Taxonomy and Digital Sequence Information

Chris Lyal

NATURAL HISTORY MUSEUM

Scope of my discussion: science

- Focus on non-commercial taxonomy
- Taxonomists and systematists have the task of documenting the natural world
  - what are its components?
  - How can they be identified?
  - How are they related?
- This work is not simply academic
- The Convention on Biological Diversity has a cross-cutting issue – the Global Taxonomy Initiative
  - To support Parties' implementation of the Convention

gti Global Taxonomy Initiative

Scope of my discussion: 'DSI'

- Focus on nucleotide sequence data
  - The aspect of DSI most used by taxonomy
  - Increasing value as a tool for identification
  - (relatively) Clearly understood
- Considered as 'DSI' only when retrieved from / used on a 3rd party source

Scope of my discussion: 'DSI'

- Almost all relevant information is stored digitally (morphological, behavioural, nomenclatural etc etc)
- And may be used by taxonomic studies

McKenna *et al. Genome Biology* (2016) 17:227
DOI 10.1186/s13059-016-1088-8

Genome Biology

RESEARCH Open Access

Genome of the Asian longhorned beetle (*Anoplophora glabripennis*), a globally significant invasive species, reveals key functional and evolutionary innovations at the beetle–plant interface

Legislative challenge

- Understanding of ABS and the Nagoya Protocol is high and still becoming more refined
  - Workflows are adapted to bring in compliance measures
- Increasing number of countries enacting legislation covering DSI, but
  - weak and varied definitions are challenging
  - few taxonomists aware of these regulations, or understand their application
  - Addressing DSI in a bilateral manner does not match how it is accessed and used
  - Adaptation of workflows potentially not realistically possible in all cases

Taxonomists and DSI

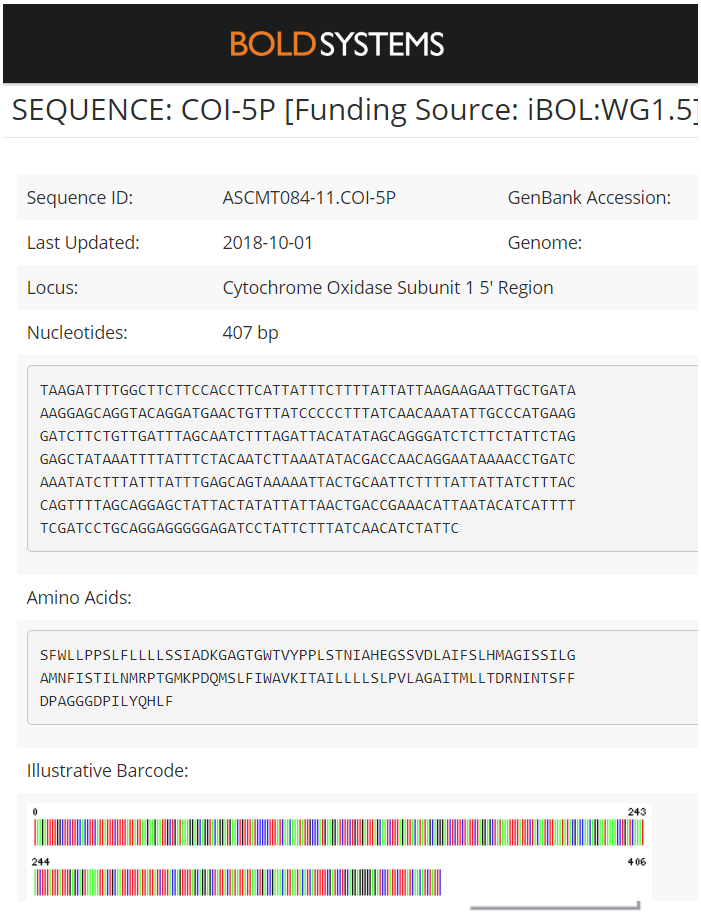
- Taxonomists:
  - Create DSI
  - Publish DSI in databases and scientific papers
  - Use DSI for:
    - Identification
    - monitoring
    - (species description)
    - phylogenetic analysis

Where do taxonomists obtain sequence data?

- Generated during research
  - from recent GR accessed with PIC & MAT
  - from older specimens in collections
- In-house databases
  - Developed from earlier sequencing activities
- Public databases, particularly:
  - International Nucleotide Sequence Database Collaboration (INSDC)
    - DNA Data Bank of Japan (DDBJ)
    - European Bioinformatics Institute (EMBL-EBI)
    - National Center for Biotechnology Information (NCBII) (Genbank)
  - Barcode of Life Data System (BOLD)

Identification

- Comparison of sequences
- Generally use ‘DNA barcodes’
  - COI gene ‘standard’ for many animals
  - Different genes for plants, bacteria etc
- Run ‘BLAST’ search
  - finds regions of similarity between sequences
  - Potentially could read all sequences in the database
  - A match suggests an identification



e-DNA and monitoring

- Check for presence or absence of endangered or invasive species
- Detect unknown species
- Assess overall biodiversity
- Increasingly important tool for environmental management
- Likely to use databases to identify sequences or develop specific probes

Sequence data in species description

- Increasingly NSD used in species description
- So DNA libraries will be required for identification

Research Article
Deutsche Entomologische Zeitschrift 66(2): 119-145
https://doi.org/10.3897/dez.66.34683 (25 Jul 2019)

A revolutionary protocol to describe understudied hyperdiverse taxa and overcome the taxonomic impediment

Zelomorpha angelsolisi Meierotto, sp. nov.

http://zoobank.org/82FC9D54-84E0-470B-8A02-2F80FBFAC5B6

Figure 2

Molecular diagnosis

Nucleotides 43–45 TTA, 54–57 CTTT, 75 G, 136–138 GTG, 165 T, 321 G, 417 G, 462 G, 477 C, 561 G, 684 G.

Phylogenetic analysis

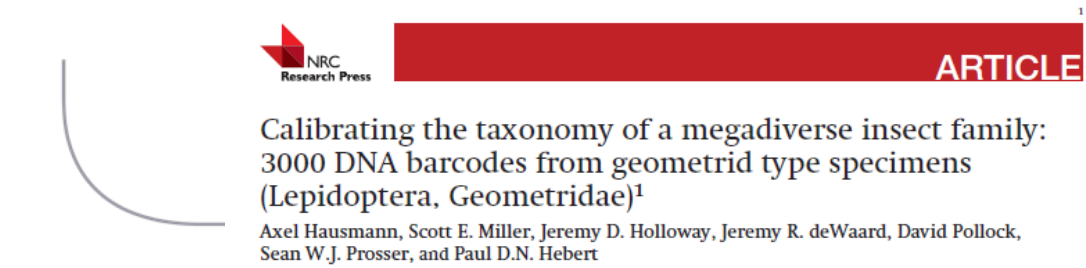
- Phylogenetic analysis
  - Use multiple genes (different genes evolve at different rates)
  - Increasing use genomes or genome skimming
  - Typically from many countries, collected over many years
  - Analysis may include many species, increasingly in hundreds
  - With standardisation of sequencing methodology, downloaded sequences increasingly useful

Phylogenetic analysis

- Geographic origin of sequences not always recorded in paper
- An individual sequence may be one of hundreds
- Like other research covering many taxa, will not necessarily have a Brazilian researcher involved

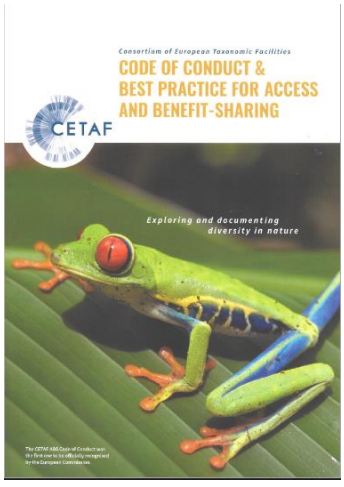
Publication

- Publication
  - Research almost always intended for publication
  - Standard condition of publication: data are made available
  - So sequences placed (published) on BOLD / INSCD etc.



How do taxonomists manage their work?

- Developing Best Practices
  - Currently focus on utilisation of GR and aTK, not DSI
  - Intended to assist sector's compliance with ABS & Nagoya Protocol



Some Questions

- What activities fall in scope of the Brazilian legislation on *in silico* genetic heritage?
- Does the inclusion of any NSD of Brazilian origin in a research paper require registration?
  - Is NSD generated from specimens collected pre 1962 considered Brazil's *in silico* genetic heritage?
  - If only one of several hundred sequences used is Brazilian in origin is registration required?
- How should we be considering non-monetary benefit-sharing?