HELIX NEBULA THESCIENCECLOUD

HNSciCloud
User Groups

Life Science

# Pan-Cancer

## Supporting large scale genomics analysis in cancer studies

The Pan-Cancer initiative primary goal is to compare 12 tumor types profiled in the context of The Cancer Genome Atlas Research Network. Cancer can take hundreds of different forms, depending on external factors such as localization and cell type. Pan-Cancer will look for shared molecular aberrations and will try to identify their impact on the evolution of the disease. Eventually, this will help extend therapies that are already known to be effective against a certain type of cancer to similar cancer types at the genomic level.

## Goal

Using HNSciCloud, the PanCancer project will be able to determine genetic variation for more than 5000 tumor samples with more coming in on a monthly basis. PanCancer currently represents the most comprehensive computational study dealing with cancer genomics, with roughly 1 PB of data to be processed. This has forced researchers to implement new pipelines able to cope with the massive quantity of data, with a focus on leveraging cloud resources provided by public and commercial clouds, including the EMBL-EBI Embassy Cloud. To process this data at the appropriate scale in a cloud environment, the Butler workflow management was developed. Computationally, HNSciCloud has been able to match the high-end environment of EMBL-EBI’s own EMBASSY cloud, leading to the establishment of a viable PanCancer deployment in the T-Systems environment.

## Preconditions

1. Data transparency layer performance must be tested to have the same performance in the other cloud providers (RHEA and Advania)
2. Data security and provenance must be guaranteed by legal agreements with each cloud provider.

## Handling human data

The very first step Pan-Cancer took was to create stable and coherent datasets comprising ~5000 tumor samples, including genomic, epigenomic and gene and protein expression data. As a result, 1PB of data is currently held at EMBL-EBI and is available for download and analysis by any authorised researcher. The same data is replicated in several different repositories worldwide. All the datasets are constituted by human data, with stringent implications on privacy protection. For this reason, data access is strictly regulated, and data requests must be authorised by the governing bodies of the project.

www.hnscicloud.eu | @HelixNebulaSC

European Commission

Helix Nebula - The Science Cloud (HNSciCloud) with Grant Agreement number 687614 is a Pre- Commercial Procurement Action funded by the EU Framework Programme for Research and Innovation Horizon 2020.

## The Challenge

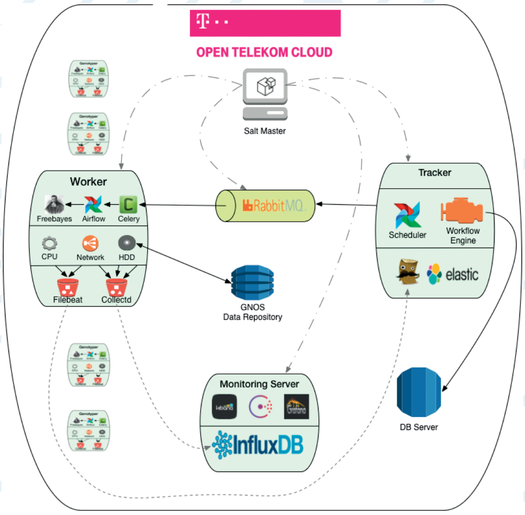
Deployment of this project onto a commercial cloud infrastructure, given the amount of data involved and the scalability needed to deliver results in a timely manner, should be able to give insights into the general feasibility of adopting commercial cloud environments to fulfil a big data challenge in the bioinformatics context. In particular, we will be able to assess two key factors on this type of deployment: network connectivity to and from the cloud provider, along with ingress and egress costs, and the feasibility and ease of scaling to thousands of VMs running concurrently that need to access both local (to the VM) and shared (to all the VMs) storage.

In the long term, the challenge is to foster adoption of cloud-based solutions as a way to cope with peak in computational demand (e.g. bursting to the cloud), even when big data sets are involved

## Benefits and impact

The extent of the Pan-Cancer project in terms of data to be processed in a reasonable time scale represents a perfect use case to explore the feasibility of bursting such workloads to commercial clouds when time or restricted availability of internal resources (available from multiple institutes) are a constraint. From a technical point of view, successfully exploiting the hybrid cloud model could foster a broader adoption of external cloud resources by research initiatives facing similar data challenges. From the cancer genetics research perspective, Pan-Cancer results will represent a key step in disentangling the very complex panorama represented by genetic alterations in cancer, with likely repercussions on the adoption of therapies.

Representatives from EMBL have leadership roles in both the project's governance and scientific activities. Results from the analysis conducted by EMBL staff will be made available to the whole Pan-Cancer consortia. Results from analysis activities that have taken place on cloud resources within the Pan-Cancer consortium will help drive the analysis undertaken within this use case.



## Procurer sponsoring the use case: EMBL-EBI

EMBL-EBI shares and archives data from Life Science experiments and performs basic computational biology. research as well as training for academia and industry. EMBL-EBI services process 27 million requests on average per day and store over 100PB of data.

EMBL-EBI

**Contacts**

Tony Wildish, EMBL-EBI

wildish@ebi.ac.uk

European Commission

Helix Nebula - The Science Cloud (HNSciCloud) with Grant Agreement number 687614 is a Pre- Commercial Procurement Action funded by the EU Framework Programme for Research and Innovation Horizon 2020.