cells

MDPI

Review

Inferencing Bulk Tumor and Single-Cell Multi-Omics Regulatory Networks for Discovery of Biomarkers and Therapeutic Targets

Qing Ye [1,2] and Nancy Lan Guo [1,3,*]

[1] West Virginia University Cancer Institute, Morgantown, WV 26506, USA
[2] Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV 26506, USA
[3] Department of Occupational and Environmental Health Sciences, School of Public Health, West Virginia University, Morgantown, WV 26506, USA
* Correspondence: lguo@hsc.wvu.edu; Tel.: +1-304-293-6455

Abstract: There are insufficient accurate biomarkers and effective therapeutic targets in current cancer treatment. Multi-omics regulatory networks in patient bulk tumors and single cells can shed light on molecular disease mechanisms. Integration of multi-omics data with large-scale patient electronic medical records (EMRs) can lead to the discovery of biomarkers and therapeutic targets. In this review, multi-omics data harmonization methods were introduced, and common approaches to molecular network inference were summarized. Our Prediction Logic Boolean Implication Networks (PLBINs) have advantages over other methods in constructing genome-scale multi-omics networks in bulk tumors and single cells in terms of computational efficiency, scalability, and accuracy. Based on the constructed multi-modal regulatory networks, graph theory network centrality metrics can be used in the prioritization of candidates for discovering biomarkers and therapeutic targets. Our approach to integrating multi-omics profiles in a patient cohort with large-scale patient EMRs such as the SEER-Medicare cancer registry combined with extensive external validation can identify potential biomarkers applicable in large patient populations. These methodologies form a conceptually innovative framework to analyze various available information from research laboratories and healthcare systems, accelerating the discovery of biomarkers and therapeutic targets to ultimately improve cancer patient survival outcomes.

Keywords: biomarkers; therapeutic targets; multi-omics regulatory networks; single cells; Prediction Logic Boolean Implication Networks (PLBINs); network centrality; electronic medical records (EMRs); SEER-Medicare

Citation: Ye, Q.; Guo, N.L. Inferencing Bulk Tumor and Single-Cell Multi-Omics Regulatory Networks for Discovery of Biomarkers and Therapeutic Targets. *Cells* **2023**, *12*, 101. https://doi.org/10.3390/cells12010101

Academic Editors: Huihui Fan, Fulong Yu and Desi Shang

Received: 6 December 2022
Revised: 22 December 2022
Accepted: 24 December 2022
Published: 26 December 2022

1. Introduction

Despite decades of efforts in cancer research, cancer ranks as the top cause of death and shortened life expectancy in every country in the world [1]. In 2040, the global cancer burden is estimated to increase by 47% from 2020, reaching 28.4 million cases [1]. The Cancer Moonshot project was launched in 2016 to accelerate scientific discovery, foster collaboration, and improve data sharing in cancer research [2]. The current unmet clinical needs in cancer treatment include a lack of biomarkers for precise assessment of cancer risk, tumor progression, recurrence, and treatment response in individual patients. More effective therapeutic targets are needed to improve patient survival outcomes.

The advent of high-throughput sequencing technology has led to the discovery of abnormal genomic variants in cancer patients as novel therapeutic targets, such as the EML4-ALK fusion gene in non-small-cell lung cancer (NSCLC) [3]. In addition, the blockade of immune checkpoint inhibitors (ICIs), including PD1, PDL1, and CTLA4, has improved cancer patient survival outcomes [4–10]. However, there are currently no established predictive biomarkers in immunotherapy, as PDL1 and tumor mutational burdens are not proven indicators [11]. Systematic disease mechanisms underlying cancer remain illusive.

The tumor immune microenvironment is a multidimensional system of immune cells, stromal cells, and host factors. Complex and interweaving signaling pathways and networks of genes and proteins function in these various cell types [12]. In the tumor microenvironment, the presence of tertiary lymphoid structures (TLSs) is linked to good cancer prognosis [13]. TLSs comprise B cells and adjacent clusters of dendritic cells and T cells [14]. TLSs and tumor-infiltrating B cells improve ICIs responses in cancer immunotherapy [15–18]. Recent studies suggest an essential role of B cells in antitumor immunity, including the determination of protective T cell responses in cancer patients [19–21]. T-cell dysfunction and therapy have been established for cancer treatment [22–28]. However, B-cell biology and therapeutic potential have not been substantiated [29–31]. The emerging single-cell sequencing technique is an effective method to better understand disease mechanisms and develop novel therapeutic interventions.

Genes and proteins form complex gene regulatory networks (GRN) in living organisms [32,33]. Perturbed gene regulation is closely related to disease and its revelation is important for developing intervention strategies [34–37]. Molecular network analysis is crucial to decipher cancer mechanisms and advance precision oncology [38]. Artificial intelligence (AI) methods are needed to reveal essential GRNs and essential hub genes at multiple regulatory levels by analyzing emerging multi-modal data in patient bulk tumors and single cells for the discovery of biomarkers and therapeutic targets.

To achieve optimal treatment selection in individual patients, it is essential to integrate patient multi-omic biomarkers with clinical, pathological, demographic, and comorbid factors using electronic medical records (EMRs) [39]. Retrospective analysis of EMRs has led to the discovery of new and repositioning drugs [40–42].

This review is focused on multi-omics data processing and integration in Section 2, common systems biology software and data resources in Section 3, molecular network inference methods in Section 4, hub genes in tumorigenesis, proliferation, and patient survival in Section 5, and integration of multi-omics data with EMRs in Section 6. Finally, we provide recommendations for bulk tumor and single-cell multi-omics network analysis for the discovery of biomarkers and therapeutic targets in Section 7.

2. Bulk Tumor and Single-Cell Multi-Omics Data Analysis

2.1. Multi-Omics Data Processing and Integration

With the rapid development of high-throughput technology, genomic, transcriptomic, proteomic, and metabolomic profiles have provided ample sources of information for researchers to understand molecular disease mechanisms. Nevertheless, data generated from various commercially available platforms and customized arrays pose tremendous challenges for processing, analysis, and integration. The Genome Analysis Toolkit (GATK) is the industry standard for processing multi-omics data in bulk tumors and single cells, including identifying single nucleotides (SNPs) and indels, somatic short variants, copy number variations (CNV), and structural variations (SV) in germline DNA and RNAseq data [43]. In addition to data generated from current sequencing technology, a huge amount of high-throughput data was generated from legacy DNA microarrays. A research group from the FDA reported that biomarkers and predictive models derived from legacy microarray data can accurately predict phenotypes in samples profiled with RNA sequencing, whereas RNA-seq-based models are less accurate in predicting microarray data [44]. This section provides a brief overview of some software packages and methods used for bulk tumor multi-omics data processing and integration.

2.1.1. Copy Number Variation

Copy number variation (CNV) is a structural variation that is either a duplication or deletion event affecting a large number of base pairs. Deletions, amplifications, gains, and losses collectively termed CNVs, are found in all humans and other mammals [45]. The number of CNVs can make up as much as 5–15% of the human genome [46]. CNVs are a significant source of genomic diversity and driver of somatic and hereditary human

diseases including cancer. However, compared to single-nucleotide variations (SNVs), CNVs are still under-investigated, despite their evolutionary significance and clinical relevance. This is a consequence of the inherent challenges in identifying CNVs in diverse populations of cells at low-to-intermediate frequencies [47]. Using a recent method of a fluorescent gene functioning as a single-cell CNV reporter, CNVs are found to occur frequently and undergo selection with predictable dynamics across independently evolving replicate populations [47]. CNVs have been applied in the molecular diagnosis of many diseases and non-invasive prenatal care. Nevertheless, CNVs have not reached their full potential as emerging biomarkers. Cancer immunotherapy targets, including *PD1*, *PDL1*, *CD27*, and *CD20* have more CNVs than SNVs in NSCLC tumors in The Cancer Genome Atlas (TCGA) [48]. Tumor mutation burdens are used in cancer management, but not CNVs. The screening, diagnosis, prognosis, and monitoring of several illnesses, including cancer and cardiovascular disease, are likely to be significantly impacted by CNVs [49].

Genomic alterations in DNA might interfere with the normal function of the genes. The genomic instability and structural dynamics of cancer cells require that CNV data be examined to discover the underlying associations between CNVs, gene/protein expression, and functional aberrations. Different platforms were used to profile genome-scale CNVs, including high-resolution SNP arrays (GeneChip Mapping 250K-Nsp array, Affymetrix), whole-genome tiling path aCGH (BCCRC whole genome tilling path array v2), and whole exome sequencing (SOLiD 5500xl) [50]. Various CNV data processing methods were developed as described below.

PennCNV-Affy [51], a Perl/C-based software tool, is the most commonly used method for CNV calling for data produced with SNP genotyping arrays. The first step is to process the raw CEL files and generate the signal intensity data. The second step is to split the signal file generated from step 1 into individual files. After the file splitting is completed, CNV calls will be generated by PennCNV. The output provides information on the CN state for each SNP probe. Normally, a CN < 2 indicates a deletion in copy number, and a CN > 2 indicates a duplication. For the SNP probes located within the same gene, the probe with the maximum intensity is used to represent the CN state for the gene.

Bioconductor packages CGHbase [52] and CGHcall [53] are often used to call the CNV in the aCGH data. The $\log_2$ normalized ratios of Cy3/Cy5 are used as inputs. In CGHcall, the number of output classes can be selected among 3 classes (loss, normal, gain), 4 classes (loss, normal, gain, amplification), or 5 classes (double deletion, loss, normal, gain, amplification).

GISTIC2.0 is a pipeline used to find genes targeted by somatic copy-number alterations (SCNAs) in human cancers [54]. GISTIC2.0 uses an iterative optimization algorithm to deconstruct each segmented copy-number profile into its most likely set of SCNAs. Compared with other methods, GISTIC2.0 is advantageous in separating arm-level and focal SCNAs explicitly by length.

CNV data generated by various platforms provide the corresponding chromosome location of each SNP. To harmonize the CNV data from various platforms, we can convert the genome assembly version from earlier versions, such as hg17, to hg38 by using the Python package *CruzDB*, a fast and intuitive tool for the UCSC genome browser [55]. Using the latest reference genome is an important step to ensure compatibility in the CNV data integration.

2.1.2. Categorization of Gene Regulation

Cancer is caused by dysregulated tumor suppressor genes or oncogenes. Due to genetic mutations or alterations in gene regulation, such genes are switched on or off and are expressed at abnormally high or low levels in tumor initiation and progression. It is important to define the up-regulation, normal, and down-regulation ranges by categorizing the gene expression data generated from high-throughput microarray or RNA sequencing. Housekeeping genes are generally used to categorize gene expression data.

Housekeeping genes are essential for the existence of the cell, regardless of their specific role in the tissue or organism. Housekeeping genes are expressed in all cells of an organism regardless of conditions (normal or pathophysiological), tissue type, developmental stage, cell cycle status, or external signals. Unlike in qRT-PCR, housekeeping genes are not generally used for normalization in RNA sequencing analysis. Therefore, the variation in gene expression measurements due to different sample preparation techniques is not accounted for in the RNA expression analysis. A set of stably expressed housekeeping genes in particular tissue types should be used for the corresponding research. For instance, a set of housekeeping genes were used for NSCLC [56–60], including *ACTB*, *B2M*, *CDKN1B*, *ESD*, *FLOT2*, *GAPDH*, *GRB2*, *GUSB*, *HMBS*, *HPRT1*, *HSP90AB1*, *IPO8*, *LDHA*, *NONO*, *PGK1*, *POLR2A*, *PPIA*, *PPIH*, *PPP1CA*, *RHOA*, *RPL13A*, *SDCBP*, *TBP*, *TFRC*, *UBC*, *YAP1*, and *YWHAZ* to define the threshold of gene expression level in multi-omics regulatory network studies [48,61]. Specifically, the total percentage of up-regulated and down-regulated samples was fixed for all the housekeeping genes to be 30%, and the average standard deviation of the normal range for the selected housekeeping genes was calculated. This average standard deviation was applied to all other genes in the genome to define their normal, up-regulation, or down-regulation ranges [48,61]. "Half SAM score" is recommended for differential gene expression analysis of data generated from microarrays and next-generation sequencing (NGS) [62]. DEseq2 is commonly used for fold change and differential gene expression analysis of NGS data [63].

2.1.3. Categorization of Protein Regulation

Protein expression represents how proteins are synthesized, modified, and regulated in an organism. The synthesis and regulation of proteins depend on the functional requirements in the cell. The blueprint for proteins is stored in DNA and decoded by a highly regulated transcriptional process that produces messenger RNA (mRNA). The information encoded by mRNA is subsequently translated into proteins as functional units of biological processes. Protein expression data generated from AQUA [56] and Nano-LC-MS/MS [64] are often log-transformed for differential expression analysis and Cox survival modeling.

The up-regulation, normal, and down-regulation ranges of protein expression also need to be defined, similar to gene expression. In a regulatory network analysis of NSCLC tumors [64], the categorization of protein regulation was performed by using the normal range defined with NSCLC housekeeping genes [56–60], including *B2M*, *ESD*, *FLOT2*, *GAPDH*, *GRB2*, *HPRT1*, *HSP90AB1*, *LDHA*, *NONO*, *POLR2A*, *PPP1CA*, *RHOA*, *SDCBP*, and *TFRC*, based on their protein expression in NSCLC tumors and non-cancerous adjacent tissues in Xu's cohort [65]. The total percentage of up-regulated and down-regulated samples was fixed for all the housekeeping proteins, and the average standard deviation of the normal range for the selected housekeeping proteins was calculated and applied to all other proteins in the genome to define their normal, up-regulation, or down-regulation ranges [64].

2.2. Single-Cell Muti-Omics Data Processing

Each cell type has its distinct function. The single-cell analysis allows the study within a cell population to reveal how cell networks function [66,67]. Ginkgo [68] is an open-source web-based platform for single-cell CNV analysis. Single-cell transcriptomics simultaneously measures the gene expression level of individual cells in a given population [69]. Single-cell whole-genome analyses by Linear Amplification via Transposon Insertion (LIANTI) can generate sufficient DNA for next-generation sequencing [70]. In processing the single-cell gene expression data from Illumina HiSeq 2000, gene features are counted with the *featureCounts* method [71] based on the Gencode v19 transcriptome annotation. In processing the data from Illumina HiSeq 4000, the reads are mapped with *STAR* aligner [72] based on human genome reference GRCh38, and SAMtools [73] is used to sort and index the mapped reads.

The dropout phenomenon, i.e., the RNA in the cell is not detected due to limitations of current experimental protocols, is severe in single-cell transcriptomic data. As a result, a large number of genes are expressed with a value of 0 in many cells. This makes it difficult to classify single-cell gene expression as in bulk tumors, and the housekeeping gene technique described above cannot achieve usable results. Thus, single-cell gene expression data is generally classified into two categories, "not expressed" for genes with a feature count of 0, and "expressed" for genes with a future count greater than 0 in regulatory networks [74]. DEsingle [75] in Bioconductor is a common method for single-cell differential expression analysis.

3. Common Systems Biology Software and Data Resources

3.1. Pathway Analysis

Molecular pathway analysis is important to translate multi-omics analysis to drug discovery [76]. Once a list of genes is identified from a study, gene set enrichment analysis can be performed to examine the relevant biological processes and canonical pathways. Enrichment is the process of classifying genes according to a priori knowledge. The following tools are used for pathway analysis.

GSEA is an online tool to evaluate the over-representation of a gene list in a comprehensive database MSigDB [77]. The input to GSEA is a gene expression matrix in which the samples are divided into two sets. All genes are first sorted from largest to smallest based on the processed differential expressions, which are used to represent the trend of gene expression changes between the two sets. GSEA analyzes whether all genes in a gene set are enriched at the top or bottom of a ranked list for a biological process. If they are enriched at the top, the gene set is considered overall up-regulated in this biological process. Conversely, if they are enriched at the bottom, this gene set is considered overall down-regulated in this biological process.

ToppFun in ToppGene Suite is a one-stop portal for enrichment analysis and candidate gene prioritization based on functional annotations and protein interaction networks [78]. ToppFun provides enrichment analysis of pathways, gene families, cytobands, drugs, diseases, etc. The input to ToppFun is a list of genes. The outputs include significant functional enrichment results with information such as *p*-values, FDRs, etc.

Qiagen Ingenuity Pathways Analysis (IPA) is an online pathway analysis tool incorporating curated molecular interactions and their involvement in diseases with confirmed information retrieved from scholarly publications. Using these data, it is possible to map interactions among a list of genes in various pathological conditions, such as cancer and immunological diseases.

Adviata iPathwayGuide computes the over-representation of an input gene list in a pathway or disease using Fisher's method. Multiple hypothesis testing is applied using FDR or Bonferroni corrections. The enrichment analysis utilizes pathways and diseases from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [79,80], gene ontologies (GO) from the Gene Ontology Consortium database [81], and miRNA-mRNA target pairs from the miRBase and MICROCOSM databases [82]. Experimentally confirmed microRNA targets can be retrieved from TarBase [83].

3.2. Proliferation Assays

Cancer cells have high rates of cell division and growth, and are very prolific. The DepMap portal provides genome-scale CRISPR-Cas9/RNA interference (RNAi) screening data in Cancer Cell Line Encylopedia (CCLE). The dependency scores in CRISPR-Cas9 [84] knockout and RNAi [85] knockdown screening data measure a gene's impact on proliferation. Essential genes significantly impact the cellular growth in a cell line in knockout/knockdown assays; otherwise, they are defined as nonessential. Gene knockout/knockdown effects, represented with dependency scores, are normalized such that the median dependency score of the non-essential genes is 0, and the median dependency score of the essential genes is –1 in each cell line. Negative dependency scores indicate the cancer

cell line growth is highly dependent on the gene; positive dependency scores indicate the cell line grows faster after the gene is knocked out or knocked down. A normalized dependency score less than –0.5 is considered a significant effect in CRISPR-Cas9 knockout or RNAi knockdown.

The current single-cell technologies, including single-cell sequencing and CRISPR-Cas9/RNAi screening, have not been widely adopted. Recent studies explored editing immune cells using CRISPR-Cas9 [86–89]. Nevertheless, there is a lack of single-cell genome-scale CRISPR-Cas9/RNAi screening data for broad research and clinical applications.

3.3. Stromal and Immune Infiltration and Cell Activity

The extracellular matrix, soluble chemicals, and tumor stromal cells constitute the tumor microenvironment. The formation of the tumor microenvironment will result in the chemotaxis of numerous immune cells (e.g., macrophages, T cells, etc.) that form part of the tumor microenvironment. In the tumor microenvironment, immune cells and stromal cells are the two main non-tumor components, which are of great potential for cancer diagnosis and prognosis assessment.

The Estimation of STromal and Immune cells in MAlignant Tumours (ESTIMATE) [90] predicts tumor purity and infers the stromal and immune infiltration in tumor tissues. The function *estimateScore* of the *ESTIMATE* package in R computes the stromal score and immune score in each sample using transcriptomic data.

The xCell tool [91] predicts the levels of 64 immune and stroma cell types based on gene expression data. The xCell scores for patient samples can be calculated using single-sample gene set enrichment analysis (ssGSEA) to analyze the immune microenvironment. Low xCell scores indicate the cell type has similar levels across all samples; whereas high xCell scores indicate the cell type has different levels across all samples.

TIMER 2.0 [92–94] and CIBERSORTx [95] are comprehensive resources for systematically analyzing the immune infiltration in tumors. They provide the abundance of immune infiltration estimated by a variety of immune deconvolution methods. TIMER 2.0 [92–94] and CIBRSORTx [95] compute the association of gene expression and immune infiltration in multiple cell types including myeloid dendritic cells, macrophages, neutrophils, CD4+ T cells, CD8+ T cells, B cells, etc. using a variety of immune deconvolution methods. Microenvironment Cell Populations-counter (MCP-counter) [96] quantifies the absolute abundance of eight immune and two stromal cell populations in heterogeneous tissues using transcriptomic data. MCP-counter estimates immune infiltrates across healthy tissues and non-hematopoietic tumors in human samples.

3.4. Drug Discovery and Repurposing

LINCS L1000 Connectivity Map (CMap) [26,27] provides an online tool to identify functional pathways and drugs based on gene expression signatures of up-regulated or down-regulated genes. CMap incorporates over 1.5M transcriptomic profiles from the treatment of ~5000 small molecules and ~3000 genetic reagents in multiple cell types. A hypothesis is considered valid for further investigation with a p-value < 0.05 and a connectivity score > 0.9. The selected compounds can be further analyzed with the drug screening data to discover potential repositioning drugs [48,61,74].

Drug screening data from PRISM [97] and GDSC1/2 [98–100] datasets contain drug activity data in CCLE. Multiple doses were tested for each drug. Cell lines are considered resistant to a drug if the IC_{50} or EC_{50} values are higher than the maximum dose; cell lines are considered sensitive to a drug if the IC_{50} or EC_{50} values are lower than the minimum dose. Using the mean $\pm$ 0.5 standard deviations of the drug sensitivity measurements, the remaining cell lines can be categorized into three groups, including sensitive, partial response, or resistant [101,102]. This in vitro drug sensitivity categorization is corresponding to RECIST 1.1 (i.e., complete response, partial response, and stable disease/disease progression) in evaluating therapeutic responses in patient solid tumors [103].

are considered sensitive to a drug if the IC_{50} or EC_{50} values are lower than the minimum dose. Using the mean ± 0.5 standard deviations of the drug sensitivity measurements, the remaining cell lines can be categorized into three groups, including sensitive, partial response, or resistant [101,102]. This in vitro drug sensitivity categorization is corresponding to RECIST 1.1 (i.e., complete response, partial response, and stable disease/disease progression) in evaluating therapeutic responses in patient solid tumors [103].

4. Common Approaches to Molecular Network Inference

Molecular networks have been widely used to understand multicellular functions in disease [104] and elucidate drug response from modulators to targets [105]. Artificial intelligence/machine learning methods are needed to construct multi-omics genome-scale networks. This section reviewed common approaches for network inference in terms of computational efficiency, scalability, and accuracy.

4.1. Relevance Networks

Relevance Networks mainly construct gene regulatory network (GRN) models by calculating the associations between genes. This method considers that genes with similar expression profiles may interact with each other and therefore may have similar functions [106]. If the expression value of gene A is increased and the expression value of gene B is simultaneously increased or decreased, the relationship between the two genes can be detected and modeled. The regulatory relationship can also be inferred by the transcriptional dependence between them. The main idea of the correlation detection method is that for a predetermined threshold if the association between genes is higher than the threshold, the genes will be connected by edges in the network. Two genes are more related if they have the same or similar regulatory mechanisms, especially for target genes of the same transcription factor or genes on the same biological pathway. The relevance between genes can be inferred with the following metrics.

4.1.1. Pearson Correlation Coefficient (PCC)

PCC [33] is a linear correlation coefficient, which reflects the degree of linear correlation between two variables. Let X and Y be two random variables, $PCC(X, Y)$ is defined as:

$$PCC(X, Y) = \frac{\sum_i (X_i - \overline{X_i})(Y_i - \overline{Y_i})}{\sqrt{\sum_i (X_i - \overline{X_i})^2 \sum_i (Y_i - \overline{Y_i})^2}} \quad (1)$$

$$PCC(X,Y) = \frac{\sum_i (X_i - \overline{X_i})(Y_i - \overline{Y_i})}{\sqrt{\sum_i (X_i - \overline{X_i})^2} \times \sqrt{\sum_i (Y_i - \overline{Y_i})^2}} \quad (1)$$

where $\overline{X_i}$, $\overline{Y_i}$ are the mean values of X and Y, respectively. $PCC(X, Y)$ takes values between −1 and 1. When $PCC(X, Y)$ is −1 or 1, it means that the two variables are completely correlated; when $PCC(X, Y)$ is 0, the two variables are linearly uncorrelated. Figure 1 shows a simple example of constructing a network model using PCC.

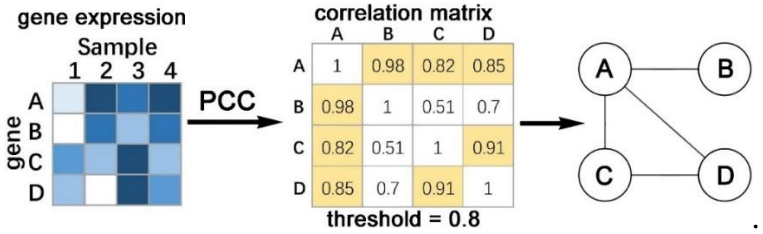


Figure 1. Constructing a relevance network using Pearson Correlation Coefficient (PCC).

Weighted gene correlation network analysis (WGCNA) is a typical method for constructing gene co-expression regulatory networks with PCC [107], where genes are first divided into clusters using hierarchical clustering, and highly co-expressed genes in each cluster are connected by correlation values. Genomic networks are established after the interrelationships of every pair of genes have been determined. Various correlation networks have been implemented for multi-omics analysis. MiBiOmics [108] implements WGCNA in R as a Shiny app for multi-omics network analysis and visualization. OmicsAnalyst [109] system models correlation networks and is hosted on Google Cloud. CorDiffViz [110] is an R package to construct and visualize multi-omics differential correlation networks. In addition to Pearson's correlation, CorDiffViz utilizes rank-based correlation metrics coping with non-Gaussian observations commonly present in omics data for more robust inferences of differential correlations. The outputs are automatically saved to a local directory

by calling a single R function *viz()* with some specified parameters. The users can then visualize the results by opening the HTML file in a browser.

Since PCC only needs to calculate the similarity of expression profiles between genes, it has the advantage of low computational time and space complexity. It is therefore able to cope with large-scale data and can be applied to both continuous and discrete data, but not categorical data. Since PCC can only measure the linear relationship between nodes, it is only capable of analyzing genes with similar expression profiles. Moreover, PCC is vulnerable to noise and random perturbations, which makes it inaccurate and less robust.

4.1.2. Gaussian Graphical Models (GGM)

GGM is an undirected probabilistic graphical model that assumes gene expression data follow a multidimensional normal distribution. A partial correlation coefficient matrix between genes is first calculated, and then the edges of the network are selected by testing whether each element of the partial correlation coefficient matrix is significantly different from zero.

For a network containing n genes with expression levels denoted as $x_1, x_2, \ldots, x_n$, assuming the genes are joint normally distributed, the partial correlation coefficient is:

$$\rho_{ij} = Corr\left(x_i, x_j \middle| x_{-(i,j)}\right) \tag{2}$$

where $x_{-(i,j)} = \{x_k | 1 \leq k \neq i, j \leq n\}$. $\rho_{ij} \neq 0$ means the two genes are conditionally dependent so that there is an edge between them. The partial correlation coefficient can be expressed as the inverse of the covariance matrix as follows:

$$\rho_{ij} = -\frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}} \tag{3}$$

where σ_{ij} is the element of the inverse covariance matrix. However, if the number of genes n is large but the sample size is small, the covariance matrix cannot be obtained. To avoid the calculation of the covariance matrix, some methods were proposed based on low-order partial correlation analysis [111–113].

The advantage of GGM is that it can eliminate a lot of indirect connections between genes to facilitate further analysis. The disadvantages are (1) the edges of the network it constructs are undirected and cannot infer causality, and (2) the static model cannot reflect the dynamic behaviors in GRNs.

4.1.3. Mutual Information (MI)

To improve the limitations of correlation coefficients in association-based methods, information theory-based methods have been proposed for the construction of GRNs. Mutual information (MI) [114] is usually used to describe the statistical correlation between two systems or to reflect the amount of information embedded in one system about the other system using entropy [115,116]. According to the definition of entropy in information theory, the mutual regulatory information between genes can be analyzed from an information theory point of view, and the gene expression information can be quantified by using the Shannon evaluation of information entropy [117]. The entropy of a gene expression pattern is a measure of the information contained, and the model describes the association of genes in terms of entropy and mutual information.

The entropy of a gene expression pattern X is a measure of the amount of information it contains and is calculated as:

$$H(X) = -\sum_{x} p(x) \log_2 p(x) \tag{4}$$

where $p(x)$ is the probability of $X = x$. The larger the entropy value, the more the gene expression level tends to be randomly distributed. Let $p(x, y)$ be the joint probability when $X = x$ and $Y = y$, then the joint entropy of X and Y is defined as:

$$H(X, Y) = -\sum_{x}\sum_{y} p(x, y) \log_2 p(x. y) \quad (5)$$

From this, we can obtain the MI of the random variables X and Y as:

$$MI(X, Y) = H(X) + H(Y) - H(X, Y) \quad (6)$$

A high MI indicates a close relationship between two genes and a low MI indicates their independence [118]. To construct a gene association network, the MI is used to (1) calculate the degree of association between all gene pairs, (2) define the existence of associations between gene pairs that pass the pre-set threshold, and (3) to connect these gene pairs with edges [119].

MI is a widely recognized metric for quantifying statistical association [120]. The most important advantage of MI is its ability to infer nonlinear relationships between genes accurately and efficiently [121–123]. Secondly, MI can handle large-scale data with a limited sample size [124–126]. The disadvantage is that MI may overestimate the interaction relationships between genes and the constructed networks tend to contain many false-positive edges. To reduce the false positive edges, the influence of other genes can be analyzed and eliminated when calculating the association degree of two genes, i.e., Conditional Mutual Information (CMI). CMI was introduced to delete these false positive edges [118,127]. However, CMI appears to underestimate the regulatory relationships between genes in some cases, increasing false negative network edges. To address these problems, Zhang et al. [128] proposed the CMI2NI algorithm, which reduces this error by introducing the concept of relative entropy by calculating the Kullback-Leribler divergence.

The association network model can only obtain whether two genes are associated or not, but cannot infer the specific regulatory relationship. To distinguish direct and indirect effects, various optimization methods based on information theory have been proposed. The Algorithm for the Reconstruction of Accurate Cellular Networks (ARCANE) by Margolin et al. [121] employs an information theory-based approach to constructing association networks by using the Data Processing Inequality (DPI) constraint. If the data processing imbalance exceeds a certain threshold, ARCANE evaluates all possible gene triplets and prunes the least significant edge in each triplet with the smallest MI among the corresponding genes. ARACNE is a relatively conservative network construction method that retains the majority of edges inferred from the network. The Context Likelihood of Relatedness (CLR) algorithm [111] proposes an adaptive background correction step to remove erroneous correlations. CLR estimates paired MI values for all gene pairs and then converts the MI values into z-scores for comparison with the sample distribution to estimate the statistical possibility of the specific gene pair. Maximum relevance/minimum redundancy (MRMR) used by MRNET [129] can infer gene interactions. The MRMR algorithm is used to choose the ideal subset of regulators, which initially treats one gene as the target gene and the rest genes as its potential regulators. C3NET [130] retains only the core causality of the network, i.e., only the MI of the gene pair that is higher than the MI with any other gene in the genome for both genes, and then the connection between these pair of genes will be established.

MI constructs undirected networks, and most applications require known gene regulation to assume the directionality between genes. Therefore, the wide use of such methods is limited by the available a priori information of the data.

4.2. Bayesian Belief Networks (BBNs)

The Bayesian belief network (BBN) model is a probabilistic graphical model describing the conditional structural independence between random variables and is used to construct

MI constructs undirected networks, and most applications require known gene regulation to assume the directionality between genes. Therefore, the wide use of such methods is limited by the available a priori information of the data.

4.2. Bayesian Belief Networks (BBNs)

The Bayesian belief network (BBN) model is a probabilistic graphical model describing the conditional structural independence between random variables and is used to construct networks by establishing a joint probability distribution of nodes using the Bayesian theory. The concept of the probabilistic graphical model was first proposed and applied to intelligent systems [131–133]. Hartemink et al. [134] proposed Bayesian network-based GRN models around 2000. The Bayesian theory was combined with graph theoretic models to quantitatively model the general properties of gene regulation [135].

Learning the BBN structure from data is all about finding a network that fits best to a given dataset. Suppose $B_1, B_2, \ldots, B_n$ denote random events in the sample space and $P(B_i)$ can be estimated based on previous data analysis or prior knowledge, then $P(B_i)$ is said to be the prior knowledge; the probability of B_i occurring in the case of event A is the posterior, i.e., $P(B_i|A)$. As the sample information keeps changing, the posterior probabilities are constantly updated. The previous posterior probability will be used as the prior probability when obtaining the new posterior probability. This is a process of constant updating and iterative adjustment.

A BBN consists of two parts: the network structure and the conditional probability distribution, which can be defined as a binary group $B = (G, P)$. $G = (\mathbb{N}, E)$ is a Directed Acyclic Graph (DAG) [136]. N is the set of nodes, and each node X_i can be regarded as a variable taking discrete or continuous values. E is the set of directed edges, and each edge represents a direct probabilistic dependency between two nodes. The degree of dependency is determined by the conditional probability. P is a set of conditional probability distributions: $P = \{P(X_i|parent(X_i)) : X_i \in N\}$ [137], where $parent(X_i)$ is the parent nodes of X_i in graph G. The edge pointing from X_i to X_j indicates that X_i is the parent node of X_j. The Markov assumption is implicit in BBNs that the probability of each node is related to its parent node only, i.e., each node is independent of its non-child nodes when the parent nodes are known [138]. Based on this conditional independence property, by applying the chain rule of probability and conditional independence, the joint probability distribution of the specified settings in the Bayesian network G can be expressed in the form of a product:

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|parent(X_i)) \quad (7)$$

The BBN model can describe the state of a gene in both discrete and continuous data, which provides an intuitive and simple way to understand and present GRNs. Figure 2 depicts an example of a discrete BBN. According to Equation (7), the probability that the state of all three (A, B, and C) nodes are 1 is: $P(A = 1, B = 1, C = 1) = P(B = 1) \times P(A = 1|B = 1) \times P(C = 1|A = 1, B = 1) = 0.6 \times 0.7 \times 0.7 = 0.294$.

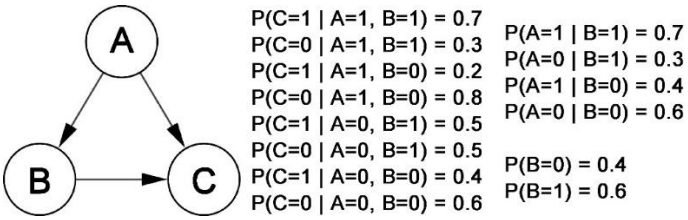


Figure 2. Example of a discrete Bayesian network.

The usage of BBNs requires the computation of conditional probabilities between child nodes and all their possible parent nodes, which grows exponentially in computational time as more variables are incorporated. It has been shown that finding the optimal graph for BBNs is an NP-hard problem [139], which poses a tremendous challenge to constructing complete gene regulatory networks for higher organisms such as humans [140]. To solve this problem, Campos et al. [141] proposed a method based on structural constraints that can reduce the search space by inferring the maximum number of potential parents of a node. Liu et al. [142] designed the Local Bayesian Networks model by (1) first constructing the initial graph with mutual information and conditional mutual information, (2) then

splitting the initial network by the K-nearest neighbors algorithm to reduce the search space, (3) using Bayesian networks to build small the sub-networks, and (4) finally integrating the generated sub-networks.

BBNs can process random data, fuse different types of data and prior information to introduce a suitable network structure, and can handle incomplete noisy data and hidden variable data [143]. The BBN model is highly interpretable and the results are accurate, but it is computationally complex. Hence, it is less capable of handling large-scale data and needs to develop appropriate methods to reduce the search space. When applied to GRNs, BBNs consider the noise of the gene expression data itself as well as the stochastic nature and allow the use of Bayesian theory to incorporate some a priori biological knowledge, such as heterogeneous information [137], in deciding gene relationships.

CBNplot [144] is an R package that uses biological pathway information curated by enrichment analysis to construct and visualize BBNs. The structural inference in CBNplot is based on the bootstrap method of the R package *bnlearn*, which uses preprocessed gene expression data to infer BBNs, and uses eigengene as the expression value for pathway inferences. The results of the CBNplot highlight the interactions between genes and pathways through knowledge mining and visualization. CBNplot can be installed in R through *devtools*. The algorithms in the R package *bnlearn* are implemented with C++ in *BayesNetty* [145,146]. Networks are drawn with the *igraph* R package. TETRAD IV [147] is another implementation of BBNs.

There are major limitations of BBNs. First, BBNs identify regulatory networks as directed acyclic graphs (DAG) that do not include feedback loops. Second, BBNs do not take into account the dynamics of regulatory relationships [133], although feedback loops and dynamics are very important features of regulatory networks.

4.3. Dynamic Bayesian Networks (DBNs)

To capture the dynamic characteristics in GRNs and the information on loop interactions between genes, dynamic Bayesian networks (DBN) were proposed to consider the time-delayed nature of gene regulation and incorporate the dimension of time information in BBNs. The value of a random variable in DBN is determined by the previous time point, and DBN is the transformation process of the random variables at all possible random discrete points [148,149]. The DBN structure is modeled at discrete time points t. Similar to the assumption of the BBN, if X_t is the expression of n genes at time points t, the DBN can be described as:

$$P(X_t|X_{t-1}) = \prod_{i=1}^{n} P(X_{i,\,t}|parent(X_{i,\,t})) \quad (8)$$

where $X_{i,\,t}$ is the expression value of the gene X_i, on time slice t, and $parent(X_{i,\,t})$ is the set of its parent nodes. Figure 3A represents a static BBN and Figure 3B represents a DBN. In the static BBN (Figure 3A), the loop $A \rightarrow B \rightarrow C \rightarrow A$ is not allowed, but this feedback mechanism can be represented in the DBN (Figure 3B).

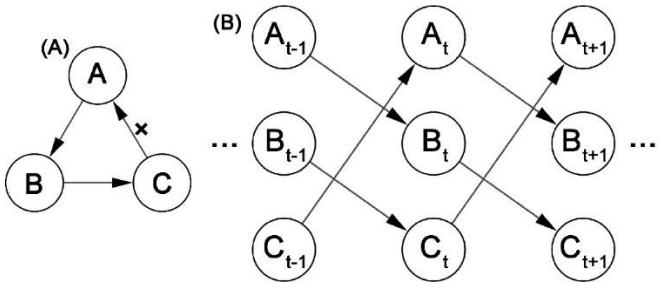


Figure 3. Example of a static Bayesian network (**A**) and a dynamic Bayesian network (**B**).

Smith et al. [150] used the DBN model to analyze microarray data, combining the negative feedback of gene regulation with the time delay factor, so it is necessary to use different nodes in the network to represent the expression of the same gene. Song et al. [151] proposed a new data integration model on DBN by combining a priori knowledge of the relationship between microarray data and genes to construct GRNs using parallel algorithms.

nodes in the network to represent the expression of the same gene. Song et al. [151] proposed a new data integration model on DBN by combining a priori knowledge of the relationship between microarray data and genes to construct GRNs using parallel algorithms.

4.4. Ordinary Differential Equation (ODE) Based Networks

Differential equation models use continuous variables to describe changes in gene expression values as a function of other genes and environmental influences, and it captures the dynamics of GRNs in a quantitative form. The flexibility of differential equation modeling enables the representation of more complex relationships between components. Ordinary Differential Equations (ODEs) are often used to model GRNs. Differential equation models regard the expression level of a gene as a function of time and therefore require the use of time-series data when constructing a GRN. In the process of using ODE to construct GRNs, each differential equation represents the relationship between target genes and their regulatory factors, and the corresponding parameters determine the topology of the network and the interrelationships between genes. The ODE of GRN is expressed as:

$$\frac{dx_i}{dt} = f_i(x),\ 1 \leq i \leq n \tag{9}$$

where x_i represents the expression level of gene X_i. $X_1, \ldots, X_n$ are the n genes that affect gene X_i, and $x = [x_1, \ldots, x_n]^T$ is their expression levels. $\frac{dx_i}{dt}$ represents the rate of change of the expression level of gene X_i at moment t in the GRN modeling. $f_i(x)$ illustrates the mode of action and the regulatory mechanism between genes, i.e., the structure of the regulatory network. The function $f_i(x)$ can be linear, segmented linear, pseudo linear, or continuous nonlinear functions. $f_i(x)$ in its simplest form is a linear function and can be expressed as:

$$\frac{dx_i}{dt} = \sum_j \omega_{ij} x_j + b_i,\ 1 \leq i \leq n \tag{10}$$

The relationship between the genes in the regulatory network can be expressed by the parameter ω_{ij}, for which the activation, repression, and no-regulation relationships take values of positive, negative, and 0, respectively. b_i denotes the basal activity of the gene X_i. Linear differential problems can be solved using singular value decomposition, least squares regression, or likelihood-based approaches [152,153].

However, the regulatory relationships in cells are not simply linear [154] and can be inscribed using nonlinear regulatory functions $f_i(x)$. The disadvantage of nonlinear functions is the computational difficulty and the high computational cost of finding the solution to the differential equation. Moreover, the number of samples is usually too small compared to the number of genes, resulting in a non-singular matrix that will have multiple solutions satisfying the differential equation, which in turn requires the selection of reasonable model parameters from multiple solutions. Therefore, the search space of the nonlinear model structure needs to be strictly limited. Sakamoto et al. [155] used genetic programming to identify small-scale networks by fitting a polynomial function f; Spieth et al. [156] used different search mechanisms such as evolutionary algorithms to infer small networks.

The advantages of ODE modeling are: (1) it is powerful and flexible; (2) it facilitates the description of complex relationships in GRNs; and (3) it is especially suitable for genes with periodic expression. ODE models are mathematically well expressed and have great potential in the analysis of local GRNs. In addition, ODEs can be used to study the effects on gene expression levels by changing environmental variables, introducing new variables, etc., and comparing the changes in the weight matrix before and afterward.

The disadvantages of differential equations are: (1) the parameters in the model are difficult to estimate, and (2) it is hard to obtain a globally optimal solution. In large networks, the ODE model is limited by sample size requirements, lacks robustness to noisy data, and does not capture the stochastic information contained in gene expression data

very well. In the absence of constraints, the ODE system will have an infinite number of solutions. Therefore, to determine the appropriate ODE model and to perform accurate parameter estimation, a thorough study of the nature of the f function and the definition of reasonable constraints based on prior knowledge are required.

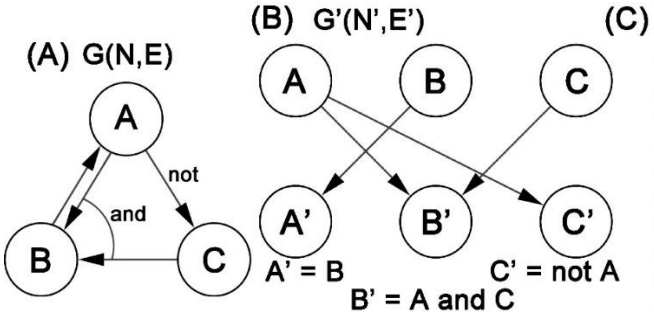
4.5. Boolean Networks

The Boolean network model, first introduced by Kauffman in 1969 [157], is a dynamic discrete model in which the network nodes have synchronous relationships [158]. The Boolean network model is one of the simplest models to reveal GRNs, which treats genes as logical elements [159]. Individual genes can be represented by Boolean variables regardless of whether they are expressed or not. The Boolean network model abstracts the expression level of a gene by clustering or threshold discretization into two states: on and off, the state "on" indicating that the gene is expressed (or overexpressed state) and the state "off" indicating that the gene is not expressed (or low expression state). The interactions in Boolean networks between genes must follow Boolean rules. A Boolean network contains n nodes (representing genes in GRNs) in the repressed or expressed states (i.e., 0 or 1), which are connected by the logical operators "*and*", "*or*", and "*not*" [160]. The expression level of a given gene is obtained by a Boolean function on the expression levels of multiple genes associated with that gene, and the states of all genes are updated using a synchronous update mechanism. The challenge of Boolean network construction for GRN lies in finding the appropriate Boolean function for each gene so that the model can accurately interpret the observed data.

A Boolean network is a directed graph, denoted by $G(N, E)$, E is the set of directed edges, where each node $Xi \in N$ is determined by a function. The next state at $t + 1$ of the network can be represented by all inputs and the functions of the nodes at a time point t:

$$X_i(t+1) = B_i[X_1(t), \ldots, X_n(t)] \quad (11)$$

$X_i(t)$ represents the expression level of gene i at moment t. The function B_i represents the Boolean function of the whole network for gene i. The interaction relationship between genes is represented by Boolean functions, and Boolean rules are expressed in the form of truth tables. Figure 4A depicts a simple Boolean network $G(N, E)$, and Figure 4B represents the state transition corresponding to the linkage graph $G'(N', E')$ of Figure 4A. The truth table of this network is shown in Figure 4C.



(C)

t			t+1		
INPUT			OUTPUT		
A	B	C	A'	B'	C'
1	1	1	1	1	0
1	1	0	1	0	0
1	0	1	0	1	0
1	0	0	0	0	0
0	1	1	1	0	1
0	1	0	1	0	1
0	0	1	0	0	1
0	0	0	0	0	1

Figure 4. Example of a Boolean network. (**A**) A simple Boolean network $G(N, E)$. (**B**) The state transition corresponding to the linkage graph $G'(N', E')$ of A. (**C**) The truth table of this network.

Boolean networks simplify the actual GRNs, providing a framework to describe the complex interactions between genes in GRNs in a biological context. Boolean networks emphasize the underlying global network rather than a quantitative biochemical model. Boolean functions can find possible gene interaction relationships, which can be used as a basis for modeling real gene regulatory networks.

The disadvantage of Boolean networks is their imprecision. Boolean networks can only be represented as a crude qualitative model that portrays the interactions between genes by combining fixed logical rules. It is difficult to accurately describe the real GRN enumerating all possible logical operations, so the Boolean network can only be used as the basis for modeling the real GRN. The update of the network state in the Boolean model is synchronous. However, biological networks are typically asynchronous. Boolean network modeling discretizes gene expression levels into two simple values. However, in

genes by combining fixed logical rules. It is difficult to accurately describe the real GRN enumerating all possible logical operations, so the Boolean network can only be used as the basis for modeling the real GRN. The update of the network state in the Boolean model is synchronous. However, biological networks are typically asynchronous. Boolean network modeling discretizes gene expression levels into two simple values. However, in real biological systems, gene expression is not a simple state, but continuous. When discretizing gene data, it will inevitably result in the loss of many important expression information [161], which can largely affect the accuracy of the model. Moreover, gene expression regulation should have at least three states: up-regulated, normal, or down-regulated, and its discretization is a difficult process. The setting of the threshold is crucial to determine the state of the node, and errors in the threshold setting will directly lead to changes in the gene state. It in turn will lead to inaccurate inferences, which is a common drawback of discrete models.

Liang et al. [162] first proposed to predict possible GRN structures from gene expression data using Boolean networks and developed a Boolean network-based software Reverse Engineering Algorithm (REVEAL) by considering the information entropy between nodes to help build the network structure. Kim et al. [163] proposed to utilize chi-squared tests to eliminate uncorrelated edges between nodes to accelerate the search for the optimal network structure. Due to the stochastic nature of biological systems and the noise contained in gene expression data, Boolean networks as deterministic models are not able to capture network regulatory relationships accurately. To solve this problem, a combination of the Boolean network and Markov chain was developed into the Probabilistic Boolean Network (PBN) model [164], which is a more flexible topology that adds stochasticity to the original network and can better handle the uncertainties among genes in the probabilistic framework. Boolean networks can be combined with MI to infer the structural and dynamic relationships between genes for time-series data [165]. The Single Cell Network Synthesis toolkit (SCNS) [166] is a computational tool for reconstructing and analyzing executable models from single-cell gene expression data. SCNS constructs a state transition graph of binary expression profiles using single-cell qPCR or RNA sequencing data acquired over the entire time course. An asynchronous Boolean network model is built by searching for rules that drive the transition from early to late cell states and thus reconstructing Boolean logical regulatory rules.

4.6. Boolean Implication Networks

The Boolean implication is the logical relationship between two Boolean variables, where the state of one variable can imply the state of the other variable. Boolean implication networks were first proposed by Sahoo et al. in 2008 [167] for building genome-wide gene relationship networks based on microarray data. The nodes in Boolean implication networks are genes and the edges are implication relationships. The implication is an if-then rule. For example, "if gene A is expressed high, then gene B is expressed high" which can be also expressed as "A high implies B high". The Boolean implication network automatically sets a separate threshold for each gene, which is used to classify the expression of a gene as "low" or "high". Then, the Boolean implications will be identified between each pair of genes in the whole genome.

There are six possible Boolean relationships in the Boolean implication network, including four asymmetric relationships: "A high $\Rightarrow$ B high", "A high $\Rightarrow$ B low", "A low $\Rightarrow$ B high", and "A low $\Rightarrow$ B low"; two symmetric relationships: if "A high $\Rightarrow$ B high" is accompanied by "B high $\Rightarrow$ A high", then gene A and gene B are "Boolean equivalent", if "A high $\Rightarrow$ B low" and "B high $\Rightarrow$ A low" at the same time, then gene A and gene B are "opposite".

The process of establishing Boolean connections between two genes A and B is shown in Figure 5. First of all, each gene has a threshold t derived using the StepMiner algorithm [168], and the interval of $t \pm 0.5$ is called "intermediate" (the gray areas in Figure 5). The values in the "intermediate" area may be misclassified, so the values in the "intermediate" range do not participate in the network creation process. Next, among the four

quadrants consisting of high-low expressions of gene A and gene B, we need to check which one is significantly sparser than the other quadrants. The sparsity can be calculated using the *statistic* and *error rate* (Equations (12) and (13)). For example, we want to detect the sparsity in quadrant IV (i.e., A high B low) and thus infer the implication rule of A high $\Rightarrow$ B high. Let a_I, a_{II},a_{III}, a_{IV} be the number of values in each quadrant.

$$\begin{aligned} total &= a_I + a_{II} + a_{III} + a_{IV} \\ n_{A\ high} &= a_I + a_{IV} \\ n_{B\ low} &= a_{III} + a_{IV} \\ expected &= \frac{n_{A\ high} + n_{B\ low}}{total} \\ observed &= a_{IV} \\ statistic &= \frac{expected - observed}{\sqrt{expected}} \end{aligned} \tag{12}$$

$$error\ rate = \frac{1}{2}\left(\frac{a_{IV}}{a_{IV} - a_I} + \frac{a_{IV}}{a_{IV} + a_{III}}\right) \tag{13}$$

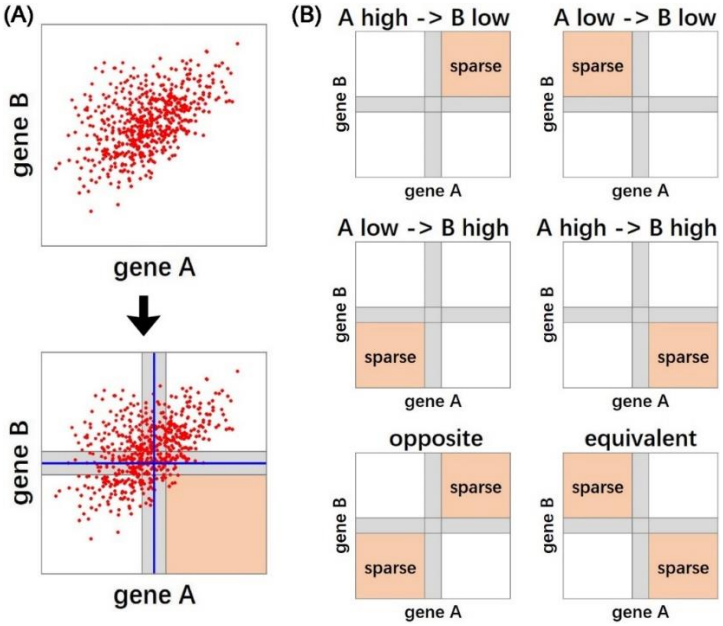


Figure 5. Example of a Boolean implication rule. (**A**) Inducing an implication rule based on quadrants of the categorized expression levels of gene A and gene B. The intermediate areas in gray are removed from the implication rule induction. (**B**). Six specific implication rules connecting gene A and gene B.

If quadrant IV has a calculated statistic greater than 3.0 and an error rate less than 0.1, we will consider that rule A high $\Rightarrow$ B high as significant. After the whole genome pairwise Boolean implication rules were generated, and the Boolean implication network was built.

Sinha et al. [169] applied the Boolean implication on mutation, copy number, methylation, and gene expression data. Their results indicated that a large number of Boolean implications exist in the data that could not be detected by other methods. Further analysis using GSEA showed that the Boolean implication networks also have biological significance. Their analysis showed that Boolean implications could be used for finding genes whose expression was regulated by copy number variations or DNA methylation changes. In the work of Çakır et al. [170], the combination of Boolean implication analysis with SOM revealed relationships between genes, metagenes, and similarly behaving metagene groups, and provided a more general and functional modularity-oriented view of the data.

The advantages of the Boolean implication networks are: (1) it can indicate the biological mechanisms between the pair of associated genes; (2) it is a directed graph in that each gene pair have a causal relationship with each other; and (3) it can incorporate feedback loops in the network. The disadvantage of these Boolean implication networks is all the gene variables have to be binary. Genes with an expression level in the "intermediate" range that is not up-regulated or down-regulated are removed from the analysis. Similarly

behaving metagene groups, and provided a more general and functional module-oriented view of the data.

The advantages of the Boolean implication networks are: (1) it can indicate the biological mechanisms between the pair of associated genes; (2) it is a directed graph in that each gene pair have a causal relationship with each other; and (3) it can incorporate feedback loops in the network. The disadvantage of these Boolean implication networks is all the gene variables have to be binary. Genes with an expression level in the "intermediate" range that is not up-regulated or down-regulated are removed from the analysis. Similarly, genes with normal copy numbers are also removed in the Boolean implication network modeling. This will result in considerable data loss and incomplete representation of real GRNs.

4.7. Prediction Logic Boolean Implication Networks (PLBINs)

We developed Boolean implication networks [171,172] based on prediction logic [173]. Using this algorithm, Boolean implication networks are inducted with logic rules connecting two binary variables. A contingency table is created for each pair of genes (Figure 6). The cells in the contingency table show the count of samples in each situation. There are six possible Boolean implication rules, and the count of error cells for each implication rule is used to calculate the precision and the scope to select the implication rules. The implication rules and their corresponding error cells are shown in Figure 6. The scope U_p and precision ∇_p for implication rule between $node_i$ and $node_j$ are calculated with the following equations:

$$U_p = U_{ij} = \frac{N_i \times N_j}{N^2} \tag{13}$$

$$U_p = \sum_i \sum_j \omega_{ij} \times U_{ij} \tag{14}$$

$$\nabla_p = \nabla_{ij} = 1 - \frac{N_{ij}}{N \times U_p} \tag{15}$$

$$\nabla_p = \sum_i \sum_j \frac{\omega_{ij} \times N_i \times N_j}{N^2} \times \nabla_{ij} \tag{16}$$

$$\nabla_p = \sum_i \sum_j \frac{\omega_{ij} \times U_{ij}}{U_p} \times \nabla_{ij} \tag{17}$$

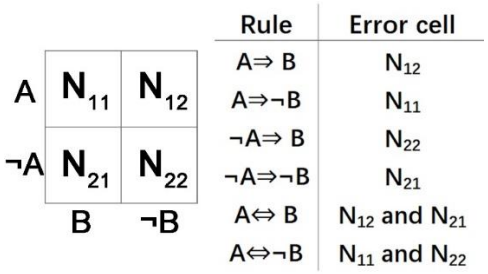


Figure 6. Contingency table of the Boolean implication rule and their corresponding error cells in prediction logic.

Equations (14) and (16) are for multiple cells, where $\omega_{ij} = 1$ for error cells, otherwise, $\omega_{ij} = 0$. To select the implication rules, thresholds for precision and scope are defined by the one-tailed z-tests based on the sample size and a preset z value. If a rule has the highest scope and precision amongst all six rules, and the scope and precision are greater than the thresholds, this implication rule will be considered a significant rule. The z value used for network construction is at least 1.645 (95% confidence interval, $\alpha = 0.05$, one-tailed z-tests).

The use of error cells enables the analysis of multivariate data in Boolean implication networks. The computational complexity of constructing genome-scale networks is $O(n^2)$, where n is the number of genes. Our PLBINs can model cyclic relations including feedback loops. PLBINs have been applied in modeling multi-omics [48,61,64] and single-cell [74] networks for the discovery of prognostic biomarkers and therapeutic targets in NSCLC.

4.8. Neural Networks

As an important branch in the field of machine learning, neural networks have been applied to systems biology and bioinformatics [174–176]. The relationships between genes and other gene products are often so complex for simple linear models to capture. Inspired by the animal central nervous system, neural networks are an effective mathematical model to learn multilayered complex patterns in linear and nonlinear functions. These advantages allow them to capture data features well and meet the requirements of higher accuracy in modeling multi-omics GRNs.

Neural Networks consist of multiple layers of neurons that are connected with other

4.8. Neural Networks

As an important branch in the field of machine learning, neural networks have been applied to systems biology and bioinformatics [174–176]. The relationships between genes and other gene products are often so complex for simple linear models to capture. Inspired by the animal central nervous system, neural networks are an effective mathematical model to learn multilayered complex patterns in linear and nonlinear functions. These advantages allow them to capture data features well and meet the requirements of higher accuracy in modeling multi-omics GRNs.

Neural Networks consist of multiple layers of neurons that are connected with other neurons in their preceding and succeeding layers. These neurons form three basic types of layers: the input layer, hidden layer, and output layer. A basic structure of the neural network is shown in Figure 7. The neural network model passes the feature representation of each level to the next level of unit modules by combining some simple nonlinear unit modules. By combining such nonlinear modules, neural networks can automatically extract higher and more abstract features from the original data and portray a more detailed biological data structure, which can provide modeling for complex systems [177,178]. Neural networks introduce nonlinear factors through activation functions such as the ReLu function, Sigmoid function, and Tanh function.

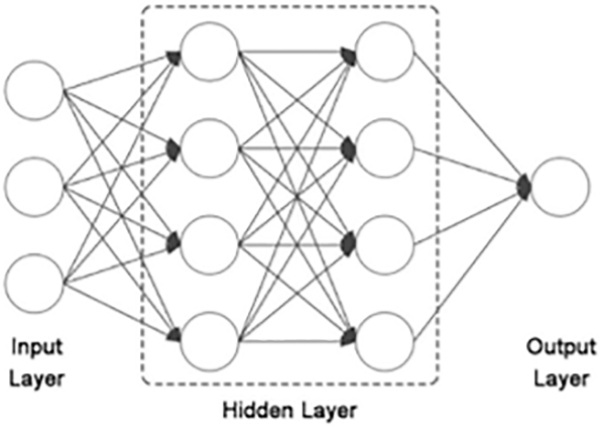


Figure 7. The basic structure of neural networks.

Figure 7. The basic structure of neural networks.

Alipanahi et al. [179] developed the DeepBind framework to predict the sequence specificities of DNA- and RNA-binding proteins using deep learning models. The DeepBind framework consists of a convolutional neural network for feature representation learning and a fully connected prediction module for feature combination, using gradient descent and backpropagation algorithms to train the model and compute nucleic acid binding interactions from different datasets. The DeepBind framework can be applied to a wide range of datasets and can improve the predictive power compared to traditional single-domain methods.

The Recurrent Neural Network model exhibits strong modeling capabilities with its nonlinear structure and can adaptively recognize and remember temporal and spatial patterns, which can more realistically simulate the working processes of real biological systems. Because of its ability to establish nonlinear and dynamic interactions between genes, RNN is also a well-established method for deriving GRNs with up to 30 genes [180].

Graph neural networks (GNN), as a generalization of neural networks, are deep learning architectures that can handle graph and graph-related problems, such as node classification, link prediction, and graph classification [181,182]. Graph Convolution Networks [183], a kind of GNN, migrate the traditional convolutional operations in deep learning to the processing of graph-structured data and specify them through complex spectral graph theory derivation. Its core idea is to learn the features of a node in a graph itself and the features of its neighbors and aggregate them to generate a new mapping of functions representing vectors for this node. Wang et al. [184] proposed an end-to-end gene regulatory graph neural network (GRGNN) to reconstruct GRNs in a supervised and semi-supervised framework using gene expression data. To obtain better inductive generalization, the GRN inference is formulated as a graph classification problem to distinguish whether a subgraph centered on two nodes contains a link between these two nodes

functions representing vectors for this node. Wang et al. [184] proposed an end-to-end gene regulatory graph neural network (GRGNN) to reconstruct GRNs in a supervised and semi-supervised framework using gene expression data. To obtain better inductive generalization, the GRN inference is formulated as a graph classification problem to distinguish whether a subgraph centered on two nodes contains a link between these two nodes. The computational complexity to construct GRGNN is exponential of the number (h-hop) of subgraphs [184]. Single-cell graph neural network (scGNN) was developed to provide a hypothesis-free deep learning framework for single-cell RNA-sequencing data imputation and clustering [185]. Major deep learning-based methods in cancer classification and clustering using multi-omics and single-cell data were benchmarked [186].

The main limitation of neural network-based GRN inference is the requirement of the training data. The network training requires benchmarks with systematically explicit, experimentally validated, gold-standard conditioning relationships. On species with complete data, the goal of inference may be easily achieved. However, it is challenging in constructing genome-scale neural network-based GRN in complex human diseases, such as cancer. Two problems exist with deep learning modeling of genomic data: (1) the insufficient amount of training data, which affects the model performance, and (2) the high data dimensionality, which leads to a huge number of model parameters and increases the training difficulty. In addition, although neural networks are very good at learning complex tasks, their internal descriptions are generally difficult to interpret, and training deeply layered models is algorithmically difficult to handle and statistically prone to over-fit.

4.9. Summary of Existing Network Inference Methods

These network methods have many applications in the discovery of biomarkers [187–193] and therapeutic targets [194–196]. They are implemented in several software packages. GeNeCK [197] is a web server that allows users to build GRNs from expression data using different network construction methods, including four partial correlation-based methods: *GeneNet*, *NS*, *SPACE*, and *ESPACE*; four likelihood-based methods: *GLASSO*, *GLASSO-SF*, *BayesianGLASSO*, and *EGLASSO*; and two mutual information-based methods: *PCACMI* and *CMI2NI*. *EGLASSO* and *ESPACE* accept hub gene specification to improve the network results. There is also an ensemble method, *ENA*, which does not require choosing or tuning parameters so it is suitable for most users. *ENA* provides a p-value for each edge in the network to indicate its statistical significance. The Weka software (Version 3.8.6) [198] implements commonly used machine learning methods for classification, including radial basis function networks and Bayesian belief networks (BBNs). A summary of software tools for multi-omics processing, pathway analysis, and network inferencing is provided in Table 1.

Despite the successful applications of these network models in classification and clustering, there are certain limitations in these methods to construct genome-scale GRNs using emerging multi-omics data. Relevance networks can only measure the linear relationship between genes and are not robust. Relevance networks cannot model categorical data such as DNA mutations or CNVs in muti-omics analysis. Bayesian networks have high computational complexity and can only be used for small and medium-sized data. The static Bayesian networks cannot represent cyclic relations such as feedback loops. The parameter learning process of ODE networks is very complex and is limited by data sample size. Other Boolean (implication) networks can only model binary variables, which do not present biological states of mutations, CNVs, or gene/protein expression. Neural networks are limited by high requirements for sample size and completeness of information in training data and the exponential complexity of the number of subgraphs, which makes it infeasible to model genome-scale multi-omics networks for complex human diseases such as cancer.

Table 1. Summary of software for multi-omics data processing, pathway analysis, and network inferencing in bulk tumors and single cells.

Purpose	Software
Data Processing	
Multi-omics data	GATK [43]
Copy number variation	PennCNV-Affy [51], CGHbase [52], CGHcall [53], GISTIC2.0 [54]
Single-cell RNA sequencing	Ginkgo [68], *STAR* aligner [72], SAMtools [73], DEsingle [75], scGNN [185]
Pathway Analysis	GSEA [77], ToppFun [78], Qiagen IPA, Adviata iPathwayGuide
Stromal and Immune Infiltration and Cell Activity	ESTIMATE [90], xCell [91], TIMER 2.0 [92–94], CIBERSORTx [95], MCP-counter [96]
Drug Discovery and Repositioning	CMap [26,27]
Network Inferencing Methods	GeNeCK [197]
Relevance networks	MiBiOmics [108], OmicsAnalyst [109], CorDiffViz [110]
Bayesian networks	CBNplot [144], TETRAD IV [147]
Boolean networks	SCNS [166]
PLBINs	Proprietary
Classification	Weka [198] (including neural networks and Bayesian networks)

Our PLBINs overcome the limitations of other methodologies. First, PLBINs can integrate discrete CNV data and continuous gene/protein expression data seamlessly that relevance networks cannot. Second, PLBINs can model cyclic molecular interactions that the acyclic Bayesian networks cannot. Third, PLBINs have a computational complexity of $O(n^2)$ and can efficiently model genome-scale GRNs that Bayesian networks, neural networks, and ODE networks cannot. Finally, PLBINs can model multinary data with robust statistical tests, whereas other Boolean networks can only analyze binary variables. This is a major advantage of PLBINs because there need to be at least three biologically relevant states without losing important information in categorized CNV (amplification/normal/deletion) andgene/protein expression data (upregulation/normal/downregulation). Our PLBINs identified gene signatures that accurately predict the risk of lung cancer risk and tumor recurrence, outperforming previous studies ones in the same patient data [171,199–201], meanwhile, revealing more biologically relevant molecular interactions than other network methodologies in comprehensive evaluation with MSigDB [171,201]. Using our PLBINs, we developed a seven-gene signature for NSCLC prognosis and prediction of the clinical benefits of adjuvant chemotherapy in early-stage NSCLC patients, including clinical trials [56]. Our 7-gene signature is unique that it (1) works on all NSCLC histological subtypes and multiple clinical testing platforms; (2) predicts the risk of tumor recurrence; and (3) classifies NSCLC tumors from non-cancerous normal adjacent tissues [56,61,64].

The comparative data of our PLBINs and other methods were published previously [171,201]. The precision and false discovery rate (FDR) of the gene coexpression networks were evaluated as follows [171]. The validity of computationally derived coexpression relations was comprehensively evaluated with five gene set collections (positional, curated, motif, computational, and Gene Oncology) and canonical pathway databases in the MSigDB [77]. A coexpression relation was labeled as a true positive (TP) if both genes were present in the same gene set or pathway in any examined database. A coexpression relation was labeled a false positive (FP) if the gene pair did not share any gene set or pathway in all the examined databases. A coexpression relation was defined as non-discriminatory (ND) if at least one gene in the pair was not annotated in a database [202]. The evaluation did not include ND coexpression relations as they were not confirmatory. The precision of a gene expression network was defined as TP/(TP + FP). The precision of our identified smoking-mediated coexpression networks in NSCLC patient tumors was 100% [171]. To test the statistical significance of the network precision, the null distribution was gener-

ated in 1000 random permutations of the class labels in the test cohort. The precision of our identified smoking-mediated coexpression networks was significant at $p < 0.001$, with no TP generated in the random tests. The FDR of gene coexpression networks was defined as the average of FP/(TP + FP) in 1000 permutations. The FDR of our identified smoking-mediated coexpression networks in NSCLC patient tumors was 0.0099. In contrast, Pearson's correlation networks did not identify any coexpression relations using the same methodology on the same datasets [171]. In the evaluation of our identified 21 NSCLC prognostic gene signatures [199] using the NCI Director's Challenge Study [203], our PLBINs-derived gene coexpression relations from the training cohort could be successfully reproduced in both test cohorts with significantly high precision (precision = 1 for 18 gene signatures) and low FDR (FDR < 0.1) for all 21 gene signatures [201]. As a comparison, the Bayesian networks implemented in TETRAD IV [147] did not identify any coexpression relations from the training cohort that were validated in both test cohorts [201]. The Boolean implication networks by Sahoo et al. [167] did not identify coexpression with many of the major NSCLC hallmarks, making it infeasible to select marker genes with concurrent crosstalk with multiple signaling pathways as we did with our PLBINs. In the genome-scale evaluation, our PLBINs achieved significantly high precision in 1000 random tests ($p < 0.05$), whereas the precision of the Boolean implication networks by Sahoo et al. [167] was not significant in 1000 random tests ($p = 0.21$) [201]. These results demonstrate that our PLBINs are more accurate in retrieving biologically relevant gene associations, in addition to other advantages such as computational scalability and efficiency.

5. Hub Genes in Multi-Omics and Single-Cell Networks

Some hub genes in multi-omics networks were shown to be promising cancer biomarkers and therapeutic targets [204,205]. Nevertheless, there were insufficient genome-scale investigations on multi-omics network hub genes and their biological and clinical relevance in human cancers. Graph theory centrality metrics can characterize hub genes. Common metrics include degree centrality (in-degree and out-degree centralities) [206], eigenvector centrality [207–209], betweenness centrality [210,211], closeness centrality [212–214], and VoteRank centrality [215]. Degree centrality is simply the total number of neighbors of each node. The eigenvector centrality of a node is the sum of the centrality of its neighbors. Betweenness centrality is the frequency of a node appearing on the shortest paths of all node pairs in the entire network. Closeness centrality is the average length of the shortest paths between the node and all other nodes in the network. VoteRank centrality is selected with a voting score that is calculated by the sum of all neighbors' voting abilities. Degree centrality and eigenvector centrality are also classified as local centrality metrics because only neighbors of each node are included in the calculation. Betweenness centrality, closeness centrality, and VoteRank centrality are categorized as global centrality metrics since the connectivity of the entire graph is used in the metrics computation. These centrality metrics are correlated in many cases [214,216]. A Python package NetworkX [217] calculates centrality metrics.

A barrier to this systematic evaluation is the limitations of existing computational methodologies in constructing genome-scale multi-omics GRNs, as summarized above. In a recent study [218], our PLBINs were used to construct 12 genome-scale GRNs of CNV, mRNA, and protein expression in NSCLC tumors. Seven centrality metrics were correlated with NSCLC tumorigenesis measured with T-statics in differential gene/protein expression between tumors vs. non-cancerous adjacent tissues (NATs), proliferation quantified with dependency scores from CRISPR-Cas9/RNAi screening of human NSCLC cell lines, and patient survival with hazard ratios from Cox modeling of The Cancer Genome Atlas (TCGA) [218]. Hub genes in multi-omics networks involving gene/protein expression were found to be associated with oncogenic, proliferative potentials and poor patient survival. Hub genes with higher co-occurrences of CNV aberrations seemed to have tumor-suppressive and anti-proliferative properties. Regulated genes in hubs were linked to proliferative potential and worse patient survival, whereas regulatory genes in hubs

were linked to anti-proliferative potential and better patient survival. Established cancer immunotherapy targets PD1, PDL1, CTLA4, and CD27 were top hub genes in most constructed multi-omics GRNs [218]. These results show that multi-omics network centrality in bulk tumors can be used in the prioritization of biomarkers and therapeutic targets.

Similarly, our PLBINs [74] were applied to genome-wide transcriptomic profiles of B cells from tumors and NATs [219], T cells from peripheral blood lymphocytes (PBL) [220], and tumor-infiltrating T-cell gene expression data of NSCLC patients. In each cell sample, a gene was defined as expressed (with a feature count > 0) or not-expressed (with a feature count = 0). The details of single-cell network construction were provided in our previously published study [74]. The results of five single-cell co-expression networks are shown in Table 2.

Table 2. Information of single-cell gene co-expression networks. The network nodes are genes and network edges are computed gene associations (one-tailed *z*-tests, $p < 0.05$, 95% confidence interval).

Patient Cohort	Network (Number of Cell Samples)	Number of Network Nodes	Number of Network Edges
GSE84789	NATs: B-cell gene co-expression ($n = 96$)	13,797	21,474,928
	Tumors: B-cell gene co-expression ($n = 96$)	13,420	6,298,276
GSE151531	Healthy donors: T-cell PBL gene co-expression ($n = 431$)	16,143	5,246,634
	NSCLC Patients: T-cell PBL gene co-expression ($n = 92$)	11,082	2,138,492
GSE151537	Tumors: T-cell gene co-expression ($n = 2950$)	20,171	7,805,674

We examined the centrality metrics of four established immune checkpoint inhibitors (ICIs), including *PD1*, *PDL1*, *CD27*, and *CTLA4*. Figure 8 shows the centrality distribution of the ICIs that were within the top 10th percentile in the constructed networks. *PD1* was ranked as a top hub gene in the T-cell PBL gene co-expression network in healthy donors. *CTLA4* was ranked as a top hub gene in the T-cell PBL gene co-expression network in NSCLC tumors. *CD27* was ranked as a top hub gene in the T-cell PBL gene co-expression network in NSCLC patients. These results are consistent with their functional involvements in T-cell immunity. *PDL1* was not ranked within the top 10th percentile of any of the examined centrality metrics in the constructed networks. None of these ICIs were ranked as top hub genes in B-cell gene co-expression networks in tumors or NATs.

In a previous study, we identified a gene co-expression network missing in NSCLC tumor B cells using PLBINs [74]. Genes in this network either promote proliferation in human NSCLC epithelial cells or are indicative of NSCLC patient outcomes at both mRNA and protein expression levels in bulk tumors. These network genes were associated with drug response to 10 therapeutic regimens in 135 human NSCLC cell lines. Based on this single B-cell co-expression network, we discovered tyrosine kinase inhibitor lestaurtinib as a new drug option for treating NSCLC [74]. Here, we examined if this clinically relevant single B-cell network had higher average centrality compared with 1000 random networks with the same number of genes selected from the genome. The results showed that the previously published B-cell network had significantly higher average centrality ($p < 0.05$) than 1000 random networks selected from genome-scale single B-cell networks in tumors and NATs, single T-cell PBL networks in NSCLC patients and healthy donors, and T-cell network in NSCLC tumors (Figure 9). These results support the relevance of single-cell network hub genes in tumor biology.

CTLA4 was ranked as a top hub gene in the T-cell PBL gene co-expression network in NSCLC tumors. *CD27* was ranked as a top hub gene in the T-cell PBL gene co-expression network in NSCLC patients. These results are consistent with their functional involvements in T-cell immunity. *PDL1* was not ranked within the top 10th percentile of any of the examined centrality metrics in the constructed networks. None of these genes were ranked as top hub genes within the top 10th percentile in networks in tumor or NATs.

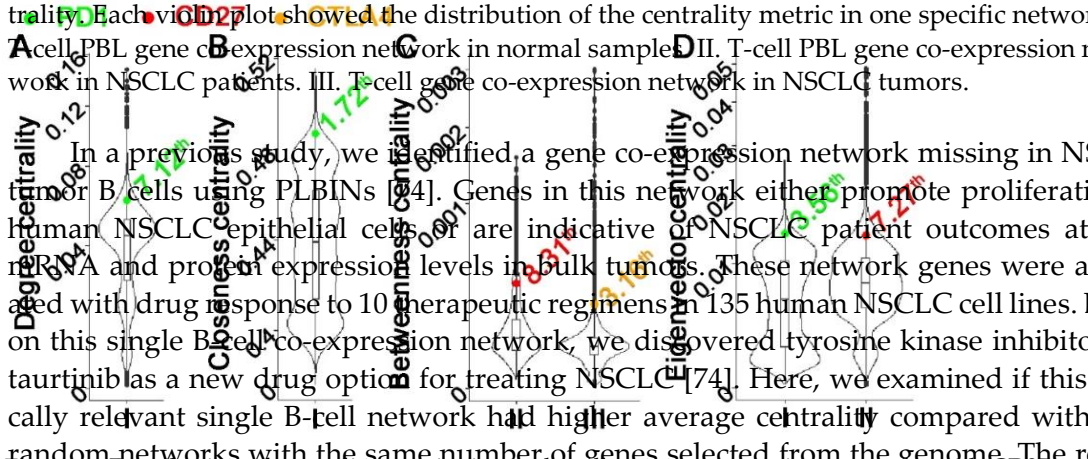


Figure 8. Distribution of centrality metrics in single-cell gene co-expression networks with *CD27*, *CTLA4*, or *PD1* ranked within the top 10th percentile. Each subplot represented a centrality metric (**A**). Degree centrality; (**B**). Closeness centrality; (**C**). Betweenness centrality; (**D**). VoteRank centrality. Each violin plot showed the distribution of the centrality metric in one specific network: I. T-cell PBL gene co-expression network in normal samples. II. T-cell PBL gene co-expression network in NSCLC patients. III. T-cell gene co-expression network in NSCLC tumors.

In a previous study, we identified a gene co-expression network missing in NSCLC tumor B cells using PLBINs [74]. Genes in this network either promote proliferation in human NSCLC epithelial cells or are indicative of NSCLC patient outcomes at both mRNA and protein expression levels in bulk tumors. These network genes were associated with drug response to 10 therapeutic regimens in 135 human NSCLC cell lines. Based on this single B-cell co-expression network, we discovered tyrosine kinase inhibitor lestaurtinib as a new drug option for treating NSCLC [74]. Here, we examined if this clinically relevant single B-cell network had higher average centrality compared with 1000 random networks with the same number of genes selected from the genome. The results showed that the previously published B-cell network had significantly higher average centrality ($p < 0.05$) than 1000 random networks selected from genome-scale single B-cell networks in tumors and NATs, single T-cell PBL networks in NSCLC patients and healthy donors, and T-cell networks in NSCLC tumors (Figure 9). These results support the relevance of single-cell network hub genes in tumor biology.

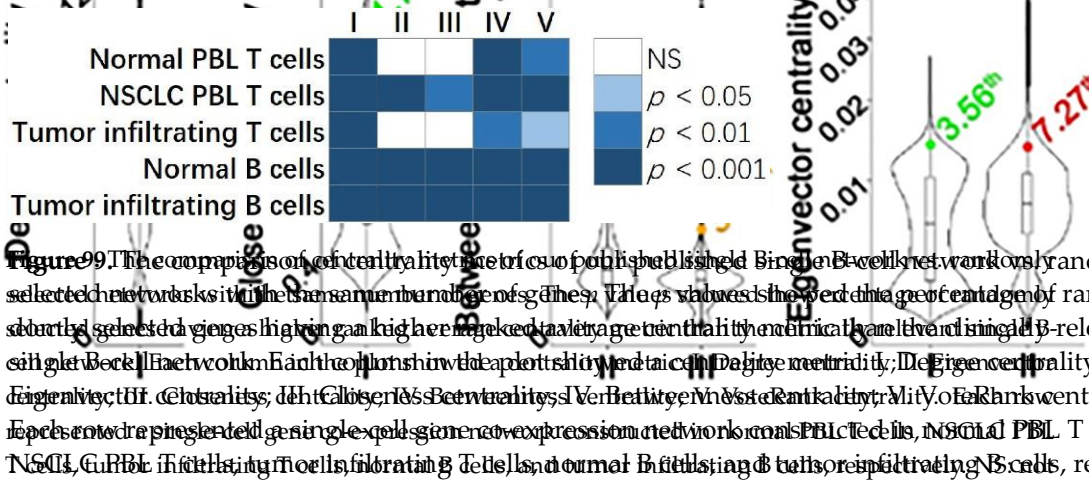


Figure 9. The comparisons of centrality metrics of published single B-cell network vs. randomly selected networks with the same number of genes. The values showed the percentage of randomly selected genes having a higher average centrality metric than the clinically relevant single B-cell network. Each column showed the plot for a centrality metric: I. Degree centrality; II. Eigenvector centrality; III. Closeness centrality; IV. Betweenness centrality; V. VoteRank centrality. Each row represented a single-cell gene co-expression network constructed in normal PBL T cells, NSCLC PBL T cells, tumor infiltrating T cells, normal B cells, and tumor infiltrating B cells, respectively. NS: statistically significant.

Our PLBIN algorithm was written in C and R and was run on Spruce Knob High-Performance Computing (HPC) Clusters. It took about 67 minutes for the algorithm to finish on an HPC node with 4 × 8 Intel(R) Xeon(R) CPU E5-4620 0 @ 2.20GHz 64GB memory 1TB HDD, in the case of whole-genome network construction which yielded approximately 20 million rules. The program for network-centrality metrics calculation was written in Python. The running time on the HPC cluster node for a multi-omics network with 20 million edges and 12 thousand nodes for each centrality metric was provided in Table 3. The network with 20 million edges and 12 thousand nodes for each centrality metric was provided in Table 3. The complexity of each method and the actual running time were consistent as shown in Table 3.

A summary of concordant significant correlations between multi-omics network centrality and tumorigenesis, proliferation, and patient survival cross NSCLC cohorts reported in our previous study [218] is provided in Table 3. According to the running time (Table 2) and the concordant significant correlations (Table 4), eigenvector centrality, closeness centrality, and degree centrality were the top three performing methods in terms of computational efficiency and recapitulation of network biological and clinical relevance. Degree centrality is the best choice if genes are regulated or regulators need to be studied with in-degree or out-degree centralities, respectively.

Table 3. Computational complexity and running time of each centrality method. N: the number of nodes (genes). E: the number of edges (gene associations).

Method	Complexity	Running Time (or a Network with 20 Million Edges)
Boolean Implication Network	$O(N^2)$	67 minutes

Table 3. Computational complexity and running time of each centrality method. *N*: the number of nodes (genes). *E*: the number of edges (gene associations).

Method	Complexity	Running Time (of a Network with 20 Million Edges)
PLBIN	$O(N^2)$	67 min
Degree Centrality	$O(N)$	0.02 s
Eigenvector Centrality	$O(N+E)$	89 s
Closeness Centrality	$O(NE)$	121 min
Betweenness Centrality	$O(N^2 logN + NE)$	24 h
VoteRank Centrality	$O\left(E + rlogN + r\frac{E^2}{N^2}\right)$	53 h

Table 4. Counts of concordant significant associations of each centrality metric with tumorigenesis, proliferation, and patient survival in the multi-omics networks.

Centrality Metric	Tumorigenesis (mRNA Expression)	Tumorigenesis (Protein Expression)	Proliferation (CRISPR-Cas9)	Proliferation (RNAi)	Patient Survival	Sum
Degree Centrality	3	2	3	3	2	13
Eigenvector Centrality	4	1	5	5	3	18
Closeness Centrality	4	1	5	4	2	16
Betweenness Centrality	0	1	2	2	1	6
VoteRank Centrality	0	0	4	3	1	8
Sum	11	5	19	17	9	61

6. Integrating Multi-Omics Data with Patient Electronic Medical Records

The successful application of biomarkers and drugs requires rigorous testing in the patient population considering diverse clinical, pathological, comorbid, and demographic factors. In certain cancer types such as lung cancer, lifestyle factors including smoking, and environmental and occupational exposures, also need to be considered. Nevertheless, it is not currently feasible to conduct multi-omics and single-cell profiling in tens of thousands of patients using a well-controlled clinical study design, due to the required costs, time, and infrastructure. When a new treatment is added to the NCCN guidelines, it may take years to collect sufficient data to establish predictive biomarkers. In current biomarker studies, candidate genes are first identified from clinical cohorts of a limited number of patient samples and are then validated by leveraging public data such as TCGA. The following gaps exist in clinical applications of biomarkers: (1) Most published patient cohorts, including TCGA, do not have complete treatment information and the number of patients in specific treatment categories is very small, making it infeasible to establish predictive biomarkers of therapeutic response for clinical utility; (2) Some sequencing facilities do not have patient treatment or outcome information on all the samples they have sequenced for predictive biomarker R&D; and (3) Large-scale patient EMRs of hospital information systems or cancer registries have enough patients with comprehensive clinical information but do not have sufficient matched patient genome-scale profiles for biomarker discovery. To determine the applicability of multi-omics biomarkers in general patient

populations, large-scale EMRs and genomic/transcriptomic profiles from specific patient cohorts must be combined.

By merging SEER-Medicare data, we created a unique technique to find prognostic and chemopredictive biomarkers with the potential to be used in large patient populations to fill this gap [221]. The SEER database is a compilation of registration information from specific geographic areas, which account for around 26% of the U.S. population [222]. Without additional natural language processing, the linked SEER-Medicare data are adequately annotated and prepared for computational analysis. A previous study identified chemopredictive genes by correlating mRNA expression profiles in solid tumors in the advanced cancer stage of a Serial Analysis of Gene Expression (SAGE) database with patient survival in SEER data [223]. In our previous study, a novel tumor progression indicator, combining AJCC cancer staging [224] T, N, and M factors with tumor grade was used to correlate miRNA expression in a lung squamous cell carcinoma (LUSC) patient cohort with SEER-medicare LUSC patient outcomes receiving different treatments. The identified chemopredictive miRNAs were then validated with extensive pubic data and our collected patient cohorts. Our study revealed miRNA-mediated transcriptional networks in NSCLC proliferation and progression using CRISPR-Cas9/RNAi screening data [221]. Our findings show that, in the absence of novel cohorts with tens of thousands of patients who have matched clinical outcomes and genome-scale transcriptomic profiles, extrapolation of miRNA expression from smaller cohorts to larger population-based data can serve as an additional confirmatory tool based on similarities in tumor progression. This method, in conjunction with stringent external validation, can discover prognostic and predictive biomarkers with concordant expression patterns in tumor development in sizable patient populations.

7. Recommendations

Multi-omics network analysis of bulk tumors and single cells can help understand molecular mechanisms in multi-dimensional tumor immune microenvironments for the identification of clinically relevant biomarkers and effective therapeutic targets. The increasing amount of data generated with various high-throughput platforms can accelerate scientific discovery, and meanwhile, pose a challenge in harmonization and computation. To integrate genomic data such as CNV and SV generated from different sources, the genome assembly version in each dataset should be converted to hg38. To define the regulation of gene and protein expression, a set of housekeeping genes with stable expression in the studied tissue type should be used for data normalization. To construct genome-scale multi-omics regulatory networks, our Prediction Logic Boolean Implication Networks (PLBINs) have advantages over other methods in terms of computational efficiency, scalability, and accuracy [48,61,64,74]. Our recent study shows that graph theory network centralities can be used for the prioritization of biomarkers and therapeutic targets [218]. Eigenvector centrality, degree centrality, and closeness centrality are top-ranked metrics regarding time complexity and performance. Finally, multi-omic biomarkers should be integrated with patient clinical, pathological, demographic, and comorbid factors for optimal treatment selection. Our approach to integrating multi-omic profiles with large-scale patient EMRs such as the SEER-Medicare cancer registry [221] can identify biomarkers with consistent expression patterns in tumor progression, with potential prognostic and predictive implications in large patient populations. Our methodologies form a conceptually innovative framework encompassing various available information from research laboratories to healthcare systems for the discovery of biomarkers and therapeutic targets, including new and repositioning drugs, ultimately improving cancer patient survival outcomes.

8. Patents

The artificial intelligence methodology using Boolean implication networks based on prediction logic for drug discovery was filed under PCT/US22/75136.

Author Contributions: Conceptualization, Q.Y. and N.L.G.; methodology, N.L.G.; software, N.L.G. and Q.Y.; formal analysis, Q.Y.; data curation, Q.Y.; writing—original draft preparation, Q.Y. and N.L.G.; writing—review and editing, N.L.G.; visualization, Q.Y.; supervision, N.L.G.; funding acquisition, N.L.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Institutes of Health R01/R56 LM009500, P20RR16440, and ARRA Supplement.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data was published and publicly available in the cited manuscripts. An earlier implementation of our Prediction Logic Boolean Implication Networks (PLBINs) was released in SourceForge: https://sourceforge.net/projects/genet-cnv/ (accessed on 21 December 2022). The current version of the software is patented for the discovery of biomarkers and therapeutic targets and is not released.

Acknowledgments: We acknowledge institutional support from the Graduate Research and Education Office at Health Sciences Center at West Virginia University.

Conflicts of Interest: N.L.G. is the inventor of the patent PCT/US22/75136 filed and owned by West Virginia University. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* **2021**, *71*, 209–249. [CrossRef]
2. Cancer Moonshot℠. Available online: https://www.cancer.gov/research/key-initiatives/moonshot-cancer-initiative (accessed on 4 December 2022).
3. Soda, M.; Choi, Y.L.; Enomoto, M.; Takada, S.; Yamashita, Y.; Ishikawa, S.; Fujiwara, S.; Watanabe, H.; Kurashina, K.; Hatanaka, H.; et al. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature* **2007**, *448*, 561–566. [CrossRef] [PubMed]
4. Patsoukis, N.; Wang, Q.; Strauss, L.; Boussiotis, V.A. Revisiting the PD-1 pathway. *Sci. Adv.* **2020**, *6*, eabd2712. [CrossRef] [PubMed]
5. Chaft, J.E.; Rimner, A.; Weder, W.; Azzoli, C.G.; Kris, M.G.; Cascone, T. Evolution of systemic therapy for stages I-III non-metastatic non-small-cell lung cancer. *Nat. Rev. Clin. Oncol.* **2021**, *18*, 547–557. [CrossRef] [PubMed]
6. Forde, P.M.; Chaft, J.E.; Smith, K.N.; Anagnostou, V.; Cottrell, T.R.; Hellmann, M.D.; Zahurak, M.; Yang, S.C.; Jones, D.R.; Broderick, S.; et al. Neoadjuvant PD-1 Blockade in Resectable Lung Cancer. *N. Engl. J. Med.* **2018**, *378*, 1976–1986. [CrossRef] [PubMed]
7. Hellmann, M.D.; Rizvi, N.A.; Goldman, J.W.; Gettinger, S.N.; Borghaei, H.; Brahmer, J.R.; Ready, N.E.; Gerber, D.E.; Chow, L.Q.; Juergens, R.A.; et al. Nivolumab plus ipilimumab as first-line treatment for advanced non-small-cell lung cancer (CheckMate 012): Results of an open-label, phase 1, multicohort study. *Lancet Oncol.* **2017**, *18*, 31–41. [CrossRef]
8. Reck, M.; Rodríguez-Abreu, D.; Robinson, A.G.; Hui, R.; Csőszi, T.; Fülöp, A.; Gottfried, M.; Peled, N.; Tafreshi, A.; Cuffe, S.; et al. Pembrolizumab versus Chemotherapy for PD-L1-Positive Non-Small-Cell Lung Cancer. *N. Engl. J. Med.* **2016**, *375*, 1823–1833. [CrossRef]
9. Reck, M.; Rodríguez-Abreu, D.; Robinson, A.G.; Hui, R.; Csőszi, T.; Fülöp, A.; Gottfried, M.; Peled, N.; Tafreshi, A.; Cuffe, S.; et al. Updated Analysis of KEYNOTE-024: Pembrolizumab Versus Platinum-Based Chemotherapy for Advanced Non-Small-Cell Lung Cancer With PD-L1 Tumor Proportion Score of 50% or Greater. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **2019**, *37*, 537–546. [CrossRef]
10. Mok, T.S.K.; Wu, Y.L.; Kudaba, I.; Kowalski, D.M.; Cho, B.C.; Turna, H.Z.; Castro, G., Jr.; Srimuninnimit, V.; Laktionov, K.K.; Bondarenko, I.; et al. Pembrolizumab versus chemotherapy for previously untreated, PD-L1-expressing, locally advanced or metastatic non-small-cell lung cancer (KEYNOTE-042): A randomised, open-label, controlled, phase 3 trial. *Lancet* **2019**, *393*, 1819–1830. [CrossRef]
11. Doroshow, D.B.; Sanmamed, M.F.; Hastings, K.; Politi, K.; Rimm, D.L.; Chen, L.; Melero, I.; Schalper, K.A.; Herbst, R.S. Immunotherapy in Non-Small Cell Lung Cancer: Facts and Hopes. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **2019**, *25*, 4592–4602. [CrossRef]
12. Emens, L.A. Predictive Biomarkers: Progress on the Road to Personalized Cancer Immunotherapy. *J. Natl. Cancer Inst.* **2021**, *113*, 1601–1603. [CrossRef] [PubMed]
13. Sautès-Fridman, C.; Petitprez, F.; Calderaro, J.; Fridman, W.H. Tertiary lymphoid structures in the era of cancer immunotherapy. *Nat. Rev. Cancer* **2019**, *19*, 307–325. [CrossRef] [PubMed]

14. Dieu-Nosjean, M.-C.; Antoine, M.; Danel, C.; Heudes, D.; Wislez, M.; Poulot, V.; Rabbe, N.; Laurans, L.; Tartour, E.; de Chaisemartin, L. Long-term survival for patients with non–small-cell lung cancer with intratumoral lymphoid structures. *J. Clin. Oncol.* **2008**, *26*, 4410–4417. [CrossRef] [PubMed]
15. Xia, L.; Guo, L.; Kang, J.; Yang, Y.; Yao, Y.; Xia, W.; Sun, R.; Zhang, S.; Li, W.; Gao, Y.; et al. Predictable Roles of Peripheral IgM Memory B Cells for the Responses to Anti-PD-1 Monotherapy Against Advanced Non-Small Cell Lung Cancer. *Front. Immunol.* **2021**, *12*, 759217. [CrossRef]
16. Cabrita, R.; Lauss, M.; Sanna, A.; Donia, M.; Skaarup Larsen, M.; Mitra, S.; Johansson, I.; Phung, B.; Harbst, K.; Vallon-Christersson, J.; et al. Tertiary lymphoid structures improve immunotherapy and survival in melanoma. *Nature* **2020**, *577*, 561–565. [CrossRef]
17. Helmink, B.A.; Reddy, S.M.; Gao, J.; Zhang, S.; Basar, R.; Thakur, R.; Yizhak, K.; Sade-Feldman, M.; Blando, J.; Han, G.; et al. B cells and tertiary lymphoid structures promote immunotherapy response. *Nature* **2020**, *577*, 549–555. [CrossRef]
18. Petitprez, F.; de Reyniès, A.; Keung, E.Z.; Chen, T.W.; Sun, C.M.; Calderaro, J.; Jeng, Y.M.; Hsiao, L.P.; Lacroix, L.; Bougoüin, A.; et al. B cells are associated with survival and immunotherapy response in sarcoma. *Nature* **2020**, *577*, 556–560. [CrossRef]
19. Stankovic, B.; Bjørhovde, H.A.K.; Skarshaug, R.; Aamodt, H.; Frafjord, A.; Müller, E.; Hammarström, C.; Beraki, K.; Bækkevold, E.S.; Woldbæk, P.R.; et al. Immune Cell Composition in Human Non-small Cell Lung Cancer. *Front. Immunol.* **2018**, *9*, 3101. [CrossRef]
20. Germain, C.; Gnjatic, S.; Tamzalit, F.; Knockaert, S.; Remark, R.; Goc, J.; Lepelley, A.; Becht, E.; Katsahian, S.; Bizouard, G.; et al. Presence of B cells in tertiary lymphoid structures is associated with a protective immunity in patients with lung cancer. *Am. J. Respir. Crit. Care Med.* **2014**, *189*, 832–844. [CrossRef]
21. Germain, C.; Devi-Marulkar, P.; Knockaert, S.; Biton, J.; Kaplon, H.; Letaïef, L.; Goc, J.; Seguin-Givelet, A.; Gossot, D.; Girard, N.; et al. Tertiary Lymphoid Structure-B Cells Narrow Regulatory T Cells Impact in Lung Cancer Patients. *Front. Immunol.* **2021**, *12*, 626776. [CrossRef]
22. Thommen, D.S.; Schumacher, T.N. T Cell Dysfunction in Cancer. *Cancer Cell* **2018**, *33*, 547–562. [CrossRef] [PubMed]
23. Labanieh, L.; Majzner, R.G.; Mackall, C.L. Programming CAR-T cells to kill cancer. *Nat. Biomed. Eng.* **2018**, *2*, 377–391. [CrossRef]
24. Depil, S.; Duchateau, P.; Grupp, S.A.; Mufti, G.; Poirot, L. 'Off-the-shelf' allogeneic CAR T cells: Development and challenges. *Nat. Rev. Drug Discov.* **2020**, *19*, 185–199. [CrossRef] [PubMed]
25. Manfredi, F.; Cianciotti, B.C.; Potenza, A.; Tassi, E.; Noviello, M.; Biondi, A.; Ciceri, F.; Bonini, C.; Ruggiero, E. TCR Redirected T Cells for Cancer Treatment: Achievements, Hurdles, and Goals. *Front. Immunol.* **2020**, *11*, 1689. [CrossRef]
26. van der Leun, A.M.; Thommen, D.S.; Schumacher, T.N. CD8(+) T cell states in human cancer: Insights from single-cell analysis. *Nat. Rev. Cancer* **2020**, *20*, 218–232. [CrossRef] [PubMed]
27. Yazdanifar, M.; Barbarito, G.; Bertaina, A.; Airoldi, I. γδ T Cells: The Ideal Tool for Cancer Immunotherapy. *Cells* **2020**, *9*, 1305. [CrossRef]
28. Singh, A.K.; McGuirk, J.P. CAR T cells: Continuation in a revolution of immunotherapy. *Lancet Oncol.* **2020**, *21*, e168–e178. [CrossRef]
29. Wang, S.S.; Liu, W.; Ly, D.; Xu, H.; Qu, L.; Zhang, L. Tumor-infiltrating B cells: Their role and application in anti-tumor immunity in lung cancer. *Cell. Mol. Immunol.* **2019**, *16*, 6–18. [CrossRef]
30. Patel, A.J.; Richter, A.; Drayson, M.T.; Middleton, G.W. The role of B lymphocytes in the immuno-biology of non-small-cell lung cancer. *Cancer Immunol. Immunother. CII* **2020**, *69*, 325–342. [CrossRef]
31. Leong, T.L.; Bryant, V.L. B cells in lung cancer-not just a bystander cell: A literature review. *Transl. Lung Cancer Res.* **2021**, *10*, 2830–2841. [CrossRef]
32. Parikshak, N.N.; Gandal, M.J.; Geschwind, D.H. Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. *Nat. Rev. Genet.* **2015**, *16*, 441–458. [CrossRef] [PubMed]
33. Stuart, J.M.; Segal, E.; Koller, D.; Kim, S.K. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **2003**, *302*, 249–255. [CrossRef] [PubMed]
34. Lee, T.I.; Young, R.A. Transcriptional regulation and its misregulation in disease. *Cell* **2013**, *152*, 1237–1251. [CrossRef] [PubMed]
35. Emmert-Streib, F.; Dehmer, M.; Haibe-Kains, B. Gene regulatory networks and their applications: Understanding biological and medical problems in terms of networks. *Front. Cell Dev. Biol.* **2014**, *2*, 38. [CrossRef] [PubMed]
36. Singh, A.J.; Ramsey, S.A.; Filtz, T.M.; Kioussi, C. Differential gene regulatory networks in development and disease. *Cell. Mol. Life Sci.* **2018**, *75*, 1013–1025. [CrossRef] [PubMed]
37. Budczies, J.; Bockmayr, M.; Denkert, C.; Klauschen, F.; Gröschel, S.; Darb-Esfahani, S.; Pfarr, N.; Leichsenring, J.; Onozato, M.L.; Lennerz, J.K.; et al. Pan-cancer analysis of copy number changes in programmed death-ligand 1 (PD-L1, CD274)—Associations with gene expression, mutational load, and survival. *Genes Chromosom. Cancer* **2016**, *55*, 626–639. [CrossRef] [PubMed]
38. Kuenzi, B.M.; Ideker, T. A census of pathway maps in cancer systems biology. *Nat. Rev. Cancer* **2020**, *20*, 233–246. [CrossRef]
39. Abul-Husn, N.S.; Kenny, E.E. Personalized Medicine and the Power of Electronic Health Records. *Cell* **2019**, *177*, 58–69. [CrossRef]
40. Yao, L.; Zhang, Y.; Li, Y.; Sanseau, P.; Agarwal, P. Electronic health records: Implications for drug discovery. *Drug Discov. Today* **2011**, *16*, 594–599. [CrossRef]
41. Ashburn, T.T.; Thor, K.B. Drug repositioning: Identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discov.* **2004**, *3*, 673–683. [CrossRef]
42. Pushpakom, S.; Iorio, F.; Eyers, P.A.; Escott, K.J.; Hopper, S.; Wells, A.; Doig, A.; Guilliams, T.; Latimer, J.; McNamee, C.; et al. Drug repurposing: Progress, challenges and recommendations. *Nat. Rev. Drug Discov.* **2019**, *18*, 41–58. [CrossRef]

43. Van der Auwera, G.A.; O'Connor, B.D. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*, 1st ed.; O'Reilly Media: Sebastopol, CA, USA, 2020.
44. Su, Z.; Fang, H.; Hong, H.; Shi, L.; Zhang, W.; Zhang, W.; Zhang, Y.; Dong, Z.; Lancashire, L.J.; Bessarabova, M.; et al. An investigation of biomarkers derived from legacy microarray data for their utility in the RNA-seq era. *Genome Biol.* **2014**, *15*, 523. [CrossRef] [PubMed]
45. Freeman, J.L.; Perry, G.H.; Feuk, L.; Redon, R.; McCarroll, S.A.; Altshuler, D.M.; Aburatani, H.; Jones, K.W.; Tyler-Smith, C.; Hurles, M.E.; et al. Copy number variation: New insights in genome diversity. *Genome Res.* **2006**, *16*, 949–961. [CrossRef] [PubMed]
46. Redon, R.; Ishikawa, S.; Fitch, K.R.; Feuk, L.; Perry, G.H.; Andrews, T.D.; Fiegler, H.; Shapero, M.H.; Carson, A.R.; Chen, W.; et al. Global variation in copy number in the human genome. *Nature* **2006**, *444*, 444–454. [CrossRef] [PubMed]
47. Lauer, S.; Gresham, D. An evolving view of copy number variants. *Curr. Genet.* **2019**, *65*, 1287–1295. [CrossRef]
48. Ye, Q.; Singh, S.; Qian, P.R.; Guo, N.L. Immune-Omics Networks of CD27, PD1, and PDL1 in Non-Small Cell Lung Cancer. *Cancers* **2021**, *13*, 4296. [CrossRef]
49. Pös, O.; Radvanszky, J.; Buglyó, G.; Pös, Z.; Rusnakova, D.; Nagy, B.; Szemes, T. DNA copy number variation: Main characteristics, evolutionary significance, and pathological aspects. *Biomed. J.* **2021**, *44*, 548–559. [CrossRef]
50. de Ligt, J.; Boone, P.M.; Pfundt, R.; Vissers, L.E.; de Leeuw, N.; Shaw, C.; Brunner, H.G.; Lupski, J.R.; Veltman, J.A.; Hehir-Kwa, J.Y. Platform comparison of detecting copy number variants with microarrays and whole-exome sequencing. *Genom. Data* **2014**, *2*, 144–146. [CrossRef]
51. Wang, K.; Li, M.; Hadley, D.; Liu, R.; Glessner, J.; Grant, S.F.; Hakonarson, H.; Bucan, M. PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **2007**, *17*, 1665–1674. [CrossRef]
52. Vosse, S.; van de Wiel, M.A. *CGHbase: CGHbase: Base Functions and Classes for arrayCGH Data Analysis*; R Package Version 1.46.0; R Foundation for Statistical Computing: Vienna, Austria, 2021.
53. van de Wiel, M.A.; Vosse, S. *CGHcall: Calling Aberrations for Array CGH Tumor Profiles*; R Package Version 2.48.0; R Foundation for Statistical Computing: Vienna, Austria, 2021.
54. Mermel, C.H.; Schumacher, S.E.; Hill, B.; Meyerson, M.L.; Beroukhim, R.; Getz, G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **2011**, *12*, R41. [CrossRef]
55. Pedersen, B.S.; Yang, I.V.; De, S. CruzDB: Software for annotation of genomic intervals with UCSC genome-browser database. *Bioinformatics* **2013**, *29*, 3003–3006. [CrossRef] [PubMed]
56. Guo, N.L.; Dowlati, A.; Raese, R.A.; Dong, C.; Chen, G.; Beer, D.G.; Shaffer, J.; Singh, S.; Bokhary, U.; Liu, L.; et al. A Predictive 7-Gene Assay and Prognostic Protein Biomarkers for Non-small Cell Lung Cancer. *EBioMedicine* **2018**, *32*, 102–110. [CrossRef] [PubMed]
57. Zhan, C.; Zhang, Y.; Ma, J.; Wang, L.; Jiang, W.; Shi, Y.; Wang, Q. Identification of reference genes for qRT-PCR in human lung squamous-cell carcinoma by RNA-Seq. *Acta Biochim. Biophys. Sin.* **2014**, *46*, 330–337. [CrossRef] [PubMed]
58. Walter, R.F.; Werner, R.; Vollbrecht, C.; Hager, T.; Flom, E.; Christoph, D.C.; Schmeller, J.; Schmid, K.W.; Wohlschlaeger, J.; Mairinger, F.D. ACTB, CDKN1B, GAPDH, GRB2, RHOA and SDCBP Were Identified as Reference Genes in Neuroendocrine Lung Cancer via the nCounter Technology. *PLoS ONE* **2016**, *11*, e0165181. [CrossRef] [PubMed]
59. Saviozzi, S.; Cordero, F.; Lo Iacono, M.; Novello, S.; Scagliotti, G.V.; Calogero, R.A. Selection of suitable reference genes for accurate normalization of gene expression profile studies in non-small cell lung cancer. *BMC Cancer* **2006**, *6*, 200. [CrossRef]
60. Chang, Y.C.; Ding, Y.; Dong, L.; Zhu, L.J.; Jensen, R.V.; Hsiao, L.L. Differential expression patterns of housekeeping genes increase diagnostic and prognostic value in lung cancer. *PeerJ* **2018**, *6*, e4719. [CrossRef]
61. Ye, Q.; Falatovich, B.; Singh, S.; Ivanov, A.V.; Eubank, T.D.; Guo, N.L. A Multi-Omics Network of a Seven-Gene Prognostic Signature for Non-Small Cell Lung Cancer. *Int. J. Mol. Sci.* **2021**, *23*, 219. [CrossRef]
62. Tzeng, I.S. Modified Significance Analysis of Microarrays in Heterogeneous Diseases. *J. Pers. Med.* **2021**, *11*, 62. [CrossRef]
63. 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]
64. Ye, Q.; Hickey, J.; Summers, K.; Falatovich, B.; Gencheva, M.; Eubank, T.D.; Ivanov, A.V.; Guo, N.L. Multi-Omics Immune Interaction Networks in Lung Cancer Tumorigenesis, Proliferation, and Survival. *Int. J. Mol. Sci.* **2022**, *23*, 14978. [CrossRef]
65. Xu, J.Y.; Zhang, C.; Wang, X.; Zhai, L.; Ma, Y.; Mao, Y.; Qian, K.; Sun, C.; Liu, Z.; Jiang, S.; et al. Integrative Proteomic Characterization of Human Lung Adenocarcinoma. *Cell* **2020**, *182*, 245–261. [CrossRef] [PubMed]
66. Hodzic, E. Single-cell analysis: Advances and future perspectives. *Bosn. J. Basic Med. Sci.* **2016**, *16*, 313–314. [CrossRef] [PubMed]
67. Saadatpour, A.; Lai, S.; Guo, G.; Yuan, G.-C. Single-Cell Analysis in Cancer Genomics. *Trends Genet.* **2015**, *31*, 576–586. [CrossRef] [PubMed]
68. Garvin, T.; Aboukhalil, R.; Kendall, J.; Baslan, T.; Atwal, G.S.; Hicks, J.; Wigler, M.; Schatz, M.C. Interactive analysis and assessment of single-cell copy-number variations. *Nat. Methods* **2015**, *12*, 1058–1060. [CrossRef]
69. Kanter, I.; Kalisky, T. Single cell transcriptomics: Methods and applications. *Front. Oncol.* **2015**, *5*, 53. [CrossRef]
70. Chen, C.; Xing, D.; Tan, L.; Li, H.; Zhou, G.; Huang, L.; Xie, X.S. Single-cell whole-genome analyses by Linear Amplification via Transposon Insertion (LIANTI). *Science* **2017**, *356*, 189–194. [CrossRef]

71. Liao, Y.; Smyth, G.K.; Shi, W. featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **2014**, *30*, 923–930. [CrossRef]
72. Dobin, A.; Davis, C.A.; Schlesinger, F.; Drenkow, J.; Zaleski, C.; Jha, S.; Batut, P.; Chaisson, M.; Gingeras, T.R. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **2013**, *29*, 15–21. [CrossRef]
73. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; Genome Project Data Processing, S. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [CrossRef]
74. Ye, Q.; Guo, N.L. Single B Cell Gene Co-Expression Networks Implicated in Prognosis, Proliferation, and Therapeutic Responses in Non-Small Cell Lung Cancer Bulk Tumors. *Cancers* **2022**, *14*, 3123. [CrossRef] [PubMed]
75. Miao, Z.; Deng, K.; Wang, X.; Zhang, X. DEsingle for detecting three types of differential expression in single-cell RNA-seq data. *Bioinformatics* **2018**, *34*, 3223–3224. [CrossRef] [PubMed]
76. Greene, C.S.; Voight, B.F. Pathway and network-based strategies to translate genetic discoveries into effective therapies. *Hum. Mol. Genet.* **2016**, *25*, R94–R98. [CrossRef] [PubMed]
77. 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]
78. Chen, J.; Bardes, E.E.; Aronow, B.J.; Jegga, A.G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* **2009**, *37*, W305–W311. [CrossRef] [PubMed]
79. Kanehisa, M.; Goto, S.; Hattori, M.; oki-Kinoshita, K.F.; Itoh, M.; Kawashima, S.; Katayama, T.; Araki, M.; Hirakawa, M. From genomics to chemical genomics: New developments in KEGG. *Nucleic Acids Res.* **2006**, *34*, D354–D357. [CrossRef]
80. Kanehisa, M.; Goto, S.; Kawashima, S.; Nakaya, A. The KEGG databases at GenomeNet. *Nucleic Acids Res.* **2002**, *30*, 42–46.
81. Creating the gene ontology resource: Design and implementation. *Genome Res.* **2001**, *11*, 1425–1433. [CrossRef]
82. Agarwal, V.; Bell, G.W.; Nam, J.W.; Bartel, D.P. Predicting effective microRNA target sites in mammalian mRNAs. *eLife* **2015**, *4*, e05005. [CrossRef]
83. Karagkouni, D.; Paraskevopoulou, M.D.; Chatzopoulos, S.; Vlachos, I.S.; Tastsoglou, S.; Kanellos, I.; Papadimitriou, D.; Kavakiotis, I.; Maniou, S.; Skoufos, G.; et al. DIANA-TarBase v8: A decade-long collection of experimentally supported miRNA-gene interactions. *Nucleic Acids Res.* **2018**, *46*, D239–D245. [CrossRef]
84. Meyers, R.M.; Bryan, J.G.; McFarland, J.M.; Weir, B.A.; Sizemore, A.E.; Xu, H.; Dharia, N.V.; Montgomery, P.G.; Cowley, G.S.; Pantel, S.; et al. Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nat. Genet.* **2017**, *49*, 1779–1784. [CrossRef]
85. McFarland, J.M.; Ho, Z.V.; Kugener, G.; Dempster, J.M.; Montgomery, P.G.; Bryan, J.G.; Krill-Burger, J.M.; Green, T.M.; Vazquez, F.; Boehm, J.S.; et al. Improved estimation of cancer dependencies from large-scale RNAi screens using model-based normalization and data integration. *Nat. Commun.* **2018**, *9*, 4610. [CrossRef] [PubMed]
86. Chu, V.T.; Graf, R.; Wirtz, T.; Weber, T.; Favret, J.; Li, X.; Petsch, K.; Tran, N.T.; Sieweke, M.H.; Berek, C.; et al. Efficient CRISPR-mediated mutagenesis in primary immune cells using CrispRGold and a C57BL/6 Cas9 transgenic mouse line. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 12514–12519. [CrossRef] [PubMed]
87. Johnson, M.J.; Laoharawee, K.; Lahr, W.S.; Webber, B.R.; Moriarity, B.S. Engineering of Primary Human B cells with CRISPR/Cas9 Targeted Nuclease. *Sci. Rep.* **2018**, *8*, 12144. [CrossRef] [PubMed]
88. Azangou-Khyavy, M.; Ghasemi, M.; Khanali, J.; Boroomand-Saboor, M.; Jamalkhah, M.; Soleimani, M.; Kiani, J. CRISPR/Cas: From Tumor Gene Editing to T Cell-Based Immunotherapy of Cancer. *Front. Immunol.* **2020**, *11*, 2062. [CrossRef]
89. Razeghian, E.; Nasution, M.K.M.; Rahman, H.S.; Gardanova, Z.R.; Abdelbasset, W.K.; Aravindhan, S.; Bokov, D.O.; Suksatan, W.; Nakhaei, P.; Shariatzadeh, S.; et al. A deep insight into CRISPR/Cas9 application in CAR-T cell-based tumor immunotherapies. *Stem Cell Res. Ther.* **2021**, *12*, 428. [CrossRef]
90. Yoshihara, K.; Shahmoradgoli, M.; Martinez, E.; Vegesna, R.; Kim, H.; Torres-Garcia, W.; Trevino, V.; Shen, H.; Laird, P.W.; Levine, D.A.; et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **2013**, *4*, 2612. [CrossRef]
91. Aran, D.; Hu, Z.; Butte, A.J. xCell: Digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.* **2017**, *18*, 220. [CrossRef]
92. Li, T.; Fu, J.; Zeng, Z.; Cohen, D.; Li, J.; Chen, Q.; Li, B.; Liu, X.S. TIMER2.0 for analysis of tumor-infiltrating immune cells. *Nucleic Acids Res.* **2020**, *48*, W509–W514. [CrossRef]
93. Li, T.; Fan, J.; Wang, B.; Traugh, N.; Chen, Q.; Liu, J.S.; Li, B.; Liu, X.S. TIMER: A Web Server for Comprehensive Analysis of Tumor-Infiltrating Immune Cells. *Cancer Res.* **2017**, *77*, e108–e110. [CrossRef]
94. Li, B.; Severson, E.; Pignon, J.C.; Zhao, H.; Li, T.; Novak, J.; Jiang, P.; Shen, H.; Aster, J.C.; Rodig, S.; et al. Comprehensive analyses of tumor immunity: Implications for cancer immunotherapy. *Genome Biol.* **2016**, *17*, 174. [CrossRef]
95. Chen, B.; Khodadoust, M.S.; Liu, C.L.; Newman, A.M.; Alizadeh, A.A. Profiling Tumor Infiltrating Immune Cells with CIBERSORT. *Methods Mol. Biol.* **2018**, *1711*, 243–259. [CrossRef] [PubMed]
96. Becht, E.; Giraldo, N.A.; Lacroix, L.; Buttard, B.; Elarouci, N.; Petitprez, F.; Selves, J.; Laurent-Puig, P.; Sautès-Fridman, C.; Fridman, W.H.; et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol.* **2016**, *17*, 218. [CrossRef] [PubMed]

97. Corsello, S.M.; Nagari, R.T.; Spangler, R.D.; Rossen, J.; Kocak, M.; Bryan, J.G.; Humeidi, R.; Peck, D.; Wu, X.; Tang, A.A.; et al. Discovering the anti-cancer potential of non-oncology drugs by systematic viability profiling. *Nat. Cancer* **2020**, *1*, 235–248. [CrossRef] [PubMed]
98. Yang, W.; Soares, J.; Greninger, P.; Edelman, E.J.; Lightfoot, H.; Forbes, S.; Bindal, N.; Beare, D.; Smith, J.A.; Thompson, I.R.; et al. Genomics of Drug Sensitivity in Cancer (GDSC): A resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.* **2013**, *41*, D955–D961. [CrossRef]
99. Iorio, F.; Knijnenburg, T.A.; Vis, D.J.; Bignell, G.R.; Menden, M.P.; Schubert, M.; Aben, N.; Gonçalves, E.; Barthorpe, S.; Lightfoot, H.; et al. A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* **2016**, *166*, 740–754. [CrossRef]
100. Garnett, M.J.; Edelman, E.J.; Heidorn, S.J.; Greenman, C.D.; Dastur, A.; Lau, K.W.; Greninger, P.; Thompson, I.R.; Luo, X.; Soares, J.; et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **2012**, *483*, 570–575. [CrossRef]
101. Ma, Y.; Ding, Z.; Qian, Y.; Shi, X.; Castranova, V.; Harner, E.J.; Guo, L. Predicting cancer drug response by proteomic profiling. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **2006**, *12*, 4583–4589. [CrossRef]
102. Ma, Y.; Ding, Z.; Qian, Y.; Wan, Y.W.; Tosun, K.; Shi, X.; Castranova, V.; Harner, E.J.; Guo, N.L. An integrative genomic and proteomic approach to chemosensitivity prediction. *Int. J. Oncol.* **2009**, *34*, 107–115. [CrossRef]
103. Schwartz, L.H.; Litière, S.; de Vries, E.; Ford, R.; Gwyther, S.; Mandrekar, S.; Shankar, L.; Bogaerts, J.; Chen, A.; Dancey, J.; et al. RECIST 1.1-Update and clarification: From the RECIST committee. *Eur. J. Cancer* **2016**, *62*, 132–137. [CrossRef]
104. Greene, C.S.; Krishnan, A.; Wong, A.K.; Ricciotti, E.; Zelaya, R.A.; Himmelstein, D.S.; Zhang, R.; Hartmann, B.M.; Zaslavsky, E.; Sealfon, S.C.; et al. Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.* **2015**, *47*, 569–576. [CrossRef]
105. Iorio, F.; Saez-Rodriguez, J.; di Bernardo, D. Network based elucidation of drug response: From modulators to targets. *BMC Syst. Biol.* **2013**, *7*, 139. [CrossRef] [PubMed]
106. Eisen, M.B.; Spellman, P.T.; Brown, P.O.; Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 14863–14868. [CrossRef] [PubMed]
107. Langfelder, P.; Horvath, S. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinform.* **2008**, *9*, 559. [CrossRef] [PubMed]
108. Zoppi, J.; Guillaume, J.F.; Neunlist, M.; Chaffron, S. MiBiOmics: An interactive web application for multi-omics data exploration and integration. *BMC Bioinform.* **2021**, *22*, 6. [CrossRef]
109. Zhou, G.; Ewald, J.; Xia, J. OmicsAnalyst: A comprehensive web-based platform for visual analytics of multi-omics data. *Nucleic Acids Res.* **2021**, *49*, W476–W482. [CrossRef]
110. Yu, S.; Drton, M.; Promislow, D.E.L.; Shojaie, A. CorDiffViz: An R package for visualizing multi-omics differential correlation networks. *BMC Bioinform.* **2021**, *22*, 486. [CrossRef]
111. Faith, J.J.; Hayete, B.; Thaden, J.T.; Mogno, I.; Wierzbowski, J.; Cottarel, G.; Kasif, S.; Collins, J.J.; Gardner, T.S. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol.* **2007**, *5*, e8. [CrossRef]
112. Wille, A.; Bühlmann, P. Low-order conditional independence graphs for inferring genetic networks. *Stat. Appl. Genet. Mol. Biol.* **2006**, *5*. [CrossRef]
113. Zuo, Y.; Yu, G.; Tadesse, M.G.; Ressom, H.W. Reconstructing biological networks using low order partial correlation. In Proceedings of the 2013 IEEE International Conference on Bioinformatics and Biomedicine, Shanghai, China, 18–21 December 2013; pp. 171–175.
114. Butte, A.J.; Kohane, I.S. Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements. In *Biocomputing 2000*; World Scientific: Singapore, 1999; pp. 418–429.
115. Altay, G.; Emmert-Streib, F. Revealing differences in gene network inference algorithms on the network level by ensemble methods. *Bioinformatics* **2010**, *26*, 1738–1744. [CrossRef]
116. Emmert-Streib, F.; Glazko, G.V.; Altay, G.; de Matos Simoes, R. Statistical inference and reverse engineering of gene regulatory networks from observational expression data. *Front. Genet.* **2012**, *3*, 8. [CrossRef]
117. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
118. Zhang, X.; Zhao, X.-M.; He, K.; Lu, L.; Cao, Y.; Liu, J.; Hao, J.-K.; Liu, Z.-P.; Chen, L. Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. *Bioinformatics* **2012**, *28*, 98–104. [CrossRef] [PubMed]
119. Butte, A.J.; Tamayo, P.; Slonim, D.; Golub, T.R.; Kohane, I.S. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 12182–12186. [CrossRef]
120. Kinney, J.B.; Atwal, G.S. Equitability, mutual information, and the maximal information coefficient. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 3354–3359. [CrossRef] [PubMed]
121. Margolin, A.A.; Nemenman, I.; Basso, K.; Wiggins, C.; Stolovitzky, G.; Favera, R.D.; Califano, A. ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinform.* **2006**, *7*, S7. [CrossRef] [PubMed]
122. Brunel, H.; Gallardo-Chacón, J.-J.; Buil, A.; Vallverdú, M.; Soria, J.M.; Caminal, P.; Perera, A. MISS: A non-linear methodology based on mutual information for genetic association studies in both population and sib-pairs analysis. *Bioinformatics* **2010**, *26*, 1811–1818. [CrossRef]

123. Zhang, X.; Liu, K.; Liu, Z.-P.; Duval, B.; Richer, J.-M.; Zhao, X.-M.; Hao, J.-K.; Chen, L. NARROMI: A noise and redundancy reduction technique improves accuracy of gene regulatory network inference. *Bioinformatics* **2013**, *29*, 106–113. [CrossRef]
124. Honkela, A.; Girardot, C.; Gustafson, E.H.; Liu, Y.-H.; Furlong, E.E.; Lawrence, N.D.; Rattray, M. Model-based method for transcription factor target identification with limited data. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 7793–7798. [CrossRef]
125. Belcastro, V.; Siciliano, V.; Gregoretti, F.; Mithbaokar, P.; Dharmalingam, G.; Berlingieri, S.; Iorio, F.; Oliva, G.; Polishchuck, R.; Brunetti-Pierri, N. Transcriptional gene network inference from a massive dataset elucidates transcriptome organization and gene function. *Nucleic Acids Res.* **2011**, *39*, 8677–8688. [CrossRef]
126. Treviño III, S.; Sun, Y.; Cooper, T.F.; Bassler, K.E. Robust detection of hierarchical communities from Escherichia coli gene expression data. *PLoS Comput. Biol.* **2012**, *8*, e1002391. [CrossRef]
127. Xiao, F.; Gao, L.; Ye, Y.; Hu, Y.; He, R. Inferring gene regulatory networks using conditional regulation pattern to guide candidate genes. *PLoS ONE* **2016**, *11*, e0154953. [CrossRef] [PubMed]
128. Zhang, X.; Zhao, J.; Hao, J.-K.; Zhao, X.-M.; Chen, L. Conditional mutual inclusive information enables accurate quantification of associations in gene regulatory networks. *Nucleic Acids Res.* **2015**, *43*, e31. [CrossRef] [PubMed]
129. Meyer, P.E.; Kontos, K.; Lafitte, F.; Bontempi, G. Information-theoretic inference of large transcriptional regulatory networks. *EURASIP J. Bioinform. Syst. Biol.* **2007**, *2007*, 79879. [CrossRef]
130. Altay, G.; Emmert-Streib, F. Inferring the conservative causal core of gene regulatory networks. *BMC Syst. Biol.* **2010**, *4*, 132. [CrossRef] [PubMed]
131. Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1988.
132. Murphy, K.; Mian, S. *Modelling Gene Expression Data Using Dynamic Bayesian Networks*; Technical Report; Computer Science Division, University of California: Berkeley, CA, USA, 1999.
133. Friedman, N.; Linial, M.; Nachman, I.; Pe'er, D. Using Bayesian networks to analyze expression data. *J. Comput. Biol.* **2000**, *7*, 601–620. [CrossRef] [PubMed]
134. Hartemink, A.J.; Gifford, D.K.; Jaakkola, T.S.; Young, R.A. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. In Proceedings of the Pacific Symposium on Biocomputing, Mauna Lani, HI, USA, 3–7 January 2001; pp. 422–433. [CrossRef]
135. Chai, L.E.; Loh, S.K.; Low, S.T.; Mohamad, M.S.; Deris, S.; Zakaria, Z. A review on the computational approaches for gene regulatory network construction. *Comput. Biol. Med.* **2014**, *48*, 55–65. [CrossRef]
136. Kaderali, L.; Radde, N. Inferring Gene Regulatory Networks from Expression Data. *Comput. Intell. Bioinform.* **2008**, *94*, 33–74. [CrossRef]
137. Liu, Z.P. Reverse Engineering of Genome-wide Gene Regulatory Networks from Gene Expression Data. *Curr. Genom.* **2015**, *16*, 3–22. [CrossRef]
138. Bøtcher, S.G.; Dethlefsen, C. deal: A Package for Learning Bayesian Networks. *J. Stat. Softw.* **2003**, *8*, 1–40. [CrossRef]
139. Chickering, M.; Heckerman, D.; Meek, C. Large-sample learning of Bayesian networks is NP-hard. *J. Mach. Learn. Res.* **2004**, *5*, 1287–1330.
140. Hill, S.M.; Heiser, L.M.; Cokelaer, T.; Unger, M.; Nesser, N.K.; Carlin, D.E.; Zhang, Y.; Sokolov, A.; Paull, E.O.; Wong, C.K.; et al. Inferring causal molecular networks: Empirical assessment through a community-based effort. *Nat. Methods* **2016**, *13*, 310–318. [CrossRef] [PubMed]
141. De Campos, C.P.; Ji, Q. Efficient structure learning of Bayesian networks using constraints. *J. Mach. Learn. Res.* **2011**, *12*, 663–689.
142. Liu, F.; Zhang, S.-W.; Guo, W.-F.; Wei, Z.-G.; Chen, L. Inference of gene regulatory network based on local Bayesian networks. *PLoS Comput. Biol.* **2016**, *12*, e1005024. [CrossRef]
143. Zhang, Y.; Deng, Z.; Jiang, H.; Jia, P. Inferring Gene Regulatory Networks from Multiple Data Sources Via a Dynamic Bayesian Network with Structural EM. In Proceedings of the International Conference on Data Integration in the Life Sciences, Philadelphia, PA, USA, 27–29 June 2007; Volume 4544, pp. 204–214. [CrossRef]
144. Sato, N.; Tamada, Y.; Yu, G.; Okuno, Y. CBNplot: Bayesian network plots for enrichment analysis. *Bioinformatics* **2022**, *38*, 2959–2960. [CrossRef]
145. Howey, R.; Shin, S.Y.; Relton, C.; Davey Smith, G.; Cordell, H.J. Bayesian network analysis incorporating genetic anchors complements conventional Mendelian randomization approaches for exploratory analysis of causal relationships in complex data. *PLoS Genet.* **2020**, *16*, e1008198. [CrossRef]
146. Howey, R.; Clark, A.D.; Naamane, N.; Reynard, L.N.; Pratt, A.G.; Cordell, H.J. A Bayesian network approach incorporating imputation of missing data enables exploratory analysis of complex causal biological relationships. *PLoS Genet.* **2021**, *17*, e1009811. [CrossRef]
147. Taylor, I.W.; Linding, R.; Warde-Farley, D.; Liu, Y.; Pesquita, C.; Faria, D.; Bull, S.; Pawson, T.; Morris, Q.; Wrana, J.L. Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat. Biotechnol.* **2009**, *27*, 199–204. [CrossRef] [PubMed]
148. Perrin, B.-E.; Ralaivola, L.; Mazurie, A.; Bottani, S.; Mallet, J.; d'Alche–Buc, F. Gene networks inference using dynamic Bayesian networks. *Bioinformatics* **2003**, *19*, ii138–ii148. [CrossRef]
149. Zou, M.; Conzen, S.D. A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics* **2005**, *21*, 71–79. [CrossRef]

150. Smith, V.A.; Yu, J.; Smulders, T.V.; Hartemink, A.J.; Jarvis, E.D. Computational inference of neural information flow networks. *PLoS Comput. Biol.* **2006**, *2*, e161. [CrossRef]
151. Song, L.; Kolar, M.; Xing, E.P. KELLER: Estimating time-varying interactions between genes. *Bioinformatics* **2009**, *25*, i128–i136. [CrossRef] [PubMed]
152. Bansal, M.; Belcastro, V.; Ambesi-Impiombato, A.; Di Bernardo, D. How to infer gene networks from expression profiles. *Mol. Syst. Biol.* **2007**, *3*, 78. [CrossRef] [PubMed]
153. Yeung, M.S.; Tegnér, J.; Collins, J.J. Reverse engineering gene networks using singular value decomposition and robust regression. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 6163–6168. [CrossRef] [PubMed]
154. Heinrich, R.; Schuster, S. *The Regulation of Cellular Systems*; Springer Science & Business Media: New York, NY, USA, 2012.
155. Sakamoto, E.; Iba, H. Inferring a system of differential equations for a gene regulatory network by using genetic programming. In Proceedings of the 2001 Congress on Evolutionary Computation (IEEE Cat. No. 01TH8546), Seoul, Republic of Korea, 27–30 May 2001; pp. 720–726.
156. Spieth, C.; Hassis, N.; Streichert, F. Comparing mathematical models on the problem of network inference. In Proceedings of the 8th Annual Conference on Genetic and Evolutionary Computation, Seattle, WA, USA, 8–12 July 2006; pp. 279–286.
157. Kauffman, S.A. Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol.* **1969**, *22*, 437–467. [CrossRef]
158. Graudenzi, A.; Serra, R.; Villani, M.; Damiani, C.; Colacci, A.; Kauffman, S.A. Dynamical properties of a Boolean model of gene regulatory network with memory. *J. Comput. Biol.* **2011**, *18*, 1291–1303. [CrossRef]
159. Thomas, R. Boolean formalization of genetic control circuits. *J. Theor. Biol.* **1973**, *42*, 563–585. [CrossRef]
160. Wang, R.-S.; Saadatpour, A.; Albert, R. Boolean modeling in systems biology: An overview of methodology and applications. *Phys. Biol.* **2012**, *9*, 055001. [CrossRef]
161. Maheshri, N.; O'Shea, E.K. Living with noisy genes: How cells function reliably with inherent variability in gene expression. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 413–434. [CrossRef]
162. Liang, S.; Fuhrman, S.; Somogyi, R. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pac. Symp. Biocomput.* **1998**, *3*, 18–29.
163. Kim, H.; Lee, J.K.; Park, T. Boolean networks using the chi-square test for inferring large-scale gene regulatory networks. *BMC Bioinform.* **2007**, *8*, 37. [CrossRef]
164. Lähdesmäki, H.; Hautaniemi, S.; Shmulevich, I.; Yli-Harja, O. Relationships between probabilistic Boolean networks and dynamic Bayesian networks as models of gene regulatory networks. *Signal Process.* **2006**, *86*, 814–834. [CrossRef] [PubMed]
165. Barman, S.; Kwon, Y.-K. A novel mutual information-based Boolean network inference method from time-series gene expression data. *PLoS ONE* **2017**, *12*, e0171097. [CrossRef] [PubMed]
166. Woodhouse, S.; Piterman, N.; Wintersteiger, C.M.; Gottgens, B.; Fisher, J. SCNS: A graphical tool for reconstructing executable regulatory networks from single-cell genomic data. *BMC Syst. Biol.* **2018**, *12*, 59. [CrossRef] [PubMed]
167. Sahoo, D.; Dill, D.L.; Gentles, A.J.; Tibshirani, R.; Plevritis, S.K. Boolean implication networks derived from large scale, whole genome microarray datasets. *Genome Biol.* **2008**, *9*, R157. [CrossRef]
168. Sahoo, D.; Dill, D.L.; Tibshirani, R.; Plevritis, S.K. Extracting binary signals from microarray time-course data. *Nucleic Acids Res.* **2007**, *35*, 3705–3712. [CrossRef]
169. Sinha, S.; Tsang, E.K.; Zeng, H.; Meister, M.; Dill, D.L. Mining TCGA data using Boolean implications. *PLoS ONE* **2014**, *9*, e102119. [CrossRef]
170. Çakır, M.V.; Binder, H.; Wirth, H. Profiling of genetic switches using boolean implications in expression data. *J. Integr. Bioinform.* **2014**, *11*, 246. [CrossRef]
171. Guo, N.L.; Wan, Y.W. Pathway-based identification of a smoking associated 6-gene signature predictive of lung cancer risk and survival. *Artif. Intell. Med.* **2012**, *55*, 97–105. [CrossRef]
172. Guo, N.L.; Wan, Y.W.; Bose, S.; Denvir, J.; Kashon, M.L.; Andrew, M.E. A novel network model identified a 13-gene lung cancer prognostic signature. *Int. J. Comput. Biol. Drug Des.* **2011**, *4*, 19–39. [CrossRef]
173. Hildebrand, D.K.; Laing, J.D.; Rosenthal, H. *Prediction Analysis of Cross Classifications*; John Wiley & Sons: New York, NY, USA, 1977.
174. Park, Y.; Kellis, M. Deep learning for regulatory genomics. *Nat. Biotechnol.* **2015**, *33*, 825–826. [CrossRef]
175. Singh, S.; Yang, Y.; Póczos, B.; Ma, J. Predicting enhancer-promoter interaction from genomic sequence with deep neural networks. *Quant. Biol.* **2019**, *7*, 122–137. [CrossRef] [PubMed]
176. Zou, J.; Huss, M.; Abid, A.; Mohammadi, P.; Torkamani, A.; Telenti, A. A primer on deep learning in genomics. *Nat. Genet.* **2019**, *51*, 12–18. [CrossRef] [PubMed]
177. Webb, S. Deep learning for biology. *Nature* **2018**, *554*, 555–558. [CrossRef]
178. Ching, T.; Himmelstein, D.S.; Beaulieu-Jones, B.K.; Kalinin, A.A.; Do, B.T.; Way, G.P.; Ferrero, E.; Agapow, P.-M.; Zietz, M.; Hoffman, M.M. Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* **2018**, *15*, 20170387. [CrossRef]
179. Alipanahi, B.; Delong, A.; Weirauch, M.T.; Frey, B.J. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **2015**, *33*, 831–838. [CrossRef]

180. Kordmahalleh, M.M.; Sefidmazgi, M.G.; Harrison, S.H.; Homaifar, A. Identifying time-delayed gene regulatory networks via an evolvable hierarchical recurrent neural network. *BioData Min.* **2017**, *10*, 1–25. [CrossRef] [PubMed]
181. Hamilton, W.; Ying, Z.; Leskovec, J. Inductive representation learning on large graphs. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–11.
182. Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y. Graph attention networks. *Stat* **2017**, *1050*, 20.
183. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv* **2016**, arXiv:1609.02907.
184. Wang, J.; Ma, A.; Ma, Q.; Xu, D.; Joshi, T. Inductive inference of gene regulatory network using supervised and semi-supervised graph neural networks. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 3335–3343. [CrossRef]
185. Wang, J.; Ma, A.; Chang, Y.; Gong, J.; Jiang, Y.; Qi, R.; Wang, C.; Fu, H.; Ma, Q.; Xu, D. scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses. *Nat. Commun.* **2021**, *12*, 1882. [CrossRef] [PubMed]
186. Leng, D.; Zheng, L.; Wen, Y.; Zhang, Y.; Wu, L.; Wang, J.; Wang, M.; Zhang, Z.; He, S.; Bo, X. A benchmark study of deep learning-based multi-omics data fusion methods for cancer. *Genome Biol.* **2022**, *23*, 171. [CrossRef] [PubMed]
187. Sherif, F.F.; Zayed, N.; Fakhr, M. Discovering Alzheimer genetic biomarkers using Bayesian networks. *Adv. Bioinform.* **2015**, *2015*, 639367. [CrossRef] [PubMed]
188. Dridi, N.; Giremus, A.; Giovannelli, J.-F.; Truntzer, C.; Hadzagic, M.; Charrier, J.-P.; Gerfault, L.; Ducoroy, P.; Lacroix, B.; Grangeat, P. Bayesian inference for biomarker discovery in proteomics: An analytic solution. *EURASIP J. Bioinform. Syst. Biol.* **2017**, *2017*, 1–14. [CrossRef]
189. Liu, F.; Aulin, L.; Kossen, S.S.; Cathalina, J.; Bremmer, M.; Foks, A.C.; van der Graaf, P.H.; Moerland, M.; van Hasselt, J.G. A system pharmacology Boolean network model for the TLR4-mediated inflammatory response in early sepsis. *J. Pharmacokinet. Pharmacodyn.* **2022**, *49*, 645–655. [CrossRef] [PubMed]
190. Villoslada, P.; Baranzini, S. Data integration and systems biology approaches for biomarker discovery: Challenges and opportunities for multiple sclerosis. *J. Neuroimmunol.* **2012**, *248*, 58–65. [CrossRef]
191. Sun, X.; Hu, B. Mathematical modeling and computational prediction of cancer drug resistance. *Brief. Bioinform.* **2018**, *19*, 1382–1399. [CrossRef] [PubMed]
192. Zafeiris, D.; Rutella, S.; Ball, G.R. An artificial neural network integrated pipeline for biomarker discovery using Alzheimer's disease as a case study. *Comput. Struct. Biotechnol. J.* **2018**, *16*, 77–87. [CrossRef]
193. Moteghaed, N.Y.; Maghooli, K.; Pirhadi, S.; Garshasbi, M. Biomarker discovery based on hybrid optimization algorithm and artificial neural networks on microarray data for cancer classification. *J. Med. Signals Sens.* **2015**, *5*, 88.
194. Gendelman, R.; Xing, H.; Mirzoeva, O.K.; Sarde, P.; Curtis, C.; Feiler, H.S.; McDonagh, P.; Gray, J.W.; Khalil, I.; Korn, W.M. Bayesian network inference modeling identifies TRIB1 as a novel regulator of cell-cycle progression and survival in cancer cells. *Cancer Res.* **2017**, *77*, 1575–1585. [CrossRef]
195. Biane, C.; Delaplace, F. Abduction based drug target discovery using Boolean control network. In Proceedings of the International Conference on Computational Methods in Systems Biology, Darmstadt, Germany, 27–29 September 2017; pp. 57–73.
196. Vo, D.T.; Ghosh, P.; Sahoo, D. Artificial Intelligence Guided Discovery of Gastric Cancer Continuum. *bioRxiv* **2022**. [CrossRef]
197. Zhang, M.; Li, Q.; Yu, D.; Yao, B.; Guo, W.; Xie, Y.; Xiao, G. GeNeCK: A web server for gene network construction and visualization. *BMC Bioinform.* **2019**, *20*, 12. [CrossRef] [PubMed]
198. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H. The WEKA data mining software. *ACM SIGKDD Explor. Newsl.* **2009**, *11*, 10–18. [CrossRef]
199. Wan, Y.W.; Beer, D.G.; Guo, N.L. Signaling pathway-based identification of extensive prognostic gene signatures for lung adenocarcinoma. *Lung Cancer* **2012**, *76*, 98–105. [CrossRef] [PubMed]
200. Wan, Y.W.; Raese, R.A.; Fortney, J.E.; Xiao, C.; Luo, D.; Cavendish, J.; Gibson, L.F.; Castranova, V.; Qian, Y.; Guo, N.L. A smoking-associated 7-gene signature for lung cancer diagnosis and prognosis. *Int. J. Oncol.* **2012**, *41*, 1387–1396. [CrossRef] [PubMed]
201. Guo, N.L.; Wan, Y.W. Network-based identification of biomarkers coexpressed with multiple pathways. *Cancer Inform.* **2014**, *13*, 37–47. [CrossRef]
202. Ucar, D.; Neuhaus, I.; Ross-MacDonald, P.; Tilford, C.; Parthasarathy, S.; Siemers, N.; Ji, R.R. Construction of a reference gene association network from multiple profiling data: Application to data analysis. *Bioinformatics* **2007**, *23*, 2716–2724. [CrossRef]
203. Shedden, K.; Taylor, J.M.; Enkemann, S.A.; Tsao, M.S.; Yeatman, T.J.; Gerald, W.L.; Eschrich, S.; Jurisica, I.; Giordano, T.J.; Misek, D.E.; et al. Gene expression-based survival prediction in lung adenocarcinoma: A multi-site, blinded validation study. *Nat. Med.* **2008**, *14*, 822–827.
204. Wang, Z.; Wei, Y.; Zhang, R.; Su, L.; Gogarten, S.M.; Liu, G.; Brennan, P.; Field, J.K.; McKay, J.D.; Lissowska, J.; et al. Multi-Omics Analysis Reveals a HIF Network and Hub Gene EPAS1 Associated with Lung Adenocarcinoma. *EBioMedicine* **2018**, *32*, 93–101. [CrossRef]
205. Fan, Y.; Kao, C.; Yang, F.; Wang, F.; Yin, G.; Wang, Y.; He, Y.; Ji, J.; Liu, L. Integrated Multi-Omics Analysis Model to Identify Biomarkers Associated With Prognosis of Breast Cancer. *Front. Oncol.* **2022**, *12*, 899900. [CrossRef]
206. Krackhardt, D. Assessing the Political Landscape: Structure, Cognition, and Power in Organizations. *Adm. Sci. Q.* **1990**, *35*, 342–369. [CrossRef]
207. Bonacich, P.; Lloyd, P. Eigenvector centrality and structural zeroes and ones: When is a neighbor not a neighbor? *Soc. Netw.* **2015**, *43*, 86–90. [CrossRef]

208. Aguirre, J.; Papo, D.; Buldú, J.M. Successful strategies for competing networks. *Nat. Phys.* **2013**, *9*, 230–234. [CrossRef]
209. Bonacich, P.; Lloyd, P. Eigenvector-Like Measures of Centrality for Asymmetric Relations. *Soc. Netw.* **2001**, *23*, 191–201. [CrossRef]
210. Freeman, L.C. Centrality in social networks conceptual clarification. *Soc. Netw.* **1978**, *1*, 215–239. [CrossRef]
211. Brandes, U. A faster algorithm for betweenness centrality. *J. Math. Sociol.* **2001**, *25*, 163–177. [CrossRef]
212. Bolland, J.M. Sorting out centrality: An analysis of the performance of four centrality models in real and simulated networks. *Soc. Netw.* **1988**, *10*, 233–253. [CrossRef]
213. Brandes, U.; Hildenbrand, J.A.N. Smallest graphs with distinct singleton centers. *Netw. Sci.* **2014**, *2*, 416–418. [CrossRef]
214. Coscia, M. The Atlas for the Aspiring Network Scientist. *arXiv* **2021**. [CrossRef]
215. Zhang, J.-X.; Chen, D.-B.; Dong, Q.; Zhao, Z.-D. Identifying a set of influential spreaders in complex networks. *Sci. Rep.* **2016**, *6*, 27823. [CrossRef]
216. Schoch, D.; Valente, T.W.; Brandes, U. Correlations among centrality indices and a class of uniquely ranked graphs. *Soc. Netw.* **2017**, *50*, 46–54. [CrossRef]
217. Hagberg, A.; Swart, P.; Chult, D.S. *Exploring Network Structure, Dynamics, and Function Using NetworkX*; Los Alamos National Lab. (LANL): Los Alamos, NM, USA, 2008.
218. Ye, Q.; Guo, N.L. Hub Genes in Non-Small Cell Lung Cancer Regulatory Networks. *Biomolecules* **2022**, *12*, 1782. [CrossRef] [PubMed]
219. Lizotte, P.H.; Ivanova, E.V.; Awad, M.M.; Jones, R.E.; Keogh, L.; Liu, H.; Dries, R.; Almonte, C.; Herter-Sprie, G.S.; Santos, A.; et al. Multiparametric profiling of non-small-cell lung cancers reveals distinct immunophenotypes. *JCI Insight* **2016**, *1*, e89014. [CrossRef] [PubMed]
220. Chiou, S.H.; Tseng, D.; Reuben, A.; Mallajosyula, V.; Molina, I.S.; Conley, S.; Wilhelmy, J.; McSween, A.M.; Yang, X.; Nishimiya, D.; et al. Global analysis of shared T cell specificities in human non-small cell lung cancer enables HLA inference and antigen discovery. *Immunity* **2021**, *54*, 586–602.e8. [CrossRef]
221. Ye, Q.; Putila, J.; Raese, R.; Dong, C.; Qian, Y.; Dowlati, A.; Guo, N.L. Identification of Prognostic and Chemopredictive microRNAs for Non-Small-Cell Lung Cancer by Integrating SEER-Medicare Data. *Int. J. Mol. Sci.* **2021**, *22*, 7658. [CrossRef]
222. Seiffert, J. *SEER Program: Comparative Staging Guide for Cancer, Version 1.1 (Rep. No. 93-3640)*; NIH Publication: Bethesda, MD, USA, 1993.
223. Stein, W.D.; Litman, T.; Fojo, T.; Bates, S.E. A Serial Analysis of Gene Expression (SAGE) database analysis of chemosensitivity: Comparing solid tumors with cell lines and comparing solid tumors from different tissue origins. *Cancer Res.* **2004**, *64*, 2805–2816. [CrossRef]
224. Edge, S.B.; Byrd, D.R.; Compton, C.C.; Fritz, A.G.; Greene, F.L.; Trotti, A.E. (Eds.) *AJCC Cancer Staging Manual*, 7th ed.; Springer: New York, NY, USA, 2010.