DIFFERENTIAL GENE EXPRESSION ANALYSIS IN MALIGNANT CANCER TISSUES

Synopsis of the PhD thesis

Áron Bartha, MD

Doctoral School of Pathological Sciences

Semmelweis University

Supervisor: Balázs Győrffy MD DSc

Official reviewers: Tímea Tőkés MD PhD

Sándor Spisák PhD

Head of the Complex Examination Committee:

Péter Nyírády MD DSc

Members of the Complex Examination Committee:

Balázs Szalay MD PhD

Budapest

2023

INTRODUCTION

Cancer arises from normal cells that mutate into pre-cancerous and eventually malignant cells due to genetic or epigenetic lesions. These lesions are primarily caused by external mutagenic factors, but hereditary mutations also contribute to cancer development. Genetic changes result in altered gene expressions within tumor cells, driving the cancerous phenotype. While most genes show similar expression profiles in cancerous and normal tissues, differentially expressed genes can serve as targets for treatment or molecular biomarkers of cancer progression. Targeting genes with increased expression of specific products has demonstrated significant clinical benefits, exemplified by the selective inhibition of overexpressed tyrosine kinases. Gene expression changes in cancer cells are associated with a limited set of distinct characteristics known as cancer hallmarks. These include sustained proliferative signaling, evading growth suppressors, resistance to cell death, replicative immortality, induction of vasculature, activation of invasion and metastasis, reprogramming cellular metabolism, and avoidance of immune destruction. As the second leading cause of death globally, identifying potential predictive and prognostic biomarkers for cancer is of utmost importance. Frequent cancers include breast, lung, colon, prostate, pancreas, and kidney carcinomas. Clear cell renal carcinoma (ccRCC) is the most common form of kidney cancer, with over 80,000 estimated cases in the United States in 2023. Mortality rates for ccRCC have plateaued in recent years, with better survival outcomes observed in patients with early-stage and no distant metastasis. For patients with advanced ccRCC, systemic therapies involving mTOR inhibitors, VEGF inhibitors, and checkpoint inhibitors (e.g., nivolumab, avelumab, pembrolizumab, ipilimumab, and interleukin 2 therapy) are used. Developing a protein abundance-based gene panel specific to ccRCC could greatly support clinical diagnostics and therapeutic decision-making. In conclusion, understanding the genetic and expression changes in cancer cells and identifying specific biomarkers for ccRCC are critical steps towards improved cancer diagnosis and treatment strategies, with the potential to enhance patient outcomes.

OBJECTIVES

1. My aim was to create an integrated database of a significant number of samples with transcriptome-level data.
    a. With the utilization of both gene chip and RNA-Seq based datasets, my goal was to establish a comprehensive set of malignant and normal samples from both adult and pediatric patients.
    b. My second objective was to investigate the difference between malignant and normal tissues.
    c. My third objective was to assess the database's robustness by employing a training-test approach to identify genes exhibiting differential expression in specific tumor types.
    d. Finally, I further aimed to establish an online analysis portal which enables the comparison of gene expression changes across all genes and multiple platforms by mining the entire integrated database.
2. My second main aim was to identify potentially clinically relevant biomarkers of ccRCC to help diagnostic and therapeutic decision-making process.

a. An important first objective was to leverage a significant volume of transcriptomic and protein data for the purpose of identifying proteins that demonstrate elevated expression in ccRCC.

b. Then, by using data form patients treated at Semmelweis University with available proteotranscriptomic and clinical data I aimed to investigate the abundance of expressed proteins and the effect of these proteins on survival.

c. By specifically focusing on markers with higher expression in tumor tissues using a machine learning approach, I sought to increase the specificity of my analysis to solidify future clinical application of the results.

METHODS

We utilized data generated by gene arrays obtained from the Gene Expression Omnibus of the National Center for Biotechnology Information (NCBI-GEO) and RNA-seq data from The Cancer Genome Atlas (TCGA), Therapeutically Applicable Research to Generate Effective Treatments (TARGET), and The Genotype-Tissue Expression (GTEx) repositories. The altered expression within each platform was analyzed separately. Statistical significance was determined using either Mann-Whitney or Kruskal-Wallis tests, while the False Discovery Rate (FDR) was calculated using the Benjamini-Hochberg method. In total, the database comprises 56,938 samples, which includes 33,520 samples from 3180 gene chip-based studies (453 metastatic, 29,376 tumorous, and 3691 normal samples), 11,010 samples from TCGA (394 metastatic, 9886 tumorous, and 730 normal samples), 1193 samples from TARGET (1 metastatic, 1180 tumorous, and 12 normal samples), and 11,215 normal samples from GTEx.

Through the utilization of in silico discovery datasets containing paired normal tissue samples from gene array and RNA-Seq data repositories, we identified the top over-expressed genes in ccRCC (clear cell renal carcinoma). To validate these findings, we obtained surgically resected ccRCC specimens from Semmelweis University and performed RNA sequencing on matched tissue sample pairs from pathologically confirmed ccRCC patients. The resulting differential expression was further assessed at the protein level using targeted mass spectrometry (MS). Subsequently, we established a support vector machine-based classification algorithm based on the protein-level data. To expand our analysis, we constructed a database comprising 558 renal tissue samples from NCBI GEO and examined them to identify the most prominently expressed genes in ccRCC. Additionally, for the protein level analysis, we obtained 162 samples each of malignant and normal kidney tissues.

RESULTS

Integrated database

In total, the entire database holds 56,938 samples, including both RNA-seq and gene array samples. These include, after pre-processing, 33,520 unique gene array samples from 38 tissue types, including 3,691 normal, 29,376 tumorous and 453 metastatic samples. For each of these samples, the mRNA expression of 12,210 genes is available. Included RNA-seq data comprise three different platforms. After curation, normalization steps and data processing, we collected data of 11,010 samples, including 730 normal, 9,886 cancerous and 394 metastatic specimens from adult cancer patients. We also added 1,193 pediatric related data from GDC, consisting of 12 normal, 1,180 cancerous, and one metastatic samples. In order to increase the number of normal samples, we further included 11,215 RNA-Seq GTEx data from non-cancerous persons.

TNMplot.com analysis platform

We established a web application to enable a real-time comparison of gene expression changes between tumor, normal and metastatic tissues amongst different types of platforms across all genes. The portal can be accessed at www.tnmplot.com and has several analysis options. The pan-cancer analysis tool compares normal and tumorous samples across 22 tissue types simultaneously. This RNA-seq-based rapid analysis serves as explanatory data to furnish comparative information for a selected gene.

The second approach directly compares tumor and normal samples by either grouping all specimens of the same category and running a Mann–Whitney U test or—in the case of the availability of paired normal and adjacent tumors—by running a paired Wilcoxon statistical test. The results are visualized by both boxplots and violin plots. We have also implemented a graphical representation of sensitivity and specificity: a diagram provides the percentage of tumor samples that show higher expression of the selected gene than normal samples at each major cutoff value. While the number of metastatic samples is generally limited, there are sufficient specimens available in the RNA-seq and gene array databases for five and twelve tissue types, respectively. The third feature of the analysis platform allows us to simultaneously compare these tumor, normal and metastatic data using a Kruskal–Wallis test and the Dunn post-hoc test.

Sensitivity and specificity

Whenever a new biomarker is developed, the two most crucial pieces of information include sensitivity (the proportion of tumors which have higher expression than normal at a given cutoff) and specificity (the proportion of tumors divided by the total sum of all tumors and normal over the given cutoff). The online analysis interface provides a graphical representation of sensitivity and specificity at the major cutoff values (minimum, Q1, median, Q3, and maximum).

Linking the most significant genes to cancer hallmarks

We performed gene ontology analysis on the 55 genes shared by all cancer types in both RNA-Seq and gene array studies. Most enriched biological processes in which these genes might be involved resulted in mainly terms which participate in cell proliferation. We further linked the best 55 genes common across all cancer types in both platforms to the cancer hallmarks based on their functions available in Entrez Gene Summary, GeneCards Summary, and UniProtKB/Swiss-Prot Summary. The majority of the genes (*n* = 21) were linked to sustained proliferative signaling. The second most common hallmark was the deregulation of cellular energetics (*n* = 13). Activation of invasion and metastasis (*n* = 5), enabling replicative immortality (*n* = 8), and avoiding immune destruction (*n* = 5) were also represented by multiple genes. Only single genes were linked to genome instability and mutation, evasion of growth suppressors, and tumor-promoting inflammation.

Genes over-expressed in ccRCC

We uncovered differentially expressed genes between paired ccRCC and adjacent normal tissues using gene chip data from NCBI-GEO and RNA-Seq data from GDC-TCGA. IGFBP3 was found to be the most upregulated gene in tumor tissues confirmed to both platforms ($FC_{\text{gene chip}}$ = 8.15, p = 1.01E-33 and $FC_{\text{RNA-Seq}}$ = 10.47, p = 2.17E-12). The most significant genes include previously established molecular targets like VEGFA ($FC_{\text{gene chip}}$ = 3.02 p = 3.1E-32 and $FC_{\text{RNA-Seq}}$ = 9.03, p = 3E-13) and CCND1 ($FC_{\text{gene chip}}$ = 4.12, p = 2.3E-32 and $FC_{\text{RNA-Seq}}$ = 5.98, p = 4.25E-13). PLIN2 is a further differentially expressed gene that showed comparable results in both array and sequencing studies with FC $_{\text{gene chip}}$ = 3.85, p = 1.59E-32, and FC $_{\text{RNA-Seq}}$ = 7.08, p = 1.1E-11 respectively.

Gene expression analysis of Semmelweis cohort

The Semmelweis cohort includes 162 samples from 81 patients. In the RNAseq analysis, we examined 32 normal and 57 tumor samples with an average sequencing yield of 7.5 million reads per sample. In these, we confirmed differential expression for 29 genes. Top differentially expressed genes like VEGFA (FC = 32, p = 1.77E-11) IGFBP3 (FC = 1.56.1, p = 6.24E-09), PFKP (FC = 13.81, p = 4.59E-09), PLIN2 (FC = 46.5, p = 2.82E-11) showed comparable results with the GDC and GEO datasets.

Proteomic analysis of Semmelweis cohort

Proteomic analysis was performed using 162 normal and malignant tissue samples. Of the complete list of the 30 selected genes from GDC and GEO results, we were able to successfully measure 22 in the targeted LC-MS/MS. Top differentially expressed genes include PLIN2 (FC = 26.09, p = 3.9E−39), PLOD2 (FC = 15.84, p = 6.51E−36), PFKP (FC = 12.78, p = 1.01E−47), IGFBP3 (FC = 3.04, p

= 7.53E−18), CCND1(FC = 7.9, p = 1.04E−24) and VEGFA (FC = 3.5, p = 1.4E−22).

Using 88 tissue samples with simultaneously available RNA-Seq and MS data we performed a correlation analysis to assess the link between RNA expression and protein expression values. Fourteen genes had a significant correlation between protein and RNA data, with a mean coefficient of 0.51.

Survival Analysis Using Proteome-Level Data

To estimate the potential effects of protein expression on patient survival, we performed a survival analysis using all available proteins. Five out of the investigated proteins showed a correlation with survival. Patients with elevated expression of PLOD2 protein showed significantly worse overall survival compared to subjects with lower expression (p = 2.42E−7, HR = 5.03). Overexpression of further proteins such as TIMP1 (p < 3E−2, HR = 4.71), VIM (p < 3E−2, HR = 2.49), LGALS1 (p < 3E−2, HR = 2.47), and P4HA1 p < 3E−2, HR = 2.6) also showed significant correlation with impaired overall survival.

ccRCC specific model creation

MS based protein abundance data of the twenty selected proteins in the 162 patient samples were used for establishing the most robust classification algorithm. We investigated multiple machine learning methods (including k-nearest neighbors, random forest, logistic regression, and support vector machines) to build a model which is capable to differentiate between normal and malignant kidney tissues. For the proper estimation of the optimal gene panel, we performed recursive feature elimination. Of the four methods, SVM delivered the best performance in both test and training cohorts using nine proteins as input. SVM was able to identify tumor tissues from MS quantification data with a classification accuracy of 0.98 in the test set (Kappa = 0.95, sensitivity = 0.95, specificity = 1).

CONCLUSIONS

1. Creation of an integrated database of a significant number of samples with transcriptome-level data

a. I have established an integrated database with nearly 57,000 samples of gene expression data suitable for the identification of differentially expressed genes.

b. In my database I included 3,691 normal, 29,376 primary tumor, and 453 metastatic tissues from gene chip based datasets.

c. I further included RNA-seq based datasets, comprising 11,957 normal tissues, 11,066 primary tumor tissues, and 395 metastatic tissues.

d. I have incorporated 1,193 pediatric and more than 55,000 adult samples from both RNA-Seq and gene chip based datasets.

e. Investigating the top ten cancers with the highest mortality rates I identified aberrantly functioning pathways related to cell proliferation as the main characteristics of malignant cells.

f. I validated the genes showing differential expression in specific tumor types using a training-test approach, thereby enhancing the established database's robustness.

    g. I created an online analysis portal that compares gene expression changes across all genes and multiple platforms.

2. Identification of potentially clinically relevant biomarkers of ccRCC to help diagnostic and therapeutic decision-making process.

    a. By the combination of transcriptomic data and targeted mass spectrometry I identified the top 22 differentially expressed proteins in ccRCC.

    b. Using the proteotranscriptomic data of 162 samples from Semmelweis University I identified the effect of this protein panel on patient survival.

    c. Using a support vector machine-based machine learning approach I identified the top nine proteins with higher expression in tumor tissues. These proteins could serve as promising biomarkers in the clinical setting.

SUMMARY

Genes showing higher expression in either tumor or metastatic tissues can help in better understanding tumor formation and can serve as biomarkers of progression or as potential therapy targets. Our goal was to establish an integrated database using available transcriptome-level datasets and to create a web platform which enables the mining of this database by comparing normal, tumor and metastatic data across all genes in real time. We utilized data generated by either gene arrays from NCBI-GEO or RNA-seq from TCGA, TARGET, and GTEx repositories. The entire database contains 56,938 samples, including 33,520 samples from 3180 gene chip-based studies 11,010 samples from TCGA, 1193 samples from TARGET and 11,215 samples from GTEx. The most consistently upregulated gene across multiple tumor types was TOP2A, the most consistently downregulated gene was ADH1B. Validation of differential expression using equally sized training and test sets confirmed the reliability of the database in breast, colon, and lung cancer. The online analysis platform enables unrestricted mining of the database and is accessible at TNMplot.com. We further

used a proteotranscriptomic approach to differentiate normal and tumor tissues in ccRCC. Using transcriptomic data of patients with malignant and paired normal tissue samples from gene array and RNA-Seq cohorts, we identified the top genes over-expressed in ccRCC. We collected surgically resected ccRCC specimens to further investigate the transcriptomic results on the proteome level. The differential protein abundance was evaluated using targeted mass spectrometry. We assembled a database of 558 renal tissue samples and used these to uncover the top genes with higher expression in ccRCC. We further investigated the expression of the identified genes on our patient data both on the transcriptomic and protein level using 162 tissue samples. The most consistently upregulated proteins were IGFBP3, PLIN2, PLOD2, PFKP, VEGFA, and CCND1. We also identified those proteins which correlate with overall survival. Finally, a support vector machine-based classification algorithm using the protein-level data was set up. We used transcriptomic and proteomic data to identify a minimal panel of proteins highly specific for clear cell renal carcinoma tissues. The introduced gene panel could be used as a promising tool in the clinical setting.

BIBLIOGRAPHY OF THE CANDIDATE’S PUBLICATIONS

Publications related to this dissertation.

Bartha A, Gyorffy B. *TNMplot.com: A Web Tool for the Comparison of Gene Expression in Normal, Tumor and Metastatic Tissues*. International journal of molecular sciences. 2021;22(5). IF: 6.2

Bartha A, Darula Z, Munkacsy G, Klement E, Nyirady P, Gyorffy B. *Proteotranscriptomic Discrimination of Tumor and Normal Tissues in Renal Cell Carcinoma.* International journal of molecular sciences. 2023;24(5).

Publications not included in the dissertation:

Á. Bartha; Győrffy, B. *Comprehensive Outline of Whole Exome Sequencing Data Analysis Tools Available in Clinical Oncology*. Cancers IF: 6.1

Antal Jobbágy, Norbert Kiss, Fanni Adél Meznerics, Klára Farkas, Dóra Plázár, Szabolcs Bozsányi, Luca Fésűs,

Áron Bartha, Endre Szabó, Kende Lőrincz, Miklós Sárdy, Norbert Miklós Wikonkál, Péter Szoldán, András Bánvölgyi: *Emergency Use and Efficacy of an Asynchronous Teledermatology System as a Novel Tool for Early Diagnosis of Skin Cancer during the First Wave of COVID-19 Pandemic*; International Journal of Environmental Research and Public Health

Szabolcs Bozsányi, Noémi Nóra Varga, Klára Farkas, András Bánvölgyi, Kende Lőrincz, Ilze Lihacova, Alexey Lihachev, Emilija Vija Plorina, **Áron Bartha**, Antal Jobbágy, Enikő Kuroli, György Paragh, Péter Holló, Márta Medvecz, Norbert Kiss, Norbert M Wikonkál: *Multispectral Imaging Algorithm Predicts Breslow Thickness of Melanoma*; Journal of Clinical Medicine IF: 3.9

Csaba Miskey, Lacramioara Botezatu, Nuri A Temiz, Andreas Gogol-Doring, **Aron Bartha**, Balazs Gyorffy, David A Largaespada, Zoltan Ivics, Attila Sebe: *In Vitro Insertional Mutagenesis Screen Identifies Novel Genes Driving Breast Cancer Metastasis*; Molecular Cancer Research IF: 5.2