ArHa: A globally standardised, relational database of arenavirus and hantavirus sampling in wild small mammal populations

David Simons[1], Ricardo Rivero[2], Grant Rickard[2], Ana Martinez-Checa[3], Harry Gordon[4], Gregory C. Milne[3], David W. Redding[3], Stephanie N. Seifert[2]

[1] Department of Anthropology, Pennsylvania State University, State College, Pennsylvania, USA [2] Paul G. Allen School of Global Health, Washington State University, Pullman, Washington, USA [3] Science Department, Natural History Museum, London, UK [4] School of Biological Sciences, University of East Anglia, Norwich, UK

Corresponding Authors: David Simons (dzs6259@psu.edu) and Stephanie N. Seifert (stephanie.seifert@wsu.edu)

1. Abstract

Arenaviruses and hantaviruses are globally distributed zoonotic pathogens primarily hosted by small mammals. Efforts to map their ecological drivers and predict spillover risk are hindered by the fragmentation of historical surveillance data across the primary literature, compounded by shifting host nomenclature and variable geographic precision. We present ArHa, a database of 123,101 host occurrence and pathogen diagnostic records derived from a systematic review of published wild small mammal sampling. The dataset standardises host and pathogen taxonomies, resolves records of the same animals published in more than one paper, and classifies the spatial resolution and reliability of each coordinate to support bias-aware spatial epidemiology. To accurately preserve complex study designs, such as multiple diagnostic assays or genetic sequences derived from a single host, ArHa is structured as a four-tier relational database. This architecture handles many-to-one relationships and supports linkage without data duplication. The dataset is distributed as compressed Apache Parquet files for programmatic analysis and as a Darwin Core Archive, so that it can be read by standard biodiversity data tools. An accompanying interactive web application, the ArHa Database Explorer, allows users to filter, visualise, and export subsets of the database without requiring programming expertise. All data, extraction scripts, and the interactive web application are openly accessible via a permanent Zenodo repository.

1.1 Keywords

Hantaviridae; Arenaviridae; rodent-borne zoonoses; wildlife disease surveillance; systematic review; relational database; spatial epidemiology; Darwin Core

2. Background & Summary

Arenaviruses and hantaviruses represent substantial global zoonotic threats, primarily circulating within small mammal reservoirs of the orders *Rodentia* and *Eulipotyphla* (Jonsson et al. 2010; Childs and Peters 1993). The frequency and geographic extent of zoonotic spillover are driven by complex interactions between climate, land-use change, and host community ecology (Keesing et al. 2010; Yanagihara et al. 2014; McMahon et al. 2018). Predictive spillover modelling and spatial epidemiology depend on robust, broad-scale ecological data to parameterise these environmental drivers (Kreuder Johnson et al. 2015; Olival et al. 2017; Gibb, Albery, Becker, et al. 2021). However, the historical surveillance data required for such analyses remain highly fragmented across decades of isolated primary literature. This fragmentation restricts the ability to construct the global datasets required for macro-ecological modelling.

Compounding this fragmentation is the challenge of harmonising historical ecological data through shifting taxonomic nomenclature and inconsistent spatial reporting. Reservoir host taxonomy has undergone substantial revision over the past half-century (Kryštufek et al. 2020). Without standardisation, these taxonomic shifts can artificially separate historical ecological baselines from contemporary genomic surveillance. Further, trapping data frequently lacks high-resolution geographic coordinates, often reporting sampling locations at the provincial or district level. Treating these aggregate, district-level reports as point-level occurrence data in spatial models risks introducing analytical bias by misaligning host occurrences with high-resolution environmental covariates (Lachish and Murray 2018; Gibb, Albery, Becker, et al. 2021).

The ArHa database bridges raw historical field trapping data and modern analytical pipelines. By systematically extracting, standardising, and centralising published small mammal sampling data, ArHa addresses the barriers posed by shifting host nomenclature and inconsistent spatial and temporal reporting. Each record is annotated with a categorical geographic coordinate resolution (from trapping site, village, town, city or study area, through the third, second and first administrative levels, to country) and a categorical temporal resolution (from a full date, through day, month and year ranges, to dates inferred from the publication), alongside the taxonomic resolution of the sampled animal, allowing users to filter or weight records according to the precision required. ArHa uses a relational database structure to capture the complexity of historical study designs: where single specimens yield multiple diagnostic assay results or genetic sequences, this structure preserves those relationships without the record duplication inherent to flat-file formats, providing a robust, FAIR-compliant foundation for future research.

3. Methods

3.1 Literature Search and Data Extraction

Searches of NCBI PubMed and Clarivate Web of Science on 16 October 2023 returned 4,020 records; with 11 records identified by other means, 2,755 unique records remained after de-duplication (Simons et al. 2025). Briefly, the extraction process captured study design, raw trapping effort (e.g., total trap nights), diagnostic assay methodologies, and NCBI GenBank accession numbers linking host and pathogen sequences. To preserve data provenance, the verbatim text from the primary literature was extracted before any harmonisation was applied.

Data were extracted into the template set out in the protocol, with linked sheets for studies, small mammal sampling, pathogen assays and sequences (Simons et al. 2025). Of the 728 papers from which data were extracted, 26 were extracted by two team members. D.S. compared the two extractions of each of these papers and kept the more complete one; the choice is recorded in the processing scripts, and the other extraction is not part of the release.

In a deviation from the registered protocol, AI tools (Claude Code, Opus 5.5, Anthropic) were used prior to finalisation of database: to find candidate cases of the same field data published in more than one paper, to check extracted records against their source papers and propose corrections, and to check coordinates that fell outside their stated country against the source papers. D.S. decided every cross-study duplicate action and every coordinate that was kept outside its stated country. The individual field corrections were not all reviewed by the authors; their accuracy was estimated from a random sample (see Technical Validation).

The release also departs from the protocol in three other ways. The protocol planned to impute non-detections for species detected elsewhere in a study; this was not done, and non-detection records come only from what the papers report, recorded at extraction (see Trapping Effort). The protocol planned to give the uncertainty of each coordinate as a radius; the release gives a categorical coordinate resolution instead (see Technical Validation). Finally, the protocol planned to re-run the search once extraction was complete; it has not been re-run for this release (see Limitations).

3.2 Data Processing and Harmonisation Pipeline

To transform the fragmented, verbatim extractions into a relational, machine-readable database, a sequential programmatic pipeline was developed in `R` (R Core Team 2023). This ensures reproducibility from raw data ingestion to the final database release. The pipeline comprises seven sequential phases:

3.2.1 Corrections to Extracted Records

The extraction sheets are not edited after extraction. Corrections are listed in tables in the repository, each with its source in the paper: 3,243 corrections to individual values, 331 deleted rows and 316 re-extracted rows. The pipeline downloads the sheets and applies the corrections to the downloaded copy. It stops if a value in a sheet no longer matches the value a correction expects, so a correction cannot silently overwrite a later change.

3.2.2 Metadata Consolidation

Raw extractions are aggregated to consolidate citation metadata and establish unique identifiers for all studies, hosts, pathogens, and sequences. Because historical literature reports temporal data in highly variable formats (e.g., specific days, month-year, multi-year spans), a parsing function was developed to translate verbatim event dates into standardised `start_date` and `end_date` bounds, generating a consistent `temporal_resolution` flag for every record.

3.2.3 Taxonomic and Spatial Resolution

Host taxonomy is resolved via the `taxize` package (Chamberlain et al. 2020), querying the GBIF API to append standardised hierarchies (Species through Class) and persistent `taxonID` keys (Chamberlain et al. 2026). Concurrently, geographical point data undergo validation against standard administrative boundaries. Coordinates are matched to their corresponding ISO-3166 country codes, and a spatial join against the Database of Global Administrative Areas (GADM, version 4.1) enriches each record with standardised sub-national administrative levels (adm1, adm2, adm3) relevant to the level of geographic resolution (Database of Global Administrative Areas 2022; Hijmans et al. 2024).

Where a paper gave no coordinates, the named locality (a village, town or administrative unit) was located using Google Maps, Wikipedia, OpenStreetMap or the GEOnet Names Server, and `coordinate_resolution_processed` records the level of that locality. One study gives its sites only as GADM version 2.8 municipality identifiers (Wang et al. 2023); these records are placed at a point inside each municipality.

3.2.4 Trapping Effort

To support downstream macro-ecological modelling, unstructured, free-text trapping descriptions are parsed to extract trap-night metrics, recorded in `trap_nights_clean` and described by `trap_nights_status`. Where source manuscripts reported data in tabular form listing multiple species and sampling sites, zero-detection records were extracted directly from blank or absent cells, preserving the structure of the original study design. Users should be aware that a zero-detection record in `host_occurrences` reflects a recorded absence within a specific sample, and should not be interpreted as ecological absence from the broader study area.

3.2.5 Pathogen and Genomic Enrichment

Pathogen nomenclature undergoes parallel harmonisation against the NCBI Taxonomy database (Schoch et al. 2020). To bridge the gap between ecological sampling and molecular epidemiology, raw NCBI GenBank accession numbers extracted from the literature are passed back through the NCBI Entrez API (Sayers et al. 2024). This automated sequence enrichment downloads and parses the metadata for each accession. The host and pathogen names in each GenBank record are then compared with the names assigned in ArHa, and the differences are listed in the quality assurance report. Differences are not corrected automatically, because GenBank host metadata are often incomplete or give a synonym, an older name or a higher taxon (see Technical Validation).

3.2.6 Cross-study Duplicate Resolution

Some field data were published in more than one paper and so entered ArHa more than once. After enrichment, the pipeline resolves these cases from a table of decisions in the processing scripts. Records are removed where a paper repeats data already extracted from another, merged where two papers report complementary results for the same animals, and linked as a duplicate group where the same animals are reported at different spatial or temporal breakdowns that cannot be reconciled (`duplicate_group` and `duplicate_role` in `sampling_events`). How the duplicates were found, and how many, is described below.

3.2.7 Quality Assurance and Schema Formatting

Prior to final compilation, the pipeline generates a quality assurance report to evaluate database dimensions, coordinate validity, and taxonomic consistency. The final output is then structured into the four-tier relational schema.

3.3 Duplicate Publication of Field Data

The same field data may be published more than once: a survey reported in several papers, a synthesis that consolidates earlier studies, or results reanalysed in a later paper. Where a paper does not say that its data have been published before, extraction can duplicate data. The same animals, sites and assay results then enter the database more than once, affecting subsequent counts, prevalences and ranges drawn from, in ways that cannot be seen in the data.

To find such cases, each assay record was given a fingerprint of host species, pathogen, number positive, number tested and country, and pairs of studies that shared these fingerprints with overlapping sampling dates were flagged; study totals were compared in the same way. Each flagged pair was checked against both papers. Of the 44 pairs checked, 19 shared counts by chance. The rest were resolved by pre-defined rules: keep the most complete source data; where two papers report the same animals at breakdowns that cannot be merged, keep both and link them; and do not treat a review, or a second copy of the same paper, as a source.

In all, 15 studies were removed entirely. Of these, 5 were one paper entered under two citation records. A further 124 host records were removed where two papers overlapped in part, and 43 host records were merged where two papers reported different assays on the same animals. Studies reporting the same animals in formats that cannot be merged are retained and linked through `duplicate_group` and `duplicate_role` in `sampling_events`: 6 studies in 3 groups (see Usage Notes). Each action is listed in the processing scripts and summarised in `duplicate_resolution.csv`.

3.4 The ArHa Database Explorer

To visualise the relational data architecture and improve accessibility for the wider ecological and public health communities, the database is accompanied by the ArHa Database Explorer (https://diddrog11.shinyapps.io/arenavirus_hantavirus_app/). Built using `R` and `Shiny` (Chang et al. 2025), with a DuckDB backend (Mühleisen and Raasveldt 2025) for relational querying and `Leaflet/plotly` for interactive spatial and statistical visualisation (Cheng et al. 2024; Sievert 2020).

Users can spatially visualise genomic sampling via the Leaflet integration, cross-filter by diagnostic modality or host taxonomy, and export subsets of the data for local modelling.

3.5 Data Records

The ArHa database is released under a dual-format strategy to maximise utility and interoperability.

This release holds 710 studies with sampling in 103 countries at 4,910 distinct localities (Figure 1). Sampling began between 1964 and 2021 (Figure 2).

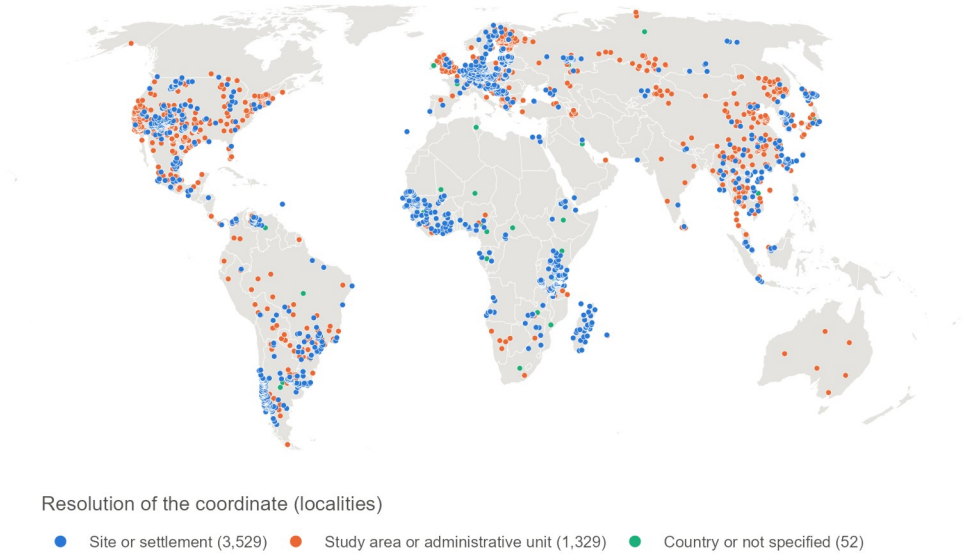


Figure 1: Sampling localities in ArHa, coloured by the resolution of the reported location. Each point is a distinct coordinate; a coordinate used at more than one resolution is shown at its most precise. The numbers in the legend are localities.

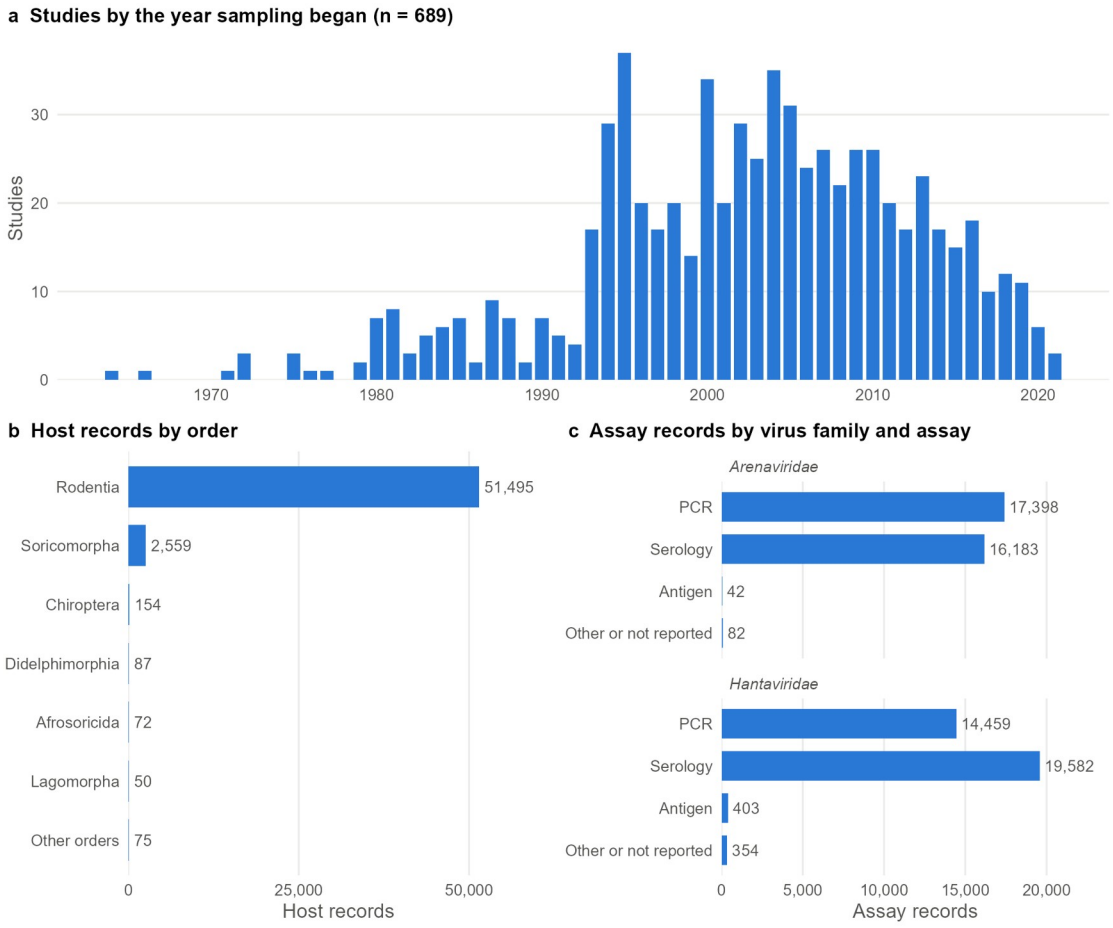


Figure 2: Contents of ArHa. (a) Studies by the year in which sampling began, for the 689 of 710 studies that report a sampling date. (b) Host records by order. (c) Assay records for Arenaviridae and Hantaviridae by assay type; "Other or not reported" includes virus culture, sequencing, Western blot and pooled serology or PCR. 106 assay records for other or unrecorded virus families are not shown.

Host occurrence records tend not to be individual animals. Each record gives the number of animals of one taxon recorded in one sampling session at one locality, and 14,333 records (26.3%) record that a taxon was not caught (`individualCount` = 0). Records are also unevenly spread across studies: the three studies with the most records, two long-term population studies and one survey of many localities, hold 52% of `host_occurrences` (Mariën et al. 2020; Cuypers et al. 2023; Guzzetta et al. 2017). Cuypers et al. (2023), for example, sampled 187 localities in Tanzania and reported results for 54 small mammal

taxa at each, contributing 10,120 records (18.6% of `host_occurrences`) from 1,224 animals. Most of these records document that a taxon was not caught at a locality rather than a capture. Analyses that treat records as equivalent units, without reference to `individualCount`, will therefore over-weight studies of this kind relative to those that report only captures.

3.5.1 Schema Definition

A primary vulnerability when archiving complex ecological sampling in flat-file formats is data expansion: forcing many-to-many relationships (e.g., multiple diagnostic assays or genetic sequences derived from a single host) into a single two-dimensional table can result in substantial record duplication. To mitigate this and ensure data integrity, the ArHa dataset is engineered as a four-tier relational database, summarised in Table 1 and Figure 3.

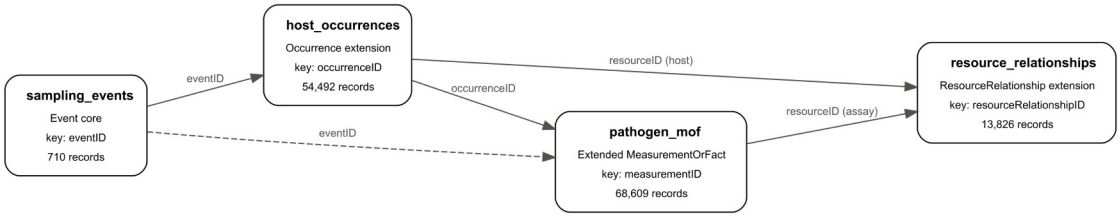


Figure 3: The four relational tables of ArHa, their roles in the Darwin Core Archive, their keys and the number of records. Arrows run from the referenced table to the table that holds the key. Assay records link to a sampling event and, where the paper allows, to a host record; sequence records link to either a host or an assay record.

Table 1: Summary of the four ArHa relational tables.

Table	Primary Key	Foreign Key(s)	Rows	Description
sampling_events	eventID	—	710	Study-level metadata, literature provenance, and aggregated sampling effort.
host_occurrences	occurrenceID	eventID	54,492	Spatiotemporal host sampling records, including zero-detection records.
pathogen_mof	measurementID	occurrenceID, eventID	68,609	Diagnostic assay results (Darwin Core MeasurementOrF

Table	Primary Key	Foreign Key(s)	Rows	Description
				act extension).
resource_relationships	resourceRelationshipID	occurrenceID or measurementID (polymorphic)	13,826	Linked GenBank sequence metadata (Darwin Core ResourceRelationship extension).

Every field of the four tables is defined in the data dictionary distributed with the release (`data_dictionary.csv`; Supplementary Tables S1–S4), with its Darwin Core term where one applies. The pipeline stops if a released field is missing from the dictionary.

3.5.1.1 `sampling_events`

This table contains the study-level metadata, literature provenance, and aggregated sampling effort metrics (e.g., total study-wide trap-nights), providing the bibliographic anchor for all subsequent records. Studies that report the same animals at different breakdowns that cannot be merged share a `duplicate_group`; within a group, `duplicate_role` marks the primary study (the finer spatial breakdown), and `eventRemarks` records what was done. `publication_year` is the year the paper was published, not the year of sampling, and `samplingProtocol` records the type of effort the paper reports rather than the trapping method.

3.5.1.2 `host_occurrences`

This table details the spatiotemporal sampling of individual small mammals or homogeneous pooled cohorts. Each record is anchored by a unique `occurrenceID` identifier. To ensure interoperability, taxonomic classifications are harmonised to the GBIF backbone (`scientificName`, `genus`, `order`, `class`) and matched to a definitive `taxonID` (Registry-Migration.Gbif.Org 2023). The depth of this resolution is captured in `taxonRank`, while the unprocessed historical nomenclature is preserved within `verbatimIdentification`. Sampling metrics include `individualCount` and specific localised effort (`trap_nights_clean`, `trap_nights_status`). The timing of these events is recorded in `eventDate`, with the precision of the historical reporting quantified in `temporal_resolution`. Spatial geometries are provided at multiple hierarchical levels, from ISO-3166 `countryCode` down to sub-national administrative boundaries (`stateProvince`, `county`, `municipality` referring to `adm1-3` in GADM (Database of Global Administrative Areas 2022)). Explicit spatial coordinates (`decimalLatitude`, `decimalLongitude`) are provided alongside `coordinate_resolution_processed`, a categorical descriptor (e.g. site, village, adm3, adm2, adm1, country) reflecting the precision at which the original sampling location was reported, rather than a continuous uncertainty radius.

The administrative fields hold GADM 4.1 names, and `higherGeographyID` gives the identifier of the finest unit. `countryCode` is the ISO 3166-1 alpha-3 code (alpha-2 in the Darwin Core Archive). `organismQuantityType` marks the one compilation that reports

minimum numbers known alive rather than animals handled. `eventDate` is expanded to whole days whatever the precision reported; `temporal_resolution` and `verbatimEventDate` give that precision and the date as reported.

3.5.1.3 pathogen_mof

Structured in alignment with the Darwin Core MeasurementOrFact (MoF) extension, this table registers the specific diagnostic assays performed on the sampled populations. It permits a one-to-many relationship with `host_occurrences` (via the foreign key `occurrenceID`) to accurately preserve instances of multi-modal diagnostic testing (e.g., parallel PCR and serological assays on a single specimen) or testing for multiple pathogens. Pathogen taxonomy is harmonised to the International Committee on Taxonomy of Viruses (ICTV) (Zerbini et al. 2025) and NCBI taxonomy databases (Schoch et al. 2020) (`scientificName_pathogen`, `taxonID_pathogen`, `taxonRank_pathogen`). To ensure bioinformatic transparency, the provenance of these taxonomic assignments is preserved as a JSON entry within the `dynamicProperties` field. The verbatim assay descriptions and original pathogen terminology reported by the authors are retained within `measurementRemarks` and `pathogen_family_original` to prevent historical data loss. Diagnostic outcomes are captured using quantitative fields: the diagnostic modality is defined in `measurementMethod`, the aggregate number of individuals tested in that specific cohort is recorded in `number_tested`, and the absolute number of positive detections is captured as the `measurementValue`. Explicit accompanying variables (`number_negative`, `number_inconclusive`) support interpretation of the extracted study results.

`measurementType` gives the finest pathogen name available (species, else genus, else family), and `measurementUnit` is individuals except for the compilation that reports minimum numbers known infected. For serological records the pathogen named is usually the one the paper reports, often the antigen used in the assay. Because antibodies cross-react between related viruses, a species name on a serological record does not confirm that species.

Assay records with `number_tested` = 0 (n = 21,766; 31.7% of assay records) represent species-assay combinations where no individuals were screened. The majority (78.7%) correspond to host records with `individualCount` = 0, where no animals were captured. A further 19.8% represent species that were captured but not screened by that assay, most commonly where a study applied molecular testing selectively rather than to all captured taxa. The remaining 1.5% (n = 321) have no corresponding host record at all, typically reflecting samples pooled or reported at a level the extraction could not resolve to an individual occurrence. These records are retained because they document the taxonomic scope of screening effort within each study, which is itself informative for characterising surveillance bias. They carry no denominator and must be excluded from prevalence calculations.

3.5.1.4 resource_relationships

Aligned with the Darwin Core ResourceRelationship extension, this table isolates genomic sequence metadata while preventing data duplication that can occur when multiple genetic targets (e.g., individual segments of a multipartite virus, host derived and pathogen derived sequences) are derived from a single specimen. Linkage is maintained through a polymorphic foreign key (`resourceID`), which maps the sequence to either the sampled host (`occurrenceID`) or a specific diagnostic event (`measurementID`). The biological context of this linkage is defined by the `relationshipOfResource` field, further categorised by `sequence_type` (e.g., Host or Pathogen) and the specific `gene_name` targeted (e.g., *Cytb*, *CO1*). To support interoperability with current genomic infrastructure, the table gives the GenBank accession as reported (`query_accession`) and as resolved by NCBI (`accession_primary`), alongside a resolvable address in the NCBI Nucleotide database (`relatedResourceID`), so that computational pipelines can link ArHa records directly to the sequences. Where NCBI did not resolve an accession, the address uses the accession as reported. A sequence linked to both a host record and an assay record has one row for each, with `resourceRelationshipID` ending in `_host` or `_pathogen`.

3.6 Formats

The ArHa database is distributed in two formats to serve the needs of end-users and data aggregators.

1. For analytical applications, the relational database is deployed as a set of compressed Apache Parquet (.parquet) files. By partitioning the database into Parquet files, users can perform multi-table relational joins across the database's 137,637 records — spanning 710 sampling events, 54,492 host occurrences, 68,609 pathogen assay records, and 13,826 sequence relationships — on standard local hardware. The same four tables are also provided as CSV files, which open in a spreadsheet or any statistical software without further tools.
2. To meet FAIR data principles, all four tables are also provided as a Darwin Core Archive (DwC-A) sampling-event dataset, so that the data can be read by standard biodiversity data tools and ingested by aggregators. `sampling_events` forms the Event core, with `host_occurrences` as an Occurrence extension, `pathogen_mof` as an Extended MeasurementOrFact extension and `resource_relationships` as a ResourceRelationship extension. In the archive each assay record gives two measurements, the number tested and the number positive, and `countryCode` is given as the ISO 3166-1 alpha-2 code recommended by Darwin Core. The archive includes the `meta.xml` file descriptor and dataset metadata in Ecological Metadata Language (`eml.xml`).

4. Limitations

Several limitations should be considered when interpreting or reusing this database.

ArHa records whether a virus, or antibodies to it, was detected in sampled animals; it does not show whether a host species maintains the virus or passes it on. Antibodies show past exposure, and a virus detected by PCR may come from a spillover infection in a host that does not transmit it further. The prevalence of a virus in a host species is therefore not a measure of that species' competence as a reservoir, or of its contribution to transmission or to spillover into people, and detection records should not be read as evidence that a species is a reservoir (Goodfellow et al. 2021).

ArHa is released in versions rather than maintained continuously. The literature search was run on 16 October 2023 and has not been updated, so at the time of publication it is several years out of date; users who need recent literature should treat ArHa as a historical baseline. Each release is archived under a single concept DOI, and because every processing step is scripted, the workflow could be rerun on an updated search.

Most constituent studies report aggregated rather than individual-level data (e.g. summary counts per site or per species, rather than a record per captured animal). Where possible, the extraction process disaggregates these summaries to the finest resolution reported by the source manuscript, but this remains a reconstruction from aggregate reporting rather than true individual-level data, and should not be treated as a substitute for individual-level datasets where those are specifically required.

Records describe sampling sessions, not individual animals. In studies that trapped the same sites repeatedly and released animals, the same individuals may be counted in more than one session; in studies that removed animals for tissue sampling, they cannot. This release does not record which sampling protocol each study used, so users cannot yet separate the two, and summing `individualCount` or `number_tested` across sessions may count some animals more than once. One compilation reports minimum numbers known alive and infected rather than animals tested; these records are identified by `organismQuantityType` and `measurementUnit`. Recording the sampling protocol for each multi-session study is planned for the next version of ArHa.

We have not produced an accompanying sampling-effort bias correction (e.g. a global rodent trapping-effort raster) to weight or adjust the occurrence data. Sampling effort documented in the assembled literature is unlikely to be driven by host distributions alone, and instead reflects research infrastructure, funding priorities, and public health interest, factors that vary depending on the disease or ecological question under consideration (Lachish and Murray 2018; Gibb, Albery, Becker, et al. 2021; Simons et al. 2023). A correction appropriate for characterising general small-mammal biodiversity sampling bias would not necessarily be appropriate for characterising disease-surveillance effort, since the two are not driven by the same factors, and a single global correction risks being inappropriate for at least some downstream uses. We consider this out of scope for the present release and see it as a valuable direction for future, question-specific methodological work rather than a single correction applied uniformly.

Spatial and temporal resolution vary substantially across the database, and this variation is a direct consequence of incomplete or imprecise reporting in the underlying literature

rather than a limitation introduced by the extraction pipeline itself: many historical surveillance papers report only administrative-level localities or multi-year sampling windows. The coordinate_resolution_processed and temporal_resolution fields are provided so that users can filter or weight records according to the precision appropriate to their analysis (see Usage Notes).

Taxonomic and pathogen nomenclature reflect the GBIF and NCBI Taxonomy backbones current at the time of processing. A number of pathogen records could not be resolved beyond family level and remain classified as unclassified Arenaviridae or Hantaviridae; this reflects the resolution reported in the primary literature and the current state of viral taxonomy databases, and may resolve further as viral taxonomy continues to be revised and backbone databases updated (Zerbini et al. 2025).

The literature search used a deliberately sensitive strategy built around the family-level terms *Arenaviridae* and *Hantaviridae* to maximise recall. This may have missed studies in which small mammals were sampled and tested for these viruses without a family-level term appearing in the title, abstract, or keywords — for example, studies reporting negative results framed only around a specific virus or disease name (Simons et al. 2025). As a result, ArHa's coverage of true negative surveillance data is likely less complete than its coverage of positive detections and studies explicitly framed around arenavirus or hantavirus surveillance.

Records of the same animals published in more than one paper were found by comparing records and study totals across studies (see Methods). Repeats that report the data at a different level of aggregation, or that overlap only in part, are hard to find this way, particularly when a paper does not cite the earlier publication, so some animals may still be counted more than once.

`eventDate` is expanded to whole days whatever the precision reported, and sessions that a paper pools are given as one interval from the first date to the last. `temporal_resolution` and `verbatimEventDate` give the precision of the original report.

The link between a sequence and its host record comes from the source paper. It could be compared with GenBank only where the GenBank record names a host, and the differences found were not reviewed one by one (see Technical Validation).

Finally, this database reflects the studies that reached publication rather than the full body of historical small-mammal-pathogen surveillance: unpublished negative surveys, internal agency reports, and datasets never submitted for peer review are not represented, which may bias the geographic and taxonomic patterns visible in the database toward regions and host taxa where publication was more likely (Gibb, Albery, Mollentze, et al. 2021).

5. Technical Validation

Data harmonisation and structural validation were executed programmatically in R, ensuring reproducibility from raw extraction to final release.

As an additional check on extraction reliability, records from a random sample of approximately 10% of included studies were cross-checked against their original source manuscripts. This confirmed complete record-level recovery — no records were missing or spuriously added — for the audited studies; concordance of individual field values within those records was reviewed informally but not systematically quantified.

The scripted corrections (see Methods) were checked against the source papers by a random spot check. On 20 October 2026, 50 of the 3,243 corrections then listed (field corrections, row deletions and added rows) were drawn at random (R seed 20261020), and D.S. checked each against the source paper. Of these, 46 were confirmed (92%, exact 95% CI 80.8–97.8%), 4 were marked unsure and 0 were found to be wrong. The unsure cases were two coordinates converted from UTM, which could not be checked by hand, one sampling date range, and one virus abbreviation. The spot check does not cover the 647 corrections added after the sample was drawn, or the cross-study duplicate actions.

Foreign keys were checked with anti-joins in the released tables. Every assay record in `pathogen_mof` links to an existing sampling event, and every sequence record in `resource_relationships` links to an existing host or assay record. 480 assay records (0.7%) have an `eventID` but no `occurrenceID`, because the source paper reported the results at a level that could not be matched to a host record (see `pathogen_mof`); every other assay record links to an existing host record.

Of the 8,078 GenBank accessions, 227 could not be resolved through the NCBI Entrez API. Where both names were available, the host name in the GenBank record matched the ArHa host name for 84.5% of 7,424 sequences, and the pathogen name matched for 71.3% of 4,481 pathogen sequences. A difference does not by itself show an error, because GenBank records often give a synonym, an older name, a strain name or a higher taxon. The differences were not reviewed one by one and are not flagged in the release. One difference that was followed up showed an extraction error, which was corrected against the source paper.

Locality data underwent spatial validation using the `terra` package, with `sf` used for visualisation (Hijmans 2025; Pebesma 2018). Each coordinate was assigned a categorical `coordinate_resolution_processed` value (`site`, `village`, `town`, `city`, `study_area`, `adm3`, `adm2`, `adm1`, `country` or `unspecified`) reflecting the precision of the geographic information reported in the source manuscript, rather than a continuous uncertainty radius. Where a paper gave no coordinates, the record was placed at its named locality (see Methods). Missing coordinates and coordinates outside the valid range are flagged in `coord_status`; 0 host records are flagged in the release.

Each coordinate was then checked against the full-resolution GADM 4.1 outline of its stated country. Every point found outside the outline was reviewed: 10 points were corrected against the source papers, and the others were kept and listed, with the reason, in `coordinate_reviews.csv`. Most kept points are offshore points for coastal and island sites or points on a border, which fall outside the outline because of the precision of the outline or of the reported coordinate. The rest are in areas that GADM codes differently

from ArHa (below). In the release, 120 host records at 26 coordinates lie outside the outline of their stated country, 20 of them by more than 50 km, and all are on the reviewed list. For these records, `dist_from_expected` gives the distance to the border of the stated country in two bands (less or more than 50 km).

`country` names the country where the animals were sampled, judged from the locality and coordinates rather than from the title of the paper. Where one state administers an area that another claims, ArHa uses the administering state: Kinmen and the Matsu Islands are coded as Taiwan, and the Paracel Islands as China. This is a convention for placing records and implies no position on sovereignty. GADM 4.1 codes the Paracel Islands as a separate unit and places all of Bolshoy Ussuriysky (Heixiazi) Island in Russia, although its western part was transferred to China in 2008, so records from these places lie outside the outline of their stated country.

In the Darwin Core Archive, `georeferenceRemarks` gives the resolution of each coordinate and, for these records, the reason the coordinate was kept, so that tools that repeat a country check can see it.

Because historical surveillance frequently aggregates trapping effort over multi-year windows without providing specific session dates, users risk extracting high-cadence environmental covariates (e.g., daily precipitation or temperature) for low-cadence sampling events. To prevent this, the pipeline isolates aggregate temporal windows into defined `start_date` and `end_date` parameters and assigns a categorical `temporal_resolution` flag (`full_date`, `day_range_resolution`, `month_year`, `month_range_resolution`, `year_only`, `year_range_resolution` or `publication_derived`, the last for studies that give no sampling date, whose `eventDate` is left empty). This encourages downstream spatial models to align environmental covariate extraction with the true temporal precision of the primary literature.

6. Usage Notes

To ensure the appropriate application of the ArHa database in downstream predictive modelling, users should adhere to the following analytical constraints.

Avoid flat-file manipulation. The multi-relational nature of this dataset means it cannot be accurately subset or manipulated in standard spreadsheet software (e.g., Microsoft Excel) without risking duplication. Users are advised to utilise the provided DuckDB/dplyr pipeline or the web application to extract specific cohorts.

Treat zero-detection records as sample-level non-detections, not ecological absences. A zero `individualCount` or zero `measurementValue` in this database indicates that the target species or pathogen was not detected within the reported sampling effort for that record. It does not indicate that the species or pathogen is absent from the broader study area or habitat. This limitation applies to the majority of constituent studies regardless of how the data were extracted: most historical surveillance papers do not

quantitatively report sampling effort, so the absence of a detection record for a given species at a given location reflects the limits of the sample, not confirmed ecological absence. Downstream spatial occupancy or spillover risk models must explicitly parameterise detection probability — for example, using `trap_nights_clean` as an offset — to avoid classifying under-sampled regions as true ecological non-habitats.

Exclude zero-denominator assay records from prevalence estimates. Prevalence should be calculated only across records where `number_tested` > 0. Retaining zero-denominator records will not produce invalid positives — no record reports positives without a denominator — but will bias any per-record prevalence summary toward zero.

Count each group of duplicated animals once. 6 studies in 3 groups report the same animals at breakdowns that cannot be merged (see Duplicate Publication of Field Data). Use one study from each `duplicate_group`, normally the one with `duplicate_role` = primary, which has the finer spatial breakdown; using both counts the same animals twice.

Check what a count measures. In 1,475 host records and 1,475 assay records from one compilation, the counts are monthly minimum numbers of animals known alive and known infected, not animals handled and tested. `organismQuantityType` and `measurementUnit` identify these records; exclude them, or treat them separately, when summing animals or estimating prevalence.

Match covariates to the precision of the date. `eventDate` is expanded to whole days whatever the precision reported, so a record known only to the year spans the whole year. Use `temporal_resolution` to choose the records, or the temporal scale of covariates, that suit the analysis. The 537 host records marked `publication_derived` have no sampling date.

Filter assay records on virus family. `pathogen_mof` keeps 12 assay records for other virus families reported alongside arenaviruses and hantaviruses, and 94 with no family recorded. Filter on `family_pathogen`.

7. Acknowledgements

We thank the reviewers of the protocol paper for their comments.

The R code that compiles the database was developed by the authors over four years. Claude Code (Anthropic) was used in the final stages to revise and improve it, and to draft and edit parts of this manuscript, under the direction of D.S., who reviewed the code, its outputs and the text. As described in Methods, it was also used to find candidate duplicate publications, to check extracted records against their source papers and to propose corrections under rules set by the authors. Not every correction was reviewed individually by the authors; their accuracy was estimated from a random sample (see Technical Validation). The authors take full responsibility for the content.

8. Author Contributions

D.S.: conceptualisation, methodology, data curation, software, formal analysis, supervision, funding acquisition, writing – original draft, writing – review and editing.
R.R.: methodology, data curation, software, formal analysis, writing – review and editing.
G.R.: data curation, writing – review and editing.
A.M.-C.: data curation, writing – review and editing.
H.G.: data curation, writing – review and editing.
G.C.M.: supervision, writing – review and editing. D.W.R.: software, formal analysis, supervision, funding acquisition, writing – original draft, writing – review and editing.
S.N.S.: conceptualisation, methodology, formal analysis, supervision, funding acquisition, writing – original draft, writing – review and editing.

9. Funding

This work was supported by funding from a joint NSF-NIH-NIFA and BBSRC Ecology and Evolution of Infectious Disease Award (2208034; D.W.R., D.S.), an NSF Biology Integration Institute grant to the Verena Institute (2213854; S.N.S., D.S., R.R.) and a Wellcome Trust/Royal Society Sir Henry Dale Research Fellowship (220179/Z/20/Z, 220179/A/20/Z; D.W.R.).

10. Conflict of interest disclosure

The authors of this article declare that they have no financial conflict of interest with the content of this article.

11. Data, script, code, and supplementary information availability

This version of the database was compiled on 23 October 2026 and is archived on Zenodo (https://doi.org/10.5281/zenodo.21978982; this DOI always resolves to the latest version). The archive contains the four tables of the database as Parquet files and as CSV files, the Darwin Core Archive, the data dictionary, the extraction data as downloaded from the extraction sheets (one CSV file per sheet), the tables of scripted corrections, of cross-study duplicate resolution and of taxonomic matching, the R scripts that compile the database from these files, and the source code for the ArHa Database Explorer web application. The development version of the pipeline is maintained on GitHub (https://github.com/DidDrog11/arenavirus_hantavirus).

12. Supplementary Tables

Supplementary Table S1. Fields of `sampling_events`.

Field	Darwin Core term	Type	Description
eventID	eventID	identifier	Identifier of the study (one sampling event per study), event_0001 onwards. Links to host_occurrences and pathogen_mof.
associatedReferences	associatedReferences	text	Citation of the source paper: authors (year), title, journal and DOI.
publication_year	NA	integer	Year the source paper was published. This is not the year of sampling, which is in eventDate in host_occurrences.
samplingProtocol	samplingProtocol	category	Type of sampling effort the paper reports for the whole study: trap_nights, sessions, time_period (a sampling period only), unspecified (effort described but not in one of these forms) or not_reported.
sampleSizeValue	sampleSizeValue	decimal	Total trap-nights for the study, where samplingProtocol is trap_nights; empty otherwise. The unit is trap-nights (sampleSizeUnit in the Darwin Core Archive).
samplingEffort	samplingEffort	decimal	Number of trapping sessions for the study, where samplingProtocol is sessions; empty otherwise. Study-level effort was poorly extracted, and this is filled for only one study.
dataAccessLevel	NA	category	Whether the paper reports data for individual animals (individual), aggregated data (summarised), or does not say (unspecified).
dataResolution	NA	category	Level at which the paper aggregates its data: individual, site-session, site-season, site, session, study or unspecified; empty where not recorded.
duplicate_group	NA	text	Shared identifier for studies that report the same animals at different breakdowns that cannot be merged; empty otherwise.
duplicate_role	NA	category	Within a duplicate_group: primary (the finer spatial breakdown) or alternative_stratification. Keep only primary studies to avoid counting the same animals twice.
eventRemarks	eventRemarks	text	How the study's records were combined with, or separated from, another paper reporting the same animals.

Supplementary Table S2. Fields of `host_occurrences`.

Field	Darwin Core term	Type	Description
occurrenceID	occurrenceID	identifier	Identifier of the host record, host_00001 onwards.
eventID	eventID	identifier	Study the record belongs to (sampling_events).
scientificName	scientificName	text	Host species name accepted in the GBIF backbone taxonomy; empty when the host is not identified to species.
genus	genus	text	Host genus in the GBIF backbone.
family	family	text	Host family in the GBIF backbone.
order	order	text	Host order in the GBIF backbone, which places shrews and moles in Soricomorpha and hedgehogs in Erinaceomorpha.
class	class	text	Host class in the GBIF backbone.
taxonRank	taxonRank	category	Finest rank to which the host is identified: species, genus, family, order or class.
taxonID	taxonID	integer	GBIF backbone taxon key of the host at taxonRank.
individualCount	individualCount	decimal	Number of animals of the taxon recorded in this sampling session at this locality. 0 means the taxon was not caught; empty when the paper gives no count.
organismQuantity	organismQuantity	decimal	The same value as individualCount, with its type given in organismQuantityType.
organismQuantityType	organismQuantityType	category	individuals, or minimum number known alive for one compilation that reports monthly minimum numbers known alive rather than animals handled.

Field	Darwin Core term	Type	Description
eventDate	eventDate	date or interval	Sampling date, or interval as start/end, in ISO 8601. Dates are expanded to whole days whatever the precision reported; temporal_resolution gives that precision and verbatimEventDate the date as reported.
country	country	text	Country where the animals were sampled, as written in the extraction (for example USA). Where one state administers an area claimed by another, the administering state is used.
countryCode	countryCode	text	ISO 3166-1 alpha-3 code of country (alpha-2 in the Darwin Core Archive); empty for records that span several countries.
stateProvince	stateProvince	text	Name of the GADM 4.1 first-level administrative unit containing the coordinate, where the coordinate is precise enough to place it.
county	county	text	Name of the GADM 4.1 second-level administrative unit containing the coordinate, where the coordinate is precise enough to place it.
municipality	municipality	text	Name of the GADM 4.1 third-level administrative unit containing the coordinate, where the coordinate is precise enough to place it.
higherGeographyID	higherGeographyID	identifier	GADM 4.1 GID of the finest administrative unit named in stateProvince, county or municipality.
decimalLatitude	decimalLatitude	decimal	Latitude of the locality in decimal degrees (WGS84). coordinate_resolution_processed gives its precision.
decimalLongitude	decimalLongitude	decimal	Longitude of the locality in decimal degrees (WGS84).

Field	Darwin Core term	Type	Description
			coordinate_resolution_processed gives its precision.
verbatimIdentification	verbatimIdentification	text	Host name as written in the source paper.
verbatimLocality	verbatimLocality	text	Site label or habitat description from the paper, where it distinguishes records at the same locality.
verbatimEventDate	verbatimEventDate	text	Sampling date as extracted from the paper.
coord_status	NA	category	valid; missing (no coordinate); or illogical (latitude or longitude out of range).
dist_from_expected	NA	category	For a coordinate outside the GADM 4.1 outline of its stated country, its distance from that country's border: less than 50km or greater than 50km; empty otherwise. Every such coordinate was checked against its source paper (coordinate_reviews.csv).
coordinate_resolution_processed	NA	category	Precision of the locality the paper reports: site, village, town, city, study_area, adm3, adm2, adm1 (third to first administrative level), country or unspecified.
temporal_resolution	NA	category	Precision of the sampling date the paper reports: full_date, day_range_resolution, month_year, month_range_resolution, year_only, year_range_resolution, or publication_derived (the paper gives no sampling date; eventDate is empty, and publication_year in sampling_events is the latest possible date).
trap_nights_clean	NA	decimal	Trap-nights for this record as reported in the paper; empty when not reported.

Field	Darwin Core term	Type	Description
trap_nights_status	NA	category	How trap_nights_clean was read: trapnight_reported; nights (a value under 10, read as nights of trapping rather than trap-nights); or not_reported.

Supplementary Table S3. Fields of `pathogen_mof`.

Field	Darwin Core term	Type	Description
measurementID	measurementID	identifier	Identifier of the assay record, path_00001 onwards.
occurrenceID	occurrenceID	identifier	Host record tested (host_occurrences); empty when the paper reports results at a level that cannot be matched to a host record.
eventID	eventID	identifier	Study the record belongs to (sampling_events).
measurementType	measurementType	text	Pathogen tested for, at the finest name available (species, else genus, else family), followed by “detection”.
measurementValue	measurementValue	decimal	Number of animals that tested positive.
measurementUnit	measurementUnit	category	individuals, or minimum number known infected for the compilation described under organismQuantityType.
measurementMethod	measurementMethod	category	Assay type: Serology, PCR, Antigen, Virus Culture, Sequencing, Western Blot, Serology and/or PCR (results pooled across both), Other, or Missing (not reported).
measurementRemarks	measurementRemarks	text	Number tested, the pathogen and assay as written in the paper, and extraction notes, separated by “|”.
scientificName_pathogen	NA	text	Pathogen species name matched in NCBI Taxonomy; empty when the record is not resolved to species.
genus_pathogen	NA	text	Pathogen genus.
family_pathogen	NA	text	Pathogen family. A few records are for families other than Arenaviridae and Hantaviridae, reported alongside them; filter on this field.
taxonRank_pathogen	NA	category	Rank to which the pathogen is resolved: species, genus, family or unclassified.
taxonID_pathogen	NA	identifier	NCBI Taxonomy identifier of the pathogen at taxonRank_pathogen.

Field	Darwin Core term	Type	Description
dynamicProperties	dynamicProperties	JSON	How the pathogen name was matched to the ICTV and NCBI taxonomies.
number_tested	NA	decimal	Number of animals tested. 0 where none were tested by this assay (for example, where none were caught); exclude these records from prevalence.
number_negative	NA	decimal	Number of animals that tested negative.
number_inconclusive	NA	decimal	Number of animals with an inconclusive result, where the paper reports it.
tested_detected	NA	logical	TRUE where number_tested is larger than the host record's individualCount (a check flag).
positive_tested	NA	logical	TRUE where measurementValue is larger than number_tested (a check flag; none in this release).
pathogen_species_cleaned	NA	text	Pathogen name after harmonisation to ICTV binomial names, before matching to NCBI Taxonomy.
pathogen_family_original	NA	text	Pathogen family as recorded at extraction.

Supplementary Table S4. Fields of `resource_relationships`.

Field	Darwin Core term	Type	Description
resourceRelationshipID	resourceRelationshipID	identifier	Identifier of the link: seq_00001_host or seq_00001_pathogen. A sequence linked to both a host and an assay record has one row for each.
resourceID	resourceID	identifier	Host record (occurrenceID) or assay record (measurementID) the sequence was obtained from.
relatedResourceID	relatedResourceID	URL	NCBI Nucleotide address of the sequence.
relationshipOfResource	relationshipOfResource	category	host has sequence, or pathogen has sequence.
sequence_type	NA	category	Host or Pathogen: whether the sequence is of the host or of the virus.
gene_name	NA	text	Gene or segment named in the GenBank record, where given.
query_accession	NA	text	GenBank accession as reported in the paper.
accession_primary	NA	text	Primary accession returned by NCBI for query_accession; empty where the

Field	Darwin Core term	Type	Description
			accession could not be resolved.

12. References

Chamberlain, Scott, Vijay Barve, Dan Mcglinn, et al. 2026. *Rgbif: Interface to the Global Biodiversity Information Facility API*. https://CRAN.R-project.org/package=rgbif.

Chamberlain, Scott, Eduard Szoecs, Zachary Foster, et al. 2020. *Taxize: Taxonomic Information from Around the Web*. https://github.com/ropensci/taxize.

Chang, Winston, Joe Cheng, JJ Allaire, et al. 2025. *Shiny: Web Application Framework for r*. https://CRAN.R-project.org/package=shiny.

Cheng, Joe, Barret Schloerke, Bhaskar Karambelkar, and Yihui Xie. 2024. *Leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library*. https://CRAN.R-project.org/package=leaflet.

Childs, James E., and Clarence J. Peters. 1993. "Ecology and Epidemiology of Arenaviruses and Their Hosts." In *The Arenaviridae*, edited by Maria S. Salvato. Springer US. https://doi.org/10.1007/978-1-4615-3028-2_19.

Cuypers, Laura N., Sophie Gryseels, Natalie Van Houtte, et al. 2023. "Subspecific Rodent Taxa as the Relevant Host Taxonomic Level for Mammarenavirus Host Specificity." *Virology* 581 (April): 116–27. https://doi.org/10.1016/j.virol.2023.02.014.

Database of Global Administrative Areas. 2022. *GADM*. Https://gadm.org/index.html.

Gibb, Rory, Gregory F. Albery, Daniel J. Becker, et al. 2021. "Data Proliferation, Reconciliation, and Synthesis in Viral Ecology." *BioScience*, no. biab080 (August). https://doi.org/10.1093/biosci/biab080.

Gibb, Rory, Gregory F. Albery, Nardus Mollentze, et al. 2021. "Mammal Virus Diversity Estimates Are Unstable Due to Accelerating Discovery Effort." *Biology Letters* 18 (1): 20210427. https://doi.org/10.1098/rsbl.2021.0427.

Goodfellow, Samuel M., Robert A. Nofchissey, Kurt C. Schwalm, et al. 2021. "Tracing Transmission of Sin Nombre Virus and Discovery of Infection in Multiple Rodent Species." *Journal of Virology* 95 (23): e01534–21. https://doi.org/10.1128/jvi.01534-21.

Guzzetta, Giorgio, Valentina Tagliapietra, Sarah E. Perkins, et al. 2017. "Population Dynamics of Wild Rodents Induce Stochastic Fadeouts of a Zoonotic Pathogen." *Journal of Animal Ecology* 86 (3): 451–59. https://doi.org/10.1111/1365-2656.12653.

Hijmans, Robert J. 2025. *Terra: Spatial Data Analysis*. https://CRAN.R-project.org/package=terra.

Hijmans, Robert J., Márcia Barbosa, Aniruddha Ghosh, and Alex Mandel. 2024. *Geodata: Download Geographic Data*. https://CRAN.R-project.org/package=geodata.

Jonsson, Colleen B, Luiz Tadeu Moraes Figueiredo, and Olli Vapalahti. 2010. "A Global Perspective on Hantavirus Ecology, Epidemiology, and Disease." *Clinical Microbiology Reviews* 23 (2): 412–41. https://doi.org/10.1128/CMR.00062-09.

Keesing, Felicia, Lisa K. Belden, Peter Daszak, et al. 2010. "Impacts of Biodiversity on the Emergence and Transmission of Infectious Diseases." *Nature* 468 (7324): 647–52. https://doi.org/10.1038/nature09575.

Kreuder Johnson, Christine, Peta L. Hitchens, Tierra Smiley Evans, et al. 2015. "Spillover and Pandemic Properties of Zoonotic Viruses with High Host Plasticity." *Scientific Reports* 5 (1): 14830. https://doi.org/10.1038/srep14830.

Kryštufek, Boris, Alexey S. Tesakov, Vladimir S. Lebedev, Anna A. Bannikova, Nataliya I. Abramson, and Georgy Shenbrot. 2020. "Back to the Future: The Proper Name for Red-Backed Voles Is *Clethrionomys* Tilesius and Not *Myodes* Pallas." *Mammalia* 84 (2): 214–17. https://doi.org/10.1515/mammalia-2019-0067.

Lachish, Shelly, and Kris A. Murray. 2018. "The Certainty of Uncertainty: Potential Sources of Bias and Imprecision in Disease Ecology Studies." *Frontiers in Veterinary Science* 5 (May): 90. https://doi.org/10.3389/fvets.2018.00090.

Mariën, Joachim, Benny Borremans, Christophe Verhaeren, et al. 2020. "Density Dependence and Persistence of Morogoro Arenavirus Transmission in a Fluctuating Population of Its Reservoir Host." *Journal of Animal Ecology* 89 (2): 506–18. https://doi.org/10.1111/1365-2656.13107.

McMahon, Barry J., Serge Morand, and Jeremy S. Gray. 2018. "Ecosystem Change and Zoonoses in the Anthropocene." *Zoonoses and Public Health* 65 (7): 755–65. https://doi.org/10.1111/zph.12489.

Mühleisen, Hannes, and Mark Raasveldt. 2025. *Duckdb: DBI Package for the DuckDB Database Management System*. https://CRAN.R-project.org/package=duckdb.

Olival, Kevin J., Parviez R. Hosseini, Carlos Zambrana-Torrelio, Noam Ross, Tiffany L. Bogich, and Peter Daszak. 2017. "Host and Viral Traits Predict Zoonotic Spillover from Mammals." *Nature* 546 (7660): 646–50. https://doi.org/10.1038/nature22975.

Pebesma, Edzer. 2018. "Simple Features for R: Standardized Support for Spatial Vector Data." *The R Journal* 10 (1): 439–46. https://doi.org/10.32614/RJ-2018-009.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/.

Registry-Migration.Gbif.Org. 2023. *GBIF Backbone Taxonomy*. GBIF Secretariat. https://doi.org/10.15468/39OMEI.

Sayers, Eric W, Mark Cavanaugh, Karen Clark, et al. 2024. “GenBank 2024 Update.” *Nucleic Acids Research* 52 (D1): D134–37. https://doi.org/10.1093/nar/gkad903.

Schoch, Conrad L, Stacy Ciufo, Mikhail Domrachev, et al. 2020. “NCBI Taxonomy: A Comprehensive Update on Curation, Resources and Tools.” *Database* 2020 (January): baaa062. https://doi.org/10.1093/database/baaa062.

Sievert, Carson. 2020. *Interactive Web-Based Data Visualization with r, Plotly, and Shiny*. Chapman; Hall/CRC. https://plotly-r.com.

Simons, David, Lauren A Attfield, Kate E Jones, Deborah Watson-Jones, and Richard Kock. 2023. “Rodent Trapping Studies as an Overlooked Information Source for Understanding Endemic and Novel Zoonotic Spillover.” *PLOS Neglected Tropical Diseases* 17 (1): e0010772. https://doi.org/10.1371/journal.pntd.0010772.

Simons, David, Ricardo Rivero, Ana Martinez-Checa Guiote, et al. 2025. “Protocol to Produce a Systematic Arenavirus and Hantavirus Host-Pathogen Database: Project ArHa.” *Wellcome Open Research* 10 (August): 227. https://doi.org/10.12688/wellcomeopenres.24037.2.

Wang, Yingying X. G., Liina Voutilainen, Mahdi Aminikhah, et al. 2023. “The Impact of Wildlife and Environmental Factors on Hantavirus Infection in the Host and Its Translation into Human Risk.” *Proceedings of the Royal Society B: Biological Sciences* 290 (1996): 20222470. https://doi.org/10.1098/rspb.2022.2470.

Yanagihara, Richard, Se Hun Gu, Satoru Arai, Hae Ji Kang, and Jin-Won Song. 2014. “Hantaviruses: Rediscovery and New Beginnings.” *Virus Research* 187 (July): 6–14. https://doi.org/10.1016/j.virusres.2013.12.038.

Zerbini, F. Murilo, Anya Crane, Jens H. Kuhn, Peter Simmonds, Elliot J. Lefkowitz, and ICTV Taxonomy Summary Consortium. 2025. “Summary of Taxonomy Changes Ratified by the International Committee on Taxonomy of Viruses (ICTV) – General Taxonomy Proposals, 2025.” *Journal of General Virology* 106 (7). https://doi.org/10.1099/jgv.0.002116.