Computers in Biology and Medicine 110 (2019) 234–243

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

CBS-miRSeq: A comprehensive tool for accurate and extensive analyses of microRNA-sequencing data

Rupesh K. Kesharwani[a,b,*,1], Mattia Chiesa[a,b], Riccardo Bellazzi[b], Gualtiero I. Colombo[a,**]

[a] Unit of Immunology and Functional Genomics, Centro Cardiologico Monzino IRCCS, Via Carlo Parea, 4, 20138, Milano, Italy
[b] Department of Electrical, Computer and Biomedical Engineering, Università degli Studi di Pavia, Via Ferrata 3 27100, Pavia, Italy

ARTICLE INFO

Keywords:
microRNA
Gene expression profiling
Color-space
Base-space
Bioinformatics pipeline

ABSTRACT

Several online and local tools have been developed to analyze microRNA-sequencing (miRNA-Seq) data, but usually they are limited by many factors including: inaccurate processing, lack of optimal parameterization, outdated references plus annotations, restrictions in uploading large datasets, and shortage of biological inferences.

In this work, we have developed a fully customized bioinformatics analysis pipeline (Color and Base-Space miRNA-Seq – CBS-miRSeq) for the seamless processing of short-reads miRNA-Seq data. The pipeline has been designed using Bash, Perl, and R scripts. CBS-miRSeq includes modules for read pre- and post-processing (quality assessment, filtering, adapter trimming and mapping) and different types of downstream analyses (identification of miRNA variants (isomiRs), novel miRNA prediction, miRNA:mRNA interaction target prediction, robust differential miRNA analysis, and target gene functional analysis). In this manuscript, we show that re-analysis of two published datasets using the CBS-miRSeq pipeline leads to better performance and efficiency in terms of their pipelines set and biomarker discovery between two biological conditions.

1. Introduction

Small RNAs are part of the wide family of non-coding RNAs, which possess many regulatory functions. In particular, microRNAs (miRNAs) have gained major interest because of their activity in gene silencing through posttranscriptional repression [1]. Mature miRNAs are about 21–23 nucleotides (nt) in length [2], regulate many biological processes including cell differentiation [3] and organ development [4], and are involved in the pathophysiology of a variety of diseases such as cancer [5] and cardiovascular diseases [6]. Next generation microRNA-sequencing (miRNA-Seq) allows for the characterization of miRNA expression profiles with higher accuracy and resolution than hybridization-based microarray technology. A number of online and local tools have been developed to analyze miRNA-Seq data, but have many limitations including: inaccurate processing, lacking of optimal parameterization, outdated reference genomes and annotations, restrictions in uploading large datasets, input format issues, and shortage of biological inferences. Web-based tools (mirTools [7], miRanalyzer [8], DARIO [9], wapRNA [10], ncPRO-seq [11], and omiRas [12]) have limited capacity in processing large datasets, support base-space formats only (except wapRNA), and depend upon outdated references and annotations. On the other hand, local tools (mirToolsv2.0 [7], miRDeep2 [13], ncPRO-seq [11], and CAP-miRSeq [14]) conduct analysis with multiple samples, but support only base-space formats and/or require pre-processed data as input. These tools have limited options for parameter adjustment and/or lack of isomiR classification capability, prediction of target genes, or functional analysis. Furthermore, also the very recent tool miARma-Seq [15] presents a few significant limitations, such as bad-quality reads filtering, statistical tests for expression comparison and prediction of novel candidates. However, miARma-Seq further requires multiple optimizations to input of color-space reads, finding miRNA editing events, and accurate novel miRNA expression and target prediction as well as option to choose or merging differential expression analysis from more than one methodology. Overall, existing tools do not perform a fully comprehensive analysis of microRNA expression changes and their potential biological meaning. Moreover, researchers need to set up their own local workflow and use multiple tools, each of them provides partial functions of the entire data analysis

* Corresponding author. Unit of Immunology and Functional Genomics, Centro Cardiologico Monzino IRCCS, Via Carlo Parea, 4, 20138, Milano, Italy.
** Corresponding author.
E-mail addresses: bioinforupesh2009.au@gmail.com, rupeshkumar.kesharwani01@universitadipavia.it (R.K. Kesharwani), gualtiero.colombo@ccfm.it (G.I. Colombo).
[1] Current address: Department of Computer Science, The Jackson laboratory for Genomic Medicine, 10 Discovery Dr, Farmington, CT, 06032, USA.

https://doi.org/10.1016/j.compbiomed.2019.05.019
Received 27 January 2019; Received in revised form 24 May 2019; Accepted 25 May 2019
0010-4825/ Published by Elsevier Ltd.

steps. Installing those tools from various sources, testing their modules and integrating them together are also tedious and time-consuming.

To overcome these limitations and problems, we present here a fully customized pipeline i.e. Color and Base-Space miRNA-Seq (CBS-miRSeq), which can be installed locally and is able to run analysis on a large number of samples from both Illumina base-space (fastq) and SOLiD short reads color-space (csfasta) formats. The CBS-miRSeq tool enables users to perform: (*i*) comprehensive miRNA profiling, using any version of species-specific reference (animal only) genomes along with their annotation; (*ii*) accurate differential expression analysis using two widely accepted statistical methods (DESeq2 [16] and edgeR [17]); (*iii*) robust prediction of novel miRNA candidates along with counts matrix; (*iv*) prediction of target genes using two different algorithms (miRanda [18] and RNAhybrid [19]); (*v*) gene set enrichment analysis based on gene ontology (GO), disease ontology (DO), and Reactome and/or KEGG pathway categories.

The pipeline is developed in a Unix/Linux environment and consists of three main modules which enable running multiple samples and controlling results produced at every analysis step. Moreover, the modular architecture of the pipeline allows starting analyses at any point in the workflow. The CBS-miRSeq performance was tested on two different publicly available datasets: (*1*) Ataxia Telangiectasia Mutated kinase (ATM)-deficient human mammary epithelial cells (HME-CCs), in base-space [21], and (*2*) human neuroblastoma tumor samples, in color-space [22,23]. Comparison with previously published studies show that our pipeline performs better in identifying differentially expressed (DE) miRNAs between experimental conditions and predicting miRNA targets along with functional enrichment analysis.

1.1. Implementation

The CBS-miRSeq pipeline is implemented with combination of *Bash*, *Perl*, and *R* scripts in a Unix/Linux environment. The pipeline can be run sequentially on a single machine or in parallel on a high-performance computing cluster/server. This allows for the analysis of a small number of samples up to hundreds of samples in a single run with in a reasonable timeframe, depending on the computing infrastructure available. The standard workflow of the pipeline comprises reads preprocessing (quality assessment, filtering, and adapter trimming), alignment of short reads, prediction of novel miRNA candidates, differential analysis using two different statistical packages, prediction of miRNA targets along with miRNA:mRNA interaction, and prediction of biological processes and pathways affected by DE miRNAs.

The tool can be installed locally with an installation script and detailed instructions. We provide a virtual machine image (VM-image) for users who are not comfortable with the installation to use the tool directly for a small dataset. CBS-miRSeq is also available as a Docker file to build Docker image. This Docker container can easily be built and deployed in cloud or local environments for small to large data analyses.

1.2. Workflow of CBS-miRSeq

CBS-miRSeq consists of three main modules (Fig. 1), each producing a separate result sheet. This flexibility allows user to review results generated at every single analytical step.

Module 1. Module 1 integrates two sub-modules.

Module 1a. This is a prerequisite step for Module 1b. Module 1a downloads the reference genome if required, and builds the bowtie index corresponding to input reads for either color-space or base-space. After index building, the pipeline automatically launches Module 1b. Optionally, the user can deploy Module 1b separately, if reference genome and index are already available.

Module 1b. This module performs preprocessing and mapping of sequence reads.

1.2.1. Pre-alignment processing and quality control

Quality control of raw reads is a critical and complex checkpoint before downstream analysis that helps eliminate sequencing errors. Sequencing artifacts include base calling errors (insertion/deletion), base missing, poor quality reads, and contaminations. These artifacts are quite common in datasets generated by deep sequencing and significantly affect mapping and subsequent gene expression analysis [24]. The quality of raw sequence data is first checked using the Solid2fastq tool from BFAST v0.6.5a [25] and FastQC v0.10.1 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc), which provides a detailed report of the quality of the reads. Next, based on the input type (color-space or base-space) reads are filtered either using 'SOLiD_preprocess_filter_v2. pl' [24] or fastq_quality_filter from the FastX–Toolkit v0.0.14 (http://hannonlab.cshl.edu/fastx_toolkit/index.html). Color-space reads with a quality score ≤ 9 and base-space reads with a quality score < 20 are discarded and excluded from further analysis.

1.2.2. Adapter removal

The Cutadapt tool v1.6 [26] is used to remove 3' adapter sequences from the short reads prior to mapping. Resulting reads with less than 15 nt after adapter removal are excluded from further analysis. Next, these trimmed reads are submitted to additional quality control analysis with FastQC to assess the length distribution of the reads (ideally between 18 and 23 nt for genuine mature miRNAs) of the processed data.

1.2.3. Mapping

The 3′ adapter trimmed reads are aligned with the widely used Bowtie aligner v1.0.0 [27] to the whole genome of the reference species provided by the user. In short, Bowtie enables ultrafast and memory-efficient alignment for sequencing reads to a reference sequence. We force the aligner to use the option 'no mismatches in the first 15 base pair seed region' of the short reads and allow up to the first best alignment hit per sequence read only. After mapping, the pipeline generates standard alignment files in the Sequence Alignment/Map (SAM) and Binary Alignment/Map (BAM) formats, which are used for subsequent analyses. Mapping statistics are plotted in hypertext files using the SAMStat tool [28].

1.2.4. Mapping color-space reads

Unlike other sequencing technologies, reads generated by the SOLiD chemistry are longer than the ideal miRNA: each read is 35 nt long where the first base of the read is the last nucleotide of the 5′ primer (usually T or C) and the miRNA and the 3′ adapter sequences are provided in color-space [29]. After alignment, Bowtie trims off the first and the last base from the read due to low confident about in decoding. Hence, to avoid the loss of the 5′ nucleotide, which may alter the recognition of the miRNA functional seed region (nucleotides 2–7 of the miRNA), we force the Bowtie aligner to keep the informative first base by enabling the '–col-keep ends' option.

1.2.5. Quantification

The featureCounts v1.4.6 program [30] is implemented in CBS-miRSeq to summarize aligned reads and quantify the expression of miRNAs annotation by miRBase [31]. The Ensembl genome coordinates are also used to (http://www.ensembl.org/info/data/ftp/index.html) computes the expression of RNA biotypes which may possibly present in the miRNA libraries. Briefly, featureCounts is a highly efficient general-use read summary program that counts mapped reads for genomic features like genes, exons, miRNAs, genomic bins, and chromosomal locations. The module generates two expression tables which further used for miRNA differential expression comparison and percentage distribution of RNA biotypes. Plots will be generated to assessing the DE results and quality of miRNA library preparation.

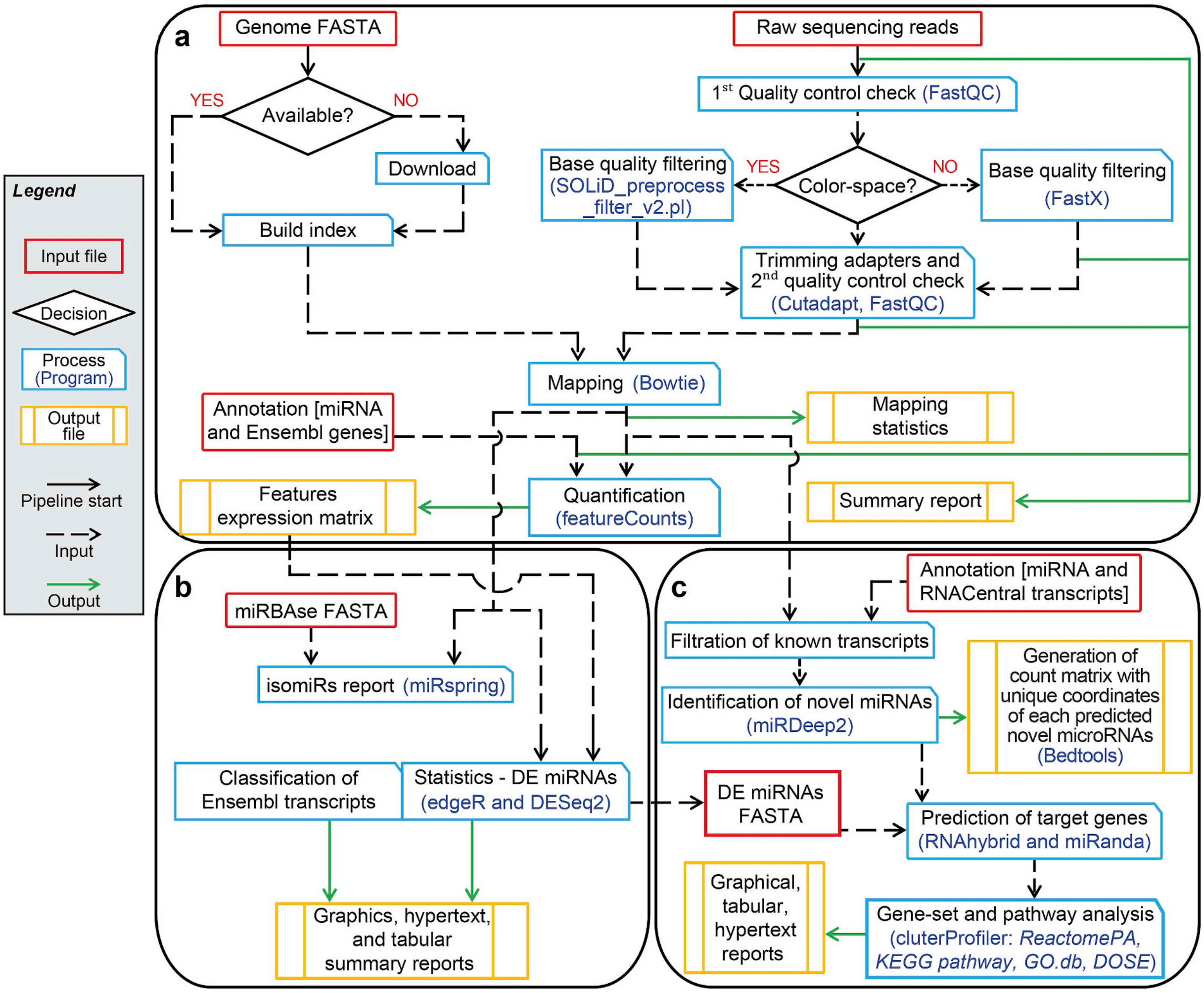

Fig. 1. Workflow diagram of the CBS-miRSeq pipeline. Each box presents one module of the pipeline, with the name of the tools integrated and required in a module highlighted in blue. (**a**) Module 1 performs preprocessing, quality control qualification, mapping, and quantification of raw reads. (**b**) Module 2 carries out differential expression analysis and isomiR detection. (**c**) Module 3 accomplishes identification of novel miRNA candidates, target gene prediction, and functional enrichment analysis of predicted targets.

1.2.6. Summary statistics

The CBS-miRSeq Module 1b produces high quality reports in Excel and hypertext files that comprise total read counts of each sample, quality control statistics, read length distribution, summary of adapter removal, mapping statistics, and raw expression count matrix of known miRNAs.

Module 2. Module 2 has two sub-modules, which can be used alone or together (recommended).

Module 2a. This module performs differential expression analysis between two experimental conditions by implementing two widely used R-Bioconductor statistical packages (DESeq2 [16] and edgeR [17]) for sequence count data. Both methods use the negative binomial distribution to model discreet counts. In order to identify DE miRNAs with the highest degree of confidence, by default only miRNAs with ≥ 1 count per million (CPM) in at least 50% of the samples are retained. Optionally, user can assign a different filtering cutoff based either on CPM or raw counts. Next, filtered counts are normalized by DESeq2 internal procedure [32] or by Trimmed Mean of M-values (TMM) in edgeR [33]. Briefly, TMM computes the weighted mean of log ratios between test and reference sample, after exclusion of the most expressed genes and the genes with the largest log ratios whereas, DESeq normalization divides the counts by sample-specific size factors determined by median ratio of gene counts relative to geometric mean per gene. Differential expression between groups is determined using the General Linear Model Likelihood Ratio Test (glmLRT) by default. False discovery rate (FDR) and log 2 fold change (log_2FC) are computed for each miRNA to select those that are deemed significantly DEP rior to differential expression analysis, the module may also control for inconsistent results in the dataset (for instance, miRNAs with hyper-variant expression values in any experimental condition), by enabling the option 'yes' for the parameter *RemoveInconsistentFeatures*. The user is asked to assign a threshold value to the parameter *ThresholdToRemoveInconsistentFeatures*, which is the coefficient of variation (CV) deemed acceptable (default is 1.5): miRNAs with a CV greater than threshold value within each class/group will be filtered out. In practice, we observed that the removal of the hyper-variant genes is useful to avoid incorrect detection of these miRNAs as DE. In addition, Module 2a allows an option to perform a data-driven filtering procedure based on the Jaccard similarity index among biological replicates [34], which helps in choosing an appropriate threshold to remove uninformative genes, by enabling the parameter *Independent Filtering* before differential expression analysis Module 2a finally produces complete data reports in tab-delimited and hypertext formats for all miRNAs including the following: normalized expression counts, log_2FC, significance level, and FDR, along with diagrams summarizing miRNAs deemed DE by the two statistical methods (separately and at the intersection), quality statistics box plots, principal component analysis (PCA) plots, MA plots, volcano plots, hyperlinks to miRBase. Module 2a produces results obtained by DESEq2 and edgeR as distinct outputs, and/or visualizes overlapping results with a Venn diagram, reporting them in a separate list. Users can choose which list of DE miRNAs use for subsequent downstream analysis.

Module 2b. This module implements miRspring v1.2 [20] to detects isomiRs (miRNA editing events) from the mapped reads. Briefly, the analysis includes discovery of miRNAs that are not annotated on in miRBase database, i.e. evaluating the complete abundance of 5′ and 3′ isomiRs, relative expression levels within genomic clustered miRNA and the list of seed-isomiRs. Predicting seed-isomiRs would have significant overlap to annotated miRNAs in miRBase database, and it may provide an alternate mechanism of biological systems to enhance the regulation of specific targets through their expression changes.

Module 3. Module 3 consists of two sub-modules, which can be used separately or sequentially.

Module 3a. This module makes use of miRDeep2 v2.0.0.5 [13] for identifying potential novel miRNAs. In brief, miRDeep2 predicts novel miRNA by assessing the miRNA compatibility of the sequence where reads are stacked, i.e. whether it forms precursor-miRNA hairpin structure and read distribution of 5′ and 3′ mature miRNAs, and is not present in miRNA repository i.e. miRBase. Next, a confidence score is assigned to each predicted miRNA. To avoid false discoveries, we implemented a number of filtration steps. First, we consider only reads longer than 18 nt. Second, the module removes miRNAs that are annotated in miRBase and noncoding RNAs (such as tRNAs, rRNAs, snoRNAs, etc.) from the annotation of RNAcentral database [35]. Third, we force miRDeep2 to analyze all predicted precursors for novel miRNAs candidates. Fourth, we select candidates with a log-odds score assigned by miRDeep2 ≥ 1. Given that genomic coordinates of predicted miRNAs can vary from sample to sample with a difference of few positions, we define new coordinates (start and end positions) merging chromosome ranges of commonly detected novel miRNAs from multiple samples. If more than 90% miRNA sequences overlap among samples, a new genomic range is generated exploiting the outermost coordinate using bedtools v2.23.0 (http://bedtools.readthedocs.org/en/latest/index.html). Finally, we compute the expression counts for each new coordinate by summing read counts within each sample in this range. At the end, two files with the count matrix and genomic coordinates of the putative novel miRNAs are created for further examination (*i.e.* wet-lab validation and downstream analysis).

Module 3b. This module performs target gene prediction, along with functional enrichment analysis. The module requires a fasta file of mature miRNAs (either DE, novel, or sequences of interest) and a fasta file of 3′ UTR sequences downloaded from the Ensembl BioMart repository (www.ensembl.org/biomart/). Module 3b integrates two algorithms for sequence-based target gene prediction, RNAhybrid v2.1.1 [19] and miRanda v3.3a [18], with the R/Bioconductor [36] package cluterProfiler [37]), which is used for functional annotation analysis based on GO, DO, KEGG and/or Reactome pathways. The module automatically downloads the required packages (if needed) for investigating functional processes associated with the target genes of the miRNAs of interest. The implementation of two miRNA target-gene prediction algorithms helps producing consistent results. Briefly, RNAhybrid first predicts secondary structure and then calculates thermodynamics, along with significance p-values to find the minimum free energy of a miRNA:mRNA interaction; miRanda calculates sequence to sequence complimentary binding energy scores by position-weighted local alignment. Module 3b requires the user to indicate a hybridization threshold for both algorithms. Finally, unique miRNA:mRNA gene pairs predicted by both algorithms are considered for subsequent functional analysis. A detailed reports are generated by CBS-miRSeq that includes tab-delimited and hypertext reports of the miRNA:mRNA hybridization.

1.3. Test datasets

1 Base-space dataset. The first test set consists of 3 ATM-deficient and 3 normal HME-CCs of a total of 208 million base-space reads, generated with an Illumina Genome Analyzer II as described in Ref. [21] in a study of biomarker discovery for breast cancer susceptibility related to ATM-deficiency. The dataset was retrieved from NCBI Sequence Read Archive (SRA) in the native fastq format using the accession no. SRP011278. This dataset is used to demonstrate the following: (a) Speed and performance of the proposed workflow i.e. CBS-miRSeq: (2) Pre and post processing of the analysis; (3) Differential miRNA expression of ATM-deficient human mammary epithelial cells vs. normal cells and to report potential miRNA biomarkers; (4) Functional and pathway analysis of miRNAs that reported by the workflow.

2 Color-space dataset. The second dataset comprises a total of approximately 188 million color-space reads from five favorable (event-free survivors) *vs.* five unfavorable (died of disease) human neuroblastoma tumor samples generated using the Applied Biosystems SOLiD system v3 as described in Ref. [22]. This dataset was retrieved in the native format from SRA repository using the accession no. SRA009986. In this study, several up (miR-181a-2-3p) and down (miR-628–5p, miR-744–5p, miR-1249, and miR-3612) miRNAs are expressed in unfavorable human neuroblastoma tumor samples relative to favorable. These were further confirmed and validated through RT-PCR [23]. We used this data set to test our workflow if the same results are replicated along with other potential biomarkers being detected against unfavorable tumor proliferation and association with neuroblastoma.

2. Results and discussion

2.1. CBS-miRSeq computational performance

The pipeline is mainly developed for a cluster environment to speed up the process. Additionally, flexibility of CBS-miRSeq allows user to run and review results generated at every single analytical step. The user can initiate each step separately or parallel using core modules i.e. CBS-miRSeq.module1.sh, CBS-miRSeq.module2.sh and CBS-miRSeq.module3.sh. We used 1 processor from our cluster together with a maximum of 10 GB of RAM to see the velocity of the steps of each module, with special consideration for the potential steps that might take longer processing time. After running all 6 libraries of dataset 1 (SRP011278), each workflow step took an average of 4 min 42 s to complete with a maximum of 2.1 GB of memory. Prediction of novel miRNAs took on average of 20 min to finish for each library. We found maximum time taken step was building genome indices from module1a. It took 1 h:21 m:37s with max uses of 5.4 GB of memories to complete the job. Therefore, we advise analyst to use pre-built genome indices in order to save time and allow for faster analysis. S1 Fig. summarizes the speed of CBS-miRSeq's potential time-consuming step.

2.2. CBS-miRSeq performance for ATM-deficient human mammary epithelial cells (base-space dataset)

We examined whether our pipeline could identify the same DE miRNAs reported in the original work [21] and/or reveal additional DE miRNAs, predict their targets, and predict novel miRNAs. To evaluate CBS-miRSeq performance, we started from raw base-space data and proceeded as follows: using Module-1 default settings: (1) pre-alignment processing, quality check and low-quality data filtering; (2) 3′ adapter clipping; (3) mapping to the reference genome (Fig. 2a–c), and (4) quantification of annotated miRNAs and classification of other annotated transcripts (Fig. 2d–f). After low-quality read filtration and 3′ adapter trimming, 63.1 $\pm$ 9.1% raw reads from each library were retained. Quality-filtered reads were aligned to the same reference hg19 human genome build (Ensembl GRCh37 v.67) used in the original work [21]. We found correct alignment for 87–94% filtered reads per library (Table S1); the unaligned reads were excluded from subsequent

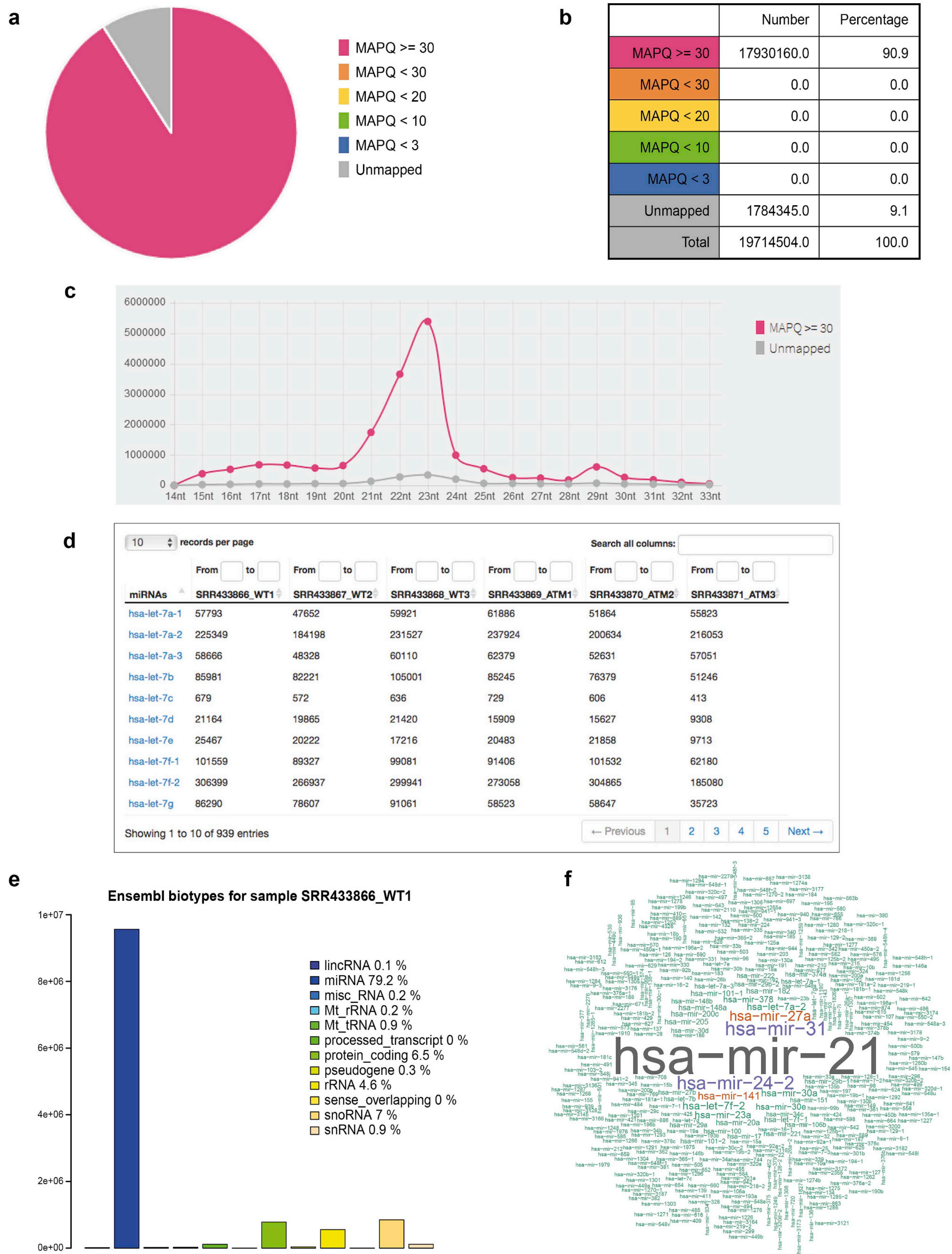

Fig. 2. Example outputs from Module 1 of CBS-miRSeq from the dataset of ATM-deficient human mammary epithelial cells. Panels (**a**) and (**b**) show mapping and quality statistics of miRNA reads after alignment with the genome. Panel (**c**) shows length distribution of the miRNA reads. Most aligned reads presented the length of a mature miRNA (21–23 nt) and had a mapping quality MAPQ ≥ 30. Panel (**d**) shows the tabular report of mature, annotated miRNA raw counts in each library. Each miRNA has a hyperlink to miRBase. Panel (**e**) displays the distribution of all detected RNA transcripts by category in a given sample. Panel (**f**) is a word cloud depicting miRNA relative expression in ATM samples.

analysis. For a fair comparison, we quantified miRNA counts using the same annotation (UCSC small genome track WgRNA on hg19) of the original work [21]: thus, the number of annotated miRNAs found (939) was the same.

2.2.1. Differential comparison

We used Module 2a to perform differential expression analysis between groups (ATM-deficient cells *vs.* wild type controls). Using the default options (filtering out miRNAs with < 1 CPM in ≥50% of the samples), we deemed expressed 371 unique miRNAs. Making comparison with the same significance cut-offs (nominal p-value < 0.05; FC ≥ |1.5|) as in the original report [21], we found 124 DE miRNAs by both statistical packages (edgeR and DESeq2). Module 2a successfully identified 60 of the 81 miRNAs reported as DE in Ref. [21] and, additionally, 64 miRNAs were not reported in the original study (Table S2a). CBS-miRSeq did not report other 21 miRNAs as indicated as DE in that study. Our pipeline did not deem expressed 8 of these miRNAs and, thus, it filtered them out prior to comparison; 13 miRNAs were not significant using either edgeR or DESeq2, *i.e.* they did not show a p-value < 0.05 and a FC ≥ |1.5| at both tests (Table S2b). Sources of these differences between our and the original analysis reside in the pre-alignment filtration steps, the normalization procedures, and statistical tests used to detect significantly DE miRNAs. Our pipeline makes use of more stringent quality checks and filtration, as well as of widely accepted statistical approaches that are more appropriate for count data: in addition, we considered only overlapping results between the two statistical algorithms, which enhance the confidence in their robustness. However, many of these miRNAs should be considered as false positives, because the significance values were not corrected for multiple testing. Applying such a correction, only 92 miRNAs resulted still significantly DE at an FDR < 0.05 by both edgeR and DESeq2 (Table S2a). The expression profile of these 92 miRNAs were able to discriminate correctly between ATM-deficient cells and wild type controls at an unsupervised hierarchical clustering analysis (Fig. 3a). Notably, all tumor suppressor miRNAs (miR-16, miR-96, miR-141, miR-200a, and members of the miR-29 family, which included miR-29b-1, miR-29b-2, and miR-29c) and 3 oncomiRs (miR-106a, miR-146a, and miR-221) were confirmed to be identical between the original study [21] and our reanalysis (Fig. 3b).

2.2.2. Discovery of miRNA editing/isomiRs

To demonstrate the functionality of Module 2a for the discovery of isomiRs within the dataset 1, the analysis was completed sequentially on each sample using information from miRBase miRNA precursor and then detailed profiling in interactive html format (Table S8) was created. It is interesting to note that results report no significance 5′ and 3' isomiRs differences (± 1 nt) between ATM-deficient vs normal HMECCs. Therefore, it shows that main effect is based on known miRNAs expressions changes between two biological conditions. However, other useful profiling of miRNAs within the sample such as list of seed abundance and counts of miRNA cluster are generated for further research concussions.

2.2.3. Prediction of novel miRNA candidates

Next, we tested the function of Module 3a for identifying novel miRNAs on dataset 1. The pipeline predicted 1039 novel miRNA candidates with at least 1 CPM in at-least 3 samples. To limit false positives and restrict the analysis to the most robust candidates, we considered truly novel only 93 miRNAs that were expressed in ≥3 samples to be truly novel, either ATM-deficient or wild type cells with ≥10 CPM. Table S3 reports these novel candidates, along with their genome coordinates and expression counts for each sample. The user may use this list for differential analysis and prediction of associated targets and pathways and/or for experimental validation. These findings may pave the way to the discovery of new biomarkers of cancer susceptibility in ATM-deficient patients.

2.2.4. Target gene prediction and functional enrichment

We tested the performance of Module 3b in predicting target genes, pathway and gene set enrichment, comparing results with those reported in Ref. [21]. We predicted target genes of the 124 miRNAs identified as DE by Module 2, using the 3' UTR sequences of human genome (Ensembl Biomart release 81) as target sequences. To reduce false predictions, we set a stringent hybridization threshold for RNAhybrid and miRanda at −20 kcal/mol. We found 14,959 unique targets.

We used coding target genes to predict the biological functions associated with ATM-deficiency related changes in miRNA expression. Analysis with clusterProfiler showed that 77 GO biological processes and 33 DO terms were associated with ATM depletion at an adjusted p-value < 0.05 (Table S4). Notably, among the most significant, there are GO and DO terms related to cell cycle, regulation of locomotion, neurological and cognitive disorders, hereditary breast ovarian cancer, which are all consistent with the clinical hallmarks of ATM-deficiency related syndrome. Additionally, the pipeline reported associations with ATM-deficiency dependent miRNA dysregulation (adjusted p-value < 0.05) 73 KEGG pathways (Table S5), which includes a number of signaling, metabolic, and cancer-related pathways. It is noteworthy to mention that the authors of the original study performed analysis of gene expression changes by microarray and correlated it with target prediction of miRNA significantly regulated. Interestingly, our analysis based only on prediction gave similar biological insights. Indeed, our findings are consistent with the original report, indicating miRNAs associated with cancer formation and progression.

2.3. CBS-miRSeq performance for favorable (event-free survivors) vs. unfavorable (died of disease) human neuroblastoma tumor samples (color-space dataset)

We used this dataset to check the performance of CBS-miRSeq whether it identifies same miRNA biomarkers and/or improving the results obtained in previous study [23]. For a fair comparison, we used the same reference genome (Human Genome Assembly GRCh37.67) and the same miRBase annotation (v.18) as in the previous analysis [23]. Starting from ~188 million reads, Module 1 filtered (low quality reads) out 57.3 ± 8.0% of raw reads from each library. Remaining reads were trimmed by removing the 3' adapter sequences and too short reads (< = 15 nt) were discarded. Clipped reads of each sample (~51 million in total) were aligned to the genome with the mapping rates of between 55.2 and 78.1%. (Table S6). The total number of unique annotated miRNAs found (1919) was similar to that of the original work [23].

Next, Module 1 computed expression of mature miRNAs for analyzing downstream differential comparison between favorable (event-free survivors) and unfavorable (died from disease) human neuroblastoma tumor samples. Using the default filtering option (miRNAs with < 1 CPM in ≥50% of the samples), we deemed expressed 646 unique miRNAs. We considered the overlapping results obtained by the two statistical packages (DESeq2 and edgeR) implemented in Module 2. Both tests found 98 miRNAs with a nominal significance p-value < 0.05 (Table S7) on which 18 of them were significantly DE at FDR < 0.05 and $\log_2$FC ≥ |1.5| (Fig. 4). CBS-miRSeq confirmed all five biomarkers (i.e. miR-181a-2-3p overexpressed and miR-628–5p, miR-744–5p, miR-1249 and miR-3612 decreased in patients with unfavorable outcome) reported in the original study [23]. In addition, we identified 13 other DE miRNAs significantly up-regulated (miR-181a-5p, miR-675–3p, miR-99a-5p, miR-325, miR-181b-5p, miR-551b-3p, miR-1179, miR-3648, and miR-575, in order of significance) and down-regulated (miR-1912, miR-196a-5p, miR-204–5p, and miR-149–5p) in unfavorable human neuroblastoma tumor samples (Fig. 4). Overall, our pipeline outperforms previous analysis, increasing the number of potential miRNA biomarkers of unfavorable tumor outcome. Interestingly, many of these miRNAs have been associated with and/or demonstrated to have a role in neuroblastoma, leukemia, cancer, and tumor

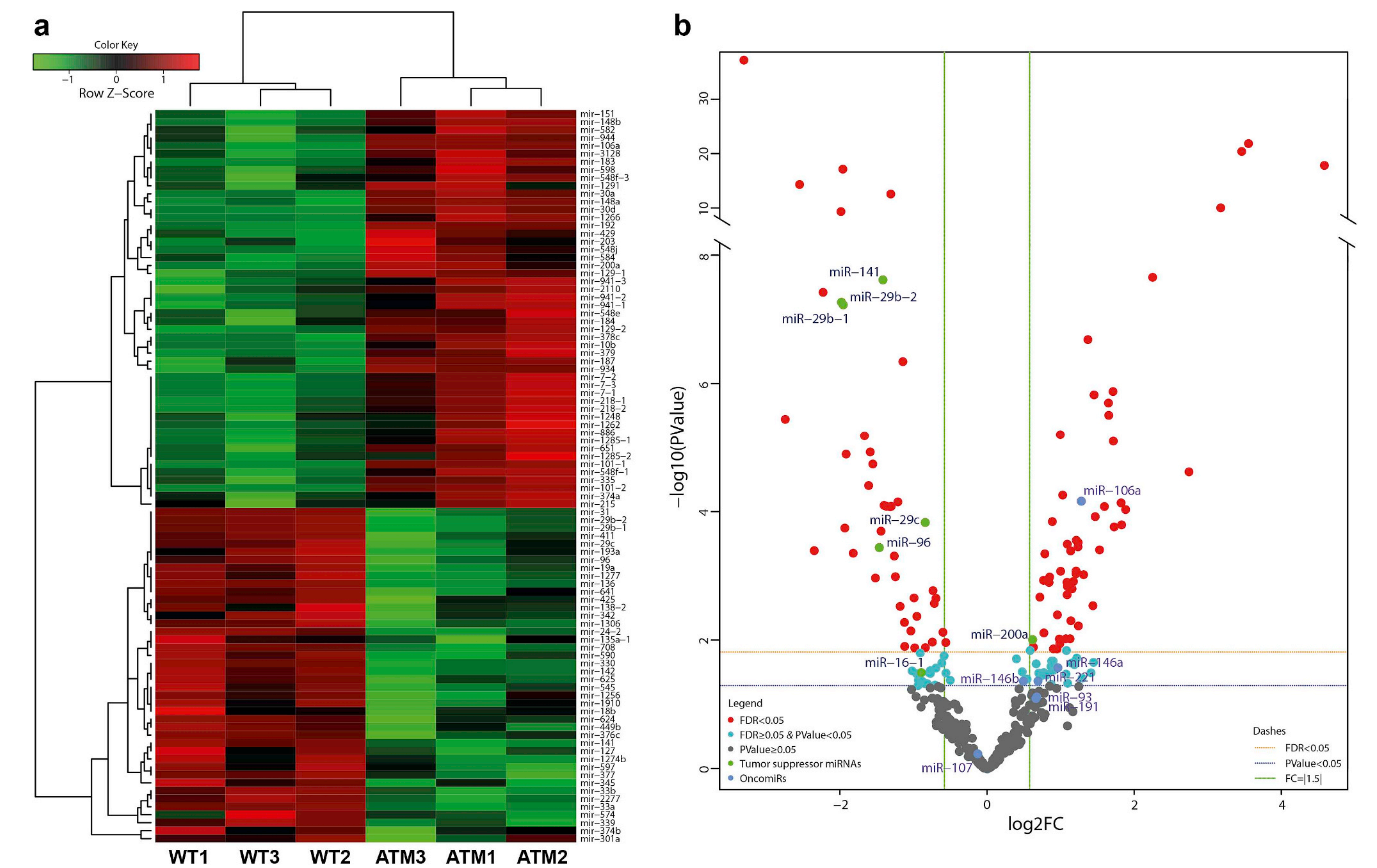

Fig. 3. Differentially expressed miRNAs detected by CBS-miRSeq in ATM-deficient human mammary epithelial cells. (**a**) Unsupervised hierarchical clustering showed that 92 differentially expressed miRNAs, significant at an FDR < 0.05 by both edgeR and DESeq2, discriminated correctly between ATM-deficient cells and wild type controls (WT). (**b**) Volcano plot of the differentially expressed miRNAs as detected by edgeR: significant (red dots), nominally significant (aqua green), and nonsignificant (grey) miRNAs are reported. Tumor suppressor miRNAs and OncomiRs are highlighted, respectively, in bright green and blue.

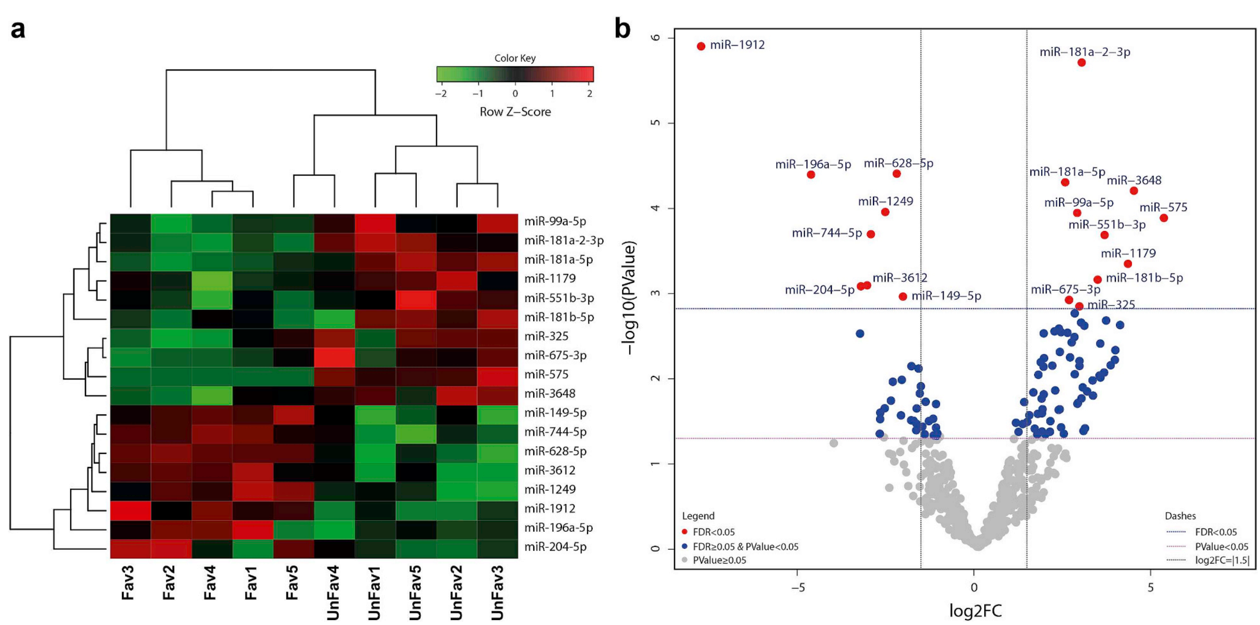

Fig. 4. Differentially expressed miRNAs detected by CBS-miRSeq in favorable (event-free survivors) vs. unfavorable (died of disease) human neuroblastoma tumor samples. (**a**) Unsupervised hierarchical clustering of the 18 miRNAs significantly different at an FDR < 0.05 by both edgeR and DESeq2 between favorable (Fav) and unfavorable (UnFav) human neuroblastoma tumor samples. (**b**) Volcano plot of the differentially expressed miRNAs as detected by DESeq2: significant (red dots), nominally significant (blue), and nonsignificant (grey) miRNAs are reported.

Table 1
Comparison of the CBS-miRSeq features with existing tools for miRNA-Sequencing data analysis.

	CBS-miRSeq	miARmaSeq	CAP-miRSeq	mirTools2	wapRNA	ncPRO-seq	miRanalyzer	DARIO	miRDeep2	omiRas
Input data										
Color-space raw reads	✓				✓	✓				
Base-space raw reads	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Mapped reads		✓				✓				
Model Organism										
Few								✓		✓
Almost any				✓	✓	✓	✓			
Any, recent releases	✓	✓	✓			✓			✓	
Reads quality control										
Visualization	✓	✓	✓			✓				✓
Filtering	✓				✓					
Differential expression analysis										
One statistical test	✓	✓	✓	✓	✓		✓			✓
More than one test	✓	✓								
Output of combined results	✓									
miRNA editing/isomiRs detection	✓		✓				✓			
Estimate of other transcripts	✓		✓	✓	✓	✓	✓	✓		✓
Novel miRNA prediction										
Filtering of known ncRNAs	✓			✓			✓	✓		✓
Differential expression analysis of novel miRNAs	✓		✓	✓		✓				✓
miRNA target prediction										
One algorithm	✓			✓	✓					
More than one algorithm combined	✓			✓	✓					
Novel miRNA target prediction	✓			✓	✓					
Functional analysis	✓	✓		✓	✓					
Modular analysis	✓	✓	✓						✓	
Usage type										
Local	✓	✓	✓	✓	✓	✓	✓		✓	
Web				✓	✓	✓	✓	✓		✓

proliferation and include miR-99 [38,39], miR-149–5p [40], miR-181a-5p [41], miR-181b-5p [42], miR-196a-5p [43], miR-204–5p [44], miR-551b-3p [45], miR-575 [46], miR-675–3p [47], miR-1179 [48], miR-3648 [49]. The results further validated the reliability of our workflow and analysis.

3. Conclusions

We developed CBS-miRSeq, a module-based pipeline that allows performing robust and comprehensive analyses for miRNA-Seq data. Alternatively, the user can choose performing separately different steps in data analysis. Table 1 illustrates the advantages of CBS-miRSeq in comparison with other publicly available tools. Compared to them, CBS-miRSeq has several key features:

1 accepts different formats of raw reads (color- and base-space)
2 analyzes sequencing data from any species (animal only)
3 performs robust differential expression analysis integrating two statistical packages, visualizing both separate and overlapping results
4 identifies isomiRs, based on provided miRNA annotations such as miRBase v21
5 conservatively predicts novel miRNA candidates
6 predicts the most consistent miRNA:mRNA unique pairs, integrating two target prediction algorithms for known and novel miRNAs
7 and performs extensive functional annotation analysis, based on different ontologies and annotated pathways.

Overall, these characteristics make CBS-miRSeq the most comprehensive and powerful tool compared to other existing tools.

We benchmarked the performance of our tool on previously published datasets and found that CBS-miRSeq replicates and outperforms previous analyses in terms of differential miRNAs, prediction of target genes of DE miRNAs, and gene functional analysis. Our tool may process small to large number of samples produced by the sequencing platforms such as Illumina and SOLiD sequencing systems. The CBS-miRSeq is a fully customized and flexible tool that provides clear and prominent results. Several outputs (in html, pdf, tiff, xls, and csv formats) generated from CBS-miRSeq allow complete visual inspection and interpretation of the miRNA biomarker as well as their targeted gene regulation mechanism and disease associations. Further, our speed test results show that using CBS-miRSeq workflow is promising when it comes to use computational resources. We believe that our workflow would be insightful towards miRNA sequence data analysis and at the same time beneficial to the computational biology community.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability and requirements

Project name: CBS-miRSeq: a comprehensive tool for accurate and extensive analyses of microRNA-sequencing data.
Project page: https://bitbucket.org/unique379/cbs-mirseq/overview.
Operating system(s): Linux/Unix.
Programming language: Bash, Perl, and R.
Other requirements: Fastx (v0.0.14), FastQC (v0.10.1), SOLiD_preprocess_filter_v2. pl, cutadapt (v1.6), BFAST (v0.6.5a), Bowtie ($\leq$v1.0), featureCounts (v1.4.6), SAMtools (v0.1.18), BEDtools (v2.23.0), miRspring (v1.2), miRDeep2 (v2.0.05), RNAhybrid (v2.1.1), miRanda (v3.3a), Bioconductor packages ($>$ v3).
License: GNU GPLv3.
Any restriction to use by non-academics: None.

Conflicts of interest

The authors have declared that no competing interests exist.

Funding

Institutional Research Funds (Italian Ministry of Health, Ricerca Corrente 2016) supported this work.

Authors' contributions

RKK designed and developed the pipeline, performed data analysis, and drafted the manuscript. MC contributed to the pipeline optimization. GIC and RB supervised the project. GIC critically revised and wrote the final version of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

List of abbreviations

ATM	Ataxia Telangiectasia Mutated kinase
CPM	count per million
CV	coefficient of variation
DE	differentially expressed
DO	Disease Ontology
FC	fold change
FDR	False Discovery Rate
GO	Gene Ontology
HME-CCs	human mammary epithelial cells
miRNA	microRNA
miRNA-Seq	microRNA-sequencing
SRA	Sequence Read Archive

Appendix ASupplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2019.05.019.

References

[1] D.P. Bartel, MicroRNAs: target recognition and regulatory functions, Cell 136 (2) (2009) 215–233.
[2] J. Winter, S. Jung, S. Keller, R.I. Gregory, S. Diederichs, Many roads to maturity: microRNA biogenesis pathways and their regulation, Nat. Cell Biol. 11 (3) (2009) 228–234.
[3] I. Bray, K. Bryan, S. Prenter, P.G. Buckley, N.H. Foley, D.M. Murphy, L. Alcock, P. Mestdagh, J. Vandesompele, F. Speleman, et al., Widespread dysregulation of MiRNAs by MYCN amplification and chromosomal imbalances in neuroblastoma: association of miRNA expression with survival, PLoS One 4 (11) (2009) e7850.
[4] V. Ambros, MicroRNAs and developmental timing, Curr. Opin. Genet. Dev. 21 (4) (2011) 511–517.
[5] A. Esquela-Kerscher, F.J. Slack, Oncomirs - microRNAs with a role in cancer, Nat. Rev. Canc. 6 (4) (2006) 259–269.
[6] C. Urbich, A. Kuehbacher, S. Dimmeler, Role of microRNAs in vascular diseases, inflammation, and angiogenesis, Cardiovasc. Res. 79 (4) (2008) 581–588.
[7] E. Zhu, F. Zhao, G. Xu, H. Hou, L. Zhou, X. Li, Z. Sun, J. Wu, mirTools: microRNA profiling and discovery based on high-throughput sequencing, Nucleic Acids Res. 38 (2010) W392–W397 Web Server issue).
[8] M. Hackenberg, M. Sturm, D. Langenberger, J.M. Falcon-Perez, Aransay AM: miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments, Nucleic Acids Res. 37 (2009) W68–W76.
[9] M. Fasold, D. Langenberger, H. Binder, P.F. Stadler, S. Hoffmann, DARIO: a ncRNA detection and analysis tool for next-generation sequencing experiments, Nucleic Acids Res. 39 (2011) W112–W117.
[10] W.M. Zhao, W.F. Liu, D.M. Tian, B.X. Tang, Y.Q. Wang, C.X. Yu, R.J. Li, Y.C. Ling, J.Y. Wu, S.H. Song, et al., wapRNA: a web-based application for the processing of RNA sequences, Bioinformatics 27 (21) (2011) 3076–3077.
[11] C.J. Chen, N. Servant, J. Toedling, A. Sarazin, A. Marchais, E. Duvernois-Berthet, V. Cognat, V. Colot, O. Voinnet, E. Heard, et al., ncPRO-seq: a tool for annotation and profiling of ncRNAs in sRNA-seq data, Bioinformatics 28 (23) (2012) 3147–3149.
[12] S. Muller, L. Rycak, P. Winter, G. Kahl, I. Koch, B. Rotter, omiRas: a Web server for differential expression analysis of miRNAs derived from small RNA-Seq data, Bioinformatics 29 (20) (2013) 2651–2652.
[13] M.R. Friedlander, S.D. Mackowiak, N. Li, W. Chen, N. Rajewsky, miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades, Nucleic Acids Res. 40 (1) (2012) 37–52.
[14] Z. Sun, J. Evans, A. Bhagwate, S. Middha, M. Bockol, H. Yan, J.P. Kocher, CAP-miRSeq: a comprehensive analysis pipeline for microRNA sequencing data, BMC Genomics 15 (2014) 423.
[15] E. Andres-Leon, R. Nunez-Torres, A.M. Rojas, miARma-Seq: a comprehensive tool for miRNA, mRNA and circRNA analysis, Sci. Rep. 6 (2016) 25749.
[16] M.I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, Genome Biol. 15 (12) (2014) 550.
[17] M.D. Robinson, D.J. McCarthy, G.K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data, Bioinformatics 26 (1) (2010) 139–140.
[18] B. John, A.J. Enright, A. Aravin, T. Tuschl, C. Sander, D.S. Marks, Human MicroRNA targets, PLoS Biol. 2 (11) (2004) e363.
[19] M. Rehmsmeier, P. Steffen, M. Hochsmann, R. Giegerich, Fast and effective prediction of microRNA/target duplexes, RNA 10 (10) (2004) 1507–1517.
[20] D.T. Humphreys, C.M. Suter, miRspring: a compact standalone research tool for analyzing miRNA-seq data, Nucleic Acids Res. 41 (15) (2013) e147.
[21] J.E. Hesse, L. Liu, C.L. Innes, Y. Cui, S.S. Palii, R.S. Paules, Genome-wide small RNA sequencing and gene expression analysis reveals a microRNA profile of cancer susceptibility in ATM-deficient human mammary epithelial cells, PLoS One 8 (5) (2013) e64779.
[22] J.H. Schulte, T. Marschall, M. Martin, P. Rosenstiel, P. Mestdagh, S. Schlierf, T. Thor, J. Vandesompele, A. Eggert, S. Schreiber, et al., Deep sequencing reveals differential expression of microRNAs in favorable versus unfavorable neuroblastoma, Nucleic Acids Res. 38 (17) (2010) 5919–5928.
[23] S. Rahmann, M. Martin, J.H. Schulte, J. Koster, T. Marschall, A. Schramm, Identifying transcriptional miRNA biomarkers by integrating high-throughput sequencing and real-time PCR data, Methods 59 (1) (2013) 154–163.
[24] A. Sasson, T.P. Michael, Filtering error from SOLiD output, Bioinformatics 26 (6) (2010) 849–850.
[25] N. Homer, B. Merriman, S.F. Nelson, BFAST: an alignment tool for large scale genome resequencing, PLoS One 4 (11) (2009) e7767.
[26] M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads, EMBnetjournal 17 (1) (2011) 10–12.
[27] B. Langmead, C. Trapnell, M. Pop, S.L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, Genome Biol. 10 (3) (2009) R25.
[28] T. Lassmann, Y. Hayashizaki, C.O. Daub, SAMStat: monitoring biases in next generation sequencing data, Bioinformatics 27 (1) (2011) 130–131.
[29] A. Biosystems, Principles of di-base sequencing and the advantages of color space analysis in the SOLiD system, Applied Biosystems Application Note 139AP (2008) 10-01.
[30] Y. Liao, G.K. Smyth, W. Shi, featureCounts: an efficient general purpose program for assigning sequence reads to genomic features, Bioinformatics 30 (7) (2014) 923–930.
[31] S. Griffiths-Jones, H.K. Saini, S. van Dongen, A.J. Enright, miRBase: tools for microRNA genomics, Nucleic Acids Res. 36 (2008) D154–D158.
[32] S. Anders, W. Huber, Differential expression analysis for sequence count data, Genome Biol. 11 (10) (2010) R106.
[33] M.D. Robinson, A. Oshlack, A scaling normalization method for differential expression analysis of RNA-seq data, Genome Biol. 11 (3) (2010) R25.
[34] A. Rau, M. Gallopin, G. Celeux, F. Jaffrezic, Data-based filtering for replicated high-throughput transcriptome sequencing experiments, Bioinformatics 29 (17) (2013) 2146–2152.
[35] A.I.K.S. Petrov, R. Gibson, et al., RNAcentral: an international database of ncRNA sequences, Nucleic Acids Res. 43 (Database issue) (2015) D123–D129.
[36] R.C. Gentleman, V.J. Carey, D.M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, et al., Bioconductor: open software development for computational biology and bioinformatics, Genome Biol. 5 (10) (2004) R80.
[37] G. Yu, L.G. Wang, Y. Han, Q.Y. He, clusterProfiler: an R package for comparing biological themes among gene clusters, OMICS A J. Integr. Biol. 16 (5) (2012) 284–287.
[38] J.H. Schulte, S. Horn, T. Otto, B. Samans, L.C. Heukamp, U.C. Eilers, M. Krause, K. Astrahantseff, L. Klein-Hitpass, R. Buettner, et al., MYCN regulates oncogenic MicroRNAs in neuroblastoma, Int. J. Cancer 122 (3) (2008) 699–704.
[39] C. Oneyama, J. Ikeda, D. Okuzaki, K. Suzuki, T. Kanou, Y. Shintani, E. Morii, M. Okumura, K. Aozasa, M. Okada, MicroRNA-mediated downregulation of mTOR/FGFR3 controls tumor growth induced by Src-related oncogenic pathways, Oncogene 30 (32) (2011) 3489–3501.
[40] F. Wang, Y.L. Ma, P. Zhang, T.Y. Shen, C.Z. Shi, Y.Z. Yang, M.P. Moyer, H.Z. Zhang, H.Q. Chen, Y. Liang, et al., SP1 mediates the link between methylation of the tumour suppressor miR-149 and outcome in colorectal cancer, J. Pathol. 229 (1) (2013) 12–24.
[41] G. Chen, W. Zhu, D. Shi, L. Lv, C. Zhang, P. Liu, W. Hu, MicroRNA-181a sensitizes human malignant glioma U87MG cells to radiation by targeting Bcl-2, Oncol. Rep. 23 (4) (2010) 997–1003.
[42] J. Ji, T. Yamashita, A. Budhu, M. Forgues, H.L. Jia, C. Li, C. Deng, E. Wauthier, L.M. Reid, Q.H. Ye, et al., Identification of microRNA-181 by genome-wide screening as a critical player in EpCAM-positive hepatic cancer stem cells,

Hepatology 50 (2) (2009) 472–480.
[43] D.M. Maru, R.R. Singh, C. Hannah, C.T. Albarracin, Y.X. Li, R. Abraham, A.M. Romans, H. Yao, M.G. Luthra, S. Anandasabapathy, et al., MicroRNA-196a is a potential marker of progression during Barrett's metaplasia-dysplasia-invasive adenocarcinoma sequence in esophagus, Am. J. Pathol. 174 (5) (2009) 1940–1948.
[44] J. Ryan, A. Tivnan, J. Fay, K. Bryan, M. Meehan, L. Creevey, J. Lynch, I.M. Bray, A. O'Meara, L. Tracey, et al., MicroRNA-204 increases sensitivity of neuroblastoma cells to cisplatin and is associated with a favourable clinical outcome, Br. J. Canc. 107 (6) (2012) 967–976.
[45] J.M. Cummins, Y. He, R.J. Leary, R. Pagliarini, L.A. Diaz Jr., T. Sjoblom, O. Barad, Z. Bentwich, A.E. Szafranska, E. Labourier, et al., The colorectal microRNAome, Proc. Natl. Acad. Sci. U.S.A. 103 (10) (2006) 3687–3692.
[46] J.E. Lee, E.J. Hong, H.Y. Nam, M. Hwang, J.H. Kim, B.G. Han, J.P. Jeon, Molecular signatures in response to Isoliquiritigenin in lymphoblastoid cell lines, Biochem. Biophys. Res. Commun. 427 (2) (2012) 392–397.
[47] B.K. Dey, K. Pfeifer, A. Dutta, The H19 long noncoding RNA gives rise to microRNAs miR-675-3p and miR-675-5p to promote skeletal muscle differentiation and regeneration, Genes Dev. 28 (5) (2014) 491–501.
[48] S. Subramanian, W.O. Lui, C.H. Lee, I. Espinosa, T.O. Nielsen, M.C. Heinrich, C.L. Corless, A.Z. Fire, M. van de Rijn, MicroRNA expression signature of human sarcomas, Oncogene 27 (14) (2008) 2015–2026.
[49] E. Meiri, A. Levy, H. Benjamin, M. Ben-David, L. Cohen, A. Dov, N. Dromi, E. Elyakim, N. Yerushalmi, O. Zion, et al., Discovery of microRNAs and other small RNAs in solid tumors, Nucleic Acids Res. 38 (18) (2010) 6234–6246.

Doctor Kesharwani is a Bioinformatician with over 3 years of computational biology experience from around the world and currently employed as Application computational scientist at The Jackson Laboratory for Genomic Medicine, CT, USA. His range of experiences includes Cancer-genomics and transcriptomics.

In Jan 2016, Dr. Kesharwani completed Ph.D. in bioengineering and bioinformatics from University of Pavia, Italy with the joint the collaboration with the unit of Immunology and Functional Genomics, hospital Monzino, Milan, Italy. During his thesis, he studied on “Role of RAGE in age-dependent cardiac remodeling: integrated time-course expression analysis of cardiac miRNome and transcriptome”. In this work, he focused on the relevance of miRNA and gene expression analysis in cardiac aging. He earned Masters in Bioinformatics (2007–09) from University of Allahabad, India. And before, he worked on Designing of putative siRNA against Geminivirus resistant Papaya crop’.