Evolutionary Bioinformatics

SOFTWARE REVIEW

PlutoF—a Web Based Workbench for Ecological and Taxonomic Research, with an Online Implementation for Fungal ITS Sequences

Kessy Abarenkov[1,2], Leho Tedersoo[1,2], R. Henrik Nilsson[2,3], Kai Vellak[2], Irja Saar[2], Vilmar Veldre[2], Erast Parmasto[1,4], Marko Prous[2], Anne Aan[1], Margus Ots[1], Olavi Kurina[4], Ivika Ostonen[2], Janno Jõgeva[1], Siim Halapuu[1], Kadri Põldmaa[2], Märt Toots[2,5], Jaak Truu[6], Karl-Henrik Larsson[7] and Urmas Kõljalg[1,2]

[1]Natural History Museum of Tartu University, 46 Vanemuise St., 51014 Tartu, Estonia. [2]Institute of Ecology and Earth Sciences, University of Tartu, 46 Vanemuise St., 51014 Tartu, Estonia. [3]Department of Plant and Environmental Sciences, University of Gothenburg, Box 461, 405 30 Göteborg, Sweden. [4]Institute of Agriculture and Environment, Estonian University of Life Sciences, 181 Riia St., 51014 Tartu, Estonia. [5]Institute of Statistics, University of Tartu, 2 Liivi St., 50409 Tartu, Estonia. [6]Institute of Molecular and Cell Biology, University of Tartu, 23 Riia St., 51010 Tartu, Estonia. [7]Natural History Museum of Oslo University, Box 1172, Blindern, N-0318 Oslo, Norway. Corresponding author email: kessy.abarenkov@ut.ee

Abstract: DNA sequences accumulating in the International Nucleotide Sequence Databases (INSD) form a rich source of information for taxonomic and ecological meta-analyses. However, these databases include many erroneous entries, and the data itself is poorly annotated with metadata, making it difficult to target and extract entries of interest with any degree of precision. Here we describe the web-based workbench PlutoF, which is designed to bridge the gap between the needs of contemporary research in biology and the existing software resources and databases. Built on a relational database, PlutoF allows remote-access rapid submission, retrieval, and analysis of study, specimen, and sequence data in INSD as well as for private datasets though web-based thin clients. In contrast to INSD, PlutoF supports internationally standardized terminology to allow very specific annotation and linking of interacting specimens and species. The sequence analysis module is optimized for identification and analysis of environmental ITS sequences of fungi, but it can be modified to operate on any genetic marker and group of organisms. The workbench is available at http://plutof.ut.ee.

Keywords: sequence management environment, data mining, metadata, thin clients, sequence identification

Evolutionary Bioinformatics 2010:6 189–196

doi: 10.4137/EBO.S6271

This article is available from http://www.la-press.com.

Introduction

Molecular (DNA-based) techniques and informatics form vital research implements in nearly all fields of the biological sciences, including ecology and taxonomy.[1–3] As more and more DNA sequences accumulate in the International Nucleotide Sequence Databases (INSD: EMBL, GenBank, and DDBJ),[4] the joint corpus of sequence data generated by the international research community gradually attains far-reaching explanatory power in the disciplines of taxonomy, ecology, and biogeography.[5–7] The analysis of such amalgamated data are of particular relevance to understanding the biology of microorganisms because of their inconspicuous and poorly understood nature, their high population sizes, and the insurmountable difficulties associated with keeping many of them in culture.[8,9] Extensive sampling in terms of sequence depth, ecological niches, and geographical regions is typically required to answer microbiological questions with any noteworthy degree of certainty, pointing to the benefits of—indeed, need for—integrating datasets and resources already generated. Studies in microbiology rely to a great extent not only on the sequence data itself but also on the associated metadata—auxiliary information on, eg, collection site, host, and soil type. Unfortunately, INSD does not require that metadata be submitted alongside the sequence data itself and offers little by way of a standardized vocabulary for specification of metadata, leaving the sequence authors free to decide what information items to give and how to do it. Thus, in spite of international standardization and data infrastructure initiatives such as the Darwin Core standard (maintained by TDWG, http://rs.tdwg.org/dwc/) and the Microbiological Common Language,[10] the INSD metadata is often given in inconsistent and irreconcilable ways (eg, specified under different headings or using synonymous wording).

Additional technical problems further complicate data mining of public sequence data. Names of species or higher taxonomic lineages are often applied in conflicting ways due to differences in taxonomic opinion or in tradition among ecologists and taxonomists.[11] A substantial proportion of the publicly available sequences are furthermore chimeric, reverse complementary, or contain numerous erroneous bases or ambiguities.[12–14] Worryingly, there is at present no straightforward way to alert other users of INSD to the presence of such defective data,[15] paving the way for the percolation of incorrect information through the databases and the scientific community at large.[16] As an example, the set of nuclear internal transcribed spacer (ITS) sequences of fungi in INSD includes an estimated 1% reverse complementary, 1.5% chimeric, and more than 10% incorrectly identified entries.[17,18] This is problematic given the weight assigned to the ITS region in contemporary mycology; it is the most commonly sequenced genetic marker for species identification from environmental samples due to its ease of amplification and its discriminative power at the species level.[19–21]

These complications notwithstanding, the INSD provides an important backbone resource for the development of more accurate, but less inclusive, databases, such as SILVA,[22] Greengenes,[23] and UNITE.[19,24] One of the main objectives of these resources is to facilitate reliable taxonomic identification of newly generated environmental and clinical sequences (ie, from samples such as soil, wood, and gut). The core set of reference sequences in these databases is composed of entries that have passed various steps of quality control and that are deemed of sufficient standard and reliability to be of true use in taxonomy and ecology. As such these initiatives often assume the role of INSD as the primary reference database in large-scale environmental sequencing studies,[25–27] and they typically feature tailored search tools and analysis modules not found in INSD. As an example, the command line-based utility MOTHUR[28] was developed to span the range of steps involved in assigning environmental sequences to species or operational taxonomic unit (OTU) level and to obtain diversity assessments of the samples at hand. However, these utilities were primarily built with prokaryotes and the ribosomal small subunit (16S) gene in mind; furthermore, many of them require that the indata be presented in the form of a joint, scientifically sound multiple alignment.[23,28,29] Thus, by their very nature, these resources are largely incompatible with fungal ITS sequence data since the high level of variability of the region precludes admissible alignment across higher taxonomic levels.

Here we describe an online workbench—PlutoF—that is designed to tackle the many issues of contemporary DNA-based research in ecology and taxonomy. PlutoF was developed in response to the need of many researchers and research networks to manage and analyze their molecular data in ways not fully supported by existing resources and databases (Table 1).

Table 1. Overview and comparison of PlutoF with INSD, mothur, and QIIME.

	INSD	PlutoF	mothur	QIIME[50]
Main idea of the platform	*Gathering and providing open access to all nucleotide sequences and their basic metadata; web-based and standalone applications*	*To upload, sort, update, and analyse biodiversity data online; web-based workbench*	*To analyse molecular data to assign environmental sequences to species level; designed for less variable datasets (ie, the bacterial 16S gene); standalone package*	*Pipeline for performing microbial community analysis; standalone package*
FEATURE				
Data sharing within workgroups	No support for creating and sharing data within workgroups	Yes, users can give specific workgroups access to their data	No	No
Third-party annotation	No	Non-anonymous, including annotation on taxonomy, ecology, and sequence quality	No	No
Automated chimera checking	No	Based on the ITS Chimera Checker and manual annotation	Multiple alignment-based options	Based on either the blast_fragments approach or ChimeraSlayer
Sequence quality tagging	No	Based on manual annotation by molecular taxonomists	Sorts low quality sequences into separate files	Low quality sequences can be filtered out
Support for multiple determinations	No	Indefinite number of determinations can be assigned to sequences non-anonymously	No	No
Merging INSD and private datasets	Yes, if the user's own dataset is part of INSD	Possible to merge INSD and workgroup data for searches and analyses	Multiple alternative reference databases	Multiple alternative reference databases
Advanced search options	To some extent; limited by widespread use of non-standardized terminology	Searches based on both the original INSD data and annotated metadata in the relational database	No	No
Analysis module	Chiefly local alignment (BLAST) and neighbour joining programs and various LinkOut resources	MassBLASTer, Chimera Checker, ITS extractor, 454 pipeline, emerencia	Command line-based programs for quality control, aligning, trimming, clustering, and biodiversity statistics for 16S rDNA sequences	Command line based programs for quality control, aligning, clustering, phylogeny reconstruction, biodiversity statistics, and visualization of the end results
Taxonomy	NCBI taxonomy	Index Fungorum, APG III, Fauna Europaea; regularly updated	Multiple alternative taxonomies	RDP or alternative taxonomies provided by the user

The ultimate goal of PlutoF is to cover all elements of the extant biodiversity, *viz.* ecological, genetic, and taxonomic diversity of all biological kingdoms. This will enable researchers to address integrated questions spanning the different fields of the biosciences, something that is in increasing demand.[30] Through the PlutoF workbench any researcher can develop an indefinite number of databases and bring together existing databases for joint analysis. Data available to the user can be searched, sorted, and analysed across all these databases. The present study addresses the procedures of rapid submission and retrieval of large sequence datasets, annotation of new and pre-existing sequences and specimens, and the sequence analysis features of PlutoF. The workbench supports tools for processing raw community sequence data from any genetic marker, but the analysis module of PlutoF is optimized for fungal ITS sequences by default.

Database Structure and Operation

Database and web design

The PlutoF workbench draws from the relational MySQL v. 5.0.77 database and has more than 150 tables for storing taxonomic, ecological, and molecular data (Suppl. Item 1). The database structure (Fig. 1) is rooted in Taxonomer[31] but with far-reaching modifications to integrate modules for storing multimedia, molecular data, and analysis results. The current database model enables users to insert, search, and browse various taxon occurrences (based on, eg, specimens, observations, or DNA sequences), literature references, and scientific collections. PlutoF has a hierarchical study/plot/sample model (Fig. 1) that enables users to manage their own projects all the way from sampling design and persistent storage of data to molecular data analysis and interpretation of the results. Users can work with their own data and form

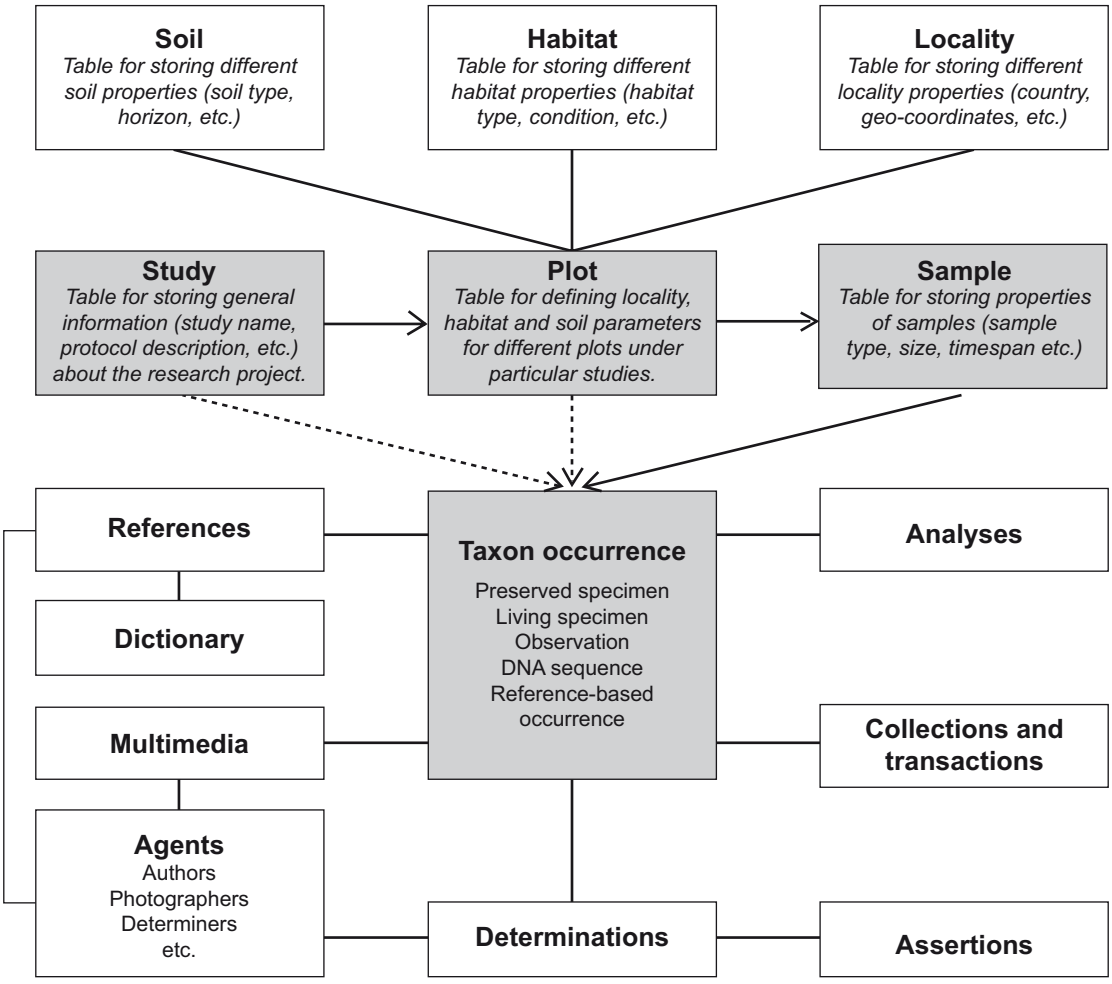

Figure 1. Simplified database scheme showing the core modules. Shaded modules and arrows illustrate the hierarchical structure of the study/plot/sample model, and lines indicate relationships among other modules.

workgroups of users that share the data. The regularly updated classification used in the central taxonomy module is largely based on Hibbett et al (2007)[32] and Index Fungorum (http://www.indexfungorum.org/) for fungi, Fauna Europaea (http://www.faunaeur.org/) for animals (higher taxonomic levels have however been updated to reflect recent phylogenetic literature),[33,34] and APG III[35] for plants.

As implemented at the University of Tartu (Estonia), the PlutoF workbench runs on a quad-core 64-bit Linux server (CentOS 5.2, Apache webserver v. 2.2.3). To communicate with the databases, the PlutoF web interface uses the PHP, HTML, CSS, AJAX, JavaScript, and SQL programming languages. The software packages of the analysis module are written in Perl. PlutoF has been tested with all major web browsers, including Mozilla Firefox (v. 2.x and 3.x), Internet Explorer (v. 6.x–8.x), and Safari (v. 5.0) on various operating systems.

Storage of sequence data in PlutoF

The core information of the PlutoF system is the sequence data, and PlutoF supports the distinction between external (eg, INSD) and internal (public or private) sequences. These datasets can be queried separately or jointly. In recognition of the explanatory power of the body of amalgamated fungal ITS sequences in INSD, PlutoF offers the possibility to mirror the INSD for the fungal ITS data (Suppl. Fig. 2a) or any other genetic marker of interest; in the UNITE database, all reasonably full-length fungal ITS sequences identified as such in INSD are downloaded on a monthly basis. As of September 2010, UNITE thus contained 160,581 INSD sequences and 6,368 native sequences of the fungal ITS region (the latter including 2,843 entries from fully identified and vouchered reference fruiting bodies). The overall corpus of sequences corresponds to about 15,000 fully identified species of fungi; about 50% of the sequences, nearly all of which stem from INSD, remain unidentified to species level however. The system furthermore supports the distinction between different classes of sequences. The present classes include INSD, native reference, native non-reference, and next generation sequencing (NGS) sequences (eg, sequences from massively parallel ("454") pyrosequencing[36] efforts). NGS entries form a challenge due to their sheer numbers and potential reduction in length and read quality.[37,38] We advocate that pyrosequencing entries be marked as being distinct from sequences obtained using traditional Sanger sequencing. Since cleaning and filtering methods of pyrosequencing raw data improve over time,[37,39–41] the availability of raw NGS data underlying scientific studies and results may prove important for ulterior analyses. PlutoF accordingly supports deposition of NGS data at two levels—i) compressed files of raw sequence data, quality scores, and barcode translation tables; and ii) quality filtered sequences—optionally in the form of majority-rule consensus sequences—with abundance and sample information added to their annotation. Templates for comma- and tab-delimited files are available for these purposes. These and other file types can be uploaded to the database through the PlutoF Digital Repository module, which recognises most common file types and formats.

The sequence data in INSD are by default retrieved with all available metadata (eg, isolation source, geographical locality, and literature reference); these data are extracted and stored in PlutoF. All INSD entries are indexed according to study of origin using the hierarchical model so that sequences belonging to the same study are separated into plots and samples based on their locality information, as available. This makes precise data retrieval possible (Suppl. Fig. 2b); for instance one could search for all studies involving fungal ITS sequences on Canadian territory in a single query. Similarly, all sequences deposited by a specific researcher or during a given year are easily retrieved. Such searches are not always straightforward in INSD itself.

Data Handling and Sequence Analysis Modules

Handling user data

The PlutoF structure supports submission of sample details and other auxiliary information along with sequence data on a sequence-per-sequence, as well as bulk, basis. For example, samples (as *Taxon occurrence* in the main menu) may comprise multiple specimens in a scientific collection, mere field observations of some given species, or DNA sequences from various genes and organisms. Similarly to INSD, direct submission of sequence data requires that the name of the study or project be given along with one or more plot as relevant. Unlike INSD, however,

PlutoF offers a standardized vocabulary for describing and defining the properties of the sequences and the conditions under which they were obtained. In accordance with contemporary research in ecology, PlutoF supports the subdivision of plots into samples to allow very specific data retrieval queries. For each plot and sample, comprehensive descriptions can be provided, including data on locality (eg, geocoordinates, altitude, and municipalities), habitat (following the IUCN habitat classification system: http://www.iucnredlist.org/technical-documents/classification-schemes/habitats-classification-scheme-ver3 including history, age, and climate), soil (the FAO classification),[42] soil horizon (chemical and physical properties), plant root (eg, biomass, turnover, and production by diameter), forest (eg, canopy height, stand density, and basal area), and general information (name, type, and size). Specimen information includes taxonomy (eg, name of the taxon and pheno/logic/typic data), collection (date, collector, and determiner), and substrate/interacting taxon (taxonomy and type of interaction). Sequence information includes ID, DNA sequence, name of the gene, PCR primers, and level of availability to other users.

While the taxonomic classification in PlutoF follows international standards, power users can add and edit taxon names directly in the workbench on subclass or lower level. Above the level of subclass, only administrators can implement changes; prior agreement between classification curators is however required. All users can apply for the right to upload and edit taxon names.

Annotating INSD entries

The PlutoF workbench allows third-party annotation of INSD, as well as native, sequences. The primary rationale is to support the addition of missing metadata, the correction of incomplete or incorrect taxonomic information, and the provision of information pertaining to the overall reliability of the sequence, such as chimeric nature. The original information is retained, and annotations are introduced as separate data layers. All annotations are by default non-anonymous. Missing metadata can be added directly to each specimen/sequence, sample, or plot in the relevant window (Suppl. Fig. 2c,d). Sequences of ectomycorrhizal fungi can be assigned to monophyletic lineages (sensu Tedersoo et al 2010)[43] to overcome paraphyly. Updating taxonomic annotations—typically by providing additional taxon names to misidentified or unnamed sequences—should only be undertaken by users with sufficient experience of the taxonomic lineage at hand, and PlutoF supports a peer-review type of process for managing such annotations.

Bioinformatics resources and the analysis module

PlutoF enables rapid sorting and retrieval of relevant sequence data by various search parameters such as sequence ID, taxon name, country, interacting taxon, sequence length, and study. Another option is to use the BLAST[44]-based search tool *emerencia*[45] which is designed to track the taxonomic affiliation of insufficiently identified ITS sequences over time. In both cases, relevant entries are marked and sent to the clipboard, where they can be checked for duplicates (data that has been submitted to both PlutoF and INSD) and exported to FASTA or comma separated (csv) files with a full set of metadata. In addition, data can be sent to an integrated Google Maps module for instant geographical visualisation (Suppl. Fig. 2e).

The analysis module includes software for extracting and classifying ITS sequences that are derived from high-throughput sequencing or cloning studies (Suppl. Fig. 2f). Based on highly conserved short signal motifs, the ITS Extractor[46] separates the ITS1 and ITS2 subregions of the ITS region from the flanking rDNA genes, a process that is much to the purpose of high-precision clustering and sequence identification.[47,48] BLASTClust of the BLAST suite performs single-linkage clustering at user-defined similarity threshold values to collapse query datasets into OTUs. The chimera checker utility identifies potentially chimeric ITS sequences through contrasting the respective taxonomic signal of the ITS1 and ITS2 subregions.[18] A serial BLAST engine to compare arbitrarily large query datasets for similarity against the sequences in UNITE/INSD is also available. A pyrosequencing pipeline allows for pyrosequencing datasets of the ITS region to be analysed in a reasonable time, providing the taxonomic results in a spreadsheet format where OTUs are separated into rows and samples into columns.[49]

Conclusions

PlutoF is a web-based workbench for the storage, editing, analysis, and overall management of ecological, taxonomic, and genetic data. It has a strong ecological and taxonomic orientation but also covers several aspects of biogeography and co-evolution. PlutoF was developed in light of the urgent need to address integrated questions in these fields through DNA sequence data. In recognition of the increasing internationalisation of biological research and the fact that different research groups and taxonomic lineages require different information items to be stored and analysed, PlutoF is flexible, scalable, and highly modularized. PlutoF is run at University of Tartu, Estonia, and it is open for public use, including data submission, annotation, and analysis. Potential users are requested to contact the curator (http://plutof.ut.ee/contact.php) for obtaining authentication information.

Disclosures

This manuscript has been read and approved by all authors. This paper is unique and is not under consideration by any other publication and has not been published elsewhere. The authors and peer reviewers of this paper report no conflicts of interest. The authors confirm that they have permission to reproduce any copyrighted material.

Acknowledgements

We thank the Estonian Science Foundation (grants 8-2/T8030PKPK, 0180012s09, 0180122s08, 0180127s08, 6939, 7434, 7558, 8235, and JD-0092), FIBIR, and Kapten Carl Stenholms Donationsfond for financial support.

References

1. Tautz D, Arctander P, Minelli A, Thomas RH, Vogler AP. A plea for DNA taxonomy. *Trends Ecol Evol*. 2003;18:70–4.
2. Peay KG, Kennedy PG, Bruns TD. Fungal community ecology: a hybrid beast with a molecular master. *BioScience*. 2008;58:799–810.
3. Stajich JE, Berbee ML, Blackwell M, et al. The fungi. *Curr Biol*. 2009;19:840–5.
4. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. *Nucl Acids Res*. 2009;37:D26–31.
5. Lozupone CA, Knight R. Global patterns in bacterial diversity. *Proc Natl Acad Sci U S A*. 2007;104:11436–40.
6. Ryberg M, Nilsson RH, Kristiansson E, Topel M, Jacobsson S, Larsson E. Mining metadata from unidentified ITS sequences in GenBank: a case study in *Inocybe* (Basidiomycota). *BMC Evol Biol*. 2008;8:50.
7. Nilsson RH, Ryberg M, Sjökvist E, Abarenkov K. Rethinking taxon sampling in the light of environmental sequencing. *Cladistics*. 2010;doi:10.1111/j.1096-0031.2010.00336.x (in press).
8. Oren A. Prokaryote diversity and taxonomy: current status and future challenges. *Phil Trans R Soc Lond B*. 2004;359:623–38.
9. Konstantinidis K, Tiedje JM. Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. *Curr Opin Microbiol*. 2007;10:504–9.
10. Verslyppe B, Kottman R, de Smet W, de Baets B, de Vos P, Dawyndt P. Microbiological Common Language (MCL): a standard for electronic information exchange in the Microbial Commons. *Res Microbiol*. 2010;161:439–45.
11. Nielsen DL, Shiel RJ, Smith FJ. Ecology versus taxonomy: is there a middle ground? *Hydrobiologia*. 1998;387–388:451–7.
12. Christen R. Global sequencing: a review of current molecular data and new methods available to assess microbial diversity. *Microb Environ*. 2008;23:253–68.
13. Ryberg M, Kristiansson E, Sjökvist E, Nilsson RH. An outlook on the fungal internal transcribed spacer sequences in GenBank and the introduction of a web-based tool for the exploration of fungal diversity. *New Phytol*. 2009;181:471–7.
14. Nilsson RH, Veldre V, Wang Z, et al. A note on the incidence of reverse complementary fungal ITS sequences in the public sequence databases and a software means for their detection and reorientation. *Mycoscience*. 2010 (in press).
15. Pennisi E. Proposal to 'Wikify' GenBank meets stiff resistance. *Science*. 2008;319:1598–9.
16. Gilks WR, Audit B, de Angelis D, Tsoka S, Ouzounis CA. Modeling the percolation of annotation errors in a database of protein sequences. *Bioinformatics*. 2002;18:1641–9.
17. Bidartondo MI. ADD 2 CO_AUTHORS. Preserving accuracy in GenBank. *Science*. 2008;319:1616.
18. Nilsson RH, Abarenkov K, Veldre V, et al. An open source chimera checker for the fungal ITS region. *Mol Ecol Res*. 2010;10:1076–81.
19. Abarenkov K, Nilsson RH, Larsson K-H, et al. The UNITE database for molecular identification of fungi—recent updates and future perspectives. *New Phytol*. 2010;186:281–5.
20. Begerow D, Nilsson RH, Unterseher M, Maier W. Current state and perspectives of fungal DNA barcoding and rapid identification procedures. *Appl Microbiol Biotechnol*. 2010;87:99–108.
21. Eberhardt U. A constructive step towards selecting a DNA barcode for fungi. *New Phytol*. 2010;187:265–8.
22. Pruesse E, Quast C, Knittel K, et al. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucl Acids Res*. 2007;35:7188–96.
23. DeSantis TZ, Hugenholtz P, Larsen N, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *App Environ Microbiol*. 2006;72:5069–72.
24. Kõljalg U, Larsson K-H, Abarenkov K, et al. UNITE: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi. *New Phytol*. 2005;166:1063–8.
25. Dethlefsen L, Huse S, Sogin ML, Relman DA. The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS Biol*. 2008;6:e280.
26. Bueé M, Reich M, Murat C, et al. 454 Pyrosequencing analyses of forest soils reveal an unexpectedly high fungal diversity. *New Phytol*. 2009;184:449–56.
27. Öpik M, Vanatoa A, Vanatoa E, et al. The online database MaarjAM reveals global and ecosystemic distribution patterns in arbuscular mycorrhizal fungi (Glomeromycota). *New Phytol*. 2010;188:223–41.
28. Schloss PD, Westcott SL, Ryabin T, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol*. 2009;75:7537–41.
29. Huber T, Faulkner G, Hugenholtz P. Bellerophon; a program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics*. 2004;20:2317–9.
30. Arnold AE, Lamit LJ, Gehring CA, Bidartondo MI, Callahan H. Interwoven branches of the plant and fungal trees of life. *New Phytol*. 2010;185:874–8.
31. Pyle RL. Taxonomer: a relational data model for managing information relevant to taxonomic research. *Phyloinformatics*. 2003;1:1–54.

32. Hibbett DS, Binder M, Bischoff JF, et al. A higher-level phylogenetic classification of the Fungi. *Mycol Res*. 2007;111:509–47.
33. De Ley P, Blaxter M. A new system for *Nematoda*: combining morphological characters with molecular trees, and translating clades into ranks and taxa. *Nemat Monogr Persp*. 2004;2:633–53.
34. Sørensen MV, Giribet G. A modern approach to rotiferan phylogeny: combining morphological and molecular data. *Mol Phyl Evol*. 2006;40:585–608.
35. Angiosperm Phylogeny Group III. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Bot J Linn Soc*. 2009;161:105–21.
36. Margulies M, Egholm M, Altman WE, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*. 2005;437:376–80.
37. Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol*. 2007;8:R143.
38. Shendure J, Ji H. Next-generation DNA sequencing. *Nature Biotechnol*. 2008;26:1135–45.
39. Quince C, Lanzen A, Curtis TP, et al. Accurate determination of microbial diversity from 454 pyrosequencing data. *Nature Meth*. 2009;6:639–41.
40. Huse SM, Welch DM, Morrison HG, Sogin ML. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ Microbiol*. 2010;12:1889–98.
41. Reeder J, Knight R. Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions. *Nature Meth*. 2010;7:668–9.
42. IUSS Working Group WRB. *World reference base for soil resources 2006. 2nd edition*. *World Soil Resources Reports No. 103*. 2006; FAO, Rome.
43. Tedersoo L, May TW, Smith ME. Ectomycorrhizal lifestyle in fungi: global diversity, distribution, and evolution of phylogenetic lineages. *Mycorrhiza*. 2010;20:217–63.
44. Altschul SF, Madden TL, Schäffer AA, et al. Gapped BLAST and PSI BLAST: a new generation in protein database search programs. *Nucl Acids Res*. 1997;25:3389–402.
45. Nilsson RH, Kristiansson E, Ryberg M, Larsson K-H. Approaching the taxonomic affiliation of unidentified sequences in public databases -an example from the mycorrhizal fungi. *BMC Bioinform*. 2005;6:178.
46. Nilsson RH, Veldre V, Hartmann M, et al. An open source software package for automated extraction of ITS1 and ITS2 from fungal ITS sequences for use in high-throughput community assays and molecular ecology. *Fung Ecol*. 2010;3:284–7.
47. Hartmann M, Howes CG, Abarenkov K, Mohn WW, Nilsson RH. V-Xtractor: An open-source, high-throughput software tool to identify and extract hypervariable regions of small subunit (16S/18S) ribosomal RNA gene sequences. *J Microbiol Meth*. 2010;83:250–3.
48. Nilsson RH, Kristiansson E, Ryberg M, Hallenberg N, Larsson K-H. Intraspecific ITS variability in the kingdom Fungi as expressed in the international sequence databases and its implications for molecular species identification. *Evol Bioinf*. 2008;4:193–201.
49. Tedersoo L, Nilsson RH, Abarenkov K, et al. 454 Pyrosequencing and Sanger sequencing of tropical mycorrhizal fungi provide similar results but reveal substantial methodological biases. *New Phytol*. 2010;188:291–301.
50. Caporaso JG, Kuczynski J, Stombaugh J, et al. QIIME allows analysis of high-throughput community sequence data. *Nature Methods*. 2010;7:335–6.