Host-pathogen evolutionary signatures reveal dynamics of future invasions of vampire bat rabies

Supporting Online Information

DG Streicker, J Winternitz, D Satterfield, RE Condori-Condori, A Broos, C Tello, S Recuenco, A Velasco-Villa, S Altizer and W Valderrama

Contents:

S1 Supporting Materials and Methods

Sequence amplification of rabies viruses. Aliquots of livestock brains dissected from fresh incisions into the cerebellum and brainstem were either homogenized in PBS using sterile wooden swabs and applied to Whatman FTA cards or subjected to total RNA extraction using TRIzol™ (Invitrogen, Carlsbad, CA) according to the manufacturer's instructions. The complete nucleoprotein gene (1350 bp) was amplified using primers and RT-PCR cycling conditions described in Streicker *et al.* (1). Primers for the amplification and sequencing of the non-coding region between the glycoprotein and polymerase (G-L) are shown in Table S3. RT-PCR amplicons were purified with ExoSAP-IT® (USB Corporation) or excised and purified from low melting agarose gels using Wizard PCR Clean-up kits (Promega). Sequencing was carried out on an Applied Biosystems 3730xl DNA Sequencer. Chromatograms were edited, assembled and aligned in Geneious v. 6.1.5 (2). Virus sequencing was carried out at the US Centers for Disease Control and Prevention and the MRC-University of Glasgow Centre for Virus Research.

Collection, sequencing and genotyping of bat tissue samples. At each colony, bats were captured using mist nets, harp traps or insect nets and marked with uniquely numbered wing bands (Porzana Inc.). Two 2mm biopsies were collected from wing membranes and stored in 95% ethanol at 4°C until they could be frozen at -20°C. The University of Georgia's Institutional Animal Care and Use Committee approved protocols for the capture and handling of bats (AUP #A2009-10003-0). Genomic DNA was isolated from wing punches using DNeasy Tissue Kits (Qiagen) following the manufacturer's instructions. The complete mitochondrial cytochrome b (*CytB*) gene (1140 bp) was amplified by PCR using primers and protocols described in (3). PCR amplicons were purified with ExoSAP-IT® (USB Corporation) or excised and purified from low melting agarose gels using the PrepEase Gel Extraction Kit (USB Corporation). Sequencing was carried out on an Applied Biosystems 3730xl DNA Sequencer at the Georgia Genomics Facility at the University of Georgia using primers from Martins *et al.* (3) and an additional internal sequencing primer, DR341F (5' – GAAACATGAAACATCGGCATCC – 3'). Forward and reverse sequences were edited by eye, assembled and aligned using Geneious v. 6.1.5 (2). Nine microsatellite loci in two multiplex reactions were used to estimate the neutral nuclear genetic variation in vampire bat populations (Table S2). Loci are described as Panels A and B in (4); fluorescent dyes distinguished primers in each panel. We amplified microsatellite loci according

to conditions of Piaggio *et al*. (4). PCR products were electrophoresed at the Georgia Genomics Facility on an Applied Biosystems 3730xl 96-capillary DNA Analyzer. Genotyping of individuals was completed using Genemarker v2.4.0. We quantified genotyping error rate across loci by re-genotyping 10-18 samples per marker.

Examining patterns of nuclear genetic structure

Testing assumptions and null alleles among microsatellite loci

The presence of null alleles, allelic stuttering, and large allele dropout was tested using MicroChecker (5). Using *FreeNA* (6), we found three loci (*Dero B10, Dero C12*, and *Dero D06*) with relatively high mean frequencies of null alleles (0.429, 0.344, and 0.417, respectively). In some but not all colonies, these loci would not amplify in the majority of samples (Table S4-S5). MicroChecker also found evidence of null alleles, and found no indication of allele scoring errors (Table S4 and S5). Because null alleles can bias population genetic analyses by underestimating within-population differentiation and overestimating between-population differentiation, we checked the consistency of our analyses with and without null alleles (6, 7).

Genotypic linkage disequilibrium (LD) between all pairs of loci was tested using randomization methods with Bonferroni corrections in FSTAT v. 2.9.3.2 (8). Tests for departure from Hardy-Weinberg Equilibrium (HWE) were conducted on all colonies and loci within samples using the F_{IS} statistic in FSTAT. There were three pairs of loci showing significant LD ($p < 0.05$) after correction: *Dero C12* and *Dero D06*, *Dero C07* and *Dero H02*, and *Dero D12* and *Dero H02*. All but four colonies (CAJ4, HUA2, LMA10/11, and UCA3, out of 23 colonies) were found to be in HWE after Bonferroni corrections based on F_{IS}, and four loci (*Dero D06*, *Dero C07*, *Dero G10*, and *Dero H02*) significantly deviated from HWE (Table S3 and S4). Paired t-test showed that observed heterozygosity was significantly less than expected heterozygosity ($t = 5.21$, $d.f. = 8$, $p = 4.06 \times 10^{-4}$). In summary, specific loci showed evidence of LD and deviations from HWE expectations, mediated by high frequencies of null alleles (both visual homozygotes and absent null homozygotes). Null alleles appear in many taxa with high effective population size or in taxa with high intra-specific phylogenetic distance (6, 7). While null alleles can slightly overestimate F_{ST} in simulation studies, loci affected by null alleles can still be useful for population genetic studies (9). Therefore, both to account for effects of null alleles and to examine the degree to which patterns of genetic structure are influenced by locus-

specific biases, we ran all clustering analyses for two sets of loci: (i) all nine loci and (ii) six loci after removing the three with the highest frequency of null alleles (missing data treated as null homozygotes).

Number of genetic groups

Nuclear population structure was assessed using two classification methods STRUCTURE (a Bayesian clustering method) and DAPC (discriminant analysis of principal components), each considering 1-10 genetic groups (K). STRUCTURE identifies the number of subpopulations that minimize Hardy-Weinberg and linkage disequilibrium (10). STRUCTURE analysis was performed using the admixture model with correlated allele frequencies. Five simulations of $1x10^6$ iterations were run for each value of K, with a burn-in of 10,000 iterations. The most likely value of K was determined using the Evanno method with STRUCTURE HARVESTER (11).

DAPC uses sequential k-means and BIC model selection and was used to account for potential effects of isolation by distance (IBD, see below) on clustering methods and to avoid relying on assumptions underlying population genetics models. Analyses were conducted using the *adegenet* package in R, using both loci sets (12). We applied the *a.score* and *a.optim.score* functions to identify the optimal number of PCs to obtain the highest proportion of successful reassignment. We compared the mean a-score signal to noise ratios (mean/standard deviation) across each value of K. The a-score is the difference between probabilities of population assignment to true clusters versus random clusters, and was determined with 100 simulations for each value of K. One concern with k-means clustering approaches is that defining the appropriate number of clusters is subjective, relying on the minimum information criterion (BIC in this case), which tends to decrease monotonically until it flattens and/or rises as the number of clusters K increases. Typically, the "elbow" indicates the appropriate number of clusters, but this can be a shallow bend for stepping stone models of population structure. Therefore, to increase our confidence in predicted groups, we also analyzed prediction error using the prediction strength statistic, implemented in the R package *fpc* (13). Using this method we partitioned the dataset into half of individuals used 500 times for each value of K, ranging from two to ten. The prediction strength for each value of K is the minimum (taken over all clusters) relative frequency of individuals that cluster into the same groups for both halves of the data, and a value

greater than 0.8 is recommended (13). This analysis identified decreasing prediction strength with increasing values of K for 9 loci and suggested an optimal K = 3 for 6 loci (Figure S1). Therefore, K= 3 was the most supported group size. We assessed whether the K value with the highest prediction strength provides valid and stable clusters using 1000 bootstraps with the function *clusterboot*. Generally, a mean Jaccard similarity value of 0.75 or more is evidence of a stable cluster and 0.85 and above is "highly stable" (14). The most strongly supported number of groups (K = 3) had “stable/highly stable” clusters with Jaccard similarity values for both 6 and 9 loci (Table S6) while higher values of K had significantly lower stability (< 0.50).

Assessing and controlling for isolation by distance (IBD)

Mantel tests for patterns of IBD between colonies were significant with and without using the excluding null allele (ENA) correction (6) (ignoring missing data: $r = 0.351$, Mantel $p = 0.0003$; missing data as null homozygotes: $r = 0.335$, Mantel $p = 0.0003$; not corrected: $r = 0.349$, Mantel $p = 0.0001$). Local density plots showed some evidence for clustering of differentiated populations, but the general pattern was of a continuous cline of differentiation with distance (Fig. S6).

Because a gradient in genetic differentiation can bias inference of clusters using assignment tests (12), we looked for congruence between our clustering methods as a sign of true population delimitation. For both clustering methods, the Apurimac colonies (API1, API14, API3, API9, API13, and API138/140) consistently grouped together and were isolated from other populations. STRUCTURE analyses indicated the presence of two main clusters based on Evanno’s ΔK method (15), using 6 loci without the highest null allele frequencies and all 9 loci. Consistent with STRUCTURE results, when colonies were clustered by DAPC into the most strongly supported number of divisions (K = 3), the Apurimac colonies again consistently grouped apart from the other colonies, and this was observed using all nine or the subset of six loci (Fig. S3).

Population genetic tests of male-biased dispersal

Intense male-biased dispersal between our colonies was detected using the heterozygote deficiency test (F_{IS}, Table S7), which assumes a sample of the dispersing sex will include immigrant and resident males coming from populations with differing allele frequencies

(Wahlund effect). No other test was significant ($p > 0.0125$). In the total population ($N_F = 229$, $N_M = 243$) there were no differences in variance between males and females for expected heterozygosity (He) or observed heterozygosity (Ho) using Bartlett tests of homogeneity of variances (He: $K^2 = 0.0002$, *d.f.*= 1, $p = 0.997$; Ho: $K^2 = 0.0116$, *d.f.*=1, $p = 0.914$). Similarly, males and females did not differ in mean He or Ho using two-sample t-tests (He: $t = -0.048$, *d.f.*=16, $p = 0.963$; Ho: $t = -0.166$, *d.f.* = 16, $p = 0.871$), demonstrating that there were no underlying differences in heterozygosity between sexes.

Power analysis of microsatellites

To confirm that the weak population structure observed east of the Andes mountains signaled real gene flow rather than insufficient resolution in microsatellite markers, we tested the statistical power of 6 and 9 microsatellites to detect population structure in vampire bats. The resolution of our modest number of microsatellite loci (nine) was high enough to detect significant levels of genetic differentiation between populations ($F_{ST} = 0.195$, 95% CI = 0.146-0.248; Hedrick's (2005) standardized $G_{ST} = 0.387$, 95% CI = 0.269 - 0.512) with 1000 bootstraps using R package *diveRsity* (16, 17). At these F_{ST} and G_{ST} levels, STRUCTURE has been shown to correctly identify over 97% of individuals using 10 co-dominant loci in simulation studies (18). Power analyses were conducted by simulating the observed degree of differentiation using empirical values (number of populations, sample sizes, number of loci and alleles, and allele frequencies), with 1000 simulations for different effective population sizes (1000 to 3000) and generations under isolation (200 to 1000) using the software POWSIM 4.1 (19). Power calculations indicated we had 98 to 100% power to detect true effects using our six and nine loci sets. Empirical pairwise-Fst values between bat colonies was calculated using HIERFSTAT (20).

Time-series analysis of the seasonality of viral expansions. In vampire bats, reproduction occurs year round, but has a peak in the wet season (October through March) (21). It has been speculated that the new cohort of births forces male dispersal (22), which could create seasonal waves of viral dispersal across the landscape. We evaluated seasonal pulses in viral expansion using a database of 1146 laboratory-confirmed livestock rabies outbreaks recorded by the passive surveillance system of Peru from 2003-2014 (23). For each month of the time series, we estimated the infected area of the country using kernel densities (bandwidth = 0.15) using the

bkde2D function in the KernSmooth package of R. Seasonality in the infected area was assessed using seasonal decomposition with loess using the *stl* function in R. We determined the season of peak spatial incidence using a generalized additive model in the *mgcv* package with a non-parametric smoothers on month and year. These analyses show that the spatial extent of viral presence on the landscape is seasonal (Fig. S10), which could be attributed to seasonality in male dispersal, rabies incidence, reporting to the passive surveillance system or a combination of these factors. Additional data on rabies reporting patterns are currently being collected to disentangle these effects. We note also that while month is statistically significant, much variation in the time series is attributed to unidentified factors.

S2 Phylogenetic and epidemiological estimates of epidemic velocities

We calculated the velocity of advancing wavefronts of vampire bat rabies using phylogenetic and time-series methods with different underlying assumptions. The phylogenetic estimates of historical wavefront velocities were calculated by sampling rooted, time and latitude/longitude annotated phylogenies from the posterior distribution and extracting the inferred time and location of ancestral nodes (24, 25). Distances between inferred and observed nodes reflect viral dispersal in any direction, accounting for the possibility of indirect spatial pathways underlying advancing wavefronts. Projecting these velocities forward in time assumes directional viral dispersal along projected pathways, which appeared to be the case in the invasion front in northern Peru (Fig. 3B). The time series wavefront speed was by estimated by optimizing a regression of time since and distance to an inferred origin point for a series of outbreaks (26, 23). The origin point was inferred by randomly sampling 10,000 latitudes and longitudes within 10km of the most western reported outbreak in 2012 and refitting linear regressions. The slope from the hypothetical origin that maximized the r^2 value was used to calculate the epidemiological wavefront velocity (Fig. S9). This linear dispersal velocity (i.e., viral spread between points occurs without detours), likely underestimates the true distance traversed between observed outbreaks and therefore the potential velocity of the wavefront in areas without barriers to spread. Therefore, the time series estimates are the minimum speed required to explain the observed distribution of cases without detours, while the phylogenetic estimates should be treated as maximum potential wavefront velocities.

The spatial scale of inference also differed between phylogenetic and time series based methods. The phylogenetic estimate infers the mean velocity across branches of the phylogenetic trees representing viral dispersal between many localities, including endemic and epidemic areas. A strength of this approach is that it averages a wide range of possible velocities across different landscape types. In contrast, the time series estimate of velocity represents a single viral invasion in a relatively small area. In the phylogenetic context, this would be equivalent to a velocity estimated from a small cluster of tips spanning a relatively small area. The time series estimate therefore does not fully capture the potential for variability in epidemic velocities across time and space. To capture the potential spatial or temporal variability in velocities, we project on the phylogenetic estimates of velocity and present our point estimates of viral invasion along with the 95% highest posterior density on arrival dates. We focused on estimates from tip branches to reduce the impact of phylogenetic uncertainty at deeper nodes and to concentrate our estimates on more recent viral dispersal events, which seemed relevant considering that VBRV Lineage 1 appeared to be decelerating relative to an initial invasion front (Fig. 2D). The consistency of the estimates between VBRV lineage 1 and lineage 3 when using this approach further suggests that this approach is robust across diverse landscapes (Fig. 2E). We interpret the time series based estimates as confirmation of an ongoing spatial expansion in the predicted region and validation that the velocity falls within the range of those expected from historical data, noting that the time series estimate is a lower bound.

S3 Supporting figures

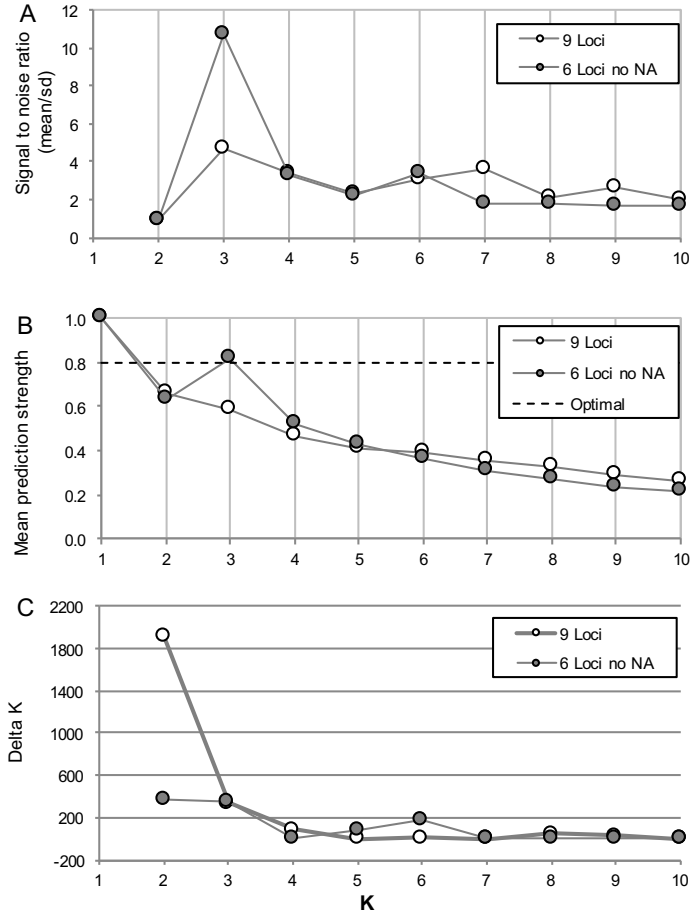

Figure S1. Evidence for different numbers of groups of vampire bats based on microsatellite data for DAPC and STRUCTURE analyses. A) Optimal number of DAPC clusters inferred by the highest signal to noise (mean/sd) ratio of a-scores calculated in Adegenet using the full dataset of 9 loci and a conservative dataset of 6 loci after excluding loci with null alleles. The a-score is the difference between probabilities of population assignment to true clusters versus random clusters, so that a higher a-score indicates stronger evidence of group assignment. B) Mean prediction strength for a given number of clusters (K) from one to ten for two sets of loci. "Optimal" line at 0.8 defines the suggested cut-off for choosing the optimal K (Tibshirani and Walther 2005). Taken together, both lines of evidence indicate that K of 3 is the most supported value from DAPC analyses. C) Plot of the Evanno method showing delta K for each set number of clusters in the STRUCTURE analysis. The most supported value for K is 2 for 9 loci and 2 or 3 for 6 loci after excluding loci with null alleles.

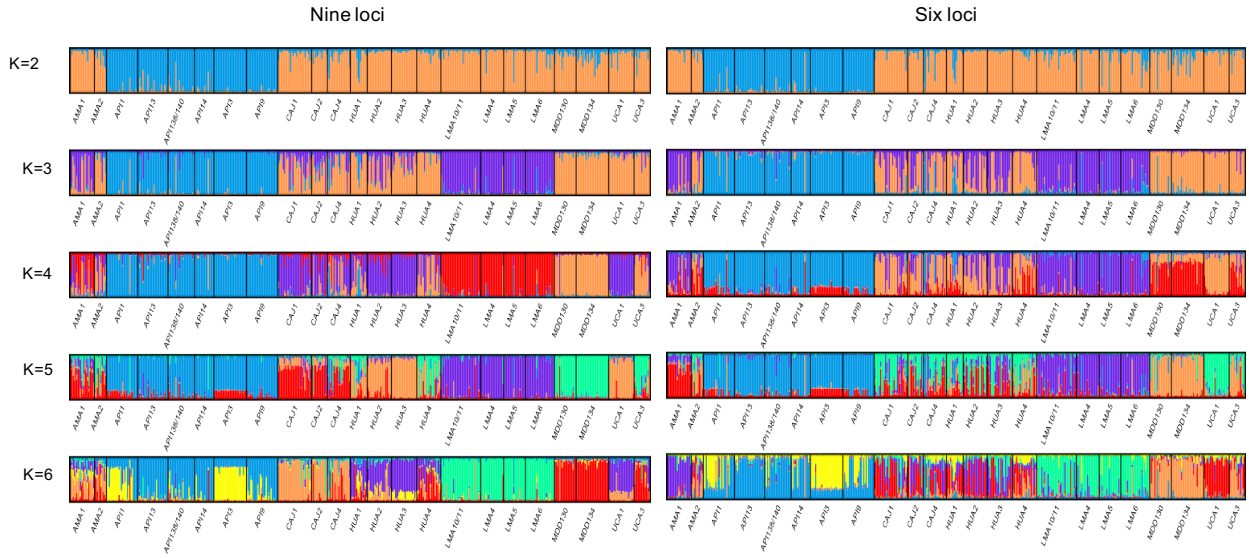

Figure S2. **Comparison of STRUCTURE results using 9 microsatellites and 6 microsatellites without null alleles from vampire bats.** Estimates from STRUCTURE analyses assuming K = 2 - 6 populations from 480 bats. Each bar represents the probability of membership assignment to each of K groups. Abbreviations correspond to sites in 1C.

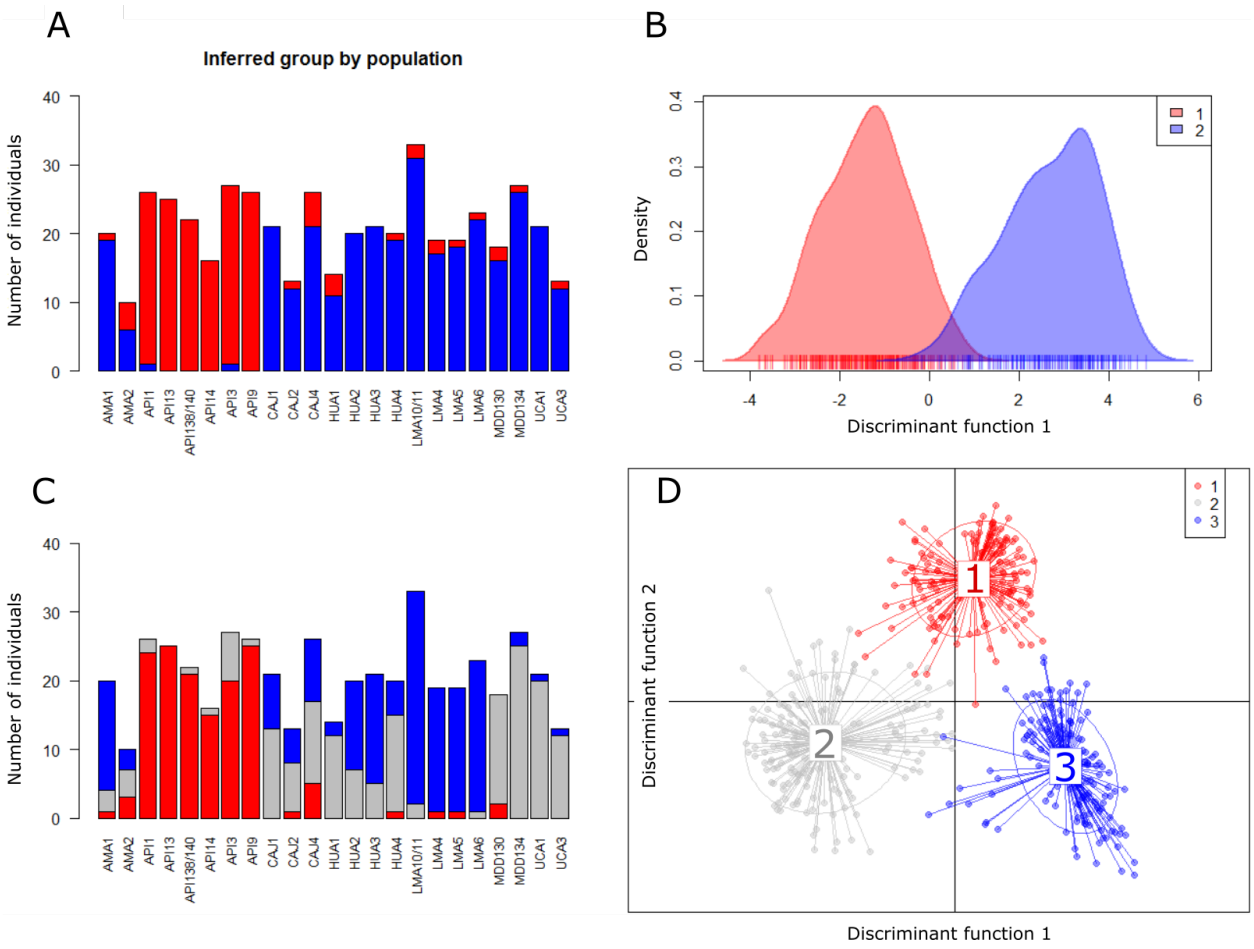

Figure S3. Discriminant Analysis of Principal Components (DAPC) results for vampire bat microsatellite data. Analyses were conducted with 9 microsatellite loci for the most probable number of groups (see Fig. S1). Top panels (A and B) show K=2 groups and bottom panels (C and D) show K = 3 groups. A and C show the proportion of individuals per population assigned to each group. B shows the density of individual assignment based on the first discriminant function. D shows individual assignments to groups based on discriminant function 1 and 2. Results indicate that the API colonies (in red) cluster apart from other colonies at the most likely values of K. As in the STRUCTURE analysis, bats from the Pacific coast (LMA) have a distinct genotype at K = 3 (here, blue), which was detected at low to moderate frequency in bats from northern and central Peru. Abbreviations correspond to sites in Fig. 1C.

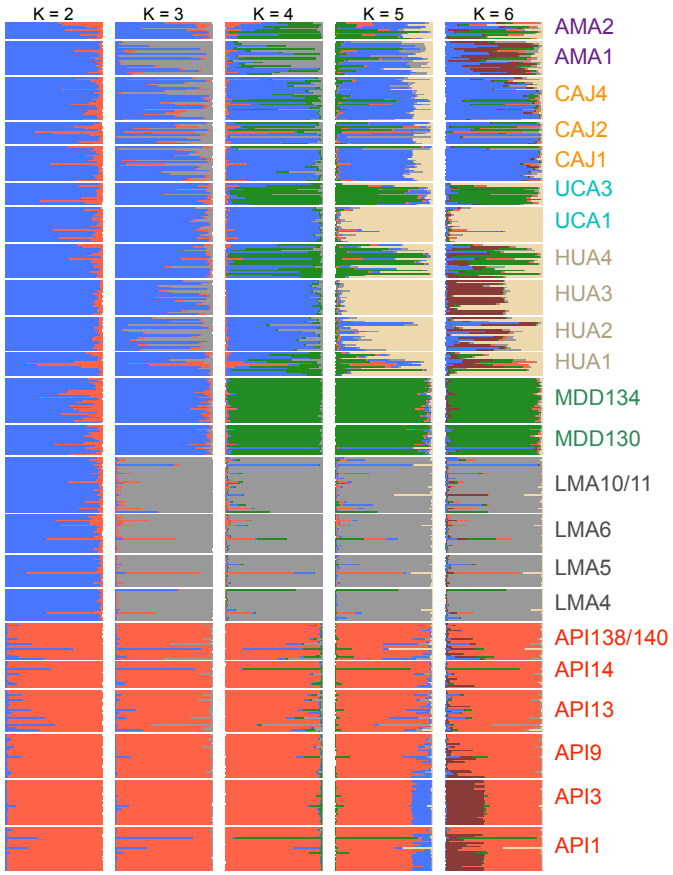

Figure S4. **STRUCTURE results without location prior.** Estimates from STRUCTURE analyses using the admixture model without location prior, assuming K = 2 - 6 populations from 480 vampire bats using 9 microsatellite loci. Each bar represents the probability of membership assignment to each of K groups. Results are qualitatively similar to STRUCTURE output using 9 and 6 loci with location prior (Fig. 1E).

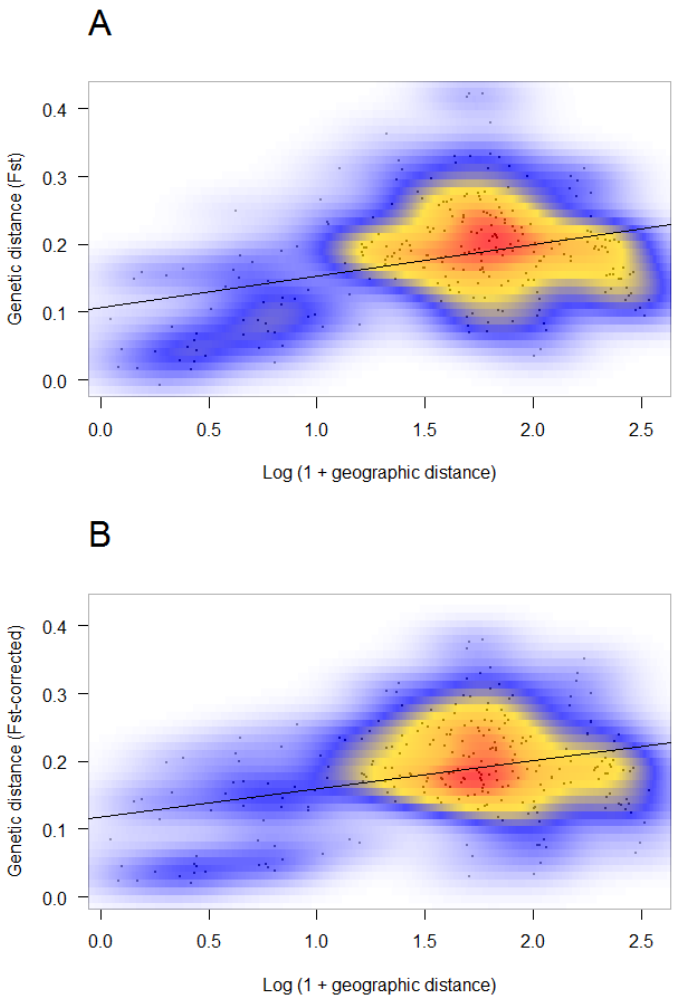

Figure S5. **Microsatellite isolation by distance (IBD) for 23 colonies visualized as a local density scatterplot.** (A) IBD plot for all nine loci and (B) six-locus set after removing loci with high frequency of null alleles. Mantel tests were performed with 10,000 permutations for nine loci: $r = 0.35$ [95% CI: 0.27 – 0.45], $p = 0.0001$ and six loci: $r = 0.34$ [95% CI: 0.27 – 0.44], $p = 0.0003$). Color scale indicates the density of observations (red>blue>white).

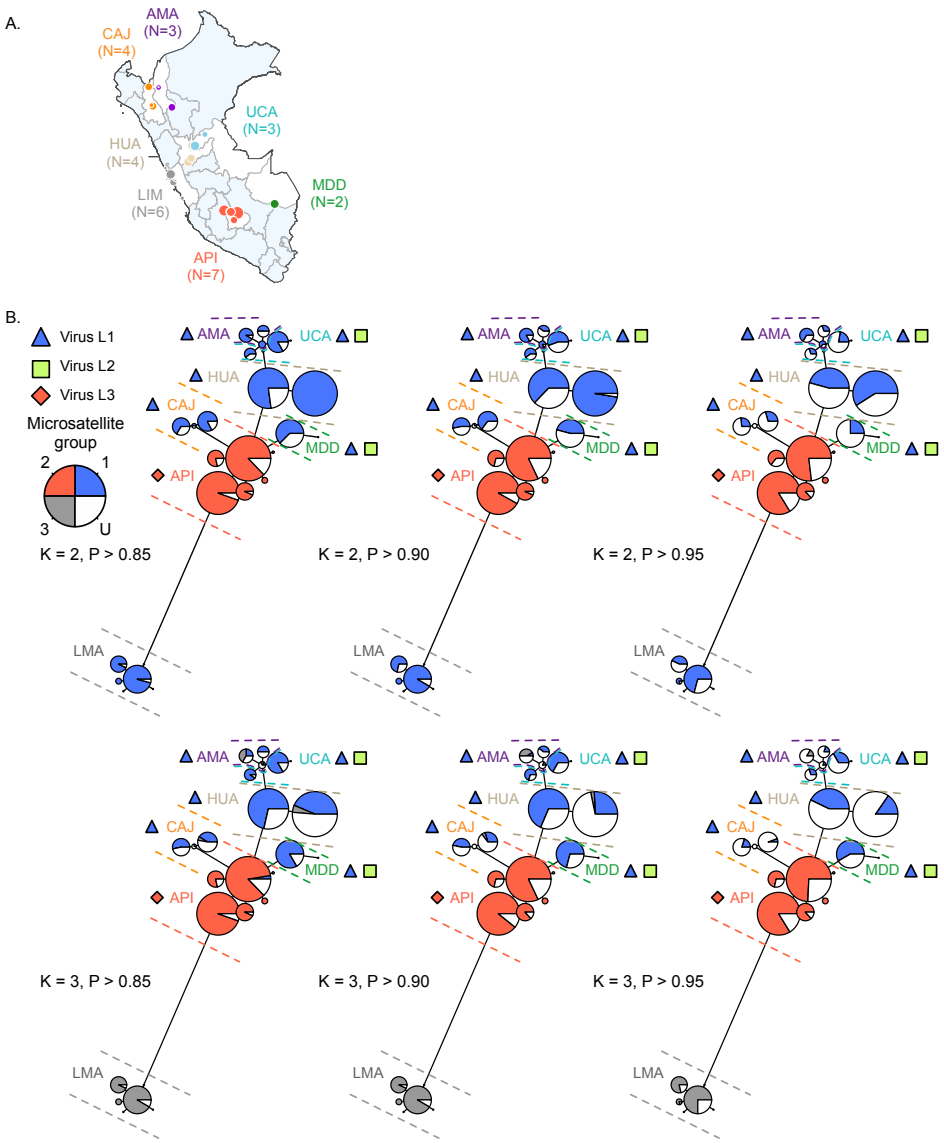

Figure S6. Mismatches between mitochondrial and nuclear genotypes in individual vampire bats. A) Sampling localities in Peru, with points scaled by numbers of individuals sequenced for mtDNA. B) Mitochondrial haplotype networks with node sizes proportionate to haplotype frequency. Pie charts are colored according to microsatellite group membership for individuals with both haplotype and microsatellite data (N = 352). Networks shows nuDNA group membership assuming K = 2 and 3 groups, under three cutoffs for the posterior probability of nuDNA group membership (PP > 0.85, 0.90, 0.95); unassigned individuals are colored white. Dashed lines indicate the geographic origins of sampled individuals, color coded as in A. Symbols indicate viral lineages present in the Departments of Peru where bats were sampled.

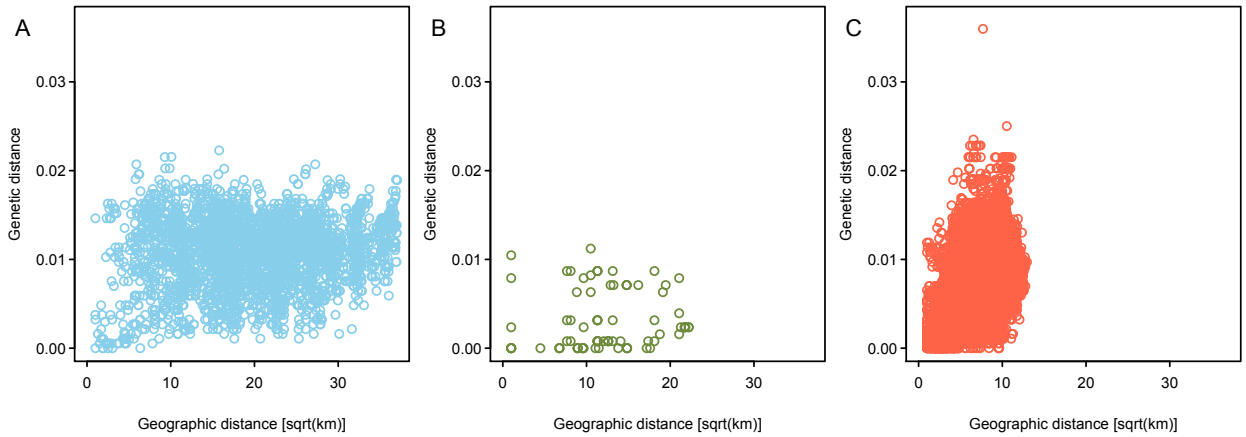

Figure S7. Genetic isolation by geographic distance in rabies virus lineages. A) Viral lineage 1 (Mantel tests; L1: $r = 0.12$ [95% CI: 0.08 – 0.16], $p = 0.008$). B) Viral lineage 2 ($r = 0.11$ [95% CI: -0.04 – 0.37], $p = 0.70$). C) Viral lineage 3 ($r = 0.52$ [95% CI: 0.49 – 0.55], $p = 0.001$).

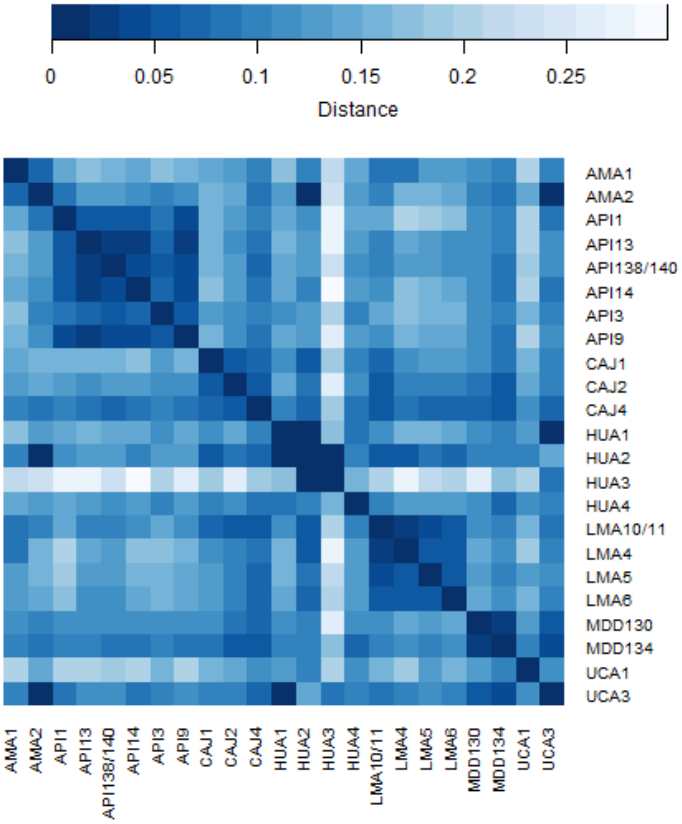

Figure S8. **Pairwise F_{st} heat map for 23 vampire bat populations using 9 microsatellite loci.** Pairwise genetic distance (as measured by F_{st}) between bat colonies, with darker shades indicating low population divergence. Blocks of very low pairwise genetic differentiation are visible with API and LMA sampling locations. Abbreviations correspond to sites in Fig. 1.

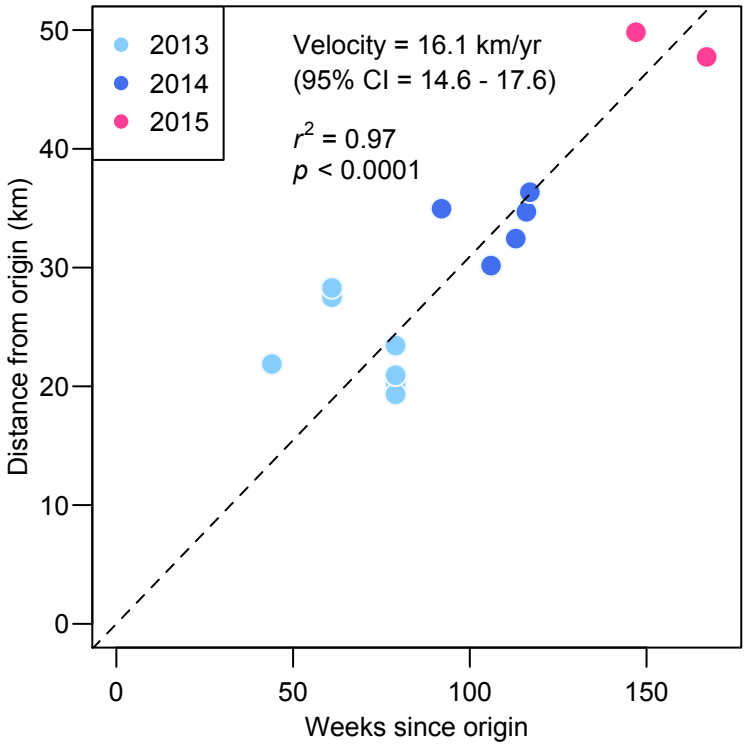

Figure S9. Linear regression used to calculate the wavefront velocity of a VBRV outbreak traveling west through the Huancabamba Depression towards the Pacific coast of Peru. Points (N = 15) are colored by year as in Fig. 3. The origin point for distance calculations was estimated by refitting the regression to 10,000 randomly sampled latitude and longitude coordinates within 10km of the the most western detected case in 2012. Regressions were forced through the origin.

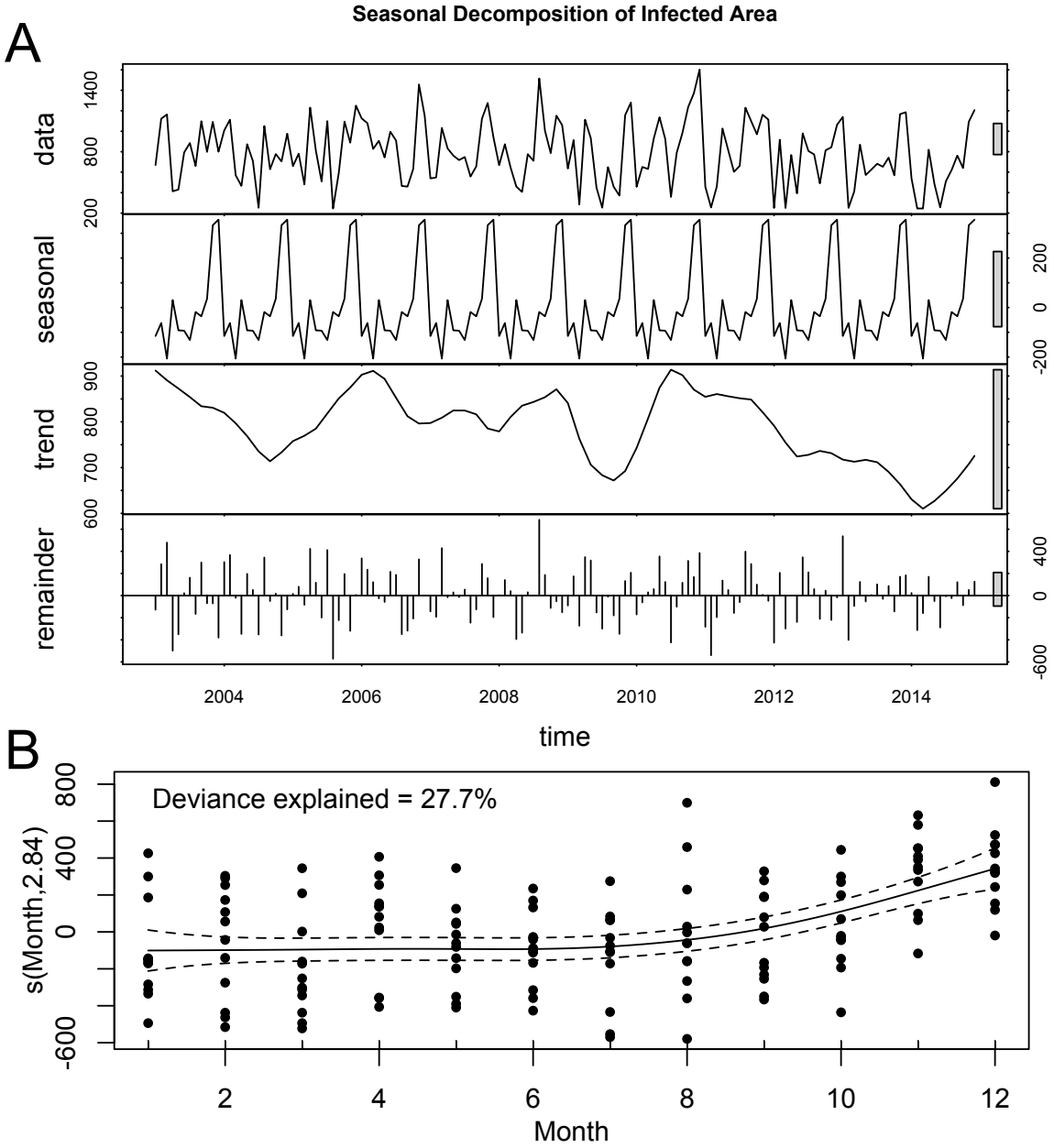

Figure S10. Seasonality in spatial expansions of vampire bat rabies. A) Seasonal decomposition of the monthly kernel density area infected with rabies from 2003 until 2014. Grey bars indicate the relative importance of each component (larger = less contribution). B) Monthly component of a generalized additive model of the area infected on the scale of the linear predictor. Dashed lines are 95% confidence intervals. The model showed a significant effect of month (edf = 2.84, $p < 0.0001$), but no strong effect of year (edf = 1.75, $p = 0.086$).

S4 Supplementary Tables

Table S1. Geographic locations of 29 vampire bat colonies used in population genetic analyses.

Site	Department	Latitude	Longitude
AMA1	Amazonas	-6.55	-77.38
AMA2	Amazonas	-5.22	-78.28
AMA3	Amazonas	-5.19	-78.29
API1	Apurimac	-13.45	-73.83
API13	Apurimac	-13.65	-72.92
API138	Apurimac	-14.09	-73.18
API14	Apurimac	-13.57	-73.46
API140	Apurimac	-14.14	-73.14
API3	Apurimac	-13.49	-73.81
API9	Apurimac	-13.59	-73.35
CAJ1	Cajamarca	-6.46	-78.68
CAJ2	Cajamarca	-6.41	-78.86
CAJ3	Cajamarca	-6.40	-78.77
CAJ4	Cajamarca	-5.17	-78.95
HUA1	Huanuco	-9.19	-75.96
HUA2	Huanuco	-10.22	-76.26
HUA3	Huanuco	-9.97	-76.07
HUA4	Huanuco	-9.12	-76.01
LMA10	Lima	-11.59	-77.28
LMA11	Lima	-11.59	-77.21
LMA4	Lima	-12.67	-76.67
LMA5	Lima	-10.64	-77.82
LMA6	Lima	-11.06	-77.46
LMA8	Lima	-11.47	-77.14
MDD130	Madre de Dios	-13.01	-70.65
MDD134	Madre de Dios	-13.05	-70.37
UCA1	Ucayali	-9.15	-75.79
UCA2	Ucayali	-8.43	-75.16
UCA3	Ucayali	-8.37	-75.12

Table S2. Molecular diversity of vampire bat colonies. For the mitochondrial DNA *cytochrome b* dataset: N, sample size; N_h, number of haplotypes; Private, percentage of haplotypes that occur exclusively at that location; H, haplotype diversity; and π, nucleotide diversity. For the microsatellite dataset: N, sample size; Na, number of alleles; Ho, observed heterozygosity; He, expected heterozygosity; Fis, index of heterozygote deficiency. Significant *p*-values after correction for multiple testing ($p < 0.0002$) in Fstat are in bold.

	Cytochrome b (mtDNA) diversity					**Microsatellite (nuDNA) diversity**					
Site	**N**	**Nh**	**Private (%)**	**H**	**π**	**N**	**Na**	**Ho**	**He**	**Fis**	**Fis p-value**
AMA1	18	2	100	0.471	1.882	20	35	0.4	0.455	0.196	0.0058
AMA2/3	16	1	100	0	0	10	35	0.4	0.466	0.203	0.0051
API1	18	4	25	0.601	1.092	26	35	0.485	0.428	-0.114	0.9872
API13	23	2	0	0.514	0.514	25	40	0.401	0.394	0.005	0.4543
API138/140	21	1	0	0	0	22	43	0.451	0.454	0.043	0.2249
API14	17	1	0	0	0	16	37	0.409	0.41	0.035	0.2814
API3	25	4	0	0.687	0.853	27	34	0.673	0.553	-0.189	0.9998
API9	21	3	0	0.338	0.581	26	32	0.444	0.42	-0.035	0.7415
CAJ1	19	1	0	0	0	28	47	0.49	0.536	0.112	0.0121
CAJ2	8	1	100	0	0	13	39	0.431	0.483	0.156	0.0159
CAJ3	11	1	0	0	0	1	NA	NA	NA	NA	NA
CAJ4	18	1	100	0	0	19	40	0.386	0.508	0.286	**0.0002**
HUA1	23	1	0	0	0	14	36	0.523	0.536	0.095	0.1493
HUA2	21	1	0	0	0	20	26	0.328	0.373	0.34	**0.0002**
HUA3	19	1	0	0	0	21	24	0.29	0.328	0.161	0.0572
HUA4	20	1	0	0	0	20	49	0.462	0.628	0.299	**0.0002**
LMA4	10	4	75	0.533	1	19	31	0.375	0.443	0.181	0.001
LMA5	8	3	33	0.536	0.536	19	36	0.406	0.503	0.22	0.0005
LMA6	22	1	100	0	0	23	34	0.497	0.497	0.022	0.3437
LMA8/10/11	28	1	0	0	0	33	41	0.44	0.492	0.122	0.0034
MDD130	8	2	0	0.25	0.5	18	55	0.609	0.629	0.063	0.0681
MDD134	20	2	50	0.1	0.7	27	77	0.731	0.705	-0.011	0.5932
UCA1	22	2	50	0.247	2.468	21	43	0.474	0.52	0.114	0.0085
UCA2/3	16	3	100	0.508	2.258	13	35	0.38	0.455	0.223	**0.0002**

Table S3. Oligonucleotide primers used in RT-PCR and sequencing of the rabies virus intergenic region between glycoprotein and polymerase (G-L).

Primer ID	Sense	Sequence	NT position †
G4765f	Forward	5' – CAAGCGAACAGAGTCRAGAAG - 3'	4,765-4,785
G4850f	Forward	5' – CATGGGAGTTRTACAAGAGTG - 3'	4,851 – 4,871
L5446R	Reverse	5' – CGATCCTATYGAGTCGGAAGC – 3'	5,445 – 5,465
L5492R	Reverse	5' – CCCAACATCTTAAGAAACTCTG – 3'	5,491 – 5,512

† Positions relative to a complete genome sequence of vampire bat rabies virus (Genbank AB519642)

Table S4. Sets of loci used for nuclear analyses. Set 1 includes all nine loci, Set 2 excludes loci with high frequencies of null alleles (NA*), calculated using FreeNA.

Set 1	**Set 2**	**NA***
Dero B03	Dero B03	0.018
Dero B10	-	0.429
Dero B11	Dero B11	0.016
Dero C12	-	0.344
Dero D06	-	0.417
Dero C07	Dero C07	0.081
Dero D12	Dero D12	0.049
Dero G10	Dero G10	0.135
Dero H02	Dero H02	0.116
Mean NA	**0.069**	**0.178**

*missing data as null homozygote

Table S5. Loci summary table. Na, number of alleles; Ho, heterozygosity observed; He, heterozygosity expected; NA, null allele frequency. Significant *p*-values after Bonferroni correction (α = 0.006) are in bold.

Locus ID	Na	Ho	He	NA*	Microchecker NA pops	FST	FST-ENA *corrected	HWE p-value
Dero B03	7	0.567	0.662	0.018	0	0.139	0.138	0.349
Dero B10	12	0.668	0.853	0.429	1	0.261	0.219	0.917
Dero B11	7	0.056	0.071	0.016	2	0.112	0.123	0.122
Dero C12	15	0.663	0.851	0.344	1	0.162	0.158	0.01
Dero D06	8	0.168	0.412	0.417	3	0.373	0.452	**0.001**
Dero C07	13	0.587	0.778	0.081	1	0.192	0.181	**0.005**
Dero D12	16	0.681	0.715	0.049	0	0.106	0.104	0.99
Dero G10	18	0.592	0.777	0.135	3	0.141	0.144	**0.001**
Dero H02	17	0.297	0.580	0.116	11	0.281	0.271	**0.001**

*missing data treated as null homozygous in FreeNA

Table S6. Cluster stability inferred from mean clusterwise Jaccard bootstrap (1000 resampling runs) for two sets of loci. Values of K for which mean Jaccard similarity value indicates cluster stability (0.75) or high stability (0.85), as recommended by Hennig (14), are in bold.

Loci Set	K value	Cluster 1	Cluster 2	Cluster 3
9 Loci	K = 2	0.65	0.68	
	K = 3	**0.88**	**0.86**	**0.76**
6 Loci no NA	K = 2	0.54	0.54	
	K = 3	**0.94**	**0.94**	**0.95**

Table S7. Tests for sex-biased dispersal implemented in Fstat. Each test used 10,000 randomizations using 1-tailed t-tests because we predicted females to be philopatric. The dataset included 23 populations, which contained at least 2 males and 2 females for at least 6 informative loci and no missing data from the 6 loci set, N = 422 individuals). Significant p-values after Bonferroni correction ($\alpha = 0.05/4 = 0.0125$) are in bold. F_{IS}, excess homozygotes, F_{ST}, differentiation; mAIc, mean assignment index test; vAIc, variance in assignment index test. We predicted males to have: (1) higher F_{IS}, (2) lower F_{ST}, (3) lower mAIc and (4) higher vAIc (27).

Category	**F_{IS}**	**F_{ST}**	**mAIc**	**vAIc**
Female (N=206)	-0.0003	0.1745	0.0717	5.734
Male (N=216)	0.1363	0.1630	-0.0684	5.545
Overall (N=422)	0.0719	0.2707		
p-value	**0.0004**	0.4854	0.2696	0.6532

Table S8. Distribution of vampire bat rabies lineages livestock by Peruvian Department and year. Dog rabies virus variants detected in livestock in the Departments of Madre de Dios (N = 1) and Puno (N = 3) are excluded. Cells are colored according to viral phylogenetic lineage. Cells show the number of cases typed using nucleoprotein sequences.

	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Amazonas	0	0	1	0	0	1	0	0	5	2	0	0	3	5	5	2
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Apurimac	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	4	0	5	7	13	3	5	2	22	10	10	6	39	45	19
Ayacucho	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	2
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	12	9	9	14	3	7	9	5	6	6	3	3
Cajamarca	0	0	0	0	0	0	0	1	0	0	1	2	3	1	3	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cusco	0	0	0	0	0	0	1	1	0	0	3	0	0	0	0	0
	0	0	0	0	0	0	1	0	0	0	4	0	0	0	0	0
	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
Huanuco	0	1	0	3	1	1	2	1	0	0	2	2	3	0	2	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Junin	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Lambayeque	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
Loreto	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Madre de Dios	0	0	0	0	0	0	2	7	1	0	0	0	0	0	0	0
	0	0	0	0	0	0	1	0	0	0	4	0	1	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Pasco	0	0	1	0	0	0	1	1	2	0	1	0	0	0	0	1
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Puno	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
San Martin	1	0	1	0	1	5	1	6	0	2	3	0	1	5	7	5
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
Ucayali	0	0	0	1	0	0	2	0	1	0	0	4	0	1	2	1
	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Rabies virus lineage 1 Rabies virus lineage 2 Rabies virus lineage 3

S5 Supporting references

1. Streicker DG, et al. (2010) Host phylogeny constrains cross-species emergence and establishment of rabies virus in bats. *Science* 329(5992):676–679.

2. Drummond AJ, et al. (2011) Geneious, version 5.4. *Geneious, Auckland, New Zeal*.

3. Martins FM, Templeton AR, Pavan ACO, Kohlbach BC, Morgante JS (2009) Phylogeography of the common vampire bat (Desmodus rotundus): Marked population structure, Neotropical Pleistocene vicariance and incongruence between nuclear and mtDNA markers. *BMC Evol Biol* 9(294). doi:10.1186/1471-2148-9-294.

4. Piaggio AJ, Johnston JJ, Perkins SL (2008) Development of polymorphic microsatellite loci for the common vampire bat, Desmodus rotundus (Chiroptera : Phylostomidae). *Mol Ecol Resour* 8(2):440–442.

5. Van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P (2004) MICRO-CHECKER: Software for identifying and correcting genotyping errors in microsatellite data. *Mol Ecol Notes* 4(3):535–538.

6. Chapuis MP, Estoup A (2007) Microsatellite null alleles and estimation of population differentiation. *Mol Biol Evol* 24(3):621–631.

7. Dakin EE, Avise JC (2004) Microsatellite null alleles in parentage analysis. *Heredity (Edinb)* 93(5):504–9.

8. Goudet J (2013) FSTAT: a computer program to calculate F-Statistics. *J Hered* 104:586–590.

9. Carlsson J (2008) Effects of microsatellite null alleles on assignment testing. *J Hered* 99(6):616–623.

10. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155(2):945–959.

11. Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv Genet*

Resour 4(2):359–361.

12. Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet* 11(1):94.

13. Tibshirani R, Walther G (2005) Cluster Validation by Prediction Strength. *J Comput Graph Stat* 14(December 2013):511–528.

14. Hennig C (2007) Cluster-wise assessment of cluster stability. *Comput Stat Data Anal* 52(1):258–271.

15. Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Mol Ecol* 14(8):2611–2620.

16. Keenan K, Mcginnity P, Cross TF, Crozier WW, Prodöhl PA (2013) DiveRsity: An R package for the estimation and exploration of population genetics parameters and their associated errors. *Methods Ecol Evol* 4(8):782–788.

17. Hedrick PW (2005) A Standardized Genetic Differentiation Measure. *Evolution (N Y)* 59(8):1633–1638.

18. Latch EK, Dharmarajan G, Glaubitz JC, Rhodes Jr OE (2006) Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. *Conserv Genet* 7(2):295–302.

19. Ryman N, Palm S (2006) POWSIM: A computer program for assessing statistical power when testing for genetic differentiation. *Mol Ecol Notes* 6(3):600–602.

20. Goudet J, Jombart T (2005) HIERFSTAT, a package for R to compute and test hierarchical F-statistics. *Mol Ecol Notes* 5(1):184–186.

21. Lord RD (1992) Seasonal Reproduction of Vampire Bats and Its Relation to Seasonality of Bovine Rabies. *J Wildl Dis* 28(2):292–294.

22. Wilkinson GS (1985) The social organization of the common vampire bat: II Mating system, genetic structure and relatedness. *Behav Ecol Sociobiol* 17:123–134.

23. Benavides JA, Valderrama W, Streicker DG (2016) Spatial expansions and travelling waves of rabies in vampire bats. *Proc R Soc B Biol Sci* 283(1832). doi:10.1098/rspb.2016.0328.

24. Pybus OG, et al. (2012) Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proc Natl Acad Sci* 109(37):15066–15071.

25. Dellicour S, Rose R, Pybus OG (2016) Explaining the geographic spread of emerging viruses: a new framework for comparing viral genetic information and environmental landscape data. *BMC Bioinformatics* 17(82):1–12.

26. Walsh PD, Biek R, Real LA (2005) Wave-like spread of Ebola Zaire. *PLoS Biol* 3(11):1946–1953.

27. Goudet J, Perrin N, Waser P (2002) Tests for sex-biased dispersal using bi-parentally inherited genetic markers. *Mol Ecol* 11(6):1103–1114.