



Research article

GeneBank Genie: An interactive toolkit for integrated multivariate analysis and visualization of GenBank records

Yash Munnalal Gupta*, Somjit Homchan

Department of Biology, Faculty of Science, Naresuan University, 99 Moo 9, Phitsanulok-Nakhonsawan Road, Phitsanulok 65000, Thailand.

Article Info

Article history:

Received 30 April 2025

Revised 18 December 2025

Accepted 14 February 2026

Available 27 February 2026

Keywords:

Bioinformatics,

GenBank,

Outlier detection,

Principal component analysis (PCA),

Software,

Taxonomy visualization

Abstract

Importance of the work: GeneBank Genie fills a critical gap by providing integrated desktop software for accessible and comprehensive analysis of GenBank data.

Objectives: To develop intuitive software facilitating simultaneous automated parsing, preliminary genomic analysis, visualization and sequence extraction from multiple GenBank records.

Materials and Methods: GeneBank Genie was built in the Python programming environment, using Tkinter for the graphical user interface (GUI) design, Biopython and custom Python scripts for sequence processing and scikit-learn for principal component analysis (PCA) and clustering analytics.

Results: The developed software called GeneBank Genie was used successfully to analyze a dataset of 333 Orthoptera mitochondrial genomes, automatically computing nucleotide compositions, the percentage of nitrogenous bases that are either guanine or cytosine (GC content) and gene annotations. The PCA revealed distinct genomic clustering patterns and Mahalanobis distances were used to identify outliers effectively. Taxonomic visualizations demonstrated robust exploratory capabilities based on interactive Sankey diagrams, dendrograms, correlation heatmaps and K-means clustering. Additionally, rapid extraction of gene sequences illustrated practical applications for molecular research workflows.

Main finding: GeneBank Genie uniquely integrates automated batch processing of GenBank records, PCA-based analytics and intuitive visualizations, greatly simplifying genomic data exploration for biologists. GeneBank Genie is freely available at: <https://github.com/yashmgupta/GenBank-Genie>

* Corresponding author.

E-mail address: yashmunnalalg@nu.ac.th (Y.M. Gupta)

online 2452-316X print 2468-1458/Copyright © 2025. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), production and hosting by Kasetsart University Research and Development Institute on behalf of Kasetsart University.

<https://doi.org/10.34044/j.anres.2026.60.1.12>

Introduction

Recent advancements in DNA sequencing technologies have rapidly accelerated the accumulation of genomic data, which are routinely deposited into publicly accessible repositories such as GenBank (Benson et al., 2018; Sayers et al., 2020). GenBank, managed by the National Center for Biotechnology Information (NCBI), stores extensive genetic information, including detailed annotations of nucleotide sequences, gene features, coding regions, regulatory elements and comprehensive taxonomic metadata (Sayers et al., 2024). These datasets represent critical resources for diverse biological research, encompassing phylogenetic studies, gene expression analyses, functional characterization, primer design and evolutionary biology investigations (Baxevanis and Bateman, 2015).

Despite their wide-ranging utility, GenBank records present analytical challenges due to their inherently multifaceted, structured data format with features and metadata. Typically, parsing GenBank files to extract meaningful biological insights requires meticulous handling of intricate metadata and nested annotations—a process often demanding specialized computational expertise and scripting (Miller et al., 2009). Existing bioinformatics tools commonly available either focus narrowly on specific tasks, such as sequence alignment or primer design, or lack the capability to manage systematically large sets of GenBank records simultaneously with analyzing metadata, limiting their efficiency and accessibility for many biologists (Fernandez-Suarez and Galperin, 2012). The suboptimal implementation of metadata parsing functionality represents a critical constraint in current bioinformatics platforms.

Furthermore, most existing software platforms require considerable computational training or scripting capabilities, posing barriers to researchers, particularly molecular biologists, who seek to leverage extensive genomic data but may not possess advanced computational skills. The broader application of genomic databases in biological research is greatly restricted by the lack of integrated tools that seamlessly combine multi-file GenBank processing, advanced analytical capabilities such as preliminary principal component analysis (PCA), statistical outlier detection and intuitive visualization techniques (De Maesschalck et al., 2000; Ringnér, 2008). Such introductory analytical processes are critical for informing downstream investigative pathways and simplifying evidence-based decision making regarding subsequent specialized analyses.

GeneBank Genie was developed to overcome these limitations, as a comprehensive desktop software package developed specifically to simplify and streamline the analysis and visualization of multiple GenBank records. GeneBank Genie integrates robust sequence extraction routines, automated computation of genomic features (including nucleotide composition, the percentage of nitrogenous bases in a DNA or RNA molecule that are either guanine or cytosine (GC content) and gene counts) and advanced statistical analytics using PCA and Mahalanobis distance-based outlier detection (De Maesschalck et al., 2000; Pedregosa et al., 2011). Furthermore, it includes a suite of intuitive visualization modules, such as interactive Sankey diagrams, for exploring taxonomic hierarchies, dendrogram-based hierarchical clustering, correlation heatmaps and cluster analyses via K-means algorithms, thus providing powerful exploratory data analysis capabilities accessible through a graphical user interface (GUI).

The software was designed explicitly to reduce barriers for biologists who may not possess extensive bioinformatics experience. By using libraries, such as Biopython (Cock et al., 2009), based on the Python programming language and custom scripts for sequence parsing, Tkinter for GUI implementation and scikit-learn for statistical analyses (Pedregosa et al., 2011), GeneBank Genie offers both versatility and ease of use, enabling efficient analysis of large genomic datasets without manual scripting or extensive computational training. Frequently, comparative genomic studies require rapid screening and preliminary characterization of sequence datasets before undertaking detailed analytical procedures. GeneBank Genie addresses this workflow need by providing streamlined, interactive analysis of GenBank records across three core dimensions: 1) nucleotide composition analysis, to quantify base frequencies and identify compositionally anomalous sequences; 2) taxonomic visualization, to hierarchically organize and visualize organism relationships; and (3) gene-centric analysis, to extract and compare annotated genes from GenBank records. Specifically, the tool processes GenBank-formatted files, leveraging their structured metadata and annotations to enable these analyses.

Materials and Methods

Software development and implementation

GeneBank Genie was developed as a desktop-based bioinformatics toolkit using Python 3.11. The GUI was created using the Tkinter library (<https://docs.python.org/3/library/>

tkinter.html), which offers a versatile framework for intuitive, user-friendly software development. The GUI is organized into distinct, modular analysis tabs: General Analysis, Gene Analysis, Taxonomic Visualization (Sankey diagram), Additional Visualizations (correlation heatmaps, K-means clustering, pair plots), Dendrogram Analysis, Sequence Extraction and a dedicated Summary tab for real-time logging

and monitoring of analysis results. Each analytical module provides customizable parameters, interactive visual elements and comprehensive logging to ensure user-friendly operation for researchers without extensive bioinformatics expertise. Fig. 1 provides an overview of the software's core architecture and user-driven analysis flow.

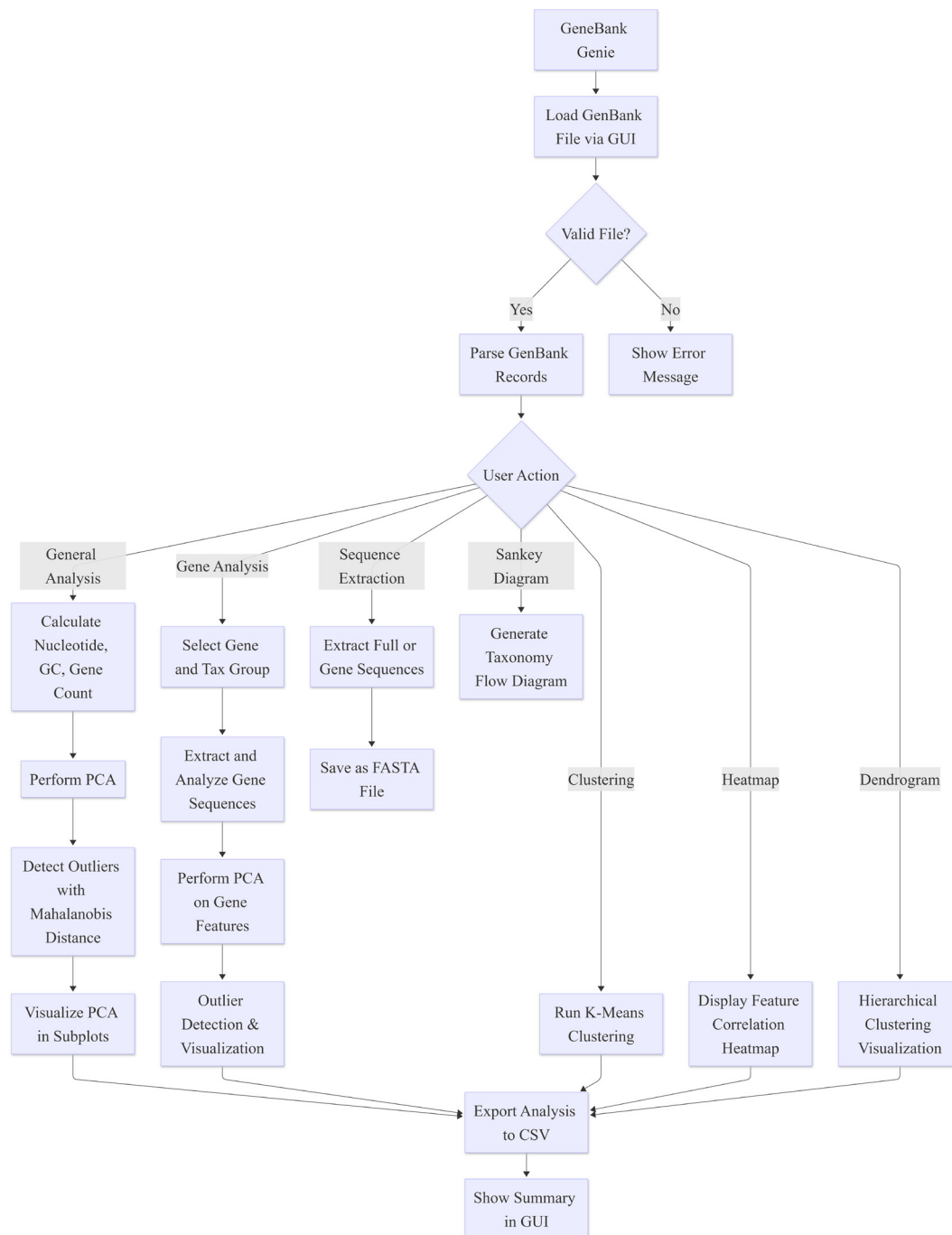


Fig. 1 Flowchart illustrating key components and analysis flow in GeneBank Genie for gene-level and preliminary genomic exploration.

Data handling and sequence parsing

The Biopython library (SeqIO and SeqRecord) (Cock et al., 2009) was used for the parsing and management of sequence data from GenBank records. Custom scripts were developed to systematically extract targeted gene sequences across multiple GenBank files based on specified gene qualifiers. Biopython facilitated automated loading, parsing and extraction of sequence information and metadata (including taxonomy annotations, gene features, nucleotide sequences and genomic annotations). Customized Python scripts were adjusted for rapid processing and preparation of sequence data for downstream molecular applications.

Statistical and genomic feature computations

NumPy (Harris et al., 2020) and Pandas libraries (McKinney, 2011) were used for efficient numerical computations and structured data management. Upon loading GenBank files, the software automatically computed a variety of genomic features for each record, including the nucleotide proportions of adenine (A), cytosine (C), guanine (G) and thymine (T), the GC content, the total sequence length and the total number of annotated genes. These features were compiled systematically into structured Pandas DataFrames, providing an organized dataset for subsequent statistical analyses and visualization tasks.

PCA, implemented using scikit-learn's PCA module, was performed on various feature subsets to reduce dimensionality and uncover underlying genomic variation (Ringnér, 2008; Pedregosa et al., 2011). PCA was applied separately to the nucleotide components (A, C, G, T), additional genomic features (sequence length, the GC content and the gene count) and combined feature sets. For single-feature visualization (GC content, sequence length, gene count), individual features were duplicated to enable two-dimensional scatter plot representation, as PCA mathematically requires at least two input variables to produce two principal components. This approach maintained visual consistency across all PCA comparisons and allowed preliminary assessment of taxonomic group separation based on individual genomic features. Outlier detection used the Mahalanobis distance metric (De Maesschalck et al., 2000), computed for each record within a selected taxonomic group, based on PCA-transformed features. Thresholds for outlier detection were determined using the chi-square (χ^2) distribution at the 95% confidence interval with two degrees of freedom ($df = 2$), corresponding to the

two-dimensional PCA-projected space in which the Mahalanobis distances were calculated, thereby identifying sequences significantly deviating from group centroids. Additionally, clustering analyses utilized K-means clustering algorithms (MacQueen, 1967) that were implemented via scikit-learn to calculate silhouette scores to evaluate clustering quality and robustness. Hierarchical clustering was executed using SciPy's hierarchical clustering (linkage) module (Virtanen et al., 2020), applying Ward's method (Ward, 1963) to standardize genomic features, with visualization based on dendrogram plots. This analysis allows users to explore genomic relationships and groupings within their datasets.

Visualization methods

GeneBank Genie incorporates multiple visualization libraries to facilitate intuitive exploration and interpretation of genomic data. Static PCA plots, dendrograms and K-means clustering visualizations were generated using Matplotlib (Hunter, 2007), while Seaborn (Waskom, 2021) was used to create detailed correlation heatmaps and pair plots, elucidating relationships among nucleotide proportions and genomic features. Plotly's interactive Sankey diagrams dynamically visualized hierarchical flows across taxonomic ranks, enabling immediate exploration and identification of annotation patterns or anomalies within the dataset. GeneBank Genie also provides flexible plot export and customization capabilities: after generating any PCA, heatmap, Sankey diagram, dendrogram or clustering figure, users can fine-tune visual parameters (color palettes, marker styles, axis labels, sizes) directly within the software GUI.

Logging and error handling

Robust error handling mechanisms were implemented throughout GeneBank Genie using Python's logging module to record detailed logs of all analysis steps, user interactions and software states at the INFO level or higher. Tkinter's messagebox alerts were integrated to inform users promptly of errors or incorrect inputs, providing clear guidance for troubleshooting and maintaining the software's accessibility and ease of use. Comprehensive logging and real-time updates were collated on the Summary tab, offering users immediate access to analytical results and detailed action logs, thereby enhancing reproducibility and auditability of analyses performed.

Dataset description

Demonstration and evaluation were assisted by using a curated dataset of 333 complete mitochondrial genome sequences from the insect order Orthoptera (grasshoppers, crickets, katydids), comprising a 13.4 MB GenBank file. This served as the primary dataset with sequences having a range of 14,851–18,255 base pairs in length and including common mitochondrial genes along with taxonomic metadata across 12 families within Orthoptera. This dataset was used for demonstrating GeneBank Genie's capabilities and is available at <https://github.com/yashmgupta/GeneBank-Genie>. In addition, GeneBank Genie was tested with a progressively larger secondary test dataset, consisting of 1,032 mixed arthropod mitochondrial genomes (37.2 MB).

Ethics statement

This study did not involve the use of human, animal, or biosafety-regulated subjects and therefore no ethical approval was required.

Results

User interface

GeneBank Genie provides a clean and intuitive GUI, as shown in the screenshot in Fig. 2, which is organized into seven tab-based analytical views: General Analysis,

Gene Analysis, Sankey Diagram, Additional Visualizations, Dendrogram Analysis, Sequence Extraction and Summary. Each tab clearly defines analytical parameters, enabling interactive customization of taxonomic levels, visualization styles and clustering methods. Descriptive labels, user-friendly dropdowns, interactive selection boxes and real-time logging in a dedicated summary pane ensure seamless navigation, even for users without bioinformatics expertise.

Example use case

GeneBank Genie's capabilities were tested using a dataset comprising 333 complete mitochondrial genome sequences from the insect order Orthoptera available through NCBI's RefSeq database (O'Leary et al., 2016; Sayers et al., 2024). Sequences were retrieved using the query “((mitochondrion, complete genome) AND Orthoptera)) AND RefSeq [Nucleotide]”. Sequences were retrieved on April 11, 2025 and the dataset (13.4 MB GenBank file) is available at <https://github.com/yashmgupta/GeneBank-Genie>.

While the current study focused on the analysis of 333 Orthoptera mitochondrial genomes as a use case demonstration, additional testing was performed using a larger dataset (37.2 MB GenBank file, $n=1,032$) to evaluate performance at scale. The larger dataset was processed successfully by all the GeneBank Genie modules with associated linear scaling in the processing time. Complete visualizations and analysis outputs from this extended testing are available in the public GitHub repository (<https://github.com/yashmgupta/GeneBank-Genie>),

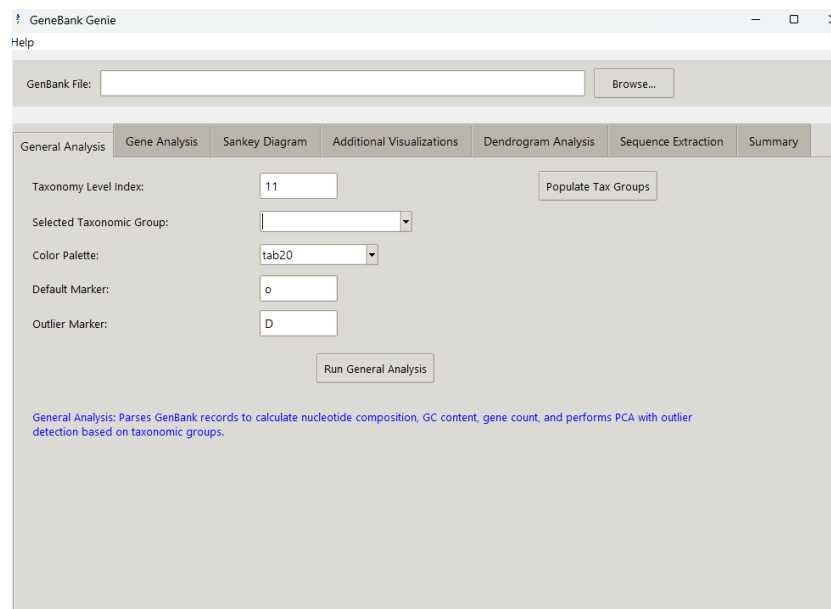


Fig. 2 GeneBank Genie Interface showing seven tabs that guide users through each stage of analysis

with these outputs demonstrating the tool's capability to handle datasets approximately three times larger than the primary test case.

Considering the test case with the primary 333 records, upon loading the dataset into GeneBank Genie's General Analysis module (Fig. S1), nucleotide composition, sequence length, GC content and annotated gene counts were computed automatically across all 333 mitochondrial records. The PCA performed on the combined genomic features revealed distinct clustering patterns among the Orthopteran species (Fig. S2). Mahalanobis distance-based outlier detection (De Maesschalck et al., 2000) identified several notable outliers, representing sequences with unusual nucleotide compositions or atypical genomic features, clearly summarized on the software's Summary tab. The Gene Analysis module lets the user select a target gene (such as COX1) from the GenBank file and then extracts and filters all matching coding sequences, computes per-sequence metrics (length, A/C/G/T proportions, GC content) and groups records based on a user-specified taxonomic level. Next, it standardizes these six features, performs PCA to reduce them to two principal components and flags outliers within the chosen taxonomic group, based on Mahalanobis distances against a chi-square threshold. Finally, it displays an interactive scatter plot colored by taxon with distinct markers for outliers.

Using the taxonomic Sankey Diagram module, we explored the hierarchical taxonomic annotation flow starting at family-level taxonomy, visualizing the detailed lineage from order down to species. This interactive diagram revealed insightful patterns, including unexpected taxonomic annotation splits and inconsistencies, allowing immediate identification of potential annotation errors or anomalies within the dataset (Fig. S3).

Further analyses using the Additional Visualizations module provided deeper insights, with the correlation heatmap clarifying significant correlations among nucleotide proportions, the GC content and pair plots indicated inter-feature relationships, while k-means clustering further subdivided the mitochondrial genomes into biologically meaningful groups, with high silhouette scores indicating robust clustering performance (Fig. S4). In addition, hierarchical clustering based on dendrogram analysis supported these findings, providing clear visualizations of genetic relationships and taxonomic groupings among the Orthopteran mitochondrial genomes (Fig. S5).

Finally, the Sequence Extraction module facilitated rapid extraction of the selected gene sequences from all 333 mitochondrial genomes into a single FASTA file, streamlining

the workflow for subsequent molecular analyses such as phylogenetic reconstructions or primer design. Currently, GeneBank Genie is optimized specifically for GenBank format files due to their rich structured metadata, including taxonomic hierarchy, gene annotations and feature tables. The current design intentionally leverages the comprehensive metadata structure unique to GenBank records. Overall, this comprehensive demonstration highlighted GeneBank Genie's powerful integration of data processing, sophisticated analytics and intuitive visualization, greatly simplifying genomic analyses for a broad research audience.

Selection and export

GeneBank Genie provides multiple, explicit ways to select or exclude records prior to analysis and to export the selected data for downstream workflows. Before running analyses, users may filter records by taxonomic level (via the Taxonomy Level index and the taxon selector), restrict analyses to records that contain a particular annotated gene (Gene Analysis using select gene), or apply numeric thresholds (for example, GC content or sequence length). Furthermore, visual clustering and outlier detection (PCA scatterplots, K-Means clusters, dendrograms) allow users to identify groups of interest and export subsets. Export options are: 1) FASTA export of either full sequences or extracted gene sequences; and 2) CSV export of computed features and PCA coordinates (such as `'pca_features.csv'` and `'selected_gene_pca_features.csv'`). The Summary tab provides quality assurance by listing any records excluded due to data issues (undefined or empty sequences), allowing users to review potential problems and decide whether to clean and re-analyze the data.

Software availability and installation

GeneBank Genie is freely available from its GitHub repository (<https://github.com/yashmgupta/GeneBank-Genie>). Detailed installation and usage instructions are provided in the project's README.

Discussion

GeneBank Genie was developed to address a critical gap in desktop bioinformatics software by providing a truly end-to-end, Python-based environment that begins with bulk GenBank parsing. Leveraging Biopython's SeqIO.parse,

the application ingests entire multi-record GenBank files in one operation, automatically harvesting each record's taxonomy metadata for dynamic filtering and visualization. While the demonstration utilizes mitochondrial genomes, GeneBank Genie's analytical framework is equally applicable to any GenBank dataset, regardless of genomic source or taxonomic scope, providing researchers with flexibility across diverse biological investigations. This contrasts with other tools, such as Geneious Prime, which, while powerful for alignment and annotation, treat GenBank files on a per-document basis and lack built-in PCA or Mahalanobis outlier detection. By centralizing metadata extraction at import, GeneBank Genie ensures that downstream analyses—whether PCA, clustering, or Sankey flow diagrams—operate on the full dataset rather than isolated entries.

Once records are loaded, GeneBank Genie extracts focus features nucleotide composition, sequence length, GC content and gene counts from every sequence in a single pass. Then, it conducts multifaceted PCA analyses, running separate PCAs on each feature set (and on a combined matrix) via scikit-learn. This rapid, integrated PCA overview highlights clustering patterns and variance contributions across taxa, enabling users to make informed decisions about which groups or sequences warrant deeper investigation. Mainstream GUIs, such as CLC Genomics Workbench, offer extensive NGS pipelines (Liu and Di, 2020; Zeden and Gründling, 2023) and statistical modules but do not natively support PCA for exploratory data analysis in this way.

A further advantage of GeneBank Genie is its statistical outlier detection. The tool objectively flags anomalous records based on computing Mahalanobis distances within any selected taxonomic group and applying a χ^2 threshold. This feature guides researchers in deciding which sequences to include or exclude from downstream phylogenetic reconstructions or targeted gene studies. Neither Geneious Prime (Kearse et al., 2012) nor CLC Genomics Workbench (Zeden and Gründling, 2023) includes an equivalent Mahalanobis-based outlier module (Paraskevopoulou et al., 2013) and SnapGene, while excellent for feature mapping and manual extraction (Shaha et al., 2022), offers no integrated clustering, PCA, or χ^2 -based outlier detection. GeneBank Genie complements rather than competes with comprehensive genomic platforms such as NCBI Genome Workbench (Kuznetsov et al., 2021) and Unipro UGENE (Okonechnikov et al., 2012). While those platforms excel at sequence alignment, phylogenetic reconstruction and protein annotation, GeneBank Genie specializes in rapid preliminary exploration and interactive visualization of taxonomic and

compositional patterns. Users seeking sequence alignment or phylogenetic tree construction should use these dedicated tools; GeneBank Genie serves as an efficient, accessible entry point for GenBank record characterization and quality control prior to in-depth downstream analysis.

GeneBank Genie also excels at visual decision support by integrating Plotly Sankey diagrams to illustrate taxonomy flow, SciPy dendrograms for hierarchical clustering, Seaborn heatmaps for correlation matrices and scikit-learn clustering plots; all accessible through a unified, seven-tab Tkinter notebook. A dedicated Summary tab logs every action and result in real time, creating an audit trail that enhances reproducibility. In contrast, most other GUI tools display status messages or console logs only, while none combines Sankey taxonomy flows with real-time logging in a single interface.

Despite these strengths, GeneBank Genie has some limitations. It does not currently support sequence alignment, BLAST integration (Ye et al., 2006), primer design, or NGS data processing features available in Geneious Prime, CLC Genomics Workbench and other specialized platforms. Additionally, the Tkinter-based GUI may face performance constraints with very large datasets, whereas tools built on optimized C++ backends can handle higher volumes more efficiently. However, the Tkinter library is popular among bioinformaticians for rapid prototyping and software development (Habib et al., 2019; Bessa-Silva, 2024). GeneBank Genie fills a unique niche by uniting bulk GenBank parsing, metadata-driven filtering, feature extraction, multifaceted PCA, statistical outlier detection and one-click sequence export within a single, pure-Python decision-support platform. While many bioinformatics GUIs excel at alignment, annotation and NGS workflows, none match GeneBank Genie's exploratory multivariate analytics, an integrated audit trail and automatic bulk feature analysis leveraging the annotated features in GenBank files to process hundreds of records at once and all within one cohesive, user-friendly platform. This integrated analytical approach enables rapid identification of taxonomic inconsistencies, gene annotation errors and sequence anomalies that might otherwise remain undetected in large genomic datasets, potentially preventing downstream analytical errors in phylogenetic reconstruction or comparative genomic analyses.

Notably, GeneBank Genie is not intended to supplant specialized bioinformatics pipelines or expert judgment, but rather to serve as an intuitive, preliminary research aid. With its low barrier to entry, automated analytical pipelines and real-time visualizations, GeneBank Genie can be seamlessly integrated

into teaching laboratories and research groups, empowering users to perform hands-on genomic data exploration, rapid hypothesis generation and preliminary sequence quality control, without requiring advanced computational expertise.

Conflict of Interest

The authors declare that there are no conflicts of interest.

Acknowledgements

This research was supported by the Department of Biology, Faculty of Science, Naresuan University, Thailand (Grant No. R2568B018) and partially supported by the Global and Frontier Research University Fund, Naresuan University (Grant No. R2566C051).

References

- Baxeavanis, A.D., Bateman, A. 2015. The importance of biological databases in biological discovery. *Curr. Protoc. Bioinform.* 50: 1–1. doi.org/10.1002/0471250953.bi0101s50
- Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Ostell, J., Pruitt, K.D., Sayers, E.W. 2018. GenBank. *Nucleic Acids Res.* 46: D41–D47. doi.org/10.1093/nar/gkx1094
- Bessa-Silva, A. 2024. Fast2Structure: A user-friendly tool for converting multiple aligned FASTA files to STRUCTURE format. *BMC Bioinform.* 25: 73.
- Cock, P.J.A., Antao, T., Chang, J.T., et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25: 1422–1423. doi.org/10.1093/bioinformatics/btp163
- De Maesschalck, R., Jouan-Rimbaud, D., Massart, D.L. 2000. The Mahalanobis distance. *Chemom. Intell. Lab. Syst.* 50: 1–18. doi.org/10.1016/S0169-7439(99)00047-7
- Fernandez-Suarez, X.M., Galperin, M.Y. 2012. The 2013 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. *Nucleic Acids Res.* 41: D1–D7. doi.org/10.1093/nar/gks1297
- Habib, P.T., Alsamman, A.M., Hamwieh, A. 2019. BioAnalyzer: bioinformatic software of routinely used tools for analysis of genomic data. *Adv. Biosci. Biotechnol.* 10: 33–41.
- Harris, C.R., Millman, K.J., van der Walt, S.J., et al. 2020. Array programming with NumPy. *Nature* 585: 357–362. doi.org/10.1038/s41586-020-2649-2
- Hunter, J.D. 2007. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* 9: 90–95. doi.org/10.1109/MCSE.2007.55
- Kearse, M., Moir, R., Wilson, A., et al. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28: 1647–1649.
- Kuznetsov, A., Bollin, C.J. 2021. NCBI genome workbench: desktop software for comparative genomics, visualization and GenBank data submission. *Methods Mol. Biol.* 2231: 261–295. doi.org/10.1007/978-1-0716-1036-7_16
- Liu, C.H., Di, Y.P. 2020. Analysis of RNA sequencing data using CLC genomics workbench. *Methods Mol Biol.* 2102: 61–113. doi.org/10.1007/978-1-0716-0223-2_4
- MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley, CA, USA. pp. 281–297.
- McKinney, W. 2011. Pandas: A Foundational Python Library for Data Analysis and Statistics. *Python for High Performance and Scientific Computing* 14: 1–9.
- Miller, H., Norton, C.N., Sarkar, I.N. 2009. GenBank and PubMed: how connected are they? *BMC Res. Notes* 2: 101. doi.org/10.1186/1756-0500-2-101
- O’Leary, N.A., Wright, M.W., Brister, J.R., et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion and functional annotation. *Nucleic Acids Res.* 44: D733–D745.
- Okonechnikov, K., Golosova, O., Fursov, M. and Ugene Team, 2012. Unipro UGENE: A unified bioinformatics toolkit. *Bioinformatics* 28: 1166–1167.
- Paraskevopoulou, M.D., Vlachos, I.S., Athanasiadis, E., Spyrou, G. 2013. BiDaS: a web-based Monte Carlo BioData Simulator based on sequence/feature characteristics. *Nucleic Acids Res.* 41: W582–W586.
- Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12: 2825–2830.
- Ringnér, M. 2008. What is principal component analysis? *Nat. Biotechnol.* 26: 303–304. doi.org/10.1038/nbt0308-303
- Sayers, E.W., Beck, J., Bolton, E.E., et al. 2024. GenBank 2024 update. *Nucleic Acids Res.* 52: D134–D137. doi.org/10.1093/nar/gkad903
- Sayers, E.W., Cavanaugh, M., Clark, K., Ostell, J., Pruitt, K.D., Karsch-Mizrachi, I. 2020. GenBank. *Nucleic Acids Res.* 48: D84–D86. doi.org/10.1093/nar/gkz956
- Shaha, M., Majumder, S., Hossain, M.S., et al. 2022. Identification of a novel variant of hepatitis B virus isolated from patient co-infected with hepatitis C virus. *Virus Res.* 319: 198859.
- Virtanen, P., Gommers, R., Oliphant, T.E., et al. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* 17: 261–272. doi.org/10.1038/s41592-019-0686-2
- Ward, J.H. 1963. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* 58: 236–244. doi.org/10.1080/01621459.1963.10500845
- Waskom, M.L. 2021. Seaborn: statistical data visualization. *J. Open Source Softw.* 6: 3021. doi.org/10.21105/joss.03021
- Ye, J., McGinnis, S., Madden, T.L. 2006. BLAST: improvements for better sequence analysis. *Nucleic Acids Res.* 34: W6–W9.
- Zeden, M.S., Gründling, A. 2023. Bacterial whole-genome-resequencing analysis: basic steps using the CLC Genomics Workbench software. *Cold Spring Harb. Protoc.* 2023: 16. doi.org/10.1101/pdb.prot107901