SEMMELWEIS EGYETEM

DOKTORI ISKOLA

# Ph.D. értekezések

## 2934.


# BARTHA ÁRON


**Onkológia**
című program

Programvezető: Dr. Bödör Csaba, egyetemi tanár

Témavezető: Dr. Győrffy Balázs, egyetemi tanár

# DIFFERENTIAL GENE EXPRESSION ANALYSIS IN MALIGNANT CANCER TISSUES

**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

**TABLE OF CONTENTS**

## LIST OF ABBREVIATIONS

BCA- Bicinchoninic Acid Assay

BP – Biological Process

ccRCC- Clear Cell Renal Carcinoma

CPTAC - Clinical Proteomic Tumor Analysis Consortium

DTT - Dithiothreitol

ECM - Extracellular Matrix

EMT - Epithelial-To-Mesenchymal Transition

ESI - Electrospray Ionization

FC – Fold-Change

FDA – U.S. Food and Drug Administration

FDR – False Discovery Rate

GDC – Genomic Data Commons

GO – Gene Ontology

GTEx - Genotype-Tissue Expression repository

HPLC - High Performance Liquid Chromatography

HR – Hazard Ratio

HTS – High-Throughput Sequencing

KM – Kaplan-Meier

KNN – K-Nearest Neighbors

KRAS - Kirsten ras oncogene

LC-MS/MS - Liquid Chromatography Coupled Tandem Mass Spectrometry

LOGIT - Logistic Regression

MALDI - Matrix-Assisted Laser Desorption/Ionization

MS – Mass Spectrometry

MW – Mann-Whitney U test

NCBI-GEO- National Center for Biotechnology Information - Gene Expression Omnibus

ncRNA – Non-Coding Ribonucleic Acid

PTM - Post-Translational Modification

RB - Retinoblastoma

RF - Random Forest

RNA-Seq – RNA Sequencing

ROC - Receiver Operating Characteristics

RT-qPCR - Reverse Transcription Quantitative Real-Time Polymerase Chain Reaction

SE – Semmelweis University

SIL – Stable- Isotope Labeled Peptide

SVM - Support Vector Machines

TARGET - Therapeutically Applicable Research to Generate Effective Treatments

TCGA– The Cancer Genome Atlas

TMT – Tandem Mass Tag

# 1 INTRODUCTION

## 1.1 Characteristics of malignant tissues

Cancer develops from normal cells, mutating first to pre-cancerous and then to malignant cells, because of genetic or epigenetic lesions. Such lesions originate mostly in external mutagenic factors, but hereditary mutations also influence their evolution. These genetic lesions lead to gene expression changes in the tumor cells which gear up the cancerous phenotype (1).

While most genes exhibit comparable expression profiles between cancerous and normal tissues, those differentially expressed can serve as either target of treatment or molecular biomarkers of cancer progression. Targeting a gene with higher expression of a certain gene product can deliver astonishing clinical benefit, as was demonstrated over two decades ago by the selective inhibition of overexpressed tyrosine kinases (2).

Gene expression changes in cancer cells are related to a limited set of special characteristics often termed as cancer hallmarks (3). These paramount differences between malignant and normal tissues include sustaining proliferative signaling, evading growth suppressors, resistance to cell death, enabling replicative immortality, inducing/accessing vasculature, activating invasion and metastasis, reprogramming cellular metabolism, and avoiding immune destruction.

## 1.2 Key characteristics of cancer cells – Hallmarks of cancer

Unrestricted cellular proliferation constitutes a fundamental characteristic exhibited by malignant cells. The perpetuation of proliferative signaling and the evasion of growth suppressors are contingent upon the presence of a sufficient number of healthy cells within the sample and/or the sustained maintenance of proliferative capacity over time (4). Notable instances of mutant driver oncogenes that sustain proliferative signaling include the epidermal growth factor (EGF) receptor and the RAS-RAF-MEK-MAPK pathway signaling transporters, which facilitate the processing and transmission of growth-promoting signals (5). In normal cells, proliferative signals are counteracted by inhibitory mechanisms that either override or impede the cell division process triggered by such signals. The genes responsible for encoding these proteins are recognized as tumor suppressor genes, exemplified by RB and tp53 (6, 7). In the majority of human tumors, genetic or epigenetic abnormalities affecting the function of the Rb, and tp53 tumor suppressor pathways are commonly observed (4).

Apoptosis, an orchestrated and active form of cellular demise, represents a programmed self-destruction process. Activation of apoptotic extrinsic and intrinsic pathways, induces alterations in cellular morphology and surface properties, ultimately resulting in genome fragmentation and mitochondrial dysfunction. Subsequently, cells fragment into apoptotic bodies (8). Tumor cells, distinguished by their resistance to apoptosis, employ various mechanisms to achieve this state. These mechanisms include the inactivation of the tumor suppressor gene TP53 and upregulation of antiapoptotic regulators (e.g., Bcl-2, Bcl-xL) or survival signals (e.g., Igf1/2) (4).

Normal cells undergo a finite number of cell cycles, with telomeres playing a vital role in preventing uncontrolled proliferation by safeguarding the ends of chromosomes. In the absence of telomeres, unprotected chromosome ends undergo fusion, resulting in karyotypic abnormalities and cellular demise. Telomerase, a specialized DNA polymerase, catalyzes the addition of telomeric hexanucleotide repeats to the terminal regions of DNA (9). The immortalization of cells, leading to tumorigenesis, is attributed to their capacity to maintain telomeric DNA at a length that evades the induction of senescence (a non-proliferative state) or apoptosis. This ability is primarily achieved by upregulating the expression of telomerase or, less frequently, through an alternative recombination-based mechanism for telomere maintenance (3, 4).

Tumor cells employ the mechanism of angiogenesis to ensure an adequate supply of nutrients and oxygen for their sustenance. Vascular endothelial growth factor-A (VEGF-A) stands as the archetypal inducer of angiogenesis, orchestrating vascular growth, maintaining endothelial cell homeostasis, and facilitating wound healing. The upregulation of VEGF and fibroblast growth factor (FGF) expression in tumors is primarily attributed to oncogenes such as KRAS and hypoxia (4, 10).

The dissemination of cancer is commonly described as a multi-step and sequential process known as the invasion-metastatic cascade. Migration denotes the directed movement of cells without encountering barriers, whereas invasion necessitates the breakdown of barriers for passage, thus involving the remodeling of the extracellular matrix (ECM) (4). During local invasion and distant metastasis, tumor cells undergo notable changes in their shape and their adhesion to other cells and the extracellular matrix. One significant alteration is the loss of the E-cadherin molecule in tumor cells, which plays a pivotal role in cell-to-cell adhesion. Through the formation of adhesion

contacts with neighboring epithelial cells, E-cadherin contributes to the assembly of epithelial cell sheets and the maintenance of cell quiescence within these sheets (11).

Tumor cells must carefully regulate their energy metabolism to sustain their growth. They achieve this by enhancing glycolysis while restricting oxidative phosphorylation, leading to a phenomenon known as aerobic glycolysis. This metabolic shift enables the redirection of glycolytic intermediates towards biosynthetic pathways necessary for the synthesis of new cellular components (4, 12). In many cases cancer cells can be divided to two main subtypes according to their metabolism, one glycolytic like and another with an oxidative like type of metabolism. The glucose-dependent cells release lactate as the final product of glycolysis, while the cells in the other subpopulation uptake lactate generated by neighboring cells and utilize it as their primary energy source, importing the lactate to the citric acid cycle (13). The glycolytic shift has been observed in numerous rapidly dividing embryonic tissues, indicating its involvement in facilitating extensive biosynthetic processes essential for active cellular proliferation and has been demonstrated to be linked with mutant tumor suppressor genes and activated oncogenes (14, 15).

The immune system plays a crucial role in identifying and eliminating early malignant cells, thereby exerting control over the majority of tumor development. Current FDA approved checkpoint inhibitors to treat malignancies include anti-CTLA4 (tremelimumab, ipilimumab), anti-PD-1 (nivolumab, pembrolizumab), and anti-PDL1 (avelumab, atezolizumab, durvalumab) agents (16). Nevertheless, tumor cells have the ability to evade the defensive mechanisms of the immune system (17). This process usually includes the mechanism of losing or altering the expression of antigenic proteins recognized by the immune system. Further mechanism of avoiding immune destruction is to create an immunosuppressive microenvironment, or by a mechanism called the abscopal effect in which the cancer cell modifies the location of the antigene, thus making it invisible to the immune system (18, 19).

Further cancer specific characteristics which enable a cancer cell to proliferate, disseminate, and survive involves genome instability and mutation; and creating a tumor-promoting inflammation in the microenvironment. Genomic instability of premalignant cells could lead to random mutations such as chromosomal rearrangements which could promote hallmark capabilities (4).

A specific inflammatory microenvironment orchestrated by various immunomodulatory cells could also promote a malignant shift in premalignant cells. Recent update of these hallmarks has proposed further potential cancer specific characteristics such as unlocking phenotypic plasticity, nonmutational epigenetic reprogramming, polymorphic microbiomes, and senescent cells (20).

Phenotypic plasticity, a phenomenon of genotypes to produce different phenotypes when exposed to different environmental conditions (21). Normally this capability is blocked in normal cells after they reached the state of terminal differentiation (22). Cancer cells however can reach this state in several ways. Pre-malignant cells with normal ancestry could reverse their way of complete differentiation resulting in dedifferentiated progenitor-like cell states. Progenitor based malignant cells might cut the process of end stage differentiation resulting in a partially differentiated state, similar to the progenitor form. Another type of differentiation switch is the process of transdifferentiation in which cells with a specific developmental fate switch to another differentiation pathway, resulting in new phenotypic characteristics (20).

Besides genetic mutations which show association with multiple forms of the beforementioned cancer hallmarks, there are approaches which highlight the importance of a different mode of genome reprograming termed as the nonmutational epigenetic reprograming, which have been proposed more than a decade ago (23). A notable example of this type of reprogramming is the hypoxia induced aberrant epigenetic regulation in pediatric ependymoma (24).

The microorganism system or microbiome which lives in the barrier tissues of human body has a non-negligible role in the health and disease equilibrium. Proper balance of heterogeneity of different bacterial species might have either protective or harmful effects on the initiation, progression of cancer and response to therapy (25). Changes in this symbiotic ecosystem of ours could influence the process of cancer hallmark gain leading to phenotypic characteristics like modulated cell growth, inflammation, immune evasion, genome instability and therapy resistance (20). The role of dysbiosis in the malignancies of the gastrointestinal tract is widely researched, especially in the case of colon tumors (26). Besides the role of bacteria in the gut, several studies proposed their important role as tumor promoting or repressing factor in oral, skin, and ovarian cancers (27-29).

Cellular senescence is considered as an irreversible cellular process in which - depending on the senescence-inducing triggers- cells end up in state where they stop proliferation and form a dormant state (30). The general assumption of cellular senescence is being protective factor against malignant transformations (31). On the other hand, by certain induction factors senescence of malignant cells could enhance tumor progression and therapy resistance (32).

Previously, several experimental methods capable of examining a variety of these hallmark genes at the gene expression level have been comprehensively reviewed (33). Currently, the most widespread and robust techniques to determine transcriptome-level gene expression include RNA-sequencing and microarray platforms, while selected genes can be measured by RT-qPCR or NanoString technologies (34).

### 1.3 Gene chip technology

Gene chip technology, a form of microarray technology, is widely employed for genome-wide expression profiling, with Affymetrix Gene chips being the most prevalent variant (35). In addition to expression profiling, Gene chips find applications in high-throughput mutation detection, single nucleotide polymorphism (SNP) genotyping, and the detection of chromosomal aberrations (36). The production process of Affymetrix Gene chips distinguishes them from other microarrays by utilizing photochemical synthesis to load DNA probes onto the chips. This technology enables the synthesis of over one million distinct probes on a small array approximately the size of a thumb. Consequently, gene chips have the capability to simultaneously capture multiple oligonucleotides (37) (38). For gene chips, the delivery of probe sequences to quartz slices is achieved through in situ synthesis. As the attachment of nucleotides to the array is light-dependent, photochemical synthesis is employed to transfer nucleotides containing a light-sensitive protecting group onto the quartz slices. The precise attachment points of the nucleotides are controlled using lithography masks (36) (39).

The gene chips are utilized in hybridization experiments, wherein the target DNA or RNA is labeled with biotin and then hybridized to the microarray. To detect the hybridization events, staining is performed using a phycoerythrin-streptavidin-antibody complex, followed by scanning the array using a high-resolution scanner (36).

Gene chip technology offers several advantages. One notable advantage is the ability to read a significant number of features from a microarray, depending on the resolution

of the scanner used in the experiment. However, in recent times, its usage in diagnostics has become less practical and is being gradually replaced by RNA sequencing. RNA sequencing offers greater flexibility, improved sensitivity, and the ability to provide more comprehensive insights into the transcriptome, leading to its increased popularity in diagnostic applications (36). When dealing with well-characterized microorganisms such as bacteria, RNA-seq and gene chip technologies yield comparable results (40). However, for more complex organisms like humans and human cancer, RNA-seq technology surpasses the gene chip method in determining the characteristics of malignant tissues (41).

## 1.4 RNA Sequencing

RNA sequencing (RNA-Seq) is a high-throughput technology to unravel transcriptome specificities. RNA-Seq is a robust technique to quantify gene expression levels, and also enables the discovery of novel transcripts, identification of genes undergoing alternative splicing, and detection of allele-specific expression. Moreover, it facilitates the examination of various RNA types, encompassing total RNA, pre-mRNA, and non-coding RNA types such as microRNA and long non-coding RNA (ncRNA) (42).

In RNA sequencing, the initial step involves extracting RNA from the target biological material, such as cells or tissues. Following this, specific protocols are employed to separate the mRNA from the ribosomal RNA which is more abundant than mRNA. Two commonly used protocols are the poly-A selection protocol, which enriches for polyadenylated transcripts, and the ribo-depletion protocol, which eliminates ribosomal RNAs from the sample (42, 43).

Subsequently, the RNA is converted into complementary DNA (cDNA), via reverse-transcription. Sequencing adaptors are then attached to the ends of the resulting cDNA fragments, and short sequences, known as reads, are generated using high-throughput sequencing (HTS) technology. These reads typically range from 30 to 40 base pairs in length. The generated reads are then compared to a reference genome and classified into different categories, including junction reads, exonic reads, and poly(A) end reads. These categories are utilized to construct a precise expression profile at the base-resolution level for a given gene (44).

Regrettably, the RNA sequencing technique encounters various technological limitations. These challenges encompass the need for substantial data storage capacity

and the complexity involved in developing efficient data processing algorithms. Additionally, the accurate identification of sequencing errors poses a significant obstacle. While minor errors of 1-2 bases are generally manageable, larger variations necessitate more profound sequencing knowledge and precise annotation for accurate interpretation (44). Nonetheless, RNA sequencing has undergone remarkable technical advancements in recent decades. These developments have expanded the capabilities of RNA-Seq to explore multiple facets of RNA biology, including single-cell gene expression, translation dynamics, and RNA structure investigations (45). Moreover, RNA-Seq exhibits several advantages over other methods, such as the ability to precisely pinpoint transcriptional boundaries down to the single-base pair level, providing valuable insights into gene regulation (44).

Both RNA-seq and microarray techniques produce a vast amount of clinically relevant data and large repositories, hosting thousands of samples which are now available. The National Cancer Institute's Genomic Data Commons (GDC) platform provides whole exome sequencing data and transcriptome-level gene expression datasets, such as The Cancer Genome Atlas (TCGA) (46) and the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) (47). The Genotype-Tissue Expression (GTEx) repository makes RNA sequencing, exome sequencing and whole genomic data available for the same patient (48). Nevertheless, the largest open resource is the Gene Expression Omnibus of National Center for Biotechnology Information (NCBI-GEO), which provides microarray, next-generation sequencing and additional high-throughput genomics data for hundreds of thousands of samples (49). In many cases, these repositories provide processed and aggregated results, and it is also common for them to offer raw data. At the same time, digesting such large sample cohorts requires complex bioinformatical analytical tools and can be time-consuming. Mining these databases could be speeded up by an openly available, validated and easily accessible online tool which enables the comparison of expression profiles between normal and cancer related data.

### 1.5 Mass spectrometry

Mass spectrometry is of utmost importance in the examination of biological specimens and has emerged as an essential instrument in the field of proteomics. Mass spectrometry has the capability to measure dynamic changes in protein expression,

interaction, and modifications, thanks to the utilization of different types of labeling techniques (50). Essentially, mass spectrometry quantifies the mass-to-charge ratio (m/z) of ions in the gas phase. A typical mass spectrometer comprises three main components: the ion source, the mass analyzer, and the detector.

The ion source is responsible for converting analyte molecules into gas-phase ions. The mass analyzer separates the ionized analytes based on their m/z ratios. Finally, the detector records the number of ions at each m/z value, providing valuable data for analysis. Usually, mass spectrometers are coupled with some separation techniques such as liquid chromatography to enable the controlled fractionation of protein mixture (51). Currently this machine complex, the HPLC coupled mass spectrometer serves as the gold standard in the field of proteomics.

Regarding the identity of examined proteins there are two main approaches, the discovery-based in which we want to identify a previously unknown set of proteins, this is the so-called shotgun method. The other approach is the targeted approach in which researchers have a predefined set of proteins (52). By the first approach one can identify new biomarkers, however usually this approach is not very specific as proteins with smaller abundance can be missed. Based on sample preparation the two main approaches are the bottom-up and top-down strategies (50, 53). The top-down approach uses intact proteins for further analysis, which can be useful in the case of proper PTM and protein isoform identification. On the other hand, this approach faces several drawbacks due to the complicated protein fractionation, ionization and fragmentation in the gas phase (52). Bottom-up proteomics, considered the standard approach, involves the detection of peptides as an indication of the presence of proteins (50, 54). In this strategy, after the fractionation by HPLC, the proteins of the protein mixture are digested by enzymes, which breaks the proteins down to peptide fragments. Enzymes, like trypsin being the most commonly used, are employed for the digestion process. Trypsin cleaves peptides at the C-terminus of lysine and arginine residues, this specific phenomenon can be used to filter the identified peptides with higher accuracy (55).

Following digestion, the resulting peptides are ionized since mass spectrometry can only measure ions. During ionization, the electrons are removed from the peptides, resulting in the formation of positively and negatively charged ions (56). Two methods are utilized for ionizing samples: MALDI (Matrix-Assisted Laser Desorption/Ionization)

and ESI (Electrospray Ionization). In the MALDI method, the sample is combined with a matrix material and then ionization is achieved through the deposition of ions via a laser pulse. In the ESI procedure, the sample is introduced into the analyzer in the form of tiny droplets using a positively or negatively charged spray. Ionization occurs during the subsequent evaporation of the liquid (57). The ionized samples are introduced into the mass spectrometer, where they undergo separation based on their mass-to-charge ratio. Within the fragmentation chamber, the ionized peptides undergo collisions with noble gases, leading to fragmentation along the weakest bond, which is typically the peptide bond. The resulting fragmented ions are then detected, and the recorded data are subsequently analyzed (56). The outcomes obtained through mass spectrometry are represented in the form of a mass spectrum, which displays the abundance of ions relative to their mass-to-charge ratios. This spectrum provides valuable information, including the molecular mass of the analyzed molecule and the masses of its fragments, which can be utilized to determine the molecule's structure. As molecules undergo specific fragmentation patterns under particular conditions, the mass spectrum obtained can be used for definitive sample identification. In addition to providing quantitative measurements, mass spectrometry can also be employed to determine peptide sequences (53).

Protein quantification in mass spectrometry can be achieved using two main methods: label-free and using labeled ions. The key distinction between these approaches lies in the use of tag molecules for fragment identification. The labeling approach employs different tag molecules, either biological or chemical, to label the fragments and enable their identification. On the other hand, label-free methods rely solely on the intensity of ions during identification, without the use of tag molecules. The TMT (Tandem Mass Tag) method is particularly valuable for determining quantitative differences between samples (53, 56).

When it comes to protein quantitation it is beneficial to consider the utilization of the targeted bottom-up approach. During this process the peptide quantification is done using stable isotope labeled standard peptides, usually labeled with heavy isotopes. If the proteins of interest is known beforehand, by the targeted addition of commercially available stable isotope labeled peptides, -which only differs in the isotopic composition- one can quantify the target peptides in the sample of interest as the SIL peptides have a

predefined mass difference. Since the concentration of SIL peptides in the sample is also known the absolute concentration of the target peptides can be also measured (58).

While mass spectrometry has certain limitations, such as the inability to definitively trace the origin of tryptic peptides to determine the encoding genes of detected proteins, it offers numerous advantages. Mass spectrometry can be used for the identification of unknown compounds by determining their molecular weight, as well as for the quantification of specific compounds. Furthermore, it allows for the determination of the structure and chemical properties of molecules. In recent years, mass spectrometry has found extensive applications in biomarker research (50).

In the clinical practice MS was introduced almost half a century ago in endocrinology and toxicology for drug, steroid, and organic acid quantitation and got its main medical application in the widespread newborn screening (59, 60). Although the setup of MS based diagnostic applications can be costly and complicated at the beginning, the versatility and reliability lead to new applications in clinical settings. In recent years, MS has been proved to be a comparatively cost-effective, precise, and quick analysis tool in microbial identification (61). With the advent of proteomics and proteogenomics, MS based techniques have an increasing role in cancer diagnostics as well (56). Recent studies have shed light on several molecular specificities using a proteogenomic approach in breast, colon, kidney cancers and several further malignancies (62-64). Precision oncology is expected to advance in the next decade through the analysis of proteogenomics data from thousands of tumors across major cancer types. This advancement enables a deeper molecular classification of cancer, guiding personalized approaches for patients and identifying new potential therapeutic targets. Hopefully, this will facilitate the study of the relationship between molecular findings and treatment outcomes, accelerating clinical trials with valuable biomarkers(65).

### 1.6 Clear cell renal carcinoma

As cancer is the second cause of death worldwide, the identification of potential predictive and prognostic biomarkers has outmost importance. Most frequent cancers includes the malignant transformation of breast, lung colon, prostate, and pancreas tissues (66). Besides the beforementioned, carcinoma of the kidney is also a relatively frequent type of cancers with an estimated incidence of more than 80,000 cases in 2023 in the United States. Clear cell renal carcinoma (ccRCC) is the malignant transformation of

epithelial cells of the kidney and is the most frequent form of kidney cancer (67). In 2020, there were 431,288 new cases and 179,368 deaths of kidney and renal pelvis cancer worldwide (68). Although the rate of new cases seems to rise, in the past decades the mortality rates are stagnating in the US (69). Risk factors of ccRCC include obesity, smoking, hypertension, older age, and male gender. Patients with a family history of ccRCC also have a higher risk of developing this disease (70).

Diagnosis of ccRCC is usually based on radiological imaging and tissue slide based histopathological examination. Histopathological confirmation is essential before systematic therapy initiation (70). Treatment of ccRCC can include surgery, percutaneous ablation (71), and targeted drugs including VEGF inhibitors (72) and mTOR inhibitors (73) and checkpoint inhibitors. In the case of localized disease, surgical intervention is the first-line therapy, and depending on the size and stage, the intervention can range from partial to radical nephrectomy. If the tumor mass is relatively small, ablative techniques (such as cryo-, thermo-, or radio-ablation) are also available (71). Patients with early-stage and lack of distant metastasis have more favorable survival rates than those with advanced disease (74). The majority of patients, specifically 93 percent with a low-grade diagnosis, experience a five-year overall survival, whereas only 15 percent of patients with distant metastasis survive for five years(66). Patients with advanced disease (stage IV) further require systemic therapy using mTOR inhibitors, VEGF inhibitors, or checkpoint inhibitors, such as nivolumab, avelumab, pembrolizumab, ipilimumab, and interleukin 2 therapy (75). Patients with locoregional disease can be also treated with pembrolizumab in the adjuvant setting (76).

Uncovering a protein abundance-based gene panel specific to ccRCC could provide valuable support for the everyday clinical diagnostic and therapeutic decision-making process.

# 2 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.

# 3 METHODS

## 3.1 Database Setup — Gene Arrays

We searched the NCBI Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) repository for datasets containing "cancer" samples. Only datasets utilizing the Affymetrix HGU133, HGU133A_2 and HGU133A platforms were considered because these platforms use identical sequences for the detection of the same gene. In total, 3,180 GEO series met these criteria, and each of these has been manually examined. We executed a filtering process to exclude datasets containing either cell line studies, pooled samples, or xenograft experiments. Samples taken after neoadjuvant therapy were also excluded. In addition, samples with incomplete description, unavailable raw data, and repeatedly published samples with distinct identifiers have been removed. For this, the expressions of the first 20 genes were compared, and samples with identical values were identified. In each case, the first published version was retained in the dataset. After the manual selection, the remaining samples were normalized using the MAS5 algorithm by utilizing the Affy Bioconductor library (77). Finally, a second scaling normalization was executed to set the mean expression on each array to 1000. JetSet correction and annotation package was used to pick the most reliable probe set for each gene (78).

## 3.2 Database Setup—RNA-seq

RNA-seq data for a total of 11,688 samples were downloaded from the Genotype-Tissue Expression (GTEx) portal (version no. 7—15 May 2019), from which two non-primary cohorts were removed. Read counts were normalized by the DESeq2 algorithm, followed by a second scaling normalization.

Using the GDC database's (https://portal.gdc.cancer.gov/) TCGA and TARGET projects (version no. 15.0 – 20 February 2019), 11,010 and 1,197 files were downloaded, respectively. We only included primary tumors, adjacent normal, and metastatic tissues. Thus, non-primary tissue samples have been excluded. HTSeq–Counts files were normalized by DeSeq2 and a second scaling normalization was also executed for both cohorts.

## 3.3 Gene annotation

In order to select the optimal probe set for each gene, we used the JetSet correction and annotation package, which delivered 12,210 unique genes in the gene-array datasets.

Appropriate genes in the RNA-seq cohorts were selected and annotated by the biomaRt and AnnotationDbi R packages. After annotation, gene names referring to Long Intergenic Non-Protein Coding RNA, MicroRNA, Small Nucleolar RNA and further non-relevant names were removed. Genes showing zero expression value or NA in any of the tissue types were removed from all datasets. Following the annotation and gene selection in the GTEx, TARGET, and TCGA databases, a total of 21,479 genes remained. After harmonization, the GTEx and GDC data were combined into a single set.

### 3.4 Statistical analysis

Data processing and analysis features of the TNM-plotter pipeline were developed in R version 3.6.1. Comparison of the normal and the tumorous samples was performed by the Mann–Whitney U test, and matched tissues with adjacent samples were compared using the Wilcoxon signed-rank test. Normal, tumorous and metastatic tissue gene comparison were analyzed using Kruskal–Wallis test. The statistical significance cutoff was set at $p < 0.01$.

### 3.5 Shiny user interface

Graphical visualization, including box plots, bar charts, and violin plots produced by the TNM-plotter algorithm, were developed using the ggplot2 R package (79). The web application and the user interface were developed by employing Shiny R packages, with the utilization of the ShinyThemes (http://rstudio.github.io/shinythemes/) and the ShinyCssLoaders (https://github.com/daattali/shinycssloaders) R packages (80).

### 3.6 Validation of differential expression

In order to show that differentially expressed genes truly present differential expression regardless of sample compilation, and to confirm the reliability of the integrated database, we conducted a validation using randomly selected training and test sets across breast, lung, and colon tissue datasets in both RNA-seq and gene array platforms. In this validation process, we compared the expression profiles of normal and tumor samples using the Mann–Whitney U test for 12,210 genes in the GEO and for 21,479 genes in the GDC datasets. After calculating and adjusting the p-values for each gene using the Benjamini-Hochberg method, a chi-squared test was conducted to compare the selection overlap between the training set and the test sets. This test was performed to validate the proportion of differentially expressed genes. Volcano plots comparing $-\log_{10}$ $p$ values and $\mathrm{Log}_2$ fold changes were generated to visualize differential expression.

### 3.7 Cancer biomarker genes

To pinpoint genes showing the highest differential expression between normal and tumor samples across multiple tumor types, we utilized the analysis pipeline and the database of the top ten cancer types with the highest mortality rate. Tumor types were selected using the 2019 mortality data from the United States (81). We compared gene expression values between normal and tumor samples for all available genes in all platforms in each selected tumor type using the Mann–Whitney U test. Then, to combat multiple hypothesis testing, we calculated the False Discovery Rate using the Benjamini–Hochberg method. Subsequently, the remaining significant genes were ranked by using the median fold change (FC) in all tissues. In other words, the significant genes were ranked based on their gene expression differences across all investigated tumor types. Finally, we selected genes with the highest FC values in both RNA-seq and gene array datasets, respectively.

### 3.8 Determining differentially expressed genes in ccRCC

Data processing and analysis were performed in R version 4.1.0 (https://www.r-project.org). Wilcoxon signed-rank test was used to compare the tumorous and adjacent normal samples. Genes showing significant differences according to the Wilcoxon test ($p < 0.01$) have been selected and ranked based on their fold-change values (FC). Finally, the top 30 genes with a FC over two and significant in both RNA-seq and gene chip cohorts were selected for further investigation.

### 3.9 Ethics statement

ccRCC samples were collected at the Department of Urology of the Semmelweis University. An institutional ethical review board approved the study under the number ID 7852-5/2014/EKU by the Semmelweis University Regional and Institutional Committee of Science and Research Ethics. All subjects were treated under the tenets of the Declaration of Helsinki and written informed consents were obtained before sample collection.

### 3.10 Sample collection for the proteomic and transcriptomic analysis

Clear cell renal carcinoma and adjacent normal samples were collected during surgical resection from patients diagnosed with ccRCC, tissue samples were stored

immediately at -80°C. Sample collection happened between 2011 and 2013, the median follow up was 1241 days.

RNA and protein isolation was performed using the AllPrep DNA/RNA/Protein Mini Kit and RNeasy Mini Kit (Qiagen, Hilden, Germany) by the manufacturer's protocol using 30 mg of tissue samples. RNA was quantified using a Qubit fluorometer (ThermoFisher, Waltham, USA), and RNA quality check was done using Fragment Analyzer Standard Sensitivity RNA Analysis Kit (Agilent, Santa Clara, USA)

### 3.11 Gene expression analysis – RNA Sequencing

According to the manufacturer's protocol, tissue samples were processed with Illumina TruSeq Stranded mRNA Sample Prep Kit (Illumina, San Diego, USA). mRNA has been enriched using oligo-dT attached magnetic beads before cDNA synthesis has been performed. Then, the fragments were adenylated, and Illumina sequencing adapters have been ligated onto them. Each sample was indexed with Illumina Truseq HT indexes. Finally, samples were cleaned up and sequencing has been performed in an Illumina NextSeq 500 instrument (Illumina, San Diego, USA) using the NextSeq500/550 High Output v2.5 (150 Cycles) sequencing kit. Before gene expression analysis, the FASTQ files were examined by FASTQC. Reads were aligned to GRch38 using the STAR alignment tools, and the reads were counted using featureCounts (82, 83). Quality control, read alignment, and counting was performed in the Galaxy platform (84). DESeq2 and second scaling normalization were done using the count files (85).

### 3.12 Targeted liquid chromatography coupled tandem mass spectrometry (LC-MS/MS) analysis

The expression of selected target proteins was verified by targeted LC/MS-MS. After isolation, protein samples were stored in guanidine isothiocyanate, and stored at -80°C. For targeted quantification we used stable isotope labeled (SIL) peptides (1-5 respectively for each protein, labeled at Arg:13C6;15N4, Lys:13C6;15N2). Protein concentration was determined by the bicinchoninic acid (BCA) test. Samples were reduced by dithiothreitol (DTT) and alkylated using iodoacetamide followed by protein precipitation, then samples were re-dissolved in 5% SDS/50 mM ammonium-bicarbonate for BCA test. Sample volumes representing 50 ug protein content were digested by trypsin

according to the S-trap protocol (https://files.protifi.com/protocols/s-trap-mini-long-4-1.pdf).

LC-MS/MS analysis was performed using an ACQUITY UPLC M-Class system (Waters, Milford, MA, USA) with HPLC coupled to an Orbitrap Fusion Lumos Tribrid (Thermo Fisher Scientific, Waltham, MA, USA) mass spectrometer on the mixture of the protein digests spiked with the mixture of the SIL peptides. Samples were loaded onto a trap column, ACQUITY UPLC M-Class Symmetry C18 Trap (100 Å, 5 µm, 180 µm × 20 mm, 2G, V/M); the sample loading time was 5 min; the flow rate was 5 µL/min, and separation was performed on an ACQUITY UPLC M-Class Peptide BEH C18 (130 Å, 1.7 µm, 75 µm × 250 mm) column with a flow rate of 400 nL/min. MS data acquisition was performed in an internal standard triggered parallel reaction monitoring fashion (86), where the presence of the corresponding SIL peptides, verified by their expected retention time and MS2 fragmentation pattern, triggers data acquisition of the targeted peptides with high sensitivity and resolution. MS signal intensities of the SIL peptides were between 1–5E7. Raw MS data were analyzed using the Skyline software and the MSstats statistical analysis tool. During the data processing steps, we performed the inbuilt normalization steps of the MSstats software package, which includes median polishing and log2 transformation.

### 3.13 Statistical and functional analysis, data visualization

T-test was used to compare the log2 transformed protein intensity values between the tumorous and adjacent normal samples. In order to examine if any of the gene candidates are affected by covariates, we performed a t-test to see if any of the proteins show differential expression between male and female patients. To examine age as a covariate factor, we performed regression analysis to see if any of the examined proteins are influenced by age. Functional analysis was performed using the clusterProfiler R package (87). For each protein, we performed Cox proportional hazard regression analysis. To estimate the best cutoff value for each protein, we examined each possible cutoff values between the lower and the upper quartiles; these cutoff values have been used for Kaplan–Meier plot visualization. The Benjamini–Hochberg method was used for p-value adjustment. For survival analysis, we used the survminer and survival R packages. Further visualization has been done using the R packages ggplot2 (79), ComplexHeatmap (88) and ggrepel (https://cran.r-

project.org/web/packages/ggrepel/index.html). Correlation analysis was done using data of 88 samples from 57 patients with simultaneously available RNA-Seq and MS results. The normalized intensity and read count values were correlated using Spearman correlation, statistical significance was set to 0.01.

### 3.14 Building a model for ccRCC detection

Using the results of the targeted LC/MS-MS $log_2$ intensity values we tried four supervised AI methods, k-nearest neighbors (KNN), random forest (RF), logistic regression (LOGIT), and support vector machines (SVM) to set up the most accurate model for cancer detection. The data matrix from MS data was the input for the classification model, and we used the "caret" R package for data preparation and model establishment (89, 90). From all available patients with MS data, we had to remove one patient due to a missing value. The entire cohort was split into training and test cohorts with a ratio of 0.7:0.3. Repeated K-fold cross-validation was used for training cohort resampling with 10 folds and five repeats. Within the resampling mechanism, we performed recursive feature elimination to specify the ideal number of used genes for each of the SVM, KNN, LOGIT, and RF algorithms. Model prediction capability was validated using the test set. The caret package's built-in methods were used to determine accuracy, specificity, sensitivity, and kappa value, as well as for visualization.

# 4 RESULTS

## 4.1 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 sample. In order to increase the number of normal samples, we further included 11,215 RNA-Seq GTEx data from non-cancerous persons. Steps of data curation and processing are summarized in **Table 1**.

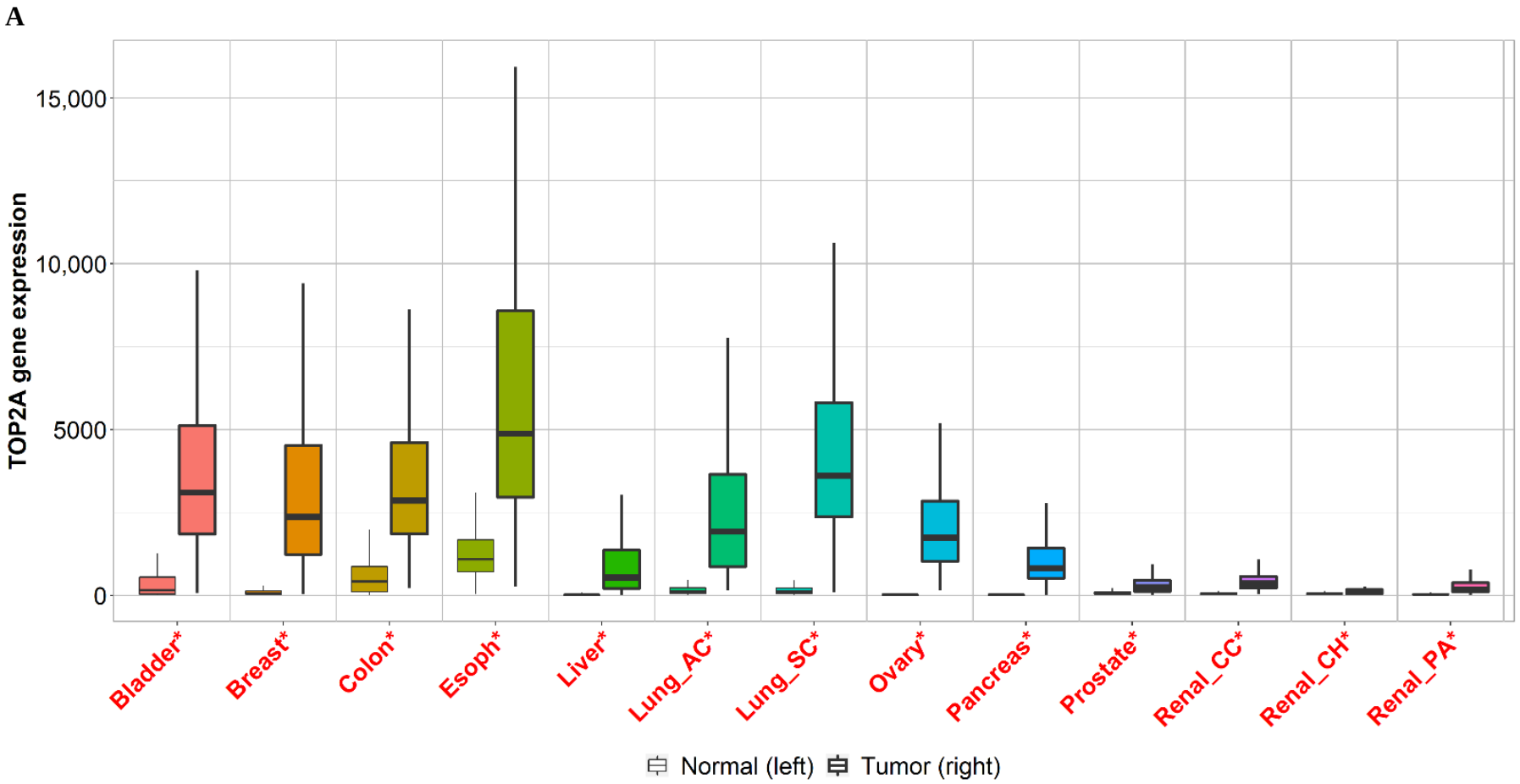
**Table 1.** Summary of datasets and data processing
(T: Tumor, N: normal, M: metastatic). Source: (91)

| | **Manual Screening** | | **Computational Screening** | | | **Result** | **T** | **N** | **M** |
|---|---|---|---|---|---|---|---|---|---|
| NCBI GEO | GSE - 3,180 datasets | Primary tissue series *n* = 554 (38,897 Samples) | Data cleaning | MAS5 (77) normalization and scaling | JetSet (78) Annotation | 38,431 Samples 38 tumor types | 29,376 | 3,691 | 453 |
| TARGET | 1,193 samples | - | Data cleaning | DESeq2 (85) normalization and scaling | AnnotationDBI (92)annotation | 1,193 samples 7 tumor types | 1,180 | 12 | 1 |
| TCGA | 11,050 samples | Removal of non-primary tissues | Data cleaning | DESeq2 normalization and scaling | AnnotationDBI annotation | 11,010 samples 33 tumor types | 9,886 | 730 | 394 |
| GTEx | 11,688 samples | Removal of non-primary tissues | Data cleaning | DESeq2 normalization and scaling | biomaRt (93) AnnotationDBI annotation | 11,215 samples 51 tissue types | - | 11,215 | - |

### 4.1.1 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 (91). 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. A representative boxplot of pan-cancer analysis using cancer types with the highest mortality rate is displayed in **Figure 1**.

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. Example outputs of normal–tumor comparison are displayed in **Figures 2 and 3.**


A
TOP2A gene expression
15,000
10,000
5000
0
Bladder*
Breast*
Colon*
Esoph*
Liver*
Lung_AC*
Lung_SC*
Ovary*
Pancreas*
Prostate*
Renal_CC*
Renal_CH*
Renal_PA*
Normal (left)
Tumor (right)

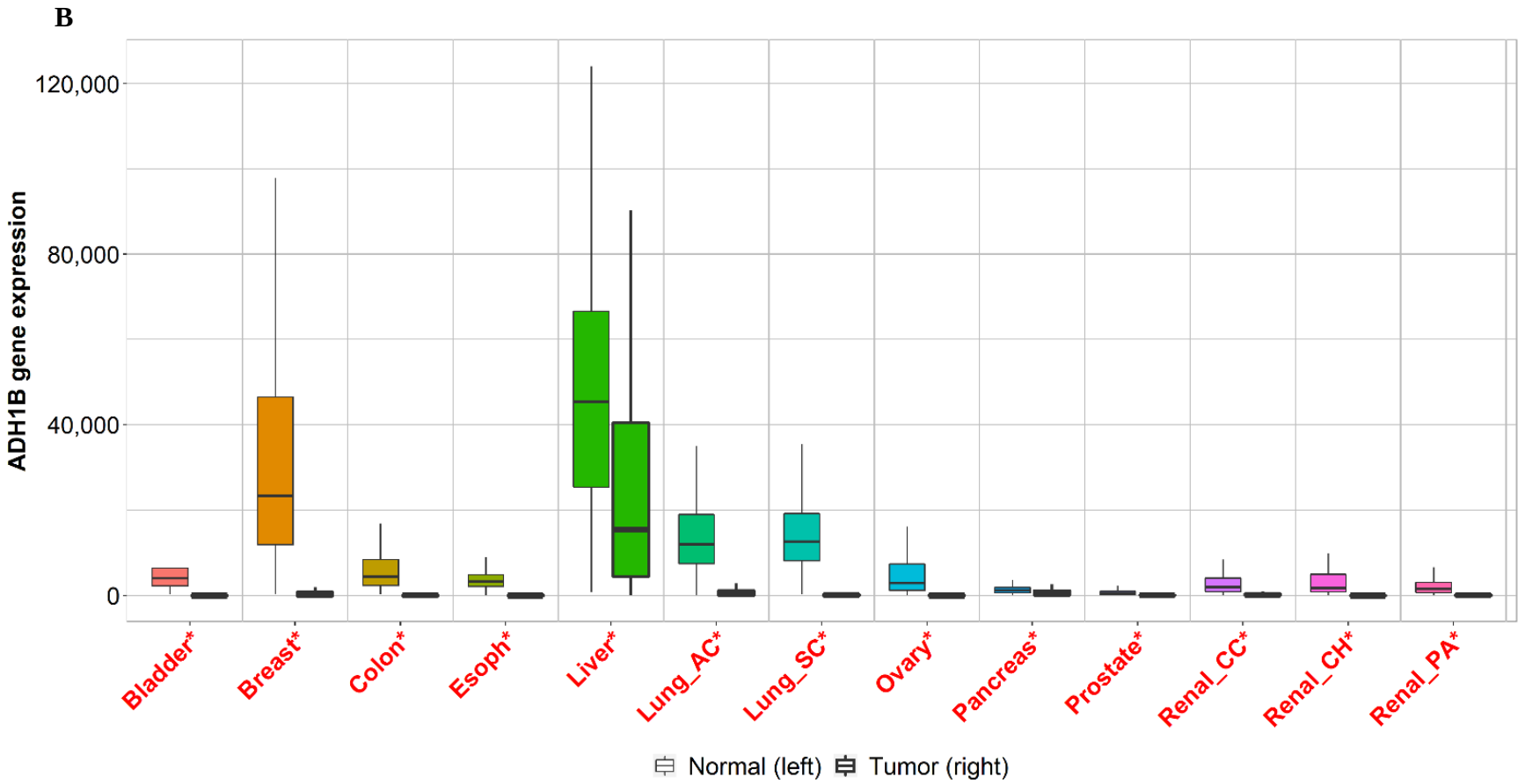



**Figure 1. A and B:** Boxplots of top two genes differentially expressed in most of the cancer types with the highest mortality rates. Significant differences by a Mann–Whitney U test are marked with red color (* $p < 0.01$), *Abbreviations*: Esoph – esophagus; Lung_AC – lung adenocarcinoma; Lung_SC – squamous cell lung cancer; Renal_CC – clear cell renal carcinoma; Renal_CH - Chromophobe renal cell carcinoma; Renal_PA - papillary renal cell carcinoma. Source: (91).

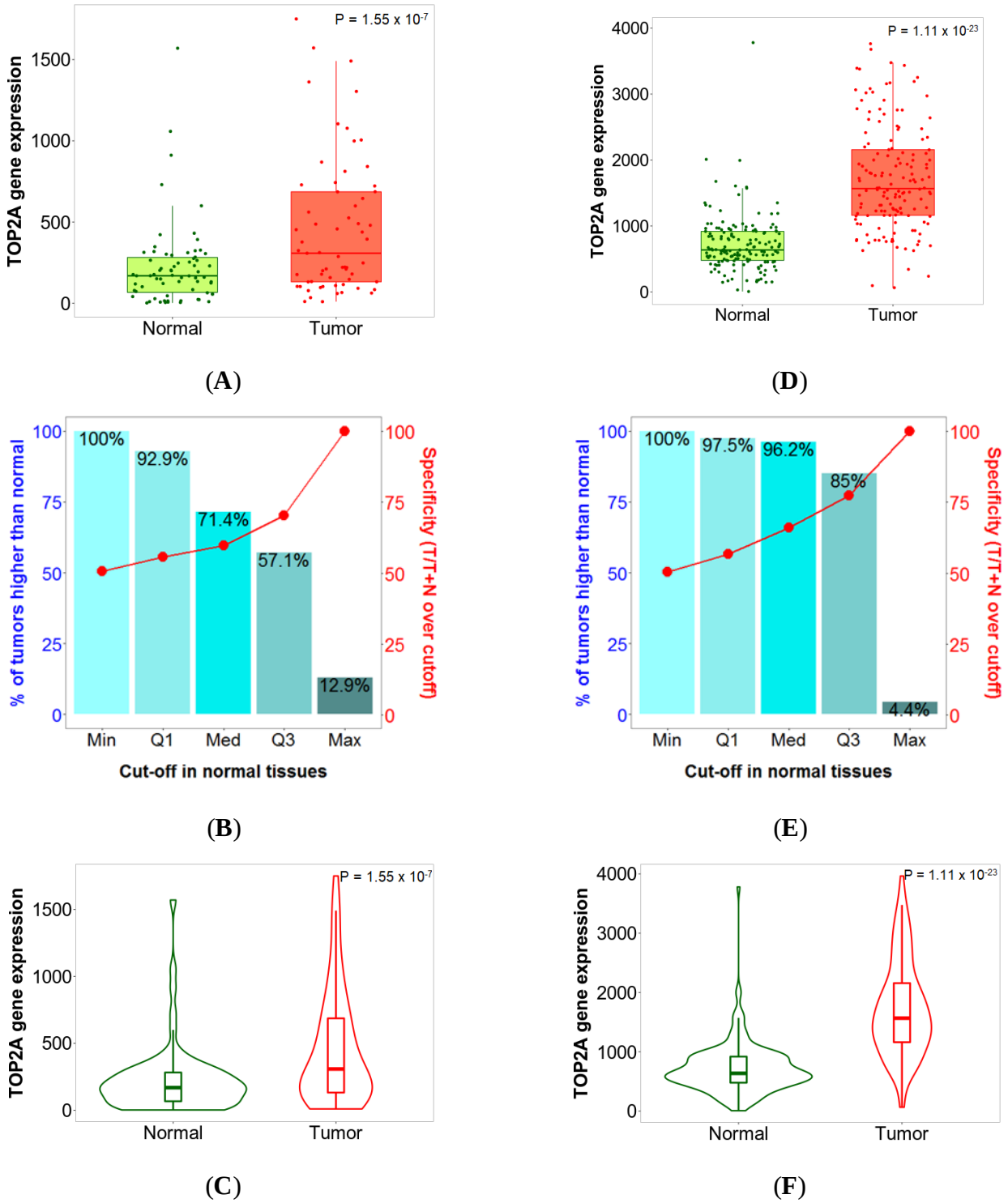



**Figure 2.** Boxplots (**A**,**D**), bar charts (**B**,**E**) and violin plots (**C**,**F**) of TOP2A gene expression in breast (left) and colon cancer (right) when comparing paired normal and tumor gene array data. The bars represent the proportions of tumor samples that show higher expression of the selected gene compared to normal samples at each of the quantile cutoff values (minimum, 1st quartile, median, 3rd quartile, maximum). Specificity is calculated by dividing the number of tumor samples with the sum of tumor and normal samples *below* each given cutoff. In cases where the fold change was over 1, those "*over*" were used instead of those "*below*". Source: (91).

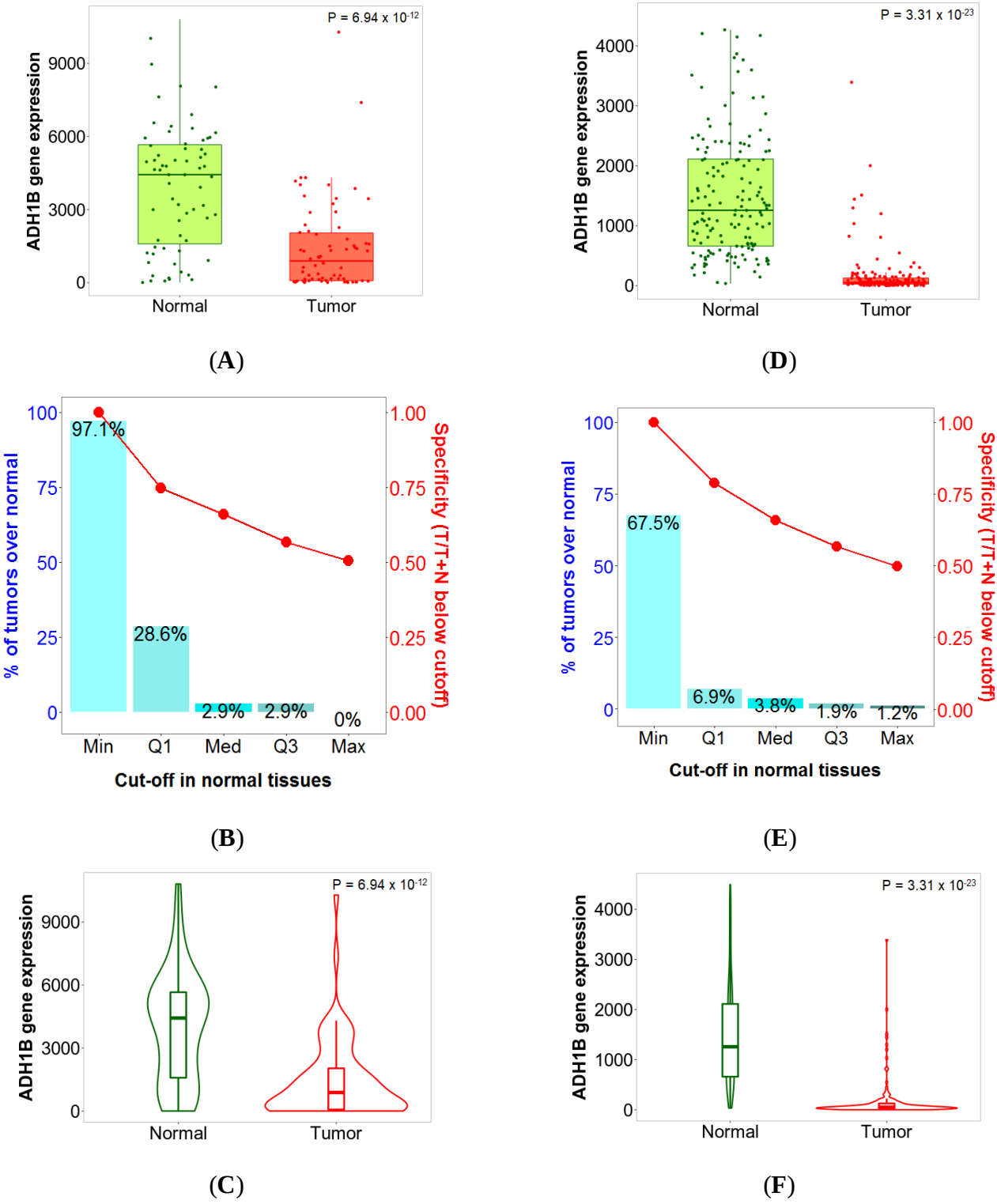



**Figure 3.** Boxplots (**A**,**D**), bar charts (**B**,**E**) and violin plots (**C**,**F**) of ADH1B gene expression in breast (left) and colon cancer (right) when comparing paired normal and tumor gene array data. The bars represent the proportions of tumor samples that show higher expression of the selected gene compared to normal samples at each of the quantile cutoff values (minimum, 1st quartile, median, 3rd quartile, maximum). Specificity is calculated by dividing the number of tumor samples with the sum of tumor and normal samples *below* each given cutoff. In cases where the fold change was over 1, those "*over*" were used instead of those "*below*". Source: (91).

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.

### 4.1.2 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).

TOP2A was the most upregulated gene in the above analysis, with a fold change of 3.26 in breast cancer and 2.54 in colon cancer, among others. In **Figure 2**, the expression boxplot, the sensitivity/specificity plot, and the violin plots for TOP2A are displayed using the breast and colon cancer datasets. The most downregulated gene was ADH1B, which had a fold change of 0.3 in breast cancer and 0.22 in colon cancer (see detailed plots in **Figure 3**).

### 4.1.3 Gene expression analysis of cancers with the highest mortality

We compared the expression of all genes in normal and tumor samples across the ten most lethal tumor types, including breast, bladder, colon, lung, liver, esophageal, prostate, pancreas, renal, and ovarian cancers. In the gene array dataset, 555 and 2,623 genes reached statistical significance at False Discovery Rate (FDR) <10% and fold change over 1.5, respectively. Similarly, in the RNA-seq cohort, 3,189 and 12,037 genes were dysregulated at FDR <10% and fold change over 1.5, respectively.

### 4.1.4 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 as presented in **Figure 4.** 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 as presented in **Figure 5**. The overlapping 55 genes are listed in **Table 2.**

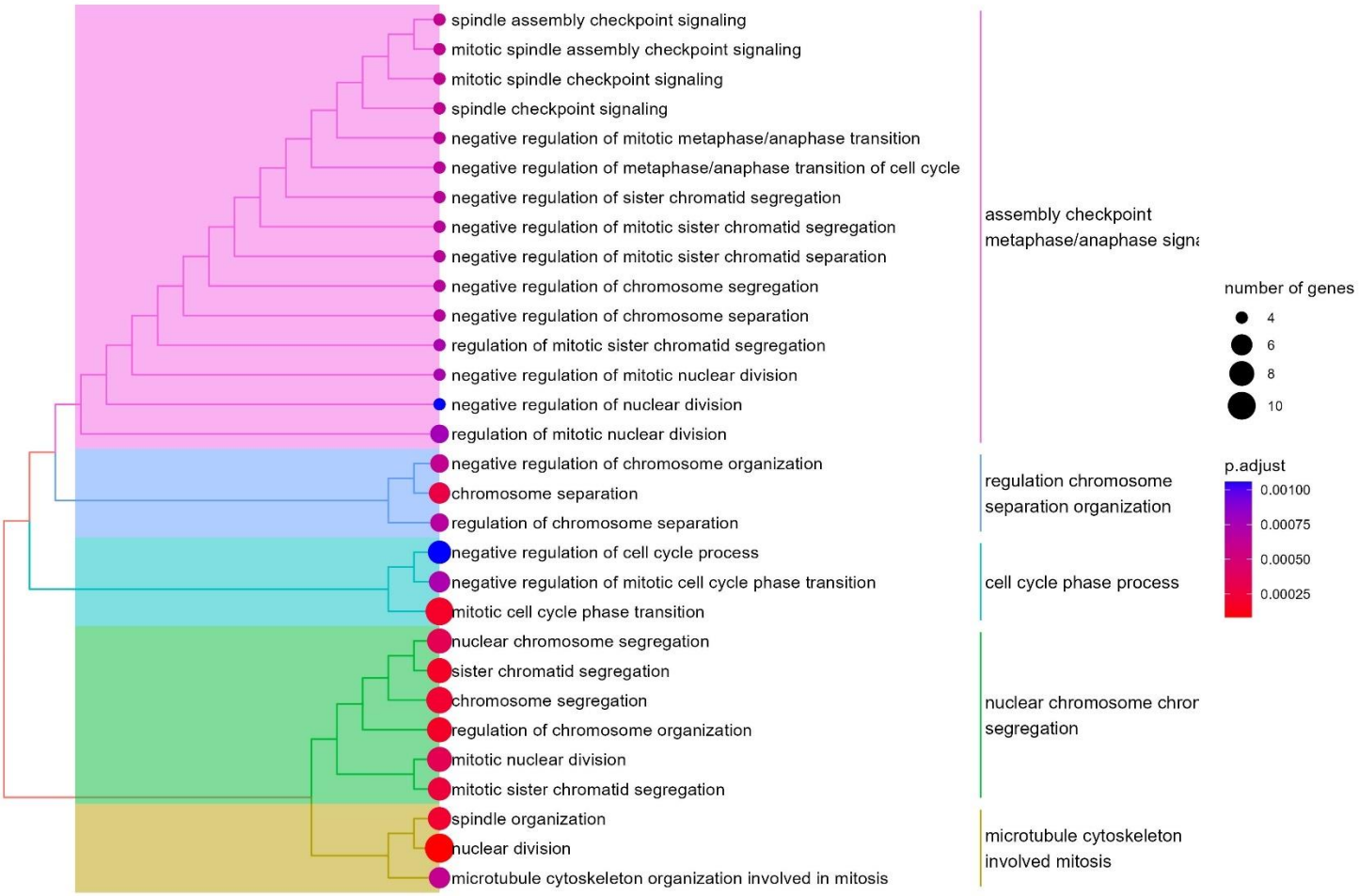



**Figure 4.** Functional representation of the mostly enriched GO terms of biological processes using a gene set commonly identified across the deadliest cancers.

**Table 2.** Top fifty-five genes differentially expressed when comparing normal and tumor samples across the ten most common tumor types in RNA-seq and gene array datasets. Fold change over one corresponds to higher expression in tumors, and fold change below one corresponds to higher expression in normal specimens (highlighted in grey). Source: (91).

| Gene | Mean Fold Change | Gene | Mean Fold Change |
|---|---|---|---|
| TOP2A | 7.8 | RUVBL2 | 1.77 |
| SPP1 | 7 | TMSB10 | 1.76 |
| CENPA | 6.03 | RPN1 | 1.75 |
| NEK2 | 5.63 | CHPF2 | 1.67 |
| MELK | 5.46 | CERS2 | 1.63 |
| HMMR | 5.29 | SH3BGRL3 | 1.61 |
| KIF20A | 4.96 | APRT | 1.6 |
| NEIL3 | 4.89 | IRAK1 | 1.56 |
| TTK | 4.85 | SEC61A1 | 1.54 |
| ASPM | 4.82 | PSME2 | 1.52 |
| CCNB2 | 4.76 | SPAST | 1.49 |
| DTL | 4.44 | DNASE1L1 | 1.42 |
| NCAPG | 4.44 | PGLS | 1.4 |
| ZWINT | 4.15 | DIRAS3 | 0.6 |
| CCNB1 | 4.14 | ECHDC3 | 0.59 |
| BUB1B | 3.79 | PDE8B | 0.56 |
| TK1 | 3.76 | PCDH9 | 0.52 |
| PRC1 | 3.72 | PEG3 | 0.46 |
| CENPU | 3.58 | PKNOX2 | 0.44 |
| KPNA2 | 3.23 | CXCL12 | 0.42 |
| CENPN | 3.03 | PHYHIP | 0.33 |
| CKAP2 | 2.62 | GPM6A | 0.32 |
| KNOP1 | 2.26 | FHL1 | 0.27 |
| SNRPB | 2 | DPT | 0.25 |
| MAGOHB | 1.9 | C7 | 0.24 |
| RPN2 | 1.83 | AOX1 | 0.22 |
| SNRPF | 1.82 | ADH1B | 0.15 |
| ENO1 | 1.79 | | |

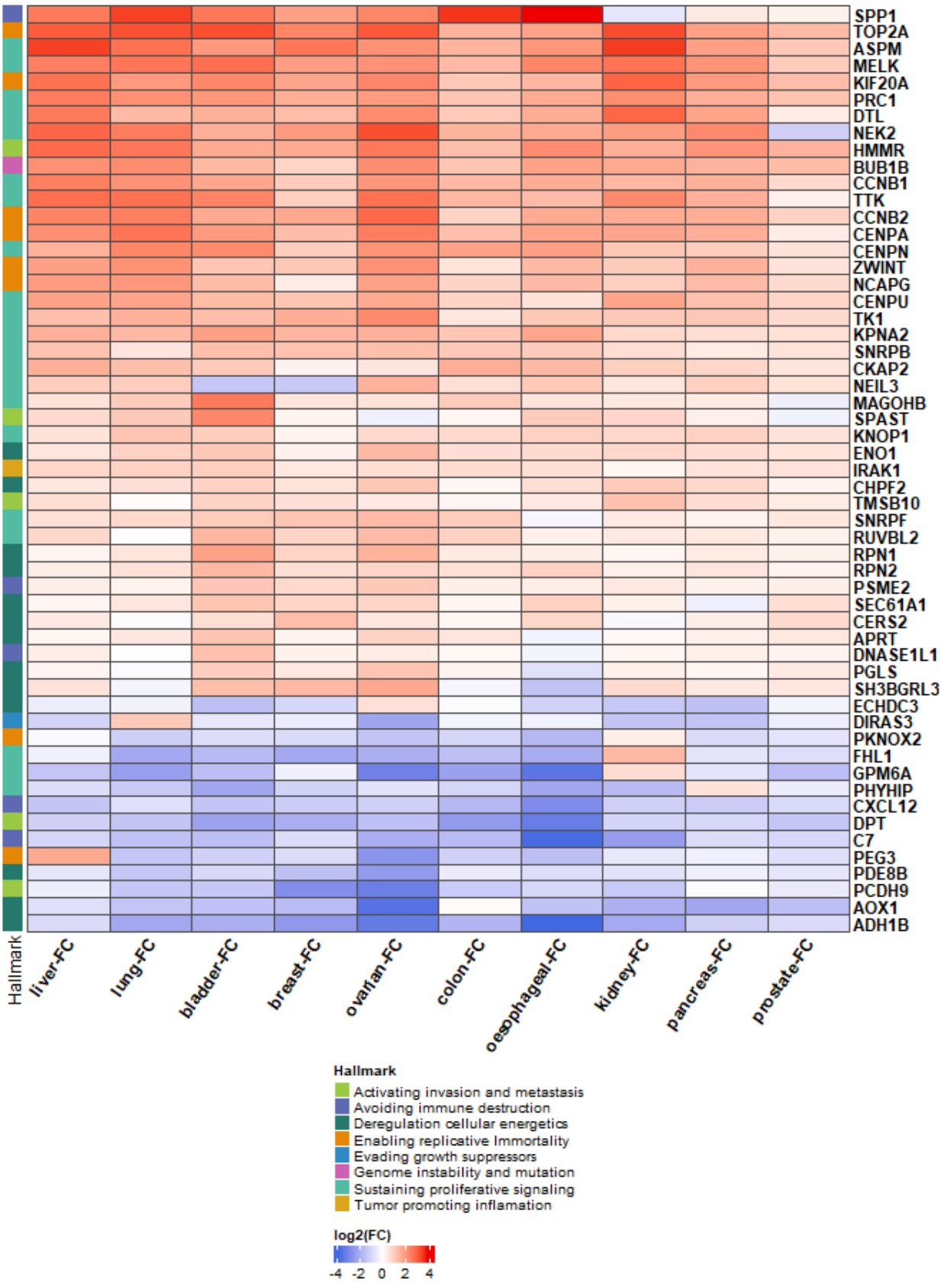
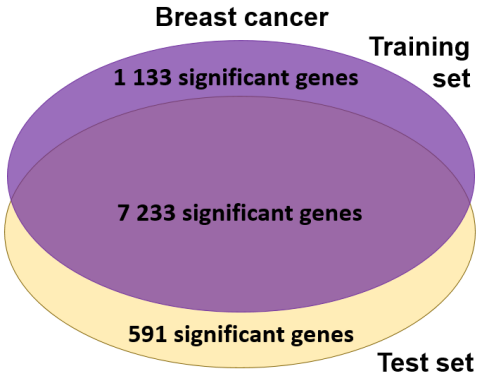



**Figure 5**. Manually curated hallmark representation of the top 55 markers across the top 10 deadliest cancer types.

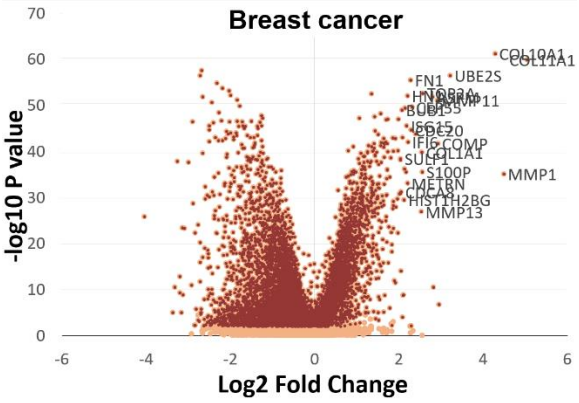
### 4.1.5 Validation of differential expression between normal and tumor samples

In order to confirm the reproducibility of differential expression, and to confirm the reliability of the integrated database, we conducted a validation using randomly selected training and test cohorts across breast, lung and colon cancers using both RNA-seq and gene array samples. During this process, we analyzed the normal–tumor gene-expression difference for all genes in these three selected tissue types. Randomly selected sample cohorts comprised the test and the training set, and we conducted differential gene-expression analysis for all genes in both training and test sets using a Mann–Whitney U test. In each setting, the training and test sets were equally sized to avoid false positive or false negative findings. Using a chi-square test, we aimed to validate the proportion of differentially expressed genes across both test and training sets. In the breast cancer gene array and RNA-seq datasets, respectively, 7,223 and 11,689 genes showed significant difference in both training and test sets. These deliver a high concordance in both cases with a chi-square test $p$ value <0.0001. Regarding colon cancer, 8,259 and 6,763 genes presented significant difference in both training and test datasets in gene array and in RNA-seq samples, respectively ($p < 0.0001$). In lung cancer, altogether, 7,846 and 8,484 overlapping genes reached significance in both examined cohorts in the gene array platform and in RNA-seq, respectively ($p < 0.0001$).

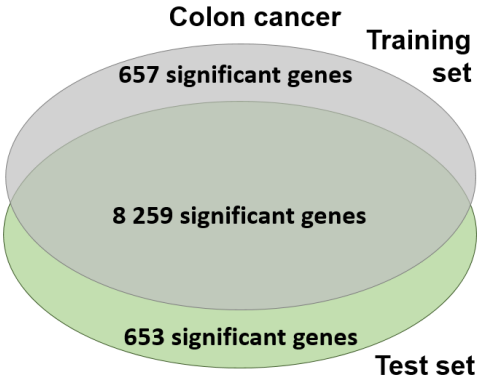
Based on the results of each analysis, which consistently showed a p-value of less than 0.0001, we concluded that the database has the potential to yield highly reproducible results in both platforms. Volcano plots and Venn diagrams depicting the results of the validation are listed in **Figure 6**.



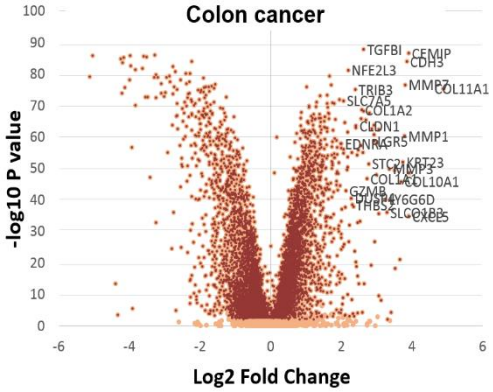

(A)



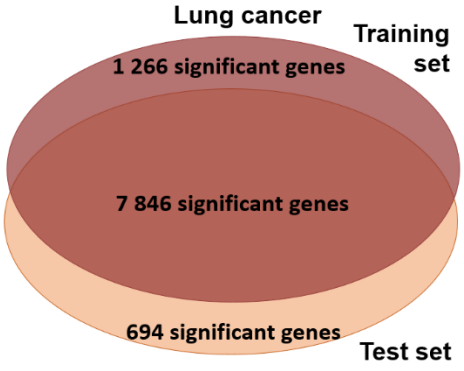

(B)



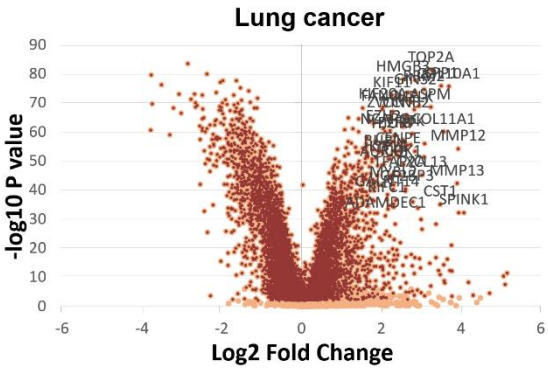

(C)




(D)




(E)




(F)

**Figure 6.** Volcano plots and Venn diagrams of differentially expressed genes in breast, colon and lung cancer using equally sized training–test sets —Venn diagram (A) and Volcano plot (B) from breast cancer; Venn diagram (C) and Volcano plot (D) from colon cancer Venn diagram (E) and Volcano plot (F) from lung cancer. Source: (91).

### 4.1.6 Top genes differing between malignant and normal breast tissues.

To establish a tissue-specific expression pattern in malignant breast tissues, we conducted an analysis to identify genes exhibiting statistically significant differential expression. To achieve this, we applied a fold change (FC) cutoff of 1.5 and a false discovery rate (FDR) cutoff of 10%. Notably, our findings revealed several highly differentially expressed genes in breast cancer, such as COL11A1 (FC = 22.34, p = 3.5E−108) and MMP1 (FC = 17.73, p = 1.6E−59) summarized in **Figure 7A**. Additionally, COL10A1 and UBE2S exhibited substantial differences in gene expression with FC values of 16.55 and 9.14, respectively, along with adjusted p-values of 2.8E−114 and 5.7E−109. Furthermore, we conducted a gene ontology analysis to explore the functional characteristics of the identified top genes. The most significantly enriched term was related to extracellular external encapsulating structure, with subterms related to matrix and structure organization. Further significantly enriched term with numerous genes involved was the spindle cytoskeleton fission division with enriched subterms related to chromosome segregation and nuclear division **Figure 7B**.

### 4.1.7 Top genes differing between malignant and normal colon tissues.

To discern a tissue-specific expression pattern specific to colorectal cancer, we identified the top genes exhibiting statistically significant differential expression. To accomplish this, we implemented stringent criteria, setting a fold change (FC) cutoff of 1.5 and a false discovery rate (FDR) cutoff of 10%. Our analysis revealed several highly differentially expressed genes, including COL11A1 (FC = 30.72, p = 7.3E−149) and CST1 (FC = 17.28, p = 6.5E−100) **Figure 8A**. Additionally, notable differences in gene expression were observed for PPBP (FC = 16.67, adjusted p = 4.4E−44) and CEMIP (FC = 16.6, adjusted p = 3E−171). Using gene ontology analysis, we successfully identified the functional terms that are associated with the upregulated genes. Our analysis resulted in significant enrichment in biological processes associated with the cell checkpoint cycle phase with enriched subterms related mostly to the regulation of nuclear division, chromosomal organization, and further positive regulation cell cycle processes. Another important enriched term with several involved genes and subterms was the sister fission chromatid segregation with subterms related to organelle fission and nuclear division **Figure 8B**.

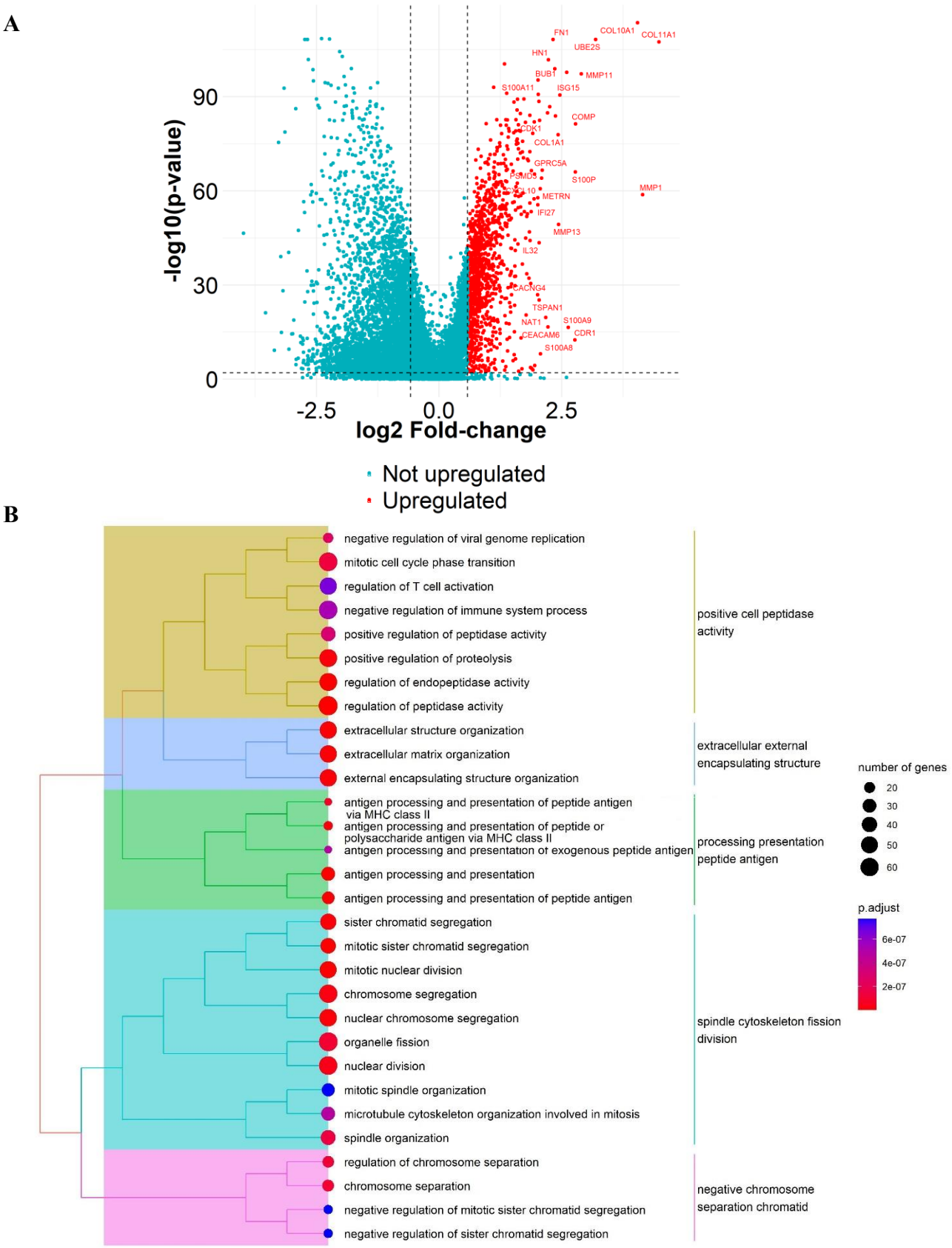



**Figure 7.** Volcano plot of highly expressed genes specific to breast cancer (A) Functional enrichment result of the mostly enriched GO terms in breast cancer (B).

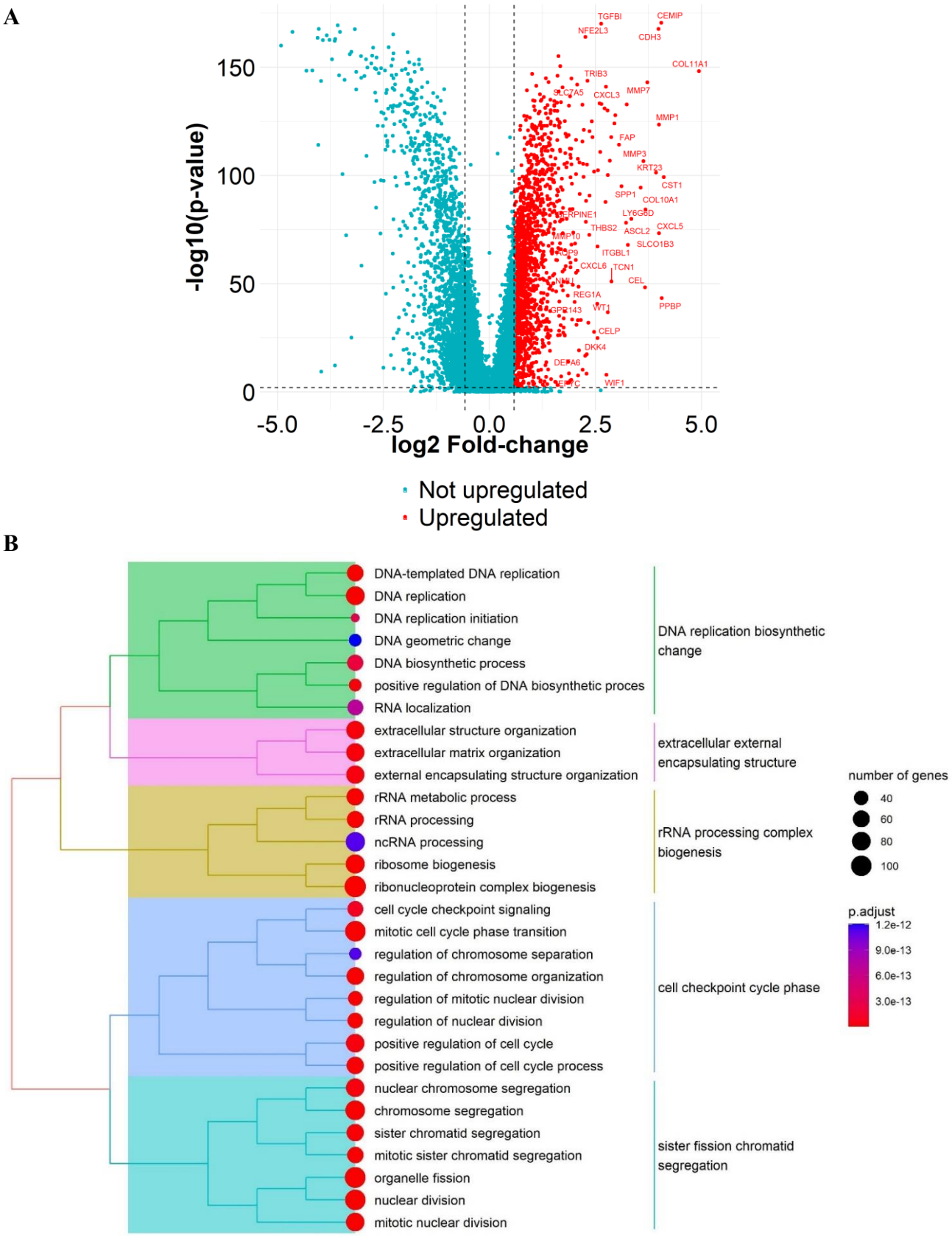



**Figure 8.** Volcano of highly expressed genes specific to colon cancer (A) Functional enrichment result of the mostly enriched GO terms in breast cancer (B).

### 4.2 Clear cell renal carcinoma transcriptomic database setup

The database including both RNA-Seq and gene chip datasets comprises 1,317 samples. The RNA-seq based data consists of 607 samples from the GDC TCGA project including 535 malignant specimens. These RNA-seq based datasets also include 72 paired normal samples which we used to identify tumor specific differences. In the gene chip dataset, we included 23 GEO series which contain 715 samples. Of these 715 samples, 277 were from normal kidney tissues, and 438 were from ccRCC. 414 samples out of the entire gene array database were paired samples (207 pairs), and we used the paired specimens for further analysis. Thus, we all used 558 (144 RNA-Seq and 414 gene chip) normal and tumor samples to identify differentially expressed genes. The entire analysis pipeline is summarized in **Figure 9**.

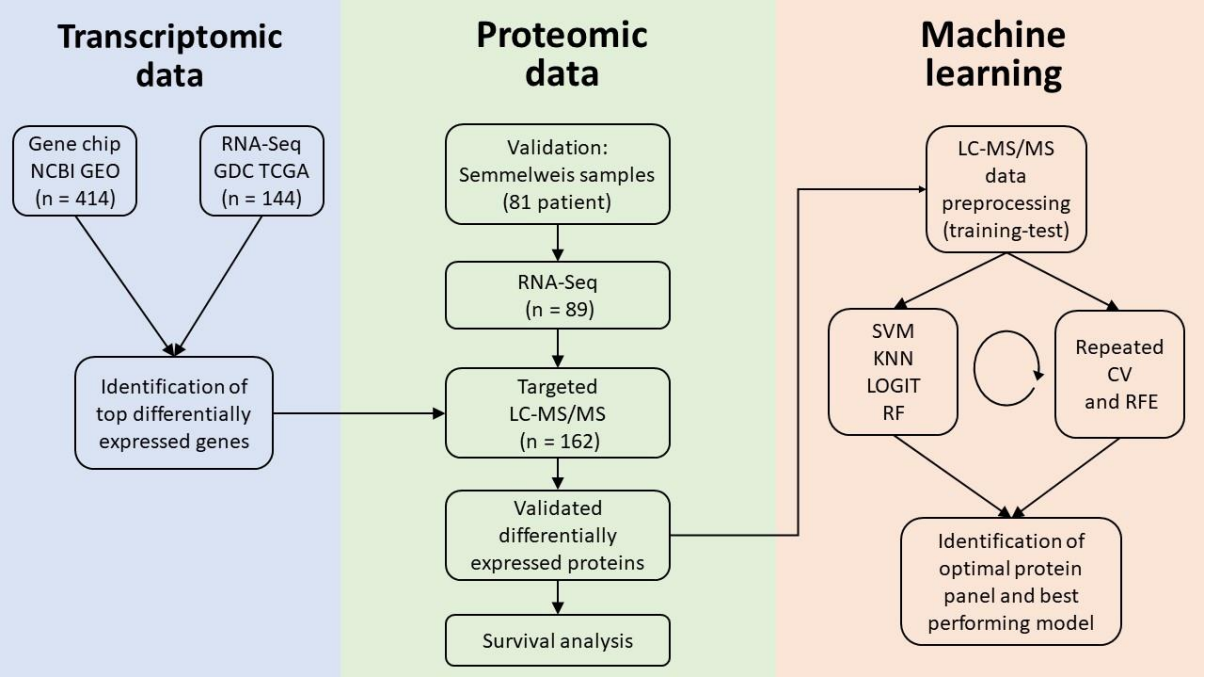



**Figure 9.** Analysis pipeline. Using transcriptomic data, we identified the top differentially expressed genes discriminating normal kidney tissue and ccRCC. Using our own patient sample data we performed RNA-Seq to measure gene expression and targeted LC-MS/MS to measure protein abundance for the selected top genes. Using proteomic data, we established an optimal gene panel and the most accurate model for ccRCC detection. CV: K-fold cross-validation, RFE: recursive feature elimination, KNN: k-nearest neighbors, RF: random forest, LOGIT: logistic regression, and SVM: support vector machines. Source: (94).

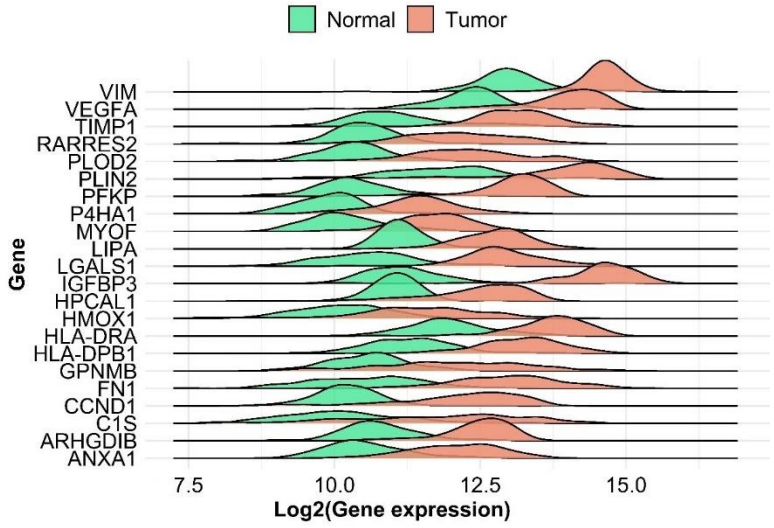
#### 4.2.1 Genes over-expressed in ccRCC

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 ($FC_{\text{gene chip}} = 8.15$, $p = 1.01E\text{-}33$ and $FC_{\text{RNA-Seq}} = 10.47$, $p = 2.17E\text{-}12$).

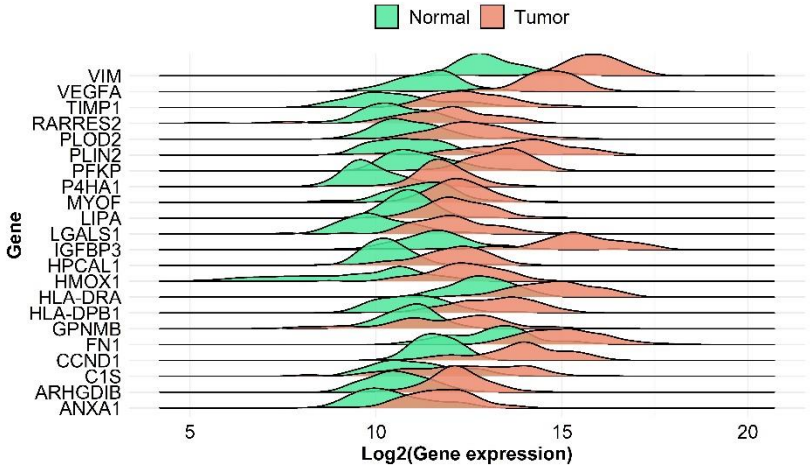
The most significant genes include previously established molecular targets like VEGFA (FC $_{\text{gene chip}}$ = 3.02 p = 3.1E-32 and FC $_{\text{RNA-Seq}}$ = 9.03, p = 3E-13) and CCND1 (FC $_{\text{gene chip}}$ = 4.12, p = 2.3E-32 and FC $_{\text{RNA-Seq}}$ = 5.98, p = 4.25E-13). PLIN2 is a further 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. Top differentially expressed genes are also shown in **Figure 10**.

**Figure 10. A and B** Differential gene expression of compared normal and ccRCC tumor samples from the gene chip data. Ridge plots of differentially expressed genes shows the distribution of log2 expression values. Source: (94). **B** ridge plots of ccRCC samples and adjacent normal tissues from TCGA data.

**A**




**B**




### 4.2.2 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

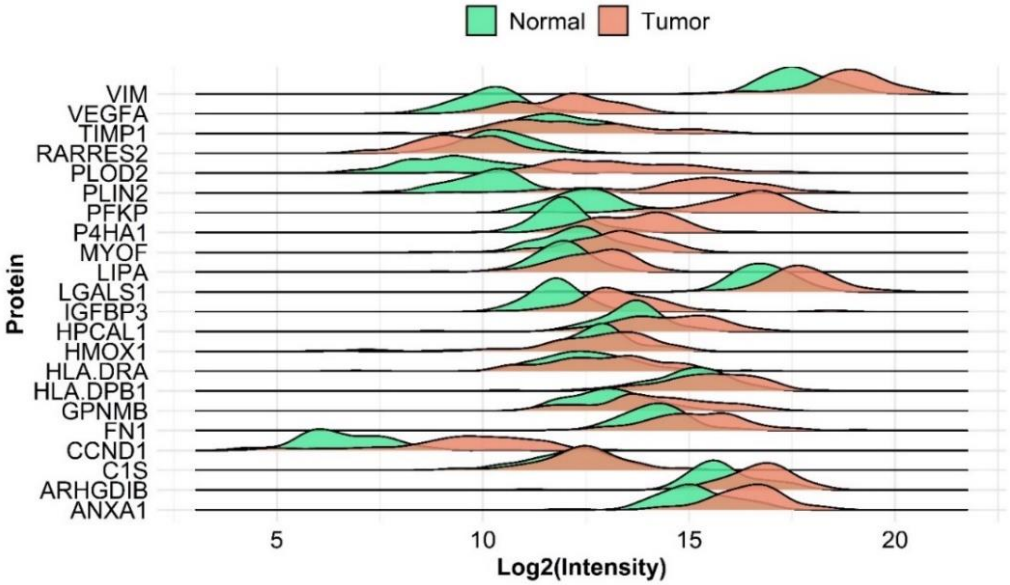
genes. Top differentially expressed genes like VEGFA (FC = 32, p = 1.77E-11) IGFBP3 (FC = 1.56.1, p = 6.24E-09), PFKP (FC = 13.81, p = 4.59E-09), PLIN2 (FC = 46.5, p = 2.82E-11) showed comparable results with the GDC and GEO datasets. Further results of the gene expression changes are presented in **Table 3**.

### 4.2.3 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) shown in **Figure 11.** Differential analysis between male and female patients resulted in no significant differences. Regression analysis of age and protein expression showed a significant result only in the case of IGFBP2, however, the adjusted R-squared value was 0.064. Thus, we can conclude that neither age nor gender can be considered as a covariate factor. Using the clusterProfiler R package, we performed an enrichment analysis; mostly enriched GO terms are connected to migration and adhesion. Results of the enrichment analysis are presented in **Figure 12**. Detailed results of the protein expression changes are also presented in **Table 3**.

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. Genes showing the most significant differential expression in both platforms also presented the highest correlation coefficients, including PLIN2 (R = 0.70), IGFBP3 (R = 0.66), PFKP (R = 0.59), PLOD2 (R = 0.59), CCND1 (R = 0.58) and VEGFA (R = 0.56).

**A**




**B**

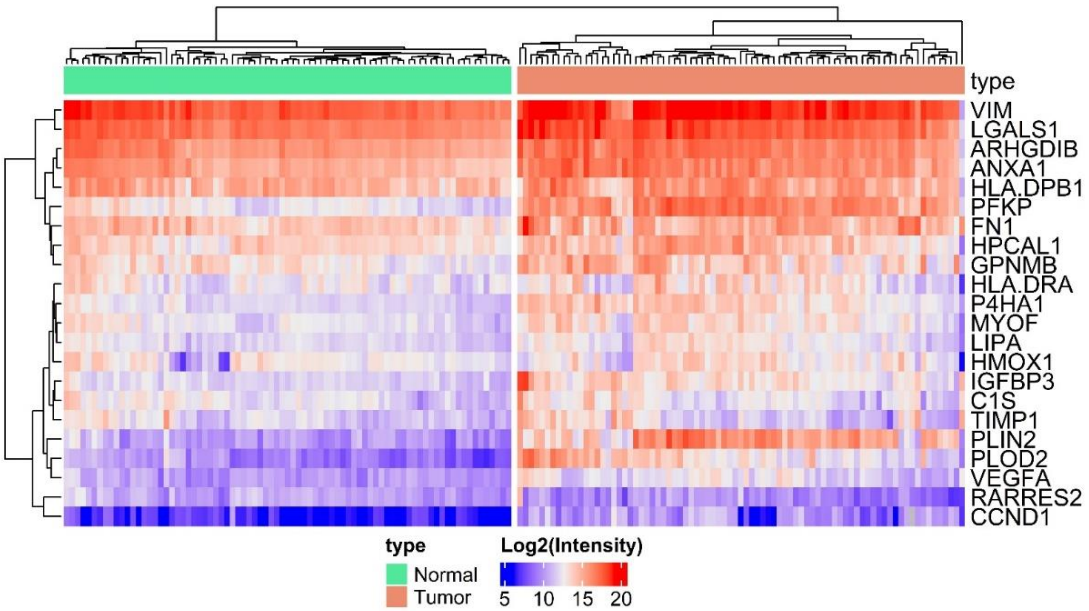



**Figure 11.** Differential protein abundances of compared normal and ccRCC tumor samples. Ridge plots of differentially expressed proteins shows the distribution of log2 intensity values (a). Heatmap of log2 intensity values (b) Source: (94).

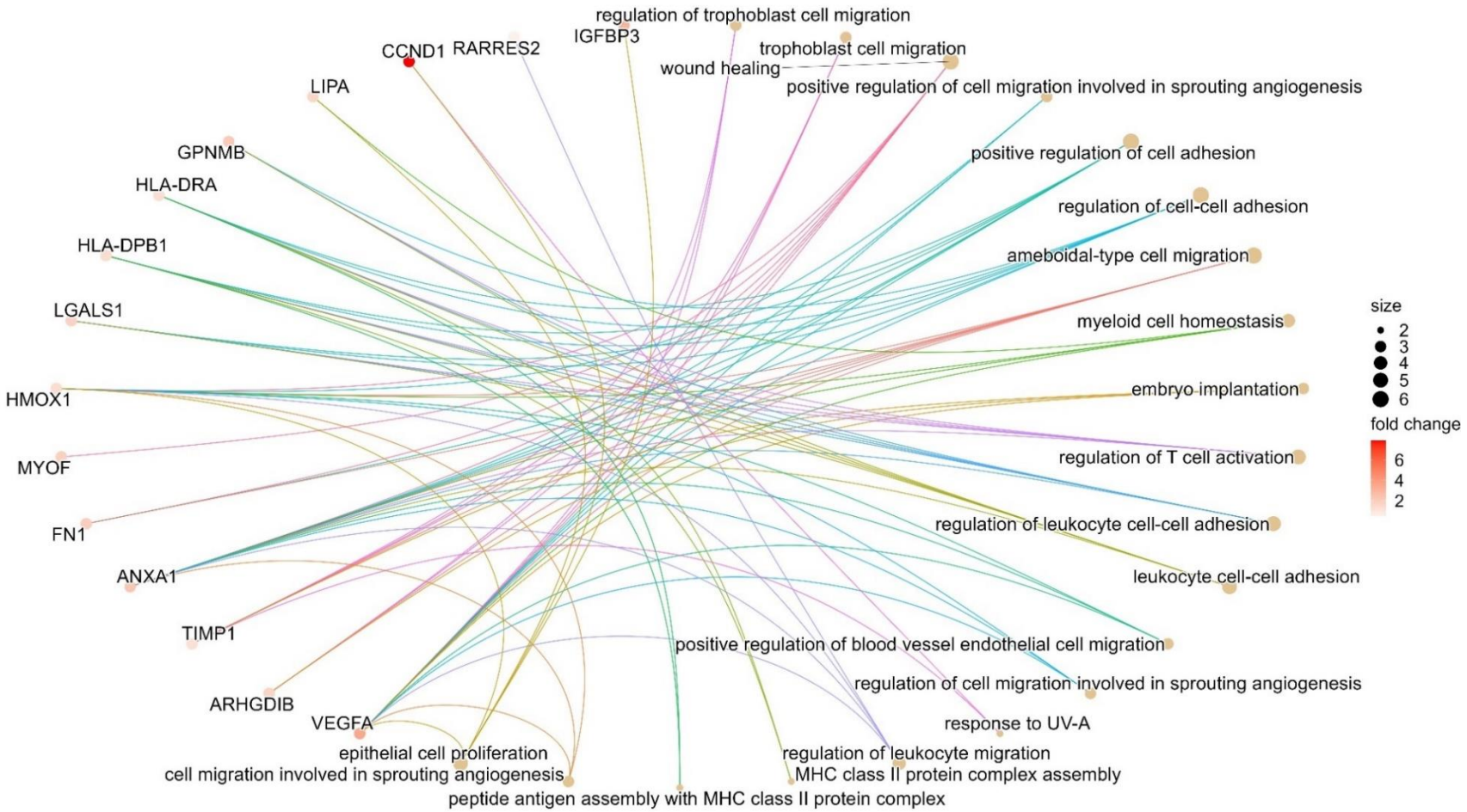


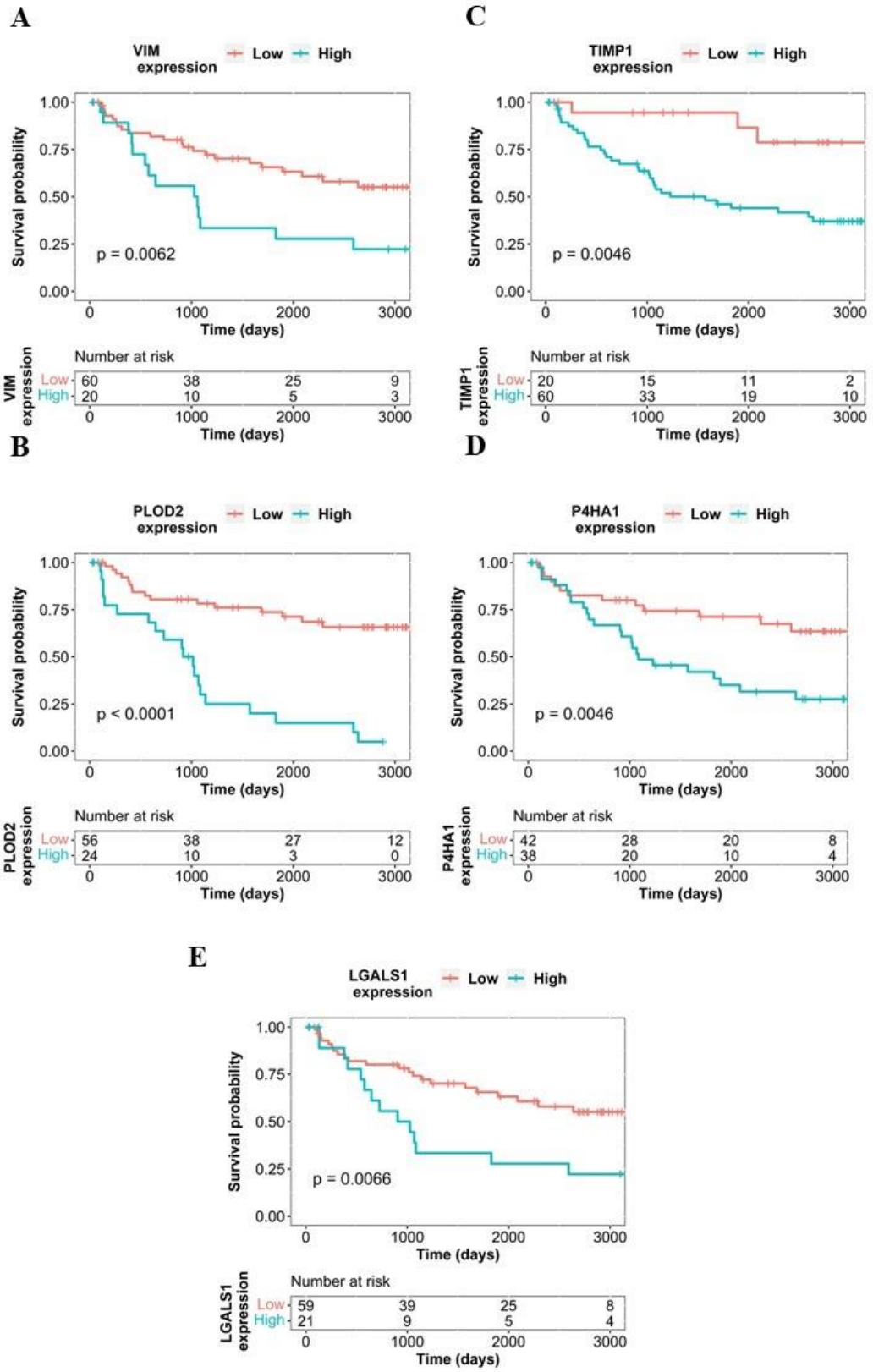

**Figure 12.** Gene ontology of the top genes. Gene ontology (GO) analysis of the strongest genes which discriminate normal kidney and ccRCC in all investigated cohorts. In the Gene-concept network plot (cnet plot) the linkages of genes and biological concepts are presented as a circular- shaped network. The color of the genes represents the FC values, and the size of the GO terms represents the associated genes. Source: (94).

**Table 3.** Summary table of differential expression analysis of the twenty genes reaching significance in all cohorts.

| | GDC cohort | | GEO cohort | | SE-RNA-seq | | SE-MS | |
|---|---|---|---|---|---|---|---|---|
| **SYMBOL** | **FC** | **P-value** | **FC** | **P-value** | **FC** | **P-value** | **FC** | **P-value** |
| ANXA1 | 2.32 | 6.87E-09 | 2.89 | 3.08E-33 | 1.41 | 8.29E-04 | 2.26 | 1.46E-13 |
| ARHGDIB | 2.96 | 1.01E-12 | 3.07 | 1.27E-33 | 0.63 | 1.23E-04 | 1.68 | 4.83E-07 |
| C1S | 2.19 | 7.33E-07 | 3.64 | 2.79E-25 | 11.48 | 3.18E-03 | 1.22 | 1.04E-01 |
| CCND1 | 5.98 | 4.25E-13 | 4.12 | 2.28E-32 | 1.09 | 1.93E-07 | 7.89 | 1.04E-24 |
| FN1 | 3.70 | 1.67E-10 | 5.21 | 1.08E-34 | 1.66 | 1.24E-04 | 1.99 | 2.31E-08 |
| GPNMB | 3.32 | 3.86E-07 | 3.48 | 2.66E-29 | 5.29 | 1.53E-01 | 2.11 | 1.02E-07 |
| HLA-DPB1 | 4.49 | 7.73E-12 | 3.45 | 1.62E-32 | 13.45 | 1.42E-06 | 1.37 | 1.20E-02 |
| HLA-DRA | 4.26 | 6.68E-13 | 3.17 | 5.31E-33 | 6.23 | 1.78E-05 | 1.31 | 5.60E-02 |
| HMOX1 | 5.01 | 5.22E-12 | 2.95 | 5.54E-29 | 51.46 | 1.81E-09 | 1.32 | 8.10E-02 |
| HPCAL1 | 3.21 | 3.35E-11 | 2.86 | 2.43E-32 | 16.96 | 2.16E-07 | 1.75 | 5.33E-06 |
| IGFBP3 | 10.47 | 2.17E-12 | 8.15 | 1.01E-33 | 1.56 | 6.24E-09 | 3.04 | 7.53E-18 |
| LGALS1 | 4.16 | 1.50E-11 | 4.57 | 1.76E-34 | 8.78 | 1.48E-05 | 1.76 | 6.03E-08 |
| LIPA | 2.78 | 1.33E-11 | 3.07 | 2.12E-34 | 12.41 | 5.35E-05 | 1.62 | 7.13E-07 |
| MYOF | 1.86 | 1.16E-07 | 2.86 | 3.86E-34 | 13.95 | 9.74E-06 | 1.87 | 5.39E-08 |
| P4HA1 | 3.78 | 3.17E-13 | 2.96 | 2.71E-34 | 12.84 | 7.70E-07 | 3.15 | 2.30E-22 |
| PFKP | 3.97 | 3.24E-12 | 5.69 | 2.12E-34 | 13.81 | 4.59E-09 | 12.78 | 1.01E-47 |
| PLIN2 | 7.08 | 1.10E-11 | 3.85 | 1.59E-32 | 46.48 | 2.82E-11 | 26.09 | 3.90E-39 |
| PLOD2 | 3.38 | 3.72E-10 | 4.21 | 3.75E-34 | 0.58 | 4.92E-07 | 15.84 | 6.51E-36 |
| RARRES2 | 2.59 | 5.16E-09 | 3.35 | 1.65E-31 | 2.79 | 2.96E-01 | 0.53 | 2.11E-07 |
| TIMP1 | 3.72 | 2.90E-09 | 3.61 | 1.87E-34 | 8.21 | 1.75E-05 | 1.21 | 2.13E-01 |
| VEGFA | 9.03 | 3.04E-13 | 3.02 | 3.09E-32 | 32.00 | 1.77E-11 | 3.49 | 1.40E-22 |
| VIM | 6.82 | 2.92E-13 | 2.88 | 1.68E-33 | 18.01 | 4.04E-10 | 2.06 | 4.09E-08 |

### 4.2.4 Survival Analysis Using Proteome-Level Data

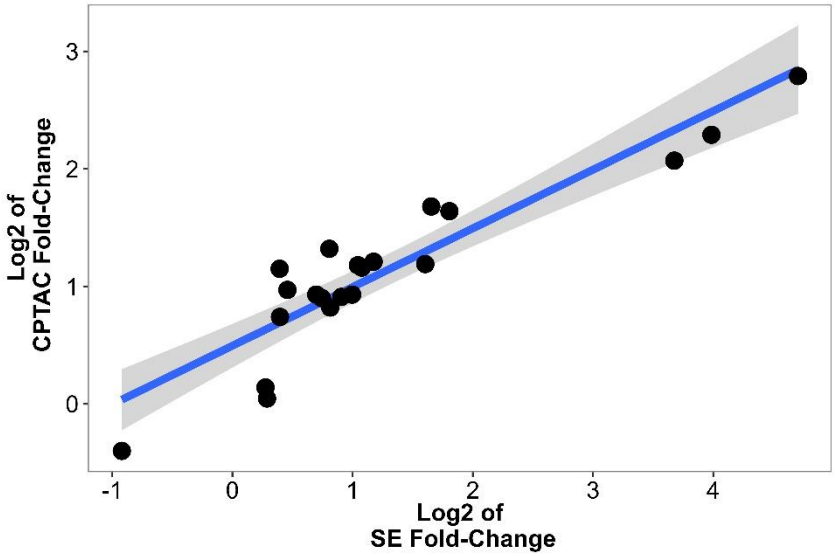
To estimate the potential effects of protein expression on patient survival, we performed a survival analysis using all available proteins. Five out of the investigated proteins showed a correlation with survival. Patients with elevated expression of PLOD2 protein showed significantly worse overall survival 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. Kaplan-Meier curves of genes associated with varying overall survival rates are shown in **Figure 13.**




**13.Figure** Kaplan–Meier plots of VIM(A), PLOD2(B), TIMP1(C), P4HA1(D), LGALS1(E), each protein shows a significant correlation with impaired overall survival. Source: (94).

### 4.2.5 Validation using data from CPTAC.

To further support our analysis, we validated our results using CPTAC data from the study of Clark at al.(64). Out of the 22 proteins identified by our current study, 21 were also available in the CPTAC dataset. The FC values between the two MS analyses had comparable results. Correlation analysis of the log2FC values of the CPTAC and SE cohorts resulted in a significant correlation (R = 0.91, p = 3.7E−9) **Figure 14**. Top proteins identified, such as PLIN2 (FC = 6.92, p = 1.7E−33), PLOD2 (FC = 4.89, p = 7.4E−33), PFKP (FC = 4.2, p = 4.3E−56), IGFBP3 (FC = 2.28, p = 2.1E−31), and VEGFA (FC = 3.12, p = 3E−32), had significant differences between normal kidney and ccRCC in the CPTAC study. Further results are displayed int **Table 4.**




**14. Figure** Correlation analysis of log-transformed CPTAC and SE Fold-change values. Each dot rep-resents a FC value of a protein, we also added a trend line using a linear model. Source: (94).

**Table 4** Summary table of own MS data and CPTAC protein expression differences Source: (94).

| | SE Data MS | | CPTAC Protein Data | |
|---|---|---|---|---|
| | **Fold-Change** | **Adjusted *p*-Value** | **Fold-Change** | **Adjusted *p*-Value** |
| ANXA1 | 2.26 | $1.46 * 10^{-13}$ | 2.31 | $6.60 * 10^{-41}$ |
| ARHGDIB | 1.68 | $4.83 * 10^{-7}$ | 1.87 | $7.10 * 10^{-42}$ |
| C1S | 1.22 | 0.10 | 1.03 | 0.49 |
| FN1 | 1.99 | $2.31 * 10^{-8}$ | 1.91 | $1.90 * 10^{-25}$ |
| GPNMB | 2.11 | $1.02 * 10^{-7}$ | 2.23 | $2.60 * 10^{-17}$ |
| HLA-DPB1 | 1.37 | 0.01 | 1.96 | $3.10 * 10^{-32}$ |
| HLA-DRA | 1.31 | 0.06 | 2.22 | $7.80 * 10^{-36}$ |
| HMOX1 | 1.32 | 0.08 | 1.67 | $1.20 * 10^{-29}$ |
| HPCAL1 | 1.75 | $5.33 * 10^{-6}$ | 2.50 | $5.00 * 10^{-45}$ |
| IGFBP3 | 3.04 | $7.53 * 10^{-18}$ | 2.28 | $2.10 * 10^{-31}$ |
| LGALS1 | 1.76 | $6.03 * 10^{-8}$ | 1.77 | $1.60 * 10^{-33}$ |
| LIPA | 1.62 | $7.13 * 10^{-7}$ | 1.91 | $9.40 * 10^{-31}$ |
| MYOF | 1.87 | $5.39 * 10^{-8}$ | 1.88 | $2.00 * 10^{-39}$ |
| P4HA1 | 3.15 | $2.30 * 10^{-22}$ | 3.20 | $9.90 * 10^{-57}$ |
| PFKP | 12.78 | $1.01 * 10^{-47}$ | 4.20 | $4.30 * 10^{-56}$ |
| PLIN2 | 26.09 | $3.90 * 10^{-39}$ | 6.92 | $1.70 * 10^{-33}$ |
| PLOD2 | 15.84 | $6.51 * 10^{-36}$ | 4.89 | $7.40 * 10^{-33}$ |
| RARRES2 | 0.53 | $2.11 * 10^{-7}$ | 0.76 | $1.20 * 10^{-13}$ |
| TIMP1 | 1.21 | 0.21 | 1.10 | 0.17 |
| VEGFA | 3.49 | $1.40 * 10^{-22}$ | 3.12 | $3.00 * 10^{-32}$ |
| VIM | 2.06 | $4.09 * 10^{-8}$ | 2.27 | $1.70 * 10^{-63}$ |
| CCND1 | 7.89 | $1.04 * 10^{-24}$ | - | - |

### 4.2.6 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).

Results of all four methods (SVM, k-nearest neighbors, random forest, and logistic regression) in test sets are displayed in **Table 5,** the list of optimal genes is provided in **Table 6.**

**Table 5** Summary table of classification accuracy, sensitivity, specificity, and Kappa values in the test set by each applied method. KNN: k-nearest neighbors, RF: random forest, LOGIT: logistic regression, and SVM: support vector machine Source: (94).

| | **RF** | **SVM** | **KNN** | **LOGIT** |
|---|---|---|---|---|
| **Accuracy** | 0.958 | 0.979 | 0.9375 | 0.958 |
| **Kappa** | 0.916 | 0.958 | 0.8750 | 0.916 |
| **Sensitivity** | 0.916 | 0.958 | 0.8750 | 0.916 |
| **Specificity** | 1.0 | 1.0 | 1.0 | 1.0 |

**Table 6** Summary table of ideal gene panels in each algorithm. KNN: k-nearest neighbors, RF: random forest, LOGIT: logistic regression, and SVM: support vector machines. Source: (94).

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **RF** | PFKP | PLOD2 | PLIN2 | | | | | | |
| **SVM** | PFKP | PLIN2 | PLOD2 | IGFBP3 | VEGFA | P4HA1 | CCND1 | VIM | ANXA1 |
| **KNN** | PFKP | PLIN2 | PLOD2 | IGFBP3 | VEGFA | P4HA1 | CCND1 | | |
| **LOGIT** | PFKP | PLIN2 | PLOD2 | | | | | | |

# 5 DISCUSSION

## 5.1 Differential expression analysis of the most malignant cancers

Our most important aim was to establish a framework for the comparison of gene expression in malignant, normal and metastatic tissues. To that end, we established a database from publicly available RNA-seq and gene array resources. Followed by a multistep manual and computational curation, we used the datasets in combination with established statistical algorithms to set up an online analysis platform. Finally, the reproducibility of the results delivered by our approach was validated using a training–test approach with multiple randomly differentiated cohorts in three distinct tumor types. Since all implemented examinations delivered high concordance, we can state that the established database provides solid results in both platforms used.

One of the major features of our approach is the generation of an expression cutoff-based sensitivity/specificity plot. This graphical representation displays a bar graph showing the proportion of tumor samples with elevated expression compared to the normal cohort at selected cutoff values (minimum, first quartile, median, third quartile, maximum).

Since pharmacologically useful targets have to be as specific to the tumor cell as possible, by looking at the graph, one can obtain easily interpretable information regarding the clinical utility of the selected gene. The conventional approach to show sensitivity and specificity would be to generate a receiver operating characteristics (ROC) plot and examine the area under the curve to assess the usefulness of a potential biomarker. Of note, our group also established the www.rocplot.org platform, capable of identifying predictive biomarkers in multiple tumor types by employing ROC analysis (95).

After completing the entire database, our paramount question was: which genes are most specific to cancer across multiple tumor types? We performed a comparative study across the top ten most deadly tumor types and ranked the common genes in these malignancies, regardless of the platform. The most consistently upregulated gene was DNA topoisomerase 2-alpha (TOP2A), a gene playing an important role in transcription and replication. Several studies highlighted the importance of TOP2A, and elevated TOP2A expression can serve as a prognostic biomarker in multiple malignancies, including lung (96), colon (97), and breast cancer (98). At present, multiple drugs, including doxorubicin, epirubicin or etoposide, are widely used in clinical practice to

target TOP2A or other topoisomerase gene products (99). These agents are now used in multiple tumor types, including breast cancer (100), leukemias and lymphomas (101, 102).

The most consistently downregulated gene across the investigated tumor types was Alcohol dehydrogenase 1B (ADH1B), a member of the alcohol dehydrogenase enzyme subgroup which serves as an important member in the ethanol, retinol and further alcoholic substance metabolization processes. In concordance with our results, earlier studies came to a comparable conclusion, as downregulation of ADH1B might have a role in multiple cancers, including colon (103), lung (104) or head and neck cancer (105).

A notable limitation of our study is the low number of available metastatic tissues. Although the total number ($n$ = 848) seems robust, these represent only 1.5% of the included specimens. Unfortunately, this is an open issue not dealt with in any of the large-scale data collection projects.

Current clinical diagnostics of cancer relies mainly on pathological examination using tissue slide staining or immune-histochemistry. The importance of tissue inspection is undoubtful, however, with the increasing burden of workload in pathological diagnostics the need for further potent diagnostic possibilities and tools capable to provide sufficient pathological decision support is necessary. While transcriptome-based methods are useful for this purpose, several studies with promising results were published recently in the proteome field as well. Establishing proteins with differential abundance in malignant samples compared to healthy tissues can provide valuable information in diagnostics and therapeutic target identification. For example, a breast cancer study comparing malignant breast cancer samples to adjacent normal samples using MS identified a novel luminal subtype (106). A comparison of normal prostate and prostate adenocarcinoma samples was performed to identify a new prognostic biomarker (107).

### 5.2 Characteristics of differentially expressed proteins in ccRCC

Like other cancer types, early surgical intervention is the best solution for total recovery in ccRCC as well. Especially in the early stages, when the disease is localized, partial or radical nephrectomy is the most frequently performed treatment option (71). In the present study, by using transcriptomic data we uncovered genes with higher expression in ccRCC and then developed an algorithm capable to identify ccRCC tissues with accuracy high enough for future clinical application. We focused on genes having

higher expression in the tumor tissues. By using targeted MS data of the selected proteins, our algorithm is capable to differentiate between normal and malignant tissues and could provide a valuable decision support during the pathological diagnostic process.

The final discriminative algorithm is based on the differential expression of nine proteins. Of these, VEGFA and CCND1 are well-known cancer biomarkers. VEGFA (vascular endothelial growth factor A) is used as a target molecule in ccRCC treatment (72). CCND1 (cyclin D1), a member of the cyclin family, acts as a regulator of cyclin-dependent kinases (CDKs). CDK inhibitors are widely used in the treatment of breast cancer (108). PLOD2 (procollagen-lysin 2-oxoglutarate 5-dioxygenase) has a role in the maintenance of intermolecular collagen cross-links (109). The aberrant function of PLOD2 might have a role in ovarian cancer (109) and gastric cancer progression (110). PFKP (phosphofructokinase platelet isoform) is responsible for one of the early steps of glycolysis (111). It might also have a crucial part in metabolic reprogramming in multiple cancer types like breast cancer (112) and non-small cell lung cancer (113). IGFBP3 (insulin-like growth factor binding protein 3) acts as a carrier protein of several types of IGF molecules, and it is related to cell growth and differentiation (114). IGFBP3 has been shown to be important in the development of colorectal and breast cancer (114, 115). PLIN2 (perilipin 2) is a member of the perilipin family and takes part in the formation of intracellular lipid storage droplets in multiple tissue types (116). It has been connected to the development of atherosclerosis (117) but it has relevance in cancer initiation and progression as well (116). Using Western blot technique, an earlier study has proposed PLIN2 as a potential plasma biomarker in ccRCC (118). As both IGFP3 and PLIN2 can be detected in the plasma, we hypothesize that they could also serve as potential diagnostic biomarkers of ccRCC. Using our current knowledge, however, we lack any robust evidence for our hypothesis.

By survival analysis, we identified five proteins with a high expression which correlates with poor survival outcomes. Out of these five, PLOD2, VIM, and P4HA1 are also highlighted by our model. Both PLOD2 and P4HA1 are enzymes involved in collagen-related pathways and proved to be a biomarker of epithelial-to-mesenchymal transition (EMT) in multiple types of cancers (119, 120). While vimentin acts as an important structural protein and a known marker of EMT, overexpression of these

proteins in patients with poor survival outcomes implies their involvement in EMT and metastasis formation in renal cell clear carcinoma.

We must note an important limitation of our approach. Although transcriptome-based examinations can provide valuable input of new potential biomarkers, due to mechanisms like alternative splicing, mutations, and post-translational modifications, RNA expression only moderately correlates with protein expression as shown both in our and further studies (121). A further limitation of our model is the incapability of tumor stage estimation, as staging is usually based on imaging, pathological examination, and further clinical characteristics.

In summary, we established the largest currently available transcriptomic cancer database, consisting of nearly 57,000 samples, by utilizing multiple RNA-seq and microarray datasets. We show that the results obtained by these specimens are highly reproducible, and we have set up an online analysis portal which enables mining of the database for any gene to assess expression differences in normal, cancer and metastatic samples (91).We further validated our results on the proteome level using a set of renal samples of paired normal and tumor tissues to identify biomarkers differentiating renal clear cell cancer (ccRCC) and normal kidney tissues. With a support vector machine-based machine learning algorithm using nine genes, we set up a model which can differentiate between normal and malignant ccRCC tissues using proteomic data. Finally, a set of proteins showed a significant correlation with poor survival outcomes and might serve as potential biomarkers of progression (94).

# 6 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.

# 7 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.

# 8 REFERENCES


1. Zhang L, Zhou W, Velculescu VE, Kern SE, Hruban RH, Hamilton SR, et al. Gene expression profiles in normal and cancer cells. Science. 1997;276(5316):1268-72.
2. Druker BJ, Tamura S, Buchdunger E, Ohno S, Segal GM, Fanning S, et al. Effects of a selective inhibitor of the Abl tyrosine kinase on the growth of Bcr-Abl positive cells. Nat Med. 1996;2(5):561-6.
3. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell. 2000;100(1):57-70.
4. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646-74.
5. Davies MA, Samuels Y. Analysis of the genome to personalize therapy for melanoma. Oncogene. 2010;29(41):5545-55.
6. Burkhart DL, Sage J. Cellular mechanisms of tumour suppression by the retinoblastoma gene. Nature reviews Cancer. 2008;8(9):671-82.
7. Junttila MR, Evan GI. p53--a Jack of all trades but master of none. Nature reviews Cancer. 2009;9(11):821-9.
8. Rich T, Watson CJ, Wyllie A. Apoptosis: the germs of death. Nat Cell Biol. 1999;1(3):E69-71.
9. Blasco MA. Telomeres and human disease: ageing, cancer and beyond. Nat Rev Genet. 2005;6(8):611-22.
10. De Palma M, Biziato D, Petrova TV. Microenvironmental regulation of tumour angiogenesis. Nature reviews Cancer. 2017;17(8):457-74.
11. Berx G, van Roy F. Involvement of members of the cadherin superfamily in cancer. Cold Spring Harb Perspect Biol. 2009;1(6):a003129.
12. Vander Heiden MG, Cantley LC, Thompson CB. Understanding the Warburg effect: the metabolic requirements of cell proliferation. Science. 2009;324(5930):1029-33.
13. Semenza GL. Tumor metabolism: cancer cells give and take lactate. J Clin Invest. 2008;118(12):3835-7.
14. Jones RG, Thompson CB. Tumor suppressors and cell metabolism: a recipe for cancer growth. Genes Dev. 2009;23(5):537-48.

15. DeBerardinis RJ, Lum JJ, Hatzivassiliou G, Thompson CB. The biology of cancer: metabolic reprogramming fuels cell growth and proliferation. Cell Metab. 2008;7(1):11-20.
16. Kovacs SA, Gyorffy B. Transcriptomic datasets of cancer patients treated with immune-checkpoint inhibitors: a systematic review. J Transl Med. 2022;20(1):249.
17. Kallingal A, Olszewski M, Maciejewska N, Brankiewicz W, Baginski M. Cancer immune escape: the role of antigen presentation machinery. Journal of cancer research and clinical oncology. 2023.
18. Messerschmidt JL, Prendergast GC, Messerschmidt GL. How Cancers Escape Immune Destruction and Mechanisms of Action for the New Significantly Active Immune Therapies: Helping Nonimmunologists Decipher Recent Advances. The oncologist. 2016;21(2):233-43.
19. Alfonso JCL, Papaxenopoulou LA, Mascheroni P, Meyer-Hermann M, Hatzikirou H. On the Immunological Consequences of Conventionally Fractionated Radiotherapy. iScience. 2020;23(3):100897.
20. Hanahan D. Hallmarks of Cancer: New Dimensions. Cancer discovery. 2022;12(1):31-46.
21. Fusco G, Minelli A. Phenotypic plasticity in development and evolution: facts and concepts. Introduction. Philos Trans R Soc Lond B Biol Sci. 2010;365(1540):547-56.
22. Yuan S, Norgard RJ, Stanger BZ. Cellular Plasticity in Cancer. Cancer discovery. 2019;9(7):837-51.
23. Huang S. Tumor progression: chance and necessity in Darwinian and Lamarckian somatic (mutationless) evolution. Prog Biophys Mol Biol. 2012;110(1):69-86.
24. Michealraj KA, Kumar SA, Kim LJY, Cavalli FMG, Przelicki D, Wojcik JB, et al. Metabolic Regulation of the Epigenome Drives Lethal Infantile Ependymoma. Cell. 2020;181(6):1329-45 e24.
25. Thomas S, Izard J, Walsh E, Batich K, Chongsathidkiet P, Clarke G, et al. The Host Microbiome Regulates and Maintains Human Health: A Primer and Perspective for Non-Microbiologists. Cancer Res. 2017;77(8):1783-812.

26. Sears CL, Garrett WS. Microbes, microbiota, and colon cancer. Cell Host Microbe. 2014;15(3):317-28.
27. Swaney MH, Kalan LR. Living in Your Skin: Microbes, Molecules, and Mechanisms. Infect Immun. 2021;89(4).
28. Healy CM, Moran GP. The microbiome and oral cancer: More questions than answers. Oral Oncol. 2019;89:30-3.
29. Xu J, Peng JJ, Yang W, Fu K, Zhang Y. Vaginal microbiomes and ovarian cancer: a review. Am J Cancer Res. 2020;10(3):743-56.
30. De Blander H, Morel AP, Senaratne AP, Ouzounova M, Puisieux A. Cellular Plasticity: A Route to Senescence Exit and Tumorigenesis. Cancers (Basel). 2021;13(18).
31. Lee S, Schmitt CA. The dynamic nature of senescence in cancer. Nat Cell Biol. 2019;21(1):94-101.
32. Wang B, Kohli J, Demaria M. Senescent Cells in Cancer Therapy: Friends or Foes? Trends Cancer. 2020;6(10):838-57.
33. Menyhart O, Harami-Papp H, Sukumar S, Schafer R, Magnani L, de Barrios O, et al. Guidelines for the selection of functional assays to evaluate the hallmarks of cancer. Biochim Biophys Acta. 2016;1866(2):300-19.
34. Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T. Transcriptomics technologies. PLoS Comput Biol. 2017;13(5):e1005457.
35. Auer H, Newsom DL, Kornacker K. Expression Profiling Using Affymetrix GeneChip Microarrays. Methods Mol Biol. 2009;509:35-46.
36. Ragoussis J, Elvidge G. Affymetrix GeneChip system: moving from research to the clinic. Expert Rev Mol Diagn. 2006;6(2):145-52.
37. Takahashi Y, Nagata T, Nakayama T, Ishii Y, Ishikawa K, Asai S. [GeneChip system from a bioinformatical point of view]. Nihon Yakurigaku Zasshi. 2002;120(2):73-84.
38. Dalma-Weiszhausz DD, Warrington J, Tanimoto EY, Miyada CG. The affymetrix GeneChip platform: an overview. Methods Enzymol. 2006;410:3-28.
39. Miller MB, Tang YW. Basic concepts of microarrays and potential applications in clinical microbiology. Clin Microbiol Rev. 2009;22(4):611-33.

40. Nookaew I, Papini M, Pornputtapong N, Scalcinati G, Fagerberg L, Uhlen M, et al. A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. Nucleic acids research. 2012;40(20):10084-97.

41. Zhang W, Yu Y, Hertwig F, Thierry-Mieg J, Zhang W, Thierry-Mieg D, et al. Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. Genome biology. 2015;16(1):133.

42. Kukurba KR, Montgomery SB. RNA Sequencing and Analysis. Cold Spring Harb Protoc. 2015;2015(11):951-69.

43. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome biology. 2016;17:13.

44. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10(1):57-63.

45. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. Nat Rev Genet. 2019;20(11):631-56.

46. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45(10):1113-20.

47. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, et al. Toward a Shared Vision for Cancer Genomic Data. N Engl J Med. 2016;375(12):1109-12.

48. Consortium GT. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45(6):580-5.

49. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids research. 2002;30(1):207-10.

50. Han X, Aslanian A, Yates JR, 3rd. Mass spectrometry for proteomics. Curr Opin Chem Biol. 2008;12(5):483-90.

51. Pedde RD, Li H, Borchers CH, Akbari M. Microfluidic-Mass Spectrometry Interfaces for Translational Proteomics. Trends Biotechnol. 2017;35(10):954-70.

52. Dapic I, Baljeu-Neuman L, Uwugiaren N, Kers J, Goodlett DR, Corthals GL. Proteome analysis of tissues by mass spectrometry. Mass Spectrom Rev. 2019;38(4-5):403-41.
53. Bensimon A, Heck AJ, Aebersold R. Mass spectrometry-based proteomics and network biology. Annu Rev Biochem. 2012;81:379-405.
54. Zhang Y, Fonslow BR, Shan B, Baek MC, Yates JR, 3rd. Protein analysis by shotgun/bottom-up proteomics. Chem Rev. 2013;113(4):2343-94.
55. Scherl A. Clinical protein mass spectrometry. Methods. 2015;81:3-14.
56. Zhang B, Whiteaker JR, Hoofnagle AN, Baird GS, Rodland KD, Paulovich AG. Clinical potential of mass spectrometry-based proteogenomics. Nat Rev Clin Oncol. 2019;16(4):256-68.
57. Shimizu H, Jinno F, Morohashi A, Yamazaki Y, Yamada M, Kondo T, et al. Application of high-resolution ESI and MALDI mass spectrometry to metabolite profiling of small interfering RNA duplex. J Mass Spectrom. 2012;47(8):1015-22.
58. Calderon-Celis F, Encinar JR, Sanz-Medel A. Standardization approaches in absolute quantitative proteomics with mass spectrometry. Mass Spectrom Rev. 2018;37(6):715-37.
59. Shackleton C. Clinical steroid mass spectrometry: a 45-year history culminating in HPLC-MS/MS becoming an essential tool for patient diagnosis. The Journal of steroid biochemistry and molecular biology. 2010;121(3-5):481-90.
60. Recent Advances in the Clinical Application of Mass Spectrometry. Ejifcc. 2016;27(4):264-71.
61. Jannetto PJ, Fitzgerald RL. Effective Use of Mass Spectrometry in the Clinical Laboratory. Clin Chem. 2016;62(1):92-8.
62. Krug K, Jaehnig EJ, Satpathy S, Blumenberg L, Karpova A, Anurag M, et al. Proteogenomic Landscape of Breast Cancer Tumorigenesis and Targeted Therapy. Cell. 2020;183(5):1436-56 e31.
63. Vasaikar S, Huang C, Wang X, Petyuk VA, Savage SR, Wen B, et al. Proteogenomic Analysis of Human Colon Cancer Reveals New Therapeutic Opportunities. Cell. 2019;177(4):1035-49 e19.

64. Clark DJ, Dhanasekaran SM, Petralia F, Pan J, Song X, Hu Y, et al. Integrated Proteogenomic Characterization of Clear Cell Renal Cell Carcinoma. Cell. 2019;179(4):964-83 e31.
65. Rodriguez H, Zenklusen JC, Staudt LM, Doroshow JH, Lowy DR. The next horizon in precision oncology: Proteogenomics to inform cancer diagnosis and treatment. Cell. 2021;184(7):1661-70.
66. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. CA: a cancer journal for clinicians. 2023;73(1):17-48.
67. Chow WH, Dong LM, Devesa SS. Epidemiology and risk factors for kidney cancer. Nat Rev Urol. 2010;7(5):245-57.
68. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians. 2021.
69. Motzer RJ, Jonasch E, Boyle S, Carlo MI, Manley B, Agarwal N, et al. NCCN Guidelines Insights: Kidney Cancer, Version 1.2021. Journal of the National Comprehensive Cancer Network : JNCCN. 2020;18(9):1160-70.
70. Hsieh JJ, Purdue MP, Signoretti S, Swanton C, Albiges L, Schmidinger M, et al. Renal cell carcinoma. Nat Rev Dis Primers. 2017;3:17009.
71. Campbell S, Uzzo RG, Allaf ME, Bass EB, Cadeddu JA, Chang A, et al. Renal Mass and Localized Renal Cancer: AUA Guideline. J Urol. 2017;198(3):520-9.
72. Choueiri TK, Kaelin WG, Jr. Targeting the HIF2-VEGF axis in renal cell carcinoma. Nat Med. 2020;26(10):1519-30.
73. Battelli C, Cho DC. mTOR inhibitors in renal cell carcinoma. Therapy. 2011;8(4):359-67.
74. Choueiri TK, Motzer RJ. Systemic Therapy for Metastatic Renal-Cell Carcinoma. N Engl J Med. 2017;376(4):354-66.
75. Buchbinder EI, Dutcher JP, Daniels GA, Curti BD, Patel SP, Holtan SG, et al. Therapy with high-dose Interleukin-2 (HD IL-2) in metastatic melanoma and renal cell carcinoma following PD1 or PDL1 inhibition. J Immunother Cancer. 2019;7(1):49.
76. Powles T, Tomczak P, Park SH, Venugopal B, Ferguson T, Symeonides SN, et al. Pembrolizumab versus placebo as post-nephrectomy adjuvant therapy for clear cell renal cell carcinoma (KEYNOTE-564): 30-month follow-up analysis of a multicentre,

randomised, double-blind, placebo-controlled, phase 3 trial. The Lancet Oncology. 2022;23(9):1133-44.

77. Gautier L, Cope L, Bolstad BM, Irizarry RA. affy--analysis of Affymetrix GeneChip data at the probe level. Bioinformatics. 2004;20(3):307-15.

78. Li Q, Birkbak NJ, Gyorffy B, Szallasi Z, Eklund AC. Jetset: selecting the optimal microarray probe set to represent a gene. BMC Bioinformatics. 2011;12:474.

79. Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag; 2016.

80. Winston Chang JC, JJ Allaire, Yihui Xie, Jonathan McPherson. shiny: Web Application Framework for R 2019 [R package version 1.4.0]. Available from: https://CRAN.R-project.org/package=shiny.

81. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. CA: a cancer journal for clinicians. 2020;70(1):7-30.

82. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30(7):923-30.

83. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21.

84. Boekel J, Chilton JM, Cooke IR, Horvatovich PL, Jagtap PD, Kall L, et al. Multi-omic data analysis using Galaxy. Nature biotechnology. 2015;33(2):137-9.

85. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology. 2014;15(12):550.

86. Gallien S, Kim SY, Domon B. Large-Scale Targeted Proteomics Using Internal Standard Triggered-Parallel Reaction Monitoring (IS-PRM). Mol Cell Proteomics. 2015;14(6):1630-44.

87. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16(5):284-7.

88. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 2016;32(18):2847-9.

89. Kuhn M. The caret Package. Journal of Statistical Software. 2012;28.

90. Kuhn M. Building Predictive Models in R Using the caret Package. Journal of Statistical Software; Vol 1, Issue 5 (2008). 2008.

91. 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).
92. Pagès H CM, Falcon S, Li N. AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor. [R package version 1.48.0.]. In press 2019.
93. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nature protocols. 2009;4(8):1184-91.
94. 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).
95. Fekete JT, Gyorffy B. ROCplot.org: Validating predictive biomarkers of chemotherapy/hormonal therapy/anti-HER2 therapy using transcriptomic data of 3,104 breast cancer patients. Int J Cancer. 2019;145(11):3140-51.
96. Kou F, Sun H, Wu L, Li B, Zhang B, Wang X, et al. TOP2A Promotes Lung Adenocarcinoma Cells' Malignant Progression and Predicts Poor Prognosis in Lung Adenocarcinoma. J Cancer. 2020;11(9):2496-508.
97. Zhang R, Xu J, Zhao J, Bai JH. Proliferation and invasion of colon cancer cells are suppressed by knockdown of TOP2A. J Cell Biochem. 2018;119(9):7256-63.
98. An X, Xu F, Luo R, Zheng Q, Lu J, Yang Y, et al. The prognostic significance of topoisomerase II alpha protein in early stage luminal breast cancer. BMC Cancer. 2018;18(1):331.
99. Delgado JL, Hsieh CM, Chan NL, Hiasa H. Topoisomerases as anticancer targets. Biochem J. 2018;475(2):373-98.
100. Jasra S, Anampa J. Anthracycline Use for Early Stage Breast Cancer in the Modern Era: a Review. Curr Treat Options Oncol. 2018;19(6):30.
101. Hallek M. Chronic lymphocytic leukemia: 2020 update on diagnosis, risk stratification and treatment. Am J Hematol. 2019;94(11):1266-87.
102. Cederleuf H, Bjerregard Pedersen M, Jerkeman M, Relander T, d'Amore F, Ellin F. The addition of etoposide to CHOP is associated with improved outcome in ALK+ adult anaplastic large cell lymphoma: A Nordic Lymphoma Group study. Br J Haematol. 2017;178(5):739-46.

103. Kropotova ES, Zinovieva OL, Zyryanova AF, Dybovaya VI, Prasolov VS, Beresten SF, et al. Altered expression of multiple genes involved in retinoic acid biosynthesis in human colorectal cancer. Pathol Oncol Res. 2014;20(3):707-17.
104. Wang P, Zhang L, Huang C, Huang P, Zhang J. Distinct Prognostic Values of Alcohol Dehydrogenase Family Members for Non-Small Cell Lung Cancer. Med Sci Monit. 2018;24:3578-90.
105. Lan J, Huang HY, Lee SW, Chen TJ, Tai HC, Hsu HP, et al. TOP2A overexpression as a poor prognostic factor in patients with nasopharyngeal carcinoma. Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine. 2014;35(1):179-87.
106. Yanovich G, Agmon H, Harel M, Sonnenblick A, Peretz T, Geiger T. Clinical Proteomics of Breast Cancer Reveals a Novel Layer of Breast Cancer Classification. Cancer Res. 2018;78(20):6001-10.
107. Iglesias-Gato D, Wikstrom P, Tyanova S, Lavallee C, Thysell E, Carlsson J, et al. The Proteome of Primary Prostate Cancer. Eur Urol. 2016;69(5):942-52.
108. O'Leary B, Finn RS, Turner NC. Treating cancer with selective CDK4/6 inhibitors. Nat Rev Clin Oncol. 2016;13(7):417-30.
109. Guo T, Gu C, Li B, Xu C. PLODs are overexpressed in ovarian cancer and are associated with gap junctions via connexin 43. Lab Invest. 2021;101(5):564-9.
110. Kiyozumi Y, Iwatsuki M, Kurashige J, Ogata Y, Yamashita K, Koga Y, et al. PLOD2 as a potential regulator of peritoneal dissemination in gastric cancer. Int J Cancer. 2018;143(5):1202-11.
111. Webb BA, Forouhar F, Szu FE, Seetharaman J, Tong L, Barber DL. Structures of human phosphofructokinase-1 and atomic basis of cancer-associated mutations. Nature. 2015;523(7558):111-4.
112. Moon JS, Kim HE, Koh E, Park SH, Jin WJ, Park BW, et al. Kruppel-like factor 4 (KLF4) activates the transcription of the gene for the platelet isoform of phosphofructokinase (PFKP) in breast cancer. J Biol Chem. 2011;286(27):23808-16.
113. Wang F, Li L, Zhang Z. Platelet isoform of phosphofructokinase promotes aerobic glycolysis and the progression of nonsmall cell lung cancer. Molecular medicine reports. 2021;23(1).

114. Jin L, Shen F, Weinfeld M, Sergi C. Insulin Growth Factor Binding Protein 7 (IGFBP7)-Related Cancer and IGFBP3 and IGFBP7 Crosstalk. Front Oncol. 2020;10:727.
115. Chan YX, Alfonso H, Paul Chubb SA, Ho KKY, Gerard Fegan P, Hankey GJ, et al. Higher IGFBP3 is associated with increased incidence of colorectal cancer in older men independently of IGF1. Clin Endocrinol (Oxf). 2018;88(2):333-40.
116. Conte M, Santoro A, Collura S, Martucci M, Battista G, Bazzocchi A, et al. Circulating perilipin 2 levels are associated with fat mass, inflammatory and metabolic markers and are higher in women than men. Aging (Albany NY). 2021;13(6):7931-42.
117. Pisano E, Pacifico L, Perla FM, Liuzzo G, Chiesa C, Lavorato M, et al. Upregulated monocyte expression of PLIN2 is associated with early arterial injury in children with overweight/obesity. Atherosclerosis. 2021;327:68-75.
118. Morrissey JJ, Mobley J, Figenshau RS, Vetter J, Bhayani S, Kharasch ED. Urine aquaporin 1 and perilipin 2 differentiate renal carcinomas from other imaged renal masses and bladder and prostate cancer. Mayo Clin Proc. 2015;90(1):35-42.
119. Xu WH, Xu Y, Wang J, Tian X, Wu J, Wan FN, et al. Procollagen-lysine, 2-oxoglutarate 5-dioxygenases 1, 2, and 3 are potential prognostic indicators in patients with clear cell renal cell carcinoma. Aging (Albany NY). 2019;11(16):6503-21.
120. Zhu X, Liu S, Yang X, Wang W, Shao W, Ji T. P4HA1 as an unfavorable prognostic marker promotes cell migration and invasion of glioblastoma via inducing EMT process under hypoxia microenvironment. Am J Cancer Res. 2021;11(2):590-617.
121. Wang D, Eraslan B, Wieland T, Hallstrom B, Hopf T, Zolg DP, et al. A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. Mol Syst Biol. 2019;15(2):e8503.

# 9 BIBLIOGRAPHY OF THE CANDIDATE'S PUBLICATIONS

## 9.1 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).

## 9.2 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

## 10 ACKNOWLEDGEMENTS

I would like to express my special thanks of gratitude to my supervisor, Prof. Balázs Győrffy, for the guidance, and advice he has provided throughout my time as his student. I have been fortunate to have a supervisor who took a good care of my work and personal development, and who responded to my questions and queries promptly.

I would like to extend my gratitude to all the former and current members of the Department of Bioinformatics whose continuous support helped me through my doctoral years. Special thanks to Gyöngyi Munkácsy and Otília Menyhart for their assistance help and advice during my projects, and to my fellow PhD student colleagues Szonja Kovács and Dalma Müller for their support.

I would like to thank Prof. Dr. Attila Szabó, Director of the Department of Pediatrics and Prof. Dr. Gábor Kovács, former Director of the II. Department of Pediatrics, for the opportunity to conduct my research at the Department of Pediatrics.

I am also grateful to my scientific collaborators – Zsuzsanna Darula, Éva Klement and Prof. Péter Nyírády - whose work was indispensable in my research.

I would also like to give special thanks to my family and my friends for their continuous support and understanding my devotion to my research.