Instructions for Sanger sequence annotation study-by-study

Objectives

This subproject of UNITE-traits involves updating and supplementing metadata to existing INSD sequences to build a new, high-resolution database of principal ecological traits of fungi and oomycetes. The benefits include: 1) a wiki-like body of reference information for researchers that want to have a rough idea about the occurrence and ecological role of certain fungi; 2) reference traits database for researchers that use molecular methods for identification and are interested in functional traits in addition to taxonomy. To achieve the latter goal, the reference data set is integrated into the bioinformatics workflow of high-throughput sequencing analyses and taxonomic reference databases such as UNITE. 3) to share the benefits among contributors by co-authoring an enormously citeable database article and to provide useful practice of work with sequences and attention to the importance of metadata.

Methods

Background

As of early January 2019, all studies that included ITS sequences in INSD were downloaded and sorted by the number of sequences they contained, resulting in a data set of >40,000 studies and >1,000,000 ITS sequences. Project coordinators first selected studies that contain at least 100 ITS sequences and filtered out HTS/NGS studies and studies involving plant, animal and protist sequences. These studies were separated by combinations of subject (substrate, pathogenicity, country of performance, and research team) into multiple groups to be shared among the participants, preferably covering their field of particular expertise. For selected critical groups, we searched for information from studies with lower number of citations and added these to the filtered data set. Altogether, our effort is focused on all sequences contained in >3000 studies.

INSD provided >100 data fields for data and metadata about the entries. While some of the metadata about sequences are present, these data are usually scattered across several fields and the terminology is used in a highly inconsistent manner. Furthermore, many studies lack any background information about its sequences, rendering these essentially as noise in INSD. Project coordinators carefully checked the data fields present in INSD and re-ordered these in a seemingly logical manner.

Based on MIMARKS and other standards, project coordinators prepared lists of principal traits and trait states that would be most useful for the scientific community intending to use these data as a reference or wiki-based information. These traits were added next to the respective fields from the INSD to facilitate supplementing metadata from these pre-existing fields to our standardized fields. In the categorical trait fields, the trait states are selectable from a drop-down list. Other fields (Updated_study, interacting_taxon, latitude, longitude, etc.) require manual entry of specific information, preferably in copy-paste form to prevent spelling mistakes.

Each participant is provided with a set of ITS sequences from 30-50 studies in an Excel file. The data is initially sorted by study (first column) and sequence accession. The first row represents the name of the trait or field of information. The second row describes this field, specifies the format of data and may provide guidelines of data entry.

Non-highlighted data fields represent pre-existing INSD data and metadata that serve as information about what is already there, but in a potentially different format. The highlighted fields represent traits in three categories of relevance. Fields with titles highlighted in red are most important and should be filled, whenever this information can be found or retrieved from the authors. Fields with titles in green

represent general ecological traits that are important and relevant to most studies and these make the basic contribution to the traits database. Fields with titles in yellow are very specific fields that are relevant to only for a fraction of the studies, but these are very important when it comes to specific information about where fungal cultures were isolated from, type status of cultures or specimens, tissue of occurrence in case of animal/human samples, putative biotrophic capacities of plant and animal-associated fungi.

In addition to co-authorship, your name as a contributor will be marked in the separate field of the database for each sequence you revised. You have a right to remain anonymous, but please do inform us for this decision.

Guidelines for annotation

1. Check the studies and data fields to understand the structure of the task and data; if necessary, ask from Dr. Sergei, sergei.polme@gmail.com. But before, read the instructions for a second time.
2. Start with the first study – check which data fields are not filled, which terms are non-standard.
3. Field Study_updated. Whether or not the study is marked as unpublished or published, copy the title and first author to Google Scholar search. if found, paste the DOI here (short form or link). If unpublished/not found, write 'unpublished'. DO download the paper. From the paper you probably need to look for specific information, which can be in the text body or table format. Programs such as Adobe Professional and some freeware programs enable converting entire tables from pdf to Excel format that may improve your work. Note that during conversion or copying from pdf, there may be mistakes because of poor recognition of text or other incompatibilities. If you need to match up accessions from the two tables, you can use the VLOOKUP Excel function.
4. Field DNA_Source. Have a look at the ~~four~~four preceding INSD fields for information and check from the paper. The DNA_Source field already contains some information from previous annotation trials, but these may be incorrect and non-corresponding to the updated terminology. SELECT source of DNA from the drop-down list. Attention! Sequences obtained from cultures should be treated as 'living_culture'. In the next field, Culture_source, you can specify the substrate from which the culture was obtained from. If the sequence or culture was obtained from human or animal tissue, fill the next field, Animal/human tissue. For e.g. corals, sponges mark 'other' for tissue type. Check the Interacting_taxon field further on.
5. The fields Guild and Growth_form may be rather speculative. Therefore, fill these from the drop-down list ONLY IF these features have been indicated in the publication. DO NOT make guesses just to fill these fields.
6. The fields Ectomycorrhiza_exploration_type, Ericoid_mycorrhiza_formation, Endophytic_interaction_capability, Plant/fungal_pathogenic_capacity, and Animal/human_biotrophic_interaction_capacity are relevant to only a fraction of studies that particularly address tissues of plants, fungi including lichens, and animals. If your set of studies does not contain these substrates or manipulation trials, or expert opinion of the authors, do not fill these to avoid mistakes
7. The fields Interacting_taxon and Co-occurring_taxa. Have a look at the preceding UNITE field for information and check from the paper. The Latin binomials and higher level names can be copied or modified from the respective GenBank fields and original paper. The Interacting_taxon represents ONLY INTIMATE associations, e.g. fungi sequenced from tissue (or culture originating from) of specific plant/animal host or victim. The co-occurring taxon/taxa structure the habitat but are NOT NECESSARILY intimately associated. This is relevant for free-living organisms but also to obligate pathogens or mutualists isolated from sources other than the intimate interacting taxon (e.g. mycorrhizal mycelium in Pinus sylvestris forest soil). Give names of the interacting/co-occurring organisms in Latin (species to kingdom level), separated by commas if there are more than one.

8. The fields Strain, Specimen.voucher and Type_status originate from INSD and are intended for supplementation with information from the publication. These are relevant only for taxonomic studies. Note the format of Collection/Herbarium acronym – space – number. In the Type_status field, information about typification should be checked from the paper (only relevant to taxonomic papers!). If this is relevant to your set of studies, please learn about the terminology of types; if unsure 'type' will do.
9. The fields Sampling.area.Country, State/Province, Locality_text rely on the two preceding INSD fields and the Sampling.area.Country field is pre-filled with data from INSD. For large countries of >1M km^2 it is important to provide State/Province information IF latitude and longitude data are unavailable. Attention! Former and present colonies of some countries (e.g. Puerto Rico, French Guyana but not Hawaii are enlisted separately from mother countries).
10. The fields Sampling.area.Latitude and Sampling.area.Longitude contain pre-filled information that should be supplemented. Please search the locality in Google Map, Google Earth or similar source - If sampling area can be located to 100 km (1 degree) precision, please fill the coordinates but only at DD.D (10 km) precision. Otherwise, fill in DD.DDDDD using this format and WGS84 projection. Attention! The system will not recognize DD.MM.SS or UTM formats, so these would need to be converted. Attention! latitudes in the Southern Hemisphere and longitude in Western Hemisphere are negative.
11. The fields Altitude (m above sea level) and Depth (m below water surface or soil surface or sea bottom) are related to two previous GenBank fields for each. If the coordinates are precise, altitude can be obtained from Google Earth.
12. The field Biome is related to two preceding fields, coordinates and habitat description in the article. Users should select the best suitable biome from a drop-down list.
13. The Remarks field is optional and is intended to include any information when authors selected 'other' in any of the fields or if they feel that important information needs to be added. Project coordinators will review these and potentially incorporate into other fields (if relevant) or edit for retention in the remarks field related to a relevant traits field.
14. Participants may re-arrange the fields if they feel that other arrangement is more logical or useful given their set of sequences. Do NOT edit sequence accessions (3rd column) or names of the data fields (1st row).
15. After finishing with information from each study, the participants are RESPONSIBLE FOR contacting the authors (find contacts from papers of institutions if the coordinators are unable to arrange these from INSD authorities) to get additional information about the mandatory (red-highlighted) fields if they see that this information can be easily provided by the author (e.g. progress with publication – make a note to REMARKS; geocoordinates, DNA source, but also interacting taxon if relevant). You can use the inquiry text given in the end of this document. It is important to contact the authors ASAP to secure them time to find the information for you.
16. Repeat these steps study-by-study
17. If contact details of some authors cannot be found anyhow, in very important cases you can contact Prof. Conrad Schoch (schoch2@ncbi.nlm.nih.gov), but do this for asking for several at the same time and providing the INSD original study name (column 1).
18. Incorporate principal information obtained from the authors. Please send a reminder to the authors after one week of no-response.
19. Re-check the data for obvious copy-paste errors and send it back to Sergei Põlme. Please add some comments about contacting researchers (how many contacted, how many responded, how many actually helped).

Recommended body of email for contacting the authors

Dear Dr. ...,

I am contacting you regarding the global collaborative initiative of fungal traits annotation that aims to provide metadata to GenBank sequences. The objectives of this initiative include preparation of species-level traits data set of fungi and oomycetes to benefit the public understanding about functions of different fungi and providing a traits assignment tool to researchers working on molecular methods of identification in many different fields of research (further information from Dr. Sergei Põlme, sergei.polme@gmail.com). To accomplish these objectives, I am turning to you regarding metadata about your GenBank submission entitled ... . I note that for sequences of this submission, very important metadata about ... and ... are missing. Please send me additional information about these important fields so that the scientific community can benefit more from your work.

Sincerely,

...

The fungal traits annotation team.