International Journal of
Molecular Sciences

MDPI

Review

A Detailed Catalogue of Multi-Omics Methodologies for Identification of Putative Biomarkers and Causal Molecular Networks in Translational Cancer Research

Efstathios Iason Vlachavas [1,*,†], Jonas Bohn [1,†], Frank Ückert [1,2] and Sylvia Nürnberg [1,2,*]

[1] Medical Informatics for Translational Oncology, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany; j.bohn@dkfz-heidelberg.de (J.B.); f.ueckert@dkfz-heidelberg.de (F.Ü.)
[2] Applied Medical Informatics, University Hospital Hamburg-Eppendorf, 20251 Hamburg, Germany
* Correspondence: Efstathios-Iason.Vlachavas@dkfz-heidelberg.de (E.I.V.); s.nuernberg@uke.de (S.N.)
† These authors contribution is equally to this work.

Abstract: Recent advances in sequencing and biotechnological methodologies have led to the generation of large volumes of molecular data of different omics layers, such as genomics, transcriptomics, proteomics and metabolomics. Integration of these data with clinical information provides new opportunities to discover how perturbations in biological processes lead to disease. Using data-driven approaches for the integration and interpretation of multi-omics data could stably identify links between structural and functional information and propose causal molecular networks with potential impact on cancer pathophysiology. This knowledge can then be used to improve disease diagnosis, prognosis, prevention, and therapy. This review will summarize and categorize the most current computational methodologies and tools for integration of distinct molecular layers in the context of translational cancer research and personalized therapy. Additionally, the bioinformatics tools Multi-Omics Factor Analysis (MOFA) and netDX will be tested using omics data from public cancer resources, to assess their overall robustness, provide reproducible workflows for gaining biological knowledge from multi-omics data, and to comprehensively understand the significantly perturbed biological entities in distinct cancer types. We show that the performed supervised and unsupervised analyses result in meaningful and novel findings.

Keywords: translational cancer research; oncology; multi-omics data integration; supervised data integration; unsupervised data integration; integrative methods; analysis tools; literature review; personalized medicine

Citation: Vlachavas, E.I.; Bohn, J.; Ückert, F.; Nürnberg, S. A Detailed Catalogue of Multi-Omics Methodologies for Identification of Putative Biomarkers and Causal Molecular Networks in Translational Cancer Research. *Int. J. Mol. Sci.* **2021**, *22*, 2822. https://doi.org/10.3390/ijms22062822

Academic Editor: George M. Spyrou

Received: 31 January 2021
Accepted: 5 March 2021
Published: 10 March 2021

1. Introduction

The last two decades can be characterized as the "Post Genomic Era", moving from hypothesis-driven approaches based on molecular and cellular methodologies (i.e., functional assays, genetic modifications of mice, animal modeling etc.) to discovery-driven approaches with the emergence of high-throughput methodologies and the area of functional genomics. Functional genomics is a field of molecular biology, which aims to understand the dynamic relationships between an organism's genome and its phenotype, by applying different omics technologies that utilize the continuously growing body of sequence information. The term omics describes a comprehensive quantitative characterization of a class of molecules in a given biological sample or specimen, aiming to understand the molecular mechanisms and underpinnings underlying the functioning of an organism [1,2]. Currently, there are numerous single omics approaches, investigating how distinct molecular layers contribute to the manifestation and progression of various diseases [3]. Table 1 below shows an overview of the most relevant omics data types used in translational cancer research.

Table 1. Different omics levels of gene-function relationship.

Level of Analysis	Definition	Method of Analysis
Genome [4]	Complete set of genes of an organism or its organelles	WGS, WES, DNA microarray
Transcriptome [5]	Complete set of messenger RNA molecules present in a cell, tissue of organ	RNA-Sequencing Expression microarray Spatially resolved transcriptomics
Proteome [6]	Complete set of protein molecules present in a cell, tissue or organ	Peptide/protein microarrays (RPPA) Mass spectrometry Imaging mass cytometry
Metabolome [7]	Complete set of metabolites (low-molecular-weight intermediates) in a cell, tissue or organ	Nuclear magnetic resonance spectrometry Mass spectrometry Infa-red spectroscopy
Methylome [8]	Complete set of methylation sites within a genome	Bisulfite-Sequencing, ChIP-Seq
Microbiome [9]	Complete set of genes of all microbes (bacteria, fungi, protozoa and viruses) in a cell, tissue or organ	DNA-Sequencing 16S rRNA-Sequencing
Lipidome [10]	Complete set of all biomolecules defined as lipids	Mass Spectrometry

WGS: Whole-genome Sequencing; WES: Wole-exome sequencing; ChIP: Chromatin Immunoprecipitation.

In the context of translational cancer research, high-throughput methodologies—and more recently the wave of NGS technologies—have highlighted significant genomic alterations in distinct solid tumors, and proposed perturbed molecular networks with potential impact on cancer pathophysiology [11–13]. Mutations in oncogenes and tumor suppressor genes, copy number alterations and other genetic aberrations, along with epigenomic modifications all contribute to the alteration of gene expression programs, the perturbation of normal cellular processes and the promotion of tumor formation [14]. Understanding these biological processes may enable the development of novel therapeutics and the faster detection of various types of cancers [15]. One representative example of published studies using cancer genomic data on a global scale is The Cancer Genome Atlas (TCGA) consortium, a landmark cancer genomics program funded by the National Cancer Institute in 2006 [16], and its Pan-Cancer Atlas initiative. The Pan-Cancer project is the largest and most comprehensive molecular analysis of multi-omics sequencing data and clinical annotation from more than 10,000 samples, comprising 33 of the most prevalent forms of cancer. In detail, the computational analysis with collectively 27 publications led to the identification of 299 cancer-driver genes and over 3400 driver mutations. These results shed light on the molecular underpinnings of cancer, such as cell-of-origin patterns and oncogenic processes, which classify distinct solid tumors and could serve as a valuable resource for precision medicine [17].

1.1. Limitations of Single-Omics Approaches in Complex Phenotypes

The majority of diseases and human disorders have extremely complex phenotypes, with confounding variables making it difficult to detect a clear causality [18]. Similarly, in the vast majority of cancer types studied through single-omics high-throughput experiments, the hidden biological variation can be represented as the metaphorical tip of the iceberg. For instance, extracting a list of differentially expressed genes, somatic point mutations or copy number alterations provides a limited understanding of the studied malignancy, not reflecting the total molecular complexity [19]. Additionally, various biases associated with each technology—based on both analytical and statistical thresholds and related to the experimental design—further confound each separate bioinformatics analysis [20]. Therefore, by interrogating only a single-omics experiment, one cannot identify the interplay between the different molecular entities, and thus unravel the causal mechanisms that comprehensively describe the diverse nature of each cancer. Ignoring the complexity of

underlying molecular mechanisms could lead to wrong assumptions or misinterpretation of results.

1.2. Multi-Omics Concept Introduction and Background

The multi-omics analysis approach follows the core principle that any biological condition or disease such as cancer constitutes multiple molecular phenomena, and only through the detailed understanding of the interactions between the different molecular layers can one understand holistically the significantly perturbed biological entities that characterize the specific disease [21,22].

These approaches have been employed to predict vaccine response [23] or to link complex phenotypes with multi-omics profiles in genome-wide association studies (GWAS) [24,25].

Personalized medicine can benefit from the implementation of multi-omics data integration methods, as the profound diversity in disease onset, progression and treatment outcome across cancer patients makes it difficult to decide on the optimal patient-specific treatment. Thus, joint analysis of multiple omics layers ('multi-view learning') may lead to a better understanding of heterogeneity and thus personalized treatment decisions [26–28]. Furthermore, single-cell multi-omics approaches can help to disentangle the different factors contributing to cell-to-cell heterogeneity [29], and therefore build a more complete snapshot of cancer biology, especially with respect to tumor clonality and treatment relapse [30].

1.3. Public Cancer Multi-Omics Data Repositories

As multi-omics approaches require high-dimensional data, portals hosting these data have high standards for normalization and transparent harmonization of different molecular omics modalities. In this section we would like to highlight different data repositories for multi-omics purposes and their utility for translational cancer research. For a more detailed and broader overview on leveraging distinct omics databases for personalized oncology see [31].

- GDC: The Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/) from the National Cancer Institute (NCI) is the largest scale consortium that enables the retrieval, download, comprehensive analysis and exploitation of multimodal cancer genomics studies. Within the GDC, scientists can access over 3 petabytes of data from programs like the NCI's Clinical Proteomic Tumor Analysis Consortium (CPTAC), the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) initiative and The Cancer Genome Atlas (TCGA). Both harmonized datasets and legacy data on older genome versions are available. Furthermore, it includes various data visualization tools to enhance the exploration of specific projects and cancer types, and available bioinformatics pipelines. Based on the latest data summary (October 27, 2020), this data-driven platform contains 67 projects, 68 different cancer types with more than 84 thousand cases, along with clinical data [32].
- ICGC: The International Cancer Genome Consortium was established in 2008 as an international effort to harmonize the large number of ongoing and future projects on cancer genomics. Members include the NIH, the Wellcome Trust Sanger Institute, Cancer Research UK, RIKEN, and many more. Its data portal (https://dcc.icgc.org/) currently holds 86 projects with more than 80 million somatic mutations (data release 28). Its flagship projects are the Pan-Cancer Analysis of Whole Genomes (PCAWG) and ARGO (Accelerating Research in Genomic Oncology). Launched in 2019, ARGO is the next phase of ICGC, which aims to uniformly analyze specimens from 100,000 cancer patients with high quality clinical data to address outstanding questions in genomic cancer research (https://www.icgc-argo.org/).
- PCAWG: The Pan-Cancer Analysis of Whole Genomes is the latest ICGC initiative with data released in 2020, and one of the biggest international collaborative studies, including 13 research institutes and more than 700 scientists from individual TCGA

and ICGC working groups. The major aim of the PCAWG consortium is to identify common mutational patterns and to investigate the nature and consequences of somatic and germline mutations of 38 cancer types collected from 48 TCGA & ICGC projects. The project data is available from five major resources [33]:

1. The ICGC Data Portal (https://dcc.icgc.org/repositories) as the main dissemination platform for ICGC
2. UCSC Xena (https://xenabrowser.net/) is the main exploration tool for the included multi-omics resource data, in order to identify any putative correlations among all primary results. Additionally, it includes the possibility of performing survival analyses
3. EBI Expression Atlas (https://www.ebi.ac.uk/gxa/) is an open science resource hub hosted by EMBL/EBI. It provides information about gene and protein expression across species and biological conditions such as different tissues, cell types, developmental stages and diseases.
4. BSC PCAWG Scout (https://pcawgscout.bsc.es) is another analysis platform to visualize and explore PCAWG data. It consists of a portal that presents the original omics data and sample annotation along with the results from different analysis working-groups, whereas its main focus lies on providing information on driver mutations and resulting proteins.
5. Chromothripsis Explorer (http://compbio.med.harvard.edu/chromothripsis/) is a tool that provides highly interactive Circos plots for all tumors in the PCAWG cohort. Each Circos plot reports the point mutations, small insertions and deletions, structural variations and copy number profiles detected in each tumor. On this premise, the user can exploit large-scale alterations such as chromosome arm deletions, and complex mutational patterns such as chromothripsis.

- CCLE: The Cancer Cell Line Encyclopedia (https://portals.broadinstitute.org/ccle) is an ongoing collaborative project between the Broad Institute and the Novartis Institutes for BioMedical Research. Established in 2008, its main goal is to conduct a thorough genetic and pharmacologic characterization of a large panel of 1457 cancer cell lines. Its aims are to capture the genomic heterogeneity of the preclinical models and link them with the molecular heterogeneity in cancer patients. Additionally, it can unravel clinically actionable molecular targets that might be associated with drug response and ultimately link them to cancer survival, enhancing personalized medicine. Collectively, the multi-omics cell lines dataset includes gene expression from microarray and RNA-Sequencing experiments, reverse-phase protein arrays, copy number, gene methylation and mutation data. In parallel, the database also stores legacy data, which include pharmacological profiles of 24 anticancer drugs across 504 cell lines. Besides data the web page also includes tools for data visualization, including box plots, scatter plots and bubble maps for methylation data [34].
- cBioPortal: The cBioPortal for Cancer Genomics (https://www.cbioportal.org/) is an open-source resource platform developed at Memorial Sloan Kettering Cancer Center, whereas the software is developed and maintained by various research institutes. Its main goal lies in the interrogation, interactive visualization and integrated analysis of clinical and complex multimodal cancer genomics datasets. While the major focus of the platform lies on genomic alterations (non-synonymous somatic mutations, DNA copy-number variations), it also hosts mRNA and microRNA expression, protein and phosphoprotein level data (RPPA or mass spectrometry based), DNA methylation and microbiome data, especially for TCGA data. Additionally to TCGA projects, cBioPortal includes other large-scale cancer genomics projects to advance translational cancer research, such as immunogenomic and pan-cancer studies. Overall, whereas cBioPortal is considered mainly as an exploratory analysis tool, GDC would be a more appropriate choice if the user requires full access to raw data from various cancer projects (TCGA, TARGET). Additionally, cBioPortal is currently using only

data aligned to the hg19/GRCh37 reference genome, and it doesn't provide normal tissue samples for any study [35,36].

- COSMIC: Besides cBioPortal, the Catalogue Of Somatic Mutations In Cancer (COSMIC)–developed at the Wellcome Sanger Institute—is the largest and most comprehensive resource for mining publicly available cancer sequence data, aiming to investigate the impact of somatic mutations on cancer progression and pathophysiology [37]. The latest version (COSMIC v. 92, August 2020) contains more than 37 million coding mutations and other clinical information from more than 1500 cancer types both on GRCh38 (hg38) and GRCh37 genomes. The impact of somatic variants can be summarized on various levels and across projects, such as clinical actionability (drug resistance), mutational processes associated with cancer progression (mutational signatures) and more. Furthermore, COSMIC includes the Cell Lines Project (CLP), a multidimensional dataset containing a detailed molecular characterization of more than 1000 cancer cell lines with copy number variation and gene expression data, including also other previously published moderate scale sequencing projects, such as the NCI-60 Human Tumor Cell Lines Screen dataset [38].

1.4. Platforms and Packages for Leveraging Multi-Omics Data Retrieval

Development of platforms and packages for accessing, configuration and preparation of data in the field of multi-omics data integration makes tools easier applicable and saves time for the major integration and analysis of data but is also limited to the use of specific runtime environments. The mainly used programming languages in this field are R and python (see following Section 2). Efforts to use a multi complex runtime environment by including both languages have led to the development of Python-R interfaces like rpy2 (https://pypi.org/project/rpy2/) and reticulate (https://rstudio.github.io/reticulate/) which have been used for multi-omics data integration, especially for combining machine learning computations and data mining approaches [39]. The main innovation and development in open source machine learning platforms like TensorFlow [40] and PyTorch [41] make python the language of choice for ML-development and related applications.

In addition, recent implementations in R led to integration of functionalities from those two big platforms to make ML-development also available for R users (https://github.com/rstudio/tensorflow, https://github.com/f0nzie/rTorch). The main source of development for bioinformatics packages in R is the Bioconductor software platform. It is an open source and open development project, which provides tools for the exploration and analysis of high-throughput omics data. It is based on the R programming language, and among its main priorities are reuse and interoperability, along with high-quality documentation. In R, the fundamental unit of sharable code is the R package, which combines code, data, tests and vignettes, which are extensive documents illustrating how to use the corresponding package. The latest version (3.12) includes 1974 packages, covering a broad range of bioinformatics and statistical applications for sequencing data (RNA-Seq, ChIP-Seq, variant annotation etc.), microarrays, flow cytometry, imaging and proteomics [42,43]. Regarding translational cancer research, there are a number of important R packages that facilitate the management, assessment and download of TCGA data from the aforementioned public data resources. These include but are not limited to GenomicDataCommons, TCGAbiolinks, cBioPortalData and curatedTCGA R packages, with varying strengths in ease-of-use, integration, and completeness of data. For example, GenomicDataCommons [44] offers full access to all available files from the TCGA and other studies. TCGAbiolinks [45] additionally reduces the burden of computational time and data processing when starting from raw or not fully transformed data, by providing a single-omics data type harmonization with the "SummarizedExperiment" data container, along with the accompanied clinical data for the selected cancer studies. Furthermore, the R package curatedTCGAData [46] aims at balancing interoperability with complexity, by offering an integrative and user-friendly representation of multimodal TCGA data for download in Bioconductor [47]. The package is based on the MultiAssayExperiment

(MAE) software, an integrative representation for multi-omics data studies. MAE is a Bioconductor object-oriented S4 class general data structure, which is modelled after the "SummarizedExperiment" representation for expression data and coordinates multi-omics experiments on a set of biological specimens [48]. Moreover, MAE can incorporate any number of assays with distinct representations and dimensions. Assays have to be either "range-based" (measurements relate to genomic ranges such as gene expression or copy number) or "ID-based" (measurements are indexed identifiers of genes, proteins, microRNAs, etc.). The package curatedTCGAData can yield and construct "on the fly" MAE representations from flat files of 33 different cancer types from the Broad GDAC Firehose (hg19 data). Finally, the cBioPortalData package provides an R/Bioconductor interface to fetch, expose and utilize cBioPortal cancer data. It imports cBioPortal datasets as MultiAssayExperiment objects into Bioconductor, in order to construct integrative representations of multi-layered studies. Moreover, cBioPortalData implements two main approaches for accessing the data: one for downloading pre-packaged and another for sending queries through the cBioPortal API. One current limitation is that the user can only query specific gene panel combinations within a study.

On the other hand, python-based development in this field has also led to numerous useful tools for accessing, preprocessing, analyzing, and integrating multi-omics data from cancer repositories. Such tools like TCGAIntegrator [49], PyGDC (https://github.com/hammerlab/pygdc), xenaPython [50] and OpenOmics [51] helped by accessing APIs, prepare and integrate multi-omics data from widely used web platforms such as TCGA/GDC or cBioPortal in various studies [52–54]. To date, based on the Python Package Index (PyPI) repository, 197 bioinformatics related projects for multi-omics data are currently stored (https://pypi.org, accessed: 4 March 2021). Finally, besides the Bioconductor and Python projects, also other web based curated platforms provide tools to perform multi-omics data analysis. The most representative example is the online platform Galaxy [55], including various interfaces for the integrative visualization and exploration of multi-modal layers, such as the Multi-omics Visualization Platform (MVP) plugin suited for proteogenomic data analysis [56].

1.5. Challenges Integrating Multi-Omics Experiments

Despite the wealth of different cancer omics layers deposited in the aforementioned databases, there are some noticeable challenges regarding their efficient integration and interpretation. Firstly, one of the major bottlenecks is the multi-layered data acquisition. Heterogeneous data collected using different techniques (i.e., data modalities) generally exhibit distinct statistical properties (discrete analytical ranges), which could be attributed also to inter-patient individual genomic diversity, cell type composition and other technical factors. Additionally, this complexity is further enhanced by the inherent correlation structures and hidden confounders (i.e., systematic errors) introduced by each different omic layer [57–59]. A representative example is the integration of proteomics with other types of omics data, such as transcriptomics, as the former are usually investigating a limited percentage of the expressed genome, are more challenging in the experimental preparation, with additional effects (post translational modifications, localization and/or degradation) further perplexing the modeling of inter-data relationships [60]. Moreover, one other limitation lies in the absence of a "standardized" protocol for sharing and storing the available multi-omics data in the various cancer data repositories, resulting in the "under-utilization" of the available molecular information being present. In particular, different web platforms, host multi-modal cancer data in distinct processing and transformation formats (different normalization pipelines, reference genome versions). This absence of "harmonized" data containers pertains a major obstacle to researchers trying to utilize or compare different studies, or even to reproduce original findings from published initiatives. Thus, this augments the necessity for reproducible, common and standard data representations pertaining multi-omics cancer studies. Finally, another bottleneck is the presence of large amounts (and in parallel different patterns) of missing values,

mainly in the clinical data, but also amongst the same patients being profiled with different high-throughput experiments. This results in sparse datasets, frequently including non-matched tumor-normal samples, missing percentages of profiled omics layers and inaccessible clinical annotations for the studied patient cohorts [39].

1.6. Research Outlook: Single-Cell Multimodal Analysis

Single-cell multimodal omics represents the recent technological advancement from single-cell RNA-sequencing (scRNA-seq) to the acquisition of multiple molecular data types such as genome, transcriptome, methylome or proteome from single cells.

This includes the combination of multiple next-generation sequencing-based methods such as DR-Seq (gDNA-mRNA sequencing) [61] and G&T-Seq (genome and transcriptome sequencing) [62], ATAC-RNA-Seq (combined assay for transposase-accessible chromatin using sequencing and RNA sequencing) [63] or the capture of three-dimensional genome structures with DNA methylome profiling (scMethyl-HiC [64] and snm3C-seq [65]). Additionally, droplet-based methods such as Perturb-Seq [66,67], and CRISP-Seq [68] have been developed, which combine CRISPR-based transcriptional interference with high-throughput single-cell RNA sequencing. For a full review on experimental methodologies see Zhu et al. 2020 [69], Ma et al. 2020 [70] or Lee et al. 2020 [71].

This powerful technology enables the investigation of complex biological states and processes of multicellular organisms. In cancer research it can be used to explore tumor heterogeneity, tumor evolution or the identity of infiltrating immune cells [72–74]. A triple omics single-cell sequencing approach in hepatocellular carcinoma for example identified two subpopulations of carcinoma cells, which significantly differed in DNA copy number, DNA methylome, and transcriptome [75]. A study in cutaneous squamous cell carcinoma combined scRNA-Seq with spatial transcriptomics and multiplexed ion beam imaging and uncovered multiple features of potential immunosuppression in the compartmentalized tumor stroma [76].

The aim of analyzing multimodal single-cell data is the unification of different data modalities to uncover complex biological mechanisms on the cellular level such as the reconstruction of gene-regulatory and signaling networks [77]. Particular challenges of this approach lie for one in the still low throughput and high cost of multimodal single-sequencing assays often leading to data sparsity. Additionally, technical noise is often high due to low sequencing coverage and missing values [69,78].

Often not all modalities of a data set stem from exactly the same cell but cells from the same sample or tissue, leading to batch effects from unmatched data. To remedy this projection into a common latent space (Feature Projection) can be applied. Canonical correlation analysis (CCV) and Manifold alignment are both feature projection-based dimensionality reduction techniques. CCV, which is a multivariate analysis technique for estimating a linear relationship between two sets of measurements, can be performed using Seurat3 [79]. VDJView [80] is a specialized tool for the multimodal analysis of data from T and B cells, and includes Seurat [81], Scater [82] and SC3 [83] as well as several additional analysis and visualization features. Manifold alignment algorithms such as MATCHER [84] or MMD-MA [85] use a type of machine learning algorithm that produces projections between sets of data, given they lie on a common manifold.

Bayesian Modeling is a stochastic variational inference method based on Bayesian modeling [86]. Clonealign [87] integrates expression and copy number data from human cancers under the paradigm that copy number is positively correlated with gene expression. BREM-SC [88] is a random effects mixture model for the joint clustering of paired single cell transcriptomic and proteomic data.

Regression Models include least absolute shrinkage and selection operator (LASSO) regression with sci-CAR [89], gradient boosting regression (GBR) modeling [90], Hidden Markov random field (HMRF) modeling with trendsceek [91] and multivariate normal modeling (MNM) with SpatialDE [92].

In addition, single-cell multimodal autoencoders for mapping to a shared latent space are emerging [93,94].

For the unsupervised integration of single-cell multimodal data a widely used method is Matrix Factorization. Here, the data matrices are decomposed into two lower dimensionality matrices. Methods include integrative non-negative matrix factorization (iNMF) by algorithms such as Wishbone [95] or LIGER [96], coupled nonnegative matrix factorization (coupleNMF) [97], group factor analysis (GFA) with algorithms such as Multi-Omics Factor Analysis (MOFA+) [98], and independent component analysis (ICA) [67].

Additionally, MIMOSCA [66] and MUSIC [99] are algorithms for the analysis of expression data after CRISPR perturbation (Perturb-Seq). While MIMOSCA is based on a regularized linear model to estimate the impact of perturbations on gene expression, MUSIC utilizes topic modeling, a decomposition method to discover the shared latent information among input matrices as used for the discovery of hidden semantic features in natural language processing.

Finally, an interesting new implementation is included in the new release of the DESeq2 R package: the function integrateWithSingleCell integrates bulk differential gene expression analysis results from DESeq2 with public single-cell datasets. This facilitates the investigation of which types of cells might be responsible for the relative expression differences in the bulk samples [100].

2. Results

2.1. Literature Mining

The general literature search in PubMed was performed in November 2020 and resulted in 753 publications about multi-omics data integration. We started our search with basic search terms to discover the entire complexity of multi-omics publications and continued with more specific filtering for supervised methods, unsupervised methods, reviews, tools, and cancer related papers. A detailed description of the literature mining parameters is available in the Supplementary Materials. Supplementary Figure S2 shows the distribution of classifications of these publications. We classified 91.5% (688) of all papers, some of which also have multiple classifications as reviews may also deal with supervised or unsupervised tools in cancer research. Most papers (75% (566)) were classified in the tool category, where they either apply tools or report new tools. Only 3% (24) were classified as dealing with supervised multi-omics data integration, whereas about 18% (135) were classified as dealing with unsupervised multi-omics data integration. Figure 1 shows the overlapping distribution of the three categories Cancer, Review, and Tool for supervised, unsupervised and other not clearly classified papers. The class called "Other" in Figure 1 contains papers which are not clearly classified as supervised or unsupervised multi-omics papers but they can deal with both supervised and unsupervised, or semi-supervised data integration. Only 3% are related to supervised multi-omics data integration which can be fully classified into the overlapping subcategories Cancer, Tool, and Review whereas the subcategories for unsupervised and other papers cover 91% and 89% of search results. We observed that publications classified as tools, which is the largest subclass, have overlap with the review subclass and predominately to the cancer subclass. The dominating trend of unsupervised tools in comparison to supervised tools can also be observed by looking at published summaries of multi-omics data integration tool classifications [101,102]. Anyhow, 8.5% (64) publications did not fit into any of these categories. Additionally, Supplementary Figure S3 shows the distribution of classified papers according to the publication year. There we see a drastic increase of multi-omics related papers starting from 2012 to 2020 where publications dealing with tools are cover the majority of publications followed by Cancer related papers and Reviews. This displays the development in the field and the importance of multi-omics research in general and in translational cancer research. A full list of mined publications can be found in Supplementary Materials S1.

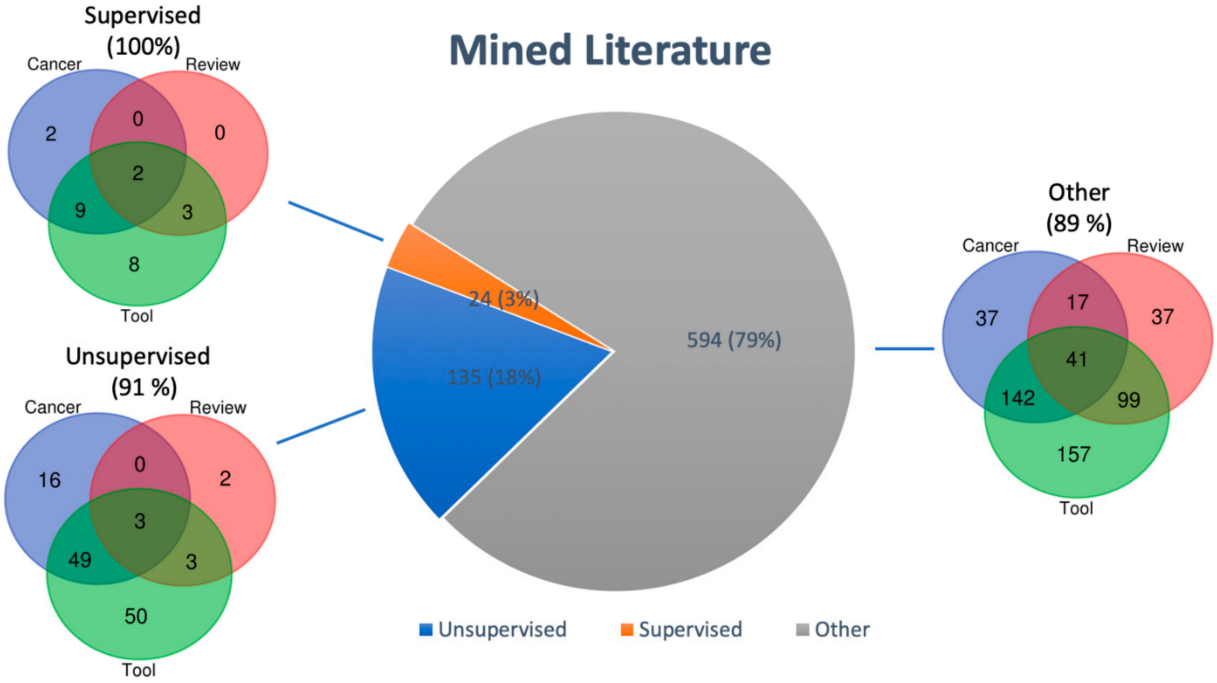

Figure 1. Summary of literature mining. The general search can be classified into three main classes Supervised, Unsupervised and Other. The majority of papers (100% in Supervised, 91% in Unsupervised, 89% in Other) within these classes are included in the overlapping subclasses Cancer, Review, and Tool (Venn diagrams have been created with http://bioinformatics.psb.ugent.be/webtools/Venn/).

2.2. Methodologies and Motivation

In this section, we will give a summary of user-friendly tools suited for defined general research purposes, using tools identified in our literature research as well as tools classified by Nicora et al. (2020) [103] and Huang et al. (2017) [101]. Criteria for selected tools were that they should integrate more than one omics layer, should show cancer-related use cases or demonstration on cancer data, and they should have a clear documentation for user-friendliness, such as a vignette or repository with sufficient information.

We used common general research purposes as defined by Nicora, et al. (2020) [103], but added "cancer subtype classification" to the list for cancer-specific analyses. The following Tables 2–7 reflect this classification and the wide range of developed tools available for multi-omics data integration. The categorization of these tools is based on the most common use case of each tool, which does not necessarily mean that the tool is limited to that research purpose but has been mainly applied for this aim. We also classified the tools based on criteria for supervised and unsupervised learning. For supervised learning methods, the tool has been trained on labeled training data in order to optimize a given hyperparameter for the defined hypothesis and to minimize a specific loss function. Unsupervised tools do not utilize labeled data. They can learn from non-labeled data with unknown non-categorized patterns. However, they are commonly based on statistical methods rather than on machine learning techniques. Additionally, the tools in Table 7 are used for multiple research purposes and can be applied to several research aims.

2.2.1. Patient Stratification

Multi-omics data integration tools for patient stratification (see Table 2) are finding groups of samples of therapeutic and clinical relevance. These groups can be defined by a multi-omics profile for specific treatment response or survival benefit.

Table 2. Summary of computational tools for patient stratification in the field of multi-omics data integration.

Model Nature Orientation	Tool Name	Programming Language	Integration Method	Used Omics	Reference
Unsupervised	R.JIVE	R	Multi-step analysis	mRNA, miRNA, MET	[104]
Unsupervised	PROFILE	R, Python	Multi-step analysis	mRNA, CNV, MUT	[105]
Supervised	SALMON	Python	Gene co-expression Analysis	mRNA, miRNA, CNV	[106]
Supervised	netDx	R	Feature network aggregation	mRNA, miRNA, CNV, MUT, MET, PROT	[107]
Supervised	Graper	R	Bayesian	mRNA, DRUGre, MET	[108]

mRNA = Gene Expression, miRNA = microRNA Expression, MET = Methylation, CNV = Copy Number Variation, PROT = Proteomics, DRUGre = Drug response, MUT = mutation.

2.2.2. Biomarker Discovery

Tools for the discovery of biomarkers (see Table 3) are aiming to find specific composite molecular signatures for clinical (prognostic and/or diagnostic) utility such as disease state, treatment response, or survival rate. Multi-Omics integrated biomarker sets can encapsulate changes and effects in different omics layers as key points for personalized medicine (e.g., mutations leading to changes in expression, protein folding, genetic regulation, or methylation). Representative examples include the derivation of composite signature sets in the field of radiogenomics and colorectal cancer [109] and in the pan-cancer classification of distinct solid tumors [110].

Table 3. Summary of computational tools for biomarker discovery in the field of multi-omics data integration.

Model Nature Orientation	Tool Name	Programming Language	Integration Method	Used Omics	Reference
Unsupervised	Joint Bayes Factor	Matlab	Matrix factorization	mRNA, MET, CNV	[111]
Unsupervised	iProFun	R	Multiple-step analysis	mRNA, CNV, MET	[112]
Unsupervised	CCA-sparse group	Matlab	Canonical correlation analysis	mRNA, SNP	[113]
Supervised	sMBPLS	Matlab	Partial Least Squares	mRNA, miRNA, CNV, MET	[114]
Unsupervised	CNAmet	R	Multi-step analysis	mRNA, CNV, MET	[115]
Supervised	iBAG	R	Multi-step analysis	mRNA, CNV, MET	[116]
Supervised	Anduril	R, Python, Shell	Multi-step analysis	aCGH, mRNA, miRNA, SNP, MET	[117]
Supervised	CapsNetMMD	Python	Capsule network model	mRNA, CNV, MET	[118]

mRNA = Gene Expression, miRNA = microRNA Expression, MET = Methylation, CNV = Copy Number Variation, aCGH: DNA microarray.

2.2.3. Pathway Analysis

Tools for pathway analysis (see Table 4) are dealing with regulatory effects (e.g., gene regulatory networks, post-translational modifications), interactions between different pathways on multiple omics layers (e.g., gene/protein interaction networks), or use databases like KEGG [119] or REACTOME [120] for pathway discovery. Except for these, some additional and widely used resources of prior knowledge include the Molecular Signatures (MSigDB) [121] and the Pathway Commons [122] databases. MSigDB is a comprehensive resource of annotated gene-sets separated into nine major collections, whereas Pathway Commons comprises of an integrated repository spanning about 4794 biochemical processes and 2.3 million interactions. In addition, the SignaLink 2 resource [123] is a signaling pathway database with multi-layered regulatory networks for the interpretation of multi-omics studies. Finally, the Omnipath database [124] is one of the richest sources regarding protein-protein interactions, including more than 100 knowledge resources for 20,000 human proteins and 16,500 complexes.

Table 4. Summary of computational tools for pathway analysis in the field of multi-omics data integration.

Model Nature Orientation	Tool Name	Programming Language	Integration Method	Used Omics	Reference
Unsupervised	ModMap	Java	Multi-step analysis	mRNA, miRNA, PROT	[125]
Unsupervised	NetICS	Matlab	Multi-step analysis	mRNA, miRNA, CNV, MUT, MET, PROT	[126]
Unsupervised	SNMNMF	Matlab	Matrix factorization	mRNA, miRNA	[127]
Unsupervised	PARADIGM	Web-app, Python	Probabilistic graphical models	mRNA, CNV	[128]
Supervised	FSMKL	Matlab	Multiple kernel learning	mRNA, CNV	[129]
Unsupervised	Sumer	R	Multi-step analysis	mRNA, PROT	[130]
Unsupervised	MOSClip	R	PCA	mRNA, CNV, MUT, MET	[131]
Supervised and Unsupervised	COCOA	R	Multi-step analysis	mRNA, ATAC-Seq, DRUGre, MUT, MET	[132]

mRNA = Gene Expression, miRNA = microRNA Expression, MET = Methylation, CNV = Copy Number Variation, PROT = Proteomics, DRUGre = Drug response, MUT = mutation, ATAC-Seq = Transposase-Accessible Chromatin.

2.2.4. Drug Analysis

The following tools in Table 5 aim at the discovery of new drugs or new drug effects and the use of existing ones in combination with others for improved drug response and survival based on data from different omics layers. These tools try to identify potential drug targets in search for better treatment with higher survival rates, and are applicable in the field of pharmacogenomics and drug repurposing, where multi-omics analysis can identify putative target regulators, which affect dynamic molecular networks (e.g., drugs targeting identified pathways resulting from analysis of differences in gene and protein expression) [133].

Table 5. Summary of computational tools for drug analysis (drug repurposing and drug discovery) in the field of multi-omics data integration.

Model Nature Orientation	Tool Name	Programming Language	Integration Method	Used Omics	Reference
Supervised	MOLI	Python	Neural networks	mRNA, CNV, MUT	[134]
Unsupervised	SNPLS	Matlab	Partial least squares	mRNA, DRUGre	[135]

mRNA = Gene Expression, CNV = Copy Number Variation, DRUGre = Drug response, MUT = mutation.

2.2.5. Cancer Subtype Classification

The tools in this category are used for the classification of molecular subtypes of specific cancer types (see Table 6).

The identification of novel disease subtypes can be improved by the application of these integrative methodologies. They can lead to the identification of improved targets for anti-cancer treatment or they could contribute additional knowledge to existing cancer subtypes from a multi-view perspective.

Table 6. Summary of computational tools for Cancer Subtype classification in the field of multi-omics data integration.

Model Nature Orientation	Tool Name	Programming Language	Integration Method	Used Omics	Reference
Unsupervised	mixKernel	R	Multiple Kernel learning	mRNA, miRNA, MET	[136]
Unsupervised	iClusterBayes	R	Bayesian clustering	mRNA, CNV, MUT, MET	[137]
Unsupervised	SNF	R, Matlab	Network fusion	mRNA, miRNA, MET	[138]
Unsupervised	iCluster	R	Matrix factorization	mRNA, CNV	[139]
Unsupervised	iCluster Plus	Matlab	Matrix factorization	mRNA, CNV, MUT	[140]
Unsupervised	JIVE	Matlab	Matrixfactorization	mRNA, miRNA	[141]
Unsupervised	PSDF	Matlab	Bayesian	mRNAs, CNV	[142]
Unsupervised	BCC	R	Bayesian	mRNA, miRNA, MET, PROT	[143]
Unsupervised	SCFA	R	Multi-step analysis	mRNA, miRNA, MET	[144]
Supervised	MAUI	Python	Autoencoder	mRNA, CNV, MUT	[145]

mRNA = Gene Expression, miRNA = microRNA Expression, MET = Methylation, CNV = Copy Number Variation, PROT = Proteomics, MUT = mutation.

2.2.6. Multi-Omics Data Discovery

The previous research aim categories are sometimes very closely related (e.g., drug discovery implies sometimes biomarker discovery for detecting effective druggable marker). Therefore, several tools can be employed in multiple of the selected research aims (see Table 7). Most of them are unsupervised and can be employed for carrying out an initial exploratory analysis on multi-omics profiles of different cancer types.

Table 7. Summary of computational methods which can be applied to several mentioned research aims in a multi-omics context.

Research Purpose	Model Nature Orientation	Tool Name	Programming Language	Integration Method	Used Omics	Reference
Biomarker discovery, Cancer subtype classification, Pathway analysis	Unsupervised	MCIA	R	Multi-step analysis	mRNA, PROT	[146]
Patient stratification, Cancer subtype analysis	Supervised	mixOmics	R	Feature transformation	mRNA, miRNA, PROT	[147]
Biomarker prediction, Pathway analysis	Unsupervised	Lemon-Tree	Java	Module network learning	mRNA, CNV	[148]
Patient stratification, Cancer subtype classification	Unsupervised	Clusternomics	R	Multi-step analysis	mRNA, miRNA, MET, PROT	[149]
Biomarker discovery, Pathway analysis	Unsupervised	AMARETTO	R	Multi-step analysis	mRNA, CNV, MET	[150]
Pathway analysis, Cancer subtype classification	Supervised	iOmicsPASS	C++	Multi-step analysis	mRNAs, CNV, PROT	[151]
Biomarker discovery, Cancer subtype classification	Unsupervised	MOGSA	R	Matrix factorization	mRNA, CNV, Phosp, PROT	[152]
Patient stratification, Pathway analysis	Unsupervised	PathME	R, Python	Matrix factorization	mRNA, miRNA, CNV, MET	[153]
Drug analysis, Pathway analysis	Supervised	DrugCombo Explorer	Java, Python	Multi-step analysis	DNA, mRNA, CNV, MET	[154]
Biomarker discovery and Patient stratification	Unsupervised	MOFA	R	Matrix Factorization	mRNA, MUT, MET, DRUGre	[155]

mRNA = Gene Expression, miRNA = microRNA Expression, MET = Methylation, CNV = Copy Number Variation, PROT = Proteomics, DRUGre = Drug response, MUT = mutation, Phosp = Phosphorylation profiles.

Generally, we propose three main criteria that could guide the appropriate selection of all the above catalogued methodologies:

A. Research aim: Firstly, the most important aspect is the research question: what is the purpose of the specific study, or which biological insights are aimed for regarding a specific cancer type or cohort? This can significantly inform the selection of the most suitable tools that match the specific research goal, as no single tool or pipeline covered above can address a complex disease like cancer in its entirety. In addition, except for the initial selection (i.e., unsupervised or supervised methodologies), benchmark studies can be further utilized and considered as guidelines for narrowing the candidate tools [156].
B. Another crucial criterion is the experimental design and the interrogated datasets: which is the relative sample size? For example, usually ML-based approaches require a higher number of samples for model training and validation in comparison to unsupervised methodologies. Also, can the relative tool cope with the percentage of missing values and/or the nature of omics layers (continuous vs. sparse genetic data)? When a large percentage of missing values is present in both omic layers and clinical data, researchers should consider various published studies covering different methodologies for missing value imputation [157].
C. Furthermore, the third selection criterion that is often underestimated is the presence of extensive documentation that accompanies an available tool. Despite the fact that an approach might be well suited for a specific research scenario, the absence of a rich vignette and detailed reproducible examples poses a significant constraint on the utilization of the respective methodology [158].

2.3. Rationale for Selection of Tools and Datasets

Overall, as already pinpointed in the literature mining process above, there is a large amount of computational tools and pipelines that can be utilized for different research goals. However, few studies or reviews provide also comprehensive examples or tutorials on how to utilize public cancer genomics repositories, and perform multi-omics data integration. On this premise, we selected two tools that can be utilized in two distinct scientific scenarios: the MOFA R package for unsupervised methodologies, and the netDx R package for the supervised ones. Initially, the main rationale of selecting R language tools is that the Bioconductor project is the largest consortium for the statistical analysis and comprehension of genomics data (https://bioconductor.org). It is comprised of a core team of more than 1200 researchers to support continuous development. It is widely used with around 3/4 million distinct IP downloads annually, and well respected (42,000 PubMed Central full text citations). Moreover, it provides detailed documentation and extensive vignettes based on high quality standards and a broad scientific community that can provide support (https://support.bioconductor.org). Additionally, each package is thoroughly tested in different computational systems for scalable and performant analysis. Furthermore, the majority of the above selected tools are based on the R language.

Concerning the non-supervised approach, MOFA/MOFA+ [98,159] was chosen as the respective methodology, as it is by design unsupervised, so it is not aimed at detecting differential changes between a predefined set of samples. It provides a well-established workflow to characterize these sources of variation, especially when analyzing datasets with complex group structure. Also, MOFA+ has been extensively used and cited in more than 80 research studies and comparative reviews [160–163]. Furthermore, the MOFA+ stable Bioconductor installation is utilizing basilisk to automatically set up the necessary Python-R connection, which facilitates interoperability.

On the other hand, netDx [107,164] is a recently published Bioconductor R package, which provides a novel methodology of implementing patient similarity networks for efficient patient classification, which has been shown to outperform other machine learning approaches. It can integrate heterogeneous patient data from clinical to omics layers, while implementing machine learning algorithms for robust feature selection. Further-

more, it uses Cytoscape (RCy3) for the efficient visualization and interpretability of the inferred biological networks. Finally, as the two aforementioned methodologies can't be directly compared, we selected two different multi-omics cancer datasets, for the different computational approaches. In particular, the TCGA-LUAD dataset was selected for the unsupervised approach, based on the absence of known molecular subtypes between the patients. For the hypothesis driven netDx approach, we selected the CLL dataset, as the IGHV mutational status is a known clinical marker that separates patients into distinct classes (see Materials and Methods Sections 4.2 and 4.3).

2.4. Unsupervised Multimodal Data Integration Case Study with MOFA

In order to disentangle the heterogeneity and unravel new biological insights regarding lung adenocarcinoma, MOFA+ analysis was applied in the processed TCGA-LUAD dataset, as described in the Materials and Methods section.

An initial overview of the trained MOFA model is illustrated in Figure 2. In detail, in Figure 2A we observe the correlation between the inferred latent factors from the model, which verifies that all factors are mostly uncorrelated, suggesting a good model fit. Figure 2B shows the percentage of variance explained by each factor across each omics layer. Interestingly, Factor 1 seems to capture a source of variation that is presented across two modalities, being gene expression (RNASeq) and protein abundance (RPPAArray). In contrast, Factor 2 seems to capture a strong source of variation that can be attributed mainly to the gene expression data, whereas Factors 3 and 4 are mainly related to the CNV data. Collectively, in this dataset using in total 15 Factors, the model explained up to ~43% of the variation in the gene expression data, around 38% in the copy number alteration data, and ~18% in the RPPA data. Overall, the above findings suggest that no single omics technology can explain holistically all the sources of variation in the dataset, further augmenting the necessity of profiling a complex disease with different molecular layers.

Next, aiming to explore the molecular landscape of lung adenocarcinoma, we initially performed a correlation analysis to associate the MOFA factor values with any included clinical sample metadata. The analysis highlighted that expression subtype, ATM mutation and gender had a significant correlation (log10 adjusted p-value <0.05) with specific factors. In detail, visualization of the samples in the latent space showed that the expression subtype had an association with Factor 1, clearly separating the terminal respiratory unit (TRU) subtype from other two (Figure 3A), whereas ATM mutation showed an interrelation with Factor 2 (Figure 3B). Notably, ATM gene somatic mutations have been illustrated to play a role in the pathophysiology of lung cancer [165,166].

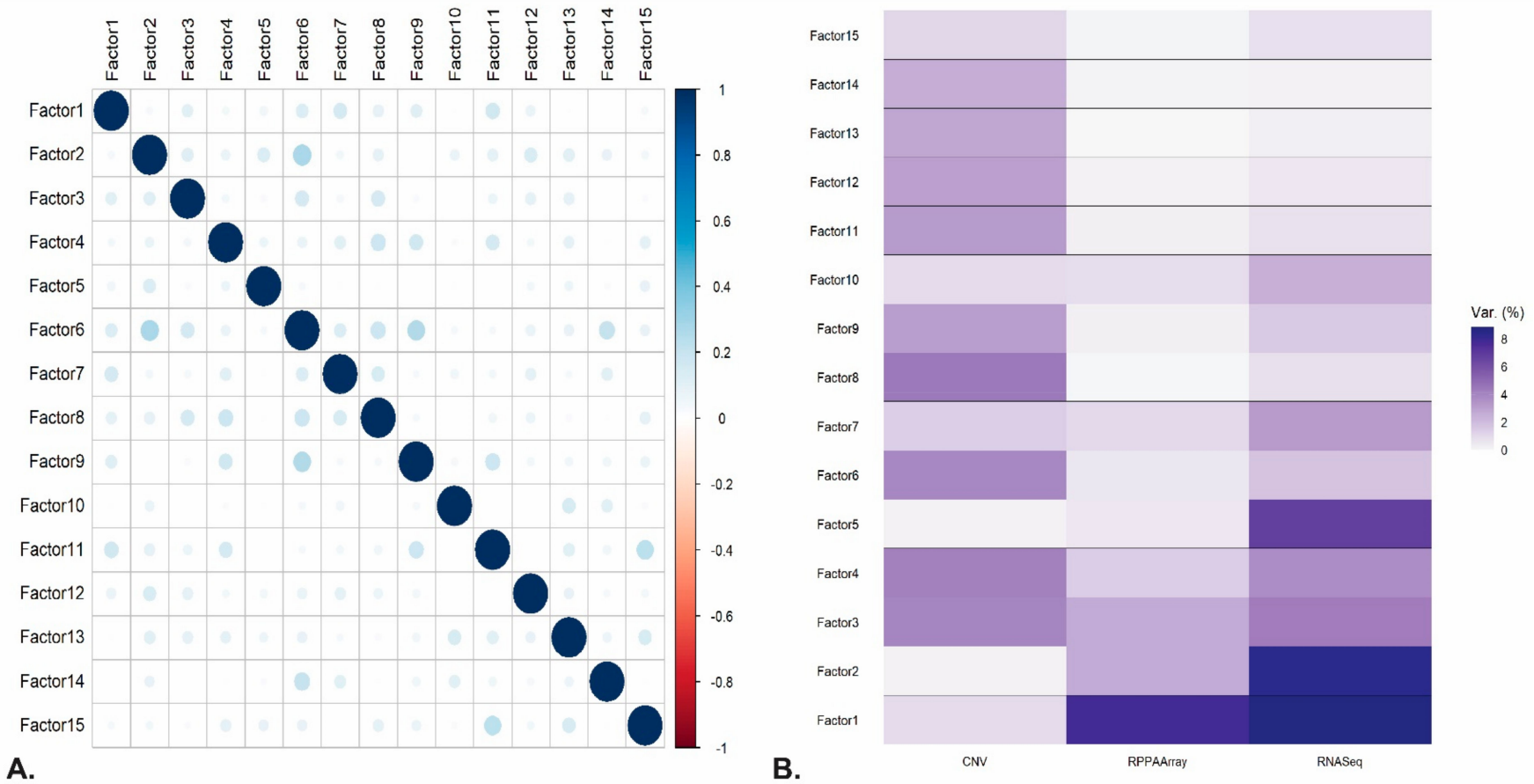

Figure 2. (**A**). Plot of the correlation matrix between the inferred Multi-Omics Factor Analysis (MOFA) latent factors, which can be used as a quality control of the fitted model. It returns a symmetric matrix with the correlation coefficient between every pair of factors. Blue color denotes positive correlation, whereas red negative, respectively. A diagonal correlation matrix is usually expected for a robust model fit, suggesting low correlation overall between the MOFA factors. (**B**). Variance decomposition analysis plot, which illustrates the variance explained (R-squared value) per factor and per layer (CNV, RPPAArray = protein, RNASeq = expression). The values are calculated using a coefficient of determination, which ranges from 0 to 1, and scaled to a percentage by multiplying by 100.

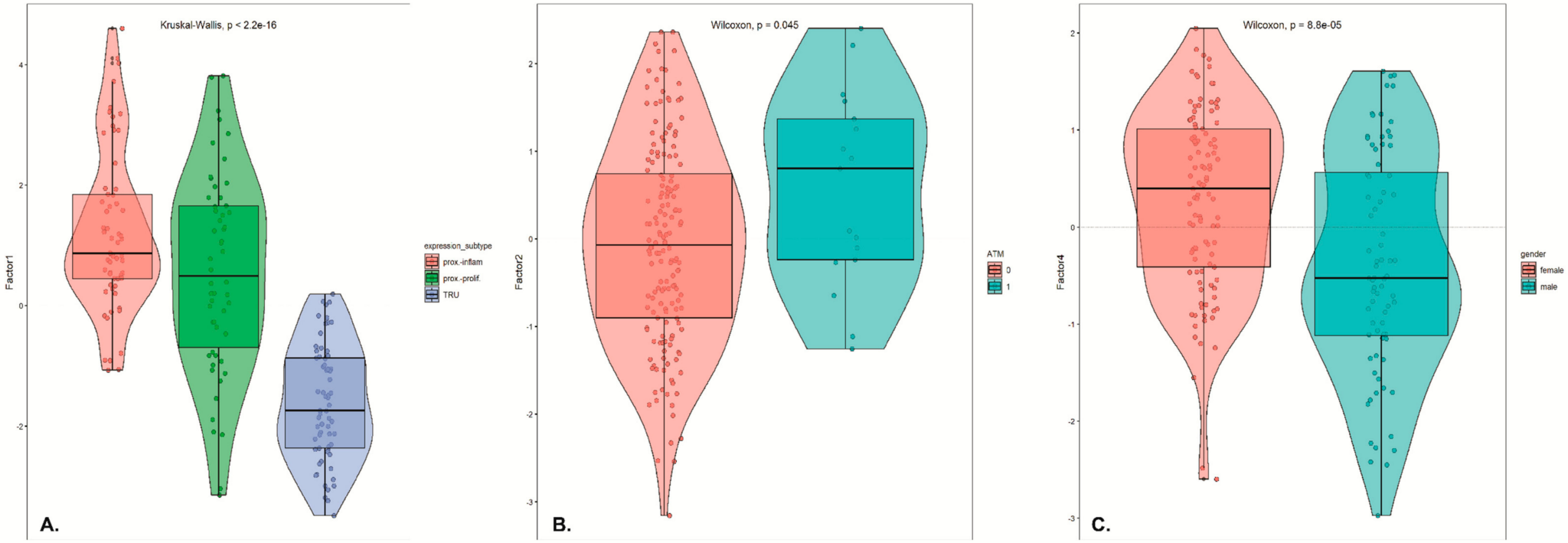

Figure 3. (**A**). Visualization of samples using Factor 1 values, which are colored by the covariate expression_subtype, which denotes the 3 available lung adenocarcinoma transcriptional subtypes. The plot shows a clear separation of the "TRU" subtype from the "proximal-proliferative" and "proximal-inflammatory". (**B**) Visualization of the samples using Factor 2 and ATM mutation status for color. Samples with positive values have the ATM mutation *blue), whereas samples with negative Factor 2 values do not have the mutation (red). (**C**) Visualization of the LUAD samples using Factor 4 and the gender covariate to color the selected factor values. From the relative plot the significant association of Factor 4 with gender is illustrated. Samples with average positive values are mostly female and samples with negative values are mostly male. In all plots, *p*-values for comparison of the means of the groups were calculated using the function stat_compare_means from ggpubr R package.

In parallel, for further exploring the biology of LUAD, we performed functional enrichment analysis to look for biological processes and pathways related to the individual MOFA factors (see Materials and Methods section). Collectively, GSEA analysis shows an overrepresentation of MAPK and AKT on the protein level in factor 1 (see Figure 4A). The MAP kinase pathway is a highly complex signaling cascade involving three kinases. The RAS-RAF-MEK-ERK pathway is altered in forty percent of all human cancers, mainly due to mutations in BRAF and its upstream activator RAS [167]. In lung cancer, KRAS mutations often play a role in activating the MAP kinase pathway [168]. Interestingly, MAP kinase activation is also underrepresented in Factor 3 on the gene expression level (see Supplementary Figure S4). AKT as part of the AKT/mTOR signaling pathway may be a downstream of the PD-L1 pathway [169]. In contrast, latent factor 2 is highly negatively enriched for gene expression pathways related to immunity (see Figure 4B). Amongst the enriched Reactome pathways are innate and adaptive immunity, interferon signaling, and notably PD-1 signaling. The PD-1/PD-L1 pathway controls the induction and maintenance of immune tolerance within the tumor microenvironment. In personalized medicine, the PD-L1 status is used as a predictor for benefit from targeted therapies or immune checkpoint blockers [170,171]. Consistently, when plotting the Reactome pathways enriched in Factor 3 CNV negative weights, it captures differences associated with the immune system, such as innate immune response, immune surveillance and inflammation (Figure 4C). For example, NFKB signaling has been demonstrated to be implicated in lung cancer manifestation, by promoting anti-tumor T cell responses [172]. Another interesting finding is that Factor 3 is also enriched in Notch related signaling pathways. It is worth noting that Notch signaling has been shown to play a pivotal role in lung cancer progression-especially in NSCLC-with genetic alterations associated with survival estimates and therapeutic significance [173,174].

As the resulting MOFA factors can be utilized to predict discrete clusters of samples, we used all the inferred factors to cluster the patients in the latent factor space, implementing collectively all information from the multi-omics layers and their differential contributions. Here, as described in the Materials and Methods section, k-means clustering resulted in three discrete groups of patients. Visualization of the three resulting clusters from the integrated analysis showed a significant overlap (using Pearson chi-squared test) with various clinicopathological parameters, such as AKAP9 gene mutational status, expression subtypes and gender (Supplementary Figure S5). Of note, cluster 3 did not contain any TRU expression subtype samples, whereas the vast majority of samples harboring AKAP9 mutations were allocated in cluster 2. Finally, cluster 2 was more enriched in female patients and cluster 1 had the smallest number of mutations in the COL3A1 gene.

Finally, in order to investigate if any of the inferred latent Factors could be associated with building clinical models of predicting patient outcome, we implemented Cox proportional hazards models (coxph function from R package survival). From the identified 15 MOFA factors, Factor 1 (p-value = 0.0358), Factor 4 (p-value = 0.04162) and Factor 9 (0.04750) were statistically significantly associated with overall survival as the response variable, using p-values derived from Wald statistic (Figure 5).

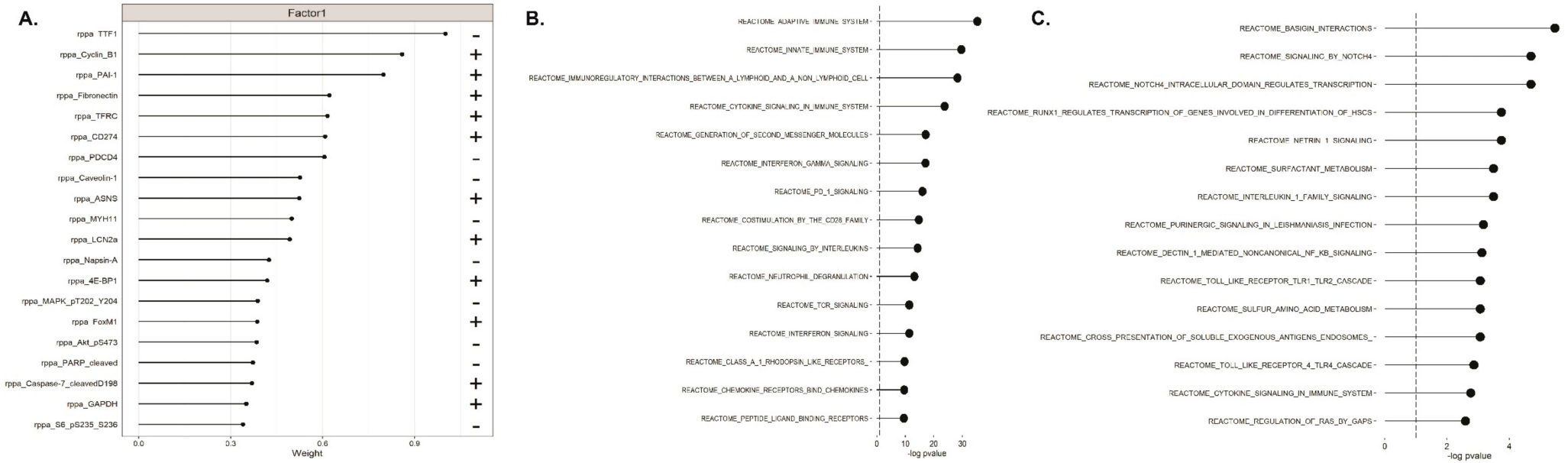

Figure 4. (**A**) Line plot displaying the absolute loading from the top 20 features of Factor 1 in the protein data. The corresponding weight sign is depicted on the right, scaled from −1 to 1. Proteins with positive weights have higher levels of expression in the samples that have Factor 1 positive values, and vice-versa. (**B**) Visualization of the enrichment analysis results, running GSEA on Multi-Omics Factor Analysis (MOFA) factor 2 with gene expression negative weights and Reactome gene sets. (**C**) Visualization of the enrichment analysis results, running GSEA on MOFA factor 3 with copy number variation negative weights and Reactome gene sets.

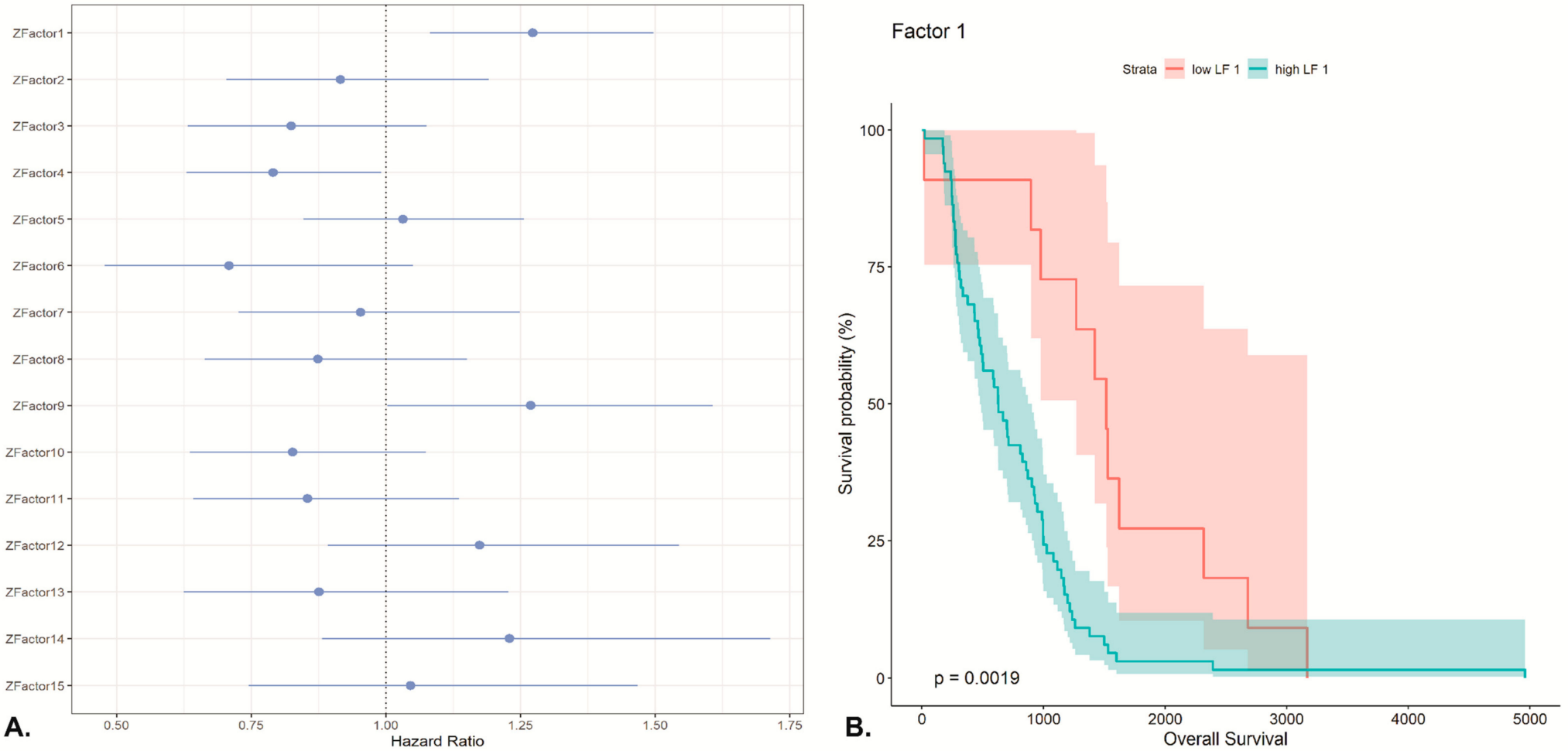

Figure 5. Putative prognostic utility of the Multi-Omics Factor Analysis (MOFA) latent factors. (**A**). Forest plot of resulting hazard ratios, illustrating the association of the tested MOFA factors with overall survival, based on Cox regression modeling (error bars representing 95% confidence intervals) (**B**) Example of a Kaplan-Meier plot for Factor 1, showing overall survival. The samples were separated into two distinct groups, based on the maximally selected rank statistics from the maxstat R package [175]. As Factor 1 has a positive coefficient, samples with high values have an increased hazard in comparison to samples with low relative values. The *p*-value was calculated using a log-rank test on the two aforementioned groups.

2.5. Supervised Multimodal Classification Case Study with netDx

The performed supervised multi-omics data integration on CLL data using netDx resulted in a performance accuracy of 93% (±1.5%). The model was able to clearly discriminate the samples into the binary classes defined by the IGHV mutation status (AUROC = 97.1 ± 2.3%, AUPR = 92.0 ± 2.7%) (see Supplementary Figure S6). Running netDx with 1 CPU took about 72 minutes for this dataset with defined settings (see Materials and Methods section). The aim of applying netDx was to obtain patient similarity networks (PSN) and group patients based on a multi-omics profile [164]. The PSN networks consist of nodes which represent the patients connected by edges representing the weighted pairwise similarities between patients. The classification of the performed analysis is based on a separation of IGHV mutation status for CLL patients which is known as a relevant prognostic factor [176]. Here we followed one suggested design of netDx developers which groups biological pathway enrichments based on gene expression measurements (Pai, et al., 2020). Selected features require a minimum feature score of 9 in at least 50% of train/test splits. The resulting pathway enrichment networks based on the expressed genes are shown for non-IGHV-mutated patients in Figure 6 and for IGHV-mutated CLL patients in Figure 7.

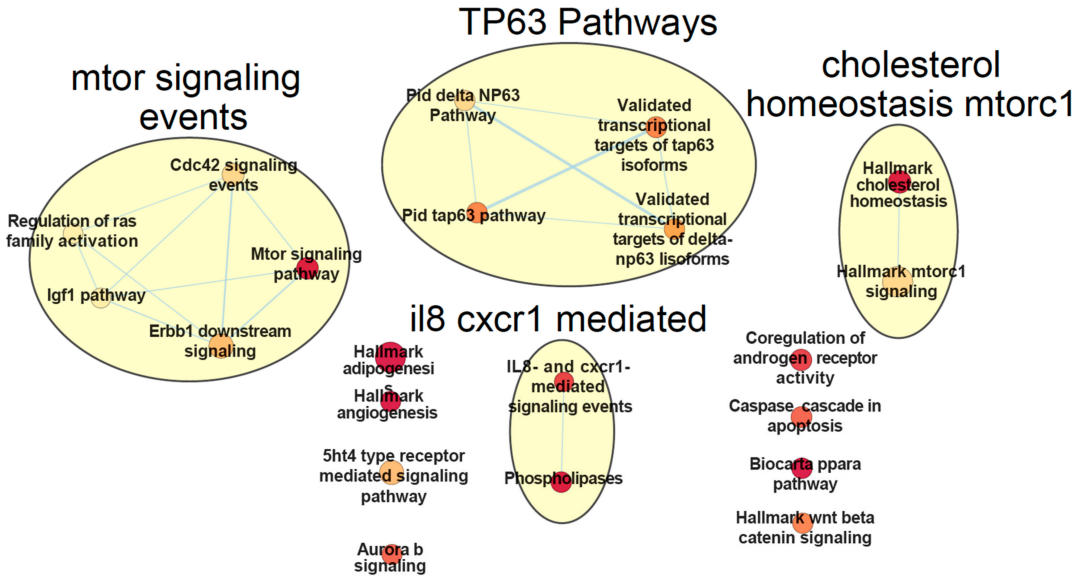

Figure 6. Top predictive features for not mutated IGHV CLL patients. Each node shows predictive pathway features and edges, which connect shared genes of pathways. Node fill uniformly indicates highest score; yellow = netDx score 3, red = netDx score 10. The size of the nodes displays the amount of genes in the underlying gene set. Selected features required a minimum feature score of 9 out of 10 in at least in 50% of train/test splits.

Each network has been manually selected as input for annotation with AutoAnnotate [177]. The titles of annotated networks in Figures 6 and 7 correspond to main themes and categories within each network (e.g., mTOR signaling events) and in some cases had to be manually curated to reflect the included nodes. Enrichments of all signaling pathways from IGHV-mutated samples are included in enrichments from samples without IGHV mutations. Non-mutated samples show 13 additional pathway enrichments, which are not present in the mutated samples. The majority of shared enrichment signals have a higher score for not mutated samples (cholesterol homeostasis, mTOR signaling, phospholipases signals, interleukin 8 (IL8) and chemokine receptor 1 (CXCR1) signals and Aurora b signaling). Only the cell division control protein 42 (CDC42) signal enrichment is higher in mutated samples. The mTOR pathways are enriched in the pathway networks of both classes (IGHV- mutated and not mutated) and connected to CDC42 signals in both classes. Non-mutated samples show more concatenation of the mTOR-pathway and higher scoring to other signals. High scoring of adipogenesis and cholesterol signaling in both classes can also be observed. Ten nodes in non-mutated samples and 6 nodes in mutated samples are

not connected with any edges. Figure 7 shows a relatively sparse representation of pathway networks for IGHV-mutated CLL patients in comparison to Figure 6. Two major networks are visible which mainly include IL8, CXCR1, mTOR-pathway and CDC42 signals. The enrichment called biocarta ppara in Figures 6 and 7 refers to the Mechanism of Gene Regulation by Peroxisome Proliferators via PPAR alpha.

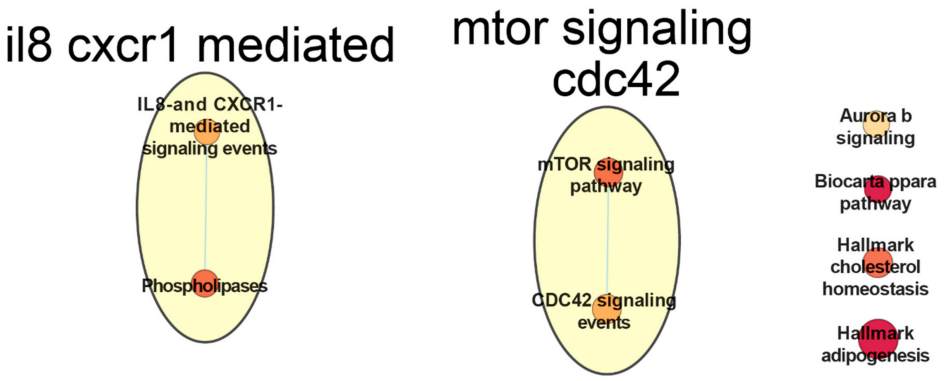

Figure 7. Top predictive features for IGHV-mutated CLL patients. Each node shows a predictive pathway features and edges connect shared members of pathways. Node fill uniformly indicates highest score; yellow = netDx score 3, red = netDx score 10. Selected features required a minimum feature score of 9 out of 10 in at least in 50% of train/test splits.

Shared enrichment between the classes show general cancer associated signals. Enrichment of adipogenesis indicates cancer-induced changes to the regulation of adipose tissue, which promotes cancer cell survival during therapy [178]. Also a breakdown of Cholesterol homeostasis is known to be linked to hypocholesterolemia in lymphocytic leukemia [179]. In addition, chemokine receptors CXCR1/2 and their ligand CXCL8 are essential for the activation and trafficking of inflammatory mediators as well as tumor progression and metastasis [180]. The IL-8 and CXCR1 related pathways are enriched in both classes. IL-8 is known for B-cell progression [181] and Chemokine receptors CXCR1/2 and their ligand CXCL8 are essential for the activation and trafficking of inflammatory mediators as well as tumor progression and metastasis [180]. Interestingly, it has been shown that leukemic B-cells neither express CXCR1 or CXCR2 nor they respond to exogenous IL-8 in CLL patients [181]. The aggregated collection of pathways highlights the mTOR –pathway, which plays a critical role in leukemia initiation [182]. Inhibitors of the mTOR-pathway are currently one line of therapies for leukemia patients. For example, CDC42 signaling as part of the mTOR-pathway and related pathways are known as key targets for CLL treatment with lenalidomide [183]. Differences in pathway related networks should highlight the driving variance of the IGHV mutation status in the gene expression layer. The Tumor Protein P63 (TP63) related pathway in non-mutated samples is completely absent in mutated samples, which demonstrates the prognostic relation between TP63- related pathways and IGHV mutation status in CLL patients [184].

Additionally to the described pathway enrichment analysis, we also performed clustering of patients based on multi-omics profiles. The following analysis is based on more strict selection of features with a minimum feature score of 9 for at least 70% of splits. Figure 8A shows the patient similarity network (PSN), which integrates the predictive features for all patient labels. Another visualization of a PSN is a tSNE plot for a cluster representation of patient classification based on the IGHV mutation status, as shown in Figure 8B. Patient similarity networks show a complex landscape of similarities where the tSNE applied clustering shows more distinct clusters of patients but not a clear separation of the IGHV mutation status. Supplementary Figure S7 shows how well the patients are clustered in the patient similarity network (PSN) by using pairwise patients' shortest distance in the classes (IGHV_0 and IGHV_1) and between the classes. Distances within the classes should be smaller than between the classes in order to separate clustering of patients in the PSN.

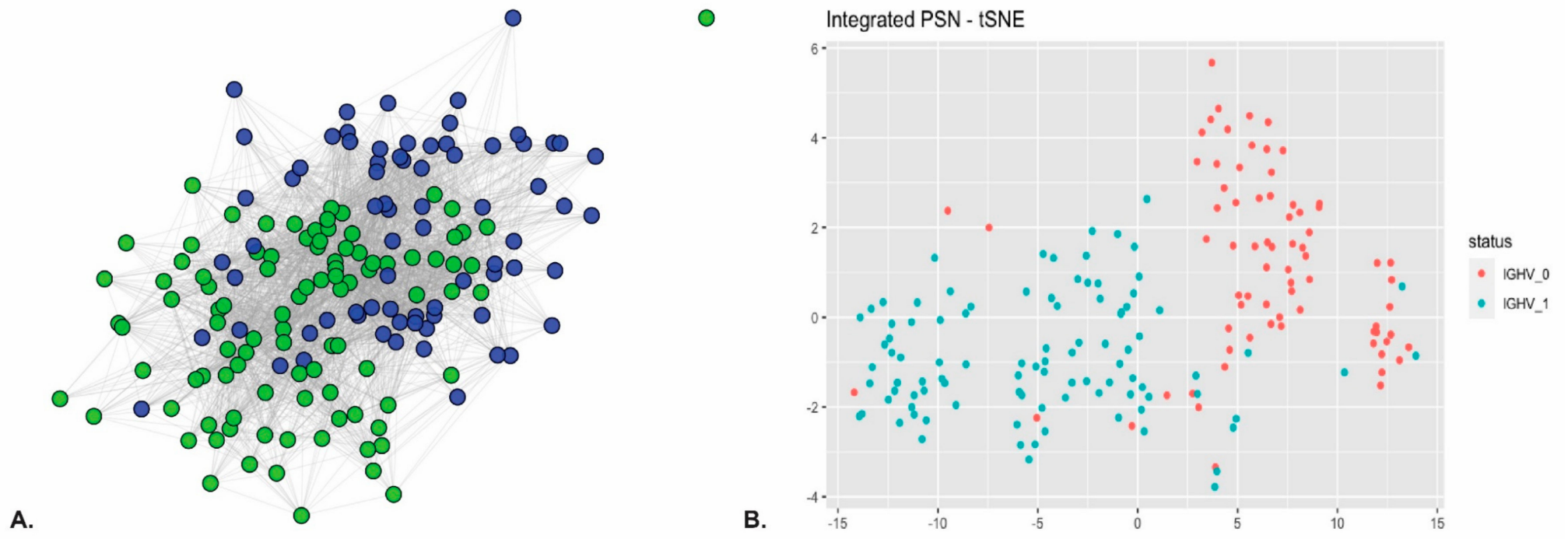

Figure 8. (**A**). Integrated patient similarity network for binary stratification of IGHV status based on multi-omic data (green—IGHV mutation, blue—no IGHV mutation). Each node in this network corresponds to a CLL patient and each edge corresponds to weights displaying the average similarity across all features passing feature selection [164]. This network was generated by implemented functions of netDx and visualized by Cytoscape. (**B**) tSNE visualization of integrated patient similarity network for binary stratification of IGHV status (IGHV_0 = no IGHV mutations, IGHV_1 = IGHV mutations). Only features which passed the feature selection step are integrated.

3. Discussion

In the last decade, new massively parallel sequencing technologies have yielded new biological insights at the RNA, DNA, cellular, and spatial resolution, resulting in the accumulation of massive amounts of genomics data. Recently, there is a growing interest in integrating diverse data from such distinct molecular layers, in order to shed light on the biology of various complex phenomena. The simultaneous examination of multi-layer views can paint in-depth molecular pictures that provide comprehensive insights into the way our "omes" interact in the manifestation of diseases like cancer. Indeed, multimodal data integration approaches are redefining precision oncology through the exploitation of different molecular entities, which characterize holistically the molecular landscape of distinct tumors and facilitate the identification of actionable targets with clinical utility. However, despite the fact that multi-omics integration is an active area of translational cancer research, it lacks established performance benchmarks and assessment standards.

On this premise, in this review we sought to create a detailed catalogue of all the available computational tools, which a researcher could utilize for the integration of heterogeneous cancer genomics data in the context of translational cancer research. Our main goal was not only to provide a rich resource of cutting edge technologies, but also to implement two reproducible case studies, that illustrate the analysis of heterogeneous cancer multi-omics data using state of the art tools, focusing on two typical research scenarios: the supervised approach, for when a researcher for example tries to find the features that characterize the taxonomy between known molecular cancer subtypes; and the unsupervised one, that tries to unravel the heterogeneity and stratify the cancer patients into new disease subgroups. These two case study examples can serve as start-to-end workflows, which a user can utilize to analyze from scratch public multi-layered cancer data. Both studies cover important parts from multi-omics data acquisition, preprocessing of individual omics layers, integration, model training and functional enrichment analysis, along with extensive documentation of the R code and can be directly obtained from github (see here: https://github.com/Jasonmbg/CaseStudy_MAE_TCGA_LUAD_Review and https://github.com/jonasboh/Case_Study_netDx_CLL). Altogether, these two reproducible pipelines can serve as a complement to the literature mining process, on how to address two different research scenarios based on the general categorization of model nature orientation (Supervised vs. unsupervised methodologies).While it lay beyond the scope of our current review to critically analyze and provide methodological insights for each mentioned tool, we addressed the main pros and cons of the unsupervised and supervised methodologies utilized in our case studies.

This was initially demonstrated in the LUAD dataset case study, where MOFA/MOFA+ was capable of recovering known sources of biological variation related to lung adenocarcinoma expression subtypes, gender and specific somatic mutations. The molecular basis of these inferred factors aligns well with previous studies, highlighting crucial signaling pathways and perturbed biological mechanisms related to immune response, cell cycle, MAPK signaling cascades and inflammation [185–187]. In addition, the model identified putative clinical markers. For example, based on the top weights, using the loadings of each feature in the gene expression data, Factor 1 was highly correlated with the pulmonary-associated surfactant protein B (SFTPB) gene. A recent study illustrated that SFTPB gene expression was correlated with tumor-infiltrating lymphocytes (TIL), defining an "inflamed" lung adenocarcinoma subtype with favorable survival estimates [188]. The methodology is fast (the model was trained in less than 30 min in "medium" mode on a laptop with 64-bit Windows 10 operating system, i7 CPU 1.8GHz and 16 GB RAM), sparse and can cope with missing values. However, MOFA+ also suffers from some general limitations of unsupervised data-integration methodologies. In detail, while matrix factorization techniques are often used to reduce the feature space from tens of thousands to a significantly lower number, they might inadvertently ignore a large amount of biological information concerning relationships between features. The same caveat is intrinsic to MOFA+, as the model assumes independence between features in its prior distribution. Furthermore,

MOFA+ by nature is limited to capture strong non-linear relationships, which could be an issue when trying to analyze noisy datasets with high non-linearities as these would result in small amounts of variance explained. Overall, while matrix decomposition methodologies are quite popular as the method of choice among the available unsupervised data integration approaches, the biological interpretation of the inferred latent factors can be a challenging process [189]. A representative example is somatic mutations: latent factors are essentially defined as linear combinations of features. Thus, for a factor to exist it requires an effect over multiple features. Sometimes, somatic mutations don't "behave" like this, as a single somatic mutation can produce a large downstream effect on the expression level, rather than having a contribution to a single factor. This was also evident in our analysis, where based on an initial exploratory training of the model, the plot of 'resulted variance explained' showed that the somatic mutations did not have a contribution over the factors, and thus seemed to behave as independent features.

The performed supervised analysis for patient classification with netDx based on patients suffering from chronic lymphocytic leukemia (CLL) illustrates a typical use case for supervised multi-omics data integration. The challenge here lies in applying multi-omics data integration on a sparse cohort of less than 1000 patients, with missing values, unequal feature sizes per layer and unequal class sizes (class imbalance). The used cohort includes clinical, methylation, drug response, and gene expression data. Unfortunately, we were not able to integrate mutation data, likely because of the sparsity of this data layer which comprised of the binary mutation status for 69 genes. The aim of this case study was to show the challenges and possibilities when working with multi-omics data for a specific research purpose. Supervised analysis in the multi-omics field is based on prior knowledge of the data and its biomedical context, which impacts both hypothesis and primary feature selection. In our study we classified CLL patients based on their IGHV mutation status in order to separate patients with better treatment response prognosis (IGHV-mutated) from those with a worse prognosis (no IGHV mutation). The majority of findings in the netDx study is based on the gene expression layer, which was used for pathway enrichment analysis (see Figures 5 and 6).

The classification of CLL patients based on IGHV status resulted in multiple clusters (see Figure 8B). The implementation of pathway enrichment analysis for methylation and drug response layers is likely to have a great effect on the separation of classes in resulting PSN. The interpretation of results also needs to take into account the differences in class imbalance.

The limitations of this study are not only based on the aforementioned data related issues but also due to the nature of supervised methodologies. Supervised approaches need to be validated on an external dataset in order to evaluate potential overfitting and the generalizability of predictions. Therefore, the good performance of netDx needs to be validated on an independent larger dataset using the same features as in the applied model. Observed differences in performance between netDx v. 1.2.2 and 1.3.1 led us to apply netDx v1.3.1 as an application under development in a Docker container. Increasing the number of performed cross validations could identify more clearly generalizable patterns in a small or very heterogeneous dataset, but it would further increase the calculation time. Fortunately, there are more enriched pathways for non IGHV mutated patients, which have a worse prognosis, than for IGHV mutated patients. Targeted therapy of these patients based on their enriched pathways could lead to better prognosis for these patients. In summary, we could identify enriched pathways which are known to be involved in the pathophysiology of CLL. Furthermore, we could highlight interconnected pathways in the mTOR and the TP63 network in non-IGHV mutated samples. After validation of these results, they could help lead to the refinement of existing treatment combinations for targeting the aforementioned enriched networks for IGHV-mutated and non-IGHV mutated patients.

Furthermore, it is essential to highlight some putative limitations of our literature review. The literature research has been performed in an automated framework by using

the Entrez Direct (EDirect) tool. General limitations of automated annotations concern the challenge to classify literature without evaluating its context. For example, publications of tools such as SIMMS [190] failed to be considered in our literature search, as they use the word multi-modal which is not used in our mining process as it is more general and increases the number of unspecific search results. Although the classification of papers in our review is partly based on occurrence of words in title and abstract, which does not necessarily correspond to the content of the paper, we can see from the results shown in Supplementary Figure S1 and categorization of multi-omics integration tools in Tables 2–7, that the classification of mined literature into five categories worked quite well. The selected papers for tool classification all refer to the corresponding general research purpose. Another limitation of this research is the sole use of PubMed as literature database, which may exclude some papers such as more technical oriented literature, However, PubMed contains cancer-related publications and thus suited our motivation of providing an overview of literature for cancer-related multi-omics data integration.

Finally, in addition to the main aforementioned challenges that govern the integration, sharing and utilization of distinct omics sources, we would like to summarize three major aspects, which facilitate the robust and successful amalgamation of heterogeneous cancer data layers:

1. Initially, one important aspect that influences the integration part includes the pre-processing steps prior to integrating any multimodal data: initially, appropriate normalization or transformation is essential to remove any technical confounders related to each omic data layer. For example, for count based data like RNA-Seq, size factor normalization and variance stabilization are generally recommended. Also, significant differences in size in at least one of the interrogated omics layers could inflate the data integration model to capture non-biological variation associated with this specific data layer, while downweighting more subtle sources of variation. In addition, it is well known that most genomic studies suffer from the "curse of dimensionality", that is the number of features being substantially higher than the number of samples. Hence, a feature selection step like selecting the top most variable features per omic modality is essential, both in supervised and unsupervised approaches. However, filtering is not trivial especially when dealing with somatic mutations or copy number alterations, where a more "sophisticated" filtering is needed. Somatic mutations can be very sparse with the vast majority of cancer genes being of low prevalence, cancer-specific and not shared among all patients of the same cancer. Intratumoral diversification adds further complexity to the application of a simple reduction based on the frequency of no events [191]. Instead, clinical data portals with prior biological knowledge should be used along with computational frameworks to identify putative driver genes, aiming to reduce the CNV/somatic mutations feature space. A quality control step is critical to investigate the percentage and distribution of missing values relative to the number of total samples. In the near future, improvements of the human reference genome (GRCh38) could increase completeness of multi-omics studies. For example, applied telomere to telomere long-read sequencing has started to fill unresolved gaps in the human reference genome for the X Chromosome [192]. Further ongoing efforts will reveal new functional landscapes by creating a human pan-genome, which would include diverse sets of individuals in order to catch the genomic variation across different populations [193]. This will provide the opportunity to study genetic similarities and differences among human populations within genomic or multi-omics studies of complex diseases.
2. Additionally, another crucial part lies in the biological interpretation of the integrative analysis: it is vital to associate any findings to molecular mechanisms and perturbed pathways, in order to identify any causal regulatory relationships between the profiled entities. For this purpose there are various recent tools and databases that perform pathway analysis and provide prior knowledge on molecular biology to construct intracellular communication networks well-suited for multi-omics functional anno-

tation. These include but are not limited to SignaLink 2.0 [123], OmniPath [124,194], ReactomeGSA [195] and ActivePathways [196]. Moreover, incorporation of prior knowledge from clinical data portals could further facilitate the prioritization of features or signatures from multimodal studies, which could serve as putative biomarkers. A representative example is the Variant Interpretation for Cancer Consortium (VICC) meta-knowledgebase [197], a harmonized effort for cancer variant interpretation by encapsulating multiple different cancer variant annotation databases. VICC can be utilized for the validation of putative biomarkers from multi-omics cancer studies. Consequently, the development of multi-omics data integration methodologies that incorporate such prior biological knowledge should be enhanced as well, in order to delineate more readable causal networks between the perturbed omics in cancer manifestation. COSMOS (Causal Oriented Search of Multi-Omics Space) for example integrates phosphoproteomics, transcriptomics, and metabolomics data sets with prior knowledge such as protein-protein interactions to create hypotheses about causal links between signaling kinase cascades, transcriptional factors and metabolites [198].

3. Finally, a researcher should strongly consider to follow specific protocols such as the FAIR guiding principles (findability, accessibility, interoperability, and reusability) [199] when publishing multi-omics cancer data, and to take into account important bioethics considerations when sharing cancer patient data [200].

4. Materials and Methods

4.1. Literature Review Workflow

A systematic and automated literature search in the PubMed database was performed with the aim to collect publications of interest and classify them into distinct meaningful classes without manual configuration. Based on fast rising numbers of publications in the field of multi-omics data integration, the here portrayed results may change heavily in the future. For literature mining we employed the Entrez Direct (EDirect) tool, which allows systematic filtering of publications when accessing the NCBI publication databases, and query for multiple molecular data types in a command-line frame [201]. EDirect facilitates a multi-step search in one single command with the use of piped command blocks. The used keywords mesh terms, and additional search criteria were chosen to be specific for multi-omics data integration in the field of multi-view learning (see Table 8). We refrained from the use of multi-view learning as a search term itself, as multi-view data refers to the general use of any kind of heterogeneous data [202] and associated techniques are also widely applied for non-clinical purposes [203]. Multi-view applications should therefore be applied carefully in the multi-omics context in the literature search [204].

Table 8. Summary of search strategies and associated keywords for collecting multi-omics data integration associated publications with EDirect (PTYP: publication type; MESH: medical subject headings).

Search Round	Search Terms
General Search	(multi AND omics) OR multi-omics OR multiomics OR (multivariate AND genomic) OR (Algorithms AND integrative AND Cluster AND Analysis)) AND data AND integration
Searching for supervised methods	General Search + supervised
Searching for unsupervised methods	General Search + unsupervised OR cluster OR (Factor AND Analysis) NOT supervised
Searching for reviews	General Search + review [PTYP] OR review
Searching for tools	General Search + Tool OR Application OR Algorithm OR method
Searching for cancer	General Search + humans [MESH] AND cancer [MESH]) OR cancer

The retrieved publications are listed in Supplementary Table S1 and include the publication date, the PubMed ID and the title. Furthermore, we defined five different classes with specific Entrez search filtering criteria in order to define the purpose of the paper (cancer, review, or tool) and the nature of the presented methodology (supervised or unsupervised). In order to keep it simple and clear, we did not use more complex classifications like semi-supervised or recurrent learning. For a detailed description of EDirect commands see the Supplementary Materials section.

4.2. LUAD Dataset and MOFA Analysis

Despite the significant advances in targeted treatments with receptor tyrosine kinase inhibitors like Sunitinib or immune checkpoint inhibitors, lung cancer remains the first leading cause of cancer-related deaths and the second most commonly diagnosed cancer worldwide in both sexes, based on the WHO GLOBOCAN database (http://gco.iarc.fr/today/fact-sheets-cancers) epidemiological data for 2020. This can be largely attributed to its propensity to metastasize to the brain, and its high lineage plasticity resulting in poor prognosis and treatment relapse [171,205,206]. Amongst the two major types of lung cancer, namely non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC), lung adenocarcinoma (LUAD) is the most common histological subtype. Overall, lung cancer is considered a highly heterogeneous disease, with complex etiology.

To use a reproducible case study example for the integration of heterogeneous public cancer multi-omics data with an unsupervised approach, we retrieved and downloaded the LUAD TCGA cohort genomic cancer multimodal dataset [168] using the R package curatedTCGAData (v. 1.12.0). The complete bioinformatics analysis was performed with custom made scripts in R-4.0.3/Bioconductor. Briefly, the four omics layers gene expression, proteomics, copy number variation, and somatic mutations were selected for an initial number of patients, utilizing the MultiAssayExperiment integrative data container R package (v. 1.16.0). The total number of patients and features for each assay are illustrated in the Supplementary Figure S1. Using the R package TCGAutils (v. 1.10.0) for initial preprocessing, the final number of common patients profiled across all the assays was 181. For the gene expression data (Upper quartile normalized RSEM TPM gene expression values) downstream analysis included normalization using the variance stabilizing transformation (VST) from DESeq2 (v1.30.0) [207]. Then, we applied a non-specific intensity filtering to remove genes that are not expressed in more than 50% of all samples. Afterwards, additional variance filtering for dimensionality reduction was performed using M3C (v. 1.12.0) [208], resulting in 6958 genes remaining in the dataset.

For an extended feature reduction, we selected only those genes from the copy number alteration and the somatic mutation data that overlapped with the aforementioned final expression genes. In addition, for the somatic mutations, we further performed an intersection of the top 100 most frequently mutated genes, with the COSMIC, Cancer Gene Census (CGC) gene list [209]. This resulted in 13 common genes, which were used as external clinical covariates for downstream analysis. From the protein data we only removed those proteins that had missing (NA) values in the majority of samples, as the downloaded RPPA data were already normalized.

After data preprocessing we used the R package MOFA+ (1.0.1) [155], an unsupervised factor analysis model to perform multi-omics data integration based on the three layers expression, copy number alterations and proteins, while the somatic mutations were used as external clinical covariates. For model training we used default parameters (number of factors = 15, convergence mode = "medium"). To investigate and interpret the output of the model, we utilized various internal package functions. Additionally, we applied principal component gene set enrichment (PCGSE) [210] with Reactome gene sets [211] downloaded from MSigDB [121,212] to interrelate the inferred latent factors to biological processes and molecular pathways. Finally, we isolated all the numeric inferred latent factors to conduct unsupervised clustering of the patients in a multi-omics fashion, with the ultimate goal of predicting discrete clusters that could resemble disease subtypes. For the selection of the

optimal number of clusters we utilized the M3C package [208] (Monte Carlo iterations = 100, resampling reps for reference-real data = 250, inner clustering algorithm = kmeans). Visualization of the resulting clusters along with available clinicopathological data was conducted using the ComplexHeatmap R package (v. 2.6.2) [213].

4.3. CLL Dataset and netDx Analysis

Chronic lymphocytic leukemia (CLL) is the most common type of adult leukemia in the western world. CLL is known as a chronic disease affecting B lymphocyte activity. B cells are activated by different stimuli of the B cell receptors (BCR) coming from cytogenetic abnormalities as well as different genetic alterations. Most of CLL patients carry at least one of four common chromosomal alterations, namely deletion 13q14, deletion 11q22-23, deletion 17p12, and trisomy 12. Frequently, mutations include genes that can be integrated into the NOTCH signaling, inflammatory receptor, MAPK, NFKB, DNA damage and cell cycle control, chromatin modification, transcription, and ribosomal processing pathways [214]. However, the underlying role of included genetic alterations for development and progression of CLL is still largely unknown. There is also growing evidence which implicates aberrant signaling through the mTOR pathway in B cell malignancies [182,215,216].

A multi-omics drug perturbation study [217] using CNV, methylation, mutation, gene expression, and drug response measurements clustered 246 CLL patients into three groups based on their drug response. These groups were separated by signals belonging to the BCR pathway, the mTOR pathway or MEK pathway. The study highlights the IGHV gene mutation status and trisomy 12 as very important markers of kinase inhibition in their integrated analysis.

Somatic mutations of the IGHV gene are known to be prognostic clinical markers for chemoimmunotherapy outcome and therefore crucial factors for patient survival [176]. Analysis of a subcohort using the MOFA algorithm reported the somatic mutation status of the IGHV and trisomy 12 as driving sources of molecular heterogeneity of CLL [155] The subcohort comprised 200 patients with CLL including gene expression data (5000 features, 136 samples), mutation data (69 features, 200 samples), methylation data (4248 features, 196 samples), and drug response data (310 features, 184 samples).

In the analysis the IGHV status was linked to the differentiation of cancer cells and the activation of B-cell receptors, and is the main factor driving the variance in the gene expression layer of the used CLL cohort.

We chose this subcohort for supervised multi-omics classification on the IGHV mutation status using netDx [107]. One motivation was to complement an unsupervised analysis with a supervised one to increase the evidence for the importance of the IGHV status in CLL. Another motivation for selecting this dataset was that it represents well the challenges for multi-omics integration tools with few samples (<1000), unequal distributed missingness (136–200 samples), and unequal feature representation for different layers (69–5000 features per layer). The analysis was performed with R/Rstudio.

We applied netDx v. 1.3.1 with R-4.0.3 by using one CPU on a MacBook Pro with macOS Big Sur 11.1, 3.1 GHz Quad-Core Intel Core i7 and 16 GB memory. We used a Docker container with preinstalled R and netDx as well as all dependencies (https://hub.docker.com/repository/docker/shraddhapai/netdx), using the gene expression, the drug response, and the methylation data layer. The input was defined by the binary IGHV mutation status (0 = not mutated, 1 = mutated). Having to remove 28 samples based on the missing mutation status of the IGHV gene, the final input dataset contained 172 samples (98 with and 74 without IGHV mutation). No further data filtering based on missingness was applied, as netDx can handle missing values in different omics layers. The design of the features for patient classification was grouped by pathways for gene expression data and one feature per layer for the others. We compiled 728 pathways containing 10–200 genes from several curated pathway databases [107]. We used 10 train/test splits with 80% of every split for training and 20% for testing. Feature selection was performed by setting the maximum feature score to 10 (featScoreMax) and the feature

selection threshold to 9 (featSelCutoff). Only features with minimum netDx scoring of 9 were further used for classification of patients in the test set. Well-performing features were selected based on performance across train/test splits. Features needed to score at least 9 in at least 50% of splits. Selected features for the creation of final patient similarity networks (PSNs) need to pass a more strict selection of a minimum score of 9 in at least 70% of splits. For visualization of enrichment maps and similarity networks we applied Cytoscape v. 3.8.2 with EnrichmentMap app v. 3.3.1 and an edge cutoff on similarity of 0.1. See corresponding GitHub repository for detailed step by step procedure.

5. Conclusions

For one, we conclude that automated literature search, although not a guarantee for accurate or comprehensive results, gives a good overview and classification of published knowledge in a specific field, in this case the integration of multi-omics data in oncology. Secondly, the findings of our case studies demonstrate that we can retrieve both known results and novel findings using predominantly the core tools with minimal tuning. Furthermore, in the future the definition and improvement of data sharing and biomedical meta-data to enhance clinical decision support will be of critical importance. Finally, bringing together interdisciplinary computational teams and researchers will help promote the development of cutting edge techniques for multi-omics integration and analysis, increasing the necessity for multi-platform analysis with common datasets to benchmark the performance of various methodologies.

Supplementary Materials: The following are available online at https://www.mdpi.com/1422-0067/22/6/2822/s1.

Author Contributions: Conceptualization, methodology, writing—original draft preparation: E.I.V., J.B. and S.N.; software, validation, formal analysis, investigation, data curation, visualization: E.I.V. and J.B.; resources and funding acquisition: F.Ü.; writing—review and editing, E.I.V., J.B., F.Ü. and S.N.; supervision and project administration: S.N. All authors have read and agreed to the published version of the manuscript.

Funding: E.I.V. was funded by the BMBF as part of the Athens Comprehensive Cancer Center (ACCC). J.B. was funded by the National Center for Tumor Diseases (NCT).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The performed Literature search is displayed in the GitHub repository: https://github.com/jonasboh/multi-omics_literature_search.

Acknowledgments: We thank Shraddha Pai for her help with the netDx v. 1.3.1 analysis and Ricard Argelaguet for his contribution in setting and applying the MOFA+ framework.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Oliver, S. Guilt-by-association goes global. *Nature* **2000**, *403*, 601–602. [CrossRef] [PubMed]
2. Bunnik, E.M.; Le Roch, K.G. An introduction to functional genomics and systems biology. *Adv. Wound Care* **2013**, *2*, 490–498. [CrossRef]
3. Perakakis, N.; Yazdani, A.; Karniadakis, G.E.; Mantzoros, C. Omics, big data and machine learning as tools to propel understanding of biological mechanisms and to discover novel diagnostics and therapeutics. *Metabolism* **2018**, *87*, A1–A9. [CrossRef] [PubMed]
4. Goldman, A.D.; Landweber, L.F. What is a genome? *PLoS Genet.* **2016**, *12*, e1006181. [CrossRef]
5. Wang, Z.; Gerstein, M.; Snyder, M. Rna-seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **2009**, *10*, 57–63. [CrossRef]
6. Timp, W.; Timp, G. Beyond mass spectrometry, the next step in proteomics. *J. Sci. Adv.* **2020**, *6*, eaax8978. [CrossRef] [PubMed]
7. Shin, S.-Y.; Fauman, E.B.; Petersen, A.-K.; Krumsiek, J.; Santos, R.; Huang, J.; Arnold, M.; Erte, I.; Forgetta, V.; Yang, T.-P.; et al. An atlas of genetic influences on human blood metabolites. *Nat. Genet.* **2014**, *46*, 543–550. [CrossRef] [PubMed]
8. Stricker, S.H.; Köferle, A.; Beck, S. From profiles to function in epigenomics. *Nat. Rev. Genet.* **2017**, *18*, 51–66. [CrossRef] [PubMed]

9. Org, E.; Parks, B.W.; Joo, J.W.; Emert, B.; Schwartzman, W.; Kang, E.Y.; Mehrabian, M.; Pan, C.; Knight, R.; Gunsalus, R.; et al. Genetic and environmental control of host-gut microbiota interactions. *Genome Res.* **2015**, *25*, 1558–1569. [CrossRef]
10. Raghu, P. Functional diversity in a lipidome. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 11191–11193. [CrossRef]
11. Ley, T.J.; Mardis, E.R.; Ding, L.; Fulton, B.; McLellan, M.D.; Chen, K.; Dooling, D.; Dunford-Shore, B.H.; McGrath, S.; Hickenbotham, M.; et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **2008**, *456*, 66–72. [CrossRef]
12. Bolton, K.L.; Chenevix-Trench, G.; Goh, C.; Sadetzki, S.; Ramus, S.J.; Karlan, B.Y.; Lambrechts, D.; Despierre, E.; Barrowdale, D.; McGuffog, L.; et al. Association between brca1 and brca2 mutations and survival in women with invasive epithelial ovarian cancer. *JAMA* **2012**, *307*, 382–390. [CrossRef]
13. Bailey, M.H.; Tokheim, C.; Porta-Pardo, E.; Sengupta, S.; Bertrand, D.; Weerasinghe, A.; Colaprico, A.; Wendl, M.C.; Kim, J.; Reardon, B.; et al. Comprehensive characterization of cancer driver genes and mutations. *Cell* **2018**, *173*, 371–385.e318. [CrossRef] [PubMed]
14. Yi, S.; Lin, S.; Li, Y.; Zhao, W.; Mills, G.B.; Sahni, N. Functional variomics and network perturbation: Connecting genotype to phenotype in cancer. *Nat. Rev. Genet.* **2017**, *18*, 395–410. [CrossRef]
15. Francies, H.E.; McDermott, U.; Garnett, M.J. Genomics-guided pre-clinical development of cancer therapies. *Nat. Cancer* **2020**, *1*, 482–492. [CrossRef]
16. Cancer Genome Atlas Research, N.; Weinstein, J.N.; Collisson, E.A.; Mills, G.B.; Shaw, K.R.M.; Ozenberger, B.A.; Ellrott, K.; Shmulevich, I.; Sander, C.; Stuart, J.M. The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* **2013**, *45*, 1113–1120.
17. Hoadley, K.A.; Yau, C.; Hinoue, T.; Wolf, D.M.; Lazar, A.J.; Drill, E.; Shen, R.; Taylor, A.M.; Cherniack, A.D.; Thorsson, V.; et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* **2018**, *173*, 291–304. [CrossRef] [PubMed]
18. Karczewski, K.J.; Snyder, M.P. Integrative omics for health and disease. *Nat. Rev. Genet.* **2018**, *19*, 299–310. [CrossRef]
19. Rheinbay, E. The genomic landscape of advanced cancer. *Nat. Cancer* **2020**, *1*, 372–373. [CrossRef]
20. Hasin, Y.; Seldin, M.; Lusis, A. Multi-omics approaches to disease. *Genome Biol.* **2017**, *18*, 83. [CrossRef]
21. Ebrahim, A.; Brunk, E.; Tan, J.; O'Brien, E.J.; Kim, D.; Szubin, R.; Lerman, J.A.; Lechner, A.; Sastry, A.; Bordbar, A.; et al. Multi-omic data integration enables discovery of hidden biological regularities. *Nat. Commun.* **2016**, *7*, 13091. [CrossRef]
22. Baldwin, E.; Han, J.; Luo, W.; Zhou, J.; An, L.; Liu, J.; Zhang, H.H.; Li, H. On fusion methods for knowledge discovery from multi-omics datasets. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 509–517. [CrossRef]
23. Shannon, C.P.; Blimkie, T.M.; Ben-Othman, R.; Gladish, N.; Amenyogbe, N.; Drissler, S.; Edgar, R.D.; Chan, Q.; Krajden, M.; Foster, L.J.; et al. Multi-omic data integration allows baseline immune signatures to predict hepatitis b vaccine response in a small cohort. *Front. Immunol.* **2020**, *11*, 578801. [CrossRef]
24. Wang, B.; Lunetta, K.L.; Dupuis, J.; Lubitz, S.A.; Trinquart, L.; Yao, L.; Ellinor, P.T.; Benjamin, E.J.; Lin, H. Integrative omics approach to identifying genes associated with atrial fibrillation. *Circ. Res.* **2020**, *126*, 350–360. [CrossRef] [PubMed]
25. Zhao, S.; Jiang, H.; Liang, Z.H.; Ju, H. Integrating multi-omics data to identify novel disease genes and single-neucleotide polymorphisms. *Front. Genet.* **2019**, *10*, 1336. [CrossRef] [PubMed]
26. Woo, H.G.; Choi, J.-H.; Yoon, S.; Jee, B.A.; Cho, E.J.; Lee, J.-H.; Yu, S.J.; Yoon, J.-H.; Yi, N.-J.; Lee, K.-W.; et al. Integrative analysis of genomic and epigenomic regulation of the transcriptome in liver cancer. *Nat. Commun.* **2017**, *8*, 839. [CrossRef] [PubMed]
27. Zhu, B.; Song, N.; Shen, R.; Arora, A.; Machiela, M.J.; Song, L.; Landi, M.T.; Ghosh, D.; Chatterjee, N.; Baladandayuthapani, V.; et al. Integrating clinical and multiple omics data for prognostic assessment across human cancers. *Sci. Rep.* **2017**, *7*, 16954. [CrossRef] [PubMed]
28. Orlandi, E.; Iacovelli, N.A.; Tombolini, V.; Rancati, T.; Polimeni, A.; De Cecco, L.; Valdagni, R.; De Felice, F. Potential role of microbiome in oncogenesis, outcome prediction and therapeutic targeting for head and neck cancer. *Oral Oncol.* **2019**, *99*, 104453. [CrossRef]
29. Stegle, O.; Teichmann, S.A.; Marioni, J.C. Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* **2015**, *16*, 133–145. [CrossRef] [PubMed]
30. Miyamoto, D.T.; Lee, R.J.; Kalinich, M.; LiCausi, J.A.; Zheng, Y.; Chen, T.; Milner, J.D.; Emmons, E.; Ho, U.; Broderick, K.; et al. An rna-based digital circulating tumor cell signature is predictive of drug response and early dissemination in prostate cancer. *J. Cancer Discov.* **2018**, *8*, 288–303. [CrossRef] [PubMed]
31. Das, T.; Andrieux, G.; Ahmed, M.; Chakraborty, S. Integration of online omics-data resources for cancer research. *Front. Genet.* **2020**, *11*, 578345. [CrossRef] [PubMed]
32. Grossman, R.L.; Heath, A.P.; Ferretti, V.; Varmus, H.E.; Lowy, D.R.; Kibbe, W.A.; Staudt, L.M. Toward a shared vision for cancer genomic data. *N. Engl. J. Med.* **2016**, *375*, 1109–1112. [CrossRef] [PubMed]
33. Campbell, P.J.; Getz, G.; Korbel, J.O.; Stuart, J.M.; Jennings, J.L.; Stein, L.D.; Perry, M.D.; Nahal-Bose, H.K.; Ouellette, B.F.F.; Li, C.H.; et al. Pan-cancer analysis of whole genomes. *Nature* **2020**, *578*, 82–93.
34. Ghandi, M.; Huang, F.W.; Jané-Valbuena, J.; Kryukov, G.V.; Lo, C.C.; McDonald, E.R.; Barretina, J.; Gelfand, E.T.; Bielski, C.M.; Li, H.; et al. Next-generation characterization of the cancer cell line encyclopedia. *Nature* **2019**, *569*, 503–508. [CrossRef]
35. Cerami, E.; Gao, J.; Dogrusoz, U.; Gross, B.E.; Sumer, S.O.; Aksoy, B.A.; Jacobsen, A.; Byrne, C.J.; Heuer, M.L.; Larsson, E.; et al. The cbio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *J. Cancer Discov.* **2012**, *2*, 401–404. [CrossRef]

36. Gao, J.; Aksoy, B.A.; Dogrusoz, U.; Dresdner, G.; Gross, B.; Sumer, S.O.; Sun, Y.; Jacobsen, A.; Sinha, R.; Larsson, E.; et al. Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal. *J. Sci. Signal.* **2013**, *6*, pl1. [CrossRef]
37. Tate, J.G.; Bamford, S.; Jubb, H.C.; Sondka, Z.; Beare, D.M.; Bindal, N.; Boutselakis, H.; Cole, C.G.; Creatore, C.; Dawson, E.; et al. Cosmic: The catalogue of somatic mutations in cancer. *Nucleic Acids Res.* **2018**, *47*, D941–D947. [CrossRef] [PubMed]
38. Abaan, O.D.; Polley, E.C.; Davis, S.R.; Zhu, Y.J.; Bilke, S.; Walker, R.L.; Pineda, M.; Gindin, Y.; Jiang, Y.; Reinhold, W.C.; et al. The exomes of the nci-60 panel: A genomic resource for cancer biology and systems pharmacology. *J. Cancer Res.* **2013**, *73*, 4372–4382. [CrossRef] [PubMed]
39. Krassowski, M.; Das, V.; Sahu, S.K.; Misra, B.B. State of the field in multi-omics research: From computational needs to data mining and sharing. *Front. Genet.* **2020**, *11*, 610798. [CrossRef]
40. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Ghemawat, S.; et al. Tensorflow: Large-scale machine learning on heterogeneous systems. *arXiv* **2016**, arXiv:1603.04467.
41. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Kileen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*; Curran Associates Inc.: New York, NY, USA, 2019; Volume 32, pp. 8024–8035.
42. Gentleman, R.C.; Carey, V.J.; Bates, D.M.; Bolstad, B.; Dettling, M.; Dudoit, S.; Ellis, B.; Gautier, L.; Ge, Y.; Gentry, J.; et al. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol.* **2004**, *5*, R80. [CrossRef] [PubMed]
43. Huber, W.; Carey, V.J.; Gentleman, R.; Anders, S.; Carlson, M.; Carvalho, B.S.; Bravo, H.C.; Davis, S.; Gatto, L.; Girke, T.; et al. Orchestrating high-throughput genomic analysis with bioconductor. *Nat. Methods* **2015**, *12*, 115–121. [CrossRef] [PubMed]
44. Morgan, M. *Genomic Data Commons: NIH/NCI Genomic Data Commons Access*; Bioconductor: Seattle, WA, USA, 2016. [CrossRef]
45. Colaprico, A.; Silva, T.C.; Olsen, C.; Garofano, L.; Cava, C.; Garolini, D.; Sabedot, T.S.; Malta, T.M.; Pagnotta, S.M.; Castiglioni, I.; et al. Tcgabiolinks: An r/bioconductor package for integrative analysis of tcga data. *Nucleic Acids Res.* **2015**, *44*, e71. [CrossRef] [PubMed]
46. Ramos, M.; Waldron, L.; Schiffer, L.; Geistlinger, L.; Obenchain, V.; Morgan, M. *Curatedtcgadata: Curated Data from the Cancer Genome Atlas (TCGA) as Multiassay Experiment Objects*, Bioconductor: Seattle, WA, USA, 2020. [CrossRef]
47. Ramos, M.; Geistlinger, L.; Oh, S.; Schiffer, L.; Azhar, R.; Kodali, H.; de Bruijn, I.; Gao, J.; Carey, V.J.; Morgan, M.; et al. Multiomic integration of public oncology databases in bioconductor. *JCO Clin. Cancer Inform.* **2020**, *4*, 958–971. [CrossRef] [PubMed]
48. Ramos, M.; Schiffer, L.; Re, A.; Azhar, R.; Basunia, A.; Rodriguez, C.; Chan, T.; Chapman, P.; Davis, S.R.; Gomez-Cabrero, D.; et al. Software for the integration of multiomics experiments in bioconductor. *J. Cancer Res.* **2017**, *77*, e39–e42. [CrossRef]
49. Yousefi, S.; Amrollahi, F.; Amgad, M.; Dong, C.; Lewis, J.E.; Song, C.; Gutman, D.A.; Halani, S.H.; Velazquez Vega, J.E.; Brat, D.J.; et al. Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Sci. Rep.* **2017**, *7*, 11707. [CrossRef]
50. Goldman, M.; Craft, B.; Hastie, M.; Repečka, K.; McDade, F.; Kamath, A.; Banerjee, A.; Luo, Y.; Rogers, D.; Brooks, A.N.; et al. The ucsc xena platform for public and private cancer genomics data visualization and interpretation. *bioRvix* **2019**, 326470.
51. Nhat, T.; pyup.io bot; DeepSource Bot; Moss, S. *Biomecis-Lab/Openomics: Bug Fixes from Pyopensci Reviewer 2*, Zenodo: Meyrin, Switzerland, 2021.
52. Nalisnik, M.; Amgad, M.; Lee, S.; Halani, S.H.; Velazquez Vega, J.E.; Brat, D.J.; Gutman, D.A.; Cooper, L.A.D. Interactive phenotyping of large-scale histology imaging data with histomicsml. *Sci. Rep.* **2017**, *7*, 14588. [CrossRef]
53. Levings, D.C.; Wang, X.; Kohlhase, D.; Bell, D.A.; Slattery, M. A distinct class of antioxidant response elements is consistently activated in tumors with nrf2 mutations. *Redox Biol.* **2018**, *19*, 235–249. [CrossRef] [PubMed]
54. Giwa, A.; Fatai, A.; Gamieldien, J.; Christoffels, A.; Bendou, H. Identification of novel prognostic markers of survival time in high-risk neuroblastoma using gene expression profiles. *Oncotarget* **2020**, *11*, 4293–4305. [CrossRef] [PubMed]
55. Ostrovsky, A.; Hillman-Jackson, J.; Bouvier, D.; Clements, D.; Afgan, E.; Blankenberg, D.; Schatz, M.C.; Nekrutenko, A.; Taylor, J.; Team, t.G.; et al. Using galaxy to perform large-scale interactive data analyses—an update. *Curr. Protoc.* **2021**, *1*, e31.
56. McGowan, T.; Johnson, J.E.; Kumar, P.; Sajulga, R.; Mehta, S.; Jagtap, P.D.; Griffin, T.J. Multi-omics visualization platform: An extensible galaxy plug-in for multi-omics data visualization and exploration. *GigaScience* **2020**, *9*. [CrossRef]
57. Ritchie, M.D.; Holzinger, E.R.; Li, R.; Pendergrass, S.A.; Kim, D. Methods of integrating data to uncover genotype–phenotype interactions. *Nat. Rev. Genet.* **2015**, *16*, 85–97. [CrossRef]
58. Haas, R.; Zelezniak, A.; Iacovacci, J.; Kamrad, S.; Townsend, S.; Ralser, M. Designing and interpreting 'multi-omic' experiments that may change our understanding of biology. *Curr. Opin. Syst. Biol.* **2017**, *6*, 37–45. [CrossRef]
59. Iacovacci, J.; Peluso, A.; Ebbels, T.; Ralser, M.; Glen, R.C. Extraction and integration of genetic networks from short-profile omic data sets. *Metabolites* **2020**, *10*, 435. [CrossRef] [PubMed]
60. Vitrinel, B.; Koh, H.W.L.; Mujgan Kar, F.; Maity, S.; Rendleman, J.; Choi, H.; Vogel, C. Exploiting interdata relationships in next-generation proteomics analysis *. *Mol. Cell. Proteom.* **2019**, *18*, S5–S14. [CrossRef]
61. Dey, S.S.; Kester, L.; Spanjaard, B.; Bienko, M.; van Oudenaarden, A. Integrated genome and transcriptome sequencing of the same cell. *Nat. Biotechnol.* **2015**, *33*, 285–289. [CrossRef] [PubMed]
62. Macaulay, I.C.; Haerty, W.; Kumar, P.; Li, Y.I.; Hu, T.X.; Teng, M.J.; Goolam, M.; Saurat, N.; Coupland, P.; Shirley, L.M.; et al. G&t-seq: Parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* **2015**, *12*, 519–522.

63. Reyes, M.; Billman, K.; Hacohen, N.; Blainey, P.C. Simultaneous profiling of gene expression and chromatin accessibility in single cells. *Adv. Biosyst.* **2019**, *3*, 1900065. [CrossRef] [PubMed]
64. Li, G.; Liu, Y.; Zhang, Y.; Kubo, N.; Yu, M.; Fang, R.; Kellis, M.; Ren, B. Joint profiling of DNA methylation and chromatin architecture in single cells. *Nat. Methods* **2019**, *16*, 991–993. [CrossRef]
65. Lee, D.-S.; Luo, C.; Zhou, J.; Chandran, S.; Rivkin, A.; Bartlett, A.; Nery, J.R.; Fitzpatrick, C.; O'Connor, C.; Dixon, J.R.; et al. Simultaneous profiling of 3d genome structure and DNA methylation in single human cells. *Nat. Methods* **2019**, *16*, 999–1006. [CrossRef]
66. Dixit, A.; Parnas, O.; Li, B.; Chen, J.; Fulco, C.P.; Jerby-Arnon, L.; Marjanovic, N.D.; Dionne, D.; Burks, T.; Raychowdhury, R.; et al. Perturb-seq: Dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *Cell* **2016**, *167*, 1853–1866.e1817. [CrossRef]
67. Adamson, B.; Norman, T.M.; Jost, M.; Cho, M.Y.; Nuñez, J.K.; Chen, Y.; Villalta, J.E.; Gilbert, L.A.; Horlbeck, M.A.; Hein, M.Y.; et al. A multiplexed single-cell crispr screening platform enables systematic dissection of the unfolded protein response. *Cell* **2016**, *167*, 1867–1882.e1821. [CrossRef]
68. Jaitin, D.A.; Weiner, A.; Yofe, I.; Lara-Astiaso, D.; Keren-Shaul, H.; David, E.; Salame, T.M.; Tanay, A.; van Oudenaarden, A.; Amit, I. Dissecting immune circuits by linking crispr-pooled screens with single-cell rna-seq. *Cell* **2016**, *167*, 1883–1896.e1815. [CrossRef] [PubMed]
69. Zhu, C.; Preissl, S.; Ren, B. Single-cell multimodal omics: The power of many. *Nat. Methods* **2020**, *17*, 11–14. [CrossRef]
70. Ma, A.; McDermaid, A.; Xu, J.; Chang, Y.; Ma, Q. Integrative methods and practical challenges for single-cell multi-omics. *Trends Biotechnol.* **2020**, *38*, 1007–1022. [CrossRef]
71. Lee, J.; Hyeon, D.Y.; Hwang, D. Single-cell multiomics: Technologies and data analysis methods. *Exp. Mol. Med.* **2020**, *52*, 1428–1442. [CrossRef] [PubMed]
72. Venteicher, A.S.; Tirosh, I.; Hebert, C.; Yizhak, K.; Neftel, C.; Filbin, M.G.; Hovestadt, V.; Escalante, L.E.; Shaw, M.L.; Rodman, C.; et al. Decoupling genetics, lineages, and microenvironment in idh-mutant gliomas by single-cell rna-seq. *J. Sci.* **2017**, *355*, eaai8478. [CrossRef] [PubMed]
73. Zhang, L.; Dong, X.; Lee, M.; Maslov, A.Y.; Wang, T.; Vijg, J. Single-cell whole-genome sequencing reveals the functional landscape of somatic mutations in b lymphocytes across the human lifespan. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 9014–9019. [CrossRef]
74. Wang, R.; Dang, M.; Harada, K.; Han, G.; Wang, F.; Pool Pizzi, M.; Zhao, M.; Tatlonghari, G.; Zhang, S.; Hao, D.; et al. Single-cell dissection of intratumoral heterogeneity and lineage diversity in metastatic gastric adenocarcinoma. *Nat. Med.* **2021**, *27*, 141–151. [CrossRef] [PubMed]
75. Hou, Y.; Guo, H.; Cao, C.; Li, X.; Hu, B.; Zhu, P.; Wu, X.; Wen, L.; Tang, F.; Huang, Y.; et al. Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Res.* **2016**, *26*, 304–319. [CrossRef]
76. Ji, A.L.; Rubin, A.J.; Thrane, K.; Jiang, S.; Reynolds, D.L.; Meyers, R.M.; Guo, M.G.; George, B.M.; Mollbrink, A.; Bergenstråhle, J.; et al. Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. *Cell* **2020**, *182*, 497–514.e422. [CrossRef]
77. Efremova, M.; Teichmann, S.A. Computational methods for single-cell omics across modalities. *Nat. Methods* **2020**, *17*, 14–17. [CrossRef]
78. Lähnemann, D.; Köster, J.; Szczurek, E.; McCarthy, D.J.; Hicks, S.C.; Robinson, M.D.; Vallejos, C.A.; Campbell, K.R.; Beerenwinkel, N.; Mahfouz, A.; et al. Eleven grand challenges in single-cell data science. *Genome Biol.* **2020**, *21*, 31. [CrossRef]
79. Stuart, T.; Butler, A.; Hoffman, P.; Hafemeister, C.; Papalexi, E.; Mauck, W.M.; Hao, Y.; Stoeckius, M.; Smibert, P.; Satija, R. Comprehensive integration of single-cell data. *Cell* **2019**, *177*, 1888–1902.e1821. [CrossRef]
80. Samir, J.; Rizzetto, S.; Gupta, M.; Luciani, F. Exploring and analysing single cell multi-omics data with vdjview. *BMC Med. Genom.* **2020**, *13*, 29. [CrossRef]
81. Satija, R.; Farrell, J.A.; Gennert, D.; Schier, A.F.; Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **2015**, *33*, 495–502. [CrossRef] [PubMed]
82. McCarthy, D.J.; Campbell, K.R.; Lun, A.T.; Wills, Q.F. Scater: Pre-processing, quality control, normalization and visualization of single-cell rna-seq data in r. *Bioinformatics* **2017**, *33*, 1179–1186. [CrossRef]
83. Kiselev, V.Y.; Kirschner, K.; Schaub, M.T.; Andrews, T.; Yiu, A.; Chandra, T.; Natarajan, K.N.; Reik, W.; Barahona, M.; Green, A.R.; et al. Sc3: Consensus clustering of single-cell rna-seq data. *Nat. Methods* **2017**, *14*, 483–486. [CrossRef] [PubMed]
84. Welch, J.D.; Hartemink, A.J.; Prins, J.F. Matcher: Manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. *Genome Biol.* **2017**, *18*, 138. [CrossRef] [PubMed]
85. Liu, J.; Huang, Y.; Singh, R.; Vert, J.-P.; Noble, W.S. Jointly embedding multiple single-cell omics measurements. *BioRxiv* **2019**, 644310. [CrossRef]
86. Simidjievski, N.; Bodnar, C.; Tariq, I.; Scherer, P.; Andres Terre, H.; Shams, Z.; Jamnik, M.; Liò, P. Variational autoencoders for cancer data integration: Design principles and computational practice. *Front. Genet.* **2019**, *10*, 1205. [CrossRef] [PubMed]
87. Campbell, K.R.; Steif, A.; Laks, E.; Zahn, H.; Lai, D.; McPherson, A.; Farahani, H.; Kabeer, F.; O'Flanagan, C.; Biele, J.; et al. Clonealign: Statistical integration of independent single-cell rna and DNA sequencing data from human cancers. *Genome Biol.* **2019**, *20*, 54. [CrossRef] [PubMed]
88. Wang, X.; Sun, Z.; Zhang, Y.; Xu, Z.; Xin, H.; Huang, H.; Duerr, R.H.; Chen, K.; Ding, Y.; Chen, W. Brem-sc: A bayesian random effects mixture model for joint clustering single cell multi-omics data. *Nucleic Acids Res.* **2020**, *48*, 5814–5824. [CrossRef] [PubMed]

89. Cao, J.; Cusanovich, D.A.; Ramani, V.; Aghamirzaie, D.; Pliner, H.A.; Hill, A.J.; Daza, R.M.; McFaline-Figueroa, J.L.; Packer, J.S.; Christiansen, L.; et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *J. Sci.* **2018**, *361*, 1380–1385. [CrossRef]
90. Lake, B.B.; Chen, S.; Sos, B.C.; Fan, J.; Kaeser, G.E.; Yung, Y.C.; Duong, T.E.; Gao, D.; Chun, J.; Kharchenko, P.V.; et al. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat. Biotechnol.* **2018**, *36*, 70–80. [CrossRef]
91. Edsgärd, D.; Johnsson, P.; Sandberg, R. Identification of spatial expression trends in single-cell gene expression data. *Nat. Methods* **2018**, *15*, 339–342. [CrossRef]
92. Svensson, V.; Teichmann, S.A.; Stegle, O. Spatialde: Identification of spatially variable genes. *Nat. Methods* **2018**, *15*, 343–346. [CrossRef]
93. Zuo, C.; Chen, L. Deep-joint-learning analysis model of single cell transcriptome and open chromatin accessibility data. *Brief. Bioinform.* **2020**. [CrossRef]
94. Yang, K.D.; Belyaeva, A.; Venkatachalapathy, S.; Damodaran, K.; Katcoff, A.; Radhakrishnan, A.; Shivashankar, G.V.; Uhler, C. Multi-domain translation between single-cell imaging and sequencing data using autoencoders. *Nat. Commun.* **2021**, *12*, 31. [CrossRef] [PubMed]
95. Setty, M.; Tadmor, M.D.; Reich-Zeliger, S.; Angel, O.; Salame, T.M.; Kathail, P.; Choi, K.; Bendall, S.; Friedman, N.; Pe'er, D. Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat. Biotechnol.* **2016**, *34*, 637–645. [CrossRef]
96. Welch, J.D.; Kozareva, V.; Ferreira, A.; Vanderburg, C.; Martin, C.; Macosko, E.Z. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* **2019**, *117*, 1873–1887.e1817. [CrossRef] [PubMed]
97. Duren, Z.; Chen, X.; Zamanighomi, M.; Zeng, W.; Satpathy, A.T.; Chang, H.Y.; Wang, Y.; Wong, W.H. Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 7723–7728. [CrossRef]
98. Argelaguet, R.; Arnol, D.; Bredikhin, D.; Deloro, Y.; Velten, B.; Marioni, J.C.; Stegle, O. Mofa+: A statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol* **2020**, *21*, 111. [CrossRef]
99. Duan, B.; Zhou, C.; Zhu, C.; Yu, Y.; Li, G.; Zhang, S.; Zhang, C.; Ye, X.; Ma, H.; Qu, S.; et al. Model-based understanding of single-cell crispr screening. *Nat. Commun.* **2019**, *10*, 2233. [CrossRef]
100. KwameForbes. Integratewithsinglecell. Available online: https://rdrr.io/bioc/DESeq2/man/integrateWithSingleCell.html (accessed on 4 March 2021).
101. Huang, S.; Chaudhary, K.; Garmire, L.X. More is better: Recent progress in multi-omics data integration methods. *Front. Genet.* **2017**, *8*. [CrossRef] [PubMed]
102. Subramanian, I.; Verma, S.; Kumar, S.; Jere, A.; Anamika, K. Multi-omics data integration, interpretation, and its application. *Bioinform Biol. Insights* **2020**, *14*, 1177932219899051. [CrossRef]
103. Nicora, G.; Vitali, F.; Dagliati, A.; Geifman, N.; Bellazzi, R. Integrated multi-omics analyses in oncology: A review of machine learning methods and tools. *Front. Oncol* **2020**, *10*, 1030. [CrossRef] [PubMed]
104. O'Connell, M.J.; Lock, E.F.R. Jive for exploration of multi-source molecular data. *Bioinformatics* **2016**, *32*, 2877–2879. [CrossRef] [PubMed]
105. Beal, J.; Montagud, A.; Traynard, P.; Barillot, E.; Calzone, L. Personalization of logical models with multi-omics data allows clinical stratification of patients. *Front. Physiol* **2018**, *9*, 1965. [CrossRef]
106. Huang, Z.; Zhan, X.; Xiang, S.; Johnson, T.S.; Helm, B.; Yu, C.Y.; Zhang, J.; Salama, P.; Rizkalla, M.; Han, Z.; et al. Salmon: Survival analysis learning with multi-omics neural networks on breast cancer. *Front. Genet.* **2019**, *10*, 166. [CrossRef] [PubMed]
107. Pai, S.; Hui, S.; Isserlin, R.; Shah, M.A.; Kaka, H.; Bader, G.D. Netdx: Interpretable patient classification using integrated patient similarity networks. *Mol. Syst. Biol.* **2019**, *15*, e8497. [CrossRef] [PubMed]
108. Velten, B.; Huber, W. Adaptive penalization in high-dimensional regression and classification with external covariates using variational bayes. *Biostatistics* **2019**. [CrossRef]
109. Vlachavas, E.-I.; Pilalis, E.; Papadodima, O.; Koczan, D.; Willis, S.; Klippel, S.; Cheng, C.; Pan, L.; Sachpekidis, C.; Pintzas, A.; et al. Radiogenomic analysis of f-18-fluorodeoxyglucose positron emission tomography and gene expression data elucidates the epidemiological complexity of colorectal cancer landscape. *Comput. Struct. Biotechnol. J.* **2019**, *17*, 177–185. [CrossRef] [PubMed]
110. González-Reymúndez, A.; Vázquez, A.I. Multi-omic signatures identify pan-cancer classes of tumors beyond tissue of origin. *Sci. Rep.* **2020**, *10*, 8341. [CrossRef]
111. Ray, P.; Zheng, L.; Lucas, J.; Carin, L. Bayesian joint analysis of heterogeneous genomics data. *Bioinformatics* **2014**, *30*, 1370–1376. [CrossRef]
112. Song, X.; Ji, J.; Gleason, K.J.; Yang, F.; Martignetti, J.A.; Chen, L.S.; Wang, P. Insights into impact of DNA copy number alteration and methylation on the proteogenomic landscape of human ovarian cancer via a multi-omics integrative analysis. *Mol. Cell Proteom.* **2019**, *18*, S52–S65. [CrossRef]
113. Lin, D.; Zhang, J.; Li, J.; Calhoun, V.D.; Deng, H.W.; Wang, Y.P. Group sparse canonical correlation analysis for genomic data integration. *Bmc Bioinform.* **2013**, *14*, 245. [CrossRef] [PubMed]
114. Li, W.; Zhang, S.; Liu, C.C.; Zhou, X.J. Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. *Bioinformatics* **2012**, *28*, 2458–2466. [CrossRef] [PubMed]
115. Louhimo, R.; Hautaniemi, S. Cnamet: An r package for integrating copy number, methylation and expression data. *Bioinformatics* **2011**, *27*, 887–888. [CrossRef]

116. Wang, W.; Baladandayuthapani, V.; Morris, J.S.; Broom, B.M.; Manyam, G.; Do, K.A. Ibag: Integrative bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics* **2013**, *29*, 149–159. [CrossRef]
117. Ovaska, K.; Laakso, M.; Haapa-Paananen, S.; Louhimo, R.; Chen, P.; Aittomaki, V.; Valo, E.; Nunez-Fontarnau, J.; Rantanen, V.; Karinen, S.; et al. Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. *Genome Med.* **2010**, *2*, 65. [CrossRef] [PubMed]
118. Peng, C.; Zheng, Y.; Huang, D.S. Capsule network based modeling of multi-omics data for discovery of breast cancer-related genes. *IEEE Acm Trans. Comput. Biol. Bioinform* **2020**, *17*, 1605–1612. [CrossRef] [PubMed]
119. Kanehisa, M. Toward understanding the origin and evolution of cellular organisms. *Protein Sci.* **2019**, *28*, 1947–1951. [CrossRef]
120. Jassal, B.; Matthews, L.; Viteri, G.; Gong, C.; Lorente, P.; Fabregat, A.; Sidiropoulos, K.; Cook, J.; Gillespie, M.; Haw, R.; et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* **2019**, *48*, D498–D503. [CrossRef]
121. Subramanian, A.; Tamayo, P.; Mootha, V.K.; Mukherjee, S.; Ebert, B.L.; Gillette, M.A.; Paulovich, A.; Pomeroy, S.L.; Golub, T.R.; Lander, E.S.; et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 15545–15550. [CrossRef] [PubMed]
122. Rodchenkov, I.; Babur, O.; Luna, A.; Aksoy, B.A.; Wong, J.V.; Fong, D.; Franz, M.; Siper, M.C.; Cheung, M.; Wrana, M.; et al. Pathway commons 2019 update: Integration, analysis and exploration of pathway data. *Nucleic Acids Res.* **2019**, *48*, D489–D497. [CrossRef]
123. Fazekas, D.; Koltai, M.; Türei, D.; Módos, D.; Pálfy, M.; Dúl, Z.; Zsákai, L.; Szalay-Bekő, M.; Lenti, K.; Farkas, I.J.; et al. Signalink 2—a signaling pathway resource with multi-layered regulatory networks. *BMC Syst. Biol.* **2013**, *7*, 7. [CrossRef]
124. Türei, D.; Valdeolivas, A.; Gul, L.; Palacio-Escat, N.; Ivanova, O.; Gábor, A.; Módos, D.; Korcsmáros, T.; Saez-Rodriguez, J. Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. *bioRvix* **2020**. [CrossRef]
125. Amar, D.; Shamir, R. Constructing module maps for integrated analysis of heterogeneous biological networks. *Nucleic. Acids Res.* **2014**, *42*, 4208–4219. [CrossRef] [PubMed]
126. Dimitrakopoulos, C.; Hindupur, S.K.; Hafliger, L.; Behr, J.; Montazeri, H.; Hall, M.N.; Beerenwinkel, N. Network-based integration of multi-omics data for prioritizing cancer genes. *Bioinformatics* **2018**, *34*, 2441–2448. [CrossRef] [PubMed]
127. Zhang, S.; Li, Q.; Liu, J.; Zhou, X.J. A novel computational framework for simultaneous integration of multiple types of genomic data to identify microrna-gene regulatory modules. *Bioinformatics* **2011**, *27*, i401–i409. [CrossRef]
128. Vaske, C.J.; Benz, S.C.; Sanborn, J.Z.; Earl, D.; Szeto, C.; Zhu, J.; Haussler, D.; Stuart, J.M. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using paradigm. *Bioinformatics* **2010**, *26*, i237–i245. [CrossRef]
129. Seoane, J.A.; Day, I.N.M.; Gaunt, T.R.; Campbell, C. A pathway-based data integration framework for prediction of disease progression. *Bioinformatics* **2013**, *30*, 838–845. [CrossRef]
130. Savage, S.R.; Shi, Z.; Liao, Y.; Zhang, B. Graph algorithms for condensing and consolidating gene set analysis results. *Mol. Cell Proteom.* **2019**, *18*, S141–S152. [CrossRef]
131. Martini, P.; Chiogna, M.; Calura, E.; Romualdi, C. Mosclip: Multi-omic and survival pathway analysis for the identification of survival associated gene and modules. *Nucleic Acids Res.* **2019**, *47*, e80. [CrossRef]
132. Lawson, J.T.; Smith, J.P.; Bekiranov, S.; Garrett-Bakelman, F.E.; Sheffield, N.C. Cocoa: Coordinate covariation analysis of epigenetic heterogeneity. *Genome Biol* **2020**, *21*, 240. [CrossRef]
133. Turanli, B.; Karagoz, K.; Bidkhori, G.; Sinha, R.; Gatza, M.L.; Uhlen, M.; Mardinoglu, A.; Arga, K.Y. Multi-omic data interpretation to repurpose subtype specific drug candidates for breast cancer. *Front. Genet.* **2019**, *10*. [CrossRef] [PubMed]
134. Sharifi-Noghabi, H.; Zolotareva, O.; Collins, C.C.; Ester, M. Moli: Multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics* **2019**, *35*, i501–i509. [CrossRef] [PubMed]
135. Chen, J.; Zhang, S. Integrative analysis for identifying joint modular patterns of gene-expression and drug-response data. *Bioinformatics* **2016**, *32*, 1724–1732. [CrossRef] [PubMed]
136. Mariette, J.; Villa-Vialaneix, N. Unsupervised multiple kernel learning for heterogeneous data integration. *Bioinformatics* **2018**, *34*, 1009–1015. [CrossRef] [PubMed]
137. Mo, Q.; Shen, R.; Guo, C.; Vannucci, M.; Chan, K.S.; Hilsenbeck, S.G. A fully bayesian latent variable model for integrative clustering analysis of multi-type omics data. *Biostatistics* **2018**, *19*, 71–86. [CrossRef]
138. Wang, B.; Mezlini, A.M.; Demir, F.; Fiume, M.; Tu, Z.; Brudno, M.; Haibe-Kains, B.; Goldenberg, A. Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* **2014**, *11*, 333–337. [CrossRef]
139. Shen, R.; Olshen, A.B.; Ladanyi, M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* **2009**, *25*, 2906–2912. [CrossRef] [PubMed]
140. Mo, Q.; Wang, S.; Seshan, V.E.; Olshen, A.B.; Schultz, N.; Sander, C.; Powers, R.S.; Ladanyi, M.; Shen, R. Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 4245–4250. [CrossRef]
141. Lock, E.F.; Hoadley, K.A.; Marron, J.S.; Nobel, A.B. Joint and individual variation explained (jive) for integrated analysis of multiple data types. *Ann. Appl Stat.* **2013**, *7*, 523–542. [CrossRef] [PubMed]
142. Yuan, Y.; Savage, R.S.; Markowetz, F. Patient-specific data fusion defines prognostic cancer subtypes. *PloS Comput. Biol.* **2011**, *7*, e1002227. [CrossRef]
143. Lock, E.F.; Dunson, D.B. Bayesian consensus clustering. *Bioinformatics* **2013**, *29*, 2610–2616. [CrossRef]
144. Tran, D.; Nguyen, H.; Le, U.; Bebis, G.; Luu, H.N.; Nguyen, T. A novel method for cancer subtyping and risk prediction using consensus factor analysis. *Front. Oncol* **2020**, *10*, 1052. [CrossRef]

145. Ronen, J.; Hayat, S.; Akalin, A. Evaluation of colorectal cancer subtypes and cell lines using deep learning. *Life Sci. Alliance* **2019**, *2*, e201900517. [CrossRef]
146. Meng, C.; Kuster, B.; Culhane, A.C.; Gholami, A.M. A multivariate approach to the integration of multi-omics datasets. *BMC Bioinform.* **2014**, *15*, 162. [CrossRef]
147. Rohart, F.; Gautier, B.; Singh, A.; Le Cao, K.A. Mixomics: An r package for 'omics feature selection and multiple data integration. *PloS Comput. Biol.* **2017**, *13*, e1005752. [CrossRef] [PubMed]
148. Bonnet, E.; Calzone, L.; Michoel, T. Integrative multi-omics module network inference with lemon-tree. *PloS Comput. Biol.* **2015**, *11*, e1003983. [CrossRef] [PubMed]
149. Gabasova, E.; Reid, J.; Wernisch, L. Clusternomics: Integrative context-dependent clustering for heterogeneous datasets. *PloS Comput. Biol.* **2017**, *13*, e1005781. [CrossRef]
150. Champion, M.; Brennan, K.; Croonenborghs, T.; Gentles, A.J.; Pochet, N.; Gevaert, O. Module analysis captures pancancer genetically and epigenetically deregulated cancer driver genes for smoking and antiviral response. *EBioMedicine* **2018**, *27*, 156–166. [CrossRef]
151. Koh, H.W.L.; Fermin, D.; Vogel, C.; Choi, K.P.; Ewing, R.M.; Choi, H. Iomicspass: Network-based integration of multiomics data for predictive subnetwork discovery. *Npj. Syst. Biol. Appl.* **2019**, *5*, 22. [CrossRef]
152. Meng, C.; Basunia, A.; Peters, B.; Gholami, A.M.; Kuster, B.; Culhane, A.C. Mogsa: Integrative single sample gene-set analysis of multiple omics data. *Mol. Cell Proteom.* **2019**, *18*, S153–S168. [CrossRef]
153. Lemsara, A.; Ouadfel, S.; Frohlich, H. Pathme: Pathway based multi-modal sparse autoencoders for clustering of patient-level multi-omics data. *BMC Bioinform.* **2020**, *21*, 146. [CrossRef] [PubMed]
154. Huang, L.; Brunell, D.; Stephan, C.; Mancuso, J.; Yu, X.; He, B.; Thompson, T.C.; Zinner, R.; Kim, J.; Davies, P.; et al. Driver network as a biomarker: Systematic integration and network modeling of multi-omics data to derive driver signaling pathways for drug combination prediction. *Bioinformatics* **2019**, *35*, 3709–3717. [CrossRef]
155. Argelaguet, R.; Velten, B.; Arnol, D.; Dietrich, S.; Zenz, T.; Marioni, J.C.; Buettner, F.; Huber, W.; Stegle, O. Multi-omics factor analysis-a framework for unsupervised integration of multi-omics data sets. *Mol. Syst Biol.* **2018**, *14*, e8124. [CrossRef]
156. Velten, B.; Braunger, J.M.; Arnol, D.; Argelaguet, R.; Stegle, O. Identifying temporal and spatial patterns of variation from multi-modal data using mefisto. *bioRxiv* **2020**. [CrossRef]
157. Cantini, L.; Zakeri, P.; Hernandez, C.; Naldi, A.; Thieffry, D.; Remy, E.; Baudot, A. Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nat. Commun.* **2021**, *12*, 124. [CrossRef] [PubMed]
158. Song, M.; Greenbaum, J.; Luttrell, J.; Zhou, W.; Wu, C.; Shen, H.; Gong, P.; Zhang, C.; Deng, H.-W. A review of integrative imputation for multi-omics datasets. *Front. Genet.* **2020**, *11*. [CrossRef]
159. Kumuthini, J.; Chimenti, M.; Nahnsen, S.; Peltzer, A.; Meraba, R.; McFadyen, R.; Wells, G.; Taylor, D.; Maienschein-Cline, M.; Li, J.-L.; et al. Ten simple rules for providing effective bioinformatics research support. *PloS Comput. Biol.* **2020**, *16*, e1007531. [CrossRef]
160. Alcala, N.; Leblay, N.; Gabriel, A.A.G.; Mangiante, L.; Hervas, D.; Giffon, T.; Sertier, A.S.; Ferrari, A.; Derks, J.; Ghantous, A.; et al. Integrative and comparative genomic analyses identify clinically relevant pulmonary carcinoid groups and unveil the supra-carcinoids. *Nat. Commun.* **2019**, *10*, 3407. [CrossRef] [PubMed]
161. Ha, K.C.H.; Sterne-Weiler, T.; Morris, Q.; Weatheritt, R.J.; Blencowe, B.J. Differential contribution of transcriptomic regulatory layers in the definition of neuronal identity. *Nat. Commun.* **2021**, *12*, 335. [CrossRef] [PubMed]
162. Sudhakar, P.; Verstockt, B.; Cremer, J.; Verstockt, S.; Sabino, J.; Ferrante, M.; Vermeire, S. Understanding the molecular drivers of disease heterogeneity in crohn's disease using multi-omic data integration and network analysis. *Inflamm. Bowel Dis.* **2020**. [CrossRef] [PubMed]
163. Forcato, M.; Romano, O.; Bicciato, S. Computational methods for the integrative analysis of single-cell data. *Brief. Bioinform.* **2020**, *22*, 20–29. [CrossRef]
164. Pai, S.; Weber, P.; Isserlin, R.; Kaka, H.; Hui, S.; Shah, M.; Giudice, L.; Giugno, R.; Nøhr, A.; Baumbach, J.; et al. Netdx: Software for building interpretable patient classifiers by multi-'omic data integration using patient similarity networks [version 2; peer review: 2 approved with reservations]. *F1000Research* **2021**, *9*. [CrossRef]
165. Weber, A.M.; Drobnitzky, N.; Devery, A.M.; Bokobza, S.M.; Adams, R.A.; Maughan, T.S.; Ryan, A.J. Phenotypic consequences of somatic mutations in the ataxia-telangiectasia mutated gene in non-small cell lung cancer. *Oncotarget* **2016**, *7*, 60807–60822. [CrossRef] [PubMed]
166. Chen, Y.; Chen, G.; Li, J.; Huang, Y.-Y.; Li, Y.; Lin, J.; Chen, L.-Z.; Lu, J.-P.; Wang, Y.-Q.; Wang, C.-X.; et al. Association of tumor protein p53 and ataxia-telangiectasia mutated comutation with response to immune checkpoint inhibitors and mortality in patients with non–small cell lung cancer. *Jama Netw. Open* **2019**, *2*, e1911895. [CrossRef]
167. Lee, S.; Rauch, J.; Kolch, W. Targeting mapk signaling in cancer: Mechanisms of drug resistance and sensitivity. *Int. J. Mol. Sci.* **2020**, *21*, 1102. [CrossRef]
168. Collisson, E.A.; Campbell, J.D.; Brooks, A.N.; Berger, A.H.; Lee, W.; Chmielecki, J.; Beer, D.G.; Cope, L.; Creighton, C.J.; Danilova, L.; et al. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **2014**, *511*, 543–550.
169. Escors, D.; Gato-Cañas, M.; Zuazo, M.; Arasanz, H.; García-Granda, M.J.; Vera, R.; Kochan, G. The intracellular signalosome of pd-l1 in cancer cells. *Signal. Transduct. Target. Ther.* **2018**, *3*, 26. [CrossRef]
170. Han, Y.; Liu, D.; Li, L. Pd-1/pd-l1 pathway: Current researches in cancer. *Am. J. Cancer Res.* **2020**, *10*, 727–742. [PubMed]

171. Herbst, R.S.; Morgensztern, D.; Boshoff, C. The biology and management of non-small cell lung cancer. *Nature* **2018**, *553*, 446–454. [CrossRef]
172. Hopewell, E.L.; Zhao, W.; Fulp, W.J.; Bronk, C.C.; Lopez, A.S.; Massengill, M.; Antonia, S.; Celis, E.; Haura, E.B.; Enkemann, S.A.; et al. Lung tumor nf-κb signaling promotes t cell–mediated immune surveillance. *J. Clin. Invest.* **2013**, *123*, 2509–2522. [CrossRef] [PubMed]
173. Westhoff, B.; Colaluca, I.N.; D'Ario, G.; Donzelli, M.; Tosoni, D.; Volorio, S.; Pelosi, G.; Spaggiari, L.; Mazzarol, G.; Viale, G.; et al. Alterations of the notch pathway in lung cancer. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 22293–22298. [CrossRef] [PubMed]
174. Yuan, X.; Wu, H.; Han, N.; Xu, H.; Chu, Q.; Yu, S.; Chen, Y.; Wu, K. Notch signaling and emt in non-small cell lung cancer: Biological significance and therapeutic application. *J. Hematol. Oncol.* **2014**, *7*, 87. [CrossRef]
175. Hothorn, T.; Lausen, B. On the exact distribution of maximally selected rank statistics. *Comput. Stat. Data Anal.* **2003**, *43*, 121–137. [CrossRef]
176. Rozovski, U.; Keating, M.J.; Estrov, Z. Why is the immunoglobulin heavy chain gene mutation status a prognostic indicator in chronic lymphocytic leukemia? *Acta Haematol.* **2018**, *140*, 51–54. [CrossRef] [PubMed]
177. Kucera, M.; Isserlin, R.; Arkhangorodsky, A.; Bader, G. Autoannotate: A cytoscape app for summarizing networks with semantic annotations [version 1; peer review: 2 approved]. *F1000Research* **2016**, *5*. [CrossRef]
178. Sheng, X.; Mittelman, S.D. The role of adipose tissue and obesity in causing treatment resistance of acute lymphoblastic leukemia. *Front. Pediatr* **2014**, *2*, 53. [CrossRef] [PubMed]
179. Sankanagoudar, S.; Singh, G.; Mahapatra, M.; Kumar, L.; Chandra, N.C. Cholesterol homeostasis in isolated lymphocytes: A differential correlation between male control and chronic lymphocytic leukemia subjects. *Asian Pac. J. Cancer Prev.* **2017**, *18*, 23–30.
180. Ha, H.; Debnath, B.; Neamati, N. Role of the cxcl8-cxcr1/2 axis in cancer and inflammatory diseases. *Theranostics* **2017**, *7*, 1543–1588. [CrossRef]
181. Risnik, D.; Podaza, E.; Almejún, M.B.; Colado, A.; Elías, E.E.; Bezares, R.F.; Fernández-Grecco, H.; Cranco, S.; Sánchez-Ávalos, J.C.; Borge, M.; et al. Revisiting the role of interleukin-8 in chronic lymphocytic leukemia. *Sci. Rep.* **2017**, *7*, 15714. [CrossRef] [PubMed]
182. Mirabilii, S.; Ricciardi, M.R.; Piedimonte, M.; Gianfelici, V.; Bianchi, M.P.; Tafuri, A. Biological aspects of mtor in leukemia. *Int. J. Mol. Sci.* **2018**, *19*, 2396. [CrossRef]
183. Ramsay, A.G.; Evans, R.; Kiaii, S.; Svensson, L.; Hogg, N.; Gribben, J.G. Chronic lymphocytic leukemia cells induce defective lfa-1-directed t-cell motility by altering rho gtpase signaling that is reversible with lenalidomide. *Blood* **2013**, *121*, 2704–2714. [CrossRef] [PubMed]
184. Humphries, L.A.; Godbersen, J.C.; Danilova, O.V.; Kaur, P.; Christensen, B.C.; Danilov, A.V. Pro-apoptotic tp53 homolog tap63 is repressed via epigenetic silencing and b-cell receptor signalling in chronic lymphocytic leukaemia. *Br. J. Haematol.* **2013**, *163*, 590–602. [CrossRef] [PubMed]
185. Gillette, M.A.; Satpathy, S.; Cao, S.; Dhanasekaran, S.M.; Vasaikar, S.V.; Krug, K.; Petralia, F.; Li, Y.; Liang, W.-W.; Reva, B.; et al. Proteogenomic characterization reveals therapeutic vulnerabilities in lung adenocarcinoma. *Cell* **2020**, *182*, 200–225.e235. [CrossRef]
186. AbdulJabbar, K.; Raza, S.E.A.; Rosenthal, R.; Jamal-Hanjani, M.; Veeriah, S.; Akarca, A.; Lund, T.; Moore, D.A.; Salgado, R.; Al Bakir, M.; et al. Geospatial immune variability illuminates differential evolution of lung adenocarcinoma. *Nat. Med.* **2020**, *26*, 1054–1062. [CrossRef]
187. Zhang, Y.; Sun, B.; Hu, M.; Lou, Y.; Lu, J.; Zhang, X.; Wang, H.; Qian, J.; Chu, T.; Han, B. Cxcl9 as a prognostic inflammatory marker in early-stage lung adenocarcinoma patients. *Front. Oncol.* **2020**, *10*. [CrossRef]
188. Pocha, K.; Mock, A.; Rapp, C.; Dettling, S.; Warta, R.; Geisenberger, C.; Jungk, C.; Martins, L.R.; Grabe, N.; Reuss, D.; et al. Surfactant expression defines an inflamed subtype of lung adenocarcinoma brain metastases that correlates with prolonged survival. *J. Clin. Cancer Res.* **2020**, *26*, 2231–2243. [CrossRef]
189. Tini, G.; Marchetti, L.; Priami, C.; Scott-Boyer, M.-P. Multi-omics integration—A comparison of unsupervised clustering methodologies. *Brief. Bioinform.* **2017**, *20*, 1269–1279. [CrossRef]
190. Haider, S.; Yao, C.Q.; Sabine, V.S.; Grzadkowski, M.; Stimper, V.; Starmans, M.H.W.; Wang, J.; Nguyen, F.; Moon, N.C.; Lin, X.; et al. Pathway-based subnetworks enable cross-disease biomarker discovery. *Nat. Commun.* **2018**, *9*, 4746. [CrossRef]
191. Martínez-Jiménez, F.; Muiños, F.; Sentís, I.; Deu-Pons, J.; Reyes-Salazar, I.; Arnedo-Pac, C.; Mularoni, L.; Pich, O.; Bonet, J.; Kranas, H.; et al. A compendium of mutational cancer driver genes. *Nat. Rev. Cancer* **2020**, *20*, 555–572. [CrossRef]
192. Miga, K.H.; Koren, S.; Rhie, A.; Vollger, M.R.; Gershman, A.; Bzikadze, A.; Brooks, S.; Howe, E.; Porubsky, D.; Logsdon, G.A.; et al. Telomere-to-telomere assembly of a complete human x chromosome. *Nature* **2020**, *585*, 79–84. [CrossRef]
193. Sherman, R.M.; Salzberg, S.L. Pan-genomics in the human genome era. *Nat. Rev. Genet.* **2020**, *21*, 243–254. [CrossRef] [PubMed]
194. Türei, D.; Korcsmáros, T.; Saez-Rodriguez, J. Omnipath: Guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods* **2016**, *13*, 966–967. [CrossRef] [PubMed]
195. Griss, J.; Viteri, G.; Sidiropoulos, K.; Nguyen, V.; Fabregat, A.; Hermjakob, H. Reactomegsa-efficient multi-omics comparative pathway analysis. *Mol. Cell. Proteom.* **2020**, *19*, 2115–2125. [CrossRef] [PubMed]
196. Paczkowska, M.; Barenboim, J.; Sintupisut, N.; Fox, N.S.; Zhu, H.; Abd-Rabbo, D.; Mee, M.W.; Boutros, P.C.; Abascal, F.; Amin, S.B.; et al. Integrative pathway enrichment analysis of multivariate omics data. *Nat. Commun.* **2020**, *11*, 735. [CrossRef] [PubMed]

197. Wagner, A.H.; Walsh, B.; Mayfield, G.; Tamborero, D.; Sonkin, D.; Krysiak, K.; Deu-Pons, J.; Duren, R.P.; Gao, J.; McMurry, J.; et al. A harmonized meta-knowledgebase of clinical interpretations of somatic genomic variants in cancer. *Nat. Genet.* **2020**, *52*, 448–457. [CrossRef]
198. Dugourd, A.; Kuppe, C.; Sciacovelli, M.; Gjerga, E.; Emdal, K.B.; Bekker-Jensen, D.B.; Kranz, J.; Bindels, E.M.J.; Costa, A.S.H.; Olsen, J.V.; et al. Causal integration of multi-omics data with prior knowledge to generate mechanistic hypotheses. *J. bioRxiv* **2020**. [CrossRef]
199. Boeckhout, M.; Zielhuis, G.A.; Bredenoord, A.L. The fair guiding principles for data stewardship: Fair enough? *Eur J. Hum. Genet.* **2018**, *26*, 931–936. [CrossRef] [PubMed]
200. Dupras, C.; Bunnik, E.M. Toward a framework for assessing privacy risks in multi-omic research and databases. *Am. J. Bioeth.* **2021**, 1–32. [CrossRef]
201. National Center for Biotechnology Information, *Entrez Programming Utilities Help*; National Center for Biotechnology Information: Bethesda, MD, USA, 2010.
202. Li, Y.; Wu, F.X.; Ngom, A. A review on machine learning principles for multi-view biological data integration. *Brief. Bioinform* **2018**, *19*, 325–340. [CrossRef]
203. Sun, S. A survey of multi-view machine learning. *Neural Comput. Appl.* **2013**, *23*, 2031–2038. [CrossRef]
204. Rappoport, N.; Shamir, R. Multi-omic and multi-view clustering algorithms: Review and cancer benchmark. *Nucleic Acids Res.* **2019**, *47*, 1044. [CrossRef]
205. Quintanal-Villalonga, Á.; Chan, J.M.; Yu, H.A.; Pe'er, D.; Sawyers, C.L.; Sen, T.; Rudin, C.M. Lineage plasticity in cancer: A shared pathway of therapeutic resistance. *Nat. Rev. Clin. Oncol.* **2020**, *17*, 360–371. [CrossRef]
206. Barta, J.A.; Powell, C.A.; Wisnivesky, J.P. Global epidemiology of lung cancer. *Ann. Glob. Health* **2019**, *85*. [CrossRef] [PubMed]
207. Love, M.I.; Huber, W.; Anders, S. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biol.* **2014**, *15*, 550. [CrossRef]
208. John, C.R.; Watson, D.; Russ, D.; Goldmann, K.; Ehrenstein, M.; Pitzalis, C.; Lewis, M.; Barnes, M. M3c: Monte carlo reference-based consensus clustering. *Sci. Rep.* **2020**, *10*, 1816. [CrossRef]
209. Sondka, Z.; Bamford, S.; Cole, C.G.; Ward, S.A.; Dunham, I.; Forbes, S.A. The cosmic cancer gene census: Describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **2018**, *18*, 696–705. [CrossRef]
210. Frost, H.R.; Li, Z.; Moore, J.H. Principal component gene set enrichment (pcgse). *Biodata Min.* **2015**, *8*, 25. [CrossRef] [PubMed]
211. Fabregat, A.; Jupe, S.; Matthews, L.; Sidiropoulos, K.; Gillespie, M.; Garapati, P.; Haw, R.; Jassal, B.; Korninger, F.; May, B.; et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* **2017**, *46*, D649–D655. [CrossRef] [PubMed]
212. Liberzon, A.; Subramanian, A.; Pinchback, R.; Thorvaldsdóttir, H.; Tamayo, P.; Mesirov, J.P. Molecular signatures database (msigdb) 3.0. *Bioinformatics* **2011**, *27*, 1739–1740. [CrossRef] [PubMed]
213. Gu, Z.; Eils, R.; Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **2016**, *32*, 2847–2849. [CrossRef]
214. Gaidano, G.; Rossi, D. The mutational landscape of chronic lymphocytic leukemia and its impact on prognosis and treatment. *Hematol. Am. Soc. Hematol. Educ. Program.* **2017**, *2017*, 329–337. [CrossRef]
215. Holroyd, A.K.; Michie, A.M. The role of mtor-mediated signaling in the regulation of cellular migration. *Immunol. Lett.* **2018**, *196*, 74–79. [CrossRef] [PubMed]
216. Ondrisova, L.; Mraz, M. Genetic and non-genetic mechanisms of resistance to bcr signaling inhibitors in b cell malignancies. *Front. Oncol.* **2020**, *10*. [CrossRef] [PubMed]
217. Dietrich, S.; Oleś, M.; Lu, J.; Sellner, L.; Anders, S.; Velten, B.; Wu, B.; Hüllein, J.; da Silva Liberio, M.; Walther, T.; et al. Drug-perturbation-based stratification of blood cancer. *J. Clin. Invest.* **2018**, *128*, 427–445. [CrossRef] [PubMed]