Biotechnology Reports 48 (2025) e00938

Contents lists available at ScienceDirect

Biotechnology Reports

journal homepage: www.elsevier.com/locate/btre

Review

Integrating multi-omics data: Methods and applications in human complex diseases

Pasquale Sibilio [a,b], Enrico De Smaele [b], Paola Paci [c,d], Federica Conte [e,*]

[a] Translational Oncology Research Unit, IRCCS—Regina Elena National Cancer Institute, Rome, Italy
[b] Department of Experimental Medicine, Sapienza University of Rome, Rome, Italy
[c] Department of Computer, Control and Management Engineering, Sapienza University of Rome, Rome, Italy
[d] Karolinska Institutet, Stockholm 17177, Sweden
[e] Institute for Systems Analysis and Computer Science "Antonio Ruberti", National Research Council, Rome, Italy

ARTICLE INFO

Keywords:
Multi-omics data
Integrative approaches
Human complex diseases

ABSTRACT

Over the past few decades, technological advancements and declining costs of high-throughput data generation have revolutionized biomedical research, enabling the collection of large-scale datasets across multiple omics layers—including genomics, transcriptomics, proteomics, metabolomics, and epigenomics. The analysis and integration of these datasets provides global insights into biological processes and holds great promise in elucidating the myriad molecular interactions associated with human diseases, particularly multifactorial ones such as cancer, cardiovascular, and neurodegenerative disorders. However, integrating multi-omics data presents significant challenges due to high dimensionality and heterogeneity. This review explores computational methods for integrating multi-omics data, with a particular focus on network-based approaches that offer a holistic view of relationships among biological components in health and disease. Furthermore, this review showcases a selection of recent, successful applications of multi-omics data integration, moving beyond theoretical methods to demonstrate their transformative potential in biomarker discovery, patient stratification, and guiding therapeutic interventions in specific human diseases.

1. Multi-omics data integration: A cornerstone for understanding human complex diseases

In the past decades, the advent of high-throughput technologies capable of quantitatively measuring the full spectrum of molecules across intracellular layers—such as the genome, transcriptome, epigenome, proteome, and metabolome—has revolutionized biomedical research (Fig. 1). Nowadays, the term "omics" refers not only to the profiling of specific classes of biomolecules but also to other types of high-dimensional biomedical data, such as *radiomics* (from imaging data) and *interactomics* (from all physical interactions between molecules) [1]. Although each omics data has individually contributed medical advances, it is their integration that may offer a more comprehensive, holistic understanding of human biology and diseases. Indeed, multi-omics data integration has emerged as a powerful approach to grasp the intricate interplay among different biomolecular layers that contribute to disease states, and thus to unveil more effective strategies for diagnosis, treatment, and prevention [1–3]. For instance, the integration of genomic and transcriptomic data allows the identification of expression quantitative trait loci (eQTLs), linking genetic variants to gene expression changes in disease [4]. Furthermore, combining epigenomic data (e.g., DNA methylation, ChIP-seq, and ATAC-seq) with transcriptomics within regulatory network frameworks enhances our understanding of how environmental exposures influence disease phenotypes [5].

However, the integration of multi-omics data presents numerous and multifaceted challenges [2,6]. A major difficulty lies in integrating datasets from different studies and laboratories, where non-biological sources of variation, or "batch effects", can compromise reproducibility and obscure true biological signals. The high dimensionality and heterogeneity of the data further complicate analyses, often amplifying noise, reducing statistical power, and raising significant issues for normalization. In addition, the continuous growth of multi-omics datasets demands computational frameworks capable of scaling efficiently without sacrificing accuracy and interpretability. Importantly, computational findings must subsequently be subjected to rigorous

* Corresponding author.
E-mail address: federica.conte@iasi.cnr.it (F. Conte).

https://doi.org/10.1016/j.btre.2025.e00938
Received 21 July 2025; Received in revised form 2 October 2025; Accepted 15 November 2025
Available online 15 November 2025

experimental validation to confirm their true biological and clinical relevance. These challenges can be effectively addressed through advanced computational strategies.

To provide a clear conceptual framework, it is essential to distinguish between the primary methods for multi-omics data integration. These methods are currently distinguished on the basis of various criteria [2]. A first key distinction is horizontal versus vertical integration [7]. Vertical integration is the most common approach and combines different omics data for the same set of biological samples or patients to build a comprehensive molecular profile for each individual. Horizontal integration, on the other hand, combines data from the same omics platform but across different studies or cohorts to increase statistical power and identify robust, reproducible biomarkers that are not specific to a single dataset, and thus validating findings on a larger scale. Other classification criteria pertain to the analytical strategy employed (statistical, artificial intelligence/machine learning, or network-based methods), the research question addressed (disease subtyping/patient stratification or biomarker prediction/disease insights), and the nature of the integrated data (specific omics or any omics data types) [2].

An alternative classification of multi-omics integration methods distinguish them between sequential (or cascade) approaches—where datasets are analysed separately and integrated at a later stage — and parallel (or simultaneous) approaches —where datasets are analysed jointly within a unified computational framework (Fig. 1). Slightly different from the distinction between sequential and parallel integration is the categorization into early versus late integration, although these terms are often used interchangeably in the literature [8,9]. Indeed, while sequential versus parallel integration pertains to the workflow of the integration process, early versus late integration primarily refers to the stage at which data from different omics layers are combined: in early integration, data from multiple omics are concatenated into a single dataset prior to analysis, whereas in late integration, each omics layer is analysed independently and the results are subsequently combined [9]. For completeness, we also refer to hybrid approaches which combine elements of both strategies, typically performing an initial separate analysis and then using the intermediate results to guide a more sophisticated joint analysis.

In general, methods integrating data sequentially are more commonly used to address problems related to disease subtyping and patient classification, whereas methods integrating data simultaneously are employed — albeit not exclusively — to predict novel biomarkers and derive disease insights [2,10,11]. Table 1 provides an overview of the state-of-art methodologies for integrating multi-omics data. Although we are aware that the field offers a vast array of sophisticated methods, we deliberately focused on a representative subset that are both widely adopted and particularly effective in illustrating the different integration strategies to a multidisciplinary audience.

Among AI/ML techniques, we presented FSMKL (Feature selection multiple kernel learning) [12], which leverages multiple kernels to capture omics layer similarity and pinpoint features crucial for disease progression; and ATHENA (Analysis Tool for Heritable and Environmental Network Associations) [13], which utilizes neural networks to build models from carefully selected omics features that are less noisy and associated with clinical outcomes. These methods, like most AI/ML approaches, offer exceptional predictive performance but their 'black-box' nature often limits biological interpretability of their predictions.

Statistical approaches that integrate specific multi-omics data sequentially are iPAC (in-trans Process Associated and Cis-correlated) [14], MCD (Multiple Concerted Disruption) [15]; and CNAmet (Copy Number Alterations and mRNA expression) [16]. Conversely, a well-established statistical methodology that integrates any kind of multi-omics data simultaneously is MOFA (Multi-Omics Factor Analysis) [17], that can be viewed as a generalization of principal component analysis. Indeed, MOFA employs a probabilistic Bayesian model to infer a set of hidden factors capturing the principal sources of data variability and defined as a linear combination of the individual molecular features from the input omics data. However, while MOFA excels at uncovering latent structures in multi-omics datasets, it lacks the ability to reveal the inter-relationships among the molecular features driving each inferred factor.

Network-based methods for multi-omics data integration have instead gained considerable attention due to their capacity to capture the intricate interactions within and across omics layers, thereby enabling the modeling of complex biological systems within a unified framework [23]. These methods are currently leveraged to uncover novel biomarkers, identify disease subtyping, and provide insights into the molecular underpinnings of biological functions and disease phenotypes. Among them, WGCNA (Weighted Gene Coexpression Network Analysis) [18] is a well-known tool to construct gene co-expression networks and to identify modules of highly correlated genes, which can be also extended to include different dataset or omics data in order to reveal shared modules (called consensus modules) among them [18]. A key drawback of WGCNA is its tendency to underrepresent negative correlations between omics features when constructing correlation-based networks [18]. Next, TransNet (Transkingdom Network) [19] is a network-based tool that, for identifying novel putative biomarkers, focuses on inter-omics interactions - particularly between host genes and microbial taxa- and disregards intra-omics interactions. Finally, notable examples of network-based methods developed mainly to enhance disease subtyping and patient stratification rather than biomarker prediction include iCluster [20], SNF (Similarity Network Fusion) [21], and NEMO (NEighborhood based Multi-Omics clustering) [22].

In this review, we want to offer a distinct perspective of multi-omics methods by focusing on their practical application in recent and notable case studies. Specifically, we selected successful examples from different human disease contexts—including colorectal cancer (Section 2), breast cancer (Section 3), chronic obstructive pulmonary disease (Section 4),

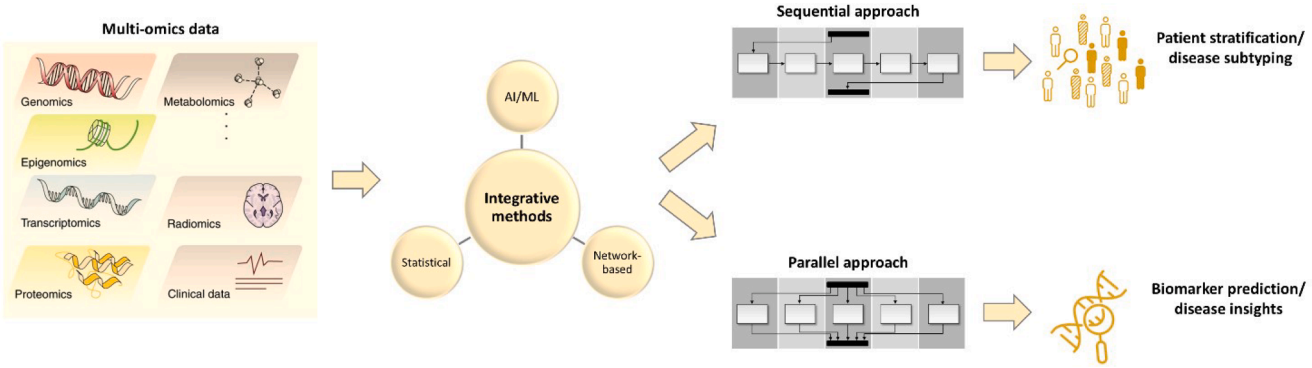

Fig. 1. Multi-omics data and methods.

Table 1
Examples of multi-omics data integration methodologies. Abbreviations: AI, Artificial Intelligence; ML, Machine Learning; CNV, Copy Number Variations; LOH, Loss Of Heterozygosity; BP, Biomarker Prediction; DI, Disease Insights; PS, Patient Stratification; DS, Disease Subtyping.

Name	Approach	Objective	Data input	Analysis	Language	Availability	Reference
FSMKL	AI/ML	BP/DI	Any omics	Parallel	Matlab	https://github.com/jseoane/FSMKL	Seoane et al., 2014 [12]
ATHENA	AI/ML	BP/DI	CNV,gene expression, DNA Methylation	Parallel	Perl	https://ritchielab.org/software/athena-downloads-1	Kim et al., 2013 [13]
iPAC	Statistical	BP/DI	CNV, gene expression	Sequential	Stand-alone	-	Aure et al., 2013 [14]
MCD	Statistical	BP/DI	CNV, LOH, DNA Methylation	Sequential	Stand-alone	-	Chari et al., 2010 [15]
CNAmet	Statistical	BP/DI	CNV, DNA Methylation, gene expression	Sequential	R	https://csbi.ltdk.helsinki.fi/CNAmet/	Louhimo et al., 2011 [16]
MOFA	Statistical	BP/DI	Any omics	Parallel	R/Python	https://github.com/bioFAM/MOFA	Argelaguet et al., 2020 [17]
WGCNA	Network-based	BP/DI	Any omics	Parallel	R	https://cran.r-project.org/web/packages/WGCNA/index.html	Langfelder et al., 2008 [18]
TransNet	Network-based	BP/DI	Any omics	Parallel	R	https://github.com/richrr/TransNetDemo	Rodrigues et al., 2018 [19]
iCluster	Network-based	PS/DS	CNV, DNA Methylation, gene expression	Parallel	R	http://www.mskcc.org/mskcc/html/85130.cfm	Shen et al., 2009 [20]
SNF	Network-based	PS/DS	Any omics	Parallel	R/Matlab	http://compbio.cs.toronto.edu/SNF/SNF/Software.html	Wang et al., 2014 [21]
NEMO	Network-based	PS/DS	Any omics	Parallel	R	https://github.com/Shamir-Lab/NEMO	Rappoport et al., 2019 [22]

glioblastoma multiforme (Section 5) and central nervous system disorders (Section 6)—to illustrate how sequential and parallel integration approaches were used to derive novel clinical and biological insights. Our goal is to bridge the gap between methodological descriptions and their real-world impact, providing a clear demonstration of the power of multi-omics data integration in understanding disease pathophysiology, guiding therapeutic interventions, and improving patient stratification.

2. Sequential multi-omics data integration for studying colorectal cancer

Colorectal cancer (CRC) ranks among the leading causes of cancer-related morbidity and mortality worldwide. The classification of CRCs has historically been based on two mutually exclusive subgroups with distinct genomic and molecular features: Microsatellite Instability (MSI, ~15 % of CRCs) and Chromosomal Instability (CIN, ~85 % of CRCs) [24]. Recent discoveries regarding the favourable response of MSI CRCs to immune checkpoint inhibitors (ICIs), in contrast to the poor response observed in Microsatellite Stable (MSS)/CIN CRCs, have highlighted the clinical importance of this classification for therapeutic decision-making [24]. MSI CRCs exhibit a high tumour mutational burden (TMB), a condition referred to as hypermutation, which can lead to the generation of numerous immunogenic neoantigens on the surface of cancer cells. This, in turn, promotes lymphocyte infiltration into the tumour microenvironment, enabling an effective immune response that can be enhanced by immunotherapy. However, the MSI CRCs classification has limitations, as it does not encompass the entire spectrum of hypermutated (HM) CRCs—for example, tumours with high TMB that retain a proficient DNA mismatch repair (MMR) system [25].

With the aim of investigating the differences between HM and non-HM tumours and potentially expanding the CRC subgroups eligible for ICIs, Sibilio et al. [26] recently performed a sequential multi-omics integration of genomic, epigenomic, and transcriptomic data (Fig. 2) of 520 CRC patients retrieved from The Cancer Genome Atlas (TCGA) [27].

The first step of this cascading multi-omics analysis involved the stratification of CRC patients based on their copy number variation (CNV) profiles. This revealed a novel cluster of CRCs, accounting for approximately 8 % of the entire cohort, which exhibited distinct molecular features compared to both HM and non-HM tumours. Notably, this new subgroup was characterized by low TMB and low CNV levels, yet shared clinicopathological features with HM tumours. Consequently, it was designated as HM-like (orange samples in Fig. 2). The second step of the analysis focused on the molecular characterization of the HM, HM-like, and non-HM subgroups, by integrating DNA methylation and single nucleotide variation (SNV) data. SNV analysis revealed a distinct mutational profile in HM-like tumours compared to both HM and non-HM tumours. Interestingly, the HM-like subgroup displayed the highest frequency of KRAS mutations and a low CpG Island Methylator Phenotype (CIMP-L). Moreover, HM-like tumours were distinguished by a lower dependence on WNT and TGF-β pathway deregulation and exhibited a high prevalence of *SOX9* mutations. Additionally, mutational signature analysis of SNV data suggested that the pathogenic mechanisms and aetiology underlying HM-like CRCs may differ from those of HM and non-HM tumours. The final step of this multi-omics integration approach involved transcriptomics data analysis to deconvolute immune cell infiltration. This revealed that HM-like tumours exhibited high expression of genes associated with inflammatory processes. Most notably, they were characterized by T-cell and NK-cell infiltration signatures, similar to those observed in HM tumours.

In summary, this study highlighted the existence of a novel CRC subgroup with distinct molecular features and a “hot” immunogenic profile, which may support the expansion of ICI therapy to a broader range of CRC patients. However, it should be acknowledged that the

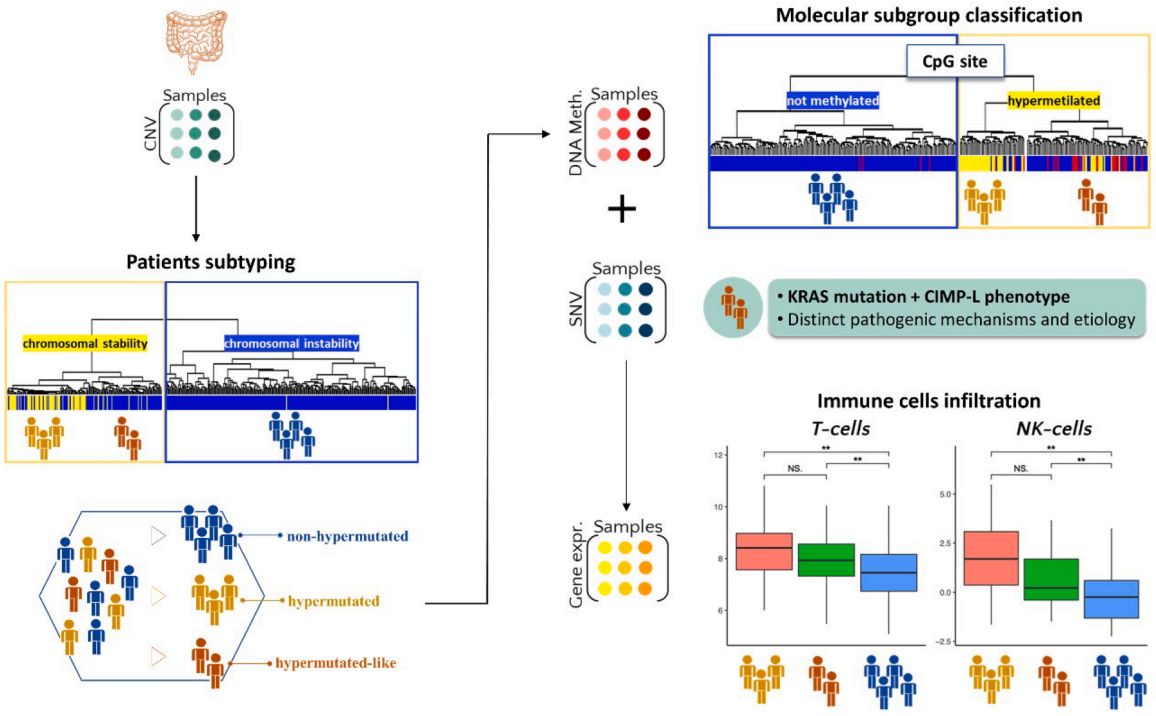

Fig. 2. Workflow of the sequential integration analysis for defining a novel subgroup of CRCs patients possibly eligible for ICIs therapy.

inherent sequential nature of this analysis may constitute a limitation. Indeed, the initial patient stratification based solely on CNV profiles could potentially miss important subgroups that are defined by a combination of genomic, epigenomic, and transcriptomic features. Furthermore, the findings are based on a single dataset (TCGA), and the results require further validation in independent cohorts to ensure they are reproducible and clinically relevant.

3. Sequential multi-omics data integration for studying breast cancer

Breast cancer (BC) is the most common cancer among women and, despite significant progress in early detection and research, it remains the second leading cause of death in women worldwide [28]. BC is inherently heterogeneous, as evidenced by the existence of various subtypes with distinct morphological characteristics and clinical

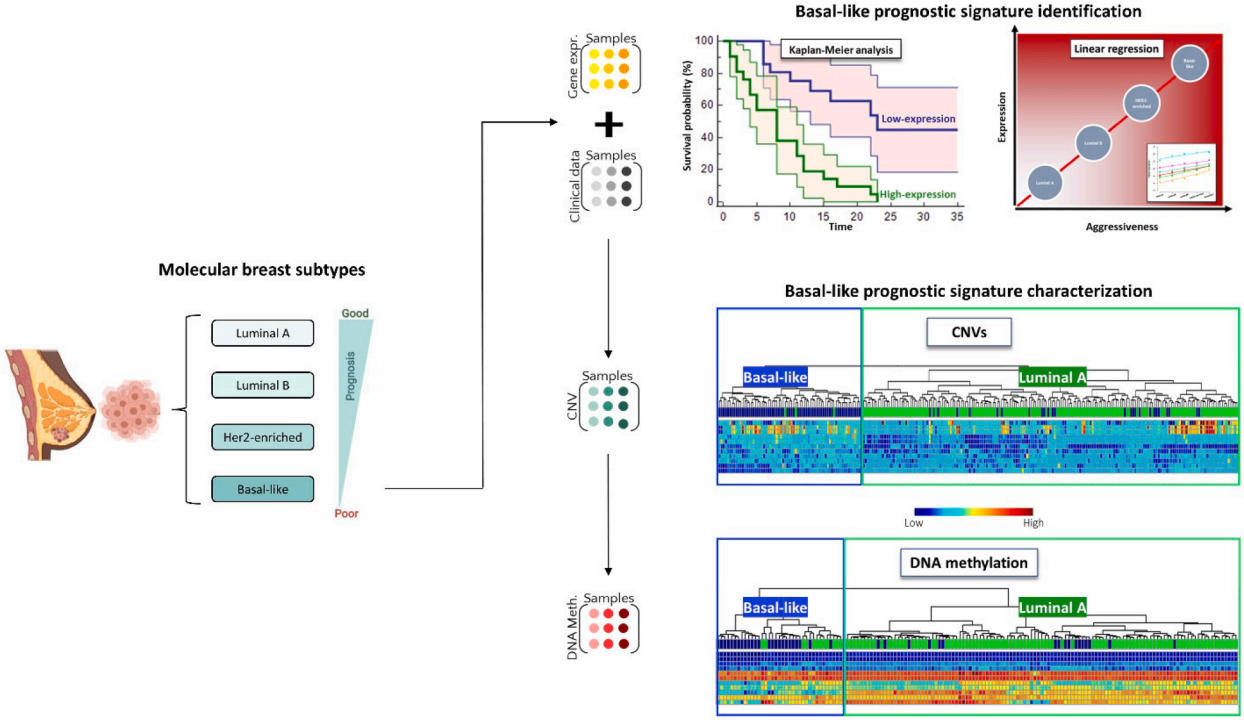

Fig. 3. Workflow of the sequential integration analysis for identifying a basal-like signature in breast cancer patients.

outcomes [29]. Genomic studies led to the so-called PAM50 classification [30,31], which recognizes four molecular intrinsic subtypes of BC, i.e., Luminal A, Luminal B, HER2 positive, and Basal-like. Among them, the most aggressive phenotypes is the Basal-like subtype, which accounts for about 15–20 % of all diagnosed BC cases [32]. This subtype differs from others because it grows and spreads more rapidly, has fewer treatment options—typically relying on chemotherapy—and exhibits a metastatic pattern that more frequently involves the brain and lungs. Relapse in basal-like patients is frequent, particularly within the first five years, leading to the worst survival outcomes among all BC subtypes [32]. Currently, there are no widely accepted prognostic markers available to predict outcomes for basal-like patients. Consequently, the identification of novel prognostic biomarkers for the basal-like subtype remains an unmet clinical need that could significantly enhance disease management.

In this context, Conte et al. [33] recently performed a sequential multi-omics integration of transcriptomic, genomic, epigenomic, and clinical data retrieved from TCGA [27] for the recognition of a basal-like prognostic gene signature (Fig. 3).

Specifically, the authors started from the application of a well-established network-based methodology on the transcriptomic data of TCGA-BC samples to unveil key genes (called switch genes) putatively associated with the transition from healthy to disease state. This step led to the identification of 108 switch genes that appeared to be specifically active in basal-like compared to other BC subtypes [33,34]. Next, by combining transcriptomic and clinical data, the authors conducted survival and linear regression analyses to associate the gene expression profiles of the identified basal-like switch genes with both the patients' clinical outcome and the disease aggressiveness (Fig. 3). These analyses revealed a signature of 11 genes —CENPN, LRP8, DSCC1, CTPS, RCOR2, GINS4, TUBA1C, PRAME, SLC7A11, CDCA7, and GSDMC— overexpressed in basal-like tumors and associated with the worst overall survival, thus acting as unfavorable prognostic biomarkers. Finally, the 11 basal-like prognostic biomarkers were further analysed to investigate whether variations in their expression could be related to genetic (copy number variations, CNV) or epigenetic (DNA methylation differences) alterations in their gene loci, or to the activities of transcription factors binding to their promoter regions. Some genes showed amplification at the DNA level, meaning that extra copies of the gene may drive higher activity, while others exhibited hypomethylation that could increase gene expression. In addition, many of these genes were found to be putatively regulated by key transcription factors such as MYC, TP53, and NFKB1.

Overall, the genes identified in this research study appeared to be involved in biological pathways linked to cell cycle regulation, DNA replication, and resistance to cell death—processes that are often disrupted in cancer cells. By highlighting these pathways, the study not only provided potential biomarkers for prognosis but also offered insight into the biological behaviour of basal-like tumours. However, the presented BC study shares a primary limitation with the CRC study: its sequential approach. The analysis first identified "switch genes" from transcriptomic data alone and then integrated them with other omics data. This method risks losing crucial information about genes whose expression may not be the primary driver of the disease but are still significantly influenced by upstream genetic or epigenetic alterations. Next, the reliance on a single dataset (TCGA) also poses a challenge for reproducibility and generalizability. Additionally, further laboratory and clinical validations are mandatory to confirm the biological and clinical relevance of the findings. The lack of this validation is another major limitation for translating the results into personalized treatment and prognosis.

4. Parallel multi-omics data integration for studying chronic obstructive pulmonary disease

Chronic Obstructive Pulmonary Disease (COPD) is a progressive lung disease characterized by persistent airflow limitation and chronic inflammation in the airways. It is a major cause of morbidity and mortality worldwide and is often associated with smoking and exposure to environmental pollutants [35]. Despite extensive research, the exact molecular mechanisms underlying COPD are still not fully understood. It is known that both genetic and epigenetic factors play roles in disease development and progression [36]. Therefore, understanding how gene expression and epigenetic modifications (like DNA methylation) interact in lung tissue could provide new insights into the biological basis of COPD and potentially uncover novel therapeutic targets.

With this aim, Sibilio et al. [37] very recently performed a parallel multi-omics integration of RNA sequencing and DNA methylation lung tissue data in a COPD case-control cohort derived from Lung Tissue Research Consortium (LTRC). The authors exploited a correlation-based network approach to integrate these two different omics data (Fig. 4). Specifically, they first used the well-established SWIM tool [38,39] to compute the differentially expressed genes (DEGs) and differentially methylated genes (DMGs) between COPD and control samples, and then to build their separate correlation networks. Next, they integrated these two single-omics correlation into a unique network (named coupled network), preserving the nodes and the edges in common between them (Fig. 4). This means that in the coupled network the nodes were genes both DEGS and DMGs and a link between two nodes occurred only if they were highly correlated or anti-correlated in both RNA and methylation-based networks. In particular, the links of the coupled network were categorized into four types (i.e., $+/+$, $-/-$, $+/-$, $-/+$), depending on the sign of pairwise correlations in the single-omics correlation networks: $+/+$ referred to gene pairs that were positively correlated in both single-omics correlation networks; $-/-$ referred to gene pairs that were negatively correlated in both single-omics correlation networks; $\pm$ and $-/+$ referred to gene pairs whose sign of the correlation coefficient was opposite in the single-omics correlation networks.

Ultimately, the integrative analysis proposed in [37] led to a coupled network consisting of 97 genes and 317 links, enabling to investigate how the gene-gene relationships vary moving from transcriptomics to epigenomics layer and which genes are mainly involved in these changes (i.e., nodes with greater size/degree in Fig. 4).The biological functional enrichment analysis of these genes revealed a strong involvement in immune system pathways, including both innate and adaptive immunity, as well as cytokine signalling. These findings were consistent with the known inflammatory and immune-related components of COPD pathogenesis. Moreover, genes not previously known to participate to COPD pathogenesis and inflammation process were identified interacting with known COPD-related genes at both transcriptomics and epigenomics layers. However, a key limitation of this parallel integration approach remains its reliance on correlation, which does not imply causation. Indeed, the identified links in the coupled network show that two genes are correlated in both transcriptomic and epigenomic layers, but they don't explain the causal mechanism behind that relationship. Additionally, the study primarily focuses on interactions between different omics layers (inter-omics), which may lead to overlooking important relationships within individual layers (intra-omics).

5. Parallel multi-omics data integration for studying glioblastoma multiforme

Glioblastoma Multiforme (GBM) is the most aggressive and lethal brain tumour, with a median survival of just 12–15 months following diagnosis [40,41]. It represents approximately 15 % of all primary brain tumours, 46 % of malignant primary brain tumours, and between 60 and 75 % of astrocytomas. GBM is characterized by rapid growth, extensive infiltration into surrounding brain tissue, and resistance to conventional therapies such as surgical removal followed by radiotherapy and chemotherapy. Another defining features of GBM is its heterogeneity—the tumour exhibits high variability at genetic, epigenetic,

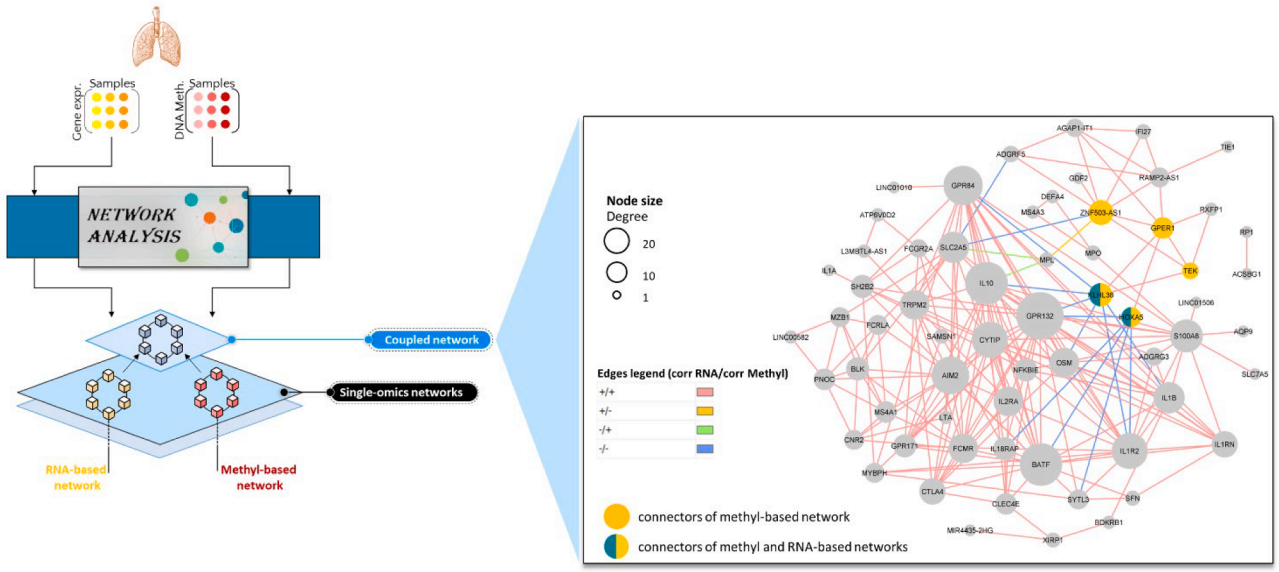

Fig. 4. Workflow of the parallel integration analysis for investigating COPD-related mechanisms. In the zoomed network view: edges are coloured according the four types of correlations; node size is proportional to number of connections (degree) in the coupled network; nodes found to be key connectors in the single-omics correlation networks are highlighted.

transcriptional, and cellular levels—, which contributes to differences in clinical behaviour and treatment response [42]. Thus, understanding and targeting this heterogeneity has become a central focus in current GBM research, with the goal of developing more personalized and effective treatment strategies.

In an effort to reach this objective, Wang et al. [21] proposed a parallel multi-omics integration of DNA methylation, mRNA expression and miRNA expression data derived from 215 GBM patients. This analysis was performed through the Similarity Network Fusion (SNF) method, which uses networks of samples as a basis for integration. Specifically, starting from multiple data types available for the same set of samples (e.g., patients), SNF first computes similarity scores between each pair of samples and uses them to simultaneously build a sample-by-sample similarity matrix for each data type (Fig. 5). Next, these matrices are converted into similarity networks, where the nodes represent samples and the weighted edges indicate the degree of similarity between them (Fig. 5). Finally, the distinct similarity networks are fused into a single comprehensive network that reflects complementary information from all data types. This fusion is achieved via a non-linear combination method based on message-passing-like process that iteratively updates every network, making it more similar to the others with every iteration. After several iterations, SNF converges to the final fused network that may reveal clearer clustering patterns (e.g., disease subtypes) than any single data type alone (Fig. 5). Additional technical details about the SNF method can be found in reference [21].

The application of SNF approach to GBM data enabled the identification of three clearly distinct clusters of patients associated to novel putative disease subtypes with differential survival profiles [21]. These

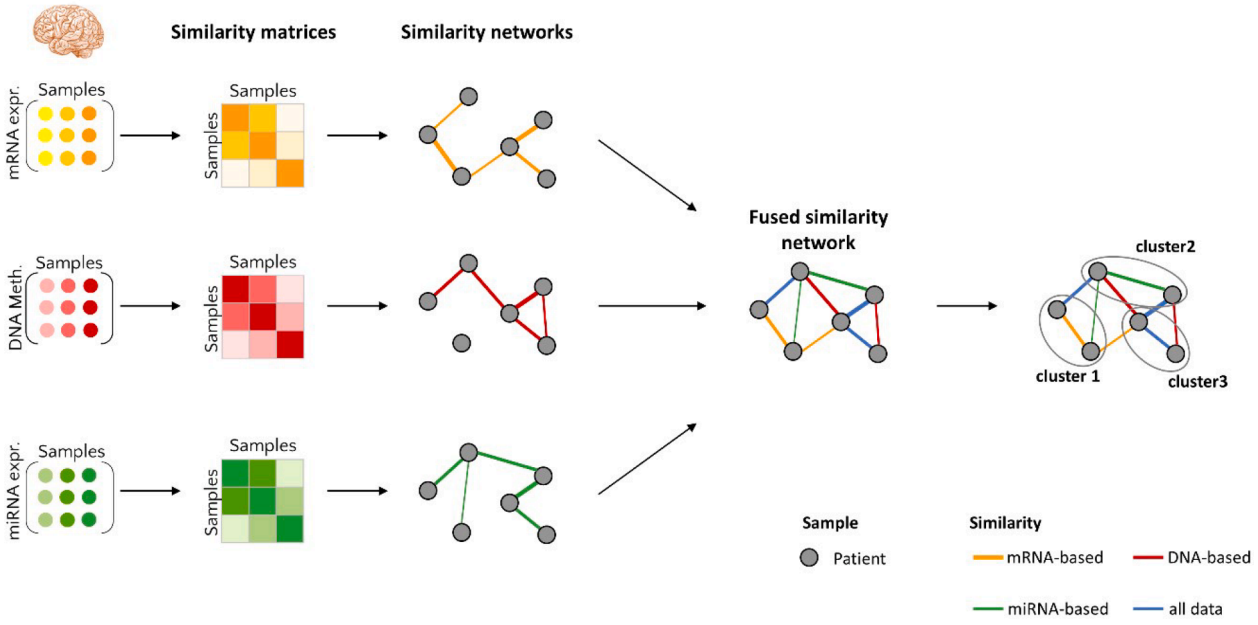

Fig. 5. Workflow of the parallel integration analysis for studying glioblastoma through SNF method. Different colours refer to different data types. Edge thickness of similarity networks reflects the strength of the similarity.

findings were subsequently compared with unified results from previous GBM subtype analyses, highlighting associations of potential biological and clinical relevance. Specifically, the authors identified: a first patient cluster (subtype 3) corresponding to the well-characterized IDH subtype [43] and consisting of younger subjects with markedly better prognosis; a second patient cluster (subtype 1) exhibiting a favourable response to temozolomide (TMZ), a standard chemotherapeutic agent used in GBM treatment; and a third patients' cluster (subtype 2) characterized by a lack of TMZ responsiveness, likely attributable to a strong association with CTSD overexpression — previously shown to impair TMZ efficacy in vitro [44]. However, it should be noted that this method can be computationally intensive and challenging to scale for even larger datasets. Moreover, a major limitation of methods like SNF is that while they can reveal clear clustering patterns, the biological interpretability of the fused network can be a "black box". It's difficult to pinpoint which specific molecular features from each omics layer are driving the observed patient clusters. The authors correctly stated that their method goes "beyond a conventional disease subtyping strategy," but the challenge of interpreting the fused network remains a key limitation for achieving deeper biological insights.

6. Single-cell multi-omics data integration for studying central nervous system disorders

Central nervous system (CNS) disorders—including ischemic, hemorrhagic, neurodegenerative, inflammatory, and developmental conditions—affect a diverse array of cell types, each with distinct functions and gene expression profiles. Traditional bulk sequencing approaches, which average gene expression across millions of cells, fail to capture the distinct molecular signatures of specialized cell types such as neurons, astrocytes, and microglia. This limitation hampers the identification of the specific cellular changes driving disease progression, leaving critical questions about underlying mechanisms and potential therapeutic targets unresolved. By contrast, single-cell RNA sequencing (scRNA-seq) has emerged as a powerful tool to dissect this cellular heterogeneity, enabling high-resolution analysis and precise characterization of individual cells [45,46].

Against this backdrop, Fang et al. [47] recently proposed a groundbreaking machine learning approach for the parallel integration of massive amounts of scRNA-seq data derived from multiple public and private sources (Fig. 6). Specifically, the authors employed a self-supervised contrastive learning method—called scCM—to combine 20 CNS datasets spanning 4 species (i.e., human, rodent, primate, fish) and 10 CNS diseases (e.g., glioblastoma, Alzheimer's disease, Huntington's disease, medulloblastoma, multiple sclerosis). In brief, scCM learns to recognize and cluster cells with similar gene expression patterns while distinguishing cells with larger expression differences across different studies, without requiring prior labels or manual annotation. By distinguishing between similar and dissimilar cell pairs, the model effectively removes technical noise (e.g., batch effects) and projects the data into a shared biological space, revealing the underlying cellular landscape. Technical details about the scCM method can be found in reference [47].

The innovative approach by Fang et al. [47] led to several significant results. The most prominent finding was the creation of a comprehensive cellular atlas of the CNS, which revealed both known and novel cell types and states in unprecedented detail. For instance, the study unveiled previously uncharacterized disease-specific cell states in microglia and astrocytes that were associated with specific pathological conditions. This provided a deeper understanding of how these cells contribute to neuroinflammation and disease progression in conditions like Alzheimer's disease and multiple sclerosis [47]. The strengths of this study are twofold: i) the sophisticated use of advanced machine learning, which proved highly effective at integrating large, heterogeneous datasets and robustly removing batch effects—a major hurdle in single-cell research; and ii) the creation of a single, unified dataset for investigating CNS disease mechanisms at the cellular level, thereby offering new opportunities for therapeutic discovery. Nevertheless, a few weaknesses should be acknowledged. The computational resources required for this type of large-scale integration are substantial, potentially restricting its accessibility to smaller research groups. Furthermore, while the method successfully identifies novel cellular populations and states, the biological validation of these findings still requires extensive follow-up experiments.

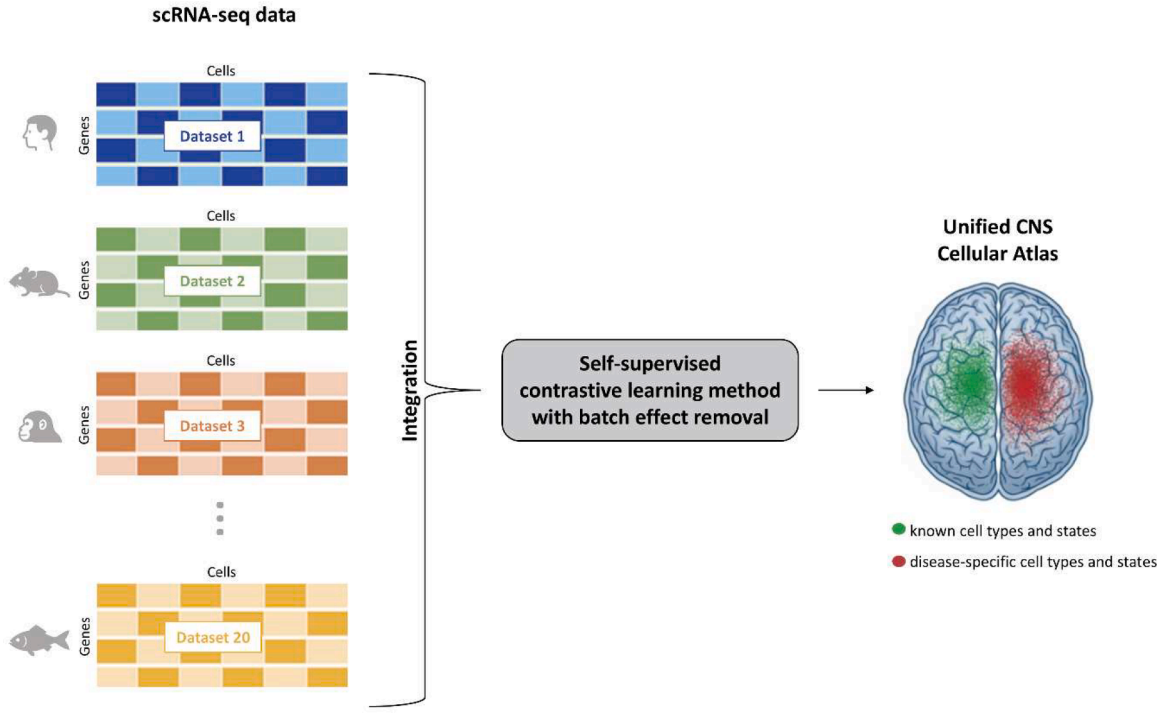

Fig. 6. Workflow of the single-cell data integration leading to the construction of a unified CNS cellular atlas.

Finally, we would like to point out that the CNS case study here presented differs substantially from the previous ones in terms of methodology, integration strategy, and data type, as it applies a machine learning–based approach to achieve horizontal integration of multiple single-cell datasets.

7. Concluding remarks

This review offers a novel perspective by going beyond a mere description of computational integration methods. Indeed, unlike many existing reviews [10,11], we highlighted their practical applications in recent, successful, real-world case studies that address specific clinical problems and provide tangible insights into complex human diseases. Specifically: i) the CRC case study showed how a sequential multi-omics integration analysis can uncover a new patient subgroup with a "hot" immunogenic profile that may be eligible for immune checkpoint inhibitor therapy; ii) the BC case study demonstrates how a sequential multi-omics integration approach may lead to the identification of a prognostic gene signature for the aggressive basal-like subtype; iii) the COPD case study showcases a parallel multi-omics integration approach that revealed new insights into the immune system pathways involved in the disease and identified genes not previously known to participate in COPD pathogenesis; iv) the GBM case study highlights the use of SNF to identify new patient clusters with different survival profiles and treatment responses, offering a deeper understanding of this highly heterogeneous cancer; v) the CNS case study illustrates how integrating single-cell data from multiple sources into a unified, large-scale dataset can reveal and validate cellular heterogeneity and its role in driving diseases. By detailing these diverse applications, our review serves as a practical example of the transformative potential of multi-omics data integration.

We also aimed to make this review accessible to a heterogeneous audience, including experimental biologists and clinicians with limited expertise in computational methodologies. Accordingly, the descriptions of the methods and case studies were intentionally kept at a non-technical level, prioritizing conceptual clarity and practical relevance over technical details.

A critical determinant of success in multi-omics integration studies is the choice of the most appropriate integration strategy, which is a non-trivial task dependent on multiple factors. A practical decision-making framework for researchers— including non-specialists —should begin with a clear definition of the study objective (e.g., patient stratification, biomarker discovery). Beyond the research question itself, characteristics of the data such as heterogeneity and dimensionality play a pivotal role in shaping methodological choices. For instance, when datasets are highly disparate or noisy, parallel integration approaches may be preferable, as they can accommodate complexity within a unified framework. Computational resources also represent another important factor. If available resources are limited, sequential approaches may offer a more feasible and less computationally intensive alternative. Another key factor to consider is the expertise of applied researchers. While many computational methods require a strong background in programming languages such as R or Python, an increasing number of user-friendly GUI-based tools and web platforms are becoming available, making such analyses more accessible to researchers with limited programming expertise. Nevertheless, for more advanced analyses, customization, and the handling of large-scale datasets, a deeper technical understanding remains essential. A further major question in study design concerns the required number of individuals or samples for a robust analysis. Although there is no universal rule, for unsupervised analyses a minimum of 50–100 samples is typically recommended to uncover meaningful patterns. In contrast, for supervised analyses—such as building a predictive model—the required sample size depends heavily on the number of features. In general, larger cohorts are crucial for ensuring that findings are reproducible and not merely artifacts of a small, underpowered study [48]. Finally, the results from any multi-omics analysis should always be validated. Validation is typically carried out in two ways: (i) independent dataset validation, the most common approach, in which the same method is applied to a different, independent dataset to assess reproducibility; and (ii) experimental validation, which tests the biological or clinical relevance of the findings in the laboratory.

As the field of multi-omics continues to evolve, future research directions will focus mainly on integrating data at a more granular level and from decentralized platforms. A first major frontier is the integration of multi-omics data at the single-cell level [9]. While bulk multi-omics often masks cellular heterogeneity, this approach allows for an unprecedented view of how molecular states (e.g., genetic, transcriptional, and epigenetic) vary from one cell to another within a tissue, providing a deeper understanding of disease mechanisms. Complementing this, spatial multi-omics technologies are emerging that preserve the physical location of molecules and cells within a tissue [49]. By integrating multi-omics data with spatial context, researchers can investigate how cell-cell interactions and tissue architecture influence disease phenotypes. However, the development of robust computational methods that can handle the high dimensionality and complexity of these datasets (single-cell and spatial) remains a crucial challenge for the field. In parallel, challenges related to data privacy and sharing between institutions are driving new computational paradigms, such as federated and distributed data integration methods [50]. These approaches allow organizations to collaboratively derive insights from large patient cohorts without transferring sensitive information, thus balancing the need for innovation with ethical constraints and accelerating research in personalized medicine.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used “chatGPT” in order to refine the text. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

CRediT authorship contribution statement

Pasquale Sibilio: Writing – original draft, Investigation, Formal analysis. **Enrico De Smaele:** Writing – review & editing, Funding acquisition. **Paola Paci:** Writing – review & editing, Funding acquisition, Conceptualization. **Federica Conte:** Writing – original draft, Visualization, Supervision, Project administration.

Declaration of competing interest

The authors declare no competing interests.

Acknowledgements

The research leading to these results has received funding from the European Union - NextGenerationEU through the Italian Ministry of University and Research under PNRR - M4C2-I1.3 Project PE_00000019 "HEAL ITALIA" (CUP B53C22004000006) to Enrico De Smaele, and from MUR - PRIN 2022 “Defining gene regulation and coregulation at single cell resolution in grapevine” (CUP B53D23008170006) to Paola Paci. The views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

Data availability

No data was used for the research described in the article.

References

[1] Y. Hasin, M. Seldin, A. Lusis, Multi-omics approaches to disease, Genome Biol. 18 (2017) 83.
[2] I. Subramanian, S. Verma, S. Kumar, A. Jere, K. Anamika, Multi-omics Data integration, interpretation, and its application, Bioinform. Biol. Insights 14 (2020) 1177932219899051.
[3] K.J. Karczewski, M.P. Snyder, Integrative omics for health and disease, Nat. Rev. Genet. 19 (2018) 299–310.
[4] GTEx Consortium, Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans, Science 348 (2015) 648–660.
[5] A. Kundaje, et al., Integrative analysis of 111 reference human epigenomes, Nature 518 (2015) 317–330.
[6] S. Tarazona, A. Arzalluz-Luque, A. Conesa, Undisclosed, unmet and neglected challenges in multi-omics studies, Nat. Comput. Sci. 1 (2021) 395–402.
[7] NN7, T. Das, G. Andrieux, M. Ahmed, S. Chakraborty, Integration of online omics-data resources for cancer research, Front. Genet 11 (2020) 578345.
[8] N. Rappoport, R. Shamir, Multi-omic and multi-view clustering algorithms: review and cancer benchmark, Nucleic Acids Res. 46 (2018) 10546–10562.
[9] N. Adossa, S. Khan, K.T. Rytkönen, L.L. Elo, Computational strategies for single-cell multi-omics integration, Comput. Struct. Biotechnol. J. 19 (2021) 2588–2596.
[10] M. Bersanelli, et al., Methods for the integration of multi-omics data: mathematical aspects, BMC Bioinform. 17 (2016). S15.
[11] N. Vahabi, G. Michailidis, Unsupervised Multi-Omics data integration methods: a comprehensive review, Front. Genet 13 (2022).
[12] J.A. Seoane, I.N.M. Day, T.R. Gaunt, C. Campbell, A pathway-based data integration framework for prediction of disease progression, Bioinformatics 30 (2014) 838–845.
[13] D. Kim, R. Li, S.M. Dudek, M.D. Ritchie, ATHENA: identifying interactions between different levels of genomic data associated with cancer clinical outcomes using grammatical evolution neural network, *BioData* Min 6 (2013) 23.
[14] M.R. Aure, et al., Identifying in-trans process associated genes in breast cancer by integrated analysis of copy number and expression data, PLoS One 8 (2013) e53014.
[15] R. Chari, B.P. Coe, E.A. Vucic, W.W. Lockwood, W.L. Lam, An integrative multi-dimensional genetic and epigenetic strategy to identify aberrant genes and pathways in cancer, BMC Syst. Biol. 4 (2010) 67.
[16] R. Louhimo, S. Hautaniemi, CNAmet: an R package for integrating copy number, methylation and expression data, Bioinformatics 27 (2011) 887–888.
[17] R. Argelaguet, et al., MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data, Genome Biol. 21 (2020) 111.
[18] P. Langfelder, S. Horvath, WGCNA: an R package for weighted correlation network analysis, BMC Bioinform. 9 (2008) 559.
[19] R.R. Rodrigues, N. Shulzhenko, A. Morgun, Transkingdom networks: a systems biology approach to identify causal members of host-microbiota interactions, Methods Mol. Biol. 1849 (2018) 227–242.
[20] R. Shen, A.B. Olshen, M. Ladanyi, Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis, Bioinformatics 25 (2009) 2906–2912.
[21] B. Wang, et al., Similarity network fusion for aggregating data types on a genomic scale, Nat. Methods 11 (2014) 333–337.
[22] N. Rappoport, R. Shamir, NEMO: cancer subtyping by integration of partial multi-omic data, Bioinformatics 35 (2019) 3348–3356.
[23] J. Yan, S.L. Risacher, L. Shen, A.J. Saykin, Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data, Brief. Bioinformatics 19 (2018) 1370–1381.
[24] H.T. Nguyen, H.-Q. Duong, The molecular characteristics of colorectal cancer: implications for diagnosis and therapy, Oncol. Lett. 16 (2018) 9–18.
[25] M. Catalano, et al., Immunotherapy-related biomarkers: confirmations and uncertainties, Crit. Rev. Oncol. Hematol. 192 (2023) 104135.
[26] P. Sibilio, F. Belardinilli, V. Licursi, P. Paci, G. Giannini, An integrative in-silico analysis discloses a novel molecular subset of colorectal cancer possibly eligible for immune checkpoint immunotherapy, Biol. Direct 17 (2022) 10.
[27] A. Colaprico, et al., TCGAbiolinks: an R/bioconductor package for integrative analysis of TCGA data, Nucleic Acids Res. 44 (2016) e71.
[28] F. Bray, et al., Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA *Cancer J. Clin.* 74 (2024) 229–263.
[29] D. Zardavas, A. Irrthum, C. Swanton, M. Piccart, Clinical management of breast cancer heterogeneity, Nat. Rev. Clin. Oncol. 12 (2015) 381–394.
[30] Comprehensive molecular portraits of human breast tumours, Nature 490 (2012) 61–70.
[31] J.S. Parker, et al., Supervised risk predictor of breast cancer based on intrinsic subtypes, J. Clin. Oncol. 27 (2009) 1160–1167.
[32] H.H. Milioli, I. Tishchenko, C. Riveros, R. Berretta, P. Moscato, Basal-like breast cancer: molecular profiles, clinical features and survival outcomes, BMC Med. Genomics 10 (2017) 19.
[33] F. Conte, et al., In silico recognition of a prognostic signature in basal-like breast cancer patients, PLoS One 17 (2022) e0264024.
[34] A.M. Grimaldi, et al., The new paradigm of network medicine to analyze breast cancer phenotypes, Int. J. Mol. Sci. 21 (2020) 6690.
[35] S.A. Quaderi, J.R. Hurst, The unmet global burden of COPD, Glob. Health Epidemiol. Genom. 3 (2018) e4.
[36] E.K. Silverman, J.D. Crapo, B.J. Make, et al., Chronic obstructive pulmonary disease, in: J.L. Jameson, et al. (Eds.), Harrison's Principles of Internal Medicine, McGraw-Hill Education, New York, NY, 2018.
[37] P. Sibilio, et al., Correlation-based network integration of lung RNA sequencing and DNA methylation data in chronic obstructive pulmonary disease, Heliyon 10 (2024) e31301.
[38] G. Fiscon, F. Conte, P. Paci, SWIM tool application to expression data of glioblastoma stem-like cell lines, corresponding primary tumors and conventional glioma cell lines, BMC Bioinform. 19 (2018) 436.
[39] P. Paci, G. Fiscon, SWIMmeR: an R-based software to unveiling crucial nodes in complex biological networks, Bioinformatics 38 (2022) 586–588.
[40] R.M. Young, A. Jamshidi, G. Davis, J.H. Sherman, Current trends in the surgical management and treatment of adult glioblastoma, Ann. Transl. Med. 3 (2015).
[41] K. Anjum, et al., Current status and future therapeutic perspectives of glioblastoma multiforme (GBM) therapy: a review, Biomed. Pharmacother. 92 (2017) 681–689.
[42] D. Eisenbarth, Y.A. Wang, Glioblastoma heterogeneity at single cell resolution, Oncogene 42 (2023) 2155–2165.
[43] D. Sturm, et al., Hotspot mutations in H3F3A and IDH1 define distinct epigenetic and biological subgroups of glioblastoma, Cancer Cell 22 (2012) 425–437.
[44] S. Sun, et al., Protein alterations associated with temozolomide resistance in subclones of human glioblastoma cell lines, J. Neurooncol. 107 (2012) 89–100.
[45] T.J. Walter, R.K. Suter, N.G. Ayad, An overview of human single-cell RNA sequencing studies in neurobiological disease, Neurobiol. Dis. 184 (2023) 106201.
[46] B. Yang, et al., Advancements in single-cell RNA sequencing research for neurological diseases, Mol. Neurobiol. 61 (2024) 8797–8819.
[47] Y. Fang, et al., Integrating large-scale single-cell RNA sequencing in central nervous system disease using self-supervised contrastive learning, Commun. Biol. 7 (2024) 1107.
[48] R. Duan, et al., Evaluation and comparison of multi-omics data integration methods for cancer subtyping, PLoS Comput. Biol. 17 (2021) e1009224.
[49] X. Liu, et al., Spatial multi-omics: deciphering technological landscape of integration of multi-omics and its applications, J. Hematol. Oncol. 17 (2024) 72.
[50] M.M. Orabi, O. Emam, H. Fahmy, Adapting security and decentralized knowledge enhancement in federated learning using blockchain technology: literature review, J. Big Data 12 (2025) 55.