Next Generation Biodiversity Monitoring Symposium

University of Cyprus
11th-13th of November 2019

Background for the Discussion sessions (Days 2 and 3)

Session 1 (big group): To reach a consensus with a working GO definition that builds optimally on the diversity of current initiatives for biodiversity surveying and monitoring

High-throughput sequencing is widely used in the characterisation and monitoring of ecosystems, and now holds the prospect for a much deeper understanding of species diversity on Earth. Deep DNA sequencing conducted at particular sites, including so-called Genomic Observatories (GOs), presents great opportunities for detailed characterisation of local biomes, and if combined with other such sites, may generate broader synthesis of biodiversity patterns and the underlying processes. Yet, studies of both local sites and their global integration, to be meaningful, require a certain level of standardisation of data generation and data analysis, including the informed choice of molecular markers and sequencing protocols, and the application of sound bioinformatics and statistical procedures. Ultimately, the choice of protocols and sampling strategies will be driven by the specific questions we are going to ask based on the deep sequencing of local sites.

In that sense, the notion of GOs and similar proposals needs to be distilled and developed further, particularly regarding spatial, temporal and taxonomic dimensions. At one extreme, selected sites could be intensively sampled for genomic sequencing and long-term biomonitoring. Alternatively, they could simply constitute a network of shallow but geographically more extensive sampling points covering the ecoregions on Earth. This session explores the GO concept in the face of what are the most exciting questions that can be addressed with deep-sequencing of local sites, in particular if such data from numerous sites can be connected. Studies might focus on particular taxonomic or ecologically defined groups; they may be based on single sites or multiple sites covering a particular bioregion or even the global level; they may be obtained in a single sampling event or over increasingly longer time periods, possibly even extending into (sub)fossil layers. What is needed is the synthesis of relevant objectives and approaches within a unified network of sampling sites and analytical approaches, that is at the same time (i) as inclusive as possible with regard to ongoing initiatives, and (ii) as useful as possible with regard to the data it generates. This symposium, by bringing together researchers involved in different regional and global initiatives, with different but complementary objectives related to the next generation biodiversity monitoring, will define the goals and methodologies of a genomics-informed biodiversity science. Specifically:

1.1 What objectives can be addressed with a spatially extensive site-based network and what are the arguments for this being the core function of GOs.

1.2 What are the arguments for long-term biomonitoring to also be a function of GOs? What are the objectives across the temporal dimension?

1.3 What existing initiatives meet the criteria of (i) collecting genomic data, at (ii) some community level, from (iii) multiple sites and/or temporal sampling, with (iv) standardised methods?

1.4 What are the objectives of these initiatives and what data are being generated to meet those objectives?

1.5 Which of these objectives are relevant for a unified network of GOs?

1.6 What other objectives are needed for a unified network of GOs?

1.7 What is our definition of a unified network of GOs?

Session 2 (big group): Core data and measures to be provided by GOs and the molecular tools to obtain them

DNA data provide additional levels of information that go beyond conventional biodiversity data, including genetic (haplotype) information that increase the resolution of biodiversity patterns and phylogenetic information providing an evolutionary framework. Additionally, deep sequencing of genomes and transcriptomes provides a functional dimension. High throughput sequencing of local sites can further establish interactions of organisms, including the use of ingested DNA (iDNA), for an ever more complete understanding of whole-ecosystem processes, while sequencing eDNA from the environment, such as lake sediments, will extend the temporal component beyond the immediately observable period. Our aim should be to identify a minimum set of data and measures to meet the objectives identified in the first session, across both spatial and temporal dimensions. This will include the use of site-based sampling for the study of regional/global patterns, and to identify the molecular data required. While single-gene sequencing may be the most efficient (and the only one currently available for large quantities of samples), new long-read genomic methods may become available that go beyond the current (meta)barcoding and mitogenomics, and it is unclear what is their potential, in particular as genome sequences become available for numerous species. Likewise, transcriptomic studies can potentially elucidate functional aspects of an ecosystem and, for example, establish the link to biogeochemical cycles.

This session will discuss the relative importance of different biodiversity measures obtained by high throughput sequencing of local sites, including taxonomic richness, relative abundance and biomass, haplotypic diversity, phylogenetic diversity, functional and interaction diversity. It will also explore the new possibilities of site-based sequencing from using a range of markers and sequencing of increasing portions of the genomes. To what degree can such studies overcome existing problems of estimating genetic diversity, species limits, taxonomic assignment and phylogenetic history? How will deep sequencing at physically well-characterised local sites increase our understanding of what explains diversity, or how it is affected by global change? On the one hand, we want the most detailed data possible, but on the other hand, if that were to involve an overly time consuming or complex set of protocols, the site-based approach would be limited in terms of sites analysed and the possibilities for comparative studies. Thus it will be important to consider how temporal and spatial sampling should be integrated within a network of GOs.

2.1 What molecular data types and biodiversity measures from ongoing initiatives are most relevant to the objectives identified across the spatial and temporal dimensions?

2.2 What level of taxonomic assignment is important? Do we need data that can (with some caveats) be interpreted at the level of species (e.g. mtDNA COI data for animal life), or are higher taxonomic assignments acceptable (e.g. 18S for Eukaryotes), or is some combination of both required?

2.3 Do we need measures of phylogenetic and or functional diversity? Do we need abundance/biomass data? Is haplotype diversity needed? Do we need measures of interaction diversity based on eDNA and iDNA?

2.4 What other types of molecular data are necessary? Do we need deeper sequencing and longer genomic fragments?

2.5 What should the temporal dimension of sampling be? What data is already generated to meet the objectives identified across the temporal dimension?

2.6 What are the practical limitations of integrating spatial and temporal sampling?

2.7 What should the minimum requirement for temporal sampling be? Does that differ among taxonomic groups and habitats?

Session 3 (big group): Parameterising the GO network definition: an integrative and modular framework for data acquisition

The universality of genomic data allows the seamless link among sites and multiple hierarchical levels (genetic, species, higher clades), for any taxonomic group and regardless of prior taxonomic knowledge. It also provides various entry points for integration with other data types used in Earth observation. To exploit the full power of these data, we require (i) a unified framework for data acquisition that maximises data comparability across sites and (ii) appropriate methods for extrapolation/interpolation of site-based data across space and time and models of biodiversity that take local-site data to infer global processes. We further envision (iii) the integration with remote sensing data, which requires the careful design of site-based sequencing studies to establish data complementarity for the estimation of biodiversity parameters.

This session will look to the future of site-based genomic sequencing, to discuss the wide range of emerging possibilities and attempt to channel them into more concrete research programs. The rapid growth of site-based data from current initiatives already provides the opportunity for large-scale compilation and analysis of regional and global biodiversity patterns, for marine, terrestrial (above and below-ground) and freshwater habitats. We will discuss how this rapidly growing database can be exploited with efficient bioinformatics procedures and high throughput analytical approaches, and whether we need novel biodiversity models to integrate local inventories with global-scale ecological and evolutionary processes. Biodiversity modelling might be also required to derive ‘essential biodiversity variables’ for the use of deep-sequencing data in ecosystem evaluation and to build a predictive framework of global change. We will also discuss the potential of linking genomic data with other data types providing information on land use change, pollution, climate change and others, which can greatly expand the power of the GO approach.

With this broader prospect in mind, a key objective of a GO network is to maximise data comparability across different types of research. What does this mean for site selection, sampling design and sequencing? Would the scientific community want to agree on common standards and goals? Ongoing regional and global initiatives could provide the basis for structuring the data generation and parameterisation of a unified framework. A “modular” approach for data acquisition may be the most promising avenue to overcome fragmented research interests, whereby site-based studies generate a minimum unit of data as the basic building blocks of large-scale analyses. A "module" thus refers to a set of protocols and pipelines for data acquisition of a particular "target group" (e.g. soil bacteria, fungi, soil mesofauna, canopy arthropods are all distinct "target groups"). The definition of such modules in general terms will be discussed in this session, while specific strategies for uniform data collection within such modules will be addressed in the smaller groups discussion (session 4).

3.1 What are the main types of non-genomic data that could provide synergy?

3.2 How can we best integrate site-based sequencing with remote sensing?

3.3 What might be the main impediments to achieve a synthesis of regional/global patterns from within a unified network of GOs?

3.4 What analytical approaches are available for extrapolation/interpolation across space and time? What biodiversity models are there/missing for extracting global patterns from site-based data?

3.5 What degree of standardisation is needed? What analytical approaches are available for integrating data from non-unified sampling sites?

3.6 What guidelines are needed for site selection for a GO to be considered within a GO network?

3.7 Define the sampling area for GOs, and use that to constrain step (i) of modules, or use step (i) of modules to define GO sampling area? Are there context dependent (e.g. abiotic gradients) sampling guidelines for comparability?

Session 4 (smaller groups): Designing a modular framework for data acquisition

We propose structuring the parameterisation of a GO network as a modular framework for data acquisition that will (i) maximise the generation of comparable data from independently sampled GOs, (ii) provide a broad measure of biodiversity within each site, and (iii) facilitate the study of regional to global-scale biodiversity patterns from the joint analysis of individual sites. In general terms, the parameterisation of each module includes four steps: (i) sampling the group in the field, (ii) processing the samples in the wet lab, (iii) amplification and sequencing of molecular regions in the molecular lab, and (iv) application of bioinformatic pipelines to generate the data required. The discussions of the previous session should, to some extent, provide a framework for how different modules may be able to share the same protocol for one or more of the four steps within their module (e.g. perhaps a common set of protocols for steps (iii) and (iv) will be achievable for aquatic and terrestrial microbes). However, there are also likely to be substantial differences across modules, particularly for steps (i), (ii) and (iii). It would also be desirable to integrate step (i) across different modules where possible, and as such this is best discussed for modules linked to a particular habitat. For that we propose to conduct this discussion in three smaller groups defined as (i) above ground, (ii) below ground and (iii) aquatic, to be able to address the following module-specific questions:

3.6. What modules (target groups) are fundamental for a GO network?

3.7. Which specific steps of the parameterisation of the different modules can be integrated?

3.8. What existing protocols, data sets or sampling programs can be tapped into for each module or group of modules?

3.9. Which degree of standardisation in protocols and pipelines is already developed for each module?

3.10. Which degree of comparability can be obtained for the data from different modules within a GO?

3.11. What other non-genomic data could provide synergy within each module and group of modules?