2 **Standards for Reporting Metabolomics Data in Phytomedicine**
3 **Research**

4 *Biswapriya B. Misra*^{1, #},

5
6 ¹Department of Genetics, Texas Biomedical Research Institute, San Antonio, 78227, TX USA

7
8
9 *#Corresponding Author*

10 Biswapriya B. Misra, Ph.D.
11 Staff Scientist I
12 Texas Biomedical Research Institute (TBRI)
13 Department of Genetics
14 7620 NW Loop 410
15 San Antonio 78227, Texas USA
16 Office: (210) 258-9623
17 E-mails: bbmisracb@gmail.com, bmisra@txbiomed.org
18 ORCID: orcid.org/0000-0003-2589-6539
19 Twitter Handle: [@BiswapriyaMisra](https://twitter.com/BiswapriyaMisra)
20
21
22
23
24
25
26
27

28 **Abstract**

29 Growing number of studies are using metabolomics as a research tool to aid in hypothesis
30 generation or in solving a question. Unsurprisingly, a vast number of ethnopharmacology, herbal
31 medicine, and phytomedicine investigations have started to rely on metabolomics approaches to
32 find the mechanisms of action or to characterize a herbal medicine. However, it is often
33 encountered that the reported study fails to report the metabolomics data in a comprehensible
34 format or accessible to a larger audience. Hence, this effort is to make the authors and readers
35 aware of the existing minimum standards for reporting metabolomics in addition to current
36 developments and advances in this exciting area of research. Moreover, data archiving is becoming
37 very important for FAIR (Findability, Accessibility, Interoperability, and Reusability) purposes.
38 Understandably, all requirements may not be met by novice authors and research groups, however,
39 best put must be put forth to meet the standards as a practice.

40 **Keywords**

41 Archiving; best practice; big data; database; mass spectrometry; metabolomics; metabolite; NMR;
42 metabolism; -omics;

43

44

45

46

47

Introduction

Metabolomics, as a research avenue is booming, resulting into thousands of research studies culminating in publications that cover almost the entire breadth of life sciences, ranging from microbiology, plant and animal sciences to biomedical engineering and systems biology (**Figure 1a**). Ethnopharmacology and ethnomedicine are no exceptions to this growing trend where a vast array of studies focus on metabolomic changes associated with a natural product-based bioresource, phytochemicals, extracts and pure compounds. However, in this rush to amalgamate ethnopharmacology with metabolomics, a lot of caveats are evident in preliminary submissions to top notch and leading journals in this area of ethnopharmacology. This treatise is a call for authors and new investigators in metabolomics research to help them meet the expectations of the metabolomics research community when dealing with ethnopharmacology-based investigations. These bioprospection efforts have been speeded with the help of –omics scale studies, esp. relying on metabolites, variously called as metabolomics, metabonomics as well as metabolite profiling for studies that deal with narrow chemical classes. Without adding another –omics term, we feel that efforts in metabolomics studies in phytomedicine and ethnopharmacology may eventually lead to explosion in this area of scientific advancements with immense bioresources hidden in Traditional Chinese Medicine (TCM), Ayurveda, and other folklore medicines. Although, it is very difficult to define standards, but a set of standard rules would certainly help authors address these points to a greater extent- thereby facilitating better peer-review and an outcome.

1. ***Study design and statistical power:*** In a collaborative research world that is highly interdisciplinary, it is not difficult to seek metabolomics services from academia, industry, collaborators, and colleagues. However, a thorough consideration must be given to the aspects of study design as to whether the proposed metabolomics study is hypothesis driven

or hypothesis-generating. A treatise is also available on quick tips to perform a metabolomics study (Misra, 2016), which touch bases on study design, platform usage for a given study, adoption of metabolomics standard initiative (MSI) guidelines, tools for data processing and annotation, statistical analyses, integration with other –omics data, pathway and network analysis, visualization and data availability among others. For any –omics scale study, where number of observations (n) are minuscule as compared to number of parameters (p), a critical number of ‘n’ is important for confidence in the study. Many a times, authors cited ‘limited resources’ for performing a study with just two biological replicates, or somehow reach the magic figure of three replicates to conduct large-scale –omics studies, which not only raises questions on the reproducibility of the study but also on other ethical issues. Although, the subject lends itself to more discussion than solutions, for instance, a minimum of 20 samples have been suggested for providing necessary power to achieve an effect size of 0.8 (Lenth et al., 2001).

2. ***Quality controls and analytical standards:*** Easier said that ‘control what you can control, and do not, what you cannot. But it is far easy to be pragmatic in approaches to record all the quality control (QC) measures during the course of experiments. One common flaw in many published studies is the non-inclusion of an *internal standard* (IS) during extraction steps. With IS (a unique compound absent or not expected in the analytes or samples of interest) a better control on extraction, normalization, instrument drift over time, batch effects and others can be accounted for.

Pooled sample mixtures can serve as excellent QCs when randomly run within and across batches. Their performance may be shown as Supplementary information to account for any batch effects or instrument drift over the period of data acquisition in a large study. In

a typical PCA output the grouped separation of such QC pools, blank runs (negative controls) or dummy sample runs must be evident. Preferably, isotope labelled standards at least one representing each chemical class ideally be used as internal standard during data acquisition.

Analytical standards used to calibrate the instruments (for ionization efficiency, mass accuracy, stability), or to prepare the libraries to obtain the spectra must be reported alongside their spectra whenever possible, as well. The current state of quality assurance and quality control processes within the metabolomics society and the recommendations can be found here (Dunn et al., 2017).

3. ***Choice of tissue or biofluids***: Of course ethical issues, institutional constraints, funding, availability, and the study itself dictates as to what tissue or sample one must subject to metabolomics analysis. But, then what is point of using a mice or rat model in the study, if for instance say, instead of liver tissues the serum/ plasma are interrogated for the effects of a phytochemical or phytoextract that is hepatoactive. The choice between performing serum or plasma metabolomics is guided by whether using targeted (Breier et al., 2014) approaches, their collection methods and storage conditions (Moriya et al., 2016). Analytes or pathways of interest should guide the solvents to be used for extraction, and analytical platform used to acquire metabolomics data. Excellent treatise on these aspects are available for plant (Kim and Verpoorte, 2010) as well as for clinical (Chetwynd et al., 2017) metabolomics. Also detailed are sample normalization methods in quantitative metabolomics (Wu and Li, 2016) or those from data-driven normalization methods (Li et al., 2016).

4. ***Statistical considerations:*** Statistical treatment of the data is paramount before drawing any conclusive biological insights. Understandably, statistical significance does not equal to biological relevance, alone, but then identification of false positives and true negatives are important to not over-read into statistical analyses. Most studies suffer from the fact that, a multiple hypothesis testing or correction is not in place, be it through Bonferroni or false discovery rate (of 5% at least) corrective measures (Benjamini and Hochberg, 1995). Clearly, the field has come along, where performing multiple t-tests for p-hacking (see Head et al., 2015), in lieu of an Analysis of Variance (ANOVA) followed by post-hoc tests, have clearly been abandoned. For instance, why certain fold change (FC) cut-off or a P-value cut off were chosen must be rationalized. Thus, consulting a biostatistician prior to study design and during data analysis is highly recommended. Furthermore, the choice of visualizations for heat maps (color choices), clustering algorithms, dendrogram versus heat maps, principal components analysis (PCA) versus partial least-squares discriminant analysis (PLSDA) analysis and such considerations should be study-specific. Depending on the study design, univariate analysis such as t-tests or ANOVA or multivariate analyses such as unsupervised PCA and orthogonal PLS-DA (OPLS-DA), sparse PCA (sPCA) can be performed to show the clustering of samples, and discrimination between groups (i.e., control vs. disease etc.). More advanced analysis such as Correspondence analysis (CA), Multiple Correspondence Analysis (MCA) can be performed using available R-packages with the same names. *BioStatFlow* version 2.7.7. (<http://biostatflow.org/>), developed by National Institute of Agronomic Research, France (INRA, <http://www.inra.fr/>), is an extremely user friendly web-tool for analyses of -omics scale data.

5. ***Tools, resources and databases:*** In recent times, a lot of reviews have focused to aggregate the software tools, databases, resources for the metabolomics community. Recent reviews have identified all tools and resources developed in 2015-2017 for the metabolomics community for (Misra and Hooft, 2016, Misra et al., 2017) which covers topics from data handling and preprocessing, tools for statistical analysis of metabolomics data, metabolite annotation and identification tools and resources, pathway analysis, network construction, visualization and biological interpretation tools, metabolomics libraries, databases, experiment repositories and meta-data storage resources followed by GC-MS-, NMR-based and multifunctional tools. In addition, efforts have also helped collate the mass spectral databases for LC-MS and GC-MS based metabolomics (Vinaixa et al., 2016), software tools and resources for NMR-based metabolomics (Eghbalnia et al., 2016), and tools for mass spectrometry imaging (Ràfols et al., 2016).

Undoubtedly, the tools, software and databases have been growing in the well-known online resources such as *OMICtools* (<http://omictools.com/>) (Henry et al., 2014) Fiehn Lab resources (<http://fiehnlab.ucdavis.edu/>), Metabolomics Society's resource pages (<http://metabolomicssociety.org/resources/metabolomics-databases>), *ms-utils.org* (<http://www.ms-utils.org/wiki/pmwiki.php/Main/SoftwareList>) Pacific Northwest National Laboratory (PNNL)'s Integrative Omics open source software collections (<https://omics.pnl.gov/software/>), and *metabomatch* (<http://www.metabomatch.com/software/>).

Just citing that an X- commercial software or database or an open-source tool yields these results is not enough- the underlying logic must be explained and substantiated from the results. Also the commercial tool's vendor name, exact version used and ditto parameters

used for analysis must be explicitly stated. There is also a growing urgency now for reproducible research, where the scripts used (be it from R, Python, Matlab) must be shared with the research community and the workflow used i.e., KNIME-based or otherwise must be shared.

6. ***Nomenclature of compounds***: Although the common or chemical names of the compounds are often used (i.e., glucose, maltose, proline), in addition to say the *International Union of Pure and Applied Chemistry* (IUPAC) names for natural products, for metabolomics community a very standard norm is to use the InChIKey or even the very recently proposed SPLASH key (Wohlgemuth et al., 2016). Once the annotations have generated the names associated with the metabolite features or assign them as metabolites, the common or chemical names derived from the searched database may be converted to their more recognizable and uniform IDs for the ease of data sharing and archiving purposes. These IDs include KEGG IDs, *HMDB* IDs, *PubChem* (<https://pubchem.ncbi.nlm.nih.gov/>) IDs, *LipidMaps* (<http://www.lipidmaps.org/>) IDs, *InChIKey* (<http://www.inchi-trust.org/>) IDs, Chemical Abstracts Service (CAS) (<http://www.cas.org/content/chemical-substances>) IDs, Chemical Entities of Biological Interest (ChEBI) (<https://www.ebi.ac.uk/chebi/>) IDs, and *ChemSpider* (<http://www.chemspider.com/>) IDs among other options available. Denoting the L- and D-forms for amino acids, or R- and S- forms of stereoisomers is also imperative to avoid any confusion for follow-up investigators in the future. For such ID conversions, the Chemical Translation Service (CTS) hosted by the Oliver Fiehn Laboratory, UC Davis, is immensely useful: <http://cts.fiehnlab.ucdavis.edu/>, and that comes with options to batch convert up to 100 IDs at a time. Web-tools such as MBRole (<http://csbg.cnbc.es/mbrole2/>) do a

similar job. *Metabolonote* (<http://metabolonote.kazusa.or.jp/>) takes the innovative approach of using a Wiki style interface to facilitate community-based metadata annotation of metabolomics datasets from the literature, delocalizing the burden of this work (Ara et al., 2015). *ChemDataExtractor* (<http://chemdataextractor.org/>) is a very convenient to use webserver for reliable and automated extraction of chemical information from the scientific literature provided as a document or pasted onto the server (Swain and Cole, 2016). Moreover, for compound/ metabolite identifications the level of identification (from 1-4) should be mentioned clearly as recommended by the MSI (Sumner et al., 2007), which demonstrate the confidence in the presented data and whether the annotation relied solely on database and literature, authentic compounds, orthogonal techniques or further validations like synthesis etc.

7. **Figure and table quality:** An image is worth thousands words- but an image should not draw a thousand comments. Figures and images generated during metabolomics studies, i.e., heat maps, PCA score-plots, S-plots, correlation plots, pathways, and networks are often complex, huge or clumsy- but then the onus is on the authors to generate high resolution images with legible legends and justified use of colors. Even recently efforts have been put to help color-blind readers by suggesting suitable color combinations (Roskoski, 2017). Similarly, total ion chromatograms (TICs) or base peak chromatograms (BPCs) do not represent compounds in a LC-MS run, and though these look discriminatory for case-controls or for samples A and B, they hardly mean anything as far as their information content are concerned. And it is inconsequential to show those forests of peaks without identified compounds. Moreover, the outputs produced from the software tools and

webservers, such as one of the most popular ones such as MetaboAnalyst (Xia et al., 2009), XCMS Online (Gowda et al., 2014) produces nice visualizations, but that needs to be made high resolution and publishable (in terms of labels and fonts used) using off line tools such as commercial ones (Adobe Photoshop) or those available as open source tools such as Inkscape (<https://inkscape.org/en/>) or Gephi (<https://gephi.org/>). For instance just drawing a network based on available tools using apps in Cytoscape (<http://www.cytoscape.org/>) is not aesthetic, clear or informative enough and needs manual efforts to make them readable and interpretable. Same is true for KEGG-based pathway templates or those templates or outputs available with many popular commercial software from the mass spectrometer or NMR vendors.

In Tables, a list of so called “features” may be informative to the study, but in the longer run hardly it can be followed up without a recorded m/z, spectra, and retention time(s). So, metabolite ID are of paramount interest.

8. ***Data sharing, and archiving efforts:*** Data must be shared with the readers, peers, colleagues, researchers interested in metabolomics research, not only for the ethical and fairness issues, but to help the field move forward in all aspects. These shared data sets, make the reviewer aware of the data openness and let them dig into the data set to find the quality of acquired data, in addition to allowing others take a peek post-publication. One can never stress enough on the urgent need to allow reuse of scholarly data associated with academia, industry, funding agencies, and publishers, and what better than the FAIR (Findability, Accessibility, Interoperability, and Reusability) data principles (Wilkinson et al., 2016). Moreover, akin to the manner in which gene, transcript and genomics sequence data are archived with National Center for Biotechnology Information (NCBI)

(<https://www.ncbi.nlm.nih.gov/>) (Pruitt et al., 2005), and proteomics data sets are submitted and disseminated at the Proteomics Identifications Database (PRIDE) and ProteomeXchange Consortium (<http://www.proteomexchange.org/>) (Vizcaíno et al., 2014) there is an ethical obligation and moral responsibility from the authors to submit the metabolomics data as well.

These public archives are growing and are developed keeping in mind the metabolomics community that are available at UK-based MetaboLights (<http://www.ebi.ac.uk/metabolights/>) (Kale et al., 2016) and USA-based Metabolomics Workbench (<http://www.metabolomicsworkbench.org/>) (Sud et al., 2015), the two most prominent resources for archiving. More recently, the Global Natural Products Social Molecular Networking (GNPS; <http://gnps.ucsd.edu>), an open-access knowledge base for community-wide organization and sharing of raw, processed or identified tandem mass (MS/MS) spectrometry based metabolomics data is gaining popularity through the Mass Spectrometry Interactive Virtual Environment (MassIVE) data repository (<https://gnps.ucsd.edu/ProteoSAFe/index.jsp>) hosted at computational infrastructure of the University of California San Diego (UCSD) Center for Computational Mass Spectrometry (CCMS; <http://proteomics.ucsd.edu/>) (Wang et al., 2016). In addition, for the mass spectrometry imaging data MetaSpace (<http://annotate.metaspac2020.eu/#/datasets>) (Palmer et al., 2016).

In addition, MassBank (<http://www.massbank.jp/?lang=en>) and PRIMEe (Platform for RIKEN Metabolomics), RIKEN, Japan, hosted at <http://prime.psc.riken.jp/> are other

resources where data archiving is possible. Sharing of large datasets at the DRYAD (<http://datadryad.org/>; White et al., 2008) and Fig Share (<https://figshare.com/>, Thelwall and Kousha, 2016).

In addition, to these data sharing efforts, all raw data sets once extracted into .csv/ .tsv formats onwards, through the intermediate processing steps (i.e., normalizations, transformation, scaling etc.) should also be presented as Supplementary materials. Just saying that they are available on demand or at one of the senior authors website (without a Digital Object Identifier –DOI) is not useful for the community. The source codes, scripts, command lines used for certain data analysis or development of packages can be shared freely at Github (<https://www.github.com>). Further, myExperiment (<https://www.myexperiment.org/home>) is another growing collaborative environment where scientists can safely publish their workflows and in silico experiments, share them with groups and find those of others (Goble et al., 2010).

Moreover, KNIME (Berthold et al., 2009)-based workflows that incorporate steps from chromatographic pre-processing to statistical analysis and visualization of –omics scale data are starting to be incorporated in metabolomics workflow (Sarpe and Schriemer, 2017), and not a long time into the future when journals would be demanding such workflows be submitted alongside the manuscript.

Concluding remarks

To summarize, the challenge does not only encompass incorporating a robust metabolomics workflow into ethnopharmacological studies, but also in a robust study design, statistical

treatment, visualization and biological interpretation of the data, but also in correct and aesthetic presentation of the findings of the study and best data sharing practices. This is summarized in **Figure 1b**. With metabolomics marrying ethnopharmacology/ ethnomedicine/ phytomedicine, we have only touched the tip of a vastly underlying ice-berg where much more remains to be explored in terms of lead compound identification (incl. dereplication) and realization of natural resources. Understandably, a lot of subjectivism goes into incorporating the above thoughts into reality, but with set standards the current workflow followed till publications can be better streamlined.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

References

1. Ara, T., Enomoto, M., Arita, M., Ikeda, C., Kera, K., Yamada, M., Nishioka, T., Ikeda, T., Nihei, Y., Shibata, D., Kanaya, S., 2015. Metabolonote: A wiki-based database for managing hierarchical metadata of metabolome analyses. *Front Bioeng Biotechnol.* 3, 38.
2. Benjamini, Y., Hochberg Y., 1995. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J R Stat. Soc. Series B Stat. Methodol.* 57, 289-300.
3. Berthold, M.R., Cebron, N., Dill, F., Gabriel, T.R., Kötter, T., Meinl, T., Ohl, P., Thiel, K., Wiswedel, B., 2009. KNIME-the Konstanz information miner: version 2.0 and beyond. *SIGKDD Explor* 11, 26-31.
4. Breier, M., Wahl, S., Prehn, C., Fugmann, M., Ferrari, U., Weise, M., Banning, F., Seissler, J., Grallert, H., Adamski, J. and Lechner, A., 2014. Targeted metabolomics identifies reliable and stable metabolites in human serum and plasma samples. *PLoS One* 9, e89728.

5. Chan, K., Shaw, D., Simmonds, M.S., Leon, C.J., Xu, Q., Lu, A., Sutherland, I., Ignatova, S., Zhu, Y.P., Verpoorte, R., Williamson, E.M., 2012. Good practice in reviewing and publishing studies on herbal medicine, with special emphasis on traditional Chinese medicine and Chinese materia medica. *J Ethnopharmacol* 140, 469-475.
6. Chetwynd, A.J., Dunn, W.B., Rodriguez-Blanco, G., 2017. Collection and Preparation of Clinical Samples for Metabolomics. In *Metabolomics: From Fundamentals to Clinical Applications* (pp. 19-44). Springer International Publishing.
7. Cos, P., Vlietinck, A.J., Berghe, D.V., Maes, L., 2006. Anti-infective potential of natural products: how to develop a stronger in vitro 'proof-of-concept'. *J Ethnopharmacol* 106-290-302.
8. Dunn, W.B., Broadhurst, D.I., Edison, A., Guillou, C., Viant, M.R., Bearden, D.W., Beger, R.D., 2017. Quality assurance and quality control processes: summary of a metabolomics community questionnaire. *Metabolomics* 13, 50.
9. Eghbalnia, H.R., Romero, P.R., Westler, W.M., Baskaran, K., Ulrich, E.L., Markley, J.L., 2017. Increasing rigor in NMR-based metabolomics through validated and open source tools. *Curr Opin Biotechnol*, 43, 56-61.
10. Froede, T.S.A., Y.S. Medeiros, Y.S., 2008. Animal models to test drugs with potential antidiabetic activity. *J Ethnopharmacol* 115, 173-183.
11. Gertsch J., 2009. How scientific is the science in ethnopharmacology? Historical perspectives and epistemological problems. *J Ethnopharmacol* 122, 177-183.
12. Goble, C.A., Bhagat, J., Aleksejevs, S., Cruickshank, D., Michaelides, D., Newman, D., Borkum, M., Bechhofer, S., Roos, M., Li, P., and De Roure, D.: myExperiment: a

- repository and social network for the sharing of bioinformatics workflows, Nucl. Acids Res., 2010.
13. Gowda, H., Ivanisevic, J., Johnson, C.H., Kurczy, M.E., Benton, H.P., Rinehart, D., Nguyen, T., Ray, J., Kuehl, J., Arevalo, B. and Westenskow, P.D., 2014. Interactive XCMS Online: simplifying advanced metabolomic data processing and subsequent statistical analyses. Anal. Chem. 86, 6931-6939.
14. Head, M.L., Holman, L., Lanfear, R., Kahn, A.T., Jennions, M.D., 2015. The extent and consequences of p-hacking in science. PLoS Biol, 13, e1002106.
15. Heinrich, M., Edwards, S., Moerman, D.E., Leonti, M., 2009. Ethnopharmacological field studies: a critical assessment of their conceptual basis and methods. J. Ethnopharmacol 124, 1-17.
16. Henry, V.J., Bandrowski, A.E., Pepin, A.S., Gonzalez, B.J., Desfeux, A., 2014. OMICtools: an informative directory for multi-omic data analysis. Database, p.bau069.
17. Kale, N.S., Haug, K., Conesa, P., Jayseelan, K., Moreno, P., Rocca-Serra, P., Nainala, V.C., Spicer, R.A., Williams, M., Li, X., Salek, R.M., 2016. MetaboLights: an open-access database repository for metabolomics data. Curr Protoc Bioinformatics 14-13.
18. Kim, H.K., Verpoorte, R., 2010. Sample preparation for plant metabolomics. Phytochem. Anal. 21, 4-13.
19. Lenth, R.V., 2001. Some practical guidelines for effective sample size determination. Am. Stat. 55, 187-193.
20. Li, B., Tang, J., Yang, Q., Cui, X., Li, S., Chen, S., Cao, Q., Xue, W., Chen, N., Zhu, F., 2016. Performance evaluation and online realization of data-driven normalization methods used in LC/MS based untargeted metabolomics analysis. Sci. Rep. 6.

21. Matteucci, E., Giampietro, O., 2008. Proposal open for discussion: defining agreed diagnostic procedures in experimental diabetes research. *J Ethnopharmacol* 115, 163-172.
22. Misra, B.B., 2016. Quick tips to perform a metabolomics study. *PeerJ Preprints*. e2002v1
23. Misra, B.B., der Hooft, J.J., 2016. Updates in metabolomics tools and resources: 2014-2015. *Electrophoresis* 37, 86-110.
24. Misra, B.B., Fahrman, J.F., Grapov, D., 2017. Review of emerging metabolomic tools and resources: 2015–2016. *Electrophoresis*.
25. Moriya, T., Satomi, Y. and Kobayashi, H., 2016. Intensive determination of storage condition effects on human plasma metabolomics. *Metabolomics* 12, 179.
26. Palmer, A., Phapale, P., Chernyavsky, I., Lavigne, R., Fay, D., Tarasov, A., Kovalev, V., Fuchser, J., Nikolenko, S., Pineau, C., Becker, M., 2016. FDR-controlled metabolite annotation for high-resolution imaging mass spectrometry. *Nature Methods*.
27. Pruitt, K.D., Tatusova, T., Maglott, D.R., 2006. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 35, D61-D65.
28. Ràfols, P., Vilalta, D., Brezmes, J., Cañellas, N., del Castillo, E., Yanes, O., Ramírez, N., Correig, X., 2016. Signal preprocessing, multivariate analysis and software tools for MA (LDI)-TOF mass spectrometry imaging for biological applications. *Mass Spectrom. Rev.*
29. Roskoski, R., 2017. Guidelines for preparing color figures for everyone including the colorblind.
30. Sarpe, V., Schriemer, D.C., 2017. Supporting metabolomics with adaptable software: design architectures for the end-user. *Curr Opin Biotechnol* 43, 110-117.

31. Sud, M., Fahy, E., Cotter, D., Azam, K., Vadivelu, I., Burant, C., Edison, A., Fiehn, O., Higashi, R., Nair, K.S., Sumner, S., 2015. Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res.* p.gkv1042.
32. Sumner, L.W., Amberg, A., Barrett, D., Beale, M.H., Beger, R., Daykin, C.A., Fan, T.W.M., Fiehn, O., Goodacre, R., Griffin, J.L., Hankemeier, T., 2007. Proposed minimum reporting standards for chemical analysis. *Metabolomics* 3, 211–221.
33. Swain, M.C., Cole, J.M., 2016. ChemDataExtractor: A toolkit for automated extraction of chemical information from the scientific literature. *J. Chem. Inf. Model.*
34. Thelwall, M., Kousha, K., 2016. Figshare: a universal repository for academic resource sharing?. *Online Inform Rev* 40, 333-346.
35. Vinaixa, M., Schymanski, E.L., Neumann, S., Navarro, M., Salek, R.M., Yanes, O., 2016. Mass spectral databases for LC/MS-and GC/MS-based metabolomics: State of the field and future prospects. *Trends Analyt Chem.* 78, .23-35.
36. Vizcaíno, J.A., Deutsch, E.W., Wang, R., Csordas, A., Reisinger, F., Rios, D., Dianes, J.A., Sun, Z., Farrah, T., Bandeira, N., Binz, P.A., 2014. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* 32, 223-226.
37. Wang, M., Carver, J.J., Phelan, V.V., Sanchez, L.M., Garg, N., Peng, Y., Nguyen, D.D., Watrous, J., Kapon, C.A., Luzzatto-Knaan, T., Porto, C., 2016. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* 34, 828-837.

38. White, H., Carrier, S., Thompson, A., Greenberg, J., Scherle, R., 2008, September. The Dryad Data Repository: A Singapore framework metadata architecture in a DSpace environment. In *Dublin Core Conference* (pp. 157-162).
39. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3.
40. Wohlgemuth, G., Mehta, S.S., Mejia, R.F., Neumann, S., Pedrosa, D., Pluskal, T., Schymanski, E.L., Willighagen, E.L., Wilson, M., Wishart, D.S., Arita, M., 2016. SPLASH, a hashed identifier for mass spectra. *Nat. Biotechnol* 34, 1099-1101.
41. Wu, Y. and Li, L., 2016. Sample normalization methods in quantitative metabolomics. *J Chromatogr A* 1430, 80-95.
42. Xia, J.G., Psychogios, N., Young, N., Wishart D.S., 2009. MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic Acids Res.* 37, W652 -W660.

Figure 1. (a) Increasing popularity of metabolomics. **(b)** A generic workflow depicting steps from metabolomics data acquisition to analysis, and data archiving.