medical sciences

MDPI

Article

Multi-Omics Data Analysis Identifies Prognostic Biomarkers across Cancers

Ezgi Demir Karaman [1] and Zerrin Işık [2,*]

[1] Department of Computer Engineering, Institute of Natural and Applied Sciences, Dokuz Eylul University, Izmir 35390, Turkey; ezgi.demir@ceng.deu.edu.tr
[2] Department of Computer Engineering, Faculty of Engineering, Dokuz Eylul University, Izmir 35390, Turkey
* Correspondence: zerrin@cs.deu.edu.tr

Abstract: Combining omics data from different layers using integrative methods provides a better understanding of the biology of a complex disease such as cancer. The discovery of biomarkers related to cancer development or prognosis helps to find more effective treatment options. This study integrates multi-omics data of different cancer types with a network-based approach to explore common gene modules among different tumors by running community detection methods on the integrated network. The common modules were evaluated by several biological metrics adapted to cancer. Then, a new prognostic scoring method was developed by weighting mRNA expression, methylation, and mutation status of genes. The survival analysis pointed out statistically significant results for *GNG11*, *CBX2*, *CDKN3*, *ARHGEF10*, *CLN8*, *SEC61G* and *PTDSS1* genes. The literature search reveals that the identified biomarkers are associated with the same or different types of cancers. Our method does not only identify known cancer-specific biomarker genes, but also proposes new potential biomarkers. Thus, this study provides a rationale for identifying new gene targets and expanding treatment options across cancer types.

Keywords: multi-omics data; network-based integration; community detection; survival analysis; cancer biomarker

check for updates

Citation: Demir Karaman, E.; Işık, Z. Multi-Omics Data Analysis Identifies Prognostic Biomarkers across Cancers. *Med. Sci.* **2023**, *11*, 44. https://doi.org/10.3390/medsci11030044

Academic Editor: Tracy Murray-Stewart

Received: 19 May 2023
Revised: 18 June 2023
Accepted: 20 June 2023
Published: 27 June 2023

1. Introduction

Cancer is a heterogeneous disease caused by changes in cell behavior, uncontrolled growth and genomic alterations such as mutations. It contains many different forms, variables and multiple subgroups. In 2020, a total of 19.3 million new cancer cases occurred in the world and there were almost 10 million cancer-related deaths [1]. The most diagnosed cancers were breast (11.7%), lung (11.4%) and colorectal (10%), while cancer-related deaths occurred most oftenwith lung (18%), colorectal (9.4%), liver (8.3%), stomach (7.7%), and breast (6.9%) cancers [1]. If the incidence rates continue at the same frequencies, it is estimated that there may be 28.4 million new cancer cases in 2040 [1]. For a better prognosis and treatment process in such a disease, it is important to categorize tumors into genetically similar subgroups and associate these subgroups with clinical outcomes. Identifying key genomic similarities shared between cancer types will allow extending effective treatments in one cancer type to others due to sharing similar genomic profiles [2].

The complex biology of cancer diseases cannot be explained by analyzing a single omic data type. A wealth of omics data from genomes, transcriptomes, proteomes, metabolomics, ionomics and epigenomes provide a comprehensive perspective for researchers to better explore cancer biology [3]. The availability of such data requires integrative methods to make further evaluations. The use of cancer informatics methods, which integrate and interpret genome-scale molecular data, may reveal possible biomarkers related to tumor prognosis, diagnosis, etc. For this purpose, various clustering algorithms and advanced analysis techniques can be applied to integrated data [4].

In recent years, biological networks, as a simple but effective representation of complex interactions and regulatory relationships between molecules, have been used extensively to understand the system-level characteristics of diseases [5]. Integrating different types of omics data on these networks and applying network-clustering methods to the integrated data may give more effective results inrevealing biomarkers of cancer development or prognosis [6].

Multi-omics data integration methods can be grouped as deep learning networks, network-based, clustering, features extraction, transformation and factorization [7]. These methods address various applications such as disease subtyping, biomarker discovery, pathways analysis and drug repurposing [7,8].

Here, we focus on specific studies that integrate different types of omics data using network-based approaches and use them for biomarker discovery. Kim et al. [9] proposed a random walk approach on an integrated gene–gene graph with expression and methylation profiles; their analysis identified cancer-specific pathways covering genes related to breast cancer. Another research identified differentially expressed and methylated genes and miRNAs for lung adenocarcinoma, integrated the common genes into the PPI network structure and determined potential target genes as a result of survival analysis [10]. There are some studies that aim to find diagnostic and prognostic biomarkers in endometrial, prostate, and colorectal cancers by applying a similar approach using DNA methylation and gene expression data [11–13]. Sun et al. [14] performed an integrated analysis of genome-wide DNA methylation and gene expression for hepatocellular carcinoma, applied a weighted gene co-expression network analysis (WGCNA) and survival analysis; and found gene signatures associated with overall survival. Champion et al. [15] developed a new algorithm and identified potential cancer drivers for eleven cancer types, including breast invasive carcinoma (BRCA), colorectal adenocarcinoma (COAD), lung squamous-cell carcinoma (LUSC), and kidney renal clear-cell carcinoma (KIRC), with the integration of copy number, DNA methylation and gene expression data. Dimitrakopoulos et al. [16] proposed a network-based integration of the multi-omics data (NetICS) method to prioritize cancer genes. The SNF method [17], which was also applied in this study to integrate gene expression and DNA methylation data, is one of the commonly used methods developed for subgroup identification. Furthermore, it has also been applied to prioritize candidate disease genes [18] and to identify candidate cancer biomarkers [19].

There are studies that apply community detection methods instead of traditional clustering algorithms for cancer biomarker discovery. Tanvir and Mondal [20] determined possible biomarkers for COAD, BRCA and glioblastoma multiforme (GBM) by running seven community detection algorithms on gene co-expression networks. Another study applied a community detection algorithm to a differential gene regulatory network created for breast cancer and suggested diagnostic biomarkers [21]. Yu et al. [22] applied the MCODE algorithm to co-expression networks of multiple cancers to find biomarkers. However, these studies only consider gene expression data rather than an integrated multi-omics network. In addition, while evaluating clustering algorithm performances in most studies, statistical metrics were used instead of biological metrics. Despite that, the extent to which genes in the same module are biologically homogeneous is important for biomarker discovery.

Although different types of omics data were used as biomarkers, the expression data of several genes wereused as a scoring value in the survival analysis [23–25]. To the best of our knowledge, there is no integrative scoring method which concurrently combines different omics data for performing the survival analysis.

In this study, different types of omics data of lung, breast, colorectal and kidney cancers, which are at the forefront in terms of both mortality and incidence, were analyzed. RNA-sequencing and DNA methylation data are integrated into a network. Various network clustering algorithms were applied to the integrated data. Biological metrics were used to evaluate clustering results. For this purpose, a metric called "bioscore" has been developed that examines only cancer-specific biological functions and pathways in clustering evaluation. The same analysis workflow was applied to the validation set and prospective biomarkers were selected. In addition, the mutation status of these biomarker genes was also investigated. Finally, survival analysis was conducted with a new prognostic scoring method developed by using different omics data. The obtained biomarkers were compared with studies in the literature. Some studies present these genes as biomarkers for lung, colorectal, breast and kidney cancers, in line with our study. On the other hand, there are other studies suggesting some genes as biomarkers for other cancer types such as prostate, gastric, hepatocellular, ovarian, and bladder. From this point of view, our study helps to reveal genomic similarities among various cancer types. Moreover, some potential novel biomarkers have been found that need to be confirmed by further wet-lab studies.

2. Materials and Methods

The data set and stages of the method are presented in this section. Gene expression, DNA methylation and somatic mutation data of four cancer types (BRCA, KIRC, LUSC, COAD) were obtained from the publicly available TCGA projects [26]. The dataset was divided into two parts for using different patient samples in training and validation of the model.

Figure 1 provides an overview of the methods used in the study. First, differentially expressed genes and methylated probes were identified. Then, probes with significant methylation changes were paired with the 10 closest upstream and downstream target genes with significant expression changes. Using these probe-gene pairs, the mean value of the probes was assigned to each gene. After that, common differentially expressed and methylated genes were identified. Co-expression and co-methylation networks were constructed with these genes. Co-expression and co-methylation networks were integrated by Similarity Network Fusion (SNF) [17]. Network clustering algorithms run on the resulting integrated networks for each cancer type. Clustering results were evaluated by using biological metrics and the most biologically significant modules were determined. The same pre-processing and analysis methods were applied to both the training and the validation data sets. Common genes (i.e., biomarkers) identified in the same module for both training and validation datasets in four cancer types were extracted. The mutation status of each biomarker gene was examined and the genes covering most mutations for all cancer types were determined. In addition, survival time analysis was applied to observe the effects of biomarker genes; eventually, a scoring method was proposed for survival forecasting.

2.1. Data Analysis

We retrieved DNA methylation, gene expression and somatic mutation data for four different cancer types available on the TCGA website: COAD, KIRC, BRCA and LUSC [26]. We selected these tumors based on analysis of the Pan-Cancer Project [2], which focused on 12 tumor types. Due to the higher number patient samples for three omics data and literature comparability, we focused on four of them (COAD, KIRC, BRCA, LUSC). The TCGAbiolinks package was used to retrieve TCGA data from the GDC data portal [27]. Then, patients having both gene expression and DNA methylation data types were determined. Data from untreated patients were used because some treatments may cause changes in omics data. To avoid misleading results, untreated patients in stage-I and stage-II were filtered out. The data were divided into two sets, training and validation, by random split. Table 1 shows the number of samples for both training and validation datasets. In addition, baseline clinical characteristics were presented in the File S1. The Chi-Square test was used to compare the differences in clinical variables between the training and validation data sets.

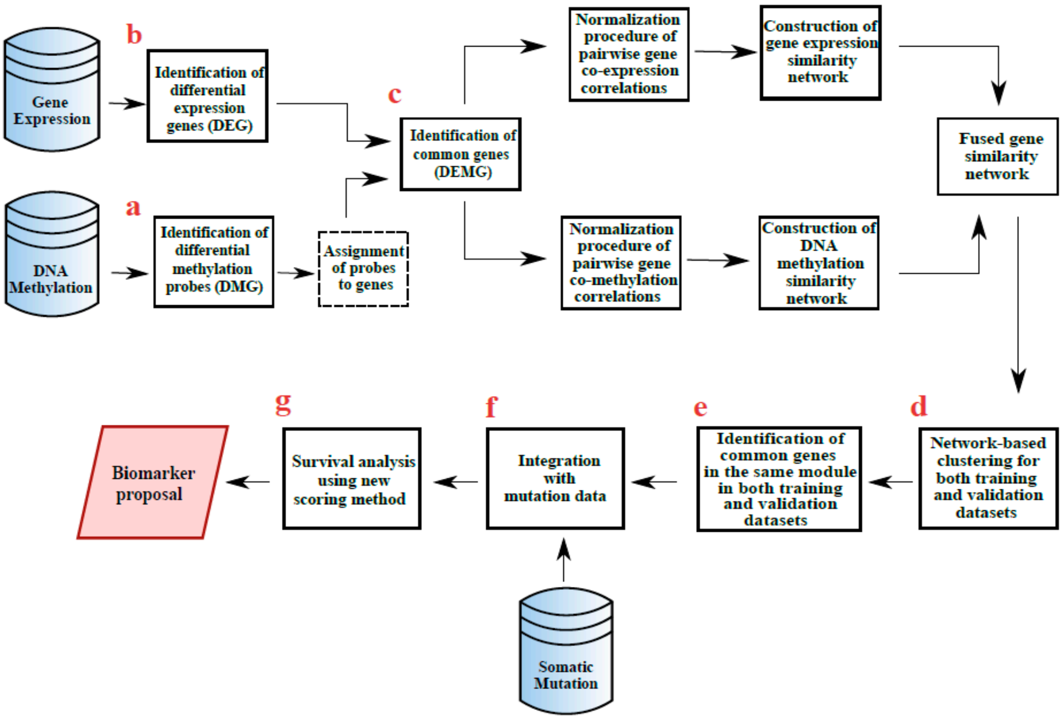
b
Gene Expression
Identification of differential expression genes (DEG)
a
DNA Methylation
Identification of differential methylation probes (DMG)
Assignment of probes to genes
c
Identification of common genes (DEMG)
Normalization procedure of pairwise gene co-expression correlations
Construction of gene expression similarity network
Normalization procedure of pairwise gene co-methylation correlations
Construction of DNA methylation similarity network
Fused gene similarity network
d
Network-based clustering for both training and validation datasets
e
Identification of common genes in the same module in both training and validation datasets
f
Integration with mutation data
Somatic Mutation
g
Survival analysis using new scoring method
Biomarker proposal

Figure 1. Overview of the study workflow. (**a**) Differentially methylated genes (DMG-hypo/hyper) were obtained. (**b**) Differentially expressed genes (DEG) were obtained. (**c**) Differentially expressed and differentially methylated genes (DEMG-hypo/hyper) were obtained by taking the common ones between these two groups. (**d**) Clustering algorithms were implemented to detect modules on the DEMG-hyper and DEMG-hypo networks for both training and validation sets. (**e**) Common genes of all cancer types and that were included in the same modules in the training and validation datasets were determined. (**f**) The mutation status of each gene was examined. (**g**) The potential biomarker genes were identified through survival analysis based on the developed prognostic scoring method.

Table 1. The number of cases by cancer types in the data set.

Cancer Type	Numberof Training Samples		Numberof Validation Samples	
	Tumor	Normal	Tumor	Normal
COAD	74	19	78	19
KIRC	90	24	91	24
BRCA	261	83	279	83
LUSC	153	7	152	7

2.1.1. Identification of Differentially Expressed Genes

Differential gene expression analysis was performed to identify gene expression changes between the tumor and normal samples. In this analysis, the edgeR package is used and both exacttest and log2 foldchange were calculated. The *p*-values were adjusted using Benjamini and Hochberg's approach [28]. Statistically significant gene lists were obtained by filtering genes with the absolute log2 foldchange value > 1.0 and FDR < 0.05.

2.1.2. Identification of Differentially Methylated Probes

We aimed to identify DNA methylation changes in distal regulatory regions and correlate these signatures with mRNA expression in nearby genes. Identification of differentially methylated probes, binding of distal probes with significant methylation changes to target genes, and selection of probe-gene pairs were performed by using the ELMER package [29].

ELMER analysis [29] uses a data structure called "MultiAssayExperiment" (MAE) which stores different assays of all samples in a single object. A "MAE" object containing

DNA methylation and gene expression data was created using the “createMAE” function. Using the “get.feature.probe” function provided by ELMER, only distal probes (at least 2 Kbp away from the transcription start site) were selected; thus, we aimed to identify distant interactions that regulate genes. In this function, the “genome” parameter is set to hg38, and the “met.platform” parameter is set to 450 K. The determined distal probes were given as the “filter.probe” parameter of the “createMAE” function. After this step, differentially methylated CpGs were identified using the “get.diff.meth” function, which performed a one-way *t*-test. The “sig.dif” parameter of this function, which indicates the smallest DNA methylation difference, is a cutoff value for selecting significant hypo-/hyper-methylated probes and it was set to 0.3. Since the group structure (tumor vs. normal) in the analysis was known in advance, the “mode” parameter was chosen as supervised. Raw *p*-values were adjusted by using the Benjamini–Hochberg method [28], and probes with adjusted *p*-value < 0.01 were selected. The next step of the analysis is to identify probe–gene pairs. Using the “get.pair”function, selected distal probes with significant methylation changes were linked to the closest 10 upstream and 10 downstream target genes with significant expression changes. Silva et al. [29] aimed to avoid systematic false positives for probes in gene-rich regions by choosing a fixed number of genes to be tested for each probe. In this function, the “filter.percentage” and “filter.portion” parameters are set to 0.05 and 0.3, respectively. This setup guarantees that at least 5% of beta values are less than 0.3 and 5% of beta values are greater than 0.3.

2.2. Construction Gene Co-Expression & Co-Methylation Networks

Using the probe–gene pairs determined in the previous step, the average methylation value of the probes was assigned to each gene. The Ensemble gene identifiers were converted to the Entrez gene identifiers by using the “org.Hs.eg.db” package [30]. Then, common differentially expressed-hypomethylated genes (DEMG_Hypo), and differentially expressed-hypermethylated genes (DEMG_Hyper) were identified.

While constructing a co-expression and co-methylation network, we used these common genes specific to each cancer type. A correlation value between two genes is computed by the normalized absolute Pearson correlation with the same method as given in a previous study [31]. First, the expression and methylation correlation coefficients between two genes were computed using Pearson correlation. The Fisher transform was applied to make comparable correlation estimates between datasets. We standardized values as *z*-scores in each dataset. Then, the standardized correlations were obtained by inverting the *z*-score. The absolute value of correlations is used as the edge weight of both co-expression and co-methylation networks. The algorithm is summarized in the ‘Algorithm 1’ section below. This method was applied to all types of cancers (i.e., BRCA, COAD, KIRC, LUSC).

Algorithm 1: Procedure for determining pairwise gene correlations.

```
Input: expression and methylation profiles of $n$ genes.
Output: pairwise gene correlations $r'_{ij}$ for any pair of genes $i$ and $j$.
Compute correlation $r_{ij}$ of each pair of genes $i$ and $j$, using Pearson correlation.
Normalize $r_{ij}$ for any $1 \leq i, j \leq n$ with the following steps:
1. Apply Fisher's $z$ transformation to $r_{ij}$, i.e., $= 0.5ln\left(\frac{1+r_{ij}}{1-r_{ij}}\right)z_{ij}$
2. Standardize $z_{ij}$, i.e., $z'_{ij} = \frac{z_{ij}-\mu}{\sigma}$, where $\mu$ and $\sigma$ are the mean and standard deviation of $z_{ij}$ for all $1 \leq i, j \leq n$.
3. Apply Fisher's inverse transformation to $z'_{ij}$, i.e., $r'_{ij} = \frac{\exp(2z'_{ij})-1}{\exp(2z'_{ij})+1}$
Return $r'_{ij}$ for any $i, j$.
```

2.3. Network-Based Data Integration

Co-expression and co-methylation networks individually created for each cancer type were used as the input of an integrative method called Similarity Network Fusion (SNF) to construct a weighted and undirected similarity network [17].

SNF is based on a certain number of similarity matrices corresponding to different layers referring to the same set of nodes. The similarity matrices are then converted into a unique similarity matrix. During this transformation, SNF has the purpose of strengthening the weaker links common to all layers as well as the very strong links found in one layer. The nodes of the obtained network are the common ones in each layer, and the edges are calculated according to the new similarity values. There are three parameters in SNF: K is the number of neighbors, α is a hyper-parameter, and t is the number of iterations. We ran the SNF algorithm with the K value as 5, 9, 21, and 30 and the t value as 5, 10, and 20. However, we obtainedmore stable results by setting $K = 9$ and $t = 20$. This setup was used for all cancer types.

After the t steps of iteration, co-expression and co-methylation networks converge to integrated gene similarity networks. We used a min-max normalization for these networks to obtain more stable results. The adjacency matrix obtained as a result of SNF was converted into a graph using the "igraph" package [32].

2.4. Network-Based Clustering

Fast Greedy [33], Infomap [34] and Louvain [35] clustering algorithms run on integrated gene similarity networks specific to each cancer type. Fast Greedy tries to find communities in graphs by optimizing the modularity score, which is based on the idea of having dense connections between nodes within modules but having sparse connections between nodes of different modules [33]. Infomap finds a community structure that minimizes the expected description length of a random walker trajectory [34]. Louvain implements the multi-level modularity optimization algorithm for finding a community structure. It is based on the modularity measure and a hierarchical approach [35]. Each clustering algorithm runs using the corresponding functions of the igraph library with its default parameters [32].

BHI and Bioscore metrics were used for the evaluation of the clustering results. The BHI measures how biologically homogeneous the clusters are [36]. The measure checks whether genes found in the same cluster also belong to the same biological function classes. The BHI is in the range of [0,1]; larger values correspond to more biologically homogeneous clusters. The "BHI" function in the "clValid" library was used to calculate the BHI score.

Another biological metric is the Bioscore, which was adapted based on the work of Bruno and Friori [37]. According to their work, this score assessed how many gene subsets showed a significant p-value considering all function classes. However, there were many functional terms that are unrelated to cancer development. Therefore, we adapted the Bioscore metric to measure the homogeneity of clusters by scoring only the cancer-related Gene Ontology (GO) Biological Processes (BP) and KEGG pathway terms. The cancer-related GO BP and KEGG pathway terms are taken from the study of [38]. Fisher'sexact test [39] was used to identify significant terms and raw p values were adjusted using the Benjamini–Hochberg method [28] and terms with adjusted $p < 0.05$ were considered significant. If a gene in a cluster is involved in a significant cancer-related GO BP or KEGG pathway, the score of this gene increases by 1, otherwise it remains 0. After calculating a score for each gene in a cluster, they are summed, and a min-max normalization is applied to ensure consistency across all clusters. The Bioscore of a cluster is

$$Bioscore = \sum_{i}^{K} \sum_{cat}^{G} \Theta_{i,cat} \tag{1}$$

where K is the number of genes in the dataset, and G is the number of cancer-related and functional categories stored in the external file. These cancer-related terms are given in Table S1 for GO BP and in Table S2 for the KEGG pathway. $\Theta_{i,cat}$ is defined as follows:

$$\Theta_{i,cat} = \begin{cases} 1, P_{i,cat} < t \\ 0, otherwise \end{cases} \tag{2}$$

where $P_{i,cat}$ is the *p*-value of the cancer-related category *cat* associated with gene *i*, and *t* is a threshold (e.g., 0.05). The most biologically homogeneous modules were determined by examining the results obtained.

2.5. Validation Analysis

The same pre-processing and analysis methods were applied to the validation samples that are given in Table 1. Statistically significant modules were obtained by applying clustering to the validation dataset. Common genes, which are found in the same module for both training and validation datasets, were identified for all cancer types. Then, these genes were selected for biomarker analysis.

2.6. Somatic Mutation Status of Biomarkers

Somatic mutation data of BRCA, LUSC, KIRC, and COAD cohorts were downloaded from the GDC Portal. The mutation data were filtered based on biomarker genes identified in the previous step for untreated patients in stageI and stage II. The mutation status of each biomarker was examined and the genes with the highest number of mutations were determined for all cancer types.

2.7. Survival Analysis

After identifying biomarker genes, the effects of these genes on the overall survival time of patients were also investigated. For this purpose, a new scoring scheme was created by taking a weighted summation of individual scores of DNA methylation, gene expression and mutation data. We called this score "prognostic score", since this score would show both positive (e.g., high prognostic score → good survival) and negative (e.g., high prognostic score → poor survival) correlation with the survival time of a patient.

The prognostic score by considering three data types is calculated by the following equation:

$$Prognosticscore(g_x) = Geneexpression(g_x) \times 0.5 + DNAmethylation(g_x) \times 0.3 + Mutationstatus(g_x) \times 0.2 \quad (3)$$

where g_x represents a gene. For this procedure, a log transformation followed by a min-max normalization was applied to the raw read counts of RNA-sequencing. Mutation status was assigned "1" if the gene has a mutation, otherwise "0". Since the beta value varies between 0 and 1 in DNA methylation, it remains the same value. For survival analysis, continuous values should be represented as categorical values. For this process, the differentially expressed and hypomethylated genes (DEMG_Hypo) and differentially expressed and hypermethylated genes (DEMG_Hyper) were compared among themselves by cancer type. Consequently, common DEMG Hypo and common DEMG Hyper genes were identified in both the training and validation sets for each cancer type. The numbers of these genes, named DEMG_Common, are shown in Table 2.

The prognostic score value was calculated for all DEMG_Common given in Table 2.

$$HighLevelforg_x = \frac{\sum Prognosticscore(g_x)}{Numberofpatientsincancertype} + SD(Prognosticscore(g_x)) \quad (4)$$

$$LowLevelforg_x = \frac{\sum Prognosticscore(g_x)}{Numberofpatientsincancertype} - SD(Prognosticscore(g_x)) \quad (5)$$

where g_x represents a gene. High and low levels were determined by taking the mean +/− 1-standard deviation of each gene's score for all patients (Equations (4) and (5)). After calculating these values for all genes, the average high and low cutoff values were obtained by dividing by the number of DEMG_Common in each cancer type *T* (Equations (6) and (7)).

$$Avg.ofHighLevels(T) = \frac{\sum HighLevelforallg_x}{NumberofDEMG_{Common}incancerT} \quad (6)$$

$$Avg.of LowLevels(T) = \frac{\sum LowLevel forall g_x}{Number of DEMG_{Common} incancer T} \tag{7}$$

According to these limits, a score of a gene less than average low level was labeled as “low”, one between average low and high level as was labeled “normal”, and one higher than average high level as was labeled “high”. Finally, we obtained a categorical prognostic score for each gene.

Table 2. The numbers of common DEMG in the training and validation set. DEMG_Common indicates this number for each cancer type.

Cancer Type	DEMG_Common
Brca_hypo	2428
Lusc_hypo	3235
Coad_hypo	3382
Kırc_hypo	3184
Brca_hyper	2288
Lusc_hyper	1749
Coad_hyper	1475
Kırc_hyper	1063

The Cox proportional hazard model and “survival” package were used to analyze the risk factors [40]. To perform survival analysis, vital status, days to last follow-up and days to death information were obtained from the clinical data files of the patients. The time variable was taken as the days to the last follow-up if the patient was alive, and as the days to death if the patient was dead. In addition, to understand the relationship between categorical variables and overall survival, the Kaplan–Meier estimator [41] was used, which is one of the most widely-used non-parametric measures in survival analysis and in medical research.

Another point we would like to mention is that gene expression (0.5), DNA methylation (0.3), and mutation (0.2) weights are not arbitrarily selected in the prognostic score equation. We also experimentally tested the version of the weights with gene expression (0.4), DNA methylation (0.4) and mutation (0.2). However, in the analysis carried out with this version (0.4, 0.4, 0.2), we obtained fewer biomarkers based on significant hazard ratio and *p*-values. Considering that there are no mutation data for each gene, we assigned the smallest weight (0.2) to the mutation data in both versions. Since survival analysis studies are mostly based on gene expression, we decided to use the weight combination to place more emphasis on gene expression.

Moreover, in order to evaluate the power of survival analysis by combining the three data types, we also computed the prognostic score (Equation (3)) by using a single data type (gene expression, DNA methylation, or mutation status). For this process, the same pipeline described above was applied to each data type. For gene expression and DNA methylation, high and low cutoff values were determined independently, and survival analysis was carried out by labeling in accordance with these cutoffs. Since the mutation status is represented as binary data (value of “1” indicates mutation, otherwise it becomes “0”), survival analysis with mutation status was performed by directly using these binary values.

2.8. MOFA Analysis

We applied the Multi-Omics Factor Analysis (MOFA), which is a computational method used to gain biological insights from multi-omics data. SNF combines multi-omics data through network fusion, whereas MOFA applies a matrix factorization for data integration. MOFA is an adaptation of Principal Component Analysis (PCA) for multi-omics data. MOFA takes data matrices from each omics type as input, and then decomposes these matrices into a factor matrix for each sample and weight matrices for each omics data type [42].

The same samples (given in Table 1) and the three omics layers of DEMG_Common (mentioned in Table 2), gene expression, DNA methylation, and somatic mutation were used in the MOFA implementation. In addition, information from patients' clinical data files was also included as metadata. For the gene expression data, a log transformation followed by a min-max normalization was applied to the raw read counts. Mutation status was assigned "1" if the gene has a mutation, otherwise "0". Since the beta value ranges between 0 and 1 in DNA methylation, it remains the same value. After data preprocessing, we used the R package MOFA [42], an unsupervised factor analysis model to perform multi-omics data integration. We employed default parameters for model training (number of factors = 15, convergence mode = "slow", maxiter = "1000", seed = "42").

Next, we aimed to understand the molecular etiology of the MOFA factors. We investigated whether any of the inferred latent factors were related to prediction of patient outcomes by using the Cox proportional hazards model. Evaluating top weights using the loadings of each feature can provide us with insights for identifying clinical biomarkers. Therefore, across all omics data types, we selected the top 30 genes with the highest weights in the significant factors identified through survival analysis. In addition, for each omics data type, we identified the 30 highest weighted genes in the first three components that were shown to be significant as a result of the variance decomposition analysis performed with MOFA. We examined the associations of these genes with the previously identified potential biomarkers.

3. Results

The results of the entire analysis are summarized in this section.

3.1. Identification of Differentially Expressed Genes/Differentially Methylated Probes

Table 3 showsthe number of significant hypo-/hyper-methylated probes, the number of 10 closest upstream and 10 downstream target genes to probes with significant methylation changes, and the number of statistically significant ones among these probe-gene pairs for the training set. The same analysis was also applied for the validation set and the statistics are given in Table 4.

Table 3. Summary of differential methylation analysis for the training set. "Hypo-M" and "Hyper-M" indicate hypomethylated and hypermethylated, respectively.

Cancer Type	Number of Differentially Methylated Probes		Number of Nearby Genes		Number of Probe-Gene Pairs	
	Hypo-M	**Hyper-M**	**Hypo-M**	**Hyper-M**	**Hypo-M**	**Hyper-M**
COAD	3103	2195	62,039	43,895	2561	6117
KIRC	1277	691	25,540	13,820	2388	2277
BRCA	1252	1048	25,040	20,953	2490	4606
LUSC	3415	1949	68,300	38,980	2588	3451

Table 4. Summary of differential methylation analysis for the validation set. "Hypo-M" and "Hyper-M" indicate hypomethylated and hypermethylated, respectively.

Cancer Type	Number of Differentially Methylated Probes		Number of Nearby Genes		Number of Probe-Gene Pairs	
	Hypo-M	**Hyper-M**	**Hypo-M**	**Hyper-M**	**Hypo-M**	**Hyper-M**
COAD	3084	1729	61,666	34,580	5324	3615
KIRC	1809	780	36,180	15,600	3440	2458
BRCA	1436	1278	28,720	18,180	2925	4225
LUSC	2121	1957	42,420	39,140	1543	4737

For the probe-gene pairs determined in the previous step, the mean methylation values ofprobes matching a gene were assigned this gene. Table 5 summarizesthe number of differentially methylated genes obtained in this way (DMG-hypo/hyper) (Figure 1a), the number of differentially expressed genes (DEG) (Figure 1b) and the number of both differentially expressed and differentially methylated genes (DEMG-hypo/hyper) (Figure 1c) obtained by taking the common ones in these two groups for the training set. The same analysis was also applied for the validation set and the same statistics are given in Table 6. The next analysis steps were continued with the genes in the DEMG-hypo/hyper group.

Table 5. Differential analysis results for RNA sequencing and methylation data in the training set.

Cancer Type	DEG	DMG_Hypo	DMG_Hyper	DEMG_Hypo	DEMG_Hyper
COAD	10,916	10,676	5012	4581	2211
KIRC	12,273	7005	2524	3556	1323
BRCA	14,294	4971	4773	2806	2812
LUSC	11,585	10,898	4666	5085	2309

DEG indicates differentially expressed gene numbers, DMG_Hypo indicates differentially hypomethylated gene numbers, DMG_Hyper indicates differentially hypermethylated gene numbers, DEMG_Hypo indicates differentially expressed and hypomethylated gene numbers, and DEMG_Hyper indicates differentially expressed and hypermethylated gene numbers.

Table 6. Differential analysis results for RNA sequencing and methylation data in the validation set.

Cancer Type	DEG	DMG_Hypo	DMG_Hyper	DEMG_Hypo	DEMG_Hyper
COAD	11,815	10,426	4417	4886	2095
KIRC	14,087	9177	2655	5325	1578
BRCA	14,667	5510	4547	3228	2747
LUSC	12,147	8442	4733	4040	2412

DEG indicates differentially expressed gene numbers, DMG_Hypo indicates differentially hypomethylated gene numbers, DMG_Hyper indicates differentially hypermethylated gene numbers, DEMG_Hypo indicates differentially expressed and hypomethylated gene numbers, and DEMG_Hyper indicates differentially expressed and hypermethylated gene numbers.

3.2. Identification of Common Genes in Different Cancer Types

Figures 2 and 3 show the distribution of the DEMG_hyper and DEMG_hypo for training and validation data before applying SNF and clustering algorithms. As seen in these figures, 49 DEMG-hyper and 151 DEMG_hypo genes were found for the training set, and 53 DEMG-hyper and 227 DEMG_hypo common genes were found for the validation set.

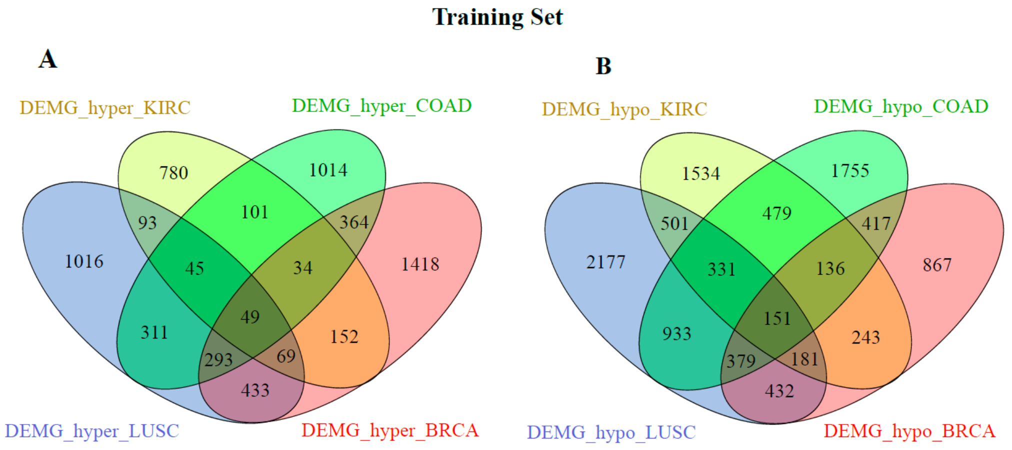

Figure 2. Distribution of training samples (**A**) DEMG_hyper and (**B**) DEMG_hypo between cancer types without applying SNF and clustering algorithms.

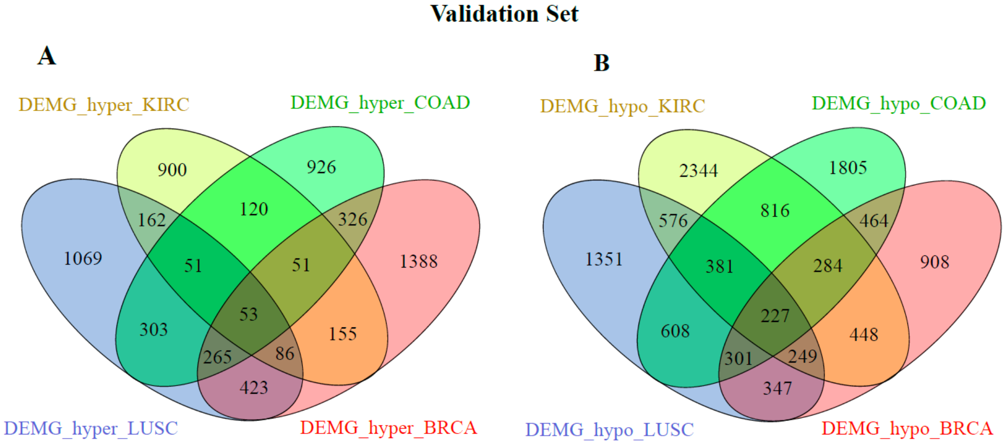

Figure 3. Distribution of validation samples (**A**) DEMG_hyper and (**B**) DEMG_hypo between cancer types without applying SNF and clustering algorithms.

In addition, for the training and validation set, we compared the DEMG_hyper and DEMG_hypo genes common to all four cancer types among themselves. The distribution of these genes is given in Figure 4. Most of the common genes were found in both the training and validation sets.

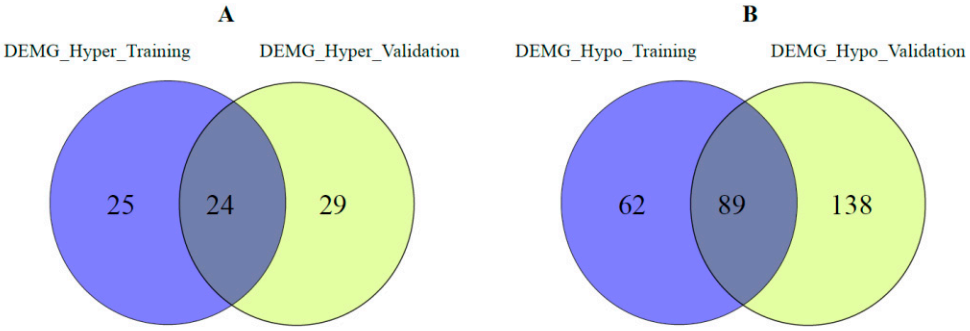

Figure 4. Comparison of (**A**) DEMG_hyper genes common to all four cancer types for validation and training sets, (**B**) DEMG_hypo genes common to all four cancer types for validation and training sets.

3.3. Network Clustering

The biologically homogeneous modules offour cancer types were compared to reveal potential common biomarker genes related to these cancers. For this purpose, we have implemented Fast Greedy, Infomap and Louvain clustering algorithms to detect modules on the DEMG-hyper and DEMG-hypo networks for both training and validation sets (Figure 1d). The performance of each algorithm was evaluated by using both BHI and Bioscore metrics; these results are summarized in Tables 7–10. As seen in Tables 7 and 8, the Fast Greedy algorithm gave higher BHI and Bioscore for all DEMG-Hyper and DEMG-Hypo data for the training set. As seen in Table 9, fast greedy algorithm for BRCA and COAD, Louvin algorithm for LUSC and KIRC gave the best results for DEMG_Hyper data for the validation set. As seen in Table 10, Fast Greedy algorithm for BRCA, COAD and KIRC and Louvin algorithm for LUSC gave the best results for all DEMG_Hypo data for the validation set.

The modules obtained by the clustering algorithm, which gave a better result for each cancer type, were compared among themselves as the DEMG_Hyper ones and the DEMG_Hypo ones. Then, genes that are common to all cancer types and that were included in the same modules in the training and validation datasets were determined (Figure 1e). These genes are listed in Table 11.

Table 7. Performance comparison of clustering algorithms between DEMG-Hyper data for the training set.

	Average-Bioscore (GO-BP)	**Average-Bioscore (KEGG)**	**Average-BHI**	**# of Cluster**
BRCA_hyper				
Fast Greedy	0.500	0.596	0.077	27
Infomap	0.229	0.144	0.055	257
Louvin	0.400	0.422	0.069	30
COAD_hyper				
Fast Greedy	0.289	0.427	0.069	20
Infomap	0.178	0.128	0.046	227
Louvin	0.358	0.328	0.065	27
KIRC_hyper				
Fast Greedy	0.449	0.126	0.067	15
Infomap	0.164	0.007	0.044	144
Louvin	0.342	0.05	0.056	20
LUSC_hyper				
Fast Greedy	0.409	0.446	0.072	24
Infomap	0.135	0.032	0.049	213
Louvin	0.304	0.217	0.065	31

Table 8. Performance comparison of clustering algorithms between DEMG-Hypo data for the training set.

	Average-Bioscore (GO-BP)	**Average-Bioscore (KEGG)**	**Average-BHI**	**# of Cluster**
BRCA_hypo				
Fast Greedy	0.427	0.539	0.081	21
Infomap	0.117	0.094	0.042	274
Louvin	0.484	0.465	0.071	30
COAD_hypo				
Fast Greedy	0.516	0.453	0.082	19
Infomap	0.132	0.064	0.042	434
Louvin	0.467	0.374	0.08	35
KIRC_hypo				
Fast Greedy	0.525	0.499	0.083	18
Infomap	0.176	0.112	0.04	377
Louvin	0.387	0.472	0.071	32
LUSC_hypo				
Fast Greedy	0.274	0.517	0.08	25
Infomap	0.08	0.026	0.043	393
Louvin	0.525	0.351	0.074	37

3.4. Somatic Mutation Status of Biomarkers

The mutation status of each gene in Table 11 was also examined (Figure 1f). Figure 5 shows the number of patients with gene mutations for the hypomethylated group. The color of the bubbles was used to represent the genes in the same modules, and bubble size represents the number of patients. For this procedure, we normalized the number of patients with mutations in that gene by the total number of patients with mutations in each cancer type. The genes having the most mutations for the hypomethylated group were PRKDC, EGFR, PTDSS1, ADGRD1 and LGR4, while SLC9A3 and BRIP1 were the most mutated ones for the hypermethylated group.

Figure 6 shows the number of patients with gene mutations for the hypermethylated group. It was observed that the mutations were mostly of the "missense" type for both groups.

3.5. Survival Analysis

Survival analysis was performed for the genes given in Table 11. The "prognostic score" described in Section 2.7 was used for this analysis. Since this is a continuous value, it must be converted into a categorical value for survival analysis. Therefore, high and low limits were determined by taking the mean +/− 1-standard deviation of each gene's score for all patients. After calculating these averages for all genes, high- and low-level cutoffs were determined based on the average of high- and low-level scores computed specifically for each cancer type. These values were summarized in Table 12. According to these limits, a score less than average low level was labeled as "low", one between average low and high level was labelled as "normal", and one higher than average high level was labelled as "high" for each cancer type.

Table 9. Performance comparison of clustering algorithms between DEMG-Hyper data for the validation set.

	Average-Bioscore (GO-BP)	Average-Bioscore (KEGG)	Average-BHI	# of Cluster
BRCA_hyper				
Fast Greedy	0.515	0.371	0.074	25
Infomap	0.187	0.096	0.066	246
Louvin	0.253	0.476	0.044	32
COAD_hyper				
Fast Greedy	0.512	0.105	0.062	17
Infomap	0.246	0.011	0.064	194
Louvin	0.359	0.057	0.050	27
KIRC_hyper				
Fast Greedy	0.346	0.186	0.077	14
Infomap	0.191	0.106	0.071	173
Louvin	0.361	0.383	0.048	22
LUSC_hyper				
Fast Greedy	0.460	0.349	0.074	23
Infomap	0.147	0.081	0.075	220
Louvin	0.543	0.379	0.050	30

Table 10. Performance comparison of clustering algorithms between DEMG-Hypo data for the validation set.

	Average-Bioscore (GO-BP)	Average-Bioscore (KEGG)	Average-BHI	# of Cluster
BRCA_hypo				
Fast Greedy	0.526	0.294	0.076	23
Infomap	0.045	0.025	0.051	295
Louvin	0.299	0.243	0.077	31
COAD_hypo				
Fast Greedy	0.305	0.567	0.089	20
Infomap	0.193	0.134	0.083	424
Louvin	0.487	0.545	0.056	29
KIRC_hypo				
Fast Greedy	0.459	0.454	0.080	21
Infomap	0.170	0.056	0.039	525
Louvin	0.415	0.199	0.074	32
LUSC_hypo				
Fast Greedy	0.322	0.229	0.076	25
Infomap	0.087	0.036	0.070	336
Louvin	0.415	0.287	0.055	33

Table 11. Common genes in the same module for both training and validation sets.

Genes Name	Methylation Group
PRKDC, MCM4, UBE2V2	Hypo-methylated
LPCAT1, mrpl36	Hypo-methylated
CDKN3, CGRRF1	Hypo-methylated
GNG11, GNGT1	Hypo-methylated
ACTR3B, IMMP2L	Hypo-methylated
SEC61G, EGFR	Hypo-methylated
PTDSS1, CPQ	Hypo-methylated
ARHGEF10, CLN8	Hypo-methylated
CBX8, CBX2	Hypo-methylated
RAN, ADGRD1	Hypo-methylated
TPRG1L, PRDM16-DT	Hypo-methylated
LGR4, BDNF-AS	Hypo-methylated
SLC9A3, PP7080	Hyper-methylated
ENPP5, CYP39A1	Hyper-methylated
RAD54L, EFCAB14	Hyper-methylated
BRIP1, TBX2-AS1	Hyper-methylated

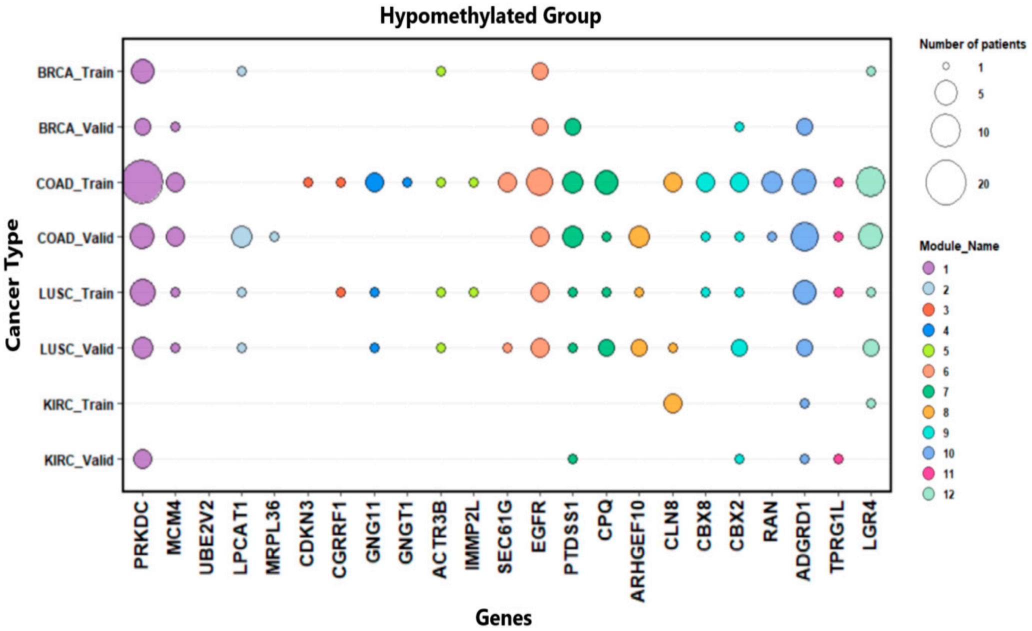

Figure 5. Distribution of patients with mutated genes in the hypomethylated group by cancer types.

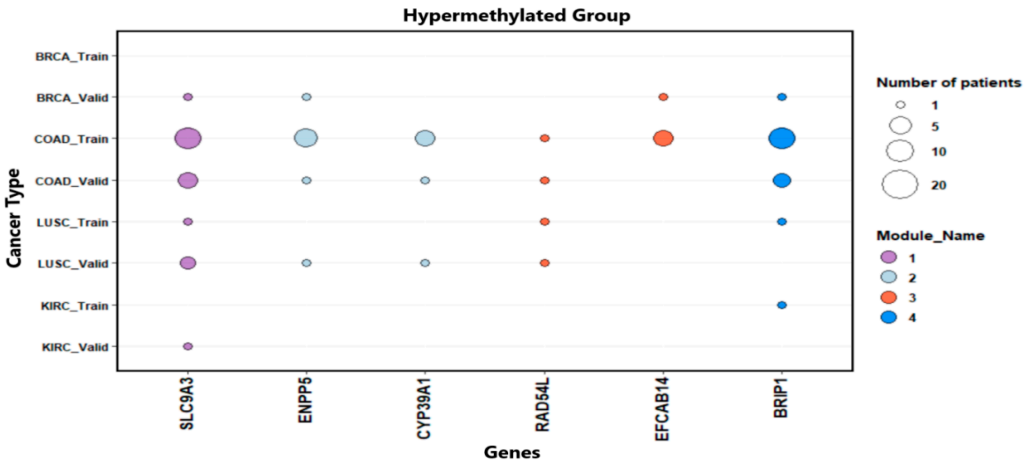

Figure 6. Distribution of patients with mutations of genes in the hypermethylated group by cancer types.

Table 12. Average cutoff values for the high and low levels.

Cancer Type	Average of Low-Level Scores	Average of High-Level Scores
Brca_hypo	0.277	0.419
Lusc_hypo	0.282	0.41
Coad_hypo	0.279	0.437
Kırc_hypo	0.276	0.381
Brca_hyper	0.314	0.457
Lusc_hyper	0.285	0.434
Coad_hyper	0.324	0.49
Kırc_hyper	0.364	0.489

The results of the survival analysis with hazard ratio > 1.0 and *p*-value < 0.05 are presented in Table 13. Among the significant results are potential biomarker genes that were determined by considering the number of patients at that level and the number of deaths according to the prognostic score (Figure 1g). In addition, Kaplan–Meier plots of these genes are presented in File S2.

Table 13. Survival analysis Cox PH model results.

Cancer Type	Gene Name	Prognostic Score Level	Hazard Rate	*p*-Value	Number of Patients at Score Level	Number of Deaths
Brca_hypo	GNG11	Low	7.7055	0.000189	11	4
	CBX2	High	2.0370	0.0138	188	27
Coad_hypo	CDKN3	High	2.577	0.0262	64	15
	ARHGEF10	High	2.855	0.0128	56	14
	GNG11	High	2.2279	0.0563	45	12
	CLN8	High	3.037	0.00823	53	14
Kırc_hypo	CBX2	High	2.8296	0.02	19	7
Lusc_hypo	SEC61G	High	1.6608	0.0541	239	99
	PTDSS1	High	2.6287	0.0217	273	111

We additionally tested the weights with gene expression (0.4), DNA methylation (0.4), and mutation (0.2). Based on a significant hazard ratio and *p*-values, we found fewer biomarkers in the analysis carried out with this version (0.4, 0.4, 0.2). We present the proof of this analysis result in File S3.

3.6. Usage of Individual Data Types for Survival Analysis

Another survival analysis was performed by using gene expression, DNA methylation and mutation data individually to compare them with the proposed multi-omics prognostic score. The survival analysis based on individual data types presented fewer significant results when the hazard ratio and *p*-values were considered. The File S4 summarizes the results of survival analysis with individual data types. We claim that the new prognostic scoring by integrating multi-omics data would empower common biomarker identification across tumor types.

3.7. MOFA Analysis

MOFA was applied to gene expression, DNA methylation and somatic mutation data of four cancer types (BRCA, COAD, KIRC, and LUSC). Significant factors in the trained models were used in the survival analysis. From the identified 15 MOFA factors, Factor 7 (*p*-value = 0.0001), and Factor 15 (*p*-value = 0.0198) for BRCA_hypo, Factor 7 (*p*-value = 0.0201) for COAD_hypo, Factor 8 (*p*-value = 0.0005) for KIRC_hypo, Factor 12 (*p*-value = 0.016) for LUSC_hypo, and Factor 4 (*p*-value = 0.04) and Factor 8 (*p*-value = 0.015) for BRCA_hyper were statistically significantly associated with overall survival. We identified the top 30 genes with the highest weight in these factors and the first three factors determined by variance decomposition analysis. We observed some similarities between these genes and the results provided in Tables 11 and 13. For example, the survival analysis of two methods identified CBX2 and GNG11 genes in BRCA_hypo phenotype. Further results are presented in File S5.

4. Discussion

In the first stage of the study, we performed a network-based integrative analysis with the SNF method using DNA methylation and gene expression data of BRCA, COAD, KIRC and LUSC. Community detection methods were applied to the integrated network and the results were evaluated using cancer-related biological metrics. The same procedure was implemented on both training and validation datasets for all cancer types. As a result of this procedure, there is a concordance between the genes identified in the same module for both training and validation data sets (Table 11). Some of these genes were also mentioned in previous studies in the literature. These studies integrated various omics profiles (e.g., gene expression, DNA methylation, somatic mutation, copy number) and applied one or more computational approaches such as a deep neural network, co-expression network, feature selection, differential expression or methylation gene analysis, or protein–protein interaction analysis on different cancer types. For instance, Fan et al. [43] identified triple-evidence genes representing differentially methylated, differentially expressed, and somatic mutation-associated genes in each of the 13 TCGA cancers. Among the triple-evidence genes they determined, the genes that were also common to all four cancer types in our study are as follows: CBX2, CBX8 genes for BRCA, LUSC and COAD; MCM4, GNG11 genes for LUAD; LGR4 gene for COAD, KIRC, LUSC; LPCAT1 gene for LUSC; EGFR gene for KIRC and ARHGEF10 gene for BRCA. In another study, Mo et al. [44] performed a statistical integrative clustering analysis (iCluster+) using exome sequence, DNA copy number, promoter methylation, and mRNA expression data of TCGA colorectal carcinoma. In this analysis to discover cancer subgroups, PTDSS1, MCM4 and PRKDC genes were identified as molecular drivers belonging to the same subgroup. Qi et al. [45] constructed a PPI network with differentially expressed and aberrantly methylated genes for breast cancer and identified MCM4, CDKN3 and EGFR as hub genes. In another breast cancer study using gene expression and copy-number alterations data in a neural network-based approach, the CDKN3 was one of the subtype-specific genes identified belonging to the LumA subtype [46]. Fiorentino et al. [47] developed a methodology that fuses omics-specific similarity networks in a single network and verified the SEC61G gene as a prognostic biomarker using gene expression, methylation, and miRNA data of GBM. Dimitrakopoulos et al. [16] identified the known EGFR gene for lung cancer by their proposed network-based integration method using somatic mutations, copy number variations, methylation, mRNA and miRNA expression data. Sheng et al. [48] identified differentially expressed mRNAs, miRNAs, and circRNAs for breast cancer and constructed a regulatory network. Then, to explore the key genes involved in the regulatory network, they established a PPI network and applied the MCODE algorithm; as a result of this analysis, LPCAT1, CBX2 and EGFR were identified as potential hub genes. Shi et al. [49] proposed an approach to identify driver genes by integrating mutation data, expression data, and gene networks and reported EGFR and PRKDC as potential driver genes for GBM.

In the next stage of our study, the somatic mutation status of selected genes for biomarker analysis was determined. In addition, a new prognostic scoring method has been developed that uses mRNA expression, methylation and mutation states of biomarkers simultaneously. Finally, we obtained statistically significant results for GNG11, CBX2, CDKN3, ARHGEF10, CLN8, SEC61G and PTDSS1 genes in the survival analysis. Previous studies found in the literature about these genes are summarized below.

G protein subunit gamma 11 (GNG11), a constituent of G-proteins, plays a vital role in the transmembrane signaling system. It has been described as a hub gene or a candidate biomarker in different cancer types. Hua et al. [50] reported in their study that GNG11 acts as a hub gene in lung adenocarcinoma. Moreover, Shi et al. [51] observed that GNG11 was downregulated in lung cancer, and low expression of GNG11 was associated with worse OS for female lung cancer patients who never smoked. Buttarelli et al. [52] generated a ten-gene signature, including the downregulated GNG11 gene, that predicts response to first-line chemotherapy in high-grade serous ovarian cancer patients. Furthermore, Jiang et al. [53] identified that high expression of GNG11 is related toa poor prognosis in ovarian cancer patients. According to Zhang et al. [54], GNG11 is downregulated in tumor tissue, and is the core gene in protein–protein interaction network analysis for triple-negative breast cancer. In addition, Xing et al. [55] stated that GNG11 is one of the eighteen key genes identified for the treatment of colorectal cancer. In line with other studies, according to our study, the GNG11 gene was downregulated and highly methylated in both breast and colon tumor tissues compared to normal tissues. Moreover, according to the prognostic scoring method we developed, GNG11 was associated with poor survival by presenting low scores in breast cancer and high scores in colorectal cancer.

Chromobox 2 (CBX2) is a polycomb repressor complex subunit, and some studies classified it as an oncogene. Clermont et al. [56] reported CBX2 as a potential drug target in their study and associated CBX2 expression with poor clinical outcomes in prostate cancer. Previous studies have shown that high expression of CBX2 is associated with worse survival in hepatocellular carcinoma, high-grade serous ovarian cancer, and lung adenocarcinoma [57–59]. Conversely, Ma et al. [60] identified that CBX2 mRNA and protein levels were significantly increased in gastric cancer tissues, but these levels were not significantly associated with the overall survival of patients. Furthermore, studies on colorectal cancer (CRC) showed that the CBX2 gene was significantly upregulated in CRC tissues compared to normal tissues, and this may be associated with poor prognosis [61,62]. There are various studies about the function of CBX2 in breast cancer. Bilton et al. [63] identified novel mechanisms by which CBX2 promotes breast cancer growth, and inhibition of CBX2 could be a novel therapeutic strategy. Zheng et al. [64] stated that there was a positive correlation between high CBX2 expression and activation of the PI3K/AKT pathway, and that CBX2 could be a potential prognostic biomarker. Li et al. [65] showed that the expression of CBX2 was strongly associated with tumor stage, and there was higher CBX2 expression in stage IV patients compared to others. Moreover, Piqué et al. [66] found that CBX2 promotes cell proliferation in breast cancer, its overexpression causes upregulation of genes involved in cell cycle progression, and CBX2 overexpression is associated with poor 5-year survival. Our results are consistent with previous studies; we found that CBX2 is upregulated in breast and clear-cell renal cell carcinoma and patients with poor survival showed higher prognostic scores in both cancer types.

Cyclin-dependent kinase 3 (CDKN3) is a member of the protein phosphatase inhibitors family and involved in regulation of the cell cycle [67,68]. Abnormal expression of CDKN3 is seen in many types of cancer. Abdel-Tawab et al. [69] suggested that CDKN3 expression could be used as a diagnostic and predictive biomarker of gastric cancer. Li et al. [70] found that CDKN3 was overexpressed in human gastric cancer tissues and associated with poor patient survival. Similarly, there are other studies associating CDKN3 overexpression with poor survival in nasopharyngeal carcinoma, lung adenocarcinoma, breast, bladder, and cervical cancer [71–75]. An immunohistochemical study for ESCC identified abnormal CDKN3 protein expression in esophageal squamous-cell cancer (ESCC) patients and confirmed its association with ESCC progression [76]. Yang and Sun [77] showed the role of CDKN3 in cellular proliferation of colorectal cancer by examining the effects of CDKN3 siRNA on the SW480 cell line; it is associated with cell cycle progression and apoptosis. Moreover, Li et al. [78] stated that CDKN3 is highly expressed in colorectal cancer, and this may be closely related to the poor prognosis of the patients. In our study using a different dataset, we found that the CDKN3 gene was highly expressed and less methylated in colorectal cancer patients compared to normal samples. In parallel with literature studies, we identified that colorectal cancer patients with poor survival showed a high prognostic score.

ARHGEF10 encodes the Rho guanine nucleotide exchange factor, and its role in cancer has not yet been clarified. However, there are studies presenting it as a candidate tumor suppressor gene for pancreatic ductal adenocarcinoma [79], hepatocellular carcinoma [80], breast [81] and urothelial carcinoma [82]. In addition, while decreased ARHGEF10 expression was observed in tumor cells in these studies, increased ARHGEF10 expression was found in colorectal cancer in our study. In addition, according to the prognostic scoring method we developed, high scores in colorectal cancer were associated with poor survival.

CLN8 encodes a transmembrane protein, and mutations in this gene are linked to progressive epilepsy with cognitive disabilities (EPMR), a subtype of neuronal ceroidlipofuscinoses (NCL) [83]. In order to reveal the role of NCL genes in cancer-related processes, Yap et al. [84] stated that the CLN8 gene showed low expression in brain cancer cells and had a tumor suppressor effect on patient survival. However, more research is needed in the future to explore the importance of CLN8 in cancers. In our study, low expression was observed in colorectal cancer tissue, and patients with high score values showed poor survival, according to the developed prognostic scoring method.

The subunit of the SEC61 translocon complex (SEC61G) participates in protein folding, post-translational modification and translocation, and plays critical roles in several cancer types [85]. Zhang et al. [86] used the expression levels of five genes to develop a prognostic model for colorectal cancer; one of these genes was SEC61G. In studies conducted for breast cancer, it has been stated that the SEC61G gene can be used as a diagnostic biomarker and therapeutic target, since high expression of SEC61G is associated with the expression of the proliferation marker Ki-67 and glycolysis. It was stated that SEC61G expression was higher and methylation level was lower in tumors compared to normal tissues, and this was associated with poor survival [87,88]. Zhang et al. [89] similarly found that SEC61G showed hypomethylation and high expression in bladder cancer cells. Meng et al. [90] stated that SEC61G is up-regulated in human kidney tumors and is associated with poor prognosis, compared with the control group. SEC61G knockdown significantly inhibits cell proliferation, migration and invasion; therefore it may serve as a biomarker for kidney cancer. In addition, some studies associated SEC61G overexpression with worse survival in hepatocellular carcinoma, head and neck squamous carcinoma, glioblastoma and lung adenocarcinoma [91–94]. In our study, we found that the SEC61G gene was highly expressed and hypo-methylated in lung cancer patients compared to normal samples. Furthermore, we identified that lung cancer patients with poor survival had high prognostic scores. Therefore, our analysis is supported by various literature studies.

There have been some cancer-related reports addressing phosphatidylserine synthase 1 (PTDSS1). Cheng et al. [95] stated that the PTDSS1 could be one of the anti-cancer targets for the treatment of colorectal cancer. Sekar et al. [96] showed that inhibiting the production of ether-phosphatidylserine by targeting PTDSS1 limits tumor-associated macrophage expansion and breast tumor growth. In a study on ESCC, it was stated that mRNA expression has a differential significance between ESCC and normal controls [97]. N'Guessan et al. [98] measured the expression of PTDSS1 at each stage of the cell cycle and found that PTDSS1 gene expression increased in the G2/M phase compared to the G1 phase in pancreatic cancer cells. They also noted that PTDSS1 gene expression was higher in pancreatic cancer patients compared to healthy tissues, and this was associated with a lower probability of survival in pancreatic cancer patients. In another research study, Li et al. [99] identified that high expression of PTDSS1 is significantly associated with a lower probability of survival in urothelial bladder carcinoma (BLCA), concluding that PTDSS1-mediated phosphatidylserine signaling is involved in the pathogenesis of BLCA. Furthermore, Wang et al. [100] concluded in their study that PTDSS1 is an oncogene in lung adenocarcinoma and its overexpression may reduce the likelihood of survival. In our study, we found that the PTDSS1 gene was upregulated and hypomethylated in LUSC compared to normal tissues. Moreover, according to the prognostic scoring method we developed, low scores in PTDSS1 were associated with poor survival.

In addition to carcinoma, potential biomarker genes in Table 13 have been associated with a wide range of other diseases, and they seem to be activated or inhibited in various biological processes. Cheng et al. [101] suggested that GNG11 could be used as a biomarker for differentiate ulcerative colitis and Crohn's disease. Moradi et al. [102] proposed that GNG11 could be a diagnostic biomarker for Parkinson's disease. GNG11 plays a key role in heart rhythm regulation and is associated with cardiac disease risk [103]. The CDKN3 gene could be used as a potential marker to identify severe COVID-19 patients [104]. Yue et al. [105] identified 10 central genes, including CDKN3, and stated that these genes may serve as new target markers for early diagnosis, prognosis and therapy in psoriasis. Yao et al. [106] constructed an index using seven genes, including CDKN3, which are associated with hypoxia, a prominent factor in the diagnosis and treatment of osteoarthritis. The ARHGEF10 mutation was associated with slowed nerve conduction velocity [107]. The ARHGEF10 gene might be associated with the pathogenesis of Behcet's disease [108]. Zhang et al. [109] revealed the candidacy of the CLN8 gene as a genetic modifier contributing to extreme phenotypic variation in Gaucher disease. A novel mutation in CLN8 may cause Northern Epilepsy cases in Turkey [110]. CBX2 gene plays a role in the human sex development process and its disorders [111]. SEC61G is among the nine circadian-related genes identified related to circadian rhythm disruption, which is critical in the pathogenesis of Alzheimer's disease [112]. The SEC61G gene was differently methylated in patients with Balkan endemic nephropathy [113]. The SEC61G gene is differentially expressed and methylated in fetal alcohol spectrum disorder patients [114]. There are many studies about relationship between the mutation in the PTDSS1 gene and Lenz-Majewski syndrome [115]. Soueid et al. [116] proposed that PTDSS1 is among the potential autism susceptibility genes in their study.

Besides the SNF method, MOFA has also been applied for the integration of gene expression, DNA methylation and mutation data. There are differences in the application of the two methods and the interpretation of the obtained results. The SNF method initially constructs a distinct similarity network for each omics, then integrates the networks using an iterative procedure. However, MOFA utilizes a matrix factorization technique. Although matrix factorization techniques are frequently employed for dimensionality reduction, they might ignore biological correlations between the features. Furthermore, because of its linearity, the MOFA model may miss non-linear correlations between features. Another challenging process was the biological interpretation of the inferred latent factors. Each feature in MOFA has a 'weight' that represents its relative relevance to the factor. We utilized these weights to assess the most informative biomarkers. The most difficult aspect of implementing the SNF method was integrating clinical data and was not included in this method. On the other hand, this capability is available in MOFA. Consequently, the two methodologies cannot be directly comparable in terms of their results due to their different computational methods. Despite all the differences of the two methods, they report similar gene outputs for some cancer phenotypes.

This study provides new insights into potential prognostic biomarkers for many tumor types; however, it has some limitations. First, gene expression, DNA methylation and mutation data of the same patient are required to calculate the proposed prognostic score. Nevertheless, in some cases, all three data types may not be available for the same patient. Second, all omics data come from the TCGA database; when other public repositories are checked, it is common to find gene expression data for specific drug treatments on cancer cell lines or gene knockout studies. In contrast, the samples in our study were selected from patients who did not receive any treatment. Furthermore, due to focusing on patients at cancer stage 1 and 2, there were relatively few samples remaining in the study. If a larger sample size is used, the predictive power of the algorithm can be more effectively verified. Although verification of the proposed biomarkers on a new patient cohort could not be currently applied, we aim to investigate the biological validity of some of these biomarkers in new patient cohorts as a future study.

5. Conclusions

We implemented an integrative network analysis approach that explores common biomarkers for lung, breast, colorectal and kidney cancers by integrating RNA-sequencing and DNA methylation data. Several network clustering algorithms were used on the integrated network data. Cancer-specific evaluation metrics were applied to evaluate clustering results, and finally, common modules were reported across four cancers. The same analysis pipeline was applied to the validation set and final prospective biomarkers were identified. Survival analysis for biomarkers was conducted with a new prognostic scoring method that integrates mRNA expression, methylation and mutation status of genes. A literature survey about significant biomarkers highlighted in survival analysis revealed that GNG11, CBX2, CDKN3, ARHGEF10, CLN8, SEC61G and PTDSS1 genes present similar survival and prognostic behaviors in the specified cancers. In summary, multi-omics and network-based analysis can help to discover new targets across cancers and to reduce treatment costs.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/medsci11030044/s1, File S1: Basic clinical characteristics of patients; File S2: Kaplan–Meier plots of potential biomarker genes; File S3: The results of survival analysis for the weights with gene expression (0.4), DNA methylation (0.4), and mutation (0.2); File S4: The results of survival analysis with individual data types; File S5: The results of MOFA analysis; Table S1: Cancer-related terms used in Bioscore for GO BP; Table S2: Cancer-related terms used in Bioscore for KEGG pathway.

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

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All patient samples are available from the GDC data portal https://portal.gdc.cancer.gov (accessed on 20 February 2022).

Acknowledgments: E. DemirKaraman is supported by the TUBITAK 2211/C Ph.D. Scholarship Program.

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

References

1. Sung, H.; Ferlay, J.; Siegel, R.; 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 A Cancer J. Clin.* **2021**, *71*, 209–249. [CrossRef] [PubMed]
2. Weinstein, J.; Collisson, E.; Mills, G.; Shaw, K.; Ozenberger, B.; Ellrott, K.; Shmulevich, I.; Sander, C.; Stuart, J. The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nat. Genet.* **2013**, *45*, 1113–1120. [CrossRef] [PubMed]
3. Haas, R. Designing and Interpreting 'Multi-Omic'Experiments That May Change Our Understanding of Biology. *Curr. Opin. Syst. Biol.* **2017**, *6*, 37–45. [CrossRef] [PubMed]
4. Huang, S.; Chaudhary, K.; Garmire, L. More Is Better: Recent Progress in Multi-Omics Data Integration Methods. *Front. Genet.* **2017**, *8*, 84. [CrossRef] [PubMed]
5. Liu, Z.; Zhang, S. Tumor Characterization and Stratification By Integrated Molecular Profiles Reveals Essential Pan-Cancer Features. *BMC Genom.* **2015**, *16*, 503. [CrossRef]
6. Cantini, L.; Medico, E.; Fortunato, S.; Caselle, M. Detection of Gene Communities in Multi-Networks Reveals Cancer Drivers. *Sci. Rep.* **2015**, *5*, 17386. [CrossRef]
7. Nicora, G.; Vitali, F.; Dagliati, A.; Geifman, N.; Bellazzi, R. Integrated Multi-Omics Analyses in Oncology: A Review of Machine Learning Methods and Tools. *Front. Oncol.* **2020**, *10*, 1030. [CrossRef]
8. Subramanian, I.; Verma, S.; Kumar, S.; Jere, A.; Anamika, K. Multi-Omics Data Integration, Interpretation, and Its Application. *Bioinform. Biol. Insights* **2020**, *14*, 117793221989905. [CrossRef]
9. Kim, S.; Kim, T.; Jeong, H.; Sohn, K. Integrative Pathway-Based Survival Prediction Utilizing the Interaction between Gene Expression and DNA Methylation in Breast Cancer. *BMC Med. Genom.* **2018**, *11*, 33–43. [CrossRef]
10. Yang, Z.; Liu, B.; Lin, T.; Zhang, Y.; Zhang, L.; Wang, M. Multiomics Analysis on DNA Methylation and the Expression of Both Messenger RNA and Microrna in Lung Adenocarcinoma. *J. Cell. Physiol.* **2018**, *234*, 7579–7586. [CrossRef]
11. Huo, X.; Sun, H.; Cao, D.; Yang, J.; Peng, P.; Yu, M.; Shen, K. Identification of Prognosis Markers for Endometrial Cancer By Integrated Analysis of DNA Methylation and RNA-Seq Data. *Sci. Rep.* **2019**, *9*, 1–10. [CrossRef] [PubMed]

12. Xu, N.; Wu, Y.; Ke, Z.; Liang, Y.; Cai, H.; Su, W.; Tao, X.; Chen, S.; Zheng, Q.; Wei, Y.; et al. Identification of Key DNA Methylation-Driven Genes in Prostate Adenocarcinoma: An Integrative Analysis of TCGA Methylation Data. *J. Transl. Med.* **2019**, *17*, 1–15. [CrossRef] [PubMed]
13. Wang, G.; Wang, F.; Meng, Z.; Wang, N.; Zhou, C.; Zhang, J.; Zhao, L.; Wang, G.; Shan, B. Uncovering Potential Genes in Colorectal Cancer Based on Integrated and DNA Methylation Analysis in the Gene Expression Omnibus Database. *BMC Cancer* **2022**, *22*, 1–13. [CrossRef] [PubMed]
14. Sun, X.; Wang, M.; Zhang, F.; Kong, X. An Integrated Analysis of Genome-Wide DNA Methylation and Gene Expression Data in Hepatocellular Carcinoma. *FEBS Open Bio* **2018**, *8*, 1093–1103. [CrossRef]
15. Champion, M.; Brennan, K.; Croonenborghs, T.; Gentles, A.; Pochet, N.; Gevaert, O. Module Analysis Captures Pancancer Genetically and Epigenetically Deregulated Cancer Driver Genes for Smoking and Antiviral Response. *EBioMedicine* **2018**, *27*, 156–166. [CrossRef]
16. Dimitrakopoulos, C.; Hindupur, S.; Häfliger, L.; Behr, J.; Montazeri, H.; Hall, M.; Beerenwinkel, N. Network-Based Integration of Multi-Omics Data for Prioritizing Cancer Genes. *Bioinformatics* **2018**, *34*, 2441–2448. [CrossRef]
17. Wang, B.; Mezlini, A.; Demir, F.; Fiume, M.; Tu, Z.; Brudno, M.; Haibe-Kains, B.; Goldenberg, A. Similarity Network Fusion for Aggregating Data Types on A Genomic Scale. *Nat. Methods* **2014**, *11*, 333–337. [CrossRef]
18. Tian, Z.; Guo, M.; Wang, C.; Xing, L.; Wang, L.; Zhang, Y. Constructing an Integrated Gene Similarity Network for the Identification of Disease Genes. *J. Biomed. Semant.* **2017**, *8*, 32. [CrossRef]
19. Pidò, S.; Ceddia, G.; Masseroli, M. Computational Analysis of Fused Co-Expression Networks for the Identification of Candidate Cancer Gene Biomarkers. *npj Syst. Biol. Appl.* **2021**, *7*, 1–10. [CrossRef]
20. Tanvir, R.B.; Maharjan, M.; Mondal, A.M. Community Based Cancer Biomarker Identification from Gene Co-Expression Network. In Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, Niagara Falls, NY, USA, 7–10 September 2019; p. 545.
21. Wang, Y.; Liu, Z. Identifying Biomarkers for Breast Cancer by Gene Regulatory Network Rewiring. *BMC Bioinform.* **2021**, *22*, 1–14. [CrossRef]
22. Yu, L.; Huang, Q.; Zhou, X. Identification of Cancer Hallmarks Based on the Gene Co-Expression Networks of Seven Cancers. *Front. Genet.* **2019**, *10*, 99. [CrossRef]
23. Marcucci, G.; Yan, P.; Maharry, K.; Frankhouser, D.; Nicolet, D.; Metzeler, K.; Kohlschmidt, J.; Mrózek, K.; Wu, Y.; Bucci, D.; et al. Epigenetics Meets Genetics in Acute Myeloid Leukemia: Clinical Impact of a Novel Seven-Gene Score. *J. Clin. Oncol.* **2014**, *32*, 548–556. [CrossRef]
24. Hu, C.; Zhou, Y.; Liu, C.; Kang, Y. A Novel Scoring System for Gastric Cancer Risk Assessment Based on the Expression of Three CLIP4 DNA Methylation-Associated Genes. *Int. J. Oncol.* **2018**, *53*, 633–643. [CrossRef]
25. Kaur, H.; Dhall, A.; Kumar, R.; Raghava, G. Identification of Platform-Independent Diagnostic Biomarker Panel for Hepatocellular Carcinoma Using Large-Scale Transcriptomics Data. *Front. Genet.* **2020**, *10*, 1306. [CrossRef]
26. The International Cancer Genome Consortium. International Network of Cancer Genome Projects. *Nature* **2010**, *464*, 993–998. [CrossRef]
27. GDC. Available online: https://portal.gdc.cancer.gov/ (accessed on 20 February 2022).
28. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **1995**, *57*, 289–300. [CrossRef]
29. Silva, T.; Coetzee, S.; Gull, N.; Yao, L.; Hazelett, D.; Noushmehr, H.; Lin, D.; Berman, B. ELMER V.2: An R/Bioconductor Package to Reconstruct Gene Regulatory Networks from DNA Methylation and Transcriptome Profiles. *Bioinformatics* **2018**, *35*, 1974–1977. [CrossRef]
30. Carlson, M. org.Hs.eg.db: Genome Wide Annotation for Human. 2019. Available online: http://bioconductor.org/packages/release/data/annotation/html/org.Hs.eg.db.html (accessed on 1 May 2023).
31. Li, W.; Liu, C.; Zhang, T.; Li, H.; Waterman, M.; Zhou, X. Integrative Analysis of Many Weighted Co-Expression Networks Using Tensor Computation. *PLoS Comput. Biol.* **2011**, *7*, 1001106. [CrossRef]
32. Csárdi, G.; Nepusz, T. The Igraph Software Package for Complex Network Research. *Inter J. Complex Syst.* **2006**, *1695*, 1–9.
33. Clauset, A.; Newman, M.; Moore, C. Finding Community Structure in Very Large Networks. *Phys. Rev. E* **2004**, *70*, 066111. [CrossRef]
34. Rosvall, M.; Bergstrom, C. Maps of Random Walks on Complex Networks Reveal Community Structure. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 1118–1123. [CrossRef] [PubMed]
35. Blondel, V.; Guillaume, J.; Lambiotte, R.; Lefebvre, E. Fast Unfolding of Communities in Large Networks. *J. Stat. Mech. Theory Exp.* **2008**, *10*, 10008. [CrossRef]
36. Datta, S.; Datta, S. Methods for Evaluating Clustering Algorithms for Gene Expression Data Using a Reference Set of Functional Classes. *BMC Bioinform.* **2006**, *7*, 397. [CrossRef] [PubMed]
37. Bruno, G.; Fiori, A.M. Microarray Clustering Analysis. *J. Parallel Distrib. Comput.* **2013**, *73*, 360–370. [CrossRef]
38. Sun, Y.; Sheng, Z.; Ma, C.; Tang, K.; Zhu, R.; Wu, Z.; Shen, R.; Feng, J.; Wu, D.; Huang, D.; et al. Combining Genomic and Network Characteristics for Extended Capability in Predicting Synergistic Drugs for Cancer. *Nat. Commun.* **2015**, *6*, 8481. [CrossRef]
39. Fisher, R.A. On the Interpretation of X2 from Contingency Tables, and the Calculation of P. *J. R Stat. Soc.* **1922**, *85*, 87–94. [CrossRef]

40. Cox, D.R.; Oakes, D. *Analysis of Survival Data*, 1st ed.; Chapman and Hall/CR: Boca Raton, FL, USA, 1984.
41. Kaplan, E.L.; Meier, P. Nonparametric Estimation from Incomplete Observations. *J. Am. Stat. Assoc.* **1958**, *53*, 457–481. [CrossRef]
42. Argelaguet, R.; Velten, B.; Arnol, D.; Dietrich, S.; Zenz, T.; Marioni, J.C.; Buettner, F.; Huber, W.; Stegle, O. Multi-Omics Factor Analysis-a Framework for Unsupervised Integration of Multi-Omics Data Sets. *Mol. Syst. Biol.* **2018**, *14*, e8124. [CrossRef]
43. Fan, S.; Tang, J.; Li, N.; Zhao, Y.; Ai, R.; Zhang, K.; Wang, M.; Du, W.; Wang, W. Integrative Analysis with Expanded DNA Methylation Data Reveals Common Key Regulators and Pathways in Cancers. *Npj Genom. Med.* **2019**, *4*, 2. [CrossRef]
44. Mo, Q.; Wang, S.; Seshan, V.; Olshen, A.; Schultz, N.; Sander, C.; Powers, R.; Ladanyi, M.; Shen, R. Pattern Discovery and Cancer Gene Identification in Integrated Cancer Genomic Data. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 4245–4250. [CrossRef]
45. Qi, L.; Zhou, B.; Chen, J.; Hu, W.; Bai, R.; Ye, C.; Weng, X.; Zheng, S. Significant Prognostic Values of Differentially Expressed-Aberrantly Methylated Hub Genes in Breast Cancer. *J. Cancer* **2019**, *10*, 6618–6634. [CrossRef]
46. Le, H.; Nguyen, V.M.; Nguyen, Q.H.; Le, D.H. A Biphasic Deep Semi-Supervised Framework for Subtype Classification and Biomarker Discovery. *bioRxiv* **2022**.
47. Fiorentino, G.; Visintainer, R.; Domenici, E.; Lauria, M.; Marchetti, L.M.O.U.S.S.E. Multi-Omics Using Subject-Specific Signatures. *Cancers* **2021**, *13*, 3423. [CrossRef]
48. Sheng, H.; Pan, H.; Yao, M.; Xu, L.; Lu, J.; Liu, B.; Shen, J.; Shen, H. Integrated Analysis of Circular RNA-Associated Cerna Network Reveals Potential Circrna Biomarkers in Human Breast Cancer. *Comput. Math. Methods Med.* **2021**, *2021*, 1732176. [CrossRef]
49. Shi, K.; Gao, L.; Wang, B. Discovering Potential Cancer Driver Genes By an Integrated Network-Based Approach. *Mol. BioSystems* **2016**, *12*, 2921–2931. [CrossRef]
50. Hua, P.; Zhang, Y.; Jin, C.; Zhang, G.; Wang, B. Integration of gene profile to explore the hub genes of lung adenocarcinoma: A quasi-experimental study. *Medicine* **2020**, *99*, 22727. [CrossRef]
51. Shi, K.; Li, N.; Yang, M.; Li, W. Identification of Key Genes and Pathways in Female Lung Cancer Patients Who Never Smoked by a Bioinformatics Analysis. *J. Cancer* **2019**, *10*, 51–60. [CrossRef]
52. Buttarelli, M.; Ciucci, A.; Palluzzi, F.; Raspaglio, G.; Marchetti, C.; Perrone, E.; Minucci, A.; Giacò, L.; Fagotti, A.; Scambia, G.; et al. Identification of a Novel Gene Signature Predicting Response to First-Line Chemotherapy in BRCA Wild-Type High-Grade Serous Ovarian Cancer Patients. *J. Exp. Clin. Cancer Res. CR* **2022**, *41*, 50. [CrossRef]
53. Jiang, M.M.; Zhao, F.; Lou, T.T. Assessment of Significant Pathway Signaling and Prognostic Value of GNG11 in Ovarian Serous Cystadenocarcinoma. *Int. J. Gen. Med.* **2021**, *14*, 2329–2341. [CrossRef]
54. Zhang, X.; Kang, X.; Jin, L. ABCC9, NKAPL, and TMEM132C Are Potential Diagnostic and Prognostic Markers in Triple-Negative Breast Cancer. *Cell Biol. Int.* **2020**, *44*, 2002–2010. [CrossRef] [PubMed]
55. Xing, S.; Wang, Y.; Hu, K.; Wang, F.; Sun, T.; Li, Q. WGCNA Reveals Key Gene Modules Regulated by the Combined Treatment of Colon Cancer with PHY906 and CPT11. *Biosci. Rep.* **2020**, *40*, 20200935. [CrossRef] [PubMed]
56. Clermont, P.; Crea, F.; Chiang, Y.; Lin, D.; Zhang, A.; Wang, J.; Parolia, A.; Wu, R.; Xue, H.; Wang, Y.; et al. Identification of the Epigenetic Reader CBX2 As A Potential Drug Target in Advanced Prostate Cancer. *Clin. Epigenetics* **2016**, *8*, 16. [CrossRef] [PubMed]
57. Mao, J.; Tian, Y.; Wang, C.; Jiang, K.; Li, R.; Yao, Y.; Zhang, R.; Sun, D.; Liang, R.; Gao, Z.; et al. CBX2 Regulates Proliferation and Apoptosis Via the Phosphorylation of YAP in Hepatocellular Carcinoma. *J. Cancer* **2019**, *10*, 2706–2719. [CrossRef] [PubMed]
58. Wheeler, L.; Watson, Z.; Qamar, L.; Yamamoto, T.; Post, M.; Berning, A.; Spillman, M.; Behbakht, K.; Bitler, B. CBX2 Identified as Driver of Anoikis Escape and Dissemination in High Grade Serous Ovarian Cancer. *Oncogenesis* **2018**, *7*, 92. [CrossRef]
59. Hu, F.; Chen, H.; Duan, Y.; Lan, B.; Liu, C.; Hu, H.; Dong, X.; Zhang, Q.; Cheng, Y.; Liu, M.; et al. CBX2 and EZH2 Cooperatively Promote the Growth and Metastasis of Lung Adenocarcinoma. *Mol. Ther. Nucleic Acids* **2022**, *27*, 670–684. [CrossRef]
60. Ma, T.; Ma, N.; Chen, J.; Tang, F.; Zong, Z.; Yu, Z.; Chen, S.; Zhou, T. Expression and Prognostic Value of Chromobox Family Members in Gastric Cancer. *J. Gastrointest. Oncol.* **2020**, *11*, 983–998. [CrossRef]
61. Li, Q.; Pan, Y.; Cao, Z.; Zhao, S. Comprehensive Analysis of Prognostic Value and Immune Infiltration of Chromobox Family Members in Colorectal Cancer. *Front. Oncol.* **2020**, *10*, 582667. [CrossRef]
62. Zhou, H.; Xiong, Y.; Liu, Z.; Hou, S.; Zhou, T. Expression and Prognostic Significance of CBX2 in Colorectal Cancer: Database Mining for CBX Family Members in Malignancies and Vitro Analyses. *Cancer Cell Int.* **2021**, *21*, 1–6.
63. Bilton, L.; Warren, C.; Humphries, R.; Kalsi, S.; Waters, E.; Francis, T.; Dobrowinski, W.; Beltran-Alvarez, P.; Wade, M. The Epigenetic Regulatory Protein CBX2 Promotes Mtorc1 Signalling and Inhibits DREAM Complex Activity to Drive Breast Cancer Cell Growth. *Cancers* **2022**, *14*, 3491. [CrossRef]
64. Zheng, S.; Lv, P.; Su, J.; Miao, K.; Xu, H.; Li, M. Overexpression of CBX2 in Breast Cancer Promotes Tumor Progression through the PI3K/AKT Signaling Pathway. *Am. J. Transl. Res.* **2019**, *11*, 1668–1682.
65. Li, X.; Gou, J.; Li, H.; Yang, X. Bioinformatic Analysis of the Expression and Prognostic Value of Chromobox Family Proteins in Human Breast Cancer. *Sci. Rep.* **2020**, *10*, 1–11. [CrossRef]
66. Piqué, D.; Montagna, C.; Greally, J.; Mar, J. A Novel Approach To Modelling Transcriptional Heterogeneity Identifies the Oncogene Candidate CBX2 in Invasive Breast Carcinoma. *Br. J. Cancer* **2019**, *120*, 746–753. [CrossRef]
67. Malumbres, M.; Barbacid, M. Cell Cycle, Cdks and Cancer: A Changing Paradigm. *Nat. Rev. Cancer* **2009**, *9*, 153–166. [CrossRef]
68. Hannon, G.; Casso, D.; Beach, D. KAP: A Dual Specificity Phosphatase That Interacts with Cyclin-Dependent Kinases. *Proc. Natl. Acad. Sci. USA* **1994**, *91*, 1731–1735. [CrossRef]

69. Abdel-Tawab, M.; Fouad, H.; Othman, A.; Eid, R.; Mohammed, M.; Hassan, A. Reyad Evaluation of Gene Expression of PLEKHS1, AADAC, and CDKN3 As Novel Genomic Markers in Gastric Carcinoma. *PLoS ONE* **2022**, *17*, 0265184. [CrossRef]
70. Li, Y.; Ji, S.; Fu, L.; Jiang, T.; Wu, D.; Meng, F. Knockdown of Cyclin-Dependent Kinase Inhibitor 3 Inhibits Proliferation and Invasion in Human Gastric Cancer Cells. *Oncol. Res. Featur. Preclin. Clin. Cancer Ther.* **2017**, *25*, 721–731. [CrossRef]
71. Chang, S.; Chen, T.; Lee, Y.; Lee, S.; Lin, L.; He, H. CDKN3 Expression Is an Independent Prognostic Factor and Associated with Advanced Tumor Stage in Nasopharyngeal Carcinoma. *Int. J. Med. Sci.* **2018**, *15*, 992–998. [CrossRef]
72. Fan, C.; Chen, L.; Huang, Q.; Shen, T.; Welsh, E.; Teer, J.; Cai, J.; Cress, W.; Wu, J. Overexpression of Major CDKN3 Transcripts Is Associated with Poor Survival in Lung Adenocarcinoma. *Br. J. Cancer* **2015**, *113*, 1735–1743. [CrossRef]
73. Jin, H.; Huang, X.; Shao, K.; Li, G.; Wang, J.; Yang, H.; Hou, Y. Integrated Bioinformatics Analysis To Identify 15 Hub Genes in Breast Cancer. *Oncol. Lett.* **2019**, *18*, 1023–1034. [CrossRef]
74. Li, M.; Che, N.; Jin, Y.; Li, J.; Yang, W. CDKN3 Overcomes Bladder Cancer Cisplatin Resistance Via LDHA-Dependent Glycolysis Reprogramming. *OncoTargets Ther.* **2022**, *15*, 299–311. [CrossRef]
75. Barrón, E.; Roman-Bassaure, E.; Sánchez-Sandoval, A.; Espinosa, A.; Guardado-Estrada, M.; Medina, I.; Juárez, E.; Alfaro, A.; Bermúdez, M.; Zamora, R.; et al. CDKN3 Mrna As A Biomarker for Survival and Therapeutic Target in Cervical Cancer. *PLoS ONE* **2015**, *10*, 0137397. [CrossRef] [PubMed]
76. Wang, W.; Liao, K.; Guo, H.; Zhou, S.; Yu, R.; Liu, Y.; Pan, Y.; Pu, J. Integrated Transcriptomics Explored the Cancer-Promoting Genes CDKN3 in Esophageal Squamous Cell Cancer. *J. Cardiothorac. Surg.* **2021**, *16*, 1–7. [CrossRef] [PubMed]
77. Yang, C.; Sun, J. Mechanistic Studies of Cyclin-Dependent Kinase Inhibitor 3 (CDKN3) in Colorectal Cancer. *Asian Pac. J. Cancer Prev.* **2015**, *16*, 965–970. [CrossRef] [PubMed]
78. Li, W.H.; Zhang, L.; Wu, Y.H. CDKN3 Regulates Cisplatin Resistance to Colorectal Cancer through TIPE1. *Eur. Rev. Med. Pharmacol. Sci.* **2020**, *24*, 3614–3623.
79. Joseph, J.; Radulovich, N.; Wang, T. Rho guanine nucleotide exchange factor ARHGEF10 is a putative tumor suppressor in pancreatic ductal adenocarcinoma. *Oncogene* **2020**, *39*, 308–321. [CrossRef]
80. Xue, W.; Kitzing, T.; Roessler, S.; Zuber, J.; Krasnitz, A.; Schultz, N.; Revill, K.; Weissmueller, S.; Rappaport, A.R.; Simon, J.; et al. A Cluster of Cooperating Tumor-Suppressor Gene Candidates in Chromosomal Deletions. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 8212–8217. [CrossRef]
81. Cooke, S.L.; Pole, J.C.; Chin, S.F.; Ellis, I.O.; Caldas, C.; Edwards, P.A. High-Resolution Array CGH Clarifies Events Occurring on 8p in Carcinogenesis. *BMC Cancer* **2008**, *8*, 288. [CrossRef]
82. Williams, S.V.; Taylor, C.; Platt, F.; Hurst, C.D.; Aveyard, J.; Knowles, M.A. Mutation and Homozygous Deletion of ARHGEF10 in Bladder Cancer; a Candidate Tumour Suppressor Gene at 8p23. 3. *Cancer Genet. Cytogenet.* **2010**, *203*, 68. [CrossRef]
83. Ranta, S.; Zhang, Y.; Ross, B. The Neuronal Ceroid Lipofuscinoses in Human EPMR and Mnd Mutant Mice Are Associated with Mutations in CLN8. *Nat. Genet.* **1999**, *23*, 233–236. [CrossRef]
84. Yap, S.Q.; Mathavarajah, S.; Huber, R.J. The Converging Roles of Batten Disease Proteins in Neurodegeneration and Cancer. *iScience* **2021**, *24*, 102337. [CrossRef]
85. Zimmermann, R.; Müller, L.; Wullich, B. Protein Transport Into the Endoplasmic Reticulum: Mechanisms and Pathologies. *Trends Mol. Med.* **2006**, *12*, 567–573. [CrossRef]
86. Zhang, Y.; Yang, F.; Peng, X.; Li, X.; Luo, N.; Zhu, W.; Fu, M.; Li, Q.; Hu, G. Hypoxia Constructing the Prognostic Model of Colorectal Adenocarcinoma and Related To the Immune Microenvironment. *Front. Cell Dev. Biol.* **2021**, *9*, 665364. [CrossRef]
87. Ma, J.; He, Z.; Zhang, H.; Zhang, W.; Gao, S.; Ni, X. SEC61G Promotes Breast Cancer Development and Metastasis Via Modulating Glycolysis and Is Transcriptionally Regulated By E2F1. *Cell Death Dis.* **2021**, *12*, 1–14. [CrossRef]
88. Jin, L.; Chen, D.; Hirachan, S.; Bhandari, A.; Huang, Q. SEC61G Regulates Breast Cancer Cell Proliferation and Metastasis By Affecting the Epithelial-Mesenchymal Transition. *J. Cancer* **2022**, *13*, 831–846. [CrossRef]
89. Zhang, Y.; Fang, L.; Zang, Y.; Xu, Z. Identification of Core Genes and Key Pathways Via Integrated Analysis of Gene Expression and DNA Methylation Profiles in Bladder Cancer. *Med. Sci. Monit.* **2018**, *24*, 3024–3033. [CrossRef]
90. Meng, H.; Jiang, X.; Wang, J.; Sang, Z.; Guo, L.; Yin, G.; Wang, Y. SEC61G Is Upregulated and Required for Tumor Progression in Human Kidney Cancer. *Mol. Med. Rep.* **2021**, *23*, 427. [CrossRef]
91. Gao, H.; Niu, W.; He, Z.; Gao, C.; Peng, C.; Niu, J. SEC61G Plays an Oncogenic Role in Hepatocellular Carcinoma Cells. *Cell Cycle* **2020**, *19*, 3348–3361. [CrossRef]
92. Lu, T.; Chen, Y.; Gong, X.; Guo, Q.; Lin, C.; Luo, Q.; Tu, Z.; Pan, J.; Li, J. SEC61G Overexpression and DNA Amplification Correlates with Prognosis and Immune Cell Infiltration in Head and Neck Squamous Cell Carcinoma. *Cancer Med.* **2021**, *10*, 7847–7862. [CrossRef]
93. Liu, B.; Liu, J.; Liao, Y.; Jin, C.; Zhang, Z.; Zhao, J.; Liu, K.; Huang, H.; Cao, H.; Cheng, Q. Identification of SEC61G as a Novel Prognostic Marker for Predicting Survival and Response to Therapies in Patients with Glioblastoma. *Med. Sci. Monit.* **2019**, *25*, 3624–3635. [CrossRef]
94. Zheng, Q.; Wang, Z.; Zhang, M.; Yu, Y.; Chen, R.; Lu, T.; Liu, L.; Ma, J.; Liu, T.; Zheng, H.; et al. Prognostic Value of SEC61G in Lung Adenocarcinoma: A Comprehensive Study Based on Bioinformatics and In Vitro Validation. *BMC Cancer* **2021**, *21*, 1216. [CrossRef]
95. Cheng, C.; Wang, T.; Chen, P.; Wu, W.; Lai, J.; Chang, P.; Hong, Y.; Huang, C.; Wang, F. Computer-Aided Design for Identifying Anticancer Targets in Genome-Scale Metabolic Models of Colon Cancer. *Biology* **2021**, *10*, 1115. [CrossRef] [PubMed]

96. Sekar, D.; Dillmann, C.; Sirait-Fischer, E.; Fink, A.; Zivkovic, A.; Baum, N.; Strack, E.; Klatt, S.; Zukunft, S.; Wallner, S.; et al. Phosphatidylserine Synthase PTDSS1 Shapes the Tumor Lipidome to Maintain Tumor-Promoting Inflammation. *Cancer Res.* **2022**, *82*, 1617–1632. [CrossRef] [PubMed]
97. Yang, T.; Hui, R.; Nouws, J.; Sauler, M.; Zeng, T.; Wu, Q. Untargeted Metabolomics Analysis of Esophageal Squamous Cell Cancer Progression. *J. Transl. Med.* **2022**, *20*, 127. [CrossRef]
98. N'Guessan, K.; Davis, H.; Chu, Z.; Vallabhapurapu, S.; Lewis, C.; Franco, R.; Olowokure, O.; Ahmad, S.; Yeh, J.; Bogdanov, V.; et al. Enhanced Efficacy of Combination of Gemcitabine and Phosphatidylserine-Targeted Nanovesicles against Pancreatic Cancer. *Mol. Ther.* **2020**, *28*, 1876–1886. [CrossRef]
99. Li, M.; Xu, D.; Lin, S.; Yang, Z.; Xu, T.; Yang, J.; Lin, Z.; Huang, Z. Transcriptional Expressions of Hsa-Mir-183 Predicted Target Genes As Independent Indicators for Prognosis in Bladder Urothelial Carcinoma. *Aging* **2022**, *14*, 3782–3800. [CrossRef] [PubMed]
100. Wang, Y.; Lin, M.; Chen, W.; Wu, W.; Wang, F. Optimization of A Modeling Platform To Predict Oncogenes from Genome-Scale Metabolic Networks of Non-Small-Cell Lung Cancers. *FEBS Open Bio* **2021**, *11*, 2078–2094. [CrossRef]
101. Cheng, C.; Hua, J.; Tan, J.; Qian, W.; Zhang, L.; Hou, X. Identification of Differentially Expressed Genes, Associated Functional Terms Pathways, and Candidate Diagnostic Biomarkers in Inflammatory Bowel Diseases by Bioinformatics Analysis. *Exp. Ther. Med.* **2019**, *18*, 278–288. [CrossRef]
102. Moradi, S.; Tapak, L.; Afshar, S. Identification of Novel Noninvasive Diagnostics Biomarkers in the Parkinson's Diseases and Improving the Disease Classification Using Support Vector Machine. *BioMed Res. Int.* **2022**, *2022*, 5009892. [CrossRef]
103. Nolte, I.M.; Munoz, M.L.; Tragante, V.; Amare, A.T.; Jansen, R.; Vaez, A.; von der Heyde, B.; Avery, C.L.; Bis, J.C.; Dierckx, B.; et al. Genetic Loci Associated with Heart Rate Variability and Their Effects on Cardiac Disease Risk. *Nat. Commun.* **2017**, *8*, 15805. [CrossRef]
104. Ou, H.; Fan, Y.; Guo, X.; Lao, Z.; Zhu, M.; Li, G.; Zhao, L. Identifying Key Genes Related to Inflammasome in Severe COVID-19 Patients Based on a Joint Model with Random Forest and Artificial Neural Network. *Front. Cell. Infect. Microbiol.* **2023**, *13*, 1139998. [CrossRef]
105. Yue, Q.; Li, Z.; Zhang, Q.; Jin, Q.; Zhang, X.; Jin, G. Identification of Novel Hub Genes Associated with Psoriasis Using Integrated Bioinformatics Analysis. *Int. J. Mol. Sci.* **2022**, *23*, 15286. [CrossRef]
106. Yao, S.; Deng, M.; Du, X.; Huang, R.; Chen, Q. A Novel Hypoxia Related Marker in Blood Link to Aid Diagnosis and Therapy in Osteoarthritis. *Genes* **2022**, *13*, 1501. [CrossRef]
107. Verhoeven, K.; De Jonghe, P.; Van de Putte, T.; Nelis, E.; Zwijsen, A.; Verpoorten, N.; De Vriendt, E.; Jacobs, A.; Van Gerwen, V.; Francis, A.; et al. Slowed Conduction and Thin Myelination of Peripheral Nerves Associated with Mutant Rho Guanine-Nucleotide Exchange Factor 10. *Am. J. Hum. Genet.* **2003**, *73*, 926–932. [CrossRef]
108. Kim, S.M.; Park, M.J.; Park, S.; Cheng, J.Y.; Lee, E.-S. Differential Expression of Novel Genes and Signalling Pathways of Senescent CD8+ T Cell Subsets in Behçet's Disease. *Clin. Exp. Rheumatol.* **2020**, *38* (Suppl. S127), 17–25.
109. Zhang, C.K.; Stein, P.B.; Liu, J.; Wang, Z.; Yang, R.; Cho, J.H.; Gregersen, P.K.; Aerts, J.M.F.G.; Zhao, H.; Pastores, G.M.; et al. Genome-Wide Association Study of N370S Homozygous Gaucher Disease Reveals the Candidacy of CLN8 Gene as a Genetic Modifier Contributing to Extreme Phenotypic Variation. *Am. J. Hematol.* **2012**, *87*, 377–383. [CrossRef]
110. Sahin, Y.; Güngör, O.; Gormez, Z.; Demirci, H.; Ergüner, B.; Güngör, G.; Dilber, C. Exome Sequencing Identifies a Novel Homozygous CLN8 Mutation in a Turkish Family with Northern Epilepsy. *Acta Neurol. Belg* **2017**, *117*, 159–167. [CrossRef]
111. Norling, A.; Hirschberg, A.L.; Iwarsson, E.; Wedell, A.; Barbaro, M. CBX2 Gene Analysis in Patients with 46,XY and 46,XX Gonadal Disorders of Sex Development. *Fertil. Steril.* **2013**, *99*, 819–826.e3. [CrossRef]
112. He, H.; Yang, Y.; Wang, L.; Guo, Z.; Ye, L.; Ou-Yang, W.; Yang, M. Combined Analysis of Single-Cell and Bulk RNA Sequencing Reveals the Expression Patterns of Circadian Rhythm Disruption in the Immune Microenvironment of Alzheimer's Disease. *Front. Immunol.* **2023**, *14*, 1182307. [CrossRef]
113. Staneva, R.; Rukova, B.; Hadjidekova, S.; Nesheva, D.; Antonova, O.; Dimitrov, P.; Simeonov, V.; Stamenov, G.; Cukuranovic, R.; Cukuranovic, J.; et al. Whole Genome Methylation Array Analysis Reveals New Aspects in Balkan Endemic Nephropathy Etiology. *BMC Nephrol.* **2013**, *14*, 225. [CrossRef]
114. Krzyzewska, I.M.; Lauffer, P.; Mul, A.N.; van der Laan, L.; Yim, A.Y.F.L.; Cobben, J.M.; Niklinski, J.; Chomczyk, M.A.; Smigiel, R.; Mannens, M.M.A.M.; et al. Expression Quantitative Trait Methylation Analysis Identifies Whole Blood Molecular Footprint in Fetal Alcohol Spectrum Disorder (FASD). *Int. J. Mol. Sci.* **2023**, *24*, 6601. [CrossRef]
115. Tamhankar, P.M.; Vasudevan, L.; Bansal, V.; Menon, S.R.; Gawde, H.M.; D'Souza, A.; Babu, S.; Kondurkar, S.; Adhia, R.; Das, D.K. Lenz-Majewski Syndrome: Report of a Case with Novel Mutation in PTDSS1 Gene. *Eur. J. Med. Genet.* **2015**, *58*, 392–399. [CrossRef] [PubMed]
116. Soueid, J.; Kourtian, S.; Makhoul, N.J.; Makoukji, J.; Haddad, S.; Ghanem, S.S.; Kobeissy, F.; Boustany, R.-M. RYR2, PTDSS1 and AREG Genes Are Implicated in a Lebanese Population-Based Study of Copy Number Variation in Autism. *Sci. Rep.* **2016**, *6*, 19088. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.