PhenoMeNal

Deliverable 6.4

Project ID	654241
Project Title	A comprehensive and standardised e-infrastructure for analysing medical metabolic phenotype data
Project Acronym	PhenoMeNal
Start Date of the Project	1st September 2015
Duration of the Project	36 Months
Work Package Number	6
Work Package Title	PhenoMeNal Virtual Research Community Gateway
Deliverable Title	D6.4 Participating Biobanks and repositories connected to the VRC
Delivery Date	M24
Work Package leader	EMBL-EBI
Contributing Partners	EMBL-EBI, BBMRI-ERIC, CRS4, CIRMMP
Authors	Ken Haug, Gianluigi Zanetti, Antonio Rosato

Abstract: PhenoMeNal VRE deployments can be used in both public and private data environments. Public data can be transferred seamlessly to and from a VRE. Biobanks are restricted by existing data governance and ELSI restrictions for each dataset, hence a local approach to integration is required.

PhenoMeNal

Contents

Executive Summary

The PhenoMeNal infrastructure has been designed from the bottom up as a powerful solution for analysing medical metabolic phenotype data. The infrastructure seamlessly harnesses the computational power needed to run complex analyses by adapting to multiple computational infrastructure scenarios, while maintaining a simple workflow based front-end.

PhenoMeNal has so far developed mechanisms to deploy the Virtual Research Environment (VRE) in OpenStack, Amazon and Google clouds. Additionally, the project has released containerised data transfer mechanisms for automated exchange between EMBL-EBI's MetaboLights[1] and PhenoMeNal. These data transfer mechanisms can be used, irrespective of which cloud provider chosen by the user to run the PhenoMeNal VRE. To maximise the value of the outputs of PhenoMeNal workflows, well annotated metadata accompanying the raw data is required. As part of this project, this metadata annotation activity is being done by curators working on the MetaboLights database to provide use cases for testing and validation of the workflows and to provide examples and documentation for users. This activity is also required for data from other resources, like Biobanks and Phenome Centres. The project has further developed the extensive metadata validation rules currently present in MetaboLights, ranging from semantic checks of the structure of the underlying ISA-Tab documents, through to new APIs and online validation of ontology and taxonomy terms used for annotations.

In this document, in addition to describing the open data transfer to EMBL-EBI MetaboLights, we cover applications that use the PhenoMeNal infrastructure for medical biobank relevant computing.

- General analysis work on biobank data.
- A specialised pipeline to do spectral analysis of plasma samples NMR spectra for quality control by a group of biobanks participating to the EXpert CEnter in METabolomics (EXCEMET), a reference BBMRI networked expert center.

Contribution towards project objectives

- **Task 6.5**: Implement PhenoMeNal Data submission guidelines
- **D6.4** Participating Biobanks and repositories connected to the VRC

[1] http://www.ebi.ac.uk/metabolights

Detailed report of the deliverable

PhenoMeNal consortia partners CIRMMP, through the da Vinci biobank in Tuscany (Italy) and ICL, through the MRC-NIHR National Phenome Centre in London[2] (UK), have extensive experience in working with biobanked data.

Data in biobanks are generally governed by existing legal compliance and ELSI protection requirements. Biobanks are therefore not likely at this point in time to be able to deposit data in public or external clouds. The compute-to-data approach however facilitates PhenoMeNal workflows to be run within local secured environments.

The analysis of data present in medical biobanks is characterised, before anything else, by three main requisites:

- privacy-related risks should be minimised;
- analysis should be standardised;
- no sophisticated IT training should be required.

The PhenoMeNal approach directly satisfies these requisites. In fact,

- the infrastructure can be directly deployed on stand-alone computational resources that can be completely insulated within private networks (*aka "Compute to data"*);
- it is designed to support reproducible, standardised data processing and analysis by providing pre-defined workflows, with automatic tracking of computational runs;
- it presents users with easy to use and ready-to-start pre-packaged workflow selections.

Moreover, the flexible deployment architecture of PhenoMeNal guarantees seamless transition, with no change in the user interface, from the use of private to public computational resources. These resources could become relevant even to medical biobanks once public clouds are considered secure enough (e.g. by using hardware secured containers on processors that support intel SGX[3] technology) to satisfy the related ELSI constraints.

[2] http://phenomecentre.org/
[3] https://software.intel.com/en-us/sgx

Metadata curation is an ongoing activity that is crucial for aligning the raw data for maximum utilisation of the respective workflows. Data from direct upload to the PhenoMeNal infrastructure, i.e. not originating from MetaboLights, can be transferred automatically to MetaboLights for final metadata curation and validation. In this scenario, primary research data and, if present, metadata is copied at the same time. PhenoMeNal users will only have to supply their MetaboLights API key to enable this integration. Data uploaded from a biobank can be transferred to MetaboLights using this same mechanism. Deliverable D9.2 describes in-depth the technical details about PhenoMeNal to MetaboLights direct data transfer[4,5].

BBMRI-ERIC - EXpert CEnter in METabolomics (EXCEMET)

The EXpert CEnter in METabolomics (EXCEMET) is a reference BBMRI-ERIC expert centre. EXCEMET participants share their sample collections in a distributed, networked virtual biobank on an European scale. One of the important issues of EXCEMET is to maintain consistent sample quality controls across participating biobanks.

An important example of quality control is the evaluation of changes in NMR spectra during the shelf life of plasma samples. After the raw NMR spectra have been acquired, they need to be processed by a specialised pipeline to extract sample specific signature information which can then be compared with equivalent data from the same sample at later times. Ideally, the procedure should be standardised both across time, in the same biobank, and across biobanks.

In principle, this task could be accomplished by sending data to a centralised computation facility. Unfortunately, since these are human related samples, this is not possible, and thus the task of running the computation has to be localised to each participating biobank. Note that this is not simply a question of installing a single, specific, hardwired, program, but rather a pipeline that will most likely evolve in time, introducing non-trivial IT complexities -- e.g. different samples could have been analysed using a different version of the quality control procedure -- that need to be managed by the biobank staff.

The PhenoMeNal approach and technology are particularly well suited to solve this class of problem and the consortium is helping EXCEMET in setting up the appropriate configuration for this specific quality control application. The actual deployment mechanism to the biobanks will leverage BBMRI-ERIC BibBox[6], a specialised toolbox for biobanks, to provide an always-ready browser-accessible starting point.

[4] http://portal.phenomenal-h2020.eu/app-library/scp-aspera
[5] http://portal.phenomenal-h2020.eu/app-library/mtbl-labs-uploader
[6] http://bibbox.org/welcome

The specific quality control process is divided in two macro steps (using plasma samples as an example)

Acquisition of the reference spectrum

1. A participating biobank receives a plasma sample.
2. An aliquot of the sample is used to measure the reference NMR spectra for the sample.
3. The reference spectrum is then cleaned, normalised and stored with all the relevant acquisition and processing information in a localised database.

Later time quality control

Two typical triggers of quality controls are: a user request for a sample aliquot; a random or programmed check on samples years after the sample entered the biobank.
In both cases, a further sample aliquot is taken and processed as follows.

1. The original data and metadata, e.g. actual workflow run and related computational history, saved at sample acquisition are retrieved.
2. A new NMR spectra is taken using the same NMR acquisition parameters.
3. The raw spectra data are processed using the saved workflow information to obtain aliquot-specific signature information.
4. A data comparison pipeline is run on the old and the new signature data.

Private OpenStack deployment at the Da Vinci European Biobank in Florence

The da Vinci European BioBank (daVEB)[7] is an infrastructure tightly linked to the activities of the Fiorgen Foundation[8], a non-profit organisation that promotes research in the field of pharmacogenomics and personalised medicine. The standard operating procedures (SOPs) concerning samples and data that have been developed at daVEB largely stem from the strong metabolomics connections of Fiorgen. Additionally, work involving scientific collaborators of the foundation in international and European projects, aimed to tackle the standardisation of pre-analytical procedures and the promotion of data standards in metabolomics[9], contributed strongly. daVEB is an ISO 9001-certified biobank, and collects samples from various hospitals and medical research institutes in Tuscany as well as from outside Italy, through collaborations and European projects.

[7] https://www.davincieuropeanbiobank.org/it
[8] http://www.fiorgen.net/
[9] doi:10.3390/jpm5020107

As part of its computational infrastructure, daVEB is using resources made specifically available by the CIRMMP consortium and the University of Florence. These resources are provided as an OpenStack cloud with the following technical specifics:

- OpenStack version Mitaka
- Service URL: https://cloud1.cerm.unifi.it
- 64 cores
- 128 GB RAM
- 1.3 TB storage space

To deploy the PhenoMeNal VRE, an instance of OpenStack running Ubuntu 16.0.4 was configured for the deployment of the service. The deployment was done by daVEB staff following online public project instructions[10]. This VRE is available to the staff of daVEB and Fiorgen foundation for research and quality control purposes and has proved to be highly beneficial in driving forward the development of PhenoMeNal.

Private OpenStack deployment at the MRC-NIHR National Phenome Centre

The MRC-NIHR National Phenome Centre[11] (NPC) led by ICL is one of the largest centres for metabolic phenotyping in Europe. NPC uses both NMR and MS analytical techniques in parallel, which gives a unique insight into the molecular data. Naturally Ethical Legal and Social Implications (ELSI) have a strong bearing on the use and possible re-use of the data produced by the NPC. This led to ICL being the first full implementation of our compute-to-data approach, which was instrumental in driving forward the development of PhenoMeNal in this area. ICL is now successfully running a private PhenoMeNal VRE using sensitive data without any need for external connections or access. This OpenStack deployed VRE is located behind ICL firewalls and is only accessible from within the ICL local network or via remote connection to the ICL VPN. Additionally, this VRE has a strict access policy as one can only connect to it via ssh-key or a strong password.

[10] https://github.com/phnmnl/cloud-deploy-kubenow/blob/master/README.md#get-phenomenal-kubenow
[11] http://phenomecentre.org/

Data annotation guidelines

Data used in the PhenoMeNal workflows will vary based on the respective tools that the workflow utilises. The PhenoMeNal data management plan and general guidelines specify that users should not publish data from the infrastructure directly, but rather use the connected MetaboLights repository. To satisfy the current metadata requirements in MetaboLights, submitters need to supply additional data before a study (experiment) can be published. This additional metadata cannot be extracted or mined from the raw or processed data used in PhenoMeNal workflows, so the infrastructure users will supply this retrospectively when data enters the MetaboLights curation system. Metadata minimum reporting standards were defined in the Metabolomics Standards Initiative[12] (MSI). The MSI working groups defined the Core Information for Metabolomics Reporting (CIMR) templates, which formed the basis for the templates currently used in MetaboLights. Data submitted/uploaded to PhenoMeNal can be submitted to MetaboLights using a set of established transfer protocols and mechanisms, described below.

MetaboLights data transfer mechanisms

The project has developed data transfer mechanisms to enable seamless integration of the MetaboLights database. Here we describe more details about the transfer mechanisms available.

MetaboLights data upload container

We developed a Python based script to upload data directly from the PhenoMeNal infrastructure. This script has been containerised and is preloaded into all PhenoMeNal VREs by default. To upload data the Docker container, *mtbl-labs-uploader,* requires the user's MetaboLights API key. (found in the user account pages of MetaboLights). Data from PhenoMeNal is then uploaded to the MetaboLights staging area for further processing, like submission as a complete study.

To Install this container locally/outside a PhenoMeNal VRE:

```
$ docker pull container-registry.phenomenal-h2020.eu/phnmnl/mtbl-labs-uploader
```

[12] http://www.metabolomics-msi.org

MetaboLights data download container

Both public and private studies can be copied from MetaboLights to PhenoMeNal using this container. The "*MTBLS Downloader"* is present in all PhenoMeNal VREs, under the Transfer category in the toolbar to the left of the Galaxy screen. Hence no installation is required.

To Install this container locally/outside a PhenoMeNal VRE:

```
$ docker pull container-registry.phenomenal-h2020.eu/phnmnl/scp-aspera
```

HTTP transfer

All data associated with a study in MetaboLights can be uploaded to a study data upload folder, using HTTP, FTP or Aspera. The traditional HTTP protocol is probably the simplest method for smaller datasets and once-off submissions. This is however fairly unreliable given the nature of HTTP. Larger transfer jobs will typically time out and data has to be resubmitted using more reliable protocols, like FTP and Aspera. HTTP transfers is only available from the web user interface of MetaboLights.

FTP transfer

FTP is a good alternative where HTTP falls short. FTP is typically used for individual, or larger, files, where the user can control what is being transferred when. However, FTP does not default support resumable, i.e. partial file, transfers.

To use FTP for upload, choose the "Request upload folder" from any private study:

```
user: <shared on request>
password: <shared on request>
server: ftp-private.ebi.ac.uk
remote folder: /prod/<MTBLS accession number>-<obfuscation code>
```

To use FTP for download, simply browse all public studies:

http://ftp.ebi.ac.uk/pub/databases/metabolights/studies/public/

Aspera transfer

The most advanced and fastest option offered by MetaboLights is Aspera[13]. Aspera offer free clients and has a raft of options that can help facilitate a very fast transfer speed. Over longer distances, i.e. where there are several networks in between the submitter and MetaboLights, this is by far the most reliable option. Aspera also supports resumable jobs, so that a submitter can pause the data transfer and resume the process at a later stage or a different geographical location. Aspera forms the basis for the MetaboLights up/download containers.

Upload

Direct Aspera container upload to MetaboLights will automatically create a project under the user's MetaboLights Workspace, aka. MetaboLights LABS

Aspera for MetaboLights study upload is available on PhenoMeNal Galaxy instances under "PhenoMeNal H2020 Tools", "Transfer".

To use Aspera outside a PhenoMeNal VRE for MetaboLights study upload, choose the "Request upload folder" from any private study. This will create a linked upload directory for a given study:

user: <*shared on request*>
password: <*shared on request*>
ascp -QT -L- -l 300M your_local_data_folder mtblight@ah01.ebi.ac.uk:/prod/<*MTBLS accession number*>-<obfuscation code>

To use Aspera outside a PhenoMeNal VRE for MetaboLights Workspace upload. A python based command line interface tool is available to enable uploads:

$ pip install git+https://github.com/EBI-Metabolights/MetaboLightsLabs-PythonCLI

[13] http://asperasoft.com/

Once this is installed, files can be uploaded to the workspace project using the following command:

```
$ uploadToMetaboLightsLabs.py -t < token > -i < filesToUpload > -p < projectId > -s < server >

User token: <available from MetaboLights website>
Files to Upload: <local files to upload>
Project Id (workspace project id): <available from workspace project interface>
Server: PROD
```

Download

Both public and private MetaboLights data can be downloaded using Aspera. Private data can only be accessed by the data owner, or when access has been by the owner. Public data can of course be downloaded without any restrictions, but private/embargoed data can only be accessed through data ownership or explicit permission granted.

Aspera for MetaboLights study download is available on PhenoMeNal Galaxy instances under “PhenoMeNal H2020 Tools”, “Transfer”. The same feature can be manually used from the command line using the following command:

```
$ docker run -v $PWD:/data -e "ASPERA_SCP_PASS=<password>" container-registry.phenomenal-h2020.eu/phnmnl/scp-aspera -QT -l 1g fasp-ml@fasp.ebi.ac.uk:/studies/public/<MTBLS accession number>/data
```

To manually download private data from the command line, use the following command:

```
$ export ASPERA_SCP_PASS=<secret password>
$ ascp -QT -l 1g mtblight@ah01.ebi.ac.uk:<folder-obfuscation-location> .
```

Delivery and schedule

The delivery is delayed: No

Conclusion

PhenoMeNal VRE deployments can be used in both public and private data environments. Data from EMBL-EBI's MetaboLights can be transferred seamlessly to and from a PhenoMeNal VRE, providing data dissemination agreements are in place for the VRE. Biobanks are restricted by existing data governance and ELSI restrictions for each dataset, hence a local approach to integration is required. Although most human clinical data is governed by ELSI restrictions, the ability to run the PhenoMeNal infrastructure behind institutional firewalls enables researchers to utilise the workflows and tools offered. Our approach to develop the infrastructure for both cloud and local deployment, ensures that a wider group of researchers will benefit from our reproducible workflows.