DOI 10.1002/pmic.201600034 *Proteomics* 2016, *16*, 2454–2460

VIEWPOINT

Strategies to improve usability and preserve accuracy in biological sequence databases

Johan Bengtsson-Palme[1,2], Fredrik Boulund[2,3,4], Robert Edström[3,5], Amir Feizi[6], Anna Johnning[2,3], Viktor A. Jonsson[3], Fredrik H. Karlsson[6], Chandan Pal[1,2], Mariana Buongermino Pereira[2,3], Anna Rehammar[3], José Sanchez[3,7]**, Kemal Sanli[8] and Kaisa Thorell[4]*

[1] Department of Infectious Diseases, Institute of Biomedicine, the Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden
[2] Centre for Antibiotic Resistance Research (CARe), University of Gothenburg, Gothenburg, Sweden
[3] Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, Gothenburg, Sweden
[4] Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Stockholm, Sweden
[5] Department of Computer Science and Engineering, Chalmers University of Technology, Gothenburg, Sweden
[6] Department of Biology and Biological Engineering, Chalmers University of Technology, Gothenburg, Sweden
[7] Bioinformatics Core Facility, University of Gothenburg, Gothenburg, Sweden
[8] Department of Biology and Environmental Sciences, University of Gothenburg, Gothenburg, Sweden

Received: February 12, 2016
Revised: July 25, 2016
Accepted: August 11, 2016

Biology is increasingly dependent on large-scale analysis, such as proteomics, creating a requirement for efficient bioinformatics. Bioinformatic predictions of biological functions rely upon correctly annotated database sequences, and the presence of inaccurately annotated or otherwise poorly described sequences introduces noise and bias to biological analyses. Accurate annotations are, for example, pivotal for correct identification of polypeptide fragments. However, standards for how sequence databases are organized and presented are currently insufficient. Here, we propose five strategies to address fundamental issues in the annotation of sequence databases: (i) to clearly separate experimentally verified and unverified sequence entries; (ii) to enable a system for tracing the origins of annotations; (iii) to separate entries with high-quality, informative annotation from less useful ones; (iv) to integrate automated quality-control software whenever such tools exist; and (v) to facilitate postsubmission editing of annotations and metadata associated with sequences. We believe that implementation of these strategies, for example as requirements for publication of database papers, would enable biology to better take advantage of large-scale data.

Keywords:
Annotation / Bioinformatics / Databases / Functional prediction / Sequencing / Standards

1 Introduction

During the last decade, many biological disciplines have become increasingly dependent on efficient bioinformatics. The advent of large-scale analysis techniques, including MS-based proteomics and massively parallel DNA sequencing, has driven costs down dramatically, and we have now entered an era in which data generation is easier and less costly than the analysis of the data generated [1]. This rapid increase in available data has created a great demand for advanced software for crunching enormous amounts of sequence data,

Correspondence: Dr. Johan Bengtsson-Palme, Department of Infectious Diseases, Institute of Biomedicine, the Sahlgrenska Academy, University of Gothenburg, Guldhedsgatan 10, SE-413 46 Gothenburg, Sweden
E-mail: johan.bengtsson-palme@microbiology.se

*Current address: Dr. Fredrik H. Karlsson, Metabogen AB, SE-411 26 Gothenburg, Sweden;
**Current address: Dr. José Sanchez, AstraZeneca, SE-431 83 Mölndal, Sweden
Colour Online: See the article online to view Fig. 1 in colour.

fast and powerful computer infrastructure, and accurate reference datasets to make comparisons to. While efforts to develop efficient software have been—and still are—very fruitful, and many labs have access to sophisticated computer infrastructure, the databases used are often defective in terms of accuracy, and several studies have shown fallacies in public databases [2–5]. Increasingly widespread use of more or less fully automated bioinformatic pipelines for proteome, genome, and metagenome annotation likely also introduces errors and escalates the rate of error propagation in large sequence databases [2, 6]. Since nearly all bioinformatic endeavors depend on that the fundamental annotation of sequences is correct, the presence of poorly—or even erroneously—annotated sequences is disturbing. Before high-throughput techniques were widely applied in biology, every individual gene and its products received thorough investigation, but the total amount of available information was scarce. Today, we face the opposite problem, with millions of annotated sequence entries, but little indication of which ones that actually have experimental support. This article outlines some current challenges related to public sequence repositories, and describes five possible strategies to improve the accuracy of new (and existing) sequence databases (Table 1). These can be implemented individually or collectively depending on the specific database use cases.

2 Distinguish between experimentally proven and unverified entries

Biological sequence databases constitute collections of nucleotide or protein sequences and knowledge (metadata) associated with them. This knowledge can either be derived from actual experimental data verifying the function of a sequence, or from bioinformatic predictions made mainly on the basis of sequence similarity. In many databases, such as the Antibiotic Resistance Genes Database [7] and the NONCODE database of noncoding RNAs [8], a distinction between experimentally verified and computationally predicted sequences does not exist, leaving the researcher in the dark about whether an entry in the database have any experimental support or not. These are not isolated cases; even among sequences in the curated segments of larger reference databases, such as the RefSeq part of GenBank [9], dubious and noninformative annotations are carried over from their noncurated counterparts. This is partly due to that curators have little ability to know what information that is appropriately supported for every single entry. Because of this, it is often impossible for researchers to know the validity of a given functional prediction.

However, even in large databases it is possible to integrate information regarding whether a sequence entry is experimentally verified or not. UniProt, for instance, has already implemented an efficient system that enables distinction between verified and unverified protein sequences at different evidence levels, such as "experimental evidence at the protein or transcript level," "inferred from homology," "predicted," and "uncertain" [10]. We suggest that use of a similar system, preferably based around ECO evidence codes [11], should be required for new databases, implemented as a prerequisite for publishing papers describing databases. New databases should be designed such that the division between experimentally verified and unverified entries is clear. A distinct division would make it considerably easier for users to evaluate the quality of their own annotations. Some databases already separate between these two categories of entries [12, 13], but this practice is still far from the norm. For this to become routine, thorough literature searches will be needed to underpin the knowledge associated with sequence entries in each database [14]. Thus, a complete verified/predicted separation may only be feasible for focused database projects, which by their nature already require substantial literature support to be of any value compared to larger sequence repositories.

3 Introduce tracing systems for sequence annotations

A closely related problem is that an increasing number of sequences tend to have their annotation inferred from similarity to another gene or protein, that in turn has its function deduced by similarity to a third, and so on. This makes it troublesome—often impossible—to trace down the experimental origin for the functional prediction of a specific entry in a database. This creates an infrastructure that begs for propagation of errors. For example, assume that a query sequence matches a database sequence with 80% identity, and that this sequence was annotated from another one with 80% identity, which in turn was annotated using a match to a third, experimentally verified sequence, with 70% identity. In this case, the actual conservation between the query sequence and the closest experimentally verified sequence might be very different from what would be anticipated given, for example, the output of a BLAST search (Fig. 1). Since a small number of amino acid changes—or sometimes even a single changed residue—can alter the substrate preferences [15, 16], binding sites [17–19], or even the overall functions of particular proteins [20, 21], such differences in percentage identity may be highly important for precise and accurate annotation. Furthermore, queries only matching part of a database sequence might be mistakenly annotated with the function of

Table 1. Strategies for preserving database accuracy

Proposed strategy	Action by database providers	Requirements for researchers and peer reviewers
(1) Clear division between experimentally verified and unverified entries	Incorporate already available data on experimental verification from, e.g., UniProtKB	Demand such a division for papers describing new databases to be accepted for publication
(2) Enable a tracing system for sequences to their closest experimentally verified entry	Incorporate a data field describing how each sequence was annotated and on which other sequences the annotation was based	Demand researches to provide the source of annotation when submitting sequences to sequence databases
(3) Separate entries with high-quality and low-quality annotation information	Separate "hypothetical" entries from actually annotated ones, preferably by a quality score for annotation information (Table 2)	Demand researchers to report the quality scores for each annotation
(4) Increased automated quality control for submissions to sequence databases	Employ appropriate automated annotation quality control tools for genes where such tools exist	Follow guidelines for database submission
(5) Postsubmission editing of annotations for entries in the databases	Change the way annotations of entries are maintained, enabling researchers to easily add and modify annotation data, and/or flagging entries as compromised	Follow suggested guidelines for annotation quality control. Possibly peer-review of sequence entries, e.g., before a sequence is moved into curated databases, such as RefSeq or UniProtKB

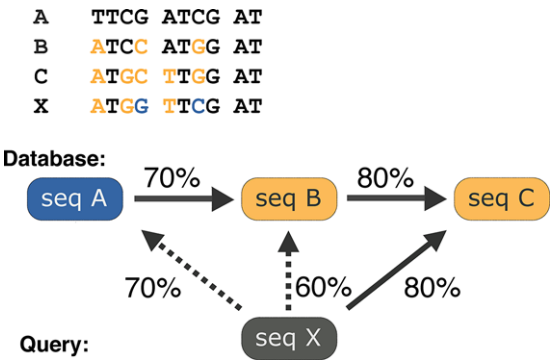

Figure 1. Query sequence X has a best-matching database sequence C (80% identity). Sequence C was annotated from sequence B, which in turn was annotated using the experimentally verified sequence A. In this particular case, the identity between the query sequence X and the experimentally verified sequence A is 70%, due to different substitutions having occurred between the sequences.

that database entry, even though the actual functional domain might be missing altogether from the query sequence [22]. The error rates for annotations based on sequence similarity have been estimated to be up to 49% in the worst cases [23]. It is thus difficult not only to determine if the database match to a given query sequence has any experimental evidence; but also how the annotation was otherwise determined, and if an error has been made somewhere in this undisclosed annotation chain.

It is clear that the problems associated with sequences being annotated by similarity to other unverified genes or proteins will persist, as it remains the most powerful tool we have to infer the functionality of predicted genes. To partially mitigate this problem, we propose a tracing system to enable tracking of the basis for each sequence entry's annotation, preferably in a stepwise fashion all the way back to a functionally verified sequence. This would make it possible to judge how well-supported a matching sequence in a similarity search is in terms of verified biological function. Additionally, applying such a stepwise tracing system would enable scoring of annotations according to reliability, for example based on the percent identity to the closest experimentally verified entry. Furthermore, it would be relatively easy to implement an automatic system that would immediately update the annotation trace of sequences with high identity to a newly submitted experimentally verified entry, and revise the reliability score accordingly. Such an automated system could also be used as a guidance tool to suggest groups of conserved database entries without a closely related experimentally verified homolog. Research could then be targeted toward verifying the function of entries where most knowledge can be gained. Importantly, we suggest that if such an automated update system is put into practice, the trace forming the basis for the original annotation should still be kept for reference purposes.

4 Separate entries with high-quality annotation information and hypothetical data

The ease of generating sequenced genomes and the recent flood of metagenomic datasets contribute to yet another concern with large databases; finding sequences with

Table 2. Proposal for a five-level scoring system for annotation information quality

Quality score	Criteria	Inference
1	No current annotation, or annotated as "unknown protein," "predicted protein," or "hypothetical protein" (or similar)	Automatic
2	Annotated as "conserved hypothetical protein," or "homologous to protein X"	Semi-automatic
3	Protein annotated with a function and/or belongs to a protein family with described function; proteins annotated using sequence motifs or hidden Markov models, such as Pfam domains	Automatic
4	Protein without experimentally verified function, but with, e.g., ≥90% sequence identity to the closest verified entry	Manual[a)]
5	Protein with experimentally verified function	Manual

a) This could be inferred automatically if the proposed annotation reliability score based on percentage identity to the closest experimentally verified database entry is implemented.

high-quality annotation information is becoming increasingly harder. The proportion of genes or proteins that are annotated to some extent, either automatically or manually, ranges in different genomes from around a few percent up to 66% for *Escherichia coli* [24]. However, finding potential genes (or open reading frames; ORFs) is easy compared to deducing their function, and therefore large quantities of unknown or hypothetical proteins have started to become enriched in the databases. This means that sequences with valid, informative, annotation often constitute a minority of the database, and thus get diluted in an ocean of hypothetical data of little or no interest. Additionally, unverified and spurious ORFs are often kept in the same databases as high-quality entries, making it complicated to distinguish between them.

In the same way as a clear separation between experimentally verified and nonverified sequence data would improve annotation quality, a separation of sequences with informative high-quality annotation (either with experimental evidence or inferred by sequence similarity, preferably with manual curation), and poorly annotated sequences (e.g., unknown and hypothetical entries) would help researchers to identify the most relevant entries, while still maintaining access to the world of yet unexplored proteins. We propose the use of a quality score for annotation information, ranging from 1 to 5 (Table 2). As a first step, all entries marked "hypothetical," "unknown," and so on, could be designated to have low-quality annotation information, and given a quality score of 1. If the separation of experimentally verified and nonverified entries is implemented, the verified records should generally be regarded as having high-quality annotation information, corresponding to a quality score of 5. Integration with the above-mentioned scoring system for annotation reliability could then guide which annotations would be considered of higher quality, and which would still to be regarded of lower value. Often, there would be substantial differences between looking at all entries regardless of the quality of annotation information compared to restricting a search to entries with high-quality annotation information only. This may be particularly important for the annotation of short polypeptide fragments, for which the precision of a similarity search is poorer due to the limited length of each fragment. Implementing a separation between entries of different information quality would not be overly difficult, as has been recently demonstrated by the UniProtKB, which now applies a system somewhat similar to the one we suggest [25].

5 Utilize software for automated quality control

Some of the problems mentioned above could easily be alleviated if well-working systems for improving and correcting annotations of database sequences existed. However, many databases are either reliant on the information provided upon sequence submission, or derive their sequence data from larger repositories such as GenBank. Since the start of these large database archives, submission systems have been designed in such a way that submitters themselves are primarily responsible for keeping annotation and sequence data accurate. This scheme, combined with insufficient personnel to perform efficient database curation—both in large and more focused database projects—make error correction and withdrawal of compromised entries slower than the pace that errors can propagate through the system. At the same time, submitting incorrect or erroneous information is, for many types of genes, as easy as submitting a good entry. Largely, this is due to that there is often no effective way for database managers to check the quality of submitted data at large scales. This makes it easier to create than to correct inaccuracies, which compromises many uses of centralised repositories for annotated sequence data.

To ease the issue with errors finding their way into databases, it is integral to implement automatic quality control of annotations in all situations when integrity can be easily checked. This is already employed for many barcoding genes, where a range of tools exist for delimiting if a sequence actually contains the supposed region [26, 27], is annotated in the appropriate organismal group [28], is chimeric [29, 30] or

is reversed [31]. Automated integrity control is fairly elemental to implement also for large-scale data, but only applies to a small fraction of sequence types. Similar tools directed at specific proteins of interest would clearly be desirable. An alternative in highly specialized databases containing few types of proteins would be to carry out automatic searches for conserved motifs and domains, and to check entries against the AntiFam database of spurious and compromised ORFs [32], which has already been adopted as a part of the quality control for Pfam [33]. Use of automated quality-control solutions could with little effort remove many of the worst annotation problems for well-studied genes.

6 Enable postsubmission editing of database entries

However, relying solely on automatic control schemes as a panacea to sequence annotation quality issues will certainly not be sufficient. Similarly, applying and enforcing guidelines for best practices regarding sequence handling before data submission [19, 34, 35] is not a complete remedy as it only applies to errors discovered *before* the submission (or inclusion) of sequence data. To account for fallacies unraveled after deposition in a database, we suggest a moderated Wiki-style annotation feature where researchers can contribute to and improve upon sequence annotations and other metadata, as well as flagging sequences as misannotated or dubious. Such a scheme already exists in the Rfam [36] and Pfam [33] databases, where it is even connected to Wikipedia. In both these instances, the user-contribution model has successfully improved the quality of annotations of biological data. There is little reason to believe that similar schemes could not work also for sequence annotations. This approach would also allow for a sort of peer-review of submitted sequences, which could easily be formalized as a sequence submission requirement where appropriate. For databases of smaller scale, active curation through submission systems might be sufficient to keep information accurate. However, since many database projects lack a clear long-term solution for maintenance beyond the current funding periods, smaller databases could potentially benefit from a Wiki-style editing system. Nevertheless, given the diversity of biological databases available, enforcing such an editing system as a publication requirement for papers describing databases would have consequences that are way too far-reaching to overview.

7 Conclusions and outlook

We have here outlined five strategies for improving the reliability of annotations in sequence databases to better accommodate for the flood of data generated by modern high-throughput platforms (Table 1). We believe that the increase in biological data generated will not cease in the near future—and that accuracy of annotation will become increasingly important to be able to draw biologically relevant conclusions. Ignoring the quality of annotation can severely impact the precision to which short polypeptides can be annotated, and can cause important biomarkers to be overlooked as seemingly irrelevant. Furthermore, construction of genome-scale networks, functional characterization of microbial communities, estimation of functional divergence in proteins, and characterization of newly sequenced organisms all depend on the reliability of reference databases. Failure to properly annotate sequences also affects fundamental properties of sequence entries such as protein and gene nomenclature. Indeed, the quality of the annotation itself should be taken at least as seriously as the quality of the generated raw data. It is also important that annotation data are presented in a manner that is easily accessible by both users and software tools. In this area, initiatives such as the PEFF format (http://www.psidev.info/node/363) for sequence entries are very important. In addition, personalized medicine will require database structures that allow a more protein-centric and per-individual view of sequences. Furthermore, mass spectrometry approaches to proteomics require specialized tools adapted for peptide-based similarity searches. All this brings additional set of problems to proteomics and other large-scale analyses. To cope with these upcoming challenges, it is fundamental to build a foundation of correctly annotated sequence data. While collaborative efforts can accomplish vast improvements in specific areas (see, e.g., [37]), employing minimal standards on annotation quality to preserve accuracy in public databases would greatly benefit this process. The Pfam and UniProt databases have in this respect acted as frontrunners in different areas, showing that changes such as those we propose are possible to implement even at large scales. Thus, it seems likely that smaller and more focused database projects would be able to implement at least some of the strategies suggested. We advocate that ultimately enforcement must be targeted at the publication process of papers describing databases and web services. The strategies proposed here could be used to form minimal criteria for publication of a database or related web service. This would foster a development of biological databases toward better accuracy, consistency, and accessibility, and thereby enable the road of excess to actually lead toward the palace of wisdom.

This article was conceived on a workshop arranged by the Gothenburg Bioinformatics Group for young scientists (GoBiG). We gratefully acknowledge support from the Gothenburg Bioinformatics Network (GOTBIN).

The authors have declared no conflict of interest.

8 References

[1] Hayden, E. C., Gene sequencing leaves the laboratory. *Nature* 2013, *494*, 290–291.

[2] Schnoes, A. M., Brown, S. D., Dodevski, I., Babbitt, P. C., Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput. Biol.* 2009, *5*, e1000605.

[3] Promponas, V. J., Iliopoulos, I., Ouzounis, C. A., Annotation inconsistencies beyond sequence similarity-based function prediction – phylogeny and genome structure. *Stand. Genomic Sci.* 2015, *10*, 108.

[4] Bidartondo, M. I., Preserving accuracy in GenBank. *Science* 2008, *319*, 1616.

[5] Nilsson, R. H., Ryberg, M., Kristiansson, E., Abarenkov, K. et al., Taxonomic reliability of DNA sequences in public sequence databases: a fungal perspective. *PLoS One* 2006, *1*, e59.

[6] Tripp, H. J., Hewson, I., Boyarsky, S., Stuart, J. M., Zehr, J. P., Misannotations of rRNA can now generate 90% false positive protein matches in metatranscriptomic studies. *Nucleic Acids Res.* 2011, *39*, 8792–8802.

[7] Liu, B., Pop, M., ARDB – Antibiotic Resistance Genes Database. *Nucleic Acids Res.* 2009, *37*, D443–447.

[8] Xie, C., Yuan, J., Li, H., Li, M. et al., NONCODEv4: exploring the world of long non-coding RNA genes. *Nucleic Acids Res.* 2014, *42*, D98–D103.

[9] O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S. et al., Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 2015, *44*, D733–D745.

[10] UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* 2015, *43*, D204–212.

[11] Chibucos, M. C., Mungall, C. J., Balakrishnan, R., Christie, K. R. et al., Standardized description of scientific evidence using the Evidence Ontology (ECO). *Database (Oxford)* 2014. doi:10.1093/database/bau075

[12] Pal, C., Bengtsson-Palme, J., Rensing, C., Kristiansson, E., Larsson, D. G. J., BacMet: antibacterial biocide and metal resistance genes database. *Nucleic Acids Res.* 2014, *42*, D737–743.

[13] Blohm, P., Frishman, G., Smialowski, P., Goebels, F. et al., Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis. *Nucleic Acids Res.* 2014, *42*, D396–400.

[14] Naumoff, D. G., Xu, Y., Glansdorff, N., Labedan, B., Retrieving sequences of enzymes experimentally characterized but erroneously annotated: the case of the putrescine carbamoyltransferase. *BMC Genomics* 2004, *5*, 52.

[15] Johnson, E. T., Ryu, S., Yi, H., Shin, B. et al., Alteration of a single amino acid changes the substrate specificity of dihydroflavonol 4-reductase. *Plant J.* 2001, *25*, 325–333.

[16] Smooker, P. M., Whisstock, J. C., Irving, J. A., Siyaguna, S. et al., A single amino acid substitution affects substrate specificity in cysteine proteinases from *Fasciola hepatica*. *Protein Sci.* 2000, *9*, 2567–2572.

[17] Rudikoff, S., Giusti, A. M., Cook, W. D., Scharff, M. D., Single amino acid substitution altering antigen-binding specificity. *Proc. Natl. Acad. Sci. USA* 1982, *79*, 1979–1983.

[18] Glaser, L., Stevens, J., Zamarin, D., Wilson, I. A. et al., A single amino acid substitution in 1918 influenza virus hemagglutinin changes receptor binding specificity. *J. Virol.* 2005, *79*, 11533–11536.

[19] Dabrazhynetskaya, A., Brendler, T., Ji, X., Austin, S., Switching protein-DNA recognition specificity by single-amino-acid substitutions in the P1 par family of plasmid partition elements. *J. Bacteriol.* 2009, *191*, 1126–1131.

[20] Atkinson, H. J., Babbitt, P. C., An atlas of the thioredoxin fold class reveals the complexity of function-enabling adaptations. *PLoS Comput. Biol.* 2009, *5*, e1000541.

[21] Bianchi, L., Díez-Sampedro, A., A single amino acid change converts the sugar sensor SGLT3 into a sugar transporter. *PLoS One* 2010, *5*, e10241.

[22] Doerks, T., Bairoch, A., Bork, P., Protein annotation: detective work for function prediction. *Trends Genet.* 1998, *14*, 248–250.

[23] Jones, C. E., Brown, A. L., Baumann, U., Estimating the annotation error rate of curated GO database sequence annotations. *BMC Bioinformatics* 2007, *8*, 170.

[24] Karp, P. D., Keseler, I. M., Shearer, A., Latendresse, M. et al., Multidimensional annotation of the *Escherichia coli* K-12 genome. *Nucleic Acids Res.* 2007, *35*, 7577–7590.

[25] Poux, S., Magrane, M., Arighi, C. N., Bridge, A. et al., Expert curation in UniProtKB: a case study on dealing with conflicting and erroneous data. *Database (Oxford)* 2014, *2014*, bau016.

[26] Bengtsson-Palme, J., Ryberg, M., Hartmann, M., Branco, S. et al., 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.* 2013, *4*, 914–919.

[27] Hartmann, M., Howes, C. G., Abarenkov, K., Mohn, W. W., Nilsson, R. H., 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. Methods* 2010, *83*, 250–253.

[28] Bengtsson-Palme, J., Hartmann, M., Eriksson, K. M., Pal, C. et al., Metaxa2: improved identification and taxonomic classification of small and large subunit rRNA in metagenomic data. *Mol. Ecol. Resour.* 2015, *15*, 1403–1414.

[29] Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., Knight, R., UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 2011, *27*, 2194–2200.

[30] Nilsson, R. H., Tedersoo, L., Ryberg, M., Kristiansson, E. et al., A comprehensive, automatically updated fungal ITS sequence dataset for reference-based chimera control in environmental sequencing efforts. *Microbes Environ.* 2015, *30*, 145–150.

[31] Hartmann, M., Howes, C. G., Veldre, V., Schneider, S. et al., V-REVCOMP: automated high-throughput detection of reverse complementary 16S rRNA gene sequences in large environmental and taxonomic datasets. *FEMS Microbiol. Lett.* 2011, *319*, 140–145.

[32] Eberhardt, R. Y., Haft, D. H., Punta, M., Martin, M. et al., AntiFam: a tool to help identify spurious ORFs in protein annotation. *Database (Oxford)* 2012, *2012*, bas003.

[33] Finn, R. D., Bateman, A., Clements, J., Coggill, P. et al., Pfam: the protein families database. *Nucleic Acids Res.* 2014, *42*, D222–230.

[34] Nilsson, R. H., Tedersoo, L., Abarenkov, K., Ryberg, M. et al., Five simple guidelines for establishing basic authenticity and reliability of newly generated fungal ITS sequences. *MycoKeys* 2012, *4*, 37–63.

[35] Yilmaz, P., Kottmann, R., Field, D., Knight, R. et al., Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat. Biotechnol.* 2011, *29*, 415–420.

[36] Gardner, P. P., Daub, J., Tate, J., Moore, B. L. et al., Rfam: wikipedia, clans and the "decimal" release. *Nucleic Acids Res.* 2011, *39*, D141–145.

[37] Nilsson, R. H., Hyde, K. D., Pawłowska, J., Ryberg, M. et al., Improving ITS sequence data for identification of plant pathogenic fungi. *Fungal Divers.* 2014, *67*, 11–19.

