Computers in Biology and Medicine 133 (2021) 104387

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: http://www.elsevier.com/locate/compbiomed

KnowSeq R-Bioc package: The automatic smart gene expression tool for retrieving relevant biological knowledge

Daniel Castillo-Secilla [a,*], Juan Manuel Gálvez [a], Francisco Carrillo-Perez [a], Marta Verona-Almeida [a], Daniel Redondo-Sánchez [b], Francisco Manuel Ortuno [c], Luis Javier Herrera [a,1], Ignacio Rojas [a,1]

[a] *Department of Computer Architecture and Technology,University of Granada. C.I.T.I.C., Periodista Rafael Gómez Montero 2, 18014, Granada, Spain*
[b] *Instituto de Investigación Biosanitaria de Granada, Non-Communicable Disease and Cancer Epidemiology Group, ibs.GRANADA, Avda. de Madrid, 15. Pabellón de Consultas Externas 2, 2a Planta, CP, 18012, Granada, Spain*
[c] *Clinical Bioinformatics Area, Fundación Andaluza Progreso y Salud (FPS), Hospital Universitario Virgen del Rocío, Avenida Manuel Siurot s/n, 41013, Sevilla, Spain*

ARTICLE INFO

Keywords:
Bioconductor
Gene expression
Classification
Enrichment
Bioinformatics

ABSTRACT

KnowSeq R/Bioc package is designed as a powerful, scalable and modular software focused on automatizing and assembling renowned bioinformatic tools with new features and functionalities. It comprises a unified environment to perform complex gene expression analyses, covering all the needed processing steps to identify a gene signature for a specific disease to gather understandable knowledge. This process may be initiated from raw files either available at well-known platforms or provided by the users themselves, and in either case coming from different information sources and different Transcriptomic technologies. The pipeline makes use of a set of advanced algorithms, including the adaptation of a novel procedure for the selection of the most representative genes in a given multiclass problem. Similarly, an intelligent system able to classify new patients, providing the user the opportunity to choose one among a number of well-known and widespread classification and feature selection methods in Bioinformatics, is embedded. Furthermore, *KnowSeq* is engineered to automatically develop a complete and detailed HTML report of the whole process which is also modular and scalable. Biclass breast cancer and multiclass lung cancer study cases were addressed to rigorously assess the usability and efficiency of *KnowSeq*. The models built by using the Differential Expressed Genes achieved from both experiments reach high classification rates. Furthermore, biological knowledge was extracted in terms of Gene Ontologies, Pathways and related diseases with the aim of helping the expert in the decision-making process. *KnowSeq* is available at Bioconductor (https://bioconductor.org/packages/KnowSeq), GitHub (https://github.com/CasedUgr/KnowSeq) and Docker (https://hub.docker.com/r/casedugr/knowseq).

1. Background

During the last decade, the importance of the DNA sequencing studies has risen significantly due to the emergence of Next Generation Sequencing (NGS) and the decrease in prices of this technology in comparison with its predecessors. As a result, the amount of available data, both public and controlled, has grown exponentially. Nowadays, the use of parallel architectures such as computer clusters or GPUs is highly recommended for an appropriate and efficient processing of the raw NGS data.

Concretely, transcriptomic studies at gene expression level are fundamental to win the battle against genetic and multifactorial diseases such as cancer. This worrying disease is still the second cause of death worldwide, just behind cardiovascular disease. Nowadays, the main medical challenge lies in the development of early diagnosis and prognosis cancer detection mechanisms. Therefore, the search for biomarkers that allow for achieving an early diagnosis of cancer is essential when addressing this type of studies.

In this sense, the design of powerful bioinformatic tools that allow processing and extracting transcriptomic information from raw data becomes a key goal in this research area. Currently, there is a number of tools that combine the different steps and technologies involved in this

* Corresponding author.
E-mail address: cased@ugr.es (D. Castillo-Secilla).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.compbiomed.2021.104387
Received 24 November 2020; Received in revised form 5 April 2021; Accepted 5 April 2021
Available online 13 April 2021

scope, not only for R language but also disseminated along other languages or platforms. Concretely, for the R language, there are tools such as GEO2RNAseq which propose a pipeline for processing RNA-seq data from FASTQ files to gene expression analysis [1]. Furthermore, this package allows for downloading data automatically from the Gene Expression Omnibus (GEO) public repository [2]. Alternatively, another tool, called RobiNA, addresses all the necessary steps to analyze RNA-seq and Microarray data (quality control, filtering, analysis of differential gene expression, and visualization of results). This software is available for both R and Java languages, including a user interface [3]. Finally, the RNAseqR package also proposes a pipeline to process RNA-seq data that implements all the steps mentioned in the previous references but also implements a functional enrichment within the steps (GO enrichment and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis) [4]. These well-known R tools are just a small set within a larger amount of currently available RNA-seq pipelines. With the aim of providing a more in-depth state of the art, a summary of the most important tools, including the previous mentioned R packages, is shown in Supplementary Table 1. Additionally, the RNA-Seq pipeline steps addressed by each tool are specified.

Moreover, the utilization of intelligent system in bioinformatics is on the rise nowadays due to the possibility of extracting knowledge from a massive quantity of data. These help in the identification of optimal biomarker sets, allowing their assessment through the design of classification models, which assist in the diagnostic process on unseen samples. Analyzing the aforementioned most frequent tools for gene expression, the integration of predictive models within their methodologies has not been exploited enough yet.

As seen, most of the well-developed tools presented in the bibliography satisfactorily perform part of the standardized phases in the treatment of bioinformatic data. However, the accomplishment of a complete pipeline under a single tool, including Machine Learning (ML), has not yet been exploited. In order to bring light in this regard, this work presents a very powerful tool sustained by a complete and public Bioconductor R package to perform comprehensive and intelligent RNA-seq studies. In fact, the analysis can be initiated from different transcriptomic processed files (FASTQ, BAM, SAM, count) from which an automatic and precise extraction of the most representative Differentially Expressed Genes (DEGs) can be obtained. This process ends up with specific steps for ML assessment and DEGs knowledge enrichment. In particular, these functionalities and steps are focused on the assistance for the design of Clinical-Decision Support Systems (CDSS) [5]. In addition, *KnowSeq* has the possibility of performing an automatic complete HyperText Markup Language (HTML) report including the results and information of all the implemented steps, only requiring the expression matrix as input. Our tool is not exclusively thought to deal with the *Homo sapiens* cancer pathologies since it is also ready to support the analysis of any other genetic or multifactorial disease or species.

Essentially, *KnowSeq* is focused on RNA-seq because it is nowadays the most powerful and widespread genetic characterization technology for transcriptome. *KnowSeq* comprises part of the tools used in our previous studies/publications using RNA-seq data. Several cancer types were addressed such as breast cancer, skin cancer, leukemia and lung cancer and in all them relevant results were achieved [6–8]. They widely confirm the validity of the tools connected through *KnowSeq* in order to carry out genetic disease analysis working at gene expression level with raw data from RNA-seq.

This paper presents a real application of *KnowSeq* to two different study cases (Breast cancer and Lung cancer) with data coming from *Genomic Data Commons (GDC) Portal* [9]. On one hand, for the breast cancer study, 180 BAM files belonging to 90 breast cancer patients were used. For each patient, two samples were collected: a primary tumor sample from Ductal and Lobular Neoplasm and a solid tissue normal sample. Thanks to this, the experiment was designed with Tumor-Normal paired samples, which guarantees better experimental conditions in terms of sample availability. There is a number of previous breast cancer studies which extracted DEGs, without a posterior ML assessment [10,11]. Recently, following the increasingly-standardized ML step among the scientific community, Sun, D. et al. developed a Deep learning method based on SVM for the prognosis prediction of human breast cancer [12]. Furthermore, Wu, J. et al. gathered samples from TCGA to extract DEGs and assessed them by implementing different ML algorithms obtaining outstanding results [13]. Finally, our group has a previous study integrating microarray and RNA-seq data from NCBI/GEO to extract biomarkers and assess them through ML [6].

On the other hand, for the lung cancer study, samples coming from three different states were collected with the aim of carrying out a complete multiclass study in the search of relevant biomarkers. Those states are Adenocarcinomas (ACC), Squamous Cell Carcinomas (SCC) and Solid Tissue Normal (Control). Previous studies, focusing on the application of ML algorithms to the lung cancer DEGs assessment, Hu, F. et al., proposed an unsupervised classification system for different subtypes of lung adenocarcinoma [14]. Podolsky, M. D. et al. implemented ML algorithms to perform a binary classification between malignant pleural mesothelioma and adenocarcinoma [15]. Tian, S. developed a classification and survival system to predict for early-stage lung adenocarcinoma and squamous cell carcinoma patients by applying different statistical metrics and SVM [16]. Within our group, we also designed and implemented a multiclass lung cancer classification pipeline from microarray samples, to distinguish 4 sub-types of lung cancer [17]. For the presented experiment, 1100 count files were retrieved, starting the *KnowSeq* pipeline from the count files instead of the BAM files. Lung cancer was selected as it is the cancer with the largest number of available and balanced transcriptomic multiclass samples among the existing ones at GDC Portal, ensuring a relevant number of samples to carry out a rigorous assessment *KnowSeq*.

The rest of the paper is structured as follow: Section 2 contains an in-depth explanation of *KnowSeq* and the different functions and modules included within it. Section 3 exposes the results for the two study cases developed to assess *KnowSeq*. Section 4 summarizes important features of *KnowSeq*, current limitations and future implementations of our tool. Finally, Section 5 presents the study conclusions which highlight the main contributions of this work.

2. Methodology

This section presents in depth the steps and functionalities implemented by the *KnowSeq* pipeline. In order to visually comprehend the pipeline functionalities, Fig. 1 represents the outline of the whole methodology, split in 4 different distinguished steps: 1) Transcriptomic Raw Data Processing, 2) Biomarker Identification and Assessment, 3) DEGs Functional Enrichment, and finally, 4) Automatic Report. On this basis, these steps are organized in subsections with the purpose of giving a deeper explanation for each of them. It should be noted that *KnowSeq* has been designed to achieve a high modularity. This means that each of the steps and sub-steps conforming this tool could be perfectly replaced, taking into account that the inputs and outputs, using the same data type. Due to this, *KnowSeq* can be easily adapted even for different species and biological data types not explicitly addressed in our tool. In order to enumerate and summarize the different functions available in the package, Table 1 summarizes the most important functions included in *KnowSeq*, together with the corresponding pipeline step and a brief description of its functionality.

2.1. Transcriptomic Raw Data Processing

Raw data treatment is one of the most crucial steps in Transcriptomic studies. In order to gather the samples, *KnowSeq* brings the opportunity to automatize the download of public and controlled samples from the most renowned web platform databases: *Gene Expression Omnibus (GEO)*, *ArrayExpress* [18] and *GDC Portal*. Then, if the pipeline starts from SRA, FASTQ or BAM/SAM files, an alignment process will be

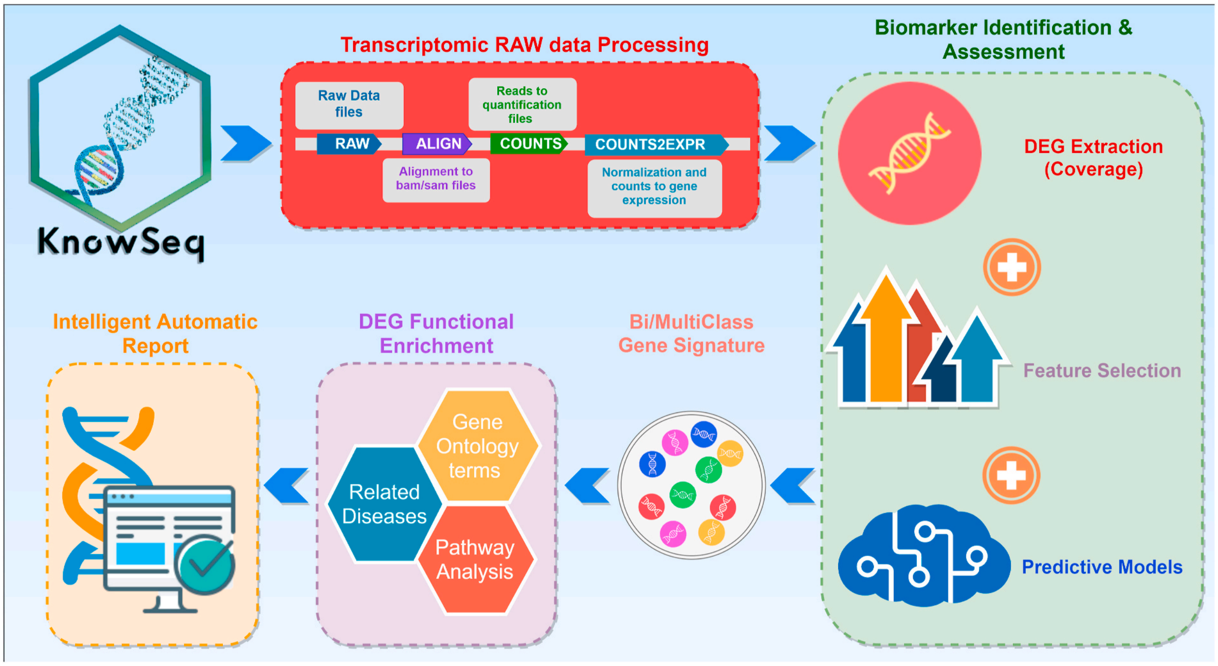

KnowSeq
Transcriptomic RAW data Processing
Raw Data files
Reads to quantification files
RAW
ALIGN
COUNTS
COUNTS2EXPR
Alignment to bam/sam files
Normalization and counts to gene expression
Biomarker Identification & Assessment
DEG Extraction (Coverage)
Feature Selection
Predictive Models
Intelligent Automatic Report
DEG Functional Enrichment
Gene Ontology terms
Related Diseases
Pathway Analysis
Bi/MultiClass Gene Signature

Fig. 1. Pipeline implemented by KnowSeq R/bioc package, considering the connection of the most standardized RNA-seq steps together with the novel steps introduced by our tool.

required by using the human reference genome in order to obtain the count files to perform the DEGs analysis. *KnowSeq* also allows for downloading the Human Reference Genome GRCh37 and GRCh38 from Ensembl, although whichever reference genome can be used by indicating the corresponding file path. This procedure is performed through the use of the *rawAlignment KnowSeq* function, for which our tool utilizes the *samtools* [19] and one of the most well-known aligners (*hisat2* [20]), providing the academics with the most renowned choices. Furthermore, the *Htseq-count* tool extracts the count files for each sample. It is highly recommended to run the raw data alignment in a computer cluster as the aforementioned tasks may be severely computationally demanding.[2]

When the raw files have been aligned, or in case the user starts the pipeline from the count files, the function *countsToMatrix* automatically combines all the count files into one aggregated matrix with *edgeR* [21]. Consequently, through the use of the function *calculateGeneExpressionValues*, the equivalent gene expression values are calculated using the *cqn* R package [22]. Although *KnowSeq* allows RNA-seq data to be processed from raw to counts, quantified as equivalent expression values, also other biological data which have been previously preprocessed and are quantified as an expression data matrix (for example, microarray or miRNA) are entirely compatible with Biomarkers Extraction and ML *KnowSeq* processes. Specifically, this has been verified by the authors through experimentation using microarray data from previous works [17] and through an in-progress miRNA research.

2.2. Biomarker Identification and Assessment

DEGs extraction is a very delicate process because the samples must pass a strong quality analysis and restrictive batch effect removal steps. If these steps are performed incorrectly, the DEGs candidates could not be true DEGs due to possible intrinsic deviations on the corrected gene expression from the considered samples. To solve that, *KnowSeq* has its own quality analysis step implemented inside the function *RNAseqQA*. During this analysis, outliers are detected using three different methods: distance between samples, Kolmogorov-Smirnov and MA plots. Furthermore, the user can decide if the outliers are only shown or also automatically removed. In this sense, they will only be removed if they previously appeared at least in two of the three outlier evaluation methods. To perform the quality analysis in a rigorous manner, it is crucial to ensure the correct development within the rest of the study. Even though the quality analysis is well-performed, there is still the possibility of having batch effects among the chosen samples or series. The batch effect is a deviation effect in the gene expression values due to several external technical factors (origin, sequencing time, lab technician, among others) and it is extremely hard to deal with [23]. *KnowSeq* allows for using two of the most relevant algorithms to remove batch effects such as ComBat for predefined batch groups, and *Surrogate Variable Analysis (SVA)* for unknown batch groups [24] through the function *batchEffectRemoval*. In addition, the user can decide to treat the batch effect correction step for a custom implementation.

With respect to the DEGs extraction, *limma R package* was included within *KnowSeq* as it is one of the most widespread methods in the literature [25]. *KnowSeq* allows limma DEGs extraction step to be replaced by other tools if the user wishes to do so [26]. It is important to note the difficulty to achieve true DEGs when there are more than two classes to be compared. For that, the Coverage (*COV*) parameter, which was introduced in our previous publication [8], has been included in the *KnowSeq* pipeline. *COV* allows for detecting DEGs that are differentially expressed in different classes, by counting the number of biclass comparisons where this differential expression takes place, thus controlling and improving the multiclass DEGs detection and the posterior multiclass ML assessment. Then, *KnowSeq* automatically detects the number of classes or labels for a given problem, and consequently applies the standardized limma method for biclass studies, or limma along with the *COV* parameter for multiclass DEGs extraction. It is important to note that *COV* parameter takes values between 1 and COV_{max} where COV_{max} is defined in Equation (1):

[2] With the aligner index previously computed, the complete human genome alignment for one RNA-seq sample took 65 min in a personal computer (MSI GP62VR 7RF Leopard Pro, i7-7820HK, 4 cores and 8 threads per core, 16 GB of RAM, 1 TB of SSD), while 18 min on average were required under a 4-node cluster use (with the following specification per node: 2 CPUs Intel Xeon Silver 4110 8c with 8 cores and 16 threads per core, 32 GB of RAM, 35 TB of NAS).

Table 1
Most relevant functions when using KnowSeq. For each of them, its name, the pipeline step to which the function belongs and its description with possible options are shown. The KnowSeq user manual contains an in-depth explanation for each function.

Function Name	Pipeline step	Description (options)
downloadPublicSeries	RNA-seq raw data processing	Automatically download series from GEO and AE
gdcClientDownload	RNA-seq raw data Processing	Automatically download data from GDC-Portal
rawAlignment	RNA-seq raw data processing	Transcriptomic raw data alignment with hisat2
countsToMatrix	RNA-seq raw data processing	Convert genes count files to matrix
calculateGeneExpressionValues	RNA-seq raw data processing	Gene expression values calculation and normalization
RNAseqQA	Biomarkers Identification & Assessment	Expression matrix Quality Analysis and Outlier Detection
getAnnotation	Biomarkers Identification & Assessment	Retrieve DEGs annotation from a given list
batchEffectRemoval	Biomarkers Identification & Assessment	Batch effect detection and removal (Combat, SVA)
DEGsExtraction	Biomarkers Identification & Assessment	Biclass and multiclass DEGs extraction
dataPlot	Biomarkers Identification & Assessment	Plots different data information and results (boxplot, orderedBoxplot, genesBoxplot, heatmap, confusionMatrix, classResults)
featureSelection	Biomarkers Identification & Assessment	Feature selection for a DEGs matrix (mRMR, RF, DARED)
knn_trn	Biomarkers Identification & Assessment	Train a k-NN by using Cross-Validation for a given DEGs matrix
knn_test	Biomarkers Identification & Assessment	Test a k-NN model for a given DEGs matrix
rf_trn	Biomarkers Identification & Assessment	Train a RF by using Cross-Validation for a given DEGs matrix
rf_test	Biomarkers Identification & Assessment	Test a RF model for a given DEGs matrix
svm_trn	Biomarkers Identification & Assessment	Train a SVM by using Cross-Validation for a given DEGs matrix
svm_test	Biomarkers Identification & Assessment	Test a SVM model for a given DEGs matrix
DEGsToDiseases	DEG Functional Enrichment	Retrieve related diseases and evidence for a DEGs list
geneOntologyEnrichment	DEG Functional Enrichment	Gene ontology enrichment for a DEGs list
DEGsToPathways	DEG Functional Enrichment	DEG Related KEGG pathways
knowseqReport	Automatic Report	Automatic and Modular HTML Report for a given dataset and disease or diseases

$$COV_{max} = \frac{N^2 - N}{2}, \quad \text{where } N \text{ is the number of classes.} \tag{1}$$

However, a value of *COV* near to COV_{max} is usually too restrictive while a low value might introduce DEGs with a poor multiclass discernment potential. Typically, a problem in which the classes present clear differences at the gene expression level enables the use of a high *COV* value, retrieving a reduced number of DEGs with the capability of discerning among the classes. Otherwise, when the differences are less evident, a lower *COV* value should be used to retrieve a sufficient number of DEGs with enough information to distinguish between the states. This DEGs extraction is carried out by using the function *DEGsExtraction*. Theoretically, the final DEGs candidates are genes with the capability to differentiate between the classes to be analyzed. In order to add a graphical assessment to the process, the function *dataPlot* includes the possibility of plotting all the required charts for that (e.g. boxplots by samples, boxplots by genes, heatmaps and others). Nevertheless, to really test and assess those hypothetical DEGs, a ML process is implemented in *KnowSeq*. This step is sub-divided in two substeps: a feature selection process and a supervised classification process. On the one hand, for a given number of candidate DEGs, the first substep applies a feature selection (FS) process which is highly recommended for precision medicine. This allows for reducing the system complexity, diminishing the number of genes, and helping to make clinical decisions [27]. For that, *KnowSeq* applies two different feature selection algorithms such as *minimum Redundancy Maximum Relevance* (mRMR) [28] and *Random Forest* as feature selector (RFfs), invoked by means of the function *featureSelection*. These algorithms create a ranking of DEGs in order to reduce the complexity of the classification models while keeping the intelligent classifier results by listing those DEGs with more discernment capability at the top.

The second step is the supervised classification process in which *KnowSeq* allows using three of the most well-known classifiers for this type of analyses: Support *Vector Machine* (SVM) [29], *k-Nearest Neighbour* (k-NN) [30] and *Random Forest* (RF) [31]. All of them are widely used to deal with genomic and transcriptomic studies by the scientific community [32–34], and any or all three of them can be combined with any of the FS algorithms and used to check the validity of the selected DEGs. It is important to note that if several classifiers are used, these can lead to slightly different results. The package implements both a training-test or training-validation subdivision and assessment, and a k-fold cross-Validation (CV) strategy, which means that the training dataset is split into k sub-training sets, leaving the rest of samples for validation. Thanks to the latter technique, all the training samples are used for both training and validation at least once, and an appropriate estimate of the expected performance can be obtained. CV will also be used in the experiments for the assessment of the optimal number of genes in the ranking extracted by the feature selection algorithms. It is important to note that this may provide an overestimation of the performance of the selected set of genes, since the feature selection procedure was already applied to the whole training dataset. Instead, different FS ranking could be calculated for the different CV partitions, but this strategy can suffer from a low stability of the FS algorithms whose rankings can be very different among the varied, though highly overlapping, CV partitions.

After the CV step, the final subset of selected DEGs can be now used to optimize the final model along with the whole training dataset. Extensively, the possibility to test the model with unseen samples (which were set aside from the beginning with this express purpose) has also been implemented. Taking this into account, there are two stages for each classifier: one stage for hyperparameter optimization and training of the models (*knn_trn*, *svm_trn* & *rf_trn*) and another stage for testing (*knn_test*, *svm_test* & *rf_test*). For the three algorithms, the following hyperparameters are optimized, searching the determination of the best model for each analysis: k for k-NN model, n-trees for RF, and for the SVM, Gaussian kernel has been implemented, thus requiring the optimization of hyperparameters (c & g). In all cases, grid-search and CV is used with the aim of finding the best combinations for the respective parameter or parameters. CV ends up also providing the expected performance on unseen samples, which can be explicitly assessed using the respective testing functions. *KnowSeq* has been designed to deal with multiclass classification problems regardless of the number of classes to address. Furthermore, the use of the coverage parameter helps to detect truly multiclass DEGs, improving the quality and robustness of the subsequent multiclass classification process. Moreover, *KnowSeq* allows the graphical representation of the results to be plotted, including the confusion matrix, the sensitivity, the specificity and the f1-score. This gives to the user the possibility of performing a very complete analysis and assessment of the tackled problem in a very simple and quick way.

2.3. DEG functional enrichment

KnowSeq assists with attaining biological knowledge related to the final DEGs candidates, which have been previously assessed in the ML process. This knowledge can be interpreted afterwards by a clinician or any person with a biological profile. *KnowSeq* allows for retrieving biological information from three different well-known sources. The first one is the *Gene Ontology (GO) enrichment* which retrieves information about the biological functions from different ontologies or domains for each DEG [35]. The three available GO domains are queried by our tool: the Biological Process (BP), the Molecular Function (MF) and the Cellular Component (CC). Thanks to this, the biological functions related with the DEGs can be acquired and facilitating a deeper study to find connections with the addressed disease. For the GO enrichment, web queries to the DAVID Platform are automatically performed, and then the retrieved information is formatted to be readable in a table format [36]. In order to carry out the GOs retrieval, *KnowSeq* has the function *geneOntologyEnrichment*.

The second source of biological information is pathway enrichment. Nowadays, it is well known that the interaction among genes can eventually activate or inhibit different biological processes. Genes interacting among themselves in the same biological function are put together in the same pathway. For that reason, it is important to know not only the expression of the DEGs but also their interactions with genes that belong to their same pathways. *KnowSeq* allows for retrieving DEG related pathways from the well-known pathway databases KEGG [37]. For this purpose, the function *DEGsToPathways* is responsible for performing this process.

The last source of biological information implemented in *KnowSeq* is the related diseases retrieval. It is performed by executing the function *DEGsToDiseases*. In this step, all the related diseases of the DEGs candidates are obtained together with the evidence that support these relations. This information helps with finding possible associations with the pathology addressed and with other possible precursor pathologies. This information is attained from the *targetValidation* [38] web platform. This platform has several scores to determine if a gene is related with the different possible diseases, based on the data collected by the web platform. The value of these scores increases when the biological relationship increases too, meaning a strong association with the selected disease. More concretely, each individual score is calculated taking into account the evidence frequency, the strength of the effect described by them, and their confidence. The Final Association Score is computed as the harmonic sum of all the available individual scores [38]. *KnowSeq* can retrieve not only each individual score but also the final score together with the evidence for a concrete disease. Then, the acquired diseases are correctly organized by *KnowSeq* to provide this information in a more readable way for the user. With the information automatically collected by *KnowSeq* from the three different sources, a strong functional enrichment process is carried out in order to build a biological profile for each of the DEGs without requiring external tools.

2.4. Automatic report

Taking into consideration those scientists who do not have a deep programming background, *KnowSeq* implements an Automatic HTML Report by simply calling the function *knowseqReport* with the expression matrix to analyze as input. This function will automatically generate an HTML report that the user can inspect in any web-browser. This report includes results of a batch execution of the pipeline, including Quality Analysis, Batch Effect Removal, Differential Expressed Genes, Feature Selection and Predictive Models and the three different functional enrichment sources (GOs, KEGG Pathways and Related diseases and evidences) along with several plots. Moreover, the report is also modular, having the possibility of deactivating any of the aforementioned steps and adapting the study to the user requirements. An example of an automatic report generated by *KnowSeq* is provided in the Supplementary Information. This example report has been created using the same Breast cancer dataset used to show the detailed application of *KnowSeq* in the next Results and Discussion section. This report includes similar outcomes to those subsequently explained, according to all the aforementioned features, but for decisions that cannot be taken in a batch execution of the pipeline (for instance the number of final DEGs selected from a feature selection algorithm).

3. Results and discussion

With the purpose of showing the operation of *KnowSeq* under a real application, two different study cases have been addressed. The section is divided in three subsections: the first one, specifies the information about the data acquisition and the description of both study cases while the following two sections introduce the obtained results and discussion for each study case separately.

3.1. Data preparation & description

All the data samples used in this research come from The Cancer Genome Atlas (TCGA) and have been acquired through the GDC Portal platform. GDC requires permission access to download BAM files from the controlled data. However, the study can also be replicated by starting from open-access count files instead of BAM files. The links to the source files for both breast and lung datasets are within the Supplementary Information in order to make the experiments totally reproducible.

Firstly, for the paired breast cancer study, 90 patients were selected with the condition of having BAM files from both solid normal and primary tumor tissues for each patient. With this condition, both the paired datasets and the best quality conditions in terms of samples for the study are ensured. The dataset was divided into a training dataset formed by 80 patients and was used to extract the DEGs. The test dataset with the 10 remaining patients was used for testing those DEGs in the ML step.

Following, for the multiclass lung cancer study, 1100 count files from three different states were collected from GDC Portal. Concretely, 495 ACC Primary Tumor, 502 SCC Primary Tumor and 103 Solid Tissue Normal samples as Control were considered. A stratified 80%–20% training-test subdivision was performed for this problem. The main motivation of this case study is the search for relevant biomarkers with the capability of discerning among the all addressed states and not exclusively for the typical case: cancer vs control. This type of studies allow for finding DEGs with significant differences even among different pathological sub-types within the same cancer type.

3.2. Paired breast cancer study

3.2.1. Gene expression analysis

The quality analysis was first performed using the 80 patients and no outlier was detected among them. Thereafter, the batch effect removal step was applied taking into account that the possible batches were unknown. The SVA algorithm [39] was performed to find the surrogate variables in order to create a model considering those variable to remove the deviations. After the quality analysis and the batch effect correction steps, DEGs candidates can now be extracted. To carry out this extraction, the thresholds imposed were very restrictive, using three statistical values for filtering: the Log Fold Change (LFC) greater or equal than 3, the *P*-value less or equal than 0.001 and COV equal to 1 due to it is a biclass problem. Applying these restrictions, a total amount of 50 DEGs candidates ordered by LFC were extracted which can be seen at Supplementary Table 2. Furthermore, Fig. 2 represents an expression heatmap that graphically shows important differences of the DEGs candidates between both tumor and normal samples.

3.2.2. Machine Learning assessment

Firstly, a 10-CV step was applied in order to assess the behavior of the classifier with the 80 patients training dataset when those DEGs are used for classification. Thereupon, all the different combination of classifiers with feature selection algorithms reached better results than applying just only limma extraction, recognizing all the training samples with a lower number of genes. SVM and RF acquired outstanding results, while k-NN had slightly better results than the other two algorithms. However, it is important to know how the classifier behaves with samples never seen before in order to simulate a real clinical case. This is the reason to create a test process with the 10 patient (20 samples) datasets. Different matches, or combinations between classifiers and feature selection algorithms, were executed with the purpose of searching the combination with the best results. Table 2 contains the results for all these combinations depending on the number of genes used to classify. It is important to highlight that with only 3 genes, k-NN reached 100% of accuracy when mRMR and RF f. s. were applied. Although all of them achieved prominent results, k-NN obtained the best results regardless of the feature selection algorithm and the number of genes used. As it can be seen, with only 3 genes selected by the feature selection process from our DEGs, all the test patients were perfectly recognized for the ML designed models. This means that *KnowSeq* brings the support necessary to create intelligent systems with the capability of extracting relevant biomarkers that are useful to discern among the addressed diseases or states.

Once the classification is done, it is very helpful to graphically visualize the gene expression differences that exist between the tumor samples and the normal samples for the three 3 DEGs that discriminate perfectly the test patients. In order to carry out this representation, *KnowSeq* contains the *dataPlot* function in mode genesBoxplot. Fig. 3 represents the genes Boxplots for the top 3 DEGs applying mRMR. In this figure, the three selected genes (COL10A1, VEGFD AND PITX1) have several differences in average expression between the addressed states.

3.2.3. DEGs enrichment

At this point of the study, our DEGs have been assessed by applying a ML process. In order to help with the biological interpretation and functional enrichment, *KnowSeq* has a last step in its pipeline created solely and exclusively for this purpose (DEGs Functional Enrichment).

Of the first three DEGs from limma, mRMR and RF f. s., only two DEGs from RF f. s. (COL10A1 & MMP11) and two DEGs from mRMR (COL10A1 & VEGFD) have a strong reported relation with breast cancer and one of them is common to limma, mRMR and RF f. s. (COL10A1). It is very interesting to note that only the first gene of the top 3 DEGs with limma has a significant relationship with breast cancer, although they all are the DEGs with the higher LFC or *P*-value. Therefore, the use of a feature selection step in this case has implied the determination of DEGs in the first positions more related with breast cancer. This fact clearly improves the classification accuracy as shown in the previous subsection. Hence, the 3 breast cancer reported DEGs will be used for the enrichment, with their scores presented in Table 3. These 4 scores provide values between 0 and 1 depending on if the association in each field is low or high, respectively. Furthermore, the *Final Association score* defined at subsection DEGs Functional Enrichment is also shown. As can be seen in the table, the three genes have a strong final association, so they are highly involved in breast cancer. From this point, the experts in the field have an important overview of the genes to continue investigating them.

The next step is the Gene Ontology enrichment. For this process the same 3 DEGs are used and the five most important GOs for them and for the three different ontologies (BP, MP & CC) will be retrieved with the function *geneOntologyEnrichment*. Supplementary Table 4 shows the top

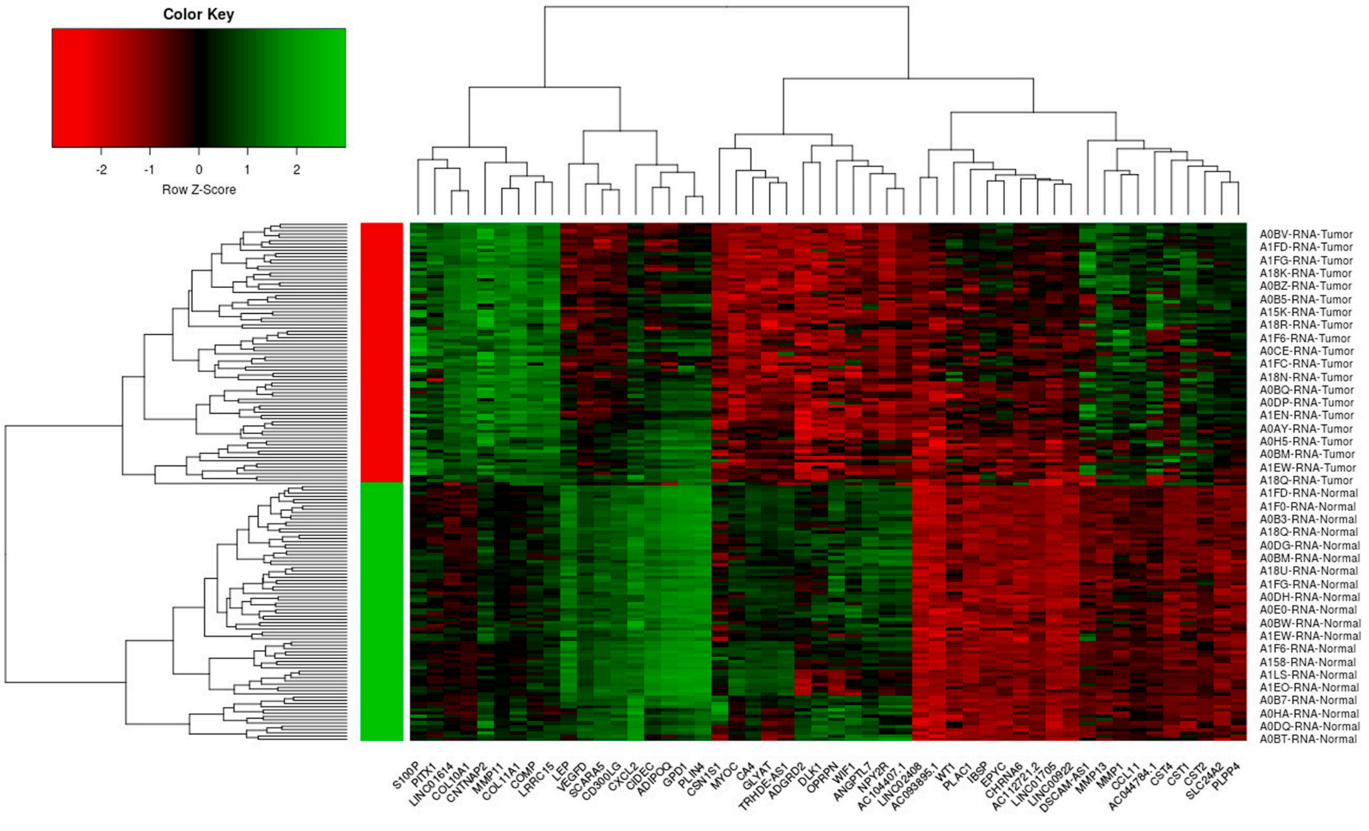


Fig. 2. Heatmap of the 50 DEGs candidates clearly showing differences between tumor and normal samples.

Table 2
Breast cancer test results for the different combinations of feature selection algorithms with the classifiers depending on the number of DEGs selected.

Method/N. Genes	Limma			mRMR			RF f.s.		
	3	5	10	3	5	10	3	5	10
SVM	85%	90%	95%	95%	95%	100%	100%	95%	100%
k-NN	90%	85%	100%	**100%**	100%	100%	100%	100%	100%
RF	85%	90%	95%	90%	70%	95%	85%	85%	100%

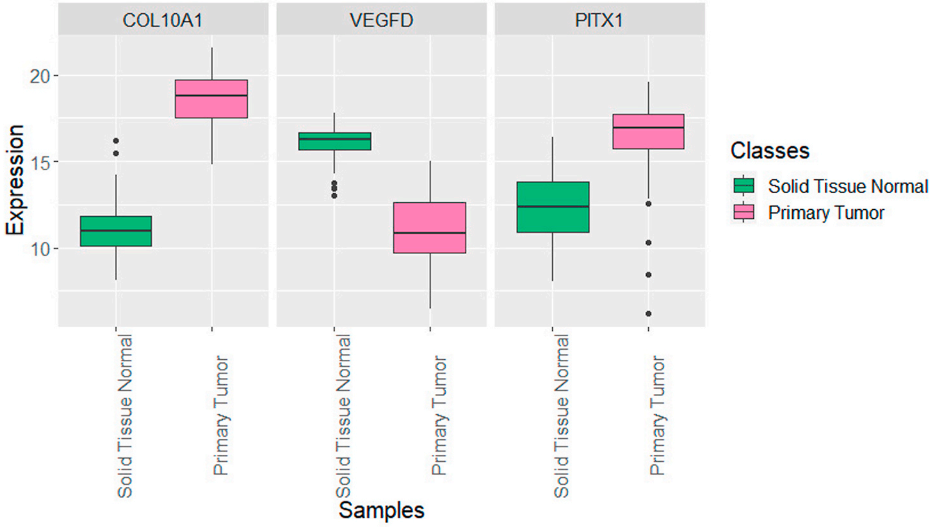


Fig. 3. Boxplots of the 3 first mRMR selected DEGs by KnowSeq for the Breast cancer study case.

Table 3
Information about the Breast cancer association scores (sc.) for the final 3 DEGs to study.

Gene	Liter. sc.	RNA Exp. sc.	Affected Paths. sc.	Final Assoc. sc.
COL10A1	0.0372	0.1787	0.6835	0.7323
MMP11	0.1935	0.1094	0.6065	0.6670
VEGFD	0.1169	0.1400	0.6948	0.7428

5 GOs for our 3 DEGs. As it can be seen, the VEGFD gene does not appear for any GO terms in the top 5, but only GOs related with COL10A1 and MMP11 genes. Only by increasing the maximum number of retrieved GOs, were GOs related to the VEGFD retrieved. Thanks to this step, the Biological Processes (BP), the Molecular Functions (MF) and the Cellular Components (CC) of the DEGs are stored by *KnowSeq* to help users know the biological domain of each DEGs and study possible relations with processes that could lead to the development of cancer.

Finally, the last biological enrichment step included in *KnowSeq* is the pathway enrichment. Pathways involving the selected DEGs are important to understand how the expression changes are affecting other genes and biological processes, as well as how theses changes can turn into cancer (breast cancer, in this case). To achieve that, *KnowSeq* includes the function *DEGsToPathways*. This function makes use of KEGG database to acquire the pathway information. For the COL10A1, there is one reported pathway affected. For the MMP11 there exists no affected pathways in KEGG. Finally, for the VEGFD gene there are nine reported pathways. Table 4 shows the nine VEGFD related pathways as well as the pathway related with COL10A1 gene.

As can be seen in Table 4, the gene VEGFD is involved in several pathways (Pathway in cancer included). The changes in its expression could produce disorders in those pathways, which could end up in the development of breast cancer and other diseases. However, since KEGG database does not reveal any pathway for MMP11 or any interaction between them, an additional search has been done into Reactome Database as KnowSeq will support that tool in the next release. In this sense, information from the Reactome database shows that both MMP11 and COL10A1 have influence in the degradation of the extracellular matrix which, in turn, contributes to the tumor growth and progression [40]. Concretely, both DEGs affect the collagen degradation while disrupting the extracellular matrix. As has been reported for other cancer diagnoses, DEGs involved in collagen degradation can help discern ovarian and breast cancer from healthy controls [41].

Table 4
Retrieved pathways and their description for the chosen Breast cancer DEGs.

KEGG Pathway	Name	Genes
hsa04974	Protein digestion and absorption	COL10A1
hsa04010	MAPK Signaling Pathway	VEGFD
hsa04014	RAS Signaling Pathway	
hsa04015	RAP1 Signaling Pathway	
hsa04151	PI3K-AKT Signaling Pathway	
hsa04510	Focal Adhesion	
hsa04668	TNF Signaling Pathway	
hsa04926	Relaxing Signaling Pathway	
hsa04933	Age-Rage Signaling Pathway in diabetic complications	
hsa05200	Pathway in cancer	

3.3. Multiclass lung cancer study

3.3.1. Gene expression analysis

The quality analysis was first performed to the whole set of patients, detecting 10 outliers among them. Then, as before, the batch effect removal step was carried out. The SVA algorithm [39] was used to find the surrogate variables in order to create a model considering those variable to remove the deviations. For the DEGs candidates extraction, the thresholds imposed were very restrictive, using three statistical values for filtering: the LFC greater than or equal to 1.5 (in order to not lose information), the *P*-value less than or equal to than 0.001 and COV equal to 2 given that it is a multiclass problem and there are 3 classes. Supplementary Table 3 shows the 50 more relevant multiclass genes of a total of 410 DEGs, taking into account both LFC and COV.

3.3.2. Machine Learning assessment

A 10-CV step was performed over the training dataset in order to observe the performance of the classifier. Thereafter, the classifier was applied to the test set. Table 5 contains multiclass test classification results for the different combinations of feature selection and classification methods for different number of DEGs. As it can be observed, using any type of FS improves the results over using just limma, as it happened for the previous problem. RF classifier obtains the best results when using mRMR feature selector, using only a subset of 6 genes. The graphic representation of the accuracy obtained on the test set, using RF and for each feature selector, is presented in Fig. 4. Even though mRMR and RF f. s. reach the same performance when using only one gene, mRMR outperforms RF f. s. in any other case, and both of them improve results over using limma. When dealing with a multiclass problem, accuracy might not be the most reliable metric, since it is affected by class imbalance, and hence the confusion matrix is a better method to measure the classifier precision. Therefore, Fig. 5 presents the confusion matrix when using mRMR and RF classifier. As it can be observed, the f1-Score value (that uses the harmonic mean and therefore is not affected by imbalanced classes) is very similar to the accuracy value, having a similar behavior in terms of both sensitivity and specificity. Therefore, the classifier is able to properly classify each class without being prone to any specific one. It is important to note, that although it is very hard to obtain the same recognition for the classifiers, the classification trends between them seem to be the same in relation to the two problems tackled in this work.

Fig. 6 shows the boxplots for the 6 selected genes by mRMR. As it can be observed in the figure, the genes are able to properly discriminate between the different classes. The first selected gene (KRT5) shows an outstanding discrimination between ACC and SCC while, for instance, the third selected gene (SH3GL3) presented this discrimination power between ACC and control. As shown for the results presented in Table 5, using these genes with the RF classifier led to obtaining impressive results for the multiclass classification in the test set. It is important to highlight that all the graphs and plots have been represented by using the *dataPlot* function available at *KnowSeq*.

3.3.3. DEGs enrichment

In this case, the enrichment is done over the 6 DEGs selected by mRMR which is composed of the GOs enrichment, KEGG pathways and related diseases retrieval.

To start with the functional enrichment step, Supplementary Table 5 shows the list of the top 5 five GO terms for the three main ontologies (BP, MF & CC). All the 6 final DEGs are involved in at least one GO term in the table. However, more GO terms can be retrieved just by modifying the parameters of the function *geneOntologyEnrichment*. An in-depth study of this information could reveal relations between those DEGs and biological considerations related to the addressed states in the multiclass study.

Table 6 contains the list of the KEGG pathways in which one or more of the selected DEGs are involved by using the function *DEGsToPathways*.

Finally, it is very useful to know which is the correlation between the final DEGs and the different addressed diseases. To that end, the different Final Association Score for each DEGs for different lung related diseases have been retrieved with the function *DEGsToDiseases*. As it is detailed in Table 7, four of the six DEGs are related with some lung pathology or even directly with cancer (KRT5, SH3GL3, TFAP2A & S1PR5) with a very strong association in some cases. Nevertheless, the other remaining DEGs lack of association with practically all the addressed cases. FAM189A2 gene only has a very low association with Lung ACC while TICRR has a low association with cancer. These results open the door for scientists and experts to perform an in-depth study about those DEGs, which have several expression changes among the states yet have been related with a very low number of diseases.

4. Features, limitations & future implementations

This section summarizes relevant information about our R package. *KnowSeq*, as it is currently implemented, offers modularity and scalability, i.e. different modules can be replaced by the obtained results from other tools, customizing or even extending our proposed automatic pipeline. For example, the number of retrieved DEGs can vary in function of the main objective of a specific study. *KnowSeq* is focused on small gene signatures as it is of utmost importance to develop diagnostic kits or interpretable models to help clinical decision-making [42]. However, a larger number of DEGs can be extracted to their posterior analysis through external tools taking into account, for example, networks or sub-networks [43].

KnowSeq allows for developing complex studies, without the need for advanced programming skills. The required lines of code decrease dramatically due to the high encapsulation and degree of abstraction (around 40 to 50 lines of code for both cases). Moreover, to develop complete study cases, starting from the count files, 143 and 447 min were required respectively in the 4-node cluster described in Section 2.1. These time metrics consolidate the fact that thanks to *KnowSeq*, complete studies can be done using an insignificant number of lines of code along with a low execution time. Qualitatively, *KnowSeq* implements and automates all the necessary steps to perform complete expression analysis. Moreover, *KnowSeq* includes different parameters to allow for a high customization, to adapt the pipeline to the users' requirements. Finally, a Docker container was designed which includes a RStudio server with the latest version of *KnowSeq* installed along with an R script of a real case to use as an example. The container allows *KnowSeq* to be run no matter the Operative System or the required dependencies. This boosts the usability while decreasing the time needed to perform a gene expression analysis. The KnowSeq Docker can be also deployed to corroborate both qualitative and quantitative assessment through the real analysis of breast cancer expression which is included within the container.

KnowSeq has a set of limitations which will be resolved in further releases. At the moment of the redaction of this manuscript, *KnowSeq* does not allow GSEA within their current three Functional Enrichment methods. Furthermore, currently there is no support for Reactome pathway information retrieval. In addition, due to the large number of Microarray platforms and the particular existing pre-processing libraries

Table 5
Mean F1-score for Multiclass Lung cancer test results for the different combinations of feature selection algorithms with the classifiers depending on the number of DEGs selected.

Method/N. Genes	Limma			mRMR			RF f.s.		
	3	6	9	3	6	9	3	6	9
SVM	78.4%	88.8%	89.2%	90.4%	94.7%	95.1%	78.0%	88.0%	89.1%
k-NN	77.7%	86.1%	87.2%	90.4%	95.1%	95.1%	79.6%	88.3%	88.2%
RF	76.2%	83.1%	85.3%	91.4%	**95.4%**	94.4%	81.1%	84.7%	90.1%

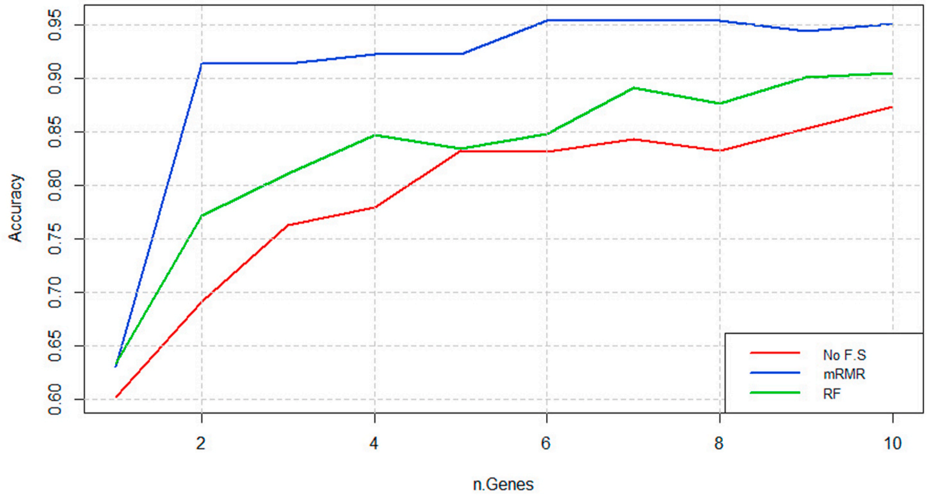
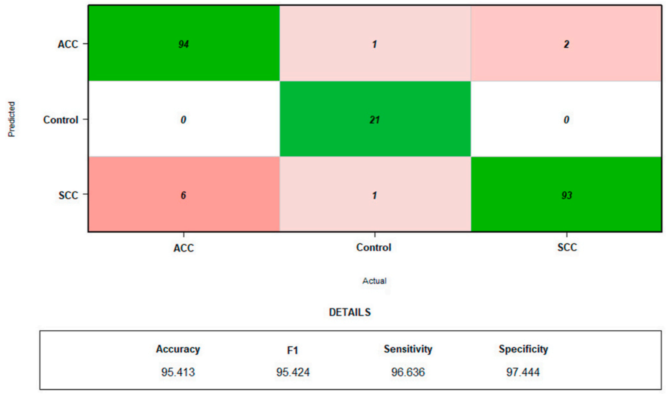


Fig. 4. Mean F1-score for the Lung cancer Test classification results by implementing RF with the three FS methods.



Predicted \ Actual	ACC	Control	SCC
ACC	94	1	2
Control	0	21	0
SCC	6	1	93

Accuracy	F1	Sensitivity	Specificity
95.413	95.424	96.636	97.444

Fig. 5. Test Confusion Matrix for the Lung cancer study case for RF classifier using mRMR 6 first selected DEGs.

each of them require, we decided to focus on RNA-Seq RAW pre-processing. Nevertheless, as aforementioned, once the Microarray gene expression matrix is obtained, *KnowSeq* can work with it. Currently, *KnowSeq* does not contain automatic training-test stratification, nor imbalance dataset treatments for classification tasks, which are expected to be implemented as soon as possible. Further releases of *KnowSeq* will also implement interpretable classification to provide even more information to help in decision making [44–48]. Finally, the advances of *KnowSeq* can be implemented in a user-friendly pipeline, similar to the presented platform by R. Kohen et al. [49]. In fact, a web version of *KnowSeq* is now under development with the aim of reducing to zero the required programming skills to launch the tool.

Table 6
Retrieved pathways and their description for the chosen Lung cancer DEGs.

KEGG Pathway	Name	Genes
hsa04071	Sphingolipid signaling pathway	S1PR5
hsa04080	Neuroactive ligand-receptor interaction	S1PR5
hsa04144	Endocytosis	SH3GL3

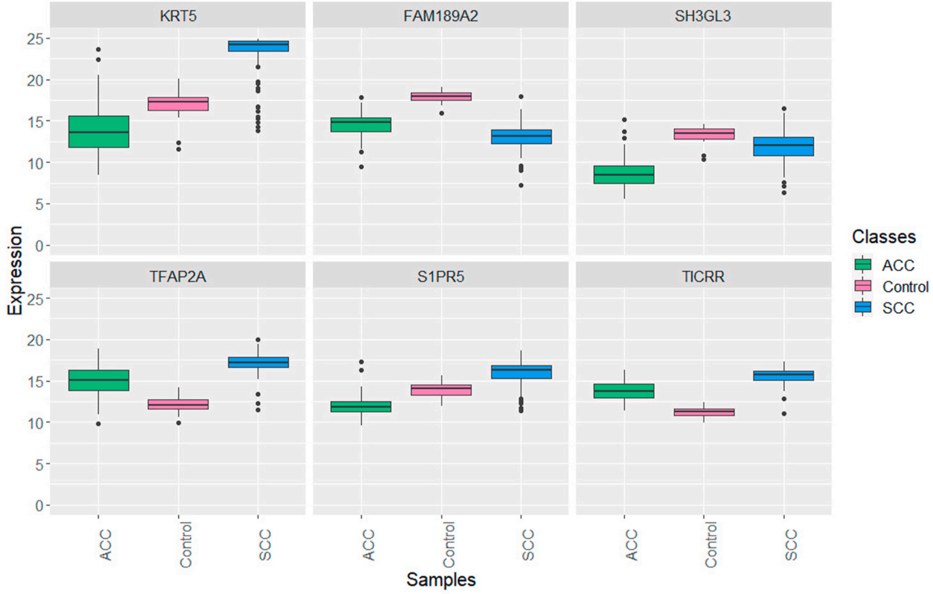


Fig. 6. Test Genes Boxplots for the Lung cancer study case for the first mRMR 6 DEGs.

Table 7
Final Association Scores for the chosen 6 Lung cancer DEGs selected by mRMR.

Genes	Lung Disease	Lung cancer	Lung ACC	Lung SCC	cancer
KRT5	0.10	0.09	0.09	0.53	1.0
FAM189A2	–	–	0.08	–	–
SH3GL3	0.68	0.68	0.68	–	0.81
TFAP2A	–	0.68	0.67	–	1.0
S1PR5	–	–	0.27	–	0.76
TICRR	–	–	–	–	0.06

5. Conclusions

In this paper, a novel tool to carry out transcriptomic gene expression analysis, publicly available at Bioconductor, has been presented. *KnowSeq* includes the most well-known steps to perform this type of study, and extends them with powerful functionalities: feature selection, classification and enrichment analysis. Thanks to this, complete analyses can be carried out from raw data treatment up to biological knowledge extraction, in an easy, modular and flexible way. In addition, our package returns an Automatic HTML report that computes all the above steps together, providing the expert a file with the results of the complete study. For this reason, *KnowSeq* expects to server as a decision-making support system that can be used by experts when diagnosing certain pathologies.

The operation of *KnowSeq* has been exemplified by addressing two case studies with different complexity levels: at the biclass level (breast cancer) and at the multiclass level (lung cancer). For both experimental analyses, the classification results confirm the validity of *KnowSeq* determining a reduced and highly informative subset of DEGs for an intelligent diagnosis (100% for breast cancer with only 3 genes and 95% for lung cancer considering 6 genes). In addition to the strong biological evidence, supported by affected pathways and specific literature associating those biomarkers with each cancerous disease, the power of *KnowSeq* resides in the effective integration of heterogeneous datasets and benefits from how the feature selection algorithms are able to determine those more highly informative biomarkers.

The *KnowSeq* R/Bioc package gives the possibility to carry out gene expression analyses in an easy and modular way. In fact, our tool presented here is openly thought to serve as an innovative assessment instrument to help experts in the field acquire robust knowledge and conclusions for the data and diseases to study. *KnowSeq* can be outlined and broadly characterized by four clear strengths: firstly, in terms of modularity, as the analyses can be started from different points (FASTQ, BAM or count files, and even from a custom expression matrix); secondly, in terms of versatility, due to the different algorithms for ML and feature selection implemented as well as the different databases taken into account in *KnowSeq*; thirdly, in terms of adaptability of the analyses, because *KnowSeq* allows for using data from different sources and even selecting different parameters that give the user a real control of the pipeline; lastly, in terms of interpretability, as *KnowSeq* allows HTML reports to be created with the aim of providing experts with one automatically generated report with full and detailed results of their studies.

Declaration of competing interest

None Declared.

Acknowledgment

This work was funded by the Spanish Ministry of Sciences, Innovation and Universities under Project RTI2018-101674-B-I00 titled “Computer Architectures and Machine Learning-based solutions for complex challenges in Bioinformatics, Biotechnology and Biomedicine”, in collaboration with the Government of Andalusia under Postdoctoral Grant P12-TIC-2082. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of this manuscript. The results published here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2021.104387.

References

[1] B. Seelbinder, T. Wolf, S. Priebe, S. McNamara, S. Gerber, R. Guthke, J. Linde, GEO2RNAseq: an Easy-To-Use R Pipeline for Complete Pre-processing of RNA-Seq Data, 2019, p. 771063, bioRxiv.
[2] T. Barrett, D.B. Troup, S.E. Wilhite, et al., NCBI GEO: mining tens of millions of expression profiles—database and tools update, Nucleic Acids Res. 35 (suppl 1) (2007) D760–D765.
[3] M. Lohse, A.M. Bolger, A. Nagel, A.R. Fernie, J.E. Lunn, M. Stitt, B. Usadel, RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics, Nucleic Acids Res. 40 (W1) (2012) W622–W627.
[4] K.H. Chao, Y.W. Hsiao, Y.F. Lee, et al., RNASeqR: an R Package for Automated Two-Group RNA-Seq Analysis Workflow, 2019 arXiv preprint arXiv:1905.03909.
[5] G. Gómez-López, J. Dopazo, J.C. Cigudosa, A. Valencia, F. Al-Shahrour, Precision medicine needs pioneering clinical bioinformaticians, Briefings Bioinf. 20 (3) (2019) 752–766.
[6] D. Castillo, J.M. Galvez, L.J. Herrera, B. San Roman, F. Rojas, I. Rojas, Integration of RNA-seq data with heterogeneous Microarray data for breast cancer profiling, BMC Bioinf. 18 (1) (2017), https://doi.org/10.1186/s12859-017-1925-0.
[7] J.M. Galvez, D. Castillo, L.J. Herrera, B. San Roman, O. Valenzuela, F.M. Ortuno, et al., Multiclass classification for skin cancer profiling based on the integration of heterogeneous gene expression series, PloS One 13 (5) (2018) 1V, https://doi.org/10.1371/journal.pone.0196836.
[8] D. Castillo, J.M. Galvez, L.J. Herrera, F. Rojas, O. Valenzuela, O. Caba, J. Prados, I. Rojas, Leukemia multiclass assessment and classification from Microarray and RNA-seq technologies integration at gene expression level, PloS One 14 (2) (2019), e0212127, https://doi.org/10.1371/journal.pone.0212127.
[9] R.L. Grossman, A.P. Heath, V. Ferretti, H.E. Varmus, D.R. Lowy, W.A. Kibbe, L. M. Staudt, Toward a shared vision for cancer genomic data, N. Engl. J. Med. 375 (12) (2016) 1109–1112.
[10] S.U. Shin, J. Lee, J.H. Kim, et al., Gene expression profiling of calcifications in breast cancer, Sci. Rep. 7 (1) (2017) 1–11.
[11] H.G. Russnes, O.C. Lingjærde, A.L. Børresen-Dale, C. Caldas, Breast cancer molecular stratification: from intrinsic subtypes to integrative clusters, Am. J. Pathol. 187 (10) (2017) 2152–2162.
[12] D. Sun, M. Wang, H. Feng, A. Li, October). Prognosis prediction of human breast cancer by integrating deep neural network and support vector machine: supervised feature extraction and classification for breast cancer prognosis prediction, in: 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), IEEE, 2017, pp. 1–5.
[13] J. Wu, C. Hicks, Breast cancer type classification using machine learning, J. Personalized Med. 11 (2) (2021) 61.
[14] F. Hu, Y. Zhou, Q. Wang, Z. Yang, Y. Shi, Q. Chi, Gene expression classification of lung adenocarcinoma into molecular subtypes, IEEE ACM Trans. Comput. Biol. Bioinf 17 (4) (2019) 1187–1197.
[15] M.D. Podolsky, A.A. Barchuk, V.I. Kuznetcov, N.F. Gusarova, V.S. Gaidukov, S. A. Tarakanov, Evaluation of machine learning algorithm utilization for lung cancer classification based on gene expression levels, Asian Pac. J. Cancer Prev. APJCP 17 (2) (2016) 835–838.
[16] S. Tian, Classification and survival prediction for early-stage lung adenocarcinoma and squamous cell carcinoma patients, Oncology letters 14 (5) (2017) 5464–5470.
[17] S. González, D. Castillo, J.M. Galvez, I. Rojas, L.J. Herrera, Feature selection and assessment of lung cancer sub-types by applying predictive models, in: International Work-Conference on Artificial Neural Networks, Springer, Cham, 2019, pp. 883–894.
[18] A. Brazma, H. Parkinson, U. Sarkans, et al., ArrayExpress—a public repository for microarray gene expression data at the EBI, Nucleic Acids Res. 31 (1) (2003) 68–71.
[19] H. Li, B. Handsaker, A. Wysoker, T. Fennell, et al., The sequence alignment/map format and SAMtools, Bioinformatics 25 (16) (2009) 2078–2079.
[20] D. Kim, B. Langmead, S.L. Salzberg, HISAT: a fast spliced aligner with low memory requirements, Nat. Methods 12 (4) (2015) 357.
[21] M.D. Robinson, D.J. McCarthy, G.K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data, Bioinformatics 26 (1) (2010) 139–140.
[22] K.D. Hansen, R.A. Irizarry, Z. Wu, Removing technical variability in RNA-seq data using conditional quantile normalization, Biostatistics 13 (2) (2012) 204–216.
[23] W.W.B. Goh, W. Wang, L. Wong, Why batch effects matter in omics data, and how to avoid them, BMC Bioinf. 6 (1) (2017) 191.
[24] J.T. Leek, W.E. Johnson, H.S. Parker, A.E. Jaffe, J.D. Storey, The sva package for removing batch effects and other unwanted variation in high-throughput experiments, Bioinformatics 28 (6) (2012) 882–883.

[25] G.K. Smyth, Limma: linear models for microarray data, in: Bioinformatics and Computational Biology Solutions Using R and Bioconductor, Springer, New York, NY, 2005, pp. 397–420.

[26] C. Soneson, M. Delorenzi, A comparison of methods for differential expression analysis of RNA-seq data, BMC Bioinf. 14 (1) (2013) 1–18.

[27] V. Bolón-Canedo, N. Sánchez-Marono, A. Alonso-Betanzos, J.M. Benítez, F. Herrera, A review of microarray datasets and applied feature selection methods, Inf. Sci. 282 (2014) 111–135.

[28] C. Ding, H. Peng, Minimum redundancy feature selection from Microarray gene expression data, J. Bioinf. Comput. Biol. 3 (2) (2005) 185–205.

[29] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.

[30] R.M. Parry, W. Jones, T.H. Stokes, et al., k-Nearest neighbor models for Microarray gene expression analysis and clinical outcome prediction, Pharmacogenomics J. 10 (4) (2010) 292.

[31] T.K. Ho, Random decision forests, in: Proceedings of 3rd International Conference on Document Analysis and Recognition, vol. 1995, IEEE, 1995, pp. 278–282.

[32] S. Huang, N. Cai, P.P. Pacheco, S. Narrandes, Y. Wang, W. Xu, Applications of support vector machine (SVM) learning in cancer genomics, CANCER GENOMICS PROTEOMICS 15 (1) (2018) 41–51.

[33] Y. Li, K. Kang, J.M. Krahn, N. Croutwater, K. Lee, D.M. Umbach, L. Li, A comprehensive genomic pan-cancer classification using the cancer Genome Atlas gene expression data, BMC Genom. 18 (1) (2017) 1–13.

[34] J.C. Almlöf, A. Alexsson, J. Imgenberg-Kreuz, et al., Novel risk genes for systemic lupus erythematosus predicted by random forest classification, Sci. Rep. 7 (1) (2017) 1–11.

[35] Gene Ontology Consortium, The gene ontology resource: 20 years and still GOing strong, Nucleic Acids Res. 47 (D1) (2018) D330–D338.

[36] D.W. Huang, B.T. Sherman, R.A. Lempicki, Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists, Nucleic Acids Res. 37 (1) (2009) 1–13.

[37] M. Kanehisa, S. Goto, KEGG: kyoto encyclopedia of genes and genomes, Nucleic Acids Res. 28 (1) (2000) 27–30.

[38] G. Koscielny, P. An, D. Carvalho-Silva, J.A. Cham, et al., Open Targets: a platform for therapeutic target identification and validation, Nucleic Acids Res. 45 (D1) (2016) D985–D994.

[39] J.T. Leek, J.D. Storey, Capturing heterogeneity in gene expression studies by surrogate variable analysis, PLoS Genet. 3 (9) (2007) e161.

[40] N. Masoud, F. Bagher, M. Keywan, Extracellular matrix (ECM) stiffness and degradation as cancer drivers, J. Cell. Biochem. 120 (3) (2019) 2782–2790.

[41] C.L. Bager, N. Willumsen, D.J. Leeming, V. Smith, M.A. Karsdal, D. Dornan, A. C. Bay-Jensen, Collagen degradation products measured in serum can separate ovarian and breast cancer patients from healthy controls: a preliminary study, Canc. Biomarkers 15 (6) (2015), 789-788.

[42] S. Narrandes, W. Xu, Gene expression detection assay for cancer clinical use, J. Canc. 9 (13) (2018) 2249.

[43] P. Shannon, A. Markiel, O. Ozier, et al., Cytoscape: a software environment for integrated models of biomolecular interaction networks, Genome Res. 13 (11) (2003) 2498–2504.

[44] P. Sabol, P. Sinčák, K. Ogawa, P. Hartono, Explainable classifier supporting decision-making for breast cancer diagnosis from histopathological images, in: *2019* International Joint Conference on Neural Networks (IJCNN), IEEE, 2019, pp. 1–8.

[45] P. Sabol, P. Sinčák, P. Hartono, et al., Explainable classifier for improving the accountability in decision-making for colorectal cancer diagnosis from histopathological images, J. Biomed. Inf. 109 (2020) 103523.

[46] G.R. Vasquez-Morales, S.M. Martinez-Monterrubio, P. Moreno-Ger, J.A. Recio-Garcia, Explainable prediction of chronic renal disease in the Colombian population using neural networks and case-based reasoning, IEEE Access 7 (2019) 152900–152910.

[47] S.M. Lundberg, B. Nair, M.S. Vavilala, et al., Explainable machine-learning predictions for the prevention of hypoxaemia during surgery, Nature biomedical engineering 2 (10) (2018) 749–760.

[48] M.O. Adebiyi, A.A. Adebiyi, O. Okesola, M.O. Arowolo, ICA learning approach for predicting RNA-seq data using KNN and decision tree classifiers, International Journal of Advanced Science and Technology 3 (29) (2020) 12273–12282.

[49] R. Kohen, J. Barlev, G. Hornung, et al., UTAP: user-friendly transcriptome analysis pipeline, BMC Bioinf. 20 (1) (2019) 1–7.