International Journal of
Molecular Sciences

MDPI

Article

TNMplot.com: A Web Tool for the Comparison of Gene Expression in Normal, Tumor and Metastatic Tissues

Áron Bartha [1,2,3] and Balázs Győrffy [1,2,3,*]

[1] Department of Bioinformatics, Semmelweis University, 1094 Budapest, Hungary; gyorffylab@gmail.com
[2] Momentum Cancer Biomarker Research Group, Research Centre for Natural Sciences, 1117 Budapest, Hungary
[3] 2nd Department of Pediatrics, Semmelweis University, 1094 Budapest, Hungary
* Correspondence: gyorffy.balazs@med.semmelweis-univ.hu; Tel.: +3630-514-2822

Abstract: 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 the Gene Expression Omnibus of the National Center for Biotechnology Information (NCBI-GEO) or RNA-seq from The Cancer Genome Atlas (TCGA), Therapeutically Applicable Research to Generate Effective Treatments (TARGET), and The Genotype-Tissue Expression (GTEx) repositories. The altered expression within different platforms was analyzed separately. Statistical significance was computed using Mann–Whitney or Kruskal–Wallis tests. False Discovery Rate (FDR) was computed using the Benjamini–Hochberg method. The entire database contains 56,938 samples, including 33,520 samples from 3180 gene chip-based studies (453 metastatic, 29,376 tumorous and 3691 normal samples), 11,010 samples from TCGA (394 metastatic, 9886 tumorous and 730 normal), 1193 samples from TARGET (1 metastatic, 1180 tumorous and 12 normal) and 11,215 normal samples from GTEx. The most consistently upregulated genes across multiple tumor types were TOP2A (FC = 7.8), SPP1 (FC = 7.0) and CENPA (FC = 6.03), and the most consistently downregulated gene was ADH1B (FC = 0.15). Validation of differential expression using equally sized training and test sets confirmed the reliability of the database in breast, colon, and lung cancer at an FDR below 10%. The online analysis platform enables unrestricted mining of the database and is accessible at TNMplot.com.

Keywords: cancer; transcriptomics; gene array; RNA-seq; differential expression

Citation: Bartha, Á.; Győrffy, B. TNMplot.com: A Web Tool for the Comparison of Gene Expression in Normal, Tumor and Metastatic Tissues. *Int. J. Mol. Sci.* **2021**, *22*, 2622. https://doi.org/10.3390/ijms22052622

Academic Editors: Henry Yang

Received: 7 January 2021
Accepted: 28 February 2021
Published: 5 March 2021

1. Introduction

Cancer emerges as 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 targets 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—among others—resistance to cell death and activating invasion and metastasis. Various experimental methods capable of inspecting these hallmark genes have been reviewed previously [4]. 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 [5].

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) [6] and the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) [7]. The Genotype-Tissue Expression (GTEx) repository makes RNA sequencing, exome sequencing and whole genomic data available for the same patient [8]. 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 [9]. A common feature of these repositories is the provision of raw data in addition to processed and aggregated results.

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. Our first aim was to establish an integrated database of a significant number of normal and tumor samples with transcriptome-level gene expression data. We sought to establish a database which includes both adult and pediatric cases and both RNA-seq and gene array datasets. Our second goal was to validate the reliability of the database by employing a training–test approach to identify genes showing differential expression in selected tumor types. Finally, we designed 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. Results

2.1. 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 3691 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, 9886 cancerous and 394 metastatic specimens from adult cancer patients. We also added 1193 pediatric related data from GDC, consisting of 12 normal, 1180 cancerous, and 1 metastatic samples. In order to increase the number of normal samples, we included a further 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.

	Manual Screening		Computational Screening			Result	T	N	M
NCBI GEO	GSE screened: 3180 datasets	Primary tissue series n = 554 (38,897 Samples)	Data cleaning	MAS5 [10] normalization and scaling	JetSet [11] Annotation	38,431 Samples 38 tumor types	29,376	3691	453
TARGET	1193 samples	-	Data cleaning	DESeq2 [12] normalization and scaling	AnnotationDBI [13] annotation	1193 samples 7 tumor types	1180	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	9886	730	394
GTEx	11,688 samples	Removal of non-primary tissues	Data cleaning	DESeq2 normalization and scaling	biomaRt [14] an-dAnnotationDBI annotation	11,215 samples 51 tumor types	-	11,215	-

2.2. 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 registration-free analysis portal can be accessed at www.tnmplot.com and has three separate analysis options. The pan-cancer analysis tool compares normal and tumorous samples across 22 tissue types simultaneously. This RNA-seq-based rapid analysis serves as explanatory data to furnish comparative information for a selected gene. A representative boxplot of pan-cancer analysis is displayed in Figure 1.

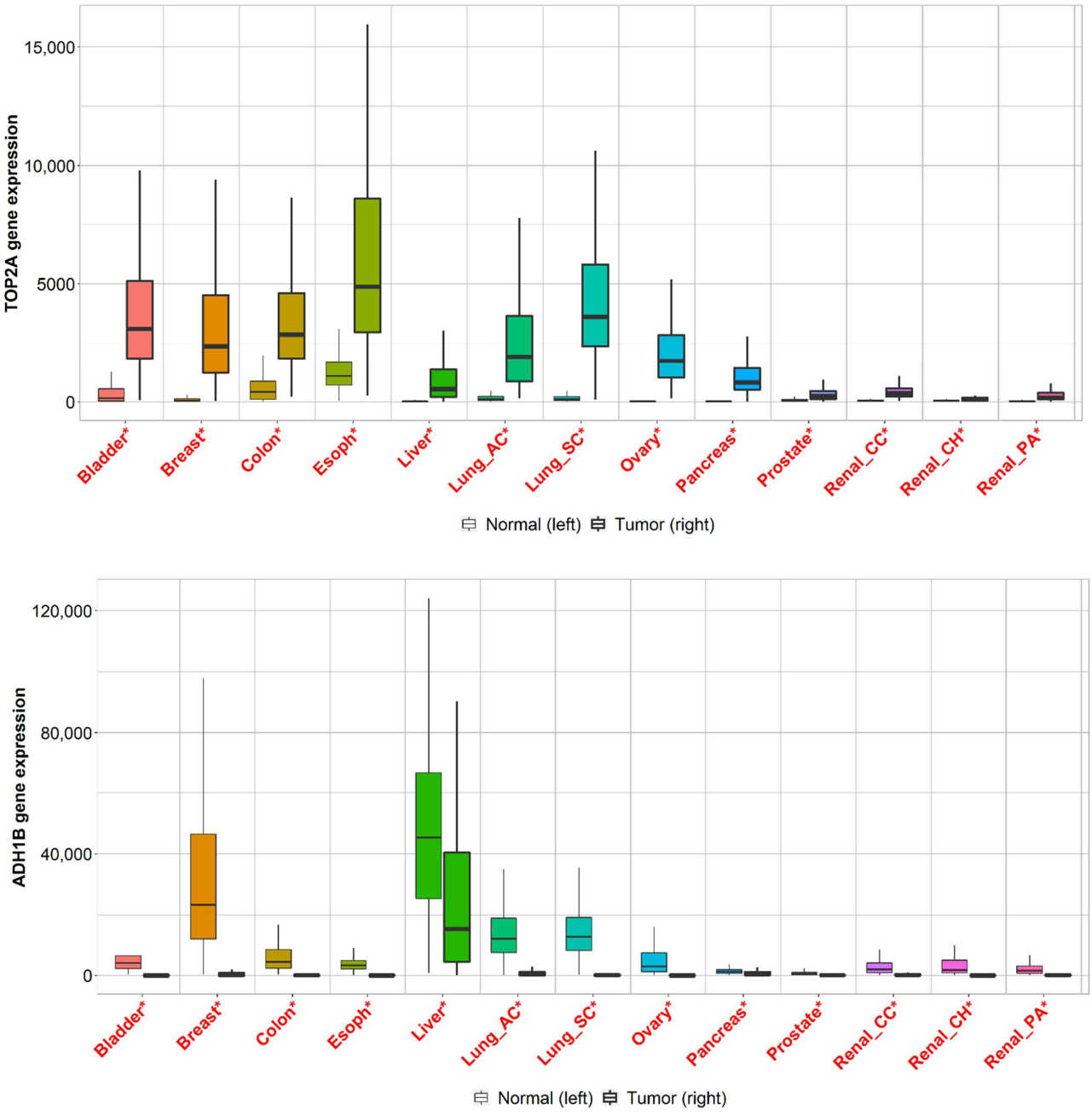

Figure 1. Boxplots of top two genes differentially expressed in most of the ten common tumor types. Significant differences by a Mann–Whitney U test are marked with red color (* $p < 0.01$).

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.

Although the number of metastatic samples is limited in most cases, five and twelve tissue types in the RNA-seq and gene array databases have a useful number of specimens. The third feature of the analysis platform allows us to simultaneously compare these tumor, normal and metastatic data using a Kruskal–Wallis test.

2.3. 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 cancer. In the gene array dataset, 555–2623 reached statistical significance at False Discovery Rate (FDR) <10% and fold change over 1.5. The entire list of all genes is presented in Supplemental Table S2. When using the RNA-seq cohort, 3189–12,037 genes were dysregulated at FDR <10% and fold change over 1.5. The entire list of all genes dysregulated in the RNA seq cohorts is presented in Supplemental Table S3.

2.4. Linking the Most Significant Genes to Cancer Hallmarks

We 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. The overlapping 55 genes are listed in Table 2.

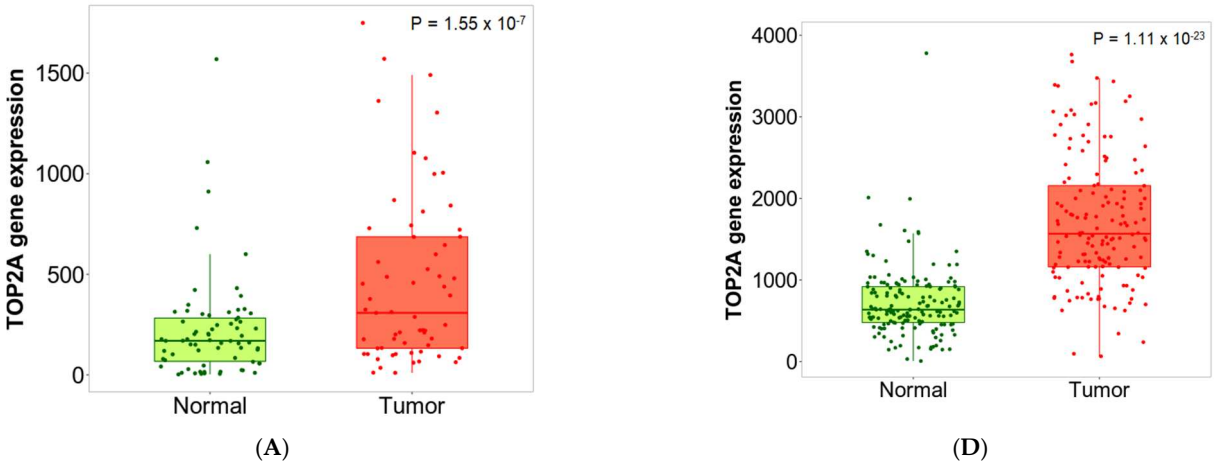

Figure 2. *Cont.*

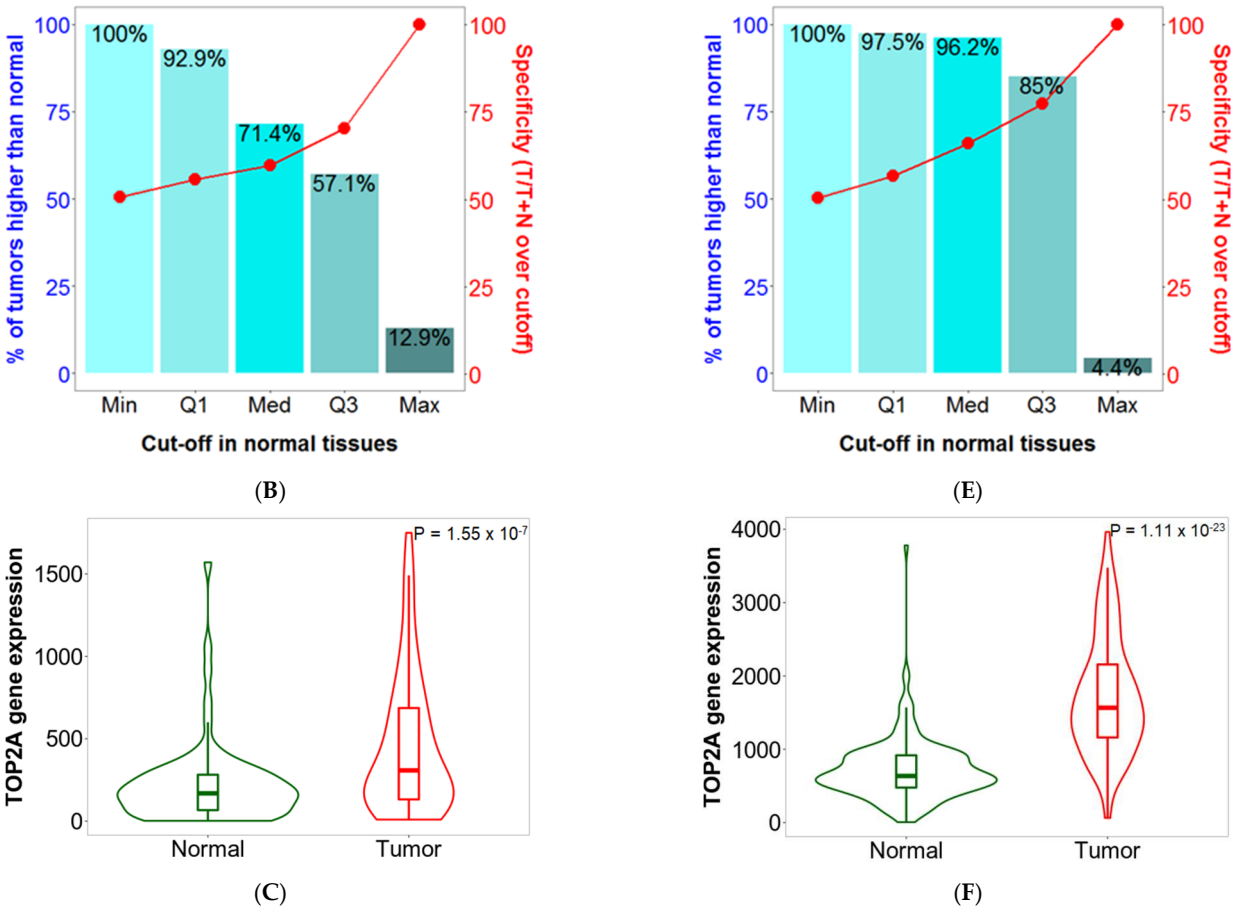

Figure 2. Boxplots (**A**,**D**), bar charts (**B**,**E**) and violin plots (**C**,**F**) of TOP2A gene expression in breast (left) and colon cancer (right) when comparing paired normal and tumor gene array data. The bars represent the proportions of tumor samples that show higher expression of the selected 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*".

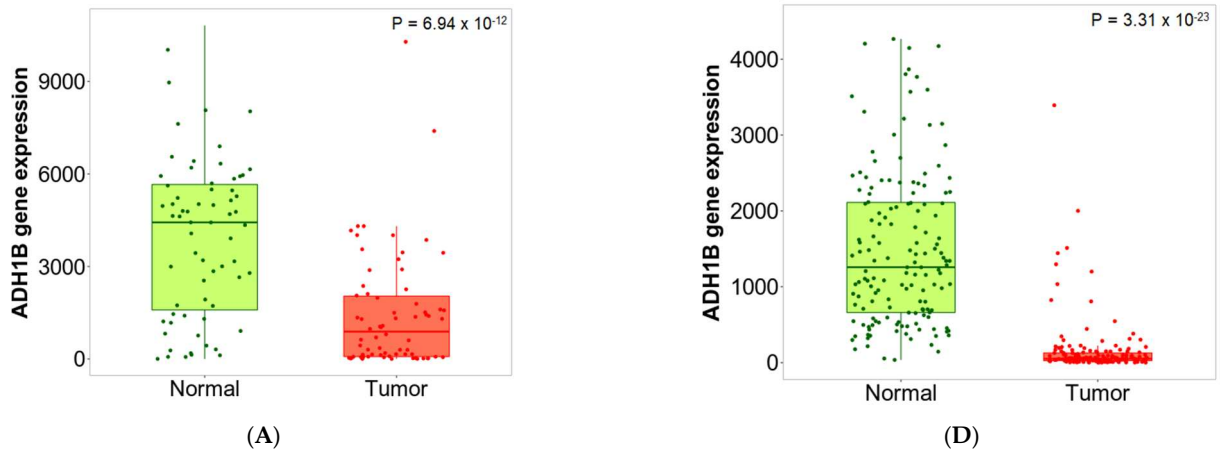

Figure 3. *Cont.*

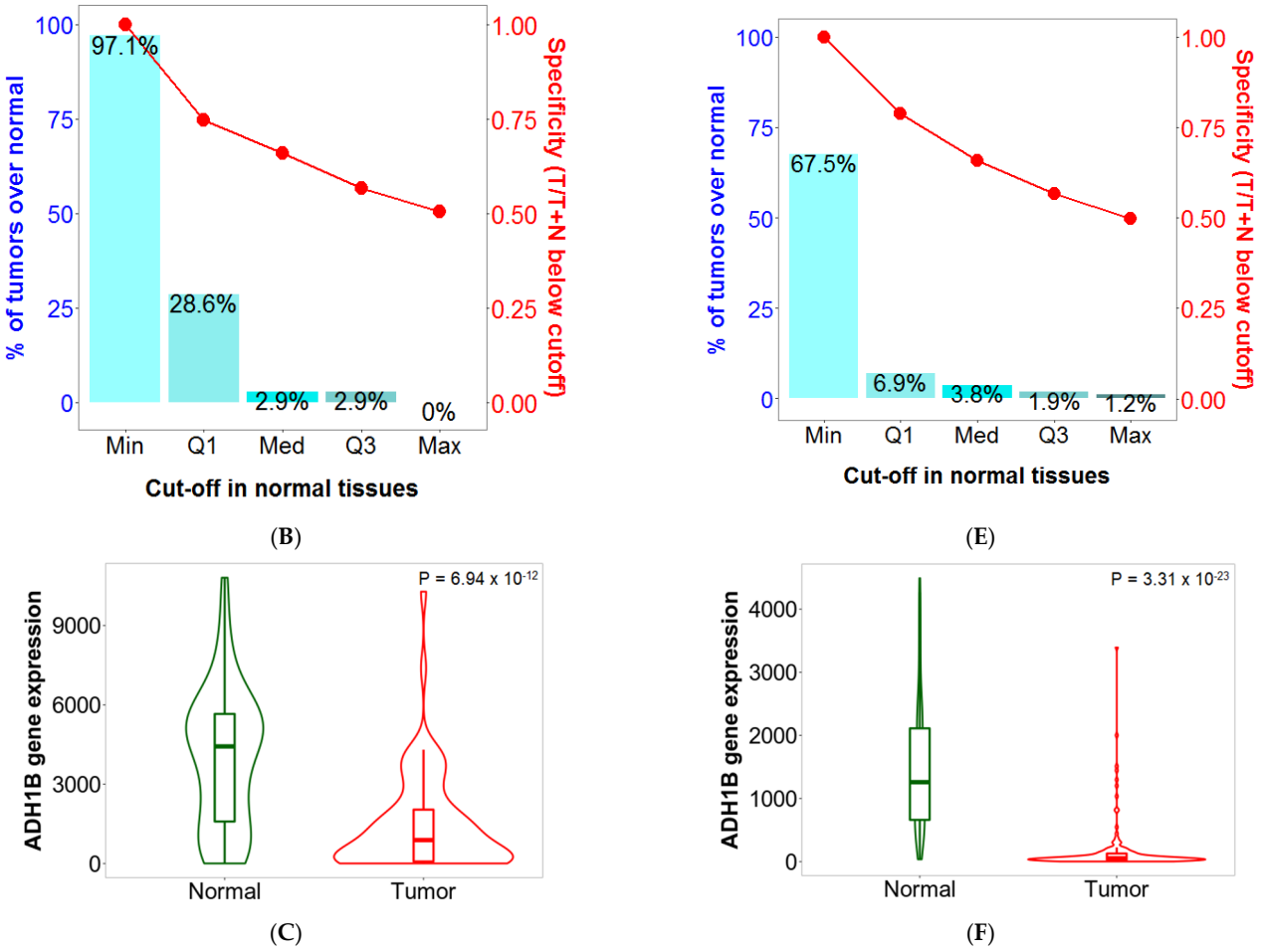

Figure 3. Boxplots (**A**,**D**), bar charts (**B**,**E**) and violin plots (**C**,**F**) of ADH1B gene expression in breast (left) and colon cancer (right) when comparing paired normal and tumor gene array data. The bars represent the proportions of tumor samples that show higher expression of the selected 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*".

2.5. 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).

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).

Gene	Mean Fold Change	Gene	Mean Fold Change
TOP2A	7.80	RUVBL2	1.77
SPP1	7.00	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	1.60
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	1.40
ZWINT	4.15	DIRAS3	0.60
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	2.00	DPT	0.25
MAGOHB	1.90	C7	0.24
RPN2	1.83	AOX1	0.22
SNRPF	1.82	ADH1B	0.15
ENO1	1.79		

2.6. 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, 7223 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, 8259 and 6763 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, 7846 and 8484 overlapping genes reached significance in both examined cohorts in the gene array platform and in RNA-seq, respectively ($p < 0.0001$). As each executed analysis showed a $p < 0.0001$, we concluded that the database could provide highly reproducible results in both platforms. Volcano plots and Venn diagrams depicting the results of the validation are listed in Figure 4.

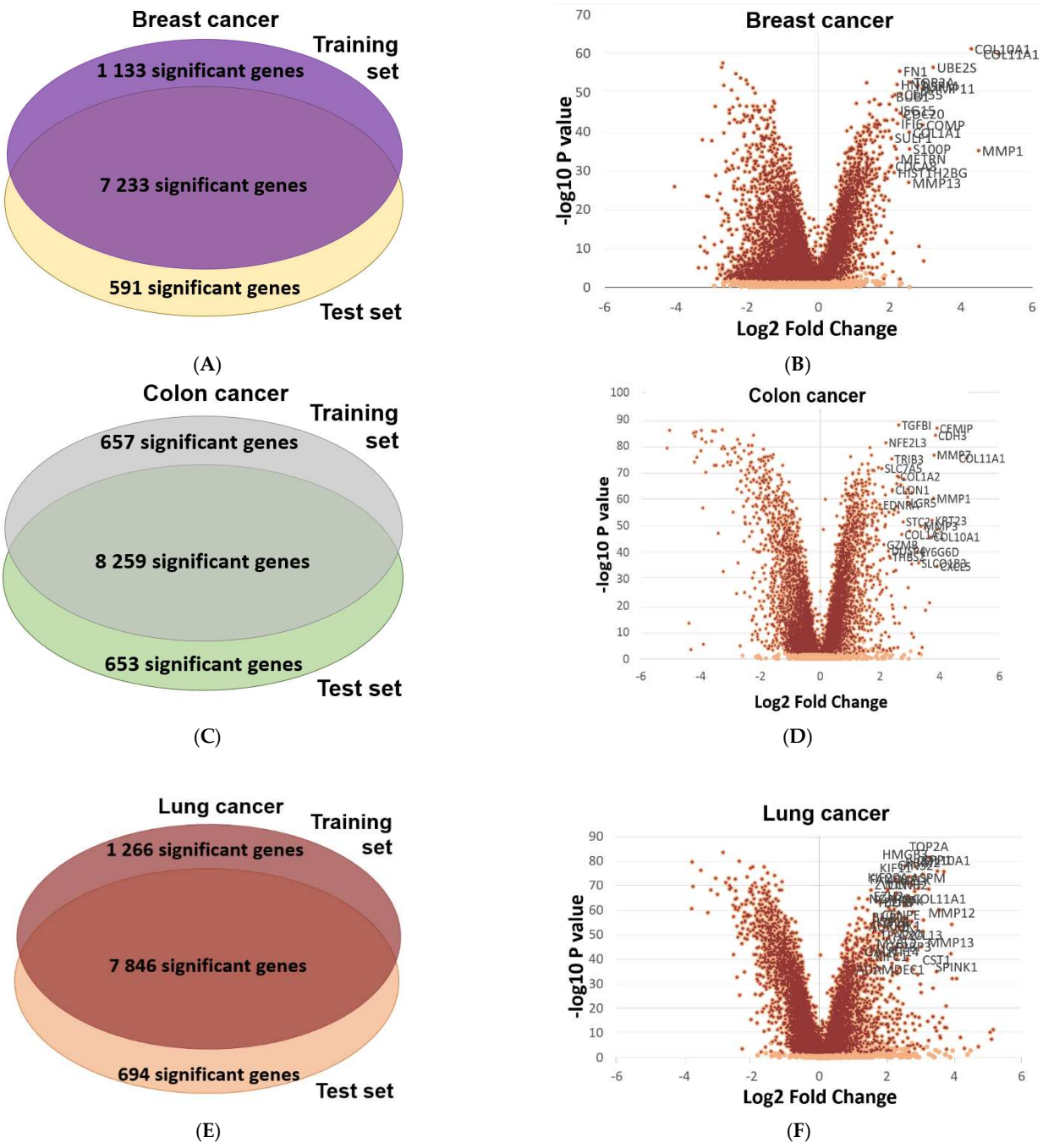

Figure 4. Volcano plots and Venn diagrams of differentially expressed genes in breast, colon and lung cancer using equally sized training–test sets —Venn diagram (**A**) and Volcano plot (**B**) from breast cancer; Venn diagram (**C**) and Volcano plot (**D**) from colon cancer Venn diagram (**E**) and Volcano plot (**F**) from lung cancer.

3. Discussion

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, we have recently established the www.rocplot.org platform, capable of identifying predictive biomarkers in multiple tumor types by employing ROC analysis [15]. However, one has to set a clinically applied cutoff, thus the overall performance of a marker in an ROC analysis is of little clinical value. Another minor drawback of the ROC plot is that the determination of the optimal cutoff value needs additional computations.

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 [16], colon [17], and breast cancer [18]. At present, multiple drugs, including doxorubicin, epirubicin or etoposide, are widely used in clinical practice to target TOP2A or other topoisomerase gene products [19]. These agents are now used in multiple tumor types, including breast cancer [20], leukemias and lymphomas [21,22].

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 [23], lung [24] or head and neck cancer [25].

A notable limitation of our study is the low number of available metastatic tissues. Although the total number (n = 848) seems useful, 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. Another limitation of our database is the lack of data on gene regulation, including alternative splicing. Alternative splicing can result in different proteins with dissimilar functions. A future employment of a multi-omic approach, in conjunction with the utilization of proteomic data, might help to circumvent these issues [26]. With the administration of other robust methods for further complex analysis, such as variable selection methods—for instance, robust loss function methods, LASSO approach or simultaneous multiple quantile regression—one can gain further insights to additional gene and gene set interactions [27].

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 a registration-free online analysis portal which enables mining of the database for any gene to assess expression differences in normal, cancer and metastatic samples.

4. Materials and Methods

4.1. 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, 3180 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. Following this manual selection, the remaining samples were normalized using the MAS5 algorithm by employing the Affy Bioconductor library. Finally, a second scaling normalization was made to set the mean expression on each array to 1000.

4.2. 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 1197 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.

4.3. 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. For the support of future data analysis, we constructed a master gene annotation data table with all the previous gene names and available synonyms for each included gene (Supplemental Table S1).

4.4. 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 test. Normal, tumorous and metastatic tissue gene comparison can be analyzed using Kruskal–Wallis test. The statistical significance cutoff was set at $p < 0.01$.

4.5. 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 [28]. 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 [29].

4.6. 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. Following the calculation of the p values for each gene, a chi-squared test was performed to compare selection overlap between the training set and the test sets and to validate the proportion of differentially expressed genes. Volcano plots comparing $-\log_{10}$ p values and Log_2 fold changes were generated to visualize differential expression.

4.7. 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. Tumor types were selected using the 2019 mortality data from the United States [30]. 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.

Supplementary Materials: The following are available online at https://www.mdpi.com/1422-0067/22/5/2622/s1.

Author Contributions: Conceptualization, B.G. and Á.B.; methodology, Á.B., B.G. software, Á.B.; validation, Á.B. and B.G.; formal analysis, Á.B.; investigation, Á.B.; resources, B.G.; data curation, Á.B.; writing—original draft preparation, Á.B.; writing—review and editing, B.G.; visualization, Á.B.; supervision, B.G.; project administration, Á.B.; funding acquisition, B.G. All authors have read and agreed to the published version of the manuscript.

Funding: The research was financed by the 2018-2.1.17-TET-KR-00001 and 2018-1.3.1-VKE-2018-00032 grants and by the Higher Education Institutional Excellence Programme (2020-4.1.1.-TKP2020) of the Ministry for Innovation and Technology in Hungary, within the framework of the Bionic thematic programme of the Semmelweis University.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data available in a publicly accessible repository that does not issue DOIs Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/4ronB/tnmplot.

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

References

1. Zhang, L.; Zhou, W.; Velculescu, V.E.; Kern, S.E.; Hruban, R.H.; Hamilton, S.R.; Vogelstein, B.; Kinzler, K.W. Gene expression profiles in normal and cancer cells. *Science* **1997**, *276*, 1268–1272. [CrossRef] [PubMed]
2. Druker, B.J.; Tamura, S.; Buchdunger, E.; Ohno, S.; Segal, G.M.; Fanning, S.; Zimmermann, J.; Lydon, N.B. Effects of a selective inhibitor of the Abl tyrosine kinase on the growth of Bcr-Abl positive cells. *Nat. Med.* **1996**, *2*, 561–566. [CrossRef]
3. Hanahan, D.; Weinberg, R.A. Hallmarks of cancer: The next generation. *Cell* **2011**, *144*, 646–674. [CrossRef] [PubMed]
4. Menyhart, O.; Harami-Papp, H.; Sukumar, S.; Schafer, R.; Magnani, L.; de Barrios, O.; Gyorffy, B. Guidelines for the selection of functional assays to evaluate the hallmarks of cancer. *Biochim. Biophys. Acta* **2016**, *1866*, 300–319. [CrossRef]
5. Lowe, R.; Shirley, N.; Bleackley, M.; Dolan, S.; Shafee, T. Transcriptomics technologies. *PLoS Comput. Biol.* **2017**, *13*, e1005457. [CrossRef] [PubMed]

6. Cancer Genome Atlas Research, N.; Weinstein, J.N.; Collisson, E.A.; Mills, G.B.; Shaw, K.R.; Ozenberger, B.A.; Ellrott, K.; Shmulevich, I.; Sander, C.; Stuart, J.M. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **2013**, *45*, 1113–1120. [CrossRef]
7. Grossman, R.L.; Heath, A.P.; Ferretti, V.; Varmus, H.E.; Lowy, D.R.; Kibbe, W.A.; Staudt, L.M. Toward a Shared Vision for Cancer Genomic Data. *N. Engl. J. Med.* **2016**, *375*, 1109–1112. [CrossRef] [PubMed]
8. Consortium, G.T. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **2013**, *45*, 580–585. [CrossRef]
9. Edgar, R.; Domrachev, M.; Lash, A.E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **2002**, *30*, 207–210. [CrossRef] [PubMed]
10. Gautier, L.; Cope, L.; Bolstad, B.M.; Irizarry, R.A. affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **2004**, *20*, 307–315. [CrossRef] [PubMed]
11. Li, Q.; Birkbak, N.J.; Gyorffy, B.; Szallasi, Z.; Eklund, A.C. Jetset: Selecting the optimal microarray probe set to represent a gene. *BMC Bioinform.* **2011**, *12*, 474. [CrossRef]
12. Love, M.I.; Huber, W.; Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **2014**, *15*, 550. [CrossRef] [PubMed]
13. Pagès, H.; Carlson, M.; Falcon, S.; Li, N. AnnotationDbi: Manipulation of SQLite-Based Annotations in Bioconductor. Available online: http://www.bioconductor.org/packages/release/bioc/html/AnnotationDbi.html (accessed on 4 March 2021).
14. Durinck, S.; Spellman, P.T.; Birney, E.; Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **2009**, *4*, 1184–1191. [CrossRef] [PubMed]
15. Fekete, J.T.; 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*, 3140–3151. [CrossRef] [PubMed]
16. Kou, F.; Sun, H.; Wu, L.; Li, B.; Zhang, B.; Wang, X.; Yang, L. TOP2A Promotes Lung Adenocarcinoma Cells' Malignant Progression and Predicts Poor Prognosis in Lung Adenocarcinoma. *J. Cancer* **2020**, *11*, 2496–2508. [CrossRef] [PubMed]
17. Zhang, R.; Xu, J.; Zhao, J.; Bai, J.H. Proliferation and invasion of colon cancer cells are suppressed by knockdown of TOP2A. *J. Cell Biochem.* **2018**, *119*, 7256–7263. [CrossRef] [PubMed]
18. An, X.; Xu, F.; Luo, R.; Zheng, Q.; Lu, J.; Yang, Y.; Qin, T.; Yuan, Z.; Shi, Y.; Jiang, W.; et al. The prognostic significance of topoisomerase II alpha protein in early stage luminal breast cancer. *BMC Cancer* **2018**, *18*, 331. [CrossRef]
19. Delgado, J.L.; Hsieh, C.M.; Chan, N.L.; Hiasa, H. Topoisomerases as anticancer targets. *Biochem. J.* **2018**, *475*, 373–398. [CrossRef] [PubMed]
20. Jasra, S.; Anampa, J. Anthracycline Use for Early Stage Breast Cancer in the Modern Era: A Review. *Curr. Treat. Options Oncol.* **2018**, *19*, 30. [CrossRef]
21. Hallek, M. Chronic lymphocytic leukemia: 2020 update on diagnosis, risk stratification and treatment. *Am. J. Hematol.* **2019**, *94*, 1266–1287. [CrossRef] [PubMed]
22. 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*, 739–746. [CrossRef]
23. Kropotova, E.S.; Zinovieva, O.L.; Zyryanova, A.F.; Dybovaya, V.I.; Prasolov, V.S.; Beresten, S.F.; Oparina, N.Y.; Mashkova, T.D. Altered expression of multiple genes involved in retinoic acid biosynthesis in human colorectal cancer. *Pathol. Oncol. Res.* **2014**, *20*, 707–717. [CrossRef] [PubMed]
24. 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*, 3578–3590. [CrossRef] [PubMed]
25. Lan, J.; Huang, H.Y.; Lee, S.W.; Chen, T.J.; Tai, H.C.; Hsu, H.P.; Chang, K.Y.; Li, C.F. TOP2A overexpression as a poor prognostic factor in patients with nasopharyngeal carcinoma. *Tumour Biol. J. Int. Soc. Oncodev. Biol. Med.* **2014**, *35*, 179–187. [CrossRef] [PubMed]
26. Sulakhe, D.; D'Souza, M.; Wang, S.; Balasubramanian, S.; Athri, P.; Xie, B.; Canzar, S.; Agam, G.; Gilliam, T.C.; Maltsev, N. Exploring the functional impact of alternative splicing on human protein isoforms using available annotation sources. *Brief. Bioinform.* **2019**, *20*, 1754–1768. [CrossRef]
27. Wu, C.; Ma, S. A selective review of robust variable selection with applications in bioinformatics. *Brief. Bioinform.* **2015**, *16*, 873–883. [CrossRef] [PubMed]
28. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*; Springer: New York, NY, USA, 2016.
29. Chang, W.; Cheng, J.; Allaire, J.J.; Xie, Y.; McPherson, J. Shiny: Web Application Framework for R. Available online: https://CRAN.R-project.org/package=shiny (accessed on 4 March 2021).
30. Siegel, R.L.; Miller, K.D.; Jemal, A. Cancer statistics, 2020. *CA Cancer J. Clin.* **2020**, *70*, 7–30. [CrossRef]