Computational and Structural Biotechnology Journal 19 (2021) 5149–5159

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Data article

Cancer DEIso: An integrative analysis platform for investigating differentially expressed gene-level and isoform-level human cancer markers

Tzu-Hsien Yang [a], Yu-Hsuan Chiang [b], Sheng-Cian Shiue [b], Po-Heng Lin [b], Ya-Chiao Yang [a,1], Kai-Chi Tu [a,1], Yan-Yuan Tseng [c], Joseph T. Tseng [d,*], Wei-Sheng Wu [b,*]

[a] Department of Information Management, National University of Kaohsiung, Kaohsiung University Rd, 811 Kaohsiung, Taiwan
[b] Department of Electrical Engineering, National Cheng Kung University, University Road, 701 Tainan, Taiwan
[c] Center for Molecular Medicine and Genetics, Wayne State University, School of Medicine, Detroit, MI, USA
[d] Department of Biotechnology and Bioindustry Sciences, National Cheng Kung University, University Road, 701 Tainan, Taiwan

ARTICLE INFO

Article history:
Received 10 June 2021
Received in revised form 6 September 2021
Accepted 6 September 2021
Available online 08 September 2021

Keywords:
TCGA
Cancer
Cancer stage differential
isoform expression analysis

ABSTRACT

Transcript isoforms regulated by alternative splicing can substantially impact carcinogenesis, leading to a need to obtain clues for both gene differential expression and malfunctions of isoform distributions in cancer studies. The Cancer Genome Atlas (TCGA) project was launched in 2008 to collect cancer-related genome mutation raw data from the population. While many repositories tried to add insights into the raw data in TCGA, no existing database provides both comprehensive gene-level and isoform-level cancer stage marker investigation and survival analysis. We constructed Cancer DEIso to facilitate in-depth analyses for both gene-level and isoform-level human cancer studies. Patient RNA-seq data, sample sheets, patient clinical data, and human genome datasets were collected and processed in Cancer DEIso. And four functions to search differentially expressed genes/isoforms between cancer stages were implemented: (i) Search potential gene/isoform markers for a specified cancer type and its two stages; (ii) Search potentially induced cancer types and stages for a gene/isoform; (iii) Expression survival analysis on a given gene/isoform for some cancer; (iv) Gene/isoform stage expression comparison visualization. As an example, we demonstrate that Cancer DEIso can indicate potential colorectal cancer isoform diagnostic markers that are not easily detected when only gene-level expressions are considered. Cancer DEIso is available at http://cosbi4.ee.ncku.edu.tw/DEIso/.

1. Introduction

Alternative splicing is a vital way to extend the protein diversity for genes with multiple exons. It is estimated that around 90% of human genes undergo alternative splicing to produce distinct transcript isoforms [1]. A transcribed pre-mRNA binds with splicing factors to generate different transcript isoforms through selective exon skipping or intron retaining [2]. Because of this critical mechanism, aberrant spicing patterns or defects in the spicing factors might lead to human diseases and cancer [3–5]. For example, the recent genomic expression analysis of the two isoforms of K-Ras, or K-Ras4A and K-Ras4B, indicated that the abnormal isoform-level expression might be associated with the initiation and progression of lung adenocarcinoma [6]. Hence besides comparing the gene-level cancer markers, investigating the malfunction of isoform distributions can also provide clues for human cancer research.

The Cancer Genome Atlas (TCGA) project [7] was launched in 2008 to enhance the collection of cancer genomic and transcriptomic data. Later, the Genomic Data Commons (GDC) Data Portal [8] was built to guard patient privacy and facilitate the data retrieval process. TCGA is a funded and coordinated project that aims to gather major cancer-related genome mutations from the human population. High-throughput sequencing and patient clinical data for over 30 different human cancer types are deposited in TCGA [9,10]. The richness of the cancer data stored in TCGA makes it

* Corresponding authors.
E-mail addresses: thyangza1025@nuk.edu.tw (T.-H. Yang), theedward285@gmail.com (Y.-H. Chiang), t50504t@gmail.com (S.-C. Shiue), bb932156bb@gmail.com (P.-H. Lin), a1073314@mail.nuk.edu.tw (Y.-C. Yang), a1073348@mail.nuk.edu.tw (K.-C. Tu), ytseng@wayne.edu (Y.-Y. Tseng), tctseng@mail.ncku.edu.tw (J.T. Tseng), wessonwu@mail.ncku.edu.tw (W.-S. Wu).
[1] These authors contributed equally.

https://doi.org/10.1016/j.csbj.2021.09.005

possible for researchers to catalog and unearth the prognostic and diagnostic signals for cancer progression detection and treatment.

While a tremendous number of cancer whole transcriptome microarray and sequencing data have been deposited in TCGA, it still requires advanced data processing and mining to obtain a comprehensive understanding of human cancer [11]. Cancer in different tissues may result from various genetic or regulation disorders [12]. If one can figure out the potential markers that bear differential expressions among normal people and cancer patients, it will be beneficial for designing precise tissue-specific clinical experiments. For cancer prognostic and diagnostic biomarker investigation, three major in-depth search functions are expected to be implemented: (1) Cancer gene/isoform candidate marker identification for different cancer types and stages; (2) Comprehensive stage analysis of the gene-level or isoform-level expression change for the biomarker; (3) Survival analysis on the expression of the biomarker. Many efforts have been made to deal with these demands in the past decade. Cancer RNA-seq Nexus [13] was first constructed to help researchers study the gene-level expression difference between normal and cancer cells. Later, different databases (KM-express [14], OncoLnc [15], and UALCAN [16]) that facilitate clinical survival analysis on gene-level expression were developed to supplement the analysis. UALCAN further integrated patient clinical data to provide a comprehensive gene-level differential comparison tool for 31 types of human cancer. Nevertheless, these databases do not provide the possibility of isoform-level analysis. ISOexpresso [17] and GEPIA2 [18] allow cancer isoform-level expression analysis. However, ISOexpresso does not offer the opportunity for clinical survival analysis. And these two platforms both lack cancer stage information and do not provide the ability to perform stage differential expression analysis. Due to the large data volume and tedious data processing steps, no one such database can provide all these three important functionalities in both gene level and isoform level.

To overcome the problem and provide an in-depth analysis tool for human cancer studies, we constructed the Cancer DEIso (Cancer differentially expressed isoform and gene) database based on advanced integration of TCGA data. In the Cancer DEIso database, patient transcriptome RNA-seq data, patient sample sheets, patient clinical data, and human genome information datasets were collected and processed. Four investigation functions were implemented in the database. First, users can glimpse the gene list or transcript isoform list containing items differentially expressed between two stages to select possible stage markers for a specified cancer type. Second, users can obtain the cancer lists potentially induced by the differential expressions of a given gene or transcript isoform. A detailed comparison result page for a selected gene/isoform and its corresponding isoforms/gene is presented for the above two investigation functions. Third, users can perform FPKM expression survival analysis for a chosen gene or transcript isoform for the query cancer type and stage. Fourth, users can visualize the FPKM comparison of a given gene/isoform between different cancer stages and cancer types via the boxplot comparison. All analysis results within Cancer DEIso can be downloaded for downstream experimental designs. We demonstrate that the database can successfully indicate potential colorectal cancer transcript isoform markers that are not easily detected when only gene-level expressions are considered. The Cancer DEIso database is available online at http://cosbi4.ee.ncku.edu.tw/DEIso/.

2. Construction and content

2.1. Data collection and processing

The construction of the Cancer DEIso database can be divided into four different steps (See Fig. 1). In the first step, the massive RNA-seq data were downloaded from the GDC Data Portal. And then, these sequencing data were categorized by the diagnosed cancer stages. In the second step, the aligned reads for each patient sample in the categorized cancer stages were mapped to the hg38 human transcriptome to compute the gene-level and isoform-level expressions by Cufflinks. In the third step, the gene-level and isoform-level differential expression between different cancer stages were computed by the Cuffdiff tool. Finally, the days-to-death information was collected from the patient clinical data and underwent the survival analysis for different cancer types and stages. Details of each stage are depicted in the following subsections.

2.1.1. Data acquisition

In Cancer DEIso, we collected three genres of data from TCGA [9] to perform the gene-level and isoform-level differential analysis between stages of each collected cancer type: patient RNA-sequencing data, patient sample sheets, and patient clinical data. We downloaded the patient RNA-sequencing sample results in the aligned BAM format from the GDC Data Portal [19]. Samples from the "Primary Tumor" and "Solid Tissue Normal" categories were selected. Then the patient sample sheets and clinical data were also downloaded from the GDC Data Portal. The linkages between RNA-sequencing samples and patient clinical data are listed in the patient sample sheets. And in the clinical data, each patient's demographic information (such as height, weight, sex, and race), the diagnosed cancer stage, diagnosed morphology, primary diagnosis code, and the treatments are recorded. Using the clinical data, we grouped the RNA-seq samples into five categories (Normal, Stage I, Stage II, Stage III, and Stage IV) and eliminated the samples tagged with "Not Report" or "Not Clinical." We also adopted the "primary_diagnosis" and "morphology" columns in Cancer DEIso to provide basic clinical information for the samples. The cancer types and corresponding patient sample numbers gathered in Cancer DEIso are listed in Table 1. Notice that the sample numbers collected in this database refer to the RNA-sequencing sample numbers instead of patient numbers. And in Cancer DEIso, cancer types with no available stage information in the collected samples were not included in the analysis. To perform the transcriptome expression analysis, we adopted the human hg38 reference genome and transcriptome (Refseq GRCh38 Dec. 2013 assembly) from the UCSC Genome Browser [20]. In this database, 56,892 transcript isoforms for 26,380 human genes were collected and analyzed.

2.1.2. Patient transcriptome expression analysis

High-throughput mRNA sequencing (RNA-seq) using the cDNA technology can reveal the abundance of alternative splice isoforms in human transcriptome for different patients and cell conditions with at least comparable accuracy to microarrays [21,22]. Based on the read count density of the RNA-sequencing results for different transcripts, we can obtain the potential cancer markers that bear differential expressions between normal and cancer samples. We downloaded the RNA-seq data in the aligned BAM format from the GDC Data Portal for our marker expression analysis. We used the Cufflinks [23] RNA-seq transcript expression analysis tool to infer the splicing structure of each gene and calculate the expression level of each transcript in the unit of fragments per kilobase per million mapped fragments (FPKM). The FPKM metric incorporates both the transcript length normalization and machine run yield bias correction [24]. Default parameters were set when applying the Cufflinks tool. The hg38 human reference genome and transcriptome were used in the expression level calculation. Since FPKM is directly proportional to the transcript and gene expression abundance [23], we summed up the FPKM values of different splice isoforms and duplicates of a given gene as the

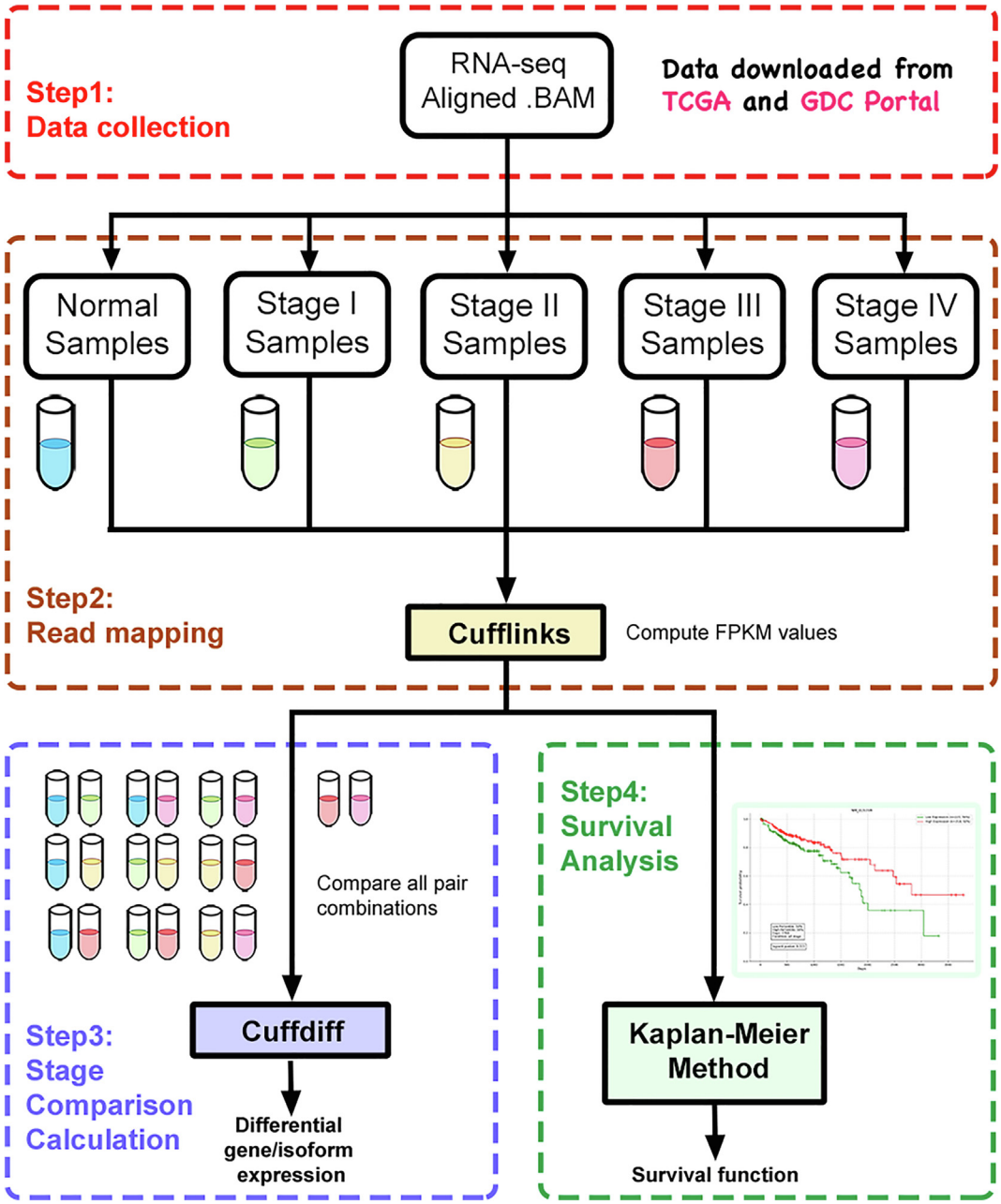

Fig. 1. Overview of the Cancer DEIso database construction. The construction of Cancer DEIso can be divided into four steps: (1) Step 1: Data collection. RNA-seq, patient sample sheets, and clinical data for all patients were collected from the GDC Data Portal. (2) Step 2: Read mapping. The downloaded aligned BAM files were processed and mapped to the human hg38 transcriptome by Cufflinks. (3) Step 3: Stage comparison calculation. The gene/isoform-level differential expressions between cancer stages were computed by Cuffdiff. (4) Step 4: Survival analysis. The survival curves are generated based on the high FPKM value group and the low FPKM value group for the selected cancer type and stage.

expression level of the gene. All RNA-seq samples of the normal tissues and cancer cells collected from different patients were processed using this same procedure.

2.1.3. Differential analysis for normal and cancer samples

Understanding the differentially expressed genes or isoforms across normal and cancer cells can help identify the markers for cancer diagnosis. Hence after calculating the FPKM values of each gene/transcript isoform in all RNA-seq samples, we performed a comprehensive analysis of the FPKM differential expressions between different cancer stages. Although the Kallisto-Sleuth [25,26] pipeline was reported to provide rapid transcript quantification while achieving near-optimal differential quantification performance among recent analysis pipelines [27], we found out that it was not well fitted for the TCGA aligned BAM files. The Cufflinks-Cuffdiff [23] data analysis procedure can provide more marker investigation results that bear significant stage-differential expressions. Hence the Cufflinks-Cuffdiff data analysis pipeline was adopted in Cancer DEIso. The Cuffdiff tool calculates the quantitative differential expression between normal and cancer cells. To speed up the computing process, we first used the Cuffquant [23] tool to save the FPKM expression values of each gene and transcript isoform in the binary .cxb format. For a comprehensive comparison of differential expression analysis amidst different cancer stages, we fed the .cxb files of all the combination of 2 stages (10 combinations between Normal, Stage I, Stage II, Stage

Table 1
Data statistics for cancer types and samples included in Cancer DEIso.

Cancer Types	Project	No. of Samples				
		Normal	Stage I	Stage II	Stage III	Stage IV
Adrenal Gland (Adrenocortical Carcinoma) cancer	TCGA-ACC	–	9	37	16	15
Bile Duct cancer	TCGA-CHOL	9	19	9	–	7
Bladder cancer	TCGA-BLCA	19	4	130	142	136
Breast cancer	TCGA-BRCA	113	182	627	248	20
Colon cancer	TCGA-COAD	41	81	186	132	66
Esophagus cancer	TCGA-ESCA	11	16	69	49	8
Eye cancer	TCGA-UVM	–	39	36	4	–
Head and Neck cancer	TCGA-HNSC	44	25	70	78	259
Kidney (Kidney Chromophobe) cancer	TCGA-KICH	24	20	25	14	6
Kidney (Kidney Renal Clear Cell Carcinoma) cancer	TCGA-KIRC	72	271	59	123	82
Kidney (Kidney Renal Papillary Cell Carcinoma) cancer	TCGA-KIRP	32	172	21	51	15
Liver cancer	TCGA-LIHC	50	171	86	85	5
Lung (Lung Adenocarcinoma) cancer	TCGA-LUAD	59	292	124	84	26
Lung (Lung Squamous Cell Carcinoma) cancer	TCGA-LUSC	49	245	162	84	7
Pancreas cancer	TCGA-PAAD	4	21	146	3	4
Pleura cancer	TCGA-MESO	–	10	16	44	16
Rectum cancer	TCGA-READ	10	30	51	51	24
Skin cancer	TCGA-SKCM	–	2	66	27	3
Stomach cancer	TCGA-STAD	32	53	111	150	38
Testis cancer	TCGA-TGCT	–	101	12	14	–
Thyroid cancer	TCGA-THCA	58	280	52	112	55

III, and Stage IV) into the Cuffdiff tool. Cuffdiff computes the $\log_2$ fold change of each gene and transcript isoform between two conditions. We used the default parameters when running the Cuffdiff software. Sometimes considering only the fold change may be misled by small quantity noises [28]. The *q*-values of differential fold changes computed by Cuffdiff were incorporated to help users identify statistically differentially expressed genes/isoforms. And we additionally applied three tests to help evaluate the statistical significance of the differential expression levels of a given gene or transcript isoform between the samples in one condition and samples in a second condition: the Kolmogorov–Smirnov (KS) test [29], the parametric one-tailed *t*-test [30], and the non-parametric one-tailed *U*-test [31–33]. The Cuffdiff-computed FPKM values were used to calibrate the sequencing data distributions when applying these tests. In brief, the KS test/t-test/U-test compares if the calibrated FPKM value distribution/average/median of patient samples in one condition is larger than the corresponding statistic of samples in the other condition, respectively. To control the multiple hypotheses bias, the test *p*-values among all genes/transcripts were calibrated using the FDR-control procedure [34]. We also consider if the isoform usage within a gene alters among different cancer stages [35]. As in IsoformSwitchAnalyzeR [36], the differential isoform usage is estimated by the isoform fractions (IFs) and the differential IF (dIF) values based on the Cuffdiff-calibrated FPKM values. For each tumor stage, the IF value of a transcript is defined as the ratio of the isoform FPKM value to the value of its related gene. And the dIF value of an isoform between any two stages is computed by $IF_2 - IF_1$ for Condition II and Condition I. Statistical tests, including the one-sided *t*-test, the one-sided *U*-test, and the one-sided KS-test, were performed to compute the *p*-values of the dIF values. In the one-side *t*-test/*U*-test/KS-test, we computed the dIF *p*-value by comparing if the average/median/distribution of IF_1 for Condition II is larger than the corresponding statistic of IF_2 for Condition I. These *p*-values were also FDR-corrected for the multiple hypotheses bias. Based on the IFs and dIFs between two stages for the specified isoform, we can further check if the differential isoform expression of a given transcript results from differential isoform usage and isoform switching.

2.1.4. Survival analysis

Survival rates are the essential indicator for cancer detection and treatment efficacy [37,38] and serve as an excellent clinical evaluation of the identified potential cancer marker. We implemented the essential survival analysis tool in the constructed database as well. The survival time data were gathered from the clinical data of the cancer patients. The lifetime unit used in the survival analysis is "day". If the "vital_status" is "Alive", the column "days_to_last_follow_up" is used as the survival time. And if the patient is not alive ("vital_status" = "Dead"), the survival time is adopted from the "days_to_death" column. Five types of samples (all cancer stages, Stage I only, Stage II only, Stage III only, and Stage IV only) for each cancer type can be used for clinical survival analysis. For a specified cancer type, the RNA-seq FPKM values generated by Cufflinks are used to divide the samples into two groups, according to user-defined FPKM percentile thresholds. The group 1 samples of the given cancer type and stage consist of patient data with FPKM values higher than the high percentile threshold. The group 2 samples include the data with FPKM values lower than the low percentile threshold in the same cancer type and stage. The survival functions for the filtered two groups are plotted based on the Kaplan–Meier method using the Python package `lifelines`. Cancer DEIso further performs the log-rank test against the null hypothesis that the survival curves between samples of the two groups are equal.

2.2. Implementation of Cancer DEIso

We used the Python scripting language (version 2.7.6 by Python Software Foundation, an open-source software) to facilitate and streamline the data preprocessing and analysis pipeline. The website query and browse interface of the Cancer DEIso database are implemented using the Python Model-View-Controller (MVC) framework Django (version 1.11.1 by Django Software Foundation, an open-source software). The processed RNA-seq data and analysis results are deposited using the Mysql database management system (version 5.7.19 by Oracle Corporation, Santa Clara, California, U.S).

3. Utility and discussion

3.1. Database interface

In Cancer DEIso, the following gene-level or isoform-level search functionalities were implemented to analyze stage

differentially expressed cancer markers: (1) Search for cancer marker candidates. Users can specify the cancer types and stages, and the differentially expressed gene list or isoform list will be provided. The stage differentially expressed gene/isoform list reveals the potential markers for the selected cancer type. (2) Search cancer types in which a given gene/isoform is differentially expressed. Users can also input the gene or transcript of interest to investigate which cancer types are potentially related to the differential expression of the input gene/isoform. The cancer types in which the gene/isoform shows differential expression are then summarized. Users can browse through each related cancer type and check the detailed expression FPKM values. (3) Survival analysis. Users can choose a gene/isoform, the cancer type, and its cancer stage to perform the survival analysis based on the expression FPKM values generated by Cufflinks. A statistical test on the survival comparison is also provided in this function. (4) Stage comparison. Users can visualize the expression difference between stages of the selected cancer and gene/isoform. A box plot will be generated to reveal the distribution information. (5) Download function. The list of potential cancer markers and the survival analysis results can all be downloaded in.csv files. These functionalities are described in detail in the following subsections.

3.1.1. Function 1 (Search Potential Cancer Markers): gene/isoform level cancer marker investigation

When users specify one cancer type and two stages, a list of genes or transcripts with differential FPKM expressions is shown (See Fig. 2). In this function, four parameters should be provided (Fig. 2-a). Users first need to select whether the gene-level or isoform-level analysis is considered. Then the query cancer type should be specified. After selecting the search cancer type, users can indicate the Condition 1 cancer stage and the Condition 2 cancer stage for performing differential expression analysis. In Cancer DEIso, differential expression analysis is computed by the average FPKM ratio between Condition 2 and Condition 1 (average FPKM of Condition2/ average FPKM of Condition1). Users can further provide the minimum fold change and statistical significance threshold to control the false discovery rate in the fourth part. We have incorporated the *q*-values computed by Cuffdiff and additionally performed the Kolmogorov–Smirnov distribution test, the one-tailed independent *t*-test, and the one-tailed Mann–Whitney U test to test for the alternative hypothesis that the distribution/average/median FPKM value of Condition 2/1 is larger than the value of Condition 1/2, respectively. The FDR correction procedure was performed for all genes or isoforms to fix the multiple-hypotheses bias. After inputting these parameters, a list of genes/isoforms satisfying the user-specified thresholds will be presented in a tabular format (Fig. 2-b). A heat map that visualized the differential FKPM values between Condition 2 and Condition 1 for every item in the list is provided. Further, the FPKM values for Condition 2 and Condition 1 (calibrated by Cuffdiff), the FPKM ratio of Condition 2 over Condition 1 (computed by Cuffdiff), and the comparison *q*-values are also given in the table. Users can investigate these potential cancer markers and click the "detail" link of some confident item for further information (Fig. 2-c). Finally, users can click the "download" button to download the list for subsequent analyses (Fig. 2-d). Users using the differential isoform analysis function can also filter the transcript to investigate only coding transcripts or non-coding transcripts (Fig. 2-e).

3.1.2. Function 2 (Search Potentially Induced Cancers): search for cancer types potentially induced by the differential expression of a selected gene/isoform

Users can also determine the cancer types and stages in which a specified gene or isoform bears differential expression (See Fig. 3). This aids users in understanding the potentially induced cancer types due to the differential expression of the chosen gene/isoform. In this function, users need to type in the gene or isoform of interest first and then set the Cuffdiff fold change threshold and the differential statistic significance level (Fig. 3-a). After clicking the search button, a table of the cancer types with stage differential expression of the specified gene/isoform is listed (Fig. 3-b). The detailed information of the differential expression for each potentially induced cancer type can be further checked in the "detail" link of every compared condition that satisfies the specified fold change and significance level (Fig. 3-c). All the information can be downloaded by the "download" button provided in the page (Fig. 3-d).

3.1.3. The detail comparison page

Whether users pick the potential cancer marker from the list of differentially expressed genes or transcript isoforms, a detail comparison page for the marker is provided (See Fig. 4). In the detail page, related information for the chosen marker is listed. In Fig. 4, we provide an example of an isoform detail page when the user selects a potential isoform cancer marker. In this isoform information page, the differential FPKM expression values between the specified cancer stages are tabulated (Fig. 4-a). At the end of the rows, a quick link to the survival analysis for this marker is provided (Fig. 4-b). Below the table of FPKM expression comparison, a boxplot comparison visualization of the markers between the selected two cancer stages is shown (Fig. 4-c). To understand the alternative splicing activity of the marker, we provide the transcript information and the genomic map of exon constitutions for all splice isoforms related to the marker (Fig. 4-d). Users can also investigate if the differential isoform expression of a given transcript results from differential isoform usage and isoform switching by comparing the IF values and the dIF values between the selected two cancer stages (Fig. 4-e). Moreover, the expression information of the related gene of the selected potential isoform marker is also summarized. By clicking the "Related gene information" tab (Fig. 4-f), the gene's information and FPKM expression summary is shown in another tab page. In the bottom part of the related gene tab page, a boxplot of the gene expressions between the specified cancer stages is also plotted to facilitate overall comparison visualization. On the other hand, if users are navigated to the detail page via a potential gene marker, the user is first directed to the gene tab page. And the related transcript info page is provided for further referencing. All the contents in the gene marker page and the related transcript info page are similar to those in Fig. 4.

3.1.4. Function 3 (Survival Analysis)

In the third function implemented in Cancer DEIso, users can perform the survival analysis on the gene/isoform of interest for the specified cancer type. Cancer DEIso supports the survival function visualization for the top M% and last M% (default M = 50) cancer samples of the chosen cancer type based on the FPKM expression of the specified gene/isoform (See Fig. 5). And the log-rank test for the two groups is performed to indicate the statistical significance of the difference between the two survival functions. Users first type in the gene or isoform of interest and then choose the query cancer type (Fig. 5-a). After pressing the search button, the survival analysis using the given gene/isoform marker on the cancer samples of the chosen cancer type is performed for the user. As a default setting, the survival functions (unit in days) for cancer samples with top 50% FPKM values (high expression group) and cancer samples with last 50% FPKM values (low expression group) are computed. Users are free to change the FPKM thresholds that define the high expression and low expression groups. Users can also update the survival functions using cancer samples diagnosed in different stages (Fig. 5-b). By pushing the submit button with the

Fig. 2. Function 1 of Cancer DEIso. In Function 1 (Search Potential Cancer Markers), users can specify the cancer type, analysis stages, and the significance threshold to obtain the list of differentially expressed genes or isoforms. (a) The query form of Function 1. (b) The differentially expressed gene list or isoform list for the given cancer type and stages. (c) The link to the detail expression information for the selected gene or isoform between the selected 2 stages. (d) The differentially expressed gene/isoform list can be download as a plain text file for further processing. (e) The users can filter the transcript list to check only coding transcripts or non-coding transcripts. This filter only appears in the "DE isoforms" mode.

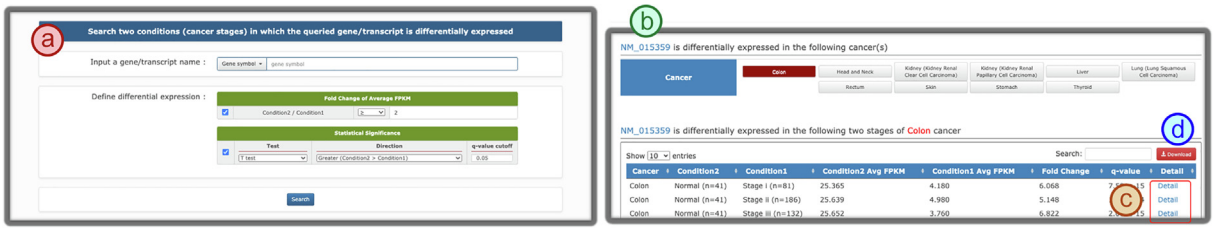

Fig. 3. Function 2 of Cancer DEIso. In Function 2 (Search Potentially Induced Cancer by the Specified Gene/Isoform), users can input a gene/transcript isoform name and the significance threshold to obtain the list of cancer types that may be potentially induced by the differential expression of the input gene/isoform. (a) The search form of Function 2. (b) The table of cancer types in which the input gene/isoform shows differential expression between two stages. (c) The link to the detail expression information for the selected gene or isoform between the specified 2 stages. (d) The potentially induced cancer type list and the differential expression of the input gene/isoform can be downloaded in a plain text file.

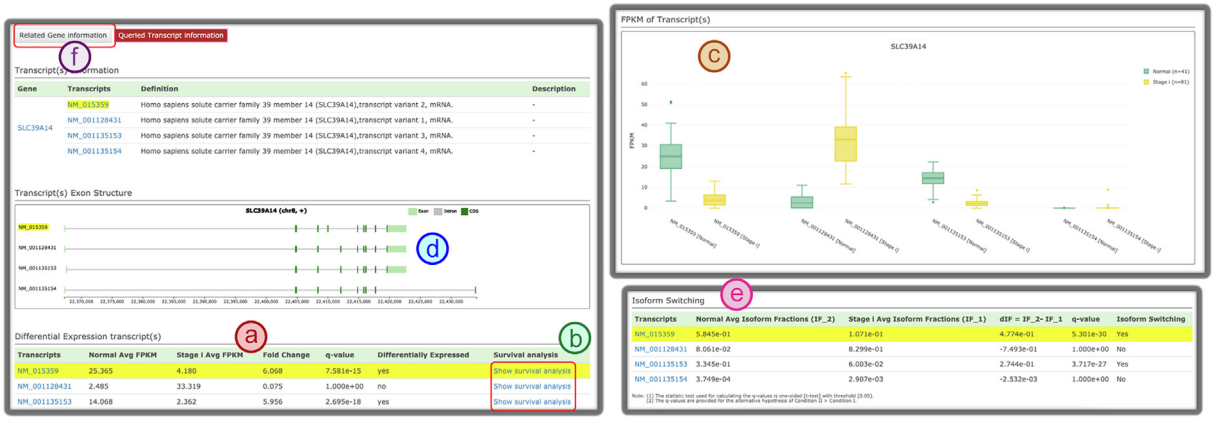

Fig. 4. The detail page in Cancer DEIso. In the detail page of a selected isoform or gene, the detailed comparison of FPKM values between the selected two stages is shown. In this figure, a differentially expressed isoform detail page is used as an example. The results are similar for a differentially expressed gene detail page. (a) The FPKM value comparison summary for the isoform. (b) A link to the survival analysis of the selected isoform is provided. (c) A boxplot visualizes the stage comparison of the related isoforms. (d) In the isoform detail tab, a visualization exon map of the alternative spliced transcripts is implemented for users to investigate the relationship between different splice isoforms. (e) In the isoform detail tab, a tabular summary of isoform switching test results is provided for identifying differential isoform usage. (f) The related gene tab can be clicked to view the comparison summary of the related gene of the specified isoform.

newly changed thresholds, survival functions of the newly defined high/low expression groups in the specified cancer type and stage are re-generated (Fig. 5-c). The log-rank test *p*-value is provided in the lower-left corner of the plot to indicate the significance of the hypothesis that the two survival functions are from different distributions (Fig. 5-d).

3.1.5. Function 4 (Stage Comparison): visualization of the cancer stage expression comparison

Sometimes users merely need to compare the RNA-seq expressions between some specified cancer stages. We implemented a function that helps visualize the distribution comparison between two or more cancer stages of a cancer type (See

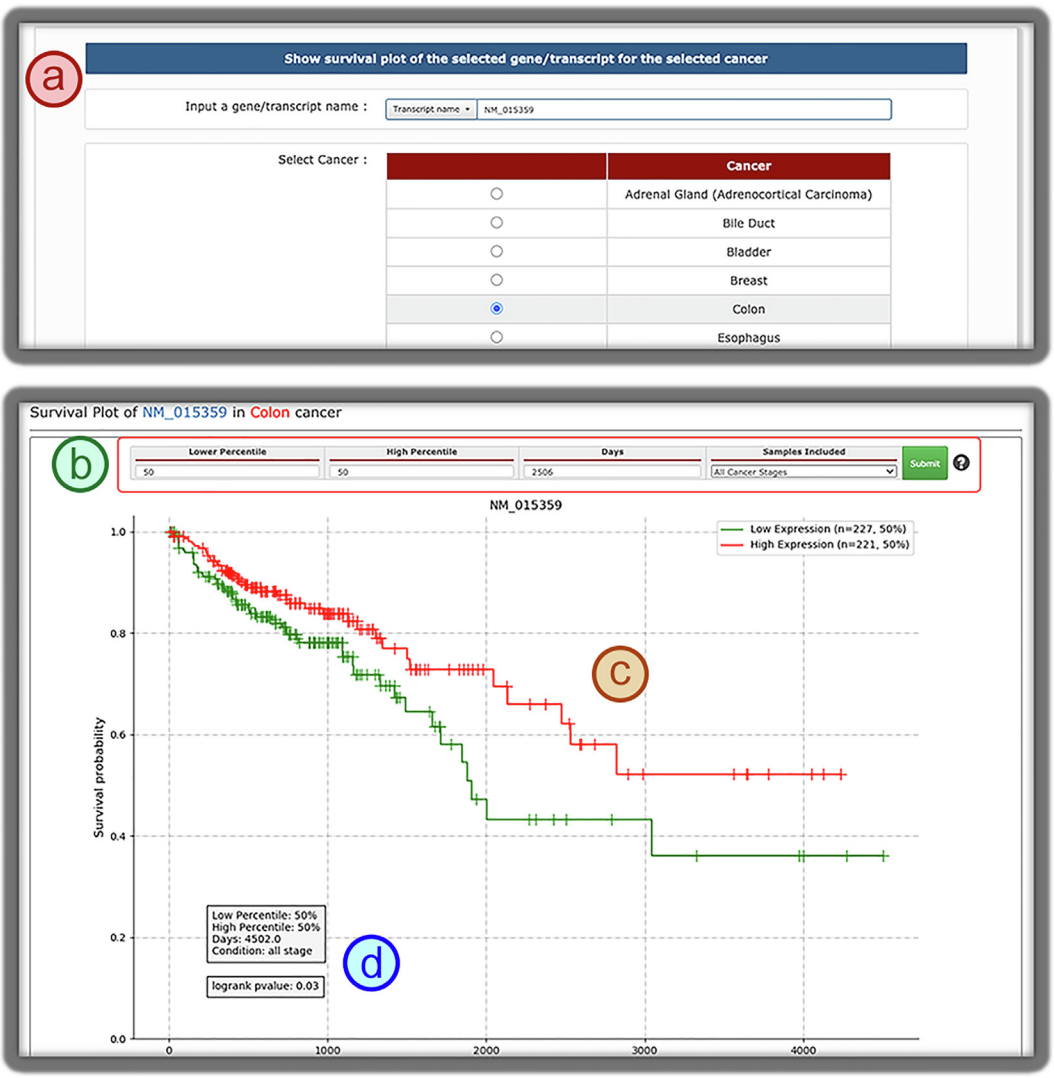

Fig. 5. Function 3 of Cancer DEIso. In Function 3 (Survival Analysis), users can input a gene/transcript isoform name and a cancer type to perform the survival analysis using the RNA-seq FPKM expression values. (a) The query form of Function 3. (b) The threshold and cancer stages for the high expression group and the low expression group can be specified. (c)(d) The survival curves of the high expression and low expression groups and the log-rank test *p*-value.

Fig. 6). In this part, users can type in the gene/isoform of interest and select the cancer stages and type (Fig. 6-a). Users can also select the primary tumor group (patient samples from all cancer stages) to compare primary tumors with normal tissues. After the "Search" button is pressed, the paired expression comparison results for the specified gene/isoform between the normal samples and primary tumors are listed in a table (Fig. 6-b). And a comparison FPKM value distribution boxplot that lines up the selected cancer stages and the primary tumor group is shown (Fig. 6-c). Both the FPKM median values and average values are marked on the plot for referencing. Moreover, all categories of the "primary_diagnosis" and "morphology" clinical data columns for the samples from the chosen stages of the given gene/isoform are listed (Fig. 6-d). Users can select the category items to filter the samples that match the chosen diagnosis codes and morphology codes. The gene/isoform expressions of the samples satisfying the selected codes in different cancer stages will then be filtered in the comparison boxplot (Fig. 6-c). Finally, the generated boxplot can be downloaded as a .png file by clicking the download function on the upper right corner (Fig. 6-e).

3.2. Case study

The constructed Cancer DEIso database can help researchers investigate the potential cancer markers in the gene level or the isoform level. We provide a walk-through example to elucidate how the database can be used. Colorectal cancer is one of the most prevalent types of cancer worldwide. And a suitable diagnostic marker for colorectal carcinogenesis can better support early patient treatment [28,39]. We investigated the differentially expressed isoforms in colorectal cancer cells (stage I) versus normal samples to search for potential markers. We used Function 1 of Cancer DEIso to filter out differentially expressed isoforms between stage I colorectal cancer and normal samples (Fig. 7). The results show that NM_015359 (alternatively spliced SLC39A14-4A transcript) has significantly higher FPKM expression in the normal samples (average FPKM = 25.365, 4.180 in normal samples and stage I cancer samples, respectively). And NM_001128431 (alternatively spliced SLC39A14-4B transcript) bears significantly higher FPKM expression in the stage I colorectal cancer samples (average FPKM = 2.485, 33.319 in normal samples and stage I cancer cells, respectively). These two transcripts differ

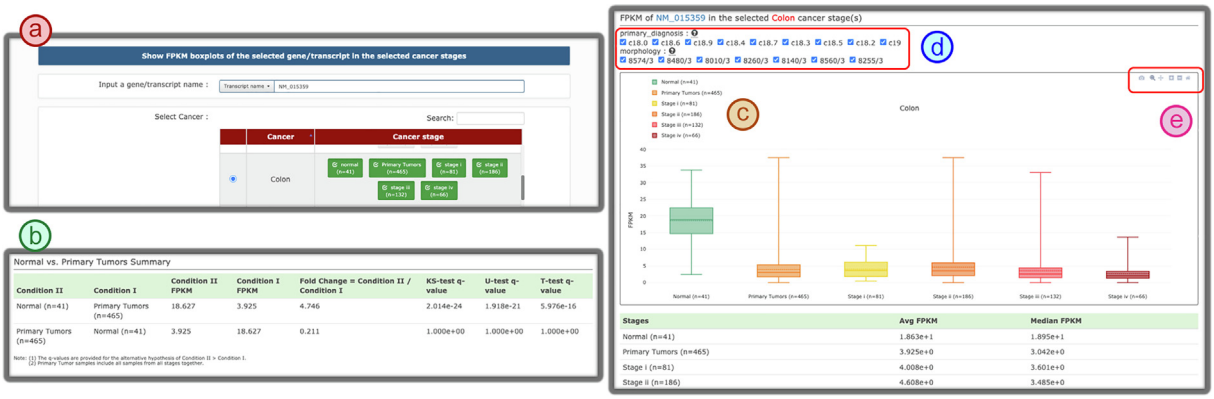

Fig. 6. Function 4 of Cancer DEIso. In Function 4 (Stage Comparison), users can input a gene/transcript isoform name, a cancer type, and stages to visualize the RNA-seq FPKM expression value comparison between stages. (a) The query form of Function 4. (b) The comparison results between the primary tumor samples and the normal group. (c) The boxplot and the average values are provided. (d) Clinical data filtering for samples of the specified gene/isoform in different cancer stages. (e) The plot can be downloaded in the.png format.

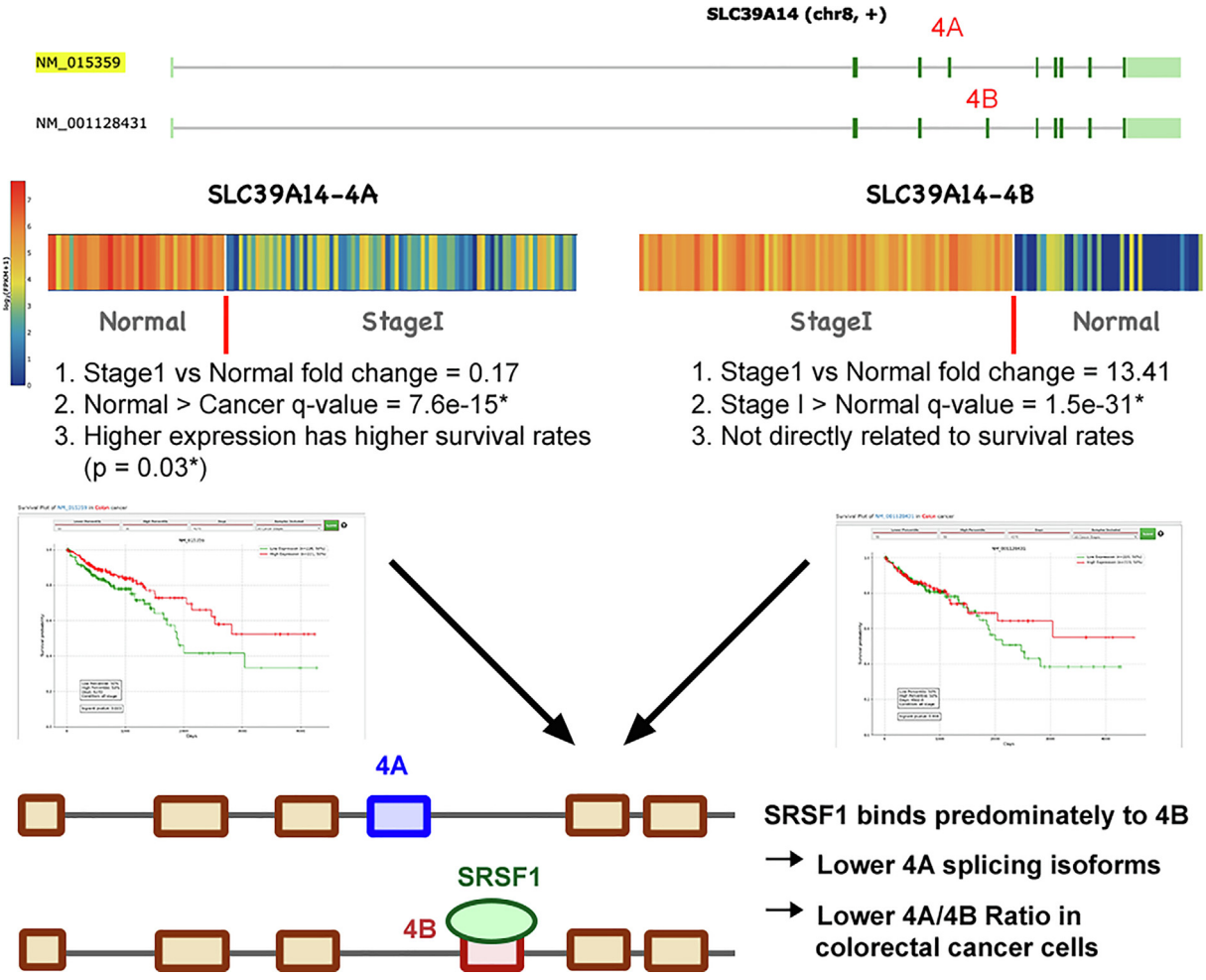

Fig. 7. Cancer DEIso can help identify potential cancer markers. In the deposited analysis results of Cancer DEIso, the SLC39A14-4A transcript isoform shows decreased FPKM expression after colorectal cancer detection while the SLC39A14-4B transcript isoform demonstrates an increase in FPKM expression in colorectal cancer cells. The survival analysis of SLC39A14-4A reveals that low SLC39A14-4A expression leads to a lower survival rate. It is now experimentally verified that the decrease ratio of SLC39A14-4A/4B can be observed after colorectal carcinogenesis and is possibly regulated by SRSF1 (Serine And Arginine Rich Splicing Factor 1) that targets the 4B exon. The experimental evidence supports the hypothesis deduced from Cancer DEIso.

only in the fourth exon. However, the SLC39A14 gene-level expression reveals no difference between normal and cancer samples. A similar FPKM trend between cancer and normal cells is also observed in the rectum samples. To further check the clinical property of these two transcripts, the survival analysis for these two transcripts on colon cancer was performed. In the survival analysis, the SLC39A14-4A transcript reveals a higher survival function when highly expressed (log-rank $p = 0.03^{*}$). Based on the information, it is hypothesized that lower SLC39A14-4A expression and higher SLC39A14-4B could detect the existence of

colorectal cancer. The SLC39A14 alternative spliced transcripts 4A and 4B have now been verified to be differentially expressed in colorectal cancer samples [40]. These two transcripts are known to have different cadmium binding affinity [41]. The dysregulation of the cadmium binding affinity caused by the decrease expression ratio between SLC39A14 4A/4B transcript results in an increased cadmium ion uptake in colorectal cancer cells [42]. Since cadmium ion may induce a conformational change in p53, the decrease expression ratio of SLC39A14 4A/4B may lead to a higher DNA damage rate and trigger colorectal carcinogenesis [43]. On the other hand, SLC39A14 gene expression has no change between cancer and normal cells by qRT-PCR. The gene-level expression similarity further indicates the importance of isoform-level marker investigation. All these experimental findings match the hypothesis generated by the database. Therefore, Cancer DEIso can serve to facilitate cancer research.

3.3. Comparison with related works

With the development of high-throughput RNA-seq techniques, many transcriptome-wide analyses can be performed to unearth candidate biomarkers for different diseases. TCGA project has been launched to help gather the cancer-related RNA-seq samples. However, detailed data analyses of the deposited datasets must be carried out to understand cancer biology further. The Cancer DEIso database is constructed to provide biologists a new repository for both potential gene-level and isoform-level cancer marker investigation based on stage differential analysis and survival analysis. Some previous researches were also conducted to devote to similar purposes. These results can be divided into two categories: the gene-level analysis platforms and the isoform-level analysis platforms. We compare the constructed Cancer DEIso database with previous works to pinpoint the novel improvements in Cancer DEIso. The overall comparison summary is listed in Table 2.

The first type of analysis platforms considers only the gene-level differential expression between the normal and cancer samples. These platforms include KM-express [14], OncoLnc [15], GEPIA [44], UALCAN [16], and Cancer RNA-seq Nexus [13]. These platforms provide the functionality of gene-level differential expression investigation. In KM-express, an online patient survival and expression analysis tool for breast and prostate cancers was developed. However, only the breast and prostate cancer analyses are available in KM-express, limiting the interest of broader audiences. The GEPIA web server broadened the cancer types in the developed platform. Gene-level cancer marker investigation and whole feature dimension reduction for specific cancer types were provided in GEPIA. However, no stage information and comparison visualization are available in GEPIA. OncoLnc and Cancer RNA-Seq Nexus incorporated the TCGA datasets with miRNA, lncRNA, and gene regulatory networks to aid the biological experiment design based on the cancer gene expression profiles. Finally, in UALCAN, RNA-seq data collected from TCGA (level 3, v2 expression data) were analyzed by the RSEM tool [45]. UALCAN visualized the differential expression between normal and cancer samples through the implemented heatmap and boxplot function. Advanced analysis based on patient's race, sex, body weight, diagnosed cancer stage, or other features were also implemented in the UALCAN database platform. Most of these platforms are equipped with gene-level survival analysis for the TCGA RNA-seq data. However, none of these platforms can help investigate the isoform-level cancer markers. It is known that various cancers might be diagnosed or caused by the malfunction of the splicing mechanisms, leading to some unusual transcript isoform expressions [6]. Hence it is of importance to survey the differential expression in the isoform level.

The ISOexpresso database [17] and the GEPIA2 web-service [18] provide the opportunity to consider splice isoform expressions. ISOexpresso collected the cancer RNA-seq data (level 3, v2 expression data) from TCGA and the Refseq gene isoform splice information from UCSC Genome Browser [46]. They performed the expression analysis using the RSEM tool [45] to obtain the TPM (transcript per million) expression value for each human transcript isoform. Two types of candidate cancer-specific isoforms were considered in ISOexpresso. Type I isoforms consist of transcripts that are expressed only in cancer cells and are absent in normal cells. And Type II isoforms represent transcripts with fold change larger than 2 in comparing cancer samples versus normal cells. While the analysis provides novel candidates, there are large portions of false positives due to the lack of significance control. And it will be better to have the survival analysis for a selected candidate isoform marker before verification experiments or clinical research. Further, cancer stage information and comparison visualization are not available in ISOexpresso, and users cannot simultaneously study the gene-level and isoform-level expression analysis. GEPIA2 extends the functionalities of GEPIA to include the isoform expression analysis and isoform-level survival analysis. Nevertheless, there is still no stage comparison functionality for genes and isoforms. In summary, there is still a lack of stage-differential isoform analysis functionalities in ISOexpresso and GEPIA2.

The Cancer DEIso database is constructed to supplement the insufficiency of all these platforms. The improvements of Cancer DEIso make it feasible to search both gene-level and isoform-level markers and provide stage comparisons. In Cancer DEIso, transcript isoform differential expression in the RNA-seq data was calculated using Cufflinks and Cuffdiff to provide potential cancer transcript markers. Besides finding the differential genes or isoforms for a given cancer type, users can investigate other cancer types that may also be induced by the differential expression of the identified gene or isoform marker. Then survival analysis for each gene or transcript isoform based on the RNA-seq FPKM values was implemented to help validate the clinical significance of the identifications. These supplemented functions to current tools can make up the need for carcinogenesis research. Therefore, Cancer DEIso is definite to broaden the add-on values of TCGA data and help the community.

3.4. Issues related to Cancer DEIso

Cancer DEIso is constructed to provide novel add-on values to the deposited TCGA RNA-seq datasets. Users can search the database for differentially expressed potential cancer markers or investigate possible cancer types induced by the differential expression of some particular gene/isoform. Survival analysis is also available in Cancer DEIso to help evaluate the clinical confidence of the selected cancer marker. Some issues should be taken awareness of when investigating the potential cancer markers or the information of affected cancers by a specific gene or isoform.

In Cancer DEIso, RNA-seq samples for the cancer-stage differential expression analysis were collected from various patients using different cDNA library protocols or analysis methods. The problem was dealt with in the original TCGA data depositing flow. Sample quality was first ensured by the Biospecimen Core Resource (BCR) in TCGA [10] to alleviate the protocol variance. The clinical data, metadata, and sequencing data were then submitted to the Genome Characterization Centers (GCCs) and Genome Sequencing Centers (GSCs) to deposit the raw BAM files. And data analysis issues were calibrated by the GDC Data Portal. In the GDC Data Portal, the submitted .bam or .fastq files were realigned using the International Cancer Genome Consortium (ICGC) STAR (Spliced Transcripts Alignment to a Reference)[47] two-pass alignment standard operation procedure (SOP). The analysis bias in the

Table 2
Comparison with other related works. KM-express, OncoLnc, GEPIA, and UALCAN provide only gene-level functionality.

	Cancer DEIso	Cancer RNA-seq Nexus	KM-express	OncoLnc	GEPIA	GEPIA2	UALCAN	ISOexpresso
Gene-level expression analysis	v	v	v	v	v	v	v	v
Isoform-level expression analysis	v					v		v
Cancer stage comparison	v						v	
Between-stage marker investigation	v							
Survival analysis	v		v	v	v	v	v	

deposited aligned BAM files is hence reduced to the minimum level. From this data traceback, the data collection biases are ensured to be controlled. To further relieve inter-individual noises, users are suggested to carry out further validation experiments of the differentially expressed cancer markers. Finally, RNA-seq only captures the transcript existence in cells. The technique does not probe the actively translated transcripts [48]. The dynamics of the genes/isoforms between cancer stages can be further investigated using the ribosome profiling (ribo-seq) techniques [49]. The information will be updated and incorporated into Cancer DEIso when the patient ribo-seq data are available.

Different patient samples are available in different cancer stages or types. Nevertheless, some cancer stages include only few RNA-seq data. We introduced three additional statistical tests in the Cuffdiff-calibrated FPKM value comparison between stages to address this problem in the analysis process. When the sample numbers of the two stages are massive, users are suggested to use the one-tailed independent *t*-test for a higher statistical power. When one of the selected stages has only few available samples, users are suggested to use the Mann–Whitney U test or the Kolmogorov–Smirnov distribution test for the comparison. These two non-parametric methods are more suitable to deal with conditions with only few samples. Therefore, besides canonical condition fold change, the difference significance is also provided in the analysis results. The sample number problem is handled in the differential analysis and should be considered when using Cancer DEIso.

4. Conclusions

In this research, we presented a database called Cancer DEIso (Cancer differentially expressed isoform and gene database). In this database, both gene-level and transcript isoform-level expressions are compared between different stages of a specified cancer type. The isoform-level analysis can reveal more precise diagnostic cancer markers. Four functions were implemented in Cancer DEIso to facilitate easy investigation of potential gene-level and isoform-level markers. Users can find the differentially expressed genes or transcript isoforms for the given cancer type and stages. Further, users can identify the cancer types potentially induced by the stage-differential expression of a specified gene or transcript. Survival analysis and comparison visualization are also provided in the database interface. Compared with previous similar tools, Cancer DEIso further provides isoform-level marker investigation, cancer stage expression comparison, and transcript isoform survival analysis. Moreover, the gene-level and isoform-level comparisons are integrated in a detail page for a better mechanistic understanding of the potential marker. We believe that Cancer DEIso can provide extra novel insights for the deposited TCGA datasets.

Data Availability

The datasets supporting the conclusions of this article are included within the article and the database website (http://cosbi4.ee.ncku.edu.tw/DEIso/).

CRediT authorship contribution statement

Tzu-Hsien Yang: Project administration, Software, Formal analysis, Writing - original draft. **Yu-Hsuan Chiang:** Data Curation, Software, Investigation. **Sheng-Cian Shiue:** Data Curation, Investigation. **Po-Heng Lin:** Software, Investigation. **Ya-Chiao Yang:** Investigation, Formal analysis, Writing - review & editing. **Kai-Chi Tu:** Investigation, Formal analysis, Writing - review & editing. **Yan-Yuan Tseng:** Conceptualization, Project administration, Resources. **Joseph T. Tseng:** Conceptualization, Project administration, Methodology. **Wei-Sheng Wu:** Conceptualization, Project administration, Methodology, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was supported by National University of Kaohsiung, National Cheng Kung University, and Ministry of Science and Technology of Taiwan (MOST 107-2218-E-390-009-MY3, MOST 107-2221-E-006-225-MY3, MOST 108-2628-E-006-004-MY3 and MOST 110-2221-E-006-198-MY3).

References

[1] Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. Alternative isoform regulation in human tissue transcriptomes. Nature 2008;456(7221):470–6.
[2] Long JC, Caceres JF. The SR protein family of splicing factors: master regulators of gene expression. Biochem J 2009;417(1):15–27.
[3] Faustino NA, Cooper TA. Pre-mRNA splicing and human disease. Genes Dev 2003;17(4):419–37.
[4] Karni R, de Stanchina E, Lowe SW, Sinha R, Mu D, Krainer AR. The gene encoding the splicing factor SF2/ASF is a proto-oncogene. Nature Struct Mol Biol 2007;14(3):185–93.
[5] Venables JP. Aberrant and alternative splicing in cancer. Cancer Res 2004;64 (21):7647–54.
[6] Yang IS, Kim S. Isoform specific gene expression analysis of KRAS in the prognosis of lung adenocarcinoma patients. BMC Bioinformatics 2018;19 (1):1–10.
[7] Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The cancer genome atlas pan-cancer analysis project. Nat Genet 2013;45(10):1113–20.
[8] Jensen MA, Ferretti V, Grossman RL, Staudt LM. The NCI Genomic Data Commons as an engine for precision medicine. Blood 2017;130(4):453–9.
[9] Hutter C, Zenklusen JC. The cancer genome atlas: creating lasting value beyond its data. Cell 2018;173(2):283–5.
[10] Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemporary Oncol 2015;19(1A): A68.
[11] Masseroli M, Canakoglu A, Pinoli P, Kaitoua A, Gulino A, Horlova O, Nanni L, Bernasconi A, Perna S, Stamoulakatou E, et al. Processing of big heterogeneous genomic datasets for tertiary analysis of Next Generation Sequencing data. Bioinformatics 2019;35(5):729–36.
[12] Hanahan D, Weinberg RA. The hallmarks of cancer. Cell 2000;100(1):57–70.
[13] Li J-R, Sun C-H, Li W, Chao R-F, Huang C-C, Zhou XJ, Liu C-C. Cancer RNA-Seq Nexus: a database of phenotype-specific transcriptome profiling in cancer cells. Nucleic Acids Res 2016;44(D1):D944–51.

[14] Chen X, Miao Z, Divate M, Zhao Z, Cheung E. 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;2018.
[15] Anaya J. OncoLnc: linking TCGA survival data to mRNAs, miRNAs, and lncRNAs. PeerJ Computer Sci 2016;2:e67.
[16] Chandrashekar DS, Bashel B, Balasubramanya SAH, Creighton CJ, Ponce-Rodriguez I, Chakravarthi BV, Varambally S. UALCAN: a portal for facilitating tumor subgroup gene expression and survival analyses. Neoplasia 2017;19(8):649–58.
[17] Yang IS, Son H, Kim S, Kim S. ISOexpresso: a web-based platform for isoform-level expression analysis in human cancer. BMC Genomics 2016;17(1):1–14.
[18] Tang Z, Kang B, Li C, Chen T, Zhang Z. GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. Nucleic Acids Res 2019;47(W1):W556–60.
[19] Grossman RL, Heath A, Murphy M, Patterson M, Wells W. A case for data commons: toward data science as a service. Computing Sci Eng 2016;18(5):10–20.
[20] Haeussler M, Zweig AS, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, Lee CM, Lee BT, Hinrichs AS, Gonzalez JN, et al. The UCSC genome browser database: 2019 update. Nucleic Acids Res 2019;47(D1):D853–8.
[21] Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 2008;18(9):1509–17.
[22] Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL. Integrative annotation of human large intergenic noncoding rnas reveals global properties and specific subclasses. Genes Dev 2011;25(18):1915–27.
[23] Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 2012;7(3):562.
[24] Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci 2012;131(4):281–5.
[25] Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 2016;34(5):525–7.
[26] Pimentel H, Bray NL, Puente S, Melsted P, Pachter L. Differential analysis of RNA-seq incorporating quantification uncertainty. Nat Methods 2017;14(7):687–90.
[27] Sahraeian SME, Mohiyuddin M, Sebra R, Tilgner H, Afshar PT, Au KF, Asadi NB, Gerstein MB, Wong WH, Snyder MP, et al. Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. Nature Commun 2017;8(1):1–15.
[28] Yang T-H, Chang H-T, Hsiao ES, Sun J-L, Wang C-C, Wu H-Y, Liao P-C, Wu W-S. iPhos: a toolkit to streamline the alkaline phosphatase-assisted comprehensive LC-MS phosphoproteome investigation. BMC Bioinformatics 2014;15(16):1–14.
[29] Massey Jr FJ. The Kolmogorov-Smirnov test for goodness of fit. J Am Stat Assoc 1951;46(253):68–78.
[30] Boneau CA. The effects of violations of assumptions underlying the t test. Psychol Bull 1960;57(1):49.
[31] Steel RG. A rank sum test for comparing all pairs of treatments. Technometrics 1960;2(2):197–207.
[32] Yang T-H, Wu W-S. Inferring functional transcription factor-gene binding pairs by integrating transcription factor binding data with transcription factor knockout data. BMC Syst Biol 2013;7(6):1–14.
[33] Yang T-H. An aggregation method to identify the RNA meta-stable secondary structure and its functionally interpretable structure ensemble. IEEE/ACM Trans Comput Biol Bioinf 2021.
[34] Yang T-H. Transcription factor regulatory modules provide the molecular mechanisms for functional redundancy observed among transcription factors in yeast. BMC Bioinformatics 2019;20(23):1–16.
[35] Vitting-Seerup K, Sandelin A. The landscape of isoform switches in human cancers. Mol Cancer Res 2017;15(9):1206–20.
[36] Vitting-Seerup K, Sandelin A. IsoformSwitchAnalyzeR: analysis of changes in genome-wide patterns of alternative splicing and its functional consequences. Bioinformatics 2019;35(21):4469–71.
[37] Haggar FA, Boushey RP. Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors. Clinics Colon Rectal Surgery 2009;22(4):191.
[38] Hassan MRA, Suan MAM, Soelar SA, Mohammed NS, Ismail I, Ahmad F. Survival analysis and prognostic factors for colorectal cancer patients in malaysia. Asian Pac J Cancer Prev 2016;17(7):3575–81.
[39] Ferlay J, Shin H-R, Bray F, Forman D, Mathers C, Parkin DM. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. Int J Cancer 2010;127(12):2893–917.
[40] Sveen A, Bakken AC, Ågesen TH, Lind GE, Nesbakken A, Nordgård O, Brackmann S, Rognum TO, Lothe RA, Skotheim RI. The exon-level biomarker SLC39A14 has organ-confined cancer-specificity in colorectal cancer. Int J Cancer 2012;131(6):1479–85.
[41] Liuzzi JP, Aydemir F, Nam H, Knutson MD, Cousins RJ. Zip14 (Slc39a14) mediates non-transferrin-bound iron uptake into cells. Proc Nat Acad Sci 2006;103(37):13612–7.
[42] Thorsen K, Mansilla F, Schepeler T, Øster B, Rasmussen MH, Dyrskjøt L, Karni R, Akerman M, Krainer AR, Laurberg S, et al. Alternative splicing of SLC39A14 in colorectal cancer is regulated by the wnt pathway. Mol Cellular Proteomics 2011;10(1):M110–002998.
[43] Wieland M, Levin MK, Hingorani KS, Biro FN, Hingorani MM. Mechanism of cadmium-mediated inhibition of Msh2-Msh6 function in DNA mismatch repair. Biochemistry 2009;48(40):9492–502.
[44] 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. Nucleic Acids Res 2017;45(W1):W98–W102.
[45] Li B, Dewey CN. RSEM: accurate transcript quantification from rna-seq data with or without a reference genome. BMC Bioinformatics 2011;12(1):1–16.
[46] Navarro Gonzalez J, Zweig AS, Speir ML, Schmelter D, Rosenbloom KR, Raney BJ, Powell CC, Nassar LR, Maulding ND, Lee CM, et al. The UCSC Genome Browser database: 2021 update. Nucleic Acids Res 2021;49(D1):D1046–57.
[47] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;29(1):15–21.
[48] Yang T-H, Wang C-Y, Tsai H-C, Liu C-T. Human IRES Atlas: an integrative platform for studying IRES-driven translational regulation in humans. Database 2021;2021.
[49] Ingolia NT, Lareau LF, Weissman JS. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell 2011;147(4):789–802.