RESPONSE TO QUESTIONS FROM THE INTERIM REVIEW COMMITTEE

QUESTION 1: Regarding taxonomic representation. Echinoderms for example are broadly represented, with many species from a very diverse set of taxa higher than genus, and with impressive global distribution. Other classes / phyla are represented by only a single genus, presumably from samplings that were provided by a group that is intensely interested in that one particular genus and collects all species from that single genus that they can get their hands on. I would really like to see an n x m table that describes the taxonomic diversity of the whole effort so far. n would be the appropriate highest level taxonomic classification, such as phylum or class, that contains barcoded specimens. m would be, in descending order, the 10 or so lower classifications, such as subclass, order, family, subfamily, genus, species. In each cell of the n x m table we would have the number of barcodes at that taxonomic level, for that phylum / class. The motivation behind asking for this table is to get a sense of how broadly the sampling is being done, across the thus-far represented phyla / classes. In addition, it would be great to have a summary of the geographic reach of the project, by an appropriately high-level taxonomic classification, though perhaps one notch lower than phylum or class. (So that there is a bit more granularity.) Perhaps these statistics could be found somewhere if one knew where to look, but my cursory searches did not turn up anything

Answer: Figure 1 provides a visual summary of known eukaryotic biodiversity, making clear the dominance of certain taxonomic lineages. Table 1 summarizes barcode coverage for representatives of the 25 phyla that compose the animal kingdom for five levels in the taxonomic hierarchy (Class, Order, Family, Genus, Species). The number at each level of the hierarchy reflects the diversity of taxa barcoded at that rank. Table 2 summarizes barcode coverage for plant and protist phyla based upon conventional taxonomy. The information in Tables 1 and 2 can be 'harvested' from the Taxonomy Browser on BOLD (www.boldsystems.org).

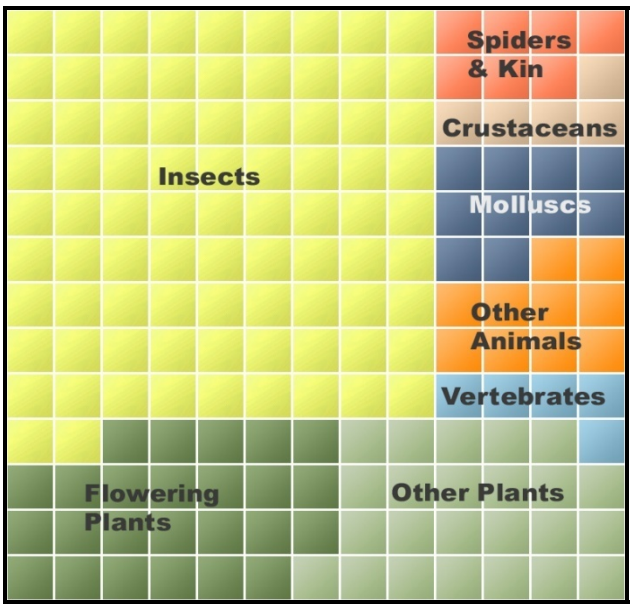

Figure 1: Taxonomic distribution of the 1.7M described species. Each cell represents 10K species.

Table 1: A summary of barcode coverage for animal phyla in BOLD (February 2011). The species count is a proxy based on BINs.

		NUMBER OF TAXA AT EACH RANK				
Animals	**INDIVIDUALS**	**CLASSES**	**ORDERS**	**FAMILIES**	**GENERA**	**SPECIES**
Arthropoda	802,596	16	110	1,136	17,497	113,561
Chordata	171,656	16	132	402	5,595	19,215
Mollusca	36,881	7	66	54	1,692	7,144
Annelida	11,548	4	19	16	478	2,497
Echinodermata	9,285	5	32	8	357	1,015
Platyhelminthes	1,702	4	17	16	119	248
Cnidaria	1,674	6	23	2	199	259
Nematoda	1,517	4	14	9	81	312
Rotifera	1,308	3	6		36	257
Porifera	486	3	12	3	59	103
Nemertina	486	2	5		35	143
Bryozoa	216	3	5		46	101
Onychophora	160	1			6	45
Acanthocephala	124	2	4		9	21
Tardigrada	113	1	1		6	10
Chaetognatha	108	1	2		9	42
Brachiopoda	88	4	6	14	26	47
Sipuncula	36	2	3		13	17
Hemichordata	9	1			2	5
Acoelomorpha	9	1			1	5
Gnathostomulida	8		2		7	8
Cephalorhyncha	8	1	1		1	6
Echiura	7	1	2	1	3	4
Entoprocta	3		1		3	3
Xenoturbellida	2				1	1
Totals	**1,040,030**	**88**	**463**	**1,661**	**26,281**	**145,069**

Table 2: A summary of barcode coverage for plant and protist phyla in BOLD (February 2011).

PLANTS	**INDIVIDUALS**	**CLASSES**	**ORDERS**	**FAMILIES**	**GENERA**	**SPECIES**
Magnoliophyta	35,124	3	56	48	2,597	9,468
Chlorophyta	451	3	5		20	64
Pinophyta	386	1	1		17	103
Pteridophyta	186	2	5		24	52
Cycadophyta	184	1	1		3	58
Marchantiophyta	83	1	1		3	10
Lycopodiophyta	17	1	2		3	9
Gnetophyta	5	1	1		2	3
Bryophyta	4	1	2		2	3
Totals	**36,440**	**14**	**74**	**48**	**2,671**	**9,770**

PROTISTS	INDIVIDUALS	CLASSES	ORDERS	FAMILIES	GENERA	SPECIES
Rhodophyta	13,243	3	23		430	2,081
Heterokontophyta	4,824	7	36		185	835
Pyrrophycophyta	1,100	1	11	1	33	98
Ciliophora	371	5	14		28	80
Cryptophycophyta	85	1	2		8	33
Haptophyta	24	2	5		10	18
Apicomplexa	11	1	1		1	8
Totals	**19,658**	**20**	**92**	**1**	**695**	**3,153**

The geographic scope of barcode coverage is already comprehensive. Figure 2 provides a heat map showing all sites with barcode coverage, while Figure 3 shows barcode coverage for the 50 nations with the greatest number of records in BOLD. These results have been assembled through customized analytics.

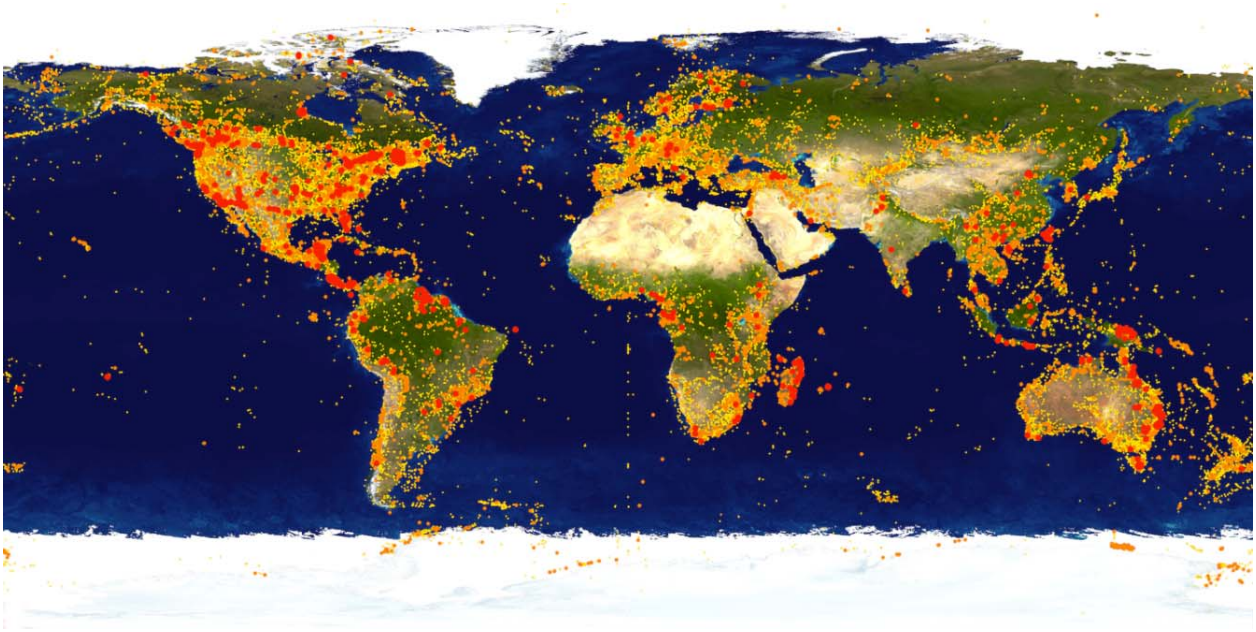

Figure 2: Heat map showing the geographic distribution of barcode records on BOLD. Red dots indicate sites with more than 100 records; orange dots indicate those with 10 - 100 records; yellow dots indicate sites with 1 - 10 records.

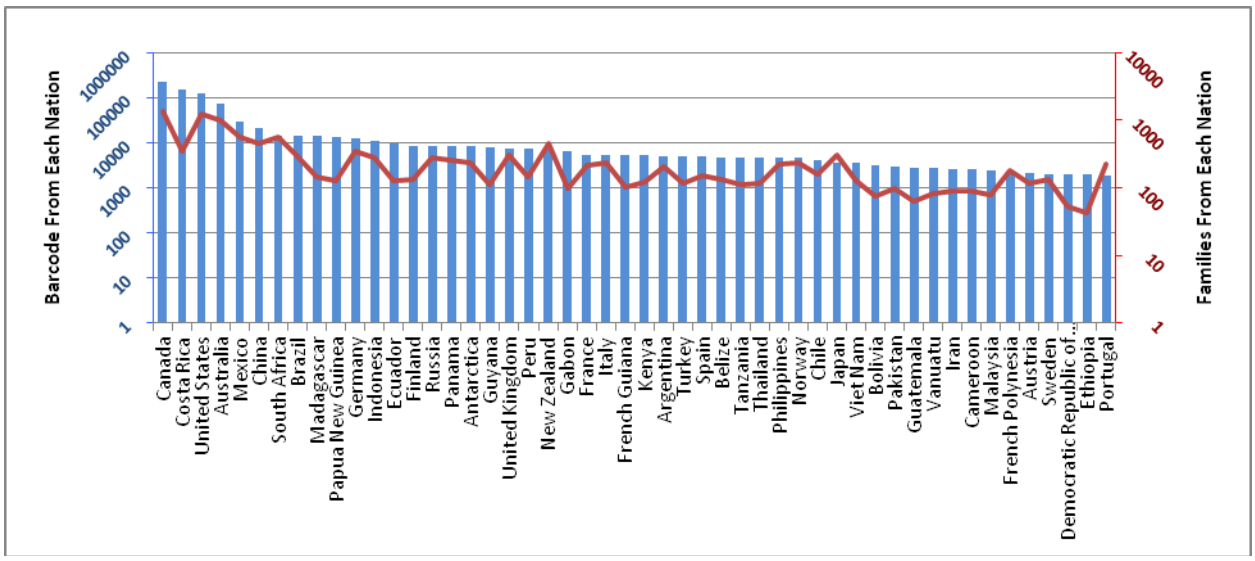

Figure 3: The 50 countries with the largest number of barcode records gathered during iBOL. The red line shows the diversity of taxonomic coverage from each nation as measured by the number of families represented.

QUESTION 2: Regarding impact of the project as a whole. I recognize that the team has provided citations of work arising directly or indirectly from the project; and also that certain google scholar searches were done that give a first impression of the impact. It would be helpful to understand what the team feels have been the biggest scientific accomplishments thus far that were enabled by barcoding, and that (preferably) were driven by a group not directly involved in the barcoding effort. To this end, I would suggest to identify five signal publications, and ask the team to provide a couple of sentences on each that describes the significance of the findings in those papers.

Answer: The primary mission of DNA barcoding is the creation of a molecular identification system for all eukaryotic species grounded in taxonomic knowledge. We see societal and scientific impacts of DNA barcoding in regulatory science, ecosystem monitoring, food web ecology, biodiversity discovery and conservation, and importantly, taxonomic practice. Interestingly, DNA barcoding is also providing new insights into speciation processes and molecular evolution. We highlight six papers in response to this question, citing only articles where all or most of the authors are outside the iBOL research team. The first two papers establish that DNA barcoding is gaining adoption in the surveillance of both the marketplace and natural environments. The next three papers illustrate the transformative effects of DNA barcoding in unraveling plant-herbivore interactions, enabling rapid biodiversity assessment in poorly-studied environments, and in strengthening the standards for taxonomic practice. The final paper explores new questions about speciation raised by the nearly universal effectiveness of DNA barcoding in animals.

1. Handy SA, Deeds JR, et al. 2011. A single-laboratory validated method for the generation of DNA barcodes for the identification of fish for regulatory compliance. ***J. AOAC International*** 94: 1-10. *This paper represents a key step in the adoption of DNA barcodes as the standard method for the identification of seafood products in the marketplace. Although the paper is led by authors at the*

US Food and Drug Administration, the Canadian Food Inspection Agency will adopt the same approach to aid consumer protection in Canada.

2. Sweeney BW, Battle JM, Jackson JK and T Dapkey. 2011. Can DNA barcodes of stream macroinvertebrates improve descriptions of community structure and water quality? ***J. N. Am. Benthol. Soc.*** 30: 195-216. *This paper represents an important step in the adoption of DNA barcoding as a tool for the assessment of community structure and water quality for both ecological and bioassessment purposes*.
3. Navarro SP, Juradoo-Rivera JA, Gomez-Zurita J, Lyal CHC, Vogler AP. 2010. DNA profiling of host-herbivore interactions in tropical forests. ***Ecol Entomol*** 35:18-32. *The diversity of insects and potential host plants in tropical forests overwhelms our ability to unravel food webs through direct observation. In this study, a DNA barcoding approach was applied to identify beetles and their host plants using taxon-specific primers to selectively amplify insect and ingested plant barcodes from DNA extracts prepared from intact beetles. The authors conclude that 'This technique provides a new means of studying species diversity and plant-herbivore interactions in tropical forests, and removes the constraints of the need for actual observations of feeding in ecological and evolutionary study.'*
4. Michida RJ, Hashiguchi Y, Nishida M, Nishida S. 2009. Zooplankton diversity analysis through single-gene sequencing of a community sample. ***BMC Genomics*** 10: 438-445. *Marine zooplankton are an ecologically important but taxonomically challenging set of organisms due to their small size, fragility, and the large number of species from deeply divergent phyla. In this paper, DNA barcoding enabled rapid assessment of marine zooplankton diversity with COI barcodes from a single water column sample indicating approximately 200 species from 11 orders of animals, only about 5% of which could be assigned to named species.*
5. Teletchea F. 2010. After 7 years and 1000 citations: Comparative assessment of DNA barcoding and the DNA taxonomy proposals for taxonomists and non-taxonomists. ***Mitochondrial DNA*** 21: 206-226. *This paper reports on the growing, transformative impact of DNA barcoding on taxonomic practice, concluding 'the indisputable success of the DNA barcoding project is chiefly due to the fact that DNA barcoding standards considerably enhance current practices in the molecular identification field, and standardization offers virtually endless applications for various users'*.
6. Lane N. 2009. On the origin of bar codes. ***Nature*** 462: 272-274. *The availability of barcode libraries documenting limited intraspecific variation in animal species with widely divergent evolutionary and life histories has attracted new scientific scrutiny with a particular consideration of its implications for the study of speciation. There are several proposed mechanisms, but none that account for the near-universality of limited intraspecific variation in animals. This report explores whether mitochondrial divergence itself may be a driver of speciation through reduced hybrid fitness resulting from mitochondrial-nuclear incompatibility.*

QUESTION 3: Finally, it would be helpful to understand how the team is planning to leverage recent massive advances in sequencing to both extend the number of loci for barcoding as well as possibly reduce the costs associated with it. This issue would probably also involve some discussion of the

computational approaches that might be necessary to adapt barcoding to the changing landscape of sequencing.

Answer: Sanger sequencing remains the most effective approach for obtaining the high-quality long reads (i.e. 650-700 base) that are required for standard DNA barcode library construction. This workflow has been adopted by the project (and approved by the TDAG) and the sequencing pipeline is in place for the generation of DNA barcode libraries from single voucher specimens. However, the project team has also been at the forefront in exploring the application of next generation sequencing (NGS) technologies via Working Group 4.1 (Environmental Barcoding). Aside from our extensive research in evaluating and using NGS approaches for biodiversity analysis from bulk environmental samples, we have been working on pilot projects to use the NGS pipeline in regular barcode analysis. In this case, sets of 96 (or more) samples are tagged using oligos attached to standard barcode PCR primers and then pooled and sequenced on a 454 FLX platform. We will soon test this approach on a larger set of samples (i.e. 10,000 specimens) in parallel to Sanger workflow. As indicated by the reviewer, a computational pipeline needs to be developed specifically for the NGS-barcode workflow and this pipeline should be integrated into the BOLD-LIMS to handle the large volumes of sequence data that will result. We also note that provisioning specimens and pooling them into large sets of species requires automation and we have applied for funds to support this workflow. Although multiplexing PCR to obtain sequence information from other loci is a possibility, DNA barcoding performs extremely well with a single gene for animals and only cases of cryptic diversity require additional loci for verification. However, plant barcoding (with two or more loci) may benefit from this multiplex approach. We note that all DNA extracts are stored for further analysis if necessary. We close this response by emphasizing that the goal for DNA barcoding as a tool for species identification and discovery is to minimize rather than maximize the amount of sequence information used to deliver identifications. By restricting the sequence length and the number of genes surveyed, sequencing costs are minimized and existing analytical tools can be employed for data interpretation.

QUESTION 4: Mobile Barcoding (WG 4.2) – I found the planned strategic priorities for WG 4.2 at the bottom of page 39 rather vague. The report basically says that it is best left as a matter of private sector development, but that developers would likely invest in technology that has broader (e.g., medical) applications as well. iBOL then is left in a kind of lobbying mode to convince end users of the importance of a barcoding device, but I'm not sure how well this all comes together. Is there any more definite plan to develop a mobile barcoding device? Is iBOL tracking any developments in this area and is such a tool a viable proposition during the timeframe of the project?

Answer: There is no allocation in our award from Genome Canada to directly support the kind of alliances with the technology sector that will be required to produce a portable DNA barcoder. However, the support from Genome Canada will speed the development of an accurate, taxonomically comprehensive barcode library that will be a powerful stimulus to the emergence of barcode technology. Although our funding from Genome Canada does not support private sector alliances, the University of Guelph has established a full-time business development position to explore commercialization opportunities linked to iBOL. Dr. Peter Miller will take up this position on May 1, 2011. In advance of his appointment, Peter has had several meetings with Life Technologies to explore

their interest in developing a bench-top instrument that integrates the functions needed to carry out barcode analysis. We also sustain efforts to build community interest. For example, we co-organized a workshop in 2009 with the Ontario Genomics Institute to discuss the feasibility of a mobile DNA barcode device. Both genomics technology developers and potential users were engaged in discussions with the following outcomes: (1) Technological challenges to miniaturize DNA barcoding workflow need to be solved through R&D in areas that require engineering and microfluidic expertise and resources. (2) In addition, a market for species identification through barcodes must be established (i.e. through offering service using available DNA sequencing facilities) before major investments would be directed towards a mobile device. The iBOL project has been instrumental in addressing both of these issues. Through our collaboration with regulatory agencies such as FDA and CFIA, we have established DNA barcode analysis as the standard method for the authentication of fish and seafood products. Hence, a market for such service is now being established and a number of companies have started offering barcode services. While we believe that these developments, coupled with potential applications of mobile sequencing in medical and forensic diagnostics, will catalyze the development of a mobile DNA barcode device, it is extremely unlikely that a handheld device will be developed by 2015.

QUESTION 5: There is some natural overlap between WG 2.3 (Methods development) and WG 4.1 (Environmental barcoding). Both seem linked to using next generation sequencing to handle environmental samples as well as to significantly reduce overall costs. My main question concerns the statement in the middle of page 26, where they talk about unidirectional reads that do not produce barcode compliant records, but could identify "single individuals ... required for regulatory compliance." That seems like a bit of a contradiction, since regulatory agencies presumably would need evidence that would be considered "compliant" by some objective standard, yet this sounds like a "quick and dirty" kind of barcode. What would become of the large amounts of sequence data obtained from such high-throughput techniques, if they don't conform to BOLD standards, and would they be treated or archived in a similar manner? Or is this merely a for-profit kind of spin-off activity for iBOL?

Answer: The generation of the reference barcode library must follow the highest QA/QC standards and must ensure that DNA extracts are of high quality. However, simplified protocols will be useful for the application of barcoding and these protocols could well lead to the establishment of companies offering analytical services to the private and public sectors. One of the earliest regulatory applications of DNA barcoding will involve the analysis of specimens gathered for aquatic biomonitoring programs. Both Environment Canada and the Environmental Protection Agency in the USA are engaged in pilot studies to test the efficacy of this approach. Bio-surveillance applications have tightly proscribed protocols that require the taxonomic identification of 300 specimens per site. These identifications are currently done through the morphological examination of specimens, an approach which often fails to deliver species-level identifications. The cost of analyzing each sample through morphology is approximately $500 and DNA barcoding will gain wide adoption if it can come close to matching this price point.

The protocols employed in the development of the DNA barcode reference library involve bidirectional sequencing and the preparation of high quality DNA extracts that can be retained for subsequent analysis. For these reasons, the costs of analysis average approximately $10 a specimen. If similar protocols were adopted for biomonitoring, this would mean that analytical costs for each sample of 300

individuals would need to be at least $5000 ($3000 to cover analytical costs, $2000 for profit). Because this high cost would deter the application of DNA barcoding, we have developed an alternate analytical pathway that can deliver a reliable identification for freshly collected specimens. This pathway is not designed as a substitute for the protocols employed to build the DNA barcode reference library. It is designed as a low cost option for the application of DNA barcoding for biosurveillance.

QUESTION 6: BIN system. It's becoming more apparent that the iBOL protocols are better suited for most groups of animals than for plants or fungi, despite the increased ability of BOLD to deal with multiple markers. The examples of BIN species pages shown in Figs. 3.7 and 3.8 are very compelling, but it sounds like all of the BINs and species pages that have been generated so far are for animals. Will there be any other way to produce the equivalent of species pages for plants or fungi? How to avoid what appears to be a growing gap in the suitability/applicability of barcode data for animals vs plants and fungi?

Answer: Single gene DNA barcodes deliver 98% species discrimination for most animal groups, a performance that will not be matched for the plant kingdom with existing markers. However, we emphasize that cases of incomplete taxonomic resolution (e.g. generic rather than species-level assignment) often represent a hugely important advance from the status quo. For example, recent work has shown the power of DNA barcoding to provide a newly detailed perspective on seedling recruitment in tropical rain forests.

The ability of BINs to provide a rapid registration system for animal diversity is a key advance because it provides a mechanism to apply DNA barcoding to animal groups where many of the species are undescribed. The approach is not universally applicable to land plants, as they do not show such clear DNA 'barcode gaps' between species. Fortunately, the existing knowledge-base on plant diversity is far higher which reduces the requirement for a BIN-type system. Although it is too early to reach conclusions about the resolution that will be obtained in other eukaryote kingdoms (fungi, protists) because standard barcode markers have not been adopted for them, early results suggest that species discrimination success will approach that achieved in animals. As a consequence, there is reason to expect that a BIN approach will also be effective for these kingdoms, and will thus be applicable to the vast majority of eukaryotic life.

We emphasize that the BIN system already provides an outstanding capacity to track the growth in barcode coverage, allowing an overview of progress among geographic regions and taxonomic groups. For example, Figure 4 shows the current status of barcode coverage for terrestrial life.

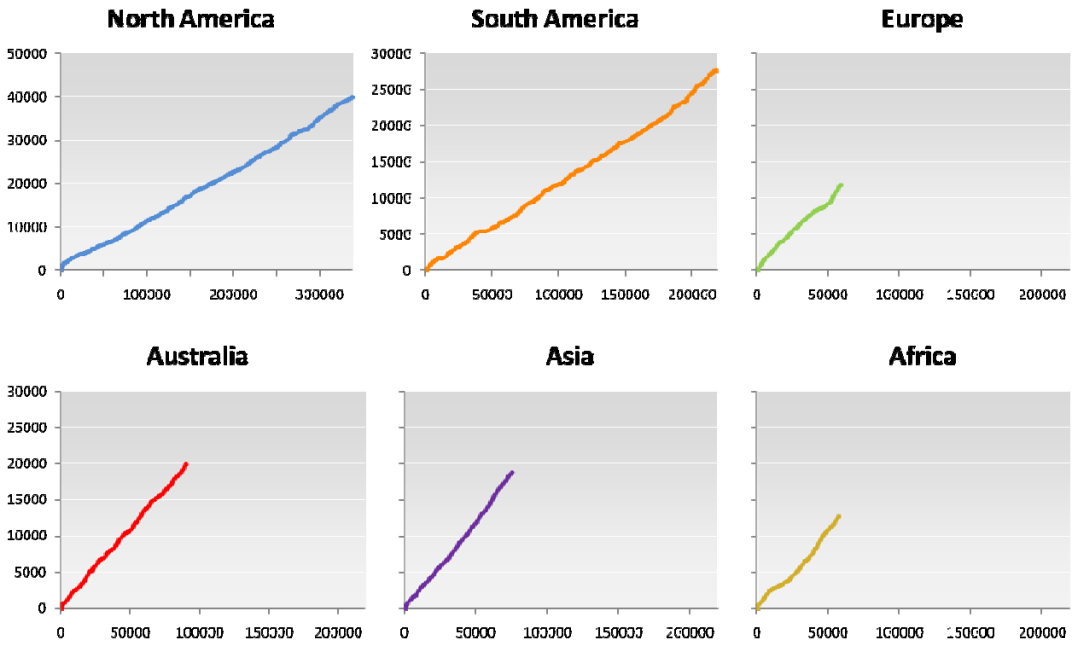

Figure 4: Accumulation curves for BINs with increasing sampling effort for six continents.

QUESTION 7: If I understand correctly, barcodes have been added to the database for 57,000 species since the onset of the project – i.e. during the last 18 months (making a total of 153,000 species). The target is to reach 500,000 by the end of the five-year duration of the project. At current progress, this would suggest an additional 133,000 species will be added during this period – totaling to 286,000 species. What factors will allow the considerable acceleration of species barcode acquisition to enable the project to meet its 500k target?

Answer: The iBOL project aims to deliver coverage for 500,000 species by December 31, 2015, a target which includes the legacy data that were available at the launch of iBOL. Barcode coverage was available for 96,000 species when iBOL began activity, and barcode coverage has been extended by 57,000 species over its first 18 months of activity. If this production rate (38,000 species per year) is sustained over the next 4.5 years, barcode records will be obtained from an additional 171,000 species by July 2015, resulting in iBOL adding coverage for 228,000 species to BOLD. When coupled with the legacy data, this will produce a total of 324,000 barcode records.

It is certain that that this total will be exceeded because plans call for a substantial rise in barcode production rates over the lifetime of the project. This increase will result, in part, from expanded production at the Canadian Centre for DNA Barcoding which has recently obtained the equipment needed to double its analytical capacity. A second major core facility is under construction (Kunming Biodiversity Centre, China) and it plans to analyze at least 200K sequences per year. Other smaller-scale facilities are either being established or are expanding their production. These facilities, which include the Natural History Museum (UK), the National Museum of Natural History (USA), Araungabad (India), and ECOSUR (Mexico), should collectively generate at least 100K sequences per year. Because this

increased analytical capacity is coupled with growing access to specimens, the members of iBOL's Scientific Steering Committee indicated their confidence when they met in September 2010 that barcode records would be obtained for 500,000 species by 2015.

QUESTION 8: With reference to the above, the variation in progress among groups, while explicable in the sense of decisions regarding the choice of sequence, would seem to be a cause for concern. The lack of progress with fungi, pathogens and marine biosurveillance species seems to require special measures. What specific measures are in place for these groups (plus the mentioned amphibians and reptiles)? Furthermore, would it not be more effective to divert some of the resources available to the other groups – predicted to exceed their targets – into these groups encountering more problems?

Answer: Because Genome Canada provides no funding to support the collection of specimens or their identification, it is not possible for us to divert resources to strengthen the supply of specimens that are in short supply. Fortunately, with the exception of the minor adjustments proposed for certain Working Groups in Theme 1, iBOL researchers are confident that they can achieve the targets for barcode coverage established in the original proposal. The lack of progress on fungi is a direct consequence of the need for 'certification' of the primary barcode marker(s) for this kingdom before work can begin in earnest. Fortunately, barcode designation is close to resolution as CBOL plans ratification of the fungal marker(s) in April 2011. The barcode target for the fungal kingdom (WG1.2) is 10,000 species, a total that will be readily achieved because one of iBOL's lead partner organizations, the CBS Fungal Biodiversity Centre maintains cultures of more than 30,000 species. Moreover, it has received the funds to gather a barcode record from each of them. The target for WG1.4 (Vertebrate pathogens, vectors and parasites) is 10,000 species and coverage is currently in place for 20% of these species. Many of the species within this WG are arthropods which can be obtained from museum collections and efforts are underway to rapidly build barcode coverage for this WG. Barcode studies on marine organisms (WG1.7) confront the complication that specimens in museums have been preserved in formalin, a reagent that seriously impedes DNA sequencing. Efforts to launch new collection programs are one component of iBOL's response to this challenge, but such programs are extremely expensive. As a consequence, our strategy involves making effective use of ongoing marine sampling programs. For example, cruises funded by NOAA are leading to the acquisition of substantial collections of marine zooplankton for barcode researchers in Mexico. However, because of the complexity in sample acquisition, plans call for a slight reduction in barcode coverage for marine life. Our efforts to build barcode coverage for amphibians and reptiles have gained momentum as a team has been charged with leading progress on these groups.

QUESTION 9: Of the 90k specimens that failed to yield barcodes during the period in question, how realistic will it be to obtain additional samples to replace them and how do these samples break down by group? It seems that sample acquisition remains the real priority.

Answer: The summary of specimens that failed to deliver a barcode record reveals that diverse taxonomic groups are represented (Figure 5). The primary factor underpinning the failure to recover a sequence record for members of most animal phyla is specimen age rather than taxonomy as success drop from more than 90% in freshly collected specimens to 60% in specimens that are 50 years old

(Figure 6). Other failures reflect analytical error or the exposure of specimens to preservatives that resulted in DNA damage. Barcode recovery for plants is still impeded by the need for better primers and the same is true for a few groups of animals (e.g., nematodes). Because of the uncertainty in sequence recovery from any particular specimen, multiple individuals of a species are sampled whenever possible. To aid the orientation of sampling efforts, iBOL Working Groups are increasingly directing their efforts toward the analysis of museum specimens. When the initial round of analyses fail to deliver barcode coverage, it is often possible to seek additional specimens for analysis from other museums. We emphasize, as well, that we continue to recruit new participation from taxonomists and that these individuals are aiding the acquisition of freshly collected specimens from increasingly diverse sites.

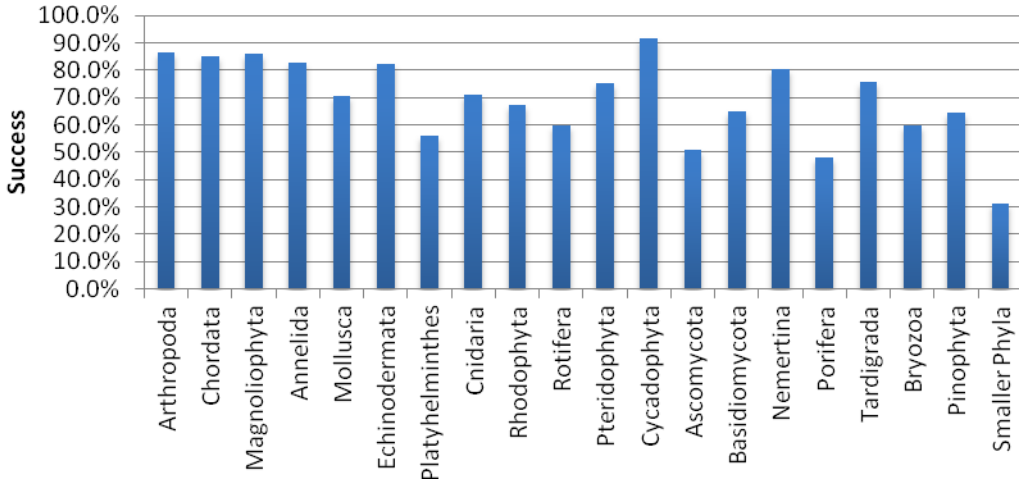

Figure 5: The percentage of specimens in 20 animal, fungal and plant phyla that generated a barcode.

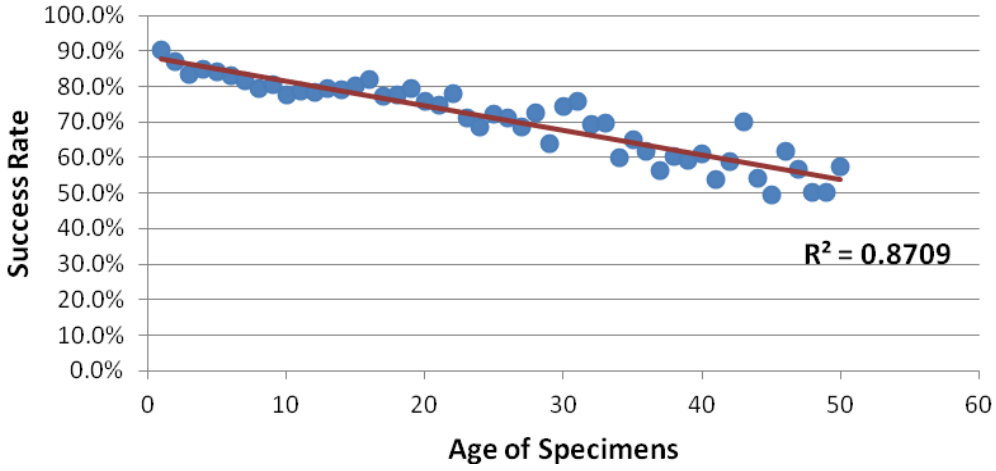

Figure 6: The success rate in barcode recovery for animals, fungi and land plant specimens.

QUESTION 10: In WG2.3 a stated aim is to move barcoding to an 'undirectional read'. What are the implications of this approach for base-call confidence (especially in recalcitrant specimens) and has this shift in emphasis been discussed with databases with bidirectional read standards (e.g. Genbank)?

Answer: There is no intention to shift the barcode reference library standard from a bidirectional to a unidirectional read. The use of unidirectional reads is only proposed as an option for **the application** of DNA barcoding in specific situations where it is necessary to identify freshly collected specimens at low cost. Such applications can often be accomplished with unidirectional sequencing because the pool of candidate species in both relatively small and well characterized. We emphasize that full-length, bi-directionally sequenced barcodes remain necessary for the construction of the barcode reference library and certain forensic applications.

QUESTION 11: In WG2.4, progress is stated primarily in terms of funding in Norway using mini-barcoding. To what extent are these mini-barcodes compatible with the sequences generated by the rest of the project?

Answer: Past work done on plants by Christian Brochmann's group has employed mini-barcodes in the trnL intron. They adopted this region because the conserved stem-loop structure in this intron has allowed the design of primers with near universality which amplify a variable but very short region. This bit of DNA is probably the best parameterized plant gene region, making it highly useful for this specific application (mini-barcoding). Unfortunately, the primers for this region are the subject of a patent filing by Taberlet et al., a fact which conflicts with iBOL's commitment to an open access identification system. Dr. Brochmann is keen to use standard COI barcodes as a basis for paleobarcoding studies on animal lineages and work is underway to develop primers with broad generality for a short fragment of the barcode region. We emphasize that mini-barcodes used for the analysis of museum specimens are fragments within the standard full-length COI DNA barcode. In this context their primary use is to link freshly collected specimens with taxonomically described museum specimens. We have done empirical and bioinformatics research (Hajibabaei et al. 2006, Hajibabaei et al. 2007, Meusnier et al. 2008) to demonstrate the applicability of different 100-200 base long fragments from standard barcodes in identification of species with ~91% resolution at species-level. Hence mini-barcodes make effective use of the reference sequence library generated by the rest of the project. As for plants, work is underway to develop similar mini-barcodes from core plants barcodes (rbcL and matK).

Hajibabaei, M., Smith, M. A., Janzen, D. H., Rodriguez, J. J., Whitfield, J. B., and Hebert, P. D. N. 2006. A minimalist barcode can identify a specimen whose DNA is degraded. Molecular Ecology Notes 6:959-964.

Hajibabaei, M., Singer, G. A., Clare, E. L. and Hebert, P.D. 2007. Design and applicability of DNA arrays and DNA barcodes in biodiversity monitoring. BMC Biology 5:24.

Meusnier, I., Singer, G.A., Landry, J. F., Hickey, D. A., et al., 2008. A universal DNA mini-barcode for biodiversity analysis, BMC Genomics 9:214.

QUESTION 12: For WG4.1, I am a little puzzled why it is necessary to carry out "*a systematic evaluation and optimization of protocols*". 454-based metagenomics of eukaryotes is now relatively routine and is powerful enough to overcome imbalances in biomass (the bias referred to). Indeed, I would go far enough to say that COX-1 sequence analysis using this platform is already fit-for-purpose and could be rolled out immediately.

Answer: We agree that the sequencing capacity of NGS platforms can aid in overcoming biases associated with biomass differences. However, for DNA barcoding at a species-level we require both long reads (i.e. >150 base) and standard barcode information (i.e. COI barcode region). The first requirement means that we need to use NGS platforms that provide long reads. The 454 FLX system has been used for this reason despite its lower throughput. We also need to use DNA barcodes as targets for amplicon-based analysis. Most past studies used 18S or other ribosomal loci because of the availability of universal primers, but these gene regions ordinarily fail to deliver species-level resolution. Over the past two years we have done extensive research on "known" mixtures and have shown that PCR bias can hugely influence qualitative and quantitative analysis of environmental samples. We have developed modified PCR regimes and new primer sets to reduce this problem and have successfully tested this method on moderately complex mixtures such as freshwater benthic samples (Hajibabaei et al. 2011). More work is now underway to adopt this approach for other environmental samples.

Hajibabaei, M., Shokralla S, Zhou X, Singer G, et al., 2011. Environmental barcoding: a next-generation sequencing approach for biomonitoring applications. PLoS ONE (Accepted).

QUESTION 13: Numts – can you say a little about how and where these problems have been most severe and what technical advances have been made to overcome them?

Answer: NUMTS are a minor problem to barcode recovery. Prior work has shown that most NUMTS are short (less than 200 bp) (e.g. Richly and Leister 2004). One of the most important defenses against NUMTS lies in the decision to bi-directionally sequence the barcode amplicon. In those cases where a short NUMT amplicon is admixed with a full-length product from the barcode region, it only impedes sequence analysis in the 'front' section of each read. Because of the bi-directional read, clean sequence can be recovered for the complete amplicon. This defense fails for NUMTS that are long enough to include the entire barcode region, but their copy number is typically low (a few copies per genome), while the authentic mitochondrial gene region is present in hundreds or even thousands of copies per cell. As a consequence, the PCR amplicon pool generated is ordinarily dominated by the mitochondrial gene region. Primer mismatches to the mitochondrial copy can result in situations where only the NUMT is amplified. To recognize such situations, all barcode sequences are automatically checked by BOLD for frameshift mutations, for stop codons and for unexpected amino acid substitutions that are good indicators that a NUMT has been sequenced. Such cases are rare, representing less than 0.09% of all sequences gathered since July 2009. In such cases, the authentic mitochondrial gene can almost always be recovered by shifting to another primer set.

Richly E and Leister D. 2004. NUMTs in sequenced eukaryotic genomes. Mol Biol Evol. 21:1081-1084.

QUESTION 14: The interim review report gives insufficient details about changes to the GE3Ls research program and progress to date. What, exactly, is the proposed new activity on taxonomy, and what social science methods would it use?

QUESTION 15: To say that the earlier Activity 5 was "not universally valued" is an understatement; I saw only serious critiques in the previous reviews. To what extent and how is it being curtailed, and what ambitions remain?

Answer: We answer questions 14 and 15 together since they relate to the same programmatic changes.

On July 5, 2010 Genome Canada's letter of award, indicated that our revised plan had been "accepted for funding at the requested level of $999,259 over five years as a component of the overall iBOL budget…." No conditions apart from the timing of funding were attached, but the GE^{3}LS team was encouraged to consider the anonymous reviewer comments in case we find them "useful in [our] execution and development of the GE^{3}LS plan…."

Reviewer 1 and 3 had useful critical comments regarding Activity 5, and the comments were seriously considered. Reviewer 2, rather than criticizing Activity 5, actually called for something similar to what we proposed, but we recognize some shortcomings in how Activity 5 was described. Turning to other comments made by Reviewer 3, we noted a request to work on the taxonomic questions in barcoding, which had been in the original GE^{3}LS plan.

In light of these comments, we proposed in a September 1, 2010 memo to the iBOL executive, to include an activity (now 6.6) on the taxonomic question and to divide the budget and reorganize the timing of Activity 5 (now 6.5).The iBOL GE^{3}LS team is working with the iBOL executive to identify the correct path to sanction the proposed revised work, including the specification of research methodology.

QUESTION 16: What are the net budget effects of the above two for the GE^{3}LS budget?

Answer: The net budget effect will be neutral.

QUESTION 17: What is the timeline for effort and outputs from the different GE^{3}LS activities, so we can gauge progress? Several (actually, most) of the activities were not mentioned in the progress report. What is their status and expected timing?

Answer: During the period covered by the interim review, no funding was provided for GE^{3}LS research. Despite this fact, all members of the GE^{3}LS team participated in the September 2010 meeting of iBOL's Scientific Steering Committee. As well, there have been ongoing communications with the iBOL executive team, and interactions with other iBOL members. Research milestones and associated timelines will be established as soon as the contract period with researchers' institutions is established.