



Dec 09, 2025

Genome-representative sampling as a framework for diversity studies: a case study on geminiviruses

DOI

<https://dx.doi.org/10.17504/protocols.io.j8nlkybjdg5r/v1>

Moshood Olamide Lateef^{1,2}, Dauda Nathaniel³, Bolaji Osundahunsi⁴, Neil Arvin Bretana¹

¹IU International University of Applied Sciences, Germany; ²International Institute of Tropical Agriculture, Nigeria;

³Department of Crop Sciences, University of Nigeria Nsukka;

⁴Department of Entomology and Plant Pathology, University of Arkansas System Division of Agriculture, United States of America



Moshood Olamide Lateef

International Institute of Tropical Agriculture

Create & collaborate more with a free account

Edit and publish protocols, collaborate in communities, share insights through comments, and track progress with run records.

Create free account

OPEN  ACCESS



DOI: <https://dx.doi.org/10.17504/protocols.io.j8nlkybjdg5r/v1>

External link: https://github.com/Latmos-G/Geminivirus_diversity

Protocol Citation: Moshood Olamide Lateef, Dauda Nathaniel, Bolaji Osundahunsi, Neil Arvin Bretana 2025. Genome-representative sampling as a framework for diversity studies: a case study on geminiviruses. **protocols.io**

<https://dx.doi.org/10.17504/protocols.io.j8nlkybjdg5r/v1>

License: This is an open access protocol distributed under the terms of the **Creative Commons Attribution License**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Protocol status: Working

We use this protocol and it's working

Created: July 03, 2025

Last Modified: December 09, 2025

Protocol Integer ID: 221649

Keywords: Geminiviruses, Diversity, Geographical distribution, Plant viruses, Evolution, geminivirus, representative range of genome, diverse group of plant, available genomic sequence, available genomic sequences alongside the need, genome metadata, using genome metadata, diversity study, studying diversity, genome, systematic method for representative sampling, diverse group, challenges for representative sampling, infecting virus, representative sampling, based evolutionary relationship, phylogeny, specific identity thresholds for each genus, geographical distribution map, evolutionary relationship

Abstract

The rapid increase in publicly available genomic sequences alongside the need for constant updates poses challenges for representative sampling in diversity studies. This protocol describes a systematic method for representative sampling of geminiviruses, a highly diverse group of plant-infecting viruses. By using specific identity thresholds for each genus, this protocol makes sure to select a wide and representative range of genomes that are appropriate for studying diversity and phylogeny-based evolutionary relationships. Additionally, it includes procedures for constructing a geographical distribution map using genome metadata.

Materials

NCBI (GenBank): A comprehensive public database for genomic sequences. <https://www.ncbi.nlm.nih.gov/>

Sequence demarcation tool (SDT) version 1.3: A tool that classifies sequences based on their percentage pairwise identity. <http://web.cbio.uct.ac.za/~brejnev/>

NCBI datasets tool: A tool designed to access and retrieve genomic data. <https://www.ncbi.nlm.nih.gov/datasets/docs/v2/>

Python: An open-source programming language that offers extensive libraries, such as Pandas for data manipulation and Matplotlib for data visualization. <https://www.python.org/>

Plotly: An open-source library for the creation of interactive charts and dashboards. <https://plotly.com/python/>

Jupyter Notebook: A tool that enables writing and execution of code. <https://jupyter.org/>

Laptops and Internet access: Enable smooth means of accessing the database and performing the entire computational analysis.

Part 1: Defining the Sampling Criteria

- 1 Species demarcation in the *Geminiviridae* family is based on genome-wide pairwise identity thresholds ranging between 78% and 91% (Brown et al., 2015). This protocol uses genus-specific thresholds above the species demarcation cutoff for all genera except *Begomovirus* to ensure adequate representative sampling. For *Begomovirus*, due to its vast diversity, a previously used 80% threshold (Bandoo et al., 2024) was adopted. The adopted grouping threshold for genome selection from each genus was based on the level of the species divergence and the number of available datasets belonging to the genus. Specifically, genome sequences were grouped using a pairwise identity of $\geq 95\%$ for *Topocuvirus*, *Topilevirus*, *Maldovirus*, and *Eragrovirus*; $\geq 92\%$ for *Capulavirus*, *Citlodavirus*, *Curtovirus*, *Becurtovirus*, *Grablovirus*, *Mulcrilevirus*, *Opunvirus*, *Turncurtovirus*, and *Welwivirus*; and $\geq 80\%$ for *Mastrevirus* and *Begomovirus*.
- 1.1 All scripts and processed metadata are available at [Geminivirus-diversity](#)

Part 2: Representative Genome Sampling Protocol

- 2 Download Sequences:
 - 2.1 Retrieve complete nucleotide sequences from GenBank, genus by genus.
 - 2.2 For large genera (e.g., *Begomovirus*, *Mastrevirus*), retrieval by taxon through the section 'Results by taxon' on the GenBank graphic user interface (GUI) was adopted.
- 3 Pairwise Identity Calculation using SDT:
 - 3.1 Import the downloaded sequences into SDT v1.3.
 - 3.2 Run pairwise identity analysis using the default alignment algorithm.
 - 3.3 Once the pairwise identity matrix is generated, click Save → Create Datasets [based on p-identities].

4 Apply Identity Thresholds:

4.1 Input the genus-specific maximum and minimum identity thresholds (e.g., 100% and 92% for *Opunvirus*).

4.2 SDT will cluster the sequences into groups.

5 Representative Selection:

5.1 From each identity group in step 4.2, select a few sequence(s) and gather to represent the genus *Opunvirus*.

5.2 Repeat this process for each genus using the described identity thresholds.

6 Handling Large Genera (*Begomovirus* & *Mastrevirus*):

6.1 Partition the taxon (species within the genus) with a large number of sequences into smaller groups.

6.2 Process each group separately using SDT as described in step 3.

6.3 Use 100% and 80% identity thresholds to generate groups. This also enables the separation of DNA-A and DNA-B for bipartite genomes.

6.4 Collect samples from each group to make the taxon representative.

7 Final Compilation:

7.1 Merge all taxon representative sequences for the particular genus.

7.2 Process these merged sequences using SDT as described in step 3.

7.3 Use 100% and 80% identity thresholds to generate groups.

7.4 Collect samples from each group to make the genus representative.

Part 3: Geographical Distribution Mapping

8 Metadata Retrieval and Processing

8.1 Download metadata using NCBI datasets tool with command: `datasets summary virus genome taxon opunvirus --as-json-lines | dataformat tsv virus-genome > opunvirus.tsv`

8.2 Replace *Opunvirus* with other genus names to retrieve their respective metadata. However, only metadata for the genus *Welwivirus* was retrieved manually.

9 Geographical Data Processing:

9.1 Remove duplicate in the column 'Geographic Location' using [duplicate-removal script](#) and save output as CSV file.

9.2 Manually retain one entry per country in the output CSV file of the duplicate-removal script (step 9.1) for each genus.

10 Generate Unified Dataset:

10.1 Compile a single CSV file containing the countries and genera.

11 Geolocation Assignment:

11.1 Use online geolocation tools such as [LatLong.net](#) to assign coordinates of different provinces or states for the same country, especially when multiple genera are reported from the same country



- 11.2 Avoid automated geocoding of the countries to maintain unique coordinates (states/provinces) for a country with the multiple genera report.
- 12 Save and Format Data:
 - 12.1 Save the dataset as CSV UTF-8 (Comma delimited) for compatibility.
- 13 Map Generation:
 - 13.1 Use the **global geminivirus distribution script** to generate a geographical distribution map of geminivirus diversity

Protocol references

1. Brown JK, Zerbini FM, Navas-Castillo J, Moriones E, Ramos-Sobrinho R, Silva JC, Fiallo-Olivé E, Briddon RW, Hernández-Zepeda C, Idris A, Malathi VG. Revision of Begomovirus taxonomy based on pairwise sequence comparisons. *Archives of virology*. 2015;160:1593–1619. <https://doi.org/10.1007/s00705-015-2398-y>
2. Bando RA, Kraberger S, Varsani A. Two Novel Geminiviruses Identified in Bees (*Apis mellifera* and *Nomia* sp.). *Viruses*. 2024;16.4:602. <https://doi.org/10.3390/v16040602>