Discovering and linking public omics data sets using the Omics Discovery Index

To the Editor:

Biomedical data are being produced at an unprecedented rate owing to the falling cost of experiments and wider access to genomics, transcriptomics, proteomics and metabolomics platforms[1,2]. As a result, public deposition of omics data is on the increase. This presents new challenges, including finding ways to store, organize and access different types of biomedical data stored on different platforms. Here, we present the Omics Discovery Index (OmicsDI; http://www.omicsdi.org), an open-source platform that enables access, discovery and dissemination of omics data sets.

In 2016, a group of researchers, publishers and research funders published the first guidelines to make data "findable, accessible, interoperable and re-usable" (FAIR; https://www.force11.org/group/fairgroup/fairprinciples)[3]. The FAIR principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse. Challenges facing joint analyses of data sets of different types include achieving a common representation for data sets and their associated metadata, and the lack of protocols and tools that enable data exchange across multiple repositories.

With respect to the first principle (to make data 'findable'[3]), most of the available resources for the scientific community nowadays are either field-specific (i.e., genomics, proteomics or metabolomics experimental data sets) or organism-specific, but including data sets from different omics technologies (e.g., the *Saccharomyces* Genome Database; SGD). Finding a data set can be frustrated by the need to search individual repositories and read numerous publications. The development of consortia that integrate resources (e.g., ProteomeXchange and MetabolomeXchange) has helped to improve findability. *Nature*[1] and *Nature Biotechnology*[4] have highlighted the need for data-set-integration frameworks to increase findability of data. And, in the context of the European ELIXIR (https://www.elixir-europe.org/) and USA Big Data to Knowledge (BD2K)[5] trans-US National Institutes of Health (NIH) initiative, the need is clear for a dedicated platform, search engines and services enabling the aggregation of omics data sets, to resources, such as PubMed[6] or Europe PubMed Central (EuroPMC)[7].

OmicsDI is an open-source platform that can be used to access, discover and disseminate omics data sets. OmicsDI can integrate proteomics, genomics, metabolomics and transcriptomics data sets (**Fig. 1**). To date, 11 repositories have agreed on a common metadata structure framework and exchange format, and have contributed to OmicsDI (**Supplementary Notes 1–3**), including proteomics databases (the Proteomics Identifications (PRIDE) database, PeptideAtlas, the Mass Spectrometry Interactive Virtual Environment (MassIVE) and the Global Proteome Machine Database; GPMDB); metabolomics databases (MetaboLights, the Global Natural Products Social Molecular Networking project (GNPS), MetabolomeExpress and the Metabolomics Workbench), the major European Genome-Phenome Archive (EGA) and transcriptomics databases (ArrayExpress and Expression Atlas). OmicsDI stores biological and technical metadata from these public data sets using an efficient indexing system (**Fig. 1b**) that can integrate

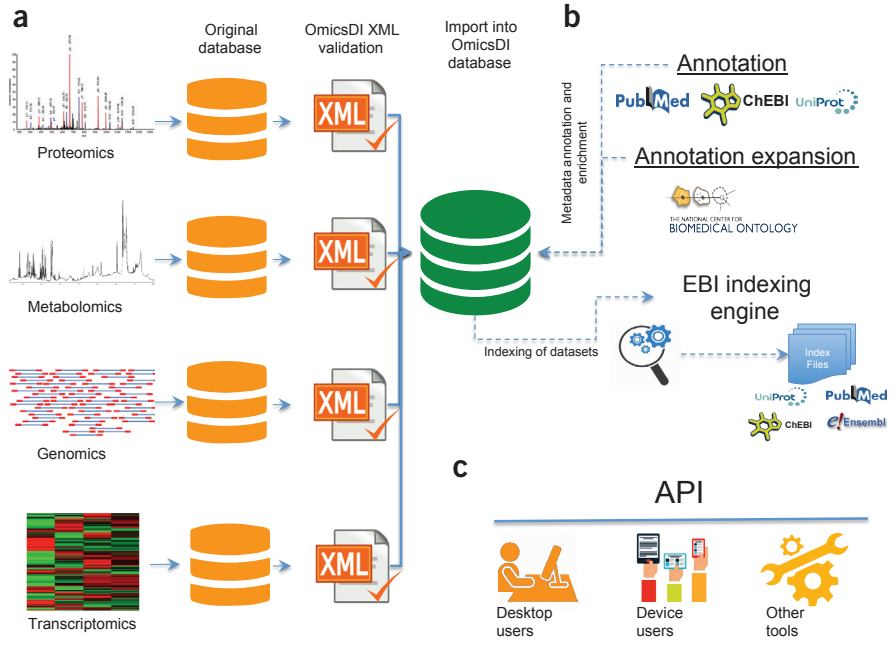

Figure 1 Omics Discovery Index: data standardization, annotation, index and presentation. (**a**) The data sets stored in public repositories are converted to a common data representation including all metadata and biological entities. The OmicsDI XML files are validated using the OmicsDI XML validator. (**b**) The OmicsDI XML files are then annotated using public services and databases like UniProt, ChEBI and PubMed, and the metadata are enriched using the Annotator service. The EBI search engine generates the indexes including other related resources such as PubMed, UniProt, Ensembl and ChEBI. (**c**) Different clients can use the OmicsDI API to retrieve data from the resource including the web interface and the ddiR package.

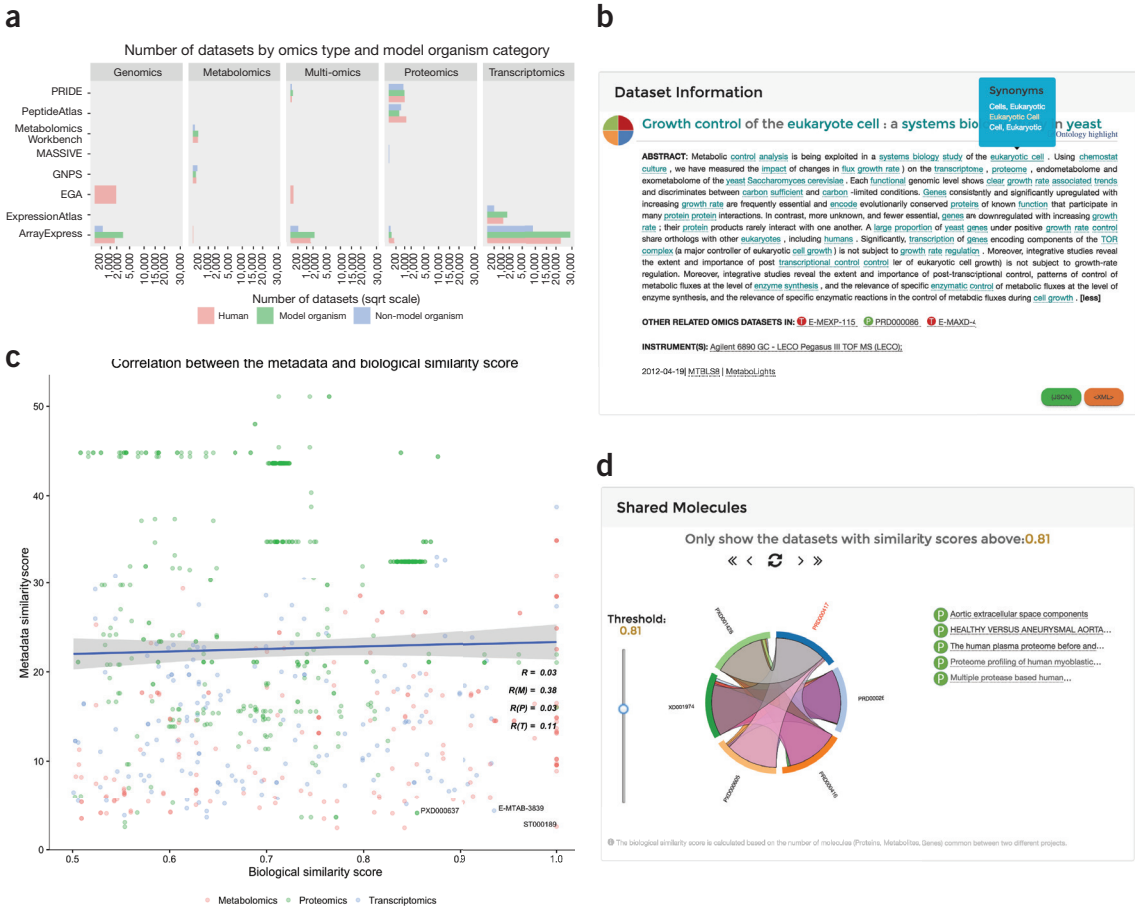

Figure 2 Distributions of OmicsDI data sets. (**a**) Distribution of data sets per omics type and organism category including model organisms, non-model organisms (excluding human) and human. (**b**) The data set view showing the other related omics data sets, including the ontology-highlighting option to extract the most relevant terms in the metadata. (**c**) Pearson-correlation plot of the metadata similarity score and the biological similarity score, across transcriptomics (T), proteomics (P) and metabolomics (M) data sets. (**d**) The shared molecules box shows all data sets with a biological similarity score of more than 0.5, with a slider allowing a user to increase the cutoff value (here set to 0.81).

different biological entities, including genes, transcripts, proteins, metabolites and the corresponding publications from PubMed.

To facilitate participation in OmicsDI by repositories and to enable the future integration of other omics fields (e.g., interactomics), we have developed a set of data integration guidelines and metadata requirements. The level of annotation required is flexible, and many repositories provide only a subset of the metadata included in our guidelines. Data with varying amounts of annotation can be made 'accessible' (the second FAIR principle[3]) in OmicsDI using a flexible metadata schema that classifies data sets as either mandatory, recommended or additional. A flexible exchange system based on the OmicsDI XML format and application programming interfaces (APIs) has been developed. Each repository needs to generate these file formats to join the OmicsDI platform. To facilitate integration, a stand-alone open-source Java tool has been developed (omicsDI XML validator). It allows the detection of metadata-related format errors as well as inconsistencies in the data set representation (**Supplementary Note 4**).

Different repositories use their own data models, metadata representation and identifiers, such as controlled vocabularies and ontologies. To address any 'interoperability' problems that arise (the third FAIR principle[3]), OmicsDI includes a metadata normalization and annotation expansion step for every data set that is integrated (**Fig. 1b**). These harmonization steps standardize experimental and technical metadata, the identifiers for the biological entities and the references to external resources. For example, for any publication named using the digital object identifier (DOI) or a citation, the matching PubMed identifier is inserted during harmonization (**Supplementary Notes 5** and **6**). If the name of an organism is provided using free-text the annotation step during harmonization converts it to a National Center for Biotechnology Information (NCBI) taxonomy identifier. Different data sets can include different terms for the same concept within the same context[8]; for example, a protein can also be referred to as a gene product. To overcome this type of problem, an ontology-based annotation expansion step is applied using the ontology tool Annotator[9], and every relevant phrase in the metadata (title, description, sample and protocols) and the corresponding publication (title and abstract) is enriched with the relevant synonyms, ontology and controlled vocabulary terms.

OmicsDI is a lightweight discovery tool that comprises >81,116 omics data sets (as of December 2016) from 11 different repositories and includes four omicsdata set types (67,361 transcriptomics, 6,281 proteomics, 8,093 genomics and 847 metabolomics). The number of data sets from human, model organisms and non-model organisms (excluding human) is uniformly distributed among repositories and omics types (**Fig. 2a**), highlighting the diversity of data sets. To the best of our knowledge, OmicsDI is the first resource that integrates data sets from different

omics fields and databases into one framework and web interface.

OmicsDI also extends FAIR's findable principle[3] by providing methods to find and link existing data sets. The annotation expansion step using synonyms enables users to find and associate data sets that cannot otherwise be found. For example, the proteomics data set PXD002530 can be found in OmicsDI with the search term 'side effects', whereas it cannot be found by searching PRIDE using that term. In PRIDE it is only possible to find that data set by inputting the term 'adverse effects', which was used in the original annotation of the data set. By indexing the biological entities information in OmicsDI, it is possible to find data sets in which the queried molecule has been reported without an exact matched term. For example, the Metabolomics Workbench data set ST000113 can be found in OmicsDI using the metabolite name 'Arg-[13C,15N]3', whereas the same search will not find ST000113 in Metabolomics Workbench.

OmicsDI links data sets by two methods. First, data sets are directly linked using explicit mentions in the metadata. If the data set is a reanalysis (e.g., PeptideAtlas data set) of a data set in a different member repository (e.g., PRIDE), a cross-reference in the OmicsDI XML is used to define this relation. This annotation can be provided by the original repository in the OmicsDI XML (e.g., PeptideAtlas) or can be inferred by OmicsDI during the annotation process. As of December 2016, the relations 'Reanalyzed by' and 'Reanalysis of' are already in use (**Supplementary Table 1**). This mechanism provides a direct link between data sets in different repositories.

Second, the publication associated with a data set can be used to link data sets that are deposited in different repositories. This enables the linking of data sets from different databases that are, however, part of the same multi-omics experiment (**Fig. 2b**) and presents them to the user as 'Other related omics data sets in'. As of December 2016, 4,476 data sets have been labeled by the OmicsDI annotation component as part of multi-omics experiments. Although still small (5% of all OmicsDI data sets), the number of multi-omics data sets is growing (**Supplementary Table 2**).

OmicsDI also uses the 'similar data set' concept (**Supplementary Note 7**). The concept of 'related article' has been applied in PubMed to explore topics[10]. In OmicsDI, similar data sets are computed at two different levels: metadata and biological entities. Both similarity levels are estimated by comparing the weighted term vectors of each data set using the dot (scalar) product. The distribution of the metadata similarity (**Supplementary Fig. 1**) and molecular similarity (**Supplementary Fig. 2**) are filtered depending on the distribution for each omics type. In this way, OmicsDI boosts the discoverability of related data sets that use similar analytical protocols or software (**Supplementary Fig. 3**), or share similar biological entities (**Supplementary Fig. 4**). To our knowledge, this enables for the first time the association of related data sets stored in different resources. For example, for the Expression Atlas data set E-GEOD-30999, OmicsDI reports 14 related data sets. In addition, the 'biological similarity' score computes the number of shared biological entities among data sets without taking into account additional metadata. For example, the same data set has five data sets with a biological similarity score >0.5 and one data set with a score of 0.7 (E-GEOD-41662). Of these, data set E-GEOD-41663 is not classified as related by the metadata-based similarity, although careful reading of the associated manuscripts reveals that E-GEOD-41663 used a subset of the samples of E-GEOD-30999. This example demonstrates the value of our approach. We determined the correlation between metadata and biological similarity scores for all OmicsDI data sets (**Fig. 2c**). The results showed no correlation ($R^2 = 0.03$) between both metrics across all types of omics data sets, with the highest correlation found in metabolomics approaches ($R^2 = 0.3$). For example, the data sets PXD000637 (PRIDE), ST000189 (Metabolomics Workbench) and E-MTAB-3839 (Expression Atlas) showed a higher biological similarity score (>0.85) and less than 5 of metadata similarity score (**Fig. 2c**). These results show that both scores are orthogonal metrics supporting discovery of related data sets through complementary methods.

The OmicsDI web interface provides different views, each of which focuses on a specific aspect of the data (**Supplementary Notes 8** and **9**). A metadata overview and access statistics provide a convenient entry point to browse a repository (**Supplementary Fig. 5**). Data sets can be searched and filtered based on annotations (e.g., species, tissue, disease), year of publication or repository. The result of each search displays all the relevant data sets sorted using a weighted scoring function (**Supplementary Fig. 6**). In addition, OmicsDI provides a data set page that includes a list of related publications and similar data sets (**Fig. 2b–d**). If the biological entities information is available for a given data set, a chord diagram presents the link to related data sets with high biological similarity scores (**Fig. 2d**). A web service interface, including a standard RESTful API to access the data programmatically, is also provided (http://www.omicsdi.org/ws). Related libraries and packages used for OmicsDI are also available at https://github.com/OmicsDI. For instance, an R-package called ddiR is provided, enabling data analysis (**Supplementary Note 10**).

OmicsDI exploits advances in metadata-based browsing to support data set findability. The original data sets are not replicated, but are referenced. In addition to fully open data sets, life science often produces valuable data sets containing personal identifiable genetic or phenotypic data. These data are deposited in controlled-access repositories, to which access is granted after application to a data access committee. However, the metadata of controlled-access repositories is accessible, and therefore OmicsDI can integrate data from EGA (the European Genome-Phenome Archive, the first controlled-access repository with open, searchable metadata). The responsibility for provision of well-formatted metadata lies with the original data providers (similar to the concept of publisher data provision in PubMed). OmicsDI displays and promotes the original data set identifiers, not only to avoid creation of another set of identifiers, but also to ensure attribution of credit to the original data providers. OmicsDI can integrate with large, broader scope efforts like the NIH BD2K DataMed (https://datamed.org) through shared a metadata format.

In conclusion, Omics DI provides an integrated search framework for data sets that introduces a range of modern features, such as access metrics and discovery of related data sets that we now take for granted in PubMed.

Editor's note: This article has been peer-reviewed.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This work has been supported by the US NIH BD2K grant U54 GM114833 and a National Natural Science Foundation of China grant (61501071). A.I.N. is supported by US National Institute of Health grant (R01-GM-094231). Y.P.-R. is supported by BBSRC 'PROCESS' grant (BB/K01997X/1). M.B. is supported by Projects of International Cooperation and Exchanges grant (2014DFB30010). M.W. is supported by an NIH grant (5P41GM103484-07). J.A.V. and N.d.-T. are supported by the Wellcome Trust (grant WT101477MA). T.T. is supported by the BBSRC 'ProteoGenomics' grant (BB/L024225/1). E.W.D. and D.S.C. are supported in part by grant (U24 AI117966-02S1). S.-A.S. is supported in part by US NIH BD2K grant (1U24AI117966-01). M.W. and N.Bandeira

were supported by NIH grant (5P41GM103484-07). N.Bandeira was also partially supported as an Alfred P. Sloan Fellow. S.Subramaniam is supported by NIH grants U01 DK097430 and U01 CA198941

AUTHOR CONTRIBUTIONS

H.H., Y.P.-R. and P.P. developed the OmicsDI concept. Y.P.-R. and M.B. designed and developed the web and annotation/enrichment framework. F.D.V.L., R.C.B. and Y.P.-R. developed the GPMDB reader. S.Squizzato, Y.M.P. and N.Buso developed the indexing system based on the EMBL-EBI (European Bioinformatics Institute) Search Server. Y.P.-R., E.F., M.S., S.Subramaniam, A.J.C., K.H., D.S., J.P., and M.W. developed the Metabolomics Workbench, Metabolome Express, MetaboLights, EGA and MassIVE readers and APIs, respectively. P.Z. helped to make the MetabolomeExpress schema OmicsDI-compatible. U.S., R.P. and M.K. contributed with the integration of ArrayExpress and Expression Atlas. N.d.-T., Y.P.-R. and T.T. developed the PRIDE reader and contributed to the web development. E.W.D., D.S.C., R.M.S., N.Bandeira, A.I.N., C.S., R.C.J., R.L. and J.A.V. contributed to the design of the system. Y.P.-R. and A.B. developed the ddiR package. Y.P.-R. and M.B., designed and implemented the biological similarity scoring system; and Y.P.-R. and A.B. performed the data analysis. Y.P.-R., J.A.V. and H.H. wrote the manuscript, with contributions from all authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Yasset Perez-Riverol[1,16], Mingze Bai[1–3,16], Felipe da Veiga Leprevost[4], Silvano Squizzato[1], Young Mi Park[1], Kenneth Haug[1], Adam J Carroll[5], Dylan Spalding[1], Justin Paschall[1], Mingxun Wang[6], Noemi del-Toro[1], Tobias Ternent[1], Peng Zhang[4,7], Nicola Buso[1], Nuno Bandeira[6], Eric W Deutsch[8], David S Campbell[8], Ronald C Beavis[9], Reza M Salek[1], Ugis Sarkans[1], Robert Petryszak[1], Maria Keays[1], Eoin Fahy[10], Manish Sud[10], Shankar Subramaniam[10], Ariana Barbera[11], Rafael C Jiménez[12], Alexey I Nesvizhskii[4], Susanna-Assunta Sansone[13], Christoph Steinbeck[1], Rodrigo Lopez[1], Juan A Vizcaíno[1], Peipei Ping[14,15] & Henning Hermjakob[1,3]

[1]*European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, United Kingdom.* [2]*Institute of Bioinformatics, Chongqing University of Posts and Telecommunications, Chongqing, China.* [3]*Beijing Proteome Research Center, National Center for Protein Sciences Beijing, Beijing, China.* [4]*Department of Pathology, University of Michigan, Ann Arbor, Michigan, USA.* [5]*Research School of Biology, Australian National University, Canberra, Australian Capital Territory, Australia.* [6]*Department of Computer Science and Engineering, University of California, San Diego, La Jolla, California, USA.* [7]*Commonwealth Scientific and Industrial Research Organization, Canberra, Australian Capital Territory, Australia.* [8]*Institute for Systems Biology, Seattle, Washington, USA.* [9]*Biochemistry & Medical Genetics, University of Manitoba, Winnipeg, Manitoba, Canada.* [10]*Department of Bioengineering, University of California, San Diego, La Jolla, California, USA.* [11]*Department of Medicine, University of Cambridge, Cambridge, United Kingdom.* [12]*ELIXIR Hub, Wellcome Genome Campus, Hinxton, United Kingdom.* [13]*Oxford e-Research Centre, University of Oxford, Oxford, United Kingdom.* [14]*Department of Physiology and Department of Medicine, Division of Cardiology, David Geffen School of Medicine at UCLA, University of California, Los Angeles, Los Angeles, California , USA.* [15]*Department of Medicine, Division of Cardiology, David Geffen School of Medicine at UCLA, University of California, Los Angeles, Los Angeles, California , USA.* [16]*These authors contributed equally to this work.*
e-mail: yperez@ebi.ac.uk or hhe@ebi.ac.uk

1. Bourne, P.E., Lorsch, J.R. & Green, E.D. *Nature* **527**, S16–S17 (2015).
2. Perez-Riverol, Y., Alpi, E., Wang, R., Hermjakob, H. & Vizcaíno, J.A. *Proteomics* **15**, 930–950 (2015).
3. Wilkinson, M.D. *et al. Sci. Data* **3**, 160018 (2016).
4. Prins, P. *et al. Nat. Biotechnol.* **33**, 686–687 (2015).
5. Bourne, P.E. *et al. J. Am. Med. Inform. Assoc.* **22**, 1114 (2015).
6. NCBI Resource Coordinators. *Nucleic Acids Res.* **44**, D7–D19 (2016).
7. Europe PMC Consortium. *Nucleic Acids Res.* **43**, D1042–D1048 (2015).
8. Blake, J. *Nat. Biotechnol.* **22**, 773–774 (2004).
9. Shah, N.H. *et al. BMC Bioinformatics* **10** Suppl 9, S14 (2009).
10. Lin, J. & Wilbur, W.J. *BMC Bioinformatics* **8**, 423 (2007).

Firmiana: towards a one-stop proteomic cloud platform for data processing and analysis

To the Editor:

Improvements in next-generation proteomics, including instrumentation[1,2], sample preparation, and computational analysis[3,4], have generated large amounts of data that cover protein profiling, post-translational modifications, and protein–protein interactions[5,6]. The first draft of the human proteome, for example, made use of 2,000 (ref. 6) and 16,000 (ref. 5) raw files. Proteomics now calls for a uniform online pipeline that can host millions of data sets with the same quality standards, analyze hundreds to thousands of experiments, and integrate multi-dimensional omics data for knowledge mining and hypothesis generation to disseminate proteomics to the scientific community. Here, we describe Firmiana (V1.0) (http://www.firmiana.org/), a one-stop proteomic data processing and integrated omics analysis cloud platform that allows scientists to deposit mass spectrometry (MS) raw files, perform proteome identification and quantification online, carry out bioinformatics analyses, extract knowledge, and visualize results using a biologist-friendly web interface without the need for programming expertise.

Current major proteomic platforms, including MaxQuant[3], SearchGUI[7], PeptideShaker[8], Perseus[9], OpenMS[10], PEAKS[11], Proteomics DB[5], ProteomeXchange[12], PRIDE[13] (member of ProteomeXchange consortium), MassIVE[14] (member of ProteomeXchange consortium), ProteoSAFe[14], PeptideAtlas[15], Trans-Proteomic Pipeline[16] (TPP), Galaxy-P[17], and Chorus[14], were built to serve at least one of the following functions: data repository, data processing (identification/quantification), and data analysis and knowledge mining.

Proteomics DB, ProteomeXchange, PRIDE, MassIVE, PeptideAtlas, and Chorus were developed mainly for exchange of MS data; they do not support identification and quantification, or data analysis and knowledge mining. MaxQuant, OpenMS, TPP, PEAKS, ProteoSAFe, SearchGUI, and Galaxy-P are excellent tools for MS data analysis, but they do not provide data repository and knowledge mining. PeptideShaker and Perseus support knowledge mining, but lack data repository and data processing capabilities (**Supplementary Fig. 1**).

Firmiana is a workflow based on the Galaxy system, which aims to facilitate for users the entire process of bioinformatics analysis from raw MS data to biological knowledge generation. It consists of multiple functional modules, including user login interface, metadata, identification and quantification, data analysis, and knowledge mining (**Fig. 1a**).

The first module is the main interface. Users apply for a laboratory account and have full authority to manage their own data (**Fig. 1a** and **Supplementary Fig. 2**), including deposition of MS files, inputting metadata, execution of bioinformatics analyses, and visualization and exporting results. All matched spectra, peptide hits, and quantitative protein IDs can be explored and annotated online. The result table includes a variety of parameters for interpreting the proteome (**Supplementary Fig. 3**) and users