5. References

Alonso A, Julia A, Beltran A, Vinaixa M, Diaz M, Ibanez L, Correig X, Marsal S. AStream: an R package for annotating LC/MS metabolomic data. Bioinformatics 2011;27:1339-1340.

Baker M. In biomarkers we trust? Nature Biotechnology 2005;23:297-304.

Balog J, Sasi-Szabo L, Kinross J, Lewis MR, Muirhead LJ, Veselkov K, Mirnezami R, Dezso B, Damjanovich L, Darzi A, Nicholson JK, Takats Z. Intraoperative Tissue Identification Using Rapid Evaporative Ionization Mass Spectrometry. Science Translational Medicine 2013;5:194ra93.

Belhajjame K, Zhao J, Garijo D, Gamble M, Hettne K, Palma R, Mina E, Corcho O, Gomez-Perez JM, Bechhofer S, Klyne G and Goble C. Using a suite of ontologies for preserving workflow-centric research objects. Web Semantics: Science, Services and Agents on the World Wide Web 2015;32:16-42.

Berthold MR, Cebron N, Dill F, Di Fatta G, Gabriel TR, Georg F, Meinl T, Ohl P, Sieb C, Wiswedel B. KNIME: the Konstanz Information Miner. In: Proceedings 4th Annual Industrial Simulation Conference (ISC), Workshop on Multi-Agent Systems and Simulation 2006.

Boekel J, Chilton JM, Cooke IR, Horvatovich PL, Jagtap PD, Kall L, Lehtio J, Lukasse P, Moerland PD, Griffin TJ. Multi-omic data analysis using Galaxy. Nature Biotechnology 2015;33:137-139.

Brown MV, McDunn JE, Gunst PR, Smith EM, Milburn MV, Troyer DA, Lawton KA. Cancer detection and biopsy classification using concurrent histopathological and metabolomic analysis of core biopsies. Genome Medicine 2012;4:33.

Cabaton NJ, Canlet C, Wadia PR, Tremblay-Franco M, Gautier R, Molina J, Sonnenschein C, Cravedi JP, Rubin BS, Soto A, Zalko D. Effects of low doses of Bisphenol A on the metabolome of perinatally exposed CD-1 mice. Environmental Health Perspectives 2013;121: 586-593.

Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, Gatto L, Fischer B, Pratt B, Egertson J, Hoff K, Kessner D, Tasman N, Shulman N, Frewen B, Baker TA, Brusniak MY, Paulse C, Creasy D, Flashner L, Kani K, Moulding C, Seymour SL, Nuwaysir LM, Lefebvre B, Kuhlmann F, Roark J, Rainer P, Detlev S, Hemenway T, Huhmer A, Langridge J, Connolly B, Chadick T, Holly K, Eckels J, Deutsch EW, Moritz RL, Katz JE, Agus DB, MacCoss M, Tabb DL, Mallick P. A cross-

platform toolkit for mass spectrometry and proteomics. Nature Biotechnology 2012;30:918-920.

Chen R, Mias G, Pook-Than JL, Jiang L, Lam H, Chen R, Miriami E, Karczewski K, Hariharan M, Dewey F, Cheng Y, Clark M, Im H, Habegger L, Balasubramanian S, O'Huallachain M, Dudley J, Hillenmeyer S, Haraksingh R, Sharon D, Euskirchen G, Lacroute P, Bettinger K, Boyle A, Kasowski M, Grubert F, Seki S, Garcia M, Whirl-Carrillo M, Gallardo M, Blasco M, Greenberg P, Snyder P, Klein T, Altman R, Butte AJ, Ashley E, Gerstein M, Nadeau K, Tang H, Snyder M. Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell 2012;148:1293-1307.

Cottret L, Wildridge D, Vinson F, Barrett MP, Charles H, Sagot MF, Jourdan F. MetExplore: a web server to link metabolomic experiments and genome-scale metabolic networks. Nucleic Acids Research 2010;38:W132-W137.

Cuperlovic-Culf M, Barnett DA, Culf AS, Chute I. Cell culture metabolomics: applications and future directions. Drug Discovery Today 2010;15:610-621.

Davidson R, Weber R, Liu H, Sharma-Oates A, Viant M. Galaxy-M: A Galaxy workflow for processing and analyzing direct infusion and liquid chromatography mass spectrometry-based metabolomics data. GigaScience 2016;5:1-9.

Degtyarenko K, de Matos P, Ennis M, Hastings J, Zbinden M, McNaught A, Alcantara R, Darsow M, Guedj M, Ashburner M. ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Research 2008;36:D344-D350.

Delabriere A, Hohenester U, Junot C, Thevenot EA. proFIA: preprocessing data from flow injection analysis coupled to high-resolution mass spectrometry. *under review*.

Dieterle F, Ross A, Schlotterbeck G, Senn H. Probabilistic Quotient Normalization as Robust Method to Account for Dilution of Complex Biological Mixtures. Application in 1H NMR Metabonomics. Analytical Chemistry 2006;78:4281-4290.

Dittami SM, Gravot A, Renault D, Goulitquer S, Eggert A, Bouchereau A, Boyen C, Tonon T. Plant, Cell & Environment 2011;34:629-642.

Dittami SM, Gravot A, Goulitquer S, Rousvoal S, Peters AF, Bouchereau A, Boyen C, Tonon T. Towards deciphering dynamic changes and evolutionary mechanisms involved in the adaptation to low salinities in Ectocarpus (brown algae). The Plant Journal 2012;71:366-377.

Dunn WB, Broadhurst D, Begley P, Zelena E, Francis-McIntyre S, Anderson N, Brown M, Knowles JD, Halsall A, Haselden JN, Nicholls AW, Wilson ID, Kell DB,

Goodacre R (2011). Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. Nature Protocols 2011;6:1060-1083.

Etalo DW, De Vos RC, Joosten MH, Hall RD. Spatially Resolved Plant Metabolomics: Some Potentials and Limitations of Laser-Ablation Electrospray Ionization Mass Spectrometry Metabolite Imaging. Plant Physiology 2015;169:1424-1435.

Fan J, Saha S, Barker G, Heesom KJ, Ghali F, Jones AR, Matthews DA, Bessant C. Galaxy integrated omics: web-based standards-compliant workflows for proteomics informed by transcriptomics. Molecular & Cellular Proteomics 2015;14:3087-3093.

Giacomoni F, Le Corguille G, Monsoor M, Landi M, Pericard P, Petera M, Duperier C, Tremblay-Franco M, Martin J, Jacob D, Goulitquer S, Thevenot E, Caron C. Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics. Bioinformatics 2015;31:1493-1495.

Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A. Galaxy: A platform for interactive large-scale genome analysis. Genome Research 2005;15:1451-1455.

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

Gonzalez-Beltran A, Li P, Zhao J, Avila-Garcia MS, Roos M, Thompson M, van der Horst E, Kaliyaperumal R, Luo R, Lee TL, Lam TW, Edmunds SC, Sansone SA, Rocca-Serra P. From peer-reviewed to peer-reproduced in scholarly publishing: The complementary roles of data models and workflows in bioinformatics. PLoS ONE 2015;10:e0127612.

Haug K, Salek RM, Conesa P, Hastings J, de Matos P, Rijnbeek M, Mahendraker T, Williams M, Neumann S, Rocca-Serra P, Maguire E, Gonzalez-Beltran A, Sansone SA, Griffin JL, Steinbeck C. MetaboLights: an open-access general-purpose repository for metabolomics studies and associated meta-data. Nucleic Acids Research 2013;41:D781-D786.

Holmes E, Loo RL, Stamler J, Bictash M, Yap IKS, Chan Q, Ebbels T, De Iorio M, Brown IJ, Veselkov KA, Daviglus ML, Kesteloot H, Ueshima H, Zhao L, Nicholson JK, Elliott P. Human metabolic phenotype diversity and its association with diet and blood pressure. Nature 2008;453:396-400.

Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K, Ojima Y, Tanaka K, Tanaka S, Aoshima K, Oda Y, Kakazu Y, Kusano M, Tohge T, Matsuda F, Sawada Y, Hirai MY, Nakanishi H, Ikeda K, Akimoto N, Maoka T, Takahashi H, Ara T, Sakurai N, Suzuki H, Shibata D, Neumann S, Iida T, Tanaka K, Funatsu K, Matsuura F, Soga T, Taguchi R, Saito K, Nishioka T. MassBank: a public repository for sharing mass spectral data for life sciences. Journal of Mass Spectrometry 2010;45:703-714.

Hull D, Wolstencroft K, Stevens R, Goble C, Pocock MR, Li P, Oinn T. Taverna: a tool for building and running workflows of services. Nucleic Acids Research 2006;34:W729-W732.

Jagtap PD, Johnson JE, Onsongo G, Sadler FW, Murray K, Wang Y, Shenykman GM, Bandhakavi S, Smith LM, Griffin TJ. Flexible and accessible Workflows for Improved Proteogenomic Analysis Using the Galaxy Framework. Journal of Proteome Research 2014;13:5898-5908.

Jagtap PD, Blakely A, Murray K, Stewart S, Kooren J, Johnson JE, Rhodus NL, Rudney J, Griffin TJ. Metaproteomic analysis using the Galaxy framework. Proteomics 2015;15:3553-3565.

Johnson CH, Ivanisevic J, Benton HP, Siuzdak G. Bioinformatics: The next frontier of metabolomics. Analytical Chemistry 2015;87:147-156.

Johnson CH, Ivanisevic J, Siuzdak G. Metabolomics: beyond biomarkers and towards mechanisms. Nature Reviews Molecular Cell Biology 2016;17:451-459.

Kale NS, Haug K, Conesa P, Jayseelan K, Moreno P, Rocca-Serra P, Nainala VC, Spicer RA, Williams M, Li X, Salek RM, Griffin JL, Steinbeck C. MetaboLights: An Open-Access Database Repository for Metabolomics Data. Current Protocols in Bioinformatics 2016;14.13.1-14.13.18.

Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research 2000;28:27-30.

Kell DB, Oliver SG. The metabolome 18 years on: a concept comes of age. Metabolomics 2016;12:148.

van der Kloet FM, Bobeldijk I, Verheij ER, Jellema RH. Analytical error reduction using single point calibration for accurate and precise metabolomic phenotyping. Journal of Proteome Research 2009;8:5132-5141.

Kopka J, Schauer N, Krueger S, Birkemeyer C, Usadel B, Bergmuller E, Dormann P, Weckwerth W, Gibon Y, Stitt M, Willmitzer L, Fernie AR, Steinhauser D.

GMD@CSB.DB: the Golm Metabolome Database. Bioinformatics 2005;21:1635-1638.

Kuhl C, Tautenhahn R, Bottcher C, Larson TR, Neumann S. CAMERA: An integrated strategy for compound spectra extraction and annotation of Liquid Chromatography/Mass Spectrometry Data Sets. Analytical Chemistry 2012;84:283-289.

Leipzig J. A review of bioinformatic pipeline frameworks. Briefings in Bioinformatics 2017;18:530-536.

Leonelli S, Smirnoff N, Moore J, Cook C, Bastow R. Making open data work for plant scientists. Journal of Experimental Botany 2013;64:4109-4117.

Li D, Heiling S, Baldwin IT, Gaquerel E. Illuminating a plant's tissue-specific metabolic diversity using computational metabolomics and information theory. Proceedings of the National Academy of Sciences 2016;113:E7610-E7618.

Nicholson J, Lindon J, Holmes E. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. Xenobiotica 1999;29:1181-1189.

Mons B, van Haagen H, Chichester C, Hoen PB, den Dunnen JT, van Ommen G, van Mulligen E, Singh B, Hooft R, Roos M, Hammond J, Kiesel B, Giardine B, Velterop J, Groth P, Schultes E. The value of data. Nature Genetics 2011;43:281-283.

Oliver SG, Winson MK, Kell DB, Baganz F. Systematic functional analysis of the yeast genome . Trends in Biotechnology 1998;16:373-378.

Patti GJ, Tautenhahn R, Siuzdak G. Meta-analysis of untargeted metabolomic data from multiple profiling experiments. Nature Protocols 2012;7:508-516.

Peng T, Royer AL, Guitton Y, Le Bizec B, Dervilly-Pinel G. Serum-based metabolomics characterization of pigs treated with ractopamine. Metabolomics 2017;13:77-91.

Peyraud R, Schneider K, Kiefer P, Massou S, Vorholt J, Portais JC. Genome-scale reconstruction and system level investigation of the metabolic network of Methylobacterium extorquens AM1. BMC Systems Biology 2011;5:189.

Rinaudo P, Boudah S, Junot C, Thevenot E. *biosigner*: A new method for the discovery of significant molecular signatures from omics data. Frontiers in Molecular Biosciences 2016;3.

Rolin D. Metabolomics Coming of Age with its Technological Diversity. Advances in Botanical Research 2013:67;2-693.

Roux A, Xu Y, Heilier JF, Olivier MF, Ezan E, Tabet JC, Junot C. Annotation of the human adult urinary metabolome and metabolite identification using ultra high performance liquid chromatography coupled to a linear quadrupole ion trap-orbitrap mass spectrometer. Analytical Chemistry 2012;84:6429-6437.

Smith CA, Maille GO, Want EJ, Qin C, Trauger SA, Brandon TR, Custodio DE, Abagyan R, Siuzdak G. METLIN: A Metabolite Mass Spectral Database. Therapeutic Drug Monitoring 2005;27:747-751.

Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Analytical Chemistry 2006;78:779-787.

Sud M, Fahy E, Cotter D, Azam K, Vadivelu I, Burant C, Edison A, Fiehn O, Higashi R, Nair KS, Sumner S, Subramaniam S. Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. Nucleic Acids Research 2016;44:D463-D470.

Sumner L, Amberg A, Barrett D, Beale M, Beger R, Daykin C, Fan T, Fiehn O, Goodacre R, Griffin J, Hankemeier T, Hardy N, Harnly J, Higashi R, Kopka J, Lane A, Lindon J, Marriott P, Nicholls A, Reily M, Thaden J, Viant M. Proposed minimum reporting standards for chemical analysis. Metabolomics 2007;3:211-221.

Tardivel PJC, Servien R, Concordet D. Familywise Error Rate Control With a Lasso Estimator. Submitted.

Tautenhahn R, Bottcher C, Neumann S. Highly sensitive feature detection for high resolution LC/MS. BMC Bioinformatics 2008;9:504.

Tautenhahn R, Patti GJ, Rinehart D, Siuzdak G. XCMS Online: A web-based platform to process untargeted metabolomic data. Analytical Chemistry 2012;84:5035-5039.

Tenenhaus M. La régression PLS. Revue de Statistiques Appliquées 1999;47:5-40.

Thevenot EA, Roux A, Xu Y, Ezan E, Junot C. Analysis of the human adult urinary metabolome variations with age, body mass index and gender by implementing a comprehensive workflow for univariate and OPLS statistical analyses. Journal of Proteome Research 2015;14:3322-3335.

Via A, Blicher T, Bongcam-Rudloff E, Brazas MD, Brooksbank C, Budd A, De Las Rivas J, Dreyer J, Fernandes PL, van Gelder C, Jacob J, Jimenez RC, Loveland J, Moran F, Mulder N, Nyrnen T, Rother K, Schneider MV, Attwood TK. Best practices in bioinformatics training for life scientists. Briefings in Bioinformatics 2013;14:528-537.

Weber RJM, Winder CL, Larcombe LD, Dunn WB, Viant MR. Training needs in metabolomics. Metabolomics 2015;11:784-786.

Wehrens R, Weingart G, Mattivi F. metaMS: An open-source pipeline for GC-MS-based untargeted metabolomics. Journal of Chromatography B 2014;966:109-116.

Weiss N, Barbier Saint Hilaire P, Colsch B, Isnard F, Attala S, Schaefer A, Amador MdM, Rudler M, Lamari F, Sedel F, Thabut D, Junot C. Cerebrospinal fluid metabolomics highlights dysregulation of energy metabolism in overt hepatic encephalopathy. Journal of Hepatology 2016;65:1120-1130.

Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, Cheng D, Jewell K, Arndt D, Sawhney S, Fung C, Nikolai L, Lewis M, Coutouly MA, Forsythe I, Tang P, Shrivastava S, Jeroncic K, Stothard P, Amegbey G, Block D, Hau DD, Wagner J, Miniaci J, Clements M, Gebremedhin M, Guo N, Zhang Y, Duggan GE, MacInnis GD, Weljie AM, Dowlatabadi R, Bamforth F, Clive D, Greiner R, Li L, Marrie T, Sykes BD, Vogel HJ, Querengesser L. HMDB: the Human Metabolome Database. Nucleic Acids Research 2007;35:D521-526.

Xia J, Psychogios N, Young N, Wishart DS. MetaboAnalyst: a web server for metabolomic data analysis and interpretation. Nucleic Acids Research 2009;37:W652-W660.

Supplementary material

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, Marie Tremblay-Franco, Gildas Le Corguillé, Jean-François Martin, Mélanie Pétéra, Pierrick Roger-Mele, Alexis Delabrière, Sophie Goulitquer, Misharl Monsoor, Christophe Duperier, Cécile Canlet, Rémi Servien, Patrick Tardivel, Christophe Caron, Franck Giacomoni, and Etienne A. Thévenot

Contents

1. Workflow Management with W4M

We provide here a brief description of the features offered by the Workflow4Metabolomics 3.0 online infrastructure (e-infrastructure). Some of these features, such as the building and sharing of workflows and histories, are provided by the Galaxy environment, which is used internally by W4M (Goecks et al., 2010). Other features are specific to the e-infrastructure, such as the computational tools themselves, the referencing of histories with DOI, the user accounts, the tutorials, the help desk, and the practical training sessions.

1.1. User Accounts

Accounts can be requested on the W4M home page. Accounts are private (i.e., only accessible to the user via his/her credentials).

1.2. Data Format

1.2.1. Raw data (for preprocessing tools)

1.2.1.1. MS

MS data should be in either the mzML, mzXML, mzData, or NetCDF open format (Smith et al., 2006). We recommend the use of centroid data, in order to reduce file size. Conversion of raw data from proprietary format to centroid open format should be performed by the user before upload into W4M, e.g., by using the open-source ProteoWizard software (Chambers et al., 2012).

Raw files should be organized into a single or multiple folder(s), corresponding to the class(es) to be considered during preprocessing. The class information is used in the ***xcms.group*** tool, through the *minfrac* argument, to discard features which are not detected in a sufficient number of samples in at least one class (Smith et al., 2006). Separate classes may correspond to distinct sample types (blank reagent, quality control pools), or experimental conditions (treatment; Smith et al., 2006).

Raw files (either in a single or multiple folders) must be zipped before upload (we recommend the use of the 7-Zip open source software; http://www.7-zip.org).

For LC-MS, it is also possible now to upload the files individually (i.e., without folders and zip file). Within the history, the files must then be grouped as a data collection

(Afgan et al., 2016) for further parallel processing by the ***xcms.xcmsSet*** tool. The use of a data collection therefore speeds up this computer intensive step. After peak detection, the collection of xset.RData outputs, together with a *sampleMetadata* file indicating the classes, are merged with the ***xcms.xcmsSet Merger*** tool before the grouping step (***xcms.group***). More details about the use of data collection for LC-MS data preprocessing can be found:

1) in the following tutorial:
http://download.workflow4metabolomics.org/docs/170510_galaxy_xcms_dataset_collection.m4v
2) and on the 'W4M_sacurine-subset_parallel-preprocessing' public history:
https://galaxy.workflow4metabolomics.org/history/list_published

1.2.1.2. NMR

NMR preprocessing tools currently work with Bruker files. Each sample directory should be organized with acquisition run and process numbered "1" (Table 1 and Fig. 8; upper right). Sample directories should then be gathered in a single parent directory, which should in turn be zipped before upload into W4M.

1.2.2. Preprocessed data (for normalization, quality control, statistical analysis, and annotation tools)

Preprocessing of the sample raw files generates a unique sample by variable data matrix of peak intensities, in addition to metadata of samples (e.g., sample ID, factor of interest) and variables (e.g., m/z and retention time). Such data and metadata are handled in W4M in a unique format consisting of 3 separate tables: *dataMatrix*, *sampleMetadata*, and *variableMetadata*. These tables, in a tabulated format (e.g., *.tsv*), are generated by the preprocessing tools from W4M, but can also be created or modified with spreadsheet editors (such as the Excel or the OpenOffice Calc software). The "3 table" format is used in all tools following the preprocessing (i.e., all tools for normalization, quality control, statistical analysis, and annotation). Details about the formats of the 3 tables can be found in the online tutorials (*HowTo* section from the front page). Once uploaded into W4M with the ***Upload File*** tool, the formats of the 3 tables can be verified with the ***Check Format*** tool (*Data Handling* section; note that the 'search tools' feature on top of the 'Tools' panel is helpful for locating a specific tool).

1.3. Data Upload

The ***Upload File*** tool allows data upload from a local computer into the W4M infrastructure. For sizes up to 2 GB (e.g., NMR raw zip or preprocessed data and metadata tables), files can be selected with a simple drag and drop. For bigger data sets (e.g., LC-MS and GC-MS raw zip), an FTP client software is required (such as Cyberduck, https://cyberduck.io, or WinSCP, https://winscp.net) to directly connect to

the *ftp://workflow4metabolomics.org* infrastructure (with your W4M credentials) and copy/paste the zip file (see the *Galaxy Initiation* tutorial in the *HowTo* section).

1.4. Building Workflows

The user-friendly Galaxy features within the W4M e-infrastructure allow to build and save complex workflows. Workflows can be built *de novo* by using the *canvas* (editor): tools can be chained, and parameter values can be selected. Alternatively, workflows can be extracted from an existing history with the *Extract workflow* option (e.g., when the tools have been sequentially tuned on specific data).

1.5. Running Workflows

W4M offers users a high-performance environment for computing (4,000 cores) and data storage. The infrastructure can be accessed via a simple web browser and does not consume local resources. Once jobs are launched, the computation will continue and the results will be saved, even if the local connection is switched off by the user.

1.6. Sharing Histories

Histories can be shared with a dedicated user by using his/her W4M email (e.g., with colleagues within a lab or a consortium, or with a member from the help desk). Alternatively, histories can be published online to give all users unrestricted access to the workflow and the associated data (e.g., for training or citation purposes, see below).

1.7. Referencing Histories

Histories (workflow and all associated input and output data and metadata) can now be published on W4M 3.0 with a reference ID and a permanent Digital Object Identifier (DOI; see the *Result* section). Reference histories can then be cited in publications (see for instance Rinaudo et al., 2016), thus giving reviewers and readers full access to the analysis (i.e., data, metadata, workflow, parameters, and results).

1.8. Tutorials

Tutorials for data preparation, upload, and analysis, in addition to history sharing and publishing, are available in the *HowTo* section from the front page (http://workflow4metabolomics.org).