Computational and Structural Biotechnology Journal 19 (2021) 949–960

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

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

Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis

Otília Menyhárt [a,b], Balázs Győrffy [a,b,*]

[a] Semmelweis University, Department of Bioinformatics and 2nd Department of Pediatrics, H-1094 Budapest, Hungary
[b] Research Centre for Natural Sciences, Cancer Biomarker Research Group, Institute of Enzymology, Magyar tudósok körútja 2., H-1117 Budapest, Hungary

ARTICLE INFO

Article history:
Received 2 June 2020
Received in revised form 5 January 2021
Accepted 8 January 2021
Available online 22 January 2021

Keywords:
Data integration
Genomics
Transcriptomics
Proteomics
Metabolomics
Driver mutation
Biomarker
Breast cancer
Lung cancer

ABSTRACT

While cost-effective high-throughput technologies provide an increasing amount of data, the analyses of single layers of data seldom provide causal relations. Multi-omics data integration strategies across different cellular function levels, including genomes, epigenomes, transcriptomes, proteomes, metabolomes, and microbiomes offer unparalleled opportunities to understand the underlying biology of complex diseases, such as cancer. We review some of the most frequently used data integration methods and outline research areas where multi-omics significantly benefit our understanding of the process and outcome of the malignant transformation.

We discuss algorithmic frameworks developed to reveal cancer subtypes, disease mechanisms, and methods for identifying driver genomic alterations and consider the significance of multi-omics in tumor classifications, diagnostics, and prognostications. We provide a comprehensive summary of each omics strategy's most recent advances within the clinical context and discuss the main challenges facing their clinical implementations.

Despite its unparalleled advantages, multi-omics data integration is slow to enter everyday clinics. One major obstacle is the uneven maturity of different omics approaches and the growing gap between generating large volumes of data compared to data processing capacity. Progressive initiatives to enforce the standardization of sample processing and analytical pipelines, multidisciplinary training of experts for data analysis and interpretation are vital to facilitate the translatability of theoretical findings.

Contents

* Corresponding author at: Semmelweis University, Department of Bioinformatics and 2nd Department of Pediatrics, Tűzoltó utca 7-9, H-1094 Budapest, Hungary.
E-mail address: gyorffy.balazs@med.semmelweis-univ.hu (B. Győrffy).

https://doi.org/10.1016/j.csbj.2021.01.009

1. Introduction

The development of malignant transformations requires molecular alterations at many levels. Single-level omics approaches interrogating entire pools of genomes, epigenomes, transcripts, proteins, microbiomes, and metabolites with increasingly affordable high-throughput technologies are attempting to untangle mechanisms of cancer development [1]. The continued reduction of cost and processing time of omics-based approaches prompted an explosion of big data within each field and transformed hypothesis-driven targeted investigations toward data-driven untargeted analyses. Nonetheless, single-level omics approaches lack the resolving power to establish causal relationships between molecular alterations and phenotypic manifestations. In contrast, systems biology integrates multidisciplinary information and holds a great promise to understand biological interactions holistically and systematically [2]. Integration of regulatory layers could be particularly suitable to dissect aberrant cellular functions behind complex diseases, such as cancer [3]. Measuring biological samples on multiple omics scales enables a better understanding of how genetic variants, the environment, and the interaction of the two perturb complex biological systems. Multi-omics data analysis improves the clustering of samples into biologically meaningful groups, provides a greater understanding of prognostic and predictive phenotypes, dissects cellular responses to therapy, and assists translational research by integrative models [4]. Here, we aim to summarize the strength of this global approach, with a particular focus on the novel insight brought by multi-omics to cancer modeling.

We discuss some of the most frequently used data integration methods in oncology and investigate the potential of multi-omics in the functional identification of driver genomic alterations, tumor classifications, prognostications, and diagnostics, predominantly based on the integration of genomics, epigenomics, transcriptomics, proteomics, and metabolomics. We summarize the most recent advances of single omics strategies within the clinical context and discuss the main challenges facing multi-omics' clinical implementations.

2. Multi-Omics data integration approaches

There is an enormous diversity in the variety of approaches for integrating multidimensional omics data [3,5,6]. Comparing technical parts of actual data integration models and statistical methods is beyond this minireview's intended scope. Excellent comprehensive descriptions from several perspectives are available elsewhere [6–11], discussing data resources [6], multi-omics fusion methods ideal for matched samples [7], comparing supervised, semi-supervised, and unsupervised integrative approaches [8], examining the development of standardized analytical pipelines [10] or highlighting critical issues regarding the use of single- vs. multi-omics strategies [11]. Numerous analysis frameworks are mutation-centered, aiming to identify genetic determinants of phenotypic traits and to distinguish driver and passenger mutations [6].

Analysis of omics data can be approached from two standpoints: a bottom-up and a top-down integration strategy [6]. According to the hypothesis-driven bottom-up approach, multiple data types are combined first, followed by manual integration of separate clusters [12]. In contrast, powerful top-down approaches incorporate all data types simultaneously and allow data integration and dimensionality reduction at the same time [6]. Integrative methods may involve unsupervised, exploratory analysis, supervised, predictive, regression analysis, or semi-supervised analysis. In the unsupervised models, inference from input variables is drawn without labeled response variables. [9,13]. Data integration algorithms can also broadly be classified as fusion-based, network-based, Bayesian, similarity-based, correlation-based, and other multivariate methods, although many tools use a combination of approaches [14].

Here we provide a brief overview of the most frequently used multi-data integration strategies involving cancer genomics, transcriptomics, epigenomics, proteomics, and metabolomics. We focus on approaches utilizing parallel integration of data sets that allow integrating at least two omics data sets derived from at least partially overlapping samples and are readily available in tools/packages.

3. Methods to reveal cancer subtypes and disease mechanisms

3.1. Multivariate methods for data integration

Joint non-negative matrix factorization (**NMF**) is the most straightforward method of unsupervised multi-omics data integration, which is based on decomposing a non-negative matrix into non-negative loadings and non-negative factors. The method projects multiple data types to a common coordinate system where heterogeneous variables projecting toward the same direction form a module. Integrating mRNA and microRNA expression with methylation data with NMF in ovarian cancer samples from the TCGA revealed novel signaling pathway perturbations and clinically distinct patient subgroups [12]. The time and memory-consuming nature of the NMF method represents its major drawback and require non-negative input matrices and proper normalization steps.

The unsupervised exploratory Joint and Individual Variation Explained (**JIVE**) method represents an extension of principal component analysis. The approach decomposes the data input, such as a gene expression data matrix into shared common factors between data types, data-specific variation within each data type, and residual noise. JIVE estimates common features more accurately, although outliers compromise robustness. A JIVE analysis

integrating miRNA and gene expression in glioblastomas improved characterization of tumor types [15]. **MoCluster** can discover joint patterns across multiple omics data by employing a multiblock multivariate analysis, followed by clustering. By integrating mRNA, protein, and methylation data, moCluster differentiated microsatellite instability-high tumors along with three novel subtypes in colorectal carcinoma [16]. In summary, the common denominator of the above methods is that all of them put multi-omics datasets into a stacked matrix, and the matrix provides input for the subsequent cluster analysis.

3.2. Statistical methods for data integration

Adopting the Bayesian framework allows assumptions on different data types with various distributions and correlations among data sets. Robust clustering strategies, such as **iCluster**, incorporate multiple genomic characteristics without non-negative constraints [17] and use a Gaussian latent variable model where the latent variables form a set of principle coordinates collectively capturing the correlative structure of multi-omics data. The iCluster method aims to obtain joint clustering of samples and identify cluster-relevant features across data sets: unsupervised clustering of paired CNV and gene expression profiles revealed novel subgroups of breast cancer with distinct clinical outcomes beyond the classic expression subtypes [18]. iCluster, initially formulated for continuous data types, has been upgraded to accommodate binary, sequential, categorical, and continuous variables with different modeling assumptions that arise from genomic, epigenomic, and transcriptomic profiling (**iCluster+**). The iCluster+ tool decomposes each omics data type into a components factor (latent cancer subtype) and loading factors (gene features) based on assumptions for different omics data types. Compared to the three other methods, iCluster+ provided the most accurate classification for simulated datasets when sample labels were not known before integration [19]. The method, though, has several limitations: it needs to tune the model parameters for optimal parameter estimation and requires lots of computations, and does not provide an evaluation of statistical significance for the selected features [19,20]. A more recent version of the iCluster+, **iClusterBayes,** uses a Bayesian integrative clustering approach to identify tumor subtypes and overcomes the limitations of the iCluster+. The method demonstrated excellent performance in revealing clinically meaningful tumor subtypes and driver omics features in glioblastoma and kidney cancer data [21].

Unsupervised network-based approaches are mostly applied to identify co-expression network modules or significant genes within signaling pathways. A popular probabilistic graphical model (PGM) based framework called PAthway Representation and Analysis by Direct Reference on Graphical Models (**PARADIGM**) incorporates curated pathway interactions among genes. Databases that include interaction-topology among genes, such as KEGG, may be exploited for data interpretation. The approach is based on factor graphs that model gene expression and activity as a set of interconnected variables, where genes are represented by nodes and links between genes by edges. The model can incorporate many types of omics data, including mutations, mRNA and miRNA expression, promoter methylation, and DNA copy number alterations [22]. However, PARADIGM requires accurate information about biochemical magnitudes of interactions that may not be available [11]. The method successfully identified altered activities in cancer-related pathways in GBM and breast cancer datasets. Based on pathway perturbations, it divided GBM patients into clinically relevant subgroups with different survival outcomes with accuracy superior to gene expression-based signatures [22]. Important to emphasize an additional finding with high clinical relevance: in high-grade serous ovarian adenocarcinomas, PARADIGM uncovered defects of homologous recombination in about half of the tumors, rendering them good candidates for PARP inhibitors [23].

The unsupervised Bayesian Consensus Clustering (**BCC**) method is based on an extended Dirichlet mixture model that seeks source-specific clusters within each data type simultaneously and performs *post-hoc* integration of separate clusters. This flexible method allows simultaneous modeling of both the dependence and heterogeneity of various data and can be utilized to integrate gene expression, miRNA expression, methylation status, and proteomics, nevertheless does not convey the critical genes associated with the clustering [24]. The Multiple Dataset Integration (MDI) method combines many different data sets and data types simultaneously and captures the underlying structural similarity based on unsupervised integrated clustering. The MDI method does not seek to find joint sample clusters; instead, datasets are modeled using a Dirichlet-multinomial allocation (DMA) mixture model. Different datasets can have a different number of clusters, and clustering of genes in one data set influences the clustering of genes in another data set [25]. The MDI separated eight distinct consensus subtypes of glioblastomas by combining gene expression, CNV, miRNA, and methylation data [26]. Both methods (BCC and MDI) perform clustering on every single omics dataset and combine the primary results into a final clustering assignment.

To identify biomarkers associated with clinical outcome, the integrative Bayesian analysis of genomics data (**iBAG**), a supervised multi-step method, considers biological relationships across data from different omics platforms and applies hierarchical modeling. The first step is a regression model partitioning data into principal components. Clinical data and survival information are included in a joint regression, including factors from the first step. Integrating gene expression and methylation data with iBAG in glioblastoma samples helped to define new methylation-regulated genes associated with patient survival [27].

3.3. Network-based integration

iOmicsPASS performs a supervised integration of DNA copy number, transcriptomics, and proteomics data by computing biological interaction scores for all molecular interactions in the network for predictive subnetwork discovery [28]. The method uses a shrunken gene-centroid algorithm to discover interactions whose joint expression patterns predict phenotypic groups the best and treats all network data as undirected. iOmicsPASS has been tested on invasive ductal breast cancer data of TCGA for the discovery of predictive subnetworks. The method successfully overcame the heterogeneity of data sets and identified the distinct molecular signatures specifying different breast cancer phenotypic groups. iOmicsPASS also discovered a new transcriptional regulatory network underlying the basal-like subtype, a result not seen through an analysis of individual omics data [28]. The method's significant advantage is that the selected predictive signatures form densely connected subnetworks limiting the search space of predictive features. It also works well for data sets with a modest sample size.

3.4. Fusion-based integration

Pattern Fusion Analysis (**PFA**) can identify significant sample patterns from different omics profiles by automated information alignment and bias correction. PFA fuses local patterns from each data type into a global sample pattern corresponding to phenotypes. The method measures each data type's contributions and identifies significant sample-patterns from different omics profiles to reveal shared sample patterns. PFA was able to identify clinically distinct subtypes in clear cell carcinoma, lung squamous cell carcinoma, and glioblastoma samples from the TCGA with clustering similar

to SNF and iCluster, but with higher prognostic efficiency [29]. However, identifying novel biomarkers is not possible with PFA and cannot reveal insights into underlying mechanisms of tumorigenesis.

3.5. Similarity-based integration

Similarity-based methods work with inter-patient similarities. Similarity network fusion (**SNF**) constructs individual networks per omic and iteratively updates these networks to increase their similarity until they converge into a single network. With each iteration, the fusion steps eliminate weak connections [30]. Analysis of DNA methylation, mRNA and miRNA expression patterns with SNF across 215 glioblastoma samples outperformed single data type analysis and identified three separate clusters, including one of younger patients with an IDH subtype and favorable prognosis and another subtype with a positive response to temozolomide [30]. SNF helps in identifying cancer subtypes but can not be applied for the identification of biomarkers.

A popular machine-learning-based biomedical data fusion method is multiple kernel learning that uses a predefined set of kernels to combine data from different sources. An unsupervised version by Speicher and Pfeifer combines multiple kernel learning with a graph embedding framework algorithm called *Locality Preserving Projections* for dimensionality reduction for the clustering of samples and the analysis of follow-up data (called Regularized Multiple Kernel Learning Locality Preserving Projections or **rMKL-LPP**). The method's advantage is that input data types can be numerical and sequence matrices, and the framework remains stable for small datasets. It is also possible to input several kernel matrices per data type. Using gene expression, miRNA, and methylation data rMKL-LPP displayed concordance to previous clustering results in glioblastoma multiform [31].

Frequently data from different omics platforms are not available for each sample, in which case clustering approaches are restricted to sub-cohort of samples. The NEighborhood Based Multi-Omics Clustering (**NEMO**) circumvents partial omics-data challenges and performs similarity-based multi-omics clustering without imputation or reducing sample numbers [32]. NEMO builds on previous similarity-based methods, such as SNF and rMKL-LPP, but does not require iterative optimization and is faster. It works in three phases: an inter-patient similarity matrix is built, followed by integrating into a single matrix. Finally, the resulting network is clustered [32]. Using partial datasets from TCGA AML samples, NEMO performed data clustering highly correlated with prognosis. Extensive testing on full data spanning over three thousand patients samples in 10 cancer types revealed results comparable to previous data integration methods [32]. However, NEMO is not suitable for biomarker discovery.

3.6. Correlation-based integration

Canonical correlation analysis (**CCA**) is typically used to assess correlation across CNV, methylation, and gene expression data and may provide insight into the mechanisms of carcinogenesis. CCA performs individual feature selection while also incorporates group effects of features into the correlation analysis [33,34]. CCA selected discriminative features from multi-omics data sources to predict survival in kidney renal clear cell carcinoma [35] but has low applicability in molecular subtype assessment and biomarker selection.

4. Methods for the identification of driver genomic alterations and cancer biomarkers

The complicated process of cancer initiation and metastasis involves multiple pathways conferring heterogeneity in patient outcomes. There is extensive genetic diversity between tumors of the same cancer types, and genetic aberrations can also be highly diverse within subclones of the same tumors [36]. Typical tumors contain between 2 and 8 driver gene mutations, encompassing about 0.1% of all mutations during cancer progression [37]. Next-generation sequencing (NGS) technology coupled with increased computing capacity can identify all mutations in a genome; nevertheless, the challenge remains to distinguish pathogenic genomic mutations from passenger alterations. Driver somatic aberrations are assumed to alter the downstream transcriptomic network providing selective advantage, while passenger mutations are not expected to modify the phenotype. Downstream alterations can also include therapeutically relevant alterations, like changes in the expression of immunotherapy targets [38]. Integration of mutational profiles and gene-expression patterns could amplify relevant signals related to tumorigenesis: integrated DNA and RNA sequencing are particularly useful for identifying relevant somatic mutations in low purity tumors [39].

Biomarker validation is a time-consuming and costly process; therefore, selecting promising candidates *in silico* is a viable concept. Various algorithmic frameworks have been developed to exploit associations between genomic aberrations and downstream alterations. Unbiased methods do not depend on pre-existing knowledge about genetic interactions and make inferences directly from data. **MuTarget** is such a model-free cancer biomarker discovery tool that identifies genes with altered expression in patient samples harboring a particular mutation. Inversely, the tool can also identify mutations related to over-or underexpressed genes of interest and provides a rapid method to filter out suitable candidates for experimental follow-up. MuTarget is broadly accessible as a registration fee, automated, online tool (www.mutarget.com). With the incorporation of 7876 solid tumor samples representing 18 different tumor types, the platform provides sufficiently robust interaction networks for data integration (Nagy and Győrffy, 2021 [40]). A similar previous analysis of KRAS mutation-driven expression profiles demonstrated high predictive power in lung cancer [41].

Masica and Kachin developed another model-free cancer biomarker discovery – they use a model-free matrix-based computational method to identify potentially cancer-specific mutations from correlations between mutations and gene expressions. The method was able to identify mutations associated with drastic changes in gene expression based on the interrogation of 149 glioblastoma samples from the TCGA [42].

A different set of bioinformatics approaches utilize known biological pathway information. **DriverNet** allows individual mutations to be associated with coincidence changes in expressions of their known interacting partners based on "influence graphs" where nodes represent genes with mutations or outlying expression status, and edges capture their interactions. Interacting partners are extracted from a known pathway or gene set databases [43]. The major drawback of the framework is its restriction to only direct interactions.

The Network-based Integration of Multi-omics Data (**NetICS)** is a graph diffusion-based model capturing the directionality of interactions. NetICS predicts how aberrant genes affect other genes' expression by identifying mediators that orchestrate downstream expression changes and are located between aberrant and differentially expressed genes. The model accommodates diverse data types, including somatic mutations and gene expressions, while CNVs, miRNA expressions, methylation patterns, and protein expressions can also be integrated. The method ranks genes proximal to upstream genetic aberrations and downstream differentially expressed genes. Proteins for each sample are subsequently combined with a robust rank aggregation technique. NetICS was

able to successfully prioritize cancer genes in five cancer types [44].

iCluster+ can identify genomic features that contribute most to the biological variation with a lasso regression [20]. Integration of copy number variation, gene expression, and mutation data of small cell lung cancer cell lines from the Cancer Cell Line Encyclopedia Data Application (CCLE) dataset identified novel potential drivers genes, including SHISA5 (Scotin), a p53-inducible ER stress protein, and gastrin-releasing peptide (GRP) [20]. Integration of transcriptomic, proteomic, genomic, and methylation data on various adult soft tissue sarcomas with **iCluster** defined prognostically distinct subsets within individual subtypes, particularly among dedifferentiated liposarcomas and soft tissue leiomyosarcomas. Immune infiltration scores in the tumor microenvironment based on the expression of genes involved in immune response and inflammation were highly associated with clinical outcome, thus offering potential biomarkers of the efficacy of immune checkpoint inhibitors [45].

Another unsupervised model-based method, the Multi-Omics Factor Analysis (**MOFA**), can detect principal sources of biological and technical variation in multi-omics data as a set of hidden factors and can cope with missing values. [46]. MOFA utilized for integrating data on somatic mutations, gene expression, methylation, and drug response in 200 chronic lymphocytic leukemia samples was able to identify major dimensions of heterogeneity and novel disease drivers, such as response to oxidative stress, enhancing prediction accuracy of clinical outcomes [46].

Altogether, the advantage of multi-omics approaches in defining driver genomic alterations is an emerging and actively developing area. The previously established lists of cancer hallmark genes [47] facilitate linking such driver events to biologically important signatures.

The above list of multi-omics frameworks is by no means exhaustive, but provides a selection of approaches that are i) suitable for the integration of the multi-data of interest, ii) gained considerable popularity by the cancer research community, and iii) were able to deliver clinically useful results. General features of each method are summarized in Table 1.

In the following sections we discuss the clinical merits of individual omics approaches in oncology and outline research areas where multi-omics data integration may facilitate our understanding of the process and outcome of malignant transformation.

5. Clinical translation of “single omics” approaches in oncology

To date, the main focus of translational research was to connect disease phenotype to the genotype. Genomics contributed to discovering major disease subtypes via the corresponding genetic mutations segregating subtypes and supports the revelation of actionable therapeutic targets that predict the effectiveness of directed interventions and altered everyday tumor-specific treatment approaches [48]. Besides individual targetable alterations, genomics can assess mutational signatures and mutational load to predict immune-checkpoint inhibitors’ effectiveness [49]. The applicability of genomic methods increases in the clinic in areas such as monitoring treatment response and characterization of resistance mechanisms [50]. Nevertheless, genetic reports focus mainly on exome data, SNPs, and pharmacogenomics risk variants that constitute only about 3% of the genome [51]. In complex diseases, it is difficult to establish a clear relationship with specific genetic variants; thus, genomics is only the starting point to tackle the cancer challenge.

In contrast to the largely identical DNA across different cells of an organism, the transcribed RNA is highly dynamic and reflects the diversity of cell types and cellular states. Detecting aberrant transcription in cancer is increasingly incorporated into clinical management: mRNA-based multigene panels relying on RT-qPCR technology such as the 21-gene expression assay OncotypeDX or the 70-gene-based MammaPrint support treatment decision in breast cancer [51,53]. RNA-seq expands beyond the measurement of expression of protein-coding genes and offers a comprehensive transcriptomics profiling to explore novel and known transcripts, isoforms, splice variants, SNPs, and chimeric gene fusions with high sensitivity and accuracy. The most immediate application of RNA-seq in cancer management is the cost-effective and unbiased detection of gene fusions: the FoundationOne Heme assay has been successfully implicated in the detection of BCR-ABL1 fusions in hematologic malignancies [54], IGH–MMSET fusions in multiple myeloma or oncogenic TRK fusions in sarcomas [55].

Chemical modifications of DNA, nuclear RNA, histones, and non-histone chromatin proteins may affect gene expression without altering the base sequence [56]. Epigenetic marks are tissue-specific and strongly depend on environmental cues or disease-related modifiers, linking genome and environment, thus providing potential biomarkers for personalized medicine [57]. The clinical applicability of epigenomics is an active field of cancer research as specific therapies may reverse some epigenetic modifications. For example, the lysine demethylase 3A (KDM3A) controls transcriptional networks, and its activity is deregulated in several cancers. Chromatin immunoprecipitation, combined with next-generation sequencing (ChIP-Seq) matched with gene expression profiles, revealed that KDM3A acts as a crucial transcriptional coactivator for the androgen receptor in prostate cancer cells [58]. The epigenetic modifier EZH2 has been implicated in silencing tumor suppressor genes. Based on 471 cases from the TCGA database, activated EZH2 was identified in about 20% of melanoma patients due to mutations, amplification, and increased transcription. These alterations were associated with DNA hypermethylation and adverse prognosis, but treatment by the EZH2 inhibitor GSK126 reversed transcriptional repression, suggesting a promising therapeutic avenue [59].

Proteomics elucidates the actual protein products and post-translational modifications present in the cell from a small amount of body fluids or tissue samples and provides information about the proteome’s temporal and spatial organization, including localization and interaction among protein products [60]. In precision cancer medicine, proteomics’ potential is increasing: in 2016, the first cell-free blood-based protein microarray diagnostic tests were introduced for early-stage breast cancer, promising to reduce the number of unnecessary breast biopsies by 67% [61]. Proteomics may enhance patient stratifications: recent quantitative proteomics and phospho-proteomic profiling enabled the classification of early-stage hepatocellular carcinomas into molecular subclasses with different clinical outcomes and potential therapeutic targets [62].

Metabolomics, a comprehensive analysis of hundreds to thousand metabolites in a biological fluid, cell, or tissue at a given instant (metabolome), started to gain importance in precision medicine, particularly cancer biomarker discovery [63]. Cells react to changing environments via the integrated actions of signaling, transcriptomic, and metabolic networks. Thus, the metabolome provides a direct readout of physiological changes while also allows inferences about upstream alterations. Metabolites reflect underlying biochemical processes related to internal (genetic) and external (environmental) factors, indicating cells/tissues’ actual state. Metabolomic profiling of cancer cells led to discovering key oncometabolites [64] and may be a non-invasive tool for discriminating cancerous tissue or subgroups of tumors. For instance, a Nuclear Magnetic Resonance (NMR) based metabolomics study identified higher concentrations of pyruvate and gluta-

Table 1
Selected methods for multi-omics data integration.

Name	Category	Method	Example (cancer type)	Results of data integration	Data type	User-friendliness	Computational platform	References
Joint NMF	unsupervised	matrix factorization	ovarian cancer	cancer subtyping	Multi-data	difficult	Python	Zhang et al., 2011, 2012
iCluster+	unsupervised	matrix factorization	colorectal carcinoma	cancer subtyping	Multi-data	difficult	R	Mo et al., 2013
iClusterBayes	unsupervised	matrix factorization	glioblastoma, kidney cancer	cancer subtyping, disease drivers	Multi-data	difficult	R	Mo et al., 2018
moCluster	unsupervised	matrix factorization	colorectal carcinoma	cancer subtyping	Multi-data	difficult	R	Meng et al., 2016
JIVE	unsupervised	matrix factorization	glioblastoma	cancer subtyping	Multi-data	difficult	MATLAB	Lock et al., 2013
MOFA	unsupervised	PCA	chronic lymphocytic leukemia	novel disease drivers	Multi-data	difficult	R/Python	Argelaguet et al., 2018
rMKL-LPP	unsupervised	multiple kernel learning, similarity-based	glioblastoma	cancer subtyping	Multi-data	difficult	available on request	Speicher and Pfeifer, 2015
NetICS	unsupervised	network-based	multiple cancers	disease drivers	Multi-data	difficult	MATLAB	Dimitrakopoulos et al., 2018
BCC	unsupervised	Bayesian	breast cancer	cancer subtyping	EXP, MET, miRNA, proteomics	difficult	R	Lock and Dunson, 2013
MDI	unsupervised	Bayesian	glioblastoma	cancer subtyping	Multi-data	difficult	MATLAB	Kirk et al., 2012; Savage et al., 2013
PARADIGM	unsupervised	pathway networks, Bayesian	glioblastoma, ovarian cancer	cancer subtyping, therapeutic opportunities	Multi-data	difficult	Python	Vaske et al., 2010
iBAG	supervised	multi-step analysis	glioblastoma	potential biomarkers of survival	Multi-data	difficult	R	Jennings et al., 2013
SNF	unsupervised	network-based, similarity-based	glioblastoma	cancer subtyping	Multi-data	difficult	R/MATLAB	Wang et al., 2014
iOmicsPASS	supervised	network-based	breast cancer	cancer subtyping, disease drivers	Multi-data	difficult	R	Koh et al., 2019
NEMO	unsupervised	similarity-based clustering	acute myeloid leukemia	cancer subtyping	Multi-data	difficult	R	Rappoport and Shamir, 2019
PFA	unsupervised	fusion-based integration	clear cell carcinoma, lung squamous cell carcinoma, glioblastoma	cancer subtyping	Multi-data	difficult	MATLAB	Shi et al., 2017
CCA	unsupervised	correlation based	kidney renal clear cell carcinoma	mechanisms of carcinogenesis	CNV, methylation, gene expression	difficult	R	Lin et al., 2013; Zhou et al., 2015;El-Manzalawy et al., 2018

mate and decreased isoleucine concentrations in the serum of untreated CLL patients compared with controls [65].

Microbiomics is an emerging field focusing on the microbial communities colonizing our body [66]. The gut microbiome is rapidly altered by diet, drugs, or additional environmental cues, transforming the metabolome and representing a direct link with the environment. The entire microbial composition of a given body site may be investigated by 16S amplicon and shotgun metagenomics sequencing. In the context of precision medicine, three independent studies confirmed that resident gut bacteria might affect responses to cancer immunotherapy. In one study, antibiotic consumption altered responses to PD-1 blockade in lung and kidney cancer patients [67], whereas decreased effectiveness of PD-1 blockade in melanoma was linked to imbalanced gut flora [67,69]. The results suggest that maintaining a healthy commensal microbiome impacts antitumor immunity; however, microbial taxa associated with responsiveness to immune checkpoint blockade differ between studies.

Overall, data generation with increasingly affordable single omics approaches is becoming less of an issue, although each "omics" approach has its limitations [70]. In cancer genomics, the interpretation of clinical variants represents a major challenge. Results from many omics-based investigations strongly depend on the presence of given cells or tissue types in the sample. Many proteins expressed in almost all tissues hinder establishing associations specific to a given disease; therefore, careful identification and selection of particular tissues are critical. In epigenomics, a limiting element is a high tissue and temporal specificity of epigenetic factors; in microbiomics, the low abundance of microbial DNA relative to the host [70]. Sample collection, handling, and storage conditions may significantly alter the abundance of RNA and metabolites; moreover, reliable identification and classification of metabolites are still not resolved.

Isolated analysis of molecular organizations is not sufficient to fully elucidate the intricate complexities across molecular layers. Large scale NGS initiatives, such as The Cancer Genome Atlas

(TCGA) by the US National Cancer Institute, became established to collect clinical and molecular and data from a diverse array of –omics platforms (including exome-sequencing, copy number variations (CNVs), gene- and miRNA expression, DNA-methylation, protein, and phosphoprotein abundance) from thousands of patients to aid the discovery of underlying molecular mechanisms [71]. The International Cancer Genome Consortium (ICGC), a comprehensive repository for cancer-specific multi-omics datasets, provides genomic, transcriptomic, and epigenomic datasets spanning 35 tumor types. However, not all omics types are available for many samples. The Clinical Proteomic Tumor Analysis Consortium (CPTAC) conducts technically advanced proteomic analyses on TCGA samples already fully characterized at the genomic level [72]. The availability of such data repositories allows us to study interactions across DNA, RNA, and protein abnormalities systematically and illuminates their complex relationship.

In subsequent sections, we outline research areas where multi-omics data integration promises to facilitate our understanding of the molecular mechanisms of a multifaceted heterogeneous disease.

6. Promising applications of the multi-omics approach

6.1. Improving functional annotation of genomic alterations and discovery of new therapeutic opportunities

Despite extensive characterization of somatic aberrations, the list of recurrent mutations with therapeutic implications is surprisingly short in numerous cancers. Moreover, the function of genomic alterations or the combined effects of mutations is frequently poorly understood. The proteomic analysis allows direct assessment of genomic alterations and provides quantitative measures of underlying signaling pathway activity by monitoring the phosphorylation status of pathway elements [73]. Linking mutations and proteomics may reveal cryptic, polygenic cancer driver genes not previously implicated in tumor samples individually displaying only low-frequency mutations.

Cancer proteogenomics offers a multi-dimensional approach to deepen our knowledge about cancer biology and therapeutic vulnerabilities [73,75]. As a joint initiative between the TCGA and CPTAC, the program implements standardized mass spectrometry on genomically fully characterized tumor samples, focusing initially on the prospectively collected colon, ovarian, and breast cancer samples with precisely designed protocols [72]. Integrating proteomics with whole-exome sequencing, CNVs, RNA-seq, and miRNA-seq data on 110 prospectively collected CRC tumor specimens revealed increased proliferation and decreased apoptosis in colon tumors with Rb (retinoblastoma) phosphorylation. Moreover, increased glycolysis in tumors with high microsatellite instability (MSI-H) was associated with decreased CD8 T cell infiltration. The method delivered a novel potential target to overcome MSI-H tumors' resistance to immune checkpoint blockade [75].

A frequent limitation of mass spectrometry-based proteomic tumor analysis is the requirement of surgically resected fresh samples, but innovative technical solutions are developed for tissue-sparing approaches. Sathapay et al. described a proteogenomic profiling pipeline as a proof-of-principle, feasible on 25 ug peptide material from a 14 G core needle biopsy with a microscaled liquid chromatography-mass spectrometry (LC-MS/MS)-based method. The integrated proteogenomic analysis of core-biopsies in ERBB2-positive breast cancer (BC) revealed different resistance mechanisms directed toward ERBB2-related therapeutics, including the overexpression of mucin proteins, active androgen signaling, and lack of antitumor immune response in trastuzumab-resistant samples [74].

Integration of genomic, transcriptomic, and proteome data across 11 non-small cell lung cancer tumor samples, matched normal tissue, and patient-derived xenografts uncovered alterations not predicted by genomics and transcriptomics alone. The findings revealed proteome remodeling and affected proteins participating in metabolism. The implicated integration-based signatures were also associated with survival [76].

In breast cancers (BC), gene expression-based clustering patterns differentiate the four distinct molecular portraits, usually referred to as mRNA-based intrinsic subtypes (luminal A, luminal B, HER2-enriched and basal-like) that provide additional significant prognostic and predictive information to histology-based parameters [76,78]. While such multigene tests provide improved prognostic power, there are still no clinically useful prognostic signatures for ER-negative cancers, and drug-specific treatment response predictors also remain elusive [79]. Nevertheless, the integration of data across platforms, including whole-genome sequencing, miRNA-expression, DNA-methylation, CNVs, and reverse-phase protein assays, confirmed the existence of the four main BC classes [80]. Based on 77 TCGA breast cancer samples, the classification scheme was investigated on the proteome level. Unsupervised clustering of global proteome and phosphoproteome data identified basal-enriched, luminal-enriched, and stromal-enriched clusters, where basal- and luminal-enriched proteome subtypes overlapped with the mRNA-based PAM50 categories, but HER2-positive samples were distributed across all three proteomic subtypes [81]. Based on the phosphorylation status of signaling pathway elements alone, the study was able to define a novel subgroup featuring a G-protein-coupled receptor cluster not identified at the mRNA level. Coexpression patterns across genes and proteins revealed subgroup-specific networks with distinct interaction patterns and identified possible druggable targets, including CDK12, TLK2, PAK1, and RIPK2. Although the patient material was limited, the study substantiated the applicability of multi-omics data integration and produced numerous hypotheses for further validations [81].

6.2. Uncovering interactions across layers of organization

6.2.1. Transcriptomics and proteomics

When integrating multiple data types, it is essential to consider the flow of information from one layer to another. There is the fundamental assumption that proteins mirror fluctuation in RNA-expression. However, the relationship between mRNA expression and protein levels is not always straightforward. The complexity of information from the genome to the transcriptome increases exponentially due to alternative splicing and further increases to the proteome due to posttranslational modifications. Four major steps determine the magnitude of protein expression in a cell: rates of transcription, mRNA degradation, and rates of translation and protein degradation. The still ongoing debate about the extent of correlation between mRNA and protein levels settled at a moderate to a poor association (with correlation coefficients $\leq$0.4) [81–85]. Utilizing a targeted proteomics approach with internal standards in contrast to previous label-free absolute protein quantifications enhanced the predictability of protein copy numbers from mRNA levels [86]. Some data suggest that gene expression is controlled at the mRNA level [81,83], while other studies indicate that translational rate is the primary factor determining protein abundance [83,85,87]. In summary, RNA-levels might correctly predict the abundance of some proteins [75,88]. Identifying genes that confer to this rule is an important step in creating frameworks for disease-specific analyses.

Moreover, genes with coordinated expression are frequently presumed to participate in the same biological processes and signaling pathways, inferring functional relationships from coexpres-

sion patterns [88], though transcriptional covariation might be accidental [89]. Global expression profiling with mass spectrometry-based technologies permits the systematic analysis of concordance between cellular mRNA levels and protein content to predict gene cofunctionality. A comparison of mRNA and protein coexpression networks in three tumor types revealed that protein profiling outperformed transcriptomic profiling in coexpression based gene function-prediction with a marked difference in network wiring: mRNA coexpression pattern was driven not only by cofunction but also by the colocalization of the genes, while protein coexpression was mainly driven by functional similarity, thus predicted biological function better. The protein coexpression network also allowed inference about novel gene-function relationships, for example, a new association between the ERBB2 gene and the lipid biosynthesis process [90].

6.2.2. Transcriptomics and epigenomics

Numerous alterations differentiate the cancer epigenome from their normal counterparts, leading to aberrant expression of tissue-specific and imprinted genes. Many studies have shown the association between DNA methylation patterns or altered histone modification and cancer progression, also reflected in transcriptome level. For instance, integrating Chip-seq and RNA-seq data from patient-derived xenografts of human papillomavirus-related head and neck squamous cell carcinoma samples revealed that H3K4me3 and H3K27ac histone marks are associated with tumor-specific expression changes in their targets, including known cancer genes such as EGFR, FGFR1, and FOXA1 [91]. However, the relation between the epigenome and transcriptome might also be discordant, and genes may exhibit unchanged expression even if their promoter is methylated. A meta-analysis integrating methylation of high-density CpG islands with gene expression across 672 matched normal and cancer samples suggests that epigenetic reprogramming by promoter hypermethylation may modify the expression of a few specific transcription factors in a tissue-dependent manner but does not necessarily induce direct inhibition of gene expression [92]. Additional multi-omics data integration studies are needed to solve the mechanisms underlying the discordance between the transcriptome and epigenome.

6.2.3. Transcriptomics and metabolomics

Integration of transcriptomics and metabolomics may yield a better understanding of tumor pathogenesis than either method alone: a joint analysis of metabolite and transcriptomic profiles of breast and hepatocellular cancer samples revealed an increase in their gene-metabolites associations compared to adjacent normal tissue. Low concentration of several cancer-related metabolites, including glucose, glycine, serine, and acetate, was associated with improved patient survival [93]. A similar approach, including metabolomics and gene expression data, was applied to reveal cancer biomarkers for prostate cancer and identified several altered metabolic pathways expressed at both metabolic and transcriptional levels. Specific metabolites such as S-adenosylhomoserine (SAH), 5-methylthioadensine (MTA), and S-adenosylmethionine (SAM), and various NAD metabolites were accumulated in prostate cancer samples compared to noncancerous tissues. Analysis of gene expression revealed elevated Glycine N-methyltransferase expression (GNMT), which is assumed to be responsible for the induction of SAH and suggested to be a tumor susceptibility gene in prostate cancer [94]. Similarly, a comprehensive analysis of metabolomics and transcriptomics identified five metabolites (bilirubin, LysoPC(17:0), n-oleoyl threonine, 12-hydroxydodecanoic acid, and tetracosahexaenoic acid) as candidate biomarkers for cervical cancer, potentially beneficial for screening and diagnosis [95].

When merged with other omics data, metabolomics may answer important questions about cancer pathophysiology. For instance, elevated levels of the oncometabolite-2-hydroxyglutarate (2HG) were identified in MYC-pathway activated, predominantly ER-negative subgroup of breast tumors and cell lines, associated with poor clinical outcome. Integration of metabolomics with genome-wide methylomics revealed a hypermethylation phenotype in breast tumors marked by elevated 2HG levels [96].

6.3. Extending tumor molecular profiling

Building tumor molecular signatures solely based on mRNA expression levels (such as Oncotype DX) miss important factors connecting genotypes and phenotypes, therefore may have limited prognostic or therapeutic relevance. Data integration across different modalities helps to connect genomic events to clinical factors and to predict the drivers of poor outcome, eventually leading to better patient stratification for therapies.

Integrating mutation, copy number, methylation, mRNA, microRNA, and proteomics datasets in colorectal cancer (CRC) identified four consensus CRC subtypes, more aligned with clinical stratification [97] compared to the previously described three transcriptomic subtypes (MSI/CIMP, invasive and CIN). The extended molecular classification may be translated to clinical tests and facilitate novel therapeutic opportunities.

Clustering of high-grade serous ovarian carcinomas (HGSC) based on the TCGA transcriptome analysis suggested four transcriptomic subtypes: differentiated, immunoreactive, mesenchymal, and proliferative, even though none showed correlation with clinical outcome [23]. Proteomic analysis of 169 HGSCs revealed exact correspondence to four of the TCGA subtypes. It also produced a fifth cluster enriched in proteins related to extracellular matrix interactions, complement cascade, erythrocyte, and platelet functions. Of note, proteome based clustering was also not associated with survival. A relatively high correlation was observed between mRNAs and proteins transcriptionally regulated in response to perturbations, such as nutrition demand. Still, a weaker relationship was observed for housekeeping and other highly stable and abundant proteins. Integrated transcriptomics and proteomics identified signaling pathways participating in angiogenesis, cell motility and migration, chemokine signaling, and adaptive immunity, differently activated in patients with diverse survival outcomes [98].

Multi-omics integration in 137 primary testicular germ cell tumors (TGCTs) with low mutational density identified distinct molecular landscapes corresponding to major histologic subtypes: seminomas, embryonal carcinomas, yolk sac tumors, and teratomas, moreover revealed a previously unappreciated diversity within seminomas. Different methylation patterns and miRNA expressions suggest a significant role of epigenetic processes across subtypes. The findings offer additional insights into TGCT tumorigenesis, providing potential new therapeutic approaches [68].

Based on cell morphology, the second most frequent (5–15%) histological subtype of breast cancer consists of invasive lobular carcinomas (ILC), with a distinct clinical course and high metastatic rate compared to invasive ductal carcinomas [99]. Multi-omics integration across genomic, transcriptomic, and proteomic data identified two robust (hormone-related and immune-related) molecular subtypes within ILC that may guide treatment decisions [100].

Integrated comparative analysis involving copy number variations, mRNA, miRNA and lncRNA expressions, and methylation data confirmed distinct patterns of genomic and transcriptomic alterations in previously identified major histologic subtypes of renal cell carcinomas (RCC). The study also revealed shared fea-

tures, including the loss of the tumor suppressor Cyclin-Dependent Kinase Inhibitor 2A (CDKN2A) gene, increased DNA hypermethylation, and increased Th2 gene expression signature, associated with poor prognosis across all histologic RCC subtypes [101].

6.4. Assisting early cancer diagnosis

Early cancer detection is crucial for the timely treatment of cancer and for preventing cancer-related deaths. Methods based on non-invasive blood tests, so-called "liquid biopsy", increase relevance in identifying tumors before the appearance of symptoms [102]. The strategy offers unparalleled advantages over surgical biopsies, as tumor tissue, if accessible at all, might be extraordinarily heterogeneous or low on cellularity. Non-invasive identification of tumor-associated mutations from the circulating tumor DNA (ctDNA) released from dying tumor cells into the bloodstream shows excellent promise. However, the most significant limitation is its low proportion among all circulating cell-free DNA. The amount of detectable ctDNA depends on tumor types and stage [103], tumor burden, and applied therapy [104], among other characteristics. Thus methods based on a single tumor-associated biomarker may produce inconsistent results with limited sensitivity. Joint detection of several biomarkers, or integration of multiple methods, e.g., combining protein-DNA mutations or RNA expression and genome alterations as biomarkers of early-stage cancers, can significantly improve the detection sensitivity of liquid biopsy-based diagnosis [104,106].

Activating RAS mutations cause permanent activation of the RAS protein, providing a continuous growth stimulus, and mutations of the KRAS Proto-Oncogene are major events in pancreatic cancer coupled with worse prognosis [107]. A non-invasive blood test combined analysis of KRAS mutations and the presence of four proteins in 221 patients with pancreatic ductal adenocarcinomas. KRAS mutations were present in the plasma of only 30% of patients. However, a combined analysis of KRAS mutations and the four protein biomarkers reached 64% sensitivity and 99.5% specificity [105]. Similarly, a multianalyte blood test, CancerSEEK combined genetic alterations and protein expression in conjunction with artificial intelligence. Based on 61 amplicons within 16 genes combined with eight protein biomarkers, the test was able to localize the cancer's organ of origin and identify the early presence of five tumor types (ovary, liver, stomach, pancreas, and esophagus cancer) with sensitivity ranging between 69 and 98% and specificity of 99%.

AFP (alfa fetoprotein) level is a potential biomarker in hepatocellular carcinoma, although the low sensitivity (39–65%) and specificity made its applicability controversial [108]. However, integrated detection of AFP and RNA-profiles of exosomes, with particular focus on miR-122 and miR-148a expression, increased the model's discriminative ability to differentiate hepatocellular cancer from liver cirrhosis with an AUC of 0.931 (95% CI, 0.857–0.973) [109]. Tumor-associated RNA derived from exosomes (exoRNA) combined with ctDNA increased sensitivity of *EGFR* mutation detection in plasma of non-small cell lung cancer (NSCLC) patients from 82% to 98% compared to ctDNA alone [110].

These first successful studies show that data integration bears an enormous potential for practical clinical utilizations, and innovative approaches are expected to appear in diagnostic practices.

6.5. Future scopes and challenges

Translation of multi-omics technologies into accessible tools in daily medical routine is slow, slower than the general public anticipates [111]. One major obstacle for the clinical application is the uneven maturity of different omics approaches; genomics is closest to routine diagnostics, followed by metabolomics mainly because metabolite screening is already routinely adopted in clinical laboratories for inborn errors of metabolism or drug monitoring, and microbiomics. Other omics, including transcriptomics, epigenomics, and proteomics, are still behind [70].

Moreover, for most cancer types, not all omics data types are generated or may not be accessible. Despite the rapidly increasing role of metabolomics or microbiome profiling in cancer research, the large data depositories, such as the TCGA or CPTAC, are behind accommodating such data. The representation of different tumor types in multi-omics investigations is uneven: in 24 multi-omics studies, breast and prostate cancer are overrepresented, although rare cancer types, such as glioblastomas, are increasingly investigated [4]. In rare or challenging tumor types, multi-omics are particularly promising to unveil novel therapeutic opportunities.

Additional technical and biological challenges need to be mitigated before the routine usage of the multi-omics approach in clinical settings. Each omics platform has its requirements for sample treatment, and streamlining coordinated sample processing poses technical challenges, such as limited accessibility of available patient material, the lack of gold standard unified sample processing workflows, and post-processing data analysis protocols including normalization, transformation, and scaling to ensure robustness, reproducibility, and comparability across studies [10].

There is also an increasing gap between data generation and interpretation [1]. Technological development keeps extending the scope and complexity of generated data, and the growing complexity of algorithmic examinations requires time and resources. Integration and interpretation of diverse layers of multi-omics outcomes into predictive computational models require enormous infrastructure, computational power, storage capacity, and multidisciplinary teams with the appropriate background to translate raw sequences into meaningful clinical interpretations. Despite technological advances, data generation and long term storage remain expensive, and classic research laboratories frequently do not possess the necessary storage and computational infrastructure for processing large and complex data volumes. Higher data costs are predicted to emerge from data analysis compared to the generation of raw data. There are valid concerns that multi-omics based personalized medicine could eventually be restricted to wealthier nations [1].

Another concern is data sharing and archiving: reference databases need to accommodate big data with rigorous format standards and appropriate data security [111]. Moreover, clinically comprehensive investigations require integrating multi-omics data with other health-related information and lifestyle choices from electronic patient records, emphasizing data security regulations. Cloud computing may offer solutions for the generation and handling of large data volumes for teams without sufficient in-house computing infrastructure. Cloud-based bioinformatics tools and workflows, such as the Galaxy-project (https://usegalaxy.org/), are becoming increasingly popular for the handling and processing high throughput data.

7. Conclusions

Multi-omics offer clear advantages for translational cancer research and reveal surprising interactions "unseen" by simple correlations. First, multi-omic biomarkers could reach specificities way over previous monogenic markers, setting future research in this area. The ultimate goal is an earlier cancer diagnosis, better patient stratification, and more efficient personalized therapeutic approaches. Nevertheless, there is a growing gap between the ability to generate large volumes of omics data compared to the capacity of data integration, processing, and interpretation. Data standardization and development of central public databases for

most omics data is yet to be implemented. At the same time, the majority of tools for multi-omics integration are not robust enough, error-prone, and only available for advanced users with expertise in programming. Hopefully, progressive collaborative initiatives, like those brought to life by ELIXIR (https://elixir-europe.org/), enforcing standardization of sample processing and analytical pipelines, multidisciplinary training of experts for data analysis and interpretation, and community computing with appropriate data security regulations will accelerate translatability of theoretical findings.

Funding

The research was financed by the 2018-2.1.17-TET-KR-00001 and 2018-1.3.1-VKE-2018-00032 grants and by the Higher Education Institutional Excellence Programme (2020-4.1.1.-TKP2020) of the Ministry for Innovation and Technology in Hungary, within the framework of the Bionic thematic program of the Semmelweis University.

Author contributions

OM and BG contributed to the concept and design of the study. OM wrote the first draft of the manuscript, and BG wrote sections of the manuscript. Both authors contributed to manuscript revision, read and approved the submitted version.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors wish to acknowledge the support of ELIXIR Hungary (www.elixir-hungary.org).

References

[1] Alyass A, Turcotte M, Meyre D. From big data analysis to personalized medicine for all: challenges and opportunities. BMC Med Genomics 2015;8:33.
[2] Yan J, Risacher SL, Shen L, Saykin AJ. Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. Brief Bioinform 2018;19:1370–81.
[3] Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol 2017;18. 83-83.
[4] Chakraborty S, Hosen MI, Ahmed M, Shekhar HU. Onco-multi-OMICS approach: a new frontier in cancer research. Biomed Res Int 2018;2018:1–14.
[5] Palsson B, Zengler K. The challenges of integrating multi-omic data sets. Nat Chem Biol 2010;6(11):787–9.
[6] Yu XT, Zeng T. Integrative analysis of omics big data. Methods Mol Biol 2018;1754:109–35.
[7] Baldwin E, Han J, Luo W, Zhou J, An L, Liu J, et al. On fusion methods for knowledge discovery from multi-omics datasets. Comput Struct Biotechnol J 2020;18:509–17.
[8] Wu C, Zhou F, Ren J, Li X, Jiang Yu, Ma S. A selective review of multi-level omics data integration using variable selection. High-throughput 2019;8 (1):4. https://doi.org/10.3390/ht8010004.
[9] Richardson S, Tseng GC, Sun W. Statistical methods in integrative genomics. Annu Rev Stat Appl 2016;3(1):181–209.
[10] Misra BB, Langefeld CD, Olivier M, Cox LA. Integrated omics: tools, advances, and future approaches. J Mol Endocrinol 2018.
[11] Rappoport N, Shamir R (2018) Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic acids research 46: 10546-10562.
[12] Zhang S, Liu CC, Li W, Shen H, Laird PW, et al. (2012) Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. Nucleic Acids Res 40: 9379-9391.
[13] Huang S, Chaudhary K, Garmire LX. More is better: recent progress in multi-omics data integration methods. Front Genet 2017;8.
[14] Subramanian I, Verma S, Kumar S, Jere A, Anamika K (2020) Multi-omics Data Integration, Interpretation, and Its Application. 14: 1177932219899051.
[15] Lock EF, Hoadley KA, Marron JS, Nobel AB. Joint and individual variation explained (Jive) for integrated analysis of multiple data types. Ann Appl Statistics 2013;7(1):523–42.
[16] Meng C, Helm D, Frejno M, Kuster B. moCluster: identifying joint patterns across multiple omics data sets. J Proteome Res 2016;15(3):755–65.
[17] Shen R, Olshen AB, Ladanyi M (2009) Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics 25: 2906-2912.
[18] Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 2012;486(7403):346–52.
[19] Sathyanarayanan A, Gupta R, Thompson EW, Nyholt DR, Bauer DC, et al. (2020) A comparative study of multi-omics integration tools for cancer driver gene identification and tumour subtyping. Brief Bioinform 21: 1920-1936.
[20] Mo Q, Wang S, Seshan VE, Olshen AB, Schultz N, Sander C, et al. Pattern discovery and cancer gene identification in integrated cancer genomic data. Proc Natl Acad Sci U S A 2013;110(11):4245–50.
[21] Mo Q, Shen R, Guo C, Vannucci M, Chan KS, et al. (2018) A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. Biostatistics 19: 71-86.
[22] Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, et al. (2010) Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics 26: i237-245.
[23] Bell D, Berchuck A, Birrer M, Chien J, Cramer DW, et al. Integrated genomic analyses of ovarian carcinoma. Nature 2011;474:609–15.
[24] Lock EF, Dunson DB (2013) Bayesian consensus clustering. Bioinformatics 29: 2610-2616.
[25] Kirk P, Griffin JE, Savage RS, Ghahramani Z, Wild DL (2012) Bayesian correlated clustering to integrate multiple datasets. Bioinformatics 28: 3290-3297.
[26] Savage R, Ghahramani Z, Griffin J, Kirk P, Wild D (2013) Identifying cancer subtypes in glioblastoma by combining genomic, transcriptomic and epigenomic data.
[27] Wang W, Baladandayuthapani V, Morris JS, Broom BM, Manyam G, et al. (2013) iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data. Bioinformatics (Oxford, England) 29: 149-159.
[28] Koh HWL, Fermin D, Vogel C, Choi KP, Ewing RM, Choi H. iOmicsPASS: network-based integration of multiomics data for predictive subnetwork discovery. npj Syst Biol Appl 2019;5(1). https://doi.org/10.1038/s41540-019-0099-y.
[29] Shi Q, Zhang C, Peng M, Yu X, Zeng T, et al. (2017) Pattern fusion analysis by adaptive alignment of multiple heterogeneous omics data. Bioinformatics 33: 2706-2714.
[30] Wang Bo, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods 2014;11(3):333–7.
[31] Speicher NK, Pfeifer N. Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. Bioinformatics 2015;31:i268–275.
[32] Rappoport N, Shamir R (2019) NEMO: cancer subtyping by integration of partial multi-omic data. Bioinformatics 35: 3348-3356.
[33] Lin D, Zhang J, Li J, Calhoun VD, Deng H-W, Wang Y-P. Group sparse canonical correlation analysis for genomic data integration. BMC Bioinf 2013;14 (1):245. https://doi.org/10.1186/1471-2105-14-245.
[34] Zhou Y, Liu Y, Li K, Zhang R, Qiu F, et al. (2015) ICan: an integrated co-alteration network to identify ovarian cancer-related genes. PLoS One 10: e0116095.
[35] El-Manzalawy Y. CCA based multi-view feature selection for multi-omics data integration; 2018 30 May-2 June 2018. pp. 1-8.
[36] Burrell RA, McGranahan N, Bartek J, Swanton C. The causes and consequences of genetic heterogeneity in cancer evolution. Nature 2013;501 (7467):338–45.
[37] Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer genome landscapes. Science 2013;339(6127):1546–58.
[38] Menyhart O, Pongor LS, Gyorffy B (2018) Mutations Defining Patient Cohorts With Elevated PD-L1 Expression in Gastric Cancer. Front Pharmacol 9: 1522.
[39] Wilkerson MD, Cabanski CR, Sun W, Hoadley KA, Walter V, et al. (2014) Integrated RNA and DNA sequencing improves mutation detection in low purity tumors. Nucleic Acids Res 42: e107.
[40] Nagy Á, Győrffy B. muTarget: A platform linking gene expression changes and mutation status in solid tumors. International journal of cancer 2021;148 (2):502–11.
[41] Nagy Á, Pongor LS, Szabó A, Santarpia M, Győrffy B. KRAS driven expression signature has prognostic power superior to mutation status in non-small cell lung cancer. Int J Cancer 2017;140(4):930–7.
[42] Masica DL, Karchin R. Correlation of somatic mutation and expression identifies genes important in human glioblastoma progression and survival. Cancer Res 2011;71(13):4550–61.
[43] Bashashati A, Haffari G, Ding J, Ha G, Lui K, Rosner J, et al. DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. Genome Biol 2012;13(12):R124. https://doi.org/10.1186/gb-2012-13-12-r124.
[44] Dimitrakopoulos C, Hindupur SK, Häfliger L, Behr J, Montazeri H, et al. (2018) Network-based integration of multi-omics data for prioritizing cancer genes. Bioinformatics 34: 2441-2448.

[45] (2017) Comprehensive and Integrated Genomic Characterization of Adult Soft Tissue Sarcomas. Cell 171: 950-965.e928.
[46] Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, et al. Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol 2018;14(6). https://doi.org/10.15252/msb.20178124.
[47] Menyhárt O, Harami-Papp H, Sukumar S, Schäfer R, Magnani L, de Barrios O, et al. Guidelines for the selection of functional assays to evaluate the hallmarks of cancer. Biochim Biophys Acta 2016;1866(2):300–19.
[48] Lehmann-Che J, Poirot B, Boyer J-C, Evrard A. Cancer genomics guide clinical practice in personalized medicine. Therapies 2017;72(4):439–51.
[49] Rizvi NA, Hellmann MD, Snyder A, Kvistborg P, Makarov V, Havel JJ, et al. Mutational landscape determines sensitivity to PD-1 blockade in non–small cell lung cancer. Science 2015;348(6230):124–8.
[50] Berger MF, Mardis ER. The emerging clinical relevance of genomics in cancer medicine. Nat Rev Clin Oncol 2018;15(6):353–65.
[51] Tsai E, Shakbatyan R, Evans J, Rossetti P, Graham C, Sharma H, et al. Bioinformatics workflow for clinical whole genome sequencing at partners healthcare personalized medicine. J Pers Med 2016;6(1):12. https://doi.org/10.3390/jpm6010012.
[52] Sparano JA, Gray RJ, Makower DF, Pritchard KI, Albain KS, Hayes DF, et al. Prospective validation of a 21-gene expression assay in breast cancer. N Engl J Med 2015;373(21):2005–14.
[53] Soliman H, Shah V, Srkalovic G, Mahtani R, Levine E, Mavromatis B, et al. MammaPrint guides treatment decisions in breast Cancer: results of the IMPACt trial. BMC Cancer 2020;20(1). https://doi.org/10.1186/s12885-020-6534-z.
[54] Sonu RJ, Jonas BA, Dwyre DM, Gregg JP, Rashidi HH. Optimal molecular methods in detecting p190^{BCR-ABL} fusion variants in hematologic malignancies: a case report and review of the literature. Case Reports Hematol 2015;2015:458052.
[55] Doebele RC, Davis LE, Vaishnavi A, Le AT, Estrada-Bernal A, Keysar S, et al. An oncogenic NTRK fusion in a patient with soft-tissue sarcoma with response to the tropomyosin-related kinase inhibitor LOXO-101. Cancer Discov 2015;5(10):1049–57.
[56] Badeaux AI, Shi Y. Emerging roles for chromatin as a signal integration and storage platform. Nat Rev Mol Cell Biol 2013;14(4):211–24.
[57] Heyn H, Méndez-González J, Esteller M. Epigenetic profiling joins personalized cancer medicine. Expert Rev Mol Diagn 2013;13(5):473–9.
[58] Wilson S, Fan L, Sahgal N, Qi J, Filipp FV. The histone demethylase KDM3A regulates the transcriptional program of the androgen receptor in prostate cancer cells. Oncotarget 2017;8(18):30328–43.
[59] Tiffen J, Wilson S, Gallagher SJ, Hersey P, Filipp FV. Somatic copy number amplification and hyperactivating somatic mutations of EZH2 correlate with DNA methylation and drive epigenetic silencing of genes involved in tumor suppression and immune responses in melanoma. Neoplasia 2016;18(2):121–32.
[60] Uzozie AC, Aebersold R. Advancing translational research and precision medicine with targeted proteomics. J Proteomics 2018;189:1–10.
[61] Lourenco AP, Benson KL, Henderson MC, Silver M, Letsios E, et al. A non-invasive blood-based combinatorial proteomic biomarker assay to detect breast cancer in women under the age of 50 years. Clin Breast Cancer 2017;17:516–525.e516.
[62] Jiang Y, Sun A, Zhao Y, Ying W, Sun H, Yang X, et al. Proteomics identifies new therapeutic targets of early-stage hepatocellular carcinoma. Nature 2019;567(7747):257–61.
[63] Armitage EG, Barbas C. Metabolomics in cancer biomarker discovery: current trends and future perspectives. J Pharm Biomed Anal 2014;87:1–11.
[64] Yang M, Soga T, Pollard PJ. Oncometabolites: linking altered metabolism with cancer. J Clin Invest 2013;123(9):3652–8.
[65] Puchades-Carrasco L, Pineda-Lucena A. Metabolomics applications in precision medicine: an oncological perspective. Curr Top Med Chem 2017;17:2740–51.
[66] Petrosino JF. The microbiome in precision medicine: the way forward. Genome Med 2018;10:12.
[67] Routy B, Le Chatelier E (2018) Gut microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors. 359: 91-97.
[68] Gopalakrishnan V, Spencer CN, Nezi L, Reuben A, Andrews MC, Karpinets TV, et al. Gut microbiome modulates response to anti–PD-1 immunotherapy in melanoma patients. Science 2018;359(6371):97–103.
[69] Matson V, Fessler J, Bao R (2018) The commensal microbiome is associated with anti-PD-1 efficacy in metastatic melanoma patients. 359: 104-108.
[70] Wang Qi, Peng W-X, Wang Lu, Ye Li. Toward multiomics-based next-generation diagnostics for precision medicine. Per Med 2019;16(2):157–70.
[71] Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 2013;45(10):1113–20.
[72] Ellis MJ, Gillette M, Carr SA, Paulovich AG, Smith RD, Rodland KK, et al. Connecting genomic alterations to cancer biology with proteomics: the NCI clinical proteomic tumor analysis consortium. Cancer Discovery 2013;3(10):1108–12.
[73] Wolf-Yadlin A, Hautaniemi S, Lauffenburger DA, White FM. Multiple reaction monitoring for robust quantitative proteomic analysis of cellular signaling networks. Proc Natl Acad Sci U S A 2007;104(14):5860–5.
[74] Satpathy S, Jaehnig EJ, Krug K, Kim B-J, Saltzman AB, Chan DW, et al. Microscaled proteogenomic methods for precision oncology. Nat Commun 2020;11(1). https://doi.org/10.1038/s41467-020-14381-2.
[75] Vasaikar S, Huang C, Wang X, Petyuk VA, Savage SR, Wen Bo, et al. Proteogenomic analysis of human colon cancer reveals new therapeutic opportunities. Cell 2019;177(4):1035–1049.e19.
[76] Li L, Wei Y, To C, Zhu C-Q, Tong J, Pham N-A, et al. Integrated omic analysis of lung cancer reveals metabolism proteome signatures with prognostic impact. Nat Commun 2014;5(1). https://doi.org/10.1038/ncomms6469.
[77] Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol 2009;27(8):1160–7.
[78] Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature 2000;406(6797):747–52.
[79] Győrffy B, Hatzis C, Sanft T, Hofstatter E, Aktas B, Pusztai L. Multigene prognostic tests in breast cancer: past, present, future. Breast Cancer Res 2015;17(1). https://doi.org/10.1186/s13058-015-0514-2.
[80] Cancer Genome Atlas N. Comprehensive molecular portraits of human breast tumours. Nature 2012;490:61–70.
[81] Mertins P, Mani DR, Ruggles KV, Gillette MA, Clauser KR, Wang P, et al. Proteogenomics connects somatic mutations to signalling in breast cancer. Nature 2016;534(7605):55–62.
[82] Li JJ, Bickel PJ, Biggin MD. System wide analyses have underestimated protein abundances and the importance of transcription in mammals. PeerJ 2014;2:e270.
[83] Liu Y, Beyer A, Aebersold R. On the dependency of cellular protein levels on mRNA abundance. Cell 2016;165(3):535–50.
[84] Vogel C, Marcotte EM. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat Rev Genet 2012;13(4):227–32.
[85] Schwanhäusser B, Busse D, Li Na, Dittmar G, Schuchhardt J, Wolf J, et al. Global quantification of mammalian gene expression control. Nature 2011;473(7347):337–42.
[86] Edfors F, Danielsson F, Hallström BM, Käll L, Lundberg E, et al. Gene-specific correlation of RNA and protein levels in human cells and tissues. Mol Syst Biol 2016;12. 883-883.
[87] Wilhelm M, Schlegl J, Hahne H, Gholami AM, Lieberenz M, Savitski MM, et al. Mass-spectrometry-based draft of the human proteome. Nature 2014;509(7502):582–7.
[88] Quackenbush J. Genomics. Microarrays–guilt by association. Science 2003;302(5643):240–1.
[89] Yanai I, Korbel JO, Boue S, McWeeney SK, Bork P, Lercher MJ. Similar gene expression profiles do not imply similar tissue functions. Trends Genet 2006;22(3):132–8.
[90] Wang J, Ma Z, Carr SA, Mertins P, Zhang H, Zhang Z, et al. Proteome profiling outperforms transcriptome profiling for coexpression based gene function prediction. Mol Cell Proteomics 2017;16(1):121–34.
[91] Kelley DZ, Flam EL, Izumchenko E, Danilova LV, Wulf HA, Guo T, et al. Integrated analysis of whole-genome ChIP-Seq and RNA-Seq data of primary head and neck tumor samples associates HPV integration sites with open chromatin marks. Cancer Res 2017;77(23):6538–50.
[92] Moarii M, Boeva V, Vert J-P, Reyal F. Changes in correlation between promoter methylation and gene expression in cancer. BMC Genomics 2015;16:873.
[93] Auslander N, Yizhak K, Weinstock A, Budhu A, Tang W, Wang XW, et al. A joint analysis of transcriptomic and metabolomic data uncovers enhanced enzyme-metabolite coupling in breast cancer. Sci Rep 2016;6(1). https://doi.org/10.1038/srep29662.
[94] Ren S, Shao Y, Zhao X, Hong CS, Wang F, Lu X, et al. Integration of metabolomics and transcriptomics reveals major metabolic pathways and potential biomarker involved in prostate cancer. Molecular amp;amp; Cellular Proteomics 2016;15(1):154–63.
[95] Yang K, Xia B, Wang W, Cheng J, Yin M, Xie H, et al. A comprehensive analysis of metabolomics and transcriptomics in cervical cancer. Sci Rep 2017;7(1). https://doi.org/10.1038/srep43353.
[96] Terunuma A, Putluri N, Mishra P, Mathé EA, Dorsey TH, Yi M, et al. MYC-driven accumulation of 2-hydroxyglutarate is associated with breast cancer prognosis. J Clin Invest 2014;124(1):398–412.
[97] Guinney J, Dienstmann R, Wang X, de Reyniès A, Schlicker A, Soneson C, et al. The consensus molecular subtypes of colorectal cancer. Nat Med 2015;21(11):1350–6.
[98] Zhang H, Liu T, Zhang Z, Payne SH, Zhang B, McDermott JE, et al. Integrated proteogenomic characterization of human high-grade serous ovarian cancer. Cell 2016;166(3):755–65.
[99] Arpino G, Bardou VJ, Clark GM, Elledge RM. Infiltrating lobular carcinoma of the breast: tumor characteristics and clinical outcome. Breast Cancer Res: BCR 2004;6:R149–56.
[100] Michaut M, Chin S-F, Majewski I, Severson TM, Bismeijer T, de Koning L, et al. Integration of genomic, transcriptomic and proteomic data identifies two biologically distinct subtypes of invasive lobular breast cancer. Sci Rep 2016;6(1). https://doi.org/10.1038/srep18517.
[101] Ricketts CJ, De Cubas AA, Fan H, Smith CC, Lang M, Reznik Ed, et al. The cancer genome atlas comprehensive molecular characterization of renal cell carcinoma. Cell Rep 2018;23(12):3698. https://doi.org/10.1016/j.celrep.2018.06.032.
[102] Mattox AK, Bettegowda C, Zhou S (2019) Applications of liquid biopsies for cancer. 11.

[103] Bettegowda C, Sausen M, Leary RJ, Kinde I, Wang Y, et al. Detection of circulating tumor DNA in early- and late-stage human malignancies. Sci Transl Med 2014;6:224ra224.
[104] Tie J, Kinde I, Wang Y, Wong HL, Roebert J, Christie M, et al. Circulating tumor DNA as an early marker of therapeutic response in patients with metastatic colorectal cancer. Ann Oncol 2015;26(8):1715–22.
[105] Cohen JD, Javed AA, Thoburn C, Wong F, Tie J, Gibbs P, et al. Combined circulating tumor DNA and protein biomarker-based liquid biopsy for the earlier detection of pancreatic cancers. Proc Natl Acad Sci U S A 2017;114 (38):10202–7.
[106] Cohen JD, Li Lu, Wang Y, Thoburn C, Afsari B, Danilova L, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. Science 2018;359(6378):926–30.
[107] Buscail L, Bournet B, Cordelier P. Role of oncogenic KRAS in the diagnosis, prognosis and treatment of pancreatic cancer. Nature Rev Gastroenterol Hepatol 2020;17(3):153–68.
[108] Tian M-M, Fan Y-C, Zhao J, Gao S, Zhao Z-H, Chen L-Y, et al. Hepatocellular carcinoma suppressor 1 promoter hypermethylation in serum. A diagnostic and prognostic study in hepatitis B. Clin Res Hepatol Gastroenterol 2017;41 (2):171–80.
[109] Wang Y, Zhang C, Zhang P, Guo G, Jiang T, Zhao X, et al. Serum exosomal microRNAs combined with alpha-fetoprotein as diagnostic markers of hepatocellular carcinoma. Cancer Med 2018;7(5):1670–9.
[110] Krug AK, Enderle D, Karlovich C, Priewasser T, Bentink S, Spiel A, et al. Improved EGFR mutation detection using combined exosomal RNA and circulating tumor DNA in NSCLC patient plasma. Ann Oncol 2018;29 (3):700–6.
[111] Tebani A, Afonso C, Marret S, Bekri S. Omics-Based Strategies in Precision Medicine: Toward a Paradigm Shift in Inborn Errors of Metabolism Investigations. Int J Mol Sci 2016;17(9):1555. https://doi.org/10.3390/ijms17091555.