Database, 2018, 1–10
doi: 10.1093/database/bay069
Original article

Original article

KM-express: an integrated online patient survival and gene expression analysis tool for the identification and functional characterization of prognostic markers in breast and prostate cancers

Xin Chen[1,2], Zhengqiang Miao[2], Mayur Divate[2], Zuxianglan Zhao[2] and Edwin Cheung[2,*]

[1]Guangdong Key Laboratory of IoT Information Technology, School of Automation, Guangdong University of Technology, No. 100 Waihuan Xi Road, Guangzhou Higher Education Mega Center, Panyu District, Guangzhou 510006, PR China and [2]Faculty of Health Sciences (E12), University of Macau, Avenida da Universidade, Room 4045, Taipa, Macau, China

*__Corresponding author:__ Tel.: +853 8822 4992; Fax: +853 8822 2314; Email: echeung@umac.mo

Citation details: Chen,X., Miao,Z., Divate,M. *et al.* KM-express: an integrated online patient survival and gene expression analysis tool for the identification and functional characterization of prognostic markers in breast and prostate cancers. *Database* (2018) Vol. 2018: article ID bay069; doi:10.1093/database/bay069

Received 4 December 2017; Revised 29 May 2018; Accepted 13 June 2018

Abstract

The identification and functional characterization of novel biomarkers in cancer requires survival analysis and gene expression analysis of both patient samples and cell line models. To help facilitate this process, we have developed KM-Express. KM-Express holds an extensive manually curated transcriptomic profile of 45 different datasets for prostate and breast cancer with phenotype and pathoclinical information, spanning from clinical samples to cell lines. KM-Express also contains The Cancer Genome Atlas datasets for 30 other cancer types with matching cell line expression data for 23 of them. We present KM-Express as a hypothesis generation tool for researchers to identify potential new prognostic RNA biomarkers as well as targets for further downstream functional cell-based studies. Specifically, KM-Express allows users to compare the expression level of genes in different groups of patients based on molecular, genetic, clinical and pathological status. Moreover, KM-Express aids the design of biological experiments based on the expression profile of the genes in different cell lines. Thus, KM-Express provides a one-stop analysis from bench work to clinical prospects. We have used this tool to successfully evaluate the prognostic potential of previously published biomarkers for prostate cancer and breast cancer. We believe KM-Express will accelerate the translation of biomedical research from bench to bed.

Database URL: http://ec2-52-201-246-161.compute-1.amazonaws.com/kmexpress/index.php

Introduction

The most commonly diagnosed cancers in men and women are prostate cancer and breast cancer, respectively (1). For these cancers, hormone-deprivation therapies are used with or without surgery as first-line treatments (2, 3). Unfortunately, these cancers often demonstrate either *de novo* resistance to hormonal therapies or subsequently acquire resistance following an initial therapeutic response (3). Thus, it is important to identify prognostic markers for disease progression and resistance to treatments, and to predict the patients' outcome. Comprehensive functional analysis is also needed to accelerate the clinical use of these markers as therapeutic targets.

To aid researchers in identifying new prognostic markers, a number of gene expression and survival analysis web-tools for patient data analysis have been developed. For example, web-tools such as Oncomine (4, 5), GENT (6), BioXpress (7) and MERAV (8) can be used for the mining of patient gene expression data. In general, these web-tools allow cross-dataset expression analysis between normal and disease states, during the progression of the disease, or in response to treatment, and thus are useful for confirming prognostic markers. For patient survival analysis, a number of options are also available and they include but not limited to PrognoScan (9), G-DOC (10), GOBO (11), SurvExpress (12), BreastMark (13), SurvMicro (14) and Kaplan–Meier Plotter (15–19). Currently, most of these web-tools house only microarray datasets but with the recent availability of next-generation sequencing (NGS) data from projects such as The Cancer Genome Atlas (TCGA), web-tools such as The cBio Cancer Genomics Portal (http://cbioportal.org) (20, 21), SurvExpress (12) and PROGgeneV2 (22) are beginning to include these datasets as well.

Besides discovering new prognostic markers for cancers, it is also important to characterize their biological functions, since they could also be potential therapeutic targets. However, in order to examine the function of biomarkers, it is necessary to find a cell model system that expresses the genes of interest in order to mimic as close to the patient as possible. For example, to study estrogen receptor (ER)-positive breast cancer most researchers will choose the model cell lines MCF7 or T47D which over-expresses ER. As of now, CellMinerHCC (23), CellLineNavigator (24), GENT (6), COSMIC (25) and MERAV (8) are web-tools that can be used for cell line gene expression analysis. CellMinerHCC is a database containing gene expression data of 18 hepatocellular carcinoma (HCC) cell lines, whereas CellLineNavigator, COSMIC and MERAV contain gene expression profiles of diverse cancer cell lines based on microarray profiles and different pathological states and tissues.

Although there are many web-tools available to aid researchers in biomarker discovery, it can still be a confusing and daunting task for them as they need to search and determine which tools are best suited for each of the specific analysis as described above. Thus, it would be useful and much easier if there exists a single tool containing the latest datasets that researchers can use to perform all of the tasks for biomarker discovery as well as functional characterization. However, to the best of our knowledge, there is currently no web-tool that contains all the necessary analysis and information in one single package. To this end, we have constructed KM-Express, a simple and easy to use web-tool that combines, (i) patient survival data, (ii) patient gene expression analysis and (iii) cell line gene expression information. KM-Express holds 36 manually curated RNA-seq data for prostate and breast cancers with 7 cell line datasets from public databases. The current version of KM-Express also includes TCGA datasets for 30 other cancer types along with cell line data support for 23 of them. In summary, KM-Express simultaneously provides survival analysis, cross-dataset and subgroup expression comparison, statistical test, experimental support from cell line and visualization of the data. We present this tool as a hypothesis generation tool to help researchers to identify potential prognostic and diagnosis biomarkers. Finally, this web-tool provides a one-stop analysis platform from bench work to clinical prospects and will greatly facilitate the process of biomarker discovery.

Materials and methods

Data acquisition

We searched in PubMed (http://www.ncbi.nlm.nih.gov/pubmed) for publications with keywords related to breast/prostate and RNA-seq. Specifically, we used the following search terms: 'prostate AND cancer AND ((RNA AND sequencing) OR RNA-seq)' for prostate cancer, and 'breast AND cancer AND ((RNA AND sequencing) OR RNA-seq)' for breast cancer. A biologist might be interested in specific questions related to drug treatment, the overexpression, or knock-down of genes in cell lines. In the current version of KM-Express, we have also included such datasets. From publications, we have obtained the accession ID and downloaded data from the Short Reads Archive (http://www.ncbi.nlm.nih.gov/sra) or dbGaP (http://www.ncbi.nlm.nih.gov/gap) for the authorized dataset (phs000310.v1.p1). Clinical data were sorted from series matrix file(s) on Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/). Gene expression datasets from TCGA was downloaded from https://gdc-portal.nci.nih.gov/projects/t. RNA-seq FPKM data were log2 transformed when comparing the expression level of the different subgroups. The accession IDs and data descriptions are shown in the web interface.

Processing of RNA-Seq data

SRA files were converted into fastq format using fastq-dump from the SRA Toolkit (https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi? view=software). The reads from each library were aligned uniquely and independently with TopHat (version 2.0.6) (26) to the hg19 human reference genome which is available at http://bowtie-bio.sourceforge.net/bowtie2/index.shtml. All of the parameters used were default except for: –splice-mismatches 1, –mate-inner-dist 200, –library-type fr-unstranded and –transcriptome-index. Based on the mapped reads, FPKM values for transcripts were calculated using Cufflinks (version 2.0.2) (26) with default parameters except for –max-bundle-frags 500 000 and –multi-read-correct. Genes were discarded if they were not expressed at appreciable levels (FPKM > 1) in at least half of all the samples in the selected dataset.

Webpage design

A summary of the webpage design for KM-Express is shown in Figure 1. Specifically, the webpage consists of four parts: ‘Input genes’, ‘Dataset Selection’, ‘Parameter Selection’ and ‘Results’. If the input genes are expressed in the dataset, then the output of their expression levels in the different groups of patients and cell lines will be shown according to the clinical or pathological status. If the selected dataset also contains follow-up information, KM-Express will perform survival analysis on this dataset. Users can select the expression level cutoff to define the high and low expressed groups of patients, choose the available patient outcome (alive or dead; recurrence or not), and check if patients with high and low expression of input genes have a significant difference in outcome. The output of the results will show the Kaplan–Meier plot containing the hazard ratio and *P*-value.

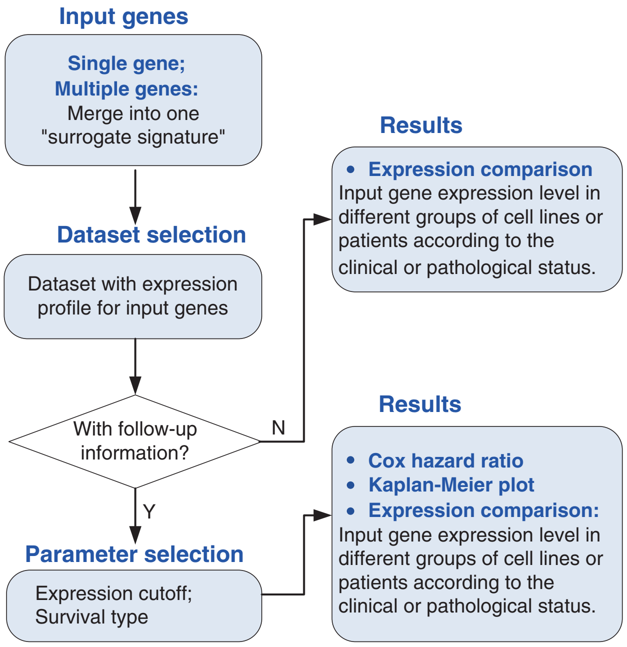

Input genes
Single gene;
Multiple genes:
Merge into one "surrogate signature"
Dataset selection
Dataset with expression profile for input genes
With follow-up information?
N
Y
Parameter selection
Expression cutoff;
Survival type
Results
• Expression comparison
Input gene expression level in different groups of cell lines or patients according to the clinical or pathological status.
Results
• Cox hazard ratio
• Kaplan-Meier plot
• Expression comparison:
Input gene expression level in different groups of cell lines or patients according to the clinical or pathological status.

Figure 1. Overview of the KM-Express web-tool. A flowchart showing how KM-Express works. If the input genes are expressed in the selected dataset, KM-Express will generate an output of their expression level in different groups of patients or cell lines according to the clinical or pathological parameters. If the selected dataset contains follow-up information, the user will be asked to select the expression cutoff to define the high and low expressed groups of patients. After the user chooses the available patient outcome, logrank test is used to determine if high and low expressed groups of patients have significantly different survival rates. The output will contain the Kaplan–Meier plot with hazard ratio and significance *P*-value. If the input contains multiple genes, the survival analysis will be done for the surrogate gene, which represents the average expression level of all the input genes.

The KM-Express web-tool was constructed using PHP and JavaScript. Survival and gene expression analyses were performed using R. PHP was used to pass the user-selected parameters from the webpage to the R program and to visualize all the output results and figures. To increase the speed of data processing by R, RData was used as the format for dataset storage, which can be fast when loading large expression profiles into R. Additionally, the expression datasets were processed and stored in the unified RData format, and thus there is no need to revise the scripts even when new datasets are added in the future. Therefore, RData formatted expression profiling makes KM-Express run faster and easier to maintain.

We have included a function panel called ‘DataSubmit’ that allows users or maintainers of KM-Express to easily update the datasets. Users simply submit two files, one for the gene expression profiling and the other for the sample annotation with or without available phenotype information. Examples of submitted files are provided. After being

Table 1. Datasets for human prostate cancer patients and cell lines

Accession ID	PMID	Phenotype	Sample size
TCGA		Tumor	384
SRP011439	21 804 560	Benign vs localized tumor	114
SRP005908	21 036 922	Prostate patients	28
GSE24283	21 261 984	Tumor vs normal	3/3
SRP002628	21 571 633	Tumors and adjacent normal tissues	20/10
ERP000550	22 349 460	Tumor and adjacent normal tissues	14/14
SRP026387	24 054 872	Pre and post treatment	7/7
GSE48812	25 044 704	PrEC, LNCaP, PC3	12/12/12
SRP013621	24 012 641	Tumor	1/1
SRP027258	26 805 894	PC3, LNCaP, PrEC	12/12/12
SRP010054	25 700 553	LNCaP	3/3
SRP014759	25 274 489	Tumor	1/1
GSE25183	21 804 560	Normal vs tumor	58

checked and curated by the administrator, the submitted files will be loaded and saved as RData format for survival and expression analysis in KM-Express. We will update the datasets for KM-Express every 6 months.

Input gene(s)

The official gene symbol from the HUGO Gene Nomenclature Committee (HGNC) is the only acceptable ID for entering genes in KM-Express. If entering multiple genes, the survival analysis will be performed for the surrogate gene, which represents the average expression level of all input genes. The expression analysis for patient samples or cell lines will show output expression comparison for each input gene. KM-Express will not process any genes with a low expression level (FPKM $<$ 1 in more than half of the samples in the selected dataset).

Hazard ratio and *P*-value

The R package ‘survival’ was used to calculate and plot Kaplan–Meier survival curves (27, 28).

Results

Data source

Currently, KM-Express contains extensive transcriptomic expression profiling based on RNA-seq data for prostate and breast cancers. Specifically, we have manually curated 13 datasets for prostate cancer and 25 datasets for breast cancer from various publications (Table 1 and Supplementary Table S1). In addition, we have begun to incorporate the expression data for 30 other cancer types from the TCGA database (Supplementary Table S2). A total of 65 cell line datasets representing 23 of these cancer types have also been included (Supplementary Table S3). The full list of the datasets can be found in the ‘Data Source’ panel of KM-Express.

Functionality

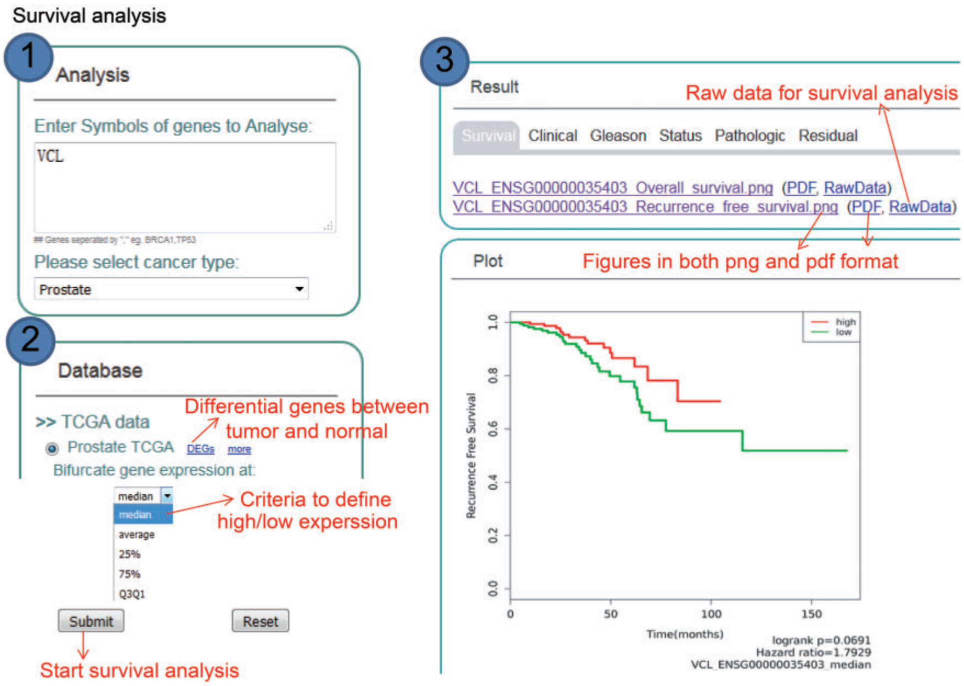
KM-Express provides a simple user-friendly web interface that enables users to perform multiple analyses including survival analysis based on patient RNA-seq datasets and gene expression analysis based on patient and cell line RNA-seq datasets. In addition, KM-Express provides a module for statistical tests to assess the differential significance of the selected gene(s). KM-Express is very simple to use and analyses can be accomplished in about one minute. The results from KM-Express are provided as high-resolution PNG format or they can be download as PDFs if users want to make modifications on them.

For a given candidate gene, KM-Express can:

- Determine if it is potential prognostic markers or not. According to the survival analysis results, users can infer its potency to predict patient outcome.
- Check if it is an oncogene or a tumor suppressor by comparing its expression level between normal and tumor samples. Multiple independent datasets from different labs enable a cross-validation of the gene to give more confidence to its role as an oncogene or a tumor suppressor.
- Help in the design of wet lab experiments. Users can refer to the gene expression level in different cell lines or under different interventions or manipulations, which facilitates the hypothesis generation on gene function.

Survival analysis

One major function of KM-Express is patient survival analysis (Figure 2). To perform survival analysis, users simply go to the ‘Analysis’ panel of the homepage and type in


Survival analysis
1
Analysis
2
Database
>> TCGA data
● Prostate TCGA DEGs more
Differential genes between tumor and normal
median
average
25%
75%
Q3Q1
Criteria to define high/low experssion
Submit
Reset
Start survival analysis
3
Result
Raw data for survival analysis
Survival Clinical Gleason Status Pathologic Residual
VCL_ENSG00000035403_Overall_survival.png (PDF, RawData)
VCL_ENSG00000035403_Recurrence_free_survival.png (PDF, RawData)
Figures in both png and pdf format
Plot
high
low
Recurrence Free Survival
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
Time(months)

Enter Symbols of genes to Analyse:
VCL
Genes seperated by "," eg. BRCA1,TP53

Please select cancer type:
Prostate

Bifurcate gene expression at:
median

logrank p=0.0691
Hazard ratio=1.7929
VCL_ENSG00000035403_median

Figure 2. Step-by-step usage of KM-Express for survival analysis. (1) Enter the gene symbol in the box provided. When users input multiple genes, gene symbols should be separated by ','. The expression level of the multiple genes will be averaged and used for survival analysis. After the cancer type is selected, the RNA-seq data available for survival analysis will be shown in 'Database' panel. (2) Select the survival measure. The survival outcome is calculated as biochemical recurrence for tumor or without tumor for prostate cancer and dead or alive status for breast cancer. Next, select the criteria to divide the patients into two groups, one group highly expressing the input gene, the other group lowly expressing the input gene. For median/average/25%/75% cutoff, higher than the threshold corresponds to highly express group, others are lowly express group. For Q3Q1, highly expressed group represents patients with input gene expression >75% patients, while the lowly expressed group represents patients with input gene expression <25% patients. (3) The output provides KM-plot figures in both PDF and PNG format. The significance *P*-value and hazard ratio are provided in the figure along with the summary of input genes and user-selected parameters.

the gene symbol of the gene(s) they are interested in (separated by a comma if more than one gene) and then select the cancer type they want to look at (e.g. prostate or breast cancer). Users then choose the survival measure and expression cutoff to divide patients into two groups. After users submit the survival job, a Kaplan–Meier plot will show if there is a significant difference in survival time between the two groups of patients. The output figures are provided to users as publication quality high-resolution PNG and PDF formats. If users prefer to reproduce or perform survival analysis using their own personalized criteria, they can download the raw data from the 'RawData' tab of the output page. If the TCGA dataset is selected, gene expression analysis is simultaneously performed with survival analysis.

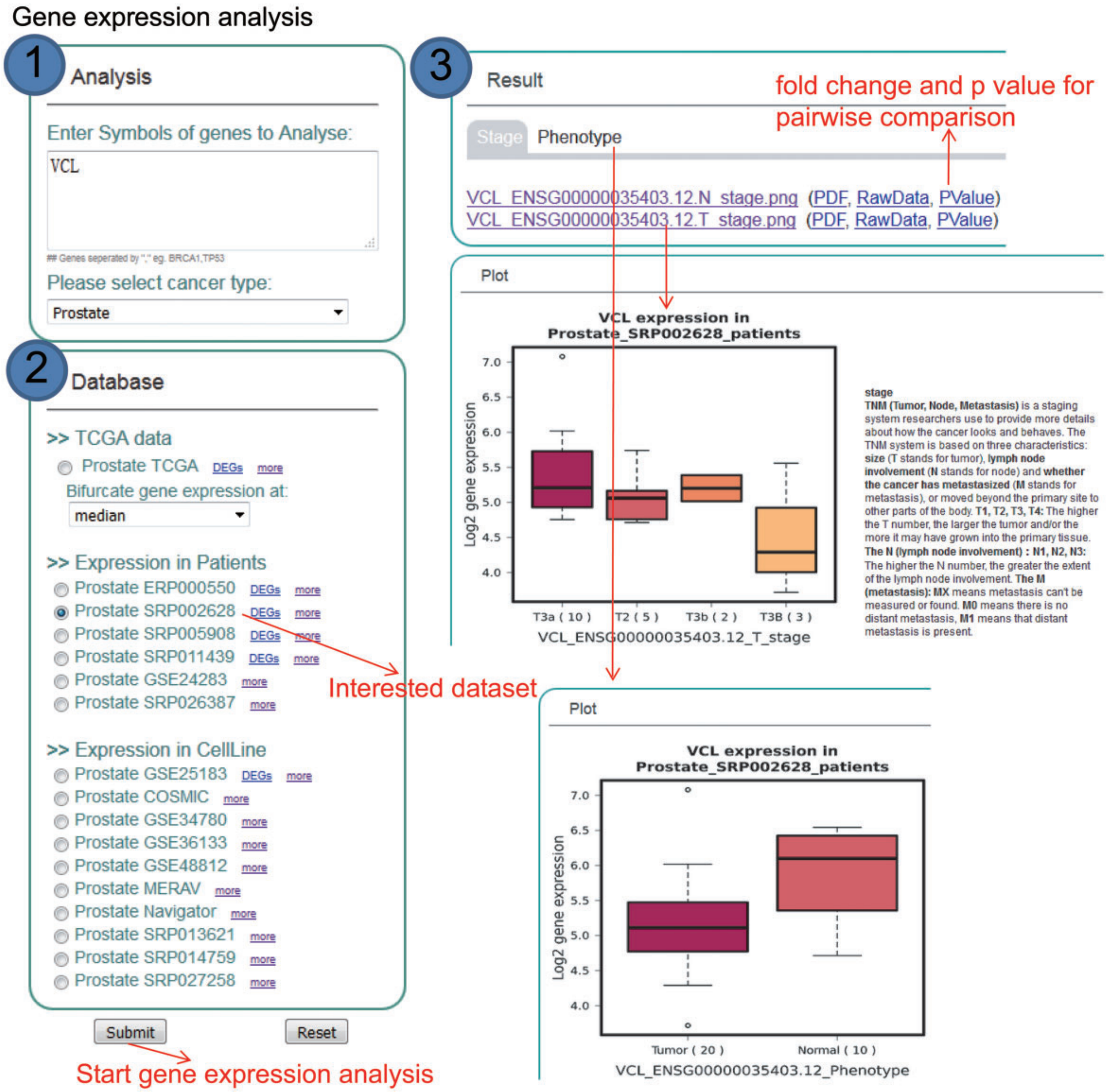
Gene expression analysis

KM-Express is also designed to perform gene expression analysis based on patients and cell line datasets (Figure 3). In the 'Analysis' panel of the homepage, users type in the gene symbol of the gene(s) they are interested in and then select the cancer type and a dataset from either patients or cell lines. After the job has been submitted and the analysis is finished, the result of the expression comparison for the input gene will be shown as a boxplot and will be categorized according to the clinical or pathological information based on the study. To help users better understand the clinical or pathological terms, interpretive statements are shown along with the boxplot.

We realize that not all potential KM-Express users will already have candidate biomarker gene(s) and thus may want to identify their own using the datasets archived in KM-Express. To facilitate this, we have included in KM-Express an option where users can obtain differentially expressed genes (DEGs) via the tab 'DEGs' (Figures 2 and 3) on any of the patient datasets that we currently have in KM-Express. Currently, KM-Express will provide a list of DEGs between normal and tumor samples with a fold change >1.2 and $P < 0.05$ (note: KM-Express will only do this for datasets that contain >3 tumors and 3 normal samples). Once users find the candidate gene(s) they are interested in, they can then perform similar survival and gene expression analysis as described above.

Statistical tests for differential significance

KM-Express provides statistical tests to determine whether a gene is significantly differentially expressed between any



Gene expression analysis

1 Analysis

Enter Symbols of genes to Analyse: VCL

Genes seperated by "," eg. BRCA1,TP53

Please select cancer type: Prostate

2 Database

Bifurcate gene expression at: median

● Prostate SRP002628

3 Result

fold change and p value for pairwise comparison

Interested dataset

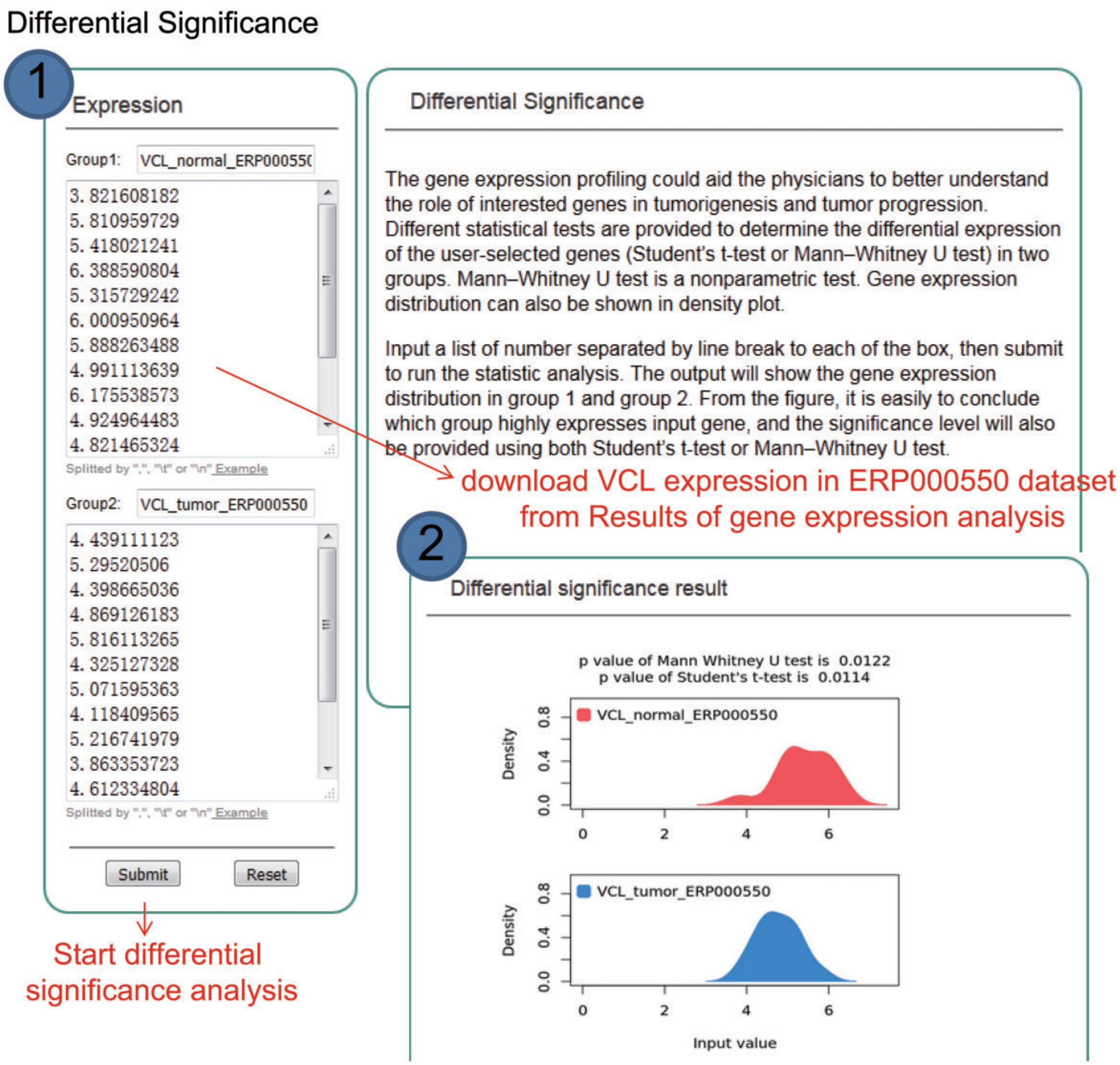
Start gene expression analysis

stage
TNM (Tumor, Node, Metastasis) is a staging system researchers use to provide more details about how the cancer looks and behaves. The TNM system is based on three characteristics: **size** (T stands for tumor), **lymph node involvement** (N stands for node) and **whether the cancer has metastasized** (M stands for metastasis), or moved beyond the primary site to other parts of the body. **T1, T2, T3, T4:** The higher the T number, the larger the tumor and/or the more it may have grown into the primary tissue. **The N (lymph node involvement) : N1, N2, N3:** The higher the N number, the greater the extent of the lymph node involvement. **The M (metastasis): MX** means metastasis can't be measured or found. **M0** means there is no distant metastasis, **M1** means that distant metastasis is present.

>> TCGA data
○ Prostate TCGA DEGs more
>> Expression in Patients
○ Prostate ERP000550 DEGs more
● Prostate SRP002628 DEGs more
○ Prostate SRP005908 DEGs more
○ Prostate SRP011439 DEGs more
○ Prostate GSE24283 more
○ Prostate SRP026387 more
>> Expression in CellLine
○ Prostate GSE25183 DEGs more
○ Prostate COSMIC more
○ Prostate GSE34780 more
○ Prostate GSE36133 more
○ Prostate GSE48812 more
○ Prostate MERAV more
○ Prostate Navigator more
○ Prostate SRP013621 more
○ Prostate SRP014759 more
○ Prostate SRP027258 more
Submit
Reset
Stage
Phenotype
VCL_ENSG00000035403.12.N_stage.png (PDF, RawData, PValue)
VCL_ENSG00000035403.12.T_stage.png (PDF, RawData, PValue)
Plot
VCL expression in Prostate_SRP002628_patients
Log2 gene expression
7.0
6.5
6.0
5.5
5.0
4.5
4.0
T3a (10)
T2 (5)
T3b (2)
T3B (3)
VCL_ENSG00000035403.12_T_stage
Plot
VCL expression in Prostate_SRP002628_patients
Log2 gene expression
7.0
6.5
6.0
5.5
5.0
4.5
4.0
Tumor (20)
Normal (10)
VCL_ENSG00000035403.12_Phenotype

Figure 3. Step-by-step usage of KM-Express for expression analyses. (1) Enter the gene symbol in the box provided. When users input multiple genes, gene symbols should be separated by ','. After selecting the cancer type, the RNA-seq data available for expression analysis will be shown in 'Database' panel. (2) Select a dataset to compare the expression of input genes in different groups of patients/cell lines, according to clinical, pathological status of patients or time-series drug treatment or different cell line types. (3) The output provides expression level of input genes in different groups of patients according to their clinical, pathological status, time-series drug treatment or different cell line types.

two groups. Since a dataset can sometime contain many different cancer types or conditions (e.g. such as breast cancer) and therefore many comparisons, we have provided a downloadable file under the 'PValue' tab (Figure 3, under Results) which contains both the fold change and the corresponding *P*-value for each pairwise comparison. If users would like to examine the expression of their gene(s) in any two clinical or pathological subgroups in more detail, they can download the raw expression data from the 'RawData' tab (Figure 3, under Results) and perform their own statistical tests or use the 'Differential Significance' function in KM-Express (Figure 4). As shown in Figure 4, users input the expression values of the gene into two groups (e.g. Group 1 = normal and Group 2 = tumor). After the statistical test is complete, a density plot of the gene expression distribution for each group is shown along with the significance using the Student's *t*-test and the Mann–Whitney *U*-test.



Differential Significance

1

Expression

Group1: VCL_normal_ERP00055(

3. 821608182
5. 810959729
5. 418021241
6. 388590804
5. 315729242
6. 000950964
5. 888263488
4. 991113639
6. 175538573
4. 924964483
4. 821465324

Splitted by ",", "\t" or "\n" Example

Group2: VCL_tumor_ERP000550

4. 439111123
5. 29520506
4. 398665036
4. 869126183
5. 816113265
4. 325127328
5. 071595363
4. 118409565
5. 216741979
3. 863353723
4. 612334804

Splitted by ",", "\t" or "\n" Example

Differential Significance

The gene expression profiling could aid the physicians to better understand the role of interested genes in tumorigenesis and tumor progression. Different statistical tests are provided to determine the differential expression of the user-selected genes (Student's t-test or Mann–Whitney U test) in two groups. Mann–Whitney U test is a nonparametric test. Gene expression distribution can also be shown in density plot.

Input a list of number separated by line break to each of the box, then submit to run the statistic analysis. The output will show the gene expression distribution in group 1 and group 2. From the figure, it is easily to conclude which group highly expresses input gene, and the significance level will also be provided using both Student's t-test or Mann–Whitney U test.

download VCL expression in ERP000550 dataset from Results of gene expression analysis

2

Differential significance result

p value of Mann Whitney U test is 0.0122
p value of Student's t-test is 0.0114

Start differential significance analysis

Figure 4. Step-by-step usage of KM-Express for differential significance test. (1) Input the gene expression level in two groups, separating each number by a line break, tab or comma, and then submit to run the statistics analysis. Two statistical tests are simultaneously provided to determine the differential significance of the user-selected genes (Student's *t*-test and Mann–Whitney *U*-test) in two groups. (2) The output provides a gene expression distribution density plot. The significance level is also provided using both Student's *t*-test and Mann–Whitney *U*-test.

While the 'Differential Significance' function was designed for analyzing data from KM-Express, we believe it can also serve as a useful tool for users if they wish to examine the statistical significance of the expression difference for their target genes from their own in-house data.

Data maintenance and update

We have implemented a function panel 'DataSubmit' that enables users or maintainers of KM-Express to submit new datasets by following the format of the example files. The datasets in KM-Express will be updated approximately every 6 months.

Documentation, tutorial and help

Documentation for KM-Express includes data source, help and detailed tutorials (http://ec2-52-201-246-161.compute-1.amazonaws.com/kmexpress/Tutorial.php). In the 'Data Source' page, a detailed description (Table 1 and Supplementary Tables S1, S2 and S3) is provided for each dataset including the accession ID, publication, sample size, normal/tumor/metastatic phenotype and clinical/pathological

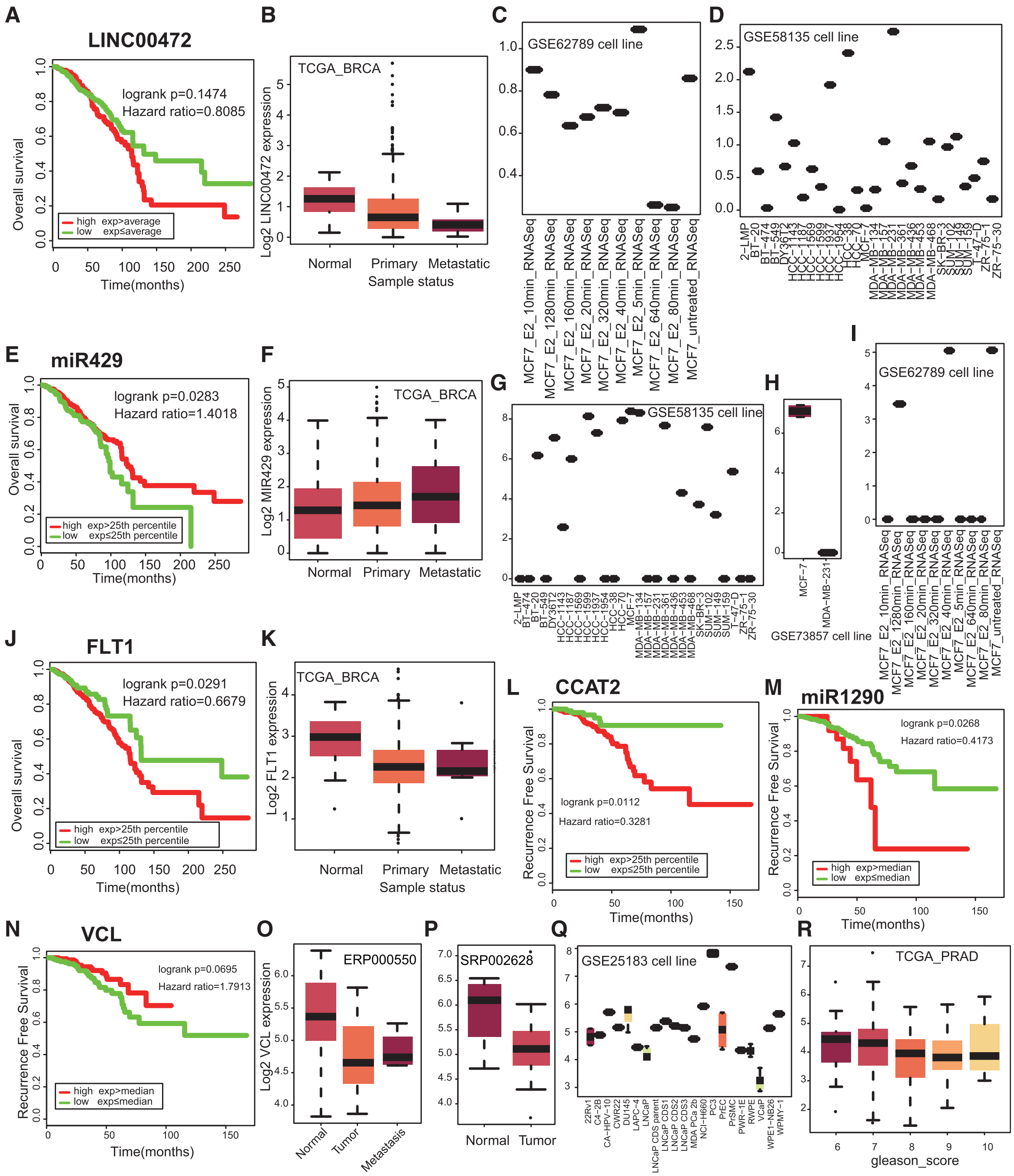


Figure 5. Validation of previously reported prognostic biomarkers. (**A**) Kaplan–Meier plot for lncRNA LINC00472. (**B**) Differential expression of LINC00472 among tumor and normal patients. Expression change of LINC00472 after estradiol (E2) treatment in MCF7 cells (**C**) and in 28 breast cancer cell lines (**D**). (**E**) Kaplan–Meier plot for miR429. (**F**) Differential expression of miR429 in metastatic patients compared with primary breast cancer patients and normal tissue. Expression of miR429 in different cell lines (**G, H**) and after E2 treatment (**I**). (**J**) Kaplan–Meier plot for the protein coding gene VEGFR1 (FLT1). Expression of FLT in normal breast samples and cancer patients (**K**). Kaplan–Meier plot for lncRNA CCAT2 (**L**) and miR1290 (**M**) and VCL (**N**). Differential expression of VCL between normal and tumor prostate (**O, P**) and in prostate cancer cell lines (**Q**). (**R**) VCL is differentially expressed in patients with different Gleason score.

characteristic. A step-by-step tutorial is provided for users to quickly execute examples using KM-Express. The ‘Help’ page allows users to have a better understanding of the clinical or pathological terms, which are used to illustrate the cancer staging.

Validation and application

To demonstrate the usefulness of KM-Express on biomarker discovery and functional validation, we examined several previously reported RNA biomarkers for breast cancer including the long non-coding RNA (lncRNA)

LINC00472 (29), miR429 (30, 31) and VEGFR1 (FLT1) (32). In general, the results from KM-Express for patient survival and gene expression data in normal vs tumor samples were consistent with previous findings for the three RNA biomarkers (Figure 5A, B, E, F, J and K). The only inconsistent result was for the patient survival finding of LINC00742, which could be attributed to the larger and more diverse cohorts included in the TCGA dataset. In addition to breast cancer biomarkers, we also examined RNA biomarkers for prostate cancer such as CCAT2 (33), miRNA1290 (34) and VCL (35). Again, KM-Express was able to obtain similar results as previously reported (Figure 5L–P).

A unique and useful feature which has been incorporated into KM-Express and is not found in other biomarker discovery tools is the inclusion of cell line gene expression information. We have added this data in KM-Express so that biologists can have the ability in the same tool to examine the gene expression profile of their RNA biomarkers not only in patient samples but also in different cell lines. This feature will negate the need to find and use other tools and thus will facilitate the design of downstream functional assays to test whether RNA biomarkers are also potential therapeutic targets.

This functionality is nicely illustrated for the breast cancer biomarkers LINC00472 and miR429 (Figure 5). Specifically, the cell line gene expression data for LINC00472 in KM-Express shows that the transcript is expressed only in MCF-7 cells after estrogen stimulation (Figure 5C) and its expression level is the highest in MDA-MB-231 cells among 28 breast cancer cell lines (Figure 5D). In contrast, KM-Express shows miR429 expression level is the highest in MCF-7 cells (but lowest in MDA-MB-231 cells) and the level of the miRNA is dramatically reduced after estrogen stimulation (Figure 5I). The cell line information for the two biomarkers would therefore suggest that MCF-7 is a good cell line model for studying the function and hormone regulation of both transcripts while MDA-MB-231 cells would make a good cell line model for LINC00472 knock-down studies and miR429 overexpressing work.

Future extensions

We will manually curate newly generated RNA-seq datasets for patients or cell lines every 6 months. Currently, users can perform both expression and survival analysis with breast and prostate cancers on KM-Express. Users can also carry out similar analysis on 30 other cancer types based on TCGA data with cell line expression analysis for most of the cancer types. We will continue to expand and include new datasets for all these cancer types.

Discussion

We have developed KM-Express, a simple user-friendly web-based tool. KM-Express is not the only web-tool which can identify prognostic markers based on RNA-seq transcriptome dataset, however, it is different from other web-tools in that it, (i) contains NGS transcriptome data in addition to TCGA datasets, (ii) provides detailed expression comparison of interested genes in cancer compared to normal, or in different subgroups of patients according to the molecular, genetic, clinical and pathological status, (iii) helps the researcher in experimental design based on the expression of gene(s) in different cell lines and (iv) includes an analysis tool to visualize and statistically determine the significance of gene-centric expression differences in two groups. To ensure quick and easy searching ability, an intuitive querying interface was implemented. The survival analysis is very quick and is accomplished in about one second.

In the coming years, expression data based on NGS will continue to accumulate for many cancers, highlighting the importance of a published web-tool providing comprehensive information about prognostic markers, cross-dataset expression and experimental support from cell lines.

Conclusions

In brief, KM-Express hosts comprehensive NGS-based expression profiles and provides convenient and user-friendly tools to explore potential prognostic markers. We have generated KM-Express as a hypothesis generation tool for researchers to identify potential prognostic RNA biomarkers to follow-up for further research. KM-Express will help accelerate translational research from bench to bedside.

Supplementary data

Supplementary data are available at *Database* Online.

Acknowledgements

The authors are grateful to Shreyas Lingadahalli, and Lingling Hu for their discussions and contributions. We would also like to acknowledge the Genomics and Bioinformatics Core at FHS. This work was performed in part at the High Performance Computing Cluster (HPCC) which is supported by Information and Communication Technology Office (ICTO) of the University of Macau. We thank Jacky Chan Hoi Kei and William Pang from HPCC for their support and help.

Funding

This work was supported in part by funding from the University of Macau (MYRG 2015-00196-FHS) and the Macau Science and Technology Development Fund (FDCT/023/2014/A1, FDCT102/2015/A3) to E.C. Funding to pay the Open Access publication charges for this article was provided by University of Macau.

Conflict of interest. None declared.

References

1. Risbridger,G.P., Davis,I.D., Birrell,S.N. *et al*. (2010) Breast and prostate cancer: more similar than different. *Nat. Rev. Cancer*, **10**, 205–212.
2. Jordan,V.C., Fan,P., Abderrahman,B. *et al*. (2016) Sex steroid induced apoptosis as a rational strategy to treat anti-hormone resistant breast and prostate cancer. *Discov. Med*., **21**, 411–427.
3. Capper,C.P., Rae,J.M. and Auchus,R.J. (2016) The metabolism, analysis, and targeting of steroid hormones in breast and prostate cancer. *Horm. Cancer*, **7**, 149–164.
4. Rhodes,D.R., Kalyana-Sundaram,S., Mahavisno,V. *et al*. (2007) Oncomine 3.0: genes, pathways, and networks in a collection of 18, 000 cancer gene expression profiles. *Neoplasia*, **9**, 166–180.
5. Rhodes,D.R., Yu,J., Shanker,K. *et al*. (2004) ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia*, **6**, 1–6.
6. Shin,G., Kang,T.W., Yang,S. *et al*. (2011) GENT: gene expression database of normal and tumor tissues. *Cancer Inform*., **10**, 149–157.
7. Wan,Q., Dingerdissen,H., Fan,Y. *et al*. (2015) BioXpress: an integrated RNA-seq-derived gene expression database for pan-cancer analysis. *Database (Oxford)*, **2015**, bav019.
8. Shaul,Y.D., Yuan,B., Thiru,P. *et al*. (2016) MERAV: a tool for comparing gene expression across human tissues and cell types. *Nucleic Acids Res*., **44**, D560–D566.
9. Mizuno,H., Kitada,K., Nakai,K. *et al*. (2009) PrognoScan: a new database for meta-analysis of the prognostic value of genes. *BMC Med. Genomics*, **2**, 18.
10. Madhavan,S., Gusev,Y., Harris,M. *et al*. (2011) G-DOC: a systems medicine platform for personalized oncology. *Neoplasia*, **13**, 771–783.
11. Ringner,M., Fredlund,E., Hakkinen,J. *et al*. (2011) GOBO: gene expression-based outcome for breast cancer online. *PLoS One*, **6**, e17911.
12. Aguirre-Gamboa,R., Gomez-Rueda,H., Martinez-Ledesma,E. *et al*. (2013) SurvExpress: an online biomarker validation tool and database for cancer gene expression data using survival analysis. *PLoS One*, **8**, e74250.
13. Madden,S.F., Clarke,C., Gaule,P. *et al*. (2013) BreastMark: an integrated approach to mining publicly available transcriptomic datasets relating to breast cancer outcome. *Breast Cancer Res*., **15**, R52.
14. Aguirre-Gamboa,R. and Trevino,V. (2014) SurvMicro: assessment of miRNA-based prognostic signatures for cancer clinical outcomes by multivariate survival analysis. *Bioinformatics*, **30**, 1630–1632.
15. Gyorffy,B., Lanczky,A., Eklund,A.C. *et al*. (2010) An online survival analysis tool to rapidly assess the effect of 22, 277 genes on breast cancer prognosis using microarray data of 1, 809 patients. *Breast Cancer Res. Treat*., **123**, 725–731.
16. Gyorffy,B., Lanczky,A. and Szallasi,Z. (2012) Implementing an online tool for genome-wide validation of survival-associated biomarkers in ovarian-cancer using microarray data from 1287 patients. *Endocr. Relat. Cancer*, **19**, 197–208.
17. Győrffy,B., Surowiak,P., Budczies,J. *et al*. (2013) Online survival analysis software to assess the prognostic value of biomarkers using transcriptomic data in non-small-cell lung cancer. *PLoS One*, **8**, e82241.
18. Lanczky,A., Nagy,A., Bottai,G. *et al*. (2016) miRpower: a web-tool to validate survival-associated miRNAs utilizing expression data from 2178 breast cancer patients. *Breast Cancer Res. Treat*., **160**, 439–446.
19. Szasz,A.M., Lanczky,A., Nagy,A. *et al*. (2016) Cross-validation of survival associated biomarkers in gastric cancer using transcriptomic data of 1, 065 patients. *Oncotarget*, **7**, 49322–49333.
20. Cerami,E., Gao,J., Dogrusoz,U. *et al*. (2012) The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov*., **2**, 401–404.
21. Gao,J., Aksoy,B.A., Dogrusoz,U. *et al*. (2013) Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal*, **6**, pl1.
22. Goswami,C.P. and Nakshatri,H. (2014) PROGgeneV2: enhancements on the existing database. *BMC Cancer*, **14**, 970.
23. Staib,F., Krupp,M., Maass,T. *et al*. (2014) CellMinerHCC: a microarray-based expression database for hepatocellular carcinoma cell lines. *Liver Int*., **34**, 621–631.
24. Krupp,M., Itzel,T., Maass,T. *et al*. (2013) CellLineNavigator: a workbench for cancer cell line analysis. *Nucleic Acids Res*., **41**, D942–D948.
25. Forbes,S.A., Beare,D., Boutselakis,H. *et al*. (2017) COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res*., **45**, D777–D783.
26. Trapnell,C., Roberts,A., Goff,L. *et al*. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc*., **7**, 562–578.
27. Therneau,T. and Grambsch,P. (2000) *Modeling Survival Data: Extending the Cox Model*. Springer, New York.
28. Therneau,T. (2015) *A Package for Survival Analysis in S, Version 2.38*. https://CRAN.R-project.org/package=survival (28 June 2018, date last accessed).
29. Shen,Y., Katsaros,D., Loo,L.W. *et al*. (2015) Prognostic and predictive values of long non-coding RNA LINC00472 in breast cancer. *Oncotarget*, **6**, 8579–8592.
30. Kong,X., Ding,X., Li,X. *et al*. (2015) 53BP1 suppresses epithelial-mesenchymal transition by downregulating ZEB1 through microRNA-200b/429 in breast cancer. *Cancer Sci*., **106**, 982–989.
31. Ye,Z.B., Ma,G., Zhao,Y.H. *et al*. (2015) miR-429 inhibits migration and invasion of breast cancer cells in vitro. *Int. J. Oncol*., **46**, 531–538.
32. Linardou,H., Kalogeras,K.T., Kronenwett,R. *et al*. (2012) The prognostic and predictive value of mRNA expression of vascular endothelial growth factor family members in breast cancer: a study in primary tumors of high-risk early breast cancer patients participating in a randomized Hellenic Cooperative Oncology Group trial. *Breast Cancer Res*., **14**, R145.
33. Zheng,J., Zhao,S., He,X. *et al*. (2016) The up-regulation of long non-coding RNA CCAT2 indicates a poor prognosis for prostate cancer and promotes metastasis by affecting epithelial-mesenchymal transition. *Biochem. Biophys. Res. Commun*., **480**, 508–514.
34. Huang,X., Yuan,T., Liang,M. *et al*. (2015) Exosomal miR-1290 and miR-375 as prognostic markers in castration-resistant prostate cancer. *Eur. Urol*., **67**, 33–41.
35. Chng,K.R., Chang,C.W., Tan,S.K. *et al*. (2012) A transcriptional repressor co-regulatory network governing androgen response in prostate cancers. *EMBO J*., **31**, 2810–2823.