

SCIENTIFIC DATA

OPEN Data Descriptor: Computational workflow to study the seasonal variation of secondary metabolites in nine different bryophytes

Received: 25 January 2018

Accepted: 23 July 2018

Published: 28 August 2018

Kristian Peters¹, Karin Gorzolka¹, Helge Bruelheide^{2,3} & Steffen Neumann^{1,3}

In Eco-Metabolomics interactions are studied of non-model organisms in their natural environment and relations are made between biochemistry and ecological function. Current challenges when processing such metabolomics data involve complex experiment designs which are often carried out in large field campaigns involving multiple study factors, peak detection parameter settings, the high variation of metabolite profiles and the analysis of non-model species with scarcely characterised metabolomes. Here, we present a dataset generated from 108 samples of nine bryophyte species obtained in four seasons using an untargeted liquid chromatography coupled with mass spectrometry acquisition method (LC/MS). Using this dataset we address the current challenges when processing Eco-Metabolomics data. Here, we also present a reproducible and reusable computational workflow implemented in Galaxy focusing on standard formats, data import, technical validation, feature detection, diversity analysis and multivariate statistics. We expect that the representative dataset and the reusable processing pipeline will facilitate future studies in the research field of Eco-Metabolomics.

Design Type(s)	time series design • database creation objective • process-based data analysis objective
Measurement Type(s)	metabolite profiling
Technology Type(s)	Ultra High-performance Liquid Chromatography/Tandem Mass Spectrometry
Factor Type(s)	Spatial Orientation • wetness of soil • degree of illumination • substrate type • season • scan polarity
Sample Characteristic(s)	Fissidens taxifolius • shoot system • Polytrichum strictum • Hypnum cupressiforme • Grimmia pulvinata • Plagiomnium undulatum • Rhytidiadelphus squarrosus • Calliergonella cuspidata • Brachythecium rutabulum • Marchantia polymorpha

¹Leibniz Institute of Plant Biochemistry, Stress and Developmental Biology, Weinberg 3, 06120 Halle (Saale), Germany. ²Institute of Biology/Geobotany and Botanical Garden, Martin Luther University Halle Wittenberg, Am Kirchtort 1, 06108 Halle (Saale), Germany. ³German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Deutscher Platz 5e, 04103 Leipzig, Germany. Correspondence and requests for materials should be addressed to K.P. (email: kpeters@ipb-halle.de).

Background & Summary

In Ecological Metabolomics (or short “Eco-Metabolomics”), metabolite profiles of organisms are studied in order to describe ecological processes such as biotic interactions or the impact of environmental changes on various biological species^{1–3}. In contrast to biochemistry, wild non-model species are typically studied in their natural environment in ecology. This often involves different individuals of one or more species from populations growing under quite heterogeneous conditions when compared to the controlled conditions in greenhouses or growth chambers. As a result, metabolite profiles are highly variable when compared to each other. Moreover, profiles of non-model species contain a large number of novel compounds (so called “unknown unknowns”) that are difficult to identify because of lacking reference compounds, which have so far been mostly elucidated in model organisms^{3,4}. Furthermore, designing ecological experiments is often complex and involves multiple factors⁵. Thus, the metabolomics data processing pipeline needs to be adapted in order to deal with the particular hypotheses and idiosyncrasies of ecological experiments.

Here, we present a descriptor for a dataset that we consider representative for the research field of Eco-Metabolomics. Our study makes use of a field campaign with a two-factorial design (seasons and species), which includes (except *Marchantia polymorpha*) non-model species of bryophytes. In order to facilitate subsequent analysis, we kept the experiment design as simple as possible. The sampling was conducted on-site at the Botanical Garden of Martin Luther University Halle-Wittenberg once in each season over a period of one year (see below). Metabolite profiles were acquired using untargeted liquid chromatography coupled with mass spectrometry (LC/MS). Raw metabolite profiles are available in the metabolomics data repository MetaboLights⁶ (Data Citation 1).

In biochemistry there are strict laboratory protocols that ensure reproducibility of the analytical methods, while in bioinformatics this function is accomplished by implementing reusable computational workflows^{7,8}. Thus, in addition to the dataset we also address the typical bioinformatic challenges that come with Eco-Metabolomics experiments by implementing a reproducible and reusable computational workflow (Fig. 1). While the analysis and ecological interpretation of the study is described in Peters *et al.*⁹, here we focus on the analytical and bioinformatic work that is required to create a computational processing pipeline that is reproducible and that can be reused by other subsequent studies.

We describe in detail the experimental methodology that was used to create the dataset as well as the methodology to make the computational workflow reproducible (to give identical results in different computational environments). By formalizing and validating the processes that led to the results^{10,11}, we expect that this approach can serve as a model for subsequent studies. We further expect that Eco-Metabolomics studies use our dataset and the computational workflow to foster reuse and improve future data processing pipelines.

Methods

These methods describe in detail the steps in producing the data, including full descriptions of the experimental design in our related work⁹, data acquisition, computational processing, diversity analysis, biostatistics and bioinformatics procedures.

Sampling campaign

Samples of the nine moss species *Brachythecium rutabulum* (Hedw.) Schimp., *Calliergonella cuspidata* (Hedw.) Loeske, *Fissidens taxifolius* Hedw., *Grimmia pulvinata* (Hedw.) Sm., *Hypnum cupressiforme* Hedw. (*H. lacunosum* was not differentiated), *Marchantia polymorpha* L., *Plagiomnium undulatum* (Hedw.) T.J. Kop., *Polytrichum strictum* Menzies ex Brid. and *Rhytidiadelphus squarrosus* (Hedw.) Warnst. were collected in the Botanical Gardens of the Martin-Luther-University Halle-Wittenberg, Germany. Sampling was performed in summer (2016/08/08), autumn (2016/11/09), winter (2017/01/27) and spring (2017/05/11) at relatively stable weather conditions as it is known that short-term climatic fluctuations and rainfall can influence secondary metabolite content and ammonium uptake of bryophytes¹². Thus, the bryophytes were only collected when there was sunshine at least two days prior to and during sampling. Furthermore, sampling was performed after mid-day between 13:00 and 15:00.

Sampling protocol

In each season, three composite samples of different individuals of each species were taken, leading to a total of $3 \times 9 \times 4 = 108$ samples. Only above-ground parts of the moss gametophytes such as leaves, branches, stems or thalloid parts were taken for sampling. From dioecious species such as *M. polymorpha*, *P. strictum* and *P. undulatum* female, male and sterile gametophytes were collected in a composite sample. Before sampling, visible archegonial and antheridial heads and any belowground parts such as rhizoids and rooting stems were removed with a sterile tweezer. The gametophytic moss parts were put in Eppendorf tubes and were frozen instantly on dry ice and later in the lab in liquid nitrogen.

Collecting ecological characteristics

In order to relate metabolomes of the bryophytes to ecology, several ecological characteristics were recorded on-site and compiled from literature. The on-site characteristics *type of substrate* with the nominal/categorical levels “soil”, “rock with lean soil cover” and “rock”; *light conditions* with the ordinal

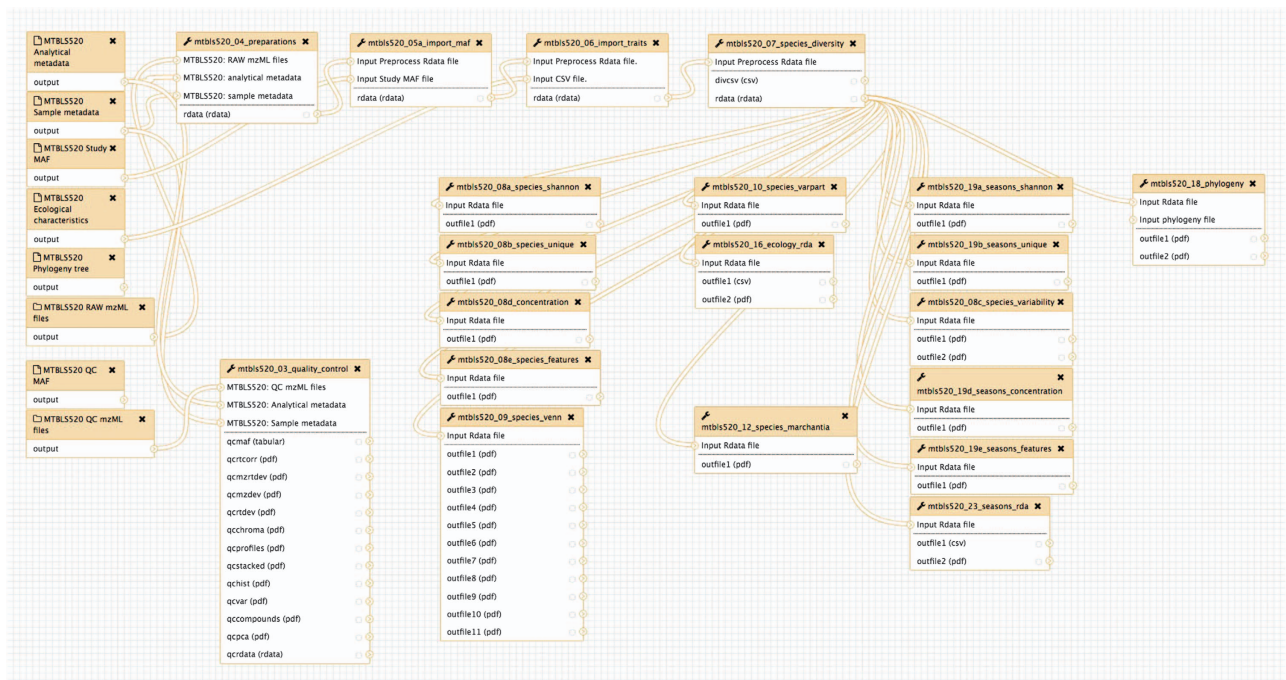


Figure 1. Computational workflow of the whole study (Data Citation 1) running in the Galaxy Workflow Management system. Each of the modules represent a particular step in the study of Peters *et al.*⁹. The modules have defined inputs, outputs and sets of parameters. The modules are connected to each other to give the resulting workflow. The function of the modules is explained in Table 1 (available online only).

levels “sunny”, “half-shade” and “shade”; *moisture of the substrate* with the ordinal levels “dry”, “fresh”, “damp” and wet; and *exposition* with the nominal levels “North”, “East”, “South”, “West”, “Northeast”, “Northwest”, “Southeast” and “Southwest” were recorded when taking the samples in the field.

The nominal characteristics *growth form*, *habitat type*, *substrate* and *life strategy*, the ordinal life-history characteristics *spore size*, *gametangia distribution* and *sexual reproduction frequency*, as well as the ordinal Ellenberg indicator values (indices for *light*, *temperature*, *continentality*, *moisture*, *reaction*, *nitrogen* and *life-form*) were collected from the literature^{13–17}. For an overview, please refer Table 1 (available online only) in Peters *et al.*⁹ or the file *m_characteristics.csv* in the dataset (see Data Citation 1, and Table 1 (available online only)).

Extraction protocol and LC/MS analysis

Frozen moss samples were homogenized by adding 200 mg ceramic beads (0.5 mm diameter, Roth) and ribolysing (Precellys 24, 2 × 20 s at 6500 r.p.m., 5 min pause in liquid nitrogen). 1 ml ice-cold 80/20 (v/v) methanol/water spiked with internal standards 5 μM biochanin A (Sigma-Aldrich), 5 μM kinetin (Sigma-Aldrich) and 5 μM N-(3-indolylacetyl)-l-valine (Sigma-Aldrich) were added. Samples were vortexed and thawed while shaking for 15 min at 1,000 r.p.m. at room temperature followed by ultrasonication for 15 min and again 15 min shaking. After 15 min centrifugation at 13,000 r.p.m. 500 μl of supernatant were dried in a vacuum centrifuge at 40 °C and reconstituted in 80/20 (v/v) methanol/water with the volume adjusted to the initial fresh weight of the sample to a final concentration of 10 mg fresh weight per 100 μl extract.

Chromatographic separations were performed at 40 °C on an Acquity UPLC system (Waters) equipped with an HSS T3 column (100 × 1 mm, particle size 1.8 μm; Waters) applying the following binary gradient at a flow rate of 150 μL min⁻¹: 0 to 1 min, isocratic 95% A (water:formic acid: 99.9:0.1 [v/v]), 5% B (acetonitrile:formic acid: 99.9:0.1 [v/v]); 1 to 18 min, linear from 5 to 95% B; 18 to 20 min, isocratic 95% B. The injection volume was 2.0 μL (full loop injection).

Ultra-performance liquid chromatography coupled to electrospray ionization quadrupole time-of-flight mass spectrometry (UPLC/ESI-QTOF-MS) was performed using a high resolution MicroTOF-Q II hybrid quadrupole time-of-flight mass spectrometer¹⁸. Data were acquired with the following MS instrument settings: nebulizer gas: nitrogen, 1.4 bar; dry gas: nitrogen, 6 L min⁻¹, 190 °C; capillary: 5000 V (+4000 V for negative mode); end plate offset: -500 V; funnel 1 radio frequency (RF): 200 Volts peak-to-peak (Vpp); funnel 2 RF: 200 Vpp; in-source collision-induced dissociation (CID) energy: 10 eV; hexapole RF: 100 Vpp; quadrupole ion energy: 3 eV (-5 eV for neg-mode); collision gas: nitrogen; collision energy: 7 eV (-7 eV for negative mode); collision cell RF: 250 Vpp (150 Vpp for negative mode); transfer time: 70 μs; prepulse storage: 5 μs; pulser frequency: 10 kHz; and spectra rate: 3 Hz. Mass spectra

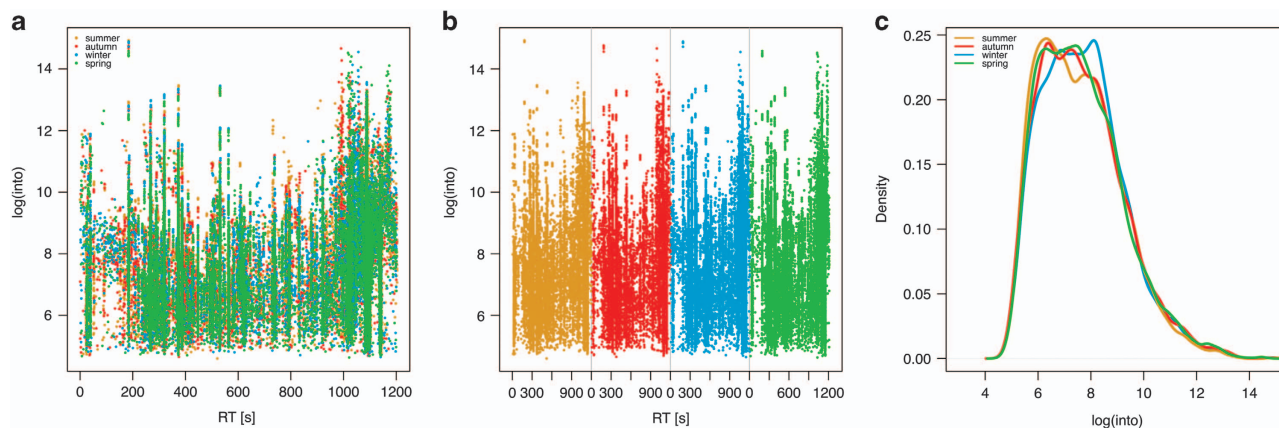


Figure 2. Plots of sets of the MM8 profiles to assess the performance of the technical setup. Green = spring, yellow = summer, red = autumn, blue = winter. $n = 28$. (a) Plot of the four sets of MM8 profiles against each other. X axis: Retention time [s]. Y axis: Logarithmic total ion current. (b) Stacked plot of the sets of MM8 profiles next to each other. X axis: Retention time [s]. Y axis: Logarithmic total ion current. (c) Density plot (histogram) of log intensities of the sets of MM8 profiles. X axis: Sample size. Y axis: Estimated kernel density.

were acquired in centroid mode. Calibration of the m/z scale was performed for individual raw data files on lithium formate cluster ions obtained by automatic infusion of 20 μL of 10 mM lithium hydroxide in isopropanol:water:formic acid, 49.9:49.9:0.2 (v/v/v) at the end of the gradient.

Quality control

In order to validate the instrument performance and to detect batch effects between the instrument runs, the following quality control (QC) protocol was realized. Samples with a lab-internal standard mix (MM8) were interspersed before and after 7 bryophyte samples in the MicrOTOF¹⁸. The following substances were used in the MM8: 2-Phenylglycine (Fluka), Kinetin (Roth), Rutin (Acros Organics), O-Methylsalicylic acid (Sigma), Phlorizin dihydrate (Sigma), N-(3-Indolylacetyl)-L-valine (Sigma), 3-Indolylacetonitrile (Fluka) and Biochanin A (Sigma). Substances in the MM8 were selected based on their ionization properties (ionization in both positive and negative mode and the differential adduct formation) and a wide coverage of known retention times throughout the gradient with our instrumental setup. Known ionization properties were used to detect shifts and effects in mass-to-charge ratios (m/z) and retention times (RT) of the respective batches and to validate RT correction made by XCMS (see below).

Raw data acquisition

Raw LC/MS data were converted to the open data format mzML¹⁹ with the software CompassXPort 3.0.9 from Bruker Daltonics (available at <http://www.bruker.com/service/support-upgrades/software-downloads.html>). In compliance with the minimum information guidelines for Metabolomics studies²⁰, metadata were recorded to ISA-Tab format²¹ using ISAcreator 1.7.10 (ref. 22) (available at <https://github.com/ISA-tools/ISAcreator/releases>) and uploaded together with the raw data to the metabolomics repository MetaboLights⁶ (Data Citation 1). Profiles of positive mode were used for the data analyses as many important and known secondary metabolites classes in bryophytes such as flavonoids, phenylpropanoids, anthocyanins, glycosides and previously characterized compounds such as Marchantins, Communins and Ohioensins ionize well in positive mode with our instrumental setup.

Peak detection

Chromatographic peak picking was performed in R 3.4.2 (available at <https://cran.r-project.org>) with the package XCMS 1.52.0 (ref. 23) using the centWave algorithm and the following parameters: ppm = 35, peakwidth = 4,21, snthresh = 10, prefilter = 5–50, fitgauss = TRUE, verbose.columns = TRUE. Grouping of chromatographic peaks was performed with two factors (in XCMS called “phenoData”): *seasons* with the levels summer, autumn, winter and spring; and *species* with the levels Brarut, Calcus, Fistax, Gripul, Hypcup, Marpol, Plaund, Polstr and Rhysqu. The following parameters were used for grouping: mzwid = 0.01, minfrac = 0.5, bw = 4. To improve subsequent data analyses, intensities in the peak table were log transformed before grouping. For further analysis, only features between the retention times 20 s and 1020 s were kept. Retention time correction was performed using the function retcor in XCMS using the parameters method = loess, family = gaussian, missing = 10, extra = 1, span = 2. The parameters were additionally optimized using the R package IPO 1.3.3 (ref. 24), but better alignment precision was achieved with manual control and knowledge of instrument settings²⁵.

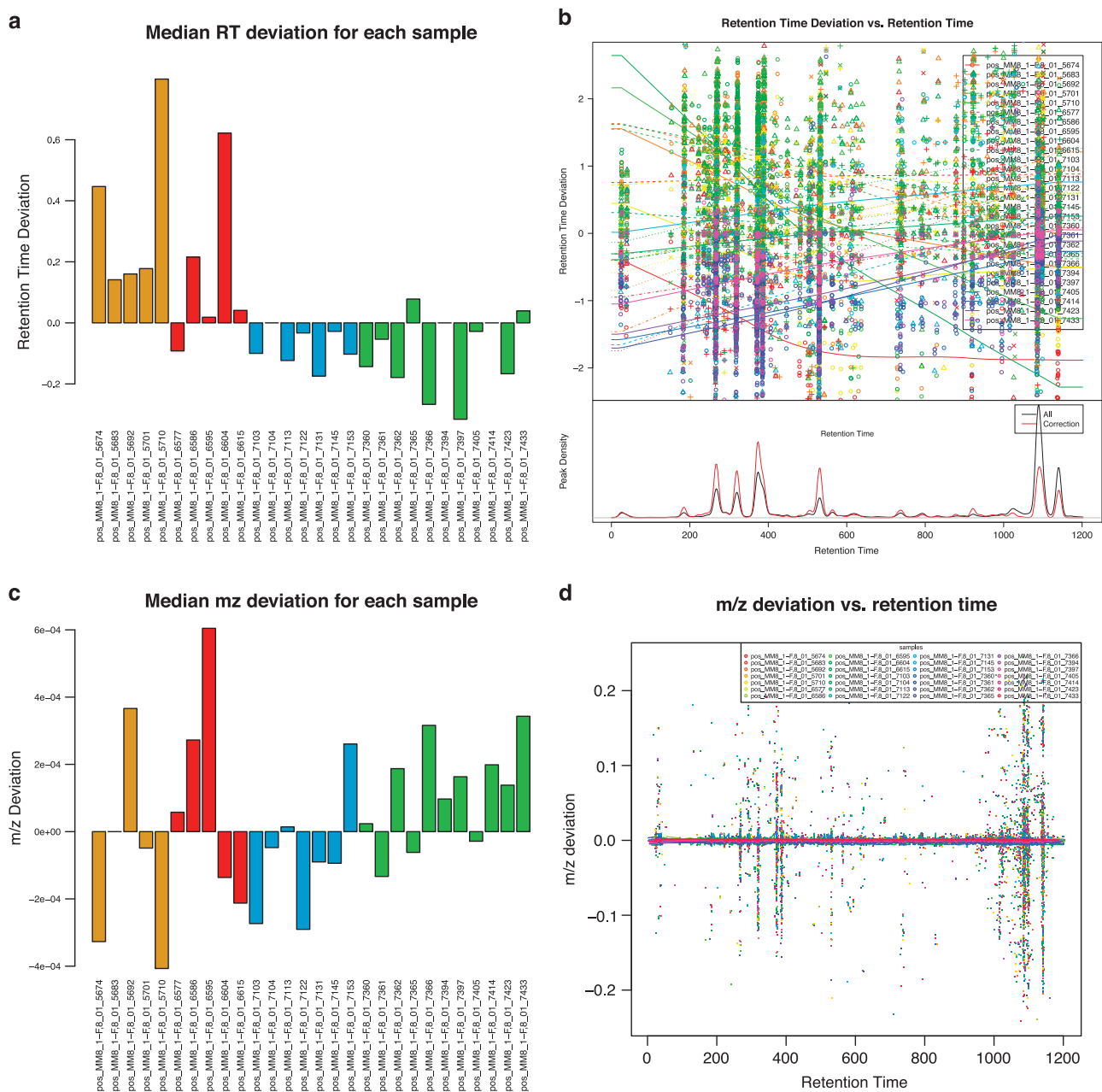


Figure 3. Quality control plots to assess shifts in retention time (RT) and mass-to-charge ratio (m/z) in the four sets of MM8 profiles. Green = spring, yellow = summer, red = autumn, blue = winter. $n = 28$. (a) Median retention time deviation for the sets of MM8 profiles. X axis: Name of MM8 profile. Y axis: Retention time deviation [s]. (b) Retention time deviation plotted against retention time. X axis: Retention time [s]. Y axis: Retention time deviation [s] per profile. (c) Median mass-to-charge deviation for each profile. X axis: MM8 profile. Y axis: m/z deviation. (d) Mass-to-charge deviation plotted against retention time. X axis: Retention time [s]. Y axis: m/z deviation per profile.

Peak annotation

Adduct annotation was performed with the R package CAMERA 1.33.3 (ref. 26) by using the following functions: `xsAnnotate`, `groupFWHM`, `findIsotopes`, `groupCorr`, `findAdducts`; with the following parameters: `perfwhm = 0.6`, `ppm = 5`, `mzabs = 0.005`, `calcIso = TRUE`, `calcCiS = TRUE`, `calcCaS = TRUE`, `graphMethod = lpc`, `pval = 0.05`, `cor_eic_th = 0.75`. In order to improve subsequent statistical analyses instead of the CAMERA function `getPeaklist` the function `getReducedPeaklist` was written that aggregates the adducts of putative compounds into a feature list with singular components (see pull request in GitHub: <https://github.com/sneumann/CAMERA/pull/16>). Since version 1.33.3 the function `getReducedPeaklist` is officially part of CAMERA. The parameter `method = median` was chosen for the study.

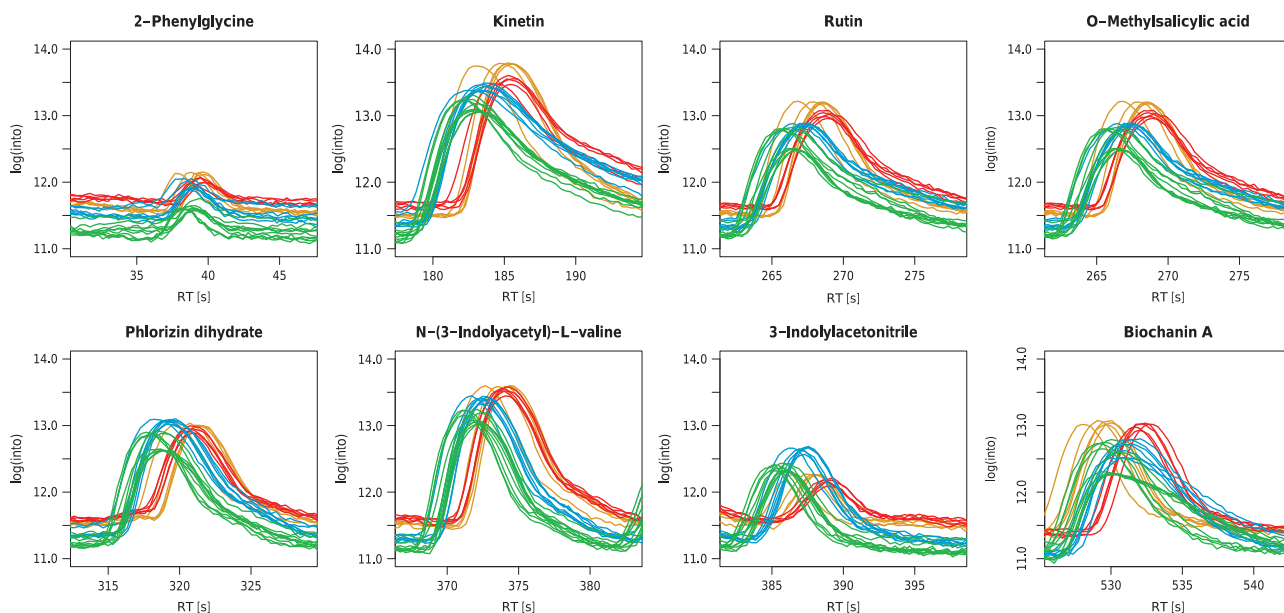


Figure 4. The eight compounds used for the internal lab standard mix (MM8) plotted next to each other. Shown are the regions of the respective compounds in the raw chromatograms before the alignment of XCMS. Green = spring, yellow = summer, red = autumn, blue = winter. X axis: Retention time [s]. Y axis: Logarithmic total ion current. $n = 28$.

Exemplary compound annotation

Compounds were putatively annotated for the follow-up validation and biochemical interpretation with the software Bruker Compass IsotopePattern 4.4. Annotation was performed by calculating accurate masses (mass-to-charge values) from known compounds in *M. polymorpha* and other liverworts found in PubChem, the KNApSACk database and Asakawa *et al.*^{27,28}. In the software Bruker Compass DataAnalysis 4.4 the mass-to-charge was matched to device-specific retention times in the metabolite profile. To validate whether the known compound was present in the profile, Extracted Ion Chromatograms (EIC) and area-under-curve (integrated intensities) were checked manually.

Diversity analysis

Statistical analyses were performed using the additional R packages: multtest, RColorBrewer, vegan, multcomp, multtest, nlme, ape, pvclust, dendextend, phangorn, Hmisc, gplots and VennDiagram. A presence-absence matrix was generated from the feature matrix to determine the differences in metabolite features between the experimental factors species and season. In accordance with the minfrac parameter in the alignment step in XCMS (see above), a feature was considered present when it was detected at least in two out of three replicates. The presence-absence matrix was used for measuring the metabolite richness for each species and season by calculating the Shannon diversity index (H') for each sample i using the function diversity in vegan with the parameter index = shannon²⁹. The following equation was used for calculation:

$$H' = \sum_{i=1}^t p_i \ln(p_i)$$

where t represents the number of samples in the particular group.

The total number of features and the number of unique features were calculated from the presence-absence matrix accordingly. To test factor levels for significant differences, the Tukey HSD on a one-way ANOVA was performed post-hoc using the multcomp package.

Variability was calculated with the Pearson Correlation Coefficient (PCC, Pearson's r) using the function rcorr in the package Hmisc. Venn diagrams were created for each species separately using the package VennDiagram. Each set in the Venn diagram represents one season and shows distinct and shared features in all possible combinations between the sets.

Multivariate statistical analysis

Variation partitioning was performed using the function varpart in the package vegan to analyze the influence of the factors species and seasons on the metabolite profiles. Distance-based redundancy analysis (dbRDA) using the function capscale with Bray-Curtis distance and multidimensional scaling in

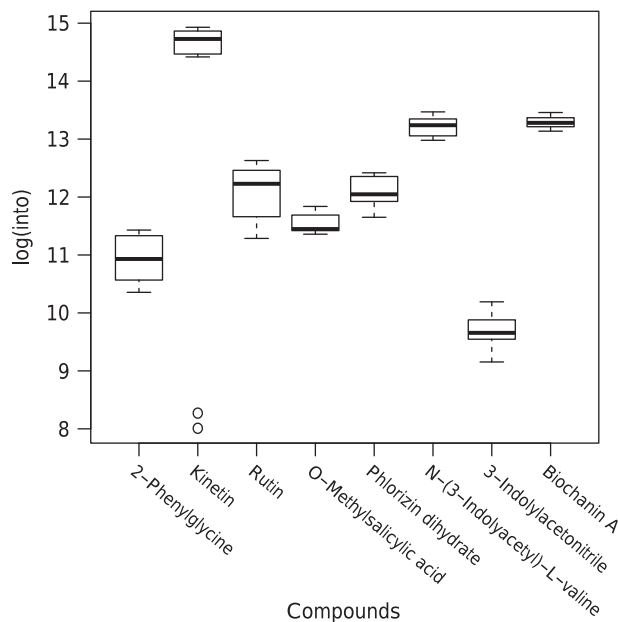


Figure 5. Boxplots of the variation in the intensities of the eight compounds used in the internal lab standard mix of all the MM8 profiles. X axis: Compound. Y axis: Logarithmic total ion current. $n = 28$ for each box.

the package `vegan` was chosen to analyse the relation of the ecological characteristics with the species metabolite profiles^{30,31}. Ordinal and categorical ecological characteristics were transformed to presence-absence matrices for the ordination. The optimal model for the dbRDA was chosen with forward and backward selection using the function `ordstep` in the package `vegan`. Ecological characteristics were added to the plots as post-hoc variables using the function `envfit` in the package `vegan`.

Chemotaxonomic comparison to phylogeny

Relationships between metabolite profiles and phylogeny were analysed by calculating dissimilarities for phylogeny and the feature matrix using Bray-Curtis distance (function `vegdist` in `vegan`) followed by hierarchical clustering using the function `hclust` and the complete linkage method. In order to improve the visual comparison between the two trees, the chemotaxonomic plot was reordered using the function `order.optimal` (package `cba`) and leaves of Polstr and Plaund were swapped using the function `reorder` in `vegan`. The similarity of the two trees was determined with the normalized Robinson-Foulds metric (function `RF.dist` in package `phangorn`). The similarity of the distance matrices was determined with the Mantel statistics (function `mantel` in `vegan`).

Computational workflow

For the computational workflow, the required software tools, their dependencies, as well as software libraries and R packages were containerized using Docker technology³². The container was based on Linux and Ubuntu 16.04 and included R version 3.4.2 from the R apt repository. The commands for building the container can be found in the Dockerfile (Table 1 (available online only)). The resulting container image was made available at DockerHub (<https://hub.docker.com/r/korseby/mtbls520/>).

The computational workflow was constructed with the Galaxy workflow management system³³. It consists of 20 modules and each individual module represents one or more dedicated steps in the Peters *et al.* study⁹, e.g. data retrieval, feature detection, alignment or statistical analysis (Fig. 1). For the workflow, individual Galaxy modules were written in XML format. Each Galaxy module executes a shell or R script with defined inputs and outputs. Scripts are only executed inside the software container. Thus, code execution is encapsulated and all required software dependencies were resolved in the software container. In order to comply with the *Interoperability* criterion in the FAIR guidelines³⁴, the PhenoMeNal cloud e-infrastructure was used to test the workflow in different computational environments (<https://phenomenal-h2020.eu>). To ensure that the workflow generates the same results in different computational environments, continuous automatic workflow testing was implemented with `wft4galaxy`³⁵.

Data Records

The primary access site for the dataset is MetaboLights (Data Citation 1), which includes the 108 metabolite profiles of the bryophytes in positive and negative mode, QC profiles, ecological data and meta-data (see Table 2 (available online only) for an overview of sample names and associated factor

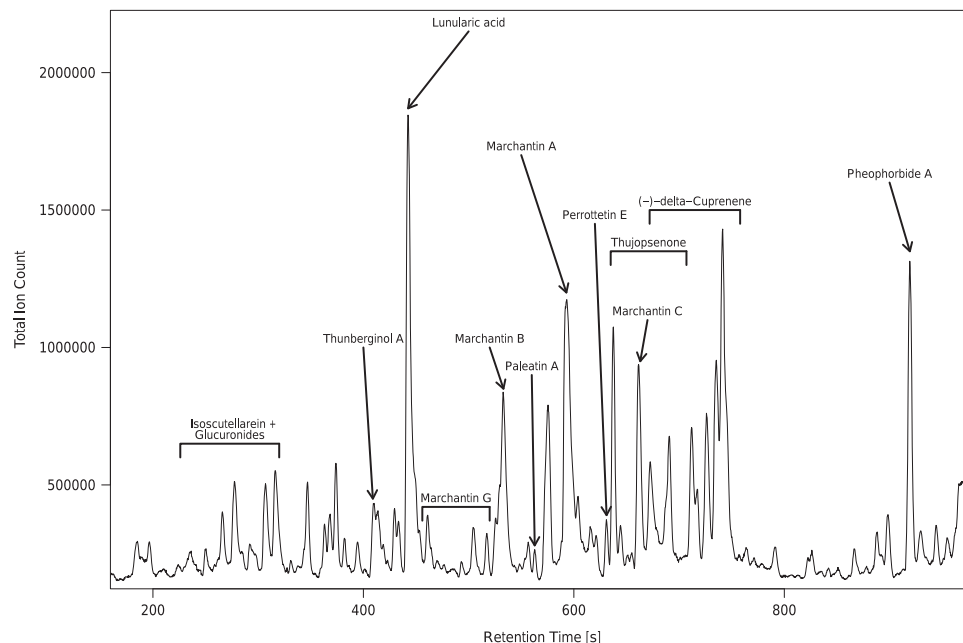


Figure 6. Total Ion Count (TIC) chromatogram obtained from the extracts of *Marchantia polymorpha*.

This exemplary chromatogram was obtained from the third sample of summer. Values of the retention times (RT values), accurate masses and sum formulas are available in Table 3 (available online only).

levels). Table 1 (available online only) provides an overview of data files, formats and functions in the computational workflow.

Code Availability

The source code (also deposited at <https://github.com/korseby/container-mtbls520/>) was published³⁶ and made available under the terms of APACHE license 2.0. Please refer Table 1 (available online only) for an overview of the function of each file of the source code.

Code for building the software container image and the workflow including Galaxy modules and scripts that are executed inside the container were published under Open Access³⁶. A pre-built binary software container image was made available at DockerHub (retrievable at <https://hub.docker.com/r/korseby/mtbls520/>).

Technical Validation

Quality control

Four sets of 27 bryophyte samples were generated in the experiment. One set for each season was analyzed with UPLC/ESI-QTOF-MS (see methods below) which resulted in a total of 108 bryophyte metabolite profiles. In order to validate the instrument performance and to detect batch effects between the four instrument runs, a quality control (QC) protocol was implemented. Sets of 27 species samples were interspersed by samples of a lab-internal standard mix (MM8) before and after 7 bryophyte samples. Peak detection in these MM8 profiles was performed with the identical parameters as for the bryophyte samples.

The four sets containing the MM8 metabolite profiles were checked visually for differences by plotting them against each other (Fig. 2a) and stacked next to each other (Fig. 2b). The density distribution of the intensities within the sets of MM8 profiles were also checked and compared to each other with a density plot (histogram) (Fig. 2c).

Mass-to-charge ratio and retention time deviation (in seconds) and correction made by XCMS were checked with diagnostic plots made by XCMS (Fig. 3). We found maximum retention time deviations within 2 s (Fig. 3a and b) which are in the expected range of the analytical setup¹⁸. The determined mass-to-charge deviations (Fig. 3c and d) are within instrument specification as well¹⁸.

The variation in the intensities of the internal lab standards was also checked for each reference compound individually as shown in Figs. 4 and 5. In general, the variation for each reference compound and the deviations between MM8 profiles are both well within the typical range of 10 to 15% (ref. 18).

We conclude that there are no significant batch effects in the technical replicates to overlap with the factor *seasons* of the experiment. Thus, the automatic retention time correction made by XCMS is validated for the parameters used in the peak detection process.

Exemplary annotation of *Marchantia polymorpha* profile

With known accurate masses (m/z values) and calculated retention time values (see methods), we confirm the annotation of many known compounds which are described in literature for the model species *Marchantia polymorpha*^{27,28} (Fig. 6). Many of these known compounds also constitute the most abundant features in the profile of *M. polymorpha* (Fig. 6).

Computational workflow

We have implemented the computational workflow in the Galaxy workflow management system³³ and have made the workflow and underlying code available as Open Source³⁶. The Galaxy workflow represents the entire computational processing pipeline that is used in the Peters *et al.* study⁹ (Fig. 1). Each of the individual modules represents a particular step in the workflow and has defined inputs (e.g. pre-processed peak table data matrix) and outputs (e.g. PDF containing the plot of a particular statistical method) (Fig. 1). We used data standards and minimum information criteria for constructing the modules of the workflow^{20,22}. Continuous automatic testing of the workflow was performed with `wft4galaxy`³⁵ in the PhenoMeNal e-infrastructure (<https://phenomenal-h2020.eu>) to ensure that the workflow generates the same results in different computational environments.

We proceeded according to the FAIR guiding principles³⁴ in order to implement a reusable computational workflow. The acronym FAIR stands for Findable, Accessible, Interoperable and Reusable and encompasses several criteria to support the reuse of scholarly data. So far, the FAIR guidelines have only been aspired to make data reusable. However, as the conceptual formulation within FAIR are quite generalized³⁷, these principles can also be applied to computational workflows. Nonetheless, there are some computational challenges involved. For example, software runs in different software environments and software dependencies need to be resolved. We tackle this by creating software containers which can be run on multiple systems and contain the software tools, all required libraries and R packages^{32,38}. As dependencies in the container have already been resolved, sharing the container image greatly facilitates to allow the software to be run in multiple environments.

We have chosen the Galaxy Workflow Management system^{33,39} to implement the whole data processing pipeline (Fig. 1) as it is already known to facilitate reproducible results⁴⁰. Several processing modules were constructed that represent the individual steps of the Peters *et al.* study⁹. Software tools are invoked from the Galaxy modules and are executed inside the container, thus, adding a level of encapsulation and eliminating the need for the user to install additional software⁴¹. Galaxy has a graphical user interface that hides the technical complexity from the end user and does not need intensive bioinformatic background knowledge to run the particular modules and workflows. This greatly contributes to the adoption by the end users (biochemists and ecologists) and facilitates future studies in the research field of Eco-Metabolomics.

Statistical analyses

With untargeted metabolomics analysis in ecology, diversity analysis is typically used to characterize the richness and the abundance of biochemical features in the metabolite profiles of biological species⁴². Metabolite richness is a simple measure that counts the individual biochemical features in the metabolite profiles of the species⁴³. The abundance of features in the metabolite profiles is usually calculated by diversity indices such as the Shannon diversity index (H') in order to characterize simple relationships with regard to the study factors⁴⁴.

Ordination methods such as Redundancy Analysis (RDA) and distance-based Redundancy Analysis (dbRDA) are frequently used in Ecology³⁰. They allow to derive correlations of specific variables between the matrix of predictors containing the measurements (X matrix) and the response matrix with the ecological traits (Y matrix)^{30,45}. These methods are also suitable for Eco-Metabolomics data as they allow the use of multiple (non-categorical) variables in a single model and allow to calculate the amount of explained variance of the model. We have chosen the dbRDA, which can also be regarded as a constrained version of metric scaling (MDS)^{46,47}. We have implemented dedicated modules for these statistical operations in our computational workflow (see Methods section and Fig. 1).

Usage Notes

Building the container image

Following are instructions to manually build the container image. The file Dockerfile in Table 1 (available online only) contains the ruleset. The container has been built using Docker version 17.05-ce under Linux Ubuntu 16.04. The following commands were run to generate the image:

```
sudo apt-get install apt-transport-https ca-certificates git
sudo echo deb http://apt.dockerproject.org/repo ubuntu-xenial main
>>/etc/apt/sources.list
sudo apt-key adv --keyserver hkp://ha.pool.sks-keyservers.net:80
--recv-keys 58118E89F3A912897C070ADBF76221572C52609D
sudo apt-get update && sudo apt-get install docker
git clone https://github.com/korseby/container-mtbls520
cd container-mtbls520
docker build -t korseby/mtbls520.
```

Installing and using Galaxy to run the workflow

The workflow was tested with Galaxy version 17.09. Instructions how to install Galaxy can be found in the training material of the Galaxy project (accessible at <https://galaxyproject.github.io/training-material/>). However, it is recommended that an official Galaxy server is used, such as those from the PhenoMeNal infrastructure (available at <https://public.phenomenal-h2020.eu/>).

After being logged into Galaxy, a click on “Workflow” in the menu bar on the top and then a click on the “Upload” button opens up a new page. In the field “Galaxy workflow URL:” enter the following address “https://raw.githubusercontent.com/korseby/container-mtbls520/develop/galaxy/mtbls520_workflow.ga” or upload the .ga file from the GitHub repository (Table 1 (available online only)) and then clicking on the button “Import”. This will import the workflow of the study into Galaxy. The workflow will now be available in Galaxy under Workflows as “Metabolights 520 Eco-Metabolomics Workflow”. From there, clicking on the drop-down menu there are options to “Edit” (visually view the complete workflow in the Galaxy workflow editor) or to “Run” the workflow. Required data can be downloaded from MetaboLights with the Galaxy module “mtbls520_01_mtbls_download” (Table 1 (available online only)). Once the download has been completed, data can be extracted with the Galaxy module “mtbls520_02_extract” (Table 1 (available online only)). The workflow can be directly run once the inputs have been assigned to the extracted data files. Processing will take approx. 40 min depending on the work load of the computational infrastructure.

References

- Sardans, J., Peñuelas, J. & Rivas-Ubach, A. Ecological metabolomics: overview of current developments and future challenges. *Chemoecology* **21**, 191–225 (2011).
- Jones, O. A. H. *et al.* Metabolomics and its use in ecology: Metabolomics in Ecology. *Austral Ecol.* **38**, 713–720 (2013).
- Peters, K. *et al.* Current Challenges in Plant Eco-Metabolomics. *Int. J. Mol. Sci.* **19**, 1385 (2018).
- van Dam, N. M. & van der Meijden, E. A Role for Metabolomics in Plant Ecology in *Annual Plant Reviews Volume 43* (ed. Hall, R. D.) 87–107 (Wiley-Blackwell, 2011).
- Rivas-Ubach, A. *et al.* Are the metabolomic responses to folivory of closely related plant species linked to macroevolutionary and plant-folivore coevolutionary processes? *Ecol. Evol.* **6**, 4372–4386 (2016).
- Haug, K. *et al.* MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res.* **41**, D781–D786 (2013).
- Gil, Y. *et al.* Examining the Challenges of Scientific Workflows. *Computer* **40**, 24–32 (2007).
- Peng, R. D. Reproducible Research in Computational Science. *Science* **334**, 1226–1227 (2011).
- Peters, K., Gorzolja, K., Bruelheide, H. & Neumann, S. Seasonal variation of secondary metabolites in nine different bryophytes. *Ecol. Evol.* <https://doi.org/10.1002/ece3.4361> (2018).
- Sandve, G. K., Nekrutenko, A., Taylor, J. & Hovig, E. Ten Simple Rules for Reproducible Computational Research. *PLoS Comput. Biol.* **9**, e1003285 (2013).
- Leipzig, J. A review of bioinformatic pipeline frameworks. *Brief. Bioinform.* **18**, 530–536 (2017).
- Cornelissen, J. H. C., Lang, S. I., Soudzilovskaia, N. A. & Doring, H. J. Comparative Cryptogam Ecology: A Review of Bryophyte and Lichen Traits that Drive Biogeochemistry. *Ann. Bot.* **99**, 987–1001 (2007).
- Urmi, E. Bryophyta (Moose) in Flora indicativa *Ecological Indicator Values and Biological Attributes of the Flora of Switzerland and the Alps* 283–310 (Haupt, 2010).
- Doring, H. J. Ecological classification of bryophytes and lichens in *Bryophytes and lichens in a changing environment* 1–31 (Clarendon Press, 1992).
- Frisvoll, A. A. Bryophytes of Spruce Forest Stands in Central Norway. *Lindbergia* **22**, 83–97 (1997).
- Smith, A. J. E. *The liverworts of Britain and Ireland* (Cambridge University Press, 1990).
- Smith, A. J. E. *The Moss Flora of Britain and Ireland* (Cambridge University Press, 2004).
- Böttcher, C. *et al.* The Multifunctional Enzyme CYP71B15 (PHYTOALEXIN DEFICIENT3) Converts Cysteine-Indole-3-Acetonitrile to Camalexin in the Indole-3-Acetonitrile Metabolic Network of *Arabidopsis thaliana*. *Plant Cell.* **21**, 1830–1845 (2009).
- Martens, L. *et al.* mzML—a Community Standard for Mass Spectrometry Data. *Mol. Cell. Proteomics* **10**(R110): 000133 (2011).
- Spicer, R. A., Salek, R. & Steinbeck, C. Compliance with minimum information guidelines in public metabolomics repositories. *Sci. Data* **4**, 170137 (2017).
- Sansone, S.-A. *et al.* Toward interoperable bioscience data. *Nat. Genet.* **44**, 121–126 (2012).
- Rocca-Serra, P. *et al.* ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics* **26**, 2354–2356 (2010).
- Tautenhahn, R., Böttcher, C. & Neumann, S. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics* **9**, 504 (2008).
- Libiseller, G. *et al.* IPO: a tool for automated optimization of XCMS parameters. *BMC Bioinformatics* **16**, 118 (2015).
- Mönchgesang, S. *et al.* Natural variation of root exudates in *Arabidopsis thaliana*-linking metabolomic and genomic data. *Sci. Rep.* **6**, 29033 (2016).
- Kuhl, C., Tautenhahn, R., Böttcher, C., Larson, T. R. & Neumann, S. CAMERA: An Integrated Strategy for Compound Spectra Extraction and Annotation of Liquid Chromatography/Mass Spectrometry Data Sets. *Anal. Chem.* **84**, 283–289 (2012).
- Nakamura, Y. *et al.* KNApSAcK Metabolite Activity Database for Retrieving the Relationships Between Metabolites and Biological Activities. *Plant Cell Physiol.* **55**, e7 (2014).
- Asakawa, Y. *et al.* *Chemical constituents of bryophytes: bio- and chemical diversity, biological activity, and chemosystematics* (Springer Verlag, 2013).
- Li, D., Heiling, S., Baldwin, I. T. & Gaquerel, E. Illuminating a plant’s tissue-specific metabolic diversity using computational metabolomics and information theory. *Proc. Natl. Acad. Sci.* **113**, E7610–E7618 (2016).
- Legendre, P. & Legendre, L. *Numerical ecology* Volume 243rd edn, (Elsevier, 2012).
- Legendre, P. & Anderson, M. J. Distance-based Redundancy Analysis: Testing Multispecies Responses In Multifactorial Ecological Experiments. *Ecol. Monogr.* **69**, 24 (1999).
- Miksa, T., Rauber, A. & Mina, E. Identifying impact of software dependencies on replicability of biomedical workflows. *J. Biomed. Inform.* **64**, 232–254 (2016).
- Afgan, E. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.* **44**, W3–W10 (2016).

34. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
35. Piras, M. E., Pireddu, L. & Zanetti, G. wft4galaxy: a workflow testing tool for galaxy. *Bioinformatics* **33**, 3805–3807 (2017).
36. Peters, K., Gorzolja, K., Bruelheide, H. & Neumann, S. Code for the computational workflow to study the seasonal variation of secondary metabolites in nine different bryophytes using the MTBLS520 dataset (Version v1.1). Zenodo <https://doi.org/10.5281/zenodo.1284246> (2018).
37. Dunning, A. C., De Smaele, M. M. E. & Böhmer, J. K. Evaluation of data repositories based on the FAIR Principles for IDCC 2017 practice paper TU Delft <https://doi.org/10.4121/uuid:5146dd06-98e4-426c-9ae5-dc8fa65c549f> (2017).
38. Merkel, D. Docker: Lightweight Linux Containers for Consistent Development and Deployment. *Linux Journal* <https://www.linuxjournal.com/content/docker-lightweight-linux-containers-consistent-development-and-deployment> (2014).
39. 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.* **11**, R86 (2010).
40. Giacomoni, F. *et al.* Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics. *Bioinformatics* **31**, 1493–1495 (2015).
41. Piccolo, S. R. & Frampton, M. B. Tools and techniques for computational reproducibility. *GigaScience* **5**, 30 (2016).
42. Richards, L. A. *et al.* Phytochemical diversity drives plant–insect community diversity. *Proc. Natl. Acad. Sci.* **112**, 10973–10978 (2015).
43. Ristok, C. *et al.* Leaf litter diversity positively affects the decomposition of plant polyphenols. *Plant Soil* **419**, 305–317 (2017).
44. Tewes, L. J., Michling, F., Koch, M. A. & Müller, C. Intracontinental plant invader shows matching genetic and chemical profiles and might benefit from high defence variation within populations. *J. Ecol.* **106**, 714–726 (2018).
45. von Wehrden, H., Hanspach, J., Bruelheide, H. & Wesche, K. Pluralism and diversity: trends in the use and application of ordination methods 1990–2007. *J. Veg. Sci.* **20**, 695–705 (2009).
46. Field, K. J. & Lake, J. A. Environmental metabolomics links genotype to phenotype and predicts genotype abundance in wild plant populations. *Physiol. Plant.* **142**, 352–360 (2011).
47. Zuppinger-Dingley, D. *et al.* Selection for niche differentiation in plant communities increases biodiversity effects. *Nature* **515**, 108–111 (2014).

Data Citation

1. Peters, K., Gorzolja, K., Neumann, S. & Bruelheide, H. *MetaboLights* MTBLS520 (2018).

Acknowledgements

K.P. acknowledges funding from the European Commission PhenoMeNal Grant EC654241. Further, we like to thank the Leibniz Foundation for supporting this study, the PhenoMeNal e-infrastructure for supplying the cloud environment for testing the Galaxy workflow, Stefanie Döll for helping with annotation, Sylvia Krüger and Julia Taubert for technical assistance and Dierk Scheel for advice and corrections to the manuscript.

Author Contributions

K.P.: Design of the experiment, Field sampling, Statistics, Quality Control, Code, Galaxy workflow, Docker container image, Writing the first draft of the manuscript K.G.: Extraction protocol and LC/MS data acquisition H.B.: Advice on multivariate statistics S.N.: Advice on the Quality Control with XCMS All authors contributed to the final version of the manuscript.

Additional Information

Tables 1, 2 and 3 are available only in the online version of this paper.

Competing interests: The authors declare no competing interests.

How to cite this article: Peters, K. *et al.* Computational workflow to study the seasonal variation of secondary metabolites in nine different bryophytes. *Sci. Data* 5:180179 doi: 10.1038/sdata.2018.179 (2018).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

The Creative Commons Public Domain Dedication waiver <http://creativecommons.org/publicdomain/zero/1.0/> applies to the metadata files made available in this article.

© The Author(s) 2018