iBOL Data & Resource Sharing Policies

Introduction

The International Barcode of Life Project will assemble a reference library composed of short, standardized gene sequence profiles (DNA barcodes) to enable the molecular identification of known species and facilitate the discovery of new ones. The Ft. Lauderdale Conference[1] on sharing data from large-scale biological research projects defined a community resource project as "a research project specifically devised and implemented to create a set of data, reagents or other material whose primary utility will be as a resource for the broad scientific community." Members of the International Barcode of Life Project (iBOL) are constructing such a community resource and in accordance with the guidelines established at Ft. Lauderdale, iBOL is committed to rapid data release and sharing. This position is typical of most large-scale collaborations in genetics (e.g. the International HapMap Project) and mirrors the data release policies of organizations such as the National Human Genome Research Institute, Genome Canada, the Gordon and Betty Moore Foundation in the USA, and The Wellcome Trust in the UK.

The iBOL data release and resource sharing policy (as outlined below) seeks to accelerate the timely development of products that will benefit humankind by providing rapid access to the primary outputs from iBOL: raw DNA sequences associated with taxonomic assignments. This goal mirrors recommendations for minimum data release associated with large-scale genomics projects (Field *et al.* 2008)[2]. Publication of more detailed sequence annotations and analyses (e.g. involving multiple sequence alignments, with expert species-level identifications attributed to each sequence) remains an equally important vehicle for data release and iBOL members are expected to publish these findings in a timely manner. Participation within iBOL requires commitment and adherence to the following polices and guidelines:

Types of Data Generated by iBOL

iBOL will generate various types of data, some of which will be similar or identical to other genomics or molecular biology projects and some of which will be unique. The primary goal of iBOL is the construction of a DNA barcode reference sequence library representing 500K species derived from 5M specimens. Each sequence entry will relate to a documented specimen-collecting event (e.g. date, place, method and agent of specimen collection) that includes a physical voucher specimen held in a reference collection. A voucher specimen, digital image of that specimen and precise geospatial coordinates of the collection locality should accompany each sequence entry except in certain circumstances (e.g. when locality data could compromise the survival of an endangered species). The key annotation element of the barcode library is an authoritative specie level identification for each specimen that has been crosschecked against a list of validated taxonomic names.

It is recognized that such a complete data set requires a concerted and oftentimes-iterative taxonomic identification process. In many cases, this requires multiple expert opinions and involves timelines that extend far beyond those normally considered for genomic data release. To delay data release until such a taxonomic validation is complete will be inconsistent with the iBOL Data Release Policy. To demonstrate project progress and foster collaboration, iBOL members will rapidly release raw sequence data (electropherogram trace files) and provisional (e.g. high level) taxonomic information, as described below. This will facilitate input from the larger research community - recognizing the need for further refinement of taxonomic

annotations and ongoing curation of the sequence database after preliminary release. The primary objective of iBOL is to generate a comprehensive DNA Barcode database that is accessible and relevant to the needs of both the public and the research community. For iBOL to succeed, it is critical that every effort is made to complete the taxonomic annotation and validation process to whatever endpoint is possible, and that complete data are released as soon as is feasible. The following is a description of the data types involved and the pipeline for generating said data.

1) **Specimen submission data**

Once a specimen is collected and submitted for analysis, it will be assigned a unique BOLD Process Identification Number and Sample Identification Number. There will also be biological and geographical data associated with the sample or specimen. The following are the minimum requirements for the initial data release:

- Country/Ocean where the specimen was collected
- Date of collection

Other data may also be submitted with the sample or specimen, but these are not mandatory for the initial data release. They may include:

- A digital image of the specimen (although this is not mandatory, it is highly recommended that an image be included at some point in the data processing pipeline)
- GPS coordinates of collection site (although this is not mandatory it is highly recommended[3] that GPS data be included, except under exceptional circumstances)

2) **Taxon Name/Identifier**

A provisional taxonomic assignment or identification is usually made upon sample submission and is required for early data release. This may be at the family level or may be a descriptive name (e.g., "environmental sample"), and is not intended to be an final identification. An accurate identification of genus and species is the goal, but this may not be achieved for some time, and data release must not be dependent on annotation of the sequence with a definitive taxonomic identification.

3) **Genetic Data**

- Gene region sequenced
- PCR primer sequence and conditions
- Electropherogram "trace" files utilized in the:
- Sequence "contig" assembly (DNA barcode)
- BOLD Barcode Index Number (BIN)

Quality Control/Quality Assurance of Genetic Data

Data must satisfy rigorous Quality Assessment/Quality Control (QA/QC) processes before release. Preliminary data as described above must pass the following QA/QC criteria:

i) Length of finished sequence must be >75% of BARCODE[4] approved marker length (e.g., 500 bp for COI), with an expectation of 2X coverage

ii) Sequence quality must be reasonable (i.e., <1% ambiguous bases in final trimmed contig assembly, with an expectation of average Phred score > 20)

iii) Sequence must not match common contaminants (e.g., human)

iv) Assessment of high-level taxonomic consistency (i.e., the DNA barcode should cluster with related taxa)

Parts i – iii could be automated processes; part iv is critical and may require human intervention

Barcode of Life Data System (BOLD)

Regardless of whether specimen processing and DNA barcoding is completed at the DNA barcoding facility within the Biodiversity Institute of Ontario, or at DNA sequencing facilities in other iBOL member institutions, the iBOL research community will use BOLD and/or, when established, BOLD International Mirror sites as the primary vehicles for assembling and releasing barcode data.

Timeframe for Data Release to Public Databases (GenBank)

Data submitted to iBOL affiliated projects in BOLD will be transferred to GenBank prior to user initiated publication. Data release will follow a two phase process to be performed on a weekly basis.

Phase I will involve the release of all generated sequence data and high level taxonomic information. This early release is intended to liberate enough information to be useful to other researchers and to monitor progress in the growth of barcode records for each iBOL Working Group. It will be performed automatically and involve data that can be released following computerized quality checks and generation of Barcode Index Numbers (BINs). In detail, the following data will be released in Phase I (within one week of the sequence generation):

- Location information: All available information
- Temporal information: date of sample collection
- Taxonomic information: order-level assignment with BIN
- Sequence information: automatically assembly sequence, trace files, primers used, and the centre that carried out sequencing
- Database identifiers: BOLD process ID and specimen identifiers (voucher number, depository, and collection code)

Phase II will involve the release of additional data elements that require manual curatorial efforts and detailed taxonomic enquiry. Phase II will ordinarily occur when manuscripts are submitted for publication. However, some researchers have indicated their intent to support rapid release of all data elements even if the early versions of the release involve substantial errors in taxonomic assignment. However, these will be corrected through an ongoing update process. The following data will be transferred from BOLD to GenBank before or at the point of manuscript submission or publication:

- Location information: GPS coordinates, elevation/depth, province/state, exact site of collection, and individuals credited for collecting the specimen
- Taxonomic information: species-level assignment (and subspecies, if appropriate) and individual credited for the identification
- Sequence information: manually assembled and curated barcode sequence

Resource Types

Just as there are several types of data generated by the iBOL project, there are several resource types that will be shared publicly by the iBOL team.
These include:

1) **Biomaterials** These could include specimens or tissue samples. Members of the iBOL consortium are committed to the regulatory framework established under the Convention

on Biological Diversity. Any transfer of resources between iBOL members will respect all restrictions in relation to biomaterial transfer and is governed by a Materials Transfer Agreement (MTA) or similar agreement which should be signed before transfer takes place. MTAs must include descriptions of terms of access and use by other researchers, storage, and curation. Some countries or institutions may choose to restrict the access and use of their biomaterials, and those terms must be clearly described in the MTA or other such agreements.

2) **The Barcode of Life Data System (BOLD).** DNA barcode records in BOLD are a community resource that will be shared publicly. No charges will be made to users of the publicly available data. The only requirement for using data from BOLD is the need to give proper accreditation to iBOL researchers who generated the data and to BOLD. iBOL strongly encourages use of public data within BOLD for development of applications that will result in technology development, improvements to public health and environmental health monitoring, or any other innovations.

3) Informatics Tools. The informatics tools used in BOLD and other aspects of the iBOL project are considered community resources within iBOL. For example, the Laboratory Information Management system used by the Barcoding facility at the Biodiversity Institute of Ontario is a part of BOLD and can be accessed and shared with all members of iBOL and the greater research community.

Accreditation of Released Data

The members of iBOL recognize that it is very important for all researchers, whether they are academic, government, or industry, graduate students, post-doctoral fellows, or professors, to be acknowledged for the data that has been generated and made available to the wider scientific community. This is important for several reasons, including the fact that publications and citations are globally recognized as performance measures for projects and for researchers. It is also important because researchers or others who wish to make use of those publicly released data benefit from knowing who generated those data, so that they may explore further partnership or collaboration opportunities, or seek further information, or they may have other data that would benefit the original researcher. Thus, it is important that the data that are released as described above, contain information about the submitter, and that the appropriate text are included with those releases to encourage citations and acknowledgments. Users of the data should acknowledge the source.

There are other mechanisms whereby researchers may receive appropriate accreditation for early data release. Two such avenues that are strongly encouraged by iBOL members are: Project Description and Data Release Publications. These serve a multitude of purposes by (a) providing information for accreditation of data submitters, so that those researchers may be cited, (b) provide the iBOL team and the greater research and public community with opportunities to provide input and fresh data that can be used to refine and improve upon preliminary data, and (c) provide opportunities for information exchange that can lead to new partnerships and new funding being leveraged.

Project Description Publications are just that: predictive descriptions of large projects undertaken by teams such as iBOL. They may contain little or no data, or preliminary data, but contain a description of the project, its objectives, the team, methodologies that will be used, timelines and deliverables, and mechanisms and timing of data release. Examples exist for other Community Resource Projects, for example see:

- The International HapMap Project. http://www.hapmap.org/ Nature. 2003 Dec 18; 426(6968):789-96.
- The ENCODE (ENCyclopedia Of DNA Elements) Project. Science. 2004 Oct 22; 306(5696):636-40.

Data Release Publications are peer-reviewed publications that describe preliminary datasets within a project. It is recognized and stated within the publication that these are preliminary data that will be refined and further analyzed at later stages in the project, e.g. Hubert *et al*., 2008[5]. Although the data and taxonomic identifications in that paper are relatively complete; additional validation derived from ongoing research in this group will provide much value to the scientific community and to the public data resource.

Proprietary Data

iBOL members consider all barcode data within BOLD a community resource to be shared publicly according to the terms and conditions outlined in this policy. There is no Intellectual Property associated with these data.

Privacy/Ethical/Sensitivity Concerns

1) Generally, there are few privacy concerns associated with DNA barcode data. However, some restrictions may apply, for example, in the case of data associated with samples within the Human Pathogen Working Group. Any iBOL researcher who engages in collection of samples from human subjects is expected to comply with national and institutional ethical requirements, and all proper documentation must be submitted before start of specific sub-projects. Consistent with applicable Privacy legislation or other policies, any data associated with such projects must be sufficiently anonymized such that no personal identification can be made.

2) As samples are collected to develop the iBOL DNA Barcode Library, some of these samples will be taken from ecologically fragile and sensitive areas. Some countries may have concerns about releasing information that may publicly identify those ecologically sensitive sites. For example, access to GPS data of orchid species within tropical forests could increase the risk to those species. If such a concern is raised, iBOL researchers will commit to holding such sensitive data confidential, or providing other means of anonymizing the data; GPS data are not mandatory.

3) Several members of the iBOL team are researchers within government departments, and they are mandated to monitor for invasive or harmful species within their country or region. Such departments and organizations have proscribed protocols for public release of information about invasive or otherwise harmful species. There may be concerns that early release of DNA barcode data, as described above, of an invasive species, may contravene government protocols. It is incumbent upon an iBOL researcher, whose first concern is with such protocols, to include an assessment of identification of invasive species in the QA/QC protocol, as described above. If preliminary data identify an invasive or harmful species, then those data must not be released until the restrictions or requirements of the specific government or department are satisfied. The Board of Directors of iBOL will be informed when such situations arise that limit the release of data.

Governance

All members of iBOL will adhere to this Policy. Any requests for extensions of timelines, advice on interpretation, other questions, will be directed to the iBOL Board of Directors.

---

[1] The definition of "community resource project" was developed at a meeting held on January 14-15, 2003 in Fort Lauderdale, Florida. The report on the conclusions of the meeting, "Sharing Data from Large-scale Biological Research Projects: A System of Tripartite Responsibility", can be found at http://www.genome.gov/Pages/Research/WellcomeReport0303.pdf

[2] Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, Tatusova T, Thomson N, Allen MJ, Angiuoli SV, Ashburner M, Axelrod N, Baldauf S, Ballard S, Boore J, Cochrane G, Cole J, Dawyndt P, De Vos P, dePamphilis C, Edwards R, Faruque N, Feldman R, Gilbert J, Gilna P, Glöckner FO, Goldstein P, Guralnick R, Haft D, Hancock D, Hermjakob H, Hertz-Fowler C, Hugenholtz P, Joint I, Kagan L, Kane M, Kennedy J, Kowalchuk G, Kottmann R, Kolker E, Kravitz S, Kyrpides N, Leebens-Mack J, Lewis SE, Li K, Lister AL, Lord P, Maltsev N, Markowitz V, Martiny J, Methe B, Mizrachi I, Moxon R, Nelson K, Parkhill J, Proctor L, White O, Sanson S-A, Spiers A, Stevens R, Swift, P, Taylor C, Tateno Y, Tett A, Turner S, Ussery D, Vaughan B, Ward N, Whetzel T, San Gil I, Wilson G, Wipat A. The minimum information about a genome sequence (MIGS) specification. 2008. *Nature Biotechnology* 26, 541 - 547) http://www.nature.com/nbt/journal/v26/n5/abs/nbt1360.html

[3] http://www.nature.com/nature/journal/v453/n7191/pdf/453002a.pdf

[4] CBOL/INSDC approved BARCODE marker (e.g. 5' COI for animals)

[5] Hubert N, Hanner R, Holm E, Mandrak NE, Taylor E, Burridge M, Watkinson D, Dumont P, Curry A, Bentzen P, Zhang J, April J, Bernatchez L. 2008. Identifying Canadian Freshwater Fishes through DNA Barcodes. *PLoS One* 3(6): e2490