Title: Discovering Uncharacterized Biological Diversity

Authors[1]**:** Henry L. Bart Jr., Dmitri R. Davydov, J. Andres Lopez, Rahul Warrior

Summmary: We know that Planet Earth is inhabited by roughly two million species because these species are formally described in scientific journals. However, it is estimated that the actual number of living species is at least 5 to10-fold higher, and potentially orders of magnitude more. The Earth is currently losing species at an alarming pace (approaching the level of a major extinction) due to human alteration of Earth's natural environments, and there are concerns that climate change will exacerbate this decline. Describing the Earth's species falls within the domain of taxonomy, a subdiscipline of biology that has itself suffered a significant contraction of expertise due to changing biological priorities. Thus, efforts to comprehend the full extent of global biodiversity face two distinct challenges, accelerated species loss and diminishing taxonomic expertise. This challenge creates an opportunity for a new, re-integrated approach to taxonomy, one that incorporates genomic science (as a means both of assessing how much of the Earth's biodiversity is uncharacterized and placing this diversity within the tree of life), computer science (specifically machine-learning, to expedite the process of phenotype data quantification), and environmental science (as a means of rapidly characterizing new species habitat). This ambitious research program requires engagement of a much broader community of researchers then presently engaged in taxonomy, and provides an opportunity to train a new generation of "re-integrated biologists", including citizen scientists, and students from groups presently underrepresented in taxonomy, genomic, computer and environmental science.

Antecedent question

How can we document uncharacterized biological variation and diversity and use this knowledge to understand how Earth's biological systems have responded to millions of years of adaptive change.

Why is this important? (background, rationale)

Roughly two million species are known to science. Estimates suggest that the actual number of living species is at least 5 to10-fold higher, and potentially orders of magnitude more (Brendan et al, 2017; May 2011; Mora et al 2011). At the same time the world faces a global crisis of loss of biological diversity, traditionally attributed to human-caused habitat destruction, pollution, and ecosystem fragmentation (NSB 1989). In addition, new concerns about present and future biodiversity loss center on climate-change related phenomena (Thomas et al. 2004). The concern is that much of the yet uncharacterized biodiversity will be extinguished before it can be discovered and described.

[1] Order of authors is alphabetical

A parallel concern is the contraction of the discipline of taxonomy over recent decades due to retirement of taxonomic specialists and a shift in academic hiring and graduate training toward more sub-organismal and molecular-level areas of biology. This shift has greatly impeded scholarly work in organismal biology and biodiversity discovery, especially involving poorly known groups such as bacteria, fungi, protists, and marine and terrestrial invertebrates. The problem of diminishing taxonomic expertise was highlighted in a National Science Board report, *Loss of Biological Diversity: A Global Crisis Requiring International Solutions* (NSB 1989). This so called "taxonomic impediment" was also addressed by the NSF program, Partnership for Enhancing Expertise in Taxonomy, which from 1995 to 2007 funded grants that trained new generations of taxonomists for studying poorly known groups of organisms and groups for which there was little expertise (Rodman and Cody 2003). Other, more recent NSF programs -- namely Planetary Biodiversity Inventories, Advancing Revisionary Taxonomy and Systematics, and Poorly Sampled and Unknown Taxa -- similarly aimed to encourage biodiversity discovery and description in poorly known groups and poorly sampled areas of the tree of life.

There have also been efforts to enumerate the global biodiversity of Earth's major environments. One such effort was the Census of Marine Life, concluded in 2010, which involved 2,700 scientists from more than 80 countries, 540 oceanic expeditions, 2,600 scientific papers and 6,000 potential new species. However, despite this project's broad geographic coverage and the breadth of its engagement of scientists in marine biodiversity exploration, only ⅕ (1,200) of the potential new species discovered during the project were described taxonomically. As with many similar inventory projects, the rate of publication of new species descriptions from the Census of Marine Life lagged far behind new species discovery. This is because of the laborious and time intensive process of traditional methods of data gathering, analysis and scholarly publication of taxonomic descriptions. This delay highlights the critical importance of devising novel approaches and organizational strategies that can accelerate taxonomic description and categorization.

This paper describes a vision for a research program that will assess uncharacterized global biodiversity using metagenomics and eDNA methods, and will explore novel, more rapid approaches to describing the discovered diversity taxonomically, using machine learning methods to expedite the pace of data gathering for character analysis and species diagnosis. A key area of innovation for this vision will be the integration of information on phenotypic traits not traditionally used in taxonomy. We envision integrative collaborations between biochemists, microbial biologists, molecular and cellular biologist to devise new, effective and high-throughput methods of characterizing physiological phenotypes that are amenable to taxonomic analyses. This would represent a critical transformation of current practice, which relies largely on morphological and genotypic features, and could be biasing the range of diversity currently cataloged.

IMPORTANT SCIENCE OPPORTUNITIES

- The project would revolutionize taxonomy by applying metagenomics to the process of assessing uncharacterized biodiversity, and machine-learning methods to taxonomic

character evaluation, both of which could radically expedite the pace of taxonomic description.
- It addresses an urgent need to understand and document global biodiversity before it is lost due to habitat destruction and climate change.
- It offers broad perspectives for uncovering functional mechanisms of biological systems (enzymes, signaling networks, etc.) through creating a strong bioinformatics background for analysis of adaptations to extreme environmental conditions.
- An important practical outcome of systematic exploration of biodiversity is creating a robust platform for a broad-range application of a comparative approach to studying the molecular mechanisms of life.
- It will identify new model systems that are applicable to studies that use comparative approaches for exploring developmental and cellular processes.

WHAT'S STOPPING US? (Key Barriers)

1. Past efforts to expedite the pace of taxonomic description, though noble, have been insufficient.
2. Lack of a program of global sampling of specimens and genetic material, orchestrated for documenting uncharacterized biodiversity.
3. Inefficient use of metagenomics to identify the uncharacterized biodiversity and to rapidly assess degrees of genetic divergence among described and undescribed species.
4. Lack of integration of databases of biodiversity and environmental information that would allow taxonomists to rapidly produce species distribution maps and identify critical species habitat.
5. Impediment #4 is of particular importance for exploring structural and mechanistic adaptations of species to extreme habitats. Although the potential of using the studies of adaptations of extremophiles, such as piezophiles, thermophiles or psychrophiles is increasingly recognized (van den Burg 2003; Morozkina et al. 2010; Davydov 2012; Ichiye 2018), its application is hampered by a limited number of characterized species and lack of systematized and properly tagged databases that document genomic information and involves specimen-vouchered records and properly organized taxonomic hierarchies.
6. Need for new computer algorithms designed to aid taxonomists in the process of extracting phenotypic data from specimens.
7. Insufficient efforts to train a new generation of taxonomy students, including students underrepresented in organismal biology, in all of the above methods, i.e., re-integrated biologists.

Strategies and tools to overcome barriers to implement this vision

1. Recruit a community of stakeholders to lead/nurture/steer the broader effort
    a. Identify common objectives and coordinate action by experts within distinct domains
    b. Link with and build on existing related efforts

    c. Coordinate/lead standards for data types, data management and annotation
    d. Lead creation of synthetic, ‘real-time’ platform for disseminating samples, raw data, processed data products, interpreted data products and results

An effort of the scale envisioned here will require continued, engaged participation from a diverse set of investigators across all stages of professional development. A potential mechanism to build this community would be the creation of a Research Coordination Network (RCN), which will have as its central mission a robust design and implementation plan for this project. Within this mission, the RCN will develop training plans that can help prepare students to fruitfully participate in this species discovery effort. The RCN will also ensure that implementation of this vision makes maximal use of past and ongoing efforts related to uncovering and organizing global biodiversity. This objective will have the dual aims of avoiding redundant use of resources and also ensuring that project outcomes can valuably contribute to existing investments. Finally, a key outcome of the RCN is to identify or design models of research dissemination that foster continued diversity discovery activities beyond those directly linked to this vision.

2. New technologies and approaches
    a. High-throughput environmental DNA sequencing involving targeted markers and metagenomics/metabarcoding
    b. Automated and reproducible DNA sequence data processing and analyses
    c. Novel technologies to assay molecular physiological phenotypes (e.g., proteins and their actions)
    d. Community-accepted standards for provisional taxonomic updates (new species discovery and synonymization)
    e. Development and implementation of ‘real time’ indicators/indices of completeness of biodiversity understanding (e.g., rarefaction curves)
    f. New methods for taxonomic discovery by a new generation of taxonomists.
    g. Enhanced, more integrated, biodiversity, genomic and environmental databases (e.g., following standards of genomic resources databases, which require assigning specimen vouchers and taxonomic hierarchies to genomic databases).
3. Synthesis
    a. Refine and standardize existing and authoritative taxonomic knowledge bodies (e.g. WORMS, Catalog of Fishes)
    b. Integrate new data and products with existing taxonomic knowledge through online platforms that maintain near real-time connections between diverse data sources
    c. Create integrated ‘real time’ platform to archive and distribute product from ongoing work and related extensions

High-throughput sequencing, involving DNA metabarcoding and metagenomics will enable us to assign new environmentally and specimen-derived DNA samples to broad taxonomic groups, thereby relating them to described species. The methods have been successfully applied to

microbial species and unicellular algae (Simon et al., 2019, Tan et al. 2019), but are equally applicable to other life forms provided that the DNA samples contain the targeted gene regions.

Machine learning methods involve using human-gathered data, generally from digital analogs of organisms, to train computer algorithms how to perform certain tasks, then testing the algorithms' performance of those tasks on separate samples of digital analogs of the same or related organisms. We envision training algorithms to aid taxonomists from each of the major domains of life, thereby expediting the laborious and time-consuming tasks of gathering phenotypic data from specimens.

What might be the broader impacts?

1. Greater appreciation of planetary diversity and the complexity of global ecosystems.
2. Potential societal impacts of the newly discovered genetic diversity to bioprospecting (material science, biomechanics, commercially valuable compounds).
3. New breakthroughs in drug discovery and developing new approaches in biomedicine enabled by exploring biomolecular mechanisms through comparative studies with newly characterized species.
4. Training of a new generation of taxonomists, skilled in re-integrated, cross-disciplinary methods of biodiversity discovery.

The success of this monumental task is contingent on engaging taxonomists currently involved in describing the earth's species and training a new generation of "re-integrated organismal biologists". One way to accomplish this is by rapidly publishing data products of the project (i.e., genomic sequences, digital analogs of specimens and machine-learning training and test datasets, data papers). This should help to attract established taxonomists who have interests in contributing to species descriptions but lack the resources to participate in the sample collection and other data gathering. Publishing preliminary project results should also attract graduate students who are looking for Master's theses and research topics for dissertations.

Broad engagement can also be accomplished by engaging citizen scientists in the task of training computer algorithms or gathering environmental data from areas where specimens/genomic resources are gathered. Efforts should also be made to engage pre-college students in the project, with a specific focus of urban, pre-college students. A focus on urban recruiting is an excellent opportunity to engage participants from groups underrepresented in Science and Engineering broadly, and in Ecology and Evolutionary Biology in particular. In addition to engagement as citizen-scientists, high school students can be directly involved in the research as summer interns or at weekend science academies. Students who show an interest in the research can be recruited to enroll in campuses where the research is occurring and further supported with summer REU experiences, putting them on a path to graduate enrollment and professional careers.

Implementing this vision will require a high degree of organization, which might best be accomplished in a series of five-year phase projects. The first phase should be a Research

Coordination Network-type effort, involving a series of workshops for engaging relevant parts of the biological and computer science communities. The first workshop could focus on metagenomics approaches, including training of PIs, postdocs and students in sampling methods and bioinformatics. Two additional workshops should engage taxonomists representing each of the Kingdoms (Domains) of Life and appropriate environmental scientists, who would be involved in field explorations. The first of these workshops could be devoted to developing a plan for regions of the world to be sampled and sampling protocols. The second taxonomist workshop could address sample prep and data analysis. A fourth workshop should involve computer scientists working alongside taxonomists representing each of the Kingdoms of life and their students, who would together explore how machines can be engaged in expediting the process of data gathering from specimens and the kinds of digital analogs of specimens that would be used (digitized slides, 2D and 3D representations of specimens. A final workshop could be devoted to promoting dissemination of project results and other broader impacts.

Subsequent five-year phases of the effort would perform the actual work of conducting explorations in unexplored/poorly samples areas of the world, collecting specimens, genomic resources and environmental data, preparing digital analogs of specimens for machine learning experiments, producing genomic data and identifying species, engaging students and citizen scientists in the project.

How does it reintegrate biology?

Discovering uncharacterized biodiversity will reintegrate biology in several ways. One mechanism is by achieving a primary goal - gaining a better understanding of how much organismal complexity currently exists both globally and in defined locations. These data will inform studies focused at the community level by delineating the range of organisms present in particular contexts. The use of genomic approaches as a first step in assaying the extent of diversity provides a link to analysis at the molecular level. Integration of genomic data with locational and environmental information that specifies the origin of samples could be a gateway to identifying molecular signatures associated with particular biomes, habitats and ecoregions. Further, we expect that developing and incorporating physiological trait phenotypes, which will be identified through the expertise of biochemists, and molecular and cell biologists. Because these types of traits have not been commonly used in species discovery, we anticipate fruitful interactions with taxonomic experts. The application of computational approaches to identifying and extracting phenotypic data and making these data accessible combines traditional taxonomic skills and approaches with data management and analysis. We anticipate that involvement of scientists from dramatically different areas of expertise and the exposure of students to scientific approaches that combine multiple levels and types of analyses will help develop a new generation of truly reintegrated biologists.

What disciplines might be needed?

Computer Science (Artificial Intelligence, Machine Learning), Engineering (Materials Science,Technology Development), environmental science, biomolecular science, Biological Taxonomy representing all of the kingdoms of Life (Animalia, Plantae, Fungi, Protista, Archaea/Archaebacteria, and Bacteria/Eubacteria), and all other areas of biology.

Intended audience of the paper

All of biology, computer and environmental science to engage participants from these communities in launching the effort.

References

Brendan B. Larsen, Elizabeth C. Miller, Matthew K. Rhodes, John J. Wiens. Inordinate Fondness Multiplied and Redistributed: the Number of Species on Earth and the New Pie of Life. *The Quarterly Review of Biology*, 2017; 92 (3): 229 DOI: 10.1086/693564

Davydov, D. R. (2012). "Merging Thermodynamics and Evolution: How the Studies of High-Pressure Adaptation may Help to Understand Enzymatic Mechanisms." J. Thermodynam. Cat. **3**(4): 1000e1110.

Ichiye, T. (2018). "Enzymes from piezophiles." Seminars in Cell & Developmental Biology **84**: 138-146.

May RM (2011) Why Worry about How Many Species and Their Loss? PLoS Biol 9(8): e1001130. doi:10.1371/journal.pbio.1001130

Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B (2011) How Many Species Are There on Earth and in the Ocean? PLoS Biol 9(8): e1001127. doi:10.1371/journal.pbio.1001127

Morozkina, E. V., E. S. Slutskaya, et al. (2010). "Extremophilic Microorganisms: Biochemical Adaptation and Biotechnological Application (Review)." Applied Biochemistry and Microbiology **46**(1): 1–14.

NSB (NATIONAL SCIENCE BOARD). 1989. Loss of biological diversity: A global crisis requiring international solutions. Report NSB 89-171. National Science Foundation, Washington.

Simon H. Ye, Katherine J. Siddle, Daniel J. Park, Pardis C. Sabeti.2019. Benchmarking Metagenomics Tools for Taxonomic Classification.
DOI:https://doi.org/10.1016/j.cell.2019.07.010

Tan, S-M, P. Yi, M. Yung, P. E. Hutchinson, C. Xie, G. H. Teo, M. H. Ismail, D. I. Drautz-Moses, P. F. R Little, R. B. H. Williams, Y. Cohen. 2019. Primer-free FISH probes from metagenomics/metatranscriptomics data permit the study of uncharacterised taxa in complex

microbial communities. NPJ Biofilms and Microbiomes (2019) 5:17, https://doi.org/10.1038/s41522-019-0090-9

Thomas, C.D., Cameron, A., Green, R.E. et al. (16 more authors) (2004) Extinction risk from climate change. Nature, 427 (6970). pp. 145-148.

van den Burg, B. (2003). "Extremophiles as a source for novel enzymes." Current Opinion in Microbiology **6**(3): 213-218.