BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 3: e37288
doi: 10.3897/biss.3.37288

OPEN ACCESS

Conference Abstract

Taxonomic Gap Analysis: A method of evaluating the taxa represented in biodiversity databases

Amanda Devine[‡], Jonathan Coddington[‡]

[‡] Smithsonian Institution, Washington, DC, United States of America

Corresponding author: Amanda Devine (devinea@si.edu)

Received: 13 Jun 2019 | Published: 21 Jun 2019

Citation: Devine A, Coddington J (2019) Taxonomic Gap Analysis: A method of evaluating the taxa represented in biodiversity databases. Biodiversity Information Science and Standards 3: e37288. https://doi.org/10.3897/biss.3.37288

Abstract

The Global Genome Initiative (GGI) endeavors to collect the Earth's genomic biodiversity, preserve this biodiversity as high quality genetic resources in Global Genome Biodiversity Network (GGBN) affiliated biorepositories, increase knowledge of biodiversity through genetic sequencing, and make resources and knowledge accessible to researchers via the GGBN Data Portal, the Global Catalogue of Microorganisms (GCM), and the National Center for Biotechnology Information (NCBI) GenBank. In GGI's seven year timespan, it is attempting to collect samples from all 9,870 families and half of the 165,683 genera of life on Earth (Roskov et al. 2019). To accomplish this, GGI must synergistically consider the following questions:

1. What life exists?
2. What has already been preserved as physical resources?
3. What is already known from genetic sequencing?
4. How will novel or legacy collections fill the gaps in resources or knowledge?

To answer the first question, GGI has explored the use of taxonomic authorities such as the Global Biodiversity Information Facility (GBIF) Backbone Taxonomy and the Catalogue of Life as taxonomic backbones to variously match taxonomic names and derive complete

lists of extant taxa at each taxonomic rank. To answer the second question, GGI utilizes the GGBN Data Portal API and the GCM website to extract lists of taxonomic names, which are then standardized to a taxonomic backbone. To answer the third question, following the recommendations of Hanner 2009 for identifying high-quality DNA barcode records, GGI employs the NCBI Entrez Programming Utilities to download GenBank records, then standardizes the associated taxa to a taxonomic backbone. Finally, GGI compares lists of taxa found in specific geographic areas or specific legacy collections to determine the amount of taxonomic novelty a new collection may supply. GGI refers to this comparison of taxonomic databases as a taxonomic gap analysis, an assessment of how well a potential collection fills the taxonomic gaps in physical collections and genetic knowledge. A gap analysis performed by GGI in March 2019 shows that 49% of families and 78% of genera still have no representation as either physical samples or genetic information (Table 1). There are substantial gaps to fill in the endeavor to capture the Earth's biodiversity, and taxonomic gap analysis will continue to be a powerful tool to identify the most promising potential new collections.

Table 1.

Taxonomic gap analysis results for all of life on Earth as of 06 March 2019. Total counts were derived from the Catalogue of Life (Roskov et al. 2019). Counts of physical samples were derived from the union of the GGBN Data Portal (GGBN 2011) and the Global Catalogue of Microorganisms (Wu et al. 2013). DNA barcode sequence data counts were downloaded from GenBank (Clark et al. 2015). Percentages represent the percent of the total in that category.

Taxonomic Rank	Total	Physical Samples Only	Sequence Data Only	Physical Samples AND Sequence Data	Missing ("Gaps")
Phylum	101	49 (49%)	3 (3%)	34 (34%)	15 (15%)
Class	339	149 (44%)	7 (2%)	105 (31%)	78 (23%)
Order	1418	460 (32%)	73 (5%)	456 (32%)	429 (30%)
Family	9870	1699 (17%)	697 (7%)	2648 (27%)	4826 (49%)
Genus	165683	10512 (6%)	11641 (7%)	13880 (8%)	129650 (78%)

Keywords

gap analysis, taxonomy, genetic resources, DNA barcodes, tissues

Presenting author

Amanda Devine

Presented at

Biodiversity_Next 2019

References

- Clark K, Karsch-Mizrachi I, Lipman D, Ostell J, Sayers E (2015) GenBank. Nucleic Acids Research 44 https://doi.org/10.1093/nar/gkv1276
- GGBN (2011) The GGBN Data Portal. http://data.ggbn.org/ggbn_portal/. Accessed on: 2019-3-01.
- Hanner R (2009) Data Standards for BARCODE Records in INSDC (BRIs). https://sibarcodenetwork.readthedocs.io/en/latest/barcode_data_standard.html. Accessed on: 2019-4-05.
- Roskov Y, Ower G, Orrell T, Nicolson D, Bailly N, Kirk PM, Bourgoin T, DeWalt RE, Decock W, Nieukerken E, Zarucchi J, Penev L (2019) Species 2000 & ITIS Catalogue of Life, 28 February 2019. www.catalogueoflife.org/col. Accessed on: 2019-4-05.
- Wu L, Sun Q, Sugawara H, Yang S, Zhou Y, McCluskey K, Vasilenko A, Suzuki K, Ohkuma M, Lee Y, Robert V, Ingsriswang S, Guissart F, Philippe D, Ma J (2013) Global catalogue of microorganisms (gcm): a comprehensive database and information retrieval, analysis, and visualization system for microbial resources. BMC Genomics 14 (1): 933. https://doi.org/10.1186/1471-2164-14-933