OmicsOne: Associate Omics Data with Phenotypes in One-Click

Yingwei Hu^{1*}, Minghui Ao¹ and Hui Zhang¹

¹Johns Hopkins University, School of Medicine, Baltimore, MD 21287, USA

*Correspondence to Yingwei Hu (yhu39@[jhmi.edu](mailto:yhu39@jhmi.edu))

Abstract

The rapid advancements of high-throughput “omics” technologies have brought huge amount of data to process during and after experiments. Multi-omic analysis facilitates a deeper interrogation of a dataset, and discovery of interesting genes, proteins, lipids, glycans, or metabolites, or pathways related to the corresponding phenotypes in a study. Many individual software tools have been developed to analyze and visualize the data. However, integrating multiple omics data analysis strategies and approaches in a single data processing pipeline is still a challenge task. OmicsOne is a software developed in R, Python and Jupyter Notebook that can achieve statistical analysis, machine learning, and data visualization on multi-‘omics’ data by taking the advantages of integrating the useful tools from individual software packages. OmicsOne can simplify “omics” data analysis, and delineate molecules, or pathways associated to interested phenotypes.

Introduction

The advancements of high-throughput “omics” technologies, such as genomics, epigenomics, transcriptomics, proteomics, protein modifications, glycomics, lipidomics, and metabolomics, have produced incredible volume of data¹⁻⁶. It is predicted that the trend of generating large datasets will continue as novel technologies are developed and current approaches advance. In this era of omics data explosion, a simple, efficient bioinformatic analysis tool is critical to work with the large amount of data.

Individual tools, such as GiaPronto⁷, PANDA-view⁸, ImageGP (<http://www.ehbio.com/ImageGP>), offer select options for data analysis and visualization. However, integrating multiple data analysis strategies and approaches in a single data processing pipeline facilitates a deeper interrogation of a dataset, and discovery of interesting molecules or pathways related to the corresponding phenotypes in a study.

Here, we present the tool OmicsOne, a software developed in R, Python and Jupyter Notebook that can preprocess and analyze the multi-“omics” data. After user selected parameters, the results could be reported and visualized in a simple, one-click format. The pipeline includes statistical analysis, machine learning, and data visualization components based on several predefined templates integrated from previous publications^{4,6}. OmicsOne can simplify “omics” data analysis, and delineate genes and pathways associated to phenotypes. OmicsOne is hosted on GitHub (<https://github.com/huizhanglab-jhu/OmicsOne>) and can be downloaded to run locally or on cloud directly via mybinder.org.

Methods

OmicsOne is developed for intra-omics studies to discover molecular changes and pathways associated with phenotypes. The OmicsOne platform is composed of several essential data processing modules, including preprocessing, profiling, quality control, differential expression analysis, dimensionality reduction, clustering, and visualized reporting (Table 1). The phenotype associations are investigated via the identification of molecules and pathways enriched in phenotypical classifications and statistical calculation of the significant ones. In the following sections, OmicsOne was applied on the public proteomic data set of ovarian cancer to demonstrate the functions⁶.

Input data and preprocess

OmicsOne requires a table of phenotype information and a table of expression matrix of quantitative “omics” data (.csv or .xlsx) both in ‘wide format’. In the phenotype information table, each row represents a sample, while each column represents a phenotype. In the expression matrix, each column represents a sample, while each row represents a feature value, such as gene, protein, peptide, or post-translational modification (PTM) expression level.

The expression matrix will be preprocessed before further analysis if necessary. OmicsOne provides common preprocessing functions, including log-transformation, noise filter, missing value filtration, and normalization.

Data profiling and quality control

Understanding the data is always the first and critical step for all the following analysis. OmicsOne supports a series of statistical functions (e.g. sum, average, median, standard deviation) to profile samples, features (e.g. proteins), and phenotype information (Figure 1), and visualize the results in various figure types, including bar chart (Figure 1A), heatmap (Figure 1B), and etc. It is also important to evaluate the quality of the data to secure starting analysis from a confident data source. The software uses the calculation of the correlation values of technical/biological replicates (Figure 1C) or coefficient of variance (CV) of selected quality control samples (Figure 1D) to evaluate the reproducibility of measured gene or protein level expression. OmicsOne also offers a correlation assessment to evaluate the concordance of the genomic, transcriptomics, or proteomic expression across multiple samples. In addition, samples or features considered as outliers relative to the larger dataset can be highlighted for additional consideration for inclusion or follow-up analyses.

Differential expression analysis (DEA)

Delineating altered expression profiles of genes, proteins, and/or PTMs offer the greatest insight into aberrant biology in comparative studies (i.e. tumor vs. normal, aggressive vs non-aggressive, treated and untreated). OmicsOne can identify the significant, differentially expressed genes or proteins, leveraging multiple statistical tests (t-test, wilcoxon and etc., Figure 2A), followed by subsequent enrichment analysis (GSEAPy⁹⁻¹¹ or WebGestaltR¹²⁻¹⁵) to discover up-/down-regulated pathways (Figure 2B). OmicsOne will further calculate the association between the discovered pathways and phenotypes to obtain a comprehensive understanding of the roles of pathways in different conditions (Figure 2C).

Dimensionality reduction and clustering analysis

Unsupervised methods are useful approaches to classify samples based on the most distinguishable features without prior knowledge; often identifying the most prominent factors driving different phenotypes. Principal component analysis (PCA) and Hierarchical clustering (HC) are two wide-spread methods, supported by Python package: scikit-learn¹⁶, to cluster samples and identify the signature gene groups associated with the corresponding sample clusters (Figure 3A and 3B). OmicsOne also implemented a customized interface for executing the R packages: CancerSubtypes¹⁷ and WebGestaltR to evaluate clustering stability and automate gene clustering interpretation. In addition, the significance of identified gene clusters respective to each phenotype or sample cluster can be calculated. Finally, OmicsOne further calculates p-values of associations between the phenotypes and enriched pathways to aid with data application towards biological interpretation.

Data visualization and report

All the analysis completed by OmicsOne report intermediate and finalized results in tables (.csv), as well as the corresponding figures. Figures are generated using matplotlib (<https://matplotlib.org/>) and seaborn (<https://seaborn.pydata.org/>) packages of Python. OmicsOne defines highly customizable template functions to generate common plotting functions (e.g. bar chart, heatmap, volcano plot, histogram, etc.) to support various approaches of data visualization.

Conclusion

In summary, OmicsOne is an efficient tool to associate the molecular changes with phenotypes. It was originally designed for isobarically labeled quantitative proteomics data (e.g. tandem-mass-

tag (TMT) but can find applications in label-free quantitation and Data Independent Acquisition (DIA) datasets, as well as other “omics” data with proper preprocessing. The software uses predefined templates to build a robust, working pipeline for standard association analyses. Overall, OmicsOne will offer end users a simple, bioinformatic pipeline to identify genes, protein, PTMs, and pathways of interest to understand aberrant biological processes.

Acknowledgement

This work was supported by the National Cancer Institute, the Clinical Proteomic Tumor Analysis Consortium (CPTAC, Grant U24CA210985) and the Early Detection Research Network (EDRN, U01CA152813). The authors would like to thank Dr. David J. Clark for his assistance with reading the manuscript and helpful discussions

Conflict of interest

The authors declare no financial or commercial conflict of interest.

References

- 1 Hoadley, K. A. *et al.* Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* **173**, 291-304.e296, doi:10.1016/j.cell.2018.03.022 (2018).
- 2 Nawy, T. A pan-cancer atlas. *Nature Methods* **15**, 407-407, doi:10.1038/s41592-018-0020-4 (2018).
- 3 Mertins, P. *et al.* Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **534**, 55, doi:10.1038/nature18003 (2016).
- 4 Zhang, B. *et al.* Proteogenomic characterization of human colon and rectal cancer. *Nature* **513**, 382-387, doi:10.1038/nature13438 (2014).
- 5 Vasaikar, S. *et al.* Proteogenomic Analysis of Human Colon Cancer Reveals New Therapeutic Opportunities. (2019).
- 6 Zhang, H. *et al.* Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. *Cell* **166**, 755-765, doi:10.1016/j.cell.2016.05.069 (2016).
- 7 Weiner, A. K., Sidoli, S., Diskin, S. J. & Garcia, B. GiaPronto: a one-click graph visualization software for proteomics datasets. *Molecular & Cellular Proteomics*, mcp.TIR117.000438, doi:10.1074/mcp.TIR117.000438 (2017).
- 8 Chang, C. *et al.* PANDA-view: an easy-to-use tool for statistical analysis and visualization of quantitative proteomics data. *Bioinformatics (Oxford, England)* **34**, 3594-3596, doi:10.1093/bioinformatics/bty408 (2018).
- 9 Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**, 15545-15550, doi:10.1073/pnas.0506580102 (2005).
- 10 Mootha, V. K. *et al.* PGC-1 α -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics* **34**, 267-273, doi:10.1038/ng1180 (2003).
- 11 Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research* **44**, W90-W97, doi:10.1093/nar/gkw377 (2016).
- 12 Zhang, B., Kirov, S. & Snoddy, J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic acids research* **33**, W741-748, doi:10.1093/nar/gki475 (2005).
- 13 Wang, J., Duncan, D., Shi, Z. & Zhang, B. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic acids research* **41**, W77-83, doi:10.1093/nar/gkt439 (2013).
- 14 Wang, J., Vasaikar, S., Shi, Z., Greer, M. & Zhang, B. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic acids research* **45**, W130-w137, doi:10.1093/nar/gkx356 (2017).
- 15 Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z. & Zhang, B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic acids research* **47**, W199-w205, doi:10.1093/nar/gkz401 (2019).
- 16 Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825-2830 (2011).
- 17 Xu, T. *et al.* CancerSubtypes: an R/Bioconductor package for molecular cancer subtype identification, validation and visualization. *Bioinformatics (Oxford, England)* **33**, 3131-3133, doi:10.1093/bioinformatics/btx378 (2017).

Table 1. Main modules and functions of OmicsOne

Module	Functions
Preprocessing	Noise filtration, Log transformation, Missing value filtration, Normalization
Data profiling (DP) and quality control (QC)	Expression profiling, Phenotype profiling, Sample reproducibility
Differential expression analysis (DEA)	Statistical testing, Enrichment analysis, Phenotype associations
Dimensionality reduction and clustering analysis	Principle component analysis, Clustering, Phenotype associations

Figure 1

Figure 1. Sample figures of data profiling and quality control module of OmicsOne.

A. A histogram of grades of all patients.

B. A heatmap of correlation values of phenotypes to help identify the relationship of known phenotype.

C. A series of scatter plots with trend lines to demonstrate the reproducibility of quality control samples

D. A histogram of coefficient of variant (CV) values of protein expression in quality control samples.

Figure 2

Figure 2. Differential expression analysis module of OmicsOne.

- A.** A volcano plot of two ovarian cancer subtypes (proliferative and mesenchymal). The proteins with differential expression were selected as fold change ≥ 2 and adjusted P values (FDR) < 0.05
- B.** A bar chart of the pathways over-represented in proliferative than mesenchymal, which were enriched from proteins with expression's fold change ≥ 1.3 by GSEApY.
- C.** A boxplot to show the significant associations of E2F_TARGETS with survival.

Figure 3

Figure 3. Differential expression analysis module of OmicsOne.

- A.** A PCA plot of all 119 samples from Johns Hopkins University (JHU) marked with subtype names and 95% confidence intervals.
- B.** A clustering map of all JHU samples with vital status and clustering information analyzed by CancerSubtypes.