Pilot: Marine metagenomics

Towards domain specific service

ELIXIR All Hands 2016, 8-9 March, Barcelona, Spain

The study of genetic material sampled directly from marine environmental sources is still in its infancy, but is rapidly expanding. To prevent that processing and analysis of these samples becomes a bottleneck, where data production is faster than the speed users are able to make use of it, there is an urgent need to establish dedicated data management e-infrastructure and bioinformatics pipelines specialized for marine research. While EBI has developed EBI-metagenomics, a generic pipeline, which aims to provide insights into the phylogenetic diversity as well as the functional and metabolic potential of the samples, the Norwegian node has developed META-pipe in the direction coupled with marine bioprospecting. In this pilot project the two pipelines will be harmonized in terms of interoperability in order to establish long-term sustainable service platforms and build a user community for marine metagenomics analysis in ELIXIR.

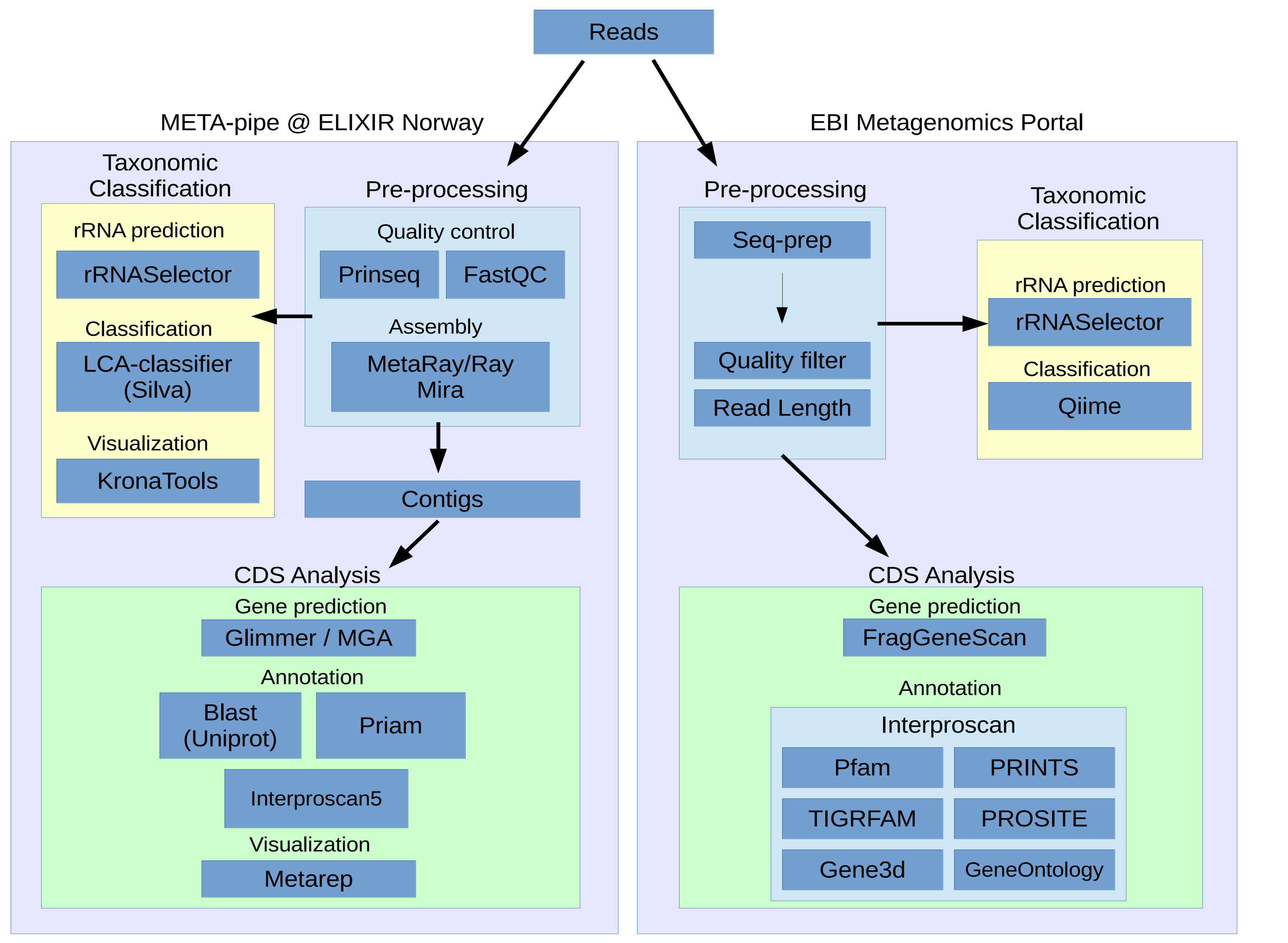

Figure 1: A comparison between the two pipelines Meta-pipe (left) and Ebi Metagenomics (right). Both pipelines provide similar functionality, but with different modules and approaches. Most noteworthy, Meta-pipe utilizes assembled contigs as input to produce mainly full length sequences, while EBI metagenomics accepts reads as input to produce a higher quantity, but mostly fragmented set of sequences

Deliverables

[1] Harmonizing existing metagenomics pipelines (EBI-metagenomics and META-Pipe) and to agree on data formats to ensure interoperability.

[2] Assessing new functionally specialized database (e.g. Interpro, MEROPS and SFLD) and evaluate if these resources can enhance or enrich the output

[3] Explore and prototype with EMBL-EBI cloud based technologies.

[4] Investigate the use of other approaches for taxonomic assignment, expanding beyond prokaryotic assignments.

[5] Gap analysis related to establishment of reference genomes for the marine environment

Results

We compared the two pipelines (Figure 1) and identified specific intermediate pipeline steps where components needed to be improved or developed, such as pre-processing and the implementation of Interproscan in META-pipe. We compared outputs of the pipelines using two marine metagenomic samples from the Barents Sea and two gut samples from sea urchin and moose taken from local spots in Northern Norway to get a better understanding of differences in biological analysis (Figure 2,3)

EMBL-EBI Embassy Cloud services was used to set up a virtual cluster with 15 nodes total and a shared NFS file system (Figure 4). To experiment with the possibilities this cloud service provides, a Blast job against UniprotKB using Slurm as a scheduler was computed successfully.

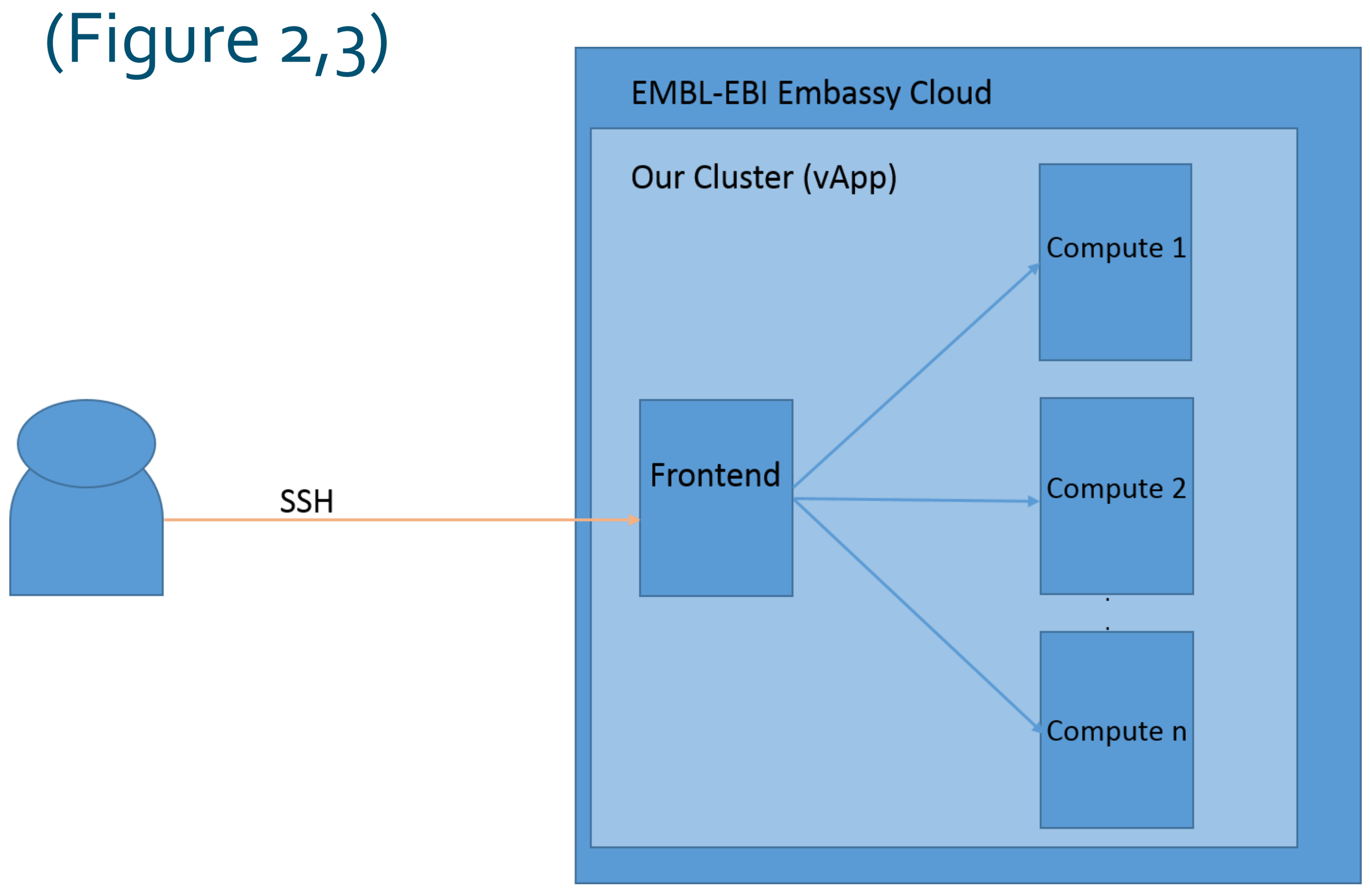

Figure 4: A schematic overview of the prototyping process with EMBL-EBI Embassy Cloud. Access to the virtual cluster via SSH, with one front end and 14 compute nodes.

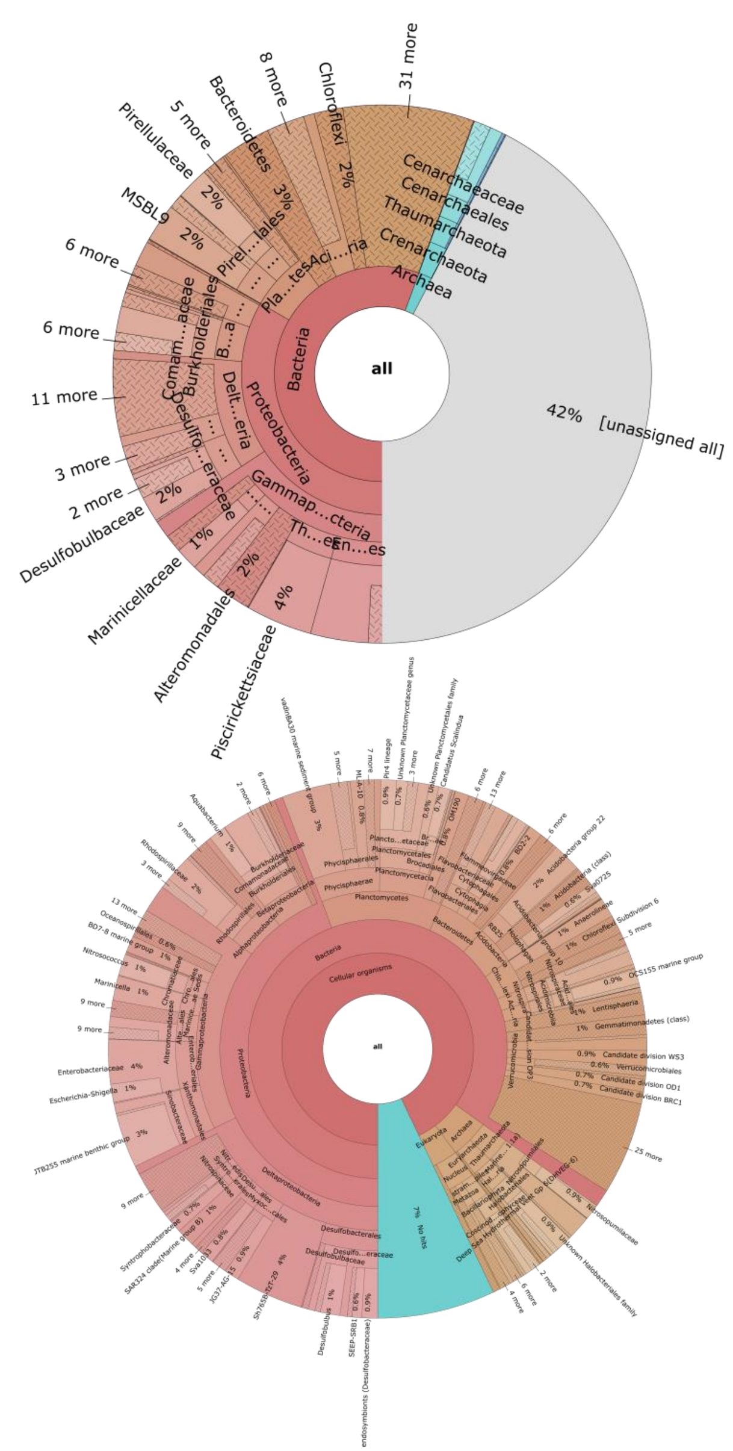

Figure 2: A taxonomic comparison of one of the Barents Sea samples coined "Muddy" between EBI Metagenomics (top) and META-pipe (bottom).

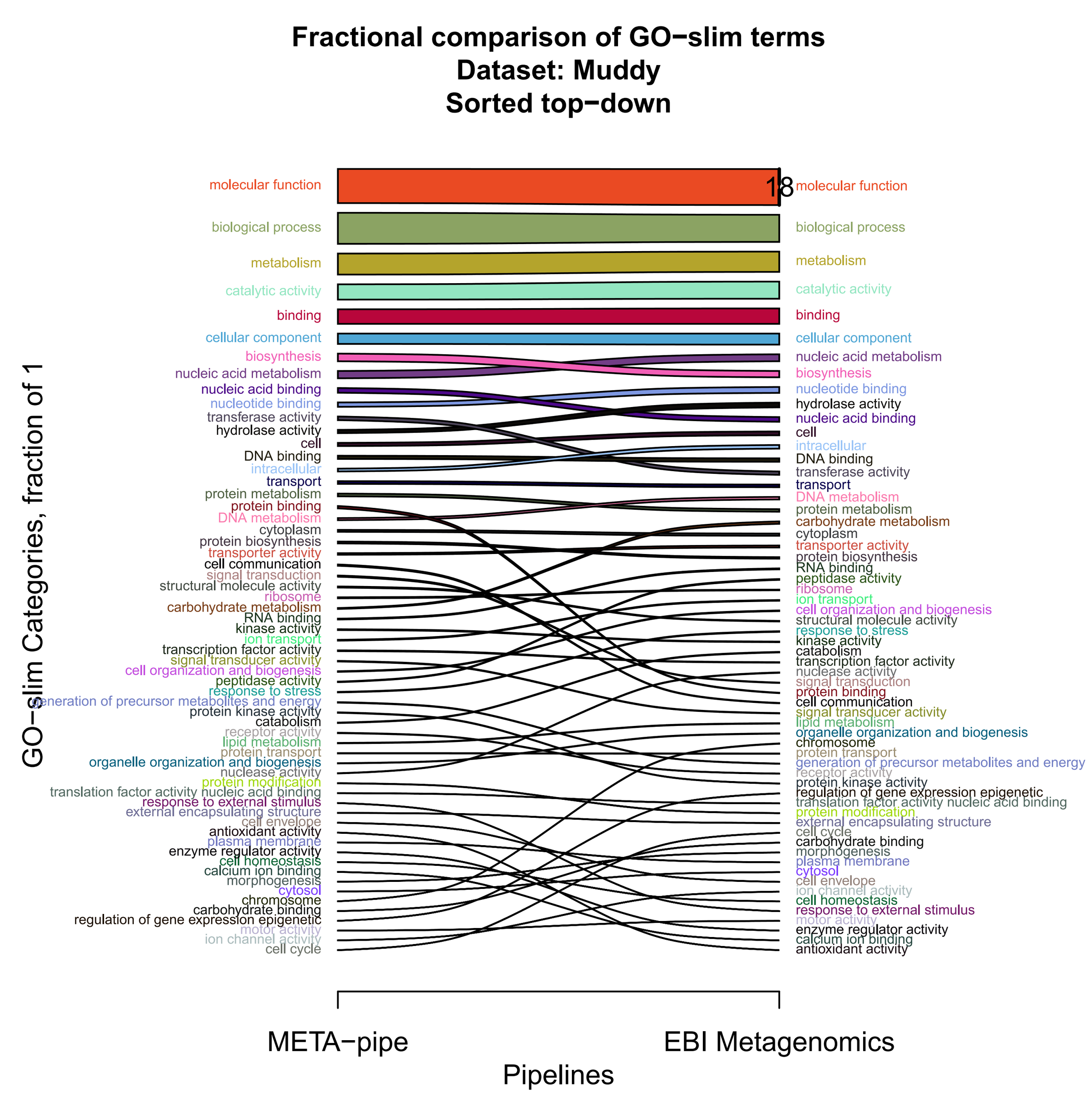

Figure 3: Fractional comparison of GO-Slim terms from predicted genes within the Barents Sea sample "Muddy" shows how an assembly effects functional analysis.

Contact:

Nils Peder Willassen
nils-peder.willassen@uit.no
Department of Chemistry
University of Tromsø

@ELIXIREurope

/company/elixir-europe