Heliyon 10 (2024) e27360

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Assessing the complementary information from an increased number of biologically relevant features in liquid biopsy-derived RNA-Seq data

Stavros Giannoukakos [a,b,c,*], Silvia D'Ambrosi [d], Danijela Koppers-Lalic [e], Cristina Gómez-Martín [f], Alberto Fernandez [g], Michael Hackenberg [a,b,c,**]

[a] Department of Genetics, Faculty of Science, University of Granada, Granada, 18071, Spain
[b] Bioinformatics Laboratory, Biomedical Research Centre (CIBM), PTS, Granada, 18100, Spain
[c] Excellence Research Unit "Modeling Nature" (MNat), University of Granada, Spain
[d] Department of Neurosurgery, Cancer Center Amsterdam, Amsterdam UMC, VU University, Amsterdam, 1081HV, the Netherlands
[e] Mathematical Institute, Leiden University, Leiden, 2333, CA, the Netherlands
[f] Department of Pathology, Cancer Center Amsterdam, Amsterdam UMC Location Vrije Universiteit Amsterdam, Amsterdam, 1081HV, the Netherlands
[g] Department of Computer Science and Artificial Intelligence, Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada, Granada, 18071, Spain

ARTICLE INFO

Keywords:
Liquid biopsy
Bioinformatics
Machine learning
Transcriptomics
RNA-Seq
lbRNA-seq
Cancer diagnostics
Normalisation
Ensemble learning

ABSTRACT

Liquid biopsy-derived RNA sequencing (lbRNA-seq) exhibits significant promise for clinic-oriented cancer diagnostics due to its non-invasiveness and ease of repeatability. Despite substantial advancements, obstacles like technical artefacts and process standardisation impede seamless clinical integration. Alongside addressing technical aspects such as normalising fluctuating low-input material and establishing a standardised clinical workflow, the lack of result validation using independent datasets remains a critical factor contributing to the often low reproducibility of liquid biopsy-detected biomarkers.

Considering the outlined drawbacks, our objective was to establish a workflow/methodology characterised by: 1. Harness the rich diversity of biological features accessible through lbRNA-seq data, encompassing a holistic range of molecular and functional attributes. These components are seamlessly integrated via a Machine Learning-based Ensemble Classification framework, enabling a unified and comprehensive analysis of the intricate information encoded within the data. 2. Implementing and rigorously benchmarking intra-sample normalisation methods to heighten their relevance within clinical settings. 3. Thoroughly assessing its efficacy across independent test sets to ascertain its robustness and potential utility.

Using ten datasets from several studies comprising three different sources of biological material, we first show that while the best-performing normalisation methods depend strongly on the dataset and coupled Machine Learning method, the rather simple Counts Per Million method is generally very robust, showing comparable performance to cross-sample methods. Subsequently,

* Corresponding author. Department of Genetics, Faculty of Science, University of Granada, Granada, 18071, Spain.
** Corresponding author. Department of Genetics, Faculty of Science, University of Granada, Granada, 18071, Spain.
E-mail addresses: sgiannoukakos@correo.ugr.es (S. Giannoukakos), s.dambrosi@amsterdamumc.nl (S. D'Ambrosi), d.koppers-lalic@lumc.nl (D. Koppers-Lalic), c.a.gomezmartin@amsterdamumc.nl (C. Gómez-Martín), alberto@decsai.ugr.es (A. Fernandez), hackenberg@go.ugr.es (M. Hackenberg).

https://doi.org/10.1016/j.heliyon.2024.e27360
Received 25 November 2023; Received in revised form 20 February 2024; Accepted 28 February 2024
Available online 12 March 2024

we demonstrate that the innovative biofeature types introduced in this study, such as the Fraction of Canonical Transcript, harbour complementary information. Consequently, their inclusion consistently enhances prediction power compared to models relying solely on gene expression-based biofeatures. Finally, we demonstrate that the workflow is robust on completely independent datasets, generally from different labs and/or different protocols. Taken together, the workflow presented here outperforms generally employed methods in prediction accuracy and may hold potential for clinical diagnostics application due to its specific design.

1. Introduction

Human body biofluids, such as blood, urine, and saliva, have proven to be a rich and valuable source of information about an individual's health status [1]. The burgeoning field of liquid biopsy (LB) research has been actively exploring these resources with the aim of unlocking their full potential for diagnosis, monitoring, prognosis, and treatment response assessment in various diseases, including cancer [2,3]. LB's most effective assets lie in its repeatability, cost-effectiveness, and minimal invasiveness.

Advancements in technology and computer science have propelled LB research. The advent of Next Generation Sequencing (NGS) technology, combined with continuous improvements in bioinformatics have deepened our understanding of the molecular landscapes in LB samples, revealing insights into disease mechanisms and the discovery of potential biomarkers [4–7]. Several studies investigating blood-based biosources like Tumour-Educated Platelets (TEPs), Extracellular Vesicles (EVs), Circulating Epithelial Cells (CECs), and Circulating Tumour Cells (CTCs) have notably unveiled a range of diagnostic signatures that hold promise for the early detection of prominent cancers, harnessing the power of mRNA sequencing (mRNA-Seq) [8–10]. For instance, Antunes-Ferreira et al. [11], report an Area Under the ROC Curve (AUC) of 0.88 through an 881 RNA biomarker panel to predict outcomes in Non-small Cell Lung Carcinoma patients.

Despite the great promise and certain advancements in the field, the current focus on LB-based transcriptomics has predominantly been centred around gene expression profiling, partly due to the lack of comprehensive pipelines tailored for laboratories with limited bioinformatics resources. Consequently, biofeatures such as isoform expression, fraction of canonical transcript (FoCT), gene fusion, RNA editing, and single nucleotide variants (SNVs) have remained largely uncharted, representing untapped sources of valuable insights. Specifically, the simultaneous usage of these biofeatures has not been explored in previous research, promising substantial potential for enhancing our understanding of LB data.

Moreover, the predominant application of cross-sample normalisation methods, while beneficial for prediction accuracy in enclosed, frequently single lab study designs, fosters challenges for clinical applications where intra-sample normalisation is mandatory to classify individual clinical samples applying a fixed prediction model. Additionally, the lack of independent test sets in many studies raises concerns about the reproducibility and generalisability of reported results, rendering metrics like AUCs potentially misleading.

The untapped wealth of information within LB-derived RNA-Seq (lbRNA-Seq) data presents an opportunity for more comprehensive and clinically relevant insights. To fully unlock the potential of LB data, a comprehensive exploration of complementary biological information available in a lbRNA-seq sample is imperative in order to gain an encircled understanding of such biosources. However, integrating this high amount of heterogeneous data into a single prediction model is challenging. Moreover, due to the high-dimensional nature of lbRNA-seq data, machine learning (ML) approaches have become indispensable for detecting patterns and gaining a deeper understanding of the underlying biological conditions [12].

In this study, we introduce ELLBA (**E**nsemble **L**earning for **L**iquid **B**iopsy **A**nalysis), a methodology designed to tackle the complexities of LB data and enhance the predictive modelling of patient, with applicability to clinical settings. ELLBA encompasses six biologically motivated feature types: gene expression, isoform expression captures alternative splicing, FoCT quantifies predominant transcript shifts, gene fusion detects structural changes, RNA editing indicates post-transcriptional modifications, and SNVs unveil potential mutations. Each biofeature type addresses different molecular properties that can be altered in pathologies like cancer, offering therefore diagnostic and prognostic value. In the context of existing literature, it is noteworthy that virtually all published LB-based studies are based on Gene Expression or, at most, SNVs. No comparable study or methodology exists harnessing the complementary information contained in six different biofeatures providing a unified decision output and applicability to clinical data. This innovative strategy distinguishes our approach, marking a significant advancement in the field of lbRNA-Seq analysis. Given the absence of a comparable workflow, our methodology focuses on comparing the final ensemble output to the standard Gene Expression results. Finally, the modelling part of the methodology utilises Ensemble Classification Methods to combine the complementary information from these features.

ELLBA was rigorously evaluated across six datasets and four independent validation sets, encompassing around 2,500 samples, covering various cancer types and biosources. Our work highlights the utility of the rather simple intra-sample Count Per Million (CPM) normalisation in clinical settings. We show that while the best normalisation method depends both on the data type and employed machine learning model, in general, CPM performs equally well compared to more sophisticated cross-sample methods. Moreover, our study demonstrates that Ensemble Learning is effectively leveraging the complementary information contained in the different biofeature types, always improving the prediction power over the best individual biofeature type. Interestingly, the improvement seems to be especially pronounced when evaluating independent test sets, which might indicate the robustness and reproducibility of the discriminative biofeatures detected by ELLBA. In summary, our workflow improves prediction accuracy and

Table 1
Implementation details on the six biofeature types utilised in this study.

Biofeature type	Biological rationale
Gene expression	Gene-level expression profiles provide information on the transcriptional activity of each gene in the sample. It is a measure of how active a gene is and determines the abundance of RNA molecules produced from that gene. Gene expression plays a crucial role in determining an organism's traits and functions. Consequently, perturbations in gene expression, driven by diseases, can lead to substantial alterations.
Isoform expression	Isoform expression is the measurement of different splice variants or isoforms of a gene's RNA transcripts. Alternative splicing allows genes to produce multiple isoforms with sometimes different functional characteristics. Isoform expression profiling reveals gene product diversity and potential disease associations.
FoCT	FoCT is designed to assess the predominant canonical transcript shift in each gene using a default group. Changes in the frequency of alternative splicing events are quantified by means of the fraction of the canonical transcript. The rationale behind this metric is that in cancer, frequently the splicing pathway is affected, increasing the transcriptional variation for at least certain genes. Under this scenario, it might be less important to correctly identify the different isoforms but to robustly quantify the existence of a differential amount of alternative transcripts.
Gene fusion	Gene fusion detection involves identifying abnormal fusion events between two genes, which can arise from chromosomal rearrangements or translocations. Fusion events can create chimeric RNA transcripts or fusion proteins that are often associated with disease.
RNA editing	RNA editing is a post-transcriptional modification process that alters the nucleotide sequence of RNA molecules, leading to changes in the encoded protein or functional non-coding RNA. This process is crucial for expanding the functional diversity of the transcriptome and can impact gene regulation, protein structure, and function.
SNV	Single Nucleotide Variants (SNVs) can alter protein structure, function, or gene regulation based on their location in the coding or regulatory sequences. SNV analysis aids in discovering disease-associated (driver) mutations.

streamlines clinical decision-making, contributing to personalised cancer care (Fig. S1).

2. Materials and methods

2.1. Workflow and implementation

The ELLBA workflow can be easily installed using a Docker image. The source code and exact installation instructions are available on GitHub. The workflow integrates established bioinformatics tools with novel algorithms for data processing, biofeature generation, and ML analysis. ELLBA was primarily developed using Python (v3.8) with supplemental R (v3.6.3) scripts. ML analysis relies on the scikit-learn (v1.2.0) Python package [13]. Table S1 presents a comprehensive summary of all the software and packages utilised in the workflow, along with their corresponding versions.

We employed data from 10 different liquid biopsy studies, encompassing a total of 2,479 publicly available samples from the SRA repository [14]. The studies span different types of LB data, including TEPs, EVs, and CECs. To initiate the ELLBA workflow, in addition to the raw fastq files, a sample sheet specifying at least the sample name and group label (e.g., control, cancer) is required. Although we conducted rigorous benchmarking in terms of normalisation and ML, this section outlines the final workflow configuration.

2.2. Data preprocessing and biofeature extraction

Artificial adapter sequences and low-quality reads (average Q below 20 or shorter than 40 nt) are automatically detected and removed by the BBDuk (v38.18) tool [15]. Furthermore, the workflow provides multi-sample quality reports through MultiQC (v1.13) [16].

Genome mapping was performed by the STAR (v2.7.6a) aligner [17] using the human reference genome GRCh38.p13 primary assembly as well as the GENCODE v35 reference gene annotation [18]. STAR was employed with 'GeneCounts' and 'TranscriptomeSAM' parameters to obtain count matrices at both gene and transcript levels. Alignment quality metrics were extracted using RSeQC (v3.0.0) and Picard tools (v2.23.3) [19,20] to evaluate the alignment process and a final summarised report is being generated.

Based on the STAR-generated BAM files, we generate a total of six biologically motivated feature types: (1) Gene expression, (2) Isoform expression, (3) FoCT, (4) Gene fusion, (5) RNA editing and, (6) SNV. A detailed overview of each biofeature type can be found in Table 1.

2.2.1. Gene expression

To quantify gene expression, we collected read abundances from the GeneCounts-based generated files and created a single expression matrix that encompassed the quantification results for all samples. Subsequently, we performed a principal component analysis (PCA) on the gene expression matrix to identify potential batch effects in the data. Furthermore, an Interquartile Range (IQR) analysis was employed to identify any potential outlier samples.

Following the exploratory analysis, we utilised the filterByExpr function from the edgeR (v3.28.1) package [21] to remove genes with low expression levels across samples. Subsequently, we applied the standard CPM normalisation followed by the MinMaxScaler function, from the sklearn package, to further transform the gene expression data ensuring that all feature values are on a comparable scale.

2.2.2. Isoform expression

For quantification at a transcript level, we employed Salmon (v1.9.0) software [22] in alignment-based mode, using STAR

TranscriptomeSAM output. Default parameters were employed, along with additional settings including seqBias and gcBias enabled for sequence and GC content biases correction, libType U for unstranded data, and 100 bootstrap iterations for robust quantification. The resulting transcript-level expression matrix was obtained by merging quantification outputs using Salmon's quantmerge script, selecting the numreads column for the final matrix.

Similar to gene expression analysis, we performed exploratory analysis, normalisation, and transformation on the transcript-level expression matrix. These steps followed the same approach as described in section 2.2.1 for gene expression analysis.

2.2.3. Fraction of canonical transcript

We randomly selected 20 control individuals and extracted the most abundant transcript of each gene (canonical transcript) based on the isoform expression levels in these samples. This list of transcripts was used to convert the transcript expression matrix into a canonical transcript matrix. The transformed matrix considered only the most abundant transcripts, dividing their expression levels by the total counts of all transcripts originating from the same gene. This yielded the fraction of canonical transcript for each sample, which was then combined into a single matrix. Any features with missing values (NA) were entirely removed. Finally, the StandardScaler function was applied to transform the frequency feature matrix.

2.2.4. Gene fusion

Gene fusions were identified using the Arriba (v2.1.0) software [23,24], with specific parameters adjusted within the STAR aligner for fusion gene detection. The Arriba software was utilised with default parameters. Following detection, gene fusions were sorted alphabetically, and gene IDs were appropriately adjusted. Specifically, the first gene ID within the fusion nomenclature was retained, while the second part was omitted. Furthermore, instances of the same first gene ID were merged into a unified entry, resulting in a count matrix. This matrix was subsequently transformed into a binary format, featuring exclusively 0 or 1 values. In this binary configuration, a value of 0 denotes the absence of fusion detection within a sample, while a value of 1 signifies its presence.

2.2.5. RNA editing events

The genome-aligned BAM files were subjected to de-duplication using the rmdup function of samtools (v1.7) [25]. Additionally, GATK BaseRecalibrator (v4.1.9.0) [26] was employed to recalibrate the base quality scores. Subsequently, BCFtools mpileup (v1.7) [27] was utilised on the pre-processed BAM files with the following parameters: min-MQ 15, min-BQ 15, redo-BAQ, per-sample-mF, and min-ireads 2. Variant calling was performed using BCFtools call with the parameters ploidy GRCh38, variants-only, and multiallelic-caller. Known common variants from the dbSNP [28] database (common_all_20180418) were excluded from the analysis.

RNA editing events were detected using REDItools (v2.0) software [29]. The software was configured with the following parameters: min-edits 2, min-read-quality 18, and min-base-quality 15. RNA editing events were filtered at the sample level based on a minimum depth per site of 10, a mean quality score per site of at least 20, and a minimum substitution frequency of 0.3. All RNA editing events that passed the quality filters were merged into an overall feature matrix and further filtered to include only events that were present in at least 20% of the samples. The resulting RNA editing event matrix was transformed into a binary format, where 0 indicated the absence of an event and 1 indicated its presence.

2.2.6. Single nucleotide variants

Up to the common variant (SNP) filtering step, an identical protocol to RNA editing analysis was followed. After this stage, positions with quality scores below 20 and a minimum depth of 2, along with previously identified RNA editing sites, were filtered out. BCFtools merge was then employed to consolidate the filtered VCF files, producing a unified matrix across all samples. Subsequently, GATK VariantsToTable (v4.1.9.0) [30] extracted the genotype field for each variant from the filtered VCF file, converting the data into a tab-delimited table format. SNV values were discretised as follows: 0 for no alternative allele, 0.5 for heterozygous calls, and 1 for homozygous positions. Lastly, variants occurring in less than 20% of samples were filtered out.

2.3. Machine learning and ensemble learning implementation

To be able to generate a robust model that may classify each input sample within the datasets, ML techniques were employed on each of the six distinct biofeature spaces extracted from the data. Each dataset was initially split into training and test sets. For datasets with an independent external validation dataset, this was designated as the test set, while the main dataset served as the training set. In the absence of an external validation set, a random 70-30 split was performed, with 70% of the data used for training and the remaining 30% as an approximation to an independent test set.

Feature selection and model training were conducted on the training set, followed by final validation on the test set. Initially, highly correlated features (Pearson's correlation above 0.8) and quasi-constant features (with 99% similarity) were removed from further analysis.

The filtered training biofeature matrices underwent feature selection using the GeneticSelectionCV function, a Python implementation of the genetic algorithm (GA) [31]. The GA operates in a wrapper-like mode, systematically searching for the optimal set of features for the classification task. The choice to utilise the GA for feature selection stems from its efficiency in managing well vast high-dimensional biofeature spaces within a relatively short timeframe. As a population-based metaheuristic algorithm, the GA employs multiple candidate solutions during the search process. It excels in exploring diverse biofeature combinations comprehensively, proving notably faster and less computationally expensive—ideal for large-scale datasets. In contrast to standard methods like Recursive Feature Elimination (RFE) or Univariate methods, the GA offers unique advantages. RFE, while effective, can be

computationally demanding and time-intensive, especially with high-dimensional data, such as Isoform Expression with nearly 34,000 features. Univariate methods, on the other hand, may overlook intricate feature interactions, limiting their ability to capture nuanced patterns in the data. The GA, with its capacity to consider feature interactions and navigate vast feature spaces efficiently, provides a more robust and holistic approach to feature selection in the context of liquid biopsy-based datasets.

The GA was utilised with stratified 5-fold cross-validation (CV), employing empirical parameters including n_population = 130, n_generations = 130, scoring = "accuracy", and max_features = 50, a choice justified by balancing model complexity and prediction performance. For each training biofeature type matrix used in GA-based feature selection, an appropriate base estimator was selected. It's worth noting that the GeneticSelectionCV function employs the selected base estimator to evaluate the fitness of different feature subsets during the genetic algorithm process. Specifically, the Random Forest classifier was selected as the base estimator for Gene Expression, the SVM classifier for Isoform Expression, and the Logistic Regression classifier for the remaining biofeatures (Table S2).

Kindly take note that, going forward, when we refer to selecting or utilising an appropriate base estimator, we specifically mean choosing or using the underlying method that is employed by algorithms like GeneticSelectionCV or AdaBoost. This subtle yet crucial distinction is pivotal for comprehending how ensemble learning harnesses the unique strengths of distinct base estimators to enhance overall predictive accuracy.

In particular, for both feature selection and modelling, we explored a range of standard and diverse classifiers to identify the most suitable ones. This set included classifiers such as AdaBoost, K-Nearest Neighbours (KNN), support vector machine with a linear kernel (LinearSV), Logistic Regression, Naive Bayes, and Random Forest. Model training was conducted using stratified 5-fold CV. Table S2 provides a detailed overview of the final feature selection and classifier combinations.

Ensemble learning was employed to combine the individual information from all six distinct biofeature types and enhance the predictive performance of each sample. To ensure the inclusion of reliable features, a minimum mean accuracy score of 0.65 during cross-validation in the training process was set as an empirical eligibility criterion. The soft voting strategy was then applied to make the final label prediction, averaging the aggregating predictions based on the probability distribution of class labels.

2.4. Functional enrichment analysis

To perform functional enrichment analysis, we utilised the online GOst tool provided by the gProfiler [32] web service. This tool facilitated the Gene Ontology (GO) and pathway enrichment analyses on the genes selected through the GA for each biofeature type. Additionally, we conducted enrichment analysis on the combined biofeature set, incorporating all selected features.

3. Results

3.1. Data collection and description

Our study design comprises data from ten different LB-based studies encompassing a total of 2,479 samples [9,10,33–40]. The data

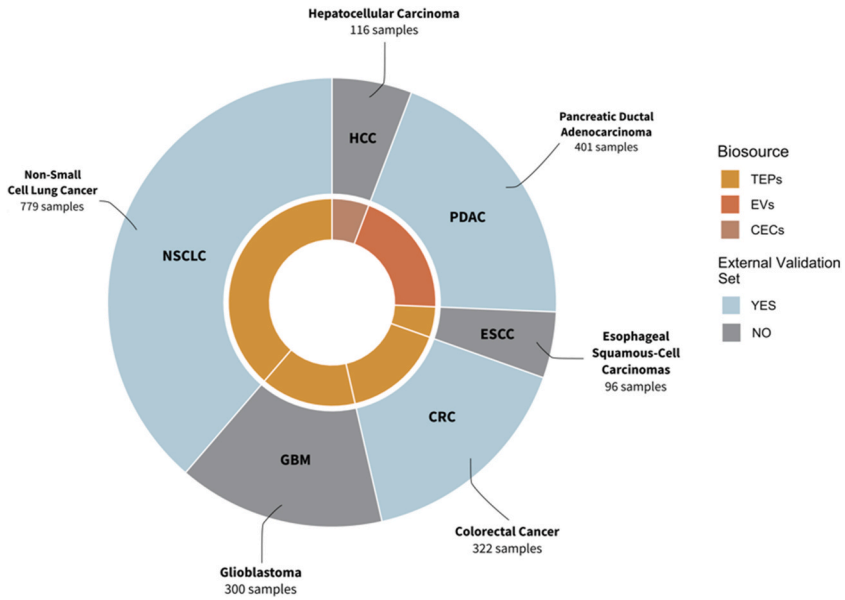
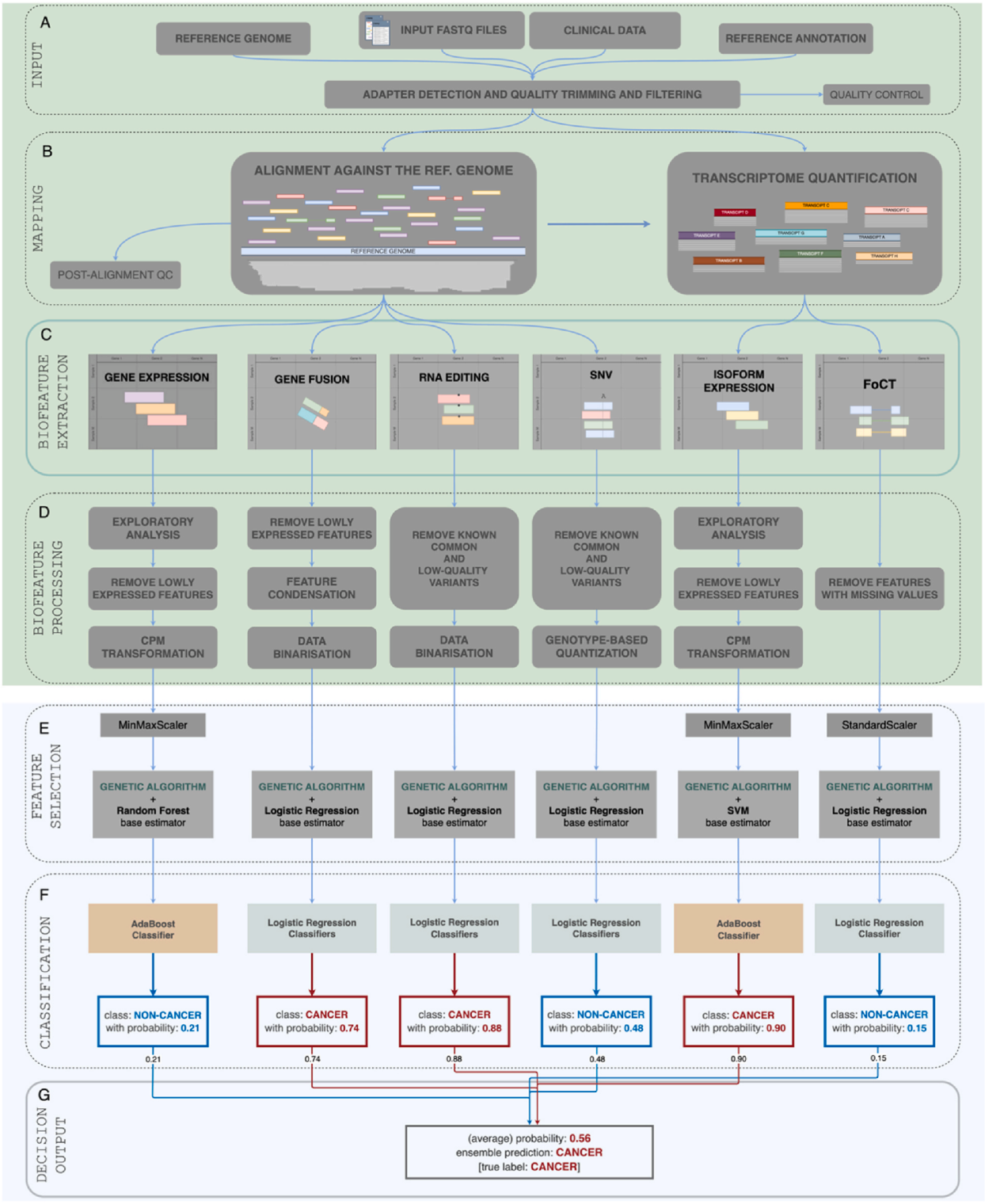


Fig. 1. A comprehensive overview of the datasets employed in this study, showcasing six distinct datasets: NSCLC, GBM, CRC, ESCC, PDAC, and HCC, as depicted in the outer donut plot. Light blue colouring (NSCLC, CRC, and PDAC) signifies datasets with independent external validation sets, while grey shading (GBM, ESCC, HCC) represents datasets without external validation. The inner circle categorises the biosource origin of each dataset: dark yellow for TEPs in NSCLC, GBM, CRC, and ESCC; cinnamon red for EVs in PDAC; and brown for CECs in HCC.



(caption on next page)

Fig. 2. Overview of the ELLBA Workflow. ELLBA methodology is divided into two key components, with bioinformatics analysis highlighted in light green and machine learning in light blue. The workflow comprises seven core modules, four of which are part of the bioinformatics analysis: Input, Mapping, Biofeature Extraction, and Biofeature Processing. The remaining three modules belong to the machine learning component and include Feature Selection, Classification, and Decision Output. The process commences with data Input, followed by alignment of raw data to the reference genome during Mapping. Subsequently, essential information is extracted across six feature types in the Biofeature Extraction stage. Each distinct biofeature then undergoes individual processing and machine learning analysis. To be more specific, following Biofeature Extraction, each biofeature is processed in the Biofeature Processing module, where data cleaning and normalisation or discretisation are managed. The remaining modules are associated with machine learning analysis. To elaborate further, each individually processed biofeature undergoes Feature Selection and Classification. In the final Decision Output, all individual classification outputs, based on the processed biofeatures, are combined using ensemble learning (soft voting) to produce a unified final decision.

were obtained from publicly available sources and consisted of short-read RNA sequencing (RNA-Seq) data. Various sequencing protocols, including mRNA-Seq, Extracellular Vesicles Long RNA Sequencing (exLR-Seq), single-cell RNA sequencing (scRNA-Seq), read lengths (100, 150, and 250), read types (single and paired-end), and sequencing methods (bulk and single-cell) were included into the analysis.

The collected data were derived from three distinct blood-extracted biosources: TEPs, EVs, and CECs. TEP-derived data comprise the majority accounting for over 1,900 samples, while approximately 450 samples were derived from EVs, and the remaining samples originated from CECs (Fig. 1).

Furthermore, our study was particularly focused on six different cancer types, each represented by a unique acronym: Non-small Cell Lung Carcinoma (NSCLC) for lung cancer, Glioblastoma multiforme (GBM) for brain cancer, Colorectal Cancer (CRC) for colon cancer, Esophageal Squamous-Cell Carcinoma (ESCC) for esophageal cancer, Pancreatic Ductal Adenocarcinoma (PDAC) for pancreatic cancer, and Hepatocellular carcinoma (HCC) for liver cancer. These well-defined cancer types serve as the foundation for our analyses, and we refer to each of the six datasets by their respective cancer type acronyms: NSCLC, GBM, CRC, ESCC, PDAC, and HCC.

To ensure the robustness of our analysis, we divided the collected studies into two main subsets: a training set and an independent external validation testing set, when available. The training set comprised samples from six of the ten studies, totalling approximately 2000 samples. The remaining four studies were exclusively used for the external validation testing set. To be more precise, within the set of six datasets, three (NSCLC, CRC, and PDAC) are accompanied by independent external validation datasets. For instance, in the NSCLC dataset, which encompasses 779 samples comparing NSCLC to non-cancer TEP samples, sequenced using an SE100 mRNA-Seq protocol, we identified an external validation dataset that perfectly aligns with the same sequencing protocol and cancer type. Similarly, the CRC dataset, comprising 322 samples, employed PE100 mRNA-Seq sequencing to distinguish between Colorectal Cancer and Non-Cancer samples. Its corresponding external validation set maintained the cancer type and utilised an SE100 mRNA-Seq protocol. In the context of PDAC, featuring 401 samples focused on Pancreatic Ductal Adenocarcinoma versus Healthy Controls, the sequencing followed a PE150 exLR-Seq protocol. The matching external validation set, derived from two publications, also maintained consistency in terms of cancer type and sequencing protocol. Table S3 provides a detailed overview of the datasets, including their configuration, utilisation, and accession information. Kindly consult the Supplementary Materials for a detailed account of the specific procedures and analyses applied to individual biofeatures in the utilised datasets.

3.2. Overview of the ELLBA methodology

The ELLBA methodology is organised into two core components: bioinformatics and machine learning. It consists a total of seven distinct modules. The initial four modules, namely Input, Mapping, Biofeature Extraction, and Biofeature Processing, constitute the bioinformatics phase of the analysis. The subsequent three modules, Feature Selection, Classification, and Decision Output, are focused on machine learning. The entire workflow is depicted in Fig. 2. Each module within this framework performs a specific set of tasks, which can be summarised as follows:

Input module: Adapter trimming and quality control. (Fig. 2A).

Mapping module: Genome alignment and gene profiling, including alignment quality controls. (Fig. 2B).

Biofeature extraction module: Six distinct biological features are extracted (Table 1): (i) gene-level expression profiles, (ii) isoform-level expression profiles, (iii) FoCT, (iv) gene fusion quantification, (v) RNA editing, and (vi) putative somatic SNV (Fig. 2C). Table S4 provides a numerical overview of the extracted biofeatures before any filtering.

Biofeature processing module: Normalisations and discretisation techniques are applied prior to filtering low-quality and non-discriminative biological features like lowly expressed genes or common germline variants (SNPs) (Fig. 2D).

Feature selection: For each specific biological feature type, feature selection is performed using the Genetic Algorithm in conjunction with a designated base estimator (Table S5) tailored to that particular biofeature type (Fig. 2E).

Classification module: Following feature selection, each biofeature classification using standard machine learning models. The class confidences generated are then retained for subsequent use in the final Decision Output module (Fig. 2F).

Decision Output module: To leverage the complementary information offered from each biofeature type, by default ensemble soft voting classification is applied. This method combines the predicted probabilities from all biofeature matrices, aggregating them into a single, consolidated average prediction. In this context (Fig. 2G), each predictive model generates a label (either "Non-cancer", highlighted in blue, or "Cancer", highlighted in red) along with an associated probability displayed beneath the respective label. During the soft voting process, all these output predictive probabilities are consolidated through averaging, culminating in the ultimate

decision (as demonstrated by "Cancer" in the figure). Additional details about soft voting can be accessed in the Supplementary Materials.

3.3. Comparative analysis of normalisation methods for gene and isoform expression data

Normalisation of gene and transcript expression data is a crucial step in analysing raw-count matrices. While the remaining biofeature types (FoCT, Gene fusion, RNA editing, and SNV) are inherently normalised as ratios or discrete values, count matrices require normalisation to account for variations in read yield and technical artefacts. Several methods with different assumptions have been implemented for this purpose [41], and their performance varies depending on whether these assumptions are met [42]. Most commonly, cross-sample normalisation methods are applied. Examples of such approaches include TMM and TMMwsp from the edgeR package, RLE from DESeq2, RUV from the RUVseq package (which can also address and rectify batch effects), as well as quantile-based methods like Full Quantile (FQ) and Upper Quartile (UQ) normalisation. While these cross-sample methods often exhibit superior performance in benchmark studies, they have a notable drawback: the normalisation outcome for a specific gene and sample is influenced by the values of other samples. This characteristic hampers their utility in clinical settings, where the objective is the normalisation of individual samples and its application to fixed prediction models.

To address this limitation, we evaluated the performance of two intra-sample normalisation methods, CPM and Reads Per Kilobase per Million mapped reads (RPKM), and the aforementioned six cross-sample normalisation methods for all six datasets. This evaluation was carried out by incorporating these normalisation methods into the "Biofeature processing module" step of the pipeline and assessing their impact on the results. To mitigate the influence of specific machine learning models, our analysis encompassed six diverse algorithms: AdaBoost, KNN, LinearSV, Logistic Regression, Naive Bayes, and Random Forest.

Performance assessment was conducted on all six training datasets applying a 5-fold cross-validation approach with exactly the same folds for each biofeature type. The average AUC is then used as the principal quality measure. The results of the gene expression normalisation, presented in Fig. 3, consistently demonstrated that CPM normalisation, despite variations between datasets and ML models, performed comparably to the more sophisticated cross-sample methods. Specifically, CPM normalisation yielded the highest collective mean AUC of 0.81 across all datasets, while RPKM exhibited the lowest performance with 0.58. RLE normalisation ranked second highest with a collective mean AUC of 0.79. On an individual dataset basis, CPM normalisation consistently exceeded other methods, except in the CRC dataset, where FQ normalisation achieved a slightly higher mean AUC (0.73) compared to CPM (0.70). It is worth noting that, despite employing RUVSeq normalisation to account for batch effect, our results showed that CPM normalisation

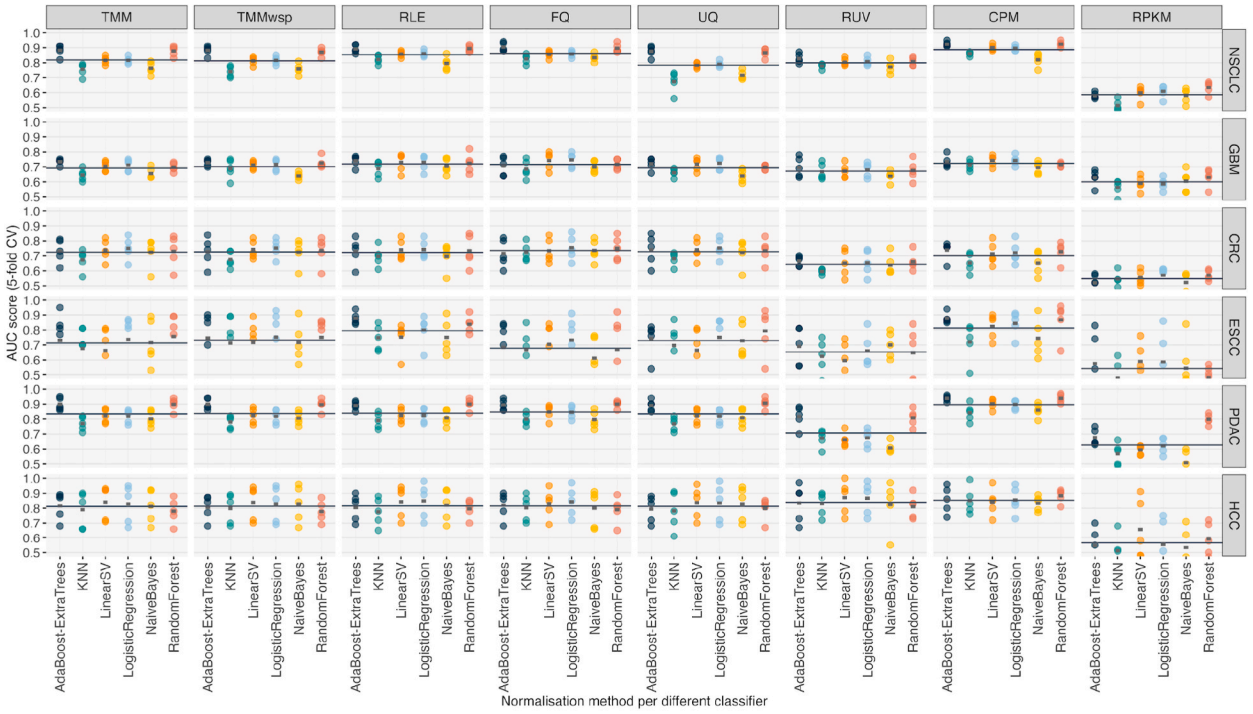


Fig. 3. Summary of Various Normalisation Methods. A total of eight normalisation techniques were assessed across all six datasets. Each column corresponds to a distinct normalisation method, while each row represents a different dataset employed. The x-axis illustrates the various models utilised within each normalisation. and the v-axis depicts the mean AUC score achieved through 5-fold CV. Each model is represented by dots indicating the AUC for each CV fold. Additionally, a dashed line indicates the mean AUC across the 5 folds. while a solid line represents the mean AUC across all models.

outperformed RUVSeq. This observation, indicating that batch effect correction did not significantly impact the ML analysis, highlights the robustness of CPM normalisation in combination with ML downstream analysis.

Similar trends were observed in the analysis of isoform expression normalisation (Fig. S2), following the same evaluation protocol as the gene expression analysis. More specifically, CPM normalisation demonstrated favourable performance, achieving the highest score with a collective mean AUC of 0.80 across all datasets, while RPKM displayed the lowest performance with 0.60. FQ normalisation was rated second, with a collective mean AUC of 0.78. Notably, when evaluating each dataset individually, CPM normalisation consistently outperformed the other methods, with the exception of the GBM and CRC datasets. In the GBM dataset, RLE normalisation achieved a slightly higher mean AUC of 0.72, compared to CPM's 0.70. Similarly, in the CRC dataset, TMMwsp exhibited a mean AUC of 0.66, while CPM achieved 0.65. Notably, CPM normalisation not only provided very robust results but also exhibited significant clinical value. Unlike other methods, CPM normalisation can normalise a single sample independently, making it more time-efficient for clinical applications.

3.4. Optimal classification models for the different biofeature types

Having established CPM as the default normalisation method for expression-based features, we extensively explored the performance of six diverse classifiers: AdaBoost, KNN, LinearSV, Logistic Regression, Naïve Bayes, and Random Forest, using all six biofeature types and datasets. These six classifiers were meticulously selected based on their algorithmic diversity, computational efficiency, and their capacity to provide a balanced approach to classification, incorporating both linear and non-linear techniques. Their selection also took into account the specific characteristics of our dataset and the nature of the features extracted. Emphasising diversity was crucial, as these classifiers comprise different modelling paradigms, extracting the most appropriate patterns and discrimination functions regarding the distinct feature spaces of the problems involved. Moreover, we have opted for the use of standard libraries and well-validated classification methods, ensuring the robustness and reliability of our software implementation. Furthermore, we acknowledge the need for diversity not only in individual classifier performance but also in the success and good behaviour of the multi-classification approach, i.e., ensemble combination. Recognising that the joint use of different methods ensures a good trade-off among their potentially different predictions, we have placed a strong emphasis on diversity to enhance the effectiveness of the ensemble method. This approach aligns with the need for a comprehensive and well-balanced strategy, considering the varying strengths of each classifier within the ensemble framework. Additionally, their widespread use in similar bioinformatics contexts enhances comparability with existing literature. The performance of each classifier was evaluated using the same stratified 5-fold cross-validation approach on each dataset, and the average AUC was reported (Fig. 4). To determine the most suitable classifier for

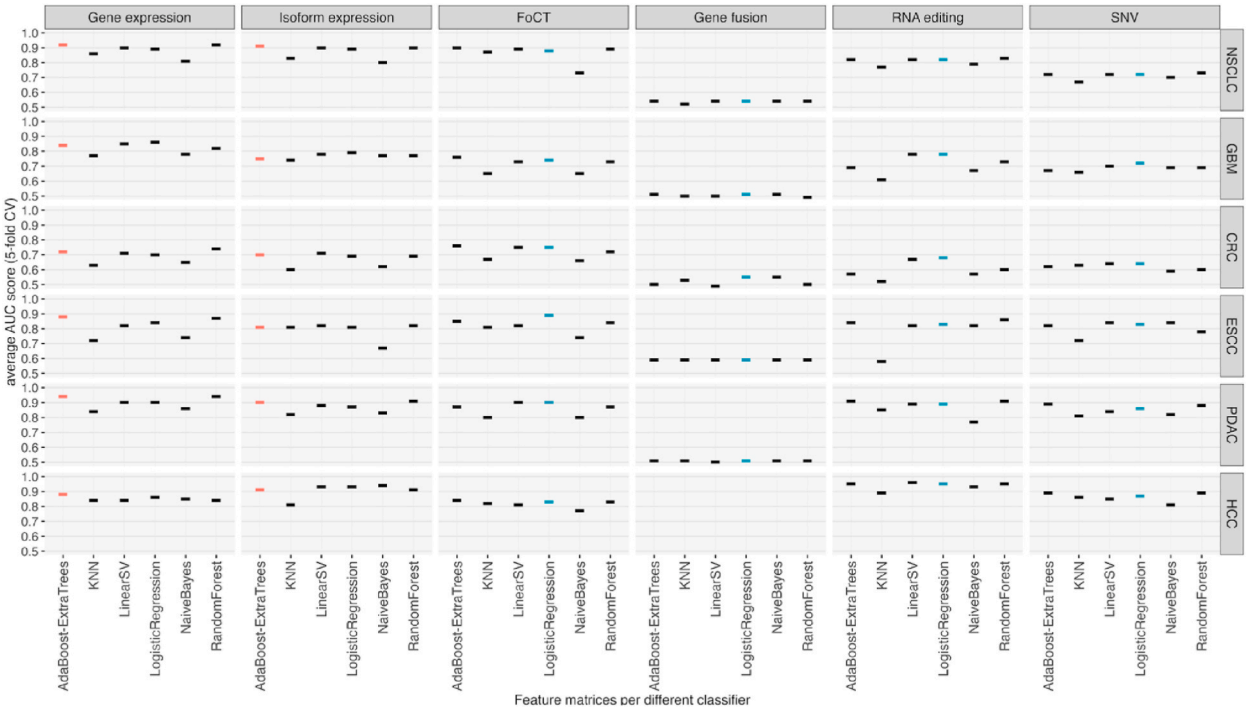


Fig. 4. Classifier Selection Overview. Each column in the plot represents a different feature type, while each row corresponds to a different dataset. On the x-axis, six classifiers are evaluated, and the y-axis displays the mean AUC score achieved through 5-fold CV. The dashed lines represent the average AUC score from the 5-fold CV evaluation. For the first two feature types (gene and isoform expression), AdaBoost with ExtraTrees, is highlighted in red, as it is better suited for these feature types. For the remaining feature types, Logistic Regression, is highlighted in blue, as it is better suited for these types of features.

each biofeature type, we calculated the average mean AUC across all datasets for that biofeature type. Our findings revealed that expression-based feature matrices yielded the highest performance when paired with the AdaBoost classifier using the ExtraTrees as base estimator. Specifically, AdaBoost achieved the highest mean AUC of 0.86 for gene expression, while KNN exhibited the lowest mean AUC of 0.78. LinearSV, Random Forest, and AdaBoost demonstrated similar performance for isoform expression with a collective mean AUC of 0.83, while KNN yielded the lowest mean AUC of 0.77. Consequently, we selected the AdaBoost classifier with the ExtraTrees base estimator as the optimal choice for both gene and isoform expression analyses.

Conversely, Logistic Regression emerged as the most suitable classifier for the remaining biofeature matrices, including FoCT, fusion genes, RNA editing, and SNVs. More specifically, Logistic Regression achieved the highest collective mean AUC of 0.83 for FoCT, while Naïve Bayes exhibited the lowest mean AUC of 0.72. In the case of fusion genes, both Logistic Regression and Naïve Bayes performed equally, yielding a collective mean AUC of 0.54, while LinearSV demonstrated the lowest performance with a collective AUC of 0.52. Regarding RNA editing and SNVs, Logistic Regression outperformed other classifiers, obtaining a collective mean AUC of 0.82 and 0.77, respectively, whereas KNN exhibited the lowest performance with a collective AUC of 0.70 and 0.72, respectively.

3.5. Enhancing predictive output through ensemble learning

Having thoroughly assessed the performance of each of the six biofeature types in isolation, we delved into the potential benefits of combining these diverse, high-dimensional biofeature spaces, each characterised by different scales and distributions. To do so, we adopted three ensemble combination techniques: soft voting, majority voting, and stacking, each with its unique approach to integrating predictions from multiple models. Soft voting relies on averaging probabilities, majority voting on majority decisions, and stacking employs a meta-model to optimise the fusion of predictions (see Supplementary Materials).

Our comprehensive evaluation spanned six datasets, assessing performance using key metrics: AUC and the percentage of misclassified samples (Fig. 5A-B). While AUC offers a broad measure of model performance, its limitation in discerning specific error types prompted us to incorporate misclassification rates for a more nuanced evaluation. Strikingly, ensemble learning consistently outshone individual models across all datasets. The selection of the most suitable ensemble technique varied depending on the dataset and the metric considered. For AUC, soft voting excelled in NSCLC, CRC, and ESCC, whereas stacking proved superior in GBM, PDAC, and HCC. Conversely, when evaluating the percentage of misclassified samples, NSCLC, CRC, PDAC, and HCC demonstrated the best outcomes. In GBM and ESCC, majority voting slightly outperformed soft voting. In some instances, multiple ensemble techniques performed equally well, such as majority voting and stacking in GBM and CRC, and soft voting and stacking in PDAC and HCC.

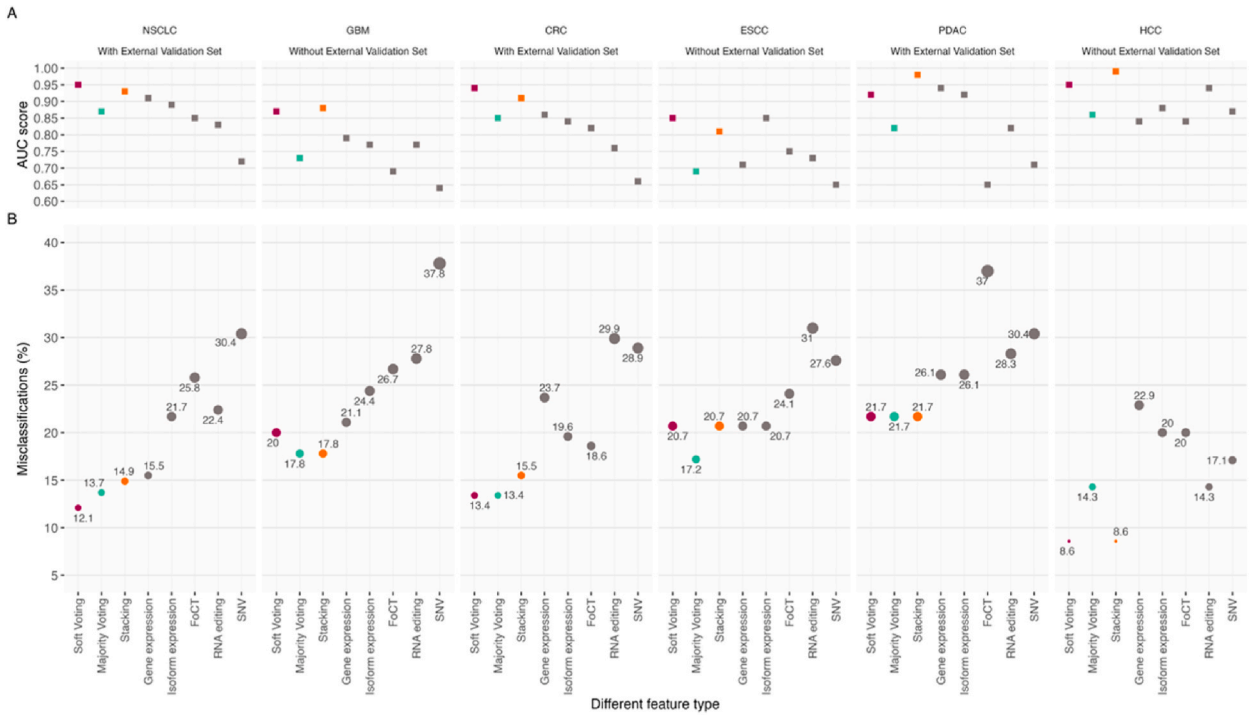


Fig. 5. Performance Overview of Feature Types and Ensemble Learning Methods Across Datasets. Each column in this figure corresponds to a specific dataset, with dataset information and the presence of an independent validation set indicated above. (A) The plot displays AUC scores for the test sets across all feature types. (B) Dot plots illustrate the percentage of misclassifications for each feature type across different datasets. The colourful squares and dots in both figures A and B represent the ensemble learning techniques, while the grey squares and dots represent the remaining feature types.

A deeper dive into specific datasets revealed intriguing findings. In NSCLC, soft voting achieved an impressive low misclassification rate of 12.1%, outperforming gene expression at 15.5%. In GBM, majority and stacking techniques led with a 17.8% misclassification rate, surpassing soft voting at 20% and gene expression at 21.1%. For CRC, both soft and majority voting achieved a low misclassification rate of 13.4%, while gene expression yielded a higher rate of 23.7%. In ESCC, majority voting reached 17.2%, while soft voting and other ensemble techniques, as well as gene and isoform expression, hovered around 20.7%. In PDAC, all ensemble techniques achieved a comparable rate of 21.7%, while gene expression exhibited a higher rate of 26.1%. Lastly, in HCC, soft voting and stacking were equally successful, both at 8.6%, while gene expression performed the worst at 22.6%. For a detailed overview of the computed misclassification rates and comprehensive data, please refer to the Supplementary Materials and Fig. S3. Additionally, various evaluation metrics are provided in Table S6.

To reach a broader consensus on which ensemble learning technique consistently outperforms the others, we aggregated the misclassification rates across all datasets and calculated their average. The same analysis was extended to the individual models for a comprehensive comparison. Fig. 6 presents our final findings, illustrating the overall performance of each model. As previously noted, all ensemble learning techniques demonstrated superiority over individual models. Soft voting, with an average rate of 16.08, emerged as the frontrunner, followed by majority voting at 16.35, and stacking at 16.53.

To further reinforce our findings, we conducted a similar analysis but limited it to datasets that possessed an independent external validation set. Once again, soft voting demonstrated superior performance with an average rate of 15.73, while majority voting and stacking followed with rates of 16.27 and 17.37, respectively. It's important to emphasise that even the top-performing results achieved by individual biofeatures fall short, or at the very least, are on par with the ensemble learning's average performance. In light of these discoveries, it becomes evident that ensemble techniques significantly enhance model performance by capitalising on the diverse nature of biofeature spaces and adeptly managing their inherent heterogeneity.

3.6. *Biological feature selected interpretation*

Genes detected as discriminative biofeatures should ideally have a causal relation to the analysed phenotype. For gene-expression based features putative relations can usually be shown by means of functional enrichment analyses, however for the other biofeature types explored in this study, this kind of analysis was not performed before. For example, genes showing RNA editing events with significant differences between control and cancer samples should act in relevant cancer pathways. To test this hypothesis, we analysed the 216 genes selected as features from the NSCLC dataset by means of the gProfiler. Fig. 7 shows the overrepresented functional annotations among these genes. Selected genes are mainly related to immune system functions, signal transduction and regulation of vesicle-mediated transport, which might indeed indicate that biologically meaningful features were extracted by means of our workflow.

The analysis of overrepresented functional annotations was specifically conducted for the NSCLC dataset alone, but we examined which genes are selected in more than one dataset (see Table S7 for complete list). Interestingly, many genes, mainly related to immune system functions, are selected from several studies and could therefore serve as pan-cancer markers. Table 2 offers a concise summary

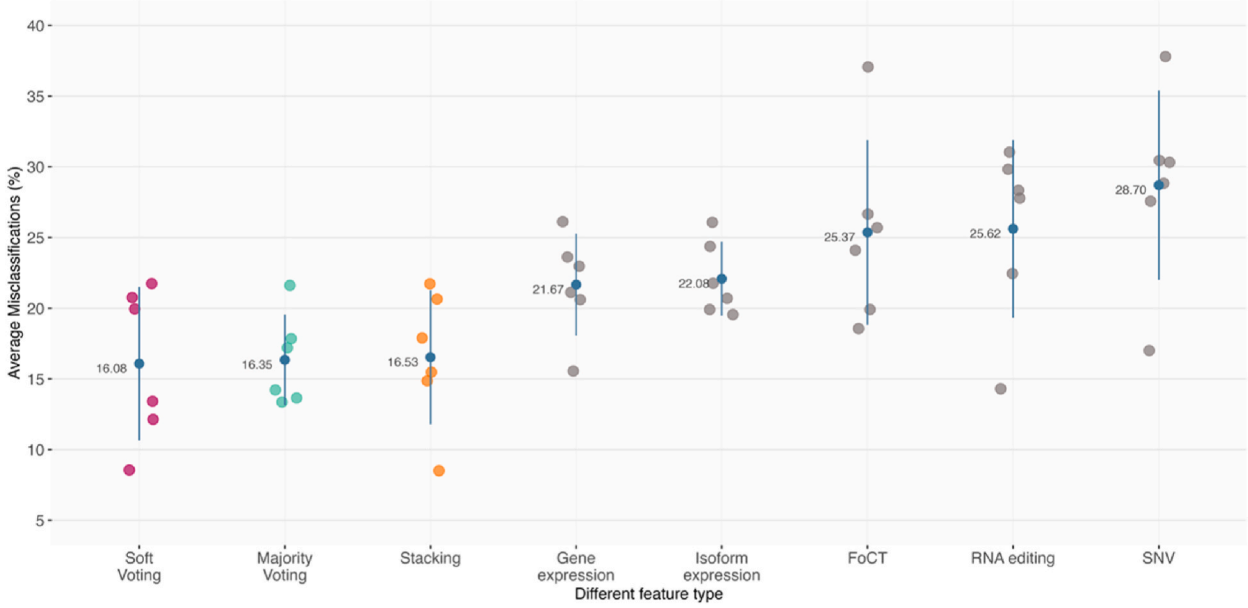
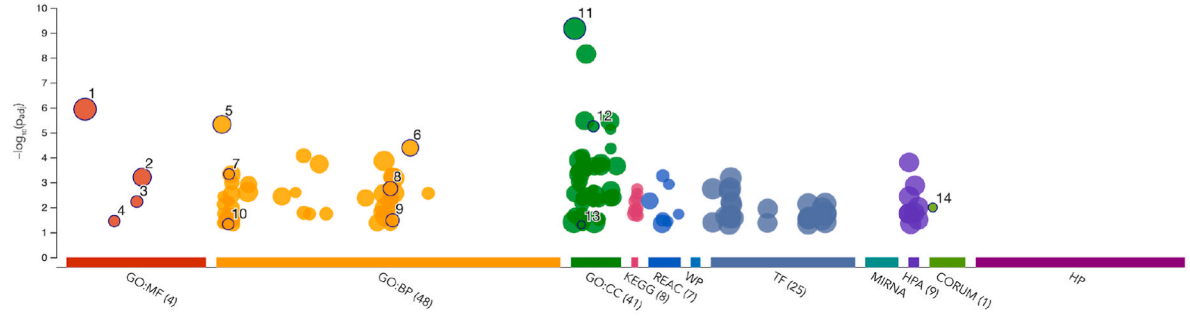


Fig. 6. Average Misclassification Percentage Across All Datasets. This graph displays the percentage of average misclassifications for each feature type on the x-axis, with feature types arranged from the least to the most misclassifications from left to right. Each dot represents the percentage of average misclassifications, providing an overview of the classification performance across all datasets.



ID	Source	Term ID	Term Name	p_{adj} (query_1)
1	GO:MF	GO:0005515	protein binding	1.158×10^{-6}
2	GO:MF	GO:0044877	protein-containing complex binding	6.221×10^{-4}
3	GO:MF	GO:0042605	peptide antigen binding	5.874×10^{-3}
4	GO:MF	GO:0023026	MHC class II protein complex binding	3.548×10^{-2}
5	GO:BP	GO:0001775	cell activation	4.681×10^{-6}
6	GO:BP	GO:0060627	regulation of vesicle-mediated transport	4.066×10^{-5}
7	GO:BP	GO:0002483	antigen processing and presentation of endogenous peptide antigen	4.673×10^{-4}
8	GO:BP	GO:0050852	T cell receptor signaling pathway	1.770×10^{-3}
9	GO:BP	GO:0051057	positive regulation of small GTPase mediated signal transduction	3.311×10^{-2}
10	GO:BP	GO:0002418	immune response to tumor cell	4.720×10^{-2}
11	GO:CC	GO:0005737	cytoplasm	6.752×10^{-10}
12	GO:CC	GO:0042611	MHC protein complex	5.564×10^{-6}
13	GO:CC	GO:0020018	ciliary pocket membrane	4.972×10^{-2}
14	CORUM	CORUM:486	WIP-WASp-actin-myosin-IIa complex	1.007×10^{-2}

Fig. 7. Presented is a Manhattan plot illustrating the results of an enrichment analysis performed on the genes selected for ensemble learning within the NSCLC dataset. The x-axis is dedicated to functional terms, which have been meticulously organised and color-coded based on their respective data sources. Simultaneously, the y-axis represents the adjusted enrichment p-values, thoughtfully presented in a logarithmic negative scale. Terms of lesser significance are discreetly depicted as faint circles, while encircled numbers within the figure denote statistically significant enriched GO terms.

of genes recurring in more than three, up to six out of six datasets.

4. Discussion

In this study, we introduced the ELLBA methodology, a robust and comprehensive multi-level bioinformatics approach tailored for analysing lbRNA-Seq data to predict patient outcomes. ELLBA's core framework centres on the extraction of six distinct biofeature types: gene expression, isoform expression, FoCT, gene fusion, RNA editing, and somatic SNVs. These diverse biofeatures aim to capture distinct molecular and functional characteristics. Our study design encompassed different cancer types and blood-based biosources, employing multi-condition samples from six datasets, followed by comprehensive evaluation across four independent validation sets.

A crucial aspect of our study was to explore the feasibility of intra-sample normalisation methods for the count-based biofeature types to improve the clinical applicability of the pipeline. Through the assessment of eight diverse normalisation methods, CPM normalisation emerged as a robust method for both gene and isoform expression analyses, performing comparably to the more

Table 2
Concise summary of genes recurring in more than three, up to six out of six datasets. The table includes the following attributes: GeneID, Gene Name, Gene Type, Occurrence (indicating in how many datasets the gene was selected as a feature), and a brief description of each gene's function.

GeneID	Gene Name	Gene Type	Occurrence	Description
ENSG00000234745.11	HLA-B	protein coding	6/6	major histocompatibility complex, class I, B
ENSG00000206503.13	HLA-A	protein coding	5/6	major histocompatibility complex, class I, A
ENSG00000213492.2	NT5C3AP1	transcribed processed pseudogene	3/6	NT5C3A pseudogene 1
ENSG00000166710.20	B2M	protein coding	3/6	beta-2-microglobulin
ENSG00000160014.17	CALM3	protein coding	3/6	calmodulin 3
ENSG00000162852.14	CNST	protein coding	3/6	consortin, connexin sorting protein
ENSG00000115956.10	PLEK	protein coding	3/6	pleckstrin
ENSG00000196126.11	HLA-DRB1	protein coding	3/6	major histocompatibility complex, class II, DR beta 1
ENSG00000117640.18	MTFR1L	protein coding	3/6	mitochondrial fission regulator 1 like
ENSG00000149781.12	FERMT3	protein coding	3/6	FERM domain containing kindlin 3
ENSG00000115310.18	RTN4	protein coding	3/6	reticulon 4
ENSG00000150867.14	PIP4K2A	protein coding	3/6	phosphatidylinositol-5-phosphate 4-kinase type 2 alpha

sophisticated cross-sample normalisation techniques like TMM or RUVSeq, which additionally corrects for batch effects. Notably, the clinical utility of intra-sample normalisation, such as CPM, deserves highlighting, as it enables rapid individual sample normalisation and seamless integration with instant predictions using pre-trained machine learning models. This property renders it highly suitable and time-efficient for clinical applications.

Further, we extensively benchmarked the choice of optimal classifiers for each biofeature type. Gene and isoform expression with continuous values demonstrated a strong compatibility with the AdaBoost classifier using the ExtraTrees as base estimator. Conversely, Logistic Regression exhibited superior performance for biofeature types with ratios (FoCT) or discrete values (remaining biofeatures). While most evaluated models performed comparably across each biofeature type, Naive Bayes and KNN consistently ranked lower. However, it is important to note that fluctuations in individual biofeature type performance might occur depending on the dataset, with no one biofeature type universally excelling. We believe that the variations in classifier performance across the different biofeature types in our benchmarking experiment can be attributed to several factors. Gene and isoform expression data might perform well with Adaboost due to their complexity and non-linear separability. On the other hand, Logistic Regression might excel with biofeatures like FoCT, gene fusion, RNA editing events, and SNVs because of their simpler, more linear relationships. Data size, feature importance, noise levels, and the presence of interaction effects could also contribute to these variations.

Our core finding underscores the superiority of combining information from several biofeatures over relying solely on standard Gene Expression. To mitigate fluctuations originating from individual biofeature space limitations and harness their complementary information, we introduced three distinct ensemble classification approaches within the ELLBA methodology. These ensemble methods, namely soft voting, majority voting, and stacking, integrate predictions derived from all six biofeature types, employing diverse strategies. Our findings showed, initially, that ensemble classification always improved predictive accuracy compared to gene expression alone. Furthermore, these ensemble learning techniques effectively reduce misclassification rates and enhance overall prediction accuracy, underscoring the value of multi-view analysis for liquid biopsy data. It is important to note that the choice of ensemble classification technique may vary depending on the dataset. In general, our findings suggest that soft voting is a robust and versatile ensemble learning method, a conclusion substantiated by datasets featuring an external validation set. This observation underscores the pivotal role of ensemble learning in enhancing the reproducibility and reliability of our results, further solidifying its importance in the context of liquid biopsy data analysis. Moreover, we believe that this superiority arises from several key advantages of ensemble methods. Firstly, they leverage the complementary information inherent in each biofeature type, capturing distinct aspects of the data and thus improving prediction accuracy. Secondly, ensemble methods reduce the impact of noise in individual models by combining multiple predictions, enhancing robustness. Additionally, ensemble techniques improve generalisation, reduce variance, and handle imbalanced data more effectively. Overall, ensemble learning emerged as a powerful strategy for harnessing the full potential of multi-feature data in cancer diagnostics.

Upon consideration of additional complementary metrics, including balanced accuracy, F1 score, and average Precision score, and comparing them with the corresponding metrics from the standard Gene Expression, as well as other biofeatures, we consistently observed unaltered results. It is crucial to emphasise that the analysis remains invariant due to the globally robust behaviour of the methodology. Our approach is dedicated to enhancing performance across diverse scenarios, irrespective of class distribution, with particular attention to cancer precision. This steadfast commitment to robustness underscores the methodology's reliability. To gain deeper insights into the biological significance of our selected features, we conducted a functional enrichment analysis, focusing on features that consistently recurred across datasets. The results highlighted that the genes identified as discriminative features, particularly those associated with immune system functions may hold significant relevance in cancer pathways. Furthermore, the recognition of genes recurring across multiple studies underscores their potential as valuable pan-cancer markers, underscoring the strength and applicability of our workflow in uncovering biologically meaningful features.

Several important considerations emerge from our study. Firstly, the restricted availability of gene fusion data may have affected the results, possibly due to the limited depth of sequencing in certain datasets or the use of single-end sequencing in cases such as NSCLC, GBM, and CRC datasets. While our fusion detection criteria were not overly stringent, the challenge of accurately capturing fusion events with high confidence warrants future exploration, particularly with the potential benefit of paired-end deep sequencing for enhancing the identification of these putatively important biomarkers. It is worth noting, that despite the depth of the CRC and the switching from PE to SE, we noticed that the Accuracy in the training is 0.60 by selecting 3 features out of the initial filtered ones which were 12. Secondly, the robust performance of our workflow on independent validation sets underscores the significance of maintaining uniform handling and sequencing protocols for validation data. This is exemplified by the closely aligned accuracy observed in the NSCLC dataset and its validation set, both managed by the same group, reinforcing the imperative for standardised procedures even across diverse laboratory settings. Finally, despite our comprehensive investigation of normalisation techniques, machine learning algorithms, and ensemble learning, there remains potential for enhancing our ensemble learning classifier's performance. This might involve exploring newer algorithms, adapting existing ones, exhaustive hyperparameter exploration, evolutionary-based feature selection, or integration of additional genomic data types. Through ongoing refinement of our machine learning pipeline, we anticipate continued progress in the precision and reliability of cancer predictions across diverse datasets.

5. Conclusions

The ELLBA workflow offers a significant contribution to liquid biopsy data bioinformatics analysis. Through multi-feature integration and ensemble learning, ELLBA presents a comprehensive avenue for patient outcome prediction in liquid biopsy-based cancer research. Its flexibility aligns well with upcoming liquid biopsy advancements, facilitating analysis across diverse cancer types and biosources. As liquid biopsy gains traction in cancer diagnostics, ELLBA holds promise for advancing precision oncology. Rigorous validation in even larger, diverse cohorts, complemented by experimental confirmation, will be pivotal to establishing ELLBA's clinical utility and reliability in liquid biopsy data analysis. With ongoing strides in liquid biopsy technologies and machine learning, ELLBA's continued evolution holds potential as an indispensable tool in liquid biopsy-based cancer research and clinical applications.

Data availability statement

All data to support the conclusions are publicly available and are located in the National Center for Biotechnology Information (NCBI) repository. They are listed under the following Bioproject Accessions: PRJNA353588, PRJNA659491, PRJNA737596, PRJNA810728, PRJNA552230, PRJNA483004, PRJNA761450, PRJNA281708, PRJNA391134, and PRJNA390988. Table S3 offers an extensive compilation of Bioproject and Data accession numbers utilised in this study.

Software availability

The source code and the extended manual of ELLBA are available in the GitHub repository: https://github.com/sgiannouk/ellba.

Funding

This project was entirely supported by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement ELBA No 765492.

CRediT authorship contribution statement

Stavros Giannoukakos: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis, Conceptualization. **Silvia D'Ambrosi:** Writing – original draft. **Danijela Koppers-Lalic:** Writing – original draft, Funding acquisition. **Cristina Gómez-Martín:** Writing – review & editing, Software. **Alberto Fernandez:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Conceptualization. **Michael Hackenberg:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT developed by OpenAI in order to improve the quality of the manuscript's language and grammar. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of competing interest

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

Acknowledgements

The authors acknowledge the usage of the computational infrastructure of the Computational Epigenomics Lab of the University of Granada. Moreover, we would like to sincerely thank Angel Martin Alganza, and Ernesto Aparicio-Puerta for maintaining such infrastructure.

We would also like to thank Sjors G.J.G. In 't Veld and Thomas Würdinger for giving early access to the PanCancer dataset.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e27360.

Abbreviations

LB	Liquid Biopsy
NGS	Next Generation Sequencing
TEP	Tumour-Educated Platelets
EV	Extracellular Vesicles
CECs	Circulating Epithelial Cells
CTCs	Circulating Tumour Cells
mRNA-Seq	mRNA Sequencing
AUC	Area Under the ROC Curve
FoCT	Fraction of Canonical Transcript
SNV	Single Nucleotide Variants
lbRNA-Seq	liquid biopsy-derived RNA-Seq
ML	Machine Learning
ELLBA	Ensemble Learning for Liquid Biopsy Analysis
CPM	Count Per Million
PCA	Principal Component Analysis
IQR	Interquartile Range
NA	Not Applicable
SNP	Single Nucleotide Polymorphisms
CV	Cross-Validation
KNN	K-Nearest Neighbours
GO	Gene Ontology
exLR-Seq	extracellular vesicles Long RNA Sequencing
scRNA-Seq	single-cell RNA sequencing
NSCLC	Non-small Cell Lung Carcinoma
GBM	Glioblastoma multiforme
CRC	Colorectal Cancer
ESCC	Esophageal Squamous-Cell Carcinoma
PDAC	Pancreatic Ductal Adenocarcinoma
HCC	Hepatocellular carcinoma
FQ	Full Quantile
UQ	Upper Quartile
RPKM	Reads Per Kilobase per Million mapped reads

References

[1] F. Ferrara, S. Zoupanou, E. Primiceri, Z. Ali, M.S. Chiriacò, Beyond liquid biopsy: toward non-invasive assays for distanced cancer diagnostics in pandemics, Biosens. Bioelectron. 196 (2022) 113698, https://doi.org/10.1016/j.bios.2021.113698.

[2] S. Bratulic, F. Gatto, J. Nielsen, The translational status of cancer liquid biopsies, Regenerative Engineering and Translational Medicine 7 (3) (2019) 312–352, https://doi.org/10.1007/s40883-019-00141-2.

[3] A. Krishnan, S. Thomas, Toward platelet transcriptomics in cancer diagnosis, prognosis and therapy, Br. J. Cancer 126 (3) (2021) 316–322, https://doi.org/10.1038/s41416-021-01627-z.

[4] B.M. Hussen, S.T. Abdullah, A. Salihi, D.K. Sabir, K.R. Sidiq, M.F. Rasul, et al., The emerging roles of NGS in clinical oncology and personalized medicine, Pathol. Res. Pract. 230 (2022) 153760, https://doi.org/10.1016/j.prp.2022.153760.

[5] E. Kilgour, D.G. Rothwell, G. Brady, C. Dive, Liquid biopsy-based biomarkers of treatment response and resistance, Cancer Cell 37 (4) (2020) 485–495, https://doi.org/10.1016/j.ccell.2020.03.012.

[6] E. Sánchez-Herrero, R. Serna-Blasco, V. Ivanchuk, R. García-Campelo, M. Dómine Gómez, J.M. Sánchez, et al., NGS-based liquid biopsy profiling identifies mechanisms of resistance to ALK inhibitors: a step toward personalized NSCLC treatment, Mol. Oncol. 15 (9) (2021) 2363–2376, https://doi.org/10.1002/1878-0261.13033.

[7] D. Shyr, Q. Liu, Next generation sequencing in cancer research and clinical application, Biol. Proced. Online 15 (1) (2013), https://doi.org/10.1186/1480-9222-15-4.

[8] M.C. Liefaard, K.S. Moore, L. Mulder, D. van den Broek, J. Wesseling, G.S. Sonke, et al., Tumour-educated platelets for breast cancer detection: biological and technical insights, Br. J. Cancer 128 (8) (2023) 1572–1581, https://doi.org/10.1038/s41416-023-02174-5.

[9] S. Yu, Y. Li, Z. Liao, Z. Wang, Z. Wang, Y. Li, et al., Plasma extracellular vesicle long RNA profiling identifies a diagnostic signature for the detection of pancreatic ductal adenocarcinoma, Gut 69 (3) (2019) 540–550, https://doi.org/10.1136/gutjnl-2019-318860.

[10] I. Bhan, M. Kelly, L. Goyal, J. Philipp, M. Kalinich, J.W. Franses, et al., Detection and analysis of circulating epithelial cells in liquid biopsies from patients with liver disease, Gastroenterology 155 (6) (2018) 2016–2018.e11, https://doi.org/10.1053/j.gastro.2018.09.020.

[11] M. Antunes-Ferreira, S. D'Ambrosi, M. Arkani, E. Post, S.G.J.G. t Veld, J. Ramaker, et al., Tumor-educated platelet blood tests for Non-Small Cell Lung Cancer detection and management, Sci. Rep. 13 (1) (2023) 9359, https://doi.org/10.1038/s41598-023-35818-w.

[12] K. Swanson, E. Wu, A. Zhang, A.A. Alizadeh, J. Zou, From patterns to patients: advances in clinical machine learning for cancer diagnosis, prognosis, and treatment, Cell 186 (8) (2023) 000946, https://doi.org/10.1016/j.cell.2023.01.035. S0092-8674(23.

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830, https://doi.org/10.5555/1953048.2078195.

[14] K. Katz, O. Shutov, R. Lapoint, M. Kimelman, J. Brister, C. O'Sullivan, The Sequence Read Archive: a decade more of explosive growth, Nucleic Acids Res. 50 (D1) (2021) D387–D390, https://doi.org/10.1093/nar/gkab1053.
[15] B. Bushnell, J. Rood, E. Singer, BBMerge – accurate paired shotgun read merging via overlap, PLoS One 12 (10) (2017) e0185056, https://doi.org/10.1371/journal.pone.0185056.
[16] P. Ewels, M. Magnusson, S. Lundin, M. Käller, MultiQC: summarize analysis results for multiple tools and samples in a single report, Bioinformatics 32 (19) (2016) 3047–3048, https://doi.org/10.1093/bioinformatics/btw354.
[17] A. Dobin, C.A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, et al., STAR: ultrafast universal RNA-seq aligner, Bioinformatics 29 (1) (2012) 15–21, https://doi.org/10.1093/bioinformatics/bts635.
[18] A. Frankish, M. Diekhans, I. Jungreis, J. Lagarde, J. Loveland, J.M. Mudge, et al., Gencode 2021, Nucleic Acids Res. 49 (D1) (2020) D916–D923, https://doi.org/10.1093/nar/gkaa1087.
[19] Picard Tools - By Broad Institute. broadinstitute.github.io. <http://broadinstitute.github.io/picard> Accessed 23.08.12.
[20] L. Wang, S. Wang, W. Li, RSeQC: quality control of RNA-seq experiments, Bioinformatics 28 (16) (2012) 2184–2185, https://doi.org/10.1093/bioinformatics/bts356.
[21] M.D. Robinson, D.J. McCarthy, G.K. edgeR. Smyth, A Bioconductor package for differential expression analysis of digital gene expression data, Bioinformatics 26 (1) (2009) 139–140, https://doi.org/10.1093/bioinformatics/btp616.
[22] R. Patro, G. Duggal, M.I. Love, R.A. Irizarry, C. Kingsford, Salmon provides fast and bias-aware quantification of transcript expression, Nat. Methods 14 (4) (2017) 417–419, https://doi.org/10.1038/nmeth.4197.
[23] S. Uhrig, J. Ellermann, T. Walther, P. Burkhardt, M. Fröhlich, B. Hutter, et al., Accurate and efficient detection of gene fusions from RNA sequencing data, Genome Res. 31 (3) (2021) 448–460, https://doi.org/10.1101/gr.257246.119.
[24] Uhrig Sebastian .suhrig/arriba. GitHub. <https://github.com/suhrig/arriba>Accessed23.08.12...
[25] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, et al., The sequence alignment/map format and SAMtools, Bioinformatics 25 (16) (2009) 2078–2079, https://doi.org/10.1093/bioinformatics/btp352.
[26] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, et al., The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, Genome Res. 20 (9) (2010) 1297–1303, https://doi.org/10.1101/gr.107524.110.
[27] P. Danecek, J.K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M.O. Pollard, et al., Twelve years of SAMtools and BCFtools, GigaScience 10 (2) (2021), https://doi.org/10.1093/gigascience/giab008.
[28] S.T. Sherry, dbSNP: the NCBI database of genetic variation, Nucleic Acids Res. 29 (1) (2001) 308–311, https://doi.org/10.1093/nar/29.1.308.
[29] E. Picardi, G. REDItools Pesole, high-throughput RNA editing detection made easy, Bioinformatics 29 (14) (2013) 1813–1814, https://doi.org/10.1093/bioinformatics/btt287.
[30] G.A. Auwera, M.O. Carneiro, C. Hartl, R. Poplin, G. del Angel, A. Levy-Moonshine, et al., From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline, Current Protocols in Bioinformatics 43 (1) (2013), https://doi.org/10.1002/0471250953.bi1110s43.
[31] Calzolari, M. sklearn-genetic. GitHub. <https://github.com/manuel-calzolari/sklearn-genetic>Accessed23.08.12..
[32] U. Raudvere, L. Kolberg, I. Kuzmin, T. Arak, P. Adler, H. Peterson, et al., Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update), Nucleic Acids Res. 47 (W1) (2019) W191–W198, https://doi.org/10.1093/nar/gkz369.
[33] M. Best, N. Sol, S. t Veld, A. Vancura, M. Muller, A. Niemeijer, et al., Swarm intelligence-enhanced detection of non-small-cell lung cancer using tumor-educated platelets, Cancer Cell 32 (2) (2017), https://doi.org/10.1016/j.ccell.2017.07.004.
[34] Myron G. Best, N. Sol, I. Kooi, J. Tannous, Bart A. Westerman, F. Rustenburg, et al., RNA-seq of tumor-educated platelets enables blood-based pan-cancer, multiclass, and molecular pathway cancer diagnostics, Cancer Cell 28 (5) (2015) 666–676, https://doi.org/10.1016/j.ccell.2015.09.018.
[35] S.G.J.G. t Veld, M. Arkani, E. Post, M. Antunes-Ferreira, S. D'Ambrosi, D.C.L. Vessies, et al., Detection and localization of early- and late-stage cancers using platelet RNA, Cancer Cell 40 (9) (2022) 999–1009.e6, https://doi.org/10.1016/j.ccell.2022.08.006.
[36] S. Li, Y. Li, B. Chen, J. Zhao, S. Yu, Y. Tang, et al., exoRBase: a database of circRNA, lncRNA and mRNA in human blood exosomes, Nucleic Acids Res. 46 (D1) (2017) D106–D112, https://doi.org/10.1093/nar/gkx891.
[37] Y. Li, Q. Zheng, C. Bao, S. Li, W. Guo, J. Zhao, et al., Circular RNA is enriched and stable in exosomes: a promising biomarker for cancer diagnosis, Cell Res. 25 (8) (2015) 981–984, https://doi.org/10.1038/cr.2015.82.
[38] T. Liu, X. Wang, W. Guo, F. Shao, Z. Li, Y. Zhou, et al., RNA sequencing of tumor-educated platelets reveals a three-gene diagnostic signature in esophageal squamous cell carcinoma, Front. Oncol. 12 (2022), https://doi.org/10.3389/fonc.2022.824354.
[39] N. Sol, S.G.J.G. t Veld, A. Vancura, M. Tjerkstra, C. Leurs, F. Rustenburg, et al., Tumor-Educated platelet RNA for the detection and (Pseudo)progression monitoring of glioblastoma. Cell reports, Medicine 1 (7) (2020) 100101, https://doi.org/10.1016/j.xcrm.2020.100101.
[40] L. Xu, X. Li, X. Li, X. Wang, Q. Ma, D. She, et al., RNA profiling of blood platelets noninvasively differentiates colorectal cancer from healthy donors and noncancerous intestinal diseases: a retrospective cohort study, Genome Med. 14 (1) (2022), https://doi.org/10.1186/s13073-022-01033-x.
[41] C. Scheepbouwer, M. Hackenberg, M. A.J, A. Gerber, D. Michiel Pegtel, C. Gómez-Martín, NORMSEQ: a tool for evaluation, selection and visualization of RNA-Seq normalization Methods, Nucleic Acids Res. 51 (W1) (2023) W372–W378, https://doi.org/10.1093/nar/gkad429.
[42] C. Evans, J. Hardin, D.M. Stoebel, Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions, Briefings Bioinf. 19 (5) (2017) 776–792, https://doi.org/10.1093/bib/bbx008.