“Digital Sequence Information” and taxonomy

Chris Lyal

NATURAL HISTORY MUSEUM

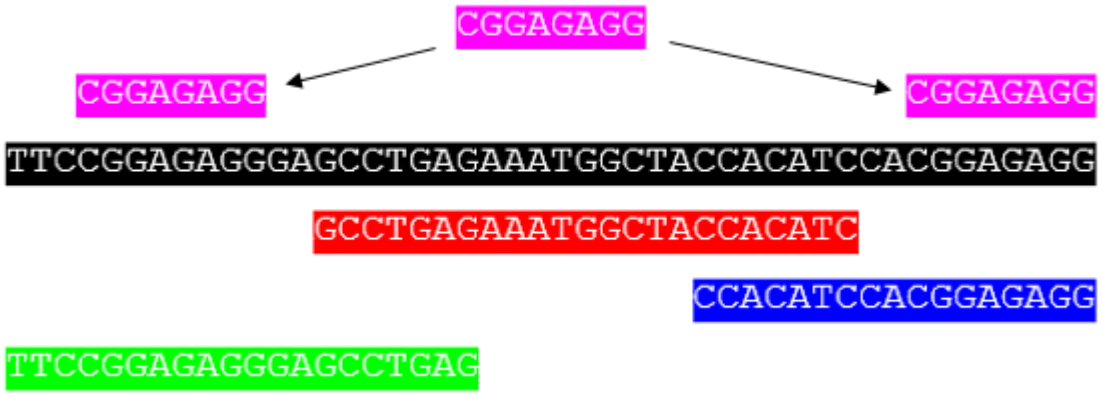
Scope of my discussion

- Focus on taxonomy
  - Particularly in a non-commercial context
  - Includes particularly
    - identification
    - species description
    - phylogenetic analysis
- CBD Parties have asked taxonomy and other non-commercial biodiversity science to deliver outputs to support implementation

Global Taxonomy Initiative

Scope of my discussion

- Digital Sequence Information
  - Sequence reads
  - Sequence assembly (putting the reads together)



Scope of my discussion

- Much more rarely used in taxonomy:
  - Gene functionality
  - Biochemistry

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

Where do taxonomists obtain molecular sequence information?

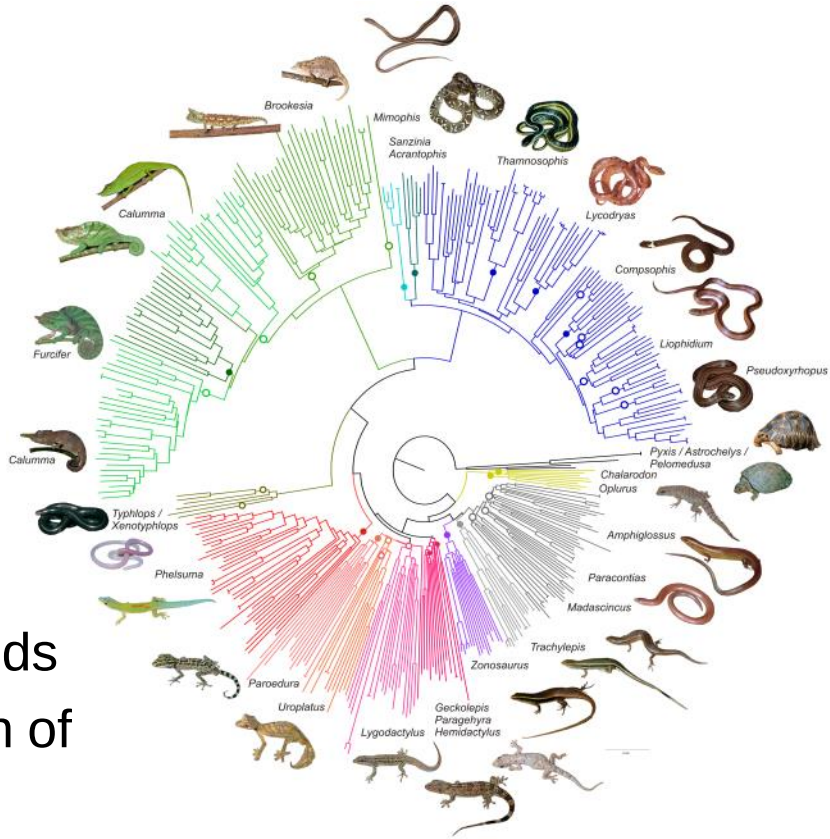
- Generated during research
  - from recent GR accessed with PIC & MAT
  - from older specimens in collections
- In-house databases
  - Developed from earlier sequencing activities
- Public databases, particularly:
  - International Nucleotide Sequence Database Collaboration (INSDC)
    - DNA Data Bank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

What do taxonomists do with DSI?

- Identification
  - Increasing use of ‘DNA barcodes’ - COI gene is ‘standard’ for many animals
  - Run ‘BLAST’ search on BOLD
    - finds regions of similarity between biological sequences
    - Potentially could read all sequences in the database
  - A match suggests an identification
    - Increasingly used for pests, invasive species etc
  - CBD (GTI) has funded training in Barcode use for relevant personnel from a number of Parties

What do taxonomists do with DSI?

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



What do taxonomists do with DSI? – e-DNA

- Check for presence or absence of endangered or invasive species
- Detect unknown species
- Increasingly important tool for environmental management
- May use Databases to identify sequences

What do taxonomists do with DSI?

- Publication
  - Research is almost always intended for publication
  - Standard condition of publication is that data are made available
  - So sequences placed on BOLD / INSCD etc.
  - INSDC databases built on principle of free availability of data

NRC Research Press

ARTICLE

Calibrating the taxonomy of a megadiverse insect family: 3000 DNA barcodes from geometrid type specimens (Lepidoptera, Geometridae)[1]

Axel Hausmann, Scott E. Miller, Jeremy D. Holloway, Jeremy R. deWaard, David Pollock, Sean W.J. Prosser, and Paul D.N. Hebert

What do taxonomists do with DSI?

- Exchange Information
  - Databases use data formats exist to exchange information about specimens / samples / sequences
  - Global Genome Biodiversity Network (GGBN) has added elements to manage permit data
  - CETAF has developed ‘data use statement’ to alert 3rd Party users of limitations in use

DATABASE
The Journal of Biological Databases and Curation

Database, 2016, 1–11
doi: 10.1093/database/baw125
Original article

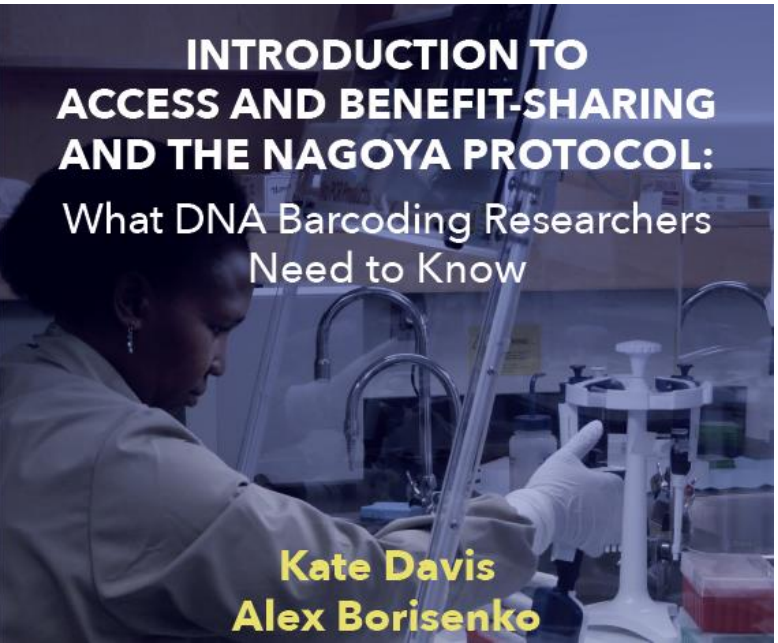
Original article

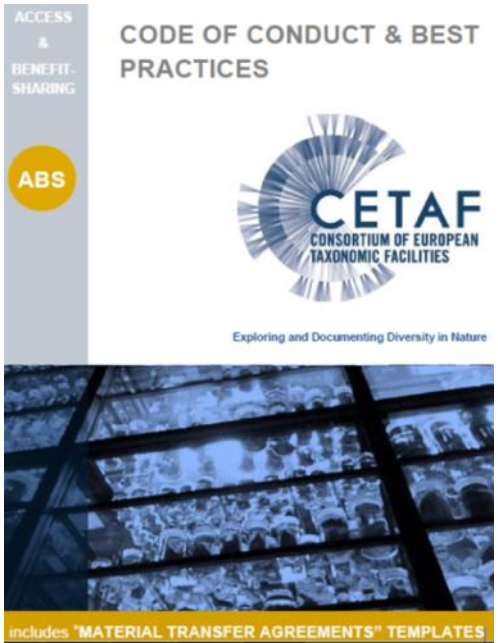
The Global Genome Biodiversity Network (GGBN) Data Standard specification

G. Droege[1,*], K. Barker[2], O. Seberg[3], J. Coddington[2], E. Benson[4], W. G. Berendsohn[1], B. Bunk[5], C. Butler[2], E. M. Cawsey[6], J. Deck[7], M. Döring[8], P. Flemons[9], B. Gemeinholzer[10], A. Güntsch[1], T. Hollowell[2], P. Kelbert[1], I. Kostadinov[11], R. Kottmann[12], R. T. Lawlor[13], C. Lyal[14],

How do taxonomists manage their work?

- Developing Best Practices
  - Currently focus on utilisation of GR and aTK
  - Intended to assist in Nagoya Protocol compliance by sector

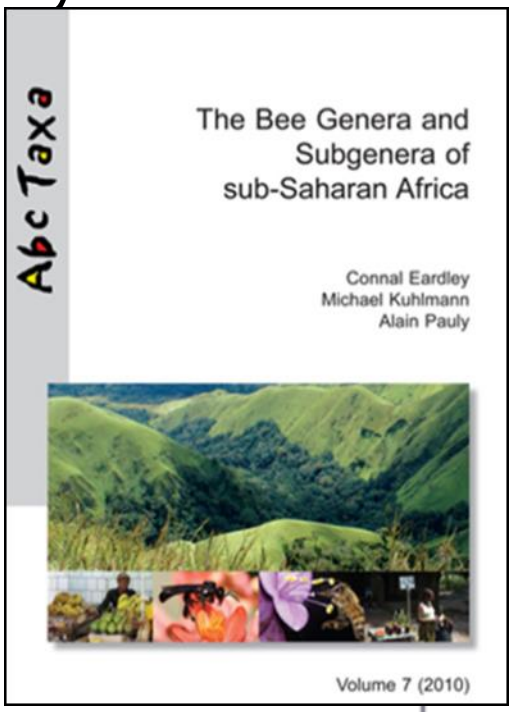




Application of PIC and MAT to DSI?

What benefits should be shared?

- Non-commercial taxonomic research typically delivers non-monetary benefits:
  - Capacity-building
    - Training
    - Equipment
    - Reference collections
  - Taxonomic information
  - Collaboration



CBD needs met by use of DSI

- Rapid species identification
  - Use of DNA barcode libraries in particular facilitates this.
- Particularly important for samples where rapid response is needed
  - Invasive species
  - Quarantine
  - Pest detection

Global Taxonomy Initiative

CBD needs met by use of DSI

- Monitoring
  - eDNA sampling can deliver indication if target species is present
  - Or the overall diversity of an environment

frontiers in Marine Science

REVIEW
published: 09 May 2017
doi: 10.3389/fmars.2017.00107

Check for updates

DNA Sequencing as a Tool to Monitor Marine Ecological Status

Kelly D. Goodwin[1], Luke R. Thompson[1,2], Bernardo Duarte[3], Tim Kahlke[4], Andrew R. Thompson[5], João C. Marques[6] and Isabel Caçador[3]*

gti Global Taxonomy Initiative

CBD needs met by use of DSI

- CBD COP has repeatedly called for greater access to information of many types, including genetic information.
- Global taxonomic information system called for under Global Taxonomy Initiative
- Strategic Plan for Biodiversity 2011-2020 (XIII/31), Aichi goals C (Target 13) and E (Target 19)

gti Global Taxonomy Initiative

CBD needs met by use of DSI

- No country holds sequence data for all of its biota and species likely to be intercepted by quarantine (Alien Species, pests etc.)

- The only way in which Parties can obtain sequence data for supporting implementation of the Convention is through freely-available global databases.

Global Taxonomy Initiative

CBD needs met by use of DSI

- Wide range of support for implementation identified in Report to CBD on DSI

Convention on Biological Diversity

GENERAL

CBD/DSI/AHTEG/2018/1/3
12 January 2018

ENGLISH ONLY

AD HOC TECHNICAL EXPERT GROUP ON
DIGITAL SEQUENCE INFORMATION ON
GENETIC RESOURCES
Montreal, Canada, 13-16 February 2018

FACT-FINDING AND SCOPING STUDY ON DIGITAL SEQUENCE INFORMATION ON GENETIC RESOURCES IN THE CONTEXT OF THE CONVENTION ON BIOLOGICAL DIVERSITY AND THE NAGOYA PROTOCOL

Note by the Executive Secretary

CBD needs met by use of DSI

CBD (GTI) - funded training in Barcoding

- 3 rounds so far
- In 2018:
  - 11 training courses (Belarus, Bhutan, Brazil, Colombia, Nigeria, Sri Lanka, Suriname, Tunisia, Turkey, Philippines, Uruguay)
  - 89 individuals will be trained in DNA barcoding, including cross-border regulatory authorities, forestry/ fisheries authorities and protected areas managers.
  - Nine DNA libraries will be established within developing countries.
  - DNA libraries on species of countries' concern (invasive alien species, threatened species, agricultural pests) will be added to two existing DNA libraries in Brazil and Colombia

Application of PIC and MAT to DSI?

- What is utilisation?
  - BLAST search, when the overwhelming majority of sequences are merely looked at and discarded as a match?
    - The match is ‘most similar to’, and the digital sequence is not used further
  - Download for inclusion in analysis?
    - Me be one in hundreds or thousands of sequences used

Application of PIC and MAT to DSI?

Could a system handle bilateral agreements?

- International Nucleotide Sequence Database Collaboration (INSDC)
  - Share data daily
  - hold quadrillions (>10 to the 15th) of nucleotides of DNA sequences
  - 201,663,568 sequences in June 2017
  - collected from over 300,000 organisms
  - EMBL-EBI search engine runs ca. 12.6 million jobs every month for users
  - scientists at over 5 million unique sites use EMBL-EBI websites every month;
  - every weekday, more than 27 million requests are made to EMBL-EBI websites
- BOLD currently holds 1.3 million public records of the COI gene

Application of PIC and MAT to DSI?

- Very large number of transactions;
- Often very low incremental value of a single sequence to the research
- Risk that requiring separate PIC and MAT halts use
  - Reducing or eliminating non-monetary benefits
  - And acting directly against CBD implementation

Application of PIC and MAT to DSI?

Can the delivery of non-monetary benefits be improved?

- While there are few if any standard clauses:
- Most permits / MAT for taxonomic research are congruent
  - Information / capacity building

Application of PIC and MAT to DSI?

- Public databases are a means of delivering information
  - Genetic sequence data (e.g. INSDC, BOLD)
  - Scientific publications (e.g. Biodiversity Heritage Library, Open Access Publications)
  - Occurrence data (e.g. GBIF)
- Information pipelines are built and being populated
- Challenge now is to build capacity to make use of them
- And adapt workflows to be able to use the information

Global Taxonomy Initiative

Summary

- Digital sequence data are of major and increasing value to taxonomists globally;
- The use of the data can directly support implementation of the CBD, and other national priorities;
- Data are open to (and used by) workers globally, not just the North;
- Applying a bilateral model to agreeing PIC and MAT would be challenging given the number of transactions and the low incremental value of each sequence;
- A rate-limiting modality would impact on CBD implementation and on research;
- A key challenge is to improve the ability of Parties to make use of data and information shared through global systems, and make use of the benefits so shared.