McIlvainea

JOURNAL OF AMERICAN AMATEUR MYCOLOGY

Fibroporia radiculosa, First Landing State Park, Virginia Beach, VA, December 29, 2016, photographed by Tom Bigelow of NYMS

Volume 31 - 2022

A Publication of the North American Mycological Association

DNA Barcode Identification of Macrofungi by Community Scientists

Emily Cantonwine*, Allison Arnold*, Todd Osmundson† Bill Sheehan‡ and Stephen Russell¶

*Department of Biology, Valdosta State University, Valdosta, GA 31698, USA, egcantonwine@valdosta.edu, aharnold@valdosta.edu

†Department of Biology, University of Wisconsin-La Crosse, La Crosse, WI 54601, USA, tosmundson@uwlax.edu

‡ bill@productpolicy.org

¶ steve@hoosiermushrooms.org

ABSTRACT

This paper reviews the scientific contributions of early barcode sequencing efforts of projects with the Fungal Diversity Survey (FunDiS). Thirty-five projects, seventeen that included a person with relevant expertise (advanced) and eighteen that did not (standard), were included in the study. Of the 1990 sequences evaluated, 241 were novel to GenBank, 412 were uncommon, 869 were common and 468 were very common. "Barcode" identification occurred for 36.7% of sequences with at least one reference match at the species-level, and 88.3% at the genus-level. Thirty-one novel sequences corresponded to one or more environmental DNA references, making the associated fruiting body collections particularly valuable. Advanced projects generated a higher percentage of novel sequences, 16% more than standard projects, 9% (P=0.04), with two NAMA forays and an active NAMA club generating >25% novel sequences. Of the 282 novel and uncommon sequences with a species identifier, thirty-seven were the first curated collection within the state of collection based on the MyCoPortal georeferenced database. These results demonstrate that collections made by amateur mycology groups in North America are likely to generate sequences of scientific value. Consultation with a taxonomic expert is recommended for groups wanting to contribute to the DNA barcoding database.

Keywords: Citizen Science, Amateur mycologist, DNA Barcoding, ITS sequencing, Fungal Diversity Survey, FunDiS, North American Mycoflora Project, NAMP

INTRODUCTION

In 2018, the North American Mycoflora Project, Inc., now known as Fungal Diversity Survey (FunDiS), was founded to support community scientist efforts to contribute high-quality vouchers of macrofungal fruiting bodies, collection data, and DNA sequences to help professional mycologists establish a complete list of fungal species for North America (Sheehan 2017, Thiers and Halling 2018). By the end of the first year of the program, there were approximately 100 community-directed projects registered. Most were individuals, clubs, or events affiliated with mushroom organizations, such as the North American Mycological Association (NAMA). Although many included someone with a high level of taxonomic expertise, or in some cases, molecular biology expertise, there were just as many projects led by amateurs with little experience in these areas. According to project registration surveys, nearly every project indicated a desire to document the fungal biodiversity of their region as a primary project objective; other objectives noted included contributing to taxonomic and systematic mycology (40%), using DNA barcoding to improve mushroom identification skills and/or to boost club participation (25%), using FunDiS as a model to teach science (6%), building collaborations with professional mycologists (4%), or just to have fun (1%).

Although FunDiS has recently narrowed its focus from documenting all macrofungi to documenting fungi of conservational interest, the initial effort of the program generated a wealth of information about how amateur mycology groups can incorporate DNA sequencing of general fungal collections for scientific impact. This paper reviews the scientific contributions of early FunDiS efforts and provides a list of lessons learned for groups interested in continuing the sequencing effort.

METHODS AND MATERIALS

Projects. All FunDiS projects with at least 17 internal transcribed spacer (ITS) sequences by June 2019 were included in the study. Of the thirty-five projects that met these criteria, most had an affiliation with amateur clubs associated with NAMA, nine of which were collections from NAMA forays or similar events. Registration survey information was used to classify projects by expertise

as follows. “Advanced” status was assigned to projects led or advised by mycologists (amateur or professional) with well-established taxonomic expertise or sequencing experience. Seventeen projects fit this category. The remaining projects were classified as “standard”. Eleven projects were located in the Northeast Region of the United States, 4 in the Midwest, 11 in the West, and 7 in the South Central Region (US Census Bureau, 2018). One project was located in Canada and one was specific to a taxonomic group rather than a location.

DNA Sequencing. DNA barcode sequences of the ITS region (Schoch et al. 2012) were generated using Sanger sequencing. DNA extraction and PCR amplification were conducted at one of three university laboratories (Purdue University, University of Wisconsin – La Crosse, or Duke University) and sequencing was conducted by Eurofins Genomics (Lexington, KY, USA). Laboratory protocols followed those described by Dirks & Russell (2020), or similar. Sequence files in FASTA format were stored in MycoMap, a data portal created and managed for community scientists (https://mycomap.com), and downloaded for use in this study.

For projects with fewer than 100 sequences, all sequences were evaluated; otherwise, 100-200 sequences were arbitrarily selected as representative samples. The mean number of sequences per project was 57.

Data Collection. Sequences were compared to the National Center for Biotechnology Information (NCBI) GenBank reference database using a standard nucleotide BLAST (megablast) search that excluded uncultured and environmental sequences. Reference sequences with ≥ 97% percent identity and ≥1000 Bit-score were accepted as species-level matches (Kõljalg et al. 2013). As a quality check, taxonomic similarities between the names associated with the FunDiS and reference sequences were compared, and a handful (~1%) of questionable FunDiS sequences – i.e., those belonging to morphologically dissimilar taxa – were removed from the study due to possible contamination or mislabeling of the sequence.

Successful, species “barcode” identifications were noted for FunDiS sequences when all reference matches (≥ 97% percent identify and ≥1000 Bit-score) had consistent taxon associations; i.e., the same specific epithet with or without cf., aff., or quotes after merging synonyms. All others were classified as having more than one species-level “barcode” identification, or a higher-level (family level or higher) “barcode” identification, when reference taxa belonged to multiple genera or lacked species-level classifications. “Barcode” is used here in quotation marks because the GenBank database includes references that do not meet sequence quality and/or specimen data standards for a

formal DNA barcode as defined by the Consortium for the Barcode of Life (Hanner 2009, Schoch et al. 2014).

Each FunDiS sequence was classified as novel, uncommon, common, or very common based on the number of GenBank sequences that fit the species-level criteria (Table 1). Novel sequences were reassessed in GenBank with a second BLAST search, this time including uncultured and environmental sequences to look for matches to environmental DNA (eDNA), which are sequences generated from DNA extracted from vegetative tissues or substrates, e.g., leaves, wood, or soils, rather than fruiting bodies. Novel sequences were also screened against the UNITE database, a curated fungal ITS sequence database (Abarenkov et al. 2010, Nilsson et al. 2019), to further assess the unique contribution of these sequences to fungal DNA databases. Upon confirmation, novel sequences were classified by family using the NCBI Taxonomic Browser (Schoch et al. 2020). Novel and uncommon sequences with a species-level identifier were screened against collections of curated Fungaria using MyCoPortal (MyCoPortal 2021) to estimate the percentage of these sequences that were first reports for the US state in which the specimen was collected.

Table 1. Criteria for classification of FunDiS sequences by existing representation in the GenBank reference database.

Category	Reference match (#) [a]	Percent Identity [b]	Bit-score [c]
Novel	0	≥97%	<1000
Uncommon	1-3	≥97%	≥1000
Common	>3 <10	≥97% ≥99%	≥1000 ≥1000
Very Common	≥10	≥99%	≥1000

[a] Number of sequences within the NCBI GenBank database that meet the criteria listed under Percent Identity and Bit-score

[b] percentage of nucleotide bases identical between the compared portions of the query (FunDiS) and reference sequence(s)

[c] a normalized statistic that weighs percent identity and sequence overlap between query and database sequences

Data Analysis. The number of sequences within each "barcode" identification classification and number of sequences matching only eDNA were tallied for all projects combined. Percentages of novel, uncommon, common, and very common sequences were computed for projects separately.

The effect of expertise on the proportion of sequences belonging to each commonness category was evaluated using univariate generalized linear models (GLMs) in the SPSS software package. Regional effects were tested using the same procedure with expertise included as a covariate. Dependent variables that were not normally distributed were natural log transformed prior to analysis. Differences were accepted as statistically significant when $P<0.05$ ($\alpha=0.05$).

RESULTS

Estimated Success Rate of Barcode Species Identification. Out of 1990 sequences evaluated, 1749 had at least one match in the reference library at the minimum species-level criterion. Of these, 36.7% were "barcoded" to a single species, 27.5% to 2 or 3 species within the same genus, and 24.1% to 4 or more species within the same genus. The remaining sequences were "barcoded" to family- or order-level (11.7%).

Sequence Contributions to GenBank. For sequences with at least one reference match (≥97% identify and ≥1000 Bit-score), 412 were classified as uncommon, 869 as common, and 468 as very common. The remaining 241 were novel to GenBank. Disregarding 16 novel sequences that lacked taxonomic classification beyond order, novel sequences came from 37 fungal families within the Basidiomycota. Families with 3 or more novel sequences were Agaricaceae (10), Amanitaceae (32), Bolbitaceae (3), Boletaceae (12), Cantharellaceae (3), Clavariaceae (4), Cortinariaceae (9), Crepidotaceae (4), Entolomataceae (3), Gomphaceae (9), Hygrophoraceae (4), Inocybaceae (10), Marasmiaceae (4), Mycenaceae (5), Polyporaceae (5), Russulaceae (15), Strophariaceae (5), Thelophoraceae (4), and Tricholomataceae (11). Thirty-one novel sequences matched at least one eDNA reference (Table 2). Fourteen came from sequenced collections from NAMA Forays, 6 from other advanced projects, and 11 from standard projects.

Table 2. Number of novel sequences grouped by family matching one or more eDNA references.

FAMILY	# Sequences	# eDNA matches
Amanitaceae	1	8
Boletaceae	4	4
Cantherellaceae	1	3

Cortinariaceae	1	12
Entolomataceae	1	32
Gomphaceae	2	2
Hygrophoraceae	2	3
Inocybaceae	5	105
Mycenaceae	2	93
Nidulariaceae	1	2
Russulaceae	6	77
Tricholomataceae	2	83
Unclassified	3	59

Of the 282 novel and uncommon sequences that included a species identifier (43% of total), 245 had 11 or more curated fungarium collections within the state of collection, 29 had one or two, and 37 had none. Families with representatives that were not previously reported within the same US state are presented in Table 3.

Table 3. Sequenced collections identified to species that lack previously reported curated fungarium collections for the US state from which the FunDiS specimen was collected.

Phylum	Order	Family	# Collections	State
Ascomycota	Helotiales	Pezizellaceae	1	NY
	Hypocreales	Hypocreaceae	2	NY, OR
Basidiomycota	Agaricales	Agaricaceae	2	NY
		Amanitaceae	3	MS, WV
		Cortinariaceae	3	MO, NY, WA
		Crepidotaceae	2	CA, PA
		Entolomataceae	2	GA, WA
		Hygrophoraceae	2	GA, MO
		Inocybaceae	2	CA
		Mycenaceae	2	NY
		Physalacriaceae	1	NY
		Psathyrellaceae	1	NY
		Tricholomataceae	3	FL, MO, NY
		Tubariaceae	1	MS
	Boletales	Boletaceae	2	MS
	Corticiales	Corticiaceae	1	NY
	Dacrymycetales	Dacrymycetaceae	1	NY
	Polyporales	Irpiaceae	1	NY
		Podoscyphaceae	1	MS

		Polyporaceae	1	WV
	Sebacinales	Sebacinaceae	1	NJ
	Stereopsidales	Stereopsidaceae	1	WV
	Thelephorales	Bankeraceae	1	PA

Project Comparisons by Expertise and Location. The percentage of novel sequences by project ranged from 0-37%, with advanced projects having a higher mean percentage of novel sequences than standard projects (P=0.04) (Figure 1). Expertise did not affect the percentages of uncommon (P=0.93), common (P=0.62), or very common (P=0.25) sequences. There were no differences observed for the proportions of novel (P=0.43), uncommon (P=0.36), common (P=0.97) or very common (P=0.17) sequences across regions (Figure 2). Expertise was not a significant cofactor in the regional analyses of sequence category proportions (P≥0.18).

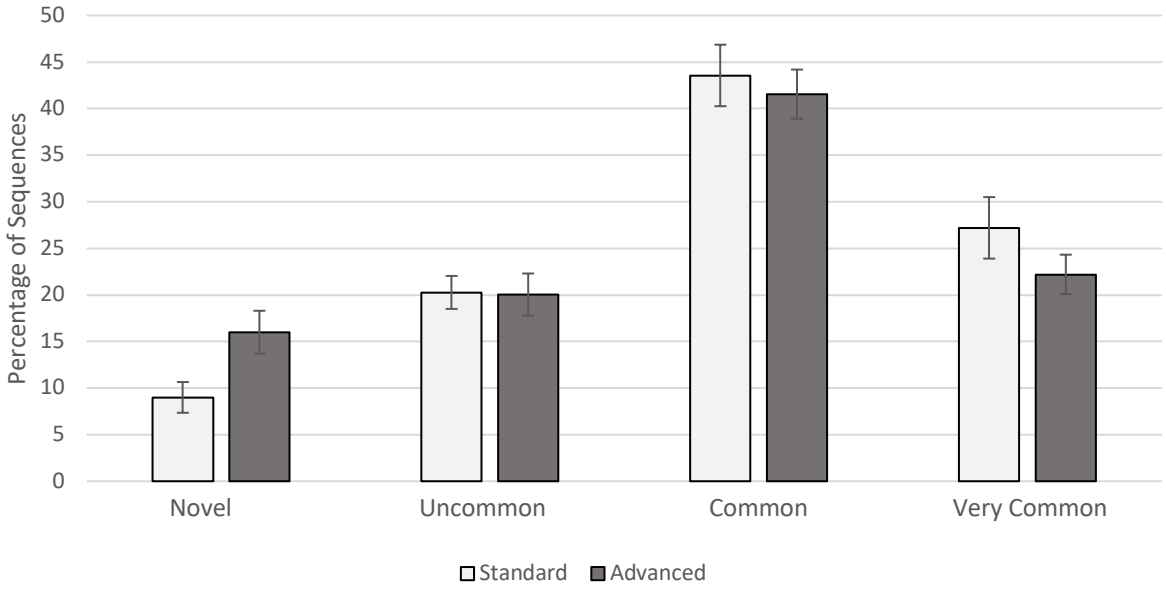

Figure 1. Mean percentages and standard error bars of sequences assigned to each commonness category for the standard and advanced projects.

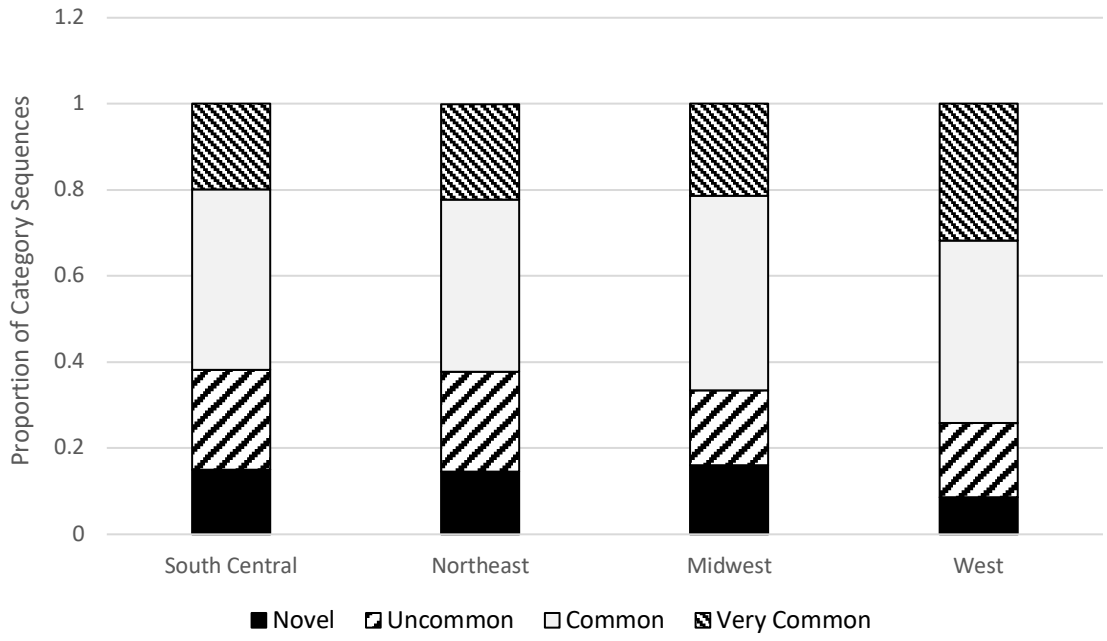

Figure 2. Mean proportions of sequences classified as novel, uncommon, common, or very common by Region.

DISCUSSION

Based on these results, community scientists in North America can expect to retrieve a species-level "barcode" identification for 1 of 3 macrofungal collections. "Barcode" identification to genus is expected for 8 of 10 collections (77.6% for all sequences and 88.3% for those with at least one match). These rates are slightly better than what might be expected based on the current number of barcoded basidiomycete species, 13,016 (BOLD data system 2021) and estimated number of basidiomycete species, ~54,700 (MycoBank 2021)[1]. Reasons for these barcoding holes include limited sequencing resources, unclear species delineations, and for some taxa, little ITS variability. Barcoding errors due to inaccurate taxonomic annotations of sequences in the reference databases can also occur (Nilsson et al. 2006, Bruns et al. 2008, Osmundson et al. 2013, Hofstetter et al. 2019, Lücking et al. 2020).

[1] The list of taxa present in MycoBank (June 8, 2021 version) was downloaded from https://www.mycobank.org and converted from Excel to comma separated (CSV) format. The statistical package, R, was used to filter the data to retain only taxa belonging to Class Agaricomycetes (Phylum Basidiomycota) and count the number of unique entries in the "Current name" field. A total of 54,739 unique current names were obtained.

For the 40% of projects that hoped to contribute to taxonomic and systematic knowledge, increasing the proportion of collections with unique ITS sequences is likely a higher priority than barcoding for species identification. DNA barcoding of well-known groups such as chanterelles (*Cantharellus* spp.; Foltz et al. 2013) and elfin saddles (*Helvella lacunosa* complex; Nguyen et al. 2013) has led to the discovery of previously undescribed diversity due to cryptic speciation. Therefore, even for previously described species, novel sequences with accurate species identifiers will add to reference databases and improve barcode species identification rates over time.

The greater frequency of novel sequences from the advanced projects suggests that having someone with experience involved improves the scientific interest of the collections and/or selection of materials to be sequenced. Although the nature of involvement by professionals and highly experienced amateurs was not reported, the three projects with >25% unique sequences provide some clues. Two were associated with NAMA Forays, one with 140 sequences and the other with 19 sequences, and one with a NAMA club that hosts numerous collection events per year, which had 27 sequences. Based on the nature of these projects, it is more likely that sequenced materials were selected from a larger set of collections than otherwise.

It is important to point out that standard groups still contributed novel sequences at a respectably high proportion (~9% on average). For projects without access to an expert, comparing collections to State and county lists, generated with MyCoPortal, Mushroom Observer or similar databases, and to the species barcode list reported by BOLD, may provide comparable guidance during the sequence selection process.

The 31 novel sequences that matched eDNA references (1.5% of all sequences evaluated) are particularly valuable, as they are the first publicly available sequences to link eDNA to a fruiting body. Six of the novel sequences that matched eDNA had over 30 matches, suggesting that some of these collections were frequently encountered and/or widely distributed fungi that had yet to be added to the reference library. Nearly half of these sequences came from NAMA forays, which again highlights the value of this sort of concentrated collecting event for improving biodiversity information. High quality vouchers of the collections are critical to capture the full scientific potential of these 31 eDNA reference matches, as well as the additional novel sequences.

Geographic region did not correspond to differences in commonness categories. Sample size was relatively low, so little can be concluded from this result; however, it does support others who have noted that even in highly collected areas, more knowledge is needed (Haelewaters et al. 2018).

Lastly, although this study did not directly assess the contributions of all FunDiS collections to improving knowledge of species distributions (reported in project registration surveys as the primary goal of most projects), distributions for unique and uncommon sequences with a species identifier – i.e., taxa deemed most likely to be infrequently collected or regionally limited – were investigated to estimate this contribution. At least 15% of unique and uncommon sequences with a species identifier – more than 2.2% of all sequenced collections in the study -- appear to be the first reports for the state of collection. The actual number is likely higher, as only half of the collections with uncommon or novel sequences included a species identifier, and none of the common or very common collections were screened. It should also be noted that quantification within state boundaries does not capture new distributional knowledge at scales that may be more biologically relevant (e.g., forest types, watersheds, or other habitat characteristics) or more important from management or conservation perspectives (counties or municipalities, parks or natural areas, threatened habitats, etc.).

CONCLUSION

The results of this study highlight the value of community-led collection and sequencing efforts and show that many of the goals reported by amateurs are within the range of possibility. Community scientists can help to populate reference barcode databases, and when accompanied by high quality vouchered collections, can extend species ranges, clarify taxonomic issues, and improve environmental biodiversity assessments. Expert involvement, particularly with the NAMA Foray model, can improve scientific outcomes, but depending on the project goals, may not be necessary.

Acknowlegements

The authors thank the following FunDiS projects whose sequenced collections were evaluated in this study: Amanitas of North America, Amanitas of Northeast Pennsylvania, Cumberland Plateau Mycoflora (TN), Eastern Penn Mycoflora Project, Fungi of New York City, Hygrophoraceae of Lower Puget Sound (WA), *Inocybe* of Quebec, Macrofungi of Lane County (OR), Michigan Mushroom Hunters Club Mycoflora, Mycoflora at Valdosta State University, Mycoflora of Greater Rochester (NY), Mycoflora of Indiana, Mycoflora of Kentucky, Mycoflora of Long Island (NY), Mycoflora of Missouri, Mycoflora of New Jersey, Mycoflora of Northern Arizona, Mycoflora of

Pennsylvania, Mycoflora of the Appalachia Region, Mycoflora of the Greater San Francisco Bay Area, Mycoflora of Vashon Island (WA), Northeast Georgia Mycoflora Project, Pioneer Valley Mycoflora (MA), PNW *Cortinarius* Study, South Sound Mycoflora Project (WA), Western New York State Mycoflora, 2014 NAMA Foray (Eatonville, WA), 2015 NAMA Foray (Black Mtn, NC), 2018 NAMA Foray (Salem, OR), 2018 Regional NAMA Foray (Canton, MS), 2018 Northeast Mycological Federation (NEMF) New York Foray, 2018 Boletaceae of NEMF Foray, 2018 A.H. Smith Foray (WI), 2018 Radical Mycology Convergence Foray, and 2018 Santa Cruz Mycoflora Foray (CA).

Additional thanks to Alejandro Rojas, Brian Looney, Rytas Vilgalys, and Catherine Aime for their support with sequencing, and Valdosta State University undergraduate students Allison Arnold, Daniel Akuoko, Courtney Taylor, Angela Rangel, Marlie Kolb and Arvin Berry who helped with data collection. Funding sources include the VSU COSM Research Scholar Award, and sequencing awards granted by FunDiS with contributions from the North American Mycological Association, the Mycological Society of America, and numerous individual supporters.

REFERENCES

Abarenkov, Kessy, R. Henrik Nilsson, Karl-Henrik Larsson, Ian J. Alexander, Ursula Eberhardt, Susanne Erland, Klaus Høiland et al. 2010. "The UNITE Database for Molecular Identification of Fungi–Recent Updates and Future Perspectives." *New Phytologist* 186, no. 2: 281-285.

BOLD: The Barcode of Life Data System. “Taxonomy Browser: Basidiomycota.” Accessed Nov 11, 2021. http://www.boldsystems.org/index.php/Taxbrowser_Taxonpage?taxid=23675.

Bruns, Thomas D., Meredith Blackwell, Ivan Edwards, Andy FS Taylor, Thomas Horton, Ning Zhang, Urmas Kõljalg et al. 2008. "Preserving Accuracy in GenBank." *Science* 319, no. 5870: 1616.

Dirks, Alden, and Stephen D. Russell. 2018. "DNA Barcoding of Macrofungi from the 2018 Smith Foray: New Fungal Records for Wisconsin and the United States of America." *Ann Arbor* 1001: 48109.

Foltz, Matthew J., Kathryn E. Perez, and Thomas J. Volk. 2013. “ Molecular Phylogeny and Morphology Reveal Three New Species of *Cantharellus* within 20 m of One Another in Western Wisconsin, USA”. *Mycologia* 105 :447–461.

Haelewaters, Danny, Alden C. Dirks, Lara A. Kappler, James K. Mitchell, Luis Quijada, Roo Vandegrift, Bart Buyck, and Donald H. Pfister. 2018. "A Preliminary Checklist of Fungi at the Boston Harbor Islands." *Northeastern Naturalist* 25(Sp9): 45-76.

Hanner, Robert. 2012. "Data Standards for BARCODE Records in INSDC (BRIs)." Accessed Dec 16, 2021. https://www.boldsystems.org/docs/dwg_data_standards-final.pdf

Hofstetter, Valérie, Bart Buyck, Guillaume Eyssartier, Sylvain Schnee, and Katia Gindro. 2019. "The Unbearable Lightness of Sequenced-Based Identification." *Fungal Diversity* 96, no. 1: 243-284.

Kõljalg, Urmas, R. Henrik Nilsson, Kessy Abarenkov, Leho Tedersoo, Andy FS Taylor, Mohammad Bahram, Scott T. Bates et al. 2013. "Towards a Unified Paradigm for Sequence-Based Identification of Fungi." *Molecular Ecology* 22 no. 21: 5271-5277.

Lücking, Robert, M. Catherine Aime, Barbara Robbertse, et al. 2020. “Unambiguous Identification of Fungi: Where do we Stand and how Accurate and Precise is Fungal DNA Barcoding?” *IMA Fungus* 11, no. 14. https://doi.org/10.1186/s43008-020-00033-z.

MycoBank Database. “Fungal Databases, Nomenclature & Species Bank”. Accessed January 2, 2021. https://www.mycobank.org.

MyCoPortal. “Mycology Collections Portal” Accessed June 06, 2021. https://mycoportal.org/portal/index.php.

Nguyen, Nhu H., Fidel Landeros, Roberto Garibay-Orijel, Karen Hansen, and Else C. Vellinga. 2013. “The *Helvella lacunosa* Species Complex in Western North America: Cryptic Species, Misapplied Names and Parasites”. *Mycologia* 105:1275–1286.

Nilsson, R. Henrik, Martin Ryberg, Erik Kristiansson, Kessy Abarenkov, Karl-Henrik Larsson, and Urmas Kõljalg. 2006. "Taxonomic Reliability of DNA Sequences in Public Sequence Databases: a Fungal Perspective." *PloS one* 1, no. 1: e59.

Nilsson, Rolf Henrik, Karl-Henrik Larsson, Andy F. S. Taylor, Johan Bengtsson-Palme, Thomas S. Jeppesen, Dmitry Schigel, Peter Kennedy et al. 2019. "The UNITE Database for Molecular Identification of Fungi: Handling Dark Taxa and Parallel Taxonomic Classifications." *Nucleic acids research* 47, no. D1: D259-D264.

Osmundson, Todd W., Vincent A. Robert, Conrad L. Schoch, Lydia J. Baker, Amy Smith, Giovanni Robich, Luca Mizzan, and Matteo M. Garbelotto. 2013. "Filling Gaps in Biodiversity Knowledge for Macrofungi: Contributions and Assessment of an Herbarium Collection DNA Barcode Sequencing Project." *PloS one* 8, no. 4: e62419.

Schoch, Conrad L., Keith A. Seifert, Sabine Huhndorf, Vincent Robert, John L. Spouge, C. André Levesque, Wen Chen, and Fungal Barcoding Consortium. 2012. "Nuclear Ribosomal Internal Transcribed Spacer (ITS) Region as a Universal DNA Barcode Marker for Fungi." *Proceedings of the National Academy of Sciences* 109, no. 16: 6241-6246.

Schoch, Conrad L., Stacy Ciufo, Mikhail Domrachev, Carol L. Hotton, Sivakumar Kannan, Rogneda Khovanskaya, Detlef Leipe et al. 2020. "NCBI Taxonomy: a Comprehensive Update on Curation, Resources and Tools." *Database* 2020.

Sheehan, Bill. 2017. "Mushroom Citizen Science in the USA: From Species Lists to Mycoflora 2.0." *Fungi Magazine* 101: 28-36.

Thiers, Barbara M., and Roy E. Halling. 2018. "The Macrofungi Collection Consortium." *Applications in Plant Sciences* 6, no. 2: e1021.

US Census Bureau. "Statistical groupings of states and counties." (2018).