EGI-ACE Open Call no.1

3rd checkpoint meeting with shepherds

eHoney

Stefano Nicotri
INFN

EGI-ACE receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 101017567.

Outline

- *Background about the scientific use case*
- *Ambition, Impact and Challenges*
- *Integration Support*
- *Capacity Requirements*
- *Timeline*

Background about the scientific use case

Objective of the use case

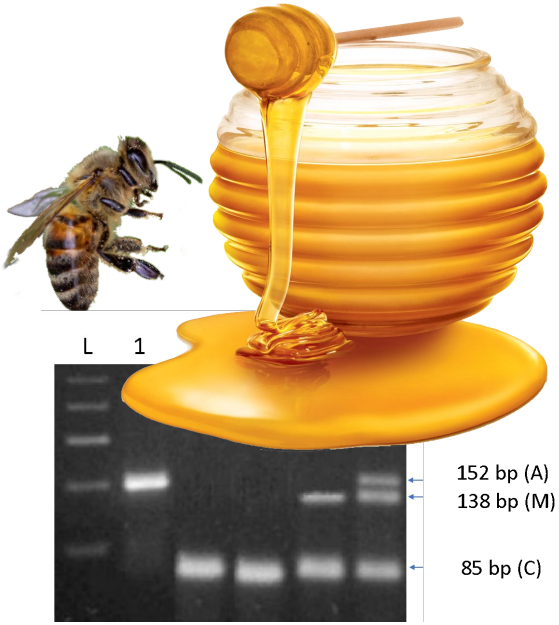
- eHoney aims at setting up a new avenue to analyse the effect of climate change on the biodiversity using retrospectively archived honey samples.

Approach

- eHoney relies on the bioinformatics analysis of environmental DNA (eDNA) information derived from more than 500 honey samples produced over the last decade.

Adopter

- The Animal and Food Genomics Group of the Department of Agricultural and Food Sciences of the University of Bologna (Bologna, Italy).
- The scientific community through Open Access and FAIR data)



Background about the scientific use case

Team members

- Located at the University of Bologna, Bologna, Italy (IT)
- Scientists are specialized in apiculture genomics and eDNA analyses
- Scientists have different backgrounds, including Bioinformatics (dry-lab) and Molecular Biology/Genetics (wet-lab)

L. Fontanesi
Professor
Group leader

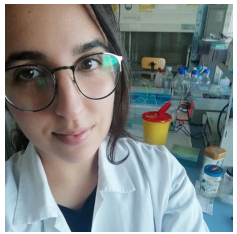
G. Schiavo
Researcher
Dry lab

S. Bovo
Researcher
Dry lab

M. Ballan
PhD student
Dry lab

A. Ribani
PostDoc
Wet lab

V.J. Utzeri
PostDoc
Wet lab



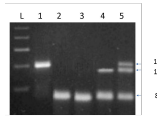
V. Taurisano
PhD student
Wet lab

Ambition, Impact, Challenge(s)

Ambition: analyse the effect of climate change on the biodiversity via an efficient process of eDNA sequence annotation from honey sample (i.e. big data processing), including algorithm redefinition and pipeline scaling up.

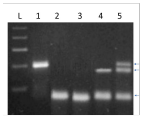
Impact: it would be possible to rapidly process eDNA sequences from honey and monitor, in real time, the level of biodiversity and the effect of climate change .

Challenge: computational analyses of the eDNA sequencing datasets represent the major bottleneck of the project (DNA sequencer produces up to billions of DNA sequences that need to be processed in order to determine the source organism) in a process called sequence annotation. The Implementation of an **efficient and scalable bioinformatics pipeline** represent a challenge in rapid data processing.



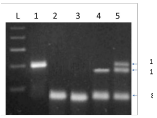
Integration Support

- Simple eDNA sequencing of one single honey sample can produce raw data for up to 2.7 million of DNA sequences. Data storage requires up to 1.6GB of physical space. Deep metagenomic sequencing can produce 10× the listed amount of data.
- Available data: 1000 and 500 sequencing datasets coming from a simple and deep metagenomic sequencing, respectively. The raw datasets include a total 8-10 TB of data that will be processed
- The analysis of eDNA sequencing data, currently relies on two small machines, each one equipped with a dual 6-core processor, around 42 GB of RAM and 18TB of storage;
- Based on this setting (using 4 threads), analyses of one eDNA sample can take up to 6 days and needs between 4 to 40 GB of physical space for data/results storage

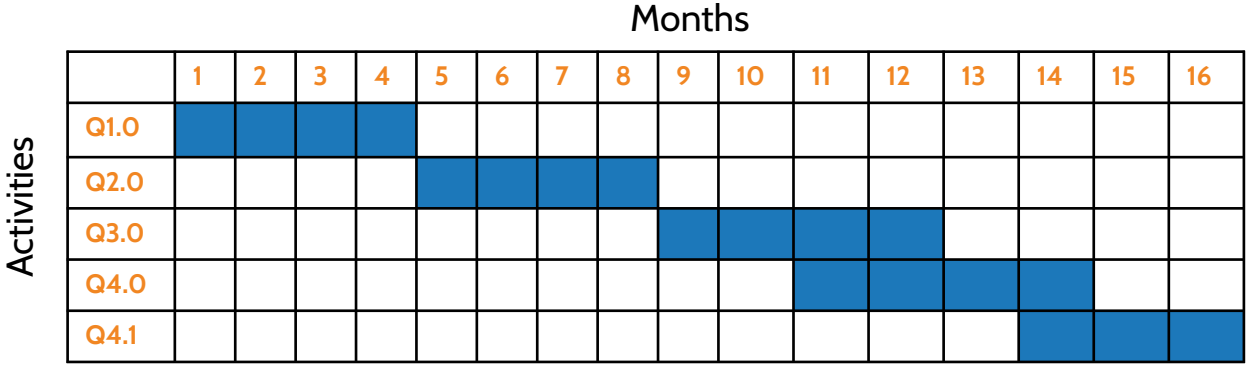


Capacity Requirements

- Datasets: about 8-10TB of metagenomics raw data to be processed (30TB is the total estimated requirement for this use case)
- Current requirements / usage: 8 threads, 20Gb of RAM for the analysis of each eDNA (honey) sample.
- A re-definition of the currently used pipeline used for data analysis is expected, to scale-up the process of DNA annotation.



Timeline



The use case will last 16 months and it will have the following activities:

- Q1.0: set up the pipelines for eDNA read annotation. Two approaches will be tested;
- Q2.0: analysis of the whole datasets;
- Q3.0: interpretation of the data;
- Q4.0: comparisons of results from the two approaches and re-definition of algorithms;
- Q4.1: construction of Open and FAIR based databases and preparation of Open Access publications

Current status (shepherd point of view)

- updated customer DB with current status
- technical plan finalized (currently respected)
- order placed in EOSC marketplace
- resources (partially) provided by INFN-CLOUD-BARI (ReCaS-Bari)
- VO creation in progress (input needed)
- local resources → accounting yet to be formalized (in progress)
- waiting for user input to correct the amount/type of provided resources

Current status (project point of view)

- more data transferred on the ReCaS-Bari (INFN) HTC infrastructure → quite large processing → the outcomes are good
- current status in sync with the timeline
- a code upgrade is planned for the near future
- next steps: finish transferring the whole pipeline on the infrastructure in order to have a precise idea of the required resources and to explore the possibility of use parallel computing techniques (which will maybe require HPC resources)

Thank you!

Contact: **egi-ace-po@mailman.egi.eu**
Website: **www.egi.eu/projects/egi-ace**

EGI Foundation

@EGI_eInfra

EGI-ACE receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 101017567.