LARGE-SCALE BARCODING OF FUNGAL COLLECTIONS

Todd W. Osmundson & Matteo M. Garbelotto
University of California, Berkeley, CA, USA
Acknowledgements:

**Venice:**
- Giovanni Robich
- Luca Mizzan
- Amy Smith
- Lydia Baker

**Moorea:**
- Sarah Bergemann
- Neil Davies
- Chris Meyer
- Lydia Smith
- Lydia Baker

Gordon & Betty Moore Foundation
- Fungi are diverse and poorly-known
- Most are cryptic over the majority of their life cycle
**Unknowns:**

- Diversity of life cycle stages (e.g., endophyte to pathogen transitions)
- Host ranges
- Geographic ranges / biogeography
- Community ecology

DNA Barcoding is a useful and important tool for identifying fungi and thereby understanding these factors...

... However, utility dependent upon existence of comprehensive sequence databases and methods for confidently assigning taxonomic identities via sequence comparisons.
Environmental DNA sequencing has generated large numbers of “insufficiently identified” sequences in INSD; will increase with high-throughput sequencing.

Causes:
- Many identified species not sequenced
- Unknown taxa

Fig. 1 – Rate of description of new species recorded in *Index Fungorum*, 2000–2009, including those with (black) and without (white) sequences of any locus now present in GenBank.

Moorea: 62% exhibit < 97% similarity to GenBank sequences; 23% exhibit best BLAST match to environmental sequence (cultured or uncultured).
Having a sequence first can be useful
- Jones et al. 2011, *Nature*; Cryptomycota
- Nilsson et al. 2011, *Cladistics*; phylogenetic utility
- Naming sequence OTUs (MOTUs) can provide a common language

2 Charges:
- Sequencing known taxa
- Gaining biological understanding for sequences that do not yet correspond to a “known” taxon (community ecology; sequence-generated autecological studies; phylogenetic understanding)
Large-scale barcoding projects can aid in closing the sequencing gap. 2 recent projects:

Barcoding the Venice Museum of Natural History Fungal Collection (Institutional)

The Moorea Biocode Project (Geographic)
Questions:

1. How can we streamline the sequencing process?
2. How can we streamline and address issues in data management?
3. How can we improve quality control?
4. How can we pay for it?
How can we streamline the sequencing process?

- (A.) Rapid extraction techniques (Osmundson et al., in review)
  - NaOH works for most applications (ITS, multicopy microsatellites); ROSE better for *Phytophthora ramorum* detection in tanoak leaves
  - Extract dilutions viable for at least 3 years
  - Improved sequence success from contaminated sporocarps
How can we streamline the sequencing process?

- (B.) Triage: What are our community priorities?
- (C.) Centralized sequencing facilities
  - Cost reduction
  - Uniform methodologies
  - High throughput
  - Centralized data management expertise
How can we streamline the sequencing process?

- (D.) High-throughput methodologies
  - Limits: read lengths, number of multiplex tags
  - Ion Torrent chip $99; ½ the cost of a Sanger plate; how many specimens can we fit on it?
  - Multi-locus barcodes are coming
- (E.) Mini-barcodes
- (F.) Volunteer lab help (e.g., local mycophiles)

<table>
<thead>
<tr>
<th></th>
<th>Complete</th>
<th>ITS1</th>
<th>ITS2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ITS1</td>
<td>0.7317</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>ITS2</td>
<td>0.6430</td>
<td>0.5433</td>
<td>1</td>
</tr>
</tbody>
</table>
How can we streamline and address issues in data management?

2 components:

- **Sequence data**
  - Venice: brute force approach to contig editing, BLAST verification, etc.
  - GenBank submission: tbl2asn; perl
  - Moorea Biocode workbench: track lab workflows and run batch BLAST searches
- Automation:
  - Feature annotation tables
  - Error checking/contig assembly
  - BLAST verification
  - Sequence submission
  - Taxon naming (eliminating “back and forth”)
How can we streamline and address issues in data management?

- **Metadata**
  - Venice: source annotation tables prepared by hand
  - Moorea Biocode workbench: FIMS + LIMS $\rightarrow$ GenBank
Certainly, more metadata can go into GB record than usually does, but still many additional types:

- Taxonomic (Mycobank, Index Fungorum)
- Range (GBIF, Mushroom Observer)
- Photographs (MO, Encyclopedia of Life, Wikipedia)
- Rarity and conservation status
- Identification (Pacific NW Key Council, Fusarium ID)

Data linkage – output of one program converted to input of another (“data permaculture”)
A field guide for the (early) 21st century: Static → Dynamic
A field guide for the (early) 21st century: Static → Dynamic
How can we improve quality control?

• Lab-level (chimeras, contaminants, etc.)
• Taxonomic (is the ID correct?)
  • No easy answers
  • An iterative, collaborative process
    • Collaborations with taxonomic specialists: important specimens get sequenced, and are well identified (e.g., Vellinga, Halling & Hibbett, Barge, Jarvis, Wenck-Reilly, Smith), and sometimes extra lab help!
• Taxonomic (continued)
  • 3rd-party annotations of sequence records

- GenBank note: “Name taken from herbarium label; specimen not examined”
How can we pay for it?

- We have the advantage of less animosity between “taxonomic” and “molecular” contingents, due to cryptic manifestations and paucity of morphological characters
- Presenting a unified front; making the case for how essential molecular data are to much of the work that we do; the Deep Hypha model
How can we pay for it?

Public goods and the “free rider problem”
• Who are the end-users of fungal biodiversity data?
  • Taxonomists
  • Ecologists
  • Mushroom clubs
  • Commercial harvesters
  • Land managers
• Potential funding sources
  • Request grant funding for barcoding component in ecological and taxonomic studies
  • Mushroom clubs
  • Publishers of field guides (sequence data as a “value-added” component)
  • Small grants from parks, etc. (e.g., Pt. Reyes)
  • Rent capture on mushroom harvests
  • PI research gift funds
A mycoflora-scale barcode database is a big job, but it can be done.

Main elements:

1. Methodological advances (sequencing, data management, IT, bioinformatics)
2. Collaboration (systematists, ecologists, mushroom clubs, knowledgeable “amateurs,” bioinformaticians, etc.)
3. Unified front for requesting funding and doing the work