# Multi-omics Pathways Workflow (MOPAW): An Automated Multi-omics Workflow on the Cancer Genomics Cloud

Cancer Informatics
Volume 22: 1–8
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/11769351231180992

Trinh Nguyen[1], Xiaopeng Bian[1], David Roberson[2], Rakesh Khanna[1], Qingrong Chen[1], Chunhua Yan[1], Rowan Beck[2], Zelia Worman[2] and Daoud Meerzaman[1]

[1]The Computational Genomics and Bioinformatics Branch, Center for Biomedical Informatics and Information Technology, National Cancer Institute, Rockville, MD, USA. [2]Seven Bridges, Charlestown, MA, USA.

## ABSTRACT

**INTRODUCTION:** In the era of big data, gene-set pathway analyses derived from multi-omics are exceptionally powerful. When preparing and analyzing high-dimensional multi-omics data, the installation process and programing skills required to use existing tools can be challenging. This is especially the case for those who are not familiar with coding. In addition, implementation with high performance computing solutions is required to run these tools efficiently.

**METHODS:** We introduce an automatic multi-omics pathway workflow, a point and click graphical user interface to Multivariate Single Sample Gene Set Analysis (MOGSA), hosted on the Cancer Genomics Cloud by Seven Bridges Genomics. This workflow leverages the combination of different tools to perform data preparation for each given data types, dimensionality reduction, and MOGSA pathway analysis. The Omics data includes copy number alteration, transcriptomics data, proteomics and phosphoproteomics data. We have also provided an additional workflow to help with downloading data from The Cancer Genome Atlas and Clinical Proteomic Tumor Analysis Consortium and preprocessing these data to be used for this multi-omics pathway workflow.

**RESULTS:** The main outputs of this workflow are the distinct pathways for subgroups of interest provided by users, which are displayed in heatmaps if identified. In addition to this, graphs and tables are provided to users for reviewing.

**CONCLUSION:** Multi-omics Pathway Workflow requires no coding experience. Users can bring their own data or download and preprocess public datasets from The Cancer Genome Atlas and Clinical Proteomic Tumor Analysis Consortium using our additional workflow based on the samples of interest. Distinct overactivated or deactivated pathways for groups of interest can be found. This useful information is important in effective therapeutic targeting.

**KEYWORDS:** Multi-omics; cloud, automatic workflow, CGC, pathway analysis, TCGA

**RECEIVED:** April 21, 2023. **ACCEPTED:** May 22, 2023.

**TYPE:** Original Research

**FUNDING:** The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: "The Seven Bridges Cancer Research Data Commons Cloud Resource has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, Contract No. HHSN261201400008C and ID/IQ Agreement No. 17X146 under Contract No. HHSN261201500003I and 75N91019D00024."

**DECLARATION OF CONFLICTING INTERESTS:** The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**CORRESPONDING AUTHOR:** Trinh Nguyen, The Computational Genomics and Bioinformatics Branch, Center for Biomedical Informatics and Information Technology, National Cancer Institute, 9609 Medical Center Drive Rockville, MD 20850, Rockville, MD 20892-2590, USA. Email: tinh.nguyen@nih.gov

## Background

Cloud computing has recently become more popular as a platform for genomic data analysis and collaboration.[1] One such cloud platform for cancer research is the Cancer Genomics Cloud (CGC),[2] powered by Seven Bridges Genomics (SBG). This is a secure platform to access data, analysis tools, and computing resources. This platform enables developers to build programs and tools that guide users from data preparation to downstream analysis. The cloud environment of this platform enables collaborators from different institutions to work together on the same project without needing to locally download and manage the datasets. This reduces the cost of data transfer, organization, and storage. To study complex biological processes, it is important to combine multi-omics datasets to find interrelationships between biomolecules and their functions.[3] For disease subtyping, there are many multi-omics data integration methods, such as the approach for transforming multi-omics data into gene similarity networks via self-organizing maps[4] or the model based on uniform manifold approximation and projection (UMAP) and convolutional neural networks (CNNs),[5] and a tool called Multivariate Single Sample Gene Set Analysis (MOGSA). MOGSA integrates omics data sets to find the most variant biomolecules, and to generate gene-set scores for each sample,[6] which are useful for finding the distinct pathways for each subgroup. Our team has recently used MOGSA to integrate transcriptome sequencing with copy number alteration at the gene level derived from the segment mean and to thereby identify distinct pathways in a

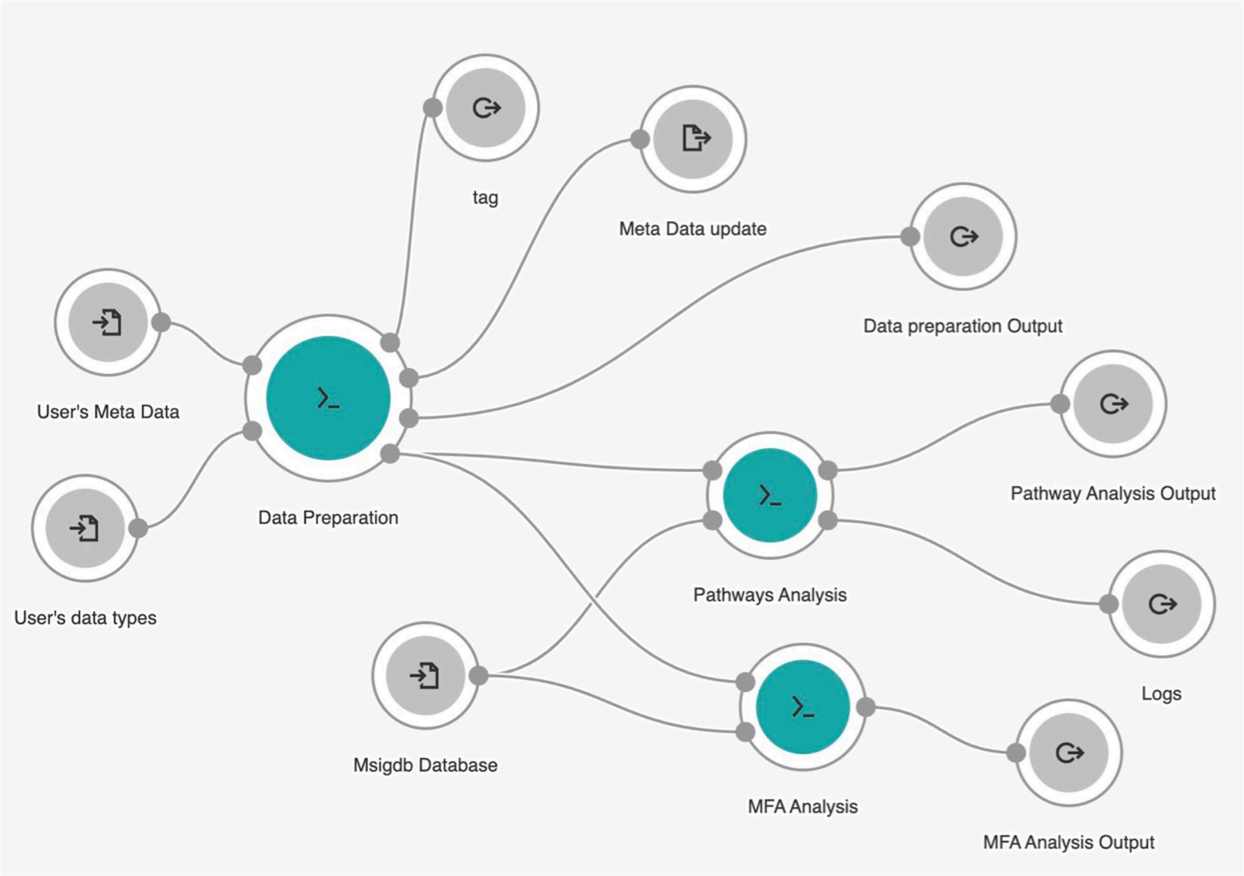


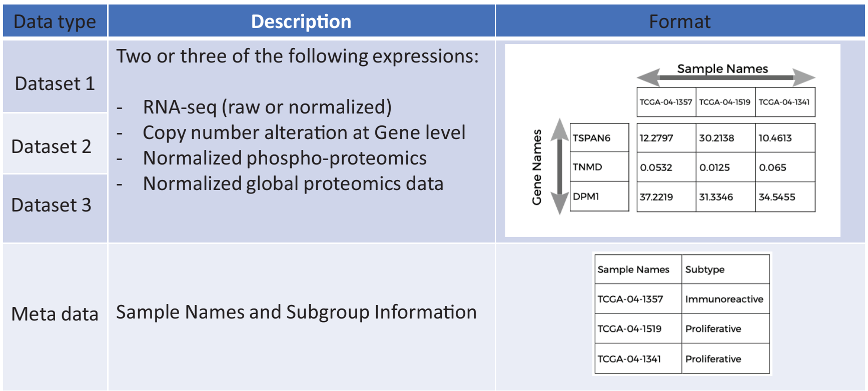
**Figure 1.** Diagram of multi-omics pathways workflow (MOPAW). This diagram illustrates the overall design of multi-omics pathways workflow

high-risk subgroup of adults with Acute Myeloid Leukemia using The Cancer Genome Atlas (TCGA) datasets.[7] An example of applying MOGSA comes from a project by the Applied Proteogenomics OrganizationaL Learning and Outcomes (APOLLO) network, in which we integrated transcriptome sequencing, and proteomics, and phosphoproteomics using MOGSA to identify distinct pathways in subtypes from 87 lung adenocarcinoma cases.[8] There is a need for a uniform framework that can process and analyze multi-omics data in an end-to-end manner. To meet that need, we implemented a workflow pipeline that extends from processing data to pathway analysis for disease subtyping. This pipeline is automated with a graphical user interface (GUI) in Multi-omics Pathway Workflow (MOPAW). There are several advantages of this MOPAW, implemented on CGC over other platform such as CBioPortal.[9] First, MOPAW allows users to easily bring their own datasets to analyze. Second, if users wish to use public datasets, they can use our additional available workflow. Third, through the implementation of this workflow on CGC, users can invite other collaborators to easily join a project. As soon as the run task is finished, all the members receive a notification to review and interpret results. In addition, users can access over 600 analytical and bioinformatics tools and workflow on the CGC platform. Finally, MOPAW allows users to do analyses based on the integration of different datatypes at the beginning.

## Implementation

### *The Multi-omics Pathways Workflow (MOPAW) on CGC*

To implement the automatic GUI MOPAW on CGC, we first created a docker image which included all required libraries, the docker image was then pushed to the CGC docker repository. Apps running the workflow wrapped in Common Workflow Language (CWL) were then created by pulling the docker image to the CGC tool visual editor and filling out the GUI template with parameters and scripts (Figure 1). Interested readers can refer to the online documentations for details (https://docs.cancergenomicscloud.org/page/bring-your-own-tools-to-the-cancer-genomics-cloud). The automatic MOPAW currently accepts 3 expression datatypes given by users to performs data preparation, multi-factorial analysis (MFA), and pathway analysis. The metadata file contains at least 2 columns, sample names and groups of interest. The format of expression data types should have rows as genes and columns as sample names. The sample names should have the same format as the metadata file. These datatypes should be 2 or 3 of the following types: RNA-seq (raw or normalized), copy number alteration, normalized phosphoproteomics, and normalized proteomics data (Figure 2). The automatic MOPAW starts with appropriate data preparation for each given data types, dimensionality reduction is then applied, and



<table>
<tr><th>Data type</th><th>Description</th><th>Format</th></tr>
<tr><td>Dataset 1</td><td rowspan="3">Two or three of the following expressions:<br>- RNA-seq (raw or normalized)<br>- Copy number alteration at Gene level<br>- Normalized phospho-proteomics<br>- Normalized global proteomics data</td><td rowspan="3">Sample Names<br>Gene Names<table><tr><td></td><td>TCGA-04-1357</td><td>TCGA-04-1519</td><td>TCGA-04-1341</td></tr><tr><td>TSPAN6</td><td>12.2797</td><td>30.2138</td><td>10.4613</td></tr><tr><td>TNMD</td><td>0.0532</td><td>0.0125</td><td>0.065</td></tr><tr><td>DPM1</td><td>37.2219</td><td>31.3346</td><td>34.5455</td></tr></table></td></tr>
<tr><td>Dataset 2</td></tr>
<tr><td>Dataset 3</td></tr>
<tr><td>Meta data</td><td>Sample Names and Subgroup Information</td><td><table><tr><td>Sample Names</td><td>Subtype</td></tr><tr><td>TCGA-04-1357</td><td>Immunoreactive</td></tr><tr><td>TCGA-04-1519</td><td>Proliferative</td></tr><tr><td>TCGA-04-1341</td><td>Proliferative</td></tr></table></td></tr>
</table>

**Figure 2.** Detail of data inputs provided by users.

finally pathway analysis is performed based on the group of interest. Below is the detail how these steps are implemented at the back end:

### *Data preparation and MFA analysis*

Data preparation: This includes the removal of low-expressed genes, data normalization, transformation, and imputation. The following steps are applied as needed depending on user's chosen app settings:

Copy number alteration (CNA): The Copy number alteration should contain the gene names with their corresponding segment mean from the segmentation file. Only genes with a sum of CNA values across samples greater than zero are retained for further analysis.

RNA-seq transcriptome: Only genes with greater than one count-per-million reads in at least 50% of the total common samples are retained for the further analysis. The RNA-seq matrix should be either raw count or normalized. In case of raw count, the matrix is normalized and then log2 transformed. In case of normalized data, the matrix is log2 transformation only.

Normalized proteomics and phosphoproteomics data: the proteomics and phosphoproteomics should be normalized before feeding to the workflow. There are many tools available to do normalization[10] depending on user's preference. In the matrices, some samples do not have abundance values available for all proteins, so we have provided an option to perform imputation. When a protein has missing abundance values for more than 50% of samples, that protein is not further considered. However, if the protein is missing abundance values for less than 50% of samples, the value for each missing sample of that protein is imputed with a k-nearest neighbor (k-NN) strategy adapted from Lazar et al.[11] This is done using the DreamAI function,[12] an algorithm for the imputation of proteomics data, used with the default parameters (Supplemental Table 1).

*MFA analysis.* Dimensionality reduction is a crucial step in many multi-omics analyses. Within our workflow, this is done with MFA analysis and is executed with the moa function of MOGSA.[6] In the output of this MFA step, users can see how much variation within their types of interest can be explained with various numbers of principal components (PCs) (Supplemental Figure 1). This allows the user to evaluate and decide the number of PCs they want to use for the pathway analysis step.

### *Pathway analysis*

This step of the analysis identifies the gene set scores (GSS) for databases selected by the user. MSigDB databases, version of 7.1,[13] are preloaded into the application. Users can download the newest version at http://www.gsea-msigdb.org/gsea/msigdb/collections.jsp. The MOGSA function, Integrative Single Sample Gene-set Analysis of Multiple Omics,[6] is used for this step with the default parameters (Supplemental Table 1). The number of PC's is specified by users at runtime.

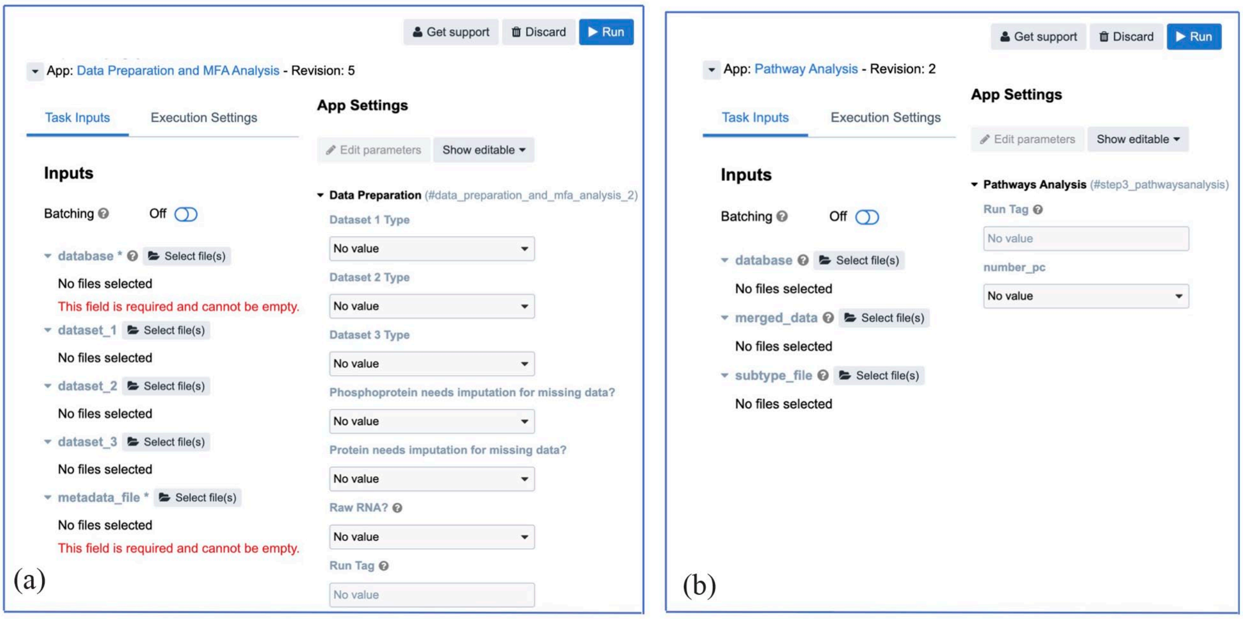
To determine what pathways are enriched in each subtype or subgroup, we first selected the pathways resulting from the MOGSA function with GSS false discovery rate (FDR) values smaller than 0.01 in 50% or more of all samples. We then used cut-off FDR < 0.01 to select the significant pathways based on t.test function for 2 groups comparison or ANOVA function for at least 3 group comparison.

Lastly, we used generalized linear models (GLM) to calculate the difference of GSS in each subgroup versus that in the rest and selected the top 5 and bottom 5 significant representative pathways ranked by GLM T values with $P < .05$.

MOPAW generates 2 heatmaps to visualize these resulting representative enriched pathways if found. With these, users can visualize the *z*-score scaled single gene-set normalized enrichment scores across all samples (Supplemental Figure 4 for case use) and z-score scaled median GSS from data types as well as the contribution of each data type (Figure 4) to the interesting subgroups.

## Results

We developed a web-based GUI application called automated MOPAW on Cloud Analysis Platforms. Our web-based workflow, built on the CGC, requires no coding or command line



(a)

App: Data Preparation and MFA Analysis - Revision: 5

Inputs

Batching Off

database * — No files selected — This field is required and cannot be empty.

dataset_1 — No files selected

dataset_2 — No files selected

dataset_3 — No files selected

metadata_file * — No files selected — This field is required and cannot be empty.

App Settings

Data Preparation (#data_preparation_and_mfa_analysis_2)

Dataset 1 Type: No value

Dataset 2 Type: No value

Dataset 3 Type: No value

Phosphoprotein needs imputation for missing data?: No value

Protein needs imputation for missing data?: No value

Raw RNA?: No value

Run Tag: No value

(b)

App: Pathway Analysis - Revision: 2

Inputs

Batching Off

database — No files selected

merged_data — No files selected

subtype_file — No files selected

App Settings

Pathways Analysis (#step3_pathwaysanalysis)

Run Tag: No value

number_pc: No value

**Figure 3.** User Interface of automated multi-omics workflow. This illustrates the corresponding user interface of MOPAW workflow: (a) data preparation and MFA analysis and (b) pathway analysis of the multi omics pathways analysis.

experience (Figure 3). With this web-based GUI application, users simply upload the input data files, and then choose desired app settings from the drop-box, and finally run their tasks solely by mouse-clicks. We encourage users to review the results from data preparation and MFA analysis before running pathway analysis module. For more detailed information how to use and run the workflow, please refer to the user manual (MOPAW User Guide).

To show the capabilities of our automated MOPAW in action, we performed an example analysis using a subset of TCGA Ovarian cancer cohort with transcriptomic data generated by the TCGA Research Network: https://www.cancer.gov/tcga and proteomics data by the Clinical Proteomic Tumor Analysis Consortium (NCI/NIH). We integrated RNA-seq and global protein datasets from 62 common TCGA samples. For RNA-seq, we used normalized version of fragments per kilobase of exon per million mapped fragments (FPKM). For Global protein, we used shared log ratio values of iTRAQTM Protein Quantitation from Pacific Northwest National Laboratory.[14] Discovery subtypes were downloaded from Verhaak et al.[15] For the molecular pathways, we used the gmt file of MSigDB hallmark version 7.1 from https://www.gsea-msigdb.org/gsea/msigdb/collections.jsp for annotation. This example utilizes data preparation, MFA analysis, and the pathway analysis applications of our workflow. The results include a Venn diagram showing the number of common samples across data types and groups of interests (Supplemental Figure 2), distribution graphs for the given datatypes (Supplemental Figure 3), *z*-score scaled median gene set scores in a heatmap (Figure 4), and the *z*-score scaled single gene-set normalized enrichment scores across all samples (Supplemental Figure 4). These 2 heatmaps show any significantly differentially expressed pathways resulting from both data types as well as the contribution of each data type to the discovery subtypes. The total running time on SBG was 7 minutes, which costs about $0.05.

### *Additional workflow for downloading and processing public data*

The availability of public datasets generated by the TCGA Research Network: https://www.cancer.gov/tcga, and the Clinical Proteomic Tumor Analysis Consortium (CPTAC), has given researchers access to omics expression data in a wide range of different cancers. Users can access, download the desired datasets, and then analyzed them with automated MOPAW, all within the CGC platform. However, this can be tedious for those who are not familiar with TCGA and CPTAC. To improve the ease of use, we also implemented a workflow (Figure 5) and its corresponding GUI interface (Figure 6) on CGC platform to help with the data downloading and preprocessing steps. The required metadata file contains at least 2 columns, sample names and groups of interest. This workflow will only download the samples provided in this required metadata file. All the final matrices contain the gene names as rows and the sample names as column. The sample names have the same format with the interest samples provided by users. The final matrices generated from this workflow can be used as input for multi-omics pathway workflow. Currently, this workflow will download and process RNA-seq and copy number segmentation datasets from TCGA for all available cancer types. Due to the current CPTAC limits of phosphoproteomics and global protein data, we have pre-downloaded and then processed data for TCGA Ovarian and TCGA Breast cancers. Depending on the desired app settings, the following might be applied as below:

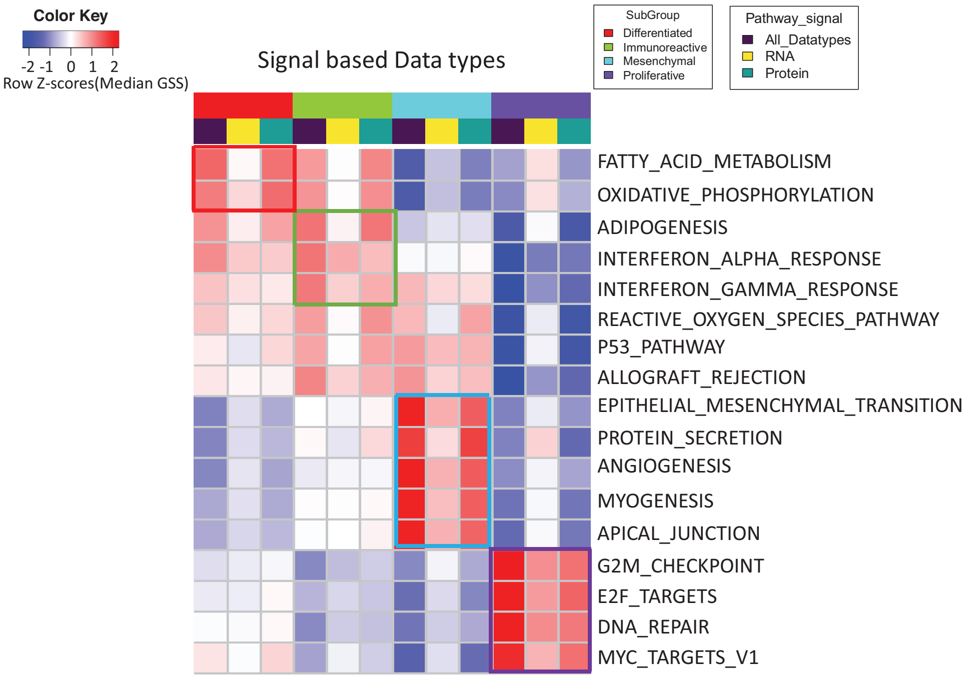


**Figure 4.** Signal based data types. An example of the unique significant Hallmark pathways for each discovery subtype within the Ovarian samples, showing the overall of both datatypes and contribution of RNA and protein signals with cut-off FDR < 0.01 across these discovery subgroups. The pathways were highlighted with the squares for their corresponding subgroups. Red represents positive scores while blue represents negative scores. The top 2 pathways have highest scores compared with the other subgroups, suggesting that these 2 enriched in the differentiated subgroup while the last 4 pathways have highest scores compared with the remaining subgroups, suggesting that these pathways enriched in the proliferation subgroup.

RNA-seq: We used Genomic Data Commons (GDC) query from TCGAbiolinks,[16] An R/Bioconductor package for integrative analysis with GDC, to obtain gene expression with these parameters setting: data.category = "Transcriptome Profiling", data.type = "Gene Expression Quantification", sample.type = "Primary Tumor", workflow.type = "STAR – Counts." Raw RNA seq as well as different normalized methods, FPKM, TPM, and FPKMUQ, along with the gene symbols generated by GDC were reported. This allows users to make decision which type they would like to use for their project.

Copy number alteration (CNA) segmentation: For copy number alteration segmentation, we used GDC query function with these parameters setting: data.category = "Copy Number Variation", data.type = "Copy Number Segment". We only selected primary samples ending with "01A" and "01B." We estimated gene level CNA as the segment mean of copy numbers of the genomic region of a gene by using TCGA-Assembler 2,[17] downloaded from https://github.com/compgenome365/TCGAAssembler-2 (version 2.0.6). Degree of CNA was calculated as log2 (tumor values/normal values). Hg38 "ensemble" was used to obtain gene position.

Global and phospho datasets: We downloaded gene-level iTRAQ log-ratios reported the *.itraq.tsv files from *The Proteomic Data Commons* (PDC) for these 2 cancer types: Ovarian cancer from Pacific Northwest National Laboratory (PNNL) study and Breast invasive (BI) Proteome study. Only the samples ending with "-01A Log Ratio" were kept for the final matrices.[18] The purpose of this workflow is to provide expression matrix for integration with other types such as RNA-seq. Therefore, the sites from phosphoproteomics dataset would not be included in the final matrix.

## Discussion

The automated MOPAW enables users to search for unique molecular pathways for subgroups of interest using various combinations of data types. By combining multiple types of data, missing or unreliable information in any single data can be compensated for, and gene sets that cannot be detected by single omics data analysis, might be found. Also, the contribution of datasets and individual biomolecules from these datasets can be observed. Users can perform analyses using databases downloaded from http://www.gsea-msigdb.org/gsea/msigdb/collections.jsp such as Hallmark, GO Biological Process, and Canonical database. Using example data, we demonstrate that this workflow was able to identify distinct pathways for each discovery subtype in ovarian cancer (Figure 4). For instance, the molecular characteristics of proliferative subgroup include the highly overactivated pathways in the categories of cell proliferation. We have pushed this project to public, users can copy and run it to get familiar with the workflow before using their own data. Moreover, there is no requirement of writing complex codes for multi-omics analysis, which is especially useful for researchers who do not have coding experience. Therefore, this workflow offers a bridge to cross the gap between bioinformaticians and clinicians. The automated MOPAW was built on the cloud platform, which enables users to access not only this

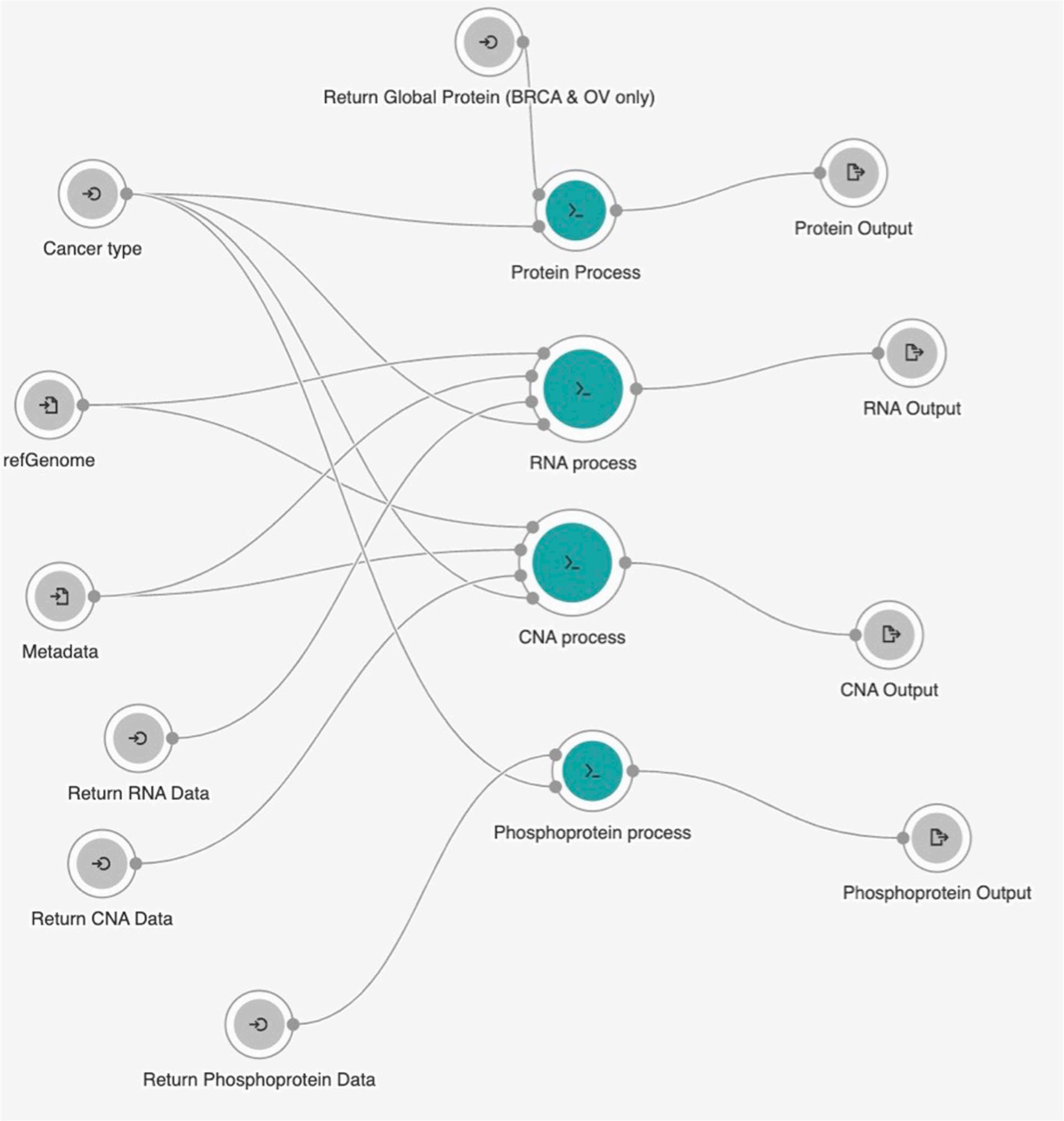
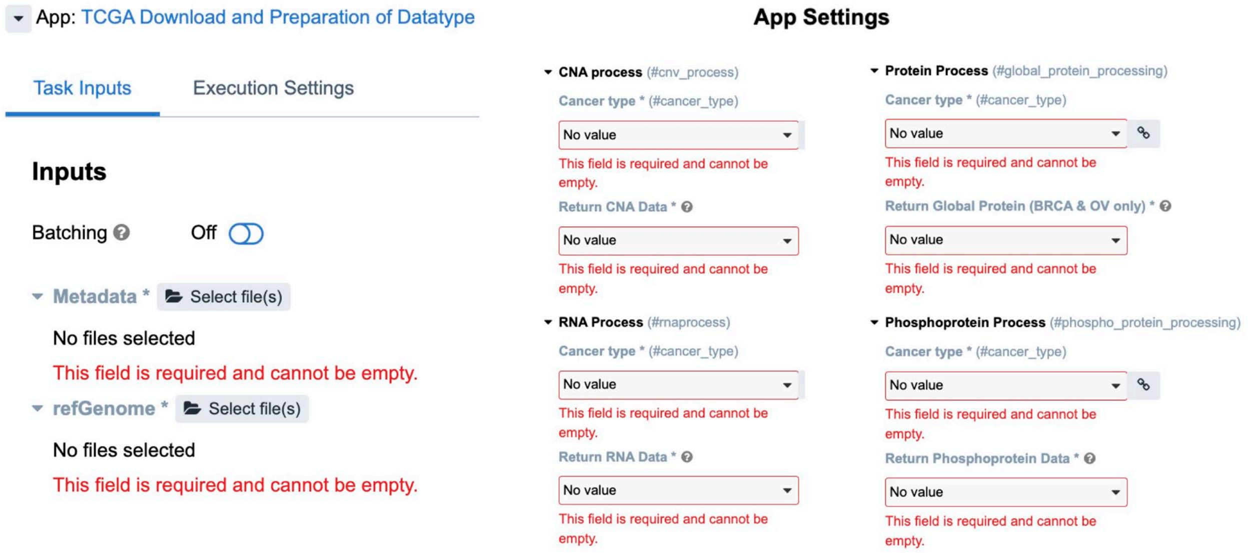


**Figure 5.** Diagram Of downloading and processing public data from TCGA and CPTAC. This diagram illustrates the overall design of downloading and processing public data from TCGA and CPTAC. Metadata is provided by users.

workflow but also many existing workflows, and to share data with other collaborators. Besides, there is additional workflow for downloading and processing public datasets if user wish to use public data for either validation purpose or research question on learning about disease subtyping. The limitation of this workflow is at least 2 and up to 3 datatypes. Also, we limit only RNA-seq, CNA, and proteomics, and phosphoproteomics. In the future, we aim to add more datatypes such as methylation and metabolites data. Also, because this workflow is implemented on the cloud, users must pay for the cost of running the workflow and the storage of data. However, the CGC platform is funded by the National Cancer Institute, new users are provided a $300 credit for use of their cloud platform if they are coming from a non-profit, academic and government institution. For members of the National Institute of Health community, these task charges for the any workflow on SBG which allow access to high performance cloud compute nodes, are fully covered. New users can also send an email to CGC support group through cgc@sbgenomics.com to request this pilot fund for their new account. Finally, to ensure a smooth experience, if users encounter issues with our MOPAW workflow, they can get technical support from the CGC team as well as our team (Computational Genomics and Bioinformatics Branch at the National Cancer Institute).



App: TCGA Download and Preparation of Datatype

Inputs

Batching Off

Metadata *
No files selected
This field is required and cannot be empty.

refGenome *
No files selected
This field is required and cannot be empty.

App Settings

CNA process (#cnv_process)
Cancer type * (#cancer_type)
No value
This field is required and cannot be empty.
Return CNA Data *
No value
This field is required and cannot be empty.

RNA Process (#rnaprocess)
Cancer type * (#cancer_type)
No value
This field is required and cannot be empty.
Return RNA Data *
No value
This field is required and cannot be empty.

Protein Process (#global_protein_processing)
Cancer type * (#cancer_type)
No value
This field is required and cannot be empty.
Return Global Protein (BRCA & OV only) *
No value
This field is required and cannot be empty.

Phosphoprotein Process (#phospho_protein_processing)
Cancer type * (#cancer_type)
No value
This field is required and cannot be empty.
Return Phosphoprotein Data *
No value
This field is required and cannot be empty.

**Figure 6.** User interface of Downloading and processing public data from TCGA and CPTAC App. This diagram illustrates the corresponding user interface of downloading and processing public data from TCGA and CPTAC.

## Conclusions

The automatic MOPAW was built to provide a complete, multi-omics data analysis pipeline that is accessible to users without programing experience. The current version of the program contains a suite of tools that help users through every step of this bioinformatic analysis. Additional workflow for downloading and processing public data is also provided. Over time, more tools and features will be added to further expand its capabilities. To use this workflow, users should have computer with access to the internet and web browser installed such as Chrome. Users can reach out to SBG and CGC for technical support, and questions or to request additional needed figures.

## Acknowledgements

Not applicable

## Author Contributions

TN, RK, and XB wrote the paper. TN wrote the scripts. XB, DR, Rowan, and TN helped to integrate the workflow to CGC platform. All authors reviewed the article.

## Ethics Approval and Consent to Participate

Not applicable

## Consent for Publication

Not applicable

## Availability of Data and Materials

User manual: MOPAW_User_Guide

For the below links, please follow the steps provided in the MOPAW's user guide.

Public project

https://cgc.sbgenomics.com/u/sevenbridges/mopaw-1

Data preparation and MFA analysis App:

https://cgc.sbgenomics.com/public/apps/bianxi/commit-meerzaman-lab-cbiit-nci/data-preparation-and-mfa-analysis-2

Pathway analysis App

https://cgc.sbgenomics.com/public/apps/bianxi/commit-meerzaman-lab-cbiit-nci/pathway-analysis

TCGA download and preprocessing App

https://cgc.sbgenomics.com/public/apps/bianxi/commit-meerzaman-lab-cbiit-nci/tcga-download-ap-prep-datatype-r

## Supplemental Material

Supplemental material for this article is available online.

## REFERENCES

1. Langmead B, Nellore A. Cloud computing for genomic data analysis and collaboration. *Nat Rev Genet*. 2018;19:208-219.
2. Lau JW, Lehnert E, Sethi A, et al. The Cancer Genomics Cloud: Collaborative, reproducible, and democratized-A new paradigm in large-scale computational research. *Cancer Res*. 2017;77:e3-e6.
3. Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics data integration, interpretation, and its application. *Bioinform Biol Insights*. 2020;14:1177932219899051.
4. Fatima N, Rueda L. iSOM-GSN: an integrative approach for transforming multi-omic data into gene similarity networks via self-organizing maps. *Bioinformatics*. 2020;36:4248-4254.
5. ElKarami B, Alkhateeb A, Qattous H, Alshomali L, Shahrrava B. Multi-omics data integration model based on UMAP embedding and convolutional neural network. *Cancer Inform*. 2022;21:11769351221124205.
6. Meng C, Basunia A, Peters B, Gholami AM, Kuster B, Culhane AC. MOGSA: integrative single sample gene-set analysis of multiple omics data. *Mol Cell Proteom*. 2019;18:S153-S168.
7. Nguyen T, Pepper JW, Nguyen C, et al. Molecular characterization of the highest risk adult patients with acute myeloid leukemia (AML) through multi-omics clustering. *Front Genet*. 2021;12:777094.
8. Soltis AR, Bateman NW, Liu J, et al. Proteogenomic analysis of lung adenocarcinoma reveals tumor heterogeneity, survival determinants, and therapeutically relevant pathways. *Cell Rep Med*. 2022;3:100819.
9. Gao J, Aksoy BA, Dogrusoz U, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal. *Sci Signal*. 2013;6:l1.
10. Dubois E, Galindo AN, Dayon L, Cominetti O. Assessing normalization methods in mass spectrometry-based proteome profiling of clinical samples. *Biosystems*. 2022;215-216:104661.

11. Lazar C, Gatto L, Ferro M, Bruley C, Burger T. Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies. *J Proteome Res*. 2016;15:1116-1125.
12. Ma W, Kim S, Chowdhury S, et al. DreamAI: algorithm for the imputation of proteomics data. 2021. doi: 10.1101/2020.07.21.214205.
13. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst*. 2015;1:417-425.
14. Zhang H, Liu T, Zhang Z, et al. Integrated proteogenomic characterization of human high-grade serous ovarian cancer. *Cell*. 2016;166:755-765.
15. Verhaak RGW, Tamayo P, Yang JY, et al. Prognostically relevant gene signatures of high-grade serous ovarian carcinoma. *J Clin Investig*. 2013;123:517-525.
16. Colaprico A, Silva TC, Olsen C, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res*. 2016;44:e71.
17. Wei L, Jin Z, Yang S, Xu Y, Zhu Y, Ji Y. TCGA-assembler 2: software pipeline for retrieval and processing of TCGA/CPTAC data. *Bioinformatics*. 2018;34:1615-1617.
18. Rudnick PA, Markey SP, Roth J, et al. A description of the Clinical Proteomic Tumor Analysis Consortium (CPTAC) common data analysis pipeline. *J Proteome Res*. 2016;15:1023-1032.