Meetings

Towards the unification of sequence-based classification and sequence-based identification of host-associated microorganisms

Sequenced-based classification of fungi, a workshop held at the Mycological Society of America Meeting, East Lansing, Michigan, USA, June 2014

Plants interact with a wide assortment of microbial organisms – taking the role of pathogens, mutualists, and commensals. Our knowledge of plant-associated microorganisms has traditionally been based on macroscopic and microscopic structures. In recent decades, the use of both DNA and RNA sequence data derived directly from the environment has been used to study both the taxonomic and functional diversity of host-associated microorganisms. More recently, an explosion of data derived from a shift in nucleotide sequencing technologies has revealed an astonishing diversity of microorganisms (Hibbett & Taylor, 2013). To elucidate taxonomic and functional microbial diversity, researchers employ distinct but not mutually exclusive techniques when using molecular data – Sequence-based Classification (SBC) and Sequence-based Identification (SBI). Those who utilize SBC are predominantly concerned with the discovery and categorization of microbial organisms on the basis of phylogenetic relationships. Researchers who engage in SBI utilize databases as references, often using similarity-based (as opposed to phylogeny-based) approaches, to taxonomically and/or functionally identify the composition of microbial communities. Together, SBC and SBI encompass a range of activities using sequence data – predominantly from nucleic acids – to identify, describe, and functionally characterize microorganisms from the plant-based environment. Marker-based and metagenomic studies, in particular, have sequenced nucleotides from thousands to millions of unidentified species and underscore the need for resources developed for taxonomic and functional characterization of microbial diversity (Hibbett *et al.*, 2011). Perhaps most importantly, new analysis techniques and resources need to integrate with existing taxonomic and systematic knowledge that is based traditionally on cultures and type-material (Lindahl *et al.*, 2013). There is a dire need to develop unified community-based resources and analysis standards for the integration of SBC and SBI of fungi and other microorganisms.

To address the challenges and best practices for SBC and SBI, a group of mycologists met after the annual meeting of the Mycological Society of America (see Kennedy & Stajich, 2014, in this issue of *New Phytologist,* pp. 23–26) for a 2-d workshop supported by the US National Science Foundation. The key aims were: (1) to identify the potential benefits and challenges of merging SBC and SBI; (2) to assess the current strengths and limitations of resources for SBC and SBI; and (3), to consider changes in nomenclatural practices that could promote the integration of traditional specimen-based identification and classification with sequence based methods. The workshop specifically focused on the use of both SBC and SBI for fungi and organisms traditionally studied by mycologists – oomycetes, slime molds, etc. – but an overarching theme could be translated to all microorganisms – including bacteria, archaea, and metazoans – and the desire to characterize microbial interactions with plants, additionally, the issues presented here apply equally to all environmental sequences (e.g. animal and human microbiomes, soil fungi, microbes of the built environment, etc.).

'. . . emphasis needs to be placed on the community involvement needed to encourage researchers to participate in open, accurate data deposition and to incentivize standardization across many resources . . .'

Technical and social challenges of integrating SBC and SBI

It is possible to visualize the optimal unification of SBC and SBI for plant-associated microorganisms in a research workflow. Ideally, a researcher would first extract nucleic acids (or another source of sequence based information, such as proteins) from a plant-based environmental sample. Data derived from either primer-based marker-selection or whole-genome-shotgun sequencing techniques would then be compared to a database of known and unknown sequences. The end result would include a list of taxa and/or genes with their putative functions with information on phylogenetic position, distribution, abundance, ecology, and biochemistry derived from experimental and sample metadata (Fig. 1). This workflow would become more accurate and robust as databases evolve to reflect more comprehensive representations, with richer information concerning taxonomy and functional properties of gene products. Perhaps simple in theory, achieving this workflow is daunting. The meeting participants identified numerous challenges described later, including the development of

New Phytologist

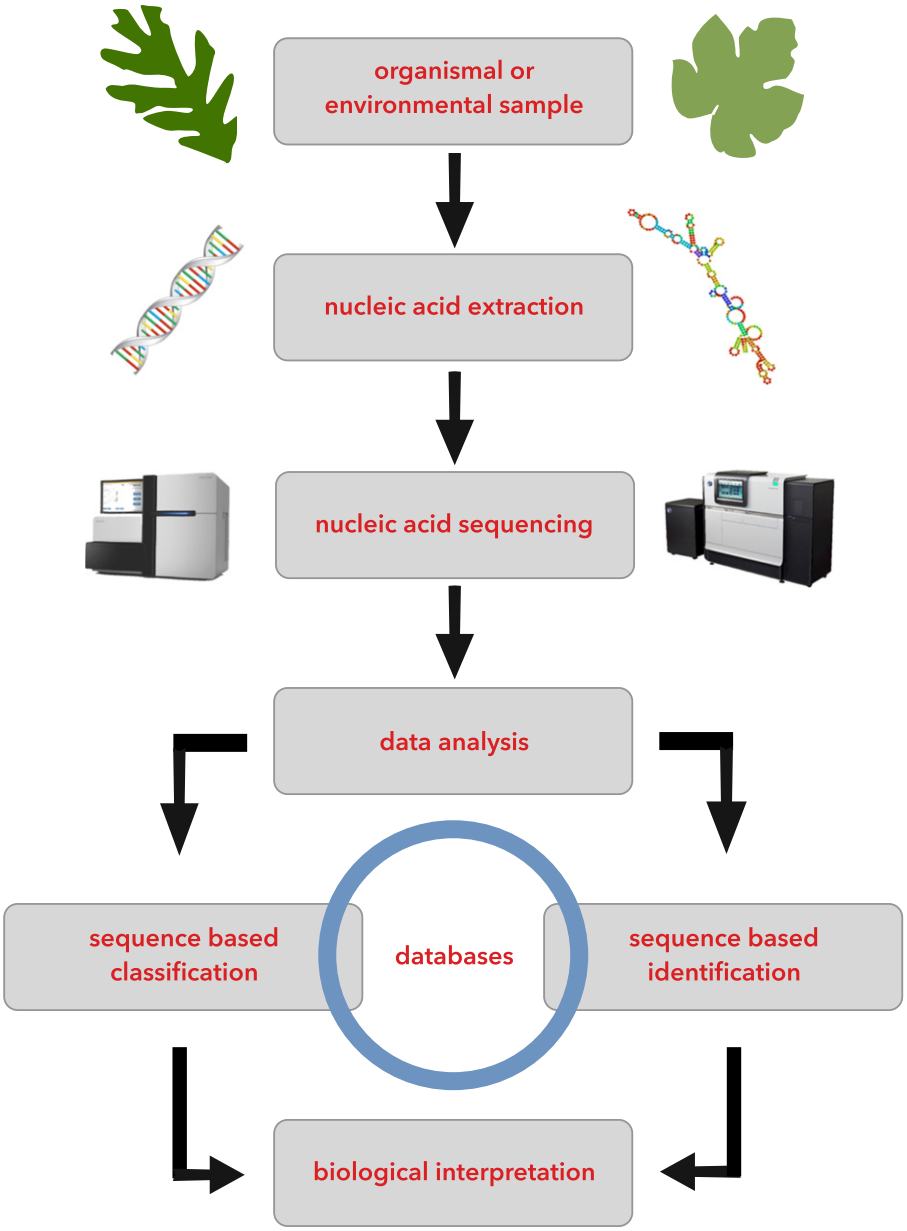

Fig. 1 Simplified workflow for the identification of biological sequence information from nucleic acids derived from an organismal or environmental sample. Please see Lindahl *et al.* (2013) for more detail on accepted data analysis protocols. Unification of sequence based classification and sequence based identification occurs physically through databases used to identify taxonomic and functionally informative sequence information.

standards, the creation and curation of databases, the linking of data and metadata, the promotion of reproducible science, establishment of best-practices, and the promotion of cultural changes rewarding those who contribute to database development and maintenance.

We are only as good as our databases

A central challenge to the unification of SBC and SBI is the development of appropriate nucleic acid sequence databases, including those devoted to the well-established ribosomal operon and promoting its integration with emerging genomic data. The International Nucleotide Sequence Database Collaboration (INSDC) has long served as the main repository for sequence data produced by the entire biological community, and it is one of the greatest successes of publicly supported science. Several excellent independent databases that largely draw on the INSDC have been created (RDP, Cole *et al.*, 2013; SILVA, Quast *et al.*, 2012; GreenGenes, McDonald *et al.*, 2011; UNITE, Abarenkov *et al.*, 2010; Maarj*AM,* Öpik *et al.*, 2010; etc.) and have been growing to accommodate community needs. All databases must be prepared for dramatic growth as new 'higher' throughput sequencing technologies are introduced. Databases increase in value as they increase in size, but large databases require resources to be maintained and may be cumbersome to query. Taxonomic assignment of sequence data deposited in the INSDC is the responsibility of those submitting the data, and third-party annotation is not possible, consequently, there is a crippling mass of misidentified sequences in the database (Bridge *et al.*, 2003). Unidentified sequences from environmental samples are also

flooding sequence databases (Hibbett *et al.*, 2011; Hibbett & Taylor, 2013). The UNITE database now facilitates annotation of sequences grouped into 'species hypotheses' (Abarenkov *et al.*, 2010), but such curation requires experts to donate their time. Specimens, cultures, and raw material, which may include host organism or environmental sample – essentially type materials – must also be maintained to fully support and complement the nucleic acid databases. The maintenance and activity of these additional resources should be placed with a high priority and their integration to existing nucleotide databases should be paramount.

Connecting data to metadata

Yet another challenge will be to link sample metadata to existing nucleotide sequence databases. Metadata, in this case, would be features of the environment that yielded the data or phenotypic data associated with a collected specimen, culture, or host organism (McDonald *et al.*, 2012). Anyone who has used the INSDC's Nucleotide database, Short Read Archive, or other popular data repositories will be unfortunately aware that there is a great deal of inconsistency among individual accessions with regards to the source and amount of metadata provided along with sequence information. Acquiring core metadata is vital for the unification of SBC and SBI. Metadata should at minimum include the origin of sampling (host or matrix), location of sampling (geographic coordinates), type of sequence data collected (marker-based or metagenomic), and sequencing technology and quality assessment (raw data in universal FASTQ format). The use of already existing well-established standards for the recognition of environmental metadata associated with sequence data, such as MIMARKS (Minimum Information about a MARKer gene sequence; Yilmaz *et al.*, 2011) for marker-based amplicon data and BIOM (http://biom-format.org; McDonald *et al.*, 2012) for metagenomics and metatranscriptomics, should be required for all projects dealing with molecular data. Used as-is or with minimal modification, methods of metadata provenance are already integrated into existing data analysis pipelines and databases, so integration into SBC and SBI workflows should be fairly easy to accomplish. Perhaps the greatest emphasis needs to be placed on the community involvement needed to encourage researchers to participate in open, accurate data deposition and to incentivize standardization across many resources.

Developing open community standards for taxon delimitation

Open community standards must be developed for taxonomic classification and species identification based on environmental sequences. However, criteria for taxon recognition vary from group to group, and different workers faced with the same data may reach different but equally valid conclusions about taxon (particularly species) limits. The ITS 'bar code' region discriminates species in many groups of fungi, but in others it is too variable or too conserved (Schoch *et al.*, 2012; Lindner *et al.*, 2013). It is unlikely that uniform standards can be codified for taxon recognition in all clades. Moreover, the standards of today, based on a single marker (ITS) or suites of markers (e.g. calmodulin, beta tubulin, etc.), will probably change as single-cell genomics and other technologies evolve. The growing number of fungal genomes should be used to supplement ITS databases and characterize genomic diversity of the rDNA operon, but these repeat regions are usually unassembled from genome sequencing projects. Some databases, such as UNITE, have begun to remedy this by including ITS regions from sequenced genomes (Abarenkov *et al.*, 2010). In any event, care must be taken to understand and recognize sequence variation from nonorthologous marker regions or those acquired through horizontal gene transfer events (Klindworth *et al.*, 2012; Chun & Rainey, 2014). Absolute standards for taxon delimitation for all groups may never be achieved, but it is important that groups of taxonomic specialists work together to determine best practices for their clades of interest. In some cases, this will require that competing researchers set aside old arguments for the sake of developing unified sequence-based classifications that best serve the users of taxonomic classifications. Some in the group were concerned that certain regulations might stifle innovation so it was recommended that regulatory approaches be carefully initiated with open data and accessible workflows as a critical requirement.

Enabling a formal sequence-based species description under the *Code*

The *International Code of Nomenclature for Algae, Fungi and Plants* does not permit formal species description based only on sequence data (a physical type specimen is required in virtually all cases, although an illustration may serve as the type in some situations). Consequently, taxa discovered only through environmental sequences cannot be validly named. If they are named, then the invalid names lack the protection of priority under the *Code*, which could create nomenclatural instability. The vast majority of taxa discovered solely through metagenomic studies are not named, and they do not enter names-based taxonomic databases. The *Code* could be modified to allow purely sequence-based taxon description, which would promote communication and raise awareness about the diversity of fungi and their ecological roles. Objections to this proposal may reflect a lack of understanding of the purpose of the *Code*, which serves only to regulate the valid publication of names, not to pass judgment on the scientific hypotheses embodied in names.

Integrating archives, databases, and people

To achieve reproducibility and standardization, not only will sequence data, specimens, cultures, and actual nucleic acids need to be archived, but computational pipelines and algorithms used to process, identify, and perform classification will also need to be documented and preserved. To be truly useful, this archived information needs to step beyond the 'Materials and Methods' section of a publication and into open resources that integrate with databases. Challenges in this area include the cost of maintaining archived materials and methods as well as encouraging the scientific

community to contribute and maintain these archives. Granting agencies might help by requiring that funded and published research follow standards and tested workflows (Wilson *et al.*, 2014).

Giving credit where credit is due

There is a lack of emphasis on rewarding contributions that benefit the common good, such as database curation and the maintenance of archives, and a challenge exists to extend rewards beyond publications into quality data contribution, database development, archive curation, and the promotion of open science (Wilson *et al.*, 2014). Perhaps the greatest challenge here lies in convincing administrators and other people responsible for job promotion and retention that data and data maintenance should be valued on an equal level with other metrics such as publications.

Reaching out to everyone

One last challenge is to encourage the scientific community as a whole to adopt approaches that unify SBC and SBI. This would entail a strategy encouraging best practices, analysis workflows, and educational development for all levels of scientists and perhaps best initiated by a 'boots on the ground' plan to promote awareness of the integration of SBC and SBI among young scientists (through the university level) and the growing number of citizen scientists who make great contributions to specimen collection and documentation. Social media could be used to stress the connection between SBC and SBI and to encourage contributions from all levels of science with regard to a unified vision for both.

Moving forward – best practices

As a community of scientists studying plant-associated microorganisms, it would benefit us to encourage extensive community resources such as integrated databases for sequence- and meta-data and the promotion of archives consisting of analysis workflows and guidelines. Those convening at this meeting came to the conclusion that we should not delay the unification of classification and identification based on sequence data by devising and promoting mechanisms to name taxa identified solely through sequence data. The merging of SBC and SBI has potential to be swift and meaningful if journals, funding agencies, meeting organizers, and the scientific community as a whole are willing to adopt and develop open-resources and best-practices.

Acknowledgements

The workshop on which this paper is based was supported by a grant from the US National Science Foundation (award number DEB1424740) to David Geiser, Andrea Porras-Alfaro, David Hibbett, and John Taylor. The authors thank the organizers, as well as other participants of the meeting, for their input. The Mycological Society of America (MSA) supported a related symposium on 'Sequence-based identification in fungi' at the Annual MSA meeting immediately preceding the workshop.

Joshua R. Herr[1]*, Maarja Öpik[2] and David S. Hibbett[3]

[1]Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI 48823, USA;
[2]Department of Botany, University of Tartu, 40 Lai Street, 51005 Tartu, Estonia;
[3]Biology Department, Lasry Biological Science Center, Clark University, 950 Main St., Worcester, MA 01610, USA
(*Author for correspondence: tel +1 517 884 5287;
email joshua.r.herr@gmail.com)

References

Abarenkov K, Nilsson RH, Larsson K-H, Alexander IJ, Eberhardt U, Erland S, Høiland K, Kjøller R, Larsson E, Pennanen T *et al.* 2010. The UNITE database for molecular identification of fungi – recent updates and future perspectives. *New Phytologist* **186**: 281–285.

Bridge PD, Roberts PJ, Spooner BM, Panchal G. 2003. On the unreliability of published DNA sequences. *New Phytologist* **160**: 43–48.

Chun J, Rainey FA. 2014. Integrating genomics into the taxonomy and systematics of the *Bacteria* and *Archaea*. *International Journal of Systematic and Evolutionary Microbiology* **64**: 316–324.

Cole JR, Wang QM, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, Tiedje JM. 2013. Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Research* **42**: D633–D642.

Hibbett DS, Glotzer D, Nilsson RH, Ohman A, Nuhn M, Kirk PM. 2011. Progress in molecular and morphological taxon discovery in Fungi and options for formal classification of environmental sequences. *Fungal Biology Reviews* **25**: 38–47.

Hibbett DS, Taylor JW. 2013. Fungal systematics: is a new age of enlightenment at hand? *Nature Reviews Microbiology* **11**: 129–133.

Kennedy P, Stajich J. 2014. Twenty-first century mycology: a diverse, collaborative, and highly relevant science. *New Phytologist* **205**: 23–26.

Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, Glockner FO. 2012. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Research* **41**: 1–11.

Lindahl BD, Nilsson RH, Tedersoo L, Abarenkov K, Carlsen T, Kjøller R, Kõljalg U, Pennanen T, Rosendahl S, Stenlid J *et al.* 2013. Fungal community analysis by high-throughput sequencing of amplified markers – a user's guide. *The New Phytologist* **199**: 288–299.

Lindner DL, Carlsen T, Henrik Nilsson R, Davey M, Schumacher T, Kauserud H. 2013. Employing 454 amplicon pyrosequencing to reveal intragenomic divergence in the internal transcribed spacer rDNA region in fungi. *Ecology and Evolution* **3**: 1751–1764.

McDonald D, Clemente JC, Kuczynski J, Rideout J, Stombaugh J, Wendel D, Wilke A, Huse SM, Hufnagle J, Meyer F *et al.* 2012. The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *GigaScience* **1**: 7.

McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P. 2011. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME Journal* **6**: 610–618.

Öpik M, Vanatoa A, Vanatoa E, Moora M, Davison J, Kalwii JM, Reier Ü, Zobel M. 2010. The online database Maarj*AM* reveals global and ecosystemic distribution patterns in arbuscular mycorrhizal fungi (Glomeromycota). *New Phytologist* **188**: 223–241.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2012. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research* **2012**: 1–7.

Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, Chen W, 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, USA* **109**: 6241–6246.

Wilson G, Aruliah DA, Brown CT, Chue Hong NP, Davis M, Guy RT, Haddock SHD, Huff KD, Mitchell IM, Plumbley MD *et al.* 2014. Best practices for scientific computing. *PLoS Biology* **12**: e1001745.

Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, Gilbert JA, Karsch-Mizrachi I, Johnston A, Cochrane G *et al.* 2011. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nature Biotechnology* **29**: 415–420.

Key words: database, fungi, genomics, metadata, Mycological Society of America, mycorrhiza, sequenced-based classification, sequenced-based identification.