Fungal Diversity (2014) 67:11–19
DOI 10.1007/s13225-014-0291-8

Improving ITS sequence data for identification of plant pathogenic fungi

R. Henrik Nilsson · Kevin D. Hyde · Julia Pawłowska · Martin Ryberg · Leho Tedersoo · Anders Bjørnsgard Aas · Siti A. Alias · Artur Alves · Cajsa Lisa Anderson · Alexandre Antonelli · A. Elizabeth Arnold · Barbara Bahnmann · Mohammad Bahram · Johan Bengtsson-Palme · Anna Berlin · Sara Branco · Putarak Chomnunti · Asha Dissanayake · Rein Drenkhan · Hanna Friberg · Tobias Guldberg Frøslev · Bettina Halwachs · Martin Hartmann · Beatrice Henricot · Ruvishika Jayawardena · Ari Jumpponen · Håvard Kauserud · Sonja Koskela · Tomasz Kulik · Kare Liimatainen · Björn D. Lindahl · Daniel Lindner · Jian-Kui Liu · Sajeewa Maharachchikumbura · Dimuthu Manamgoda · Svante Martinsson · Maria Alice Neves · Tuula Niskanen · Stephan Nylinder · Olinto Liparini Pereira · Danilo Batista Pinho · Teresita M. Porter · Valentin Queloz · Taavi Riit · Marisol Sánchez-García · Filipe de Sousa · Emil Stefańczyk · Mariusz Tadych · Susumu Takamatsu · Qing Tian · Dhanushka Udayanga · Martin Unterseher · Zheng Wang · Saowanee Wikee · Jiye Yan · Ellen Larsson · Karl-Henrik Larsson · Urmas Kõljalg · Kessy Abarenkov

Received: 28 March 2014 / Accepted: 18 April 2014 / Published online: 15 May 2014

Summary Plant pathogenic fungi are a large and diverse assemblage of eukaryotes with substantial impacts on natural ecosystems and human endeavours. These taxa often have complex and poorly understood life cycles, lack observable, discriminatory morphological characters, and may not be amenable to in vitro culturing. As a result, species identification is frequently difficult. Molecular (DNA sequence) data have emerged as crucial information for the taxonomic identification of plant pathogenic fungi, with the nuclear ribosomal internal transcribed spacer (ITS) region being the most popular marker. However, international nucleotide sequence databases are accumulating numerous sequences of compromised or low-resolution taxonomic annotations and substandard technical quality, making their use in the molecular identification of plant pathogenic fungi problematic. Here we report on a concerted effort to identify high-quality reference sequences for various plant pathogenic fungi and to re-annotate incorrectly or insufficiently annotated public ITS sequences from these fungal lineages. A third objective was to enrich the sequences with geographical and ecological metadata. The results – a total of 31,954 changes – are incorporated in and made available through the UNITE database for molecular identification of fungi (http://unite.ut.ee), including standalone FASTA files of sequence data for local BLAST searches, use in the next-generation sequencing analysis platforms QIIME and mothur, and related applications. The present initiative is just a beginning to cover the wide spectrum of plant pathogenic fungi, and we invite all researchers with pertinent expertise to join the annotation effort.

Keywords Phytopathogenic fungi · Molecular identification · ITS · Taxonomy · Annotation

Anders Bjørnsgard Aas, Siti A. Alias, Artur Alves, Cajsa Lisa Anderson, Alexandre Antonelli, A. Elizabeth Arnold, Barbara Bahnmann, Mohammad Bahram, Johan Bengtsson-Palme, Anna Berlin, Sara Branco, Putarak Chomnunti, Asha Dissanayake, Rein Drenkhan, Hanna Friberg, Tobias Guldberg Frøslev, Bettina Halwachs, Martin Hartmann, Beatrice Henricot, Ruvishika Jayawardena, Ari Jumpponen, Håvard Kauserud, Sonja Koskela, Tomasz Kulik, Kare Liimatainen, Björn D. Lindahl, Daniel Lindner, Jian-Kui Liu, Sajeewa Maharachchikumbura, Dimuthu Manamgoda, Svante Martinsson, Maria Alice Neves, Tuula Niskanen, Stephan Nylinder, Olinto Liparini Pereira, Danilo Batista Pinho, Teresita M. Porter, Valentin Queloz, Taavi Riit, Marisol Sánchez-García, Filipe de Sousa, Emil Stefańczyk, Mariusz Tadych, Susumu Takamatsu, Qing Tian, Dhanushka Udayanga, Martin Unterseher, Zheng Wang, Saowanee Wikee and Jiye Yan contributed equally to the project and are listed in alphabetical order.

Electronic supplementary material The online version of this article (doi:10.1007/s13225-014-0291-8) contains supplementary material, which is available to authorized users.

R. H. Nilsson · C. L. Anderson · A. Antonelli · S. Martinsson · F. de Sousa · E. Larsson
Department of Biological and Environmental Sciences, University of Gothenburg, Box 461, 405 30 Gothenburg, Sweden

K. D. Hyde · A. Dissanayake · R. Jayawardena · J.-K. Liu · S. Maharachchikumbura · D. Manamgoda · Q. Tian · D. Udayanga
Institute of Excellence in Fungal Research, Mae Fah Luang University, Chiang Rai 57100, Thailand

K. D. Hyde · P. Chomnunti · A. Dissanayake · R. Jayawardena · J.-K. Liu · S. Maharachchikumbura · D. Manamgoda · Q. Tian · D. Udayanga · S. Wikee
School of Science, Mae Fah Luang University, Chiang Rai 57100, Thailand

J. Pawłowska
Department of Plant Systematics and Geography, Faculty of Biology, University of Warsaw, Al. Ujazdowskie 4, 00-478 Warsaw, Poland

M. Ryberg
Department of Organismal Biology, Uppsala University, Norbyvägen 18D, 75236 Uppsala, Sweden

L. Tedersoo · M. Bahram · T. Riit · U. Kõljalg
Institute of Ecology and Earth Sciences, University of Tartu, Lai 40, Tartu 51005, Estonia

A. B. Aas · H. Kauserud
Microbial Evolution Research Group, University of Oslo, Blindernveien 31, 0371 Oslo, Norway

S. A. Alias
Institute of Biological Sciences, University of Malaya, 50603 Kuala Lumpur, Malaysia

A. Alves
Department of Biology, CESAM, University of Aveiro, 3810-193 Aveiro, Portugal

A. E. Arnold
School of Plant Sciences, The University of Arizona, 1140 E South Campus Drive, Forbes 303, Tucson, AZ 85721, USA

B. Bahnmann
Laboratory of Environmental Microbiology, Institute of Microbiology ASCR, Vídeňská 1083, 14220 Prague 4, Czech Republic

J. Bengtsson-Palme
Department of Infectious Diseases, Institute of Biomedicine, Sahlgrenska Academy, University of Gothenburg, Guldhedsgatan 10, 413 46 Göteborg, Sweden

A. Berlin · H. Friberg · B. D. Lindahl
Department of Forest Mycology and Plant Pathology, Swedish University of Agricultural Sciences, Box 7026, 750 07 Uppsala, Sweden

S. Branco
University of California at Berkeley, 321 Koshland Hall University of California, Berkeley, CA 94720-3102, USA

A. Dissanayake · R. Jayawardena · J. Yan
Institute of Plant and Environment Protection, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China

R. Drenkhan
Institute of Forestry and Rural Engineering, Estonian University of Life Sciences, Kreutzwaldi, 5, 51014 Tartu, Estonia

T. G. Frøslev
Natural History Museum of Denmark, Øster Voldgade 5-7, 1350 København K, Denmark

B. Halwachs
Institute for Genomics and Bioinformatics, Graz University of Technology, 8010 Graz, Austria

B. Halwachs
Core Facility Bioinformatics, Austrian Centre of Industrial Biotechnology, 8010 Graz, Austria

M. Hartmann
Forest Soils and Biogeochemistry, Swiss Federal Research Institute WSL, Zuercherstrasse 111, 8903 Birmensdorf, Switzerland

M. Hartmann
Molecular Ecology, Institute for Sustainability Sciences, Agroscope, Reckenholzstrasse 191, 8046 Zurich, Switzerland

B. Henricot
Plant Pathology, The Royal Horticultural Society, Wisley, Woking, Surrey GU23 6QB, UK

A. Jumpponen
Division of Biology, Kansas State University, Manhattan, KS 66506, USA

S. Koskela
Metapopulation Research Group, Department of Biosciences, University of Helsinki, PO Box 65, 00014 Helsinki, Finland

T. Kulik
Department of Diagnostics and Plant Pathophysiology, University of Warmia and Mazury in Olsztyn, Plac Lodzki 5, Olsztyn 10-957, Poland

K. Liimatainen · T. Niskanen
Plant Biology, Department of Biosciences, University of Helsinki, P.O. Box 65, 00014 Helsinki, Finland

D. Lindner
US Forest Service, Northern Research Station, Center for Forest Mycology Research, One Gifford Pinchot Drive, Madison, WI, USA

M. A. Neves
Departamento Botânica, PPG Biologia de Fungos, Algas e Plantas, Universidade Federal de Santa Catarina, Florianópolis, SC, Brazil

S. Nylinder
Department of Botany, Swedish Natural History Museum, Svante Arrhenius väg 7, 10405 Stockholm, Sweden

O. L. Pereira · D. B. Pinho
Departamento de Fitopatologia, Universidade Federal de Viçosa, Viçosa, Minas Gerais 36570-900, Brazil

T. M. Porter
Department of Biology, McMaster University, Hamilton, ON L8S 4K1, Canada

V. Queloz
ETH Zürich, Institute for Integrative Biology, CHN G 68.3, Universitätsstrasse 16, 8092 Zürich, Switzerland

M. Sánchez-García
Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, TN 37996-1610, USA

E. Stefańczyk
Plant Breeding and Acclimatization Institute-National Research Institute, Młochów Research Centre, Platanowa 19, 05831 Młochów, Poland

M. Tadych
Department of Plant Biology and Pathology, School of Environmental and Biological Sciences, Rutgers, The State University of New Jersey, 59 Dudley Rd., New Brunswick, NJ 08901, USA

S. Takamatsu
Laboratory of Plant Pathology, Faculty of Bioresources, Mie University, 1577 Kurima-Machiya, Tsu-city 514-8507, Japan

M. Unterseher
Institute of Botany and Landscape Ecology, Ernst-Moritz-Arndt University, Soldmannstr. 15, 17487 Greifswald, Germany

Z. Wang
Biostatistics Department, Yale School of Public Health, New Haven, CT 06520, USA

K.-H. Larsson
Natural History Museum, P.O. Box 1172, Blindern 0318, Oslo, Norway

U. Kõljalg · K. Abarenkov (✉)
Natural History Museum, University of Tartu, Vanemuise 46, Tartu 51014, Estonia
e-mail: kessy.abarenkov@ut.ee

Introduction

Plant pathogenic fungi are a large assemblage distributed across the fungal tree of life (Stajich et al. 2009). They share a nutritional strategy that adversely affects their plant hosts, sometimes in ways that have negative repercussions for human activities. Precise knowledge of the identity of the causal agent(s) of any given plant disease is the first step toward meaningful countermeasures and disease surveillance (Rossman and Palm-Hernández 2008; Kowalski and Holdenrieder 2009; Fisher et al. 2012). In addition, recent reports of emerging plant pathogens and their cross-kingdom infections to animals and immunocompromised humans accentuate the need for accurate and quick identification in potential outbreaks (Cunha et al. 2013; Gauthier and Keller 2013; Samerpitak et al. 2014). However, it is not always easy to identify plant pathogenic fungi to the species level, as they often lack discriminatory morphological characters or cultivable life stages (Kang et al. 2010; Udayanga et al. 2012). Molecular (DNA sequence) data have emerged as a key resource in the identification of plant pathogenic fungi and carry the benefit that all fungi, regardless of life stage, morphological plasticity, and degree of cultivability, can be analyzed (Shenoy et al. 2007; Sharma et al. 2013). As a result, recent years have seen substantial progress towards a comprehensive understanding of phytopathogenic fungi in terms of taxonomy, systematics, and ecology (Dean et al. 2012; Maharachchikumbura et al. 2012; Manamgoda et al. 2012; Woudenberg et al. 2013).

DNA data, however, are not a panacea for species identification. On the contrary, taxonomically and technically compromised DNA sequences are common in the international nucleotide sequence databases (Bidartondo et al. 2008; Kang et al. 2010). This makes their use as reference data for molecular species identification difficult, particularly because many users of newly generated sequence data may not be in a position to assess whether a proposed taxonomic affiliation is reliable. As a consequence, errors and mistakes propagate over time as users adopt incorrect species names and ecological properties retrieved from sequence similarity searches (Ko Ko et al. 2011; Nilsson et al. 2012). This is especially problematic for phytopathogens, where even closely related species may differ dramatically in terms of pathogenicity, host preference, and effective countermeasures (e.g., Barnes et al. 2004; Queloz et al. 2011). Although end users do have options to propose changes in the data and metadata in the public sequence databases, few users take action when they encounter compromised sequences (Pennisi 2008; Nilsson et al. 2012).

Molecular identification of fungi usually relies, at least in the first attempts, on sequencing the nuclear ribosomal internal transcribed spacer (ITS) region, the formal fungal barcode (Schoch et al. 2012). The largest database tailored for fungal ITS sequences is UNITE (http://unite.ut.ee; Abarenkov et al. 2010a). UNITE mirrors and curates the International Nucleotide Sequence Database Collaboration (INSDC:

GenBank, ENA, and DDBJ; Nakamura et al. 2013) for fungal ITS sequences and offers extensive capacities for analysis and third-party annotation of sequences to its users. It has been the subject of several annotation efforts (Tedersoo et al. 2011; Bengtsson-Palme et al. 2013; Kõljalg et al. 2013), but these have in part been biased towards basidiomycetes and mycorrhizal fungi. A similar effort for plant pathogenic fungi was initiated at the symposium "Classical and molecular approaches in plant pathogen taxonomy" (10–11 September 2013, Warsaw). In addition to several of the symposium participants, other experts on various fungal lineages known to harbour plant pathogens were invited as contributors through personal networking, email, and ResearchGate (http://www.researchgate.net/). Several experts on epiphytic and endophytic fungi also participated in the effort; while these fungi may not be plant pathogenic, they are often isolated alongside, or mistaken for, plant pathogenic fungi (Unterseher et al. 2013). Moreover, many fungi showing pathogenicity in certain plants represent common endophytes in other host plants (Delaye et al. 2013). This paper reports on the outcome of the annotation effort.

Materials and methods

Using third-party sequence annotation facilities provided by the PlutoF workbench (http://plutof.ut.ee, Abarenkov et al. 2010b), the participants examined fungal lineages and ecological groups of their respective expertise in UNITE for four parameters: (i) selection of representative sequences for species, (ii) improvement of taxonomic annotations, (iii) addition of ecological metadata (chiefly host and country of collection), and (iv) compromised sequence data.

(i) Selection of representative sequences for species

UNITE clusters all public fungal ITS sequences to approximately the genus/subgenus level. A second round of clustering inside each such cluster seeks to produce molecular operational taxonomic units at approximately the species level; these are called *species hypotheses* (SHs; Fig. 1; Kõljalg et al. 2013). The species hypotheses are open for viewing and querying (http://unite.ut.ee/SearchPages.php) through uniform resource identifiers (URIs) such as "http://unite.ut.ee/sh/SH158651.06FU". As a proxy for the species hypothesis, a representative sequence is chosen automatically from the most common sequence type in the species hypothesis. Through these representative sequences, UNITE assigns a unique, stable name of the accession number type – SH158651.06FU in its shortest form for the example above – to all species hypotheses to provide a means for unambiguous reference to species-level lineages even in the absence of formal Latin names. The representative sequences are also used for non-redundant BLAST databases for molecular identification in several next-generation sequencing analysis pipelines. Depending on the algorithm, including all available fungal ITS sequences in the reference database slows down sequence similarity searches significantly, and the use of downsized, non-redundant databases with only one sequence per taxon of interest is a common solution. The representative sequences of UNITE fulfill these criteria, since they comprise a single sequence from all fungal species hypotheses recovered to date

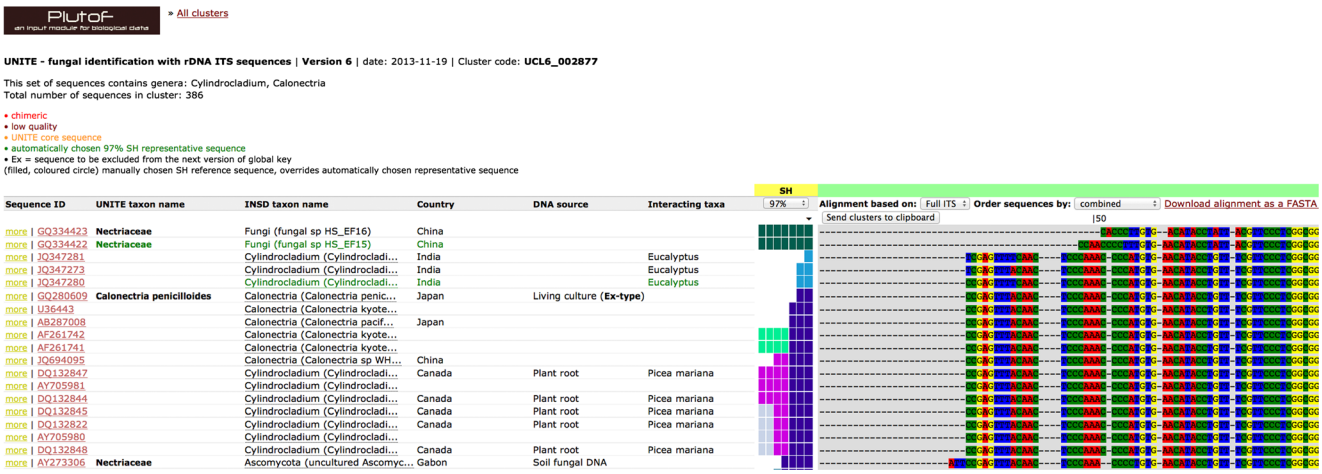

Fig. 1 A screenshot from the web-based PlutoF sequence management environment showing a *Nectriaceae* cluster, with the individual species hypotheses at different similarity levels indicated by the coloured vertical bars. Country of collection and host/interacting taxa are specified together with taxonomic re-annotations. Sequences from type material are indicated. For species hypotheses where no user has designated a reference sequence, the clustering program chooses a sequence from the most common sequence type to represent that species hypothesis (*shown in green font*). The species hypotheses are mirrored by GenBank through a LinkOut function, making it possible to go from a BLAST search in GenBank to the corresponding species hypothesis in UNITE through a single click

through ITS sequences by the scientific community. However, there are situations where one would like to influence which sequence is chosen to represent a species hypothesis. In ideal cases, the type specimen or an ex-type culture has been sequenced. Such "type sequences" form the best possible proxy for the species hypothesis, as long as they are sufficiently long and of high technical quality.

To increase the proportion of plant pathology-related fungal taxa represented by sequences from types, we scanned the 27 largest journals in plant pathology (and 12 mycological journals known for an inclination towards plant pathology or fungi otherwise associated with plants) for descriptions of new (or typifications of existing) plant pathogenic or plant-associated species of fungi (Supplementary Item 1). For all descriptions where an ITS sequence was generated from the type specimen/ex-type culture by the original authors, we examined the sequence in the corresponding UNITE cluster for read quality and length. All type sequences deemed to be of high technical quality and sufficient length were designated as reference sequences for their respective species hypothesis.

(ii) Correction of taxonomic affiliations

Taxonomic misidentifications are rife in the public nucleotide sequence databases. Similarly, more than half of all public fungal ITS sequences are not annotated to the level of species, and most of these carry little or no taxonomic annotation save, e.g., "Uncultured fungus" (cf. Hibbett et al. 2011). This makes molecular identification difficult and can lead to an incorrect name or no name at all, even when full (e.g., *Colletotrichum melonis*) or partial (e.g., *Colletotrichum* sp. or Glomerellales) naming would have been possible. Clearly it is important to avoid the common mistake of over-estimating taxonomic certainty based solely on BLAST searches, which often yield many top hits with similar quality scores and can obscure sister-level relationships to the taxa represented in the top matches. BLAST results may also differ over time according to database content, and differ markedly when, e.g., the full ITS vs. partial ITS sequences or ITS sequences with non-trivial lengths of the ribosomal small and/or large subunits for the same strain are submitted to searches (U'Ren et al. 2009). Indeed, a substantive portion of misidentified sequences in public databases appear to have resulted from spurious applications of taxonomic names to sterile mycelia, environmental samples, or otherwise unknown strains, often being studied by non-taxonomists. However, careful evaluation of database matches can provide additional information about taxonomic placement that can be applied judiciously by experts to better serve the scientific community. In addition, sequences without taxonomic annotations (e.g., "Uncultured fungus") are often unfairly disregarded in phylogenetic studies (Nilsson et al. 2011). Another reason to improve the taxonomic annotation of public ITS sequences is therefore to highlight their existence and availability for use in phylogenetic and systematic studies. Such enhanced taxon sampling carries many advantages (Heath et al. 2008). We scanned our fungal lineages of expertise in UNITE to make sure the sequences carried the most accurate name possible, viz. the full species name for fully identified sequences, and the genus, family, order, class, or phylum name for sequences that could not be fully assigned.

(iii) Addition of geographical and ecological metadata

Although DNA sequences form the core of molecular identification of fungi, additional data are often needed for final, informed decisions on the taxonomic affiliation of newly generated sequences. For plant pathogenic fungi, the identity of the host and the geographical origin of the sequences are often critical information (Britton and Liebhold 2013). Yet these metadata are usually not included with sequence data in public sequence databases; Tedersoo et al. (2011) showed, for instance, that a modest 43 % of the public fungal ITS sequences were annotated with the country of origin. To the same effect, Ryberg et al. (2009) found that host of collection was reported for less than 25 % of all public fungal ITS sequences (although not all fungi necessarily have a host). We made sure that the sequences of our core expertise were as richly annotated as possible in UNITE through recursions to the original publications.

(iv) Technical quality of sequences

Detecting sequences of substandard quality in public databases is difficult because sequence chromatograms or other original data are not present for verification of nucleotide identity, and sequencing technologies have different error rates and types of errors (e.g., 454 pyrosequencing vs. Sanger sequencing). Standards also differ among researchers and computer programs with regard to quality thresholds and what is deemed acceptable for individual nucleotides or whole-sequence reads. The extent to which sequence depositors take measures to ensure that their sequence data are of satisfactory integrity also seems to differ markedly. To discriminate with full certainty among publicly deposited sequences of high and substandard quality is simply not possible in all situations (Nilsson et al. 2012). To remove all sequences that are putatively substandard is certain to lead to many instances of false-positive removals (i.e., removal of authentic albeit poorly known biodiversity), and in this study we settled for removing entries we could prove were compromised. We evaluated sequence quality on the basis of length, evidence of chimera formations or poor read quality, and mislabelling of the genetic marker that the data represent.

Results

The participants implemented a total of 31,954 changes, including 5,135 taxonomic re-annotations, 25,028 specifications of geographical and ecological metadata, 1,368 designations of reference sequences, and 401 exclusions of substandard sequences, distributed over some 48 fungal orders. The results were incorporated in UNITE for all its users. In addition, they are made publicly available through the UNITE release of all public fungal ITS sequences (http://unite.ut.ee/repository.php) for use in, e.g., local sequence similarity searches and sequence processing pipelines such as QIIME (Caporaso et al. 2010; Bates et al. 2013), mothur (Schloss et al. 2009), SCATA (http://scata.mykopat.slu.se/), CREST (Lanzén et al. 2012), and other downstream applications. UNITE also serves as one of the data providers for BLAST (Altschul et al. 1997) searches in the EUBOLD fungal barcoding database (http://www.cbs.knaw.nl/eubold/).

(i) Selection of representative sequences for species

The extraction of sequences from type material from the literature resulted in 965 designations of reference sequences (for as many species hypotheses and a total of 194 genera of fungi; Table 1). We also designated 403 additional reference sequences based on our expertise; 174 of these stemmed from type material and 229 were from other authentic material. The latter cases involved fungal taxa of our core expertise where we knew the type material was missing or too old for DNA sequencing and where we knew that the selected sequences were as close to the type as possible in terms of morphology, country, and/or substrate of collection. A total of 202 genera were designated with at least one reference sequence.

(ii) Correction of taxonomic affiliations

The process of verifying taxonomic names given to sequences resulted in a total of 5,135 changes (Table 1), notably for the orders Hypocreales (459 changes), Glomerellales (404 changes), and Botryospheriales (393 changes). In addition, 22 ITS sequences were found to stem from kingdoms other than Fungi and were re-annotated accordingly.

(iii) Addition of geographical and ecological metadata

Our effort to complement the sequences with metadata from the literature resulted in a total of 14,478 specifications of host and 10,550 specifications of country of origin (Table 1).

Table 1 Summary of the changes made in the UNITE database. The 15 orders that saw the largest number of changes are specified separately; all other lineages are amalgamated into the "Others" category

Order	Taxonomic re-annotations	Country	Host	Reference sequences	Count
Hypocreales	459	3,751	2,960	118 (116)	7,288
Pleosporales	129	860	4,344	76 (76)	5,409
Capnodiales	200	960	1,696	181 (181)	3,037
Diaporthales	79	1,374	855	28 (28)	2,336
Glomerellales	404	814	824	148 (148)	2,190
Botryosphaeriales	393	428	626	70 (67)	1,517
Mucorales	90	630	631	87 (63)	1,438
Eurotiales	420	411	226	168 (168)	1,225
Xylariales	90	225	823	19 (19)	1,157
Helotiales	333	301	290	108 (46)	1,032
Chaetothyriales	22	121	521	17 (17)	681
Puccinales	134	313	194	9 (1)	650
Agaricales	442	31	8	21 (21)	502
Pezizales	297	0	97	1 (1)	395
Erysiphales	143	55	66	129 (4)	393
Others	1,500	276	317	188 (183)	2,281

Taxonomic re-annotations = The number of taxonomic (re)annotations implemented. Country = The number of specifications of country of collection. A total of 94 different countries were added. Host = The number of host specifications added in the system. Reference sequences = The number of reference sequences designated through manual inspection (of which sequences from type material are indicated in parentheses). Count = Total number of changes

(iv) Technical quality of sequences

We detected a total of 363 sequences of substandard technical quality. These were marked as compromised, which precludes them from being used in molecular identification procedures while still keeping them open to direct searches in the system. This included 84 cases of chimeric sequences and 279 cases of low read quality. Another 38 sequences were annotated as ITS sequences by their submitters but were found to represent other genes and markers (notably the ribosomal small and large subunits) and were re-annotated accordingly.

Discussion

Fungal pathogens of agricultural, silvicultural, horticultural, and wild plants can compromise ecosystem health and cause considerable economic loss globally. Correct identification of these fungi and subsequent understanding of their biology and ecology are key elements in protecting their host plants (Rossman and Palm-Hernández 2008). However, identification of plant pathogenic fungi to the species level is relevant to more than just studies of plant pathology. Because of the ease and moderate cost at which large amounts of sequence data can be generated, fungi and fungal communities are now being studied by an increasing number of non-mycologists, notably soil biologists, molecular ecologists, and researchers in the medical sciences (e.g., Ghannoum et al. 2010; La Duc et al. 2012; Pautasso 2013). Phytopathogenic fungi also occur in these substrates and ecosystems in various life stages, including sterile mycelia, resting stages, and propagules. Although some plant pathogenic fungi have been studied in great detail, the biology of the majority of phytopathogenic fungi remains poorly known. Therefore, information stemming from non-mycological or non-pathological research efforts may increase our understanding of these taxa. As a consequence, it is important that all researchers, regardless of expertise and extent of mycological knowledge, can obtain reliable estimates of the taxonomic identity of plant pathogenic – and all other – fungi in whatever form they are recovered.

Molecular identification of plant pathogenic fungi can be challenging due to differing sequence and annotation quality of the available reference sequences. We have gone through a large number of plant pathogenic fungal groups within our collective expertise. A total of 31,954 changes in 48 fungal orders were implemented in UNITE for these groups (Table 1). However, not all plant pathogenic lineages of fungi – or, indeed, even the groups covered by the present effort – are satisfactorily resolved in UNITE. In addition, new sequences (of both known and unknown species) are continuously generated and deposited in the INSDC by the scientific community, such that a limited group of people can never stay abreast of the data deposition. A community effort is clearly required. UNITE offers third-party annotation capacities to all its registered users. Registration is free, and contributions from all relevant scientific communities are most welcome. Even small edits – such as designating a reference sequence for a single species hypothesis, correcting and improving a handful of taxonomic annotations, or adding metadata that can be used for comparative studies (Supplementary Item 2) – will improve the database significantly and may be of substantial importance to other researchers. Going through the alignments and metadata for one's fungi of expertise in the web-based system is furthermore a good way to visualize and explore patterns in the data and identify new research questions.

Many of the corrections brought about by the present effort would have been unnecessary if the original sequence authors had taken the time to examine and annotate their sequences properly prior to submission. Lack of time and awareness of these issues are the presumed culprits. Guidelines on how to process newly generated sequences in a way to establish their integrity and maximize their usefulness to the scientific community are given in Seifert and Rossman (2010), Nilsson et al. (2012), Hyde et al. (2013), and Robbertse et al. (2014). In addition, to facilitate future assessments of sequence quality and other pursuits, we urge sequence depositors in INSDC to archive chromatograms and other relevant data in UNITE or in other resources that support long-term data storage and availability. The present initiative will contribute to more accurate molecular identification of plant pathogenic fungi for three sets of users: UNITE users, anyone using the ~350,000-sequence downloadable FASTA file of the UNITE/INSDC fungal ITS sequences (http://unite.ut.ee/repository.php) for local BLAST searches or similar, and researchers using any of the major next-generation sequencing analysis pipelines or the EUBOLD database to process newly generated fungal ITS datasets. In addition, following the data sharing history between UNITE and the INSDC, the results were made available to the INSDC to reach the widest possible audience. Fungal barcoding is in a state of constant development, but it should be clear that collaboration and data sharing among resources are necessary for the future development of the field. Mycology struggles for funding in competition with fields that are often deemed larger or more fashionable, and we simply cannot afford public fungal DNA sequences to remain in a suboptimal state. On the contrary, we hope mycologists will work together to make fungal sequence data as richly annotated and as easily interpreted as possible because, after all, many of the end users of those data will not be mycologists. The present study is a small step in that direction, and we hope that others will follow.

Acknowledgments RHN acknowledges financial support from Swedish Research Council of Environment, Agricultural Sciences, and Spatial Planning (FORMAS, 215-2011-498). ArA acknowledges financial support from European Funds through COMPETE and by National Funds through the Portuguese Foundation for Science and Technology (FCT) within projects PTDC/AGR-FOR/3807/2012 - FCOMP-01-0124-FEDER-027979 and PEst-C/MAR/LA0017/2013. SB is supported by National Science Foundation Grant DBI 1046115. The Austrian Centre of Industrial Biotechnology (ACIB) contribution (BH) was supported by FFG, BMWFJ, BMVIT, ZIT, Zukunftsstiftung Tirol, and Land Steiermark within the Austrian COMET program FFG Grant 824186. Financial support to JP was partially provided by the Polish Ministry of Science and Higher Education (MNiSW), grant no. NN303_548839. OLP acknowledges financial support from FAPEMIG and CNPq. TMP was funded by the Government of Canada through Genome Canada and the Ontario Genomics Institute through the Biomonitoring 2.0 project (OGI-050). The GenBank staff is acknowledged for helpful discussions and data sharing. The NEFOM network is acknowledged for infrastructural support. The authors have no conflict of interests to report.

References

Abarenkov K, Nilsson RH, Larsson K-H, Alexander IJ, Eberhardt U, Erland S, Høiland K, Kjøller R, Larsson E, Pennanen T, Sen R, Taylor AFS, Tedersoo L, Ursing BM, Vrålstad T, Liimatainen K, Peintner U, Kõljalg U (2010a) The UNITE database for molecular identification of fungi - recent updates and future perspectives. New Phytol 186:281–285

Abarenkov K, Tedersoo L, Nilsson RH, Vellak K, Saar I, Veldre V, Parmasto E, Prous M, Aan A, Ots M, Kurina O, Ostonen I, Jõgeva J, Halapuu S, Põldmaa K, Toots M, Truu J, Larsson K-H, Kõljalg U (2010b) PlutoF - a web-based workbench for ecological and taxonomic research, with an online implementation for fungal ITS sequences. Evol Bioinform 6:189–196

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402

Barnes I, Crous PW, Wingfield BD, Wingfield MJ (2004) Multigene phylogenies reveal that red band needle blight is caused by two distinct species of *Dothistroma*, *D. septosporum* and *D. pini*. Stud Mycol 50:551–565

Bates ST, Ahrendt S, Bik HM, Bruns TD, Caporaso JG, Cole J, Dwan M, Fierer N, Gu D, Houston S, Knight R, Leff J, Lewis C, Maestre JP, McDonald D, Nilsson RH, Porras-Alfaro A, Robert V, Schoch C, Scott J, Taylor DL, Wegener Parfrey L, Stajich JE (2013) Meeting report: fungal ITS workshop (October 2012). Stand Genomic Sci 8: 118–123

Bengtsson-Palme J, Ryberg M, Hartmann M, Branco S, Wang Z, Godhe A, De Wit P, Sánchez-García M, Ebersberger M, de Sousa F, Amend A, Jumpponen A, Unterseher M, Kristiansson E, Abarenkov K, Bertrand YJK, Sanli K, Eriksson MK, Vik U, Veldre V, Nilsson RH (2013) Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data. Methods Ecol Evol 4:914–919

Bidartondo M, Bruns TD, Blackwell M et al (2008) Preserving accuracy in GenBank. Science 319:5870

Britton KO, Liebhold AM (2013) One world, many pathogens! New Phytol 197:9–10

Caporaso JG, Kuczynski J, Stombaugh J et al (2010) QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7:335–336

Cunha KCD, Sutton DA, Fothergill AW, Gené GJ, Cano J, Madrid H, Hoog SD, Crous PW, Guarro J (2013) In vitro antifungal susceptibility and molecular identity of 99 clinical isolates of the opportunistic fungal genus *Curvularia*. Diagn Microbiol Infect Dis 76:168–174

Dean R, Van Kan JA, Pretorius ZA, Hammond-Kosack KE, Di Pietro A, Spanu PD, Rudd JJ, Dickman M, Kahmann R, Ellis J, Foster GD (2012) The top 10 fungal pathogens in molecular plant pathology. Mol Plant Pathol 13:414–430

Delaye L, García-Guzmán G, Heil M (2013) Endophytes versus biotrophic and necrotrophic pathogens – are fungal lifestyles evolutionarily stable traits? Fungal Divers 60:125–135

Fisher MC, Henk DA, Briggs CJ, Brownstein JS, Madoff LC, McCraw SL, Gurr SJ (2012) Emerging fungal threats to animal, plant and ecosystem health. Nature 484:186–194

Gauthier G, Keller N (2013) Crossover fungal pathogens: the biology and pathogenesis of fungi capable of crossing kingdoms to infect plants and humans. Fungal Genet Biol 61:146–57

Ghannoum MA, Jurevic RJ, Mukherjee PK, Cui F, Sikaroodi M, Naqvi A, Gillevet PM (2010) Characterization of the oral fungal microbiome (mycobiome) in healthy individuals. PLoS Pathog 6: e1000713

Heath TA, Hedtke SM, Hillis DM (2008) Taxon sampling and the accuracy of phylogenetic analyses. J Syst Evol 46:239–257

Hibbett DS, Ohman A, Glotzer D, Nuhn M, Kirk P, Nilsson RH (2011) Progress in molecular and morphological taxon discovery in fungi and options for formal classification of environmental sequences. Fungal Biol Rev 25:38–47

Hyde KD, Udayanga D, Manamgoda DS, Tedersoo L, Larsson E, Abarenkov K, Bertrand YJK, Oxelman B, Hartmann M, Kauserud H, Ryberg M, Kristiansson E, Nilsson RH (2013) Incorporating molecular data in fungal systematics: a guide for aspiring researchers. Curr Res Environ Appl Mycol 3:1–32

Kang S, Mansfield MAM, Park B, Geiser DM, Ivors KL, Coffey MD, Grünwald NJ, Martin FN, Lévesque CA, Blair JE (2010) The promise and pitfalls of sequence-based identification of plant pathogenic fungi and oomycetes. Phytopathology 100:732–737

Ko Ko TWK, Stephenson SL, Bahkali AH, Hyde KD (2011) From morphology to molecular biology: can we use sequence data to identify fungal endophytes? Fungal Divers 50:113–120

Kõljalg U, Nilsson RH, Abarenkov K et al (2013) Towards a unified paradigm for sequence-based identification of Fungi. Mol Ecol 22: 5271–5277

Kowalski T, Holdenrieder O (2009) The teleomorph of *Chalara fraxinea*, the causal agent of ash dieback. For Pathol 39:304–308

La Duc MT, Vaishampayan P, Nilsson RH, Torok T, Venkateswaran K (2012) Pyrosequencing-derived bacterial, archaeal, and fungal diversity of spacecraft hardware destined for Mars. Appl Environ Microbiol 78:5912–5922

Lanzén A, Jørgensen SL, Huson DH, Gorfer M, Grindhaug SH, Jonassen I, Øvreås L, Urich T (2012) CREST – classification resources for environmental sequence tags. PLoS One 7:e49334

Maharachchikumbura SSN, Guo LD, Cai L, Chukeatirote E, Wu WP, Sun X, Crous PW, Bhat DJ, McKenzie EHC, Bahkali AH, Hyde KD (2012) A multi-locus backbone tree for *Pestalotiopsis*, with a polyphasic characterization of 14 new species. Fungal Divers 56:95–129

Manamgoda DS, Cai L, McKenzie EHC, Crous PW, Madrid H, Chukeatirote E, Shivas RG, Tan YP, Hyde KD (2012) A phylogenetic and taxonomic re-evaluation of the *Bipolaris*, *Cochliobolus*, *Curvularia* complex. Fungal Divers 56:131–144

Nakamura Y, Cochrane G, Karsch-Mizrachi I (2013) The international nucleotide sequence database collaboration. Nucleic Acids Res 41: D21–D24

Nilsson RH, Ryberg M, Sjökvist E, Abarenkov K (2011) Rethinking taxon sampling in the light of environmental sequencing. Cladistics 27:197–203

Nilsson RH, Tedersoo L, Abarenkov K, Ryberg M, Kristiansson E, Hartmann M, Schoch CL, Nylander JAA, Bergsten J, Porter TM, Jumpponen A, Vaishampayan P, Ovaskainen O, Hallenberg N, Bengtsson-Palme J, Eriksson KM, Larsson K-H, Larsson E (2012) Five simple guidelines for establishing basic authenticity and reliability of newly generated fungal ITS sequences. MycoKeys 4:37–63

Pautasso M (2013) Fungal under-representation is (slowly) diminishing in the life sciences. Fungal Ecol 6:129–135

Pennisi E (2008) "Proposal to 'wikify' GenBank meets stiff resistance". Science 319:1598–1599

Queloz V, Grunig CR, Berndt R, Kowalski T, Sieber TN, Holdenrieder O (2011) Cryptic speciation in *Hymenoscyphus albidus*. For Pathol 41: 133–142

Robbertse B, Schoch CL, Robert V et al. (2014) Finding needles in haystacks: linking scientific names, reference specimens and molecular data for *Fungi*. Database, in press

Rossman AY, Palm-Hernández ME (2008) Systematics of plant pathogenic fungi: why it matters. Plant Dis 10:1376–1386

Ryberg M, Kristiansson E, Sjökvist E, Nilsson RH (2009) 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 181:471–477

Samerpitak K, Van der Linde E, Choi HJ, Gerrits van den Ende AHG, Machouart M, Gueidan C, de Hoog GS (2014) Taxonomy of *Ochroconis*, a genus including opportunistic pathogens on humans and animals. Fungal Divers 65:89–126. doi:10.1007/s13225-013-0253-6

Schloss PD, Westcott SL, Ryabin T et al (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75:7537–7541

Schoch CL, Seifert KA, Huhndorf S et al (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for fungi. Proc Natl Acad Sci U S A 109:6241–6246

Seifert K, Rossman AY (2010) How to describe a new fungal species. IMA Fungus 1:109–116

Sharma G, Kumar N, Weir BS, Hyde KD, Shenoy BD (2013) Apmat gene can resolve *Colletotrichum* species: a case study with *Mangifera indica*. Fungal Divers 61:117–138

Shenoy BD, Rajesh J, Hyde KD (2007) Impact of DNA sequence-data on the taxonomy of anamorphic fungi. Fungal Divers 26:1–54

Stajich JE, Berbee ML, Blackwell M, Hibbett DS, James TY, Spatafora JW, Taylor JW (2009) The fungi. Curr Biol 19:R840–R845

Tedersoo L, Abarenkov K, Nilsson RH, Schussler A, Grelet G-A, Kohout P, Oja J, Bonito GM, Veldre V, Jairus T, Ryberg M, Larsson K-H, Kõljalg U (2011) Tidying up international nucleotide sequence databases: ecological, geographical, and sequence quality annotation of ITS sequences of mycorrhizal fungi. PLoS One 6:e24940

Udayanga D, Liu XX, Crous PW, McKenzie EHC, Chukeatirote E, Hyde KD (2012) A multi-locus phylogenetic evaluation of *Diaporthe* (*Phomopsis*). Fungal Divers 56:157–171

Unterseher M, Peršoh D, Schnittler M (2013) Leaf-inhabiting endophytic fungi of European Beech (*Fagus sylvatica* L.) co-occur in leaf litter but are rare on decaying wood of the same host. Fungal Divers 60: 43–54

U'Ren JM, Dalling JW, Gallery RE, Maddison DR, Davis EC, Gibson CM, Arnold EA (2009) Diversity and evolutionary origins of fungi associated with seeds of a neotropical pioneer tree: a case study for analyzing fungal environmental samples. Mycol Res 113:432–449

Woudenberg JHC, Groenewald JZ, Binder M, Crous PW (2013) *Alternaria* redefined. Stud Mycol 75:171–212