Create, run, share, publish, and reference your LC-MS, FIA-MS, GC-MS, and NMR data analysis workflows with the Workflow4Metabolomics 3.0 Galaxy online infrastructure for metabolomics

Yann Guitton[1,§], Marie Tremblay-Franco[2,§], Gildas Le Corguillé[3], Jean-François Martin[2], Mélanie Pétéra[4], Pierrick Roger-Mele[5], Alexis Delabrière[5], Sophie Goulitquer[6], Misharl Monsoor[3], Christophe Duperier[4], Cécile Canlet[2], Rémi Servien[2], Patrick Tardivel[2], Christophe Caron[7], Franck Giacomoni[4,*], and Etienne A. Thévenot[5,*]

[1]LUNAM Université, Oniris, Laboratoire d'Etude des Résidus et Contaminants dans les Aliments (LABERCA), Nantes, F-44307, France

[2]Toxalim (Research Centre in Food Toxicology), Université de Toulouse, INRA, ENVT, INP-Purpan, UPS, MetaboHUB, Toulouse, France

[3]UPMC, CNRS, FR2424, ABiMS, Station Biologique, 29680, Roscoff, France

[4]INRA, UMR 1019, PFEM, MetaboHUB, 63122, Saint Genes Champanelle, France

[5]CEA, LIST, Laboratory for Data Analysis and Systems' Intelligence, MetaboHUB, F-91191 Gif-sur-Yvette, France

[6]INSERM-UBO UMR1078-ECLA, IBSAM, Faculty of Medicine, University of Brest, 29200 Brest, France

[7]INRA, Ingenum, Toulouse, France

[§]These authors contributed equally to this work.

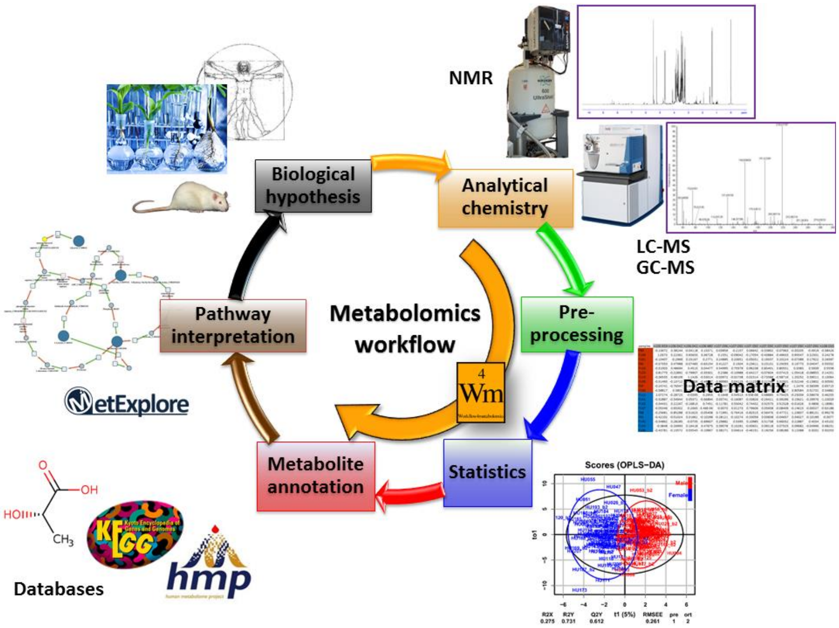
[*]**Corresponding authors:** contact@workflow4metabolomics.org

Abstract

Metabolomics is a key approach in modern functional genomics and systems biology. Due to the complexity of metabolomics data, the variety of experimental designs, and the variety of existing bioinformatics tools, providing experimenters with a simple and efficient resource to conduct comprehensive and rigorous analysis of their data is of utmost importance. In 2014, we launched the Workflow4Metabolomics (W4M; http://workflow4metabolomics.org) online infrastructure for metabolomics built on the Galaxy environment, which offers user-friendly features to build and run data analysis workflows including preprocessing, statistical analysis, and annotation steps. Here we present the new W4M 3.0 release, which contains twice as many tools as the first version, and provides two features which are, to our knowledge, unique among online resources. First, data from the four major metabolomics technologies (i.e., LC-MS, FIA-MS, GC-MS, and NMR) can be analyzed on a single platform. By using three studies in human physiology, alga evolution, and animal toxicology, we demonstrate how the 40 available tools can be easily combined to address biological issues. Second, the full analysis (including the workflow, the parameter values, the input data and output results) can be referenced with a permanent digital object identifier (DOI). Publication of data analyses is of major importance for robust and reproducible science. Furthermore, the publicly shared workflows are of high-value for e-learning and training. The Workflow4Metabolomics 3.0 e-infrastructure thus not only offers a unique online environment for analysis of data from the main metabolomics technologies, but it is also the first reference repository for metabolomics workflows.

249 Words

Graphical abstract



Highlights

- A single online resource for LC-MS, FIA-MS, GC-MS and NMR metabolomics data analysis
- 40 tools for data processing, statistical analysis, and metabolite identification
- The user-friendly Galaxy interface for building, running, saving and sharing workflows
- The first repository for the publication of workflows and histories with a permanent DOI
- Key materials and interactive environment for e-learning and teaching

Keywords: metabolomics, data analysis, e-infrastructure, workflow, Galaxy, repository

Abbreviations:

BPA, Bisphenol A

DOI, Digital Object Identifier

e-infrastructure, online infrastructure

FIA, Flow Injection Analysis

GC, Gas Chromatography

HR, High-Resolution

LC, Liquid Chromatography

MS, Mass Spectrometry

NMR, Nuclear Magnetic Resonance

(O)PLS-DA, (Orthogonal) Partial Least Squares - Discriminant Analysis

PCA, Principal Component Analysis

ppm, parts-per-million

QC, Quality control sample (pool of all samples)

SVM, Support Vector Machine

VIP, Variable Importance in Projection

W4M, Workflow4Metabolomics

Acknowledgements

The Workflow4Metabolomics infrastructure is supported by the French Institute of Bioinformatics (IFB; ANR-11-INBS-0013) and by the French Infrastructure for Metabolomics and Fluxomics (MetaboHUB; ANR-11-INBS-0010). The authors thank Dominique Rolin, Florence Castelli and Jessica Wedow for critical reading of the manuscript, in addition to the reviewers for their valuable comments. We are also grateful to Simon Dittami for sharing the GC-MS data, and acknowledge the support from the IDEALG project (ANR-10-BTBR-04, “Investissements d’avenir, Biotechnologies-Bioressources”).

1. Introduction

Metabolomics is the comprehensive quantification and characterization of the small molecules involved in metabolic chemical reactions (Oliver et al., 1998; Nicholson et al., 1999). It is a promising approach in functional genomics and systems biology for phenotype characterization and biomarker discovery, and has been applied to agriculture, biotechnology, microbiology, environment, nutrition, and health (Holmes et al., 2008; Chen et al., 2012; Rolin, 2013; Kell and Oliver, 2016; Johnson et al., 2016). Complementary analytical approaches, such as Nuclear Magnetic Resonance (NMR) or High-Resolution Mass Spectrometry (HRMS) coupled to Liquid Chromatography (LC) or Gas Chromatography (GC), can be used after minimal sample preparation. These technologies allow routine detection of hundreds to thousands of signals in a variety of biological samples such as cell cultures, organs, biofluids, or biopsies (Cuperlovic-Culf et al., 2010; Brown et al., 2012). Due to the high complexity and large amount of signals generated, however, data analysis remains a major challenge for high-throughput metabolomics (Johnson et al., 2015).

Analysis of metabolomics data (i.e., computational metabolomics) can be divided into three steps: preprocessing of raw data to generate the sample by variable matrix of intensities, statistical analysis to detect variables of interest and build prediction models, and annotation of variables to provide insight into their chemical and biological functions (Fig. 1). The two latter steps (statistics and annotation) can also be performed in the reverse order to get a first-pass overview of the dataset content by performing automatic query of metabolite databases. Furthermore, each step is subdivided into multiple successive, or alternative tasks. For example, preprocessing includes peak detection, denoising, and alignment. Statistical analysis involves normalization, univariate hypothesis testing and multivariate modeling. Finally, annotation relies on peak or spectrum matching with in-house and public databases of metabolites and spectra. This results in a high number of possible combinations of individual tasks to analyze in a dedicated data set. In addition, new methods and software tools constantly emerge to further expand, or refine, metabolomics analysis. Each of them has specific parameters and installation requirements. Typical data analysis by successive use of various software is time-consuming, repetitive, and error-prone: switching from one software to the other requires multiple steps of data manipulation (import/export, up/download, format conversion). Additionally, the workflow (i.e., the sequence of software tools and the parameter values) is not saved, thus preventing efficient and reproducible analysis.

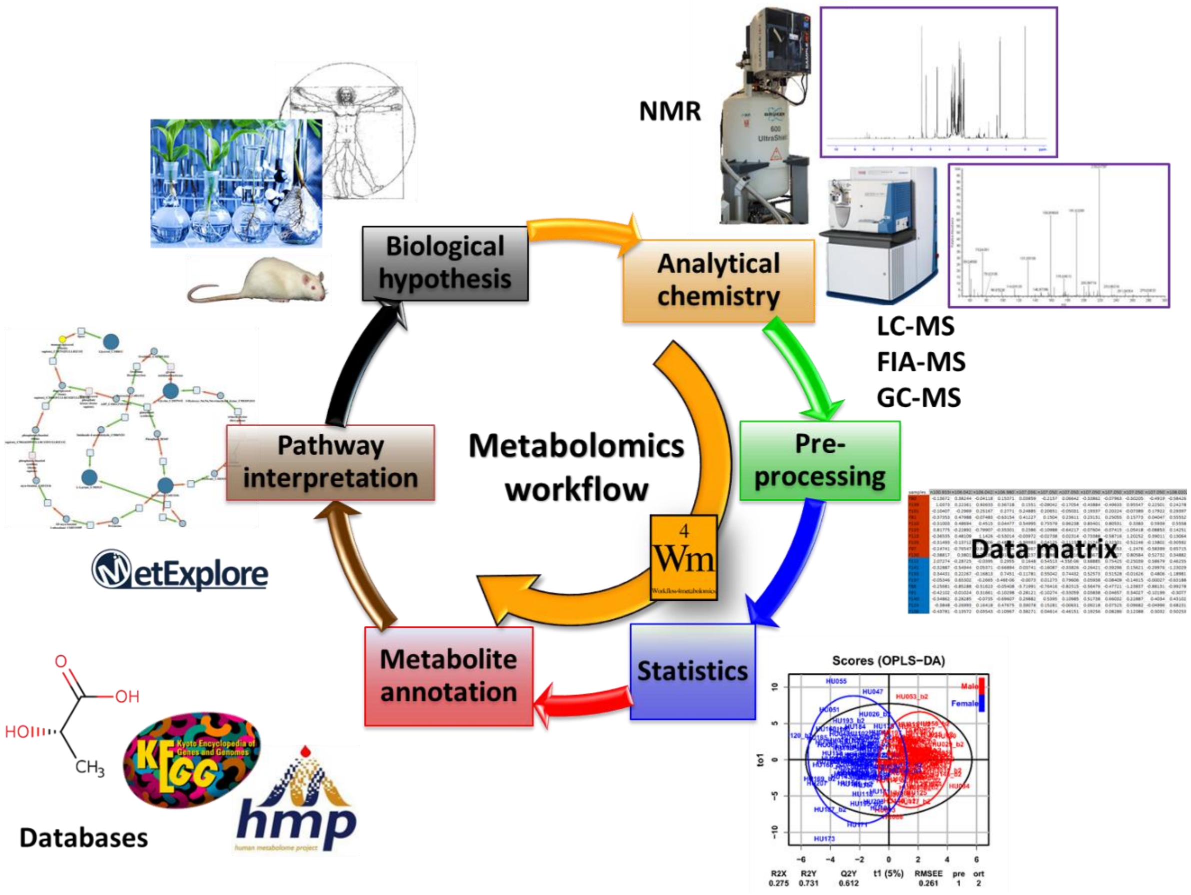
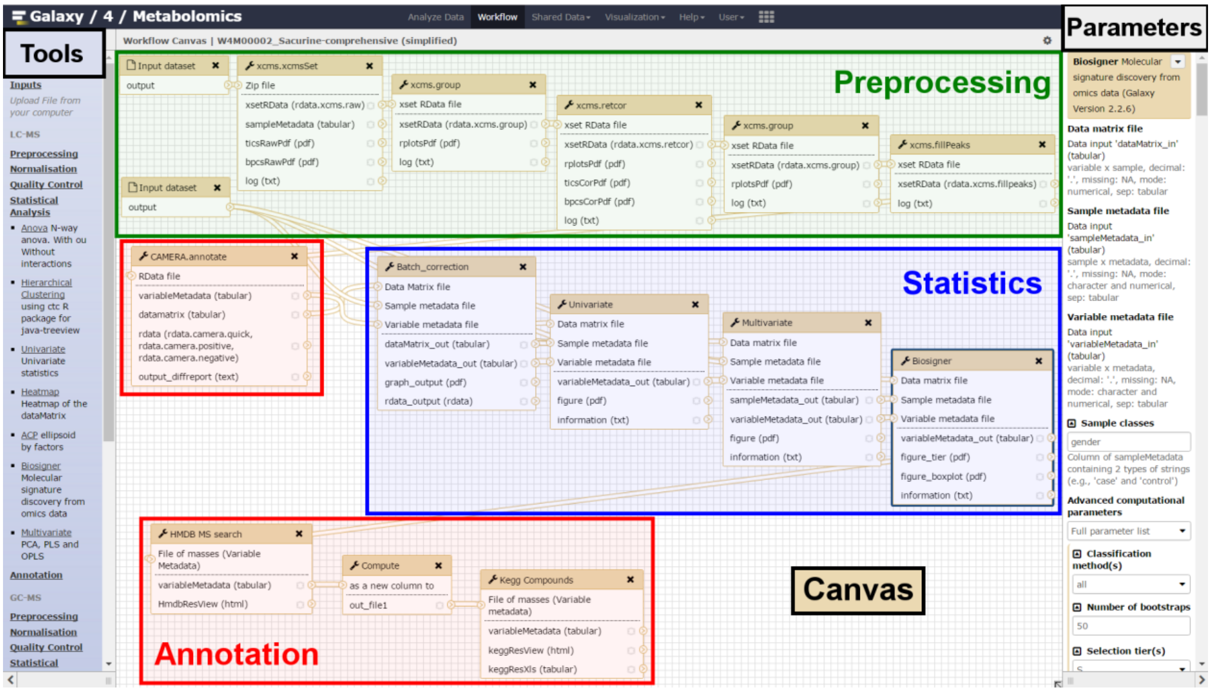


Fig. 1. The Workflow4Metabolomics 3.0 online infrastructure (e-infrastructure) for data analysis of metabolomics experiments. Metabolomics experiments start with the biological question to be addressed, which, in turn, defines the experimental design. Sample collection, preparation, and analysis with NMR or MS instruments are then performed with the appropriate quality controls (reagent blanks, sample pools, etc.). The preprocessing step generates the sample by variable matrix of peak intensities. Statistical analysis includes normalization and batch-effect correction, univariate hypothesis testing, multivariate modeling, and feature selection. Annotation relies on the query of compounds and spectral databases, such as KEGG (Kanehisa and Gotto, 2000) and HMDB (Wishart et al., 2007). Identified metabolites can then be linked within genome-scale reconstructed networks, such as MetExplore (Cottret et al., 2010). More than 40 modules (*tools*) are currently available on the Workflow4Metabolomics e-infrastructure for building comprehensive LC-MS, FIA-MS, GC-MS, and NMR preprocessing, statistical analysis, and metabolite annotation (Table 1).

Workflow management systems are software tools to compose and execute a series of computational tasks in a reproducible way (Leipzig, 2017). In the last decade, software environments with user-friendly features for creating, running, and sharing workflows have been developed, such as Galaxy (Giardine et al., 2005), Taverna (Hull et al., 2006), or KNIME (Berthold et al., 2006). Their graphical interfaces enable users who are not familiar with programming to build their workflow, by selecting tools and their parameters, and chain them in the desired order. Experimenters can therefore concentrate on the scientific design of the analysis and the interpretation of results, without worrying about software installation, command lines, scripts, data format and data management.



Biosigner Molecular signature discovery from omics data (Galaxy Version 2.2.6)

Data matrix file
Data input 'dataMatrix_in' (tabular)
variable x sample, decimal: '.', missing: NA, mode: numerical, sep: tabular

Sample metadata file
Data input 'sampleMetadata_in' (tabular)
sample x metadata, decimal: '.', missing: NA, mode: character and numerical, sep: tabular

Variable metadata file
Data input 'variableMetadata_in' (tabular)
variable x metadata, decimal: '.', missing: NA, mode: character and numerical, sep: tabular

Sample classes
gender
Column of sampleMetadata containing 2 types of strings (e.g., 'case' and 'control')

Advanced computational parameters
Full parameter list

Classification method(s)
all

Number of bootstraps
50

Selection tier(s)
S

Fig. 2. Galaxy features to create workflows on the W4M e-infrastructure. Selected computational *tools* (left panel), with the specific parameters values (right panel), can be chained to build the *workflow* (middle panel). Alternatively, workflows can be directly extracted from the current *history* by selecting *Extract workflow* from the *History options* menu.

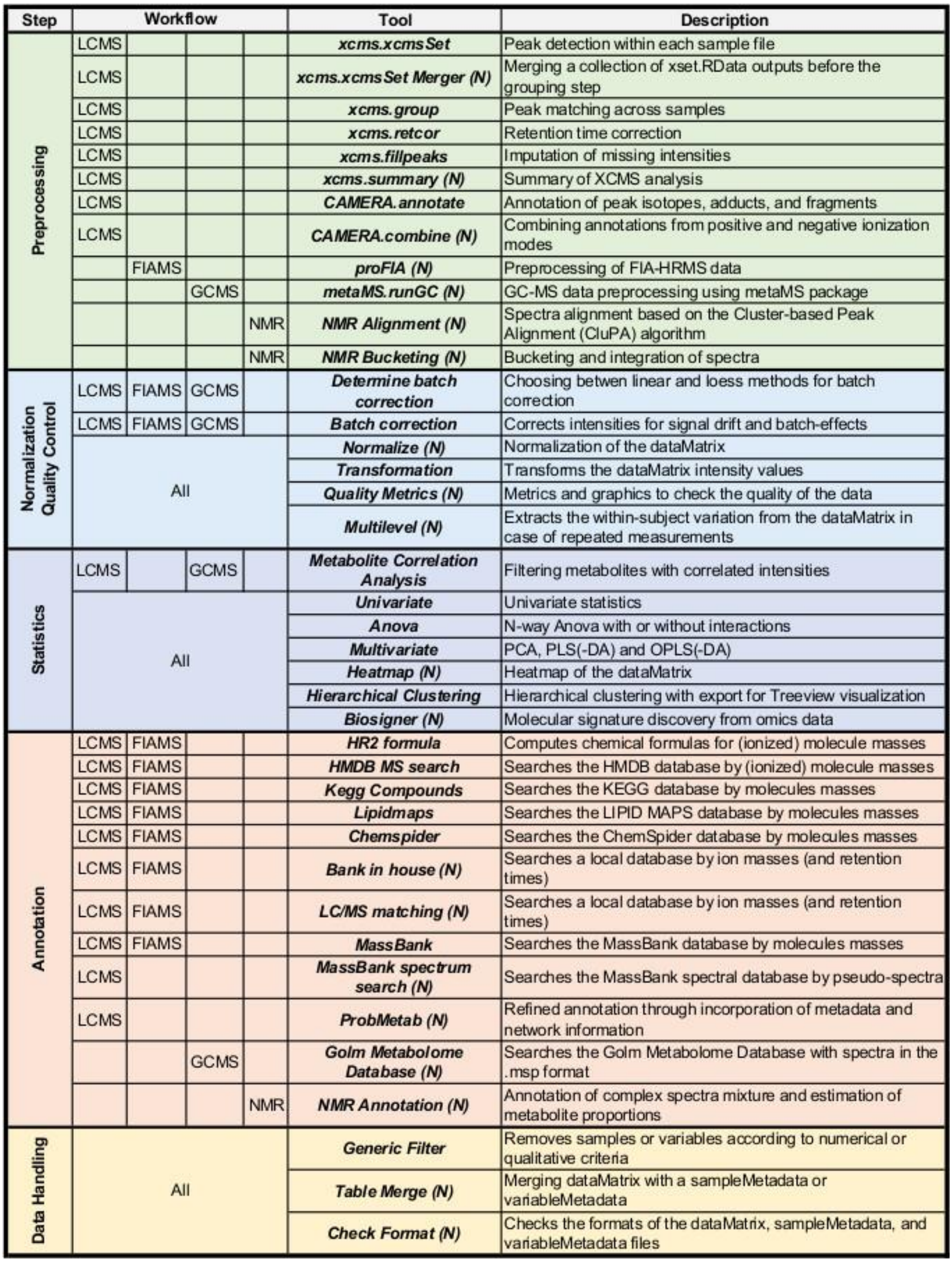
Galaxy is a major environment for workflow management available through a classic web-browser, with more than 50,000 users worldwide and hundreds of tools available (Boekel et al., 2015). The project started in the genomic community (Giardine et al., 2005; Goecks et al., 2010), and further expanded to other omics fields such as proteomics (Boekel et al., 2015; Jagtap et al., 2014; Jagtap et al., 2015). The Galaxy environment provides intuitive and powerful features that enable the experimenter to build and run complex workflows. For example, a tool can be re-run after changing a single parameter value with only two mouse clicks. The *workflow* (chained tools + selected parameter values) can be designed within a graphical editor interface (named *canvas*; Goecks et al., 2010; Fig. 2), and then run on the input data to obtain a *history* (workflow + attached input and output data). Alternatively, a history can be built by tuning the successive tools sequentially on the selected data set, and, once the history is completed, by extracting the corresponding workflow. A key feature is that both the workflow and the history can be saved and shared between users.

To develop Galaxy pipelines for metabolomics, and make them available to the user community worldwide, we created the Workflow4Metabolomics online infrastructure (W4M e-infrastructure; Giacomoni et al., 2015). At that time, W4M already provided 20 tools (some of them being newly developed for the infrastructure, while others corresponded to the integration of existing tools), which enabled the build of comprehensive workflows for LC-HRMS data analysis (Table 1). W4M not only gave access to the Galaxy environment to build workflows, but also offered a high-performance computing environment to run the analyses, online documentations, and a help-desk served by 8 bioinformaticians from the Core Team.

Here, we present the new 3.0 version of the W4M e-infrastructure. The 20 new tools (Table 1) enable advanced workflows not only for MS technologies (LC-MS, GC-MS and Flow Injection Analysis: FIA-MS) but also for NMR data. Furthermore, the complete histories can be referenced online with a permanent DOI, thus enabling fully reproducible analyses. To demonstrate how the new computational features from W4M 3.0 can be used to address biological issues, we selected three real LC-MS, GC-MS, and NMR case studies from published studies in human physiology, mouse toxicology, and algae evolution. In the next section (*Case Studies*), we briefly recall the objective of the three studies and the analytical methods used to generate the raw data. In the *Results* section, we then create three workflows to analyze the data, and compare the outputs with the published results.

Table 1

List of available tools on the W4M online infrastructure. The 20 new tools from the 3.0 release are indicated with (N). The source code of the W4M tools is available on the Galaxy toolshed (https://toolshed.g2.bx.psu.edu).



Step	Workflow				Tool	Description
Preprocessing	LCMS				***xcms.xcmsSet***	Peak detection within each sample file
	LCMS				***xcms.xcmsSet Merger (N)***	Merging a collection of xset.RData outputs before the grouping step
	LCMS				***xcms.group***	Peak matching across samples
	LCMS				***xcms.retcor***	Retention time correction
	LCMS				***xcms.fillpeaks***	Imputation of missing intensities
	LCMS				***xcms.summary (N)***	Summary of XCMS analysis
	LCMS				***CAMERA.annotate***	Annotation of peak isotopes, adducts, and fragments
	LCMS				***CAMERA.combine (N)***	Combining annotations from positive and negative ionization modes
		FIAMS			***proFIA (N)***	Preprocessing of FIA-HRMS data
			GCMS		***metaMS.runGC (N)***	GC-MS data preprocessing using metaMS package
				NMR	***NMR Alignment (N)***	Spectra alignment based on the Cluster-based Peak Alignment (CluPA) algorithm
				NMR	***NMR Bucketing (N)***	Bucketing and integration of spectra
Normalization Quality Control	LCMS	FIAMS	GCMS		***Determine batch correction***	Choosing betwen linear and loess methods for batch correction
	LCMS	FIAMS	GCMS		***Batch correction***	Corrects intensities for signal drift and batch-effects
	All				***Normalize (N)***	Normalization of the dataMatrix
					Transformation	Transforms the dataMatrix intensity values
					Quality Metrics (N)	Metrics and graphics to check the quality of the data
					Multilevel (N)	Extracts the within-subject variation from the dataMatrix in case of repeated measurements
Statistics	LCMS		GCMS		***Metabolite Correlation Analysis***	Filtering metabolites with correlated intensities
	All				***Univariate***	Univariate statistics
					Anova	N-way Anova with or without interactions
					Multivariate	PCA, PLS(-DA) and OPLS(-DA)
					Heatmap (N)	Heatmap of the dataMatrix
					Hierarchical Clustering	Hierarchical clustering with export for Treeview visualization
					Biosigner (N)	Molecular signature discovery from omics data
Annotation	LCMS	FIAMS			***HR2 formula***	Computes chemical formulas for (ionized) molecule masses
	LCMS	FIAMS			***HMDB MS search***	Searches the HMDB database by (ionized) molecule masses
	LCMS	FIAMS			***Kegg Compounds***	Searches the KEGG database by molecules masses
	LCMS	FIAMS			***Lipidmaps***	Searches the LIPID MAPS database by molecules masses
	LCMS	FIAMS			***Chemspider***	Searches the ChemSpider database by molecules masses
	LCMS	FIAMS			***Bank in house (N)***	Searches a local database by ion masses (and retention times)
	LCMS	FIAMS			***LC/MS matching (N)***	Searches a local database by ion masses (and retention times)
	LCMS	FIAMS			***MassBank***	Searches the MassBank database by molecules masses
	LCMS				***MassBank spectrum search (N)***	Searches the MassBank spectral database by pseudo-spectra
	LCMS				***ProbMetab (N)***	Refined annotation through incorporation of metadata and network information
			GCMS		***Golm Metabolome Database (N)***	Searches the Golm Metabolome Database with spectra in the .msp format
				NMR	***NMR Annotation (N)***	Annotation of complex spectra mixture and estimation of metabolite proportions
Data Handling	All				***Generic Filter***	Removes samples or variables according to numerical or qualitative criteria
					Table Merge (N)	Merging dataMatrix with a sampleMetadata or variableMetadata
					Check Format (N)	Checks the formats of the dataMatrix, sampleMetadata, and variableMetadata files

2. Case Studies

We have selected three case studies which illustrate the diversity of the biological issues, experimental designs, and analytical technologies in metabolomics. Here, we briefly describe the context, objective, and analytical methods of these three published studies.

2.1. Sacurine human physiological study (LC-HRMS)

Human urine has been used since Antiquity for disease prediction. Nowadays, this biofluid shows great promise for biomarker discovery in metabolomics. Characterization of the variations of the urine metabolome with age, BMI, and gender, is therefore critical not only to better understand human physiology, but also to avoid confounding effects in biomarker studies. Since physiological information about urine concentrations is scarce in metabolomics databases, a cohort of 184 volunteers from the CEA research institute was studied (Roux et al., 2012; Thevenot et al., 2015). Urine samples were analyzed by Ultra-high Performance Liquid Chromatography (Hypersil GOLD C18 column) coupled to High-Resolution Mass Spectrometry (LTQ-Orbitrap Discovery, Thermo Fisher Scientific).

The work by Roux et al. (2012) focused on the identification of the metabolites in urine, while the second study (Thevenot et al., 2015) analyzed the variations of their concentrations with age, body mass index (BMI), and gender. Raw data were preprocessed with XCMS (Smith et al., 2006) and annotated with CAMERA (Kuhl et al., 2012), and a selection of metabolites of putative interest were identified at levels 1 and 2 (Metabolomics Standard Initiative; Sumner et al., 2007) by matching with the KEGG (Kanehisa and Gotto, 2000), HMDB (Wishart et al., 2007) and METLIN (Smith et al., 2005) databases, followed by interpretation of additional MS/MS fragmentation experiments (Roux et al., 2012). Quantification of their intensities was refined by visual determination of the peak limits in the raw data (Quan Browser tool from the Xcalibur software; Thevenot et al., 2015).