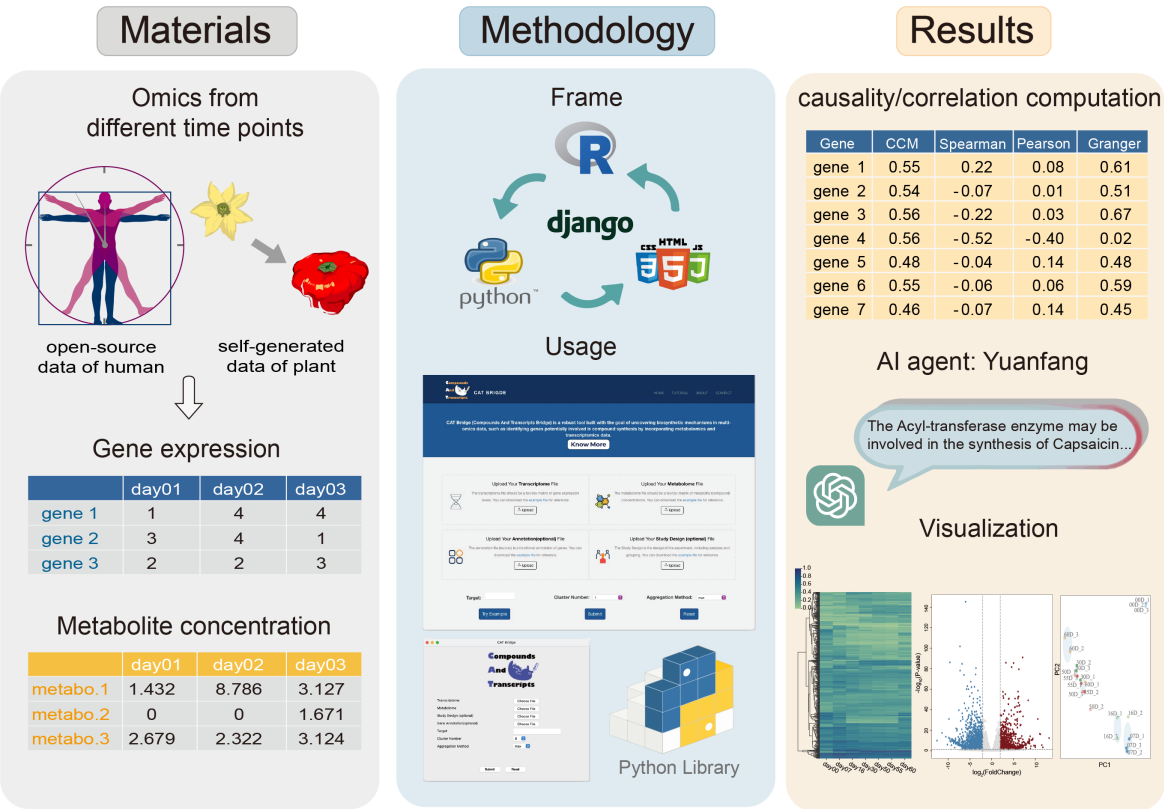
CAT Bridge: an efficient toolkit for compound-transcript association mining from multi-omics data

Bowen Yang[1,2], Tan Meng[1], Xinrui Wang[1], Jun Li[1], Shuang Zhao[3], Yingheng Wang[4], Shu Yi[1], Yi Zhou[1], Yi Zhang[1], Liang Li[2,3,*], Li Guo[1,*]

[1]Peking University Institute of Advanced Agricultural Sciences, Shandong Laboratory of Advanced Agricultural Sciences at Weifang, Weifang, 261325, China
[2]Department of Chemistry, University of Alberta, Edmonton, AB T6G 2G2, Canada
[3]The Metabolomics Innovation Centre, University of Alberta, Edmonton, AB T6G 1C9, Canada
[4]Department of Computer Science, Cornell University, Ithaca, NY 14853, USA

*Correspondence should be addressed to Guo Li. Tel: +86 13325308113; Email: li.guo@pku-iaas.edu.cn
*Correspondence may also be addressed to Liang Li. Tel: +1 (780) 492-3250; Email: liang.li@ualberta.ca

GRAPHICAL ABSTRACT



	day01	day02	day03
gene 1	1	4	4
gene 2	3	4	1
gene 3	2	2	3

	day01	day02	day03
metabo.1	1.432	8.786	3.127
metabo.2	0	0	1.671
metabo.3	2.679	2.322	3.124

Gene	CCM	Spearman	Pearson	Granger
gene 1	0.55	0.22	0.08	0.61
gene 2	0.54	-0.07	0.01	0.51
gene 3	0.56	-0.22	0.03	0.67
gene 4	0.56	-0.52	-0.40	0.02
gene 5	0.48	-0.04	0.14	0.48
gene 6	0.55	-0.06	0.06	0.59
gene 7	0.46	-0.07	0.14	0.45

ABSTRACT

With advancements in sequencing and mass spectrometry technologies, multi-omics data can now be easily acquired for understanding complex biological systems. Nevertheless, substantial challenges remain in determining the association between gene-metabolite pairs due to the complexity of cellular networks. Here, we introduce Compounds and Transcripts Bridge (abbreviated as CAT Bridge, freely available at http://catbridge.work), a user-friendly platform for longitudinal multi-omics analysis to efficiently identify transcripts associated with metabolites using time-series omics data. To evaluate the association of gene-metabolite pairs, CAT Bridge is the first pioneering work benchmarking a set of statistical methods spanning causality estimation and correlation coefficient calculation for multi-omics analysis. Additionally, CAT Bridge featured an artificial intelligence (AI) agent to assist users interpreting the association results. We applied CAT Bridge to self-generated (chili pepper) and public (human) time-series transcriptome and metabolome datasets. CAT Bridge successfully identified genes involved in the biosynthesis of capsaicin in *Capsicum chinense*. Furthermore, case study results showed that the convergent cross mapping (CCM) method outperforms traditional approaches in longitudinal multi-omics analyses. CAT Bridge simplifies access to various established methods for longitudinal multi-omics analysis, and enables researchers to swiftly identify associated gene-metabolite pairs for further validation.

INTRODUCTION

Recent advancements in sequencing and mass spectrometry (MS) technologies have made the acquisition of multi-omics data more cost-efficient and feasible, and comprehensive multi-omics data analysis is crucial for understanding intricate biological mechanisms from a more comprehensive perspective (1-3). For integrated data analysis of transcriptomics and metabolomics, a crucial task is to examine the associated gene-metabolite pairs. Existing strategies bifurcate primarily into two classes: knowledge-driven approaches and data-driven approaches (4). Knowledge-driven approaches have shown their inadequacies for non-model organisms due to the lack of knowledge, restrictions in revealing *de novo* mechanisms, and difficulties in quantifying and ranking their outcomes (4). Data-driven strategies mainly depend on statistical methods that model the correlation of gene-metabolite pairs or sophisticated machine learning methods (5,6). However, due to the severe batch effects in omics data, machine learning approaches usually lack generalizability. Meanwhile, they are also prone to overfit when applied to relatively small omics datasets, making them harder to transfer to different scenarios and less interpretable (4). In terms of statistical methods, people usually calculate the correlation coefficient to match gene-compound pairs (7,8) such as in the studies of the growth cycles of tomato (9) and rice (10), where Pearson correlations were utilized to study metabolic regulatory networks by integrating transcriptomics and metabolomics data. However, these methods face reliability issues, especially when dealing with longitudinal omics data. This is because of the time lag in the expression of genes and metabolites, and the inherent complexity of biological systems, which is a dynamical system with non-linear interactions between different molecules. Furthermore, purely data-driven strategies can occasionally lead to biologically naive conclusions (7). Therefore, integrating two methodologies may offer a more comprehensive and accurate interpretation of multi-omics data.

To address the existing limitations, we have introduced Compounds and Transcripts Bridge (CAT Bridge), a comprehensive cross-platform toolkit that provides a novel analysis pipeline for integrative analysis linking upstream and downstream omics (typically transcriptomics and metabolomics). The novel pipeline encompasses three essential steps, data preprocessing,

computing association between gene-metabolite pairs, and result presentation. For measuring the association, we integrated seven different statistical algorithms on causality estimation and correlation calculation and benchmarked them on human and plant datasets. It also offers three ways to display results that are generated from both data-driven approach and knowledge-driven approach, including common omics statistical analysis and visualization, heuristic ranking of candidate genes based on causality/correlation, and an AI agent driven by large language models (LLMs) to identify associated gene-metabolite pairs through prior knowledge.

Besides, we offer three different access options for CAT Bridge, including a web server, a standalone application, and a Python library. We also provide a detailed tutorial and a sample dataset to help the users get started easily.

MATERIAL AND METHODS

Overview of CAT Bridge

The workflow of CAT Bridge consists of three primary steps, data processing, statistical modeling (causality estimation and correlation calculation), and visualization and interpretation (Figure 1A). Users are required to upload two processed files, gene expression and metabolite concentration matrices, and specify a metabolite of interest as the target. After data pre-processing, seven different causality/correlation algorithms are available for selection to measure the relationships between each gene and the target metabolite. The algorithm chosen by the user will then generate vectors representing gene-metabolite pairwise association (Figure 1B). Subsequently, a vector magnitude is employed for heuristic ranking, with the top 100 genes being reviewed by the AI agent, to utilize prior knowledge to offer inspiration to users. Finally, commonly used omics visualization will be employed to assist users in interpreting the overall expression pattern and facilitating the selection of potential candidate genes.

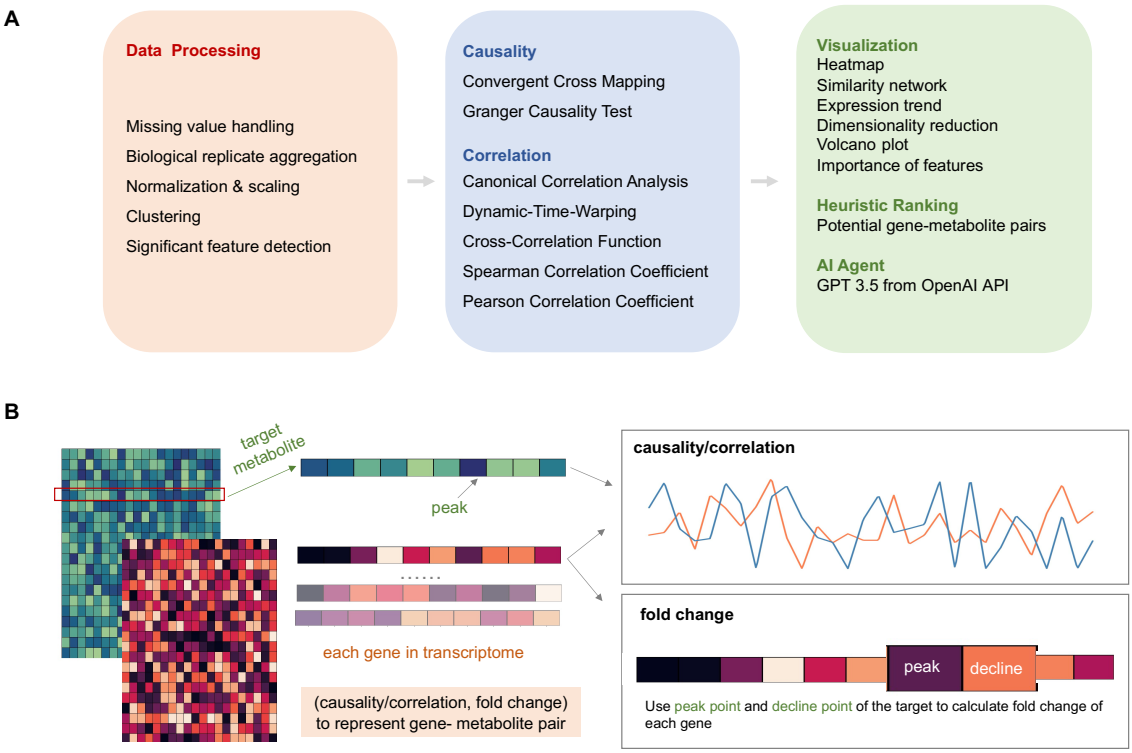


Figure 1. The features and overall workflow of CAT Bridge. (A) The workflow of CAT Bridge consists of three primary steps: data preprocessing, estimation of cause-effect relationships or computation of correlation coefficient, and the presentation of results, which includes visualization, heuristic ranking, and responses from an AI agent. (B) The computation of CAT Bridge involves: extracting the target metabolite from the metabolite concentration matrix, and pinpointing the time point of its maximum concentration as the peak time. Next, the causality or correlation between each gene in the gene expression matrix and the target metabolite is obtained. Then along with the fold change between the peak and decline time points to compose a vector to represent the association between gene-metabolite pairs.

Gene-metabolite association computing and heuristic ranking

For gene-metabolite pair identification such as inferring the biosynthetic genes for particular metabolites, correlation is often used to imply association, such as Spearman Correlation Coefficient (Spearman) and Pearson Correlation Coefficient (Pearson). However, such correlation-based methods have substantial limitations (7,8) because they tend to overlook the non-linearity and lag issues of gene expression leading to metabolite changes in complex biological systems. Therefore, besides considering Spearman and Pearson, we have integrated into CAT Bridge various distinct statistical methods, including Convergent Cross Mapping

(CCM) and Granger Causality (Granger) to identify causality between gene-metabolite pairs, as well as Canonical Correlation Analysis (CCA), Dynamic-Time-Warping (DTW), Cross-Correlation Function (CCF) for calculating correlation coefficient (The implementation methods are provided in the Supplementary text). These algorithms were based on different assumptions, so that some of them allow compatibility with time series data and complex systems. Among them, correlation-based strategies have been widely applied in genomics and multi-omics analysis (11-14). The CCM and Granger, which estimate causality from time series data, are already used in some areas of biology such as ecology and neurobiology but leave a gap in the omics analysis (15-18). The integration of diverse computational methods provides users with the flexibility to select the approach that best suits their needs. Furthermore, our benchmarking results (as detailed in the Results section) suggest that in longitudinal multi-omics studies, causal relationships may provide a more accurate depiction of the association between genes and metabolites.

Fold change (FC) is another measurement frequently used in omics analyses to identify differentially expressed genes (19). CAT Bridge pinpoints the peak time of the target metabolite and calculates each gene's log2 normalized FC of this peak time point and the subsequent decline time point (that is, the next sampling time point after the peak time point) using DESeq2 (20). Then, causality/correlation and FC are combined into a vector to represent the gene-metabolite pair. After the min-max normalization of values (The details are provided in the Supplementary text), the magnitude of this vector is calculated as the CAT score (Figure 1B). This score heuristically ranks the strength of association between each gene and the metabolite. Users can filter putative genes based on thresholds (e.g., 0.5 for causality/correlation, 1 for normalized FC) or manually review them in descending order.

Optionally, if users provide a gene function annotations file, typically derived from homology annotations using tools like InterProScan (21) or eggNOG-mapper (22) for non-model organisms), the CAT score will be adjusted with an additional value. This value is determined by a scoring rule based on the gene's description. By default, genes identified as enzymes receive a score of 0.2, while those with unknown functions get a score of 0.1. Users can

customize this scoring rule based on their specific requirements, depending on the presence of target annotations and their importance.

Finally, knowledge-driven strategies are applied, the top 100 genes in heuristically ranking will be evaluated by the GPT-3 Turbo based AI agent, to identify putative genes on gene's annotation and prior knowledge.

In summary, CAT Bridge provides a platform with a novel pipeline that allows for the rapid identification of putative genes for further investigation and validation.

Visualization and other features

To enhance data interpretation, the CAT Bridge workflow offers a visual ranking of genes based on computation result, and also incorporates a spectrum of widely utilized graphical outputs. Firstly, heatmap is used to present the abundance levels of various genes and metabolites. Such visualization facilitates the discernment of inherent patterns and prevailing trends across the dataset. Secondly, principal component analysis (PCA) is used for reducing dimensions to demonstrate the consistency of biological replicates in single-omics and multi-omics scenarios. This approach is also beneficial in multi-omics integration, ensuring that a matrix with a larger number of features (typically it is transcriptomic data), does not overshadow the combined matrix. Thirdly, the software generates variable importance in projection (VIP) plots for both metabolites and genes, highlighting features that significantly influence the data's variability. Moreover, correlation networks are designed to identify metabolites that show concentration patterns similar to the target metabolite. Finally, the platform also deploys volcano plots for displaying statistical significance against fold change for each gene between peak and decline point. For gene clustering, inspired by Mfuzz (23), the fuzzy c-means algorithm was adopted. The primary aim of this approach was to categorize genes based on expression profiles, thereby deeper insights into their interrelated functions and possible regulatory interplays.

Plant Material Cultivation and Sampling

To test the effectiveness of CAT Bridge across different species, especially its applicability to non-model organisms, we collected transcriptome sequencing and metabolic profiling data from chili peppers (*Capsicum chinese* L.), focusing on one of its trademark natural products, capsaicin. Pepper seedlings were grown in a greenhouse of Peking University Institute of Advanced Agricultural Sciences with a controlled environment of 25°C temperature, a light-dark cycle of 16 hours light and 8 hours dark, and 70% relative humidity. The fruits of peppers were sampled at seven distinct time points, starting from the day of flowering, i.e. 0 day post-anthesis (DPA) during which flowers were collected. Following this, fruits were harvested on days 7, 16, 30, 50, 55, and 60 DPA. The samples were ground and freeze-dried in liquid nitrogen, followed by extraction using 1.0 mL of 70% aqueous methanol for every 50 mg of the sample, and an ultrasonic process was employed for 30 minutes. The preparation of standards was conducted as follows: a mixed standard in the range of 20-50μg/mL was prepared using methanol of MS grade. For the amino acid standard solution, a 1mg/mL stock solution was prepared in water and subsequently diluted with 50% methanol to achieve a concentration of 50μg/mL. Three biological replicates were used in later transcriptome and metabolome analysis.

Metabolome Profiling using HPLC-MS and Data Pre-processing

The metabolome profiling was carried out using untargeted metabolomics based on liquid chromatography coupled with mass spectrometry (LC-MS). The samples were filtered through a 0.22 μm membrane and transferred into the lining tube of a sampling vial. Subsequent centrifugation was carried out at 12000 rcf and 4℃ for 10 minutes. The processed samples were then analyzed using Thermo Scientific Orbitrap Exploris™ 240 (Thermo Fisher Scientific, USA). Chromatographic separation was achieved on a T3 C18 (1.7 μm, 2.1 mm × 150 mm column, USA) maintained at 40℃. The mobile phase consisted of A: 1% formic acid in water and B: 1% formic acid in acetonitrile, with a flow rate of 300 μL/min. A 3 μL sample was injected at an autosampler temperature of 10℃. The elution gradient was set as follows: 0-2.5 min, 3-10% B; 2.5-6 min, 10-44% B; 6-14 min, 44-80% B; 14-20 min, 80-95% B; 20-23 min, 95% B; 23-23.1 min, 95-3% B; 23.1-28 min, 3% B. MS was performed using both positive and

negative ion scans, with a precursor ion scan mode. The auxiliary gas heater temperature was set at 350°C, and the ion transfer tube temperature was also maintained at 350°C. The sheath gas flow rate and auxiliary gas flow rate were set to 35 arb and 15 arb, respectively. The voltages were set to 3.5 KV for the positive spectrum and 3.2 KV for the negative spectrum. For MS1, the scan resolution was 60000, with a scan range of 80-1200. For MS2, the scan resolution was 15000, with a stepped collision energy of 20, 40, and 60 eV. Metabolite identification and quantification were performed using the Compound Discoverer software 3.3 (Thermo Fisher Scientific, USA).

RNA extraction and transcriptome sequencing

Total RNA was isolated from the above collected plant materials using Trizol Reagent (Thermo Fisher, USA) following manufacturer recommended protocol. The quality of RNA extracts was evaluated using RNA Nano 6000 Assay Kit of the Bioanalyzer 2100 system (Agilent Technologies, USA) following manufacturer's recommendation and samples with a RIN value >7 were used in downstream sequencing library construction and sequencing. The library construction was conducted using Illumina True-seq transcriptome kit (Illumina, USA) following standard protocols. Transcriptome sequencing was carried out by Novogene Co., Ltd. The sequencing reads were procured from the Illumina NovaSeq 6000 platform. For pre-processing, fastp (24) was employed to conduct quality control and clean the data. Subsequently, these reads were mapped to the Capsicum chinense cultivar PI159236 genome (25) using STAR (26). StringTie (27) was utilized to quantify and assess the expression levels of the genes that were successfully mapped. The biological function annotation of genes is obtained through the eggNOG-mapper.

Acquisition and processing of public human dataset

We also collected human multi-omics data from a public dataset to further examine the performance of CAT Bridge. The data was sourced from a published precision medicine study (28) that conducted a deep longitudinal multi-omics analysis of 105 individuals over four years. We downloaded the transcriptomic and metabolomic data from the dataset, and generated functional annotations for each gene using Entrez Programming Utilities. Then we chose the

participant with the most extensive time points to provide the most time points. Furthermore, as the intervals of the sampling times were not uniform, only time points where both transcriptomics and metabolomics were sampled were retained, and only the earliest time point for the month was preserved if more than one time point was sampled in one month. Then, we calculated the slopes for all metabolites using linear regression to select target metabolite for the following analysis.

RESULTS

CAT Bridge offers three distinct usage modes and caters to a wide range of user requirements. Firstly, it features a web server, designed for user-friendliness and accessibility, and is open to all users without any login requirements. This is particularly beneficial to those who are less familiar with programming language. Secondly, a standalone application is available for users handling large data files, as it lifts the constraints on file sizes. Finally, a Python library is available for bioinformaticians with complete features and customizable workflows (The implementation of the software is provided in the Supplementary text).

To showcase the utility and features of CAT Bridge, we applied it to both a self-generated dataset collected from chili peppers and a publicly available human dataset in two case studies. Our analysis revealed that in the context of longitudinal multi-omics, causality-based strategies tend to outperform those solely based on correlation. As such, we advocate for the adoption of causality rather than correlation in longitudinal multi-omics analysis such as co-mining the transcriptomics and metabolomics data. The capsaicin dataset generated in this study has been made open source for user exploration.

Case Study 1: Identifying genes associated with capsaicin biosynthesis in chili pepper

In our inaugural case study, we leveraged self-collected non-model organism data to examine the performance of CAT Bridge. This data comprised the transcriptome and metabolome of peppers at seven different developmental stages after bloom (Figure 2A). Capsaicin, an important natural product produced by chili peppers that gives fruit pungency and has potential anti-cancer and analgesic activity (29), was selected as the target metabolite for this study.

Time-series transcriptome and metabolic profiling of developing chili pepper fruits were used as input data to test CAT Bridge.

Through examination using the CCM method for hypothetical ranking, BC332_05016 encoding an Acyl-transferase was ranked first, regardless of whether an annotation file was provided. The Result suggests that this gene was more likely to be the synthetic gene associated with capsaicin in *C. chinense* (Figure 2B). The complete heuristic ranking results are provided in Supplementary Table S1. BLAST search revealed that BC332_05016 was homologous of PUN1 (sequence identity: 100%), a.k.a AT3 (Acyl-transferase 3) or CS (capsaicin synthesis) gene (30). Moreover, when common thresholds were applied for screening, only CCM passed the criteria. The causality modeled based on CCM was 0.55, implying a strong association between BC332_05016 and capsaicin. By contrast, the conventional Pearson correlation method produced a result of 0.08, which would fall below the commonly used threshold and potentially lead to an overlook of this gene-metabolite pair (Figure 2C). The AI agent also accurately found BC332_05016 among the top 100 genes based on functional annotation (Figure 2D). Furthermore, CAT Bridge visualization result showed that capsaicinoids such as nonivamide, dihydrocapsaicin, and homocapsaicin have high similarity to capsaicin (Figure 2E) and may play a significant role in response to the variable (Figure 2F). The rest of the visualization results are provided in the Supplementary text. These results show that the CAT Bridge is a valuable tool in multi-omics analysis to reliably identify associated gene-metabolite pairs.

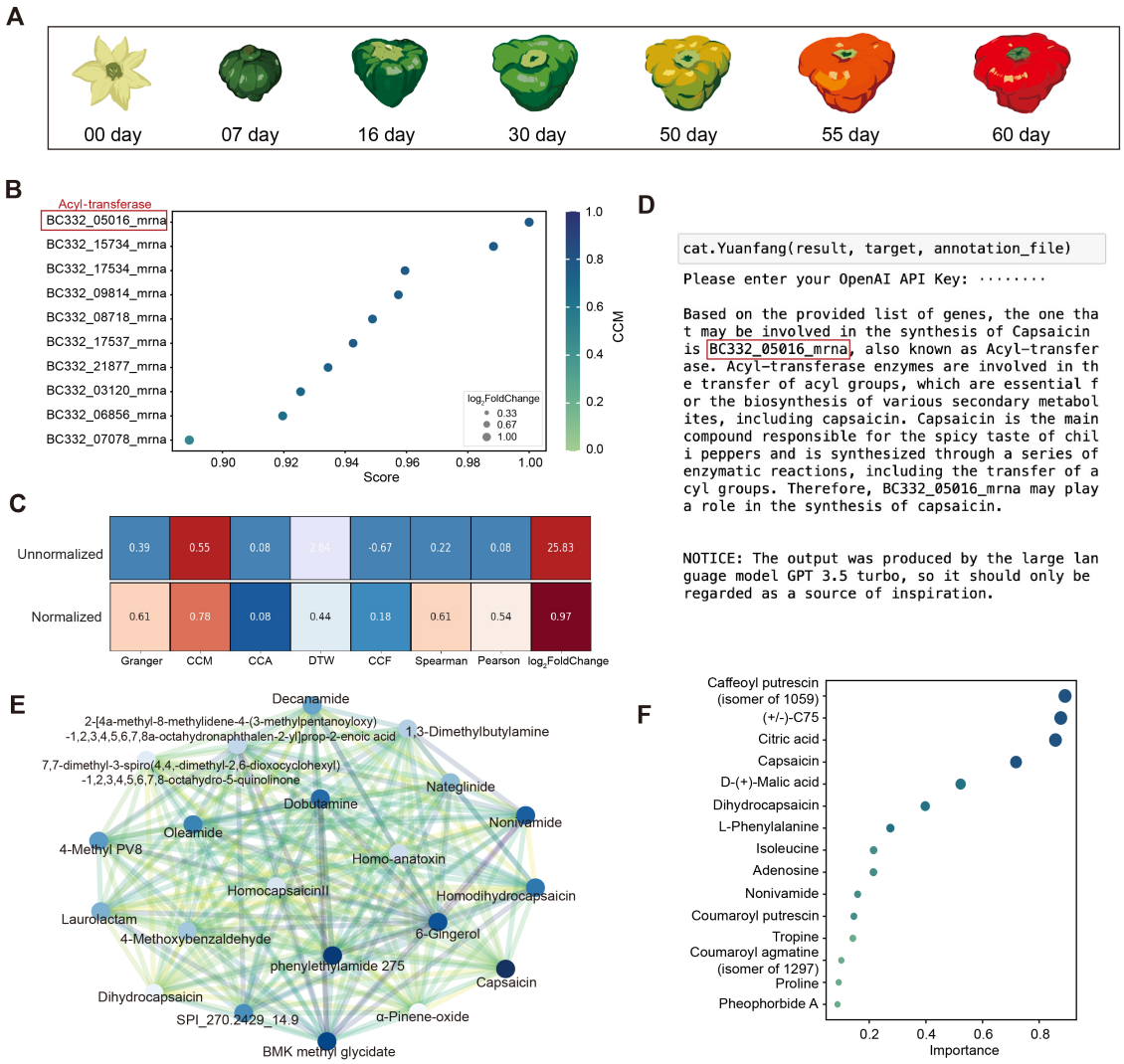


Figure 2. Application of CAT Bridge to mine transcript-capsaicin association. (A) Diagram showing the time points sampled for transcriptome and metabolic profiling during fruit development of chili peppers in case study 1. (B) Heuristic ranking produced using the CCM-based method. (C) Comparative values across different methods. For unnormalized values: red indicates a strong association; original denotes medium association; blue suggests values that are below the commonly used threshold, show no association, or are negatively associated (depending on the method); light blue means this method does not adhere to a common threshold. For normalized values: red signifies values that are high after min-max normalization; blue represents low normalized values. (D) Interpretation results derived from the AI agent (E) The correlation network of capsaicin. (F) The significance of metabolites.

Case Study 2: Identifying association of creatinine and GAMT using human data

For case study 2, we used the publicly available human datasets, a total of 48 time points were retained for validation (Figure 3A). After the removal of lipids, creatinine was selected as the target metabolite, due to having the highest slope.

Previous studies have shown that the enzyme guanidinoacetate N-methyltransferase (GAMT) can methylate guanidinoacetate to creatine, and creatine spontaneously converts to creatinine subsequently (31). And the consistent reduction in urinary creatinine excretion is an unspecific indicator of GAMT deficiency (32). These findings indicate that GAMT is a key gene involved in creatinine regulation. The result from the CAT Bridge revealed that 61 genes, including GAMT, were suggested to have a moderate to strong association with creatinine based on CCM and FC (complete heuristic ranking results are provided in Supplementary Table S2). However, computed values of GAMT-creatinine from other methods indicate a lack of association (Figure 3B). The AI agent also successfully identified GAMT as a candidate gene among the top 100 genes based on domain knowledge (Figure 3C). Interestingly, even if the values for CCM and FC are lower than those for the BC332_05016-capsaicin pair in case study 1, their values after min-max normalization are higher. The complexities inherent in human regulatory mechanisms and the stability of molecular levels during the sampling timeframe might account for this observation. In addition, creatinine is also shown as an important determinant of variance in the metabolome (Figure 3D).

These results demonstrate the CAT Bridge's potential to extract meaningful insights from multi-omics data across diverse species. By integrating time-series analysis methods, particularly CCM, it offers superior performance in longitudinal omics compared to common methods.

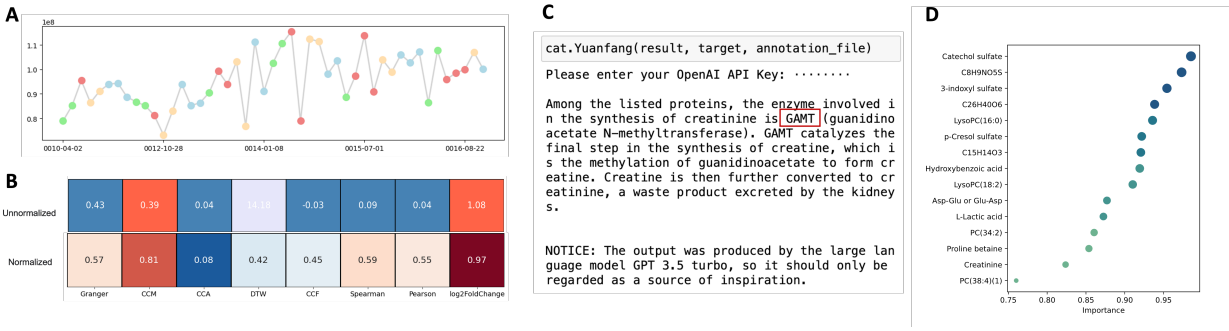


Figure 3. **Application of CAT Bridge to mine transcript-creatinine association in human** (A) Concentration of creatinine across sampling time points. The color of the dots indicates the season: green for spring, red for summer, yellow for fall, and blue for winter. (B) Comparative values across different methods. For color interpretation, refer to the annotation in Figure 2. (C) Interpretation results derived from the AI agent. (D) The significance of the metabolites.

Comparison with other web-based tools

Table 1 displays the function coverage comparisons between CAT Bridge and other data-driven multi-omics analysis web-based tools, including OmicsAnalyst (4), 3omics (33), IntLIM (34), and CorDiffViz (35). In association identification, IntLIM, 3omics, and CorDiffViz integrate either Pearson or Spearman correlations, or both, to aid in the discovery of feature relationships. What sets CAT Bridge apart is its assembly of various algorithms that handle time-series data and causality, and incorporate an AI agent to inspire users. Notably, the performance of CCM has been found from two previous case studies to be potentially more suitable for longitudinal multi-omics analysis compared to traditional methods.

Table 1. Comparison of CAT Bridge with other web-based multi-omics tools. Symbols indicating feature assessments: '✓' denotes presence, '×' signifies absence.

	CAT Bridg	OmicsAnalyst	3omics	IntLIM	CorDiffViz
Pre-processing	✓	✓	×	×	×
Visual analytics	✓	✓	✓	✓	✓
Cross-omics association	✓	×	✓	✓	✓
Cross-platform	✓	×	×	✓	✓
Longitudinal study	✓	×	×	×	×
AI agent	✓	×	×	×	×

OmicsAnalyst: https://www.omicsanalyst.ca; 3omics: https://3omics.cmdm.tw; IntLIM: https://intlim.ncats.io; CorDiffViz: https://diffcornet.github.io/CorDiffViz/demo.html.

DISCUSSION AND CONCLUSIONS

In recent years, there has been a surge in multi-omics research. A critical aspect often overlooked in such studies is the unique nature of the longitudinal experimental design. Longitudinal omics analysis is particularly important in research on the developmental cycle of plants and investigations related to chronic diseases and aging (36-39). However, many studies tend to use generic methodologies for analysis (9,10). This may inadvertently miss key discoveries. CAT Bridge provides a platform specifically for longitudinal multi-omics analysis, by drawing insights from disciplines where time series data is more prevalent, and validating them with data. Through the gene-metabolite causality/correlation modeling method, combined with visualization tools and AI assistance, researchers can more quickly identify putative genes for experimental validation.

In two case studies, CCM showed its superiority compared to other methods, which is probably due to its modeling capability of dynamical systems. Thus, it can better capture the complex non-linear interactions within biological systems(15). We advocate for modeling cause-and-

effect relationships in longitudinal omics analyses, instead of more widely used Pearson or Spearman Correlations. However, this doesn't mean that CCM is always appropriate. Factors such as sampling intervals and the number of samples also need to be considered. More precise methods for estimating cause-and-effect relationships, as well as post-processing for vector represent gene-metabolite pairs are both required to explore and validate by using more data.

Aside from computational methods, the reliability of analytical results is also influenced by experimental design and data acquisition methods. Increasing the number of sampling time points and setting a reasonable interval between them can enhance the power of association and credibility of the results. On the data acquisition front, it is recommended to annotate the transcriptome with an updated, high-quality reference genome, and employ advanced metabolomics techniques such as chemical isotope labeling liquid (CIL) LC-MS (40) to ensure a high coverage and more accurate relative quantification of metabolome.

DATA AVAILABILITY

The CAT Bridge web server is freely available to all users at http://www.catbridge.work, the source code and standalone version of CAT Bridge can be found at https://github.com/Bowen999/CAT-Bridge. Sequencing data of case study 1 have been deposited in the Small Read Archive (SRA) (http://www.ncbi.nlm.nih.gov/sra) with the BioProject accession code PRJNA1030882.

SUPPLEMENTARY DATA

Supplementary Data includes a text and two tables.

AUTHOR CONTRIBUTIONS

Conceptualization, B Yang, L Guo; Software, B Yang, T Meng, Y Zhou, Y Zhang; Methodology, B Yang, Y Wang; Visualization, B Yang, X Wang; Sample and data collection: J Li, B Yang, S Yi; Writing—original draft, B Yang; Writing—review & editing, L Guo, Y Wang, S Zhao, L Li; Supervision, L Guo, L Li; Funding acquisition, L Guo. All authors have read and agreed to the published version of the manuscript.

ACKNOWLEDGEMENTS

We would like to thank the Bioinformatics Platform at Peking University Institute of Advanced Agricultural Sciences for providing the high-performance computing resources.

FUNDING

This work was supported by Shandong Provincial Science and Technology Innovation Fund and Natural Science Foundation for Distinguished Young Scholars (ZR2023JQ010) of Shandong Province. LG is also supported by Taishan Scholars Program of Shandong Province.

CONFLICT OF INTEREST

There is no conflict of interest.

REFERENCES

1. Wörheide, M.A., Krumsiek, J., Kastenmüller, G. and Arnold, M. (2021) Multi-omics integration in biomedical research - A metabolomics-centric review. *Anal Chim Acta*, 1141, 144-162.
2. Hasin, Y., Seldin, M. and Lusis, A. (2017) Multi-omics approaches to disease. *Genome Biology*, 18, 83.
3. Subramanian, I., Verma, S., Kumar, S., Jere, A. and Anamika, K. (2020) Multi-omics Data Integration, Interpretation, and Its Application. *Bioinform Biol Insights*, 14, 1177932219899051.
4. Zhou, G., Ewald, J. and Xia, J. (2021) OmicsAnalyst: a comprehensive web-based platform for visual analytics of multi-omics data. *Nucleic Acids Research*, 49, W476-W482.
5. Krassowski, M., Das, V., Sahu, S.K. and Misra, B.B. (2020) State of the Field in Multi-Omics Research: From Computational Needs to Data Mining and Sharing. *Front Genet*, 11, 610798.
6. Athieniti, E. and Spyrou, G.M. (2023) A guide to multi-omics data collection and integration for translational medicine. *Computational and Structural Biotechnology Journal*, 21, 134-149.
7. Cavill, R., Jennen, D., Kleinjans, J. and Briedé, J.J. (2015) Transcriptomic and metabolomic data integration. *Briefings in Bioinformatics*, 17, 891-901.
8. Chong, J. and Xia, J. (2017) Computational Approaches for Integrative Analysis of the Metabolome and Microbiome. *Metabolites*, 7, 62.
9. Li, Y., Chen, Y., Zhou, L., You, S., Deng, H., Chen, Y., Alseekh, S., Yuan, Y., Fu, R. and Zhang, Z. (2020) MicroTom metabolic network: rewiring tomato metabolic regulatory network throughout the growth cycle. *Molecular plant*, 13, 1203-1218.
10. Yang, C., Shen, S., Zhou, S., Li, Y., Mao, Y., Zhou, J., Shi, Y., An, L., Zhou, Q. and Peng, W. (2022) Rice metabolic regulatory network spanning the entire life cycle. *Molecular Plant*, 15, 258-275.
11. Rockwood, A.L., Crockett, D.K., Oliphant, J.R. and Elenitoba-Johnson, K.S. (2005) Sequence alignment by cross-correlation. *J Biomol Tech*, 16, 453-458.
12. Skutkova, H., Vitek, M., Babula, P., Kizek, R. and Provaznik, I. (2013) Classification of genomic signals using dynamic time warping. *BMC Bioinformatics*, 14, S1.
13. Seoane, J.A., Campbell, C., Day, I.N., Casas, J.P. and Gaunt, T.R. (2014) Canonical correlation

analysis for gene-based pleiotropy discovery. *PLoS Comput Biol*, 10, e1003876.

14. Jiang, M.Z., Aguet, F., Ardlie, K., Chen, J., Cornell, E., Cruz, D., Durda, P., Gabriel, S.B., Gerszten, R.E., Guo, X. *et al.* (2023) Canonical correlation analysis for multi-omics: Application to cross-cohort analysis. *PLoS Genet*, 19, e1010517.
15. Yuan, A.E. and Shou, W. (2022) Data-driven causal analysis of observational biological time series. *eLife*, 11, e72518.
16. Heerah, S., Molinari, R., Guerrier, S. and Marshall-Colon, A. (2021) Granger-causal testing for irregularly sampled time series with application to nitrogen signalling in Arabidopsis. *Bioinformatics*, 37, 2450-2460.
17. Ye, H., Deyle, E.R., Gilarranz, L.J. and Sugihara, G. (2015) Distinguishing time-delayed causal interactions using convergent cross mapping. *Scientific Reports*, 5, 14750.
18. Stokes, P.A. and Purdon, P.L. (2017) A study of problems encountered in Granger causality analysis from a neuroscience perspective. *Proc Natl Acad Sci U S A*, 114, E7063-e7072.
19. Arora, S., Pattwell, S.S., Holland, E.C. and Bolouri, H. (2020) Variability in estimated gene expression among commonly used RNA-seq pipelines. *Scientific Reports*, 10, 2734.
20. Love, M.I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15, 550.
21. Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G. et al. (2014) InterProScan 5: genome-scale protein function classification. Bioinformatics, 30, 1236-1240.
22. Cantalapiedra, C.P., Hernández-Plaza, A., Letunic, I., Bork, P. and Huerta-Cepas, J. (2021) eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution*, 38, 5825-5829.
23. Kumar, L. and M, E.F. (2007) Mfuzz: a software package for soft clustering of microarray data. *Bioinformation*, 2, 5-7.
24. Chen, S., Zhou, Y., Chen, Y. and Gu, J. (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34, i884-i890.
25. Kim, S., Park, J., Yeom, S.I., Kim, Y.M., Seo, E., Kim, K.T., Kim, M.S., Lee, J.M., Cheong, K., Shin, H.S. *et al.* (2017) New reference genome sequences of hot pepper reveal the massive evolution of plant disease-resistance genes by retroduplication. *Genome Biol*, 18, 210.

26. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29, 15-21.

27. Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T. and Salzberg, S.L. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, 33, 290-295.

28. Sailani, M.R., Metwally, A.A., Zhou, W., Rose, S.M.S.-F., Ahadi, S., Contrepois, K., Mishra, T., Zhang, M.J., Kidziński, Ł., Chu, T.J. *et al.* (2020) Deep longitudinal multiomics profiling reveals two biological seasonal patterns in California. *Nature Communications*, 11, 4933.

29. Fattori, V., Hohmann, M.S., Rossaneis, A.C., Pinho-Ribeiro, F.A. and Verri, W.A. (2016) Capsaicin: Current Understanding of Its Mechanisms and Therapy of Pain and Other Pre-Clinical and Clinical Uses. *Molecules*, 21.

30. Kim, S., Park, M., Yeom, S.-I., Kim, Y.-M., Lee, J.M., Lee, H.-A., Seo, E., Choi, J., Cheong, K., Kim, K.-T. *et al.* (2014) Genome sequence of the hot pepper provides insights into the evolution of pungency in Capsicum species. *Nature Genetics*, 46, 270-278.

31. da Silva, R.P., Nissim, I., Brosnan, M.E. and Brosnan, J.T. (2009) Creatine synthesis: hepatic metabolism of guanidinoacetate and creatine in the rat in vitro and in vivo. *Am J Physiol Endocrinol Metab*, 296, E256-261.

32. Stockler-Ipsiroglu, S., van Karnebeek, C., Longo, N., Korenke, G.C., Mercimek-Mahmutoglu, S., Marquart, I., Barshop, B., Grolik, C., Schlune, A., Angle, B. *et al.* (2014) Guanidinoacetate methyltransferase (GAMT) deficiency: Outcomes in 48 individuals and recommendations for diagnosis, treatment and monitoring. *Molecular Genetics and Metabolism*, 111, 16-25.

33. Kuo, T.C., Tian, T.F. and Tseng, Y.J. (2013) 3Omics: a web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data. *BMC Syst Biol*, 7, 64.

34. Siddiqui, J.K., Baskin, E., Liu, M., Cantemir-Stone, C.Z., Zhang, B., Bonneville, R., McElroy, J.P., Coombes, K.R. and Mathé, E.A. (2018) IntLIM: integration using linear models of metabolomics and gene expression data. *BMC Bioinformatics*, 19, 81.

35. Yu, S., Drton, M., Promislow, D.E.L. and Shojaie, A. (2021) CorDiffViz: an R package for visualizing multi-omics differential correlation networks. *BMC Bioinformatics*, 22, 486.

36. Kudryashova, K.S., Burka, K., Kulaga, A.Y., Vorobyeva, N.S. and Kennedy, B.K. (2020) Aging Biomarkers: From Functional Tests to Multi-Omics Approaches. *Proteomics*, 20, e1900408.

37. Cellerino, A. and Ori, A. (2017) What have we learned on aging from omics studies? *Seminars in Cell & Developmental Biology*, 70, 177-189.

38. Allegri, M., Gregori, M.D., Minella, C.E., Klersy, C., Wang, W., Sim, M., Gieger, C., Manz, J., Pemberton, I.K., MacDougall, J. *et al.* (2016) 'Omics' biomarkers associated with chronic low back pain: protocol of a retrospective longitudinal study. *BMJ Open*, 6, e012070.

39. Mars, R.A.T., Yang, Y., Ward, T., Houtti, M., Priya, S., Lekatz, H.R., Tang, X., Sun, Z., Kalari, K.R., Korem, T. *et al.* (2020) Longitudinal Multi-omics Reveals Subset-Specific Mechanisms Underlying Irritable Bowel Syndrome. *Cell*, 182, 1460-1473.e1417.

40. zhao, S., Li, H., Han, W., Chan, W. and Li, L. (2019) Metabolomic Coverage of Chemical-Group-Submetabolome Analysis: Group Classification and Four-Channel Chemical Isotope Labeling LC-MS. Analytical Chemistry, 91, 12108-12115.