IWEG

International Workshop on Environmental Genomics

Workshop Report 5

Generating Engagement with Genomic Data

June 24th-25th, 2020

Thank you to our 2020 sponsor:

Contents

EXECUTIVE SUMMARY 1

INTRODUCTION 2

SESSION SUMMARIES 4

Day 1: Wednesday, June 24th, 2020 4

Keynote Talk: Bringing Ecogenomic Surveillance and Monitoring Data to End-users (John Darling, United States Environmental Protection Agency) 4

Session I 4

Session II 5

Day 1 Discussion 5

Day 2: Thursday, June 25th, 2020 7

Session III 7

Session IV 8

BEST PRACTICES FOR DATA VISUALIZATION: INTERACTIVE SESSION SUMMARY 9

Closing Remarks 14

Day 2 Discussion 14

GENERATING ENGAGEMENT 17

EXECUTIVE SUMMARY

Participants from twelve countries attended the 5th International Workshop on Environmental Genomics (IWEG), which was hosted by the Centre for Environmental Genomics Applications (CEGA) and eDNAtec Inc. in St. John's, NL. Because of travel restrictions and health concerns related to the global COVID-19 pandemic, this conference was held virtually for the first time. The theme of the workshop, "Generating Engagement with Genomic Data," aimed to promote discussion about 1) how to better communicate environmental genomics to diverse stakeholders, 2) how to make environmental genomics tools more accessible to these stakeholders, and 3) how to better engage these stakeholders through communication and collaboration. These ideas were explored over two days through a keynote talk (John Darling, U.S. Environmental Protection Agency), four sessions of presentations by workshop participants, and one interactive session. Participants represented stakeholders across many sectors, including the oil and gas industry, environmental consulting companies, regulatory agencies, and academia. Specific themes of the presentations included: 1) existing nation-wide research programs, 2) technological innovations, 3) opportunities created by these innovations for larger scale projects, 4) communicating the importance of environmental genomics tools, 5) improving cross-stakeholder networking, collaboration, and coordination to achieve larger research goals, and 6) standardization.

INTRODUCTION

The 2020 International Workshop on Environmental Genomics (IWEG), "Generating Engagement with Genomic Data", was held virtually through WebEx, and hosted in St. John's, NL, Canada on June 24th and 25th. Ninety-eight participants registered from twelve countries representing an array of sectors from environmental regulatory and consulting agencies, industry, and academia. Most participants (83%) had experience working with environmental genomics tools or were leading their own environmental genomics research programs (see Appendix A for details). The workshop was organized through multiple sessions, including an introductory session where John Darling (USEPA) gave the keynote talk, four presentation sessions, one interactive session, and concluding remarks given by Mehrdad Hajibabaei, Scientific Director of the Centre for Environmental Genomics Applications (CEGA). Roundtable discussions capped each day of the conference, and eight key themes emerged:

The need to responsibly address methodological conservativism.

Considering existing regulations when exploring the adoption of environmental genomics tools is critical to expedite uptake of these tools by regulatory agencies. To move the bioassessment field forward, many agencies are trying to demonstrate how DNA metabarcoding compares to traditional bioassessment approaches. However, comparing DNA metabarcoding to traditional methods in some ways creates a false equivalence. Rather than thinking about environmental genomics approaches as tools to replace existing methods, a better paradigm is to view these approaches as complimentary tools with unique capabilities and advantages.

Building trust among stakeholders.

Key to convincing various stakeholder groups of the promise of next-generation tools is to consistently give clear, accurate results, and to ensure these results are discernable to managers, decision makers, and policy makers. It is essential that environmental genomics practitioners fully understand and carefully communicate limitations and uncertainties to end-users through the development of a standardized reporting language and minimum reporting standards. Effective communication builds trust in environmental genomics and sharing data facilitates this by making the processes transparent. A final way of fostering trust among stakeholders is through strategic initiatives for improving community engagement.

Big data will provide new bioassessment tools, but regulations need to be developed to take advantage of these tools.

Most of the approaches to analyzing environmental genomic data simply apply old tools (e.g., those generated for traditional morphological data) to genomic data, which is a failure to realize the true potential of DNA data. One area that is generating a lot of excitement is the development of taxonomy-free analytical approaches for scaling up biodiversity surveys. These approaches will circumvent the need for associating sequences with taxonomic IDs, allowing researchers to capitalize on untapped information found in genomic datasets (i.e. measuring biodiversity directly from the genetic sequences through OTUs, ASVs, or similar) and enabling, for example, development of new ecological indicators based on metabarcoding data. Additionally, despite the fact that it is not currently possible to reliably assess species abundance and biomass from sequence read counts, there are plenty of existing and emerging applications that do not require quantitative abundance information for robust ecological assessments.

Increasing uptake of eDNA methods by regulatory agencies.

Regulatory agencies, including Fisheries and Oceans Canada, Environment and Climate Change Canada, the U.S. Environmental Protection Agency, the Scottish Environment Protection Agency, and others, have begun to adopt environmental genomics approaches. Many regulatory agencies are planning for much larger uptake of these approaches within the next five years, and they want to see—and are

actively working on—standardization of methods, lab accreditation, minimum reporting standards, validation, and other refinements of environmental genomics approaches. Given the legal mandates binding many of these agencies, the results must be easy to interpret and any uncertainties should be measurable and clearly stated.

Importance of standardization.

The consensus from workshop participants was clear: it is time to make concerted efforts towards standardization in this field. Though it is challenging to agree on a set of initial guidelines, standardization allows for validation and proliferation of protocols for the current generation of environmental genomics technologies. There are reservations that standardized methods will not keep up with technological advances and stagnate once the lines are drawn. On the contrary, standards and a central governing body can lead to more focused, unified efforts and accelerate advancements to the next generation of environmental genomic technology. New service providers that comply with the standards can quickly demonstrate competency by using externally validated methods and standards can instill confidence for all stakeholders leading to increased uptake.

End users face challenges requiring environmental genomics solutions.

Even if practitioners feel more needs to be worked out related to environmental genomics tools, managers, policy makers, and decision makers demand these new bioassessment tools now, given the biodiversity crisis we face. Rapid detection of changes in biodiversity, including invasive and threatened species detection, is necessary for faster, more effective responses to environmental problems. eDNA approaches can detect many more species than traditional methods, making them more cost effective. Additionally, for many applications (e.g., offshore, arctic, or remote sampling) traditional methods are infeasible or are too costly, and so many environmentally sensitive areas are monitored infrequently or not at all. Environmental genomics approaches, such as eDNA metabarcoding, give rapid results, necessary for rapid response, and these tools are essential right now, given persistent threats to global biodiversity, including human-caused global change.

The scalability of DNA metabarcoding is a key advantage over traditional approaches.

Relatively low (and decreasing) sample collecting, sample processing, and archival costs mean that DNA metabarcoding will reveal opportunities for larger scale spatial and temporal sampling. This will translate to more frequent monitoring intervals, meaning that managers, decision makers, and policy makers will be afforded more rapid response times for dealing with environmental problems. Already, DNA metabarcoding is leading to efforts to massively scale-up spatial and temporal coverage of bioassessment programs. Another element of DNA that enables environmental genomics to scale up is its ability to easily be archived and stored indefinitely, which is a key advantage over traditional approaches that require large spaces (e.g., in museums) and many people to curate and upkeep these taxonomic libraries. Archived DNA samples take up relatively little space and allow researchers to go back to samples for secondary analyses later, as genomic methods improve and change.

Need for high-capacity environmental genomics labs.

A key barrier to adoption of genomic approaches is the lack of commercial labs with capacity to accommodate national or international DNA metabarcoding biomonitoring campaigns. Though the cost of environmental genomics sampling and sample processing remains high, improved coordination among stakeholders, the advent of automated samplers, and rapidly decreasing costs of sequencing suggest that the cost of conducting environmental genomics will not long be a limiting factor.

SESSION SUMMARIES

DAY 1: WEDNESDAY, JUNE 24TH, 2020

KEYNOTE TALK: BRINGING ECOGENOMIC SURVEILLANCE AND MONITORING DATA TO END-USERS (JOHN DARLING, UNITED STATES ENVIRONMENTAL PROTECTION AGENCY)

Keynote speaker John Darling set the IWEG theme of "Generating Engagement with Genomic Data" by tackling the issue of how to bring environmental genomics tools and data to end-users ranging from citizen scientists to high-level government regulators and decision makers. A central theme of the talk was that not only is it the duty of scientists to bring science to end-users, but it is essential to recognize that good science alone will not suffice; rather it depends on the relationships among the researchers, technical experts, decision-makers, and other stakeholders. Darling suggested four virtues to be esteemed and followed by scientists: honesty (e.g., being honest about what your approach can and cannot do), humility (e.g., recognizing that your view is not the only correct view), empathy (e.g., listening to the views of others, that they might be receptive to your own), and patience (e.g., exhibiting patience toward methodological conservativism or policy change). Ascribing to these virtues provides a way of creating inroads to end-users and helps overcome challenges for the transition to ecogenomic tools. Darling concluded by recalling an observation from the past, one that rings as true as ever, especially for emerging environmental genomics tools: "Every revolutionary idea seems to evoke three stages of reaction. They may be summed up by the phrases: 1) It's completely impossible. 2) It's possible, but not worth doing. 3) I said it was a good idea all along" (Arthur C. Clarke).

SESSION I

The first presentation session featured talks by Peter Lawless (Top of the South Marine Biodiversity Partnership, New Zealand), Mauro de Rebelo (Federal University of Rio de Janeiro), Jordan Angle (Exxon Mobil), Francisco Chavez (Monterey Bay Aquarium Research Institute), and Beverly McClenaghan and Hoda Rajabi (Centre for Environmental Genomics Applications, eDNAtec). The talks highlighted several eDNA metabarcoding studies in marine and freshwater systems. Three of these talks explored ways in which environmental genomics tools, including DNA metabarcoding, are being used in large government and industrial projects to enhance ongoing bioassessment efforts and improve detection of environmental disturbances, including both pollutants and invasive species. While these projects, which are being conducted in coastal New Zealand, Guyana, and California, all provide compelling case studies for the power and utility of eDNA methods for environmental monitoring and surveillance, each speaker emphasized the need for further development of these genomic approaches. The efficient implementation of genomics methods using eDNA requires data standardization and better reference databases, which could facilitate partnerships to improve adoption of genomics approaches. Further, eDNA metabarcoding can be leveraged to create networking opportunities with multiple stakeholders. For example, metabarcoding can be coupled with AUVs for onboard sampling and sequencing to facilitate measuring the "complete" ocean food web at the same time, enabling large-scale datasets tracking commercially and socially relevant species (e.g., endangered or threatened species) at reduced costs; the large-scale, species-rich datasets collected from these efforts will be of interest to a wide range of stakeholders interested in a diverse array of questions. The two remaining talks of this session highlighted methodological and technological challenges and advances in DNA metabarcoding, including how many samples or replicates are needed to detect species in a system, how to sample in remote locations with limited capacity for sample freezing, and how metabarcoding assays can be more sensitive than qPCR assays.

SESSION II

The second presentation session featured talks by Susanna Theroux (Southern California Coastal Water Research Project), Willie Duncan (Scottish Environmental Protection Agency), Shauna Baillie (Fisheries and Oceans Canada), Jeffrey Morisette (U.S. Department of the Interior), and a joint talk by Marc Skinner, Jeffrey Pollock, and Jordan Angle (Stantec Consulting Ltd. / Chevron / ExxonMobil). This session focused on ways of promoting the use of environmental genomics tools to diverse stakeholders, including the government, regulatory agencies, and industry partners. Several of the talks explored mechanisms to secure buy-in of environmental genomics approaches by government and regulatory agencies, including adoption of statewide standard operating procedures, collaboration across agencies, understanding the strengths and limitations of genomic data, adapting existing monitoring methods for use with genomic data, and turning genomic data into actionable information. An overarching, essential part of buy-in is trust: both trusting your evidence and demonstrating your methods to various stakeholders in ways that are clear and transparent. One of the ways of securing this trust is making data open and accessible. Several talks emphasized the necessity to develop standards and highlighted considerations for national adoption, efforts that would build trust among stakeholders and facilitate buy-in of genomic methods. If eDNA sampling approaches are to be applied more broadly at the national level, governments need a series of accredited national standards. Finally, speakers emphasized the need for robust research and development for broader, industry-wide uses of eDNA—an area currently lacking in the oil and gas industry—that will also be facilitated by the development of standardized protocols.

DAY 1 DISCUSSION

For the first discussion session, panelists Willie Duncan, John Darling, Jordan Angle, Jeff Morisette, Susanna Theroux, and Mehrdad Hajibabaei were asked four key questions related to environmental genomics, the information these tools produce, and how best to communicate the associated results.

Question 1: What application do you prioritize for adoption of environmental genomics?

Panelists emphasized that environmental genomics applications need to be prioritized to fulfill regulator requirements for bioassessment. Specific applications for these bioassessments include understanding how biodiversity responds to impacts, anticipating and predicting how these impacts influence biodiversity and associated processes, indicator measurement and development, and assessing biological condition. Many of the panelists emphasized that applications need to become more intuitive, accessible, and standardized. Techniques like PCR and qPCR are appealing because these methods have already been developed for detection of many species of regulatory interest, while more development and standardization may be required before widespread adoption of other environmental genomics applications, like metabarcoding. Despite the need for further development of many genomic applications, these approaches have exciting possibilities for regulation and enforcement because DNA metabarcoding and genomics approaches will provide more robust datasets across a wide range of taxa rather than just a few indicator taxa, and this will likely make these approaches more robust, especially as novel genomic indicators are developed. With sequencing technology moving so fast, one thing is clear: this is an exciting time for all stakeholders to be involved in the field of environmental genomics.

Question 2: The complexity of metabarcoding data makes it difficult to communicate with non-experts, including people in decision-making position and the general public. Do you have any examples of cases where you think you were able to communicate to non-experts effectively?

Regarding the question about dealing with the complexity of genomic data, panelists challenged the premise of this question by asking if DNA metabarcoding data are really more complex than other ecological data and if end-users are really that much more versed in other, non-genomic ecological metrics and the science behind them. Scientists need to keep communication simple with both traditional and genomic approaches, making sure to tell a simple story that managers and policy makers can follow. Panelists also acknowledged the complexity of this issue: all data are inherently complex and require experts to distill and communicate emergent patterns, but genomic data can be particularly complex, requiring many high-quality personnel—including laboratory geneticists, bioinformaticians, and ecologists—to generate, curate, analyze, and interpret. Distilling genomic information down to a clear form is essential to help people understand the premise of one's work and, thus, its potential impact. Finally, panelists discussed what the advent of "big data" in other areas means for the general acceptance of taxonomy-free metrics and the big data associated with DNA metabarcoding. We need to understand the environmental conditions under which taxonomy-free approaches work for environmental monitoring; in this respect, it doesn't really matter if an approach uses taxonomy or is taxonomy-free.

Question 3: Most studies still show discrepancies between morphological and metabarcoding surveys of the same sites, including failure to detect organisms that are known to be present. As a result, many researchers are proposing we use metabarcoding to complement traditional surveys, not as a replacement.

What do you think causes this discrepancy? Does this negate many of the benefits of metabarcoding?

Panelists challenged the notion of DNA metabarcoding data being complimentary to traditionally derived data by questioning what we mean by "complimentary" and by stating that it is a fallacy to think we need to verify DNA metabarcoding results to morphological results. Metabarcoding is simply a different approach, with many inherent advantages. For instance, metabarcoding can be conducted with much more coverage and frequency than most traditional methods; consequently, with decreasing costs for DNA metabarcoding, this approach will enable us to process a lot more samples, which is a promising prospect. Panelists also pointed out that it will be a problem down the road to ask people to do biomonitoring using both traditional and genomics approaches but acknowledged that it is valuable to rely on the morphology data we still have. Discrepancies between DNA and morphological data are only important when we have a small number of taxa: when we have whole community assessments, it probably does not matter because both approaches will likely have enough discriminator power, which is why the future of biomonitoring is likely the cheaper, faster DNA metabarcoding approach. It is also good to think about what organism or organism group we are considering when comparing DNA metabarcoding with traditional methods. For example, DNA provides a lot more biodiversity that is not normally accepted by regulators. Microbial taxa, for instance, are easier to obtain abundance data compared to large taxa that might only have trace amounts of DNA in a sample. Further, just because an approach is taxonomy-free doesn't mean we know nothing about the ecology of the associated information. Consequently, it is important to consider perspectives from the microbial world when moving biomonitoring forward.

Question 4: How do we use environmental genomics to evaluate and understand temporal change? Does scale matter? Can we evaluate 'rapid change'? How can we address temporal biases in our analyses or study designs?

Panelists were quite unified in their responses and discussion related to how we can use environmental genomics to understand temporal change. In general, most panelists agreed that one of the key advantages of DNA metabarcoding is that it provides a rapid, cost-effective way to easily get more samples, and so this approach is well suited for temporal studies and scaling up. Metabarcoding is already improving the temporal frequency with which some panelists can sample certain organism groups (e.g., algae). An added advantage of DNA metabarcoding, which could greatly enhance researchers' ability to tackle issues related to scaling up temporally and spatially, is that it can be archived. Archiving requires data standards, data infrastructure, and storage facilities, but this represents a huge advantage over traditional sampling methods because it allows repeated analyses using various markers or analytical tools, and should lead to more large-scale studies and biodiversity monitoring efforts moving into the future. Additionally, panelists pointed out that DNA metabarcoding allows paleogenomics researchers to look back in time, as many successful attempts have been made to recover DNA from sediments or ice cores to reconstruct pre-industrial environmental states. Regardless of how much spatial and temporal resolution can be achieved with DNA metabarcoding and other environmental genomic approaches, several panelists pointed out that having other environmental data (e.g., temperature, water chemistry, etc.) are essential for providing proper context for temporal biodiversity data. For example, morphological data along with genomic data helped one research group understand the rich DNA metabarcoding time series data they collected. Finally, panelists pointed out that the future will be even brighter when we start to move towards transcriptomics and metagenomics.

DAY 2: THURSDAY, JUNE 25TH, 2020

SESSION III

The third presentation session included talks by Michael Bunce (New Zealand Environmental Protection Agency), Carolina Berdugo (Imperial Oil), Jan Pawlowski (University of Geneva / ID-Gene), Jon Thomassen Hestetun (NORC Environment), and David Coté (Fisheries and Oceans Canada; DFO). Talks from this session explored the use of eDNA metabarcoding for a variety of applications ranging from assessing habitat reclamation to biodiversity assessments of offshore and deep-sea habitats.

Bunce commenced the session with a conceptual talk about how eDNA can be used as a community engagement tool, highlighting the citizen science project Open Waters Aotearoa, which aims to empower citizens with resources to connect with their local ecosystems in New Zealand. The remaining four talks in this session outlined case studies using eDNA metabarcoding. Carolina Berdugo demonstrated how environmental genomics can assess reclamation success of lands in Cold Lake oil and gas operations in northeastern Alberta. Jon Thomassen Hestetun and Jan Pawlowski both examined the utility and challenges of applying high-throughput metabarcoding to asses eukaryotic diversity in benthic ocean sediments near offshore oil and gas platforms. David Coté concluded the session with a talk on the use of eDNA metabarcoding in ISECOLD (Integrated Studies & Ecosystem Characterization of the Labrador Sea Deep Ocean), a collaborative project exploring fish biodiversity and ecological processes in the deep waters of the Labrador Sea. Coté showcased promising eDNA results using DNA metabarcoding and stressed the need for integrative approaches. Other sampling tools provide complementary

data (e.g., abundance, biomass, fish condition) that can be merged with metabarcoding information for novel bioassessment tools, as indicated by some preliminary work with heuristic food web models.

Overall, talks from this session illustrated the power of eDNA metabarcoding as a bioassessment tool that—when used to supplement existing bioassessments—can provide novel ways to address challenging biodiversity questions. These talks underscored that the future of biodiversity assessment will include rapid measurement of broad swaths of the biodiversity in a variety of ecosystems, allowing practitioners to rapidly detect environmental change from human activities, potentially leading to environmentally responsible decisions that protect ecosystems and the services they provide.

SESSION IV

The final presentation session featured talks by Kahlil Lawless (Illumina), Cathryn Abbott (Fisheries and Oceans Canada; DFO), Michael Weise (U.S. Office of Naval Research), and Zacchaeus Compson (Centre for Environmental Genomics Applications, eDNAtec). These talks focused on the promising future of environmental genomics.

Kahlil Lawless commenced the session with a forward-thinking talk outlining interdisciplinary lessons from environmental genomics tools related to eDNA testing. The talk explored both shotgun sequencing and DNA metabarcoding, illustrating that while both approaches recover incredible amounts of biodiversity, they are very different. Lawless stressed the need for careful validation of different approaches by comparing differences in results from identical samples to uncover biases in these methods. Lawless also pointed out that innovative techniques like “Loop Genomics” enable generation of synthetic long reads from short read data (one of the limitations of metabarcoding data, which typically uses reads from 150-300bp) without compromising base-calling accuracy. The next two talks of this session outlined progress made in two national environmental DNA programs: 1) DFO’s eDNA program for managing Aquatic Invasive Species and Species at Risk, and 2) U.S. Office of Naval Research’s eDNA program. These talks underscored challenges faced by governments for integrating eDNA into management decision-making processes, including lack of fundamental research in environmental genomics approaches, incomplete reference sequence databases, a lack of guidelines and reporting standards, and poor implementation and operational research. Key mechanisms for advancing eDNA research include operational procedures and efforts for institutional—and eventually national and global level—data standards, further experiments to develop and standardize these methods, and technological developments to scale up DNA metabarcoding to global levels using automation (e.g., autonomous water collection and filtration units or portable DNA units that can extract and sequence DNA in the field, such as Biomeme and MinIon).

In the final talk of the session, Zacchaeus Compson outlined a systematic literature review performed by the CEGA team and highlighted three case studies that show exciting new directions for environmental genomics. The systematic literature search revealed that while DNA metabarcoding has achieved near global coverage in sampling over the last ten years, the promise of “total biodiversity assessment” using multiple primer sets (to capture most of the biodiversity in a sample) has not yet been widely realized. However, with the advent of ultra high throughput sequencing, like Illumina’s NovaSeq platform, this premise could quickly be realized in various applications. Finally, coupling DNA metabarcoding data with deeper sequencing tools enables novel applications, ranging from abundance quantification from DNA samples, phylogenetics, population genetics, food webs, and functional gene analysis.

BEST PRACTICES FOR DATA VISUALIZATION: INTERACTIVE SESSION SUMMARY

IWEG 2020 included an interactive session entitled, “Determining Best Practices for Data Visualization”. Using a series of surveys and focused discussion groups, participants identified important criteria for effective environmental genomics data visualizations and assessed common data visualizations using these criteria. The 37 participants in the live session represented academia (8 %), industry (13%), service (38%), and government (41%). Here, we integrate these perspectives across different sectors and job roles to create recommendations for best practices in environmental genomics data visualization.

Environmental genomics data have many properties that make results difficult to visualize and communicate to users who are not familiar with the technique. Results can be presented with or without taxonomy. New concepts or terminology may be introduced (e.g., amplicon sequence variant, ASV; operational taxonomic unit, OTU). Data are often hierarchical and non-linear, with several levels of replication (e.g., sampling replicates, PCR replicates) that may have overlapping results (e.g., multiple DNA markers). Several types of data visualizations have emerged in the literature as routine methods of presenting these results to address different research or management questions. We grouped these questions into three broad themes: targeted species detection, biodiversity inventory (or alpha diversity), and comparison of community composition (or beta diversity).

Participants evaluated the usefulness (1 – not useful at all, 5 – very useful) of common visualization types for the three research themes in environmental genomics (Box 1). Not all visualizations were equally effective in addressing each of these research or management questions. Maps and box plots of DNA quantity were voted the most useful visualizations for targeted detection where the goal is determining where target species are present in a sampling area (mean rating: 3.48). Detection summary tables were considered the least useful for targeted detection (2.70). Figures incorporating phylogenetic trees rated the highest for biodiversity inventories presenting a summary of the taxa detected (4.00); conversely, treemaps were considered the least effective visualization for an inventory (2.45). Ordinations were identified as the preferred visualization to compare community composition between groups of samples (4.2), while pie charts were deemed the least useful visualization to compare communities (2.92). On average, across all research themes figures had a higher rating (3.43, n = 16) compared to tables (2.84, n = 3). Fewer tables were included in the survey, which was due to the greater variety of figure types present in the literature compared to tables. Examples of each visualization type presented to survey participants can be found in Appendix B.

Many factors can enhance or detract from a visualization. Not only is the choice of visualization type important, but how the visualization is presented can have implications for communication to end-users. We gathered input from participants to identify the visualization criteria that are most important when communicating environmental genomics data. By grouping participants’ responses into categories, we identified six criteria that can be used to evaluate the quality of a data visualization (Box 2). Simplicity was by far the most frequently mentioned category (n = 78). This is perhaps unsurprising given that environmental genomics can generate vast amounts of information that need to be conveyed in easy to interpret plots. It also points to the perceived complexity of environmental genomics data, making simplicity an important, yet challenging, criteria. All six categories should be used to guide the generation of visualizations and improve data presentation.

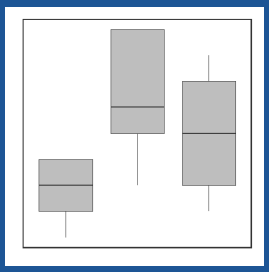
During the focused discussion session, a series of example figures (Appendix B) were presented to participants to prompt discussion. Several recommended best practices emerged for each criterion (Box 2).

Box 1. Average rating (1 – not useful at all, 5 – very useful) of visualization types for three research themes.

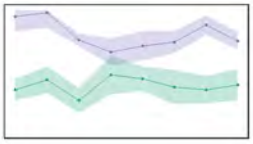
Theme 1

Targeted Detection

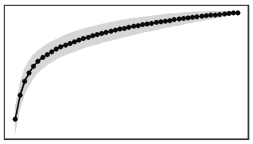
3.48 Detection Maps



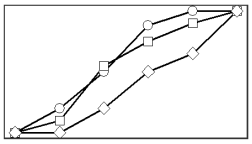
3.48 DNA Quantity Boxplots



3.30 Occupancy & Detection Plots



3.27 Accumulation Curves



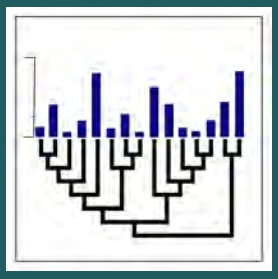
3.20 Assay Sensitivity Plot

Site	# Samples	# Detections
A	16	0
B	13	11
C	18	14
D	15	14

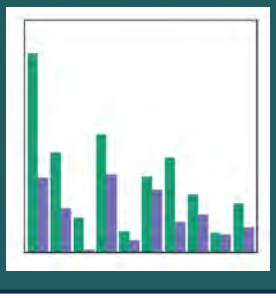
2.70 Detection Summary Tables

Theme 2

Biodiversity Inventory



4.00 Phylogenetic Trees

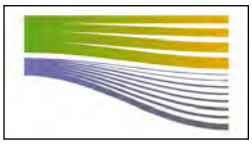


3.65 Bar Plots

3.48 Pie Charts

Species	Sites			
	1	2	3	4
Amia calva		X		
Esox americanus	X	X		
Lepomis gibbosus		X	X	X
Micropterus dolomieu	X			
Perca flavescens	X			X
Umbra limi	X			
Total Number of Species	4	3	1	2

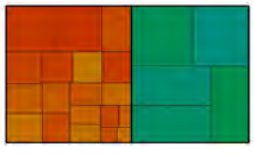
3.33 Species by Site Tables



3.00 Sankey Diagrams

Kingdom	Phylum	Family	Genus	Species
Metazoa	Arthropoda	29	41	14
	Chordata	31	48	32
	Echinodermata	3	3	2
	Mollusca	17	22	7
Fungi	Ascomycota	1	1	
	Basidiomycota	1	1	

2.50 Biodiversity Summary Table

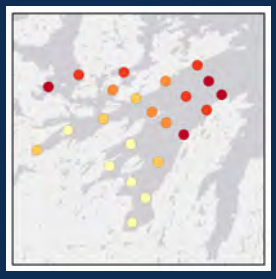


2.45 Treemaps

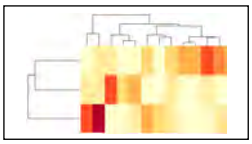
Theme 3

Community Composition

4.20 Ordination Plots



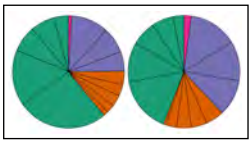
3.85 Maps



3.62 Heatmaps

3.50 Stacked Bar Plot

3.40 Venn Diagrams



2.92 Pie Charts

Many of the recommendations are not specific to environmental genomics data but are particularly important when visualizing large multivariate datasets that contain information that may be unfamiliar to users and that encompass many decision-making steps. These recommendations are intended to improve data visualizations and facilitate communication between environmental genomics practitioners and end-users.

SIMPLICITY. Many visualizations have too much information, making it difficult to see and interpret the data. Additionally, unnecessary information is included in captions, adding to the difficulty in interpretation. To simplify data visualizations, remove information not directly related to the main point of the visualization. This includes extraneous information within the figure or table and within the caption. Breaking down complex figures into multiple, simpler figures makes interpretation easier by including less data in each individual figure. Alternatively, including larger figures and tables that contain all the data in appendices—and presenting a simplified version in the main text—facilities interpretation while ensuring that all data remain discoverable. The phylogenetic tree used to visualize biodiversity inventory data was identified as a good example of a figure that communicated a lot of data in a simplified way. Simpler visuals help support the main conclusions without creating distractions.

CLARITY. Poor formatting and inappropriate stylistic choices often lead to a lack of clarity in visualizations. These mistakes are relatively easy to avoid and greatly improve the appeal and understanding of a visualization. To ensure clarity, include a title or caption to make the main point of the visualization obvious. This guides viewers through a visualization, which is especially important for technical reports compared to academic papers. Labelling should always be complete and easy to understand, with labels included on the figure as opposed to in the caption when possible. Visualizations should include all the information needed to interpret and understand the results they present, either in the figure and/or the caption. When comparing data in different categories, avoid using too many categories and choose an appropriate method for differentiating categories from one another. Avoid using multiple methods for distinguishing the same categorical variable (e.g., size and colour of points in an ordination plot), as this is redundant information and overcomplicates the figure. If using colour, choose your colour palette to be visually appealing and to make the trends in your data easily visible. Do not use colour when it is not needed, and consider using a palette that is visible to colour-blind audience members when possible. Appealing visualizations that are easy to understand will have a greater impact on your audience.

CONTEXT. While making visualizations appealing is beneficial, it is equally important to ensure the data being presented are contextualized with the necessary information for interpretation. Formatting choices or oversimplification can lead to misleading representations of the data and influence viewers' conclusions. When presenting data for multiple locations or sites, the geographic distance between sampling sites can be an important factor in interpreting the data. Whenever possible, plot data on a map to help viewers qualitatively account for spatial variation, which may be overlooked in other types of figures and tables. Include error bars in figures and include a description of information presented in a plot in the caption, when necessary (e.g., the statistical values represented in a boxplot). Define acronyms and specific terminology used in the figure. An honest and accurate presentation with all the necessary data is essential to facilitate viewers making appropriate conclusions.

EASE OF INTERPRETATION. Many visualizations are not easy to interpret and take some effort for non-technical viewers to understand without specialized training or detailed explanations (e.g., nMDS plot for community composition) compared to other visualizations that are more intuitive for broad audiences (e.g.,

Box 2. Criteria for effective data visualizations, including the number of mentions during workshop and recommended best practices for each criterion.

Simplicity
78

Whittle it down to the essentials
- Remove extraneous information from caption and visualization.
- Break down complex figures into multiple smaller ones.
- Move large, complex figures to appendix and present simplified version in main text to facilitate interpretation.

Clarity
56

Format to improve understanding
- Include a title or caption title to highlight the main point.
- Labeling should be complete and easy to understand.
- All information needed for interpretation should be included.
- When using categories, don't use too many and choose an appropriate methods to differentiate categories.
- If using colour, choose a visually appealing palette that highlights trends in the data.

Context
54

Present all necessary information
- Include relevant statistical values and a description of information presented in each plot.
- Account for geographic distance between sites when presenting site specific results (i.e. use a map).
- Define acronyms and terminology used in the figure.

Audience
38

Tailor to your viewers
- Always consider the users and viewers when determining how to visualize data.
- Visualization need to be simpler, clearer and easier to interpret in reports that will be read by managers and non-scientists.
- Tailor your visualizing to your audience based on the format of presentation (e.g. written report vs. oral presentation).
- Use colour-blind friendly palettes to make your visualization readable by a broader audience.

Ease of Interpretation
36

Make it easy for viewers to draw conclusions
- Simplifying and clarifying a visualization will improve the ease of interpretation.
- Tables are more time consuming to interpret than figures.
- Ensure consistency between multi-plot figures to avoid misleading viewers.
- Avoid using colour when it's unnecessary.
- Interactive plots may facilitate interpretation.
- For a broad audience, choose figures that are easier to understand (e.g. maps).

Relevance
28

Link visualizations with objectives and conclusions
- Visualizations should demonstrate the main points of the paper/report.
- Move to the appendix or omit visualizations that show results that are not relevant to the main point (e.g. methodological figures).

maps, stacked bar plots). Ease of interpretation is also highly affected by the simplicity and clarity of the visualization. Simplifying a plot, by including fewer categories for example, and clarifying a plot with improved formatting will make the interpretation much more straightforward for viewers. Ensure consistency between plots in multi-paneled figures (e.g., use the same axis scales and same colour scheme for all panels). Avoid using colour when it is unnecessary. The inclusion of meaningless colour can imply a relationship when there is none, which may be misleading to users. Tables are more time consuming to interpret than figures, where trends and patterns in the data are often more obvious. Interpretation may be facilitated using interactive plots. With complicated datasets containing many variables, interactive plots can allow users to explore biodiversity data at the level of detail with which they are comfortable or familiar. Visualizations are more effective at illustrating their main points when interpretation is easy and intuitive.

RELEVANCE.. The link between research questions, the data presented, and the conclusions should be evident for effective visualizations. However, some plots do not lead to clear conclusions (e.g., Sankey diagrams and summary biodiversity tables for biodiversity inventories). Visualizations should be chosen to help a report or paper meet the goals of the project. Consider omitting visualizations that do not demonstrate the main points of a report or paper. Visualizations that are not directly relevant to the conclusions but provide important context, such as methodological figures, should be included in an appendix. For technical reports, limiting the visualizations in the main text to the most central points is especially important. Clear, concise conclusions supported by relevant visualizations makes a powerful message.

AUDIENCE. The audience viewing the visualizations is a key consideration in determining how to visualize your data and implementing the previously discussed recommendations. Visualizations will be received differently depending on the expertise, experience, and goals of the viewer. The distinction between audiences of academic researchers compared to audiences of managers and policymakers was mentioned most often during the focused discussion groups. The most effective visualizations will differ between academic papers and technical reports. Visualizations need to be simpler, clearer, and easier to interpret in technical reports to facilitate decision-making by managers and policymakers, who have limited time and may lack analytical expertise. As experts, scientists need to interpret and communicate the data to enable stakeholders to act upon the results. Every visualization in an applied technical report should be linked to a point or conclusion. Audiences are not only defined by their sector or job role. The format of your data presentation (e.g., a written document, an oral presentation) should also determine how your audience receives your data and, therefore, the best ways to visualize your data. Best practices for visualizing data in an oral presentation may be different than a written document. Colour-blind viewers are also an audience that should be accommodated by using colour-blind friendly palettes. While these are specific examples of viewers and users that need to be considered, this is not a complete list of the different types of audiences exposed to environmental genomics data. No matter who your audience is, it is important to communicate effectively by tailoring your visualizations to your audience.

These recommendations for visualizing environmental genomics data were established by integrating perspectives from a range of environmental genomics stakeholders in attendance at IWEG 2020. This is not a comprehensive list of best practices, but it presents those addressing some of the challenges when visualizing eDNA-based data. Applying these best practices and using visualization types that are most effective will facilitate communication among users, practitioners, and managers, and will advance the use of environmental genomics across these sectors.

CLOSING REMARKS

Mehrdad Hajibabaei (University of Guelph and Centre for Environmental Genomics Applications. eDNAtec), concluded the formal IWEG sessions with a keynote talk summarizing important workshop themes and providing a vision for how to progress environmental genomics adoption and advancement. An overarching theme framing Hajibabaei's talk was that the field of environmental genomics is now at the stage where we can let the data speak. Hajibabaei stressed that there is a trove of accessible genomic and metabarcoding data that we can work with and use to develop, validate, and promote genomic tools to managers, decision makers, and policy makers. To highlight this point, Hajibabaei showcased the Biomonitoring 2.0 program, which started over a decade ago in collaboration with Environment and Climate Change Canada and various partners, and has now generated one of the largest, long-term DNA metabarcoding datasets in existence. Long-term, large-scale projects empower practitioners with data needed to work out methods and demonstrate to managers that a new technology or sampling approach works. Given the immense success of the Biomonitoring 2.0 program, Hajibabaei emphasized the need for a marine equivalent of Biomonitoring 2.0, which would provide essential information about offshore oil exploration efforts and other ocean related activities. Hajibabaei also highlighted the Sequencing the Rivers for Environmental Assessment and Monitoring (STREAM) project, which aims to assess Canada's watersheds and address gaps in biomonitoring data. STREAM involves collaborators from NGOs (e.g., WWF and Living Lakes Canada) and regulatory agencies (e.g., ECCC) and has built a network for community-based monitoring including several indigenous groups across Canada. This model allows sampling a vast geographic area encompassing most of Canada. Given the urgent need for rapid bioassessments, STREAM has used a model of delivering rapid results to decision makers in the form reports that are circulated to community groups to assist in local management and decision making. This and other ideas explored in the STREAM project will help overcome challenges resource managers face, accelerating adoption of these new tools more broadly.

Hajibabaei concluded his talk by emphasizing the need for specialized laboratories to meet the surging demand for genomic data and facilitate scaling up bioassessments to regional and global levels. Hajibabaei highlighted the Centre for Environmental Genomics Applications (CEGA), a global centre of excellence in ecological genomics. Currently, CEGA has standard operating procedures that have been hardened and tested in various field conditions and remote locations, and CEGA is working with other stakeholders to overcome the challenges of environmental genomics and provide additional applications. Finally, Hajibabaei celebrated the CEGA-organized International Workshop on Environmental Genomics (IWEG), which showcases the promise of environmental genomics by bringing together stakeholders across sectors, featuring a balanced representation of sectors, including academia, government, NGOs, and industry. The interdisciplinary nature of IWEG facilitates addressing real-world problems in an inclusive, responsible, proactive way.

DAY 2 DISCUSSION

For the second discussion session, panelists Dave Coté (DFO), Erik Pilgrim (US EPA), Sherry Walker (DFO), Sunniva Aagaard (Norwegian EPA), Carolina Berdugo (Imperial Oil), and Zacchaeus Compson (CEGA) were asked the following four questions related to the future of environmental genomics, including how it can be used to in novel ways and to deal with pressing environmental issues.

Question 1: What do you see as the greatest barriers to widespread adoption of metabarcoding by regulatory agencies?

Panelists' views varied greatly regarding the barriers to widespread adoption of DNA metabarcoding by regulatory agencies, emphasizing the complexities of its adoption despite its clear advantages over traditional

methods. A key barrier to adoption of metabarcoding by regulatory agencies was the need for standardization of methodologies and protocols. For instance, we need a more widespread understanding of DNA metabarcoding and training in the use of eDNA. This includes training end-users about what DNA metabarcoding data means, using targeted communication. Further, the field needs to move toward method and protocol standardization, especially for early steps in the DNA metabarcoding pipeline, like sampling and DNA extraction. Eventually, a program needs to be developed for the accreditation of environmental genomics labs to ensure consistent reporting, repeatability of results, and reproducibility of studies. Along with this, panelists identified the need for more case studies using DNA metabarcoding, which will help normalize it as a method as well as provide more context in novel systems. Another major barrier to metabarcoding adoption by regulators is overcoming the inertia from existing monitoring programs: this methodological conservativism slows down innovation and stifles adoption of newer, superior methods. One way of bridging the gap between established methods and newer methods, like metabarcoding, is to conduct more concurrent studies where traditional methods are run side-by-side with metabarcoding surveys, enabling researchers to demonstrate that metabarcoding results are both comparable and reliable. Careful consideration, however, should be given to these comparisons, as genomic methods are fundamentally different from traditional bioassessment methods (see "Day 1 Discussion" for details). A final barrier to the adoption of metabarcoding by regulatory agencies is the issue of financing metabarcoding campaigns, which for many agencies requires supplemental funding beyond what is already allotted for legislatively mandated biomonitoring using established methods. Thus, discussions about how to prioritize available biomonitoring funds and strategize the most efficient way forward are essential. For many agencies, funds are already committed to established methods, and leftover funds for research and development are scant. Fortunately, advances in sequencing technology coupled with massive efforts to scale up next-generation sequencing efforts will mean that, long-term, sampling costs for DNA metabarcoding will come down, making these agency-sponsored environmental genomics campaigns more plausible.

> Question 2: Measurements of organism abundance are a critical component of many biodiversity surveys, but current metabarcoding assays only return sequence abundance, which cannot be directly interpreted as organism abundance or biomass.
>
> Are there any technological advancements on the horizon that could overcome this limitation? Are there analytical techniques that reduce or eliminate the need for relative abundances in some cases?

Discussion around technological and analytical ways of inferring taxonomic abundances from DNA metabarcoding data was dynamic and impassioned. On the one hand, panelists noted the constant progress in this area of study, with the potential to eventually reach a point where routine quantitative assessments can be done with metabarcoding surveys. This will first require a better understanding of the factors affecting relative read abundances from metabarcoding samples, particularly as it relates to PCR steps. This may require standardizations to be done at the taxonomic level, using, for example, mock communities. Analytically, there are already some exciting ways that biodiversity quantification can be done using occupancy modeling, which can utilize repeated sampling of presence-absence data to estimate probability of occupancy, which likely correlates with abundance. While this will require more extensive sampling, which can be costly, this will be less of a concern as the costs of sequencing continue to decline. On the other hand, many panelists pointed out that abundance data are not necessarily essential for many bioassessments. Most commonly, abundance is used to add refinement for analysis of datasets with few taxa (e.g., ordinations with a small number of taxa); however, with greater taxonomic breadth and resolution, which you generally get with DNA metabarcoding surveys, you probably do not need abundance

information. Further, one panelist pointed out that we do not need to "take over the world" with metabarcoding: there is room for other techniques for different applications (e.g., studies that require abundance), and eDNA metabarcoding does not have to answer all of our questions or solve all of our problems. Rather, by viewing metabarcoding as a complementary tool, coupling this approach with other techniques can lead to exciting synergies, including food web assessment and functional gene analysis.

Question 3: What is the most effective way to generate and interpret data when lacking a reference library? How do you communicate results that do not have taxonomy?

The issue of lacking reference sequence databases is a critical one, especially in marine systems, but panelists provided insight about what types of assessments can be done without taxonomy. Currently, there is enough information in reference databases to sort organisms to at least higher taxonomic levels, and groups can be assigned that are recognizable and usable, even if that is not to the species level. For most bioassessment applications, it is often sufficient to go to the family level, or to the lowest possible taxonomic level, as is the case for most benthic macroinvertebrate biomonitoring programs. In many ways, the lack of coverage of reference databases means that there still is much work to be done to link DNA sequences to their associated taxa. As one participant pointed out, this could mean that the eDNA field has the ability to become one of the best advocates for "classical taxonomy," which should assuage the fear of many that DNA metabarcoding will trigger the waning of classical taxonomy; rather, the ESVs that are bioindicators will have more meaning if we can better understand the biology and morphology of the organisms from which they come. Still, there remains hope that reference databases will be rapidly fleshed out and that taxonomy will be more certainly assigned. Some scientists in Canada's DFO, for instance, are now using RDP classifiers, either independently or alongside BLAST, to identify sequences. Further, others have been using MEGAN's "lowest common ancestor" algorithm to interpret BLAST hits. All of this means that while we should not be limited to taxonomy-based results and analyses, we might soon benefit from this rich taxonomic information as well, should we need it.

Question 4: In the face of the novel coronavirus pandemic, how can we enable the continuation of environmental and ecological surveys on vessels or at remote sites? How can environmental genomics help streamline ongoing monitoring projects that may be disrupted by the pandemic?

Discussion around how monitoring programs can proceed in the face of the COVID-19 pandemic were illuminating, highlighting how environmental genomics can help streamline ongoing monitoring projects that have been disrupted by the pandemic. It was clear that most of the bioassessment programs that panelists were involved with are still proceeding—albeit in a limited capacity—despite the limitations imposed by the global pandemic. Many of these programs will proceed with reduced crews to ensure social distancing, and some are implementing reduced sampling programs to spend less time in the field. These actions not only allow organizations to be more proactive with respect to COVID-19, but they also provide managers with the opportunity to demonstrate to upper administration how eDNA sampling can reduce the costs of monitoring efforts. In this regard, eDNA sampling methods can provide timely answers to logistical sampling problems brought on by the pandemic, ensuring programs persist that would otherwise have sampling gaps, especially for programs using labor-intensive traditional sampling methods. For example, programs that require a safety analysis might not get the greenlight without innovative solutions to sampling; environmental genomics sampling, which can be automated and used on autonomous systems, can be offered as a clear solution to this problem. Despite the

promising ability to scale up sampling efforts using environmental genomics approaches, most sequencing centers are backed up with coronavirus work, which is causing a bottleneck for some research programs. Still, during these trying times, environmental genomics provides a rare ray of hope, and practitioners that can leverage the problems of the pandemic to highlight the power of environmental genomic tools for bioassessment may fare well in the long run.

GENERATING ENGAGEMENT

Prior to IWEG 2020, each presenter was sent a list of the following three questions, with the directive to address them in their presentations:

1) Who are your stakeholders or your primary audience?

2) What methods do you use to effectively communicate environmental genomics results and why?

3) How do you address the limitations of using an environmental genomic approach to biodiversity monitoring?

The intent of these questions was to emphasize the theme of IWEG 2020: "Generating Engagement with Genomic Data." A common sentiment from the presenters' discussion around these questions was that environmental genomic stakeholders are diverse and growing, including managers, industry, and even engaged members of the public. Consequently, it is important to know your audience and relate environmental genomics information to what people know. Speakers indicated that this can be done using simple language and concise results, that—if possible—are presented using statistical or graphical techniques that are familiar to a particular audience. It is also important to explain what results mean and their applicability to stakeholders. When possible, it is effective to extrapolate results to real situations to which stakeholders can relate. Presenters noted that case studies are important for creating comfort with new techniques, like DNA metabarcoding. Additionally, emphasizing how genomic tools can be used to supplement existing, operational bioassessment programs can enable these tools to reach a broader range of stakeholders. This can be particularly effective when compelling cases are made demonstrating how genomic tools can be cost effective compared to other monitoring methods, especially when scaled up to regional, national, and global levels. Further, speakers pointed out that buy-in can be enhanced by emphasizing how traditional methods (e.g., abundance data collected with morphological samples) can be merged with genomic data (e.g., gut content metabarcoding, functional gene analysis) to greatly expand the range of assessments possible (e.g., detailed diet studies or food web analysis). Finally, presenters reiterated one of the recurring themes of this workshop: it is imperative that environmental genomics users recognize limitations of these approaches and transparently communicate them to their audience, clearly sharing what can and cannot be done with these approaches. Being up front with stakeholders about these trade-offs will foster trust and open-mindedness with potential users of these exciting new approaches.