This is a provisional PDF only. Copyedited and fully formatted version will be made available soon.

REPORTS OF PRACTICAL
ONCOLOGY AND
RADIOTHERAPY

The RNA world: from experimental laboratory to "in silico" approach. Part 1: User friendly RNA expression databases portals

Authors: Tomasz Kolenda, Marianna Iga Śmiełowska, Julia Lipowicz, Julia Ostapowicz, Paula Pacześna, Monika Rosochowicz, Paulina Poter, Joanna Kozłowska-Masłoń, Kacper Guglas, Klaudia Dudek, Nina Grzejda, Katarzyna Regulska, Anna Florczak, Urszula Kazimierczak, Katarzyna Lamperska, Anna Teresiak

This article has been peer reviewed and published immediately upon acceptance.

The RNA world: from experimental laboratory to "in silico" approach. Part 1: User friendly RNA expression databases portals

Tomasz Kolenda[1], Marianna Iga Śmiełowska[2,3], Julia Lipowicz[3, 4], Julia Ostapowicz[5, 6], Paula Pacześna[3], Monika Rosochowicz[67], Paulina Poter[89], Joanna Kozłowska-Masłoń[1, 10], Kacper Guglas[11], Klaudia Dudek[1, 3, 13], Nina Grzejda, Katarzyna Regulska, Anna Florczak, Urszula Kazimierczak, Katarzyna Lamperska, Anna Teresiak

[1]Laboratory of Cancer Genetics, Greater Poland Cancer Centre, Poznan, Poland

[2]Department of Genome Engineering, Institute of Bioorganic Chemistry Polish Academy of Sciences, Poznan, Poland

[3]Research and Implementation Unit, Greater Poland Cancer Centre, Poznan, Poland

[4]Department of Histology and Embryology, Poznan University of Medical Sciences, Poznan, Poland

[5]Department and Clinic of Head and Neck Surgery and Laryngological Oncology, Poznan University of Medical Sciences, Poznan, Poland

[6]Radiobiology Laboratory, Department of Medical Physics, Greater Poland Cancer Centre, Poznan, Poland

[7]Department of Orthopaedic and Traumatology, W. Dega University Hospital, Poznan University of Medical Sciences, Poznan, Poland

[8]Department of Tumor Pathology and Prevention, Poznan University of Medical Sciences, Poznan, Poland

[9]Department of Tumor Pathology, Greater Poland Cancer Centre, Poznan, Poland

[10]Institute of Human Biology and Evolution, Faculty of Biology, Adam Mickiewicz University, Poznan, Poland

[11]Brachytherapy Department, Greater Poland Cancer Centre, Poznan, Poland

[12]Laboratory of Cancer Genetics, Greater Poland Cancer Centre, Poznan, Poland

[13]Department of Food Quality and Safety Management, Faculty of Food Science and Nutrition, Poznan University of Life Sciences, Poznan, Poland

[14]Uniwersytet im Adama Mickiewicza w Poznaniu, Poznan, Poland

[15]Pharmacy, Greater Poland Cancer Centre, Poznan, Poland

[16]Department of Diagnostics and Cancer Immunology, Greater Poland Cancer Centre, Poznan, Poland

[17]Chair of Medical Biotechnology, Department of Cancer Immunology, Poznan University of Medical Sciences, Poznan, Poland

[18]Chair of Medical Biotechnology, Faculty of Medicine, Poznan University of Medical Sciences, Poznan, Poland

Correspondence to: Tomasz Kolenda, Greater Poland Cancer Centre, Laboratory of Cancer Genetics, Garbary 15, 61–866 Poznan, Poland; e-mail: kolenda.tomek@gmail.com

Abstract

Cellular information about "life instruction" is stored, transferred, and modified using different types of RNA molecules. During the last decades, a growing number of RNA data has been generated thanks to the development of microarray chips and next-generation sequencing (NGS) methods. Improvement of bioinformatics contributed to the discovery of many types of new non-coding RNAs (ncRNAs), mostly with regulatory functions that supplemented the knowledge about the world of RNA. All of it, as well as the Human Genome Project (HGP) and the Cancer Genome Atlas (TCGA) project, has resulted in the formation of data storage and analysis portals which are widely used in cancer research and moved science from *in vitro* to *in silico* research. In this review we presented and discussed the data storage and analysis portals used by us, such as cBioPortal, UALCAN, ENCORI, and others. During the revision of these sites, we paid attention to data integration, simplicity of analysis, and results visualization, which are important for users without bioinformatic or statistical skills. In our opinion, the RNA analysis online tools will rapidly develop during the next decade and it seems to be a way for personalization of cancer treatment.

Key words: databases; RNA; coding- and non-coding RNA; TCGA; analysis; cancer; biomarkers

Different types of transcripts create the RNA world

RNA examination has become a central issue in contemporary transcriptome research, molecular pathomechanisms assessments of many diseases, including cancer, as well as evolutionary and developmental studies. This approach can be justified by RNA as the main source of information regarding dynamic changes in gene expression in response to different stimuli. On the other hand, DNA only provides a static picture of cell genomics [1–3]. The RNA world hypothesis renewed the importance of RNA research by underlining the ability of RNA to store data about past eras. The next decade is likely to witness an enormous rise in

2

interest in new RNA types, their functions and possible applications in innovative therapies and diagnostics. Despite scientists facing a real possibility of acquiring a complete knowledge about the *RNA world*, the characteristics of some RNAs have not been dealt with in depth. The aim of this work is to explain the probable advantages of RNA examining and to extend current knowledge to include technologies and tools used in this field [1–3].

Although the HGP occurred in 2003, its reverberation can still be felt today. The groundbreaking programme aiming at genome sequencing established that only 20,000–25,000 genes are protein-coding, accounting for less than 2% of the total genomic sequence [4, 5]. The fact that over 90% of the human genome is transcribed, significantly influenced the present understanding of molecular genetics, confirming that the RNA world is far more complex than assumed in the past century [5–7]. Classification of RNA particles depends on whether they are protein-coding or non-coding RNAs as well as on their length [8].

Even though, ncRNAs, such as transfer RNA (tRNA), ribosomal RNA (rRNA), and small nuclear RNAs (snRNAs), have been known for over 60 years, an abundant number of new functional ncRNAs along with their new features is still being described [8,9]. The majority of RNAs in a cell are ncRNAs [9]. They play a pivotal role as structural, catalytic, or regulatory RNAs [10–13]. ncRNAs are divided into two categories depending on their length: small non-coding RNAs (sncRNAs), which are shorter than 200 nucleotides, and long non-coding RNAs (lncRNAs) with a sequence longer than 200 bp [9]. rRNAs and tRNAs are usually excluded from this classification [14]. The features of RNA molecules are presented in Figure 1. A rapid development of RNA isolation techniques, sequencing capacity, and computational analysis facilitated expanding knowledge of the *RNA world*. Increasingly, more types of RNA and their functions are being discovered and tested for suitability in medicine as diagnostic or therapeutic measures. They are derived from different sources of biomaterials [15–19]. The uses of RNA-seq data in research are presented in Figure 2.

How to find the RNA maps?

The first step in research based on RNA-seq data is finding its proper source, which at the same time is the most problematic one. In fact, some databases are useless or inaccessible because of analysis inability. Fortunately, in recent years more and more portals have become "user-friendly" with tools for quick analysis mostly by the "one click" method. It should be underlined that most of them are based on TCGA data and in our opinion they are also the most developed and comprehensive ones.

The first RNA map: cBioPortal Platform

cBioPortal is an open platform for search, visualization, and multifaceted analysis of cancer genomics (mainly mutations and gene expression) and clinical databases. Cancer genomics data present in the public database are based on the TCGA project. The public version of cBioPortal contains over 200 studies, all of which are the TCGA projects and published results from the literature. The available data are divided according to the organs which are affected by cancer. It is also possible to search information by the exact tumor type. The programmes' convenient interface provides many useful analytical functions including a graphical summary of gene-level data across multiple platforms, analysis of the correlation between genes, survival analysis, and visualization of individual patient characteristics [20]. The aim of the cBioPortal is to significantly reduce the barriers between complex genomic data and cancer researchers by providing fast, intuitive, and high-quality access to molecular profiles and clinical features from large cancer genomics projects, enabling researchers to translate these intricate data sets into biological knowledge and medical applications [21]. The portal converts molecular profiling data from cancer tissues and cell lines into easily understandable genetic, epigenetic, and proteomic data. Owing to the easy-to-use interface, it is possible to customize data storage options and interactively study genetic changes in samples, genes, and pathways and even link them to clinical results. This site provides graphical summaries of results from multiple platforms, web visualization, and analysis, survival analysis, and software access. Complex cancer genomic profiles are available to researchers and clinicians without any need for bioinformatics knowledge, which is a great advantage [22]. The platform supports and stores non-synonymous mutations, DNA copy number data (presumed discrete values for each gene), messenger RNA (mRNA) and microRNA (miRNA) expression data, protein, and phosphoprotein level data (based on RPPA or mass spectrometry), DNA methylation data and unidentified clinical data. In the Data Kits section, the user can obtain a breakdown of the available types of cancer research data [21].

In addition to an easy-to-use interface and the ability to employ multiple data visualization tools by the OncoPrinter and MutationMapper, a great advantage of cBioPortal is that the platform incorporates data from large consortia, such as TCGA and the Therapeutically Applicable Research to Generate Effective Treatments (TARGET), which guarantees their reliability. The platform also provides an opportunity to customize data storage for each user. Searching for information does not require bioinformatic skills, which is a great benefit over similar platforms. cBioPortal also ensures that public data is available to

all users and can serve as an academic tool, while private data is available only to an owner or authorized users. All data is frequently updated. Most of it comes from the literature, but more accurate data is sometimes collected directly from researchers. The platform also has a group comparison option, which is a set of analytic tools enabling a user to compare the clinical and genomic characteristics of the user-defined samples. For mRNA and miRNA expression studies, the program uses a Z-score value that indicates the number of standard deviations from the average expression in the reference population. This parameter is useful to determine whether the gene is adjusted up or down from normal samples or all other tumor samples. RNASeqV2 results from the TCGA are processed and normalized using RSEM, which is a software package for estimating gene and isoform expression levels based on RNA-seq data. To check the miRNA data in the portal, either mature or precursor miRNAs can be entered. Then, the portal creates one internal ID for each pair of ID precursors and maps the mature IDs, which will also be displayed in Oncoprint. The open-source software is handled by an external company, Hyve, which provides commercial support for cBioPortal. It helps with implementation, data loading, development, advice, and training [21].

Like any similar portal, cBioPortal has several disadvantages. Firstly, installing the local version requires system management skills, including the installation and configuration of Tomcat and MySQL, in which no technical support from the site is provided. Secondly, for many studies, only somatic mutation data and limited clinical data are available, which may not deliver the full knowledge needed. For the TCGA studies, different types of data are available than for other studies. Also, germline mutations are supported by the cBioPortal, but with a few exceptions they are not available for public use. Additionally, the portal does not store raw or probe-level data, it only contains data at the gene level. The data for the different isoforms of a given gene is combined, and raw data and probe level data are available through the Gene Expression Omnibus (GEO) and the Genotypes and Phenotypes (dbGaP) from the National Center for Biotechnology Information (NCBI), or the Genomic Data Commons (GDC), from National Cancer Institute (NCI). Another disadvantage is that the cBioPortal does not store any expression data from normal tissue samples, which precludes some useful comparisons. Finally, the database has miRNA data for only a few tests that are not up-to-date. To download more recent miRNA data, the cBioPortal refers to Broad Firehose or GDC [21–23].

How to read the RNA maps?

The growing number of RNA-seq data is not correlated with the expertise of specialists who are able to analyze and interpret the results as well as generate their graphical presentation. In spite of the popularity of the R programming language, frequently used with tools from Bioconductor [24, 25], and the Python programming language with SciPy library [26], analysis of data is problematic. Most scientists learn these coding languages on their own, mostly by trial and error. However, in the majority of cases, R and Python programs are complicated for biologists and a lot of beginners quickly give up. Due to the problematic nature of data analysis and the constant increase in the capacity of transcriptomic data, on-demand analysis portals have emerged. Most of them are completely free of charge and enable statistical analysis by generating results in a graphical form. Usually, these portals use data derived from the TCGA project, but in some cases uploading their own data set is possible. The advantages and disadvantages of databases with RNA expression data are summarized in Table 1.

UALCAN

UALCAN is an interactive web portal for quick and easy evaluation of publicly available cancer transcriptome sequencing data to identify therapeutic targets and cancer biomarkers. Additionally, UALCAN helps to visualize the data in a simple, downloadable format. This portal uses TCGA level 3 RNA-seq and clinical data from 31 cancer types, and enables researchers to analyze a relative expression of an investigated gene across tumor and normal samples. Moreover, it allows browsing subgroups of cancers depending on the pathological stage, patient race, tumor grade and other clinicopathologic features. This information can be crucial to correlate gene expression with patient survival.
Data analysis: The UALCAN analysis page contains three panels. The left side panel shows a list of cancer types, which are hyperlinked to a web page showing heatmaps of top differentially expressed genes. The upper-right side panel allows a user to look through official gene symbols and types of cancer of interest, while the bottom-right side panel guides the user by gene classes The clinical patient data and gene expression were collected from the TCGA and processed to generate three major types of graphical outputs, described as follows: 1) Box and whisker plot showing gene expression level in different cancers and their subtypes/sub-stages. Primary tumor samples were categorized on clinical patient data and box plots were generated of the expression level of each gene across various subgroups. The categories of boxplots are: individual cancer stages, tumor grade, patient age, race, gender, body weight, smoking status, drinking habit, and molecular signature; 2) Heatmap showing

differential gene expression between normal and tumor samples; and 3) Kaplan–Meier survival plot [27].

ENCORI

ENCORI is an open-source platform and an updated version of Starbase, which contains information about the miRNA-ncRNA, ncRNA-RNA, miRNA-mRNA, RNA binding protein-ncRNA (RBP-ncRNA), RNA-RNA, and RBP-mRNA interactions from CLIP-seq, degradome-seq, and RNA-RNA interactome data. That platform contains data on 32 types of cancers derived from 10,882 RNA-seq and 10,546 miRNA-seq data. Moreover, it facilitates the discovery of miRNA–target interaction maps from CLIP-seq and degradome-seq data from six organisms: human, mouse, *Caenorhabditis elegans, Arabidopsis thaliana, Oryza sativa,* and *Vitis vinifera*.

Data analysis: At the upper part of that page users can choose from many the interfaces to display RNA interactions. After that, they can either enter a gene symbol or select a gene from the list box. To increase the quality of a search, users can apply filters such as interaction number, genome, or clade. A connected link launches external resources such as NCBI, University of California Santa Cruz (UCSC) and The Arabidopsis Information Resource (TAIR) databases, from which one can obtain more comprehensive information. Users are also allowed to download, analyze, and visualize large-scale data sets [28, 29].

The Kaplan-Meier Plotter (KM)

The Kaplan-Meier Plotter is a database that provides detailed expression comparisons of interesting genes in cancer and normal tissue and helps the researcher to assess the correlation between 54,000 genes expression and survival in 21 cancer types. The largest datasets include breast, lung, ovarian, and gastric cancer. The miRNA subsystems contain 11,000 samples from 20 different cancer types. Gene expression data and overall survival information are downloaded from the European Genome-phenome Archive (EGA), GEO, and TCGA. To analyze the prognostic value of a chosen gene, patient samples are split into two groups according to various quantile expressions of the proposed biomarker. The two patient cohorts are compared using a Kaplan-Meier survival plot, and the hazard ratio with 95% confidence intervals and log-rank p-value are calculated.

Data analysis: Specifically, the web page consists of four panels: 'protein', 'miRNA', 'RNA-seq', and 'mRNA gene chips'. To perform survival analysis, users simply choose one of them and type in the gene symbol they are interested in. They can also set required analysis

parameters to increase the accuracy of it. The output of gene expression levels in the different groups of patients and cell lines is shown according to the clinical or pathological status. The methodology behind the generation of survival curves is rather unclear. The „How does it work?" section of the website informs that a log-rank p-value is provided which suggests that it is calculated using the log-rank test. On the other hand, the reference cited as a main source of information regarding the tool describes the Cox regression model, called also Cox's proportional hazards model, which could be used as the equivalent of a log-rank test. Both of these methods assume proportional hazards. Interestingly, an alternative, Gehan-Breslow-Wilcoxon test, during calculation associates early time-point deaths with higher weights. Results of the KM-plotter analysis tool include the Kaplan–Meier plot containing the hazard ratio and p-value without defined information regarding the employed statistical test [30–32].

Oncomine

Oncomine is a cancer-profiling database that integrates microarray data analysis with other resources, including gene ontology annotations, to facilitate rapid interpretation of a genes' potential role in cancer. This website currently hosts gene expression and sample data from 500 cancer types and a wide quantity of cancer-related cell lines. It is made not only for individual researchers, but also for multinational companies, and it enables them to visualize and assess the diversified expression of a selected gene across all available datasets. After choosing a gene of interest, users receive a list of differential expression analyses in which the gene was included. For the chosen analyses, the statistical results are provided and linked to graphical representations of microarray data.

Data analysis: Oncomine data can be searched by gene, multiple genes or by a number of cancer-related terms. The visualizations are then classified by data samples, such as result, Gleason's grade, molecular change, stage, treatment, survival and relapse, to identify genes that are deregulated in particular cancer subtypes or that are related to specific clinical or pathological parameters. Such solutions make this database very effective in studying genes that are overexpressed in many cancer datasets to verify the relationship between transcription and disease. Many studies regarding the cancer microarray focus on identifying potential therapeutic targets or diagnostic markers. To achieve this, genes are presented with their ontological descriptors. By collecting information from the Gene Ontology (GO) consortium, three ontological categories have been created: 1) kinases which could be inhibited by small molecule kinase inhibitors; 2) membrane-bound molecules which could be targeted by antibody therapies; and 3) secreted molecules which could serve as serum biomarkers.

Finally, it needs to be mentioned that Oncomine integrates information from other bioinformatic resources, including Swiss-Prot, Unigene, and LocusLink as well as provides direct links to Human Protein Reference Database (HPRD) or the pathway resources Kyoto Encyclopedia of Genes and Genomes (KEGG) [33, 34].

Cancer-specific circRNAs database (CSCD)

Cancer-specific circular RNAs (circRNAs) database (CSCD) was created to collect and describe these molecules that have been associated with particular malignancies. CSCD gathered available RNA-seq datasets (total RNA with rRNA depleted or poly A- enriched) from 87 cancer cell line samples. The updated CSCD version 2.0 is created by over 1000 samples which include 825 tissues and 288 cell lines. An integrated analysis was performed using 4 popular algorithms: circRNA identifier (CIRI), circRNA_finder, find_circ, and circexplorer. In CSCD, 272,152 cancer-specific circRNAs were deposited. Currently, in CSCD 2.0, there are 1,013,461 circRNAs from cancer samples, 1,533,704 circRNAs from normal samples, and 354,422 circRNAs from both types of samples. This database provides the following components which could make a significant contribution to circRNA research in cancer: 1) miRNA target sites in cancer-specific circRNAs (CS-circRNAs); 2) CS-circRNAs; 3) RBPs binding sites in CS-circRNAs; 4) Cancer-specific alternative splicing associated with CS-circRNAs; 5) Potential Open Reading Frames (ORFs) in CS-circRNAs.

Data analysis: The main page of CSCD is composed of three panels: 1) Query Panel in which users can search circRNAs ID by selecting sample name, sample type, gene symbol, and cellular localization and search circRNA ID in a searching box. The output list displays all information for each cancer-specific circRNA, including genomic coordinates, host gene, location, circbase ID, annotation, spliced exon, ratios (circRNA/linear gene), prediction algorithm, and abundance. By selecting a gene symbol, users are redirected to the Gene Panel with all circRNAs in different samples; 2) Gene Panel to view all circRNAs for a selected gene. The image displays gene structures with different colors of rectangles for exons and black lines for introns, while circRNAs are shown as colored curves. To obtain a zoom of the high resolution image, the upper right corner of the panel should be selected. Users can also view detailed information about cancer-specific circRNA, including algorithm, junction reads, location, and sample date by clicking the "CircRNA". Other information can be found in the transcript tab, gene tab, circRNA tab, and splicing tab. 3) CircRNA Panel which visualizes specific circRNA by connecting exons in a colored circle. It enables users to view the position and number of RBP, microRNA response element (MRE), and open reading frame (ORF)

elements. Then, they can check the detailed information through the RBP, circRNA, MRE, and ORF tabs [35–37].

YM500 portal

      YM500 is a database that includes more than 8000 cancer-related small RNA (smRNA) datasets, arm-switching discovery, integrated pipelines for novel miRNA prediction, miRNA quantification from smRNA-seq, isomiR identification and functional sncRNAs in smRNA-seq datasets. Researchers can gain miRNA-related information from 141 mouse and 468 human smRNA-seq datasets by using a user-friendly web interface. Data analysis: YM500 provides interactive query interfaces, such as Expression, isomiRs, Novel miRNAs, arm-switching, Meta-analysis, Survival, Cancer, Result, Download, or Documentation. The analysis of data is divided into main blocks: 1) expression: this feature helps users search for miRNAs or sncRNA according to ID lists or miRNA cluster definitions. Several statistical charts are added to help researchers depict the expression profile of a given sncRNA across distinct cancer types. They can explore the Tissue Expression, Histogram, or Cancer Expression in the Expression profile or browse for an individuals' cancer boxplots; 2) survival: this section has two features: 'All Cancer Types' and 'Specific Sample Group'. The first one presents the survival analysis of a specific sncRNA (or miRNA) in all types of cancer, including a table summarizing all cancers and a Kaplan-Meier plot for each individual cancer type. The 'Specific Sample Group' helps researchers to identify a subset of samples in one type of cancer, such as triple negative breast cancer, in order to perform survival analysis according to dozens of clinical characteristics; 3) cancer: this section contains a calculated outcome of differential expression analyses, miRNA–gene interactions, and cancer miRNA-related pathways for a specific type of cancer that includes smRNA-seq and RNA-seq data of tumor and normal tissues for the same individuals. It also provides information on "individual miRNA-gene-specific pairs" to help scientists study the interactions between miRNAs and genes according to user-defined criteria. To illustrate the relationships between multiple miRNA-genes interactions, users can apply the Cytoscape Web tool for interactive web visualization. The correlations of each pair of miRNA-genes between differently expressed miRNAs and genes are calculated and classified into three groups: "Predicated", "Validated" and "Without any evidence"; 4) meta-analysis: it is the new function that allows researchers to identify differentially expressed miRNAs and arm-switching events from two user-defined, specific sets of samples. Users can pick one or more datasets and define a sample type they would like to study, such as "Primary Solid Tumor" or "Solid Tissue Normal". This subpage

also includes a list of clinical criteria, such as tumor stage, distant metastases, and lymph node status, to help scientists select a subset of well-defined cancer samples according to one or more clinical parameters; 5) novel miRNAs: they can be searched according to the exact sequence of a mature miRNA or genomic location. Users gain information about provided mature and precursor miRNA, including predicted target genes, prediction algorithms, RNA secondary structures, expression profiles, and hyperlinks to three commonly used genome browsers; 6) isomiRs: this section enables researchers to find the isomiRs of the known miRNAs in miRBase. To that end, users shall define the criteria of isomiRs according to the number of read counts, the number of mismatches, the number of expressed samples, and isomiR types (addition or trimming at 5′/3′ end); 7) arm switching: this option provides two ways to study the process. YM500 allows users to select a specific miRNA precursor and profile the expression of two arms between samples and tissues. Thus, a quick review of arm-switching events in a specific miRNA species can be done. Using another illustration method, users can select two groups of samples, as annotated in the database, and YM500 identifies miRNA precursors miRNAs, whose dominant expression switches from one arm to the other between the two groups [38-41].

OncoMir Cancer Database (OMCD)

The OncoMir Cancer Database allows a simple, systematic and comparative analysis of miRNA expression sequencing data derived from more than 10,000 cancer patients. It also provides clinical information and organ-specific controls present in the TCGA database. Control *vs*. control, tumor *vs*. control, and tumor *vs*. tumor statistical analyses are included in the OncoMir Cancer Database to generate clear and defensible hypotheses. The OMCD database contains miRNA profiling information from samples representing 33 different sets of tumors and corresponding to normal tissue data, if available.

Data analysis: Users can search for miRNAs by name or type of cancer to get a list of important relationships between miRNAs and malignancy. In addition, this search interface can be used to retrieve RNA seq data of one or more miRNAs across multiple types of cancer to gain a detailed view of cancer-dependent miRNA expression profiles. Users can also search for miRNA statistical results based on user-defined thresholds or search for miRNA statistical results by specific miRNA [42–44].

BioXpress

BioXpress is a database for gene expression and cancer associations, where expression levels are mapped to genes using RNA sequence data from the ICGC, TCGA, Expression Atlas and publications. It includes expression data from 6,361 patients, 64 cancer types and 17,469 genes with 9,513 of the genes displaying differential expression between tumor and normal samples. All types of cancer are mapped to Disease Ontology terms to facilitate a uniform pan-cancer analysis.

Data analysis: Users can search BioXpress using UniProtKB/ RefSeq accession, HUGO Gene Nomenclature Committee gene symbol, or RefSeq accessions. Moreover, differentially expressed genes for a specific cancer type can be also retrieved. Query submission redirects to the results page, organized into the chart and table components. The default chart for a BioXpress transcription search lists the frequencies of patients corresponding to each expression trend (under- or overexpression) for the queried transcript across all relevant cancer types. Here, patients with log2 fold change values higher than zero are considered to follow an overexpression trend, while those with less than zero follow an under-expression trend. The second chart illustrates the percentage of patients whose individual expression trend matches the significant trend reported for the queried transcript across different cancer types, with two colored series denoting different thresholds. The third chart shows a box plot of tumor sample expression, involving expression from those samples with matched normal data and all unpaired tumor samples from the TCGA. In addition to expression values and various statistics, the hit table contains a variety of identifiers and hyperlinks to relevant BioMuta entries and other external resources. As an alternative, the advanced cancer type search enables users to combine multiple search terms, including transcript and cancer types, in one query while simultaneously filtering hits by trend and significance threshold. The default view for the cancer type search displays the significance of the patient's expression for the top 20 transcripts matching the queried cancer type and meeting other search criteria. All data in BioXpress can be downloaded, including a list of genes that differ significantly in two or more types of cancers [45–47].

Table 1. Advantages and disadvantages of databases with RNA expression data

Database	Advantages	Disadvantages	Link to database
The cBio Cancer Genomics Portal	The portal complements existing tools (ICGC and TCGA, IGV) by	Does not currently support	http://www.cbioportal.org/

	offering a unique focus on analyzing discrete genomic events across integrated data types. The database contains a tutorial and allows users to download or visualize data.	synonymous mutations. Does not include information on normal tissues.	
UALCAN	Included data is helpful in identifying candidate biomarkers. Obtained information (box plots, KM-plots, and heatmaps) can be printed directly. The database contains a tutorial and allows users to download or visualize data. The database contains information on normal tissues.	A lot of information is required to be collected from external links.	https://ualcan. path.uab.edu/
ENCORI	The database contains the miRNA-pseudogene interaction networks and miRNA-lncRNA interactions. Drafts interaction maps between miRNAs and circRNAs. ENCORI contains a tutorial and allows users to download or visualize data. The database contains information on normal tissues.	Majority of CLIP-Seq and Degradome-Seq clusters could not be clearly predicted to be miRNA targets. The availability is restricted to selected hours.	https://rnasysu. com/encori/
The Kaplan-Meier Plotter	The database determines if a given gene is a potential marker, oncogene, or tumor suppressor. It contains a tutorial and allows users to download or visualize data.	Microarrays and RNA-seq datasets are not combined.	https://kmplot. com/analysis/

	The database contains information on normal tissues.		
Oncomine	Premium account allows expanded analyses, multi-gene search and custom concept upload. The database contains a tutorial and allows users to download or visualize data. It contains information on normal tissues.	User account is required to perform the analyses.	http://www.oncomine.org
Cancer-specific circRNAs database	The database allows users to determine whether circRNAs are cancer-specific. Users can admire potential functional regulation and translation of circRNAs by predicting the MRE, RBP, and ORF. The website contains a tutorial and allows users to download or visualize data. The database contains information on normal tissues.	The database does not include any primary tumor samples.	http://geneyun.net/CSCD2
YM500	The website provides novel findings including IsomiRs, novel miRNAs, or miRNA expression. The database contains an extensive number of smRNA-Seq datasets of various tissue/cell types and information on arm-switching. The portal contains a tutorial and allows users to download or	Some of comparisons depicted on charts lack p-values and description of employed statistical tests.	http://140.128.69.154:8787/ym500v3/index.php

	visualize data. It contains information on normal tissues.		
OncoMir Cancer Database	The database allows users to analyze miRNA-derived survival outcome signatures for many cancer types or to analyze miRNA clusters. Signature on predicting overall cancer survival is presented through a Kaplan-Meier survival curve. The website allows users to download or visualize data. The database contains information on normal tissues.	Some recently identified new miRNAs may not be available. The database does not contain a tutorial.	https://www.oncomir.umn.edu/omcd/
BioXpress	Unification by a central cancer disease ontology. The database allows users to download or visualize data. It contains information on normal tissues.	The database does not contain a tutorial.	https://hive.biochemistry.gwu.edu/bioxpress/transcriptview/

ICGC — International Cancer Genome Consortium; TCGA — The Cancer Genome Atlas; IGV — Integrative Genomics Viewer (tool for the visual exploration of genomic data); circRNA — circular RNA; MRE — microRNA response element; RBP — RNA-binding protein; ORF — open reading frame; miRNA — microRNA; smRNA — small RNA

Conclusions and future directions

In this review we presented and discussed the commonly used data storage and analysis portals, such as cBioPortal, UALCAN, ENCORI, and others. In our opinion, biomarkers based on various RNA molecules, both coding and non-protein coding, will contribute to the personalization of oncology. However, there is a need to describe these transcriptome profiles and link them with the clinicopathological data of patients. A huge

limitation to achieving this goal is the lack of high-throughput data analysis skills acquired during university classes. Most transcriptomics researchers do not have bioinformatics training. For them, the fastest option is to use websites where quick data analysis is possible. However, these databases also have many disadvantages. The first one being that portals based on the TCGA data are not connected with each other. There is no possibility of analyzing the same samples in the context of, for example, miRNA:lncRNA nor miRNA:mRNA. Moreover, the selection of the samples by their ID number is also impossible. What is more, there is no ability to select nor exclude specified patients using their clinical-pathological parameters. It should be noted that no expression level for healthy tissue for some databases or datasets can be found as well as there is no ability for analysis of the adjacent matched healthy and cancer samples from the same patient. This problem can be solved only by downloading data from the Xena Platform [48].

Inconsistent gene nomenclature is the second issue. Unfortunately, some authors still use outdated synonyms or improper nomenclature and unification of gene names used in databases with the HUGO Gene Nomenclature Committee (HGNC) (https://www.genenames.org/) should be a good practice. This portal sets the standards for human gene nomenclature and is organized by the HUGO [49]. The HGNC portal gives tools like: i) BioMart server, which is the filter for selection of proper data from a large set of data; ii) HCOP tool for orthology predictions search and iii) Multi-symbol checker which is for comparison of genes' symbols. It should be noted that HGNC enables grouping genes according to gene types or loci [49]. In our opinion, HUGO should obligate the authors of portals to standardize gene names in accordance with the guidelines, and the authors of newly created portals should follow the directions. This will allow the data to be organized and mistakes and understatements to be avoided.

The next issue is that the different normalization methods in specified databases were applied [50]. However, in the case of the TCGA-based websites there is no possibility of changing the normalization and the only solution for the user is to download raw RNA-seq data from the Xena Platform and normalize it by themselves.

It should be also emphasized that it is not always clear what statistical test was used. In most cases information about it is not included on the database webpage, the reference is not clearly explained or there is only a note about the script used, which is not understandable for users without programming knowledge. Moreover, there is often no statistical analysis presented as an expotrable table, nor a possibility of graphic modification of figures, for example as on the UALCAN database [27].

The last issue is that some databases, such as ENCORI, have limited access due to a lot of users and server overload, or full access is paid as in the case of Oncomine. Despite these inconveniences, it is possible to perform preliminary data analysis, visualize the results, download the data from the Xena Platform, and analyze it using other tools.

Different types of RNA databases are created and choosing a proper one is one of the keys to success. For this reason, the visiting of Database Commons (DC) portal is recommended. This large catalog of biological databases has been updated since 2015 and gives much information about databases and opportunities for scaling their reliability and usefulness, for example the DC portal described 16 active databases connecting with the TCGA project [51]. In the case of testing resistance to therapeutics, databases have been created and are being developed to determine the effectiveness of a given molecule and to identify individual molecular pathways involved in the metabolism or resistance to a given drug. However, there are no databases similar to those discussed above in the context of chemotherapy drugs [52–54]. Moreover, to our knowledge, there are no RNA-seq databases enabling rapid analysis of genetic changes in a cell exposed to ionizing radiation. Such a database would expand knowledge about radiogenomics and accelerate the introduction of biomarkers in radiotherapy-based therapy [55–57].

Finally, we can conclude that despite described imperfections and inconveniences regarding RNA databases, these databases enable crucial preliminary analyses for testing hypotheses, in an application for grants as well as for cancer research.

Conflict of interests

Funding

Sciences in Poznań. Joanna Kozłowska-Masłoń received a PhD program scholarship in the time of writing this manuscript from the Adam Mickiewicz University, Poznan; Kacper Guglas received a scholarship in the time of writing this manuscript from the European Union POWER PhD program and from the Medical University of Warsaw.

Authors' contributions

Authors' individual contributions: conceptualization: T.K., A.T.; writing — original draft preparation: M.K., J.L., J.O., P.P., M.R., P.P., J.K.M., K.G., K.D., N.G., K.R., A.F., U.K., T.K.; writing — review and editing: T.K., U.K., A.T., K.R.; visualization: T.K., M.K., J.L., P.P.; supervision: T.K., A.T.; funding acquisition: A.T., K.G., K.L. All authors read and approved the final manuscript.

Availability of data and materials

The datasets used during the current study are available online from described websites and from PubMed.

Ethics approval

All data is available online, access is unrestricted. The use of the data does not violate the rights of any person or any institution.

Consent for publication

All authors read and approved the final manuscript.

References

1. Li F, Wu T, Xu Y, et al. A comprehensive overview of oncogenic pathways in human cancer. Brief Bioinform. 2020; 21(3): 957–969, doi: 10.1093/bib/bbz046, indexed in Pubmed: 31155677.

2. Pompella L, Tirino G, Pappalardo A, et al. Pancreatic Cancer Molecular Classifications: From Bulk Genomics to Single Cell Analysis. Int J Mol Sci. 2020; 21(8), doi: 10.3390/ijms21082814, indexed in Pubmed: 32316602.

3. Mercatelli N, Palombo R, Paronetto MP. Emerging Contribution of PancRNAs in Cancer. Cancers (Basel). 2020; 12(8), doi: 10.3390/cancers12082035, indexed in Pubmed: 32722129.

4. Heilig M, Sommer W. Functional genomics strategies to identify susceptibility genes and treatment targets in alcohol dependence. Neurotox Res. 2004; 6(5): 363–372, doi: 10.1007/BF03033310, indexed in Pubmed: 15545019.

5. Wilusz JE, Sunwoo H, Spector DL. Long noncoding RNAs: functional surprises from the RNA world. Genes Dev. 2009; 23(13): 1494–1504, doi: 10.1101/gad.1800909, indexed in Pubmed: 19571179.

6. Braicu C, Buiga R, Cojocneanu R, et al. Connecting the dots between different networks: miRNAs associated with bladder cancer risk and progression. J Exp Clin Cancer Res. 2019; 38(1): 433, doi: 10.1186/s13046-019-1406-6, indexed in Pubmed: 31665050.

7. Braicu C, Zimta AA, Harangus A, et al. The Function of Non-Coding RNAs in Lung Cancer Tumorigenesis. Cancers (Basel). 2019; 11(5), doi: 10.3390/cancers11050605, indexed in Pubmed: 31052265.

8. Guglas K, Kozłowska-Masłoń J, Kolenda T, et al. Midsize noncoding RNAs in cancers: a new division that clarifies the world of noncoding RNA or an unnecessary chaos? Rep Pract Oncol Radiother. 2022; 27(6): 1077–1093, doi: 10.5603/RPOR.a2022.0123, indexed in Pubmed: 36632289.

9. Palazzo AF, Lee ES. Non-coding RNA: what is functional and what is junk? Front Genet. 2015; 6: 2, doi: 10.3389/fgene.2015.00002, indexed in Pubmed: 25674102.

10. Yan H, Bu P. Non-coding RNA in cancer. Essays Biochem. 2021; 65(4): 625–639, doi: 10.1042/EBC20200032, indexed in Pubmed: 33860799.

11. Goodall GJ, Wickramasinghe VO. RNA in cancer. Nat Rev Cancer. 2021; 21(1): 22–36, doi: 10.1038/s41568-020-00306-0, indexed in Pubmed: 33082563.

12. Stasiak M, Kolenda T, Kozłowska-Masłoń J, et al. The World of Pseudogenes: New Diagnostic and Therapeutic Targets in Cancers or Still Mystery Molecules? Life (Basel). 2021; 11(12), doi: 10.3390/life11121354, indexed in Pubmed: 34947885.

13. Kozłowska J, Kolenda T, Poter P, et al. Long Intergenic Non-Coding RNAs in HNSCC: From "Junk DNA" to Important Prognostic Factor. Cancers (Basel). 2021; 13(12), doi: 10.3390/cancers13122949, indexed in Pubmed: 34204634.

14. Sobala A, Hutvagner G. Transfer RNA-derived fragments: origins, processing, and functions. Wiley Interdiscip Rev RNA. 2011; 2(6): 853–862, doi: 10.1002/wrna.96, indexed in Pubmed: 21976287.

15. Kolenda T, Guglas K, Baranowski D, et al. cfRNAs as biomarkers in oncology - still experimental or applied tool for personalized medicine already? Rep Pract Oncol Radiother. 2020; 25(5): 783–792, doi: 10.1016/j.rpor.2020.07.007, indexed in Pubmed: 32904167.

16. Kolenda T, Guglas K, Kopczyńska M, et al. Quantification of long non-coding RNAs using qRT-PCR: comparison of different cDNA synthesis methods and RNA stability. Arch Med Sci. 2021; 17(4): 1006–1015, doi: 10.5114/aoms.2019.82639, indexed in Pubmed: 34336028.

17. Kolenda T, Paszkowska A, Braska A, et al. Host gene and its guest: short story about relation of long-noncoding transcript and microRNA . Rep Pract Oncol Radiother. 2023; 28(1): 114–134, doi: 10.5603/RPOR.a2023.0006, indexed in Pubmed: 37122913.

18. Kozłowska-Masłoń J, Guglas K, Kolenda T, et al. miRNA in head and neck squamous cell carcinomas: promising but still distant future of personalized oncology. Rep Pract Oncol Radiother. 2023; 28(5): 681–697, doi: 10.5603/rpor.96666, indexed in Pubmed: 38179293.

19. Kolenda T, Poter P, Guglas K, et al. Biological role and diagnostic utility of ribosomal protein L23a pseudogene 53 in cutaneous melanoma. Rep Pract Oncol Radiother. 2023; 28(2): 255–270, doi: 10.5603/RPOR.a2023.0030, indexed in Pubmed: 37456695.

20. Wu P, Heins ZJ, Muller JT, et al. Integration and Analysis of CPTAC Proteomics Data in the Context of Cancer Genomics in the cBioPortal. Mol Cell Proteomics. 2019; 18(9): 1893–1898, doi: 10.1074/mcp.TIR119.001673, indexed in Pubmed: 31308250.

21. cBioPortal for Cancer Genomics. https://www.cbioportal.org/ (03.08.2023).

22. Gao J, Aksoy BA, Dogrusoz U, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal. 2013; 6(269): pl1, doi: 10.1126/scisignal.2004088, indexed in Pubmed: 23550210.

23. Cerami E, Gao J, Dogrusoz U, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. 2012; 2(5): 401–404, doi: 10.1158/2159-8290.CD-12-0095, indexed in Pubmed: 22588877.

24. Giorgi FM, Ceraolo C, Mercatelli D. The R Language: An Engine for Bioinformatics and Data Science. Life (Basel). 2022; 12(5), doi: 10.3390/life12050648, indexed in Pubmed: 35629316.

25. Sepulveda JL. Using R and Bioconductor in Clinical Genomics and Transcriptomics. J Mol Diagn. 2020; 22(1): 3–20, doi: 10.1016/j.jmoldx.2019.08.006, indexed in Pubmed: 31605800.

26. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020; 17(3): 261–272, doi: 10.1038/s41592-019-0686-2, indexed in Pubmed: 32015543.

27. Chandrashekar DS, Karthikeyan SK, Korla PK, et al. UALCAN: An update to the integrated cancer data analysis platform. Neoplasia. 2022; 25: 18–27, doi: 10.1016/j.neo.2022.01.001, indexed in Pubmed: 35078134.

28. Li JH, Liu S, Zhou H, et al. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. Nucleic Acids Res. 2014; 42(Database issue): D92–D97, doi: 10.1093/nar/gkt1248, indexed in Pubmed: 24297251.

29. Cai L, Zhou KR, Liu S et al. ENCORI Tutorial: cover Pan-Cancer analysis. https://rna.sysu.edu.cn/encori/img/tutorial/ENCORI_Tutorial.pdf (04.08.2023).

30. Zee J, Xie SX. The Kaplan-Meier Method for Estimating and Comparing Proportions in a Randomized Controlled Trial with Dropouts. Biostat Epidemiol. 2018; 2(1): 23–33, doi: 10.1080/24709360.2017.1407866, indexed in Pubmed: 30234179.

31. Győrffy B. Discovery and ranking of the most robust prognostic biomarkers in serous ovarian cancer. Geroscience. 2023; 45(3): 1889–1898, doi: 10.1007/s11357-023-00742-4, indexed in Pubmed: 36856946.

32. Kaplan Meier plotter. https://kmplot.com/analysis/index.php?p=background (03.08.2023).

33. Oncomine. https://www.oncomine.org/resource/login. (03.09.2023).

34. Rhodes DR, Yu J, Shanker K, et al. ONCOMINE: a cancer microarray database and integrated data-mining platform. Neoplasia. 2004; 6(1): 1–6, doi: 10.1016/s1476-5586(04)80047-2, indexed in Pubmed: 15068665.

35. Xia S, Feng J, Chen Ke, et al. CSCD: a database for cancer-specific circular RNAs. Nucleic Acids Res. 2018; 46(D1): D925–D929, doi: 10.1093/nar/gkx863, indexed in Pubmed: 29036403.

36. Feng J, Chen W, Dong X, et al. CSCD2: an integrated interactional database of cancer-specific circular RNAs. Nucleic Acids Res. 2022; 50(D1): D1179–D1183, doi: 10.1093/nar/gkab830, indexed in Pubmed: 34551437.

37. Cancer-Specific CircRNA Databse 2.0. http://geneyun.net/CSCD2 (03.08.2023).

38. Cheng WC, Chung IF, Huang TS, et al. YM500: a small RNA sequencing (smRNA-seq) database for microRNA research. Nucleic Acids Res. 2013; 41(Database issue): D285–D294, doi: 10.1093/nar/gks1238, indexed in Pubmed: 23203880.

39. Cheng WC, Chung IF, Tsai CF, et al. YM500v2: a small RNA sequencing (smRNA-seq) database for human cancer miRNome research. Nucleic Acids Res. 2015; 43(Database issue): D862–D867, doi: 10.1093/nar/gku1156, indexed in Pubmed: 25398902.

40. Chung IF, Chang SJ, Chen CY, et al. YM500v3: a database for small RNA sequencing in human cancer research. Nucleic Acids Res. 2017; 45(D1): D925–D931, doi: 10.1093/nar/gkw1084, indexed in Pubmed: 27899625.

41. YM500v3 — A smRNA-Seq Database for Cancer Research. http://140.128.69.154:8787/ym500v3/index.php (03.08.2023).

42. Sarver AL, Sarver AE, Yuan Ce, et al. OMCD: OncomiR Cancer Database. BMC Cancer. 2018; 18(1): 1223, doi: 10.1186/s12885-018-5085-z, indexed in Pubmed: 30522456.

43. OncoMir Cancer Database
   (OMCD). https://www.oncomir.umn.edu/omcd/basic_search. (03.08.2023).

44. Wong NW, Chen Y, Chen S, et al. OncomiR: an online resource for exploring pan-cancer
   microRNA dysregulation. Bioinformatics. 2018; 34(4): 713–715,
   doi: 10.1093/bioinformatics/btx627, indexed in Pubmed: 29028907.

45. Dingerdissen HM, Torcivia-Rodriguez J, Hu Yu, et al. BioMuta and BioXpress: mutation and
   expression knowledgebases for cancer biomarker discovery. Nucleic Acids Res. 2018;
   46(D1): D1128–D1136, doi: 10.1093/nar/gkx907, indexed in Pubmed: 30053270.

46. Wan Q, Dingerdissen H, Fan Yu, et al. BioXpress: an integrated RNA-seq-derived gene
   expression database for pan-cancer analysis. Database (Oxford). 2015; 2015,
   doi: 10.1093/database/bav019, indexed in Pubmed: 25819073.

47. The HIVE Lab. https://hive.biochemistry.gwu.edu/bioxpress/transcriptview/ (03.08.2023).

48. UCSC Xena. https://xenabrowser.net/datapages/ (03.08.2023).

49. HGNC Gene Data. https://www.genenames.org/ (27.09.2023).

50. RNA-SEQ. https://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/ (27.09.2023
).

51. National Genomics Data Center. https://ngdc.cncb.ac.cn/databasecommons/ (10.10.2023).

52. Regulska K, Kolenda T, Michalak M, et al. Impact of ramipril nitroso-metabolites on cancer
   incidence - in silico and in vitro safety evaluation. Rep Pract Oncol Radiother. 2023; 28(5):
   612–622, doi: 10.5603/rpor.97433, indexed in Pubmed: 38179284.

53. Regulska K, Michalak M, Kolenda T, et al. Angiotensin-converting enzyme inhibitors for
   ovarian cancer? - a new adjuvant option or a silent trap. Rep Pract Oncol Radiother. 2023;
   28(4): 551–564, doi: 10.5603/RPOR.a2023.0059, indexed in Pubmed: 37795232.

54. Ganter B, Tugendreich S, Pearson CI, et al. Development of a large-scale chemogenomics
   database to improve drug candidate selection and to understand mechanisms of chemical
   toxicity and action. J Biotechnol. 2005; 119(3): 219–244,
   doi: 10.1016/j.jbiotec.2005.03.022, indexed in Pubmed: 16005536.

55. Zhu Y, Li H, Guo W, et al. Deciphering Genomic Underpinnings of Quantitative MRI-based
   Radiomic Phenotypes of Invasive Breast Carcinoma. Sci Rep. 2015; 5: 17787,
   doi: 10.1038/srep17787, indexed in Pubmed: 26639025.

56. Pavlopoulou A, Bagos PG, Koutsandrea V, et al. Molecular determinants of radiosensitivity
   in normal and tumor tissue: A bioinformatic approach. Cancer Lett. 2017; 403: 37–47,
   doi: 10.1016/j.canlet.2017.05.023, indexed in Pubmed: 28619524.

57. Andreassen CN, Schack LM, Laursen LV, et al. Radiogenomics — current status, challenges
   and future directions. Cancer Lett. 2016; 382(1): 127–136,
   doi: 10.1016/j.canlet.2016.01.035, indexed in Pubmed: 26828014.

Figure 1. Different elements of the RNA world. Percentage of total RNA in a mammalian cell
and an average length in kilobases (kb) for each RNA type. The small non-coding RNAs
(sncRNA) category, which consists mainly of micro RNA (miRNA), small nucleolar RNA
(snoRNA), small nuclear RNA (snRNA), and Single-cell RNA (scRNA), is further analyzed
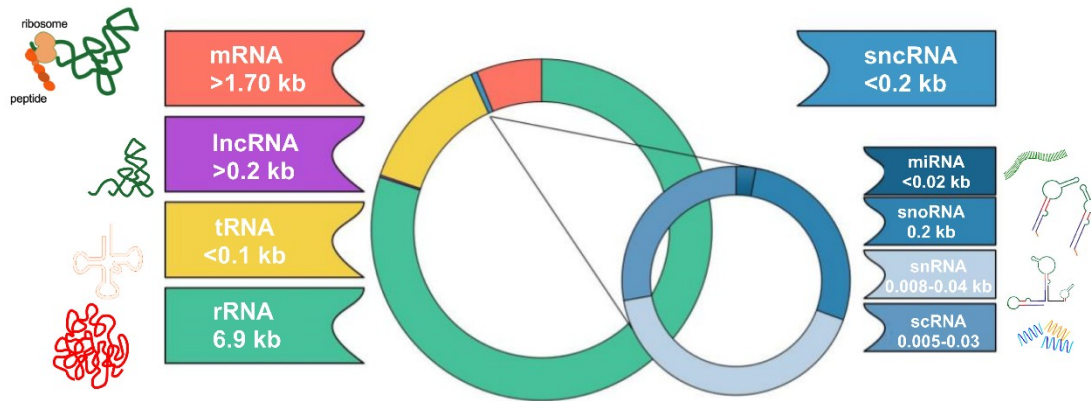on the blue graph on the right side

Figure 2. RNA molecules usefulness in biological and medical sciences. Development of computer science and analysis data portals caused the quicker and common uses of RNA-seq data in research for analysis of RNAs' interactions, Genes' expression pattern, co-expression and pathway analysis as well as in biomarkers discovery