Peer-reviewed version available at International Journal of Molecular Sciences 2020, 21, 293; doi:10.3390/ijms21010293

Docker4Circ: A Framework for a Reproducible Characterization of CircRNAs from RNA-Seq Data

Giulio Ferrero[1,#], Nicola Licheri[1,#], Lucia Coscujuela Tarrero[2], Carlo De Intinis[1], Valentina Miano[2,3], Raffaele Adolfo Calogero[4], Francesca Cordero[1], Michele De Bortoli[2,#,*] and Marco Beccuti[1,#]

[1] Department of Computer Science, University of Turin, Turin, Italy.
[2] Department of Clinical and Biological Sciences, University of Turin, Turin, Italy
[3] Division of Cellular and Molecular Pathology, Department of Pathology, University of Cambridge, Addenbrooke's Hospital, Cambridge
[4] Department of Molecular Biotechnology and Health Sciences, University of Turin, Turin, Italy

* Correspondence: michele.debortoli@unito.it

Abstract: Recently the increased cost-effectiveness of high-throughput technologies has made available a large number of RNA sequencing datasets to identify circular RNAs (circRNAs). However, despite many computational tools were developed to predict circRNAs, a limited number of workflows exists to predict and to characterize circRNAs. Moreover, to the best of our knowledge, these available workflows do not ensure computational reproducibility and require advanced bash scripting skills to be correctly installed and used.
To cope with these critical aspects we present Docker4Circ, a new computational framework designed for a comprehensive analysis of circRNAs composed of: circRNAs prediction, classification and annotation using public databases, the back-splicing sequence reconstruction; the internal alternative splicing of circularizing exons; the alignment-free circRNAs quantification from RNA-Seq reads, and, finally, their differential expression analysis. Docker4Circ was specifically designed for making easier and more accessible circRNAs analysis thanks to the following features: (i) its R interface; (ii) the encapsulation of its computational tasks into a docker image; (iii) an available user-friendly Java GUI Interface. Furthermore, Docker4Circ ensures a reproducible analysis because all its tasks were embedded into a docker image following the guidelines provided by Reproducible Bioinformatics Project (RBP, http://reproducible-bioinformatics.org/). The effectiveness of Docker4Circ was demonstrated on a real case study whose goal is to characterize the circRNAs predicted in colorectal cancer cell lines and quantified in public RNA-Seq experiments performed on primary tumor tissues.
In conclusion, we propose Docker4Circ as a framework for reproducible and comprehensive analyses of circRNAs to efficiently exploit their biological role.

Keywords: circRNA; reproducible analysis; pipeline, Docker images

1. Introduction

CircRNAs are circular RNA transcripts formed by Back-Splicing (BS) events involving a downstream 5′ splice site and an upstream 3′ splice site [1]. To date five categories of circRNAs have been defined: exonic, arising from one or more exons of the linear transcript; intronic, arising entirely from an intron of the linear transcript; antisense, arising from exons of the linear transcript on the opposite strand; intragenic, arise from same gene locus of the linear transcript but is not exonic neither intronic; and intergenic, arising from non genic regions [2]. In 2013, Jeck et al. described two possible models for circRNA formation [3]. The first one, named "lariat driven circularization" suggests the generation of a linear RNA with skipped exons and a long intron lariat containing these skipped exons, which are then back-spliced to generate a circRNA. The second

model, called "intron-pairing-driven circularization" suggests the generation of a circRNA together with an exon-intron(s)-exon intermediate.

Many computational tools were developed for predicting circRNAs from RNA-Seq data [2,4], and different tools were proposed for the post-prediction analyses including FUCHS [5] and CIRI-AS [6] for defining circRNAs internal structures, Sailfish-circ [7], and circTest [8] for circRNAs quantification and differential expression analysis and CircView for visualization of circRNAs predictions [9]. Furthermore, our group recently proposed the CircHunter algorithm for the characterization and quantification of circRNAs using public RNA-Seq datasets [10].

Several circRNAs databases are also becoming available including circBase [11], Tissue-Specific CircRNA Database (TSCD) [12], circRNADb [13], and Circ2Traits [14] increasing the accessible information on annotated circRNAs.

Concurrently with the increase of computational resources dedicated to the circRNAs research, a large amount of RNA-Seq datasets suitable for circRNAs detection is nowadays available. Specifically, many public polyA-depleted or total RNA-Seq can be used to quantify circRNAs expression in different biological contexts. For this purpose, direct quantification of circRNAs levels will be relevant to expand rapidly the knowledge about the circRNAs regulation and functions in different experimental or pathological contexts. All these aspects clearly highlight the need for workflows able to provide reproducible analysis to achieve a comprehensive characterization of circRNAs. However, most of the published circRNAs analysis pipelines are focused on the circRNAs prediction based on a specific algorithm. More extensive pipelines like CirCompara [15], Ularcirc [16], and circtools [17] provide instead multiple functions, but they do not guarantee the reproducibility of the analysis as suggested by Sandve and colleagues [18].

Here, we present Docker4Circ a comprehensive framework for circRNAs analysis merging four different modules into a reproducible analysis framework from circRNAs prediction to their expression analysis. The distinctive features of Docker4Circ are (i) the usability and portability on all Unix-like systems achieved through docker containerization, an R interface and a Java GUI; and (ii) the computational reproducibility since it follows the guidelines provided by Reproducible Bioinformatics Project (RBP) [19,20].

2. Results

2.1 A framework to create modular workflows for reproducible analysis of circRNA data

Docker4Circ is a computational framework composed of four modules embedded into Docker images (see the chart in Figure 1). In detail, the modules are designed for circRNAs prediction (Module 1), circRNAs classification and annotation (Module 2), the BS sequence analysis (Module 3), and circRNAs expression analysis (Module 4). Each module-analysis is embedded in a Docker image which is called through a specific R function which was designed following the guidelines of the RBP project (http://reproducible-bioinformatics.org) and included in the Docker4Seq R package [19]. These functions can be executed by R/bash scripts or through a GUI based on Java Swing Class which is a part the 4SeqGUI project https://github.com/mbeccuti/4SeqGUI (Figure 2A). Details on the usage of each function are reported in the Supplementary Material of the manuscript. Docker4Circ functions and associated test data can be downloaded from https://github.com/kendomaniac/docker4seq.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 July 2019 doi:10.20944/preprints201907.0219.v1

Peer-reviewed version available at International Journal of Molecular Sciences 2020, 21, 293; doi:10.3390/ijms21010293

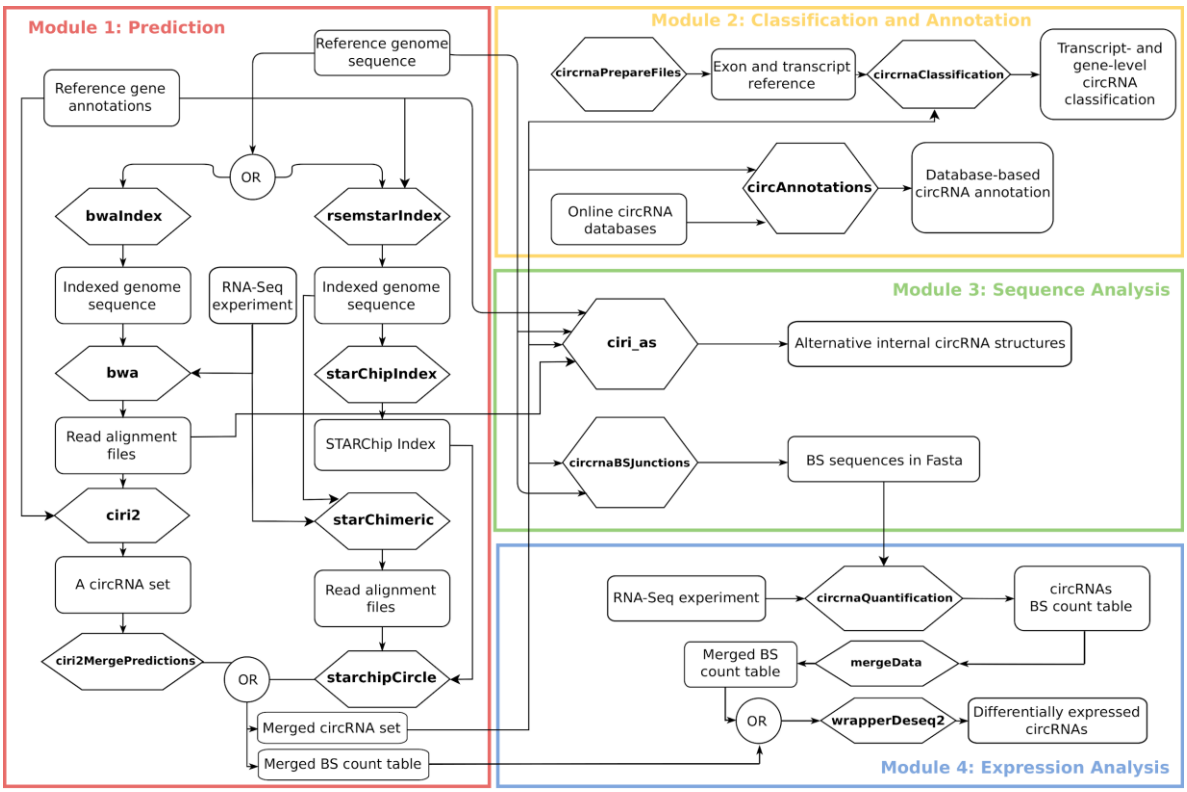

Figure 1. Schematic representation of the Docker4Circ modules with an indication of all the functions (reported in bold in the octagons) and the input/output files involved (reported in the squares). The different modules implemented in the framework are reported with different colors.

2.1.1. Module 1: circRNAs prediction.

This module is designed to predict the circRNAs using the CIRI2 [21] or the STAR Chimeric Post (STARChip) tools [22].

The CIRI2 prediction analysis is implemented by the *ciri2* function which takes as input the BWA alignment files, the fasta of the reference genome, the gene annotation files (GTF or GFF), and the CIRI2 algorithm parameters. The function *wrapperCiri* embeds the *ciri2* function with two other functions: the *fastqc* function for the evaluation of the quality control of the input RNA-Seq reads using FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) and the *bwa* function for BWA read alignment [23]. This wrapper function provides an automated approach to predict circRNAs starting from the raw RNA-Seq reads. Finally, the function *ciri2MergePredictions* can be used to merge the predicted circRNAs from each sample. This function creates a joined BS count table from different CIRI2 runs and removes those circRNAs characterized by a user-defined minimum number of BS-supporting reads in each sample and an average number of BS-supporting reads among biological replicates of the same experimental condition.

Alternatively, the STARChip prediction-pipeline is implemented in the functions *rsemstarIndex*, *starChimeric*, *starChipIndex*, and *starchipCircle*. The genomic sequence is first indexed using STAR through the function rsemstarIndex, then the raw RNA-Seq reads are aligned against indexed genomes using the *starChimeric* function which applies the chimeric alignment mode of STAR algorithm. The chimeric alignments are then evaluated for candidate circRNA junctions using the function *starChipIndex* to pre-process the reference STAR genome index. Finally, the *starchipCircle* function returns a list of circRNAs supported by a user-defined minimum number of BS-supporting reads detected in a defined number of samples. An additional filter based on the count per million reads, the linear splicing information, as well as the circRNA annotation, can be also provided by the function.

2.1.2. Module 2: the circRNAs classification and annotation.

In this module, an extensive annotation of the circRNAs predicted in Module 1 is performed through the following two functions: *circrnaClassification* and *circAnnotations*. The *circrnaClassification* considers the Ensembl transcriptome annotations by overlapping the exons genomic coordinates against circRNAs genomic coordinates. Each overlap is classified on BS position within the annotations and the number of exons involved. This function is able to distinguish unconventional circRNAs classes including intronic and intergenic circRNAs or whose BS sites do not coincide with exon boundaries (putative exon circRNAs), as reported in [10]. Specifically, if both circRNA BS sites coincide with the exon boundaries then circRNAs are classified as monoexonic or multiexonic depending on the number of exons involved. Conversely, the circRNAs are classified as intronic, intergenic, or putative exon if at least one BS site coincides with intronic, intergenic, or intra-exonic regions, respectively. *circrnaClassification* takes as input the list of predicted circRNAs, the exons/transcripts data, and the selected genome assembly. The output of this function is thus a circRNAs classification at the transcript and gene level. The *circrnaPrepareFiles* function can be used to retry the exons/transcript data, in the case are not available, using biomaRt R package [24].

On the other side, the *circAnnotations* function compares the list of predicted circRNAs with a set of online circRNA databases. Based on the genomic coordinates of the input circRNAs, the function provides to the user their related information from two databases: circBase [11] and TSCD [12]. Each circRNA identified in the databases is associated with the following information: genomic coordinates and length, the cell lines in which a circRNA was detected, the best overlapping gene and transcript, the study in which it was discovered (from circBase), the information about the fetal or adult tissue in which the circRNA was detected, the algorithm applied with the number of BS junctions, the overlapped genes, miRNAs, and proteins predicted to bound its sequence (from TSCD).

Observe that if the version of the genome assembly used for the circRNA prediction is not compatible with those present in circBase and TSCD, the genomic coordinates are converted using the UCSC LiftOver program [25].

2.1.3. Module 3: the circRNAs sequence analysis.

This module is designed for the analysis of the back-splicing circRNAs sequence. The *circrnaBSJunctions* function takes as input the list of selected circRNAs and the human genome sequence providing the reconstructed BS sequences of the input circRNAs. The function exploits a python script which identifies two sets of genomic coordinates of 35 bp starting from the boundaries of the exons involved in the circularization. To reconstruct BS sequences, these genomic coordinates are then used as input for the functions getSeq and xscat provided by the R package GenomicRanges [26].

Moreover, *ciri_as* function implements the CIRI-AS analysis to detect internal Alternative Splicing (AS) events involving the exons composing the predicted circRNAs [6]. Specifically, the function takes as input BWA alignment files used for the CIRI2 circRNAs prediction, the reference genomic sequence (in fasta format), the gene annotations (in GTF/GFF), and the list of CIRI2-predicted circRNAs. Its output is the list of alternative splicing events involving the exons of the studied circRNAs.

2.1.4. Module 4: circRNAs expression analysis.

This module provides the expression analysis of circRNAs. The differential expression analysis is based on the DESeq2 R package [27]. The function *wrapperDeseq2* takes as input the circRNA BS count table created at the end of Module 1 functions and computes the differential expression analysis for each circRNA with respect to a specific covariate. If the user is interested in quantifying and analyzing the expression of the predicted circRNAs in an independent RNA-Seq experiment it

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 July 2019 doi:10.20944/preprints201907.0219.v1

Peer-reviewed version available at International Journal of Molecular Sciences 2020, 21, 293; doi:10.3390/ijms21010293

is necessary to run the function *circrnaQuantication*. This function applies a two-step procedure to count the sequencing reads supporting a BS junction.

Practically, the first step of the function takes as input the sample reads, the set of sequences and a threshold N and it returns the corresponding set of reads which contains at least N sub-sequences of length k (called k-mer) shared with the set of pre-defined BS sequences. The k-mers are stored in RAM exploring an ad-hoc C++ hash table class implementation to optimize the trade-offs between the memory utilization and the execution time.

The second step takes as input the reads selected in the first step and directly align them against the pre-defined BS sequences. For this step, the Smith-Waterman algorithm provided by SIMD C++ library is used. The *circrnaQuantication* function can be applied by providing the input circRNAs BS junctions in fasta format, the RNA-Seq data to analyze in fastq format, and six parameters: the k-mer length, the number of threads, the dimension of the hash table, the dimension of the collision list, the number of k-mers that must be matched to the sequence and the number of perfect matches required to consider the sequence represented in the RNA-Seq data. The circRNAs BS count table obtained can be then analyzed by the *wrapperDeseq2* function.

Test of Docker4Circ for the analysis of circRNA expression in colorectal cancer cell lines. In the following section, we used Docker4Circ to identify the set of circRNAs expressed from RNA-Seq data of Normal Colon Mucosa (NCM) NCM460 cell line and Colorectal Cancer (CRC) SW620 and SW480 cell lines [28]. In detail, the Docker4Circ modules were used to predict, classify, annotate, and analyze the expression of the circRNAs in NCM and CRC cell lines. Moreover, the expression analysis (Module 4) was exploited to measure the expression of the detected circRNAs in an independent RNA-Seq experiment obtained from primary CRC tissues and adjacent normal tissue.

The prediction of circRNAs in NCM and CRC cell lines was performed using CIRI2 pipeline, i.e. in the *wrapperCiri* function. This function merges the FASTQC quality control of the reads, the read alignment with BWA and the CIRI2 prediction. As reported in Table 1, among the different samples the lowest number of circRNAs predicted by CIRI2 was 4,624, while the highest circRNAs predicted was 16,006 (average value = 9,474). The complete list of circRNAs associated with the number of BS-supporting reads for each sample is reported in Supplementary Table 1. The number of circRNAs decreases from NCM to CRC cell lines, as previously reported by the authors [28].

Dataset ID	Reads	circRNAs	AS events
NCM460_R1	66,144,999	14,003	1,482
NCM460_R2	70,945,094	16,006	1,790
NCM460_R3	73,804,226	12,413	1,078
SW480_R1	88,915,933	8,627	532
SW480_R2	97,303,573	5,688	335
SW480_R3	66,144,999	7,154	470
SW620_R1	91,406,400	1,0216	790
SW620_R2	67,013,355	4,624	214
SW620_R3	69,789,394	6,541	332
Average	76,829,774.78	9,474.67	780.33

Table 1. Table reporting the number of RNA-Seq paired reads analyzed, the number of detected circRNAs, and the number of AS events predicted by CIRI-AS

Using the *ciri2MergePredictions* function we merged the circRNAs predicted in each sample into a single circRNAs list. This list is composed of 7,086 out of 31,694 circRNAs characterized by more than two BS-supporting reads in at least two replicates and an average value of BS-supporting reads higher than 10 (Supplementary Table 2). 99.80% (n=7,072) of the circRNAs belonging to our

list was previously detected by Jiang et al [28]. Starting from the same datasets, the circRNAs prediction was also performed with the STARChip pipeline. The analysis predicted 2,933 circRNAs of which 94.7% overlapped with CIRI2 (Supplementary Table 3).

As previously observed by our group, circRNAs can be synthesized by complex splicing patterns involving exonic, intronic, and intergenic regions [10]. To better characterize the genomic regions involved in the back-splicing process of the 7,086 circRNAs predicted by CIRI2, we applied the function *circrnaClassification* that provides an accurate classification of each circRNA based on BS site location with respect to Ensembl transcript annotations. At the transcript level, the exons of 15,454 transcripts were associated with at least one circRNA, while 7,200 unique classifications were provided at the gene level. The discrepancy between the number of input circRNAs and the unique classification results is due to those circRNAs that can be attributed equally to two or more genes sharing the same exons. Supplementary Table 4 provides the classification results at the gene level with information of involved exons.

Given our circRNAs classification, we observed that exonic circRNAs were the class associated with the highest expression and the circRNAs predicted on the HIPK3 (exon 2), CAMSAP1 (exon 2-3), and ASXL1 (exon 2-4) were the most expressed circRNAs in both normal and cancer cell lines (Supplementary Table 4). These circRNAs were also identified by Jiang and coworkers as the most expressed circRNAs [28]. To further characterize the structural properties of the circRNAs in our list, we applied the sequence analysis module of Docker4Circ to identify AS events involving the exons composing the circRNAs. This analysis was performed with CIRI-AS algorithm implemented in the *ciri_as* function and generated an average of 780.3 AS events involving our circRNAs (Table 1 and Supplementary Table 5). Subsequently, to describe our circRNA set based on public database information, we applied the *circAnnotations* function of Docker4Circ. The overlap with the CircBase and TSCD databases highlighted 4,950 circRNAs (69.9%) annotated in CircBase while 540 (7.6%) and 508 (7.2%) circRNAs are annotated in the adult and fetal tissue section of TSCD database respectively (Supplementary Table 6). Interestingly, 83 circRNAs were detected in adult colon tissue samples.

Finally, using the functions implemented in the expression analysis module 4, it was possible to assess the differential expression level of our list of circRNAs among different experimental conditions or in data from independent RNA-Seq experiments. Then, we performed a differential expression analysis (*wrapperDeseq2* function) considering the number of BS-supporting reads measured in NCM and CRC cell lines. As reported in Supplementary Table 7, we identified 705, 655, and 430 circRNAs differentially expressed (adj. p-value < 0.001) between NCM460 and SW480 cell lines, NCM460 and SW620 cell lines, and SW480 and SW620 cell lines (Supplementary Table 7). Among them, 639 (90.64%), 613 (93.59%), and 352 (81.86%) were detected as differentially expressed also by Jiang and coworker [28]. Finally, 208 circRNAs were significantly differentially expressed in all the comparisons. Among them, the circRNAs mapped in EXOC6B (exons 1-3), DCBLD2 (exon 1-2), and ASAP1 (exon 2) were the most significant dysregulated circRNAs (Supplementary Table 7 and Supplementary Table 3).

2.2. Application of Docker4Circ to directly quantify circRNAs expression from CRC tissue RNA-Seq data

The expression analysis module was also used to quantify the expression of our circRNAs expression in an independent RNA-Seq dataset. Specifically, the expression of the 7,086 circRNAs identified using the cell lines data was quantified in an RNA-Seq dataset of primary CRC and paired NCM tissue data set (GSE104178) [29]. For this purpose, the sequences of the circRNAs back-splicing junctions were reconstructed using the Docker4Circ *circrnaBSJunctions* function. Using the hash table-based approach the reconstructed BS sequences were searched directly in total RNA-Seq reads bypassing the circRNAs prediction in each tissue sample. Using this quantification method, 1,758 circRNAs were associated with at least one read. The most expressed circRNAs in NCM samples was chr17_45043900_45047675 an intronic circRNA of RP11-156P1.2 gene, whereas a circRNA from the RPA3-AS1 gene (exon 2 and 3) was the most expressed circRNA in CRC samples.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 July 2019 doi:10.20944/preprints201907.0219.v1

Peer-reviewed version available at International Journal of Molecular Sciences 2020, 21, 293; doi:10.3390/ijms21010293

Then, the BS read count table was used as input of a differential expression analysis (*wrapperDeseq2* function) between the CRC and the paired NCM datasets. We identified six circRNAs differentially expressed between CRC and normal colonic mucosa samples (p-value < 0.01) (Figure 2c and Supplementary Table 8).

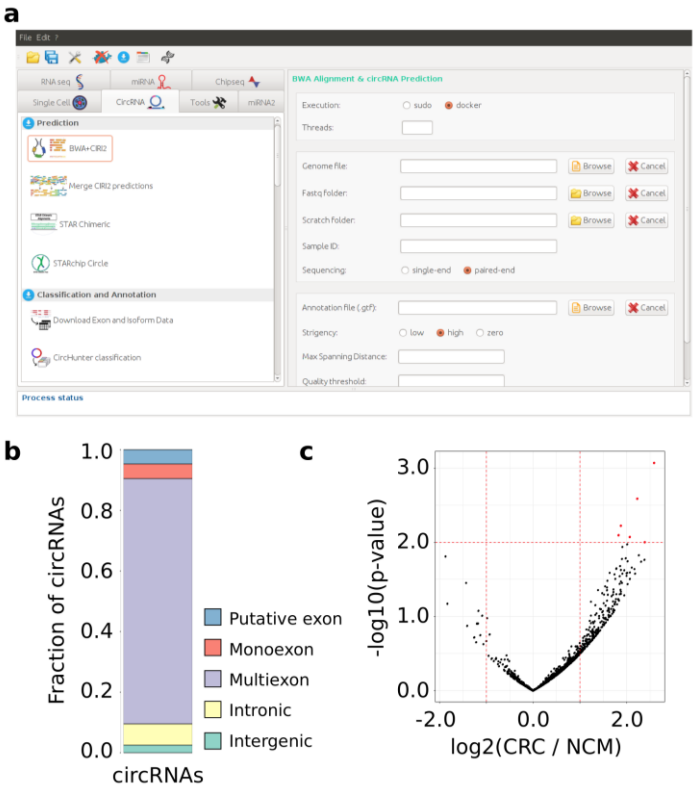

Figure 2. (**a**) The Docker4Circ Graphical User Interface. Each module implemented in the framework is accessible using the panel on the left, the right panel reports the fields and parameters of each function. (**b**) Bar plot reporting the Docker4Circ classification of circRNAs identified in the analysis of RNA-Seq datasets from CRC cell lines. **(c)**Volcano plot reporting the -log10 p-value and the log2 expression fold change (red dashed lines) computed between Docker4Circ counts of BS supporting reads from RNA-Seq datasets of NCM and CRC tissue samples.

3. Discussion

CircRNAs are widely expressed in both cancerous and normal tissues [30,31] and an increased number of sequencing experiments is becoming accessible to explore circRNAs expression in a specific biological context. To deal with the increased number of computational resources and public datasets available for circRNAs analysis, in this work, we proposed Docker4Circ as a user-friendly framework to guarantee reproducible analysis of circRNAs data.

The framework was designed for users with different levels of expertise in computational analysis. Specifically, a Java graphic user interface was designed to provide a user-friendly framework for the analysis or each Docker4Circ R-function can be applied by the more expert users through the command line user interface. In both cases, the user can set its custom analysis pipeline by setting each function properly. Indeed, despite here we described an analysis pipeline composed of the sequential application of the four modules, the modular architecture of the framework guarantees no constraints on the structure of the pipeline selected by the user.

We tested Docker4Circ to reproduce analyses performed by Jiang and colleagues [28] on circRNAs expression in CRC cell lines showing that our framework is able to extensively reproduce their results. Furthermore, we added novel evidence on these circRNAs by performing a

quantification and differential analysis of their expression level considering RNA-Seq performed on CRC and adjacent colonic tissues. This analysis showed two circRNAs differentially expressed both in tissue and cell lines models (Supplementary Table 8B). These circRNAs were annotated, respectively to the HNRNPC (exon 1-2) and the PSMA3 genes.

As reported in Supplementary Table 9 all the modules and functions implemented in Docker4Circ can be run in a limited amount of time and the overall running time of the workflow was around seven hours per dataset or twelve hours if the STARchip prediction is performed. The workflow was performed on an Intel NUC6I7KYK mini-PC with 8 threads confirming that all the Docker4Circ functions can be executed on a standard workstation because the only requirement is 32 Gb of RAM available if the STAR Chimeric analysis is performed. Indeed only the circRNAs prediction or the hash table-based quantification of circRNA expression benefited from the multithreading while all the other analyses require a minimal amount of computational resources.

The most important aspect of our framework is that it guarantees the computational reproducibility of the analysis which is obtained by embedding each analysis step into a specific docker image. Following the good-practice bioinformatics roles proposed by Sandve and colleagues [18], our framework provided analyses whose results can be fully tracked on how they were produced by recording each analysis steps and version of tools applied without any data manipulation step.

In conclusion, Docker4Circ provides an efficient analysis framework to identify and characterize the circRNAs on a large number of sequencing experiments. The usage of Docker images ensures a reproducible circRNAs analyses to easily harmonize and combine the study of these molecules in different experimental and biological contexts.

4. Materials and Methods

The detailed analysis protocol that was followed for the analyses reported in this manuscript was released on the protocol.io portal at dx.doi.org/10.17504/protocols.io.zrrf556. The specific use of each function and associated parameters are reported in Supplementary Materials of the manuscript.

CircRNAs prediction

To test Docker4Circ, RNase-R RNA-Seq datasets from PRJNA393626 were selected. These data consist of a triplicate paired-end total and RNase-R treated RNA-Seq performed on NCM460 (normal colon cells), SW480 (primary CRC cells), and SW620 cell lines (metastatic CRC cells).

CircRNAs prediction was performed using the *wrapperCiri* function of Docker4Circ with the following parameters: max.span = 200000, stringency.value = "high", and quality.threshold = 10. Using this function, reads alignment was performed using the mem mode of BWA v.0.6.1 in default settings. CircRNAs prediction was performed using CIRI2 algorithm v.2.06 [21]. Gencode v28 was used as reference transcriptome while Ensembl hg19 (GRCh37) as Human reference genome. CircRNAs predicted in at least two out of the three biological replicates in each condition and associated with an average number of BS-supporting reads > 10 were selected. This prediction overlap was performed with the *ciri2MergePredictions* function of Docker4Circ using the options min_reads = 2, min_reps = 2, and min_avg = 10. The list of circRNAs was overlapped with those predicted in [28] by converting the circRNA genomic coordinates from hg19 to hg38 human genome assembly using LiftOver algorithm [25].

For the circRNAs prediction with the STARChip pipeline, the reference genome was indexed using the function *rsemstarIndex* and *starChipIndex*. Subsequently chimeric read alignments for each dataset were detected using the function *starChimeric* with parameters chimSegmentMin = 20 and chimJunctionOverhangMin = 15. Finally, the function *starchipCircle* was applied to predict the circRNAs using the STAR Chimeric alignments. The function was applied with parameters reads.cutoff = 1, min.subject.limit = 2, do.splice = "True", cpm.cutoff = 0, subjectCPM.cutoff = 0, annotation = "true". The *ciri2MergePredictions* function was used to filter the circRNAs read count

table using the same parameters exploited during the CIRI2 analysis. The overlap between CIRI2 and STARChip circRNA predictions was performed by considering their genomic coordinates.

CircRNAs classification and annotations.

The circRNAs classification was performed using *circClassification* function (with option assembly="hg19") applied on the list of circRNAs predicted by CIRI2 and on the reference hg19 transcript annotations from Ensembl (Ensembl v93) downloaded using the *circrnaPrepareFiles* function of Docker4Circ (with option assembly = "hg19").

The circRNA annotation was performed using the *circAnnotations* function with option genome.version = "hg19".

CircRNAs sequence analysis.

The 70 bp sequences representing the reconstructed circRNA BS junctions were obtained using the function *circrnaBSJunctions* of Docker4Circ. The resulting Fasta file was used for the quantification analysis. Prediction of AS events involving the circRNAs was performed using the *ciri_as* function on each list of circRNAs predicted by CIRI2.

Quantification of circRNAs in RNA-Seq datasets.

The *circrnaQuantication* function of Docker4Circ was applied using six RNA-Seq datasets from GSE104178 [29]. These data consist of total RNA-Seq performed on three matched pairs of CRC samples and matched Normal Colonic Mucosa (NCM). For each dataset, the two mates of a paired-end read were joined using the cat command. The quantification analysis was performed selecting a k-mer length equal to 21 based on the read length (75 bp); a minimum number of matching k-mer equal to 17, and a minimum number of perfect matches equal to 30. The maximum number of the element stored in the hash table was set to 1,000,003. DESeq2 v1.20.0 [27] was applied for BS read count normalization and differential expression analysis. The algorithm was applied in default settings. The analysis was performed using the *wrapperDeseq2* function of Docker4Circ on the result of the *mergeData* function which was used to join different circRNA count tables with the covariates indicating the samples classes.

Availability of source code and requirements.

The Docker4Circ R functions were integrated into the Docker4Seq R package available at https://github.com/kendomaniac/docker4seq.

The Java GUI can be downloaded from https://github.com/mbeccuti/4SeqGUI.

The analysis is independent of the operating system applied, while Docker software is required. All Docker4Circ modules are already integrated into a Docker image. Each docker image tag is then created following rule defined by RBP: Docker image tags are labelled with the extension YYYY.NN, where YYYY is the year of insertion in the stable version and NN a progressive number. YYYY changes only if any update on the program(s), implemented in the docker image, is done. This because any such updates will affect the reproducibility of the workflow. Previous version(s) will be also available in the repository. NN refers to changes in the docker image, which do not affect the reproducibility of the workflow.

Docker4Circ running time estimation.

The running time of each analysis was computed by considering the execution using 8 threads of an Intel NUC6I7KYK mini-PC [20]. Multi-threading was applied only for the execution of the functions *ciri2*, *circrnaQuantification*, *starChimeric*, and *starchipCircle*.

Supplementary Materials: Supplementary materials can be found at www.mdpi.com/xxx/s1.

Author Contributions: N.L, G.F., and C.D.I. implemented Docker4Circ functions, the GUI and the Docker containers. G.F. conducted empirical experiments. G.F., V.M., F.C, and M.B drafted the manuscript. G.F., L.C.T., and V.M. designed the case study. R.A.C, M.B., M.D.B. coordinated the empirical study. All authors edited the manuscript.

Funding: This work was supported by Associazione Italiana per la Ricerca sul Cancro [Grant IG 15600 to MDB]; by Fondazione CRT [grant 2014.1854 to MDB and 2017.2025 to FC]; by University of Torino [2016 Local Research funding to MDB]; Consiglio Nazionale delle Ricerche [Flagship projects EPIGEN to RAC]; Fondazione Umberto Veronesi [Postdoctoral Fellowship to VM]. GF was supported by a FIRC-AIRC fellowship for Italy.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

circRNA	circular RNAs
CRC	colorectal cancer
NCM	normal colonic mucosa
AS	alternative splicing
BS	back-splicing

References

1. Salzman, J. Circular RNA Expression: Its Potential Regulation and Function. *Trends Genet.* **2016**, *32*, 309–316.
2. Szabo, L.; Salzman, J. Detecting circular RNAs: bioinformatic and experimental challenges. *Nat. Rev. Genet.* **2016**, *17*, 679–692.
3. Jeck, W.R.; Sorrentino, J.A.; Wang, K.; Slevin, M.K.; Burd, C.E.; Liu, J.; Marzluff, W.F.; Sharpless, N.E. Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA* **2013**, *19*, 141–157.
4. Gao, Y.; Zhao, F. Computational Strategies for Exploring Circular RNAs. *Trends in Genetics* **2018**, *34*, 389–400.
5. Metge, F.; Czaja-Hasse, L.F.; Reinhardt, R.; Dieterich, C. FUCHS-towards full circular RNA characterization using RNAseq. *PeerJ* **2017**, *5*, e2934.
6. Gao, Y.; Wang, J.; Zheng, Y.; Zhang, J.; Chen, S.; Zhao, F. Comprehensive identification of internal structure and alternative splicing events in circular RNAs. *Nat. Commun.* **2016**, *7*, 12060.
7. Li, M.; Xie, X.; Zhou, J.; Sheng, M.; Yin, X.; Ko, E.-A.; Zhou, T.; Gu, W. Quantifying circular RNA expression from RNA-seq data using model-based framework. *Bioinformatics* **2017**, *33*, 2131–2139.
8. Cheng, J.; Metge, F.; Dieterich, C. Specific identification and quantification of circular RNAs from sequencing data. *Bioinformatics* **2016**, *32*, 1094–1096.
9. Feng, J.; Xiang, Y.; Xia, S.; Liu, H.; Wang, J.; Ozguc, F.M.; Lei, L.; Kong, R.; Diao, L.; He, C.; et al. CircView: a visualization and exploration tool for circular RNAs. *Brief. Bioinform.* **2018**, *19*, 1310–1316.
10. Coscujuela Tarrero, L.; Ferrero, G.; Miano, V.; De Intinis, C.; Ricci, L.; Arigoni, M.; Riccardo, F.; Annaratone, L.; Castellano, I.; Calogero, R.A.; et al. Luminal breast cancer-specific circular RNAs uncovered by a novel tool for data analysis. *Oncotarget* **2018**, *9*, 14580–14596.
11. Glažar, P.; Papavasileiou, P.; Rajewsky, N. circBase: a database for circular RNAs. *RNA* **2014**, *20*, 1666–1670.
12. Xia, S.; Feng, J.; Lei, L.; Hu, J.; Xia, L.; Wang, J.; Xiang, Y.; Liu, L.; Zhong, S.; Han, L.; et al. Comprehensive characterization of tissue-specific circular RNAs in the human and mouse genomes. *Brief. Bioinform.* **2017**, *18*, 984–992.
13. Chen, X.; Han, P.; Zhou, T.; Guo, X.; Song, X.; Li, Y. circRNADb: A comprehensive database for human circular RNAs with protein-coding annotations. *Sci. Rep.* **2016**, *6*, 34985.
14. Ghosal, S.; Das, S.; Sen, R.; Basak, P.; Chakrabarti, J. Circ2Traits: a comprehensive database for circular RNA potentially associated with disease and traits. *Front. Genet.* **2013**, *4*, 283.
15. Gaffo, E.; Bonizzato, A.; Kronnie, G.T.; Bortoluzzi, S. CirComPara: A Multi-Method Comparative Bioinformatics Pipeline to Detect and Study circRNAs from RNA-seq Data. *Noncoding RNA* **2017**, *3*.
16. Humphreys, D.T.; Fossat, N.; Tam, P.P.L.; Ho, J.W.K. Ularcirc: Visualisation and enhanced analysis of circular RNAs via back and canonical forward splicing.
17. Jakobi, T.; Uvarovskii, A.; Dieterich, C. circtools - a one-stop software solution for circular RNA research. *Bioinformatics* **2018**.
18. Sandve, G.K.; Nekrutenko, A.; Taylor, J.; Hovig, E. Ten simple rules for reproducible computational research. *PLoS Comput. Biol.* **2013**, *9*, e1003285.
19. Kulkarni, N.; Alessandrì, L.; Panero, R.; Arigoni, M.; Olivero, M.; Ferrero, G.; Cordero, F.; Beccuti, M.; Calogero, R.A. Reproducible bioinformatics project: a community for reproducible bioinformatics analysis

pipelines. *BMC Bioinformatics* **2018**, *19*, 349.
20. Beccuti, M.; Cordero, F.; Arigoni, M.; Panero, R.; Amparore, E.G.; Donatelli, S.; Calogero, R.A. SeqBox: RNAseq/ChIPseq reproducible analysis on a consumer game computer. *Bioinformatics* **2018**, *34*, 871–872.
21. Gao, Y.; Zhang, J.; Zhao, F. Circular RNA identification based on multiple seed matching. *Brief. Bioinform.* **2018**, *19*, 803–810.
22. Akers, N.K.; Schadt, E.E.; Losic, B. STAR Chimeric Post for rapid detection of circular RNA and fusion transcripts. *Bioinformatics* **2018**, *34*, 2364–2370.
23. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760.
24. Durinck, S.; Spellman, P.T.; Birney, E.; Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols* 2009, *4*, 1184–1191.
25. Hinrichs, A.S.; Karolchik, D.; Baertsch, R.; Barber, G.P.; Bejerano, G.; Clawson, H.; Diekhans, M.; Furey, T.S.; Harte, R.A.; Hsu, F.; et al. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* **2006**, *34*, D590–8.
26. Lawrence, M.; Huber, W.; Pagès, H.; Aboyoun, P.; Carlson, M.; Gentleman, R.; Morgan, M.T.; Carey, V.J. Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **2013**, *9*, e1003118.
27. Love, M.I.; Huber, W.; Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **2014**, *15*, 550.
28. Jiang, W.; Zhang, X.; Chu, Q.; Lu, S.; Zhou, L.; Lu, X.; Liu, C.; Mao, L.; Ye, C.; Timko, M.P.; et al. The Circular RNA Profiles of Colorectal Tumor Metastatic Cells. *Front. Genet.* **2018**, *9*, 34.
29. Yamada, A.; Yu, P.; Lin, W.; Okugawa, Y.; Boland, C.R.; Goel, A. A RNA-Sequencing approach for the identification of novel long non-coding RNA biomarkers in colorectal cancer. *Sci. Rep.* **2018**, *8*, 575.
30. Ji, P.; Wu, W.; Chen, S.; Zheng, Y.; Zhou, L.; Zhang, J.; Cheng, H.; Yan, J.; Zhang, S.; Yang, P.; et al. Expanded Expression Landscape and Prioritization of Circular RNAs in Mammals. *Cell Rep.* **2019**, *26*, 3444–3460.e5.
31. Vo, J.N.; Cieslik, M.; Zhang, Y.; Shukla, S.; Xiao, L.; Zhang, Y.; Wu, Y.-M.; Dhanasekaran, S.M.; Engelke, C.G.; Cao, X.; et al. The Landscape of Circular RNA in Cancer. *Cell* **2019**, *176*, 869–881.e13.