GenomoBase: an integrated resource for *Genomoviridae* genomics and metagenomics

Yair Cárdenas-Conejo[1,2,3]*

1 Laboratorio de Biología Sintética, Estructural y Molecular, Universidad de Colima, Coquimatlán, Colima, México.

2 Secretaría de Ciencia, Humanidades, Tecnología e Innovación, Gobierno de México, Ciudad de México, México

3 Centro Universitario de Investigaciones Biomédicas, Universidad de Colima. Colima, Colima, México

Corresponding Author:

Yair Cárdenas-Conejo*

ORCID: 0000-0002-0190-244X

email address: ycardenas11@ucol.mx

Abstract

The family *Genomoviridae* (phylum Cressdnaviricota) comprises viruses with circular single-stranded DNA (ssDNA) genomes that typically encode a replication protein and a capsid protein. High-throughput sequencing has greatly expanded the number of known genomoviruses, but sequences and annotations remain dispersed across multiple repositories and publications, and no family-focused resource has been available. Here I present GenomoBase (https://www.genomobase.org), a curated web resource that integrates genomic, ecological and bibliographic data for all 237 *Genomoviridae* species currently recognized by the International Committee on Taxonomy of Viruses (ICTV). GenomoBase leverages a PostgreSQL backend to organize data according to standardized taxonomy, summarizing viral occurrences across ten genera, twelve host categories and nineteen countries. An "Explore Metagenomes" module incorporates more than 600 Sequence Read Archive (SRA) runs identified via Serratus, linking taxonomic records to metagenomic detections and highlighting the global distribution of genomovirus-related reads. To support research, the platform also provides automated literature updates via PubMed and local BLAST (Basic Local Alignment Search Tool) services for nucleotide and protein queries. GenomoBase offers a user-friendly entry point for exploring *Genomoviridae* diversity and is designed to facilitate taxonomic curation, ecological and evolutionary analyses, and the discovery of novel genomoviruses in both environmental and host-associated niches.

Keywords

Genomoviridae; CRESS-DNA viruses; metagenomics; viral database; Cressdnaviricota

Introduction

Circular replication-associated protein encoding single-stranded (CRESS) DNA viruses are ubiquitous components of environmental and host-associated viromes and are currently classified within the phylum Cressdnaviricota. Members of this phylum possess small circular ssDNA genomes that replicate by rolling-circle replication mediated by a conserved replication-associated protein (Rep), a hallmark of all CRESS-DNA viruses, and they infect a broad range of eukaryotic hosts, including plants, fungi, invertebrates and vertebrates [1]. The best-known families include *Geminiviridae*, which infect mono- and dicotyledonous plants; *Circoviridae*, which infect pigs, poultry and other vertebrates; and *Nanoviridae*, which infect mainly crop plants [1]. More recently, within Cressdnaviricota, the family *Genomoviridae* was established following the discovery of *Sclerotinia sclerotiorum* hypovirulence-associated DNA virus 1 (SsHADV-1) [2], a circular ssDNA virus infecting the plant-pathogenic fungus *Sclerotinia sclerotiorum* [3]. Subsequent taxonomic work has expanded *Genomoviridae* to ten genera and more than two hundred species recognized by the International Committee on Taxonomy of Viruses (ICTV) [4]. Genomoviruses typically have compact genomes of approximately 1.9–3.9 kb that encode two major proteins in an ambisense arrangement: a rolling-circle replication initiation protein (Rep) and a capsid protein (CP). The Rep protein is homologous to the corresponding Rep of members of the family *Geminiviridae*, sharing conserved domains and sequence motifs required for viral replication [2]. The origin of replication is also similar in both families, consistent with replication via rolling-circle mechanisms [2]. In contrast, the CP forms a non-enveloped, isometric T=1 icosahedral capsid and shows no recognizable amino acid sequence similarity to capsid proteins of other known virus

73 groups. Some members additionally encode small hypothetical open reading frames of
74 unknown function. To date, no canonical movement protein has been identified in
75 genomoviruses, consistent with the etymology of the family name *Genomoviridae* (Ge- for
76 "geminivirus-like", nomo- for "no movement protein") [2]. Although genomoviruses were
77 initially described in fungi, subsequent metagenomic studies have detected them in a wide
78 range of contexts, including plants, birds, mammals, invertebrates and diverse
79 environmental matrices such as wastewater, sediments, soils and food-associated
80 microbiomes. Experimental evidence of infection remains limited to *Sclerotinia*
81 *sclerotiorum* [3] and successful replication in the sciarid fly *Lycoriella ingenua* [5], and the
82 true host range of many lineages remains unclear.

83 The rapid expansion of genomovirus sequence data has been driven almost entirely by
84 metagenomic studies of human, animal and environmental viromes and by large-scale
85 surveillance projects. Gemykibiviruses and related taxa have now been reported in a variety
86 of human clinical samples, including cerebrospinal fluid from acute encephalitis patients
87 [6], vaginal secretions [7] and blood donors in the Brazilian Amazon [8]; the same virus has
88 also been identified in bats [9], suggesting a possible zoonotic link. More recently, a human
89 gemykibivirus was shown to actively replicate in HEK293T cells [10], further underscoring
90 the need to better understand the biology and potential medical relevance of these viruses.

91 Despite the growing interest, genomovirus sequences and annotations remain dispersed
92 across taxonomic resources, GenBank records and individual metagenomic studies, and no
93 dedicated, curated database has been available for this rapidly expanding family. To
94 address this gap, I developed GenomoBase (https://www.genomobase.org), an online
95 resource focused specifically on the family *Genomoviridae*. GenomoBase integrates ICTV-

96 based taxonomy with curated genome annotations, standardized host and geographic
97 metadata, Serratus-filtered metagenomic detections from the Sequence Read Archive
98 (SRA), local BLAST services and a PubMed-derived literature catalogue. Here we describe
99 the construction, content and main features of GenomoBase and provide a concise
100 overview of the patterns that can be readily explored through the portal.

101 Materials and methods

102 Reference taxonomy and genome retrieval

103 ICTV-recognized genomovirus species were retrieved from the Virus Metadata Resource
104 (VMR) file VMR_MSL40.v2.20251013 (https://ictv.global/vmr), which lists all species
105 and genera formally assigned to the family *Genomoviridae*. From this file we extracted the
106 scientific name and the corresponding GenBank accession number for each species. The
107 curated table underlying GenomoBase contains 239 genome records: one representative
108 genome for each of the 237 species, plus the three genomic segments (A, B and C) of the
109 tripartite *Gemytripvirus fugra1*. Using a custom Python script (fetch_genomes.py) built on
110 Biopython v1.86 [11], we downloaded the corresponding GenBank records via NCBI E-
111 utilities and parsed the following information: accession, genus, species name, virus name,
112 sequence definition line, genome length, full genomic sequence and all annotated coding
113 sequences (CDSs). For each genome we identified the canonical Rep and CP ORFs based
114 on gene annotations and product descriptions, extracted their nucleotide and amino acid
115 sequences and stored them as separate fields. Any additional predicted CDS were
116 concatenated into “other proteins” fields (names, amino acid sequences and nucleotide
117 sequences).

118 The resulting dataset, including host and geographic information, was manually inspected
119 to ensure that all entries had complete circular genomes and coherent annotation of Rep and
120 CP. Where necessary, minor inconsistencies in the genus or species names were corrected
121 to match the latest ICTV taxonomy. The final curated table (**Supplementary Table 1**) was
122 used as the master source for populating the GenomoBase database.

123 Host, isolation source and country information were also extracted from the GenBank
124 records. For each genome we collected the host, isolation_source, country and
125 collection_date qualifiers when present. Host names were standardized and manually
126 grouped into 11 ecological categories, including mammals, birds, plants, insects, fungi,
127 molluscs, reptiles, arachnids, actinopterygian fishes, humans and environmental samples.
128 This manual curation step allowed us to separate human-associated records from other
129 mammalian hosts and to distinguish non-living matrices (e.g. wastewater or environmental
130 swabs) from host organisms.

131 Country strings were cleaned to remove subnational descriptors and harmonized to ISO-
132 like country names (e.g. "USA: California" to "USA"). When possible, additional locality
133 information (regions, cities or sampling sites) was kept as free-text descriptors in the
134 database. These standardized fields are used in the statistics and visualization pages of the
135 portal to summarize host and geographic patterns.

136 **Metagenomic survey and SRA/BioSample integration**

137 Because most genomoviruses have been discovered by metagenomics, we implemented an
138 "Explore Metagenomes" module to connect reference genomes with public high-
139 throughput sequencing projects. As a starting point we used Serratus, a large-scale

alignment platform for viral sequence discovery [12], to identify SRA runs with reads matching genomovirus references. We queried the Serratus explorer for alignments against genomovirus targets and retained runs whose best hits satisfied the specific filters on alignment quality: genome alignment scores and percent identity between 75% and 100% (query date November 2025). This range was chosen to capture both close relatives and more divergent genomovirus-like sequences while excluding low-similarity matches that are likely spurious.

We extracted the unique run identifiers and obtained 613 distinct SRA runs with putative genomovirus signal. To enrich these with standardized metadata we developed a custom Python script (fetch_sra_metadata.py) that queries the NCBI SRA and BioSample API for NCBI-hosted runs and the ENA browser API for European accessions (e.g. SAMEA and ERS identifiers). For each run, the script retrieves: SRA accession, BioSample accession, release date, SRA “organism” or “environment” label, BioSample “host” and “isolation source” attributes, when available location string and, when possible, latitude/longitude coordinate. All retrieved fields are cleaned to remove control characters and harmonized into a JSON structure with one dictionary per SRA run. For BioSamples hosted at ENA (typically SAMEA, ERS or ERP accessions), the script parses XML records to extract host and geographic information, including coordinates embedded in free-text fields. The final JSON file (**Supplementary file 1**) is both downloadable from the portal and used internally to populate the metagenomes table in PostgreSQL. A small fraction of runs lack either explicit location names or coordinates; in these cases, only the available fields are stored. The Explore Metagenomes web page uses these curated attributes to generate summary plots and an interactive map of sampling locations.

163 **Literature retrieval and curation**

164 To provide an up-to-date overview of genomovirus research, we implemented a literature
165 module driven by a dedicated PubMed query. A custom Python script (fetch_pubmed.py)
166 uses NCBI E-utilities to search PubMed for articles containing the terms "Genomoviridae",
167 "genomovirus" and the names of each of the ten ICTV-recognized genera in titles, abstracts
168 or Medical Subject Headings terms. The script retrieves the PubMed IDs of all matching
169 records, then uses esummary to obtain titles, journal names, publication years and DOIs,
170 and efetch to access additional metadata where necessary. Each record is annotated with
171 one or more broad categories (e.g. "environmental metagenomics", "animal viromes",
172 "human-associated viromes", "fungal genomoviruses", "reviews") based on keyword
173 matching in titles and abstracts. The output is stored in a JSON file, which is rendered on
174 the Literature page as an interactive, searchable table with filters by category, year and
175 journal.

176 **BLAST services**

177 To enable sequence similarity searches, GenomoBase provides local BLASTN and
178 BLASTP services based on NCBI BLAST+ v2.17.0 [13]. From the curated genomovirus
179 table we generated FASTA files for complete genomes, Rep proteins and CP proteins,
180 which were converted into dedicated BLAST databases. User-submitted nucleotide or
181 protein queries are run against these databases with a default parameters and E-value
182 threshold of 1×10^{-3}, and the results are returned as sortable tables with basic alignment
183 statistics and direct links from hit accessions to the corresponding GenomoBase virus detail
184 pages.

185 **Database implementation, BLAST services and web interface**

186 GenomoBase is implemented using the Django web framework (Python v3.10+) with a
187 PostgreSQL backend. A primary database model stores curated genome-level information,
188 including accession, taxonomy, host and geographic metadata, as well as the nucleotide and
189 amino acid sequences of Rep, CP and additional CDS. Separate tables store metagenomic
190 run metadata and literature entries. The application is deployed on an Amazon Web
191 Services (AWS) EC2 instance, which provides a stable, scalable environment with regular
192 backups and straightforward future upgrades of both the database and the web front-end.
193 The web interface is styled using Bootstrap and custom CSS, with dedicated views for the
194 index page, master species list, species detail, statistics, BLAST search, metagenome
195 explorer, downloads, literature and external resources. Summary statistics (e.g. numbers of
196 species per genus, host category counts and country distributions) are computed
197 dynamically using Django's ORM and rendered as charts (bar plots and maps) using
198 JavaScript-based visualizations.

199 **Results**

200 **Resource content and example analyses**

201 GenomoBase organizes the currently ICTV-recognized *Genomoviridae* diversity into ten
202 genera, with *Gemycircularvirus* (126 species) and *Gemykibivirus* (50 species) clearly
203 dominating the dataset, and the remaining genera represented by between 1 and 16 species
204 each (Fig. 1A). Genome lengths range from ~1.99 to 3.9 kb (mean ~ 2.19 kb), consistent
205 with compact CRESS-DNA genomes (Fig. 1B). The upper bound of ~3.9 kb corresponds to
206 the combined size of the three genomic segments of the tripartite *Gemytripvirus fugra1*.

Beyond taxonomic composition and genome size, host annotations reveal associations with 12 ecological categories, including mammals, birds, plants, insects, fungi, molluscs, reptiles, arachnids, actinopterygian fishes and non-living matrices. Mammals represent the largest category (37.66%, excluding humans) in terms of species–host combinations, followed by plants (14.64%), birds (13.39%) and insects (10.88%) (Fig. 1C). A discrete subset of seven genomoviruses belonging to *Gemykibivirus* (5 spp.), *Gemykrogvirus* and *Gemyvongvirus* genera has been recovered from human-associated samples, highlighting a small but potentially important group of lineages at the interface between environmental and medical virology. Geographic metadata indicate that genomoviruses have been reported from at least ~19 countries, with especially high counts from the USA (97 spp.), New Zealand (38 spp.), China (26 spp.) and Brazil (21), illustrating their broad global distribution (Fig. 1D).

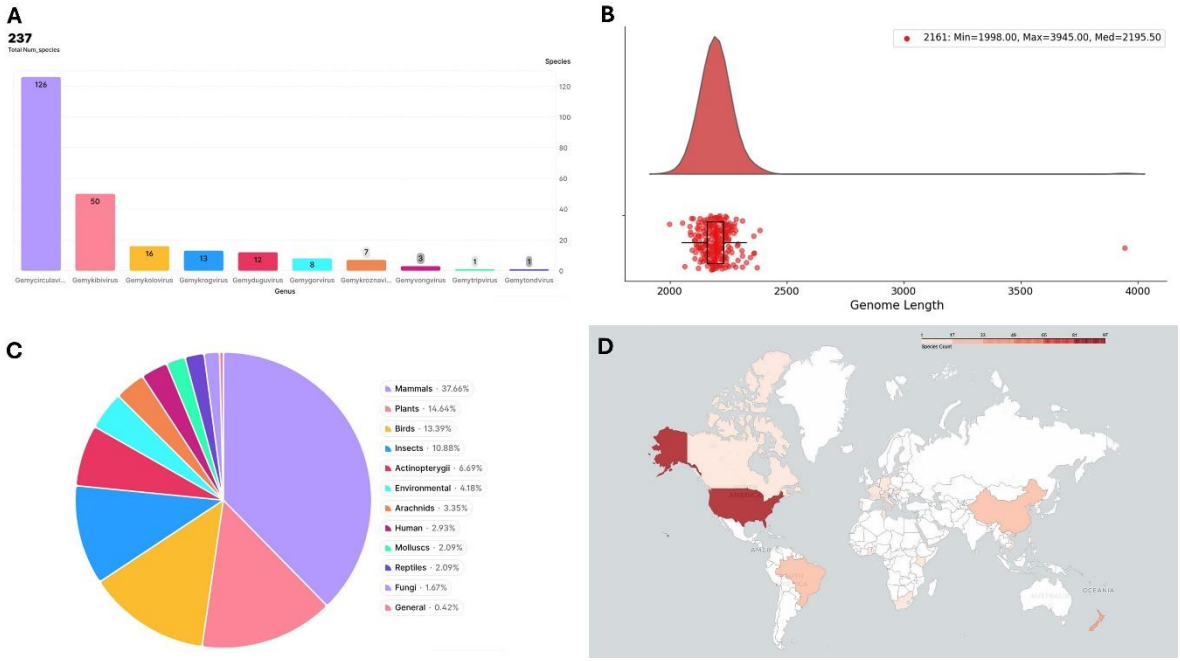


Figure 1. Overview of *Genomoviridae* species and associated metadata in GenomoBase. (A) Bar chart showing the distribution of currently classified species across ICTV-recognized genera. (B) Genome length

222 distribution visualized as a raincloud plot combining a density curve, individual genomes plotted as points,
223 and a boxplot. (C) Broad host categories for genomoviruses, summarized as a pie chart based on curated host
224 metadata. (D) Global map indicating the number of genomovirus species reported per country according to
225 available sampling location information.

226 Based on the harmonized host annotations for the 613 *Genomoviridae*-positive SRA runs,
227 most detections fall into a set of recurrent host categories (Fig. 2A). The largest single class
228 corresponds to "environmental samples", which account for 172/613 runs (28.1%) and
229 encompass marine and freshwater systems, wastewater, air, soils and food-associated
230 microbiomes. Among runs with biological hosts, human-associated samples (human gut,
231 general human metagenomes, oral/skin/vaginal metagenomes) constitute the most frequent
232 category, with 99 runs (16.0%), followed by metagenomes from birds (85 runs, 13.9%).
233 Additional mammalian hosts include giant pandas (*Ailuropoda melanoleuca*; 29 runs,
234 4.7%), domestic pigs (*Sus scrofa*; 23 runs, 3.8%), mice (*Mus musculus*; 21 runs, 3.4%),
235 bovines (*Bos* spp.; 17 runs, 2.8%) and *Macaca mulatta* (7 runs, 1.1%). There are also
236 smaller contributions from *Xenopus laevis* (11 runs, 1.8%) and samples with unknown host
237 information (50 runs, 8.2%). Beyond these dominant categories, the remaining runs form a
238 long tail of individually represented wildlife and domestic animal hosts. Collectively, these
239 patterns indicate that genomovirus-related reads are detected across a broad spectrum of
240 vertebrate hosts and environmental matrices, while a substantial fraction of SRA runs still
241 lacks precise host information.

242 Geographic metadata indicate that 462 of 613 runs (75.3%) include a textual location
243 descriptor, corresponding to 180 distinct place names distributed across 47 countries or
244 territories. In addition, 282 runs (46.0%) carry explicit latitude/longitude coordinates,
245 enabling map-based visualization of sampling sites. Within this set, the largest

contributions come from projects in China (approximately 165 runs) and the USA (approximately 96 runs), followed by individual efforts in India, Ireland, Singapore, South Korea, Norway, Canada and several European and Latin American countries (Fig. 2B).

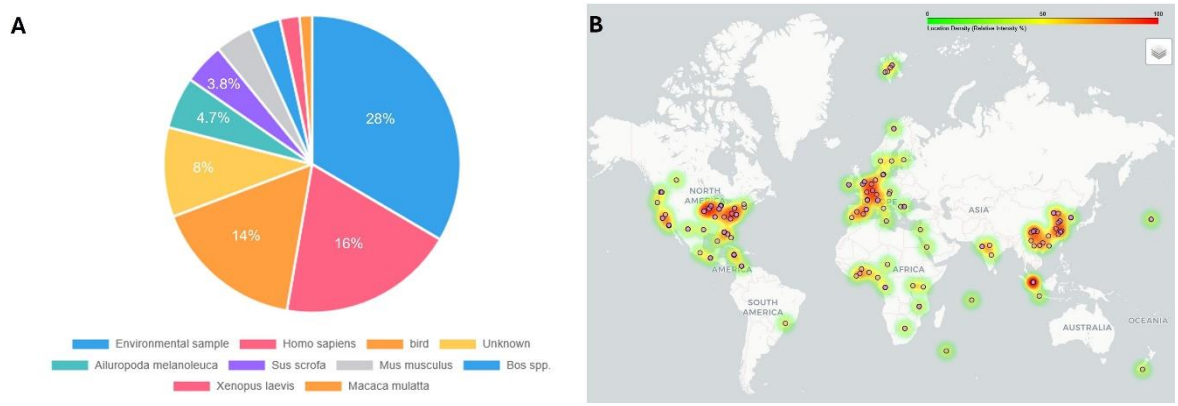


Figure 2. Overview of metagenomic runs with genomovirus detections in GenomoBase. (A) Pie chart summarizing SRA runs with genomovirus hits, grouped into broad host or sample categories (e.g. birds, humans, domestic pigs, environmental and others). (B) Global heat map of sampling locations for the 282 runs with latitude/longitude metadata, illustrating the relative density of genomovirus-positive metagenomes. Warmer colours indicate higher local density across countries and regions.

Discussion

The rapid expansion of genomovirus sequence data, largely driven by metagenomic studies, has created a challenge for data integration and standardization. Until now, genomovirus sequences and annotations have been scattered across general taxonomic resources, GenBank (https://www.ncbi.nlm.nih.gov/datasets/taxonomy/1910928/), ViralZone (https://viralzone.expasy.org/7382), ICTV (VMR_MSL40.v2.20251013) and individual metagenomic surveys, without a family-focused curation framework. GenomoBase addresses this gap by providing, to our knowledge, the first integrated resource that aligns *Genomoviridae* genomes with the official ICTV taxonomy, harmonized host and geographic metadata and systematically filtered metagenomic detections. By centralizing

265 these data within a dedicated web interface, the platform facilitates taxonomic curation and
266 enables ecological and evolutionary analyses that are difficult to perform from isolated
267 records.

268 The aggregated view provided by GenomoBase makes current sampling biases in
269 Genomoviridae immediately apparent. Geographic metadata for formally described species
270 indicate detections in 19 countries across all inhabited continents, while the metagenomic
271 module reveals SRA runs potentially containing genomoviruses from around 47 countries
272 worldwide. This pattern likely reflects sequencing effort rather than the true biogeography
273 of the family, even as it clearly indicates that members of the family *Genomoviridae* are
274 widespread across continents and environments. GenomoBase also highlights a broad
275 apparent host spectrum in both the species and metagenomic datasets, including mammals,
276 birds, invertebrates, plants and environmental samples; although infection has not been
277 experimentally confirmed for all of these hosts [4], many of these associations remain
278 plausible and merit further investigation. Of particular interest, seven formally recognized
279 species belonging to three genera have been reported in human clinical samples [6, 8], and
280 genomoviruses are detected in at least ~90 human-associated metagenomes, including gut
281 [14], vaginal [7] and brain [15]. These observations suggest that genomovirus are recurrent
282 components of the human virome and underline the value of continued surveillance of this
283 family in a public-health context, especially in light of recent evidence that at least one
284 gemykibivirus can actively replicate in human cells [10].

285 A second key contribution of GenomoBase is that it does not merely summarize existing
286 detections, but explicitly links ICTV-aligned reference genomes with Serratus-derived
287 metagenomic reads and contigs. Connecting these “orphan” SRA sequences to curated

288 *Genomoviridae* genomes shows that members of this family are ubiquitous components of
289 environmental and host-associated viromes, often detected in studies that were not designed
290 to target them. Making such hidden detections visible and searchable within a family-
291 specific context allows researchers to identify under-sampled hosts and environments, and
292 to prioritize public SRA runs for targeted re-analysis or assembly of novel genomovirus
293 genomes. In this sense, GenomoBase functions not only as a static catalogue but also as a
294 hypothesis-generating platform for future work on genomovirus diversity, ecology and
295 evolution.

296 **Conclusions and future directions**

297 GenomoBase provides a focused, user-friendly platform for the *Genomoviridae* research
298 community and for researchers interested more broadly in circular ssDNA viruses. By
299 integrating ICTV-based reference genomes, harmonized host and geographic metadata,
300 Serratus-filtered SRA detections, local BLAST services and an automatically updated
301 literature catalogue, the resource enables users to (i) search and download curated
302 genomovirus genomes and genes, (ii) explore host and geographic patterns of diversity, and
303 (iii) investigate the occurrence of genomoviruses across public metagenomic datasets. At
304 the same time, the current implementation makes explicit several limitations of the
305 available data, such as incomplete host annotations and strong geographic sampling biases,
306 which can guide future sequencing surveys. GenomoBase is freely available and will be
307 periodically updated in line with ICTV releases and Serratus-based screenings, and its
308 underlying framework can be readily extended to other CRESS-DNA virus families and
309 related viral taxa.
310

References

1. Krupovic M, Varsani A, Kazlauskas D, et al (2020) *Cressdnaviricota* : a Virus Phylum Unifying Seven Families of Rep-Encoding Viruses with Single-Stranded, Circular DNA Genomes. J Virol 94:e00582-20. https://doi.org/10.1128/JVI.00582-20

2. Krupovic M, Ghabrial SA, Jiang D, Varsani A (2016) Genomoviridae: a new family of widespread single-stranded DNA viruses. Arch Virol 161:2633–2643. https://doi.org/10.1007/s00705-016-2943-3

3. Yu X, Li B, Fu Y, et al (2010) A geminivirus-related DNA mycovirus that confers hypovirulence to a plant pathogenic fungus. Proc Natl Acad Sci USA 107:8387–8392. https://doi.org/10.1073/pnas.0913535107

4. Varsani A, Krupovic M (2021) Family Genomoviridae: 2021 taxonomy update. Arch Virol 166:2911–2926. https://doi.org/10.1007/s00705-021-05183-y

5. Liu S, Xie J, Cheng J, et al (2016) Fungal DNA virus infects a mycophagous insect and utilizes it as a transmission vector. Proc Natl Acad Sci USA 113:12803–12808. https://doi.org/10.1073/pnas.1608013113

6. Tuladhar ET, Shrestha S, Vernon S, et al (2024) Gemykibivirus detection in acute encephalitis patients from Nepal. mSphere 9:e00219-24. https://doi.org/10.1128/msphere.00219-24

7. Ramos EDSF, Tozetto-Mendoza TR, Bortoletto P, et al (2024) Characterization of CRESS-DNA viruses in human vaginal secretions: An exploratory metagenomic investigation. Journal of Medical Virology 96:e29750. https://doi.org/10.1002/jmv.29750

8. Bezerra RS, Bitencourt HT, Covas DT, et al (2020) Metagenomic identification of human Gemykibivirus-2 (HuGkV-2) in parenterally infected blood donors from the Brazilian Amazon. International Journal of Infectious Diseases 98:249–251. https://doi.org/10.1016/j.ijid.2020.06.096

9. Silvério BS, Sanz Duro RL, De Sousa LLF, et al (2025) Detection and Phylogenetic Analysis of Emerging Human-Associated Gemykibivirus-2 in *Molossus molossus* Bat From Brazil. Journal of Medical Virology 97:e70154. https://doi.org/10.1002/jmv.70154

10. Yang R, Yan H, Wang Y, et al (2025) Transcriptome Analysis Reveals Gemykibivirus Infection Induces Mitochondrial DNA Release in HEK293T Cells. Viruses 17:1331. https://doi.org/10.3390/v17101331

11. Cock PJA, Antao T, Chang JT, et al (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25:1422–1423. https://doi.org/10.1093/bioinformatics/btp163

12. Edgar RC, Taylor B, Lin V, et al (2022) Petabase-scale sequence alignment catalyses viral discovery. Nature 602:142–147. https://doi.org/10.1038/s41586-021-04332-2

13. Camacho C, Coulouris G, Avagyan V, et al (2009) BLAST+: architecture and applications. BMC Bioinformatics 10:421. https://doi.org/10.1186/1471-2105-10-421

14. Dao TK, Pham TTN, Nguyen HD, et al (2025) Metagenomic Analysis of the Gastrointestinal Phageome and Incorporated Dysbiosis in Children with Persistent Diarrhea of Unknown Etiology in Vietnam. Pathogens 14:985. https://doi.org/10.3390/pathogens14100985

348 15. Lamberto I, Gunst K, Müller H, et al (2014) Mycovirus-Like DNA Virus Sequences from Cattle Serum
349 and Human Brain and Serum Samples from Multiple Sclerosis Patients. Genome Announc 2:e00848-
350 14. https://doi.org/10.1128/genomeA.00848-14

351

352

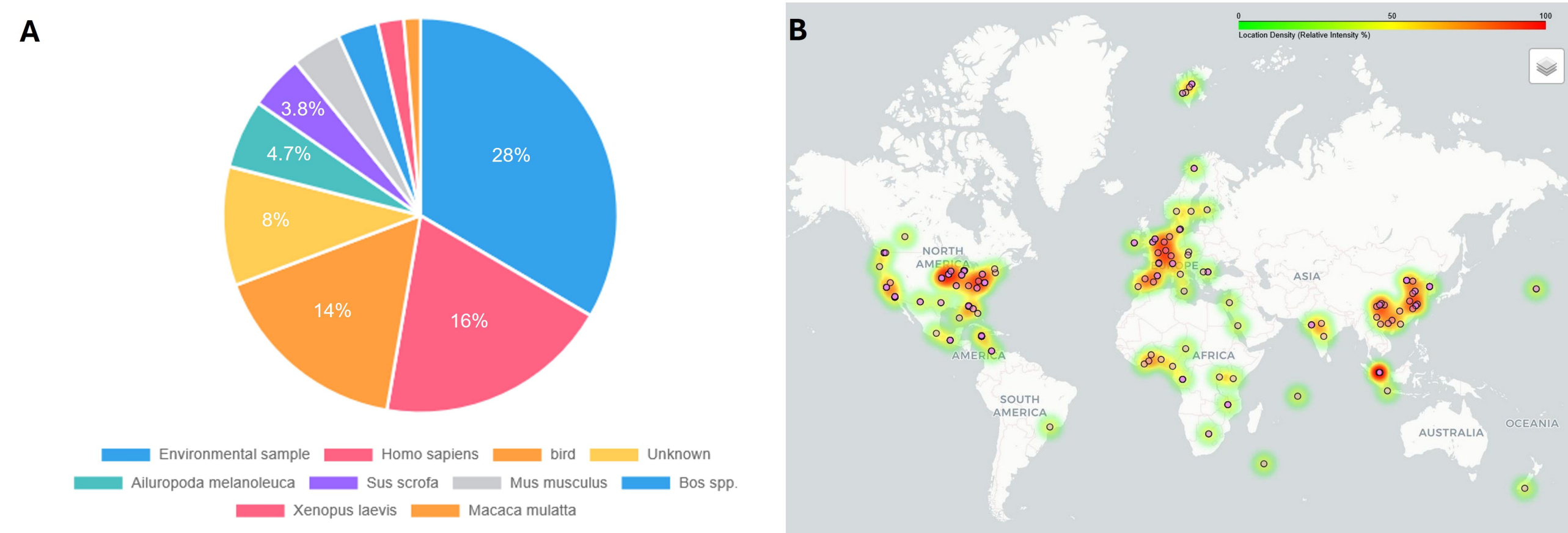
Statements & Declarations

Funding: This study was supported by grants from the Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI), Mexico (Ciencia de Frontera 2025; grant CBF-2025-G-902).

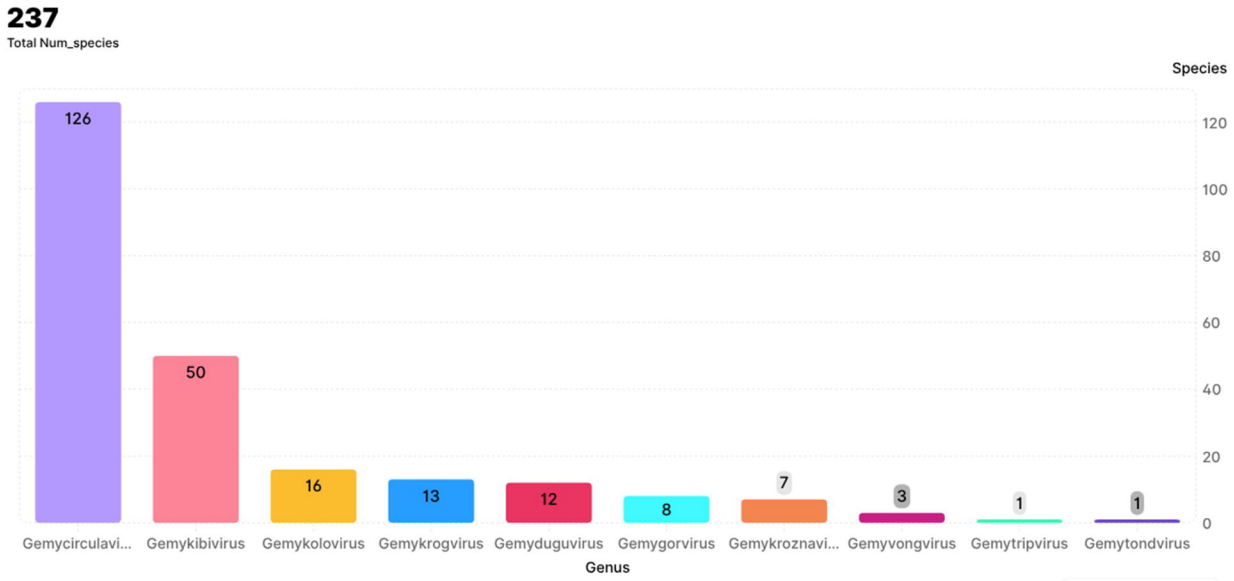
Competing interests: The author declares no competing interests.

Data, material and/or code availability: The online instance of GenomoBase is freely accessible without registration at https://www.genomobase.org. Curated datasets, processed tables, and the custom scripts used to build and update GenomoBase, together with sample input data and corresponding output, are publicly available via Zenodo (**DOI: 10.5281/zenodo.17860246**).

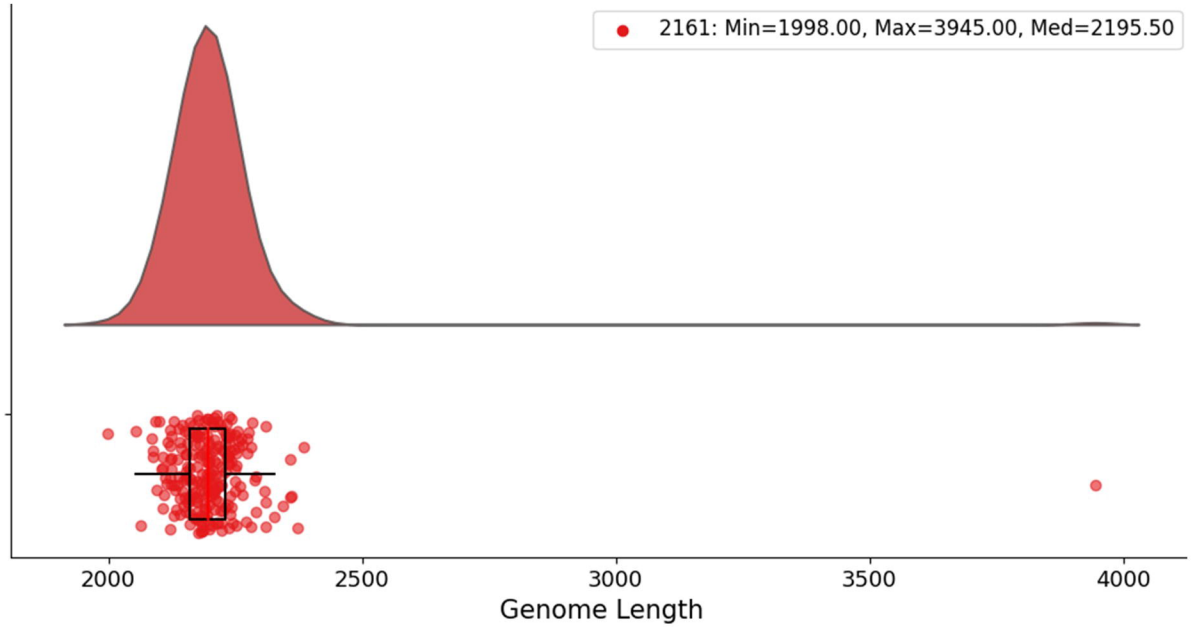
Authors' contributions: Y.C.-C. conceived and designed the study, developed the database and web server, curated and harmonized the data, performed the analyses, generated figures and tables, and wrote and revised the manuscript.


A
28%
16%
14%
8%
4.7%
3.8%
Environmental sample
Homo sapiens
bird
Unknown
Ailuropoda melanoleuca
Sus scrofa
Mus musculus
Bos spp.
Xenopus laevis
Macaca mulatta
B
0
50
100
Location Density (Relative Intensity %)
NORTH AMERICA
AMERICA
SOUTH AMERICA
ASIA
AFRICA
AUSTRALIA
OCEANIA

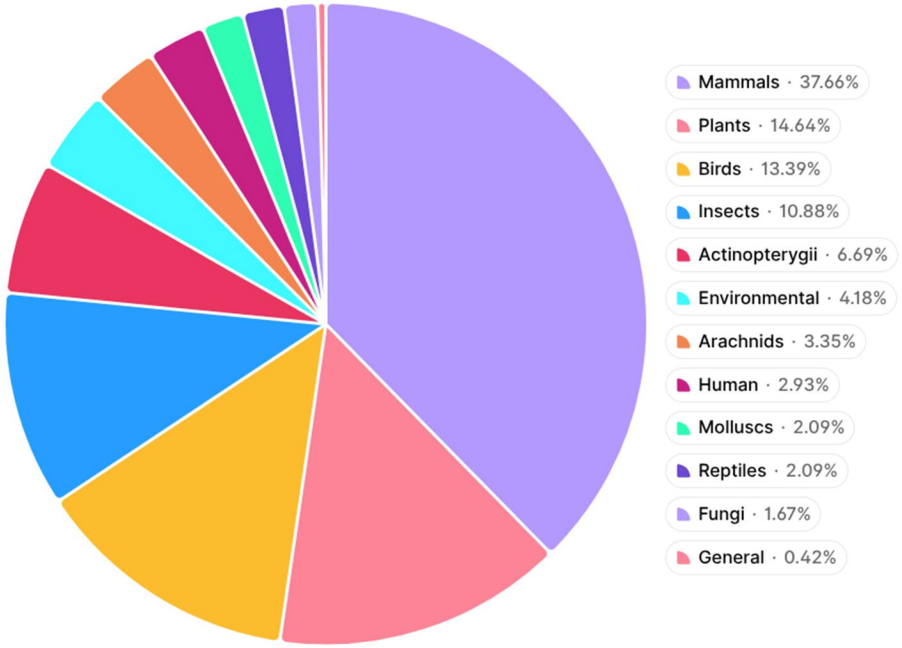
A


237
Total Num_species
Species
126
50
16
13
12
8
7
3
1
1
120
100
80
60
40
20
0
Gemycirculavi...
Gemykibivirus
Gemykolovirus
Gemykrogvirus
Gemyduguvirus
Gemygorvirus
Gemykroznavi...
Gemyvongvirus
Gemytripvirus
Gemytondvirus
Genus

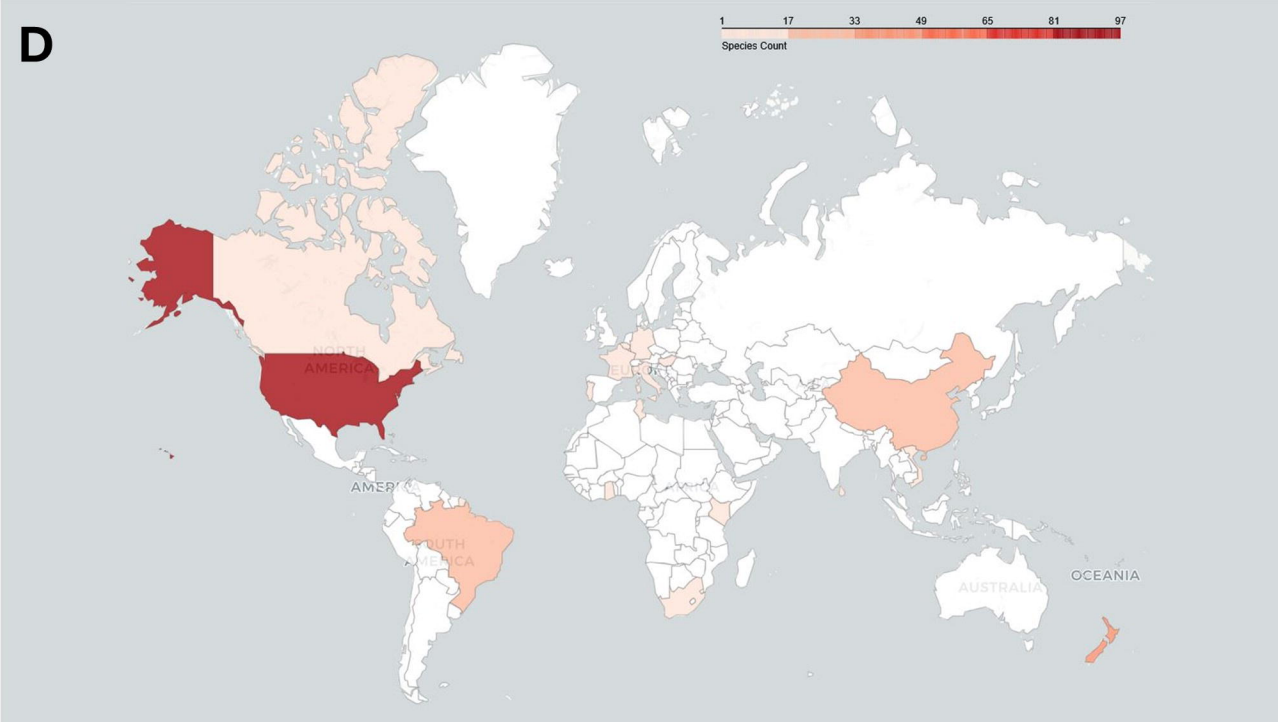
B


2161: Min=1998.00, Max=3945.00, Med=2195.50
2000
2500
3000
3500
4000
Genome Length

C


Mammals · 37.66%
Plants · 14.64%
Birds · 13.39%
Insects · 10.88%
Actinopterygii · 6.69%
Environmental · 4.18%
Arachnids · 3.35%
Human · 2.93%
Molluscs · 2.09%
Reptiles · 2.09%
Fungi · 1.67%
General · 0.42%

D


1
17
33
49
65
81
97
Species Count
OCEANIA