Web-based visual analytics for multi-omics data integration

Guangyan Zhou

Institute of Parasitology

McGill University,

Montreal, Quebec

A thesis submitted to McGill University in partial fulfillment

of the requirements of the degree of

Doctor of Philosophy

Abstract

With advances in high-throughput molecular profiling technologies, there is an increase in types and quantities of omics data available for researchers to gain insights for biomedical research. However, effective extraction of information and integration of these data remain challenging due to the complex nature of multi-omics data. This thesis aims to address the challenges of transcriptomics data analysis and multi-omics data integration by developing easy-to-use, web-based platforms to support advanced statistics and visual analytics for broad bench scientists without programming expertise.

Firstly, I developed the version 3.0 of NetworkAnalyst, a web-based platform for comprehensive analysis and interpretation of transcriptomics data. It supports both thorough data processing and comparative analysis including nested comparisons and time series. It offers a rich set of visual analytics methods encompassing network, volcano, heatmap, chord diagram, Venn diagram and scatter plot visualization coupled with molecular interaction and enrichment analysis for functional interpretation of transcriptomics data. NetworkAnalyst also allows meta-analysis of multiple gene expression tables or gene lists using a combination of advanced statistical meta-analysis methods and integrative visual analytics.

Secondly, I developed OmicsNet, a web-based visual analytics platform dedicated for network-based multi-omics integration and visual exploration. The tool distinguishes itself by enabling web-based 3D visualization of biological networks in various innovative graphical layouts. By leveraging known molecular interactions from public databases, OmicsNet is able to build multi-omics interaction network from user supplied lists of molecules encompassing genes, proteins, miRNA, transcription factors and metabolites to facilitate holistic data understanding.

Lastly, I developed OmicsAnalyst, a web-based platform that implements data-driven multi-omics integration methods coupled with advanced visual analytics. The tool supports three distinct strategies for multi-omics integration including correlation-based, clustering-based and dimension reduction-based approaches, coupled with network, heatmap and 3D scatter plot visual analytics respectively. OmicsAnalyst was able to integrate proteomics and metabolomics datasets to reveal important expression patterns and key biomarker signatures from a recent multi-omics study on human pregnancy.

Overall, this thesis shows how web-based visual analytics frameworks can be used to facilitate omics data analysis processes and expedite data exploration process for hypothesis generation and more targeted studies.

Abrégé

Avec les progrès des technologies de profilage moléculaire à haut débit, on assiste à une augmentation des types et des quantités de données omiques disponibles pour les chercheurs dans le cadre de la recherche biomédicale. Cependant, l'extraction efficace d'informations et l'intégration de ces données restent difficiles en raison de la nature complexe des données multi-omiques. Cette thèse vise à relever les défis de l'analyse des données transcriptomiques et de l'intégration des données multi-omiques en développant des plates-formes web faciles à utiliser en utilisant statistiques avancées et analyse visuelle, pour faciliter la tâche des scientifiques sans expertise en programmation.

Tout d'abord, j'ai développé la version 3.0 de NetworkAnalyst, une plateforme web pour l'analyse et l'interprétation complètes des données transcriptomiques. Elle prend en charge à la fois le traitement complet des données et l'analyse comparative, y compris les comparaisons imbriquées et les séries chronologiques. Elle offre un riche ensemble de méthodes d'analyse visuelle comprenant la visualisation de réseaux, de volcans, de cartes thermiques, de diagrammes d'accord, de diagrammes de Venn et de diagrammes de dispersion, ainsi que l'analyse des interactions moléculaires et de l'enrichissement pour l'interprétation fonctionnelle des données transcriptomiques. NetworkAnalyst permet également la méta-analyse de plusieurs tableaux d'expression génique ou de listes de gènes en utilisant une combinaison de méthodes statistiques avancées de méta-analyse et d'analyse visuelle intégrative.

Deuxièmement, j'ai développé OmicsNet, une plateforme d'analyse visuelle basée sur le web et dédiée à l'intégration et à l'exploration visuelle de données multi-omiques basées sur les réseaux. L'outil se distingue en permettant la visualisation en 3D de réseaux biologiques sur le web dans diverses dispositions graphiques innovantes. En exploitant les interactions moléculaires connues des bases de données publiques, OmicsNet est capable de construire un réseau d'interactions multi-omiques à partir de listes de molécules fournies par l'utilisateur, comprenant des gènes, des protéines, des miRNA, des facteurs de transcription et des métabolites, afin de faciliter la compréhension holistique des données.

Enfin, j'ai développé OmicsAnalyst, une plateforme web qui met en œuvre des méthodes d'intégration multi-omique axées sur les données, couplées à des analyses visuelles avancées. L'outil utilise trois stratégies distinctes d'intégration multi-omique, notamment les approches basées sur la corrélation, le regroupement et la réduction des dimensions, couplées respectivement à des analyses visuelles de réseaux, de cartes thermiques et de diagrammes de dispersion en 3D. Dans une étude de cas, OmicsAnalyst a été en mesure d'intégrer des ensembles de données protéomiques et métabolomiques pour révéler d'importants modèles d'expression et des signatures de biomarqueurs clés à partir d'une étude multi-omique récente sur la grossesse humaine.

Dans l'ensemble, cette thèse montre comment les outils d'analyse visuelle basés sur le web peuvent être utilisés pour faciliter l’analyse des données omiques et accélérer l’exploration des données pour la génération d'hypothèses et des études plus ciblées.

Acknowledgements

I would like to thank all my committee members, friends and family for their support and help throughout my entire PhD.

Firstly, I would like to express my deepest gratitude to my advisor, Dr. Jianguo Xia, for his guidance, patience and spending countless hours helping me through the whole research process. Without his visions and inputs, the research process would be much more laborious. He also spent significant time in reviewing and revising my papers. I would also like to thank my committee member Dr. Robin Beech and Dr. Reza Salavati for their invaluable insights and suggestions regarding my research.

I am thankful for the wonderful people I worked with in my lab: Achal Dhariwal, Jasmine Chong, Jessica Ewald, Le Chang, Lu Yao, Dr Othman Soufan, Dr. Orcun Hacariz, Dr Peng Liu, Dr. Peter Lee, Yannan Fan and Zhiqiang Pang.

I am thankful for my girlfriend for her support and help. Most importantly, I am very grateful to my parents for their kindness and support. Without their support, I would not be able to complete my education.

Preface and Contribution of Authors

The work described here was performed under supervision of Dr Jianguo Xia. It follows the manuscript-based format described in the Thesis Preparation. This thesis contains a total of six chapters. The first chapter introduces the thesis and describes the background, previous works, and motivation of this thesis. Chapters 2, 3 and 4 are works that have been published in the journal of Nucleic Acids Research in 2019, 2018 and 2021 respectively. Chapter 5 is a general discussion of the findings and future directions followed by an overall conclusion.

Chapter 2 is a manuscript authored by Guangyan Zhou, Othman Soufan, Jessica Ewald, Robert E W Hancock, Niladri Basu and Jianguo Xia. It was published in Nucleic Acids Research in July 2019. Conceived project: JX. Designed and developed the project: GZ JX OS. Wrote the paper: GZ, OS, JE, RH, NB, JX. All authors reviewed and approved the final manuscript.

Chapter 3 is a manuscript authored by Guangyan Zhou and Jianguo Xia. It was published in Nucleic Acids Research in July 2018. Conceived the project: JX. Designed and developed the project: GZ JX. Wrote the paper: GZ, JX. All authors reviewed and approved the final manuscript.

Chapter 4 is a manuscript authored by Guangyan Zhou, Jessica Ewald, and Jianguo Xia. It was published in Nucleic Acids Research in May 2021. Conceived the project: JX. Designed and developed the project: GZ JX. Wrote the paper: GZ, JE, JX. All authors reviewed and approved the final manuscript.

Contribution to Original Knowledge

The proposed methods aim to lower the barrier of entry to transcriptomics and multi-omics data analysis and interpretation which remain a critical bottleneck in current biomedical research where data generation capabilities significantly outpacing the development of bioinformatics tools. By developing intuitive and easy-to-use web-based platforms, we enable and empower bench scientists with advanced visual analytics so that they can focus on data understanding rather than learning how to write scripts for data analysis. More specific contributions are listed below:

1. NetworkAnalyst supports multi-list meta-analysis, a feature that is lacking in the current landscape of bioinformatics tools. It facilitates integration and comparison of differentially expressed genes identified in multiple studies.
2. NetworkAnalyst supports enrichment network visualization for both over-representation analysis (ORA) and gene set enrichment (GSEA) results. The novelty comes from meta-node feature that enables the expanding of nodes representing enriched terms to display the underlying genes involved. This feature allows a simplified overview of the overall enriched terms by default while providing more details on demand.
3. OmicsNet is a web-based tool dedicated for visualizing biological networks in 3D space. Previous 3D-based network visualization tools for biological networks are stand-alone and require local installations.
4. OmicsNet implements 3D spherical layout to facilitate visual interpretation of complex network.
5. Implementing edge bundling functionality in 3D network visual analytics to alleviate hairball effect and highlight connection patterns. Edge bundling is not a novel concept, but it is not seen in 3D-based network biology tools.

6. Web-based tool dedicated for data-driven multi-omics integration is lacking. OmicsAnalyst is an intuitive and easy-to-use platform that offers advanced visual analytics solution to multi-omics data analysis.
7. OmicsAnalyst supports advanced scatter plot visual analytics framework for joint dimension reduction results. It is complemented with flexible clustering analysis, comparative analysis, and enrichment analysis to allow users to gain insights and to test their hypothesis.
8. OmicsAnalyst offers flexible biplot visualization in its 3D scatter plot visual analytics to quickly assess feature contributions to overall sample separation.
9. OmicsAnalyst supports a dual view heatmap visual analytics to enable simultaneous visualization of two different omics data. Both heatmaps are synchronized with each other upon reordering columns facilitating the visual assessment of clustering patterns.

List of publications

1. Zhou, G., Stevenson, M. M., Geary, T. G., & Xia, J. (2016). Comprehensive transcriptome meta-analysis to characterize host immune responses in helminth infections. PLoS neglected tropical diseases, 10(4), e0004624.
2. Kang, E, Zhou, G., Yousefi, M., Cayrol, R., Xia, J., & Gruenheid, S. (2018). Loss of disease tolerance during Citrobacter rodentium infection is associated with impaired epithelial differentiation and hyperactivation of T cell responses. Scientific reports, 8(1), 1-14.
3. Zhou, G., & Xia, J. (2018). OmicsNet: a web-based tool for creation and visual analysis of biological networks in 3D space. Nucleic acids research, 46(W1), W514-W522.
4. Zhou, G., & Xia, J. (2019). Using OmicsNet for network integration and 3D visualization. Current protocols in bioinformatics, 65(1), e69.
5. Zhou, G., Soufan, O., Ewald, J., Hancock, R. E., Basu, N., & Xia, J. (2019). NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. Nucleic acids research, 47(W1), W234-W241.
6. Zhou, G., Li, S., & Xia, J. (2020). Network-Based Approaches for Multi-omics Integration. In Computational Methods and Data Analysis for Metabolomics (pp. 469-487). Humana, New York, NY.
7. Chang, L., Zhou, G., Soufan, O., & Xia, J. (2020). miRNet 2.0: network-based visual analytics for miRNA functional analysis and systems biology. Nucleic Acids Research.
8. Chong, J., Liu, P., Zhou, G., & Xia, J. (2020). Using MicrobiomeAnalyst for comprehensive statistical, functional, and meta-analysis of microbiome data. Nature Protocols, 15(3), 799-821.
9. Zhou, G., Ewald, J., & Xia, J. (2021). OmicsAnalyst: a comprehensive web-based platform for visual analytics of multi-omics data. Nucleic Acids Research.
10. Pang, Z., Zhou, G., Chong, J., & Xia, J. (2021). Comprehensive meta-analysis of COVID-19 global metabolomics datasets. Metabolites, 11(1), 44.
11. Pang, Z., Chong, J., Zhou, G., de Lima Morais, D.A., Chang, L., Barrette, M., Gauthier, C., Jacques, P.É., Li, S. and Xia, J. (2021). MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights. Nucleic acids research.

Table of Contents

List of Figures

Page

List of Tables

Page

Chapter 1: Introduction

With the rapid progress of high-throughput sequencing technologies, multi-omics datasets have become increasingly available. Understanding complex biological processes requires taking more holistic approaches that integrate multiple datasets in order to identify key biomolecules involved, understand their biological functions and interactions among them. However, analyzing and integrating these datasets pose significant challenges due to the high heterogeneity and complexity of the datasets. It is safe to say that the current bottleneck to multi-omics studies remains in the lack of tools and methods dedicated for such integrative analysis. This dissertation attempts to address the challenges associated with multi-omics analysis by proposing and developing a series of visual analytics tool designed to facilitate this task. The proposed tools can be used to explore high-dimensional omics data, identify relevant biomolecules, explore their interactions, and derive actionable insights for translational applications.

In this chapter, I first discuss the motivation of this work (Section 1.1), followed by an overview to the research areas, background information and previous works related to the thesis (Section 1.3 and 1.4). In the next section, I try to summarize the overall contributions and achievements of this thesis in the research context (Section 1.3). Finally, I describe the overall structure of the following chapters.

1.1 Motivation

Complex diseases often involve dysregulations of a plethora of biological processes. Using conventional molecular biology approach where one or few molecules are studied in isolation is insufficient to understand the whole picture (1). Systems biology approaches, where interactions between molecules are considered within the whole system, are more suitable to study these diseases. With the advent of high-throughput omics technologies and increasing computational storage and power, omics-based studies have become more and more prevalent in biomedical research over the last two decades. For instance, since 2008, the decrease in cost of genomics data sequencing has been outpacing Moore's Law by a factor of 4 (2). The proliferation of omics data has arguably led to a paradigm change from molecular biology to systems biology aiming for comprehensive understanding the molecular underpinnings of health and diseases (3). It has led to numerous landmark studies where large-scale multi-omics data has been generated to characterize biological conditions or diseases. For instance, The Cancer Genome Atlas (TCGA) program (4), led by National Cancer Institute (NCI) and National Human Genome Research Institute, catalogues over 20,000 cancer and matched normal samples spanning 33 cancer types, across genomics, epigenomic, transcriptomic and proteomic layers. Large publicly available repositories also greatly facilitate access to multi-omics datasets. OmicsDI (Omics Discovery Index) provides datasets from 11 different repositories in a standard format (5).

Although data generation technology has become more and more accessible, the subsequent analysis of high-throughput omics data is still lagging, especially data interpretation (6). Additionally, issues associated with misuse of statistical methods and over-reliance of P-values have played important roles in the ongoing reproducibility crisis in biomedical science research (7,8). To successfully convert omics data into meaningful biological knowledge, there remain multiple challenges, some of them are listed below:

(i) Heterogeneity of omics data of different types and platforms, and lack of standardized formats pose significant hurdles to multi-omics integration.

(ii) It is hard to differentiate signals from noise based on statistical methods alone.

(iii) High dimensionality of omics data results in high computational costs and poses challenges to statistical analysis. Feature prioritization is necessary to filter out noise and focus on relevant features.

(iv) Omics data analysis is a complex task that often requires both advanced bioinformatics skills and domain knowledge on the biological problems under study. Interdisciplinary knowledge of both domains is required to perform in-depth analysis.

(v) Most omics bioinformatics tools are stand-alone software or web services that is designed for specific analytical task. Additionally, users often need to have programming language knowledge and utilize multiple tools to analyze their data thoroughly.

(vi) Biological networks are often large and complex which causes undesired “hairball effect” when visualized. New graph layout and visualization techniques need to be developed to attenuate the infamous “hairball effect”.

To address the increasing need for analyses of large omics datasets and multi-omics integration, this thesis aims to develop bioinformatics applications that tackle some of the challenges mentioned above. I have taken a visual analytics approach to empower and engage researchers. The objective is to take advantage of their domain knowledge and cognitive reasoning to further improve the data analysis and interpretation process. Specifically, thesis presents three overall aims:

Aim 1 Develop web-based platform to address functional profiling, network integration and meta-analysis of transcriptomics data.

Aim 2 Develop web-based platform for multi-omics network integration and 3D visualization.

Aim 3 Develop web-based platform to enable data-driven integration on multi-omics datasets.

To address **Aim 1**, I have worked on improving NetworkAnalyst, an existing web-based platform for transcriptomics data analysis. For **Aim 2**, I presented OmicsNet, a web-based visual analytics tool to build multi-omics interaction network from lists of features. Lastly, **Aim 3** was addressed by developing OmicsAnalyst, a web-based visual analytics platform allowing users to perform multi-omics integration through correlation analysis, integrative clustering and dimension reduction. Our proposed approaches are supported by case studies to illustrate their usage.

1.2 Background and scope of research

1.2.1 Transcriptomics

Transcriptomics is defined as the study of transcriptome – the whole set of RNA transcripts produced by the genome, using high-throughput methods such as microarray or RNA sequencing. The measured transcripts can come from a large population of input cells or a single cell. One of the main approaches of transcriptomics is to perform gene expression profiling. This is performed on the expression matrix obtained from preprocessing raw sequencing data. This type of experiment aims to simultaneously measure expression levels of thousands of genes to study the change in gene expression caused by certain conditions (treatment, disease, infection, etc.). It is also referred as comparative analysis because the gene expression level of experimental group is compared with the gene expression of control group.

The current form of microarray dates from late 1990s and early 2000s, although related array technologies have appeared as early as in the mid 1970s (9). It is currently still widely used to measure relative abundance of nucleic acid sequence, but it is slowly getting replaced by RNA-seq (RNA-sequencing) starting from the last decade. The strategy is to use hybridization followed by quantification of hybridization events by using fluorescent detection. It involves in adding nucleic acid mixture to wells containing probes consisting of thousands of nucleic acid sequences attached to a solid surface.

As for RNA-seq, the technique is to convert RNA population into a complementary RNA (cDNA) library through reverse transcription. This is done by fragmenting the cDNA, adding adapters at the end of each fragment, and amplifying these fragments. The next step is to analyze cDNA library by Next-Generation Sequencing (NGS) to generate sequences corresponding to these fragments. The resulting reads are then to be aligned either to a reference genome if available or to be assembled *de novo* to produce a sequence map spanning the whole transcriptome.

NGS is the key technology propelling the rapid decrease in cost of DNA sequencing. There exists multiple commercially available NGS platforms using different sequencing technologies, of which detailed description of these technologies is beyond the scope of this thesis. Generally, the sequencing procedure is conducted in the following manner. First step consists of generating the DNA sequencing libraries using clonal amplification. Second, sequencing of DNA is done using synthesis approach where nucleotides are added to complementary strands. Lastly, millions of amplified DNA fragments are sequenced in a massively parallel fashion. NGS can be used both to sequence entire genomes but also specific parts of the genomes. Additionally, NGS can sequence each base multiple times (referred as sequencing depth), providing insights into DNA variation and more accurate sequencing data.

There are several reasons why RNA-seq is considered superior to microarray technology. Foremost, the possibility of using *de novo* assembly approach means that RNA-seq is not limited by the current knowledge of genomic sequences. It can detect novel transcripts, small nucleotide polymorphisms and other alterations. Second, mapping cDNA sequences to targeted genomic regions can significantly reduce experimental noise. Additionally, hybridization issues commonly affecting microarrays do not apply to RNA-seq. Finally, RNA-seq is a more quantifiable technology where counts are obtained in contrast to microarray where expression values are relative to other signals detected on the array.

In transcriptomics data analysis, there are typically three levels of analysis: (i) raw data preprocessing; (ii) data processing and statistical testing (iii) data interpretation by visual analytics and functional profiling. The first two levels are currently well-established. In contrast, data interpretation requires interactive implementation and is relatively lacking in the current bioinformatics landscape. Raw data preprocessing steps corresponds to processing raw data file (i.e. FASTQ) into feature count matrix which can be used for subsequent statistical analysis. Data processing steps try to remove low quality data and technical variations while preserving biological signals. In statistical testing, the overall objective is to compare different experimental conditions to identify a set of features that are differentially abundant. Most straightforward approach is to use single pairwise comparison (i.e. control against diseased groups) in contrast to more complex multiple contrasts or time series. Lastly, visual analytics such as network and heatmap are used to facilitate functional interpretation.

1.2.2 Multi-omics data

Omics data are generated from different levels of biological systems and are used to assess different aspects of biological processes. Multi-omics integration attempts to understand the interaction between different omics layers and have a more holistic understanding of their functions.

Table 1.1 List of selected multi-omics data repositories

Data	Description	Type	Links
NCI60 (US National Cancer Institute 60 human tumour cell lines anticancer drug screen)	Drug screening on 60 human tumor cell lines.	Transcriptome, proteome	https://dtp.cancer.gov/discovery_development/nci-60/
TCGA (The Cancer Genome Atlas)	Cancer genomics project containing ~20000 samples from 33 cancer types and normal samples	Genome, transcriptome, proteome	https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga
TopMed (Trans-Omics for Precision Medicine)	Large project to study heart, lung, blood, and sleep disorders using omics data from ~ 155k participants across more than 80 studies	Genome, transcriptome, proteome, metabolome	https://www.nhlbi.nih.gov/science/trans-omics-precision-medicine-topmed-program
jMorp (Japanese Multi Omics Reference Panel)	Large multi-omics database generated using samples from 5000 healthy Japanese volunteers	Genome, transcriptome, proteome, metabolome	https://jmorp.megabank.tohoku.ac.jp
iHMP (Integrative Human Microbiome Project)	Multi-omics data characterizing microbiome-host profiles in health and disease (pregnancy and preterm birth, inflammatory bowel disease, type II diabetes)	Genome, transcriptome, metabolome, proteome, microbiome	https://www.hmpdacc.org/ihm/
OmicsDI (Omics Discovery Index)	A meta-database for multi-omics data.	Genome, transcriptome, proteome, metabolome	https://www.omicsdi.org/
MuTHER (Multiple Tissue Human Expression Resource)	Multi-omics data from a range of tissues collected from a set of ~850 UK twins	Genome, transcriptome	http://www.muther.ac.uk/
METSIM (Metabolic Syndrome In Man)	Population-based study to investigate nongenetic and genetic factors associated with the risk of T2D and CVD	Genome, transcriptome, metabolomic	http://www.nationalbiobanks.fi/index.php/studies2/10-metsim

ICGC (International Cancer Genome Consortium)	Multi-omics data from mutational abnormalities in in 21 primary cancer sites from ~20 000 donors	Genome, genomic variation (somatic and germline mutation)	https://dcc.icgc.org/
TARGET (Therapeutically Applicable Research To Generate Effective Treatments)	Multi-omics resources of childhood cancers collected from ~1700 paediatric leukaemia and solid tumors.	Genome, transcriptome, metabolomic	https://ocg.cancer.gov/programs/target

With the advent of high-throughput data generation, there is an accumulation of large volumes of multi-omics data from different types. Large-scale projects such as The Cancer Genome Atlas (TCGA) (6) and Human Microbiome Project (HMP) (10) have taken systematic approach where large number of individuals are profiled across multiple omics layers, resulting in numerous publicly available multi-omics datasets. The availability of multi-omics data resources is a first stage towards more comprehensive characterization of healthy individuals and disease conditions at molecular level. Please refer to Table 1.1 for a list of multi-omics data repositories available to the public. Next stage is to develop the tools necessary for in-depth analysis in order to move towards personalized medicine (11).

1.2.3 Integrative analysis

Integrative analysis refers to the use of multiple sources of data to gain additional insights on the studied system. This integration can be performed both within a single omics type and across omics (Figure 1.2). The former, also known as meta-analysis, attempts to integrate samples from different studies from a single omics sharing same research question to increase statistical power and eliminate bias from individual studies (12). The latter aims to integrate different levels of omics measures of same samples to better understand the biological system. The shortfalls of single omics analysis lay in its inability to explain the etiology of complex diseases and to understand the causative mechanisms behind the biological processes (13). Most processes involve more than one omics layers and interrelations between them play an important role (14). Although multi-omics integration is promising, the heterogeneity between the different data types and their data complexity makes the integration process challenging.

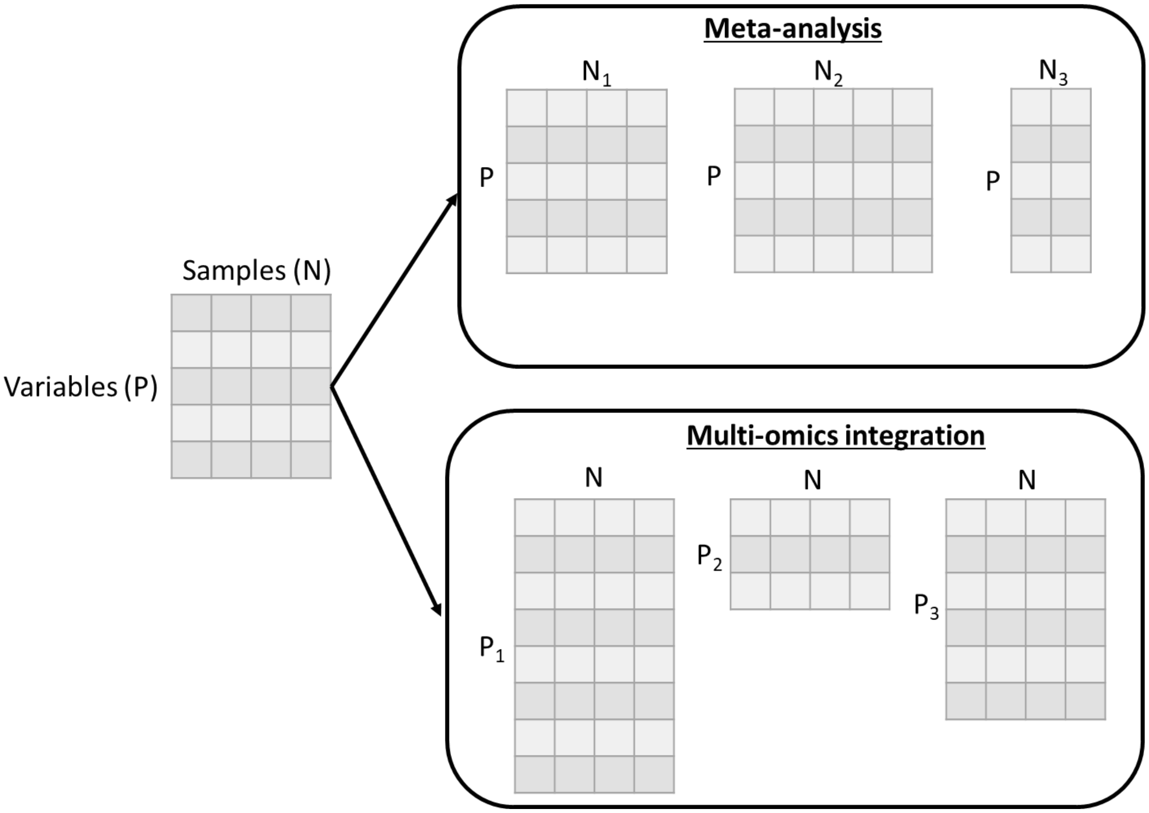


Figure 1.1 A schematic view of two types of omics integration: meta-analysis and multi-omics.

Meta-analysis is a term that refers to the combination results from multiple previous studies to confirm reliability and generalizability of individual studies. In omics studies, such as transcriptomics, the combination is more specifically applied to expression matrices from individual studies. This approach is to address some of the issues associated with individual studies such as lack of reproducibility (15), lack of robustness to data perturbations (16) and small sample size. Using meta-analysis, researchers can obtain more precise estimate of expression/abundance values with additional sample size and identify core molecular signature shared across studies. In a study I performed in 2016, I was able to identify a core host gene expression signature associated with immune response against helminth infection by integrating datasets from nine different transcriptomics studies (17). Overall, the process increases both reliability and generalizability of the results. Additionally, meta-analysis is relatively inexpensive since they take as input already available datasets.

Multi-omics approach has become increasingly used to fill gaps in understanding health and disease conditions (13). Studies range from understanding host-pathogen interactions (18) to complex and chronic diseases (19). In recent years, multi-omics approach has enabled the rise of personalized medicine (20). Overall, there are multi-omics integration can be used to achieve four different objectives:

- Mechanistic insights: Identify novel molecular interactions between different omics data types to improve our current understanding of certain biological processes or disease conditions.
- Patterns and trends: Explore overall global structure of samples or patients and compare similarities of different clusters. Leverage existing knowledge to understand the molecular context. This can be useful to improve sample classification and disease subtyping.
- Systems biology: Build models such as genome-scale reconstruction of metabolic networks where flux-balance analysis can be performed. This leads to deeper systems-level understanding and novel knowledge discovery.
- Predictive modeling: use multi-omics datasets to classify new samples or patients into different “healthy” and “disease” categories.

In my research, I focus on pattern and trends discovery within the datasets using advanced visual analytics, leading to data-driven hypothesis generation, and discovering novel insights. I also provide methods to identify potential molecular interactions using correlation analysis, but it remains peripheral. Deriving mechanistic insights leans more towards molecular biology approach requiring different experimental design. The other two objectives are beyond the scope of this thesis.

Multi-omics integration can be further categorized into three main approaches: conceptual integration, knowledge-driven and data-driven integration (21). Conceptual integration is a natural extension from single omics analysis. In this approach, the analysis of individual omics datasets is performed separately, and the resulting findings are compared without analyzing the datasets together. This can be achieved incrementally where results from a single omics data analysis is complemented with results from additional omics layers. The idea is to use the additional omics layers to refine the molecular context and have a better grasp of the pathways involved. The core omics is often from an omics technology that is more mature and complete compared to the others (i.e transcriptomics). Alternatively, equal weights can be attributed to different omics layers.

In knowledge-driven network integration, the objective is to better understand the molecular context of the omics datasets. It can be achieved by either mapping them into existing molecular interaction data (i.e protein-protein interaction, metabolic pathways, regulatory networks, etc.) or pathway knowledge to form context relevant subnetworks or attempt to infer interactions by assessing pairwise relationships using measures such as correlation or mutual information. The mapping process can be summarized into three steps:

1. Perform data processing and comparative statistical analysis on individual omics data to identify significant features.

2. Map the features to the current knowledge universe defined by the available molecular interaction databases.
3. Visualize and analyze the resulting multi-omics subnetwork.

The resulted subnetwork serves as a framework for downstream functional and topological analysis. It allows researchers to explore the connection patterns of their molecules of interest and visualize them in the context of current network knowledge.

Data-driven integration relies on statistical methods and algorithms to extract information from the datasets. Unlike the other two approaches, this has the advantage of being more conductive to discover new knowledge and being relatively free of bias associated with prior knowledge and researcher domain knowledge. In the recent years, machine learning algorithms, such as deep neural networks and random forest have found use in building highly predictive models using omics datasets. Additionally, aside from being used for predictive modeling such as classifying samples and patients, algorithms adapted to uncover underlying mechanisms of biological processes have become increasingly available (22). On the other side, one complaint that is often voiced by the research community is the “black-box” phenomenon, referring to the lack of transparency and understanding about the inner mechanisms of the models (23). There is an ongoing effort to alleviate this problem, but it remains an active research topic in machine learning communities (24).

In this thesis, I focus on integrative methods that are more intuitive to life scientists and scalable for large data. Emphasis is put on exploratory data analysis of multi-omics integration, aiming to reveal inherent patterns and trends from the datasets using three distinct approaches: correlation, clustering and joint dimension reduction. Correlation analysis focuses on assessing pairwise associations between biological features using different measures such as correlation, mutual information and co-expression. Clustering analysis leverages the recent progress in multi-view clustering to identify groups of similar samples using multi-omics datasets. Lastly, joint dimension reduction analysis uses advanced multivariate statistics to summarize the datasets into lower-dimensional space while minimizing information loss and integrate them. This is both useful in exploring inherent data structure and clustering of samples and identify features of interest contributing to sample variation.

1.2.4 Network analysis

To understand biological processes, it is essential to not only study the biological entities themselves but also how they interact with each other. Using network framework is a natural and intuitive way to display data-derived interconnections between features (i.e correlation and mutual information) or model biological data such as ecological, metabolic, molecular interaction and gene regulatory networks. Network analysis is not only useful for visual explorations, but it also makes use of graph theories to analyze and to derive novel biological insights from the underlying topological structures. Networks can be in the form of undirected or directed network. Protein-protein interaction or co-expression networks are undirected as there is no specific direction associated with each edge. On the other hand, regulatory network and metabolic network are directed.

Progresses in graph theory have also unraveled insights on topological properties of biological networks more generally. For instance, studies have found that biological networks are not random but tend to form scale-free network. Scale-free network is a type of network which has a skewed degree distribution which follows the power law. One of the most noticeable consequences is that the network contains small number of hub nodes that are responsible for connecting most of the network together (25). In protein-protein interaction networks, genes encoding hub proteins tend to be essential genes. Genes are considered essential if their encoded functions are essential to early development.

In studying human diseases, network-based approach offers a useful framework to identify potential drug targets by not only gain better understanding the intrinsic molecular interplays involved in disease but also identify disease-related pathways and genes. Concretely, network-based drug-repurposing has identified four potential drugs for treating SARS-CoV-2 in a recent study (26). The authors used a plethora of algorithms based on artificial intelligence, network diffusion and network proximity to rank over 6000 drugs for their effectiveness against SARS-CoV-2 virus. Experimental validation was performed on the top ranked candidates, and they were able to observe reduction in viral infection. More interestingly, among the 77 drugs identified to reduce viral load, 76 of them do not bind to the proteins directly targeted by SARS-CoV-2 virus. It would not be possible to identify them using the conventional docking-based strategies.

Leveraging acquired knowledge and concepts from network biology and network theory, researchers can better understand the molecular context of their omics data and gain novel biological insights. For instance, a key assumption in network analysis is guilt-by-association where direct interacting partners tend to be involved in the same biochemical process or similar roles (27). Network-based approach in omics data analysis can be further divided into two main methods relating to network building: (i) knowledge-based network by mapping omics data onto known interactions, usually obtained from publicly available databases or previous studies. (ii) data-driven networks formed using statistical methods such as correlation or co-expression analysis (28);

Knowledge-based network relies on the existing interaction databases to provide molecular context to the multi-omics data. The main objective is to map the experimental data onto known interaction network to study the relationships between them and other biological entities. Among biological networks, most used are protein-protein interaction and metabolic networks. In omics data analysis, the main approach is to identify subnetworks enriched for biological entities of interest, which corresponds to context-specific functional modules characterizing the biological system under study. The resulting subnetwork provides a framework for further downstream analysis including module detection, functional enrichment analysis, identify relevant molecules undetected from omics data analysis using the principle of guilt-by-association.

Data-driven networks are often computed based on similarities between each entity (expression level, concentration, count, etc.). A straightforward and somewhat naïve approach consists of computing pairwise correlation between entities and assign weighted edges between entities with a correlation score above a certain predefined threshold. Network edges would represent corresponding correlation score between the two connecting nodes. The resulting network can then be used to infer novel potential interactions, identify candidate biomarkers and therapeutic targets.

1.2.5 Visual analytics

I choose to use visual analytics approach to tackle the complexity and size of multi-omics data. Visual analytics is defined as “the science of analytical reasoning facilitated by interactive visual interfaces” (29). It is an emerging field that comes from information visualization and scientific visualization that integrates data analysis with interactive visualization with the goal of facilitating analytical reasoning. Visual analytics approach is well suited for omics data analysis because of the following points:

1. Domain knowledge, intuition and creativity of users can be incorporated in the analytical process using interactive visualization.
2. Data visualization takes advantage of human cognitive reasoning and perception to facilitate insights generations within large datasets.
3. Integration of statistical analysis algorithms and computational models make use of well-established methods to complement visual assessment.
4. Interactivity enables iterative data analysis processes and facilitate incremental data understanding in which initial hypothesis generation and insight discovery can be used as a starting point for further in-depth studies.

Overall, visual analytics can be summarized into four main components: data, statistical modeling, interactive visualization, and knowledge discovery from users (Figure 1.2). It allows users to enter in a sense-making loop where visualization and statistical parameters can be interactively tuned to gain better understanding and help knowledge discovery (30). This thesis presents three applications to explore, integrate and analyze omics data, where visual analytics components play a central role in their design and implementation. Our work highlights the usefulness of visual analytics approach by presenting practical use cases in gene expression profiling and multi-omics integration.

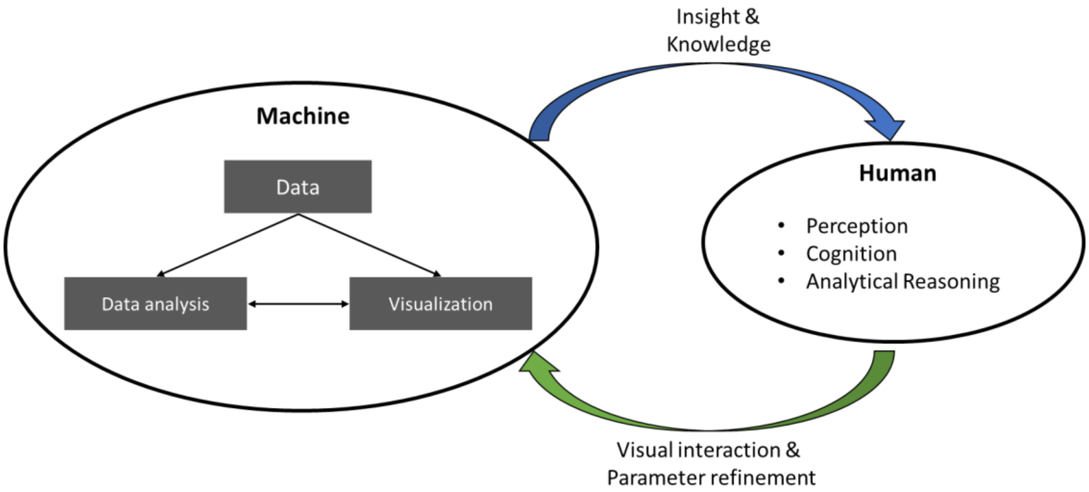


Figure 1.2 Schematic view of the visual analytics process. It is characterized as an interaction loop composed of data, visualization, statistical models of the data and knowledge discovery from users.

1.2.6 Web-based interactive visualization

The main reason why visual analytics is so important lies in its ability to include the users in the sense-making loop. Data visualization itself can reveal previously hidden data characteristics but combined with interactivity, it can allow users to dissect the data more thoroughly and to perform targeted analysis on the fly. Adequate visual analytics addressing current data challenges in life science research is lacking in the current landscape of bioinformatics tools available for omics data analysis (31).

In the past decade, there has been a significant improvement in browser technologies, giving rise to impressive web-based interactive graphics. JavaScript libraries like sigma.js (www.sigmajs.org) and D3.js (d3js.org) are supported by all major browsers and can render complex graphics and networks with thousands of nodes and edges while providing an interactive experience to the users. More recently, WebGL technology has enabled the modern browsers to leverage processing from dedicated graphical processor units (GPU), effectively enabling advanced web-based 3D interactive visualization. However, low-level nature of WebGL programming requires advanced knowledge of mathematics behind 3D graphics and presents a steep learning curve for developers (32). In this case, low-level means that WebGL codes are sent directly to the hardware (i.e. GPU) as instructions to render the graphics. In other words, it focuses on rendering speed and efficiency rather than ease of use. This has led to the development of graphical libraries such as Three.js (threejs.org) and Babylon.js (www.babylonjs.com) which have significantly lowered the barrier of entry to developing WebGL-based applications. In summary, it can be argued that web-based interactive visualization has become much more adapted for developing visual analytics tools for sense-making purpose during interactive data analysis process coupled with R-based visualization for publication quality static images.

1.2.7 Democratizing omics data analysis

Although there exist several bioinformatics tools dedicated for omics data analysis, they are often inaccessible to average life scientists due to their highly technical nature and requiring basic programming knowledge. For instance, most of the current bioinformatics tools are in the form of R packages from Bioconductor (33), requiring users to have a minimum of knowledge of R language. Additionally, most tools are designed to accomplish specific data analysis tasks and researchers often need to use a combination of different tools to achieve a thorough analysis of their data. This makes omics data analysis quite tedious due to the different data formatting standards of each tool.

To address this issue, development of intuitive and easy-to-use bioinformatics platforms that are easily accessible is needed. In this case, a bioinformatics platform refers to a collection of simpler tools dedicated for specific tasks organized in a well-designed system. The objective is for the platform to be more useful than the sum of individual tools. For instance, in the field of software development, programmers have been using Integrated Developer Environment (IDE) for developing their applications. IDE is composed of a series of basic tools such as source code editor, file manager, syntax highlighting, version control system and much more. As a system, IDE can increase the productivity of programmers significantly.

In bioinformatics, MetaboAnalyst is a good example of bioinformatics platform, as it provides an end-to-end solution of metabolomics data processing, analysis and visual analytics (34). Specifically, it supports spectral processing, exploratory statistical analysis, functional analysis (network, pathway, enrichment) and functional meta-analysis. Each of these modules are designed for specific tasks but, as a whole, it becomes a one-stop-shop for metabolomics data analysis and interpretation. Instead of using numerous different bioinformatics tools or web servers, user can simply input their metabolomics data and use the methods described above to interpret their data. This not only saves researchers' time but also presents multiple facets of the metabolomics data which may help in generating biological insights. It also closes the gap between bioinformatics and average life scientists by providing an easy-to-use and intuitive user interface that removes the need for scripting knowledge.

1.3 Previous works

The emergence of computational biology and bioinformatics has led to a paradigm shift in biology from reductionist approach to holistic approach where multiple levels of omics data are collected and analyzed. This section serves to review a selection of visual analytics tools and methods used for omics data analysis and integration. A summary of selected recent approaches in transcriptomics and multi-omics are going to be presented.

Visual analytics for omics data

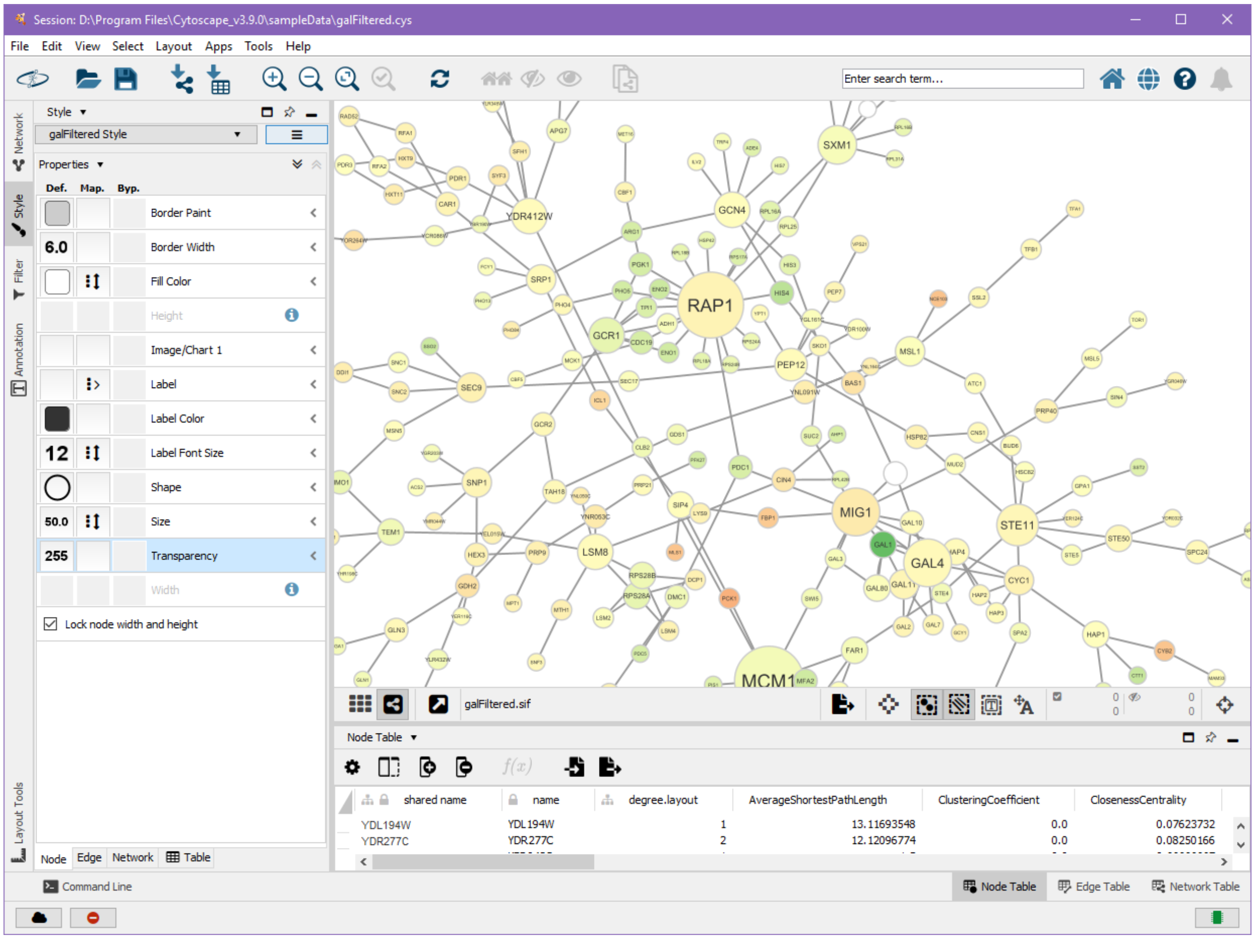
Visual analytics has been an important tool to decipher complex data from different fields and omics data is not an exception. The benefits of using visual analytics in biomedical science has been further demonstrated in recent studies. For instance, network visual analytics using NetworkAnalyst has been useful in elucidating functional pathways involved in disease progression of patients infected with Sars-CoV-2 (35). In this study, the authors were able to identify key genes and pathways related to immune responses in COVID-19 using protein-protein interaction network analysis. For instance, they identified genes related to T-cell-receptor signaling pathway to be involved and they were able to perform CyTOF mass cytometry to detect a reduction of T cells proportions. In the following chapter, I am going to review three visualization techniques that are relevant to omics data analysis. Each approach has their own strengths and is designed to tackle different analytical tasks.

1.3.1 Networks

Visualization of biological network focuses on highlighting interconnections between biological entities. It is a key method bridging statistical analysis and biological understanding. Fundamentally, most network visualizations use vertices and edges approach where vertices represent biological entities and edges represent relationship between two biological entities (protein interaction, regulatory relationship, physical binding, correlation, etc.).

Numerous tools have been developed for visualization of biological networks. Cytoscape is arguably the most widely used bioinformatics tools in visualization and analysis of biological data (Figure 1.3) (36). It is a Java stand-alone software with an intuitive interface that is accessible to users without programming knowledge and supports a wide variety of community-developed plug-ins to address specific tasks. It also has an extensive customization support for nodes, edges and network layouts. The 3Omics web application supports correlation network analysis of human transcriptomics, metabolomics, and proteomics data (37). The correlation network is computed from a pairwise similarity matrix based on abundance data (Figure 1.4). It also enables other analysis approaches including heatmap co-expression analysis, phenotype analysis and functional enrichment analysis. InnateDB, an interaction database focused on curated protein-protein interactions involved in the innate immune system, also supports network analysis (38). It allows users to build context-specific networks from a list of differentially expressed genes by mapping them into the interaction database. For interactive network visualization, InnateDB delegates the task to third-party applications such as Cytoscape, NetworkAnalyst and BioLayout Express 3D.

With the growing complexity and size of networks, graph layout has become important to provide meaningful and intuitive network visualization. In larger networks, the conventional force-directed layout becomes less effective and results in the so-called “hairball effect”, where the visualized network resembles a hairball due to the excessive number of nodes and edges. Development of novel network layouts remains critical to improve interpretability of network visualization. Cerebral, a plug-in developed for Cytoscape, incorporates cellular location information to separate nodes into different layers, presenting a biologically intuitive layout (39). The concept of multi-layered layout has also been applied in a 3D space (40). This layout organizes nodes from different data types into different layers, effectively dividing a larger network into several smaller ones, reducing visual complexity. Another example is the hive plot, a novel network layout which emphasizes on connection patterns and reproducible network visualizations (41). The nodes are positioned on radially distributed linear axes based on their topological property and connected with curved edges. In a recent paper, this layout was used to visualize connectivity of specific lung cell types and highlight the differences across different cell types (42).



Session: D:\Program Files\Cytoscape_v3.9.0\sampleData\galFiltered.cys

galFiltered Style

Def.	Map.	Byp.	
			Border Paint
6.0			Border Width
□			Fill Color
			Height
			Image/Chart 1
			Label
■			Label Color
12			Label Font Size
○			Shape
50.0			Size
255			Transparency
			Width

☑ Lock node width and height

galFiltered.sif

shared name	name	degree.layout	AverageShortestPathLength	ClusteringCoefficient	ClosenessCentrality
YDL194W	YDL194W	1	13.11693548	0.0	0.07623732
YDR277C	YDR277C	2	12.12096774	0.0	0.08250166

Figure 1.3 A screenshot of Cytoscape software's user interface displaying protein–protein and protein–DNA interactions related to the galactose-utilization pathway in yeast. On the left panel there is ma

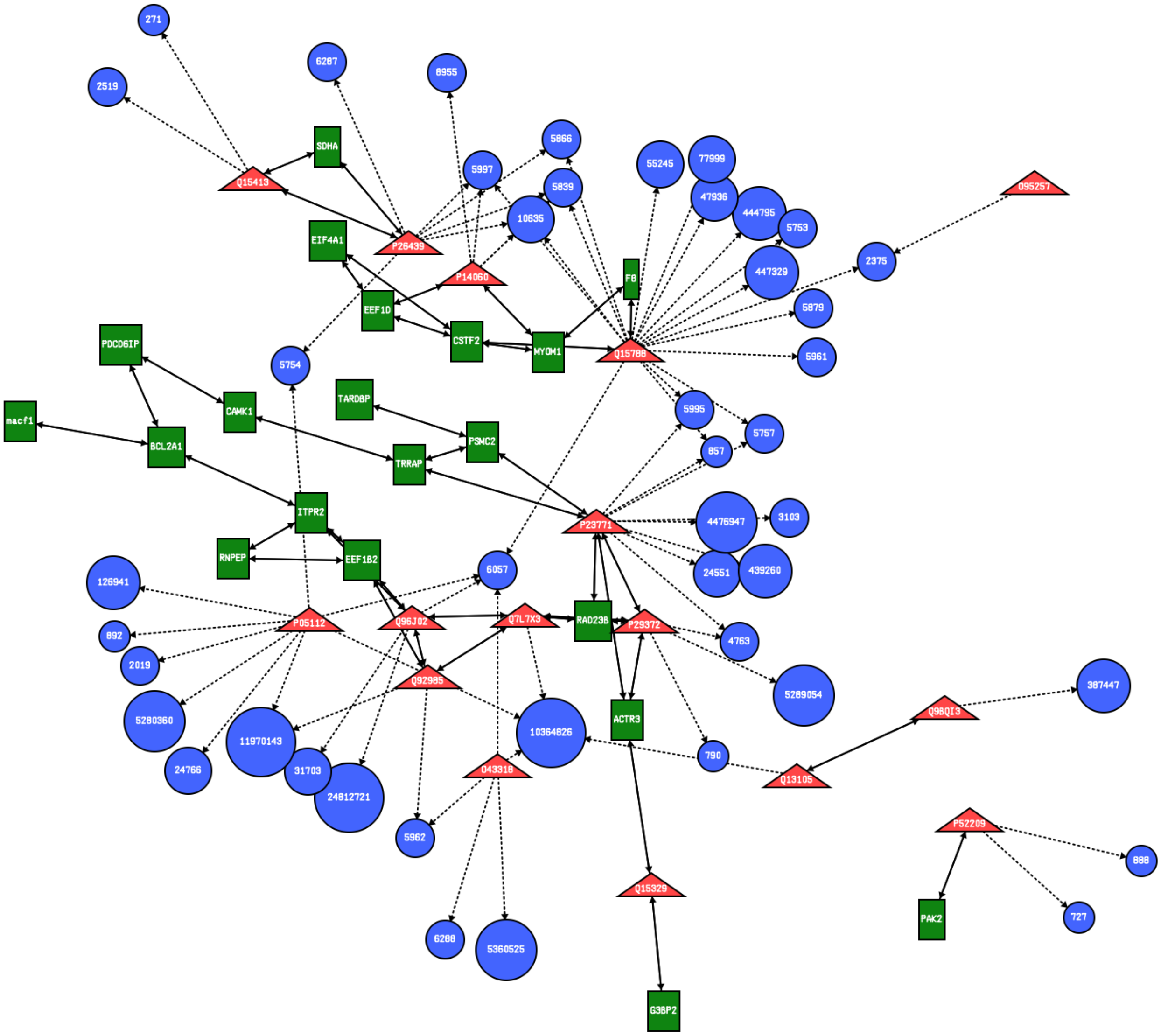


Figure 1.4 An example of correlation network of transcriptomics and proteomics data generated by 3Omics.

1.3.2 Scatter plot

A scatter plot is a type of plot that uses Cartesian coordinates to explore dependencies of two or three sets of variables from a series of data values. As omics datasets are high dimensional, dimensionality reduction methods such as principal component analysis (PCA) must be first applied to the data before scatter plot visualization. These methods are used derive a two- or three-dimensional representation of the overall data profile while preserving as much variability as possible. Their overall objective is to summarize the dataset into fewer components that explain most of the variance, facilitating visualization of general patterns in the data (43). Some of the most widely used methods for this purpose include PCA (44) and multidimensional scaling or MDS (45). This approach is widely used in the exploratory data analysis of single omics data sets. Exploratory data analysis is an important component in the early stage of omics data analysis where the researchers can assess the main characteristics of the dataset and identify outliers or potential issues regarding batch effects and overall data quality.

PCA is arguably one of the most widely used dimension reduction methods in omics data analysis (46). A typical visualization of PCA result and many other dimension reduction methods consists of score and loading plots (Figure 1.5). The score plot projects sample data points into lower dimensions while the loading plot visualizes the overall feature contributions to the separation of sample points. These two plots can be further combined into a biplot where vectors representing feature contributions are drawn on top of the score plot. As a whole, they can be used to quickly identify features driving the overall sample separation.

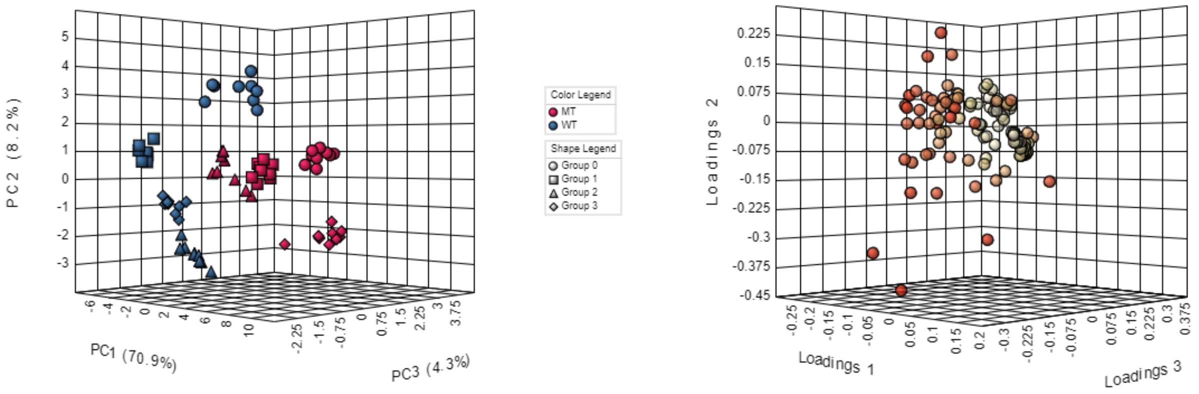


Figure 1.5 Example of a PCA score and a loading plot generated from the MetaboAnalyst platform. In the score plot, color and shape are used to represent different metadata factors. In the loading plot, color hue intensity is used to represent relative contribution of feature to sample separation.

Single-cell RNA (scRNA) data widely relies on dimension reduction techniques for both visualization and analytical purposes due to its high-dimensional and complex nature. It is especially useful in visually identifying cell clusters. Aside from classical PCA, nonlinear dimension reduction methods such as t-SNE and UMAP are also commonly used techniques in single-cell data (47,48). Nonlinear methods, in contrast to linear methods such as PCA, have the advantage of distinguishing distinct clusters that are overlapping with each other.

In the past decade, with the increasing availability of multi-omics datasets, there has been much effort in extending dimension reduction approaches to integrate multiple datasets (43). These methods integrate two or more data matrices of same samples measuring different features. Some of these methods are generalized SVD (49), Co-Inertia Analysis (50), Procrustes Analysis (51) and Canonical Correlation Analysis (CCA) (52). Notably, several research groups have developed variations of CCA for omics data integration (53,54).

Although various dimension reduction methods are available, there is a lack of web-based tools supporting interactive visualization and analysis. Many of the current bioinformatics tools mainly support basic score plot visualization for data overview purposes such as viewing sample separation, identifying outliers and assessing data quality. Babelomics, a web-based platform for transcriptomics functional profiling, offers PCA viewer among others (55). Omics Playground, another web-based platform for transcriptomics data analysis, provides graphical representation of the data using PCA and t-SNE (56). Similarly, in MiBiOmics, an interactive web application for multi-omics data integration and integration, PCA and Principal Coordinates Analysis (PCoA) are proposed to visualize each individual omics. In addition, MiBiOmics also proposes visualization using integrative methods MCIA and Procrustes Analysis. On the other hand, in scRNA data analysis, scatter plot visualization plays a more important role as it is important for visualizing both cell clustering and expression pattern. In this case, each data point represents individual cells instead of samples, resulting in a much denser scatter plot. In ASAP, a web-based pipeline for comprehensive analysis of scRNA data analysis, interactive visualization of dimension reduction results from PCA, t-SNE, Multidimensional Scaling (MDS) and Zero-Inflated Factor Analysis (ZIFA) is supported, with the ability to manually select cells for further downstream analysis (57).

1.3.3 Heatmap

Heatmap is widely used to represent omics data (Figure 1.6). It is well suited for visualizing data in the form of matrix such as a gene expression table. In omics field, the columns of heatmap usually represent samples (patient, tissue) while the rows correspond to features (gene, transcripts, metabolites, etc.). Each colored cell represents expression or abundance value which can vary by hues or intensity. Clustering is a key component of heatmap visualization for pattern discovery. This process consists of reordering the columns or rows so that similar samples or features are positioned next to each other. The samples can also be ordered according to metadata groups, facilitating group comparisons (58,59). Overall, heatmap is useful in visualizing the overall expression pattern and how the samples or feature clusters over space.

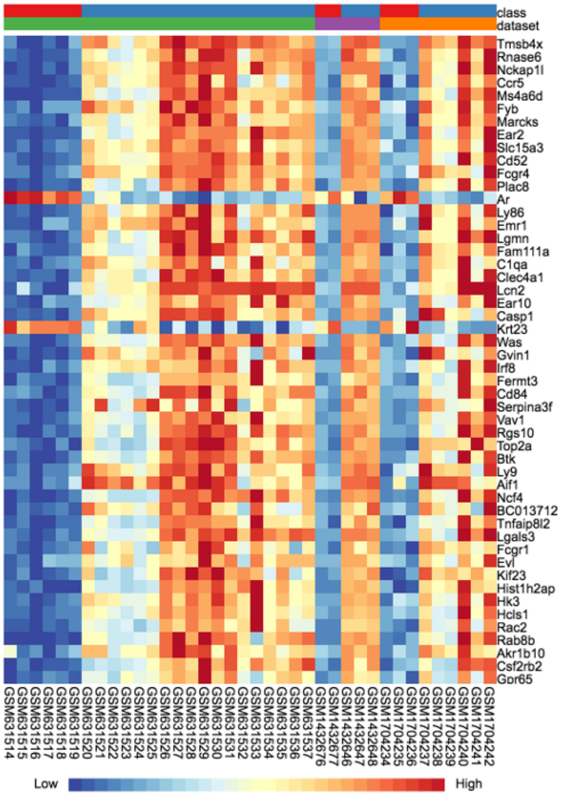


Figure 1.6 An example of heatmap visualization of gene expression meta-analysis from NetworkAnalyst**.** The columns represent samples and rows represent genes. Annotation bar located on the top indicate the experimental group and dataset origin from which the samples are from.

1.3.4 Single omics data analysis

High-throughput technologies generate huge amounts of omics data that remain uninterpretable without an analytical framework dedicated for functional profiling. The data analysis procedures differ according to the omics type of the data but there exists a general approach to analyze abundance or count-based data. It is usually the data preprocessing step that remains unique to different data generation platforms. After the feature abundance/expression matrix is obtained, the remaining steps remain relatively similar across omics types. In the following section, transcriptomics data analysis will be reviewed as an example of single omics data analysis.

As mentioned in the introductory chapter, omics data analysis can be stratified into three levels: (i) raw data preprocessing; (ii) data processing and statistical testing; (iii) visual analytics and functional interpretation. Since raw data preprocessing is more straightforward and omics-specific, most of the publicly available bioinformatics tools focus on the last two levels of analysis. Usually, most comprehensive tools take as input a count or expression matrix generated from preprocessing steps and they offer various filtering and normalization methods to prepare the data for statistical testing. Other tools focus on interpreting list of differentially expressed genes. In statistical testing, identification of DEGs remain one of the most important steps in single omics analysis. Finally, data interpretation consists of approaches such as functional enrichment analysis, pathway analysis and network analysis to help understand which biological processes are involved. Note that these tools require a built-in knowledge base to support the approaches mentioned above. Please refer to Figure 1.7 for an example of transcriptomics data analysis workflow.

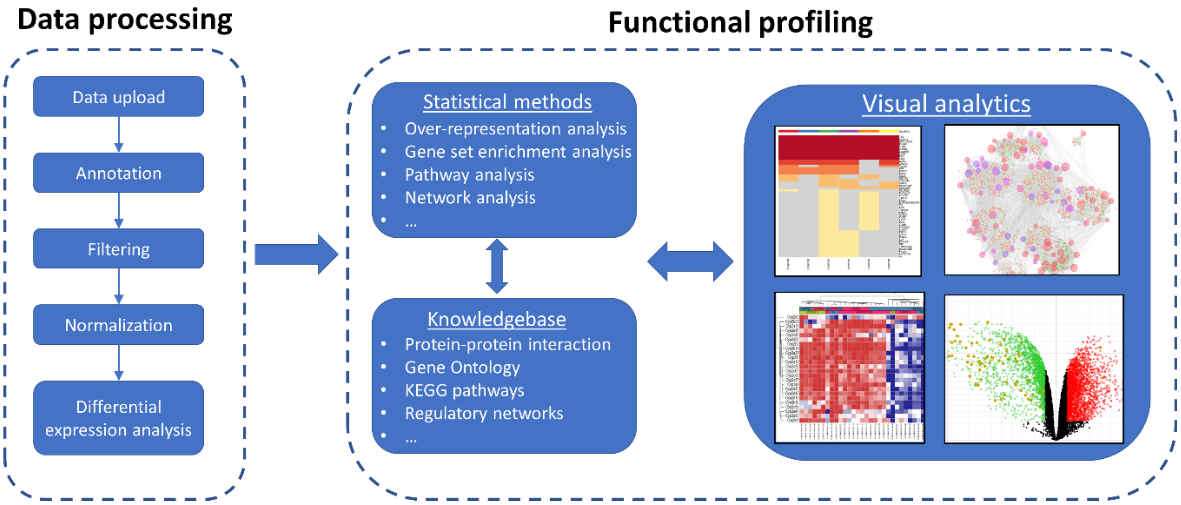


Figure 1.7 An example workflow of transcriptomics data analysis starting from preprocessed count or gene expression data matrix.

Babelomics, first published in 2005, is a web-based platform that provides a comprehensive and easy-to-use solution to genomics, transcriptomics and proteomics data analysis. It supports primary data processing such as filtering and normalization, differential expression analysis accommodating different experimental designs, and advanced functional profiling such as network analysis, enrichment analysis and *de novo* functional annotation (60). Similarly, the earlier versions of NetworkAnalyst, developed in 2014, was a comprehensive platform for transcriptomics data analysis which primarily focuses on protein-protein network analysis but also supports primary data processing, flexible differential expression analysis, heatmap visualization and meta-analysis (61,62). Omics Playground, a web-based platform developed in 2019, supports comprehensive functional profiling support for proteomics, bulk and single-cell transcriptomics (56). It provides a rich set of visual analytics such as volcano plot, heatmap and scatter plots to support functional modules such as DEG analysis, gene set enrichment analysis, functional analysis, and single cell profiling. Other tools, such as EnrichR, g:profiler and WebGestalt, focuses on functional enrichment analysis of gene lists (63-65). Although they are focused on a narrower task, they excel in the size and quality of their knowledge base and, in the case of g:profiler, number of organisms covered.

Multi-omics data analysis

In multi-omics studies, the objective is to integrate multiple datasets from different omics to study their relationships and generate novel insights. In the last decade, there have been significant developments in integrative methods. In this section, I am going to discuss three general categories of multi-omics integration: conceptual integration, network-based and data-driven integration (Figure 1.8).

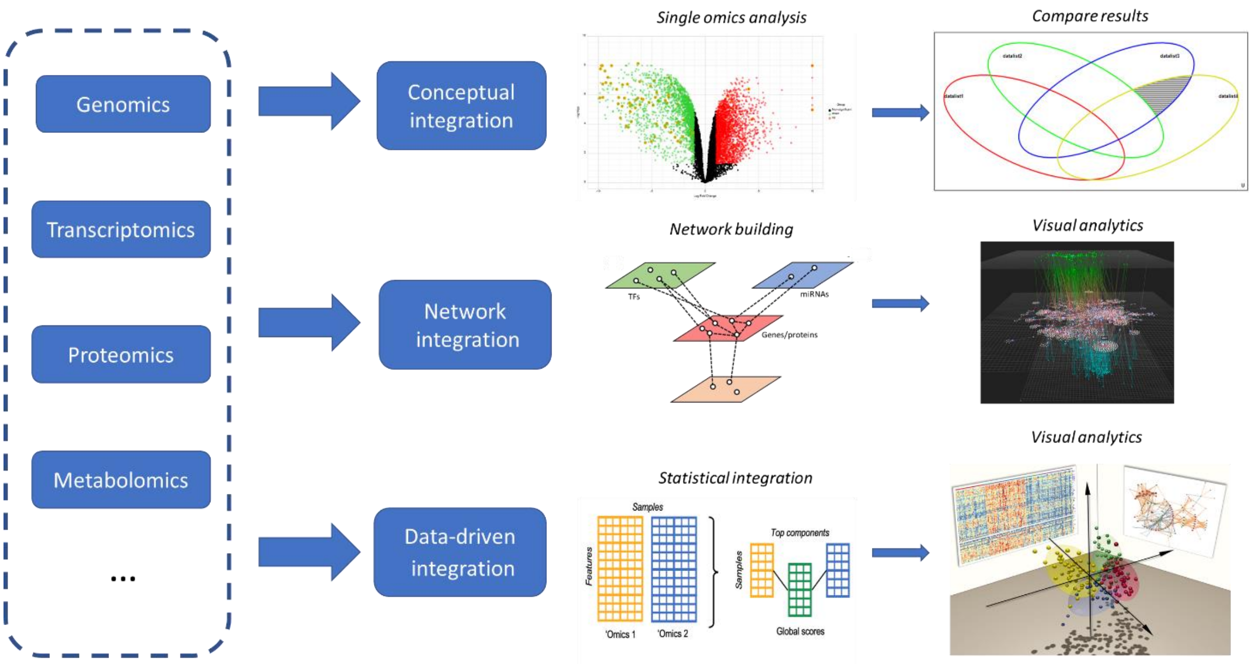


Figure 1.8 Overview of multi-omics integration approaches

1.3.5 Conceptual integration

Conceptual integration uses a straightforward approach of performing individual data analysis on each omics type and then aggregate the results obtained individually without analyzing the different datasets as a whole. This whole process relies on researchers' domain knowledge and capacity to synthesize findings. Although considered as a naïve approach, this type of integration has the advantage of being flexible and not limited by the heterogeneity of different types of omics data. For instance, the approach allowed the authors to link variation in metabolites to gene expression change by using a priori pathway information (66,67). The key drawbacks of this approach are time consuming, subjective, and will miss those associations that can only be identified from analyzing the datasets together.

1.3.6 Network-based integration

Network-based integration takes advantage of graph theory framework to visualize and analyze multi-omics datasets (19). The overall objective is to better understand the relationships between different omics features. This approach can be further divided into two subcategories: data-driven or knowledge-driven integration. The objective of data-driven integration is to infer significant associations between measurements of individual features from different omics type using statistical methods such as correlation analysis. It aims to use the datasets to infer overall patterns and shared signatures across multiple omics datasets. Knowledge-driven integration aims to map experimental data from multi-omics studies into the context of prior network interaction knowledge in the form of trans-omics networks. Most commonly used biological networks are protein-protein interaction networks and metabolic networks (68).

The most straightforward approach to build data-driven network is to infer relationships between features. This is achieved by assessing pairwise association measures by using methods such as correlation or mutual information. Pearson's and Spearman's correlation methods are designed to identify linear relationships for parametric and non-parametric data, respectively. Alternatively, nonlinear relationships can be detected using mutual information method (69). These methods have the advantage of not being affected by confirmation biases from prior knowledge. However, heterogeneity in omics datasets, the lack of correlation and varying correlation strength between certain pairs of associated omics features are some of the drawbacks (70). Also, they often yield very dense networks (high edge number) due to its inability to discriminate direct and indirect effects (71). Additionally, a study by Bradley and al. has found that correlation between metabolites and related genes can vary significantly depending of the study conditions (72).

Knowledge-based approach to network integration aims to study the known relationships between biological entities using prior knowledge framework. One of the strategies is to start with identifying molecules of interest (seeds) from the dataset using comparative analysis for instance, followed by identifying context relevant subnetwork by mapping these molecules into the interaction databases. This process is not only useful to connect the seeds but can also reveal additional molecules (their direct interaction partners) which may not be detected as significantly altered from experiments but may still play an important role. Overall, the process contains two main components: the underlying interaction database and the network building algorithm. Foremost, the quality of interaction data is primordial as the current landscape contains both computationally predicted and experimentally validated interaction data. For well-studied organisms, there is an abundance of experimentally validated interaction data, although they remain incomplete in most of the cases. This is not the case for non-model organisms which mainly contains computationally predicted interactions. Secondly, the choice of a network building procedure is crucial to control network size and to only include relevant nodes for subsequent analysis. The main objective of network building is to identify a subnetwork from the whole interactome that is enriched with seeds, to better understand the molecular context.

Network building can resort to straightforward approaches such as building zero-order or first-order networks. Zero-order approach identifies interactions between molecules of interests (seeds) and connect them. Main limitation of this approach is the possible lack of direct interactions between seeds, resulting in many orphan nodes and very sparse subnetwork. First-order approach extends zero-order network by including direct interacting partners of seeds in the network. The latter approach would connect more seeds together through these intermediate nodes. Please refer to Figure 1.9 for an example of first-order network. Prize Collecting Steiner Forest (PCSF) is an example of more advanced algorithms for network building. The algorithm itself is a well-known problem in graph theory. The goal is to identify subnetwork enriched with seeds in an undirected network in which the vertices are assigned prizes, and edges are assigned costs. The optimization objective is to minimize total cost assigned to edges while maximizing the total node prizes within the network.

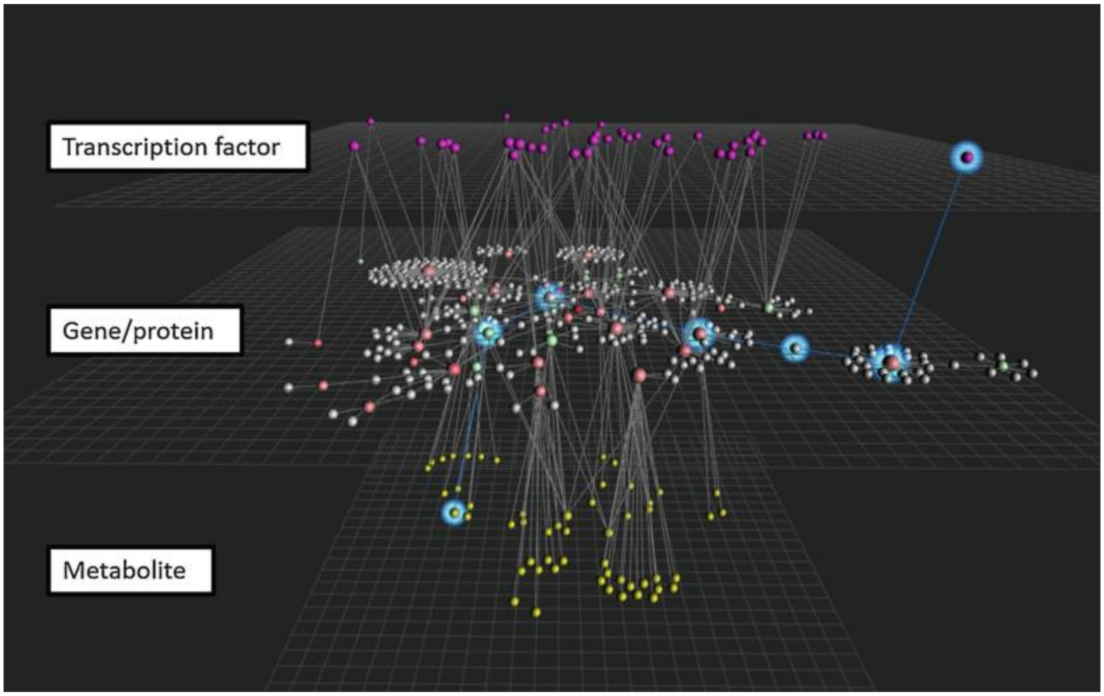


Figure 1.9 An example of first order multi-omics network generated using OmicsNet. The seeds are gene/protein nodes colored in red and green. Predicted transcription factors are colored in purple. Genes/proteins are colored by their expression value using a green-red gradient. Grey nodes refers to predicted interacting partner of seeds. Predicted metabolites are colored in yellow. A shortest path between a transcription factor and metabolite is highlighted in blue

Although knowledge-based approach is widely used, there remains numerous drawbacks from which I will name a few. First, the fact that this approach is confined to prior knowledge makes it unsuitable in studying non-model organisms. Another issue is the inherent bias of the knowledge domain towards well-studied molecules, granting them excessive importance. Finally, this approach is simply unsuitable for discovering novel relationships between biomolecules. Depending on the biological question of interest, data-driven integration or using a mix of both approaches may be better suited.

1.3.7 Data-driven integration

Data-driven integration relies on various statistical methods to integrate multi-omics datasets simultaneously. It is outside the reach of this thesis to review the whole literature of multi-omics integration methods. In the following section, a brief overview will be provided on dimension reduction and multi-view clustering methods that are used in multi-omics integration.

The different dimension reduction methods model features as a set and considers relationships between different features unlike conventional univariate statistical methods such as ANOVA or t-tests. One of its key features is the ability to perform dimension reduction, which is to project data into subspace with lower dimensions while capturing largest sources of variation. This is especially useful for visualization purposes when the subspace is reduced to two or three dimensions. Additionally, multivariate methods usually do not have strict assumptions on data distribution, making it flexible for omics data analysis (43).

Multivariate statistical methods such as Principal Component Analysis (PCA) (73) and Non-Matrix Factorization (NMF) (74) have been widely used on single data matrix. Many of these single dataset methods have been extended or adapted to integrate a pair of data matrices. Such methods include Procrustes analysis, Co-Inertia Analysis (CIA), Partial Least Squares (PLS) and Canonical Correlation Analysis (CCA). Procrustes analysis attempts to conform points from both matrices by applying linear transformation on one of the data matrices (51). It is widely used in the domains of ecological science and microbiology (75,76). CIA achieves integration by constraining projections of orthogonal axes while maximizing covariance of lower dimensional representations of the initial datasets (77). PLS integrates two data matrices by maximizing covariances between sets of variables. On the other hand, CCA maximizes correlation between components instead. There are many implementations derived from the methods mentioned above to deal with specific characteristics of omics datasets and generalized to integrate more than two datasets (78,79).

Overall, model-based integration is used for exploratory data analysis, exposing global structure across datasets, and highlighting batch effects within individual datasets. It provides an intuitive framework and visualization for further in-depth analysis of the elements within, both samples and features (genes, metabolites, etc.). Similarly, this type of framework facilitates downstream enrichment or pathway analysis. For instance, Multiple Co-inertial Analysis (MCIA) can facilitate the identification of features contributing most to global data structure by making them comparable across datasets using a transformation procedure.

In clustering analysis, the main objective is to identify cluster of samples or patients that are more similar to each other than the rest of the data. Single omics clustering remains effective, but it is missing out from the complementary and consensus information across different omics types. Multi-view clustering algorithms, a research topic investigated independently in machine learning community, is applicable to multi-omics datasets. Recent multi-omics studies have seen application of such algorithms (22,80).

In a review by N. Rappoport and R. Shamir (81), the authors classify multi-view clustering into three main categories:

1. Early integration: Concatenate omics data matrices to form a single multi-omics matrix and applies clustering algorithms on it.
2. Late integration: Each omics data matrix is clustered separately, and the results are integrated to obtain a single solution.

3. Intermediate integration: Use original omics matrices to build shared and omics-specific models.

Early integration is arguably the most straightforward approach where clustering is applied on an aggregated multi-omics matrix. Any clustering algorithm can be used but it suffers from several drawbacks such as different weights attributed to each omics layers due to different number of features, different data distributions and worsening the data dimension (aggregating features). Late integration is equally flexible as any clustering algorithm can be used to perform clustering on individual omics matrix and does not suffer from the drawbacks of early integration. However, there is a loss of weak signals from each individual omics. Intermediate integration includes several different types of methods: sample similarities-based, joint dimension reduction and statistical modeling. In the following section, several well-known implemented methods are going to be covered.

The iCluster algorithm is one of the earliest integrative methods. It performs multi-omics integration by mapping the individual data into a joint low-dimensional latent space (82). It was applied in lung and breast cancer data for the purpose of patient stratification using copy number variation and mRNA expression data (82). PINS is a late integration method that leverages connectivity matrices of different omics to performing cluster integration. Perturbations are applied on the data to test robustness of clustering results and help identify the optimal number of clusters. Similarity Network Fusion (SNF) was among the first sample similarity methods developed for multi-omics data (83). The algorithm first construct individual similarity network for each omics. The resulting networks are then fused together using message passing algorithms through an iterative process (84). Please refer to Figure 1.10 for a schematic view of SNF algorithm.

It has seen applications in the integration of cancer subtypes data from TCGA composed of gene expression, methylation and miRNA expression (83).

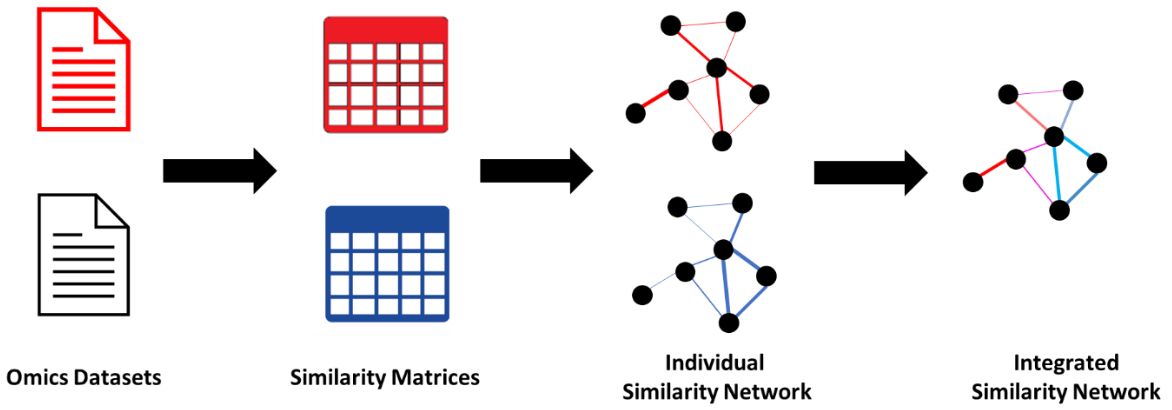


Figure 1.10 Schematic representation illustrating SNF approach on integrating two datasets from different omics. Edge color represents which dataset is contributing to the given similarity.

1.4 Theis statement

Multi-omics data analysis can reveal novel biological insights which lead to better understanding of health and disease. By coupling visual analytics with advanced statistics, analyzing single omics data and integrate omics data from different biological layers can address some of the current challenges in omics data analysis including large data size, high heterogeneity, and complexity. The tools and case studies described subsequently will propose potential biomarkers and biological processes involved in specific developmental, physiological or disease conditions.

1.5 Outline of achievements

The following section presents the three main projects composing this dissertation.

1. NetworkAnalyst: a comprehensive network visual analytics platform for gene expression analysis

I have updated NetworkAnalyst, a web-based platform dedicated for functional profiling of transcriptomics data. It offers a comprehensive workflow from raw RNA-seq data pre-processing to advanced comparative analysis and visual analytics. I have expanded the interaction databases considerably from generic protein-protein interaction (PPI) to cell-type or tissue-specific PPI, gene regulatory networks, gene co-expression networks, drug-protein and chemical-protein networks. A new module dedicated for meta-analysis of gene lists have been added. It is coupled with a series of visual analytics tools including interactive heatmap, Venn diagrams, chord diagrams, enrichment network and interaction network to further explore, compare and analyze the gene lists.

2. OmicsNet: a web-based tool for creation and visual analysis of biological networks in 3D space

   OmicsNet has been developed to create multi-omics interaction networks and visualize them in 3D space. It accepts as input one of multiple lists of genes, miRNA, transcription factors and metabolites to create and merge different types of biological networks. OmicsNet supports three different graph layouts to facilitate network navigation: force-directed layout, multi-layered layout and spherical layout. The network viewer is supported with a rich set of functions to perform coloring, shading, topology analysis, and enrichment analysis.

3. OmicsAnalyst: navigating complex landscapes of multi-omics data integration via intuitive visual analytics.

   OmicsAnalyst has been developed to integrate multi-omics datasets using well-established methods coupled with advanced visual analytics features. It proposes three different analytic tracks: (i) correlation network analysis, where users choose one of the proposed feature selection methods and explore relationships of important features in 2D or 3D network. (ii) cluster heatmap analysis, to visualize and analyze results from multi-view clustering methods on a pair of omics datasets using interactive dual heatmap. (iii) dimension reduction analysis, coupling integrative multivariate statistical methods with interactive 3D scatter plots to explore global data structures via corresponding score and loading plots.

Preface to Chapter 2

This chapter described an extensive update to NetworkAnalyst, a web-based visual analytics tool designed to perform functional profiling of transcriptomics data. It offers a comprehensive workflow of data processing along with differential expression analysis functions and various visual analytics options to explore and analyze transcriptomics data. It is composed of four different modules, depending on what type of input the user has: 1) Gene List Input; 2) Gene Expression Table; 3) Multiple Gene Expression Tables; 4) Raw RNA-seq data. It offers a wide array visual analytics tools to facilitate data interpretation and analysis. In this update, network visual analytics have been improved by expanding the interaction knowledge base from generic PPI to other interaction types and adding new network layouts. Secondly, enrichment network visual analytics were added to assess the overall relationships between enriched terms, and their associated genes. Lastly, to address the need of tools supporting meta-analysis of multiple gene lists, the existing visual analytics have been adapted to facilitate their simultaneous analysis.

Chapter 2: NetworkAnalyst - a comprehensive network visual analytics platform for gene expression profiling

This chapter has been published in Nucleic Acids Research

Guangyan Zhou[1], Othman Soufan[1], Jessica Ewald[2], Robert E.W Hancock[3], Niladri Basu[2] and Jianguo Xia[1,4*]

[1]Institute of Parasitology, McGill University, Montreal, Quebec, Canada, [2]Department of Natural Resource Sciences, McGill University, Montreal, Quebec, Canada, [3]Department of Microbiology and Immunology, University of British Columbia, Vancouver, British Columbia, Canada and [4]Department of Animal Science, McGill University, Montreal,
Quebec, Canada

*To whom correspondence should be addressed:

Tel: 1-514-398-8668

Email: jeff.xia@mcgill.ca

Zhou, G., Soufan, O., Ewald, J., Hancock, R. E., Basu, N., & Xia, J. (2019). NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. Nucleic acids research, 47(W1), W234-W241.**ch**

2.1 Introduction

The scientific community is in the midst of a boom of transcriptomics yet there are few accepted and standardized bioinformatics tools to organize, analyze, visualize and interpret the resulting big data. To deal with the challenges from such datasets, new-generation bioinformatics tools must be high performance (i.e. scalable for large data or user traffic), intuitive to use (i.e. to enable complex analytics via a simple interface) and universally accessible (i.e. web/cloud-based). Here, we introduce NetworkAnalyst 3.0 as a powerful web-based visual analytics platform for comprehensive profiling, meta-analysis and systems-level interpretation of gene expression data. NetworkAnalyst was first released in 2014 centered on PPI network analysis and visualization (61). It was soon updated (version 2.0) in mid 2015, with a completely revamped user interface and enhanced workflow for statistical meta-analysis of multiple gene expression studies (62). Over the years, we have made continuous updates and feature enhancements based on community feedback. According to Google Analytics, the public server has performed >220 000 data analysis jobs submitted from >14 000 users worldwide over the past 12-month period.

The development of NetworkAnalyst, and subsequent updates, have been driven by the practical data analysis challenges facing researchers from a wide variety of different areas. Addressing these needs has required different levels of effort and expertise. At the basic level, we have expanded support from the initial five model organisms to currently 17 species covering mammals, birds, bacteria, plants and parasites. In addition, many researchers do not have access to high-end computational infrastructure, and thus we have developed and made available a public Galaxy server to support raw RNAseq processing for all the 17 species. At the intermediate level, we have spent significant efforts in curating high-quality, comprehensive molecular interaction data to allow users to create gene regulatory networks, tissue or cell-type specific networks as well as gene co-expression networks to enable more biologically meaningful analysis. For gene expression analysis, we have implemented an interactive volcano plot and added the widely used gene set enrichment analysis (GSEA) method (85). At the advanced level, we have spent most of our efforts on developing and improving visual analytics methods to address several key challenges in big data analysis of omics data. To address the 'hairball' effect associated with large network visualization, we have implemented 3D and VR network visualization. For networks with hierarchical structures such as enrichment network, we have developed a 'meta-node' feature which can be expanded to show more details upon user click. To overcome the limitations of Venn diagrams and chord diagrams, we have developed simple yet powerful heatmaps to allow users to intuitively compare gene lists of varying sizes for meta-analysis. Finally, NetworkAnalyst now allows users to save their data analysis projects and resume analysis later. Meanwhile, we have performed thorough code refactoring, updated the framework, and enhanced the user interface to significantly improve its efficiency and user experience. We have also updated the frequently asked questions (FAQs) and have added new tutorials for first-time users. All these changes and updates have been released as NetworkAnalyst 3.0. It is now available freely at https://www.networkanalyst.ca.

2.2 Overall Design

NetworkAnalyst accepts five types of data inputs - one or multiple gene lists, a single gene expression data table, multiple gene expression data tables, raw RNAseq reads as well as common network files. To start analysis, users can click the corresponding circular menu from the NetworkAnalyst home page. Each data input corresponds to a data analysis module with specific data processing steps. The analysis results will be presented in several highly interactive visual analytics methods with built-in support for functional enrichment analysis against multiple libraries available from our knowledge base. The main workflow of NetworkAnalyst is summarized in Figure 2.1. In the following sections, we will focus primarily on the new or improved features introduced in the NetworkAnalyst 3.0. Other features can be found in our prior publications (59,61,62,85,86).

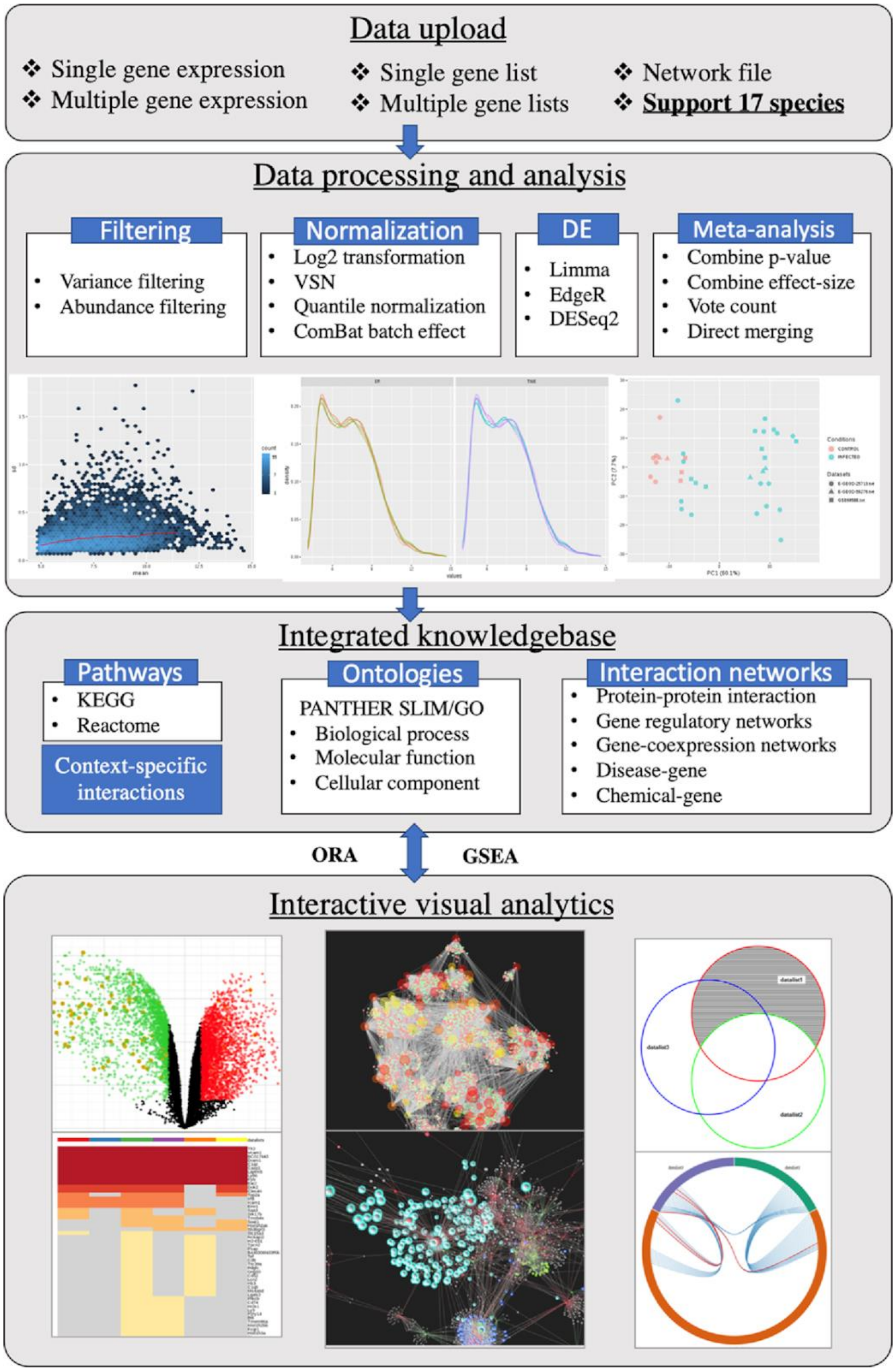


Figure 2.1 Overview of the workflow of NetworkAnalyst 3.0

2.3 Program description and methods

2.3.1 Enhancing gene expression analysis

Given the prevalence of transcriptome studies across life sciences, we have spent substantial effort in improving both the capacity and the workflow for gene expression analysis, with a particular focus on RNAseq data analysis and interpretation.

Raw RNAseq processing

NetworkAnalyst now features a Galaxy-based pipeline for processing raw RNAseq, which includes trimming, quality checking, read mapping and quantification. In particular, we have implemented both the classical spliced aligner—HISAT2 (87), as well as the ultra-fast pseudoalignment based method—kallisto (88) to support raw RNAseq mapping for the 17 species. The resulting gene count tables can then be used for gene expression analysis as described below.

Gene expression profiling

To enable more refined data analysis and to improve the user experience, we have expanded the previous single-page gene expression analysis module into multiple pages spanning data upload, quality check, normalization and differential expression analysis steps. Both the quality check page and the normalization page include a number of diagnostic plots to provide different perspectives on the data. For instance, users can view the distributions of gene expression values across samples (box plots) or experimental factors (density plots), and the effects of different normalization methods on sample clustering can be visualized via PCA plots. All these figures can be downloaded as high-resolution images for publication. Differentially expressed genes (DEG) can be identified using limma (89), edgeR (90) or DESeq2 (91). Users can further select different parameters based on their study designs and comparisons of interest.

Interactive volcano plot

This is a simple yet powerful visualization method that integrates statistical significance (p values) and biological significance (fold changes) to allow users to quickly identify the most promising gene candidates from differential expression analysis results. The interactive volcano plot was implemented based on the canvasDesigner package (92). Users can directly click any data point to view the corresponding gene name and its expression profile as a boxplot. Users can perform enrichment analysis on all DEG, up-regulated DEG, down-regulated DEG, as well as genes in the current selection. Double clicking any returned function name will highlight the corresponding genes in the volcano plot. A screenshot is shown in Figure 2.2A.

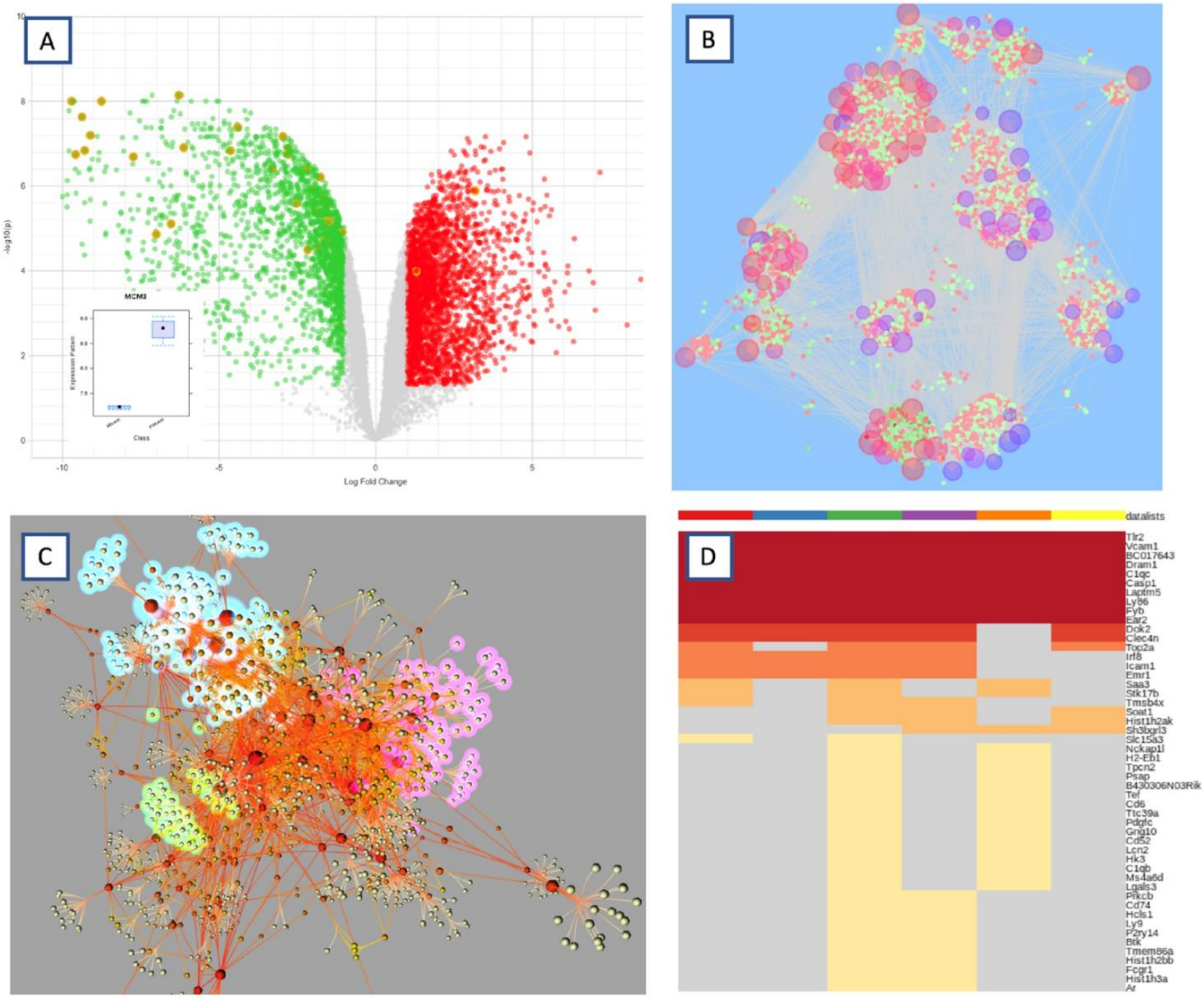


Figure 2.2 Screenshots of selected features introduced in NetworkAnalyst 3.0. (**A**) Interactive volcano plot. Users can click a data point to view the corresponding boxplot or click a function name to highlight the corresponding genes (shown in orange border). (**B**) Enrichment network with meta-nodes. Users can double click a meta-node (large semitransparent circles) to view all its associated genes (small solid circles). (**C**) 3D network viewer displaying a force-directed tissue-specific PPI network with several modules highlighted; D) Multi-list heatmap viewer. Users can intuitively identify and select shared or unique gene subsets and then perform enrichment analysis.

Gene Set Enrichment Analysis (GSEA)

In the previous versions of NetworkAnalyst, enrichment analysis was limited to over-representation analysis (ORA) on DEGs identified based on user selected cut-offs. Cut-off free methods, such as GSEA (85), utilizing the entire list of genes to compute functional enrichment, allows the detection of subtle yet consistent changes in gene expression profiles. GSEA in NetworkAnalyst is based on the high-performance *fgsea* R package (93). As GSEA requires a list of ranked genes as input, how to order the genes is an important parameter. NetworkAnalyst offers four robust gene ranking methods (moderated T-test, signal-to-noise ratio, fold change and statistics from the current DE method) based on a recent benchmark study (94). The results can be visualized as interactive heatmaps or enrichment networks. The heatmap visualization tool shows detailed gene expression patterns underlying individual functions; while the enrichment network tool (discussed further below) provides an overview of all enriched functions with similar ones connected by edges. A screenshot of an enrichment network is given in Figure 2.2B.

2.3.2 Expanding molecular interaction knowledgebase

Biological networks provide an intuitive framework to help understand complex molecular interactions. While PPI networks are widely used to aid in the interpretation of gene expression data, it is clear that other types of networks are also needed to obtain deeper mechanistic insights. For example, gene regulatory networks incorporating transcription factors (TFs) or microRNAs (miRNAs) are critical to infer causal link of molecular interactions, while applying tissue or cell-type specific PPI can greatly reduce false positives. In addition, gene co-expression networks based on large-scale gene expression studies can complement networks based on experimental evidence to facilitate novel hypothesis generation (95). We have spent extensive efforts to expand the underlying molecular interaction knowledge base as discussed below.

Context-specific networks

Our newly added human tissue-specific PPI data comes from the DifferentialNet database (96), covering 42 different tissues. The interaction data were generated by mapping tissue-specific co-expression data from GTEx (97) to experimentally detected PPI data from four major interaction databases. The tissue-specific co-expression data comes from the TCSBN database (98), covering 46 tissues. The cell type specific co-expression data comes from Immuno-Navigator (99), covering 24 immune cell types in mouse and human.

Gene regulatory networks

Transcriptional and post-transcriptional gene regulation plays important roles in many biological processes and cellular functions. We have added two key players in gene regulation: TFs and miRNAs. TarBase (100) and miRTarbase (101) have been used to obtain experimentally validated miRNA–gene target information, while ENCODE (102), JASPAR (103) and CHEA (104) have been used to obtain TF–gene target information. We also included the TF–miRNA–gene coregulatory networks built by RegNetwork.

Other biological networks

To address growing needs in toxicogenomics and pharmacogenomics, NetworkAnalyst now also includes protein-chemical interactions from the Comparative Toxicogenomics Database (CTD) (105) and protein-drug interactions from DrugBank (106). The CTD is a comprehensive public database of toxicogenomic information manually curated from the literature, providing key information on the effects of environmental chemicals. DrugBank is a public database specialized in drug molecular information, mechanisms of action and drug-target information for >10 000 drugs. Additionally, we have included gene-disease association networks for humans from

DisGeNet (107) which is a comprehensive database covering most of the known human disease-specific genotype–phenotype relationships.

2.3.3 Addressing the ‘Hairball’ issue

As biological networks become increasingly large and complex, they often suffer from the well-known ‘hairball’ effect which greatly reduces their practical utilities and uptake. Two general approaches can be performed to overcome this issue: trimming the default network to retain only those significant nodes/edges; and developing better visualization methods to reduce edge and node occlusions. In NetworkAnalyst 3.0, we have implemented new functions employing both approaches.

Network customization

The default networks are created by searching for direct interaction partners in the molecular interaction knowledge base. They are generally known as the first-order interaction networks. For very small networks, users can further expand the networks to create the second-order networks. When there are a large number of query genes (‘seeds’), it is reasonable to focus only on the interactions among those seeds (i.e. zero-order networks). However, many seeds could become orphan nodes when switching to zero-order networks. A ‘gentle’ approach is to extract, from the first-order network, a minimal subnetwork that maximally connects those seeds, a process known as Prize-collecting Steiner Forest (PCSF) algorithm. In NetworkAnalyst 3.0, we have added the support for efficient PCSF-based subnetwork extraction (108), as well as many other empirical trimming methods (available under ‘Network Tools’) based on shortest paths, node degree or betweenness values.

Network visualization in 3D and VR

Shifting from 2D to 3D can be a potential solution as it provides users a larger space for network layout and additional viewing angles. Although the hairball effect may still be problematic in 3D visualization, users will have more interaction freedom. Additionally, it may help expose some patterns otherwise undetectable in 2D visualization. Our implementation enables highly interactive network exploration in 3D space and allows extensive customization in terms of color, opacity, shading, etc. A screenshot of a 3D network generated by NetworkAnalyst is shown in Figure 2.2C. We have also added a virtual reality (VR) version of the 3D network based on the A-Frame framework (https://aframe.io/). VR brings to the table not only an immersive experience but also a much larger field of view that will not be limited by the size of the computer screen. Users with a compatible VR device (such as Oculus Rift) can view the network through web browsers. Please note our current implementation of the VR network is still in its prototype stage. We intend to develop a fully featured VR environment for 3D network visual exploration in the near future.

2.3.4 Powering multi-list comparisons through visual analytics

NetworkAnalyst supports comprehensive meta-analysis of multiple gene expression tables through various statistical methods. In many cases, however, researchers may simply have a number of different lists of DEGs generated from different studies or different comparisons from the same studies, for which they wish to compare and analyze. This observation has been demonstrated by the tremendous popularities of several web-based tools dedicated to functional interpretation of a given gene list, such as WebGestalt (109), g:Profiler (64) and Enrichr (63). The research community is increasingly interested in comparing results across multiple studies, a key feature missing in the aforementioned tools is the ability to perform meta-analysis of multiple gene lists to identify their shared as well as unique functions. There is an unmet need for intuitive yet flexible bioinformatics tools to allow researchers to easily compare multiple gene lists to gain biological insights.

In NetworkAnalyst 3.0, multiple gene lists can be easily uploaded using the gene list module. Users simply insert a '//' line to separate different gene lists when using the text area directly, or upload each gene list as an individual file. Please note NetworkAnalyst can also accept gene lists submitted programmatically as external requests based on our specified RESTful API (https://www.networkanalyst.ca/faces/docs/Resources.xhtml). After ID checking and conversion, users can visually compare different lists and perform enrichment analysis on a subset of genes generated from different set operations (i.e. unique, union, intersections for selected gene lists) using multiple visual analytics tools. A Venn diagram is probably the most straightforward way to compare a few gene lists - up to four gene lists are supported in our current implementation of Venn diagram. A chord diagram is also a popular visualization method to show pair-wise relationships between genes in multiple gene lists. However, a chord diagram can become too crowded when there are large number of genes and connections (>1000). To address these limitations, we have implemented two new methods to support the meta-analysis of genes and gene lists of arbitrary sizes.

Multi-list heatmaps

Heatmaps are a very popular visualization method for gene expression data. When used for visualizing multiple gene lists, heatmaps are able to show the presence or absence of genes in particular gene list in addition to the fold-change patterns. This form of presentation provides an overall picture of how the DEGs are shared across multiple lists. Our implementation allows users to directly click-and-drag to select a 'patch' of interest and perform enrichment analysis on the selected genes. Figure 2.2D shows a screenshot of the multi-list heatmaps in which different colors represent the frequencies of the genes appearing across all gene lists.

Enrichment network

To improve the interpretability of the results from enrichment analysis, we have implemented an interactive enrichment network viewer based on a similar concept introduced by ClueGO and EnrichmentMap (110,111). Users can now visualize the relationships among enriched function terms and their associated genes in a similarity network (Figure 2.2B). By default, the viewer shows a global enrichment network in which nodes represent functions and edges are determined by the overlap ratio between genes associated with the two functions. These nodes are implemented as meta-nodes. Users can double click to expand any meta-node to view its associated genes. Our implementation allows users to easily customize the style of the network (colors, layout, etc.) or to extract a subnetwork based on selected functions of interest.

2.3.5 Enabling resumable and reproducible data analysis

There is a growing interest in the bioinformatics research community to develop solutions for sharing data and analysis steps to support publications and scientific claims (112,113). Due to the wide array of visual analytics methods available in NetworkAnalyst, users may not be able to complete their analyses in a single session. This can partially explain the >220 000 data analysis jobs submitted to NetworkAnalyst from ∼14 000 users over the past year—it is likely that many jobs were re-analyzing the same datasets submitted from the same users. There is a need for NetworkAnalyst to store user data and analysis steps to allow users to resume their data analysis later.

In NetworkAnalyst 3.0, we have developed a project management component as an initiative to address the challenges associated with reproducible research. Users can now create up to 10 projects. These projects can be loaded, updated or deleted. Within each project, all key analysis steps are tracked. Since these projects need to be stored securely, users need to create an account to manage their projects. It should be noted, however, that creating an account, is not required for using any data analysis module in NetworkAnalyst 3.0.

2.4 Implementation

NetworkAnalyst 3.0 was implemented based on the PrimeFaces (v6.2) component library (http://primefaces.org/) and R (version 3.5.1). The various visual analytics methods have been developed based on several powerful JavaScript libraries including sigma.js (http://sigmajs.org) for 2D interactive network visualization, three.js (https://threejs.org) for 3D network visualization and canvasXpress (https://canvasxpress.org) for heatmaps and volcano plots. The system is hosted on a Google Cloud *n1-highmem-8* instance (52GB RAM and eight virtual CPUs with 2.6 GHz each). The project management component has been developed as a microservice hosted on a separate server using Spring Boot and Spring Security. As a web-based tool, NetworkAnalyst is mainly designed to support analysis of gene expression data generated from small to medium-sized studies. For raw RNAseq data processing, our Galaxy Server offers 100 GB disc space per user by default (∼30–50 samples dependent on the organisms and sequencing depth); For gene expression table, users can upload files with a maximum size of 50 MB (∼200 samples with ∼25 000 genes for each sample). For meta-analysis, users can upload up to 1000 samples in total. For large-scale studies, we recommend users to first process their data locally and upload gene lists for network analysis and visual exploration.

Table 2.1 Comparison with other web-based network analysis tools. Symbols used for feature evaluations with '√' for present, '-' for absent, and '+' for a more quantitative assessment (more '+' indicate better support). The URLs for each tool are given below.

Tools	**NetworkAnalyst**	**WebGestalt**	**g:Profiler**	**Enrichr**
Inputs	Gene lists, gene expression data, network files	Gene lists	Gene lists	Gene lists, BED file
Organisms	17 species	12 species	213 species	6 species
Gene expression analysis				
RNAseq processing	√	-	-	-
DE analysis	√	-	-	-
Enrichment analysis	++	+++	++	+++
Knowledgebase	++	+++	+++	+++
Network construction and visualization				
Network types	+++	++	+	+
3D/VR network	√	-	-	-
Meta-analysis and visual analytics				
Multiple lists	Enrichment analysis on any sets (union, intersection, unique)	Enrichment analysis on individual lists		-
Multiple tables	√	-	-	-
Heatmap view	√	√	√	√
Chord diagram	√	-	-	-
Venn diagram	√	-	-	-
Enrichment network	√	√	-	√
Volcano plot	√(genes)	√ (gene sets)	-	-

2.4.1 Comparison with other web-based tools

Table 2.1 shows the comparisons between NetworkAnalyst 3.0 and several other well-known web-based tools dedicated to functional profiling of transcriptomics data, including WebGestalt (109), g:Profiler (64) and Enrichr (63). The WebGestalt web application, first released in 2005, provides comprehensive enrichment analysis for 12 selected organisms and also supports user-supplied functional enrichment categories. The g:Profiler tool suite, first released in 2007, provides the broadest species coverage by supporting >200 species and corresponding gene ID conversions. Additional features include mapping human single nucleotide polymorphism (SNP) to gene name as well as ortholog search. The Enrichr web server, first released in 2013, provides the broadest functional coverage by supporting enrichment analysis against >100 gene set libraries. A key contribution of Enrichr is its curation effort, and allowing users to download their curated gene sets. Another unique feature of Enrichr is its support for BED file as input for enrichment analysis. These three tools are powerful web-based platforms that offer rich annotations for a given gene list. In contrast, NetworkAnalyst distinguishes itself from other web-based tool by providing cutting-edge network visualization, versatile visual analytics, comprehensive support for gene expression profiling, meta-analysis and multi-list comparisons. NetworkAnalyst 3.0 offers an end-to-end solution for RNAseq analysis - from raw reads mapping to differential expression analysis and identification of important pathways and functions.

2.5 Conclusion

NetworkAnalyst 3.0 is a unique online visual analytics platform specialized in transcriptome profiling, network analysis, and meta-analysis for gene expression data. NetworkAnalyst has been developed to address three unique gaps in the current landscape of bioinformatics tools. Firstly, NetworkAnalyst aims to provide a web-based tool for creating and visualizing biological networks to complement the widely used stand-alone tools such as Cytoscape (114). We will continue to add new features with a special consideration of emergent revolutions in web technologies (i.e. cloud, WebVR and browser computing) in the coming years. Secondly, NetworkAnalyst has filled a unique gap by enabling web-based meta-analysis of gene expression data. Gene expression meta-analysis is a very complex process and is usually performed by statisticians using R and Bioconductor packages rather than by average life science researchers. Curating as well as uploading and processing multiple datasets using online tools can be a challenging, unreliable (i.e. unstable connections) and time-consuming task. The implementation of the project management component is a first step towards addressing these concerns. Users can now save their projects (including datasets and steps) and resume analysis at a later time. In addition, the new multi-list comparison feature enables more flexible meta-analysis by accepting gene lists generated using users' own favorite tools and methods. Finally, it is now widely accepted that the over-reliance on p-values (among other statistical missteps) have contributed to the current crisis in reproducible research (8,115). With the proliferation of datasets that are increasingly large and complex, there is a great need to design and develop novel and intuitive bioinformatics tools to better educate, empower and engage users. We believe that integrating a range of visual analytics tools together with 'conversational' data analysis steps as used by NetworkAnalyst is a promising approach towards addressing this issue.

Preface to Chapter 3

Chapter 3 presents, OmicsNet, a visual analytics tool that performs multi-omics network integration using a knowledge-based approach. OmicsNet is designed to explore the overall molecular context of features of interest by leveraging well-established interaction databases. In contrast to NetworkAnalyst which is centered for the analysis of transcriptomics data, OmicsNet takes as input lists of genes, metabolites, transcription factors and miRNA and outputs a multi-omics interaction network that links them. Other key features include the use of 3D space to visually represent the interactive network and employment of multi-layered and module-based layout to facilitate data interpretation. Please refer to the following link for a video demo of 3D network visualization by OmicsNet (https://youtu.be/4q8LFQmlYNk).

Chapter 3: OmicsNet: a web-based tool for creation and visual analysis of biological networks in 3D space

This chapter has been published in Nucleic Acids Research

Guangyan Zhou[1], and Jianguo Xia[1,2]

[1]Institute of Parasitology, McGill University, Montreal, Quebec, Canada and [2]Department of Animal Science, McGill University, Montreal,
Quebec, Canada

*To whom correspondence should be addressed:

Tel: 1-514-398-8668

Email: jeff.xia@mcgill.ca

Zhou, G., & Xia, J. (2018). OmicsNet: a web-based tool for creation and visual analysis of biological networks in 3D space. Nucleic acids research, 46(W1), W514-W522.

3.1 Background and motivation

The growing applications of large-scale multi-omics studies in current life sciences have generated vast amounts of molecular measurements at DNA, RNA, protein, and metabolite levels. Novel bioinformatics tools and computational methods are urgently needed to help researchers analyze these complex datasets to facilitate systems-level understanding. Two general approaches have emerged - the statistical approach and the network-based approach. The statistical approach aims to identify overall patterns or shared signatures across multiple datasets by employing various multivariate statistical methods (43,116), while the network-based approach views the biological system as interconnected networks of molecular entities, and is primarily concerned with creating and computing on such networks (117,118)(3,4). Multivariate statistics are inherently complex. Although numerous methods have been developed to deal with multi-omics datasets, there is a general lack of well-established guidelines and strong use cases to promote their wide adoption and application (21). In contrast, the network-based approach is particularly appealing as networks can easily integrate new data into current knowledge framework and visually engage researchers to facilitate data understanding. Over the past decade, large-scale experiments have enabled comprehensive collection of high-quality molecular interaction data. Many excellent public databases and bioinformatics tools have been developed for storage, visualization, and analysis of such data (36,38,119-122). These expansive resources have made the network-based approach the preferred choice in current multi-omics data integration and systems biology.

The first step in the network-based approach is to create a subnetwork (or a few subnetworks) that connects significant molecules identified from individual omics data analysis. Protein–protein interactions (PPI) and metabolic reactions have been widely used for building such subnetworks. In general, there is a lack of easy-to-use bioinformatics tools that permits facile incorporation of important regulators such as microRNAs (miRNAs) or transcription factors (TFs) into biological networks. These two types of molecules are important players in gene regulations and are integral components in systems biology. High-quality public resources housing gene regulator data have become readily available in recent years. For instance, TarBase (100) and miRTarBase (101) are two comprehensive databases that host experimentally validated miRNA-target interactions. Meanwhile, the ENCODE (102), JASPAR (123) and TRRUST (124) databases have provided high-quality information on TFs and their potential target genes. Integrating these resources to allow users to easily include these important players into widely-supported PPI or metabolic networks would therefore enable deep insights for systems understanding.

After the creation of subnetworks that can, ideally, connect a significant portion of the molecules of interest, the next step is to analyze the subnetworks. Although graph theory is often used to help identify important patterns and links, a key strength of network analysis lies in organizing and visualizing the considerable knowledge about the interplay among biological molecules to help researchers make informed decisions or to develop new hypotheses (68). Therefore, an important goal of network visualization is to facilitate easy interpretation and absorption of large quantities of information without being overwhelmed by it. However, as networks become larger, it often leads to the well-known ‘hairball’ effect, which is caused by a large number of overlapping nodes and edges. Many empirical methods have been developed to address this issue such as trimming uninformative nodes, edge bundling or applying different layouts (39,125). One potential solution is to increase the visualization space from the conventional 2D to 3D space, thereby providing more viewing perspectives and reducing intersections between nodes and edges. In addition, the extra dimension can present critical information unique in multi-omics and time-series data to facilitate systems-level understanding (126,127).

Most current network visualization tools are standalone programs focusing primarily on 2D network visualization, such as Cytoscape (36) and Gephi (128). A few of them also support 3D visualizations such as BioLayout3D (129), iCAVE (130), NAViGaTOR (121), Arena3D (127) and 3DScapeCS (131). Over the past several years, there is a clear trend to move away from standalone applications towards integration of visualization within web browsers (62,132,133). To the best of our knowledge, no dedicated web-based tools are currently available to support 3D visualization of biological networks. There are some technical reasons behind this. Early 3D rendering was often implemented using Flash or Java 3D, both of which require plugins in order to work within a web browser. In addition, 3D rendering is inherently a computationally intensive task, and displaying large networks in 3D could easily exceed the computing capacity of early web browsers. The situation has significantly changed over the past few years. Modern web browsers are much more powerful. Browser-based applications with hardware acceleration using graphics processing units (GPUs) can deliver excellent user experience through their interactive, media-rich interfaces. The recent arrival of WebGL technology, now standard in all modern web browsers, has made it possible to implement interactive 3D graphics directly in a web browser. When properly implemented, WebGL can deliver higher performance as compared to other existing technologies such as canvas or scalable vector graphics (SVG) (134). Leveraging this new web technology to enable intuitive online 3D network visualization represent a promising direction to address the current challenges in large network visualization and multi-omics integration.

We introduce OmicsNet, a novel web-based tool for biological network creation and visual exploration in 3D space. OmicsNet was developed using the state-of-the-art WebGL technology to enable 3D network visual analytics, with built-in support for flexible creations of composite networks. The key features of OmisNet include:

- Accepting lists of genes/proteins, transcription factors, miRNAs, metabolites, as well as network files (.txt, .sif or .graphml);
- Supporting ten molecular interaction databases on protein-protein, miRNA-target, TF-target and enzyme-metabolite interactions, with multiple procedures for network customization;
- Fully-featured 3D network visualization system supporting three layouts (force-directed layout, multi-layered perspective layout and spherical layout) and a wide array of 3D visual effects and interactions (shading, zooming, highlighting, rotating, drag-and-drop, etc.);
- Comprehensive support for functional analysis based on GO, KEGG, Reactome and PANTHER (135-138), as well as network topology analysis including module detection, computing shortest paths and node centralities;

OmicsNet contains a comprehensive list of frequently asked questions (FAQs) and multiple tutorials on different use cases to help researchers navigate common analysis tasks. The public server is freely available at http://www.omicsnet.ca.

3.2 Program description and methods

OmicsNet is mainly composed of three web pages corresponding to the three tasks - data input, network creation, and network visual analytics. Figure 3.1 shows the overall design and workflow of OmicsNet. Users can upload lists of genes/proteins, TFs, miRNAs or metabolites to search different molecular interaction databases. The results will be used to create different subnetworks, which can be explored in our 3D visualization system. Each component is furnished with various options to facilitate users' tasks. The key features of each page are described in the sections below.

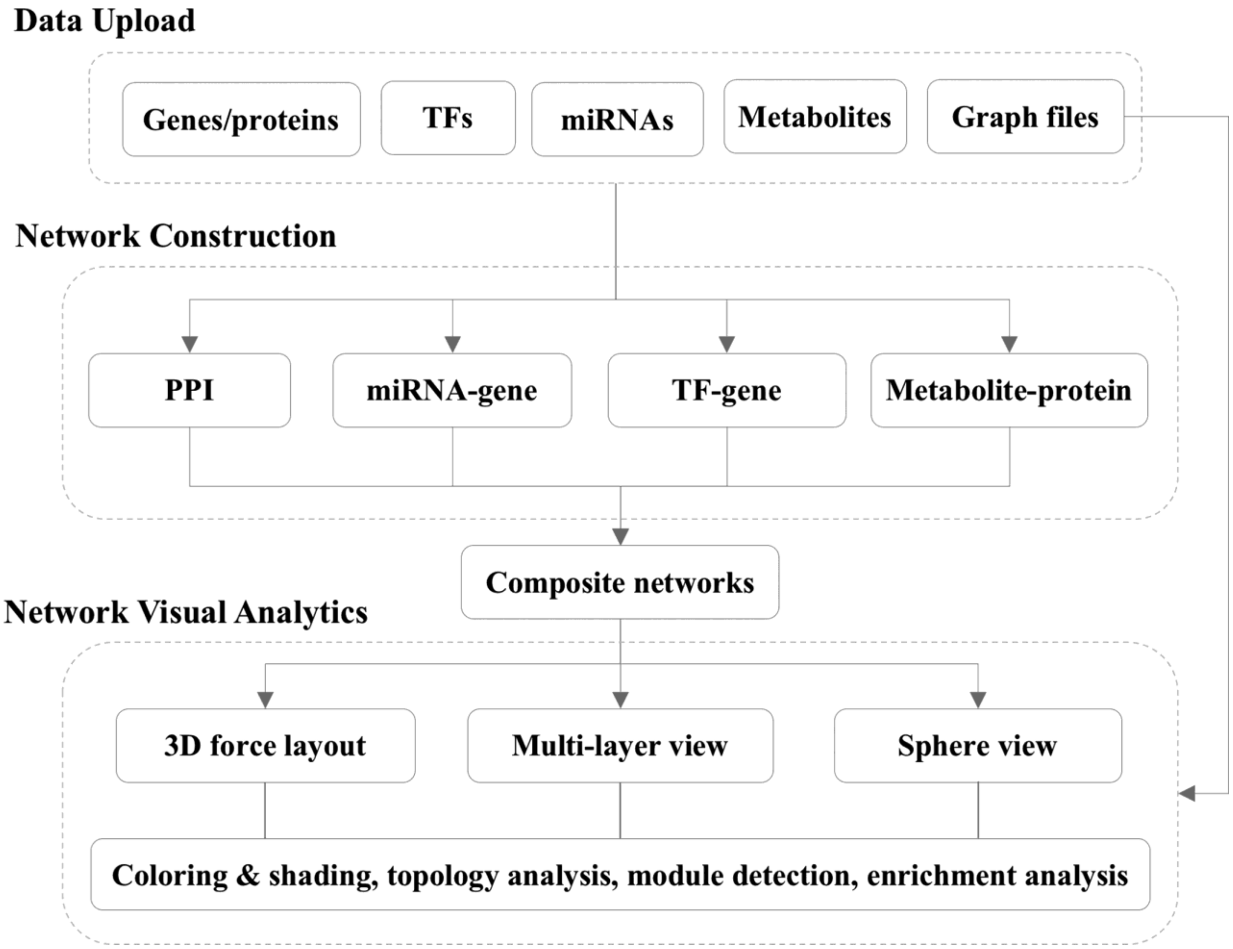


Figure 3.1 The overall workflow of OmicsNet. Users can upload lists of genes/proteins, TFs, miRNAs or metabolites to search different molecular interaction databases. The results will be used to create composite networks, which can be explored in a powerful 3D visualization system with comprehensive built-in support for different layouts, topology analysis and functional analysis.

3.2.1 Creation of knowledge base on molecular interactions

To support the construction of biological networks for different types of molecules, the first task is to create a comprehensive knowledge base on molecular interactions. In addition to PPI and metabolic interactions, we have also included transcriptional and post-transcriptional regulations.

Together they represent the four main types of molecular interactions in a simplified biological system. In total, data from ten different databases were collected, including three PPI databases (STRING (119), InnateDB (38), and IntAct (120)), two miRNA-target database (TarBase (100) and miRTarBase (101)), two metabolic databases (KEGG (136) and Recon2 (139)), and three TF-target databases (TRRUST (124), JASPAR (123), and ENCODE (102)). These publicly available databases are well maintained. We will perform annual check to synchronize our knowledge base with the major releases of these databases.

3.2.2 Data upload and processing

The query input can be one or multiple lists of genes/proteins, miRNAs, TFs or metabolites. OmicsNet currently supports nine organisms (human, mouse, rat, cattle, chicken, zebra fish, fruit fly, Caenorhabditis elegans and Schistosoma mansoni). In addition to supporting creation of conventional PPI, miRNA-gene, TF-gene and metabolic networks, OmicsNet has been designed to support three general use cases for systems biology and multi-omics integration: (i) starting from a list of genes, proteins or metabolites to build PPI or metabolic networks and further include miRNAs or TFs that target these nodes; (ii) starting from a few miRNAs or TFs to identify their target genes and further add interactions between these target genes/proteins based on PPI information. Note it is not advisable to start from a long list of TFs or miRNAs as primary queries because they tend to have large numbers of interaction partners, making it impossible to identify any meaningful connections through visual inspection of the resulting networks; (iii) starting from multiple lists of molecules (genes, miRNAs, and TFs) to identify known interactions among them. Finally, users can also directly upload their own networks in several common graph file formats (.txt, .sif, or .graphml) for 3D visual exploration.

3.2.3 Network construction

After users have uploaded one or more lists of molecules of interest, they can proceed to the next page for network building. The interface allows users to select one or more (up to three) types of interactions (PPI, miRNA-gene, TF-gene or protein–metabolite) to be included in the network. For building composite network containing more than one interaction types, users need to specify the order of network creation (primary, secondary, or tertiary interactions). The primary interaction should be selected to build networks consisting of molecular entities of main interest and their immediate interactors. The secondary and tertiary interactions are mainly to 'enrich' the information contained in the primary network through: (a) adding new edges - when the PPI database is chosen as secondary, the process will introduce new edges between gene/protein nodes in the current network; or (b) adding new nodes - when the TF or miRNA database is chosen, the process will introduce new regulator nodes that target gene nodes in the current network. If multiple lists are uploaded, the lists corresponding to the secondary and/or tertiary interactions will serve as constraints to make the resulting networks more context-specific by filtering out nodes that are not in the input lists. In addition, we have implemented the 'targeted node search' function, which allows users to search for higher-order interactions for a selected node during the network visualization stage. The details will be described later in the corresponding section.

Once interaction types are chosen and submitted for network building, a table will be displayed indicating the number of edges and nodes of the resulting networks to help users make decisions regarding whether to perform network filtering or proceed to network visualization. The purpose of network filtering is to reduce the network size by excluding less-informative nodes based on their topological properties, such as degrees or betweenness. Users can also compute and extract a minimum network that connects all current seed nodes.

3.2.4 Network visual analytics

OmicsNet offers comprehensive options for network visualization, customization, topology analysis, and functional analysis. A screenshot of the Network Viewer page is shown in Figure 3.2A. The top tool bar contains various menu items for network viewing and customization, the left panel displays node-related information, the center panel shows the network, and the right panel consists of various functions for enrichment analysis and network topology analysis.

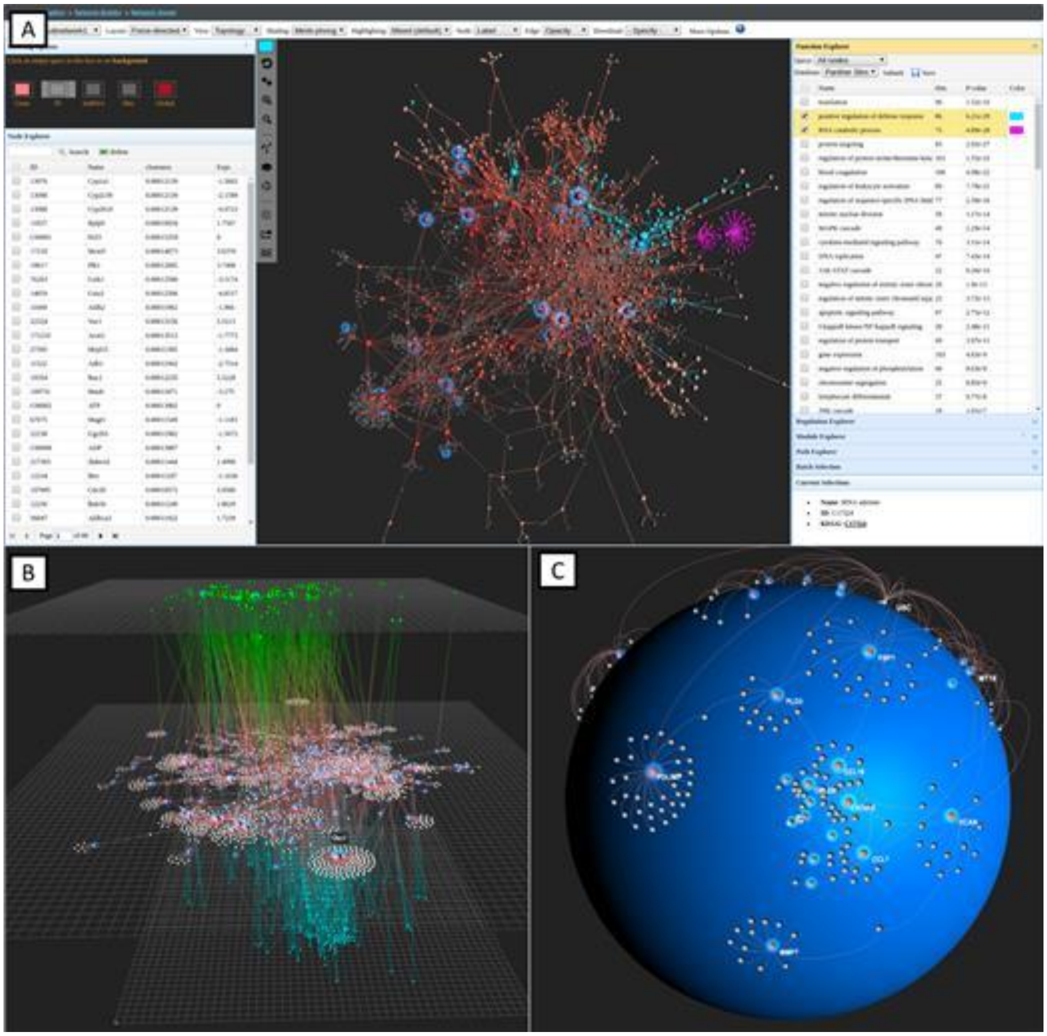


Figure 3.2 Some screenshots of the Network Viewer showing the main features and different network layouts. (A) A force-directed subnetwork composed of ~2000 nodes and ~4000 edges. Seed nodes are indicated using halo effect, and nodes from two enriched pathways are highlighted in different colors. (B) A 2D perspective view of PPI subnetwork further enhanced with TFs and miRNAs targeting the key genes; (C) A spherical layout showing a module extracted from a large PPI network

Visual exploration through mouse controls

OmicsNet allows users to intuitively navigate 3D network using a mouse or trackpad. The basic mouse controls are described below:

- Zoom in/out: scroll the mouse wheel in the middle. Node labels will show up automatically based on the zoom levels;
- Rotate the current view: press the left mouse button and drag. The network will stay in the center;
- Obtain node information: move the mouse over a node to show its label; click a node to display more detailed information about the node in the 'Current selections' panel on the bottom right;
- Drag and drop: in the 3D force-directed layout, users can directly drag a single node or a group of highlighted nodes depending on the current scope selection. In the 2D perspective mode, users can drag and drop individual layers using the grey triangle located at one corner of the layer. Node dragging is not yet supported at the moment in the spherical layout.
- Other advanced options: users can right-click on a node to search for interaction partners for this particular node against several databases (targeted node search will be discussed later), or select two nodes (source and target) and search for shortest paths between them.

Coloring

Coloring is probably the most important factor for effective visualizations. OmicsNet provides three places for users to adjust the colors of their networks. The 'Coloring Options' panel on the top-left corner allow users to set the background color, as well as to customize the node colors for different molecular types. The 'Node' option on the top tool bar provides a comprehensive list of coloring schemes based on different node topology measures or node expression values (if available). Some of the most commonly used functions are provided in the vertical toolbar located inside the network view. Located on the top is the color picker, which is used to set the current highlight color that will be applied in subsequent highlighting when users double click a node or click the halo icon (a circle with rays) to indicate the ‘seed’ nodes.

Shading

Shading is a unique feature in 3D visualization. When applied, the colors of a node surface will vary based on its angle and distance to the light source to produce more realistic 3D effects. To minimize memory load, the default network nodes are generated using premade texture mapped to point primitives to simulate 3D effect. OmicsNet supports six different shading options under the ‘Shading’ drop-down menu. Note the ‘Mesh-phong’ shading was implemented based on 3D mesh objects, which is more memory intensive thus only suitable for small and medium networks. For visualization of very large networks, it is recommended to turn off the shading effects for better performance.

Node highlighting

This is an important function to help bring out important nodes and connections. OmicsNet currently supports three options for node highlighting - mixed mode (default), halo effect and node color. In the default mixed mode, halo effect is used for node searching (when users click a node name in the node table), and for highlighting seed nodes (when users click the halo icon); while the node coloring is used for direct node highlighting (when user double click a node in the network) and for highlighting functions, modules, or shortest paths (when users click an item from the results of enrichment analysis, module detection, or shortest path finding). The highlighting color (including the color for halo effect) is controlled by the color palette located on the top-left corner of the center panel.

Network layout

Network layout (arrangements of nodes) plays a critical role in revealing important patterns during network visualization. Unlike the 2D layout where numerous algorithms have been implemented, very few ready-to-use algorithms are available for 3D network layout. OmicsNet offers the standard 3D force-directed layout as default. We have also spent significant efforts to implement two other layouts - a multi-layered 2D perspective layout and a 3D spherical layout. These three layouts are described below.

Force-directed layout. This algorithm was adapted from the standard 2D force-directed layout algorithm (37). It rearranges nodes in the current network using a physical model where all pairs of nodes repulse and adjacent nodes attract each other with edges acting as springs. It often results in an aesthetically pleasing graph with reasonable node distribution and clustering. An example is shown in Figure 3.2A. In some cases, the default force-directed layout in 3D may seem even more cluttered than the 2D view. There are several options to help partially resolve this issue including edge bundling, manual drag-and-drop of nodes to reduce overlap, decreasing edge opacity using the ‘Edge’ option in the top toolbar, or rotating the network to a different viewing angle.

Multi-layered perspective layout. When networks contain more than one node type (i.e. bipartite or tripartite graphs), it is often more intuitive to apply a multi-layered layout that takes advantage of the best of both 2D and 3D. This layout, first introduced by Arena3D (127), separates the network into an array of 2D networks using existing context information (i.e. types of molecules). This arrangement can greatly reduce the number of edge-crossings and emphasizes the source data type of each node. This feature is also available in iCAVE (130). An example is shown in Figure 3.2B. Users can use their mouse to move each layer by dragging the grey triangle at one corner to improve the layout. The type of layer (grid, plane, or none) can also be specified using the corresponding option under the ‘More Options’ menu.

Spherical layout. The spherical layout is inspired by flight paths around the globe, which is implemented by projecting a 2D force-directed network onto the surface of a sphere. This layout improves the visual experience in some cases by reducing visual occlusions and avoids information overload by showing only a part of the network. An example is shown in Figure 3.2C. Users can change both the color and opacity of the globe using the corresponding option under the ‘More Options’ menu.

Functional and topology analysis

OmicsNet supports functional enrichment analysis on genes displayed in the current network. It uses hypergeometric tests for over-representation analysis (140), and can be performed against GO, PANTHER GO-Slim, Reactome or KEGG pathways (135-138) . OmicsNet supports three network topology analyses including node centrality analysis, module detection, and shortest path finding. Five different node centrality measures can be computed (degree, betweenness, closeness, eigenvalue, and transitivity), with degree being the default. To view different centrality measures, users can use the ‘Topology’ option under the ‘View’ menu or the ‘Color’ option under the ‘Node’ menu. Module analysis aims to find subsets of nodes that are more closely connected than expected by chance. Three module detection algorithms are supported in OmicsNet including InfoMap (141), Walktrap (142) and Label Propagation (143). Finally, users can use the ‘Path Explorer’ panel to search for the shortest paths between any two nodes of interest. Users can either enter the corresponding node IDs or right click the two nodes to define the source and target. Click a returned path will highlighted it in the current network.

Targeted node search

Due to practical reasons, the network creation interface does not allow users to introduce high-order interaction partners in batches (i.e. for all nodes). To address this limitation, we added the ‘Targeted node search’ to allow users to search higher-order interactions for a particular node displayed in the current network. To do this, users must right click a node of interest to show a drop-down menu containing different databases, then click to search a particular database. The detailed results will be displayed in the ‘Regulation Explorer’ panel on the right. Users can then use checkboxes to select one or more hits, and then click the ‘Add nodes’ button located directly above the result table. These new nodes will be added to the current network via connections to the target node.

Other features

The top menu bar contains most of the functions related to network viewing and customization. From the left side, the ‘Network’ menu allows users to access the other subnetworks created during network building; the ‘Layout’ menu contains the three different layout options; the ‘Shading’ menu allow users to select different shading effects or turn off the shading; the ‘Node’ and ‘Edge’ menus allow users to customize the node style (size, color and label) and edge style (opacity, color and bundling). Finally, the ‘More Options’ menu contains various advanced functions to customize the scope of selection for highlighting, dragging as highlighting styles. The network can be exported as a PNG image or graph files (.txt, .sif or .graphml) in the ‘Download’ menu.

4.3 Case study: Understanding complex immune regulations during helminth infection

Parasitic nematodes (helminths) are known to employ a wide array of immunomodulatory mechanisms in order to maintain their long-time survival in the host (144). To better understand the effects of helminth infection, we recently performed a meta-analysis of multiple gene expression datasets from helminth-infected mice, and revealed a core signature of genes that are differentially expressed across multiple independent studies (17). It is of great interest to further identify potential regulators (i.e. miRNAs or TFs) involved in the host immune response. To achieve this, we first built a PPI network using the InnateDB database from the signature gene list that maximally connects all seed genes and then further included miRNA-gene and TF-gene regulatory relationships using the miRNet and TRRUST databases. From the 2D perspective view, Sp1 and mir-9-5p clearly stands out as the key regulatory hub nodes in the composite network. Literature search indicates that both molecules play important roles in the immune system. Sp1 is a transcription factor involved in the regulation of Il-10 (145), a key effector in regulatory T cell response that mediates helminth-mediated immunoregulation (146), while the miRNA-9 family is involved in the regulation of the immune response (147,148). More detailed step-by-step analysis together with screenshots are available as Tutorial #4 on the OmicsNet website.

4.4 Implementation and tools comparison

4.4.1 Implementation

OmicsNet was developed using a server-client design. The server side was implemented using the PrimeFaces component library (version 6.1) for the web framework, and R (version 3.4.3) for back-end computing. The client side was implemented based on JavaScript using the Three.js library (https://threejs.org/) as an interface to WebGL. WebGL can take advantage of the GPU acceleration by sending and executing code directly on the GPU to render graphics. This type of code is termed shader. To minimize the memory load and computational resources required, we used material rendered using low-level custom shader to represent nodes as opposed to memory intensive meshes. Additionally, to minimize the instances of time-consuming data passing to CPU, we store the geometry data of nodes and edges in buffers before sending them. Our empirical testing shows that OmicsNet can display large networks with ∼10,000 nodes. A key limiting factor in terms of performance is the high interactivity of the current implementation. Supporting features such as drag-and-drop and dynamic updating visual properties (color, size, etc.) of nodes/edges requires a large amount of event listeners which will negatively impact the performance in the cases of larger networks. We are developing a specialized version for 3D viewing only (zoom and rotate) that will allow visualization of up to one million nodes with same size and color. We intend to add this option in the near future. Meanwhile, we recommend users to keep network size between 200 and 2000 for practical reasons. OmicsNet also supports retina display by automatically adjusts the pixels of rendered networks depending on the user's screen resolution. Since most of the network visualization functions come from browser-side JavaScript functions, its performance is dependent on the user's browser and graphics card. The public server is hosted on a Google Cloud Engine with 30GB of RAM and eight virtual CPUs with 2.6 GHz each. OmicsNet has been tested in most major web browsers such as Chrome 50+, Firefox 47+, Safari 10.1+ and Edge 12+ with WebGL enabled.

4.4.2 Comparison with other tools

OmicsNet is a 3D network visualization and integrative analysis tool containing a comprehensive built-in molecular interaction knowledge base that supports an array of different organisms. To the best of our knowledge, it is currently the only web-based application dedicated for visualizing biological network in 3D space. Table 3.1 compares OmicsNet with several well-known stand-alone tools that support 3D biological network visualization, including 3DScapeCS, BioLayout3D, Arena3D, and NAViGaTOR. Compared to these tools, OmicsNet distinguishes itself as being the only web-based tool with comprehensive built-in support for generation of different types of molecular interaction networks and a fully-featured 3D visualization system.

Table 3.1 Comparison of OmicsNet with other network visualization tools

Tools	**OmicsNet**	**NetworkAnalyst**	**BioLayout3D**	**Arena3D**	**NAViGaTOR**
Platform	Web	Web	Standalone	Standalone	Standalone
Inputs	One or more lists of genes, proteins, TFs, miRNAs, metabolites; Or graph files (.sif, graphml)	List of genes, expression matrix	List of genes, Multiple graph files, expression matrix	Multiple graph files; time-series data	List of genes, multiple graph files
Network Construction and Integration					
Built-in database support	Yes	Yes	Yes	-	Yes
Network integration	Up to three types of interactions	-	-	-	-
Network Visualization and Analysis					
3D visualization	Yes	-	Yes	Yes	Yes
2D perspective layout	Yes	-	-	Yes	-
Spherical layout	Yes	-	-	-	-
Node drag-drop	Yes	Yes	2D only	-	Yes
Enrichment Analysis	GO, KEGG, Reactome, PANTHER	GO, KEGG, Reactome	-	-	-
Module Detection	Yes	Yes	-	-	-

The URL for each tool is given below the table (note, evaluation for 3DScapeCS is based on functions offered by the plug-in itself).

- OmicsNet: http://omicsnet.ca/.
- 3DScapeCS: http://scape3d.sourceforge.net/.
- BioLayout3D: https://kajeka.com/graphia-professional/.
- Arena3D: http://arena3d.org/.

- NAViGaTOR: http://ophid.utoronto.ca/navigator/.

4.5 Conclusion

Current limitations and future perspectives

While the standard 3D visualization increases the viewing space and provides greater freedom in navigation, new issues are introduced such as edge occlusion and lack of perceptual reproducibility due to excessive numbers of viewing perspectives. To address these issues, we have implemented two enhanced layout options by dividing the nodes into multiple layers based on node types (the multi-layered perspective view), and by projecting the network on a globe's surface to mask network complexity while maintaining connectivity and ease of navigation (the sphere view). To further reduce edge occlusion, we implemented a force-directed edge bundling. In the future, we will implement additional network layouts and editing options to improve both the performance and visualization experience. Meanwhile, we also intend to increase its interoperability with community network visualization tools such as Cytoscape (36) and Gephi (128). At the moment, OmicsNet does not support directed or weighted edges, both features are important for many biological network visualization and interpretation. This will be our focus in the next updates. Other features to be added is the support of time-series, dynamic networks and general functionalities to perform differential network analysis. Indeed, as tremendous progresses have been made in the field of personalized medicine, there is an increasing need in the processing and visualization of -omics data from a single source over a period of time (149). This remains a huge challenge in the field and novel features such as integrating animation and other additional dimensions could facilitate its visualization and analysis (150). Another extension of the current work is to explore the effects of virtual reality (VR) through browsers using the WebVR API. This is already achievable with either Firefox or Chrome using a VR device such as the Oculus Rift.

Conclusion

Driven by the growing numbers of studies on multi-omics data integration and systems biology, there are strong demands for user-friendly web-based tools to allow researchers to easily create, integrate and visualize different types of biological networks. To address this need, we have developed OmicsNet to support intuitive network construction from a single or multiple lists of molecules. To facilitate data visualization experience, OmicsNet leverages the powerful WebGL technology to enable native 3D rendering of complex biological networks within modern web browsers. Three graph layouts have been implemented to provide different perspectives of the same network. The interface allows users to easily customize their visualizations through coloring, shading, highlighting, drag-and-drop, etc. In addition, users can also perform targeted node search, functional enrichment analysis, module detection, and shortest path computing. OmicsNet therefore fills an important gap by providing an easy-to-use web-based tool for 3D network visual analytics.

Preface to Chapter 4

Chapter 4 presents OmicsAnalyst, a visual analytics platform focusing on data-driven multi-omics integration. The platform offers three different integrative tracks: correlation network analysis, heatmap clustering analysis and dimension reduction analysis. Correlation network analysis aims to identify correlative relationships between features across two omics layers by using both univariate and multivariate methods. The result is visualized using interactive 2D/3D network. Heatmap clustering analysis uses cutting-edge multi-view clustering method coupled with dual-heatmap viewer. Finally, the last track couples dimension reduction methods with interactive visualization of score plots, loading plots and biplots in 3D space. Please refer to the following links for a video showcasing dimensionality reduction analysis (https://youtu.be/3_no0nCH2uE) and joint heatmap visualization (https://youtu.be/DWoeL1y9FHU).

Chapter 4: OmicsAnalyst - a comprehensive web-based platform for visual analytics of multi-omics data

This chapter has been published in Nucleic Acids Research

Guangyan Zhou[1], Jessica Ewald[2] and Jianguo Xia[1,3]

[1]Institute of Parasitology, McGill University, Montreal, Quebec, Canada, [2]Department of Natural Resource Sciences, McGill University, Montreal, Quebec, Canada and [3]Department of Animal Science, McGill University, Montreal, Quebec, Canada

*To whom correspondence should be addressed:

Tel: 1-514-398-8668

Email: jeff.xia@mcgill.ca

Zhou, G., Ewald, J., & Xia, J. (2021). OmicsAnalyst: a comprehensive web-based platform for visual analytics of multi-omics data. Nucleic Acids Research.

4.1 Introduction

The rapid development and increasing accessibility of various omics profiling technologies such as massive parallel sequencing and mass spectrometry have made multi-omics data collection more routine practices in recent years. These multi-omics studies promise to provide more holistic pictures to enable comprehensive understanding of complex diseases and biological processes (13,151). As a result, the last few years have witnessed a growing number of bioinformatics tools and statistical methods developed for multi-omics integration (21,81). These computational approaches can be largely classified as either knowledge-driven or data-driven strategies. The knowledge-driven strategy is well established. A typical example is to map genes and metabolites of interest into known metabolic pathways or networks and then visually explore the results for hypothesis generation (152-154). A key limitation of this strategy is its dependency on a prior knowledge base. Data analysis and interpretation will be conducted within the confines of this knowledge domain, making it unsuitable for novel discoveries and applications to non-model organisms. The data-driven strategy, on the other hand, depends primarily on the datasets themselves, and can be applied in a more general and unbiased manner (155).

Many different data-driven approaches have been proposed and practiced for multi-omics integration. They can be loosely put into three categories based on their main themes, including (i) *Feature correlation analysis* -this theme aims to identify features that are correlated across different omics layers and/or co-vary under the conditions of interest. These correlated features provide more detailed delineations of underlying biological processes than those obtained from a single omics layer; (ii) *Sample clustering analysis* -this theme aims to leverage multiple molecular profiles to improve sample characterization, such as to identify subsets of cancer patients for more targeted treatments (156); (iii) *Understanding global structure* -this theme aims to gain a high-level overview of multi-omics data by extracting and examining their shared structural variations and local patterns. Compared to the knowledge-driven strategy where many user-friendly tools are available, most data-driven methods are in the form of complex multivariate statistics or machine learning algorithms, available mainly in the form of command line programs (78,83,157-159). For most researchers, they are harder to use, and the results are harder to interpret. User-friendly bioinformatics tools supporting data-driven strategy are urgently needed to help convert the complex multi-omics data into meaningful patterns and insights.

Here, we introduce OmicsAnalyst, a web-based visual analytics platform dedicated for data-driven multi-omics integration. It currently supports more than a dozen well-established methods through three visual analytics tracks - correlation network analysis, cluster heatmap analysis, and dimension reduction analysis. These three visualization tracks are equipped with comprehensive functions and menus to allow users to perform parameter customization, visual exploration and interactive analysis. To help users navigate the tool, we have compiled a comprehensive list of frequently asked questions (FAQs), four different screenshot tutorials, and a case study. The main features of OmicsAnalyst are described below.

4.2 Overview of OmicsAnalyst

The workflow of OmicsAnalyst is shown in Figure 4.1. It consists of three main phases to help users to navigate the complex procedures of multi-omics analysis. In the Phase 1 (data processing), users go through the conventional single omics data analysis workflow including data upload, annotation, missing value estimation, data filtering, and identification of significant features. After basic quality check and optional data normalization for multi-omics integration, users enter the Phase 2 (method selection). OmicsAnalyst offers a wide array of approaches organized under three categories: correlation network analysis, cluster heatmap analysis, and dimension reduction analysis. After method selection, users are presented with an overview and diagnostic plots to decide whether the default parameters (if any) should be updated. Finally, users enter the Phase 3 (visual analytics) and explore the results through interactive visualization coupled with various statistical and functional analysis.

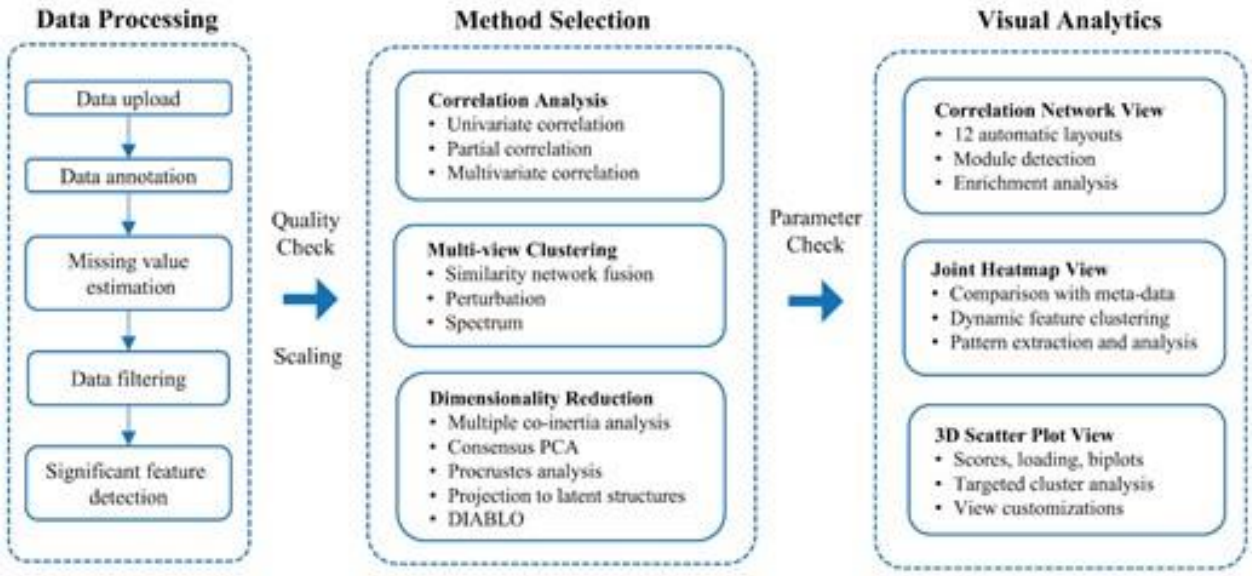


Figure 4.1 Overall workflow of OmicsAnalyst. Multi-omics integration is divided into three main phases - data processing, method selection and visual analytics. Each phase contains multiple steps and options to allow comprehensive analysis and customization.

4.2.1 Data processing

Data upload and annotation

OmicsAnalyst accepts data tables containing feature abundance values (raw or normalized) generated from different omics platforms. They must share the same sample names and metadata information. For data from human and mouse, users can further perform feature annotation for transcriptomics, proteomics, metabolomics and miRNA. The annotation is required for enrichment analysis in the visual analytics stage. *Missing value estimation*. Omics data often contain missing values which could cause potential issues in downstream analysis. Users can exclude features with too many missing values or perform missing value estimation based on several widely used methods. *Data filtering*. Given the high-dimensional nature of omics data, it is strongly recommended to perform unspecific data filtering to exclude features that are unlikely to be useful in downstream analysis. In particular, features that are relatively consistent can be safely excluded based on their inter-quantile ranges (IQRs) or other variance measures. Features that are of very low abundance should also be excluded, as they contribute little to the overall variance-covariance structure in multi-omics integration. *Differential analysis*. Users can perform conventional statistical comparisons to identify significant features within individual omics data. These features will be available for correlation network creation or highlighted in heatmaps or scatter plots. *Quality checking and normalization/scaling*. The goal is to make different omics data more 'integrable' by sharing similar distributions. Users can visually examine the distribution of individual omics data through density plot, principal component analysis (PCA) plot, and t-distributed stochastic neighbor embedding (t-SNE) plot. Based on the visual assessment, users can choose among a variety of data transformation, centering and scaling options to improve the integrability.

4.2.2 Correlation network analysis track

The objective of the correlation network analysis is to identify and visualize relationships between key features from two omics datasets. It consists of three main steps, detailed below.

Network creation

This step involves selecting the key features and computing their pairwise correlations. By default, significant features identified by differential analysis during the data processing phase will be used for network creation. However, users can also select top features based on the loading scores from the multivariate dimension reduction methods. Details on the dimension reduction techniques can be found in the ‘Dimension Reduction Analysis Track’ section. The next step is to compute pairwise similarities between selected features. Due to their simplicity and widespread familiarity, univariate methods, such as Pearson correlation, are usually computed as a first line of analysis. However, these methods can produce many false connections due to presence of highly collinear features in omics data. Partial correlation, a multivariate method that measures the correlation between two variables while controlling for all others, has been successfully applied to omics data to detect connections between features that are more likely to represent true dependencies (160).

Network customization

Networks with a large number of nodes and edges are too complex and overwhelming for visualization and interpretation. OmicsAnalyst partially addresses this issue by allowing users to control network sizes based on the strengths of correlations. However, applying a single threshold can often produce networks with the majority of edges existing between nodes of the same omics type. This is because in many cases, correlations between features of the same omics type are categorically higher than those of different omics types, likely due to technical differences between platforms. To address this issue, OmicsAnalyst offers two filters to control correlation strengths, one for within-omics and the other for between-omics, with a more stringent default threshold for the former. In addition, users can also apply degree or betweenness filters to control network size based purely on the topological properties of the nodes.

Network visual analytics

In addition to providing different filters to allow users to refine the nodes and edges that comprise the network, OmicsAnalyst offers a variety of simple and advanced functions to facilitate visual identification of important network structures. For instance, binary edge coloring is used to differentiate positive and negative correlations, and edge thickness is used to reflect strengths of the correlation to enable quick identification of feature pairs that are highly correlated. OmicsAnalyst also offers 3D network visualization for a deeper perspective of the relationships. Advanced graph layout algorithms, for example edge bundling can be applied to aggregate similar edges into groups to reduce clutter in visualization. Other features such as the concentric circular layout facilitate the evaluation of focal nodes and hierarchical relationships within network. When features are annotated during data processing, users can perform enrichment analysis on a group of nodes selected either manually or through automatic module detection algorithms.

4.2.3 Cluster heatmap analysis track

The objective of the track is to identify and visually explore relationships between samples and key features in side-by-side heatmaps, each displaying data from one omics type. It consists of two main steps, detailed below.

Sample cluster detection

In multi-omics data, each omics type is a separate representation of the same samples, making it suitable for multi-view clustering (81). One main advantage of multi-view clustering is that it tends to reduce spurious correlations that are due to random noise or platform-specific technical artifacts, as it is highly unlikely that exact same erroneous effects are present across multiple datasets. OmicsAnalyst currently supports three multi-view clustering algorithms: spectral clustering (159), perturbation-based clustering (161) and similarity network fusion (83). The distinguishing features of these three methods are as follows. Spectral clustering makes use of eigenvalues derived from a similarity matrix to perform clustering based on fewer dimensions, which greatly increases the speed (162). OmicsAnalyst employs the *Spectrum* R package, which combines the advantages of spectral clustering with several other advanced techniques (159). Perturbation clustering assumes that reliable clusters are robust to small alterations to the data (81). OmicsAnalyst uses the perturbation clustering for data integration and disease subtyping (PINSplus) R package to support this approach (161). The similar network fusion (SNF) method involves fusing individual sample similarity matrices together using a rapid nearest neighbour approach (83). Since the associated *SNFtool* package does not support cluster detection, the spectral clustering is applied to the learned status matrix for this purpose.

Heatmap visual analytics

The results of clustering analysis can be intuitively explored via heatmaps, which use visual cues to show how samples are clustered and how feature abundances vary across samples. OmicsAnalyst implements an interactive joint-heatmap viewer where two different omics datasets can be visualized and analyzed simultaneously. The interactive visualization was implemented based on the INVEX heatmap viewer (59). It is organized into two main views consisting of an overview and a focus view for each omics data. The overview heatmap displays the overall abundance patterns for all features. Users can click-and-drag to select a region of interest to be displayed in the focus view for a more detailed inspection. The annotation bars along the top indicate the original group memberships as well as the cluster memberships based on the selected multi-view clustering algorithm. Similar to the correlation network analysis, users can perform enrichment analysis on the features displayed in the focus view for each omics type, when features are annotated during data processing.

4.2.4 Dimension reduction analysis track

The objective of this track is to perform dimension reduction, and then visually explore corresponding scores, loadings and biplots in interactive 3D scatter plots to understand high-level trends and associated key features. It consists of two main steps, detailed below.

Multi-omics dimension reduction

Many standard multivariate dimension reduction techniques do not perform well on multi-omics datasets, which typically have many more features than observations (*p* >> *n*) and a multicollinear structure. Multivariate regression, the foundation of many multivariate dimension reduction techniques, performs poorly in these cases and so special care has been taken to develop more robust techniques for multi-omics data integration (163-165). OmicsAnalyst provides five different methods including multiple co-inertia analysis (MCIA), consensus PCA (CPCA), projection to latent structures (PLS), Procrustes analysis, and data integration analysis for biomarker discovery using latent components (DIABLO) (76,78,157,166). In general, these algorithms aim to identify sets of components that capture maximum variance within individual datasets and maximum association across datasets. They can be distinguished by individual optimization and constraint criteria used to identify component sets across the omics datasets. More detailed information and comparisons on these methods are provided in our FAQs under the 'Dimension Reduction Analysis' tab.

Visual analytics based on 3D scatter plots

OmicsAnalyst offers an interactive 3D scatter plot viewer that can display sample space (score plot), feature space (loading plot), as well as a 'merged' space (biplot) that overlays sample and feature spaces in the same plot to showcase contributions of key feature to the overall patterns. The 3D scatter plot viewer is divided into four different sections. The left panel contains a top section ('Settings') for controlling the overall visual environment of the scatter plots. The middle section ('Overall Pattern') allows users to change the grouping of nodes based on different metadata or clustering analysis. It offers extensive options such as colors, shapes, and highlighting effects for group visualization. The bottom section displays information related to the current selections. The main scatter plot viewer in the center displays the current view - score plot, loading plot or biplot which allow users to specify features of interest to be shown as arrows on top of sample space. Users can also overlay different metadata groups as ellipsoids on top of the feature space. The right panel is divided into top ('Comparison Test') and bottom ('Enrichment Analysis') sections to allow users to perform targeted statistical and functional analysis on the current selected groups or clusters, respectively. Click a row of the result tables, the corresponding feature(s) will be displayed as arrows in the current score plot.

4.3 Case study: multi-omics analysis of human pregnancy

To facilitate users to explore different features of OmicsAnalyst, three example multi-omics datasets have been provided including one from the Cancer Genome Atlas (TCGA, https://www.cancer.gov/tcga), one from the STATegra (167) and one from a recent multi-omics study on human pregnancy (168). Here, we provide a case study using the proteomics and metabolomics datasets from the pregnancy study.

Various physiological systems are known to change predictably throughout pregnancy (169). This study was conducted to collect comprehensive molecular data (repeated samples from the first three trimesters and 6 weeks postpartum for baseline levels; *n* = 17 women) to build a predictive model for gestational age (168). Here, we re-analyze the proteomics and metabolomics data sets as a case study. Differential analysis was performed using ANOVA/t-tests with thresholds chosen to give ~30% significant features ($|\log_2 FC| > 1$; adjusted *P*-value < 0.005), and datasets were auto-scaled before integration. All three visual analytics tracks were used to gain complementary perspectives of the data. First, we used the 'Free Exploration' mode of the Heatmap Visual Analytics track to understand patterns present in individual omics. While the baseline samples form a weak cluster, samples from the three trimesters are very mixed. Next, we computed the multidimensional components that best separated the sample groups using DIABLO and explored the results with 3D scatter plots. The global structure confirms what we expect, with baseline samples well distinguished from those collected during pregnancy, and samples collected during later trimesters located further away from the baseline (Figure 4.2A). The biplot overlays the sample space with the top features that most contribute to the separation (Figure 4.2B), in this case highlighting several proteins and metabolites that are consistent with the biology of pregnancy. Three out of the five top metabolites are associated with hormones that are elevated during pregnancy (thyroxine, pregnanediol-3-glucuronide, and cortisol). One of the top proteins (ADAM12) is a serum marker for pregnancy, two (GDF15 and GPC3) are encoded by genes that have high expression in placenta relative to other tissues, and one is angiotensin (AGT), a hormone known to be elevated during pregnancy (170). All feature arrows point in the same direction, except for the DL-2-aminooctinoic acid metabolite. Finally, we used correlation networks to visualize relationships between key features from the top three DIABLO components. The network has a central cluster of proteins that are positively correlated with the proteins and metabolites on the left, and negatively correlated with the metabolites on the right (Figure 4.2C). Inspecting several individual features shows that the structure is consistent with Figure 4.2B: the central proteins and positively correlated metabolites contain many of the previously highlighted biplot features (ADAM12, Cortisol, Sunitinib, and Pregnanediol-3-glucuronide) while one of the negatively correlated metabolites is DL-2-aminooctinoic acid, the lone biplot feature that pointed in the opposite direction. Network module analysis with the 'WalkTrap' algorithm resulted in three modules, all of which contained both proteins and metabolites (Figure 4.2D). The blue module was statistically significant, and enrichment analysis revealed that it is significantly enriched for the Reactome pathway 'Regulation of Insulin-like Growth Factor (IGF) transport and uptake by

Insulin-like Growth Factor Binding Proteins (IGFBPs)’. IGF is known to be elevated during pregnancy (171). This case study has illustrated the improved insights and rich biological context when multi-omics data and visual analytics are used together. More details and figures from the case study are available from the ‘Tutorial’ page (under the ‘Case Study’ tab) of OmicsAnalyst.

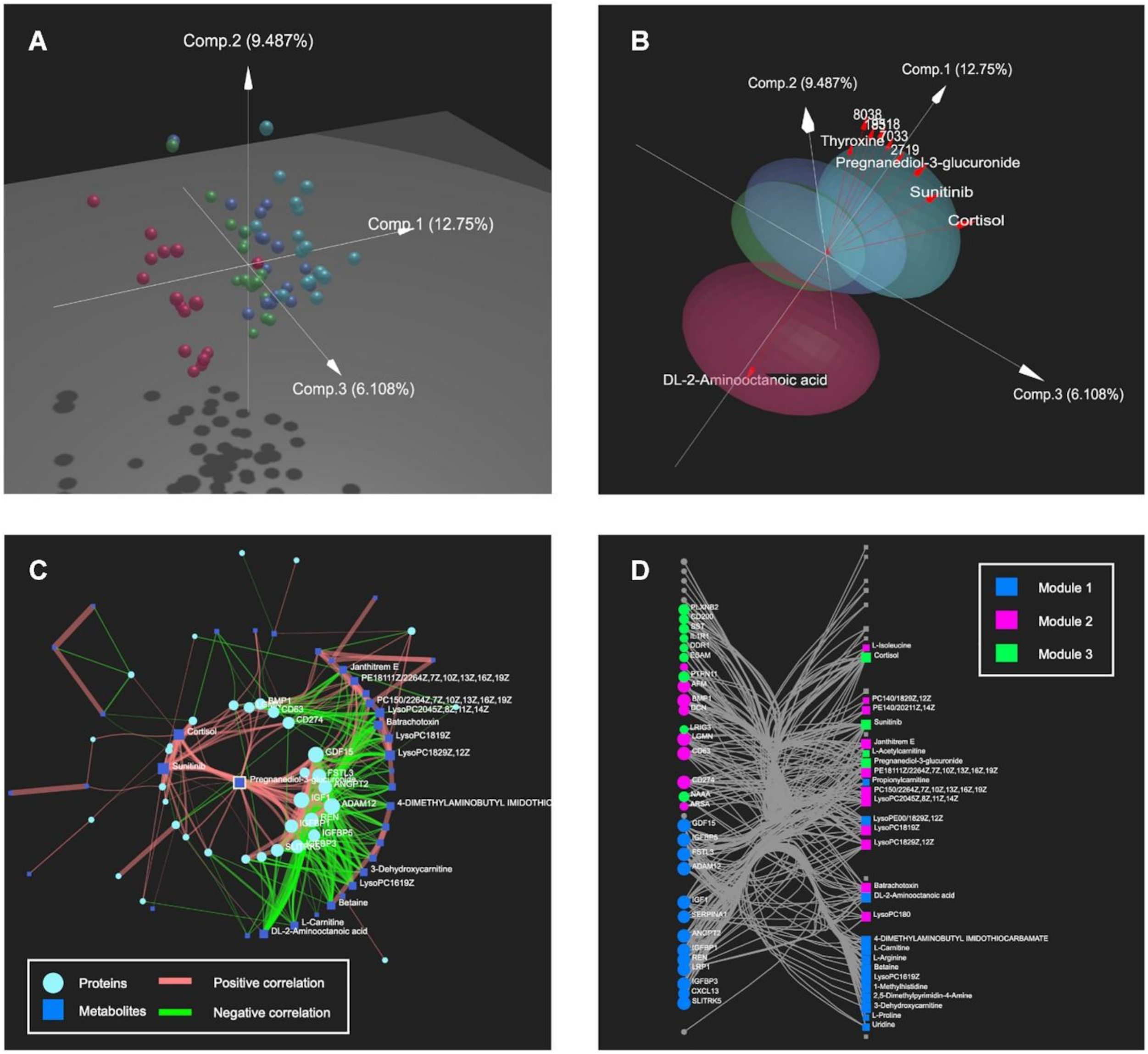


Figure 4.2 Example outputs from the case study. Dimension reduction was performed with DIABLO and results visualized with (A) 3D scatter plot of score plot, and (B) 3D biplot with elliptical summaries of sample groups (red = baseline, green = first trimester, dark blue = second trimester, light blue = third trimester) and the contributions of top five differentially expressed proteins and metabolites (red arrows). Correlation networks of features selected from the top three DIABLO components in (C) concentric circular layout, and (D) linear bipartite/tripartite layout, with modules detected by the 'WalkTrap' algorithm.

4.4 Implementation

OmicsAnalyst was implemented based on JavaServer Faces (JSF) using the PrimeFaces (v10.0) library (http://primefaces.org/) and R (version 4.0.2). The visual analytics methods have been developed based on several JavaScript libraries including sigma.js (http://sigmajs.org) for 2D network visualization, and three.js (https://threejs.org) for 3D network and scatter plot visualization. The system is hosted on a Google Cloud *n1-highmem-8* instance (64 GB RAM and eight virtual CPUs with 2.6 GHz each).

Comparison with other web-based tools

Table 4.1 shows the comparisons between OmicsAnalyst and three other web-based tools dedicated for multi-omics integration and analysis, including 3Omics (37), MiBiOmics (172) and OmicsNet (154). The 3Omics supports analysis of transcriptomics, proteomics and metabolomics data from human. It includes modules for correlation analysis, co-expression profiling, phenotype mapping and functional enrichment analysis. MiBiOmics tackles multi-omics integration through correlation analysis using WGCNA-based approach and dimension reduction analysis using MCIA and Procrustes analysis. Finally, OmicsNet uses *a priori* interaction information to construct multi-omics networks for genes, proteins, metabolites, miRNA, and transcription factors. The resulting network is interactively visualized in 3D space. OmicsAnalyst distinguishes itself by bringing together multivariate, data-driven feature selection and integration with innovative visual analytics for unbiased exploration and interrogation of complex multi-omics datasets.

Table 4.1 Comparison of OmicsAnalyst with other web-based tools. Symbols used for feature evaluations with ‘ √ ’ for present, ‘-’ for absent and ‘+’ for a more quantitative assessment (more ‘+’ indicating better support). The URLs for each tool are given below.

	OmicsAnalyst	**3Omics**	**MiBiOmics**	**OmicsNet**
Input format	Matrix	List, matrix	Matrix	List
Data processing				
Annotation	+++	+++	-	+++
Filtering	+++	-	+	-
Normalization	+++	-	+	-
Scaling	+++	-	+	-
Differential expression	+++	-	-	-
Integration methods				
Univariate correlation	✓	✓	✓	-
Partial correlation	✓	-	-	-
Similarity network fusion	✓	-	-	-
Spectral clustering	✓	-	-	-
Perturbation-based clustering	✓	-	-	-
MCIA	✓	-	✓	-
CPCA	✓	-	-	-
Procrustes	✓	-	✓	-
PLS	✓	-	-	-
DIABLO	✓	-	-	-
Visual analytics				
Scatter plot	+++	-	+	-
Heatmap	+++	++	++	-
Network	+++	-	++	+++
Contextual enrichment analysis				
Metabolite sets	++	++	-	++
Gene sets	++	++	-	++
miRNA sets	++	-	-	-

- OmicsAnalyst: https://www.omicsanalyst.ca/
- 3Omics: https://3omics.cmdm.tw/
- MiBiOmics: https://shiny-bird.univ-nantes.fr/app/Mibiomics
- OmicsNet: https://www.omicsnet.ca/

4.5 Conclusion

The motivation for OmicsAnalyst was to create an intuitive, web-based platform for multi-omics integration that allows researchers to fuse statistical and visual streams of evidence together to make more informed judgements. In particular, we implemented three distinct visual analytics tracks - feature correlation analysis coupled with networks, sample clustering analysis coupled with heatmaps, and dimension reduction analysis coupled with 3D scatter plots. In doing so, OmicsAnalyst enables users to dissect large and complex multi-omics datasets by facilitating pattern recognition and cognitive reasoning through powerful yet intuitive visual analytics.

Chapter 5: Conclusion and Future Works

Omics data analysis has become essential in biomedical science for both basic research and clinical practices. Integrative analysis using data from multiple omics layers facilitates the understanding of interplays between biological entities to gain novel insights into the etiology and pathogenesis of complex diseases. However, bioinformatics tools for multi-omics data analysis are lacking and inaccessible to average researcher without programming knowledge. There is an urgent demand for development of new tools and methods coupled with easy-to-use interface to facilitate hypothesis generation and knowledge discovery from multi-omics data.

This thesis aims to develop bioinformatics tools to overcome some of the challenges associated with omics data analysis. Throughout my thesis, one of the main themes is to extend static results from existing methods with interactive visualization to enable iterative and conversational analytic processes. Three specific tasks were achieved in this thesis:

Aim 1 Developing web-based visual analytics platform to address functional profiling, network integration and meta-analysis of transcriptomics data.

Aim 2 Developing web-based platform for multi-omics network integration and 3D visualization.

Aim 3 Developing web-based visual analytics platform to enable data-driven integration of multi-omics datasets.

Chapter 2 presents the version 3.0 of NetworkAnalyst, a comprehensive web-based visual analytics platform designed to perform functional profiling of transcriptomics data. In the new update, I have expanded the molecular interaction beyond generic protein-protein interaction with the addition of gene regulatory network, protein-chemical and protein-drug interactions.

Additionally, I have collected protein-protein interaction data with more refined context such as tissue-specific and cell-specific PPI data. Another addition is enrichment network which allows users to quickly assess the relationships between enriched terms and corresponding genes. This update also added support for multi-list comparisons through heatmap, Venn diagram and chord diagram visual analytics. This addresses an unmet need for intuitive and flexible bioinformatics tools for performing meta-analysis of multiple gene lists.

Chapter 3 presents OmicsNet, a web-based visual analytics platform, which enables network-based multi-omics integration and 3D network visualization. This tool allows researchers to map their lists of genes, proteins, transcription factors, miRNA and metabolites within context of current knowledge framework of molecular interactions. OmicsNet also offers network visualization in 3D space to address the issue of hairball effect associated with large and complex networks. The network viewer seamlessly integrates visualization and functional analysis. OmicsNet also proposes a module-based analysis approach which aims to facilitate biological interpretation of complex networks. It aims to simplify the network into a series of graph modules and focus on interpreting the individual modules.

Chapter 4 proposes OmicsAnalyst, a web-based visual analytics platform dedicated for data-driven multi-omics integration. As a tool primarily designed for exploratory data analysis, it offers three analytics tracks allowing researchers to explore different facets of their multi-omics datasets via correlation network, heatmap clustering and dimension reduction analysis to facilitate hypothesis generation. To explore results from dimension reduction methods, OmicsAnalyst proposes an innovative 3D scatter plot viewer that enables users to effectively explore and analyze their datasets. The scatter plot viewer is supported by a set of functions that allows users to perform in-depth and targeted analysis of their multi-omics datasets.

More generally, the projects described in this thesis have addressed several important challenges in current omics data analysis:

- Most bioinformatics tools for omics data analysis and integration are command line based and require users to have some degree of programming knowledge. Additionally, they tend to be designed for more specific tasks and have limited outreach. This leads to a gap between bioinformatics analysis and average life science researchers. In this dissertation, all the projects are web-based platforms integrating multiple approaches otherwise harder to access, they aim to be intuitive and easily accessible by a larger audience, from bench researchers to clinicians. They offer user-friendly graphical user interface to lower the barrier of entry to omics data analysis.
- Data interpretation is currently the bottleneck in omics data analysis. In contrast to the "black-box" models seen in machine learning, my focus is to implement interpretable and intuitive data analysis process complemented with visual analytics. Throughout the different projects, I always attempt to create engaging analytical process with interactive visualization to help users make sense of their data using cognitive reasoning and domain knowledge.
- Reproducibility remains an important issue in life science. To alleviate this problem, NetworkAnalyst has implemented a project management system to allow users to save the progress of their analysis and resume later. Additionally, sharable links can be created on visual analytics page to foster collaborations and sharing results between researchers.
- "Hairball effect" plagues network visualization and reduces its effectiveness in exploratory data analysis. Throughout the thesis, several features were implemented to address this issue.

  - In NetworkAnalyst, the use of meta-nodes was added to simplify networks with hierarchical structure such as enrichment network. Meta-nodes can be expanded on node click events to show genes associated with functional terms in an enrichment network.
  - Edge bundling has been implemented to alleviate the visual occlusion issue associated with large quantities of edge crossings in both 2D and 3D networks. In edge bundling, similar edges are deformed and grouped together. The procedure also highlights connection patterns within the network.
  - Using different network layouts in 2D and 3D help understand different aspects of complex networks. For instance, concentric circular layout, first implemented in NetworkAnalyst, enables rapid assessment of how one focal node relates to the rest of the network in terms of shortest path lengths. In OmicsNet, multi-layered layout in 3D facilitates the visualization of multi-omics networks by layering nodes from each omics separately.
- Network visualization in 3D space is not well supported in the web. Visualizing networks in 3D space provides additional visual perspectives and layout space. This may facilitate the interpretation of biological networks depending on the user preferences and/or the layout used. For instance, layered layout excels in displaying networks with hierarchical structure such as multi-omics network. I believe it is beneficial to provide such option for researchers. OmicsNet facilitates access to 3D-based network visual analytics of biological networks.
- In dimension reduction analysis, linking features to sample separation requires the use of biplots. Additionally, these biplots can be used to assess how features correlate with each

other. OmicsAnalyst has implemented interactive and intuitive biplot functionalities in its 3D-based scatter plot visual analytics.

There remains much room for future research in omics data analysis. In the following section, I list some future directions for the tools described in this thesis.

- *Time series analysis.* Current omics studies mainly involve snapshot omics datasets. This approach has significant limitations in their scope due to different time scales of change associated with different omics layers. Although single data point omics can be used to identify correlations between omics layers at steady state, it cannot be used for studying causative and mechanistic aspects of the system. For instance, when integrating transcriptomics and metabolomics datasets, it is known that changes in metabolic concentrations will occur after transcriptional changes. Therefore, capturing time course of samples from each omics type is better suited to study the dynamic behavior and regulatory relationships of cellular constituents.

  In time series approach, data points are collected successively at a pace that is usually equally spaced in time between each of them. Classical statistical approaches are often unsuitable for the analysis of time series data due to some specific characteristics to time series data including low number of replicates, importance of sample synchronization and uneven sampling due to missing data (173). With the advent of high throughput and single-cell sequencing technologies, time series studies will become more and more prevalent. There is a high demand for bioinformatics tools to explore and analyze time series data, especially in the context of joint analysis of multiple omics layers (174,175). Appropriate visual analytics will greatly help in the tasks of exploring and understanding this type of data.

- *Network analysis*. Rarely does a biomolecule act alone and most biological processes are mediated by cellular components interacting with other cellular components. Network is a useful framework to understand the molecular context of biological molecules and it is the key to understand how they contribute to phenotypes (176). Although biological networks can be come extremely complex, graph theory comes in handy in deriving information from them. In this thesis, there is a need to complement the visual analytics with more advanced network mining algorithms.

The first challenge is to address the problem of visualizing large network (more than ten thousand nodes). The approaches proposed throughout this thesis mainly center on trimming network size, but it can cause bias and may not be desirable in cases where analysis needs to be performed on the whole interactome. Additional efforts need to be spent on implementing visual analytics solutions able to handle interactive visualization of large networks and proposing effective visualization techniques to interpret the data. There exists multiple stand-alone applications supporting up to millions of nodes and edges (128,177) but web-based implementation presents inherent challenges due to limited access to the computing resource of the operating system (i.e. max of 2G memory for most laptop browsers).

This thesis mainly focuses on active subnetwork approach where the objective is to identify a subnetwork enriched with biomolecules of interest, which is key to understand the molecular basis of diseases or phenotypes. There are many other approaches that can be used to investigate their biological roles and molecular context. For instance, diffusion-based approaches aim to identify pathways closest to these biomolecules. This type of algorithm uses random walks starting from biomolecules of interest and diffuse along the edges to visit the rest of the network with equal probability. The closest nodes and edges will be visited more often by the random walkers. This is useful in prioritizing nodes and interactions based on network proximity to biomolecules of interest (68).

- *Additional omics data types and clinical data.* The tools described in this thesis are restricted to transcriptomics, miRNA, proteomics, metabolomics, and microbiome data. The next step is to extend support to single nucleotide polymorphism and epigenomics data to enable a more complete genotype-phenotype integrative approach. Additionally, with the rise of precision medicine, clinical data has become another layer to be considered. Integrating phenotypic and clinical information or metadata in general with omics data remains an active research topic (178,179). There is a clear need for bioinformatics tools that can address omics and non-omics (such as phenomics data)integration.

  In addition to data complexity and heterogeneity that also plague omics data, non-omics data especially suffers from lack of uniformity and standardization: a same descriptor can take form of both quantitative and qualitative variables to characterize the same samples depending on the study. The subjective nature of such data can also severely limit their integration with omics data. Moreover, the relationship between omics and non-omics data can also be problematic. Ascertainment bias in case-control studies can also result in unwanted correlations between omics and non-omics data. For instance, DNA methylation measurements may be associated with age and gender (180). Additionally, extra care needs to be taken to address collinearity and redundancy between metadata themselves to avoid statistical missteps.

- *Support virtual reality (VR) based visual analytics.* Shifting from standard computer screens to virtual reality will provide an alternative environment for data analysis that may be more intuitive, immersive, and interactive for users. Current visual analytics mainly consists of multiple isolated "point-and-click" interfaces that leads to a fragmented analytic experience. In recent years, extending visual analytics to VR-related technology, also referred as immersive analytics, has become an active research area. In a paper by Chandler et al., the authors defined the term immersive analytics as investigating how new interaction and display technologies can be used to support analytical reasoning and decision-making (181). VR promotes an integrated and linked multi-view environment with a unifying graphical interface. Most importantly, VR provides stereoscopic visualization to create immersive environment (182). Furthermore, VR can support multimodal interaction interface in which multiple senses are involved including haptic input and olfactory feedback.

  Recent progress in VR and graphics processing technologies have significantly lowered the entry barrier to virtual reality access and increased its market share. The global market of VR has increased by 42.2% from US$1.03 billion to US$1.47 billion in 2020 (183). Additionally, WebVR technology and frameworks such as A-frame and Unity has greatly facilitated the development of browser-based VR experience (184-186). The 3D-based visual analytics systems described in the thesis translate well in VR environment. In NetworkAnalyst, I have implemented a prototype version of VR-based network visualization.

- *Artificial intelligence and machine learning.* With the age of big data, artificial intelligence (AI) and machine learning (ML) are key to address the complexity of omics data. ML algorithms excel in predictive modeling and, more and more, they are being used to understand the underlying mechanisms of biological processes (187). With the rapid advancement in computational speed, capacity, and software programming, it is a matter of time that ML and AI will supplement traditional regression-based methods in analyzing health data. Additionally, there has been significant progress in standardized and open source implementations of ML algorithms from software packages such as scikit-learn (188), Weka (189), or TensorFlow, leading to much more accessibility for researchers.

In life and biomedical science, the adoption of ML has been mainly limited due to the lack of understanding of the algorithms and the black-box nature of predictive models (190). Indeed, compared to other fields, the rate of adoption is low. There remain many issues to be addressed. Foremost, there is the issue of generalizability due to study-specific technical bias or other confounding factors that make ML findings unreliable (191). Another key challenge in machine learning is the selection of the right algorithm for the problem. To address some of the challenges, coupling machine learning with interactive visualization can facilitate both algorithm selection and interpretation of different machine learning models (192). Visual analytics can also help with building machine learning models to complement conventional automated algorithms or AI-driven platforms for such tasks (193). Complementing machine learning with visual analytics will be the next state-of-art framework for omics data analysis.

Multi-omics data analysis is still an emerging research field. It is recommended to use multiple integrative methods and visualizations rather than relying on single method. My work aims to provide intuitive and easy-to-use bioinformatics platforms for researchers as new options to explore and analyze their datasets. Well thought-out tool suites enable more complex analysis normally restricted to experienced bioinformaticians, leading to democratization of omics data analysis and integration. The future directions mentioned above can improve the current tools leading to more insights in multi-omics studies and, in my opinion, represent the next stage of multi-omics data analysis.

Bibliography

1. Yugi, K., Kubota, H., Hatano, A. and Kuroda, S.J.T.i.b. (2016) Trans-omics: how to reconstruct biochemical networks across multiple 'omic'layers. *Trends in biotechnology*, **34**, 276-290.

2. O'Driscoll, A., Daugelaite, J. and Sleator, R.D.J.J.o.b.i. (2013) 'Big data', Hadoop and cloud computing in genomics. *Journal of biomedical informatics*, **46**, 774-781.

3. Ideker, T., Galitski, T., Hood, L.J.A.r.o.g. and genetics, h. (2001) A new approach to decoding life: systems biology. *Annual review of genomics and human genetics*, **2**, 343-372.

4. Tomczak, K., Czerwińska, P. and Wiznerowicz, M.J.C.o. (2015) The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary oncology*, **19**, A68.

5. Perez-Riverol, Y., Bai, M., da Veiga Leprevost, F., Squizzato, S., Park, Y.M., Haug, K., Carroll, A.J., Spalding, D., Paschall, J. and Wang, M.J.N.b. (2017) Discovering and linking public omics data sets using the Omics Discovery Index. *Nature biotechnology*, **35**, 406-409.

6. Alyass, A., Turcotte, M. and Meyre, D.J.B.m.g. (2015) From big data analysis to personalized medicine for all: challenges and opportunities. *BMC medical genomics*, **8**, 1-12.

7. Begley, C.G. and Ioannidis, J.P.J.C.r. (2015) Reproducibility in science: improving the standard for basic and preclinical research. *Circulation research*, **116**, 116-126.

8. Benjamin, D.J., Berger, J.O., Johannesson, M., Nosek, B.A., Wagenmakers, E.-J., Berk, R., Bollen, K.A., Brembs, B., Brown, L. and Camerer, C.J.N.h.b. (2018) Redefine statistical significance. *Nature human behaviour*, **2**, 6-10.

9. Bumgarner, R.J.C.p.i.m.b. (2013) Overview of DNA microarrays: types, applications, and their future. *Current protocols in molecular biology*, **101**, 22.21. 21-22.21. 11.

10. Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-Liggett, C.M., Knight, R. and Gordon, J.I.J.N. (2007) The human microbiome project. *Nature*, **449**, 804-810.

11. Chin, L., Andersen, J.N. and Futreal, P.A.J.N.m. (2011) Cancer genomics: from discovery science to personalized medicine. *Nature medicine*, **17**, 297.

12. Ramasamy, A., Mondry, A., Holmes, C.C. and Altman, D.G.J.P.m. (2008) Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS medicine*, **5**, e184.

13. Hasin, Y., Seldin, M. and Lusis, A.J.G.b. (2017) Multi-omics approaches to disease. *Genome biology*, **18**, 1-15.

14. Evangelou, E. and Ioannidis, J.P.J.N.R.G. (2013) Meta-analysis methods for genome-wide association studies and beyond. *Nature Reviews Genetics*, **14**, 379-389.

15. Ntzani, E.E. and Ioannidis, J.P.J.T.L. (2003) Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. *The Lancet*, **362**, 1439-1444.

16. Ein-Dor, L., Kela, I., Getz, G., Givol, D. and Domany, E.J.B. (2005) Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, **21**, 171-178.

17. Zhou, G., Stevenson, M.M., Geary, T.G. and Xia, J.J.P.n.t.d. (2016) Comprehensive transcriptome meta-analysis to characterize host immune responses in helminth infections. *PLoS neglected tropical diseases*, **10**, e0004624.

18. Khan, M.M., Ernst, O., Manes, N.P., Oyler, B.L., Fraser, I.D., Goodlett, D.R. and Nita-Lazar, A.J.A.i.d. (2019) Multi-omics strategies uncover host–pathogen interactions. *ACS infectious diseases*, **5**, 493-505.

19. Yan, J., Risacher, S.L., Shen, L. and Saykin, A.J.J.B.i.b. (2018) Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. *Briefings in bioinformatics*, **19**, 1370-1381.

20. Delhalle, S., Bode, S.F., Balling, R., Ollert, M., He, F.Q.J.N.s.b. and applications. (2018) A roadmap towards personalized immunology. *NPJ systems biology and applications*, **4**, 1-14.

21. Chong, J. and Xia, J.J.M. (2017) Computational approaches for integrative analysis of the metabolome and microbiome. *Metabolites*, **7**, 62.

22. Kuan, V., Fraser, H.C., Hingorani, M., Denaxas, S., Gonzalez-Izquierdo, A., Direk, K., Nitsch, D., Mathur, R., Parisinos, C.A. and Lumbers, R.T.J.S.r. (2021) Data-driven identification of ageing-related diseases from electronic health records. *Scientific reports*, **11**, 1-17.

23. Rai, A. (2020) Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, **48**, 137-141.

24. Ribeiro, M.T., Singh, S. and Guestrin, C. (2016), *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135-1144.

25. Jeong, H., Mason, S.P., Barabási, A.-L. and Oltvai, Z.N.J.N. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41-42.

26. Gysi, D.M., Do Valle, Í., Zitnik, M., Ameli, A., Gan, X., Varol, O., Ghiassian, S.D., Patten, J., Davey, R.A. and Loscalzo, J.J.P.o.t.N.A.o.S. (2021) Network medicine framework for identifying drug-repurposing opportunities for COVID-19. *Proceedings of the National Academy of Sciences*, **118**.

27. Hartwell, L.H., Hopfield, J.J., Leibler, S. and Murray, A.W.J.N. (1999) From molecular to modular cell biology. *Nature*, **402**, C47-C52.

28. Zhou, G., Li, S. and Xia, J. (2020), *Computational Methods and Data Analysis for Metabolomics*. Springer, pp. 469-487.

29. Thomas, J.J. (2005) *Illuminating the path: The research and development agenda for visual analytics*. IEEE Computer Society.

30. Keim, D., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J. and Melançon, G. (2008), *Information visualization*. Springer, pp. 154-175.

31. O'Donoghue, S.I. (2021) Grand Challenges in Bioinformatics Data Visualization. *Frontiers in Bioinformatics*, **1**, 13.

32. Parisi, T. (2012) *WebGL: up and running*. O'Reilly Media, Inc.

33. Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y. and Gentry, J.J.G.b. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, **5**, 1-16.

34. Pang, Z., Chong, J., Zhou, G., de Lima Morais, D.A., Chang, L., Barrette, M., Gauthier, C., Jacques, P.-É., Li, S. and Xia, J.J.N.a.r. (2021) MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights. *Nucleic acids research*.

35. Ouyang, Y., Yin, J., Wang, W., Shi, H., Shi, Y., Xu, B., Qiao, L., Feng, Y., Pang, L. and Wei, F.J.C.I.D. (2020) Downregulated Gene Expression Spectrum and Immune Responses Changed During the Disease Progression in Patients With COVID-19. *Clinical Infectious Diseases*, **71**, 2052-2060.

36. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T.J.G.r. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, **13**, 2498-2504.

37. Kuo, T.-C., Tian, T.-F. and Tseng, Y.J.J.B.s.b. (2013) 3Omics: a web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data. *BMC systems biology*, **7**, 1-15.

38. Breuer, K., Foroushani, A.K., Laird, M.R., Chen, C., Sribnaia, A., Lo, R., Winsor, G.L., Hancock, R.E., Brinkman, F.S. and Lynn, D.J.J.N.a.r. (2013) InnateDB: systems biology of innate immunity and beyond—recent updates and continuing curation. *Nucleic acids research*, **41**, D1228-D1233.

39. Barsky, A., Gardy, J.L., Hancock, R.E. and Munzner, T.J.B. (2007) Cerebral: a Cytoscape plugin for layout of and interaction with biological networks using subcellular localization annotation. *Bioinformatics*, **23**, 1040-1042.

40. Pavlopoulos, G.A., O'Donoghue, S.I., Satagopam, V.P., Soldatos, T.G., Pafilis, E. and Schneider, R.J.B.s.b. (2008) Arena3D: visualization of biological networks in 3D. *BMC systems biology*, **2**, 1-7.

41. Krzywinski, M., Birol, I., Jones, S.J. and Marra, M.A.J.B.i.b. (2012) Hive plots—rational approach to visualizing networks. *Briefings in bioinformatics*, **13**, 627-644.

42. Raredon, M.S.B., Adams, T.S., Suhail, Y., Schupp, J.C., Poli, S., Neumark, N., Leiby, K.L., Greaney, A.M., Yuan, Y. and Horien, C.J.S.a. (2019) Single-cell connectomic analysis of adult mammalian lungs. *Science advances*, **5**, eaaw3851.

43. Meng, C., Zeleznik, O.A., Thallinger, G.G., Kuster, B., Gholami, A.M. and Culhane, A.C.J.B.i.b. (2016) Dimension reduction techniques for the integrative analysis of multi-omics data. *Briefings in bioinformatics*, **17**, 628-641.

44. Hotelling, H.J.J.o.e.p. (1933) Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, **24**, 417.

45. Kruskal, J.B.J.P. (1964) Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, **29**, 1-27.

46. Ringnér, M.J.N.b. (2008) What is principal component analysis? *Nature biotechnology*, **26**, 303-304.

47. Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I.W., Ng, L.G., Ginhoux, F. and Newell, E.W.J.N.b. (2019) Dimensionality reduction for visualizing single-cell data using UMAP. *Nature biotechnology*, **37**, 38-44.

48. Van der Maaten, L. and Hinton, G.J.J.o.m.l.r. (2008) Visualizing data using t-SNE. *Journal of machine learning research*, **9**.

49. Alter, O., Brown, P.O. and Botstein, D.J.P.o.t.N.A.o.S. (2003) Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proceedings of the National Academy of Sciences*, **100**, 3351-3356.

50. Dray, S., Chessel, D. and Thioulouse, J.J.E. (2003) Co‐inertia analysis and the linking of ecological data tables. *Ecology*, **84**, 3078-3089.

51. Gower, J.C.J.P. (1975) Generalized procrustes analysis. *Psychometrika*, **40**, 33-51.

52. Ter Braak, C.J.J.E. (1986) Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, **67**, 1167-1179.

53. Lê Cao, K.-A., Martin, P.G., Robert-Granié, C. and Besse, P.J.B.b. (2009) Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC bioinformatics*, **10**, 1-17.

54. Mandal, A. and Maji, P.J.I.t.o.c. (2017) FaRoC: fast and robust supervised canonical correlation analysis for multimodal omics data. *IEEE transactions on cybernetics*, **48**, 1229-1241.

55. Medina, I., Carbonell, J., Pulido, L., Madeira, S.C., Goetz, S., Conesa, A., Tï¿ ½rraga, J.n., Pascual-Montano, A., Nogales-Cadenas, R. and Santoyo, J.J.N.a.r. (2010) Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. *Nucleic acids research*, **38**, W210-W213.

56. Akhmedov, M., Martinelli, A., Geiger, R., Kwee, I.J.N.g. and bioinformatics. (2020) Omics playground: a comprehensive self-service platform for visualization, analytics and exploration of big omics data. *NAR genomics and bioinformatics*, **2**, lqz019.

57. Gardeux, V., David, F.P., Shajkofci, A., Schwalie, P.C. and Deplancke, B.J.B. (2017) ASAP: a web-based platform for the analysis and interactive visualization of single-cell RNA-seq data. *Bioinformatics*, **33**, 3123-3125.

58. Perez-Llamas, C. and Lopez-Bigas, N.J.P.o. (2011) Gitools: analysis and visualisation of genomic data using interactive heat-maps. *PloS one*, **6**, e19541.

59. Xia, J., Lyle, N.H., Mayer, M.L., Pena, O.M. and Hancock, R.E.J.B. (2013) INVEX—a web-based tool for integrative visualization of expression data. *Bioinformatics*, **29**, 3232-3234.

60. Alonso, R., Salavert, F., Garcia-Garcia, F., Carbonell-Caballero, J., Bleda, M., Garcia-Alonso, L., Sanchis-Juan, A., Perez-Gil, D., Marin-Garcia, P. and Sanchez, R.J.N.a.r. (2015) Babelomics 5.0: functional interpretation for new generations of genomic data. *Nucleic acids research*, **43**, W117-W121.

61. Xia, J., Benner, M.J. and Hancock, R.E.J.N.a.r. (2014) NetworkAnalyst-integrative approaches for protein–protein interaction network analysis and visual exploration. *Nucleic acids research*, **42**, W167-W174.

62. Xia, J., Gill, E.E. and Hancock, R.E.J.N.p. (2015) NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data. *Nature protocols*, **10**, 823-844.

63. Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M. and Lachmann, A.J.N.a.r. (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*, **44**, W90-W97.

64. Reimand, J., Arak, T., Adler, P., Kolberg, L., Reisberg, S., Peterson, H. and Vilo, J.J.N.a.r. (2016) g: Profiler—a web server for functional interpretation of gene lists (2016 update). *Nucleic acids research*, **44**, W83-W89.

65. Liao, Y., Wang, J., Jaehnig, E.J., Shi, Z. and Zhang, B.J.N.a.r. (2019) WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic acids research*, **47**, W199-W205.

66. Cavill, R., Sidhu, J.K., Kilarski, W., Javerzat, S., Hagedorn, M., Ebbels, T., MD, Bikfalvi, A. and Keun, H.C.J.J.o.p.r. (2010) A combined metabonomic and transcriptomic approach to investigate metabolism during development in the chick chorioallantoic membrane. *Journal of proteome research*, **9**, 3126-3134.

67. Fan, T.W., Bandura, L.L., Higashi, R.M. and Lane, A.N.J.M. (2005) Metabolomics-edited transcriptomics analysis of Se anticancer action in human lung cancer cells. *Metabolomics*, **1**, 325-339.

68. Barabási, A.-L., Gulbahce, N. and Loscalzo, J.J.N.r.g. (2011) Network medicine: a network-based approach to human disease. *Nature reviews genetics*, **12**, 56-68.

69. Song, L., Langfelder, P. and Horvath, S.J.B.b. (2012) Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC bioinformatics*, **13**, 1-21.

70. Fendt, S.M., Buescher, J.M., Rudroff, F., Picotti, P., Zamboni, N. and Sauer, U.J.M.s.b. (2010) Tradeoff between enzyme and metabolite efficiency maintains metabolic homeostasis upon perturbations in enzyme capacity. *Molecular systems biology*, **6**, 356.

71. Schäfer, J. and Strimmer, K.J.B. (2005) An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, **21**, 754-764.

72. Bradley, P.H., Brauer, M.J., Rabinowitz, J.D. and Troyanskaya, O.G.J.P.C.B. (2009) Coordinated concentration changes of transcripts and metabolites in Saccharomyces cerevisiae. *PLoS computational biology*, **5**, e1000270.

73. Wold, S., Esbensen, K., Geladi, P.J.C. and systems, i.l. (1987) Principal component analysis. *Chemometrics and intelligent laboratory systems* **2**, 37-52.

74. Lee, D.D. and Seung, H.S.J.N. (1999) Learning the parts of objects by non-negative matrix factorization. *Nature*, **401**, 788-791.

75. Johnson, A.J., Vangay, P., Al-Ghalith, G.A., Hillmann, B.M., Ward, T.L., Shields-Cutler, R.R., Kim, A.D., Shmagel, A.K., Syed, A.N., Students, P.M.C.J.C.h. *et al.* (2019) Daily sampling reveals personalized diet-microbiome associations in humans. *Cell host & microbe*, **25**, 789-802. e785.

76. Dixon, P.J.J.o.V.S. (2003) VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science*, **14**, 927-930.

77. Culhane, A.C., Perrière, G. and Higgins, D.G.J.B.b. (2003) Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. *BMC bioinformatics*, **4**, 1-15.

78. Meng, C., Kuster, B., Culhane, A.C. and Gholami, A.M.J.B.b. (2014) A multivariate approach to the integration of multi-omics datasets. *BMC bioinformatics*, **15**, 1-13.

79. Hanafi, M., Kohler, A., Qannari, E.-M.J.C. and systems, i.l. (2011) Connections between multiple co-inertia analysis and consensus principal component analysis. *Chemometrics and intelligent laboratory systems*, **106**, 37-40.

80. Shafi, A., Nguyen, T., Peyvandipour, A., Nguyen, H. and Draghici, S.J.F.i.g. (2019) A multi-cohort and multi-omics meta-analysis framework to identify network-based gene signatures. *Frontiers in genetics*, **10**, 159.

81. Rappoport, N. and Shamir, R.J.N.a.r. (2018) Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic acids research*, **46**, 10546-10562.

82. Shen, R., Olshen, A.B. and Ladanyi, M.J.B. (2009) Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, **25**, 2906-2912.

83. Wang, B., Mezlini, A.M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B. and Goldenberg, A.J.N.m. (2014) Similarity network fusion for aggregating data types on a genomic scale. *Nature methods*, **11**, 333-337.

84. Pearl, J. (2014) *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier.

85. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R. and Lander, E.S.J.P.o.t.N.A.o.S. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, **102**, 15545-15550.

86. Xia, J., Fjell, C.D., Mayer, M.L., Pena, O.M., Wishart, D.S. and Hancock, R.E.J.N.a.r. (2013) INMEX—a web-based tool for integrative meta-analysis of expression data. *Nucleic acids research*, **41**, W63-W70.

87. Kim, D., Langmead, B. and Salzberg, S.L.J.N.m. (2015) HISAT: a fast spliced aligner with low memory requirements. *Nature methods*, **12**, 357-360.

88. Bray, N.L., Pimentel, H., Melsted, P. and Pachter, L.J.N.b. (2016) Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology*, **34**, 525-527.

89. Law, C.W., Chen, Y., Shi, W. and Smyth, G.K.J.G.b. (2014) voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome biology*, **15**, 1-17.

90. Robinson, M.D., McCarthy, D.J. and Smyth, G.K.J.B. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139-140.

91. Love, M.I., Huber, W. and Anders, S.J.G.b. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, **15**, 1-21.

92. Zhang, B., Zhao, S. and Neuhaus, I.J.B. (2018) canvasDesigner: a versatile interactive high-resolution scientific multi-panel visualization toolkit. *Bioinformatics*, **34**, 3419-3420.

93. Sergushichev, A.A.J.B. (2016) An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *BioRxiv*, 060012.

94. Zyla, J., Marczyk, M., Weiner, J. and Polanska, J.J.B.b. (2017) Ranking metrics in gene set enrichment analysis: do they matter? *BMC bioinformatics*, **18**, 1-12.

95. Roy, S., Bhattacharyya, D.K. and Kalita, J.K.J.B.b. (2014) Reconstruction of gene co-expression network from microarray data using local expression patterns. *BMC bioinformatics*, **15**, 1-14.

96. Basha, O., Shpringer, R., Argov, C.M. and Yeger-Lotem, E.J.N.a.r. (2018) The DifferentialNet database of differential protein–protein interactions in human tissues. *Nucleic acids research*, **46**, D522-D526.

97. Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F. and Young, N.J.N.g. (2013) The genotype-tissue expression (GTEx) project. *Nature genetics*, **45**, 580-585.

98. Lee, S., Zhang, C., Arif, M., Liu, Z., Benfeitas, R., Bidkhori, G., Deshmukh, S., Al Shobky, M., Lovric, A. and Boren, J.J.N.a.r. (2018) TCSBN: a database of tissue and cancer specific biological networks. *Nucleic acids research*, **46**, D595-D600.

99. Vandenbon, A., Dinh, V.H., Mikami, N., Kitagawa, Y., Teraguchi, S., Ohkura, N. and Sakaguchi, S.J.P.o.t.N.A.o.S. (2016) Immuno-Navigator, a batch-corrected coexpression database, reveals cell type-specific gene networks in the immune system. *Proceedings of the National Academy of Sciences*, **113**, E2393-E2402.

100. Karagkouni, D., Paraskevopoulou, M.D., Chatzopoulos, S., Vlachos, I.S., Tastsoglou, S., Kanellos, I., Papadimitriou, D., Kavakiotis, I., Maniou, S. and Skoufos, G.J.N.a.r. (2018) DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA–gene interactions. *Nucleic acids research*, **46**, D239-D245.

101. Hsu, S.-D., Lin, F.-M., Wu, W.-Y., Liang, C., Huang, W.-C., Chan, W.-L., Tsai, W.-T., Chen, G.-Z., Lee, C.-J. and Chiu, C.-M.J.N.a.r. (2011) miRTarBase: a database curates experimentally validated microRNA–target interactions. *Nucleic acids research*, **39**, D163-D169.

102. Science, E.P.C.J. (2004) The ENCODE (ENCyclopedia of DNA elements) project. *Science*, **306**, 636-640.

103. Mathelier, A., Zhao, X., Zhang, A.W., Parcy, F., Worsley-Hunt, R., Arenillas, D.J., Buchman, S., Chen, C.-y., Chou, A. and Ienasescu, H.J.N.a.r. (2014) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic acids research*, **42**, D142-D147.

104. Lachmann, A., Xu, H., Krishnan, J., Berger, S.I., Mazloom, A.R. and Ma'ayan, A.J.B. (2010) ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics*, **26**, 2438-2444.

105. Liu, Z.-P., Wu, C., Miao, H. and Wu, H.J.D. (2015) RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database*, **2015**.

106. Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C. and Sayeeda, Z.J.N.a.r. (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic acids research*, **46**, D1074-D1082.

107. Piñero, J., Queralt-Rosinach, N., Bravo, A., Deu-Pons, J., Bauer-Mehren, A., Baron, M., Sanz, F. and Furlong, L.I.J.D. (2015) DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database*, **2015**.

108. Akhmedov, M., Kedaigle, A., Chong, R.E., Montemanni, R., Bertoni, F., Fraenkel, E. and Kwee, I.J.P.c.b. (2017) PCSF: An R-package for network-based interpretation of high-throughput data. *PLoS computational biology*, **13**, e1005694.

109. Wang, J., Vasaikar, S., Shi, Z., Greer, M. and Zhang, B.J.N.a.r. (2017) WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic acids research*, **45**, W130-W137.

110. Merico, D., Isserlin, R., Stueker, O., Emili, A. and Bader, G.D.J.P.o. (2010) Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PloS one*, **5**, e13984.

111. Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., Fridman, W.-H., Pagès, F., Trajanoski, Z. and Galon, J.J.B. (2009) ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*, **25**, 1091-1093.

112. Kim, Y.-M., Poline, J.-B. and Dumas, G.J.G. (2018) Experimenting with reproducibility: a case study of robustness in bioinformatics. *GigaScience*, **7**, giy077.

113. Leipzig, J.J.B.i.b. (2017) A review of bioinformatic pipeline frameworks. *Briefings in bioinformatics*, **18**, 530-536.

114. Su, G., Morris, J.H., Demchak, B. and Bader, G.D.J.C.p.i.b. (2014) Biological network exploration with Cytoscape 3. *Current protocols in bioinformatics*, **47**, 8.13. 11-18.13. 24.

115. Amrhein, V. and Greenland, S.J.N.h.b. (2018) Remove, rather than redefine, statistical significance. *Nature human behaviour*, **2**, 4-4.

116. Ritchie, M.D., Holzinger, E.R., Li, R., Pendergrass, S.A. and Kim, D.J.N.R.G. (2015) Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics*, **16**, 85-97.

117. Kitano, H.J.s. (2002) Systems biology: a brief overview. *Science*, **295**, 1662-1664.

118. Robinson, J.L. and Nielsen, J.J.M.B. (2016) Integrative analysis of human omics data using biomolecular networks. *Molecular BioSystems*, **12**, 2953-2964.

119. Szklarczyk, D.J.D.D. (2017) The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible Nucleic Acids Res., 45. *Nucleic acids research*.

120. Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N.H., Chavali, G., Chen, C. and Del-Toro, N.J.N.a.r. (2014) The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic acids research*, **42**, D358-D363.

121. Brown, K.R., Otasek, D., Ali, M., McGuffin, M.J., Xie, W., Devani, B., Toch, I.L.v. and Jurisica, I.J.B. (2009) NAViGaTOR: network analysis, visualization and graphing Toronto. *Bioinformatics*, **25**, 3327-3329.

122. Hu, Z., Snitkin, E.S. and DeLisi, C.J.B.i.b. (2008) VisANT: an integrative framework for networks in systems biology. *Briefings in bioinformatics*, **9**, 317-325.

123. Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J.A., Van Der Lee, R., Bessy, A., Cheneby, J., Kulkarni, S.R. and Tan, G.J.N.a.r. (2018) JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic acids research*, **46**, D260-D266.

124. Han, H., Cho, J.-W., Lee, S., Yun, A., Kim, H., Bae, D., Yang, S., Kim, C.Y., Lee, M. and Kim, E.J.N.a.r. (2018) TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic acids research*, **46**, D380-D386.

125. Holten, D. and Van Wijk, J. (2009) Force ‐ directed edge bundling for graph visualization. *Computer graphics forum*.

126. De Domenico, M., Porter, M.A. and Arenas, A.J.J.o.C.N. (2015) MuxViz: a tool for multilayer analysis and visualization of networks. *Journal of Complex Networks*, **3**, 159-176.

127. Secrier, M., Pavlopoulos, G.A., Aerts, J. and Schneider, R.J.B.b. (2012) Arena3D: visualizing time-driven phenotypic differences in biological systems. *BMC bioinformatics*, **13**, 1-11.

128. Bastian, M., Heymann, S. and Jacomy, M.J.I. (2009) Gephi: an open source software for exploring and manipulating networks. *Third international AAAI conference on weblogs and social media*, **8**, 361-362.

129. Theocharidis, A., Van Dongen, S., Enright, A.J. and Freeman, T.C.J.N.p. (2009) Network visualization and analysis of gene expression data using BioLayout Express 3D. *Nature protocols*, **4**, 1535.

130. Liluashvili, V., Kalayci, S., Fluder, E., Wilson, M., Gabow, A. and Gümüş, Z.H.J.G. (2017) iCAVE: an open source tool for visualizing biomolecular networks in 3D, stereoscopic 3D and immersive 3D. *GigaScience*, **6**, gix054.

131. Wang, Q., Tang, B., Song, L., Ren, B., Liang, Q., Xie, F., Zhuo, Y., Liu, X. and Zhang, L.J.B.b. (2013) 3DScapeCS: application of three dimensional, parallel, dynamic network visualization in Cytoscape. *BMC bioinformatics*, **14**, 1-8.

132. Salavert, F., García-Alonso, L., Sánchez, R., Alonso, R., Bleda, M., Medina, I. and Dopazo, J.J.B. (2016) Web-based network analysis and visualization using CellMaps. *Bioinformatics*, **32**, 3041-3043.

133. Fan, Y., Siklenka, K., Arora, S.K., Ribeiro, P., Kimmins, S. and Xia, J.J.N.a.r. (2016) miRNet-dissecting miRNA-target interactions and functional associations through network-based visual analysis. *Nucleic acids research*, **44**, W135-W141.

134. Kee, D.E., Salowitz, L. and Chang, R. (2012), *Poster Proc. IEEE Conf. InfoVis*.

135. biology, R.G.G.o.t.G.O.C.J.P.c. (2009) The Gene Ontology's Reference Genome Project: a unified framework for functional annotation across species. *PLoS computational biology*, **5**, e1000431.

136. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. and Morishima, K.J.N.a.r. (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research*, **45**, D353-D361.

137. Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F. and May, B.J.N.a.r. (2018) The reactome pathway knowledgebase. *Nucleic acids research*, **46**, D649-D655.

138. Mi, H., Poudel, S., Muruganujan, A., Casagrande, J.T. and Thomas, P.D.J.N.a.r. (2016) PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic acids research*, **44**, D336-D342.

139. Swainston, N., Smallbone, K., Hefzi, H., Dobson, P.D., Brewer, J., Hanscho, M., Zielinski, D.C., Ang, K.S., Gardiner, N.J. and Gutierrez, J.M.J.M. (2016) Recon 2.2: from reconstruction to model of human metabolism. *Metabolomics*, **12**, 1-7.

140. Cho, R.J., Huang, M., Campbell, M.J., Dong, H., Steinmetz, L., Sapinoso, L., Hampton, G., Elledge, S.J., Davis, R.W. and Lockhart, D.J.J.N.g. (2001) Transcriptional regulation and function during the human cell cycle. *Nature genetics*, **27**, 48-54.

141. Rosvall, M. and Bergstrom, C.T.J.P.o.t.n.a.o.s. (2007) An information-theoretic framework for resolving community structure in complex networks. *Proceedings of the national academy of sciences*, **104**, 7327-7331.

142. Pons, P. and Latapy, M. (2005), *International symposium on computer and information sciences*. Springer, pp. 284-293.

143. Raghavan, U.N., Albert, R. and Kumara, S.J.P.r.E. (2007) Near linear time algorithm to detect community structures in large-scale networks. *Physical review*, **76**, 036106.

144. Maizels, R.M. and Yazdanbakhsh, M.J.N.R.I. (2003) Immune regulation by helminth parasites: cellular and molecular mechanisms. *Nature Reviews Immunology*, **3**, 733-744.

145. Zhang, X., Edwards, J.P. and Mosser, D.M.J.T.J.o.I. (2006) Dynamic and transient remodeling of the macrophage IL-10 promoter during transcription. *The Journal of Immunology*, **177**, 1282-1288.

146. Sher, A., Gazzinelli, R.T., Oswald, I.P., Clerici, M., Kullberg, M., Pearce, E.J., Berzofsky, J.A., Mosmann, T.R., James, S.L. and Morse 3rd, H.J.I.r. (1992) Role of T-cell derived cytokines in the downregulation of immune responses in parasitic and retroviral infection. *Immunological reviews*, **127**, 183-204.

147. Bazzoni, F., Rossato, M., Fabbri, M., Gaudiosi, D., Mirolo, M., Mori, L., Tamassia, N., Mantovani, A., Cassatella, M.A. and Locati, M.J.P.o.t.N.A.o.S. (2009) Induction and regulatory function of miR-9 in human monocytes and neutrophils exposed to proinflammatory signals. *Proceedings of the National Academy of Sciences*, **106**, 5282-5287.

148. Nie, K., Gomez, M., Landgraf, P., Garcia, J.-F., Liu, Y., Tan, L.H., Chadburn, A., Tuschl, T., Knowles, D.M. and Tam, W.J.T.A.j.o.p. (2008) MicroRNA-mediated down-regulation of PRDM1/Blimp-1 in Hodgkin/Reed-Sternberg cells: a potential pathogenetic lesion in Hodgkin lymphomas. *The American journal of pathology*, **173**, 242-252.

149. Schork, N.J.J.N.N. (2015) Personalized medicine: time for one-person trials. *Nature News*, **520**, 609.

150. Pavlopoulos, G., Malliarakis, D., Papanikolaou, N., Theodosiou, T., Enright, A. and Iliopoulos, I. (2015) Visualizing genome and systems biology: technologies, tools, implementation techniques and trends, past, present and future. *GigaScience*, **4**, s13742-13015-10077-13742.

151. Integrative, H.J.C.h. and microbe. (2014) The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell host & microbe*, **16**, 276-289.

152. Kamburov, A., Cavill, R., Ebbels, T.M., Herwig, R. and Keun, H.C.J.B. (2011) Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. *Bioinformatics*, **27**, 2917-2918.

153. Hernández-de-Diego, R., Tarazona, S., Martínez-Mira, C., Balzano-Nogueira, L., Furió-Tarí, P., Pappas Jr, G.J. and Conesa, A.J.N.a.r. (2018) PaintOmics 3: a web resource for the pathway analysis and visualization of multi-omics data. *Nucleic acids research*, **46**, W503-W509.

154. Zhou, G. and Xia, J.J.N.a.r. (2018) OmicsNet: a web-based tool for creation and visual analysis of biological networks in 3D space. *Nucleic acids research*, **46**, W514-W522.

155. Huang, S., Chaudhary, K. and Garmire, L.X.J.F.i.g. (2017) More is better: recent progress in multi-omics data integration methods. *Frontiers in genetics*, **8**, 84.

156. Akavia, U.D., Litvin, O., Kim, J., Sanchez-Garcia, F., Kotliar, D., Causton, H.C., Pochanard, P., Mozes, E., Garraway, L.A. and Pe'er, D.J.C. (2010) An integrated approach to uncover drivers of cancer. *Cell*, **143**, 1005-1017.

157. Rohart, F., Gautier, B., Singh, A. and Lê Cao, K.-A.J.P.c.b. (2017) mixOmics: An R package for ‘omics feature selection and multiple data integration. *PLoS computational biology*, **13**, e1005752.

158. Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J.C., Buettner, F., Huber, W. and Stegle, O.J.M.s.b. (2018) Multi ‐ Omics Factor Analysis—a framework for unsupervised integration of multi ‐ omics data sets. *Molecular systems biology*, **14**, e8124.

159. John, C.R., Watson, D., Barnes, M.R., Pitzalis, C. and Lewis, M.J.J.B. (2020) Spectrum: fast density-aware spectral clustering for single and multi-omic data. *Bioinformatics*, **36**, 1159-1166.

160. Kayano, M., Imoto, S., Yamaguchi, R. and Miyano, S.J.J.o.C.B. (2013) Multi-omics approach for estimating metabolic networks using low-order partial correlations. *Journal of Computational Biology*, **20**, 571-582.

161. Nguyen, H., Shrestha, S., Draghici, S. and Nguyen, T.J.B. (2019) PINSPlus: a tool for tumor subtype discovery in integrated genomic data. *Bioinformatics*, **35**, 2843-2846.

162. Ng, A.Y., Jordan, M.I. and Weiss, Y. (2002), *Advances in neural information processing systems*, pp. 849-856.

163. Cantini, L., Zakeri, P., Hernandez, C., Naldi, A., Thieffry, D., Remy, E. and Baudot, A.J.N.c. (2021) Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nature communications*, **12**, 1-12.

164. Chauvel, C., Novoloaca, A., Veyre, P., Reynier, F. and Becker, J.J.B.i.b. (2020) Evaluation of integrative clustering methods for the analysis of multi-omics data. *Briefings in bioinformatics*, **21**, 541-552.

165. Stein-O'Brien, G.L., Arora, R., Culhane, A.C., Favorov, A.V., Garmire, L.X., Greene, C.S., Goff, L.A., Li, Y., Ngom, A. and Ochs, M.F.J.T.i.G. (2018) Enter the matrix: factorization uncovers knowledge from omics. *Trends in Genetics*, **34**, 790-805.

166. Meng, C., Basunia, A., Peters, B., Gholami, A.M., Kuster, B., Culhane, A.C.J.M. and Proteomics, C. (2019) MOGSA: integrative single sample gene-set analysis of multiple omics data. *Molecular & Cellular Proteomics*, **18**, S153-S168.

167. Gomez-Cabrero, D., Tarazona, S., Ferreirós-Vidal, I., Ramirez, R.N., Company, C., Schmidt, A., Reijmers, T., von Saint Paul, V., Marabita, F. and Rodríguez-Ubreva, J.J.S.d. (2019) STATegra, a comprehensive multi-omics dataset of B-cell differentiation in mouse. *Scientific data*, **6**, 1-15.

168. Ghaemi, M.S., DiGiulio, D.B., Contrepois, K., Callahan, B., Ngo, T.T., Lee-McMullen, B., Lehallier, B., Robaczewska, A., Mcilwain, D. and Rosenberg-Hasson, Y.J.B. (2019) Multiomics modeling of the immunome, transcriptome, microbiome, proteome and metabolome adaptations during human pregnancy. *Bioinformatics* **35**, 95-103.

169. Aghaeepour, N., Ganio, E.A., Mcilwain, D., Tsai, A.S., Tingle, M., Van Gassen, S., Gaudilliere, D.K., Baca, Q., McNeil, L. and Okada, R.J.S.i. (2017) An immune clock of human pregnancy. *Science immunology*, **2**.

170. Irani, R. and Xia, Y. (2011) Seminars in nephrology.

171. Yang, M.-J., Tseng, J.-Y., Chen, C.-Y. and Yeh, C.-C.J.J.o.t.C.M.A. (2013) Changes in maternal serum insulin-like growth factor-I during pregnancy and its relationship to maternal anthropometry. *Journal of the Chinese Medical Association*, **76**, 635-639.

172. Zoppi, J., Guillaume, J.-F., Neunlist, M. and Chaffron, S.J.B.b. (2021) MiBiOmics: an interactive web application for multi-omics data exploration and integration. *BMC bioinformatics*, **22**, 1-14.

173. Grigorov, M.G. (2011), *Bioinformatics for Omics Data*. Springer, pp. 153-172.

174. Spies, D., Renz, P.F., Beyer, T.A. and Ciaudo, C. (2019) Comparative analysis of differential gene expression tools for RNA sequencing time course data. *Briefings in bioinformatics*, **20**, 288-298.

175. Conard, A.M., Goodman, N., Hu, Y., Perrimon, N., Singh, R., Lawrence, C. and Larschan, E. (2021) TIMEOR: a web-based tool to uncover temporal regulatory mechanisms from multi-omics data. *Nucleic Acids Research*, **49**, W641-W653.

176. Schadt, E.E. (2009) Molecular networks as sensors and drivers of common human diseases. *Nature*, **461**, 218-223.

177. Ellson, J., Gansner, E., Koutsofios, L., North, S.C. and Woodhull, G. (2001), *International Symposium on Graph Drawing*. Springer, pp. 483-484.

178. Zanfardino, M., Franzese, M., Pane, K., Cavaliere, C., Monti, S., Esposito, G., Salvatore, M. and Aiello, M.J.J.o.t.m. (2019) Bringing radiomics into a multi-omics framework for a comprehensive genotype–phenotype characterization of oncological diseases. *Journal of translational medicine*, **17**, 1-21.

179. Ahmed, Z.J.H.G. (2020) Practicing precision medicine with intelligently integrative clinical and multi-omics data analysis. *Human Genomics*, **14**, 1-5.

180. Spiegl-Kreinecker, S., Lötsch, D., Ghanim, B., Pirker, C., Mohr, T., Laaber, M., Weis, S., Olschowski, A., Webersinke, G. and Pichler, J. (2015) Prognostic quality of activating TERT promoter mutations in glioblastoma: interaction with the rs2853669 polymorphism and patient age at diagnosis. *Neuro-oncology*, **17**, 1231-1240.

181. Czauderna, T., Haga, J., Kim, J., Klapperstück, M., Klein, K., Kuhlen, T., Oeltze-Jafra, S., Sommer, B. and Schreiber, F. (2018), *Immersive Analytics*. Springer, pp. 289-330.

182. Sommer, B., Hamacher, A., Kaluza, O., Czauderna, T., Klapperstück, M., Biere, N., Civico, M., Thomas, B., Barnes, D.G. and Schreiber, F. (2016) Stereoscopic space map–semi-immersive configuration of 3D-stereoscopic tours in multi-display environments. *Electronic Imaging*, **2016**, 1-9.

183. (2021). Fortune Business Insights, Vol. 2021.

184. Hadjar, H., Meziane, A., Gherbi, R., Setitra, I. and Aouaa, N. (2018), *Proceedings of the 2nd International Conference on Web Studies*, pp. 56-63.

185. Dibbern, C., Uhr, M., Krupke, D. and Steinicke, F.J.M.u.c.-u.p. (2018) Can WebVR further the adoption of Virtual Reality? *Mensch und computer 2018-usability professionals*.

186. Sommer, B., Baaden, M., Krone, M. and Woods, A. (2018) From virtual reality to immersive analytics in bioinformatics. *Journal of integrative bioinformatics*, **15**.

187. Bhaskar, H., Hoyle, D.C. and Singh, S. (2006) Machine learning in bioinformatics: A brief survey and recommendations for practitioners. *Computers in biology and medicine*, **36**, 1104-1125.

188. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R. and Dubourg, V. (2011) Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, **12**, 2825-2830.

189. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H. (2009) The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, **11**, 10-18.

190. Vellido, A.J.N.C. and Applications. (2019) The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural computing and applications*, 1-15.

191. Kelly, C.J., Karthikesalingam, A., Suleyman, M., Corrado, G. and King, D. (2019) Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, **17**, 1-9.

192. Chatzimparmpas, A., Martins, R.M., Jusufi, I. and Kerren, A.J.I.V. (2020) A survey of surveys on the use of visualization for interpreting machine learning models. *Information Visualization*, **19**, 207-233.

193. Chegini, M., Bernard, J., Berger, P., Sourin, A., Andrews, K. and Schreck, T. (2019) Interactive labelling of a multivariate dataset for supervised machine learning using linked visualisations, clustering, and active learning. *Visual Informatics*, **3**, 9-17.