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
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).
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 (Inc-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
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

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

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


Integrative Biology Approaches


