2

Integrative Biology Approaches Applied to Human Diseases

Alysson H. Urbanski[1] • José D. Araujo[1] • Rachel Creighton[2] • Helder I. Nakaya[1,3]

[1]Department of Clinical and Toxicological Analyses, School of Pharmaceutical Sciences, University of Sao Paulo, Sao Paulo, Brazil; [2]Department of Bioengineering, University of Washington, Seattle, WA, USA; [3]Scientific Platform Pasteur/USP, University of Sao Paulo, Sao Paulo, Brazil

Author for correspondence: Helder I. Nakaya, Department of Clinical and Toxicological Analyses, School of Pharmaceutical Sciences, University of Sao Paulo, Sao Paulo, SP, 05508, Brazil. Email: hnakaya@usp.br

Doi: http://dx.doi.org/10.15586/computationalbiology.2019.ch2

Abstract: The study of multifactorial and complex interactions in human diseases has been transformed by the omics revolution. The speed and scale of omics analysis have increased exponentially in the past decades, and it is now easier and faster to generate large amounts of biological data. However, extracting meaningful information from this "sea of data" remains a major challenge. The field of integrative biology utilizes a holistic approach to integrate multilayer biological data. In this chapter, we introduce concepts and techniques for the analysis of single-layer omics data and for integrating multilayer omics datasets to extract meaningful and relevant biological insights. Integrative biology is a promising approach for the study of a wide range of human diseases. We also highlight some current challenges in the field, such as the need for more specialized and interpretable methods, while increasing the accessibility of integrative analysis for the scientific community.

Keywords: integrative biology; multi-omics; proteogenomics; single-layer high-throughput data; systems biology

In: *Computational Biology.* Holger Husi (Editor), Codon Publications, Brisbane, Australia. ISBN: 978-0-9944381-9-5; Doi: http://dx.doi.org/10.15586/computationalbiology.2019

INTRODUCTION

Human diseases involve complex interactions between genes, environment and lifestyle (1). For example, in type 2 diabetes mellitus, there are many behavioral, lifestyle, and genetic risk factors and other pathophysiological abnormalities contributing to hyperglycemia. Major mechanisms of the disease are impaired insulin secretion and insulin resistance in muscle and liver; however, other genes and signaling pathways in different tissues are also involved, such as increased kidney malfunction, inflammation, and neurotransmitter dysfunction (2). Other well-known examples of complex, multigenic, or multifactorial diseases are tumors (3), infectious diseases (4), and cardiovascular diseases (5).

Life sciences research has been revolutionized in past decades by a series of genome-wide technologies, starting with the Human Genome Project in 1990. The speed and scale of genomics analysis increased exponentially after this, facilitated by technologies such as microarrays and high-throughput sequencing (6). Genomics is classified as discovery science, along with other omics such as transcriptomics, miRNAomics, epigenomics, cistromics, proteomics, metabolomics, and microbiomics. The goal of discovery science is to collect and store data describing all the elements of a system (6, 7). As it has become easier and faster to generate large amounts of biological data, new challenges in data analysis and interpretation are emerging (8).

High-throughput data allow us to visualize processes in a certain layer of biological information in an organism or at the single-cell level. A recent example is the association of CD177+ neutrophils to Kawasaki disease through genome-wide transcriptome analysis (9). Additionally, analyzing the metabolome of coronary atherosclerosis patients enabled discovery of several biomarkers of lipid metabolism dysfunctions (10). At a proteomic level, researchers have identified proteins in the brain which are associated with the cognitive trajectory in the elderly (11). Finally, the evolution of single-cell sequencing has allowed the evaluation of these different layers in greater detail (12). The analysis of omics data has advanced the understanding of human diseases, but it is important to remember that these studies represent only one layer of a more complex system.

Network science analyzes the interactions between biomolecules (proteins, RNA, gene sequences), pathways, cells, organs, and even individuals using graph theory methods, and it is an efficient way of extracting information from omics data. Through network analysis, it is possible to identify complex patterns among different components to generate scientific hypotheses regarding the interactions present in health and disease events (13). For example, a recent gene expression network analysis study identified a membrane receptor as a potential therapeutic target for an antiepileptic drug (14). Although the integration of genes into networks gives us a lot of information, it describes only one omics level. Therefore, there is a growing interest in the integration of different omics data (15). In this chapter, we introduce concepts and tools for the analysis of single-layer biological data and integration of multilayer biological data to extract meaningful and relevant biological insights of various human diseases (Figure 1).

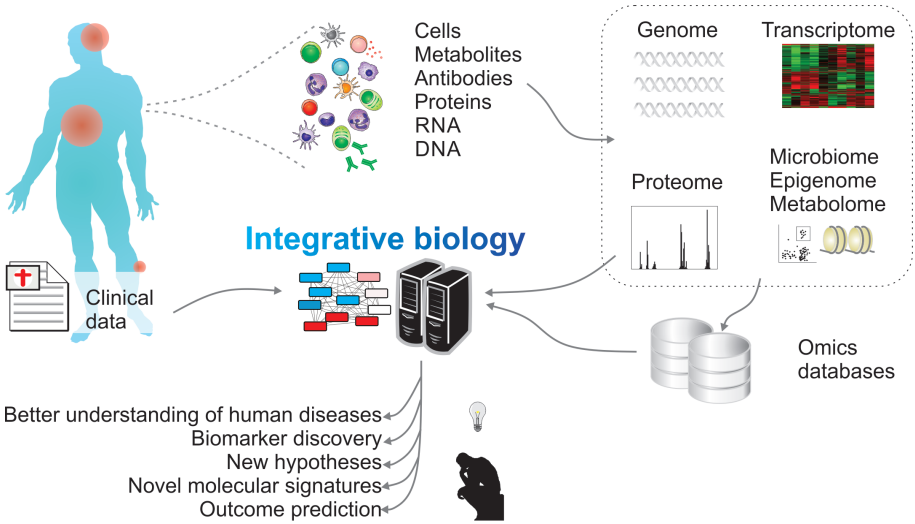

Figure 1 A framework for integrative biology. High-throughput techniques such as transcriptomics, proteomics and metabolomics, in addition to clinical data and other databases, can be used to investigate human diseases through an integrative approach.

APPLICATIONS OF SINGLE-LAYER HIGH-THROUGHPUT DATA

Since the popularization of next-generation sequencing (NGS) and high-throughput mass spectrometry methods, there has been an exponential increase in the generation of biological data, and it is likely that the amount of biological data available will continue to increase. The evolution of high-throughput mass spectrometry has enabled high-resolution visualization of the proteome and metabolome of cells, tissues, and fluids. These data are useful to understand the pathogenic mechanisms, contributing to diagnoses, prognoses, and potential therapeutic interventions.

DNA genomes and exomes can be elucidated using NGS. NGS-based techniques have already overcome the use of microarrays for RNA transcriptome sequencing by enabling the identification of virtually any transcript present in the sample, including unknown transcripts. NGS techniques can also identify differentially expressed genes (DEGs) by applying statistical methods to the expression data (16). Recently, long noncoding RNA (lnc-RNA) (17) and circular RNA (18) molecules have been implicated in the regulation of the innate immune response and can potentially elucidate infectious, autoimmune, and inflammatory disease mechanisms. Despite this, it is important to remember the limitations of studying a heterogeneous mixture of cells. Although the cells may be similar in morphology, localization or other classificatory factors, it is impossible to understand individual cellular features such as metabolic states, transcriptional levels, and metabolic activation using traditional bulk transcriptome sequencing (19).

Thus, RNA sequencing at single-cell level (scRNA-seq) allows a more accurate reconstruction of intracellular and intercellular network interactions (20). Since the first scRNA-seq a decade ago (21), the technology has improved and several protocols and platforms have been developed to respond to the most diverse biological problems, including those related to immune system in health and disease (22, 23). Recently, ultra-high-throughput scRNA-seq techniques based on the droplets strategy, such as Drop-Seq (24), InDrop (25), and 10X Genomics Chromium (26), have gained popularity. These techniques can reduce the cost of sequencing while increasing the throughput by allowing a parallel mRNA profiling of thousands of individual cells by encapsulating them in droplets (27). Raw and processed high-throughput data are stored in several online repositories, making them valuable resources for discovery science approaches (7). The content of the data repositories ranges from genomics and transcriptomics to epigenetics, protein–protein interaction, metabolomics, and microbiome data (Table 1).

Examples of big data generation in specific human disease applications are numerous. Although we do not focus on any specific disease in this chapter, we provide several relevant examples. Zhao et al. performed the transcriptomic profiling of glioma, generating 30 billion reads, from 325 samples in different stages of malignant progression (28). There have also been efforts to investigate in vitro and in vivo response to viral infections, such as influenza and severe acute respiratory syndrome coronavirus, generating dozens of transcriptome and proteome datasets (29). More specific events have also been investigated, such as the methylome of brain metastases that may help to predict individual responses to therapies (30) or the profiling of long non-coding RNA in human hypertrophic cardiomyopathy (31). Data generated from a large-scale multi-omic study, including genome and transcriptome sequencing and proteomic profiling of a large cohort of Alzheimer's disease patients, could improve our knowledge about this pathology (32). In another study, the characterization of *post-mortem* microbial diversity in 188 individuals allowed a better understanding of the *ante-mortem* health condition of some individuals, suggesting that it is possible to estimate the health conditions in living populations from these data (33).

TOOLS FOR THE ANALYSIS OF SINGLE-LAYER HIGH-THROUGHPUT DATA

Ensuring data quality is an essential step in the analysis and integration of omics data. When artifacts and noise are not handled correctly, they can influence the results of the analysis (34). The term "garbage in, garbage out," a common concept in computer science and mathematics, is also applicable in bioinformatics. This phrase means that the output data quality is determined by the input data quality. Several methods can be used to evaluate and control input data quality. One strategy is to determine the statistical significance to avoid false positives, known as the false discovery rate (FDR). Despite a recent debate about the appropriate use of statistical significance, an FDR value of 0.05 or smaller has been generally accepted in academia (35). In addition to the statistical analysis of individual layers, it is important to ensure that the data are biologically meaningful. In this case, the fold-change cut-off is used. The fold-change describes how

TABLE 1 Biological repositories

Database	Description	Reference website
ArrayExpress	Functional genomics data from microarray or NGS. Data types include transcription profiling (mRNA and miRNA), SNP genotyping, chromatin immunoprecipitation (ChIP), and comparative genomic hybridization	https://www.ebi.ac.uk/arrayexpress/
BioGRID	Curated database. Data types include protein–protein, genetic and chemical interactions, and post-translational modifications	https://thebiogrid.org/
dbGAP	Data and results from the interaction of genotype and phenotype	https://www.ncbi.nlm.nih.gov/gap/
ENCODE	Whole-genome database	https://encodeproject.org/
GDC	Genomic, epigenomic, transcriptomic, and proteomic data from cancer samples	https://portal.gdc.cancer.gov/
GEO	Gene expression, hybridization arrays, chips, and microarrays database	https://www.ncbi.nlm.nih.gov/geo/
GTEx	The genotype–tissue expression includes data of tissue-specific gene expression and regulation	https://gtexportal.org/home/
HMDB	Human metabolome database	http://www.hmdb.ca/
ICGC	Cancer genomics database	https://dcc.icgc.org/
IMGT	Immune-related genes sequence database	http://www.imgt.org/
InnateDB	Genes, proteins, interactions, and pathways involved in the innate immune response	https://www.innatedb.com/
MethylomeDB	DNA methylation profiles	http://habanero.mssm.edu/methylomedb/index.html
MGnify	Microbiome database	https://www.ebi.ac.uk/metagenomics/
miRbase	miRNA sequences and annotation	http://www.mirbase.org/
PHISTO	Pathogen–human protein–protein interaction database	http://www.phisto.org/
Reactome	Curated pathway database	https://reactome.org/
SRA	Sequencing and alignment data	https://www.ncbi.nlm.nih.gov/sra
STRING	Protein–protein interaction networks	https://string-db.org/

These databases store raw or processed, and sometimes curated, data derived from different studies and omics technologies.

much a gene or pathway is up- or down-regulated, for example, 2 or 0.5, respectively (36). This kind of analysis allows further downstream integration of the data, since it is possible to associate, for example, a group of DEGs and the metabolic pathways that they belong to (37).

Numerous tools are used to analyze different types of data. Although it is not the focus of this chapter to describe these tools, the concepts of some techniques are described here. Bioconductor is a robust software platform used in the analysis

of omics data (https://www.bioconductor.org/). In bioconductor, there are several packages, mainly in the R scripting language, that provide metrics and methods to evaluate reproducibility, identify outliers and noise. For example, the EdgeR package for gene expression analysis calculates the difference in gene expression for different samples and conditions, considering both the FDR and fold-change of each gene (38). Bioconductor can also be used to analyze high-dimensional mass cytometry (CyTOF) datasets. CyTOF is a platform for collecting high-dimensional phenotypic and functional data for single cells (39). For example, CyTOF can be used to uncover tissue- and disease-associated immune cell subsets (40). A review by Nowicka et al. presents a detailed workflow for CyTOF analyses using the bioconductor platform (41).

Metabolomics provides quantification of metabolites in cells, tissues or biological fluids (42). Several tools are available for the analysis of metabolomics data, including the web tool MetaboAnalyst (43) and the R package MetaboAnalystR (44). Both carry out analyses with the same workflow: (i) Exploratory data analysis; (ii) Metabolic enrichment analysis and metabolic pathway activity prediction; and (iii) Data integration, such as biomarker meta-analysis, joint path analysis, and network explorer. The data input for these tools can be a list of genes or KEGG orthologs.

Single-cell RNA-seq (scRNA-seq) methods are also widely used in studies involving human health (23). To ensure a biologically significant analysis, it is necessary to consider the intrinsic variations of the technique, called batch effects (45). There are several tools that assist in the batch correction process, most of which are based on linear regression, including limma (46), RUVseq (47, 48), and svaseq (49). Other promising approaches for batch correction are based on the detection of mutual nearest neighbors in the high-dimensional gene expression space (50).

The high-dimensional gene expression space is a matter of concern when analyzing scRNA-seq gene expression data. The problem with this high-dimensional space is that it is hard to differentiate the variability between cell populations from the variability between cells within a population, as the distances between cells become more homogenous. High-dimensional data are handled through dimensionality reduction and feature selection. Dimensionality reduction is a process to project data in a smaller dimensional space, preserving some key characteristics of the sample enough to distinguish differences between populations (51). While principal component analysis (PCA) is the recommended tool for RNA-seq, T-distributed stochastic neighbor embedding (tSNE) is the most popular method for dimensionality reduction of scRNA-seq data. PCA is not recommended for scRNA-seq datasets because it is a linear dimensionality reduction algorithm and assumes approximately normally distributed data, while tSNE uses different probability distributions that are more suitable to scRNA-seq data (51). Nonetheless, a recently developed nonlinear dimensionality-reduction technique named uniform manifold approximation and projection (UMAP) outperformed other dimensionality-reduction methods for cell clustering (52). Feature selection reduces the number of dimensions by excluding uninformative genes and identifying the most relevant features for analysis (53). Feature selection in scRNA-seq can be based on correlated expression, highly variable genes (HVG), Michaelis–Menten modeling of dropouts (M3Drop) or spike-in methods (51).

As already mentioned, scRNA-seq enables the identification of transcriptionally distinct cell subpopulations in an otherwise homogeneous cell population. Identification of these groups is typically accomplished through clustering analysis. Clustering approaches can be supervised or unsupervised. If the method uses a known set of gene markers for clustering, it is supervised. Alternatively, unsupervised clustering methods can identify groups without prior information (53). There are many algorithms designed for unsupervised clustering, but the main classes of them are k-means, hierarchical, density-based, and graph clustering (51). For example, through transcriptional clustering analysis of $CD127^{+}$ innate lymphoid cells (ILCs), Björklund et al. uncovered four different cell subpopulations: three different ILCs and natural killer (NK) cells. The group further subdivided the ILC3 group into three new transcriptionally and functionally distinct populations, contributing to the knowledge of ILC biology, and associated inflammatory processes (54).

Clustering analyses in scRNA-seq data can be very useful and informative, but they are not always able to describe dynamic biological processes involved in transitions between different states, such as cellular proliferation and maturation (12). Such events can be computationally modeled through the reconstruction of the cell trajectory and pseudotime estimation (53). Because the cells in a scRNA-seq experiment are unsynchronized, there are different instantaneous timepoints captured that together may represent an entire cell trajectory (55). The term pseudotime refers to an ordering of the cells according to some dynamic process of interest, such as development processes occurring over time. Through pseudotime estimation, cells in different states of a trajectory can be identified, permitting identification of transcriptional changes, branching points in trajectories, and reconstruction of gene regulatory networks (56). Recent efforts have used trajectory and pseudotime methods to better understand human diseases, including hepatitis B (57), osteoarthritis (58), muscular dystrophy (59), and Parkinson's disease (60). As bulk tissue RNA-seq data is more accessible than scRNA-seq data, there is a great interest in the development of deconvolution tools capable of describing the cellular composition of tissue samples, especially in the study of tumors (61).

RNA-seq techniques are also useful for studying the high variability of the immune system and how this may influence disease progression. The immune repertoire is defined as the set of B-cell receptors (BCR) and T-cell receptors (TCR) of an organism. The former directly binds antigen to initiate differentiation of B cells into plasma cells, which then secrete antibodies. The latter recognizes antigens bound to major histocompatibility complex (MHC) molecules displayed on antigen-presenting cells. A robust adaptive immune system relies on the generation of a wide variety of BCRs and TCRs to recognize a varied range of antigens. A highly diverse immune repertoire is generated through V(D)J recombination. Additionally, the BCRs undergo somatic hypermutation, which increases the antigen binding specificity and affinity. Several bioinformatics tools have been developed to accurately determine the immune repertoires from genomic or RNA sequencing data, with a focus on the hypervariable complementarity-determining region 3 (CDR3) sequences. Some of these tools are specific to BCR or TCR, such as TRUST (62) and V'Djer (63), while others can work with both receptor types, such as MiXCR (64). There are also specific tools for scRNA-seq data, such as BASIC (65).

APPLICATIONS OF INTEGRATIVE BIOLOGY TO HUMAN DISEASES

Diseases are accompanied by many simultaneous changes in cell and molecular dynamics, such as gene and protein expression, metabolic pathways, and tissue cell population composition, that can be the cause or consequence of the disease state. An integrative approach to investigate these complex changes and interactions can enable a more holistic understanding of immunology, including inhibition of viral replication, generation of protective immune responses, pathogen evasion of innate and adaptive immunity, and differences in susceptibility between individuals and populations (66).

The central dogma of molecular biology states that the information is transferred sequentially from mRNA to proteins (67). However, this does not always mean there is a perfect correlation between mRNA and protein expression, highlighting the importance of analyzing multiple layers of biological data (68). In fact, now it is clear that the correlation between mRNA and protein expression depends on the cell state. In steady-state conditions, mRNA and protein levels have a strong positive correlation, but during dynamic conditions, including stress responses that are cause or consequence of disease, post-transcriptional processes cause deviations from an ideal positive correlation (69).

MicroRNAs (miRNAs) are short and endogenous RNAs that play important regulatory roles by suppressing mRNA translation by directing mRNA degradation. Again, we might expect a negative correlation between miRNA levels and target protein expression, but the correlation patterns are more complex than expected (70). Nunez et al. observed positively correlated miRNA and mRNA in a mouse model during early stages of alcohol dependence, suggesting that early miRNA activation may play an important role to limit the effect of alcohol-induced genes (71). Recently, an extensive investigation revealed the miRNA–mRNA correlation profile in human peripheral blood mononuclear cells (PBMC) in a rheumatoid arthritis cohort (70), leading to a better understanding of this and other autoimmune diseases (72). Similar efforts are being applied to profile the miRNA-mRNA correlation in tumorigenesis (73).

As personalized and precision medicine evolves, integration of metabolomics data with other layers of information becomes increasingly important. Nakaya et al. (74) used a systems analysis approach to uncover shared molecular signatures that predict influenza antibody response after vaccination. Briefly, they were able to identify transcriptomic signatures of innate immunity that could predict influenza vaccine-induced antibody titers. In addition, they uncovered many miRNA regulators of the response after vaccination. Another example study showing metabolomics integration with proteomics data uncovered signatures of innate immunity, T-cell signaling, and platelet activation related to clinical tolerance to *Plasmodium vivax* (75). Another study showed the association between metabolic pathways and chronic obstructive pulmonary disease (COPD) phenotypes, applying an unbiased metabolomics and transcriptomics approach, enabling the determination of phenotypic and outcome differences (76).

The study of genetic variability is important in the context of human health, since it may be related to differential disease risk in a population. Genome-wide association studies showed that approximately 80% of single-nucleotide polymorphisms (SNPs) associated with human phenotypes are located within non-coding regions, showing the potential association between these regions and the regulation of differential gene expression in health and disease (77) or in pharmacologic susceptibility (78). These non-coding regions may explain part of the variation and tissue-specificity in mRNA expression levels (79). By integrating genomic and transcriptomic data, scientists can find other expression quantitative trait loci (eQTLs) responsible for partial or complete alteration of gene expression (80).

Proteogenomics is an integrative approach between genomic and transcriptomic data, which has greatly advanced the study of several pathologies, especially cancer (81). This approach includes two methods of extracting information. In one method, data from transcriptomics and genomics are used to create protein databases with new peptides that are not present in reference databases. Alternatively, transcriptomics data can be used to validate genomics data and refine gene models (82). For example, Mun et al. performed an extensive proteogenomic characterization of patients with gastric cancer by integrating transcriptional, protein, phosphorylation, and N-glycosylation data (83). The group identified markers that predict a patient's prognosis and how they would respond to treatment. Similarly, this integration of proteogenic data has allowed a better understanding of colon cancer pathology and identification of potential therapeutic targets (84). Integration of metabolome, proteome, and clinical data has also been a powerful approach in fields other than oncology. For example, potential biomarkers for sepsis prognosis have been identified, which may aid in the development of new therapies for patients at higher risk of death (85).

To understand the response to herpes zoster vaccine, Li et al. (86) conducted a multi-layered study combining different datasets including transcriptomics, blood cell population flow cytometry, and plasma cytokine analysis to identify molecular networks correlated with adaptive immunity responses. The analysis revealed high correlations between distinct molecular signatures and biological convergence between the pathways identified by the metabolomic and transcriptomic data. These convergences suggested that the transcription program of blood cells is potentially regulatory in response to metabolic stimuli. For example, the same gene network, consisting of heme biosynthesis, BCR signaling, and inositol phosphate metabolism, was highly expressed in subjects with higher viral load. There were also significant differences between young and old adults, including NK cells frequency and expression of inflammatory genes. This contextualization of immune responses related to vaccination provides a good example of how these new integrative biology techniques may aid in research involving complex molecular responses such as biomarker identification and development of new immunization protocols.

The integration of omics data in health and disease has enabled a more detailed understanding of molecular interactions. This approach has improved the ability to study highly complex diseases including psychiatric diseases (87), pulmonary diseases (88), cardiovascular diseases (89), and the role of the microbiota in inflammatory bowel diseases (90).

TOOLS FOR INTEGRATIVE ANALYSIS

The molecular complexity of many diseases and advances in data integration have popularized studies that integrate different levels of biological data. However, integrative data analysis depends on the data types available and the aims of the study. Consequently, with the emergence of multi-omic data, new challenges have appeared for the development of appropriate statistical computational methods to integrate these data. Methods are required for the integration of the same type of data collected from different studies and the integration of different types of data collected from the same sample, termed horizontal and vertical data integration, respectively (Figure 2) (91). Although not discussed in detail, we briefly review some concepts of omics data integration.

In addition to horizontal and vertical data integration, multiple layers of data can be integrated using top–down and bottom–up approaches. Bottom–up integration consists of associating genomics and/or transcriptomics data with proteomics, metabolomics and/or clinical data in order to predict global changes in a cell or organism, such as phenotypic responses and key pathways. In contrast, a top–down approach consists of parallel clustering of different categories of data for automated and unified integration (92).

One bottom–up method used frequently in the integration of multiple omic layers is the search for correlations (93). This approach is based on regression methods and seeks to find elements that vary simultaneously in different layers, such as the search for SNPs and eQTLs that influence gene expression and are responsible for disease phenotypes (94). Co-expression network analysis is an informative bottom–up approach that can improve our knowledge in functional annotation and disease gene prediction (95). Recently, an integrative tool, CEMiTool, for the identification of co-expression modules was developed (95). In addition to unsupervised identification of co-expression modules, this tool allows automated integration with gene set enrichment analysis (96), which can identify whether the co-expression gene module is enriched for some relevant biological pathway and associated with a phenotype. This tool can also integrate co-expression modules with protein–protein interaction data, which is useful to identify the key regulators of a network (95). Other bottom–up approaches include clustering of DNA, mRNA, miRNA, protein, metabolite, epigenetic, network, and manual annotation data for later integration. These approaches are concisely described in a review by Yu and Zeng (92).

MixOmics is a multi-omic integrative computational tool based on the R language that is useful in a wide variety of omic studies. It is dedicated to the multivariate analysis of biological datasets with a specific focus on data exploration, dimensionality reduction, and data visualization (97). It offers a wide range of supervised statistical analysis methods that integrate multiple omic data to analyze relationships between these data. The methods include canonical correlation analysis, partial least squares regression, and PCA to perform discriminant analysis, horizontal or vertical integration, and the identification of molecular signatures (98, 99). Assuming the data have been normalized by specific methods (depending on its nature), mixOmics can explore and integrate different types of biological data. The input can be based on both discrete and continuous data such

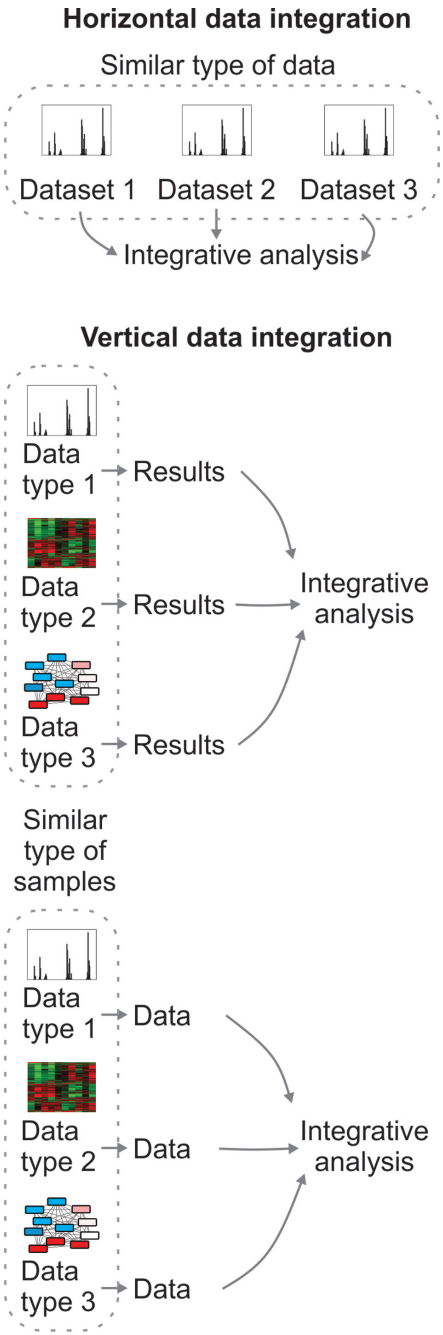

Figure 2 Horizontal and vertical data integration. Horizontal integration joins the similar data type of *n* datasets for analysis, while vertical integration combines different data types from the similar types of samples. Vertical analysis can integrate individually generated results (middle panel) or extract complex patterns directly from the data in parallel (bottom).

as mass spectrometry, microarray, proteomics, and metabolomics, or data generated by sequencing, such as RNA-seq, 16S, and metagenomic shotgun.

In contrast, a top–down approach consists of the parallel clustering of different categories of data for automated and unified integration (92). Top–down methods consist of statistical and machine learning tools such as joint models (100), Bayesian analysis (101), factor analysis (102), multiple kernel learning (103), deep learning (104), and simultaneous clustering (105). There are many useful data integration methods, and the method selection depends on the nature of the data to be analyzed. With the increasing availability of data on public databases and the development of new methods, the tendency is for greater use of omic data integration.

CHALLENGES

With the continued advancement of NGS technologies, omics science is expected to move towards an increasingly integrative approach. With this shift, managing the vast amount of data generated and integrating these data in a significant way remains a challenge (106, 107). There are concerns about the data reproducibility and accessibility (108) and efforts to overcome this, such as the FAIR principles (109). The FAIR guideline suggests ways to data become Findable, Accessible, Interoperable, and Reusable. Additionally, curated databases and improved software-database interoperability would facilitate data integration (110). Another part of the solution is the popularization of open source sharing platforms, such as GitHub, enabling developers and users to share and review their codes and scripts, as well as develop tools in collaboration with other researchers (111). A particular issue is to go beyond finding correlations to infer causality between two or more elements, such as concentration of metabolites and levels of gene expression (112). This remains a great challenge for integrative biology, which relies on molecular studies, both in vitro and in vivo, to attest the causation (93). It is important to develop new analytical methods to produce results that are easy to interpret, since the interpretation of the results can be another challenge as great as the creation of new tools (110). Finally, the evolution of integrative biology also depends on massive computational resources, both for data storage and analysis (113).

CONCLUSION

Although a huge amount of biological data is being generated at incredible pace, this is not being translated to knowledge. A large fraction of the data has the potential to be applied in clinical practice, but they are idle in repositories or waiting for the development of proper methods for data integration and interpretation. Traditionally, these data are generated by conventional hypothesis-driven methodologies. In this approach, the hypothesis is stated, tested and then accepted or refuted, based on the outcome. Alternatively, the popularization of high-throughput technologies spreads the data-driven hypothesis, or hypothesis-free, approach. In data-driven hypothesis definition, models are created after data

analysis and only then a hypothesis is formulated and tested. This integrative and systems approach can reproduce complex disease states and, therefore, has higher chances of clinical implementation. Hypothesis-driven generation and data-driven hypothesis generation are non-exclusive, since the latter can use the data produced by the former to create useful models for new hypothesis-driven studies. In this context, collaboration between bioinformatics and wet lab experts is essential for integrating multiple layers of information, which is, and will continue to be, very useful for elucidating how disease processes occur and for the development of new therapeutic interventions.

Acknowledgement: This work was supported by the São Paulo Research Foundation (FAPESP; grants 2018/14933-2, 2018/21934-5 and 2013/08216-2) and a grant from the Innovative Medicines Initiative 2 Joint Undertaking (IMI2 JU) under the VSV-EBOPLUS (grant number 116068) project.

Conflict of interest: The authors declare no potential conflict of interest with respect to research, authorship, and/or publication of this chapter.

Copyright and permission statement: To the best of our knowledge, the materials included in this chapter do not violate copyright laws. All original sources have been appropriately acknowledged and/or referenced. Where relevant, appropriate permissions have been obtained from the original copyright holder(s).

REFERENCES

1. Hunter DJ. Gene–environment interactions in human diseases. Nat Rev Genet. 2005 Apr;6(4): 287–98. http://dx.doi.org/10.1038/nrg1578
2. DeFronzo RA, Ferrannini E, Groop L, Henry RR, Herman WH, Holst JJ, et al. Type 2 diabetes mellitus. Nat Rev Dis Prim. 2015 Dec 23;1(1):15019. http://dx.doi.org/10.1038/nrdp.2015.19
3. Hanahan D, Weinberg RA. Hallmarks of cancer: The next generation. Cell. 2011 Mar 4;144(5):646–74. http://dx.doi.org/10.1016/j.cell.2011.02.013
4. Khor CC, Hibberd ML. Shared pathways to infectious disease susceptibility? Genome Med. 2010 Aug 10;2(8):52. http://dx.doi.org/10.1186/gm173
5. Cosselman KE, Navas-Acien A, Kaufman JD. Environmental factors in cardiovascular disease. Nat Rev Cardiol. 2015 Nov 13;12(11):627–42. http://dx.doi.org/10.1038/nrcardio.2015.152
6. Weaver JMJ, Ross-Innes CS, Fitzgerald RC. The "–omics" revolution and oesophageal adenocarcinoma. Nat Rev Gastroenterol Hepatol. 2014 Jan 27;11(1):19–27. http://dx.doi.org/10.1038/nrgastro.2013.150
7. Ideker T, Galitski T, Hood L. A new approach to decoding life: Systems Biology. Annu Rev Genomics Hum Genet. 2001 Sep 28;2(1):343–72. http://dx.doi.org/10.1146/annurev.genom.2.1.343
8. Nakaya HI, Li S, Pulendran B. Systems vaccinology: Learning to compute the behavior of vaccine induced immunity. Wiley Interdiscip Rev Syst Biol Med. 2012;4(2):193–205. http://dx.doi.org/10.1002/wsbm.163
9. Ko T-M, Chang J-S, Chen S-P, Liu Y-M, Chang C-J, Tsai F-J, et al. Genome-wide transcriptome analysis to further understand neutrophil activation and lncRNA transcript profiles in Kawasaki disease. Sci Rep. 2019 Dec 23;9(1):328. http://dx.doi.org/10.1038/s41598-018-36520-y
10. Gao X, Ke C, Liu H, Liu W, Li K, Yu B, et al. Large-scale metabolomic analysis reveals potential biomarkers for early stage coronary atherosclerosis. Sci Rep. 2017 Dec 18;7(1):11817. http://dx.doi.org/10.1038/s41598-017-12254-1

11. Wingo AP, Dammer EB, Breen MS, Logsdon BA, Duong DM, Troncosco JC, et al. Large-scale proteomic analysis of human brain identifies proteins associated with cognitive trajectory in advanced age. Nat Commun. 2019 Dec 8;10(1):1619. http://dx.doi.org/10.1038/s41467-019-09613-z
12. Chen H, Albergante L, Hsu JY, Lareau CA, Lo Bosco G, Guan J, et al. Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM. Nat Commun. 2019 Dec 23;10(1):1903. http://dx.doi.org/10.1038/s41467-019-09670-4
13. Gosak M, Markovič R, Dolenšek J, Slak Rupnik M, Marhl M, Stožer A, et al. Network science of biological systems at different scales: A review. Phys Life Rev. 2018 Mar;24:118–35. http://dx.doi.org/10.1016/j.plrev.2017.11.003
14. Srivastava A, George J, Karuturi RKM. Transcriptome analysis. In: Ranganathan S, Gribskov M, Nakai K, Schönbach C, editors. Encyclopedia of bioinformatics and computational biology, vol. 3. 1st ed., Cambridge, MA: Elsevier, 2019. p. 729–805.
15. Yan J, Risacher SL, Shen L, Saykin AJ. Network approaches to systems biology analysis of complex disease: Integrative methods for multi-omics data. Brief Bioinform. 2017 Jun 30;19(6):1370–81. http://dx.doi.org/10.1093/bib/bbx066
16. Costa-Silva J, Domingues D, Lopes FM. RNA-Seq differential expression analysis: An extended review and a software tool. Wei Z, editor. PLoS One. 2017 Dec 21;12(12):e0190152. http://dx.doi.org/10.1371/journal.pone.0190152
17. Jiang M, Zhang S, Yang Z, Lin H, Zhu J, Liu L, et al. Self-recognition of an inducible host lncRNA by RIG-I feedback restricts innate immune response. Cell. 2018 May;173(4):906–19.e13. http://dx.doi.org/10.1016/j.cell.2018.03.064
18. Liu C-X, Li X, Nan F, Jiang S, Gao X, Guo S-K, et al. Structure and degradation of circular RNAs regulate PKR activation in innate immunity. Cell. 2019 May;177(4):865–80.e21. http://dx.doi.org/10.1016/j.cell.2019.03.046
19. Cristinelli S, Ciuffi A. The use of single-cell RNA-Seq to understand virus–host interactions. Curr Opin Virol. 2018 Apr;29:39–50. http://dx.doi.org/10.1016/j.coviro.2018.03.001
20. Wu AR, Wang J, Streets AM, Huang Y. Single-cell transcriptional analysis. Annu Rev Anal Chem. 2017 Jun;10(1):439–62. http://dx.doi.org/10.1146/annurev-anchem-061516-045228
21. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-seq whole-transcriptome analysis of a single cell. Nat Methods. 2009 May 6;6(5):377–82. http://dx.doi.org/10.1038/nmeth.1315
22. Stubbington MJT, Lönnberg T, Proserpio V, Clare S, Speak AO, Dougan G, et al. T cell fate and clonality inference from single-cell transcriptomes. Nat Methods. 2016 Apr 7;13(4):329–32. http://dx.doi.org/10.1038/nmeth.3800
23. See P, Lum J, Chen J, Ginhoux F. A Single-cell sequencing guide for immunologists. Front Immunol. 2018 Oct 23;9:2425. http://dx.doi.org/10.3389/fimmu.2018.02425
24. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell. 2015 May;161(5):1202–14. http://dx.doi.org/10.1016/j.cell.2015.05.002
25. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015 May;161(5):1187–201. http://dx.doi.org/10.1016/j.cell.2015.04.044
26. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017 Apr 16;8(1):14049. http://dx.doi.org/10.1038/ncomms14049
27. Zhang X, Li T, Liu F, Chen Y, Yao J, Li Z, et al. Comparative analysis of droplet-based ultra-high-throughput single-cell RNA-seq systems. Mol Cell. 2019 Jan;73(1):130–42.e5. http://dx.doi.org/10.1016/j.molcel.2018.10.020
28. Zhao Z, Meng F, Wang W, Wang Z, Zhang C, Jiang T. Comprehensive RNA-seq transcriptomic profiling in the malignant progression of gliomas. Sci Data. 2017 Apr 14;4(1):170024. http://dx.doi.org/10.1038/sdata.2017.24
29. Aevermann BD, Pickett BE, Kumar S, Klem EB, Agnihothram S, Askovich PS, et al. A comprehensive collection of systems biology data characterizing the host response to viral infection. Sci Data. 2014 Dec 14;1(1):140033. http://dx.doi.org/10.1038/sdata.2014.33
30. Salomon MP, Orozco JIJ, Wilmott JS, Hothi P, Manughian-Peter AO, Cobbs CS, et al. Brain metastasis DNA methylomes, a novel resource for the identification of biological and clinical features. Sci Data. 2018 Nov 6;5:180245. http://dx.doi.org/10.1038/sdata.2018.245

31. Liu X, Ma Y, Yin K, Li W, Chen W, Zhang Y, et al. Long non-coding and coding RNA profiling using strand-specific RNA-seq in human hypertrophic cardiomyopathy. Sci Data. 2019 Dec 13;6(1):90. http://dx.doi.org/10.1038/sdata.2018.245
32. Wang M, Beckmann ND, Roussos P, Wang E, Zhou X, Wang Q, et al. The Mount Sinai cohort of large-scale genomic, transcriptomic and proteomic data in Alzheimer's disease. Sci Data. 2018 Sep 11;5:180185. http://dx.doi.org/10.1038/sdata.2018.185
33. Pechal JL, Schmidt CJ, Jordan HR, Benbow ME. A large-scale survey of the postmortem human microbiome, and its potential to provide insight into the living health condition. Sci Rep. 2018 Dec 10;8(1):5724. http://dx.doi.org/10.1038/s41598-018-23989-w
34. Haile S, Corbett RD, Bilobram S, Bye MH, Kirk H, Pandoh P, et al. Sources of erroneous sequences and artifact chimeric reads in next generation sequencing of genomic DNA from formalin-fixed paraffin-embedded samples. Nucleic Acids Res. 2019 Jan 25;47(2):e12–e12. http://dx.doi.org/10.1093/nar/gky1142
35. Baker M. Statisticians issue warning over misuse of P values. Nature. 2016 Mar 7;531(7593):151. http://dx.doi.org/10.1038/nature.2016.19503
36. Wang Y, Sun M. Transcriptome data analysis: Methods and protocols. New York: Humana Press, Springer; 2018. 238 p.
37. Kuleshov M V., Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016 Jul 8;44(W1):W90–7.
38. Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010 Jan 1;26(1):139–40. http://dx.doi.org/10.1093/bioinformatics/btp616
39. Olsen LR, Leipold MD, Pedersen CB, Maecker HT. The anatomy of single cell mass cytometry data. Cytometry A. 2019 Feb;95(2):156–72. http://dx.doi.org/10.1002/cyto.a.23621
40. van Unen V, Li N, Molendijk I, Temurhan M, Höllt T, van der Meulen-de Jong AE, et al. Mass cytometry of the human mucosal immune system identifies tissue- and disease-associated immune subsets. Immunity. 2016 May;44(5):1227–39. http://dx.doi.org/10.1016/j.immuni.2016.04.014
41. Nowicka M, Krieg C, Crowell HL, Weber LM, Hartmann FJ, Guglietta S, et al. CyTOF workflow: Differential discovery in high-throughput high-dimensional cytometry datasets. F1000Research. 2019 May 24;6:748. http://dx.doi.org/10.12688/f1000research.11622.3
42. Azad RK, Shulaev V. Metabolomics technology and bioinformatics for precision medicine. Brief Bioinform. 2018 Jan 3;bbx170:1–15. http://dx.doi.org/10.1093/bib/bbx170
43. Chong J, Soufan O, Li C, Caraus I, Li S, Bourque G, et al. MetaboAnalyst 4.0: Towards more transparent and integrative metabolomics analysis. Nucleic Acids Res. 2018 Jul 2;46(W1):W486–94. http://dx.doi.org/10.1093/nar/gky310
44. Chong J, Yamamoto M, Xia J. MetaboAnalystR 2.0: From raw spectra to biological insights. Metabolites. 2019 Mar 22;9(3):57. http://dx.doi.org/10.3390/metabo9030057
45. Hicks SC, Townes FW, Teng M, Irizarry RA. Missing data and technical variability in single-cell RNA-sequencing experiments. Biostatistics. 2018 Oct;19(4):562–78. http://dx.doi.org/10.1093/biostatistics/kxx053
46. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015 Apr;43(7):e47–e47. http://dx.doi.org/10.1093/nar/gkv007
47. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007 Jan 1;8(1):118–27. http://dx.doi.org/10.1093/biostatistics/kxj037
48. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. Nat Biotechnol. 2014 Sep;32(9):896–902. http://dx.doi.org/10.1038/nbt.2931
49. Leek JT. svaseq: Removing batch effects and other unwanted noise from sequencing data. Nucleic Acids Res. 2014 Dec;42(21):e161. http://dx.doi.org/10.1093/nar/gku864
50. Haghverdi L, Lun ATL, Morgan MD, Marioni JC. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol. 2018 May 2;36(5):421–7. http://dx.doi.org/10.1038/nbt.4091
51. Andrews TS, Hemberg M. Identifying cell populations with scRNASeq. Mol Aspects Med. 2018 Feb;59:114–22. http://dx.doi.org/10.1016/j.mam.2017.07.002
52. Becht E, McInnes L, Healy J, Dutertre C-A, Kwok IWH, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol. 2019 Jan;37(1):38–44. http://dx.doi.org/10.1038/nbt.4314

53. Chen G, Ning B, Shi T. Single-Cell RNA-Seq Technologies and related computational data analysis. Front Genet. 2019 Apr 5;10:317. http://dx.doi.org/10.3389/fgene.2019.00317
54. Björklund ÅK, Forkel M, Picelli S, Konya V, Theorell J, Friberg D, et al. The heterogeneity of human CD127+ innate lymphoid cells revealed by single-cell RNA sequencing. Nat Immunol. 2016 Apr;17(4):451–60. http://dx.doi.org/10.1038/ni.3368
55. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med. 2018 Aug 7;50(8):96. http://dx.doi.org/10.1038/s12276-018-0071-8
56. Griffiths JA, Scialdone A, Marioni JC. Using single-cell genomics to understand developmental processes and cell fate decisions. Mol Syst Biol. 2018 Apr 16;14(4):e8046. http://dx.doi.org/10.15252/msb.20178046
57. Cheng Y, Zhu YO, Becht E, Aw P, Chen J, Poidinger M, et al. Multifactorial heterogeneity of virus-specific T cells and association with the progression of human chronic hepatitis B infection. Sci Immunol. 2019 Feb 8;4(32):eaau6905. http://dx.doi.org/10.1126/sciimmunol.aau6905
58. Ji Q, Zheng Y, Zhang G, Hu Y, Fan X, Hou Y, et al. Single-cell RNA-seq analysis reveals the progression of human osteoarthritis. Ann Rheum Dis. 2019 Jan;78(1):100–10. http://dx.doi.org/10.1136/annrheumdis-2017-212863
59. van den Heuvel A, Mahfouz A, Kloet SL, Balog J, van Engelen BGM, Tawil R, et al. Single-cell RNA sequencing in facioscapulohumeral muscular dystrophy disease etiology and development. Hum Mol Genet. 2019 Apr 1;28(7):1064–75. http://dx.doi.org/10.1093/hmg/ddy400
60. Lang C, Campbell KR, Ryan BJ, Carling P, Attar M, Vowles J, et al. Single-cell sequencing of iPSC-dopamine neurons reconstructs disease progression and identifies HDAC4 as a regulator of Parkinson cell phenotypes. Cell Stem Cell. 2019 Jan;24(1):93–106.e6. http://dx.doi.org/10.1016/j.stem.2018.10.023
61. Finotello F, Trajanoski Z. Quantifying tumor-infiltrating immune cells from transcriptomics data. Cancer Immunol Immunother. 2018 Jul 14;67(7):1031–40. http://dx.doi.org/10.1007/s00262-018-2150-z
62. Li B, Li T, Wang B, Dou R, Zhang J, Liu JS, et al. Ultrasensitive detection of TCR hypervariable-region sequences in solid-tissue RNA–seq data. Nat Genet. 2017 Apr 1;49(4):482–3. http://dx.doi.org/10.1038/ng.3820
63. Mose LE, Selitsky SR, Bixby LM, Marron DL, Iglesia MD, Serody JS, et al. Assembly-based inference of B-cell receptor repertoires from short read RNA sequencing data with V'DJer. Bioinformatics. 2016 Dec 15;32(24):3729–34. http://dx.doi.org/10.1093/bioinformatics/btw526
64. Bolotin DA, Poslavsky S, Davydov AN, Frenkel FE, Fanchi L, Zolotareva OI, et al. Antigen receptor repertoire profiling from RNA-seq data. Nat Biotechnol. 2017 Oct 11;35(10):908–11. http://dx.doi.org/10.1038/nbt.3979
65. Canzar S, Neu KE, Tang Q, Wilson PC, Khan AA. BASIC: BCR assembly from single cells. Bioinformatics. 2016 Oct 2;btw631. http://dx.doi.org/10.1093/bioinformatics/btw631
66. Nakaya HI. Systems biology of infectious diseases and vaccines. In: Eils R, Kriete A, editors. Computational Systems Biology. 2nd ed., San Diego, CA: Academic Press, 2014. p. 331–58.
67. CRICK F. Central dogma of molecular biology. Nature. 1970 Aug;227(5258):561–3. http://dx.doi.org/10.1038/227561a0
68. Anderson L, Seilhamer J. A comparison of selected mRNA and protein abundances in human liver. Electrophoresis. 1997;18(3–4):533–7. http://dx.doi.org/10.1002/elps.1150180333
69. Liu Y, Beyer A, Aebersold R. On the dependency of cellular protein levels on mRNA abundance. Cell. 2016 Apr;165(3):535–50. http://dx.doi.org/10.1016/j.cell.2016.03.014
70. Wang L, Zhu J, Deng F-Y, Wu L-F, Mo X-B, Zhu X-W, et al. Correlation analyses revealed global microRNA–mRNA expression associations in human peripheral blood mononuclear cells. Mol Genet Genomics. 2018 Feb 6;293(1):95–105. http://dx.doi.org/10.1007/s00438-017-1367-4
71. Nunez YO, Truitt JM, Gorini G, Ponomareva ON, Blednov YA, Harris R, et al. Positively correlated miRNA-mRNA regulatory networks in mouse frontal cortex during early stages of alcohol dependence. BMC Genomics. 2013;14(1):725. http://dx.doi.org/10.1186/1471-2164-14-725
72. Olsen NJ, Moore JH, Aune TM. Gene expression signatures for autoimmune disease in peripheral blood mononuclear cells. Arthritis Res Ther. 2004;6(3):120–8. http://dx.doi.org/10.1186/ar1190
73. Kumar V, Kumar V, Chaudhary AK, Coulter DW, McGuire T, Mahato RI. Impact of miRNA-mRNA profiling and their correlation on medulloblastoma tumorigenesis. Mol Ther Nucleic Acids. 2018 Sep;12:490–503. http://dx.doi.org/10.1016/j.omtn.2018.06.004

74. Nakaya HI, Hagan T, Duraisingham SS, Lee EK, Kwissa M, Rouphael N, et al. Systems analysis of immunity to influenza vaccination across multiple years and in diverse populations reveals shared molecular signatures. Immunity. 2015 Dec;43(6):1186–98. http://dx.doi.org/10.1016/j.immuni.2015.11.012
75. Gardinassi LG, Arévalo-Herrera M, Herrera S, Cordy RJ, Tran V, Smith MR, et al. Integrative metabolomics and transcriptomics signatures of clinical tolerance to Plasmodium vivax reveal activation of innate cell immunity and T cell signaling. Redox Biol. 2018 Jul;17:158–70. http://dx.doi.org/10.1016/j.redox.2018.04.011
76. Cruickshank-Quinn CI, Jacobson S, Hughes G, Powell RL, Petrache I, Kechris K, et al. Metabolomics and transcriptomics pathway approach reveals outcome-specific perturbations in COPD. Sci Rep. 2018 Dec 20;8(1):17132. http://dx.doi.org/10.1016/j.redox.2018.04.011
77. Manolio TA. Genomewide association studies and assessment of the risk of disease. N Engl J Med. 2010 Jul;363(2):166–76. http://dx.doi.org/10.1056/NEJMra0905980
78. Gamazon ER, Huang RS, Cox NJ, Dolan ME. Chemotherapeutic drug susceptibility associated SNPs are enriched in expression quantitative trait loci. Proc Natl Acad Sci. 2010 May 18;107(20):9287–92. http://dx.doi.org/10.1073/pnas.1001827107
79. Gerrits A, Li Y, Tesson BM, Bystrykh L V., Weersing E, Ausema A, et al. Expression quantitative trait loci are highly sensitive to cellular differentiation state. PLoS Genet. 2009 Oct;5(10):e1000692. http://dx.doi.org/10.1073/pnas.1001827107
80. Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. Mapping complex disease traits with global gene expression. Nat Rev Genet. 2009 Mar;10(3):184–94. http://dx.doi.org/10.1038/nrg2537
81. Kumar D, Bansal G, Narang A, Basak T, Abbas T, Dash D. Integrating transcriptome and proteome profiling: Strategies and applications. Proteomics. 2016 Oct;16(19):2533–44. http://dx.doi.org/10.1002/pmic.201600140
82. Nesvizhskii AI. Proteogenomics: Concepts, applications and computational strategies. Nat Methods. 2014 Nov 30;11(11):1114–25. http://dx.doi.org/10.1038/nmeth.3144
83. Mun D-G, Bhin J, Kim S, Kim H, Jung JH, Jung Y, et al. Proteogenomic characterization of human early-onset gastric cancer. Cancer Cell. 2019 Jan;35(1):111–24.e10. http://dx.doi.org/10.1016/j.ccell.2018.12.003
84. Vasaikar S, Huang C, Wang X, Petyuk VA, Savage SR, Wen B, et al. Proteogenomic analysis of human colon cancer reveals new therapeutic opportunities. Cell. 2019 May;177(4):1035–49.e19.
85. Langley RJ, Tsalik EL, Velkinburgh JC, Glickman SW, Rice BJ, Wang C, et al. An integrated clinico-metabolomic model improves prediction of death in sepsis. Sci Transl Med. 2013 Jul 24;5(195):195ra95. http://dx.doi.org/10.1126/scitranslmed.3005893
86. Li S, Sullivan NL, Rouphael N, Yu T, Banton S, Maddur MS, et al. Metabolic Phenotypes of Response to Vaccination in Humans. Cell. 2017 May;169(5):862–77.e17. http://dx.doi.org/10.1126/scitranslmed.3005893
87. Johnson E, Bierut L, Cox N. Integrative omics in psychiatric diseases: Tools for discovery and understanding biology. Eur Neuropsychopharmacol. 2019;29:S741–2. http://dx.doi.org/10.1016/j.euroneuro.2017.06.073
88. Hobbs BD, Chimakurthi L, Morrow JD, Wang X, Liu Y-Y, Celli BR, et al. Integrative omics to discover novel subtypes in a chronic obstructive pulmonary disease lung tissue cohort. Am J Respir Crit Care Med. 2019;199:A6092. http://dx.doi.org/10.1164/ajrccm-conference.2019.199.1_MeetingAbstracts.A6092
89. Leon-Mimila P, Wang J, Huertas-Vazquez A. Relevance of multi-omics studies in cardiovascular diseases. Front Cardiovasc Med. 2019 Jul 17;6:91. http://dx.doi.org/10.3389/fcvm.2019.00091
90. Segal JP, Mullish BH, Quraishi MN, Acharjee A, Williams HRT, Iqbal T, et al. The application of omics techniques to understand the role of the gut microbiota in inflammatory bowel disease. Therap Adv Gastroenterol. 2019 Jan 24;12:175628481882225. http://dx.doi.org/10.1177/1756284818822250
91. Wu C, Huang BE, Chen G, Lovenberg TW, Pocalyko DJ, Yao X. Integrative analysis of diseaseland omics database for disease signatures and treatments: A bipolar case study. Front Genet. 2019 Apr 30;10:396. http://dx.doi.org/10.3389/fgene.2019.00396
92. Yu XT, Zeng T. Integrative analysis of omics big data. In: Huang T, editor. Computational Systems Biology. Methods Mol Biol. New York, NY: Humana Press; 2018;1754:109–35.
93. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol. 2017 Dec;18(1):83. http://dx.doi.org/10.1186/s13059-017-1215-1

94. Sun YV, Hu Y-J. Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases. Adv Genet. 2016;93:147–90.
95. Russo PST, Ferreira GR, Cardozo LE, Bürger MC, Arias-Carrasco R, Maruyama SR, et al. CEMiTool: A bioconductor package for performing comprehensive modular co-expression analyses. BMC Bioinformatics. 2018 Dec 20;19(1):56. http://dx.doi.org/10.1186/s12859-018-2053-1
96. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci. 2005 Oct;102(43):15545–50. http://dx.doi.org/10.1073/pnas.0506580102
97. Rohart F, Gautier B, Singh A, Lê Cao K-A. mixOmics: An R package for 'omics feature selection and multiple data integration. PLoS Comput Biol. 2017 Nov 3;13(11):e1005752. http://dx.doi.org/10.1371/journal.pcbi.1005752
98. Singh A, Shannon CP, Gautier B, Rohart F, Vacher M, Tebbutt SJ, et al. DIABLO: An integrative approach for identifying key molecular drivers from multi-omics assays. Bioinformatics. 2019 Sep 1; 35(17):3055–62. http://dx.doi.org/10.1093/bioinformatics/bty1054
99. Rohart F, Eslami A, Matigian N, Bougeard S, Lê Cao K-A. MINT: A multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms. BMC Bioinformatics. 2017 Dec 27;18(1):128. http://dx.doi.org/10.1186/s12859-017-1553-8
100. Geng P, Tong X, Lu Q. An integrative U method for joint analysis of multi-level omic data. BMC Genet. 2019 Dec 10;20(1):40. http://dx.doi.org/10.1186/s12863-019-0742-z
101. Ickstadt K, Schäfer M, Zucknick M. Toward integrative Bayesian analysis in molecular biology. Annu Rev Stat Its Appl. 2018 Mar 7;5(1):141–67. http://dx.doi.org/10.1146/annurev-statistics-031017-100438
102. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, et al. Multi-Omics factor analysis—A framework for unsupervised integration of multi-omics data sets. Mol Syst Biol. 2018 Jun 20;14(6): e8124. http://dx.doi.org/10.15252/msb.20178124
103. Speicher NK, Pfeifer N. Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. Bioinformatics. 2015 Jun 15;31(12):i268–75. http://dx.doi.org/10.1093/bioinformatics/btv244
104. Zhang L, Lv C, Jin Y, Cheng G, Fu Y, Yuan D, et al. Deep learning-based multi-omics data integration reveals two prognostic subtypes in high-risk neuroblastoma. Front Genet. 2018 Oct 18;9:477. http://dx.doi.org/10.3389/fgene.2018.00477
105. Narayanan M, Vetta A, Schadt EE, Zhu J. Simultaneous clustering of multiple gene expression and physical interaction datasets. PLoS Comput Biol. 2010 Apr 15;6(4):e1000742. http://dx.doi.org/10.1371/journal.pcbi.1000742
106. D'Argenio V. The high-throughput analyses era: Are we ready for the data struggle? High Throughput. 2018 Mar 2;7(1):8. http://dx.doi.org/10.1371/journal.pcbi.1000742
107. Palsson B, Zengler K. The challenges of integrating multi-omic data sets. Nat Chem Biol. 2010 Nov 18;6(11):787–9. http://dx.doi.org/10.1038/nchembio.462
108. Nekrutenko A, Taylor J. Next-generation sequencing data interpretation: Enhancing reproducibility and accessibility. Nat Rev Genet. 2012 Sep 17;13(9):667–72. http://dx.doi.org/10.1038/nrg3305
109. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data. 2016 Dec 15;3(1):160018.
110. Pinu FR, Beale DJ, Paten AM, Kouremenos K, Swarup S, Schirra HJ, et al. Systems biology and multi-omics integration: Viewpoints from the metabolomics research community. Metabolites. 2019 Apr;9(4):76. http://dx.doi.org/10.3390/metabo9040076
111. Misra BB, Langefeld CD, Olivier M, Cox LA. Integrated omics: Tools, advances, and future approaches. J Mol Endocrinol. 2019 Jan 13;62:R21-R45. https://doi.org/10.1530/JME-18-0055.
112. Altman N, Krzywinski M. Association, correlation and causation. Nat Methods. 2015 Oct 29; 12(10):899–900. http://dx.doi.org/10.1038/nmeth.3587
113. Yin Z, Lan H, Tan G, Lu M, Vasilakos AV, Liu W. Computing platforms for big biological data analytics: Perspectives and challenges. Comput Struct Biotechnol J. 2017;15:403–11. http://dx.doi.org/10.1016/j.csbj.2017.07.004