Fig. 6. Quality control and annotation of the “Unknown 2” pseudospectrum. Top: The EIC of the pseudospectrum clearly indicates that all grouped ions belong to a single compound (no co-elution). Only ions with correlated chromatographic profiles are grouped by the metaMS algorithm, resulting in a cleaned pseudospectrum for further spectral database annotation (Wehrens et al., 2014). Bottom: After matching to the Golm and NIST databases, the “Unknown 2” pseudospectrum was annotated as citric acid, which is the correct identification (confirmed by injection of the pure compound).

Prior to data upload, the 12 GC-MS raw files (4 biological replicates of the 3 cultures) were converted from the Agilent commercial format (*.D) into the open NetCDF format, by using the Agilent Chemstation software (‘export as AIA/ANDI Files’ menu; Note that the freely available ProteoWizard software can also be used for file conversion; Chambers et al., 2012). The converted files from the W4M00004 study are available for download from W4M as a unique zipped file (260.8 MB; http://workflow4metabolomics.org/datasets).

Application of the workflow to the raw data resulted in a history containing 21 files generated in 11 minutes (7 min for peak picking and 4 min for statistics). The data matrix contained 52 pseudospectra. The quality of each pseudospectrum was checked visually on the *GCMS_EIC.pdf* output (Fig. 6, top). All pseudospectra were then matched against the Golm and NIST spectral databases, by using the *peakspectra.msp* file (Fig. 6, bottom). As a control, the internal standard ribitol was confirmed to be the “Unknown 5”. Importantly, “Unknown 2” and “Unknown 4” were annotated as citric acid and mannitol, respectively (Fig. 6, bottom). Surprisingly, mannitol was not detected in two samples (alg 2 and alg 3), because the retention time shift (15 s) was superior to the default threshold in ***metaMS.runGC*** (3 s). We therefore used an option from ***metaMS.runGC*** which enables to refine the alignment of pseudospectra between samples by providing a database of reference spectra (Wehrens et al., 2014): first, we created a spectral database with ribitol, citric acid, and mannitol, from the *peakspectra.msp* file generated previously; second, we re-run the ***metaMS.runGC*** tool with this additional spectral information, and successfully detected mannitol in all samples (feedback step on Fig. 5). We could then perform the downstream statistical analyses. The matrix of intensities (*dataMatrix*) was first normalized by dry weight of sample. Exploratory data analysis was then performed by

PCA: distinct clusters corresponding to three algal cultures were observed on the score plot. In particular, a clear metabolic shift was observed for the FWS cultures from the low versus seawater saline conditions. The decreased concentrations of mannitol, which is a putative osmoprotector, in the FS strain, were in accordance with the results from Dittami et al. (2012). Finally, PLS-DA discrimination between the 3 cultures resulted in a classifier with a significant ($p < 0.05$, when compared with models built after random permutation of the labels) and high Q2 value (Q2Y = 0.84). These multivariate analyses, which were not included in the original publication, therefore provide complementary information.

3.1.3. Brain toxicity of Bisphenol A: The *W4M00006 NMR BPAMmusculus* history

The aim of this study was to assess the effect on the brain metabolome of perinatal exposure to Bisphenol A (BPA), an endocrine disruptor widely used in plastics and resins (Cabaton et al. 2013). To analyze NMR data on W4M, specific preprocessing and annotation modules have been developed for the 3.0 version (see below). The implemented workflow (Fig. 7) corresponds to a complementary pairwise comparison study (BPA0.025 vs BPA0.25), which was not presented in the original publication. The 0.25 and 0.025 µg BPA/kg body weight/day treatments correspond to 1/100 and 1/000 of the tolerable daily intake (TDI: “Estimated maximum amount of an agent, expressed on a body mass basis, to which individuals may be exposed daily over their lifetimes without appreciable health risk”; World Health Organization, 2004) and were picked up to demonstrate that even at very low doses of exposure, BPA is still modulating differently the brain metabolome of the CD1 mice.

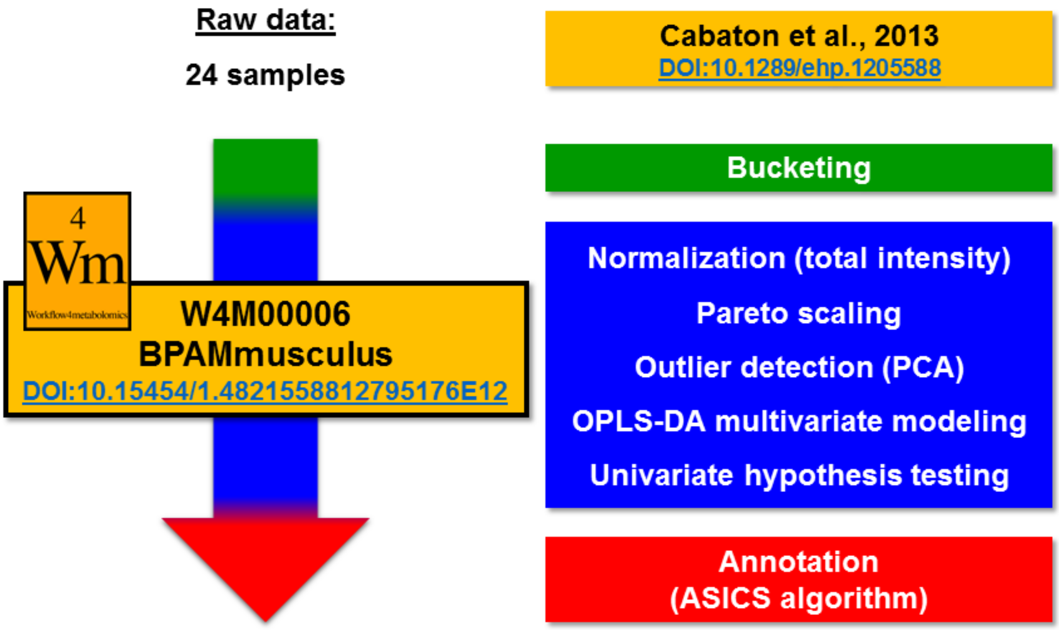
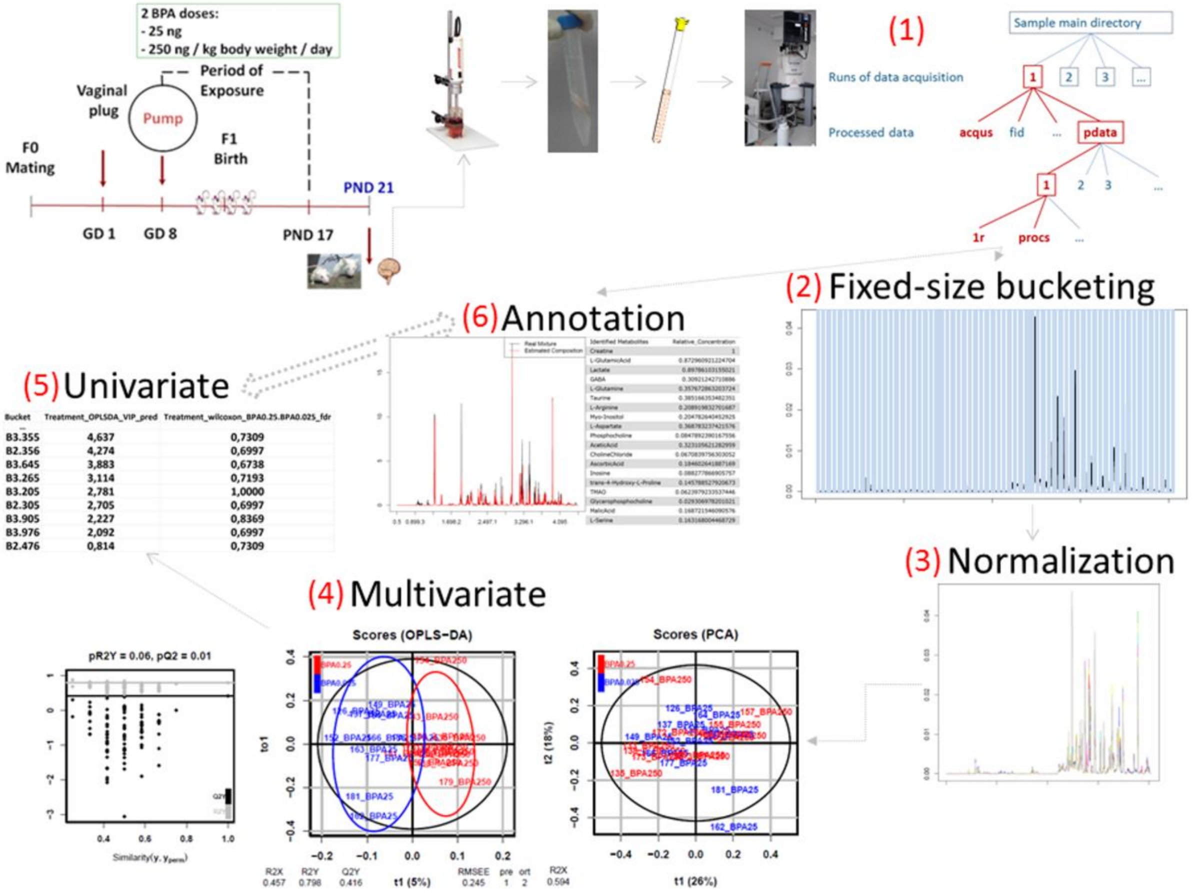


Fig. 7. *W4M00006 BPAMmusculus* reference workflow (DOI:10.15454/1.4821558812795176E12). The input data (24 samples) were acquired using ^{1}H NMR spectroscopy. TopSpin (Bruker) preprocessed files (Bruker format) are available on W4M for download (http://workflow4metabolomics.org/datasets).

The 7 tools from the *BPAMmusculus* workflow are detailed in Table S3 and illustrated on Fig. 8. First, the 24 TopSpin-preprocessed spectra were uploaded into W4M as a single zipped file (see the *Workflow Management with W4M* section). Second, spectra were segmented into 809 buckets by using the ***NMR Bucketing*** tool: this tool divides the whole spectrum into “small” fixed-size windows (e.g., 0.01 ppm). In addition, spectrum regions corresponding to water, solvent or contaminant resonances can be excluded. Finally, the sum of intensities inside each bucket (area under the curve) is computed by using the trapezoidal method. Third, spectra were normalized to the Total Intensity with the ***NMR Normalization*** tool. The objective of sample normalization is to make the data from all samples directly comparable with each other (i.e., to remove systematic biological and technical variations). The ***NMR Normalization*** tool includes 3 normalization methods: Total intensity (each bucket integration is divided by the integration of the total spectrum), quantitative variable (e.g., sample weight, osmolality), and Probabilistic Quotient Normalization (PQN; Dieterle et al., 2006), where each spectrum is compared to a reference sample (e.g., the median spectrum of control samples). Fourth, exploratory data analysis and multivariate modeling were performed with the ***Multivariate*** tool. PCA was first used to detect outliers: two observations were excluded for subsequent analyses. Then, an OPLS-DA classifier of the two treatment doses was built: the model was significant (permutation test *p*-value < 0.05), and 157 variables (buckets) having a VIP value > 0.8 were selected. In parallel, univariate analysis of differences between the two doses was performed with the Wilcoxon-Mann-Whitney test (***Univariate*** tool). No significant difference was observed after correction for multiple testing (False Discovery Rate set to 5%).



Bucket	Treatment_OPLSDA_VIP_pred	Treatment_wilcoxon_BPA0.25.BPA0.025_fdr
B3.355	4,637	0,7309
B2.356	4,274	0,6997
B3.645	3,883	0,6738
B3.265	3,114	0,7193
B3.205	2,781	1,0000
B2.305	2,705	0,6997
B3.905	2,227	0,8369
B3.976	2,092	0,6997
B2.476	0,814	0,7309

Fig. 8. *W4M00006 BPAMmusculus*: From experiment to metabolite annotation. Top: Experimental design, sample preparation, and acquisition of NMR spectra (Bruker files). (1-3) Data reduction (bucketing + normalization). (4) Multivariate analysis (PCA and OPLS-DA score plots, and permutation tests). (5-6) Annotation of significant metabolites.

Sixth, buckets were annotated with the ***NMR Annotation*** tool. This tool decomposes any input spectrum from a complex biological matrix into a mixture of spectra from pure compounds provided as a reference database. The internal database currently contains 175 spectra of pure compounds, which were acquired on a Bruker Avance III

600 MHz NMR spectrometer, at pH 7.0. The deconvolution algorithm uses a penalized regression to compute a parsimonious list of non-zero coefficients (i.e., proportions) for the reference spectra detected in the mixture (Tardivel et al., submitted). The output *proportionEstimation* table contains the identified metabolites with their estimated relative concentration (the highest concentration being arbitrarily set to 1). The ***NMR Annotation*** tool was applied to one spectrum from each of the two classes (BPA0.025 and BPA0.25), and resulted in the annotation of 39 metabolites. Among them, the Glutamic Acid and the GABA neurotransmitters had high VIP predictive values (4.1 for the 2.35 ppm bucket, and 2.6 for the 2.3 ppm bucket, respectively), which confirmed the previous results showing a decrease of the concentrations of these metabolites following exposure to BPA (Cabaton et al., 2013). Furthermore, these new analyses showed that two other neurotransmitters, Taurine (3.44-3.41, 3.26-3.24 ppm) and Aspartate (3.91-3.89, 2.7-2.69 ppm), had high discriminative values (VIP of 2.9 for the 3.26 ppm bucket, and 2.2 for the 3.90 ppm bucket, respectively). To our knowledge, this is the first time that comprehensive NMR workflows (from preprocessing to identification) can be created, run, and published online.

3.2. Referencing histories

The *histories* from each case study (i.e., workflow and the associated input data and output result files) were further *published* on W4M (i.e., shared with the community) and referenced with a digital object identifier (DOI) which can be cited in publications (Table 2). Referenced histories can be imported by any user into his/her account, and the workflow can be extracted from the history with the *Extract Workflow* functionality, e.g., for application to new data sets.

Making workflows and associated data available to the community is essential to demonstrate the value and the reproducibility of the analysis (Mons et al., 2011). As for raw data, journal editors will increasingly recommend that the process of generating the results (code, parameter values, output data) is made available on reference repositories. Funding agencies such as European Programs also require that the generated data are made public on reference online resources. Finally, sharing analyses gives experimenters the opportunity to receive feedback on their results, get cited, and initiate new collaborations. While databases for raw data and metabolites already exist, the W4M e-infrastructure is the first repository for data analysis workflows dedicated to metabolomics.

Referencing histories on W4M is straightforward (see the “Referenced Workflows and Histories” section on the home page). Authentication is required to access the shared histories in Galaxy. Extra anonymous credentials for reviewers, however, may be provided to authors from reference histories when submitting their manuscript. Six histories have already been referenced (Table 2), and have been cited in publications such as Thevenot et al. (2015), Rinaudo et al. (2016), and Peng et al. (2017).

Table 2

Publicly referenced histories (workflows and associated data and result files; http://workflow4metabolomics.org/referenced_W4M_histories). The DOI points to a landing page which details the main steps, data size, and name of the maintainer, and gives access the whole history online.

ID	Raw data	Technol.	Species	Matrix	Factor(s)	Nb samples	Size	Publication
W4M00001_ Sacurine-statistics DOI:10.15454/1.4811121736910142E12	MTBLS404	LC-MS	H. sapiens	urine	age, BMI, gender	210	4 MB	Thevenot et al., 2015 DOI:10.1021/acs.jproteome.5b00354
W4M00002_ Sacurine-comprehensive DOI:10.15454/1.481114233733302E12	MTBLS404	LC-MS	H. sapiens	urine	age, BMI, gender	234	18 GB	Thevenot et al., 2015 DOI:10.1021/acs.jproteome.5b00354
W4M00003_ Diaplasma DOI:10.15454/1.4811165052113186E12	N/A	LC-MS	H. sapiens	plasma	diabetic type	63	11 MB	Rinaudo et al., 2016 DOI:10.3389/fmolb.2016.00026
W4M00004_ GCMS-Algae DOI:10.15454/1.4811272313071519E12	N/A	GC-MS	E. siliculosus	alga	salinity	12	260 MB	Dittami et al., 2012 DOI:10.1111/j.1365-313X.2012.04982.x
W4M00005_ Ractopamine-Pig DOI:10.1545	MTBLS384	LC-MS	S. Scrofa	serum	Ractopamine	164	327 MB	Peng et al., 2017 DOI:10.1007/s11306-017-1212-0

4/1.4811287270056958E12								
W4M00006_BPA-Mmusculus DOI:10.15454/1.4821558812795176E12	W4M	NMR	Mus Musculus	brain	BPA dose	24	7 MB	Cabaton et al., 2013 DOI:10.1289/ehp.1205588

4. Discussion

High-throughput analysis of the metabolic phenotype has a profound impact on the understanding of biochemical reactions and physiology, and prediction of disease (Peyraud et al., 2011; Balog et al., 2013; Etalo et al., 2015; Li et al., 2016; Weiss et al., 2016). To cope with the volume and complexity of data generated by modern high-resolution MS and NMR instruments, adapt to the diversity of experimental designs, and ensure rigorous and reproducible analyses, there is a strong need for user-friendly, modular, and computationally efficient software platforms. To demonstrate how the Workflow4Metabolomics 3.0 e-infrastructure meets this workflow challenge, we designed, run, and referenced online three comprehensive analyses of previously published LC-MS, GC-MS, and NMR data studies from human physiology, mouse toxicology, and alga evolution. The workflow sequence, parameter values, and critical points are detailed to provide examples of critical design and interpretation of analyses. The outputs of W4M analyses confirmed the published results: for each case study, key metabolites were selected and annotated, with significant concentration differences either between gender, BMI, and age (human study), or after perinatal exposition to Bisphenol A (mouse study), or in response to salinity stress (algae study). Furthermore, they shed new lights on the datasets, e.g., by suggesting variations of neurotransmitter concentrations even at the lowest doses of BPA exposure (mouse study). Workflow building and running was very efficient: the *Sacurine-comprehensive* workflow, which contains 29 tools for preprocessing, statistical analysis and annotation, could be run on 234 raw files in a few hours. Workflow management was straightforward, as illustrated by the comparison between the *matchedFilter* and *centWave* approaches for LC-MS preprocessing. Together, these results demonstrate that comprehensive LC-MS, GC-MS, and NMR data analyses can be readily designed and run on the W4M e-infrastructure.

The number of available tools on W4M 3.0 for pre-processing, statistical analysis and annotation, 40, has doubled since the previous release (Giacomoni et al., 2015), and now allows to analyze LC-MS, FIA-MS, GC-MS, and NMR data. Importantly, several tools implement original methods from the Core Team which provide unique features to the W4M workflows, such as the ***Biosigner*** tool for selection of significant molecular signatures for PLS-DA, Random Forest, or SVM classifiers, the ***NMR Annotation*** tool for annotation of NMR spectra, or the ***proFIA*** tool for the preprocessing of data from Flow Injection Analysis coupled to High-Resolution Mass Spectrometry (FIA-HRMS; Delabriere et al., *under review*). Furthermore, to cope with the computer intensive preprocessing of LC-MS data, the ***xcms.xcmsSet*** and ***CAMERA.annotate*** tools can now be run in parallel (see the supplementary material). All W4M tools are implemented by a large core team of bioinformaticians and biostatisticians based on five metabolomics facilities, and supported in the long term by two national infrastructures, namely the French Institute of Bioinformatics (IFB; French Elixir node) and the National Infrastructure for Metabolomics and Fluxomics (MetaboHUB). In addition to ensuring a sustainability and high-performance computing environment, these large clusters provide cutting-edge technologies, know-how, and scientific expertise from both the experimental and computational fields. In the near future, new unique tools will be available on W4M (e.g., to extend the NMR workflow and to analyze MS/MS data). Moreover, complementary Galaxy tools have been recently described in metabolomics (e.g., for Direct Infusion MS data; Davidson et al., 2016), but also in complementary omics communities (Boekel et al., 2015) such as proteomics (Jagtap et al., 2015; Jagtap et al., 2014; Fan et al., 2015). Due to the modularity of the Galaxy environment, and the relative ease of wrapping existing code into Galaxy tools, the number of W4M tools and contributors should continue to expand rapidly. To help developers integrating their tools, an updated virtual machine and the code of the Galaxy modules are publicly available on the W4M GitHub (https://github.com/workflow4metabolomics) and the Galaxy Toolshed (https://toolshed.g2.bx.psu.edu/; '[W4M]' tag) repositories.

Workflow4Metabolomics brings the workflow management features of the Galaxy environment to the metabolomics user community through its online infrastructure. Online availability has many advantages compared with local installation: no local computing resources are needed, local software installation and update is not required, and the infrastructure can be directly accessed from anywhere. Two online platforms have recently emerged for LC-MS processing and annotation (XCMS Online; Tautenhahn et al., 2012), and statistical analysis (MetaboAnalyst; Xia et al., 2009), respectively. In contrast, W4M provides for the first time a single resource for the comprehensive analysis of either LC-MS, FIA-MS, GC-MS, or NMR metabolomics data. In addition, by building on the Galaxy environment, W4M provides the user with unique features to build, run, and share workflows and histories (e.g., with remote collaborators in multi-center or transdisciplinary projects).

In particular, W4M 3.0 now offers to reference a history publicly, by assigning a unique ID and DOI permanent link, which can be cited in publications. To our knowledge, this is the first time that workflows (and associated data) can be referenced. Furthermore, since the source code of the W4M tools is also publicly available (as discussed above), referenced analyses can be fully dissected and reproduced online (or locally) by the scientific community. The W4M infrastructure thus fills a gap between existing repositories for raw data, such as MetaboLights (Haug et al., 2013) or the Metabolomics Workbench (Sud et al., 2016), and the spectral and metabolite databases such as KEGG (Kanehisa and Goto, 2000), HMDB (Wishart et al., 2007), ChEBI (Degtyarenko et al., 2008) or MassBank (Horai et al., 2010). Further interoperability between the W4M workflow resource and the MetaboLights data repository is ongoing within the PhenoMeNal European consortium. In the open data era (Leonelli et al., 2013), the need for peer-reproduced workflows (Gonzalez-Beltran et al., 2015) and workflow storage (Belhajjame et al., 2015) is pivotal for good science and robust transfer to the clinic (Baker, 2005). Besides, funding agencies and journal editors already require data to be made publicly available (the MetaboLights repository is already recommended by scientific journals such as Metabolomics, the EMBO Journal, and Nature Scientific Data; Kale et al., 2016). W4M should therefore become the reference repository for metabolomics workflows.

To help users cope with data analysis concepts, parameter tuning, and critical interpretation of diagnostics and results, training is of major importance (Via et al., 2013; Weber et al., 2015). On the W4M infrastructure, remote e-learning is possible through many tutorials (http://workflow4metabolomics.org/howto). In addition, the reference histories provide detailed examples of workflows, and of table and figure outputs (http://workflow4metabolomics.org/referenced_W4M_histories). Furthermore, “hands-on” sessions using W4M can be readily organized since only an internet connection is needed to access the infrastructure (for users wishing to use W4M for training, please contact us at contact@workflow4metabolomics.org). Based on our experience, optimal results are achieved when users analyze their own data. We therefore regularly organize one-week courses combining practical presentations in the mornings, and tutoring sessions in the afternoons (Workflow4Experimenters, W4E; http://workflow4metabolomics.org/events). Such trainings with about 25 participants offer unique opportunities to discuss the designs, methods, and tools for comprehensive and rigorous data analysis.

In conclusion, the Workflow4Metabolomics 3.0 e-infrastructure provides experimenters with unique features to learn, design, run, share, and reference comprehensive LC-MS, FIA-MS, GC-MS, and NMR metabolomics data analyses.