# GeneConnector: Unlocking the Full Potential of Genbank Metadata

Samuel Galvão Elias, Débora Cervieri Guterres, Robert Weingart Barreto, and Helson Mário Martins do Vale

***Abstract*—The Genbank database serves as a pivotal global repository for genetic information, housing an extensive and diverse array of data. Nonetheless, a significant proportion of its existing records suffer from fragmented and often inadequate metadata, thereby failing to furnish the requisite contextual information regarding their acquisition. In response to this challenge, we introduce GeneConnector, a novel tool designed to harness shared information within multiple records of the same specimen in Genbank, with the ultimate objective of augmenting the completeness of inadequately annotated nodes spanning various information domains. To exemplify the capabilities of this tool, we conducted a comprehensive review and aggregation of available data, utilizing the database for Genera of Phytopathogenic Fungi (GOPHY) as a testbed. Our evaluation revealed substantial improvements in information retrieval through the analysis of shared data among nodes that connect Genbank specimen records, yielding notable enhancements ranging from 2% to an astonishing 60%. Our approach equips users with the means to conduct precise, facile, and accurate assessments of the contextual associations of results, facilitated by two distinct metrics that assess the current level of data annotation and the potential information enhancement achievable through our evaluation, the Observed Completeness Score (OCS) and the Reachable Completeness Score (RCS).**

**Link to graphical and video abstracts, and to code: https://latamt.ieeer9.org/index.php/transactions/article/view/8241**

***Index Terms*—Genbank, NCBI, Gene-Connector, Mycology, Phytopathology, GOPHY.**

## I. Introduction

Genomic data has become increasingly important in many fields of research, including medicine [1], biotechnology, and agriculture [2]–[4]. However, the sheer volume and complexity of this data can make it challenging to extract meaningful insights. Public databases, such as GenBank [5], Ensembl [6], and UniProt [7], provide a wealth of information on genes, genomes, and their products. However, accessing and analyzing this information can be a time-consuming and daunting task. Therefore, the development of tools that can perform the aggregation of genomic metadata in public databases is critical to advancing research in genomics.

One example of such a tool is BioMart [8], a widely used data management system that allows users to query "data" (only) from multiple biological databases simultaneously. BioMart has been used in many studies, including the identification of genetic markers associated with disease and the exploration of gene expression patterns in different tissues. Another example is Ensembl's Biomart (see details), which allows users to query Ensembl's databases using the same interface as BioMart. These tools are just a few examples of the many resources available to researchers looking to access and analyze genomic "data".

Box 1| The *Ceratocystis mangicola* [9] study-case.

| | |
|---|---|
| **ITS** | Submitted in Mar 4, 2005 (available under the Genbank accession nº AY953382, [10]). |
| **EF1** | Feb 13, 2007 (EF433316, [11]). |
| **TUB1** | Feb 13, 2007 (EF433307, [11]). |
| **RPB2** | Mar 11, 2014 (KJ601618, [12]). |
| **MS204** | Mar 11, 2014 (KJ601582, [12]). |

A pathogen originally described as member of the *Ceratocystis fimbriata sensu lato* complex causing the *Mangifera indica* disease, known as mango blight, murcha, or *seca da mangueira* in Brazil. Records of *C. mangicola* were registered in three different submission events, with a large time lag between the first (the Internal Transcribed Spacer [ITS] submission) and the latest submission events (RNA polymerase subunit II [RPB2] and guanine nucleotide-binding protein subunit beta-like protein [MS204], nine years after). Such time lag allowed the information associated to the *C. mangicola* to be gradually extended. Since the first registration of the ITS marker, the associated information was upgraded, starting from basic source modifiers as isolate and organism name to a well documented record including strain, specimen-voucher, type-materials, host, country, and others (see the Fig. 1 for details of the information gain associated to *C. mangicola*).

As attempted readers can see, the "data" has been the center of attention when it comes to data aggregation, while metadata is much more often overlooked. Therefore, the aggregation of genomic metadata is important because it enables researchers to integrate data from multiple sources and make more comprehensive and accurate analyses (important examples includes [13]–[16]). For example, by combining

Samuel Galvão Elias and Helson Mário Martins do Vale are with University of Brasilia, Brasilia, Brazil (e-mail: sgelias@outlook.com and helson@unb.br).

Débora Cervieri Guterres and Robert Weingart Barreto are with Federal University of Viçosa, Brazil (e-mail: debora.guterres@gmail.com and rbarreto@ufv.br).

genomic data with clinical data, researchers can identify genetic markers associated with disease and develop more effective treatments. The aggregation of genomic metadata also enables the identification of patterns and trends that may not be apparent when examining individual datasets. These patterns and trends can provide insights into the most variable scientific domains.

Despite the existence and importance of the tools that performing aggregation of genomic metadata from single records, and focused in high-throughput sequencing data (examples include Metagenote [17], and ffq [18]), there are no tools that aggregate multi-loci data. There are still challenges associated with accessing and analyzing such data, and information consistency is maybe the most important of these (see [19] for a important study-case about Genbank information consistency). GeneConnector works around such challenges. Our proposal is to create connections between unique Genbank records and use the "unique + shared" information between records to improve single gene annotations.

When specifically dealing nucleotide data stored in Genbank, it is common to observe events of information increment associated with advancements in knowledge regarding taxonomic groups (see Box 1 for an example). The phenomenon of information increment events can be attributed to the dynamic nature of scientific research. As researchers delve deeper into the genetic makeup of various organisms, they uncover novel data points and identify previously unrecognized patterns.

These discoveries, when incorporated into the Genbank database, enhance the breadth and depth of information available for taxonomic analysis. Consequently, with each advancement in our understanding of taxonomic groups, a ripple effect occurs, influencing future studies, expanding our knowledge base, and fostering further scientific breakthroughs (the natural stepping stones of science [20]).

Finally, GeneConnector was designed precisely to absorb this intrinsic characteristic of Genbank data during metadata acquisition campaigns. Therefore, our aim is to illustrate how this tool can enhance the metadata quality of a comprehensive database solely by leveraging the shared information within Genbank records. Furthermore, we introduce our novel approach, the Observed Completeness Score (OCS) and the Reachable Completeness Score (RCS), for quantifying the level of completeness in records associated with specimens with available information in Genbank. To accomplish this objective, we employed the comprehensive database for Genera of Phytopathogenic Fungi (GOPHY) as a case study ( [21]–[24]).

## II. Problem Statement

Genbank, a widely used repository for nucleotide sequence data, contains an immense amount of valuable genomic information. However, the lack of consistent and standardized metadata across the records poses a significant challenge for researchers aiming to extract meaningful insights from this vast collection. Existing approaches for metadata extraction and aggregation from Genbank records are often limited, inefficient, or require manual curation, hindering the ability to comprehensively exploit the data for scientific research.

To address this problem, a novel software solution has been developed to automate the process of populating and aggregating metadata from Genbank nucleotide records. The software aims to extract diverse metadata attributes, including taxonomy, organism properties, sequencing techniques, geographical location, and biological features, among others, from the extensive Genbank database. By automating this labor-intensive task, researchers will be empowered to efficiently access and analyze metadata associated with nucleotide sequences, enhancing their ability to conduct comprehensive studies and accelerate advancements in various biological and genomic fields.

The research article aims to evaluate the effectiveness and reliability of the GeneConnector in extracting and aggregating metadata from a large-scale sample of Genbank nucleotide records. The outcomes of this research will contribute to improving the accessibility and quality of metadata associated with Genbank nucleotide records.

## III. Proposed Solution

### A. Concepts and Information Modelling

Our tool was developed to modelling the information contained in Genbank records based in three basic data models: Metadata, Nodes, and Connections (see Fig. 1). A *Connection* is a top-level object centralizing *Nodes*. A single *Node* carries information of the accession number that originated the object, and the gene marker from which the record was extracted, connecting all metadata related to the original Genbank record. *Metadata* objects abstract the Genbank raw qualifiers information.

Raw Genbank qualifiers includes a list of key/value pairs describing the context associated to given nucleotide record. Our tool was developed to turn qualifiers into importance groups (from here on we will call them Metadata Indicator Groups, MIG) that mirrors the information relevance which turn a desired specimen unique (information importance are available in Table I, and visually explored in Fig. 1). For example, a taxonomic key (e.g. organism [with value *Bipolaris victoriae*]) is shared between multiple real world specimes, so it must have less importance than a specimen related key (e.g. isolate [with value *CBS:327.64*]).

TABLE I
Metadata Indicator Groups Used to Rank Keys and Calculate the Gene Connector Completeness Scores

| Group | Score | Description |
|---|---|---|
| SPECIMEN | 8 | Unique identifiers of specimen. |
| TAXONOMY | 5 | Taxonomy related keys. |
| HOST_SUBSTRATE | 3 | Identifiers for host interactions. |
| TIME_REFERENCES | 2 | Time milestones. |
| GEO_REFERENCES | 2 | Geographical indicators. |
| ASSAY | 0 | Related to gene sequencing methods. |
| EXTERNAL_LINKS | 0 | References to external databases. |
| ACTORS | 0 | Human actors related to the record. |
| OTHER | 0 | Not already mapped keys. |

Based on these principles, our tool systematically punctuates metadata from independent Genbank records, and calculates

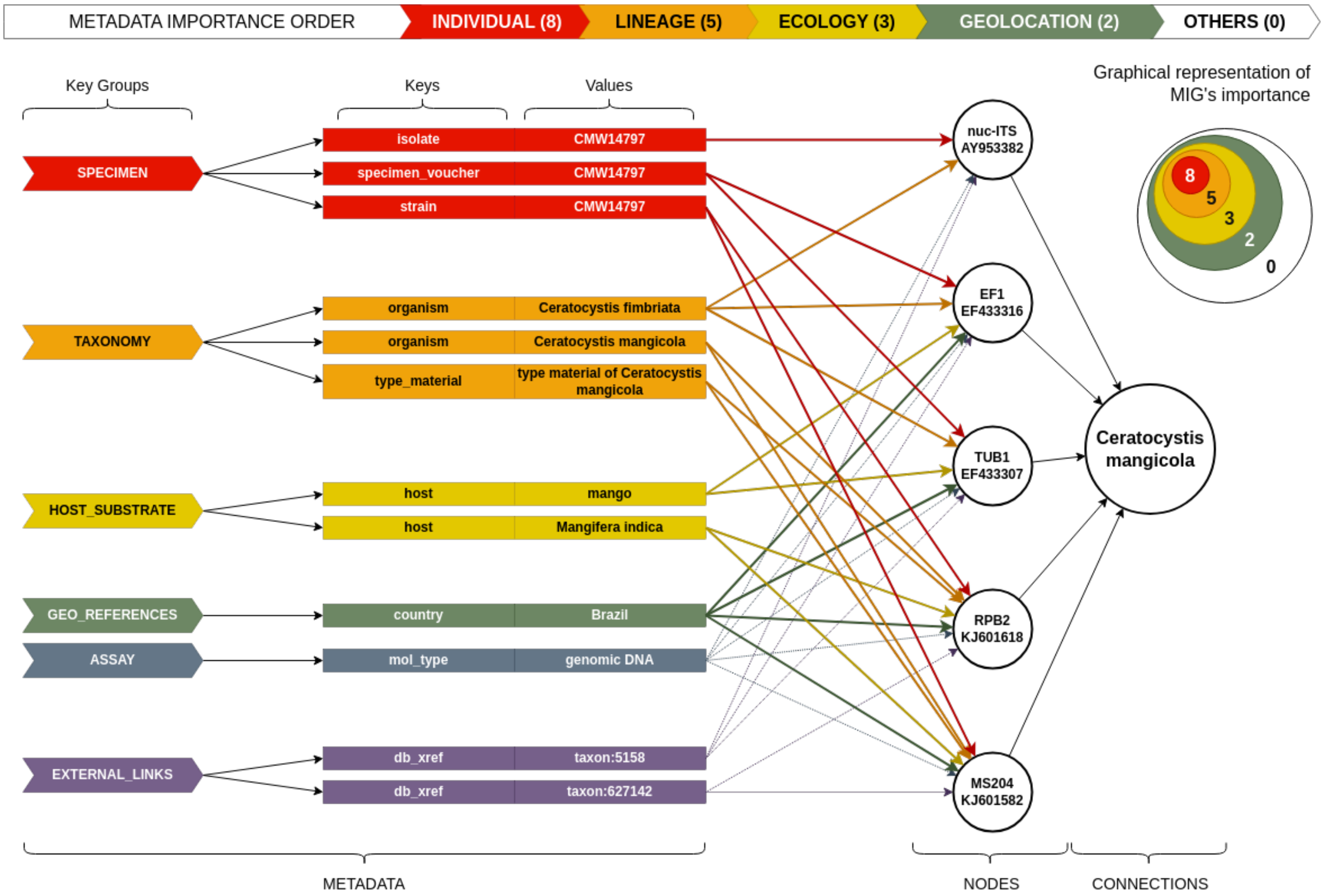


Fig. 1. GeneConnector information diagram. The diagram represents the main information models used by our tool to dealings with Genbank records information: Metadata, Nodes, and Connections. Dotted arrows connecting metadata and nodes layers indicating zero scored MIG's. Warmer colors indicate more specific information about the specimen to which the nodes (genes) belong. Therefore, the closer to red indicates metadata with greater power to approximate nodes.

the information completeness associated to a set of records that represents real world specimens, improving the information usability.

As already demonstrated in Table I, the MIG importance score is expressed on a Fibonacci scale and enables the differentiation of important metadata from spurious ones, thereby facilitating the calculation of two completeness scores: the Observed Completeness Score (OCS) and the Reachable Completeness Score (RCS).

The OCS measures how well a connection is annotated in terms of information domains taking into account parameters as uniqueness (SPECIMEN), tree of life placemant (TAXONOMY), ecological placement (HOST_SUBSTRATE), temporal marks (TIME_REFERENCES), spatial marks (GEO_REFERENCES), and less relevant ones (ASSAY, EXTERNAL_LINKS, ACTORS, and OTHER). The OCS is calculated independently for each node.

The RCS is a metric used to assess the completeness of connections based on the nodes they connect. Unlike the OCS, which considers all nodes independently, the RCS takes into account the dependencies between nodes that composes a connection. Specifically, the RCS is calculated as the ratio of the number of observed connections between nodes to the number of possible connections within a given MIG. This score reflects the degree to which the available metadata within a given MIG is interconnected and should be used to complete another nodes.

The three main steps of an hypothetical calculation of the OCS and RCS are described below where the results of individual steps are shown in Table II.

- Step 1. Finding nodes with at least one occurrence of qualifiers of each MIG with a score greater than zero:

  Let $N$ be the set of nodes.
  Let $Q$ be the set of qualifiers.
  Let $M$ be the set of MIGs.

  The condition to find such nodes can be represented as:

  $\forall n \in N, \exists m \in M, \exists q \in Q$ with score$(m) > 0$ such that occurs$(n, q)$

- Step 2. Annotating nodes with the MIG score (no more than one key per MIG scored by a node):

  Let $S$ be the function that assigns a score to a node.

The annotation can be represented as:

$$\forall n \in N, \exists m \in M, S(n) = \text{score}(m)$$

- Step 3. Calculating the expected score during the second step:

  Let $E$ represent the expected score.

  The calculation can be represented as the sum of products:

  $$E = \sum_{m \in M} \text{score}(m) \cdot \text{number of nodes}$$

  Finally we filtering expected scores to find MIGs with at least one member (penalizing MIGs with a zero score if not represented):

  Let $F$ be the set of MIGs with at least one member.

  The filtering can be represented as:

  $$F = \{m \in M \mid \exists n \in N \text{ such that occurs}(n, m)\}$$

  The penalization of MIGs not represented can be represented as:

  $$\forall m \in M \setminus F, \text{score}(m) = 0$$

TABLE II
THE FIRST THREE STEPS OF THE COMPLETENESS SCORES CALCULATION WITH HYPOTHETICAL NODES A, B, C, AND D

| | | Step 1 | | | Step 2 | Step 3 |
|---|---|---|---|---|---|---|
| Group | A | B | C | D | E-score† | 0-score‡ |
| SPECIMEN | 8 | 8 | - | 8 | 32 | 32 |
| TAXONOMY | 5 | 5 | 5 | 5 | 20 | 20 |
| HOST_SUBSTRATE | - | 3 | 3 | - | 12 | 12 |
| TIME_REFERENCES | - | - | - | - | 8 | 0 |
| GEO_REFERENCES | 2 | - | 2 | - | 8 | 8 |
| | 15 | 16 | 10 | 13 | 80 | 72 |
| Conn. Obs. score | 54 | | | | | |
| OCS | 0.68 | | | | | |
| RCS | 0.90 | | | | | |

Step 1 contain four hypothetical nodes A, B, C, and D with group scores, respective. Dash indicate groups not represented in Node. Step 2 and Step 3 includes expected scores, and non-zero group scores, respectively. † Expected score by group. The product of the nodes number and the score value of the given group. ‡ Non-zero score by group. The same as expected score if at last one Node contains a given group. Otherwise is zero.

After execute the above steps we can calculate the Connection Observed Score by sum individual node scores (conn. obs. score = 54 in Table II). Next, the OBS is calculated being the ratio between the Connection Observed Score and the sum of Expected scores (54 / 80 = 0.68 in Table II), and finally the RCS should be calculated as the ratio of the step 3's values sum and the sum of Expected scores (72 / 80 = 0.90 in Table II). This is a simple and elegant way to represents the information completeness of arbitrary Genbank records.

### B. Technologies and Code Availability

GeneConnector was developed in Python (3.11+ [25]) adopting the hexagonal architecture [26]. The complete logic for the calculation of scores, data parsing, data validations, and the data collection from Genbank are centered at the package core sub-module. For curious readers, a complete metadata list by MIG should be found at the Github repository sgelias/geneconnector-cli) whithin the 'metadata' file src/gcon/core/domain/dtos/metadata.py). Our tools is Open Source and the codebase is available under the MIT license (see details).

### C. Study Case: GOPHY Data Completeness

To demonstrate the performance and value proposition of GeneConnector we downloaded and evaluate the complete GOPHY's database containing seventeen gene markers and 1,246 specimen records. The complete database is available as a Supplementary material into the GeneConnector Github directory (files named gcon-input-gophy.xlsx in docs/manuscript/supplementary-material).

We value simplicity, so we make running the GeneConnector possible through a single command named **resolve** available after the tool installation on the host system. Currently our tool was tested only using Linux systems, thus, over Windows or Macintosh systems we recommend to run using a Docker environment [27]. See below the execution command of GeneConnector CLI:

The Code snipped of Listing 1 exemplifies our package execution. After installed GeneConnector should be called using the *gcon* callable and the *resolve* command used to execute the full package pipeline. Required arguments are shown in lines 5, 6, and 7 of the previous code snippet. A comprehensive user guide is available at the GeneConnector Github directory.

```
1 # GeneConnector execution in Linux environments
2 # using the ‘resolve‘ command of the ‘gcon‘ package.
3
4 $ gcon resolve \
5       --input-table input-table.tsv \
6       --temporary-directory /tmp/gcon/ \
7       --output-file gcon-out
```

Listing 1: GeneConnector execution command example. Lines started with hashtag are code commnets, so they are not executed.

The output generated by the aforementioned command encompasses a tabular file (TSV) that amalgamates several crucial components: (i) input table information, (ii) OCS and RCS scores, (iii) a statistical percentage depicting the information gain, which quantifies the quantity of information salvaged following the evaluation of metadata under the *Nodes* category, (iv) signatures, and ultimately, (v) all metadata associated with individual connections.

Signatures offer a streamlined mechanism enabling researchers to trace, index, or effortlessly compare results across multiple analyses conducted at disparate times. Our tool incorporates two distinct levels of signatures: the *connection-level* and *node-level* signatures, grounded in standard Universal

Unique Identifiers (UUID) of version 3 hashes. These hashes are derived by compressing the most pivotal data elements that constitute *Nodes* (comprising Genbank accession, source genome, gene name, and metadata keys and values) and *Connections* (encompassing identifiers and node signatures). Such an approach empowers users with the capability to replicate results and swiftly compare records when necessary.

Metadata columns are composed of the MIG keys concatenated to metadata keys (as example SPECIMEN.isolate). Such way turn the further integration and indexing as a simple and natural way to store GeneConnector results. In addition to the above cited tabular file, the GeneConnector results includes at default a JSON[1] formatted output file as a optimal format to be inputs into ETL[2] pipelines and web integrations.

## IV. Results

### A. MIG's Representativity and Distribution

Analysis of the complete GOPHY's database resulted in 414 events of information gain[3] from the total of 1,246 specimen records. These amount comprises 33% of the database records suffering information gains. Gains ranged from 2% up to 60%, widely distributed along all fungal genus included in our analysis. Twenty-five of the twenty-nine genera present in GOPHY were contemplated with information gains. The complete tabular results is available as a supplementary material.

The most important MIG obviously was SPECIMEN, with *strain* and *culture_collection* as the most populated keys, with 86.3% and 69.3% of coverage. It was not surprisingly due to the nature of the GOPHY database proposal itself, including only high quality records, mainly belonging to type materials.

Next, the TAXONOMY MIG with *type_material* as the second[4] most important key, with >69% of coverage in records. Both SPECIMEN and TAXONOMY are the most important keys to make each Connection record unique. And precisely for this reason that both are the best scored in our tool (see Table I).

The third most important MIG for GeneConnector approach is HOST_SUBSTRACT. Both keys *host* and *isolation_source* covered approximately 62% and 40%, respectively, of the full GOPHY dataset.

The next important MIG's is GEO_REFERENCES. It was present in about 76% of records, however in the most cases refereed to only as country. This key in most cases is not so geographically resolute when dealing with countries of continental dimensions, such as Brazil or Australia. From 1,246 records, just one included information of Latitude/Longitude, completely inhibiting the performance of geographic analyzes.

A world scale map indicating the geographic range of the records, and including the maximum information gain reachable by country is shown in Figure 2. The 10 countries covered with the highest number of information gain events were China, United States, Australia, South Africa, Brazil, Thailand, Netherlands, Indonesia, Japan, and Ecuador. The maximum information gain reached by such countries ranges between 30% and 50%. This is proportional to the country contribution to the state of the art of the phytopathogenic fungi records, a fully expected scenario.

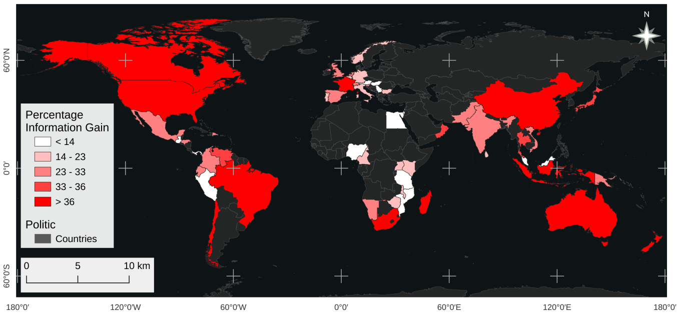


Fig. 2. Information gain at the globe scale. Records with some information gain registered in this study are highlighted in white-to-red scale (see scale legend).

Different from the previous cited MIG's the TIME_REFERENCE was an exception. Only about 6.9% (86 records) of the GOPHY database contains time milestones. Despite such MIG is not highly scored in GeneConnector (score = 2), the absence of this information inhibits temporal interpretation of the collection effort on the phytopathogenic important fungi around the world.

### B. Phytopathogenic Completeness Along GOPHY Genus

Information gains by genus are shown in Figure 3. As above cited, information gains ranged from 2% up to 60%. The top ten phytopathogenic fungal genus with most number of specimen records suffering information gains were *Calonectria* with 73 events, *Diaporte* (55), *Curvularia* (48), *Colletotrichum* (41), *Ceratocystis* (23), *Bipolaris* (21), *Boeremia* (19), *Neofusicoccum* (17), *Phyllosticta* and *Huntiella* with (16).

Using our approach 8 of 25 genus with information gains (GOPHY database include information of 29 genus) reached the full information completeness (100% of completeness, RCS = 1.0), grouping at last one of each MIG qualifier key per connection. A significative information gain in terms of the complete database. As can be seen in Fig. 3, median values of RCS were up to 90% in nine of ten most representative genus of GOPHY (cited in the previous paragraph).

## V. Conclusions

In this study, we showcase the remarkable ability of GeneConnector to substantially enhance the data completeness of specimens in Genbank by exclusively leveraging shared information within the records. Our findings demonstrate that utilizing our tool can yield gains of up to 60% in shared information among Genbank records, particularly for specific phytopathogenic genera. Furthermore, on a global scale, the data aggregation process holds the potential to benefit records from approximately 55 countries across the globe.

[1] Javascript Object Notation format.

[2] Extract, Transform, and Load pipelines.

[3] Calculated as the percentage of the Reachable Completeness Score which the Observed Completeness Score comprises.

[4] Organism is a required field, so it has full coverage in Genbank nucleotide records.

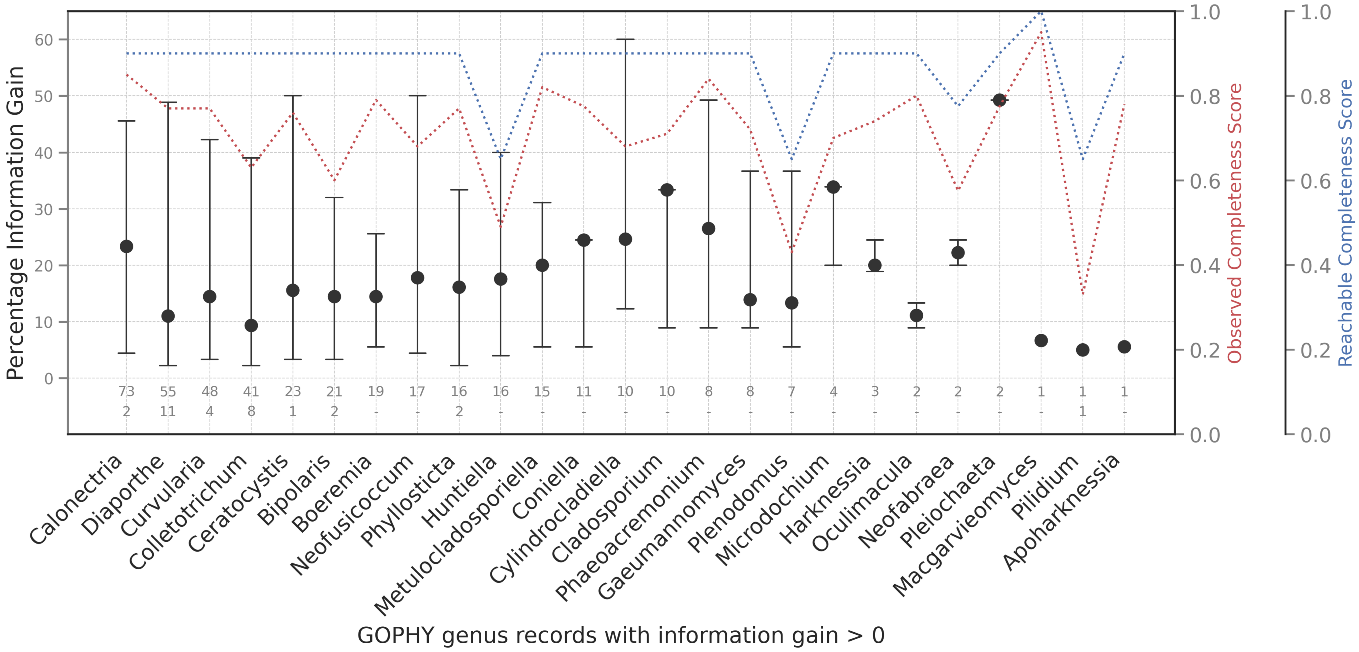


Fig. 3. Information gain by genus of phytopathogenic fungi registered in GOPHY database. Median with max/min values are presented in first Y-axis (left). Median values of Observed Completeness Scores and Reachable Completeness Scores are shown in 2nd and 3rd Y-axis (right), respectively. Only records with information gains greater zero were kept in chart. Numbers below zero in the X-axis indicates the number of records evaluated for each genus (upper number), and the number of record reaching the maximum reachable completeness (100%, lower number).

Moreover, our data aggregation process is both auditable and interpretable through two scores: the Observed Completeness Score (OCS) and the Reachable Completeness Score (RCS). These scores provide insights into the current level of information completeness and the attainable information based on shared metadata among nodes of the same specimen in Genbank.

With these comprehensive metrics, our aim is to make a significant contribution to the ongoing improvement of the information accumulation process, benefiting scientists worldwide and fostering continuous advancements in knowledge acquisition.

## Acknowledges

Thanks to the grant from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES). We also thank the anonymous reviewers who dedicated their valuable time to improving this work. Thanks to Biotrop, Solutions in Biological Technologies, for the support during the development of this work.

## References

[1] J. Shendure, G. M. Findlay, and M. W. Snyder, "Genomic medicine–progress, pitfalls, and promise," *Cell*, vol. 177, no. 1, pp. 45–57, 2019.

[2] R. Jeyasri, P. Muthuramalingam, L. Satish, S. K. Pandian, J.-T. Chen, S. Ahmar, X. Wang, F. Mora-Poblete, and M. Ramesh, "An overview of abiotic stress in cereal crops: negative impacts, regulation, biotechnology and integrated omics," *Plants*, vol. 10, no. 7, p. 1472, 2021.

[3] C. Juma, *The gene hunters: Biotechnology and the scramble for seeds*, vol. 996. Princeton University Press, 2014.

[4] E. J. Gilchrist, S. Wang, and T. D. Quilichini, "The impact of biotechnology and genomics on an ancient crop: Cannabis sativa," in *Genomics and the Global Bioeconomy*, pp. 177–204, Elsevier, 2023.

[5] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers, "Genbank," *Nucleic acids research*, vol. 41, no. D1, pp. D36–D42, 2012.

[6] K. L. Howe, P. Achuthan, J. Allen, J. Allen, J. Alvarez-Jarreta, M. R. Amode, I. M. Armean, A. G. Azov, R. Bennett, J. Bhai, *et al.*, "Ensembl 2021," *Nucleic acids research*, vol. 49, no. D1, pp. D884–D891, 2021.

[7] U. Consortium, "Uniprot: a worldwide hub of protein knowledge," *Nucleic acids research*, vol. 47, no. D1, pp. D506–D515, 2019.

[8] D. Smedley, S. Haider, B. Ballester, R. Holland, D. London, G. Thorisson, and A. Kasprzyk, "Biomart–biological queries made easy," *BMC genomics*, vol. 10, no. 1, pp. 1–12, 2009.

[9] M. Van Wyk, B. D. Wingfield, A. O. Al-Adawi, C. J. Rossetto, M. F. Ito, and M. J. Wingfield, "Two new ceratocystis species associated with mango disease in brazil," *Mycotaxon*, vol. 117, no. 1, pp. 381–404, 2011.

[10] M. Van Wyk, A. Al-Adawi, B. Wingfield, A. Al-Subhi, M. Deadman, and M. Wingfield, "Dna based characterization of ceratocystis fimbriata isolates associated with mango decline in oman," *Australasian Plant Pathology*, vol. 34, pp. 587–590, 2005.

[11] M. Van Wyk, A. O. Al Adawi, I. A. Khan, M. L. Deadman, A. A. Al Jahwari, B. D. Wingfield, R. Ploetz, and M. J. Wingfield, "Ceratocystis manginecans sp. nov., causal agent of a destructive mango wilt disease in oman and pakistan," *Fungal Divers*, vol. 27, pp. 213–230, 2007.

[12] A. Fourie, M. J. Wingfield, B. D. Wingfield, and I. Barnes, "Molecular markers delimit cryptic species in ceratocystis sensu stricto," *Mycological Progress*, vol. 14, pp. 1–18, 2015.

[13] A. Canakoglu, A. Bernasconi, A. Colombo, M. Masseroli, and S. Ceri, "Genosurf: metadata driven semantic search system for integrated genomic datasets," *Database*, vol. 2019, 2019.

[14] Z. Chen, A. S. Azman, X. Chen, J. Zou, Y. Tian, R. Sun, X. Xu, Y. Wu, W. Lu, S. Ge, *et al.*, "Global landscape of sars-cov-2 genomic surveillance and data sharing," *Nature genetics*, vol. 54, no. 4, pp. 499–507, 2022.

[15] U. Kõljalg, K.-H. Larsson, K. Abarenkov, R. H. Nilsson, I. J. Alexander, U. Eberhardt, S. Erland, K. Høiland, R. Kjøller, E. Larsson, *et al.*, "Unite: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi," *New Phytologist*, vol. 166, no. 3, pp. 1063–1068, 2005.

[16] K. Abarenkov, R. H. Nilsson, K.-H. Larsson, I. J. Alexander, U. Eberhardt, S. Erland, K. Høiland, R. Kjøller, E. Larsson, T. Pennanen, *et al.*,

"The unite database for molecular identification of fungi–recent updates and future perspectives," *The New Phytologist*, vol. 186, no. 2, pp. 281–285, 2010.

[17] M. Quiñones, D. T. Liou, C. Shyu, W. Kim, I. Vujkovic-Cvijin, Y. Belkaid, and D. E. Hurt, "Metagenote: a simplified web platform for metadata annotation of genomic samples and streamlined submission to ncbi's sequence read archive," *BMC bioinformatics*, vol. 21, pp. 1–12, 2020.

[18] Á. Gálvez-Merchán, K. H. Min, L. Pachter, and A. S. Booeshaghi, "Metadata retrieval from sequence databases with ffq," *Bioinformatics*, vol. 39, no. 1, p. btac667, 2023.

[19] Q. Chen, J. Zobel, and K. Verspoor, "Duplicates, redundancies and inconsistencies in the primary nucleotide databases: a descriptive study," *Database*, vol. 2017, 2017.

[20] S. Reining, F. Ahlemann, B. Mueller, and R. Thakurta, "Knowledge accumulation in design science research: ways to foster scientific progress," *ACM SIGMIS Database: the DATABASE for Advances in Information Systems*, vol. 53, no. 1, pp. 10–24, 2022.

[21] Y. Marin-Felix, J. Groenewald, L. Cai, Q. Chen, S. Marincowitz, I. Barnes, K. Bensch, U. Braun, E. Camporesi, U. Damm, *et al.*, "Genera of phytopathogenic fungi: Gophy 1," *Studies in mycology*, vol. 86, pp. 99–216, 2017.

[22] Y. Marin-Felix, M. Hernández-Restrepo, M. J. Wingfield, A. Akulov, A. Carnegie, R. Cheewangkoon, D. Gramaje, J. Z. Groenewald, V. Guarnaccia, F. Halleen, *et al.*, "Genera of phytopathogenic fungi: Gophy 2," *Studies in mycology*, vol. 92, pp. 47–133, 2019.

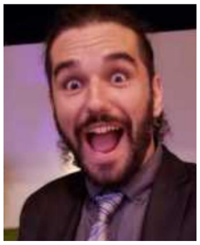
[23] Y. Marin-Felix, M. Hernández-Restrepo, I. Iturrieta-González, D. García, J. Gené, J. Z. Groenewald, L. Cai, Q. Chen, W. Quaedvlieg, R. Schumacher, *et al.*, "Genera of phytopathogenic fungi: Gophy 3," *Studies in mycology*, vol. 94, pp. 1–124, 2019.

[24] Q. Chen, M. Bakhshi, Y. Balci, K. Broders, R. Cheewangkoon, S. Chen, X. Fan, D. Gramaje, F. Halleen, M. Horta Jung, *et al.*, "Genera of phytopathogenic fungi: Gophy 4," *Studies in Mycology*, vol. 101, no. 1, pp. 417–564, 2022.

[25] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009.

[26] A. Cockburn, "Ports and adapters architecture," 2006. http://wiki.c2.com/?PortsAndAdaptersArchitecture [Accessed: 2022-11-20].

[27] D. Merkel, "Docker: lightweight linux containers for consistent development and deployment," *Linux journal*, vol. 2014, no. 239, p. 2, 2014.

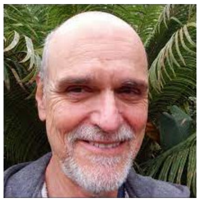


**Samuel Galvão Elias** I am a biologist/microbiologist and I value multi- and interdisciplinary approaches. As a biologist, my main focus is on mycology, and I have a strong knowledge of bacteriology as well. As a bioinformatician, I have experience in analyzing molecular data of various types. I also have expertise in analyzing microbial diversity, including community experimentation across a wide range of taxonomic groups. Additionally, I have extensive knowledge in molecular phylogenetics of eukaryotes and prokaryotic groups, along with experience in post-phylogenetics. In terms of Data Science, I have expertise in analyzing diverse classes of data, including univariate to multivariate, unifactorial to multifactorial, and categorical to continuous data.

As a developer, I have experience in web application development (monolithic and distributed), embedded systems, desktop applications, and data pipelines both within and outside the field of bioinformatics. My language stack includes Python, R, Rust, Golang, JavaScript, and TypeScript, with experience in single and multithreaded development, single and multicore programming, concurrent programming, and parallel programming. I am involved in the architecture and development of stateful and stateless applications, native to cloud environments, with a focus on Kubernetes. Some of my main open-source projects include Mycelium (Rust), an API gateway currently under development that focuses on permissioning in distributed environments, and Blutils (Rust), a tool for optimizing the execution process and analysis of Blast results.

**Débora Cervieri Guterres** holds a Ph.D. in Phytopathology from the University of Brasília (UnB, 2018), a Master's degree in Environmental Sciences from the Federal University of Bahia (UFBA, 2013), and specializes in Environmental Management from FJC (2009). Additionally, earned a degree in Agronomist Engineering from FASB (2012) and holds a Bachelor's degree in Business Administration with a focus on Foreign Trade from FASB (2007).

With a diverse academic background, Debora has expertise in the fields of Phytopathology, Mycology, Etiology, and the Diversity and Taxonomy of Fungi. She has conducted extensive research in these areas, contributing to the understanding and management of plant diseases. Currently, Debora is engaged in a postdoctoral internship at the Federal University of Viçosa, further expanding her knowledge and expertise in the field.



**Robert Weingart Barreto** is an agronomist (UFRRJ) with a strong academic background in mycology. He obtained an MSc in Pure and Applied Taxonomy (Mycology) from the University of Reading in 1986 and went on to complete his Ph.D. in Botany (Mycology) at the same institution, along with the International Institute of Mycology (currently CAB International) in 1991. Following his doctoral studies, he pursued postdoctoral research focusing on molecular taxonomy of fungi at the esteemed Centraalbureau voor Schimmelcultures.

Currently, he holds the position of full professor in the Department of Phytopathology at the Federal University of Viçosa, where he is actively involved in teaching various courses in the field of mycology, plant disease diagnosis, and biological control. Since its establishment in 1998, he has been the dedicated Coordinator of the Plant Disease Clinic - DFP/UFV, ensuring effective management and treatment of plant diseases.

With extensive experience in mycology, his research interests encompass a wide range of topics. His expertise lies in the areas of biological control of weeds, fungal taxonomy, phytopathology, diagnosis of fungal diseases in plants, and the study of fungal biodiversity in Brazilian ecosystems. As an accomplished researcher, he has made significant contributions to these fields and is recognized as a leading figure in the discipline.

Furthermore, he holds the esteemed position of the current president of the Brazilian Society of Mycology, where he actively promotes collaboration and advances in mycological research. Through his leadership, he plays a pivotal role in shaping the direction of mycology in Brazil and fostering connections within the scientific community.

**Helson Mário Martins do Vale** holds a degree in Agricultural Sciences from the Federal Rural University of Rio de Janeiro (2002), a master's degree in Agricultural Microbiology from the Federal University of Lavras (2005) and a PhD in Agricultural Microbiology from the Federal University of Viçosa (2009), post-doctorate in metagenomics of endophytic fungi at the Ruhr-Universität Bochum, Germany. He is currently Associate Professor D, Level II at the University of Brasília (UnB) and Head of the Department of Phytopathology. He works in undergraduate disciplines of Agronomy courses (Microbiology and Phytopathogenic Micro-organisms); Biology (Mycology) and Environmental Sciences (Microbial Diversity and Biological Collections) and postgraduate disciplines in the Phytopathology (Molecular Techniques) and Microbial Biology (Microbial Ecology) courses at UnB. He has experience in the area of Agronomy and Biology, with emphasis on Agricultural Microbiology, working mainly on the following topics: Biological Nitrogen Fixation, Microbial Ecology, Metagenomics, Next Generation Sequencing (NGS), Yeast Diversity in Brazilian Ecosystems, Molecular Diversity and Characterization of Epiphytic and Endophytic Microorganisms.