10-9-2020

Argonaut: A web platform for collaborative multi-omic data visualization and exploration

Dain R Brademan

Ian J Miller

Nicholas W Kwiecien

David J Pagliarini

Michael S Westphall

See next page for additional authors

Authors

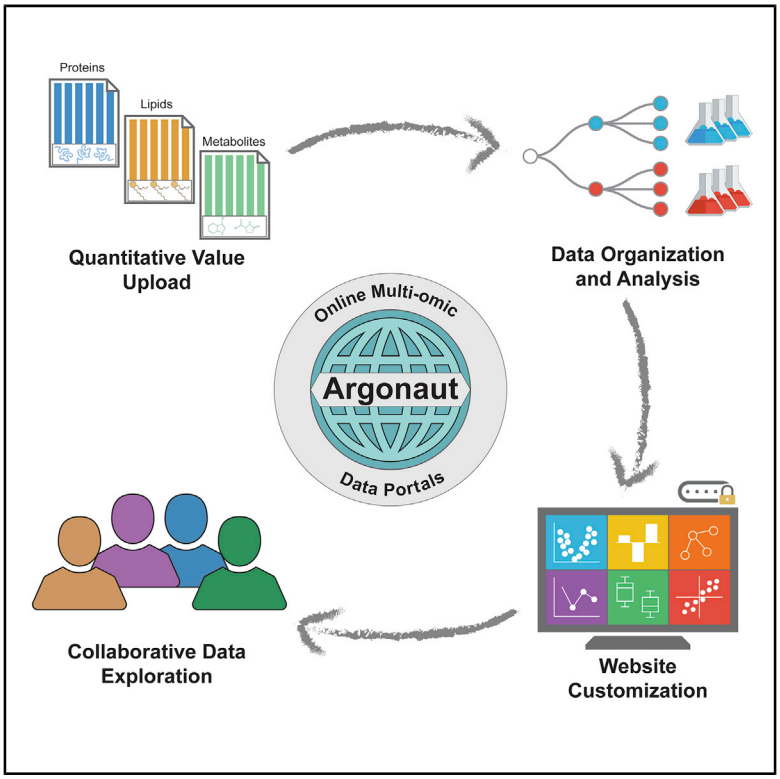
Dain R Brademan, Ian J Miller, Nicholas W Kwiecien, David J Pagliarini, Michael S Westphall, Joshua J Coon, and Evgenia Shishkova

Descriptor

Patterns

Argonaut: A Web Platform for Collaborative Multi-omic Data Visualization and Exploration

Graphical Abstract



Highlights

- We present Argonaut, an online platform for collaborative exploration of omic data
- Argonaut can be implemented with little programming expertise to create data portals
- Data are presented through intuitive, interactive visualizations via popular plots
- Data portals can be securely shared with other team members worldwide

Authors

Dain R. Brademan, Ian J. Miller, Nicholas W. Kwiecien, David J. Pagliarini, Michael S. Westphall, Joshua J. Coon, Evgenia Shishkova

Correspondence

coon@wisc.edu (J.J.C.), shishkova@wisc.edu (E.S.)

In Brief

High-throughput biomolecule profiling experiments have become more routine as quantitative technologies come to maturity. However, challenges in interpreting and broadly disseminating the generated biological "Big Data" sets have not been comprehensively addressed. We have developed Argonaut, a simple, code-free, and user-friendly platform for creation of interactive data-hosting websites. Argonaut conducts real-time statistical analysis of measured biomolecules, visualizes data using popular plots, and can be securely shared with and explored by collaborators across the globe.

Brademan et al., 2020, Patterns *1*, 100122
October 9, 2020
https://doi.org/10.1016/j.patter.2020.100122

CellPress

Patterns

CellPress
OPEN ACCESS

Descriptor

Argonaut: A Web Platform for Collaborative Multi-omic Data Visualization and Exploration

Dain R. Brademan,[1,2] Ian J. Miller,[3] Nicholas W. Kwiecien,[1] David J. Pagliarini,[2,3,4,5,6] Michael S. Westphall,[3] Joshua J. Coon,[1,2,3,*] and Evgenia Shishkova[3,7,*]

[1]Department of Chemistry, University of Wisconsin–Madison, Madison, WI 53706, USA
[2]Morgridge Institute for Research, Madison, WI 53715, USA
[3]Department of Biomolecular Chemistry, University of Wisconsin–Madison, Madison, WI 53706, USA
[4]Department of Cell Biology and Physiology, Washington University School of Medicine, St. Louis, MO 63110, USA
[5]Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, St. Louis, MO 63110, USA
[6]Department of Genetics, Washington University School of Medicine, St. Louis, MO 63110, USA
[7]Lead Contact
*Correspondence: coon@wisc.edu (J.J.C.), shishkova@wisc.edu (E.S.)
https://doi.org/10.1016/j.patter.2020.100122

THE BIGGER PICTURE Modern systems biology experiments profile thousands of biomolecules across many experimental conditions to generate insights about the biological system. While data collection for these experiments can be routine, interpretation of the resultant datasets often requires interlaboratory collaboration of scientists with diverse expertise and is hindered by challenges inherent to sharing and exploring "Big Data." We have developed Argonaut, a web-based platform purpose-built to accommodate large-scale, multi-omic experiments and to enable intuitive and interactive exploration of the associated data. Argonaut presents the experimental results in an online code-free environment, empowering both experts and non-experts worldwide to easily interact with and share the data. Our platform aims to streamline derivation of impactful experimental conclusions by overcoming the hurdles of working with large datasets and lowering the barrier to entry for biological and clinical collaborators.

Production: Data science output is validated, understood, and regularly used for multiple domains/platforms

SUMMARY

Researchers now generate large multi-omic datasets using increasingly mature mass spectrometry techniques at an astounding pace, facing new challenges of "Big Data" dissemination, visualization, and exploration. Conveniently, web-based data portals accommodate the complexity of multi-omic experiments and the many experts involved. However, developing these tailored companion resources requires programming expertise and knowledge of web server architecture—a substantial burden for most. Here, we describe Argonaut, a simple, code-free, and user-friendly platform for creating customizable, interactive data-hosting websites. Argonaut carries out real-time statistical analyses of the data, which it organizes into easily sharable projects. Collaborating researchers worldwide can explore the results, visualized through popular plots, and modify them to streamline data interpretation. Increasing the pace and ease of access to multi-omic data, Argonaut aims to propel discovery of new biological insights. We showcase the capabilities of this tool using a published multi-omics dataset on the large mitochondrial protease deletion collection.

INTRODUCTION

Multi-omics is a powerful and versatile approach for probing biological systems. Encompassing many layers of biological information, multi-omics data can holistically describe a living system and its response to perturbations, as metabolites, lipids, and proteins co-function to orchestrate responses to various stimuli.[1,2] Recent advances in mass spectrometry (MS) profiling

technologies have revealed this coordination by enabling simultaneous measurement of multiple molecular classes.[3–9] Specifically, improvements in experimental throughput of multi-omic analyses have opened the door to large-scale MS-based profiling studies, where the analysis of diverse biomolecules in many samples under dozens of different conditions is considered nearly routine.[10–16]

The rapid creation of these large and complex datasets, however, presents a new challenge: quickly processing raw MS data into sets of quantified biomolecules and extracting rigorous biological insights from these results. To this end, tools for processing mass spectral data—primarily proteomic data, such as Perseus and MSstats[17,18]—have enabled a number of analyses and visualizations. Nonetheless, major challenges persist: (1) designed for use by MS experts, these tools require both a thorough understanding of statistics and knowledge of common nuances in MS data; (2) because data processing is not fully streamlined, considerable hands-on and potentially taxing interaction with large datasets is required, and (3) these tools' tabular outputs are not conducive for dissemination to and exploration by a non-expert user base. Making results accessible to a broader scientific community is essential to realizing the full potential of biological MS, particularly as MS technologies become increasingly application-driven and therefore collaborative.[19,20]

Online data analysis and visualization tools have become increasingly popular in other areas of science as they stand to alleviate many of the issues associated with analysis and communication of large datasets.[21–23] These online tools also avoid issues commonly associated with software distribution, eliminating the need for version control by centralizing the software to a standardized web server environment. Functional web-based utilities thus provide an efficient means to share results with collaborators, minimize the challenges of data transfer between laboratories, and improve scientific discussion. In fact, to augment dissemination of study findings, many large-scale resource projects feature tailored companion websites that facilitate interactive data exploration.[11,24–26]

Although ideal, such custom web-based interfaces are tedious and time-consuming to develop—even for a single research project. Construction of these tools requires programming experience and familiarity with web server architecture. Recently, Torre et al. presented BioJupies, a web-based utility that greatly augmented the analysis and distribution of transcriptomic data.[27,28] Other research groups have released web applications that facilitate online exploration and sharing of MS datasets.[29–34] The next generation of tools should be available to non-programmers, able to convert general multi-omics MS data into a cloud-friendly format, comprehensively interfaced with interactive visualizations, and sharable with collaborators for intuitive hands-on exploration.

To fulfill this need, we present a new platform called Argonaut. Our tool enables rapid and codeless generation of MS data exploration portals, allowing users to create project-specific websites hosted on standalone web servers using the Docker environment.[35] We describe this process in detail below, demonstrating its use with a large multi-omic dataset generated by a study on yeast mitochondrial protease deletion. In brief, users upload their quantitative data (formatted in simple generic spreadsheets) directly to a browser. Argonaut then performs on-the-fly statistical analyses of those data and allows users to select several interactive visualizations, which are automatically embedded into the custom website. Once created, the data portals can be securely shared with researchers worldwide in just a few clicks.

RESULTS

Creation of the Multi-omic Data Portal Argonaut

Overview

Argonaut allows users to easily generate standalone, password-protected websites containing analyses and visualizations of their MS data (hereafter "data portals"). Argonaut contains JavaScript, PHP code, and all necessary files and architecture to run an interactive data portal on a public or private web server. Each data portal holds project descriptors, interactive visualizations, and detailed annotations describing the types of data inside the project. Individual data portals can be created in a "containerized" Docker environment, allowing users to save and export snapshots of a fully functional data portal at any time.[35]

The portal creation process can be completed through a series of intuitive steps (Figure 1). First, a new data portal is initialized using Docker. Then, the project owner can log into the data portal through their preferred web browser and begin to customize the newly created portal by providing a project title and description and uploading hierarchically organized quantitative data in accordance with their experimental design. The upload procedure allows for experimental or technical replicates to be easily grouped into separate branches under experimental treatment and an ome classifier. Argonaut utilizes an HTML upload page that accepts files containing quantitative data in a post-processed form, e.g., tabular sets of biomolecule abundances. While many pipelines use a variety of standardized file formats to store data,[36,37] Argonaut supports solely tabular, text-based spreadsheets; thanks to their simplicity and flexibility, many search algorithms and processing pipelines are capable of exporting results in this format. Following the initial data upload, users can select individual visualizations to add to their custom web portal from a menu of options, such as volcano plots and correlations, among many others. Based on these selections, Argonaut constructs a complete data exploration webpage with all associated functionality embedded. These custom web portals can then be shared with other researchers—at the discretion of the creator—via a tiered-permission sharing scheme.

Use with Testbed Dataset

To demonstrate how our tool creates interactive data portals for improved analysis and exploration, we acquired data from a multi-omics study investigating the biological functions of mitochondrial proteases and their substrates in *Saccharomyces cerevisaie*.[38] In this study, 19 single-gene deletion yeast strains and a control wild-type strain were analyzed in biological triplicate under two growth conditions for a total of 120 unique samples per ome. We reasoned that this rich dataset with validated biological insights could serve as a suitable and sufficiently challenging testbed for our tool. The publicly hosted data portal containing this dataset can be accessed at https://coonlabdatadev.com with the username "guest" and password "password."

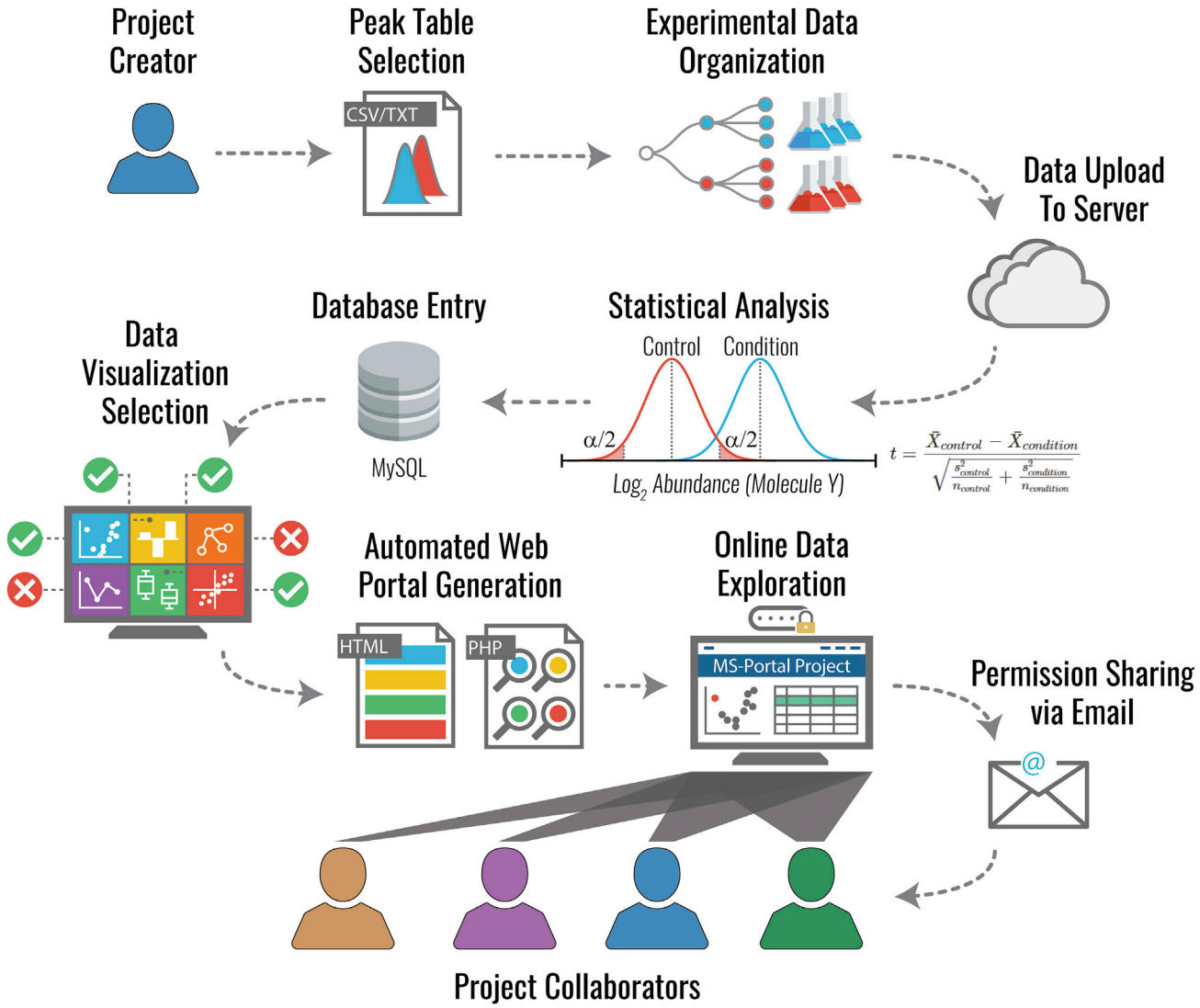


Figure 1. The Argonaut Workflow

Argonaut is designed as a portable platform to share multi-omics data in an online environment using customizable interactive visualizations. Processed quantitative measurements from case/control-style experiments are uploaded to the online platform in a variety of text-based formats. Uploaded data are then categorized according to the uploader's experimental design. Common data transformations, such as missing value imputation, filtering missing values, control normalization, or $\log_2$ transformations, can be conducted. Inferential statistics are used to determine the significance of molecular perturbations. Data portals can be customized in a variety of ways, allowing detailed project and data descriptions, selection of visualization options, and project management. Data portal access can also be shared directly with collaborators using a secure permission sharing scheme, allowing multiple laboratories to concurrently explore large datasets to rapidly generate biological insight.

The Veling et al. study[38] identified and quantified >3,000 biomolecules. Highlighting the compatibility of this dataset with our platform, we converted the abundance measurements of these biomolecules into three comma-separated value tables, each corresponding to one of the omes profiled in the study, i.e., the proteome, metabolome, and lipidome. The columns of the files included non-redundant biomolecule identifiers (e.g., UniProt, KEGG, or HMDB), unique names of experimental conditions, including names of the strains, condition and replicate information, and optional sample metadata (alternative biomolecule names, FASTA headers, etc.), as illustrated in Figure S1A. A category of information in the columns must be specified during the data upload. The rows contained quantitative values of each biomolecule in the respective sample conditions. An in-depth description of file structure along with example quantitative files can be found in the Supplemental Information and at Argonaut's GitHub.

As in many other large-scale studies, samples in this study were processed in experimental batches with a designated batch control (the wild-type strain) and included biological replicates. To accommodate this common experimental design, Argonaut utilizes a tree-based hierarchy to organize replicates of experimental conditions in a batch-based format (Figure S2B). Within each branch, samples are grouped to determine their

relationship to the rest of the samples within the uploaded dataset, including designation of replicate sets. Here, three replicates of the wild-type strain were denoted as batch controls, and average molecular abundances in the three replicates of each deletion strains were normalized to those in the selected control. (Note if the batch control is not specified, Argonaut automatically normalizes the condition measurements to the population mean.)

During data processing, the significance of molecular perturbations between the control and the experimental conditions are calculated on-the-fly using homoscedastic two-sided t test and stored in a database for later querying. Correction for multiple hypothesis testing is available upon request using either the Bonferroni or Benjamini-Hochberg procedures. Once data are processed, the project creator can navigate to a list of predefined options to choose which visualizations and analyses are presented to the portal's users. If downstream gene ontology (GO) enrichment is desired and the uploaded file includes GO-compatible unique identifiers, GO enrichment analysis can be enabled (as it was here) by specifying the sample organism, the column containing the unique identifier, and the identifier type.

As a completed study, Veling et al.[38] had already conducted data transformations and filtered quantitative values. Note, however, that users can enable these operations during upload. For example, raw quantitative values can be $\log_2$ transformed to facilitate fold-change visualizations; missing values can be imputed utilizing a modified left-censored imputation algorithm (Figure S2); or biomolecules can be filtered from downstream analysis if their abundance measurements are missing in a user-defined number of samples (refer to Experimental Procedures, Table S1, or the GitHub repository for more details).

Interactive Data Examination

The tree-based architecture of Argonaut is well suited to multi-omics datasets because it segregates data from a single experiment into distinct branches (e.g., proteomics, lipidomics, and metabolomics). When exploring the data, users can rapidly switch between branches to compare trends in biomolecular abundances across samples and different biomolecule classes. On entering the data portal, users are presented with an overview, highlighting each branch (i.e., "ome") of the project to briefly summarize the number of experimental conditions, replicates, quantified biomolecules, and the average biomolecular coefficient of variation. From there, users can navigate to data visualization tabs to explore their dataset through six staple bioinformatic analyses (shown in Figure 2): volcano plots, principal-component analysis (PCA), condition-condition correlation, bar charts of biomolecule abundances, GO enrichment, and the outlier analysis. All visualizations are generated using the JavaScript library D3.js, which enables real-time customization and interactivity.[39] Significance thresholds can be modified by the user, and many plots support data point lookups by unique identifiers. Any visualization can be downloaded in scalable vector graphics format, permitting easy integration into publication-quality figures, such as Figure 2. In addition, any uploaded or processed data used to generate the visualizations can be exported from the data portal. All interactive visualization options can be inspected on our publicly hosted portal. Further details explaining each visualization can be found on the Argonaut GitHub wiki.

Using our platform, we rapidly recapitulated and visualized several key findings of the Veling et al. study[38] using their quantitative data. All visualizations contained in Figure 2 are representative of a selection of main text and supplemental figures from the Veling et al. manuscript. Volcano plots are used in many studies to quickly highlight large biomolecule perturbations resulting from experimental conditions.[15,40] For example, the volcano plot in Figure 2A demonstrates the upregulation of the iron sulfur assembly protein ISA1 in the absence of the *pim1* gene, revealing a novel relationship between the two proteins in the Veling et al. study. PCA is commonly used to segregate groups of samples or explore origins of variance in a dataset.[41,42] Veling et al. used PCA (2B) to demonstrate clear separation between the respiration-deficient and respiration-competent deletion strains. Biomolecule abundance correlation analysis is often used in gene deletion experiments to (2C) demonstrate a functional relationship between two gene deletion strains.[11,41,43] Veling et al. show the highly correlated biomolecular profiles of two strains with deleted genes encoding inner membrane proteases IMP1 or IMP2.

Visualizing a biomolecule's abundance across experimental conditions can quickly highlight expression trends and relationships.[44] Figure 2D visualizes the $\log_2$ fold-change abundance of 3-hexaprenyl-4-hydroxybenzoate across all respiration-competent deletion strains and uncovers an increase in the Oct1 mitochondrial protease deletion strain, as reported in the Veling et al. study. GO analysis is an established technique to determine enrichment for specific biological functions in transcriptomic and proteomic datasets.[45] Veling et al. demonstrated that the term "ATP synthesis coupled protein transport" (2E) recapitulates known relations between the mitochondrial protease PIM1 and ATP. The outlier analysis used in Argonaut quickly pinpoints experimental conditions where a biomolecule is characteristically and uniquely up- or downregulated.[11] Veling et al. used an outlier analysis (2F) to reveal the Oct1 mitochondrial protease deletion condition to be a significant outlier regarding the abundance of 4-hydroxybenzoate, a cytosolic precursor to coenzyme Q. These and potentially many other novel biological insights, hidden in the Veling et al. dataset, are readily accessible for exploration by all portal users. As the website navigation is intuitive, engaging with the hosted data requires almost zero prior guidance from the project creator, lowering barriers to entry into the world of omics data for non-experts.

Collaborative Data Exploration

To facilitate the collaborative aspect of Argonaut, we have developed a three-tier accessibility scheme that allows for flexible utilization of a generated data portal's functionality. The first tier provides read-only access to a portal, allowing users with this permission level to view the created portal, interact with the visualizations, and download the processed data. The second tier upgrades the user accesses with additional edit permissions, which allow addition, removal, and editing of the uploaded data. The third permission tier corresponds to that of the project creator, allowing the users to invite collaborators, select visualizations, or to delete the portal website entirely. Only project owners and tier three users have comprehensive discretion

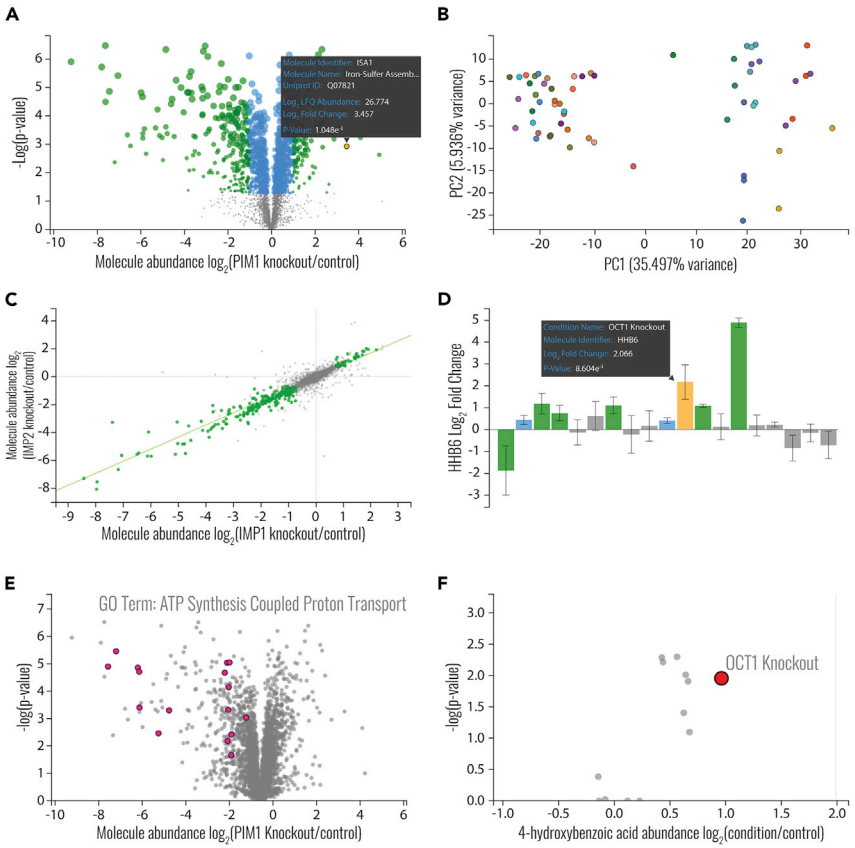


Figure 2. Bioinformatic Analyses Visualized by Argonaut

The website generated using the multi-omic data from the Veling et al. study[38] features a set of six analyses that are commonly used in omics experiments. All visualizations are fully interactive and generated on-demand using queries from the uploaded data. Significance and fold-change thresholds for data highlighting can be adjusted as desired. Visualizations and data can be exported from the portal as vector graphics, such as the ones used to produce this figure, and text-based spreadsheets, respectively.

over access to the portal data. Upon initial data portal creation, the portal creator is the only user with permission to access the site, but with only a few clicks, they can send any collaborator an automatic invitation via email, which grants the new user access to the associated project with the predefined permission level. As with the Veling et al. dataset here, upon publication project-specific data portals can be made public through the creation of generic usernames and passwords with the tier one permission level. This approach enables online exploration of the study findings worldwide, while preserving the integrity of the underlying dataset.

DISCUSSION

As MS researchers increasingly leverage "Big Data" offered by high-throughput studies to answer complex biological questions, multi-omics has become a particularly powerful approach—one that generates deep, multi-faceted descriptions of the biological systems. However, tools to quickly interpret the results of multi-omics experiments across laboratories have not kept pace. Furthermore, in the absence of means for enhancing data accessibility, the discovery potential of rapidly evolving MS techniques remains untapped for many researchers unfamiliar with systems biology data.

We thus developed Argonaut to provide the scientific community with a much-needed tool for the analysis of complex, data-rich resources. To the best of our knowledge, there is no other online platform enabling users to compile their multi-omics data into a single resource and present it in the easy-to-explore manner offered by Argonaut. The codeless generation of web portals for data analysis, visualization, and sharing makes this tool uniquely accessible. The only processing steps required to take advantage of the platform are fold-change normalization and statistical testing. These requirements are compatible with the most generic batch-based experimental designs that contain biological or technical replicates. For additional flexibility, the platform allows users to conduct different data filtering, missing value imputation, or data transformation operations external to Argonaut, if desired. Note that due to the considerable computational overhead, Argonaut does not currently support processing of raw MS data. We also elected to decouple Argonaut from any "searching" operations such as those offered by the TransProteomic Pipeline.[46] In doing so, we aim to keep our platform lightweight to increase performance for both data retrieval and visualization and to widen the platform's utility for the broadest MS and omics community.

Although the analyses conducted using Argonaut are not exhaustive or exclusive to our platform, we believe the ability to securely share experimental omics data in a unified and intuitive format is transformative. By encouraging broad data sharing among the research community, Argonaut is directed to two goals. First, to leverage the expertise of individual researchers from different fields, it allows data portals to be hosted on a public server as companion resources for manuscripts using a few simple Docker commands. Second, to sharpen the significance of novel biological findings, our tool allows multiple portals from our platform to be hosted in a singular location that permits facile comparisons across multiple datasets. Indeed, Argonaut is an agnostic platform that can be used to host the transcriptomic, epigenetic, and phenotypic data that are often generated in the course of comprehensive multi-omic studies. While researchers with the relevant expertise may adapt Argonaut to serve specific projects, applications, or frameworks, for the broader scientific community it provides a stable platform for teams of researchers to concurrently conduct in-depth analyses of their datasets and readily share their data in an intuitive, accessible format.

EXPERIMENTAL PROCEDURES

Resource Availability

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Evgenia Shiskova (shishkova@wisc.edu).

Materials Availability

This study did not generate any new unique reagents or materials.

Data and Code Availability

The example Argonaut data portal can be accessed at https://coonlabdatadev.com using the username "guest" and the password "password." Argonaut's source code and guides can be found at https://github.com/coongroup/Argonaut. Several example upload datasets are included in Data S1. The Argonaut Docker image is made freely available at https://hub.docker.com/r/coonlabs/argonaut.

Method Details

The Argonaut Website

The Argonaut platform is served using a Dockerized Linux, Apache, MySQL, and PHP (LAMP) web server. The client-side platform was built using the HTML templating framework Bootstrap (3.0) and the Angular JavaScript framework (1.3). Server-side scripts written in PHP (7.0) conduct database operations and relay data from the server to client. On container initialization, a blank data portal is assembled with a predefined administrative user account for data portal management. When an administrative user makes edits to a data portal's architecture (e.g., name, project description, or visualization options), the server utilizes the new data portal settings and embeds the new settings together with required HTML templates (Figure S1B) to generate an updated data portal. Only administrative user accounts are permitted to upload data, edit project architecture, or invite new users. A running data portal can be accessed by any web browser capable of communicating with the Docker machine.

Docker Image

The Docker container consists of a base Ubuntu image (16.04) with an Apache web server (2.4.33), a MySQL relational database (1.6), and the server-side scripting language PHP (7.2). The *PHP Data Object* extension is used for abstracted database accession and automatic query sanitization. The MySQL database contains 47 tables that serve to rapidly store and retrieve user-submitted data. To enable easier project database management, phpMyAdmin, a common web server administration platform, is installed to enable database management (phpMyAdmin, https://www.phpmyadmin.net). Details on how to access the Docker phpMyAdmin administrator account can be found in the Argonaut wiki.

Submitting Data to Argonaut

Administrative users can submit text-based spreadsheets containing quantified biomolecules from an experiment using the *angular-file-upload* directive (https://github.com/nervgh/angular-file-upload). Argonaut is capable of parsing text-based quantitative spreadsheets delimited by tabs. For example, when a new text file is uploaded, the file is temporarily saved and the column headers from the file are extracted. A selection of these column headers can be assigned as either unique identifiers, metadata, or quantitative values (Figure S1A). The user must also indicate which quantitative columns belong to the same condition (i.e., are experimental replicates). The uploaded data are then organized into a hierarchical structure to bin the uploaded experimental replicates into experimental conditions, experimental batches, and branches (Figure S2B). Branches can be used to separate data generated from the measurement of different biomolecule classes (i.e., different omes). Each uploaded file is denoted as by the keyword batch (i.e., batch of samples), and as such, missing value imputation, control normalization, and $\log_2$ transformation is conducted on each uploaded set of quantitative values independently. If standard molecular identifiers are included in the uploaded file's metadata, specifying the sample organism, the column containing the standard identifier, and the type of standard identifier can enable optional downstream GO enrichment analysis. UniProt identifiers are currently supported for GO enrichment analysis, although support for other identifiers will be added in the future.

Once all settings are finalized, the data's tree-based hierarchy is presented to the user for review, and raw quantitative values are then uploaded to the database to begin data processing. Submitted files are queued for processing using the PHP *Client URL* library to enable the asynchronous processing of concurrent file uploads and appropriately meter computational resources. After data processing, the uploaded spreadsheet is preserved on the server to allow retrieval of the stored data at any time. Examples of Argonaut-compatible files can be found in the Supplemental Information, downloaded from the Argonaut GitHub, or downloaded from the example portal provided at: https://coonlabdatadev.com using the username "guest" and the password "password."

Quantification and Statistical Analysis

Data Organization

When a new set of experimental data is uploaded, the server begins a multistage process to group the user-provided quantitative measurements into conditions, conduct an optional $\log_2$ transform for raw quantitative values, optionally filter and/or impute missing values, and conduct significance testing and other analyses. Initially, the raw quantitative values are loaded into memory. Any non-numeric or negative entries found in the quantitative value columns are initially set to 0. Biomolecular identifiers are checked for uniqueness. If duplicates in the unique identifier values are found, the duplicates are appended with an additional text qualifier. The experimental condition classifiers provided by the user are used to group experimental replicates into conditions using custom PHP objects. These grouped experimental replicates then undergo optional data filtration and missing value imputation.

Data Filtering and Imputation

Before statistical testing can be conducted, the dataset first must be considered complete, meaning there can be no missing values. The best strategy to account for missing quantitative values is an active area of debate regarding large-scale MS profiling experiments.[47,48] To provide a one-size-fits-most solution to this issue, Argonaut provides functionality for the user to remove sparsely quantified biomolecules that are missing in a user-specified proportion of the experimental replicates and in addition offers a left-censored missing value imputation algorithm. Alternatively, data filtering and missing value imputation can be conducted externally to Argonaut if other data-cleaning approaches are more appropriate for a particular dataset.

Missing values in MS profiling experiments often arise from low-abundance molecular species below the limit of quantification.[47,48] Argonaut's missing value imputation algorithm is adapted from the imputation strategy implemented in Perseus,[49] where a set of the smallest quantitative values are leveraged to impute missing data. Argonaut's imputation algorithm is visualized in Figures S2A and S2C. For each condition, $\log_2$ transformed quantitative measurements are placed into an array and placed in ascending order, generating a roughly normal distribution of quantitative values. An iterative loop is used to subset the smallest x% of existing quantitative values, ($1\% \leq x \leq 100\%$), as demonstrated in Figure S2C. A Gaussian distribution is drawn for the mean and standard deviation of the subset data. This Gaussian is then randomly sampled to populate all missing values for this condition, and the average biomolecule coefficient of variation (CV) within the condition is calculated for the cutoff x. The ideal cutoff x is selected by minimizing the average biomolecular CV. This calculation is then iteratively applied to all other conditions. If data filtration and missing imputation are not conducted and missing values remain, they will be excluded from further statistical analysis.

Normalization and Statistical Testing

After data filtration and missing value imputation are completed, the mean and standard deviation of all remaining quantified biomolecules are calculated within each condition. Calculated means then undergo a linear control normalization to better scale raw abundances for visualizations. If a control condition was not specified for an uploaded batch, this calculation uses the mean abundance across all conditions for a biomolecule instead.

These quantitative data then undergo statistical testing against the newly uploaded batch's control if one was specified. Otherwise the testing is conducted against the batch's $\log_2$ transformed average biomolecular abundance for each respective biomolecule. The test statistic of differential biomolecule expressions are calculated using an unpaired two-tailed Student's or Welch's t test.

The test statistic is converted to a p value and stored in the MySQL database. Multiple hypothesis corrected p values are then calculated using both the Bonferroni[50] and Benjamini-Hochberg[49] correction methods. These

corrected p values can be selected for use in Argonaut's interactive visualizations to enable users to be more stringent in what is labeled as statistically significant. Finally, outlier analysis and PCA are conducted as described previously by Stefely et al.[11]

Uploaded proteomics datasets that contain either UniProt identifiers or systematic gene names can undergo GO enrichment analysis. A contingency table is derived from the quantified biomolecules in an experimental condition using user-specified fold-change and p value cutoffs. GO enrichment statistical testing is conducted using a Fisher's exact test followed by a Benjamini-Hochberg correction for multiple hypothesis testing correction. A list of supported organisms for GO enrichment can be found at the Argonaut GitHub wiki.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.patter.2020.100122.

ACKNOWLEDGMENTS

We gratefully acknowledge support from NIH grants P41 GM108538 and R35 GM131795. D.R.B. was supported by an NHGRI training grant awarded to the Genomic Sciences Training Program (5T32HG002760).

AUTHOR CONTRIBUTIONS

Software Development, D.R.B., I.J.M., and N.W.K.; Formal Analysis, D.R.B., E.S., and I.J.M.; Funding Acquisition, J.J.C.; Investigation, D.J.P.; Writing – Original Draft, D.R.B., I.J.M., J.J.C., and E.S.; Conceptualization, N.W.K., D.R.B., M.S.W., E.S., and J.J.C.; Supervision, J.J.C., E.S., and M.S.W.; Validation, D.J.P.; Visualization, D.R.B. and E.S.

DECLARATION OF INTERESTS

N.W.K., M.S.W., and J.J.C. filed a patent, entitled "Web-Based Data Upload and Visualization Platform Enabling Creation of Code-Free Exploration of MS-Based Omics Data" (US20190034047A1; status 9.6.2020 "Pending"), related to the work described in this manuscript. The other authors declare no competing financial interest.

Received: May 28, 2020
Revised: August 15, 2020
Accepted: September 14, 2020
Published: October 9, 2020

REFERENCES

1. Hasin, Y., Seldin, M., and Lusis, A. (2017). Multi-omics approaches to disease. Genome Biol. *18*, 1–15.
2. Pinu, F.R., Beale, D.J., Paten, A.M., Kouremenos, K., Swarup, S., Schirra, H.J., and Wishart, D. (2019). Systems biology and multi-omics integration: viewpoints from the metabolomics research community. Metabolites *9*, 76.
3. Gillet, L.C., Navarro, P., Tate, S., Röst, H., Selevsek, N., Reiter, L., Bonner, R., and Aebersold, R. (2012). Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. Mol. Cell Proteomics *11*, https://doi.org/10.1074/mcp.O111.016717.
4. Hebert, A.S., Richards, A.L., Bailey, D.J., Ulbrich, A., Coughlin, E.E., Westphall, M.S., and Coon, J.J. (2013). The one hour yeast proteome. Mol. Cell Proteomics *13*, 339–347.
5. Meier, F., Geyer, P.E., Virreira Winter, S., Cox, J., and Mann, M. (2018). BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. Nat. Methods *15*, 440–448.
6. Rosenberger, G., Koh, C.C., Guo, T., Röst, H.L., Kouvonen, P., Collins, B.C., Heusel, M., Liu, Y., Caron, E., Vichalkovski, A., et al. (2014). A repository of assays to quantify 10,000 human proteins by SWATH-MS. Sci. Data *1*, 140031.
7. Wiśniewski, J.R., Zougman, A., Nagaraj, N., and Mann, M. (2009). Universal sample preparation method for proteome analysis. Nat. Methods *6*, 359–362.
8. Zhang, Y., Vera, J.M., Xie, D., Serate, J., Pohlmann, E., Russell, J.D., Hebert, A.S., Coon, J.J., Sato, T.K., and Landick, R. (2019). Multiomic fermentation using chemically defined synthetic hydrolyzates revealed multiple effects of lignocellulose-derived inhibitors on cell physiology and xylose utilization in *Zymomonas mobilis*. Front. Microbiol. *10*, 2596.
9. Overmyer, K.A., Rhoads, T.W., Merrill, A.E., Ye, Z., Westphall, M.S., Acharya, A., Shukla, S.K., and Coon, J.J. (2020). Proteomics, lipidomics, metabolomics and 16S DNA sequencing of dental plaque from patients with diabetes and periodontal disease. bioRxiv. https://doi.org/10.1101/2020.02.25.963967.
10. Chick, J.M., Munger, S.C., Simecek, P., Huttlin, E.L., Choi, K., Gatti, D.M., Raghupathy, N., Svenson, K.L., Churchill, G.A., and Gygi, S.P. (2016). Defining the consequences of genetic variation on a proteome-wide scale. Nature *534*, 500–505.
11. Stefely, J.A., Kwiecien, N.W., Freiberger, E.C., Richards, A.L., Jochem, A., Rush, M.J.P., Ulbrich, A., Robinson, K.P., Hutchins, P.D., Veling, M.T., et al. (2016). Mitochondrial protein functions elucidated by multi-omic mass spectrometry profiling. Nat. Biotechnol. *34*, 1191–1197.
12. Wang, D., Eraslan, B., Wieland, T., Hallström, B., Hopf, T., Zolg, D.P., Zecha, J., Asplund, A., Li, L.-H., Meng, C., et al. (2019). A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. Mol. Syst. Biol. *15*, e8503.
13. Williams, E.G., Wu, Y., Wolski, W., Kim, J.Y., Lan, J., Hasan, M., Halter, C., Jha, P., Ryu, D., Auwerx, J., et al. (2018). Quantifying and localizing the mitochondrial proteome across five tissues in a mouse population. Mol. Cell Proteomics *17*, 1766–1777.
14. Meyer, J.G., Liu, S., Miller, I.J., Coon, J.J., and Gitter, A. (2019). Learning drug functions from chemical structures with convolutional neural networks and random forests. J. Chem. Inf. Model. *59*, 4438–4449.
15. Lapointe, C.P., Stefely, J.A., Jochem, A., Hutchins, P.D., Wilson, G.M., Kwiecien, N.W., Coon, J.J., Wickens, M., and Pagliarini, D.J. (2018). Multi-omics reveal specific targets of the RNA-binding protein Puf3p and its orchestration of mitochondrial biogenesis. Cell Syst. *6*, 125–135.e6.
16. Weisenhorn, E.M.M., van 'T Erve, T.J., Riley, N.M., Hess, J.R., Raife, T.J., and Coon, J.J. (2016). Multi-omics evidence for inheritance of energy pathways in red blood cells. Mol. Cell Proteomics *15*, 3614–3623.
17. Choi, M., Chang, C.Y., Clough, T., Broudy, D., Killeen, T., MacLean, B., and Vitek, O. (2014). MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. Bioinformatics *30*, 2524–2526.
18. Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M.Y., Geiger, T., Mann, M., and Cox, J. (2016). The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat. Methods *13*, 731–740.
19. Palmblad, M., and van Eck, N.J. (2018). Bibliometric analyses reveal patterns of collaboration between ASMS members. J. Am. Soc. Mass Spectrom. *29*, 447–454.
20. Sidoli, S., Kulej, K., and Garcia, B.A. (2017). Why proteomics is not the new genomics and the future of mass spectrometry in cell biology. J. Cell Biol. *216*, 21–24.
21. Dennis, G., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C., and Lempicki, R.A. (2003). DAVID: database for annotation, visualization, and integrated discovery. Genome Biol. *4*, P3.
22. James Kent, W., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res. *12*, 996–1006.
23. Severin, J., Lizio, M., Harshbarger, J., Kawaji, H., Daub, C.O., Hayashizaki, Y., Bertin, N., Forrest, A.R.R., Beckhouse, A., Wells, C., et al. (2014).

Interactive visualization and analysis of large-scale sequencing datasets using ZENBU. Nat. Biotechnol. *32*, 217–219.

24. Krogan, N.J., Lippman, S., Agard, D.A., Ashworth, A., and Ideker, T. (2015). The cancer cell map initiative: defining the hallmark networks of cancer. Mol. Cell *58*, 690–698.
25. Schweppe, D.K., Huttlin, E.L., Harper, J.W., and Gygi, S.P. (2018). BioPlex display: an interactive suite for large-scale AP-MS protein-protein interaction data. J. Proteome Res. *17*, 722–726.
26. Thul, P.J., and Lindskog, C. (2018). The human protein atlas: a spatial map of the human proteome. Protein Sci. *27*, 233–244.
27. Kluyver, T., Ragan-kelley, B., Pérez, F., Granger, B.E., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., et al. (2016). Jupyter Notebooks—A Publishing Format for Reproducible Computational Workflows (Positioning and Power in Academic Publishing: Players, Agents and Agendas), pp. 87–90.
28. Torre, D., Lachmann, A., and Ma'ayan, A. (2018). BioJupies: automated generation of interactive notebooks for RNA-seq data analysis in the cloud. Cell Syst. *7*, 556–561.e3.
29. Xia, J., Psychogios, N., Young, N., and Wishart, D.S. (2009). MetaboAnalyst: a web server for metabolomic data analysis and interpretation. Nucleic Acids Res. *37*, W652–W660.
30. Sharma, V., Eckels, J., Taylor, G.K., Shulman, N.J., Stergachis, A.B., Joyner, S.A., Yan, P., Whiteaker, J.R., Halusa, G.N., Schilling, B., et al. (2014). Panorama: a targeted proteomics knowledge base. J. Proteome Res. *13*, 4205–4210.
31. Norris, J.L., Farrow, M.A., Gutierrez, D.B., Palmer, L.D., Muszynski, N., Sherrod, S.D., Pino, J.C., Allen, J.L., Spraggins, J.M., Lubbock, A.L.R., et al. (2017). Integrated, high-throughput, multiomics platform enables data-driven construction of cellular responses and reveals global drug mechanisms of action. J. Proteome Res. *16*, 1364–1375.
32. Vizcaíno, J.A., Deutsch, E.W., Wang, R., Csordas, A., Reisinger, F., Ríos, D., Dianes, J.A., Sun, Z., Farrah, T., Bandeira, N., et al. (2014). ProteomeXchange provides globally coordinated proteomics data submission and dissemination. Nat. Biotechnol. *32*, 223–226.
33. Shamsaei, B., Chojnacki, S., Pilarczyk, M., Najafabadi, M., Niu, W., Chen, C., Ross, K., Matlock, A., Muhlich, J., Chutipongtanate, S., et al. (2020). piNET: a versatile web platform for downstream analysis and visualization of proteomics data. Nucleic Acids Res. *48*, W85–W93.
34. Peng, G., Wilson, R., Tang, Y., Lam, T.K.T., Nairn, A.C., Williams, K., and Zhao, H. (2019). Proteomicsbrowser: MS/proteomics data visualization and investigation. Bioinformatics *35*, 2313–2314.
35. Merkel, D. (2014). Docker: lightweight linux containers for consistent development and deployment docker: a little background under the hood. Linux J. *2014*, 2–7.
36. Griss, J., Jones, A.R., Sachsenberg, T., Walzer, M., Gatto, L., Hartler, J., Thallinger, G.G., Salek, R.M., Steinbeck, C., Neuhauser, N., et al. (2014). The mzTab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. Mol. Cell Proteomics *13*, 2765–2775.
37. Jones, A.R., Eisenacher, M., Mayer, G., Kohlbacher, O., Siepen, J., Hubbard, S.J., Selley, J.N., Searle, B.C., Shofstahl, J., Seymour, S.L., et al. (2012). The mzIdentML data standard for mass spectrometry-based proteomics results. Mol. Cell Proteomics *11*, https://doi.org/10.1074/mcp.M111.014381.
38. Veling, M.T., Reidenbach, A.G., Freiberger, E.C., Kwiecien, N.W., Hutchins, P.D., Drahnak, M.J., Jochem, A., Ulbrich, A., Rush, M.J.P., Russell, J.D., et al. (2017). Multi-omic mitoprotease profiling defines a role for Oct1p in coenzyme Q production. Mol. Cell *68*, 970–977.e11.
39. Bostock, M., Ogievetsky, V., and Heer, J. (2011). D^3 data-driven documents. IEEE Trans. Vis. Comput. Graph. *17*, 2301–2309.
40. Reidenbach, A.G., Kemmerer, Z.A., Aydin, D., Jochem, A., McDevitt, M.T., Hutchins, P.D., Stark, J.L., Stefely, J.A., Reddy, T., Hebert, A.S., et al. (2018). Conserved lipid and small-molecule modulation of COQ8 reveals regulation of the ancient kinase-like UbiB family. Cell Chem. Biol. *25*, 154–165.e11.
41. Jha, P., McDevitt, M.T., Halilbasic, E., Williams, E.G., Quiros, P.M., Gariani, K., Sleiman, M.B., Gupta, R., Ulbrich, A., Jochem, A., et al. (2018). Genetic regulation of plasma lipid species and their association with metabolic phenotypes. Cell Syst. *6*, 709–721.e6.
42. Wang, H., Diaz, A.K., Shaw, T.I., Li, Y., Niu, M., Cho, J.H., Paugh, B.S., Zhang, Y., Sifford, J., Bai, B., et al. (2019). Deep multiomics profiling of brain tumors identifies signaling networks downstream of cancer driver genes. Nat. Commun. *10*, 3718.
43. Jha, P., McDevitt, M.T., Gupta, R., Quiros, P.M., Williams, E.G., Gariani, K., Sleiman, M.B., Diserens, L., Jochem, A., Ulbrich, A., et al. (2018). Systems analyses reveal physiological roles and genetic regulators of liver lipid species. Cell Syst. *6*, 722–733.e6.
44. MacGilvray, M.E., Shishkova, E., Place, M., Wagner, E.R., Coon, J.J., and Gasch, A.P. (2020). Phosphoproteome response to dithiothreitol reveals unique versus shared features of *Saccharomyces cerevisiae* stress responses. J. Proteome Res. *19*, 3405–3417.
45. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29.
46. Deutsch, E.W., Mendoza, L., Shteynberg, D., Farrah, T., Lam, H., Tasman, N., Sun, Z., Nilsson, E., Pratt, B., Prazen, B., et al. (2010). A guided tour of the trans-proteomic pipeline. Proteomics *10*, 1150–1159.
47. Lazar, C., Gatto, L., Ferro, M., Bruley, C., and Burger, T. (2016). Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies. J. Proteome Res. *15*, 1116–1125.
48. Do, K.T., Wahl, S., Raffler, J., Molnos, S., Laimighofer, M., Adamski, J., Suhre, K., Strauch, K., Peters, A., Gieger, C., et al. (2018). Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies. Metabolomics *14*, 128.
49. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B Methodol. *57*, 289–300.
50. Dunn, O.J. (1961). Multiple comparisons among means. J. Am. Stat. Assoc. *56*, 52–64.

PATTER, Volume *1*

Supplemental Information

Argonaut: A Web Platform for Collaborative Multi-omic Data Visualization and Exploration

Dain R. Brademan, Ian J. Miller, Nicholas W. Kwiecien, David J. Pagliarini, Michael S. Westphall, Joshua J. Coon, and Evgenia Shishkova

Table S1, Key Resources Table. (Related to Experimental Procedures)

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
Analyzed Data	Veling et al., 2017	https://chorusproject.org/pages/dashboard.html#/projects/all/1366/experiments
Software and Algorithms		
Argonaut Source Code	This Paper	https://github.com/coongroup/Argonaut
Argonaut Example Website	This Paper	https://coonlabdatadev.com
Docker Environment	Merkel, 2014	https://www.docker.com
Argonaut Docker Image	This Paper	https://hub.docker.com/r/coonlabs/argonaut
D3.JS, Data Driven Documents	Bostock et al., 2011	https://d3js.org
Outlier Analysis	Stefely et al., 2016; This Paper	https://github.com/coongroup/Y3K-Software/Y3K_Data_Analyses; https://github.com/coongroup/Argonaut
Volcano Plots	Stefely et al., 2016; This Paper	https://github.com/coongroup/Y3K-Software/Y3K_Data_Analyses; https://github.com/coongroup/Argonaut
Molecular perturbations by condition	Stefely et al., 2016; This Paper	https://github.com/coongroup/Y3K-Software/Y3K_Data_Analyses; https://github.com/coongroup/Argonaut
Condition/Condition Correlations	Stefely et al., 2016; This Paper	https://github.com/coongroup/Y3K-Software/Y3K_Data_Analyses; https://github.com/coongroup/Argonaut
Principle Component Analysis	Stefely et al., 2016; This Paper	https://github.com/coongroup/Y3K-Software/Y3K_Data_Analyses; https://github.com/coongroup/Argonaut
Gene Ontology Enrichment Analysis	Ashburner et al., 2000; Gene Ontology Consortium, 2019; This Paper	http://geneontology.org/; https://github.com/coongroup/Argonaut

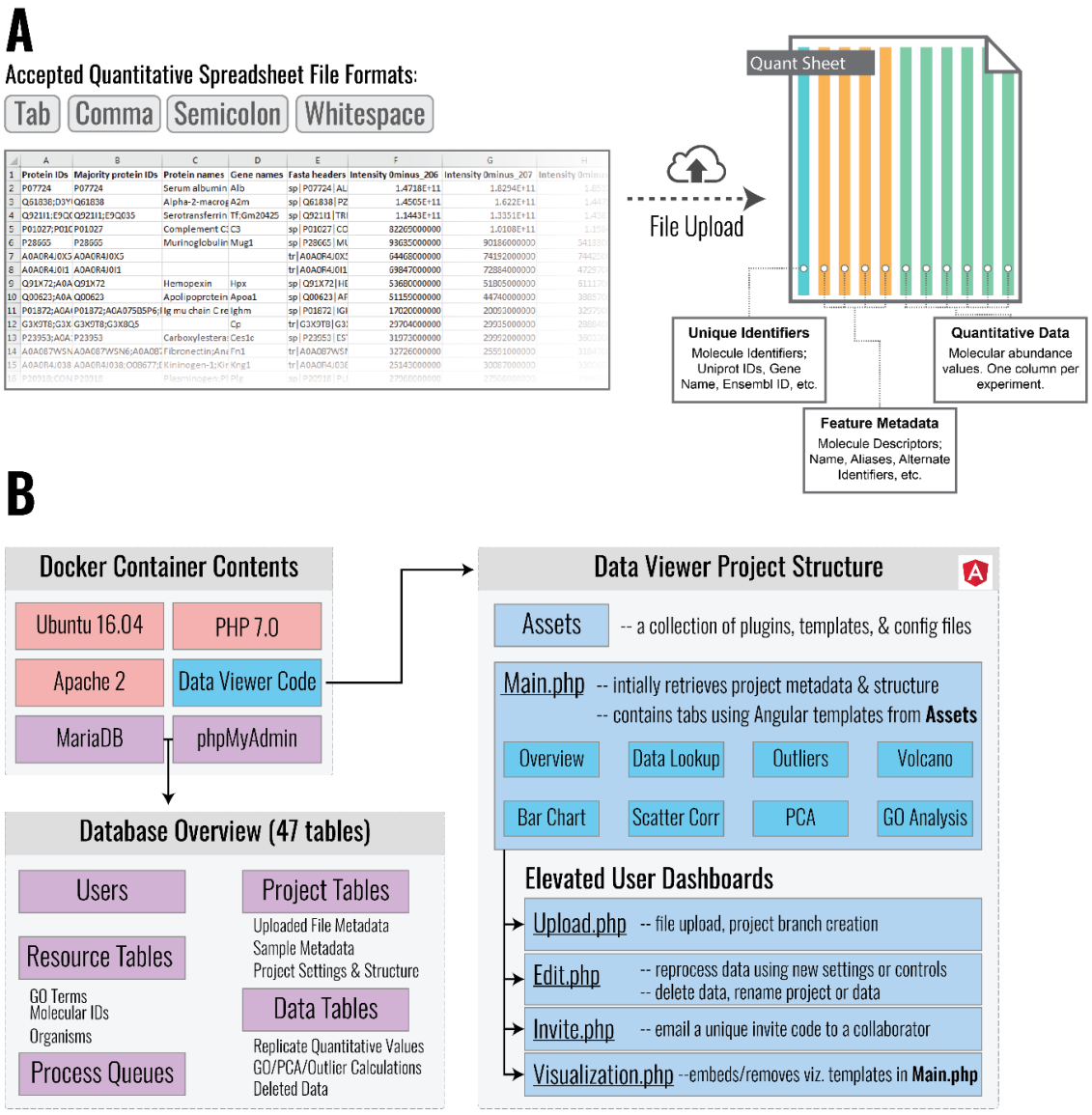


Figure S1: **Data Upload and Docker Container Structures (Related to Figure 1).** (**A**) Currently, Argonaut supports the upload and processing of four formats of plain-text files to supply a data portal with new quantitative data: tab, comma, semicolon, and whitespace. Once uploaded, the user can select which columns contain a unique identifier, metadata, or quantitative values. In the figure, the user would select Protein IDs as a unique identifier; Majority Protein IDs, Protein Names, Gene Names, and Fasta Headers as feature metadata; and the remaining columns as quantitative data. (**B**) A brief graphical breakdown of the backend structure of an Argonaut project running in a Docker container. The project database contains 47 individual tables used to store data for five general purposes: user account information, static processing resources, job processing queues, project-specific metadata, and uploaded quantitative values. The Angular JavaScript framework drives the client-side application. Only elevated users can create or change projects, upload new data, and edit existing data in a data portal.

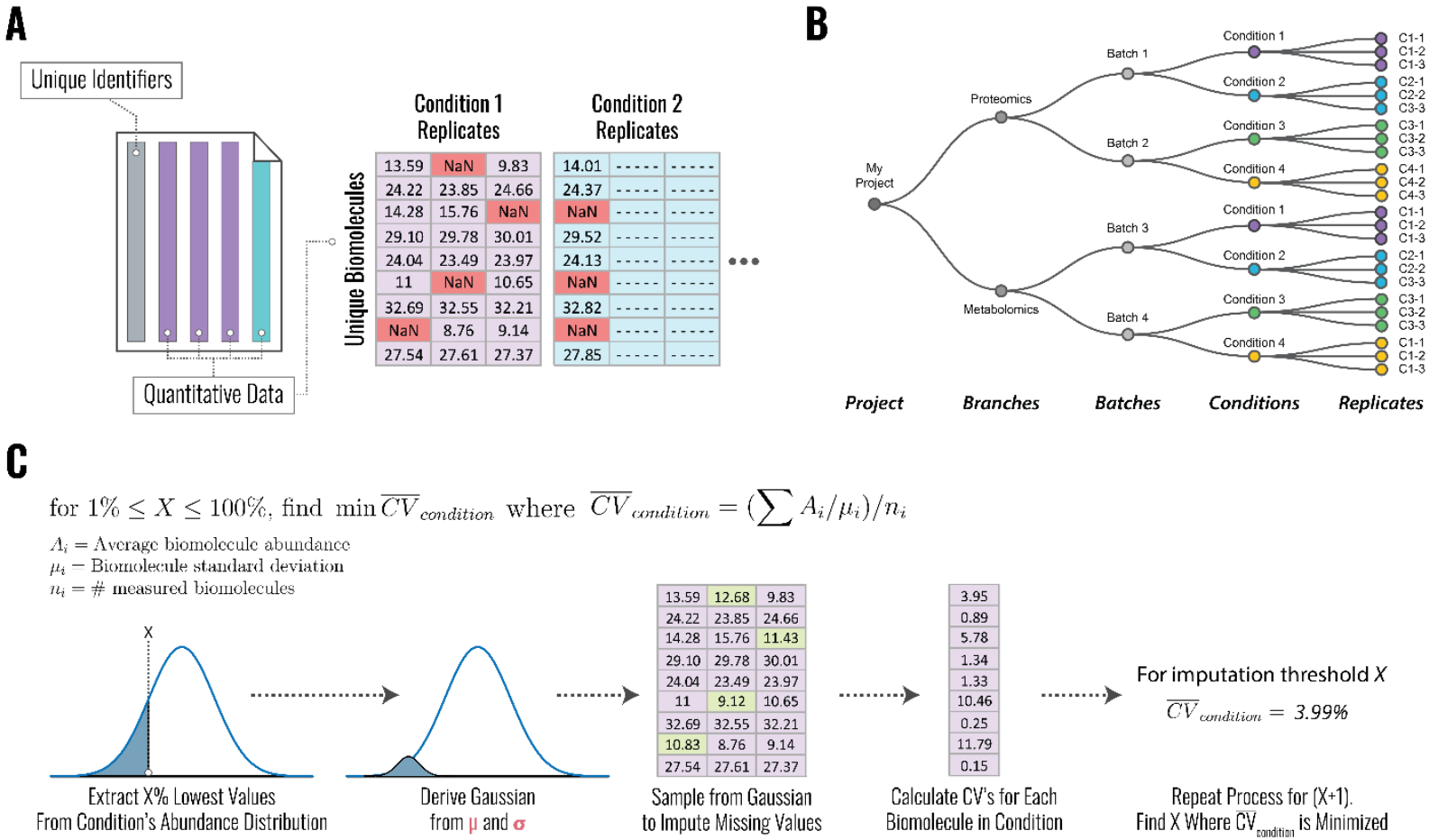


Figure S2: **Portal Data Organization and Missing Value Imputation (Related to Figure 1).** (**A**) After a new dataset has been uploaded to the portal by an elevated user, experimental replicates are grouped into their respective condition classifiers. However, some biomolecules may be missing quantitative values. **Panel B** presents a detailed representation of the tree-based hierarchy that Argonaut utilizes to classify experimental replicates into different categories. **Panel C** shows a visual representation of the modified left-censored imputation algorithm used by Argonaut. Imputation is done on each condition separately. The algorithm iteratively selects a percentage of the lowest biomolecule measurements to find a sampling threshold to minimize the average biomolecule coefficient of variation for the specified condition.

REFERENCES

1. Veling, M.T., Reidenbach, A.G., Freiberger, E.C., Kwiecien, N.W., Hutchins, P.D., Drahnak, M.J., Jochem, A., Ulbrich, A., Rush, M.J.P., Russell, J.D., et al. (2017). Multi-omic Mitoprotease Profiling Defines a Role for Oct1p in Coenzyme Q Production. Molecular Cell *68*, 970-977.e11.

2. Merkel, D. (2014). Docker : Lightweight Linux Containers for Consistent Development and Deployment Docker : a Little Background Under the Hood. Linux Journal *2014*, 2–7.

3. Bostock, M., Ogievetsky, V., and Heer, J. (2011). D^3 Data-Driven Documents. IEEE Transactions on Visualization and Computer Graphics *17*, 2301–2309.

4. Stefely, J.A., Kwiecien, N.W., Freiberger, E.C., Richards, A.L., Jochem, A., Rush, M.J.P., Ulbrich, A., Robinson, K.P., Hutchins, P.D., Veling, M.T., et al. (2016). Mitochondrial protein functions elucidated by multi-omic mass spectrometry profiling. Nature Biotechnology *34*, 1191–1197.

5. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: Tool for the unification of biology. Nature Genetics *25*, 25–29.

6. Carbon, S., Douglass, E., Dunn, N., Good, B., Harris, N.L., Lewis, S.E., Mungall, C.J., Basu, S., Chisholm, R.L., Dodson, R.J., et al. (2019). The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Research *47*, D330–D338.