ASSEMBLE

ASSOCIATION OF EUROPEAN MARINE BIOLOGICAL LABORATORIES EXPANDED

Acronym: ASSEMBLE Plus

Title: ASSEMBLE PLUS - ASSOCIATION OF EUROPEAN MARINE BIOLOGICAL LABORATORIES EXPANDED

Grant Agreement: 730984

Deliverable [D7.2]

Standardization of Metabarcoding

March 2020

Lead parties for Deliverable: [10 - HCMR]
Due date of deliverable: M[12]
Actual submission date: M[15]

GENERAL DATA

Acronym:	**ASSEMBLE Plus**
Contract N°:	**730984**
Start Date:	**1st October 2017**
Duration:	**48 months**

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730984. This output reflects the views only of the author(s), and the European Union cannot be held responsible for any use which may be made of the information contained therein.

Deliverable number	7.2

Author(s)	Georgios Kotoulas, Melathia Stavroulaki
Editor(s)	
Version	

Deliverable title	Standardization of Metabarcoding
Submission due date	
Actual submission date	1 March 2020
WP number & title	WP7 – JRA1 Genomics Observatories
WP Lead Beneficiary	10 - HCMR
Participants (names & institutions)	HUJI, IL: Genin Amatzia, HUJI, IL, Frada Miguel, HCMR, GR: Kotoulas G., Polymenakou P., Stavroulaki M., Gerovasileiou V., Pavloudi Ch, Kasapidis Panagiotis, Arvanitidis Ch. NIB, SI: Ramsak Andreja , SZN, IT: Zingone A. Casotti R., Piredda R. OBS-VLFR, FR, Guidi L. i CCMAR-UALG, PT, Canario Adelino, Louro B. ECIMAT-UVIGO, ES, Villanueva A., Troncoso J. UPV/EHU, ES, Cancio Ibon, SBR, FR Not F., Garczarek L., Viard F., Vaulot D. Comtet T. VLIZ, BE, Deneudt K., Hablutzel P. Exter K., MPI-BREMEN, DE, Glöckner F.O., Kostadinov I. IOPAN, PL , Wenne R., Kuklinski P., UG, PL, Wolowicz M., Lasota R., TZS-UH, FI, Norkko T.Joanna, SYKE, FI Hermanni Kaartokallio, Kremp Anke GU, SE, Obst M., MBA, UK, Cunliffe M., Chrismas N., BAS, UK Clark S.M., NUI, IE, Allcock L., Power A.

Dissemination Type

Report	□
Websites, patent filling, etc.	□
Ethics	□
Open Research Data Pilot (ORDP)	□
Demonstrator	□
Other	X

Dissemination Level

Public	X
Confidential, only for members of the consortium (including the Commission Services)	□

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 730984. This output reflects the views only of the author(s), and the European Union cannot be held responsible for any use which may be made of the information contained therein.

Document properties

Abstract

The present document originally can be used as an introduction to DNA metabarcoding technology for potential users who have not encountered it before. The main purpose it serves though, is to attract the attention of laboratory practicians to the many possible sources of biases that can lead, on one hand to the impossibility of cross-study comparisons and on the other to erroneous conclusions in biology, ecology and evolution in marine ecosystems. The discussion aims to address the need for standardized procedures that implement broadly distributed Genomics Observatories. It describes, in a historical context, the evolution of Next-Generation Sequencing technologies in parallel with the decreasing investment in taxonomy. It further highlights the value of Next Generation Sequencing coupled with environmental data for the assessment and monitoring of community structure and evolution and for the quantitative scaling up of biodiversity assessments. The report try to acknowledge all known sources of biases, i.e. sampling, sample storage, DNA extraction, choice and use of primers, PCR amplification, "library" construction, DNA sequencing, as well as batch effects due to individual labs, and it just touches to the biases due to data analysis which would require a separate treatment. It also brings out the importance of DNA barcoding reference databases, in adding value to DNA metabarcoding data. Concluding, the document is meant to be a living document in order to be continuously enriched with the expertise and protocols of ASSEMBPE Plus and other communities using standards and investing in the development of distributed Genomics Observatories.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730984. This output reflects the views only of the author(s), and the European Union cannot be held responsible for any use which may be made of the information contained therein.

Table of Contents

1. Introduction

The possibility to combine Genomics with environmental data under appropriate designs of environmental sampling, offers an effective way to assess marine biodiversity, discover patterns, monitor change, assess the evolutionary history and interpret associations between physical drivers and biological communities. Based on Next Generation Sequencing, named “highthroughput” technology, it gets a highthroughput-monitoring approach that scales from local to global. When properly applied, environmental genomics allows extracting signal from noise and achieve assessment of environmental health, facilitating sustainable management of natural resources (Adamowicz *et al*., 2019). Next Generation Sequencing (NGS) brought big data to biology and ecology and is considered a revolution paralleling the discovery of the microscope, keeping promise for new discoveries and new insights into biological complexity. Technology and data on their own do not constitute science; making sense out of data requires well thought sampling and experimental designs and processes in the way that data are produced, described, managed, made accessible, analyzed, and communicated. These are complex steps, hindering knowledge production and prone to erroneous outcomes and/or interpretations.
Knowledge production from data requires knowing the potential and limitations of methods, the sources of biases and how to avoid or quantify them. Application of new technologies in biological research do not necessarily ease or exempt from challenges related to sampling strategy, sample handling, sample archiving, organizing and keeping track of sample experimental processing, and equally importantly describing, managing and deciding among the many possible ways of analyzing data. On the contrary they may and they do add new ones. To step towards a global science, beyond the technical difficulties, there are also social and organizational prerequisites such as building consensus and standards about use of concepts, methods and processes and making research transparent and reproducible. Biodiversity assessment by Next Generation Sequencing evolved from DNA barcoding of single individuals, to DNA metabarcoding assessing the structure of complex biological communities by analyzing environmental samples, to DNA metagenomics and meta-Transcriptomics assessing the structure, the functional potential and the functional instantiation of communities. In this report, we will mainly address DNA metabarcoding.

2. Objective

The main objective of the present deliverable is to offer an overview of DNA metabarcoding as a tool for scaling biodiversity assessment, and to raise awareness about the limits and sources of possible biases. It is also meant to serve as a living document accessible over the ASSEMBPLE Plus Portal, hosting, renewing and sharing experience of ASSEMBPLE Plus, EMBRC and collaborating Genomics Observatories practitioners, and informing on good practice principles and protocols.

3. DNA barcoding

A DNA barcode is a DNA marker, in the form of a DNA sequence of limited length, which is able to amplify DNA from any taxon, and sufficiently informative to differentiate a species from any other closely related species. The original aim of DNA barcoding is to assign a molecular identifier to each taxon, allowing overcoming limitations of morphological approaches in species identification. Paul Hebert’s vision

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 730984. This output reflects the views only of the author(s), and the European Union cannot be held responsible for any use which may be made of the information contained therein.

aimed at addressing the collapsing taxonomic expertise by “the construction of systems that employ DNA sequences as taxon 'barcodes” (Hebert P.D et al., 2003). Following those lines, the ultimate objective of the CBoL consortium (http://www.ibol.org/phase1/cbol/) is to ideally make available, under the above rule, a reference database comprising a barcode for each existing species. Not only species identification requires expert taxonomic knowledge, but it also requires good biological knowledge of life cycles and species ecology, in order to be able to sample biodiversity as exhaustively as possible. Usually it is more difficult to taxonomically assign early developmental stages of most organisms, while this is not an issue for DNA based methods. Indeed with the expansion of molecular approaches, cases of morphology-based misidentification and misclassification of species due either to evolutionary convergence or to different dynamics between molecular and morphological divergence, have become commonplace unraveling the existence of many sibling (cryptic) species. The barcoding of life, which has been made possible thanks to Sanger sequencing technology, is quite demanding as it requires expertise in taxonomy, as well as, in high quality data curation. DNA barcoding initially did not attract so much interest as it does today, as most questions asked in ecology, population genetics and evolution required multi-gene approaches. This trend has been increasing while technologies were evolving, until our days, where whole genome sequencing data are produced even at the level of populations. This reality was unimaginable at 2001, where the first whole human genome sequence has been published, having cost millions and having involved many Institutions around the world from the public as well as from private sector.

Today, expertise in taxonomy continues its decreasing trend. The molecular revolution and the human-centered research have completely washed out of focus taxonomy so that today taxonomists are getting scarce. While DNA barcoding is considered as a way to partially mitigate this gap, its progress dramatically needs taxonomic expertise to establish links between DNA sequences and taxa and bring human expertise into artificial intelligence algorithms. Once this huge task is sufficiently mature, the gap will somehow be covered, although the need for taxonomic expertise will never vanish, as naming species is inherent to human way of perceiving and dealing with the reality of biological diversity.

The possibility of applying DNA barcoding is due to the common origin of species and is dependent on the way that evolutionary process has been realized and has been operating on both genomes and phenotypes. The neutral theory of molecular evolution (Kimura M., 1968) is our basic hypothesis when reconstructing phylogenetic relationships of organisms by using a broad range of approaches and methods. Taxonomy does not always reflect phylogenetic relationships, but thanks to molecular evolution and the availability of molecular data, new trends in taxonomy tend to take into account phylogenetic approaches. DNA barcoding data can be, and are used for phylogenetic analysis. Nevertheless, they do not always contain sufficient information to resolve challenging phylogenetic questions or to resolve taxonomy. In brief, Molecular Evolution (Kimura 1968, 1983, Graur, D. & Li, W.-H. 2000), and subsequently Molecular Systematics (Hillis et al., Phylogenies (https://www.journals.elsevier.com/molecular-phylogenetics-and-evolution) have preceded the practice and protocols of DNA barcoding, which is just one application of the above scientific fields. However, just this one use of molecular evolution has in recent years become more known to the public than its broader use in molecular evolution or molecular systematics. The application of DNA barcoding would remain very useful in any case; nevertheless, its broadest usefulness comes from the value it adds to DNA metabarcoding. DNA metabarcoding which has been introduced with the advent of Next Generation Sequencing allows assessing biodiversity by bulk analysis of environmental samples offering insights into the structure and dynamics

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 730984. This output reflects the views only of the author(s), and the European Union cannot be held responsible for any use which may be made of the information contained therein.

of biological communities across habitats and ecosystems at a highthroughput rate. DNA barcoding when comprehensively developed in an ecosystem transforms quantitative information of unknown sequences, to high value community ecology data. By this process, the combined DNA barcoding and metabarcoding data greatly impacts ecological studies, and biodiversity monitoring offering operational tools for environmental management.

4. DNA metabarcoding

DNA metabarcoding is a highthroughput DNA-based method for rapid identification of species from environmental samples, which may contain hundreds or thousands of species. This is achieved by bulk DNA extraction from the sample, followed by the amplification, of a targeted DNA fragment ("Marker gene") from the population of DNAs in the sample. Next follows the massively parallel sequencing of the amplified fragments (amplicons) and the bioinformatic analysis of the outcome (Taberlet et al., 2012). DNA metabarcoding has served two distinct research communities: prokaryotic and eukaryotic, which today jointly address holistic approaches on ecosystem functioning. Cytochrome c oxidase subunit one (COI) is a well-known gene used as a DNA barcode of eukaryotic taxa. Nevertheless, there is no DNA barcode that can amplify all existing eukaryotic taxa (Elbrecht and Leese, 2015), and to be able to do so with equal efficiency for all of them (PCR amplification biases). Moreover, there is no single marker that can differentiate all existing taxa, especially to distinguish among closely related ones. Therefore different DNA barcodes have been used depending on the questions been asked, and in many cases combination of them are used to increase resolution. There are different strategies followed when designing a study of biodiversity assessment by DNA metabarcoding. There are available apparently, universal primers that is, oligonucleotide primer pairs that have been designed to amplify a marker gene across a broad phylogenetic range, allowing to capture the complexity of the communities. In the case that a group of taxa (.e.g. diatoms, or nematodes), are not amplified appropriately and are absent or underrepresented following PCR amplification, then new primer pairs specific to the given group can be designed. Such approach addresses the need to get data from a given marker gene. We also need to cope with the case when we do get amplification of marker genes across all taxa addressed by a study, but this marker gene is not suitable for part of the phylogenetic tree. This happens in many cases, as there is no ideal marker gene, good for all kind of questions. In such a case we can turn around such an impasse by use of other marker genes. As a result of such a process today several marker genes (DNA barcodes) are used, addressing with different effectiveness different phylogenetic groups and different questions. By combining different DNA barcodes we can address a broad range of questions related to biodiversity assessment. Even population genetics can be served by means of technologies like eDNA (Parsons KM et al., 2018). Whatever method we use, users need to be aware of the possible sources of biases, of the power and limitations of each method and approach used. In the case of DNA metabarcoding, which involves a series of technical steps, each step is prone to its specific sources of bias. Also, even under fully unbiased data, the methods get full value when there are rich and high-quality DNA barcoding reference databases, so that ideally, every species encountered in the given ecosystem has a DNA barcode and by consequence it gets a valid taxonomic assignment.

Extracellular DNA (eDNA)

Extracellular environmental DNA (eDNA) is DNA "escaping" from organisms in the environment, found either in free cells, diluted DNA. This may originate from skin, scales, mucus, faces, gametes and so on, being offered to highly noninvasive

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730984. This output reflects the views only of the author(s), and the European Union cannot be held responsible for any use which may be made of the information contained therein.

technology of eDNA. It has been defined as "DNA that is not associated with living biomass" (Corinaldesi et al., 2008), and more explicitly: "We define extracellular, naked, free, ambient, or environmental DNA as: those molecules present in, or released from, cells in which energy production has permanently ceased, viral DNA, and DNA secreted from metabolically active cells" (Nielsen et al., 2007). The methodology of eDNA owes its existence in the stability of DNA as a molecule, and the possibility of its detection to PCR technology and NGS. In the same way with DNA metabarcoding, eDNA technology uses the sequence of DNA fragments targeted and amplified by PCR to reach detectable levels, as is used as proxy for the presence of organisms. The mechanisms of extracellular DNA release have been well studied in the case of evolved mechanisms (physiological cells) or dead cells in microbial communities (Ibáñez de Aldecoa, *et al*., 2017).

It has been a great positive surprise that DNA can survive and be detected in the most variable environmental settings, under specific environmental conditions, surviving even for thousands of years (Corinaldesi *et al*., 2008). On the other hand, the dilution in the marine environment and the rates of decay are key issues affecting estimation of biodiversity (Harrison et al., 2019). DNA persistence is defined as the time of preservation of DNA in the environment once the source of origin of this DNA is removed (Díaz-Ferguson and Moyer, 2014). We mainly count in term of hours, usually up to 48 hours, while it behaves differently in fresh water versus marine ecosystems and it varies depending on the conditions (Collins, R. A., *et al*., 2018, Sassoubre *et. al*., 2016). This makes geographically precise estimations of occurrence of organisms difficult to standardize for one more reason; DNA sampled in a site may originate from more or less distant sites. Here we will not develop further eDNA technology.

The study of biodiversity by means of analyzing eDNA, represents a very "hot" research activity as it may unravel treasures of information about biodiversity structure (presence and abundance or organisms), although it still require many steps to standardization of different services. This technology differs from DNA metabarcoding in the sampling procedure, and in the knowledge of the mechanisms and dynamics leading from living organisms to samples of extracellular DNA, while it is about the same for the steps of PCR and downstream. In the interpretation of data one should take into account the consequences of the differences between the two technologies on biodiversity estimations.

All comments on sources of biases and processes presented in this document for DNA metabarcoding apply also for eDNA-based assessment of biodiversity.

5. DNA metabarcoding vs. DNA barcoding; mind the differences

Although the two approaches seem to differ mainly in the sequencing technology used and in DNA extraction from single individuals versus DNA extraction from environmental samples, there are different requirements in primer design, and therefore DNA regions targeted for amplification. DNA barcoding requires long reads with high discrimination capacity, and therefore pose different primer design constraints than DNA metabarcodes that are much shorter and which might offer broader range of solutions for primer design. As an example, SILVA database is a reference database which hosts full and highly curated 16S rRNA sequences of bacteria and the CBoL does the same for CoI gene for animals or rbcL and matK genes for plants, (Guillou et al., 2013). On the other hand, metabarcoding, which is based on bulk DNAs of environmental samples of any type, faces different challenges. Generally, environmental samples are not found under ideal conditions, as it may be the case for single organisms and the DNA may be degraded. Therefore in practical terms only short amplicons can be amplified for collections of

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730984. This output reflects the views only of the author(s), and the European Union cannot be held responsible for any use which may be made of the information contained therein.

environmental samples, and highly conserved primers need to be used to capture biodiversity by minimizing the amplification biases in mixed-template reactions (Riaz et al., 2011). Although DNA metabarcodes aimed at comprehensive biodiversity capture are designed around hypervariable regions of DNA barcodes, being small offer smaller resolution; nevertheless, for many questions, it is sufficient to assign a DNA barcode to higher taxonomic level (genus or family). This means that there are cases where DNA metabarcoding cannot distinguish between some species having small divergence (e.g. sibling species). Some answer to such a deficiency could be given by the creation and use of **local reference databases**. To the best of our knowledge, there are only few species with global distribution, so that geography may be a discriminatory factor for barcodes that cannot differentiate between two closely related species having different geographic range.

An additional way to increase resolution is to design primers amplifying only a subset of taxa that address a specific biological question. In this case, there isn't such a strong constraint as if addressing global biodiversity that requires universally conserved regions to design primers. By consequence there are more options to identify regions that at the same time are conserved within the targeted taxa and that amplify the most variable and discriminatory marker genes.

6. DNA barcodes and DNA reference databases

DNA metabarcoding sequences in the absence of **reference databases,** are lists of sequences named Operational taxonomic Units (OTUs), that cannot be assigned to biological taxa they come from, and therefore the interpretation of data, although still useful, looses much of its value. More formally, an Operational Taxonomic Unit (OTU) is a cluster of organisms grouped by DNA sequence similarity of a specific taxonomic marker gene (Blaxter *et al*., 2005). It is widely used in case of a DNA sequence of the organism with no proper taxonomy established.

In the actual state of NGS technology applications, there are thousands of complete genomes being decoded, which enrich public domain databases. We can figure out a not too distant future where any DNA sequence, could be assigned to a taxonomically annotated full genome and therefore to a species. Within such a framework, almost any sequence would be a marker gene, allowing species identification by simply mapping it against the annotated genome data collections. Since, science is not yet so advanced, it is built instead an «encyclopedia» for any extant species with data of sequences, for a series of specific genomic regions that we call marker genes, and tag the origin of each marker to the species it originated from. In this way, whenever this sequence appears in a sample, by queering the proper reference database of sequences from specimens taxonomically annotated by experts it unravels the presence of that species in the sample, and therefore the anonymous OTUs survey gets all the organismal biology knowledge of the organism. This makes clear in what ways by going through those DNA reference databases, we capitalize over the taxonomic assignment work that has been done by taxonomic experts, when populating the reference databases. This taxonomic expertise is then used time and again in DNA metabarcoding biodiversity scans.

We also need to stress that we need Global DNA reference database need to be enriched with data from respective local reference databases. Without such local databases when scientists try to associate biodiversity patterns to ecosystem functioning, ecosystem services and their changes, they will depend on taxonomic assignments from other areas introducing errors in estimations at several levels such as community composition, species invasiveness, population and community connectivity etc. (Bohan et al., 2017, et al., 2019). The geographic and Ecosystemic distribution and coverage of local reference databases should take into account

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730984. This output reflects the views only of the author(s), and the European Union cannot be held responsible for any use which may be made of the information contained therein.

oceanographic historical and actual connectivity data, biogeographic considerations and phylogeographic and population genetic data (Carr et al., 2017, Hillman et al., 2018).
List and presentation of DNA reference databases, is beyond the scope of the present report, and will be addressed within the project's portal.

7. Possible sources of biases

Knowledge and presentation of known sources of biases and suggestions about how to avoid them may help reduce them. This knowledge is very useful even in case that it fails to fully prevent biases, as it helps interpreting the data, which should be seen under the assumption that each one bias or combination of individual biases have occurred, against the assumption of no biases. If we can exclude biases as causes of the data at hand, then we can look for biological parameters behind the data produced. In case of doubts, additional experiments, or approaches may be needed.

7.1. Sampling and sampling intensity

Population sampling is primordial in statistics as it is in ecology. Here we just take the opportunity to remind that sampling needs to be standardized among stations in order to allow for meaningful comparisons. A tricky issue on that respect is sampling intensity. Sampling intensity may affect estimations of diversity and community composition even when the very same sampling method is used. It has been shown that richness estimates are very sensitive to sampling intensity (Grey et al., 2018). In OSD the same method and volume of sampled water across sampling sites are used, following a standard protocol, while sampling takes place in the mid-day to exclude biases due to day-night differences in community occupancy of the water column due to the well-known Diel Vertical Migration, DVM (e.g. Häfker et al., 2017). Nevertheless, population densities differ between different ecosystems, e.g. between extreme oligotrophic Mediterranean type and the much more productive Atlantic ecosystems or eutrophic sites. Such differences affect both population census sizes and biomass in ways usually not investigated or accounted for. Therefore in the case of plankton sampling, sampled water volume is not always a good measure of sampling intensity, as it does not reflect the number of sampled cells, in the case of bacterial sampling, or number of organisms in the case of sampled multicellular organisms. In the case of bacteria, equalizing samples at the level of quantity of DNA to be sequenced is a way to address the issue. In multicellular eukaryotes it is more challenging to equalize sampling intensity between ecosystems.

7.2. Storage of samples

Sample storage may have significant impact on the assessment of bacterial community structure, either prokaryotic or eukaryotic. In the case of bacteria, due to differences in cell wall composition and in differences in creation of specific structures, such as spores for some free living bacteria, or biofilms, some species are more prone to degradation than others, so that suboptimal sampling process (e.g. time lag between sampling and sample storage) or type of storage, will enrich the community towards the most resistant to degradation. There are several references for some community types, further work may still needed in the case of different marine habitats.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730984. This output reflects the views only of the author(s), and the European Union cannot be held responsible for any use which may be made of the information contained therein.

In eukaryotes we know from population genetics of individual species or related studies that sample storage impacts DNA quantity and quality, while this is much more prominent in the case of RNA. Transferring this to community level we can safely assume that sample storage is critical in the assessment of community composition. If we expand the issue to shotgun metagenomic analysis where relatively high molecular weight DNA is needed, sample storage becomes a really critical issue. Comparing the two approaches, important biases to shotgun metagenomics studies may come from sample storage, which although also important source of biases for DNA metabarcoding, it is less critical, while on the other hand, DNA metabarcoding important biases may introduced by the PCR step.

7.3. DNA extraction

Taking as an example the theoretically easy case of bacteria, as being unicellular, where there are similarities in bacterial cell walls of different species involving lipids, lipopolysaccharides, membrane proteins, etc., but at the same time there are striking differences among others, such as between bacteria and Achaea, between Gram positive and Gram negative, as well as within and between the Firmicutes and Actinobacteria, etc. (Kuhn A, 2019). This has as consequence that different bacterial species have differences on how easily they are lysed by different methods, impacting DNA yield and quality for different bacterial species. As a consequence by using different DNA extraction methods on the same community, we may come up with significantly diverging community structure estimations concerning either differences in abundance, or/and in diversity estimations. Indeed DNA extraction may be a very significant source of biases in estimation of community structure, in some cases accounting for over 10 or greater fold differences - for the same sample - in relative frequency of a taxon (Costea et al., 2017), while in other cases errors rates from biases of over 85% have been observed in some samples when different extraction kits had been used on the same mock samples (Brooks et al., 2015).

Unless the same protocols are used, the observed differences may be due to the differences in the protocols used rather than to the original community composition. This may be more pronounced in the case of multicellular organisms, where additional treatments are required and a number of alternative protocols could be used (e.g. depending on the organisms and non organismal components contained in the environmental sample to get read of humid acids, chitins, silica, calcium carbonate, cellulose, proteins, carbohydrates, fatty acids, pigments etc.). Although DNA metabarcoding methods are designed to be able to work with material of poor DNA quality, environmental DNA extraction should target isolation of high quality, high molecular weight DNA as this may serve other purposes, such as production of shotgun metagenomics data, or to be subject to long read sequencing technologies. The performance of different DNA extraction methods can be assessed by use of mock communities, created from know species. Scott Tighe et al., 2017 presents an interesting model process for comparing different DNA extraction protocols by use of mock communities. In this study, eight different protocols, among which known commercial protocols, have been compared for DNA yield, and on assessment of the initial variation by subsequent sequencing. On the other hand, mock samples do not have any environmental contaminants that may interfere with DNA extraction and/or with down stream processing (PCR, library construction, sequencing reactions), so that standardization of methods also need to be tested on real environmental samples. Such an example is given for bacterial communities of marine sediments by use of nine protocols using CTAB aiming at the removal of humic acids (Kachiprath et al., 2018). Although DNA extraction method is critical for an unbiased estimation of microbial diversity, there are still no standard procedures even within the oldest and

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 730984. This output reflects the views only of the author(s), and the European Union cannot be held responsible for any use which may be made of the information contained therein.

broader microbial research community, which is the Human microbiome community, so that still in 2019 there is a call for establishing Minimum Standards Requirements for DNA extraction (Greathouse et al., 2019).

Concerning microbial protocols for environmental samples a major resource for microbial diversity is the Earth Microbiome (http://www.earthmicrobiome.org/). In the case of Ocean Sampling Day, a protocol has been standardized and used following optimization and benchmarking.

7.4. Primer choice

16S rRNA gene may have several copies within a genome and there are differences in copy number between different bacterial species. Therefore this may impact they quantity of PCR products biasing estimations of abundance (Kormas 2011). With increasing number of whole genome data for bacterial species across the phylogenetic tree, we now have a good overview of copy number per species, so that we may correct estimations. It has been shown that incorporation of such information improves estimates of microbial diversity and abundance (Kembel et al., 2012).

Primer choice can have a very important impact on biodiversity estimation as well as on species assignments. The later depends not only of the discriminatory capacity of any marker but also at how comprehensive the respective reference databases are. Today there are available various software applications that facilitate primer design, in ways that allow the choice to be adapted to the specific question asked, or to accommodate global biodiversity assessment of taxa specific assessment. On the other hand, only primers that have been experimentally tested are informative on the capacity to assess a part of biodiversity. As literature increases in size, there is no easy way to link all ever tested primer pairs with the range of taxa they are capable of detecting. In addition, primer effectiveness also depends on the experimental conditions, quality of DNA template etc. so that it is not always straightforward to assess the value of a primer pair. The number and quality of published studies is a criterion vs. single case publications, for instance. A software or text mining approach that can inform on range of taxa and other biodiversity indices, which is estimated by using every primer pair ever published would be very useful. Certainly, new primers designing, with the use of software will continue to be developed and used. Although many of DNA metabarcoding applications may be replaced by whole genome or whole organelle DNA sequencing data, DNA metabarcoding will continue to be a useful method, as it is today, and as it will be improved in the future.

7.5. Next Generation Sequencing

There are known batch effects in DNA sequencing so that even the same library run at different times by the same machine, gives different community composition. Ideally, if sample sizes allow, samples from different communities to be compared should be managed in the same run. Considering the multiplexing capacity of sequencing technologies in the case of DNA metabarcoding, hundreds of samples can be accommodated within the same run. Nevertheless, we wish to be able to compare data from different studies. When possible, it is important to use the same sequencing platform and machine, operated by the same experienced engineer to reduce batch effects as much as possible. There are additional sources of errors such as within the same run well-to-well contamination, a quite common problem in sequencing analysis, surprisingly is more pronounced when sample handling is automatic. This type of contamination impacts more importantly samples with smaller DNA quantity. The fact that there are not only external sources of contamination, but

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730984. This output reflects the views only of the author(s), and the European Union cannot be held responsible for any use which may be made of the information contained therein.

also between sample of the same run, makes so that decontamination by simple removal of variants detected in negative controls is not appropriate (Walker A.W., 2019) but more careful decontamination is needed, taking for instance into account the proximity of the wells of samples in the plate.

7.6. Individual lab experimental vs. centralized sample processing

By now it has been made clear that lab specific biases can be introduced at any step along the multistep experimental suites, therefore, it is not easy to disentangle the source of variation between lab specific effects and real patterns. In addressing the experimenter effect usually blind experiments are performed (Holman et al., 2015) and should be considered in further developments. The use of the same mock community samples as controls in all labs, is also a useful tool allowing to assess only some of the possible biases; different mock samples may reveal different biases, so mock sample is not a general solution in assessing between labs or experimenters biases. Certainly we may expect that the biases will be smaller when procedures are well standardized and the personnel performing the experiments are experienced. In ASSEMBPLE Plus, being very demanding to benchmark different methods for such multistep experimental procedures we have been conservative and have centralized to a single laboratory and been processed by the same experimenter the use cases that we have introduced as first steps to the implementation of Genomics Observatories. This is the case for both, Ocean Sampling Data (OSD) and the Autonomous Reef Monitoring Structures (ARMS). The experience of OSD2018 to OSD2019 events brought out a number of issues. Some are already resolved while others need to be discussed before changing the standard procedures followed so far. One drawback of the procedure followed is the cost of shipping the samples with dry ice. There were also cases that samples originating from non-European countries were lost due to unpredictable operations of the shipping companies or due to inflexibility over the in-between Custom's checkpoints. Summing up, the samples shipping issues are mostly related to labeling of expenditure and in using the right shipping boxes and quantity of dry ice. Other methods of maintaining and shipping samples can be tested in the future to adopt less expensive ones. Before changing any procedure systematic benchmarking for EMBRC stations need to be performed and assessment of the impact of new versus older methods on the estimation of community composition. One such test may be to compare data produced when DNA is extracted using the same protocol, by each one lab and be sent centrally vs. the until now used approach to send samples in the same lab that performs DNA extraction. By sharing samples in two, one part to be treated locally and the other to be sent on a central point, we can assess the impact of lab effect on the parameters of DNA extraction. This should be repeated several times in order to get an appreciation of possible biases. Such experiments can be done on the whole process too. Mock samples should also be included in the exercise.

8. Data issues - Bioinformatics

Data analysis is both a science and to some extend an art, and the conclusions reached may differ depending on the platforms, the tools, the parameters, etc. This cannot be addressed here, we just need to be aware that big data present big challenges and reproducibility issues in their analysis. Analysis of high quality data require detailed data description using standards, data management, data curation,

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730984. This output reflects the views only of the author(s), and the European Union cannot be held responsible for any use which may be made of the information contained therein.

data processing at different levels, and data analysis and visualization. Reproducibility of results is very challenging in the case of bioinformatics as there are many ways to process all the above with a large choice of software. Even by use of the same workflow and analysis pipeline, parameters can be set in many different ways, based on the assumptions of each researcher. Even early stages of analysis, before processing NGS data for proper community ecology analysis, e.g. treating raw data for their quality, can greatly affect the quality of research outcomes (Bokulich *et al*., 2013). This makes of it even more difficult to get reproducible results. Indeed, data related issues is a hot subject in research and efforts are made for data to be Findable Accessible Interoperable and Reproducible (FAIR data Principle) but also efforts are made so that procedures of Analytics be fully traceable.

We need to be aware that still today "…for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias" (Ioannidis 2015).

9. Glances to the close future

MGS: Marker-gene and Metagenomic Sequencing (jointly MGS) are two methods for the assessment of community structure by Next Generation Sequencing and Bioinformatics analysis of environmental samples following bulk DNA extraction. The two methods differ in many ways, but both follow a similar for most steps experimental workflow. Since Shotgun Metagenomic is still expensive we have focused on DNA metabarcoding, which is Marker-gene based sequencing. As sequencing becomes cheaper and cheaper and while shotgun metagenomics produces much richer information on environmental communities while it is exempt from bias-prone PCR amplification step, we may wander if DNA metabarcoding will survive further evolution of technologies overdriven by whole genome sequencing methods. The answer here depends on the specific questions that DNA metabarcoding is meant to answer. The main use of DNA metabarcoding is the assessment and monitoring of community structure, its change in space and time and search for underpinning factors. This information is valuable on its own as it can inform policies and assist environmental management. For instance, based on a combination of community structure and gene expression levels, it appears that "in polar regions, alterations in community activity in response to ocean warming will be driven more strongly by changes in organismal composition than by gene regulatory mechanisms" (Salazar *et al*., 2019), something that can very well be monitored by DNA metabarcoding methods. The evolution of additional technologies will enrich the value of DNA metabarcoding. For instance, under the hypothesis that in a close future, databases will host high quality annotated genomes for all extant taxa associated with rich contextual data (e.g. environmental data), the one-to-one relation between marker gene and taxa targeted by DNA barcoding, will to a great extend become one-to-one relation between marker genes and annotated genomes. This will make of DNA metabarcoding a great tool, allowing to scale up monitoring capacity and effectiveness, at tractable data management and compute cost. Such a hypothesis, already today seems realistic, with the main bottleneck being access to high quality, high molecular weight DNA rather than sequencing capacity.

For now, evolution of sequencing technologies can help getting functional information about communities, such as metabolic processes and physiological capabilities, together with community structure. An example of the perspectives been opened by such technological developments are viral and bacterial shotgun metagenomics studies. These studies not only allow taxonomic assignment and assessment of community structure but also offer functional annotation and allow a dialogue

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730984. This output reflects the views only of the author(s), and the European Union cannot be held responsible for any use which may be made of the information contained therein.

between candidate functionalities and environmental parameters. We are already at the point where whole genomes are reconstituted from metagenomic data known as metagenome-assembled genomes -MAGs (e.g. Delmont *et al.*, 2018), without that meaning that there are no challenges still remaining with data analytics of metagenomes (Shaiber and Eren, 2019).

When we address eukaryotes, shotgun metagenomics is not a very effective method due to larger genome sizes, which means lower coverage level for the same sequencing effort. Nevertheless, technologies are still in transition phase, and long-read sequencing technologies are evolving day-after-day, while even today, metagenomics analysis of some eukaryotic communities may come up with low coverage whole genome sequencing, something like partial MAGs (Olm *et al*., 2019), which is valuable information. For instance, low coverage genome sequencing is valuable for phylogenomic analysis (Zhang et al., 2018), which can shed light on the biology and ecology of organisms, by mapping their functional traits on phylogenetic trees. Such knowledge allows making predictions of response to environmental changes, and together with community ecology can advice ecosystem and biodiversity management, or can help interpret the observed reality.

Another hope, already applied for getting more that single marker DNA metabarcoding information, is genome skimming as a universal 'extended barcode', which is a low-coverage shotgun sequencing of genomic DNA which allows recovering the complete organelle genomes (either mitochondrial or plastid) and nuclear ribosomal regions for an increasing number of specimen. As a result are recovered all standard barcoding regions together with many other genes producing phylogenetically and functionally rich information (Coissac et al., 2016).

With the evolution of long read sequencing technologies towards better performance, certainly steps will also be improved in this direction for organisms with larger genomes. Challenges about data management and compute capacity that HPC platforms face with the actual rate of data production (Zekun Yin et al., 2017), will further raise. On the other hand, the need for well-designed sampling, and adoption of best practice protocols will remain part of responsible way of doing science.

10. Conclusions

Here we have presented a global appreciation of the major factors that one needs to be aware of when designing a DNA metabarcoding project. It is different to address a research project than to apply a standardized protocol in a long-term observatory. The two need to crosstalk in order to add improved components in the latter, without loosing capacity of comparing with earlier stages of a time series scientific observation.

Exact protocols and operational tools for DNA metabarcoding will be given in the GOs Portal as we progress with OSD and ARMS. The analysis of the ARMS optimization process and analysis of results, which is an ongoing process, will be an important component and this will also be the case following the comparative analysis of OSD2018, OSD2019 data and their comparison with OSD2014 data of the Micro B3 project. The portal will be enriched over the course of the project with the contribution of all teams, and this will also be the case for survey of literature. Such a base of knowledge shall be used in further developing the A+ roadmap for GOs. Although DNA metabarcoding is the inexpensive way to assess biodiversity today and although it has many limitations when compared to access to whole genome data, beyond its own interest, it is a very useful tool to inform and educate on good research practices related to ecological research and long term biodiversity observation.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730984. This output reflects the views only of the author(s), and the European Union cannot be held responsible for any use which may be made of the information contained therein.

11. References

Adamowicz Sara J., Cicanovski I., Ratnasingham S., Hebert P.D.N., Braukmann T.W.A., and Kremer S.C., 2019. Extracting signal from noise: big biodiversity analysis from high-throughput sequence data. *The 8th International Barcode of Life Conference, Trondheim, Norway, 17–20 June 2019.*

Blaxter, M., Mann, J., Chapman, T., Thomas, F., Whitton, C., Floyd, R., & Abebe, E. (2005). Defining operational taxonomic units using DNA barcode data. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *360*(1462), 1935–1943. doi:10.1098/rstb.2005.1725

Bohan, D. A., Vacher, C., Tamaddoni-Nezhad, A., Raybould, A., Dumbrell, A. J., and Woodward, G. (2017). Next-generation global biomonitoring: largescale, automated reconstruction of ecological networks. Trends Ecol. Evol. 32, 477–487. doi: 10.1016/j.tree.2017.03.001

Bokulich N.A., Subramanian S., Faith J.J., Gevers D., Gordon J.I., Knight R., Mills D.A., Caporaso J.G., 2013. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nat Methods 10(1): 57–59. doi: 10.1038/nmeth.2276

Brooks JP, Edwards DJ, Harwich MD, Rivera MC, Fettweis JM, Serrano MG, Reris RA, Sheth NU, Huang B, Girerd P, Strauss JF (2015) The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. BMC Microbiol 15:66. https://doi.org/10.1186/s12866-015-0351-6.

Carr M.H., Robinson S.P., Wahle Ch., Davis G., Kroll S., Murray S., Schumacker E.J., Williams M., 2017. The central importance of ecological spatial connectivity to effective coastal marine protected areas and to meeting the challenges of climate change in the marine environment. Aquatic Conservation Marine and Freshwater Ecosystems. 2017;27(S1):6–29.

Coissac, E., Hollingsworth, P.M., Lavergne, S. and Taberlet, P., 2016. From barcodes to genomes: extending the concept of DNA barcoding. Mol Ecol, 25: 1423-1428. doi:10.1111/mec.13549

Collins, R. A., Wangensteen, O. S., O'Gorman, E. J., Mariani, S., Sims, D. W., & Genner, M. J., 2018. Persistence of environmental DNA in marine systems. *Communications biology*, *1*, 185. https://doi.org/10.1038/s42003-018-0192-6

Corinaldesi, C., Beolchini, F. and Dell'Anno, A., 2008. Damage and degradation rates of extracellular DNA in marine sediments: implications for the preservation of gene sequences. Molecular Ecology, 17: 3939-3951. doi:10.1111/j.1365-294X.2008.03880.x

Costea PI, Zeller G, Sunagawa S, Pelletier E, Alberti A, Levenez F, Tramontano M, Driessen M, Hercog R, Jung FE, Kultima JR, Hayward MR, Coelho LP, Allen-Vercoe E, Bertrand L, Blaut M, Brown JRM, Carton T, CoolsPortier S, Daigneault M, et al. 2017. Towards standards for human fecal sample processing in metagenomic studies. Nature Biotechnology 35:1069–1076. DOI: https://doi.org/10.1038/nbt.3960, PMID: 28967887

Díaz-Ferguson E., Moyer G.R., 2014. History, applications, methodological issues and perspectives for the use environmental DNA (eDNA) in marine and freshwater environments. *Revista de biologia tropical* 62(4):1273-84. DOI: 10.15517/rbt.v62i4.13231

Elbrecht V, Leese F, 2015. Can DNA-Based Ecosystem Assessments Quantify Species Abundance? Testing Primer Bias and Biomass—Sequence Relationships with an Innovative Metabarcoding Protocol. *PLoS ONE*, 10(7): e0130324. doi.org/10.1371/journal.pone.0130324.

Graur, D. & Li, W.-H., 2000. Fundamentals of molecular evolution. Sinauer. ISBN 0-87893-266-6.

Greathouse, K. L., Sinha, R., & Vogtmann, E. (2019). DNA extraction for human microbiome studies: the issue of standardization. *Genome biology*, *20*(1), 212. doi:10.1186/s13059-019-1843-8

Häfker, N. S. et al. Circadian clock involvement in Zooplankton diel vertical migration. *Curr. Biol.* **27**, 2194–2201.e3 (2017).

Harrison J.B., Sunday J.M., and Rogers S.M., 2019. Predicting the fate of eDNA in the environment and implications for studying biodiversity. Proc. R. Soc. B. 286: 20191409. http://doi.org/10.1098/rspb.2019.1409.

Hebert PD, Cywinska A, Ball SL, 2003. Biological identifications through DNA barcodes. Proc R Soc Lond B Biol Sci., 270:313–21.

Hillis D.M., Moritz C., Mable B.K. (eds.), 1996. Molecular Systematics. Sinauer, Sunderland, Massachsetts. 655 pp.

Hillman J.R., Lundquist C.J., Thrush S.F., 2018. The Challenges Associated With Connectivity in Ecosystem Processes. Frontiers in Marine Science, 5:364. 10.3389/fmars.2018.00364.

Holman L, Head ML, Lanfear R, Jennions MD (2015) Evidence of Experimental Bias in the Life Sciences: Why We Need Blind Data Recording. PLoS Biol 13(7): e1002190. doi:10.1371/journal. pbio.1002190

Ibáñez de Aldecoa, A. L., Zafra, O., & González-Pastor, J. E. (2017). Mechanisms and Regulation of

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730984. This output reflects the views only of the author(s), and the European Union cannot be held responsible for any use which may be made of the information contained therein.

Extracellular DNA Release and Its Biological Roles in Microbial Communities. *Frontiers in microbiology*, *8*, 1390. https://doi.org/10.3389/fmicb.2017.01390

Kachiprath, B., Jayanath, G., Solomon, S., and Sarasan, M. (2018). CTAB influenced differential elution of metagenomic DNA from saltpan and marine sediments. *3 Biotech* 8:44. doi: 10.1007/s13205-017-1078-x

Kembel SW, Wu M, Eisen JA, Green JL (2012) Incorporating 16S Gene Copy Number Information Improves Estimates of Microbial Diversity and Abundance. PLoS Comput Biol 8(10): e1002743. https://doi.org/10.1371/journal.pcbi.1002743

Kimura, M., 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge. ISBN 0-521-23109-4.

Kimura, Motoo (1968). Evolutionary rate at the molecular level. Nature, 217: 624-626.

Kormas KA (2011) Interpreting diversity of Proteobacteria based on 16S rRNA gene copy number. Proteobacteria: Phylogeny, Metabolic Diversity and Ecological Effects, ed Sezenna ML (Nova Publishers, Hauppauge, NY), pp 73–89

Kuhn A, 2019. The Bacterial Cell Wall and Membrane-A Treasure Chest for Antibiotic Targets. Subcell Biochem. 2019;92:1-5. doi: 10.1007/978-3-030-18768-2_1.

McGee KM, Robinson CV and Hajibabaei M (2019) Gaps in DNA-Based Biomonitoring Across the Globe. Front. Ecol. Evol. 7:337. doi: 10.3389/fevo.2019.00337

Nielsen K.M. , Johnsen P.J. , Bensasson D. , Daffonchio D., 2007. Release and persistence of extracellular DNA in the environment. Environ. Biosaf. Res., 6 :37-53.

Olm, M.R., West, P.T., Brooks, B. *et al.* Genome-resolved metagenomics of eukaryotic populations during early colonization of premature infants and in hospital rooms. *Microbiome* **7,** 26 (2019) doi:10.1186/s40168-019-0638-1.

Parsons KM, Everett M, Dahlheim M, Park L. 2018 Water, water everywhere: environmental DNA can unlock population structure in elusive marine species. R. Soc. open sci.5: 180537. http://dx.doi.org/10.1098/rsos.180537

Ramond P, Sourisseau M, Simon N, Romac S, Schmitt S, Rigaut-Jalabert F, Henry N, de Vargas C, Siano R, 2019. Coupling between taxonomic and functional diversity in protistan coastal communities. *Environ Microbiol*., 21(2):730-749. doi: 10.1111/1462-2920.14537.

Salazar et al., 2019. Global Trends in Marine Plankton Diversity across Kingdoms of Life. Cell, Volume 179, Issue 5, 14 November 2019, Pages 1084-1097.e21

Sassoubre, Lauren M. and Yamahara, Kevan M. and Gardner, Luke D. and Block, Barbara A. and Boehm, Alexandria B., 2016. Quantification of Environmental DNA (eDNA) Shedding and Decay Rates for Three Marine Fish. Environmental Science \& Technology, 50 (19): 10456—10464.

Shaiber, A., & Eren, A. M. (2019). *Composite Metagenome-Assembled Genomes Reduce the Quality of Public Genome Repositories. mBio, 10(3).* doi:10.1128/mbio.00725-19. Yin Zekun, Lan Haidong, Tan Guangming, Lu Mian, Vasilakos Athanasios, Liu Weiguo, 2017. Computing Platforms for Big Biological Data Analytics: Perspectives and Challenges. Computational and Structural Biotechnology Journal, 15: 403-411.

Tighe, S., Afshinnekoo, E., Rock, T. M., McGrath, K., Alexander, N., McIntyre, A., … Mason, C. E. (2017). Genomic Methods and Microbiological Technologies for Profiling Novel and Extreme Environments for the Extreme Microbiome Project (XMP). *Journal of biomolecular techniques : JBT*, *28*(1), 31–39. doi:10.7171/jbt.17-2801-004

Walker AW. 2019. A lot on your plate? Well-to-well contamination as an additional confounder in microbiome sequence analyses. mSystems 4:e00362-19. https://doi.org/ 10.1128/mSystems.00362-19.

Zhang Feng, Ding Yinhuan , Zhu Chao-Dong , Zhou Xin ,. Orr Michael C, Scheu Stefan, Luan Yun-Xia, 2018. Phylogenomics from low-coverage whole-genome sequencing Methods Ecol. Evol., 10: 507-517. DOI: 10.1111/2041-210X.13145.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730984. This output reflects the views only of the author(s), and the European Union cannot be held responsible for any use which may be made of the information contained therein.