REPORT ON REVIEW OF IBOL TARGETS AND MILESTONES

Karen James
23 April 2012, updated 30 August 2012

Executive Summary

This is a report on a review of iBOL targets and milestones at the project's mid-point. The review was carried out in consultation with the iBOL Scientific Steering Committee (SSC) and over 65 other iBOL participants and other DNA barcoding stakeholders. Acknowledging that this review is based on information provided by a cross-section of global DNA barcoding stakeholders at a single point in time, and cannot therefore be viewed as comprehensive, the key findings and recommendations are summarized as follows:

Findings

- The DNA barcoding stakeholders consulted in this review affirm iBOL's goals (i.e. to build a global accessible library of DNA barcodes for eukaryotes and promote applications for science and society), but also raise concerns and note conditions for success. These include concerns about the tension between data quality and quantity.

- As part of iBOL's numerical targets, approximately 1 million specimens will need to be barcoded to support applications. There is a higher quality requirement for these specimens, particularly in relation to how well they are identified.

- The extent to which these 1 million specimens overlap with the growing DNA barcode reference library is unknown. What is the identity of these specimens? If and when the numerical target of 5 million specimens is reached, will it include them? If not, the success of iBOL's Goal B – the promotion of applications of DNA barcode data fro science and society – is potentially at risk.

- The combined, planned efforts of the DNA barcoding stakeholders consulted for this review will result in the barcoding of approximately 4 million preserved specimens and 2.8 million newly collected specimens. Well over 200,000 additional preserved specimens and approximately 1 million additional newly collected specimens could (and would) be made available for DNA barcoding at an external sequencing facility, if funding to support that sequencing could be identified. Thus the provision of specimens is unlikely to be a rate-limiting factor in meeting iBOL's numerical targets.

- The sequencing infrastructures of the existing DNA-barcoding facilities are sufficient to meet iBOL's goals – both the numerical targets and in terms of supporting applications – but these infrastructures are not operating at full capacity. Funding is the limiting factor.

Recommendations

- Subsequent to this review, a more in-depth follow-up activity should be undertaken to generate the information and tools needed to establish a stronger and more deliberate connection between iBOL's goals and the specimen-to-barcode supply chain. This "matchmaking service" should enable the use of wish-lists of species needed to support applications to identify sources of priority specimens. The development of such a service – which would need to be done at the level of species names – is well

beyond the scope of this review. It will require contracting a bioinformatics-savvy postdoctoral level research assistant for perhaps 6-12 months, full-time, to create databases on both 'goals' and 'supply chain' sides, and a tool to match them.

- To use this matchmaking service in support of iBOLs goals, a rigorous and transparent mechanism will need to be put into place to facilitate the movement of priority specimens identified through the service through the specimen-to-barcode supply chain, and to promote and ensure the higher standard of quality required for specimens that support applications.

- Barcoding stakeholders who participated in this review affirm that an important iBOL priority is broad phylogenetic coverage across eukaryotic life. Thus, in terms of the definition of targets and milestones under the SSC's Theme 1 in support of iBOL's Goal A, this review recommends the establishment of a a new "breadth target" on top of existing numerical targets for each Working Group.

- Finally, this review recommends that iBOL explore opportunities for securing funding to support the full utilization of existing but dormant sequencing infrastructures for DNA barcoding. The establishment of a "matchmaking service" as recommended above will support and inform any funding proposals that might emerge from this review.

Introduction

The iBOL board of directors has requested a review of targets and milestones to be presented at the board meeting on April 24th, 2012. This review has been undertaken as part of the SSC's 'Theme 1', i.e. 'to define and achieve targets and milestones for the DNA barcode library ("5M, 500K") which are consistent with the project's scientific and societal priorities (cf. Theme 4) and are achievable with respect to available resources'. This Theme is associated with iBOL's Goal A, i.e. 'build a global, accessible library of DNA barcodes for eukaryotic species'. It is expected, however, that emergent findings and recommendations will underline the importance of Goals B and C, which relate to applications for science and society, and coordination, respectively, as well as highlighting the relevance of Theme 2 (technology) and Theme 3 (quality concerns related to bioinformatics) to Goal A.

Two questions drive this review. First, is iBOL on target to reach its Goal A, that is, to build a global accessible library of 5 million DNA barcodes corresponding to 500,000 eukaryotic species? Important factors include the provision of tissue samples from voucher specimens and the receptivity of core facilities to generating barcodes from these samples. Second, will the resulting barcode reference library satisfy iBOL's Goal B, i.e. to promote applications of barcode data for science and society including "economically, socially or environmentally important species" for "pest and disease control, food production and safety, resource management, conservation, research, education and recreation"? Important factors include the prioritization of sets of specimens from which reference sequences are required to support these activities.

The purpose of the review is to gather and analyze information from a selection of iBOL participants and other DNA barcoding stakeholders in order to address these questions. It is hoped that this will in turn prompt and support actions by the iBOL board of directors to enable the fulfillment of iBOL's mission: "To develop a rigorous DNA-based identification system for all eukaryotes, and to apply this new tool to better manage, discover and protect global biodiversity."

Methods and instruments

On a consultancy basis and working closely with Pete Hollingsworth, Peter Freeman and other Scientific Steering Committee members, Karen James (KJ) gathered and analyzed information and opinions from iBOL participants and other DNA barcoding stakeholders. Of 136 stakeholders approached, 68 responded

by either agreeing to be interviewed by phone/internet call (13) or by filling in an online survey (55). A more detailed breakdown of the numbers of stakeholders approached, and the proportion of those who responded to the interview/survey request ("respondents" from here forward), listed by stakeholder role is shown in Table 1 below. Stakeholder roles were assigned by KJ using information from the iBOL website and various iBOL documentation (see References). For respondents' self-reported roles, see the Results section. For a complete list of stakeholders' names, roles, email addresses and whether each responded to requests for input by filling in the survey/being interviewed, see "Stakeholders.xls", attached.

Table 1. Number of stakeholders approached, and respondents, by stakeholder role.

	Approached	Interviewed by phone/internet call	Filled in survey	Overall turnout
WG chair/vice-chair	21	3	8	52%
Node representative/alternate	64	4	28	50%
Campaign/collection initiative coordinator	26	4	16	77%
Major museum representative	4	1	0	25%
Sequencing capacity provider	11	7	2	82%
User (applications of barcoding)	20	3	8	55%
Connect group creator	9	0	2	22%
ECBOL signatory/representative	7	0	2	29%
Other/unknown	8	0	4	50%
Total (stakeholder roles)*	**170**	**22**	**70**	**54%**
Total (stakeholders)*	**136**	**13**	**55**	**50%**

**The difference between the number of stakeholder roles and the number of stakeholders is a consequence of some stakeholders having multiple roles.*

The online survey (see "Printable Survey.pdf" attached) was designed to gather information about the number and kind of specimens that could be made available or collected for DNA barcoding through 2015, and the capacity for sequencing those specimens. Hosted by SurveyMonkey (http://www.surveymonkey.com/), the survey was lengthy – 81 questions in total – but each survey respondent was presented with a series of yes/no questions about their roles and interests that directed them to just those subsets of the survey relevant to them. For example, a respondent with an interest in DNA sequencing capacity would have been presented with just 28 questions.

Phone/internet call interviews were carried out following along with those survey questions relevant to each individual interviewee, and typically also explored a range of related topics including a greater depth of discussion on iBOLs priorities and roles. The online survey was filled in by KJ on behalf of each interviewee using a combination of minutes and call recordings for reference. Thus the results described below include the combined information and opinions from both survey and interview respondents. Minutes of interviews are available on request from KJ. Recordings were made with permission for the purposes of review by KJ alone and on the condition of confidentiality and are therefore not available.

To generate the summary tables and charts shown in Results below, survey and interview results were exported from SurveyMonkey as Excel spreadsheets ("Survey Results Summary Report.xls" and "Survey Results All Responses.xls", attached), cleaned up as and when necessary, then summed for inclusion in tables or selected as reference cells for pie charts. In the case of open-ended survey responses, sums were calculated either from respondents' exact entries or, if the entry was vague or unclear, from KJ's interpretation of their entries from respondents' entries. In the case of multiple-choice survey questions

offering a series of ranges (e.g. 0-1,000; 1,000-10,000; 10,000-100,000; etc.), the number of respondents selecting each range was multiplied by the median of the two range endpoints (e.g. 500; 5,000; 50,000; etc.).

The geographical locations of survey respondents were extracted from the IP addresses of survey respondents using Batch IP Locator (http://batchiplocator.webatu.com/). These were combined with the geographical locations of interview respondents, manually entered by KJ, then mapped using BatchGeo (http://batchgeo.com/).

Results

Selected results from the online survey (which, as noted above, includes data from interviews as well) are described here in text and graphs. For complete survey results, see "Survey Results Summary Report.xls" and "Survey Results All Responses.xls", attached.

1. **Survey and interview respondents**. The information about stakeholders and respondents under "Methods and Instruments" above was assigned by KJ prior to the outset of the survey and interview period based on prior iBOL documentation. The information about respondents below is self-reported.

   Figure 1. The geographical locations of interview and survey respondents.

   The pool of respondents included 13 self-reported iBOL working group (WG) chairs, 6 WG vice-chairs, 30 iBOL node representatives and 12 iBOL campaign coordinators. In addition to these iBOL roles, 31 respondents self-reported as involved or interested in the management/curation preserved specimens, 44 in collection initiatives, 38 in DNA sequencing and 37 in applications of DNA barcoding.

   Respondents' taxonomic interests ranged from animals (including moths and butterflies, beetles, flies and other insects; birds, fish, mammals and other vertebrates; mollusks, sponges, crustaceans and other marine invertebrates; and nematodes and other soil and aquatic invertebrates) to plants (including trees and other vascular plants; and mosses), marine algae, lichens and fungi. Respondents' geographical interests (as opposed to their physical locations as reported in Figure 1 above) ranged across all continents and oceans. The relationships between barcoding stakeholders'

geographical interests and their physical locations can be further explored in “country originating vs contributing.xls”, an Excel pivot table prepared by Greg Singer, attached). Respondents’ interests in the applications and implications of DNA barcoding included ecology, phylogenetics, systematics and taxonomy, evolution, biogeography, biomonitoring, biodiversity assessment and management, citizen science, conservation, governance issues, trade legality, collections management and curation and DNA sequencing.

2. Sources of specimens for DNA barcoding

a. **Preserved specimens.** 56 respondents answered “Yes” to the question, “Do you work at a institution, department or collection that holds – or have an interest in – specimens?” About half of these respondents answered the subsequent questions in this section of the survey on behalf of their institution, a quarter on behalf of a department or section and a quarter on behalf of an individual collection. The tables and charts below summarize the respondents answers to these questions, which relate to the provision of preserved specimens for DNA barcoding.

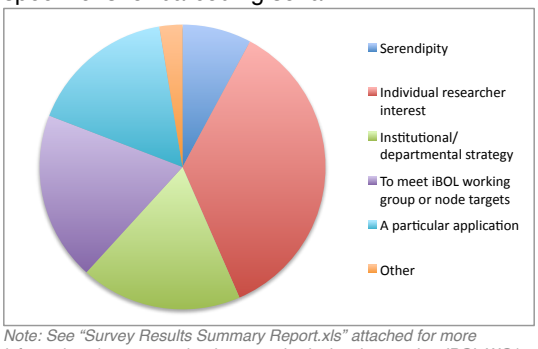
Table 2. Number of preserved specimens for DNA barcoding (all respondents combined). For more taxonomic detail about these specimens, and for information about those specimens that would be unavailable for barcoding and why, see “Survey Results Summary Report.xls” attached.

Held in the institutions, departments and collections represented by respondents	~100 million
Already barcoded	~1.5 million
Not yet barcoded, but in the barcoding pipeline	~0.7 million
Not yet in the pipeline, but earmarked for barcoding	~1.8 million
Not earmarked for barcoding, but could (and would) be made available for barcoding by another facility, if funding to support that sequencing could be identified.	>200,000*

**This figure needs to be re-confirmed for two reasons: 1) several of the responses were vague but potentially very large, such as the respondent who said they could send “huge numbers”, 2) the language used in this survey question was potentially confusing.*

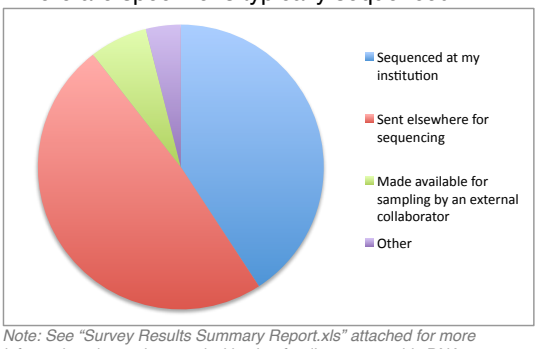
Figure 2. Charts summarizing information about the provision of preserved specimens for DNA barcoding.

What factors have guided the selection of specimens for barcoding so far?



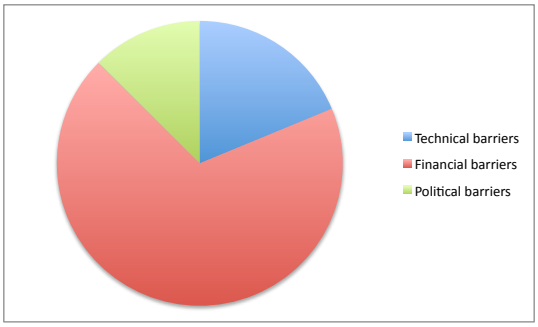
Note: See "Survey Results Summary Report.xls" attached for more information about researcher interests, institutional strategies, iBOL WG/ node targets and applications.

Where are specimens typically sequenced?

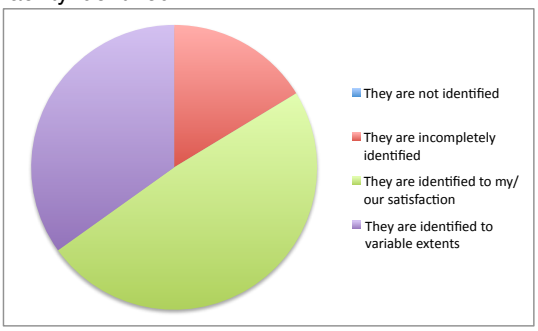


Note: See "Survey Results Summary Report.xls" attached for more information about where and with what funding support this DNA sequencing was done.

What are the major barriers to barcoding at your institution?



How well are those specimens that might be made available for barcoding at an external facility identified?



Note: Respondents report under a separate question that a large majority of these specimens can be reasonably expected to yield DNA.

b. Field collection initiatives. 47 respondents answered "Yes" to the question "Are you involved/ interested in the collection of fresh specimens?" The charts and tables below summarize the information collected from these respondents about the provision of specimens for DNA barcoding through collections initiatives.

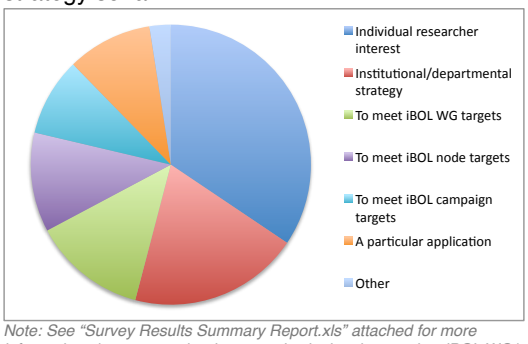
Table 3. Number of specimens to be collected through 2015 (all respondents combined). For more taxonomic detail about these specimens, and for information about those specimens that would be unavailable for barcoding and why, see "Survey Results Summary Report.xls" attached.

	Intend to collect	Intend to barcode
Vertebrates	~100,000	~61,000
Invertebrates	~5.7 million	~2.7 million
Plants	~1.1 million	~34,000
Fungi	~14,000	~7,000
Total	~6.9 million	~2.8 million

Notes: For more information about where and with what funding support this barcoding will be done, see "Survey Results Summary Report.xls" attached.

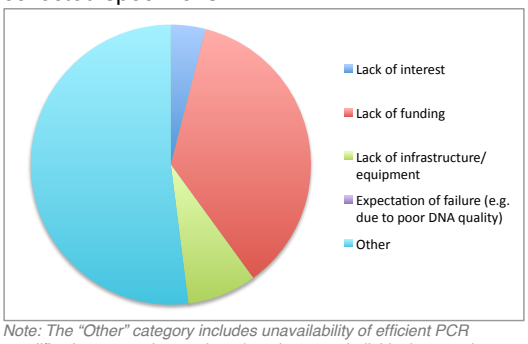
Figure 3. Charts displaying summary information about the provision of specimens for DNA barcoding through collections initiatives.

What factors have guided your collection strategy so far?



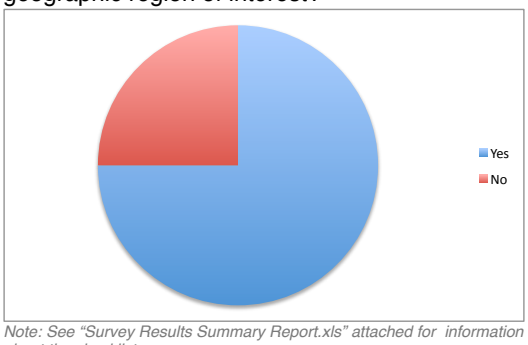
Note: See "Survey Results Summary Report.xls" attached for more information about researcher interests, institutional strategies, iBOL WG/ node targets and applications.

What is the primary reason for not barcoding collected specimens?



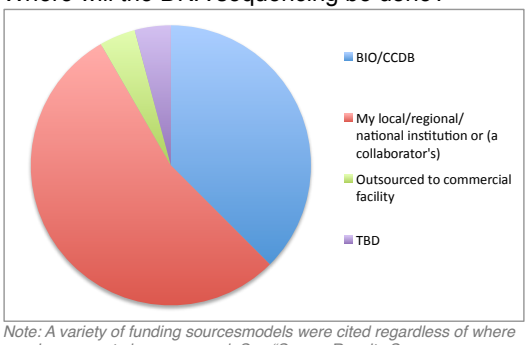
Note: The "Other" category includes unavailability of efficient PCR amplification protocols, specimen/species caps, individual researcher interest, poor identification, avoidance of redundancy and perceived limits on the number of specimens/species allowed in BOLD.

Do you have a reliable checklist for your taxa/ geographic region of interest?



Note: See "Survey Results Summary Report.xls" attached for information about the checklists.

Where will the DNA sequencing be done?



Note: A variety of funding sourcesmodels were cited regardless of where specimens are to be sequenced. See "Survey Results Summary Report.xls" attached for more information about sequencing facilities and funding support for collected specimens.

3. **Sequencing capacity for DNA barcoding.** 36 respondents answered "Yes" to the question, "Do you work in a laboratory or facility that does – or have an interest in – DNA sequencing?" The charts and tables below summarize the information collected from these respondents about DNA sequencing capacity for DNA barcoding.

Table 4. Sequencing capacity for DNA barcoding (figures in terms of sequencing reads i.e. single runs, so divide by two if you want equivalent number of specimens).

		CCDB*	SI	NHM	Kunming	All other respondents combined***	Total***
Theoretical capacity**		~1.4 million	~350,000	~350,000	100K-1M	~5.8 million	~8.4 million
Actual capacity 2011		~960,000	~250,000	~70,000	~100,000	~1.6 million	~3 million
Barcode sequences 2011		~960,000	~70,000	~10,000	~100,000	~1 million	~2 million
Capacity not yet committed	**2012**	0	200,000	240,000	N/A	~140,000	~580,000
	2013	~140,000	200,000	240,000	N/A	~180,000	~760,000
	2014	~140,000	200,000	240,000	N/A	~200,000	~780,000
	2015	~140,000	200,000	240,000	N/A	~200,000	~780,000
	Total	~420,000	800,000	960,000	N/A	~700,000	~2.9 million
Spare capacity available to external users "for free", but w/ strings attached (e.g. certain specimens/users)		Yes	No	No	Yes	Yes (30,000)	
Spare capacity available to external users "for free", with no strings attached		No	No	No	No	Yes (80,000)	
Spare capacity available for external users, if funding to support that sequencing could be identified.		Yes	Yes	Yes	?	?	

**Modified from CCDB Receptivity Statement to reflect single sequencing reads rather than bidirectional "barcodes".*

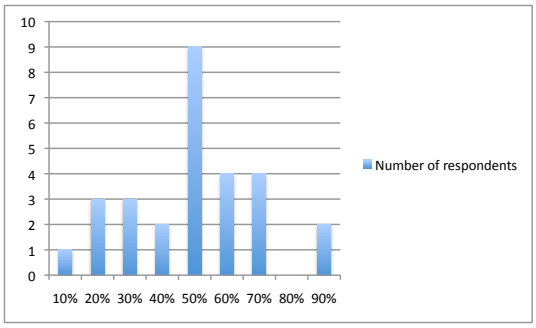
***Theoretical capacity is reduced by 50%, on average, for archival specimens (Figure 4).*

****The figures for 'All other respondents combined' and 'Total' are likely to include some redundancies and therefore should be considered provisional, pending verification. Two biggest contributions to the '1 million' figure for barcode sequences in 2011, for example, were from Dirk Steinke (500,000) and Chris Meyer (350,000). Their figures seem likely to overlap with CCDB and SI's figures, respectively. If so, this number (and the total) would be reduced by ~850,000.*

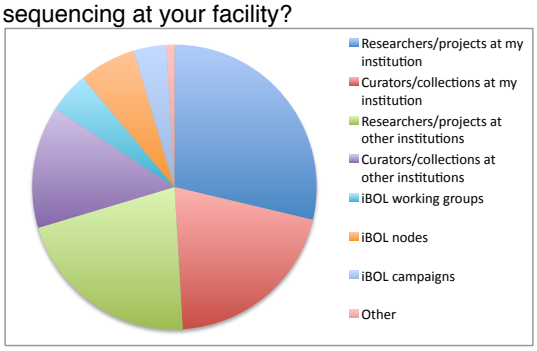
*****See "Survey Results Summary Report.xls" attached for more information.*

Figure 4. Charts displaying summary information about DNA sequencing capacity for barcoding.

How is capacity at your facility affected by the age of specimens? If all specimens were archival (i.e., >10 years old), by what % would capacity be reduced?

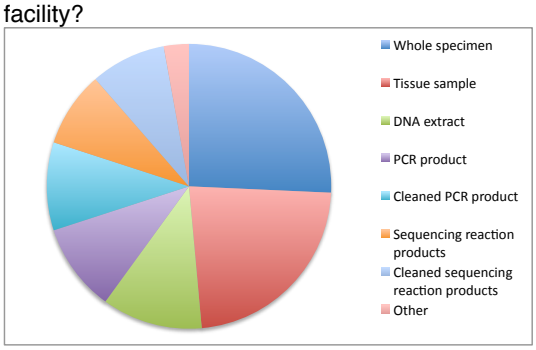


Who/what provides specimens for DNA barcode sequencing at your facility?

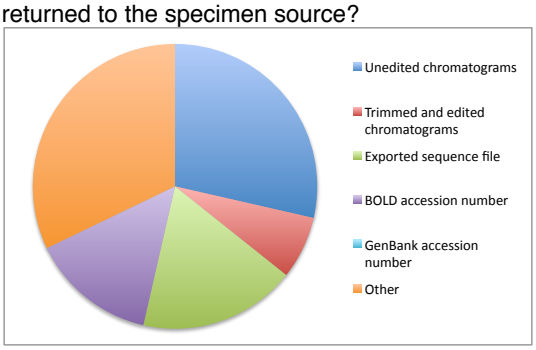


Note: See "Survey Results Summary Report.xls" attached for more information about the specific iBOL WGs, nodes and campaigns.

In what state do specimens normally enter your facility?



In what state are DNA barcodes normally returned to the specimen source?



4. **Requirements of specific applications and user communities.** 35 respondents answered "Yes" to the question "Are you involved/interested in using DNA barcoding for a particular application?" These respondents included David Schindel, who answered interview questions on behalf of CBOL's seven target user communities including agricultural and forestry pests in quarantine, water quality indicators, endangered species, consumer protection (especially food fraud), medicinal plants, use of barcodes in species descriptions and biotech companies. The charts and tables below summarize the information collected from these respondents about the data needed to support their applications and communities.

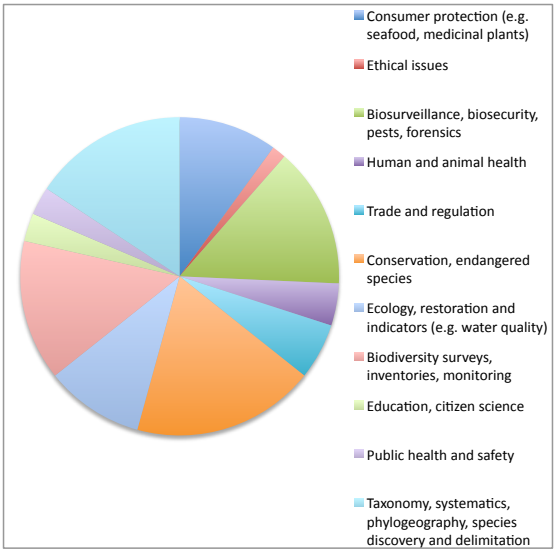
Table 5. Number of specimens that need to be DNA barcoded to to support applications and user communities.

Consumer protection (e.g. seafood, medicinal plants)	~5,000
Biosurveillance, biosecurity, pests, forensics	~160,000
Human and animal health	~10,000
Trade and regulation	~3,000
Conservation, endangered species	~100,000
Ecology, restoration and indicators (e.g. water quality)	~230,000
Biodiversity surveys, inventories, monitoring	not available
Education, citizen science	not available
Public health and safety	not available
Taxonomy, systematics, phylogeography, species discovery and delimitation	~120,000
Total	~630,000

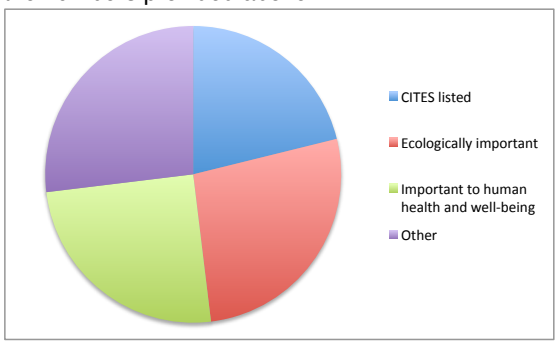
Note: These numbers are almost certainly a very significant underestimate. Not all respondents were willing or able to provide numbers for their particular applications, and those that were given were often vague. Moreover, a majority of respondents did not include allies that would need to be ruled out in an identification procedure (see Figure 5 below).

Figure 5. Charts displaying summary information about the applications of DNA barcoding.

In what application of DNA barcoding are you involved/interested?

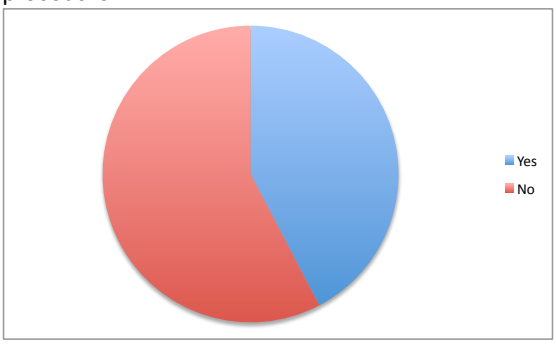


What criteria have been/will be used to generate the numbers provided above?

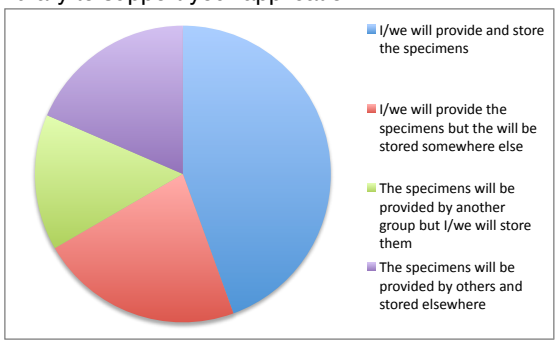


Note: The "other" category was dominated by the assembly of a taxonomically comprehensive barcode library for a particular region.

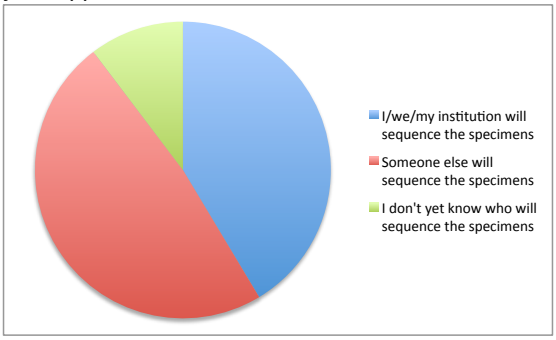
Do the numbers provided above include allies that would need to be ruled out in an identification procedure?



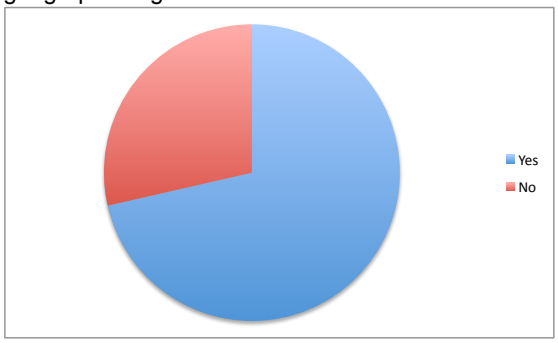
Who are you expecting to provide/collect and store the specimens required to generate the reference library to support your application?



Who are you expecting to provide DNA sequencing capacity to generate the reference library to support your application?



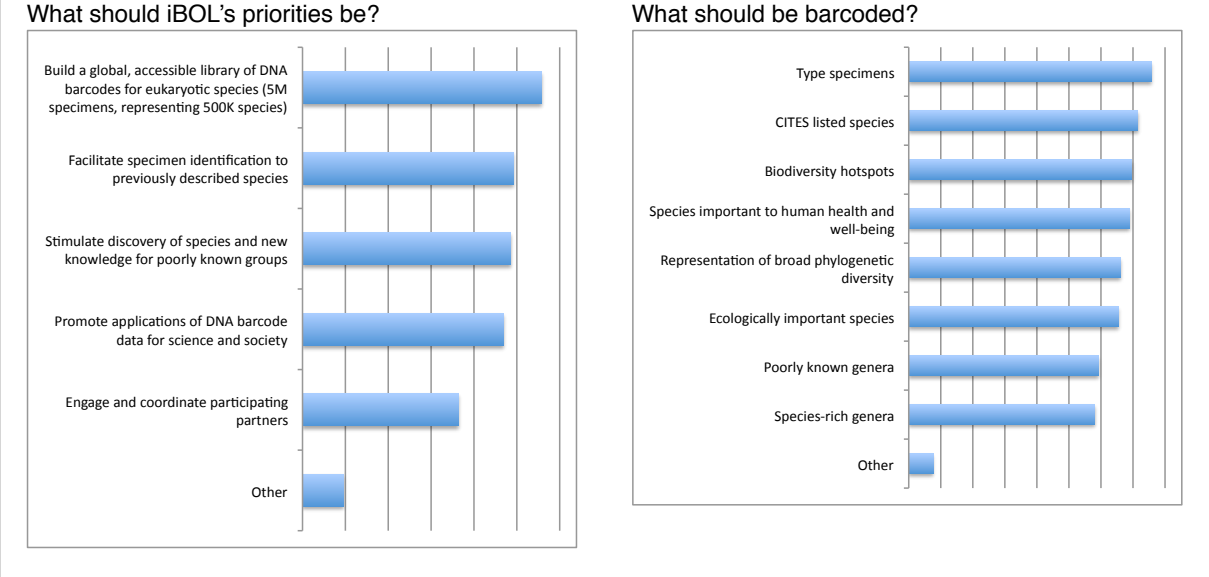
Do you have a reliable checklist for your taxa/ geographic region of interest?



Note: See "Survey Results Summary Report.xls" attached for checklists

5. **Perceptions about iBOL's priorities.** All respondents were asked to comment on iBOL's priorities and targets; 53 respondents did so. The charts below summarize their opinions.

Figure 6. Charts displaying summary information about survey and interview respondents' opinions about iBOL.



Finally, survey and interview respondents were given an opportunity to make any other comments about iBOL priorities or the rest of the survey. Here is a sampling of their paraphrased remarks.

<u>On quality vs. quantity</u>

- iBOL should shift its focus towards quality instead of quantity, with reliable identifications and high quality photographs.
- *Genus sp.* is a "massive problem".
- The reason the BARCODE flag and BOLD were invented in the first place was to solve the problems of low quality IDs and low quality sequences. These problems persist in the form of poorly- or un-identified specimens and an inflexible approach to sequence editing in BOLD that disenfranchises users who wish to shepherd their own sequences through the process.
- BARCODE records are not considered reliable by stringent users like regulatory agencies; e.g. the FDA has signaled its intention to collect all new data.
- 350,000 records got "kicked out" of GenBank for not meeting data standards; more could follow, and that's not even addressing the specimen identification problems.
- iBOL should shift away from numerical targets and towards improving existing records.
- "Numbers mean nothing, if quality isn't there."

<u>On supporting the DNA barcoding community</u>

- "I would expect more cooperation", supported by funding from iBOL.
- "DNA Barcoding is an important taxonomic tool whose importance in Africa still needs further promotion and support."
- "Who is the customer that IBOL is serving?"

On technological development

- “Next generation sequencing is moving fast. DNA extraction for DNA Barcoding need to be done with this in mind, extracting and saving DNA in a way so that the day it cost nothing to sequence genomes they can be used.”
- There should be investment in multi-plex next generation sequencing for DNA barcoding.
- Why aren’t we developing a handheld device?
- The 'hand-held barcoder' may not be there yet but “is it indeed true that bioliteracy has globally increased, can we measure that - or is it too early to conclude?”

On what should be barcoded and for what purpose

- “The key thing is not to fill up the library with the 'easy to obtain' at the expense of the 'things that matter'.”
- “In all cases, we should be going to where data are needed, not where the supply of specimens is rich.”
- Though it is important that a few groups (e.g. Leps and birds) be barcoded comprehensively, as this will illuminate the properties, opportunities and challenges posed by a well-sampled reference library, the most important overriding goal should continue to be constructing a library with the broadest possible phylogenetic coverage, with good, vouchered, well-identified material
- “I support a focus on known species of recognized importance--species we already know but sometimes can't identify…”
- “Type specimens are critically important, BUT they are likely not worth the time and money. …This should be marginalized to a working group trying to establish techniques and best practices for acquiring barcodes from type collections. Reports of high success in current literature are, in my opinion, grossly overblown...”
- For African species, “collections are critical given the state of old reference collections in museums”.
- Museum sampling is preferred over major collections initiatives (related to quality vs. quantity).
- “iBOL could have a real impact on species discovery and sharpening taxonomic boundaries.”
- Collecting from under-sampled habitats is important but hasn’t been mentioned much here.
- There needs to be a feedback mechanism that will make sampling, processing and analysis responsive to evolving priorities, i.e. where we want and need data.

On practicalities and realities ‘on the ground’ including funding

- “Our situation is that there is a lot of enthusiasm towards barcoding, but very little technical assistance…. We have therefore focused our efforts on finding funding for these tasks and not for lab infrastructure….”
- “Make the message of the DNA barcoding movement more relevant to a broader public and business at large. If they find it interesting the money will follow and so will the researchers.”
- Problematic genera are a potential strength of barcoding; this should be used to leverage funding.

Discussion and recommendations

Around half of the DNA barcoding stakeholders approached for input on this review responded. Even if they had all responded, this would still not represent a comprehensive view of the specimens and sequencing capacities available worldwide. Additionally, not all respondents answered every question in the survey, even in those areas for which they reported a role or interest. Moreover, it is likely that most DNA barcoding stakeholders wear multiple ‘hats’ and thus respondents may or may not have provided information for their whole institution. Similarly, multiple respondents from the same institution, or sharing the same sequencing facility, might have provided redundant information, though in the case of CCDB in particular, the potential for redundancy was monitored and removed as and where appropriate. Finally, as described in the Methods and Instruments section, information provided by respondents’ was sometimes

vague or unclear and had to be averaged or interpreted before it could be incorporated into sums for tables and charts.

Bearing these disclaimers in mind, the findings and recommendations of this review, undertaken in consultation with the iBOL Scientific Steering Committee and over 68 iBOL participants and other DNA barcoding stakeholders, are discussed below.

Findings

The DNA barcoding stakeholders consulted in this review affirm that three important priorities for iBOL are:

1. Build a global, phylogenetically (and geographically) diverse DNA barcode reference library of 5 million specimens representing 500,000 species
2. Facilitate specimen identification, particularly in the context of applications of DNA barcoding important to science and society
3. Stimulate discovery of new species and new knowledge

These priorities map nicely to iBOL's existing goals, in particular A (build a global accessible library of DNA barcodes for eukaryotic species) and B (promote applications of barcode data for science and society). In relation to each of these priorities, however, DNA barcoding stakeholders also raise concerns and conditions.

The facilitation of specimen identification, for scientific, social and commercial applications, for example, will require closer attention to quality over quantity. And though the stimulation of discovery requires exploratory collection initiatives and comprehensive sampling of a few groups (e.g. Lepidoptera, fish, birds) to model the properties, opportunities and challenges of a well-sampled DNA barcode reference library, this should not be done at the expense of meeting the targets for phylogenetic breadth, (i.e. the Working Group targets, 40,000-50,000 vertebrates, 100,000 plants, 10,000 fungi, etc.) or barcoding the wish-lists (or "shopping lists") required to support the applications of DNA barcoding.

As part of iBOL's numerical targets, approximately 630,000 well-identified, high-quality specimens will need to be DNA barcoded to support applications important for science and society (Table 5). Since approximately half of the wish-lists on which these numbers are based do not include allies that would need to be ruled out in an identification (Figure 5), the number of specimens that need to be barcoded for applications likely approaches 1 million.

Unlike the overall goal of 5 million specimens across a range of taxonomic groups (represented by the Working Group targets), the specific identity of these 1 million specimens is pre-determined by the applications; i.e. they must correspond to particular species wish-lists (or "shopping lists") as determined by the user communities associated with each particular application. What is the identity of these specimens? If and when the numerical target of 5 million specimens is reached, will it include them? If not, the success of iBOL's Goal B – the promotion of applications of DNA barcode data fro science and society – is potentially at risk. Furthermore, in addition to the need for specimens to support specific applications, there is also a higher quality requirement for these specimens, particularly in relation to how well they are identified.

Even though those who contributed to this review represent less than half of the global DNA barcoding community, their combined, planned efforts will result in the barcoding of 4 million well-identified preserved specimens (Table 2). Moreover, of the ~100 million other specimens held in respondents' repositories, well over 200,000 could (and would) rapidly be made available for DNA barcoding at an external sequencing facility (Table 2), if funding to support that sequencing could be identified.

In addition to preserved specimens, respondents also plan to collect approximately 6.9 million new specimens from the field through 2015 (Table 3). Of these, respondents are planning to DNA barcode 2.8 million, mainly at respondents' own institutions and/or with their own resources. There are no plans to barcode the remaining 4.1 million specimens, partly because of lack of funding, partly anticipated technical difficulties and partly lack of interest or to avoid unnecessary redundancy. Thus perhaps 1 million newly collected specimens could (and would) be made available for DNA barcoding at an external sequencing facility, if funding to support that sequencing could be identified.

The combined plans for barcoding preserved and newly collected specimens described above suggest that the provision of specimens is unlikely to be a rate-limiting factor in meeting iBOL's numerical targets.

At the outset of this review, it was expected that DNA sequencing capacity would be the rate-limiting factor in the generation of DNA barcodes to meet iBOL's goals. This review finds that though the sequencing infrastructures of existing DNA-barcoding facilities are sufficient to meet iBOL's targets. The total theoretical capacity of respondents' sequencing facilities is ~8.4 million sequencing reads per year (Table 4), corresponds to roughly 4 million specimens per year. Funding limitations mean, however, that those capacities are not being even remotely close to fully utilized; the total actual operating capacity is currently ~3 million reads per year, with approximately 2/3 of that capacity representing DNA barcoding-related sequencing (Table 4).

Recommendations

Following this review, a more in-depth follow-up project should be undertaken to generate the information and tools needed to establish a stronger and more deliberate connection between iBOL's goals and the specimen-to-barcode supply chain. On the 'goals' side, this project should include the identification and coalescence of wish-lists of well-identified (i.e. 'Phase 2') barcodes to support applications, and also reasonably well-identified (i.e. 'Phase 1' + 'Phase 2') barcodes to meet the Working Group targets. On the 'supply chain' side, this project should include a mechanism to query against these wish lists the specimen lists and checklists associated with repositories and field initiatives so that priority specimens can be identified and directed towards DNA barcode sequencing facilities. The development of such a "matchmaking service" – which would need to be done at the level of species names – is well beyond the scope of this review. It will require contracting a bioinformatics-savvy postdoctoral level research assistant for perhaps 6-12 months, full-time, to create databases on both 'goals' and 'supply chain' sides, and a tool for querying these databases against each other.

To use this matchmaking service in support of iBOLs goals, a rigorous and transparent mechanism will need to be put into place to facilitate the movement of priority specimens identified through the service through the specimen-to-barcode supply chain, and to promote and ensure the higher standard of quality required for specimens that support applications.

Barcoding stakeholders who participated in this review affirm that an important iBOL priority is broad phylogenetic coverage across eukaryotic life. Thus, in terms of the definition of targets and milestones under the SSC's Theme 1 in support of iBOL's Goal A, this review recommends the establishment of a a new "breadth target" on top of existing numerical targets for each Working Group.

Finally, this review recommends that iBOL explore opportunities for securing funding to support the fuller utilization of untapped sequencing infrastructures at existing sequencing facilities including both well-recognized DNA barcode sequencing facilities (e.g. BIO/CCDB, SI, NHM, etc.) and perhaps also additional, 'barcoding-ready' facilities (e.g. the South Australian Regional Facility for Molecular Ecology and Evolution and perhaps a sequencing facility in the Netherlands and/or Norway). The concept of a new funding proposal specifically to support sequencing capacity is not novel, but, armed with the information

provided by this review (once finalized) and any follow-up analyses, such as “matchmaking service” recommended above, there may be a greater chance of success.

References i.e. supporting documents and other information sources

- SSCE Targets and Milestones.docx (Peter Hollingsworth)
- iBOL Management Plan Outline.xlsx (Peter Freeman)
- Draft Management Plan.docx (Peter Freeman)
- Global Operations Report.docx (Peter Freeman)
- SSC Project Progress.pptx (Peter Freeman)
- iBOL WG1 limiting factors.xlsx (Pete Hollingsworth)
- CCDB sample receptivity statement 2012.docx (Greg Singer)
- CCDB Analytical Capacity.docx (Greg Singer)
- LAB-STRI Marine Fish Barcoding.docx (David Schindel)
- iBOL Project Dashboards (http://ibol.org/project-dashboard-wg-view/ http://ibol.org/project-dashboard-node-view)
- GenBank (http://www.ncbi.nlm.nih.gov/nuccore/)
- SurveyMonkey (http://www.surveymonkey.com/)
- Batch IP Locator (http://batchiplocator.webatu.com/)
- BatchGeo (http://batchgeo.com/)

Appendices attached as separate documents

- “Stakeholders.xls”, a complete list of barcoding stakeholders approached for input into this review, their roles, email addresses and whether they completed an interview or filled in the survey
- “Printable survey.pdf”, a printable version of the online survey hosted on SurveyMonkey
- “Survey Results Summary Report.xls”, a summary report of survey and interview responses as downloaded from SurveyMonkey
- “country originating vs contributing.xls”, an Excel pivot table prepared by Greg Singer exploring the relationships between barcoding stakeholders’ geographical interests and their physical locations/ affiliations

Appendices available on request

- “Survey Results All Responses.xls”, all survey and interview responses as downloaded from SurveyMonkey