**Sex influences the mosquito virome in a host specific way**

**Côme Morel1, Patricia Gil1, Didier P. Alexandre Kaboré2,3, Antoni Exbrayat1, Florian Charriat1, Dieudonné Diloma Soma2,3, Simon P. Sawadogo2, Georges Anicet Ouédraogo3, Philippe Van de Perre4, Thierry Baldet1, Roch K. Dabiré2 and Serafin Gutierrez1,5\***

*1ASTRE, Cirad, INRAe, Univ Montpellier, Montpellier, France.*

*2**Institut de Recherche en Sciences de la Santé (IRSS), Bobo-Dioulasso 01BP 545 Burkina Faso.*

*3Université Nazi BONI, Bobo-Dioulasso, Burkina Faso.*

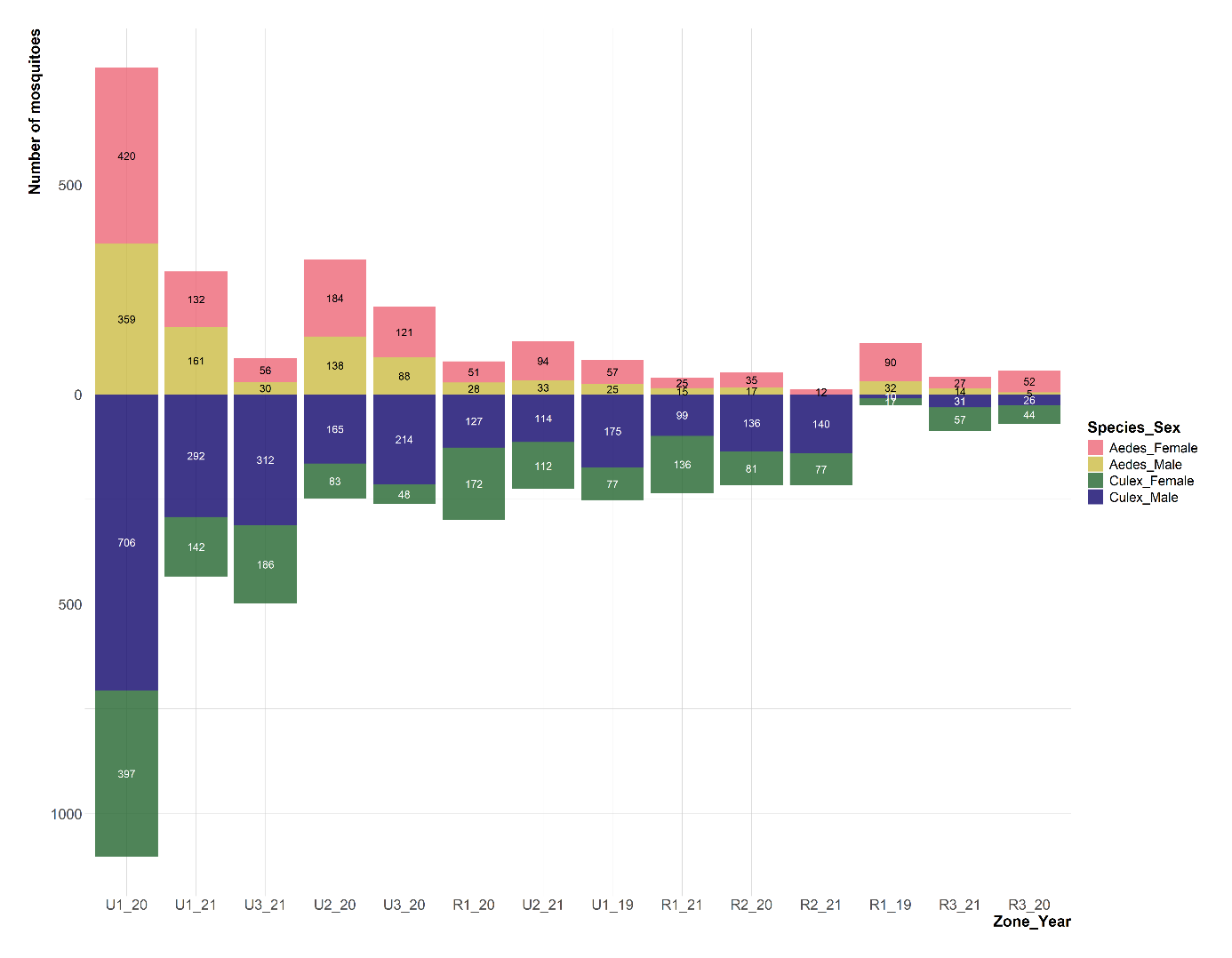
*4Pathogenesis and Control of Chronic and Emerging Infections, INSERM, University of Montpellier, Établissement Français du Sang, Montpellier, France.*

*5Laboratory of Virology, Montpellier University Hospital, Montpellier, France.*

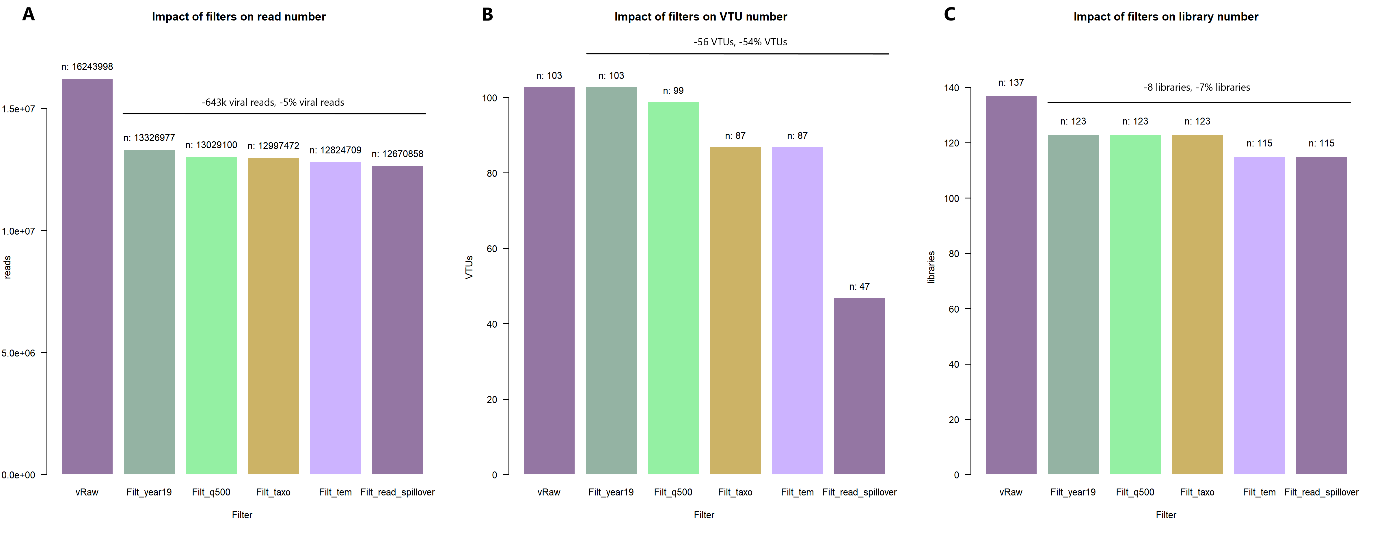
\*Corresponding author: Serafin Gutierrez

E-mail: [serafin.gutierrez@cirad.fr](mailto:serafin.gutierrez@cirad.fr)

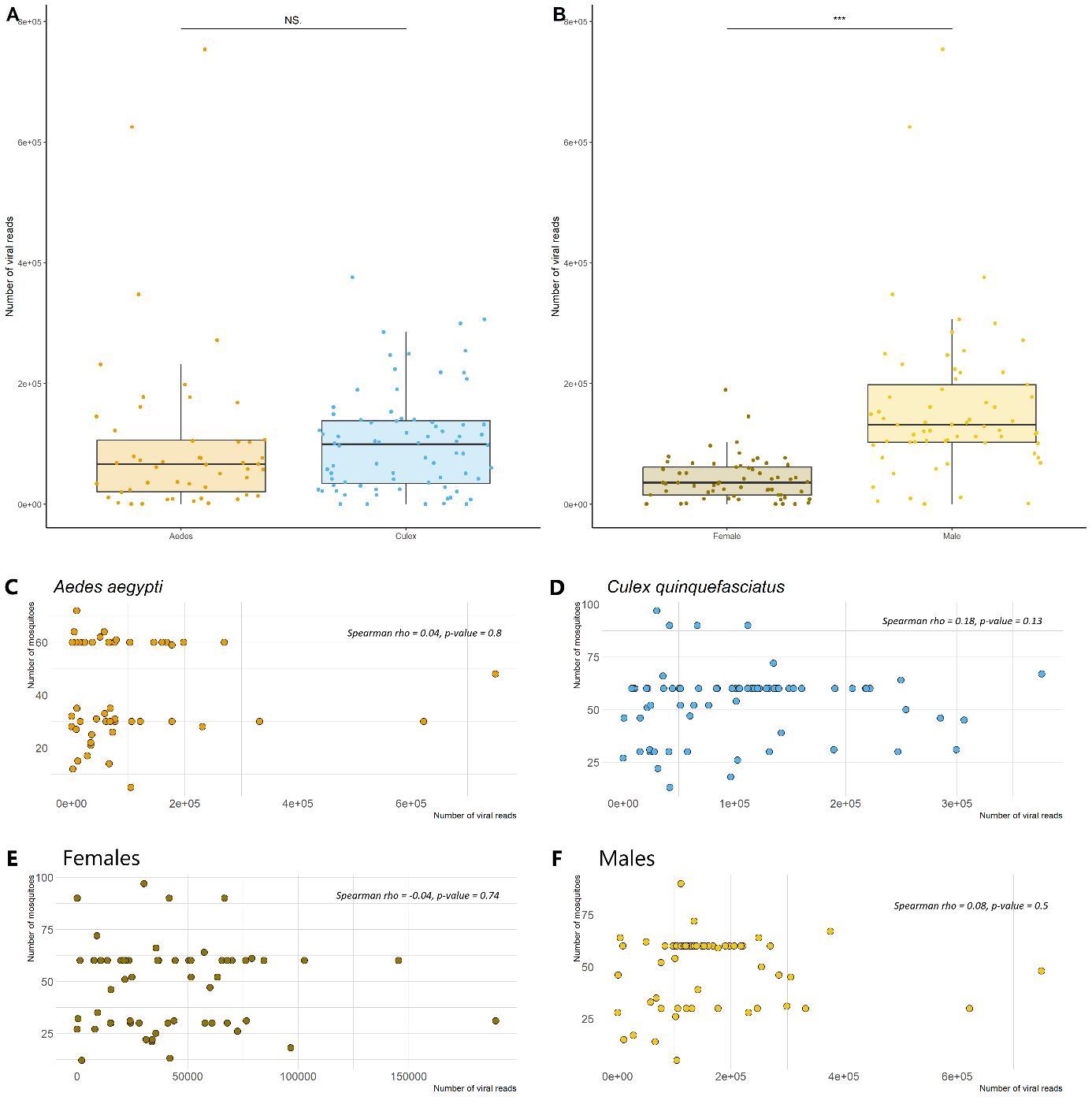
## **Supplemental figures and supplemental figures legends**

****

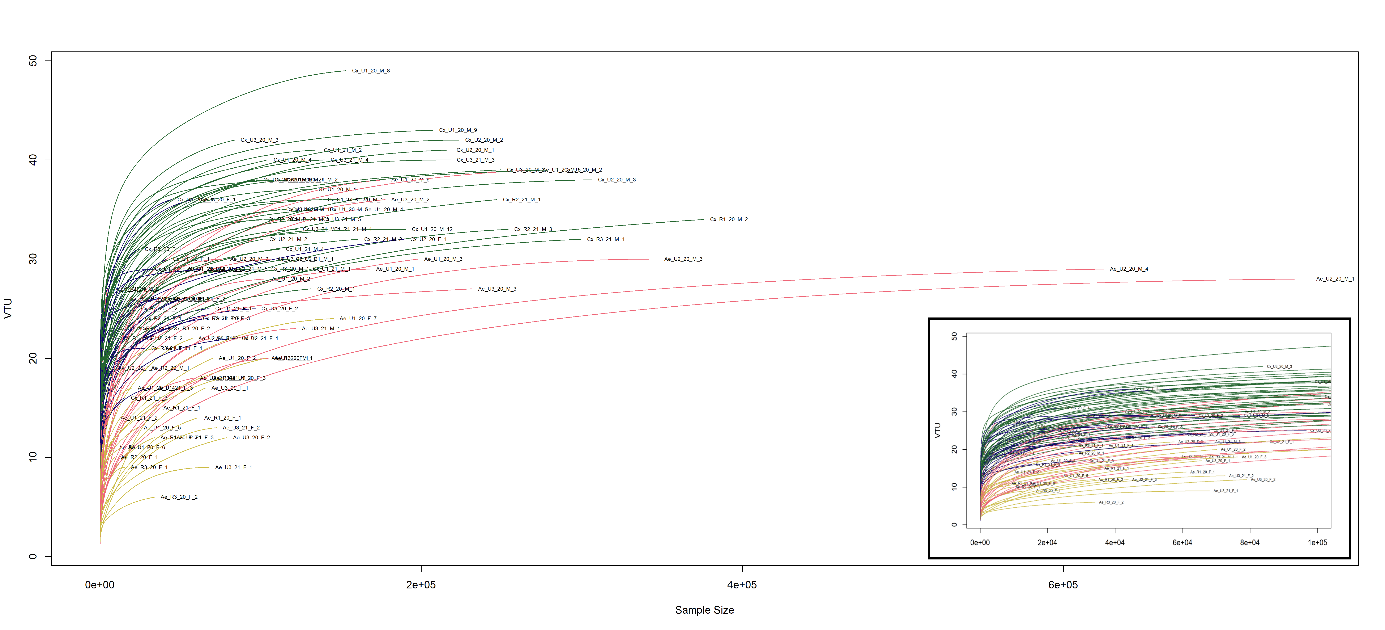
**Figure S1:** Number of individuals collected by mosquito species, sex, zone and year. Bar color stands for the couple mosquito species/sex (red: *Aedes* females, yellow: *Aedes* males, green: *Culex* females, blue: *Culex* males). The number of individuals of each category is shown in the center of bar.



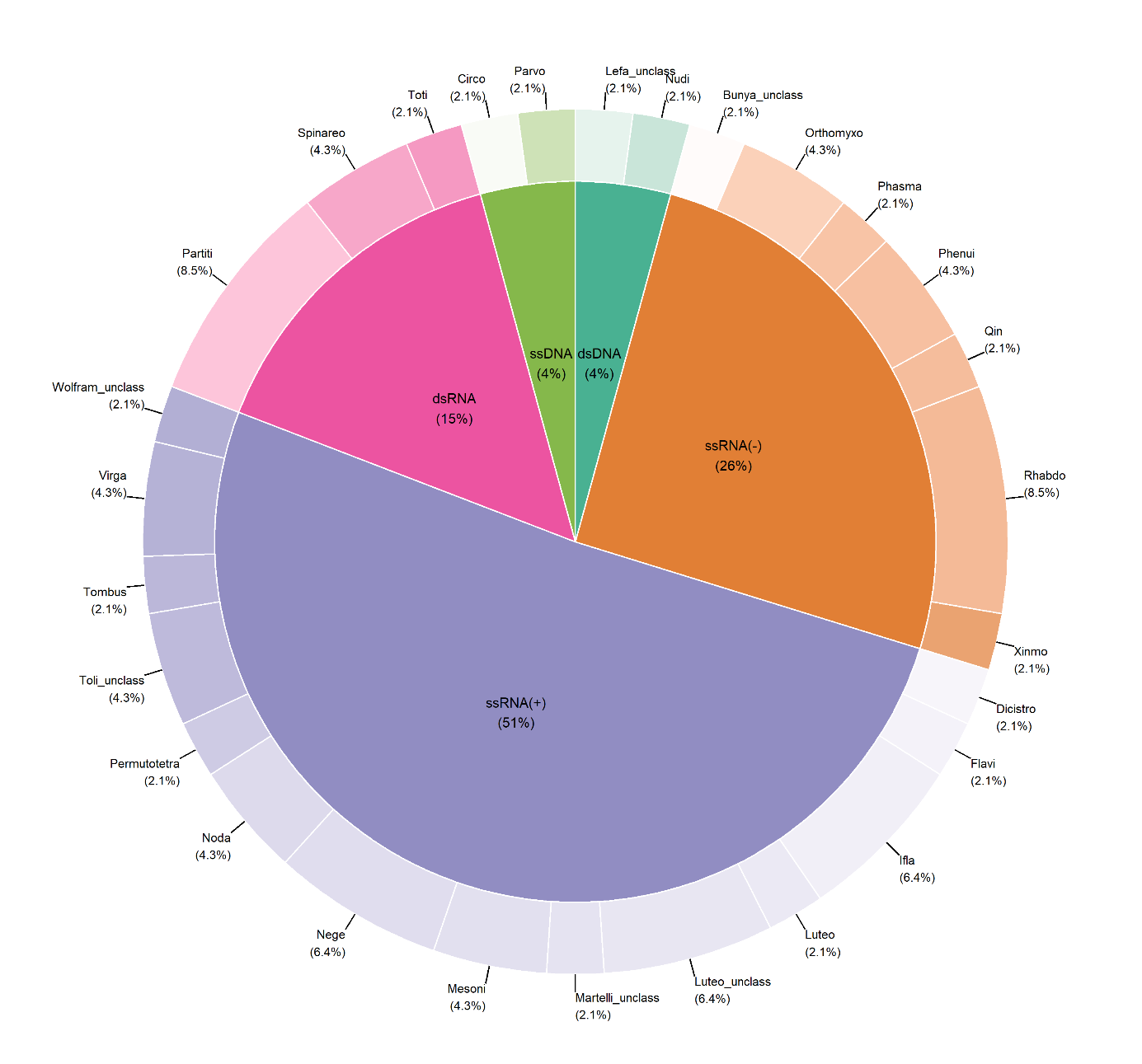
**Figure S2:**Impact of the filters performed on the dataset on the number of (A) viral reads, (B) VTUs and (C) libraries. Filters are shown on the x-axis (vRaw: raw numbers, Filt\_year19: removal of samples from year 2019, Filt\_q500: filter on contig length, Filt\_taxo: filter on taxonomy, Filt\_tem: filter based on controls, Filt\_read\_spillover: filter to avoid read spillover). The numbers on top of the bars indicate the amount of (A) viral reads, (B) VTUs and (C) libraries remaining after the corresponding filtration step. The total loss after all filtration steps, except the 2019 elimination stage (the first one), is provided on the horizontal bar above the bars.

**

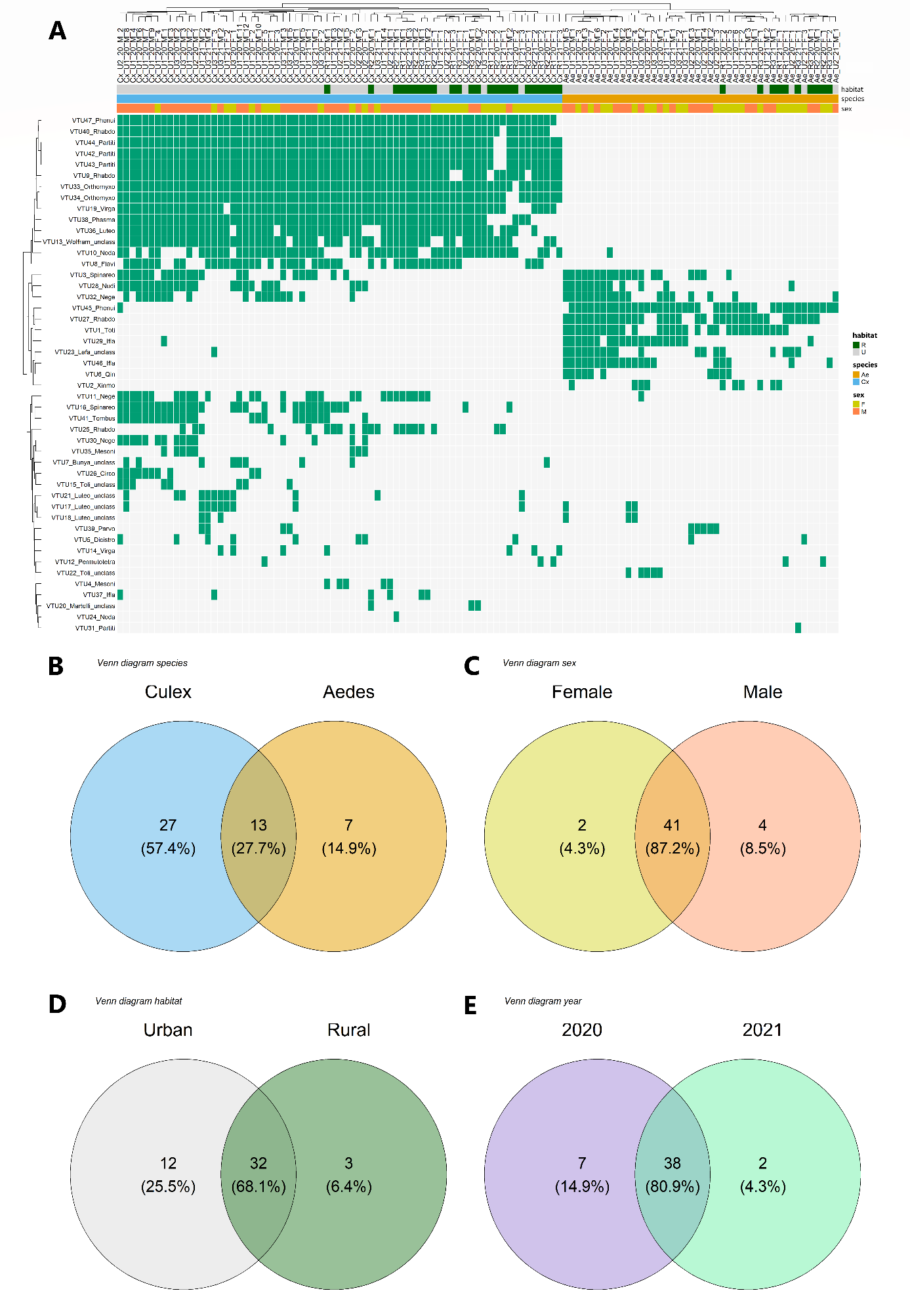
**Figure S3:** Panels A and B:Distribution of the number of viral reads per library according to mosquito species (A) and sex (B). The bar above the boxplots for each panel represents the value and degree of significance between the two groups presented, according to the wilcoxon test values. Significance codes: 0 '\*\*\*', 0.001 '\*\*', 0.01 '\*', 0.05 '.', 0.1 'NS'. Panels C, D, E and F: Correlation between the number of viral reads and the number of mosquitoes per library in Aedes or Culex mosquitoes (panels C and D, respectively), or in females (E), and males (F). Each point indicates the value for a library.



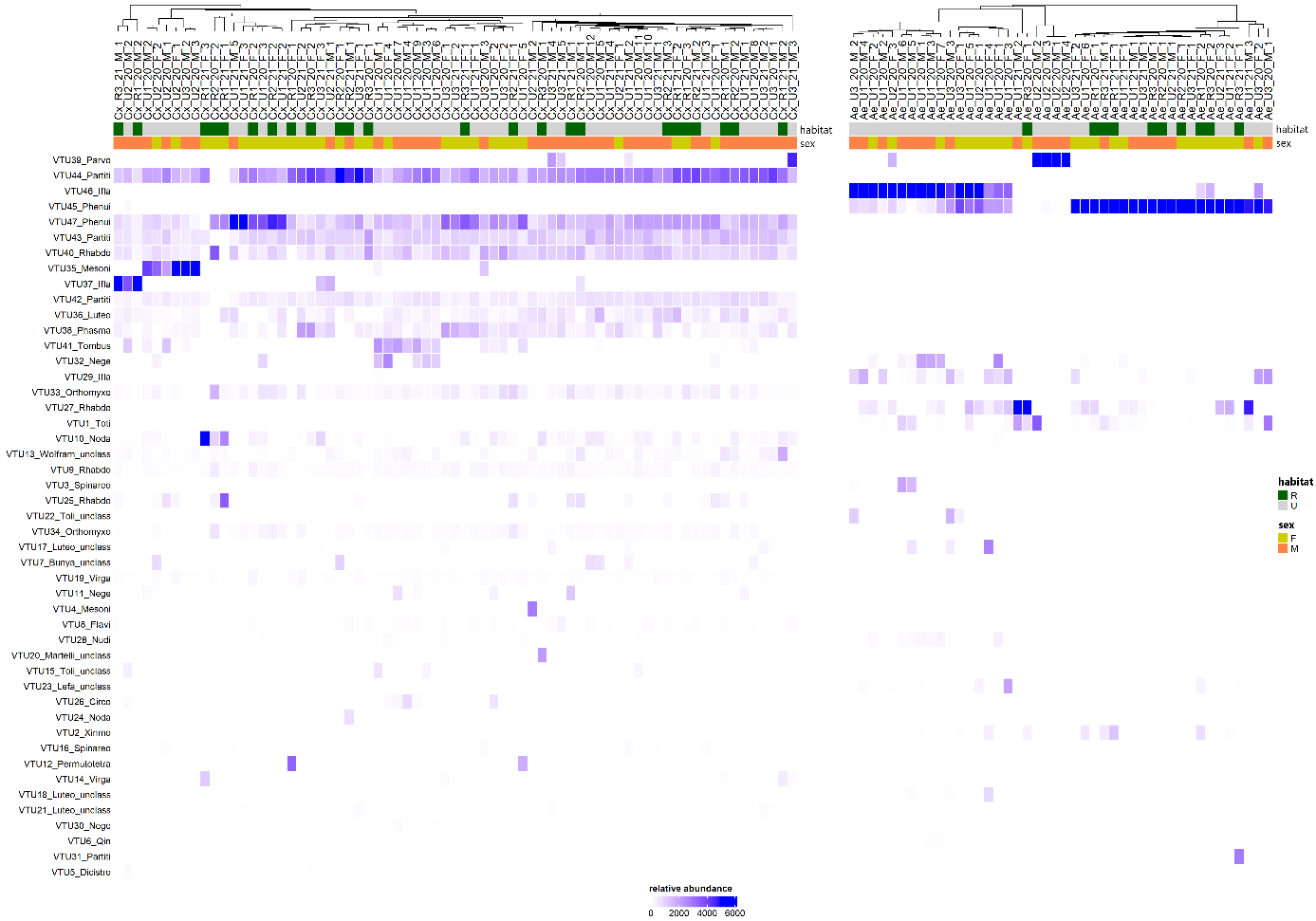
**Figure S4:** Rarefaction curves of the libraries. The x-axis indicates the number of viral reads per library and the y-axis indicates the number of viral taxonomic units (VTUs) per library. Curve color indicates species-sex pair (green: *Culex* males; blue: *Culex* females; pink: *Aedes* males; yellow: *Aedes* females). A zoom on the part below 100 000 reads has been made and is represented in the frame at the bottom right of the graph in order to facilitate a more precise understanding of the phenomena occurring at this scale.

**

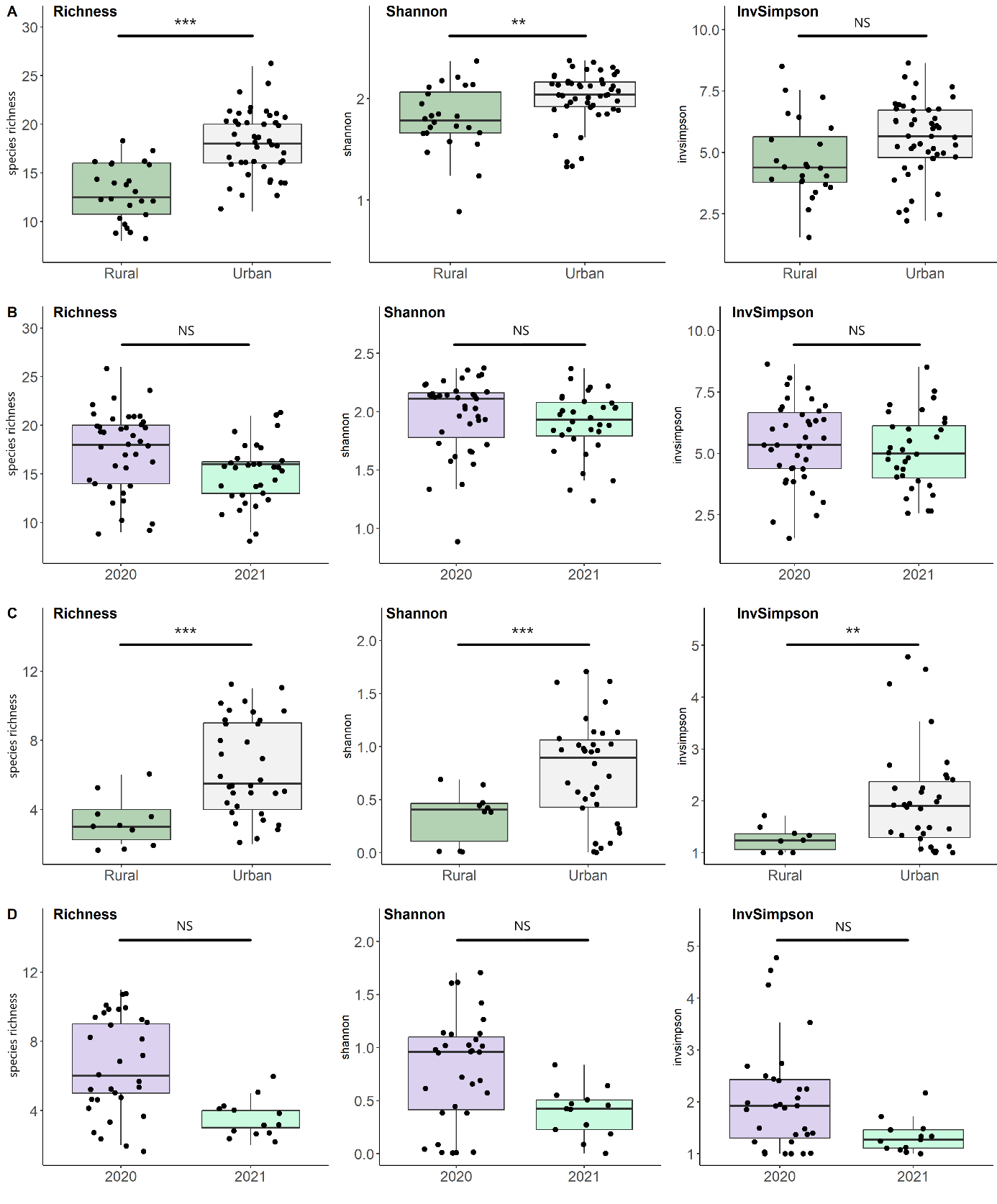
**Figure S5.** Distribution of viral taxonomic units (VTUs) among family-like clusters (external donut chart) and type of genome (inner pie chart). Percentages between brackets represent the proportion of VTUs in each group.



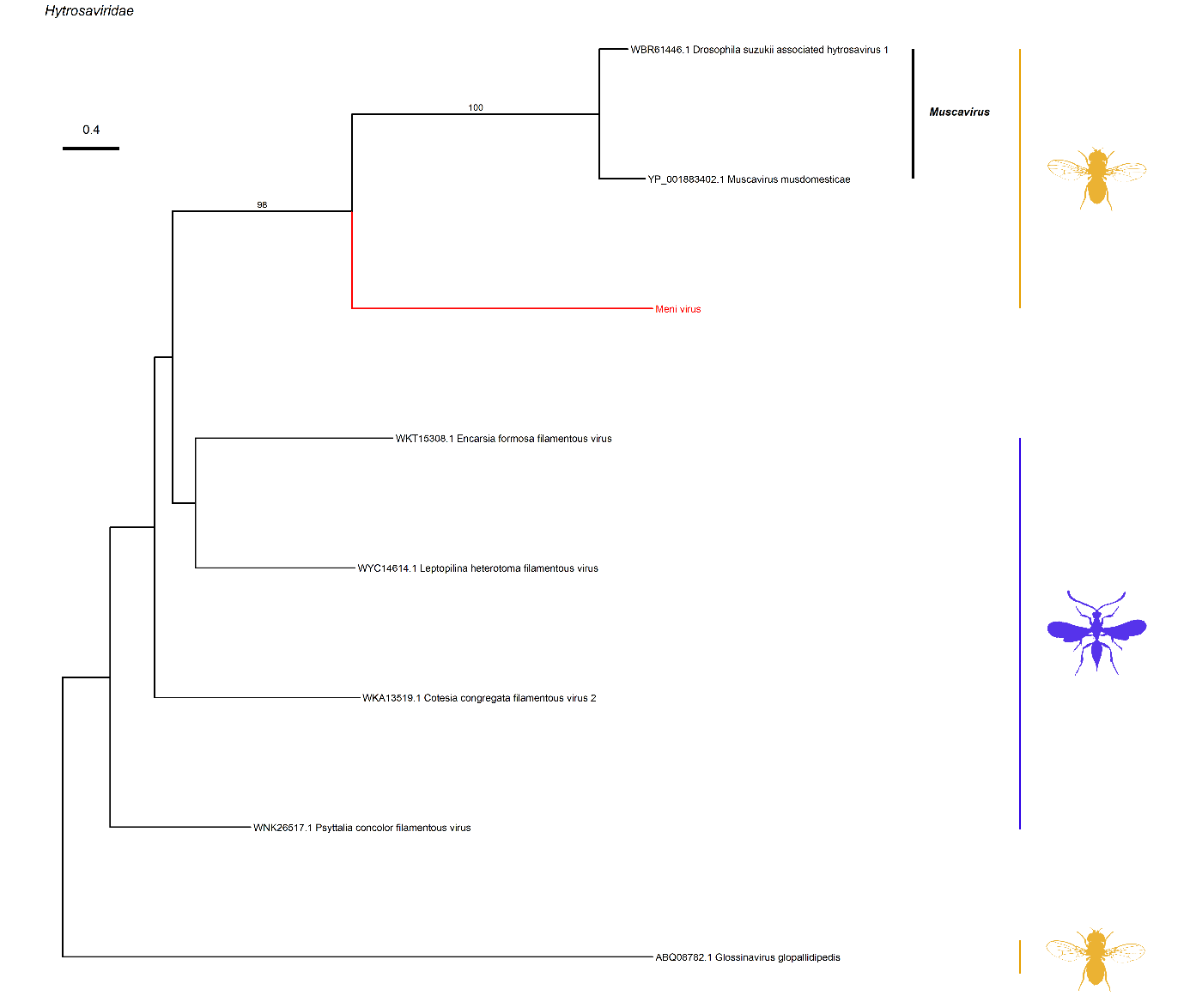
**Figure S6:** Presence-absence of viral taxonomic units (VTUs) among libraries and different ecological factors. (A) Library names are shown at the top of the heatmap (see Table S1 for explanation of names), together with a hierarchical grouping. The top three lines of the heatmap show the habitat, mosquito species and sex of each library respectively. The variable names are shown on the right. The tiles in these lines are coloured as follows: habitat (green: rural; grey: urban), species (blue: *Culex*; orange: *Aedes*) and sex (pink: male; yellow: female). Below the top three lines, lines indicate the presence (shown in green) or absence (shown in grey) of VTUs per library. VTUs are indicated on the left of the heatmap. The VTU presence is ranked according to the total number of samples in which they were detected. The VTU presence is separated into the following categories: Venn diagram separated by mosquito species (B), by sex (C), by habitat (D), by year (E). The numbers between brackets represent the proportion of each group among the total number of VTUs.



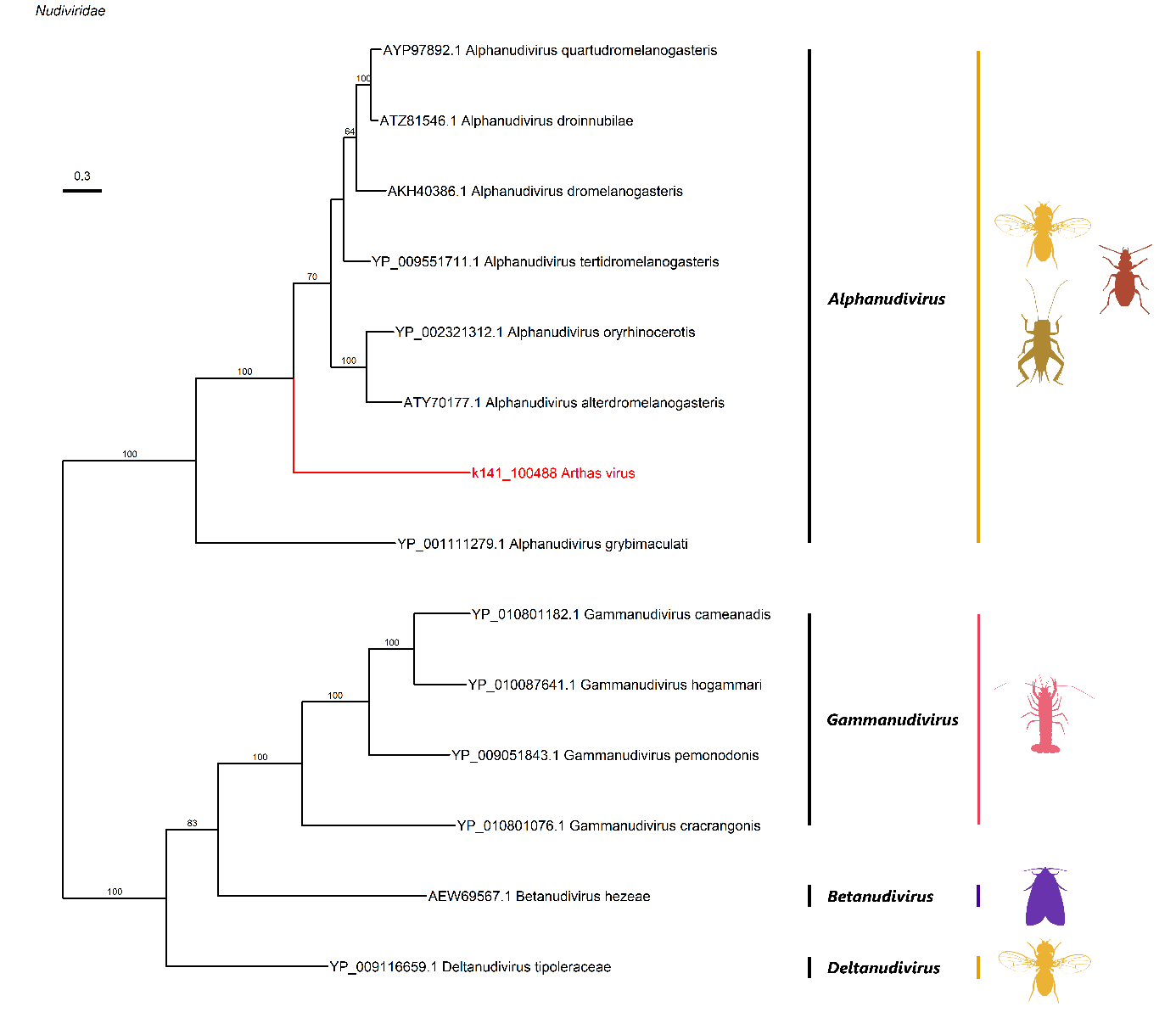
**Figure S7:** Relative abundance of viral taxonomic units (VTUs) among libraries separated by mosquito species. The heatmap for *Culex* libraries is shown on the left, while that for *Aedes* libraries is shown on the right. Library names are indicated on the top of both heatmaps (see Table S1 for name explanation), along with a hierarchical clustering. The two upper rows in the heatmap show habitat and sex of each library respectively. Variable names are shown on the right. Tiles in these rows are colored as follows: habitat (green: rural, gray: urban) and sex (orange: male; yellow: female). Below the upper three rows, rows show VTU abundance per library. VTUs are provided on the left of the heatmap. The VTUs are ranked following total read abundance. Tile color in VTU rows stands for read abundance as shown in the legend at the bottom.



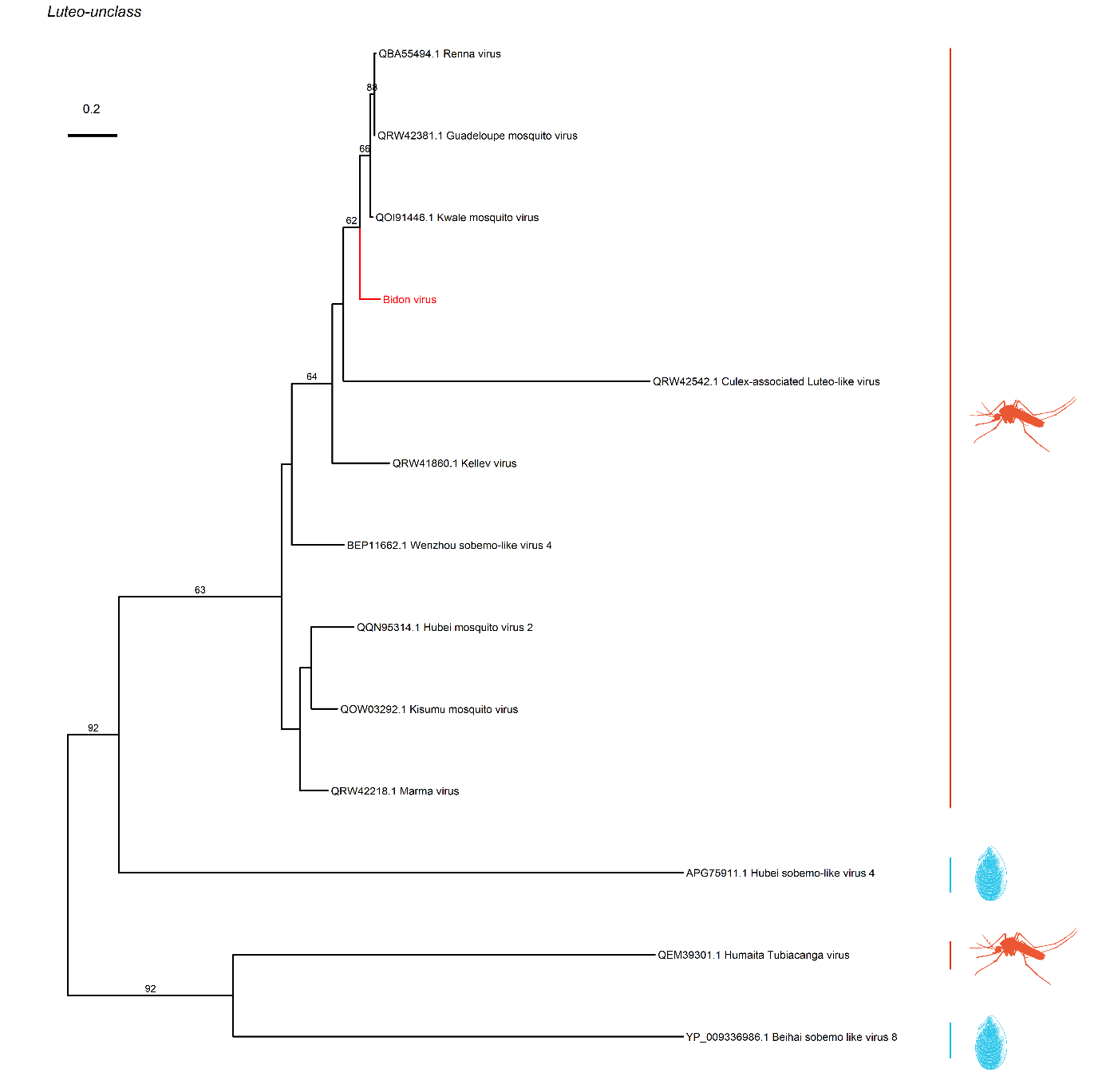
**Figure S8:** Alpha diversity indexes of viromes from *Culex quinquefasciatus* (A, B) and *Aedes aegypti* (C, D). From left to right boxplots show values for Species (VTU) richness, Shannon index and Inverse Simpson index for each sex in *C. quinquefasciatus* (A, B) and *A. aegypti* (C, D). Boxplot color indicates habitat anthropization (green: rural, grey: urban) in panels A and C. Boxplot color indicates year (purple: 2020, light green: 2021) in panels B and D. Dots indicate the index values for each library. The degree of significance between the two groups is provided on the bar above the boxplots (Table S4). Significance codes: 0 '\*\*\*', 0.001 '\*\*', 0.01 '\*', 0.05 '.', 0.1 'NS'.



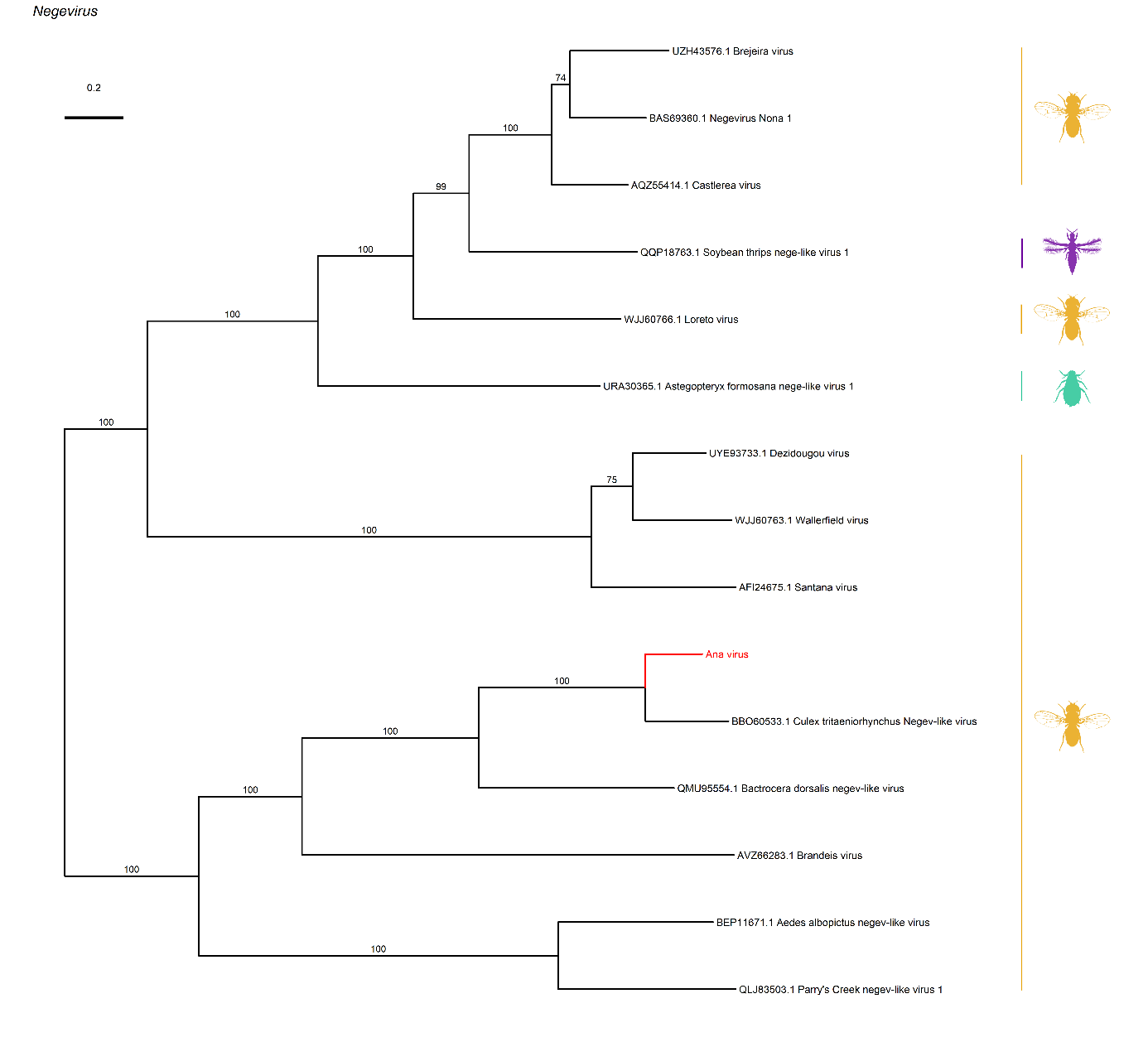
**Figure S9:** Maximum likelihood phylogenetic tree of the *Hytrosaviridae* (DNA viruses). Each phylogeny provided in Figures S9 to S16 was based on an analysis of the viral DNA polymerase (for DNA viruses), the viral RNA-dependent RNA polymerase (RdRP), or the capsid protein (for RNA viruses). The phylogenetic reconstruction for one potential new virus, designated VTU43\_Partiti, was not feasible due to the absence of a hallmark gene in the available sequences. Consequently, the creation of a robust phylogeny was unattainable for this virus. The newly identified viruses are distinguished from previously published viruses by red staining of the branch and associated names. The main clades of each family are represented opposite by black bars and names. The coloured animal silhouettes represent the host taxonomy of the viruses included in the analysis (here: yellow: *Diptera*, purple: *Hymenoptera*). Several host silhouettes are provided for viral taxa identified in hosts belonging to different groups (*e.g*., arboviruses). The values on the branches represent the bootstrap values of the internal nodes.



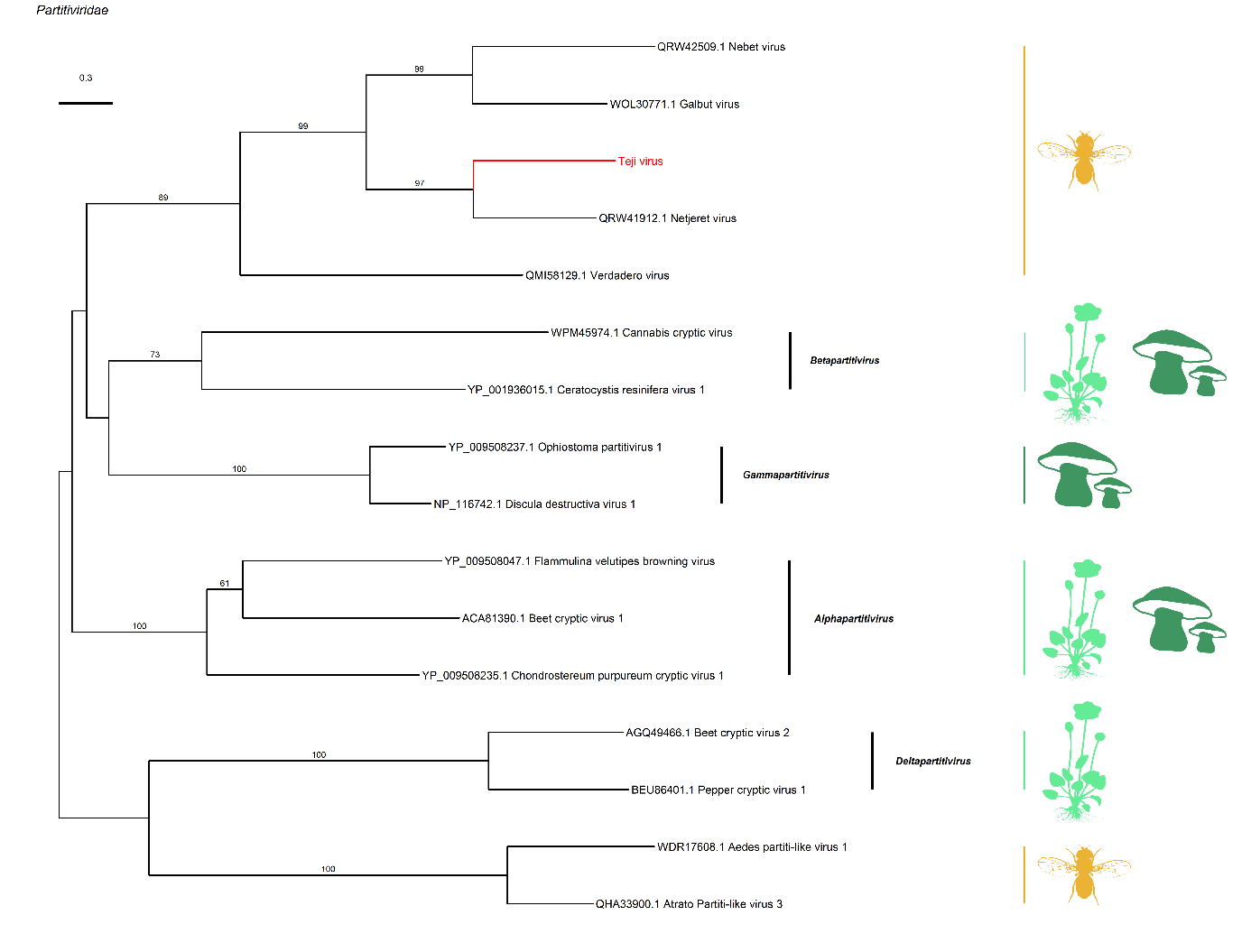
**Figure S10:** Maximum likelihood phylogenetic tree of the *Nudiviridae* (DNA viruses). Host taxonomy: yellow: *Diptera*, brown: *Orthoptera*, red: *Coleoptera*, pink: *Decapoda*, purple: *Lepidoptera*. Figure legend follows Figure S9.



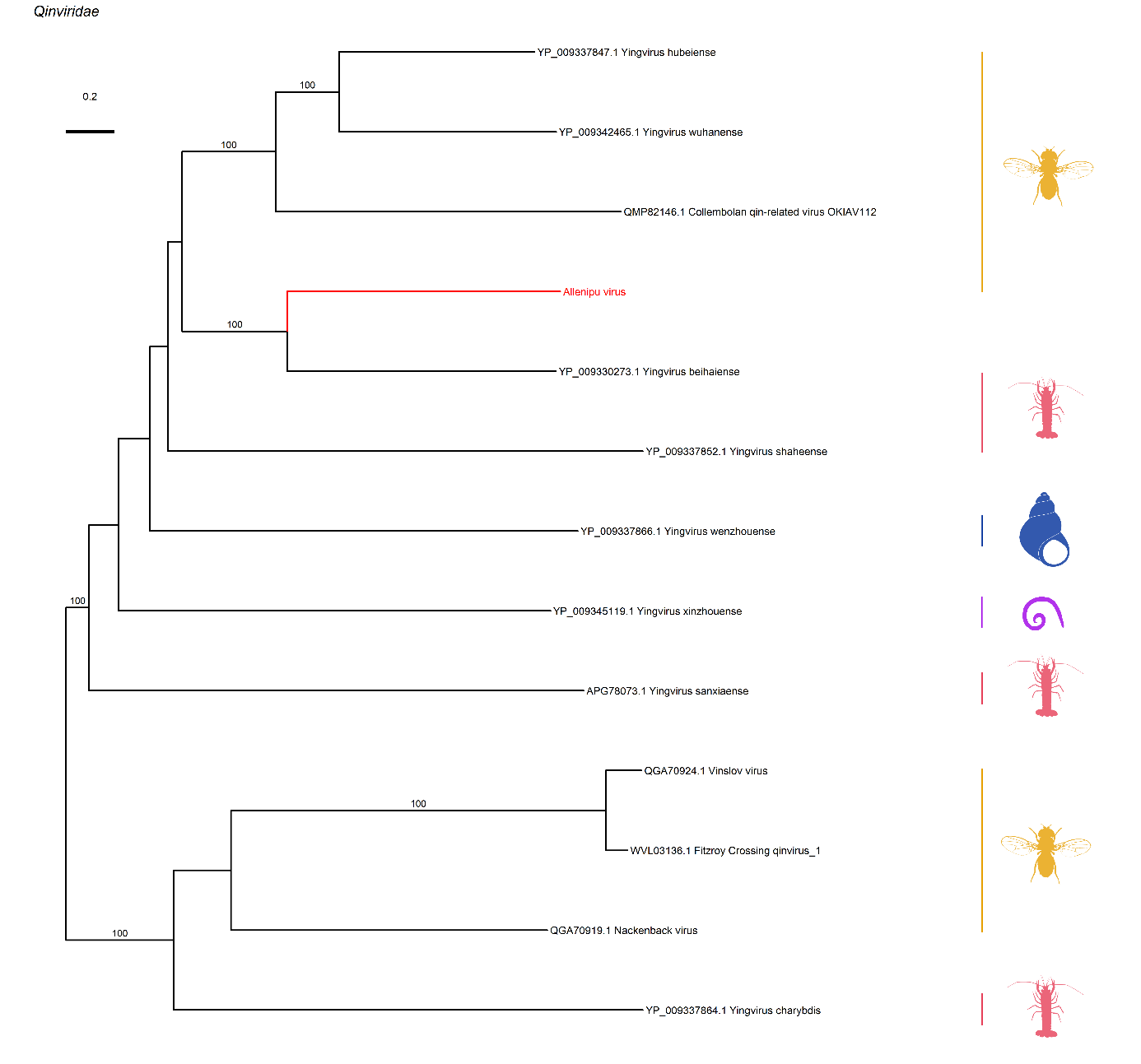
**Figure S11:** Maximum likelihood phylogenetic tree of the unclassified group of *Luteo-Sobemo* (Shi et al. 2016) (positive-strand RNA viruses). Host taxonomy: red: *Culicidae*, blue: *Bivalva*. Figure legend follows Figure S9.



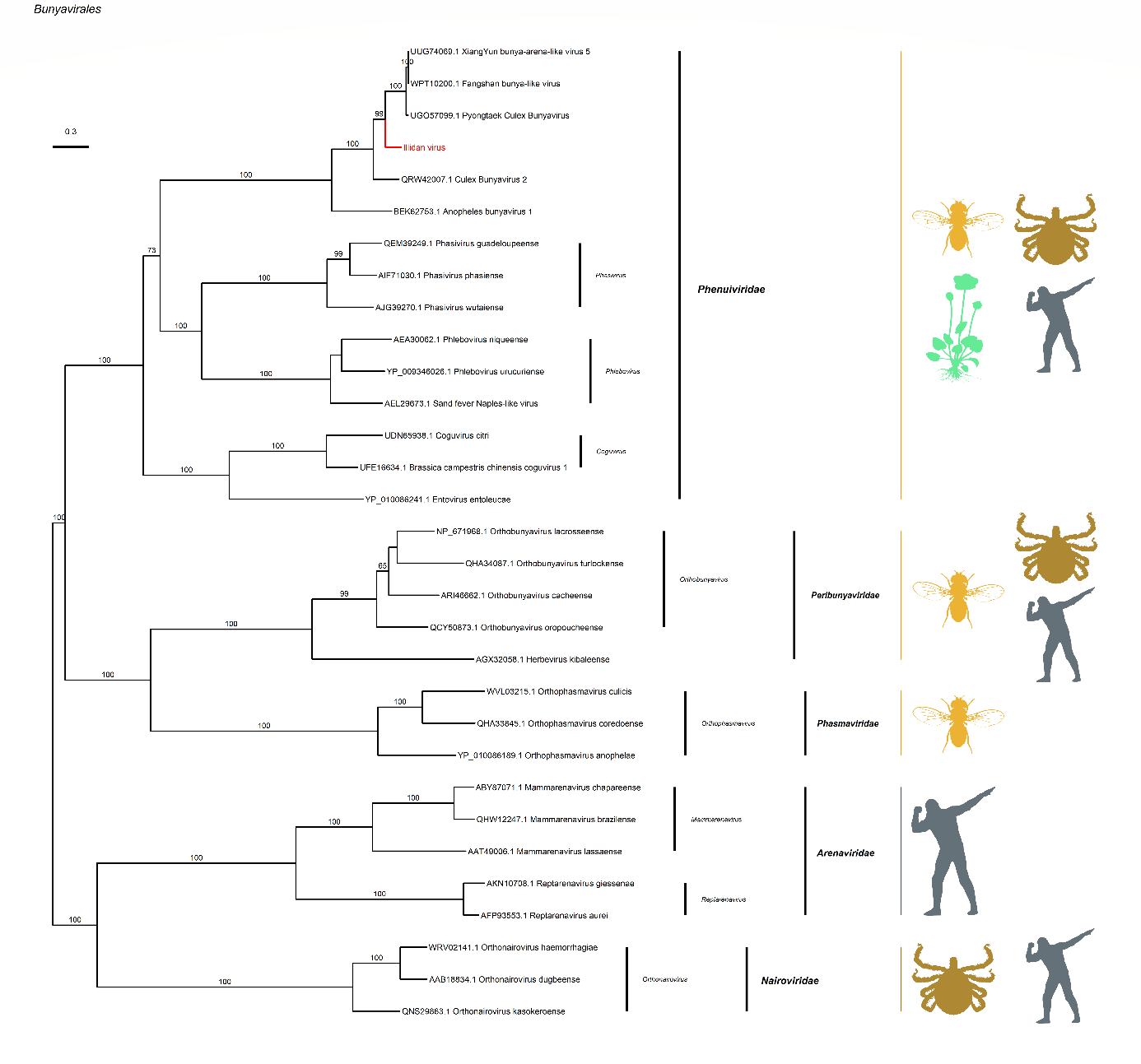
**Figure S12:** Maximum likelihood phylogenetic tree of the *Negeviruses* (positive-strand RNA viruses). Host taxonomy: yellow: *Diptera*, purple: *Thysanoptera*, green: *Hemiptera*. Figure legend follows Figure S9.



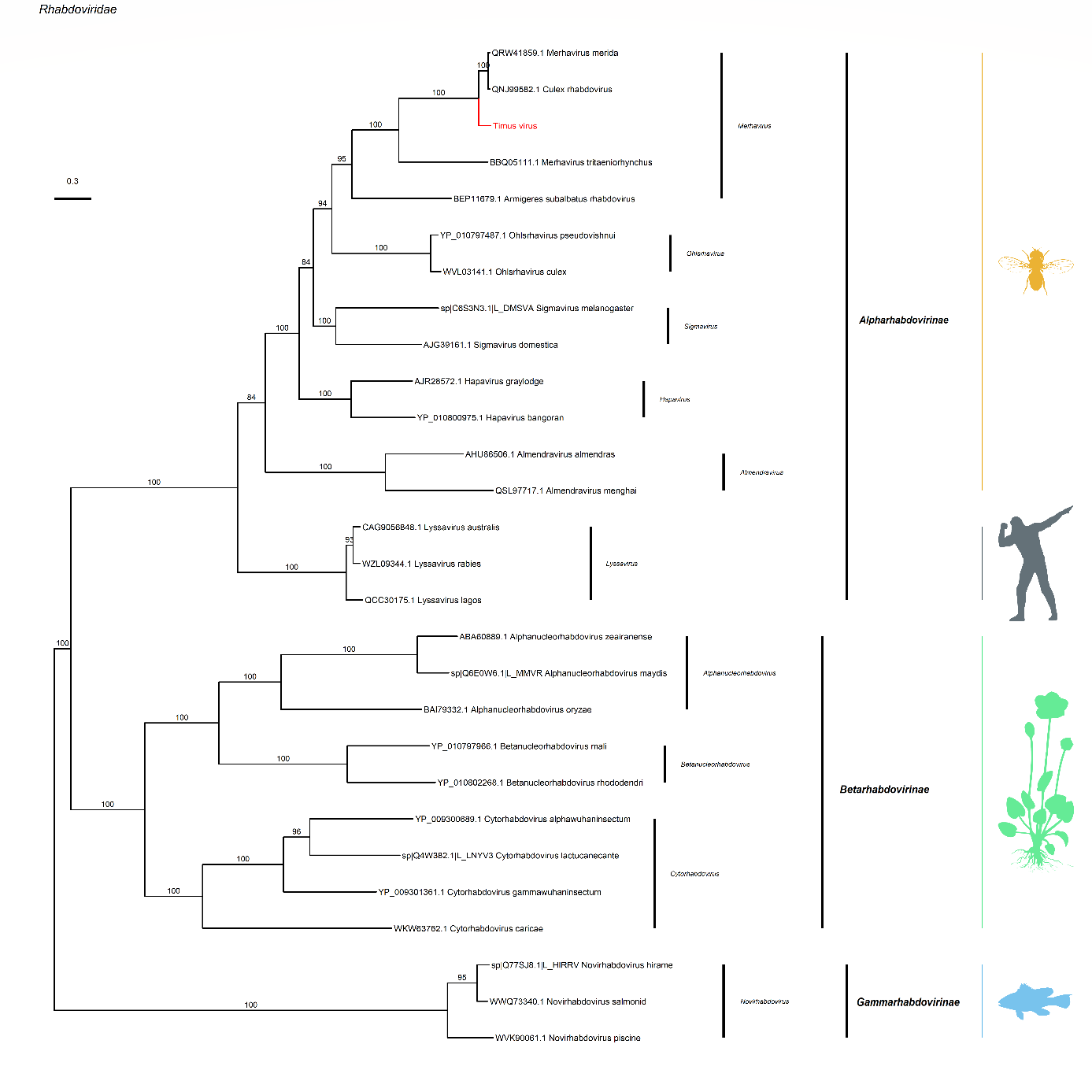
**Figure S13:** Maximum likelihood phylogenetic tree of the *Partitiviridae* (double-stranded RNA viruses). Host taxonomy: yellow: *Diptera*, light green: *Viridiplantae*, dark green: *Fungi*. Figure legend follows Figure S9.



**Figure S14:** Maximum likelihood phylogenetic tree of the *Qinviridae* (negative-sense RNA viruses). Host taxonomy: yellow: *Diptera*, red: *Malacostraca*, blue: *Gastropoda*, purple: *Nematoda*. Figure legend follows Figure S9.



**Figure S15:** Maximum likelihood phylogenetic tree of the *Bunyavirales* order (negative-strand RNA viruses). Host taxonomy: yellow: *Diptera*, brown: *Ixodidae*, grey: *Vertebrata*, green: *Viridiplantae*. Figure legend follows Figure S9.



**Figure S16:** Maximum likelihood phylogenetic tree of the *Rhabdoviridae* (negative-strand RNA viruses). Host taxonomy: yellow: *Diptera*, grey: *Vertebrata*, green: *Viridiplantae*, blue: *Actinopterygii*. Figure legend follows Figure S9.

## **Supplemental tables legends**

**Table S1:** Sampling sites of the study. The name of the site is provided in the "Site" column. The remaining columns describe the site in the following order (from left to right): anthopization level, sampling zone, the number of inhabitants, the latitude, the longitude, and a brief description of the habitat.

**Table S2:** Output of the sequencing and the bioinformatics analysis. The names of the libraries are provided in the "pool" column. The library names were generated to include the main variables of the study: mosquito species (Culex: Cx, Aedes: Ae), sampling zone, year, sex (female: F, male: M) and replicate (number of the replicate library from the same species, zone, year and sex). The "number" column provides the number of mosquitoes pooled in the library. The remaining columns represent the output of the bioinformatics pipeline, with the number of total reads (“Total\_read”), the number of reads with a viral hit (“Reads\_with\_viral\_hit”), the porportion of viral reads (%\_viral\_reads), the number of reads that mapped on insecta (“Map\_on\_Insecta”) or insect (“Map\_on\_Bacteria”), the number of contigs (“Nb\_of\_contigs”), the information on the size of the contigs (“Min\_contigs\_length”, “Max\_contigs\_length”, “Avg\_contigs\_length”) and the number of contigs with a viral hit (“Contigs\_with\_viral\_hit”).

**Table S3:** Read count by VTU and taxonomy. The "VTU" column provides the name given to the VTU. The "Best-hit" column provides the accession with the lowest e-value found by Diamond. The "Species name / New name" column provides information on the name of the virus associated with this VTU, with species names corresponding to the best-hit for VTUs belonging to known viruses (percentage of identity > 90% at amino acid level). The "Nucleic\_acid" and "genome" columns provide information on the type of VTU genome. The "cluster" column was created to add a taxonomic rank similar to a family, in order to avoid the lack of taxonomic information for certain viruses. The "clade" to "genus" columns provide the VTU taxonomy as stated by the ICTV. All columns from "qlen" to "bitscore" provide information on the contigs of each VTU and are Blastn output. All other columns represent count data by libraries for each VTU. The "sum\_reads" column provides information on the total number of reads detected for the VTU (across all libraries).

**Table S4:** Output statistics of Generalized Linear Mixed models. The upper table displays the model output using the “glmmTMB” function implemented in R. The lower table presents the results of pairwise comparisons, obtained using the emmeans function. The following significance codes are used to indicate the level of statistical significance: 0, '\*\*\*', 0.001, '\*\*', 0.01, '\*', 0.05, '.' and 1. For each table, the models were run on either a dataset containing all libraries or subsets of libraries of each mosquito species (Culex quinquefasciatus or Aedes aegypti).

**Table S5:** Output statistics of the PERMANOVA models. The table on the top shows the results for the complete dataset. The second and third tables show the results for a subset of the data only including libraries of Culex quinquefasciatus or Aedes aegypti, respectively. The following significance codes are used to indicate the level of statistical significance: 0, '\*\*\*', 0.001, '\*\*', 0.01, '\*', 0.05, '.' and 1.

**Table S6:** Output statistics of the IndVal function and data used in defining new viral species. Each indicator VTU of one of the two sexes in Culex (top table) and Aedes (bottom table) is listed here (column “VTU”) with the statistic (column “stat”) and the associated p-value (column “p-value”). The following significance codes are used to indicate the level of statistical significance: 0, '\*\*\*', 0.001, '\*\*', 0.01, '\*', 0.05, '.' and 1. The species name (or new species name for new virus species - pident < 90% at amino acid level) is also given (column “Species name / New name”), together with the statistics for the length of the best contig. The column designated as "qlen" represents the length of the query sequence, the column labeled "slen" represents the length of the subject sequence, while the column identified as "length" represents the length of the alignment (sequence overlap).

**Table S7:** Output statistics from the R “binGroup” package. All VTUs are listed in the table in the column designated "VTU," along with the calculation of the infection rate ("p") by mosquito species ("species”) and the sex of the host mosquito ("sex"). The values of the lower confidence limit and the upper confidence limit are displayed in the columns designated "lcl" and "ucl," respectively.