Supplementary Information for

CaPSSA: Visual evaluation of cancer biomarker genes for patient stratification and survival analysis using mutation and expression data

Yeongjun Jang[1,2], Jihae Seo[1], Sun Kim[2] and Sanghyuk Lee[1,*]

[1]Ewha Research Center for Systems Biology (ERCSB), Ewha Womans University, Seoul, Korea.

[2]Interdisciplinary Program in Bioinformatics, College of Natural Science, Seoul National University, Seoul, Korea.

Contents

List of Tables

List of Figures

1. Background

Patient stratification and predictive biomarkers are essential components for realizing the paradigm of precision medicine. Molecular characteristics, such as somatic mutations and expression signatures, are often used to identify putative biomarker genes for patient stratification. There is an immediate requirement for interactive tools to evaluate such candidate biomarkers using the vast public omics data.

Because of NGS technologies and publicly available omics datasets, such as The Cancer Genome Atlas (TCGA), it is now possible to validate putative biomarkers in large cancer cohorts based on patient stratification and survival analysis. Indeed, cBioPortal (Gao et al., 2013) can handle basic queries on multiple TCGA studies and allows users to submit sets of genes; however, it shows limitations when identifying patient sub-groups specific to candidate biomarkers with distinct molecular changes and performing survival analysis based on such patient stratification. Therefore, there is a need for a new, comprehensive, easy-to-use tool in order to determine whether DNA mutation, copy number, and RNA expression are informative in stratifying cancer patients and validating genes as biomarkers per their prognostic role in patient survival. SurvExpress (Aguirre-Gamboa et al., 2013) was developed for a similar purpose but only encompasses gene expression data; all outputs are static plot images, lacking in interactive user experience.

Here, we present an interactive open-access web-based application called CaPSSA (Cancer Patient Stratification and Survival Analysis) for evaluation of putative biomarkers; this application has a novel visualization scheme to allow an integrative in-depth analysis incorporating mutation, copy-number, and transcription data from TCGA cohorts. This can be used to evaluate the predictive role of putative biomarkers. CaPSSA dynamically and interactively distributes patients from a large cohort (e.g., TCGA) into two groups according to the molecular signature of query genes (putative biomarkers). CaPSSA allows users to examine the prognostic value of resulting patient groups based on survival analysis and associate this information with clinical features and previously annotated molecular subtypes. Finally, CaPSSA possesses a rich and interactive visualization. A schematic overview of CaPSSA is provided in Figure S1.

2. Methods

2.1. Cancer omics and clinical data

We retrieved somatic mutations, copy number variations, and gene expression data, as well as clinical metadata corresponding to various human cancers available in TCGA, via the Firehose Broad Genome Data Analysis Center (https://gdac.broadinstitute.org) on June 8th, 2016, and the Genomic Data Commons (GDC) Data Portal (Grossman et al., 2016; https://gdc-portal.nci.nih.gov) on Apr 20th, 2018. Although data collection will continue, to date we have collected approximately 10,701 patients out of 26 cancer cohorts (Table S1).

Somatic mutations were called and annotated by Oncotator (Ramos et al., 2015). RNA-Seq counts were estimated by RSEM (Li and Dewey 2011) and log2 transformed. CNVs harboring high-level copy number amplifications and homozygous deletions only were estimated by gene level using GISTIC 2.0 (Mermel et al., 2011). Additional patient data, absent in TCGA, included genetic and clinically relevant subtypes (i.e. Luminal A, Luminal B, Triple Negative/Basal-like, and HER2-type in breast cancer) and were obtained for all cancers from UCSC Xena (Goldman et al., 2018; http://xena.ucsc.edu) platform.

2.2. Patient stratification and survival analysis

We provide three approaches for performing on-the-fly interactive patient stratification with survival outcome difference; these approaches are based on somatic mutations, copy number variations, and gene expression data. In all methods, Kaplan-Meier was used to estimate survival curves, and the log-rank test was used to evaluate differences between different patient groups.

2.2.1. Patterns of genomic alterations

Cancer evolution is driven by interdependent alterations determining functions together, rather than individually, in the machinery of a cancer cell. Indeed, functionally redundant alterations rarely occur together and are instead likely to occur exclusively of each other (Kim et al., 2016; Babur et al., 2015; Ciriello et al., 2012). Whereas, synergistic mutations frequently co-occur, and affect disease progression and survival rate (Oricchio et al., 2014). Therefore, studying patient groups with patterns of interdependent oncogenic mutations will provide insights into functional interactions among these mutations and will propose putative mutational prognosis markers (Mina et al., 2017). For this purpose, we employed a visualization scheme similar to CBioPortal OncoPrint (Gao et al., 2013), but with improved ability to stratify patients by co-occurring genetic alterations, mutual exclusivity, or a

combination of these factors (see the section on "3.2. Case Studies").

2.2.2. Gene expression-based risk estimation

A recent study has shown that the majority of predictable differential dependencies among genes in cancer cell lines is best predicted by RNA expression levels rather than by DNA mutations or copy number (Tsherniak et al., 2017). Using the TCGA transcriptome and clinical data, CaPSSA enables users to examine how a set of genes of interest can separate patients into low- and high-risk groups, and whether these groups show a difference in survival outcomes.

By default, we split patients in the cohort of interest into two risk groups (high and low risk) of the same size, determined by ordered Prognostic Indexes (PI, higher value for higher risk). The Prognostic Index (PI) indicates a linear combination of expression levels in a set of genes. PI uses the Cox proportional hazards model for multivariate analysis, computed using the gene expression value multiplied by the regression coefficient of each gene estimated via Cox fitting (Collett 2015). It can be formularized as following: PI (Prognostic Index) = $\beta 1x1 + \beta 2x2 + ... + \beta pxp$, where xi is the expression value of ith gene, and βi (regression coefficeint) is obtained from Cox fitting (Aguirre-Gamboa et al., 2013). The fitting was performed using the JavaStat library (http://www2.thu.edu.tw/~wenwei).

We also employed an alternative approach, called nearest centroid classification (Guo et al., 2008; Figure S7), for expression-based patient classification. For this, we calculated the correlation coefficients of expression-level distributions for given potentially prognostic genes using a centroid (average expression for each gene) of patients with good prognosis (patients in upper 10th percentile calculated by days of survival or patients who have survived over 5 years). Then, we divided patients into two equal groups (median), where each group was determined by ordered Pearson correlation coefficients from positive correlation (lower risk) to negative correlation (higher risk). Simple average and mean of expression levels in a given set of genes can also be used as risk scores by which patients are divided.

2.2.3. Gene expression-based hierarchical clustering

We also incorporated gene-expression-based hierarchical clustering. This approach divides the patient population into subgroups sharing similar expression patterns. This method has been widely and successfully used for molecular stratification of human tumors. We employed Ward's linkage hierarchical clustering with Euclidean distance measure using fastclust Python library (http://danifold.net/fastcluster.html). This library allows users to run clustering on a cancer cohort of

interest using a given set of genes. Furthermore, interactive zoom-in-out and cluster selection enables the user to focus on particular sub-clusters, and refine the choice of a sub-group, with a more significant association with survival outcome.

3. Results

3.1. Web interface

Firstly, users have to upload a set of genes of interest to be used for the analysis via the page labeled 'Upload'. One can paste one or more gene symbols in the text area or can import gene sets from MSigDB (http://software.broadinstitute.org/gsea/msigdb/index.jsp). To analyze the query genes on a patient cohort, user can choose a gene set and a cancer type of interest from drop-down menus on top of the page. Once the genes and cancer cohort are selected, CAPSSA offers a three main section which performs the patient stratification and help users evaluate the impact on survival outcome. Sections can be accessed from the top menu, listed below:

3.1.1. Exclusivity and co-occurrence of simple nucleotide alterations (SNVs) and copy number variations (CNVs)

Distinct genomic alterations, including somatic mutations and copy number alterations, in patients of a given cancer cohort are visualized by a heatmap. Individual genes are represented as rows, and individual patients are represented as columns. Each bar above an alteration heatmap shows representative color-coded clinical features (i.e., expression subtype) used for sorting patients. Similar to CBioPortal (Gao et al., 2013) OncoPrint, this type of plot is useful for visually identifying alteration patterns in mutual exclusivity or co-occurrence between genes in a gene set across all patients.

The most distinctive feature of CaPSSA is that patients can be stratified into two groups by co-occurrence, mutual exclusivity, or a combination of these. For example, patients can be separated by *KRAS* mutations that possess *CDKN2A/B* inactivation by homozygous deletion versus patients with *KRAS* mutations having wildtype *CDKN2A/B* locus (see section "3.2. Case Studies"). The significance of prognostic genes in a given cancer type is highlighted with Kaplan-Meier plots, which are based on overall survival difference between two groups, as illustrated in Figure S2. In addition to survival analysis, CaPSSA provides a plot showing how concordant the stratification results are with known molecular or clinical sub-groups, such as PAM50 subtype in breast cancer or the expression subtype in lung and brain cancer (Figure S3).

3.1.2. Risk groups stratified using gene expression data

The outputs of gene-expression-based risk estimation consist of eight plots, as shown in Figure S4. 1) A Kaplan-Meier plot shows the prognosis of each patient group, together with log-rank test *p*-values.

2) A heatmap on the bottom depicts normalized and log2 transformed TPM (transcripts per million), estimated by RSEM (Li and Dewey 2011). 3) A scatter plot shows vitality status and survival day of each patient. 4) A waterfall plot of risk scores, calculated using prognostic index by Cox regression (Collett 2015) or nearest centroid classification (Guo et al., 2008; Figure S7). Bar colors in the waterfall plot represent color-coded values of the relevant clinical information. The user can manually choose an arbitrary PI cut-off point to separate patients in which survival difference is more significant, and group size can be adjusted by sliding the division line while seeing changes interactively. 5) A box plot shows differential expression levels of the genes in low and high risk groups. The user can correlate the expression direction of each gene with patient survival. 6) A box plot represents relative expression of genes in each clinical sub-group of low and high risk groups. 7) A Sankey plot shows the coherence of stratified patients with other clinical subtypes such as expression subtype, race, or smoking status, where applicable. 8) Cox regression outputs in tabular form. Negative regression coefficients indicate favorable prognostic genes for which higher expression of a given gene is correlated with longer patient survival outcome. Positive regression coefficients indicate unfavorable prognostic genes for which higher expression of a given gene is correlated with poor patient survival outcome.

3.1.3. Hierarchical clustering using gene expression data

The results of hierarchical clustering are shown as a tree structure called a dendrogram. The dendrogram shows the arrangement of individual clusters, a heatmap that is used for visualizing gene expression patterns in a grid panel (rows and columns represent genes and patients, respectively), and upper bars showing color-coded clinical features (Figure S5). Our cluster heatmap visualization and exploration is highly interactive. It facilitates patient stratification and comparison of survival outcomes between groups by enabling the user to select clustered patients or gene groups, and zoom in and out of clusters. Once a patient cluster is selected, various plots, similar to "Risk groups stratified using gene expression data", are available on the right panel. These include a Kaplan-Meier plot, two box plots of gene expression values across risk groups for each gene and clinical sub-group, and a Sankey plot showing visual coherence of available clinical information with clustered groups.

3.2. Case studies

CaPSSA enables users to conduct multiple types of analyses for the validation of putative biomarkers. Examples of analyses, provided below, illustrate the utility of CaPSSA.

3.2.1. *KRAS*-mutant lung adenocarcinoma dominated by co-occurring genetic events of

CDKN2A/B deletions coupled with low expression of the *NKX2-1* transcription factor

KRAS, an oncogene found on the short arm of chromosome 12 (12p), shows mutations in approximately 20% of non–small cell lung cancers, and gene amplification in a smaller set of cancers. Activating *KRAS* mutations are present in a variety of tumors. *KRAS* mutations activate the phosphorylation of mitogen-activated protein kinase (MAPK), inducing downstream signals of cellular proliferation and tumor progression via cell-cycle progression (Figure S12). *KRAS*-mutant lung adenocarcinomas have greater molecular diversity and therapeutic responsiveness, and possess subgroups with biologically and therapeutically relevant differences (Skoulidis et al., 2015). One of these subsets is *KRAS*-mutant lung adenocarcinoma, harboring either *CDKN2A* or *CDKN2B* deletions, coupled with low expression of the *NKX2-1* gene.

Two-hit inactivation of *CDKN2A* and *CDKN2B* occurs frequently in patients with mutant *KRAS* lung adenocarcinoma. The complete loss of *CDKN2A/B* accelerates mutant *KRAS*–driven lung tumorigenesis and progression, leading to loss of differentiation, increased metastatic disease, and decreased overall survival (Schuster et al., 2014). Cell proliferation depends on progression through four distinct phases of the cell cycle — G0/G1, S, G2, and M —regulated by several cyclin-dependent kinases (CDKs) that act in complex with their cyclin partners. Aberrant activity of CDKs, involved in cell cycle regulation and leading to uncontrolled proliferation, is commonly observed in human cancers. The inhibition of CDK activity in cancer not only leads to cell-cycle arrest, but also triggers senescence or apoptosis of tumor cells. As negative regulators of the cell cycle, CDK inhibitor proteins (CKIs) act as tumor suppressors. A deficiency in CKIs increases susceptibility to tumorigenesis (Otto and Sicinski, 2017; Figure S12). Examples of such tumor suppressor proteins, inactivated in cancer, are the INK4 proteins, particularly $p16^{INK4A}$ and $p15^{INK4B}$ (encoded by *CDKN2A* and *CDKN2B*, respectively); these are located within the chromosomal region 9p21, which is frequently deleted in human cancers (Zhao et al., 2016). *CDKN2A/B* inactivation is common in lung cancer and occurs via homozygous deletion; its expression is also commonly silenced by promoter methylation (Tam et al., 2013).

As shown in Figure S2, our results are in agreement with those of a separate study (Schuster et al., 2014); our results confirm that the overall median survival of patients, incorporating both somatic mutations in *KRAS* and homozygous deletions of *CDKN2A/B*, is significantly decreased compared with that of *KRAS*-mutants with wild type *CDKN2A/B* locus. These results suggest that mutant *KRAS* lung cancers with two-hit inactivation of *CDKN2A/B* identify a subset of patients with high-risk disease.

Thyroid Transcription Factor 1 (TTF1; also known as *NKX2-1*), a lineage-specific, homeobox-containing transcription factor, is almost uniformly suppressed among tumors in *KRAS*-mutant lung adenocarcinoma with loss of *CDKN2A/B*. TTF1 represents a clinical biomarker that facilitates the identification of these tumors (Skoulidis et al., 2015). In lung tumorigenesis, *NKX2-1* functions in a context-dependent manner and inhibits *KRAS*-driven lung adenocarcinoma (Maeda et al., 2012; Winslow et al., 2011). *NKX2-1* is a master regulator of lung differentiation and is expressed in a majority of human lung adenocarcinomas (Barletta et al., 2009). Human lung tumors, downregulated for *NKX2-1*, are poorly differentiated, show increased metastatic ability, and display high-grade histological features and aggressive behavior (Schuster et al., 2014). *NKX2-1* restrains lung-cancer progression by enforcing a lung epithelial differentiation program by repressing the chromatin regulator *HMGA2*, which controls the metastatic potential of lung adenocarcinoma (Winslow et al., 2011).

As shown in Figure S4, Cox regression-based risk estimation shows that lower expression of *NKX2-1* is associated with significantly reduced survival (log-rank test p = 3.83e-7 and regression coefficient p = 0.001). This means that *NKX2-1* is a favorable prognostic factor for survival in patients with lung adenocarcinoma. A study by Winslow et al., (2011) has shown the molecular and cellular basis for the association of *NKX2-1* expression with good patient outcome and that of *HMGA2* with poor patient outcome. We obtained concordant results with that study using our expression-based risk estimation.

Copy-number loss of *CDKN2A/B* is associated with higher stage, higher incidence of metastasis, and decreased overall survival in mutant KRAS lung cancer patients. Interestingly, a survival curve analysis obtained in another study indicated that patients with lung adenocarcinoma and high expression of *CDKN2A* show poor survival rates (Hsu et al., 2017), which agrees with our results (Figure S4.5). Although the exact biological function needs to be explored further, higher gene expression level of *CDKN2A/B* may potentially affect tumor proliferation and functions of tumor suppressors in the specific molecular environment of lung adenocarcinoma.

Our hierarchical clustering analysis indicates that overall survival time was significantly longer (log-rank p = 1.02e-7) for patients in cluster with higher TTF-1 expression (median 77.8 months) than for patients in cluster with lower TTF-1 expression (Figure S5). Patients with high and low levels of *CDKN2A/B* expression coexisted in the identified cluster; interestingly, these patients were also divided into sub-clusters with distinct low and high expression levels of *CDKN2A/B*.

Finally, we confirmed that high-risk patients designated using a mutation-based analysis, and those

designated using two expression-based analyses, significantly overlapped (p = 4.56e-5; Figure S6). This suggests that *KRAS* mutation in lung cancers with both *CDKN2A/B* deletions and low *NKX2-1* expression is a prognostic marker to identify a subset of patients at high-risk.

3.2.2. Co-occurrence and mutual exclusiveness of *TP53* alterations with amplifications of the *CCNE1* and *SKP2* oncogenes

SKP2 is an oncogenic protein that targets tumor suppressor proteins for degradation. Cyclin-dependent kinase (CDK) inhibitor p27, a positive regulator of cell-cycle progression, is a major target of SKP2, which has been shown in vivo and in vitro. Increased levels of SKP2 and reduced levels of p27 are observed in many types of cancer. In several cases, these levels are used as independent prognostic markers (Frescas and Pagano, 2008). Cyclin E, coded by the genes *CCNE1* and *CCNE2*, is the main regulator of transition from the G1 phase to the S phase of the cell cycle (Figure S13). *CCNE1* in particular has frequently been reported as a putative oncogene because it is amplified in various types of cancer. *SKP2* and *CCNE1* are oncogenic and overexpressed in human cancers (Gstaiger et al., 2001); their expression is an independent and significant prognostic factor for overall survival (Pils et al., 2014), as confirmed by our analyses of gene-expression data using Cox regression (Figure S8) and hierarchical clustering (Figure S9).

The activities of CCNE1 and SKP2 proteins are associated with the status of TP53 in the cell cycle, suggesting functional dependency (Loeb et al., 2005; Lin et al., 2010; Figure S13). *CCNE1* and *SKP2* show amplification-dependent overexpression in lung cancer (Ohshima et al., 2017). A recent study reported that increased expression of *CCNE1* and *SKP2* co-occur with high frequency in human samples (Mina et al., 2017); however, the exact biological function and difference in clinical outcomes remain undetermined.

Generally positive correlations with other genetic events may indicate functional synergies, whereas anti-correlations may indicate functional redundancies, because redundant events would not be required by the same cancer. Amplifications of *CCNE1* and *SKP2* are expected to be mutually exclusive. This is because either change provides a path to tumor development via uncontrolled cell-cycle progression (Figure S13), with no selective advantage to having both mutations. This is confirmed in our results, as shown in Figure S10. Similarly, co-occurrence of *TP53* alterations with amplifications of *CCNE1* and *SKP2* is expected to be synergistic with tumor progression. However, our results show no difference between each event with respect to the risk of death in patients with TCGA lung adenocarcinoma (Figure S10). Instead, patients with wild-type *TP53* and amplification of *CCNE1* and *SKP2* show better overall survival rates (Figure S11). Co-occurring alterations in the

regulators of cell-cycle and apoptosis are consistently associated with significantly increased mRNA expression of proliferation-associated genes (Gatza et al., 2014) in multiple types of tumors (Mina et al., 2017). This generates disease vulnerabilities, thereby increasing tumor sensitivity to specific drugs (Dey et al., 2017). *TP53*-deficient patients, concurrently exhibiting amplification of both *CCNE1* and *SKP2*, show higher cell-cycle progression and proliferation of cancer cells. These events may occur at the cost of exposing therapeutically-actionable disease vulnerabilities, resulting in moderately increased survival rates.

4. Comparison with existing tools

Existing tools having capability similar to that of CaPSSA can be divided into two categories according to their specific application and strengths: 1) Integrative visualization tools for cancer multi-omics datasets; these include cBioPortal (Gao et al., 2013), UCSC Xena (Goldman et al., 2018), and ICGC Portal (Zhang et al., 2011). 2) Tools for patient stratification and survival analysis that use a biomarker gene list as input; these include iGPSe (Ding et al., 2014), UALCAN (Chandrashekar et al., 2017), and SurvExpress (Aguirre-Gamboa et al., 2013).

Integrative visualization tools for cancer multi-omics datasets provide basic and exploratory analyses of cancer genomics data. These tools feature an intuitive web interface, biologically relevant abstraction of genetic alterations at the gene level, integrative analysis of genomic data sets and clinical attributes, interactive network analysis, and a list of co-expressed and mutually expressed genes. These tools are useful for exploring gene-level associations across different cancers involving mutation frequency or gene expression. However, there remains a need for tools that would allow the user to examine survival associations across different cancer subsets as defined by prognostic or predictive biomarker genes with distinct molecular changes. Such tools should also allow the user to examine survival associations across different cancer subsets as defined by clinical pathologic features (e.g., pathological stages, tumor grade, and patient drinking and/or smoking history).

Compared with these omics-level large cohort visualization tools, CaPSSA extends and complements available functionality with its ability to stratify patients based on molecular patterns of putative biomarker genes. This allows for performance comparisons and validations of prognostic and predictive biomarkers using survival analysis and risk assessment across tumors. More specifically, it allows the users to identify and assess genomic signatures in cancer subtypes, assess the role of these signatures in stratifying patients into different risk groups, and compare and contrast with previously established clinical or molecular subtypes.

Several tools have been proposed to facilitate the evaluation of putative biomarkers and stratify patients based on risk assessment (i.e., survival analysis) (Aguirre-Gamboa et al., 2013; Ding et al., 2014; Chandrashekar et al., 2017). However, these tools encompass only gene-expression data, and all outputs are static images lacking in interactive user experience. CaPSSA incorporates data on somatic mutations and copy number variations, as well as gene-expression data. Therefore, CaPSSA provides various ways to interactively define and select sub-populations and clusters. This allows the user to focus on details of interest and to compare patient outcomes across sub-populations dynamically, based on molecular characteristics of the user-supplied gene list.

The comparison of CaPSSA with existing tools is summarized in Table S2 and Table S3, which focus on visualization of cancer multi-omics data and patient stratification, respectively.

Table S1. TCGA cancer cohorts

Cancer Type	Cases	Data Source
Bladder Urothelial Carcinoma (BLCA)	412	Broad Genome Data Analysis Center (http://gdac.broadinstitute.org)
Breast invasive carcinoma (BRCA)	1,097	
Cervical Cancer (CESC)	308	
Colon Adenocarcinoma (COAD)	462	Genomic Data Commons (GDC) Data Portal (https://portal.gdc.cancer.gov)
Esophageal Carcinoma (ESCA)	185	
Glioblastoma multiforme (GBM)	595	
Head and Neck Squamous Cell Carcinoma (HNSC)	528	UCSC Xena TCGA data hub (https://tcga.xenahubs.net)
Kidney Renal Clear Cell Carcinoma (KIRC)	537	
Kidney Renal Papillary Cell Carcinoma (KIRP)	291	
Acute Myeloid Leukemia (LAML)	200	
Brain Lower Grade Glioma (LGG)	516	
Liver Hepatocellular Carcinoma (LIHC)	377	
Lung adenocarcinoma (LUAD)	522	
Lung Squamous Cell Carcinoma (LUSC)	506	
Ovarian Serous Cystadenocarcinoma (OV)	595	
Pancreatic Adenocarcinoma (PAAD)	185	
Pheochromocytoma and Paraganglioma (PCPG)	179	
Prostate Adenocarcinoma (PRAD)	498	
Rectum Adenocarcinoma (READ)	169	
Sarcoma (SARC)	261	
Skin Cutaneous Melanoma (SKCM)	471	
Stomach Adenocarcinoma (STAD)	478	
Testicular Germ Cell Tumors (TGCT)	150	
Thyroid Carcinoma (THCA)	507	
Thymoma (THYM)	124	
Uterine Corpus Endometrial Carcinoma (UCEC)	548	
Total	**10,701**	

Table S2. Comparison of integrative visualization tools for multi-omics datasets

Tool	Gene set management *	Patient stratification based on molecular signature	Survival analysis	Comparison of risk groups with clinical subtype
CaPSSA	O	Various **	O	O
cBioPortal	X	Altered vs. not altered patients	O	X
UCSC Xena	X	According to previously-defined subtypes	O	O
ICGC Portal	X	X	X	X

* Management of a list of gene sets allowing users to add, modify, and delete a gene set.

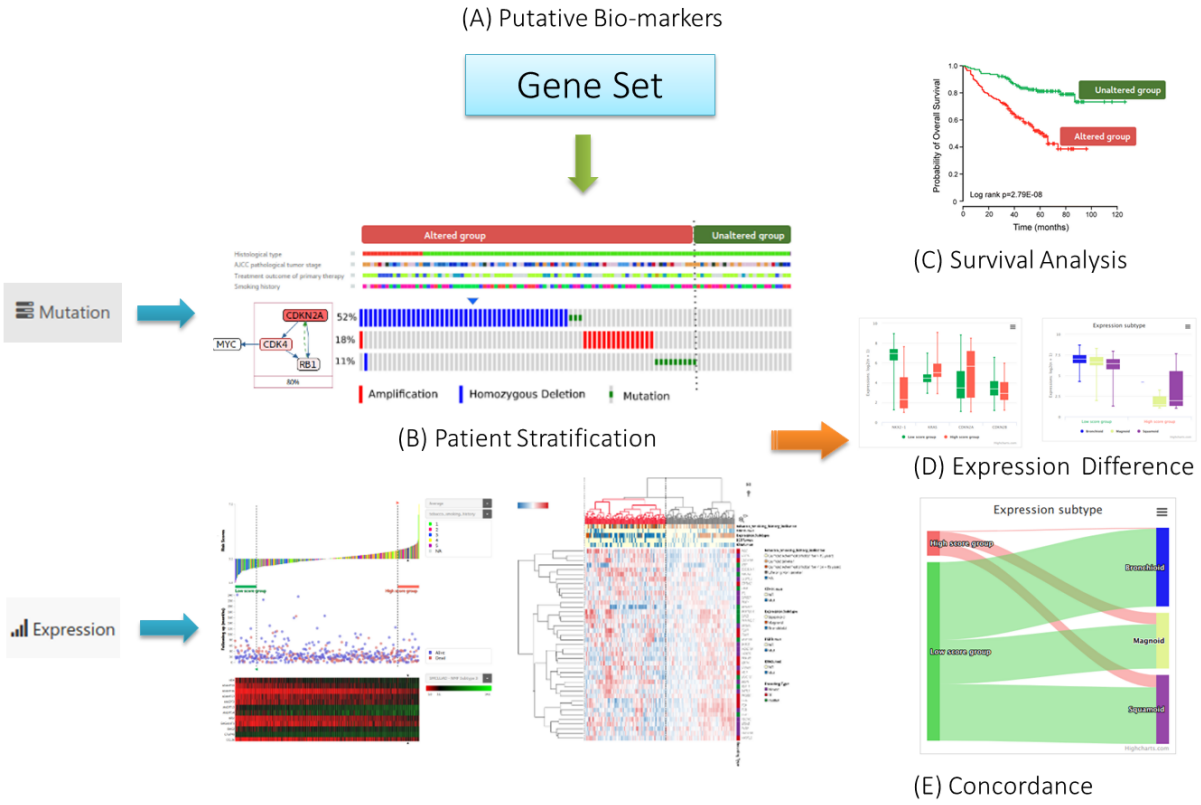
** See stratification methods in Table S3.

Table S3. Comparison of patient stratification and survival analysis tools

Tool	Stratification methods	Interactive sub-grouping	Concordance to other subtypes	Data types
CaPSSA	Mutual exclusivity of mutations Co-occurrence of mutations CoxPH regression Nearest centroid classification Hierarchical clustering	O	O	Somatic mutations Copy number alterations Gene expression
SurvExpress	Cox fitting User-specified weights	X	O *	Gene expression
iGPSe	K-means clustering Spectral clustering Community detection	X	X	Gene expression
UALCAN	High vs. low expression	X	X	Gene expression

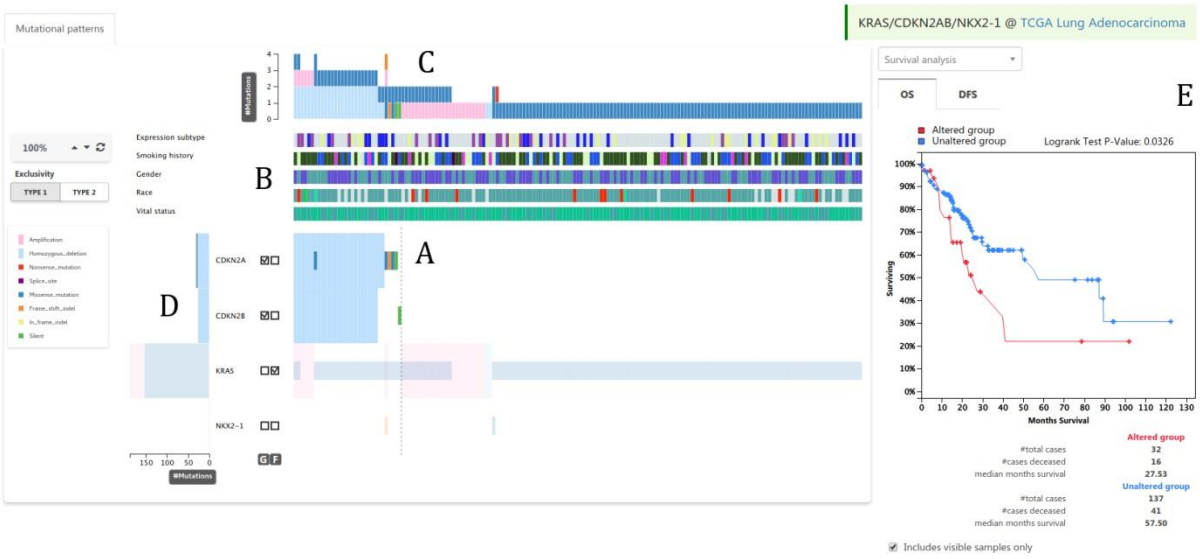
* Static graphics lacking interactive user experiences.

Figure S1. Schematic overview of CaPSSA



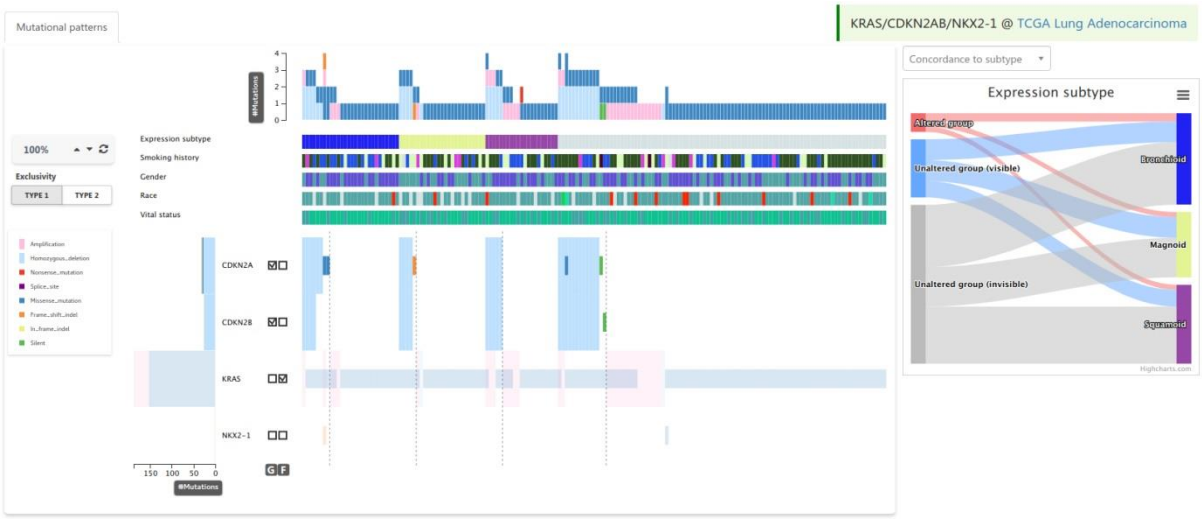
(A) Users upload a set of genes as putative biomarkers to be used in further analyses. **(B)** CaPSSA provides three approaches to stratify a patient. The approaches are based on molecular signatures of the given putative gene set and use somatic mutations, copy number variations, and gene expression data. Patient stratification schemes include mutual exclusivity and co-occurrence of mutations (upper), risk estimation based on gene-expression profiles (lower left), and hierarchical clustering based on gene-expression data (lower right). **(C)** Evaluation of survival differences between patient groups. **(D)** Differential gene expression in groups of patients, defined by molecular signature or clinical subtype. **(E)** Concordance with previously established tumor subtypes.

Figure S2. Patient stratification and survival analysis based on mutational patterns



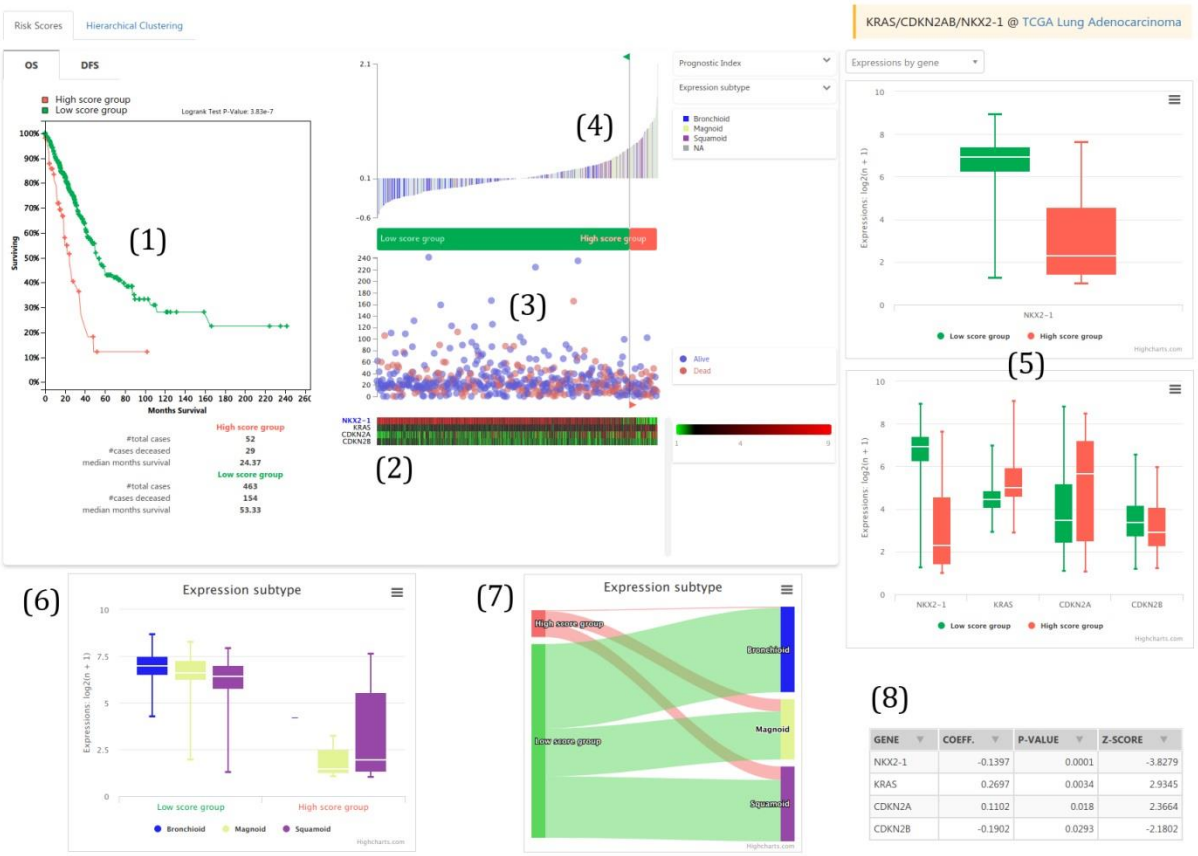
(A) A heatmap in the center depicts distinct color-coded genetic alterations, including somatic mutations or copy number alterations across patients. Individual genes and patients are represented as rows and columns, respectively. The order of genes can be changed by dragging a gene name into the desired position. The user can interactively define patient groups, based on combinations of alteration patterns between genes, via checkbox UIs beside the gene symbol labels. Check-boxes in the first column, tagged by 'G' (Grouping), are for choosing whether to designate patients with alterations in a given gene into an alteration group. Checkboxes tagged by 'F' (Filtering) are for choosing whether to display patients having alterations in a given gene only, or all the patients. Each alteration group is visually divided by a vertical dashed line. **(B)** Each bar above the alteration heatmap shows a representative clinical feature (i.e., Expression Subtype). The user can sort patients by their clinical values by clicking the feature's name. **(C)** and **(D)** Horizontally and vertically stacked bars on the left and top of the heatmap indicate the number of alterations and their color-coded types across patients and genes. **(E)** A Kaplan-Meier plot highlighting the significance of prognostic genes in a given cancer type (in this case, TCGA Lung Adenocarcinoma) based on overall survival. Log-rank *p*-values are shown in the upper right corner of the Kaplan-Meier plot. The red and blue colors of the lines indicate altered and unaltered patient groups in selected genes, respectively.

Figure S3. Coherence of differentially altered groups to other subtypes



Patients are sub-grouped by selected clinical feature (in this case, by clicking “Expression Subtype”), then ordered again based on mutational patterns in each sub-group. In the right panel, a Sankey plot shows how concordant the stratification results are with known molecular or clinical subtypes. In each stacked box on each side, a subgroup is encoded by a box whose height is proportional to the number of patients within that subgroup. Colors of the boxes indicate corresponding subgroups. On the left side, the red/light-blue/grey bars represent patients with alterations, patients without alterations initially loaded, and patients without alterations not initially loaded, respectively. On the right side, the blue/yellow/purple bars indicate three groups of expression subtypes (Bronchoid, Magnoid, and Squamoid) in TCGA Lung Adenocarcinoma, which is consistent with the color of the sub-group shown in the bars for clinical features on the main panel. The bands connecting boxes represent matched patients in different groups. The width of the bands is proportional to the number of patients. The main goal of this plot is to show how consistent are the results of patient stratification with other molecular or clinical subtypes.

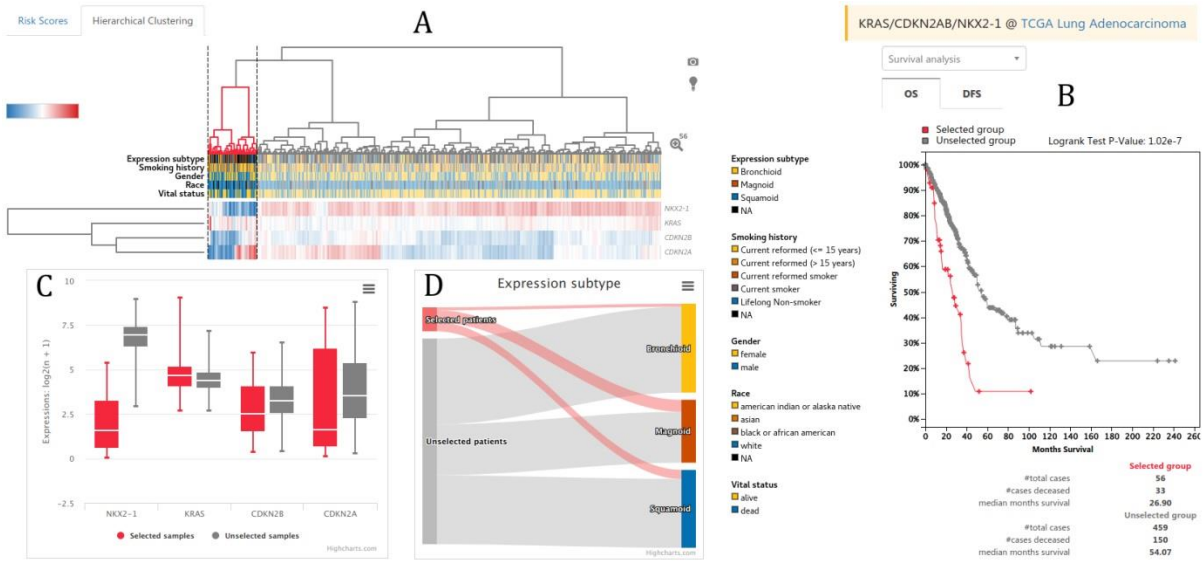
Figure S4. Gene expression-based risk estimation using Cox regression



Red and green colors in the following plots represent patient groups with high and low risk scores, respectively. Sample names and associated values can be visualized by placing the cursor over any part of the plot. **(1)** Prognosis of each risk group of patients was examined by Kaplan-Meier survival estimators, and the survival outcomes of different groups were compared by log-rank tests; **(2)** Lower heatmap depicts normalized and log2 transformed TPM (transcripts per million) of putative bio-marker genes, estimated by RSEM (Li and Dewey 2011) from low to high expression (green to red) across patients with TCGA Lung Adenocarcinoma. **(3)** A scatter plot shows the vitality status and survival days of patients. **(4)** A waterfall plot represents ordered risk scores calculated using prognostic index by Cox regression. Bar colors in the waterfall plot represent clinical values of a chosen feature, in this case "Expression Subtype". To choose a group with significant risk, patients with risk score in upper 10th percentiles are classified into the high-risk group and the remaining patients into the low-risk group. This is accomplished by dragging arrows on both ends of the vertical division line. **(5)** Expression levels of a chosen gene (upper plot; *NKX2-1*) and all genes (lower plot) across all patients are each shown as a box plot for low- and high-risk groups. **(6)** A box plot shows the relative expression of a chosen gene (*NKX2-1*) in Bronchoid, Magnoid, and Squamoid LUAD

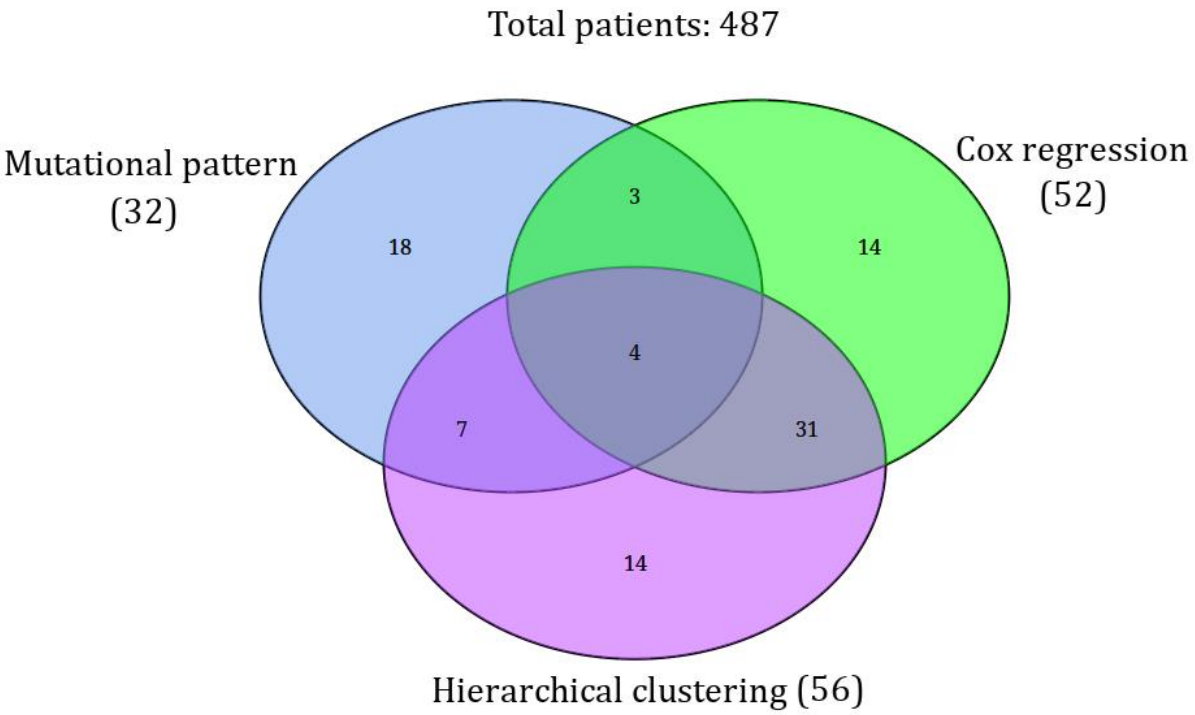
patients of low- and high-risk groups. **(7)** A Sankey plot shows the coherence of risk groups with other subtypes (see Figure S3 for details). **(8)** Cox regression outputs in tabular form (see "3. Results").

Figure S5. Hierarchical clustering using gene expression data



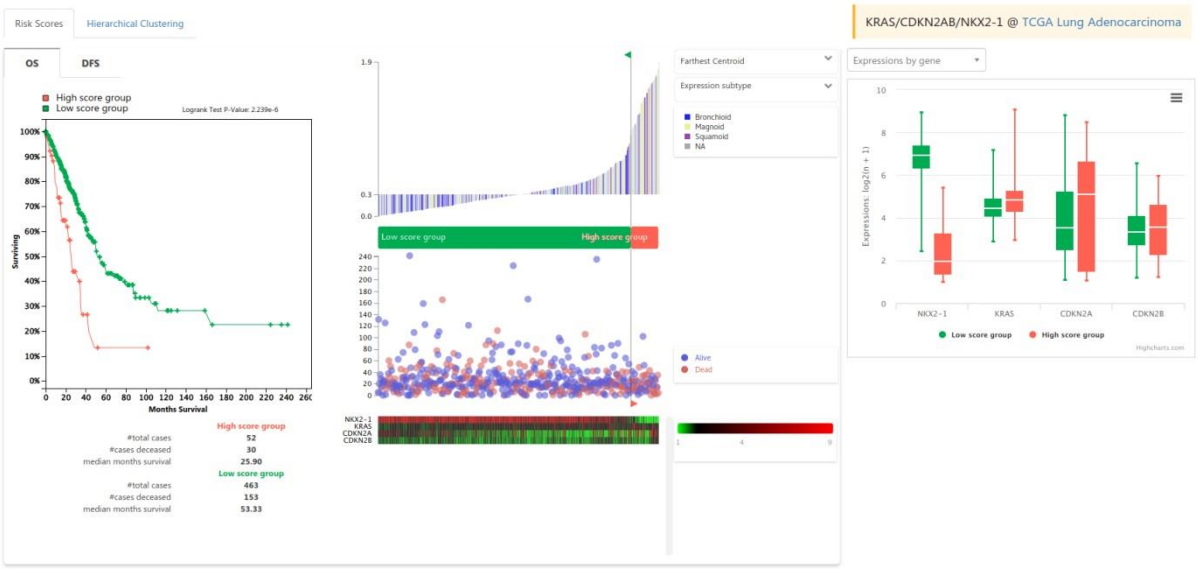
(A) Hierarchical clustering analysis of the expression patterns of *NKX2-1*, *KRAS*, and *CDKN2A/B* genes in a TCGA Lung Adenocarcinoma cohort. Interactive zoom-in-out of a cluster enables choosing particular sub-clusters. Clicking the magnifier icon in the upper right corner of the heatmap panel allows the user to choose a sub-group with a more significant association with survival outcome. **(B)** When a patient (column) cluster (dendrogram) is selected, a KM-plot displays survival difference between patients (red color) in the cluster and other patients (grey color) in the cohort. **(C)** The expression level of *NKX2-1* (left most) is substantially decreased across patients in the selected cluster (red). **(D)** A Sankey plot shows visual coherence of available clinical information with clustered groups. Patients in the selected cluster (light red) tend to match with Magnoid (red) and Squamoid (blue), rather than with Bronchoid (yellow), expression subtypes in TCGA Lung Adenocarcinoma.

Figure S6. Overlaps between high-risk groups



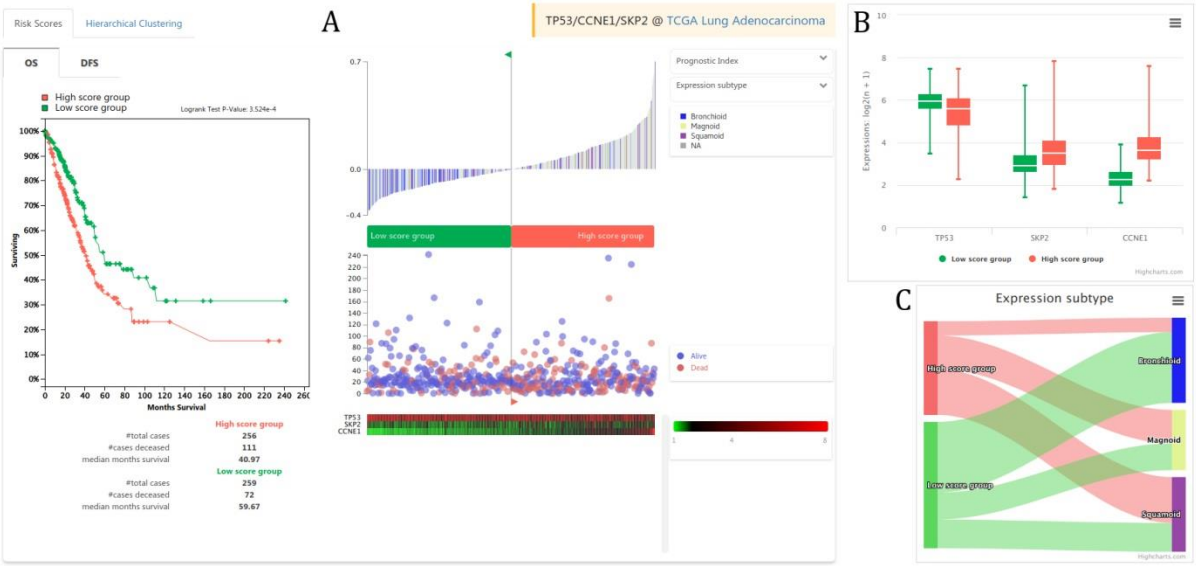
A Venn diagram showing overlaps between high-risk patient groups obtained using mutation-based analysis and those obtained using two expression-based analyses. Each circle is labeled by the method used for the identification of each risk group. Overlaps are statistically significant, with p = 4.56e-5 calculated using hypergeometric distribution (http://nemates.org/MA/progs/overlap_stats.html).

Figure S7. Nearest centroid classification



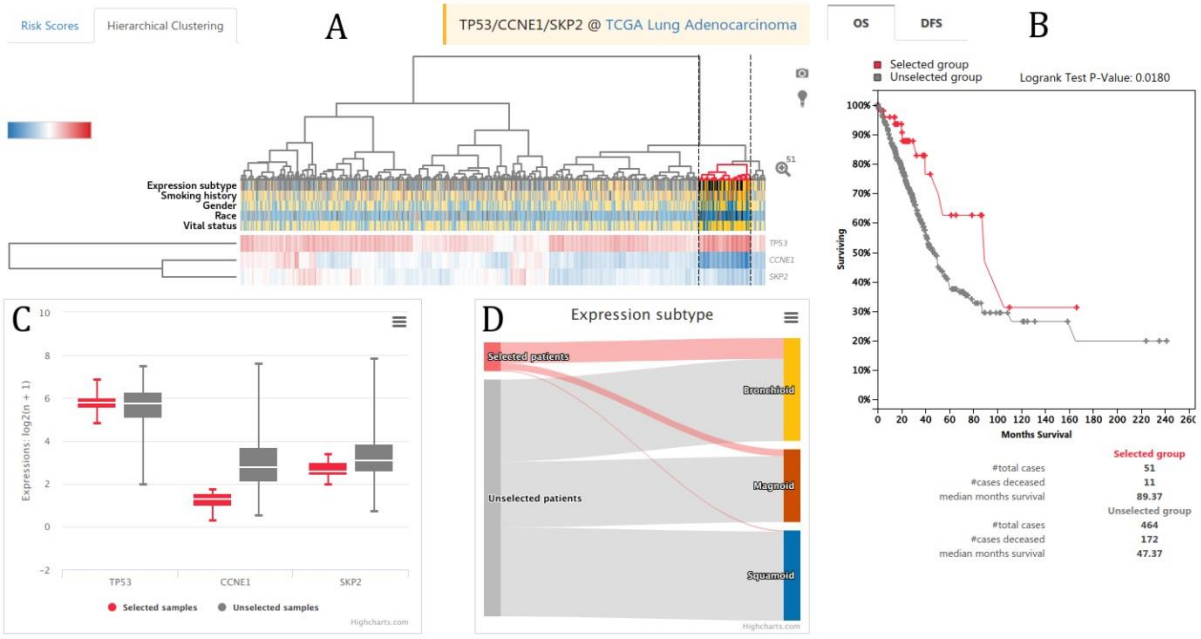
Another risk-score based stratification approach ("Farthest Centroid" menu in the upper drop-down menu, on the right of the waterfall plot) uses correlation coefficients of gene expression in each patient and centroids of patients with good prognosis. Positive correlation implies lower risk, whereas negative correlation implies higher risk. Although patients are divided into equal groups (median) by default, we adjusted the division for more significant results by manually moving the vertical division line.

Figure S8. Expression of *SKP2* and *CCNE1* oncogenes in risk groups



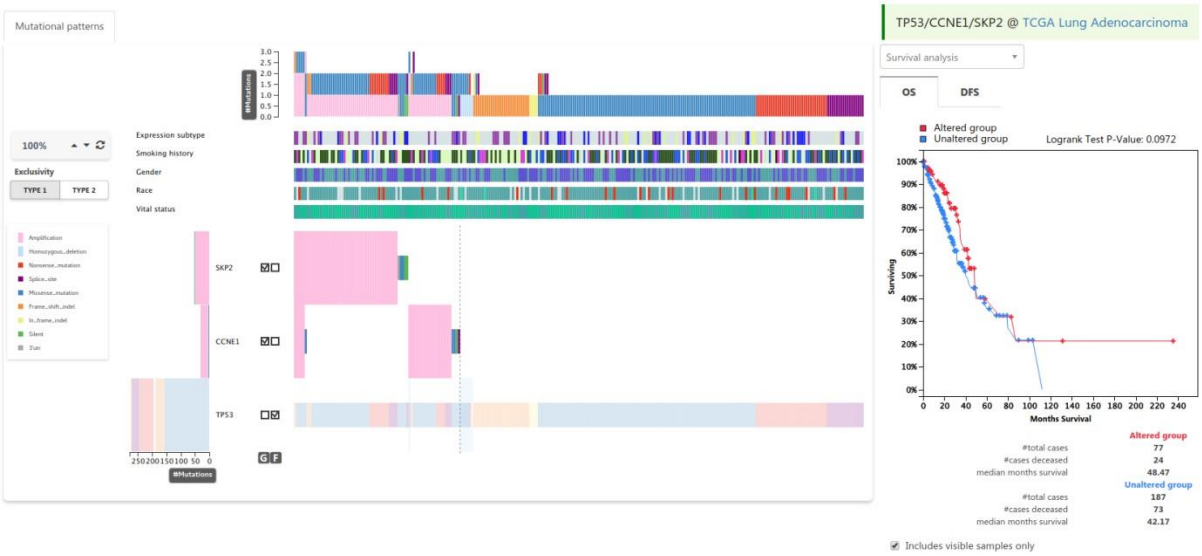
Risk estimation of patients using expression levels of *SKP2* and *CCNE1* genes. **(A)** High (red color) and low (green color) risk groups are stratified using Cox regression. K-M plot shows significant survival difference between groups, with log-rank p = 3.524e-4. **(B)** Patients at higher risk scores (red color) show higher expression levels of *SKP2* and *CCNE*, whereas patients at lower risk (green color) show lower expression levels. **(C)** Higher-risk group (red) tends to match with Magnoid (light yellow) and Squamoid (purple), rather than Bronchoid (blue), expression subtypes in TCGA Lung Adenocarcinoma; the opposite applies in the low-risk group.

Figure S9. Hierarchical clustering analysis of expression levels of *CCNE* and *SKP2* genes



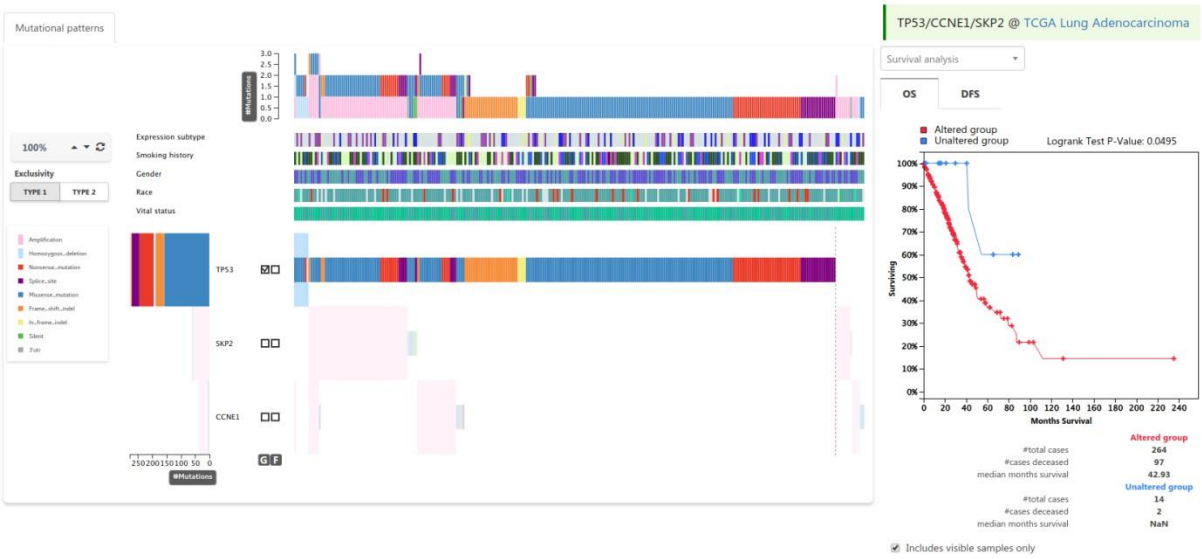
(A) Selection of a cluster with lower expression of *CCNE1* and *SKP2*. **(B)** K-M plot depicting significantly longer survival outcome for patients (red color) in the selected cluster (log-rank p = 0.018). **(C)** A box plot showing lower expression of the two oncogenes *CCNE1* and *SKP2* across patients in the selected cluster (light red). **(D)** Patients in the selected cluster (red) tend to match with Bronchoid (yellow), rather than with Magnoid (red) or Squamoid (blue), expression subtypes in TCGA Lung Adenocarcinoma.

Figure S10. Co-occurrence of *TP53* alterations with amplifications of *CCNE1* and *SKP2*



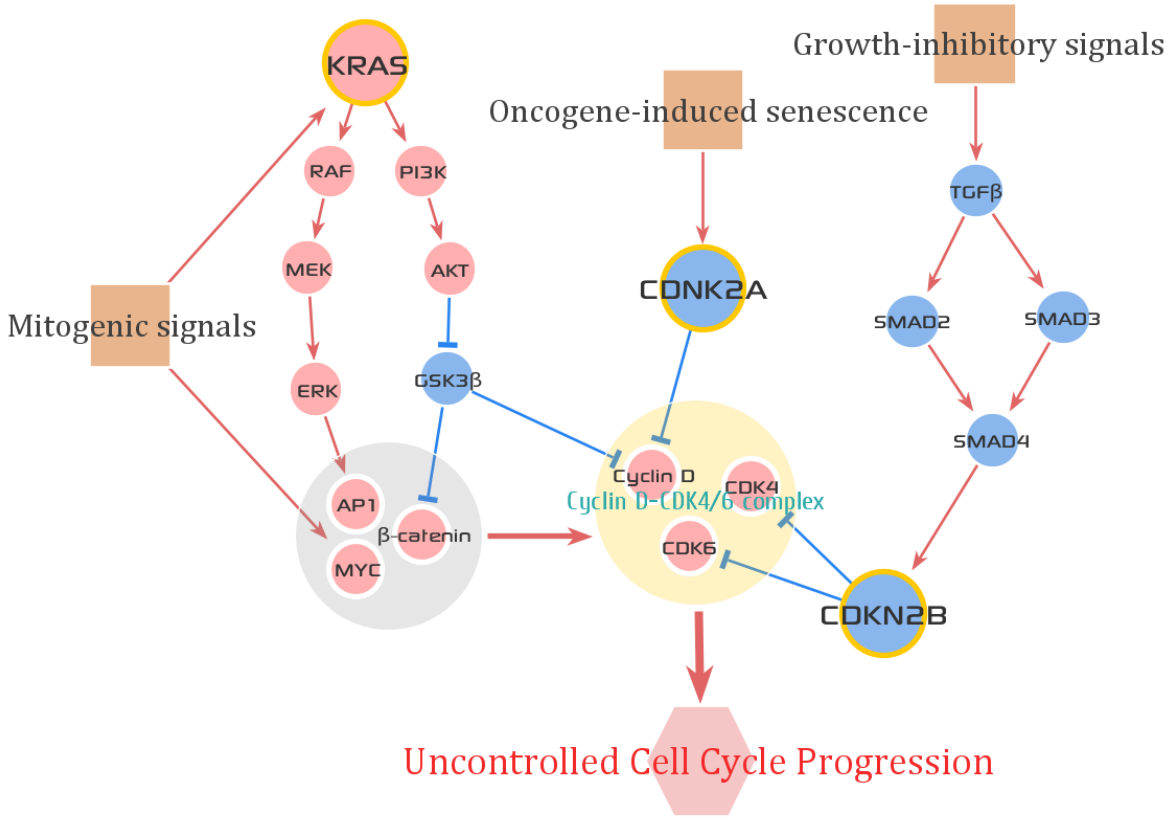
The co-occurrence of *TP53* alterations with amplifications of *CCNE1* and *SKP2* is expected to be synergistic with tumor progression in lung adenocarcinoma. However, our results show no difference in survival rates (log-rank p = 0.0972) between patients having both *TP53* mutations and *CCNE1*/*SKP2* amplification (left of dashed division line) versus *TP53*-mutants only (right of dashed division line). Additionally, the occurrences of amplification events in *SKP2* and *CCNE1* genes are almost mutually exclusive across tumors.

Figure S11. Amplification of *CCNE1* and *SKP2* without *TP53* mutations



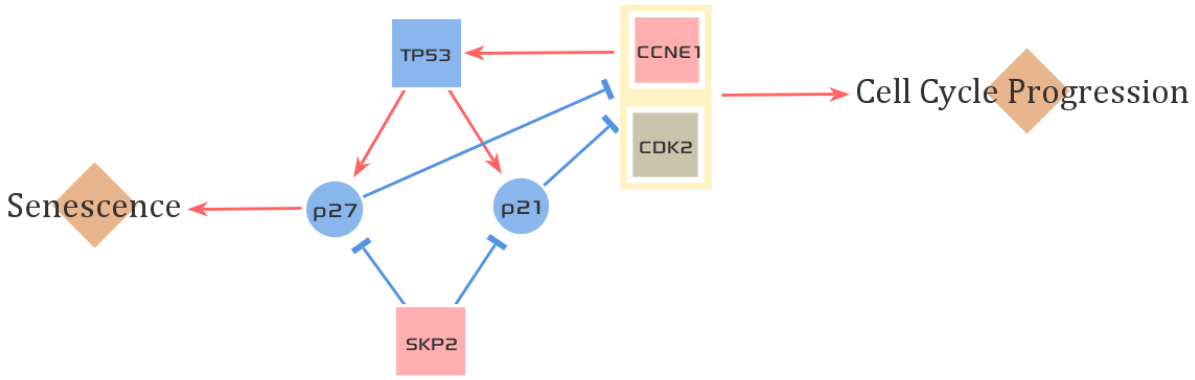
Patients with wild-type *TP53* and amplified *CCNE1* and *SKP2* (on right of the guide line) show better overall survival rates than patients with *TP53* mutations (log-rank $p = 0.0495$).

Figure S12. The biology of *KRAS* mutation-induced signaling that triggers cell-cycle progression, leading to uncontrolled cellular proliferation



Entry into the cell cycle is typically induced in response to mitogenic signals that activate signaling pathways such as RAS, MAP kinase, and PI3K. These pathways eventually impinge on transcription factors such as MYC, activator protein 1 (AP1), or β-catenin, and lead to induction of several cell cycle proteins, including D-type cyclins. Formation of active complexes of D-type cyclins with cyclin-dependent kinase 4 (CDK4) or CDK6 is antagonized by the INK4 family (p16INK4A and p15INK4B, encoded by *CDKN2A* and *CDKN2B*, respectively) in response to senescence-inducing or growth-inhibitory signals, such as transforming growth factor-β (TGFβ) (Otto and Sicinski, 2017). Oncogenes and tumor suppressor genes are represented by pink and cyan nodes, respectively. Red and blue edges depict activating and inhibitory relationships, respectively. Genes with somatic mutations or copy number loss are indicated using a thick orange border (*KRAS*, *CDKN2A*, and *CDKN2B*). This network was generated and visualized using MONGKIE (Jang et al., 2016).

Figure S13. Functional interactions between *TP53*, *CCNE1*, and *SKP2* genes in the senescence and cell -cycle pathways



Oncogenes and tumor suppressors are represented by pink and cyan nodes, respectively. Red and blue edges depict activating and inhibitory relations, respectively. This signaling network was generated and visualized using MONGKIE (Jang et al., 2016).

Bibliography

Aguirre-Gamboa, R. et al. (2013) SurvExpress: an online biomarker validation tool and database for cancer gene expression data using survival analysis. PLoS One, 8, e74250.

Babur, Ö. et al. (2015) Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations. Genome Biol., 16, 45.

Barletta, J.A. et al. (2009) Clinical significance of TTF-1 protein expression and TTF-1 gene amplification in lung adenocarcinoma. J. Cell. Mol. Med., 13, 1977–86.

Chandrashekar, D.S. et al. (2017) UALCAN: A Portal for Facilitating Tumor Subgroup Gene Expression and Survival Analyses. Neoplasia, 19, 649–658.

Ciriello, G. et al. (2012) Mutual exclusivity analysis identifies oncogenic network modules. Genome Res., 22, 398–406.

Collett, D. (2015) Modelling Survival Data in Medical Research, Third Edition. CRC press.

Dey, P. et al. (2017) Genomic deletion of malic enzyme 2 confers collateral lethality in pancreatic cancer. Nature, 542, 119–123.

Ding, H. et al. (2014) iGPSe: a visual analytic system for integrative genomic based cancer patient stratification. BMC Bioinformatics, 15, 203.

Frescas, D. and Pagano, M. (2008) Deregulated proteolysis by the F-box proteins SKP2 and β-TrCP: tipping the scales of cancer. Nat. Rev. Cancer, 8, 438–449.

Gao, J. et al. (2013) Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci. Signal., 6.

Gatza, M.L. et al. (2014) An integrated genomics approach identifies drivers of proliferation in luminal-subtype human breast cancer. Nat. Genet., 46, 1051–1059.

Goldman, M. et al. (2018) The UCSC Xena Platform for cancer genomics data visualization and interpretation. bioRxiv, 326470.

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

Gstaiger, M. et al. (2001) Skp2 is oncogenic and overexpressed in human cancers. Proc. Natl. Acad. Sci. U. S. A., 98, 5043–8.

Guo, N.L. et al. (2008) Confirmation of gene expression-based prediction of survival in non-small cell lung cancer. Clin Cancer Res, 14, 8213–8220.

Hsu, Y.-L. et al. (2017) Identification of novel gene expression signature in lung adenocarcinoma by using next-generation sequencing data and bioinformatics analysis. Oncotarget, 8, 104831–104854.

Jang, Y. et al. (2016) MONGKIE: An integrated tool for network analysis and visualization for multi-omics data. Biol. Direct, 11.

Kim, J.W. et al. (2016) Characterizing genomic alterations in cancer by complementary functional associations. Nat. Biotechnol., 34, 539–546.

Li, B. and Dewey, C.N. (2011) RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics, 12.

Lin, H.K. et al. (2010) Skp2 targeting suppresses tumorigenesis by Arf-p53-independent cellular senescence. Nature, 464, 374–379.

Loeb, K.R. et al. (2005) A mouse model for cyclin E-dependent genetic instability and tumorigenesis. Cancer Cell, 8, 35–47.

Maeda, Y. et al. (2012) Kras(G12D) and Nkx2-1 haploinsufficiency induce mucinous adenocarcinoma of the lung. J. Clin. Invest., 122, 4388–400.

Mermel, C.H. et al. (2011) GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol., 12, R41.

Mina, M. et al. (2017) Conditional Selection of Genomic Alterations Dictates Cancer Evolution and Oncogenic Dependencies. Cancer Cell, 32, 155–168.e6.

Ohshima, K. et al. (2017) Integrated analysis of gene expression and copy number identified potential cancer driver genes with amplification-dependent overexpression in 1,454 solid tumors. Sci. Rep., 7, 641.

Oricchio, E. et al. (2014) Frequent disruption of the RB pathway in indolent follicular lymphoma suggests a new combination therapy. J. Exp. Med., 211, 1379–91.

Otto, T. and Sicinski, P. (2017) Cell cycle proteins as promising targets in cancer therapy. Nat. Rev. Cancer, 17, 93–115.

Pils, D. et al. (2014) Cyclin E1 (CCNE1) as independent positive prognostic factor in advanced stage serous ovarian cancer patients – A study of the OVCAD consortium. Eur. J. Cancer, 50, 99–110.

Ramos, A.H. et al. (2015) Oncotator: Cancer variant annotation tool. Hum. Mutat., 36, E2423–E2429.

Schuster, K. et al. (2014) Nullifying the CDKN2AB locus promotes mutant K-ras lung tumorigenesis. Mol. Cancer Res., 12, 912–23.

Skoulidis, F. et al. (2015) Co-occurring genomic alterations define major subsets of KRAS-mutant lung adenocarcinoma with distinct biology, immune profiles, and therapeutic vulnerabilities. Cancer Discov., 5, 860–77.

Tam, K.W. et al. (2013) CDKN2A/p16 inactivation mechanisms and their relationship to smoke exposure and molecular features in non-small-cell lung cancer. J. Thorac. Oncol., 8, 1378–88.

Tsherniak, A. et al. (2017) Defining a Cancer Dependency Map. Cell, 170, 564–576.e16.

Winslow, M.M. et al. (2011) Suppression of lung adenocarcinoma progression by Nkx2-1. Nature, 473, 101–4.

Zhang, J. et al. (2011) International Cancer Genome Consortium Data Portal--a one-stop shop for cancer genomics data. Database (Oxford)., 2011, bar026.

Zhao, R. et al. (2016) Implications of Genetic and Epigenetic Alterations of CDKN2A (p16(INK4a)) in Cancer. EBioMedicine, 8, 30–39.