KnowSeq R/bioc package: Beyond the traditional RNA-seq pipeline. A breast cancer case study.

Daniel Castillo-Secilla (cased@ugr.es)
Universidad de Granada https://orcid.org/0000-0002-7380-1023
Juan Manuel Galvez
Universidad de Granada
Francisco Manuel Ortuno
Hospital Universitario Virgen del Rocio
Luis Javier Herrera
Universidad de Granada
Ignacio Rojas
Universidad de Granada

Software

Keywords: RNA-Seq, bioconductor, breast cancer, cancer, computational biology, Bioinformatics, gene expression, classification, enrichment

Posted Date: November 8th, 2019

DOI: https://doi.org/10.21203/rs.2.16962/v1

RESEARCH

KnowSeq R/bioc package: Beyond the traditional RNA-seq pipeline. A breast cancer study case.

Daniel Castillo-Secilla[1*†], Juan Manuel Gálvez[1], Francisco Manuel Ortuno[2], Luis Javier Herrera[1] and Ignacio Rojas[1]

[*]Correspondence: cased@ugr.es
[1]Department of Computer Architecture and Technology, University of Granada, Periodista Rafael Gómez Montero, 2, 18014 Granada, Spain
Full list of author information is available at the end of the article
[†]Equal contributor

Abstract

Background: The number of gene expression analyses has grown exponentially over the last years. The main triggers of this increase are the reduction in the sequencing cost per sample and the technological advances, specially in the computing scope. Those analyses generally involve a number of steps. Firstly, a raw samples alignment and a quality analysis are needed. After that, a Differentially Expressed Genes (DEGs) extraction and a subsequent gene enrichment can be performed. The development of intelligent predictive tools results essential in bioinformatics given that there exists a real need of assistance for decision-making systems towards precision medicine. Therefore, *KnowSeq* incorporates novel steps of feature selection and classifier design in the traditional RNA-seq pipeline. No tool exists in the research community that achieves this complete RNA-seq analysis, encapsulating all those steps in one single tool.

Results: In order to show the functionalities provided by the general pipeline designed for the KnowSeq package, an application to a real problem is presented. Concretely, an analysis of a breast cancer set of patients collected from the controlled repository *GDC portal* is performed, keeping paired samples between tumour and control. As results show, *KnowSeq* achieves extracting more relevant biological knowledge related with breast cancer from the RNA raw data acquisition. *KnowSeq* is available through Bioconductor.

Conclusions: *KnowSeq* R/bioc package is born with the purpose of providing an integrative tool, containing the necessary steps to address complex RNA-seq analyses in a modular and flexible way. In this paper a breast study case is addressed with *KnowSeq*, obtaining outstanding results and demonstrating the validity of *KnowSeq* to carry out gene expression analyses.

Keywords: RNA-Seq; bioconductor; breast cancer; cancer; computational biology; Bioinformatics; gene expression; classification; enrichment

Background

During the last decade, the importance of the DNA sequencing studies severely raised due to the emergence of Next Generation Sequencing (NGS) and the consequent decrease in prices of this technology in comparison with its predecessors. As a result, the number of public or controlled available data has grown exponentially, together with the computational cost needed to process this increasing amount of information. 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.

DNA sequencing studies are fundamental to win the battle against genetic diseases like cancer. Cancer is still the second cause of death worldwide, just behind cardiovascular disease. Although the survival rate is increasing gradually thanks to the medical researches and advances, breast cancer is still one of the five most dangerous cancers in the world according to World Health Organisation (WHO). Breast cancer is also the cancer with the highest impact among the female population, causing the 15% of the total women death by cancer [1]. Nowadays, most of the breast cancer diagnoses are done when the patients present symptoms, which increases the mortality risk. The worst case occurs when the cancer has spread and the patients suffer from metastasis, as the treatment becomes more difficult, and the chances of surviving are significantly lower. Therefore, it is primordial in this type of studies to search for biomarkers that allow achieving an early diagnosis of breast cancer.

Because of that, the design of novel bioinformatic tools that allows processing and extracting multi-omics information from raw data is a crucial objective in this research area. Currently, there exist different tools that combine the different steps and technologies involved in this scope ([2–4]). Nevertheless, to the best of our knowledge, there are no tools that integrate the traditional DEGs extraction steps with further, and nowadays essential steps dealing with the intelligent predictive model design and biological enrichment processes. Those steps are focused on the assistance to decision-making system applied to precision medicine ([5]). In this sense, our group presents a very powerful pipeline sustained by a complete and public Bioconductor R package to perform a complete RNA-seq study, starting from the automatic alignment of raw data and ending up in the DEGs knowledge enrichment [6]. *KnowSeq* is thought to deal with the Homo Sapiens genetic diseases but it is prepared to support any other species.

Specifically, the manuscript addresses from the application of *KnowSeq* to the search of relevant biomarkers for breast cancer detection as study case, together with their related biological information. This means that the pipeline applied for breast cancer data here could be applied for any other study with genetic source no matter the pathology to address.

KnowSeq is focused on RNA-seq as it is the most powerful and widespread genetic characterisation technology for transcriptome nowadays. *KnowSeq* comprises a large part of the tools used in our previous studies/publications involving 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 [7–10].

They widely confirm the validity of *KnowSeq* to carry out genetic diseases analysis working at gene expression level with raw data from RNA-seq.

In this scope, *KnowSeq* can be very helpful to perform these types of analysis to find and assess biomarkers. For that, this paper addresses a real application of *KnowSeq* to a set of raw breast cancer controlled data coming from *GDC Portal* [11]. Although *KnowSeq* allows the SRA/FASTQ alignment, GDC Portal does not supply those files, thus, for the analysis, 180 BAM files belonging to 90 breast cancer patients were used. For each patient, two samples were collected, a primary tumour sample and a solid tissue normal sample. Thanks to this, the experiment was designed with Tumour-Normal paired samples, which ensures the best experiment quality in terms of samples.

A whole evolved pipeline was designed combining the traditional RNA-seq pipeline with new functionalities and improvements, some of them already tested in previous works, all now under the *KnowSeq* package. Although the methodology will be deeply explained in the next sections, a brief summary of *KnowSeq* giving basic information about its operation and possibilities is given herein: The download and alignment of the samples is performed automatically. Then the gene expression values are estimated and the quality analysis and batch effect removal is carried out. When the quality is checked, the DEGs between two or more conditions established by the user are extracted (e.g. treated vs non treated, normal vs control, etc). At this point, the traditional primary pipeline is over. Nevertheless, *KnowSeq* adds a set of steps to provide depth to the studies. In these new steps, a feature selection approach is included to estimate those genes that contain more information to discern between conditions (in our study case, normal vs tumoral tissue). Furthermore, it also includes a machine learning step with different algorithms and configurations to assess those DEGs. Finally, the most useful step at biological level added by *KnowSeq* is the DEGs enrichment step. In this sense, the tool allows retrieving information about the Gene Ontologies terms (GOs) of the DEGs, the involved pathways coloured according to the gene expression level of the samples and a list of diseases related with the DEGs and different combination of those DEGs. Due to all of these reasons, *KnowSeq* is the only R/bioc package that allows performing a complete RNA-seq study by using the same single tool and programming language during all the process.

Material and Methods. KnowSeq Pipeline.

This section describes all the steps implemented by the *KnowSeq* pipeline, being also applied to this research. Figure 1 represents the whole pipeline and four different steps can be clearly distinguished: Webdata Resources gathering, RNA-seq RAW data processing, Biomarkers identification and assessment and DEGs enrichment methodology. On this basis, each step is presented in one subsection with the purpose of giving a deeper explanation for each of them. It is to be highlighted that *KnowSeq* is designed to achieve a high modularity. This means that each of the steps and sub-steps conforming *KnowSeq* can be perfectly replaced, provided that the inputs maintain the same data type. Because of this, *KnowSeq* can be easily adapted even for different species and biological data types not explicitly addressed in this first version of our tool. Furthermore, the pipeline can be launched from

different steps depending on the type of input files (e.g. SRA/FASTQ, BAM/SAM or counts). In order to summarise the different functions available in the package, Table 1 shows for each function the name, the pipeline step where this function is used, the description of the functionality and the different options implemented inside the function. Furthermore, the functions inside the table are ordered by the steps in Figure 1.

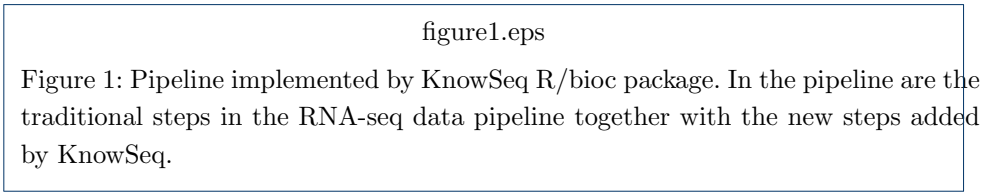

Figure 1: Pipeline implemented by KnowSeq R/bioc package. In the pipeline are the traditional steps in the RNA-seq data pipeline together with the new steps added by KnowSeq.

Webdata Resources

One of the hardest step in any biological study is the data gathering. *KnowSeq* allows to automatize the download of public and controlled samples from the most renowned web platform databases: *NCBI/GEO*, *ArrayExpress* and *GDC Portal* [12, 13]. The data from *NCBI/GEO* and *ArrayExpress* are publicly available and from these web platforms *KnowSeq* only requires the series ID to download. However, the raw data in *GDC Portal* are under restricted access and an authorisation is required via token file. If the user has this token file, GDC Portal raw data could be automatically downloaded by calling the function *gdcClientDownload*.

Users need then to construct or download an CSV/TSV files with the information of each data/series. Using the specific function *downloadPublicSeries*, an automatic download of these supporting files are made with series that belong to *NCBI/GEO* and *ArrayExpress*. Thanks to this, it is very simple to specify and gather the series and samples required to perform an analysis. All the data used to carry out our study, were downloaded by using this method.

RNA-seq RAW data processing

Once the data raw files have been acquired, an alignment process is required by using the human reference genome in order to obtain the count files to perform the DEGs analysis. *KnowSeq* allows to download the Human Reference Genome GRCh37 and GRCh38 from Ensembl, although whichever reference genome can be used if the user indicates the path to the file. In this step, the raw files in SRA or FASTQ [14] formats are processed to obtain the BAM/SAM files. This is performed through the use of *rawAlignment KnowSeq* function. For this process, *KnowSeq* counts on the *samtools* [15] and four of the most well-known aligners with the purpose of giving to the academics not only one option to apply. The aligners are *tophat2*, *hisat2*, *salmon* and *kallisto* [16–19]. Furthermore, the *Htseq-count* tool extracts the count files for each samples [20].

Finally, through the function *countsToMatrix*, all the count files are merged in one aggregated matrix with *edgeR*. By the use of the function *calculateGeneExpressionValues*, the equivalent gene expression values are calculated with *cqn* R package [21, 22]. By applying this step to the counts files, the desired number of

Table 1: Table that contains the most important functions in KnowSeq. For each function, the name, the pipeline step where this function is, the description and the options inside the function are showed.

Function Name	**Pipeline step**	**Description (options)**
downloadPublicSeries	Webdata resources	Download series from GEO and AE
gdcClientDownload	Webdata resources	Download data from GDC-Portal
rawAligment	RNA-seq RAW data processing	Raw data aligment with different algorithms (tophat2, salmon, hisat2, kallisto)
countsToMatrix	RNA-seq RAW data processing	Convert count files to matrix
calculateGeneExpressionValues	RNA-seq RAW data processing	Gene expression values calculation
RNAseqQA	Biomarkers Identification & Assessment	Expression matrix QA
getAnnotationFromEnsembl	Biomarkers Identification & Assessment	Retrieve information for a DEGs list
batchEffectRemoval	Biomarkers Identification & Assessment	Batch effect treatment (Combat, SVA)
limmaDEGsExtraction	Biomarkers Identification & Assessment	Biclass and multiclass DEGs extraction
dataPlot	Biomarkers Identification & Assessment	Plots different data information and results (boxplot, orderedBoxplot, genesBoxplot, heatmap, optimalClusters, knnClustering, confusionMatrix, classResults)
featureSelection	Biomarkers Identification & Assessment	Feature selection for a DEGs matrix (mRMR,RF)
knn_CV	Biomarkers Identification & Assessment	Run a knn-CV for a DEGs matrix
knn_test	Biomarkers Identification & Assessment	Run a knn-test
rf_CV	Biomarkers Identification & Assessment	Run a rf-CV for a DEGs matrix
rf_test	Biomarkers Identification & Assessment	Run a rf-test
svm_CV	Biomarkers Identification & Assessment	Run a svm-CV for a DEGs matrix
svm_test	Biomarkers Identification & Assessment	Run a svm-test
DEGsToDiseases	DEGs Enrichment methodology	Related diseases for a DEGs list (targetValidation, genes2Diseases)
geneOntologyEnrichment	DEGs Enrichment methodology	Gene ontology for a DEGs list
DEGsPathwayVisualization	DEGs Enrichment methodology	Pathway visualization for a DEGs list

samples can be automatically processed. It is highly recommended to run the raw data alignment in a computer cluster as the use of the tools involves high computational cost for this task.

Biomarkers identification & assessment

The main goal of this type of studies is the search of potential biomarkers with the capability to discern samples among different conditions like control and disease or several diseases groups. To achieve this goal, *KnowSeq* counts with a step that allows to do that task for any specie and genetic disease. Moreover, our tool incorporates mechanisms to study the quality of the samples and the batch effects. It also includes the possibility of plotting all the required charts for the graphical assessment of the samples (e.g. boxplots by samples, boxplots by genes, heatmaps...) in a unique function named as *dataPlot*.

Although the output of the *KnowSeq* aligner step can be used as input of this step, the user can also introduces its own samples matrix. *KnowSeq* has been designed as a modular tool, this meaning that the user can carry out all the study by using *KnowSeq* or can use only the steps in which the user has interest.

The DEGs extraction is a very delicate process because the samples must pass a strong quality analysis and batch effect removal steps. If these steps are wrongly performed, the DEGs candidates would not be true DEGs due to the possible intrinsic deviations of the samples. To solve that, *KnowSeq* has a quality analysis step by using *arrayQualityMetrics* package adapted to RNA-seq by running the function *RNAseqQA*. This package counts on several statistical analysis to detect possible outliers in the samples [23]. Furthermore, our tool also has graphical representation such as gene expression boxplots disordered and ordered by class or label, heatmaps and gene by gene boxplot even allowing multiclass representation. It is very crucial to perform the quality analysis in a rigorous manner to ensure the correct development within the rest of the study. Even though the quality analysis is well done, there still exists the possibility of having batch effect 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 hour, lab technician, among others) and it is very hard to treat [24]. Nowadays there are different algorithms depending on whether the possible batch effect groups are known or not. *KnowSeq* allows to use two of the most relevant algorithms to treat batch effect such as ComBat for predefined batch groups and *sva* for unknown batch groups [25] through the function *batchEffectRemoval*. The correct quality analysis and batch effect assessment ensure the robustness of the DEGs candidates, because those DEGs would be expressed or inhibited due to the biological effect and not by possible deviations in the samples.

In order to perform the DEGs extraction, *limma R package* is used, with the peculiarity that *KnowSeq* automatically detects the number of different classes or labels and consequently applies limma biclass or multiclass [26]. Limma is the well-known library to determine DEGs by applying an rma statistical model. For the multiclass, we introduce the coverage parameter that allows to detect DEGs that are expressed for more than one biclass comparison. This coverage is deeply explained in one of our previous publication about leukemia multiclass biomarkers assessment [9]. This DEGs extraction is carried out by using the function *limmaDEGsExtraction*.

Theoretically, the final DEGs candidates are genes with the capability to discern among the classes to study. However, to assess and improve this consideration, a machine learning process is implemented in *KnowSeq*. This step is sub-divided in two, a feature selection process and a supervised machine learning classification process.

Next, a feature selection process is highly recommended for precision medicine to reduce the system complexity, diminishing the number of genes and, helping to make clinical decisions [27–29]. For that, *KnowSeq* allows to apply with the function *featureSelection*, two different feature selection algorithms, *minimum Redundancy Maximum Relevance* (mRMR) [30] and *Random Forest* as feature selector (RFfs) [31]. These algorithms create a ranking of DEGs in order to increase the classification rate by putting the DEGs with more information for the classifier listed at the top.

Finally, for the supervised machine learning process, *KnowSeq* allows to use three of the most relevant classifiers: *Support Vector Machine* (SVM) [32, 33], *k-Nearest Neighbour* (k-NN) [34] and *Random Forest* (RF) [35]. There exist two versions for each classifier in *KnowSeq*, one version with cross-validation, also known as CV (*knn_CV*, *svm_CV* & *rf_CV*) in which the user decides the number of fold and the data partitions always considering the representation of all the classes, and other version for testing (*knn_test*, *svm_test* & *rf_test*), by using a test dataset without CV independent from the dataset used for the DEGs extraction and CV assessment. Furthermore, for the three algorithms the hyperparameters are optimised, searching the acquisition of the best model for each analysis. Moreover, *KnowSeq* allows to plot the graphical representation of the results, including the confusion matrix, the sensitivity, the specificity and the f1-score. This gives to the user the possibility to perform a complete analysis and assessment of the addressed problem in a very simple and quick way.

DEGs enrichment methodology

Our tool is designed to automatise the knowledge extraction whatever is being the disease and for that, the last step of *KnowSeq* pipeline attains biological knowledge related to the final DEGs candidates. This knowledge must be interpreted by a clinician or a person with biological profile. In this sense, *KnowSeq* can retrieve information from three different sources to help with that interpretation. One of these sources is the *Gene Ontology (GO) enrichment* with information about the biological functions and locations of the DEGs [36, 37]. 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 in order to perform a more deeply study trying to find connections with the addressed disease. For the GOs enrichment, the *topGO* R package is used [38] and, in order to carry out the GOs retrieval, *KnowSeq* has the function *geneOntologyEnrichment*.

The second source of biological information is the pathway visualisation. Nowadays it is well known that the interaction of several genes whether can lead to a genetic disorder or not. Genes interacting among them in the same biological function are distributed 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 the sames pathways of those DEGs. The *pathview* package allows to colour the pathways depending on the expression values of the genes inside the pathways [39]. *KnowSeq* has kept this idea to automatically retrieve and colour all the pathways related with the final DEGs candidates and listed in the *KEGG database* [40]. With this implementation, it is easy to know if the expression of the DEGs and the surrounding genes are affecting a critical function in the disease development. The function *DEGsPathwayVisualization* takes care of this process.

The last source of biological information implemented in *KnowSeq* is the related diseases retrieval and it is performed executing the function *DEGsToDiseases*. In this step, all the related diseases of the DEGs candidates listed in the literature are obtained with the purpose of finding possible relation with the pathology addressed and with other possible precursor pathologies. Furthermore, the diseases related with a set of DEGs are also obtained in order to find possible DEGs that are related with the same pathology. This information can be attained from two different sources: the first one is the *Gene Set to Diseases* web platform [41] and the second one is the *targetValidation* [42] web platform. Then, the acquired diseases are correctly formatted by *KnowSeq* to do more readable this information for the user.

With the information collected by *KnowSeq* automatically from the three different sources, a strong biological enrichment process is done in order to build a biological profile for each of the DEGs without requiring external tools.

Results and Discussion. Breast Cancer Application

After the application of *KnowSeq* to the collected breast cancer data, very promising results were retrieved and will be discussed in this section. The section is divided in four subsections, one for the information about the data acquisition and three representing categories of results that were obtained for this study. For the last three subsections, the first one is focused on the final candidate DEGs extraction and the restrictions imposed to achieve them. The second one shows the assessment of those DEGs by using machine learning techniques with the main goal of finding a smaller sub-set of DEGs. Finally, the last one describes the enrichment of the sub-set of DEGs in order to find relevant biological information about them in an easy way by using *KnowSeq*.

Data preparation & description

To describe the main functionality of *KnowSeq*, a study case has been developed. All the data or samples used in this research come from *The Cancer Genome Atlas* (TCGA) and have been acquired through 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. For this breast cancer study, 90 patients were selected with the condition of having BAM files from both solid normal and primary tumour tissues for each patient. With this condition the paired datasets are ensured, achieving the best quality conditions in terms of samples for the study. Primary breast cancer is a tumour that still remains inside the breast or the lymph nodes (glands) under the arm. On the

other hand, the solid normal tissue is collected from the adjacent healthy tissue to the primary breast tumour. A table with all patient data to replicate the study is available in the supplementary information file named as Supplementary Table 1. In order to perform a more robust study, two different datasets will be taken into account. The first dataset is formed by 80 patients and will be used to extract the DEGs. The second dataset is conformed for the 10 remaining patients and will be only used for testing those DEGs in a machine learning step. Thanks to this division, the DEGs extracted will be independent of the samples used to assess them.

Gene expression analysis

The importance of achieving robust biomarkers is crucial for this type of problems. A mistake in the process could lead to a wrong selection and assessment of the DEGs and thus turn into an error in the machine learning diagnosis process. To avoid or minimise this error rate, it is necessary to carefully follow a strong pre-processing and quality analysis step. Also, it is important to correctly select the imposed restriction to extract the set of DEGs. To find them, as mentioned before, 80 patients were used and the 10 remaining patients were kept only for testing those DEGs in the machine learning step. This separation is very important to test the DEGs in patient never seen before in the process, bringing robustness to the results and avoiding overfitting. The quality analysis was first performed to the 80 patients and no outlier was detected among them. Then, the batch effect removal step was applied taking into account that the possible batches were unknown. The SVA algorithm [43] was performed to find the surrogate variables in order to create a model considering those variable to remove the deviations. It is critical to remove the batch effect in order to correct the data but without removing any possible deviations caused by biological processes. 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 two well-known statistics values for filtering: the Log Fold Change (LFC) greater or equal than 3 and the P-value less or equal than 0.001. Applying these restrictions, a total amount of 50 DEGs candidates ordered by LFC were extracted.

Table 2 shows those DEGs with several statistical values that describe at numerical level why those genes have been selected as DEGs. Those values are five statistics: The log-fold change ($logFC$) that represents the average difference in expression between the two groups to study (tumour and normal). If $\mid logFC \mid \geq 2$ it means that one of the group is at least over-expressed an average of 1 time more than the other group, thus exists statistical differences between both groups. The second value is the the moderated t-statistic, which has the same interpretation than the normal t-statistic but the standard errors have been reduced between the genes, effectively obtaining information from the set of genes to help with inference about each individual gene. The next value is the P-Value ($P.Value$) which represents the probability of obtaining a result equal or higher than what it was observed when the null hypothesis is true. The adjusted P-Value indicates which proportion of comparisons within a family of comparisons (hypothesis tests) are significantly different. The B-statistic (B) is the log-odds that a given gene is differentially expressed. As it can be seen, all the DEGs candidates pass the imposed restrictions and they will be assessed to corroborate their validity.

figure2.eps

Figure 2: Heatmap of the 50 DEGs candidates clearly showing differences between tumour and normal samples.

Furthermore, the Figure 2 represents an expression heatmap that graphically shows important differences of the DEGs candidates between both tumour and normal samples.

Machine Learning assessment

Traditionally, the gene expression studies were only focused on the DEGs candidates extraction. However, *KnowSeq* also includes a machine learning step to assess those DEGs and their capability to discern among the considered pathologies. Through this process, a smaller subset of DEGs can be achieved with the purpose of finding a more reduced gene signature candidate. For that, *KnowSeq* has three different supervised machine learning algorithms and two different feature selection methods as was explained before. This machine learning step has two different approaches. The first one is the application of a CV process to assess the DEGs with the training patients. The second one is the test process in which our DEGs are evaluated by using the 10 test patients previously chosen only for this purpose.

Firstly, a 10-CV step was applied in order to see the behaviour of the classifier with the 80 patients training dataset when those DEGs are used for classify. Thereupon, all the different combination of classifiers with feature selection algorithms reached better results than without applying feature selection, recognising all the training samples with a few number of genes. SVM and RF acquired outstanding results, but 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 patients (20 samples) datasets. These patients were left out at the beggining for all the study to be used now to assess the DEGs. Different matches, or combinations between classifiers and feature selection algorithm, were executed with the purpose of searching the combination with the best results. Those combinations are the possible permutations resulting from the different classifiers (SVM, k-NN and RF) with none feature selection (No F.S.) and with the different feature selection algorithms (mRMR and RF f.s.). Table 3 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. SVM also reached 100% with RF f.s. but no with mRMR. For its part, RF only achieved 100% with 10 genes and by using RF f.s. algorithm. Although all of them achieved prominent results, k-NN obtained the best results whatever being 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 recognised for the machine learning designed models. This means that *KnowSeq* brings the support to create intelligent systems with the capability of extracting relevant biomarkers that are useful to discern among the addressed diseases or states.

Table 2: Table with the 50 DEGs candidates extracted for this study and several statistical values for those DEGs.

	logFC	AveExpr	t	P.Value	adj.P.Val	B
COL10A1	-7.1720	14.8062	-23.9116	1.3885e-20	2.0014e-18	37.0692
CST1	-6.8780	12.0813	-15.2775	2.2569e-15	4.2166e-14	24.9740
MMP13	-6.6652	12.6433	-23.5212	2.1842e-20	2.9420e-18	36.6168
LINC01614	-6.5380	14.5261	-20.4209	1.0349e-18	6.4411e-17	32.7475
SLC24A2	-6.2606	10.8512	-19.2851	4.8451e-18	2.2617e-16	31.1924
COL11A1	-5.8136	16.1869	-26.8715	5.4897e-22	2.0147e-19	40.2812
MMP11	-5.5748	15.7695	-27.1724	4.0272e-22	1.6695e-19	40.5878
CA4	5.4495	12.3041	22.7570	5.4093e-20	5.8144e-18	35.7099
IBSP	-5.4158	9.7764	-17.3262	8.4221e-17	2.4182e-15	28.3072
PLPP4	-5.4069	11.0745	-20.3021	1.2119e-18	7.3562e-17	32.5887
MMP1	-5.2966	12.5466	-17.7029	4.7652e-17	1.4990e-15	28.8834
LEP	5.2911	15.4111	12.7145	2.3314e-13	2.3836e-12	20.2601
MYOC	5.2690	11.2076	15.8110	9.2770e-16	1.9577e-14	25.8761
NPY2R	5.1723	12.0594	14.2605	1.3176e-14	1.9586e-13	23.1819
EPYC	-5.1495	9.8030	-16.9509	1.5009e-16	4.0510e-15	27.7222
LINC00922	-4.9341	9.0520	-19.3696	4.3080e-18	2.0622e-16	31.3109
CST2	-4.8586	11.9503	-15.2360	2.4212e-15	4.4841e-14	24.9026
CST4	-4.7626	10.3939	-14.0907	1.7854e-14	2.5427e-13	22.8731
CD300LG	4.7297	15.0205	27.8248	2.0795e-22	1.1035e-19	41.2410
ANGPTL7	4.6405	13.1176	14.4855	8.8454e-15	1.3878e-13	23.5869
SCARA5	4.6061	14.4664	23.1422	3.4131e-20	4.0065e-18	36.1706
ADGRD2	4.5781	9.9209	19.5686	3.2720e-18	1.6478e-16	31.5882
AC044784.1	-4.5405	12.4070	-14.0387	1.9606e-14	2.7528e-13	22.7780
OPRPN	4.4570	11.9701	10.6963	1.4700e-11	9.6064e-11	16.0423
GLYAT	4.4221	11.4900	14.8858	4.4038e-15	7.5592e-14	24.2953
LINC01705	-4.4206	8.2294	-21.5106	2.5216e-19	2.0538e-17	34.1667
AC093895.1	-4.4142	7.2190	-13.6822	3.7519e-14	4.8207e-13	22.1182
DLK1	4.4100	11.1670	11.1259	5.8440e-12	4.1775e-11	16.9813
DSCAM-AS1	-4.3949	12.0297	-8.8010	1.1385e-09	5.0128e-09	11.6176
PLAC1	-4.3843	8.8461	-15.0985	3.0582e-15	5.4733e-14	24.6655
COMP	-4.3195	15.8240	-17.6860	4.8879e-17	1.5286e-15	28.8577
LINC02408	-4.3143	7.0213	-13.9943	2.1243e-14	2.9515e-13	22.6965
AC104407.1	4.2835	13.0120	13.0863	1.1424e-13	1.2825e-12	20.9859
PITX1	-4.2759	14.3922	-18.0804	2.7203e-17	9.3295e-16	29.4503
CXCL2	4.2568	17.2496	17.7874	4.1995e-17	1.3449e-15	29.0112
WIF1	4.2474	12.1155	13.5579	4.7191e-14	5.8834e-13	21.8850
PLIN4	4.2328	18.3265	16.4032	3.5554e-16	8.5226e-15	26.8485
CCL11	-4.2158	12.6651	-16.2890	4.2684e-16	9.9858e-15	26.6633
VEGFD	4.1739	13.6021	21.6667	2.0704e-19	1.7581e-17	34.3646
CSN1S1	4.1379	11.1700	5.9985	1.6242e-06	4.1544e-06	4.2666
LRRC15	-4.1185	16.6620	-18.0938	2.6671e-17	9.2062e-16	29.4702
CIDEC	4.0943	16.1264	12.6589	2.5971e-13	2.6166e-12	20.1503
AC112721.2	-4.0904	9.5582	-19.0034	7.1918e-18	3.1717e-16	30.7939
CNTNAP2	-4.0853	17.0082	-12.7531	2.1636e-13	2.2371e-12	20.3361
S100P	-4.0752	15.6872	-10.5751	1.9148e-11	1.2194e-10	15.7732
ADIPOQ	4.0540	17.3106	10.6432	1.6498e-11	1.0663e-10	15.9248
WT1	-4.0456	9.7736	-9.5566	1.9007e-10	9.7380e-10	13.4377
GPD1	4.0364	18.0172	13.6989	3.6389e-14	4.7058e-13	22.1493
CHRNA6	-4.0287	8.9468	-13.2380	8.5741e-14	9.9333e-13	21.2777
TRHDE-AS1	4.0243	12.2240	12.4349	4.0261e-13	3.8499e-12	19.7042

Table 3: Table that contains the test results for the different combinations of feature selection algorithms with the classifiers depending on the number of DEGs selected.

	No F.S.			**mRMR**			**RF f.s.**		
n. Genes	**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%

Once the classification is done, it is very helpful to see graphically the gene expression differences that exist between the tumour samples and the normal samples for the three 3 DEGs that discriminate perfectly the test patients. In order to carry out this representation, *KnowSeq* counts contains the *dataPlot* function in mode genesBoxplot. Figure 3 represents the genes Boxplots for the top 3 DEGs without apply feature selection (ordered by LFC), applying mRMR and applying RF f.s. respectively. In this figure, the first gene selected by the three methods (No F.S, mRMR and RF f.s.) is the same (COL10A1). However, the second gene selected by mRMR and RF f.s.(VEGFD & MMP11), both are different than the second gene with more LFC (CST1). The third and last gene selected by mRMR and RF f.s. (PITX1, LINC01614), are also different again than the third gene with higher LFC (MMP13). Nevertheless, even though the genes selected by LFC have more differences in average expression between the states, the genes selected by the feature selection algorithms discern better between such states, thus reaching better classification results as can be seen in Table 3. Consequently, adding a refined feature selection as well as a classification algorithm based on machine learning technology proved that the selected DEGs potentially improve the differentiation of states against classical metrics like LFC.

It is a priority in this research to minimise the number of genes and maximise the final achieved accuracy. This way, a very small subset of DEGs can be found to have the capability of discerning among the studied states. Nevertheless, *KnowSeq* is flexibly prepared to use and analyse as many genes or DEGs as the user requires. Also, it is important to highlight that, even though a bi-class problem was taken into account for this study, *KnowSeq* is designed to analyse any multiclass problem. In this sense, the confusion matrix, the f1-score, the sensitivity and the specificity metrics calculation are considered by our package.

figure3.eps

Figure 3: Boxplots of the 3 first DEGs selected by KnowSeq without feature selection algorithm and with mRMR and RF.

DEGs enrichment

At this point of the study, our DEGs have been assessed by applying a machine learning process. Nevertheless, those DEGs must be interpreted at biological level by experts in the field. In order to help with the biological interpretation, *KnowSeq*

has a last step in its pipeline created solely and exclusively to this purpose (DEGs Enrichment). Although this study searches a very small subset of DEGs, the enrichment step in *KnowSeq* does not depend on the number of DEGs, because the package can compute all of them.

Previously, in the machine learning results, the 10 test patients were totally recognised with only 3 genes selected by both RF f.s. and mRMR in conjunction with the k-NN classifier. Firstly, the relationship between those 6 DEGs and breast cancer will be searched by using the function *DEGsToDiseases* with the *targetValidation* platform selected. This platform has several scores to determine if a gene is related with the different possible diseases based on the information collected by the web platform. Those scores increase when the association increases too, that meaning a strong association with the selected disease.

From the 9 DEGs commented before, 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 No F.S., mRMR and RF f.s. (COL10A1). The 6 remaining DEGs have no relation or the relation is poor (a very low association score). It is very interesting to note that only the first gene of the top 3 DEGs without feature selection has important relations with breast cancer, although they are the DEGs with the higher LFC or P-value. Therefore, the use of a feature selection step in this case has remarkably supposed 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 sub-section. Hence, the 3 breast cancer reported DEGs will be used for the enrichment. For these DEGs, a set of scores are showed in Table 4. These 4 scores acquire values between 0 and 1: the Literature score is calculated based on the evidences in the literature of the involvement of a gene with the corresponding cancer (breast cancer in this case); the RNA Expression score uses data from Expression Atlas to see if a gene has differences at expression level for a disease; the Affected Pathways score evaluates from the reactome platform if the gene is involved in relevant pathways for the disease. Lastly, the final association score is calculated from the previous scores. 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.

Table 4: Table with the information about the association scores for the final 3 DEGs to study.

Gene	Liter. Score	RNA Exp. Score	Affected Paths. Score	Final Score
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

Once the disease relationship process has been carried out, the next step is the Gene Ontology enrichment. For this process the same 3 DEGs are used and the five most important GOs for the three DEGs and for the three different ontologies (BP, MP & CC) will be retrieved with the function *geneOntologyEnrichment*. Table 5 shows the top 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 increasing the maximun number of retrieved GOs, GOs related to the VEGFD were 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 knowing the biological domain of each DEGs and studying possible relations with processes that could lead to develop cancer.

Table 5: Table that contains top 5 GOs for the three different ontologies for the 3 final DEGs

Ontology	GO.ID	Term	GO_Genes	Description
BP	GO:0001501	skeletal system development	COL10A1	The process whose specific outcome is the progression of the skeleton over time, from its formation to the mature structure. The skeleton is the bony framework of the body in vertebrates (endoskeleton) or the hard outer envelope of insects (exoskeleton or dermoskeleton).
	GO:0016043	cellular component organization	COL10A1,MMP11	A process that results in the assembly, arrangement of constituent parts, or disassembly of a cellular component.
	GO:0030198	extracellular matrix organization	COL10A1,MMP11	A process that is carried out at the cellular level which results in the assembly, arrangement of constituent parts, or disassembly of an extracellular matrix.
	GO:0043062	extracellular structure organization	COL10A1,MMP11	A process that is carried out at the cellular level which results in the assembly, arrangement of constituent parts, or disassembly of structures in the space external to the outermost structure of a cell. For cells without external protective or external encapsulating structures this refers to space outside of the plasma membrane, and also covers the host cell environment outside an intracellular parasite.
	GO:0071840	cellular component organization or bioge...	COL10A1,MMP11	A process that results in the biosynthesis of constituent macromolecules, assembly, arrangement of constituent parts, or disassembly of a cellular component.
MF	GO:0005198	structural molecule activity	COL10A1	The action of a molecule that contributes to the structural integrity of a complex or its assembly within or outside a cell.
	GO:0005201	extracellular matrix structural constitu...	COL10A1	The action of a molecule that contributes to the structural integrity of the extracellular matrix.
	GO:0030020	extracellular matrix structural constitu...	COL10A1	A constituent of the extracellular matrix that enables the matrix to resist longitudinal stress.
	GO:0043167	ion binding	COL10A1,MMP11	Interacting selectively and non-covalently with ions, charged atoms or groups of atoms.
	GO:0043169	cation binding	COL10A1,MMP11	Interacting selectively and non-covalently with cations, charged atoms or groups of atoms with a net positive charge.
CC	GO:0005581	collagen trimer	COL10A1	A protein complex consisting of three collagen chains assembled into a left-handed triple helix. These trimers typically assemble into higher order structures.
	GO:0005783	endoplasmic reticulum	COL10A1	The irregular network of unit membranes, visible only by electron microscopy, that occurs in the cytoplasm of many eukaryotic cells. The membranes form a complex meshwork of tubular channels, which are often expanded into slitlike cavities called cisternae. The ER takes two forms, rough (or granular), with ribosomes adhering to the outer surface, and smooth (with no ribosomes attached).
	GO:0005788	endoplasmic reticulum lumen	COL10A1	The volume enclosed by the membranes of the endoplasmic reticulum.
	GO:0031012	extracellular matrix	COL10A1,MMP11	A structure lying external to one or more cells, which provides structural support, biochemical or biomechanical cues for cells or tissues.
	GO:0032991	protein-containing complex	COL10A1	A stable assembly of two or more macromolecules, i.e. proteins, nucleic acids, carbohydrates or lipids, in which at least one component is a protein and the constituent parts function together.

Finally, the last biological enrichment step included in *KnowSeq* is the pathways enrichment. Pathways involving our DEGs are interesting 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 *DEGsPathwayVisualization*. This function makes use of KEGG database to acquire the pathways 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. Figure 4 shows the pathway hsa04974 that is related with the colagen gene (COL10A1) and performs the Protein digestion and absorption process. In the figure, the collagen box shows a clearly difference between the tumour samples (red) and the normal samples (green). This means that the COL10A1 gene could activate erroneous processes inside the pathway depending on its expression. Table 6 shows

figure4.eps

Figure 4: Pathway hsa04974 in which the COL10A1 gene is involved. As can be seen in the pathway, the collagen box indicates a strong expression change in the tumour samples in comparison to the normal samples.

the nine VEGFD related pathways as well as the pathway related with COL10A1 gene.

Table 6: Table that contains the retrieved pathways with their description for the final DEGs.

DEGs	ID	Description
COL10A1	hsa04974	Protein digestion and absorption
VEGFD	hsa04010	MAPK Signaling Pathway
	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

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.

When all the enrichment pipeline of *KnowSeq* is over, this information is used to find and learn more about those DEGs and their relation with breast cancer. For that, it is very important that all these details will be studied and analysed by experts in bioinformatics and biology. *KnowSeq* expects to provide a very powerful and useful tool for those experts that could retrieve the most crucial information for the DEGs based on their expression in an easy and adaptable way.

Conclusions

In this paper, a new tool publicly available at Bioconductor to carry out RNA-seq raw and pre-procesed data analysis has been presented. *KnowSeq* includes the traditional steps in this type of studies but also implements a feature selection and machine learning step and an enrichment step. Thanks to this, complete analyses can be done from RAW data up to the biological knowledge extraction in a easy, modular and flexible way.

Furthermore, in order to present a case of study with the tool, a breast cancer problem has been addressed with BAM files automatically downloaded with *KnowSeq* from GDC Portal. A total of 80 patients with paired samples (Normal-Tumour) were used to extract the DEGs candidates. Those DEGs candidates were

assessed through a machine learning step with the purpose of finding a very reduced sub-set of DEGs with the capability to discern between normal and tumour samples. Furthermore, different feature selection algorithms were applied in order to find a better order of those DEGs to improve the classification rate. Finally, the DEGs were assessed by using 10 patients never seen before, achieving outstanding results since all the patients were totally recognised with only three DEGs for several combinations of classifiers and feature selection algorithms.

Then, a final sub-set of three DEGs were enriched by using the *KnowSeq* functions designed with this purpose. Those DEGs have a strong relation with breast cancer, there exist evidences at gene expression level, in the literature and in affected pathways that link the final three enriched DEGs with the disease. Furthermore, in order to know more about them, a list of GO terms were retrieved and a list of pathways, in which the expression changes of those DEGs could lead to several biological disorders.

In view of these considerations and by way of conclusion, *KnowSeq* is a R package that gives the possibility to carry out RNA-seq analyses in an easy way with all the required steps included in the pipeline. *KnowSeq* expects to serve as a novel tool to help to the experts in the field to acquire robust knowledge and conclusions for the data and diseases to study. *KnowSeq* has three clear strengths: the first one is the modular design, because the analyses can be started from different points (fastq, bam, count and even a custom expression matrix); the second one is the versatility due to the different algorithms for machine learning and feature selection and the different databases implemented in *KnowSeq*; and the last one is the adaptability of the analyses, because *KnowSeq* allows to use data from different sources and, even select different parameters that give to the user a real control of the pipeline.

Declarations

Funding

This research has been supported by the project: RTI2018-101674-B-I00 (from the Spanish Ministry of Economy and Competitiveness –MINECO– and the European Regional Development Fund. –ERDF).

Acknowledgment

The results published here are in part based upon data generated by The Cancer Genome Atlas managed by the NCI and NHGRI. Information about TCGA can be found at http://cancergenome.nih.gov

Ethics

Not applicable.

Consent to publish

Not applicable.

Competing interest

Not applicable.

Authors' contributions

DCS is the main author of this research and the manuscript. JMGG and DCS analyzed the data. LJHM, FMO and IRR conducted the experiments. All authors have read and approved the final manuscript.

Availability of data and materials

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Author details

[1]Department of Computer Architecture and Technology, University of Granada, Periodista Rafael Gómez Montero, 2, 18014 Granada, Spain. [2]Clinical Bioinformatics Area, Fundación Andaluza Progreso y Salud (FPS), Hospital Universitario Virgen del Rocío, Avenida Manuel Siurot s/n, 41013 Sevilla, Spain.

References

1. WHO: Breast Cancer (2018). https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/
2. Seelbinder, B., Wolf, T., Priebe, S., McNamara, S., Gerber, S., Guthke, R., & Linde, J. (2019). GEO2RNAseq: An easy-to-use R pipeline for complete pre-processing of RNA-seq data. bioRxiv, 771063.
3. Lohse, M., Bolger, A. M., Nagel, A., Fernie, A. R., Lunn, J. E., Stitt, M., & Usadel, B. (2012). RobiNA: A user-friendly, integrated software solution for RNA-Seq-based transcriptomics. Nucleic acids research, 40(W1), W622-W627.
4. Chao, K. H., Hsiao, Y. W., Lee, Y. F., Lee, C. Y., Lai, L. C., Tsai, M. H., ... & Chuang, E. Y. (2019). RNASeqR: an R package for automated two-group RNA-Seq analysis workflow. arXiv preprint arXiv:1905.03909.
5. Gómez-López, G., Dopazo, J., Cigudosa, J. C., Valencia, A., & Al-Shahrour, F. (2019). Precision medicine needs pioneering clinical bioinformaticians. Briefings in bioinformatics, 20(3), 752-766.
6. Castillo D., Galvez JM., Ortuno FM., Herrera LJ., Rojas I. (2019). KnowSeq: A R package to extract knowledge by using RNA-seq raw files. R package version 0.99.51. Bioconductor. DOI:10.18129/B9.bioc.KnowSeq.
7. Castillo D., Galvez JM., Herrera LJ., San Roman B., Rojas F. and Rojas I. (2017). Integration of RNA-seq data with heterogeneous Microarray data for breast cancer profiling. *BMC Bioinformatics*. **18**(1). DOI:10.1186/s12859-017-1925-0.
8. Galvez JM., Castillo D., Herrera LJ., San Roman B., Valenzuela O., Ortuno FM., et al (2018). Multiclass classification for skin cancer profiling based on the integration of heterogeneous gene expression series. *PLoS ONE*. **13**(5):1V. DOI:10.1371/journal.pone.0196836.
9. Castillo D., Galvez JM., Herrera LJ., Rojas F., Valenzuela O., Caba O., Prados J. and Rojas I. (2019). Leukemia multiclass assessment and classification from Microarray and RNA-seq technologies integration at gene expression level. *PLoS ONE*. **14**(2), e0212127. DOI:https://doi.org/10.1371/journal.pone.0212127.
10. Gonzalez S., Castillo D., Galvez JM., Rojas I. and Herrera LJ. (2019). Feature Selection and Assessment of Lung Cancer Sub-types by Applying Predictive Models. International Work-Conference on Artificial Neural Networks, 883–894, Springer.
11. Grossman, Robert L., Heath, Allison P., Ferretti, Vincent, Varmus, Harold E., Lowy, Douglas R., Kibbe, Warren A., Staudt, Louis M. (2016) Toward a Shared Vision for Cancer Genomic Data. New England Journal of Medicine, 375:12, 1109-1112.
12. Barrett T., Troup DB., Wilhite SE., Ledoux P., Rudnev D., Evangelista C., et al (2007). NCBI GEO: mining tens of millions of expression profiles—database and tools update. *Nucleic Acids Research*, **35**(suppl 1):D760–D765.
13. Brazma A., Parkinson H., Sarkans U., Shojatalab M., Vilo J., Abeygunawardena N., Holloway E., Kapushesky M., Kemmeren P., Lara G. et al (2003). ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Research*, **31**(1), 68-71.
14. Cock, P. J., Fields, C. J., Goto, N., Heuer, M. L., and Rice, P. M. (2009). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic acids research*, **38**(6), 1767-1771.
15. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... and Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**(16), 2078-2079.
16. Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology*, **14**(4), R36.
17. Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature methods*, **12**(4), 357.
18. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods*, **14**(4), 417.
19. Nicolas L Bray, Harold Pimentel, Páll Melsted and Lior Pachter (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, **34**, 525–527.
20. Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**(2), 166-169.
21. Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**(1), 139-140.
22. Hansen, K. D., Irizarry, R. A., and Wu, Z. (2012). Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics*, **13**(2), 204-216.

23. Kauffmann, A., Gentleman, R., and Huber, W. (2008). arrayQualityMetrics—a bioconductor package for quality assessment of microarray data. *Bioinformatics*, **25**(3), 415-416.
24. Goh, W. W. B., Wang, W., and Wong, L. (2017). Why batch effects matter in omics data, and how to avoid them. *BMC Bioinformatics*. 2005; **6**(1):191.
25. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., and Storey, J. D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, **28**(6), 882-883.
26. Smyth, G. K. (2005). Limma: linear models for microarray data. In Bioinformatics and computational biology solutions using R and Bioconductor (pp. 397-420). Springer, New York, NY.
27. Bolón-Canedo, V., Sánchez-Marono, N., Alonso-Betanzos, A., Benítez, J. M., & Herrera, F. (2014). A review of microarray datasets and applied feature selection methods. Information Sciences, 282, 111-135.
28. Li, T., Zhang, C., & Ogihara, M. (2004). A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics, 20(15), 2429-2437.
29. Liu, H., Li, J., & Wong, L. (2002). A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. Genome informatics, 13, 51-60.
30. Ding C, Peng H. (2005). Minimum redundancy feature selection from Microarray gene expression data; *Journal of Bioinformatics and Computational Biology*. **3**(2), p. 185-205.
31. R. Diaz Uriarte, S. Alvarez de Andres (2006). Gene Selection and classification of microarray data using Random forest. *BMC Bioinformatics*. **7**(1), p. 3.
32. Cortes C, Vapnik V. (1995). Support-vector networks. *Machine Learning*. **20**(3):273–297.
33. Noble WS. (2006). What is a support vector machine? *Nature Biotechnology*. **24**:1565 – 1567.
34. Parry R, Jones W, Stokes T, Phan J, Moffitt R, Fang H, et al. (2010). k-Nearest neighbor models for Microarray gene expression analysis and clinical outcome prediction. *The Pharmacogenomics Journal*. **10**(4):292.
35. Ho TK. (1995). Random decision forests. Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on. vol. 1. IEEE; p. 278–282.
36. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., ... and Harris, M. A. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, **25**(1), 25.
37. Gene Ontology Consortium. (2018). The gene ontology resource: 20 years and still GOing strong. *Nucleic acids research*, **47**(D1), D330-D338.
38. Alexa A, Rahnenfuhrer J (2019). topGO: Enrichment Analysis for Gene Ontology. R package version 2.36.0.
39. Luo, W., and Brouwer, C. (2013). Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics*, **29**(14), 1830-1831.
40. Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, **28**(1), 27-30.
41. Fontaine, J. F., and Andrade-Navarro, M. A. (2016). Gene set to diseases (gs2d): Disease enrichment analysis on human gene sets with literature data. *Genomics and Computational Biology*, **2**(1), e33-e33.
42. Koscielny, G., An, P., Carvalho-Silva, D., Cham, J. A., Fumis, L., Gasparyan, R., ... and Pierleoni, A. (2016). Open Targets: a platform for therapeutic target identification and validation. *Nucleic acids research*, **45**(D1), D985-D994.
43. Leek, J. T., and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics*, **3**(9), e161.

Figures

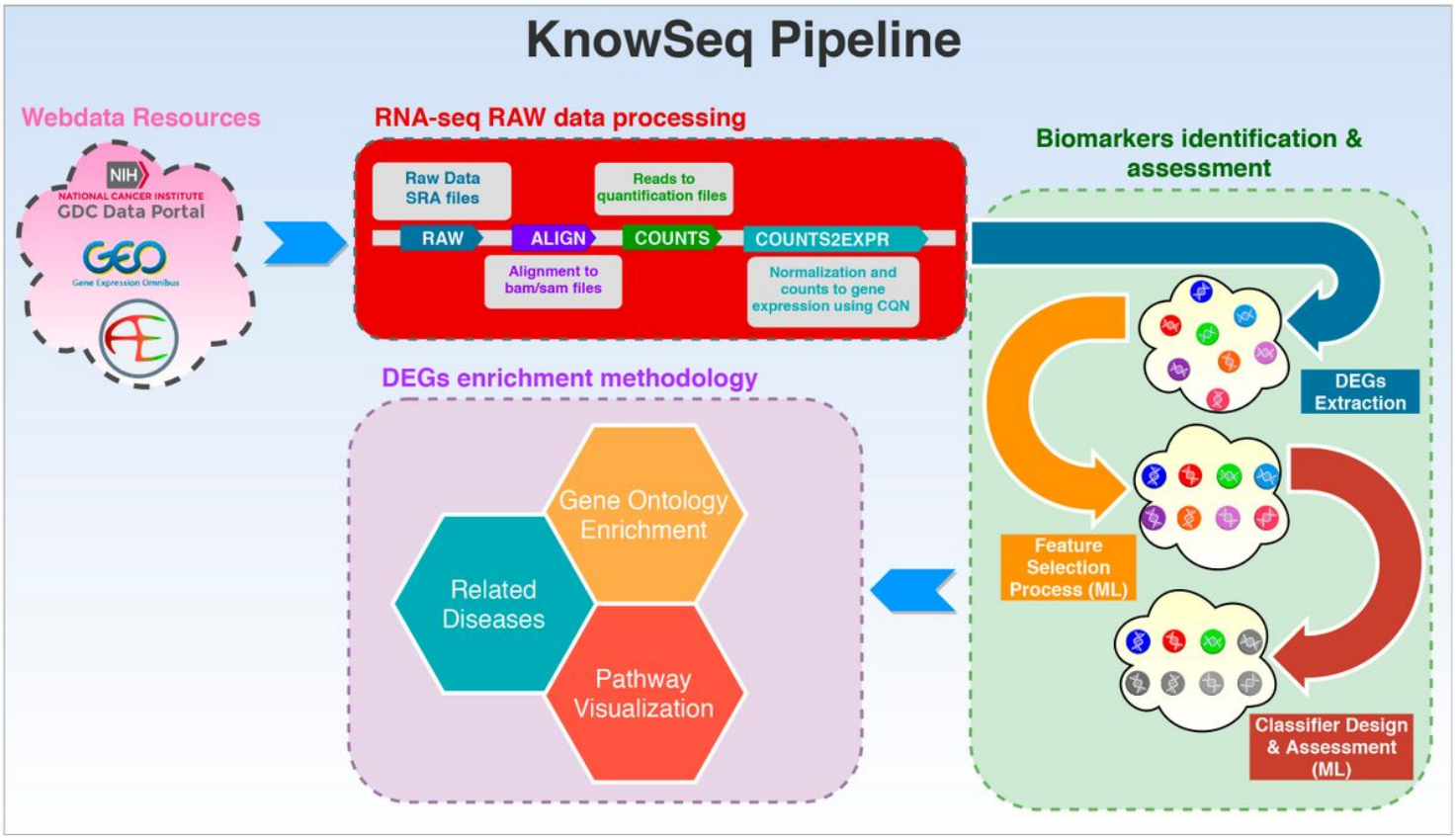

Figure 1

Pipeline implemented by KnowSeq R/bioc package. In the pipeline are the traditional steps in the RNA-seq data pipeline together with the new steps added by KnowSeq.

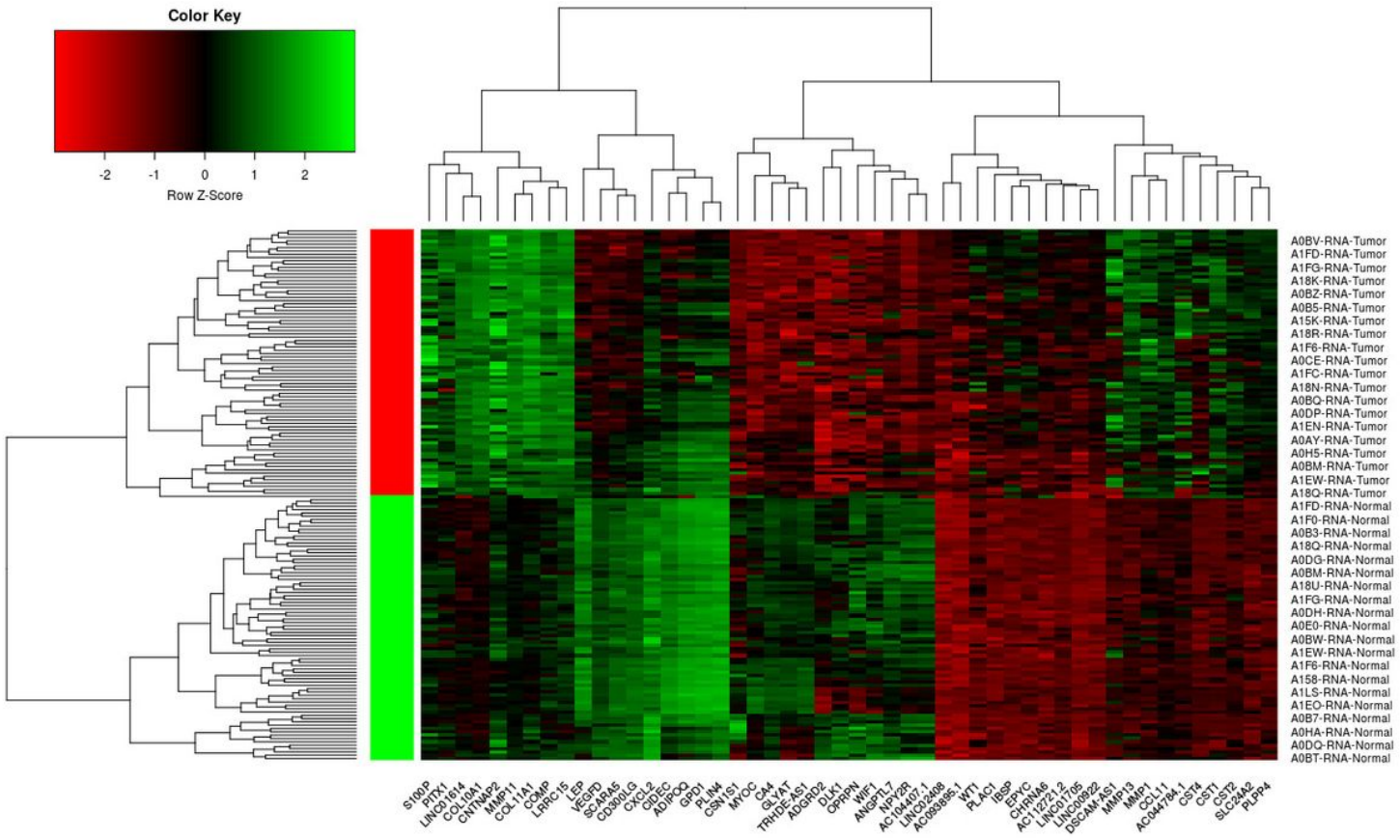

Figure 2

Heatmap of the 50 DEGs candidates clearly showing differences between tumour and normal samples.

Figure 3

Boxplots of the 3 rst DEGs selected by KnowSeq without feature selection algorithm and with mRMR and RF.

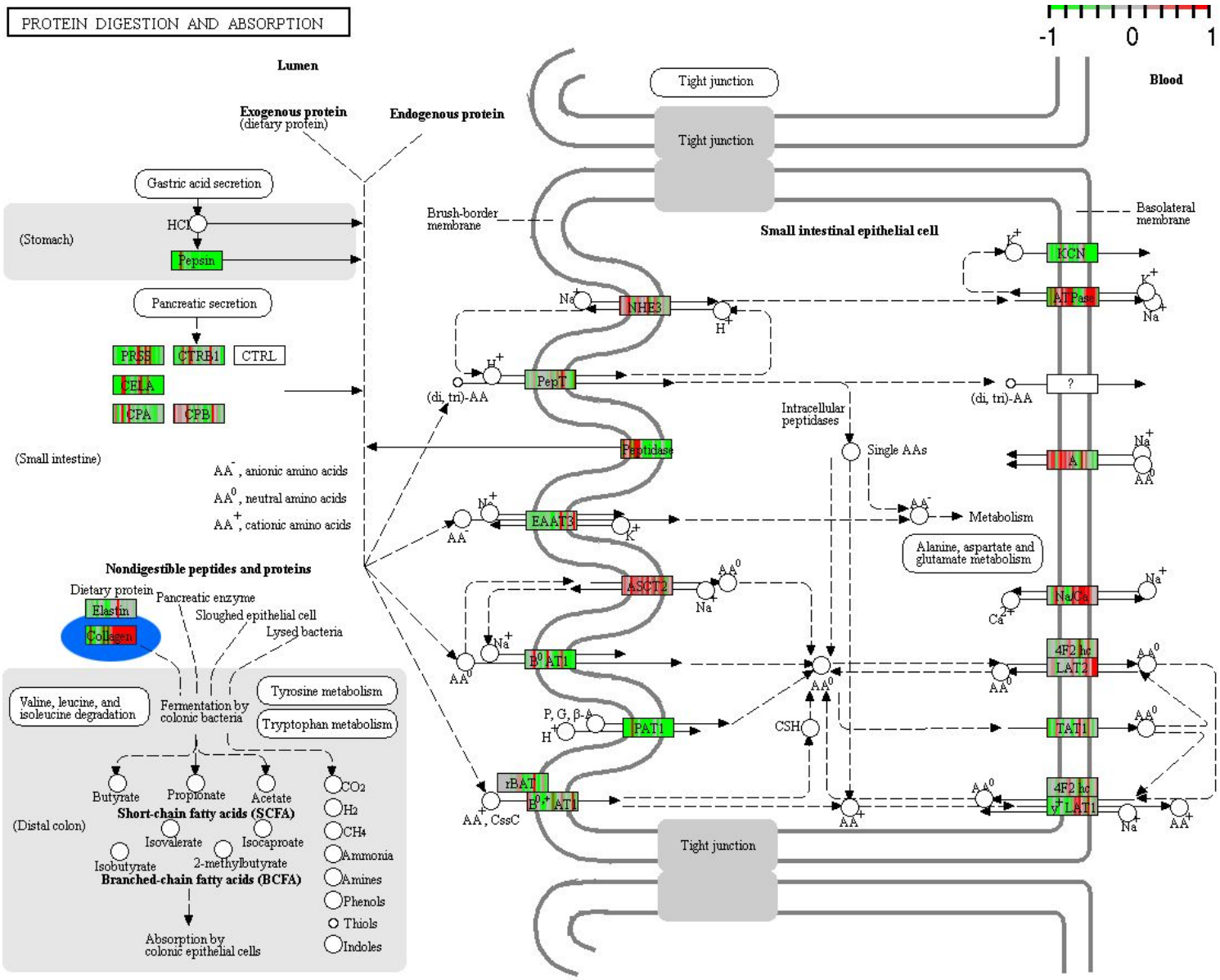

Figure 4

Pathway hsa04974 in which the COL10A1 gene is involved. As can be seen in the pathway, the collagen box indicates a strong expression change in the tumour samples in comparison to the normal samples.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.pdf