Evergene documentation

VERSION 1.0 (14TH MAY, 2024)

Introduction 2
Quick start 2
Tutorial 3
Example workflow A (with genes of interest) 3
Example workflow B (without genes of interest) 6
User input 9
Data origin 10
Source code 12

Introduction

Evergene is a webtool for interactive principal component analysis (PCA), survival analysis and correlation analysis using data from The Cancer Genome Atlas (TCGA) project or uploaded user inputs. The analysis focuses on multi-gene input/output and integration of sample annotation in data visualisation.

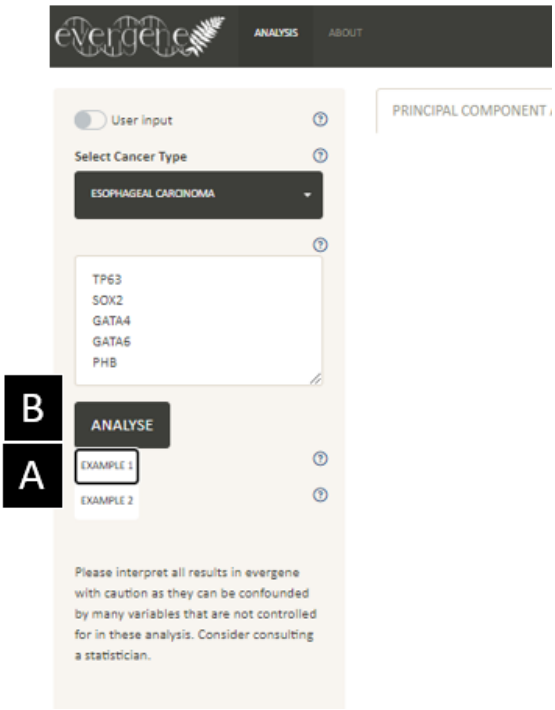
PCA is performed on all genes. In PCA, the expression of input genes are illustrated and compared to sample annotation in the various graphs.

Quick start

Evergene can be accessed at https://bshihlab.shinyapps.io/evergene/.

A. You can click on the "Example 1" and "Example 2" buttons on the sidebar to load an example input using TCGA data.
B. The analysis can be performed by using the button “Analyse”.
C. Detailed helpfiles can be found by clicking on the question marks next to each section. Links to helpfiles for the results outputs are listed below.

User input
PCA - top section
PCA - middle section
PCA - download buttons
Survival - plots
Survival - download buttons
Correlation



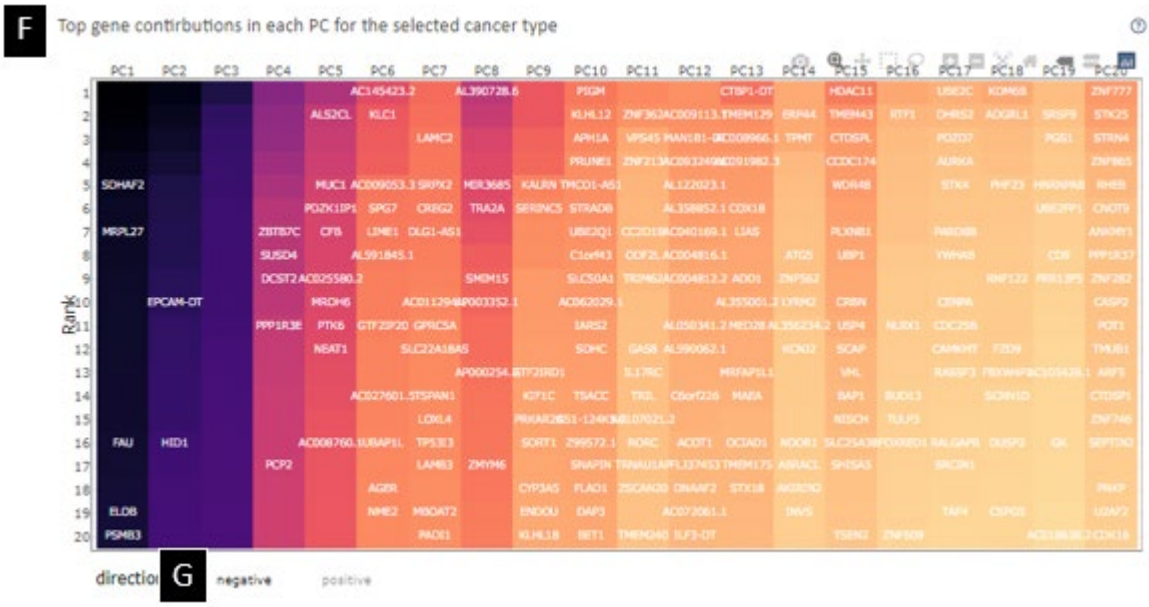
Tutorial

Example workflow A (with genes of interest)

You can use Evergene to explore the potential association between your gene of interest and cancer subtypes. Take KRT6A in Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma as an example.

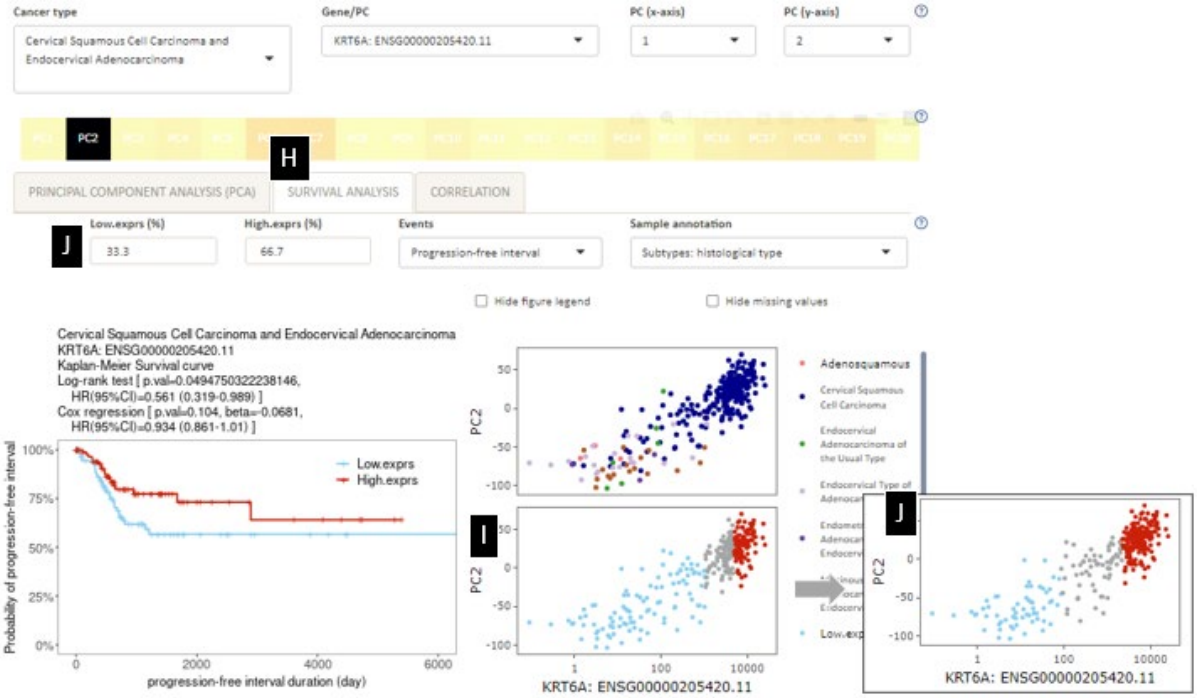
A. Enter KRT6A as input **Genes** and Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma as input **Cancer type**.
B. Click **Analyse**.
C. Find a PC the gene contributes strongly to. Note that earlier PCs (i.e. PC1, PC2… etc.) explain larger variations within the dataset. In this case, KRT6A contributes strongly to PC2.
D. Select PCs you are interested in. Here we have selected PC1 and PC2.

E. Go through the **Sample annotation** to see if the sample separation corresponds to a known sample annotation. Here the separation of PC2 corresponds to Subtype: histological type, separating squamous cell carcinoma from adenocarcinoma.
F. You can go to **Top gene contribution in each PC** for the selected cancer type to see what other genes contribute strongly to PC2.



G. By clicking on **negative** or **positive**, you can hide/show genes that contribute to each PC in the positive or negative direction. In this case, HID1 is a potential negative marker for the same grouping.

H. By using the **Survival analysis** tab, you can compare survival between patients with high or low expression of the input gene KRT6A.



I. You can see the distribution of the data that are defined as KRT6A-low/high groups.

J. You may want to change the thresholds for defining low/high groups to reflect the data spread. Here we changed the **Low.exprs (%)** and **High.exprs (%)** from the default 33.3% and 66.7% to 20% and 40%.

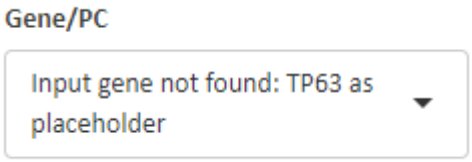
K. You can use the **Correlation** tab to quickly explore the correlation between a variable of interest and the other variables.

L. In this example, we have put KRT6A as **X-axis**, and Selected all for **Y-axis**. Note we re-ran the analysis including both KRT6A and HID1 (negative contributor to PC2 identified in Step G) in the input gene list. Variables strongly correlated to the KRT6A are highlighted by an asterisk.

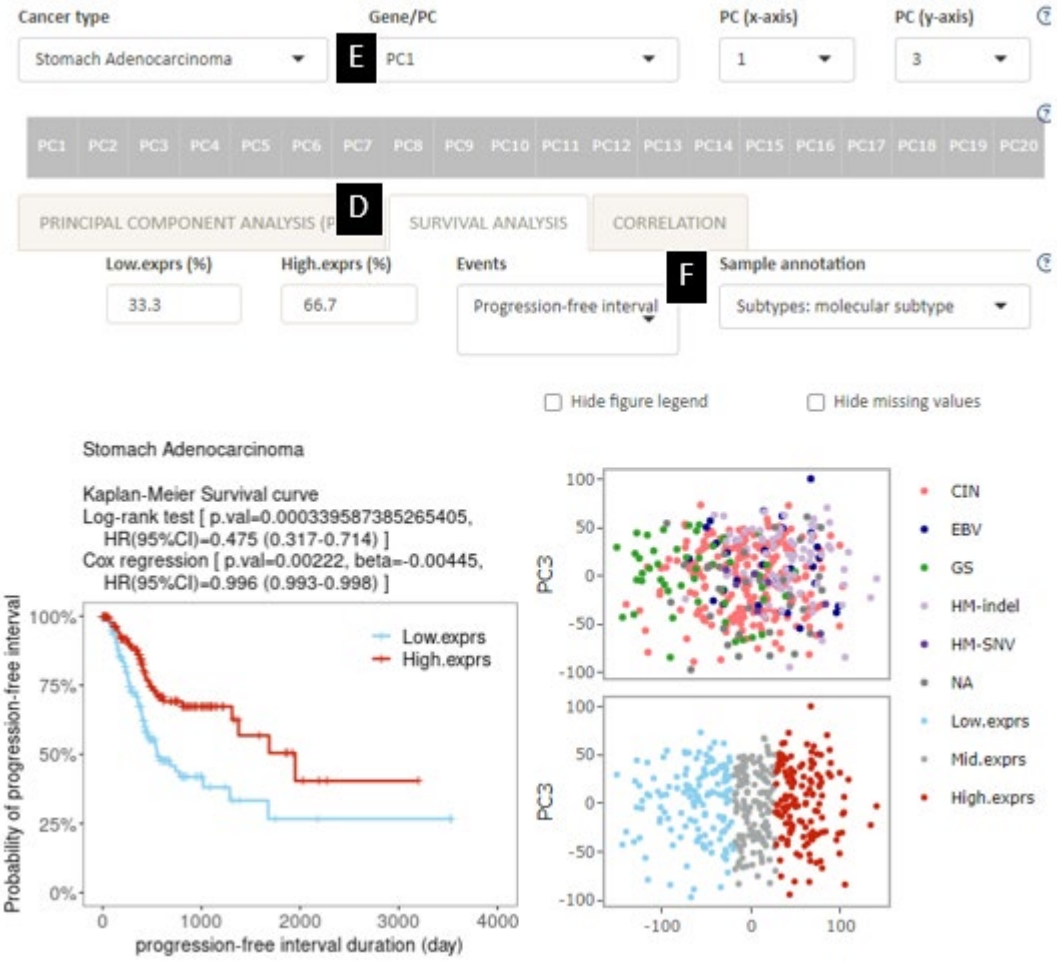
Example workflow B (without genes of interest)

You can use Evergene to explore genes that can potentially be used to define cancer subgroups, followed by evaluating their respective clinical outcomes. Take Stomach Adenocarcinoma as an example.

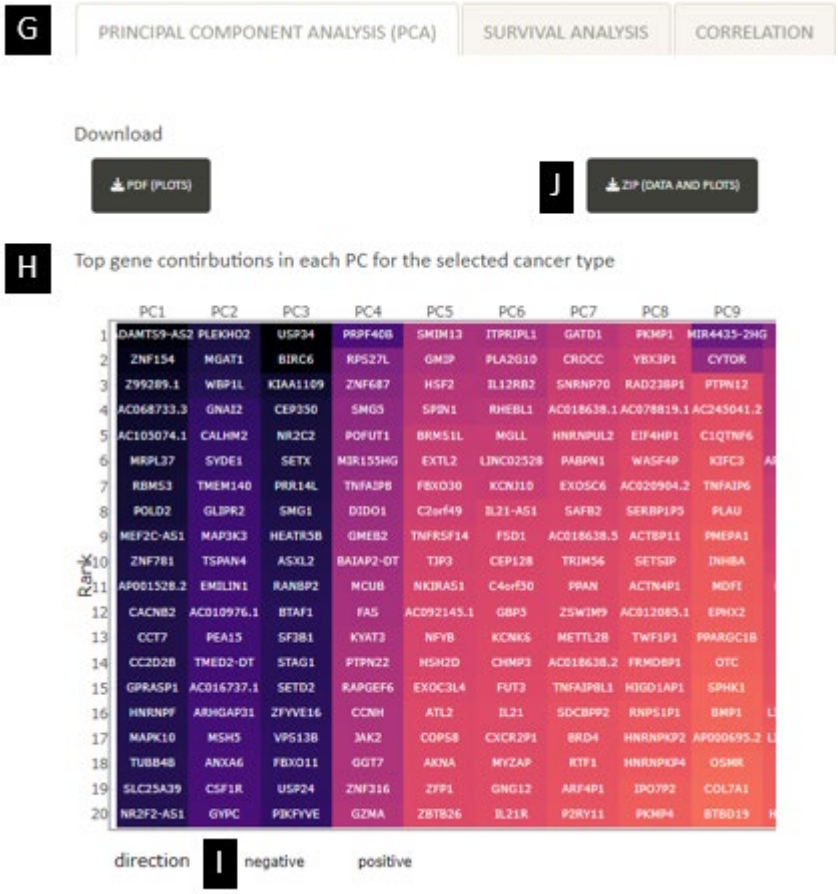
A. Leave input **Genes** empty and select Stomach Adenocarcinoma as input **Cancer type**.



B. Because there was no input gene, TP63 was selected as a place holder to allow you to access the other PCA graphs.
C. You can ignore this graph for now because this graph is made from the placeholder due to no input gene (see Step B).
D. You can select Survival analysis tab to use PC to define high-/low expressing groups.
E. You can use PCs to explore if any of PCs result in patient grouping with significantly different clinical outcomes.
F. You can change the sample annotation to see if the selected PC1 (indicated by E, plotted in x-axis) is separated by any of the sample annotations. In this case, PC1 corresponds to Subtypes: molecular subtype, with the left-hand side corresponding to the GS subtype (labelled green; genomically stable).



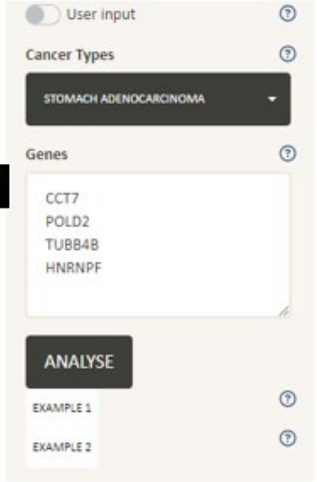
G. Going back to the **Principal Component analysis (PCA)** tab.



H. You can identify genes of interest from PC1 using **Top gene contributions in each PC for the selected cancer type** at the bottom of the PCA analysis tab.

I. You can click on **negative** or **positive** to show/hide the negative and positive contributing genes. For example, CCT7 is a positive contributor and ADAMTS9-AS2 is a negative contributor to PC1.

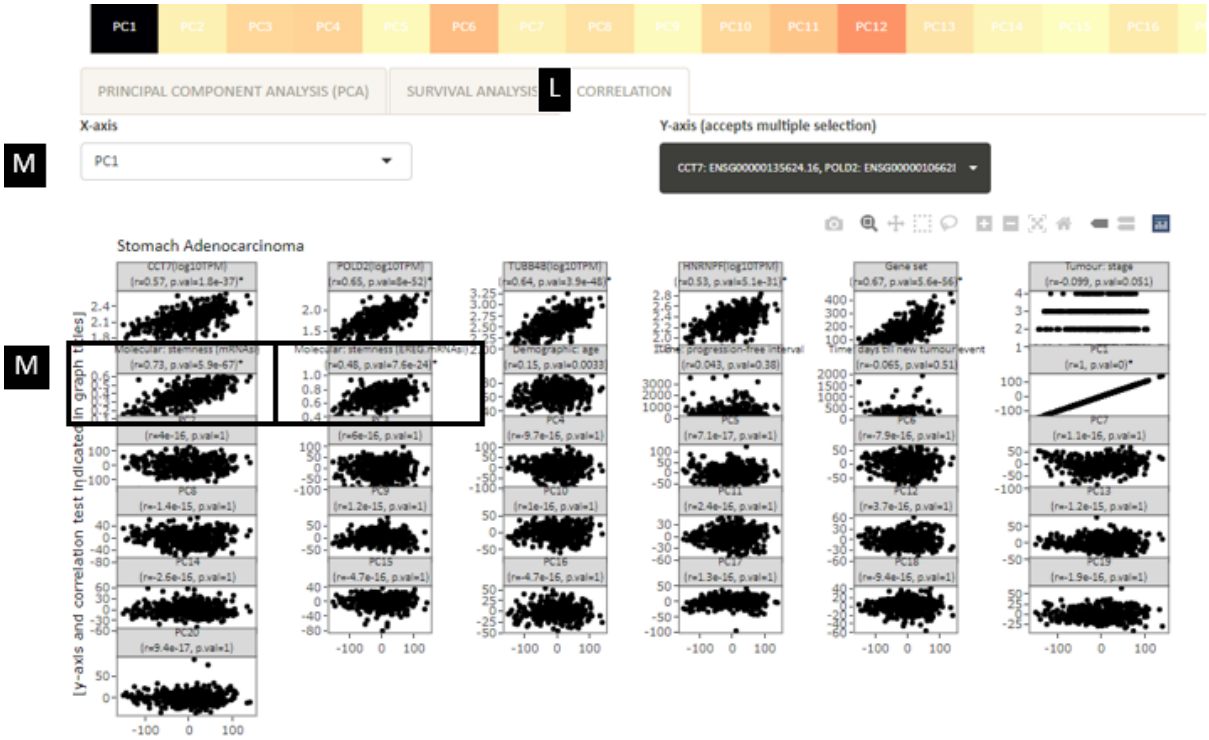
J. You can download all data using the **Zip (data and plots)** download button. This gives you the top 100 contributing genes for each PC, instead of just the top 20 in the graph. You can potentially submit these top 100 contributing genes to a gene set enrichment tool, such as Metascape https://metascape.org/, to determine the potentially corresponding gene pathways.



K

K. We would recommend you restart the analysis with these genes of interests to determine how well each can be used to explain differences in survival or sample annotations. You may want to enter positive and negative contributing genes separately if you want to explore the use of gene set (log average expression) of all input genes. Click on **Analyse** again to redo the analysis with the new gene list.

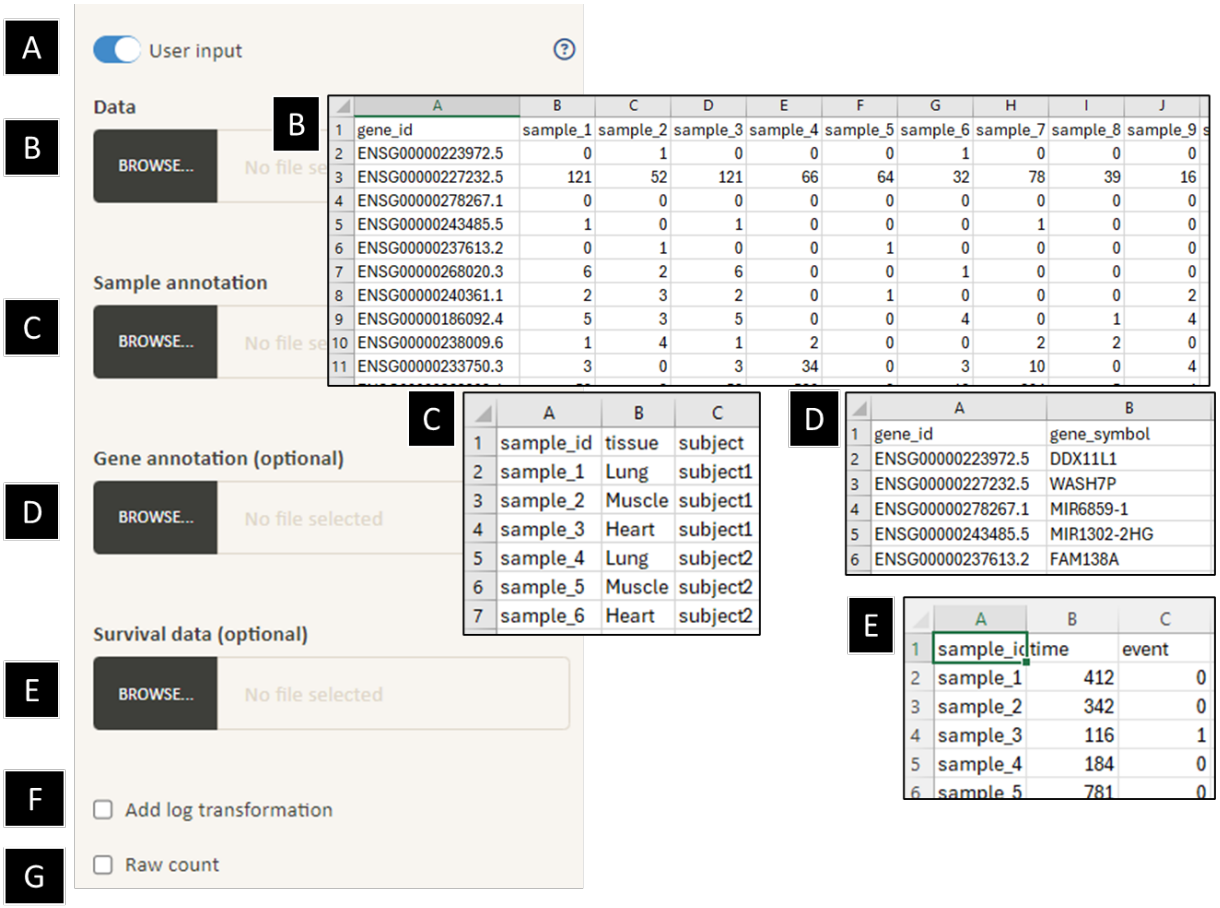
L. You can select the **Correlation** tab to perform correlation analysis.



M. By selecting PC1 for **X-axis** and Select all for **Y-axis**, you can explore other variables that are strongly correlated to PC1. In this example, Molecular: stemness (mRNAsi) and Molecular: stemness (EREG.mRNAsi) both strongly correlate with PC1, as well as the positive contributing genes we have selected from Step H.

User input

Users can input their own data and perform PCA or integrated analysis between PCA and SA. The inclusion of gene annotation and clinical events for SA is optional. You can download example user input data for Evergene from https://github.com/bshihlab/evergene/tree/main/example_user_input. To analyse your own data on Evergene, please ensure that it follows the same format as these example datasets.



	A	B	C	D	E	F	G	H	I	J
1	gene_id	sample_1	sample_2	sample_3	sample_4	sample_5	sample_6	sample_7	sample_8	sample_9
2	ENSG00000223972.5	0	1	0	0	0	1	0	0	0
3	ENSG00000227232.5	121	52	121	66	64	32	78	39	16
4	ENSG00000278267.1	0	0	0	0	0	0	0	0	0
5	ENSG00000243485.5	1	0	1	0	0	0	1	0	0
6	ENSG00000237613.2	0	1	0	0	1	0	0	0	0
7	ENSG00000268020.3	6	2	6	0	0	1	0	0	0
8	ENSG00000240361.1	2	3	2	0	1	0	0	0	2
9	ENSG00000186092.4	5	3	5	0	0	4	0	1	4
10	ENSG00000238009.6	1	4	1	2	0	0	2	2	0
11	ENSG00000233750.3	3	0	3	34	0	3	10	0	4

C

	A	B	C
1	sample_id	tissue	subject
2	sample_1	Lung	subject1
3	sample_2	Muscle	subject1
4	sample_3	Heart	subject1
5	sample_4	Lung	subject2
6	sample_5	Muscle	subject2
7	sample_6	Heart	subject2

D

	A	B
1	gene_id	gene_symbol
2	ENSG00000223972.5	DDX11L1
3	ENSG00000227232.5	WASH7P
4	ENSG00000278267.1	MIR6859-1
5	ENSG00000243485.5	MIR1302-2HG
6	ENSG00000237613.2	FAM138A

E

	A	B	C
1	sample_id	time	event
2	sample_1	412	0
3	sample_2	342	0
4	sample_3	116	1
5	sample_4	184	0
6	sample_5	781	0

A. By clicking on the **User input** toggle button, you can switch on and off the input menu.
B. You need to include the data you want to perform PCA on. Typically this would be gene expression data, with the first column being gene IDs, and the first row indicating sample IDs. The values in the first column and the first row must be unique (i.e. no duplicated values). Please avoid using spaces or special characters in them; dashes will be replaced by underscore.
C. You also need to include a table with sample annotation. The first column must be the samples included in B, and each column after this is a sample annotation.
D. As Evergene has been designed to work with gene expression data, you can optionally include a table to annotate the genes (i.e. the alternative names for the genes IDs). Please label them as **gene_id** and **gene_symbol**. By including this conversion table, you can plot genes in Evergene by referring to either **gene_id** or **gene_symbol**.
E. You can include clinical information for performing survival analysis. Please put sample IDs that matches the sample IDs in B and C in the first column, followed by time, and the clinical events (indicated by 0 and 1).

Data origin

The TCGA projects included in Evergene are listed below.

Project	Study Name	Number of samples
TCGA-BLCA	Bladder Urothelial Carcinoma	412
TCGA-BRCA	Breast invasive carcinoma	1111
TCGA-CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	304
TCGA-COAD	Colon adenocarcinoma	481
TCGA-ESCA	Esophageal carcinoma	184
TCGA-GBM	Glioblastoma multiforme	157
TCGA-HNSC	Head and Neck squamous cell carcinoma	520
TCGA-KIRC	Kidney renal clear cell carcinoma	541
TCGA-KIRP	Kidney renal papillary cell carcinoma	290
TCGA-LGG	Brain Lower Grade Glioma	516
TCGA-LIHC	Liver hepatocellular carcinoma	371
TCGA-LUAD	Lung adenocarcinoma	539
TCGA-LUSC	Lung squamous cell carcinoma	502
TCGA-MESO	Mesothelioma	87
TCGA-OV	Ovarian serous cystadenocarcinoma	421
TCGA-PAAD	Pancreatic adenocarcinoma	178
TCGA-PRAD	Prostate adenocarcinoma	501
TCGA-READ	Rectum adenocarcinoma	166
TCGA-SARC	Sarcoma	259
TCGA-SKCM	Skin Cutaneous Melanoma	103
TCGA-STAD	Stomach adenocarcinoma	412
TCGA-TGCT	Testicular Germ Cell Tumors	150
TCGA-THCA	Thyroid carcinoma	505
TCGA-THYM	Thymoma	120
TCGA-UCEC	Uterine Corpus Endometrial Carcinoma	553
TCGA-UVM	Uveal Melanoma	80

The sample and project selection criteria are as follows:

- Primary tumours
- 80 or more samples within a project
- Have at least 1 clinical survival outcome that has been recommended by TCGA-CDR

PCA processing is detailed below:

PCA was performed independently for each cancer project with all samples and all genes detected in at least 20% of the samples. When input data is count data, TMM normalisation was performed followed by log-normalised count per million (CPM) values. For log-scaling user input data that is not count data, if the lowest value in the input data is 0 and the input data is not count data, 1% of the minimum value is added to all values prior to log-transformation. PCA was performed with scaling. Gene contribution to each PC is described by PCA loadings.

The origins for sample annotations are as follows:

Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, Kovatich AJ, Benz CC, Levine DA, Lee AV *et al*: **An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics**. *Cell* 2018, **173**(2):400-416 e411.

- Clinical survival outcomes
- Clinical and demographic annotation (histology, stage, age, race, gender, histological type, treatment outcome).

Malta TM, Sokolov A, Gentles AJ, Burzykowski T, Poisson L, Weinstein JN, Kamińska B, Huelsken J, Omberg L, Gevaert O *et al*: **Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation**. *Cell* 2018, **173**(2):338-354.e315.

- mRNAsi (based on mRNA expression)
- EREG-mRNAsi (epigenetically regulated mRNAsi; based on both mRNA expression and DNA methylation)

Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang TH, Porta-Pardo E, Gao GF, Plaisier CL, Eddy JA *et al*: **The Immune Landscape of Cancer**. *Immunity* 2018, **48**(4):812-830 e814.

- Immune subtype

Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, Shen R, Taylor AM, Cherniack AD, Thorsson V *et al*: **Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer**. *Cell* 2018, **173**(2):291-304.e296.

- Molecular subtype

Source code

The source code for data pre-processing and Evergene shinyapp can be found on https://github.com/bshihlab/evergene.