International Journal of
Molecular Sciences

MDPI

Article

New Transcriptomic Biomarkers of 5-Fluorouracil Resistance

János Tibor Fekete [1,2] and Balázs Győrffy [3,4,*]

[1] Research Center for Natural Sciences, Cancer Biomarker Research Group, Institute of Enzymology, Magyar Tudósok Krt. 2, H-1117 Budapest, Hungary
[2] National Laboratory for Drug Research and Development, Magyar Tudósok Krt. 2, H-1117 Budapest, Hungary
[3] Department of Pediatrics, Semmelweis University, H-1094 Budapest, Hungary
[4] Department of Bioinformatics, Semmelweis University, H-1094 Budapest, Hungary
* Correspondence: gyorffy.balazs@med.semmelweis-univ.hu; Tel.: +36-30-514-2822

Abstract: The overall response rate to fluoropyrimidine monotherapy in colorectal cancer (CRC) is limited. Transcriptomic datasets of CRC patients treated with 5-fluorouracil (5FU) could assist in the identification of clinically useful biomarkers. In this research, we aimed to analyze transcriptomic cohorts of 5FU-treated cell lines to uncover new predictive biomarker candidates and to validate the strongest hits in 5FU-treated human colorectal cancer samples with available clinical response data. We utilized an in vitro dataset of cancer cell lines treated with 5FU and used the reported area under the dose–response curve values to determine the therapeutic response to 5FU treatment. Mann–Whitney and ROC analyses were performed to identify significant genes. The strongest genes were combined into a single signature using a random forest classifier. The compound 5-fluorouracil was tested in 592 cell lines (294 nonresponders and 298 responders). The validation cohort consisted of 157 patient samples with 5FU monotherapy from three datasets. The three strongest associations with treatment outcome were observed in *SHISA4* (AUC = 0.745, p-value = 5.5×10^{-25}), *SLC38A6* (AUC = 0.725, p-value = 3.1×10^{-21}), and *LAPTM4A* (AUC = 0.723, p-value = 6.4×10^{-21}). A random forest model utilizing the top genes reached an AUC value of 0.74 for predicting therapeutic sensitivity. The model correctly identified 83% of the nonresponder and 73% of the responder patients. The cell line cohort is available and the entire human colorectal cohort have been added to the ROCPlot analysis platform. Here, by using in vitro and in vivo data, we present a framework enabling the ranking of future biomarker candidates of 5FU resistance. A future option is to conduct an independent validation of the established predictors of resistance.

Keywords: pharmacology; proliferation; receiver operator characteristics; RNAseq; gene expression

Citation: Fekete, J.T.; Győrffy, B. New Transcriptomic Biomarkers of 5-Fluorouracil Resistance. *Int. J. Mol. Sci.* **2023**, *24*, 1508. https://doi.org/10.3390/ijms24021508

Academic Editor: Carmine Stolfi

Received: 15 November 2022
Revised: 9 January 2023
Accepted: 10 January 2023
Published: 12 January 2023

1. Introduction

Globally, colorectal cancer (CRC) accounts for 10% (1.9 million) of the overall cancer incidence and 9.4% (935,000) of deaths caused by cancer. In 2020, it placed third in incidence and second in mortality among all cancer types [1]. In terms of pathogenesis, approximately 70% of the diagnosed cases are sporadic, 25% are familial, and the remaining 5% are inherited [2]. The primary treatment option is surgery, but in advanced stages, systemic chemotherapy is the standard treatment. The adjuvant therapy for resected CRC patients without metastases includes capecitabine, oxaliplatin, 5-fluorouracil (5FU), and leucovorin [3]. Treatment for metastatic patients is based on 5FU, oxaliplatin, and irinotecan, and, depending on the genetic profile, may also involve biological therapies such as those targeting the VEGF or EGFR pathways [4].

Predictive biomarkers can be used to predict the outcome of a therapeutic intervention and can provide information on the expected success rate of a particular therapy, the existence of therapeutic resistance, or the development of a serious adverse reaction. Although biomarkers can contribute to personalized treatment with improved outcomes, the number of predictive biomarkers in colorectal cancer is still limited. One of the few biomarkers that influence therapeutic decision-making is microsatellite instability (MSI). The 5FU monotherapy treatment is not effective in patients with high MSI, but oxaliplatin-based treatment can offer a therapeutic benefit. Another biomarker affecting therapeutic decisions is the presence of KRAS/NRAS mutations, which decrease the efficacy of EGFR therapies [5]. An important predictive biomarker for avoiding serious adverse reactions to 5FU-based treatment is the test for dihydropyrimidine dehydrogenase (DPD) deficiency in patients [6].

Both 5FU and capecitabine are fluoropyrimidines, which are members of the class of agents known as antimetabolites. As their chemical structure shares a number of common traits with the substrate of enzymes essential for DNA synthesis, antimetabolites can disrupt the DNA structure and ultimately lead to tumor cell death. Besides colorectal cancer, fluoropyrimidines are also used to treat breast, head and neck, ovarian, and gastrointestinal tumors, as well as basal cell carcinomas. The anti-tumor effects of fluoropyrimidine analogs are complex and involve three different mechanisms of action: (a) the 5FU metabolite fluorodeoxyuridine monophosphate (FdUMP) inhibits thymidylate synthase, which is essential for thymidine synthesis, (b) fluorouridine triphosphate (FUTP) is incorporated into RNA, and (c) fluorodeoxyuridine triphosphate (FdUTP) is incorporated into DNA resulting in DNA strand breaks [7].

The overall response rate to fluoropyrimidine monotherapy in colorectal cancer is limited and is in the range of 10% to 15% [8]. A higher response rate of up to 45–50% can be achieved with additional oxaliplatin or irinotecan treatment; however, in these cases, toxicity will also increase [9]. The incidence of CRC is rising and 5FU is considered a key drug; therefore, there is an urgent need to be able to identify patients who may benefit from 5FU-based therapies. Transcriptomic datasets profiling tumors of CRC patients treated with 5FU could provide help with making a significant advancement in this area.

In this study, we aimed to analyze transcriptomic cohorts of 5FU-treated cell lines to uncover new predictive biomarker candidates and validate the strongest hits in 5FU-treated colorectal human samples with clinical information and response data. Our overall goal was not only to identify single genes that could serve as biomarkers in patients treated with fluoropyrimidine but also to combine the top candidates into a single predictive tool by employing machine learning.

2. Results

2.1. Features Significant in the In Vitro Dataset

The compound 5-fluorouracil was tested in 907 cell lines, with a minimum screening concentration of 0.125 μM and a maximum concentration of 32 μM. Based on the reported AUDRC values, 294 were categorized as nonresponders (AUDRC range: 0.931–0.991), 298 as responders (AUDRC range: 0.099–0.801), and 315 (AUDRC between 0.931 and 0.801) were excluded from the analysis. The most sensitive solid tumor cell lines were found to be PSN1 (pancreas cancer), CAL148 (breast cancer), and JHU011 (head and neck cancer), whereas the most resistant cell lines were LN18 (glioblastoma), HEC1 (uterus cancer), and ASH3 (cervix cancer) (Table 1).

All available genes (n = 15,791) were tested by the Mann–Whitney U test and ROC analysis, and we found statistically significant differences between nonresponders and responders in 2484 genes. The strongest associations with treatment outcome were observed in *SHISA4* (shisa family member 4 gene, n = 592, ROC AUC: 0.745, Mann–Whitney U test *p*-value: 5.5×10^{-25}), *SLC38A6* (solute carrier family 38 member 6 gene, ROC AUC = 0.725, *p*-value = 3.1×10^{-21}), and *LAPTM4A* (lysosomal protein transmembrane 4 alpha gene, ROC AUC = 0.721, *p*-value = 1.2×10^{-20}). The top ten most significant genes are presented in Table 2, and the complete list of all genes is provided in Supplemental Table S1.

Table 1. The top ten 5-fluorouracil sensitive (A) and resistant (B) solid tumor cell lines in the GDSC database. A lower AUDRC (area under the dose–response curve) value indicates higher sensitivity, while a higher AUDRC value indicates greater resistance.

(A)

Cell Line CCLE Name	Tumor Type	Normalized AUDRC
PSN1_PANCREAS	pancreas exocrine adenocarcinoma	0.352
CAL148_BREAST	breast ductal carcinoma	0.353
JHU011_UPPER_AERODIGESTIVE_TRACT	upper aerodigestive squamous carcinoma	0.39
PCI4B_UPPER_AERODIGESTIVE_TRACT	upper aerodigestive squamous carcinoma	0.423
CAL27_UPPER_AERODIGESTIVE_TRACT	upper aerodigestive squamous carcinoma	0.439
PA1_OVARY	ovary mixed germ cell	0.452
MRKNU1_BREAST	breast carcinoma	0.46
22RV1_PROSTATE	prostate adenocarcinoma	0.483
HUPT4_PANCREAS	pancreas exocrine adenocarcinoma	0.488
A2780_OVARY	ovary endometrioid adenocarcinoma	0.496

(B)

Cell Line CCLE Name	Tumor Type	Normalized AUDRC
LN18_CENTRAL_NERVOUS_SYSTEM	glioblastoma	0.991
HEC1_ENDOMETRIUM	uterus endometrial adenocarcinoma	0.99
ASH3_THYROID	thyroid carcinoma	0.988
SCH_STOMACH	choriocarcinoma	0.988
CCFSTTG1_CENTRAL_NERVOUS_SYSTEM	astrocytoma	0.988
NCIH2444_LUNG	non-small cell lung cancer	0.988
HSC2_UPPER_AERODIGESTIVE_TRACT	upper aerodigestive squamous carcinoma	0.988
NCIH524_LUNG	small cell lung cancer	0.988
SW13_ADRENAL_CORTEX	adrenal cortex	0.987
FTC133_THYROID	thyroid carcinoma	0.987

Table 2. The top ten genes associated with response in the fluoropyrimidines treated samples of the in vitro database.

Gene Symbol	Approved Name	ROC AUC	Mean Expression (Nonresponder)	Mean Expression (Responder)	Mean Fold Change	Mann–Whitney U Test *p*-Value
SHISA4	shisa family member 4	0.745	1096.7	714.4	1.54	5.52×10^{-25}
SLC38A6	solute carrier family 38 member 6	0.725	1072.0	735.3	1.46	3.07×10^{-21}
PRPF38B	pre-mRNA processing factor 38B	0.723	753.0	1056.3	0.71	6.39×10^{-21}
LAPTM4A	lysosomal protein transmembrane 4 alpha	0.721	1027.9	712.3	1.44	1.23×10^{-20}
MT2A	metallothionein 2A	0.720	1038.1	699.6	1.48	2.34×10^{-20}
FAM114A1	family with sequence similarity 114 member A1	0.719	1058.4	687.9	1.54	2.73×10^{-20}
FAM127A	retrotransposon Gag-like 8C	0.718	1031.1	678.9	1.52	4.75×10^{-20}
SLC35F5	solute carrier family 35 member F5	0.715	1091.4	732.7	1.49	1.17×10^{-19}
TSPAN4	tetraspanin 4	0.715	1090.0	703.4	1.55	1.25×10^{-19}
VAMP3	vesicle-associated membrane protein 3	0.715	1051.6	719.4	1.46	1.44×10^{-19}

2.2. Clinical Sample Database Construction

Our search criteria were met by a total of 805 CRC patients' samples from 12 datasets with available treatment information including response and raw gene expression data (Figure 1B). The detailed characteristics of all screened datasets are presented in Table 3. After normalization of the raw gene expression data and standardization between the different measurement platforms (Figure 1C), mRNA expression for a total of 19,890 genes was available. Out of 805 patients, we filtered those who had received multiple treatments: 180 patients had received bevacizumab, 221 had received irinotecan, and 438 oxaliplatin. Note that some patients had received multiple combinations of these; thus, the number of samples remaining—of patients who had received monotherapies of either 5FU or capecitabine—was reduced to 157. As our goal was to directly link the in vitro data (which was based on 5FU monotherapy) and the clinical samples, we used only these 157 samples as the validation cohort in the present study. However, the entire cohort is available for further research at the www.rocplot.com/colorectal website.

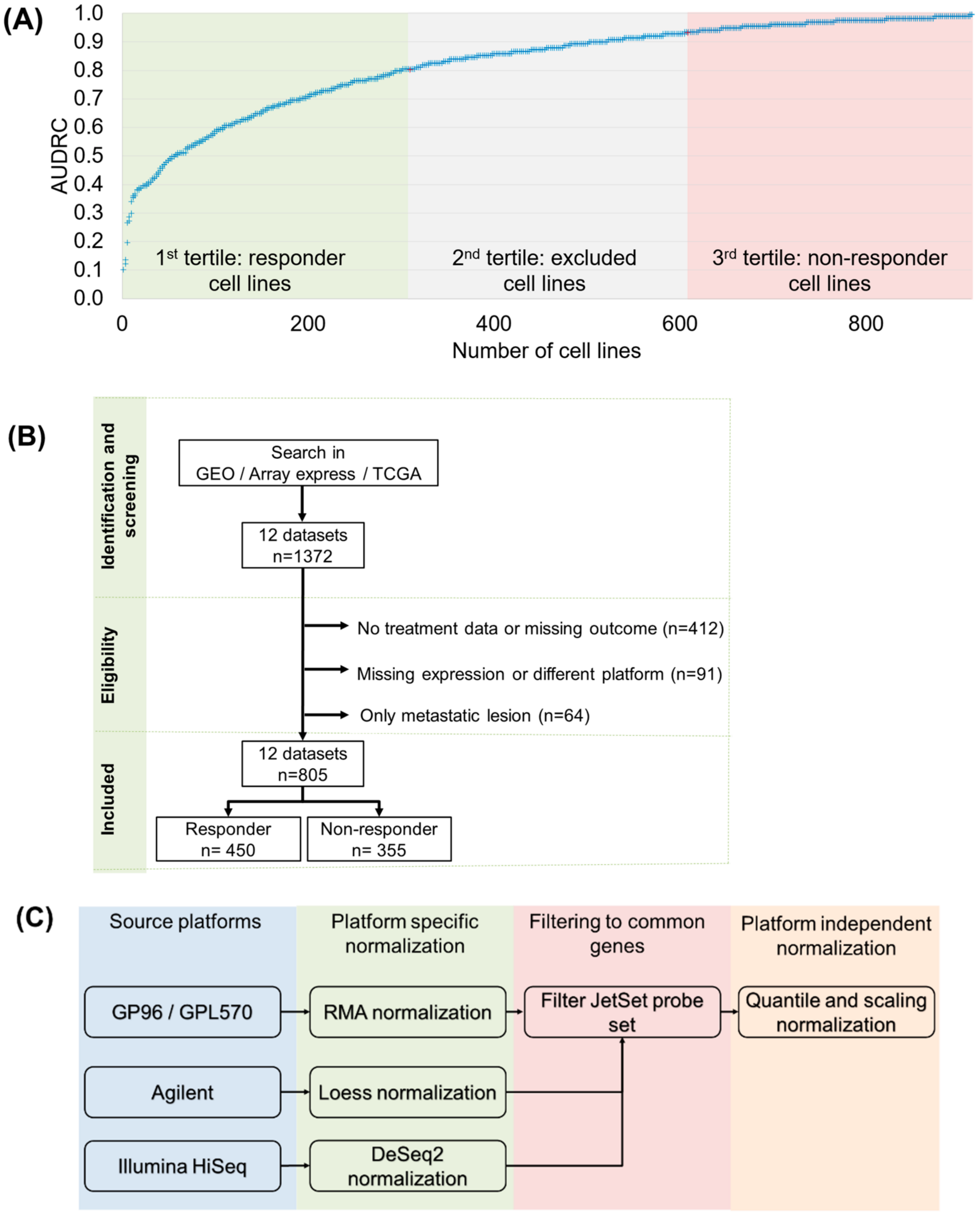
(A)
AUDRC
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1st tertile: responder cell lines
2nd tertile: excluded cell lines
3rd tertile: non-responder cell lines
0
200
400
600
800
Number of cell lines
(B)
Identification and screening
Eligibility
Included
Search in GEO / Array express / TCGA
12 datasets n=1372
No treatment data or missing outcome (n=412)
Missing expression or different platform (n=91)
Only metastatic lesion (n=64)
12 datasets n=805
Responder n= 450
Non-responder n= 355
(C)
Source platforms
Platform specific normalization
Filtering to common genes
Platform independent normalization
GP96 / GPL570
Agilent
Illumina HiSeq
RMA normalization
Loess normalization
DeSeq2 normalization
Filter JetSet probe set
Quantile and scaling normalization

Figure 1. Overview of the data processing steps. Summary of response classification using the area under the dose–response curve (AUDRC) values in the 5FU-treated cell lines (**A**). Outline of the clinical database setup (**B**) and the pipeline employed for transcriptomic data pre-processing (**C**).

Table 3. Overview of datasets screened for the clinical database.

Dataset (Reference)	Platform	Sample Size	Age in Years, (Mean ± SD)	5FU/ Capecitabine	Irinotecan	Bevacizumab	Oxaliplatin	Of These: 5FU/ Capecitabine Monotherapy
GSE19860	Affymetrix HGU133 Plus 2.0 Array	29	-	29	-	12	29	-
GSE19862	Affymetrix HGU133 Plus 2.0 Array	14	-	-	-	14	-	-
GSE28702 [10]	Affymetrix HGU133 Plus 2.0 Array	56	-	56	-	-	56	-
GSE45404 [11]	Affymetrix HGU133 Plus 2.0 Array	42	59.9 ± 11.20	42	-	-	-	42
GSE49355 [12]	Affymetrix HGU133A Array	20	59.3 ± 7.83	20	20	-	-	-
GSE52735 [13]	Affymetrix HGU133 Plus 2.0 Array	37	-	37	-	-	37	-
GSE62080 [14]	Affymetrix HGU133 Plus 2.0 Array	21	-	21	21	-	-	-
GSE69657 [15]	Affymetrix HGU133 Plus 2.0 Array	16	53.6 ± 12.88	16	-	-	16	-
GSE72970 [16]	Affymetrix HGU133 Plus 2.0 Array	124	61.7 ± 11.39	124	88	28	40	-
GSE104645 [17]	Agilent-014850 Whole Human Genome Microarray 4x44K	167	-	140	54	83	113	-
GSE119409 [18]	Affymetrix HGU133 Plus 2.0 Array	56	56.8 ± 11.17	56	-	-	-	56
TCGA	Illumina HiSeq	223	60.7 ± 12.04	210	38	43	147	59
Total		805	60.1 ± 11.69	751	221	180	438	157

2.3. Validation of Significant Genes in the Clinical Sample Database

Of a total of 2484 genes previously identified in the in vitro database, 742 genes reached a statistically significant association in the clinical samples as well. The complete list of all significant genes is presented in Supplemental Table S2. The radar chart of AUC values and gene expression boxplots in resistant and sensitive samples of the strongest genes are presented in Figure 2.

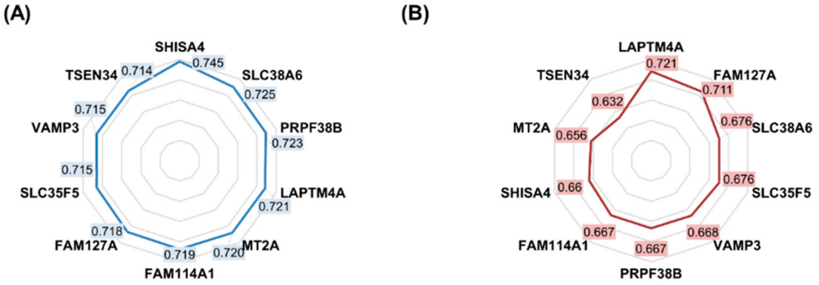

Figure 2. *Cont.*

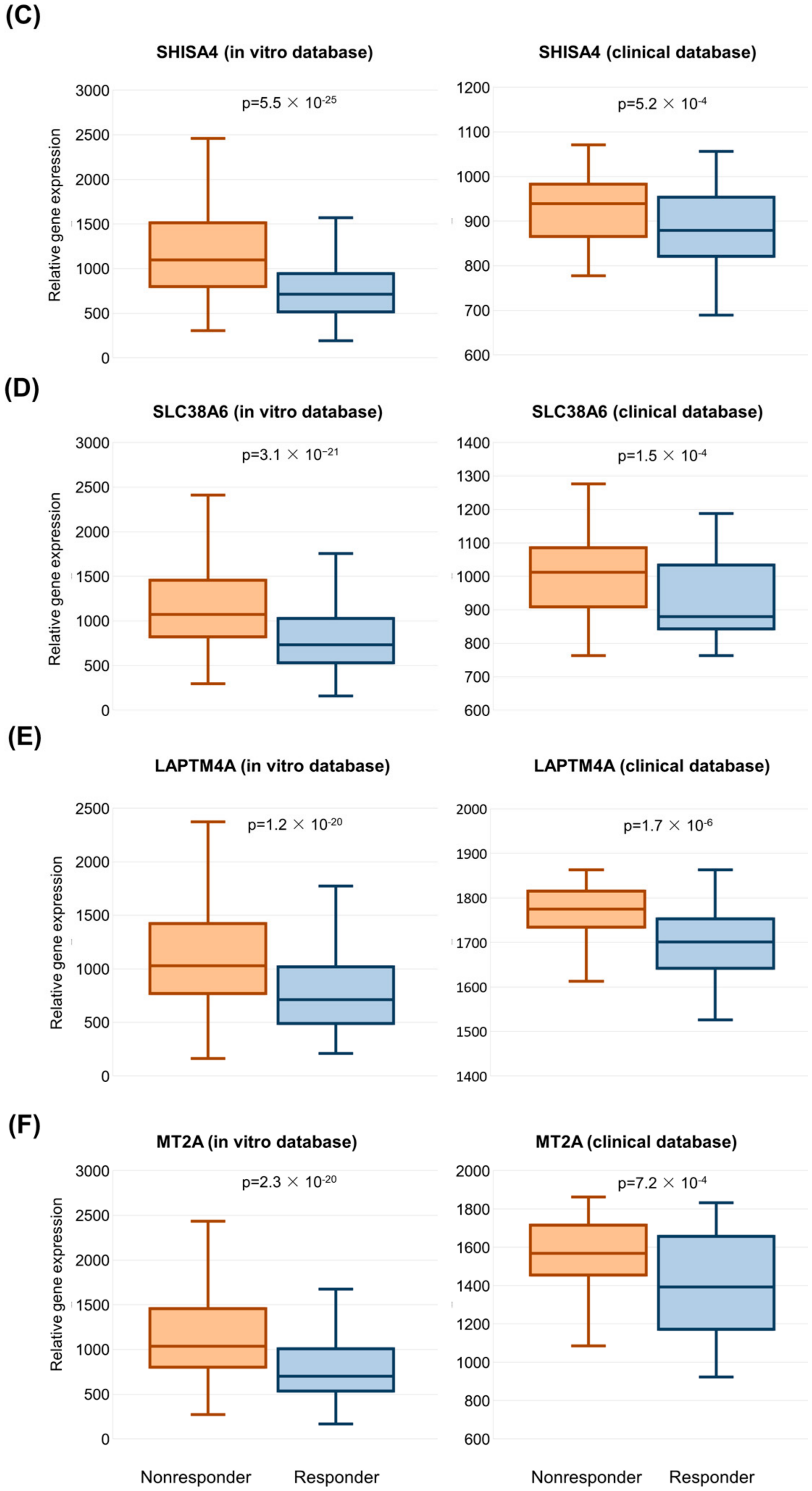

Figure 2. Radar chart of the most significant genes correlated with 5FU therapy response in the in vitro database (**A**) and the human samples (**B**). Boxplots of the most significant genes in both databases (**C**–**F**). The values presented in the radar chart are ROC AUC values for the specific genes.

2.4. Functional Annotation of Validated Genes

Significant genes were further examined by gene ontology biological process and molecular function overrepresentation analysis. The most significant terms were response to oxidative stress (35 genes, adjusted *p*-value: 3.90×10^{-2}) for biological process and GTPase activity (28 genes, adjusted *p*-value: 9.79×10^{-3}) for molecular function. The strongest significant GO categories are presented in Table 4.

Table 4. Result of the functional gene annotations of the new biomarker candidates identified in in vitro and validated human samples. The table shows only the top five genes in terms. The first value is the ROC AUC in vitro, and the second value is the ROC AUC in human samples.

Class	Gene Ontology ID	Gene Ontology Term	Gene Ratio	Adjusted *p*-Value	Top Five Genes included in the Term
Biological process	GO:0006979	response to oxidative stress	35/679	3.90×10^{-2}	*NFE2L2* (0.619; 0.714); *HNRNPD* (0.645; 0.698); *MGST1* (0.610; 0.696) *EGLN1* (0.631; 0.695); *STX4* (0.636; 0.689)
	GO:0000302	response to reactive oxygen species	22/679	4.01×10^{-2}	*NFE2L2* (0.619; 0.714); *HNRNPD* (0.645; 0.698); *EGLN1* (0.631; 0.695); *EEF2* (0.668; 0.689); *PPP2CB* (0.659; 0.689)
	GO:1903311	regulation of mRNA metabolic process	28/679	4.01×10^{-2}	*SF1* (0.670; 0.731); *GIGYF2* (0.624; 0.702); *HNRNPD* (0.645; 0.698); *HNRNPC* (0.621; 0.688); *THRAP3* (0.636; 0.684)
	GO:0048762	mesenchymal cell differentiation	22/679	4.01×10^{-2}	*PDCD6* (0.624; 0.718); *TAPT1* (0.610; 0.708); *CLASP2* (0.651; 0.697); *RTN4* (0.675; 0.690); *SDCBP* (0.644; 0.680)
Molecular function	GO:0003924	GTPase activity	28/693	9.79×10^{-3}	*RND3* (0.701; 0.726); *ARF4* (0.652; 0.703); *KRAS* (0.614; 0.694); *EEF2* (0.668; 0.689); *TUBB2A* (0.673; 0.681)
	GO:0005525	GTP binding	31/693	9.89×10^{-3}	*RND3* (0.701; 0.726); *ARL11* (0.630; 0.703); *ARF4* (0.652; 0.703); *KRAS* (0.614; 0.694); *EEF2* (0.668; 0.689)
	GO:0032550	purine ribonucleoside binding	31/693	9.89×10^{-3}	
	GO:0001883	purine nucleoside binding	31/693	9.89×10^{-3}	
	GO:0032549	ribonucleoside binding	31/693	1.10×10^{-2}	

2.5. Machine Learning Approach to Set Up an Integrated Predictive Tool

In the machine learning analysis, we used the entire combined database by employing the in vitro samples as the training set (total n = 592, nonresponder n = 294, responder n = 298), and the 5FU/capecitabine treated human samples as the test set (total n = 157, nonresponder n = 80, responder n = 77). A total of 12 genes remained in the final model and the genes that had the highest influence on the performance of the model were among the top ten genes identified in the in vitro databases, such as *SHISA4*, *SLC38A6*, *PRPF38B*, and *LAPTM4A*. Other genes with a significant effect on performance and overexpressed in the resistant phenotype include *LPP* and *FOXF2*, whereas genes overexpressed in the sensitive group were *RPL3*, *MAP2K7*, *PCF11*, *INTS7*, and *TAGAP* (Figure 3).

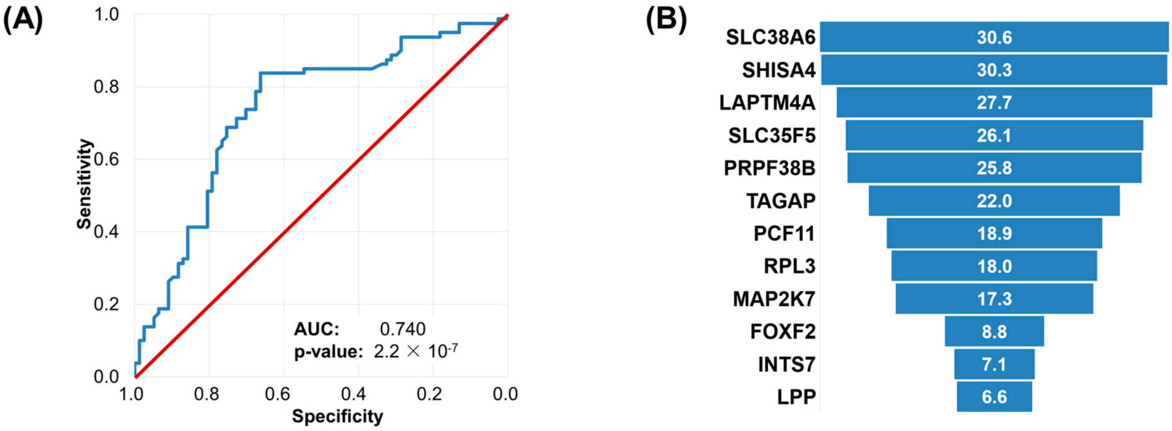

Figure 3. Performance of the random forest-based integrated predictor in the test set comprising of the human samples. The ROC curve of the random forest model tested in the human samples (**A**) and feature importance of the included genes determined by the random forest model (**B**). The red diagonal line depicts a random classifier model, whereas blue line displays our classifier model.

The accuracy of the random forest classifier in the test set was 0.745 (95% CI: 0.67–0.81), sensitivity was 0.83 (95% CI: 72–90%), and specificity was 0.66 (95% CI: 55–77%). The model correctly identified 66 out of 80 nonresponders and 56 out of 77 responders in the test set. The ROC AUC in the test was 0.74 (95% CI: 0.660–0.819, *p*-value: 2.17×10^{-7}).

3. Discussion

In our study, we had two major goals. On the one hand, since the backbone of systemic chemotherapy treatment in colorectal cancer is based on fluoropyrimidines, we first attempted to identify a set of genes that have a potential role in the chemoresistance against these agents. On the other hand, we developed a machine learning-based predictor for the identification of sensitive and resistant colorectal tumors using gene expression data.

Some information is already available for the best-performing genes involved in chemoresistance identified in the in vitro database. *SHISA4* is a member of the Wnt pathway that has been previously linked to 5FU resistance in colorectal cancer [19,20]. The gene *SLC38A6* is involved in glutamine transport and the silencing of the gene resulted in the inhibition of cell cycle progression and cell viability in a hepatocellular cancer cell line [21]. The overexpression of *LAPTM4A* negatively regulates the function of human organic cation transporter 2 [22] and overexpression of the transporter is associated with improved survival in colorectal patients [23]. Remarkably, we have found that all the most important genes determined by the random forest model were also significant in our integrated database.

A strength of our study is the independence of the in vitro training set and the test set of patient samples. We have to emphasize that here we directly linked cell line data which were based on 5FU monotherapy to clinical samples with fluoropyrimidine monotherapy. To our knowledge, this is the first such study with clinically meaningful sample numbers and with a sole focus on fluoropyrimidine monotherapy.

The robustness of our results is also supported by the fact that multiple significant genes of the most important gene ontology categories also contain several markers that have been previously linked to chemoresistance. For example, silencing the *NFE2L2* gene increased 5FU sensitization in hepatocellular carcinoma cells [24], *RTN4* knockdown promoted higher cytotoxic events in paclitaxel-treated cancer cells [25], and *RND3* overexpression promoted drug resistance in gastric cell lines [26].

We have to note a limitation of our study. Although the sensitivity of the random forest model for identifying non-responder patients was 0.83, the achieved specificity was only 0.66. In other words, while the model had a high power for predicting resistance, this did not mean that the alternative estimate automatically confersi therapeutic sensitivity. A future study with an improved prediction model based on higher sample numbers could strengthen the classification sensitivity as well. Secondly, we have to note that we used tertiles of the AUDRC values for classification so that there were no cell lines with similar resistance values. In the nonresponder cell lines, the dose–response curves were markedly shifted to the right. While there could be a difference in response among these cells, the clinical utility of such high concentrations is negligible.

In summary, we identified a set of genes related to resistance against fluoropyrimidines by using gene expression profiles from cell lines and 5FU- and capecitabine-treated patients. Our results will help to rank 5FU biomarker candidates in future studies. We also constructed a machine learning model capable of linking the in vitro and in vivo models, which was able to correctly identify a high proportion of non-responder patients. By utilizing gene expression data from the primary tumor, the random forest model could be used to help therapeutic decision-making.

4. Materials and Methods

4.1. In Vitro Biomarker Discovery

To investigate the genes which may potentially influence the efficacy of 5FU therapy in vitro, we used the Genomics of Drug Sensitivity in Cancer portal (GDSC) version 1 drug screening database as a discovery dataset [27]. The preprocessing and normalization steps for generating the gene expression data tables were executed as described previously [28].

In the in vitro dataset, we used the reported area under the dose–response curve (AUDRC) values to determine the therapeutic response in 5FU-treated cell lines. We defined a cell line as a responder to the treatment if the reported AUDRC value was in the lower tertile considering all 5FU-treated cell lines. Cell lines with an AUDRC value in the upper tertile were categorized as nonresponders. Cell lines with AUDRC values in the middle tertile were excluded from the analysis (Figure 1A). The IC50 values for the cell lines are provided in Supplemental Figure S1.

Gene expression across all genes among nonresponders and responders was compared using Mann–Whitney and receiver operating characteristic (ROC) tests in the R statistical environment (https://www.r-project.org/, accessed on 30 October 2022) using Bioconductor libraries (www.bioconductor.org, accessed on 30 October 2022). The significance cutoff for *p*-values was set at $p < 0.01$. The false discovery rate (FDR) was calculated using the web application www.multipletesting.com (accessed on 30 October 2022) [29], and only results with an FDR of less than 1% were accepted as significant.

The 'clusterProfiler version 4.0.5' R package [30] was used to perform the functional gene annotations of the new biomarker candidates.

4.2. Validation in a Dataset Comprising Clinical Samples

We searched the GEO (https://www.ncbi.nlm.nih.gov/geo/, accessed on 30 October 2022) and GDC (https://portal.gdc.cancer.gov/, accessed on 30 October 2022) portals to identify datasets suitable for the analysis. During this search, the keywords "colorectal", "cancer", "treatment", "response", and "survival" were used. We considered only those publications which included raw gene expression data, information on clinical treatment, response to the treatment or survival information, and involved at least ten patients.

The integrated database comprised two Affymetrix platforms (GPL96: Affymetrix Human Genome U133A, GPL570: Affymetrix Human Genome U133 Plus 2.0), one Agilent platform (GPL6480: Whole Human Genome Microarray 4 × 44K), and the Illumina HiSeq platform.

The raw gene expression data were processed according to standard practices. For the samples measured by Affymetrix gene arrays, normalization was performed with the affy package [31], for the samples with Agilent-based chip, pre-processing was performed with the limma package [32], and samples measured by Illumina HiSeq data were processed using DESeq2 [33]. The integrated complete gene expression dataset consisting of samples from all three platforms was quantile normalized. Finally, a scaling normalization was applied to set the mean expression across all genes in each sample to 1000.

Since most samples were measured with the Affymetrix GPL570 platform, we used this platform as the basis for gene mapping between different platforms. Samples measured by other platforms were mapped to the GPL570 platform via the official gene symbols. We used JetSet [34] to select the most reliable probe set for genes measured by multiple probes. In cases where multiple probes matched a single gene, we selected the probe with the highest interquartile range [35].

Clinical samples were divided into responder and nonresponder groups according to the authors' categorization using the clinical annotation of the source dataset. If the outcome of therapeutic response presented four classes, they were combined into two categories: tumors with "progression" and "stable disease" were categorized as nonresponders, whereas samples with "partial response" or "complete response" were categorized as responders. In the clinical cohort, all genes significant in the training set were analyzed for correlation with resistance.

4.3. Machine Learning Approach to Set Up an Integrated Predictive Tool

We aimed to combine all available genes into a single predictive algorithm. For this, we had to combine the gene expressions of cell lines and human samples into a single database. As the gene expression dynamic ranges yielded a marked difference, we performed an additional normalization using the combat command of the SVA (version: 3.40.0) R package [36]. As for the analysis of the single genes, in the model-building process, the in vitro samples were assigned to the training set and the clinical samples to the test set. For the analysis, we retained only genes upregulated in nonresponder samples if the *p*-value in the Mann–Whitney test was below 0.01 and the FDR was below 1%. Next, we applied the boruta R package as a second feature selection method to detect the most important significant genes, which were then combined into a single predictor using a random forest algorithm. The randomForest (version: 4.6.14) [37] and caret (version: 6.0.90) [38] R packages were used in the analysis. To evaluate the overall performance of the final model, we performed a receiver operating characteristic analysis and determined the area under the curve (AUC).

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/ijms24021508/s1.

Author Contributions: Conceptualization, J.T.F. and B.G.; methodology, J.T.F., B.G. writing—original draft preparation, J.T.F. and B.G.; writing—review and editing, J.T.F. and B.G. All authors have read and agreed to the published version of the manuscript.

Funding: This project was supported by the National Research, Development and Innovation Office (PharmaLab, RRF-2.3.1-21-2022-00015 and 2020-1.1.6-JÖVŐ-2021-00013).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The cell line cohort and the entire human colorectal cancer cohort have been added to the ROCplot.com analysis platform.

Acknowledgments: The authors acknowledge the support of ELIXIR Hungary (www.bioinformatics.hu). This project was supported by the National Research, Development, and Innovation Office (PharmaLab, RRF-2.3.1-21-2022-00015 and 2020-1.1.6-JÖVŐ-2021-00013). The results shown here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. The authors thank James Njoroge for help with the website PHP programming and Viktoria Lakatos and Gabriella Harsányi for the careful English editing of the manuscript.

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

References

1. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* **2021**, *71*, 209–249. [CrossRef] [PubMed]
2. Mármol, I.; Sánchez-de-Diego, C.; Pradilla Dieste, A.; Cerrada, E.; Rodriguez Yoldi, M.J. Colorectal Carcinoma: A General Overview and Future Perspectives in Colorectal Cancer. *Int. J. Mol. Sci.* **2017**, *18*, 197. [CrossRef] [PubMed]
3. Benson, A.B.; Venook, A.P.; Al-Hawary, M.M.; Arain, M.A.; Chen, Y.J.; Ciombor, K.K.; Cohen, S.; Cooper, H.S.; Deming, D.; Farkas, L.; et al. Colon Cancer, Version 2.2021, NCCN Clinical Practice Guidelines in Oncology. *J. Natl. Compr. Cancer Netw.* **2021**, *19*, 329–359. [CrossRef] [PubMed]
4. Dekker, E.; Tanis, P.J.; Vleugels, J.L.A.; Kasi, P.M.; Wallace, M.B. Colorectal cancer. *Lancet* **2019**, *394*, 1467–1480. [CrossRef] [PubMed]
5. Koncina, E.; Haan, S.; Rauh, S.; Letellier, E. Prognostic and Predictive Molecular Biomarkers for Colorectal Cancer: Updates and Challenges. *Cancers* **2020**, *12*, E319. [CrossRef]
6. Fidai, S.S.; Sharma, A.E.; Johnson, D.N.; Segal, J.P.; Lastra, R.R. Dihydropyrimidine dehydrogenase deficiency as a cause of fatal 5-Fluorouracil toxicity. *Autops. Case Rep.* **2018**, *8*, e2018049. [CrossRef]
7. Thorn, C.F.; Marsh, S.; Carrillo, M.W.; McLeod, H.L.; Klein, T.E.; Altman, R.B. PharmGKB summary: Fluoropyrimidine pathways. *Pharmacogenet. Genom.* **2011**, *21*, 237–242. [CrossRef]
8. Longley, D.B.; Harkin, D.P.; Johnston, P.G. 5-Fluorouracil: Mechanisms of action and clinical strategies. *Nat. Rev. Cancer* **2003**, *3*, 330–338. [CrossRef]
9. De Gramont, A.; Figer, A.; Seymour, M.; Homerin, M.; Hmissi, A.; Cassidy, J.; Boni, C.; Cortes-Funes, H.; Cervantes, A.; Freyer, G.; et al. Leucovorin and fluorouracil with or without oxaliplatin as first-line treatment in advanced colorectal cancer. *J. Clin. Oncol.* **2000**, *18*, 2938–2947. [CrossRef]
10. Tsuji, S.; Midorikawa, Y.; Takahashi, T.; Yagi, K.; Takayama, T.; Yoshida, K.; Sugiyama, Y.; Aburatani, H. Potential responders to FOLFOX therapy for colorectal cancer by Random Forests analysis. *Br. J. Cancer* **2012**, *106*, 126–132. [CrossRef]
11. Agostini, M.; Zangrando, A.; Pastrello, C.; D'Angelo, E.; Romano, G.; Giovannoni, R.; Giordan, M.; Maretto, I.; Bedin, C.; Zanon, C.; et al. A functional biological network centered on XRCC3: A new possible marker of chemoradiotherapy resistance in rectal cancer patients. *Cancer Biol. Ther.* **2015**, *16*, 1160–1171. [CrossRef]
12. Del Rio, M.; Molina, F.; Bascoul-Mollevi, C.; Copois, V.; Bibeau, F.; Chalbos, P.; Bareil, C.; Kramar, A.; Salvetat, N.; Fraslon, C.; et al. Gene expression signature in advanced colorectal cancer patients select drugs and response for the use of leucovorin, fluorouracil, and irinotecan. *J. Clin. Oncol.* **2007**, *25*, 773–780. [CrossRef]
13. Estevez-Garcia, P.; Rivera, F.; Molina-Pinelo, S.; Benavent, M.; Gómez, J.; Limón, M.L.; Pastor, M.D.; Martinez-Perez, J.; Paz-Ares, L.; Carnero, A.; et al. Gene expression profile predictive of response to chemotherapy in metastatic colorectal cancer. *Oncotarget* **2015**, *6*, 6151–6159. [CrossRef]
14. Cherradi, S.; Martineau, P.; Gongora, C.; Del Rio, M. Claudin gene expression profiles and clinical value in colorectal tumors classified according to their molecular subtype. *Cancer Manag. Res.* **2019**, *11*, 1337–1348. [CrossRef]
15. Tong, M.; Zheng, W.; Lu, X.; Ao, L.; Li, X.; Guan, Q.; Cai, H.; Li, M.; Yan, H.; Guo, Y.; et al. Identifying clinically relevant drug resistance genes in drug-induced resistant cancer cell lines and post-chemotherapy tissues. *Oncotarget* **2015**, *6*, 41216–41227. [CrossRef]
16. Cherradi, S.; Ayrolles-Torro, A.; Vezzo-Vié, N.; Gueguinou, N.; Denis, V.; Combes, E.; Boissière, F.; Busson, M.; Canterel-Thouennon, L.; Mollevi, C.; et al. Antibody targeting of claudin-1 as a potential colorectal cancer therapy. *J. Exp. Clin. Cancer Res. CR* **2017**, *36*, 89. [CrossRef]
17. Okita, A.; Takahashi, S.; Ouchi, K.; Inoue, M.; Watanabe, M.; Endo, M.; Honda, H.; Yamada, Y.; Ishioka, C. Consensus molecular subtypes classification of colorectal cancer as a predictive factor for chemotherapeutic efficacy against metastatic colorectal cancer. *Oncotarget* **2018**, *9*, 18698–18711. [CrossRef]
18. Ji, D.; Song, C.; Li, Y.; Xia, J.; Wu, Y.; Jia, J.; Cui, X.; Yu, S.; Gu, J. Combination of radiotherapy and suppression of Tregs enhances abscopal antitumor effect and inhibits metastasis in rectal cancer. *J. Immunother. Cancer* **2020**, *8*, e000826. [CrossRef]
19. Intarajak, T.; Udomchaiprasertkul, W.; Bunyoo, C.; Yimnoon, J.; Soonklang, K.; Wiriyaukaradecha, K.; Lamlertthon, W.; Sricharunrat, T.; Chaiwiriyawong, W.; Siriphongpreeda, B.; et al. Genetic Aberration Analysis in Thai Colorectal Adenoma and Early-Stage Adenocarcinoma Patients by Whole-Exome Sequencing. *Cancers* **2019**, *11*, 977. [CrossRef]
20. He, L.; Zhu, H.; Zhou, S.; Wu, T.; Wu, H.; Yang, H.; Mao, H.; SekharKathera, C.; Janardhan, A.; Edick, A.; et al. Wnt pathway is involved in 5-FU drug resistance of colorectal cancer cells. *Exp. Mol. Med.* **2018**, *50*, 101. [CrossRef]
21. Huang, L.; Li, L.; Cheng, B.; Xing, T. SLC38A6, regulated by EP300-mediated modifications of H3K27ac, promotes cell proliferation, glutamine metabolism and mitochondrial respiration in hepatocellular carcinoma. *Carcinogenesis* **2022**, *43*, 885–894. [CrossRef] [PubMed]
22. Grabner, A.; Brast, S.; Sucic, S.; Bierer, S.; Hirsch, B.; Pavenstädt, H.; Sitte, H.H.; Schlatter, E.; Ciarimboli, G. LAPTM4A interacts with hOCT2 and regulates its endocytotic recruitment. *Cell. Mol. Life Sci.* **2011**, *68*, 4079–4090. [CrossRef] [PubMed]
23. Tatsumi, S.; Matsuoka, H.; Hashimoto, Y.; Hatta, K.; Maeda, K.; Kamoshida, S. Organic cation transporter 2 and tumor budding as independent prognostic factors in metastatic colorectal cancer patients treated with oxaliplatin-based chemotherapy. *Int. J. Clin. Exp. Pathol.* **2014**, *7*, 204–212. [PubMed]

24. Chen, F.; Wang, H.; Zhu, J.; Zhao, R.; Xue, P.; Zhang, Q.; Nelson, M.B.; Qu, W.; Feng, B.; Pi, J. Camptothecin suppresses NRF2–ARE activity and sensitises hepatocellular carcinoma cells to anticancer drugs. *Br. J. Cancer* **2017**, *117*, 1495–1506. [CrossRef] [PubMed]
25. Pathak, G.P.; Shah, R.; Kennedy, B.E.; Murphy, J.P.; Clements, D.; Konda, P.; Giacomantonio, M.; Xu, Z.; Schlaepfer, I.R.; Gujar, S. RTN4 Knockdown Dysregulates the AKT Pathway, Destabilizes the Cytoskeleton, and Enhances Paclitaxel-Induced Cytotoxicity in Cancers. *Mol. Ther.* **2018**, *26*, 2019–2033. [CrossRef]
26. Jie, W.; Andrade, K.C.; Lin, X.; Yang, X.; Yue, X.; Chang, J. Pathophysiological Functions of Rnd3/RhoE. *Compr. Physiol.* **2015**, *6*, 169–186.
27. Yang, W.; Soares, J.; Greninger, P.; Edelman, E.J.; Lightfoot, H.; Forbes, S.; Bindal, N.; Beare, D.; Smith, J.A.; Thompson, I.R.; et al. Genomics of Drug Sensitivity in Cancer (GDSC): A resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.* **2013**, *41*, D955–D961. [CrossRef]
28. Tibor Fekete, J.; Győrffy, B. A unified platform enabling biomarker ranking and validation for 1562 drugs using transcriptomic data of 1250 cancer cell lines. *Comput. Struct. Biotechnol. J.* **2022**, *20*, 2885–2894. [CrossRef]
29. Menyhart, O.; Weltz, B.; Győrffy, B. MultipleTesting.com: A tool for life science researchers for multiple hypothesis testing correction. *PLoS ONE* **2021**, *16*, e0245824. [CrossRef]
30. Wu, T.; Hu, E.; Xu, S.; Chen, M.; Guo, P.; Dai, Z.; Feng, T.; Zhou, L.; Tang, W.; Zhan, L.; et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation* **2021**, *2*, 100141. [CrossRef]
31. Gautier, L.; Cope, L.; Bolstad, B.M.; Irizarry, R.A. affy—Analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **2004**, *20*, 307–315. [CrossRef]
32. Ritchie, M.E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C.W.; Shi, W.; Smyth, G.K. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **2015**, *43*, e47. [CrossRef]
33. Love, M.I.; Huber, W.; Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **2014**, *15*, 550. [CrossRef]
34. Li, Q.; Birkbak, N.J.; Gyorffy, B.; Szallasi, Z.; Eklund, A.C. Jetset: Selecting the optimal microarray probe set to represent a gene. *BMC Bioinform.* **2011**, *12*, 474. [CrossRef]
35. Walsh, C.J.; Hu, P.; Batt, J.; Santos, C.C.D. Microarray Meta-Analysis and Cross-Platform Normalization: Integrative Genomics for Robust Biomarker Discovery. *Microarrays* **2015**, *4*, 389–406. [CrossRef]
36. Leek, J.T.; Johnson, W.E.; Parker, H.S.; Jaffe, A.E.; Storey, J.D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **2012**, *28*, 882–883. [CrossRef]
37. Liaw, A.; Wiener, M. Classification and Regression by randomForest. *R News* **2002**, *2*, 5.
38. Kuhn, M. Building Predictive Models in R Using the caret Package. *J. Stat. Softw.* **2008**, *28*, 1–26. [CrossRef]

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.