INTERNATIONAL BARCODE OF LIFE PROJECT (iBOL)
INTERNATIONAL CONSORTIUM INITIATIVE (ICI)
INTERIM REVIEW WRITTEN REPORT FORM

ASSESSMENT

1. RESEARCH PROGRESS

Progress against the original research plan is assessed as:

☐	Excellent	☒	Good	☐	Satisfactory	☐	Not Satisfactory

Please provide comments to justify your assessment. If there are significant variances in the progress of specific components of the ICI, provide specific comments by sub-project or activity while taking into account factors such as the:

- level of funding received to date;
- duration of the funding period;
- problems encountered;
- progress relative to other research groups working on similar projects; and,
- research team's responsiveness to advances in the field and/or availability of new technologies since the project's inception.

iBOL has had a number of experimental and informatics challenges to overcome since its inception and they have done a good of identifying the issues and developing plans to overcome them. They also seem quite responsive to critiques from various groups.
Number of barcodes to BOLD is ramping up well but progress has been slower in non animal species in part due to experimental issues. This could create problems in their expected deliverables and timelines should be revisited to reflect realistic completion dates.
The iBOL informatics team has experienced significant difficulties in its ability to appropriately submit data GenBank. This is reflected in the initial low numbers of records accepted by GenBank. Although much work has been done to overcome these challenges continued careful attention to QA/QC processes as well as data synchronization are paramount for this project to have useful, accessible and meaningful data. This will be of increasing importance as additional mirrors come on line. Apart from the excel document – BOLD GenBank tasks – no other mention is made of the QA/QC processes and not much additional information is provided in the progress report (synchronization was covered). It will be key for this group to develop clear documentation and automated pipelines to improve these processes and make them efficient.
Data access and mining enhancements developed by the team are making the data much more easily digestible by both power (informatics) and casual users. Data visualization will be of particular importance to this project and although still nascent, a number of positive steps have been made to make BOLD a much more interactive site for

various types of users. The BIN framework is also a key enhancement but should be monitored carefully. Non-animal species could be a problem because and alignments may be meaningless due to the nucleotide substitutions. This is noted by the iBOL team but is worthy challenge to try and solve. It will be important to ensure users understand the limits of any analysis tools and be instructed in their use.
The majority of the funding for this project is split between the wet lab (experimental) - Theme 1 and informatics (Theme 3) and this seems appropriate, for now. It will be interesting to see if lab costs can be reduced with the use of NGS technologies especially those providing longer read lengths and therefore requiring no fragment assemblies.
It is the experience of this reviewer that large complex genome projects usually underestimate the bioinformatics costs and often do not allocate sufficient funds for these activities. It is understandable that at the outset of a project these needs are not well understood. However as projects progress both infrastructures needs (compute power, storage etc) and analysis needs usually grow exponentially. I predict this project will experience similar issues and it would be wise to undertake an audit of bioinformatics needs on a regular basis and consider reallocation of funds as needed. Due to the complexity and geographical diversity of this project professional software engineers (SE) Human Factors / User Interface engineers and project managers would also be of great benefit to this project. This will likely increase costs but will result in a much more robust and easily usable system.

2. CHANGES TO THE RESEARCH PLAN

If significant changes to the research plan have been made or are proposed for the future, what is your assessment of these changes?

☐	Recommended as proposed	☒	Recommended with modifications	☐	Not recommended

Please provide comments to justify your recommendation. Provide details of any modifications or additional changes that you would recommend.

The changes submitted in the research plan are all aligned with improvement of the project and do not represent major deviations from the initial plans.
Informatics (Theme 3) has been actively working on making modification to how they work with GenBank and although this is productive (as stated above) more robust methods for QA/QC and automated methods need to be put in place. Formalizing the interaction with GenBank would be of great importance.
Administration (Theme 5) the project team acknowledges the need for proper project management but this is not described in any detail. iBOL encompasses a fine collection of scientists but project management skills and scientific skills usually do not overlap. Furthermore, scientists (especially academics) are usually very resistant to project management skills or professional project managers. Given the complexity of this project it is important that this issue be addressed and high quality project management be deployed.

3. GE^{3}LS

The project leaders were asked to identify the ethical, environmental, economic, legal and social aspects (GE^{3}LS) arising from their proposed research and to develop and implement a plan to address them.

The progress towards achieving the goals of the GE^{3}LS subproject is assessed as:

	Excellent		Good	X	Satisfactory		Not Satisfactory		N/A

Please provide comments to justify your assessment including comments on the progress towards achieving the GE^{3}LS milestones and the research team in place. Please take into account factors such as the level of funding received to date, duration of the funding period and problems encountered.
This reviewer is not an expert in GE^{3}LS but has worked closely with a number of these teams in both government and commercial settings. This is an essential component of this project however many issues remain unresolved and it is unclear how this group will effectively engage the other more biological and computing iBOL groups. 5 GE^{3}LS themes have been described and cover a lot of ground, most need to be developed in parallel but some may not. It would be useful to have a timeline outlining how this will be achieved as well as major milestones. Communicating with the other iBOL members early and often is important.

4. ABILITY TO ACHIEVE THE OVERALL OBJECTIVES

Is the research team likely to achieve the overall objectives of the project?

X	Very likely		Possibly		Unlikely

Please provide comments to justify your response. In cases where you identify issues that would prevent the team from meeting their objectives, please state the issues and where possible, propose solutions.
I have some concerns about informatics (robustness of processes, interactions with GenBank and sufficient budget), specimen acquisition and analysis – especially for non animal species and general project management practices. However, the team is robust, seems flexible, enthusiastic and willing to embrace change (at least from the written documents). A strong and effective SAB, regular reviews with useful feedback and incorporation of these recommendations by iBOL leadership should keep iBOL on track.

5. BENEFITS FOR CANADA

Based on progress to date and the future plans, please assess whether the anticipated results of the research are likely to, a) contribute to job creation and economic growth in Canada, b) social benefits c) improvements in the quality of life, health, and/or the environment, and d) contribute to the creation of new policies in these areas. If commercialization is proposed, please comment on the strategy for IP management and ownership, technology transfer and benefit sharing.

The progress towards realizing the Benefits for Canada is assessed as:

[x]	Excellent	[ ]	Good	[ ]	Satisfactory	[ ]	Not Satisfactory

Please provide comments to justify your assessment. Have any opportunities been missed?
This is a multinational project with a lot of visibility and impact on a number of scientific, education and social areas. Attempting to integrate these areas is hard but an important step in making science more directly relevant to a community. Complex genome projects are becoming the norm and this project (and presumably others) could showcase Canada's expertise in doing these projects and doing them well. I am very enthusiastic and impressed with the educational/outreach and GE^{3}LS programs. These are substantial programs and not just tacked on to the end of a scientific endeavor. The informatics component of this project is complex and substantial and becoming aligned with many of the major bioinformatics resources in the world is an excellent way to foster international relationships and develop other collaborations. Genome Canada should be proud of the efforts of iBOL and seek to gain as much knowledge as it can in the biological, genome technology, GE^{3}LS, education and bioinformatics arenas. There are commercialization opportunities to be had from this project but these need to be carefully assessed. The addition of a business development FTE is to be applauded but a well thought out business plan needs to be developed. Commercial ventures arising from iBOL could bring many economical benefits to Canada.

6. GOVERNANCE & MANAGEMENT

Are the established governance and management plans, processes, and structures appropriate and effective?

[X]	Yes	[ ]	Remediable Concerns	[ ]	No

Please provide comments to justify your assessment including comments on the:

- governance and management structures and processes;
- decision-making processes – do these ensure that critical decisions about the research direction can be made and provide the ability to respond to unanticipated difficulties?
- effectiveness of communication mechanisms within the project.

I give this a qualified "yes" rating. iBOL recognizes the need for governance and management practices and has put a number of these in place. The appointment of an SAB is of key importance to this group but this group must be strong, well organized and effective. Of utmost importance is the SAB chair as this person sets the tone of meetings and can be strongly influence the direction of iBOL.
As noted previously, I believe that professional project management skills and practices would be of great importance to this group. The management structures described in the progress report do not provide sufficient detail about how this will be achieved. Given the complexity and increasing geographic distribution of this project strong and effective project management is key.

7. HANDLING OF SCIENTIFIC DATA & RESOURCES

Are the plans for handling scientific data and sharing of data and resources created by the project appropriate and effective?

[x] Yes [ ] Remediable Concerns [ ] No

Please provide comments to justify your assessment and, where appropriate, make suggestions for the improvement of the plans.

A qualified "yes". As noted in section 1 the iBOL informatics team has experienced significant difficulties in a number of areas. I've copied some of that text here and expanded on a few areas.

The iBOL informatics team has experienced significant difficulties in its ability to appropriately submit data GenBank. This is reflected in the initial low numbers of records accepted by GenBank. Although much work has been done to overcome these challenges continued careful attention to QA/QC processes as well as data synchronization are paramount for this project to have useful, accessible and meaningful data. This will be of increasing importance as additional mirrors come on line.

Apart from the excel document (xls) March 2011– BOLD GenBank tasks – no other mention is made of the QA/QC processes and not much additional information is provided in the progress report.
It will be key for this group to develop clear documentation and automated pipelines to

improve these processes and make them efficient. I am not entirely convinced that these issues have been settled. For example in the xls document BOLD agrees that an update channel to GenBank needs to be established and it appears this started in October 2010 but it is still in a rudimentary phase. Data submissions to GenBank can be tricky and processes can take quite some time to be developed and implemented.
Given that data is moving between these groups I wonder about the efficiency of the current system and also the inherent errors that will arise if this process is not worked out quickly. This will also affect synchronization issues and ultimately the transparency and cohesiveness of the data held in BOLD and in GenBank. I highly recommend that BOLD and GenBank meet very regularly to establish best practices and to ensure a rapid and clean transfer of data.

Synchronization of data with GenBank is another area that will need careful attention. In particular if BOLD decides to store *adhoc* data from users that may not hold up to the standards of GenBank. This will create discrepancies in the data sets between sites. Data fragmentation and versioning could be a very real problem for this project and being vigilant is important.

I looked at a BOLD record held at GenBank and have some comments to provide. It would be helpful if the GenBank database cross-reference for BOLD would go directly to a specific BOLD entry as opposed to the main page. It may be the iBOL requirement to log into the system before this occurs but even after having logged in the user is still directed to the BOLD home page.

The comments section in the GenBank record seems devoid of richer taxonomic data that would be useful for the user. I wonder if additional work needs to be done with metadata exchanges between the groups.

GenBank has been developing the BioProject Pages concept that act as a type of “home page” for a specific project in GenBank. A BioProject page for iBOL could provide general information about the project, a description and links to data contained within GenBank and links back to the BOLD home page. This type of portal may be a very useful way for user to transparently see what is held in GenBank and BOLD.

Data access and mining enhancements developed by the team are making the data much more easily digestible by both power (informatics) and casual users. Data visualization will be of particular importance to this project and although still nascent, a number of positive steps have been made to make BOLD a much more interactive site for various types of users.

As this site grows screen real estate will become a premium it would be very easy for the BOLD web site to become cluttered and user unfriendly. I would recommend the utilization of a Human Factors engineer that is trained in designing user interfaces. Scientists and software engineers are notoriously bad at user interface design. Wonderful data that is not coherently presented and easily accessible will not be used.

Additional to this, workflows describing how a user would mine data would be of value. It

would be preferable for these to be online rather than as documentation alone.

The BIN framework is also a key enhancement but should be monitored carefully. Non-animal species could be a problem because and alignments may be meaningless due to the nucleotide substitutions. This is noted by the iBOL team but is worthy challenge to try and solve. It will be important to ensure users understand the limits of any analysis tools and be instructed in their use.

Web services are being developed for BOLD and I certainly applaud these efforts although it seems that both eSearch and eFetch are not implemented at the moment. These should prove useful but I think additional exposure of data might also be a good thing. If I understand the system correctly a user will be able to query the BOLD system via knowledge of the exposed schema and pull back data records based on this input. This is a pull approach and you must understand the schema to be able to pull back desired entries. It may be of use to expose all data to the user and make that data available via something like a BioMart system. This would be a push approach.

Lastly, I noted that appendix III - Board and Advisory committee member bios and reports - describes possible future hardware solutions. Currently iBOL uses localized CPU and storage but as the data and analytical tools grow there will be increasing pressure to expand the hardware to meet these needs. Data storage and CPU costs are skyrocketing whilst sequencing technologies are decreasing in cost. Scaling to meet these needs is important. The GRID approach taken by iBOL is sound but this may need to be rethought as data sets and analysis tools increase in size and complexity. I noted that the cloud was considered as an alternative but this was rejected for a number of good reasons, for now. It is true that refactoring software for the cloud is a major obstacle but processes are beginning to take shape that have lowered the barrier and make large complex pipelines much more amenable to the cloud. Bandwidth, as noted, is also an issue but there may be a trade off where data in the cloud is much more useful and accessible to a community. Especially when bringing tools to the data is much more useful than trying to shift very large data sets.

An important comment was made that IT personnel requirements will probably increase over time and these people are critical to the success of the project. These personnel are often the unsung heroes of projects and I wish to highlight the importance of their work. The IT systems provide the needed plumbing of a project and I am sure we are all aware of the outcome and urgency of plumbing emergencies when they occur.

8. TRAINING, RECRUITMENT AND PROFESSIONAL DEVELOPMENT

Has the project team been and/or will they continue to be successful in training, recruiting, and professional development of highly qualified personnel?

[x] Yes [ ] Remediable Concerns [ ] No

Provide comments to justify your assessment and, where appropriate, suggest possible new approaches. Please comment on the responses to any challenges faced to date.
This project has obviously attracted a lot of interest and clearly the iBOL leadership has been active in providing community workshops and training opportunities to quite a number of graduate students. It would have been helpful if the progress report had provided some user feedback from workshops and how the iBOL team plans to improve or change their workshops. In addition, plans for future workshops and their goals would also have been useful. The iBOL team should consider ways to scale training. Workshops are very useful but there are many electronic technologies available now to disseminate information and training than just face-to-face meetings.

9. PUBLIC OUTREACH & COMMUNICATION ACTIVITIES

Do the activities that have been undertaken or are planned ensure that the research is communicated to the public and other interested parties?

[x] Yes [ ] Remediable Concerns [ ] No

Please provide comments to justify your assessment and, where appropriate, make suggestions for improvement.
Again, there seems to be a lot of activity in this area but it was somewhat difficult to assess what exactly had been achieved. Some clear metrics and progress against these would be useful.

10. FINANCIAL

Is the research progress commensurate with how funds have been spent to date?

[x] Yes [ ] Remediable Concerns [ ] No

Please provide comments to justify your response. Include comments on the following: • whether explanations for budget variances are reasonable in the context of

research progress to date;

- the level of success in securing co-funding from other sources;
- if there is a co-funding shortfall, suggest alternate sources of co-funding and comment on how the shortfall will impact the research.

Budget seems appropriate and expenditure seems on target with research goals. As noted previously, informatics needs may increase as the project progress so reallocation of funds may be necessary upon a later review. Project management costs may increase if professional project managers are employed.

11. Progress towards addressing the issues raised at the time of the initial ICI review

☐	Excellent	☒	Good	☐	Satisfactory	☐	Not Satisfactory

Please comment specifically on the progress towards addressing the issues that were raised by the Expert Committee at the initial review listed below:

- the resolution of outstanding management, budget and co-funding issues;
- the revision of the GE^{3}LS component of the proposal;
- the establishment of a Technology Development Advisory Group (TDAG) and a Science Advisory Board (SAB);
- the implementation of a quality control/quality assurance monitoring system.
- the resolution of outstanding issues with the public release of data in the ***International Nucleotide Sequence Database Collaboration (INSDC);***

Most of the issues raised at the previous meeting have been addressed.
The GE^{3}LS component still needs work but it seems the groups are tackling this.
The bioinformatics issues are being addressed but continued vigilance is important and necessary.

OVERALL RECOMMENDATION

☐ Further funding for the ICI is recommended

☒ Further funding for the ICI is recommended with modifications

☐ Further funding for the ICI is not recommended

Please provide a brief summary of the status of the project and a justification for your recommendation. Where issues have been identified, state whether these are major or minor, what actions should be taken (e.g., activities that should be reduced, abandoned or strengthened), alternate approaches to be considered and avenues to strengthen the project.
This is a complex project with many moving parts and will only get more complex with time. iBOL has started well and as is to be expected not all WG are up to speed. It will be imperative to establish excellent project management to keep track of all the components and to help align goals with concrete outcomes. The bioinformatics to support such a project will also increase in complexity and scale it will be important to keep a careful watch of the processes so that data can be easily accessible by all types of users.