# PROJECT DELIVERABLE REPORT

| | |
|---|---|
| DELIVERABLE NUMBER AND TITLE | D3.3.1 |
| TITLE | SPECIFICATION OF THE TOOLS |
| AUTHOR(S) | DR. M. SANTAMARIA, DR. B. BALECH, G. MAGGI, R. LOMBARDO |
| WORK PACKAGE | WP 3 |
| TASK | TASK 3.3 |
| WP LEADER | ROSARIO LOMBARDO, CORRADO PRIAMI |
| BENEFICIARIES CONTRIBUTING TO THE DELIVERABLE | IBBE, POLIBA |
| STATUS – VERSION | FINAL - VERSION 2.0 |
| DELIVERY DATE (MONTH) | M12 |
| SUBMISSION DATE | M14<br>M28 – FINAL REVISION |
| DISSEMINATION LEVEL – SECURITY* | PU |
| DELIVERABLE TYPE** | R |

* Security: PU – *Public*; PP – *Restricted to other programme participants (including JPI Services)*: RE – *Restricted to a group specified by the consortium (including JPI Services)*; CO – *Confidential, only for members of the consortium (including JPI Services)*

** Type: R – *Report*; P – *Prototype*; D – *Demonstrator*; - O - *Other*

## CONTENTS

Scope and significant results ..... 3

Short description of activities and intermediate results ..... 3

References ..... 8

Problems/ challenges/ deviations from proposal/work plan ..... 9

## SCOPE AND SIGNIFICANT RESULTS

In Task 3.3 we deal with the identification of the relevant tools and resources in the framework[1] of human molecular nutrition studies. These are adopted according to the case studies (WP2) and the experimental design (i.e. Metagenomics) and then standardized to cover vast types of datasets[2]. The tools are designed in a way to be easily accessible by a web browser, reused, updated and permanently maintained. For that, we identified several available databases, datasets and bioinformatic analysis tools concerning genomic and metagenomic data in relation to human nutrition. In addition, we are conducting the implementation of curated genomic and metagenomic databases and analysis tools. In particular, in order to design the database structure and populate it, we are referring to the use cases defined in WP2 and ontology terms from WP4. Tasks 3.5 developed the database infrastructure and, at IBBE, we are regularly optimizing and updating the databases and various bioinformatic pipelines[3]aimed at the study of environmental and human microbiomes, as one the many study-storage and analysis possibilities offered by the DASH-IN infrastructure (see Deliverable "D3.2.1 Functional/technical requirement data infrastructure"). These resources are exposed and used as web services and/or web applications. Finally, a metadata search tool has been designed from scratch based on the suggestions and requirements from a multidisciplinary task force drawn from the ENPADASI project. The metadata search features in fact combines the needs and information from biological case studies (WP2), technical infrastructure (WP3) and ontologies (WP4), and allows the identification of the proper data stored in the DASH-IN infrastructure that best matches any given research query. Other available third party tools for genomics and metagenomics data analysis will also be investigated and proposed to be included in the infrastructure.

## SHORT DESCRIPTION OF ACTIVITIES AND INTERMEDIATE RESULTS

The goal for the activities carried out and the corresponding intermediate results regarded the identification of bioinformatic pipelines and their reference databases suitable for the analysis of genomic and metagenomic datasets from nutritional sequencing studies. In this context, we recognized the available resources needed to setup the bioinformatics genomic and metagenomics data analysis pipelines, namely BioMaS, MetaShot and MSA-PAD. The first two pipelines are aimed at taxonomic analysis of microbial communities through Metagenomics. Indeed the application of metagenomic approach in the field of nutrition, from the monitoring of food supply chains to the study of different diets effect on human health and welfare, are experiencing an unprecedented expansion. The third pipeline provides a generic multiple sequence alignment workflow[4], useful for SNP (Single Nucleotide Polymorphism) calling and phylogenetic analysis of target genomic fragments or genes involved in the response on nutritional interventions/treatments.

The implementation of these pipelines and their integrated reference databases on the Bari-ReCaS e-Infrastructure aims to provide user-friendly interface with easily executable workflows able to process genomic and metagenomic sequence data produced in nutrition studies, already collected in the ENPADASI project or provided by the user. Their adaptation to interacting with Opal and Phenotype Database resources available at the ReCaS Datacenter will allow to store and share the analysis results directly through these

[1] a supporting structure around which something can be built.
[2] Collection of data.
[3] a chain of data-processing stages.
[4] a series of computational or data manipulation steps executed in order.

databases. As mentioned above, BioMaS and MSA-PAD can be already executed on Bari-ReCaS e-Infrastructure, the first to profile microbiomes composition from amplicon-based Metagenomics datasets, as those produced from human microbiomes in nutritional studies, and the second to multiple align target DNA coding sequences. The implementation of Metashot, a recently developed pipeline aiming at taxonomic profiling of shotgun sequenced human-associated microbiomes, is currently in progress. The output of BioMaS analysis has already been tailored to the formats required to populate Opal and Phenotype DB. The adaptation of MSA-PAD and MetaShot outputs to these resources are currently in progress.
Additional value was added by designing a metadata search tool intended to identify relevant studies for analyses, located in the geographically distributed infrastructure DASH-IN (see also D3.2.1) established throughout the European countries participating to the ENPADASI project. Once relevant studies are located in the infrastructure they can be analyzed via the genomic/metagenomic tools, described in this report, but also with the DataShield and R[5] tools mentioned here but described more in detail in D3.2.1.

**BioMaS** – (Bioinformatic analysis of Metagenomic AmpliconS) (Fosso et al., 2015) carries out a sequential[6] and entirely automated workflow structured in consecutively running pre-built modules which basically accomplish the assessment of the Meta-barcoding[7]High-Throughput Sequencing (HTS) data quality, their clustering according to the original samples, the removal of sequence errors noise, the comparison with reference databases and, finally, the taxonomic binning[8] and deep annotation. BioMas has been benchmarked with QIIME[9] and Mothur[10]. The benchmark results demonstrated that BioMaS outperforms both the other pipelines mainly at deeper taxonomic levels and highlighted its accuracy in revealing also the quantitative differences between microbial species represented in multiple samples. This property is very important in metagenomic studies in which microbial population dynamics are correlated to a number of variables, as generally happens in observational and interventional nutrition studies.

**MetaShot** – (Metagenomics Shotgun) is a complete pipeline designed for the taxonomic classification of the human microbiota starting from shotgun Metagenomics HTS data. It runs pre-built Python[11] and BASH[12] scripts performing the following procedure:

(i) low-quality and low-complexity sequences removal.

(ii) Automatic comparison with Prokaryotes, Virus, Fungi and Protista custom collections and Homo Sapiens genome and transcriptome data and filtering of alignments according to identity percentage (≥ 97%) and query coverage (≥ 70%). Sequences mapping on only one reference collection are taxonomically classified, the others are considered ambiguous.

(iii) Ambiguous sequences mapping on both Human and Virus are taxonomically classified on the basis of their match with the viral reference collection in order to identify HERV sequences.

(iv) Report generation: a CSV file, an HTML interactive table summarizing the taxonomic assignment and a krona[13] graph of the obtained taxonomy are provided for Prokaryotes, Virus, Fungi and

[5] R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS.
[6] following a particular order.
[7] A rapid method of biodiversity assessment that combines DNA based identification with high-throughput DNA sequencing.
[8] The process of grouping sequencing reads and assigning them to operational taxonomic units
[9] An open-source bioinformatics pipeline for performing microbiome analysis from raw DNA sequencing data.
[10] An open-source bioinformatics workbench for microbial ecology DNA data analysis.
[11] Programming language.
[12] Bash is a command processor that typically runs in a text window, where the user types commands that cause actions.
[13] Multi-layered pie charts.

Protista. MetaShot has been benchmarked with Kraken (Wood and Salzberg, 2014) and MetaPhlAn2 (Truong et al., 2015).

The benchmarking results of MetaShot versus Kraken and MetaPhlAn2 showed the outperformance of MetaShot over the other pipelines in taxonomic assignment accuracy of microbial sequences at Genus and Species levels either qualitatively or quantitatively.

**MSA-PAD** – (Balech et al., 2015) is a DNA multiple sequence alignment (MSA) framework that uses protein domain information to align DNA sequences encoding either single or multiple protein domains. It conceptually translates DNA sequences into amino acids (based on user-defined genetic code and reading frame/s), uses information from protein domains, PFAM (Finn et al., 2014) and/or user's profiles, to assign the translated sequences to known protein domains, accounts for frameshifts when domain regions are split by introns, performs a domain-based protein alignment and then uses protein alignment information to generate the relevant nucleotide multiple alignment. The final MSA can be generated following two different strategies: (i) Gene or (ii) Genome mode. Gene mode alignment respects domain order organization from 5' to 3', and resolves the alignment of repetitive domains even when they are repeated in tandem. Genome mode alignment provides a super-gene-like alignment ignoring domain order constraints.

In collaboration with POLIBA, BioMaS and MSA-PAD were deployed[14] and tested on ReCaS e-infrastructure, in particular the Bari ReCaS datacentre. Among the identified data resources we identified RDP II (Ribosomal Database Project II) (Cole et al., 2009) and GreenGenes (DeSantis et al., 2006), two collections of 16S rRNA sequences suitable for prokaryotic taxa identification, ITSoneDB (Santamaria et al., 2012), a collection of ITS1 sequences designed for supporting the taxonomic characterization of Fungi, GenBank and RefSeq (Sayers et al., 2009), Hg19 (Human Gernome hg19, GRCh37. 2009), UCSC refseq transcript (Speir et al., 2016) and PFAM (Finn et al., 2014), a database of curated protein families.

Some of the above tools, algorithms and pipelines (such as BioMaS and MSA-PAD) are available for the whole ENPADASI scientific community and can be already executed on the Bari-ReCaS e-Infrastructure by means of the Galaxy workflow manager or the Liferay portal (purl: https://recasgateway.cloud.ba.infn.it/), both of them making use of the Job Submission Tool (JST). The Bari ReCaS Datacenter is a medium sized computer farm with more than 8000 CPU cores, 5 PB of disk storage and 2.5 PB of tape storage, where the input and the output data produced by the above described algorithms can be stored and managed using "Cloud Storage" services based on WebDav[15] and ownCloud[16]. The Job Submission Tool (JST) was initially developed in order to simplify the submission, management and bookkeeping of large number of jobs required by particular applications. JST has been adapted to the ENPADASI requirements: in particular a WEB services interface has been added. In this way it is now possible to access JST also from within workflow managers like Taverna, LONI, Galaxy and other similar tools.

Within the parallel development with the ongoing work in complementary efforts in ELIXIR we evaluated the cross-integrations with the EBI Metagenomic portal and the activities to be carried out in the EXCELERATE project within the ELIXIR RI framework.

[14] Making available for use.
[15] Web Distributed Authoring and Versioning allowing the users to move and change documents on a web server.
[16] An open source file synchronization and share software.

A road map concerning the pipelines integration in the workflows for genomic and metagenomic data analysis has been agreed. The idea behind that was to implement the needed web services that can be easily accessed by a simple web browser or command line client even by non-experts users. In addition, the possibility to expose data collections or analysis tools directly as services developed on ReCaS computational platform would be further pursued as it would allow taking great advantage from an already well-established expertise and collaboration.

ReCaS Data centre provides also an access to the entire ENPADASI community to Opal, DataShield, Mica and Phenotype Database systems. These applications can host structured observation and intervention studies data and their associated metadata. In details, Opal and Phenotype Database can store processed data coming from several laboratory platforms (e.g. biochemical analyses, metabolomics, metagenomics, etc…) and allow the visualization of their summary statistics. DataShield gives the possibility to use additional or more complex statistical features such as the generalized linear models to interpret the data. Mica can be connected to many Opal instances[17] at the back-end side and allows mainly the annotation using metadata and the experimental design of studies present in Opal. The Mica instance available at ReCaS has been selected as the centralized instance for ENPADASI e-infrastructure.

**Metadata search** – As the DASH-IN infrastructure developed by WP3 is growing in number of studies and storage sites across Europe, the need for a specific metadata search tool has arisen to allow finding studies whose characteristics best match the user's research questions. The metadata search tool is presented to the user as a simple search-box available at a central DASH-IN portal. The central portal interacts with 2 families of servers spread across Europe:

a) a number of independent Phenotype Databases (DBNP, for intervention and observational studies)

b) a central Mica server which contain metadata information referring to studies stored on multiple Opal servers (for observational studies).

Credentials-based search access to the different systems is achieved on the same basis of the Web-based Dash-In federated analysis systems extensively described in Deliverable "D3.2.1 Functional/technical requirement data infrastructure". A brief description of the search tool is described as follows:

a) Studies from different instances of DBNP can be found by a "search" access credential setup for each DBNP server. Metadata for public studies are accessible even without credentials. Appropriate users can search account-protected studies. Studies data cannot be accessed by the search tool itself.

b) Studies from the central MICA are found by providing a MICA credential. Only studies-related metadata are stored on MICA server and not the data itself. Therefore, studies cannot be accessed from the search tool.

The central search box will allow to search user-provided query terms contained both in the Mica server (acting a central metadata point for OPAL servers) and in each independent DBNP, which have no federation/centralized concept. In Figure 1 is reported a screenshot of the metadata search tool working with the MICA server Bari (currently the only metadata API available to the search box). Data-uploaders/ data-curators can annotate their studies with the metadata terms for each of the studies in the respective DB system (either DBNP or MICA). This action makes the corresponding studies visible from the metadata search-box at the central Dash-In server.

[17] An instance is a concrete occurrence of any object (in this case Opal), existing usually during the runtime of a computer program.

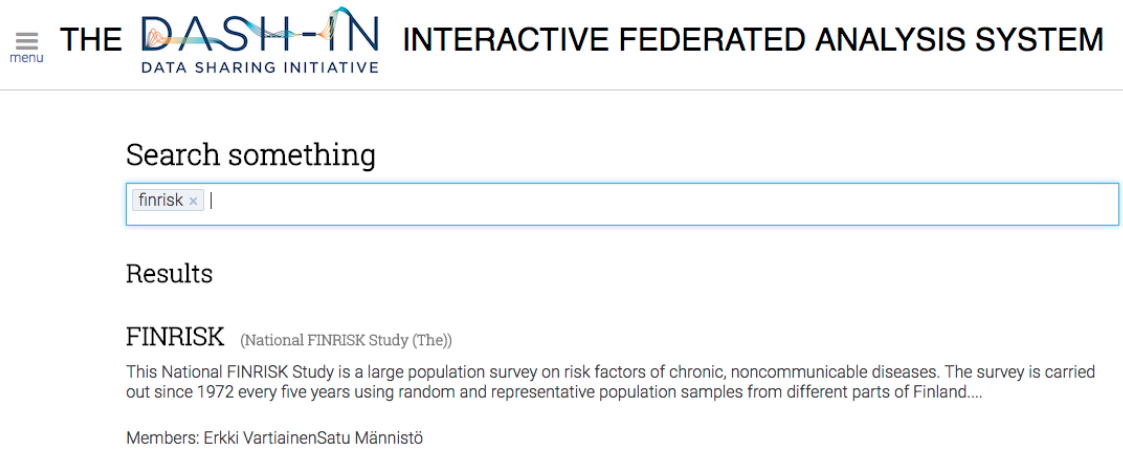
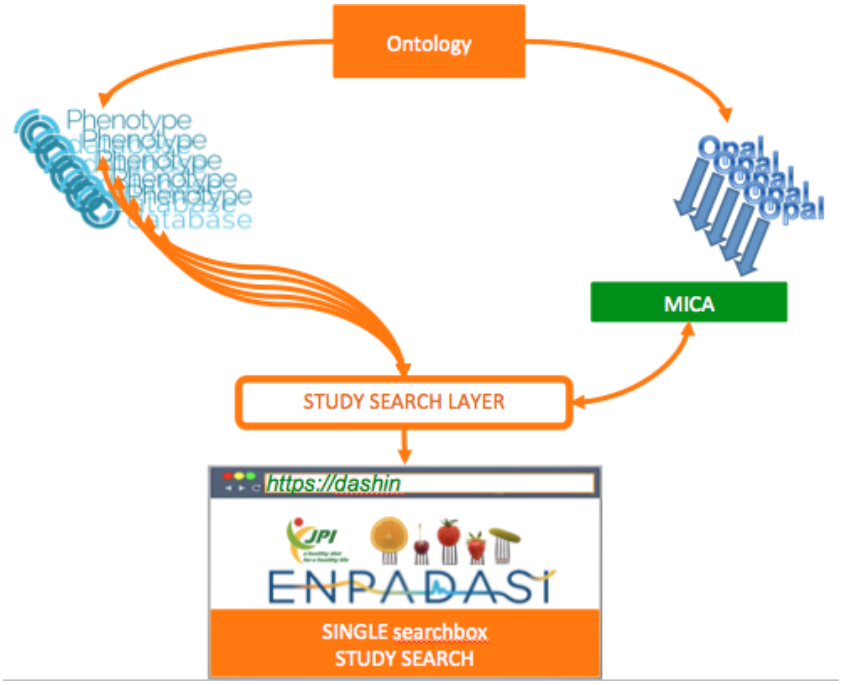



**Figure 1**: Study metadata search performed using the DASH-IN infrastructure looking for a sample data set loaded on the MICA central metadata server in Bari.

The search results consist of a list of study entries each containing only metadata information on the studies, and particularly the owner/server allowing to get in contact and obtain credentials to access the data (if not already available to the user) and perform analytical tasks such as those described above. Additional user-friendly statistical analyses such as regressions, generalized linear models and exploratory data-undisclosing plots can be setup and applied from the web-based Dash-In portal (see more details in “Deliverable D3.2.1 Functional/technical requirement data infrastructure”).

The task force identified a set of study search terms to be used for both study annotation and study search which are based on WP4 ontologies but might be extended with terms derived from: minimal study requirements (WP2), study quality appraisal tool (WP2) and potential biomarker ontologies from shared partnerships with the FoodBALL project.

Graphical overview of the metadata search tool where the constituent components from WP3 infrastructure are interacting together to provide the best data for the user research needs.




## REFERENCES

Balech, B., et al. (2015) **MSA-PAD: DNA multiple sequence alignment framework based on PFAM accessed domain information**, Bioinformatics, 31, 2571-2573.

Cole, J.R., et al. (2009) **The Ribosomal Database Project: improved alignments and new tools for rRNA analysis**, Nucleic acids research, 37, D141-145.

DeSantis, T.Z., et al. (2006) **Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB**, Applied and environmental microbiology, 72, 5069-5072.

Finn, R.D., et al. (2014) **Pfam: the protein families database**, Nucleic acids research, 42, D222-D230.
Fosso, B., et al. (2015) **BioMaS: a modular pipeline for Bioinformatic analysis of Metagenomic AmpliconS**, BMC bioinformatics, 16, 203.

Santamaria, M., et al. (2012) **Reference databases for taxonomic assignment in metagenomics**, Briefings in bioinformatics, 13, 682-695.

Speir, M.L., et al. (2016) **The UCSC Genome Browser database: 2016 update**, Nucleic acids research, 44, D717-725.

Truong, D.T., et al. (2015) **MetaPhlAn2 for enhanced metagenomic taxonomic profiling**, Nat Methods, 12, 902-903.

Wood, D.E. and Salzberg, S.L. (2014) **Kraken: ultrafast metagenomic sequence classification using exact alignments**, Genome Biol, 15:R46.

## PROBLEMS/ CHALLENGES/ DEVIATIONS FROM PROPOSAL/WORK PLAN

The main challenge faced during this project period is the funding delay. This has limited human resources to accomplish the study itself and in conjunction with the other partners (WP2/WP4) facing the same situation. The activities were readjusted and conducted according the new plan.