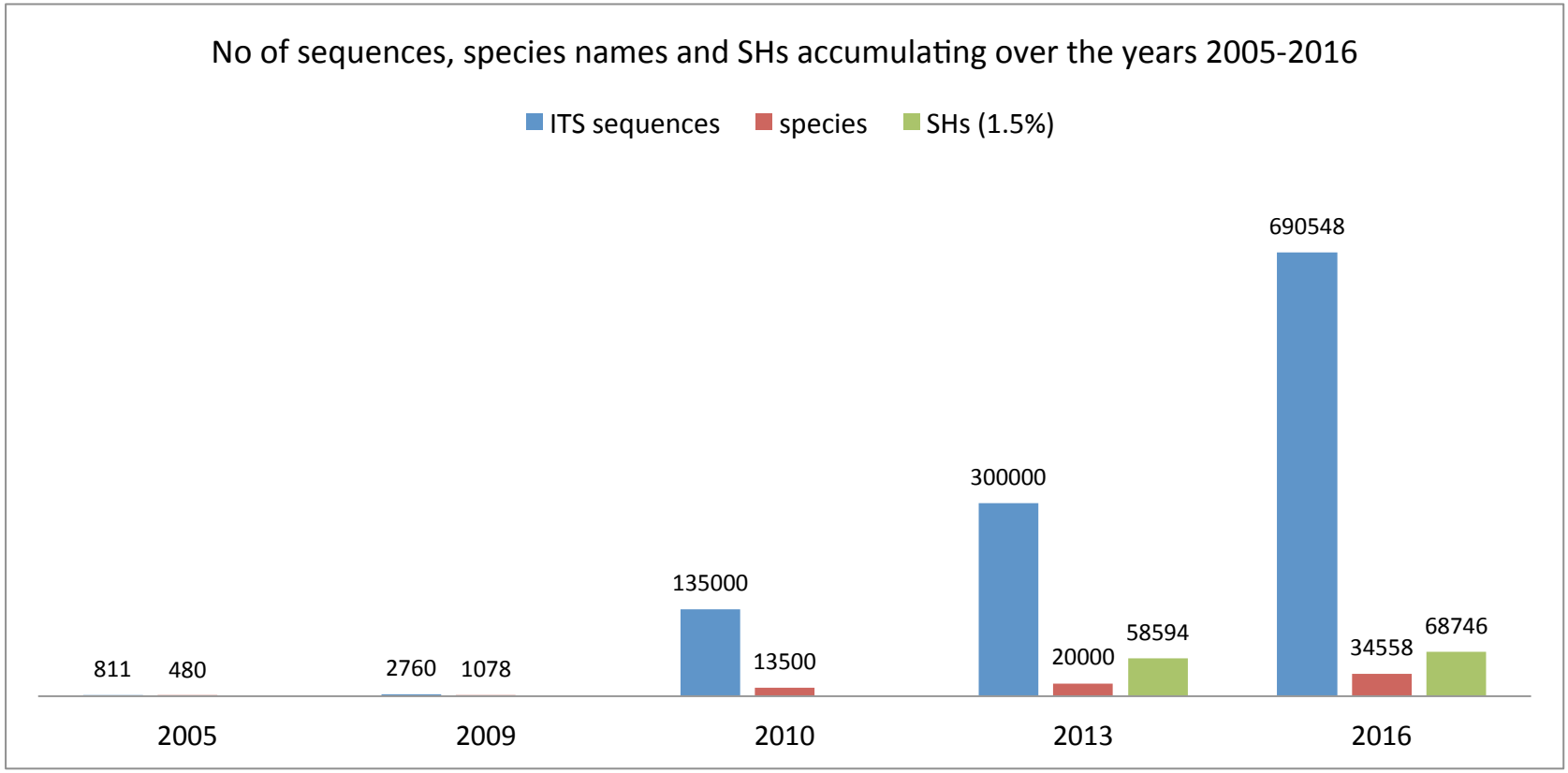
# INSD sequence annotations in PlutoF

Kessy Abarenkov
University of Tartu

“Biodiversity Data Management and Open Data” course, Tartu, Nov 2 – 5, 2016

## History of UNITE + INSD data




Addition of INSD sequences

Implementing SH system

**Main reasons for third party annotation:**

- Identified vs unidentified sequences
- Large proportion (up to 20%) incorrectly identified (Nilsson et al., 2006)
- Lack (and heterogeneity) of metadata on, e.g. country of collection, interacting taxon, source of identification
- Sequence quality problems (chimeric, low quality)

**Main reasons for third party annotation:**

- Identified vs unidentified sequences
- Large proportion (up to 20%) incorrectly identified (Nilsson et al., 2006)
- Lack (and heterogeneity) of metadata on, e.g. country of collection, interacting taxon, source of identification
- Sequence quality problems (chimeric, low quality)

**Results by taxon**

Top Organisms [Tree]

uncultured fungus *(216473)*
fungal endophyte *(6217)*
uncultured Ascomycota *(6160)*
uncultured Glomus *(4387)*
Colletotrichum gloeosporioides *(3547)*
All other taxa *(423191)*
More...

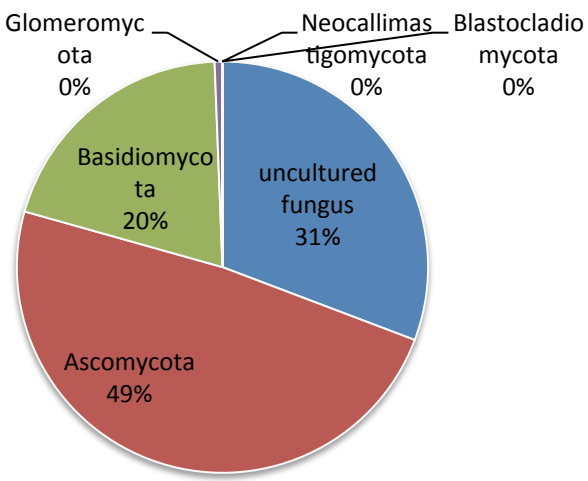
*https://www.ncbi.nlm.nih.gov*

**Main reasons for third party annotation:**

INSD sequences from the first 6 months in 2016 – **56,689**

- Country specified – 56%
- Isolation source specified – 63%
- Host specified – 33%
- MixS metadata specified – 0%

Distribution of sequences based on phylum:




```
/isolation_source="sunflower leaf surface"
/isolation_source="air"
/isolation_source="Atlantic sea water"
/isolation_source="hemp rope"
/isolation_source="exudate of Betula maximowicziana"
/isolation_source="dead pustule of raspberry yellow rust
/isolation_source="tropical foliage"
/isolation_source="forest ecosystem"
/isolation_source="smut-infected leaves"
/isolation_source="pasture grass"
/isolation_source="soil in the permafrost area"
/isolation_source="female olive fly Bactrocera oleae"
/isolation_source="lake water"
/isolation_source="air"
/isolation_source="male olive fruit fly"
/isolation_source="olive Olea europeae flower"
/isolation_source="Araucaria araucana tree"
/isolation_source="soil"
/isolation_source="auxotrophic mutant of strain UCDFST
/isolation_source="Opuntia sp."
/isolation_source="Turks head cactus Melocactus intortus"
/isolation_source="leaf of Desmodium repens"
/isolation_source="green lacewing Chrysoperla carnae"
/isolation_source="Opuntia ficus-indica"
/isolation_source="exudate of Betula ermanii"
/isolation_source="sea water"
/isolation_source="insect frass in Alder tree Alnus sp."
/isolation_source="drying sap of olive tree Olea europaea"
/isolation_source="female olive fly Bactrocera oleae"
/isolation_source="wrist from female"
```

**Main reasons for third party annotation:**

- Identified vs unidentified sequences
- Large proportion (up to 20%) incorrectly identified (Nilsson et al., 2006)
- Lack (and heterogeneity) of metadata on, e.g. country of collection, interacting taxon, source of identification
- Sequence quality problems (chimeric, low quality)

**List of annotation efforts:**

- **2016** — **Built mycobiome** sequence metadata annotation workshop (Abarenkov et al. 2016)
- **2016** — Top 50 most wanted – annotating the **"dark fungal diversity"** (Nilsson et al. 2016)
- **2014** — Improving ITS sequence data for identification of **plant pathogenic fungi** (Nilsson et al. 2014)
- **2013** — **UNITE jamboree** – fungal ITS sequence annotation workshop (Kõljalg et al. 2013)
- **2011** — Ecological, geographical and sequence quality annotation of ITS sequences of **mycorrhizal fungi** (Tedersoo et al. 2011)

**List of annotation efforts:**

| | |
|---|---|
| 2016 | **Built mycobiome** sequence metadata annotation workshop (Abarenkov et al. 2016) |
| 2016 | Top 50 most wanted – annotating the **"dark fungal diversity"** (Nilsson et al. 2016) |
| 2014 | Improving ITS sequence data for identification of **plant pathogenic fungi** (Nilsson et al. 2014) |
| 2013 | **UNITE jamboree** – fungal ITS sequence annotation workshop (Kõljalg et al. 2013) |
| 2011 | Ecological, geographical and sequence quality annotation of ITS sequences of **mycorrhizal fungi** (Tedersoo et al. 2011) |

## 2011

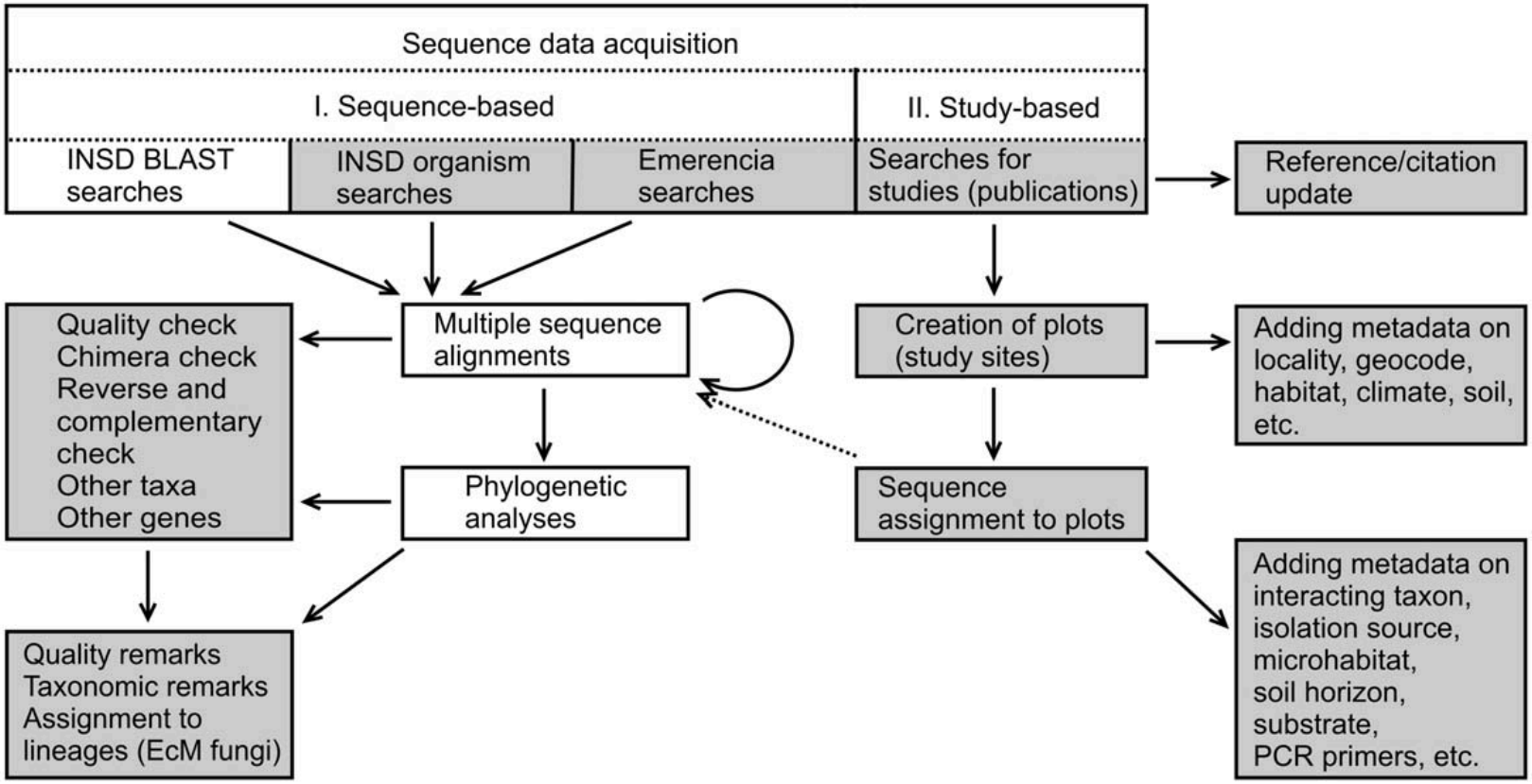



Figure 1. Scheme of the metadata annotation workflow. Shaded boxes indicate procedures performed and/or saved over the PlutoF workbench (http://plutof.ut.ee/). *Tedersoo et al., 2011.*

**List of annotation efforts:**

- **2016** — **Built mycobiome** sequence metadata annotation workshop (Abarenkov et al. 2016)
- **2016** — Top 50 most wanted – annotating the **"dark fungal diversity"** (Nilsson et al. 2016)
- **2014** — Improving ITS sequence data for identification of **plant pathogenic fungi** (Nilsson et al. 2014)
- **2013** — **UNITE jamboree** – fungal ITS sequence annotation workshop (Kõljalg et al. 2013)
- **2011** — Ecological, geographical and sequence quality annotation of ITS sequences of **mycorrhizal fungi** (Tedersoo et al. 2011)

**List of annotation efforts:**

**2016** **Built mycobiome** sequence metadata annotation workshop (Abarenkov et al. 2016)

**2016** Top 50 most wanted – annotating the **"dark fungal diversity"** (Nilsson et al. 2016)

**2014** Improving ITS sequence data for identification of **plant pathogenic fungi** (Nilsson et al. 2014)

**2013** **UNITE jamboree** – fungal ITS sequence annotation workshop (Kõljalg et al. 2013)

**2011** Ecological, geographical and sequence quality annotation of ITS sequences of **mycorrhizal fungi** (Tedersoo et al. 2011)

**2014**

Methods for contacting: symposia, personal networking, ResearchGate, …

27 largest journals in plant pathology (and 12 mycological journals) were scanned for descriptions of new (or typifications of existing) plant pathogenic or plant-associated species of fungi.

Types of annotations:

- Selection of representative sequences for species
- Improvement of taxonomic annotations
- Addition on ecological metadata (host, country of collection)
- Identifying compromised sequence data
- A total of **31,954** changes were implemented

# 2014

PlutoF
an input module for biological data

» All clusters

**UNITE - fungal identification with rDNA ITS sequences** | **Version 6** | date: 2013-11-19 | Cluster code: **UCL6_002877**

This set of sequences contains genera: Cylindrocladium, Calonectria
Total number of sequences in cluster: 386

- chimeric
- low quality
- UNITE core sequence
- automatically chosen 97% SH representative sequence
- Ex = sequence to be excluded from the next version of global key

(filled, coloured circle) manually chosen SH reference sequence, overrides automatically chosen representative sequence

SH 97% | Alignment based on: Full ITS | Order sequenc | Send clusters to clipboard

| Sequence ID | UNITE taxon name | INSD taxon name | Country | DNA source | Interacting taxa |
|---|---|---|---|---|---|
| more \| GQ334423 | **Nectriaceae** | Fungi (fungal sp HS_EF16) | China | | |
| more \| GQ334422 | **Nectriaceae** | Fungi (fungal sp HS_EF15) | China | | |
| more \| JQ347281 | | Cylindrocladium (Cylindrocladi... | India | | Eucalyptus |
| more \| JQ347273 | | Cylindrocladium (Cylindrocladi... | India | | Eucalyptus |
| more \| JQ347280 | | Cylindrocladium (Cylindrocladi... | India | | Eucalyptus |
| more \| GQ280609 | **Calonectria penicilloides** | Calonectria (Calonectria penic... | Japan | Living culture (**Ex-type**) | |
| more \| U36443 | | Calonectria (Calonectria kyote... | | | |
| more \| AB287008 | | Calonectria (Calonectria pacif... | Japan | | |
| more \| AF261742 | | Calonectria (Calonectria kyote... | | | |
| more \| AF261741 | | Calonectria (Calonectria kyote... | | | |
| more \| JQ694095 | | Calonectria (Calonectria sp WH... | China | | |
| more \| DQ132847 | | Cylindrocladium (Cylindrocladi... | Canada | Plant root | Picea mariana |
| more \| AY705981 | | Cylindrocladium (Cylindrocladi... | | | |
| more \| DQ132844 | | Cylindrocladium (Cylindrocladi... | Canada | Plant root | Picea mariana |
| more \| DQ132845 | | Cylindrocladium (Cylindrocladi... | Canada | Plant root | Picea mariana |
| more \| DQ132822 | | Cylindrocladium (Cylindrocladi... | Canada | Plant root | Picea mariana |
| more \| AY705980 | | Cylindrocladium (Cylindrocladi... | | | |
| more \| DQ132848 | | Cylindrocladium (Cylindrocladi... | Canada | Plant root | Picea mariana |
| more \| AY273306 | **Nectriaceae** | Ascomycota (uncultured Ascomyc... | Gabon | Soil fungal DNA | |

**List of annotation efforts:**

- **2016** — **Built mycobiome** sequence metadata annotation workshop (Abarenkov et al. 2016)
- **2016** — Top 50 most wanted – annotating the **"dark fungal diversity"** (Nilsson et al. 2016)
- **2014** — Improving ITS sequence data for identification of **plant pathogenic fungi** (Nilsson et al. 2014)
- **2013** — **UNITE jamboree** – fungal ITS sequence annotation workshop (Kõljalg et al. 2013)
- **2011** — Ecological, geographical and sequence quality annotation of ITS sequences of **mycorrhizal fungi** (Tedersoo et al. 2011)

**2016 I**

Nilsson et al. 2016. Top 50 most wanted fungi. MycoKeys.

Top 50 search in UNITE

**List of annotation efforts:**

- **2016** — **Built mycobiome** sequence metadata annotation workshop (Abarenkov et al. 2016)
- **2016** — Top 50 most wanted – annotating the **"dark fungal diversity"** (Nilsson et al. 2016)
- **2014** — Improving ITS sequence data for identification of **plant pathogenic fungi** (Nilsson et al. 2014)
- **2013** — **UNITE jamboree** – fungal ITS sequence annotation workshop (Kõljalg et al. 2013)
- **2011** — Ecological, geographical and sequence quality annotation of ITS sequences of **mycorrhizal fungi** (Tedersoo et al. 2011)

**2016 II**

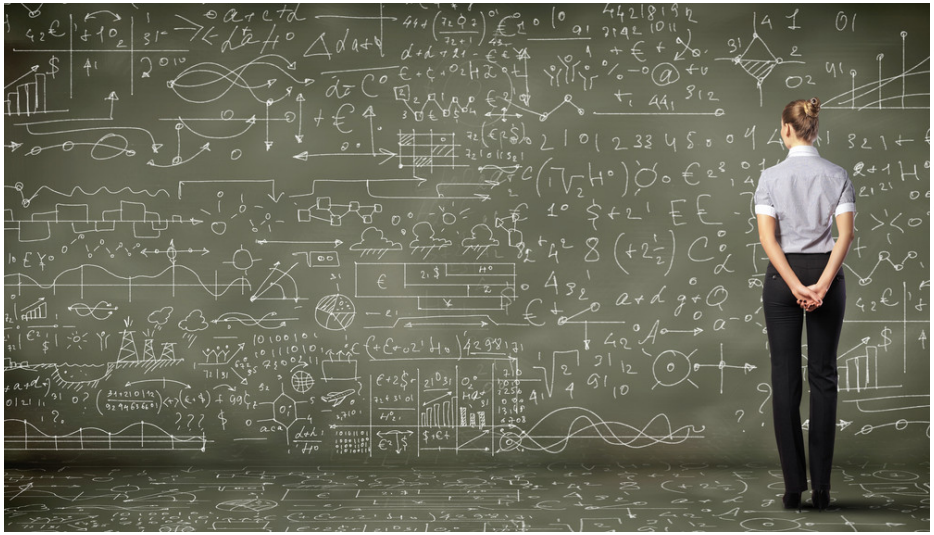
Annotating public fungal ITS sequences from the built environment according to the MIxS-Built environment standard – a report from a May 23-24, 2016 workshop (Gothenburg, Sweden)

**255** studies, ~**18,000** sequences
**45,488** annotations made

## Statistics on metadata annotations in UNITE

| | |
|---|---|
| **22,217** | taxonomic re-annotations |
| **56,720** | specifications of ectomycorrhizal lineage |
| **101,035** | specifications of country of collection |
| **52,720** | specifications of host and interacting taxa |
| **2,977** | chimeric and **7,224** low read quality sequences found |
| **6,281** | reference sequence specifications |
| **9,560** | sequences with specimen/culture metadata annotated |
| **4,518** | sequences linked to type specimens/cultures |

# What next?



http://www.vitaminam.com.mx/tag/marketing/

**Resources**

UNITE homepage (https://unite.ut.ee)

- Searches
- Reference datasets
- Custom queries (e.g. Top 50 most wanted)

PlutoF platform (https://plutof.ut.ee)

- Searches
- Analysis
- Export
- RESTful web services

## Lessons learned and recommendations for future annotation efforts

- Collaborate
- Give something in return
- Use user-friendly ways to do the job
- Automate tasks
- Most valuable resource – human

## Lessons learned and recommendations for future annotation efforts

- Most valuable resource – human (the rare taxonomist)