

Active Data Canvas: web-based visual analytic tool to link data to knowledge

Joon-Yong Lee¹, Ryan Wilson¹, Gary R. Kiebel¹, Grant M. Fujimoto¹, Michael G. Degan¹, Vladislav A. Petyuk¹, Richard D. Smith¹, Nick Cramer², and Samuel H. Payne¹
¹Biological Sciences Division and ²National Security Directorate, Pacific Northwest National Laboratory, Richland, WA

Overview

- A myriad of high-throughput multi-omics data is continuously generated.
- Domain knowledge and intuition are essential to derive and test hypotheses.
- Analysis requires diverse expertise and collaboration.
- **Active Data Canvas** makes it easy: a web-based intuitive, interactive and collaborative visual analytic tool to enable knowledge discovery from multi-omic data.

Introduction

- Data deluge in multi-omic experiments (transcriptome, proteome, metabolome, etc.)
- Requiring computational acumen limits productive browsing to explore/form conclusions.
- Logistical complications among collaborators with different domain knowledge can limit effective data analysis and scientific discovery.
- Visual analytic tools allow users to interact with data, and greatly improve the speed and quality of analysis.
- Some tools offer web-based interactive heatmap (e.g. Next-Generation Clustered Heat Map and IntOGen), but they lack the ability to link data to external knowledge-bases and to share coworker's analysis.

Results

Customizable Data Import

- In order to visually display the interactive heatmap, users upload either a **tabular formatted file (CSV)** or a **RData file** including the clustering information as well as raw data.
- Users can customize the list of genesets for enrichment tests.
- Metadata files describing the experiment and samples (e.g. clinical measurements) are used to identify statistically significant attributes of a group of samples.

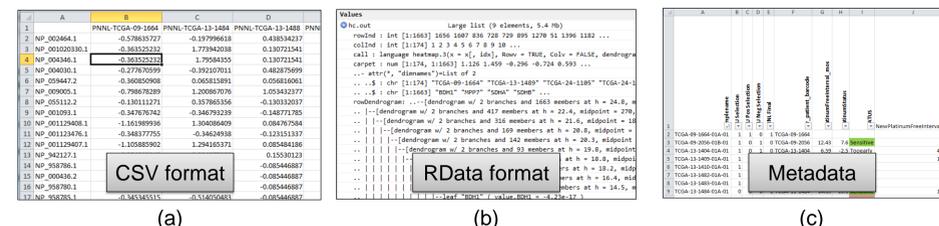


Figure 2 – Example of input data files. (a) CSV format of protein abundance levels across tumor samples. Column and row names indicate the tumor sample IDs and protein IDs, respectively. (b) RData format of protein abundance levels across tumor samples. (c) CSV format of metadata for tumor samples.

Interactive Data Viewers

- Every node in heatmap cluster can be clicked for statistical significant tests.
- Pathway viewer overlays data on the familiar visual context.

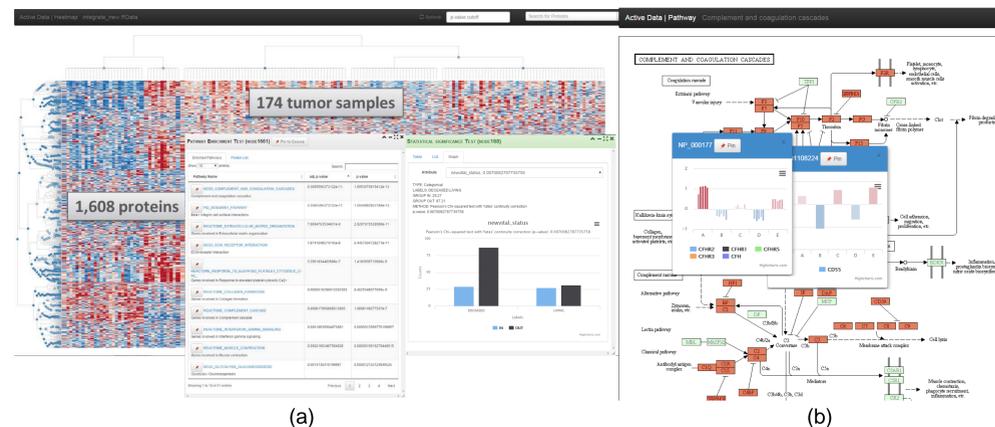


Figure 3 – Interactive Data Viewers: (a) Heatmap viewer. It displays the clickable dendrograms as well as heatmap. If you click the nodes in row, you can get the enriched pathways associated with the selected protein group. When you click the column node, you can get the statistical results and its graphs. (b) Pathway viewer. On the top of a pathway (Complement and coagulation cascades), raw data is integrated.

Data Canvas

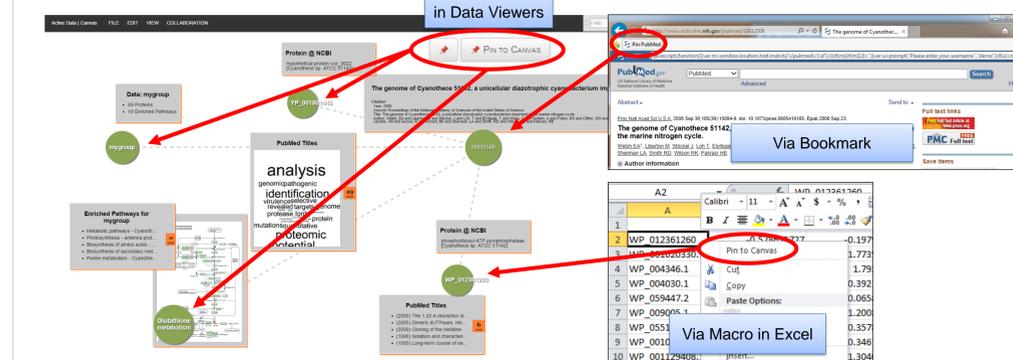


Figure 4 – Data Canvas. From other viewers, users can pin the entities as they want through RESTful APIs. In canvas, a green seed appears. Data canvas has diverse assistant modules to recommend the knowledge by fetching the relevant data associated with the entity. It provides the recommendation cards for PubMed, NCBI, KEGG, and so on.

Data Sharing

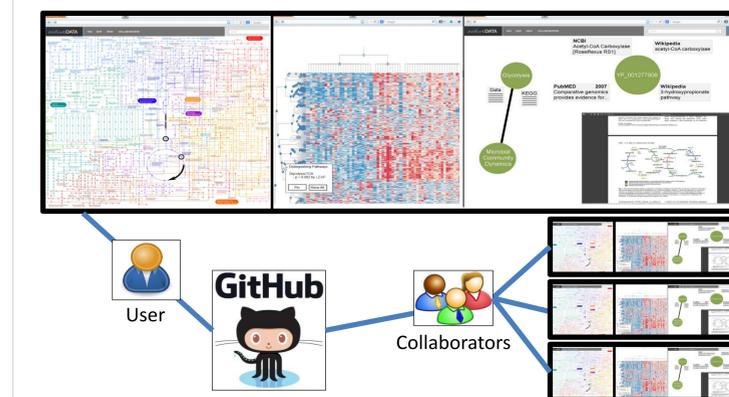


Figure 5 – For enhancing collaboration, active data canvas employs GitHub which is the largest code hoster in the world and is usually used to host open-source software projects. GitHub is used to control revisions of user canvas as well as data itself and to share canvases with each other.

Conclusions

- Active data canvas facilitates more effective collaborations based on the interactive visual analytic tools and data sharing/versioning strategy.
- Pinning data can dramatically reduce the time to proactively explore the extensive external resources for gathering the relevant information.
- Speed hypothesis testing and scientific discovery from large-scale omics data.

Methods

Active Data Canvas

- Provides multiple data viewers implemented as web applications.
- For data analysis, gene-set enrichment tests and statistical tests are conducted across rows and columns.
- Data viewers communicate with the Canvas via API to pin (save) interesting data.
- Software assistants proactively research external knowledge sources associated with pinned items.
- All data and analysis are versioned via GitHub.

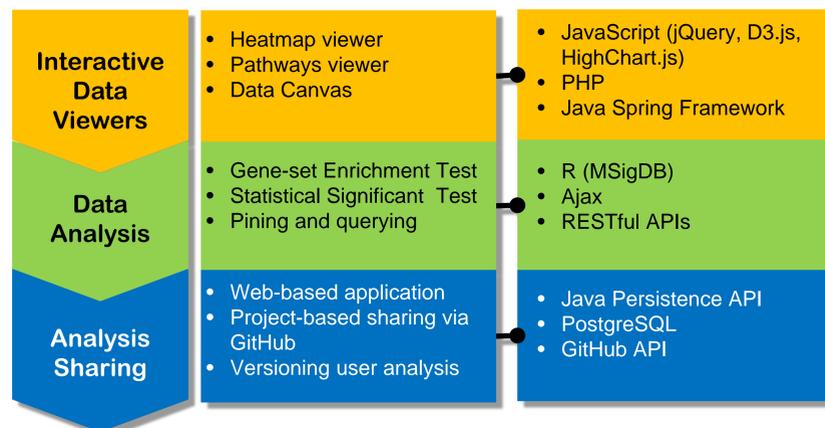


Figure 1 – Active Data Canvas has three main features/components. First, the interactive data viewers are web applications so that collaborators can easily access and share data. Second, it proactively conducts statistical analysis, and generates queries to research the domain knowledge to assist analyses. Third, all data and analysis is versions and shared via GitHub.

CONTACT: Joon-Yong Lee, Ph.D.
Biological Sciences Division
Pacific Northwest National Laboratory
E-mail: joonyong.lee@pnnl.gov

Acknowledgements

This work was funded by the Laboratory Directed Research and Development Program (LDRD) at Pacific Northwest National Laboratory (PNNL) and an Early Career Award from the US Department of Energy (DOE) Office of Biological and Environmental Research (OBER) to SHP. Data for the project was created using the National Cancer Institute (NCI) CPTAC awards U24-CA-160019. Samples were analyzed using capabilities developed under the support of NIH National Institute of General Medical Sciences (GM103493) and the Pan-omics program supported by DOE/OBER, and performed in the Environmental Molecular Sciences Laboratory, a DOE OBER national scientific user facility on the PNNL campus. PNNL is a multiprogram national laboratory operated by Battelle for the DOE under contract DE-AC05-76RL01830.

www.omics.pnl.gov

