cancers

MDPI

Article

TCGAnalyzeR: An Online Pan-Cancer Tool for Integrative Visualization of Molecular and Clinical Data of Cancer Patients for Cohort and Associated Gene Discovery

Talip Zengin [1,†], Başak Abak Masud [2,†] and Tuğba Önal-Süzek [2,*]

[1] Department of Molecular Biology and Genetics, Mugla Sitki Kocman University, Mugla 48000, Türkiye; talipzengin@mu.edu.tr
[2] Department of Bioinformatics, Mugla Sitki Kocman University, Mugla 48000, Türkiye; basakkabakk@gmail.com
* Correspondence: tugbasuzek@mu.edu.tr
† These authors contributed equally to this work.

Simple Summary: TCGAnalyzeR provides a novel web site integrating 123 pre-computed pan-cancer cohorts (i.e., microsatellite instability, immune, metastasis, PAM50, Triple Negative breast cancer, idh1-mutated glioblastoma, etc.), along with our own iCluster+ subcohorts, computed based on pre-processed single-nucleotide variations, copy number variations, differential expression, miRNA, methylation, and clinical data. TCGAnalyzeR interface provides an optimized and fully customizable experience to each user, enabling the selection of their own "My patients" or "My genes" to a clipboard. Several use cases of the web site are presented as Help documents.

Abstract: For humans, the parallel processing capability of visual recognition allows for faster comprehension of complex scenes and patterns. This is essential, especially for clinicians interpreting big data for whom the visualization tools play an even more vital role in transforming raw big data into clinical decision making by managing the inherent complexity and monitoring patterns interactively in real time. The Cancer Genome Atlas (TCGA) database's size and data variety challenge the effective utilization of this valuable resource by clinicians and biologists. We re-analyzed the five molecular data types, i.e., mutation, transcriptome profile, copy number variation, miRNA, and methylation data, of ~11,000 cancer patients with all 33 cancer types and integrated the existing TCGA patient cohorts from the literature into a free and efficient web application: TCGAnalyzeR. TCGAnalyzeR provides an integrative visualization of pre-analyzed TCGA data with several novel modules: (i) simple nucleotide variations with driver prediction; (ii) recurrent copy number alterations; (iii) differential expression in tumor versus normal, with pathway and the survival analysis; (iv) TCGA clinical data including metastasis and survival analysis; (v) external subcohorts from the literature, curatedTCGAData, and BiocOncoTK R packages; (vi) internal patient clusters determined using an iClusterPlus R package or signature-based expression analysis of five molecular data types. TCGAnalyzeR integrated the multi-omics, pan-cancer TCGA with ~120 subcohorts from the literature along with clipboard panels, thus allowing users to create their own subcohorts, compare against existing external subcohorts (MSI, Immune, PAM50, Triple Negative, *IDH1*, miRNA, metastasis, etc.) along with our internal patient clusters, and visualize cohort-centric or gene-centric results interactively using TCGAnalyzeR.

Keywords: clinical data integration; cancer subcohort analysis; TCGA data visualization; driver mutations prediction; copy number variations in cancer; transcriptome analysis; oncology research platforms

Citation: Zengin, T.; Masud, B.A.; Önal-Süzek, T. TCGAnalyzeR: An Online Pan-Cancer Tool for Integrative Visualization of Molecular and Clinical Data of Cancer Patients for Cohort and Associated Gene Discovery. *Cancers* **2024**, *16*, 345. https://doi.org/10.3390/cancers16020345

Academic Editors: Katrien Remaut and Wim P. Ceelen

Received: 6 November 2023
Revised: 8 January 2024
Accepted: 11 January 2024
Published: 13 January 2024

1. Introduction

With the decreasing cost and increased availability of the new generation of sequencing techniques and their power to simultaneously detect more than one gene variant in the

clinic, many genetic tests have been released to the market and approved for clinical diagnosis by the FDA. Examples of tests in clinical use are the Oncomine Dx Target Test [1], which tests the sequence variations of 46 genes on DNA and RNA for lung cancer, the Oncotype DX test [2], which tests the 21 mutations for breast cancer, and the PAM50 test [3], which tests the expression data of 50 genes in breast cancer. Yet, due to several recent clinical studies showing that the powers of these tests are not comparable to conventional single gene tests, there is an open question on how to analytically compare the clinical performance of personalized oncology tests.

The sheer scale and complexity of The Cancer Genome Atlas (TCGA) data [4] offers great potential for scientific discovery, but the challenges to effective use of this valuable resource by biologists and clinicians have led to the development of several visualization tools such as cBioPortal [5,6], Firebrowse [7], and University of California, Santa Cruz (UCSC) Xena [8]. These visualization tools do not serve as data repositories but rather aim to create an integrated visualization of TCGA. Among these tools, cBioPortal is the most preferred due to its interactive exploration of larger and up-to-date cancer datasets. OncoKB [9] is another precision oncology knowledge base that allows searching and comparing of drug response data from different TCGA cohorts, yet the visualizations are limited by variant effects. Although the ICGC web portal [10] allows patient/gene subsetting of TCGA cohorts and provides survival and set operation visualization of cohorts, it does not allow comparison of cohorts compiled from the literature generated by other research groups. Only the Coral web application [11], similar to TCGAnalyzeR, incorporates a few (MSI, tumor-purity and immune) subcohorts from the literature, yet it does not allow for comparison against each other and it does not allow for their projection onto Oncoplots. Most of these comparable tools provide access to raw data only, with limited additional pre-processing. TCGAnalyzeR, in contrast, enables users to project any cohort out of an extensive set of 123 pre-loaded patient subcohorts onto Oncoplots/Oncogrids of patient mutations, in addition to survival and subsetting options. Oncoplot/Oncogrid visualizations are especially critical for oncologists who use FDA-approved diagnostic gene panels, such as the Oncotype DX test or PAM50, and want to validate these tests' predictive power against a wide spectrum of existing TCGA cohorts. The comparison of the existing TCGA visualization web tools against TCGAnalyzeR is summarized in Table 1.

Table 1. Comparison of TCGA visualization tools.

TCGA Visualization Tool	Login or Subscription Required for Advanced Features	Survival and Set Operation Diagrams of Selected Patient Subcohorts	Caching of Both User's Gene and Patient Selection	Number of External TCGA Cohorts from Literature
TCGAnalyzeR http://tcganalyzer.mu.edu.tr (accessed on 28 December 2023)	No	Yes	Yes	123
cBioPortal [5,6] https://www.cbioportal.org/ (accessed on 28 December 2023)	Yes	No	Yes	0
Xena [8] https://xenabrowser.net/ (accessed on 28 December 2023)	Yes	Yes	No	0
OncoKB [9] https://www.oncokb.org/ (accessed on 28 December 2023)	Yes	No	No-only genes	0
ICGC cohort analysis [10] https://dcc.icgc.org/analysis (accessed on 28 December 2023)	No	Yes	Yes	0
Coral web application [11] https://coral.caleydoapp.org (accessed on 28 December 2023)	Yes	No	No	3

To address this limitation, we developed an interactive R Shiny web application for the analysis and visualization of four data categories across 33 cancer types. Users can visualize the results of preprocessed analysis of Simple Nucleotide Variations (SNVs), Copy Number Variations (CNVs), differential gene expression in tumor versus normal samples, and clinical data of TCGA projects from the National Cancer Institute's (NCI) Genomic Data Commons (GDC) [12]. Moreover, users can compare patient clusters determined using an iClusterPlus R package [13] with expression-based survival risk groups [14,15] and curated subtypes, such as immune subtypes [16], Triple Negative Breast Cancer (TNBC) subtypes [17], PAM50 subtypes [18], Microsatellite Instability (MSI)-related subgroups and several data type clusters from BiocOncoTK [19,20], and curatedTCGAData (version 1.20.1) R packages [21]. While gathering many of these subcohorts in a clinical setting can present challenges, such as metastasis patients with primary tumor data or MSI-H in non-endometrial cancers like BRCA, the vast size of TCGA enables the analysis of these rare subcohorts. Furthermore, users can create custom subcohorts based on genomic analyses and/or clinical data, including metastasis organs/tissues to subset data visualization. Users can also create gene sets for data type and/or pan-cancer comparisons. For each cancer, whenever available, sample types, survival risk groups (Low-risk/High-risk), and pre-computed or curated patient clusters can be used for filtering patients. The main novelty of our tool is its ability to integrate many published subcohorts at a single pan-cancer interface, allowing the users to generate their own custom patient sub-cohorts and/or gene sets using interactive graphical representations via clipboard functionality.

2. Materials and Methods

2.1. TCGA Data

Publicly available hg38 data, including SNV, CNV, Transcriptome Profiling, microRNA, Methylation, and clinical data of 33 cancer types from The Cancer Genome Atlas (TCGA) projects, were downloaded on 6 March 2022 from NCI GDC [12] using TCGAbiolinks R package [22].

2.2. Pre-Computed Molecular Data Analysis

2.2.1. SNV Analysis

Potential driver mutated genes, with their roles as a tumor suppressor or oncogene, were determined by SomInaClust R package [23] using a mutation annotation format (maf) file generated by mutect2 pipeline. With the "Somatic Driver Mutations" option, the user can see the significant mutated genes ranked by their q-value. This option is only available for the "SNV Analysis" category. Statistical methods implemented for SNV analysis are described in more detail in our previous publications [14,15] and the R codes are provided in the GitHub repository.

2.2.2. CNV Analysis

Significant recurrent copy number variations were identified by GAIA R package [24]. NCBI IDs and Hugo Symbols of the genes on chromosomal regions with altered copy numbers were determined using GenomicRanges [25] and biomaRt [26] R packages. Statistical methods implemented for CNV analysis are described in more detail in our previous publications [14,15], and the R codes are provided in the GitHub repository.

2.2.3. Gene Expression Analysis

Two different analyses were performed using paired tumor samples against tumor-adjacent normal samples of patients with both sample types (Paired), or tumor samples of all patients against normal samples of patients who have both sample types (All), if it was available for a particular cancer. For cancers with paired tumor samples, differentially expressed genes were determined using normalized HTseq counts, by limma-voom method with a 'duplicate correlation' function from edgeR [27] and limma [28] R packages. Ensemble IDs were converted to NCBI IDs and Hugo Symbols using the biomaRt package [26].

For 11 cancers out of 33 cancer types, there is no patient sample with tumor adjacent normal tissue that exists, therefore we created the (All) option for all cancers using the normalized HTseq counts by the TMM method, followed by a Log2 transformation. Genes with consistently zero or low counts were filtered out. Statistical methods implemented for paired differential expression analysis are described in more detail in our previous publication [14,15] and the R codes are provided in the GitHub repository.

2.2.4. Pathway Enrichment

Pathway enrichment and visualization was performed for each analysis using a clusterProfiler R package [29]. Statistical methods implemented for CNV analysis are described in more detail in our previous publication [14,15], and the R codes are provided in the GitHub repository.

2.2.5. Pre-Computed Patient Clusters and Curated Subcohorts from the Literature

Although many pan-cancer subcohorts based on TCGA data, such as Microsatellite Instability (MSI) clusters [19,20] and immune clusters [16], have been published in high-impact journals, only few tools (i.e., Coral) have integrated them into their visualizations. The Coral web application [11] integrated only a few of the literature-curated cohorts, such as MSI, tumor-purity and immune subcohorts, yet it does not allow comparison of these subcohorts against each other and it does not allow projection of these subcohorts onto Oncoplots/Oncogrid visualizations. As of January 2024, there was no web tool in the literature that allowed for visual comparison of a large number of patient subcohorts with each other. To address this need, TCGAnalyzerR integrated 123 external patient cohorts from the literature into the web interface, enabling efficient filtering and facilitating cross-comparative analysis of multiple subcohorts in parallel. TCGAnalyzeR provides an interactive visual analysis of several patient cohorts. (i) Survival Risk Groups: we provide low-risk or high-risk patient groups determined by expression-based gene signature analysis for Lung Adenocarcinoma (LUAD), Lung Squamous Cell Carcinoma (LUSC), and Colon Adenocarcinoma (COAD) [14,15]. (ii) iClusters: we clustered patients using their raw SNV, CNV, gene expression, miRNA expression and methylation data of tumor samples which have all 5 types of data using the iClusterBayes method [30]. (iii) Curated subcohorts: Immune, TNBC, MSI, PAM50 subtypes are downloaded from original publications [16–18] and for fifteen cancer types, previously published TCGA cohorts of the individual tumor types were retrieved from the curatedTCGAData R package (version 1.20.1) [21]. Patient clusters based on Microsatellite Instability (MSI) were compiled using BiocOncoTK [19,20], and Immune clusters [16] were compiled from its original publication for all 33 cancers. Metastasis site information is brought together in-house from BCR Biotab clinical patient data, BCR Biotab clinical new tumor event data, and BCR Biotab clinical new tumor event follow-up data using the TCGAbiolinks R package.

2.2.6. Survival Analysis

Real-time Kaplan–Meier (KM) survival analysis is conducted using the survival R package [31] and is based on overall survival data for patients of interest with selected clinical features. Data input and tabular reading are facilitated by the readr [32] R package.

2.2.7. Visualization

The TCGAnalyzeR front-end was implemented using javascript-based R packages with an interactive dashboard enabling users to select cancer types, data types, risk groups, and patient cohorts using heatmaply, g3viz, and highcharter R packages [33–35]. All visualizations are interactive and customizable by the user through the filtration options with “My genes” and/or “My patients” panels, enabling them to copy genes and/or patients of interest to the clipboard. For BRCA, OncotypeDX gene identifiers are provided to users as an example use case of the “My genes” clipboard. TCGAnalyzeR currently supports the tab separated values (TSV) file type for downloading tables and a high-

resolution PNG format for downloading figures. The help page provides links to the original publications of the external subcohorts. An animated image of usage examples has been placed in the help section.

2.2.8. Performance Optimization

For the web performance profiling of all the tools, the profvis package is utilized to inspect the call stack, identify, and optimize the most memory- and (computationally) time-consuming parts of each module. Since its launch on 1 January 2022, although not published, TCGAnalyzeR has been accessed by an average of 79 unique user IPs per day.

3. Results

The TCGAnalyzeR web application offers simple nucleotide (SNV) analysis as its first step. We present two data sets for SNV analysis: "Somatic Driver Mutations" predicted by the SomInaClust R package and "All" mutations from the original maf file without any analysis. The Oncoplot in Figure 1 shows candidate driver genes with their percentages in tumor samples of Breast Invasive Carcinoma (BRCA) with annotations regarding patient iClusters, PAM50, TNBC, immune subtypes, and metastasis organs/tissues. iCluster 1 is highly correlated with the Basal and TNBC subtype. Wound-healing and IFNɤ-dominant immune subtypes gather around iCluster 1. iCluster 2 is mostly correlated with the Luminal A subtype and Inflammatory immune subtype. iCluster 3 seems to be a mixture of estrogen receptor positive Luminal A and Luminal B subtypes and heterogenous immune subtypes. Moreover, both iCluster 2 and iCluster 3 are not TNBC subtypes. On the other hand, iCluster 1 shows a highly different mutation pattern than other clusters. iCluster 1, together with basal and triple-negative subtypes, has a higher prevalence of TP53 mutations with very few mutations of *PIK3CA*, *CDH1*, *GATA3*, *KMT2C*, or *MAP3K1* genes. In addition, mutations of *TP53*, *CDH1*, and *GATA3* genes are mutually exclusive. Furthermore, the presence of metastases is heterogeneous across iClusters, PAM50, TNBC subtypes, and immune subtypes (Figure 1), while it is highly correlated with higher tumor stages.

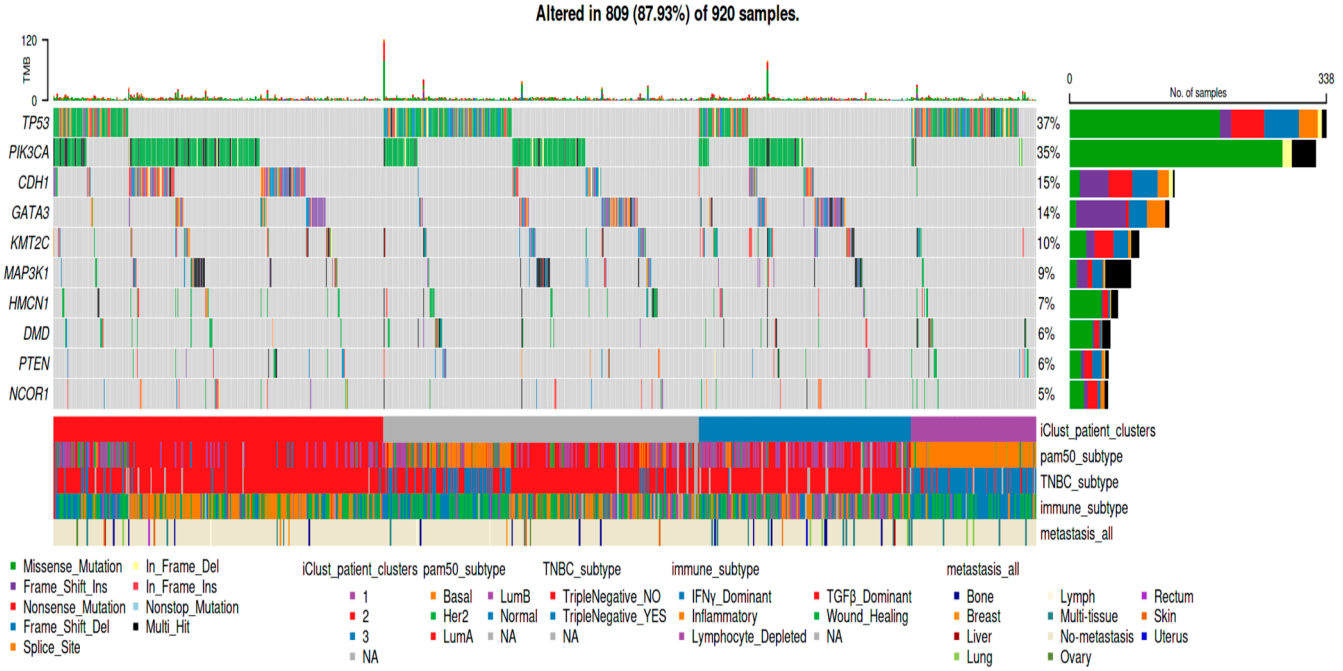
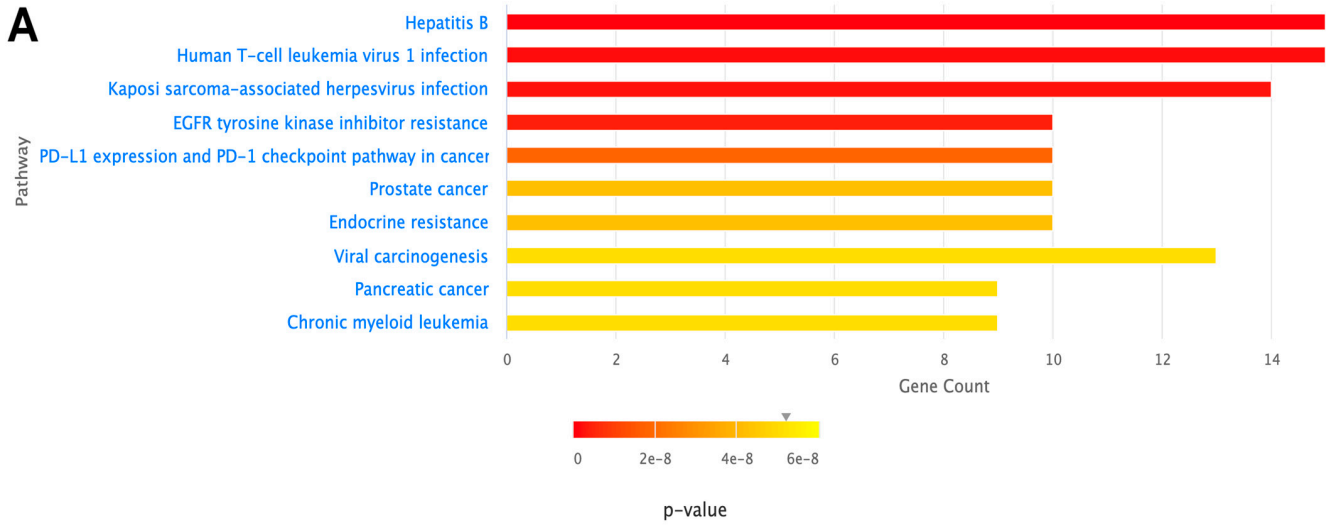


Figure 1. Oncoplot of candidate driver genes with our own pre-computed patient clusters and curated subcohorts from the literature. Top 10 significant candidate driver genes with mutations determined by SomInaClust R package. Bottom annotations show the iClusters, metastasis organs, and subcohorts curated from the literature.

Pathway enrichment of candidate driver mutated genes is shown as a bar graph in Figure 2A. Significant pathways of driver genes are highly cancer-related pathways, such as EGFR tyrosine kinase inhibitor resistance, PD-L1, and PD-1 pathways in cancer, prostate cancer, pancreatic cancer and chronic myeloid leukemia pathways. Pathway enrichment analysis also supplies a table showing KEGG IDs, with related genes and p/q-values (Figure 2B).



B

KEGG ID	Pathway Name	Gene Ratio	p.adjust	q value	Number of Genes
hsa05161	Hepatitis B	15/66	0.000000000332825901032664	0.000000000116781017906198	15
hsa05166	Human T-cell leukemia virus 1 infection	15/66	0.00000001574749393474 02	0.00000000552543646832989	15
hsa05167	Kaposi sarcoma-associated herpesvirus infection	14/66	0.000000019451391809 6028	0.00000000682504975775536	14
hsa01521	EGFR tyrosine kinase inhibitor resistance	10/66	0.000000026068192204 4995	0.00000000914673410684192	10
hsa05235	PD-L1 expression and PD-1 checkpoint pathway in cancer	10/66	0.000000069094441116 5716	0.00000002424366354 96743	10
hsa05215	Prostate cancer	10/66	0.000000128403619028508	0.000000045053901413 5117	10
hsa01522	Endocrine resistance	10/66	0.000000128403619028508	0.000000045053901413 5117	10
hsa05203	Viral carcinogenesis	13/66	0.000000148057644482855	0.000000051950050695 7387	13
hsa05212	Pancreatic cancer	9/66	0.000000148057644482855	0.000000051950050695 7387	9
hsa05220	Chronic myeloid leukemia	9/66	0.000000148057644482855	0.000000051950050695 7387	9

Showing 1 to 10 of 108 entries Previous 1 2 3 4 5 … 11 Next

Figure 2. A sample pathway enrichment plot of candidate BRCA driver genes as downloaded from the TCGAnalyzeR website. (**A**) Bar plot showing top 10 significant pathways of BRCA candidate driver genes determined by SomInaClust R package. (**B**) Pathway enrichment table presenting KEGG ID, genes in significant pathways with adjusted p-value and q-value.

The SomInaClust R package determines candidate driver mutated genes with their potential roles as tumor suppressors (TSG) or oncogenes (OG) with predicted scores [23]. The pyramid plot in Figure 3A summarizes the TSG score and OG score of candidate driver genes ranked by their analysis q-values. Some genes may have both an OG score and a TSG score over the threshold score of 40, in that case, SomInaClust considers the COSMIC cancer gene census (CGC) information (Figure 3B).

CGC	Gene	# Mutations	qDG	OG Score	TSG Score	Classification
Dom	PIK3CA	357	0	89.5	0.3	OG
Rec	TP53	344	2e-198	88.7	39.1	TSG
Rec	GATA3	136	3.26e-143	33.3	90.2	TSG
Rec	CDH1	140	1.82e-135	36.4	85.4	TSG
Rec	MAP3K1	117	5.87e-87	0	75	TSG
Rec	KMT2C	104	3.91e-63	9.1	69.1	TSG
Dom	PTEN	55	7.52e-37	47.1	65.5	TSG
Rec	ARID1A	43	3.79e-27	0	69	TSG
Dom	AKT1	29	1.24e-26	94.1	0	OG
Rec	MAP2K4	37	2.83e-23	55.6	63.9	TSG

Showing 1 to 10 of 104 entries Previous 1 2 3 4 5 … 11 Next

Figure 3. Sample visualization plots of SomInaClust predictions of candidate driver genes as downloaded from the TCGAnalyzeR website. (**A**) Sample web site output showing the pyramid plot of the Oncotype DX genes which were predicted as candidate driver genes with calculated oncogene (OG) and tumor-suppressor (TSG) scores for BRCA by SomInaClust R package. (**B**) Sample web site output showing the SomInaClust analysis results for BRCA with number of mutations, OG score, TSG score, red dashed line representing the SominaClust score threshold of 40 and *q*-value (qDG). CGC: COSMIC cancer gene census, Rec: Recessive (TSG), Dom: Dominant (OG).

The "My genes" clipboard panel of TCGAnalyzeR allows for the modification of plots in order to show genes of interest. For example, Figure 4 shows the mutation pattern of the Oncotype DX gene set together with clinical annotations. iCluster 2, Luminal A subtype, and Her2 subtypes are highly related with *ERBB2* (*HER2*) mutations. Additionally, iCluster 1 has fewer mutations than the other two iClusters. Moreover, mutations of Oncotype DX genes are mostly mutually exclusive (Figure 4).

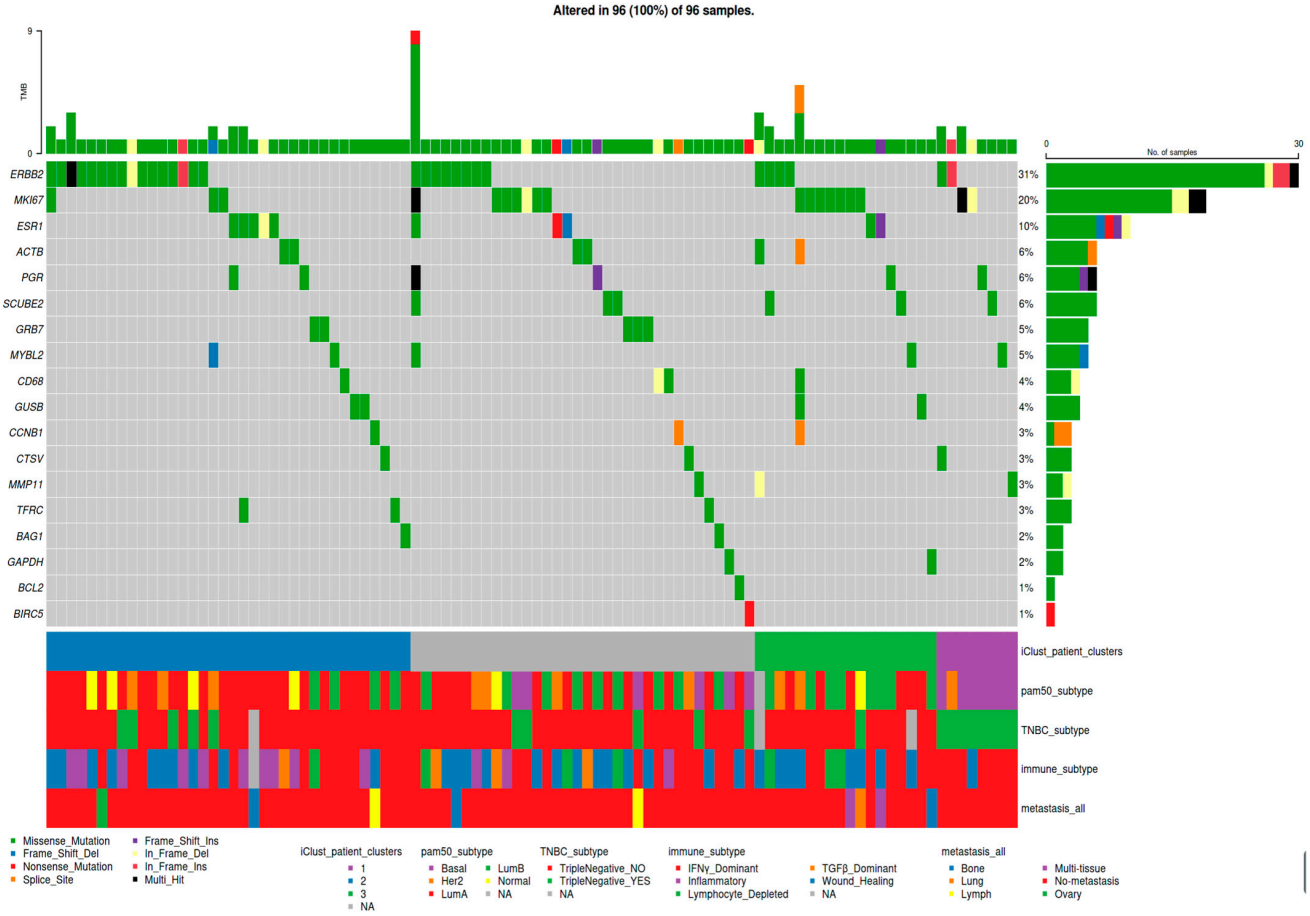


Figure 4. A sample Oncoplot of BRCA Oncotype DX genes with patient clusters and pre-computed subcohorts as downloaded from the TCGAnalyzeR website. Mutations of Oncotype DX genes with annotations showing the patient iClusters and pre-computed BRCA patient subcohorts curated from the literature.

Transcriptome analysis module provides differential expression analysis (DEA) of RNAseq data by comparing the expression of genes in primary tumor samples against adjacent normal samples. We present two result options for this analysis: "Paired" as a comparison of tumor samples against their own paired normal or "All" as a comparison of tumor samples against a normal sample subset of patients, if such is available for the particular cancer. The volcano plot in Figure 5A summarizes the differential expression analysis of paired BRCA samples and Oncotype DX genes that are highlighted through the "My Genes" panel. Figure 5B presents the table showing the details of DEA with gene symbols, fold change (logFC), and *p* values of significantly differentially expressed genes ranked by *p*-value. Pathway enrichment of differentially expressed genes showed that these genes play a role in focal adhesion and ECM-receptor interaction, which can be related with metastasis, Ras signaling, PI3K-Akt signaling, cAMP signaling, and Phenylalanine metabolism pathways, which are related with cell growth (Figure 5C).

The metastasis-related gene *MMP11* and proliferation-related genes *BIRC5*, *MYBL2*, *MKI67* (Ki67), *AURKA* (*STK15*), *CCNB1*, and *ERBB2* (*HER2*) from the Oncotype DX gene set exhibit significant up-regulation in tumor samples of breast cancer (BRCA). However, hormone-related genes (*BAG1*, *BCL2*, *CD68*, *ESR1* (*ER*), *GSTM1*, *PGR*, *SCUBE2*) do not show significant differential expression among all tumor samples (see Figure 5A).

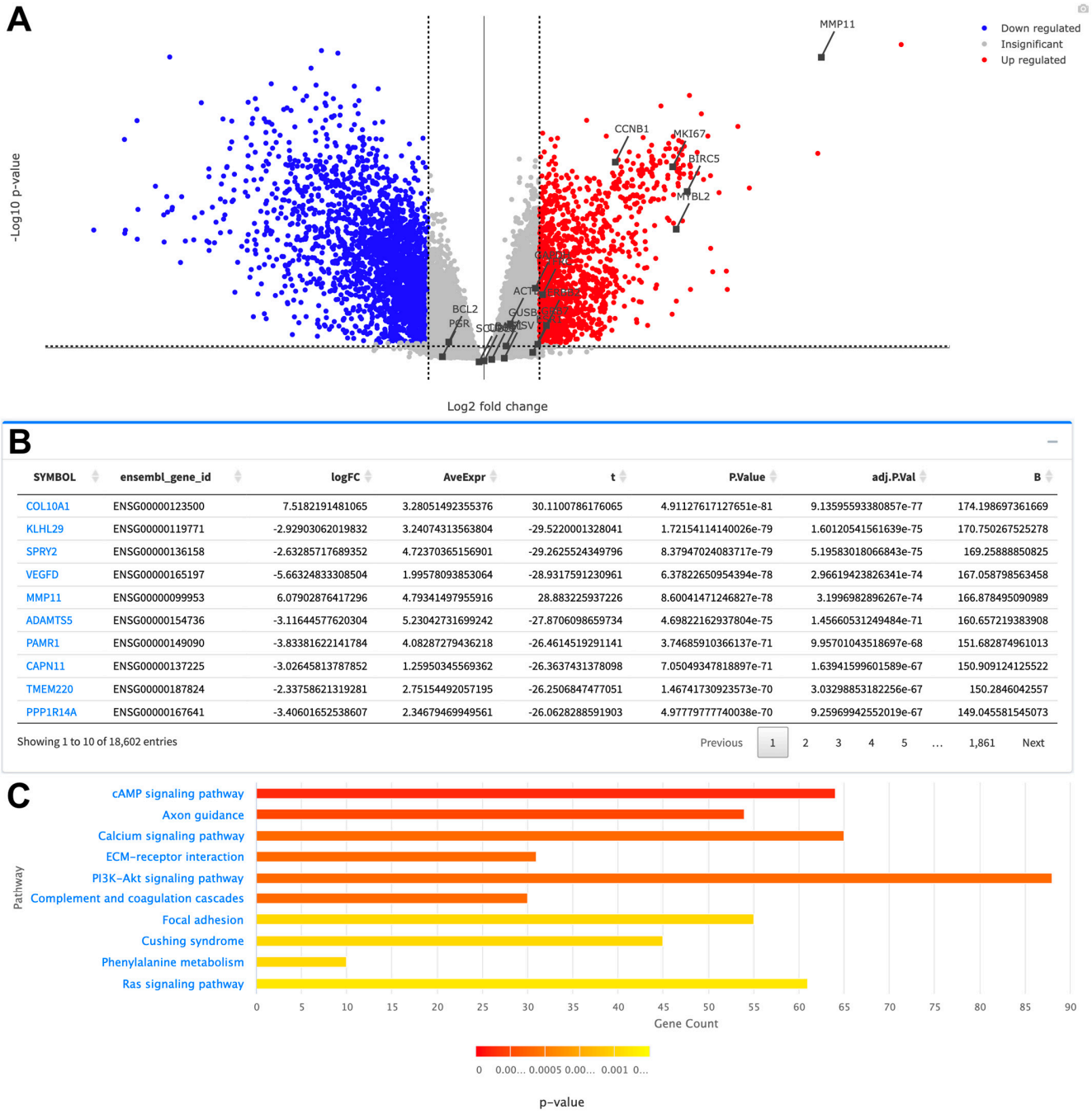


Figure 5. Sample visualizations of differential expression of genes in BRCA tumor samples versus normal samples as downloaded from the TCGAnalyzeR website. (**A**) Volcano plot showing up-regulated and down-regulated genes with -log10 conversion of *p*-values. Oncotype DX genes are highlighted on the graph with gene identifiers connected to the black dots. Black dashed line designates the logFC threshold of 1. Camera image on the upper right corner lets the users download the figure in png format. (**B**) Differential expression results table presenting gene symbols, fold changes (logFC) and adjusted-*p*-values. (**C**) Bar plot showing pathway enrichment of differentially expressed genes.

Focusing on the *ERBB2* gene, which was predicted as a driver oncogene, the positions of mutations can be visualized using the lollipop plot in Figure 6A. Most of the mutations in the *ERBB2* gene are in the kinase domain (see Figure 6A). These mutations are mostly missense on protein positions 755 (n = 7), 767 (n = 2), 769 (n = 3), 777 (n = 4), 797 (n = 1), 842

(n = 1), and 939 (n = 1) and in frame insertion on protein position 885 (n = 1). From these mutations, D769H (n = 1), D769Y (n = 2), V777L (n = 4), and V842I (n = 1) mutations are activating mutations and L755S (n = 5) causes lapatinib resistance [36]. Mutations on the *ERBB2* gene in tumor samples cause lower survival probability with a 1.43 hazard ratio (p = 0.08) (Figure 6B). This is related to the finding that the existence of a mutation in the *ERBB2* gene is one of the prognostic indicators of survival for patients with a primary invasive lobular carcinoma subtype of breast cancer [37].

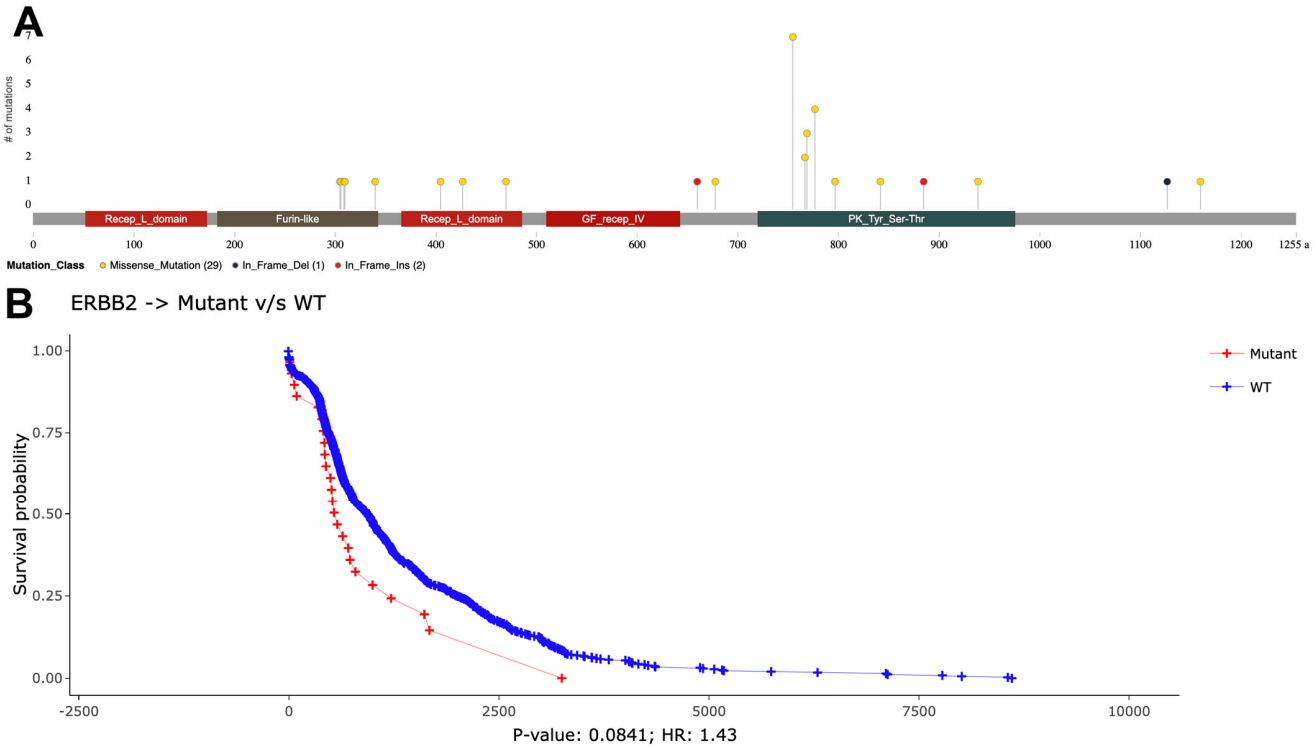


Figure 6. A sample analysis of *ERBB2* (*HER2*) mutation visualizations as downloaded from the TCGAnalyzeR website. (**A**) Lollipop plot showing mutations of *ERBB2* gene among BRCA tumor samples. (**B**) Overall survival analysis of BRCA wild-type versus mutated *ERBB2* in BRCA tumor samples.

We checked the expression levels of *CCNB1*, which is one of the upregulated OncotypeDX genes in tumor samples, versus adjacent normal samples (from paired DEA) (Figure 5A). *CCNB1* is expressed in tumor samples at a significantly higher rate than in their adjacent normal samples ($p = 1.565 \times 10^{-49}$) (Figure 7A). Moreover, patients with higher expression of *CCNB1* have significantly higher survival probability (p = 0.011) (Figure 7B), which is correlated with the finding that high CCNB1 protein expression was associated with poor clinical outcomes [38].

Clinical data analysis comprises pie chart visualization and survival analysis of clinical features using our patient clusters and pre-computed patient subcohorts gathered from the literature. Figures 8 and 9 depict the visualization of proportions and the survival status of PAM50, TNBC, iClusters, and immune subtypes. iClusters exhibited a differential survival probability close to the significance level (p = 0.057); however, PAM50, TNBC, and immune subtypes did not show differential survival probabilities (p = 0.68) (see Figures 8 and 9).

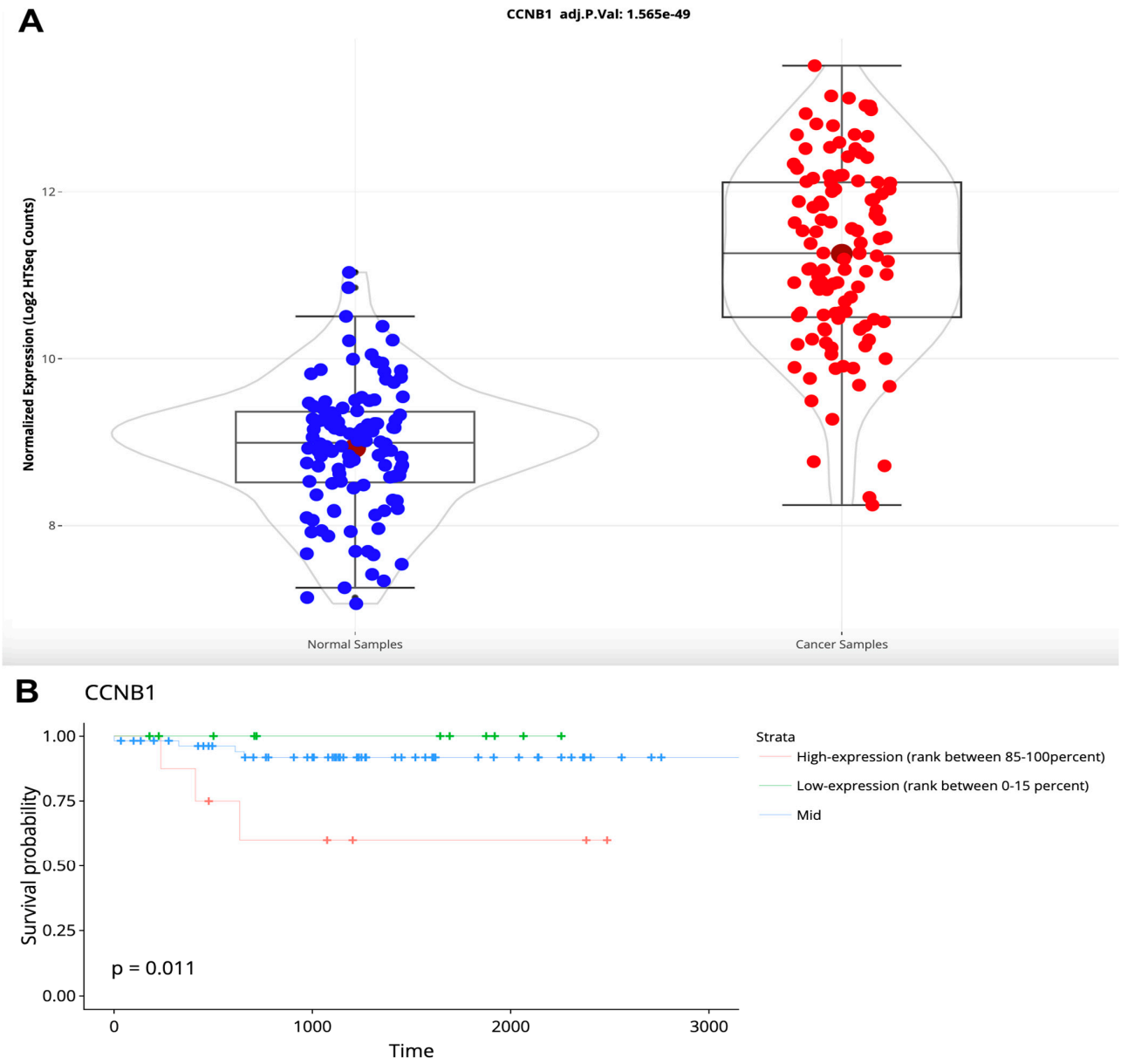


Figure 7. A sample expression analysis of *CCNB1* expression visualizations as downloaded from the TCGAnalyzeR website. (**A**) Violin plot presenting log2 transformed normalized mRNA expression of *CCNB1* in adjacent normal and BRCA tumor samples with adjusted *p*-value. Blue and red circles represent the normalized expression level of each patient's tumor adjacent normal tissue and tumor tissue respectively (**B**) Overall survival analysis of expression levels of *CCNB1* in BRCA tumor samples. + sign indicates a censored patient. Green, blue and red colors represent the survival day of patients with CCNB1 expression between 0–15%, 15–85% and 85–100% of all sorted normalized CCNB1 expression values respectively.

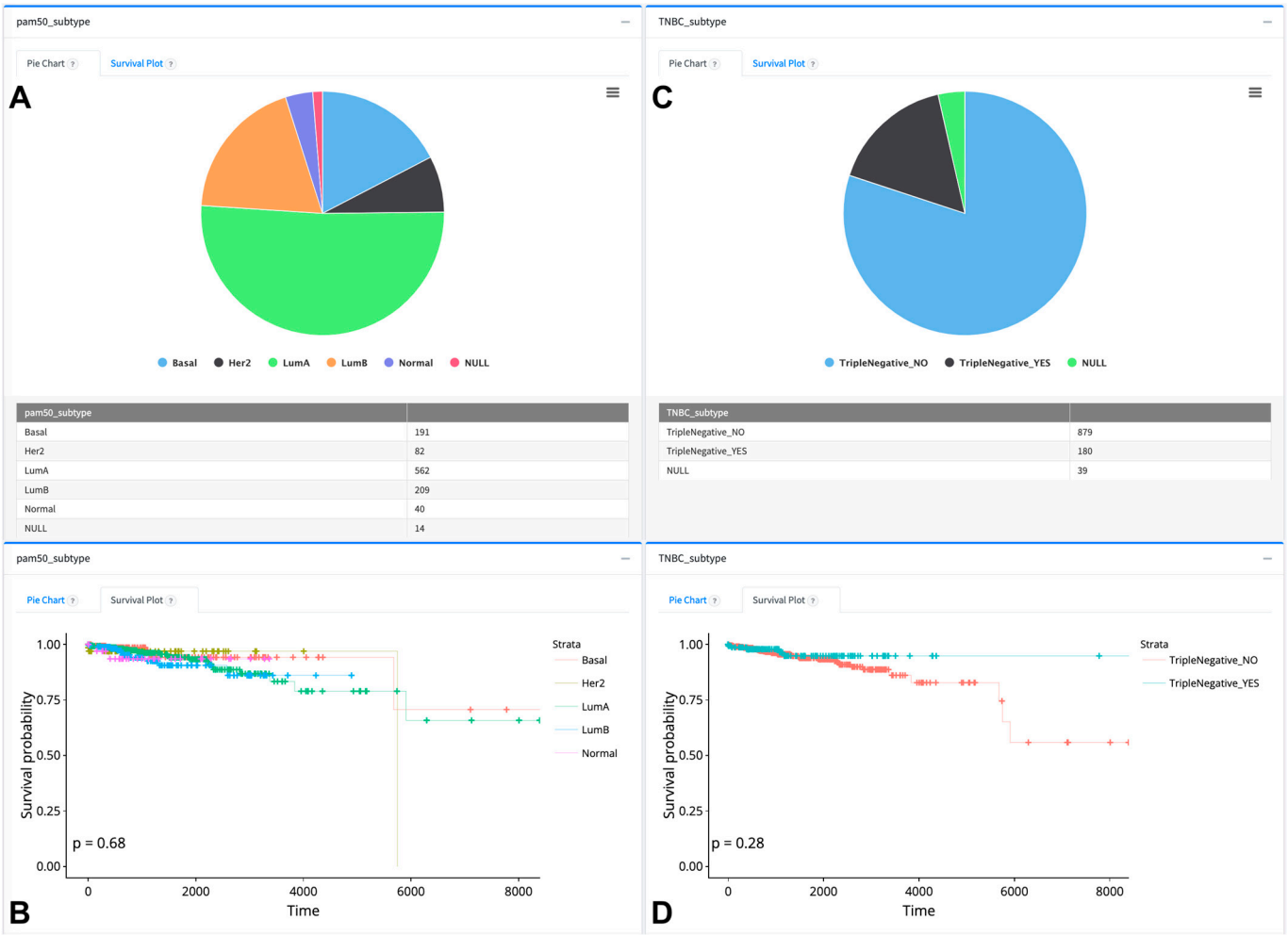


Figure 8. Sample of pie chart and survival analysis visualizations of BRCA PAM50 and TNBC subtypes as downloaded from the TCGAnalyzeR website. (**A**,**C**) Pie chart representation and number of BRCA patients in subtype groups. (**B**,**D**) Survival analysis of subtype groups. Each color dot in the survival analysis (**B**,**D**) represent a patient's number of survived days for each dynamically selected patient group. + sign indicates a censored patient.

When we parsed the metastasis organs in breast cancer clinical data, 71 patients with primary tumors contain metastasized organ information. Using the TCGAnalyzeR clinical tab, one can filter the pie chart and survival analysis using metastasis conditions by excluding the "No-metastasis" data. The final filtered pie chart shows that most of the 71 breast cancer patients have metastasis to bone or multi-tissue, and these patients have significantly less overall survival probability (Figure 10).

Radial slices of the pie charts are clickable, letting the user add the corresponding patient subsets to the "My Patients" clipboard panel. Furthermore, users can customize a variety of plots such as survival plot, volcano plot, box plot, heatmaps, lollipop plot, and pie charts for the purpose of discovering common molecular profiles for precision oncology. Each plot and data table are downloadable for use in articles.

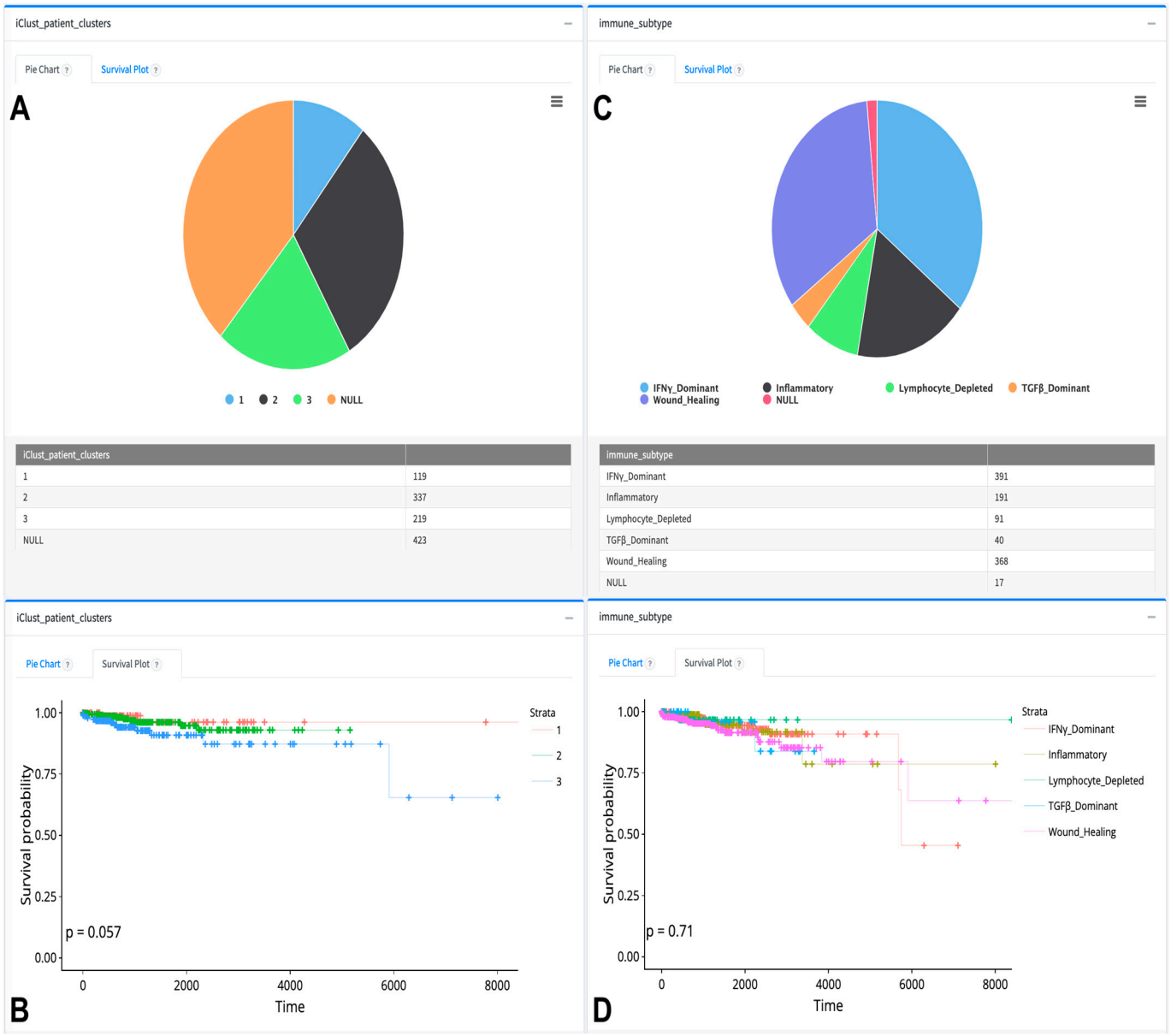


Figure 9. Sample of pie charts and survival analysis of iCluster and Immune subtypes as downloaded from the TCGAnalyzeR website. (**A**,**C**) Pie chart representation of BRCA patients who have metastasis information. (**B**,**D**) Survival analysis of BRCA patient groups with different tissue metastasis. Each color dot in the survival analysis (**B**,**D**) represent a patient's number of survived days for each dynamically selected patient group. + sign indicates a censored patient.

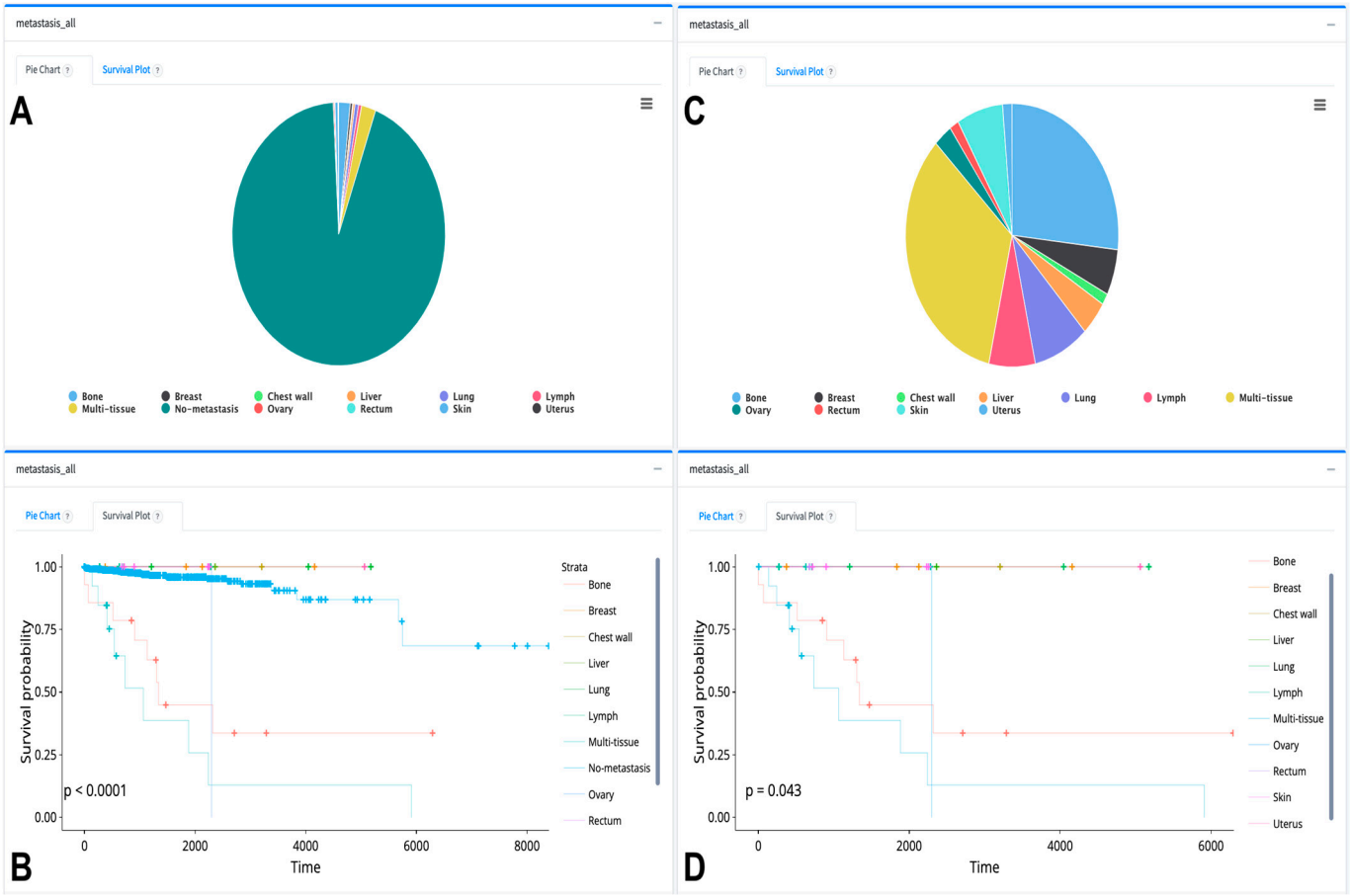


Figure 10. Sample of pie chart and survival analysis visualizations of patients with metastasis as downloaded from the TCGAnalyzeR website. (**A**,**C**) Pie chart representation of BRCA patients who have metastasis information. (**B**,**D**) Survival analysis of BRCA patient groups with different tissue metastasis. Each color dot in the survival analysis (**B**,**D**) represent a patient's number of survived days for each dynamically selected patient group. + sign indicates a censored patient.

4. Discussion

Several web portals facilitating analysis on TCGA data have been developed and widely used, such as the Genomic Data Commons (GDC) data portal [12], ICGC data portal [10], and CPTAC data portal [39]. The cBioPortal is an open-access, open-source resource for interactive exploration of multidimensional cancer genomics data sets [5,6] providing gene-centered query and visualization functions across multiple cancers. IntOGen is another similar framework for automated comprehensive knowledge extraction based on mutational data from sequenced tumor samples from TCGA patients [40]. However, we provide pre-performed SNV, CNV, and differential expression analyses with large sets of our own patient clusters and pre-computed patient subcohorts. We present signature-based clustering using the Generalized Linear Model for three cancer types (LUAD, LUSC, and COAD). For all 33 cancer types, the immune and MSI-sensor scores of all patients are retrieved from their original publications. For breast cancer (BRCA), PAM50 and TNBC patient cohorts are retrieved from their original publications and metastasis data is retrieved from BCR Biotab. For fifteen cancer types, previously published TCGA cohorts of the individual tumor types are retrieved by a curatedTCGAData R package [21]. iClusterPlus-based patient cohorts are generated for 32 cancer types based on five data dimensions: miRNA, methylation, single nucleotide variation, transcriptome, and copy number variation. A re-runnable parallel Linux pipeline is implemented, enabling a scalable update of the pan-cancer data at the backend.

TCGAnalyzeR provides a user-friendly web framework for integrative, large-scale analyses of genomic and clinical data of 33 cancer types from TCGA. The TCGAnalyzeR web interface allows cancer researchers to create subcohorts and/or gene sets of interest to filter through visualizations of the analyses.

5. Conclusions

TCGAnalyzeR provides a user-friendly web framework for integrative, large-scale analyses of the genomic and clinical data of 33 cancer types from TCGA. The TCGAnalyzeR web interface allows cancer researchers to create subcohorts and/or gene sets of interest to filter through visualizations of the analyses. For future work, we aim to integrate the subcohort targeting drug repurposing, miRNA, and methylation interfaces to TCGAnalyzeR. TCGAnalyzeR is freely available on the web at tcganalyzer.mu.edu.tr (accessed on 28 December 2023).

Author Contributions: Conceptualization, T.Z.; methodology, T.Z.; software, T.Z. and B.A.M.; validation, T.Z. and T.Ö.-S.; resources, T.Ö.-S.; data curation, T.Z.; writing—original draft preparation, T.Z.; writing—review and editing, T.Z. and T.Ö.-S.; visualization, B.A.M. and T.Z.; supervision, T.Ö.-S.; project administration, T.Ö.-S.; funding acquisition, T.Z. and T.Ö.-S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Turkish National Institutes of Health (TÜSEB) grant number 4583.

Institutional Review Board Statement: All the TCGA datasets are publicly available for anyone to use under the terms provided by the dataset source (https://cancergenome.nih.gov/ (accessed on 28 December 2023)) and are provided "AS IS" without any warranty, express or implied, from TCGAnalyzeR web site. All external patient subcohort data are downloaded from their original references and the references are indicated in the TCGAnalyzeR help section. The patient cohorts integrated in this current study are included in the PhD thesis of Talip Zengin and the M.Sc. thesis of Başak Abak Masud approved by Mugla Sitki Kocman University, Graduate School of Natural and Applied Sciences.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data and the codes presented in this study are available online at portal.gdc.cancer.gov (accessed on 28 December 2023), tcganalyzer.mu.edu.tr (accessed on 28 December 2023), and https://github.com/talipzengin/TCGAnalyzeR (accessed on 28 December 2023).

Conflicts of Interest: T.Z. and B.A.M. declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results. T.Ö.-S. was partially supported by the company Kedi Mobil Uygulamalar Limited Sirketi, Muğla, Türkiye during the study.

References

1. Sakaguchi, T.; Iketani, A.; Esumi, S.; Esumi, M.; Suzuki, Y.; Ito, K.; Fujiwara, K.; Nishii, Y.; Katsuta, K.; Yasui, H.; et al. Clinical Importance of the Range of Detectable Variants between the Oncomine Dx Target Test and a Conventional Single-Gene Test for EGFR Mutation. *Sci. Rep.* **2023**, *13*, 13759. [CrossRef]
2. Paik, S.; Shak, S.; Tang, G.; Kim, C.; Baker, J.; Cronin, M.; Baehner, F.L.; Walker, M.G.; Watson, D.; Park, T.; et al. A Multigene Assay to Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer. *N. Engl. J. Med.* **2004**, *351*, 2817–2826. [CrossRef] [PubMed]
3. Parker, J.S.; Mullins, M.; Cheang, M.C.U.; Leung, S.; Voduc, D.; Vickery, T.; Davies, S.; Fauron, C.; He, X.; Hu, Z.; et al. Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *J. Clin. Oncol.* **2009**, *27*, 1160–1167. [CrossRef]
4. The Cancer Genome Atlas Research Network; Weinstein, J.N.; Collisson, E.A.; Mills, G.B.; Shaw, K.R.M.; 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] [PubMed]
5. Cerami, E.; Gao, J.; Dogrusoz, U.; Gross, B.E.; Sumer, S.O.; Aksoy, B.A.; Jacobsen, A.; Byrne, C.J.; Heuer, M.L.; Larsson, E.; et al. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discov.* **2012**, *2*, 401–404. [CrossRef]

6. Gao, J.; Aksoy, B.A.; Dogrusoz, U.; Dresdner, G.; Gross, B.; Sumer, S.O.; Sun, Y.; Jacobsen, A.; Sinha, R.; Larsson, E.; et al. Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. *Sci. Signal.* **2013**, *6*, pl1. [CrossRef] [PubMed]
7. Deng, M.; Brägelmann, J.; Kryukov, I.; Saraiva-Agostinho, N.; Perner, S. FirebrowseR: An R Client to the Broad Institute's Firehose Pipeline. *Database* **2017**, *2017*, baw160. [CrossRef]
8. Goldman, M.J.; Craft, B.; Hastie, M.; Repečka, K.; McDade, F.; Kamath, A.; Banerjee, A.; Luo, Y.; Rogers, D.; Brooks, A.N.; et al. Visualizing and Interpreting Cancer Genomics Data via the Xena Platform. *Nat. Biotechnol.* **2020**, *38*, 675–678. [CrossRef]
9. Chakravarty, D.; Gao, J.; Phillips, S.; Kundra, R.; Zhang, H.; Wang, J.; Rudolph, J.E.; Yaeger, R.; Soumerai, T.; Nissan, M.H.; et al. OncoKB: A Precision Oncology Knowledge Base. *JCO Precis. Oncol.* **2017**, *2017*, PO.17.00011. [CrossRef]
10. Zhang, J.; Baran, J.; Cros, A.; Guberman, J.M.; Haider, S.; Hsu, J.; Liang, Y.; Rivkin, E.; Wang, J.; Whitty, B.; et al. International Cancer Genome Consortium Data Portal--a One-Stop Shop for Cancer Genomics Data. *Database* **2011**, *2011*, bar026. [CrossRef]
11. Adelberger, P.; Eckelt, K.; Bauer, M.J.; Streit, M.; Haslinger, C.; Zichner, T. Coral: A Web-Based Visual Analysis Tool for Creating and Characterizing Cohorts. *Bioinformatics* **2021**, *37*, 4559–4561. [CrossRef] [PubMed]
12. 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]
13. Mo, Q.; Shen, R. iClusterPlus: Integrative Clustering of Multi-Type Genomic Data. 2022. Available online: https://bioconductor.org/packages/release/bioc/html/iClusterPlus.html (accessed on 28 December 2023).
14. Zengin, T.; Önal-Süzek, T. Analysis of Genomic and Transcriptomic Variations as Prognostic Signature for Lung Adenocarcinoma. *BMC Bioinform.* **2020**, *21*, 368. [CrossRef] [PubMed]
15. Zengin, T.; Önal-Süzek, T. Comprehensive Profiling of Genomic and Transcriptomic Differences between Risk Groups of Lung Adenocarcinoma and Lung Squamous Cell Carcinoma. *J. Pers. Med.* **2021**, *11*, 154. [CrossRef]
16. Thorsson, V.; Gibbs, D.L.; Brown, S.D.; Wolf, D.; Bortone, D.S.; Ou Yang, T.-H.; Porta-Pardo, E.; Gao, G.F.; Plaisier, C.L.; Eddy, J.A.; et al. The Immune Landscape of Cancer. *Immunity* **2018**, *48*, 812–830.e14. [CrossRef]
17. Lehmann, B.D.; Jovanović, B.; Chen, X.; Estrada, M.V.; Johnson, K.N.; Shyr, Y.; Moses, H.L.; Sanders, M.E.; Pietenpol, J.A. Refinement of Triple-Negative Breast Cancer Molecular Subtypes: Implications for Neoadjuvant Chemotherapy Selection. *PLoS ONE* **2016**, *11*, e0157368. [CrossRef]
18. Berger, A.C.; Korkut, A.; Kanchi, R.S.; Hegde, A.M.; Lenoir, W.; Liu, W.; Liu, Y.; Fan, H.; Shen, H.; Ravikumar, V.; et al. A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. *Cancer Cell* **2018**, *33*, 690–705.e9. [CrossRef]
19. Carey, V. BiocOncoTK 2018. Available online: https://www.bioconductor.org/packages/release/bioc/html/BiocOncoTK.html (accessed on 28 December 2023).
20. Ding, L.; Bailey, M.H.; Porta-Pardo, E.; Thorsson, V.; Colaprico, A.; Bertrand, D.; Gibbs, D.L.; Weerasinghe, A.; Huang, K.; Tokheim, C.; et al. Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. *Cell* **2018**, *173*, 305–320.e10. [CrossRef]
21. Ramos, M.; Geistlinger, L.; Oh, S.; Schiffer, L.; Azhar, R.; Kodali, H.; de Bruijn, I.; Gao, J.; Carey, V.J.; Morgan, M.; et al. Multiomic Integration of Public Oncology Databases in Bioconductor. *JCO Clin. Cancer Inform.* **2020**, *4*, 958–971. [CrossRef]
22. Colaprico, A.; Silva, T.C.; Olsen, C.; Garofano, L.; Cava, C.; Garolini, D.; Sabedot, T.S.; Malta, T.M.; Pagnotta, S.M.; Castiglioni, I.; et al. TCGAbiolinks: An R/Bioconductor Package for Integrative Analysis of TCGA Data. *Nucleic Acids Res.* **2016**, *44*, e71. [CrossRef]
23. Van den Eynden, J.; Fierro, A.C.; Verbeke, L.P.; Marchal, K. SomInaClust: Detection of Cancer Genes Based on Somatic Mutation Patterns of Inactivation and Clustering. *BMC Bioinform.* **2015**, *16*, 125. [CrossRef]
24. Morganella, S.; Pagnotta, S.M.; Ceccarelli, M. Finding Recurrent Copy Number Alterations Preserving Within-Sample Homogeneity. *Bioinformatics* **2011**, *27*, 2949–2956. [CrossRef] [PubMed]
25. Lawrence, M.; Huber, W.; Pagès, H.; Aboyoun, P.; Carlson, M.; Gentleman, R.; Morgan, M.T.; Carey, V.J. Software for Computing and Annotating Genomic Ranges. *PLOS Comput. Biol.* **2013**, *9*, e1003118. [CrossRef]
26. 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]
27. Robinson, M.D.; McCarthy, D.J.; Smyth, G.K. edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data. *Bioinformatics* **2010**, *26*, 139–140. [CrossRef] [PubMed]
28. Ritchie, M.E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C.W.; Shi, W.; Smyth, G.K. Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies. *Nucleic Acids Res.* **2015**, *43*, e47. [CrossRef] [PubMed]
29. Yu, G.; Wang, L.-G.; Han, Y.; He, Q.-Y. clusterProfiler: An R Package for Comparing Biological Themes Among Gene Clusters. *OMICS A J. Integr. Biol.* **2012**, *16*, 284–287. [CrossRef]
30. Mo, Q.; Shen, R.; Guo, C.; Vannucci, M.; Chan, K.S.; Hilsenbeck, S.G. A Fully Bayesian Latent Variable Model for Integrative Clustering Analysis of Multi-Type Omics Data. *Biostatistics* **2018**, *19*, 71–86. [CrossRef]
31. Therneau, T. A Package for Survival Analysis in R 2022. Available online: https://cran.r-project.org/web/packages/survival/ (accessed on 28 December 2023).
32. Wickham, H.; Hester, J.; Bryan, J. Readr: Read Rectangular Text Data 2022. Available online: https://cran.r-project.org/web/packages/readr/ (accessed on 28 December 2023).

33. Galili, T.; O'Callaghan, A.; Sidi, J.; Sievert, C. Heatmaply: An R Package for Creating Interactive Cluster Heatmaps for Online Publishing. *Bioinformatics* **2018**, *34*, 1600–1602. [CrossRef]
34. Guo, X.; Zhang, B.; Zeng, W.; Zhao, S.; Ge, D. G3viz: An R Package to Interactively Visualize Genetic Mutation Data Using a Lollipop-Diagram. *Bioinformatics* **2019**, *36*, 928–929. [CrossRef]
35. Kunst, J. Highcharter: A Wrapper for the "Highcharts" Library 2022. Available online: https://cran.r-project.org/web/packages/highcharter/ (accessed on 28 December 2023).
36. Bose, R.; Kavuri, S.M.; Searleman, A.C.; Shen, W.; Shen, D.; Koboldt, D.C.; Monsey, J.; Goel, N.; Aronson, A.B.; Li, S.; et al. Activating HER2 Mutations in HER2 Gene Amplification Negative Breast Cancer. *Cancer Discov.* **2013**, *3*, 224–237. [CrossRef] [PubMed]
37. Kurozumi, S.; Alsaleem, M.; Monteiro, C.J.; Bhardwaj, K.; Joosten, S.E.P.; Fujii, T.; Shirabe, K.; Green, A.R.; Ellis, I.O.; Rakha, E.A.; et al. Targetable ERBB2 Mutation Status Is an Independent Marker of Adverse Prognosis in Estrogen Receptor Positive, ERBB2 Non-Amplified Primary Lobular Breast Carcinoma: A Retrospective in Silico Analysis of Public Datasets. *Breast Cancer Res.* **2020**, *22*, 85. [CrossRef] [PubMed]
38. Aljohani, A.I.; Toss, M.S.; Green, A.R.; Rakha, E.A. The Clinical Significance of Cyclin B1 (CCNB1) in Invasive Breast Cancer with Emphasis on Its Contribution to Lymphovascular Invasion Development. *Breast Cancer Res. Treat.* **2023**, *198*, 423–435. [CrossRef] [PubMed]
39. Edwards, N.J.; Oberti, M.; Thangudu, R.R.; Cai, S.; McGarvey, P.B.; Jacob, S.; Madhavan, S.; Ketchum, K.A. The CPTAC Data Portal: A Resource for Cancer Proteomics Research. *J. Proteome Res.* **2015**, *14*, 2707–2713. [CrossRef]
40. Martínez-Jiménez, F.; Muiños, F.; Sentís, I.; Deu-Pons, J.; Reyes-Salazar, I.; Arnedo-Pac, C.; Mularoni, L.; Pich, O.; Bonet, J.; Kranas, H.; et al. A Compendium of Mutational Cancer Driver Genes. *Nat. Rev. Cancer* **2020**, *20*, 555–572. [CrossRef]