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)


¹ University of Tartu, Natural History Museum, Tartu, Estonia  ² University of California, Plant & Microbial Biology, Berkeley, CA 94720, USA  ³ Sydney Medical School-Westmead Hospital, Molecular Mycology Research Laboratory, Centre for Infectious Diseases and Microbiology, Sydney, Australia  ⁴ University of Sydney, Marie Bashir Institute for Infectious Diseases and Biosecurity, Sydney, Australia  ⁵ Westmead Institute for Medical Research, Westmead, Australia  ⁶ Institute of Ecology and Earth Sciences, University of Tartu, Tartu, Estonia  ⁷ Via Calamandrei 2, 53035 Monteriggioni, Siena, Italy  ⁸ University of Gothenburg, Department of Biological and Environmental Sciences, Box 461, 405 30 Göteborg, Sweden  ⁹ Gothenburg Botanical Garden, Carl Skottsbergs gata 22A, 413 19 Göteborg, Sweden  ¹⁰ Department of Organismal Biology, Evolutionary Biology Centre, Uppsala University, SE 75236 Uppsala, Sweden  ¹¹ University of Gothenburg, Department of Infectious Diseases, Sahlgrenska Academy, Guldhedgatan 10, 413 46 Göteborg, Sweden  ¹² SP Technical Research Institute of Sweden, Box 857, 501 15 Borås, Sweden  ¹³ Universidade Federal de Pernambuco (UFPE), Departamento de Micologia, Centro de Ciências Biológicas (CCB), Av. Prof. Nelson Chaves, s/n, 50670-901 Recife, Pernambuco, Brazil  ¹⁴ University of Tuscia, Department of Ecological and Biological Sciences, 01100, Viterbo, Italy  ¹⁵ University of Gothenburg, Herbarium GB, Box 461, 405 30 Göteborg, Sweden  ¹⁶ South China Botanical Garden, The Chinese Academy of Sciences, 723 Xingke Road, Tianhe District, Guangzhou 510650, China  ¹⁷ Chalmers University of Technology, Department of Mathematical Sciences, 412 96 Göteborg, Sweden  ¹⁸ University of Gothenburg, Department of Marine Sciences, Box 460, 405 30 Göteborg, Sweden  ¹⁹ University of Gothenburg, Department of Marine Sciences - Tjärnö, 452 96, Strömstad, Sweden  ²⁰ University of California Riverside, Department of Plant Pathology and Microbiology and Institute of Integrative Genome Biology, Riverside, California 92521 USA  ²¹ University of Gothenburg, Department of Earth

* Co-authors contributed equally.

Copyright Kessy Abarenkov et al. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
Abstract
Recent molecular studies have identified substantial fungal diversity in indoor environments. Fungi and fungal particles have been linked to a range of potentially unwanted effects in the built environment, including asthma, decay of building materials, and food spoilage. The study of the built mycobiome is hampered by a number of constraints, one of which is the poor state of the metadata annotation of fungal DNA sequences from the built environment in public databases. In order to enable precise interrogation of such data – for example, “retrieve all fungal sequences recovered from bathrooms” – a workshop was organized at the University of Gothenburg (May 23–24, 2016) to annotate public fungal barcode (ITS) sequences according to the MiXS-Built Environment annotation standard (http://gensc.org/mixs/). The 36 participants assembled a total of 45,488 data points from the published literature, including the addition of 8,430 instances of countries of collection from a total of 83 countries, 5,801 instances of building types, and 3,876 instances of surface-air contaminants. The results were implemented in the UNITE database for molecular identification of fungi (http://unite.ut.ee) and were shared with other online resources. Data obtained from human/animal pathogenic fungi will furthermore be verified on culture based metadata for subsequent inclusion in the ISHAM-ITS database (http://its.mycologylab.org).

Key words
Built environment, Indoor fungi, ITS, Annotation, Mycobiome

Introduction
Fungi are found throughout the biosphere, and the built environment is no exception. The taxonomic composition of indoor fungal communities tends to reflect the local outdoor communities, although the majority of fungal particles found indoors is
thought to represent spores, hyphal fragments, and other dormant and passively distributed stages (Seo et al. 2015). Although most of the fungi recovered from indoor environments would not be able to live in the built environment for any extended period of time, a minority of these species are able to cope with, and will even thrive in, the harsh conditions that the built environment presents (Hamada and Abe 2010; Nevalainen et al. 2015; Zupančič et al. 2016). These species are mainly saprotrophic, and their degree of active growth largely depends on water availability (Adams et al. 2013). They can be a serious cause of decay and other concerns in water-damaged buildings, but they are also found in buildings not subject to moisture issues – and even in buildings where very strict sanitization and filtration regimes are applied (e.g., La Duc et al. 2012; Checinska et al. 2015). Exposure to aerosolized fungal particles has been linked to asthma onset in humans and may furthermore play a role in eczema development and other issues in human health (Reijula et al. 2003; Knutsen et al. 2012). Indoor fungi may also contribute to other unwanted processes, such as food spoilage and wall staining (Varga et al. 2014). The built mycobiome is thus of interest to a range of scientific fields, including mycology, medicine, food biology, construction, and engineering.

Traditional, morphology-based studies of fungal spores and cultures derived from indoor sampling have recognized ca. 90 species of common indoor fungi (Flannigan et al. 2002). Efforts based on high-throughput DNA sequencing, in contrast, have revealed a vast and hitherto unknown diversity of indoor fungi. In a global study of indoor dust samples, Amend et al. (2010) using next-generation sequencing found ca. 4,500 fungal operational taxonomic units (OTUs; Blaxter et al. 2005) approximately at the species level. Similarly, another next-generation sequencing-powered study – Nonnenmann et al. (2012) – recovered 450 fungal species from 50 indoor dust samples in Yakima valley, WA (USA). Although precise species delimitation and species counts from next-generation sequencing data remain challenging (Nguyen et al. 2016), the taxonomic span of the fungal assemblages recovered in Amend et al. (2010) and Nonnenmann et al. (2012) is far larger than that occupied by the fungi traditionally thought of as common indoor fungi (cf. Flannigan et al. 2002). Thus, whereas these studies should not be used as estimates of the total number of indoor fungi, they do testify to the substantial diversity of fungi in the built environment. The lack of taxonomic reference sequences makes precise identification of many of these species problematic, and it is not unusual that a sizable proportion of the OTUs in environmental sequencing studies remain unassigned beyond the kingdom or phylum levels (e.g., Tedersoo et al. 2014; Grossart et al. 2016; Fouquier et al. 2016; Nilsson et al. 2016). There is clearly a need to generate reliable reference sequences, most notably from type material, to address this issue (cf. Schoch et al. 2014). However, the estimated number of extant species of fungi – 1.5-6 million (Hawksworth 2001; Taylor et al. 2014) – stands in stark contrast to the number of described species (~130,000 as of March 2016; www.speciesfungorum.org), and strongly suggests that molecular identification of fungi will remain challenging for the foreseeable future. In some cases, even reference barcode (nuclear
ribosomal internal transcribed spacer, ITS) sequences from type material will not be enough. Several fungal genera regularly recovered from built environment samples - such as *Aspergillus*, *Cladosporium*, *Fusarium*, and *Penicillium* - show little or no ITS variation across sets of two to several species (Bensch et al. 2012; Samson et al. 2014; Visagie et al. 2014; O’Donnell et al. 2015). Additional genetic markers are needed for robust species-level identification in these cases.

A second problem that compounds the scientific understanding of the built mycobiome has been the lack of a standardized vocabulary for sequence annotation. The International Nucleotide Sequence Database Collaboration (INSDC; Cochrane et al. 2016) holds more than 5,000 Sanger-derived fungal ITS (barcode) sequences from the built environment, but their level of metadata annotation differs widely. This unfortunately applies to most available fungal ITS sequences (cf. Nilsson et al. 2014); for example, a modest 43% are known to be annotated with something as simple and straightforward as country of collection (Tedersoo et al. 2011). In addition, where metadata exist they are not always provided in standardized and searchable formats, making precise queries difficult. There is, for instance, no straightforward way to download all fungal ITS sequences from bathrooms, or to target the substrate of gypsum board. It is reasonable to think that analysis of fungi recovered from bathrooms may prove a rewarding scientific enterprise, as indeed should be the case for fungi collected on specific building materials, under different moisture regimes, or any other particular parameter or setting. The full potential of such searches cannot presently be utilized due to the poor state of sequence annotation – primarily omitted by the original sequence authors – in the public sequence databases.

The new MIxS-Built Environment annotation standard (Glass et al. 2014; http://gensc.org/mixs/) addresses the need for a thorough, standardized vocabulary for microbiological analysis of the built microbiome. If all relevant fungal ITS sequences in the INSDC were annotated according to this standard, then this would open up the body of extant molecular data to detailed, precise scientific queries in the context of the built mycobiome. Going through and annotating large sequence sets is a daunting effort for any researcher, but fortunately such efforts are easy to split among a set of individual researchers. This paper presents the outcome of a sequence metadata annotation workshop (University of Gothenburg, May 23-24, 2016) to annotate the ~6,500 public fungal ITS sequences from the built environment according to the most relevant parts of the MIxS-Built Environment annotation standard. In recognition of the fact that fungi found indoors are typically found outdoors as well, the workshop also annotated closely related outdoor sequences according to basic geo-ecological parameters. The workshop was organized jointly with the UNITE and ISHAM databases (Kõljalg et al. 2013; Irinyi et al. 2015). UNITE is a general-purpose sequence management environment seeking to reconcile molecular ecology and taxonomy of fungi and fungal communities. The ISHAM database centers on identification of human and animal pathogenic fungi to guide antifungal treatment choices. Both databases focus, at least for the time being, on the ITS region and share views on the importance of openness, free accessibility, and community participation.
**Materials and methods**

The workshop comprised 20 physical participants, mainly local Ph.D. students and post-docs – but also other researchers – in systematics and ecology. In addition, another 16 researchers participated remotely through Skype, Google Docs, and email. The participants focused on the public fungal ITS sequences of the INSDC as mirrored in the UNITE and ISHAM databases.

To single out INSDC sequences associated with the built environment, we used a set of 24 keywords such as “dust”, “gypsum”, and “floor” (Suppl. material 1). Keyword matches were made to the title of the underlying publication (the INSDC field “title”), the INSDC fields “source” and “tissue type”, and the UNITE field “sequence source”. We refer to this set of sequences as the **built mycobiome set** (BMS). To single out outdoor sequences with a direct relation to the BMS, we extracted all UNITE species hypotheses with at least one BMS sequence. We then built the **outdoor mycobiome set** (OMS) from all sequences that did not match any of our keywords but that were found in the same species hypothesis as at least one BMS sequence. Sequences that initially were assigned to the BMS set, but that on closer inspection turned out not to qualify as the built mycobiome (“collected outside hospital”, for example), were transferred to the OMS set.

For each BMS sequence we tried to locate any underlying publication through the INSDC fields TITLE, JOURNAL, and PUBMED. If these were not informative, we resorted to ISI Thompson, Google/Google Scholar, and ResearchGate searches. We examined the publications for the nine items of the MIxS-Built Environment annotation standard that we felt were the most relevant and the most likely to be covered by the studies: building occupancy type, indoor space, indoor surface, surface material, surface-air contaminant, space typical state, substructure type, ventilation type, and filter type (http://gensc.org/mixs/). In addition we also targeted the country and host of collection and the nature of the fungus-host association (e.g., “plant: wood”, “plant: leaf”, and “human/animal: skin”), as applicable, for all sequences. We only targeted metadata and information that was clearly and unequivocally specified in the paper. A research professional (G. Bok) from a building-related technical institute was present to assist with technical, analytical, and construction-related questions in the context of the built environment. For the OMS we similarly retrieved the underlying publications and annotated the sequences to country and host of collection plus host association (as applicable, and if and when these data were missing). All results were entered into an Excel sheet for upload into UNITE and ISHAM (after culture-based verification in the case of the latter), and for sharing with other online resources.

**Results**

A total of 6,526 BMS and 11,574 OMS sequences from a total of 255 separate studies were annotated with at least one metadata item. A total of 45,488 annotations were made during the workshop. For example, “building occupancy type” was established for 5,801 sequences, and “ventilation type” was established for 2,235 sequences (Table 1;
Table 1. Results of the annotation workshop, specified for the built mycobiome sequence set (BMS) and the outdoor mycobiome sequence set (OMS). Countries and hosts of collections plus host association were assembled for both of these. The number of sequences processed, plus the number of underlying published and unpublished scientific studies, are also provided. For the BMS, the nine MIxS-Built Environment annotation standard items targeted at the workshop are specified in separate columns. The sequence numbers shown in the table refer to the number of sequences annotated for each data item.

<table>
<thead>
<tr>
<th></th>
<th>Number of sequences (annotated)</th>
<th>Number of studies</th>
<th>Country of collection</th>
<th>Different countries</th>
<th>Host of collection</th>
<th>Different hosts</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMS</td>
<td>6550 (6526)</td>
<td>144</td>
<td>2447</td>
<td>29</td>
<td>881</td>
<td>15</td>
</tr>
<tr>
<td>OMS</td>
<td>16766 (11574)</td>
<td>128</td>
<td>5983</td>
<td>83</td>
<td>5632</td>
<td>859</td>
</tr>
<tr>
<td>Total</td>
<td>23316 (18100)</td>
<td>255 unique</td>
<td>8430</td>
<td>83 unique</td>
<td>6513</td>
<td>865 unique</td>
</tr>
<tr>
<td>Host association</td>
<td>Comments</td>
<td>Building occupancy type</td>
<td>Indoor space</td>
<td>Indoor surface</td>
<td>Surface material</td>
<td></td>
</tr>
<tr>
<td>BMS</td>
<td>764</td>
<td>2348</td>
<td>5801</td>
<td>1223</td>
<td>1207</td>
<td>1318</td>
</tr>
<tr>
<td>OMS</td>
<td>2892</td>
<td>1293</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Total</td>
<td>3656</td>
<td>3641</td>
<td>5801</td>
<td>1223</td>
<td>1207</td>
<td>1318</td>
</tr>
<tr>
<td>Surface-air contaminant</td>
<td>Space typical state</td>
<td>Substructure type</td>
<td>Ventilation type</td>
<td>Filter type</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BMS</td>
<td>3876</td>
<td>5618</td>
<td>96</td>
<td>2235</td>
<td>1874</td>
<td></td>
</tr>
<tr>
<td>OMS</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>3876</td>
<td>5618</td>
<td>96</td>
<td>2235</td>
<td>1874</td>
<td></td>
</tr>
</tbody>
</table>
Figures 1–3). The results were uploaded into UNITE via its data management system PlutoF (https://plutof.ut.ee; Abarenkov et al. 2010) for open query by the scientific community and was shared with the INSDC as an Excel sheet (Suppl. material 2).

Discussion

The workshop compiled a total of 45,488 metadata items, making them available for scientific query through UNITE and other venues. These metadata, although typically “published” and thus “available”, were previously not open for direct query. This highlights the wealth of relevant scientific information that lies buried in the last few decades’ worth of scientific publications – formally available, yet only available to those who know where to look, and reachable only to those with access to that literature. Fortunately, we live in a digital age where the infrastructure for recovering and sharing such information is falling into place (Martin and Martin 2010). Furthermore, there is a growing awareness of the need to annotate newly generated sequences beyond the barest minimum when these are first deposited into public sequence databases (Hyde et al. 2013; Schoch et al. 2014). Such annotations unlock significant scientific potential of those molecular data, increase the citability of the underlying scientific studies, and
Figure 2. Krona chart of the taxonomic affiliation of the BMS sequences down to order level. The Krona chart lists all annotated BMS sequences except those classified as Fungi sp. (36.4%) and those of non-fungal origin (0.9%). An interactive version of the Krona chart is provided as Suppl. material 3. The figure includes pre-existing data plus the data added during the workshop, such that these charts indicate the scientific state of ITS-based Sanger-derived sequencing of the built mycobiome as of spring 2016. Sequences that were not annotated with a single built environment-related term in the INSDC were not included in this effort, and are not represented in these charts.

fulfill funding agencies’ demands for openness and maximum scientific use of research funding. We certainly hope that the mycological community will be quick to embrace a more integrative approach to sequence annotation. The public sequence databases can similarly make it even easier and faster to provide such metadata upon sequence submission. We speculate that excessive time consumption is the primary reason why some sequence depositors do not annotate their sequences as well as they could have.

We managed to process nearly all BMS sequences – for which we could retrieve the underlying publication(s) – for at least one metadata item. A total of 4,985 sequences
were false positives – our keywords indicated them to belong to the BMS whereas in reality they did not. A sequence could stem from “outside city hospital” (keyword “hospital”), for instance. These sequences were annotated for country and host of sampling, plus the nature of the relation to the host, whenever the underlying scientific study could be retrieved and interpreted. It is reasonable to assume that our initiative suffered from a fair number of false negatives as well – sequences that should have been a part of the BMS, but that were not. Although we used no fewer than 24 keywords in our efforts to capture the built environment, we presumably missed one or more important terms in the field. We similarly missed out on all built-environment sequences that featured no relevant annotation whatsoever – perhaps just a species name and the country of origin were available. Thus, whereas we managed to do at least something about nearly all BMS sequences we recovered, we do not claim to have annotated all public fungal ITS sequences from the built environment.

The workshop identified several potential venues for amendments to the MIxS-BE standard. For example, “floor” was found to be a common place for sampling of, e.g., dust, yet the data point of “floor” could not easily be fitted into any extant MIxS-BE category. Similarly, “air” could not be represented in a straightforward way in the MIxS-BE standard (but rather applied to other packages of the MIxS standard). We also felt the need for a “laboratory” flag to indicate that a sequence stemmed from sampling in a laboratory. In addition, we were surprised by the number of fungal sequences generated from environments that must be considered to qualify as “built” or at least altered by man, but that nevertheless were difficult to fit into the present MIxS-BE categories. The examples included tombs, crypts, and mummies (Šimonovičová et
al. 2015), tumuli and other prehistoric remains (e.g., Kiyuna et al. 2011; Fernandez-Cortes et al. 2011), spacecraft (Satoh et al. 2011), and indoor historical paintings or artifacts such as the Turin shroud (López-Miras et al. 2013; Barcaccia et al. 2015). In these cases, we tried to capture the essence of the underlying sequence entries to the extent that the MIxS-BE standard allowed. We used our free-text field “Comment” to provide additional information that we felt was important with respect to future queries of these entries. These potential venues for improvements of the MIxS-BE standard have been communicated to MIxS-BE representatives from the Genomic Standards Consortium’s MIxS Compliance and Implementation working group (http://gensc.org/mixs/mixs-compliance-and-implementation/).

Conclusions

The present study used a workshop-style approach to accomplish a task that would have taken several months for a single researcher to accomplish. Costs were kept low by recruiting many of the participants among local Ph.D. students and postdocs in systematics and ecology, and workshop participation was made attractive by providing the opportunity to contribute to this workshop report. We can recommend this model when tackling projects of a similar kind, such as data assembly and analysis in molecular ecology and systematics. As an added benefit, the more junior participants obtain experience in scientific collaboration and communication as well as in carrying out scientific projects (cf. Ryberg et al. 2016). The workshop was funded by an Alfred P. Sloan foundation grant to improve the support for the built mycobiome in UNITE and elsewhere. Other events include a forthcoming (2017) taxonomic sequence annotation workshop and the generation and public release of sequences from type material. We invite feedback and participation in these events, and we welcome any other idea to take molecular identification of the built mycobiome to the next level.

Acknowledgements

We gratefully acknowledge financial support from the Alfred P. Sloan Foundation and the Swedish Research Council of Environment, Agricultural Sciences, and Spatial Planning (FORMAS, 215-2011-498). VRMC thanks CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico (SWE 232695/2014-8) for providing a shared Ph.D. scholarship. A.A. is funded by the Swedish Research Council (B0569601), the European Research Council under the European Union’s Seventh Framework Programme (FP/2007-2013, ERC Grant Agreement n. 331024), and a Wallenberg Academy Fellowship.
References


Annotating public fungal ITS sequences from the built environment according to...


**Supplementary material 1**

**Keywords used to identify fungal sequences from the built environment in the INSDC**


Data type: text

Explanation note: Keywords used to identify fungal sequences from the built environment in the INSDC.

Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

---

**Supplementary material 2**

**Annotations made during the workshop**


Data type: metadata

Explanation note: The annotations made during the workshop shown with original INSDC data. For the BMS, we targeted nine MIxS-BE items plus country of collection, host of collection, host association, and a general “Comment” field. For the OMS, we targeted country of collection, host of collection, host association, and a general “Comment” field.

Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.
Supplementary material 3

Krona chart
Data type: html
Explanation note: Interactive Krona chart for visualizing the taxonomic distribution of annotated BMS sequences down to order level. Sequences classified as Fungi sp. (36.4%) or non-fungal (0.9%) were excluded from this dataset.
Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.