Computational Molecular Biology 2024, Vol.14, No.2, 64-75
http://bioscipublisher.com/index.php/cmb

Feature Review Open Access

Emerging Trends in Multi-Omics Data Integration: Challenges and Future Directions

Jie Zhang
Institute of Life Sciences, Jiyang Colloge of Zhejiang A&F University, Zhuji, 311800, Zhejiang, China
Corresponding email: jie.zhang@jicat.org
Computational Molecular Biology, 2024, Vol.14, No.2 doi: 10.5376/cmb.2024.14.0008
Received: 09 Feb., 2024
Accepted: 20 Mar., 2024
Published: 07 Apr., 2024

Preferred citation for this article:
Zhang J., 2024, Emerging Trends in multi-omics data integration: challenges and future directions, Computational Molecular Biology, 14(2): 64-75 (doi: 10.5376/cmb.2024.14.0008)

Abstract This study analyzed the latest trends, challenges, and future directions of multi omics data integration. High throughput technology enables the generation of large amounts of data at multiple omics levels, including genomics, transcriptomics, proteomics, and metabolomics. However, integrating these heterogeneous datasets faces significant challenges due to differences in data types, dimensions, and a lack of standardized analysis protocols. We discussed various integration methods, including data-driven, knowledge driven, and machine learning approaches, with a focus on their applications in disease subtype classification, biomarker discovery, and precision medicine. In addition, we also analyzed the computational and visualization challenges associated with single-cell multi omics data and proposed future directions for developing stronger and more interpretable integration strategies, hoping to provide a comprehensive overview of the current status of multi omics data integration and demonstrate its potential in translational biomedical research and clinical practice.
Keywords Multi-omics integration; High-throughput technologies; Machine learning; Precision medicine; Single-cell analysis

1 Introduction

The advent of high-throughput technologies has revolutionized the field of biological research, enabling the comprehensive profiling of various molecular layers within biological systems. These layers include genomics, transcriptomics, proteomics, metabolomics, and more recently, single-cell omics (Misra et al., 2019; Miao et al., 2021; Wörheide et al., 2021). Multi-omics approaches aim to integrate these diverse datasets to provide a holistic view of the complex molecular interactions and regulatory mechanisms that underpin biological processes and disease states (Ebrahim et al., 2016; Colomé-Tatché and Theis, 2018; Wörheide et al., 2021). The integration of multi-omics data allows researchers to uncover hidden biological regularities and gain deeper insights into cellular functions and physiological responses (Ebrahim et al., 2016; Santiago-Rodriguez and Hollister, 2021).

The integration of multi-omics data is crucial for advancing our understanding of complex biological systems. By combining data from different omics layers, researchers can achieve a more comprehensive and nuanced understanding of the molecular underpinnings of health and disease (Misra et al., 2019; Wörheide et al., 2021; Agamah et al., 2022). This integrative approach facilitates the identification of novel biomarkers, disease subtypes, and therapeutic targets, thereby enhancing the precision and efficacy of medical interventions (Graw et al., 2020; Santiago-Rodriguez and Hollister, 2021). Moreover, multi-omics data integration helps in overcoming the limitations of individual omics datasets, such as data heterogeneity and high dimensionality, by providing a more robust and contextually relevant analysis (Misra et al., 2019; Sokač et al., 2023).

This study provides a comprehensive overview of emerging trends in multi omics data integration, with a focus on the current challenges and future directions in this rapidly developing field. Researchers discuss various integration methods, including data-driven, knowledge-based, simultaneous, and step-by-step approaches, and their applications in recent multi omics studies. In addition, computational and statistical tools developed for the integration of multi omics data will be explored, emphasizing their advantages, limitations, and potential for standardization.

2 Overview of Multi-Omics Data Types

2.1 Genomics

Genomics is the study of the complete set of DNA (the genome) in an organism, including its structure, function, evolution, and mapping. Genomic data typically involve DNA sequences, single nucleotide polymorphisms (SNPs), copy number variations (CNVs), and other genetic variations. The advent of high-throughput sequencing technologies has revolutionized genomics, enabling the generation of vast amounts of data that can be used to identify genetic factors associated with diseases, understand evolutionary relationships, and explore genetic diversity within and between populations (Ritchie et al., 2015; Wörheide et al., 2021) Genomic data serve as the foundation for other omics layers, providing the blueprint for the synthesis of RNA and proteins. Integrating genomics with other omics data can reveal how genetic variations influence gene expression, protein function, and metabolic pathways, thereby offering insights into the molecular mechanisms underlying phenotypic traits and disease states (Manzoni et al., 2016; Misra et al., 2019).

2.2 Transcriptomics

Transcriptomics involves the study of the complete set of RNA transcripts produced by the genome under specific circumstances or in a specific cell. This includes messenger RNA (mRNA), ribosomal RNA (rRNA), transfer RNA (tRNA), and non-coding RNAs. Transcriptomic data provide a snapshot of gene expression levels, reflecting which genes are active and to what extent in different tissues, developmental stages, or environmental conditions (Wörheide et al., 2021; Santiago-Rodriguez and Hollister, 2021). High-throughput RNA sequencing (RNA-seq) has become the standard method for transcriptomic analysis, allowing for the quantification of gene expression and the identification of novel transcripts and alternative splicing events. Integrating transcriptomic data with genomic and proteomic data can help elucidate the regulatory mechanisms controlling gene expression and how these are altered in disease states (Manzoni et al., 2016; Zhang et al., 2019).

2.3 Proteomics

Proteomics is the large-scale study of proteins, which are the functional molecules in cells. Proteomic data include information on protein expression levels, post-translational modifications, protein-protein interactions, and protein localization. Proteins are the direct effectors of cellular functions, and their study is crucial for understanding the biochemical activities within cells (Zhang et al., 2019; Wörheide et al., 2021). Mass spectrometry (MS) and protein microarrays are commonly used techniques in proteomics. These methods can identify and quantify thousands of proteins in a single experiment. Integrating proteomic data with genomic and transcriptomic data can provide insights into how genetic and transcriptional changes are translated into functional outcomes at the protein level. This integration is essential for understanding the complex regulatory networks and pathways involved in cellular processes and disease mechanisms (Ritchie et al., 2015; Jendoubi, 2021).

2.4 Metabolomics

Metabolomics is the study of the complete set of metabolites (small molecules) within a biological sample. Metabolites are the end products of cellular processes and provide a direct readout of the biochemical activity within cells. Metabolomic data can reveal changes in metabolic pathways and networks in response to genetic, environmental, or physiological changes (Pinu et al., 2019; Wörheide et al., 2021).

Techniques such as nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS) are used to profile metabolites. Metabolomics is particularly valuable in multi-omics studies because it reflects the downstream effects of changes in the genome, transcriptome, and proteome. Integrating metabolomic data with other omics layers can help identify biomarkers for disease, understand metabolic dysregulation, and uncover the biochemical basis of phenotypic traits (Jendoubi, 2021; Santiago-Rodriguez and Hollister, 2021). Each type of omics data provides a unique layer of information about the biological system. Genomics offers insights into the genetic blueprint, transcriptomics reveals gene expression patterns, proteomics uncovers protein functions and interactions, and metabolomics reflects the biochemical activities within cells. Integrating these diverse data types is essential for a holistic understanding of biological systems and for advancing precision medicine (Manzoni et al., 2016; Misra et al., 2019; Zhang et al., 2019).

3 Methods for Multi-Omics Data Integration

3.1 Statistical methods

Statistical methods for multi-omics data integration are foundational and often serve as the first step in understanding complex biological systems. These methods include various forms of regression analysis, Bayesian statistics, and factor analysis. For instance, Bayesian models can incorporate prior knowledge and handle different data distributions, making them suitable for integrating heterogeneous omics data (Li et al., 2016). Additionally, methods like Multiple Canonical Correlation Analysis (MCCA) and Multiple Factor Analysis (MFA) are used to identify correlations and shared variations across different omics layers, facilitating a more comprehensive understanding of biological interactions (Tini et al., 2019).

3.2 Machine learning approaches

Machine learning (ML) approaches have gained significant traction in multi-omics data integration due to their ability to handle large, complex datasets and uncover hidden patterns. These approaches can be broadly categorized into supervised learning, unsupervised learning, and deep learning techniques. Supervised learning techniques are employed when the outcome variable is known, and the goal is to predict this outcome based on input features from multiple omics datasets. Common methods include support vector machines (SVMs), random forests, and various forms of regression analysis. For example, SVMs and random forests have been effectively used to integrate genomic, proteomic, and metabolomic data for disease prediction and biomarker discovery (Feldner-Busztin et al., 2023). These methods are particularly useful in precision medicine, where they can help in patient stratification and personalized treatment plans (Reel et al., 2021).

Unsupervised learning techniques are used when the outcome variable is unknown, and the goal is to uncover the underlying structure of the data. Clustering methods like k-means, hierarchical clustering, and more advanced techniques like Similarity Network Fusion (SNF) are commonly used. SNF, for instance, integrates multiple types of omics data by constructing similarity networks for each data type and then fusing them into a single network, which can reveal complex biological relationships (Figure 1) (Tini et al., 2019). Other methods like Multiple Co-Inertia Analysis (MCIA) and Joint and Individual Variation Explained (JIVE) are also employed to identify shared and unique variations across different omics datasets (Tini et al., 2019).

Deep learning methods, particularly neural networks, have shown great promise in multi-omics data integration due to their ability to model complex, non-linear relationships. Autoencoders, a type of neural network, are frequently used to transform multi-omics data into latent representations that capture essential features while reducing dimensionality (Hauptmann and Kramer, 2022). For example, methods like MOLI, Super.FELT, and OmiEmbed have been developed to integrate multi-omics data for drug response prediction, showing that deep learning can outperform traditional methods in certain contexts (Hauptmann and Kramer, 2022). Additionally, novel architectures like Omics Stacking combine the advantages of intermediate and late integration, further enhancing predictive performance (Figure 2) (Hauptmann and Kramer, 2022). Customizable deep learning strategies, such as CustOmics, adapt training to each data source independently before learning cross-modality interactions, providing interpretable results and high performance in tasks like tumor classification and survival outcome prediction (Benkirane et al., 2023).

3.3 Network-based methods

Network-based methods are another powerful approach for integrating multi-omics data. These methods construct networks where nodes represent biological entities (e.g., genes, proteins) and edges represent interactions or associations between them. Network-based diffusion/propagation methods can exploit information captured in each omics dataset to infer associations between different data types (Cominetti et al., 2023). For instance, network-based models have been used to integrate genomic, transcriptomic, and proteomic data to identify key regulatory pathways and potential therapeutic targets in cancer (Nicora et al., 2020). These methods can also incorporate external knowledge from biological databases, enhancing the robustness and interpretability of the results (Vahabi and Michailidis, 2022).

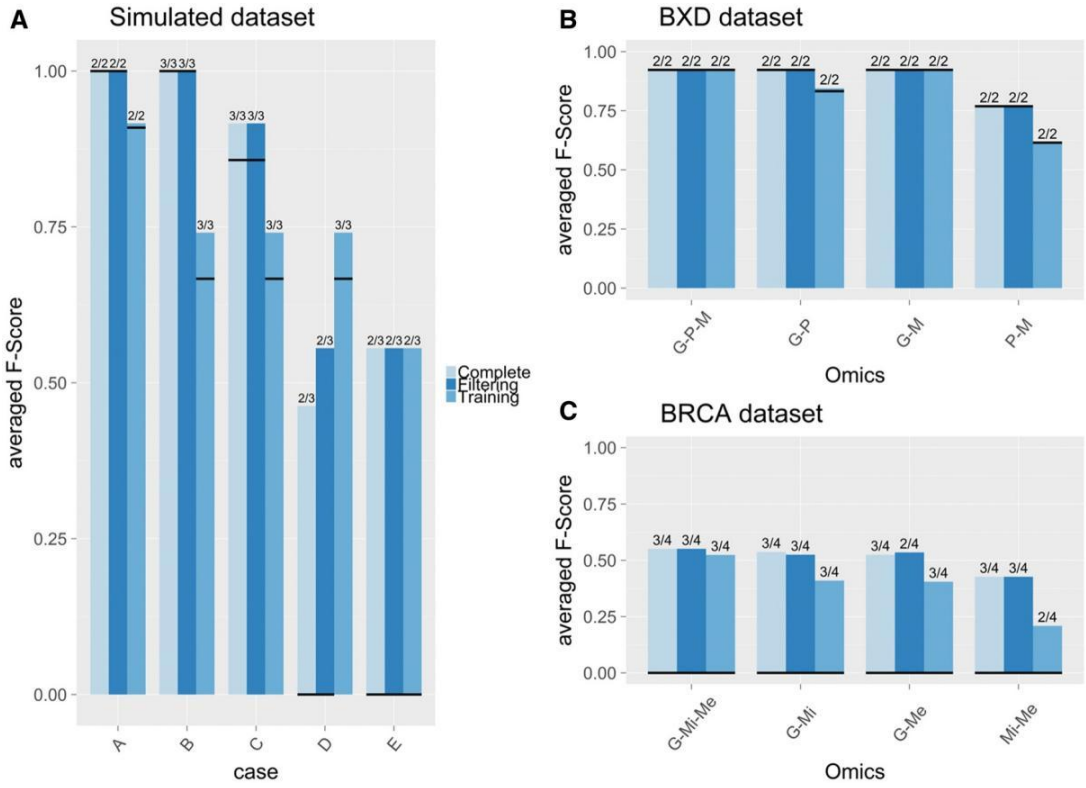

Figure 1 Comparison of SNF results on the validation sets by using default parameters before and after feature selection and by using trained parameters without feature selection (Adopted from Tini et al., 2019)

Image caption: Averaged F-scores obtained from the three analyses are represented with light-, dark- and medium-shaded bars, respectively. Minimum F-scores are represented with black lines. A minimum F-score equal to 0 indicates that not all the subtypes have been recognized. The number of subtypes recognized for each trial is added above the bars. (A) Simulated scenarios. (B) BXD data set (G: gene expression, P: proteins, M: metabolites). (C) BRCA data set (G: gene expression, Mi: miRNA, Me: methylation).

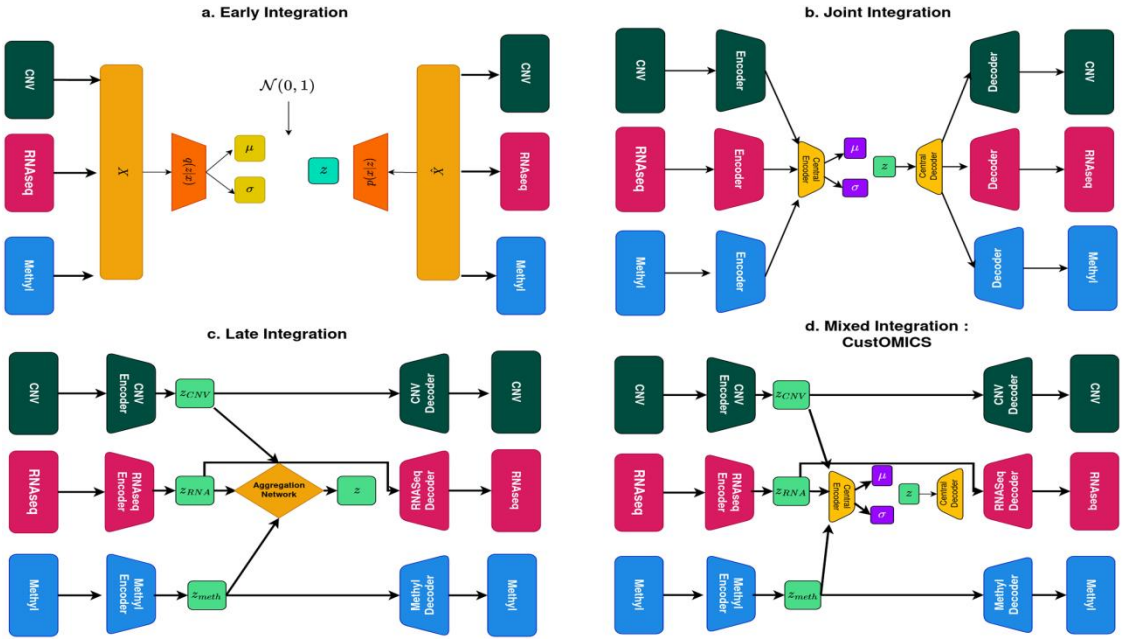

Figure 2 Four Strategies for Multi-Omics Data Integration: Early Integration, Joint Integration, Late Integration, and Mixed Integration (Adapted from Benkirane et al., 2023)

Image caption: a. Early Integration VAE: Variational Autoencoder architecture with early integration strategy. b. Joint Integration VAE: Variational Autoencoder architecture with joint integration strategy. c. Late Integration VAE: Variational Autoencoder architecture with late integration strategy.d. Mixed-Integration/CustOmics: This is a hierarchical architecture composed of specific per-source autoencoders that converges into a central variational autoencoder (Adapted from Benkirane et al., 2023)

4 Applications of Multi-Omics Data Integration

4.1 Precision medicine

Multi-omics data integration has revolutionized the field of precision medicine by enabling the development of personalized treatment strategies. By combining data from genomics, transcriptomics, proteomics, and metabolomics, researchers can gain a comprehensive understanding of the molecular mechanisms underlying individual diseases. This holistic approach allows for the identification of patient-specific therapeutic targets and the design of tailored treatment regimens. For instance, in oncology, multi-omics analyses have facilitated the stratification of patients based on molecular subtypes, leading to more effective and personalized cancer therapies (Nicora et al., 2020; Kang et al., 2021; Raufaste-Cazavieille et al., 2022). Additionally, the integration of multi-omics data with clinical information has been shown to improve the prediction of treatment responses and disease outcomes, further enhancing the precision of medical interventions (Reel et al., 2021; Terranova and Venkatakrishnan, 2023).

The discovery of reliable biomarkers is crucial for the early diagnosis, prognosis, and monitoring of diseases. Multi-omics data integration has emerged as a powerful tool for biomarker discovery, as it allows for the simultaneous analysis of multiple molecular layers. This approach can reveal complex interactions and regulatory networks that are not apparent when examining single-omics data alone. For example, in cancer research, multi-omics analyses have identified novel biomarkers that can predict disease progression and response to therapy, thereby guiding clinical decision-making (Olivier et al., 2019; Turanli et al., 2019; Demirel et al., 2021). Moreover, the integration of multi-omics data has led to the identification of biomarkers for various other diseases, including autoimmune disorders and cardiovascular diseases, highlighting its broad applicability in precision medicine (Subramanian et al., 2020; Reel et al., 2021).

Predictive modeling is a key component of precision medicine, as it enables the anticipation of disease trajectories and treatment outcomes. The integration of multi-omics data with advanced machine learning algorithms has significantly enhanced the accuracy and robustness of predictive models. These models can capture the complex, nonlinear relationships between different molecular entities and clinical variables, providing valuable insights into disease mechanisms and therapeutic responses (Kang et al., 2021; Cominetti et al., 2023; Terranova and Venkatakrishnan, 2023). For instance, in oncology, predictive models based on multi-omics data have been used to forecast patient survival, treatment efficacy, and potential adverse effects, thereby informing personalized treatment plans (Nicora et al., 2020; Raufaste-Cazavieille et al., 2022). Additionally, the use of multi-omics data in predictive modeling has been extended to other therapeutic areas, such as neurology and immunology, demonstrating its versatility and potential to transform clinical practice (Reel et al., 2021; Terranova and Venkatakrishnan, 2023).

4.2 Drug discovery and development

The integration of multi-omics data has also made significant contributions to drug discovery and development. By providing a comprehensive view of the molecular landscape of diseases, multi-omics analyses can identify novel drug targets and elucidate the mechanisms of action of existing drugs. This information is invaluable for the development of new therapeutic agents and the repurposing of existing drugs for new indications (Turanli et al., 2019; Demirel et al., 2021).

For example, network-based approaches that integrate multi-omics data have been used to identify key regulatory nodes in cancer and other diseases, leading to the discovery of potential drug targets (Turanli et al., 2019; Subramanian et al., 2020). Furthermore, multi-omics data integration can facilitate the optimization of drug dosing and the identification of biomarkers for patient stratification in clinical trials, thereby improving the efficiency and success rate of drug development (Kang et al., 2021; Terranova and Venkatakrishnan, 2023).

4.3 Understanding complex diseases

Complex diseases, such as cancer, cardiovascular diseases, and neurodegenerative disorders, are characterized by intricate molecular interactions and regulatory networks. Multi-omics data integration provides a powerful approach to unravel these complexities by combining information from different molecular layers. This holistic

perspective can reveal novel insights into disease etiology, progression, and heterogeneity, which are essential for the development of effective therapeutic strategies (Olivier et al., 2019; Nicora et al., 2020; Raufaste-Cazavieille et al., 2022). For instance, in cancer research, multi-omics analyses have uncovered the molecular diversity within tumors, leading to the identification of distinct molecular subtypes and the development of targeted therapies (Demirel et al., 2021; Raufaste-Cazavieille et al., 2022). Similarly, in neurodegenerative diseases, multi-omics data integration has shed light on the molecular mechanisms underlying disease onset and progression, paving the way for the development of novel diagnostic and therapeutic approaches (Reel et al., 2021; Terranova and Venkatakrishnan, 2023).The integration of multi-omics data is transforming our understanding of complex diseases and driving the advancement of precision medicine.

5 Challenges in Multi-Omics Data Integration

5.1 Data heterogeneity

One of the primary challenges in multi-omics data integration is the inherent heterogeneity of the data. Different omics layers, such as genomics, transcriptomics, proteomics, and metabolomics, each have unique characteristics, data formats, and scales, making their integration complex. For instance, genomic data is often discrete and categorical, while proteomic and metabolomic data are typically continuous and quantitative. This disparity necessitates sophisticated normalization and transformation techniques to harmonize the data before integration (Misra et al., 2019; Subramanian et al., 2020; Kaur et al., 2021).

Moreover, the nomenclature and identifiers used across different omics datasets can vary significantly, complicating the process of matching corresponding entities across datasets. For example, gene identifiers in genomic data may not directly correspond to protein identifiers in proteomic data, requiring extensive cross-referencing and mapping efforts (Misra et al., 2019). The heterogeneity also extends to the experimental designs and conditions under which the data are collected, adding another layer of complexity to the integration process (Bodein et al., 2020).

5.2 Computational complexity

The integration of multi-omics data is computationally intensive due to the high dimensionality and large volume of the datasets involved. High-throughput technologies generate vast amounts of data, often in the tera- to peta-byte range, which poses significant challenges for data storage, processing, and analysis (Misra et al., 2019; Lee et al., 2020). The computational burden is further exacerbated by the need to perform complex operations such as data cleaning, normalization, and dimensionality reduction before meaningful integration can occur.

Advanced computational methods, including machine learning and deep learning techniques, have been developed to address these challenges. However, these methods themselves are computationally demanding and require substantial computational resources and expertise to implement effectively (Nicora et al., 2020; Benkirane et al., 2023). For instance, deep learning models, while powerful, necessitate extensive training on large datasets, which can be time-consuming and resource-intensive (Benkirane et al., 2023). Additionally, the integration process often involves constructing and analyzing complex network models to represent the relationships between different omics layers. These network models, such as heterogeneous multi-layered networks (HMLNs), are computationally challenging to build and analyze due to their complexity and the need to infer novel biological relations from the integrated data (Lee et al., 2020).

5.3 Data interpretation and biological relevance

Even after successful integration, interpreting the integrated multi-omics data and deriving biologically relevant insights remain significant challenges. The complexity of the integrated data can obscure meaningful biological signals, making it difficult to draw clear conclusions about the underlying biological processes (Ebrahim et al., 2016; Subramanian et al., 2020). For example, while integrated data may reveal correlations between different omics layers, determining the causal relationships and biological significance of these correlations requires careful analysis and validation (Ebrahim et al., 2016). Furthermore, the interpretation of multi-omics data often relies on sophisticated visualization techniques to make the data comprehensible. However, current visualization methods

are still underdeveloped and may not adequately capture the complexity of the integrated data, limiting their utility in biological interpretation (Miao et al., 2020). Effective visualization tools are essential for identifying patterns and regularities in the data that can lead to new biological insights (Miao et al., 2020).

Another challenge is the need for biologically interpretable integration strategies that can account for the specific noise and variability associated with each omics modality. Developing such strategies requires a deep understanding of the biological context and the specific characteristics of each data type, which can be a daunting task (Bodein et al., 2020; Miao et al., 2020). Additionally, the integration process must be designed to ensure that the resulting models are not only accurate but also biologically meaningful and interpretable (Benkirane et al., 2023).

6 Technological Advances Supporting Integration

6.1 High-throughput sequencing technologies

High-throughput sequencing technologies have revolutionized the field of multi-omics by enabling the generation of vast amounts of data across various biological layers, including genomics, transcriptomics, proteomics, and metabolomics. Platforms such as Illumina, PacBio, and 10X Genomics have significantly reduced the cost and time required for sequencing, making it feasible to conduct large-scale studies that integrate multiple omics datasets (Miao et al., 2020). These technologies allow for the comprehensive profiling of biological systems, providing insights into the complex molecular mechanisms underlying health and disease (Wörheide et al., 2021).

The integration of high-throughput sequencing data poses several challenges, including data storage, normalization, and interpretation. The sheer volume of data generated can reach tera- to peta-byte scales, necessitating robust computational infrastructure for effective data management (Misra et al., 2019). Additionally, differences in data formats and nomenclature across various omics layers further complicate the integration process. Despite these challenges, high-throughput sequencing remains a cornerstone of multi-omics research, driving advancements in precision medicine and systems biology (Nicora et al., 2020; Wörheide et al., 2021).

6.2 Advanced bioinformatics tools

The rapid accumulation of multi-omics data has spurred the development of advanced bioinformatics tools designed to facilitate data integration, analysis, and visualization. Tools such as OmicsSuite offer comprehensive solutions for multi-omics analysis, integrating various statistical and computational methods to handle diverse data types (Miao et al., 2023). These tools are essential for addressing the complexities associated with multi-omics data, including high dimensionality, heterogeneity, and the need for accurate biomolecule identification and data normalization (Misra et al., 2019).

Machine learning and deep learning algorithms have emerged as powerful methods for multi-omics data integration. These algorithms can capture nonlinear and hierarchical features within the data, providing predictive insights and uncovering complex relationships between different molecular layers (Kang et al., 2021). For instance, deep learning has been successfully applied to integrate genomics, transcriptomics, and metabolomics data, enhancing our understanding of disease mechanisms and aiding in biomarker discovery (Nicora et al., 2020; Kang et al., 2021).

Network-based methods, such as heterogeneous multi-layered networks (HMLNs), have also proven effective in representing the hierarchical relationships within biological systems. HMLNs facilitate the integration of diverse omics data, enabling the inference of novel biological interactions and the establishment of causal genotype-phenotype associations (Lee et al., 2020). These advanced bioinformatics tools are crucial for transforming raw multi-omics data into actionable biological insights, driving forward the field of integrative biology (Ritchie et al., 2015; Subramanian et al., 2020).

6.3 Cloud computing and big data analytics

The advent of cloud computing and big data analytics has provided a scalable and flexible solution for managing the enormous datasets generated by high-throughput sequencing technologies. Cloud platforms offer significant

advantages in terms of data storage, processing power, and accessibility, making them ideal for multi-omics data integration and analysis (Koppad et al., 2021). By leveraging cloud computing, researchers can perform complex computational tasks without the need for extensive local infrastructure, reducing costs and accelerating the pace of discovery (Koppad et al., 2021).

Cloud-based bioinformatics applications have been developed to handle various aspects of multi-omics data analysis, including RNA sequencing, metabolomics, and proteomics. These applications facilitate the integration and interpretation of phenotypic data, providing a holistic view of biological systems (Koppad et al., 2021). Additionally, cloud computing enables the sharing and collaboration of large datasets across the global research community, fostering a more collaborative and open scientific environment (Koppad et al., 2021).

Big data analytics, combined with cloud computing, allows for the efficient processing and analysis of high-dimensional multi-omics data. Advanced analytical techniques, such as machine learning and deep learning, can be deployed on cloud platforms to uncover hidden patterns and relationships within the data (Kang et al., 2021). This integration of cloud computing and big data analytics is essential for overcoming the challenges associated with multi-omics data, paving the way for new discoveries in systems biology and precision medicine (Koppad et al., 2021; Kang et al., 2021).

7 Case Studies of Successful Multi-Omics Integration

7.1 Cancer research and treatment

The integration of multi-omics data has significantly advanced cancer research and treatment, providing a comprehensive understanding of the molecular underpinnings of various cancers. One notable example is the use of multi-omics approaches to achieve precision medicine in oncology. By integrating genomics, transcriptomics, proteomics, and metabolomics data, researchers have been able to classify tumors not just by their site of origin but by their molecular characteristics, leading to the concept of pan-cancer molecular classification (Table 1). This has opened new therapeutic opportunities and allowed for the identification of prognostic and treatment-specific biomarkers, which are crucial for personalized therapy (Nicora et al., 2020; Raufaste-Cazavieille et al., 2022).

For instance, the integration of multi-omics data has been pivotal in understanding the spatial and temporal heterogeneity of tumors. This approach has revealed the incredible complexity and molecular diversity within the same tumor type, which traditional single-omics studies could not capture. By combining different layers of biological data, researchers can now dissect the tumor immune environment and host-tumor interactions, providing insights that guide therapeutic decisions in immuno-oncology (Ning and He, 2021; Raufaste-Cazavieille et al., 2022).

7.2 Metabolic disease studies

Multi-omics integration has also been instrumental in studying metabolic diseases. The holistic approach of combining genomics, transcriptomics, proteomics, and metabolomics data allows for a comprehensive understanding of the metabolic pathways and their regulation. This is particularly important in diseases like diabetes and obesity, where multiple biological processes are dysregulated.

For example, a metabolomics-centric review highlighted the potential of integrating metabolomics data with other omics layers to uncover the complex molecular relationships within metabolic diseases. This approach has enabled researchers to identify novel biomarkers and therapeutic targets, which are essential for developing effective treatments (Wörheide et al., 2021). The integration of multi-omics data has also facilitated the study of metabolic fluxes and the interactions between different metabolic pathways, providing deeper insights into disease mechanisms. Furthermore, the development of computational tools and methods for multi-omics data integration has advanced the field of metabolic disease research. These tools help in addressing the challenges of data dimensionality and heterogeneity, enabling researchers to derive meaningful insights from complex datasets (Subramanian et al., 2020).

Table 1 Description of different tools for multi-omics integration with their application and their major strength and limits (Adopted from Raufaste-Cazavieille et al., 2022)

Method	Principle	Aim	Omics element	Pros	Cons
JIVE	Matrix factorization	Disease subtyping systemic knowledge, module detection	Genomics and epigenomics	Integrate large amount of data	Sensitive tooutliers and missing values
NMF	-	Disease subtyping module detection, biomarker discovery	Genomics and epigenomics	Filtering weak signal.Integrate large amount of data. Detection of cluster of small size	Time and memory consuming. Underperforming on missing values
nNMF	-	-	-	-	-
jNMF	-	-	-	-	-
intNMF	-	-	-	-	-
SLIDE	-	Disease subtyping, module detection, biomarker discovery	Genomics,epigenomics and proteomics	Integrate large amount of data	Underperforming with missing values. Optimum solution is not guaranteed
MALA	Logicdata mining	Sample classification	Genomics and transcriptomics	Works well on experimental data Integrate large amount of data	Phenotype number must be delivered with data. Sensitive to missing values
iCluster	Gaussian latent variable model	-	Genomics,epigenomics and transcriptomics	-	Needs to test a large amount of solution tofind the most relevant
iCluster+	Generalized linear regression	Disease subtyping	Genomics,transcriptomics, proteomics and epigenomics	Handle missing values	No evaluation of statistical significancefor selected features
iClusterBayes	Bayesian integrative clustering	Biomarker discovery	Genomics,transcriptomics, and epigenomics	Good performance in thepresence of explicative data	Underperform with outliers
MOFA	Bayesian factor analysis	Biomarker discovery, systemic knowledge	Proteomics,metabolomics and lipidomics	Handle well missing values	Linear model can miss linear relation
MOFA+	-	-	Genomics and epigenomics	The use of continuous learning enabling MOFA to recover different trajectory	Need of multi-moda measurement for the same set of cells

Table caption: JIVE:joint and individual variation explained; (n,j,int) NMF:(network,joint,integrative) non-negative matrix factoization; SLIDE: structural learning and integrative decomposition; MALA: micro array logic analyzer; MOFA:multi-omics factor analysis (Adopted from Raufaste-Cazavieille et al., 2022)

7.3 Microbiome and host interactions

The study of microbiome and host interactions has greatly benefited from multi-omics integration. By combining metagenomics, metatranscriptomics, metaproteomics, and metabolomics data, researchers can gain a comprehensive view of microbial communities and their functions. This approach has been particularly useful in understanding the role of the microbiome in health and disease.

One successful example is the use of dynamic Bayesian networks (DBNs) to integrate multi-omics data from longitudinal microbiome studies. This method has been applied to data collected from patients with inflammatory bowel disease (IBD), allowing researchers to identify known and novel interactions between microbial taxa, their genes, metabolites, and host genes. The resulting models have provided insights into the temporal interactions and their impact on host expression, which are crucial for understanding disease progression and developing targeted

therapies (Ruiz-Perez et al., 2019). Additionally, multi-omics integration has been used to study the role of hydrogen sulfide in colon cancer. By combining 16S rRNA microbial community data with metabolomics and metabolic models, researchers have been able to track the metabolic flux of hydrogen sulfide and identify microbial interactions involved in its production.

8 Concluding Remarks

The integration of multi-omics data has emerged as a powerful approach to understanding complex biological systems. Key findings from recent research highlight the significant advancements and persistent challenges in this field. High-throughput technologies have enabled the generation of vast amounts of data across various omics layers, including genomics, transcriptomics, proteomics, and metabolomics, which can be integrated to provide a holistic view of biological processes. However, the integration of these diverse datasets remains challenging due to issues such as data heterogeneity, differences in nomenclature, and the need for robust computational methods. Recent studies have developed various tools and methodologies to address these challenges, focusing on applications such as disease subtyping, biomarker discovery, and precision medicine. Despite these advancements, there is still a need for standardized analytical pipelines and improved data visualization techniques to fully realize the potential of multi-omics integration.

Future research in multi-omics data integration is likely to focus on several key areas. One promising direction is the development of more sophisticated computational methods, including deep learning algorithms, which have shown great potential in capturing complex, nonlinear relationships within multi-omics data. Additionally, there is a growing interest in integrating single-cell multi-omics data, which can provide unprecedented insights into cellular heterogeneity and the molecular mechanisms underlying various biological processes. Another important area is the improvement of data visualization techniques, which are crucial for interpreting the results of multi-omics analyses and making them accessible to a broader scientific community. Furthermore, there is a need for more comprehensive and standardized data repositories and visualization portals to facilitate data sharing and collaboration among researchers.

To overcome the challenges associated with multi-omics data integration, several recommendations can be made. The development of standardized protocols for data cleaning, normalization, and integration is essential to ensure the consistency and reproducibility of multi-omics studies. The adoption of advanced computational methods, such as deep learning and network-based approaches, can help to address the complexity and high dimensionality of multi-omics data. Third, improving data visualization techniques and developing user-friendly tools for data exploration and interpretation will be crucial for making multi-omics analyses more accessible and actionable. Fostering collaboration and data sharing among researchers through the establishment of comprehensive data repositories and standardized analytical pipelines will be key to advancing the field of multi-omics data integration.

Acknowledgments

Special thanks to Ms. Jelena for providing relevant materials for this research.

Conflict of Interest Disclosure

The author affirms that this research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest.

References

Agamah F.E., Bayjanov J.R., Niehues A., Njoku K.F., Skelton M., Mazandu G.K., Ederveen T.H.A., Mulder N., Chimusa E., and Hoen P., 2022, Computational approaches for network-based integrative multi-omics analysis, Frontiers in Molecular Biosciences, 9: 967205. https://doi.org/10.3389/fmolb.2022.967205

Arbas S.M., Busi S.B., Queirós P., Nies L., Herold M., May P., Wilmes P., Muller E.E.L., and Narayanasamy S., 2021, Challenges strategies and perspectives for reference-independent longitudinal multi-omic microbiome studies, Frontiers in Genetics, 12: 666244. https://doi.org/10.3389/fgene.2021.666244

Benkirane H., Pradat Y., Michiels S., and Cournède P.H., 2023, CustOmics: a versatile deep-learning based strategy for multi-omics integration, PLOS Computational Biology, 19(3): e1010921. https://doi.org/10.1371/journal.pcbi.1010921

Bodein A., Scott-Boyer M.P., Périn O., Cao K.A., and Droit A., 2020, Interpretation of network-based integration from multi-omics longitudinal data, Nucleic Acids Research, 50(5): e27-e27.
https://doi.org/10.1093/nar/gkab1200

Colomé-Tatché M., and Theis F., 2018, Statistical single cell multi-omics integration, Current Opinion in Systems Biology, 7: 54-59.
https://doi.org/10.1016/J.COISB.2018.01.003

Cominetti O., Agarwal S., and Oller-Moreno S., 2023, Editorial: advances in methods and tools for multi-omics data analysis, Frontiers in Molecular Biosciences, 10: 1186822.
https://doi.org/10.3389/fmolb.2023.1186822

Demirel H.C., Arici M.K., and Tuncbag N., 2021, Computational approaches leveraging integrated connections of multi-omic data toward clinical applications, Molecular omics, 18(1): 7-18.
https://doi.org/10.1039/d1mo00158b

Ebrahim A., Brunk E.J., Tan J., O'Brien E., Kim D., Szubin R., Lerman J., Lechner A., Sastry A., Bordbar A., Feist A., and Palsson B., 2016, Multi-omic data integration enables discovery of hidden biological regularities, Nature Communications, 7(1): 13091.
https://doi.org/10.1038/ncomms13091

Feldner-Busztin D., Nisantzis P.F., Edmunds S.J., Boza G., Racimo F., Gopalakrishnan S., Limborg M.T., Lahti L., and Polavieja GG.., 2023, Dealing with dimensionality: the application of machine learning to multi-omics data, Bioinformatics, 39(2): btad021.
https://doi.org/10.1093/bioinformatics/btad021

Graw S., Chappell K., Washam C.L., Gies A., Bird J., Robeson M.S., and Byrum S., 2020, Multi-omics data integration considerations and study design for biological systems and disease, Molecular Omics, 17(2): 170-185.
https://doi.org/10.1039/d0mo00041h

Hauptmann T., and Kramer S., 2022, A fair experimental comparison of neural network architectures for latent representations of multi-omics for drug response prediction, BMC Bioinformatics, 24(1): 45.
https://doi.org/10.1186/s12859-023-05166-7

Jendoubi T., 2021, Approaches to integrating metabolomics and multi-omics data: a primer, Metabolites, 11(3): 184.
https://doi.org/10.3390/metabo11030184

Kang M., Ko E., and Mersha T.B., 2021, A roadmap for multi-omics data integration using deep learning, Briefings in Bioinformatics, 23(1): bbab454.
https://doi.org/10.1093/bib/bbab454

Kaur P., Singh A., and Chana I., 2021, Computational techniques and tools for omics data analysis: state-of-the-art challenges and future directions, Archives of Computational Methods in Engineering, 28(7): 4595-4631.
https://doi.org/10.1007/s11831-021-09547-0

Koppad S., Gkoutos G.V., and Acharjee A., 2021, Cloud computing enabled big multi-omics data analytics, Bioinformatics and Biology Insights, 15: 11779322211035921.
https://doi.org/10.1177/11779322211035921

Lee B., Zhang S., Poleksic A., and Xie L., 2020, Heterogeneous multi-layered network model for omics data integration and analysis, Frontiers in Genetics, 10: 381.
https://doi.org/10.3389/fgene.2019.01381

Li Y., Wu F.X., and Ngom A., 2016, A review on machine learning principles for multi-view biological data integration, Briefings in Bioinformatics, 19(2): 325-340.
https://doi.org/10.1093/bib/bbw113

Manzoni C., Kia D., Vandrovcova J., Hardy J., Wood N., Lewis P., and Ferrari R., 2016, Genome transcriptome and proteome: the rise of omics data and their integration in biomedical sciences, Briefings in Bioinformatics, 19: 286-302.
https://doi.org/10.1093/bib/bbw114

Miao B.B., Dong W., Gu Y.X., Han Z.F., Luo X., Ke C.H., and You W.W., 2023, OmicsSuite: a customized and pipelined suite for analysis and visualization of multi-omics big data, Horticulture Research, 10(11): uhad195.
https://doi.org/10.1093/hr/uhad195

Miao Z., Humphreys B., McMahon A., and Kim J., 2021, Multi-omics integration in the age of million single-cell data, Nature Reviews Nephrology, 17: 710-724.
https://doi.org/10.1038/s41581-021-00463-x

Misra B.B., Langefeld C., Olivier M., and Cox L.A., 2019, Integrated omics: tools advances and future approaches, Journal of Molecular Endocrinology, 62(1): R21-R45.
https://doi.org/10.1530/JME-18-0055

Nicora G., Vitali F., Dagliati A., Geifman N., and Bellazzi R., 2020, Integrated multi-omics analyses in oncology: a review of machine learning methods and tools, Frontiers in Oncology, 10: 1030
https://doi.org/10.3389/fonc.2020.01030

Ning L., and He H.X., 2021, Topic evolution analysis for omics data integration in cancers, Frontiers in Cell and Developmental Biology, 9: 631011.
https://doi.org/10.3389/fcell.2021.631011

Olivier M., Asmis R., Hawkins G.A., Howard T.D, and Cox L., 2019, The need for multi-omics biomarker signatures in precision medicine, International Journal of Molecular Sciences, 20(19): 4781.
https://doi.org/10.3390/ijms20194781

Pinu F.R., Beale D.J., Paten A.M., Kouremenos K., Swarup S., Schirra H.J., and Wishart D., 2019, Systems biology and multi-omics integration: viewpoints from the metabolomics research community, Metabolites, 9(4): 76.
https://doi.org/10.3390/metabo9040076

Raufaste-Cazavieille V., Santiago R., and Droit A., 2022, Multi-omics analysis: paving the path toward achieving precision medicine in cancer treatment and immuno-oncology, Frontiers in Molecular Biosciences, 9: 962743.
https://doi.org/10.3389/fmolb.2022.962743

Reel P.S., Reel S., Pearson E., Trucco E., and Jefferson E., 2021, Using machine learning approaches for multi-omics data analysis: a review, Biotechnology Advances, 49: 107739.
https://doi.org/10.1016/j.biotechadv.2021.107739

Ritchie M., Holzinger E., Li R., Pendergrass S., and Kim D., 2015, Methods of integrating data to uncover genotype–phenotype interactions, Nature Reviews Genetics, 16: 85-97.
https://doi.org/10.1038/nrg3868

Ruiz-Perez D., Lugo-Martinez J., Bourguignon N., Mathee K., Lerner B., Bar-Joseph Z., and Narasimhan G., 2019, Dynamic bayesian networks for integrating multi-omics time series microbiome data, mSystems, 6(2): 10.1128.
https://doi.org/10.1128/mSystems.01105-20

Santiago-Rodriguez T.M., and Hollister E.B., 2021, Multi 'omic data integration: a review of concepts considerations and approaches, Seminars in Perinatology, 45(6): 151456.
https://doi.org/10.1016/j.semperi.2021.151456

Sokač M., Kjær A., Dyrskjøt L., Haibe-Kains B., Aerts H., and Birkbak N., 2023, Spatial transformation of multi-omics data unlocks novel insights into cancer biology, eLife, 12: RP87133.
https://doi.org/10.7554/eLife.87133

Subramanian I., Verma S., Kumar S., Jere A., and Anamika K., 2020, Multi-omics data integration interpretation and its application, Bioinformatics and Biology Insights, 14: 1177932219899051.
https://doi.org/10.1177/1177932219899051

Terranova N., and Venkatakrishnan K., 2023, Machine learning in modeling disease trajectory and treatment outcomes: an emerging enabler for model-informed precision medicine, Clinical Pharmacology and Therapeutics, 115(4): 720-726.
https://doi.org/10.1002/cpt.3153

Tini G., Marchetti L., Priami C., and Scott-Boyer M., 2019, Multi-omics integration-a comparison of unsupervised clustering methodologies, Briefings in Bioinformatics, 20(4): 1269-1279.
https://doi.org/10.1093/bib/bbx167

Turanli B., Karagoz K., Gulfidan G., Sinha R., Mardinoğlu A., and Arğa K., 2019, A network-based cancer drug discovery: from integrated multi-omics approaches to precision medicine, Current Pharmaceutical Design, 24(32): 3778-3790.
https://doi.org/10.2174/1381612824666181106095959

Vahabi N., and Michailidis G., 2022, Unsupervised multi-omics data integration methods: a comprehensive review, Frontiers in Genetics, 13: 854752.
https://doi.org/10.3389/fgene.2022.854752

Wörheide M., Krumsiek J., Kastenmüller G., and Arnold M., 2021, Multi-omics integration in biomedical research-a metabolomics-centric review, Analytica Chimica Acta, 1141: 144-162.
https://doi.org/10.1016/j.aca.2020.10.038

Zhang B., and Kuster B., 2019, Proteomics is not an island: multi-omics integration is the key to understanding biological systems, Molecular and Cellular Proteomics, 18: S1-S4.
https://doi.org/10.1074/mcp.E119.001693