Automating the Integration of Viral Genomic Data: A Hierarchical Multi-Layered Approach for De-duplicating GenBank and GISAID Repositories

Andrêza Leite de Alencar[1], **Laise de Moraes**[2,3], **Ricardo Khouri**[2,3]

[1]Departamento de Ciência da Computação,
Universidade Federal Rural de Pernambuco (UFRPE)
Recife, PE, Brazil

[2]Laboratório de Medicina e Saúde Pública de Precisão,
Instituto Gonçalo Moniz,Fundação Oswaldo Cruz,
Salvador, BA, Brazil

[3]Instituto Nacional de Ciência e Tecnologia em Saúde Digital DigiSaude
Salvador, BA, Brazil

andreza.leite@ufrpe.br

Abstract. *The rapid expansion of viral genomic sequencing has led to data fragmentation across major repositories like GenBank and GISAID. The same isolate could be deposited in both databases without any cross-reference linking the two entries. Consequently, there is no systematic way to identify this redundancy, which compromises the compilation of representative, non redundant large scale datasets. For Neglected Tropical Diseases (NTDs), this redundancy complicates real time genomic surveillance. This paper presents a computational framework designed to de-duplicate and integrate these databases using a hierarchical strategy: sequence identity, fuzzy string matching of isolate identifiers, and metadata reconciliation. Our results demonstrate that the tool effectively identifies redundant records for CHIKV-ECSA lineage and filters out laboratory derived sequences, providing a high-quality dataset for phylogeographic analysis.*

1. Introduction

Genomic sequencing is a powerful tool for molecular epidemiology of viral infections, allowing us to identify new viruses [Wu et al. 2020, Zhou et al. 2020], characterize emerging lineages and recombinant forms [Greninger et al. 2010, Naveca et al. 2024], trace their origins and natural reservoirs [Pekar et al. 2025], thereby revealing how infectious diseases spread across spatial and temporal scales, from localized outbreaks [Weaver 2014] to broader scenarios [Worobey et al. 2016, Santa Ardisson et al. 2025].

However, researchers often face the ”double-counting” problem, where identical biological isolates are distributed across multiple database sources. This phenomenon typically arises from independent submission workflows: a laboratory may upload a sequence to GISAID [Khare S and S. 2021] for immediate public health sharing while simultaneously submitting it to GenBank [Benson et al. 2014] as part of a formal publication requirement. The most frequent overlap occurs between GenBank and GISAID.

Both are essential for comprehensive sampling, as each captures unique subsets of viral diversity and epidemiological context. Nevertheless, each database employs distinct data-sharing policies and metadata reporting practices.

Discrepancies in isolate nomenclature, varied levels of metadata granularity, and differing sequence trimming protocols make it difficult to recognize when two records correspond to the same viral isolate, leading to a fragmented and redundant landscape.

This redundancy biases molecular clock analyses and epidemiological modeling, since duplicate sequences distort estimates of evolutionary rates and population size dynamics, directly compromising the accuracy of molecular clock calibrations and phylogeographic inference [Garamszegi and Møller 2010, Vakulenko et al. 2019, He et al. 2025]. In the context of Neglected Tropical Diseases, where resources are limited, automated tools to clean and integrate these data are essential. Addressing this, we present the G2G matcher, a toll for the systematic de-redundancy and harmonization of viral genomic metadata across GenBank and GISAID. Using the CHIKV-ECSA lineage as a proof-of-concept, we demonstrate how the application helps to provide a curated dataset for downstream phylogeographic inference.

2. Methodology

The G2G Matcher framework was developed as a modular Python-based pipeline designed to streamline the integration of heterogeneous viral genomic data. The architecture is divided into five specialized stages, prioritizing computational efficiency and biological accuracy. The Figure 1 illustrates the G2G Matcher framework pipeline.

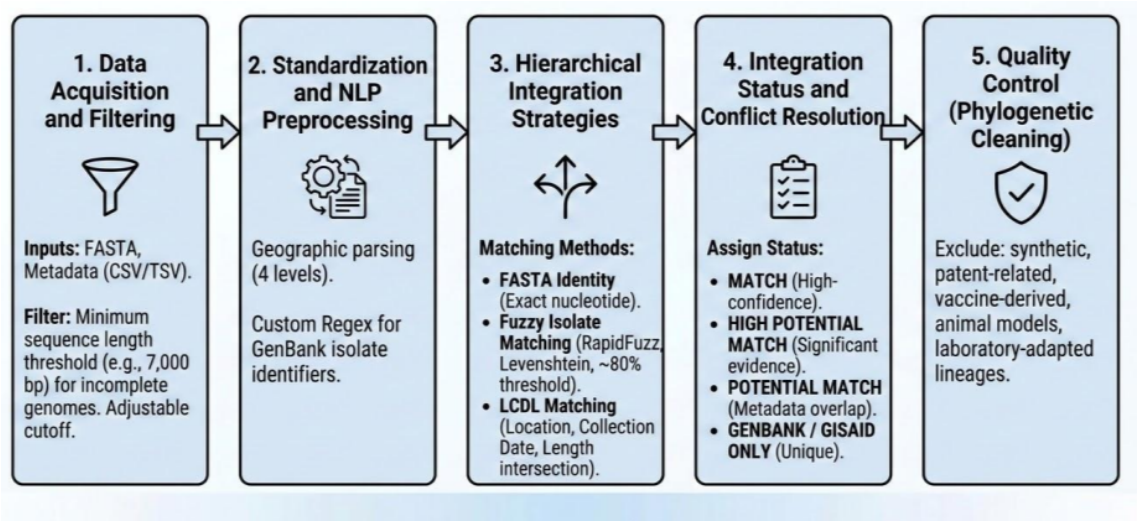

Figure 1. Architectural Workflow of the G2G Matcher Framework Pipeline. The pipeline illustrates the five stage hierarchical process for genomic data integration

2.1. Data Acquisition and Filtering

The pipeline initiates with the ingestion of raw FASTA sequences and their corresponding metadata in CSV or TSV formats from GenBank and GISAID. To ensure that subsequent phylogenomic inferences are based on high-quality genetic information, a rigorous filtering step is applied. A minimum sequence length threshold (defaulting to 7,000 bp - for

CHIKV) is utilized to exclude fragmented or partial genomes that might introduce noise into molecular clock estimates. This threshold is a configurable parameter, allowing the tool to be adapted to different viral species or study designs.

2.2. Standardization and NLP Preprocessing

A critical challenge in database integration is the lack of standardized nomenclature. Our framework addresses this through two main processes:

- **Geographic Parsing**: Location metadata is decomposed into a four-level hierarchy (Continent, Country, State/Province, and City) to resolve discrepancies in spatial granularity. Recognizing that metadata completeness varies across repositories, the system is designed to handle missing fields while parsing these levels only where available. The unavailable levels can be automatically inferred based on the available fields using a standardized geographical database. This ensures a consistent categorization even when original records provide limited spatial information.
- **NLP-based Extraction**: While GISAID provides structured isolate names, GenBank titles often embed this information within non-standardized strings. We implemented a series of custom Regular Expressions (Regex) to programmatically isolate and extract these identifiers, creating a unified key for cross-repository comparison.

2.3. Hierarchical Integration Strategies

The core of the framework is a multi-layered matching engine that employs three distinct strategies to identify redundancies:

- **FASTA Identity**: The tool performs an exact nucleotide-to-nucleotide comparison/matching. Sequences with 100% identity across both GenBank and GISAID repositories are flagged as high-confidence matches, regardless of metadata differences.
- **Fuzzy Isolate Matching**: To account for typographical errors or naming variations, (e.g., ”isolate/01” vs. ”isolate_01”), we utilize RapidFuzz library to calculate similarity scores between isolate names. By applying Levenshtein distance-based metrics and partial ratio algorithms, the degree of string overlap. A default similarity threshold of 80% was established to reconcile these variations, providing a robust mechanism to align records that would otherwise be treated as distinct due to minor orthographic or formatting differences across platforms.
- **LCDL Matching (Metadata-based Matching)**: As a final step an exact matching approach focusing on the intersection of Location, Collection Date, and sequence Length is executed to identify records that correspond to the same sample but exhibit minor sequence differences. This serves as a safety mechanism for records where sequence data might differ slightly due to due to varying laboratory trimming (cutting) quality-filtering protocols but metadata remains identical.

2.4. Integration Status and Conflict Resolution

Upon accumulation of evidence from the matching engine, the framework assigns an Integration Status to each record to facilitate automated curation:

- **MATCH**: High-confidence redundant entries supported by genomic (sequence data) and/or multiple metadata overlaps.
- **HIGH POTENTIAL MATCH**: Records with significant evidence (typically sequence identity) but slight metadata discrepancies requiring minor validation.
- **POTENTIAL MATCH**: Entries with metadata overlaps but lacking sequence identity or lower-score fuzzy matches.
- **GENBANK ONLY / GISAID ONLY**: Unique biological records present in only one of the source repositories.

2.5. Quality Control (Phylogenetic Cleaning)

To ensure biological relevance of the final dataset and refine the dataset for downstream biological analysis, the tool performs an automated cleaning. The system parses metadata fields for specific keywords to exclude sequences that do not represent natural viral circulation. This includes the removal of:

- Experimental Artifacts: Synthetic constructs, clones, mutants, and patent-related sequences.
- Laboratory Passages: Vaccine-derived strains and laboratory-adapted lineages (adapted in cell lines, e.g., Vero, C6/36).
- Non-human Hosts: Murine or other animal models, unless specifically requested by the user.

3. Implementation and Results for Chikungunya virus

The tool is implemented in Python [van and FL 2010] available at the Github repository [1]), utilizing the Pandas library [W 2010] for efficient vectorization of metadata comparisons, Biopython [Cock PJA and B 2009] for sequence parsing, and tqdm [Casper da Costa-Luis and Rodrigues 2026] for progress monitoring in large-scale integration tasks. The hierarchical matching protocol successfully reconciled the genomic data, identifying 3,030 redundant records shared between GenBank and GISAID. This deduplication process prevented a 26.6% inflation of the global dataset (based on the original overlapping records), ensuring that subsequent phylogenomic analyses are grounded in unique biological samples rather than repository-specific duplicates.

3.1. Genomic Data Reconciliation

The distribution of unique and redundant records is illustrated in 2. Out of the total initial pool, the Matcher identified 3,030 sequences as redundant across both databases, which were subsequently integrated into single non-redundant entries. The final curated dataset comprised 8,347 unique viral genomes, consisting of 2,211 GenBank-exclusive records, 3,106 GISAID-exclusive records, and the 3,030 reconciled consensus entries.

3.1.1. Performance of Matching Strategies

The protocol demonstrated exceptional robustness by integrating genomic and documentary evidence, as detailed in Table (1). A significant portion of the reconciled dataset (38.38%) was supported by *Genomic & Metadata Consensus*, where absolute nucleotide

[1] https://github.com/CIDA-UFRPE/Integration-of-Viral-Genomic-Data.git

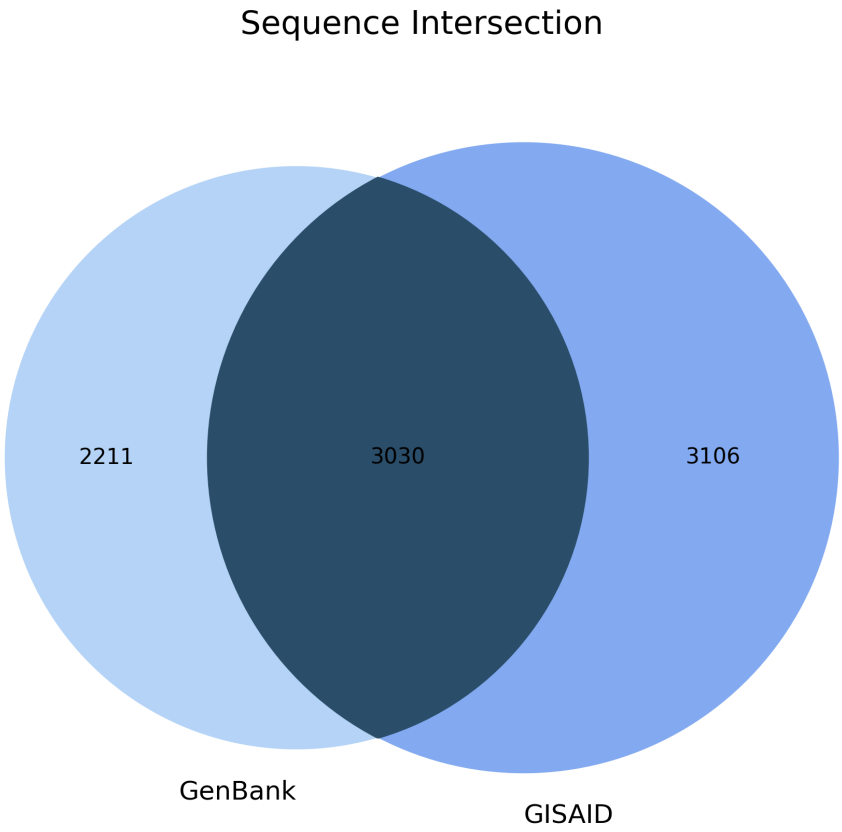

Figure 2. Genomic Data Reconciliation and De-duplication. Venn diagram illustrating the intersection of CHIKV genomic records between GenBank and GISAID. Intersection (Matches): 3,030 sequences were identified as redundant and integrated. GenBank Exclusive: 2,211 sequences. GISAID Exclusive: 3,106 sequences.

identity (FASTA) converged with isolate nomenclature and temporal-spatial metadata (LCDL). This dual-validation layer provides a high-confidence backbone for as it confirms that both the genetic material and the epidemiological documentation are in perfect agreement across repositories.

Interestingly, 61.39% of the matches relied solely on *Genomic Identity (FASTA)*. This highlights a common challenge in genomic surveillance: identical viral sequences are frequently submitted to different repositories with non-standardized or missing metadata, making the nucleotide sequence the only universal link for reconciliation.

Conversely, only a minor fraction (0.23%, $n = 7$) required integration based exclusively on *Metadata Consensus*. These cases are particularly noteworthy for data quality; they typically represent records where the sequence data differs slightly due to divergent laboratory trimming or quality-filtering protocols (e.g., one version having a few more nucleotides at the 5' end than the other), yet the metadata identifies them as the same biological isolate. By reconciling these via metadata, the framework prevents the artificial inclusion of identical duplicates that could bias molecular clock calibrations.

Table 1. Performance of Hierarchical Matching Strategies

Matching Strategy	Unique Records	Contribution (%)	Description
Genomic & Metadata Consensus	1,163	38.38%	Supported by FASTA + Isolate Names and/or Temporal-Spatial data.
Genomic Identity Only (FASTA)	1,860	61.39%	Resolved solely via nucleotide identity.
Metadata Consensus Only	7	0.23%	Resolved via Isolate Names without strict FASTA identity.
Total Unique Overlaps	**3,030**	**100.00%**	

3.1.2. Geographic Distribution and Impact on Genomic Surveillance

The geographic impact of the de-duplication is illustrated in Figure 3. The comparison reveals that the Matcher effectively mitigates artificial sampling bias—most visible in high-surveillance regions like Brazil—without compromising spatial representation. All countries present in the initial aggregate (Map A) remained in the consolidated dataset (Map B), confirming that the tool specifically targets technical redundancy while preserving the full breadth of the virus's global circulation.

By automating this convergence analysis, the Matcher resolved over 99.7% of redundancies through high-confidence validation. This allows genomic surveillance programs to maintain rigorous data quality standards, focusing human expert assessment only on the negligible fraction of truly ambiguous cases (e.g., identical sequences from different locations). Ultimately, this ensures that the global CHIKV landscape is represented by a balanced, non-redundant dataset essential for accurate phylogeographic reconstruction.

4. Discussion

The integration of viral genomic data from multiple repositories is a critical challenge for real-time surveillance. The Matcher addresses the redundancy problem through a hierarchical strategy that prioritizes genomic identity while leveraging the richness of available metadata. Our results demonstrate that relying exclusively on metadata for de-duplication would be insufficient, as 61.39% of redundancies were resolved solely via sequence identity (FASTA). This occurs because the same biological isolate is frequently deposited with distinct names or identifiers across GenBank and GISAID. Conversely, the consensus layer between genome and metadata (38.38%) provides a high-confidence result for phylogeographic studies. The normalization of geographic hotspots, as observed in the case of Brazil, is perhaps the most practical contribution of this framework. By mitigating an artificial 26.6% inflation of records, the tool reduces sampling biases that could compromise molecular clock analyses and dispersal route inferences. Furthermore, the automated cleaning of laboratory artifacts ensures the final dataset reflects natural viral circulation, removing biological noise that does not contribute to public health monitoring. We emphasize that the consolidated dataset provides a reference point for manual curation, an essential step that this pipeline does not replace. Manual review remains critical for validating ambiguous matches, verifying metadata accuracy, and resolving conflicts that exceed algorithmic resolution, such as distinguishing between genuine de-redundancy and independently sampled isolates from the same spatiotemporal context.

5. Conclusion

This study introduced the G2G Matcher, a computational framework developed to automate the integration and cleaning of global viral genomic repositories. By identifying 3,030 redundant records, the tool generated a consolidated dataset of 8,347 unique

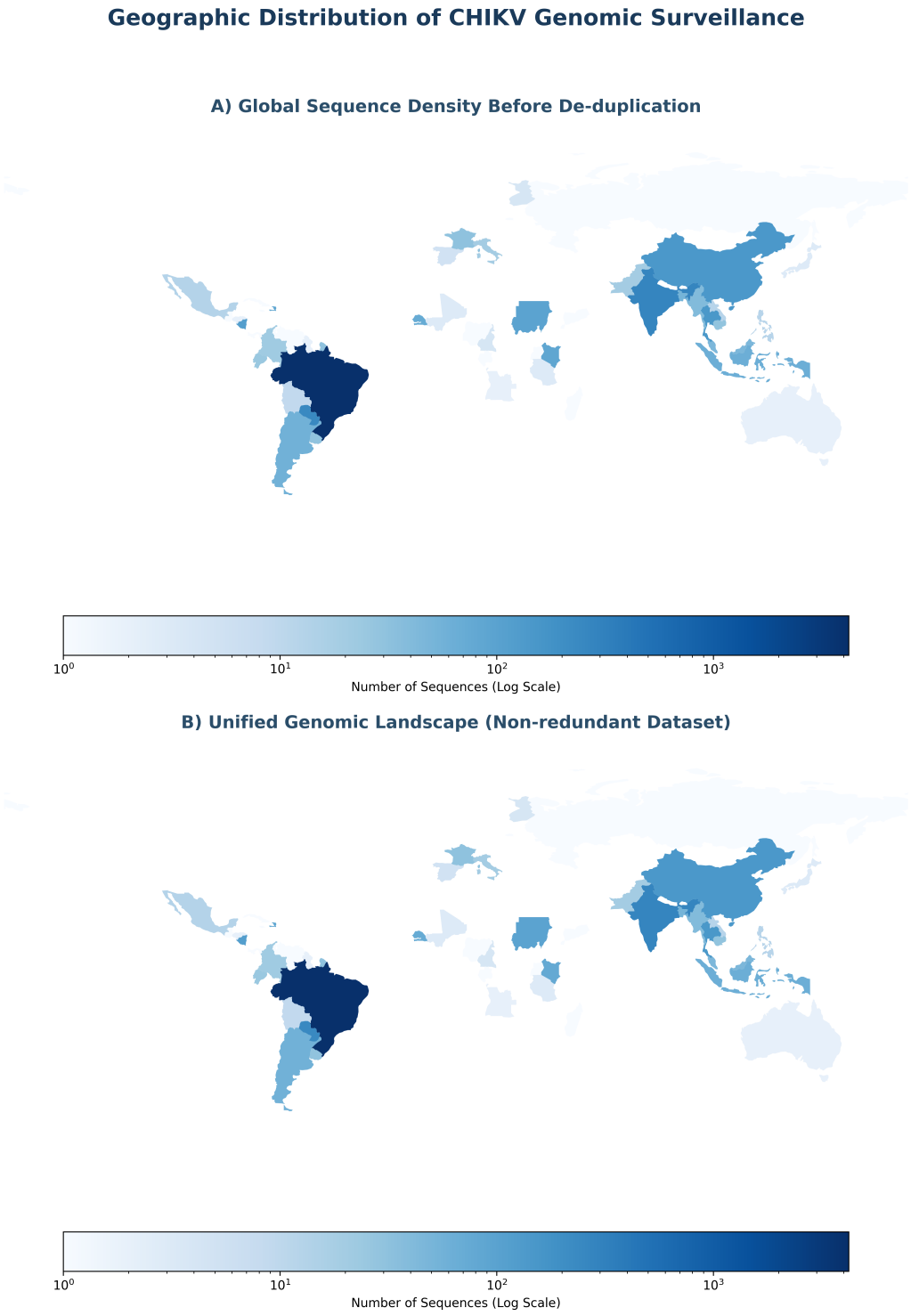

Figure 3. Global Geographic Distribution of CHIKV Sequences Before and After Integration. Heatmap representing the density of genomic records per country. (A) Raw aggregate of GenBank and GISAID records. (B) Consolidated dataset after Matcher de-duplication. The color gradient (Number of Sequences) indicates that while the total volume of data is normalized to remove cross-database duplicates, the spatial coverage remains intact, providing a balanced input for subsequent phylogeographic modeling.

genomes, effectively preventing a 26.6% artificial inflation of the global CHIKV landscape. This process is essential for accurate epidemiological modeling and mitigating sampling biases that often distort phylogeographic inferences.

The hierarchical approach proved highly effective in overcoming the challenges of data fragmentation and metadata heterogeneity between GenBank and GISAID. By applying genomic identity and NLP-based metadata extraction, the G2G Matcher ensures that genomic surveillance, especially for Neglected Tropical Diseases (NTDs), is conducted with increased transparency, biological rigor, and computational efficiency providing a robust solution for researchers and public health agencies.

While this work focused on the Chikungunya virus, the modular architecture of the G2G Matcher is intentionally designed for cross-viral scalability. Future developments will focus on:

- **Multiviral Framework Calibration:** Expanding the pipeline to other arboviruses of high public health impact, such as Dengue, Zika, and Mayaro. This involves fine-tuning biological filters, such as minimum sequence length thresholds, to accommodate the specific genomic architectures of these viruses.
- **Advanced Metadata Extraction via LLMs:** We aim to incorporate Large Language Models (LLMs) to automate the parsing of unstructured metadata. By moving beyond traditional Regex, the system will be able to interpret complex isolate descriptions, significantly reducing the 'potential match' ambiguity and increasing de-duplication accuracy.
- **Accessibility and Cloud-Native Deployment:** Transitioning the tool into a python library or a containerized (Docker) and cloud-ready solution. This ensures that public health laboratories in resource-limited settings can use and/or deploy the framework without advanced local computing infrastructure, fostering global equity in genomic surveillance.
- **API-Driven Real-Time Integration:** Implementing direct API connections to NCBI and GISAID to enable automated, continuous data synchronization. Such advancements will transform the G2G Matcher into a surveillance hub, strengthening the global capacity to monitor and respond to emerging viral threats.

AI Usage Declaration

Gemini was utilized for text editing, grammatical refinement, image generation and enhancing clarity. All suggestions generated by the tool were critically analyzed and entirely reviewed by the authors, who assume full responsibility for the final content presented.

References

Benson, D. A., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2014). GenBank. *Nucl. Acids Res.*, 42(D1):D32–D37.

Casper da Costa-Luis, Stephen Karl Larroque, K. A. H. M. r. M. K. N. R. I. I. M. B. and Rodrigues, N. (2026). tqdm: A fast, extensible progress bar for python and cli.

Cock PJA, Antao T, C. J. C. B. C. C. D. A. F. I. H. T. K. F. and B, W. (2009). Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25:1422–1423.

Garamszegi, L. Z. and Møller, A. P. (2010). Effects of sample size and intraspecific variation in phylogenetic comparative studies: a meta-analytic review. *Biological Reviews*, 85(4):797–805.

Greninger, A. L., Chen, E. C., Sittler, T., Scheinerman, A., Roubinian, N., Yu, G., Kim, E., Pillai, D. R., Guyard, C., Mazzulli, T., Isa, P., Arias, C. F., Hackett, J., Schochetman, G., Miller, S., Tang, P., and Chiu, C. Y. (2010). A Metagenomic Analysis of Pandemic Influenza A (2009 H1N1) Infection in Patients from North America. *PLoS ONE*, 5(10):e13381.

He, C., Chen, M.-Y., and Zhu, H. (2025). Population Size Differences Can Lead to Biases in Phylogenetic Inference and Introgression Detection in the Presence of Purifying Selection. *Systematic Biology*, page syaf085.

Khare S, Gurry C, F. L. S. M. B. G. D. A. A. N. H. J. L. R. Y. W. C. T. G. and S., M.-S. (2021). Gisaid's role in pandemic response. *China CDC Wkly*, 3(3(49)):1049–1051.

Naveca, F. G., Almeida, T. A. P. D., Souza, V., Nascimento, V., Silva, D., Nascimento, F., Mejía, M., Oliveira, Y. S. D., Rocha, L., Xavier, N., Lopes, J., Maito, R., Meneses, C., Amorim, T., Fé, L., Camelo, F. S., Silva, S. C. D. A., Melo, A. X. D., Fernandes, L. G., Oliveira, M. A. A. D., Arcanjo, A. R., Araújo, G., André Júnior, W., Carvalho, R. L. C. D., Rodrigues, R., Albuquerque, S., Mattos, C., Silva, C., Linhares, A., Rodrigues, T., Mariscal, F., Morais, M. A., Presibella, M. M., Marques, N. F. Q., Paiva, A., Ribeiro, K., Vieira, D., Queiroz, J. A. D. S., Passos-Silva, A. M., Abdalla, L., Santos, J. H., Figueiredo, R. M. P. D., Cruz, A. C. R., Casseb, L. N., Chiang, J. O., Frutuoso, L. V., Rossi, A., Freitas, L., Campos, T. D. L., Wallau, G. L., Moreira, E., Lins Neto, R. D., Alexander, L. W., Sun, Y., Filippis, A. M. B. D., Gräf, T., Arantes, I., Bento, A. I., Delatorre, E., and Bello, G. (2024). Human outbreaks of a novel reassortant Oropouche virus in the Brazilian Amazon region. *Nat Med*, 30(12):3509–3521.

Pekar, J. E., Lytras, S., Ghafari, M., Magee, A. F., Parker, E., Wang, Y., Ji, X., Havens, J. L., Katzourakis, A., Vasylyeva, T. I., Suchard, M. A., Hughes, A. C., Hughes, J., Rambaut, A., Robertson, D. L., Dellicour, S., Worobey, M., Wertheim, J. O., and Lemey, P. (2025). The recency and geographical origins of the bat viruses ancestral to SARS-CoV and SARS-CoV-2. *Cell*, 188(12):3167–3183.e18.

Santa Ardisson, J., Vedovatti Monfardini Sagrillo, M., Ramos Athaydes, B., Corredor Vargas, A. M., Torezani, R., Ribeiro-Rodrigues, R., Cruz Spano, L., Gaburro Paneto, G., Delatorre, E., Ventorin Von Zeidler, S., and Freire Bastos Filho, T. (2025). Comparative spatial–temporal analysis of SARS-CoV-2 lineages B.1.1.33 and BQ.1.1 Omicron variant across pandemic phases. *Sci Rep*, 15(1):10319.

Vakulenko, Y., Deviatkin, A., and Lukashev, A. (2019). The Effect of Sample Bias and Experimental Artefacts on the Statistical Phylogenetic Analysis of Picornaviruses. *Viruses*, 11(11):1032.

van, R. G. and FL, D. (2010). *The Python language reference*. Hampton, NH: Python Software Foundation, release 3.0.1 edition.

W, M. (2010). Data structures for statistical computing in python. page 56–61.

Weaver, S. C. (2014). Arrival of Chikungunya Virus in the New World: Prospects for Spread and Impact on Public Health. *PLoS Negl Trop Dis*, 8(6):e2921.

Worobey, M., Watts, T. D., McKay, R. A., Suchard, M. A., Granade, T., Teuwen, D. E., Koblin, B. A., Heneine, W., Lemey, P., and Jaffe, H. W. (2016). 1970s and 'Patient 0' HIV-1 genomes illuminate early HIV/AIDS history in North America. *Nature*, 539(7627):98–101.

Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., Hu, Y., Tao, Z.-W., Tian, J.-H., Pei, Y.-Y., Yuan, M.-L., Zhang, Y.-L., Dai, F.-H., Liu, Y., Wang, Q.-M., Zheng, J.-J., Xu, L., Holmes, E. C., and Zhang, Y.-Z. (2020). A new coronavirus associated with human respiratory disease in China. *Nature*, 579(7798):265–269.

Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., Si, H.-R., Zhu, Y., Li, B., Huang, C.-L., Chen, H.-D., Chen, J., Luo, Y., Guo, H., Jiang, R.-D., Liu, M.-Q., Chen, Y., Shen, X.-R., Wang, X., Zheng, X.-S., Zhao, K., Chen, Q.-J., Deng, F., Liu, L.-L., Yan, B., Zhan, F.-X., Wang, Y.-Y., Xiao, G.-F., and Shi, Z.-L. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, 579(7798):270–273.