InBIO RESEARCH NETWORK
Biodiversity and Evolutionary Biology

EnvMetaGen

Deliverable 4.2 (D4.2)

Protocol for building and organising reference collections of DNA sequences

Project acronym:	ENVMETAGEN
Project name:	Capacity Building at *InBIO* for Research and Innovation Using Environmental Metagenomics
Work Programme Topics Addressed:	H2020-WIDESPREAD-2014-2 (ERA CHAIRS)
Grant agreement:	668981
Project duration:	01/09/2015 – 31/08/2020 (60 months)
Co-ordinator:	ICETA - Instituto de Ciências e Tecnologias Agrárias e Agro-Alimentares
Delivery date from Annex I:	M36 (August 2018)
Actual delivery date:	M37 (September 2018)
Lead beneficiary:	ICETA
Project's coordinator:	Pedro Beja

Dissemination Level		
PU	Public	✓
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	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 668981

TABLE OF CONTENTS

SUMMARY

The overall goal of ERA Chair/EnvMetaGen project No 668981 is to expand the research and innovation potential of InBIO – Research network in Biodiversity and Evolutionary Biology, through the creation of an ERA Chair in Environmental Metagenomics. This field was selected as the focus of the ERA Chair, because Environmental DNA (eDNA) analysis is increasingly being used for biodiversity assessment, diet analysis, detection of rare or invasive species, population genetics and ecosystem functional analysis. In this context, the work plan of EnvMetaGen includes one work package dedicated to the Deployment of an eDNA Lab (WP4), which involves the training of InBIO researchers and technicians for implementing best practice protocols for the analysis of eDNA (Task 4.2). These protocols are essential for key application areas and to the development of research projects in association with business partners and other stakeholders, and thus to the strengthening of InBIO triple-helix initiatives (InBIO-Industry-Government; WP5). This report provides an overview of the current state of the art for the development of best practice for building DNA reference collections of voucher specimens identified by specialised taxonomists, and how these practices are being implemented at InBIO. Building and organising reference collections of DNA sequences is an essential component of this task because eDNA studies require that DNA sequences recovered from the environment are compared to reference collections, which thus need to be developed following consistent and repeatable procedures. Such reference collections are likely to become a tool with significant relevance to the InBIO-Industry-Government triple-helix activities (WP5) by promoting the development of partnerships in all key areas: Monitoring of freshwater eDNA for species detection; Assessing natural pest control using faecal metagenomics; and Next-generation biomonitoring using DNA metabarcoding. Protocols described herein were developed in close connection and due to the accomplishment of the other WPs in the project, including: i) the recruitment of the ERA Chair team (WP2), ii) secondments researcher exchanges through collaborations with international networks (WP3), iii) the enhancement of the computational infrastructure at InBIO (WP4) and iv) participation and organization in workshops and conferences (WP6). Future directions in the improvement to enlarge the reference collection of DNA sequences of InBIO are identified. Together, Deliverables D4.2-D4.5 (this document; Egeter et al., 2018; Galhardo et al., 2018; Paupério et al., 2018) form a detailed account of the successful deployment of a fully functional eDNA lab under the EnvMetaGen project, and provide a valuable resource for eDNA practitioners in all spheres of the triple-helix model.

1. INTRODUCTION TO REFERENCE COLLECTIONS OF DNA SEQUENCES

1.1. Overview on reference collections of DNA sequences and their relevance

The storage and accessibility of DNA sequences have become crucial aspects in current research pipelines (Hsi-Yang Fritz et al. 2011, Muir et al. 2016). Since the early 1970s, when the first DNA sequences were obtained (Sanger et al. 1977), an enormous number of sequences have been generated worldwide as a result of quickly evolving sequencing technology. Currently, most available DNA sequence information can be found in major reference collections like the GenBank (https://www.ncbi.nlm.nih.gov/genbank/) and the Barcode of Life Data System - BOLD (http://www.barcodeoflife.org/) databases. However, initiatives focused on specific taxa or geographical regions have also led to the development of separate reference collections, such as Fredie - Freshwater Diversity Identification for Europe (https://www.fredie.eu/), FishBase - global species database of fish species (https://www.fishbase.de/) or the Fungal Barcoding Database (http://www.fungalbarcoding.org/).

Presently, the amount of data stored in GenBank exhibits a strong taxonomical bias towards animal taxa, with vertebrates being considerably better represented than invertebrates, although vertebrates are likely to represent less than 5% of the described animal species (IUCN, 2014). Additionally, there are geographic biases, with some regions being better represented than others, and often the better represented regions do not correspond to areas with higher biological diversity levels (Vernooy et al. 2010). As a result, most of the invertebrate taxa present in biodiversity hotspots around the world are currently poorly represented.

The importance of reference collections of DNA sequences surpasses the need of detailed documentation of global biological diversity. Reference collections are nowadays used as a crucial baseline for further biological research, but also have direct utility in answering practical questions related to forensics, agriculture, food industry and other aspects of day to day life, by allowing the identification of organisms from almost any kind of biological tissue and, more recently, environmental DNA samples.

The majority of projects developed at InBIO to date have been storing DNA sequence data in GenBank or Dryad upon publication of manuscripts. In most of these works, the generation of DNA sequence data constituted a mean to reach a research objective. With the advent of Next

Generation sequencing, it became possible to develop other kind of studies, namely the development of diet analysis and monitoring methods based on DNA metabarcoding. While such studies can include a range of environmental DNA sample types, such as faeces, saliva, blood meal, stomach contents, hair, water, air, pollen/natural by-products (e.g. honey), soil, bulk samples (or preservative), all demand the availability of a reference collection of DNA sequences in order to allow the correct identification of taxa found in each sample. This need was identified during the early stages of the EnvMetaGen project conception and for this reason the Task 4.2. *Building capacity for eDNA analysis* includes the construction and organisation of reference collections of DNA sequences as one of the pivotal capacity-building aspects. This component involves the development of best practice for building DNA reference collections of voucher specimens identified by specialised taxonomists, which is essential to develop and conduct consistent, reliable and repeatable research studies boosting the future performance of InBIO in environmental genomics.

1.2. Overview of the existence and development of reference collections of DNA sequences at InBIO

Since the creation of CIBIO in the 1990's, its researchers have been producing a considerable amount of DNA sequence data. The information produced has been systematically deposited on GenBank and Dryad upon publication of each project manuscripts. In general, there was no need to develop a reference collection of DNA sequences for the institution given the targeted taxonomic focus of the projects in place, which covered specific taxa phylogenies and phylogeographic history, with data being managed within the scope of each project. Furthermore, prior to Next Generation sequencing platforms DNA was only sequenced from isolated organisms, and in most cases with known taxonomic identity. After publication, this data would become available to everyone, and could be easily accessed from the public databases. Nevertheless, a few works are exceptions, Barbosa et al. 2013, Oliveira et al. 2010 and Silva et al. 2015, develop genetic tools for species identification. Two events had disrupted this line of functioning. On the one hand, DNA metabarcoding projects became possible with the technological developments allowing the sequencing of DNA from a single sample consisting of many species/individuals. Many opportunities of research in ecology and other disciplines had therefore been created. Simultaneously, this methodological advance significantly reinforced the importance of the establishment of DNA sequence reference

collections, which had become a priority itself, in order to allow the development of such studies.

On the other hand, the massive increase of data produced by Next Generation Sequencing has created the need of new ways of data processing and storage within each project. What once was a manageable amount of data, had become a challenge to handle using the same basic tools, and the complexity of the information regarding the quality of data has also defied existing databasing procedures. Most importantly, it also increased the potential of data redundancy if each project created and managed its own DNA reference collections. The main goal of DNA barcoding is to identify a certain organism based on a short DNA sequence previously sequenced from morphologically identified specimens. This requires the construction of comprehensive reference collections of DNA sequences that represent the existing biodiversity.

1.3. The projects need: why building and organising reference collections of DNA sequences at InBIO

Within the scope of InBIO, the need for building and organising reference collections of DNA sequences has become obvious to directly assist projects of freshwater monitoring and diet analysis, among other potential applications. DNA barcoding has become a powerful tool to researchers interested in population and community ecology, and is used as a conceptual framework to solve practical problems in ecosystem management, conservation biology, impact assessment and mitigation, and the sustainable use of biological resources. Nevertheless, its applicability is hampered by the lack of comprehensive reference collections, particularly of invertebrates that are underrepresented in reference databases. This knowledge gap becomes greater in biodiversity hotspots, as it is the Mediterranean Basin, the region of development and implementation of many of InBIO research projects. For this reason, InBIO has launched the InBIO Barcoding Initiative (IBI), which aims to develop a reference collection of DNA barcoding sequences covering Portuguese invertebrate taxa. Within this taxon, special focus is afforded to insects, given their relevance to food webs, ecosystems functioning, high diversity and limited number of available DNA barcoding sequences. The reference collection of DNA sequences conforms to the H2020 programme's principles of FAIR (findable, accessible, interoperable and reusable) data dissemination and also follows an Open Research Data (ORD) approach, as mentioned in the Deliverable 1.4 "*Data Management Plan*".

The EnvMetaGen project aims to develop three main strategic initiatives in key areas for

biodiversity conservation, the provision of ecosystem services, and environmental biomonitoring:

- Monitoring of freshwater eDNA for species detection;
- Assessing natural pest control using faecal metagenomics;
- Next-generation biomonitoring using DNA metabarcoding.

In all of these, the existence of a reference collection of DNA sequences is a crucial step for its successful development as it is needed to perform the correct assignment of sequences obtained by DNA metabarcoding methods to taxonomic entities. While the two first initiatives require a reference collection of DNA sequences of target taxa (either freshwater taxa or pest species), the third initiative requires a much more comprehensive reference collection of DNA sequences of the taxa present in each ecosystem.

In addition to invertebrate DNA barcoding, there are also ongoing projects that include DNA barcoding of vertebrate species not yet represented in public databases and for which DNA sequences are needed for its successful development: FILTURB (Amphibians), SABOR and TUA (Bats) and FRESHING (Fish) (Appendix).

1.4. Structure of the report

This report details the protocols being developed at InBIO laboratories for building and organising reference collections of DNA sequences, following the best practices. First, a discussion on the general aspects concerning challenges in building and organising reference collections of DNA sequences is provided. Then the procedures implemented are detailed and, finally, planned future work within the frame of the project is described. Figure 1 provides an overview of the EnvMetaGen eDNA Lab workflow, with steps arranged according to the reporting structure of Deliverables D4.2 - D4.5 (this document; Egeter et al., 2018; Galhardo et al., 2018; Paupério et al., 2018). Within this framework, the present report focus on the procedures related to specimen collection techniques, species morphological identification procedures, and sample processing protocols for DNA barcoding. The bioinformatics protocols for processing the sequencing data generated are detailed in Galhardo et al. 2018. For details of current EnvMetaGen affiliated projects, including their applicability to the triple-helix initiatives and EnvMetaGen objectives, see Appendix.

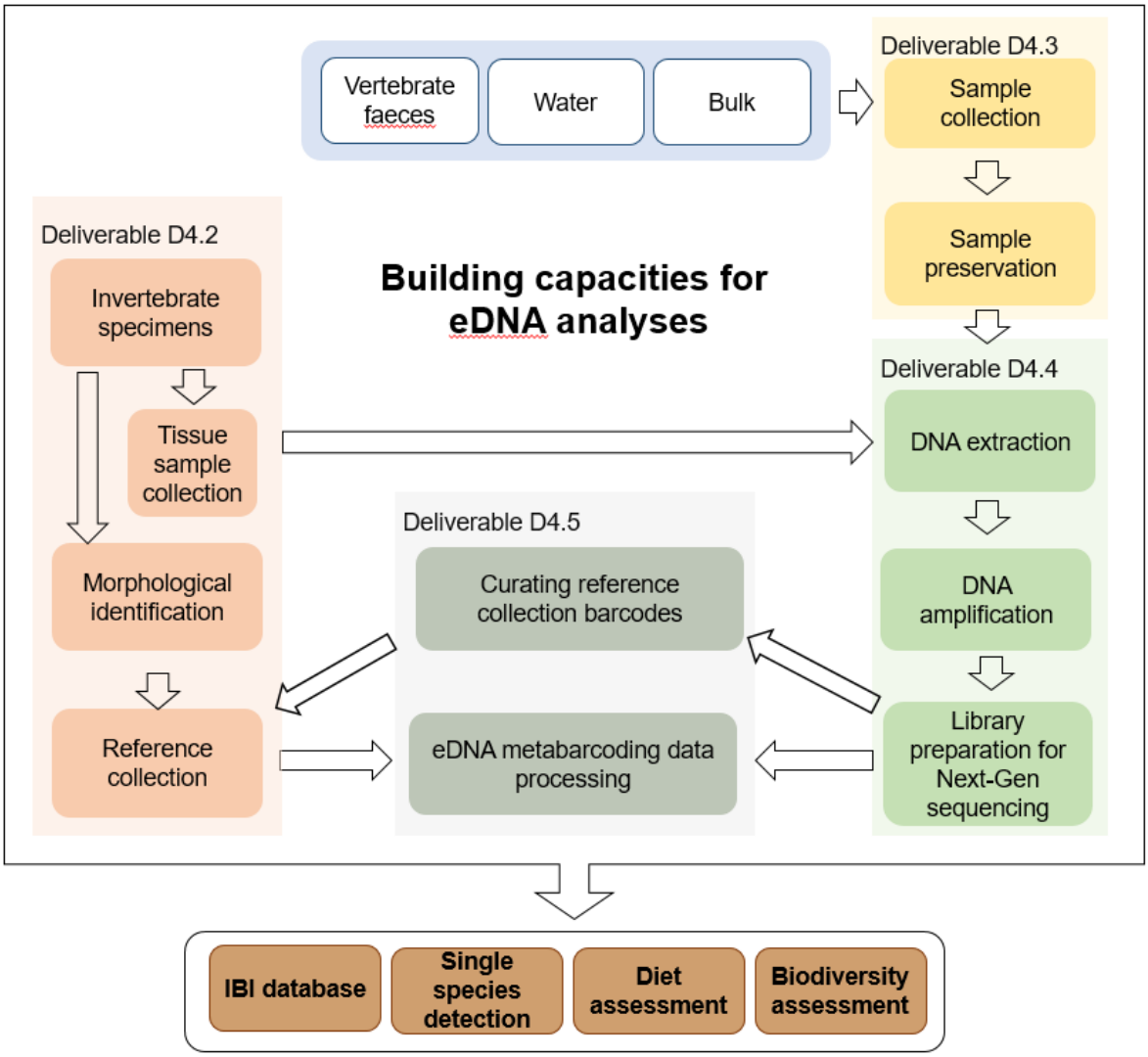

Figure 1. EnvMetaGen eDNA Lab workflow: steps are grouped by colours, as reported in Deliverables D4.2 - D4.5 (this document; Egeter et al., 2018; Galhardo et al., 2018; Paupério et al., 2018). The type of eDNA samples (blue) and project applications (brown) require a range of tailored protocols within workflow steps, which are detailed in Deliverables D4.2 - D4.5 (this document; Egeter et al., 2018; Galhardo et al., 2018; Paupério et al., 2018).

2. CHALLENGES IN BUILDING AND ORGANISING REFERENCE COLLECTIONS OF DNA SEQUENCES

Most of the ENVMETAGEN affiliated projects have their geographic scope in the Iberian Peninsula and focus on the identification of invertebrates, either for the characterisation of vertebrate diets (SABOR, TUA, GALEMYS, CHASCOS) or to be used as indicators for environmental monitoring (FRESHING). For this reason, the IBI has focused on DNA barcoding invertebrates from Portugal, prioritizing the taxonomic orders that are known to be more represented in bat diets (like Lepidoptera), and that are key for freshwater monitoring (Ephemeroptera, Plecoptera, Trichoptera and Odonata). More recently, it has become obvious that there was a need to expand the target groups to hyperdiverse orders such as Hymenoptera, Diptera and Coleoptera, as a significant number of the sequences obtained in current projects could not be assigned to a taxon, even at the family level within these orders.

Moreover, during the development of the ENVMETAGEN affiliated projects, it was considered pertinent to obtain DNA barcoding sequences for some vertebrate taxa present in Portugal. This is the case of bats (SABOR and TUA projects), amphibians (FILTURB) and fishes (AZORES and FRESHING). An additional collaboration with the University of Natural Resources and Life Sciences, Vienna, has been established, within the scope of the recently approved AGRIVOLE project, for building a reference collection for plants using high throughput sequencing, fundamental for the diet studies and vegetation surveys.

The future work to be developed regarding the reference collection of DNA sequences aims to cover taxonomical and geographic gaps, improving the representation of invertebrate species. To fulfil this aim it is crucial to strengthen the collaborations already established with taxonomists and develop new collaborations, focusing especially on taxonomic groups that are poorly represented in public databases. Nevertheless, this task is expected to be challenging in some groups due to poorly resolved taxonomy and the concomitant need of taxonomic revisions and/or the lack of available taxonomists.

3. STEPS IN THE BUILDING AND ORGANISING REFERENCE COLLECTIONS OF DNA SEQUENCES

The work related with building and organising a reference collection of DNA sequences within the scope of EnvMetaGen includes four main components: i) the reference collection development; ii) the specimen collection; iii) the morphological identification of specimens, and iv) the processing of the samples for DNA barcoding. In this section we detail each of this components and describe how they relate to each other.

3.1. Reference collection development

The collection of specimens to incorporate the InBIO Barcoding Initiative reference collection of DNA sequences began prior the formal start of EnvMetaGen project in 2015. For this reason, there was the need since the early stages of the project to store the specimens' data in a preliminary database that would gather all available information in an interoperable way with available public databases, in order to allow future deposition of data. Next, the structure of the database currently in use, and the characteristics of the database that is being designed to accommodate the InBIO Barcoding Initiative reference collection of DNA sequences, are described.

3.1.1. Database structure: Tables, Records, and Fields

The preliminary version of the database used in IBI project relies on an Excel document composed by 3 datasheets/Tables: *Specimens*, *Sequence data* and *Primer info*. In the *Specimens* datasheet each specimen has a unique entry and a specific unique code that is composed by the letters INV (to indicate it belongs to the IBI collection) and followed by a serial number composed of 5 digits. To each specimen information is stored in 10 sections:

- Taxonomy: Order, Family, Genus, Species;
- Field data and storage data associated: Collection Date, Country, State / Province, Municipality / Region, Locality / Exact site, Basin, Watercourse, Habitat, UTM (1km), LAT, LONG, Elevation, Gender, Life stage, Reproduction, Field ID, Museum ID, Museum Collection code, Institution storing, Collection event ID, Sampling protocol, Provisional ID, Storage, Identification method, Taxonomy notes;
- Identification data: Identifier, Collector, Identifier email, Identifier institution;

- Voucher data: Voucher, Voucher status, Voucher Box, Sample Box, Sample Type / Tissue descriptor;
- DNA extraction data: Sample Size, Sample status, DNA extraction, DNAPlate, DNA Well, Extraction Date;
- DNA Amplification: 16S Amplification, 16S Primer set, COI Amplification, LC, BH, Whole COI, COI Primer set;
- DNA sequencing: 16S Sequencing, 16S Sanger seq, COI Sequencing, COI Sanger seq, COI Sanger Seq date, Miseq, Miseq run date, Miseq comment;
- Sequencing data: MID, INDEX SET P5, INDEX SET P7, Miseq run folder;
- Observations: Obs.;
- BOLD inclusion: BOLD;

In the *Sequence data* datasheet each specimen may have a unique entry or several depending on how many DNA fragments have been sequenced from it. The information is presented in 6 sections:

- INV code: INV code - the data link this datasheet to the *Specimen* datasheet;
- Taxonomy: Order, Family, Species - these first fields have direct origin in the *Specimen* datasheet and are automatically filled in this datasheet once information is entered in the *Specimen* datasheet;
- Sequence data: Sequence ID, Marker, Primer set, Sequencing primer, Read Direction, Length (bp), Sequencing method, Sequence date, Alignment / Folder;
- Quality info: Quality info, LC coverage, LC Reads, BH coverage, BH Reads;
- Sequence: Sequence;
- Observations: Obs.

The third datasheet, designated *Primer info*, includes information on the primers used along the work, and the combinations and Polymerase Chain Reaction conditions used.

The current version of the preliminary database includes all fields required to deposit the data in BOLD database in an almost automatic way, a commitment assumed in the task 4.2 where all DNA sequence data generated in the project will be deposited to GenBank and/or BOLD databases.

The routine usage of the preliminary database allowed to make adjustments in the fields of the database and in the procedures to collect and register data, according to the needs detected along the project implementation, providing the baseline for the development of the final database.

3.1.2. Database development

To accommodate the data requirements of the InBIO Barcoding Initiative project a relational database is being developed using MySql. The database is normalized, thus reducing data redundancy, and a client side front end is being developed using Java. The database will be hosted on a server acquired under the scope of WP4 – Task 4.1: Enhancing the computational and data storage capacity. Access will be provided to users via the Java front end that will run independently on the individual laptops and desktops of project members. The front end will also be downloadable to third party users through the project website, and differing levels of access will permit differing features, for example administrators can modify and add new data, while third party users can view or extract data but not modify it, in line with what has been described as future evolutions of the Knowledge Management System in Deliverable 6.3.

The database itself is being designed to manage aspects of data management ranging from, specimen collectors, institutes they belong to, specimens themselves including a wide range of biological data including collection date, geographical coordinates, Species, Genus, Order, etc and eventually down to detailed sequencing information including primer combinations. Sequencing products will also be stored within the database and a range of methods for clustering, querying and extracting these sequences will be provided as a series of plugins that will be related to different aspects of the project.

3.2. Specimen collection

Specimens sequenced within the scope of the IBI can have two sources: specimens captured in the field, and specimens that have been previously deposited in an entomological collection.

3.2.1. Sampling techniques

Most specimens sequenced to date for IBI were collected in the field. The technique used for specimen detection and capture vary with the taxonomic group. Invertebrate species exhibit a remarkable diversity of life histories, abundance and phenology, therefore is not surprising that

there are several sampling techniques for this group. The diversity is so high that, even if we consider a particular taxonomic group, it would not be possible to sample all species by using a single sampling technique. Also, each sampling technique has its own limitations and targets distinct characteristics of species resulting in very distinct samples regarding taxonomical content. For this reason, it is frequent to use at least two or three sampling techniques to obtain a representative sample of the species richness in each study area. Knowledge of the biology, preferred habitats and activity patterns of each taxonomical group are crucial auxiliaries for the informed choice of sampling techniques. In the scope of IBI the aim is to cover Portuguese invertebrate taxa and in order to achieve this broad taxonomical scope several sampling techniques have been selected: direct search, light traps, aerial nets, aquatic nets, sweep nets, coloured pan traps, Malaise traps and pitfall traps.

3.2.1.1. Direct search

Direct search consists in the direct and active observation of surfaces in which the specimens might occur. Although it is often considered a technique with low efficiency, given the low numbers of specimens obtained, it can be one of the most useful techniques in exploratory studies, as it allows to learn the whereabouts of species and uncover aspects of their life history, providing data for sampling design. It includes searches under rocks, bark and other structures that other sampling techniques are not able to cover. When performed in an area not previously studied, it might provide valuable information regarding where to install traps or perform other sampling techniques. The taxa collected so far in the context of EnvMetaGen using direct search includes: Annelida, Araneae, Archaeognatha, Blattodea, Colembola, Coleoptera, Decapoda, Dermaptera, Diptera, Ephemeroptera, Hemiptera, Hymenoptera, Isoptera, Lepidoptera, Mantodea, Mecoptera, Megaloptera, Metastigmata, Neuroptera, Odonata, Opiliones, Orthoptera, Phasmatodea, Plecoptera, Pseudoscorpiones, Psocoptera, Pulmonata, Raphidioptera, Scolopendromorpha, Scutigeromorpha, Solifugae, Strepsiptera, Trichoptera, Trombidiformes and Zygentoma.

3.2.1.2. Light traps

Light traps have been used for inventories and monitoring invertebrate diversity of certain taxonomic groups (especially moths) for many years. In the scope of IBI two types of light traps

are being used in the field: Heath traps and mercury vapour lights. In the case of Heath traps the invertebrates are attracted to a UV light (Figure 2A). Once they arrive at the trap they collide with the trap flight interception structure and fall into a collection receptacle, where they remain until the trap is disassembled. The traps are portable and autonomous. Furthermore, they can be used for a standard time period in each sampling point. However, due to the characteristics of the light it is not safe for researchers to stay near the trap continuously and observe the arrival of animals, making selective collection of specimens. In contrast, mercury vapour lights (Figure 2B and 2C) allows continuous surveillance and detection of animals when they come to the light, even if they are not captured by the trap. This allows direct capture of a higher number of specimens and selection of samples prior to capture. However, these procedures greatly vary between researchers and therefore are less prone to standardization. The taxa collected so far in the context of EnvMetaGen using light traps includes: Blattodea, Coleoptera, Diptera, Ephemeroptera, Hemiptera, Hymenoptera, Lepidoptera, Mantodea, Neuroptera, Opiliones, Orthoptera, Plecoptera and Trichoptera.

Figure 2. Examples of light traps used during fieldwork in the scope of EnvMetaGen affiliated projects: A) UV LED light trap, B and C) mercury vapour lights setting.

3.2.1.3. Aerial nets

Aerial nets (also known as butterfly nets – Figure 3A) are one of several types of nets used to collect flying insects. The use of these nets implies an active search and chase of flying specimens, in order to intersect their flight. The bag of the net is generally constructed from a lightweight mesh to minimize damage of specimens. The taxa collected so far in the context of EnvMetaGen using aerial nets includes: Coleoptera, Diptera, Ephemeroptera, Hemiptera,

Hymenoptera, Mecoptera, Megaloptera, Neuroptera, Odonata, Orthoptera, Plecoptera, Raphidioptera and Trichoptera.

3.2.1.4. Aquatic nets

Aquatic nets are used to collect insects and other invertebrates that live underwater. Usually they have a wider mesh than aerial nets. One well known sampling technique is kick-net sampling, often used for stream and small river habitats and involves placing a net on the stream/river bed while disturbing the area immediately upstream of the net. This results in invertebrates on rocks and other sediments being caught in the net. The taxa collected so far in the context of EnvMetaGen using aquatic nets includes: Annelida, Coleoptera, Decapoda and Hemiptera.

3.2.1.5. Sweep nets

The use of sweep net (Figure 3B) can yield specimens of several taxonomical groups. The results can vary with the shape and size of the net but also with the sweeping technique. In this case, the collector does not necessarily observe the target specimens prior to net capture, but carries out a random sampling of the vegetation and/or soil surface. Observation is made after sweeping and all sample can be collected or the collection can be directed to key specimens chosen by the collector. It is a versatile technique as the collector can adjust the duration of sampling depending on the objectives. The taxa collected so far in the context of EnvMetaGen using sweep nets includes: Araneae, Archaeognatha, Blattodea, Colembola, Coleoptera, Diptera, Ephemeroptera, Hemiptera, Hymenoptera, Lepidoptera, Mantodea, Mecoptera, Neuroptera, Orthoptera, Phasmatodea, Plecoptera, Pseudoscorpiones, Psocoptera, Pulmonata, Trichoptera and Zygentoma.

3.2.1.6. Malaise traps

A Malaise trap is a tent-like structure (Figure 4A-B) used for trapping flying insects for a relatively long period of time (several days), and that tend to trap particularly Hymenoptera and Diptera species, but many other groups can be found represented in the samples. Insects fly into the tent wall and are funnelled into a collecting vessel that is filled with a preservative liquid and attached to the highest point. Two Malaise traps have been implemented within the scope

Figure 3. Examples of nets used for invertebrate sampling used during fieldwork in the scope of EnvMetaGen affiliated projects: A) aerial nets, B) sweep nets.

of the collaboration with the Malaise Global Program: A) site in the Eurosiberian region, B) site in the Mediterranean region. So far no capture has been done with other Malaise traps as these have been recently acquired (July 2018) and are currently being set.

Figure 4. Malaise traps implemented within the scope of the collaboration with the Malaise Global Program: A) site in the Eurosiberian region, B) site in the Mediterranean region.

3.2.1.7. Coloured pan traps

Coloured pan traps are a very used sample technique for flying insects (Figure 5A) along Malaise traps and sweeping nets. The coloured pan traps are remarkably versatile as any device that holds a certain amount of preserving liquid and that features a colour attractive to a certain taxon becomes a suitable pan trap. Colours frequently used are yellow, white and blue, but others may be used. The taxa collected so far in the context of EnvMetaGen using coloured pan traps includes: Diptera and Hymenoptera.

3.2.1.8. Pitfall traps

A pitfall trap is a trapping pit for small animals (Figure 5B-D), trapping mostly invertebrates but occasionally also vertebrates, as amphibians, reptiles and small mammals. Pitfall traps consist of containers (tin, jar or drum) buried in the ground with its rim at surface level used to trap mobile animals that fall into it. They can be dry (no liquid inside) or contain a solution designed to kill and preserve the trapped animals. These traps are often used in standardized studies but can also be used for exploratory studies, although their setting can be a time constrain. The implementation of pitfall traps to obtain specimens for the IBI reference collection of DNA sequences will take place in 2019. Specimens of several groups are likely to be collected in big numbers specially ground-dwelling Coleoptera, Hemiptera and Orthoptera.

Figure 5. Other examples of traps used for invertebrate sampling during fieldwork in the scope of EnvMetaGen affiliated projects: coloured pan traps (A); pitfall traps: set in the field (B), detail of the setting (C) and schematic illustration (D): 1 - top of stone or tree bark, 2 – tin, jar or drum, 3 – preservative liquid, 4 – soil.

3.2.2. Entomological collections

In the case of relatively rare species, for which it was not possible to obtain fresh specimens, the samples of specimens sequenced were obtained from material deposited in collections. This approach was applied to specimens from multiple insect orders, namely Lepidoptera from the private collection of Martin Corley, Trichoptera from Universidad de Santiago de Compostela, and Odonata from Universidad de Vigo.

3.3. Morphological identification

Taxonomical identification of specimens to be integrated in the reference collection of DNA sequences has been performed at species level by experienced taxonomists. Surveys have been done to identify experienced taxonomists for key taxonomical groups, and contacts have been made in order to establish collaborations.

Martin Corley, a collaborator of CIBIO since 2005, has been the responsible for moth identification (Heterocera, Lepidoptera) since the beginning of the project. Additional collaborations regarding moths have involved Sasha Vasconcelos (CIBIO-InBIO) and Jorge Rosete (amateur lepidopterist). Regarding freshwater invertebrates, four taxonomic groups have been given priority due to their key role in biomonitoring methods: Ephemeroptera, Plecoptera, Trichoptera and Odonata. Collaborations were established in 2016 with Dr. Michael T. Monaghan (Leibniz – Institute of Freshwater Ecology and Inland Fisheries; Ephemeroptera), Dr. José Manuel Tierno de Figueroa (Granada – Sciences Faculty; Plecoptera) and Dr. Marcos González, Dr. Jesús Martinez and Dr. Luis Martín González (Santiago de Compostela – Faculty of Biology; Trichoptera). The Odonata specimens are been identified by the ENVMETAGEN team member, Dr. Sónia Ferreira, a postdoc researcher funded by CIBIO – InBIO. An additional collaboration has been established with Dr. Adolfo Cordero Ribera (Vigo – Laboratory of Evolutionary and Conservation Ecology). Besides the taxonomical identification of specimens, these collaborations have resulted in fruitful exchange of data and access to specimens of taxa not previously barcoded. In February 2018, Dr. González became further integrated with the EnvMetaGen team with a postdoc position funded by CIBIO - InBIO, and is currently responsible for the taxonomic work regarding Trichoptera and Ephemeroptera.

The identification work on Orthoptera is performed mostly by Silvia Pina and Dr. Sónia Ferreira. The initial work on the DNA barcoding of bees (Hymenoptera: Apoidea) contemplated the specimens collected during Dr. Andrea Penado PhD (Penado, 2018) and that were identified by David Baldock (Surrey), Thomas J. Wood (East Lansing – Michigan State University), Ian Cross (Dorset) and Jan Smit (Duiven). Regarding ants, contact for future collaboration have been established with Dr. Mário Boieiro (Lisbon - cE3c). As for the Diptera order, a considerable increase in species and in families represented in the database is being obtained in result of the collaboration with Rui Andrade (Porto). Additionally, specimens of the families Tipulidae and Limoniidae are being identified by Dr. Pjotr Oosterbroek (Leiden – Naturalis).

Additional collaborations have been established with Dr. Josefina Garrido-González (Vigo – Grupo de Investigacion Entomologia Acuatica de la Universidad de Vigo: freshwater Coleoptera) and Dr. José Manuel Grosso-Silva (Porto - Museu de História Natural e da Ciência da Universidade do Porto).

3.4. Processing samples for DNA barcoding

All specimens collected in the field are preserved in 96% ethanol in labelled tubes, and stored in a room with regulated temperature (Figure 6A-C). A sample of tissue is then removed from every specimen for DNA barcoding. In the great majority of specimens one leg is removed and holds enough biological material for successful DNA extraction. Two DNA extraction kits are currently used: ExtractMe Genomic DNA 96-well kit (BLIRT S.A., Gdansk, Poland), for specimens bigger than 0.5cm, (over 85% of the cases), and the QIAamp DNA Micro Kit (Qiagen, Hilden, Germany) for specimens smaller than 0.5cm and/or older material preserved in collections. After DNA extraction the workflow includes amplification of the mitochondrial gene fragments (e.g. 658 bp of the COI gene, mitocondrial 16S) with group specific primers followed either by Sanger sequencing or by a high throughput sequencing (HTS) protocol, similar to that used for DNA metabarcoding (see Paupério et al. (2018) for details on the HTS workflow for DNA metabarcoding).

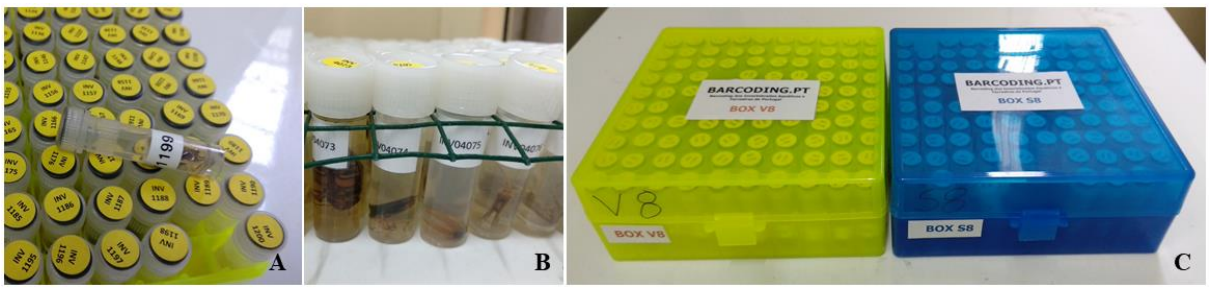

Figure 6. Specimens collected in the field preserved in 96% ethanol in labelled tubes (A) detail of tubes of 2 ml (B), detail of the tubes of 8 ml (note the high proportion of ethanol volume in each tube, and (C) example of Voucher boxes (yellow box) and sample boxes (blue box).

Briefly, Polymerase Chain Reactions (PCR) are performed in a 10ul volume, using 5ul of Qiagen© PCR Multiplex Kit Master Mix (Qiagen, Hilden, Germany), 0.3 to 0.4ul of each primer and 1ul of DNA. PCR thermo profiles vary with the amplified fragment (see details in Table 1, for the COI and 16S fragments amplified). The obtained PCR products are then purified with ExoSAP-IT® PCR clean-up Kit (GE Healthcare, Piscataway, NJ, USA).

In the Sanger sequencing workflow, the purified PCR products are sequenced with the amplification primers. Sequencing reactions are carried out using BigDye® Terminator v1.1 Cycle Sequencing Kit (Applied Biosystems, Carlsbad, CA, USA), and samples are subsequently sequenced for both strands on a 3130xl Genetic Analyser Sequencer (Applied Biosystems/HITASHI).

The HTS workflow for sequencing the 658bp COI fragment, follows the general approach described by Shokralla et al. 2015, and matches closely the workflows used within EnvMetaGen for DNA metabarcoding studies (for details see Paupério et al. 2018). Due to the maximum length allowed by the Illumina sequencing platforms, the COI fragment is amplified and sequenced in two overlapping fragments. The workflow is based on a double indexing, two-step PCR approach, where small barcodes ('indexes') are incorporated in both the forward and reverse Illumina adapter sequences (see Paupério et al. 2018, Sections 2.3 and 3.3, for details). Hence, briefly, for the barcode sequencing through HTS, the first PCR is performed with primers that target the region of interest (Table 1) but that already include an overhang adapter for the indexing and sequencing adapters. When Illumina indexes are used, a small multiplex identifier (MID, following Shokralla et al. 2015) is added to the overhang for allowing additional pooling of samples for sequencing. Indexes adapted from Kircher et al 2012 and Gansauge and Meyer 2013 are also used as an alternative protocol, without the need of adding the MID to the overhang sequence. After the purification of the first PCR with ExoSAP-IT® PCR clean-up Kit (GE Healthcare, Piscataway, NJ, USA), the second PCR is performed for incorporating the sequencing adapters and indexes. Then, the samples are cleaned with AMPure beads (Beckman Coulter), quantified with NanoDrop 1000 (Thermo Scientific), normalized and pooled for sequencing (see Paupério et al. 2018, sections 3.3 for details). Sequencing is performed on an Illumina Miseq platform, using V2 Miseq sequencing kits (2x 250 bp). The bioinformatics procedure that processes the Ilumina MiSeq sequencing run are described in detail in Galhardo et al. 2018, Section 4, "EnvMetaGen protocol for processing InBIO Barcoding Inititiative (IBI) DNA data".

Table 1. Primer sets and PCR conditions used for DNA barcoding.

Taxa	Gene	Forward primer	Reverse primer	TM	Reference	Seq[1]
Invertebrates	**COI**	LCO1490	Ill_C_R	Touch-up 47º-51ºC	Folmer et al. 1994 Shokralla et al. 2015	HTS
Invertebrates	**COI**	Ill_B_F	HCO2198	Touch-up 47º-51ºC	Shokralla et al. 2015 Folmer et al. 1994	HTS
Invertebrates	**COI**	LCO1490	HCO2198	45ºC	Folmer et al. 1994	Sanger
Invertebrates	**16S**	16Sar	16Sbr	50ºC	Palumbi et al. 1991	Sanger
Amphibians	**12S**	12SV5.1F	12SV5.1&2R	47/51ºC	Riaz et al. 2011	Sanger
Bats	**COI**	BF2	BR2	51ºC	Elbrecht and Leese, 2017	HTS
Bats	**COI**	fwhF1	Ill_C_R	45ºC	Vamos et al 2017 Shokralla et al. 2015	HTS
Fish	**COI**	LCO	HCO	53ºC	Folmer et al. 1994	Sanger
Fish	**CYTB**	Gludg-L	H16460, CB3-H	54ºC	Palumbi et al., 1991; Perdices & Doadrio 2001	Sanger
Fish	**12S**	MiFish-F	MiFish-R	65ºC	Miya et al., 2015	Sanger, HTS

[1] – Sequencing workflow

4. STATE OF THE ART OF INBIO BARCODING INITIATIVE

4.1. Implementation and growth of the collection

During the first three years of the EnvMetaGen project much effort has been devoted towards the constitution of a reference collection of DNA sequences representative of the biodiversity of Portugal, with special focus in geographic areas where the EnvMetaGen affiliated projects are implemented. So far, over 6200 specimens have been included in the reference collection. Most of these specimens originated from direct captures in the field, conducted specifically for increasing the reference collection. To this end, over 120 days of fieldwork have been carried out in northern Portugal, involving at least 3 researchers devoted to intensive invertebrate sampling using several techniques, adding more than 3800 specimens to the collection. This effort has been complemented with over 250 days in which specimen collecting activities took place in a less intensive way (less than 10 samples collected per day), as a result of using a single sampling technique or by taking advantage of fieldwork ongoing in the context of other projects. Supplementary additions to the reference collection are DNA tissue samples from museum or private collections.

From the about 6200 specimens that integrate currently the reference collection, more than 4900 have been already identified based on morphology by team members or project collaborators. These specimens represent 23 orders, 247 families and 2411 species included in the reference collection. Regarding the sequencing status of specimens, 4282 specimens were sequenced until July 2018, while nearly 1300 are currently being processed in the lab to be sequenced this summer. Figure 7 illustrates the number of specimens of the 10 most represented insect orders in the collection (n=5708), and the status of sequencing process of those specimens.

Of the hyperdiverse insect orders (Coleoptera, Diptera, Hymenoptera and Lepidoptera), priority was given to the inclusion of Lepidoptera species in the reference collection of DNA sequences due to their relevance as bat prey, and to the existence of species with potential as agriculture pests with high negative impact in the geographical region of EnvMetaGen affiliated projects (ECOLIVES, SABOR and TUA), complementing available information in public databases (*e. g.* BOLD and GenBank). Other orders that were given priority were those considered key for freshwater monitoring (Ephemeroptera, Plecoptera, Trichoptera and Odonata) (FRESHING). These orders are still underrepresented in public databases and include a considerable number of Iberian endemic species to which there is poor taxonomical resources. For this reason, efforts

have been directed to fill the identified gap in public databases, by generating DNA barcodes of unrepresented species.

Currently, plans are being made for the consistent inclusion of representatives from other hyperdiverse orders (Coleoptera, Diptera and Hymenoptera), that involve the identification of taxonomist collaborators and the design of specific fieldwork. Regarding less diverse insect orders, like the Dermaptera, Mantodea, Mecoptera, Neuroptera, Phasmatodea and Rhaphidioptera, it is expected to achieve comprehensive representation of Portuguese fauna in the reference collection of DNA sequences as result of Daniel Oliveira´s the master thesis currently ongoing under the supervision of Dr. Sónia Ferreira (2018-2019).

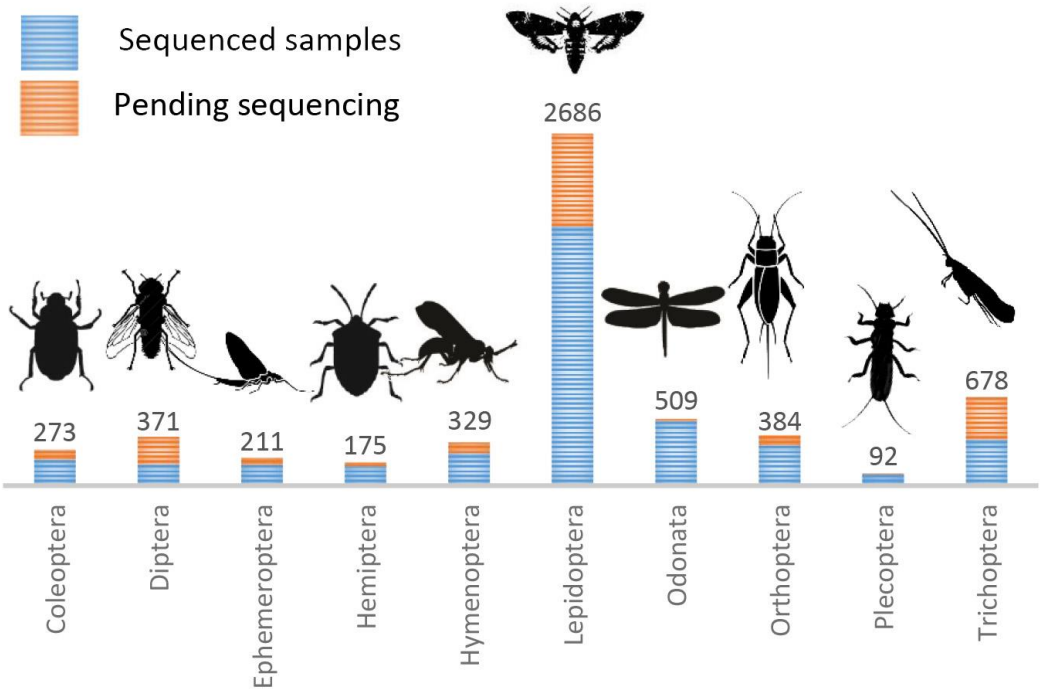

Figure 7. Number of specimens of the ten most represented insect orders currently incorporated in the Inbio Barcoding Initiative reference collection of DNA sequences.

4.2. Reference collection impact and outreach

As a component of EnvMetaGen project, the reference collection of DNA sequences follows standard "best practice" for the field of environmental metagenomics. It´s development takes into account the conformity with the H2020 programme's principles of FAIR (findable, accessible, interoperable and reusable) data dissemination and that also follows an Open Research Data (ORD) approach, as mentioned in the Deliverable 1.4 "*Data Management Plan*".

Active data dissemination to specialists, regarding the reference collection of DNA sequences, is being done primarily through published papers and oral and poster presentations at scientific conferences. The data generated is being stored in a compatible way to facilitate its deposition in BOLD and GenBank databases in due time. In all scientific conferences, special efforts have been made to invite researchers to collaborate with the reference collection of DNA sequences, in particular to establish new collaborations with taxonomical experts on groups that are poorly represented in the public databases.

4.2.1. Major findings

Along the development of the IBI reference collection of DNA sequences many have been the discoveries regarding the Portuguese fauna, and some of them with implications on a wider geographic and/or methodological scope. Herein are summarized the findings which have already been presented in scientific meetings and/or published.

4.2.1.1. New species in Europe

One remarkable finding was the identification of a moth species present in central Portugal for which the family was difficult to ascertain using morphology. *Borkhausenia crimnodes* Meyrick, 1912 (Lepidoptera, Oecophoridae), a species described from Argentina, was found to be resident in Beira Litoral, Portugal, constituting its first records in Europe. COI barcode sequencing has shown that a Portuguese specimen has 100% similarity with specimens collected in Eastern Cape – South Africa and deposited in the National Museum of Natural History, Smithsonian Institution, but whose identity was not known at that stage. By providing a link between these specimens it was possible to: 1) identify the Portuguese specimens as *Borkhausenia crimnodes* Meyrick, 1912; 2) record for the first time the species in Europe; 3) synonymise *Borkhausenia intumescens* Meyrick, 1921 described from South Africa with *B. crimnodes*, described from Argentina; 4) identify specimens deposited in the National Museum of Natural History, Smithsonian Institution through BOLD; and 5) make available morphological characters to identify the species. This finding has been focus of media attention: https://cibio.up.pt/media-clippings/details/there-is-a-new-exotic-moth-in-portugal

CORLEY MFV, FERREIRA S, LVOVSKY AL, ROSETE J. (2017) *Borkhausenia crimnodes* Meyrick, 1912 (Lepidoptera, Oecophoridae), a southern hemisphere species resident in Portugal. *Nota Lepidopterologica*: 40(1): 15–24.

2017 – SÓNIA FERREIRA, JOANA PAUPÉRIO, MARTIN CORLEY, JOANA VERÍSSIMO, FILIPA M.S. MARTINS, PAMELA PUPPO, VANESSA A MATA, HUGO REBELO, ANTÓNIO MUÑOZ-MERIDA, JOHN ARCHER, PAULO C. ALVES, SIMON JARMAN & PEDRO BEJA. "Large-scale barcoding of Portuguese moths: accelerating species inventories while revealing exotic species, sexual dimorphism and cryptic diversity" 7th International Barcode of Life Conference. Skukuza, South Africa 20-24 November. Oral communication

2016 – JOANA PAUPÉRIO, SÓNIA FERREIRA, FILIPA M.S. MARTINS, JOANA VERÍSSIMO, MARTIN CORLEY, VANESSA MATA, JOHN ARCHER, JOSE MANUEL GROSSO-SILVA, ANTÓNIO MUNOZ, SASHA VASCONCELOS, HUGO REBELO, PAULO C. ALVES & PEDRO BEJA. "InBIO Barcoding Initiative: moth reference database using Next Generation Sequencing" XVII Congresso Ibérico de Entomologia. Lisboa 5-8 September 2016. Oral communication

4.2.1.2. Sexual dimorphism

One of the strengths of DNA barcoding is the possibility to allow taxonomic identification of specimens irrespective of life stage or gender, which is particularly relevant in species with strong sexual dimorphism. COI barcode sequencing has associated unidentified female specimens with males of *Isotrias penedana* Trematerra, 2013 (Lepidoptera: Tortricidae, Chlidanotinae) revealing sexual dimorphism in the species, that was described from north Portugal based on males alone. The characters that distinguish males from females were described and illustrated making the identification of specimens accessible to everyone. Additionally, the species occurrence was also documented in northern Spain.

CORLEY MFV, FERREIRA S (2017) DNA Barcoding reveals sexual dimorphism in *Isotrias penedana* Trematerra, 2013 (Lepidoptera: Tortricidae, Chlidanotinae). Zootaxa, 4221 (5): 594–600.

2017 – SÓNIA FERREIRA, JOANA PAUPÉRIO, MARTIN CORLEY, JOANA VERÍSSIMO, FILIPA M.S. MARTINS, PAMELA PUPPO, VANESSA A MATA, HUGO

REBELO, ANTÓNIO MUÑOZ-MERIDA, JOHN ARCHER, PAULO C. ALVES, SIMON JARMAN & PEDRO BEJA. "Large-scale barcoding of Portuguese moths: accelerating species inventories while revealing exotic species, sexual dimorphism and cryptic diversity" 7th International Barcode of Life Conference. Skukuza, South Africa 20-24 November. Oral communication

2016 – JOANA PAUPÉRIO, SÓNIA FERREIRA, FILIPA M.S. MARTINS, JOANA VERÍSSIMO, MARTIN CORLEY, VANESSA MATA, JOHN ARCHER, JOSE MANUEL GROSSO-SILVA, ANTÓNIO MUNOZ, SASHA VASCONCELOS, HUGO REBELO, PAULO C. ALVES & PEDRO BEJA. "InBIO Barcoding Initiative: moth reference database using Next Generation Sequencing" XVII Congresso Ibérico de Entomologia. Lisboa 5-8 September 2016. Oral communication

4.2.1.3. DNA Barcoding as a useful tool in Sialis *sp.*

As mentioned in the previous case, DNA barcoding allows identification of specimens despite life stage or gender. During the development of the IBI reference collection of DNA sequences, DNA barcoding has proved to be a useful tool to identify Iberian species of *Sialis* in an automated way. Until now only *Sialis fuliginosa* Pictet, 1836, was known to be present in Portugal, but there are also numerous records of *Sialis* sp. larvae that have remained unidentified to species level. We detected two other species (*S. lutaria* Linnaeus, 1758 and *S. nigripes* Pictet, 1865) in Portugal based on morphology and DNA sequencing. The three species can be unequivocally assigned to the species level using DNA barcodes, and therefore we suggest its use in freshwater biodiversity assessments and environmental monitoring schemes.

FERREIRA S, PAUPÉRIO J, GROSSO-SILVA JM, BEJA P (submitted) DNA barcoding of Sialis sp. (Megaloptera) in Portugal: the missing tool to species identification.

2017 – JOANA PAUPÉRIO, SÓNIA FERREIRA, FILIPA M.S. MARTINS, JOANA VERÍSSIMO, ADOLFO CORDERO-RIVERA, JOSEFINA GARRIDO-GONZÁLEZ, LUIS MARTIN, JESUS MARTÍNEZ, MARCOS A. GONZÁLEZ, JOSE MANUEL TIERNO DE FIGUEROA, JOSE MANUEL GROSSO-SILVA, LORENZO QUAGLIETTA, ANTÓNIO MUÑOZ-MERIDA, MICHAEL T. MONAGHAN, ANA F. FILIPE, PAULO C. ALVES, SIMON JARMAN, PEDRO BEJA. "InBIO Barcoding Initiative: building a reference database

for freshwater insects using Next-Generation Sequencing". DNAqua-Net Kick-Off Conference. Essen, 7-9 March 2017. Poster

4.2.1.4. NUMTs and over estimation of OTUs

Dragonflies and Damselflies (order Odonata) have been one of the priority groups for DNA barcoding. Along the development of the work it was possible to identify some of the challenges posed by insect identification using DNA barcoding. While many species can be easily identified using the mitochondrial COI gene fragment, this is not always true. Not all species possess a specific DNA barcode (*e. g. Ischnura graellsii* and *I. elegans*), hindering the correct assignment of taxonomic names to unidentified specimens. Other species possess multiple copies of COI in the genome, impeding successful Sanger sequencing, which can be overcome using Next Generation Sequencing. This has implications on DNA metabarcoding studies analysis. Given that more than one copy of the target DNA barcode is obtained, it is likely that species diversity could be overestimated in some cases, and underestimated in cases of lack of specific DNA barcoding. These results have been presented in two oral communications, one poster and a manuscript is currently under preparation.

2017 – SÓNIA FERREIRA, ADOLFO CORDERO-RIVERA, JOANA PAUPÉRIO, JOANA VERÍSSIMO, FILIPA M.S. MARTINS, VANESSA A. MATA, ANTÓNIO MUÑOZ-MERIDA, PAULO C. ALVES, SIMON JARMAN, PEDRO BEJA. "Odonata Barcoding: The good, the bad and the ugly" DNAqua-Net Kick-Off Conference. Essen, 7-9 March. Oral communication

2017 – JOANA PAUPÉRIO, SÓNIA. FERREIRA, FILIPA M.S. MARTINS, JOANA VERÍSSIMO, ADOLFO CORDERO-RIVERA, JOSEFINA GARRIDO-GONZÁLEZ, LUIS MARTIN, JESUS MARTÍNEZ, MARCOS A. GONZÁLEZ, JOSE MANUEL TIERNO DE FIGUEROA, JOSE MANUEL GROSSO-SILVA, LORENZO QUAGLIETTA, ANTÓNIO MUÑOZ-MERIDA, MICHAEL T. MONAGHAN, ANA F. FILIPE, PAULO C. ALVES, SIMON JARMAN, PEDRO BEJA. "InBIO Barcoding Initiative: building a reference database for freshwater insects using Next-Generation Sequencing". DNAqua-Net Kick-Off Conference. Essen, 7-9 March 2017. Poster

2017 – SÓNIA FERREIRA, ADOLFO CORDERO-RIVERA, JOANA PAUPÉRIO, JOANA VERÍSSIMO, FILIPA M.S. MARTINS, VANESSA A. MATA, ANTÓNIO MUÑOZ-

MERIDA, PAULO C. ALVES, SIMON JARMAN, PEDRO BEJA. “Odonata Barcoding: The good, the bad and the ugly” 7th International Barcode of Life Conference. Skukuza, South Africa 20-24 November. Oral communication

4.2.1.5. Undescribed species distribution and phenology

Simultaneously to the development of the IBI reference collection of DNA sequences, there are projects using DNA metabarcoding being implemented focusing on the diet of selected organisms (GALEMYS, SABOR, TUA). Due to the incompleteness of the public databases there are a proportion of DNA haplotypes obtained that lack identification. However, by using the limited dataset of the IBI reference collection of DNA sequences it has already been possible to identify more than 100 species. One of the most remarkable species found was a moth species of the genus *Aproaerema* that has not been described yet, but was already represented in IBI reference collection of DNA sequences. It has been found in one bat diet sample and in 11 prey samples from the TUA project. The first specimen was collected in 2000 and due to limited data regarding number of specimens and number of localities, the description of the species has been postponed. With the current information on distribution and phenology it is possible to catalyse this species description.

2018 – SÓNIA FERREIRA, VANESSA A MATA, REBECA M CAMPOS, JOANA VERÍSSIMO, PEDRO BEJA. “DNA metabarcoding of hidden biodiversity in the Mediterranean Basin” 5th European Congress of Conservation Biology, Jyväskylä, Finland 12-15 of June. Oral communication

4.2.2. Participation in International projects

Along the development of the EnvMetaGen project, the team became involved with two international projects that have direct connection to the reference collection of DNA sequences and major impact in the achievement of its objectives: the Global Malaise Program (http://biodiversitygenomics.net/projects/gmp/) and the DNAqua-Net (http://dnaqua.net/).

4.2.2.1. Global Malaise Program

In the scope of EnvMetaGen, CIBIO-InBIO became the first Portuguese institution participating in the Global Malaise Program (http://biodiversitygenomics.net/projects/gmp/), and the only one in the Iberian Peninsula. The project is an initiative from the Centre for Biodiversity Genomics (CBG), the global leader in the field of DNA barcoding, which is based in the University of Guelph. Currently includes 47 sampling sites across 28 countries, and the program represents the first step towards the acquisition of detailed temporal and spatial information on terrestrial arthropod communities across the globe. It aims to study the immense diversity of terrestrial arthropods distributed across the world, using DNA barcoding as a technique of identification through DNA, and using a standardized sampling technique – the Malaise trap. Due to their easiness to deploy, cost-effectiveness and wide taxonomical scope (insects from various groups are captured), when used in combination with DNA barcoding, Malaise traps allow to carry out large-scale sampling programs and enables a time- and cost-efficient approach for biodiversity assessments around the world. In the scope of this program, animals are collected using Malaise traps, and then DNA sequencing technique is used to produce DNA Barcodes that will allow to identify each species present in the samples. Since January 2018 two sampling sites are active in Portugal, one in the Eurosiberian region and another in the Mediterranean region, where samples are being collected every week for a year period. The samples will be sent to CBG facilities for subsequent processing. It is expected that EnvMetaGen team participation in this initiative will contribute to a giant advance for arthropod knowledge in Portugal. The program moves forward through collaborative efforts in which the EnvMetaGen team is creating a species inventory and barcode library for Portugal. That will support species identification of specimens. The existence of shared Barcode Index Numbers (BINs) between countries provides additional opportunities for concurrent identifications between collections. Sample collection in Portugal will be completed in January 2019 and sequence data is likely to become available for analysis by the end of 2019. The participation in this international project has been focus of media attention: https://cibio.up.pt/media-clippings/details/portuguese-team-integrates-a-global-initiative-to-study-biodiversity-through-dna

4.2.2.2. DNAqua-net

The goal of DNAqua-Net (http://dnaqua.net/) is to nucleate a group of researchers across disciplines with the task to identify gold-standard genomic tools and novel eco-genomic indices and metrics for routine application for biodiversity assessments and biomonitoring of European water bodies. The EU-Water Framework Directive (WFD; 2000/60/EC) and the Marine Strategy Framework Directive (2008/56/EC) legally regulates in the European states the protection, preservation and restoration of aquatic ecosystems and their functions. Jointly with water managers, politicians and other stakeholders, the group of researchers develops a conceptual framework for the standard application of eco-genomic tools as part of legally binding assessments. The assessment of the ecological status of a given water body, implies that aquatic biodiversity data are obtained and compared to a reference water body. Therefore, to be able to implement monitoring plans through DNA analysis it is of high relevance the existence of DNA barcoding reference collections. It is in this context the EnvMetaGen team got involved with the Working Group 1 – DNA Barcode References of the DNAqua-Net Cost Action to promote the completion of a DNA barcode database for European freshwater biota. Furthermore, DNAqua-Net provides a platform for training of the next generation of European researchers, a relevant opportunity for additional capacity building in the context of EnvMetaGen project.

4.3. Public databases

Taxonomic information relating to DNA sequences being generated in this project will be archived in custom databases with links to the sequences and organisms associated with them. All sequences of specimens related to manuscripts already published have been deposited in GenBank and accession numbers are available in the manuscripts. Furthermore, as a first trial of data archive in BOLD, 464 records of Lepidoptera, 140 records of Orthoptera and 11 records of Trichoptera have been deposited in BOLD, where they are currently stored as private records. Nevertheless, this option allows for their use on the Identification tool (http://www.boldsystems.org/index.php/IDS_OpenIdEngine) to any user, since the date of the deposition.

5. FUTURE DIRECTIONS

5.1. Identified priorities and lines of action

Despite the significant advances achieved in building and organising reference collections of DNA sequences within the scope of EnvMetaGen project, continued efforts will be devoted to this task and several aspects have been identified as priorities and future lines of action.

5.1.1. Implementation of the relational database

The fully implementation of the relational database that has been developed to accommodate the data of the InBIO Barcoding Initiative reference collection of DNA sequences is a fundamental step, which will largely contribute for a faster advancement of data curation and reduction of data redundancy. To fully implement the relational database, a preliminary set of data from the database in use will be selected and used for testing, aiming to identify the need of adjustments. Once the tests demonstrate its robustness, the entire dataset will be uploaded and another validation step will follow. After this stage data entry should be done directly in the relational database via the Java front end that will run independently on the individual laptops and desktops of project members. Two levels of access will be given, with administrators being able to modify and add new data, while third party users will be able to view or extract data but not modify it. The relational database is expected to be fully implemented and routinely used by the end of 2019.

5.1.2. Reinforcement of taxonomical sampling scope

Currently, there are 23 orders and 247 families already represented in the reference collection of DNA sequences, though several of these orders have less than half of their diversity represented. This is particularly relevant in hyperdiverse orders such as Coleoptera, Diptera and Hymnoptera, which represent an enormous part of the invertebrate diversity. Therefore, future work regarding this aspect will focus on two aspects: i) the establishment and strengthening of collaborations with taxonomists, and ii) the intensification of the use of sampling techniques that target these taxonomical groups. The establishment and strengthening of collaborations will continue to be made both by open calls to collaborations in scientific meetings and direct contacts with specialists. The intensification of the use of sample techniques that target

hyperdiverse orders will consist in: i) the continuity of the participation in the Global Malaise Program - a key initiative to expand the reference collection; ii) the implementation in 2019 of a sampling network using Malaise traps, pitfall traps and coloured pan traps, in key geographical areas where EnvMetaGen affiliated projects are based, namely in the North-east Portugal (CHASCOS, CRAYFISH, FRESHING, GALEMYS, MANTIDS, SABOR and TUA projects).

5.1.3. Reinforcement of geographic and temporal sampling scope

Biological diversity is not homogeneously distributed along the landscape and when it comes to invertebrates many microhabitat features impact significantly the number of species and taxonomic groups present in a certain area. Sampling should therefore be planned in a way that the highest number of habitats and microhabitats are surveyed in order to yield a representative sample of the species diversity of the areas sampled. In future work the aim is to represent more habitats and microhabitats in the reference collection complementing the fieldwork performed previously. Additionally, in many invertebrate species, adult forms live for a very short period of time in comparison with their larvae. Nevertheless, it is at adult stage that most species have conspicuous morphological characters that allow accurate taxonomic identification. For this reason, it is relevant to sample throughout the year, so all the seasonal species composition is covered and the highest numbers of species becomes represented in the reference collection. These two aspects are being considered during the design of the sampling network using Malaise traps, pitfall traps and coloured pan traps to implement in 2019.

5.1.4. DNA sequence data deposition on public databases

The aim of the work developed so far regarding the building and organising reference collections of DNA sequences is to directly assist EnvMetaGen affiliated projects (*e. g.* freshwater monitoring and diet analysis) providing accurate identification of DNA sequences. Additionally, it is intended to make the DNA sequences generated available in conformity with the H2020 programme's principles of FAIR (findable, accessible, interoperable and reusable) data dissemination and following an Open Research Data (ORD) approach. In this context, the DNA data deposition on public databases as BOLD and GenBank is one of the priority actions for future work. The inclusion of DNA sequences generated within the scope of EnvMetaGen

affiliated projects in BOLD is planned to be done systematically towards the end of 2018 and during all of 2019. Data will be publicly available following the preparation of scientific papers documenting the barcodes for particular taxonomic groups.

5.1.5. Scientific publications

As a result of the work developed in the last three years, some taxonomic groups have now relatively high representation on the reference collection of DNA sequences in terms of species and specimens. Given the substantial data gathered so far, manuscripts on Lepidoptera, Odonata and Trichoptera are currently being prepared and are scheduled to be submitted during 2019. Other taxonomical groups are currently being thoroughly studied, namely Dermaptera, Mantodea, Mecoptera, Neuroptera, Phasmatodea and Rhaphidioptera and the submission of results for publication is expected towards the end of 2019.

6. CONCLUDING REMARKS

This report provides an overview of the current state of the art for the development of best practices for building DNA reference collections of voucher specimens identified by specialised taxonomists, and how these practices are being implemented at InBIO following consistent and repeatable procedures. The implementation and optimization of best-practices for the analyses of eDNA is one of the main goals of EnvMetaGen. Environmental DNA studies require that eDNA sequences are assigned to taxonomical units by comparison with DNA sequences stored in reference collections. Therefore, the construction and organisation of reference collections of DNA sequences is an essential component for any of the key areas of the EnvMetaGen project: Monitoring of freshwater eDNA for species detection; Assessing natural pest control using faecal metagenomics; and Next-generation biomonitoring using DNA metabarcoding. An overview of the challenges in building and organising reference collections of DNA sequences was provided in Section 2. Hence, the EnvMetaGen project, along with its numerous affiliated projects and collaborators, has been developing efforts to identify key taxonomical groups to incorporate in the reference collection, and to do so, in order to fulfil the need of accurate taxa assignation of DNA sequences. Such reference collection is likely to become a tool with significant relevance to the InBIO-Industry-Government triple-helix activities (WP5) by promoting the development of partnerships in all EnvMetaGen key areas. Additionally, it generates valuable knowledge *per se* regarding species taxonomy, ecology and distribution, and also promotes the understanding of the markers characteristics and its methodological limitations as described in section 4.2. This report (Deliverable 4.2) describes best practice protocols for building and organising reference collections of DNA sequences. Together, Deliverables D4.2-D4.5 (this document; Egeter et al., 2018; Galhardo et al., 2018; Paupério et al., 2018) form a detailed account of the successful deployment of a fully functional eDNA lab under the EnvMetaGen project, achieving Task 4.2 and providing a valuable resource for eDNA practitioners in all spheres of the triple-helix model.

7. HOW TO CITE

Ferreira S, Fonseca N, Egeter B, Paupério J, Galhardo M, Oxelfelt F, Aresta S, Archer J, Corley M, Penado A, Pina S, Jarman S and Beja P (2018). Deliverable 4.2 (D4.2): Protocol for building and organising reference collections of DNA sequences, EnvMetaGen project (Grant

Agreement No 668981). European Union Horizon 2020 Research & Innovation Programme - H2020-WIDESPREAD-2014-2 doi: 10.5281/zenodo.2586893

8. REFERENCES

Barbosa S, Paupério J, Searle JB, Alves PC (2013) Genetic identification of Iberian rodent species using both mitochondrial and nuclear loci: application to noninvasive sampling. Molecular Ecology Resources, 13: 43–56.

Egeter B, Fonseca NA, Paupério J, Galhardo M, Ferreira S, Oxelfelt F, Aresta S, Martins FMS, Mata VA, da Silva LP, Peixoto S, Garcia-Raventós A, Vasconcelos S, Gil P, Khalatbari L and Beja P (2018). *Deliverable 4.3 (D4.3): Protocol for field collection and preservation of eDNA samples, EnvMetaGen project (Grant Agreement No 668981). European Union Horizon 2020 Research & Innovation Programme - H2020-WIDESPREAD-2014-2* doi: 10.5281/zenodo.2579806

Galhardo M, Fonseca NA, Egeter B, Paupério J, Ferreira S, Oxelfelt F, Aresta S, Muñoz-Merida A, Martins FMS, Mata VA, da silva L, Peixoto S, Garcia-Raventós A, Vasconcelos S, Gil P, Khalatbari L, Jarman S and Beja P (2018). *Deliverable 4.5 (D4.5): Protocol for the processing of DNA sequence data generated by next-gen platforms, EnMetaGen project (Grant Agreement No 668981). European Union Horizon 2020 Research & Innovation Programme – H2020-WIDESPREAD-2014-2* doi: 10.5281/zenodo.2586889

Hsi-Yang Fritz M, Leinonen R, Cochrane G, Birney E (2011) Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Research*, 21(5): 734–740. http://doi.org/10.1101/gr.114819.110

IUCN 2014. IUCN Red List of Threatened Species, 2014.3. Summary Statistics for Globally Threatened Species. Table 1: Numbers of threatened species by major groups of organisms (1996–2014).

Muir P, Li S, Lou S, Wang D, Spakowicz D J, Salichos L, Zhang J, Weinstock GM, Isaacs F, Rozowsky J, Gerstein M (2016). The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biology*, 17, 53. http://doi.org/10.1186/s13059-016-0917-0

Oliveira R, Castro D, Godinho R, Luikart G, Alves PC (2010) Species identification using a small nuclear gene fragment: Application to sympatric wild carnivores from South-western Europe. *Conservation Genetics*, 11: 1023-1032.

Paupério J, Fonseca N, Egeter B, Galhardo M, Ferreira S, Oxelfelt F, Aresta S, Martins F, Mata V, Veríssimo J, Puppo P, Pinto JC, Chaves C, Garcia-Raventós A, Peixoto S, da Silva LP, Vasconcelos S, Gil P, Khalatbari L, Jarman S and Beja P (2018). *Deliverable 4.4 (D4.4):*

Protocol for next-gen analysis of eDNA samples, EnMetaGen project (Grant Agreement No 668981). European Union Horizon 2020 Research & Innovation Programme – H2020-WIDESPREAD-2014-2 doi: 10.5281/zenodo.2586885

Penado, Andreia de Barros Mendes (2018) *The effects of climate and land abandonment on Iberian bees.* Doctoral thesis (PhD), University of Sussex.

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A.; 74:5463–5467.

Silva T, Godinho R, Castro D, Abáigar T, Brito JC, Alves PC (2015) Genetic identification of endangered North African ungulates using noninvasive sampling. *Molecular Ecology Resources*, 15: 652-661.

Vernooy R, Haribabu E, Muller MR, Vogel JH, Hebert PDN, Schindel DE, Shimura J, Singer GAC (2010) Barcoding life to conserve biological diversity: beyond the taxonomic imperative. PLoS Biology 8: e1000417

APPENDIX: DESCRIPTION OF ENVMETAGEN-AFFILIATED PROJECTS

This section provides a description of current EnvMetaGen-affiliated projects. At present, there are 20 ongoing EnvMetaGen-affiliated projects. Through the development of field, laboratory and data analysis pipelines, each of the projects contributes to the deployment of an eDNA Lab, which is the main goal of Work Package 4 and the focus of Deliverables 4.2 to 4.5 (this document; Egeter et al., 2018; Galhardo et al., 2018; Paupério et al., 2018).

All of the projects are highly collaborative involving a total of six other InBIO research groups, five research groups from other Portuguese institutions and fourteen overseas research groups. Twelve of the projects are being led by the EnvMetaGen team. These collaborations build relationships with key national and international organisations and networks in the environmental area, fostering the establishment of long-term partnerships with leading research institutions, helping to fulfil the objectives of Work Package 3 *Development of Capacities to Participate in the ERA.*

All projects are within the focus of one or more of the three key areas being developed under the triple-helix model of innovation (WP5):

1. Monitoring of freshwater eDNA for species detection

2. Assessing natural pest control using faecal metagenomics

3. Next-generation biomonitoring using DNA metabarcoding

The applicability of each project to EnvMetaGen Work Packages and Objectives is highlighted. Overall, the projects´ contributions to the deployment of an eDNA Lab, by developing analyses within the scope of the triple-helix key areas, as well as fostering networks among institutional, national and international collaborators, substantially increase InBIO´s capacity for research and innovation using environmental metagenomics.

AGRIVOLE

The role of voles in agroecosystems: linking pest management to biodiversity conservation under environmental change

Agroecosystem services are being threatened worldwide by biodiversity loss. Biological pest management is one of the main ecosystem services often supported by agroecosystems, as non-

crop habitats can provide resources for species that may act as natural controllers of agricultural pests, responsible for huge losses in crop yields. However, there is still limited understanding on how biodiversity levels relate with biological control, particularly considering current trends in agricultural land use change. AGRIVOLE project aims to assess the responses of vole communities to agroecosystem structure and management practices, by combining ecological tools and high throughput DNA sequencing techniques. The project will analyse the effects of different population regulatory processes and evaluate how community responses may affect the potential for pest outbreaks or impact the resilience of vole species of conservation concern. The focus will be on the vole community of northeastern Portugal agroecosystems, a species rich system where vole pests have significant economic impact on fruit tree orchards. The project will use data previously collected on voles' distribution in the region, complemented with detailed plant and vole surveys across agroecosystems with different structures and management treatments. We will also use high-throughput sequencing techniques, namely DNA metabarcoding, to determine voles' trophic niches based on their droppings. Overall, it is expected that the results obtained in this project contribute significantly to foster sustainable agricultural techniques linking pest management to biodiversity conservation. This project begun recently, but its progress will boost the development of the laboratory methods for analysing herbivore diets, using a metabarcoding approach, as well as the methods for collecting and analysing soil samples for determining plant diversity. Moreover, this project involves a collaboration with the University of Natural Resources and Life Sciences, Vienna, for building a reference collection for plants using high throughput sequencing, fundamental for the diet studies and vegetation surveys. Therefore, this project will contribute significantly for building capacity on the eDNA analyses in InBIO, while expanding its network of collaborations (WP3). AGRIVOLE is aligned with one of the key application areas of EnvMetaGen, *Assessing natural pest control using faecal metagenomics*, and it is expected that it provides relevant outcomes for practical applications in crop management. This may lead to the development of services, relevant to the farmers and Regional Agricultural Institutions, thereby fostering the triple helix (WP5).

AZORES

Assessing fish diversity in Azores freshwater lagoons using a metabarcoding approach

Eutrophication is a relevant issue for water quality in lagoons and is considered one of the main environmental problems in the Azorean archipelago, with high impacts on landscape, economy and the conservation of natural resources. Landscape changes and anthropogenic activities in general are considered as the main causes for eutrophication, and the lagoons in the island of São Miguel, are considered a good example of this situation, where land use changes have been associated with water quality degradation. Water quality of the Azorean lagoons has been monitored since 2003, and within this frame the development of efficient and cost-effective methods for monitoring biodiversity in the lagoons has become highly relevant. This project aims at developing a cost-effective monitoring program for fish diversity in the Azores freshwater lagoons. The main goal is the optimization of field and laboratory protocols for assessing the diversity of fish communities from environmental samples, using a metabarcoding approach. Samples have been collected by the University of Azores InBIO team, using both water filtering and precipitation techniques. The data is helping to refine best practices in collecting eDNA samples from water, while the optimisation of extraction and amplification protocols contribute to the development of capacities at InBIO. This project is aligned with the one of the key application areas of EnvMetaGen, *Next-generation biomonitoring using DNA metabarcoding*, and it is expected that it will help progress monitoring programs for fish diversity in freshwater ecosystems. The developed methodology is of relevance for the Regional Government of Azores, and applicable to other areas, with potential for application by other regional institutions and companies, thereby fostering the triple helix (WP5), and contributing to the expansion of InBIO´s collaboration network.

CHASCOS

Diet analysis of black wheatears (*Oenanthe leucura*)

The black wheatear (*Oenanthe leucura*) is the most threatened passerine in Portugal. Its distribution used to range from the Portuguese coast to the French Pyrenees. Nowadays it is extinct in France, while in Portugal it is restricted to the remote inner Douro and Tagus valleys, and in Spain its population decreased more than one third in recent years. To help understand the reasons for this severe decline, this project aims to study in detail the diet of this threatened

bird. High throughput sequencing techniques have been shown to be able to characterise the diet of several animals in unprecedented detail. However, to study the diet of passerines and other large feeding spectrum animals is challenging for metabarcoding techniques due to several constraints, such as molecular marker selection and secondary predation detection. High throughput sequencing is being used on droppings from captured birds in the Douro valley. As well as using traditional morphological analysis, several commonly used molecular markers are being used. All the information obtained from the molecular markers and the morphological identification are being compared. This has allowed the detailed description of the feeding requirements of the black wheatear, and given the observed large feeding spectrum and plasticity found, it has become apparent that it is unlikely that its decline is directly related to shortage of food. The project also identified the main problems and biases of some of the most commonly used molecular markers used in metabarcoding diet studies, and allowed for the development of techniques to minimize these problems. The project focuses on protecting biodiversity (identified as a societal challenge to be tackled by InBIO, EnvMetaGen Objectives) thereby contributing to the triple-helix initiatives (WP5). It focuses on identification of critical food resources for endangered species (identified as an emerging eDNA research line, EnvMetaGen Objectives). By comparing diet analysis protocols and molecular markers, it contributes substantially to the development of an eDNA lab by making technical advancements that have implications for eDNA best practices (WP4) and help to build capacity at InBIO.

CRAYFISH

Assessing the impact of invasive crayfish through diet analysis

The invasion of freshwater ecosystems by exotic species is a cause of concern worldwide due to their negative environmental and economic impacts. Invasive crayfish are one of the most detrimental alien species occurring in European freshwater ecosystems. Among the known, negative effects are bioturbation, competition with native species, predation on native biodiversity, effects on leaf and algae abundance, and trophic subsidizing for predators (which in turn can enhance predation on native species). To adequately assess the impact of these species, including their potential overlap with the trophic niche of native, threatened fauna, and provide information on their control and management, knowledge of their trophic ecology is essential. This project aims to characterize the diet of two invasive crayfish species in Northern Portugal (*Procambarus clarkii* and *Pacifastacus leniusculus*) using metabarcoding. As both

species are thought to have a varied generalist diet, the project will involve conducting assays targeting a number of mitochondrial metabarcoding markers across multiple prey groups. The project will provide high resolution diet information for improved management of these invasive species, which pose a widespread global threat to biodiversity. It should be noted that this project is in the early stages of development, and as such detailed protocols are not provided in these deliverables. The project will focus on biodiversity conservation and invasive species control (identified as an emerging eDNA research line, EnvMetaGen Objectives), producing data to inform governmental management for protecting biodiversity (identified as a societal challenge to be tackled by InBIO, EnvMetaGen Objectives) thereby contributing to the triple-helix initiatives (WP5). The project already has an associated InBIO MSc student, who will receive training in metagenomic techniques, helping to build InBIO's capacity (WP4).

ECOLIVES

Fostering sustainable management in Mediterranean olive farms: pest control services provided by wild species as incentives for biodiversity conservation

Efficient pest management is recognized as a major challenge for fostering economically profitable agroecosystems worldwide. Biocontrol services provide clear incentives for biodiversity conservation in agroecosystem as naturally occurring species can efficiently reduce populations of pests, thus reducing both crop losses to pests and the need for agrochemicals. Yet, the ecology of biocontrol services is poorly known, thus limiting our ability to understand its value and to plan their conservation and management. Using Mediterranean olive farms as case study, the overarching research goal of this project is to estimate the value of natural biological control of the Olive fruit fly (*Bactrocera oleae*) and the Olive fruit moth (*Prays oleae*) –the two major pests in olive farms worldwide–, in farms following distinct pest management strategies. The overall hypothesis is that the abundance and diversity of biocontrol providers will decline with increasing pest management intensity, which will be expressed in a non-negligible economic impact. Specifically, the project will focus on predatory insects (parasitoid wasps) as well as insectivorous vertebrates (birds and bats) as biocontrol providers. This is particularly relevant because, although birds and bats are thought to provide high levels of pest suppression, knowledge about their role as biocontrol providers is negligible compared to insect predators in Mediterranean olive farms in particular and in agroecosystems worldwide in general. The hypothesis will be tested by quantifying occurrence and abundance patterns

both of biocontrol providers and insect pests in 2 olive farms following distinct types of pest management strategies: IPM (Integrated Pest Management), where producers apply agrochemicals when pest populations reach the economic threshold; and organic, where producers rely completely on biocontrol services. The relative importance of each biocontrol provider on levels of pest infection will be investigated, and their economic value calculated. The data obtained at this local scale will be used to model potential scenarios of biocontrol services provision in olive farms at the whole Iberian Peninsula, with the aim to select priority conservation-management in the face of global environmental change. This project is based in Évora University and the EnvMetaGen team will participate on the development of molecular tools to identify prey items of key predators/parasitoids present in olive farms and to perform diet analysis. The project is likely to provide data to assist farmers finding better solutions to pest control than using high loads of pesticides. This project is of high relevance to existing and future InBIO-Industry-Government triple-helix initiatives (WP5), as it uses faecal eDNA samples to assess natural species as a form of pest control, addressing the provision of ecosystem services (identified as a promising eDNA research theme, WP2). The associated InBIO PhD student will receive training in metagenomic techniques, helping to boost InBIO's capacity (WP4).

FILTURB

Comparing methods to filter turbid water and modelling site occupancy based on eDNA detections

eDNA survey methods have been applied mainly in freshwater ecosystems, focusing on water without a high sediment load. This is largely due to difficulties with sampling suitable volumes of turbid water. One of the objectives of this project is to test the efficiency of different DNA capture methods in turbid waters, evaluating their performance on eDNA recovery and species detection. The project will compare the most common filtering and DNA precipitation methods with newer high-capacity filtering approaches. The latter have the potential to filter much higher volumes of water than the former, even in turbid environments. Using the information from this objective a second aspect of eDNA sampling will be investigated: modelling site occupancy based on eDNA detections. Once shed into the environment, the probability of detecting DNA of a target species will vary depending on environmental factors. By collecting eDNA samples multiple times at many sites, the probability of detection of amphibians will be estimated using

site occupancy models. This will inform future studies on the number of samples that are required to detect a given species. The project is focussed on making technical advancements for cost-effective species detection and biodiversity assessment, contributing to existing and future triple-helix initiatives in different areas (WP5). By comparing existing and emerging protocols, it will also help to implement best practice protocols for eDNA analysis (WP4). The project already has an associated InBIO MSc student, who will receive training in eDNA sampling and metagenomic techniques, helping to boost InBIO's capacity (WP4). This project is closely linked with GUELTA.

FRESHING

Next-generation biomonitoring: freshwater bioassessment and species conservation improved with metagenomics

Data collection of freshwater habitats is essential, allowing countries to fulfil legislation requirements, such as the European Union Habitat and Water Framework directives. However, collecting biotic data for freshwater monitoring implies extensive effort. This project aims to investigate the value of using latest metagenomic approaches and applied ecological tools to improve freshwater bioassessments and detection of species of conservation concern, and ultimately optimize monitoring programs. Objectives include: 1) developing metagenomic approaches to obtain reliable biodiversity data and species detections; 2) building metagenomic multimetric indexes for bioassessment of ecological quality; 3) validating rapid landscape predictions for monitoring bioassessment indices, and threatened and invasive species; and 4) designing a next-generation biomonitoring framework for freshwaters for an early warning system to alert authorities. The project will focus on fishes and macroinvertebrates, in the Douro Basin (North Portugal), because they are informative freshwater indicators and include many species of conservation concern. Ultimately, the project will use decision making and conservation tools to perform a cost-efficiency analysis, and design a framework for next-generation monitoring programs in freshwaters. The project is focussed on making technical advancements for cost-effective species detection, biodiversity assessment and biomonitoring. It will have implications for biodiversity conservation and invasive species control, contributing to the triple-helix initiatives (WP5) and the development of an emerging eDNA research line (EnvMetaGen Objectives), producing data to inform governmental management for protecting biodiversity (identified as a societal challenge to be tackled by InBIO, EnvMetaGen

Objectives). The project tackles the pressing societal challenge of the loss of biodiversity (EnvMetaGen Objective). The project has an associated InBIO PhD student, who will receive training in metagenomic techniques, and will include the comparison of existing and emerging protocols, helping to boost InBIO's capacity (WP4).

GALEMYS

Conservation genetics of a threatened semi-aquatic mammal: The Iberian desman (*Galemys pyrenaicus*) in northeast Portugal

The Iberian desman (*Galemys pyrenaicus*) is a threatened, elusive mammal endemic of the Iberian Peninsula and the Pyrenees. In Portugal, the species is restricted mostly to the North of the country and a recent survey revealed a marked reduction in the species distribution in Northeast Portugal. Besides the paucity of distributional data, baseline information relative to the ecology, genetic diversity and structure in Portugal is also scarce. However, this knowledge is crucial for understanding how river connectivity shapes the species ecology, particularly considering the threat posed by the recent construction of large hydroelectric infrastructures. Therefore, this project aims at determining the degree of genetic diversification and structuring of the desman population in Portugal and examining how species traits and trophic requirements together with river connectivity and other landscape features influence the species persistence in fragmented areas. This information is vital for an efficient conservation of this endangered, poorly known, semiaquatic mammal. For achieving this main goal, a set of microsatellites is being optimized using high throughput sequencing (HTS) for analysing the population genetic structure and diversity with tissues and non-invasive samples (faeces). Moreover, faeces collected in two river basins are being analysed using metabarcoding for assessing the species trophic niche in the study area. Therefore, this project is contributing for building capacities at InBIO, namely for the optimization of methods for genotyping microsatellites using HTS and for refining best practices in the diet analyses of insectivores using metabarcoding. GALEMYS project is related with one of the key application areas of EnvMetaGen, *Next-generation biomonitoring using DNA metabarcoding*, as it is expected that the results obtained with this project will help define conservation actions for this endangered species. Therefore, we expect this project to contribute with relevant information to the Portuguese administration strengthening the relation between InBIO and administration (WP5).

GUELTA

Assessing vertebrate diversity in turbid Saharan water-bodies using environmental DNA

The Sahara Desert is the largest warm desert in the world and a poorly-explored area. Small water-bodies occur across the desert, which are crucial habitats for vertebrate biodiversity, as well as providing resources for local human activities. The long-term conservation of these habitats requires a better assessment of local biodiversity and potential human-related conflicts. There is potential to use eDNA for monitoring vertebrate biodiversity in these areas. However, there are a number of difficulties with sampling eDNA from such turbid water-bodies and it is often not feasible to rely on electrical tools in remote desert environments. This project is trialling novel, manually-powered, water filtering methods in Mauritania to obtain eDNA samples. The project is focussed on making technical advancements for cost-effective biodiversity assessment, contributing to triple-helix initiatives in identified key areas (WP5), in poorly explored regions (identified as a promising eDNA research theme, WP2). As well as contributing to the deployment of an eDNA lab, it provides training for InBIO researchers as it involves the investigation and comparison of multiple field eDNA sampling methods (WP4). This project is also closely linked to FILTURB.

ICVERTS

Providing an eDNA tool for rapid assessment of ecological integrity through detection of rare indicator species in Western Africa

This project focuses on the detection of two iconic West African wetland species as bio-indicators: the Critically Endangered West African slender-snouted crocodile (*Mecistops cataphractus*) and the Endangered pygmy hippopotamus (*Choeropsis liberiensis*). The goal of the project is to assess whether an eDNA approach can provide a rapid assessment tool of ecological integrity by detecting the presence of these important indicator species. Such a tool would greatly reduce manpower and costs associated with traditional survey methods. High sensitivity qPCR species-specific assays have been developed to detect the DNA of these two high-value species. Water samples were collected throughout protected areas of Cote d'Ivoire, the last strongholds for these species in the Upper Guinea forests of West Africa. Although qPCR is often regarded as the most sensitive method of species detection, there is a current ideological shift towards the idea that metabarcoding methods may in fact detect rare species

in eDNA samples with a similar efficacy. The project will compare both approaches of species detection. The project is focussed on developing biodiversity assessment tools, contributing to triple-helix initiatives in identified key areas (WP5), in a poorly-explored tropical region (identified as a promising eDNA research theme, WP2), to be used by researchers and government for protecting biodiversity (identified as a societal challenge to be tackled by InBIO, EnvMetaGen Objectives).

IBI

InBIO Barcoding Initiative

DNA barcoding is an essential tool in a vast array of ecological and conservation studies. With the advent of Next Generation sequencing, it became possible to implement diet analysis and monitoring methods based on DNA metabarcoding. While such studies can include a range of environmental DNA sample types, such as faeces, saliva, blood meal, stomach contents, hair, water, air, pollen/natural by-products (e.g. honey), soil, bulk samples (or preservative), all demand the availability of a reference collection of DNA sequences in order to allow the correct identification of taxa found in each sample. Therefore, its applicability is hampered by the lack of comprehensive reference collections, particularly of invertebrates that are underrepresented in reference databases and this knowledge gap becomes greater in biodiversity hotspots. During the early stages of the EnvMetaGen project conception the need of developing a reference collection of DNA sequences for Portuguese invertebrates was identified and for this reason the Task 4.2. - Building capacity for eDNA analysis includes the construction and organisation of reference collections of DNA sequences as one of the pivotal capacity-building aspects. The InBIO Barcoding Initiative consists in the development of a DNA reference collection of voucher specimens identified by specialised taxonomists following the best practices, which is essential to develop and conduct consistent, reliable and repeatable research studies boosting the future performance of InBIO in environmental genomics. By combining field work and networking with taxonomists and ecologists, the project aims to produce DNA barcodes for thousands of species, covering over one hundred families of insects. The reference library will be a fundamental tool for long-term and large scale monitoring programs in Portugal and serve as base for ecological studies related with loss of biodiversity, degradation of ecosystem services, and sustainable development (EnvMetaGen Objectives) and to promising eDNA research themes (WP2). Along its construction the project contributes for the training in

taxonomy and metagenomic techniques, helping to boost InBIO's capacity (WP4). Furthermore, it is likely to become a tool with significant relevance to the InBIO-Industry-Government triple-helix initiatives (WP5) by promoting the development of partnerships in all key areas: Monitoring of freshwater eDNA for species detection; Assessing natural pest control using faecal metagenomics; and Next-generation biomonitoring using DNA metabarcoding.

IRANVERTS

Assessing diet of large felids in central deserts of Iran

Information on population structure, hormones, parasites and diets can all be produced using non-invasive faecal samples. Such information is highly valuable for conservation of elusive species such as Asiatic cheetah (*Acinonyx jubatus venaticus*). For this project scat samples are being collected from large carnivores across the distribution range of Asiatic cheetah. Using metabarcoding, scats will firstly be assigned to the predator species and secondly used to assess the diets of large felids. Two different extraction methods are being trialled to test for their efficacy in producing DNA suitable for predator species identification. Extracted DNA will be subject to PCR using a number of vertebrate-targeting PCR primers. Possible prey items include wild sheep (*Ovis orientalis*), wild goat (*Capra aegagrus*), gazelles (*Gazella bennettiii* and *Gazella subgutturosa*) and domestic livestock. This project is of relevance to the agricultural industry sector as well as for conservation of a threatened species, contributing to two key areas targeted for triple-helix initiatives (WP5). It tackles the pressing societal challenge of sustainable development (EnvMetaGen Objective) and includes assessment of habitat loss on trophic interactions in human-modified landscapes and management of wild and domestic herbivores (identified as promising eDNA research themes, WP2).

MANTIDS

Diet analysis of mantids

Modern molecular techniques have made it possible to assess species composition of complex samples, almost independently of individual density. In the last decades, DNA Metabarcoding together with High Throughput Sequencing (HTS) has allowed for diet assessment in several groups of animals, including insects. Although major developments have been made for assessing vertebrate diets using metabarcoding, it is the field of invertebrate ecology that has

largely pioneered research in this area of molecular ecology. One of the reasons for this is that many invertebrates either heavily masticate their prey or are fluid feeders, precluding morphological analysis. This EnvMetaGen-affiliated project aims to utilise metabarcoding methods to characterise the diet of selected species of mantids in Portugal. Mantids (Order: Mantodea) are highly-adapted predatory insects. Their diet is thought to be varied but no DNA-based assessment has been performed so far. This project will assess mantid diets in nature, through the collection of mantid faecal samples, focussing on their potential as agricultural pest controllers. This exploratory project might prove to be of high relevance to the InBIO-Industry-Government triple-helix activities (WP5), as it uses faecal eDNA samples to assess natural species as a form of pest control, addressing the provision of ecosystem services (identified as a promising eDNA research theme, WP2). The associated InBIO master student, will receive training in metagenomic techniques, helping to boost InBIO's capacity (WP4).

MATEFRAG

Impacts of habitat fragmentation on social and mating systems: testing ecological predictions for a monogamous vole through non-invasive genetics

Intensification of agriculture has caused severe loss and fragmentation of semi-natural habitats worldwide. Studies of the effects of habitat fragmentation on biodiversity have revealed large impacts on species distribution and abundance patterns. However, understanding demographic and behavioural processes that determine species vulnerability to fragmentation is important to properly understand population viability in human-dominated landscapes. Key, relevant, within-population processes affecting reproductive success and thus population persistence include social interactions, mating systems, and the formation of Kin-structures. In this project we aim to assess the effects of habitat fragmentation on mammalian social and mating systems, and how this affects population persistence. As it is expected that monogamous species are more susceptible to stochasticity and prone to extinction events, we have focused this project on a monogamous Iberian endemic mammal, the Cabrera vole (*Microtus cabrerae*). To achieve this main goal, this project is using genetic non-invasive sampling (faeces) for individual identification and for estimating kin-structure. The methods being used for species and individual identification from faeces were already optimized at InBIO (see Paupério et al. 2018 for details), hence this project has provided a relevant contribution in capacity building of eDNA (WP4).

NZFROG

Determining the impact of invasive mammals on frogs in New Zealand

Since the arrival of mammals, New Zealand's endemic frogs (*Leiopelma* spp.) have undergone a number of species extinctions and range contractions. Only two species now persist on the mainland. One of these, *Leiopelma archeyi*, is Critically Endangered and ranked as the world´s most evolutionarily distinct and globally endangered amphibian. Ship rats (*Rattus rattus*) have often been implicated in the decline of amphibians in New Zealand and worldwide, but prey from rodent stomach contents are notoriously difficult to identify. This project utilises metabarcoding to survey for predation by ship rats on the remaining mainland *Leiopelma* species. New PCR primers were developed that target all anuran species. This study has provided the first evidence of these frog species in mammalian stomach contents and this, along with evidence from other studies, has led to the the New Zealand government including certain important sites in their rodent control program. It should be noted that field samples for this project were collected as part of a separate project and as such the field collection protocols are not explicitly detailed, but the treatment of the eDNA samples and subsequent data are included in Paupério et al. 2018 and Galhardo et al. 2018. The project focuses on biodiversity conservation and invasive species control, contributing to the triple-helix initiatives (WP5) and an emerging eDNA research line (EnvMetaGen Objectives), producing data to inform governmental management for protecting biodiversity (identified as a societal challenge to be tackled by InBIO, EnvMetaGen Objectives). It also contributes to the deployment of an eDNA lab (WP4) by providing a new and validated primer set.

SABOR

Assessment of the role of bats as pest regulators in Mediterranean agriculture

Small vertebrate insectivores are judged to provide important ecosystem services by controlling insect pests. Bats, in particular, are major insect predators, suggesting that they play a vital role in protecting crops from pests. However, there’s a lack of basic information regarding bats’ diet and foraging behaviour. Traditional diet analyses use visual identification of arthropod fragments present in faecal or stomach contents, and are limited to order or family level identifications, not allowing the identification of possible pest species. When species level identifications are possible, these are usually restricted to hard-bodied insects, like Coleoptera.

Recently, with the advancement of molecular methods, it became possible to identify at the species level both hard and soft-bodied insects, present in bat guano. In particular, the emergence of HTS techniques allows the barcoding of multiple insect species in complex samples – metabarcoding. These novel methods are revolutionizing dietary studies and can give us precious insights into the role of bats as pest regulators. This project consists of a PhD thesis and aims to answer the following questions: i) What's the diet of a Mediterranean bat community? ii) How do bats group in terms of diet composition? iii) Is there a relationship between bat diet and bat/insect traits? IV) Which bats prey on pest insects and how often? This study will help enlightening the role of bats as pest regulators in Mediterranean agricultural fields. This will not only promote bat populations, but also help farmers finding better solutions to pest control than using high loads of pesticides. This project is of high relevance to develop InBIO-Industry-Government triple-helix initiatives (WP5), as it uses faecal eDNA samples to assess natural species as a form of pest control, addressing the provision of ecosystem services (identified as a promising eDNA research theme, WP2). The associated InBIO PhD student, has been receiving training in metagenomic techniques, helping to boost InBIO's capacity (WP4).

SOILPHOS

Assessing diversity of phosphorus-cycling bacteria in response to fertiliser treatments

Phosphorus is essential to crop and pasture growth and is added to soil in large volumes around the world. However, phosphorus is a scarce, finite resource with peak phosphorus expected as early as 2030 and high-quality rock phosphate estimated to be exhausted within 80 years. It has long been established that bacteria are involved in making phosphorus available to plants, but only recently have DNA-based technologies developed enough to study 1) bacterial soil community and 2) the prevalence of 'phosphorus-freeing' genes in the soil. The aim of this project is to investigate the prevalence and diversity of phosphorus-freeing genes in soil experimentally subjected to various phosphorus levels. The objective is to inform practitioners and researchers as to whether the global community should be trying to foster certain bacterial communities that will allow us to continue food production at its current rate whilst lowering the amount of phosphorus currently applied to agricultural land. This project is of high relevance to develop InBIO-Industry-Government triple-helix initiatives (WP5) as well as tackling the pressing societal challenge of sustainable development (EnvMetaGen Objective)

and addressing the provision of ecosystem services (identified as a promising eDNA research theme, WP2). It should be noted that eDNA sampling and PCRs for this project were part of a separate project and as such are not explicitly detailed, but the data processing is included in Galhardo et al. 2018.

TUA

Promotion of ecosystem services in the Vale do Tua Regional Natural Park: Control of agricultural and forest pests by bats

The Vale do Tua Regional Natural Park (PNRVT) is an excellent example of the natural and patrimonial values that exist in the northern region of Portugal. Here the landscape is dominated by a mosaic of natural and semi-natural vegetation and agricultural areas with predominance of vineyards, olive groves and cork oak forests. Thus, as in other regions of the interior of Portugal, the region's economy is very dependent on agricultural productivity. In this context, one of the most relevant Ecosystem Services (ESs) potentially provided by biodiversity in the region may be the control of agricultural and forestry pests. Due to the high diversity of birds and bats in the region, it is expected that these groups may have great relevance in the provision of these ESs. Several studies have shown that large numbers of these flying vertebrates associated with high prey consumption (mostly insects) make birds and bats one of the most significant natural controllers of agricultural and forest pests populations, thus providing a high economic value, reduced use of pesticides and increased productivity. Therefore, this project aims to create conditions for the intensification of the provision of pest control services (identified as a promising eDNA research theme, WP2) by promoting the populations of the respective predators, focusing essentially on bats. In order to increase the number of bat colonies in the areas of interest, shelter boxes were placed in the most important agricultural and forestry systems in the PNRVT area, specifically vineyards, olive groves and cork oak forests. The evaluation of the effectiveness of this measure will be done by analysing the diet of bats in the shelters, checking which bat species are using the shelters and if they consume (and when) the existing agricultural and forest pests in the region. This project is a prime example of an InBIO-Industry-Government triple-helix initiative (WP5), as it involves stakeholders from administration (the Agency for Regional Development of the Tua Valley, in charge of the management of the park), academia (InBIO) and industry (landowners within the geographical limits of the park). Its results will allow the development of management plans optimizing the

ESs provided by bats in the region, giving an example where the promotion and preservation of biodiversity will translate into economic gains for the stakeholders involved, thus waiting for the PNRVT's management model to be disseminated at the regional and national levels, fostering sustainable development (EnvMetaGen Objective).

WOLFDIET

Describing the diet of African golden wolf (*Canis anthus*) and assessing human conflict

The African golden wolf (*Canis anthus*), previously considered as Golden jackal (*Canis aureus*), is now recognized as a new canid species occurring in North and East Africa. There is a lack of knowledge regarding most of the ecological traits of this medium-sized canid, particularly regarding feeding ecology. African wolves are reported as generalist feeders, consuming plants, insects and vertebrates, including livestock and poultry which raise important conflicts with humans. However, the few available studies are based on identification of macro-components found in scats rarely genetically validated, which may bias the results and underestimate some prey items. Based on 150 scats of African wolves collected in NW Senegal (comprising Djoudj National Park and a neighboring agricultural area) already available and genetically identified in a scope of another InBIO project, this study aims to adequately characterize the diet of African wolves using metabarcoding. The project will involve targeting metabarcoding markers across multiple prey groups and a methodological assay involving two different extractions performed for each scat. By using a high resolution approach, this project is expected to assess the diet of African wolves and their potential impact on threatened fauna (e.g. breeding and migrating birds) and domestic animals, providing essential information for an efficient management. This project is of relevance to the agricultural industry sector as well as for conservation of a threatened species, contributing to key areas identified for triple-helix initiatives (WP5). It tackles the pressing societal challenge of sustainable development (EnvMetaGen Objective) and includes assessment of habitat loss on trophic interactions in human-modified landscapes and management of wild and domestic herbivores (identified as promising eDNA research themes, WP2).

XENOPUS

Detecting the presence of invasive frogs (*Xenopus laevis*) in Portugal

The African clawed frog (*Xenopus laevis*) is a species that has been introduced to many parts of the world. Invasions are due to both accidental escape and voluntary release of laboratory animals in many cases. The predatory impacts of *X. laevis* on native populations of amphibians and fish have been well documented. The species has been implicated in the global transmission of disease including chytridiomycosis, a disease cited as one of the principal causes for the global decline in amphibians. Under a protocol established between Portugal´s governmental conservation agency (ICNF), the Environmental Biology Centre of the Faculty of Sciences of the University of Lisbon and the Gulbenkian Institute of Science, a plan was developed that aims to control *X. laevis*. In order to assess whether the control protocol is effective, an eDNA experiment was planned which aims to detect *X. laevis* at sites where the species is present, sites where it has never been observed and sites where populations have been the subject of the control protocol. The aim is to simultaneously provide a reliable species detection tool and assess the efficacy of current control protocols. This project involves all three groups of the InBIO-Industry-Government triple-helix model (WP5). It focusses on invasive species detection and control (identified as an emerging eDNA research line, EnvMetaGen Objectives) as well as tackling the pressing societal challenge of the loss of biodiversity (EnvMetaGen Objective) and addressing the provision of ecosystem services (identified as a promising eDNA research theme, WP2).