Gene expression and survival analysis in cancer research using online open access platforms: a comparative analysis

Sofia Ramos, Ana Sofia Fernandes, Nuno Saraiva

CBIOS, Universidade Lusófona Research Center for Biosciences & Health Technologies Universidade Lusófona, Av Campo Grande, 376, 1749-024, Lisboa, Portugal

*corresponding author: nuno.saraiva@ulusofona.pt

Abstract

The development of genomics and transcriptomics and the potential associated with sharing data related with cancer, led to a growing understanding of cancer biology and to the identification of new biomarkers.

Analysis of tumor gene expression and associated patient survival rate enables the dissection of the impact of certain genes in cancer patient’s survival. For that purpose, it is essential to choose user-friendly platforms, where it is easy to analyze, compare and collect information for a certain set of genes. The goal of this article is to compare the content and utility of five open access online platforms for tumor gene expression and patient survival analysis from TCGA datasets – cBioPortal, UCSC Xena, GEPIA, UALCAN and ONCOLNC. We explore these platforms from the point of view of a lay user, assessing their applicability to study differences in gene expression in tumor vs normal tissues, or according to cancer stage, and the impact of such expression patterns on patient survival. Although all five platforms are very intuitive and access to the data is easy, they vary in the information available, results visualization, and statistical tests performed. Therefore, the choice of a platform must take into account the study goals. For some purposes, a combination of platforms may be required.

Keywords: TCGA datasets, open access platforms, tumor gene expression, patient survival

Resumo

O desenvolvimento da genómica e transcriptómica e o potencial associado à partilha de dados relacionados com o cancro levaram a uma crescente compreensão da sua biologia e à identificação de novos biomarcadores.

A análise da expressão génica do tumor e a taxa de sobrevivência do doente associada, permitem explorar o impacto de um determinado gene na sobrevivência do doente com cancro. Para isso, é essencial escolher plataformas simples de usar, onde seja fácil analisar, comparar e recolher informações para um determinado conjunto de genes de interesse. O objetivo deste artigo é comparar o conteúdo e utilidade de cinco plataformas on-line de acesso aberto para expressão de genes tumorais e análise da sobrevivência de doentes a partir de conjuntos de dados do TCGA - cBioPortal, UCSC Xena, GEPIA, UALCAN e ONCOLNC. Exploramos as plataformas acima mencionadas do ponto de vista de um utilizador médio, avaliando a sua aplicabilidade no estudo de diferenças na expressão génica em tecidos tumorais vs tecidos normais, ou de acordo com o estádio do cancro, e o impacto desse padrão de expressão na sobrevivência do doente. Embora todas as cinco plataformas sejam muito intuitivas e o acesso aos dados seja fácil, elas variam nas informações disponibilizadas, na visualização dos resultados e nos testes estatísticos realizados. Logo, a escolha de uma plataforma deve ter em consideração os objetivos do estudo. Para alguns propósitos, pode ser necessária uma combinação de plataformas.

Palavras-chave: Bases de dados do TCGA; plataformas de open access; expressão génica tumoral; sobrevida do doente
Introduction

Due to the high rate of cancer morbidity and mortality, several studies on etiology and therapeutics have been developed, including studies focused on the genome and transcriptome. This resulted in a wide range of data for each patient, collected in databases, which may include mRNA, miRNA and protein levels, gene copy number, and mutation variations, in addition to clinical information (1). One good example of such type of database is The Cancer Genome Atlas (TCGA), a program which aggregates data from over 11000 cancer patients and 30 different types of cancer (2). These raw findings have been shared in open access international platforms and made available to the entire scientific community to broaden the spectrum of cancer information.

There are cancer-specific platforms (eg., for cervical cancer - CCDB, http://crdd.osdd.net/raghava/ccdb/index.php, or for glioblastoma – IVY GAP https://glioblastoma.alleninstitute.org/), and more comprehensive ones that provide information on various types of cancers (eg., cBioPortal, http://www.cbioportal.org/).

Platforms specialised in a given type of data are also available. This is the case of microRNA profiles (mirCancer, http://mircancer.ecu.edu/), gene expression (GEPIA2, http://gepia2.cancer-pku.cn/#index) or genomic and epigenomic data of diverse origin (eg., UCSC Xena, https://xena.ucsc.edu/, which provides data on copy number variation, DNA methylation, gene and protein expression, etc.), or even the relation between survival data and mRNA, miRNA or IncRNA expression levels (OncoLnc, http://www.oncolnc.org/).

Other platforms are focused on drug sensitivity and response (eg., Genomics of cancer drug sensitivity, https://www.cancerrxgene.org/), or are oriented to the study of drug biomarkers (eg., The Institute of Cancer Research, https://cansarblack.icr.ac.uk/).

This article focuses on the analysis of gene expression levels, since it allows the dissection of gene dysregulations and the assessment of its consequences for cancer progression and survival, and may thus provide clues to the pathways involved in cancer pathogenesis, as well as on potential therapeutic targets (3). Additionally, it is possible to estimate prognosis from the survival rates, by comparing high and low risk scores using a ratio, such as the widely used hazard ratio. Patients with high risk scores have poor survival rates compared with those with low risk scores, which

Introdução

Devido à alta taxa de morbimortalidade por cancro, vários estudos sobre etiologia e terapêutica têm sido desenvolvidos, incluindo estudos focados no genoma e transcriptoma, o que resultou numa ampla gama de dados para cada doente, que pode incluir níveis de mRNA, miRNA e proteína, número de cópias de genes e variações de mutações, além de informações clínicas (1). Um bom exemplo deste tipo de banco de dados é o The Cancer Genome Atlas (TCGA), um programa que agrega dados de mais de 11000 doentes com cancro e 30 tipos diferentes de cancro (2). Estes dados em bruto têm sido compartilhados em plataformas internacionais de acesso aberto e disponibilizados a toda a comunidade científica para ampliar o espectro de informações sobre o cancro.

Existem plataformas específicas para determinados tipos de cancro (por exemplo, para cancro do colo do útero - CCDB, http://crdd.osdd.net/raghava/ccdb/index.php, ou para glioblastoma - IVY GAP https://glioblastoma.alleninstitute.org/) e outras mais abrangentes que fornecem informações sobre vários tipos de cancro (por exemplo, cBioPortal, http://www.cbioportal.org/).

Plataformas especializadas num determinado tipo de dados também estão disponíveis. É o caso de perfis de microRNA (mirCancer, http://mircancer.ecu.edu/), ou expressão de genes (http://gepia2.cancer-pku.cn/#index), ou dados genómicos e epigenómicos de origem diversa (por exemplo, UCSC Xena , https://xena.ucsc.edu/, que fornece dados sobre variação do número de cópias, metilação do DNA, expressão de genes e proteínas, etc.), ou mesmo a relação entre dados de sobrevida e níveis de expressão de mRNA, miRNA ou IncRNA (OncoLnc, http://www.oncolnc.org/).

Outras plataformas estão focadas na sensibilidade e resposta a medicamentos (por exemplo, Genomics of cancer drug sensitivity, https://www.cancerrxgene.org/) ou estão direcionadas para o estudo de biomarcadores de medicamentos (por exemplo, The Institute of Cancer Research, https://cansarblack.icr.ac.uk/).

Este artigo foca-se na análise dos níveis de expressão génica, uma vez que esta abordagem permite avaliar desregulações génicas e as suas consequências para a progressão do cancro e sobrevida, podendo, assim, fornecer pistas sobre as vias envolvidas na patogénese do cancro, bem como sobre possíveis alvos terapêuticos (3). Adicionalmente, é possível estimar um prognóstico a partir das taxas de sobrevida, comparando os scores de alto e baixo risco utilizando um ratio, como o amplamente utilizado hazard ratio. Doentes com scores
contribute to determine risk score without using other clinical variables and information (4).

In this work, we describe and analyse the cBioPortal, UCSC Xena, GEPIA and UALCAN platforms, which provide data on gene expression of several types, and stages of cancer and patient survival analysis. OncoLnc, a platform focused in survival analysis was also included.

The applicability of the different platforms was assessed using the expression of the Thioredoxin-1 (TXN) gene in breast cancer as an example. Thioredoxin-1 is an enzyme that contributes to many hallmarks of several types of cancer, including increased tumor cell proliferation and growth, resistance to cell death, and increased angiogenesis (5, 6). Thioredoxin-1 is associated with aggressive tumor growth, resistance to standard therapy, and decreased patient survival (5), and may therefore constitute a potential therapeutic target and a biomarker for breast cancer aggressiveness.

By comparing these online resources/platforms we aim at providing a starting point for future users to familiarize themselves with the different types of resources available, and to facilitate the decision-making process of selecting the most adequate for a specific purpose.

**Materials and Methods**

*Platform selection criteria*

To achieve our aims, we tested each platform and described our experience using the TXN gene as an example. Online open access platforms that allow the analysis of gene expression and the impact of this expression on patient survival from the TCGA original dataset, were considered in this analysis. Platform selection was performed using the keywords “gene expression database” or “gene expression platform” on Google and PubMed. Several original and review articles where these platforms are presented and discussed were also considered in our initial selection (1,7).

The inclusion criteria used to select the platforms demand that these allow TCGA dataset analysis, were to be free of charge, not require advanced training, de risco elevado apresentam taxas de sobrevida baixas em comparação com os que apresentam scores de baixo risco, o que contribui para determinar o score de risco sem utilizar outras variáveis e informações clínicas (4).

Neste trabalho, descrevemos e analisamos as plataformas cBioPortal, UCSC Xena, GEPIA e UALCAN, que fornecem dados sobre a análise da expressão génica em vários tipos de câncer, e respetivos estádios e análise da sobrevida dos doentes. Também foi incluída a OncoLnc, uma plataforma focada na análise da sobrevida.

A aplicabilidade das diferentes plataformas foi avaliada usando como exemplo a expressão do gene da Tiorredoxina-1 (TXN) no câncer de mama. A Tiorredoxina-1 é uma enzima que contribui para muitas características-chave de vários tipos de câncer, incluindo o aumento da proliferação e crescimento de células tumorais, a resistência à morte celular e o aumento da angiogênese (5, 6). A Tiorredoxina-1 está associada ao crescimento agressivo do tumor, resistência à terapêutica e diminuição da sobrevida do doente (5), podendo, portanto, constituir um potencial alvo terapêutico e um biomarcador da agressividade do câncer de mama.

Ao comparar esses diferentes recursos/plataformas online, pretendemos fornecer um ponto de partida para os futuros utilizadores se familiarizarem com os diferentes tipos de plataformas disponíveis e facilitar o processo de tomada de decisão ao selecionar a plataforma mais adequada para uma finalidade específica.

**Material e Métodos**

*Critérios de seleção das plataformas*

Para atingir os objetivos a que nos propomos, foi testada cada plataforma e descrita a experiência do utilizador tendo como exemplo o gene TXN. Foram consideradas plataformas on-line de open access (gratuitas) que permitem a análise da expressão génica e do impacto dessa expressão na sobrevida do doente a partir do conjunto de dados original do TCGA. Para selecionar as plataformas a analisar, utilizaram-se como palavras-chave “gene expression database” ou “gene expression platform” no Google e PubMed e também foram considerados, na nossa seleção inicial, vários artigos originais e de revisão onde este tipo de plataformas é apresentado e discutido (1,7).

Os critérios de inclusão utilizados para selecionar as plataformas exigem que estas permitam a análise do
not require a login, and to present gene expression for different types of cancers, gene expression by cancer stage, and allow patient survival analysis using Kaplan-Meyer curve and related statistics, i.e., Logrank tests.

The platforms cBioPortal (http://www.cbioportal.org/), UCSC Xena (https://xena.ucsc.edu/), GEPIA (http://gepia.cancer-pku.cn/) and UALCAN (http://ualcan.path.uab.edu/analysis.html) fulfilled all our inclusion criteria and were hence selected for this study. OncoLnc (http://www.oncolnc.org/) was also included since it provides an easy and quick method to obtain a list of gene expression and survival data for each patient (with TCGA ID).

Utility testing

The dataset named The Cancer Genome Atlas Breast Invasive Carcinoma (TCGA - BRAC) was used and TXN was selected as a model gene to test platform utility.

After analyzing articles from the developers of the different platforms and the tutorials available, each platform was tested to evaluate its utility. Information related with gene expression, gene expression by cancer stage and patient survival data, relevant statistic tests available, and specific limitations were collected for each platform. The selected analysis criteria were divided in three groups: i) general features of the platform; ii) gene expression analysis; iii) differential survival analysis. We considered various factors in each group, and for each factor, a classification of 1, 2 or 3 was given, as described in Table 1.

For “general features of the platform”, the classification of 1, 2 or 3 was ascribed, for user-friendliness, quality of the graphics generated by the platform, both in online and downloaded versions, and possibility of extracting the data attributes, respectively.

In the other two groups of characteristics, related with gene expression analysis and differential survival analysis, 1 corresponded to an unavailable analysis, 2 for cases where the analysis was available but with some limitations, and 3 when analysis was available.

 conjunto de dados do TCGA, sejam gratuitas, não requeiram treino avançado, não precisem de login, apresentem expressão génica para diferentes tipos de cancro e por estádio de cancro e permitam a análise da sobrevida do doente, usando a curva de Kaplan-Meyer e estatísticas relacionadas, ou seja, testes de Logrank.

A cBioPortal (http://www.cbioportal.org/), a UCSC Xena (https://xena.ucsc.edu/), a GEPIA (http://gepia.cancer-pku.cn/) e a UALCAN (http://ualcan.path.uab.edu/analysis.html) preencheram todos os nossos critérios de inclusão e, portanto, foram selecionados para este estudo. A OncoLnc (http://www.oncolnc.org/) também foi incluída, pois fornece um método fácil e rápido para obter uma lista dos doentes (com TCGA ID) com dados de expressão génica e sobrevida.

Teste de utilidade

Foi utilizado a base de dados The Cancer Genome Atlas Breast Invasive Carcinoma (TCGA - BRAC), tendo sido selecionado o TXN como gene modelo para testar a utilidade da plataforma.

Após analisar os artigos dos responsáveis/autores das plataformas, que as descreveram e disponibilizaram tutoriais, cada plataforma foi testada para avaliar a sua utilidade. Foram recolhidas informações relacionadas com a expressão génica, expressão génica por estádio do cancro e dados de sobrevida dos doentes, testes estatísticos relevantes disponíveis e limitações específicas de cada plataforma. Os critérios de análise selecionados foram divididos em três grupos: i) características gerais da plataforma; ii) análise de expressão génica; iii) análise diferencial da sobrevida. Em cada grupo foram considerados vários fatores e, para cada fator, foi dada uma classificação de 1, 2 ou 3, conforme descrito na Tabela 1.

As “características gerais da plataforma”, foram classificadas em 1, 2 ou 3, tendo sido atribuída esta classificação à facilidade de utilização, à qualidade dos gráficos gerados pela plataforma, nas versões on-line e após download, e à possibilidade de extração dos dados, respectivamente.

Nos outros dois grupos de características, relacionados com a análise de expressão génica e análise diferencial da sobrevida, 1 corresponde a uma análise que não está disponível; 2 aos casos em que a análise está disponível, mas com algumas limitações; e 3, quando a análise está disponível.
Table 1 - Score for the different criteria used to classify the platform.

| Criteria | Score |
|----------|-------|
| Feature-based comparison | 3 |
| Performance analysis | 2 |
| Cost-effectiveness | 1 |

Sofia Ramos et al
Results and Discussion

The analysis of gene expression according to tumor stage and the impact of under- and overexpression in patient survival is important to identify new targets that can be used as biomarkers for cancer development. Online platforms that can easily and quickly provide such analysis are essential, as many of the specialists working in this area lack the bioinformatics skills needed to process this type of complex data. Our study analysed and compared five open access platforms, in three different perspectives: gene expression quantification and analysis, gene expression by cancer stage and patient survival analysis.

Comparison of platform gene expression analysis possibilities

Although quantifying the expression levels of a gene set in one pathway does not provide information on which pathways were responsible for altered expression (8), nor does it reflect the effect of post-transductional modifications (9), it is the easiest method to infer the signalling activity of a pathway (8, 9). Therefore, analysing gene interactions in signalling pathways opens up a range of possibilities for studying a particular oncological pathology as it can identify its gene expression signature (10).

The cBioPortal provides data from The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC) including a set of studies that can be chosen for each cancer type. Data is presented in the form of graphics that quantify the percentage of samples or patients that present a high or low expression of a selected gene in a certain study using the Onco Query Language (OQL) tool (Figure 1A) (11). In this analysis, we can add heatmap tracks for more than one gene, using mRNA/miRNA expression Z-scores or mRNA expression z-scores, and clinical tracks, such as sample type, cancer type, diagnosis age, overall survival, tumor stage and others. All data can be downloaded easily in SVG, PNG or PDF format, as well as raw data. However, this platform does not allow the comparison between cancer and normal tissues, providing only data from tumors.

UCSC Xena also includes data from several studies and sources, and presents a visual worksheet, survival analysis, scatter plots, bar graphs, statistical tests and genomic signatures. Filtering allows to dynamically compare data from different genomic contexts and types, cancer type, diagnosis age, overall survival, tumor stage and others. All data can be downloaded easily in SVG, PNG or PDF format, as well as raw data. However, this platform does not allow the comparison between cancer and normal tissues, providing only data from tumors.

A UCSC Xena também inclui dados de vários estudos e fontes, e apresenta uma worksheet visual, análise da...
Figure 1 - Comparative analysis between the graphical output of expression of TXN. The expression levels of the TXN gene in the TCGA Breast Cancer (BRCA) dataset were analysed by the different platforms. (A) cBioportal, using ± 0.5 z-score threshold. By default, the platform selects only those cases where mRNA expression was analyzed. (B) Boxplot of TXN expression in UCSC Xena using sample type, which compares normal and tumor tissues. (C) Dot plot of TXN expression in normal and tumoral cells in GEPIA. Boxplot with comparison between normal and tumoral cells presented in GEPIA (D) and UALCAN (E).

Figura 1 - Análise comparativa entre o output gráfico da análise de expressão de TXN. Os níveis de expressão do gene TXN no conjunto de dados TCGA Breast Cancer (BRCA) foram analisados pelas diferentes plataformas. (A) Apresentação cBioportal, usando como limites o Z-score de ± 0,5. Por defeito, a plataforma seleciona apenas os casos em que a expressão de mRNA foi analisada. (B) Diagrama de caixa da expressão de TXN no UCSC Xena usando o tipo de amostra, que compara tecidos normais e tumoriais. (C) Gráfico de pontos da expressão de TXN em células normais e tumorais em GEPIA. Diagrama de caixa com a comparação entre células normais e tumorais apresentadas em GEPIA (D) e UALCAN (E).
subgroups (tumor vs normal, under-expression, overexpression, different cancer stages, different types of classification, etc.), using various hubs simultaneously (12). Gene expression of normal and tumor tissues is quantified using boxplots, with maximum, minimum, quartiles and median in a log2(norm_count+1) scale or norm_count+1 and Welch’s t-test associated (difference between means) (Figure 1B). All data can be downloaded and graphs are presented in PDF format.

GEPIA (Gene Expression Profiling Interactive Analysis) only includes data from two databases, TCGA and GTEx (Genotype-Tissue Expression Portal), that can be selected together or each on the same analysis. Multiple genes as well as gene expression in multiple types of cancer can be compared simultaneously. This platform allows differential analysis, to help defining the gene expression profile in tumor vs normal tissues, to perform correlational analysis, survival analysis, and detection of genes with a similar expression pattern (13). The results of differential analysis can be presented using a dot plot, with ANOVA or LIMMA differential methods, or boxplot, both using transcript per million (TPM) or log2 (TPM+1) units (Figures 1C and 1D). Although it is not possible to collect values of expression, plots in PDF format can be downloaded.

UALCAN provides a detailed analysis of gene expression using the TCGA database. It enables the identification of the main genes under- and overexpressed in different types of cancer and to analyze the relative expression of a gene in tumors and normal tissues. It also allows gene expression analysis in several subpopulations, based on cancer type, race, gender, age, stage of menopause (for female breast cancer or gynecological cancers such as endometrial or uterine), and the histological subtype (14). These data are presented using boxplots with TPM, maximum, minimum, quartiles and median values and t-test statistical comparison for normal vs tumor tissues (Figure 1E), that can be downloaded in SVG or PDF format. A file with the data produced is also available. Additionally, UALCAN presents a heatmap, listing genes with a differential expression for tumor and normal samples (14). Another very interesting and peculiar function of this platform is the gene scan, which allows the identification of a set of genes that show a differential expression and/or survival (14).
This scan is defined according to the class of genes (kinase coding and related genes, genes associated with metastasis, oncogenes and related, etc.).

These platforms offer different forms of presenting gene expression. Boxplot is a common type of presentation, used by UCSC Xena, GEPIA and UALCAN (Figure 1B, 1D and 1E). Oncoprint is used by cBioPortal (Figure 1A). GEPIA can also present data as dot plots (Figure 1C).

For comparison purposes, the same database (TCGA) and TCGA Breast Cancer (BRCA) dataset, was used to collect all the data that each platform provides (descriptive analysis, statistical tests).

The cBioPortal platform provides an easy way to identify the source of the study and allows the quantification of the proportion of under- and overexpression of each gene (Figure 1A). Although quartiles are not defined, to set a threshold z-score values can be used. Independently of the unit or statistical methodology used, it is easy to understand that TXN is globally overexpressed in breast cancer tissue.

Recovery of values for maximum, minimum, quartiles and median and statistical comparison between gene expression in normal and cancer tissue is only provided in UCSC Xena and UALCAN, which can be interesting to know if the differences between means have significance. GEPIA only estimates expression levels and does not provide the exact values.

Gene expression analysis by cancer stage: platform comparison

Tumor staging is important for prognosis because it allows categorization of the evolution of malignant disease of a patient at various time points after diagnosis, and is critical for the selection of the most suitable cancer therapy. It is used in cancer-related clinical research to define the targeted study population or as a stratification factor in randomized studies to reduce bias in comparisons between treatment groups (15).

tumorais e normais (14). Outra função muito interessante e peculiar dessa plataforma é o scan dos genes, que permite a identificação de um conjunto de genes que mostram a expressão diferencial e/ou a sobrevida (14). Este scan é definido de acordo com a classe de genes (codificação de cinases e genes relacionados, genes associados a metástases, oncogenes e afins, etc.).

Estas plataformas oferecem diferentes formas de apresentar a expressão génica. Os diagramas de caixa são um tipo comum de apresentação, utilizadas por UCSC Xena, GEPIA e UALCAN (Figura 1B, 1D e 1E). O Oncoprint é usado pelo cBioPortal (Figura 1A). O GEPIA também pode apresentar dados a partir de gráficos de pontos (Figura 1C).

Com o propósito de as comparar, o mesmo estudo TCGA Breast Cancer (BRCA) foi utilizados para recolher todos os dados que cada plataforma fornece (análise descritiva, testes estatísticos).

A plataforma cBioPortal fornece um modo fácil de identificar a fonte do estudo e permite a quantificação da proporção de sub e sobre-expressão de cada gene (Figura 1A). Embora os quartis não estejam definidos, é possível limitá-los usando os valores do z-score. Independentemente da unidade ou metodologia estatística usada, é fácil entender que TXN é globalmente mais expresso no tecido do cancro de mama.

A extração de valores de máximos, mínimos, quartis e medianas e a estatística comparativa entre a expressão génica em tecidos normais e tumorais é fornecida apenas na UCSC Xena e na UALCAN, o que pode ser interessante para perceber se as diferenças entre as médias são significativas. A GEPIA permite apenas estimar o valor da expressão, pois não fornece os valores exatos.

Análise da expressão génica por estádio de cancro: comparação das plataformas

O estadiamento do tumor é importante para o prognóstico, pois permite categorizar a evolução da doença maligna de um doente em vários momentos após o diagnóstico e é fundamental para a seleção da terapia mais adequada. É usado na pesquisa clínica relacionada com o cancro para definir a população alvo do estudo ou como fator de estratificação em estudos randomizados para reduzir o viés nas comparações entre os grupos de tratamento (15).
In cBioPortal, the expression level by tumor stage is presented in the form of a boxplot, which can be further explored from the raw data. Several forms of classification (Converted Stage, Tumor Stage, American Joint Committee on Cancer Tumor Stage Code, Neoplasm Disease Stage American Joint Committee on Cancer Code) as clinical attribute are available. In the example below (Figure 2A) the Neoplasm Disease Stage American Joint Committee on Cancer Code was used. It is also possible to select mRNA profile based on mRNA expression, microarray of mRNA expression, mRNA/miRNA expression z-Scores and others, depending on the type of analysis.

USCS Xena also offers several types of classification, depending on the selected study, since not all studies use the same classification system. It is also possible to collect maximum, minimum, quartiles and median for each stage and the corresponding one-way ANOVA statistical analysis (difference between means) (Figure 2B).

GEPIA presents pathological stages (I to IV stages) or “sub-stages”, based on the American Joint Committee on Cancer Code, and the F test with p value, to compare means between stages. It is also possible to select (or not) a log scale. Similar to gene expression, it is not possible to collect precise values of expression. Plots can be downloaded and compared visually (Figure 2C).

UALCAN provides gene expression analysis according to pathological stages (stages I to IV), with maximum, minimum, quartiles and median, and also important statistical comparisons between stages (Figure 2D).

There are different pathological stage classifications. The most common one is composed by four overall stages that result from the values according to the TNM classification system (primary Tumor, lymph Nodes, Metastases). However, not all studies used this classification. For example, the TCGA Firehose Legacy by cBioPortal presents others, such as Neoplasm Disease Stage American Joint Committee on Cancer Code, as shown in Figure 2A. For that reason, it is difficult to establish a comparison with other platforms. It is possible to quantify the number of patients in each stage using data, and it shows that Stages IIA, IIIA and IIB contain more patients. Comparing to the others platforms, the median of expression in GEPIA remains

Na cBioPortal, o nível de expressão por estádio do tumor é apresentado na forma de um diagrama de caixa, que pode ser mais explorado a partir dos dados brutos. É possível escolher várias formas de classificação (Converted Stage, Tumor Stage, American Joint Committee on Cancer Tumor Stage Code, Neoplasm Disease Stage American Joint Committee on Cancer Code), assim como atributos clínicos. No exemplo abaixo (Figura 2A), foi utilizada a classificação Neoplasm Disease Stage American Joint Committee on Cancer Code. Também é possível selecionar o perfil de mRNA com base na expressão z de mRNA, microarray da expressão de mRNA, z-scores de expressão de mRNA/miRNA entre outros, dependendo do tipo de análise.

A UCSC Xena também oferece vários tipos de classificação, dependendo do estudo selecionado, pois nem todos os estudos possuem o mesmo sistema de classificação. Também é possível recolher o máximo, o mínimo, os quartis e a mediana para cada estádio e a análise estatistica ANOVA unidirecional respetiva (diferença entre médias) (Figura 2B).

A GEPIA apresenta estádios patológicos (estádios I a IV) ou “subestádios”, com base no American Joint Committee on Cancer Code, e o teste F com o valor-p para comparar as médias entre os estádios. Também é possível selecionar ou não a escala logarítmica. Da mesma forma que a expressão génica, não é possível recolher valores precisos de expressão para os diferentes estádios. Os gráficos podem ser obtidos através de download e comparados visualmente (Figura 2C).

A UALCAN fornece a análise da expressão génica de acordo com os estádios patológicos (estádios I a IV), com máximo, mínimo, quartis e mediana, além de importantes comparações estatísticas entre os estádios (Figura 2D).

Existem diferentes classificações para os estádios patológicos. O mais comum é composto por quatro estádios gerais que resultam dos valores de acordo com o sistema de classificação TNM (Tumor primário, linfonodos, Metástases). No entanto, nem todos os estudos usaram essa classificação. Por exemplo, o TCGA Firehose Legacy, da cBioPortal, apresenta outros, como o Neoplasm Disease Stage American Joint Committee on Cancer Code, como pode ser visto na Figura 2A. Por esse motivo, é difícil estabelecer uma comparação com as outras plataformas. É possível quantificar o número de doentes em cada estágio utilizando os dados, o que mostra que os estádios IIA, IIIA e IIB contêm mais doentes. Comparando com as outras plataformas, a
Figure 2 - Comparative overview of TXN expression level analysis according to tumor stage. The levels of TXN expression by tumor stage were explored with cBioportal, using Neoplasm Disease Stage American Joint Committee on Cancer Code (A), UCSC Xena, using classification defined as Tumor_nature2012 (B), GEPIA, with F statistical (C) and UALCAN (D) using TPM unit.

Figura 2 - Visão geral comparativa da análise do nível de expressão de TXN de acordo com o estádio do tumor. Os níveis de expressão de TXN expressos nos diferentes estádios de tumor foram explorados com cBioportal, utilizando o Neoplasm Disease Stage American Joint Committee on Cancer Code (A) UCSC Xena, utilizando a classificação definida como Tumor_nature2012 (B), GEPIA, com estatística F (C) e apresentação da UALCAN (D) usando a unidade TPM.
unchanged, despite the fact that it includes more cases with overexpression until third grade, which decreased in last stage (Figure 2C). However, the number of cases in each tumor stage is not available, which does not allow a direct comparison of the results. Comparing UCSC Xena with UALCAN, the means show little variation, as GEPIA, but quartiles are well defined in both and follow the same trend (Figures 2B and 2D).

**Differential survival analysis based on the expression level of a specific gene: platform comparison**

The identification of new cancer biomarkers related to gene expression levels is important to estimate the survival in cancer patients. The under- or overexpression of specific genes has an increasing impact in prognosis and consequently in therapeutics selection (16). Survival and hazard functions can be calculated based on specific survival data for subgroups of patients according to the expression levels of selected genes (17). The survival function determines the probability of a patient to survive past a specific time. The hazard function provides the rate measured at a specific instant given that an individual has survived up to that time (17).

Survival data are usually presented as Kaplan-Meier curves, to allow visualization of survival over time. This type of curve assumes that each event occurs independently and each event time creates a new time interval. So, in each interval, the number of patients with an event is calculated along the time with the number at risk (18).

There are several statistical models commonly applied to survival data analysis. The Cox proportional hazards model is the most common. This model calculates the ratio of the hazard rates from two groups and if there is no difference in survival, the hazard ratio is equal to 1. However, if the ratio is less than 1 the effect is considered protective and if it is greater than 1, the effect is considered a risk factor. It provides also a statistical test with a corresponding P value for each variable of the model (18).

Considering the abovementioned parameters, the cBioPortal provides a comparison between gene expression levels. This portal can be used to categorize cases as under- and over gene expression, but it is necessary to define a z-score threshold and condition (Figure 3A was generated with a specific condition: EXP<-0.5 and >-0.5 to include cases that would approximately be included in upper and lower quartiles).

**Análise diferencial da sobrevida com base no nível de expressão de um gene específico: comparação de plataformas**

A identificação de novos biomarcadores de cancro associados com os níveis de expressão génica é importante para estimar a sobrevida em doentes com cancro. A sub ou sobre-expressão de genes específicos tem um impacto crescente no prognóstico e consequentemente na seleção terapêutica (16). As funções de sobrevida e risco podem ser calculadas com base em dados de sobrevida específicos para subgrupos de doentes, de acordo com os níveis de expressão dos genes selecionados (17). A função de sobrevida determina a probabilidade de um doente sobreviver após um tempo específico. A função de risco fornece a taxa medida num instante específico, dado que um indivíduo sobreviveu até aquele momento (17).

Os dados de sobrevida são geralmente apresentados como curvas de Kaplan-Meier, para permitir a visualização da sobrevida ao longo do tempo. Este tipo de curva assume que cada evento ocorre independentemente e cria um novo intervalo de tempo. Assim, em cada intervalo, o número de doentes com um evento é calculado ao longo do tempo com o número em risco (18).

Existem vários modelos estatísticos comumente aplicados à análise de dados de sobrevida. O modelo de risco proporcional de Cox é o mais comum. Este modelo calcula o hazard ratio de dois grupos e, se não houver diferença na sobrevida, o hazard ratio é igual a 1. No entanto, se o ratio for menor que 1, o efeito é considerado protetor e se for maior que 1, o efeito é considerado um fator de risco. Ele também fornece um teste estatístico com um valor-p correspondente para cada variável do modelo (18).

Considerando os parâmetros supramencionados, a cBioPortal fornece uma comparação entre os níveis de expressão génica. Este portal pode ser usado para categorizar os casos com sub e sobre-expressão de genes, mas é necessário definir um limite de z-score e
Figure 3 - Comparison between the typical survival analysis output from different platforms. Kaplan-Meier survival curves for high and low TXN expression levels were generated with (A) cBioportal, using EXP<-0.5 and EXP>-0.5. (B) UCSC Xena Kaplan-Meier using quartiles as cut-offs. (C) GEPIA with Logrank test and hazard ratio customized using quartiles as stratification. (D) UALCAN with upper quartile and below upper quartile division and (E) OncoLnc with Logrank p-value and quartiles cut-offs customization.

Figura 3 - Comparação entre o resultado típico do output da análise da sobrevida no conjunto de plataformas analisadas. As curvas de sobrevivência de Kaplan-Meier para níveis de expressão elevados e baixos de TXN foram geradas com (A) apresentação cBioportal, utilizando EXP<-0.5 e EXP>-0.5. (B) Curvas de sobrevida de Kaplan-Meier da UCSC Xena usando quartis como pontos de corte. (C) GEPIA com teste de Logrank e hazard ratio customizado usando quartis como estratificação. (D) UALCAN com divisão superior e inferior do quartil superior e (E) OncoLnc com valor-p do teste Logrank e customização dos pontos de corte por quartis.
It provides overall survival in months, in a range that can be changed for a more detailed analysis, Logrank test p-value and number of cases, number of deceased and median months of overall survival in each group.

UCSC Xena generates a Kaplan-Meier plot, that allows the division into two groups (under- and over gene expression), three groups (under-, over and medium expression) or quartiles (25% and 75%, as showed in Figure 3B). UCSC Xena adds Logrank test statistics to the survival analysis. The survival time is given in days, and can also be customized for a more detailed analysis. Note that it is common to have duplicate values (primary + relapse), and in this situation the platform gives an alert signal and these cases must be filtered.

GEPIA provides survival analysis using Logrank and Mantel-Cox test, which can include hazard ratio, and is based on the level of gene expression (under- and overexpression), with the option of overall or disease-free survival. It is also possible to analyse survival using the expression ratio of two genes, as well as providing a list of genes most associated with the survival of a selected type of cancer (13). For a differential survival analysis, median, quartile or custom cut-offs, confidence interval and temporal unit (days or months) can be selected (Figure 3C).

With UALCAN, it is possible to estimate the effect of the level of gene expression or clinical-pathological characteristics resources (race, gender, type of cancer, molecular signature) on the survival of patients and a Logrank statistical test. For survival analysis between under- and over-expression, the groups generated were pre-categorized in high expression (TPM values in the upper quartile) and low/medium expression (TPM values bellow upper quartile) (Figure 3D). For clinical parameters, the samples were first divided into the referred groups and then stratified (14).

Although OncoLnc presents some differences compared with the other selected platforms, it was included within our study because it can be used as a complement to the other platforms. OncoLnc main goal is to explore survival correlations, but can also provide clinical data associated with gene expression performed by TCGA (19). After selecting a gene of interest, OncoLnc provides overall survival using the Logrank and Mantel-Cox tests, with options to include hazard ratio and to customize survival analysis with median, quartile or custom cut-offs. The platform also generates Kaplan-Meier plots, allowing for a detailed analysis of survival distribution. When duplicate values are encountered, the platform alerts and filters these cases. For a differential analysis, options for median, quartile or custom cut-offs are provided, along with confidence intervals and temporal units. The platform is customizable to the researcher's needs.

A GEPIA foi gerada um gráfico de Kaplan-Meier, que permite a divisão em dois grupos (sub e sobre-expressão) e a opção de sobrevida global ou livre de doença. Também é possível analisar a sobrevida usando o hazard ratio, que é baseado no nível de expressão génica (sub e sobre-expressão), com opção de sobrevida global ou livre de doença. Além disso, é possível selecionar pontos de corte na mediana, quartis ou personalizados, assim como o intervalo de confiança e a unidade temporal (dias ou meses). Com o UALCAN, é possível estimar o efeito do nível de expressão génica e características clinico-patológicas nos doentes e no teste estatístico de Logrank. Para a análise da sobrevida entre sub e sobre-expressão, os grupos gerados foram pré-categorizados para a expressão elevada (valores de TPM no quartil superior) e expressão baixa/média (valores de TPM abaixo do quartil superior). Para os parâmetros clínicos, as amostras foram divididas primeiro nos grupos referidos e depois estratificadas.

Embora a OncoLnc apresente algumas diferenças em relação às outras plataformas selecionadas, foi incluída porque pode ser usada como um complemento. O objetivo principal da OncoLnc é explorar as correlações da sobrevida, mas também pode fornecer dados clínicos associados à expressão génica baseados no TCGA. Após selecionar um gene de interesse, a OncoLnc fornece uma análise de sobrevida utilizando o teste de Logrank e Mantel-Cox, que pode incluir o hazard ratio, e é baseado no nível de expressão génica (sub e sobre-expressão), com opção de sobrevida global ou livre de doença. Além disso, é possível analisar a sobrevida usando o ratio de expressão de dois genes, assim como o intervalo de confiança e a unidade temporal (dias ou meses). Com a UALCAN, é possível estimar o efeito do nível de expressão génica e características clinico-patológicas nos doentes e no teste estatístico de Logrank. Para a análise da sobrevida entre sub e sobre-expressão, os grupos gerados foram pré-categorizados para a expressão elevada (valores de TPM no quartil superior) e expressão baixa/média (valores de TPM abaixo do quartil superior). Para os parâmetros clínicos, as amostras foram divididas primeiro nos grupos referidos e depois estratificadas.
calculates Cox regression related to different types of cancer simultaneously and can include a Kaplan-Meier plot with customized percentiles (lower/upper) and data from patients divided for both groups (ID, days, status and expression). All data can be downloaded in a PDF format and Kaplan-Meier curves include Logrank related statistic that needs to be transcribed for PDF (Figure 3E).

Regarding survival analysis, cBioPortal parameters definition is less intuitive than others. UALCAN and UCSC Xena do not allow customized adjustments. Despite that, UCSC Xena permit the selection of quartiles, while UALCAN does not. With quartiles selection, survival analysis shows that there are not significant differences between groups in all platforms. However, the analysis from GEPIA presents a p value close to the significant level (Figure 3C), suggesting that patient’s survival tends to be different when TXN gene expression is high and low. GEPIA does not allow timeline customization; only UCSC Xena and cBioportal provide this option and in both cases the result was the same (p<0.05).

**Overall platform comparative analysis**

Figure 4 shows a comparative analysis applying the classification of 1 to 3 above described, in a graduation of colours, where the darker ones represent the better scores. Considering the general features of the platform, a blue graduation was chosen and in gene expression and differential survival analysis the option was green graduation. All the platforms presented are user-friendly and have tutorials that help with the most general questions. However, all have clear advantages above others and some limitations if a more comprehensive analysis is chosen. Table 2 summarizes the main highlights for each one.

CBioPortal does not have data from normal tissue samples, rendering the differential expression analysis limited to the comparison between tumor cases. Another disadvantage is the Kaplan-Meier survival curve presentation, because it is not possible to define the limits in quartiles, median or other, and the only statistical test is Logrank. On the other hand, cBioPortal allows the analysis of a wide range of cancer attributes (grade, stage, different classifications, type of cancer, survival, among others), while some other databases do not allow the analysis of mRNA levels, or do not have survival data.

calcula a regressão de Cox associada a diferentes tipos de cancro simultaneamente e pode incluir um gráfico de Kaplan-Meier com percentis customizados (inferior/superior) e dados de doentes divididos entre os dois grupos (ID, dias, status e expressão). Todos os dados podem ser obtidos através de download em formato PDF e as curvas de Kaplan-Meier incluem estatísticas Logrank que precisam de ser transcritas para o PDF (Figura 3E).

Em relação à análise da sobrevida, a definição dos parâmetros da cBioPortal é menos intuitiva do que as outras. A UALCAN e a UCSC Xena não permitem ajustes personalizados. Apesar disso, a UCSC Xena permite a seleção de quartis, enquanto a UALCAN não. Com a seleção dos quartis, a análise da sobrevida mostra que não há diferenças significativas entre os grupos em todas as plataformas. No entanto, a análise da GEPIA apresenta um valor-p próximo do nível significativo (Figura 3C), sugerindo que a sobrevida dos doentes tende a ser diferente quando a expressão do gene TXN é elevada ou baixa. A GEPIA não permite a personalização da linha do tempo, apenas a UCSC Xena e a cBioportal fornecem essa opção e em ambos os casos o resultado foi o mesmo (p <0,05).

**Análise comparativa global das plataformas**

A Figura 4 apresenta uma análise comparativa aplicando a classificação de 1 a 3 supra descrita, numa graduação de cores, onde as mais escusas representam as melhores pontuações. Foi escolhida uma graduação azul para as características gerais das plataformas, e para a expressão gênica e a análise diferencial da sobrevida, a opção foi a graduação verde. Todas as plataformas apresentadas são fáceis de usar e têm tutoriais que ajudam nas dúvidas mais comuns. No entanto, todas têm vantagens claras sobre as outras e algumas limitações se for feita uma análise mais abrangente. A Tabela 2 resume os principais pontos fortes e fracos de cada uma.

A cBioPortal não possui dados de amostras de tecido normal, tornando a análise de expressão diferencial limitada à comparação entre casos de tumor. Outra desvantagem é a apresentação da curva de sobrevida de Kaplan-Meier, porque não é possível definir os limites em quartis, mediana ou outros, e o único teste estatístico disponível é o Logrank. Por outro lado, a cBioPortal permite a análise de uma ampla gama de características do cancro (grau, estádio, diferentes classificações, tipo de cancro, sobrevida, entre outros), mas algumas bases de dados não permitem a análise dos níveis de mRNA ou não têm dados de sobrevida.
Table 2 - Main advantages and disadvantages of each platform.

| Platform  | Advantages | Disadvantages | Vantagens | Desvantagens |
|-----------|------------|---------------|-----------|--------------|
| cBioPortal | Includes clinical tracks, diagnosis age, different stage classifications, patient’s data. Several data sources | Only cancer tissue samples. Need to select patient/cases to analyze overall survival related to under and overexpression. | Inclui atributos clínicos, idade aquando do diagnóstico, diferentes classificações dos estádios, dados dos pacientes. Diversas fontes de dados. | Apenas amostras de tecido tumoral. Necessidade de selecionar pacientes/casos para analisar o sobrevivência global relacionada com a sub e sobre-expressão. |
| UCSC Xena | Several hubs and genomic contexts simultaneously. Various data sources and clinical tracks. Detailed results. | Absence of statistical significance in downloaded graphs. Requires some tutorials to understand. | Diversos hubs e contextos genômicos simultaneamente. Várias fontes de dados e atributos clínicos. Resultados detalhados. | Ausência de significância estatística no gráfico obtidos através de download. Requer algumas tutoriais para compreender. |
| GEPIA 2 | Overall survival with under and overexpression, Hazard ratio, median, quartile and custom cut-off. | Gene expression and statistical related not quantified. Unable to visualize patient/case data. | Sobrevivência global com sub e sobre-expressão, Hazard ratio, mediana, quartil e pontos de corte customizados. | Falta da quantificação da expressão génica e respetiva estatística. Impossibilidade de visualizar dados dos pacientes/casos. |
| UALCAN | Data from boxplot detailed (min, max, quartiles). Statistical comparisons (TtN and NxS). | Cut-off in survival analysis is fixed. Unable to obtain patient/case data. | Dados detalhados nos diagramas de caixa (min, max, quartis). Comparações estatísticas (TtN and NxS). | Pontos de corte na análise da sobrevida são fixos. Impossibilidade de obter dados dos pacientes/casos. |
| OncoLnc | Patient data (gene expression, status, survival days). | Gene expression by stage not presented. | Dados dos pacientes (expressão génica, status, dias de sobrevivência). | Não apresenta expressão génica por estádio. |

Figure 4 - Global comparative analysis of the platforms. General features related with gene expression and differential survival analysis were considered. A colour gradient was used to classify platforms using three points scale. More intense colour reflects better classifications. General features (blue) were classified as “good”, “average” or “poor” in terms of user-friendliness, quality of the graphs (presented online and downloaded versions) and data extraction availability/flexibility. Gene expression and differential survival analysis (green) were classified as “available”, “available with limitations” or as “non-available”. For gene expression analysis, comparison between normal and cancer tissue, availability of cancer stage analysis and different types of output (different types of graphical outputs, descriptive and statistical analysis) were considered. For survival analysis, the customization of Kaplan-Meier parameters, test and p-value of Logrank and hazard ratio were analysed.

Figura 4 - Análise comparativa global das plataformas. Foram consideradas as características gerais das plataformas, as da expressão génica e as da análise diferencial da sobrevivência. Um gradiente de cores foi usado para classificar as plataformas usando uma escala de três níveis, com a cor mais intensa a refletir melhores classificações. As características gerais (azul) foram classificadas como “bom”, “médio” ou “mau” em termos de facilidade de uso, qualidade dos gráficos (apresentados nas versões on-line e por download) e disponibilidade/flexibilidade de extração de dados. A expressão génica e a análise diferencial da sobrevivência (verde) foram classificadas como "disponíveis", “disponíveis com limitações” ou “não disponíveis”. Para a análise da expressão génica, considerou-se a comparação entre tecido normal e tumoral, disponibilidade da análise do estádio do cancro e diferentes tipos de outputs (diferentes tipos de outputs de gráficos, análise descritiva e estatística). Para a análise da sobrevida, foram analisados a customização dos parâmetros de Kaplan-Meier, o teste e valor-p de Logrank e hazard ratio.
UCSC Xena is the least intuitive platform, requiring some tutorials and tests for the user to take full advantage of its potential. Nonetheless, this platform includes more studies, data sources, classifications and clinical tracks. Note that not all studies include the same type of information. The possibility to examine several hubs and genomic contexts simultaneously allows the analysis of the expression of one or more genes in parallel or to observe various genomic characteristics simultaneously. The boxplots offer detailed information (with maximum, minimum, quartiles) and plots are easy to download, however, the related statistical information needs to be transcribed.

The main limitation of GEPIA is that it does not allow the acquisition of the gene expression level values. The comparison between tumor and normal gene expression is only visual. Some information, such as the histological subtype of cancer, patient data or the existence of metastasis are not available. However, in an analysis focused on the comparison between gene expression levels, it can be a tool to quickly and easily select genes of interest. The main advantage of this platform, aside from the great graphical quality of the website, is their more complete survival analysis, which adds hazard ratio to traditional Logrank test and the possibility to customize cut-offs of high and low expression.

In contrast, UALCAN offers a complete quantification of gene expression, with detailed boxplot (maximum, minimum, quartiles) and statistical comparison between normal and tumor tissues, normal and tumor tissues by stage and between stages. It also provides the correlation between genes and several survival curves, which allow to compare different histological subtypes of cancer, ages, ethnicities, etc. However, the survival analysis has a fixed cut-off and statistics only shows Logrank’s p-value, which can be a limitative factor in several researches.

Finally, OncoLnc simplifies querying TCGA data by showing some patient data, such as gene expression, status and survival over time and it allows the user to choose the cut-offs (lower and upper percentile of gene expression) that are best for each specific case. However, because it is a platform focused on survival, gene expression analysis is not presented.

A UCSC Xena é a plataforma menos intuitiva, exigindo alguns tutoriais e testes para que o utilizador tenha a possibilidade de aproveitar ao máximo o potencial da plataforma. No entanto, esta plataforma inclui mais estudos, fontes de dados, classificações e atributos clínicos. De notar que nem todos os estudos incluem o mesmo tipo de informação. A possibilidade de examinar vários hubs e contextos genómicos simultaneamente permite a análise da expressão de um ou mais genes em paralelo ou observar várias características genómicas simultaneamente. Os diagramas de caixa oferecem informações detalhadas (com máximo, mínimo, quartis) e os gráficos são fáceis de obter através de download, no entanto, é necessário transcrever as respetivas informações estatísticas.

A principal limitação da GEPIA é não permitir obter os valores do nível de expressão génica. Só é possível comparar visualmente a expressão do gene nos tecidos tumoral e normal. Também não é possível extrair algumas informações, como o subtipo histológico do cancro, dados do doente ou a existência de metástases. No entanto, numa análise focada na comparação entre os níveis de expressão génica, pode ser uma ferramenta para selecionar rápida e facilmente genes de interesse. A principal vantagem desta plataforma, além da excelente qualidade gráfica do website, é a análise mais completa da sobreviva, que adiciona o hazard ratio ao tradicional teste de Logrank e a possibilidade de personalizar pontos de corte com a expressão elevada e baixa.

Em contrapartida, a UALCAN oferece uma quantificação completa da expressão génica, com um diagrama de caixa detalhado (máximo, mínimo, quartis) e comparação estatística entre tecidos normais e tumorais, tecidos normais e tumorais por estádio e entre estádios. Também fornece a correlação entre genes e várias curvas de sobrevia, que permitem comparar diferentes subtipos histológicos de cancro, idades, etnias, etc. No entanto, a análise de sobrevia possui um ponto de corte fixo e as estatísticas mostram apenas o valor-p de Logrank, características que precisam ser mais completas em muitas pesquisas.

Finalmente, a OncoLnc simplifica a consulta de dados do TCGA, apresentando alguns dados de doentes, como expressão génica, status e dias de sobreviva, e permite escolher os pontos de corte (percentil inferior e superior da expressão génica) que são melhores para cada caso específico. Por ser uma plataforma focada na sobrevia, a análise da expressão génica não é apresentada.
Conclusion

All the platforms analysed allow an easy access to the data and explore individual and multiple gene expression patterns. Some platforms relate gene expression to disease stages and analyse survival associated with the under and overexpression of a given gene. Some platforms also allow the user to filter data by selecting specific attributes, such as the type of cancer, molecular alteration, recurrence, histological aspects, etc. The data can be visualized through graphs, showing the distribution of the expression of a given gene or set of genes, also allowing the quantification of expression levels and establishing statistical relationships.

The diversity of available information contributes to the study of the complex mechanisms involved in the pathogenesis of cancer in a very relevant way, since the investigative approaches can be multiple. The choice of a given platform must take in consideration the type of information sought for each purpose. Additionally, if one platform does not provide all the information/resources needed, it is possible to combine data analysis from more than one platform. In this case, it is fundamental to select the same database as source.

Authors Contributions Statement

NS, conceptualization and study design; SR, platform experimentation; SR, data analysis and drafting; NS and ASF, editing and reviewing; SR and NS, figures and graphics; NS and ASF, supervision and final writing.

Funding

Research developed with funding from Universidade Lusófona/ILIND (Grant Program FIPID 2019/2020).

Conflict of Interests

The authors declare there are no financial and/or personal relationships that could present a potential conflict of interests (or what applies for each case).

Conclusões

Todas as plataformas analisadas são intuitivas, permitem um fácil acesso aos dados e exploram padrões individuais e múltiplos de expressão gênica. Alguns relacionam a expressão gênica a estádios da doença e analisam a sobrevida associada a sub e sobre-expressão de um determinado gene. Algumas plataformas também permitem ao utilizador filtrar os dados selecionando atributos específicos, como o tipo de cancro, alteração molecular, recorrência, aspetos histológicos, etc. Os dados podem ser visualizados através de gráficos, mostrando a distribuição da expressão de um determinado gene ou conjunto de genes, permitindo também a quantificação dos níveis de expressão e o estabelecimento de relações estatísticas.

A diversidade de informações disponíveis contribui de maneira muito relevante para o estudo dos complexos mecanismos envolvidos na patogénese do cancro, uma vez que as abordagens investigativas podem ser múltiplas. A escolha de uma determinada plataforma deve levar em consideração o tipo de informação necessária para cada finalidade. Adicionalmente, se uma plataforma não fornecer todas as informações/recursos necessários, é possível combinar a análise de dados de mais de uma plataforma. Nesse caso, é fundamental selecionar a mesma fonte do banco de dados.

Declaração sobre as contribuições do autor

NS, conceção e desenho do estudo; SR, experimentação das plataformas; SR, análise de dados e redação; NS e ASF, edição e revisão; NS e SR, figuras e gráficos; ASF e NS, supervisão e redação final.

Financiamento

Este trabalho foi financiado pelo Grant Program FIPID 2019/2020 (Universidade Lusófona/ILIND).

Conflito de Interesses

Os autores declaram que não há relações financeiras e/ou pessoais que possam representar um potencial conflito de interesses (ou o que for aplicável).
References / Referências

1. Pavlopoulou A, Spandidos DA, Michalopoulos I. Human cancer databases (Review). (2015) Oncology Reports, 33, 3-18. https://doi.org/10.3892/or.2014.3579
2. Wang Z, Jensen MA, Zenklusen JC. A Practical Guide to The Cancer Genome Atlas (TCGA). (2016) Methods in Molecular Biology, 418, 111-41. https://doi.org/10.1007/978-1-4939-3575-9_6
3. Hussain SA, Palmer DH, Syn WK, Sacco JJ, Greensmith RM, Elmetwali T, et al. Gene expression profiling in bladder cancer identifies potential therapeutic targets. (2017) International Journal of Oncology, 50(4), 1147-59. https://doi.org/10.3892/ijo.2017.3893
4. Liu L, Chen Z, Shi W, Liu H, Pang W. Breast cancer survival prediction using seven prognostic biomarker genes. (2019) Oncology Letters, 18, 2907-16. https://doi.org/10.3892/ol.2019.10635
5. Powis G, Kirkpatrick DL. Thioredoxin signaling as a target for cancer therapy. (2007) Current Opinion in Pharmacology, 7(4), 392-7. https://doi.org/10.1016/j.coph.2007.04.003
6. Arnér ES, Holmgren A. The thioredoxin system in cancer. (2006) Seminars in Cancer Biology, 16(6), 420-6. https://doi.org/10.1016/j.semcancer.2006.10.009
7. Yang Y, Dong X, Xie B, Ding N, Chen J, Li Y, et al. Databases and Web Tools for Cancer Genomics Study. (2015) Genomics, Proteomics Bioinformatics, 13, 46-50. https://doi.org/10.1016/j.gpb.2015.01.005
8. Parikh JR, Klinger B, Xia Y, Marto JA, Blüthgen N. Discovering causal signaling pathways through gene-expression patterns. (2010) Nucleic Acids Research, 38, W109-W17. https://doi.org/10.1093/nar/gkq424
9. Schubert M, Klinger B, Klünemann M, Sieber A, Uhlitz F, Sauer S, et al. Perturbation-response genes reveal signaling footprints in cancer gene expression. (2018) Nature communications, 9(1), 20. https://doi.org/10.1038/s41467-017-02391-6
10. Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim JS, et al. A novel signaling pathways impact analysis. (2009) Bioinformatics, 25(1), 75-82. https://doi.org/10.1093/bioinformatics/btn577
11. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. (2013) Science Signaling, 6(269), pl1. https://doi.org/10.1126/scisignal.2004088
12. Goldman M, Craft B, Kamath A, Brooks A, Zhu J, Haussler D. The UCSC Xena Platform for cancer genomics data visualization and interpretation. (2019) bioRxiv. https://doi.org/10.1101/326470
13. Tang Z, Li C, Kang B, Gao G, Li C, Zhang Z. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. (2017) Nucleic Acids Research, 45(W1), 98-102. https://doi.org/10.1093/nar/gkw247
14. Chandrashekar DS, Bashel B, Balasubramanya SAH, Creighton CJ, Ponce-Rodriguez I, Chakravarthi BV, et al. UALCAN: A Portal for Facilitating Tumor Subgroup Gene Expression and Survival Analyses. (2017) Neoplasia, 19(8), 649-58. https://doi.org/10.1016/j.neo.2017.05.002
15. Hu Z-D, Zhou Z-R, Qian S. How to analyze tumor stage data in clinical research. (2015) Journal of Thoracic Disease, 7(4), 566-75. https://doi.org/10.3978/j.issn.2072-1439
16. Mirza Y, Ali SM, Awan MS, Idress R, Naeem S, Zahid N, et al. Overexpression of EGFR in Oral Premalignant Lesions and OSCC and Its Impact on Survival and Recurrence. (2018) Oncomedicine, 3, 28-36. https://doi.org/10.7150/oncm.22614
17. Clark TG, Bradburn MJ, Love SB, Altman DG. Survival analysis part I: basic concepts and first analyses. (2003) British Journal of Cancer, 89(2), 232-8. https://doi.org/10.1038/sj.bjc.6601118
18. Pugh SL. Essence of survival analysis. (2016) Neuro-Oncology Practice, 4(2), 77-81. https://doi.org/10.1093/neo/npw012
19. Anaya J. OncoLnc: linking TCGA survival data to mRNAs, miRNAs, and lncRNAs. (2016) Peer Journal Computer Science, 2, e67 https://doi.org/10.7717/peerj-cs.67