OXFORD (GIGA)ⁿSCIENCE

TECHNICAL NOTE

PhenoMeNal: processing and analysis of metabolomics data in the cloud

Kristian Peters [1,*], James Bradbury [2,*], Sven Bergmann [3,4], Marco Capuccini [5,6], Marta Cascante [7], Pedro de Atauri [7], Timothy M. D. Ebbels [9], Carles Foguet [7], Robert Glen [9,10], Alejandra Gonzalez-Beltran [11], Ulrich L. Günther [12], Evangelos Handakas [9], Thomas Hankemeier [14], Kenneth Haug [15], Stephanie Herman [6,16], Petr Holub [17], Massimiliano Izzo [11], Daniel Jacob [18], David Johnson [11,19], Fabien Jourdan [20], Namrata Kale [15], Ibrahim Karaman [21], Bita Khalili [3,4], Payam Emami Khonsari [16], Kim Kultima [16], Samuel Lampa [6], Anders Larsson [6,22], Christian Ludwig [23], Pablo Moreno [15], Steffen Neumann [1,24], Jon Ander Novella [6,22], Claire O'Donovan [15], Jake T.M. Pearce [9], Alina Peluso [9], Marco Enrico Piras [25], Luca Pireddu [25], Michelle A.C. Reed [12], Philippe Rocca-Serra [11], Pierrick Roger [26], Antonio Rosato [27], Rico Rueedi [3,4], Christoph Ruttkies [1], Noureddin Sadawi [8,9], Reza M. Salek [15], Susanna-Assunta Sansone [11], Vitaly Selivanov [7], Ola Spjuth [6], Daniel Schober [1], Etienne A. Thévenot [26], Mattia Tomasoni [3,4], Merlijn van Rijswijk [13,28], Michael van Vliet [14], Mark R. Viant [2,29], Ralf J. M. Weber [2,29], Gianluigi Zanetti [25] and Christoph Steinbeck [30,*]

[1]Leibniz Institute of Plant Biochemistry, Stress and Developmental Biology, Weinberg 3, 06120 Halle (Saale), Germany, [2]School of Biosciences, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom, [3]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland, [4]Swiss Institute of Bioinformatics, Lausanne, Switzerland, [5]Division of Scientific Computing, Department of Information Technology, Uppsala University, Sweden, [6]Department of Pharmaceutical Biosciences, Uppsala University, Box 591, 751 24 Uppsala, Sweden, [7]Department of Biochemistry and Molecular Biomedicine,

Universitat de Barcelona; Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBEREHD), Instituto de Salud Carlos III (ISCIII), Spain, [8]Department of Computer Science, College of Engineering, Design and Physical Sciences, Brunel University, London, UK, [9]Department of Surgery & Cancer, Imperial College London, South Kensington, London, SW7 2AZ, United Kingdom, [10]Centre for Molecular Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, CB21EW, United Kingdom, [11]Oxford e-Research Centre, Department of Engineering Science, University of Oxford, 7 Keble Road, OX1 3QG, Oxford, United Kingdom., [12]Institute of Cancer and Genomic Sciences, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom, [13]Netherlands Metabolomics Center, Leiden, 2333 CC, Netherlands, [14]Division of Systems Biomedicine and Pharmacology, Leiden Academic Centre for Drug Research (LACDR), Leiden University, Leiden, 2333 CC, The Netherlands, [15]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom, [16]Department of Medical Sciences, Clinical Chemistry, Uppsala University, 751 85 Uppsala, Sweden, [17]BBMRI-ERIC, Graz, Austria, [18]INRA, University of Bordeaux, Plateforme Métabolome Bordeaux-MetaboHUB, 33140 Villenave d'Ornon, France, [19]Department of Informatics and Media, Uppsala University, Box 513, 751 20 Uppsala, Sweden, [20]INRA - French National Institute for Agricultural Research, UMR1331, Toxalim, Research Centre in Food Toxicology, Toulouse, France, [21]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, St. Mary's Campus, Norfolk Place, W2 1PG, London, United Kingdom, [22]National Bioinformatics Infrastructure Sweden, Uppsala University, Uppsala, Sweden, [23]Institute of Metabolism and Systems Research (IMSR), University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom, [24]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Deutscher Platz 5e, 04103 Leipzig, Germany, [25]Distributed Computing Group, CRS4, Pula, Italy, [26]CEA, LIST, Laboratory for Data Analysis and Systems' Intelligence, MetaboHUB, Gif-Sur-Yvette F-91191, France, [27]Magnetic Resonance Center (CERM) and Department of Chemistry, University of Florence and CIRMMP, 50019 Sesto Fiorentino, Florence, Italy, [28]ELIXIR-NL, Dutch Techcentre for Life Sciences, Utrecht, 3503 RM, Netherlands, [29]Phenome Centre Birmingham, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom, [30]Cheminformatics and Computational Metabolomics, Institute for Analytical Chemistry, Lessingstr. 8, 07743 Jena, Germany, [31] and [32]

***Correspondence address.** Kristian Peters, Leibniz Institute of Plant Biochemistry, Stress and Developmental Biology, Weinberg 3, 06120 Halle (Saale), Germany. E-mail: kpeters@ipb-halle.de http://orcid.org/0000-0002-4321-0257; James Bradbury, E-mail: j.bradbury@bham.ac.uk; Christoph Steinbeck, E-mail: christoph.steinbeck@uni-jena.de http://orcid.org/0000-0001-6966-0814

Abstract

Background: Metabolomics is the comprehensive study of a multitude of small molecules to gain insight into an organism's metabolism. The research field is dynamic and expanding with applications across biomedical, biotechnological, and many other applied biological domains. Its computationally intensive nature has driven requirements for open data formats, data repositories, and data analysis tools. However, the rapid progress has resulted in a mosaic of independent, and sometimes incompatible, analysis methods that are difficult to connect into a useful and complete data analysis solution. **Findings:** PhenoMeNal (Phenome and Metabolome aNalysis) is an advanced and complete solution to set up Infrastructure-as-a-Service (IaaS) that brings workflow-oriented, interoperable metabolomics data analysis platforms into the cloud. PhenoMeNal seamlessly integrates a wide array of existing open-source tools that are tested and packaged as Docker containers through the project's continuous integration process and deployed based on a kubernetes orchestration framework. It also provides a number of standardized, automated, and published analysis workflows in the user interfaces Galaxy, Jupyter, Luigi, and Pachyderm. **Conclusions:** PhenoMeNal constitutes a keystone solution in cloud e-infrastructures available for metabolomics. PhenoMeNal is a unique and complete solution for setting up cloud e-infrastructures through easy-to-use web interfaces that can be scaled to any custom public and private cloud environment. By harmonizing and automating software installation and configuration and through ready-to-use scientific workflow user interfaces, PhenoMeNal has succeeded in providing scientists with workflow-driven, reproducible, and shareable metabolomics data analysis platforms that are interfaced through standard data formats, representative datasets, versioned, and have been tested for reproducibility and interoperability. The elastic implementation of PhenoMeNal further allows easy adaptation of the infrastructure to other application areas and 'omics research domains.

Keywords: metabolomics; data analysis; e-infrastructures; NMR; mass spectrometry; computational workflows; galaxy; cloud computing; standardization; statistics

Findings

Background

The field of metabolomics has seen remarkable progress over the last decade and has enabled fascinating discoveries in many different research areas. Metabolomics is the study of small molecules in organisms that can reveal detailed insights into metabolic biochemistry, e.g., changes in concentrations of specific molecules, metabolic fluxes between cells or compartments, identification of molecules that are involved in the pathogenesis of a disease, and the study of the biochemical phenotype of animals, plants, and even soil microorganisms [1–3].

The principal metabolomics technologies of mass spectrometry (MS) and nuclear magnetic resonance spectroscopy (NMR) typically generate large datasets that require computationally intensive analyses [4]. Biomedical investigations can involve large cohorts with many thousands of metabolite profiles and can produce hundreds of gigabytes of data [5–8]. With such large datasets, processing becomes impracticable and unmanageable on commodity hardware. Cloud computing can offer a solution by enabling the outsourcing of calculations from local workstations to scalable cloud data centers, with the possibility to allocate thousands of central processing unit (CPU) cores simultaneously. Furthermore, cloud computing allows for resources to be instantiated on-demand (CPUs, random access memory, network, storage) and allows access to computational tools in the form of microservices that can dynamically grow or shrink.

MS and NMR data processing usually involves selection of parameters (that are often specific to the analytical instrumentation), algorithmic peak detection, peak alignment and grouping, annotation of putative compounds, and extensive statistical analyses [9, 10]. Many open-source tools have been developed that address these different steps in data processing and analysis. These tools, however, usually come with their own software dependencies, resource requirements, and scripting languages. As a consequence, configuring and running them is often complicated, especially for researchers who are untrained in computer science [4]. Furthermore, many tools require users to input parameters that can significantly affect results and performance, and reporting of these parameters is not always clear [11].

A number of infrastructures and integration efforts have been initiated in the past five years, including metabolomics data repositories with a global scope [6, 12], platforms for reproducible workflow analysis [13, 14], as well as initiatives to integrate and coordinate data standards [15]. Simultaneously, multiple networks of service centers such as the international Phenome Centers [16] and MetaboHub [17] have formed with the goal to facilitate the acquisition, processing, and analysis of metabolomics data [6–8] at ever increasing scales.

Currently, several web-based metabolomics data processing platforms are available. XCMSOnline provides a platform based on XCMS for downstream data analysis, visualization, data sharing, and access to Metlin to facilitate metabolite identification and pathway analysis [18]. MetaboAnalyst presents a wide variety of data processing and analysis tools including statistical analysis, time-series analysis, functional analysis, and pathway analysis [19]. Workflow4Metabolomics is based on Galaxy and provides various metabolomics processing workflows, including NMR [13, 20]. These common tools for analyzing metabolomics data provide web-based graphical user interfaces (GUIs) with different functionality.

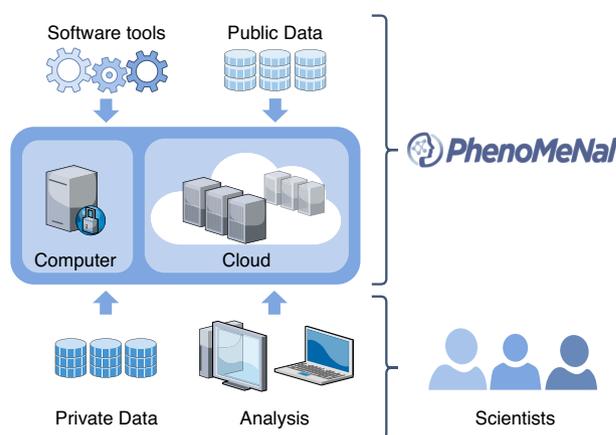Here, we present PhenoMeNal (Phenome and Metabolome aNalysis), a unique, easy-to-use, complete, robust, and per-

Figure 1: Conceptual design of the PhenoMeNal cloud e-infrastructure, which brings compute to the data for any large number of data scientists.

formant cloud e-infrastructure that provides a large suite of standardized and interoperable metabolomics data processing tools as a complete data analysis solution. In contrast to current metabolomics processing platforms, PhenoMeNal provides Infrastructure-as-a-Service (IaaS) and seamlessly integrates a wide array of existing open-source tools.

A major advantage over other platforms is that PhenoMeNal make it possible to instantiate many different services in the cloud and provides a number of standardized, automated, and published analysis workflows in the user interfaces Galaxy, Jupyter, Luigi, and Pachyderm (Fig. 1). Moreover, the PhenoMeNal e-infrastructure can be easily deployed onto public and private cloud environments and can be configured elastically to fit into any cloud-based environment, thus enabling scalable and cost-effective high-performance metabolomics data analysis in a way that hides the technical complexity from the user. PhenoMeNal further facilitates reproducible analyses through automated, sharable, and citable workflows.

Overview

The features of the PhenoMeNal e-infrastructure are encapsulated as a cloud research environment (CRE). The PhenoMeNal CRE can be instantiated on major commercial public cloud providers, including Amazon web services (AWS) and Google cloud platform (GCP), as well as OpenStack-based private clouds and in custom environments. Technical complexity is hidden from the users, simplifying setting up the cloud infrastructure for administrators (Fig. 2).

From a web-based portal, users can deploy the CRE, which includes several web services and software tools (Fig. 2). Data can be processed directly in the e-infrastructure without the need to install additional software. Scientific workflows can be executed via user-friendly web-based platforms such as Galaxy, as well as programmatic interfaces and notebooks. Each service has been supplied with a rich source of documentation and training material to assist researchers.

The PhenoMeNal Portal

The PhenoMeNal Portal [21] allows users to deploy, manage, and delete PhenoMeNal CREs simply through a web interface. Deployments to major commercial cloud platforms (AWS and GCP) as well as OpenStack, an open-source cloud platform, can be made using an easy-to-follow wizard (Fig. 2). OpenStack deploy-
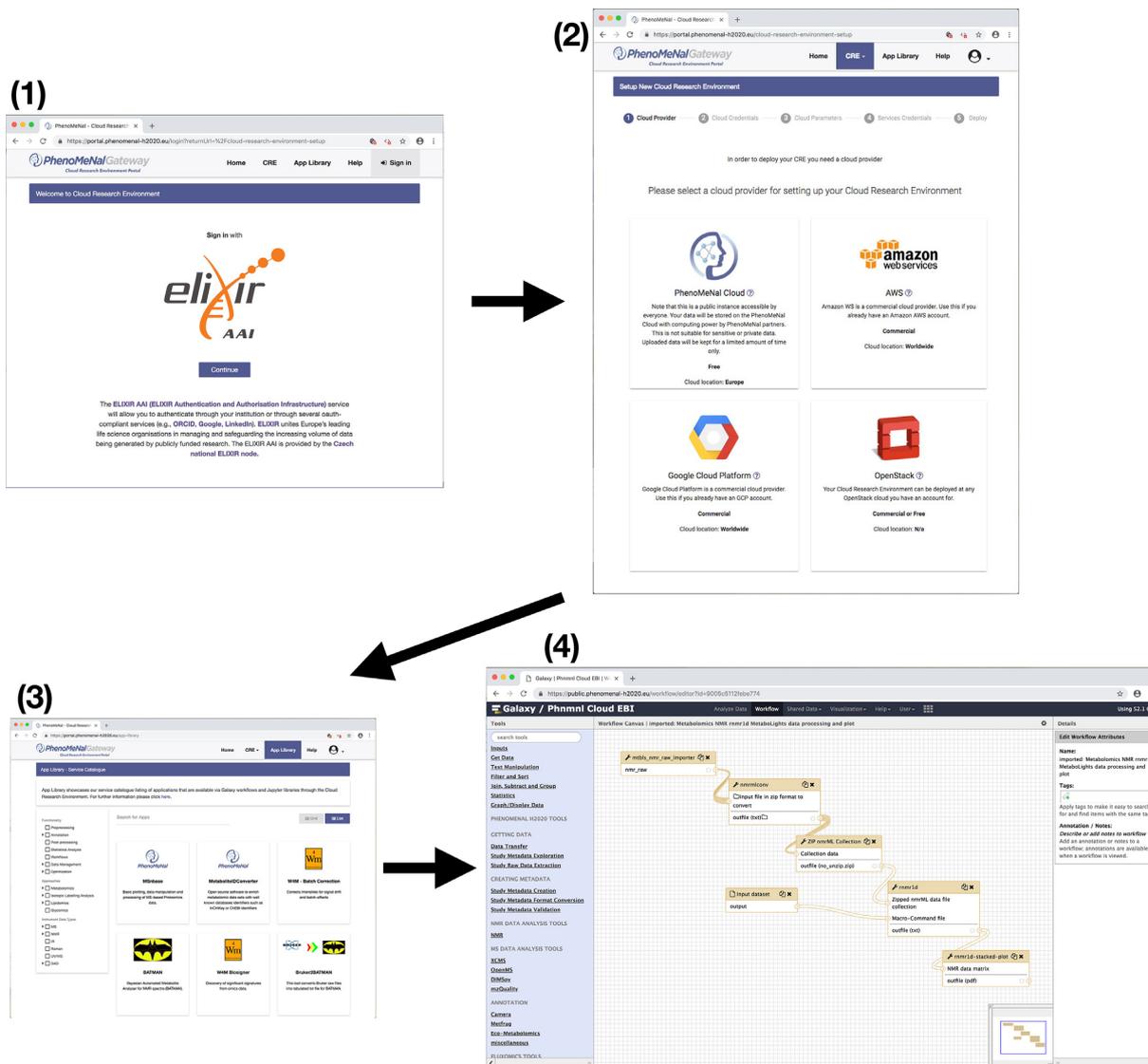
Figure 2: Screenshots of creating and using the PhenoMeNal cloud e-infrastructure. First, log in with ELIXIR to the cloud research environment (CRE) portal. Second, select a public or private cloud provider. After entering cloud credentials and setting up parameters in the dedicated portal, the deployment of the PhenoMeNal e-infrastructure into the cloud environment can be made. Third, in the PhenoMeNal Portal app library there are several services ready to be deployed and used in the set-up infrastructure. Fourth, dedicated web services such as Galaxy are readily available in the cloud e-infrastructure. All steps can be operated from an easy-to-use web interface that is accessible from any standard web browser.

ments can be deployed behind clinical firewalls, which is especially pertinent when dealing with sensitive (i.e., patient) data.

The PhenoMeNal public instance allows users to test-run a CRE without the need to deploy on a cloud platform. It can be deployed and accessed through the portal. Once credentials for users have been generated, analyses can be run through a Galaxy instance containing the tools and workflows present in any deployed CRE. The portal also includes user and developer documentation, workflow tutorials, and links to training videos.

Scientific workflows

A scientific workflow is a set of computational steps that are carried out to process and analyze data [22]. Usually, a workflow is comprised of several linked software tools that are each executed during a particular step of the workflow. In order to manage and automate scientific workflows, in PhenoMeNal the well-established dedicated workflow management system Galaxy

can be deployed, which presents the user with an easy-to-use graphical user interface as well as providing a programmatic interface [20, 23]. Galaxy facilitates collaborative exchange, reproducibility, and traceability of data analysis by enabling users to share entire workflows and analysis histories [24]. In addition to Galaxy, programmatic executable notebooks (Jupyter) and the workflow tools exposed as programmatic interfaces Luigi and Pachyderm are also supported [25].

In order to cover typical use cases in metabolomics and to illustrate the usage and applicability of given analytical pipelines and software tools, five representative scientific workflows are available in the PhenoMeNal Galaxy (Table 1), each having different computational demands and purposes. More than 250 individual modules have been integrated in Galaxy (see the subsection Scientific Workflows in the Methods section).

Table 1: List of workflows that are representative for their respective metabolomics domains (identification in NMR, Fluxomics, Annotation, and identification in MS and eco-metabolomics)

Workflow name	Description	Reference
1D NMR	Processes 1D NMR experiments from raw data to a data matrix required for visualization and statistical analysis, building on nmrML and NMRProcFlow. The automatic workflow is based on the MTBLS1 dataset, describing urinary changes in type 2 diabetes in humans.	[26, 27, 28]
Fluxomics	Quantifies steady-state fluxes following ^{13}C metabolic flux analysis. The workflow was first based on the analysis of the MTBLS412 dataset with 13C tracer data of human umbilical vein endothelial cells under hypoxia.	[29, 30]
LC-MS/MS	Processes, quantifies, and annotates/identifies features in mass spectra using MetFrag — a tool that annotates molecules from compound databases of tandem mass spectrometry (MS/MS) spectra. The workflow is based on MTBLS558.	[31, 32, 33]
Univariate and Multivariate Statistics	Applies univariate and multivariate statistical analysis and illustrates how datasets may be explored, enabling the identification of variables of interest and the construction of predictive models. The workflow is based on MTBLS404.	[13, 34]
Eco-Metabolomics	Implementation of a resource demanding metabolomics use case in ecology, used in large field experiments to describe interactions between different species of organisms in remarkable detail. The workflow is based on MTBLS520.	[35]
ISA-Create-Validate-Upload	A workflow to create Investigation, Study, and Assay data model framework-compliant metadata files based on study design information, augmented with semantic markup as source, implementing UK Phenome center naming conventions. Following validation, the workflow also allows visualization of overall study design and deposition to EMBL-EBI.	

Software tools

The Portal App Library [36] shows the software tools packaged in PhenoMeNal that are available through the CRE deployment (Fig. 2). The range of software tools available covers several metabolomics domains, making PhenoMeNal relevant for use in a wide range of data analysis scenarios. The domains covered include clinical metabolomics, plant metabolomics, fluxomics, and eco-metabolomics. Data from both targeted and untargeted analysis can be analyzed for metabolite profiling and fingerprinting approaches [1, 2]. NMR and MS (liquid chromatography coupled with mass spectrometry, gas chromatography coupled with mass spectrometry, direct infusion mass spectrometry) data can be processed.

PhenoMeNal also provides tools for data management (e.g., via the Investigation, Study, and Assay data model framework [ISA] format and application programming interface [API]), metabolite feature detection (e.g., XCMS, CAMERA, nmrProcFlow), metabolite identification (MetFrag, BATMAN, MetaboMatching), and (bio)statistics (e.g., univariate, multivariate, and power analyses) (Supplementary Table S1). Tools can be filtered for functionality, approaches, and instrument (data) types to readily find the most appropriate software tools. Some tools that implement specific functionality (e.g., Rnmr1D, which performs baseline correction of NMR spectra as part of nmrProcFlow) are available through dedicated Galaxy modules or through software containers (Supplementary Table S1).

Study design

PhenoMeNal was designed to use standardized protocols and software tools and to comply with state-of-the-art dedicated specifications and data formats across the entire project. Development was geared toward implementation of open standards for tracking provenance of both data and metadata generated by clinical phenotyping projects. In PhenoMeNal, the ISA model and specifications were implemented using the ISA format to generate, annotate, validate, and deposit experimental metadata information of datasets and studies to public repositories such as MetaboLights [37, 38]. ISA-based metadata tracking is used for the different analysis pipelines that are specific to the distinct metabolomics domains. PhenoMeNal reached native support for the ISA format by developing a dedicated Galaxy composite data type. Such component affords direct recognition of the ISA format by the Galaxy environment, thus ensuring seamless integration with the downstream workflow component.

Data deposition

PhenoMeNal encourages the metabolomics data repository MetaboLights as a primary source of data deposition [39]. Private and public datasets are supported, as are download and upload to MetaboLights. If the storage in a data repository such as MetaboLights is not possible, data can be stored locally or in the cloud e-infrastructure. Access to the data is strictly controlled and secured. To support data deposition, ISA-based Galaxy modules are available making it possible to publish and disseminate scientific results in standard compliant ways.

Reproducibility

One of the challenges of cloud computing is that analyses need to be run continuously and successfully in different environments [40]. Specifically, it has to be ensured that, given the same input, workflows and tools produce identical results regardless of the underlying environment [4, 40]. When these requirements are fulfilled, end users can be confident that their data will be analyzed correctly. PhenoMeNal has implemented three major testing strategies to ensure technical reproducibility using a continuous integration framework [41]. Tests were implemented for the infrastructure components, individual software containers, and data involved in computational workflows.

Sustainability

PhenoMeNal is part of a number of initiatives (BioMedBridges, COSMOS, and ELIXIR) to foster the role of metabolomics and to harmonize experimental data and metadata usage [15, 42]. Col-

laboratories were established with EGI [43] and Indigo Datacloud [44] infrastructure providers and initiatives [45, 46] to ensure that PhenoMeNal uses technologies that are well supported and ensure their widespread usage, continuity, and further development. For example, the development of KubeNow and contributions to the Galaxy and Workflow4Metabolomics community are essential for PhenoMeNal [47]. Core development will continue on GitHub and is fostered by collaborations with tool developers.

Dependencies on specific technologies and frameworks were avoided by focusing on open standards such as ISA-Tab/ISA-JSON, mzML and nmrML, and widely accepted software [48]. By being able to deploy PhenoMeNal on multiple types of cloud environments, lock-ins to specific computing resource providers are avoided. PhenoMeNal implemented continuous integration and delivery, validated by extensive testing and with clear maintenance responsibilities (see Methods section).

Privacy and security

With human or animal material, the collection, storage, and analysis of metabolomics data introduce a number of constraints due to ethical, legal, and social implications (ELSI) [49]. In particular, data initially derived from human clinical studies may be identifiable and will require consent for use, usually for a defined objective, such as diagnosis, or be related to a particular disease study. Where data is identifiable or pseudonymized, users can deploy PhenoMeNal on local secure resources, thus avoiding the export of data. In this scenario, access to the e-infrastructure should be strictly controlled through local access and authorization. It is recommended that clinical data be fully anonymized before analysis in PhenoMeNal [49, 50].

The PhenoMeNal portal provides substantial guidance to enable users to comply with ELSI and general data protection regulation (GDPR) requirements. Users must register in order to use the individual parts of the e-infrastructure. PhenoMeNal was implemented to use secured and encrypted transport and network communications.

Documentation and training materials

Extensive user documentation and tutorials are provided via the PhenoMeNal Wiki page [51]. The Wiki includes detailed developer resources including information about the PhenoMeNal release schedule; guidelines for tool, workflow, and portal developers; continuous integration; and testing. Further documentation is also provided detailing, creating, and managing PhenoMeNal CREs and tutorials for the Galaxy modules and preconfigured workflows, as well as Galaxy tours that provide step-by-step guidance for inexperienced users.

Community engagement

The PhenoMeNal project is open source and is hosted on GitHub [52]. Developers can contribute tools to PhenoMeNal and are encouraged to do so. To add a tool to PhenoMeNal, it must be containerized using Docker and then integrated into the build process. Detailed documentation is available in the project's Wiki for developers who wish to add their tools to PhenoMeNal.

Collaborations with other projects have been actively encouraged during the development of PhenoMeNal, including Workflow4Metabolomics [13] and the developers of both nmrML and nmrProcFlow [26]. These collaborations are essential to fostering greater standardization within PhenoMeNal and to increasing compatibility with other metabolomics data processing infrastructures.

Availability

Information on how to access PhenoMeNal can be found at the project's website [53]. The GitHub repository hosts the source code of all development projects [52]. The project container-galaxy-k8s-runtime contains all of the developments regarding Galaxy. The Wiki containing documentation is also hosted on GitHub [51]. The PhenoMeNal Portal can be reached at [21]. The public instance of Galaxy is accessible at [54]. Source code and documentation are available under the terms of the Apache 2.0 license. Integrated open-source projects are available under the respective licensing terms.

Conclusions

PhenoMeNal has succeeded in increasing the robustness and coverage of representative metabolomics data processing in scientific cloud e-infrastructures. The presented cloud e-infrastructure covers a wide range of analysis pipelines including data generation and download, data pre- and post-processing, (bio)statistics, and result deposition in data repositories. A large effort has been made to introduce lower-level changes to cloud e-infrastructures (e.g., the cloud deployment software KubeNow) to meet the demands of the biomedical domain. Furthermore, Galaxy has been enriched with metabolomics data standards, in particular, the ISA format for study metadata and mzML and nmrML for acquired data files, as well as support for Kubernetes. PhenoMeNal has fostered the visibility of new metabolomics tools and has enabled the development of more sophisticated data analysis workflows. Our efforts were also guided by feedback from real-life test scenarios collected at workshops with users from the clinical domain.

PhenoMeNal constitutes a keystone solution in cloud platforms available for metabolomics data analysis. The platform was designed to deliver optimal performance and functionality for typical use cases in the metabolomics domain. While the needs of clinicians and researchers in the biomedical and biochemical domains have been targeted, PhenoMeNal is not limited to a specific domain as the cloud infrastructure, tools, and workflows can be adapted to other use cases as demonstrated with the inclusion of the eco-metabolomics workflow. The technological advancements can be reused in other scientific cloud environments and could be integrated with solutions from other 'omics domains in the future.

Methods

Cloud e-infrastructure

The PhenoMeNal CRE is designed as a microservice architecture, with services being implemented as virtual machine images and software containers. Containers are used to provide microservices for metabolomics data analysis tools and also long-running services such as workflow management systems. A container orchestrator runs containers on top of the scalable infrastructure. The orchestrator takes a group of machines that act as a distributed cluster and receives requests for tools as well as service executions. PhenoMeNal implements various layers to providea container orchestrator on top of either bare metal hardware or IaaS given by a cloud provider [55] (Supplementary Fig. S1).

During the setup process and while PhenoMeNal is deployed, data storage and CPU limits can be configured and dynamically scaled to fit any cloud environment. Deployments can be made

to GCE, AWS, and OpenStack-based private clouds from the PhenoMeNal portal. Deployments are also supported from the command line to Microsoft Azure [56], the European Science Cloud [57], and local servers (bare metal) [58]; we provide step-by-step instructions for these solutions.

PhenoMeNal provides IaaS for three different cloud environments:

"local cloud": local workstations or bare metal clusters where data are not allowed to leave the facility.
"public cloud": the flexible use of commercial cloud providers such as GCE and AWS.
"shared cloud": using OpenStack—a free and open-source software platform for cloud computing, ideal for custom environments and research networks.

Software tools

The PhenoMeNal portal has an application library that allows users to deploy tools as microservices into the cloud infrastructure (Fig. 2, Supplementary Table S1). The portal is packaged into frontend and backend engines on top of Kubernetes.

Most software tools in PhenoMeNal are compiled from source code and use a variety of programming languages. Linux versions of software tools and user interfaces such as Galaxy are supported in dedicated encapsulated Docker containers that are implemented as minimum-sized microservices. PhenoMeNal currently hosts 100 such projects in its GitHub repository [59] (Supplementary Table S1). Projects are indicated by the trailing Àcontainer-À name and include a ruleset to build and run the containerized tools, as well as datasets for testing and other necessary files.

PhenoMeNal provides tutorials for developers who want to integrate their tools into our e-infrastructure [60].

Scientific workflows

In PhenoMeNal, a number of options are available for running reproducible and standardized workflows (Table 1).

Galaxy
The Galaxy workflow management system is widely regarded as one of the most popular scientific workflow platforms [20, 61]. It provides a user-friendly web-based GUI to make it easy for the end user to configure and run individual modules and entire workflows without programming experience. Command-line tools and scripts are encapsulated into modules that are launched via the web interface. Galaxy also supports more powerful features such as programmatic access through a REST API and helper libraries to access the running instance of Galaxy [62].

PhenoMeNal has been able to adapt Galaxy for use with a microservices-based architecture [31]. To this end, modules are encapsulated into Docker containers that can be flexibly launched within the cloud e-infrastructure. Galaxy is available in all deployed PhenoMeNal CREs and contains more than 250 modules that have been implemented as part of PhenoMeNal.

Six representative metabolomics Galaxy workflows have been fully integrated into PhenoMeNal (Table 1), and more workflows (mzQuality, NMR-BATMAN) are available for testing.

Jupyter
Jupyter, which started its history as the IPython notebook, is the most popular among the tools commonly referred to as exe-

cutable notebooks or computational notebooks [63]. Jupyter lets users combine executable code with results from code executions such as text, tables, and figures. Usually, Jupyter notebooks are enriched with extended information that explains what the code does. As a result, they are often used for training material and for tutorials. Also, computational notebooks can, to some extent, be used as a way to document code executions and to make executions more reproducible [64].

Luigi and pachyderm
Luigi is a Python workflow programming library that was originally developed by the company Spotify. It manages pipelines of computations primarily on "big data" systems such as Hadoop and Apache Spark but also supports local execution [63, 64]. Luigi is a very flexible library that facilitates building complex pipelines of batch jobs handling dependency resolution, workflow management, and visualization.

Similarly, Pachyderm makes it possible to process distributed data and to keep track of the data from every stage of the analysis pipeline [25]. With Pachyderm, it is possible to track the provenance of results and to accurately reproduce scientific workflows. Luigi and Pachyderm are well suited for complex scientific tasks and are easy to use from the python environment in Jupyter notebooks without additional integration tooling needed.

In PhenoMeNal, we have extended Galaxy, Jupyter, Luigi, and Pachyderm in such a way that they can be orchestrated throughout the cloud infrastructure together with the data analysis tools themselves [31]. Six important metabolomics workflows have been fully integrated into PhenoMeNal (Table 1), and more (mzQuality, NMR-BATMAN) are available for testing.

Reproducibility

Three strategies are realized to ensure technical reproducibility. They are implemented in the continuous integration (CI) software development framework Jenkins [41] which is accessible at [65]. These strategies are implemented as tests in our Jenkins and a tutorial guide is available at [66].

- Infrastructure testing: Procedures were implemented to ensure that each individual component (e.g., the deployment process of software containers, resource management, APIs/application binary interfaces [ABIs]) within the infrastructure is interacting correctly with the other components.
- Container testing: Verification that tools, which are packaged into software containers, build and run correctly in the infrastructure. Dependencies within one container and across several interdependent containers are tested.
- Data testing: The output of tools, which process demonstration data, is checked against a data set that is known to contain the expected result. This is being done for both individual tools and for several tools running in a workflow using the workflow testing tool for Galaxy called wft4galaxy [67].

Standardization

PhenoMeNal has implemented several dedicated Galaxy modules that directly retrieve and store ISA-Tab data set descriptors from and to MetaboLights, and can convert between other formats. Native Galaxy composite data types to support ISA-Tab and ISA-JSON have also been integrated, building upon the ISA API [38, 48]. The ISA data type allows for the upload of an ISA-Tab archive (a zip file containing the ISA set of files and raw

Table 2: Overview of the most important FAIR criteria and implementations suggested for PhenoMeNal data, tools and workflows

	Data	Tools	Workflows
(F)indability	Indexing in domain relevant databases (e.g., MetaboLights)	Indexing in domain relevant software repositories (e.g., the PhenoMeNal App Library, GitHub)	Indexing in workflow management systems such as Galaxy (e.g., PhenoMeNal, W4M), or libraries such as [69]
	Rich descriptions of metadata (e.g., ISA-Tab)	Tool descriptions follow the EDAM ontology	Persistent identifier (e.g., W4M ID, DOI) and intuitive naming patterns
(A)ccessibility	Data access and rights management based on e.g., data use ontology (DUO)	Accessible open-source licenses	Access to workflow systems can be configured to be shared or restricted
(I)nteroperability	Standard formats for experimental metadata (ISA-Tab/ISA-JSON)	Standardized tool descriptions	Standardized workflow format (e.g., Galaxy GA format, Common Workflow Language CWL)
	Domain specific standards for raw data (e.g., mzML, nmrML)	Containerization of software tools	Execution in various software environments (e.g., through the use of containers)
	OboFoundry vocabularies and established domain ontologies to annotate data	EDAM ontology to annotate tools	Workflow annotation ontologies (e.g., Ontology of workflow motifs for annotating workflow specifications [70])
(R)eusability	Deposition in data repositories (e.g., MetaboLights) and data indexing sites (e.g., OmicsDI)	Rich documentation and usage guides	Rich documentation and tutorials (e.g., Galaxy tours)

data when available), which is displayed to the users as a single Galaxy history data set. The integrated Galaxy modules include a MetaboLights downloader and uploader (for ingestion and submission), an ISAcreate module for the creation of ISA compliant archives, modules to explore study metadata through queries on study factors, ISA-Tab "slicing" where queries are used to select subsets of data files of interest, as well as format conversion (export to ISA-JSON and Workflow4Metabolomics [W4M]) and study metadata validation (Supplemental Table S1).

PhenoMeNal also advanced the specification of the nmrML standard data format [27] and contributed a dedicated composite data type for nmrML to Galaxy. nmrML is used extensively throughout the NMR 1D workflow and conversion from raw format into nmrML is supported via dedicated Galaxy modules (Table 1).

Throughout the entire analysis pipeline, modules of computational workflows were designed to accept standard formats such as mzML, XML or CSV whenever possible.

Standardized APIs/ABIs are being used for the programmatic interfaces as well as for deploying services. To this end, modern and standardized programming, scripting and meta languages were selected such as Go, HCL, Python, Shell, XML and YAML that are widely used in cloud computing.

Reusability

In an ongoing effort, PhenoMeNal is actively advancing the criteria for good data management and stewardship based on findability, accessibility, interoperability and reusability (FAIR) for good data management and stewardship [68] to be applied not only to data, but also to software tools and computational workflows (Table 2).

Privacy

PhenoMeNal supports fully anonymized data, which cannot be traced back to individuals in any way [50] and treats pseudonymized data as identifiable. As pseudonymized data are anonymous to the investigator, third parties may be able to link pseudonymized data back to identifiable individuals through mappings such as a hash or code [49]. In these cases, e.g., in a hospital environment, users must deploy PhenoMeNal within a private cloud or bare metal cluster behind their institution's firewall.

PhenoMeNal provides guidance on ethical and technical frameworks to regulate and secure the use of private or sensitive data [49, 50]. It is possible to combine data and metadata within an ELSI compliant framework [50] and in such cases users can follow the example of the European Genome Phenome Archive (EGA) [71]. In public installations of PhenoMeNal, the ELIXIR policy on privacy has been implemented within a technically secure environment to process data [42].

Security

Open-source tools are used throughout the entire e-infrastructure. This promotes community efforts to discover and resolve bugs and security issues. The container build process is steered by the continuous integration (CI) service Jenkins, which continuously builds the containers and generates reports. On success and through authentication, container images are pushed to the PhenoMeNal container registry, which is publicly available but read-only. Cloud provider credentials are not stored in the cloud but only on the deployer host. The Kubernetes cluster running the Jenkins-CI and the container registry, as well as the portal, runs on a CoreOS container, which is a self-updatable, cluster-aware system with most portions being read-only. It reboots nodes sequentially to avoid lack of availability.

KubeNow is a key component that initializes the cloud infrastructure and configures access to it via Cloudflare [72], providing dynamic Domain Name Services (DNS) and encryption for all network communication. The flexible implementation of PhenoMeNal allows the user to decide to not use Cloudflare, in

which case encryption is disabled. KubeAdm, which manages the setup of Kubernetes, is not reachable at runtime by default. The only way to access it is by having access to the private key stored on the computer on which it was launched. PhenoMeNal only allows access to standard ports (ssh, http, https, and port 44 for the Galaxy Downloader) and implements a cloud-specific firewall for all supported cloud providers.

Microservices are designed to be launched on-demand and terminated after completed analysis. If security issues are reported for the microservices, tool, or dependencies or if incremental security patches are available, new builds are automatically triggered in the CI system and developers and the release manager are notified to take additional actions if required. Images are built on a daily basis and tested for deployment to avoid security patches from introducing any abnormality in the deployment process.

User resources

There are many user resources for both PhenoMeNal users and developers in the form of documentation, tutorials, and training videos. The PhenoMeNal Wiki [51] contains detailed documentation on all aspects of PhenoMeNal, including general user guides, workflow and tool tutorials, developer documentation, and general information on topics such as security and the e-infrastructure landscape. The PhenoMeNal portal contains help pages generated from the Wiki [73], which are categorized as User Documentation, Developer Documentation, and Workflow Tutorials. Interactive Galaxy tours are directly integrated in Galaxy [74]. Training videos are available at the project's YouTube page [75].

Availability of source code and requirements

Project name: PhenoMeNal,
Project home page: http://phenomenal-h2020.eu
Operating system(s): Platform independent
Programming language: Go, HCL, Java, JavaScript, Python, R, Shell, XML, YAML
Other requirements: Linux, Docker, Kubernetes, Terraform, Ansible, Helm
License: MIT license for all code written by the PhenoMeNal project. Individual, Open Source Foundation approved licenses for all containerized tools.
RRID:SCR_016605

Availability of supporting data

The following MetaboLights datasets are integrated into PhenoMeNal and are used to demonstrate the cloud integration and reproducibility of Galaxy workflows: MTBLS1 (NMR1D), MTBLS404 (Uni- and multivariate statistics), MTBLS412 (Fluxomics), MTBLS520 (Eco-Metabolomics), MTBLS558 (MetFrag). Datasets are available at https://www.ebi.ac.uk/metabolights. Snapshots of the code and additional supporting data are available in the *GigaScience* repository, GigaDB [76].

Additional files

Supplemental Figure 1: PhenoMeNal implements various layers to provision containers on top of the e-infrastructure.

Supplemental Table 1: List of external software tools that were incorporated into PhenoMeNal.

Abbreviations

ABI: application binary interface; API: application programming interface; AWS: Amazon web services; CI: continuous integration; CPU: central processing unit; CRE: cloud research environment; ELSI: ethical, legal, and social implications; FAIR: criteria for good data management and stewardship based on findability, accessibility, interoperability, and reusability; GCP: Google cloud platform; GUI: graphical user interface; IaaS: Infrastructure-as-a-Service; ISA: Investigation, Study, and Assay data model framework; MS: mass spectrometry; NMR: nuclear magnetic resonance (spectroscopy); PhenoMeNal: Phenome and Metabolome aNalysis; W4M: Workflow4Metabolomics.

Competing interests

The authors declare that they have no competing interests.

Declarations

Human-derived samples in the datasets MTBLS404 and MTBLS412 were processed according to ELSI guidelines.

Author contributions

Writing original draft: K.P. and J.B.
Conceptualization: C.S.
Supervision: R.G., U.L.G., K.H., S.N., A.R., M.vR., C.S., O.S., P.R.-S., R.W.
Project administration: N.K.
Technical lead: P.M.
Review and editing: J.B., M.C., M.Cap., M.Cas., P.dA., T.M.D.E., R.G., A.G.-B., K.H., S.H., D.Ja., D.Jo., F.J., K.K., N.K., P.E.K., A.L., S.L., P.M., S.N., C.O.D., K.P., L.P., M.E.P., M.A.C.R., P.R.-S., P.R.-M., A.R., R.R., C.R., M.vR., N.S., R.M.S., S.-A.S., D.S., O.S., V.S., E.A.T., M.T., T.H., M.vV., M.R.V., R.J.M.W., G.Z., C.S.
Software: J.B., M.Cap., M.Cas., P.dA., A.G.-B., U.L.G., K.H., S.H., D.Jo., F.J., P.E.K., A.L., C.L., P.M., S.N., C.O.D., K.P., L.P., M.E.P., M.A.C.R., P.R.-S., P.R.-M., A.R., R.R., C.R., M.vR., N.S., R.M.S., S.-A.S., O.S., V.S., E.A.T., M.T., T.H., M.vV., R.J.M.W., G.Z.
External software: S.B., C.F., E.H., S.H., M.I., D.Ja., B.K., I.K., K.K., P.E.K., S.L., J.A.N., J.T.M.P., A.P., L.P., R.R.
Data curation: K.H., S.-A.S., P.R.-S.
Funding acquisition: R.G., U.L.G., K.H., S.N., A.R., M.vR., C.S., O.S., P.R.-S., R.W.
Support was provided by the Nordice-Infrastructure Collaboration (NeIC) via the Glenna2 and Tryggve2 projects. All authors contributed to, read, and approved the final manuscript.

Funding

The project was funded by the European Commission PhenoMeNal (grant EC654241). The consortium members include J.B., M.Cap., M.Cas., P.dA., T.M.D.E., R.G., A.G.-B., K.H., M.I., D.Jo., F.J., N.K., P.E.K., A.L., P.M., S.N., C.O.D., K.P., L.P., M.A.C.R., P.R.-S., P.R.-M., A.R., R.R., C.R., T.H., M.vR., M.vV., N.S., R.M.S., S.-A.S., D.S., O.S., V.S., E.A.T., M.T., M.R.V., and R.J.M.W. C.S. received funding from the European Commission PhenoMeNal (grant EC654241).

References

1. Gowda GN, Zhang S, Gu H, et al. Metabolomics-based methods for early disease diagnostics. Expert Rev Mol Diagn 2008;**8**:617–33.

2. Bundy JG, Davey MP, Viant MR. Environmental metabolomics: a critical review and future perspectives. Metabolomics 2009;**5**:3–21.

3. Peters K, Worrich A, Weinhold A, et al. Current challenges in plant eco-metabolomics. Int J Mol Sci 2018;**19**:1385.

4. Weber RJM, Lawson TN, Salek RM, et al. Computational tools and workflows in metabolomics: an international survey highlights the opportunity for harmonisation through Galaxy. Metabolomics 2017;**13**:12.

5. Joyce AR, Palsson BØ. The model organism as a system: integrating "omics" data sets. Nat Rev Mol Cell Biol 2006;**7**:198–210.

6. Haug K, Salek RM, Conesa P, et al. MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. Nucleic Acids Res 2013;**41**:D781–6.

7. Lindon JC, Nicholson JK. The emergent role of metabolic phenotyping in dynamic patient stratification. Expert Opin Drug Metab Toxicol 2014;**10**:915–9.

8. Sumner LW, Hall RD. Metabolomics across the globe. Metabolomics 2013;**9**:258–64.

9. Rosato A, Tenori L, Cascante M, et al. From correlation to causation: analysis of metabolomics data using systems biology approaches. Metabolomics Off J Metabolomic Soc 2018;**14**:37.

10. Vignoli A, Ghini V, Meoni G, et al. High-throughput metabolomics by 1D NMR. Angew. Chem. Int. Ed., 2018, **57**, 2–29, doi:10.1002/anie.201804736.

11. Goodacre R, Broadhurst D, Smilde AK, et al. Proposed minimum reporting standards for data analysis in metabolomics. Metabolomics 2007;**3**:231–41.

12. Sud M, Fahy E, Cotter D, et al. Metabolomics Workbench: an international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. Nucleic Acids Res 2016;**44**:D463–70.

13. Giacomoni F, Le Corguille G, Monsoor M, et al. Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics. Bioinformatics 2015;**31**:1493–5.

14. Haug K, Salek RM, Steinbeck C. Global open data management in metabolomics. Curr Opin Chem Biol 2017;**36**:58–63.

15. Salek RM, Neumann S, Schober D, et al. COordination of Standards in MetabOlomicS (COSMOS): facilitating integrated metabolomics data access. Metabolomics 2015;**11**:1587–97.

16. IPCN. International Phenome Centre Network. http://phenomenetwork.org. 2018. Accessed 25 Oct 2018.

17. French Ministry of Research, Higher Education and the National Agency for Science. MetaboHUB. http://www.metabohub.fr/metabohub.html. 2018. Accessed 25 Oct 2018.

18. Tautenhahn R, Patti GJ, Rinehart D, et al. XCMS Online: a web-based platform to process untargeted metabolomic data. Anal Chem 2012;**84**:5035–9.

19. Chong J, Soufan O, Li C, et al. MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. Nucleic Acids Res 2018;**46**:W486–94.

20. Afgan E, Baker D, van den Beek M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res 2016;**44**:W3–10.

21. PhenoMeNal: The PhenoMeNal Portal. https://portal.phenomenal-h2020.eu. 2018. Accessed 25 Oct 2018.

22. Hoffa C, Mehta G, Freeman T, et al. On the Use of Cloud Computing for Scientific Workflows. 2008 IEEE Fourth Int Conf EScience. Indianapolis, IN, USA: IEEE; 2008 [cited 2018 Sep 3]. p. 640–5. Available from: http://ieeexplore.ieee.org/document/4736878/.

23. Digan W, Countouris H, Barritault M, et al. An architecture for genomics analysis in a clinical setting using Galaxy and Docker. GigaScience 2017;**6**:1–9.

24. Goecks J, Nekrutenko A, Taylor J, Galaxy Team T. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 2010;**11**:R86.

25. Novella JA, Khoonsari PE, Herman S, et al. Container-based bioinformatics with Pachyderm, Wren J . editor. Bioinformatics 2018, 1–8; doi:10.1093/bioinformatics/bty699/5068160.

26. Jacob D, Deborde C, Lefebvre M, et al. NMRProcFlow: a graphical and interactive tool dedicated to 1D spectra processing for NMR-based metabolomics. Metabolomics 2017;**13**:36.

27. Schober D, Jacob D, Wilson M, et al. nmrML: a community supported open data standard for the description, storage, and exchange of NMR Ddta. Anal Chem 2018;**90**:649–56.

28. Salek RM, Maguire ML, Bentley E, et al. A metabolomic comparison of urinary changes in type 2 diabetes in mouse, rat, and human. Physiol Genomics 2007;**29**:99–108.

29. Buescher JM, Antoniewicz MR, Boros LG, et al. A roadmap for interpreting 13 C metabolite labeling patterns from cells. Curr Opin Biotechnol 2015;**34**:189–201.

30. Niedenführ S, Wiechert W, Nöh K. How to measure metabolic fluxes: a taxonomic guide for 13 C fluxomics. Curr Opin Biotechnol 2015;**34**:82–90.

31. Emami Khoonsari P, Moreno P, Bergmann S, et al. Interoperable and scalable data analysis with microservices: Applications in Metabolomics, Journal: bioRxiv. 2018, **bioRxiv:213603**, 1–29 bioRxiv doi:10.1101/213603.

32. Ruttkies C, Schymanski EL, Wolf S, et al. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. J Cheminformatics 2016;**8**:3. http://www.jcheminf.com/content/8/1/3.

33. Herman S, Khoonsari PE, Tolf A. et al. Integration of magnetic resonance imaging and protein and metabolite CSF measurements to enable early diagnosis of secondary progressive multiple sclerosis. Theranostics 2018;**8**:4477–90.

34. Thévenot EA, Roux A, Xu Y. 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. J Proteome Res 2015;**14**:3322–35.

35. Peters K, Gorzolka K, Bruelheide H, et al. Computational workflow to study the seasonal variation of secondary metabolites in nine different bryophytes. Sci Data 2018;**5**:180179.

36. PhenoMeNal. The Portal App Library. https://portal.phenomenal-h2020.eu/app-library. 2018. Accessed 25 Oct 2018.

37. Rocca-Serra P, Salek RM, Arita M, et al. Data standards can boost metabolomics research, and if there is a will, there is a way. Metabolomics. 2016;**12**:14.

38. Smith B, Ashburner M, Rosse CThe OBI Consortium,, et al., The OBI Consortium, The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 2007;**25**:1251–5.

39. Steinbeck C, Conesa P, Haug K, et al. MetaboLights: towards a new COSMOS of metabolomics data management. Metabolomics 2012;**8**:757–60.

40. Gil Y, Deelman E, Ellisman M, et al. Examining the challenges of scientific workflows. Computer 2007;**40**:24–32.

41. Moutsatsos IK, Hossain I, Agarinis C, et al. Jenkins-CI, an

open-source continuous integration system, as a scientific data and image-processing platform. SLAS Discov Adv Life Sci RD 2017;**22**:238–49.

42. van Rijswijk M, Beirnaert C, Caron C, et al. The future of metabolomics in ELIXIR. F1000Research 2017;**6**:1649.

43. EGI Foundation. EGI: Advanced Computing for Research. https://www.egi.eu. 2018. Accessed 25 Oct 2018.

44. INIGO Datacloud. INtegrating Distributed data Infrastructures for Global ExplOitation. https://www.indigo-datacloud.eu. 2018. Accessed 25 Oct 2018.

45. Viljoen M, Dutka L, Kryza B, et al. Towards European Open Science Commons: the EGI Open Data Platform and the EGI DataHub. Procedia Comput Sci 2016;**97**:148–52.

46. Salomoni D, Campos I, Gaido L, et al. INDIGO-DataCloud: a Platform to Facilitate Seamless Access to E-Infrastructures, J Grid Computing, 2018, **16**, 381–408. ArXiv160309536 Cs. doi: 10.1007/s10723-018-9453-3.

47. Capuccini M, Larsson A, Carone M, et al. On-demand virtual research environments using microservices, 10.1093/bioinformatics/bty699/5068160, **arXiv:1805.06180**, 1–31. ArXiv180506180 Cs. 2018; http://arxiv.org/abs/1805.06180.

48. Rocca-Serra P, Brandizi M, Maguire E, et al. ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. Bioinformatics 2010;**26**:2354–6.

49. Sariyar M, Schluender I, Smee C, et al. Sharing and reuse of sensitive data and samples: supporting researchers in identifying ethical and legal requirements. Biopreservation Biobanking 2015;**13**:263–70.

50. Heatherly R, Rasmussen LV, Peissig PL, et al. A multi-institution evaluation of clinical profile anonymization. J Am Med Inform Assoc 2016;**23**:e131–7.

51. PhenoMeNal. Wiki. https://github.com/phnmnl/phenomenal-h2020/wiki. 2018. Accessed 25 Oct 2018.

52. PhenoMeNal. GitHub Project Repository. https://github.com/phnmnl/. 2018. Accessed 25 Oct 2018.

53. PhenoMeNal. Phenome and Metabolome aNalysis. https://phenomenal-h2020.eu. 2018. Accessed 25 Oct 2018.

54. PhenoMeNal. Public Galaxy Instance. https://public.phenomenal-h2020.eu. 2018. Accessed 25 Oct 2018.

55. Mell PM, Grance T. The NIST definition of cloud computing. In: Gaithersburg MD . National Institute of Standards and Technology; 2011. Report No.: NIST SP 800-145. Available from: https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf.

56. PhenoMeNal. Deploy on Microsoft Azure. https://github.com/phnmnl/phenomenal-h2020/wiki/Deploy-on-Microsoft-Azure. 2018. Accessed 25 Oct 2018.

57. PhenoMeNal. Deploy on European Open Science Cloud (EOSC). https://github.com/phnmnl/phenomenal-h2020/wiki/Deploy-on-European-Open-Science-Cloud-(EOSC). 2018. Accessed 25 Oct 2018.

58. PhenoMeNal. Deploy on a local server (bare metal). https://github.com/phnmnl/phenomenal-h2020/wiki/Deploy-on-a-local-server-(bare-metal). 2018. Accessed 25 Oct 2018.

59. Phnmnl GitHub https://github.com/phnmnl/?q=container.

60. PhenoMeNal. How to make your software tool available through PhenoMeNal. https://github.com/phnmnl/phenomenal-h2020/wiki/How-to-make-your-software-tool-available-through-PhenoMeNal. 2018. Accessed 25 Oct 2018.

61. Nekrutenko A, Taylor J. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. Nat Rev Genet 2012;**13**:667–72.

62. Sloggett C, Goonasekera N, Afgan E. BioBlend: automating pipeline analyses within Galaxy and CloudMan. Bioinformatics 2013;**29**:1685–6.

63. Thomas K, Benjamin R-K, Fernando P, et al. Jupyter Notebooks - a publishing format for reproducible computational workflows. Stand Alone. 2016;87–90.

64. Lampa S, Alvarsson J, Spjuth O. Towards agile large-scale predictive modelling in drug discovery with flow-based programming design principles. J Cheminformatics 2016;**8**:67.

65. PhenoMeNal. Jenkins-CI Instance. http://phenomenal-h2020.eu/jenkins/. 2018. Accessed 25 Oct 2018.

66. PhenoMeNal. Jenkins Guide. https://github.com/phnmnl/phenomenal-h2020/wiki/Jenkins-Guide. 2018. Accessed 25 Oct 2018.

67. Piras ME, Pireddu L, Zanetti G. wft4galaxy: a workflow testing tool for galaxy. Bioinformatics 2017;**33**:3805–7.

68. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data 2016;**3**:160018.

69. myExperiment www.myexperiment.org. Accessed 25 Oct 2018

70. Cohen-Boulakia S, Belhajjame K, Collin O, et al. Scientific workflows for computational reproducibility in the life sciences: status, challenges and opportunities. Future Gener Comput Syst 2017;**75**:284–98.

71. Lappalainen I, Almeida-King J, Kumanduri V, et al. The European genome-phenome archive of human data consented for biomedical research. Nat Genet 2015;**47**:692–5.

72. Cloudflare Inc. Cloudflare. https://www.cloudflare.com/. 2018. Accessed 25 Oct 2018.

73. PhenoMeNal. Portal Help. https://portal.phenomenal-h2020.eu/help. 2018. Accessed 25 Oct 2018.

74. PhenoMeNal. Interactive Galaxy Tours. https://public.phenomenal-h2020.eu/tours. 2018. Accessed 25 Oct 2018.

75. PhenoMeNal. The PhenoMeNal YouTube page. https://www.youtube.com/channel/UCXGAvsVNQk-aUpckjRC8Ang. 2018. Accessed 25 Oct 2018.

76. Peters K, Bradbury J, Bergmann S, et al. Supporting data for "PhenoMeNal: Processing and analysis of Metabolomics data in the Cloud." GigaScience Database 2018. http://dx.doi.org/10.5524/100528.

77. Brikman Y. Terraform: Writing Infrastructure as Code. Sebastopol: O'Reilly Media; 2017. Available from: http://public.eblib.com/choice/publicfullrecord.aspx?p=4822376.

78. Hanwell MD, de Jong WA, Harris CJ. Open chemistry: RESTful web APIs, JSON, NWChem and the modern web application. J Cheminformatics. 2017;**9**:55.

79. Newman S. Building microservices: designing fine-grained systems. First Edition. Beijing Sebastopol, CA: O'Reilly Media; 2015.

80. Erl T (Ed.). SOA with REST: principles, patterns & constraints for building enterprise solutions with REST. Upper Saddle River, NJ: Prentice Hall; 2012.

81. Bandrowski A, Brinkman R, Brochhausen M, et al. The Ontology for Biomedical Investigations. PLoS One 2016;**11**:e0154556.

82. Sansone S-A, Rocca-Serra P, Field D, et al. Toward interoperable bioscience data. Nat Genet 2012;**44**:121–6.

83. Sansone S-A, Schober D, Atherton HJ, et al. Metabolomics standards initiative: ontology working group work in progress. Metabolomics 2007;**3**:249–56.

84. Dyke SOM, Philippakis AA, Rambla De Argila, J et al. Consent Codes: upholding standard data use conditions. PLoS Genet 2016;**12**:e1005772.

85 Selivanov VA, Benito A, Miranda, A et al. MIDcor, an R-program for deciphering mass interferences in mass spectra of metabolites enriched in stable isotopes. BMC Bioinformatics 2017;**18**:88.

86 Hao J, Liebeke M, Astle W, et al. Bayesian deconvolution and quantification of metabolites in complex 1D NMR spectra using BATMAN. Nat Protoc 2014;**9**:1416–27.

87 Rinaudo P, Boudah S, Junot C, et al. biosigner: a new method for the discovery of significant molecular signatures from omics data. Front Mol Biosci 2016;**3**:26.

88 Kuhl C, Tautenhahn R, Böttcher C, et al. CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. Anal Chem 2012;**84**:283–9.

89 Dührkop K, Shen H, Meusel M, et al. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. Proc Natl Acad Sci 2015;**112**:12580–5.

90 Southam AD, Weber RJM, Engel J, et al. A complete workflow for high-resolution spectral-stitching nanoelectrospray direct-infusion mass-spectrometry-based metabolomics and lipidomics. Nat Protoc 2017;**12**:255–73.

91 King ZA, Dräger A, Ebrahim A, et al. Escher: a web application for building, sharing, and embedding data-rich visualizations of biological pathways. PLOS Comput Biol 2015;**11**:e1004321.

92 Cottret L, Frainay C, Chazalviel M, et al. MetExplore: collaborative edition and exploration of metabolic networks. Nucleic Acids Res 2018;**46**:W495–502.

93 Libiseller G, Dvorzak M, Kleb U, et al. IPO: a tool for automated optimization of XCMS parameters. BMC Bioinformatics 2015;**16**:118.

94 González-Beltrán A, Neumann S, Maguire E, et al. The Risa R/Bioconductor package: integrative data analysis from experimental metadata and back again. BMC Bioinformatics 2014;**15**:S11.

95 Sansone S-A, Rocca-Serra P, Field D, et al. Toward interoperable bioscience data. Nat Genet 2012;**44**:121–6.

96 Selivanov VA, Vizán P, Mollinedo F, et al. Edelfosine-induced metabolic changes in cancer cells that precede the overproduction of reactive oxygen species and apoptosis. BMC Syst Biol 2010;**4**:135.

97 Perez F, Granger BE. IPython: a system for interactive scientific computing. Comput Sci Eng 2007;**9**:21–9.

98 Ludwig C, Günther UL. MetaboLab - advanced NMR data processing and analysis for metabolomics. BMC Bioinformatics 2011;**12**:366.

99 Wohlgemuth G, Haldiya PK, Willighagen E, et al. The Chemical Translation Service–a web-based tool to improve standardization of metabolomic reports. Bioinformatics 2010;**26**:2647–8.

100 Rueedi R, Mallol R, Raffler J, et al. Metabomatching: using genetic association to identify metabolites in proton NMR spectroscopy. PLOS Comput Biol 2017;**13**:e1005839.

101 Helmus JJ, Jaroniec CP. Nmrglue: an open source Python package for the analysis of multidimensional NMR data. J Biomol NMR 2013;**55**:355–67.

102 Mohamed A, Nguyen CH, Mamitsuka H. NMRPro: an integrated web component for interactive processing and visualization of NMR spectra. Bioinformatics 2016;**32**:2067–8.

103 Sturm M, Bertsch A, Gröpl C, et al. OpenMS – an open-source software framework for mass spectrometry. BMC Bioinformatics 2008;**9**:163.

104 Blaise BJ, Correia G, Tin A, et al. Power analysis and sample size determination in metabolic phenotyping. Anal Chem 2016;**88**:5179–88.

105 Scheubert K, Hufsky F, Petras D, et al. Significance estimation for large scale metabolomics annotations by spectral matching. Nat Commun 2017;**8**, 1–24. doi:10.1038/s41467-017-01318-5.

106 Chambers MC, Maclean B, Burke R, et al. A cross-platform toolkit for mass spectrometry and proteomics. Nat Biotechnol 2012;**30**:918–20.

107 Lewis IA, Schommer SC, Markley JL. rNMR: open source software for identifying and quantifying metabolites in NMR spectra. Magn Reson Chem 2009;**47**:S123–6.

108 Rodriguez N, Thomas A, Watanabe L, et al. JSBML 1.0: providing a smorgasbord of options to encode systems biology models: Table 1. Bioinformatics 2015;**31**:3383–6.

109 Benton HP, Wong DM, Trauger SA, et al. XCMS ^{2}: processing tandem mass spectrometry data for metabolite identification and structural characterization. Anal Chem 2008;**80**:6382–9.