7

The Information Systems for DNA Barcode Data

Di Liu and Juncai Ma
Network Information Center, Institute of Microbiology, Chinese Academy of Sciences
WFCC-MIRCEN World Data Centre for Microorganisms (WDCM)
China, People's Republic

1. Introduction

DNA barcoding is a novel concept for the taxonomic identification, in that it uses a specific short genetic marker in an organism's DNA to discriminate species. In 2003, professor Paul D. N. Hebert, "the father of DNA barcoding", of the University of Guelph, Ontario, Canada first proposed the idea to identify biological species using DNA barcode, where the mitochondrial gene cytochrome *c* oxidase subunit I (COI) was supposed to be the first candidate for animals (Hebert et al. 2003a). Their studies of COI profiling in both higher taxonomic categories and species-level assignment demonstrated that COI gene has significant resolutions across the animal kingdom except the phylum Cnidaria (Hebert et al. 2003b, Ward et al. 2005, Hajibabaei et al. 2006). From then on, a wide broad of taxonomic groups (i.e. birds, fish, butterflies, spiders, ants, etc) were examined by COI gene for its usability as the barcode (i.e. Hebert et al. 2004a, Hebert et al. 2004b, Greenstone et al. 2005, Smith et al. 2005, Barber and Boyce 2006, Meier et al. 2006, Kerr et al. 2007, Kumar et al. 2007, Pfenninger et al. 2007, Stahls and Savolainen 2008, Zhou et al. 2009). Meanwhile, other candidate genes, including Internal Transcribed Spacer (ITS), trnH-psbA intergenic spacer (trnH-psbA), Ribulose-bisphosphate carboxylase (rbcL) and Maturase K (matK) were analysed by different research groups (Jaklitsch et al. 2006, Evans et al. 2007, Ran et al. 2010, de Groot et al. 2011, Liu et al. 2011, Piredda et al. 2011, Yesson et al. 2011). Till recently, there are about 30 DNA barcode candidates are tested, and 4 to 8 of them are widely used for the identification of diversified taxonomic groups with a relatively good resolution.

It has been estimated that there are 10 to 100 million species of living creatures in the earth, while what we know is very limited. Knowing the biodiversity is one of the crucial biological issues of ecology, evolutionary biology, bio-security, agro-biotechnology, bio-resources and many other areas. For very long period, taxonomists have provided a nomenclatural hierarchy and key prerequisites for the society. However, the needs for species identification requested by non-taxonomists require the knowledge held by taxonomists. Therefore, a standardized, rapid and inexpensive species identification approach is needed to establish for the non-specialists. There had some attempts on the molecular identification systems based on polymerase chain reaction (PCR), especially in bacterial studies (Woese 1996, Zhou et al. 1997, Maiden et al. 1998, Wirth et al. 2006), but no successful solutions for broader scopes of eukaryotes (reviewed in Frezal and Leblois 2008). The DNA Barcode of Life project is another attempt to create a universal eukaryotic identification system based on molecular approaches. Following studies by Hebert et al.

(Hebert et al. 2003a, Hebert et al. 2003b), the Consortium for the Barcode of Life (CBOL) was initiated in 2004, and aimed to produce a DNA barcode reference library and diagnostic tools based on the taxonomic knowledge to serve taxonomists and non-taxonomists (Schindel and Miller 2005). It should note that the DNA Barcode of Life project is neither to build the tree of life nor molecular taxonomy (Ebach and Holdrege 2005, Gregory 2005). From the establishment of CBOL, more than 50 countries have been participated in and devoted themselves into this ultimate mission. One of the important projects is the International Barcode of Life project (iBOL) sponsored by Canada government (details will be described below). Till now, DNA barcoding is accepted by a great range of scientists and has achieved indisputable success (Teletchea 2010).

One of the major aims of bioinformatics is to finely store and manage the huge amount of biological data. Apart from genome sequencing projects, DNA barcoding projects are going to establish another important biological data resource to the public. Until now, there are about half a million DNA barcodes are submitted to GenBank from the Barcode of Life Data System (BOLD) (Ratnasingham and Hebert 2007), the most essential data center for barcode of life projects. Besides, large amount of DNA barcode data are under producing and to be released worldwide. It has been estimated that more than 100 million barcode records will be generated for the animal kingdom(Ratnasingham and Hebert 2007), and that size is comparable to the current GenBank release (Benson et al. 2011). Unlike the traditional nucleotide sequences deposit in the international nucleotide sequence databases collaboration (INSDC), DNA barcode data comprises comprehensive data types, including photos, DNA chromatogram (trace files), geographic data and structured morphological information of each specimen. Therefore novel information systems are required to be developed to collect, store, manage, visualize, distribute, and utilize these data for species identification, clustering/classification as well as evolutionary studies. Moreover, applying the second-generation sequencing technology (e.g. Roche 454) for DNA barcoding, especially for those environmental samples (e.g. mud, water) is under developing, and this will generate a large amount of DNA barcodes a time, with the data files different from those from traditional DNA analyser implementing Sanger sequencing approach. Hence, methods to manage and utilize the output from 2nd-generation sequencers are also to be developed. Besides, it is still a great challenge to integrate the DNA barcode data into the studies of metagenomics (Venter et al. 2004, Rusch et al. 2007).

In this chapter, we will first review the current progresses of DNA barcode of life projects, and then we will describe the data schema and the information systems of DNA barcode data. Particularly, three types of DNA barcode information systems are to be introduced: BOLD, by now the best information system for DNA barcoding with highly integration; Global Mirror System of DNA Barcode Data (GMS-DBD), the mirror system for the distribution of the publicly available DNA barcode data worldwide; and the management system for Chinese DNA barcode data, which is a manageable information system for DNA barcoding groups.

2. DNA barcode and the international collaborations

In order to make the concept that using DNA barcodes to identify species being reality, great efforts need to be contributed by the nations. After the launch in 2004, CBOL has gathered more than 150 organizations around the world, including natural history museums, zoos, herbaria, botanical gardens, university departments, governmental

organizations and private companies. The goals of CBOL include building up a DNA barcode library of the eukaryotic lives in 20 years. In July 2009, the Ontario Genomics Institute (OGI), Canada initiated the iBOL project, which was considered as the extension and expansion of the previous project, Canadian Barcode of Life Network (http://www.bolnet.ca) launched in 2005. Nowadays, there are 27 countries as partner nations participated in this international collaboration and nearly 20 established campaigns were co-working for iBOL on some specific creatures.

2.1 The concept of DNA barcode and the commonly used ones

DNA barcode is a segment of DNA that possesses the following features. a) DNA barcode is conserved in a broad range of species, so that a conserved pair (or several conserved pairs) of primers can be designed and applied for DNA amplification; b) it is orthologous; c) DNA barcode must evolve rapidly enough to represent the differences between species; d) DNA barcode needs to be short, so that a single DNA sequencing reaction is enough to obtain the sequence; and e) DNA barcode needs to be long to be capable of holding all substitutions within higher taxonomic groups (For example, a 500-base pair (bp)-long DNA barcode has the capability to hold 4^{500} possible differences to discriminate species.).

The first DNA barcode is a 658-bp long region DNA segment of COI gene within mitochondria, with primers LCO1490 (5′-GGTCAACAAATCATAAAGATATTGG-3′) and HCO2198 (5′-TAAACTTCAGGGTGACCAAAAAATCA-3′) used for DNA amplification (Hebert et al. 2003a). COI gene as the primary DNA barcode has been proven to be useful in broad ranges of animal species, despite of the limitations in some taxa (Meyer and Paulay 2005, Vences et al. 2005). In fungi, ITS was chosen as the main DNA barcode and was confirmed by the sequences within the international nucleotide sequence databases (Nilsson et al. 2006), though COI was examined applicable in Penicillium (Seifert et al. 2007). In plants, mitochondrial DNA shows intra-molecular recombination and COI gene has lower evolutionary rate (Mower et al. 2007), so that genes on the plasmid genome were examined, e.g. *rpoB*, *rpoC1*, *rbcL* and *matK*. Meanwhile, some intergenic spacers (e.g. trnH-psbA, atpF-atpH and psbK-psbI (Fazekas et al. 2008)), and markers' recombination (e.g. *rbcL* and trnH-psbA (Kress and Erickson 2007)) were tested, too. Nevertheless, those choices either meet the amplification problems or standardization problems. Recently, CBOL Plant Working Group recommended the combination of *rbcL* and *matK* for plant DNA barcoding (CBOL Plant Working Group, 2009).

2.2 The international Barcode of Life project

The main mission of iBOL is "extending the geographic and taxonomic coverage of the barcode reference library -- Barcode of Life Data Systems (BOLD) -- storing the resulting barcode records, providing community access to the knowledge they represent and creating new devices to ensure global access to this information." (http://ibol.org/about-us/what-is-ibol/). To accomplish the mission step by step, iBOL announced the first 5-year plan that is to collect and process 5 million samples covering 500 thousand species with $150 million budget. Then six working groups were established to work on barcode library construction, methodology, informatics, technology, administration and social issues (Table 1). The first two working groups are mainly focusing on the collection and production of DNA barcodes in various living creatures, biotas and specimens in museums. The third working group is dedicated to the construction of the informatics, including the core functionality and mirror

sites. Core functionality comprises at least a sophisticated bioinformatics platform with the integration of a robust IT infrastructure (computational note, storage and network), DNA barcode databases and analytical tools. Meanwhile, the mirror sites help to strengthen data security and accessibility. Working group 4 is focusing on future technologies, either applying the latest sequence techniques or developing the portable mobile devices. Although working groups 5 and 6 are not purely on the science and technology of DNA barcoding, the administration and dealing with social aspects are far more important to the success of the project.

iBOL Working Group	Sub-Working Group
WG 1. Barcode Library: Building the digital library of life on Earth	WG 1.1, Vertebrates
	WG 1.2, Land plants
	WG 1.3, Fungi
	WG 1.4, Animal Parasites, Pathogens & Vectors
	WG 1.5, Agriculatural and Forestry Pests and Their Parasitoids
	WG 1.6, Pollinators
	WG 1.7, Freshwater Bio-surveillance
	WG 1.8, Marine Bio-surveillance
	WG 1.9, Terrestrial Bio-surveillance
	WG 1.10, Polar Life
WG 2. Methods: Extending the horizons of barcoding	WG 2.1, Barcoding Biotas
	WG 2.2, Museum Life
	WG 2.3, Methodological Innovation
	WG 2.4, Paleobarcoding
WG 3. Informatics: Storing and analyzing barcode data	WG 3.1, Core Functionality
	WG 3.2, Mirrors
WG 4. Technologies	WG 4.1, Environmental Barcoding
	WG 4.2, Mobile Barcoding
WG 5. Administration: Consolidating the matrix	WG 5.1, Project Management
	WG 5.2, Communications
WG 6. GE^{3}LS	WG 6.1, Equitable Use of Genetic Resources
	WG 6.2, Regulation and International Trade
	WG 6.3, Intellectual Property and Knowledge Management
	WG 6.4, Education Initiatives for Schools and Media
	WG 6.5, Governance of Knowledge Mobilization

Table 1. iBOL working groups.

By the end of the year 2010, iBOL reported the progresses of each working group in the iBOL Project Interim Review (http://ibol.org/interim-review/). During the first 18 months, iBOL has produced DNA barcodes for 153K species from 326K specimens collected worldwide, and obtained exciting results from barcoding biotas of the locales Moorea and Churchill, where a great number of additional species were revealed by DNA barcoding. BOLD as the core functionality of iBOL has increased the number of records to 1.1 million and the number of users to 6000. The power of storage and computing was also improved dramatically. Conclusively, all working groups have made substantial progresses towards the final goals.

The achievements made by iBOL are with the help of the campaigns of barcode of life, which consists of researchers with similar interests on specific families and regions of life (e.g. birds, fish, etc.). Most of the campaigns are working closely with the relevant iBOL working groups and/or BOLD. Below lists some useful websites and campaigns of the international collaborations (Table 2).

Short Name	Description	URL
CBOL	The consortium for the barcode of life	http://www.barcoding.si.edu; http://www.barcodeoflife.org
iBOL	The international barcode of life project	http://www.ibol.org
CCDB	Canadian centre for DNA barcoding	http://www.danbarcoding.ca
BOLD	Barcode of life data systems	http://www.boldsystems.org
GMS-DBD	Global mirror system of DNA barcode data	http://www.boldmirror.net
Fish-BOL	Fish barcode of life initiative	http://www.fishbol.org
ABBI	All birds barcoding initiative	http://www.barcodingbirds.org
PolarBOL	Polar barcode of life	http://www.polarbarcoding.org
Bee-BOL	Bee barcode of life initiative	http://www.bee-bol.org
MarBOL	Marine barcode of life	http://www.marinebarcoding.org
	Lepidoptera barcode of life	http://lepbarcoding.org
	Trichoptera barcode of life	http://trichopterabol.org
	Formicidae barcode of life	http://www.formicidaebol.org
	Coral beef barcode of life	http://www.reefbarcoding.org
	Mammal barcode of life	http://www.mammaliabol.org
	Sponge barcoding project	http://www.spongebarcoding.org

Table 2. Websites of DNA barcode of life projects worldwide

2.3 DNA barcode of life projects in China

There are three categories of the participated nations of iBOL, the National Nodes, the Regional Nodes and the Central Nodes. The National Node is primarily to collect, identify and curate the specimens from their territory, and the Regional Node has additional duties to participate in DNA barcode acquisition. As for a Central Node, it has not only National Node and Regional Node's missions, but also to maintain core sequencing facilities and the bioinformatics facilities, as well as to help share DNA barcode records with all nations. Of the current 27 nations participated in iBOL (27 nations are shown in iBOL website, where there are 33 nations in the iBOL Project Interim Review), China is acting as one of the four

Central Nodes, while the others are Canada, United States, and the European Union (France, Germany, Netherlands, etc.).

To better support the international collaborations and to take great part in the iBOL project, China has established the China National Committee for iBOL project. Prof. Jiangyang Li, Vice President of Chinese Academy of Sciences (CAS), acts as President, and Prof. Zhibin Zhang of the Bureau of Life Sciences and Biotechnology, CAS and Prof. Yaping Zhang of the Kunming Institute of Zoology, CAS are taking the roles of Vice President. From then on, the constructions of the core sequencing facilities and bioinformatics facilities for DNA barcoding were initiated, and varied foundations of China, including CAS, the Ministry of Science and Technology (MOST) and the Natural Science Foundations of China (NSFC) have issued projects of DNA barcode of life (Table 3). The projects covered the studies on diversified creatures, including animal, plant and fungus, and the studies on specimen collection, data production, theory and methodology, database and information system, etc. To date, China has established three main research campaigns, working on animal, plant and fungal DNA barcoding respectively, and initiated the constructions of China DNA Barcoding Centre and China DNA Barcode of Life Information Centre by the institutes of CAS.

Projects issued by	Projects aims at
Chinese Academy of Sciences (CAS)	Specimen collection; DNA barcode data production; Basic research of DNA barcoding; Construction of information centre and central database of China National Committee for iBOL Project.
Ministry of Science and Technology (MOST)	Studies on animal, plant and fungal barcoding; Construction of DNA barcode databases and information systems
Natural Science Foundations of China (NSFC)	Basic research on DNA barcoding theories and methodologies for animal, plant and fungal barcoding

Table 3. DNA barcode of life projects in China

In terms of the construction of China DNA Barcoding Information Centre, we are now running projects from CAS and responsible for the initiation and implementation. With the help from the research campaigns of China and iBOL, we have designed the architecture of the Chinese DNA Barcode of Life Information Systems. Briefly, the entire systems consist of two essential components, the Mirror System of BOLD and the Management System for China DNA barcoding projects (Fig. 1). Each component is an independent system based on the data it contains, and serves as separated services. Nevertheless, the China DNA Barcode of Life Information Centre maintains a DNA barcode database that integrates the data of BOLD mirror and the Chinese DNA barcode data. In views of functions, the mirror system will mainly focus on data synchronization, data presentation, and statistical and analytical tools, while the data management system focuses on data submission, data verification, and data publishing. In sections below, we will describe these two kinds of systems in details.

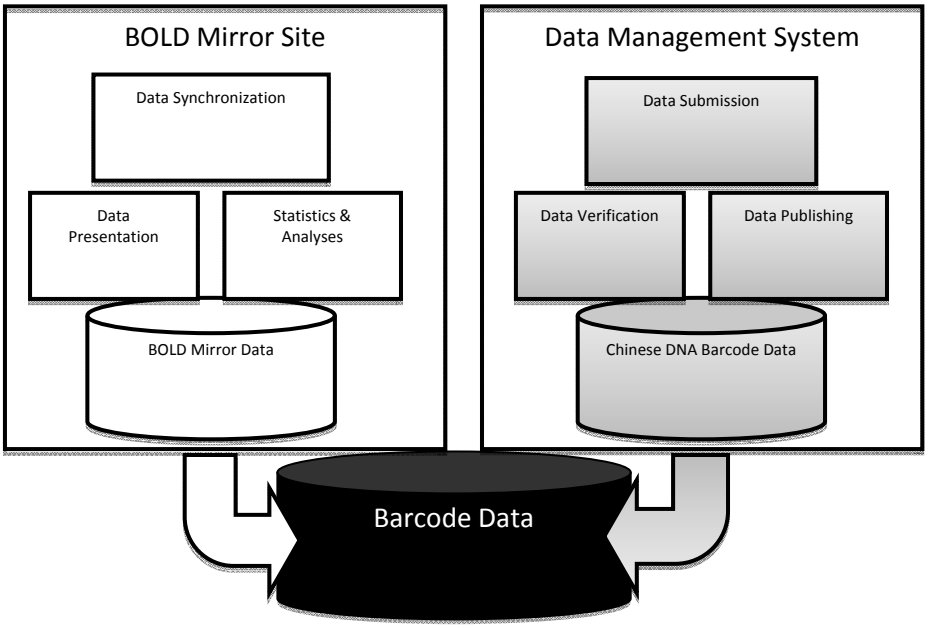

Fig. 1. The architecture of the Chinese DNA Barcode of Life Information Systems. The left square represents the functions of the Mirror System of BOLD, and the right square shows the functions of Data Management System. The DNA barcode data of both systems are further integrated into a centralized DNA barcoding database.

Fig. 2. Homepage of Barcode of Life Data System (BOLD).

3. DNA barcode data schema and the Barcode of Life Data systems

The Barcode of Life Data System (BOLD) is by far the best information system for DNA barcode data, aiming at the collection, management, analysis and use of DNA barcodes (Ratnasingham and Hebert 2007). Started from 2004, BOLD has developed to not only the authorized DNA barcode data resource, but also a global workbench for the assembly, analysis and publication of DNA barcode records. By now, BOLD has become a major contributor to INSDC, with the contribution of about 500,000 records into GenBank during 2010. Millions of DNA barcode records will be deposited into INSDC along with the proceedings of iBOL project. Each record submitted into GenBank has the keyword "BARCODE" and an ID back to BOLD.

Since BOLD uses different ways to store and present DNA barcoding data to GenBank, in the following sections we will give a brief dissection of BOLD, from data schema to the functions. To some extent, this will also help to understand another information systems to be introduced in this chapter.

3.1 Data schema of barcode of life data

In order to collect and manage the DNA barcode data effectively, a data schema is required. BOLD has built up its data schema according the Darwin Core 1 standard, which is applied by the Global Biodiversity Information Facility (GBIF) and other biodiversity alliances, for the data fields related to specimen description. Meanwhile, the data schema describes the format for the sequence (barcode or marker) information as well as the primers and trace files. In brief, there are three categories of information, specimen related, sequence related and primer related (summarized in Table 4, and example in Fig. 3). Specimen related information includes the voucher info, collection info, taxonomic info and details, as well as some multi-media files (mostly photos in the current stage). Sequence related information consists of sequence file (in FASTA format), primer codes and trace files. Considered that primer pairs to the markers (DNA barcodes) are to be standardized and the dataset is relatively small and constant, the detailed info of the primer is separated from specimen and sequence.

As far as a DNA barcode record is concerned, the *Sample ID* is one of the most important key fields. It is the identifier of the sample of specimen used for sequencing, so that it is unique in the whole system. Different *Sample IDs* may refer to one specimen. For example, two legs of butterfly are treated separated for experiments. Since there is not a single field like "Specimen ID" to uniquely mark specimens, BOLD schema uses *Field ID*, *Museum ID* and *Collection Code* instead. At least one of the *Field ID* and *Museum ID* must be appeared and must associate with *Institution Storing* to exclusively locate the specimen in the world. A *Collection Code* is required whenever it has to be combined with *Museum ID* to discriminate specimen. That is the basis to link a DNA barcode to a real specimen. The taxonomic info is for the link between DNA barcode and taxon assigned. Considered that some samples are difficult to identify (e.g. damaged organisms, immature specimen), a full taxonomic assignment is not mandatory, but the phylum level assignment is a prerequisite. Collection info is essential for knowing the global distribution of a specific taxon and the variations among different areas, so that detailed geographical information is encouraged to provide. Additionally, the Details describe the specimen in detail and the Images show the morphological natures. Another key field is *Process ID*, which is used to identify the experimental process that produces a DNA barcode. One *Sample ID* is uniquely referred to one *Process ID*, vice versa. This ensures the connection between sample and produced DNA. *Process ID* is also known as *Barcode ID* in the sequence record view. Finally, trace files are

essential to qualify the results of DNA sequencing, and primer info is required when a process needs to be repeated.

1st Category	2nd category	Data fields	Description
Specimen	Voucher info	Sample ID	ID associated with the sample being sequenced. It is unique.
		Field ID	Field number from a collection event or specimen identifier from a private collection.
		Museum ID	Catalog number in curated collection for a vouchered specimen.
		Collection Code	Code associated with given collection.
		Institution Storing	Full name of the institution where specimen is vouchered.
		Sample Donor	Full name of individual responsible for providing specimen or tissue sample.
		Donor Email	E-mail of the sample donor.
	Taxonomic info	Taxonomy	Full taxonomy. Phylum is mandatory.
		Identifier	Primary individual responsible for the taxonomic identification
		Identifier Email	Email address of the primary identifier
		Identifier Institution	Institution of the primary identifier
	Collection Info	Collectors	List of collectors
		Collection Date	Data of collection
		Continent/Ocean	Continent or ocean name
		Country	Country name
		State/Province	State and/or province
		Region & Sector & Exact site	Detailed description of place
		GPS	GPS coordinates
		Elevation/Depth	Elevation or depth
	Details	Sex	Male/female/hermaphrodite
		Reproduction	Sexual/asexual/cyclic pathogen
		Life Stage	Adult/immature
		Extra Info & Notes	User specified, free text
	Images	Image File	Name of image
		Original Specimen	If the image is from the original specimen

1st Category	2nd category	Data fields	Description
		View Metadata	Dorsal/Lateral/Ventral/Frontal/etc.
		Caption	Short description of image
		Measurement	Measurement that was taken
		Measurement Type	Body length, wing span, etc.
		Copyright	The copyright
Sequence	Sequence	Process ID (Barcode ID)	The ID of a process that produce a sequence
		Sequence	DNA sequence
	Trace Files	Trace File	Complete name of trace file
		Score File	Complete name of score file
		Read Direction	Forward or reverse
		Marker	COI-5P, ITS, rbcLa, matK, etc.
		PCR Primer Codes	PCR primers used
		Sequence Primer Codes	Sequence primers used
Primers		Primer Code	Unique code for a primer
		Primer Description	A description of what the primer is used for
		Alias Codes	Any other known codes
		Target Marker	COI-5P, ITS, etc.
		Cocktail Primer	If it is a cocktail primer
		Primer Sequence	Sequences
		Direction	The direction of the sequence
		Reference/Citation	References and/or citations
		Notes	Some notes

Table 4. Summary of the main data fields in BOLD data schema.

3.2 Barcode of Life Data system

BOLD system consists of three main modules, the Management and Analysis System (MAS), Identification System (IDS)/identification engine, and External Connectivity System (ECS). MAS is responsible for data repository, data management, data uploads, downloads and searches and some integrated analytics (Ratnasingham and Hebert 2007). With no doubt, it comprises the most important functions. According to the data schema described above, it stores specimen related information, sequences, trace files and images. All data of records was uploaded and organized by project that is created by the user. Once a user creates a project for a set of DNA barcode records, at least two data fields, *Sample ID* and *Phylum*, for each record are to be filled. Then additional information including voucher data, collection info, taxonomic assignment, identifier of the specimen, >500-bp sequence of DNA barcode,

PCR primers, and trace files is needed for a full data record. Among all DNA barcode records stored in BOLD, not all are complete and in high quality. For example, there are sequences with more than 1% Ns or less then 500-bp. Hence the integrated analytic tools are useful to help find out those records with low quality. In brief, BOLD employs Hidden Markov Model (Eddy 1998) (on amino acids) to align sequences and then to verify if the correct gene sequence was uploaded; consequently, scripts are used to check for stop codon and to compare against possible contaminant sequence. For trace file, a mean Phred score (Ewing and Green 1998) for the full sequence is determined, and this is used for the quality categorization. After these processing, a record will be flagged if has missing fields, sequence error or low quality.

```
<record>
<recordID>1224108</recordID>
<processid>ASANR583-09</processid>
<specimen_identifiers>
  <sampleid>CASENT0042697-D01</sampleid>
  <catalognum>CASENT0042697-D01</catalognum>
  <fieldnum>BLF09080</fieldnum>
  <institution_storing>California Academy of Sciences</institution_storing>
</specimen_identifiers>
<taxonomy>
  <identification_provided_by>Brian Fisher</identification_provided_by>
  <phylum>
    <taxon>
      <taxID>20</taxID>
      <name>Arthropoda</name>
    </taxon>
  </phylum>
  <class>
    <taxon>
      <taxID>82</taxID>
      <name>Insecta</name>
    </taxon>
  </class>
  <order>
    <taxon>
      <taxID>125</taxID>
      <name>Hymenoptera</name>
    </taxon>
  </order>
</taxonomy>
<collection_event>
  <collectors>B.L.Fisher</collectors>
  <collectiondate>2003-11-18</collectiondate>
  <coordinates>
    <lat>-14.443</lat>
    <lon>49.743</lon>
    <coordsource></coordsource>
    <accuracy></accuracy>
  </coordinates>
  <region>Malagasy</region>
  <exactsite>Parc National de Marojejy, Antranohofa, 26.6 km 31deg NNE Andapa, 10.7 km 318deg NW Manantenina</exactsite>
  <country>Madagascar</country>
  <province>Antsiranana</province>
</collection_event>
<sequences>
  <sequence>
    <sequenceID>3378849</sequenceID>
    <markercode>COI-5P</markercode>
    <genbank_accession>GU711306</genbank_accession>
    <nucleotides>----------------------------------------------------------------------------------------
    ------ATTCACTAATTAATAATGACCAAATTTATAACTCTCTAATTACTAGGCACGCCTTAATTATAATTTTTTTTATAATTATACCTTTTA
    TAATTGGAGGATTTGGAAATTTCCTAGTCCCACTAATACTAGGGGCCCCTGATATAGCCTACCCTCGTATAAATAACATAAGATTCTGACTAT
    TGCCCCCTTCCCTAATACTTTTAATTAGAGGAAGATTTATTAGAGATGGAGTAGGAACAGGATGAACCATCTATCCCCCCCTTTCATCAAATA
    TTTTCCATAACGGCCCTTCTGTAGACCTTTCAATTTTCTCACTTCATATCGCAGGAATATCTTCTATTTTAGGAGCTATTAACTTTATTTCAA
    CTATTATTAATATAAAAAACTCTGGCCTATCATTAGACAAAATTTCATTACTAATCTGATCAATCAACATCACCGCTATTCTCTTACTTCTCT
    CCTTACCAGTCTTAGCCGGAGCAATTACTATATTATTTACGGATCGTAATTTAAACACTTCTTTTTTTGACCCATCAGGAGGGGGAGATCCTA
    TTTTATTCCAACATTTATTT</nucleotides>
    <last_updated>2011-04-10T13:04:27Z</last_updated>
  </sequence>
</sequences>
<last_updated>2011-04-10T13:04:27Z</last_updated>
<notes></notes>
</record>
```

Fig. 3. An example of an XML format DNA barcode records according to BOLD data schema.

The IDS is one of the most commonly used analytic tools of BOLD. It uses all sequences uploaded, both public and private ones, to locate the closest match. Note that the details of the private ones are not exposed. As for animal identification, COI gene set is used as database to compare against. The Basic Local Alignment Search Tool (BLAST) (Altschul et al. 1990) is employed to detect single base indels, and Hidden Markov Model for COI protein is used for sequence alignment. There are four databases are used for COI identification in BOLD, including All Barcode Records Database, Species Level Barcode Database, Public Record Barcode Database and Full Length Record Database. They are comprised of different quality levels of sequences (http://boldsystems.org/docs/handbook.php?page=idengine). Fungal identification is based on ITS, and plant identification is on *rbcL* and *matK*. The Fungal Database and Plant Database respectively are for the identification and only BLAST algorithm is employed. By now, the data records within fungal and plant databases are much fewer than those in COI databases.

Besides IDS, there are other useful tools developed and integrated in BOLD. The Barcode Index Number system (BINs) is designed as an alternate method for species identification. In BINs, a set of operational taxonomic units (OTUs; putative species) was generated using a novel clustering algorithm based on graph methods, and a BIN ID is given for each OTU. BINs and OTUs help to solve the problem that many BOLD records have only interim species name or without fully taxonomic assignment. Another tool, the Taxon ID Tree employs varied distance metrics to build neighbour-joining tree with at most 5000 species a time. This is powerful toolbox for online phylogenetics analysis. More functions including distance summary, sequence composition, nearest neighbour summary, DNA degradation test, accumulation curve and alignment viewer are available in BOLD. These tools implemented the bioinformatics and statistic approaches for data analyses.

ECS is served as the interface for other developers to access the barcode data via web services. Currently, BOLD opens two services e-Search and e-Fetch following the Representational State Transfer (REST) architecture. Programmes may use e-Search to get a list of records, and use e-Fetch to obtain the details. Below lists the parameters for e-Search and e-Fetch. Anther web service called eTrace is also developed for the retrieval of trace files for a given Sample ID. We have tested for the use of this service, and obtained thousands of trace files from BOLD. This service will be exposed to public in the near future.

Service Name	Parameters	Description
e-Search & e-Fetch	id_type	Sample_id, process_id, specimen_id, sequence_id, tax_id, record_id
	ids	Comma separated ids
	Return_type	Text, xml, json
	File_type	Zip
e-Search	Geo_inc	Country/province to be included
	Geo_exc	Country/province to be excluded
	Taxon_inc	Taxonomy to be included
	Taxon_exc	Taxonomy to be excluded
e-Fetch	Record_type	Specimen, sequence, full

Table 5. Parameters of the web services of BOLD.

4. The Global Mirror System of DNA Barcode Data (GMS-DBD)

One of the main tasks of iBOL project is to setup global mirror sites for DNA barcode data, and this is assigned as the mission of working group 3.2 (that is chaired by the author Juncai Ma). Mirror sites play roles not only for data security but also for the global access and use of DNA barcode data fast and stably. In addition to iBOL's task, the Chinese DNA Barcode of Life Information Systems requires to mirror BOLD as well. For the current stage, the entire BOLD system is difficult to mirror, in both the storage and the analytical workbench. For this reason, we developed the mirror site of BOLD data (http://www.boldmirror.net) (Fig. 4) in China in 2010, and served it as one of the major components of the Chinese DNA Barcode of Life Information Systems. In late 2010, we started to encapsulate the mirror site into a distributable mirror system, namely the Global Mirror System of DNA Barcode Data (GMS-DBD). Different from BOLD systems, GMS-DBD is designed and currently served as a system for the presentation and analysis of DNA barcode data only, but not the management of DNA barcode projects. Moreover, GMS-DBD is designed to feature in the fast deployment and fast use of the DNA barcode data.

Fig. 4. Homepage of the mirror of BOLD data.

4.1 Design and implementation of GMS-DBD

For the purposes mentioned above, GMS-DBD was designed with three main components including data synchronization module, data presentation module, and statistics and analysis module. As for data synchronization, the module functions to obtain data from BOLD and then to distribute them to each mirror site deployed by GMS-DBD. Right after data transferred over Internet, all updated data are imported automatically into mirror site's local database management system (DBMS). Data presentation module comprises the following functions, browsing by records, browsing by taxonomy and searching by keywords. These are the key functions for a DNA barcode data mirror. Data statistics module aims at the statistical presentation of the entire dataset, including the barcode data statistics by country, by taxon or by organization. (Fig. 5 Left) For the use of DNA barcode data, especially for sequence similarity based identification, we embedded a form to submit the sequences to be identified to a BLAST server in China mirror site.

The whole system was implemented as a software package, in which the database, applications, web pages were encapsulated. This software package was for the Linux platform, and tested on some main distributes, like Fedora, Ubuntu. The installer was written in Perl, and will guide the administrator to setup and configure the mirror site step-by-step. Particularly, Apache web server and PHP scripting language were employed for the Web layer, and MySQL DBMS was used for the management of database. The applications for database search, records presenting and sorting, and statistics calculation were written using PHP scripts. For better visualization, Java applet (viewing chromatogram files), Adobe FLEX (statistical presentation) and Google Maps (geographical view of locations) were employed and embedded into the web pages. In addition to presenting record by specimen information and by sequence information as BOLD does, we developed a full record view to browse all information in a single page (Fig. 5 right).

In additional to the installer, data presentation module, and statistics and analysis module, data synchronization module was developed separately. From late 2009, BOLD and our centre were beginning to test the transfer of DNA barcode data between Canada and China, and finally defined a method. Every day, BOLD dumps the daily-updated data and daily-completed data into XML format files and put them on a HTTP server, and for the mirror site, we run a daemon process to download the latest files. After data transfer finished, the daemon process will invoke consequently another processes to parse the XML format files and to import the parsed files into MySQL database. Perl is used for the parser and Structured Query Language (SQL) is used for data import. Specific for the data import process, it will execute the insertion and updating of the new entries and log every modification of the whole dataset. In parallel, another process will run to extract the DNA sequences of the new records and then to index them for BLAST search. All these procedures are scheduled and run automatically, and the updates and logs will be shown on the web pages immediately after the procedures finished.

4.2 Distribution of GMS-DBD and DNA Barcode data

GMS-DBD is freely distributed as the DNA barcode data. Nowadays, the University of Waikato, New Zealand has firstly built up their mirror site (http://nz.boldmirror.net) using the GMS-DBD distributes, and gave us great suggestions on the improvement and the further development of GMS-DBD. The Centro de Excelencia em Bioinformatica (CEBio) of Brazil has also contacted us and is setting up the mirror site using GMS-DBD.

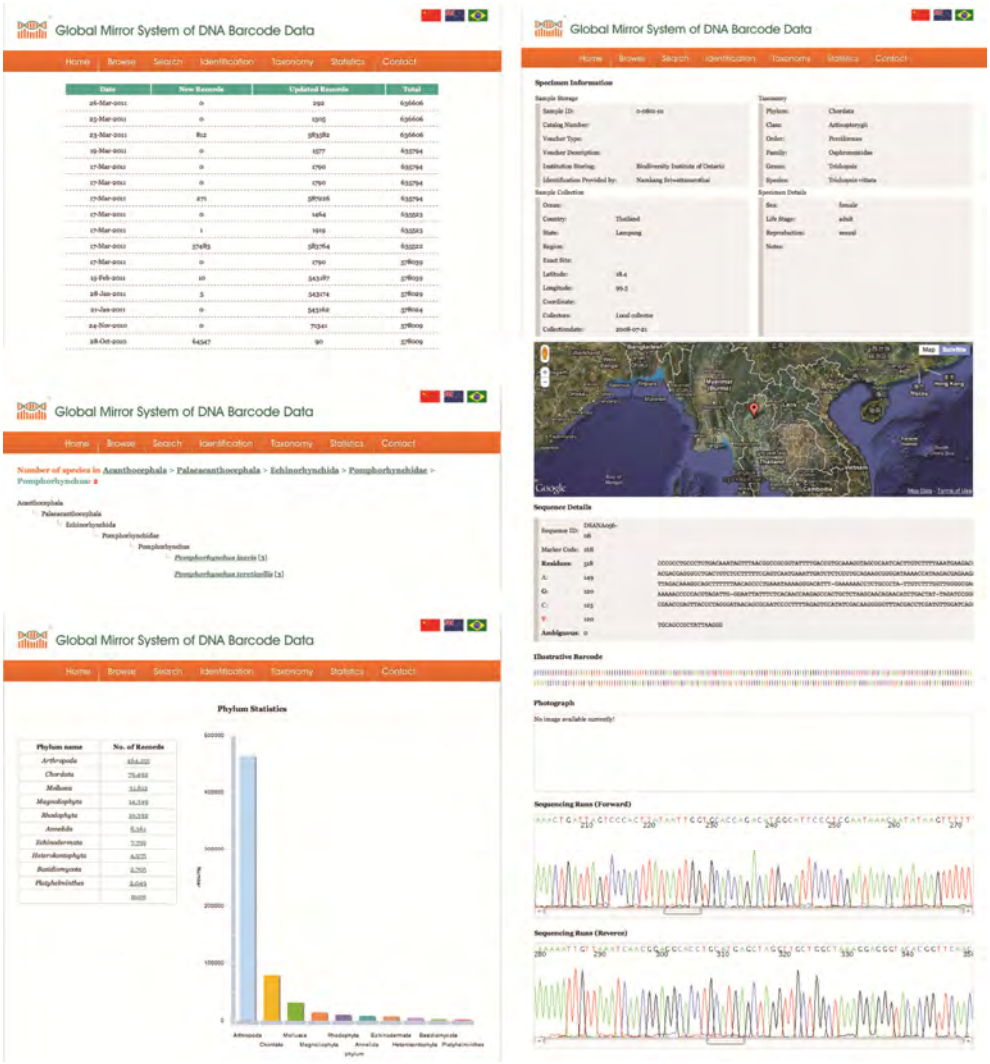

Fig. 5. Functions of GMS-DBD. Top-left, logs of data updating. Middle-left, view by taxonomic tree. Bottom-left, statistics by phylum. Right, full DNA barcode record.

To date, there are ~636,000 DNA barcode records available for the mirror sites, though more than 1.2 million records deposit in BOLD. One reason is that mirror sites stored the public available data records only, while BOLD has some private data to be released in the future. The second is that there are incomplete records held in BOLD that did not distribute to the mirror sites. Among all the records within the mirror sites, more than 200 thousand of them have trace files. The photos for each specimen are not available for mirror sites by now, because there are copyright problems to be solved.

5. Design and implementation of the DNA barcode management system in China

As described in the previous sections, China has launched several projects to contribute the construction of the global DNA barcode library. An information system is thus needed to collect, manage and publish the produced data. Although BOLD is a good choice for the management of users' DNA barcode projects and data records, many scientists are still willing to hold their data before their results published in scientific journals. Therefore, we designed and developed a DNA barcode management system for the Chinese scientists to manage their barcode data. First, the system is also designed as a distributable version, and could be downloaded and installed locally. Second, the user can hold the data for privacy for long time. Third, it supports modifications for the data schema. Fourth, for its simplicity, it lacks the connectivity to any LIMS, but using a unified data format to exchange data from LIMS.

In views of function, this system has similar aspects to the BOLD system, i.e. user identification system, project based management, data input forms for different data types, data review and analyses platform. The entire system is developed with the LAMP (Linux+Apache+MySQL+PHP) architecture, and of no need to mount on very heavy computing infrastructures. The system manager has the privileges to choose the hardware ands scale the capability of the system. For data's safety, only the registered users are allowed to use. The registered users have their data records organized by project, and only their own projects or authorized ones are allowed to visit (Fig. 6 left). The data input forms are like BOLD's as well, in that there are forms for specimen info, taxonomy info, sequences, photo, trace files, etc., except that all the labels have Chinese translations.

Fig. 6. Snapshots of DNA barcode management system in China. Left, user's project page, summarizing the projects and records. Right, Quick Upload Page.

BOLD has exemplified the management of DNA barcode data as a centralized data centre and shall be the reference for the development of core functionality of the Central Nodes. However, the development of such a comprehensive system needs a long period, and inapplicable for the immediate use. Moreover, things are to some extent different in China than in Canada. First, there is no constructed barcoding centre like CCDB before varied DNA barcoding projects were issued. Second, the research campaigns studying animal, plant and fungus barcoding respectively have already defined the working structure for

each studies, especially for the specimen and collection information. Additionally, the DNA sequences produced by each campaign are either stored in their LIMS or simply stored in personal computers. In this situation, the tough work is becoming how to make the DNA barcode data management system suitable for every data structure, and how to easily collect those data already there. To meet this need, we designed different data schema for animal, plant and fungus. Note that all data schema are following the latest BOLD data schema, but with some data fields modified to fit each species groups. For example, we omitted the data fields "sex", "reproduction" and "life stage" for fungi, but added on data fields "habitat" and "storing method". This was the result after discussion with scientists doing fungus research. Additionally, we also added on some fields for Chinese, like "Chinese common name", "taxon name in Chinese", etc. Moreover, we encouraged the users to use the Chinese characters for "provider name" and "location names", as they might be ambiguous in English.

Another feature of this data management system is that we implemented a gateway for quick upload (Fig. 6 right). In terms of the data already produced, they are stored and organized either by a local DBMS or by Excel datasheet with files. The providers would like to upload them into the system in batch mode, but not form by form. Then we developed the quick upload gateway, according to the data schema and created different templates for batch upload. In brief, the template is in Excel format, and every data field is in one column while every record is in one row. What the user needs to do first is to fill in the template for their data or slightly modify the datasheet in use. Then three files are required to prepare, one is zipped photos, another is zipped trace files, and the other is FASTA format sequence file. Note that the file names of photos and trace files, and the sequence name (*Process ID*) of the sequence should be in the right place of the Excel template. Then these four files may be uploaded onto the system and records are imported into the database in batch mode. This mode is also suitable for those who are in charge of the whole procedures of the production of DNA barcode data, from specimen to barcode.

Functions of the GMS-DBD were applied for this management system, in that we integrated Google Maps for geographical presentation, Java applet for viewing trace files, and BLAST for sequence identification. Besides, the management system has a data publishing function, which can generate XML format data following BOLD data schema. Within data transformation (from MySQL to XML), a language translation module will automatically invoked to translate those Chinese "city names", "institution names" and "person names" into English.

By now, this DNA barcode data management system is implemented using Chinese language, and has been developing for multi-language uses. To date, it is used for the collection and management of DNA barcodes of fungi, fish, birds, amphibian, and plants in China.

6. Perspectives of DNA barcode information system and the underneath bioinformatics

Along with the success of DNA barcode of life projects worldwide, huge amount of data will be produced on purpose. The traditional sequence database like GenBank seems not suitable for the storage of the DNA barcode records with multiple data types, so that novel systems are required to develop for the management and utilization of those data. Besides

the information systems described above, DNA Data Bank of Japan (DDBJ) (Kaminuma et al. 2011) maintains the Japanese repository for barcode data and employed the its BLAST server for identification, and the Korean BioInformation Center of Korea Research Institute of Bioscience and Biotechnology (KRIBB), Korea developed the BioBarcode platform for Asian biodiversity resources (Lim et al. 2009). The CBS culture collection of Netherlands is attempting to integrate the DNA barcode data into the BioloMICS software, which bundles comprehensive bioinformatics tools for data analyses. This will provide better experiences in some aspects for the bioinformatics researches applying barcode data, though the online workbench of BOLD provides sets of approaches for data analysis.

When bioinformatics is mentioned, the algorithms and software are first recalled. The commonly used method for species discrimination is the Neighbour Joining (NJ) algorithm with Kimura 2 parameters (K2P) corrections. Though this approach was claimed as the best DNA substitution model for close genetic relations (Nei and Kuman 2000), the maximum likelihood methods and Bayesian Inference are more and more used for DNA barcoding analysis (e.g. Mueller 2006, deWaard et al. 2010, Kim et al. 2010). Coalescent-based methods for phylogenetics were also examined for DNA barcoding (Nielsen and Matz 2006, Abdo and Golding 2007). Casiraghi *et al.* (Casiraghi et al. 2010) summarized bioinformatics approaches for the analyses of barcode data and proposed the use of varied methods for different scenarios.

Although DNA barcoding technology has been largely improved and the DNA barcode data has rapidly accumulated, there are still some concerns on the use of DNA barcode. First, varied paired of primers might be used to identify an unknown sample or a mixture. Although COI gene was proved to be efficient in almost all animals and some groups of fungi, the identification of unknown specimen may need several paired of primers and this will increase the complexity of the automation of DNA barcoding process. Prospectively, this would be one of the major issues on the development and implementation of the handy device for DNA barcoding, as planned in iBOL WG4.2. Secondly, every DNA barcode has the limitation of resolution in specific species groups, so that auxiliary markers need to be discovered. This problem exists mainly in plant and fungi, though some animal groups have met same one. Currently, plant and fungi data deposited in BOLD are still limited, and a lot of groups need to be examined. Additionally, the approaches for sharing and using barcode data need to be improved. A user-friendlier interface to access the barcode dataset is needed. For instance, the well-examined DNA barcodes and/or the consensus of barcodes of each taxon are organized and prepared, and user needs only to select the interested ones and download only a small dataset. This will be convenient for the users (i.e. identification of specimen using DNA barcoding) and some researchers on barcoding. As a matter of fact, that is one of the tasks what BOLD and our group are working on.

7. Conclusions

With no doubt, DNA barcoding is becoming a popular approach for knowing the biodiversity on the earth, by utilizing the accumulative knowledge of taxonomy, the modern techniques of molecular biology and bioinformatics. Bioinformatics played prominent role in the construction and the employment of the global barcode of life library, from the management of data to the development of novel methods.

8. Acknowledgments

The authors thank the team members in the Network Information Center, Institute of Microbiology, Chinese Academy of Sciences. The work in our centre is supported, in part, by the projects from the Bureau of Life Sciences and Biotechnology of CAS, the Project of Informatization of CAS, and the Project of Scientific Databases of CAS (http://www.csdb.cn). Our centre is also obtained support from the State Key Laboratory of Microbial Resources (SKLMR), Institute of Microbiology, CAS. DL would also like to appreciate the support from Natural Science Foundations of China (NSFC) (Grant no. 30800640). The authors appreciate the supports and advices from iBOL, CCDB and BOLD.

9. References

Abdo, Z. and G. B. Golding. (2007). A step toward barcoding life: a model-based, decision-theoretic method to assign genes to preexisting species groups. *Systematic biology,* 56:44-56.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215:403-410.

Barber, P. and S. L. Boyce. (2006). Estimating diversity of Indo-Pacific coral reef stomatopods through DNA barcoding of stomatopod larvae. *Proceedings of the Royal Society of London. Series B, Biological sciences*, 273:2053-2061.

Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. (2011). GenBank. *Nucleic acids research*, 39:D32-37.

Casiraghi, M., M. Labra, E. Ferri, A. Galimberti, and F. De Mattia. (2010). DNA barcoding: a six-question tour to improve users' awareness about the method. *Briefings in bioinformatics*, 11:440-453.

CBOL Plant Working Group. (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences of the United States of America*, 106:12794-12797.

de Groot, G. A., H. J. During, J. W. Maas, H. Schneider, J. C. Vogel, and R. H. Erkens. (2011). Use of rbcL and trnL-F as a two-locus DNA barcode for identification of NW-European ferns: an ecological perspective. *PLoS one*, 6:e16371.

deWaard, J. R., A. Mitchell, M. A. Keena, D. Gopurenko, L. M. Boykin, K. F. Armstrong, M. G. Pogue, J. Lima, R. Floyd, R. H. Hanner, and L. M. Humble. (2010). Towards a global barcode library for Lymantria (Lepidoptera: Lymantriinae) tussock moths of biosecurity concern. *PLoS one*, 5:e14280.

Ebach, M. C. and C. Holdrege. (2005). DNA barcoding is no substitute for taxonomy. *Nature,* 434:697.

Eddy, S. R. 1998. Profile hidden Markov models. *Bioinformatics,* 14:755-763.

Evans, K. M., A. H. Wortley, and D. G. Mann. (2007). An assessment of potential diatom "barcode" genes (cox1, rbcL, 18S and ITS rDNA) and their effectiveness in determining relationships in Sellaphora (Bacillariophyta). *Protist,* 158:349-364.

Ewing, B. and P. Green. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome research*, 8:186-194.

Fazekas, A. J., K. S. Burgess, P. R. Kesanakurti, S. W. Graham, S. G. Newmaster, B. C. Husband, D. M. Percy, M. Hajibabaei, and S. C. Barrett. (2008). Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. *PLoS one*, 3:e2802.

Frezal, L. and R. Leblois. (2008). Four years of DNA barcoding: current advances and prospects. *Infection, genetics and evolution,* 8:727-736.

Greenstone, M. H., D. L. Rowley, U. Heimbach, J. G. Lundgren, R. S. Pfannenstiel, and S. A. Rehner. (2005). Barcoding generalist predators by polymerase chain reaction: carabids and spiders. *Molecular ecology*, 14:3247-3266.

Gregory, T. R. (2005). DNA barcoding does not compete with taxonomy. *Nature,* 434:1067.

Hajibabaei, M., D. H. Janzen, J. M. Burns, W. Hallwachs, and P. D. Hebert. (2006). DNA barcodes distinguish species of tropical Lepidoptera. *Proceedings of the National Academy of Sciences of the United States of America*, 103:968-971.

Hebert, P. D., A. Cywinska, S. L. Ball, and J. R. deWaard. (2003a). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B, Biological sciences,* 270:313-321.

Hebert, P. D., E. H. Penton, J. M. Burns, D. H. Janzen, and W. Hallwachs. (2004a). Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly Astraptes fulgerator. *Proceedings of the National Academy of Sciences of the United States of America*, 101:14812-14817.

Hebert, P. D., S. Ratnasingham, and J. R. deWaard. (2003b). Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London. Series B, Biological Sciences,* 270 Suppl 1:S96-99.

Hebert, P. D., M. Y. Stoeckle, T. S. Zemlak, and C. M. Francis. (2004b). Identification of Birds through DNA Barcodes. *PLoS biology*, 2:e312.

Jaklitsch, W. M., M. Komon, C. P. Kubicek, and I. S. Druzhinina. (2006). Hypocrea crystalligena sp. nov., a common European species with a white-spored Trichoderma anamorph. *Mycologia,* 98:499-513.

Kaminuma, E., T. Kosuge, Y. Kodama, H. Aono, J. Mashima, T. Gojobori, H. Sugawara, O. Ogasawara, T. Takagi, K. Okubo, and Y. Nakamura. (2011). DDBJ progress report. *Nucleic acids research*, 39:D22-27.

Kerr, K. C., M. Y. Stoeckle, C. J. Dove, L. A. Weigt, C. M. Francis, and P. D. Hebert. (2007). Comprehensive DNA barcode coverage of North American birds. *Molecular ecology notes*, 7:535-543.

Kim, M. I., X. Wan, M. J. Kim, H. C. Jeong, N. H. Ahn, K. G. Kim, Y. S. Han, and I. Kim. (2010). Phylogenetic relationships of true butterflies (Lepidoptera: Papilionoidea) inferred from COI, 16S rRNA and EF-1alpha sequences. *Molecules and cells*, 30:409-425.

Kress, W. J. and D. L. Erickson. (2007). A two-locus global DNA barcode for land plants: the coding rbcL gene complements the non-coding trnH-psbA spacer region. *PLoS one*, 2:e508.

Kumar, N. P., A. R. Rajavel, R. Natarajan, and P. Jambulingam. (2007). DNA barcodes can distinguish species of Indian mosquitoes (Diptera: Culicidae). *Journal of medical entomology*, 44:1-7.

Lim, J., S. Y. Kim, S. Kim, H. S. Eo, C. B. Kim, W. K. Paek, W. Kim, and J. Bhak. (2009). BioBarcode: a general DNA barcoding database and server platform for Asian biodiversity resources. *BMC genomics*, 10 Suppl 3:S8.

Liu, J., M. Moller, L. M. Gao, D. Q. Zhang, and D. Z. Li. (2011). DNA barcoding for the discrimination of Eurasian yews (Taxus L., Taxaceae) and the discovery of cryptic species. *Molecular ecology resources*, 11:89-100.

Maiden, M. C., J. A. Bygraves, E. Feil, G. Morelli, J. E. Russell, R. Urwin, Q. Zhang, J. Zhou, K. Zurth, D. A. Caugant, I. M. Feavers, M. Achtman, and B. G. Spratt. (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences of the United States of America*, 95:3140-3145.

Meier, R., K. Shiyang, G. Vaidya, and P. K. Ng. (2006). DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. *Systematic biology*, 55:715-728.

Meyer, C. P. and G. Paulay. (2005). DNA barcoding: error rates based on comprehensive sampling. *PLoS biology*, 3:e422.

Mower, J. P., P. Touzet, J. S. Gummow, L. F. Delph, and J. D. Palmer. (2007). Extensive variation in synonymous substitution rates in mitochondrial genes of seed plants. *BMC evolutionary biology*, 7:135.

Mueller, R. L. (2006). Evolutionary rates, divergence dates, and the performance of mitochondrial genes in Bayesian phylogenetic analysis. *Systematic biology*, 55:289-300.

Nei, M. and S. Kuman. (2000). *Molecular Evolution and Phylogenetics*. Oxford University Press, New York.

Nielsen, R. and M. Matz. 2006. Statistical approaches for DNA barcoding. *Systematic biology*, 55:162-169.

Nilsson, R. H., M. Ryberg, E. Kristiansson, K. Abarenkov, K. H. Larsson, and U. Koljalg. 2006. Taxonomic reliability of DNA sequences in public sequence databases: a fungal perspective. *PLoS one,* 1:e59.

Pfenninger, M., C. Nowak, C. Kley, D. Steinke, and B. Streit. (2007). Utility of DNA taxonomy and barcoding for the inference of larval community structure in morphologically cryptic Chironomus (Diptera) species. *Molecular ecology,* 16:1957-1968.

Piredda, R., M. C. Simeone, M. Attimonelli, R. Bellarosa, and B. Schirone. (2011). Prospects of barcoding the Italian wild dendroflora: oaks reveal severe limitations to tracking species identity. *Molecular ecology resources,* 11:72-83.

Ran, J. H., P. P. Wang, H. J. Zhao, and X. Q. Wang. (2010). A test of seven candidate barcode regions from the plastome in Picea (Pinaceae). *Journal of integrative plant biology*, 52:1109-1126.

Ratnasingham, S. and P. D. Hebert. (2007). BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). *Molecular ecology notes*, 7:355-364.

Rusch, D. B., A. L. Halpern, G. Sutton, K. B. Heidelberg, S. Williamson, S. Yooseph, D. Wu, J. A. Eisen, J. M. Hoffman, K. Remington, K. Beeson, B. Tran, H. Smith, H. Baden-Tillson, C. Stewart, J. Thorpe, J. Freeman, C. Andrews-Pfannkoch, J. E. Venter, K. Li, S. Kravitz, J. F. Heidelberg, T. Utterback, Y. H. Rogers, L. I. Falcon, V. Souza, G. Bonilla-Rosso, L. E. Eguiarte, D. M. Karl, S. Sathyendranath, T. Platt, E. Bermingham, V. Gallardo, G. Tamayo-Castillo, M. R. Ferrari, R. L. Strausberg, K. Nealson, R. Friedman, M. Frazier, and J. C. Venter. (2007). The Sorcerer II Global

Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS biology*, 5:e77.

Schindel, D. E. and S. E. Miller. (2005). DNA barcoding a useful tool for taxonomists. *Nature*, 435:17.

Seifert, K. A., R. A. Samson, J. R. Dewaard, J. Houbraken, C. A. Levesque, J. M. Moncalvo, G. Louis-Seize, and P. D. Hebert. (2007). Prospects for fungus identification using CO1 DNA barcodes, with Penicillium as a test case. *Proceedings of the National Academy of Sciences of the United States of America*, 104:3901-3906.

Smith, M. A., B. L. Fisher, and P. D. Hebert. (2005). DNA barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: the ants of Madagascar. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 360:1825-1834.

Stahls, G. and E. Savolainen. (2008). MtDNA COI barcodes reveal cryptic diversity in the Baetis vernus group (Ephemeroptera, Baetidae). Molecular phylogenetics and evolution 46:82-87.

Teletchea, F. (2010). After 7 years and 1000 citations: comparative assessment of the DNA barcoding and the DNA taxonomy proposals for taxonomists and non-taxonomists. Mitochondrial DNA 21:206-226.

Vences, M., M. Thomas, R. M. Bonett, and D. R. Vieites. (2005). Deciphering amphibian diversity through DNA barcoding: chances and challenges. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 360:1859-1868.

Venter, J. C., K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y. H. Rogers, and H. O. Smith. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, 304:66-74.

Ward, R. D., T. S. Zemlak, B. H. Innes, P. R. Last, and P. D. Hebert. (2005). DNA barcoding Australia's fish species. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences,* 360:1847-1857.

Wirth, T., D. Falush, R. Lan, F. Colles, P. Mensa, L. H. Wieler, H. Karch, P. R. Reeves, M. C. Maiden, H. Ochman, and M. Achtman. (2006). Sex and virulence in Escherichia coli: an evolutionary perspective. *Molecular microbiology*, 60:1136-1151.

Woese, C. R. (1996). Whither microbiology? Phylogenetic trees. *Current biology*, 6:1060-1063.

Yesson, C., R. T. Barcenas, H. M. Hernandez, M. De La Luz Ruiz-Maqueda, A. Prado, V. M. Rodriguez, and J. A. Hawkins. (2011). DNA barcodes for Mexican Cactaceae, plants under pressure from wild collecting. *Molecular ecology resources*, in publishing.

Zhou, J., M. E. Davey, J. B. Figueras, E. Rivkina, D. Gilichinsky, and J. M. Tiedje. (1997). Phylogenetic diversity of a bacterial community determined from Siberian tundra soil DNA. *Microbiology,* 143 (Pt 12):3913-3919.

Zhou, X., S. J. Adamowicz, L. M. Jacobus, R. E. Dewalt, and P. D. Hebert. (2009). Towards a comprehensive barcode library for arctic life - Ephemeroptera, Plecoptera, and Trichoptera of Churchill, Manitoba, Canada. *Frontiers in zoology*, 6:30.