Deliverable 9.1

Project ID	654241
Project Title	A comprehensive and standardised e-infrastructure for analysing medical metabolic type data.
Project Acronym	PhenoMeNal
Start Date of the Project	1st September 2015
Duration of the Project	36 Months
Work Package Number	9
Work Package Title	WP9 Tools, Workflows, Audit and Data Management
Deliverable Title	D9.1 Report on existing software tools, workflows and analytical pipelines initially supported in the PhenoMeNal grid
Delivery Date	M6
Work Package leader	IPB
Contributing Partners	EMBL-EBI, IPB, UU, SIB, CEA, ICL, UB, INRA

Authors: Kristian Peters, Daniel Schober, Steffen Neumann, Reza Salek, Pablo Moreno, Rico Rueedi, Roger Mallol, Fabien Jourdan, Ibrahim Karaman, Tim Ebbels, Vitaly Selivanov, Pedro de Atauri, Etienne Thévenot

Contents

1. Executive Summary

The PhenoMeNal project will support several of the most common workflows in metabolomics, and will provide an infrastructure to host and run them in public or, for privacy sensitive patient data, private cloud environments.

This deliverable reports several Use Cases and the initially planned architecture to support running them on the PhenoMeNal e-Infrastructure. We will cover tools reported by a survey on workflow components initiated, among others, by PhenoMeNal partners. We also prepare to collect all requirements for the re-analysis of high-profile studies that will comprise the milestone "MS9.2 Re-analysis of several high-profile data analyses with the PhenoMeNal infrastructure (M24)".

2. Project Objectives

No.	Objective	Yes	No
1	Specify and integrate software pipelines and tools utilised in the PhenoMeNal e-Infrastructure into VMIs, adhering to data standards developed in WP8 and supporting the interoperability and federation middleware developed in WP5. We will develop new applications only to complete ‘missing links’ in pipelines. We will use public repositories and continuous integration to always provide development snapshots of the infrastructure VMIs.	X	
2	Develop methods to scale-up software pipelines for high-throughput analysis, supporting distributed execution on e.g. local clusters, private clouds, federated clouds, or GRIDs.	X	
3	Add quality control and quality assurance to pipelines to ensure high quality and reliable data, keep an audit trail of intermediate steps and results.	X	
4	Develop methods to present and summarize the results of the pipelines in biomedical and disease contexts.	X	

3. Detailed report on the deliverable

3.1. Background

In metabolomics, a multitude of tools is commonly used, which are implemented in multiple programming languages, ranging from Java, C++, to Python, R and Matlab. For metabolomics to achieve its full potential, the accessibility, reporting, reproducibility, and overall harmonisation of computational metabolomics tools must be improved significantly. Computational workflows provide one route to achieving such harmonisation. Galaxy is a widely used workflow platform that has helped to transform genomics research by massively increasing the accessibility to powerful data analysis tools[1]. It is intuitive to use and highly flexible allowing non-programmers to create analysis pipelines from a broad and expanding suite of tools[2].

3.2. Strategy for the collection of tools, workflows and analytical pipelines initially supported in the PhenoMeNal grid

We have collected a minimum list of tools to be used within PhenoMeNal. An initial list of required tools was gathered at the WP 6 Virtual User Community (VRC UX) workshop (https://docs.google.com/document/d/1ns9nXkid5qDORtCSprLZXluPTXbUCtSY14Y8mHEF-xg/edit#) along with an external collection of tools[3]. The list will be extended in future based on the WP8 and Elixir survey results and by our five primary use cases described below in order to add a data-driven bottom up approach. We have started to collate a table based on the following scheme (Figure 1). The work on the list of software tools is in progress and can be found: https://docs.google.com/spreadsheets/d/1Xagl8F6bnub2QUX24ymMzSZ-gWElXLBr-SVgngZHHnk/edit#gid=0 in our WP9 workspace Drive folder.

---

[1] Goecks, Jeremy, Anton Nekrutenko, and James Taylor. "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences." *Genome Biol* 11.8 (2010): R86.

[2] Giacomoni, Franck et al. "Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics." *Bioinformatics* 31.9 (2015): 1493-1495.

[3] http://omictools.com/data-preprocessing3-category

Technology	Functionality	Tool Name	Open Source	Virtualisation status	As required by
NMR	nmrML&FID2processedData conversion	nmRIO	yes	scheduled	Own tool involvement
NMR	assignment & quantification of metabolites	rNMR	yes	Add constructors for nmrML, Galaxy based docker	VRC Meeting
NMR	automated metabolite identification for mGWAS	MetaboMatching	Yes?	MatLab dependencies?	CoLaus Use Case
MS	Vendor to mzML conversion	msconvert	yes	Proteowizard VM / Docker image with msconvert pre-installed	Use cases
...	...	...	...	...	...

Figure 1: Open Source tools (here a snapshot) to be integrated into the PhenoMeNal infrastructure by virtualization. This list was collected from different sources as mentioned above.

3.3. Collection of Use Cases to define the required tools and workflows in PhenoMeNal

We have obtained a set of high-profile metabolomics studies from the PhenoMeNal consortium (see below) and the collaboration partners. These studies have either been analysed before, so we can reproduce them with the PhenoMeNal workflows, and demonstrate that we obtain the same outcome, or they are currently being obtained, and will be beta-testers of the PhenoMeNal workflows. Some of these also include phenome information and genomics data. Initially, we will support at least the tools required to analyse these use case data sets. Different workflows will be developed that are aimed at different user roles, which were collected and described at the VRC workshop and was reported in D6.1. Being able to perform the analyses for these Use Cases will allow us to reach the milestone "MS9.2 Re-analysis of several high-profile data analyses with the PhenoMeNal infrastructure (M24)".

The use cases were selected based on either involvement of project partners, typicality/representativeness of approaches and coverage of assay method across clinical metabolomics. This lead to the Use Case 1, covering MS and NMR, Use Case 2 covering NMR, Use Case 3 and 4 covering MS, and Use Case 5 covering Fluxomics:

Use Case 1: Analysis of the MESA research study on subclinical cardiovascular diseases

The Multi-Ethnic Study of Atherosclerosis (MESA, http://www.mesa-nhlbi.org/) is a medical research study involving more than 6,000 men and women in the United States. The study focuses on the characteristics of subclinical cardiovascular diseases. As part of COMBI-BIO project (Development of combinational biomarkers for subclinical atherosclerosis, http://www.combi-bio.eu/), metabolomics data were produced in two phases for 4,000 MESA participants from serum samples. There are 7 datasets available for each phase generated using ^{1}H-NMR (NOESY, CPMG, and J-resolved pulse programs) spectroscopy and UPLC-MS (HILIC +/- and LIPID +/-) mass spectrometry platforms.

Pre-processing of NOESY and CPMG NMR data involved the following steps:

a) Phasing and baseline correction of the spectra (using Bruker TopSpin^{TM}). Due to the proprietary software used in step 1, the PhenoMeNal workflow will handle the data processing steps from 2 onwards. However, other workflow components will be available later that can effect similar processing to that currently performed with TopSpin (see section 3.5 NMR processing tools).
b) Data set were calibrated to the glucose doublet peaks at δ 5.23 ppm (in-house-written Matlab functions)
c) Spectral peak alignment was done using Recursive Segment-wise Peak Alignment method (in-house-written Matlab functions)
d) Removing interfering spectral regions, such as those arising from residual water signals or contaminants (in-house-written Matlab functions)
e) Normalization of each spectrum was carried out using Probabilistic Quotient Normalization method (in-house-written Matlab functions)
f) PCA with Hotelling's T^2 distribution test was used for removing the outlier samples (in-house-written Matlab functions)
g) Spectral binning by Statistical Recoupling of Variables method[4] (in-house-written Matlab functions)

Pre-processing of UPLC-MS data involves the following steps:

a) Data sets were converted from vendor format into open XML based mzML file format.

---

[4] Blaise, Benjamin J et al. "Statistical recoupling prior to significance testing in nuclear magnetic resonance based metabonomics." *Analytical chemistry* 81.15 (2009): 6242-6251.

b) Data trimming, peak detection, chromatographic deconvolution and retention time alignment was performed using XCMS (R)
c) Features not present in any of the QC samples were removed (in-house-written Matlab functions)
d) Removing features that did not exhibit robust linearity with respect to standard dilution series of QC samples (in-house-written Matlab functions)
e) Normalization of each spectrum by Median-Fold-Change Normalization (in-house-written Matlab functions)
f) Correction of instrument drift across runs that influences the features by computing a function using locally-weighted polynomial regression based on the features of QC (in-house-written Matlab functions)
g) Removing features that have coefficient of variation higher than a defined threshold in the QC samples (in-house-written Matlab functions)
h) Log transformation of the features (in-house-written Matlab functions)
i) Where in-house MATLAB functions were used to process data, we will translate these steps into another language (R based for example) to allow more ease of use within the Galaxy workflow environment on Phenomenal compute architecture. We will demonstrate that the PhenoMeNal workflows are able to achieve the same processed data output as the original routines did.

Use Case 2: Analysis of the CoLaus population-based study on cardiovascular diseases and risk factors

CoLaus (Cohorte Lausannoise) is a monocentric, longitudinal and population-based study of 6,733 participants, focusing on the epidemiology and genetic determinants of cardiovascular diseases and risk factors. The participants were genotyped (Affymetrix GeneChip Human Mapping 500k array; HapMap and 1000 genome panel imputations), and extensively phenotyped. A subset of 983 individuals underwent metabolomics analysis. The metabolomics data set includes two types of NMR assays (NOESY + CPMG) for serum samples and one (NOESY) for urine samples.

CoLaus metabolome data will be processed and quantified using PhenoMeNal workflows to produce metabolome phenotypes for genome-wide association studies.

Use Case 3: Data processing for the Uppsala Fibromyalgia Use Case

At the Uppsala clinic, we are going to analyse metabolite-profiling data from a cohort of 120 participants, consisting of fibromyalgia patients and several matched controls. The samples are cerebrospinal fluid (CSF) measured in positive and negative ionisation mode in duplicates on Thermo Orbitrap LC-MS instruments. We are going to use a workflow including several of the OpenMS tools for the data analysis in PhenoMeNal.

Use Case 4: Data processing, statistical analysis, and annotation of the "Physiological Variations of the Urine Metabolome" Use Case

Characterization of the physiological variations of the metabolome in biofluids is critical for biomarker discovery, to avoid confounding effects in cohort studies. In this study, conducted by the CEA platform from the MetaboHUB French Infrastructure for Metabolomics, urine samples from 183 adults were analyzed by liquid chromatography coupled to high-resolution mass spectrometry (LC-HRMS). After pre-processing of the raw files, a total of 258 metabolites were identified at confidence levels provided by the metabolomics standards initiative[5],[6] (MSI) levels 1 (directly identified via reference standards) or 2 (identified by similarity using in-house and public databases). Physiological variations of these metabolites with age, body mass index, and gender, were further analyzed by using univariate and multivariate statistics[7]. Raw files are publicly available on the MetaboLights repository (accession number: MTBLS20) and the complete statistical workflow is publicly available on the Workflow4Metabolomics computational infrastructure (reference W4M00001).

Use Case 5: Data processing for fluxomic analyses

Determination of metabolic flux distributions is fundamental in order to have a complete characterization of metabolic phenotypes. In this analysis, cells are incubated in the presence of isotope-enriched substrates to describe its biochemical processing through the main metabolic pathways. This type of analyses is not normally[8] applied in-vivo, but it can be done in primary cultures isolated from patients. The use of cells in primary culture from patients is a promising tool looking to evaluate the differential adaptation of control and patient cells to drugs; e.g. drugs against cancer. Currently there are no cohort studies covering patient samples. However, files covering the analysis of cultured

---

[5] Sumner, Lloyd W et al. "Proposed minimum reporting standards for chemical analysis." *Metabolomics* 3.3 (2007): 211-221.

[6] Fiehn, Oliver et al. "The metabolomics standards initiative (MSI)." Metabolomics 3.3 (2007): 175-178.

[7] Thévenot, Etienne A et al. "Analysis of the human adult urinary metabolome variations with age, body mass index, and gender by implementing a comprehensive workflow for univariate and OPLS statistical analyses." *Journal of proteome research* 14.8 (2015): 3322-3335.

[8] There are examples in the literature where this type of measurement is done in-vivo, ie. Blood sampling before and after it goes through a particular organ.

cells associated with cancer and other diseases are available on the MetaboLights repository. We expect that these will be publically available during the course of the project, and as such will be used as examples for our analyses.

3.4. Galaxy workflow survey as source of tools used by the metabolomics user community

From November 2015 to February 2016, the consortium of UK universities, including University of Birmingham, MRC Human Nutrition Research Unit, the Francis Crick Institute, Imperial College London, University of Manchester and the European Bioinformatics Institute, performed a survey. This work, in partnership with ELIXIR-UK, obtained input from the international community to identify which additional metabolomics data processing and analysis tools should be prioritised for incorporation into Galaxy workflows. This includes tools associated with the full breadth of analytical technologies in the metabolomics toolbox. The survey was sent to mailing lists and individuals who considers themselves to be actively engaged in metabolomics science, from any scientific backgrounds, any career stage, and importantly, from any country or type of employment (academia, industry, government or any other groups). We were interested in both those who have recently started metabolomics research as well as those that might consider themselves expert users. By “metabolomics science” we include the underpinning science and technology (experimental design, analytical measurements, computational analyses and informatics) and any field of application, from microbes to plants and animals, including humans.

The survey has been closed, but is available for reference from: https://www.surveymonkey.co.uk/r/metabolomics-galaxy. Preliminary results indicate that approximately half of all metabolomics practitioners do not have access to dedicated bioinformatics support, and stated that both “Data processing and statistical analyses” and “Data curation” are the slowest steps in the entire metabolomics workflow, underlining the pressing need for user-friendly and powerful metabolomics infrastructures. The full result of this survey will be distributed widely to the metabolomics community, for example through the *Metabolomics* journal and on-line MetaboNews as well as helping to guide development of the tools within PhenoMeNal.

3.5. Tools and software supported

We have collated a detailed list of tools and knowledge representation formats required for PhenoMeNal, for each tool capturing its domain, functionality, virtualisation status and data standards requirements. In this section we briefly describe the main tools and

why they required for PhenoMeNal. The detailed tool list will be expanded and can be accessed under

https://docs.google.com/spreadsheets/d/1Xagl8F6bnub2QUX24ymMzSZ-gWElXLBr-SVgngZHHnk/edit#gid=0

MS Data processing

Support conversion of vendor file formats to open data formats

The conversion of MS data to open formats is necessary because we use vendor agnostic or open source data analysis tools. Usually, the vendors provide software libraries to access their own formats. The downside is that these often have rather complex application programming interfaces (APIs), and worse, each vendor has their own proprietary API. Currently, most of these interfaces require Windows dynamic link libraries (DLLs) for the actual file access, which are not compatible with other operating systems such as MacOSX or Linux. One of the two main routes to mzML-formatted data is using Open Source converters such as the **msConvert tool** developed by the Proteowizard team[9], which is one of the reference implementations for mzML. It can convert to mzML from Sciex, Bruker, Thermo, Agilent, Shimadzu, Waters and also the earlier file formats like mzData or mzXML and is consequently widely used.

In PhenoMeNal we are going to provide a data conversion service that can be called from the workflows in case the data is not yet available in open formats. To that end, we will create a Proteowizard VM / Docker image with **msconvert** pre-installed, as running **msconvert** under Linux/Wine requires a complex setup. This is risky because it involves multiple software components not under our control, and creating this service might fail. In that case, it will remain the responsibility of the user to upload data already converted locally on a Windows computer.

Support R/Bioconductor and Galaxy/W4M based tools on the PhenoMeNal e-Infrastructure

The partner CEA — together with INRA and CNRS within the MetaboHUB and IFB French infrastructures for metabolomics and bioinformatics, respectively — has developed the **Workflows4Metabolomics.org** computational infrastructure (W4M). It builds upon the Galaxy environment, with more than 30 additional tools. It includes an LC/MS data processing workflow that employs the R packages **XCMS** and **CAMERA** for processing of LC/MS data, but also unique modules (e.g., ***ropls*** and ***biosigner***

9 http://proteowizard.sourceforge.net/

R/Bioconductor packages and corresponding Galaxy tools for multivariate analysis and feature selection, respectively**)**. In PhenoMeNal we are going to provide docker images for the Galaxy server and for each of the tools. This will facilitate maintenance, extension and scale-up of the LC/MS workflow to large data sets. This has been prototyped by running the tools in a workflow not locally on the Galaxy server, but instead offloading that computation (each tool) to Kubernetes cluster on the EMBL-EBI EMBASSY cloud, permiting parallel remote execution of the tools. We aim at virtualizing / dockerifying each of the tools in order to create a uniform PhenoMeNal infrastructure, and to allow easy deployment within the PhenoMeNal and partner infrastructures.

We are going to "adopt" the metabolomics Docker image[10] in **Bioconductor**. With this adoption, PhenoMeNal will continuously use up-to-date tools in the Workflows. It can also be expected that the Bioconductor docker images will be maintained in the future, helping with the sustainability of the PhenoMeNal workflows.

Support OpenMS on the PhenoMeNal e-Infrastructure

OpenMS is one of the prominent Open Source Platforms for computational mass spectrometry, mainly in Proteomics, but increasingly also in Metabolomics. OpenMS consists of C++ libraries, many command line tools, and graphical interfaces. In PhenoMeNal, we will provide workflows for MS data processing[11] with Galaxy, and collaborate with the OpenMS team to integrate OpenMS Docker Images[12] compatible with the PhenoMeNal e-Infrastructure. The functionality will be complete enough to partially execute examples from the OpenMS Tutorial (ftp://ftp.mi.fu-berlin.de/pub/OpenMS/release-documentation/OpenMS_tutorial.pdf) in the PhenoMeNal cloud. The OpenMS support in PhenoMeNal will be released as part of D9.2.1 in M12.

Support of MetFrag on the PhenoMeNal e-Infrastructure

Running some of the large-scale data processing can easily overwhelm public resources if many requests are executed in parallel. To avoid that the workflows can bring down an important public server, we will provide simple options to set up mirrors

---

10 https://www.bioconductor.org/help/docker/

11 https://github.com/bgruening/docker-recipes/tree/master/galaxy-openms

12 https://github.com/hroest/docker_files/tree/master/openms

inside the PhenoMeNal e-Infrastructure.[13] As part of this effort, we have created a **MetFrag** docker image, which can be embedded into the Galaxy workflows.

NMR Data processing

Similar to MS we will convert the NMR data to open formats to be vendor agnostic, improve accessibility and readability by open source data analysis tools. Ideally, we plan to use nmrML format, a vendor-neutral open exchange and storage format used to describe NMR-based metabolomics data that was developed by The **COSMOS** COordination Of Standards In MetabOlomicS (http://www.cosmos-fp7.eu/). An alternative option would be the data exchange format for NMR data by the Joint Committee on Atomic and Molecular Physical Data (**JCAMP-DX**), currently widely used by most open source NMR data analysis software.

The already mentioned **W4M** consortium (CEA and external partners in France) has developed a **W4M Galaxy NMR workflow** (http://workflow4metabolomics.org/the-nmr-workflow packaged for NMR data processing and analysis. For the first instant we will use this workflow and contribute to its development by adding extra packages and tools as required by our use cases, i.e. ensuring coverage for commonly used data processing and NMR metabolite identification tasks. One option would be Integration of the **rNMR** (http://rnmr.nmrfam.wisc.edu/) for NMR data processing, analysis and identification of metabolites into a Galaxy based docker. Some Issues that we foresee to overcome are:

a) rNMR only reads processed data, therefore we need tools to process raw NMR files (FIDs), we can use existing tools currently implemented and in use by W4M NMR workflow for processing NMR raw files and then use rNMR for subsequent steps.
b) Explore integration of nmrML format within the rNMR pipeline. One approach would be integration of nmRIO, which is a prototype R package for input and output of NMR data in the nmrML format, available from the nmrML github code repository.

Integration of **Bayesian Automated Metabolite Analyser for NMR spectra (BATMAN** http://batman.r-forge.r-project.org/) into workflow pipeline of the W4M Galaxy.

---

[13] Merlet, Benjamin et al. "Computational solution to automatically map metabolite libraries in the context of genome scale metabolic networks." *Frontiers in Molecular Biosciences* 3 (2016): 2.

In PhenoMeNal we are going to provide docker images of the Galaxy server, and to maintain, extend and scale the BATMAN to handle data sets. PhenoMeNal partner, Imperial College London, developed BATMAN and our close collaboration should facilitate integration Docker Images compatible within the PhenoMeNal e-Infrastructure. There are other tools, some based on Matlab or Python, used for NMR processing and analysis e.g. **Dolphin -** a tool for automatic targeted metabolite profiling using 1D and 2D 1H-NMR data (Matlab and R in development). Dolphin can be used to automatically quantify a set of target metabolites in multiple sample measurements using an approach based on 1D and 2D NMR techniques to overcome the inherent limitations of 1D (1)H-NMR spectra in metabolomics. Dolphin takes advantage of the 2D J-resolved NMR spectroscopy signal dispersion to avoid inconsistencies in signal position detection, enhancing the reliability and confidence in metabolite matching.

Genomics data integration tools

Metabomatching is a Matlab based tool for automated metabolite identification in metabolome-wide genome-wide association studies (mGWAS) that use untargeted NMR metabolome data as phenotypes. Metabomatching uses the association statistics (P-value, beta coefficient and standard-error) between a genetic variant, or SNP (single nucleotide polymorphism), and all metabolome features comprising the untargeted metabolome to find the best candidate metabolite underlying an observed SNP-feature association, among a provided set of experimental NMR spectra (currently from HMDB and BMRB). For each SNP, metabomatching outputs a graphical tool that includes a list of metabolite candidates to aid the user in identifying the metabolite(s) involved in the given association(s). For PhenoMeNal, the tool will be ported into R or python to avoid reliance on commercially licensed software.

NMR tools above could also integrate or be part of a metabomatching pipeline within PhenoMeNal for metabolite identification of genome-wide association studies, more details see below.

MetExplore is a tool allowing mapping metabolomics data (and other ‘omics data) in the context of genome scale network. This can be achieved interactively via the web server (www.metexplore.fr) or via web service interaction as described in the article[14] published in 2016 where PhenoMeNal is acknowledged.

---

[14] Merlet, Benjamin et al. "Computational solution to automatically map metabolite libraries in the context of genome scale metabolic networks." *Frontiers in Molecular Biosciences* 3 (2016): 2.

In the framework of PhenoMeNal, a JavaScript package for visualization of metabolic networks had been implemented and will be available for integration in any browser based analysis pipeline of metabolomics data.
In particular, in the framework of PhenoMeNal project, some efforts were done on development of visualization. A staff exchange took place at the EBI on early February, where MetExplore developers worked with MetaboLights software engineers in order to incorporate MetExplore visualisation module (MetExploreViz) in MetaboLights.

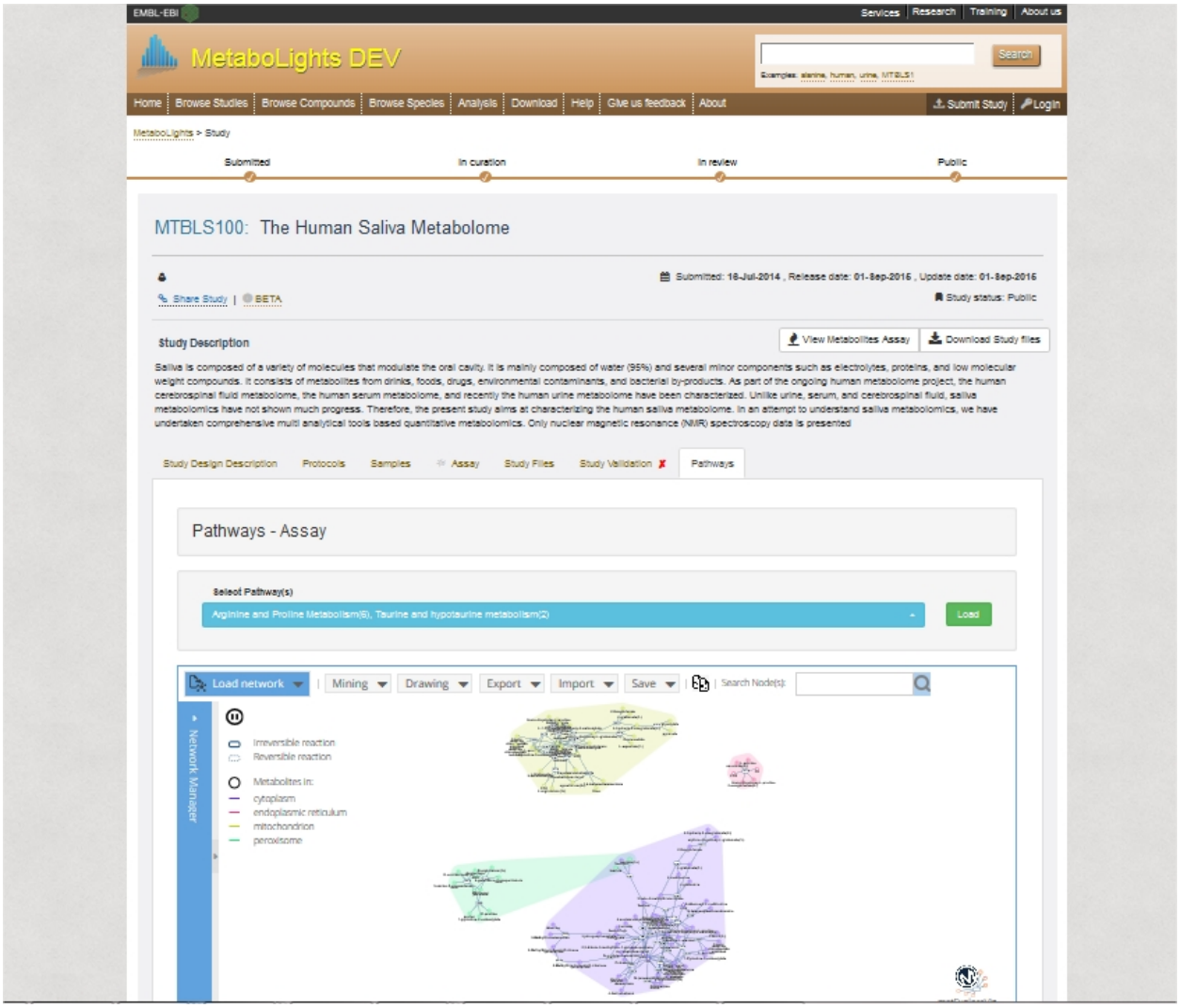

Figure 2: MetExploreViz package used for integrated visualisation in Metabolights.

Integration of the initially supported tools into the PhenoMeNal build and deployment infrastructure

Using the above approaches, we collected a list of distributed tools for phenomics, metabolomics and bioinformatics processing pipelines and workflows suitable for packaging into

docker images. We have also started to use public repositories and continuous integration to always provide development snapshots of the images. In order to make optimal use of the underlying architecture and to facilitate the interoperability, we are closely working together with WP5 and WP8. A primary goal is to hide the complexity of the underlying infrastructure to the actual user (e.g. biologists, clinicians), while giving easy to understand technical instructions to bioinformaticians for installing the supplied PhenoMeNal images in a short time while preserving data privacy and security. We have already begun to produce development-snapshots of the VMIs. They are available in the PhenoMeNal Continuous Integration Service Jenkins (see Figure 3), which has been set up as part of the deliverable D5.1. At the moment we are using the technology Docker to build container images (which are a better option than VMIs). All built images can be retrieved for execution from our Docker registry http://docker-registry.phenomenal-h2020.eu/ (machine-readable by a Docker installation, not meant for 'humans').

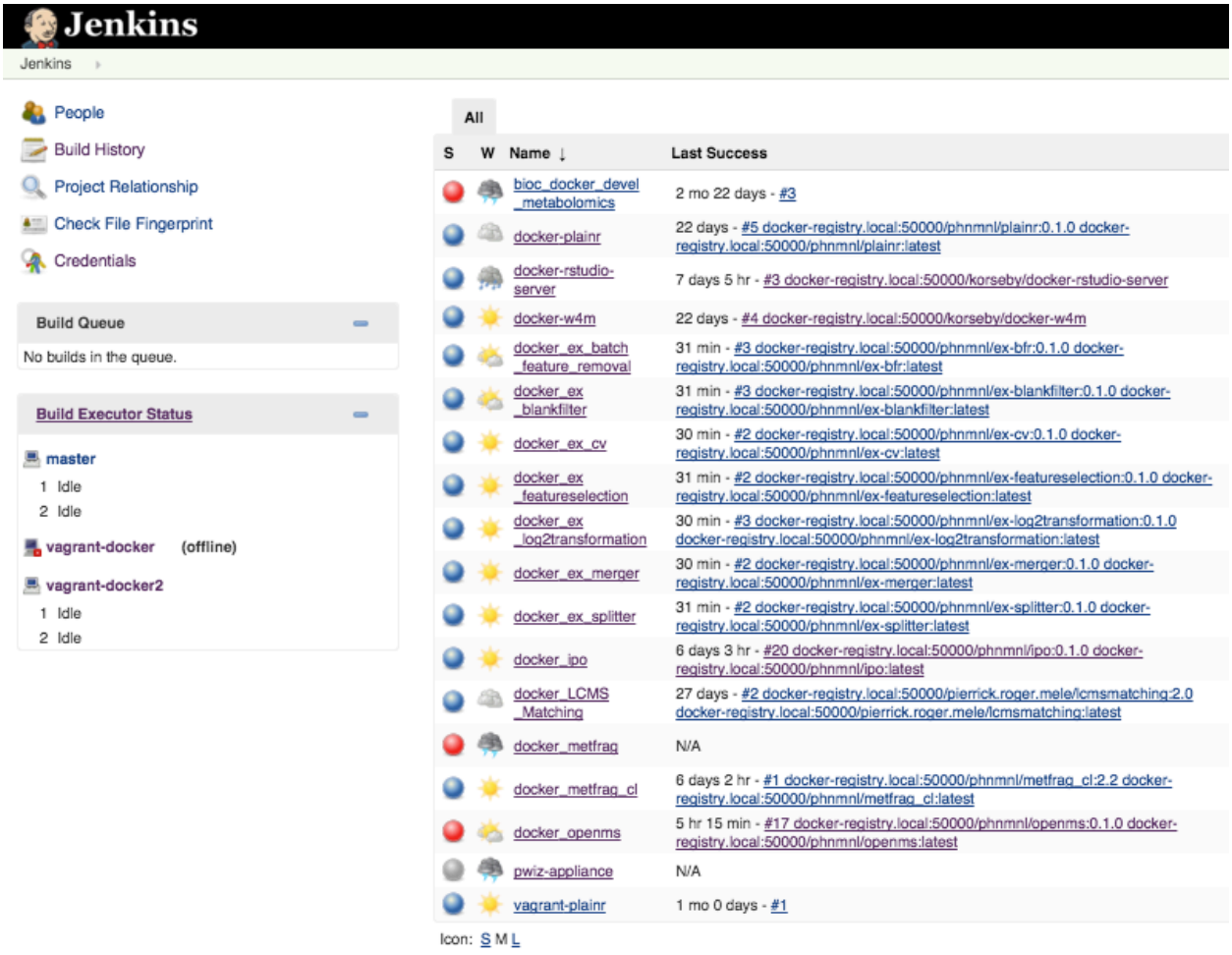

Figure 3: A first set of docker images is already available in the Continuous Integration Service Jenkins at http://phenomenal-h2020.eu/jenkins/

Tools for Fluxomic analysis

A total of three tools associated with the analysis of metabolic flux distributions developed by the UB partner.

Label2Flux (provisional name): a Python-tool for analysis of flux distribution based on experimentally measured distributions of labelled isotopomers.

The reliability of hypotheses regarding intracellular reaction fluxes can be evaluated by comparing measured and predicted label (e.g. ^{13}C) positional and mass isotopomer distributions. Computational estimation of metabolic fluxes is based on measured labelling patterns in intracellular metabolites resulting from metabolized labelled substrates, together with measured cellular uptake and secretion rates and prior knowledge of the biochemical reaction network (reviewed in Buescher, et al., Curr Opin Biotechnol, 2015. 34C:189-201; Niedenfuhr et al., Curr Opin Biotechnol, 2015. 34C:82-90.). In our analysis, two computer routines are developed and applied iteratively until a flux distribution is found that is compatible with the measured labelling patterns. In summary, taking the known reaction network and the measured uptake and secretion rates and assuming steady state, a first computer routine estimates by linear programming the range of possible values for each reaction flux, where each flux is maximised or minimised while leaving all other fluxes free (Mahadevan et al., Metab Eng, 2003. 5:264-76; Llaneras and Pico, J Theor Biol, 2007. 246:290-308; Orman et al., Biotechnol Bioeng., 2010. 107:825-35). A second routine solves a nonlinear problem to predict positional or mass isotopomer abundances by solving a system of balance equations around isotopomers, which take into account label transitions and the refined solution for fluxes obtained by applying the first routine. The two routines are repeatedly applied to estimate flux distributions and to predict the associated label distribution in experiments using labelled substrates. For each flux distribution, the enrichment in labelled products can be predicted and then compared with the measured enrichments, where comparisons are made at the positional or mass isotopomer level.

An original program based on Mathematica was translated in a version based on Python, which satisfies the open source requirement of the Project PhenoMeNal. Additional improvements are being introduced (EMU, optimization tools) and the program is being adapted to additional project requirements as scalability and connectivity with other programs and databases. The program has been designed to be integrated with databases (especially with MetaboLights) and with our other tools Midcor and Mitodyn.

Midcor: an R-tool for primary ^{13}C data correction for natural isotope enrichment. A correction of ^{13}C isotopomer data for natural isotope enrichment is used as a routine procedure needed to initiate the ^{13}C data analysis. We developed an R-program that performs such a correction for a corresponding workflow of ^{13}C data analysis in PhenoMeNal. The program reads raw gas chromatography / mass spectrometry m/z data, evaluates isotopomer distribution of specific substances, and corrects it taking into account natural distribution of isotopes, such as ^{13}C, ^{34}S, ^{29}Si, ^{30}Si. ^{13}C isotopomer data corrected by Midcor will be analyzable with Label2Flux.

Mitodyn: is a tool for investigation of possible multistationary / bifurcation behavior of mitochondrial respiration under specific metabolic conditions. According to our findings (Selivanov et al., PLoS Comput Biol., 2012. 8:e1002700; Selivanov et al., PLoS Comput Biol., 2011. 7:e1001115; Selivanov et al., PLoS Comput Biol., 2009. 5:e1000619), some specific metabolic conditions can provoke a switch of mitochondrial metabolism from a normal state of effective ATP synthesis into a state characterized by ineffective ATP synthesis and high rate of reactive oxygen species (ROS) production. Our program simulates mitochondrial metabolism and evaluates the possible factors that can switch it into the states of high ROS production rate.

Among the factors determining ROS production are the flux distributions affecting mitochondrial metabolism. Accordingly, the flux distribution resulting from the application of Labe2Flux will be one of the inputs for Mitodyn.

Midcor and Mitodyn are to be adapted to the requirements of PhenoMeNal, as with Label2Flux.

4. Delivery and Schedule

The delivery is delayed: No

5. Background information

<table>
<tr><td colspan="4">Patient and research subject data is very sensitive, and it is paramount to establish a robust governance framework for overall information management including sensitive data. The PhenoMeNal e-infrastructure will be able to cope with data generated from comprehensive clinical, genotypic, ‘omics and analytic sources including medical records, electronic health records, clinical measurements, genotypic data, phenotypic data from tissue and biofluid analysis, image and pathology data. Primarily, all data collected and held within the project will comply with all local laws, regulations and ethics. All personal information will be processed in accordance with accepted Data Protection Principles outlined above. Responsibility for data will be with the host institution/data provider.</td></tr>
<tr><td>Work package number</td><td>WP9</td><td>Start date or starting event:</td><td>M1</td></tr>
<tr><td>Work package title</td><td colspan="3">Tools, Workflows, Audit and Data Management</td></tr>
<tr><td>Participants</td><td colspan="3">IPB, EMBL-EBI, ICL, UB, UoB, CIRMMP, UL, UOXF, SIB, UU, BBMRI, CEA, INRA</td></tr>
<tr><td colspan="4">Our goal is to develop and maintain the primary scientific- and technological tools and corresponding interfaces. We will support the data standards defined by WP8 and facilitate</td></tr>
</table>

the interoperability of tools both within this consortium and those externally developed by the community.

We will establish distributed tools for phenomics, metabolomics and bioinformatics processing pipelines and workflows, including longitudinal primary research data management (continuous availability to avoid data lock-in) and data audit mechanisms, as well as quality assurance schemes. Thus, this work package will produce several tailored VMIs, which will be the basis for the service activities in WP5.

Objective 9.1 Specify and integrate software pipelines and tools utilised in the PhenoMeNal e-Infrastructure into VMIs, adhering to data standards developed in WP8 and supporting the interoperability and federation middleware developed in WP5. Most tools will be already available (see table 1.3.5.1) and we will develop new applications to complete 'missing links' in pipelines. Although two explicit releases for VMIs are listed as deliverables below, we will use public repositories and continuous integration to always provide development snapshots of the infrastructure VMIs.

Objective 9.2 Develop methods to scale-up software pipelines for high-throughput analysis, supporting distributed execution on e.g. local clusters, private clouds, federated clouds, or GRIDs.

Objective 9.3 Add quality control and quality assurance to pipelines to ensure high quality and reliable data, keep an audit trail of intermediate steps and results.

Objective 9.4 Develop methods to present and summarize the results of the pipelines in biomedical and disease contexts.

Task 9.1: Data processing pipelines (IPB, ICL, UL, UB, UOXF, UALB, CEA)

We will develop pipelines to process the data from analytical instruments (e.g. MS, NMR). It is of great importance to define sustainable workflows for processing omics data in a reliable, robust and reproducible way. This task will ensure that only high quality and reliable data is used for the subsequent data analysis pipelines, enabling reproducibility with a clean outcome. This includes e.g. normalisation steps, methods used for feature selection, calibration curves, QC, QA etc. This task also includes capturing important parameters used commonly in established data processing pipelines and workflows. Software will be integrated and/or developed for both the PhenoMeNal-Preprocess and PhenoMeNal-Data virtual machines.

Task 9.2: Data analysis pipelines (UALB, ICL, UL, IPB, ISB, UOXF, CEA, INRA)

This task will implement the analysis pipelines following the initial data processing using statistical analysis and further data annotation (aligned with Task 8.4 in WP8), and will cover commonly accepted use-cases. For both metabolite profiles and genomics data, we will explore existing supervised, unsupervised, regression and other available and commonly used data analysis techniques. For annotation or metabolite identification we will utilise reference databases and computational analysis methods. In addition, we capture the categorical / continuous parameters mostly used for analysis of different metabolomics

datasets and integrative cross -omics modeling. Existing and new pipelines and workflows will be integrated into the PhenoMeNal-Data, PhenoMeNal-Services and PhenoMeNal-Compute virtual machines.

Task 9.3: Data integration and Management (ISB, UB, EMBL-EBI, SRI, UB, ICL)

The identified data processing and data analysis pipelines from task 9.1 and 9.2 will form the basis for the data integration steps. Current bottlenecks in metabolomics and genomics data integration will help guide our efforts. We will explore various established GWAS and MWAS methodologies, correlating genomic data to metabolomic, pathway based output, interactome based analysis, enrichment analysis etc that will make the data integration pipeline. Information visualization will be used as a key paradigm to facilitate this data integration. Finally, we need to link the results to public resources like pathway databases and genetic disease databases.

Task 9.4: Data Integrity, Audit Trail and Quality Control (UL, EMBL-EBI, ICL, UOXF, UL)

We need to implement a generic and general audit trail framework to make sure all requirements have been carried out for data processing and data analysis steps. To support the audit trail requirements, all data formats will support checksums of the data leading to a particular result. The checksums and audit trail capturing will be required for later use of the developed infrastructure in regulated environments. Audit management is imperative for matrices used and code parameters and versions. The audit will be application based and dependent on the analysis platform. We will adapt and extend auditing functionality from data analysis frameworks such as Galaxy or R, where existing capabilities can provide capturing and logging of data processing and analysis steps. Also, primary research and meta-data that has been made publicly available will be subject to change management. Data integrity control is implemented via validation software which check data in open standards to check if it is error free and sufficiently complete, e.g. in terms of BioSharing/MIBBI minimal information standards for a particular domain.

Task 9.5: Provide continuous development snapshots with life-cycle management for building, testing, and deploying code, tools, and workflows (UU, IPB, EMBL-EBI)

In Task T5.5 we will set up and maintain a continuous integration instance (using e.g. Jenkins, http://jenkins-ci.org/) for building and testing tools and VMIs on a shared PhenoMeNal build system. This task T9.5 will populate the build system with all artefacts (operating system, tools and workflows) required to build regular development snapshots and create complex test scenarios (e.g. re-analysis of high-profile data analyses) for the pipelines to ensure the results are consistent after updates to individual components, and check for graceful degradation in case of corrupt or incomplete input. The system will then be used as a trigger to update VMIs with the latest version of tools, and also to facilitate automated and continuous testing of developed workflows; e.g. assessing that results from workflows are consistent after updates to individual components. The Jenkins instance will also store artefacts of previous versions to allow for rolling back to earlier versions in events of errors introduced, or allow for reproducing results on earlier versions of workflows and tools.

Task 9.6: Migrate existing prototype software from proprietary environments to open, distributable software of high performance and scalability. (ICL, IPB, UL, UB, UOXF)

Initially this will involve surveying the tools to be migrated and examining their current robustness and scalability, possibly providing extra documentation and worked examples. These tools will then be ported from the proprietary environments (e.g. MATLAB) to open source, distributed software systems. If necessary, they will be redesigned to work with large scale data. Finally they will be integrated with the processing and analysis pipelines of tasks 9.1 & 9.2 and validated for use on grid / distributed compute environments by remote users.