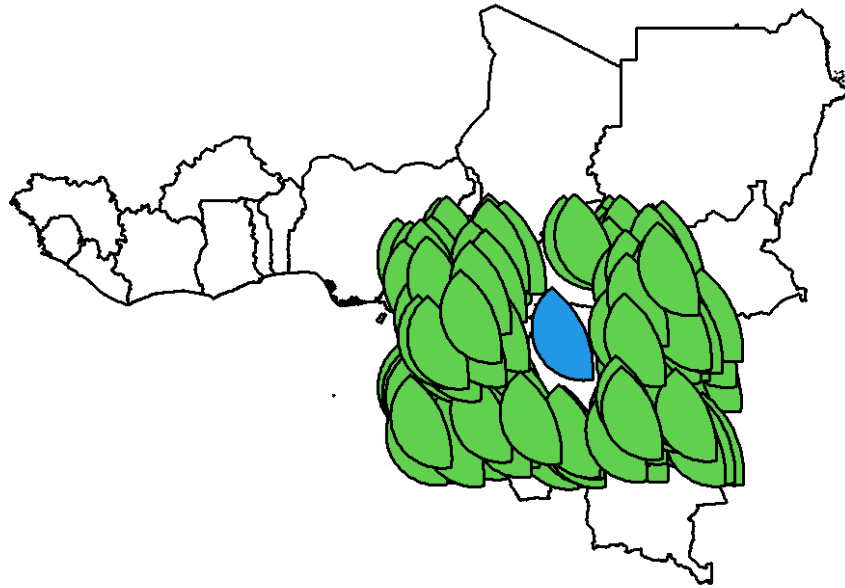


Figure S8. Comparison of fragmentation measures at coarse and fine spatial scales (continued). ... (c) HYDE Remote Semi-natural Scenario vs. GFW, with woodlands with human population density $> 1/\text{km}^2$ considered anthropogenic (Wald test, $P=0.012$), showing how increased human population densities in remote forests is associated with increased fragmentation at small spatial scales on average (and in 7 of 9 polygons, compared to 8 of 9 and 6 of 9 in (a) and (b), respectively). Statistical analyses were conducted using linear mixed-effects model in R. Each black dot represents both measures for a given year. Regression lines (blue) are shown with 95% confidence bands (gray).

a



b

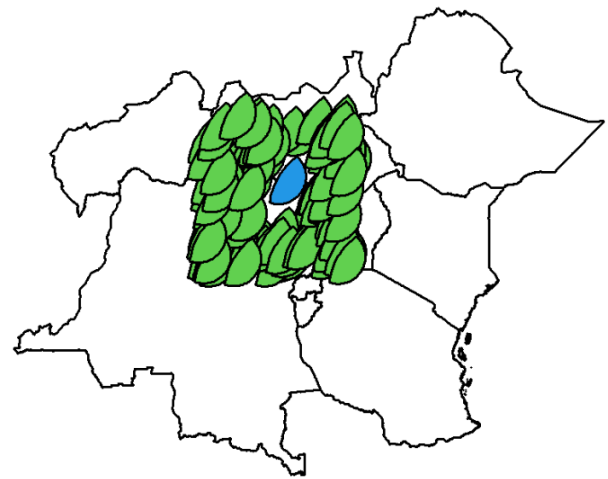
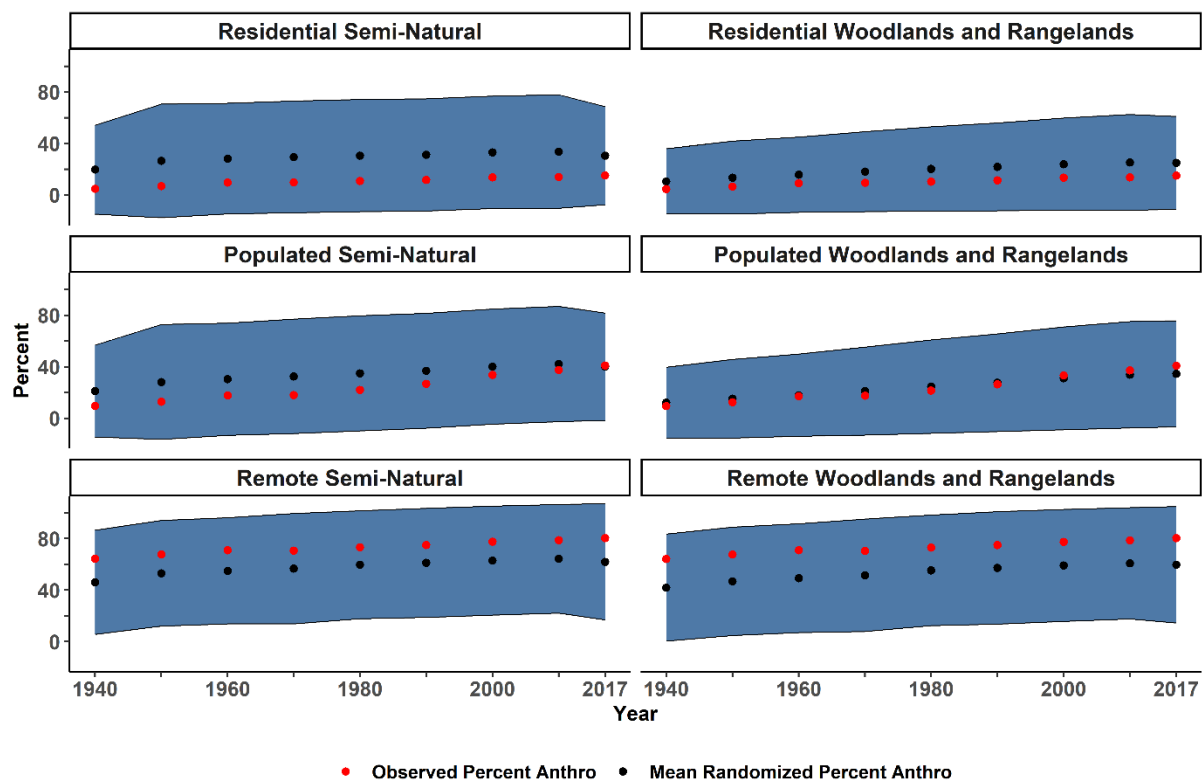


Figure S9. Comparison of origin polygons with 100 randomly placed controls. (a) Locations of 100 randomly placed polygons within the Congo basin of identical shape and size to the EBOV origin most recent common ancestor polygon. (b) Locations of 100 randomly placed polygons northeast of the Congo basin of identical shape and size to the SUDV origin most recent common ancestor polygon. The percentage of (c) anthropogenic land use and (d) anthropogenic fragmentation within the EBOV origin polygon (red dots) are not significantly different from the distribution of the 100 randomly placed polygons (blue band represents 95% confidence interval, black dots are the means), for any land use change scenarios. The (e) percentage of anthropogenic land use and (f) anthropogenic fragmentation within the SUDV origin polygon (red dots) are not significantly different than the distribution of the 100 randomly placed polygons (orange band represents 95% confidence interval, black dots are the means), for any land use change scenarios.

c



d

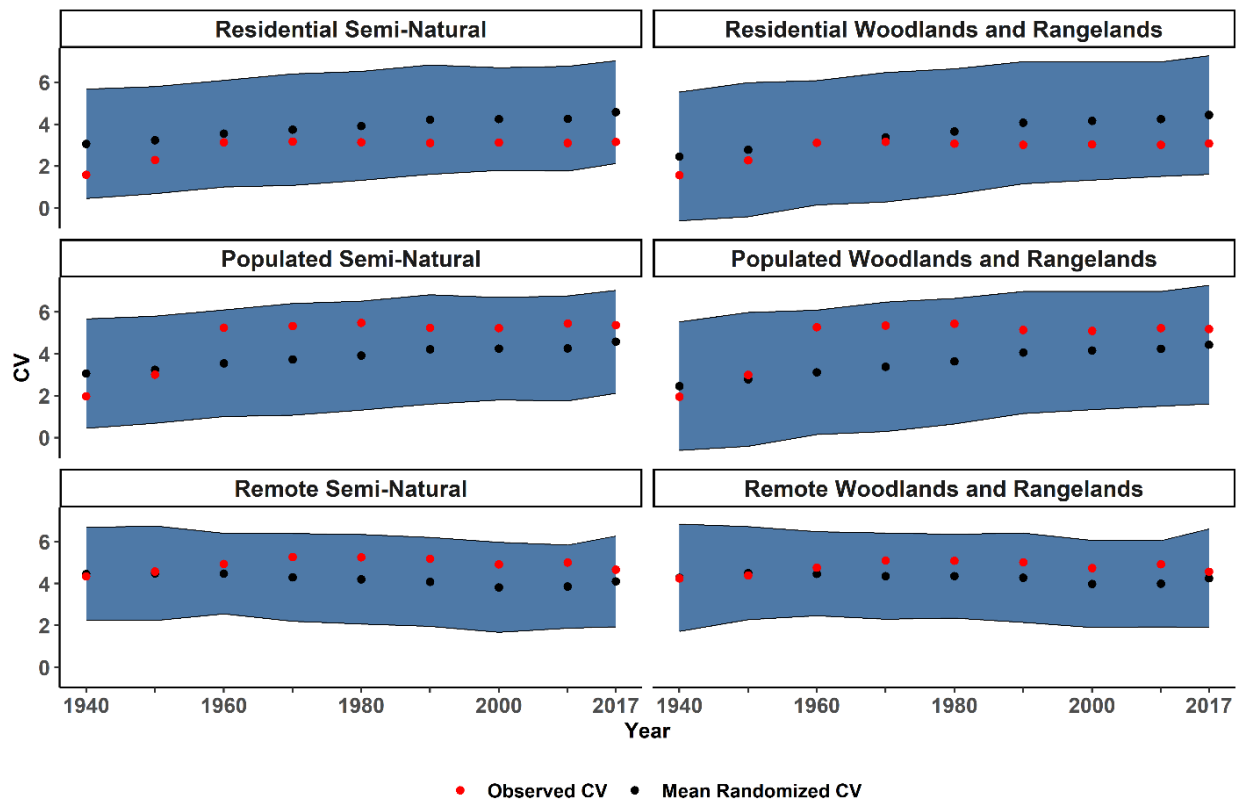
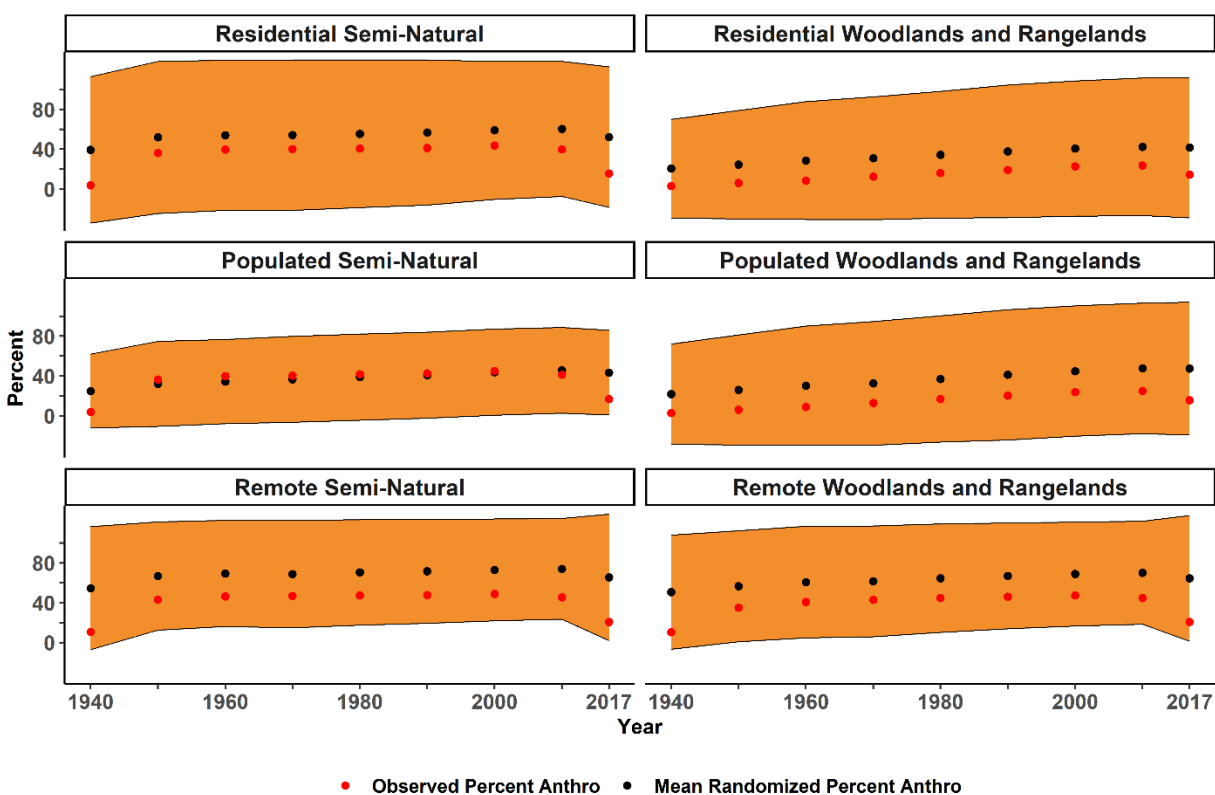


Figure S9 (continued). EBOV origin polygon randomization analysis.

e



f

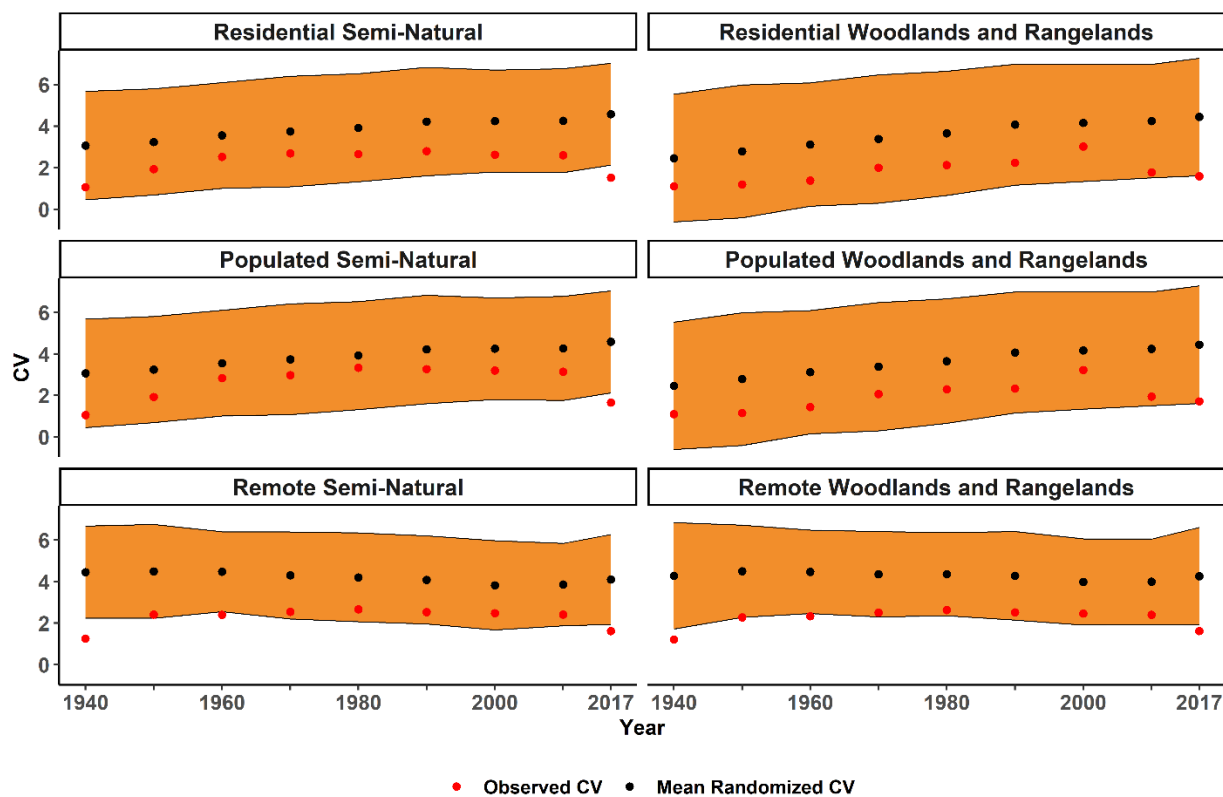


Figure S9 (continued). SUDV origin polygon randomization analysis.

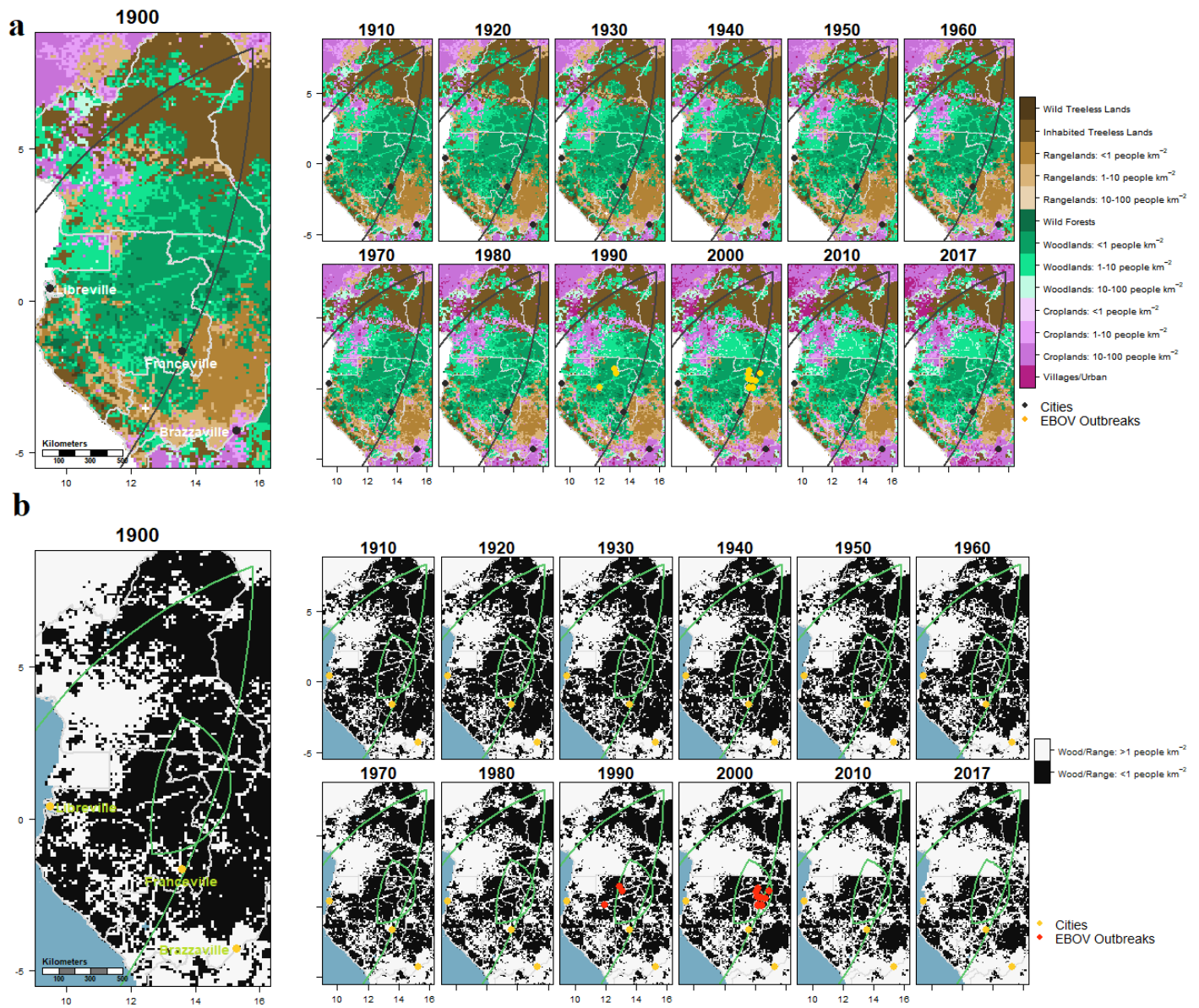


Figure S10. Time series of LUC on the Cameroon/Gabon/ROC border region. Two polygons, F-G/N 1987 (small) and J-Lue/Gui 1996 (large), show large spikes in LUC associated with EBOV outbreaks, with prominent fragmentation and increased human population density visible in southern Cameroon and on the Gabon/ROC border especially from 1970-2000 (see Fig. S7). Latitude and longitude are shown on the axes. (a) All anthrome classifications as defined in Klein Goldewijk et al. 2017. (b) The Remote Woodlands and Rangelands scenario, in which woodlands and rangelands > 1 person per km² are considered anthropogenic.

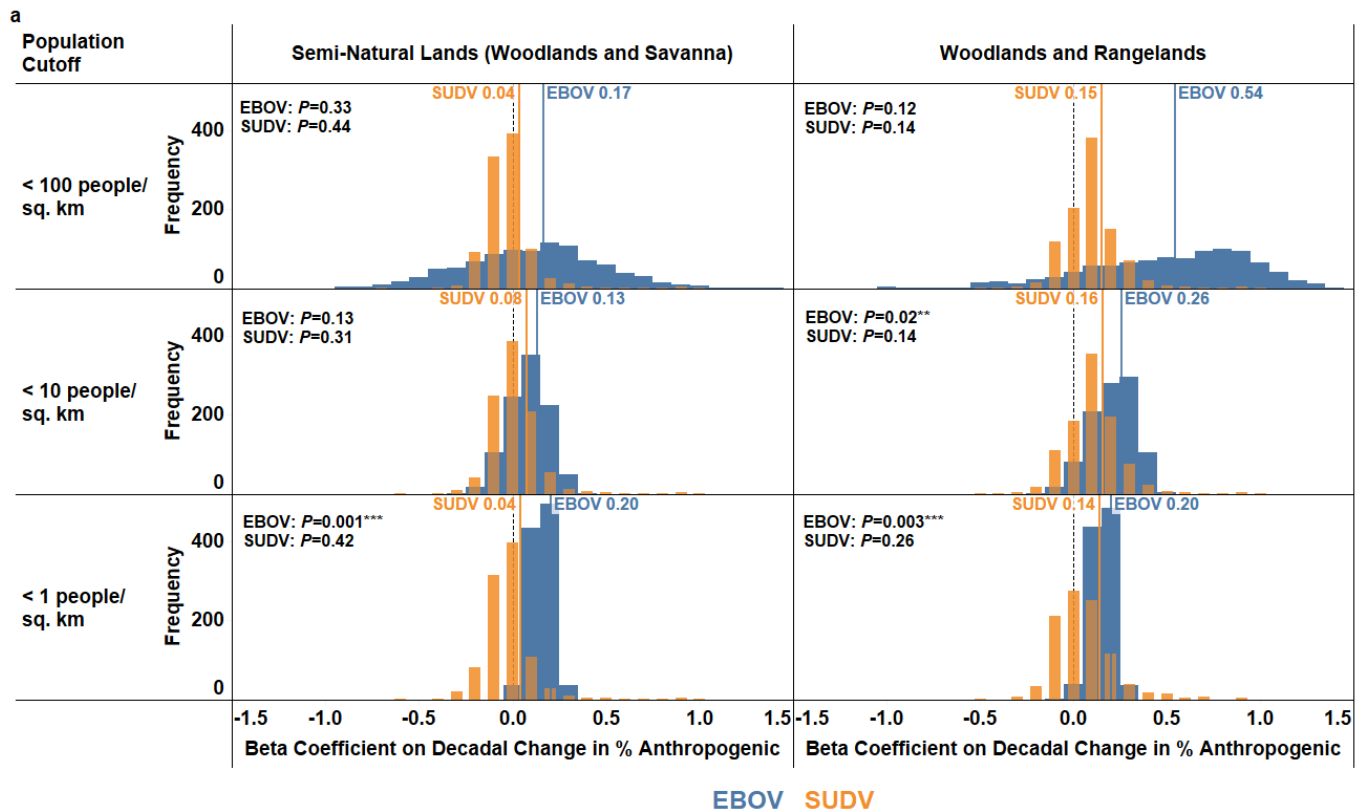


Figure S11. Sensitivity analysis of Monte Carlo simulation. The base model assessed through Monte Carlo simulation was varied in order to test the robustness of the association between land use change and viral evolution. In general, land use change was associated with increased *Ebolavirus* evolution under a range of model specifications. (a) In this model, coefficient of variation was removed, and the relationship between the percentage of land considered anthropogenic in each land use change scenario and viral substitution rate was assessed. (b) In this model, instead of viral substitution rate, a binary outcome variable (yes/no) was used, representing the emergence of a new variant, as described in Methods. (c) In this final model, the maximum of any type of fragmentation detected in a given polygon and decade was the independent variable (without differentiation by land use change scenario), and the outcome variable was the binary variable representing the emergence of a new variant. $*P<0.1$, $**P<0.05$, $***P<0.01$.

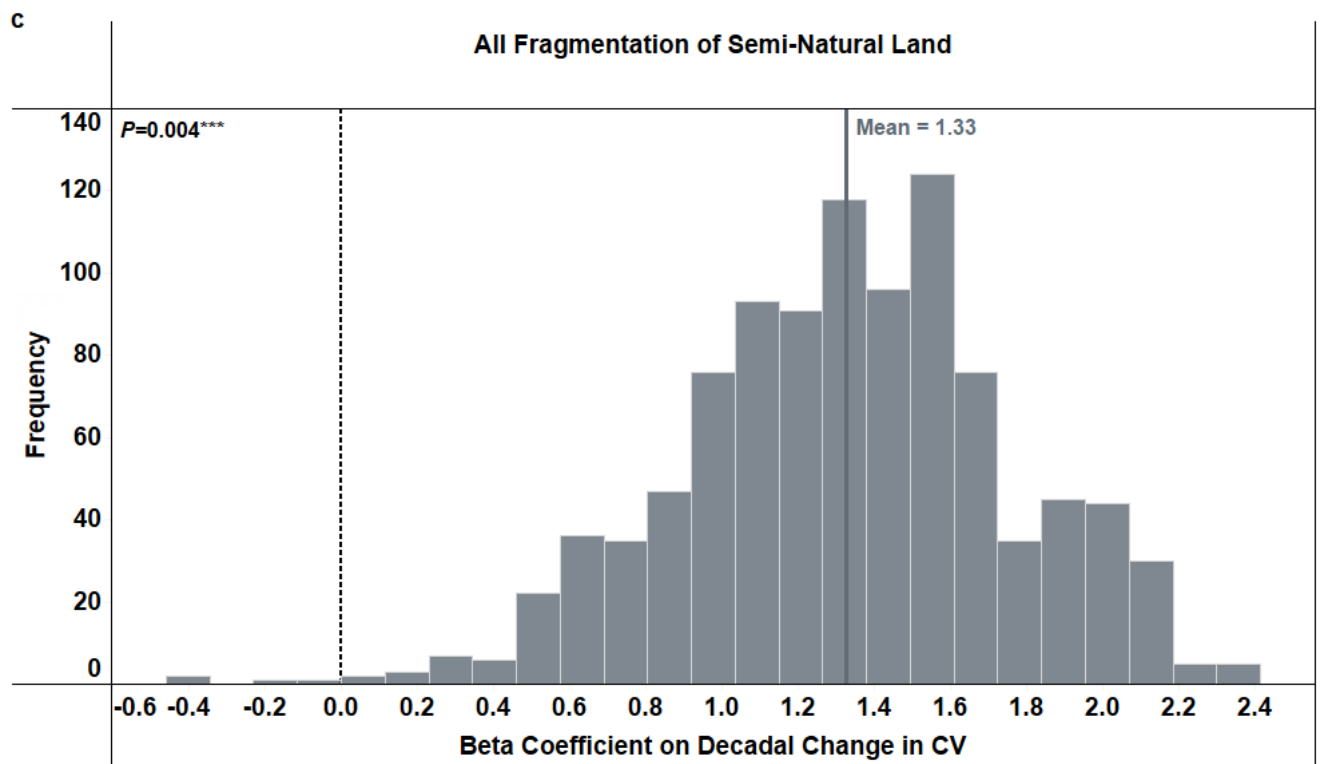
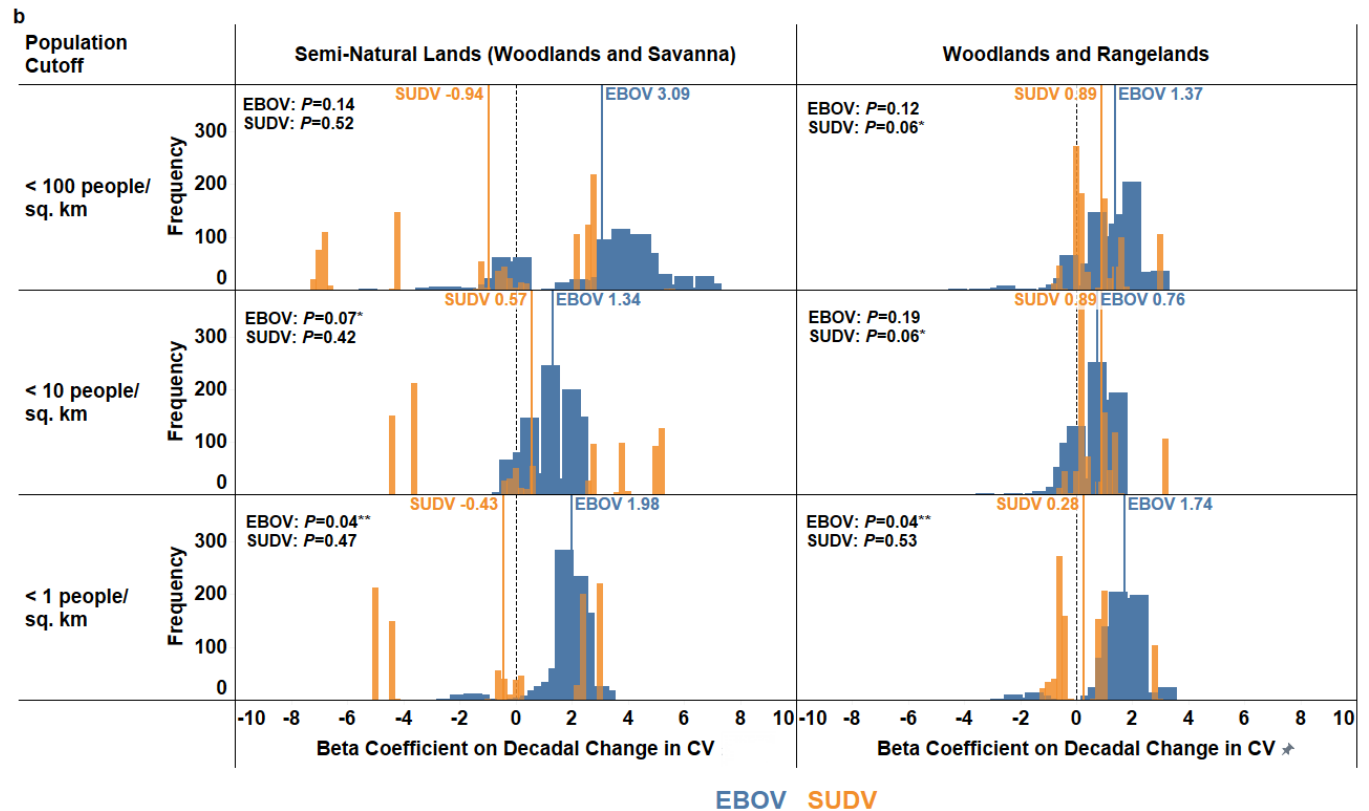


Figure S11 (continued). Sensitivity analysis of Monte Carlo simulation, logit outcome variable.

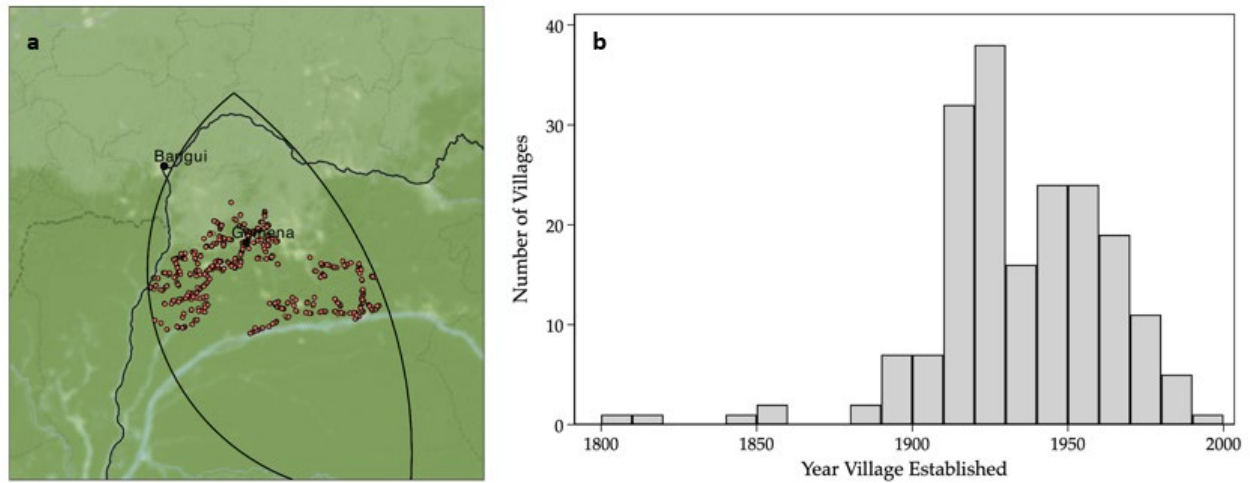


Figure S12. Results of interviews with village sages (knowledgeable elders) in 292 villages in Sud-Ubangi and Mongala provinces. (a) Location of villages surveyed within the EBOV origin polygon. (b) Histogram of year that sages reported their village was established. 55% of villages were established in the four decades prior to the calculated EBOV MRCA of 1960. The distribution is bimodal, with peaks in 1920 and 1950, corresponding to two documented changes in local land use policy.

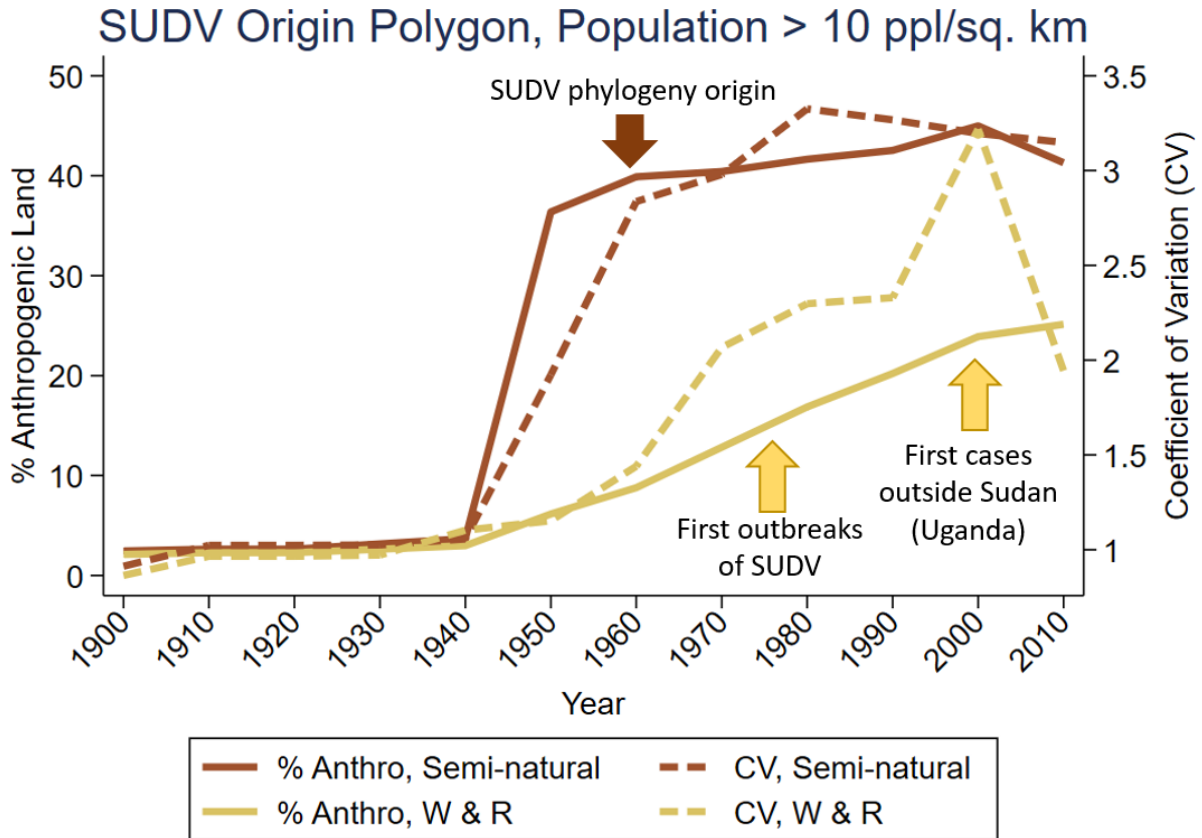


Figure S13. LUC in SUDV origin polygon is associated with important events in its phylogeographic history. Trends for percent land conversion (solid line) and coefficient of variation (CV, dashed line) for the SUDV origin polygon, for the populated semi-natural and populated woodlands and rangelands (W&R) scenarios, as defined in Table 1 and methods.

Table S1. Spatial spread regression analysis. Associations between distance traversed, phylogenetic distance, time since 1976, and virus species, using ordinary least squares regression with robust standard errors. Dependent variables for each model are the column headers, independent variables are rows, with beta coefficients and *P*-values (parentheses). The strongest association is between phylogenetic distance and distance traversed, for both EBOV and SUDV (column (2), (*P*<0.01)).

VARIABLES	(1) Distance (Thousands of km)	(2) Phylogenetic Distance from Origin (%)	(3) Distance (Thousands of km)
Sudan (Dummy Variable)	-0.552*** (0.00649)	-0.482 (0.288)	-0.166 (0.458)
Distance (Thousands of km)		0.639*** (0.000111)	
Distance*Sudan (Interaction)		5.513*** (3.71e-05)	
Years since 1976	0.0523*** (0.000541)	0.000706 (0.942)	0.0234* (0.0566)
Years since 1976*Sudan	-0.0299** (0.0296)	-0.00770 (0.731)	-0.0180 (0.159)
Phylogenetic Distance from Origin (%)			0.846** (0.0437)
Phylogenetic distance from Origin*Sudan (Interaction)			-0.716* (0.0824)
Constant	0.587*** (0.00313)	0.0474 (0.858)	0.230 (0.295)
Observations	20	20	20
R-squared	0.742	0.918	0.885
Adjusted R-squared	0.693	0.888	0.843

Robust pval in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table S2. Comparison of amino acid variation between GP and other *Ebolavirus* genes. Percentages of variation and *P*-values from two-tailed Chi-square tests with Yates' correction comparing the number of altered amino acids relative to the consensus sequence for GP with other genes in the *Ebolavirus* genome.

Variation (<i>P</i>-value)	GP	NP	VP35	VP40	L	VP30	VP24
EBOV	8.3% (ref.)	5.7% (0.054)	4.7% (0.036)	3.7% (0.007)	3.0% (<0.001)	2.8% (0.002)	2.4% (0.001)
SUDV	5.6% (ref.)	4.3% (0.27)	4.3% (0.36)	0.6% (<0.001)	2.1% (<0.001)	1.4% (0.003)	0.8% (0.001)

Table S3. Distributions for split years for each node in 1,000 trials used in the Monte Carlo simulation. As described in Methods, rejection sampling was used in the simulation to approximate the 95% HPD for each split, and to maintain the same order of splits as in the baseline tree (Fig. 1).

Virus	Node (Fig. 1)	Baseline Split Estimate	Monte Carlo Split Median	Baseline 95% HPD Range	Monte Carlo 99% CI
EBOV	A	1960	1959.6	(1932-1976)	(1923-1975)
EBOV	B	1975	1974.7	(1973-1976)	(1973-1977)
EBOV	C	1976	1976.5	(1962-1987)	(1960-1987)
EBOV	D	1978	1978.7	(1966-1989)	(1966-1988)
EBOV	E	1984	1983.5	(1973-1991)	(1974-1991)
EBOV	F	1987	1987.0	(1979-1992)	(1980-1992)
EBOV	H	1991	1991.0	(1979-1999)	(1978-1999)
EBOV	J	1996	1994.8	(1987-2004)	(1987-2002)
EBOV	M	2005	2003.9	(1993-2016)	(1992-2015)
EBOV	N	2007	2006.3	(1999-2014)	(1999-2013)
SUDV	A	1958	1957.3	(1887-1976)	(1913-1975)
SUDV	C	1976	1975.6	(1972-1979)	(1972-1979)
SUDV	D	1982	1981.5	(1975-1992)	(1975-1990)
SUDV	E	1991	1990.8	(1977-1999)	(1979-1999)
SUDV	F	1996	1996.2	(1989-2000)	(1989-2000)
SUDV	G	2003	2000.6	(1994-2011)	(1993-2009)

Table S4. Monte Carlo simulation results. Mean, standard deviation, and mean standard error across 1,000 trials for six LUC scenarios each, for the β coefficient measuring the spatiotemporal association between the decadal change in CV ($\beta_{\Delta CV}$) and the nucleotide substitution rate for EBOV and SUDV. β coefficients in this table are not exponentiated. There is a statistically significant relationship between the decadal change in CV and the rate of virus evolution across a range of LUC scenarios (distributions shown in Fig. 4). NS = Not statistically significant. MC = Monte Carlo. SE = Standard error.

EBOV, Decadal Change in CV

Scenario	mean(beta)	stdev(beta)	mean(stderr)	# trials w/beta < 0	P-val from MC betas	P-val from beta + SE	P-val from empirical MC	P-val from baseline data	Statistical Inference
Residential semi-natural	0.39	4.96	1.71	295	0.47	0.41	0.30	<0.01	NS
Populated semi-natural	1.37	0.82	0.48	25	0.05	<0.01	0.03	<0.01	** $p < 0.05$
Remote semi-natural	2.41	1.35	0.67	50	0.04	<0.01	0.05	<0.01	** $p < 0.05$
Residential woodlands and rangelands	1.59	1.90	1.23	117	0.20	0.10	0.12	0.09	NS
Populated woodlands and rangelands	0.85	0.85	0.58	72	0.16	0.07	0.07	0.06	* $p < 0.1$
Remote woodlands and rangelands	2.46	1.50	1.01	48	0.05	0.01	0.05	<0.01	** $p < 0.05$

SUDV, Decadal Change in CV

Scenario	mean(beta)	stdev(beta)	mean(stderr)	# trials w/beta < 0	P-val from MC betas	P-val from beta + SE	P-val from empirical MC	P-val from baseline data	Statistical Inference
Residential semi-natural	-1.29	5.59	2.91	627	0.59	0.67	0.63	<0.01	NS
Populated semi-natural	0.58	3.77	1.25	373	0.44	0.32	0.37	0.08	NS
Remote semi-natural	0.56	3.87	1.38	357	0.44	0.34	0.36	0.06	NS
Residential woodlands and rangelands	2.38	1.60	2.04	25	0.07	0.12	0.03	0.43	* $p < 0.1$
Populated woodlands and rangelands	2.26	1.64	1.28	22	0.08	0.04	0.02	0.47	** $p < 0.05$
Remote woodlands and rangelands	0.26	1.46	1.67	319	0.43	0.44	0.32	0.45	NS

References

- Brooks, Steve, Andrew Gelman, Galin Jones, and Xiao-Li Meng, eds. 2011. *Handbook of Markov Chain Monte Carlo*. 1st edition. Boca Raton: Chapman and Hall/CRC.
- Cui, James. 2007. “QIC Program and Model Selection in GEE Analyses.” *The Stata Journal* 7 (2): 209–20. <https://doi.org/10.1177/1536867X0700700205>.
- Dellicour, Simon, Rebecca Rose, Nuno R. Faria, Philippe Lemey, and Oliver G. Pybus. 2016. “SERAPHIM: Studying Environmental Rasters and Phylogenetically Informed Movements.” *Bioinformatics* 32 (20): 3204–6. <https://doi.org/10.1093/bioinformatics/btw384>.
- Food and Agriculture Organization. 2020. “FAOSTAT - Land Use.” Food and Agriculture Organization of the United Nations. FAOSTAT - Land Use. September 10, 2020. <http://www.fao.org/faostat/en/#data/RL/visualize>.
- Hansen, M. C., P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, et al. 2013. “High-Resolution Global Maps of 21st-Century Forest Cover Change.” *Science* 342 (6160): 850–53. Data accessed March 28, 2022. <https://doi.org/10.1126/science.1244693>.
- Heymann, D. L., J. S. Weisfeld, P. A. Webb, K. M. Johnson, T. Cairns, and H. Berquist. 1980. “Ebola Hemorrhagic Fever: Tandala, Zaire, 1977-1978.” *The Journal of Infectious Diseases* 142 (3): 372–76. <https://doi.org/10.1093/infdis/142.3.372>.
- Hin, Lin-Yee, and You-Gan Wang. 2009. “Working-Correlation-Structure Identification in Generalized Estimating Equations.” *Statistics in Medicine* 28 (4): 642–58. <https://doi.org/10.1002/sim.3489>.
- Johansson, Markus, Craig R. Primmer, and Juha Merilä. 2007. “Does Habitat Fragmentation Reduce Fitness and Adaptability?: A Case Study of the Common Frog (*Rana Temporaria*).” *Molecular Ecology* 16 (13): 2693–2700.
- Klein Goldewijk, Kees, A. Beusen, Jonathan Doelman, and E. Stehfest. 2017. “Anthropogenic Land Use Estimates for the Holocene; HYDE 3.2.” *Earth System Science Data Discussions*, 9: 927-53. <https://doi.org/10.5194/essd-9-927-2017>. Accessed November 28, 2020.
- Mylne, Adrian, Oliver J. Brady, Zhi Huang, David M. Pigott, Nick Golding, Moritz U. G. Kraemer, and Simon I. Hay. 2014. “A Comprehensive Database of the Geographic Spread of Past Human Ebola Outbreaks.” *Scientific Data* 1 (1): 140042. <https://doi.org/10.1038/sdata.2014.42>.
- R Core Team. 2019. “R: A Language and Environment for Statistical Computing.” Vienna, Austria: R Foundation for Statistical Computing. <https://www.r-project.org/>.
- “Stata Statistical Software: Release 16.” 2019. Stata Statistical Software. StataCorp.
- Tilman, David. 1996. “Biodiversity: Population Versus Ecosystem Stability.” *Ecology* 77 (2): 350–63. <https://doi.org/10.2307/2265614>.
- Walsh, P. D., R. Biek, and L. A. Real. 2005. “Wave-like Spread of Ebola Zaire.” *PLoS Biol* 3 (11): e371. <https://doi.org/10.1371/journal.pbio.0030371>.
- Wang, Han, Yi Shi, Jian Song, Jianxun Qi, Guangwen Lu, Jinghua Yan, and George F. Gao. 2016. “Ebola Viral Glycoprotein Bound to Its Endosomal Receptor Niemann-Pick C1.” *Cell* 164 (1–2): 258–68. <https://doi.org/10.1016/j.cell.2015.12.044>.

- Wang, Ming. 2014. "Generalized Estimating Equations in Longitudinal Data Analysis: A Review and Recent Developments." *Advances in Statistics* 2014 (December): e303728. <https://doi.org/10.1155/2014/303728>.
- Wang, Yishu, Orla Murphy, Maxime Turgeon, ZhuoYu Wang, Sahir R. Bhatnagar, Juliana Schulz, and Erica E. M. Moodie. 2015. "The Perils of Quasi-Likelihood Information Criteria." *Stat* 4 (1): 246–54. <https://doi.org/10.1002/sta4.95>.
- Wang, You-Gan, and L. Hin. 2010. "Modeling Strategies in Longitudinal Data Analysis: Covariate, Variance Function and Correlation Structure Selection." *Comput. Stat. Data Anal.* <https://doi.org/10.1016/j.csda.2009.11.006>.