TCGAnalyzeR: a web application for integrative visualization of molecular and clinical data of cancer patients for cohort and associated gene discovery

Talip Zengin[1,2,☯], Başak Abak Masud[2,☯], Tuğba Önal-Süzek[2,3,*]

[1]Department of Molecular Biology and Genetics, Mugla Sitki Kocman University, Mugla, 48000, Turkiye [2]Department of Bioinformatics, Mugla Sitki Kocman University, Mugla, 48000, Turkiye, [3]Department of Computer Engineering, Mugla Sitki Kocman University, Mugla, 48000, Turkiye.

*Corresponding Author.

☯Joint first authors.

Abstract

Motivation: The vast size and complexity of The Cancer Genome Atlas (TCGA) database with multidimensional molecular and clinical data of ~11,000 cancer patients of 33 cancer types challenge the effective utilization of this valuable resource. Therefore, we built a web application named TCGAnalyzeR with the main idea of presenting an integrative visualization of mutations, transcriptome profile, copy number variation and clinical data allowing researchers to facilitate the identification of customized patient cohorts and gene sets for better decision-making for oncologists and cancer researchers.

Results: We present TCGAnalyzeR for integrative visualization of pre-analyzed TCGA data with the several novel modules: (i) Simple nucleotide variations with driver prediction; (ii) Recurrent copy number alterations; (iii) Differential expression in tumor versus normal, with pathway enrichment and the survival analysis; (iii) TCGA clinical data and survival analysis; (iv) External subcohorts from literature, curatedTCGAData and BiocOncoTK R packages; (v) Internal patient clusters determined using iClusterPlus R package or signature-based expression analysis. TCGAnalyzeR provides clinical oncologists and cancer researchers interactive and integrative representations of these multi-omic, pan-cancer TCGA data with availability of subcohort analysis and visualization. TCGAnalyzeR can be used to create their own custom gene sets for pan-cancer comparisons, to create custom patient subcohorts comparing external subcohorts (MSI, Immune, PAM50, Triple Negative, IDH1, miRNA, etc) along with our internal patient clusters, to visualize cohort-centric or gene-centric results along with pathway enrichment and survival analysis graphically on an interactive web tool.

Availability: TCGAnalyzeR is freely available on the web at http://tcganalyzer.mu.edu.tr.

Contact: tugbasuzek@mu.edu.tr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The sheer scale and complexity of The Cancer Genome Atlas (TCGA) data (The Cancer Genome Atlas Research Network. *et al.*, 2013) 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 (Cerami *et al.*, 2012; Gao *et al.*, 2013), Firebrowse (Deng *et al.*, 2017), and University of California, Santa Cruz (UCSC) Xena (Goldman *et al.*, 2020). Among these tools, cBioPortal is the most preferred due to its interactive exploration of larger and up-to-date cancer datasets. OncoKB (Chakravarty et al., 2017) is another precision oncology knowledge base that allows searching and comparing drug response data from different TCGA cohorts. Although the ICGC web portal (Zhang et al., 2019) and Coral web application (Adelberger *et al.*, 2021) allow patient/gene subsetting of TCGA cohorts and provide survival and Venn diagram visualization of cohorts, they do not allow comparison of cohorts pre-generated by other research groups. These tools address the reuse needs of users. However, they only provide raw data without allowing the patient/gene subsetting of the statistical analyses for pan-cancer subcohort and associated gene discovery.

We built an interactive Shiny (Chang *et al.*, 2022) 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 National Cancer Institute (NCI) Genomic Data Commons (GDC) (Grossman *et al.*, 2016). Moreover, users can compare patient clusters determined using iClusterPlus R package (Mo and Shen, 2022) with expression-based survival risk groups (Zengin and Önal-Süzek, 2020, 2021), and curated subtypes such as immune subtypes (Thorsson *et al,* 2019), Triple Negative Breast Cancer (TNBC)

subtypes (Lehmann *et al*, 2016), PAM50 subtypes (Berger *et al*, 2018) and Microsatellite Instability (MSI) related subgroups and several data type clusters from BiocOncoTK (Carey, 2022; Ding *et al*, 2018) and *curatedTCGAData* (Ramos *et al.*, 2020) R packages. Furthermore, users can create custom subcohorts based on genomic analyses and/or clinical data 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 allowing the users to generate custom patient sub-cohorts and/or gene sets using interactive graphical representations via clipboard functionality.

2 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 March 6, 2022 from NCI GDC (Grossman *et al.*, 2016) using TCGAbiolinks R package (Colaprico *et al.*, 2016).

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 (Van den Eynden *et al.*, 2015) using 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.

2.2.2 CNV Analysis

Significant recurrent copy number variations were identified by *GAIA* R package (Morganella *et al.*, 2011). NCBI IDs and Hugo Symbols of the genes on chromosomal regions with altered copy numbers were determined using *GenomicRanges* (Lawrence *et al.*, 2013) and *biomaRt* (Durinck *et al.*, 2009) R packages.

2.2.3 Differential Gene Expression Analysis

Differentially expressed genes were determined using normalized HTseq counts, by limma-voom method with duplicate-correlation function from *edgeR* (Robinson *et al.*, 2010) and *limma* (Ritchie *et al.*, 2015) R packages. Ensembl IDs were converted to NCBI IDs and Hugo Symbols using the *biomaRt* package (Durinck *et al.*, 2009).

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 is available for a particular cancer.

2.2.4 Pathway Enrichment

Pathway enrichment and visualization was performed for each analysis by *clusterProfiler* R package (Yu *et al.*, 2012).

2.3 Pre-computed Patient Clusters and Sample Subtypes

TCGAnalyzerR 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) (Zengin and Önal-Süzek, 2020, 2021), ii) iClusters: We clustered patients using their raw SNV, CNV, gene expression, miRNA expression and methylation data of tumor samples which have all types of data by iClusterBayes method (Mo *et al.*, 2018), iii) Curated subtypes (immune subtypes, TNBC subtypes, PAM50 subtypes) from original publications (Thorsson *et al,* 2019; Lehmann *et al*, 2016; Berger *et al*, 2018). For fifteen cancer types, previously published TCGA cohorts of the individual tumor types are retrieved from curatedTCGA R package curatedTCGAData R package (Ramos *et al.*, 2020). Patient clusters based on Microsatellite Instability (MSI) were compiled using BiocOncoTK (Carey, 2022, Ding *et al.*, 2018) and Immune clusters (Thorsson et al, 2019) were compiled for all 33 cancers. In total, 123 external patient cohorts are integrated into the web interface allowing efficient filtering and cross-comparative analysis of multiple subcohorts in parallel.

2.4 Survival Analysis

Kaplan-Meier (KM) survival analysis is performed by *readr* (Wickham *et al.*, 2022) and *survfit* (Therneau, 2022) R packages in real-time based on overall survival data of patients of interest for selected clinical features.

2.5 Visualization

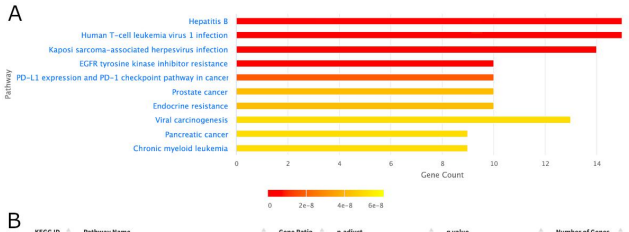
TCGAnalyzeR front-end was implemented by javascript-based R packages with an interactive dashboard enabling users to select cancer types, data types, risk groups and patient cohorts using *heatmaply, g3viz* and *highcharter* R packages (Galili *et al.*, 2018; Guo *et al.*, 2019; Kunst, 2022). All visualizations are interactive and customizable by the user through the filtration options with "My genes" and/or "My patients" panels enabling to copy genes and/or patients of interest to the clipboard. TCGAnalyzeR currently supports TSV for downloading tables; and high-resolution PNG format for downloading figures.

3 Results

TCGAnalyzeR web application offers simple nucleotide (SNV) analysis as the 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. Oncoplot in Figure 1 shows candidate driver genes with their percentage in tumor samples of Breast Invasive Carcinoma (BRCA) with annotations of patient iClusters, PAM50, TNBC and immune subtypes. iCluster #1 is highly correlated with the Basal and TNBC subtype. Wound-healing and IFNɣ-dominant immune subtypes gather around iCluster #1. iCluster #2 is mostly correlated with 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 #1 and iCluster #2 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 have higher prevalence of TP53 mutations with very few mutations of PIK3CA, CDH1, GATA3, KMT2C, MAP3K1 genes. Moreover, mutations of TP53, CDH1 and GATA3 genes are mutually exclusive (Figure 1).

Fig. 1. Oncoplot of candidate driver genes with patient clusters and sample subtypes. Top10 significant candidate driver genes with mutations determined by SomInaClust R package. Bottom annotations show the sample subtypes curated from the literature and iClusters.

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 pathway 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).



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	0.000000015747493934740 2	0.000000005525436468329 89	15
hsa05167	Kaposi sarcoma-associated herpesvirus infection	14/66	0.000000019453193809602 8	0.000000006825049757755 36	14
hsa01521	EGFR tyrosine kinase inhibitor resistance	10/66	0.000000026068192204499 5	0.000000091467341068419 2	10
Genes in the pathway: PIK3CA/NF1/PTEN/JAK2/AKT1/HRAS/ERBB2/KRAS/ERBB3/JAK1					
hsa05235	PD-L1 expression and PD-1 checkpoint pathway in cancer	10/66	0.000000069094441165716	0.000000024243663549674 3	10
hsa05215	Prostate cancer	10/66	0.000000128403619028508	0.000000045053904135117	10
hsa01522	Endocrine resistance	10/66	0.000000128403619028508	0.000000045053904135117	10
hsa05203	Viral carcinogenesis	13/66	0.000000148057644482855	0.000000051950050695738 7	13
hsa05212	Pancreatic cancer	9/66	0.000000148057644482855	0.000000051950050695738 7	9
hsa05220	Chronic myeloid leukemia	9/66	0.000000148057644482855	0.000000051950050695738 7	9

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

Fig. 2. Pathway enrichment of candidate driver genes. A. Bar plot showing top 10 significant pathways of 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 "My genes" clipboard panel of TCGAnalyzeR allows modifying plots to show genes of interest. For example, Figure 3 shows the mutation pattern of Oncotype DX gene set together with clinical annotations. iCluster #2, Luminal A subtype and Her2 subtypes are related with ERBB2 (HER2) mutations. Besides, iCluster #1 have fewer mutations than the other two iClusters. Moreover, mutations of Oncotype DX genes are mostly mutually exclusive (Figure 3).

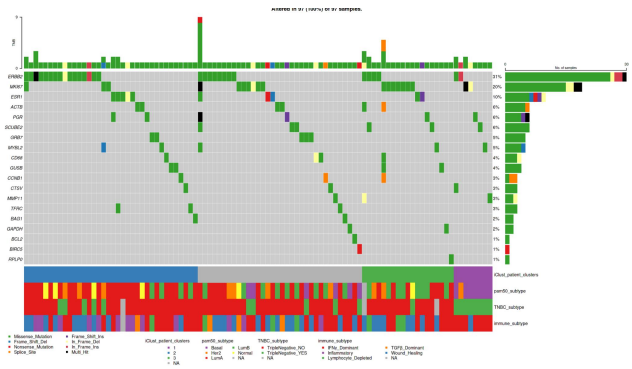
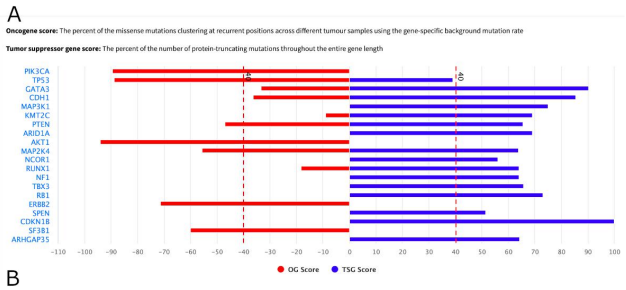


Fig. 3. Oncoplot of Oncotype DX genes with patient clusters and sample subtypes. Mutations of Oncotype DX genes with annotations showing the patient iClusters and sample subtypes curated from the literature.

SomInaClust R package determines candidate driver mutated genes with their potential roles as a tumor suppressor (TSG) or oncogene (OG) with predicted scores (Van den Eynden *et al.*, 2015). Pyramid plot in Figure 4A summarizes OG score and TSG scores of candidate driver genes ranked by their analysis q-values. Some genes may have both OG score and TSG score over threshold, in that case, SomInaClust considers the COSMIC cancer gene census information. Only 1 (ERBB2) of 21 Oncotype DX genes were predicted as significant driver mutated genes. ERBB2 is predicted as an oncogene by SomInaClust as listed Dominant (OG) in COSMIC cancer gene census (Figure 4B).

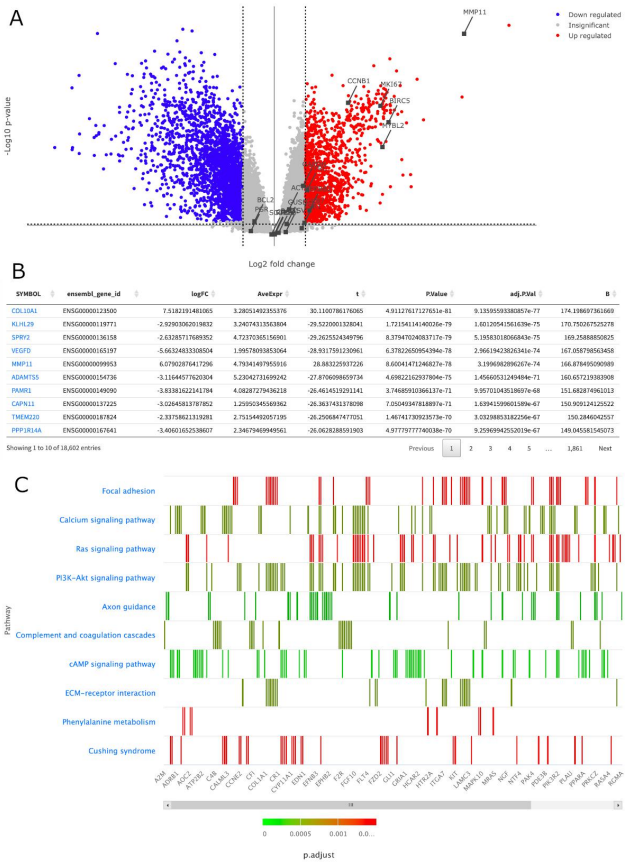


CGC	Gene	# Mutations	qDG	OG Score	TSG Score	Classification
Rec	NCOR1	55	1.9e-22	0	56	TSG
Dom	RUNX1	40	4.9e-22	18.2	64.1	TSG
Rec	NF1	41	4.9e-22	0	64.1	TSG
Dom	TBX3	38	1.53e-20	0	65.7	TSG
Rec	RB1	26	4.07e-18	0	73.1	TSG
Dom	ERBB2	32	5.71e-18	71.4	0	OG
Rec	SPEN	42	4.15e-13	0	51.4	TSG
Rec	CDKN1B	11	7.5e-13		100	TSG
Dom	SF3B1	19	1.48e-11	60	0	OG
	ARHGAP35	16	6.62e-7	0	64.3	TSG

Showing 11 to 20 of 104 entries Previous 1 2 3 4 5 ... 11 Next

Fig. 4. SomInaClust prediction of candidate driver genes. A. Pyramid plot showing Oncotype DX genes of which were predicted as candidate driver genes with calculated oncogene (OG) and tumor-suppressor (TSG) scores by SomInaClust R package. B. Detailed results of SomInaClust analysis with number of mutations, OG score, TSG score and q-value (qDG). CGC: COSMIC cancer gene census, Rec: Recessive (TSG), Dom: Dominant (OG).

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 results options for this analysis: 'Paired' as comparison of tumor samples against their own paired normal or 'All' as comparison of tumor samples against a normal sample subset of patients if such is available for the particular cancer. Volcano plot in Figure 5A summarizes the differential expression analysis of paired BRCA samples and Oncotype DX genes highlighted through the 'My Genes' panel. Table in Figure 5B presents the details of DEA with gene symbol, fold change (logFC) and p values of top 10 significantly differentially expressed genes ranked by p-value.



SYMBOL	ensembl_gene_id	logFC	AveExpr	t	P.Value	adj.P.Val	B
COL10A1	ENSG00000123500	7.5182191481065	3.280514923553376	30.11007861760265	4.91127617127651e-81	9.13595593380857e-77	174.198697361669
KLHL29	ENSG00000119771	-2.92903062019832	3.240743135638 04	-29.52200013280041	1.721545114140026e-79	1.60120541561639e-75	170.750267525278
SPRY2	ENSG00000136158	-2.63285717689352	4.723703651569 01	-29.26255243497796	8.37947024083717e-79	5.19583018066843e-75	169.258888505825
VEGFD	ENSG00000165197	-5.66324833308504	1.995780938530 64	-28.93175912303961	6.37822650954394e-78	2.96619423826341e-74	167.058798563458
MMP11	ENSG00000099953	6.07902876417296	4.793414979559 16	28.883225937226	8.600414171246827e-78	3.199698289622 67e-74	166.878495090989
ADAMTS5	ENSG00000154736	-3.11644577620304	5.230427316992 42	-27.87609865973 4	4.698221629378 04e-75	1.456605312494 84e-71	160.657219383908
PAMR1	ENSG00000149090	-3.83381622141784	4.082872794362 18	-26.46145192914 1	3.746859103661 37e-71	9.957010435186 97e-68	151.682874961013
CAPN11	ENSG00000137225	-3.02645813787852	1.259503455693 62	-26.36374313780 98	7.050493478188 97e-71	1.639415996015 89e-67	150.909124125522
TMEM220	ENSG00000187824	-2.33758621319281	2.751544920571 95	-26.25068474770 51	1.467417309235 73e-70	3.032988531822 56e-67	150.2846042557
PPP1R14A	ENSG00000167641	-3.40601652538607	2.346794699495 61	-26.06282885919 03	4.977797777400 38e-70	9.259699425520 19e-67	149.045581545073

Showing 1 to 10 of 18,602 entries Previous 1 2 3 4 5 ... 1,861 Next

Fig. 5. Differential expression of genes in tumor samples versus normal samples. A. Volcano plot showing up-regulated and down-regulated genes with -log10 conversion of p-values. Oncotype DX genes are highlighted on the graph. B. Differential expression results table presenting gene symbols, fold changes (logFC) and adjusted-p-values. C. Heatmap showing pathway enrichment of differentially expressed genes.

Pathway enrichment of differentially expressed genes showed that these genes play 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). Genes related to these pathways can be observed with their p value color representation at heat plot of pathway enrichment analysis (Figure 5C).

Metastasis related gene MMP11, proliferation related genes BIRC5, MYBL2, MKI67 (Ki67), AURKA (STK15), CCNB1 and ERBB2 (HER2) from Oncotype DX gene set are highly up-regulated significantly in tumor samples of BRCA (Figure 5A). However hormone related genes (BAG1, BCL2, CD68, ESR1 (ER), GSTM1, PGR, SCUBE2) are not significantly differentially expressed among all tumor samples.

When we concentrate on the ERBB2 gene because it was predicted as a driver oncogene, we can visualize positions of mutations by the Lollipop plot in Figure 6A. Most of the mutations of ERBB2 gene are located on the kinase domain of HER2 (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), 939 (n=1) and in frame insertion on protein position 885 (n=1). Mutations on ERBB2 gene in tumor samples cause lower survival probability with 1.43 hazard ratio (not significant, p=0.084) (Figure 6B).

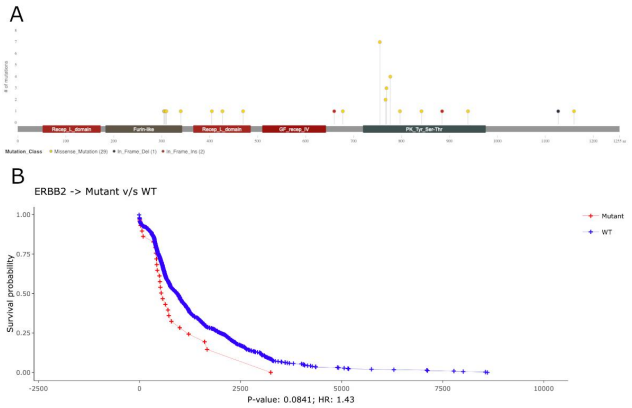


Fig. 6. Detailed analysis of ERBB2 (HER2) mutations. A. Lollipop plot showing mutations of ERBB2 gene among tumor samples. B. Overall survival analysis of wild type versus mutated ERBB2 in tumor samples.

When we checked the expression levels of ERBB2 in tumor samples versus normal samples, from paired DEA, ERBB2 is expressed in tumor samples significantly higher than their adjacent normal samples (p=3.521E-10) (Figure 7A). However, patients with higher expression of ERBB2 have higher survival probability significantly (p=0.045) (Figure 7B).

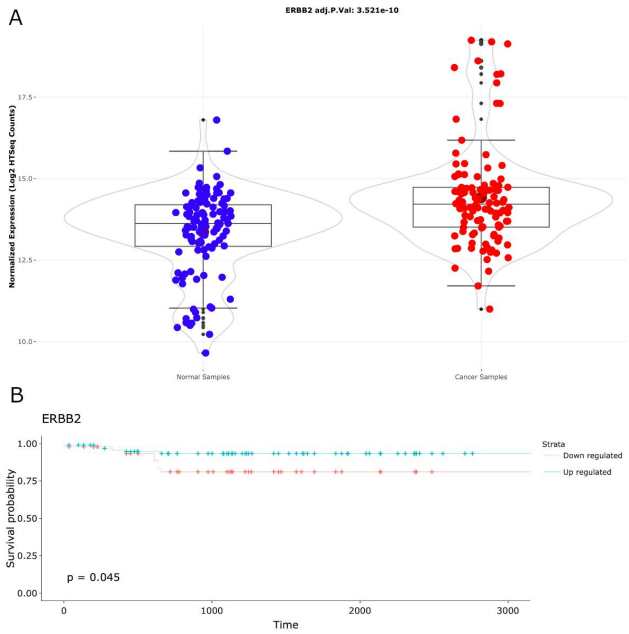


Fig. 7. Detailed analysis of ERBB2 (HER2) expression. A. Violin plot presenting log2 transformed normalized mRNA expression of ERBB2 in normal and tumor samples with adjusted p-value. B. Overall survival analysis of expression levels of ERBB2 in tumor samples.

Clinical data analysis is composed of pie chart visualization and survival analysis of clinical features with curated patient clusters and sample subtypes. Figure 8 shows the visualization of iClusters, PAM50 subtypes and MSI-sensor subtypes. iClusters showed differential survival probability close to significance level (p=0.057) (Figure 8A), however PAM50 subtypes do not have differential survival probabilities. (p=0.68) (Figure 8B). Patients who have tumors with MSI have lower survival probability than patients with MSI stable tumors (p=0.0022) (Figure 8C).

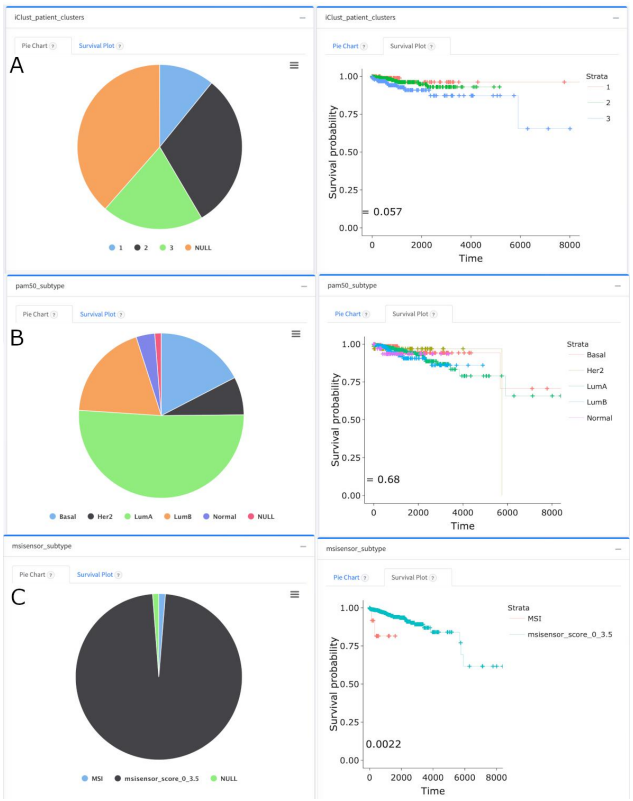


Fig. 8. Pie charts and survival analysis of clinical features. A. Pie chart representation and survival analysis of iClusters. B. Pie chart representation and survival analysis of PAM50 subtypes. C. Pie chart representation and survival analysis of MSI subtypes.

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

4 Conclusion

Several web portals facilitating analysis on TCGA data have been developed and widely used such as Genomic Data Commons (GDC) data portal (Grossman, 2016), ICGC data portal (Zhang, 2019) and CPTAC data portal (Edwards, 2015). The cBioPortal is an open-access, open-source resource for interactive exploration of multidimensional cancer genomics data sets (Cerami, 2012; Gao, 2013) 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 (Francisco, 2020). However, we provide pre-performed SNV, CNV and differentially expression analyses with large sets of patient clusters and sample subtype information. We present signature based clustering using Generalized Linear Model for three cancer types (LUAD, LUSC and COAD). For all 33 cancer types immune and MSI-sensor scores of all patients are retrieved from their original publications. For the breast cancer (BRCA), PAM50 and TNBC patient cohorts are retrieved from their original publications. For fifteen cancer types, previously published TCGA cohorts of the individual tumor types are retrieved by curatedTCGAData R package (Ramos et al, 2020). By the time this manuscript is written iClusterPlus based patient cohorts are generated for fifteen cancers 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. We plan to generate iClusters for 33 cancer types, and integrate results of miRNA and methylation analyses, too. Since its initial inception to public in January 1st of 2022, TCGAnalyzeR has been regularly accessed by ~79 unique users/day.

TCGAnalyzeR provides a user-friendly web framework for integrative, large-scale analyses of genomic and clinical data of 33 cancer types from TCGA. TCGAnalyzeR web interface allows clinical oncologists and cancer researchers to create subcohorts and/or gene sets of interest to filter visualization of analyses. TCGAnalyzeR help page includes a demonstration of the app with the two use-cases of subcohort discovery and can be used as a manual.

Funding

This work has been supported by Turkish National Institutes of Health (TÜSEB) [4583].

Conflict of Interest: none declared.

References

Berger AC, Korkut A, Kanchi RS, Hegde AM, Lenoir W, Liu W, Liu Y, Fan H, Shen H, Ravikumar V, Rao A, Schultz A, Li X, Sumazin P, Williams C, Mestdagh P, Gunaratne PH, Yau C, Bowlby R, Robertson AG, Tiezzi DG, Wang C, Cherniack AD, Godwin AK, Kuderer NM, Rader JS, Zuna RE, Sood AK, Lazar AJ, Ojesina AI, Adebamowo C, Adebamowo SN, Baggerly KA, Chen TW, Chiu HS, Lefever S, Liu L, MacKenzie K, Orsulic S, Roszik J, Shelley CS, Song Q, Vellano CP, Wentzensen N; Cancer Genome Atlas Research Network; Weinstein JN, Mills GB, Levine DA, Akbani R. A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. Cancer Cell. 2018 Apr 9;33(4):690-705.e9. doi: 10.1016/j.ccell.2018.03.014. Epub 2018 Apr 2. PMID: 29622464; PMCID: PMC5959730.

Carey V (2022). _BiocOncoTK: Bioconductor components for general cancer genomics_. R package version 1.18.0.

Cerami, E. et al. (2012) The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. Cancer Discov., 2, 401–404.

Chakravarty, D. et al (2017), OncoKB: A Precision Oncology Knowledge Base. JCO Precis Oncol. 2017 Jul;2017:PO.17.00011. doi: 10.1200/PO.17.00011. Epub 2017 May 16.

Colaprico,A. et al. (2016) TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res., 44, e71–e71.

Deng,M. et al. (2017) FirebrowseR: an R client to the Broad Institute's Firehose Pipeline. Database, 2017, baw160.

Ding L, Bailey MH, Porta-Pardo E, Thorsson V, Colaprico A, Bertrand D, Gibbs DL, Weerasinghe A, Huang KL, Tokheim C, Cortés-Ciriano I, Jayasinghe R, Chen F, Yu L, Sun S, Olsen C, Kim J, Taylor AM, Cherniack AD, Akbani R, Suphavilai C, Nagarajan N, Stuart JM, Mills GB, Wyczalkowski MA, Vincent BG, Hutter CM, Zenklusen JC, Hoadley KA, Wendl MC, Shmulevich L, Lazar AJ, Wheeler DA, Getz G; Cancer Genome Atlas Research Network. Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. Cell. 2018 Apr 5;173(2):305-320.e10.

Durinck,S. et al. (2009) Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat. Protoc., 4, 1184–1191.

Galili,T. et al. (2018) heatmaply: an R package for creating interactive cluster heatmaps for online publishing. Bioinformatics, 34, 1600–1602.

Gao,J. et al. (2013) Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. Sci. Signal., 6, pl1–pl1.

Goldman,M.J. et al. (2020) Visualizing and interpreting cancer genomics data via the Xena platform. Nat. Biotechnol., 38, 675–678.

Grossman,R.L. et al. (2016) Toward a Shared Vision for Cancer Genomic Data. N. Engl. J. Med., 375, 1109–1112.

Guo,X. et al. (2019) G3viz: an R package to interactively visualize genetic mutation data using a lollipop-diagram. Bioinformatics, btz631.

Kassambara A. (2022) survminer: Drawing Survival Curves using 'ggplot2'

Kunst,J. (2022) highcharter: A Wrapper for the 'Highcharts' Library.

Lawrence,M. et al. (2013) Software for Computing and Annotating Genomic Ranges. PLOS Comput. Biol., 9, e1003118.

Lehmann BD, Jovanović B, Chen X, Estrada MV, Johnson KN, Shyr Y, Moses HL, Sanders ME, Pietenpol JA. Refinement of Triple-Negative Breast Cancer Molecular Subtypes: Implications for Neoadjuvant Chemotherapy Selection. PLoS One. 2016 Jun 16;11(6):e0157368. doi: 10.1371/journal.pone.0157368. PMID: 27310713; PMCID: PMC4911051.

Mo,Q. et al. (2018) A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. Biostatistics, 19, 71–86.

Mo,Q. and Shen,R. (2022) iClusterPlus: Integrative clustering of multi-type genomic data.

Morganella,S. et al. (2011) Finding recurrent copy number alterations preserving within-sample homogeneity. Bioinformatics, 27, 2949–2956.

Ramos,M. et al. (2020) Multiomic Integration of Public Oncology Databases in Bioconductor. JCO Clin. Cancer Inform., 958–971.

Ritchie,M.E. et al. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res., 43, e47–e47.

Robinson,M.D. et al. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26, 139–140.

Sakaguchi T, Iketani A, Furuhashi K, et al. Comparison of the analytical performance between the Oncomine Dx Target Test and a conventional single gene test for epidermal growth factor receptor mutation in non-small cell lung cancer. Thorac Cancer. 2021;12(4):462-467. doi:10.1111/1759-7714.13767

The Cancer Genome Atlas Research Network. et al. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet., 45, 1113–1120.

Therneau,T. (2022) A Package for Survival Analysis in R.

Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang TH, Porta-Pardo E, Gao GF, Plaisier CL, Eddy JA, Ziv E, Culhane AC, Paull EO, Sivakumar IKA, Gentles AJ, Malhotra R, Farshidfar F, Colaprico A, Parker JS, Mose LE, Vo NS, Liu J, Liu Y, Rader J, Dhankani V, Reynolds SM, Bowlby R, Califano A, Cherniack AD, Anastassiou D, Bedognetti D, Mokrab Y, Newman AM, Rao A, Chen K, Krasnitz A, Hu H, Malta TM, Noushmehr H, Pedamallu CS, Bullman S, Ojesina AI, Lamb A, Zhou W, Shen H, Choueiri TK, Weinstein JN, Guinney J, Saltz J, Holt RA, Rabkin CS; Cancer Genome Atlas Research Network; Lazar AJ, Serody JS, Demicco EG, Disis ML, Vincent BG, Shmulevich I. The Immune Landscape of Cancer. Immunity. 2018 Apr 17;48(4):812-830.e14. doi: 10.1016/j.immuni.2018.03.023. Epub 2018 Apr 5. Erratum in: Immunity. 2019 Aug 20;51(2):411-412. PMID: 29628290; PMCID: PMC5982584.

Van den Eynden,J. et al. (2015) SomInaClust: detection of cancer genes based on somatic mutation patterns of inactivation and clustering. BMC Bioinformatics, 16, 125.

Wickham,H. et al. (2022) readr: Read Rectangular Text Data.

Yu,G. et al. (2012) clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. OMICS J. Integr. Biol., 16, 284–287.

Zengin,T. and Önal-Süzek,T. (2020) Analysis of genomic and transcriptomic variations as prognostic signature for lung adenocarcinoma. BMC Bioinformatics, 21, 368.

Zengin,T. and Önal-Süzek,T. (2021) Comprehensive Profiling of Genomic and Transcriptomic Differences between Risk Groups of Lung Adenocarcinoma and Lung Squamous Cell Carcinoma. J. Pers. Med., 11, 154.