Identification of Protein Markers Predictive of Drug-Specific Survival Outcome in Cancers

Shuting Lin[1], Jie Zhou[1], Yiqiong Xiao[1], Bridget Neary[1], Yong Teng[2], and Peng Qiu[1,2(✉)]

[1] Georgia Institute of Technology, Atlanta, USA
shutinglin@gatech.edu, peng.qiu@bme.gatech.edu
[2] Emory University, Atlanta, USA

Abstract. Novel discoveries of biomarkers predictive of drug-specific responses not only play a pivotal role in revealing the drug mechanisms in cancers, but are also critical to personalized medicine. In this study, we identified drug-specific biomarkers by integrating protein expression data, drug treatment data and survival outcome of 7076 patients from The Cancer Genome Atlas (TCGA). We first defined cancer-drug groups, where each cancer-drug group contains patients with the same cancer and treated with the same drug. For each protein, we stratified the patients in each cancer-drug group by high or low expression of the protein, and applied log-rank test to examine whether the stratified patients show significant survival difference. We examined 336 proteins in 98 cancer-drug groups and identified 65 protein-cancer-drug combinations involving 55 unique proteins, where the protein expression levels are predictive of drug-specific survival outcomes. Some of the identified proteins were supported by published literature. Using the gene expression data from TCGA, we found the mRNA expression of ∼11% of the drug-specific proteins also showed significant correlation with drug-specific survival, and most of these drug-specific proteins and their corresponding genes are strongly correlated.

Keywords: Protein · Drug response · Survival analysis

1 Introduction

The high inter-individual variability in drug response makes it a great challenge to develop personalized treatment strategies for individual patients [1]. Therefore, personalized medicine is a research area of great interest in terms of optimizing therapeutic options and improving patient clinical outcomes. One essential aspect for personalized medicine is to identify biomarkers that are predictive of drug treatment responses [2]. Rapid technological advances in cancerogenic research have facilitated the discovery of genetic variants as predictive and prognostic biomarkers associated with drug efficacy and patient clinical outcomes [3].

Y. Wei et al. (Eds.): ISBRA 2021, LNBI 13064, pp. 58–67, 2021.
https://doi.org/10.1007/978-3-030-91415-8_6

In the literature, numerous pharmacogenetic studies have investigated the relationship between molecular expression profiles and patient survival outcomes, and identified prognostic biomarkers in cancers [4]. Most of the existing studies chose to include as many subjects relevant to their scopes as possible, while frequently ignoring the fact that these patients might receive different treatments. In our opinion, there are two main reasons for such choices. First, population-based studies with a larger sample size often have increased statistical power for identification of biomarkers [5]. Second, drug treatment data is often either unavailable or in non-standardized formats that are difficult to incorporate. As a result, cancer survival biomarkers identified in existing studies are often general to the cancer being studied, but not specific to any drug treatments. However, studying biomarkers in a drug-specific manner has the potential to reveal the underlying cancer mechanisms and inform designs of personalized medicine.

The Cancer Genome Atlas (TCGA) is one of the most powerful cancer genomics programs to date [6], which provides massive data to expand our knowledge of tumourigenesis. TCGA has generated a large public collection of multiple types of omic data on ∼11,000 cancer patients across 33 different cancer types. The omic data types include mutation, copy number variation, methylation, gene expression, miRNA expression and protein expression data. TCGA also provides drug treatment data and survival outcomes of the patients. In the drug treatment data in TCGA, there are nomenclature problems (i.e., alternative names, abbreviations and misspellings), making it difficult for bioinformatics analysis. In our previous study [7–9], we manually standardized the drug names in the drug treatment data, which enabled us to examine the potential for gene copy number and gene expression as biomarkers of drug-specific survival.

Here, we focused on investigating the potential of proteins as drug-specific predictive biomarkers, since proteins are the functional units in the central dogma of molecular biology. And it has already been hypothesized that proteomic profiling more directly addresses biological and pharmacologic problems in cancer [10]. In recent years, proteomics efforts have led to proteins that can serve as cancer biomarkers. Several lines of evidence have shown that the expression level of proteins is frequently associated with drug response. One example is MRP1, which is associated with drug resistance or poor patient outcomes in a variety of cancers [23]. MRP3 is the ABC transporter that is most closely related to MRP1. For both MRP3 and MRP1, their protein expression levels correlated with decreased sensitivity of lung cancer cell lines to doxorubicin [11]. Another well-characterized example is eight protein signatures that were identified for the prediction of drug response to 5-fluorouracil, including CDH1, CDH2, KRT8, ERBB2, MSN, MVP, MAP2K1, and MGMT. All of these proteins, except for KRT8, are involved in the pathogenesis of colon cancer [12].

In this study, we performed survival analyses on patients with the same cancer and were exposed to the same drug, and identified proteins whose expression levels are associated with drug-specific survival outcome. Some of the identified protein markers were further supported in the literature, where we found multiple published papers indicating their relationship with drug response in cancers.

However, we also found a few drug-specific proteins that are inconsistent with previously reported findings in terms of the direction of correlation with survival outcomes. In addition, using the gene expression data in TCGA, we explored the regulatory mechanism of predictive protein markers by examining their coding genes.

2 Results

2.1 Significant Proteins Predictive of Drug-Specific Survival

To identify proteins correlated with drug-specific survival outcomes, we grouped patients who suffered from the same cancer and received the same drug together, which we call cancer-drug groups. Across the 33 cacner types in TCGA and the 254 unique drug names from our previous manual standardization of the drug treatment data [7–9], a large number of cancer-drug groups contained 0 or very small number of patients, because not all drugs were applied to treat all cancer types. We imposed a minimum sample size requirement of 15, and only considered cancer-drug groups whose number of patients exceeded this threshold. Therefore, a total of 98 cancer-drug groups were considered for the subsequent analysis to identify protein markers for drug-specific survival.

Next, we binarized the protein expression data in TCGA, which was needed in our survival analysis. For each of the 336 proteins measured by TCGA, we applied StepMiner [13] to binarize its expression data across all patients in all cancer types. Specifically, for each protein, we sorted its expression data for all patients and then fitted a step function to the sorted data that minimizes the square error between the original and the fitted values. The step function provided a threshold to binarize the expression of the protein.

Finally, we performed survival analysis to evaluate each protein's ability to predict the survival outcome of patients in each cancer-drug group. Patients in the cancer-drug group were stratified into a highly-expressed class and a lowly-expressed class based on the binarized data of the protein. To minimize undesired statistical bias, we only performed survival analysis on proteins in cancer-drug groups with at least 5 lowly-expressed patients and 5 highly-expressed patients. In total, 17,812 protein-cancer-drug combinations were tested in our analysis, which involved 23 cancer types and 41 drugs. We applied log-rank test to determine the statistical significance of survival difference between highly-expressed class and lowly-expressed class. 90 proteins exceeding an FDR threshold of < 0.1 were selected as predictive markers whose expression levels were related to patients' survival outcome in a drug-specific manner. In order to identify proteins that are specifically related to individual drugs, we performed the same analysis on all patients in each cancer type, and identified proteins that are predictive of cancer-specific survival irrespective of drug treatment. Among the 90 proteins significant for drug-specific survival, 25 were also identified in the cancer-specific analysis. In our subsequent analysis, we excluded the 25, and only included the protein markers that were significant in cancer-drug groups but not significant in the corresponding cancer types.

Table 1. 65 Significant protein-cancer-drug combinations identified in cancer-drug groups

Cancer-drug groups	Protein markers for drug-specific survival	Number of patients
BLCA-Gemcitabine	YWHAE,AKT1-3,CDK1,MAPK1	75
BRCA-Carboplatin	AKT1	24
BRCA-Doxorubicin	CDK1	282
CESC-Cisplatin	TP53BP1,Tubulins,BAP1,CTNNB1,COL6A1, CDH1,EIF4G1,HSPA1A,KU80,MRE11,CDH2, SERPINE1,TSC2	61
COAD-Oxaliplatin	EIF4EBP1	82
HNSC-Carboplatin	BCL2,CLDN7,FOXM1,MSH2	22
KIRC-Sorafenib	KDR	16
LGG-Bevacizumab	EIF4EBP1,ETS1,FASN,TIGAR,TSC2	41
LGG-Irinotecan	ETS1,KU80,SRSF1,FYN	20
LGG-Lomustine	CTNNA3,AR,CDKN1A,BRAF	31
LUAD-Cisplatin	NF2	66
LUSC-Gemcitabine	ABL1	26
LUSC-Cisplatin	CCNE2,PEA15	54
OV-Docetaxel	CCNE1,RBM15	69
OV-Carboplatin	CASP7,RICTOR	290
OV-Doxorubicin	CDH1,GAB2,RBM15	106
OV-Vinorelbine	AKT1-3,CDK1,ERCC5	19
PAAD-Gemcitabine	CDKN1B,MAPK11	73
PAAD-Fluorouracil	BECN1,GAPDH,ERBB3,CDKN1B,MAPK11, MAPK12,PTEN,SRC,PARP1	24
STAD-Cisplatin	DPP4	48
STAD-Etoposide	INPP4B	19

As showed in Table 1, a total of 65 significant protein-cancer-drug combinations were identified in 21 cancer-drug groups, which involved 55 unique proteins. We found 13 significant proteins predictive of cisplatin response in cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), and 9 protein markers are associated with fluorouracil response in pancreatic adenocarcinoma (PAAD). Interestingly, there are 11 proteins that turned out to be significant in multiple cancer-drug groups, which may potentially serve as key biomarkers to drug responses in multiple cancer types. Among the proteins that were significant in more than one cancer-drug group, we observed that CDH1 was related to the sensitivity of cisplatin in CESC and also associated with the overall survival of Doxorubicin-treated patients in ovarian serous cystadenocarcinoma (OV). We also found that CDK1 was correlated with drug response to gemcitabine in bladder urothelial carcinoma (BLCA), doxorubicin in Breast invasive carcinoma (BRCA), and vinorelbine in OV.

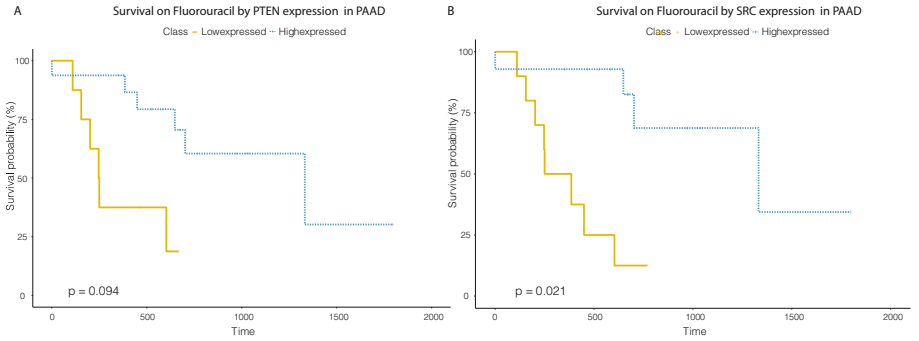

Fig. 1. Kaplan-Meier curves of overall survival for patients treated with fluorouracil at low or high expressed classes stratified by PTEN or SRC in the PAAD.

2.2 Literature Support of Predictive Protein Markers

To assess whether there are previous research that supported the identified protein markers predictive of drug response, we conducted a comprehensive literature survey on the PubMed database for each of the 65 protein-drug combinations. We found supportive evidence for multiple protein-drug combinations in various cancer contexts.

In particular, our analysis suggested that high-expressed CDKN1B was able to increase drug response to gemcitabine in pancreatic adenocarcinoma (PAAD). This is consistent with previous studies that the re-expression of CDKN1B was related to the sensitization of pancreatic cancer cells to gemcitabine leading to a significant induction of apoptosis, which could be a superior potential treatment for pancreatic cancer [14,15]. Another literature support is about PTEN. PTEN was first discovered as a tumor suppressor, and its loss of function is strongly associated with tumor growth and survival. Figure 1A shows how PTEN expression correlated with PAAD patients in TCGA, that PTEN over-expression led to increased sensitivity of fluorouracil. This observation is supported by previous studies which showed that PTEN was involved in promoting 5-Fluorouracil-induced apoptosis, and the reduced expression of PTEN was associated with increased malignancy grade in PAAD, whereas maintenance of PTEN expression showed a trend toward a longer survival [16]. In addition, it has been shown that the inhibition of TAP subsequently promoted the expression of PTEN that increase sensitivity to chemotherapeutic agents in cancer [17]. A third example is VEGFR2, which was previously reported to be predictive of sorafenib efficacy in patients with metastatic renal cell carcinoma (mRCC) and was associated with longer overall survival of patients those treated with sorafenib [18]. In our analysis, we found that the repressed VEGFR2 resulted in prolonged survival outcomes of patients exposed to sorafenib in Kidney renal clear cell carcinoma (KIRC), which reveals the potential prediction of VEGFR2 on gemcitabine in other diseases.

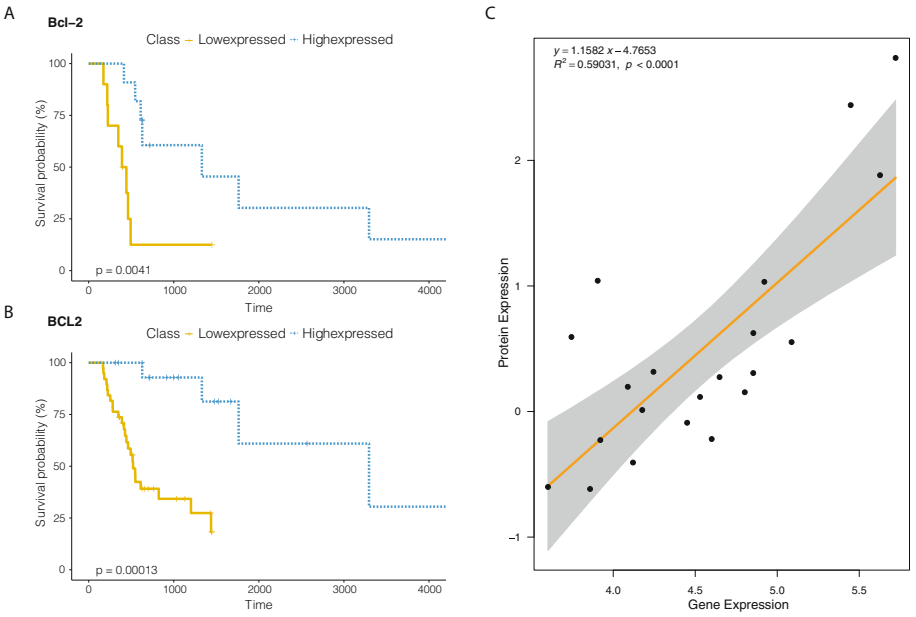

Fig. 2. Correlation between Bcl-2 and BCL2 in expression level and survival outcomes. A-B. Kaplan-Meier curves of overall survival for patients treated with fluorouracil at low or high expressed classes stratified by Bcl-2 or BCL2 in the HNSCC. C. Consistency of Bcl-2 and BCL2 correlation in expression levels for patients in HNSCC-carboplatin group.

We also found literature that showed inconsistent direction of survival impact compared to two of the protein-drug combinations we identified. Our analysis suggested that decreased SRC was related to poor overall survival of patients in PAAD and treated with Fluorouracil, shown in Fig. 1B. However, a recent study that involved fluorouracil and a few other drugs showed that SRC expression up-regulation in some Pancreatic ductal adenocarcinoma (PDAC) patients was associated to relatively poor patient outcome [19]. The second inconsistency was related to BCL2. Our analysis demonstrated that the over-expression of BCL2 resulted in better survival outcomes of patients with Head and Neck squamous cell carcinoma (HNSCC) and exposed to carboplatin (Fig. 2A), and the high-expressed BCL2-coding gene was also associated with prolonged overall survival of patients with HNSCC (Fig. 2B). In contrast, a previous study observed that BCL2 could inhibit apoptosis induced by cisplatin, carboplatin and paclitaxel, making HNSCC that express BCL2 resistant to rapamycin, carboplatin and paclitaxel [20]. Despite of these inconsistencies in the direction of correlation with survival, the literature did indicate the relevance of our identified proteins to drug responses in cancer patients.

2.3 Correlation Between Predictive Proteins and Their Coding Genes

To understand the roles of drug-specific proteins during carcinogenesis and pharmacotherapy, we investigated the regulatory mechanism of identified protein markers by examining their corresponding coding genes. We performed survival analysis on the genes coded the protein markers, in the same cancer-drug context where the protein markers were identified. Specificity, for each of the 65 identified protein-cancer-drug combinations, we extracted the binarized gene expression data of the corresponding gene for patients in that cancer-drug group, stratified the patients to high-expressed and low-expressed classes according to the binarized gene expression data, and performed log-rank test to examine whether there is a significant difference in survival outcome between the two classes. We applied p-value threshold of < 0.05 to identify genes whose expression were also predictive of drug-specific survival outcomes. Similar to our analysis of proteins, survival analysis was only performed on the corresponding genes if there were at least 5 highly-expressed patients and 5 lowly-expressed patients in the corresponding cancer-drug group. 7 genes were identified whose expression were also predictive of drug-specific survival, in the same context as its associated protein-cancer-drug combinations. Therefore, this result suggests $\sim$11% of the identified proteins also showed significance in their corresponding genes.

To elucidate the relationships between significant proteins and their corresponding genes in each cancer-drug group, we examined the correlation between the expression levels of protein and gene in each of the 7 significant protein-gene pair. Correlation analysis was performed between log-transformed gene expression data and protein expression data by using R package 'lm'. Among the 7 protein markers whose gene expression also correlated with survival in the same cancer-drug groups, we noticed that 4 (BCL2, CCNE2, ETS1, GAB2) showed positive correlation between gene expression level and protein expression level, while the remaining 3 (MAPK3, TIGAR, CTNNB1) showed negative correlation.

We also examined the consistency between the survival analyses based on the proteins and the genes. For example, for a particular protein-cancer-drug combination whose corresponding gene was also predictive of drug-specific survival, we examined the direction of their correlation with survival outcome. If high expression of the protein led to better survival in the cancer-drug group, we considered the protein to be positively correlated with survival outcome. If high expression of the corresponding gene also led to better survival in the cancer-drug group, the gene was also positively correlated with survival outcome. In this case, the protein and its corresponding gene showed consistency in terms of their directions of the survival outcome. However, if high protein expression and low gene expression led to better survival, or low protein expression and high gene expression led to better survival, the protein and its corresponding gene were inconsistent in their directions of the survival outcome. Similar to the correlation analysis above, out of the 7 proteins whose corresponding genes were also predictive of drug-specific survival outcomes, 4 showed consistent survival

directions between genes and proteins, whereas the remain 3 showed inconsistent survival direction. This is not surprising, given mixed reports in the literature on the concordance and discordance between gene expression and protein expression in various contexts [21,22].

3 Materials and Methods

3.1 Data Access

TCGA protein expression data and gene expression data were obtained from Genomic Data Commons (GDC) database using the GDC Data Transfer Tool. Clinical data were also downloaded from GDC, which included patients' drug treatment records and survival outcomes. After removing duplicates in the molecular data and filtering for samples with treatment and survival data, we finally used a total of 31 cancer types in this study.

3.2 Data Preprocessing

The gene expression data downloaded from TCGA have been normalized by FPKM-UQ, and we subsequently preprocessed the gene expression data by log-transformation. The protein expression data available from TCGA have already been properly normalized and transformed. For each gene and each protein, we used the StepMiner algorithm [13] to compute a global threshold for all patients across all cancer types. Specifically, we sorted the expression data across all patients from low to high for each gene or protein, and then a step function was fitted to minimize the square error between the original and the fitted values. Using the threshold, the normalized protein and gene expression data are binarized, so that patients can be divided into two classes (high-expressed class vs. low-expressed class) based on expression levels of each individual protein and gene features.

3.3 Survival Analysis

For each protein, patients who suffered from the same cancer and received the same drug were stratify into highly- or lowly-expressed classes according to the binarized data of that protein. We used log-rank test to compare the survival differences between patients in highly- and lowly-expressed classes. Benjamini-Hochberg multiple tests were used to calibrate the false discovery rate (FDR) for the significance. Proteins with FDR < 0.1 were identified as drug-specific markers whose expression expression levels were predictive of patients' survival outcome in a drug-specific manner. Kaplan-Meier analysis and log-rank test in this study were conducted using the R package 'survival'.

3.4 Literature Search

We performed literature searches on PubMed database to find articles that mentioned proteins interacting with drugs in the cancer-drug context from which the proteins were identified. We used a Python script with the Bio.Entrez package from Biopython, and programmatically searched the National Library of Medicine PubMed database (http://www.ncbi.nlm.nih.gov/pubmed). Keywords for the searches were drug AND protein markers in all fields, including the title, abstract and main texts of the articles.

4 Conclusion

In this study, we integrated multiple data types in TCGA to perform survival analysis for patients who belonged to the same type of cancer and exposed to the same drug. This analysis identified predictive protein markers whose expression levels are associated with drug-specific survival outcomes in various cancer types. Notably, our results included proteins that have been previously reported to be predictive biomarkers for drug sensitivity and resistance in cancers, as well as the novel ones that have not been proposed in the literature. In addition, we examined gene expression of the identified proteins in terms of both the correlations between their expression levels and their correlations with survival. Overall, the drug-specific proteins identified in this analysis may be effective biomarkers predictive of drug response and survival outcomes in cancers. Further validation investigation on these protein markers can help guide clinical decisions for individual patients.

Acknowledgement. This work was supported by funding from the National Science Foundation (CCF1552784 and CCF2007029). P.Q. is an ISAC Marylou Ingram Scholar, a Carol Ann and David D. Flanagan Faculty Fellow, and a Wallace H. Coulter Distinguished Faculty Fellow.

References

1. Latini, A., Borgiani, P., Novelli, G., Ciccacci, C.: miRNAs in drug response variability: potential utility as biomarkers for personalized medicine. Pharmacogenomics **20**(14), 1049–1059 (2019)
2. Li, B., He, X., Jia, W., Li, H.: Novel applications of metabolomics in personalized medicine: a mini-review. Molecules **22**(7), 1173 (2017)
3. Arbitrio, M., et al.: Pharmacogenomics biomarker discovery and validation for translation in clinical practice. Clin. Transl. Sci. **14**(1), 113–119 (2021)
4. Fu, Q., et al.: miRomics and proteomics reveal a miR-296-3p/PRKCA/FAK/Ras/c-Myc feedback loop modulated by HDGF/DDX5/β-catenin complex in lung adenocarcinoma. Clin. Cancer Res. **23**(20), 6336–6350 (2017)
5. Hong, E.P., Park, J.W.: Sample size and statistical power calculation in genetic association studies. Genom. Inf. **10**(2), 117 (2012)

6. Tomczak, K., Czerwińska, P., Wiznerowicz, M.: The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp. Oncol **19**(1A), A68 (2015)
7. Spainhour, C., Qiu, P.: Identification of gene-drug interactions that impact patient survival in TCGA. BMC Bioinf **17**(1), 409 (2016)
8. Spainhour, C., Lim, J., Qiu, P.: GDISC: integrative TCGA analysis to identify gene-drug interaction for survival in cancer. Bioinformatics **33**(9), 1426–1428 (2017)
9. Neary, B., Zhou, J., Qiu, P.: Identifying gene expression patterns associated with drug-specific survival in cancer patients. Sci. Rep. **11**(1), 5004 (2021)
10. Ma, Y., et al.: Predicting cancer drug response by proteomic profiling. Clin. Cancer Res. **12**(15), 4583–4589 (2006)
11. Young, L.C., Campling, B.G., Cole, S.P., Deeley, R.G., Gerlach, J.H.: Multidrug resistance proteins mrp3, mrp1, and mrp2 in lung cancer: correlation of protein levels with drug response and messenger rna levels. Clin. Cancer Res. **7**(6), 1798–1804 (2001)
12. Ginsburg, G.S., Willard, H.F.: Essentials of Genomic and Personalized Medicine. Academic Press, Cambridge (2009)
13. Sahoo, D., Dill, D.L., Tibshirani, R., Plevritis, S.K.: Extracting binary signals from microarray time-course data. Nucl. Acids Res. **35**(11), 3705–3712 (2007)
14. Khan, M.A., Zubair, H., Srivastava, S.K., Singh, S., Singh, A.P.: Insights into the role of microRNAs in pancreatic cancer pathogenesis: potential for diagnosis, prognosis, and therapy. In: Santulli, G. (ed.) microRNA: Cancer. AEMB, vol. 889, pp. 71–87. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-23730-5_5
15. Chen, X., et al.: Therapeutic effects of argyrin f in pancreatic adenocarcinoma. Cancer Lett. **399**, 20–28 (2017)
16. Ni, S.: Cbx7 suppresses cell proliferation, migration, and invasion through the inhibition of pten/akt signaling in pancreatic cancer. Oncotarget **8**(5), 8010 (2017)
17. Tian, Y., et al.: Metformin mediates resensitivity to 5-fluorouracil in hepatocellular carcinoma via the suppression of yap. Oncotarget **7**(29), 46230 (2016)
18. Hutson, T.E., et al.: Randomized phase III trial of temsirolimus versus sorafenib as second-line therapy after sunitinib in patients with metastatic renal cell carcinoma. J. Clin. Oncol. **32**(8), 760 (2014)
19. Abrams, S.L., et al.: Introduction of wt-tp53 into pancreatic cancer cells alters sensitivity to chemotherapeutic drugs, targeted therapeutics and nutraceuticals. Adv. Biol. Regul. **69**, 16–34 (2018)
20. Aissat, N., et al.: Antiproliferative effects of rapamycin as a single agent and in combination with carboplatin and paclitaxel in head and neck cancer cell lines. Cancer Chemother. Pharmacol. **62**(2), 305–313 (2008)
21. Gygi, S.P., Rochon, Y., Franza, B.R., Aebersold, R.: Correlation between protein and MRNA abundance in yeast. Molec. Cell. Biol. **19**(3), 1720–1730 (1999)
22. Chen, G., et al.: Discordant protein and MRNA expression in lung adenocarcinomas. Molec. Cell. Proteom. **1**(4), 304–313 (2002)
23. Munoz, M., Henderson, M., Haber, M., Norris, M.: Role of the MRP1/ABCC1 multidrug transporter protein in cancer. IUBMB Life **59**(12), 752–757 (2007)