


EnvMetaGen

Deliverable 4.5 (D4.5) Protocol for the processing of DNA sequence data generated by next-gen platforms

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

All intellectual property rights are owned by the EnvMetaGen consortium members and protected by the applicable laws. Except where otherwise specified, all document contents are: "© EnvMetaGenproject". This document is published in open access and distributed under the terms of the Creative Commons Attribution License 3.0 (CC-BY), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.



TABLE OF CONTENTS

SUMMARY	4
1. INTRODUCTION	5
1.1. The context.....	5
1.2. Overview of eDNA next-gen data processing.....	7
1.3. Overview on the deployment of next-gen data processing at InBIO.....	9
1.4. Structure of the report.....	10
2. GENERAL WORKFLOW FOR PROCESSING eDNA METABARCODING DATA	11
2.1. Introduction.....	12
2.2. Data filtering.....	13
2.2.1. Demultiplex	13
2.2.2. Error correction	14
2.2.3. Merge read pairs	14
2.2.4. Quality filtering	15
2.2.5. Chimera filtering	15
2.2.6. Dereplication	15
2.3. Sequence classification.....	16
2.3.1. Clustering sequences based on similarity	16
2.3.2. Querying reference databases	16
2.3.3. Filtering reference database query results for taxonomic assignment	17
2.3.4. Subsequent analyses	17
2.4. Other considerations for eDNA metabarcoding data processing.....	17
3. ENVMETAGEN PROTOCOL FOR PROCESSING eDNA METABARCODING DATA	18
3.1. Introduction.....	18
3.2. Protocol used for the processing of metabarcoding data.....	19
3.2.1. Data filtering	19
3.2.2. Taxonomic assignment	20
4. ENVMETAGEN PROTOCOL FOR PROCESSING InBIO BARCODING INITIATIVE (IBI) DNA DATA	22
4.1. Introduction.....	22
4.2. Protocol for processing IBI DNA sequence data.....	22
4.2.1. Data filtering	23
4.2.2. Curating DNA sequences for reference collection	23
5. OTHER DEVELOPING AREAS FOR THE PROJECT	24
6. CONCLUDING REMARKS	25
7. CONTRIBUTING AUTHORS	26
8. REFERENCES	27
APPENDIX A: DESCRIPTION OF ENVMETAGEN-AFFILIATED PROJECTS	34
APPENDIX B: EnvMetaGen CURRENT PROTOCOLS FOR NEXT-GEN DATA PROCESSING	51

SUMMARY

The overall goal of the 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 to the development of research projects in association with business partners and other stakeholders in key application areas identified in the project, and thus to the strengthening of InBIO triple-helix initiatives (InBIO-Industry-Government; WP5).

This report (Deliverable D4.5) builds upon previous ones (Deliverables D4.2-D4.4, respectively Ferreira et al. (2018), Egeter et al. (2018), Paupério et al. (2018)) and provides an overview of the processing protocols for DNA sequence data generated by next-gen platforms within EnvMetaGen-affiliated projects. Deliverables D4.2-D4.5 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. This development was made possible through the recruitment of the ERA Chair team (WP2), secondments and Junior Researcher exchanges through the collaboration with international networks (WP3), an enhancement of computational infrastructure at InBIO (WP4) and participation of team members in workshops and conferences (WP6).

1. INTRODUCTION

1.1. The context

The overall goal of EnvMetaGen 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. The project strengthens the research potential of human resources, lab facilities and next-generation sequencing equipment funded by a previous FP7 CAPACITIES project (No 286431). Through research, innovation, and knowledge transfer, EnvMetaGen will increase the capacity of InBIO to tackle pressing societal challenges related to the loss of biodiversity, degradation of ecosystem services and sustainable development.

The EnvMetaGen project is structured around seven interconnected Work Packages. Each Work Package has a number of Tasks designed to meet the respective Work Plan objectives. The primary objective of Work Package 4, *Deployment of an eDNA Lab*, is to deploy a fully functional environmental DNA (eDNA) lab, building upon the extant Illumina genomic platform funded by a previous FP7 CAPACITIES project (No 286431). To achieve this objective, the Work Package aims: to enhance the computational infrastructure to accommodate the massive amounts of data generated by the next-generation sequencing (Task 4.1 and Deliverable D4.1, submitted) and to train InBIO researchers and technicians for implementing best practice protocols for the analysis of eDNA (Task 4.2 and Deliverables D4.2, D4.3, D4.4 and D4.5 (this document), respectively Ferreira et al. (2018), Egeter et al. (2018), Paupério et al. (2018), current submission). Together, these activities contribute to unlocking the full research potential of InBIO in the field of environmental metagenomics.

This report constitutes the Deliverable D4.5 – Protocol for the processing of DNA sequence data generated by next-gen platforms, from Work Package 4 – *Deployment of an eDNA lab*, of the EnvMetaGen project. It reports one of the four aspects of capacity building considered pivotal to boost the future performance of InBIO in environmental genomics, which are the protocols for the processing of DNA sequence data generated by next-gen platforms. Together with the protocols for building and organizing reference collections of DNA sequences (Deliverable D4.2, Ferreira et al. (2018)), for field collection and preservation of eDNA samples (Deliverable D4.3, Egeter et al. (2018)) and for next-gen analysis of eDNA samples (Deliverable D4.4, Paupério et al. (2018)), it constitutes a standardized set of knowledge and skills that will be widely adopted in InBIO's genomic lab, achieving in this way Task 4.2 and a major objective of the EnvMetaGen project, and reaching in due time two

of the project's milestones: MS6 – Collections from sampling campaigns, and MS7 – Metagenomics protocols and tools developed.

The development of the protocols herein was made possible through a combination of activities planned within other Work Packages of the EnvMetaGen project, namely the Recruitment of the ERA Chair team (WP2; see completed Deliverables D2.1-D2.6), Secondments and Junior Researcher Exchanges through the collaboration with international networks (WP3; see completed Deliverables D3.3 & D3.5 and upcoming Deliverables D3.4 & D3.6, due at M48), an enhancement of computational infrastructure at InBIO (WP4; see above) and participation of team members in workshops and conferences (WP6; see completed Deliverable D6.6 and upcoming Deliverable D6.7, due at M48).

The protocols were designed considering the interests of stakeholders from academia, in particular InBIO, but also from industry and governmental organizations, to allow mainstreaming of environmental metagenomics to solve problems in the different domains, and in this way contributing to a major objective of Work Package 5, *Strengthening the triple helix: InBIO – Government – Industry relations*. This is expected to foster the contribution of InBIO for innovation and economic development, as one of the ways to ensure its long-term sustainability (WP5; see completed Deliverable D5.3 and upcoming Deliverables D5.4 & D5.5, due at M48).

EnvMetaGen focus on three key application areas: 1) Monitoring of freshwater eDNA for species detection; 2) Assessing natural pest control using faecal metagenomics and; 3) Next-generation biomonitoring using DNA metabarcoding. These key areas were proposed for the strategic triple helix initiatives and have been taken into account when designing eDNA projects and protocols, and that is why they are directed to samples taken from freshwater, bulk invertebrate samples and vertebrate faecal samples. Metabarcoding, the identification of species present in a sample using next-generation sequencing, has been the primary approach. For details of current EnvMetaGen-affiliated projects, including their applicability to the triple-helix initiatives and EnvMetaGen objectives, see Appendix A.

Within the context of this report and the activities associated with EnvMetaGen-affiliated projects, DNA next-generation (next-gen) data are produced by high-throughput sequencing (HTS) of both environmental DNA (eDNA) and DNA extracted from individuals (i.e. tissue) on Illumina MiSeq and HiSeq platforms, primarily those available at InBIO from a previous FP7 CAPACITIES project (No 286431).

1.2. Overview of eDNA next-gen data processing

eDNA next-gen data are generally sub-divided into two categories – i) metagenomics and ii) metabarcoding. Both survey the various organisms within environmental samples, metagenomics with a broader scope targeting whole genomes or large genomic portions including mitogenomes, and metabarcoding with a narrower scope using carefully selected fragments of marker genes that allow the identification of the taxonomic composition of samples. Technically they employ different approaches. Metagenomics, in its full extent, uses untargeted direct shotgun sequencing of the total DNA from a sample, potentially allowing the reconstitution of the genomic diversity therein through both taxonomical and functional identification (identification of biological functions and genes) (Kim et al. 2013; Porter and Hajibabaei 2018; Zepeda Mendoza, Sicheritz-Pontén, and Gilbert 2015). It has been widely used for characterizing microbial communities in the most diverse environments and conditions. On the other hand, metabarcoding is a targeted method, usually using PCR amplification coupled to high-throughput sequencing of one or more DNA marker sequences (barcodes), that allows the different species in a sample to be genetically distinguished (Epp et al. 2012; Yoccoz et al. 2012).

Due to the untargeted approach of metagenomics, low abundance species or entities could be less easily detected due to saturation with those that are more abundant. Indeed, the sequencing depth required to capture a whole community using metagenomics is much higher than the sequencing depth required to capture taxon diversity using DNA metabarcoding. In this regard, the two approaches provide complementary results, with metabarcoding being more suitable for addressing “what is there?”, while metagenomics is more powerful to investigate “what are they doing?” (Porter and Hajibabaei 2018; Zepeda Mendoza et al. 2015).

Regardless of the approach, eDNA next-gen data are most commonly text-based files in a fastq format. Generally, each sample from an experiment yields thousands or millions of raw DNA sequence reads that are processed in a standardized way in order to answer the initial question or hypothesis. The main goal of eDNA next-gen data processing is to generate reliable data that can provide the building blocks to answering ecological and environmental questions, starting from the raw sequences and most commonly involving the comparison of taxonomic diversity among samples from different environments and/or conditions (Taberlet et al. 2018).

Most existing methods for taxonomic identification in metabarcoding studies rely on pre-existing annotation linking sequences to species, stored in reference sequence databases that are used to identify the source organism based on unclassified eDNA sequences. The dependency of taxonomic identification on previously described species information has several limitations, primarily the inability to assign sequences to poorly described or undescribed species. In addition, the number of reference sequences per species in different databases is highly variable. This can lead to biased results tending to favour species that are more represented in databases, to the detriment of those less represented.

Estimating the sequencing depth that will be sufficient to recover all taxa in a sample or to answer the biological questions at hand is another difficult task. This relates, for instance, with the total sequence diversity in a sample, their lengths and relative abundances. In addition, metagenomics and metabarcoding face many additional challenges, including a high risk of contamination, the degradation of eDNA, inhibition from co-extracted molecules, erroneous sequences caused by PCR and sequencing errors, genome sequencing bias and *de novo* genome assembly (particularly metagenomics), amplification bias and chimera formation (particularly metabarcoding) (Thomsen and Willerslev 2015; Zepeda Mendoza et al. 2015). These issues pose substantial difficulties for accurate diversity estimations, requiring the inclusion of carefully selected controls, robust experimental design and consistent quality control.

The processing steps for eDNA next-gen data allow the extraction of a subset of reliable sequences (those without errors, sequencing artefacts and contaminants) that are commonly searched against reference sequence databases like the NCBI Nucleotide in order to identify the organism they originate from (Agarwala et al. 2018). In addition, sequence dereplication or clustering methods provide a measure of read quantification per sample allowing to obtain a matrix with the frequency of each unique or representative sequence in each sample. Clustering sequences into operational taxonomic units (OTUs) based on a nucleotide similarity threshold has been a common method to reduce data complexity and also produce quantification tables. These OTU tables or taxon tables can be used for downstream processing, such as statistical methods for diversity and differential analysis.

To date, there is no single universal processing procedure providing a unified and streamlined manner for satisfactorily treating eDNA data from raw sequences to taxonomic identification and diversity analysis. On the contrary, there are many bioinformatic pipelines that have been separately developed and are being used and improved by the eDNA research community. Tools such as *obitools* (Boyer et al. 2016), *USEARCH* (Edgar and Flyvbjerg 2015),

VSEARCH (Rognes et al. 2016), *qiime* (Caporaso et al. 2010), *mothur* (Schloss et al. 2009), *SWARM* (Mahé et al. 2015) and several others have been developed to facilitate data analysis. Community guidelines and efforts to systematize and unify data analysis procedures for metabarcoding and metagenomics are thus highly required. For instance, this issue is tackled within the DNAqua-Net framework for aquatic ecosystems (Leese et al. 2016).

The NCBI Nucleotide reference database is a sequence database that has historically been most used for searching nucleotide sequences, storing data from several sources including the GenBank, RefSeq and PDB databases (Agarwala et al. 2018). In more recent years, other sequence databases appeared that provide additional resources for sequence matching, namely the BOLD database (Ratnasingham and Hebert 2007), which contains DNA barcode records for the mitochondrial cytochrome c oxidase subunit I (COI) gene as well as molecular, morphological and distributional data for eukaryotes. Many reference sequence databases exist, commonly focussing in a taxonomic group or particular environment, for example the SILVA database for rRNA (Quast et al. 2013) or the ITSoneDB for fungi (Santamaria et al. 2018).

Several reviews of methods and databases for metagenomics have been reported (Breitwieser, Lu, and Salzberg 2017; Fosso et al. 2015; Kim et al. 2013; Pavlopoulos et al. 2015; Santamaria et al. 2012). Newer approaches relying on machine learning and trying to circumvent the limitations of the existing databases are increasingly employed for metabarcoding and metagenomics (Pasolli et al. 2016; Rangwala, Charuvaka, and Rasheed 2014; Soueidan and Nikolski 2015; Vacher et al. 2016).

1.3. Overview on the deployment of next-gen data processing at InBIO

Research projects generating and using next-gen data are common at InBIO, for instance on transcriptomics (Albert et al. 2012; Azevedo et al. 2016; Castro-Nallar et al. 2015; Loire et al. 2013; Machado et al. 2018; Pereira-Leal et al. 2014; Pérez-Losada et al. 2015), genomics and population genomics (Cosart et al. 2011; Cahais et al. 2012; Gayral et al. 2013; Gargani et al. 2015; Fontanesi et al. 2016; Crawford et al. 2017), mitogenomics (Gibb et al. 2016; Marques et al. 2017), viral metagenomics (Conceição-Neto et al. 2017), for defining the cystic fibrosis lung microbiome (Hahn et al. 2016), for diet analysis (Aizpurua et al. 2018; Mata et al. 2016, 2018; McInnes, Alderman, Deagle, et al. 2017; McInnes, Alderman, Lea, et al. 2017; McInnes, Jarman, et al. 2017; Sousa et al. 2016; de Vos et al. 2018) and species identification (Vasconcelos et al. 2016; Corley et al. 2017; Corley and Ferreira 2017) studies.

EnvMetaGen increases InBIO's capacity for ecological studies using eDNA data, with particular focus on three application areas for triple helix strategic initiatives: (i) monitoring of freshwater eDNA for species detection, (ii) assessing natural pest control using faecal metagenomics, and (iii) next-generation biomonitoring using DNA metabarcoding, as can be seen in more detail in Deliverables D4.3 and D4.4 (Egeter et al. (2018) and Paupério et al. (2018)) and through the affiliated projects described in Appendix A. Standardized processing pipelines are used within the EnvMetaGen project aiming at the most reproducible routines, with particular protocol differences relating to specificities of each project's application.

Most next-gen sequencing data generated within EnvMetaGen-affiliated projects are from Illumina MiSeq or HiSeq platforms, from genetic markers (e.g. COI and genes coding for ribosomal RNA such as 12S, 16S and 18S) that allow the characterisation of the taxonomic composition of a sample. Reference fragments of such markers are paired-end sequenced and allow the ascertainment of taxonomic diversity of a sample based on sequence matching to existing databases. In addition, they can serve to create reference collections, databases containing sequence identifiers and annotations for multiple organisms – see Deliverable D4.2 (Ferreira et al. (2018)) for details on the InBIO Barcoding Initiative (IBI).

Faecal, water and bulk samples are the major raw materials used within the project, serving four main applications: i) single species detection, ii) diet assessment, iii) biodiversity assessment and iv) reference collection barcoding.

1.4. Structure of the report

For accomplishing WP4 Task 4.2, the deployment of an eDNA laboratory, a workflow has been set up at InBIO and is outlined in Figure 1. The different steps of the workflow are shown according to the reporting structure of Deliverables D4.2-D4.5 (Ferreira et al. (2018), Egeter et al. (2018), Paupério et al. (2018) and this document). Within this workflow, the preceding steps allow the production of the next-gen data, namely through the InBIO Barcoding Initiative (IBI), an invertebrate reference collection (Deliverable D4.2, Ferreira et al. (2018)), the field collection and preservation of eDNA samples (Deliverable D4.3, Egeter et al. (2018)) and the analyses of environmental samples from DNA extraction, through amplification of targeted sequences and library preparation to sequencing (Deliverable D4.4, Paupério et al. (2018)).

This report constitutes Deliverable D4.5 and focuses on the processing steps for the next-gen data produced within EnvMetaGen-affiliated projects (Figure 1). Section 2 provides a general

description of the current workflow of the data processing for eDNA metabarcoding, in the context of the existing literature. Section 3 details the processing steps used for metabarcoding data produced within the EnvMetaGen scope, which serve three primary applications: i) single species detection, ii) diet assessment and iii) biodiversity assessment. Section 4 details the processing of data generated within the scope of the IBI, the fourth primary application: iv) reference collection barcoding.

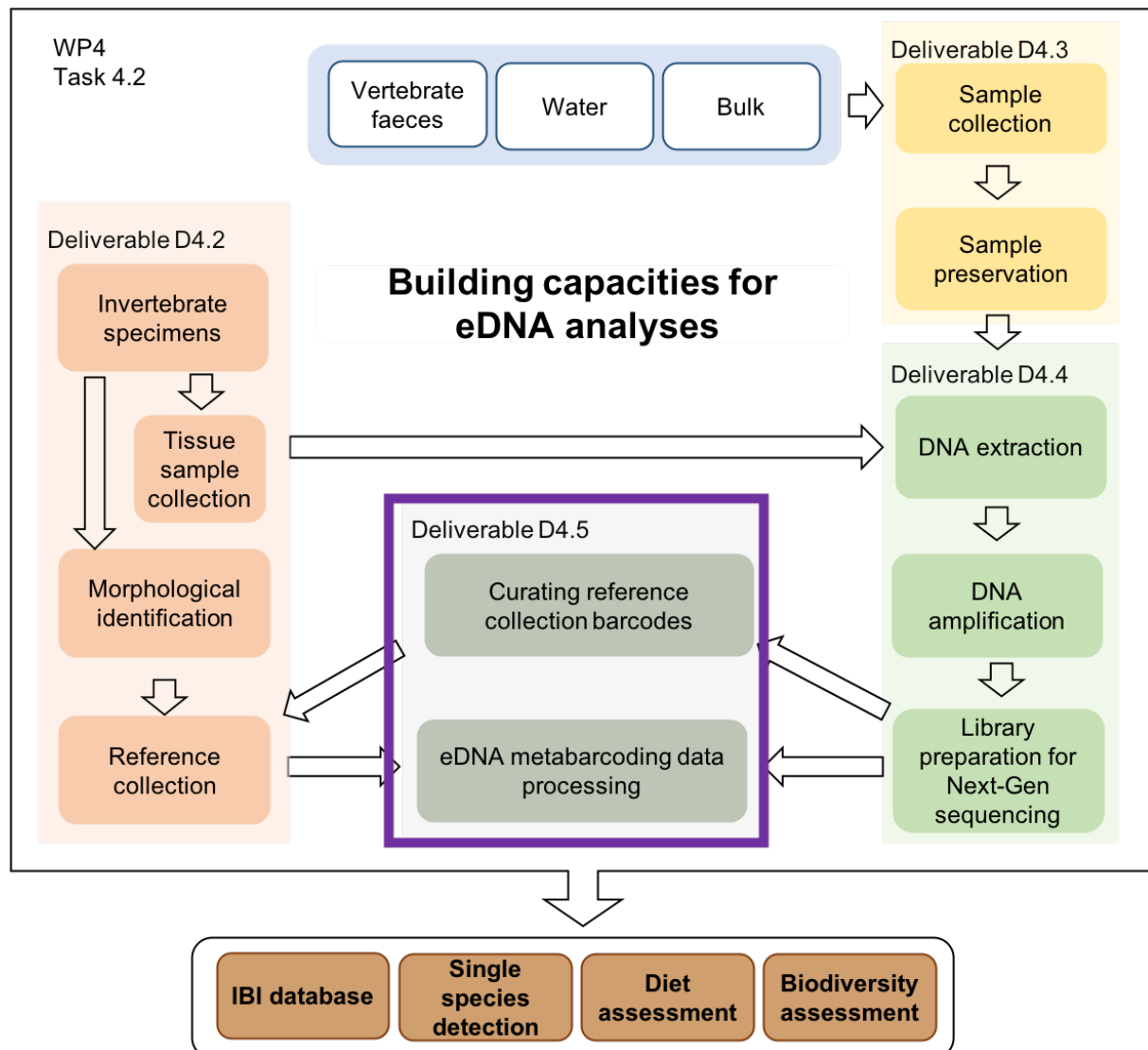


Figure 1. EnvMetaGen eDNA Lab workflow – steps are grouped according to the deliverable in which they are addressed (Deliverables D4.2 – D4.5, respectively Ferreira et al. (2018), Egeter et al. (2018), Paupério et al. (2018) and this document). The type of eDNA samples (blue) and project applications (brown) require a range of tailored protocols within workflow steps, which are detailed in Deliverables D4.2 - D4.5. The current report, Deliverable D4.5 (purple box), focuses on the processing steps for the next-gen data produced within EnvMetaGen-affiliated projects.

2. GENERAL WORKFLOW FOR PROCESSING eDNA METABARCODING DATA

This Section details the general workflow for processing eDNA metabarcoding data, as a large proportion of data produced within the EnvMetaGen project falls within this data type. Literature context is given and the pros and cons of different steps are discussed.

2.1. Introduction

The processing pipeline for eDNA metabarcoding data starts with the raw sequencing data from next-gen sequencers and produces taxonomic composition descriptions for each sample. The metabarcoding workflow produces sequencing libraries for all samples (see Deliverable D4.4, Paupério et al. (2018), for details), which can be sequenced in a multitude of platforms such as those from Illumina, Roche, PacBio, Ion Torrent, SOLiD and Nanopore. The choice of the sequencing platform will highly depend on the study goal and should be well considered, as they apply different chemical methods for the sequencing and provide data with distinct characteristics (Glenn 2011; D'Amore et al. 2016; Allali et al. 2017; Cao et al. 2017). Within EnvMetaGen, the Illumina platforms MiSeq and HiSeq are used for next-gen sequencing. For a recent thorough description of eDNA for biodiversity research and monitoring please refer to Taberlet et al. (2018). The general steps for processing next-gen metabarcoding sequencing data are depicted in Figure 2.

Processing pipelines apply some or all of the presented steps, and the order of the steps can vary among studies. Common workflows include sample demultiplexing, merging read pairs, quality filtering, error correction and chimera filtering, sequence dereplication and singleton removal, sequence clustering by similarity into Operational Taxonomic Units (OTUs), taxonomic annotation using *BLAST* (Altschul et al. 1990) against reference databases, and post-*BLAST* taxonomic assignment for instance using *MEGAN* (Huson et al. 2016) and the lowest common ancestor (LCA) algorithm. Throughout the process of taxonomic assignment, manual curation is usually necessary prior to diversity and differential analyses which are frequently done using R/Bioconductor packages such as *vegan* (Jari Oksanen, F. Guillaume Blanchet, Michael Friendly et al. 2017), *phyloseq* (McMurdie and Holmes 2013), *iNEXT* (Hsieh, Ma, and Chao 2016) and *DESeq2* (Love, Huber, and Anders 2014).

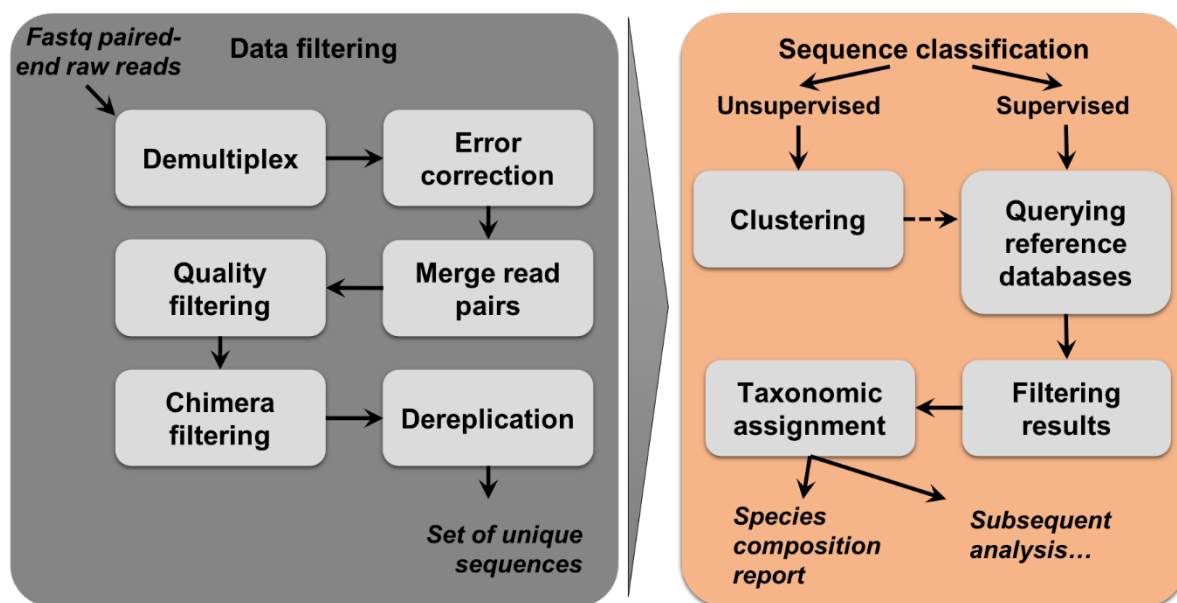


Figure 2. General metabarcoding data analysis workflow of eDNA data. An initial data filtering step takes the raw sequencing reads (commonly paired-end) and processes them into a set of ideally high quality and low error unique reads. Not all steps are mandatory and their choice impacts the results. In a second phase, sequence classification, often via taxonomic assignment through reference databases, allows the characterization of species composition details for a group of samples.

2.2. Data filtering

The main goal of data filtering is the production of a subset of reliable sequences, the most error-free possible, which better represent the real sequence diversity and abundance in a sample (Figure 2). This is a critical step of data processing that can have a strong impact on results and thus should be well considered on processing pipelines.

2.2.1. Demultiplex

Prior to sequencing, samples are generally pooled together in order to maximize the sequencing output from one run. For this, known unique specific short sequences, termed indexes, are added to each sample prior to sequencing, allowing resultant reads to be assigned to the sample from which they originated (see Deliverable D4.4, Paupério et al. (2018), for more details). Demultiplexing is the process of organizing reads by sample using these indexes and amplification primers. However, this process is not error free and can lead to the incorrect assignment of sequences to samples, for instance via index cross contamination or from sequencing errors (Pedersen et al. 2015; O'Donnell et al. 2016). Therefore, allowing 1 or 2 mismatches on the indexes during the demultiplexing step might be advantageous or necessary. Several existing tools have been reported to overcome demultiplexing difficulties

(Renaud et al. 2015; Yi et al. 2015; Zepeda-Mendoza et al. 2016; Murray, Borevitz, and Berger 2018).

2.2.2. Error correction

Erroneous DNA sequences can arise at different stages in the metabarcoding process. Some important factors that have been noted are the very long DNA preservation times, PCR errors causing point mutations and chimeric fragments, and incorrect base calling during sequencing (Thomsen and Willerslev 2015). During the sequencing process the nucleotide composition of template sequences are generated including errors. These affect the reported read composition in variable proportions depending on the sequencing technology and underlying chemistry. Errors are problematic, for instance, because they can contribute to overestimation of sample diversity (Kunin et al. 2010; Sefc, Payne, and Sorenson 2007). In addition, they can hamper assemblies and correct sequence alignments. To account for these ubiquitous error rates from sequencing, error correction programs have been developed that take into account the sequencing technology and prior knowledge to find and correct sequencing errors (Alic et al. 2016). In (Alic et al. 2016) the authors compared 50 of these methods, providing guidance on which methods perform better for each sequencing technology. For example, programs such as *USEARCH* allow error correction to be performed based on the expected number of errors for each read by using the quality scores.

2.2.3. Merge read pairs

In paired-end sequencing, both forward and reverse ends of a fragment are sequenced. In order to simplify and reduce redundancies in the sequencing data, several programs exist that take the forward and reverse reads of a sample and attempt to merge them based on an existing overlap or quality score (Zhang et al. 2014; Edgar and Flyvbjerg 2015; Schubert, Lindgreen, and Orlando 2016), one of which is the *obitools* ‘illuminapaireded’ command. Fastq files contain read IDs and the read sequence (A, C, T, G or N for uncalled bases), as well as quality scores per base, giving the probability of that base call being incorrect. The forward and reverse reads of the same fragment are at the same line position in the two fastq files obtained after paired-end sequencing. The assembly of the forward and reverse reads is done by alignment and returns the reconstructed sequence as well as an alignment score that is used to filter out reads. The fraction of reads merged and kept for further processing depend on the specified minimum read overlap. The consensus quality score calculation method will additionally influence downstream quality score filtering.

2.2.4. Quality filtering

The quality information contained in the fastq files are encoded with ASCII characters that correspond to numeric Phred quality scores. The estimated probability (P) of a base call being incorrect is given by $P = 10^{-Q/10}$, with Q being the Phred quality score of a base. Therefore, a quality score Q of 30 represents a 1 in 1000 chance of that base call being incorrect. These Phred quality scores are generally used to filter out reads or parts of reads for which the probability of incorrect base calling is higher than desirable (Sunyoung Kwon et al. 2013; Wright and Vetsigian 2016). A fairly strict quality filtering uses 30 as minimum quality score in the entire read. Softer filtering use 20 as minimum quality score and/or a percentage of the read having to fulfil that imposition (Deiner et al. 2017). Several programs allow the filtering of reads by quality scores including the *FASTX-Toolkit* () and *PRINSEQ* (Schmieder and Edwards 2011).

A cleaning step for low abundance sequences is also often employed in processing workflows, namely removing sequences with a count under a defined threshold, which might be the result of errors or chimera formation.

2.2.5. Chimera filtering

Chimeras arise during sample preparation (e.g. PCR steps), from the partial joining of two or more fragments, for instance when closely related sequences are amplified. In the simplest case, the chimeric sequence contains one part from one fragment and a second part from a different fragment. Several methods for identifying and removing chimeras have been developed (Schloss et al. 2009; Caporaso et al. 2010; Edgar and Flyvbjerg 2015; Rognes et al. 2016), either relying on filtering out known chimeric sequences (using a reference chimera database) or via *de novo* chimera filtering, which models possible chimeric formations based on the given sequences.

2.2.6. Dereplication

Dereplication produces unique sequences with abundance counts. It is essential to reduce data redundancy and minimize computational effort. It also provides a measure of the sequence diversity relative to sequence abundance, allowing to get a sense of the total number of different sequences present in a sample and their relative amounts within the sample. Most tools for sequencing data analysis perform a dereplication step by grouping and counting

identical reads (Caporaso et al. 2010; Edgar and Flyvbjerg 2015; Fosso et al. 2015; Rognes et al. 2016; Mysara et al. 2017; Allali et al. 2017). This is equivalent to clustering reads with 100% similarity.

2.3. Sequence classification

The main goal of sequence classification is to provide a simplified but comprehensive list of unique sequences grouped by common attributes that ideally cannot be further sub-divided. This can be done for instance using reference databases that in some cases allow the classification of sequences by species (supervised approach). In cases for which no sufficient annotation information and database info exists, it might only be possible to group sequences by nucleotide similarity, using clustering methods (unsupervised approach). For reference-based taxonomic assignment (Figure 2), a species composition report for a group of samples is commonly obtained and further statistical analysis done to answer ecological questions.

2.3.1. Clustering sequences based on similarity

After dereplication there is an optional clustering step. Dereplicated sequences can be directly matched against reference databases (e.g., using BLAST) or serve for differential and diversity analysis. However, clustering sequences prior to sequence database matching is useful to reduce the number of input searches, and thereby the computational effort. Clustering also decreases potentially spurious results, by joining similar sequences based on a similarity threshold and retaining a representative sequence only, commonly termed OTU (operational taxonomic unit), which theoretically represent the same taxonomic unit. In addition, this step facilitates the construction of quantification tables, which are useful to understand and compare sequence abundances among a group of samples. *obitools*, *USEARCH*, *VSEARCH*, *qiime*, *mothur* and *SWARM* all perform such tasks, providing different and adjustable parameters.

2.3.2. Querying reference databases

BLAST is by far the most common tool for the retrieval of sequence information by similarity. In the context of analysis workflows, the command line *BLAST+* suite (Camacho et al. 2009) is used to perform sequence searches against NCBI reference databases or any database converted into a *BLAST*-readable format.

The BOLD database provides an increasingly relevant resource for metabarcoding. Despite a highly manual sequence search system, recent tools are providing easier access to the BOLD database information, including programmatic access (Vesterinen et al. 2016).

With the increase of genome sequencing projects, alignment tools have been developed that allow for fast and accurate mapping of reads to genomes or target databases (Langmead and Salzberg 2012; Li and Durbin 2009; Liu et al. 2012), extending the options for sequence searching against a reference database.

2.3.3. Filtering reference database query results for taxonomic assignment

The raw outputs of sequence database searches often contain spurious results that need to be filtered in order to keep only the set of matches that are most plausible. This task frequently involves text and table processing methods using custom definitions such as the percentage identity, the alignment length, query coverage, and many other options that may be common to most projects or project-specific. In essence, this step reduces the search results into a higher confidence subset that can be further analysed. Besides manual curation, programs such as *MEGAN* are useful in this task. *MEGAN* is a toolbox for taxonomic analysis of sequences, commonly used through a graphical user interface but also available via command-line. It uses a lowest common ancestor (LCA) algorithm to assign taxonomy based on database search results, for instance from a *BLAST* results file. For each query sequence, it considers all the search results that pass the user-defined parameters and assigns the query to the highest taxonomic resolution possible. The default taxonomy used by the toolbox is the NCBI taxonomy tree, but custom taxonomy can also be defined by the user. *MEGAN* also provides tools for a variety of analyses and visualizations for comparing samples.

2.3.4. Subsequent analyses

The following steps are generally dependent on project's goals and involve statistical analysis. For instance, a common step is diversity analysis, which provides a quantitative measure of differences in the ecological traits in study with a focus on taxonomic units, functional types or communities. Statistical and community analysis tools are used in this step, including *phyloseq* (McMurdie and Holmes 2013), *vegan* (Jari Oksanen, F. Guillaume Blanchet, Michael Friendly et al. 2017), *BiodiveristyR* (Kindt and Coe 2005), *DESeq2* (Love et al. 2014) and *iNEXT* (Hsieh et al. 2016).

2.4. Other considerations for eDNA metabarcoding data processing

The vast data amounts, the varied questions and motivations underlying different studies and the lack of a community-based effort for standardizing procedures and providing guidelines is reflected in the multitude of approaches reported in the metabarcoding and metagenomics fields.

No single optimal workflow exists that allows for an easy to follow but comprehensive dataset comparison, and many challenges remain, including reproducibility. Most of the time, researchers struggle with the literature and the different tools setting up their own analysis pipeline satisfying the intended criteria. Nonetheless, several tools exist that provide means to bundle several analysis steps (Schloss et al. 2009; Caporaso et al. 2010; Edgar and Flyvbjerg 2015; Fosso et al. 2015; Zepeda Mendoza, Sicheritz-Pontén, and Gilbert 2015; Boyer et al. 2016; Rognes et al. 2016; Mysara et al. 2017; Allali et al. 2017).

3. ENVMETAGEN PROTOCOL FOR PROCESSING eDNA METABARCODING DATA

3.1. Introduction

In this Section, the processing steps for metabarcoding data used within the EnvMetaGen project are presented in context of the affiliated projects (see Appendix A). Currently, the analyses pipelines implemented within EnvMetaGen-affiliated projects are mainly based on *obitools* (Boyer et al. 2016), which allows the streamlining of several steps commonly employed in metabarcoding data analysis (Figure 2). Taxonomic assignment is achieved with *blastn* from the command-line *BLAST+* suite and *MEGAN* or with *Geneious* and manual inspection. The NCBI Nucleotide database ('nt'), BOLD, the IBI invertebrate reference collection and additional public reference and private databases are frequently used. Manual sequence query on the BOLD website (<http://www.boldsystems.org/index.php/databases>) is performed to include their private sequences on the results.

The projects AZORES, FRESHING, FILTURB, GUELTA, ICVERTS, IRANVERTS, NZFROG and WOLFDIET follow mainly the *obitools* workflow, local *BLAST* against the reference Nucleotide database and *MEGAN* for secondary taxonomic assignment. The projects CHASCOS, ECOLIVES, SABOR and TUA also follow an *obitools*-based workflow, including one cleaning step for removing sequences that are likely errors followed by alignment in *Geneious* to select unique sequences to search against BOLD and IBI, in order

to build taxa composition tables which are further processed with statistical methods. Other projects are in an early phase and will follow the described workflows with project-specific adjustments.

3.2. Protocol used for the processing of metabarcoding data

The processing of metabarcoding data produced within EnvMetaGen's scope follows the workflow detailed in Figure 3.

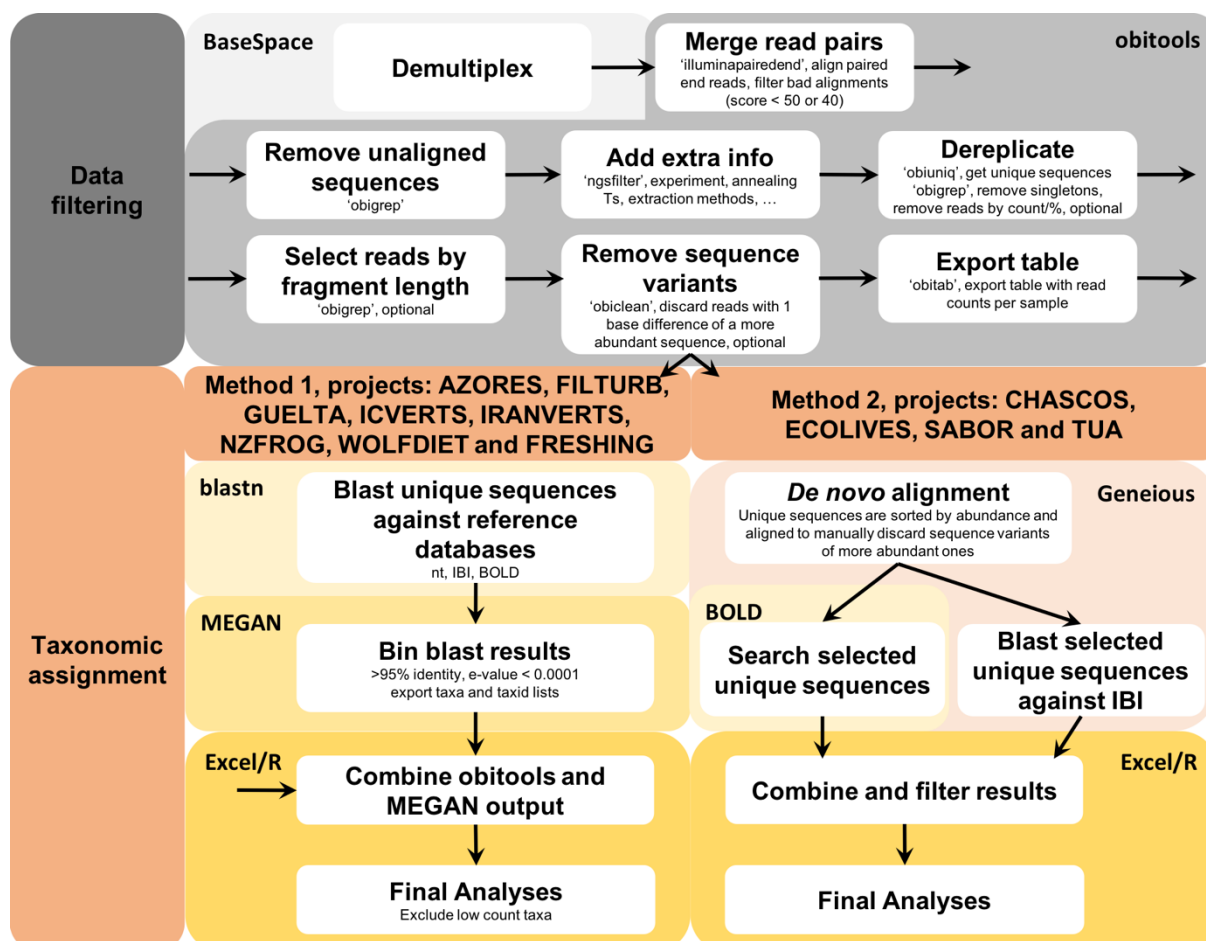


Figure 3. Data processing workflow currently used in EnvMetaGen for metabarcoding data, based on obitools for data filtering and BLAST against reference databases for taxonomic assignment.

3.2.1. Data filtering

Upon sequencing on MiSeq or HiSeq Illumina platforms, usually using paired-end reads, several gigabytes of raw sequencing data are produced and stored in fastq files. As sequencing libraries most commonly contain several samples pooled together and identifiable through multiplexing indexes (see Deliverable D4.4, Paupério et al. (2018), – Sections 2.3

and 3.3), the first step is to obtain two fastq files per sample, that is, containing all sequenced reads that originate from one sample. The Illumina program BaseSpace is used for this step, together with a user-provided identity file matching samples and/or primer sets to the respective unique combination of indexes. At the end, two fastq files are generated per sample, one for forward reads and one for the reverse reads. Next, the forward and reverse read pairs of a read are merged using the *obitools* ‘*illumina-paired-end*’ command. A minimum alignment score (commonly between 40 and 50) is used and unaligned sequences presenting no overlap between the forward and reverse reads are discarded. Visualizing the distribution of alignment scores is useful to get a sense of which threshold to use and of the merging success. Then the reads are processed through the command ‘*ngsfilter*’ to add labelling info, using a description file. Chimera filtering is carried out using the *de novo* option of *USEARCH* or *VSEARCH* (e.g. FRESHING) or by visualizing alignments and looking for sub-fragments of more abundant sequences that appear in less abundant ones, which are then removed (e.g. CHASCOS, ECOLIVES, SABOR and TUA). The following step, dereplication, is performed with the command ‘*obiuniq*’, discarding sequences with an abundance of only one read (singletons), as they most likely result from sequencing errors and no biological interpretation can be done with such low occurrence. *USEARCH* or *VSEARCH* are also commonly used for this step. Additional processing steps are done mainly to exclude reads that represent artefacts. For instance, a length range is defined for keeping sequences using the ‘*obigrep*’ command, based on the expected fragment size which depends on the primers used (e.g. CHASCOS, ECOLIVES, FRESHING, NZFROG, SABOR, TUA and WOLFDIET). Another option is removing reads differing 1 base pair from a more abundant sequence using ‘*obiclean*’ (e.g. CHASCOS, ECOLIVES, FRESHING, SABOR and TUA). In addition, depending on the project, a number of options are used to remove spurious reads, for example reads with count < 50 or reads with a count < 1% of the total reads in a sample (e.g. FRESHING, NZFROG, WOLFDIET). At this point, one has a set of unique reads and their abundances in a fasta file and can generate a quantification table, commonly referred to as an OTU table, summarizing the abundance of each read, either in the total dataset or per sample, using the ‘*obitab*’ command. In addition to those previously mentioned, FRESHING is currently comparing the output of *obitools*, *USEARCH*, *VSEARCH* and *SWARM* for sequence clustering by similarity.

A step-by-step protocol for the most often used data filtering method for metabarcoding data within EnvMetaGen is provided in Appendix B, Section B1.

3.2.2. Taxonomic assignment

The next steps detail the assignment of taxonomy to filtered sequences following two different approaches, Method 1 and Method 2 (see also Figure 3).

Method 1

The first method is based on locally blasting the filtered sequences against reference databases using the *blastn* tool and most often the ‘megablast’ option in order to search for very similar sequences. The NCBI Nucleotide database is used, often including additional sequences obtained from target taxa expected to be observed in the samples. Depending on the project, the IBI invertebrate reference collection is used. Manual searches on BOLD are often also performed in a project basis. *MEGAN* is used to filter and bin the blast results and fine tune the taxonomic assignment (using parameters such as >95% identity, e-value < 0.0001, considering the blast top hit and other hits within 5% blast score). *MEGAN* taxonomic assignments are then merged with the quantification table from ‘obitab’ and further processing (e.g. Excel, R) and manual inspection is done, including filtering of taxa based on low counts (often using data obtained from negative controls, see Deliverable D4.4, Paupério et al. (2018), for details).

EnvMetaGen-affiliated projects using this approach: AZORES, FILTURB, FRESHING, GUELTA, ICVERTS, IRANVERTS, NZFROG, WOLFDIET.

Method 2

The second method uses *Geneious* to perform *de novo* alignment of the unique sequences, sorted by read count. Manual inspection is necessary to determine which sequences are variants of more abundant sequences, which are discarded. The remaining unique sequences are searched against the IBI invertebrate reference collection or against BOLD. Manual filtering of the database query results produces a final table with a taxonomic description for each unique sequence and sequence occurrence per sample. From these, species richness and diversity analyses are commonly carried out using R/Bioconductor packages on a project basis.

EnvMetaGen-affiliated projects using this approach: CHASCOS, ECOLIVES, SABOR, TUA.

Other projects including AGRIVOLE, CRAYFISH, GALEMYS and MANTIDS are currently in early phase, as well as MATEFRAG for next-gen data analyses. They will follow similar workflows, with adjustments as needed. The project XENOPUS uses eDNA for the qPCR-based detection of an invasive frog species and will not involve next-gen data.

4. ENVMETAGEN PROTOCOL FOR PROCESSING InBIO BARCODING INITIATIVE (IBI) DNA DATA

4.1. Introduction

Deliverable D4.2 (Ferreira et al. (2018)), contains details on the InBIO Barcoding Initiative (IBI). In this section, we present the processing steps utilized for the standardized analysis of data produced within IBI's scope for the invertebrate reference collection of DNA sequences. Data filtering is mainly *obitools*-based. The software *Geneious* is used to assemble the reads from one sample (DNA amplified from a single specimen), providing one or more overlapping fragments depending on the primers used, which are then manually matched to morphological identification and added to the database.

4.2. Protocol for processing IBI DNA sequence data

The processing of next-gen DNA data produced within the IBI's scope follows the workflow detailed in Figure 4.

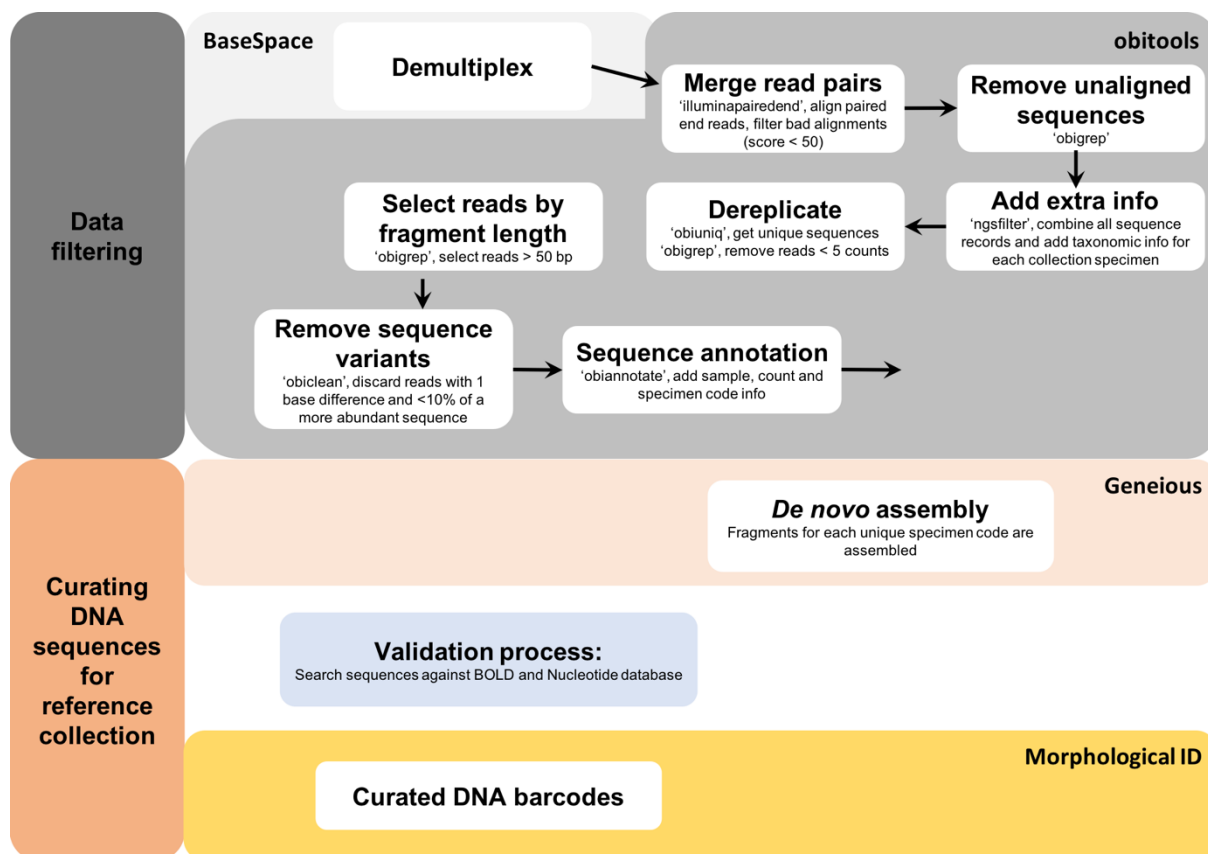


Figure 4. Data processing workflow currently used in EnvMetaGen for the IBI invertebrate reference collection, based on obitools for data filtering and using Geneious and BLAST against reference databases to curate and validate sequence data annotation with the morphological identification.

4.2.1. Data filtering

Next-gen data produced for the reference collection are generated on a MiSeq Illumina platform using paired-end reads (Deliverable D4.2, Ferreira et al. (2018), Section 3.4), with many gigabytes of raw sequencing data produced and stored in fastq files. As sequencing libraries of the reference collection contain samples from different specimens pooled together and identifiable through multiplexing indexes (see Deliverable D4.4, Paupério et al. (2018), Sections 2.3 and 3.3 for general library preparation steps), the first step is to assign all sequenced reads that originate from one specimen. The Illumina program *BaseSpace* is used for this step, together with a user provided identity file matching specimens and/or primer sets to the respective unique combination of indexes. At the end, fastq files are generated per sample, one for forward reads and one for the reverse reads. Next, the forward and reverse read pairs of a read are merged using the *obitools* ‘illuminapairedend’ command. Currently, a minimum alignment score of 50 is used and unaligned sequences presenting no overlap between the forward and reverse reads are discarded. Then, reads are processed through the command ‘ngsfilter’ to combine all sequence records in one fastq file including taxonomic

info for each collection specimen, using a description file. If needed, additional demultiplexing can be done in this step, for instance to assign each sequence record to marker combination. The following step, dereplication, is performed with the command ‘obiuniq’ and only sequences with > 5 reads are kept. Count statistics for each sample are obtained with the ‘obistat’ command. Additional processing steps are done mainly to exclude reads that represent artefacts. For instance, reads with a length < 50 bp are discarded using the ‘obigrep’ command. Fields with sample and count information are added to the reads using the ‘obiannotate’ command. Then, reads differing 1 base pair and with an abundance lower than 10% of a ‘head’ sequence are discarded using the ‘obiclean’ command. Sequences are then sorted by decreasing count number with the ‘obisort’ command and the sequence ID is converted to the IBI specimen code using ‘obiannotate’. Depending on the dataset and sequencing run, sequences are split by insect order using the ‘obisplit’ command. At this point, one has a set of unique sequences with specimen codes in a fasta file (no longer containing base quality information).

A step-by-step protocol for filtering next-gen data produced through the IBI for the invertebrate reference collection is provided in Appendix B, Section B2.

4.2.2. Curating DNA sequences for reference collection

Fasta files containing specimen codes and their sequences are uploaded to *Geneious* (Figure 4). In case of the 658bp COI barcode, which is amplified and sequenced in two overlapping fragments due to the maximum length allowed by the Illumina MiSeq platform (see Deliverable D4.2, Ferreira et al. (2018), Section 3.4 for details), *de novo* assembly is performed for each unique specimen code to assemble the two fragments and obtain a consensus sequence spanning the overall length of the COI barcode. Next, a step of manual visualization and curation is performed. Sequence variants presenting very low coverage and 1 or 2 bp differences from more abundant sequences are dismissed from the analyses. As a validation step, the consensus sequences for each specimen code are blasted against the BOLD and the NCBI Nucleotide databases following the steps described previously for the metabarcoding data processing (Section 3.2). They are then matched to the morphological identification of the specimen. Curated sequences and additional data are inserted into the reference collection of DNA sequences database, which currently comprises over 6200 specimens (Deliverable D4.2, Ferreira et al. (2018)).

5. OTHER DEVELOPING AREAS FOR THE PROJECT

Reproducibility is a general issue in research and within EnvMetaGen there is an ongoing effort to standardize data processing pipelines and analyses reports, in order to facilitate their comprehension and reproducibility. Besides *obitools*, other processing programs such as *USEARCH*, *VSEARCH*, *SWARM*, *qiime* and *mothur* are used on a project by project basis. In later stages of the project, we foresee publicly sharing the code developed through GitHub or other suitable platform. Currently we aim at developing configurable and fully automated pipelines to address the above issues. In addition, we are also working on the development of several scripts to facilitate analysis, namely to filter *BLAST* outputs and report the taxonomic assignment based in a set of rules or *a priori* defined criteria.

Data processing workflows may vary according to the specific needs of the project. For instance, SOILPHOS, an ongoing project that uses DNA extracted from an agricultural plant growth experiment (see Appendix A), uses sequence clustering by similarity and protein blast against the NCBI non-redundant database. This project is still in the early phases and additional fine tuning of the data processing will be done as the project progresses.

Reference collection of DNA sequences is a key area in the project fostered by the InBIO Barcoding Initiative (IBI), including more than 6200 invertebrate specimens. To accommodate, organize and facilitate the access to specimen information, a relational database (in MySQL) and a graphical user interface frontend (implemented in JAVA) for easy access are under development (Deliverable D4.2, Ferreira et al. (2018), Section 3.1.2).

The availability of longer barcodes for reference collection of DNA sequences could improve their usefulness for taxonomic identification by expanding the range of single species discriminated. In this regard, newer technologies that allow for sequencing longer reference barcodes are of interest, such as the Oxford Nanopore MinIon sequencer, a portable device which allows to sequence reads up to hundreds of kilobases in real-time. Indeed, DNA barcoding using the MinIon technology has already been reported in the field (Menegon et al. 2017) and a workflow including an analysis pipeline for achieving high accuracy presented (Srivathsan et al. 2018). Such workflows are being investigated and could potentially be implemented within the EnvMetaGen project, for expanding the IBI invertebrate reference collection and/or for new barcode reference collections.

6. CONCLUDING REMARKS

This report provides a description of the current processing steps for the next-gen data produced within EnvMetaGen affiliated projects, which focus on eDNA collected from three main sample types (vertebrate faeces, water samples and bulk samples) and targeting three main applications (single species detection, diets and biodiversity assessments). An overview of the current state of the art for the processing of metabarcoding data was provided in Section 2, as this is the main data type generated within the project. The current metabarcoding data processing protocols used within EnvMetaGen-affiliated projects were described in Section 3. In addition, the processing of data generated from tissue samples of specimens for the invertebrate reference collection of DNA sequences of the InBIO Barcoding Initiative (IBI) was presented in Section 4. Other relevant areas for the project were mentioned in Section 5.

All EnvMetaGen-affiliated projects are generating ecological and environmental data to tackle pressing societal challenges related to the loss of biodiversity, degradation of ecosystem services, and sustainable development. These data feed into the triple-helix initiatives in the context of the strategic key areas of freshwater species detection, natural pest control services and biomonitoring. This report (Deliverable D4.5) describes current state of the art protocols for the processing of DNA sequence data generated by next-gen platforms. Together, Deliverables D4.2-D4.5 (Ferreira et al. (2018), Egeter et al. (2018), Paupério et al. (2018) and this document) 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

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, EnvMetaGen project (Grant Agreement No 668981). European Union Horizon 2020 Research & Innovation Programme – H2020-WIDESPREAD-2014-2 doi: 10.5281/zenodo.2586889

8. REFERENCES

- Agarwala, Richa et al. 2018. "Database Resources of the National Center for Biotechnology Information." *Nucleic Acids Research* 46(D1):D8–13. Retrieved July 9, 2018 (<https://academic.oup.com/nar/article/46/D1/D8/4621330>).
- Aizpurua, Ostaizka et al. 2018. "Agriculture Shapes the Trophic Niche of a Bat Preying on Multiple Pest Arthropods across Europe: Evidence from DNA Metabarcoding." *Molecular Ecology* 27(3):815–25. Retrieved July 25, 2018 (<http://doi.wiley.com/10.1111/mec.14474>).
- Albert, Frank W. et al. 2012. "A Comparison of Brain Gene Expression Levels in Domesticated and Wild Animals" edited by J. M. Akey. *PLoS Genetics* 8(9):e1002962. Retrieved July 10, 2018 (<http://dx.plos.org/10.1371/journal.pgen.1002962>).
- Alic, Andy S., David Ruzafa, Joaquin Dopazo, and Ignacio Blanquer. 2016. "Objective Review of de Novo Stand-Alone Error Correction Methods for NGS Data." *Wiley Interdisciplinary Reviews: Computational Molecular Science* 6(2):111–46. Retrieved January 15, 2018 (<http://doi.wiley.com/10.1002/wcms.1239>).
- Allali, Imane et al. 2017. "A Comparison of Sequencing Platforms and Bioinformatics Pipelines for Compositional Analysis of the Gut Microbiome." *BMC Microbiology* 17(1):194. Retrieved December 4, 2017 (<http://www.ncbi.nlm.nih.gov/pubmed/28903732>).
- Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. 1990. "Basic Local Alignment Search Tool." *Journal of Molecular Biology* 215(3):403–10. Retrieved August 13, 2018 (<http://www.ncbi.nlm.nih.gov/pubmed/2231712>).
- Azevedo, Herlânder et al. 2016. "Transcriptomic Profiling of Arabidopsis Gene Expression in Response to Varying Micronutrient Zinc Supply." *Genomics Data* 7:256–58. Retrieved July 10, 2018 (<https://www.sciencedirect.com/science/article/pii/S2213596016300216?via%3Dihub>).
- Boyer, Frédéric et al. 2016. "Obitools: A Unix -Inspired Software Package for DNA Metabarcoding." *Molecular Ecology Resources* 16(1):176–82. Retrieved January 19, 2018 (<http://www.ncbi.nlm.nih.gov/pubmed/25959493>).
- Breitwieser, Florian P., Jennifer Lu, and Steven L. Salzberg. 2017. "A Review of Methods and Databases for Metagenomic Classification and Assembly." *Briefings in Bioinformatics*. Retrieved July 10, 2018 (<http://academic.oup.com/bib/article/doi/10.1093/bib/bbx120/4210288/A-review-of-methods-and-databases-for-metagenomic>).
- Cahais, V. et al. 2012. "Reference-Free Transcriptome Assembly in Non-Model Animals from next-Generation Sequencing Data." *Molecular Ecology Resources* 12(5):834–45. Retrieved July 10, 2018 (<http://doi.wiley.com/10.1111/j.1755-0998.2012.03148.x>).
- Camacho, Christiam et al. 2009. "BLAST+: Architecture and Applications." *BMC Bioinformatics* 10(1):421. Retrieved May 29, 2018 (<http://www.ncbi.nlm.nih.gov/pubmed/20003500>).
- Cao, Yu, Séamus Fanning, Sinéad Proos, Kieran Jordan, and Shabarinath Srikumar. 2017. "A Review on the Applications of Next Generation Sequencing Technologies as Applied to Food-Related Microbiome Studies." *Frontiers in Microbiology* 8:1829. Retrieved August 3, 2018 (<http://journal.frontiersin.org/article/10.3389/fmicb.2017.01829/full>).
- Caporaso, J. Gregory et al. 2010. "QIIME Allows Analysis of High-Throughput Community Sequencing Data." *Nature Methods* 7(5):335–36. Retrieved July 11, 2018 (<http://www.ncbi.nlm.nih.gov/pubmed/20383131>).

- Castro-Nallar, Eduardo et al. 2015. “Integrating Microbial and Host Transcriptomics to Characterize Asthma-Associated Microbial Communities.” *BMC Medical Genomics* 8(1):50. Retrieved July 10, 2018 (<http://bmcmmedgenomics.biomedcentral.com/articles/10.1186/s12920-015-0121-1>).
- Conceição-Neto, Nádia et al. 2017. “Viral Gut Metagenomics of Sympatric Wild and Domestic Canids, and Monitoring of Viruses: Insights from an Endangered Wolf Population.” *Ecology and Evolution* 7(12):4135–46. Retrieved July 10, 2018 (<http://doi.wiley.com/10.1002/ece3.2991>).
- Corley, Martin and Sónia Ferreira. 2017. “DNA Barcoding Reveals Sexual Dimorphism in *Isotrias Penedana* Trematerra, 2013 (Lepidoptera: Tortricidae, Chlidanotinae).” *Zootaxa* 4221(5):594. Retrieved July 25, 2018 (<http://biotaxa.org/Zootaxa/article/view/zootaxa.4221.5.7>).
- Corley, Martin, Sónia Ferreira, Alexander Lvovsky, and Jorge Rosete. 2017. “*Borkhausenia Crimnodes* Meyrick, 1912 (Lepidoptera, Oecophoridae), a Southern Hemisphere Species Resident in Portugal.” *Nota Lepidopterologica* 40(1):15–24. Retrieved July 26, 2018 (<http://nl.pensoft.net/articles.php?id=10938>).
- Cosart, Ted et al. 2011. “Exome-Wide DNA Capture and next Generation Sequencing in Domestic and Wild Species.” *BMC Genomics* 12(1):347. Retrieved July 10, 2018 (<http://bmccgenomics.biomedcentral.com/articles/10.1186/1471-2164-12-347>).
- Crawford, Jacob E. et al. 2017. “Population Genomics Reveals That an Anthropophilic Population of *Aedes Aegypti* Mosquitoes in West Africa Recently Gave Rise to American and Asian Populations of This Major Disease Vector.” *BMC Biology* 15(1):16. Retrieved July 10, 2018 (<http://bmcbiol.biomedcentral.com/articles/10.1186/s12915-017-0351-0>).
- D’Amore, Rosalinda et al. 2016. “A Comprehensive Benchmarking Study of Protocols and Sequencing Platforms for 16S rRNA Community Profiling.” *BMC Genomics* 17(1):55. Retrieved August 3, 2018 (<http://www.ncbi.nlm.nih.gov/pubmed/26763898>).
- Deiner, Kristy et al. 2017. “Environmental DNA Metabarcoding: Transforming How We Survey Animal and Plant Communities.” *Molecular Ecology*. Retrieved September 20, 2017 (<http://doi.wiley.com/10.1111/mec.14350>).
- Edgar, Robert C. and Henrik Flyvbjerg. 2015. “Error Filtering, Pair Assembly and Error Correction for next-Generation Sequencing Reads.” *Bioinformatics* 31(21):3476–82. Retrieved July 11, 2018 (<https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv401>).
- Egeter, Bastian et al. 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
- Epp, Laura S. et al. 2012. “New Environmental Metabarcodes for Analysing Soil DNA: Potential for Studying Past and Present Ecosystems.” *Molecular Ecology*.
- Ferreira, Sónia et al. 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
- Fontanesi, Luca et al. 2016. “LaGomiCs—Lagomorph Genomics Consortium: An International Collaborative Effort for Sequencing the Genomes of an Entire Mammalian Order.” *Journal of Heredity* 107(4):295–308. Retrieved July 10, 2018

(<https://academic.oup.com/jhered/article-lookup/doi/10.1093/jhered/esw010>).

Fosso, Bruno et al. 2015. “BioMaS: A Modular Pipeline for Bioinformatic Analysis of Metagenomic AmpliconS.” *BMC Bioinformatics* 16(1):203. Retrieved January 18, 2018 (<http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0595-z>).

Gargani, Maria et al. 2015. “Microsatellite Genotyping of Medieval Cattle from Central Italy Suggests an Old Origin of Chianina and Romagnola Cattle.” *Frontiers in Genetics* 6:68. Retrieved July 10, 2018 (<http://journal.frontiersin.org/article/10.3389/fgene.2015.00068/full>).

Gayral, Philippe et al. 2013. “Reference-Free Population Genomics from Next-Generation Transcriptome Data and the Vertebrate–Invertebrate Gap” edited by J. J. Welch. *PLoS Genetics* 9(4):e1003457. Retrieved July 10, 2018 (<http://dx.plos.org/10.1371/journal.pgen.1003457>).

Gibb, Gillian C. et al. 2016. “Shotgun Mitogenomics Provides a Reference Phylogenetic Framework and Timescale for Living Xenarthrans.” *Molecular Biology and Evolution* 33(3):621–42. Retrieved July 10, 2018 (<https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msv250>).

Glenn, Travis C. 2011. “Field Guide to Next-Generation DNA Sequencers.” *Molecular Ecology Resources* 11(5):759–69. Retrieved August 3, 2018 (<http://doi.wiley.com/10.1111/j.1755-0998.2011.03024.x>).

Hahn, Andrea et al. 2016. “Different next Generation Sequencing Platforms Produce Different Microbial Profiles and Diversity in Cystic Fibrosis Sputum.” *Journal of Microbiological Methods* 130:95–99. Retrieved July 10, 2018 (<https://www.sciencedirect.com/science/article/pii/S0167701216302524?via%3Dihub>).

Hsieh, T. C., K. H. Ma, and Anne Chao. 2016. “INEXT: An R Package for Rarefaction and Extrapolation of Species Diversity (Hill Numbers)” edited by G. McInerney. *Methods in Ecology and Evolution* 7(12):1451–56. Retrieved July 17, 2018 (<http://doi.wiley.com/10.1111/2041-210X.12613>).

Huson, Daniel H. et al. 2016. “MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data” edited by T. Poisot. *PLOS Computational Biology* 12(6):e1004957. Retrieved July 12, 2018 (<http://dx.plos.org/10.1371/journal.pcbi.1004957>).

Jari Oksanen, F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, Gavin L. Pierre Legendre, Dan McGlenn, Peter R. Minchin, R. B. O’Hara, Eduard Szoecs and Helene Simpson, Peter Solymos, M. Henry H. Stevens, and Wagner. 2017. “Vegan: Community Ecology Package.” Retrieved (<https://cran.r-project.org/package=vegan>).

Kim, Mincheol et al. 2013. “Analytical Tools and Databases for Metagenomics in the Next-Generation Sequencing Era.” *Genomics & Informatics* 11(3):102. Retrieved December 4, 2017 (<http://www.ncbi.nlm.nih.gov/pubmed/24124405>).

Kindt, R. and Richard. Coe. 2005. *Tree Diversity Analysis: A Manual and Software for Common Statistical Methods for Ecological and Biodiversity Studies*. World Agroforestry Centre. Retrieved July 13, 2018 (<https://cran.r-project.org/web/packages/BiodiversityR/citation.html>).

Kunin, Victor, Anna Engelbrekton, Howard Ochman, and Philip Hugenholtz. 2010. “Wrinkles in the Rare Biosphere: Pyrosequencing Errors Can Lead to Artificial Inflation of Diversity Estimates.” *Environmental Microbiology* 12(1):118–23. Retrieved July 11, 2018 (<http://doi.wiley.com/10.1111/j.1462-2920.2009.02051.x>).

Langmead, Ben and Steven L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie

- 2.” *Nature Methods* 9(4):357–59. Retrieved July 12, 2018 (<http://www.nature.com/articles/nmeth.1923>).
- Leese, Florian et al. 2016. “DNAqua-Net: Developing New Genetic Tools for Bioassessment and Monitoring of Aquatic Ecosystems in Europe.” *Research Ideas and Outcomes* 2:e11321. Retrieved July 10, 2018 (<http://riojournal.com/articles.php?id=11321>).
- Li, H. and R. Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform.” *Bioinformatics* 25(14):1754–60. Retrieved July 12, 2018 (<http://www.ncbi.nlm.nih.gov/pubmed/19451168>).
- Liu, C. M. et al. 2012. “SOAP3: Ultra-Fast GPU-Based Parallel Alignment Tool for Short Reads.” *Bioinformatics* 28(6):878–79. Retrieved July 12, 2018 (<https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts061>).
- Loire, Etienne et al. 2013. “Population Genomics of the Endangered Giant Galápagos Tortoise.” *Genome Biology* 14(12):R136. Retrieved July 10, 2018 (<http://genomebiology.biomedcentral.com/articles/10.1186/gb-2013-14-12-r136>).
- Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” *Genome Biology* 15(12):550. Retrieved November 18, 2016 (<http://www.ncbi.nlm.nih.gov/pubmed/25516281>).
- Machado, André M. et al. 2018. “De Novo Assembly of the Kidney and Spleen Transcriptomes of the Cosmopolitan Blue Shark, *Prionace glauca*.” *Marine Genomics* 37:50–53. Retrieved July 10, 2018 (<https://www.sciencedirect.com/science/article/pii/S1874778717303173?via%3Dihub>).
- Mahé, Frédéric, Torbjørn Rognes, Christopher Quince, Colomban de Vargas, and Micah Dunthorn. 2015. “Swarm v2: Highly-Scalable and High-Resolution Amplicon Clustering.” *PeerJ* 3:e1420. Retrieved July 12, 2018 (<https://peerj.com/articles/1420>).
- Marques, João P. et al. 2017. “Comparative Mitogenomic Analysis of Three Species of Periwinkles: *Littorina fabalis*, *L. obtusata* and *L. saxatilis*.” *Marine Genomics* 32:41–47. Retrieved July 10, 2018 (<https://www.sciencedirect.com/science/article/pii/S1874778716301337?via%3Dihub>).
- Mata, Vanessa et al. 2016. “Female Dietary Bias towards Large Migratory Moths in the European Free-Tailed Bat (*Tadarida teniotis*).” *Biology Letters* 12(3). Retrieved May 18, 2017 (<http://rsbl.royalsocietypublishing.org/content/12/3/20150988>).
- Mata, Vanessa et al. 2018. “How Much Is Enough? Effects of Technical and Biological Replication on Metabarcoding Dietary Analysis.” *Molecular Ecology* 1–11. Retrieved July 25, 2018 (<http://doi.wiley.com/10.1111/mec.14779>).
- McInnes, Julie C., Simon N. Jarman, et al. 2017. “DNA Metabarcoding as a Marine Conservation and Management Tool: A Circumpolar Examination of Fishery Discards in the Diet of Threatened Albatrosses.” *Frontiers in Marine Science* 4:277. Retrieved July 26, 2018 (<http://journal.frontiersin.org/article/10.3389/fmars.2017.00277/full>).
- McInnes, Julie C., Rachael Alderman, Mary-Anne Lea, et al. 2017. “High Occurrence of Jellyfish Predation by Black-Browed and Campbell Albatross Identified by DNA Metabarcoding.” *Molecular Ecology* 26(18):4831–45. Retrieved July 26, 2018 (<http://doi.wiley.com/10.1111/mec.14245>).
- McInnes, Julie C., Rachael Alderman, Bruce E. Deagle, et al. 2017. “Optimised Scat Collection Protocols for Dietary DNA Metabarcoding in Vertebrates” edited by M. Bunce. *Methods in Ecology and Evolution* 8(2):192–202. Retrieved July 26, 2018 (<http://doi.wiley.com/10.1111/2041-210X.12677>).

- McMurdie, Paul J. and Susan Holmes. 2013. “Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data” edited by M. Watson. *PLoS ONE* 8(4):e61217. Retrieved July 12, 2018 (<http://dx.plos.org/10.1371/journal.pone.0061217>).
- Menegon, Michele et al. 2017. “On Site DNA Barcoding by Nanopore Sequencing” edited by S. R. Gadagkar. *PLOS ONE* 12(10):e0184741. Retrieved July 16, 2018 (<http://dx.plos.org/10.1371/journal.pone.0184741>).
- Murray, Kevin D., Justin O. Borevitz, and Bonnie Berger. 2018. “Axe: Rapid, Competitive Sequence Read Demultiplexing Using a Trie” edited by B. Berger. *Bioinformatics*. Retrieved July 11, 2018 (<https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/bty432/5026649>).
- Mysara, Mohamed, Mercy Njima, Natalie Leys, Jeroen Raes, and Pieter Monsieurs. 2017. “From Reads to Operational Taxonomic Units: An Ensemble Processing Pipeline for MiSeq Amplicon Sequencing Data.” *GigaScience* 6(2):1–10. Retrieved January 15, 2018 (<http://www.ncbi.nlm.nih.gov/pubmed/28369460>).
- O’Donnell, James L., Ryan P. Kelly, Natalie C. Lowell, and Jesse A. Port. 2016. “Indexed PCR Primers Induce Template-Specific Bias in Large-Scale DNA Sequencing Studies” edited by A. R. Mahon. *PLOS ONE* 11(3):e0148698. Retrieved July 11, 2018 (<http://www.ncbi.nlm.nih.gov/pubmed/26950069>).
- Pasolli, Edoardo, Duy Tin Truong, Faizan Malik, Levi Waldron, and Nicola Segata. 2016. “Machine Learning Meta-Analysis of Large Metagenomic Datasets: Tools and Biological Insights” edited by J. A. Eisen. *PLOS Computational Biology* 12(7):e1004977. Retrieved July 17, 2018 (<http://dx.plos.org/10.1371/journal.pcbi.1004977>).
- Paupério, Joana et al. 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
- Pavlopoulos, Anastasis et al. 2015. “Metagenomics: Tools and Insights for Analyzing Next-Generation Sequencing Data Derived from Biodiversity Studies.” *Bioinformatics and Biology Insights* 9:75. Retrieved September 26, 2017 (<http://www.la-press.com/metagenomics-tools-and-insights-for-analyzing-next-generation-sequenci-article-a4809>).
- Pedersen, Mikkel Winther et al. 2015. “Ancient and Modern Environmental DNA.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 370(1660):20130383. Retrieved July 10, 2018 (<http://www.ncbi.nlm.nih.gov/pubmed/25487334>).
- Pereira-Leal, José B. et al. 2014. “A Comprehensive Assessment of the Transcriptome of Cork Oak (*Quercus Suber*) through EST Sequencing.” *BMC Genomics* 15(1):371. Retrieved July 10, 2018 (<http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-15-371>).
- Pérez-Losada, Marcos, Eduardo Castro-Nallar, Matthew L. Bendall, Robert J. Freishtat, and Keith A. Crandall. 2015. “Dual Transcriptomic Profiling of Host and Microbiota during Health and Disease in Pediatric Asthma” edited by B. A. Wilson. *PLOS ONE* 10(6):e0131819. Retrieved July 10, 2018 (<http://dx.plos.org/10.1371/journal.pone.0131819>).
- Porter, Teresita M. and Mehrdad Hajibabaei. 2018. “Scaling up: A Guide to High-Throughput Genomic Approaches for Biodiversity Analysis.” *Molecular Ecology* 27(2):313–38. Retrieved July 3, 2018 (<http://www.ncbi.nlm.nih.gov/pubmed/29292539>).

- Quast, Christian et al. 2013. "The SILVA Ribosomal RNA Gene Database Project: Improved Data Processing and Web-Based Tools." *Nucleic Acids Research* 41(Database issue):D590-6. Retrieved July 17, 2018 (<http://www.ncbi.nlm.nih.gov/pubmed/23193283>).
- Rangwala, Huzefa, Anveshi Charuvaka, and Zeehasham Rasheed. 2014. "Machine Learning Approaches for Metagenomics." Pp. 512–15 in Springer, Berlin, Heidelberg. Retrieved July 17, 2018 (http://link.springer.com/10.1007/978-3-662-44845-8_47).
- Ratnasingham, Sujeevan and Paul D. N. Hebert. 2007. "Bold: The Barcode of Life Data System (<Http://Www.Barcodinglife.Org>)." *Molecular Ecology Notes* 7(3):355–64. Retrieved July 6, 2018 (<http://www.ncbi.nlm.nih.gov/pubmed/18784790>).
- Renaud, Gabriel, Udo Stenzel, Tomislav Maricic, Victor Wiebe, and Janet Kelso. 2015. "DeML: Robust Demultiplexing of Illumina Sequences Using a Likelihood-Based Approach." *Bioinformatics* 31(5):770–72. Retrieved July 11, 2018 (<http://www.ncbi.nlm.nih.gov/pubmed/25359895>).
- Rognes, Torbjørn, Tomáš Flouri, Ben Nichols, Christopher Quince, and Frédéric Mahé. 2016. "VSEARCH: A Versatile Open Source Tool for Metagenomics." *PeerJ* 4:e2584. Retrieved July 11, 2018 (<http://www.ncbi.nlm.nih.gov/pubmed/27781170>).
- Santamaria, M. et al. 2012. "Reference Databases for Taxonomic Assignment in Metagenomics." *Briefings in Bioinformatics* 13(6):682–95. Retrieved July 9, 2018 (<http://www.ncbi.nlm.nih.gov/pubmed/22786784>).
- Santamaria, Monica et al. 2018. "ITSoneDB: A Comprehensive Collection of Eukaryotic Ribosomal RNA Internal Transcribed Spacer 1 (ITS1) Sequences." *Nucleic Acids Research* 46(D1):D127–32. Retrieved July 9, 2018 (<http://www.ncbi.nlm.nih.gov/pubmed/29036529>).
- Schloss, P. D. et al. 2009. "Introducing Mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities." *Applied and Environmental Microbiology* 75(23):7537–41. Retrieved July 11, 2018 (<http://www.ncbi.nlm.nih.gov/pubmed/19801464>).
- Schmieder, R. and R. Edwards. 2011. "Quality Control and Preprocessing of Metagenomic Datasets." *Bioinformatics* 27(6):863–64. Retrieved July 11, 2018 (<https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr026>).
- Schubert, Mikkel, Stinus Lindgreen, and Ludovic Orlando. 2016. "AdapterRemoval v2: Rapid Adapter Trimming, Identification, and Read Merging." *BMC Research Notes* 9(1):88. Retrieved July 11, 2018 (<http://www.ncbi.nlm.nih.gov/pubmed/26868221>).
- Sefc, Kristina M., Robert B. Payne, and Michael D. Sorenson. 2007. "Single Base Errors in PCR Products from Avian Museum Specimens and Their Effect on Estimates of Historical Genetic Diversity." *Conservation Genetics* 8(4):879–84. Retrieved July 11, 2018 (<http://link.springer.com/10.1007/s10592-006-9240-8>).
- Soueidan, Hayssam and Macha Nikolski. 2015. "Machine Learning for Metagenomics: Methods and Tools." Retrieved July 17, 2018 (<http://arxiv.org/abs/1510.06621>).
- Sousa, Lara L. et al. 2016. "DNA Barcoding Identifies a Cosmopolitan Diet in the Ocean Sunfish." *Scientific Reports* 6(1):28762. Retrieved July 10, 2018 (<http://www.nature.com/articles/srep28762>).
- Srivathsan, Amrita et al. 2018. "A MinION-Based Pipeline for Fast and Cost-Effective DNA Barcoding." *BioRxiv* 253625. Retrieved July 16, 2018 (<https://www.biorxiv.org/content/early/2018/01/25/253625>).
- Sunyoung Kwon, Sunyoung, Seunghyun Seunghyun Park, Byunghan Byunghan Lee, and

- Sungroh Sungroh Yoon. 2013. “In-Depth Analysis of Interrelation between Quality Scores and Real Errors in Illumina Reads.” Pp. 635–38 in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, vol. 2013. IEEE. Retrieved July 11, 2018 (<http://www.ncbi.nlm.nih.gov/pubmed/24109767>).
- Taberlet, Pierre, Aurélie Bonin, Lucie Zinger, and Eric Coissac. 2018. *Environmental DNA For Biodiversity Research and Monitoring*. Oxford University Press. Retrieved July 19, 2018 (<http://www.oxfordscholarship.com/view/10.1093/oso/9780198767220.001.0001/oso-9780198767220>).
- Thomsen, Philip Francis and Eske Willerslev. 2015. “Environmental DNA – An Emerging Tool in Conservation for Monitoring Past and Present Biodiversity.” *Biological Conservation* 183:4–18. Retrieved January 18, 2018 (<https://www.sciencedirect.com/science/article/pii/S0006320714004443>).
- Vacher, Corinne et al. 2016. *Learning Ecological Networks from Next-Generation Sequencing Data*.
- Vasconcelos, Raquel et al. 2016. “Unexpectedly High Levels of Cryptic Diversity Uncovered by a Complete DNA Barcoding of Reptiles of the Socotra Archipelago” edited by U. Joger. *PLOS ONE* 11(3):e0149985. Retrieved July 10, 2018 (<http://dx.plos.org/10.1371/journal.pone.0149985>).
- Vesterinen, Eero J. et al. 2016. “What You Need Is What You Eat? Prey Selection by the Bat *Myotis Daubentonii*.” *Molecular Ecology* 25(7):1581–94. Retrieved May 29, 2018 (<http://www.ncbi.nlm.nih.gov/pubmed/26841188>).
- de Vos, Asha, Cassandra E. Faux, James Marthick, Joanne Dickinson, and Simon N. Jarman. 2018. “New Determination of Prey and Parasite Species for Northern Indian Ocean Blue Whales.” *Frontiers in Marine Science* 5:104. Retrieved July 26, 2018 (<http://journal.frontiersin.org/article/10.3389/fmars.2018.00104/full>).
- Wright, Erik Scott and Kalin Horen Vetsigian. 2016. “Quality Filtering of Illumina Index Reads Mitigates Sample Cross-Talk.” *BMC Genomics* 17(1):876. Retrieved July 11, 2018 (<http://www.ncbi.nlm.nih.gov/pubmed/27814679>).
- Yi, Haisi, Zhe Li, Tao Li, and Jindong Zhao. 2015. “Bayexer: An Accurate and Fast Bayesian Demultiplexer for Illumina Sequences.” *Bioinformatics* 31(24):btv501. Retrieved July 11, 2018 (<http://www.ncbi.nlm.nih.gov/pubmed/26315903>).
- Yoccoz, N. G. et al. 2012. “DNA from Soil Mirrors Plant Taxonomic and Growth Form Diversity.” *Molecular Ecology*.
- Zepeda-Mendoza, Marie Lisandra, Kristine Bohmann, Aldo Carmona Baez, and M. Thomas P. Gilbert. 2016. “DAMe: A Toolkit for the Initial Processing of Datasets with PCR Replicates of Double-Tagged Amplicons for DNA Metabarcoding Analyses.” *BMC Research Notes* 9(1):255. Retrieved July 11, 2018 (<http://www.ncbi.nlm.nih.gov/pubmed/27142414>).
- Zepeda Mendoza, Marie Lisandra, Thomas Sicheritz-Pontén, and M. Thomas P. Gilbert. 2015. “Environmental Genes and Genomes: Understanding the Differences and Challenges in the Approaches and Software for Their Analyses.” *Briefings in Bioinformatics* 16(5):745–58. Retrieved January 31, 2018 (<http://www.ncbi.nlm.nih.gov/pubmed/25673291>).
- Zhang, J., K. Kobert, T. Flouri, and A. Stamatakis. 2014. “PEAR: A Fast and Accurate Illumina Paired-End ReAd MergeR.” *Bioinformatics* 30(5):614–20. Retrieved July 11, 2018 (<http://www.ncbi.nlm.nih.gov/pubmed/24142950>).

APPENDIX A: 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 D4.2-D4.5 (Ferreira et al. (2018), Egeter et al. (2018), Paupério et al. (2018) and this document).

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 bennettii* 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 Deliverable 4.4, Egeter 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 Deliverables D4.4 and D4.5 (Paupério et al. (2018) and this document). 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 Deliverable 4.5 (this document).

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).

APPENDIX B: EnvMetaGen CURRENT PROTOCOLS FOR NEXT-GEN DATA PROCESSING

B1. EnvMetaGen protocol for processing eDNA metabarcoding data

1. Demultiplex

- 1.1. Obtain fastq files per sample and by index or barcode using Illumina's BaseSpace program.
- 1.2. Assign each sequence record to the corresponding sample/marker combination (extra demultiplex if needed) and add labelling information on a project basis using obitools 'ngsfilter'.

2. Merge read pairs

- 2.1. Merge read pairs using obitools 'illuminapairedend' with minimum alignment score of 50 (in some cases 40, plots of the distribution of the alignment scores based on their frequency are helpful for checking the threshold).
- 2.2. Remove unaligned sequence records using 'obigrep', 'mode!="joined"'.

3. Dereplication

- 3.1. Dereplicate reads into unique sequences using 'obiuniq'.
- 3.2. Remove reads with a count of 1 (singletons) using 'obigrep'.

4. Additional processing and cleaning

- 4.1. Keep reads within the expected fragment length range using 'obigrep'.
- 4.2. Remove reads with 1 base difference from a more abundant sequence using 'obiclean' (optional).
- 4.3. Remove spurious reads, for instance reads with count < 50 or reads with a count < 1% of the total reads in a sample, using 'obigrep'.
- 4.4. Generate a table with read counts per sample using 'obitab'.

At this stage, fasta files of the filtered unique sequences are produced in addition to their quantification table.

B2. EnvMetaGen protocol for processing InBIO Barcoding Initiative (IBI) DNA data

1. Demultiplex

- 1.1. Obtain fastq files per sample and by index or barcode using Illumina's BaseSpace program.
- 1.2. Assign each sequence record to the corresponding sample/marker combination (extra demultiplex if needed) and combine all sequence records in one fastq file including taxonomic info for each collection specimen using obitools 'ngsfilter'.

2. Merge read pairs

- 2.1. Merge read pairs using obitools 'illuminapairedend' with minimum alignment score of 50.
- 2.2. Remove unaligned sequence records using 'obigrep', 'mode!="joined"'.

3. Dereplication

- 3.1. Dereplicate reads into unique sequences per sample using 'obiuniq -c sample'.

4. Additional processing and cleaning

- 4.1. Obtain count statistics for each sample using 'obistat'.
- 4.2. Keep unique sequences with count > 5 and length > 50bp using 'obigrep -p 'count>=6' -l 50'.
- 4.3. Add field with sample and count information using 'obiannotate'.
- 4.4. Remove reads with 1 base difference and abundance lower than 10% of a more abundant sequence using 'obiclean -H -r 0.1' (likely PCR/sequencing errors).
- 4.5. Remove unused attributes (e.g. experiment, seq_length, obiclean_count, etc.) using 'obiannotate'.
- 4.6. Sort sequences by decreasing counts using 'obisort'.
- 4.7. Change sequence ID to specimen code using 'obiannotate'.
- 4.8. Split sequences by insect order using 'obisplit, -t order' (optional).

At this stage, fasta files of the filtered unique sequences containing the specimen codes are produced.