International Journal of
Molecular Sciences

MDPI

Article

Unveiling Prognostic RNA Biomarkers through a Multi-Cohort Study in Colorectal Cancer

Zehwan Kim [1,†], Jaebon Lee [2,†], Ye Eun Yoon [2] and Jae Won Yun [2,*]

[1] Department of Laboratory Medicine, Yeungnam University College of Medicine, Daegu 42415, Republic of Korea
[2] Veterans Health Service Medical Research Institute, Veterans Health Service Medical Center, Seoul 05368, Republic of Korea
* Correspondence: jwyunmd@gmail.com
† These authors contributed equally to this work.

Abstract: Because cancer is a leading cause of death and is thought to be caused by genetic errors or genomic instability in many circumstances, there have been studies exploring cancer's genetic basis using microarray and RNA-seq methods, linking gene expression data to patient survival. This research introduces a methodological framework, combining heterogeneous gene expression data, random forest selection, and pathway analysis, alongside clinical information and Cox regression analysis, to discover prognostic biomarkers. Heterogeneous gene expression data for colorectal cancer were collected from TCGA-COAD (RNA-seq), and GSE17536 and GSE39582 (microarray), and were integrated with Entrez Gene IDs. Using Cox regression analysis and random forest, genes with consistent hazard ratios and significantly affecting patient survivability were chosen. Predictive accuracy was evaluated using ROC curves. Pathway analysis identified potential RNA biomarkers. The authors identified 28 RNA biomarkers. Pathway analysis revealed enrichment in cancer-related pathways, notably EGFR downstream signaling and IGF1R signaling. Three RNA biomarkers (ZEB1-AS1, PI4K2A, and ITGB8-AS1) and two clinical biomarkers (stage and age) were chosen for a prognostic model, improving predictive performance compared to using clinical biomarkers alone. Despite biomarker identification challenges, this study underscores integration of heterogenous gene expression data for discovery.

Keywords: prognostic biomarkers; colorectal cancer; gene expression; RNA

check for updates

Citation: Kim, Z.; Lee, J.; Yoon, Y.E.; Yun, J.W. Unveiling Prognostic RNA Biomarkers through a Multi-Cohort Study in Colorectal Cancer. *Int. J. Mol. Sci.* **2024**, *25*, 3317. https://doi.org/10.3390/ijms25063317

Academic Editor: Peter J. K. Kuppen

Received: 14 January 2024
Revised: 12 March 2024
Accepted: 12 March 2024
Published: 14 March 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Cancer is a leading cause of death [1] and colorectal cancer is the third deadliest cancer according to the epidemiology of colorectal cancer (WHO) and the second leading cause in the United States [2–4]. For the treatment of cancer, epidemiological studies, identification of causes, factors affecting prognosis, refinement of diagnosis, and selection of the most effective treatment for each diagnosed subtype of cancer are important. In many cases, cancer is thought to be caused by genetic errors or genomic instability [5]. In this regard, research has been ongoing to obtain information on which genes affect diagnosis, treatment, and prognosis, and precision medicine using genomic information is playing an important role [4,6–8].

An important part of precision medicine using genomic information is the study of gene expression. Experimental methods to check gene expression in tissues obtained from colorectal cancer have been developed, and methods with relatively low cost and effort have been developed. Among the various methods, microarray technology [9–16] or next generation sequencing (NGS)-based RNA-Seq [17–20] are the main examples.

Microarray technology, pioneered by companies such as Affymetrix and Illumina, is vital for studying mRNA abundance and gene expression. The technology employs numerous probes on a microarray platform to identify genes with significant mRNA production, forming probe sets with multiple designed probes per gene. Utilized extensively

in experiments, microarray data are well-supported by robust statistical methods and established protocols. The majority of microarray data from reported studies are publicly accessible, systematically archived in databases like Gene Express Omnibus (GEO) [21] and ArrayExpress [22,23].

The RNA-seq method uses NGS technology to measure the amount of mRNA. Using the mRNA produced by the cells of a tissue of interest, complementary DNA is transcribed and the sequence of the complementary DNA can be sequenced to indirectly determine which genes are producing the mRNA. The Cancer Genome Atlas (TCGA) is a large-scale project that collects patient information as well as tissues from many types of cancer disease [24]. It also provides publicly available experimental data in terms of various aspects, such as copy number, DNA methylation, and gene expression RNA-seq.

In this study, we developed a systematic approach to discover prognostic biomarkers using a multi-cohort study and an artificial intelligence method. The aims of our study are, first, to establish a methodology to systematically identify prognostic biomarkers using multiple multi-cohort data with minimal false positives (Figure 1). The second is to use this methodology to systematically identify and describe RNA biomarkers that reflect prognosis independently of clinical parameters such as age and stage in colorectal cancer.

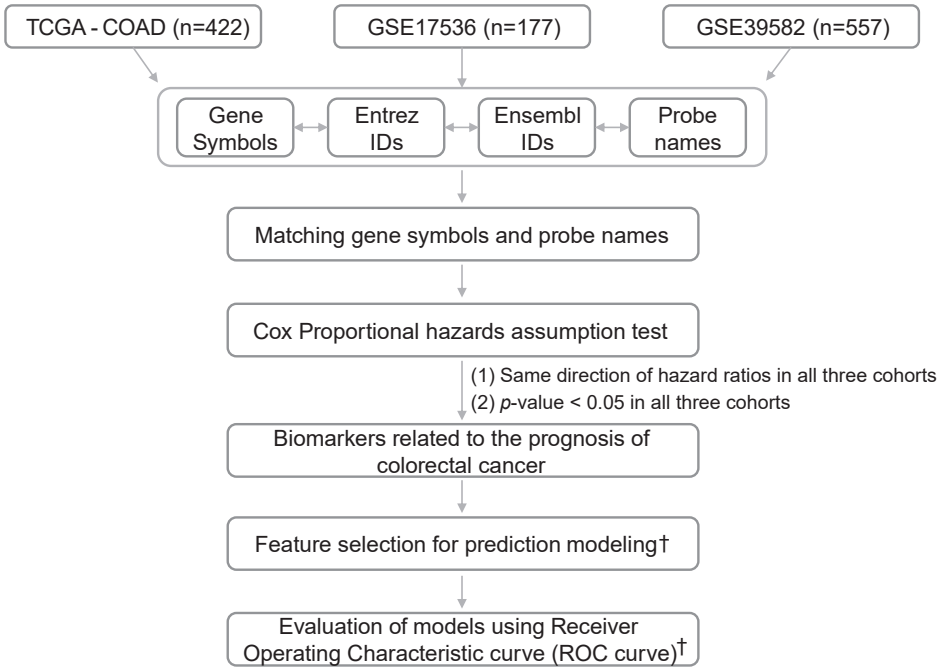

Figure 1. Flow chart of this study. A total of three cohorts were selected for the further analysis of finding biomarkers related to the prognosis of colorectal cancer and modeling the prediction of survival events. †: Details are described in Section 2.4.

2. Results

2.1. Characteristics of Each Cohort

Patient characteristics and pathologic information of the three cohorts are summarized in Table 1. The mean age of patients from all three cohorts was similar, i.e., 66.4, 65.5, and 66.8 years, respectively. Regarding the gender distribution, all cohorts showed a similar proportion of female patients of around 45%. In terms of tumor stage, lower stages (stages I and II) exhibited comparable proportions of 56.4%, 45.8%, and 52.6%, respectively. Pathologic information was available only in TCGA data. According to the TCGA cohort, the most frequent histologic type was adenocarcinoma, NOS, followed by mucinous adenocarcinoma. These comprised 356 (84.4%) and 61 (14.5%) of the 422 specimens, respectively.

Table 1. Clinical and histological characteristics of three cohorts.

Features		TCGA (n = 422)	GSE17536 (n = 177)	GSE39582 (n = 557)
		Case Number (Proportion)		
Age (Mean and SD)		66.4 ± 12.9	65.5 ± 13.1	66.8 ± 13.3
Gender	Male	256 (53.6%)	96 (54.2%)	305 (54.8%)
	Female	196 (46.4%)	81 (45.8%)	252 (45.2%)
Stage	I	73 (17.3%)	24 (13.6%)	31 (5.6%)
	II	165 (39.1%)	57 (32.2%)	262 (47.0%)
	III	123 (29.1%)	57 (32.2%)	204 (36.6%)
	IV	61 (14.5%)	39 (22.0%)	60 (10.8%)
MSI	Yes (or dMMR)	11 (2.6%)	N/A	71 (12.7%)
	No (or pMMR)	79 (18.7%)	N/A	440 (79.0%)
Histologic type	Adenocarcinoma	356 (84.4%)	N/A	N/A
	Mucinous adenocarcinoma	61 (14.5%)	N/A	N/A
	Other types	5 (1.1%)	N/A	N/A

N/A, not available.

2.2. Cox Regression Analysis Test Result

Utilizing a multiple Cox regression model that accounted for factors such as age, sex, and tumor stage, 28 RNA biomarkers were identified by filtering for those with significant *p*-values and consistent hazard ratio direction (HR > 1 in all cohorts or HR < 1 in all cohorts) across the three cohorts (Table 2).

Table 2. Results of the Cox regression analysis. After excluding genes that did not satisfy the Cox proportional hazard assumptions, 28 genes were selected based on the following criteria: (1) the direction of hazard ratios is the same in all three cohorts; and (2) the *p*-value of gene expression levels is less than 0.05 in all three cohorts. The hazard ratios (HRs) and *p*-values shown in the table are derived from a multiple Cox regression model with four variables: sex, age, stage, and gene expression level.

Gene Symbol (Probe Name)	TCGA-COAD (n = 422)		GSE17536 (n = 177)		GSE39582 (n = 557)	
	HR	*p*-Value	HR	*p*-Value	HR	*p*-Value
PI4K2A (209345_s_at)	1.766	0.03	9.269	0	1.571	0.022
BAHD1 (203051_at)	1.936	0.011	4.052	0.005	1.67	0.031
MFAP1 (203406_at)	1.84	0.022	2.561	0.006	1.407	0.049
OAZ2 (201365_at)	1.829	0.041	2.392	0.026	1.51	0.043
FAM219B (224804_s_at)	1.815	0.011	2.348	0.016	1.51	0.019
KCTD1 (226246_at)	1.3	0.004	2.583	0.009	1.303	0.014
DNAJA4 (1554334_a_at)	1.6	0.049	2.069	0.002	1.297	0.024
ZEB1-AS1 (229090_at)	1.383	0.018	1.839	0.015	1.463	0.001
OSBPL1A (209485_s_at)	1.262	0.036	1.723	0.006	1.238	0.007
MYRIP (214156_at)	0.901	0.016	0.69	0.002	0.866	0.017
TOX3 (215108_x_at)	0.854	0.033	0.726	0.009	0.844	0.017
F2RL2 (230147_at)	0.869	0.026	0.687	0.016	0.815	0.001
FGFR4 (204579_at)	0.831	0.046	0.673	0.049	0.852	0.047
KRTAP4-1 (234635_at)	0.937	0.043	0.547	0.009	0.762	0.025
CHN2 (207486_x_at)	0.774	0.013	0.527	0.013	0.78	0.023
NFE2L3 (236471_at)	0.7	0.041	0.625	0.004	0.754	0.001

Table 2. *Cont.*

Gene Symbol (Probe Name)	TCGA-COAD (n = 422)		GSE17536 (n = 177)		GSE39582 (n = 557)	
	HR	*p*-Value	HR	*p*-Value	HR	*p*-Value
TTTY14 (207063_at)	0.899	0.013	0.384	0.046	0.774	0.05
CACNA1D (1555993_at)	0.738	0	0.518	0.037	0.776	0.021
SLC18A1 (207074_s_at)	0.939	0.045	0.251	0.007	0.775	0.025
CMC1 (228283_at)	0.644	0.028	0.519	0.035	0.706	0.012
ITGB8-AS1 (230446_at)	0.8	0.021	0.416	0.002	0.638	0.001
LARS2 (204016_at)	0.621	0.019	0.459	0.027	0.661	0.017
RAB11FIP4 (225739_at)	0.681	0.042	0.466	0.003	0.576	0
LINC00511 (230812_at)	0.798	0.043	0.302	0.036	0.611	0.015
CHDH (1559591_s_at)	0.739	0.01	0.314	0.004	0.648	0.008
HNF1B (208135_at)	0.842	0.037	0.053	0	0.598	0.008
PDPK1 (221244_s_at)	0.469	0.001	0.181	0.012	0.582	0.027
EIF1B (237988_at)	0.602	0.017	0.025	0.006	0.333	0.015

Next, pathway analysis was conducted to assess the enrichment of 28 RNA biomarkers in cancer-related pathways using CancerCompass (https://cancercompass.newgenes.org/, accessed on 6 January 2024). The top 10 cancer-related pathways included EGFR (ERBB1) downstream signaling, FGF signaling pathway, IGF1R signaling cascade, IRS-mediated signaling, and PI3K cascade (Supplementary Figure S1A,B).

Interestingly, a number of genes were found to be part of the EGFR downstream signaling pathway. *CHN2* (chimerin 2), *F2RL2* (coagulation factor II thrombin receptor like 2), and *PDPK1* (3-phosphoinositide dependent protein kinase 1) were genes in the EGFR downstream signaling pathway. The association of the genes with cancer was analyzed through a literature review and gene ontology analysis. The *CHN2* gene encodes a GTP-metabolizing protein crucial for cell proliferation and migration. The *F2RL2* gene showed a significant correlation with colorectal cancer initiation and progression in a prior study [25]. *PDPK1* has been implicated in treatment resistance and cancer cell growth across several cancer types [26,27]. Additionally, multiple insulin-like growth factor-related pathways were identified, and *FGFR4* and *PDPK1* were identified as key genes involved in the IGF1R signaling cascade and IRS-mediated signaling. FGFR4 was also reported for its important role in tumor progression, oncogenesis, and the development of treatment resistance [28].

2.3. Identification of Interaction among Hub Genes Using Network Analysis

The results from the Gephi 0.10.1 software used for network analysis of RNA biomarkers or genes are shown in Figure 2. Of the 28 genes obtained by Cox regression analysis, only 22 genes had an edge with an absolute value of the Pearson correlation R-value greater than or equal to 0.3, which were categorized into four classes. *KCTD1*, *FGFR4*, and *CACNA1D* were the most important hub genes as they were related to more than eight genes, and *HNF1B*, *LARS2*, *MYRIP*, *CHDH*, *OAZ2*, *ITGB8-AS1*, *CHN2*, and *TOX3* were also hub genes as they were related to more than four genes. *KCT1D* was highly associated with other genes in a negative correlation, and seven of the nine edges tended to be negatively co-expressed with other genes.

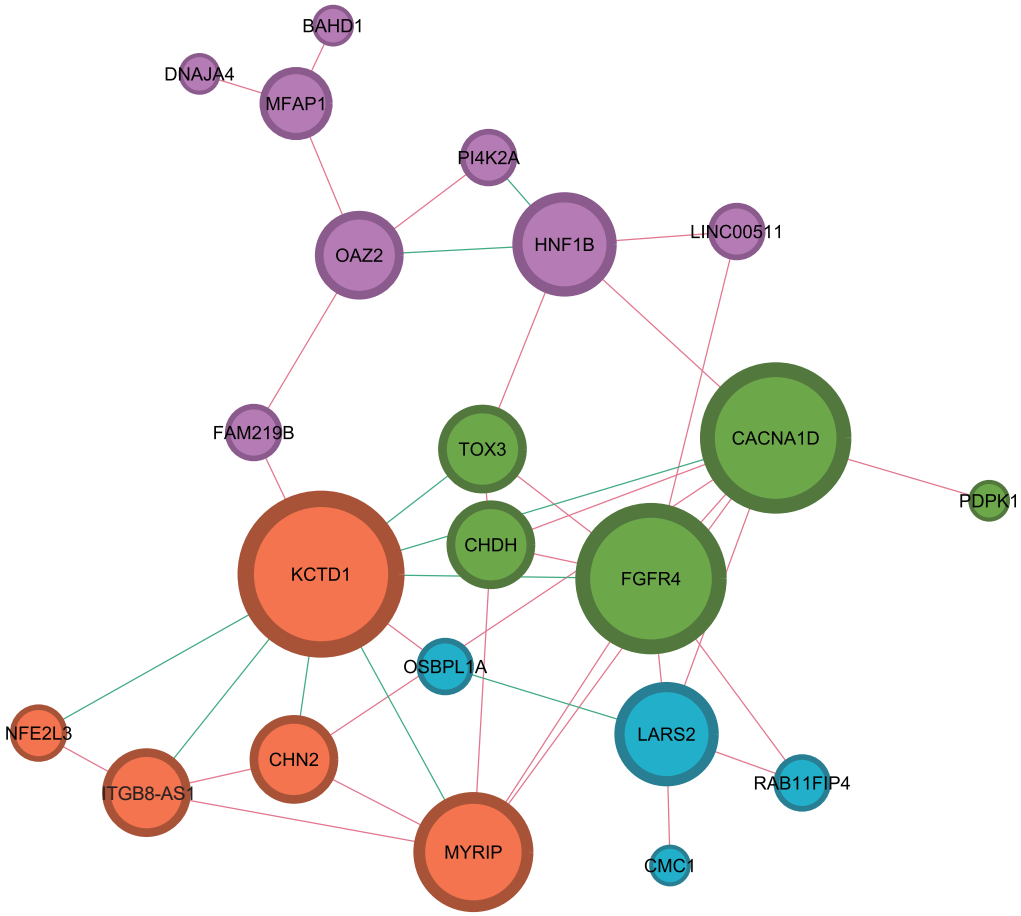

Figure 2. The outcome of the network analysis using the Gephi platform. The network was drawn by selecting all the relationships among the 28 genes identified using Cox regression analysis that satisfied having an absolute value of the Pearson correlation coefficient above 0.3, and a total of 22 genes were drawn. Nodes are divided into 4 classes, represented by 4 colors: purple, green, red, and cyan. The edge color in the network is orange when the correlation coefficient is higher than 0.3 and green when it is lower than −0.3.

2.4. Development of Prognostic Prediction Model and Performance Evaluation Using Pivotal RNA Biomarkers

To develop a prognostic prediction model for two-year survival, three pivotal RNA biomarkers were selected. In brief, MeanDecreaseGini (MDG) was computed for every biomarker in each cohort through a random forest model using the labeled information for patients based on whether they survived for more than two years or not. The MDGs for each cohort were ranked and three biomarkers from the lowest rank sum across the three cohorts were selected (Figure 3A, and details are provided in Section 4). Finally, three biomarkers, namely, ZEB1-AS1, PI4K2A, and ITGB8-AS1 (also known as CTA-293F17.1), were selected. Additionally, MDGs were calculated for three clinical biomarkers (stage, age, and gender). Among these, stage showed the lowest rank sum, while age showed the fifth lowest rank sum. In the end, three RNA biomarkers (ZEB1-AS1, PI4K2A, and ITGB8-AS1) and two clinical biomarkers (stage and age) were selected as pivotal features for the development of the model.

Next, a multiple logistic regression model for predicting two-year survival was developed using the selected five biomarkers and its performance was evaluated through ROC curve analysis (Figure 3B, and details are provided in Section 4). When using a model that included only two clinical biomarkers (stage and age) in the three cohorts, the AUC values were 0.713, 0.826, and 0.687, respectively (Figure 3C). Upon the addition of three

RNA biomarkers (ZEB1-AS1, PI4K2A, and ITGB8-AS1), the AUC values improved to 0.753, 0.847, and 0.725 in the three cohorts, confirming enhancement in predictive performance.

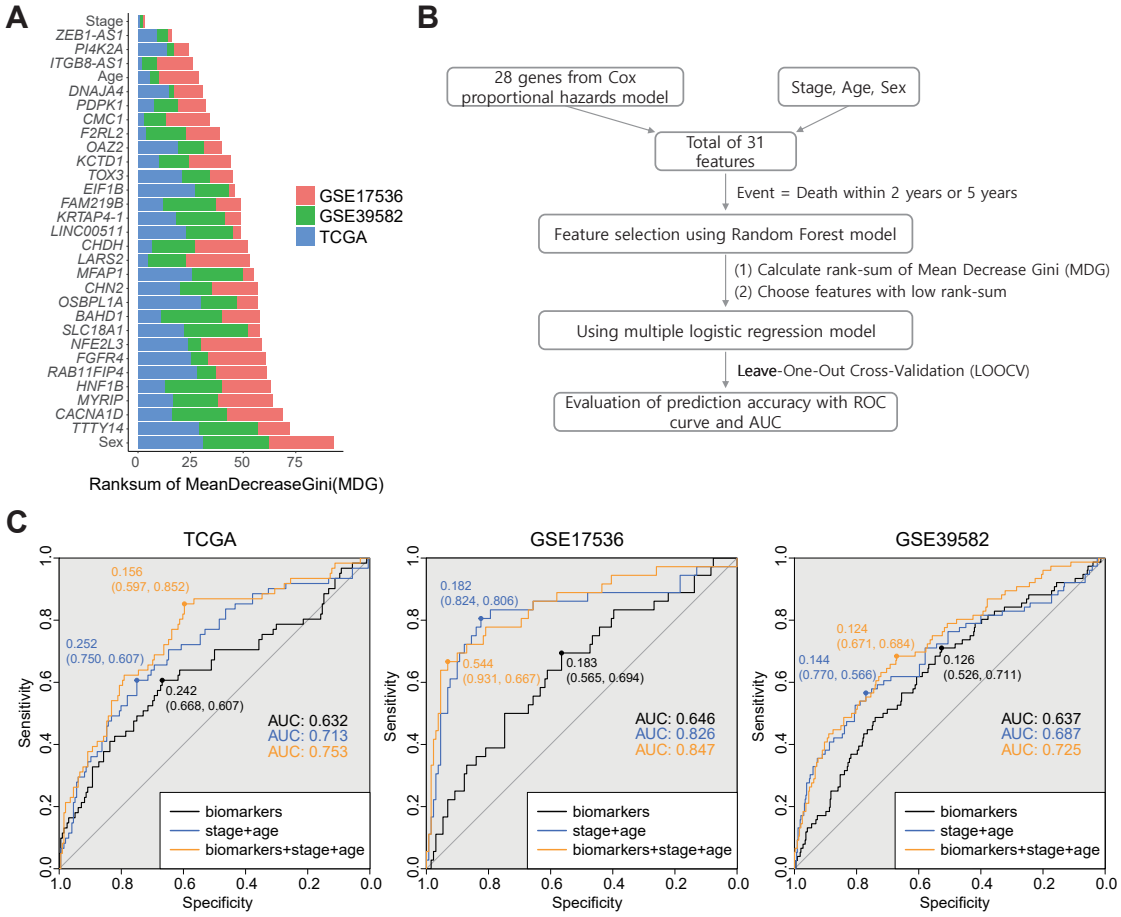

Figure 3. (**A**) The Ranksum of the MeanDecreaseGini (MDG) index. In each cohort, MDG for each feature was calculated using a random forest model and then ranked in descending order. After obtaining the MDG rankings in each cohort, the sum of rankings was calculated for each variable. The graph displays the sum of rankings in ascending order. (**B**) With the five features selected based on MeanDecreaseGini, a multiple logistic regression model was used to predict events such as death within 2 years. Leave-one-out cross-validation (LOOCV) was performed to evaluate prediction accuracy. (**C**) The three ROC curves were plotted based on the following criteria: (1) including only stage and age, (2) including only RNA biomarkers, and (3) including both stage, age, and RNA biomarkers. Area under the curve (AUC) values were calculated for each cohort.

2.5. The Log-Rank Test Results for the Three Biomarkers with the Highest MDG Values and the RNA Expression Heatmaps

The results of the log-rank test performed on the three biomarkers with the highest MDG values showed that PI4K2A satisfied the p-value < 0.05 in two out of three cohorts, while ZEB1-AS1 satisfied the p-value < 0.05 in all three cohorts. For ITGB8-AS1, similar results to those of the Cox regression analysis were observed in two out of three cohorts (Figure 4A–C). The different results of the log-rank test and Cox regression analysis are discussed in the Discussion section. Next, the results of the Wilcoxon rank-sum test and RNA expression heatmaps between individuals who survived more than 2 years and those who did not showed that 12 out of 28 genes had p-values below 0.05 in the TCGA cohort, while 4 out of 28 genes in GSE17536, and 17 out of 28 genes in GSE39582, had p-values less than 0.05 (Figure 4D). ZEB1-AS1 had Wilcoxon rank-sum test p-values less than 0.05 in three cohorts, PI4L2A in one of three cohorts, and ITGB8-AS1 in two of three cohorts. The

discrepancy between the Wilcoxon rank-sum test results and the Cox regression analysis results will also be discussed in Section 3.

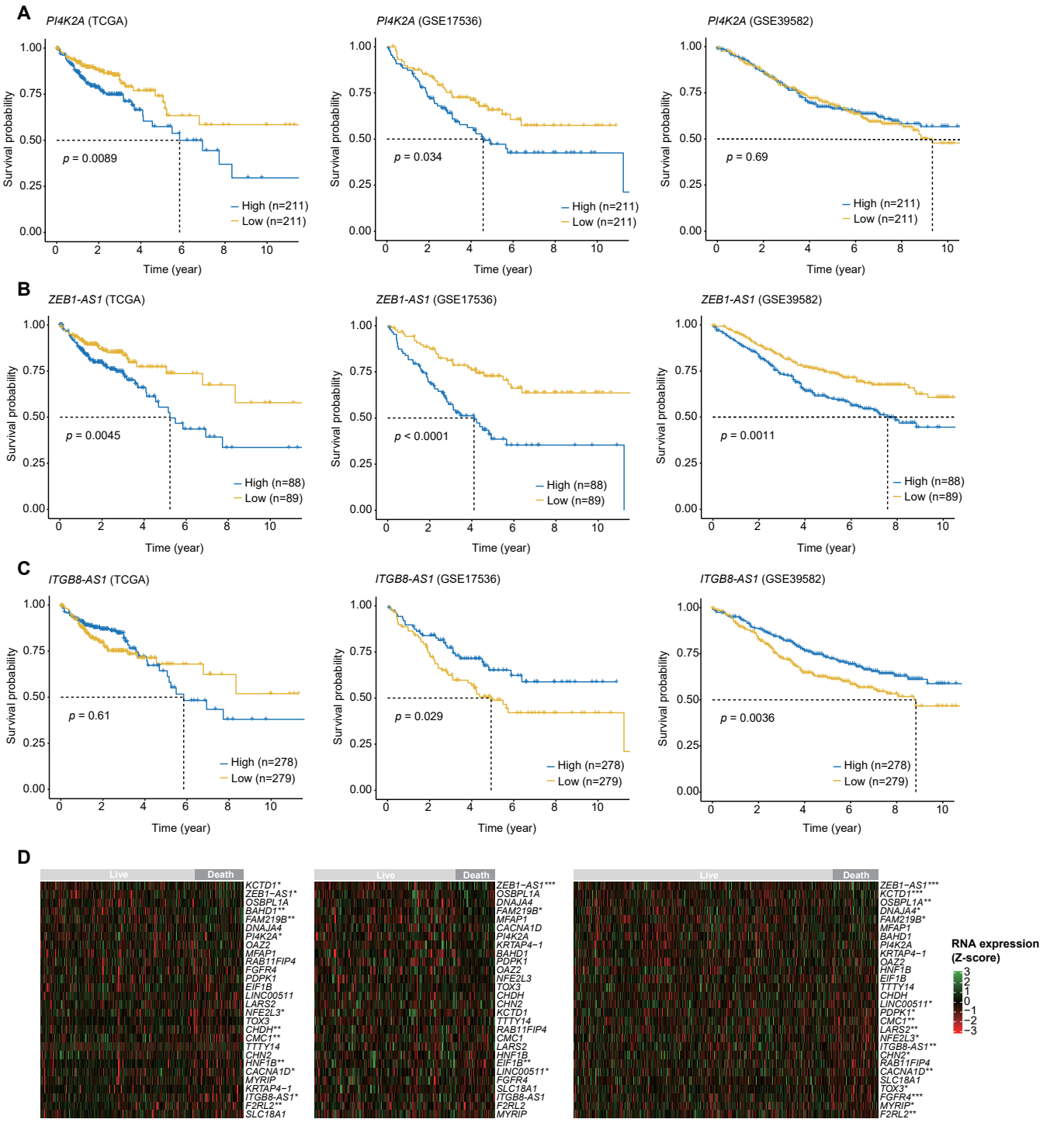

Figure 4. Kaplan–Meier plots of three genes and heatmaps of RNA expression levels in three cohorts. (**A**–**C**) Kaplan–Meier plot of three RNA biomarkers used in a prediction model. The high-expression and low-expression groups were distinguished using the median value of RNA expressions. (**D**) A heatmap of RNA expression levels. The RNA expression of 28 potential biomarkers selected from Cox regression analysis was transformed into Z-scores. 'Death within 2 years' was considered as the event. To compare RNA expression between the live group and the death group, a Wilcoxon rank-sum test was performed (***: *p*-value < 0.001, **: *p*-value < 0.01, *: *p*-value < 0.05). Left: TCGA, Middle: GSE17536, Right: GSE39582.

3. Discussion

This study proposes a method for identifying prognostic RNA biomarkers while considering clinical parameters in a multi-cohort study. In total, 28 RNA biomarkers were identified, including 9 known biomarkers previously reported as prognostic factors for colorectal cancer, and 19 novel biomarkers that are, to the best of our knowledge, newly discovered (Supplementary Table S1). While some RNA biomarkers have shown associations with the carcinogenesis or metastasis of colorectal cancer, their role as prognostic biomarkers for predicting survival has not been proven. Considering consistency with previous research, the use of the proposed method with multiple well-organized cancer cohorts could enhance statistical power for discovering potential prognostic biomarkers in future studies.

The data used to measure the AUC are limited as they are categorized based on whether patients survived more than two years. Therefore, they do not provide a detailed reflection of how long each individual patient survived or the exact time of their death. Nevertheless, each biomarker underwent rigorous statistical validation in all three colorectal cancer cohorts, demonstrating independent associations with prognosis, even after accounting for stage and age. While the effect sizes of these markers may be somewhat smaller compared to well-known prognostic factors such as stage and age, which have remarkably significant effect sizes, it is still considered that these markers hold statistical significance in contributing to prognosis. In addition, while there have been recent attempts in this direction, we would assert that our work is distinct in terms of identifying prognostic biomarkers independent of stage and age [29]. The association between 28 RNA biomarkers and prognosis has been statistically evaluated using public cohorts. However, a prospective study utilizing large-scale data is needed to validate the survival effects of these biomarkers. Furthermore, there is a need to identify factors that may influence biomarkers or pathways, such as epigenetic factors or non-coding regions, in addition to RNA expression.

Conventional methods for discovering biomarkers include the log-rank test for survival analysis and statistical methods for identifying differentially expressed genes (DEGs) such as the Wilcoxon rank-sum test, DESeq2, and Limma [30–32]. To compare the proposed method with conventional methods, KM plots and heatmaps were generated using the three RNA biomarkers that play a pivotal role in the study. Although all three biomarkers are statistically significant in predicting prognosis with Cox regression analysis, ITGB8-AS1 and PI4K2A did not show statistical significance in one cohort each (Figure 4A–C). Cox regression analysis can be considered to have increased statistical validity by accounting for the effects of clinical information, including age, stage, and sex. On the other hand, the log-rank test has limitations for such analysis. First, without a clear biological cutoff of expression values, the challenge arises from dividing these values into two groups. Second, the log-rank test does not consider effects of other clinicopathological factors such as stage or age. In terms of the heatmap, whether the 28 potential RNA biomarkers were differentially expressed in two groups based on 2-year survival was evaluated. ZEB1-AS1 showed statistical significance in all three cohorts, whereas ITGB8-AS1 showed significance in two cohorts, and PI4K2A showed statistical significance in only one cohort, representing discrepant results compared to the results of Cox regression analysis. Like survival analysis, DEG analysis has limitations in developing prognostic biomarkers. Since case and control are selected based on survival over a certain period of time, it is difficult to perform a full evaluation that reflects the survival of individual patients. Additionally, the method for identifying DEGs does not adjust the effects of various clinical factors (age, gender, and stage). Overall, it is thought that Cox regression analysis is a favorable methodology for biomarker discovery that overcomes many of these limitations.

The EGFR downstream signaling involves the RAS/RAF/MEK/ERK, PI3K/Akt, JAK/STAT, and PLC/PKC pathways. Through these diverse downstream signaling pathways, EGFR plays an important role in the carcinogenesis of colorectal cancers [33,34]. EGFR mutation is not common in colorectal cancer patients, whereas upregulation of EGFR is common in this disease [35]. The overexpression of EGFR in colorectal cancer is associ-

ated with an advanced stage of colorectal cancer [36]. Some researchers highlighted the association between high EGFR expression level and TNM stage, especially stage T3 [37]. In mouse experiments, cells with high EGFR expression showed a higher incidence of liver metastasis [38]. However, the role of EGFR expression as a prognostic factor remains controversial [37]. EGFR also plays a pivotal role in colorectal cancer treatment. The first targeted therapeutic agent approved by the Food and Drug Administration (FDA) for colorectal cancer was cetuximab, a monoclonal antibody designed to target EGFR [39]. Furthermore, FDA approved panitumumab, another EGFR-targeting monoclonal antibody in 2006 [40]. Many studies have been conducted to identify biomarkers for selecting favorable patients for EGFR-targeted therapy, including upstream molecules, EGFR amplification, molecules involved in downstream signaling pathways, miRNAs, and methylations. However, further investigation and evaluation are still needed for clinical use of these biomarkers [35].

IGF signaling is considered to be an important factor for pathogenesis of tumors, including CRC [41]. Numerous studies revealed the correlation between IGF2 signaling and CRC [41]. IGF2 signaling accelerates cell growth and survival by activating both IGF1R and IR-A signaling [42]. In the autocrine/paracrine signaling loops of cancer cells, in particular, IGF2 working through IGF-1R and/or IR-A is frequently observed [42]. In vitro studies showed the increase in IGF2 production in diverse colon cancer cell lines [43]. In these cell lines, IGF2 overexpression was one of the key signals for the cancer cell to maintain the tumorigenic features including proliferation and differentiation [43,44]. Studies based on the publicly available data sets, including TCGA, showed that copy number changes of IGF2 and ERBB2 were observed, as well as the association between IGF2 and IGF1R with the stronger beta-catenin/TCF responsive promoter activation [41,45]. IGF2 is also one of the probable therapeutic targets of CRC, along with IGFR, ERBB2, and ERBB3 [46]. The xenografted mice were treated with IGF2R/CI-M6PR, an inhibitor of IGF2, and showed a decrease in Igf2-dependent adenoma phenotype [41,47]. In addition, commercial tissue microarray and univariate survival analysis were performed with paraffin-embedded CRC samples, and, as a result, IGF-2 expression was significantly related to a worse prognosis [48]. Moreover, since the risk of CRC development is increased with obesity and insulin resistance, the development of therapeutic technologies that target IGF signals and related proteins is warranted via further studies and clinical trials [49].

4. Materials and Methods

4.1. Data Acquisition and Preprocessing of Gene Expression Data

To obtain RNA-seq data related to colorectal cancer disease, we obtained gene expression RNA-seq data of GDC TCGA-COAD Colon Cancer (TCGA-COAD) together with clinical information (phenotype and survival data) from UCSC XENA (https://xena.ucsc.edu/, accessed on 6 January 2024) to form one cohort (n = 422) (Figure 1). For microarray data, we used two cohorts of NCBI GEO's colon cancer from more than 100 patients, including clinical information such as age, stage, sex, and overall survival (OS) data, GSE17536, and GSE39582. We used 177 samples from GSE17536 and 557 samples from GSE39582, excluding those without age, sex, stage, or survival information (Figure 1). The microarray platforms of GSE17536 and GSE39582 were both GPL570 (Affymetrix Human Genome U133_Plus_2). Expression levels of genes were represented according to gene symbols for TCGA-COAD and probe names for GSE17536 and GSE39582 (Figure 1). Probe names were annotated with Entrez Gene IDs (Entrez IDs) and gene symbols. Probes linked to multiple gene symbols were excluded from the analysis. In cases where a single probe matched to both mRNA and miRNA, mRNA was only included in the analysis. Then, to match the gene names between cohorts, the associations between TCGA-COAD gene symbols and Ensembl gene IDs from the European Bioinformatics Institute (EBI), and Entrez IDs, were used (Figure 1).

4.2. Selecting a List of Genes Satisfying the Cox Proportional Hazard Assumption

Cox regression analysis was used to identify the correlation between gene expression and overall survival period. Before identifying gene expression that affects prognosis, we checked whether the proportional hazard assumption was satisfied (Figure 1). For genes that satisfied the proportional hazard assumption in a multivariate model using gene expression, age, sex, and stage, we checked the statistical significance of their hazard ratio (p-value less than 0.05 in all three cohorts). After assessing statistical significance, we selected a list of genes with consistent directionality of the hazard ratio (HR) in the three cohorts (HR is less than 1 or greater than 1 in all three cohorts) (Figure 1).

4.3. Network Analysis of Genes with Statistical Significance for Survival

We used Gephi software platform for network analysis [50]. Gephi 0.10.1 (https://gephi.org/, accessed on 6 January 2024) facilitates the analysis of interactivity interpretation of networks, leading to the identification of hub genes. Layout options include ForceAtlas 2 and overview statistics parameters include Approximate Repulsion, Dissuade Hubs, LinLog mode, Prevent Overlap, Average Degree, Average Weighted Degree, and Modularity. In the Gephi analysis, nodes were assigned gene names identified as being statistically significant using Cox regression analysis. Subsequently, we computed the average gene expression correlation R (Pearson correlation) values between genes across the three cohorts, and considered an edge to be present if the absolute value of R exceeded 0.3 for a given gene combination.

4.4. Feature Selection for Survival Prediction

The random forest machine learning method was employed to select features from gene expression and clinical information, including stage, age, and sex, that significantly affect the survival of patients. After events were defined as death within 2 years, survival prediction models were generated. To select features for survival models, MeanDecreaseGini (MDG) scores were calculated using the randomForest R package (https://cran.r-project.org/web/packages/randomForest/, accessed on 6 January 2024). Then, for each feature, the MDGs in each cohort were ranked and the features were prioritized based on the smallest sum of the ranks.

4.5. Evaluation of Predicting Accuracy

In addition to selecting features for survival prediction, receiver operating characteristic (ROC) curve analysis was conducted to assess whether the selected biomarkers could enhance survival prediction. The multiple logistic regression method was used as the prediction model. Using the model, the leave-one-out cross-validation (LOOCV) method was used. Three ROC curves were plotted using the following features: (1) stage and age; (2) multiple biomarkers (gene expression levels); and (3) stage, age, and multiple biomarkers. The area under the curve (AUC) was calculated for each curve.

4.6. Pathway Analysis

Pathway analysis was performed using CancerCompass web-based tool (https://cancercompass.newgenes.org/, accessed on 6 January 2024) to assess the enrichment of 28 potential RNA biomarkers in cancer-related pathways. Cancer-related genes were collected from multiple cohorts, and consensus cancer genes were defined as those present in more than two databases. Cancer-related pathways were identified using a hypergeometric test. In the process of pathway analysis, a cutoff of false discovery rate control was set to 0.01, and after removing duplicate pathways among the top 10 pathways satisfying p-value < 0.01, cancer-relevant pathways were selected. A Sankey plot was provided by CancerCompass and a waterfall plot was generated using R software 4.3.0.

4.7. Statistical Analysis and Visualization

R software (version 4.3.0, R Foundation for Statistical Computing, Vienna, Austria) was used for the study. The Cox regression analysis was performed using the survival packages. The Kaplan–Meier plot and log-rank test were performed using the survminer (https://cran.r-project.org/web/packages/survminer/index.html, accessed on 6 January 2024) package. The randomForest R package (https://cran.r-project.org/web/packages/randomForest/, accessed on 6 January 2024) was used for the random forest machine learning analysis. The stats R package was employed for fitting the logistic regression model (https://www.R-project.org, accessed on 6 January 2024). The porch R package was utilized for generating the ROC curve and calculating AUC values [51]. Finally, the ComplexHeatmap 2.17.0 package was used for the heatmap analysis.

5. Conclusions

Through Cox regression analysis, which considers multiple variables across multiple cohorts, both novel and known prognostic biomarkers were identified. The results of the study will contribute to precision medicine research in determining patient prognosis in colorectal cancer in the future.

Supplementary Materials: The supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijms25063317/s1.

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

Funding: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT, No. 2022R1C1C1012986), by a VHS Medical Center Research Grant, Republic of Korea (VHSMC23032), and by the 2023 Yeungnam University Research Grant.

Institutional Review Board Statement: The Institutional Review Board of the Veterans Health Service Medical Center (Seoul, Republic of Korea) approved this study (IRB no. BOHUN 2023-01-018).

Informed Consent Statement: Not applicable.

Data Availability Statement: The RNA expression data and clinical data are available in the GEO under the accession number GSE17536 and GSE39582 and in the UCSC XENA website (https://xena.ucsc.edu/, accessed on 6 January 2024).

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

References

1. Bray, F.; Ferlay, J.; Soerjomataram, I.; Siegel, R.L.; Torre, L.A.; Jemal, A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **2018**, *68*, 394–424. [CrossRef]
2. Marley, A.R.; Nan, H. Epidemiology of colorectal cancer. *Int. J. Mol. Epidemiol. Genet.* **2016**, *7*, 105–114. [PubMed]
3. Siegel, R.L.; Miller, K.D.; Sauer, A.G.; Fedewa, S.A.; Butterly, L.F.; Anderson, J.C.; Cercek, A.; Smith, R.A.; Jemal, A. Colorectal cancer statistics, 2020. *CA Cancer J. Clin.* **2020**, *70*, 145–164. [CrossRef] [PubMed]
4. Binefa, G.; Rodriguez-Moranta, F.; Teule, A.; Medina-Hayas, M. Colorectal cancer: From prevention to personalized medicine. *World J. Gastroenterol.* **2014**, *20*, 6786–6808. [CrossRef] [PubMed]
5. Burrell, R.A.; McGranahan, N.; Bartek, J.; Swanton, C. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* **2013**, *501*, 338–345. [CrossRef] [PubMed]
6. Deng, X.; Nakamura, Y. Cancer precision medicine: From cancer screening to drug selection and personalized immunotherapy. *Trends Pharmacol. Sci.* **2017**, *38*, 15–24. [CrossRef] [PubMed]
7. Salgado, R.; on behalf of the IBCD-Faculty; Moore, H.; Martens, J.W.M.; Lively, T.; Malik, S.; McDermott, U.; Michiels, S.; Moscow, J.A.; Tejpar, S.; et al. Steps forward for cancer precision medicine. *Nat. Rev. Drug Discov.* **2018**, *17*, 1–2. [CrossRef] [PubMed]

8. Amin, M.B.; Greene, F.L.; Edge, S.B.; Compton, C.C.; Gershenwald, J.E.; Brookland, R.K.; Meyer, L.; Gress, D.M.; Byrd, D.R.; Winchester, D.P. The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. *CA Cancer J. Clin.* **2017**, *67*, 93–99. [CrossRef]
9. Parmigiani, G.; Garrett, E.S.; Irizarry, R.A.; Zeger, S.L. The Analysis of Gene Expression Data: An Overview of Methods and Software. In *The Analysis of Gene Expression Data: Methods and Software*; Parmigiani, G., Garrett, E.S., Irizarry, R.A., Zeger, S.L., Eds.; Springer New York: New York, NY, USA, 2003; pp. 1–45.
10. Kathleen Kerr, M.; AChurchill, G. Statistical design and the analysis of gene expression microarray data. *Genet. Res.* **2001**, *77*, 123–128. [CrossRef]
11. Smyth, G.K. limma: Linear Models for Microarray Data. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*; Gentleman, R., Carey, V.J., Huber, W., Irizarry, R.A., Dudoit, S., Eds.; Springer New York: New York, NY, USA, 2005; pp. 397–420.
12. Zahurak, M.; Parmigiani, G.; Yu, W.; Scharpf, R.B.; Berman, D.; Schaeffer, E.; Shabbeer, S.; Cope, L. Pre-processing Agilent microarray data. *BMC Bioinform.* **2007**, *8*, 142. [CrossRef]
13. Archer, K.J.; Reese, S.E. Detection call algorithms for high-throughput gene expression microarray data. *Brief. Bioinform.* **2009**, *11*, 244–252. [CrossRef]
14. Schurmann, C.; Heim, K.; Schillert, A.; Blankenberg, S.; Carstensen, M.; Dörr, M.; Endlich, K.; Felix, S.B.; Gieger, C.; Grallert, H.; et al. Analyzing Illumina Gene Expression Microarray Data from Different Tissues: Methodological Aspects of Data Analysis in the MetaXpress Consortium. *PLoS ONE* **2012**, *7*, e50938. [CrossRef] [PubMed]
15. Gohlmann, H.; Talloen, W. *Gene Expression Studies Using Affymetrix Microarrays*; CRC Press: Boca Raton, FL, USA, 2009.
16. Jiang, N.; Leach, L.J.; Hu, X.; Potokina, E.; Jia, T.; Druka, A.; Waugh, R.; Kearsey, M.J.; Luo, Z.W. Methods for evaluating gene expression from Affymetrix microarray datasets. *BMC Bioinform.* **2008**, *9*, 284. [CrossRef] [PubMed]
17. Wang, Z.; Gerstein, M.; Snyder, M. RNA-Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **2009**, *10*, 57–63. [CrossRef]
18. Wilhelm, B.T.; Landry, J.R. RNA-Seq-quantitative measurement of expression through massively parallel *RNA-sequencing*. *Methods* **2009**, *48*, 249–257. [CrossRef]
19. Barnes, M.; Freudenberg, J.; Thompson, S.; Aronow, B.; Pavlidis, P. Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms. *Nucleic Acids Res.* **2005**, *33*, 5914–5923. [CrossRef]
20. Narrandes, S.; Xu, W. Gene Expression Detection Assay for Cancer Clinical Use. *J. Cancer* **2018**, *9*, 2249–2265. [CrossRef] [PubMed]
21. Barrett, T.; Wilhite, S.E.; Ledoux, P.; Evangelista, C.; Kim, I.F.; Tomashevsky, M.; Marshall, K.A.; Phillippy, K.H.; Sherman, P.M.; Holko, M.; et al. NCBI GEO: Archive for functional genomics data sets—Update. *Nucleic Acids Res.* **2013**, *41*, D991–D995. [CrossRef]
22. Kapushesky, M.; Emam, I.; Holloway, E.; Kurnosov, P.; Zorin, A.; Malone, J.; Rustici, G.; Williams, E.; Parkinson, H.; Brazma, A. Gene expression atlas at the European bioinformatics institute. *Nucleic Acids Res.* **2010**, *38*, D690–D698. [CrossRef]
23. Athar, A.; Füllgrabe, A.; George, N.; Iqbal, H.; Huerta, L.; Ali, A.; Snow, C.; A Fonseca, N.; Petryszak, R.; Papatheodorou, I.; et al. ArrayExpress update—From bulk to single-cell expression data. *Nucleic Acids Res.* **2019**, *47*, D711–D715. [CrossRef]
24. Morera, D.S.; Hasanali, S.L.; Belew, D.; Ghosh, S.; Klaassen, Z.; Jordan, A.R.; Wang, J.; Terris, M.K.; Bollag, R.J.; Merseburger, A.S.; et al. Clinical Parameters Outperform Molecular Subtypes for Predicting Outcome in Bladder Cancer: Results from Multiple Cohorts, Including TCGA. *J. Urol.* **2020**, *203*, 62–72. [CrossRef]
25. Ren, Q.; Zhang, P.; Zhang, X.; Feng, Y.; Li, L.; Lin, H.; Yu, Y. A fibroblast-associated signature predicts prognosis and immunotherapy in esophageal squamous cell cancer. *Front. Immunol.* **2023**, *14*, 1199040. [CrossRef]
26. Du, J.; Yang, M.; Chen, S.; Li, D.; Chang, Z.; Dong, Z. PDK1 promotes tumor growth and metastasis in a spontaneous breast cancer model. *Oncogene* **2016**, *35*, 3314–3323. [CrossRef]
27. Li, D.; Mullinax, J.E.; Aiken, T.; Xin, H.; Wiegand, G.; Anderson, A.; Thorgeirsson, S.; Avital, I.; Rudloff, U. Loss of PDPK1 abrogates resistance to gemcitabine in label-retaining pancreatic cancer cells. *BMC Cancer* **2018**, *18*, 772. [CrossRef]
28. Levine, K.M.; Ding, K.; Chen, L.; Oesterreich, S. FGFR4: A promising therapeutic target for breast cancer and other solid tumors. *Pharmacol. Ther.* **2020**, *214*, 107590. [CrossRef]
29. Park, S.W.; Kang, J.; Kim, H.S.; Yoon, S.; Kim, B.S.; Lim, C.; Lee, D.; Kim, Y.H. Predicting prognosis through the discovery of specific biomarkers according to colorectal cancer lymph node metastasis. *Am. J. Cancer Res.* **2023**, *13*, 3221–3233.
30. García-Alfonso, P.; García-González, G.; Gallego, I.; Peligros, M.I.; Ortega, L.; Pérez-Solero, G.T.; Sandoval, C.; Martin, A.M.; Codesido, M.B.; Ferrándiz, A.C.; et al. Prognostic value of molecular biomarkers in patients with metastatic colorectal cancer: A real-world study. *Clin. Transl. Oncol.* **2021**, *23*, 122–129. [CrossRef]
31. Shahjaman, M.; Rahman, M.R.; Islam, S.M.S.; Mollah, M.N.H. A Robust Approach for Identification of Cancer Biomarkers and Candidate Drugs. *Medicina* **2019**, *55*, 269. [CrossRef] [PubMed]
32. Wang, H.; Han, X.; Gao, S. Identification of potential biomarkers for pathogenesis of Alzheimer's disease. *Hereditas* **2021**, *158*, 23. [CrossRef] [PubMed]
33. Lo, H.W.; Hung, M.C. Nuclear EGFR signalling network in cancers: Linking EGFR pathway to cell cycle progression, nitric oxide pathway and patient survival. *Br. J. Cancer* **2006**, *94*, 184–188. [CrossRef] [PubMed]
34. Wee, P.; Wang, Z. Epidermal Growth Factor Receptor Cell Proliferation Signaling Pathways. *Cancers* **2017**, *9*, 52. [CrossRef] [PubMed]

35. Yang, J.; Li, S.; Wang, B.; Wu, Y.; Chen, Z.; Lv, M.; Lin, Y.; Yang, J. Potential biomarkers for anti-EGFR therapy in metastatic colorectal cancer. *Tumour Biol.* **2016**, *37*, 11645–11655. [CrossRef] [PubMed]
36. Gross, M.E.; Zorbas, M.A.; Danels, Y.J.; Garcia, R.; Gallick, G.E.; Olive, M.; Brattain, M.G.; Boman, B.M.; Yeoman, L.C. Cellular growth response to epidermal growth factor in colon carcinoma cells with an amplified epidermal growth factor receptor derived from a familial adenomatous polyposis patient. *Cancer Res.* **1991**, *51*, 1452–1459. [PubMed]
37. Spano, J.-P.; Lagorce, C.; Atlan, D.; Milano, G.; Domont, J.; Benamouzig, R.; Attar, A.; Benichou, J.; Martin, A.; Morere, J.-F.; et al. Impact of EGFR expression on colorectal cancer patient prognosis and survival. *Ann. Oncol.* **2005**, *16*, 102–108. [CrossRef]
38. Radinsky, R.; Risin, S.; Fan, D.; Dong, Z.; Bielenberg, D.; Bucana, C.D.; Fidler, I.J. Level and function of epidermal growth factor receptor predict the metastatic potential of human colon carcinoma cells. *Clin. Cancer Res.* **1995**, *1*, 19–31.
39. Jonker, D.J.; O'Callaghan, C.J.; Karapetis, C.S.; Zalcberg, J.R.; Tu, D.; Au, H.-J.; Berry, S.R.; Krahn, M.; Price, T.; Simes, R.J.; et al. Cetuximab for the Treatment of Colorectal Cancer. *N. Engl. J. Med.* **2007**, *357*, 2040–2048. [CrossRef]
40. Gibson, T.B.; Ranganathan, A.; Grothey, A. Randomized phase III trial results of panitumumab, a fully human anti-epidermal growth factor receptor monoclonal antibody, in metastatic colorectal cancer. *Clin. Color. Cancer* **2006**, *6*, 29–31. [CrossRef]
41. Kasprzak, A.; Adamek, A. Insulin-Like Growth Factor 2 (IGF2) Signaling in Colorectal Cancer-From Basic Research to Potential Clinical Applications. *Int. J. Mol. Sci.* **2019**, *20*, 4915. [CrossRef]
42. Blyth, A.J.; Kirk, N.S.; Forbes, B.E. Understanding IGF-II Action through Insights into Receptor Binding and Activation. *Cells* **2020**, *9*, 2276. [CrossRef]
43. Lamonerie, T.; Lavialle, C.; de Galle, B.; Binoux, M.; Brison, O. Constitutive or Inducible Overexpression of the IGF-2 Gene in Cells of a Human Colon Carcinoma Cell Line. *Exp. Cell Res.* **1995**, *216*, 342–351. [CrossRef]
44. Zarrilli, R.; Pignata, S.; Romano, M.; Gravina, A.; Casola, S.; Bruni, C.B.; Acquaviva, A.M. Expression of insulin-like growth factor (IGF)-II and IGF-I receptor during proliferation and differentiation of CaCo-2 human colon carcinoma cells. *Cell Growth Differ.* **1994**, *5*, 1085–1091.
45. Lee, H.; Kim, N.; Yoo, Y.J.; Kim, H.; Jeong, E.; Choi, S.; Moon, S.U.; Oh, S.H.; Mills, G.B.; Yoon, S.; et al. β-catenin/TCF activity regulates IGF-1R tyrosine kinase inhibitor sensitivity in colon cancer. *Oncogene* **2018**, *37*, 5466–5475. [CrossRef]
46. The Cancer Genome Atlas (TCGA) Research Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **2012**, *487*, 330–337. [CrossRef]
47. Harper, J.; Burns, J.L.; Foulstone, E.J.; Pignatelli, M.; Zaina, S.; Hassan, A.B. Soluble IGF2 Receptor Rescues $Apc^{Min/+}$ Intestinal Adenoma Progression Induced by Igf2 Loss of Imprinting. *Cancer Res.* **2006**, *66*, 1940–1948. [CrossRef]
48. Peters, G.; Gongoll, S.; Langner, C.; Mengel, M.; Piso, P.; Klempnauer, J.; Rüschoff, J.; Kreipe, H.; von Wasielewski, R. IGF-1R, IGF-1 and IGF-2 expression as potential prognostic and predictive markers in colorectal-cancer. *Virchows Arch.* **2003**, *443*, 139–145. [CrossRef] [PubMed]
49. Vigneri, P.G.; Tirro, E.; Pennisi, M.S.; Massimino, M.; Stella, S.; Romano, C.; Manzella, L. The Insulin/IGF System in Colorectal Cancer Development and Resistance to Therapy. *Front. Oncol.* **2015**, *5*, 230. [CrossRef] [PubMed]
50. Bastian, M.; Heymann, S.; Jacomy, M. (Eds.) Gephi: An open source software for exploring and manipulating networks. In Proceedings of the International AAAI Conference on Web and Social Media, San Jose, CA, USA, 17–20 May 2009. [CrossRef]
51. Robin, X.; Turck, N.; Hainard, A.; Tiberti, N.; Lisacek, F.; Sanchez, J.-C.; Müller, M. pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinform.* **2011**, *12*, 77. [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.