Perspectives

Clinical Cancer Research

Curation of the Pancreatic Ductal Adenocarcinoma Subset of the Cancer Genome Atlas Is Essential for Accurate Conclusions about Survival-Related Molecular Mechanisms

Ivana Peran[1], Subha Madhavan[1,2], Stephen W. Byers[1], and Matthew D. McCoy[1,2]

Abstract

Purpose: Publicly available databases, for example, The Cancer Genome Atlas (TCGA), containing clinical and molecular data from many patients are useful in validating the contribution of particular genes to disease mechanisms and in forming novel hypotheses relating to clinical outcomes.

Experimental Design: The impact of key drivers of cancer progression can be assessed by segregating a patient cohort by certain molecular features and constructing survival plots using the associated clinical data. However, conclusions drawn from this straightforward analysis are highly dependent on the quality and source of tissue samples, as demonstrated through the pancreatic ductal adenocarcinoma (PDAC) subset of TCGA.

Results: Analyses of the PDAC-TCGA database, which contains mainly resectable cancer samples from patients in stage IIB, reveal a difference from widely known historic median and 5-year survival rates of PDAC. A similar discrepancy was observed in lung, stomach, and liver cancer subsets of TCGA. The whole transcriptome expression patterns of PDAC-TCGA revealed a cluster of samples derived from neuroendocrine tumors, which have a distinctive biology and better disease prognosis than PDAC. Furthermore, PDAC-TCGA contains numerous pseudo-normal samples, as well as those that arose from tumors not classified as PDAC.

Conclusions: Inclusion of misclassified samples in the bioinformatic analyses distorts the association of molecular biomarkers with clinical outcomes, altering multiple published conclusions used to support and motivate experimental research. Hence, the stringent scrutiny of type and origin of samples included in the bioinformatic analyses by researchers, databases, and web-tool developers is of crucial importance for generating accurate conclusions. *Clin Cancer Res; 24(16); 3813–9. ©2018 AACR.*

Introduction

Completion of the Human Genome Project in 2003 (1, 2) allowed for association studies mapping genomic variation, particularly mutations, with disease phenotypes. Besides mutations, disease incidence can also be a consequence of aberrant gene expression or protein activation status. Fortunately, completion of the human genome also laid the foundation for mapping of expressed mRNA transcripts to their genomic sequences, which has subsequently revealed the diversity of gene expression patterns responsible for the underlying physiology of healthy and diseased tissues.

At the same time, the field of molecular medicine recognized an opportunity for the development of targeted therapies for patients with similar gene mutations, expression, and protein activation patterns. Eventually, the scientific and medical community turned toward personalized therapy, specifically accounting for the molecular mechanism at work in an individual patient. Technical developments and high-throughput screening allowed us to gather information on gene mutation and expression, as well as protein activation across a large number of patients. The resulting availability of datasets that include information about phenotypic and molecular features has been a significant driver of translational research over the last decade.

The Cancer Genome Atlas (TCGA) began cataloging molecular datasets for various cancer types in 2005, and has since grown to include extensive molecular and clinical information for individual cancer patient samples (3). Today, the data for 33 different tumor types based on tissues collected from over 11,000 patients is available (https://cancergenome.nih.gov). To support access and use of the data, many publically available web tools use processed TCGA datasets to enable rapid analysis of gene expression pattern, mutation analysis, and other molecular and clinical features associated with a specific cancer type. This allows researchers to compare their findings with the TCGA datasets, or correlate particular molecular features with a clinical outcome.

Here, we focus on the pancreatic ductal adenocarcinoma (PDAC) dataset of TCGA (annotated as PAAD within TCGA),

[1]Georgetown-Lombardi Comprehensive Cancer Center, Department of Oncology, Georgetown University Medical Center, Washington, DC. [2]Innovation Center for Biomedical Informatics, Georgetown University, Washington, DC.

Note: Supplementary data for this article are available at Clinical Cancer Research Online (http://clincancerres.aacrjournals.org/).

Corresponding Authors: Ivana Peran, Georgetown-Lombardi Comprehensive Cancer Center, Department of Oncology, Georgetown University Medical Center, 3970 Reservoir Road, NW, New Research Building, Room E415, Washington, DC 20057. Phone: 202-687-1891; Fax: 202-687-7505; E-mail: ip62@georgetown.edu; and Matthew D. McCoy, Matthew.McCoy@georgetown.edu

doi: 10.1158/1078-0432.CCR-18-0290

Translational Relevance

Publicly available molecular databases have significantly contributed to translational research, especially when validating the impact of a particular gene from a preclinical to clinical setting or generating hypotheses related to prognostic biomarkers. In response to the arrival of databases containing annotated information on molecular and clinical features across a variety of diseases, numerous web tools for big data analysis have emerged. Here we illustrate the importance of scrutinizing the underlying content when using common analysis methods to validate laboratory findings and generate a novel hypothesis predicting phenotypic associations with molecular signatures. We advise researchers to carefully consider the origin and characteristics of the individual samples included in the bioinformatic analyses and recommend exclusion of data originating from (pseudo)-normal or other cell origins with distinct underlying biology. A failure to properly curate the sample pool can lead to inaccurate conclusions about clinically relevant biomarkers and misinformed hypotheses about disease-related mechanisms.

which contains genomic, transcriptomic, and proteomic analyses of 185 patient samples (https://tcga-data.nci.nih.gov/docs/publications/tcga/). In addition to molecular data, the PDAC-TCGA annotated dataset contains numerous other pieces of clinical information, including overall survival, records of therapy and surgery performed, metastatic status, history of diabetes, smoking, alcohol consumption, etc. However, researchers utilizing the dataset for validation and hypothesis generation should be wary of conclusions drawn using the entire collection of samples. Despite its name, the PDAC dataset includes samples of various cell origin, as well as (pseudo)-normal samples. As shown in Table 1, some of the cancer samples did not arise from the pancreas; belong to neuroendocrine tumors, or to different subtypes of pancreatic cancer not classified as adenocarcinoma. We show here that inclusion of those samples into analyses skews the data and undermines the subsequent application of the conclusions to future studies. Considering the potential consequences that researchers may run into while using publically available software with preloaded datasets, it is important to highlight the impact of sample tissue type and/or cell of origin on molecular biomarker discovery.

Materials and Methods

The publicly available PDAC dataset of the TCGA (annotated as PAAD within TCGA) was initially analyzed using Morpheus (Broad Institute, Cambridge, MA; https://software.broadinstitute.org/morpheus/) and Xenabrowser software (University of California Santa Cruz; Xena, http://xena.ucsc.edu). The processed gene expression data were downloaded from Morpheus, and analyzed using hierarchal clustering to identify subsets of samples with related gene expression patterns. Fold changes for genes commonly associated with PDAC progression (4–7) were normalized to their median expression to quantify the distribution across the entire dataset. The data on pancreatic cancer statistics was obtained through the NIH – NCI – Surveillance, Epidemiology, and End Results Program; (https://seer.cancer.gov/statfacts/html/pancreas.html). Kaplan–Meier graphs were generated for population subsets showing differential expression for a given PDAC-associated gene and statistical significance was calculated (log-rank test and Gehan–Breslow–Wilcoxon test) by using GraphPad Prism version 5.0 for Mac OS X (GraphPad Software, La Jolla California). Patients were divided into low versus high expression groups based on the data published in the original article to the best of our knowledge.

Results

Discrepancy between PDAC-TCGA and SEER's data on pancreatic ductal adenocarcinoma statistics

According to the National Cancer Database, 55% of pancreatic cancer patients between 1992 and 1998 were diagnosed in stage IV (Fig. 1A; ref. 8). However, in the PDAC-TCGA dataset, about two-thirds of patients were diagnosed in stage IIB, with less than 3% of patients being in stage IV (Fig. 1B). The SEER registry also stratifies patients diagnosed with PDAC between 2007 and 2013 based on the spread of disease into localized, regional, or distant groups. Comparable with the National Cancer Database statistics from the previous decade, more than half of the patients in the SEER database were diagnosed with distant disease (Fig. 1C). Most PDAC-TCGA patients had localized or regionally spread disease (25% and 72%, respectively) in comparison with 52% of patients diagnosed with distant disease in the SEER database (Fig. 1C and D). Naturally, the 5-year survival rate of the PDAC population depends on the disease stage at the time of diagnosis (Fig. 1E). Importantly, patients in the PDAC-TCGA database have an unusually high 5-year survival of around 23%, and a median survival of approximately 1 year and 8 months (Fig. 1F). This

Table 1. Sample classification in PDAC subset of TCGA

(A) Normal and pseudonormal samples			(B) Non-PDAC samples		
Adjacent normal pancreas	TCGA-H6-8124-11 TCGA-L1-A7W4-01	TCGA-YB-A89D-11	Did not arise from pancreas	TCGA-FB-A7DR-01 TCGA-HZ-7289-01	TCGA-HZ-8638-01
Pseudonormal, <1% neoplastic cellularity	TCGA-F2-6880-01 TCGA-F2-7273-01 TCGA-F2-7276-01 TCGA-H8-A6C1-01 TCGA-HZ-7920-01 TCGA-HZ-7923-01	TCGA-HZ-7924-01 TCGA-IB-AAUV-01 TCGA-IB-AAUW-01 TCGA-RL-AAAS-01 TCGA-US-A77J-01	Neuroendocrine	TCGA-2L-AAQM-01 TCGA-3A-A9IJ-01 TCGA-3A-A9IL-01 TCGA-3A-A9IN-01	TCGA-3A-A9IO-01 TCGA-3A-A9IR-01 TCGA-3A-A9IS-01 TCGA-3A-A9IV-01
			Acinar cell carcinoma	TCGA-HZ-7918-01	
Solid tissue normal	TCGA-H6-A45N-11	TCGA-HV-A5A3-11	Intraductal papillary mucinous neoplasm Metastatic	TCGA-FB-AAPP-01 TCGA-HZ-A9TJ-06	TCGA-HV-A7OP-01
			Undifferentiated	TCGA-2J-AABP-01	
			Systemic treatment given to the prior/other malignancy	TCGA-IB-7654-01	

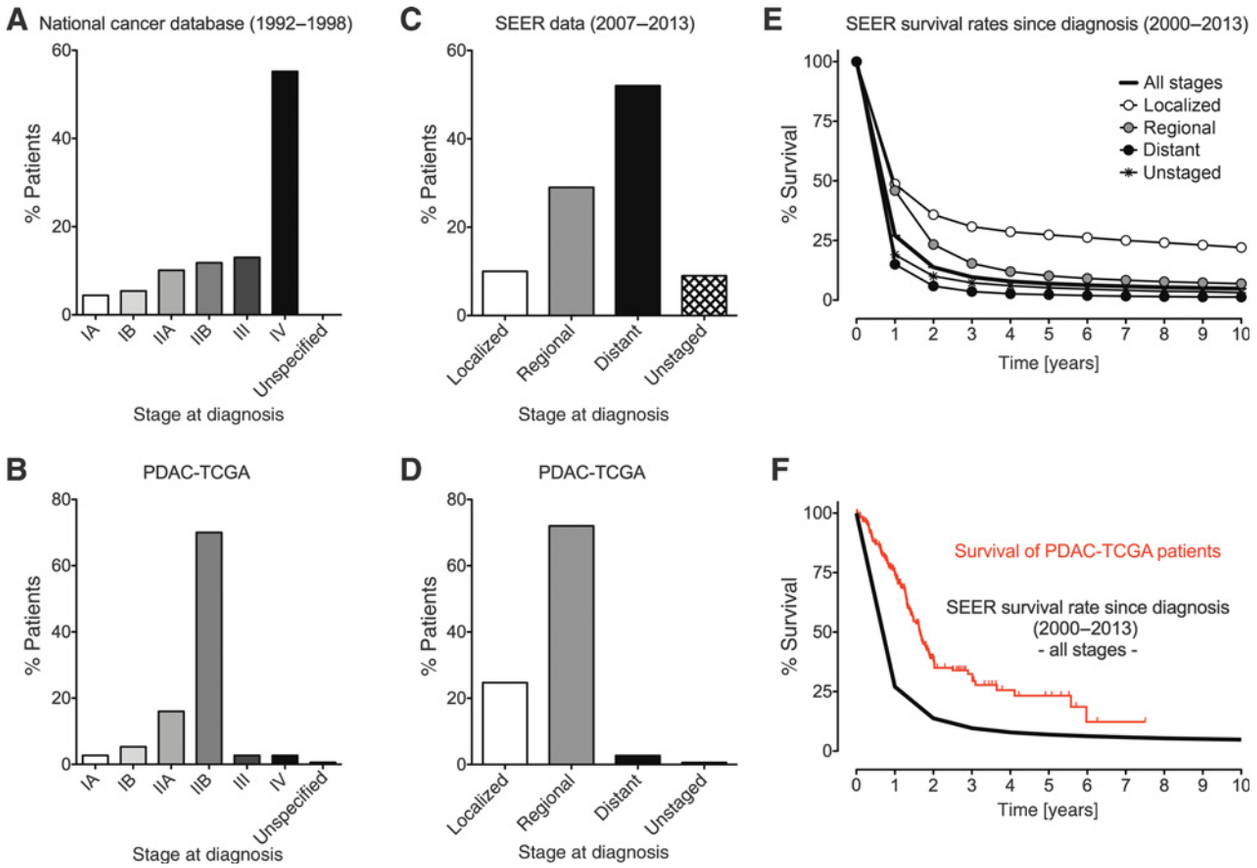

Figure 1.

Characteristics of the PDAC-TCGA dataset compared with the overall population of pancreatic cancer patients. Distribution of patients according to the disease stage at the time of diagnosis based on the data collected by National Cancer Database between 1992 and 1998 (**A**) and SEER data collected between 2007 and 2013 (**C**). For comparison, distribution of patient population from PDAC-TCGA database by disease stage is shown in **B** and **D**. **E,** SEER survival rates since diagnosis based on the disease stage at diagnosis. SEER data on pancreatic cancer represented here were collected between 2000 and 2013. **F,** Comparison of survival rate of overall pancreatic cancer patient population (SEER data collected between 2000 and 2013) with the survival rate of all patients included in the PDAC-TCGA dataset.

is in contrast to the literature and pancreatic cancer statistics, where 5-year survival rate is about 8%, with the overall median survival between 6 and 8 months (Fig. 1F; refs. 8, 9). This difference could confound conclusions that apply the TCGA survival data to the general PDAC population. A similar discrepancy from the SEER registry in cancer stage distribution at the time of diagnosis and median survival was observed in several other cancer subsets of TCGA, such as lung, stomach and liver cancers (Supplementary Fig. S1).

The differences between the TCGA cohort and the overall PDAC population may be explained by the limited availability of tumor tissue. Surgery, and tissue collection, is only possible in approximately 20% of patients whose cancer is confined in one area, does not involve major blood vessels and did not metastasize to distant secondary sites (8). Thus, the availability of pancreatic tissue samples for analysis by the TCGA consortium was necessarily limited to patients with resectable cancer, which is usually in its earlier stages and imposes a bias toward localized/early-stage tumors, even though most patients are diagnosed with the late/metastatic stage of the disease.

PDAC-TCGA is comprised of samples originating from different cell types as well as (pseudo)-normal tissues

Hierarchical clustering of quantified RNAseq gene expression data from the entire patient population included in the PDAC-TCGA, identified a cluster of samples showing a divergent pattern of gene expression (Fig. 2A). Examination of the histologic annotations revealed subtypes other than PDAC or tumors that metastasized from other tissue. Eight of these patients had neuroendocrine tumors, a cancer with a significantly higher overall 5-year survival rate of 42% (10). The PDAC-TCGA database also includes three samples that did not arise from pancreas, one acinar cell carcinoma, two intraductal papillary mucinous neoplasms, a metastatic cancer, an undifferentiated cancer, and a sample that may be excluded because of systemic treatment given for a prior/other malignancy (Table 1). In addition, a number of samples are categorized as pseudo-normal due to less than 1% of neoplastic cellularity (Table 1). In a recent TCGA research network publication using the PDAC dataset, the consortium used 150 pancreatic samples that were indeed classified as PDAC (11). When limiting analysis to include only

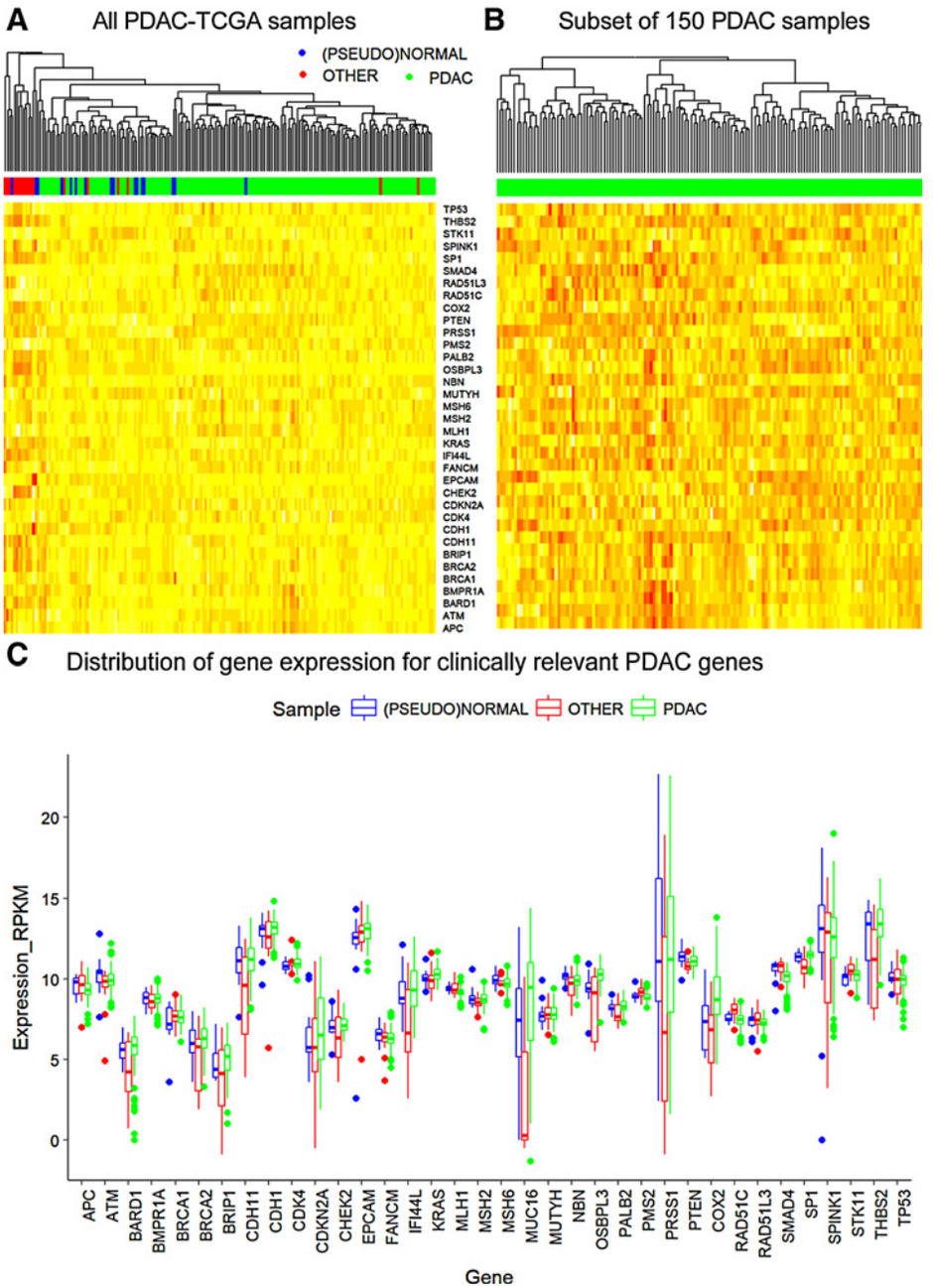

Figure 2.
Heatmaps of RNAseq gene expression data from PDAC-TCGA. Hierarchical clustering of the entire PDAC-TCGA dataset (**A**) based on their gene expression and after curating samples for 150 true pancreatic ductal adenocarcinomas (**B**). **C,** Distribution of gene expression across (pseudo)-normal versus other versus PDAC samples of PDAC-TCGA subset.

the curated dataset, the global picture of gene expression is reorganized (Fig. 2B), and highlights how the samples derived from other cell origins skew the distribution of the quantified gene expression across the sample population. Many of the genes associated with clinical outcomes in PDAC are differentially expressed in non-PDAC samples (Fig. 2C). When they are removed from the analysis, the expression profiles are more evenly distributed across the cohort.

The PDAC-TCGA sample composition affects hypotheses and study conclusions

Failure to exclude (pseudo)-normal and non-PDAC samples from the dataset translates to skewed survival curves that segregate the population based on gene expression signatures. Comparing the Kaplan–Meier plots of patients segregated by the relative mRNA expression to the source of the tissue sample reveals that using an uncurated PDAC-TCGA dataset can lead to false associations with disease progression and survival. For genes commonly associated with PDAC, normalized gene expression ($\log_2$ fold change relative to the median) shows that the extremes of the distribution often contain non-PDAC samples (Fig. 3). The associated survival curves, plotted on the basis of high versus low gene expression, are highlighted in several published examples where statistical significance was lost when using the curated PDAC-TCGA dataset. In one example, high expression of COX-2 (PTGS2) and Sp1 was found to negatively influence the

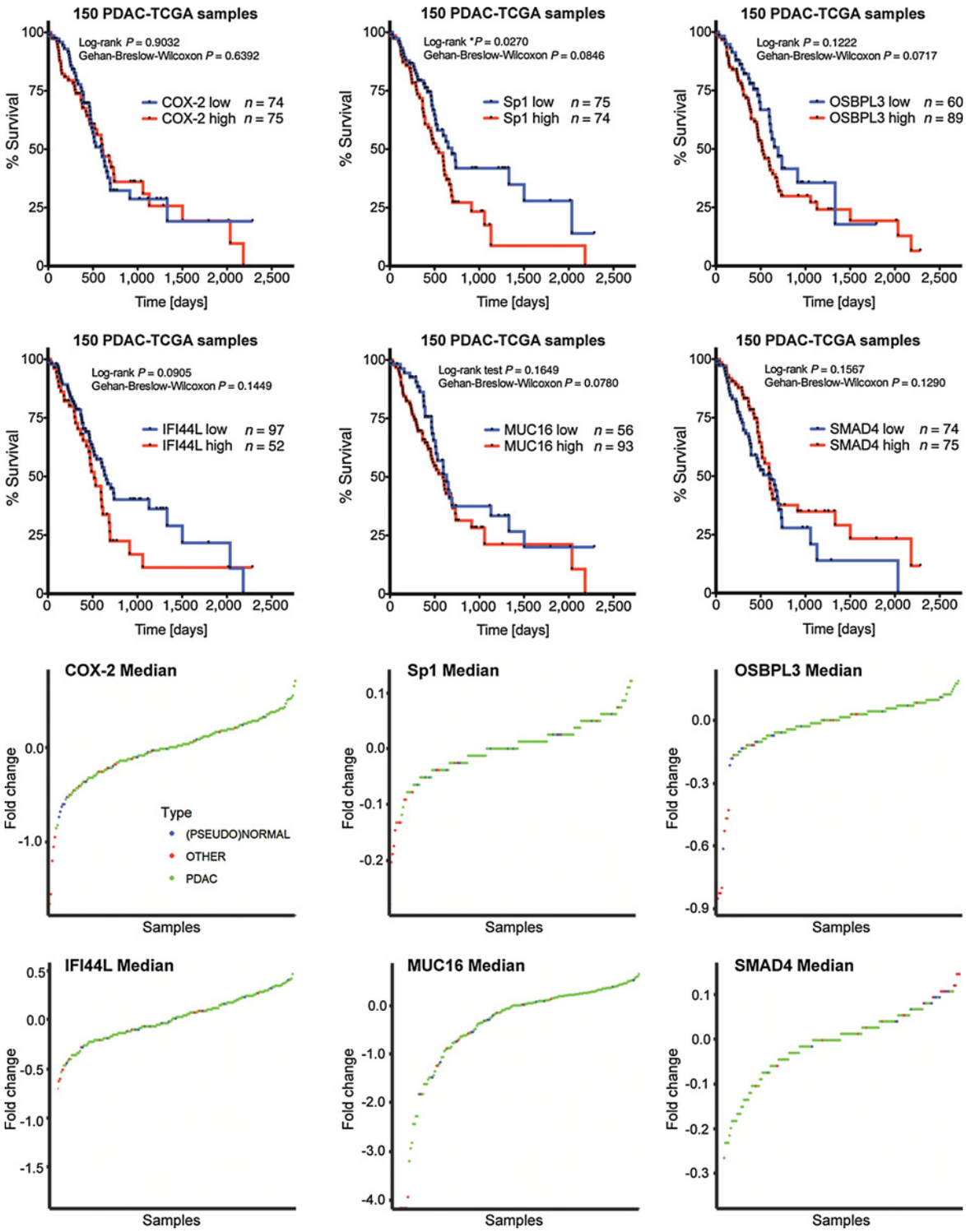

Figure 3.

Examples of the impact of sample origin on associations between gene expression level and survival. Kaplan–Meier plots for COX-2, Sp1, and SMAD4 were divided into low versus high groups based on the median gene expression of 150 PDAC-TCGA samples; MUC16 —low versus high groups based on the mean gene expression of 150 PDAC-TCGA; OSBPL3 and IFI44L —low versus high groups based on the cutoff expression level as in the original paper (top). All 150 PDAC-TCGA samples had available gene expression data; however, one sample had no available survival data. Distribution of gene expression level according to the sample origin (bottom).

survival (12). While high levels of Sp1 remained significantly associated with poor survival, there was no difference in survival based on COX-2 expression when using the curated 150 PDAC samples from TCGA dataset (Fig. 3). Careful reexamination is especially needed when the results of integrative bioinformatic analysis predict novel biomarkers, such as OSBPL3 and IFI44L (13). After reanalysis using only 150 PDAC-TCGA samples, there was no statistically significant difference between low and high expressors of OSBPL3 or IFI44L with regards to survival; hence based on the PDAC-TCGA cohort these are not good predictive survival biomarkers (Fig. 3). Liang and colleagues showed that high expression of MUC16 (CA125) was associated with shorter survival time in PDAC-TCGA cohort (14). However, the correlation was lost when we analyzed the survival plot on 150 PDAC-TCGA samples (Fig. 3). In contrast, higher serum CA125 protein levels and MUC16 tissue expression was associated with poor survival in a different cohort described by Liang and colleagues (14). This discrepancy may result from differences in assay/detection method used for quantification of MUC16/CA125. Another study depicts correlation between high SMAD4 expression and better outcome in PDAC patients (15). When the Kaplan–Meier plot was repeated on the 150 PDAC only samples, this correlation between low SMAD4 expression and poor survival was lost (Fig. 3).

Besides the publications highlighted here, there are dozens of others that used the PDAC-TCGA database and proposed new PDAC biomarkers or new mechanisms essential for cancer progression and survival. In some of them, researchers were aware of the pitfalls in the sample collection and excluded non-PDAC samples, but in many others, investigators did not properly curate the sample pool. Once again, we urge researchers to scrutinize the source of samples before making any associations with clinical outcome.

Discussion

Web-based analysis tools provide access to the wealth of information contained in publically available datasets like TCGA. They are especially useful to researchers lacking the computational background to exploit these resources on their own. When validating preclinical data in a patient setting, or expanding laboratory observations to the datasets from large consortium efforts, using all of the available data without attention to individual sample characteristics can lead to false conclusions. It is the responsibility of the investigator to ensure the underlying data fits their assumptions based on the curated patient population. In some instances, web-tools such as "OncoLnc.org" provide Kaplan–Meier survival plots and statistics based on the gene of interest, but do not display other characteristics of the samples besides coded sample ID. Others, such as Morpheus or Xenabrowser, provide various outputs and allow you to curate the molecular analysis using clinical information. However, the process is not straightforward and requires an awareness of the importance of cleaning the data before use. With the PDAC-TCGA data as an example, we illustrate the impact that sample selection can have on identifying survival-associated gene signatures or biomarkers, and hope to drive a more rigorous and responsible analysis of publicly available datasets.

Comparison between the SEER registry and the TCGA shows that some other TCGA datasets are also enriched in early stage cancer samples and produce divergent survival plots. Lung adenocarcinoma, stomach adenocarcinoma, and liver hepatocellular carcinoma subsets of the TCGA all show significant differences in cancer stage distribution at the time of diagnosis and median survival, compared with the SEER registry (Supplementary Fig. S1). In contrast, survival of patients diagnosed with cervical squamous cell carcinoma and endocervical adenocarcinoma, rectum adenocarcinoma or ovarian serous cystadenocarcinoma from TCGA, follows survival observed by SEER registry (Supplementary Fig. S1). Knowing how closely a dataset represents the disease in the larger population provides important context when extending the results of bioinformatic analysis to future translational research.

Differentiating between similar tumor subtypes is another important consideration, as demonstrated by TCGA separation of brain lower grade glioma from glioblastoma multiforme dataset (https://tcga-data.nci.nih.gov/docs/publications/tcga/). However, segregating by subtype is only one consideration when curating the underlying dataset. A systematic analysis of 21 cancer subtypes in TCGA demonstrated how controlling for tumor purity also altered the results of bioinformatic analysis (16). Clearly, understanding the context of the sample origin is critically important to the interpretation of molecular datasets. This is especially true when considering how the results drawn from bioinformatic analyses inform hypotheses about the contribution of biomolecular markers on prognosis and survival in the general population. Careful examination of the biological context of publically available samples should extend beyond the example presented here, and the lessons should be applied to analyses of other molecular databases. In the case where the histology and origin of the sample data are not available, the resulting analysis should be met with a healthy dose of skepticism and conclusions should be supported with independent lines of experimental evidence.

The PDAC-TCGA database contains (pseudo)-normal tissues and tumor samples other than pancreatic ductal adenocarcinoma, which affects the conclusions that support outcomes of validation studies and biomarker discovery. In addition to the curation of individual samples used for bioinformatic analysis, broader features of the dataset must also be considered when making associations to clinical outcomes. The PDAC-TCGA dataset is comprised of mainly resectable cancer samples, diagnosed in stage IIB, which is reflected in the greater median and 5-year survival rates. Therefore, this dataset does not reproduce the larger clinical population of PDAC patients, which has one of the poorest 5-year survival rates among all cancer types of only 8% (9). However, this discrepancy does not negate the usefulness of the resource, especially for identifying signatures related to early detection. This context must be included in the interpretation of the bioinformatic analysis. It is also the responsibility of those maintaining data repositories and analysis tools to include extensive sample metadata and educate their users on identifying and controlling for the variability of the underlying samples.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Authors' Contributions

Conception and design: I. Peran, S. Madhavan, S.W. Byers, M.D. McCoy
Development of methodology: I. Peran, M.D. McCoy

Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): M.D. McCoy
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): I. Peran, M.D. McCoy
Writing, review, and/or revision of the manuscript: I. Peran, S. Madhavan, S.W. Byers, M.D. McCoy
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): M.D. McCoy
Study supervision: S. Madhavan, S.W. Byers, M.D. McCoy

Acknowledgments

This work was supported by the 2017 AACR-AstraZeneca Fellowship in Immuno-oncology Research, grant number 17-40-12-PERA (to I. Peran); The Ruesch Center for the Cure of Gastrointestinal Cancers grant award (to I. Peran); and NIH/NHGRI 3U41HG007822-02S1 (to M.D. McCoy and S. Madhavan) and R01 CA170653 (to S.W. Byers).

Received January 24, 2018; revised April 4, 2018; accepted May 3, 2018; published first May 8, 2018.

References

1. Collins FS, Morgan M, Patrinos A. The Human Genome Project: lessons from large-scale biology. Science 2003;300:286–90.
2. Collins FS, Green ED, Guttmacher AE, Guyer MS, US National Human Genome Research Institute. A vision for the future of genomics research. Nature 2003;422:835–47.
3. The Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 2013;45:1113–20.
4. Hruban RH, Goggins M, Parsons J, Kern SE. Progression model for pancreatic cancer. Clin Cancer Res 2000;6:2969–72.
5. Bardeesy N, DePinho RA. Pancreatic cancer biology and genetics. Nat Rev Cancer 2002;2:897–909.
6. Petersen GM. Familial pancreatic cancer. Semin Oncol 2016;43: 548–53.
7. Chaffee KG, Oberg AL, McWilliams RR, Majithia N, Allen BA, Kidd J, et al. Prevalence of germ-line mutations in cancer genes among pancreatic cancer patients with a positive family history. Genet Med 2018;20:119–27.
8. Bilimoria KY, Bentrem DJ, Ko CY, Ritchey J, Stewart AK, Winchester DP, et al. Validation of the 6th edition AJCC pancreatic cancer staging system. Cancer 2007;110:738–44.
9. American Cancer Society. Cancer facts & figures 2017. Atlanta, GA: American Cancer Society; 2017.
10. Ries LAG, Young JL Jr, Keel GE, Eisner MP, Lin YD, Horner M-JD, editors. Cancer survival among adults: U.S. SEER Program, 1988–2001 [Internet]. Bethesda, MD: National Cancer Institute; 2007. p. 1–286. Available from: http://www.seer.cancer.gov.
11. The Cancer Genome Atlas Research Network. Integrated genomic characterization of pancreatic ductal adenocarcinoma. Cancer Cell 2017;32: 185–203.
12. Hu H, Han T, Zhuo M, Wu L-L, Yuan C, Wu L, et al. Elevated COX-2 expression promotes angiogenesis through EGFR/p38-MAPK/Sp1-dependent signalling in pancreatic cancer. Sci Rep 2017;7:470.
13. Li H, Wang X, Fang Y, Huo Z, Lu X, Zhan X, et al. Integrated expression profiles analysis reveals novel predictive biomarker in pancreatic ductal adenocarcinoma. Oncotarget 2017;8:52571–83.
14. Liang C, Qin Y, Zhang B, Ji S, Shi S, Xu W, et al. Oncogenic KRAS targets MUC16/CA125 in pancreatic ductal adenocarcinoma. Mol Cancer Res 2017;15:201–12.
15. Pickup MW, Owens P, Gorska AE, Chytil A, Ye F, Shi C, et al. Development of aggressive pancreatic ductal adenocarcinomas depends on granulocyte colony stimulating factor secretion in carcinoma cells. Cancer Immunol Res 2017;5:718–29.
16. Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. Nat Commun 2015;6:8971.