UNIVERSITY OF CATANIA

PhD IN COMPLEX SYSTEMS FOR PHYSICAL SOCIO-ECONOMIC AND LIFE SCIENCES

DEPARTMENT OF PHYSICS AND ASTRONOMY "ETTORE MAJORANA"

DEPARTMENT OF CLINICAL AND EXPERIMENTAL MEDICINE, BIOINFORMATICS SECTION

ALESSANDRO LA FERLITA

User-friendly software for coding and non-coding RNA-Seq data analysis: from raw sequencing data to biological pathway analysis

---

Ph.D. Thesis

***Tutor*:**
Prof. Alfredo Pulvirenti

***Co-Tutors*:**
Prof. Alfredo Ferro
Dr. Salvatore Alaimo

2021

Contents

Abstract

RNA-Seq is a well-established technology extensively used for transcriptome profiling, allowing the analysis of coding and non-coding RNA molecules. However, this technology produces a huge amount of data that require more sophisticated computational approaches for their analysis than other traditional technologies such as Real-Time PCR or microarrays, strongly discouraging non-expert users. For this reason, dozens of pipelines have been deployed for the analysis of RNA-Seq data. Although interesting, these present several limitations and their usage require a technical background, which may be uncommon in small research laboratories. Therefore, the application of these technologies in such contexts is still limited and, indeed, causes a clear bottleneck in knowledge advance. Motivated by these considerations, in this PhD thesis I present *RNAdetector*, a new free stand-alone, cross-platform, and user-friendly RNA-Seq data analysis software that can be used completely offline by means of an easy-to-use Graphical User Interface (GUI) allowing the analysis of coding and ncRNAs from RNA-Seq datasets of any sequenced biological species.

1. Introduction

1.1 Next Generation Sequencing (NGS)

Starting with the discovery of the DNA structure, great advancements have been made in understanding the complexity and diversity of genomes [1]. For sure, one of the most important milestones in human genetics was when after more than one decade of intense work we obtained the complete sequence of the human genome in 2003 with the conclusion of the human genome project. Since then, many new versions of the human genome have been released and this effort pushed for the extraordinary progress which has been made in genome sequencing technologies. These new technologies, which have been developed after the first generation of sequencing (Sanger sequencing) used for the determination of the human genome, were called Next Generation Sequencing (NGS) [1]. Over the past decade, NGS technologies have continued to evolve. In fact, today they allow sequencing of entire genomes in a few days, instead of 10 years as it was originally needed to complete the first version of the human genome, and they have brought the cost of sequencing a human genome down to around US$1,000 (as reported by Veritas Genomics) [2]. In addition, NGS technologies are now also used in clinics to detect gene mutations or polymorphisms (e.g., CNV, SNPs, INDEL, STR) potentially associated with disease predisposition and support diagnosis confirmation [3,4]. Although exciting, these advancements are not without limitations. NGS platforms provide vast quantities of data, but the associated error rates (~0.1–15%) are higher and the read lengths generally shorter (35–700 bp for short-read approaches) than those of traditional Sanger sequencing platforms, requiring careful examination of the results [1]. Although long-read sequencing

overcomes some limitations of other NGS approaches, it remains considerably more expensive and has lower throughput than other platforms, limiting the widespread adoption of this technology in favor of less-expensive approaches [1].

1.1.1 Short reads NGS

Short-read sequencing approaches are classified in two broad categories: sequencing by ligation (SBL) and sequencing by synthesis (SBS) [1]. In SBL approaches, a probe sequence, which is bound to a fluorophore, hybridizes to a DNA fragment and is ligated to an adjacent oligonucleotide for imaging. The emission spectrum of the fluorophore indicates the identity of the base or bases complementary to specific positions within the probe [1]. In SBS approaches, a polymerase is used and a signal, such as a fluorophore or a change in ionic concentration, identifies the incorporation of a nucleotide into an elongating strand [1]. In most SBL and SBS approaches, DNA is clonally amplified on a solid surface. Having many thousands of identical copies of a DNA fragment in a defined area ensures that the signal can be distinguished from background noise [1]. Massive parallelization is also allowed by millions of individual SBL or SBS reaction centers, each with its own clonal DNA template. A sequencing platform can collect information from many millions of reaction centers simultaneously, allowing sequencing of many millions of DNA molecules in parallel. In the last decade, several NGS platforms have been released. Below follows a brief description of the most used short-read NGS platforms.

1.1.1.1 Pyrosequencing

Pyrosequencing is an SBS based method for DNA sequencing described for the first time in 1993 [5] and subsequently optimized till when the final variant was finally presented in 2005 by the company 454 Life Sciences becoming the first NGS

instrument [6]. The methodology consists of a DNA library preparation first, and second in the sequencing reaction. The DNA is fragmented in several random fragments, denatured in single-strand DNA (ssDNA), and finally ligated to magnetic beads. There, ssDNA molecules are amplified by emulsion PCR (emPCR) in order to have millions of clones of the same DNA sequence in a single bead [7]. This amplification is important in order to amplify the signal during the sequence reaction. Then, the ssDNA template is hybridized to a sequencing primer and incubated with several enzymes such as DNA polymerase, ATP sulfurylase, luciferase, and apyrase, and with the substrates adenosine 5´ phosphosulfate (APS) and luciferin. The addition of one of the four deoxynucleotide triphosphates (dNTPs) (dATPαS, which is not a substrate for luciferase, is added instead of dATP to avoid noise) initiates the second step. DNA polymerase incorporates the correct, complementary dNTPs onto the template. This incorporation releases pyrophosphate (PPi). ATP sulfurylase converts PPi to ATP in the presence of APS. This ATP acts as a substrate for the luciferase-mediated conversion of luciferin to oxyluciferin that generates visible light in amounts that are proportional to the amount of incorporated nucleotides. After that, the light produced in the luciferase-catalyzed reaction is detected by a camera and analyzed in a program. Unincorporated nucleotides and ATP are degraded by the apyrase, and the reaction can restart with another nucleotide (Fig. 1) (QIAGEN "Pyrosequencing Technology and Platform Overview").

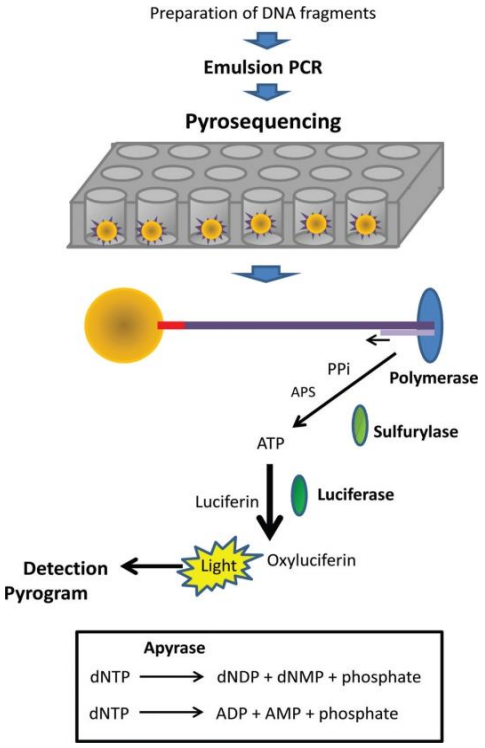

Fig 1. Schematic picture of pyrosequencing workflow (Source Siqueira JF. et al. Journal of Oral Microbiology. 2012).

After several years, in 2013 Roche (which acquired the company 454 Life Sciences) announced to stop the production of the pyrosequencing platform because it was not competitive with the other newer NGS platforms already available on the market.

1.1.1.2 Ion Torrent

Ion Torrent is the evolution of pyrosequencing released in February 2010 by Ion Torrent Systems Inc [8]. This methodology uses an approach similar to pyrosequencing to detect the DNA sequence. However, rather than using an enzymatic cascade to generate a signal, the Ion Torrent platform detects the H+ ions that are released after that each dNTP is incorporated. The resulting change in pH is detected by an integrated complementary metal oxide-semiconductor (CMOS) and an ion-sensitive field-effect transistor (ISFET) [9]. In more details, the incorporation of a dNTP into a growing DNA strand involves the formation of a covalent bond and the release of PPi and a H+. A dNTP will only be incorporated if it is complementary to the leading unpaired template

nucleotide. Ion semiconductor sequencing exploits these facts by determining if a H+ is released upon providing a single species of dNTP to the reaction. Microwells on a semiconductor chip, which each contain many copies of one ssDNA molecule to be sequenced and DNA polymerase, are sequentially flooded with unmodified dNTP [8]. If an introduced dNTP is complementary to the next unpaired nucleotide on the template strand it is incorporated into the growing complementary strand by the DNA polymerase. If the introduced dNTP is not complementary there is no incorporation and no biochemical reaction. The H+ that is released in the reaction changes the pH of the solution, which is detected by an ISFET. The unattached dNTP molecules are washed out before the next cycle when a different dNTP species is introduced [8] (Fig. 2). Each chip contains an array of microwells with corresponding ISFET detectors. Each released H+ then triggers the ISFET ion sensor. The series of electrical pulses transmitted from the chip to a computer is translated into a DNA sequence, with no intermediate signal conversion required [8]. Because nucleotide incorporation events are measured directly by electronics, the use of labeled nucleotides and optical measurements are avoided.

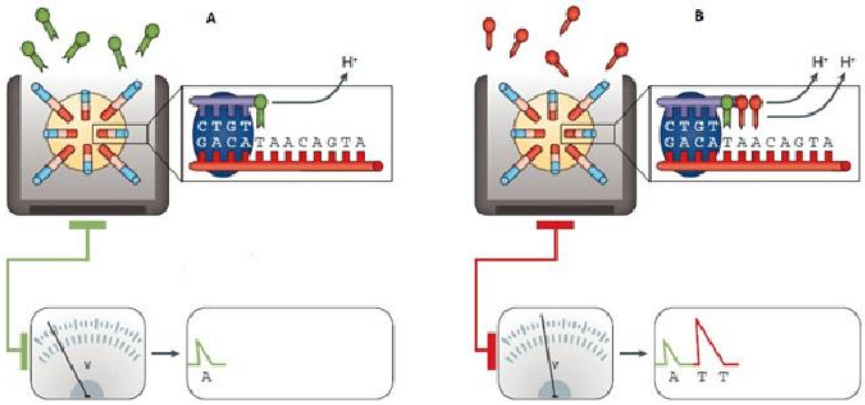

Fig 2. Detection of the incorporated nucleotides with Ion Torrent technologies (Source Kchouk M. et al. Biology and Medicine. 2017)

The major benefits of Ion Torrent are that it is faster and cheaper than other NGS platforms. This has been enabled by avoiding modified nucleotides and optical

measurements [10]. However, the major limitation is that the pH change detected by the sensor is imperfectly proportional to the number of nucleotides detected, allowing for limited accuracy in measuring homopolymer lengths.

1.1.1.3 ABI Solid Sequencing

SOLiD (Sequencing by Oligonucleotide Ligation and Detection) is an NGS platform with an SBL approach developed by Life Technologies and available since 2006. The method consists, first of all, of a DNA library preparation where the DNA is fragmented in several random fragments which are then attached to magnetic beads with a universal P1 adapter sequence attached on its surface. Universal adapter sequences are needed in order that the starting sequence of every fragment is known and identical. After the ligation of the DNA fragments with the universal adaptors present on the surface of the magnetic beads, the emPCR can start [7]. emPCR takes place in microreactors containing all the necessary reagents for PCR. As a result of the emPCR reaction, the DNA template is amplified and, therefore, million clonal DNA fragments are immobilized on a single magnetic bead. The resulting PCR products attached to the beads are then covalently bound to a glass slide where the sequencing reaction can start. First, primers hybridize to the P1 adapter sequence within the library template. Second, a set of four fluorescently labeled di-base probes compete for ligation to the sequencing primer. In fact, SOLiD platforms utilize two-base-encoded probes, in which each fluorometric signal represents a dinucleotide, and therefore, the raw output is not directly associated with the incorporation of a known nucleotide [11]. Because the 16 possible di-nucleotide combinations cannot be individually associated with specific fluorophores, four fluorescent signals are used, each representing a subset of four dinucleotide combinations [11]. The specificity of the di-base probe is achieved by interrogating every 1st and 2nd base in each ligation reaction. Multiple

cycles of ligation, detection, and cleavage are performed with the number of cycles determining the eventual read length. Following a series of ligation cycles, the extension product is removed and the template is reset with a primer complementary to the n-1 position for the second round of ligation cycles. Five rounds of primer reset are completed for each sequence tag. Through the primer reset process, each base is interrogated in two independent ligation reactions by two different primers [11] (Fig. 3).

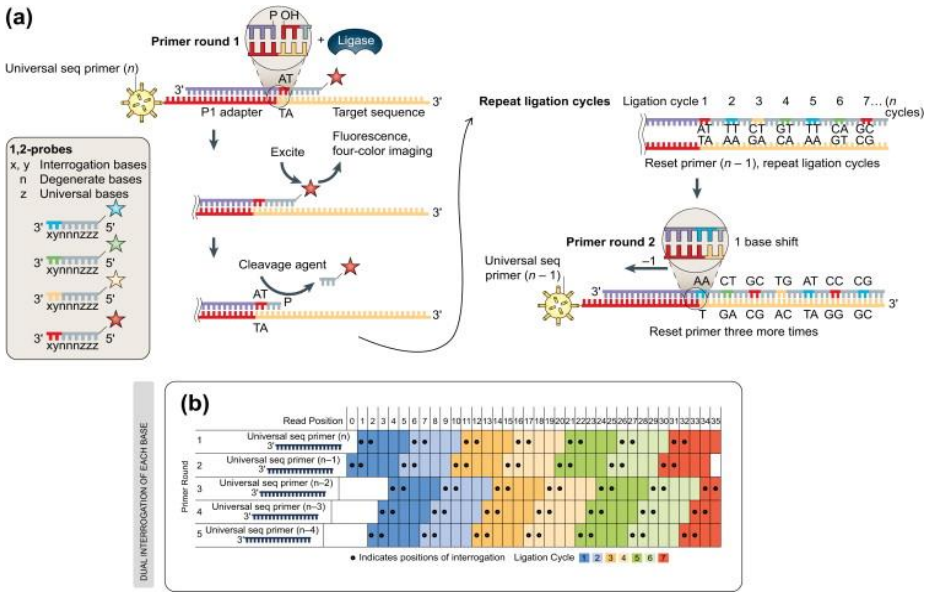

Fig 3. Schematic representation of SOLiD sequencing (Source Pereira DM. et al. Principles of Translational Science in Medicine (Second Edition). 2015)

Due to the two base encoding system, this technology offers about 99.94% of accuracy [12]. Although its high accuracy, the very short read lengths (75 bp) is limiting its use for genome assembly and structural variant detection applications.

1.1.1.4 Illumina

The Illumina system is definitely the most used NGS platform today. The secret of its success is due first to the bigger Illumina's suite of instruments, which range from the short-read sequencer for small low-throughput benchtop units to large ultra-high-throughput instruments dedicated to population-level whole-genome sequencing (WGS), and second to the better balance between economic and efficiency. Illumina

sequencing technology works in three basic steps: amplification, sequencing, and analysis. The process begins with purified DNA. The DNA gets cutted into smaller pieces and loaded onto a specialized chip where amplification and sequencing will take place. On the surface of this chip there are thousands of oligonucleotides which are anchored to the chip and able to grab DNA fragments that have complementary sequences. Once the fragments have attached, a phase called cluster generation begins. This step makes about a thousand copies of each DNA fragment. Next, primers and modified nucleotides enter the chip. These nucleotides have reversible 3' blockers that force the polymerase to add only one nucleotide at a time as well as fluorescent tags. After each round of synthesis, a camera takes a picture of the chip. A computer determines what base was added by the wavelength of the fluorescent tag and records it for every spot on the chip. After each round, non-incorporated molecules are washed away. A chemical deblocking step is then used to remove the 3' terminal blocking group and the dye in a single step. The process continues until the full DNA molecule is sequenced [13]. With this technology, thousands of places throughout the genome are sequenced at once via massively parallel sequencing (Fig. 4).

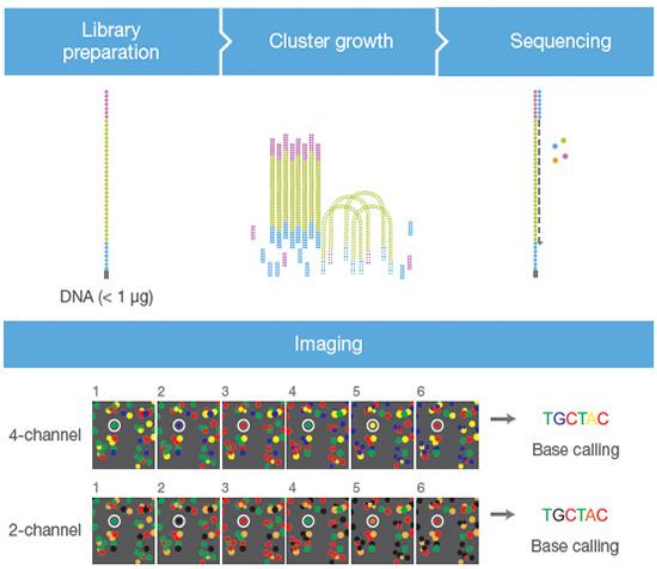

Fig 4. Schematic representation of Illumina sequencing (Source Illumina website)

1.1.2 Long reads NGS

It is widely known that genomes are highly complex and contain many long repetitive elements, and structural variations that are relevant to evolution, adaptation, and disease [14–16]. However, many of these complex elements are so long that short-read NGS technologies are not able to resolve them. Long-read sequencing tries to solve this problem by producing reads of several kilobases, allowing for the identification of large genomic features. In addition, long reads can also be useful for transcriptome research. Indeed, they are capable of sequencing entire mRNA transcripts, allowing researchers to identify exon junctions and identify alternative splicing isoforms [1]. Currently, there are two main types of long-read technologies: (1) single-molecule real-time (SMRT) sequencing approaches and (2) synthetic approaches. The SMRT approaches differ from short-read approaches because they do not rely on clonal amplified DNA fragments to generate a detectable signal, and they do not require chemical cycling for each dNTP added. On the other hand, the synthetic approaches do not generate actual long-reads; rather, they are an approach to library preparation that leverages barcodes to allow computational assembly of a larger fragment [1]. However, at this moment, the most widely used long-read platform is the SMRT sequencing approach. Below follows a brief description of the main SMRT sequencing platforms.

1.1.2.1 Pacific Biosciences (PacBio)

Pacific Biosciences is an American biotechnology company founded in 2004 that develops systems for DNA sequencing. Their first product, the PacBio RS, was commercially released in 2011. After that, a subsequent version called the PacBio RS II was released in 2013 and during these years several optimized versions of this

instrument were released. The instrument uses a special flow cell with many thousands of individual picolitre wells with transparent bottoms called zero-mode waveguides (ZMW) [17]. On the contrary of short-read SBS technologies that bind the DNA and allow the polymerase to travel along with the DNA template, PacBio fixes the polymerase to the bottom of the well and allows the DNA strand to progress through the ZMW. dNTP incorporation on each single-molecule template per well is continuously visualized with a laser and camera system that records the color and duration of emitted light as the labeled nucleotide momentarily pauses during incorporation at the bottom of the ZMW. The polymerase cleaves the dNTP-bound fluorophore during incorporation, allowing it to diffuse away from the sensor area before the next labeled dNTP is incorporated [1,18,19]. This platform also uses a unique circular template that allows each template to be sequenced multiple times as the polymerase repeatedly traverses the circular molecule (Fig. 5). Although it is difficult for DNA templates longer than ~3 kb to be sequenced multiple times, shorter DNA templates can be sequenced many times [1,18,19].

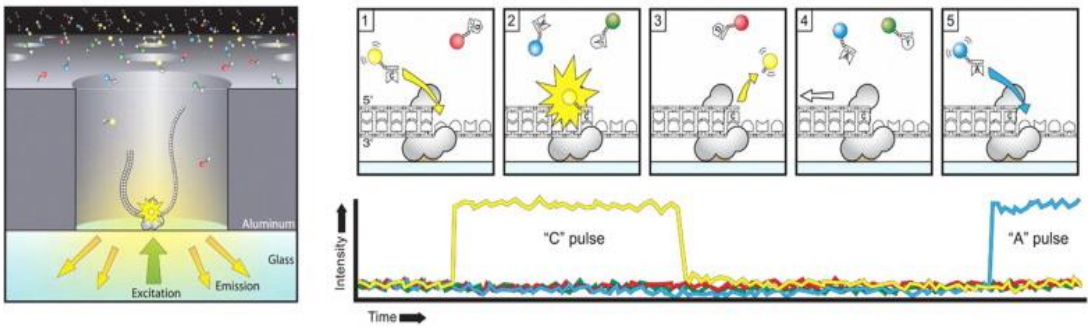

Fig 5. Schematic representation of PacBio sequencing (Source Rhoads A and Au KF. Genomics, Proteomics & Bioinformatics. 2015)

1.1.2.2 Oxford Nanopore Technologies (ONT)

Oxford Nanopore Technologies is a UK company that is developing and selling nanopore sequencing products for the sequencing of single DNA molecules [20,21].

Unlike other platforms, nanopore sequencers do not monitor incorporations or hybridizations of nucleotides guided by a template DNA strand. Whereas other platforms use a secondary signal, nanopore sequencers directly detect the DNA composition of a native ssDNA molecule. To carry out sequencing, DNA is passed through a protein pore as the current is passed through the pore [22]. As the DNA translocates through the action of a secondary motor protein, a voltage blockade occurs that modulates the current passing through the pore. The temporal tracing of these charges is called squiggle space, and shifts in voltage are characteristic of the particular DNA sequence in the pore, which can then be interpreted as a k-mer. Rather than having 1–4 possible signals, the instrument has more than 1,000 — one for each possible k-mer, especially when modified bases present on native DNA are taken into account [22] (Fig. 6).

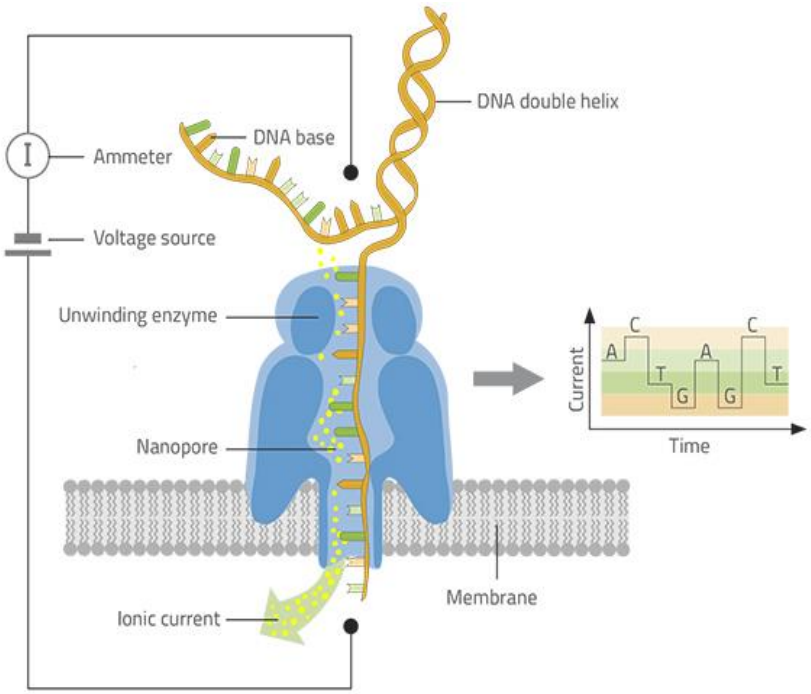

Fig 6. Schematic representation of ONT sequencing (Source Göpfrich K. Science in School. 2018)

1.2 RNA sequencing (RNA-Seq)

1.2.1 Transcriptome

Although NGS technologies were primarily developed for genomic analysis, during the last years, they have been always more used also for transcriptome analysis. By the term transcriptome, we mean the complete set of transcripts present in a cell or tissue for a specific developmental stage or physiological condition [23]. The transcriptome includes all RNA species from RNA messengers (mRNAs) to non-coding RNAs (ncRNAs). ncRNAs are RNA molecules that do not encode for proteins but represent a considerable amount of the transcriptome [24]. They are involved in many aspects of cell physiology and regulate a broad spectrum of cellular processes, controlling gene expression, and contributing to genome organization and stability [24]. ncRNAs can be classified according to their size in small ncRNAs (< 200 nucleotides) and long ncRNAs or lncRNAs (≥ 200 nucleotides) [24]. Alternatively, they can also be classified according to their function in housekeeping and regulatory ncRNAs [24] (Fig. 7). Housekeeping ncRNAs include ribosomal RNA (rRNA), transfer RNA (tRNA), small nuclear RNA (snRNA), and small nucleolar RNA (snoRNA) (Fig. 7). These ncRNAs are expressed in all cell types and they carry out essential functions in eukaryotic cells [24]. On the other hand, regulatory ncRNAs include several classes of small and long molecules, such as microRNAs (miRNAs), small interfering RNAs (siRNAs), Piwi-associated RNAs (piRNAs), long non-coding RNAs (lncRNAs), circular RNAs (circRNAs), and tRNA-derived ncRNAs (Fig. 7).

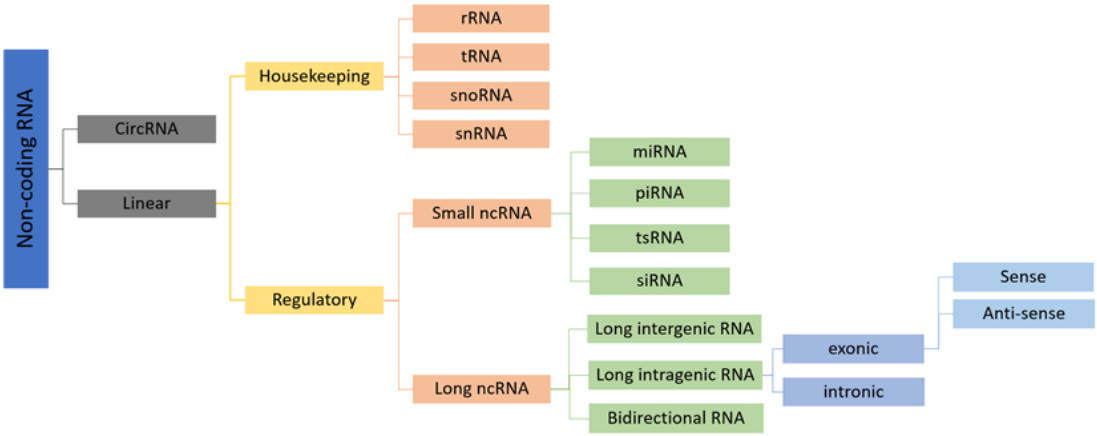

Fig 7. Classification of non-coding RNAs

This latter group represents a novel class of small regulatory ncRNAs, which derive from pre-tRNA and tRNA processing [25], and they have been reported to have crucial roles in different biological processes, such as ribosome biogenesis, retrotransposition, virus infections, apoptosis, and cancer pathogenesis [26–33]. In the last few years, several kinds of tRNA-derived ncRNAs have been discovered. However, a unique classification is still missing. A common grouping of such molecules is based on the location they originate from within the tRNA gene. tRNA-derived ncRNAs can therefore be divided into two main classes: (i) tRNA-derived small RNAs (tsRNAs), which derive from pre-tRNA; (ii) stress-induced tRNA fragments (tiRNAs), together with tRNA-derived fragments (tRFs), which derive from mature tRNA [25] (Fig. 8). tsRNA are produced inside the nucleus and result from the cleavage of the pre-tRNAs 3’ trailer sequence by RNases Z. They usually begin after the 3’-end of mature tRNAs and are characterized by a polyuracil sequence at their 3’-ends [25] (Fig. 8). tiRNAs, which have a length of ~28–36 nt, are produced in the cytoplasm via specific cleavage of the anticodon loop of mature tRNAs by Rny1p and angiogenin (ANG) in yeast and mammalians cells, respectively [34–36]. This class comprises 5’-tiRNA and 3’-tiRNA, in reference to the 5' or 3' half of the mature tRNA they derive from, respectively [34] (Fig. 8). tRFs, ranging from 14–30 nt in length, are derived from mature tRNA [34,37,38].

Three types of tRFs have been discovered to date: (i) tRF-5s, (ii) tRF-3s, and (iii) i-tRFs [39,40]. tRF-5s are generated in the cytoplasm by Dicer-mediated cleavage of the mature tRNA D-loop [39,41] (Fig. 8). tRF-3s are produced in the cytoplasm via cleavage of the T-loop in mature tRNAs operated by Dicer, angiogenin and other members of the RNase A superfamily. They are fragments originating from mature tRNA 3'-ends, and include the final CCA sequence [38,39,42] (Fig. 8). Finally, i-tRFs are enriched within the internal regions of mature tRNAs, usually straddling the anticodon region [39,43]. It is important to highlight that in literature and in some databases, tsRNAs (which derive from 3' trailer sequence of pre-tRNAs) are also termed tRF-1s [40,44,45].

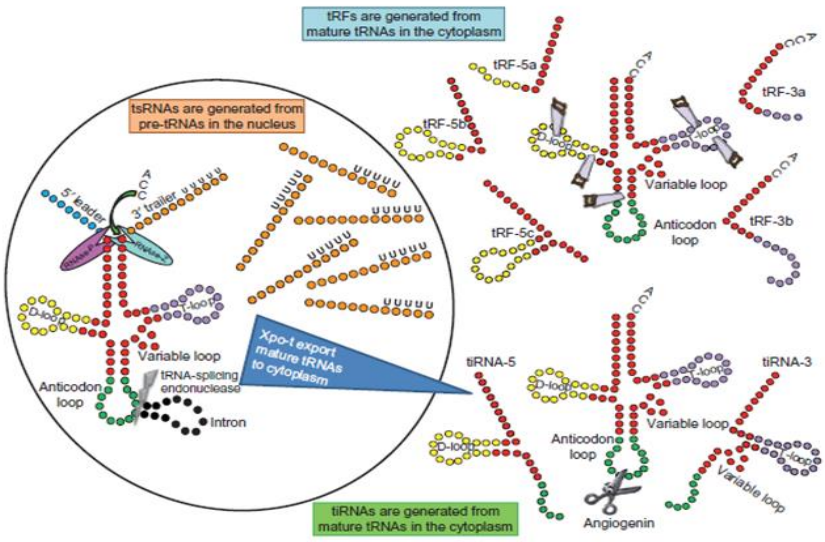

Fig 8. Classification of the different tRNA-derived ncRNA subclasses (Source Balatti V. et al. Advances in Cancer Research. 2017)

Various technologies have been developed to identify and quantify the transcriptome, including hybridization or sequence-based approaches. Hybridization-based approaches typically involve incubating fluorescently labeled cDNA with custom-made microarrays or commercial high-density oligo microarrays. Specialized microarrays have also been designed to detect and quantify distinct spliced isoforms [46]. Hybridization-based approaches are high throughput and relatively inexpensive, except for high-resolution arrays [23]. However, these methods have several limitations,

which include: (1) reliance upon existing knowledge about RNA sequences to be analyzed; (2) high background levels noise due to cross-hybridization; (3) and a limited dynamic range of detection owing to both background and saturation of signals. Moreover, comparing expression levels across different experiments is often difficult and can require complicated normalization methods [23]. In contrast to microarray methods, sequence-based approaches directly determine the cDNA sequence. Recently, the development of novel high-throughput sequencing methods has provided a new method for both mapping and quantifying transcriptomes. This method, termed RNA-Seq, has clear advantages over existing approaches and is expected to revolutionize the manner in which eukaryotic transcriptomes are analyzed [23].

1.2.2 RNA-Seq technology overview

As we previously said, RNA-Seq relies on NGS platforms for RNA identification and quantification. Generally, first of all, a population of RNA (total or fractionated) is first purified and second converted to a library of cDNA fragments with adaptors attached to one or both ends [23]. Each molecule, with or without amplification, is then sequenced in a high-throughput manner to obtain short sequences from one end (single-end sequencing) or both ends (pair-end sequencing) (Fig. 9)

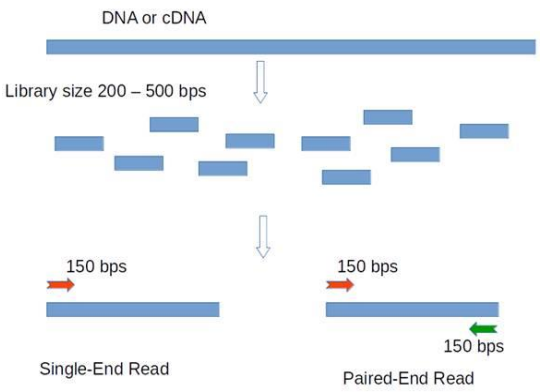

Fig 9. Single-end and paired-end reads (Source https://www.1010genome.com/illumina-sequencing/)

The reads are typically 30–400 bp, depending on the NGS platform used. In principle, any high-throughput sequencing technology can be used for RNA-Seq, but Illumina platforms are definitely the most used for transcriptome analysis. Following sequencing, the resulting reads are either aligned to a reference genome or transcriptome or assembled *de novo* without the genomic sequence to produce a genome-scale transcription map that consists of both the transcriptional structure and/or level of expression for each gene (Fig. 10).

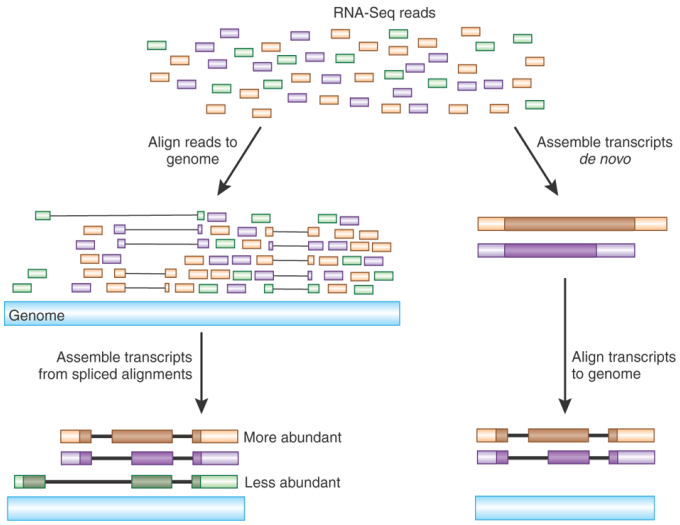

Fig 10. Alignment of reads to a reference genome and de novo transcriptome assembly (Source Haas and Zody Nat. Biotech. 2010).

Although RNA-Seq is still a technology under active development, it offers several key advantages over existing technologies. First, unlike hybridization-based approaches, RNA-Seq is not limited to detect only known transcripts [23]. This makes RNA-Seq particularly attractive for non-model organisms with genomic sequences that are not determined yet. RNA-Seq can also reveal the precise location of transcription boundaries, to a single-base resolution. Furthermore, RNA-Seq can also give information about how two exons are connected or identify sequence variations (for example, SNPs) in transcripts [23] making this technology very helpful for studying complex transcriptomes. Finally, another important advantage of RNA-Seq relative to

microarrays is that RNA-Seq has very low, if any, background signal because the reads can be unambiguously mapped to unique regions of the genome. RNA-Seq does not have an upper limit for quantification, which correlates with the number of sequences obtained. Consequently, it has a large dynamic range of expression levels over which transcripts can be detected [23]. By contrast, microarrays lack sensitivity for genes expressed either at low or very high levels and therefore have a much smaller dynamic range [23]. More importantly, RNA-Seq has also been shown to be highly accurate for quantifying expression levels, as determined using quantitative PCR (qPCR) [47] and spike-in RNA controls of known concentration [48]. Taking all of these advantages into account, RNA-Seq is the first sequencing-based method that allows the entire transcriptome to be analyzed in a very high-throughput and quantitative manner. This method offers both single-base resolution for annotation and 'digital' gene expression levels at the genome-scale, often at a much lower cost than microarray.

1.2.3 RNA-Seq experimental design

A crucial prerequisite for a successful RNA-seq study is that the data generated have the potential to answer the biological questions of interest. This is achieved by first defining a good experimental design, that consist of choosing the library type, sequencing depth, and the number of replicates appropriate for the biological system under study, and second by planning an adequate execution of the sequencing experiment itself, ensuring that data acquisition does not become contaminated with unnecessary biases [49]. One important aspect of the experimental design is the RNA-extraction protocol used to remove the highly abundant ribosomal RNA (rRNA), which typically constitutes over 90 % of total RNA in the cell, leaving the other RNA species (mRNAs and ncRNAs) that we are normally interested in [49]. Another

important step is whether to generate strand-preserving libraries. The first generation of Illumina-based RNA-seq used random hexamer primers to reverse-transcribe RNAs. This methodology did not retain the information contained on the DNA strand that is actually expressed, and therefore complicates the analysis and quantification of antisense or overlapping transcripts [49]. Several strand-specific protocols, such as the widely used dUTP method, extend the original protocol by incorporating UTP nucleotides during the second cDNA synthesis step, prior to adapter ligation followed by digestion of the strand containing dUTP [50] (Fig. 11).

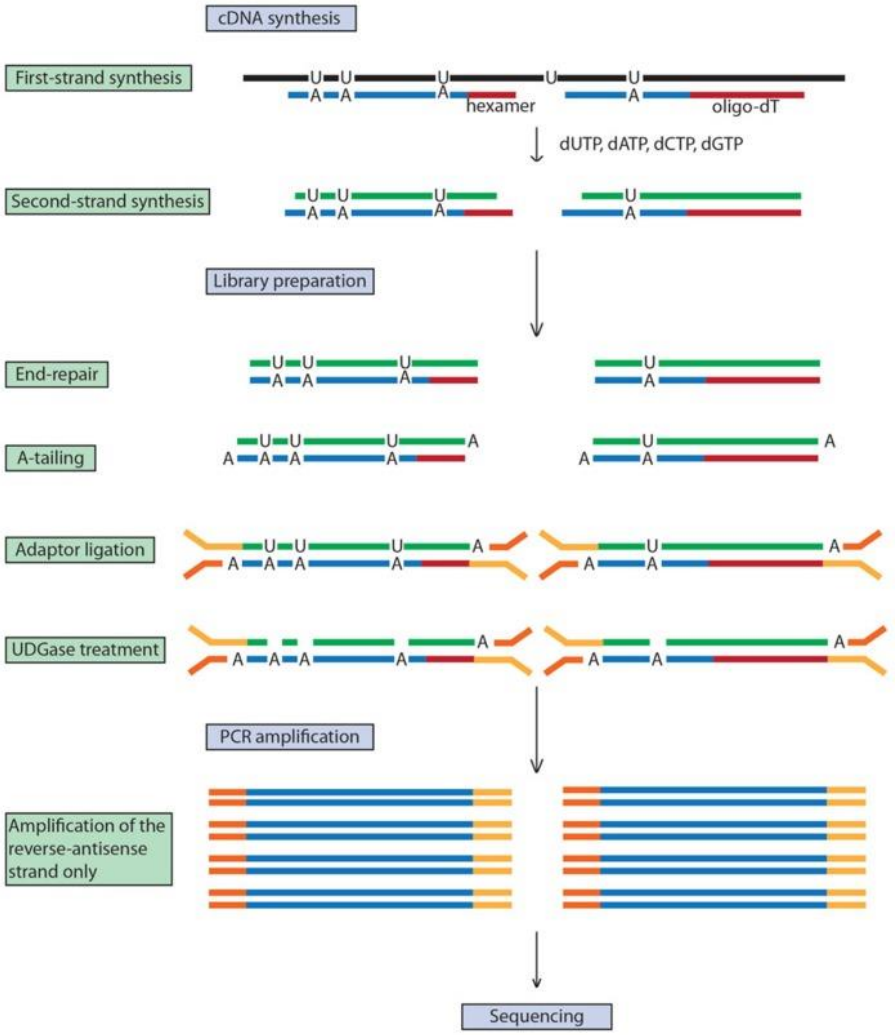

Fig 11. Strand-specific sequencing (Source Martin L. et al. Frontiers in Plant Science. 2013)

Furthermore, sequencing can involve single-end (SE) or paired-end (PE) reads, although the latter is preferable for *de novo* transcript discovery or isoform expression

analysis [51,52]. The best sequencing option depends on the analysis goals. The cheaper, short SE reads are normally sufficient for studies of gene expression levels in well-annotated organisms or for the analysis of small ncRNAs, whereas longer and PE reads are preferable to characterize poorly annotated transcriptomes [49]. Another important factor is sequencing depth, which is the number of sequenced reads for a given sample. More transcripts will be detected and their quantification will be more precise as the sample is sequenced to a deeper level. Nevertheless, optimal sequencing depth again depends on the aims of the experiment. While some authors will argue that as few as five million mapped reads are sufficient to quantify accurately medium to highly expressed genes in most eukaryotic transcriptomes, others will sequence up to 100 million reads to quantify precisely genes and transcripts that have low expression levels [53]. Finally, a crucial design factor is the number of replicates. The number of replicates that should be included in an RNA-seq experiment depends on both the amount of technical variability in the RNA-seq procedures and the biological variability of the system under study, as well as on the desired statistical power (that is, the capacity for detecting statistically significant differences in gene expression between experimental groups).

1.2.4 RNA-Seq data analysis

The power of RNA-Seq lies in the fact that the twin aspects of discovery and quantification can be combined in a single high-throughput assay. The use of RNA-seq has spread well beyond the genomics community and has become a standard tool used by the life sciences research community. Many variations of RNA-seq protocols and analyses have been published, making it challenging for new users to appreciate all of the steps necessary to conduct an RNA-seq study properly [49]. There is no optimal

pipeline for the variety of different applications and analysis scenarios in which RNA-seq can be used. Scientists plan experiments and adopt different analysis strategies depending on the organism being studied and their research goals. For example, if a genome sequence is available for the studied organism, it should be possible to identify transcripts by mapping RNA-seq reads onto the genome. By contrast, for organisms without sequenced genomes, quantification would be achieved by first assembling *de novo* reads into contigs and then mapping these contigs onto the transcriptome. For well-annotated genomes such as the human genome, researchers may choose to base their RNA-seq analysis on the existing annotated reference transcriptome alone, or might try to identify new transcripts and their differential expression. Furthermore, investigators might be interested only in mRNAs splicing-variants expression or small ncRNAs levels. Both the experimental design and the analysis procedures are very different in each of these cases [49]. Every RNA-seq experimental scenario could potentially have different optimal methods for transcript quantification, normalization, and ultimately differential expression analysis. Moreover, quality control checks should be applied at different stages of the analysis to ensure both reproducibility and reliability of the results [49]. Below follow a description of the several steps which are typically involved in a standard RNA-Seq data analysis.

1.2.4.1 Quality Control

As previously said, the analysis of RNA-seq data requires several steps such as obtaining raw reads, performing read alignment, and their quantification. At each of these steps, specific checks should be applied to monitor the quality of the data. Quality control for the raw reads involves the analysis of sequence quality, GC content, the presence of adaptors, overrepresented k-mers, and duplicated reads in order to detect sequencing errors, PCR artifacts or contaminations [49]. FastQC [54] is a

popular tool to perform these analyses on Illumina reads, whereas NGSQC [55] can be applied to any platform. As a general rule, read quality decreases towards the 3' end of reads, and if it becomes too low, bases should be removed to improve mappability (trimming). Software tools such as Cutadapt [56] and Trimmomatic [57] can be used to discard low-quality reads, remove adaptor sequences, and eliminate poor-quality bases. Concerning quality control on the read alignment step, an important parameter to look at is the percentage of mapped reads, which is a global indicator of the overall sequencing accuracy and of the presence of contaminating DNA. For example, we expect between 70 and 90 % of regular RNA-seq reads to map onto the human genome. Instead, when reads are mapped against the transcriptome, we expect slightly lower total mapping rates because reads coming from unannotated transcripts will be lost. Tools generally used for the alignment quality control are Picard, RSeQC [58], and Qualimap [59]. Finally, once those transcripts are mapped and quantified, they should be checked for GC content and gene length biases so that correcting normalization methods can be applied if necessary. For this purpose, several R packages (such as NOISeq [60] or EDASeq [61]) have been developed to provide useful plots for quality control of count data.

1.2.4.2 Alignment

When a reference genome is available, RNA-seq analysis will normally involve the mapping of the reads onto the reference genome or transcriptome to infer which transcripts are expressed. Mapping to the reference transcriptome of a known species precludes the discovery of new, unannotated transcripts and focuses the analysis on quantification alone. On the other hand, if the organism does not have a sequenced genome, then the analysis consists first to assemble reads into longer contigs and then to treat these contigs as the expressed transcriptome to which reads are mapped back

again for quantification [49]. However, if the organism under study has a sequenced genome, we can choose two different read alignment approaches: 1) align the reads to a reference genome, 2) or align the reads to a reference transcriptome. Independently if it is used a genome or transcriptome as a reference sequence, reads may map uniquely (they can be assigned to only one position in the reference) or could be multi-mapped reads (multireads) [49]. Multimapping is primarily due to repetitive sequences or shared domains of paralogous genes. They normally account for a significant fraction of the mapping output when mapped onto the genome and should not be discarded [49]. When the reference is the transcriptome, multi-mapping arises even more often because a read that would have been uniquely mapped on the genome would map equally well to all gene isoforms in the transcriptome that share the exon [49]. In both cases, transcript identification and quantification is very challenging for alternative splicing genes [49]. Several tools for the alignment of reads on a reference genome or transcriptome have been developed during the last years. For most common RNA-Seq applications, STAR [62] and HISAT 2 [63] are very used for genome-based alignment while for transcriptome-based alignment a novel and powerful tool is SALMON [64].

1.2.4.3 Novel transcripts discovery

Identifying novel transcripts using the short reads provided by Illumina technology is one of the most challenging tasks in RNA-seq. Short reads rarely span across several splice junctions and thus make it difficult to directly infer all full-length transcripts. In any case, PE reads and higher coverage help to reconstruct lowly expressed transcripts, and replicates are essential to resolve false-positive calls (that is, mapping artifacts or contaminations) at the low end of signal detection. Several methods, such as Cufflinks [65], iReckon [66], SLIDE [67], and StringTie [68], incorporate existing annotations by adding

them to the possible list of isoforms. In general, accurate transcript reconstruction from short reads is difficult, and methods typically show substantial disagreement [69].

1.2.4.4 De novo transcript reconstruction

When a reference genome is not available or is incomplete, RNA-seq reads can be assembled *de novo* into a transcriptome using packages such as SOAPdenovo-Trans [70], Oases [71], Trans-ABySS [72] or Trinity [73]. In general, PE strand-specific sequencing and long reads are preferred because they are more informative [49]. Although it is impossible to assemble lowly expressed transcripts that lack enough coverage for a reliable assembly, too many reads are also problematic because they lead to potential misassembly and increased runtimes. Therefore, in silico reduction of the number of reads is recommended for deeply sequenced samples. For comparative analyses across samples, it is advisable to combine all reads from multiple samples into a single input in order to obtain a consolidated set of contigs (transcripts), followed by mapping back of the short reads for expression estimation [49]. Either with a reference or *de novo*, the complete reconstruction of transcriptomes using short-read Illumina technology remains a challenging problem, and in many cases, *de novo* assembly results in tens or hundreds of contigs accounting for fragmented transcripts. Emerging long-read technologies, such as SMRT from Pacific Biosciences, provide reads that are long enough to sequence complete transcripts for most genes and they are a promising alternative to Illumina short reads technology [49].

1.2.4.5 Transcript quantification

The most common application of RNA-seq is to estimate gene and transcript expression. This can be easily done by counting the number of reads that map to each transcript sequence. The easiest way to do it is to summarize raw counts of mapped

reads using programs such as HTSeq-count [74] or featureCounts [75] that allow a gene-level (rather than transcript-level) quantification using Gene Transfer Format (GTF) files containing the genome coordinates of exons and genes. However, raw read counts alone are not sufficient to compare expression levels among samples, as these values are affected by factors such as transcript length, the total number of reads, and sequencing biases. For this reason, raw read counts must be first normalized. Several normalization methods are available, and they are used to remove technical biases in sequenced data such as depth of sequencing (more sequencing depth produces more read count for a gene expressed at the same level) and gene length (differences in gene length generate unequal reads count for genes expressed at the same level; longer the gene more the read count). One of these methods is to calculate the number of Read Per million Mapped reads (RPM).

$$RPM = \frac{Number\ of\ reads\ mapped\ to\ gene\ *\ 10^{6}}{Total\ number\ of\ mapped\ reads}$$

RPM is used to normalize the difference in sequence depth across the samples, but it does not take into account differences in transcript length. Another very common way to normalize the raw counts is to calculate the Transcript Per Million (TPM).

$$TPM = \frac{Number\ of\ reads\ mapped\ to\ gene\ *\ read\ length\ *\ 10^{6}}{Total\ number\ of\ transcripts\ *\ gene\ length\ in\ bp}$$

Instead of normalizing for sequencing depth only, TPM allows normalization for the difference in gene lengths. Correcting for gene length is not necessary when comparing changes in gene expression within the same gene across samples, but it is necessary for correctly ranking gene expression levels within the sample to account for the fact that longer genes accumulate more reads. Finally, several R packages for differential expression analysis such as LIMMA [76], DESeq 2 [77] and edgeR [78] have their

own normalization methods that are used prior to the analysis for the identification of the differentially expressed transcripts.

1.2.4.6 Differential expression analysis

Differential expression analysis consists of comparing the gene expression values between groups of samples (for example disease vs controls; treated vs controls) in order to find genes involved in that biological scenario. Many statistical methods have been developed for detecting differentially expressed genes or transcripts from RNA-seq data, and a major challenge is how to choose the most suitable tool for a particular data analysis. Most comparison studies have focused on simulated datasets [79–81] or on samples to which exogenous RNA ('spike-in') has been added in known quantities [82,83]. This enables a direct assessment of the sensitivity and specificity of the methods. However, no clear consensus has been reached regarding the best practices yet. Among these tools, limma [76] has been shown to perform well under many circumstances and it is also the fastest to run [79,82,84]. DESeq [77] and edgeR [78] perform similarly in ranking genes but are often relatively conservative or too liberal, respectively, in controlling FDR [81,82,85]. SAMseq performs well in terms of FDR but presents an acceptable sensitivity when the number of replicates is relatively high at least 10 [61,81]. NOISeq and NOISeqBIO (the adaptation of NOISeq for biological replication) are more efficient in avoiding false-positive calls at the cost of some sensitivity but perform well with different numbers of replicates [61,86,87]. Cuffdiff and Cuffdiff2 have performed surprisingly poorly in the comparisons [79,82]. All this limitation probably reflects the fact that detecting differentially expressed genes remains challenging. Considering the drop in the price of sequencing, it is recommended that RNA-seq experiments have a minimum of three biological replicates. Recent independent comparison studies have demonstrated that the choice

of the method can dramatically affect the outcome of the analysis and that no single method is likely to perform favorably for all datasets [79,82,84]. Therefore, it is recommended using more than one package to calculate the differentially expressed genes for important analysis. In addition, users should take in mind that while some of these differential expression tools can only perform a pairwise comparison, others such as LIMMA [76], DESeq2 [77], and edgeR [78] can perform multiple comparisons, include different covariates or analyze time-series data.

1.2.4.7 Pathway analysis

The last step in a standard transcriptomics analysis is often the characterization of the molecular functions in which the differentially expressed genes (DEGs) are involved. The two main approaches for functional characterization that were first developed for microarray technology are (a) comparing a list of DEGs against the rest of the genome for overrepresented functions, and (b) gene set enrichment analysis (GSEA), which is based on ranking the transcriptome according to a measurement of differential expression. However, functional analysis requires the availability of sufficient functional annotation data for the transcriptome under study.

Examples of resources that contain such functional annotation for several model species are Gene Ontology [88], and DAVID [89]. Unfortunately, novel transcripts discovered during *de novo* transcriptome assembly or some newly discovered small ncRNA classes would lack functional annotation. In any case, only the functional annotation of genes and ncRNAs is not sufficient to understand the biological mechanisms underlying the phenotype under study. Indeed, we know that alteration in the expression profile of protein-coding genes and ncRNAs have significant impacts on the function of signaling and metabolic pathways. For this purpose, pathway analysis is becoming a more crucial step during RNA-Seq data analysis to understand

which roles the differentially expressed genes and ncRNAs are involved in. For pathway analysis we mean an extensive class of methods, which are able to determine the status of biological processes, identifying altered functionalities involved in complex diseases [90]. The pathway-based analysis uses knowledge from pathway databases, providing insight on genes and how they interact [90]. Specifically, pathway analysis originally identified a class of techniques for (1) the analysis of ontological terms and protein-protein interaction (PPI) networks; (2) the inference of gene regulatory networks from expression data. The aim was to use pathways as knowledge bases for grouping genes or proteins into smaller subsets according to some relationships, thus reducing the dimensionality of expression data. More recently, research effort has been devoted to deploying a novel class of methods called knowledge base-driven pathway analysis. Such methods leverage existing databases such as the Kyoto Encyclopedia of Gene and Genomes (KEGG) [91], and Reactome [92] to identify perturbed pathways associated with a specific phenotype or condition. The degree of perturbation can be measured starting from several parameters, including the number of DEGs belonging to the pathway, the magnitude of their expression changes, and their interaction type, direction, and strength. A typical knowledge base-driven pathway analysis method starts from two input data: (1) a set of pathways representing the molecular interaction knowledge base, and (2) experimental data containing measurements of gene expressions, protein abundance, or metabolite concentration in two or more conditions. Some methods preprocess input data to select only a subset of genes considered to be differentially expressed based on a predefined cutoff, which is typically applied on fold-change, statistical significance, or both. However, the usage of cutoffs could be critical and data-dependent, influencing the quality of the results. A graph model is then built to represent pathways. Models

depend on pathway type: (1) signaling pathway where nodes are genes (or gene products) and edges represent signals, such as activation or repression, (2) metabolic pathway in which nodes are biochemical compounds and edges represent reactions that transform one or more compounds into another one. Pathways are then ranked according to the level of perturbation which is computed through to a scoring scheme. Following a temporal criteria, knowledge base-driven pathway analysis methods can be classified into three generations of approaches: (1) Overrepresentation analysis (ORA), (2) functional class scoring (FCS), and (3) pathway topology-based (PT). Unlike ORA and FCS methods that only consider the presence of specific sets of DEGs for each pathway to identify the dysregulated ones, PT systems fully exploit the topological information encoded by pathways when computing perturbation scores. Indeed, pathways are modeled as complex graphs where each node is a gene or a protein and each edge is an interaction between them. Even though thousands of genes are not annotated in pathways and existing annotations may be inaccurate, graphs contained in these databases provide a more detailed view of biological processes within the cell, helping the interpretation of high-throughput experiments [90]. This is a significant advancement than the previous pathway analysis systems which do not consider the topology of pathways to calculate their perturbation but only the presence of a set of DEGs for each pathway. Finally, more recently, new approaches have been proposed to analyze pathways enriched with missing regulatory elements, such as miRNAs and their post-transcriptional regulatory interactions with genes. One of them is the approach proposed by MITHrIL [93]. However, for other post-transcriptional regulators, such as tRNA-derived ncRNAs, biological pathways are still incomplete, and therefore, they have not been considered yet.

1.2.5 Pipelines for RNA-Seq

The analysis of RNA-Seq data is very challenging and it requires specific bioinformatics and computer programming skills which may be uncommon in small research laboratories. Moreover, as we previously discussed, several steps (quality control; trimming and adaptors removing; alignment; read counting; normalization; differential expression analysis; etc) are usually required to perform a standard RNA-Seq analysis. Therefore, in order to promote the use of RNA-Seq technologies and expand the community of scientists able to analyze such data, many pipelines have been developed to simplify the RNA-Seq analysis. Relevant examples include BioJupies [94], BioWardrobe [95], DEWE [96], easyRNASeq [97], ExpressionPlot [98], FX [99], GENE-counter [100], GeneProf [101], Grape RNA-Seq [102], MAP-RSeq [103], RAP [104], RobiNA [105], RSEQtools [106], RseqFlow [107], S-MART [108], TCW [109], TRAPLINE [110] and wapRNA [111]. In addition, other pipelines have also been developed for the analysis of different ncRNA classes. Examples include DSAP [112], miRanalyzer [113], miRExpress [114], miRNAkey [115], iMir [116], CAP-miRSeq [117], mirTools 2.0 [118], sRNAtoolbox [119], miRDeep 2 [120], and MapMi [121] for miRNA analysis; piPipes [122], PILFER [123], piRNAPredictor [124] and PIANO [125] for piRNA analysis; and UClncR [126] for lncRNA analysis. Although interesting, some of these tools present several limitations and shortcomings which have negatively impacted their usage among non-expert users such as (i) no Graphical User Interface but only command line shell; (ii) software dependencies prior to the pipeline installation; (iii) support only for UNIX operating systems; (iv) static workflow (they do not allow to select which tool has to be used for each step of the pipeline); (v) not suitable for the analysis of the whole transcriptome (e.g. mRNAs or few ncRNA classes supported); (vi) no downstream analysis modules (i.e. differential expression analysis or pathway analysis); (vii) only few species

supported. These limitations must not be underestimated since RNA-Seq technologies are always more used both in research and biomedical laboratories and, therefore, the number of scientists interested to analyze such data is rising. For these reasons, the request for easy-to-use pipelines for the analysis of RNA-Seq data has become very urgent.

2. Aim of the project

The aim of the Ph.D. research project funded by the Italian MIUR “PON RI FSE-FESR 2014-2020” in collaboration with Nerviano Medical Sciences and Department of Cancer Biology and Genetics of The Ohio State University was to implement a user-friendly software for the analysis of RNA-Seq data. The idea was to develop a software that could be used by users with no computer programming background in order to promote its use in research and biomedical laboratories by expanding the community of life scientists able to analyze RNA-Seq data. In order to achieve this goal, we implemented *RNAdetector* a completely offline, cross-platform and stand-alone software with an easy-to-use graphical user interface capable of analyzing coding and ncRNAs of any sequenced biological species. Precisely, *RNAdetector* has been designed to be able not only to identify and quantify such RNA molecules but also to be able to perform differential expression analysis, and miRNA-sensitive topological pathway analysis allowing users to infer important biological information from their data.

3. Project milestones

In order to complete the proposed project, several steps were required:

1. First of all, our software aimed to analyze an extensive repertoire of different classes of ncRNAs from RNA-Seq data such as miRNA, piRNAs, snoRNAs, snRNAs, tUCRs, lncRNAs, circRNAs, and tRNA-derived ncRNAs. However, for the latter, no previous system for their detection from small RNA-Seq data was released when the project started in 2018. Also, at that time there were not extensive databases that covered all the different subclasses (already available databases such as tRFdb [44] and MINTbase [127] were primarily focused on tRFs deriving from mature tRNAs). Therefore, we had to implement our own method to detect all the different subclasses of tRNA-derived ncRNAs in small RNA-Seq datasets and include them in a novel database [128]. Finally, this database would collect all the tRNA-derived ncRNAs that it would be possible to analyze with *RNAdetector* (we named this database *tRFexplorer* and we published it in 2019 [128])
2. Secondly, we had to test some recent pipelines for the analysis of ncRNAs from RNA-Seq data [129]. The idea was to identify their strengths and weaknesses and, therefore, optimize *RNAdetector* in order to fill the gaps of the previous methodologies [129] (we published the results of this benchmark in 2019 [129])
3. Finally, we had to implement *RNAdetector*.

The project lasted 3 years and was carried out in collaboration with Nerviano Medical Sciences and the Department of Cancer Biology and Genetics of The Ohio State

University where I spent one year and six months, respectively. In the end, two papers were published [128,129] and another one is under review in a well-known bioinformatics journal. The methods and results discussed in this PhD thesis are related to these 3 our papers. A summary of the PhD project milestones is shown in Fig. 12.

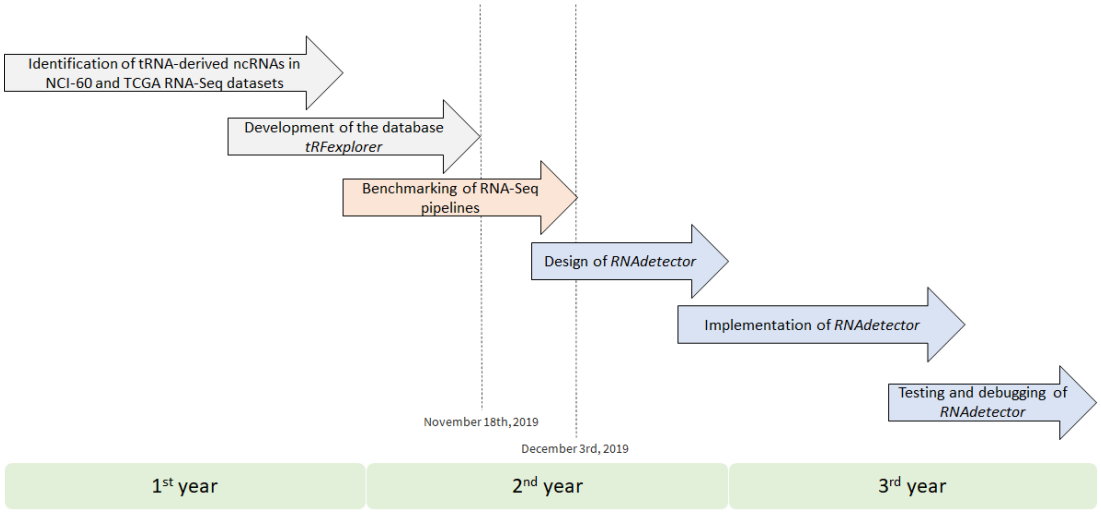

Fig 12. PhD project milestones

4. Materials and methods

4.1 Identification of tRNA-derived ncRNAs and database implementation

As I previously discussed, our software aims to analyze an extensive repertoire of different classes of ncRNAs including also all the different tRNA-derived ncRNAs subclasses. However, for the latter, no previous system for their detection was released when the project started in 2018. Also, at that time there were not extensive databases that covered all the different subclasses (already available databases such as tRFdb [44] and MINTbase [127] were primarily focused on tRFs deriving from mature tRNAs). Therefore, we had to implement our own method to detect all the different tRNA-derived ncRNAs subclasses in small RNA-Seq datasets and include them in a novel database. Finally, this database will collect all the tRNA-derived ncRNAs that it will be possible to analyze with *RNAdetector*. For this reason, we selected the public small RNA-Seq datasets of the NCI-60 cancer cell lines and The Cancer Genome Atlas (TCGA) samples in order to identify all the expressed tRNA-derived ncRNAs and we included them in a novel database that we named *tRFexplorer*. In the next two sections, I discuss the methods used for the detection of the tRNA-derived ncRNAs and for the database implementation (these methods are also reported in our paper titled *Identification of tRNA-derived ncRNAs in TCGA and NCI-60 panel cell lines and development of the public database tRFexplorer* published in 2019 [128]).

4.1.1 Identification of the tRNA-derived ncRNAs subclasses

The identification of tsRNAs, 5' leader RNAs, and tRFs in small RNA-Seq datasets is a complex process, since such small fragments may be mapped to multiple DNA regions. For this purpose, we implemented a conservative pipeline to get an accurate estimation of tsRNAs, 5' leader RNAs, and tRFs expression. First, we assembled a custom annotation of the reference human genome (hg19) containing only known tsRNAs and tRFs. We included all tRF-5s, tRF-3s, and tRF-1s from tRFdb (http://genome.bioch.virginia.edu/trfdb/) [44], all tsRNA identified by Carlo Croce's Lab [31], and the 20nt upstream region of tRNA human genes for the 5' leader RNAs. Human tRNA genes were taken from GtRNAdb (http://gtrnadb.ucsc.edu/) [130]. Subsequently, we examined sncRNA-seq datasets of NCI-60 cell lines as provided by the Sequence Read Archive (SRA) (PRJNA390643) [131], as well as small RNA-seq datasets on TCGA. In the supplementary table 1 we provide a list of NCI-60 cell lines and the SRA datasets while the supplementary table 2 lists the analyzed TCGA cancer types with their relative numbers of tumor and control samples. Raw FASTQ files were pre-processed for adaptor removal and quality filtering by applying Trim Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) tuned for sncRNA-seq (Phred quality score ≥ 20). Trim Galore is a wrapper for Cutadapt [56] and FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), which is used as a consistent method to apply quality filtering and adaptor trimming. Filtered FASTQ files were then aligned to a reference human genome (hg19) using TopHat version 2.1.0 [132] as well as to our custom annotation file. Read quantification has been performed with HTSeq version 0.10.0 [74]. In this phase, all ambiguously mapped reads were removed for a more accurate and conservative analysis. Data analysis was performed with R version 3.5.1. Raw counts were normalized with 2 different

normalization methods: transcripts per million mapped reads (TPM) [133], and reads per million mapped reads (RPM) [48].

$$RPM = \frac{number\ of\ reads\ mapped\ to\ a\ gene\ \times\ 10^6}{total\ number\ of\ mapped\ reads\ for\ a\ given\ library}$$

$$TPM\ = \frac{number\ of\ reads\ mapped\ to\ a\ gene}{gene\ length\ in\ bp} \times \left(\sum \frac{1}{\frac{number\ of\ reads\ mapped\ to\ a\ gene}{gene\ length\ in\ bp}} \right) \times 10^6$$

All tsRNAs, 5' leader RNAs, and tRFs with average log2 TPM less than 1 were removed. A summary of our full pipeline is illustrated in Fig. 13.

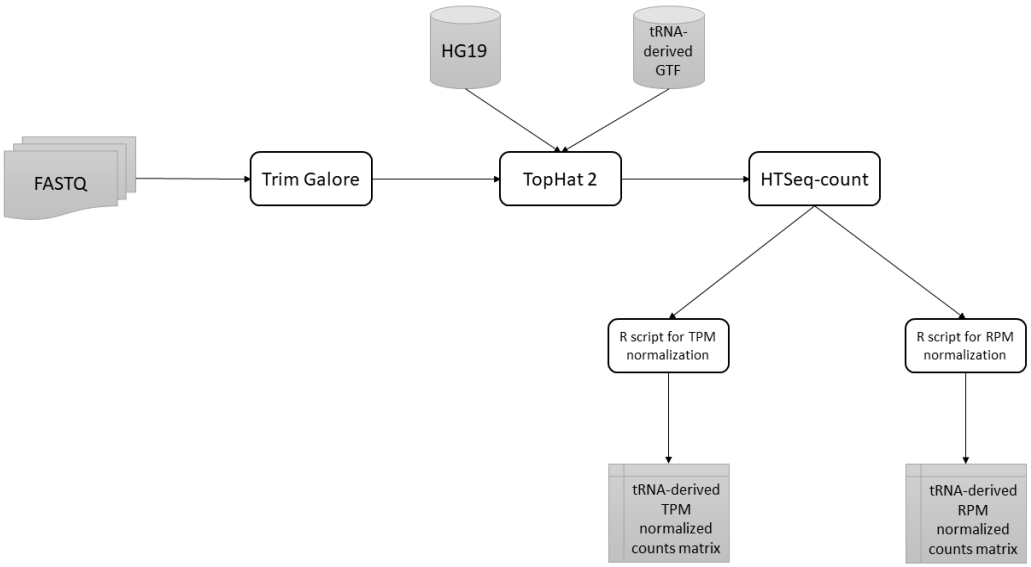

Fig 13. Pipeline for the identification of the tRNA-derived ncRNAs in NCI-60 and TCGA small RNA-Seq datasets

4.1.2 Database implementation

All identified tsRNAs, 5' leader RNAs, and tRFs with their expression profiles have been integrated into a novel database named *tRFexplorer* [128]. *tRFexplorer* enables users to visualize the expression profile of each tRNA-derived ncRNA in both NCI-60 cell lines and TCGA samples (33 cancer types). Furthermore, it uses the R package limma [76] to perform differential expression analysis on TCGA data. Interactive visualization of its results have been implemented through the R package Glimma [134].

Our database allows users to conduct correlation analyses of tRNA-derived ncRNAs expressions in NCI-60 with all data available on CellMiner [135,136]. Correlation analysis with genes and miRNA expression profiles, as well as patient survival, of TCGA samples has also been implemented. *tRFexplorer* was developed by employing PHP and R for its backend, while Javascript, and React for the main user interface. All omics data and compound activities used for the correlation analysis were obtained from CellMiner [135,136]. In supplementary table 3 we list all CellMiner datasets. Genomic viewer for tRNA-derived ncRNAs visualization is based on JBrowse [137]. JBrowse is a fast and interactive genomic viewer built entirely with new HTML5 technology. We customized our browser by allowing users to search for both tRNAs or tRNA-derived ncRNAs using both genomic coordinates or identifiers.

4.2 Testing of previous ncRNAs pipelines

After the implementation of *tRFexplorer* [128] and before proceeding with the development of our software for the analysis of RNA-Seq data named *RNAdetector*, we wanted to evaluate the state of the art of current ncRNA pipelines in order to identify their strengths and weaknesses, and therefore, optimizing *RNAdetector* by filling the gaps of the previous methodologies. For this purpose, the performances of eight ncRNA pipelines enabling the processing of RNA-Seq data, published between 2015 and 2019, were compared. In particular, we evaluated the easiness of installation and usage together with their accuracies to identify ncRNAs and their expression levels by using both synthetic and real RNA-Seq datasets. A detailed description of the methods used for this benchmark is described in the next sections (these methods are also reported in our paper titled *A benchmarking of pipelines for detecting ncRNAs from RNA-Seq data* published in 2019 [129]).

4.2.1 Design of synthetic RNA-seq datasets

Two synthetic human RNA-Seq datasets were generated and used for the comparison of the different ncRNAs pipelines. The first dataset simulates small non-coding RNA-Seq data and includes miRNAs, piRNAs, snoRNAs, and tRNA-derived ncRNAs, while the second one simulates standard RNA-Seq data and covers mRNAs, circRNAs, and lncRNAs. Synthetic RNA-Seq data were obtained by using Flux Simulator [138]. Briefly, Flux Simulator simulates a transcriptome starting from the genomics sequences of specific species and the corresponding gene structure annotation. Then, the obtained *in silico* transcriptome undergoes RT/fragmentation according to the chosen experimental technique. Flux Simulator pipeline provides optional steps for modeling the final library preparation, involving in silico ligation of adapter sequences, fragment size selection, and PCR amplification. Finally, Flux Simulator produces a FASTQ file as output in which reads associated with each ncRNA are annotated. In this way, it is possible to precisely calculate the species and the number of reads associated with each RNA molecule present in the synthetic dataset [138]. The sequences of simulated reads are retrieved from the reference genome in correspondence to the genomic coordinates reported in the GTF file which is provided in input together with the reference genome. We choose Flux Simulator to generate RNA-Seq data since it comprises explicit models for the processes that determine the abundance and distribution of reads according to specified experimental protocols [138]. Specifically, to build the two datasets, we first created custom annotation files (GTF format) for small RNAs and long RNAs, respectively. These files were created by using an R script which permits to randomly select molecules from the original genomics coordinates present in the following databases: UCSC [139] for mRNAs, miRNAs, and snoRNAs; our database tRFexplorer for all tRNA-derived

ncRNAs classes [128]; piRBase [140] for piRNA; circBase [141] for circRNAs and LNCipedia [142] for lncRNAs. This process yielded one GTF file for small RNA, containing genomics coordinates for 193 miRNAs, 100 snoRNAs, 500 piRNAs and 200 tRNA-derived ncRNAs (equally split in 5 leader RNAs, tsRNAs, tRF-5s and tRF-3s), and one annotation file for long RNAs including coordinates for 100 genes encoding for proteins (895 exon genomic coordinates), 500 circRNAs and 500 lncRNAs. Concerning circRNAs, however, Flux simulator was not able to correctly simulate their sequences. A possible explanation could be that circRNAs genomics coordinates, which were used for the generation of simulated circRNAs reads, were huge, and the genes transcribed within circRNAs were not annotated for introns and exons in circBase. Therefore, Flux Simulator might have generated reads which straddle between exons and introns of coding-protein genes, which are not components of circRNA molecules. Therefore, circRNAs were not included in the evaluation. GTF files, together with the human HG19 reference genome, were then submitted to Flux Simulator [138] and synthetic FASTQ files were built. The parameters used for the generation of the two synthetic RNA-seq data sets are reported in Table 1.

Small RNA-seq		**Standard RNA-seq**	
NB_MOLECULES	5000000	NB_MOLECULES	5000000
READ_NUMBER	15000000	READ_NUMBER	5000000
TSS_MEAN	NaN	TSS_MEAN	NaN
FRAG_SUBSTRATE	RNA	FRAG_SUBSTRATE	RNA
POLYA_SCALE	NaN	POLYA_SCALE	NaN
POLYA_SHAPE	NaN	POLYA_SHAPE	NaN
PCR_PROBABILITY	0.5	PCR_PROBABILITY	0.5
FRAG_METHOD	UR	FRAG_METHOD	UR
FRAG_EZ_MOTIF	NlaIII	FRAG_EZ_MOTIF	NlaIII
READ_LENGTH	35	READ_LENGTH	75
PAIRED_END	false	PAIRED_END	false
FASTA	YES	FASTA	YES

Table 1. FluxSimulator's parameters used for the generation of the two synthetic RNA-seq datasets.

4.2.2 Real RNA-Seq datasets

Additionally to the synthetic data, we used a real dataset of small RNA-Seq produced using Illumina HiSeq 2500 technology on MDA-MB-231 breast cancer cell line, obtained from the Sequence Read Archive (SRA) (SRR5689212) [131]. This dataset contains RNA molecules shorter than 200 nucleotides, suitable for the evaluation of all types of small ncRNAs considered in this testing. A second RNA-Seq dataset on the same cell line obtained from GDC Legacy Archive (https://portal.gdc.cancer.gov/legacy-archive/files/0f5ba7d3-6f43-44af-9bbc-f9b4c09bbfeb) was used for lncRNAs and circRNAs evaluation.

4.2.3 Selected pipelines for the analysis of ncRNAs

The purpose of this testing was to evaluate the performances and usage of currently available pipelines, allowing the processing of more than one ncRNA class and not limited to a single ncRNA type. Thus, we compared eight ncRNAs pipelines published between 2015 and 2019, enabling the analysis of more than one ncRNA class, specifically: miRNAs, piRNAs, snoRNAs, tRNA-derived ncRNAs, lncRNAs and circRNA. Tools or pipelines specifically developed for the analysis of a single ncRNA class were not included in this testing. In Table 2, the classes of ncRNAs analyzed for each pipeline are reported. A brief description of the features of each pipeline is provided below.

	miRNA	piRNA	snoRNA	tRNA-derived	lncRNA	circRNA
iSRAP	X	X	X			
iSmaRT	X	X				
sRNApipe	X	X	X			
miARma-seq	X		X			X
sRNAnalyzer	X	X	X		X	
SPORTS1.0	X	X	X	X		
Oasis 2	X	X	X			
sRNA workbench	X					

Table 2. Classes of ncRNAs analyzed by each tested pipeline.

iSmaRT is a bioinformatics pipeline with a Graphical User Interface (GUI) for the analysis of miRNAs and piRNAs from small RNA-seq data [143]. *iSmaRT* enables a comprehensive analysis, including quality control, identification of miRNAs and piRNAs expressed in each sample, differential expression analysis, identification of RNA editing events on miRNA, RNA target prediction and Reactome Pathway Analysis (ReactomePA) [143].

iSRAP is a tool provided with a command-line interface (CLI) for the profiling of small RNAs (miRNAs, piRNAs, and snoRNAs) [144]. A YAML configuration file permits to define options and optimize small RNA profiling in different data sets [144]. The pipeline can be executed starting from either FASTQ or BAM alignment files. Results are reported as PDF files and HTML documents, completed with graphical elements to illustrate the results.

miARma-Seq is a pipeline enabling the identification and differential expression analysis of mRNAs, miRNAs, snoRNAs and circRNAs. It also allows for miRNA target prediction and the analysis of gene ontologies (GO) in any organism

with a sequenced genome available [145]. *miARma-Seq* comes with several preconfigured parameters and can be executed through a CLI [145]. It takes as input different files: raw FASTQ files, BAM alignment files or raw count txt files. Thus, the pipeline can start at different steps. The tool generates PDF documents as output. Boxplots and density plots of normalized and non-normalized data, multidimensional scaling (MDS) plots, principal component analysis (PCA) plots, heatmaps and clustering plots are also provided [145].

Oasis 2 is a web tool for the analysis of miRNAs, piRNAs, snRNAs, snoRNAs, and rRNAs [146]. It takes as input FASTQ or compressed FASTQ files. It performs the identification, quantification and differential expression analysis of all the ncRNAs classes mentioned above. The results are reported by means of text tables and plots. *Oasis 2* can perform adaptor removal and can be used with several reference genomes [146].

SPORTS1.0 is a CLI pipeline for the identification and quantification of miRNAs, piRNAs, snoRNAs and tRNA-derived ncRNAs [147]. It also allows for the analysis of rRNA small derived RNAs (rsRNA) [147]. *SPORTS1.0* can be used with a wide range of species with an available reference genome. It takes as input the following files: SRA data set, FASTQ and FASTA files. The output is provided as txt and PDF files, with annotation details for each sequence, length distribution along with other statistics and figures [147].

sRNAnalyzer is a CLI pipeline with a text-based configuration file for the identification of miRNAs, piRNAs, snoRNAs, and lncRNAs [148]. The pipeline allows

for processing the reads (upon adaptors removal), performing quality filters, read mapping and counting. It also filters the exogenous RNAs. For the endogenous sequences, *sRNAnalyzer* uses 'map and remove' structure (i.e., only unmapped reads go to the next steps), with a progressive alignment strategy to sequentially map the reads against various databases [148]. *sRNAnalyzer* can be used with samples from different species by appropriately modifying the configuration files. Currently, configuration files for human, mouse, rat, horse, macaque, and plant are available [148]. *sRNAnalyzer* takes as input FASTQ files and produces as output txt files describing the matches between mapped reads and the reference genome.

sRNApipe is a web-based pipeline, available on Galaxy, which performs small RNA mapping, counting, normalization, and analysis of signatures for ping-pong amplification in the case of piRNAs [149]. The pipeline allows for the identification of mRNAs, transposable elements, miRNAs, piRNAs, snRNAs, snoRNAs, rRNAs and tRNAs. *sRNApipe* takes as input single-end sequencing data in FASTQ format (Phred +33) with no adaptors and a list of FASTA reference files such as genome, mRNAs, transposable elements, rRNAs, tRNAs, snRNAs/snoRNAs and miRNAs [149]. In the 1.0 version only, the pipeline can be run without rRNAs, tRNAs and snRNAs/snoRNAs reference files. For the analysis of small ncRNAs in the synthetic dataset, the maximum read length, which by default is 29 nt, was set to 35 nt.

sRNA workbench is a pipeline reported to allow for the analysis of miRNAs and other small RNAs (sRNAs). It performs identification, quantification, normalization and differential expression analysis of miRNAs. The mapping of miRNAs and sRNA loci on the reference genome is also possible [150].

4.2.4 Pipeline comparison

The eight ncRNAs pipelines were installed on a machine equipped with Ubuntu 18.04 Operating System, Intel Core i7-6700 CPU (4 cores at 3.40 GHz), and 32 GB of RAM. The small RNA synthetic dataset was used to test *iSmaRT*, *iSRAP*, *Oasis 2*, *sRNApipe*, *SPORTS1.0*, and *sRNA workbench*. Instead, both small and standard RNA synthetic datasets were used to test *miARma-Seq* and *sRNAnalyzer*. To evaluate the performances of the different pipelines against the synthetic RNA-Seq datasets, we analyzed the true positive (TP), true negative (TN), false positive (FP), false negative (FN) rates, and then, we computed Precision, Sensitivity, and F-measure for each pipeline as follows:

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{FP + TP}$$

$$F - measure = \frac{2\,Precision * Sensitivity}{Precision + Sensitivity}$$

Moreover, we drew the scatterplots and calculated the R^2 on the TP to compare the ncRNAs expression profile identified by each pipeline with the real RNAs expression values included in the RNA-Seq dataset. Pipeline performances were also tested against a real RNA-Seq dataset. To estimate the similarity of the different pipelines in detecting the same ncRNAs, we calculated the Jaccard similarity coefficient between each couple of pipelines for all the analyzed small ncRNA classes as follows:

$$J(A, b) = \frac{|A \cap B|}{|A \cup B|}$$

where A and B are the subset of ncRNAs identified by the pipeline P_A and P_B, respectively. Next, to compare the read count estimation among the different tools, we

computed the Pearson correlation matrix considering per each ncRNA type the subset shared by all pipelines as follows:

$$\rho A, B = \frac{\sigma_{X_A, Y_B}}{\sigma_{X_A} \sigma_{Y_B}}$$

where x_A and y_B are the vectors of the expression values of ncRNAs identified by pipelines A and B, respectively, σ_{xA} is the covariance between the vectors x_A and y_B and σ_{xA} and σ_{xB} are the standard deviations of the two vectors. Statistical analysis has been performed using R (version 3.5.2), scatterplots, Jaccard index and Pearson correlation matrix were generated using the ggplot2 library [151].

4.3 RNAdetector design and implementation

Once the implementation of the database *tRFexplorer* [128] and the benchmark of previous non-coding RNA-Seq pipelines [129] were completed, we started the development of our software *RNAdetector*. A detailed description of its design and implementation follows below.

4.3.1 RNAdetector design

RNAdetector has been designed with the idea to be extremely easy-to-use, cross-platform, completely offline, remotely controllable, and with a dynamic workflow able to analyze coding and ncRNAs. Briefly, *RNAdetector* allows starting the analysis with different input files such as FASTQ, BAM, or SAM files. By using Trim Galore (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), the system allows also to perform quality trimming and adaptors removal on FASTQ files. Accordingly with the type of input file, alignment strategy, and the types of RNAs which users want to analyze (small ncRNAs, mRNAs and\or lncRNA, or circRNAs) the proper pipeline

will be executed. For mRNAs, small ncRNAs, and lncRNAs, the alignment can be executed on a reference genome by using STAR [62] / HISAT2 [63], or transcriptome by using SALMON [64]. Read counting can also be performed by choosing one of the several available algorithms such as HTseq [74], FeatureCounts [75], or SALMON [64]. Concerning circRNAs, reads are first mapped on the reference genome with BWA [152] and second they can be quantified (for circRNAs already annotated on circBase [141]) or de-novo identified and quantified by using CIRI 2 [153,154] or CIRIquant [155]. Finally, *RNAdetector* can perform normalization and differential expression analysis on such RNAs by using DESeq2 [77], edgeR [78], and LIMMA [76] algorithms and, in addition, topological pathway analysis on protein-coding genes and miRNAs can also be performed by using MITHrIL [93]. A final report based on metaseqR [156] with a summary, and interactive tables, and figures are automatically generated for an easier result interpretation obtained from the differential expression analysis module. Moreover, a second report has also been developed to show the results obtained from the optional pathway analysis module. Finally, an offline and interactive genome browser based on JBrowse 2 [137] is also available in order to visualize the depth of coverage of mapped reads obtained by *RNAdetector*. A summary of the pipeline is shown in Fig. 14.

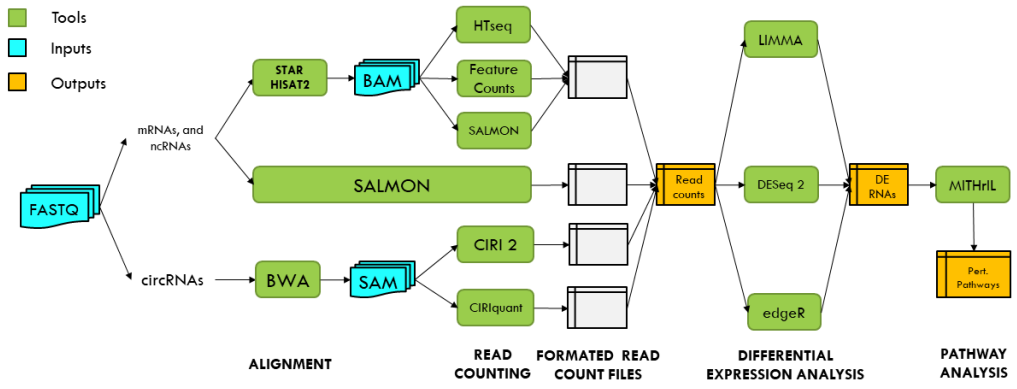

Fig 14. Schematic representation of RNAdetector's pipeline

After this short summary of RNAdetector's workflow, below follow a detailed description of each step of the pipeline.

4.3.1.1 Quality control and adaptors removing

If users start the analysis from FASTQ as input files, quality control on raw reads could be performed (if users select this option). Quality control for the raw reads involves the analysis of sequence quality, GC content, the presence of adaptors, overrepresented k-mers, and duplicated reads in order to detect sequencing errors, PCR artifacts, or contaminations. In addition, as a general rule, read quality decreases towards the 3' end of reads, and if it becomes too low, bases should be removed to improve mappability (trimming). Finally, if sequencing adaptors are still present on raw reads, they must be removed. For quality control analysis, trimming and adaptors removing two popular tools are FASTQC [54] and Cutadapt [56] which have been combined in a wrapper tool called Trim Galore (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). In *RNAdetector*, we included Trim Galore to perform an automated quality control and adapters trimming before to proceed with read alignment. For adapter trimming, Trim Galore uses the first 13 bp of Illumina standard adapters ('AGATCGGAAGAGC') by default (suitable for both ends of paired-end libraries) but accepts other adapter sequences, too. If it is not specified Trim Galore automatically detects the presence of possible adapters and proceeds with their trimming. The Phred quality of basecalls and the stringency for adapter removal can also be individually specified. Quality control, trimming, and adaptor removal can be performed on both single-end and paired-end FASTQ files.

4.3.1.2 Alignment to a reference genome or transcriptome

After that the quality control and adaptors removing have been performed, trimmed reads can now be aligned to a reference genome or transcriptome for their identification. In *RNAdetector*, we give users the opportunity to align reads on a reference genome using STAR [62], or HISAT 2 [157] or align reads on a reference transcriptome using SALMON [64]. Instead, for circRNAs analysis reads are mapped on the reference genome by using BWA [152]. The choice of one strategy over the other depends if the user wants to perform a gene-based or transcript-based analysis. In fact, it is known that protein-coding genes are not transcribed as a single transcript but many human genes go through a process called alternative splicing that allows a single gene to produce different transcripts which encode for different protein isoforms. Therefore, if users are interested in analyzing the expression profile of splicing variant transcripts, the alignment of the reads on a reference transcriptome is the suggested choice. On the contrary, if users are interested to summarize transcript expression at gene-level or they are interested to analyze small ncRNAs, the alignment of the reads on a reference genome is the suggested choice. For both approaches, *RNAdetector* stores in its remote repository human, mouse, and *C.elegans* indexed genomes and transcriptomes together with their relatives' custom GTF and FASTA files which can be downloaded directly from our repository by means of the user interface. Concerning the genome-based alignment, human (HG19 and HG38), mouse (mm9 and mm10) and C.elegans (ce11) genomes have been indexed by using STAR [62], HISAT2 [63], and BWA [152] and included in *RNAdetector*. Genome annotation for Human, Mouse, and C.elegans is also allowed by including several GTF files. Specifically, we included (1) GTF files with the genomic coordinates of protein-coding genes, snoRNAs, and lncRNAs retrieved from GENCODE for human and mouse (HG19 v19, HG38 v33, mm9 vM1,

mm10 vM26) and from ENSEMBL (ce11 WBcel235) for C.elegans (2) custom GTF files with the genomic coordinates of miRNAs (retrieved from miRBase [158]), piRNAs (retrieved from piRBase [140]), and tRNA-derived ncRNAs (retrieved from tRFexplorer [128] for human and from tRFdb [44] for mouse and C.elegans) (3) GTF files with the genomic coordinates of human, mouse and *C.elegans* circRNAs retrieved from circBase [141] (4) and a GTF file with the genomic coordinates of human t-UCRs retrieved from UCbase [159]. Genome-based alignment can be finally performed by using STAR [62], HISAT2 [157] (both for mRNAs, small ncRNAs, lncRNAs analysis) or BWA [152] (for circRNAs analysis). Concerning transcriptome-based alignment, *RNAdetector* has custom human, mouse, and C.elegans transcriptomes indexed by SALMON [64] which were built by retrieving the mRNAs and lncRNAs FASTA sequence from GENCODE for human and mouse (HG19 v19, HG38 v33, mm9 vM1, mm10 vM26) and from ENSEMBL (ce11 WBcel235) for *C.elegans*. Finally, transcriptome-based alignment is performed by aligning reads present in the FASTQ files on the indexed transcriptome using SALMON [64]. Although *RNAdetector* has been pre-built for human, mouse, and *C.elegans* RNA-Seq data analysis, other species can be analyzed by uploading their genomes and\or transcriptomes (in FASTA format) following the step-by-step procedure detailed in the user interface. In that case, fasta genomes and/or transcriptomes are automatically indexed by *RNAdetector* avoiding users to use any command-line tools for their indexing. Finally, additional ncRNA classes can be analyzed by providing their genomic coordinates in GTF or BED format.

4.3.1.3 Read quantification

After that reads have been aligned to a reference genome or transcriptome, they have to be quantified in order to infer protein-coding genes and\or ncRNAs expression

levels. For this purpose, *RNAdetector* allows users to select among several tools and options to perform the read quantification step. Specifically, If users choose the genome-based alignment strategy read quantification can be executed by choosing HTseq [74], FeatureCounts [75], or SALMON [64] (alignment-based mode). Instead, if users choose the transcriptome-based alignment strategy, reads are aligned and quantified by SALMON [64] in a single step for a faster and RAM memory saving analysis. Concerning circRNAs, a different workflow is executed for their identification and quantification. Precisely, reads are first mapped on the reference genome by BWA [152] and then they can be quantified (for circRNAs already annotated on circBase [152]) or de-novo identified and quantified by using CIRI 2 [153,154] or CIRIquant [155].

4.3.1.4 Differential expression analysis

Once the read quantification step is performed, an optional step could be to perform a differential expression analysis in order to identify which mRNAs and\or ncRNAs are differentially expressed in a case vs control study. For this purpose, we included in *RNAdetector* three of the most common tools for the differential expression analysis such as DESeq2 [77], edgeR [78], and LIMMA [76]. As we discussed in the introduction, these three methods use different assumptions, normalization methods, and statistics to identify differentially expressed genes. Therefore, they can produce different results from the same datasets. However, we included these three different methods in order that users can choose the most suitable tool for their analysis. In addition, if users want to perform a more stringent analysis, it is also possible to perform the analysis by using a combination of these 3 methods or all of them and get only the differentially expressed genes and\or ncRNAs that are in common.

4.3.1.5 miRNA-sensitive topological pathway analysis

A final optional step of *RNAdetector* is the pathway analysis. Pathway analysis could be very helpful to understand the effect of the differentially expressed genes on metabolic and signaling pathways, allowing a more comprehensive knowledge of the biological mechanisms which are in place in the samples under study. For this purpose, we have included MITHrIL [93] in *RNAdetector*. MITHrIL is a topology-based and miRNA-sensitive pathway analysis system that allows for the identification of perturbed metabolic and signaling pathways starting from the LogFC values obtained by the differential expression analysis step. As all the other topology-based pathway analysis systems, MITHrIL fully exploits the topological information encoded by pathways when computing perturbation scores but in addition to the other methods, MITHrIL uses the information of validated miRNA-mRNA interaction to infer the impact of miRNAs on pathways. Pathways are then modeled as complex graphs where each node is a gene or miRNA and each edge is an interaction between them. Even though thousands of genes are not annotated in pathways and existing annotations may be inaccurate, graphs contained in these databases provide a more detailed view of biological processes within the cell, helping the interpretation of high-throughput experiments [90].

4.3.2 Implementation and software architecture

RNAdetector is a client-server application developed to simplify deployment and usage. The server has been developed in PHP, Bash, and R. All server code and dependencies are deployed through a Docker container for easy installation. Communication between client and server is based on an HTTP REST API specifically developed for *RNAdetector*. An internal Mysql database is used to store

all server data. Authentication, API Security, and the data abstraction layer has been provided by the Laravel framework (https://laravel.com/). The Graphical User Interface (GUI) has been developed in Javascript using the Electron framework (https://electronjs.org/). Electron is an open-source framework developed and maintained by GitHub, allowing the development of desktop GUI applications using web technologies. Indeed, *RNAdetector* is completely offline and it can be used as a desktop application with several operating systems such as Windows Professional, macOS, and Linux. *RNAdetector* is distributed as a Docker container in order to guarantee an easy deployment and dependencies management and it can also be installed in servers and remotely controlled by installing RNAdetector's app in laptops and\or tablets. To install *RNAdetector*, it is only necessary to install Docker in users' machines and then download the installer specific for the user's operating systems from our repository. After that, users have only to follow the instructions on the installation wizard to authorize the installer and proceed with the installation. A detailed explanation of how to install Docker and *RNAdetector* on Windows Professional, macOS, and Linux machines is reported at the following link https://github.com/alessandrolaferlita/RNAdetector/wiki/Requirements-and-Setup. *RNAdetector* is an open-source tool and it is available for download at https://rnadetector.atlas.dmi.unict.it/download.html. Source code and issue reporting is available at https://github.com/alessandrolaferlita/RNAdetector.

4.3.3 Case study analysis

In order to show an example of *RNAdetector's* analysis, we chose a public small RNA-Seq project available on NCBI SRA (SRP183064) and we performed a complete analysis identifying the differentially expressed small ncRNAs and the impacted

biological pathways. Precisely, we used very recent small RNA-Seq datasets of Colon Rectal Cancer (CRC) [160] and we compared the expression profiles of the CRC samples against the adjacent normal tissue samples of the same patients in order to identify the differentially expressed miRNAs, snoRNAs, and tRNA-derived ncRNAs. We started the analysis from the FASTQ files, raw reads were trimmed and adapters were removed by selecting Trim Galore from the user interface. Trimmed reads were then aligned to the reference human genome (HG38) and counted by selecting from the user interface HISAT 2 [157] and featureCounts [75] respectively. Prior to the statistical testing procedure, the read counts were filtered for possible artifacts that could affect the subsequent statistical testing procedures. After that, the count table was normalized for inherent systematic or experimental biases selecting edgeR [78] from the user interface as a normalization method after removing features that had zero counts over all the RNA-Seq samples. The normalized count matrix was then used for the differential expression analysis by selecting limma [76] and edgeR [78] from the RNAdetector's user interface. Finally, in order to combine the statistical significance from multiple algorithms and perform a meta-analysis, the Simes correction and combination method was applied. Concerning the pathway analysis, it was performed by selecting the MITHrIL algorithm [93] which used the LogFC values of miRNAs obtained from the differential expression analysis step for its analysis. Pathways with FDR or adjusted p-values < 0.01 were considered impacted.

5. Results

5.1 tRNA-derived ncRNAs and their novel database tRFexplorer

Our software, *RNAdetector*, aims to analyze an extensive repertoire of different classes of coding and ncRNAs including also all the different tRNA-derived ncRNAs subclasses. However, for the latter, no previous system for their detection was released when the project started in 2018. In addition, at that time there were not extensive databases that covered all the different subclasses (already available databases such as tRFdb [44] and MINTbase [127] were primarily focused on tRFs deriving from mature tRNAs). Therefore, we had to find a strategy to detect all the different tRNA-derived ncRNAs subclasses from small RNA-Seq datasets and include them in a novel database. Finally, this database will be used to collect all the tRNA-derived ncRNAs that it will be possible to analyze with *RNAdetector*. For this reason, we selected the public small RNA-Seq datasets of the NCI-60 cancer cell lines and TCGA in order to identify all the expressed tRNA-derived ncRNAs and we included them in a novel database called *tRFexplorer* [128]. The results of the analysis and the presentation of *tRFexplorer* are described in more details in the next sections (these results are also reported in our paper titled *Identification of tRNA-derived ncRNAs in TCGA and NCI-60 panel cell lines and development of the public database tRFexplorer* published in 2019 [128]).

5.1.1 tRNA-derived ncRNAs in NCI-60 cell lines and TCGA samples

In order to identify the human tRNA-derived ncRNAs, we assessed the expression of tsRNAs, 5' leader RNAs, and tRFs from small RNA-Seq datasets of the NCI-60 cell lines [131] and TCGA samples. In these datasets, we were able to identify *322* expressed tRNA-derived ncRNAs in NCI-60 (11 tRF-5s, 55 tRF-3s, 107 tsRNAs, and 149 5' leader RNAs) and 232 expressed tRNA-derived ncRNAs (53 tRF-5s, 58 tRF-3s, 63 tsRNAs, and 58 5' leader RNAs) in TCGA. A number of tsRNAs, 5' leader RNAs, and tRFs identified across NCI-60 cell lines and TCGA samples present noticeable expression levels. Moreover, all small RNA sequences mapped within 4 specific regions: 5' end (tRF-5), 3' end (tRF-3) of mature tRNA, and 3' trailer (tsRNA) and 5' leader (5' leader RNAs) regions of primary tRNA genes. If these small RNA sequences were the result of a random degradation process, their ends would be equally distributed along the lengths of tRNA genes with a comparable frequency [44,45]. In addition, we can observe that each TCGA cancer type (whose control samples are available) displays a different pattern of dysregulated tRNA-derived ncRNAs. Taken together, these results may suggest that these small RNAs are not fragments derived from the random cleavage of precursor and mature tRNAs, rather they are actively expressed and produced by specific ribonucleases and may be dysregulated in several human cancers. Indeed, recent evidence has shown dysregulated tRNA-derived ncRNAs in Chronic lymphocytic leukemia (CLL), colon, breast, ovary, lung, and prostate cancers [30,31,33,161–163].

5.1.2 tRFexplorer

All identified tRNA-derived ncRNAs have been integrated into a novel database named *tRFexplorer* [128]. *tRFexplorer* is an easy-to-use, web-based database (https://trfexplorer.cloud/) containing tRNA-derived ncRNAs expression profiles for NCI-60 cell lines and TCGA samples, together with all omics and compound activities data available on CellMiner. Leveraging CellMiner data, *tRFexplorer* enables users to perform correlation analysis inferring knowledge on the biological function of such molecules [128]. Furthermore, a module allowing DE analysis for all tRNA-derived ncRNAs in TCGA samples has been released [128]. A detailed explanation of *tRFexplorer* functions is provided in the following sections.

5.1.2.1 Browse

In the "Browse" section, users can search for tsRNAs, 5' leader RNAs, and tRFs by "location" or "expression". Browsing by location enables users to search and visualize all tRNA-derived ncRNAs in the reference human genome (Fig. 15). Specifically, through the custom genome browser, it is possible to interactively search tRNA-derived ncRNAs either by genomic coordinates or by identifier.

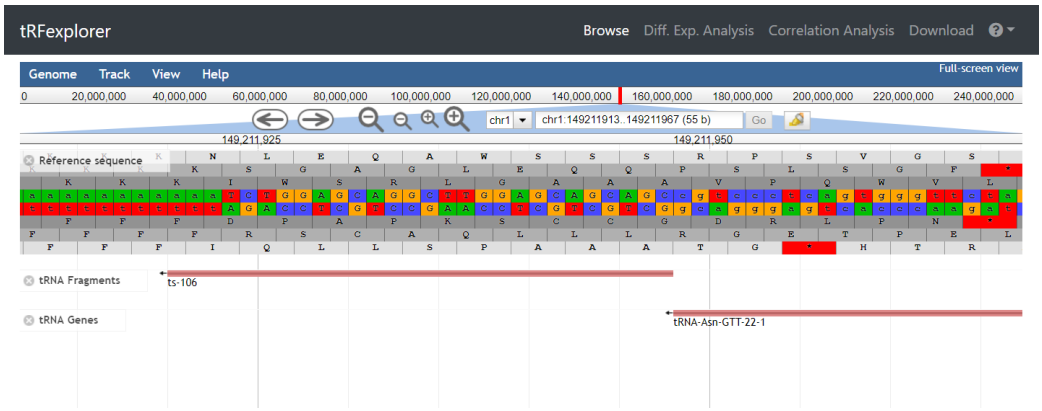

Fig 15. Genome browser available in tRFexplorer

Browsing by expression section enables users to filter data using at least one of the following options: (i) the type of fragment (tRF-3, tRF-5, tsRNA, and 5' leader RNAs); (ii) the amino acid carried by the precursor tRNA; (iii) the anticodon sequence; (iv) the dataset in which the fragment is expressed (TCGA tumor types or NCI-60 cell lines); (v) the tissue subtype (normal, tumor, metastatic, recurrent, etc). It is also possible to set a minimum RPM threshold for tRNA-derived ncRNAs. The search procedure will scan our database looking for all tRNA-derived ncRNAs matching users' criteria, and the results will be reported in a table. Once results become available, users may view a page with detailed information by selecting any single result. Such a page will show plots for assessing RPM expression levels in both NCI-60 (Fig. 16), and TCGA (Fig. 17). A genomic viewer will show genomic locations for each tRNA-derived fragment.

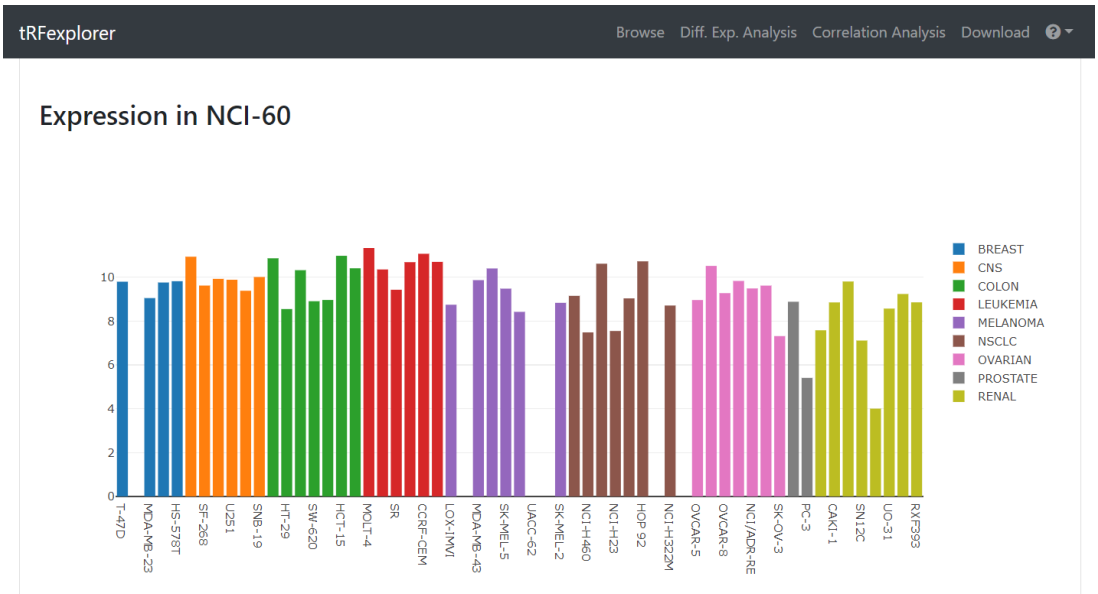

Fig 16. Expression of a tsRNA across all the NCI-60 cell lines

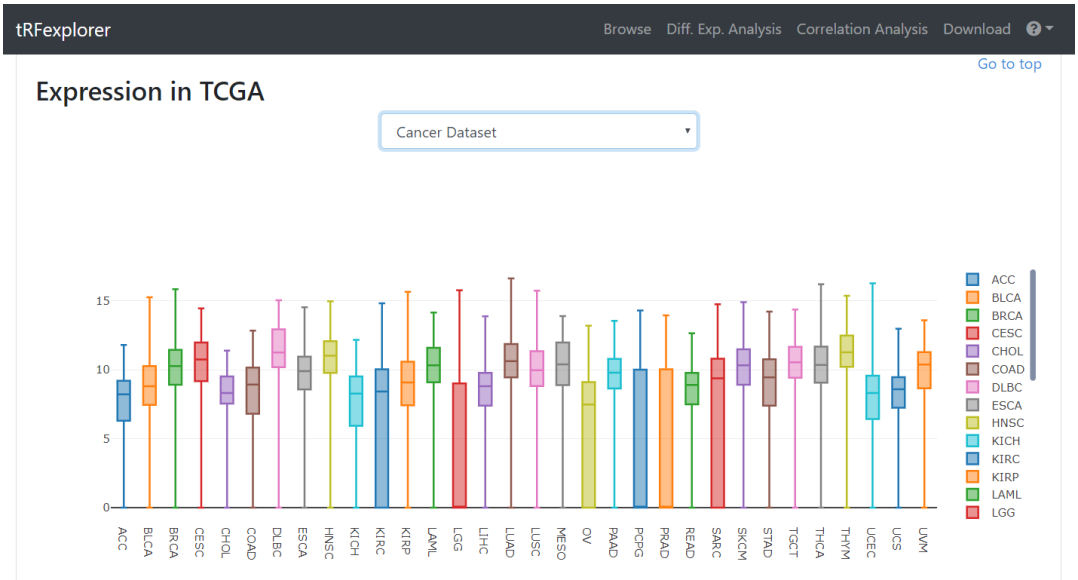

Fig 17. Expression of a tsRNA across all the TCGA cancer types

5.1.2.2 Correlation Analysis

In the "Correlation Analysis" section, users can perform correlation analyses of all identified tsRNAs, 5' leader RNAs, and tRFs in NCI-60, with the omics and compound activities data available on CellMiner [135]. Correlation analysis can also be performed with mRNA/miRNA expression profiles, as well as patient survival data, of TCGA samples. Specifically, the user selects the correlation measure (Pearson or Spearman) and which dataset to consider. A list of correlated and anticorrelated tRNA-derived ncRNAs will be shown. The results can then be filtered by: (i) ncRNA name; (ii) genes, miRNAs, compound names; (iii) the genomic coordinates, when available; (iv) the minimum correlation value. By clicking on each result, an interactive scatter plot with the data of the selected molecules will appear (Fig. 18).

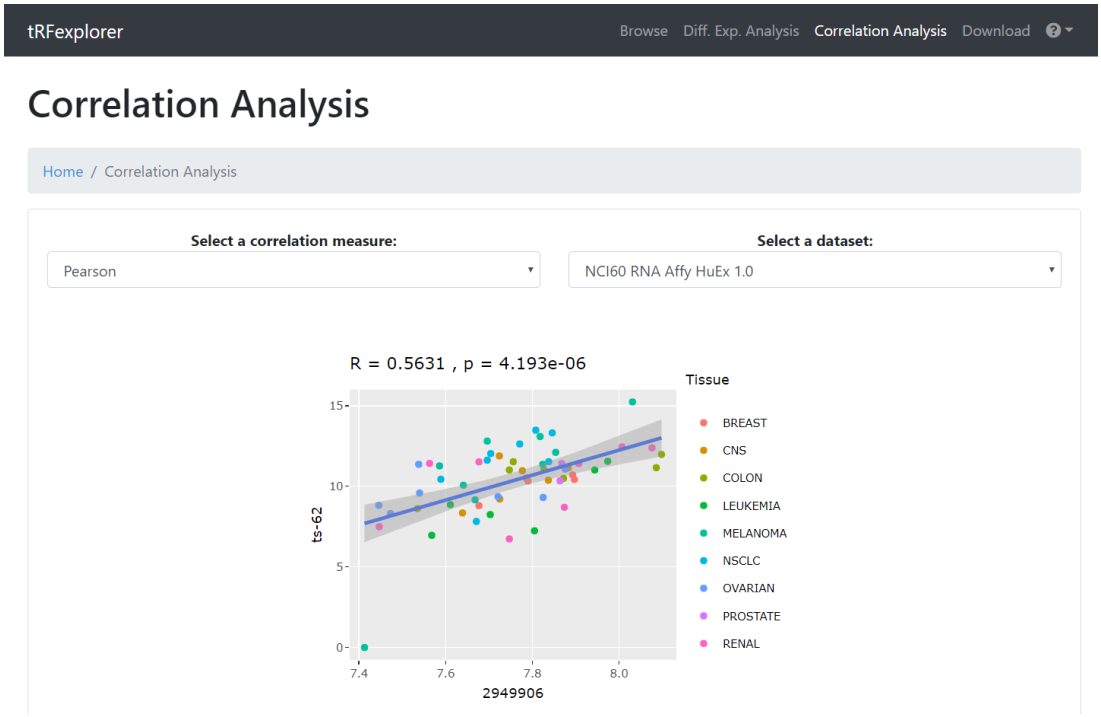

Fig 18. Scatter plot which shows the correlation between the expression of ts-62 in the NCI-60 cell lines and the expression of the mRNAs in the same cell lines

5.1.2.3 Differential expression analysis

In the "differential expression analysis" section, users can perform DE analyses to discover which tsRNAs, 5' leader RNAs and tRFs are dysregulated in TCGA tumor types. To start the analysis, users select the cancer type and one of the available covariates (gender, race, vital status, sample type, or classification). It is also possible to set the maximum p-value and minimum log-fold-change (logFC) for the analysis. After selecting all the parameters, users must select at least one contrast to perform for the DE analysis, in association with the selected covariate. Once the analysis is launched, a list of differentially expressed tRNA-derived ncRNAs with their logFC and FDR adjusted p-value will be shown together with an interactive volcano plot to better visualize their differential expression (Fig. 19). By clicking on a specific point in the plot or row in the table, a swarm plot of the expression values will be also shown (Fig. 19).

Fig 19. Example of a volcano plot and swarm plot generated by tRFexplorer after the differential expression analysis performed on TCGA samples

5.2 Evaluation of pre-existing ncRNA pipelines

After the implementation of *tRFexplorer* and before proceeding with the development of *RNAdetector*, we wanted to evaluate the state of the art of current ncRNA pipelines in order to identify their strengths and weaknesses, and therefore, optimize *RNAdetector* by filling the gaps of previous methodologies. For this purpose, the performance of eight ncRNA pipelines enabling the processing of RNA-Seq data, published between 2015 and 2019, were compared. In particular, we evaluated the easiness of installation and usage together with their accuracy to identify ncRNAs and their expression levels by using both synthetic and real RNA-Seq datasets. The results of this benchmark are described in the next sections (these results are also reported in our paper titled *A benchmarking of pipelines for detecting ncRNAs from RNA-Seq data* published in 2019 [129]).

5.2.1 Installation and usage

Installation, usage easiness and flexibility are crucial characteristics of bioinformatics pipelines. Moreover, these characteristics became even more important for the distribution of the application to non-expert users. Thus, we evaluated a number of features which might influence user experience, such as: (i) setup process, (ii) amount and quality of the documentation, (iii) presence of a GUI, (iv) possibility of using different input file formats, (v) possibility of analyzing more than one class of ncRNAs in a single run, (vi) pipeline flexibility, and (vii) output file formats (txt, pdf, image, etc.). In Table 3, we have reported a schematic evaluation of the main features of each pipeline. A summary of the criteria used for their evaluation is also reported in the supplementary table 4.

	iSmaRT	iSRAP	miARma-Seq	Oasis 2	SPORTS	sRNAnalyzer	sRNApipe	sRNA workbench
Installation	++	-	++	++	-	+	+	++
Documentation	++	++	++	++	+	+	+	++
GUI	++	-	-	++	-	-	+	++
Different Input Types	-	+	++	-	+	-	-	-
Report generation	++	++	++	++	++	+	++	-
Multi-ncRNAs in single analysis	-	++	-	++	+	++	+	-
Flexibility	-	+	++	-	-	-	-	-
Usability\configuration	++	+	++	++	-	+	++	++

Table 3. Some features possessed by the tested ncRNAs pipelines and their evaluation.

iSmaRT is provided with a comprehensive documentation, including several examples, which guide the user in each step of the installation process and pipeline use. A unique file is provided for installation and configuration. The tool comes with a user-friendly GUI developed in Python. *iSmaRT* takes as input FASTQ files only and does not allow for starting the analysis from alignment files. Moreover, users cannot choose among several tools and options for each step of the pipeline. Different

ncRNA classes cannot be analyzed together in a single run. Output is provided as txt file and tiff plots.

iSRAP requires manual installation of dependencies. The documentation is adequate to describe the tool usage. The configuration file is well structured, providing users with the flexibility of selecting at which step initiating the analysis. *iSRAP* takes as input both FASTQ and BAM files. The output is organized in independent folders, one for each step, containing txt and pdf files. *iSRAP* is a CLI tool; therefore, no GUI is provided.

miARma-Seq comes with a modular configuration file, permitting users to select which step performs. Indeed, it accepts FASTQ, BAM and txt files with raw counts. *miARma-Seq* guide is organized as a tutorial, with an exhaustive documentation covering the major features of the pipeline. Although *miARma-Seq* can identify miRNAs, snoRNAs and circRNAs, it is not able to identify circRNAs together with miRNAs and snoRNAs in a single run. It is not equipped with a GUI.

Oasis 2 is a web-based tool. No installation of the software or of dependencies is required. It is provided with a GUI which makes the tool very user friendly. The setting of few parameters, such as the reference genome and specification of which adaptors have to be removed, is required. However, *Oasis 2* has a very strict workflow since users cannot select among different tools and options for each step of the pipeline. *Oasis 2* user guide is provided as a video tutorial or PDF file. The output is organized in several folders. Results are reported as txt tables and plots. *Oasis 2* can

process simultaneously several classes of ncRNAs (miRNAs, piRNAs and snoRNAs) in a single run.

SPORTS1.0 has no setup scripts, therefore, all the parameters must be passed directly in the CLI. The provided documentation lists all the dependencies and the configuration necessary for users' machine setup. The workflow is strict, predefined and customizability is not allowed. Outputs are provided as pdf documents for each ncRNAs class which can be analyzed in the same run. Moreover, a txt file summarizes the result. *SPORTS1.0* does not have a GUI.

sRNAnalyzer comes with a YAML configuration file, which permits the users to customize pre-processing and alignment options. A separate configuration file lists the paths for internal database setup. The workflow is very strict since users cannot choose among several tools and options for each step of the pipeline. The documentation is modest and only lists the required dependencies for the setup. The output is organized in profiles and features txt files. *sRNAnalyzer* does not have a GUI.

sRNApipe requires a Galaxy Server installed on the users' machine. Galaxy handles dependencies' installation. The documentation is exhaustive regarding the pipeline usage; however, it is based on an old release. *sRNApipe* is very easy-to-use: the users only need to upload FASTQ and reference FASTA files. The output is organized in html format with text and plots. However, *sRNApipe* has a static and predefined workflow and users cannot choose which tools and parameters must be used for the alignment and read counting steps.

sRNA workbench can be used with both CLI and GUI. It is provided with a comprehensive documentation (PDF manuals and video tutorials), available through the website (http://srnaworkbench.cmp.uea.ac.uk/). *sRNA workbench* has been developed in Java. It requires the installation of one single dependency (Java FX). The tool can be launched by a jar file. *sRNA workbench* requires FASTQ or FASTA as input files. It is not possible to start the analysis at different steps of the pipeline. Although, in the *sRNA workbench* website, it is claimed that it can perform the analysis of sRNA, we only found pipelines for the identification and quantification of miRNAs. The mapping of miRNAs and sRNA loci on the reference genome is also possible. Finally, *sRNA workbench* presents a static and predefined workflow.

5.2.2 Pipeline accuracy on synthetic datasets

The identification of the correct RNA molecules and their quantification is one of the crucial tasks which an RNA-Seq analysis pipeline should accomplish. To test the ability of the different pipelines in recovering ncRNAs, we prepared two synthetic FASTQ files using Flux Simulator [138]. Specifically, we built a FASTQ file simulating small RNA-Seq data (miRNAs, snoRNAs, piRNAs and tRNA-derived ncRNAs) and a FASTQ file for standard RNA-Seq data (mRNAs and lncRNAs). For each ncRNA class, we assessed the ability of the pipelines to identify the correct RNA molecules in terms of Precision, Sensitivity, and F-measure (Table 4). To determine the accuracy in expression profile measure, we computed the R^2 coefficient between the real counts and the pipeline-quantified ones. Per each tool, we reported the scatterplots to visualize the relationship between real counts and predicted ones.

	Precision	Sensitivity	F-measure	
iSmaRT	0.76	0.98	0.85	**miRNA**
iSRAP	0.84	1.00	0.91	
miARma	0.81	0.96	0.88	
Oasis 2	0.36	0.92	0.52	
SPORTS1.0	0.90	0.97	0.93	
sRNAnalyzer	0.18	0.98	0.30	
sRNApipe	0.86	0.92	0.89	
sRNA workbench	0.82	0.79	0.81	
iSmaRT	0.37	0.73	0.49	**piRNA**
iSRAP	0.17	1.00	0.30	
Oasis 2	0.19	0.72	0.30	
sRNAnalyzer	0.71	0.89	0.79	
iSRAP	0.52	1.00	0.68	**snoRNA**
miARma	0.58	0.99	0.73	
Oasis 2	0.37	0.65	0.47	
sRNAnalyzer	0.81	0.95	0.87	
sRNAnalyzer	0.25	0.99	0.40	**lncRNA**

Table 4. Statistics obtained for each ncRNAs pipeline by using our synthetic RNA-seq dataset.

miRNAs. Our synthetic small RNA-seq data set comprises 193 miRNA sequences, which were analyzed by all eight pipelines. Using this dataset, we calculated Sensitivity, Precision and F-measure values for each pipeline, as summarized in Table 4. The scatterplots and the R^2 computed on the TPs between miRNAs expression values identified by each pipeline and the real counts present in the simulated data set are reported in Figure 20. *SPORTS1.0* is the pipeline accomplishing the best performance in detecting miRNAs, followed by *iSRAP*, *sRNApipe*, *miARma-Seq*, *iSmaRT*, *sRNA workbench*, *Oasis 2* and *sRNAnalyzer* (Table 4), while *sRNA workbench* (R2 = 0.96), *SPORTS1.0* (R2 = 0.96) and *iSRAP* (R^2 = 0.96) are the most accurate tools in read count estimation, followed by *sRNApipe* (R^2 = 0.94), *miARma-Seq* (R^2 = 0.94), *iSmaRT* (R^2 = 0.58), *sRNAnalyzer* (R^2 = 0.52) and

Oasis 2 (R^2 = 0.38) (Fig. 20). It is noteworthy that all pipelines share a common set of FP miRNAs. Although these false positives are covered by a high number of read counts, there are specific biological reasons which can explain these unexpected results. Indeed, miR-4521 has been recently re-annotated as a tRNA-derived small RNAs (tsRNAs), specifically ts-101 [30]. ts-101 is present in our synthetic small RNA-seq data set, so it was correctly identified by *sRNAnalyzer*, *SPORTS1.0*, *Oasis 2*, *iSmaRT*, and *sRNApipe* although as a miRNA. Similarly, miR-3182 and miR-6516 were identified by a significant number of counts by *iSmaRT*, *miARma-Seq*, *Oasis 2*, *SPORTS1.0* and *sRNAnalyzer* because they share sequence similarity with the tRNA-fragment (tRF) tRFdb-5026a (reported in tRFdb database [44]) and the snoRNA ACA 47, respectively, both contained in our simulated dataset. In addition, miR-214, miR-522, miR-550b and miR-103b were also identified as FP miRNAs with high counts. This artifact could be explained by performing a blastn analysis (version 2.6.0+ [164]). These miRNAs showed sequence identity (>= 91%) with the following miRNAs present in our dataset: miR-3120, miR-519a, miR-550a and miR-103a, respectively.

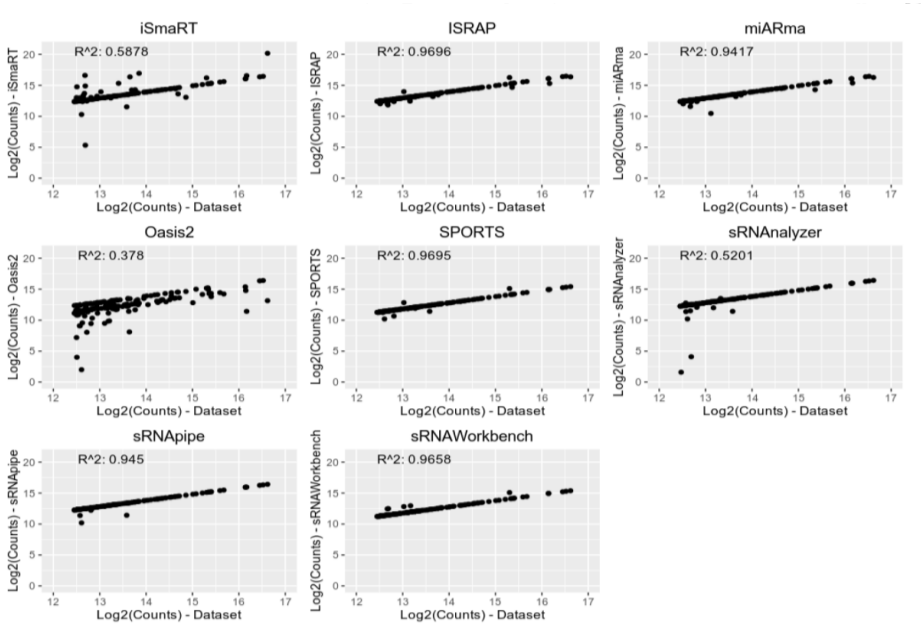

Fig 20. Scatterplots and the R^2 computed on the TPs between miRNAs expression values identified by each pipeline and the real counts present in the simulated dataset.

piRNAs. To determine the accuracy of the pipelines for piRNAs detection, we evaluated their ability to detect the 500 different piRNA sequences contained in our synthetic data set. Six of eight tested pipelines (*iSmaRT, iSRAP, Oasis 2, SPORTS1.0, sRNAnalyzer* and *sRNApipe*) are reported to allow for the identification of piRNAs. However, *SPORTS1.0* does not annotate piRNA sequences, thus, it just reports the total number of piRNA mapped reads without detailing the results. *sRNApipe* instead identifies the piRNA sequences mapping on transposable elements (TE) and protein-coding genes only, and it does not report which are the identified molecules in the output. Therefore, we could only consider four of six ncRNAs pipelines for piRNA comparison. The pipeline showing the best performance in terms of Sensitivity, Precision and F-measure in piRNAs detection is *sRNAnalyzer*, followed by *iSmaRT*, *iSRAP* and *Oasis 2* (Table 4). In terms of read count accuracy, the best performing one is *iSRAP* (R^2 = 0.83), followed by *iSmaRT* (R^2 = 0.66), *Oasis 2* (R^2 = 0.28) and *sRNAnalyzer* (R2 = 0.08) (Fig. 21).

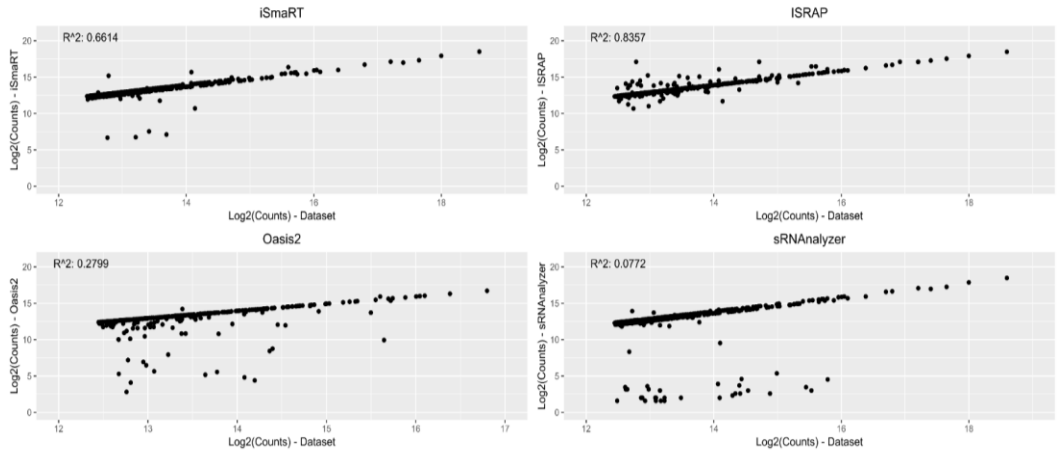

Fig 21. Scatterplots and the R^2 computed on the TPs between piRNAs expression values identified by each pipeline and the real counts present in the simulated dataset.

snoRNAs. To determine the accuracy in snoRNAs detection, we used 100 different snoRNA sequences present in our synthetic dataset. Six of eight tested pipelines (*iSRAP, miARma-Seq, Oasis2, SPORTS1.0, sRNAnalyzer, sRNApipe*) are

claimed to detect snoRNAs. However, also in this case, *SPORTS1.0* does not annotate snoRNA molecules, providing a summarization of the total number of snoRNAs mapped reads, which does not allow for a comparison, while sRNApipe could not detect any of the 100 snoRNAs present in the synthetic dataset. Thus, we could calculate the statistics for *iSRAP*, *miARma-Seq*, *Oasis 2* and *sRNAnalyzer* only. The pipeline with the best performance in terms of Sensitivity, Precision and F-measure is *sRNAnalyzer*, followed by *miARma-Seq*, *iSRAP* and *Oasis 2* (Table 4), while for read counts estimation, the best is *iSRAP* ($R^2 = 0.81$) followed by *miARma-Seq* ($R^2 = 0.73$), *sRNAnalyzer* ($R^2 = 0.50$) and *Oasis 2* ($R^2 = 0.04$) (Fig. 22).

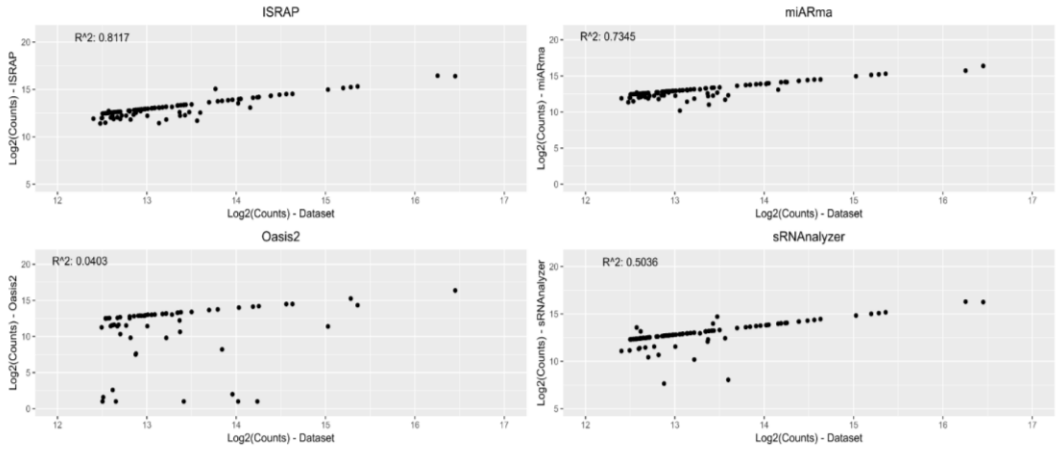

Fig 22. Scatterplots and the R^2 computed on the TPs between snoRNAs expression values identified by each pipeline and the real counts present in the simulated dataset.

tRNA-derived ncRNAs. *SPORTS1.0* is the only pipeline reported to allow for tRNA-derived ncRNAs analysis. Although *SPORTS1.0* can identify reads mapped on tRNA genes, it does not annotate their specific type and just reports the number of mapped reads for each tRNA gene. tRNA-derived ncRNAs are typically classified according to their origin within the tRNA gene and belong to two main classes: (i) tsRNA, arising from pre-tRNA; (ii) tiRNAs and tRFs, deriving from mature tRNA [25]. tRF can be further classified in tRF-5 and tRF-3 according to the ribonuclease cleavage site within mature tRNA D-loop or T-loop, respectively [25]. This annotation is

commonly used in tRNA-derived ncRNAs databases, such as tRFdb [44], MINTbase [127] and also in our *tRFexplorer* [128]. Therefore, we used this annotation for our small RNA-seq simulated dataset. Specifically, we selected 50 5' leader tsRNAs, 50 3' trailer tsRNAs, 50 tRF-5 and 50 tRF-3. *SPORTS1.0* did not annotate the specific types of tRNA-derived ncRNAs, but it only retrieved the number of mapped reads for each tRNA gene. For this reason, it was not possible to establish which tRNA-derived ncRNAs present in our synthetic data set were detected by the pipeline and an accurate performance evaluation could not be performed.

lncRNAs. Among the tested pipelines, *sRNAnalyzer* is the only one reported to analyze lncRNAs. To evaluate its performance, we selected 500 different lncRNA sequences from our synthetic long RNA-seq dataset. *sRNAnalyzer* identified lncRNAs with high Sensitivity (0.99) and low precision (0.25) (Table 4). *sRNAnalyzer* can efficiently estimate the lncRNAs expression profile ($R^2 = 0.96$) (Fig. 23).

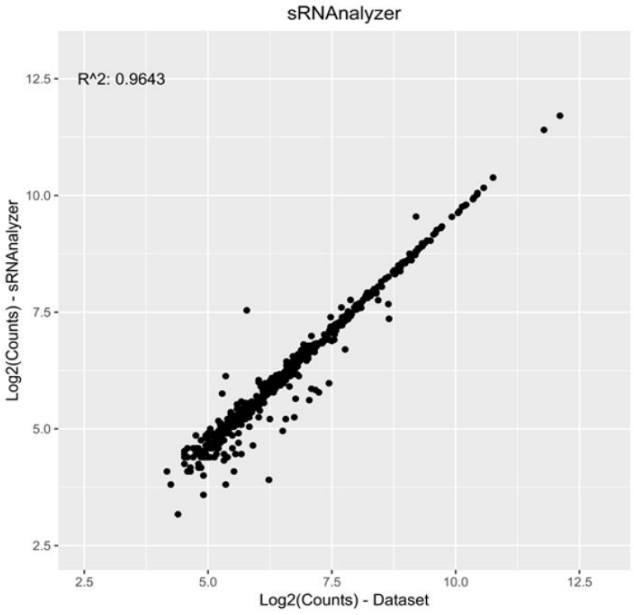

Fig 23. Scatterplot and the R^2 computed on the TPs between lncRNAs expression values identified by sRNAnalyzer and the real counts present in the simulated dataset.

5.2.3 Pipeline similarity and correlation on real datasets

To go beyond the limited complexity of the RNA species present in a synthetic dataset, we also performed a comparative analysis using real data. Specifically, we selected a small RNA-Seq dataset retrieved from Sequence Read Archive (SRA) (SRR5689212), which belongs to a breast cancer cell line (MDA-MB-231) of the NCI-60 panel [131]. This dataset covers RNA molecules shorter than 200 nucleotides, thus including all the small ncRNAs analyzed in this testing. In addition, to cover lncRNAs and circRNAs, we used another RNA-Seq dataset from GDC (https://portal.gdc.cancer.gov/legacy-archive/files/0f5ba7d3-6f43-44af-9bbc-f9b4c09bbfeb), covering the same breast cancer cell line. To evaluate similarities and differences in ncRNAs identification among the different pipelines, we used the Jaccard similarity coefficient between each couple of pipelines for all small ncRNAs classes assessed in this testing (Fig. 24). Next, we calculated the Pearson correlation matrix on the common small ncRNAs identified by each pipeline to establish their ability in estimating read counts.

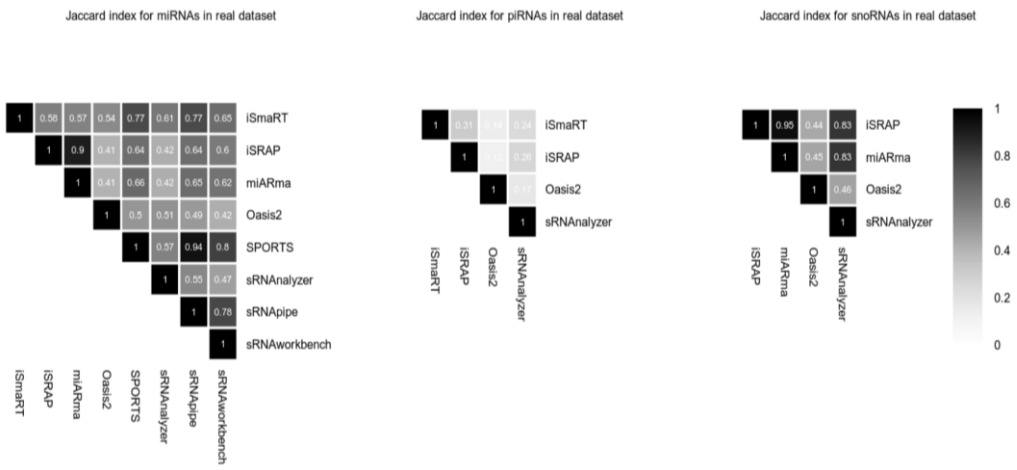

Fig 24. Jaccard similarity matrix for small ncRNAs identification

miRNAs. Concerning miRNAs identification, we observed a high similarity between *SPORTS1.0* and *sRNApipe* (J = 0.94), *iSRAP* and *miARma-Seq* (J = 0.9) and,

to a lesser extent, *SPORTS1.0* and *sRNA workbench* (J = 0.8), *sRNApipe* and *sRNA workbench* (J = 0.78), *iSmaRT* and *sRNApipe* (J = 0.77) and *iSmaRT* and *SPORTS1.0* (J = 0.77) (see Jaccard similarities in Fig. 24). Pearson correlation matrix calculated among miRNAs identified by all the pipelines also showed high correlations in read count estimation for all tools (Fig. 25).

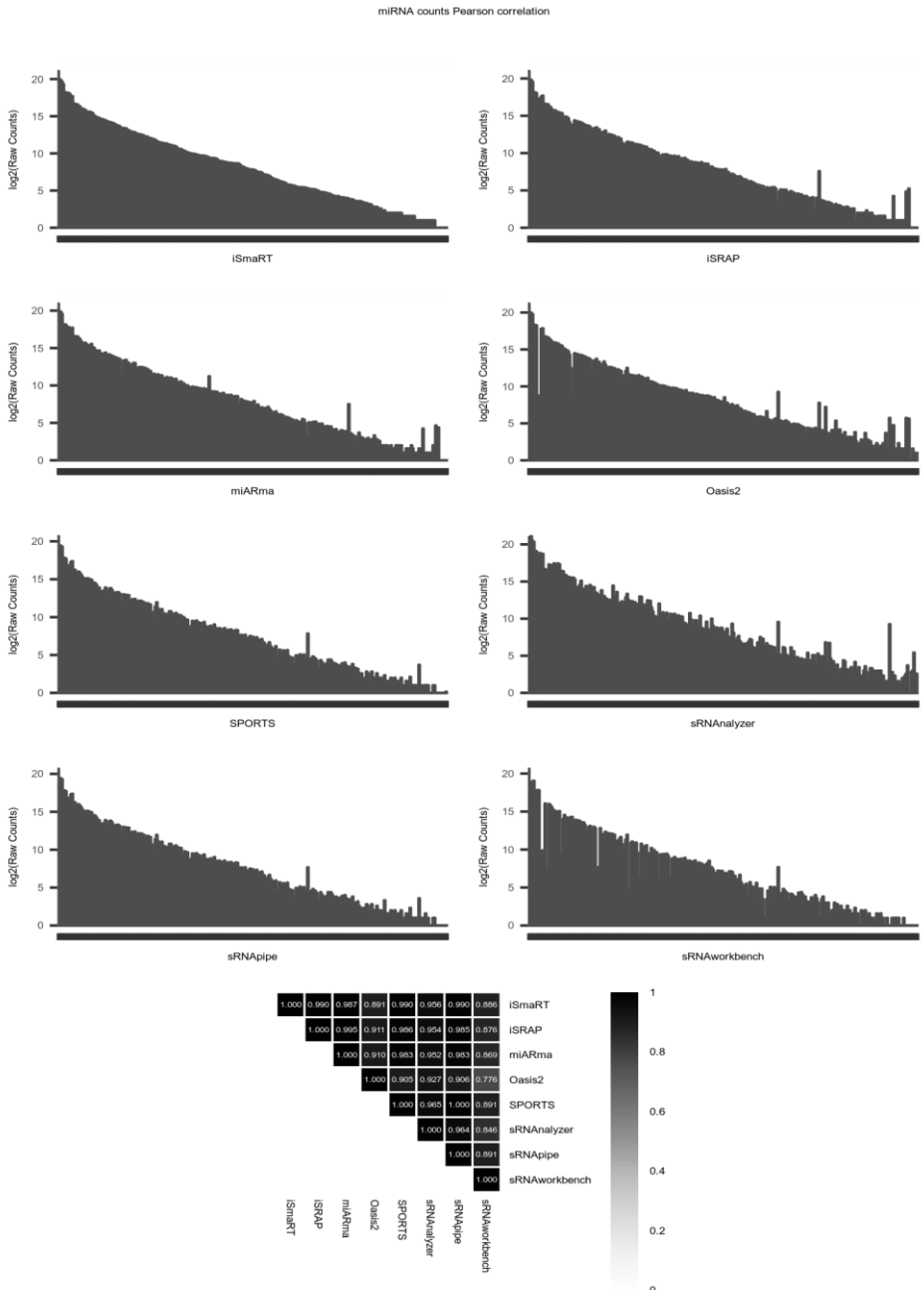

Fig 25. Pearson correlation of the common identified miRNA raw counts for each couple of pipelines.

piRNAs. Concerning piRNAs identification, we observed generally low similarities among the tools, possibly due to the high numbers of FP piRNAs identified by each method. *iSmaRT* and *iSRAP* are the most similar ones, although at a low level (J = 0.31) (Fig. 24). Concerning piRNAs counts estimation, *iSRAP*, *iSmaRT* and *sRNAnalyzer* show high correlations, while *Oasis 2* seems to be less consistent with the others (Fig. 26).

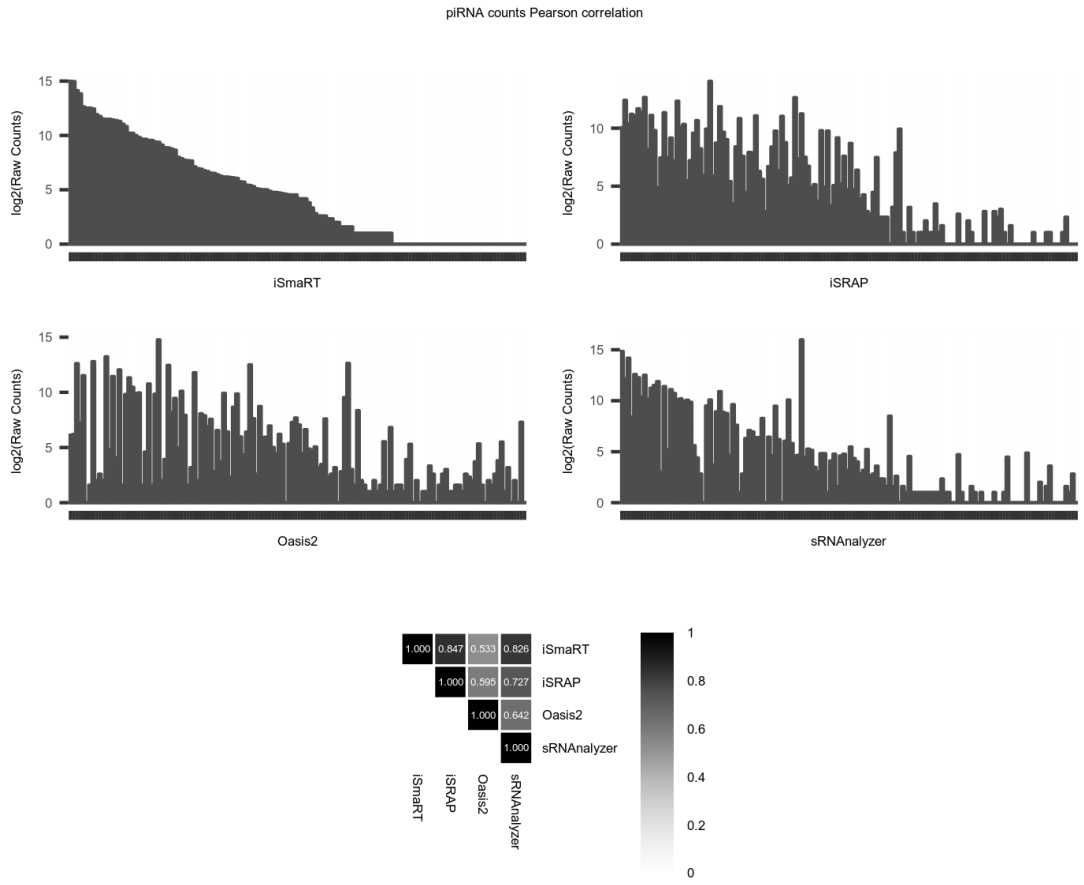

Fig 26. Pearson correlation calculated on the read counts of the common identified piRNAs for each couple of pipelines.

snoRNAs. Concerning snoRNAs identification and quantification, *iSRAP*, *miARma-Seq* and *sRNAnalyzer* show very high similarities and correlation. On the other hand, *Oasis 2* seems to be less consistent with the other tools both in terms of similarity and read count estimation (Fig. 24 and Fig. 27).

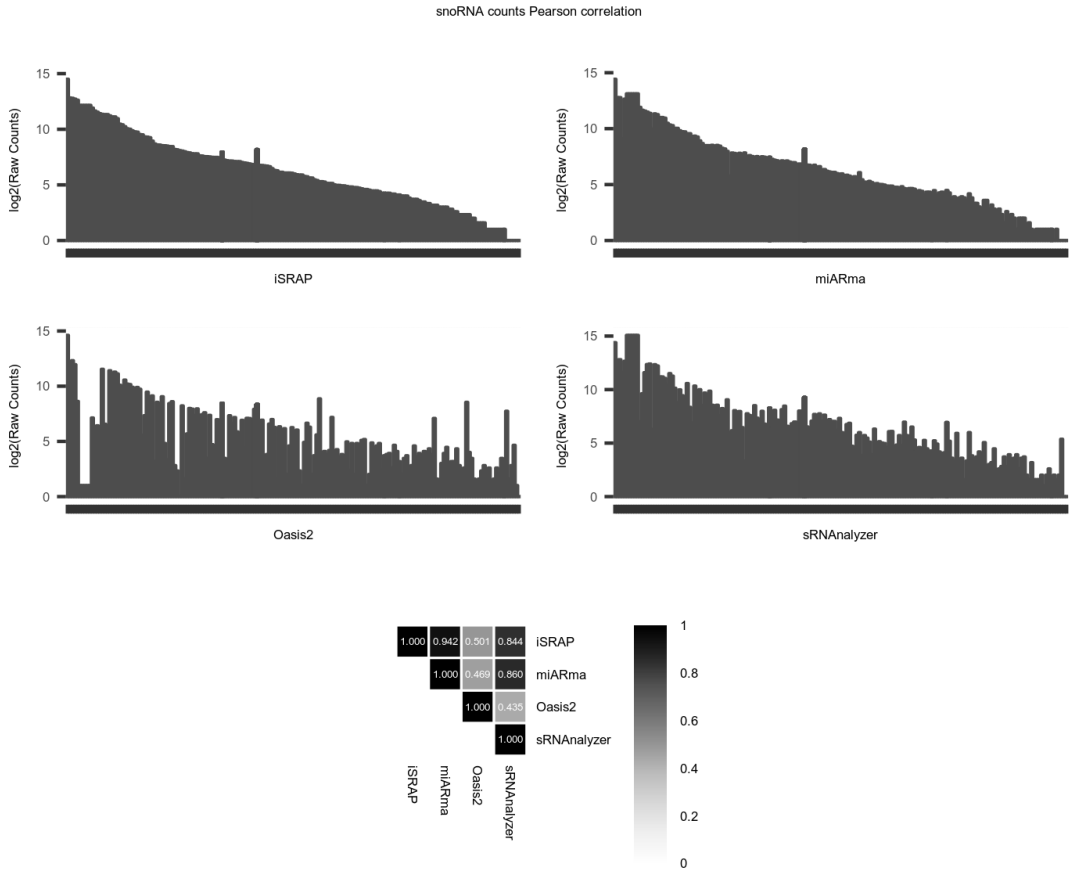

Fig 27. Pearson correlation calculated on the read counts of the common identified snoRNAs for each couple of pipelines.

lncRNAs could be analyzed only with *sRNAnalyzer*, while circRNAs could be evaluated only with *miARma-Seq*; therefore, comparative statistics could not be calculated. Nevertheless, we executed both pipelines with the real RNA-Seq data set obtained from GDC and identified 11.715 lncRNA and 819 circRNAs using sRNAnalyzer and miARma-Seq, respectively.

5.3 RNAdetector

The benchmark of the previous pipelines for the analysis of ncRNAs from RNA-Seq data has shown variable accuracies to identify and quantify the different classes of ncRNAs. However, major concerns come from their usage. Indeed, we believe that

some limitations and shortcomings may have negatively impacted their usage by non-expert users. Among them, we highlight (i) no Graphical User Interface but only command line shell; (ii) software dependencies prior to the pipeline installation; (iii) support only for Linux operating systems; (iv) strict workflow; (v) not suitable for the analysis of the whole transcriptome (e.g. mRNAs, and\or few classes of ncRNAs supported); (vi) no downstream analysis modules (i.e. differential expression analysis or pathway analysis); (vii) only few species supported. To overcome these limitations, we have developed *RNAdetector,* a free stand-alone, cross-platform, and user-friendly RNA-Seq data analysis software which can be used completely offline by mean of an easy-to-use GUI allowing the analysis of coding and ncRNAs from RNA-Seq datasets of any sequenced biological species. A detailed description of *RNAdetector* is described in the next sections.

5.3.1 Software introduction

RNAdetector has been designed to be extremely easy-to-use, flexible, cross-platform, and highly comprehensive, allowing users to analyze not only mRNAs but also different classes of ncRNAs. Precisely, several classes of human, mouse, and *C.elegans* ncRNAs such as miRNAs, piRNAs [only for human at this moment], snoRNAs, lncRNAs, t-UCR [only for human at this moment], circRNAs, and all tRNA-derived ncRNAs classes reported in tRFexplorer [128] and tRFdb [44] are already stored in the remote repository of *RNAdetector* and they can be downloaded directly by mean of the user interface allowing an easier analysis. However, additional species can also be analyzed by uploading their genomes and\or transcriptomes (in FASTA format) and the genomic coordinates (in GTF or BED format) of the RNA molecules to be analyzed following the step-by-step procedure detailed in the user interface.

More importantly, RNAdetector allows not only the identification and quantification of the aforementioned classes, but it also provides downstream analysis modules such as differential expression analysis and miRNA-sensitive topological pathway analysis [93] giving users the opportunity to infer import biological information from their RNA-Seq data.

5.3.2 Deployment and installation

In order to promote a wide use of this tool, we believed that an easy installation and dependency management had to be one of its features. For this purpose, *RNAdetector* is distributed as Docker container and automatically installed after its first execution. No previous dependencies are needed to be installed in users' machines and it can be used as a simply offline desktop application with several operative systems such as Windows Professional, macOS, and Linux. To install *RNAdetector*, it is only necessary to install Docker in users' machines and then download the installer specific for the user's operating systems from our repository https://rnadetector.atlas.dmi.unict.it/download.html. After that, users have only to follow the instructions on the installation wizard to authorize the installer and proceed with the installation. A detailed explanation of how to install Docker and RNAdetector in Windows, macOS and Ubuntu machines is reported at the following link of our GitHub page https://github.com/alessandrolaferlita/RNAdetector/wiki/Requirements-and-Setup. Moreover, *RNAdetector* can be installed in servers and it can be remotely controlled by installing our application in laptops or tablets. No internet connection is needed to perform the analysis. In fact, *RNAdetector* is a completely offline stand-alone software developed to handle not only public RNA-Seq datasets but also private patient-derived RNA-Seq data which are covered by patients' privacy and they cannot

be analyzed by using other web-based pipelines. However, internet connection is needed to update the software. A summary of its system requirements is shown in Table 5.

Feature	Description
Supported operating systems	Windows Professional; macOS; Linux
Dependencies	Docker
Connectivity	Offline (Internet connection is only required for the installation and updates)
Minimum System Requirements	Processor: 6 cores processor RAM: 16GB Hard drive: 1Tb (space is required to store the analysis of multiple samples)
Recommended System Requirements	Processor: 8 cores processor or greater RAM: 32GB or more Hard drive: 2Tb or more (space is required to store the analysis of multiple samples)

Table 5. RNAdetector's system requirements

5.3.3 Functionalities

One of the different strengths of *RNAdetector* is its interactive and easy-to-use GUI. Our GUI has been implemented to be used by users with no computer programming background in order to promote its use both in small research and biomedical laboratories. Our GUI allows users to select among several tools and options to perform the most suitable analysis for their data. In fact, users can select which input files they want to use to start their analysis (e.g., FASTQ, SAM, or BAM), and accordingly with RNA-Seq strategy, which class of RNAs they want to analyze. Precisely, in addition to the mRNAs, several classes of human, mouse, and *C.elegans* ncRNAs such as miRNAs, piRNAs [only for human at this moment], snoRNAs, lncRNAs, t-UCRs [only for human at this moment], circRNAs, and all tRNA-derived ncRNAs classes reported in tRFexplorer [128] and tRFdb [44] are already stored in the remote repository of *RNAdetector* and they can be downloaded directly by mean of the user interface allowing an easier analysis. However, additional species can also be analyzed by uploading their genomes and\or transcriptomes and the genomic

coordinates (in GTF or BED format) of the RNA molecules to be analyzed following the step-by-step procedure detailed in the user interface. In order to give an extreme flexibility to our software, users can also select which tool they want to use for each step of the pipeline and its parameters. For the alignment, users can choose to be executed on a reference genome by using STAR [62] / HISAT 2 [157] or transcriptome by using SALMON [64]. Indeed, the alignment strategy is a critical point for RNA-Seq data analysis, and it must be evaluated accordingly with the purpose of the analysis. For example, the alignment of reads to a reference transcriptome with SALMON is the suggested strategy to analyze the expression profile of splicing-variant transcripts while for other RNA molecules which are not subject to alternative-splicing or to summarize read counts at gene-level expression the alignment on a reference genome could be a good option. In addition, in order to see the depth of coverage of the mapped reads across the entire genome, an offline interactive genome browser based on JBrowse 2 [137] was integrated in the user interface. Concerning read counting, it can also be performed by choosing one of the several available tools such as HTseq [74], FeatureCount [75], or SALMON [64]. However, for circRNAs the pipeline has a strict workflow which consists of aligning the reads on a reference genome with BWA [152] and then novel or already annotated circRNAs on circBase [141] can be identified and quantified by using CIRI 2 [153,154] or CIRIquant [155]. Optional downstream analysis modules on the identified and quantified mRNAs and ncRNAs are also available. Specifically, *RNAdetector* allows users to perform differential expression analysis and miRNA-sensitive topological pathway analysis. Normalization and differential expression analysis can be performed by DESeq2 [77], edgeR [78] , LIMMA [76] or by the combination of these three methods accordingly with user preference while miRNA-sensitive topological pathway analysis is executed by MITHrIL [93] algorithm. MITHrIL

fully exploits the topological information encoded by pathways when computing perturbation scores and then pathways are modeled as complex graphs where each node is a gene or miRNA, and each edge is an interaction between them. Even though thousands of genes are not annotated in pathways and existing annotations may be inaccurate, graphs contained in these databases provide a more detailed view of biological processes within the cell, helping the interpretation of high-throughput experiments [90]. All the tools used by *RNAdetector* for each step of the pipeline are all well-known and widely used freeware tools with tested and proven efficiency individually used by bioinformaticians for the analysis of RNA-Seq data and integrated in *RNAdetector* in order to simplify users' experience. At this moment, *RNAdetector* supports several species such as human, mouse and *C.elegans* that are already available for download in our remote repository. However, it can also be easily used with any other sequenced organisms by uploading their genomes and\or transcriptomes in FASTA format following the step-by-step procedures detailed in the user interface. A summary of RNAdetector's functionalities, supported species, RNA types, and input files is shown in Table 6. A complete user's guide is available at the Wiki section of our GitHub page at the following link https://github.com/alessandrolaferlita/RNAdetector/wiki.

Feature	***Description***
Input Files	FASTQ; BAM; SAM
Supported Analysis	Quantification; Differential expression analysis; Pathway analysis
Supported Species	Human; Mouse; *C.elegans*. Additional sequenced species can be analyzed by uploading their genome and\or transcriptome in FASTA format following the step-by-step procedure detailed in the user interface.
Supported RNA types	mRNAs; miRNAs; snRNAs; snoRNAs; piRNA [only for human at this moment]; tsRNAs; tUCR [only for human at this moment]; lncRNAs; circRNAs. Additional ncRNAs classes can be analyzed by uploading their genomic coordinates in GTF or BED format following the step-by-step procedure detailed in the user interface.
Output Files	Graphical final report for both Differential Expression Analysis and Pathway Analysis with summary of the results, figures, and tables. Text files with raw counts, normalized counts, differentially expressed genes, and impacted pathways can also be downloaded.

Table 6. RNAdetector's supported analysis, species, RNA types and files.

5.3.4 Final report and output files

To guarantee an easy interpretation of the results, we believed that an interactive and exhaustive report with a summary of the results, tables, and several plots must be crucial. Specifically, we developed two different automatic reports for differential expression and pathway analysis modules respectively. The report for the differential expression analysis is based on metaseqR [156] but we modified it in order to better show RNAdetector's results. It shows a summary of the results with all parameters and input options used for the analysis (in order to allow an easier experimental reproducibility), and several figures which show the quality of the sequencing and its results such as Multidimensional scaling plots, RNA-Seq reads noise plots, Correlation plots, Pairwise scatterplots, Box Plots, RNA composition plots, Gene/transcript length bias plots, Mean-difference plots, Mean-variance plots, Volcano plots, DEG heatmaps, and Meta-analysis Venn diagrams. All the pictures generated by *RNAdetector* in its final report are high quality pictures which can be used for publications and easy results interpretation. In addition, an interactive table for each comparison is also present with all the results obtained from the analysis. Finally, the entire report for the differential expression analysis can also be downloaded as a PDF file for user's convenience or viewed directly through the user interface. Similarly to the differential expression analysis report, the report for the miRNA-sensitive topological pathway analysis presents a summary with the results and several interactive figures and tables that show the biological pathways that have been found perturbed. Also in this case, the entire report can be downloaded as a PDF file or viewed directly through the user interface. In addition to the final reports, users can also download specific figures shown in the report and text files with raw and normalized read count matrices, differentially expressed mRNAs\ncRNAs and perturbed pathways.

5.3.5 Case study

In order to show an example of *RNAdetector's* analysis, we chose a public small RNA-Seq project available on NCBI SRA (SRP183064) and we performed a complete analysis identifying the differentially expressed small ncRNAs and the impacted biological pathways. Precisely, we used very recent small RNA-Seq datasets of Colon Rectal Cancer (CRC) [160] and we compared the expression profiles of the CRC samples against the adjacent normal tissue samples of the same patients in order to identify the differentially expressed miRNAs, snoRNAs, and tRNA-derived ncRNAs and the impacted biological pathways. The total number of samples used was 12 (6 CRC samples and 6 adjacent normal tissue samples). In this case study, two conditions are presents (tumor samples and adjacent normal tissue). However, studies with a more complex experimental design which present more than two conditions are also supported by *RNAdetector*. Before starting the differential expression analysis, *RNAdetector* performs some quality control analyses whose results are included in the final report. For example, through a Multi-Dimensional Scaling (MDS) analysis performed by *RNAdetector*, it is evident that (except for two samples) the CRC samples and the normal adjacent tissue samples identify two distinct clusters (Fig. 28).

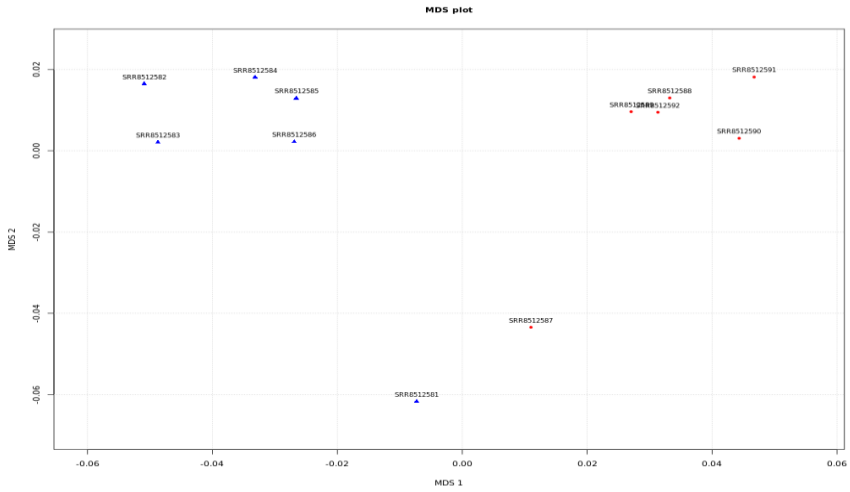

Fig 28. MDS plot showing the CRC (blue triangles) and adjacent normal tissue (red circles) samples.

In addition to the MDS analysis, the correlation analysis also showed high correlations between the samples of the same biological condition (Fig. 29) confirming the good quality of the samples used for the analysis.

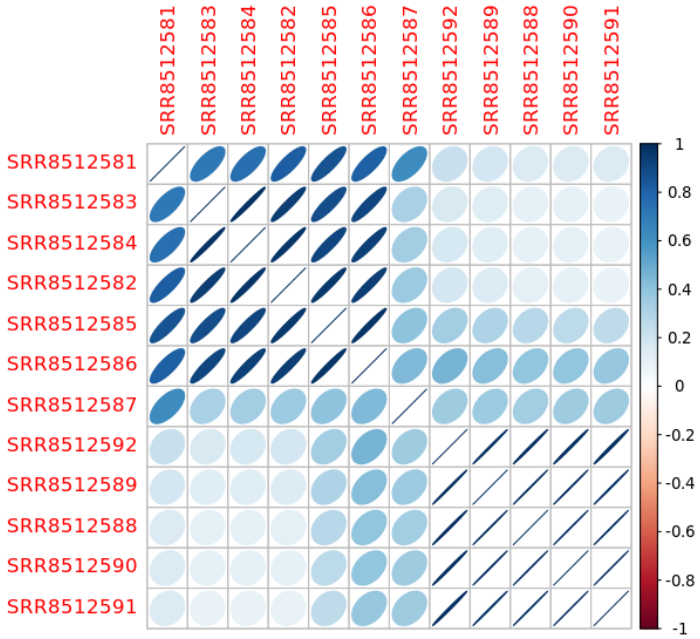

Fig 29. This figure shows a 'correlogram' plot generated by RNAdetector in its final report, where the samples are hierarchically clustered and the correlations between samples are presented as ellipses inside each cell. Each cell represents a pairwise comparison and each correlation coefficient is represented by an ellipse whose 'diameter', direction, and color depict the accordance for that pair of samples. Highly correlated samples are depicted as ellipses with narrow diameters, while poorly correlated samples are depicted as ellipses with wide diameters. Also, highly correlated samples are depicted as ellipses with a left-to-right upwards direction while poorly correlated samples are depicted as ellipses with a right-to-left upwards direction. From the correlogram plot is evident how CRC and normal tissue samples form two distinct groups (samples are named with their SRR identifiers).

Through the differential expression analysis, *RNAdetector* identified 426 statistically significant small ncRNAs with a p-value threshold of 0.05 (357 out of 426 with an FDR or adjusted p-value < 0.05) and of these, 215 (191 with an FDR or adjusted p-value < 0.05) were up-regulated, 153 (140 with an FDR or adjusted p-value < 0.05) were down-regulated and 58 (26 with an FDR or adjusted p-value < 0.05) were not

differentially expressed according to an absolute fold-change cutoff value of 1 in log2 scale. Precisely, a tRNA-fragment 3' (tRF-3) named tRFdb-3033a, a tsRNAs named ts-112, 87 snoRNAs, and 337 miRNAs were found differentially expressed. The complete list of the differentially expressed small ncRNA can be found in the supplementary table 5 while in Fig. 30 they are displayed in a volcano plot generated by *RNAdetector* in its final report.

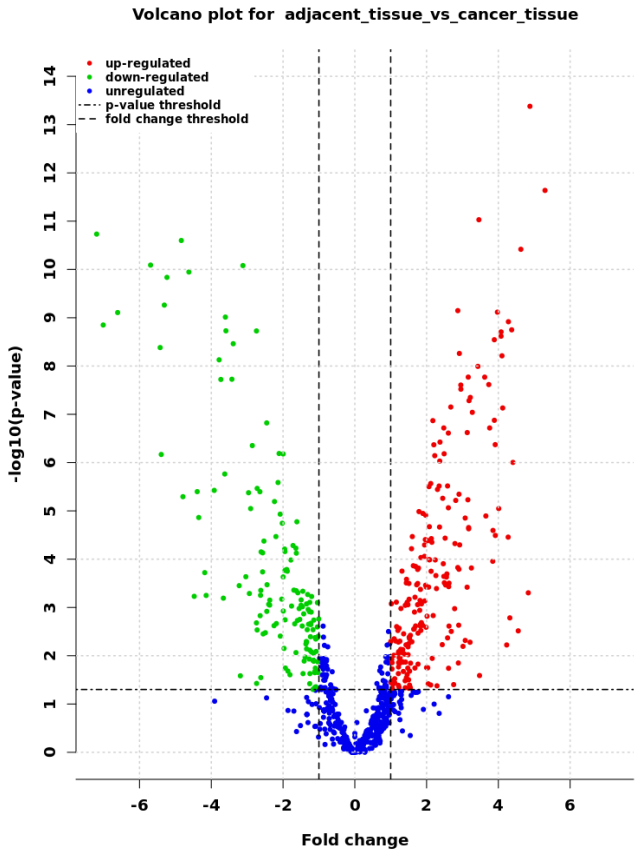

Fig 30. This figure shows a volcano plot generated by RNAdetector in its final report with the up-regulated (red) and down-regulated (green) small ncRNAs identified after the comparison between CRC samples vs adjacent normal tissue samples.

The aforementioned numbers refer to the combined analysis performed by LIMMA and edgeR selecting only the small ncRNAs which have been found differentially expressed by both approaches. A heatmap generated by *RNAdetector* with the top 100

differentially expressed small ncRNAs is also shown in Fig. 31 confirming the presence of two distinct clusters.

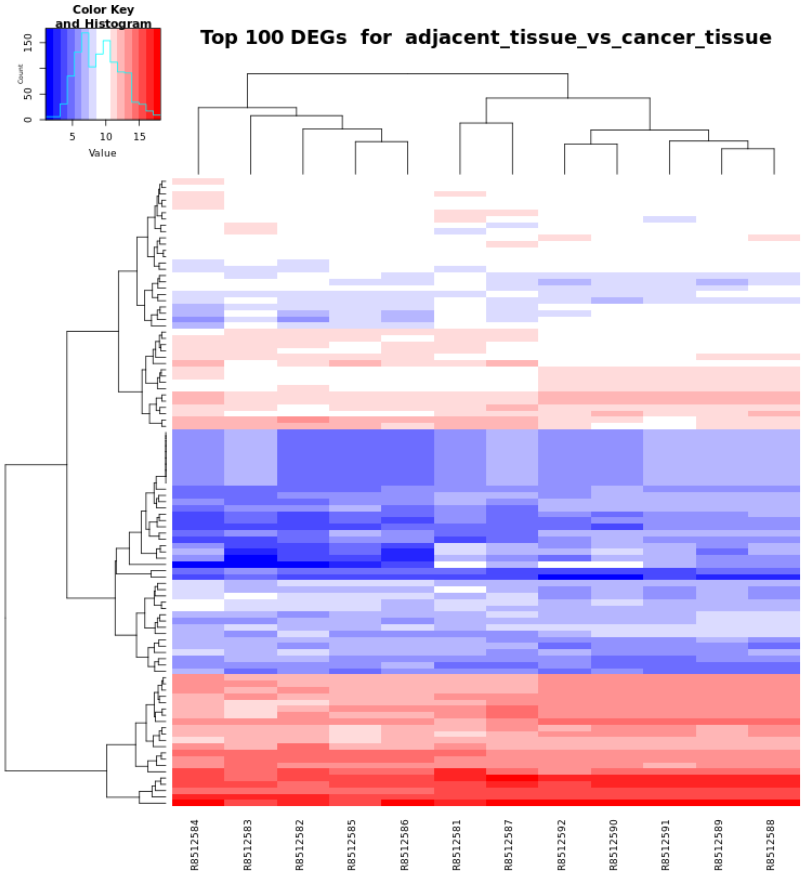

Fig 31. This figure shows a heatmap generated by RNAdetector with the top 100 differentially expressed small ncRNAs. The top 100 deregulated small ncRNAs were selected for their statistical significance in terms of smaller adjusted p-value. Also with the top 100 deregulated small ncRNAs, CRC and normal tissue samples form two distinct clusters (samples are named with their SRR identifiers).

After the differential expression analysis, the deregulated miRNAs were used for the pathway analysis. In fact, in *RNAdetector* there is the possibility to perform miRNA-sensitive topological pathway analyses by using MITHrIL algorithm [93]. In this experiment, 166 pathways were found significantly impacted (FDR or adjusted p-value threshold of 0.01) in the CRC samples compared with adjacent normal tissue samples due to the alteration in the expression profiles of miRNAs. The complete list of the impacted pathways can be found in the supplementary table 6 while in Fig. 32 they are shown in a volcano plot generated by *RNAdetector* in its final pathway

analysis report. In the end, in order to support the experimental reproducibility, all the parameters and input options used for each step of the pipeline are reported in the final reports generated by *RNAdetector*.

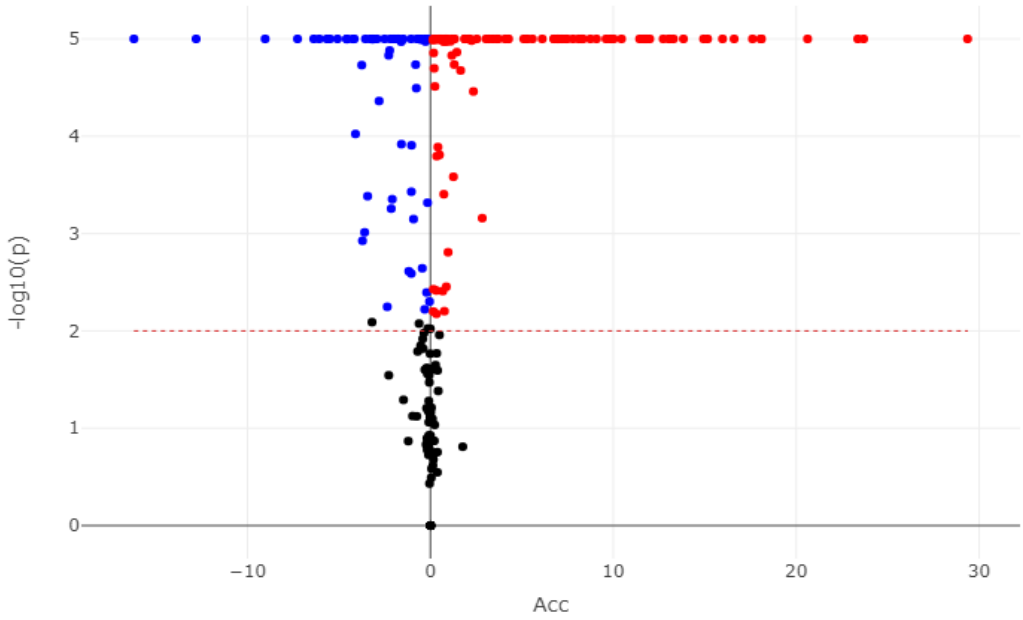

Fig 32. This figure shows a volcano plot generated by RNAdetector in its pathway analysis report with the significantly impacted pathways. All significantly impacted pathways are represented in terms of their measured accumulation (x-axis) and the significance (y-axis). The significance is represented in terms of the negative log (base 10) of the p-value so that more significant genes are plotted higher on the y-axis. The dotted lines represent the thresholds used to select significantly impacted pathways. Significantly impacted pathways with positive accumulation are shown in red, while the negative ones are in blue.

5.3.6 Feature comparison of RNAdetector against pre-existing pipelines

To highlight the extensive feature' set of *RNAdetector*, we compared our software against some other relevant examples. Specifically, in the next two sections we compared *RNAdetector* against 19 pipelines for RNA-Seq data analysis and 7 pipelines for ncRNAs analysis, respectively.

5.3.6.1 Feature comparison with previous RNA-Seq pipeline

Among the RNA-Seq analysis pipelines, we selected ArrayExpressHTS (https://www.bioconductor.org/packages/release/bioc/html/ArrayExpressHTS.html), BioJupies [94], BioWardrobe [95], DEWE [96], easyRNASeq [97], ExpressionPlot [98], FX [99], GENE-counter [100], GeneProf [101], Grape RNA-Seq [102], MAP-RSeq [103], RAP [104], RobiNA [105], RSEQtools [106], RseqFlow [107], S-MART [108], TCW [109], TRAPLINE [110] and wapRNA [111]. Although interesting, some of them present shortcomings that may have negatively impacted their usage among non-expert users (a table that shows the features of *RNAdetector* compared with the other methods is presented in the supplementary table 7). For instance, with the exception of web-based and cloud-based pipelines that do not require a local installation (e.g. BioJupies [94], FX [99], GeneProf [101], RAP [104], TRAPLINE [110], and wapRNA [111]), all of them have dependencies that have to be previously installed in the user's computer or they require the installation and setup of virtual machines. In addition, some of these pipelines do not have GUIs (e.g. ArrayExpressHTS, easyRNASeq [97], GENE-counter [100], Grape RNA-Seq [102], MAP-RSeq [103], RSEQtools [106], and RseqFlow [107]). This limits their usage by users who are confident with the command-line shell. Another limiting aspect of such pipelines is their low flexibility. In fact, some of these pipelines have no customizable work-flows (e.g. BioJupies [94], BioWardrobe [95], ExpressionPlot [98], FX [99], Grape RNA-Seq [102], MAP-RSeq [103], RobiNA [105], RseqFlow [107], S-MART [108], TCW [109], TRAPLINE [110], and wapRNA [111]) and, therefore, they do not allow users to select the proper tools and options in each step of the pipeline (e.g. alignment, read quantification, differential expression analysis, etc.). Finally, important features of an RNA-Seq analysis pipeline include 1) the presence of downstream analysis modules, 2) the presence of a graphical and interactive final report for an easy interpretation of

the results and 3) the availability of ncRNA analysis settings. Concerning the downstream analysis modules, ArrayExpressHTS, easyRNASeq [97], Grape RNA-Seq [102], RSEQtools [106], do not present any downstream analysis module. On the contrary, BioWardrobe [95], ExpressionPlot [98], RobiNA [105], and S-MART [108] include at least one tool for the differential expression analysis module while BioJupies [94], DEWE [96], GENE-counter [100], GeneProf [101], RAP [104], RseqFlow [107], TCW [109], TRAPLINE [110], and wapRNA [111] allow to perform differential expression analysis and other different downstream analyses (see supplementary table 7 for further details). Other pipelines do not generate any interactive graphical final report with a summary of the results together with figures and tables (e.g. ArrayExpressHTS, easyRNASeq [97], FX [99], GENE-counter [100], RSEQtools [106], RseqFlow [107], and TRAPLINE [110]) making more difficult the interpretation of the obtained results. Finally, as an extremely limiting aspect, none of these pipelines allows specific settings for ncRNA analyses. Only TRAPLINE [110] and wapRNA [111] enable the analysis of miRNAs and their targets. Lastly, some of these compared pipelines such as BioWardrobe [95], DEWE [96], ExpressionPlot [98], FX [99], GeneProf [101], RseqFlow [107], and wapRNA [111] are no longer maintained. *RNAdetector* overcomes all these limitations by including all these aforementioned features, which might be individually present in specific pipelines, with new additional ones in a single integrated solution in order to simplify the user's experience.

5.3.6.2 Feature comparison with previous ncRNA-Seq pipeline

We also compared the features of *RNAdetector* against some recent ncRNA pipelines which are able to analyze more than one class of ncRNAs from RNA-Seq data. These pipelines are iSmaRT [143], iSRAP [144], miARma-Seq [145], Oasis 2 [146], SPORTS1.0 [147], sRNAnalyzer [148], and sRNApipe [149]. All these pipelines are able to identify and

quantify different sets of ncRNAs classes with variable accuracy [129]. However, many of them present similar limitations to those of the previously discussed RNA-Seq pipelines (further details are reported in the supplementary table 8). All but miARma-Seq [145] (that is deployed by docker container), Oasis 2 [146] (that is a web-based application), and sRNApipe [149] (that is a Galaxy server application) are standalone tools that need several dependencies to be previously installed on users' machines. Moreover, only iSmaRT [143], Oasis 2 [146], and sRNApipe [149] have a GUI (for the last two is web interface). In addition, none of them generate a graphical final report with a summary of the results and figures that can help users to interpret the results. However, all but sRNAnalyzer [148] generate text files containing the results of the analysis together with several plots. Also for such pipelines, users have no chance to customize the workflows by selecting the suitable aligners and read-counting tool along with several parameters and options. Finally, only iSmaRT [143], miARma-Seq [145], and Oasis 2 [146] allow performing differential expression analysis, miRNA target predictions, and GO\pathways enrichment analyses while iSRAP [144] supports only a differential expression analysis module. As a final consideration, none of the tested ncRNA pipelines are able to analyze a comprehensive list of different classes of regulatory ncRNAs (e.g. miRNAs, piRNAs, snoRNAs, tUCRs, lncRNAs, circRNAs, and tRNA-derived ncRNAs). Indeed, they are restricted to analyze a small set of different classes of ncRNAs which mainly include miRNAs, piRNAs, and snoRNAs (for further details see supplementary table 8).

6. Discussion

As extensively discussed in this PhD thesis, in the last years, NGS technologies are boosting our understanding of the molecular mechanisms underlying prokaryotic and eukaryotic cell signaling, development, and organization [165]. In fact, these technologies allow the sequencing of entire genomes in a few days, yielding the possibility to detect gene mutations or polymorphisms (e.g., CNV, SNPs, INDEL, STR) potentially associated with different diseases [165] (read *1.1 Next Generation Sequencing* for more details). In addition to the DNA sequencing, NGS platforms are also extensively used for transcriptome profiling (RNA-Seq), allowing the identification of differentially expressed genes, splicing variants, or complex gene rearrangements which could represent driver events in specific diseases [23]. Unlike other technologies for transcriptome analysis such as Real Time PCR or microarray, RNA-Seq allows the analysis of the whole transcriptome. On the contrary, hybridization-based approaches have several limitations. Among them, they need existing knowledge about the RNA sequences to be analyzed. Therefore, non-annotated and novel transcripts cannot be analyzed by these technologies. On the other hand, RNA-Seq technologies directly identify all the cDNA sequences produced after the library preparation. As a consequence, all the RNA species that are present in the biological samples are sequenced and analyzed (read *1.2 RNA sequencing* for more details). This aspect is very important for transcriptome analyses since the real complexity of the transcriptome is still to be elucidated. An important factor of such complexity is the huge variety of ncRNAs that are produced by the eukaryotic cells. As discussed in the introduction, ncRNAs are RNA molecules which do not encode

for proteins but represent a considerable amount of the transcriptome involved in many aspects of cell physiology [23,24] and they can be classified according to their size or function [24] (read *1.2.1 Transcriptome* for more details). Upon the increasing research interest in ncRNAs, the identification of the different subclasses has emerged as a critical issue. Indeed, RNA-Seq produces a dramatically higher amount of data than other traditional technologies demanding for fast and effective computational approaches [166]. For this purpose, several pipelines have been deployed for the analysis of gene expression from RNA-Seq data. Relevant examples include BioJupies [94], BioWardrobe [95], DEWE [96], easyRNASeq [97], ExpressionPlot [98], FX [99], GENE-counter [100], GeneProf [101], Grape RNA-Seq [102], MAP-RSeq [103], RAP [104], RobiNA [105], RSEQtools [106], RseqFlow [107], S-MART [108], TCW [109], TRAPLINE [110] and wapRNA [111]. In addition, other pipelines have been developed for the analysis of different ncRNA classes: DSAP [112], miRanalyzer [113], miRExpress [114], miRNAkey [115], iMir [116], CAP-miRSeq [117], mirTools 2.0 [118], sRNAtoolbox [119], miRDeep 2 [120], and MapMi [121] for miRNA analysis; piPipes [122], PILFER [123], piRNAPredictor [124] and PIANO [125] for piRNA analysis; and UClncR [126] for lncRNA analysis. More recent pipelines have also been released to analyze small RNA-Seq data allowing the analysis of more than one class of ncRNAs such as iSmaRT [143], iSRAP [144], miARma-Seq [145], Oasis 2 [146], SPORTS1.0 [147], sRNAnalyzer [148], sRNApipe [149], and sRNAworkbench [150]. However, many of these tools present several limitations and shortcomings which have negatively impacted their usage by non-expert users highlighting the need for more comprehensive, flexible, and easy-to-use free tools that could be used either for research or clinical purposes [129]. In particular, within a biomedical research setting, the availability of stand-alone offline software is crucial to guarantee data safety and patient privacy. Therefore, the need for such tools is clearly urgent. To overcome these

limitations, my PhD research project was focused on the development of a free stand-alone, cross-platform, and user-friendly RNA-Seq data analysis software which can be used completely offline by mean of an easy-to-use GUI allowing the analysis of coding and ncRNAs from RNA-Seq data of any sequenced biological species. However, in order to achieve this goal several steps were required.

First of all, our software aimed to analyze an extensive repertoire of different classes of ncRNAs from RNA-Seq data. However, for tRNA-derived ncRNAs, no previous system for their detection was released when the project started in 2018. In addition, there were not extensive databases that covered all the different subclasses (already available databases such as tRFdb [44] and MINTbase [127] were primarily focused on tRFs deriving from mature tRNAs). Therefore, we had to implement our own database that collects all the tRNA-derived ncRNAs classes that it will be possible to analyze with *RNAdetector*. For this reason, we selected the public small RNA-Seq datasets of the NCI-60 cell lines and TCGA samples in order to identify all the expressed tRNA-derived ncRNAs and we included them in a novel database called *tRFexplorer* [128] (for more details read *5.1 tRNA-derived ncRNAs and their novel database tRFexplorer*).

Secondly, we wanted to evaluate the state of the art of current ncRNA pipelines in order to identify their strengths and weaknesses, and therefore, optimize our software by filling the gaps of the previous methodologies. For this purpose, the performances of eight ncRNA pipelines enabling the processing of RNA-Seq data, published between 2015 and 2019, were compared [129]. In particular, we evaluated the easiness of installation and usage together with their accuracy to identify ncRNAs and their expression levels by using both synthetic and real RNA-Seq datasets [129]. The benchmark of these pipelines showed variable accuracy to identify and quantify different ncRNA classes [129]. However, major concerns were related to their

functionalities and usage [129]. In fact, many of them presented (i) no Graphical User Interface but only command line shell; (ii) software dependencies prior to the pipeline installation; (iii) support only for Linux operating systems; (iv) not suitable for the analysis of the whole transcriptome (e.g. few ncRNA classes supported); (v) static workflow which does not allow to select among different tools and parameters for each step of the pipeline; (vi) no downstream analysis modules (i.e. differential expression analysis or pathway analysis); (vii) only few species supported (for more details read *5.2 Evaluation of pre-existing ncRNA pipelines*).

After taking note of these limitations, we started the development of our software *RNAdetector*. As extensively discussed, *RNAdetector* was designed as an easy-to-use, flexible, cross-platform, and comprehensive pipeline, allowing users to analyze mRNAs and ncRNAs of any sequenced biological species (for more details read *5.3.1 Software introduction*). Specifically, *RNAdetector* allows not only the identification and quantification of coding and ncRNAs, but it also provides downstream analysis modules such as differential expression analysis and miRNA-sensitive topological pathway analysis [93] giving users the opportunity to infer import biological information from their RNA-Seq data (for more details read *5.3.3 Functionalities*). In order to manage the several dependencies, *RNAdetector* is distributed as a Docker container and automatically installed after its first execution. No previous dependencies are needed to be installed in users' machines and it can be used as a simple offline desktop application with several operating systems such as Windows Professional, macOS, and Linux (for more details read *5.3.2 Deployment and installation*). Moreover, *RNAdetector* can be installed in servers and it can be remotely controlled by installing our application in laptops or tablets. No internet connection is needed to perform the analysis. In fact, *RNAdetector* is a completely offline stand-alone software developed

to handle not only public RNA-Seq datasets but also private patient-derived RNA-Seq data which are usually covered by patients' privacy and they cannot be analyzed by using other web-based pipelines. Moreover, to guarantee an easy interpretation of the results, we thought that an interactive and complete report with a summary of the results, tables, and several plots must be crucial. Therefore, we developed two different automatic reports for differential expression analysis and pathway analysis respectively that show a summary of the results together with several interactive figures and tables. In addition, an interactive genome browser is also present in order to visualize the depth of coverage of mapped reads produced by *RNAdetector* (for more details read *5.3.4 Final report and output files*). Finally, by comparing the features of *RNAdetector* against some relevant RNA-Seq and ncRNA-Seq analysis pipelines, we showed that some shortcomings are shared between the previous RNA-Seq and ncRNA-Seq pipelines. However, *RNAdetector* fills these important gaps by combining several features with new additional ones in a single one-stop-shop software to simplify the user's experience allowing, at the same time, a complete analysis of RNA-Seq data (for more details read *5.3.6 Feature comparison of RNAdetector against pre-existing pipelines*).

7. Conclusions

In conclusion, in this PhD thesis, I have presented my research project that led the development of *RNAdetector*, a free stand-alone, cross-platform and user-friendly software for the analysis of coding and ncRNAs from RNA-Seq data of any sequenced biological species. Among its key features we stress: (i) it is freely available for non-commercial usage; (ii) thanks to our Docker-based backend, RNAdetector can be easily installed and deployed in any operating system; (iii) it has an intuitive GUI that allows researchers with no programming background to be shortly productive; (iv) our internal repository contains the latest updates to all supported genomes and transcriptomes; (v) it is omni-comprehensive, in fact, all ncRNAs classes already discovered for sequenced organisms can be analyzed; (vi) it is flexible, indeed it allows users to select among several tools and options for each step of the pipeline; (vii) finally our integrated reporting solution can be used to easily visualize and share results. In the end, we believe that *RNAdetector* is a timely system that could fill an important gap between the needs of biomedical and research labs to process RNA-Seq data and their common lack of technical background in performing such analyses which usually relies in outsourcing through third parties bioinformatics facilities or using expensive commercial software.

Future perspectives

RNAdetector is open-source software that will undergo future developments with the aim of increasing its functionalities. Specifically, we are planning to 1) add new downstream analysis modules such as differential exon usage analysis, alternative polyadenylation analysis, and detection of fusion-transcript; 2) implement a cloud-version of *RNAdetector* for iOS and Android mobile devices. Finally, with the advancement in the functional annotation of the newly emergent classes of ncRNAs, we hope to add additional downstream analysis modules for their functional characterization.

Availability

tRFexplorer

-Website https://trfexplorer.cloud/browse

RNAdetector

-Home page https://rnadetector.atlas.dmi.unict.it/index.html

-Download https://rnadetector.atlas.dmi.unict.it/download.html

-User guide https://github.com/alessandrolaferlita/RNAdetector/wiki

-Operating system(s): Windows Professional, macOS, Linux.

-Programming language: JavaScript, PHP, Perl, Shell, R.

-Other requirements: Docker.

-License: Creative Commons Attribution-ShareAlike 4.0 International license.

-Any restrictions to use by non-academics: no restrictions.

References

1. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).
2. Pevzner, P. A. & Compeau, G. T. P. E. C. Veritas Genetics. Veritas genetics launches $999 whole genome and sets new standard for genetic testing—Press Release. (2016).
3. Di Resta, C., Galbiati, S., Carrera, P. & Ferrari, M. Next-generation sequencing approach for the diagnosis of human diseases: open challenges and new opportunities. *EJIFCC* **29**, 4–14 (2018).
4. Lee, H., Martinez-Agosto, J. A., Rexach, J. & Fogel, B. L. Next generation sequencing in clinical diagnosis. *Lancet Neurol.* **18**, 426 (2019).
5. Nyrén, P., Pettersson, B. & Uhlén, M. Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay. *Anal. Biochem.* **208**, 171–175 (1993).
6. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
7. Dressman, D., Yan, H., Traverso, G., Kinzler, K. W. & Vogelstein, B. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 8817–8822 (2003).
8. Rusk, N. Torrents of sequence. *Nat. Methods* **8**, 44–44 (2011).
9. Rothberg, J. M. *et al.* An integrated semiconductor device enabling non-optical

genome sequencing. *Nature* **475**, 348–352 (2011).

10. Perkel, J. Making contact with sequencing's fourth generation. *Biotechniques* **50**, 93–95 (2011).
11. Valouev, A. *et al.* A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome Res.* **18**, 1051–1063 (2008).
12. Liu, L. *et al.* Comparison of next-generation sequencing systems. *J. Biomed. Biotechnol.* **2012**, 251364 (2012).
13. Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* **2010**, db.prot5448 (2010).
14. McCarroll, S. A. & Altshuler, D. M. Copy-number variation and association studies of human disease. *Nature Genetics* vol. 39 S37–S42 (2007).
15. Mirkin, S. M. Expandable DNA repeats and human disease. *Nature* **447**, 932–940 (2007).
16. Stankiewicz, P. & Lupski, J. R. Structural variation in the human genome and its role in disease. *Annu. Rev. Med.* **61**, 437–455 (2010).
17. Levene, M. J. Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations. *Science* vol. 299 682–686 (2003).
18. Korlach, J. *et al.* Real-Time DNA Sequencing from Single Polymerase Molecules. *Methods in Enzymology* 431–455 (2010) doi:10.1016/s0076-6879(10)72001-2.
19. Loomis, E. W. *et al.* Sequencing the unsequenceable: Expanded CGG-repeat alleles of the fragile X gene. *Genome Research* vol. 23 121–128 (2013).
20. Eisenstein, M. Oxford Nanopore announcement sets sequencing sector abuzz.

Nat. Biotechnol. **30**, 295–296 (2012).

21. Mikheyev, A. S. & Tin, M. M. Y. A first look at the Oxford Nanopore MinION sequencer. *Mol. Ecol. Resour.* **14**, 1097–1102 (2014).
22. Clarke, J. *et al.* Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.* **4**, 265–270 (2009).
23. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).
24. La Ferlita, A. *et al.* Non-Coding RNAs in Endometrial Physiopathology. *Int. J. Mol. Sci.* **19**, 2120 (2018).
25. Balatti, V., Pekarsky, Y. & Croce, C. M. Role of the tRNA-Derived Small RNAs in Cancer: New Potential Biomarkers and Target for Therapy. *Adv. Cancer Res.* **135**, 173–187 (2017).
26. Kim, H. K. *et al.* A transfer-RNA-derived small RNA regulates ribosome biogenesis. *Nature* **552**, 57–62 (2017).
27. Schorn, A. J., Gutbrod, M. J., LeBlanc, C. & Martienssen, R. LTR-Retrotransposon Control by tRNA-Derived Small RNAs. *Cell* **170**, 61–71.e11 (2017).
28. Ivanov, P. Emerging Roles of tRNA-derived Fragments in Viral Infections: The Case of Respiratory Syncytial Virus. *Molecular therapy: the journal of the American Society of Gene Therapy* vol. 23 1557–1558 (2015).
29. Saikia, M. *et al.* Angiogenin-cleaved tRNA halves interact with cytochrome c, protecting cells from apoptosis during osmotic stress. *Mol. Cell. Biol.* **34**, 2450–2463 (2014).
30. Pekarsky, Y. *et al.* Dysregulation of a family of short noncoding RNAs, tsRNAs, in human cancer. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 5071–5076

(2016).

31. Balatti, V. *et al.* tsRNA signatures in cancer. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 8071–8076 (2017).
32. Slack, F. J. Tackling Tumors with Small RNAs Derived from Transfer RNA. *N. Engl. J. Med.* **378**, 1842–1843 (2018).
33. Shao, Y. *et al.* tRF-Leu-CAG promotes cell proliferation and cell cycle in non-small cell lung cancer. *Chem. Biol. Drug Des.* **90**, 730–738 (2017).
34. Li, S., Xu, Z. & Sheng, J. tRNA-Derived Small RNA: A Novel Regulatory Small Non-Coding RNA. *Genes* **9**, (2018).
35. Fu, H. *et al.* Stress induces tRNA cleavage by angiogenin in mammalian cells. *FEBS Lett.* **583**, 437–442 (2009).
36. Thompson, D. M. & Parker, R. The RNase Rny1p cleaves tRNAs and promotes cell death during oxidative stress in Saccharomyces cerevisiae. *J. Cell Biol.* **185**, 43–50 (2009).
37. Lee, Y. S., Shibata, Y., Malhotra, A. & Dutta, A. A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). *Genes Dev.* **23**, 2639–2649 (2009).
38. Kumar, P., Anaya, J., Mudunuri, S. B. & Dutta, A. Meta-analysis of tRNA derived RNA fragments reveals that they are evolutionarily conserved and associate with AGO proteins to recognize specific RNA targets. *BMC Biol.* **12**, 78 (2014).
39. Xu, W.-L., Yang, Y., Wang, Y.-D., Qu, L.-H. & Zheng, L.-L. Computational Approaches to tRNA-Derived Small RNAs. *Noncoding RNA* **3**, (2017).
40. Shen, Y. *et al.* Transfer RNA-derived fragments and tRNA halves: biogenesis, biological functions and their roles in diseases. *J. Mol. Med.* **96**, 1167–1176 (2018).

41. Cole, C. *et al.* Filtering of deep sequencing data reveals the existence of abundant Dicer-dependent small RNAs derived from tRNAs. *RNA* **15**, 2147–2160 (2009).
42. Kumar, P., Kuscu, C. & Dutta, A. Biogenesis and Function of Transfer RNA-Related Fragments (tRFs). *Trends Biochem. Sci.* **41**, 679–689 (2016).
43. Telonis, A. G. *et al.* Dissecting tRNA-derived fragment complexities using personalized transcriptomes reveals novel fragment classes and unexpected dependencies. *Oncotarget* **6**, 24797–24822 (2015).
44. Kumar, P., Mudunuri, S. B., Anaya, J. & Dutta, A. tRFdb: a database for transfer RNA fragments. *Nucleic Acids Res.* **43**, D141–5 (2015).
45. Zheng, L.-L. *et al.* tRF2Cancer: A web server to detect tRNA-derived small RNA fragments (tRFs) and their expression in multiple cancers. *Nucleic Acids Res.* **44**, W185–93 (2016).
46. Clark, T. A., Sugnet, C. W. & Ares, M., Jr. Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science* **296**, 907–910 (2002).
47. Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344–1349 (2008).
48. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
49. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17**, 13 (2016).
50. Parkhomchuk, D. *et al.* Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res.* **37**, e123 (2009).

51. Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods* vol. 7 1009–1015 (2010).
52. Garber, M., Grabherr, M. G., Guttman, M. & Trapnell, C. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat. Methods* **8**, 469–477 (2011).
53. Sims, D., Sudbery, I., Ilott, N. E., Heger, A. & Ponting, C. P. Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.* **15**, 121–132 (2014).
54. Andrews, S. & Others. FastQC: a quality control tool for high throughput sequence data. (2010).
55. Dai, M. *et al.* NGSQC: cross-platform quality analysis pipeline for deep sequencing data. *BMC Genomics* **11 Suppl 4**, S7 (2010).
56. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
57. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
58. Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184–2185 (2012).
59. García-Alcalde, F. *et al.* Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics* **28**, 2678–2679 (2012).
60. Tarazona, S. *et al.* Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Res.* **43**, e140 (2015).
61. Risso, D., Schwartz, K., Sherlock, G. & Dudoit, S. GC-content normalization for RNA-Seq data. *BMC Bioinformatics* **12**, 480 (2011).

62. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

63. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).

64. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).

65. Roberts, A., Pimentel, H., Trapnell, C. & Pachter, L. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* **27**, 2325–2329 (2011).

66. Mezlini, A. M. *et al.* iReckon: Simultaneous isoform discovery and abundance estimation from RNA-seq data. *Genome Research* vol. 23 519–529 (2013).

67. Li, J. J., Jiang, C.-R., Brown, J. B., Huang, H. & Bickel, P. J. Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 19867–19872 (2011).

68. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).

69. Engström, P. G. *et al.* Systematic evaluation of spliced alignment programs for RNA-seq data. *Nature Methods* vol. 10 1185–1191 (2013).

70. Xie, Y. *et al.* SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics* **30**, 1660–1666 (2014).

71. Schulz, M. H., Zerbino, D. R., Vingron, M. & Birney, E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* **28**, 1086–1092 (2012).

72. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).

73. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).

74. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).

75. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

76. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

77. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

78. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

79. Soneson, C. & Delorenzi, M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics* **14**, 91 (2013).

80. Kvam, V. M., Liu, P. & Si, Y. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. *American Journal of Botany* vol. 99 248–256 (2012).

81. Robles, J. A. *et al.* Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. *BMC Genomics* **13**, 484 (2012).

82. Rapaport, F. *et al.* Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol.* **14**, R95 (2013).

83. Su, Z. *et al.* A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.* **32**, 903 (2014).

84. Seyednasrollah, F., Laiho, A. & Elo, L. L. Comparison of software packages for detecting differential expression in RNA-seq studies. *Brief. Bioinform.* **16**, 59–70 (2015).

85. Nookaew, I. *et al.* A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. *Nucleic Acids Res.* **40**, 10084–10097 (2012).

86. Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A. & Conesa, A. Differential expression in RNA-seq: a matter of depth. *Genome Res.* **21**, 2213–2223 (2011).

87. Bi, Y. & Davuluri, R. V. NPEBseq: nonparametric empirical bayesian-based procedure for differential expression analysis of RNA-seq data. *BMC Bioinformatics* **14**, 262 (2013).

88. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nature Genetics* vol. 25 25–29 (2000).

89. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).

90. Alaimo, S., Micale, G., La Ferlita, A., Ferro, A. & Pulvirenti, A. Computational Methods to Investigate the Impact of miRNAs on Pathways. *Methods Mol. Biol.*

1970, 183–209 (2019).

91. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
92. Joshi-Tope, G. *et al.* Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.* **33**, D428–32 (2005).
93. Alaimo, S. *et al.* Post-transcriptional knowledge in pathway analysis increases the accuracy of phenotypes classification. *Oncotarget* **7**, 54572–54582 (2016).
94. Torre, D., Lachmann, A. & Ma'ayan, A. BioJupies: Automated Generation of Interactive Notebooks for RNA-Seq Data Analysis in the Cloud. *Cell Syst* **7**, 556–561.e3 (2018).
95. Kartashov, A. V. & Barski, A. BioWardrobe: an integrated platform for analysis of epigenomics and transcriptomics data. *Genome Biology* vol. 16 (2015).
96. López-Fernández, H., Blanco-Míguez, A., Fdez-Riverola, F., Sánchez, B. & Lourenço, A. DEWE: A novel tool for executing differential expression RNA-Seq workflows in biomedical research. *Comput. Biol. Med.* **107**, 197–205 (2019).
97. Delhomme, N., Padioleau, I., Furlong, E. E. & Steinmetz, L. M. easyRNASeq: a bioconductor package for processing RNA-Seq data. *Bioinformatics* **28**, 2532–2533 (2012).
98. Friedman, B. A. & Maniatis, T. ExpressionPlot: a web-based framework for analysis of RNA-Seq and microarray gene expression data. *Genome Biol.* **12**, R69 (2011).
99. Hong, D. *et al.* FX: an RNA-Seq analysis tool on the cloud. *Bioinformatics* **28**, 721–723 (2012).
100. Cumbie, J. S. *et al.* GENE-counter: a computational pipeline for the analysis of

RNA-Seq data for gene expression differences. *PLoS One* **6**, e25279 (2011).

101. Halbritter, F., Vaidya, H. J. & Tomlinson, S. R. GeneProf: analysis of high-throughput sequencing experiments. *Nat. Methods* **9**, 7–8 (2011).
102. Knowles, D. G., Röder, M., Merkel, A. & Guigó, R. Grape RNA-Seq analysis pipeline environment. *Bioinformatics* **29**, 614–621 (2013).
103. Kalari, K. R. *et al.* MAP-RSeq: Mayo Analysis Pipeline for RNA sequencing. *BMC Bioinformatics* **15**, 224 (2014).
104. D'Antonio, M. *et al.* RAP: RNA-Seq Analysis Pipeline, a new cloud-based NGS web application. *BMC Genomics* **16**, S3 (2015).
105. Lohse, M. *et al.* RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Res.* **40**, W622–7 (2012).
106. Habegger, L. *et al.* RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries. *Bioinformatics* **27**, 281–283 (2011).
107. Wang, Y. *et al.* RseqFlow: workflows for RNA-Seq data analysis. *Bioinformatics* **27**, 2598–2600 (2011).
108. Zytnicki, M. & Quesneville, H. S-MART, A Software Toolbox to Aid RNA-seq Data Analysis. *PLoS ONE* vol. 6 e25988 (2011).
109. Soderlund, C., Nelson, W., Willer, M. & Gang, D. R. TCW: transcriptome computational workbench. *PLoS One* **8**, e69401 (2013).
110. Wolfien, M. *et al.* TRAPLINE: a standardized and automated pipeline for RNA sequencing data analysis, evaluation and annotation. *BMC Bioinformatics* **17**, 21 (2016).
111. Zhao, W. *et al.* wapRNA: a web-based application for the processing of RNA sequences. *Bioinformatics* **27**, 3076–3077 (2011).

112. Huang, P.-J. *et al.* DSAP: deep-sequencing small RNA analysis pipeline. *Nucleic Acids Res.* **38**, W385–91 (2010).

113. Hackenberg, M., Rodríguez-Ezpeleta, N. & Aransay, A. M. miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments. *Nucleic Acids Res.* **39**, W132–8 (2011).

114. Wang, W.-C. *et al.* miRExpress: analyzing high-throughput sequencing data for profiling microRNA expression. *BMC Bioinformatics* **10**, 328 (2009).

115. Ronen, R. *et al.* miRNAkey: a software for microRNA deep sequencing analysis. *Bioinformatics* **26**, 2615–2616 (2010).

116. Giurato, G. *et al.* iMir: an integrated pipeline for high-throughput analysis of small non-coding RNA data obtained by smallRNA-Seq. *BMC Bioinformatics* **14**, 362 (2013).

117. Sun, Z. *et al.* CAP-miRSeq: a comprehensive analysis pipeline for microRNA sequencing data. *BMC Genomics* **15**, 423 (2014).

118. Wu, J. *et al.* mirTools 2.0 for non-coding RNA discovery, profiling, and functional annotation based on high-throughput sequencing. *RNA Biol.* **10**, 1087–1092 (2013).

119. Rueda, A. *et al.* sRNAtoolbox: an integrated collection of small RNA research tools. *Nucleic Acids Res.* **43**, W467–73 (2015).

120. Friedländer, M. R., Mackowiak, S. D., Li, N., Chen, W. & Rajewsky, N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res.* **40**, 37–52 (2012).

121. Guerra-Assunção, J. A. & Enright, A. J. MapMi: automated mapping of microRNA loci. *BMC Bioinformatics* **11**, 133 (2010).

122. Han, B. W., Wang, W., Zamore, P. D. & Weng, Z. piPipes: a set of pipelines for

piRNA and transposon analysis via small RNA-seq, RNA-seq, degradome- and CAGE-seq, ChIP-seq and genomic DNA sequencing. *Bioinformatics* **31**, 593–595 (2015).

123. Ray, R. & Pandey, P. piRNA analysis framework from small RNA-Seq data by a novel cluster prediction tool - PILFER. *Genomics* **110**, 355–365 (2018).
124. Zhang, Y., Wang, X. & Kang, L. A k-mer scheme to predict piRNAs and characterize locust piRNAs. *Bioinformatics* **27**, 771–776 (2011).
125. Wang, K. *et al.* Prediction of piRNAs using transposon interaction and a support vector machine. *BMC Bioinformatics* **15**, 419 (2014).
126. Sun, Z. *et al.* UClncR: Ultrafast and comprehensive long non-coding RNA detection from RNA-seq. *Sci. Rep.* **7**, 14196 (2017).
127. Pliatsika, V. *et al.* MINTbase v2.0: a comprehensive database for tRNA-derived fragments that includes nuclear and mitochondrial fragments from all The Cancer Genome Atlas projects. *Nucleic Acids Res.* **46**, D152–D159 (2018).
128. La Ferlita, A. *et al.* Identification of tRNA-derived ncRNAs in TCGA and NCI-60 panel cell lines and development of the public database tRFexplorer. *Database* **2019**, (2019).
129. Di Bella, S. *et al.* A benchmarking of pipelines for detecting ncRNAs from RNA-Seq data. *Brief. Bioinform.* (2019) doi:10.1093/bib/bbz110.
130. Chan, P. P. & Lowe, T. M. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Res.* **44**, D184–9 (2016).
131. Marshall, E. A. *et al.* Small non-coding RNA transcriptome of the NCI-60 cell line panel. *Sci Data* **4**, 170157 (2017).
132. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of

insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).

133. Wagner, G. P., Kin, K. & Lynch, V. J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* **131**, 281–285 (2012).
134. Su, S. *et al.* Glimma: interactive graphics for gene expression analysis. *Bioinformatics* **33**, 2050–2052 (2017).
135. Reinhold, W. C. *et al.* CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. *Cancer Res.* **72**, 3499–3511 (2012).
136. Reinhold, W. C., Sunshine, M., Varma, S., Doroshow, J. H. & Pommier, Y. Using CellMiner 1.6 for Systems Pharmacology and Genomic Analysis of the NCI-60. *Clin. Cancer Res.* **21**, 3841–3852 (2015).
137. Buels, R. *et al.* JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.* **17**, 66 (2016).
138. Griebel, T. *et al.* Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Res.* **40**, 10073–10083 (2012).
139. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
140. Wang, J. *et al.* piRBase: a comprehensive database of piRNA sequences. *Nucleic Acids Res.* **47**, D175–D180 (2019).
141. Glažar, P., Papavasileiou, P. & Rajewsky, N. circBase: a database for circular RNAs. *RNA* **20**, 1666–1670 (2014).
142. Volders, P.-J. *et al.* LNCipedia 5: towards a reference set of human long non-coding RNAs. *Nucleic Acids Res.* **47**, D135–D139 (2019).
143. Panero, R. *et al.* iSmaRT: a toolkit for a comprehensive analysis of small RNA-

Seq data. *Bioinformatics* **33**, 4050 (2017).

144. Quek, C., Jung, C.-H., Bellingham, S. A., Lonie, A. & Hill, A. F. iSRAP - a one-touch research tool for rapid profiling of small RNA-seq data. *J Extracell Vesicles* **4**, 29454 (2015).

145. Andrés-León, E., Núñez-Torres, R. & Rojas, A. M. miARma-Seq: a comprehensive tool for miRNA, mRNA and circRNA analysis. *Sci. Rep.* **6**, 25749 (2016).

146. Rahman, R.-U. *et al.* Oasis 2: improved online analysis of small RNA-seq data. *BMC Bioinformatics* **19**, 54 (2018).

147. Shi, J., Ko, E.-A., Sanders, K. M., Chen, Q. & Zhou, T. SPORTS1.0: A Tool for Annotating and Profiling Non-coding RNAs Optimized for rRNA- and tRNA-derived Small RNAs. *Genomics Proteomics Bioinformatics* **16**, 144–151 (2018).

148. Wu, X., Kim, T. K., Baxter, D. & Scherler, K. sRNAnalyzer—a flexible and customizable small RNA sequencing data analysis pipeline. *Nucleic acids* (2017).

149. Pogorelcnik, R., Vaury, C., Pouchin, P., Jensen, S. & Brasset, E. sRNAPipe: a Galaxy-based pipeline for bioinformatic in-depth exploration of small RNAseq data. *Mob. DNA* **9**, 25 (2018).

150. Stocks, M. B. *et al.* The UEA sRNA Workbench (version 4.4): a comprehensive suite of tools for analyzing miRNAs and sRNAs. *Bioinformatics* **34**, 3382–3384 (2018).

151. Wickham, H., Chang, W. & Others. ggplot2: An implementation of the Grammar of Graphics. *R package version 0. 7, URL: http://CRAN. R-project. org/package= ggplot2* (2008).

152. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows--Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).

153. Gao, Y., Wang, J. & Zhao, F. CIRI: an efficient and unbiased algorithm for de novo circular RNA identification. *Genome Biol.* **16**, 4 (2015).

154. Gao, Y., Zhang, J. & Zhao, F. Circular RNA identification based on multiple seed matching. *Brief. Bioinform.* **19**, 803–810 (2018).

155. Zhang, J., Chen, S., Yang, J. & Zhao, F. Accurate quantification of circular RNAs identifies extensive circular isoform switching events. *Nat. Commun.* **11**, 90 (2020).

156. Moulos, P. & Hatzis, P. Systematic integration of RNA-Seq statistical algorithms for accurate detection of differential gene expression patterns. *Nucleic Acids Res.* **43**, e25 (2015).

157. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).

158. Kozomara, A. & Griffiths-Jones, S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* **42**, D68–73 (2014).

159. Lomonaco, V. *et al.* UCbase 2.0: ultraconserved sequences database (2014 update). *Database* **2014**, (2014).

160. Zhou, F. *et al.* Identification of microRNAs and their Endonucleolytic Cleavaged target mRNAs in colorectal cancer. *BMC Cancer* **20**, 242 (2020).

161. Huang, B. *et al.* tRF/miR-1280 Suppresses Stem Cell-like Cells and Metastasis in Colorectal Cancer. *Cancer Res.* **77**, 3194–3206 (2017).

162. Feng, W. *et al.* Identification of tRNA-derived small noncoding RNAs as potential biomarkers for prediction of recurrence in triple-negative breast

cancer. *Cancer Medicine* vol. 7 5130–5144 (2018).

163. Olvedy, M. *et al.* A comprehensive repertoire of tRNA-derived fragments in prostate cancer. *Oncotarget* **7**, 24766–24777 (2016).
164. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
165. van Dijk, E. L., Auger, H., Jaszczyszyn, Y. & Thermes, C. Ten years of next-generation sequencing technology. *Trends Genet.* **30**, 418–426 (2014).
166. Malone, J. H. & Oliver, B. Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biol.* **9**, 34 (2011).

Supplementary tables

Supplementary table 1 - NCI-60 cell lines and their public small RNA-Seq datasets.

Cell line	Type of cancer	SRA dataset
T-47D	BREAST	SRR5689215
MCF-7	BREAST	SRR5689213
MDA-MB-231	BREAST	SRR5689212
BT-549	BREAST	SRR5689211
HS-578T	BREAST	SRR5689210
SF-295	CNS	SRR5689217
SF-268	CNS	SRR5689216
SF-539	CNS	SRR5689214
U251	CNS	SRR5689209
SNB-75	CNS	SRR5689208
SNB-19	CNS	SRR5689175
HCT-116	COLON	SRR5689179
HT-29	COLON	SRR5689178
KM12	COLON	SRR5689177
SW-620	COLON	SRR5689176
COLO 205	COLON	SRR5689174
HCT-15	COLON	SRR5689173
HCC2998	COLON	SRR5689172
MOLT-4	LEUKEMIA	SRR5689193
K-562	LEUKEMIA	SRR5689192
SR	LEUKEMIA	SRR5689191
RPMI 8226	LEUKEMIA	SRR5689190
CCRF-CEM	LEUKEMIA	SRR5689183
HL-60(TB)	LEUKEMIA	SRR5689182
LOX-IMVI	MELANOMA	SRR5689197
MALME-3M	MELANOMA	SRR5689196
MDA-MB-435	MELANOMA	SRR5689195
M14	MELANOMA	SRR5689194
SK-MEL-5	MELANOMA	SRR5689189

SK-MEL-28	MELANOMA	SRR5689188
UACC-62	MELANOMA	SRR5689167
UACC-257	MELANOMA	SRR5689164
SK-MEL-2	MELANOMA	SRR5689165
NCI-H522	NSCLC	SRR5689201
NCI-H460	NSCLC	SRR5689200
HOP 62	NSCLC	SRR5689171
NCI-H23	NSCLC	SRR5689170
EKVX	NSCLC	SRR5689169
HOP 92	NSCLC	SRR5689168
A549	NSCLC	SRR5689166
NCI-H322M	NSCLC	SRR5689163
NCI-H226	NSCLC	SRR5689162
OVCAR-5	OVARIAN	SRR5689207
OVCAR-4	OVARIAN	SRR5689206
OVCAR-8	OVARIAN	SRR5689205
OVCAR-3	OVARIAN	SRR5689204
NCI/ADR-RES	OVARIAN	SRR5689203
IGR-OV1	OVARIAN	SRR5689202
SK-OV-3	OVARIAN	SRR5689198
DU-145	PROSTATE	SRR5689199
PC-3	PROSTATE	SRR5689187
A498	RENAL	SRR5689186
CAKI-1	RENAL	SRR5689185
786-0	RENAL	SRR5689184
SN12C	RENAL	SRR5689181
TK-10	RENAL	SRR5689180
UO-31	RENAL	SRR5689161
ACHN	RENAL	SRR5689160
RXF393	RENAL	SRR5689159

Supplementary table 2 - Analyzed TCGA samples.

Tumor type	Tumor name	Tumor samples	Control samples
ACC	Adrenocortical Carcinoma	79	
BLCA	Bladder Urothelial Carcinoma	408	19
BRCA	Breast invasive carcinoma	1101	113
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	306	3
CHOL	Cholangiocarcinoma	36	9
COAD	Colon adenocarcinoma	459	41
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	48	
ESCA	Esophageal carcinoma	185	11
GBM	Glioblastoma multiforme	167	
HNSC	Head and Neck squamous cell carcinoma	522	44
KICH	Kidney Chromophobe	66	25
KIRC	Kidney renal clear cell carcinoma	534	72
KIRP	Kidney renal papillary cell carcinoma	291	32
LAML	Acute Myeloid Leukemia	173	
LGG	Brain Lower Grade Glioma	533	
LIHC	Liver hepatocellular carcinoma	374	50
LUAD	Lung adenocarcinoma	517	59
LUSC	Lung squamous cell carcinoma	501	51
MESO	Mesothelioma	87	
OV	Ovarian serous cystadenocarcinoma	309	
PAAD	Pancreatic adenocarcinoma	179	4
PCPG	Pheochromocytoma and Paraganglioma	184	3
PRAD	Prostate adenocarcinoma	498	52
READ	Rectum adenocarcinoma	166	10
SARC	Sarcoma	263	2
SKCM	Skin Cutaneous Melanoma	472	1
STAD	Stomach adenocarcinoma	414	35
TGCT	Testicular Germ Cell Tumors	139	
THCA	Thyroid carcinoma	513	59
THYM	Thymoma	120	2
UCEC	Uterine Corpus Endometrial Carcinoma	546	23
UCS	Uterine Carcinosarcoma	57	
UVM	Uveal Melanoma	80	
		10327	**720**

Supplementary table 3 - CellMiner's datasets included in tRFexplorer.

CellMiner Dataset	Description
DNA CNV - Roche NibleGen 385K aCGH	385K element tiling array based on NCBI Build 35 of the human genome (HG17) and re-mapped to NCBI Build 35 (HG19); 50-mer tiling with a median probe spacing of 6,000 bp.
DNA CNV - Combined aCGH	Probe intensities combined from four platforms: Agilent Human Genome CGH Microarray 44A, Nimblegen HG19 CGH 385K WG Tiling v2.0, Affymetrix GeneChip Human Mapping 500k Array Set and Illumina Human1Mv1_C Beadchip
DNA SNP per Gene - Affy 500K	This platform is used for whole-genome association studies. It is comprised of two arrays which enable genotyping of more than 500,000 single nucleotide polymorphisms (SNPs).
DNA SNP per Gene - Illumina 1M SNP	BeadChip array based on Illumina's Infinium Assay with probes for 1072820 SNPs
DNA Methylation - Illumina 450K	Approximately 450,000 probes querying the methylation status of CpG sites within and outside of genes.
RNA Affy HG-U133_AB	Human Genome U133. 44,000 probeset 2-chip set. Gene expression.
RNA Affy HG-U133 Plus 2.0	Aproximately 47,000 transcripts
RNA Affy HuEx 1.0	1432155 probesets for all human gene exons
RNA Agilent Human mRNAs	44,000 Probes for approximately 41,000 genes, with 4 arrays spotted on each slide.
RNA Expression Combined z-scores	Gene expressions
RNA Agilent Human miRNAs	15,000 probes for 723 human and 76 human viral miRNA's. Each slide contains 8 arrays.
RNA microRNA OSU V3 Chip	Custom microarray developed at Microarray Shared Resource Comprehensive Cancer Center, OSU microarray facility. It contains 11k probes (2 technical replicates) for murin and human microRNAs together with hypothetical microRNAs and control probes.
RNA ABC Transporters Array	47 specific oligonucleotide probes were designed for each of the ABC transporters using DNAStar Primer Select. Expression levels were measured by real-time quantitative RT-PCR using the LightCycler RNA Amplification SYBR Green kit and a LightCycler machine.
RNA OSU Transporter Array	Spotted 70-mer microarray
Protein Lysate Array	Reverse-phase lysate arrays (RPLA) for 162 antibodies for 94 genes. Each array included 64 lysates (60 cancer cells and 4 replicate control pools) in 10 serial two-fold dilutions.
Compound Activities	Negative log10 (GI50) values of sulforhodamine B assay for ~ 50K compounds, including more than 20,000 that passed quality control, 158 Food and Drug Administration approved and 79 clinical trial drugs. Higher values equate to higher sensitivity of cell lines.

Supplementary table 4 - Criteria used for the evaluation of the ncRNA pipelines.

Criteria	**-**	**+**	**++**
Installation	Manually Dependencies Installation.	Acceptable difficulty. It required manually installation of some tools, automatically dependency installation.	Very simple, all in-one installer package, run with single command, or installation is not required.
Documentation	Unclear, not available or very incomplete.	Not fully clear, incomplete (not explain every step).	Comprehensive, focused and clear (documentation explains every steps).
GUI	Absent, only command line.	Present, GUI of third-party tool.	Present, proprietary GUI or Web Interface.
Different Input Types	Accepted single input type (e.g. FASTQ files).	Accepted FASTQ or BAM files.	Accepted FASTQ or BAM or txt files.
Report generation	Report output in a single file type.	Report output in different file types.	Report output in different file types and plot graphs.
Multi-ncRNAs in single analysis	Single ncRNA class analyzed in a single run.	Multi ncRNA classes analyzed in a single run but some of them are not annotated.	Multi ncRNA classes analyzed in a single run.
Flexibility	It is not possible to start the analysis at different steps of the pipeline.	It is possible to start the analysis at different steps of the pipeline.	It performs different types of analysis, and it is also possible to start the analysis at different steps of the pipeline.
Usability/Configuration	Configuration file missed or few parameters to set.	Configuration file present but not well organized.	Configuration file is very simple to use and set.

Supplementary table 5 - Differentially expressed small ncRNAs found in the case study datasets.

gene_id	meta_p-value	meta_FDR	log2_normalized_fold_change
MIMAT0003215	4,18E-14	3,57E-11	4,880170987
MIMAT0026615	2,31E-12	9,86E-10	5,300302331
MIMAT0004598	9,35E-12	2,66E-09	3,458329062
MIMAT0000770	1,85E-11	3,96E-09	-7,194756854
MIMAT0000437	2,51E-11	4,30E-09	-4,835322132
MIMAT0001536	3,82E-11	5,45E-09	4,62771123
MIMAT0000427	8,12E-11	8,89E-09	-5,692386084
MIMAT0000250	8,31E-11	8,89E-09	-3,118014746
MIMAT0000435	1,13E-10	1,07E-08	-4,62398685
MIMAT0000416_1	1,46E-10	1,25E-08	-5,235100391
MIMAT0000416	5,47E-10	4,25E-08	-5,309290158
MIMAT0000072	7,13E-10	4,77E-08	2,870442925
MIMAT0001620	7,65E-10	4,77E-08	3,97402926
MIMAT0004764	7,81E-10	4,77E-08	-6,609604962
MIMAT0004601	9,70E-10	5,53E-08	-3,607127298
MIMAT0000682	1,21E-09	6,46E-08	4,280589831
MIMAT0002806	1,41E-09	7,11E-08	-7,011742202
MIMAT0000758	1,78E-09	8,00E-08	4,371877391
MIMAT0004552	1,87E-09	8,00E-08	-3,592017258
MIMAT0004614	1,88E-09	8,00E-08	-2,740969939
MIMAT0000252_2	1,96E-09	8,00E-08	4,076859939
MIMAT0004958	2,41E-09	9,36E-08	4,07570966
MIMAT0000318	2,85E-09	1,06E-07	3,887672204
MIMAT0000103	3,45E-09	1,23E-07	-3,388236653
MIMAT0000427_1	4,15E-09	1,42E-07	-5,425182057
MIMAT0004549	5,49E-09	1,81E-07	2,913633385
MIMAT0003242	6,14E-09	1,95E-07	4,101143416
MIMAT0000707	7,44E-09	2,27E-07	-3,781719876
MIMAT0000432	1,02E-08	3,00E-07	3,429979447
MIMAT0000243	1,70E-08	4,69E-07	3,161058641
MIMAT0000259	1,70E-08	4,69E-07	3,619512847
MIMAT0001413	1,88E-08	4,91E-07	-3,425606741
MIMAT0004599	1,90E-08	4,91E-07	-3,732768282
MIMAT0031177	2,42E-08	6,05E-07	3,734188509
U72	2,48E-08	6,05E-07	2,953751323

MIMAT0000073	3,02E-08	7,16E-07	2,955734092
MIMAT0005899	4,47E-08	1,03E-06	3,219215598
MIMAT0000095	5,21E-08	1,17E-06	3,183166164
MIMAT0003321	7,06E-08	1,55E-06	2,67516031
MIMAT0000261	7,39E-08	1,58E-06	4,121599658
MIMAT0004571	9,10E-08	1,90E-06	3,275041707
MIMAT0004698	1,33E-07	2,69E-06	3,888628389
MIMAT0004494	1,35E-07	2,69E-06	2,174769461
MIMAT0000443	1,51E-07	2,93E-06	-2,455595591
MIMAT0000267	1,92E-07	3,56E-06	2,479324126
MIMAT0004560	1,92E-07	3,56E-06	3,75826539
MIMAT0000281	2,39E-07	4,35E-06	3,132996332
ACA43	2,45E-07	4,36E-06	2,606821935
HBII-95	3,76E-07	6,56E-06	2,374065718
MIMAT0004608	4,25E-07	7,17E-06	3,913016698
MIMAT0000071	4,28E-07	7,17E-06	2,202091642
MIMAT0000088	4,44E-07	7,31E-06	-2,856701713
U23	6,57E-07	1,03E-05	2,489178962
MIMAT0005796	6,62E-07	1,03E-05	-1,997011196
MIMAT0000738	6,79E-07	1,04E-05	-5,394768902
HBII-99B	7,19E-07	1,08E-05	2,227930605
MIMAT0000691	9,30E-07	1,37E-05	2,366970571
MIMAT0019814	9,92E-07	1,44E-05	4,407139762
MIMAT0004548	1,72E-06	2,45E-05	-3,624490865
MIMAT0004615	2,58E-06	3,62E-05	-2,139789283
MIMAT0002891	2,72E-06	3,75E-05	2,116409492
MIMAT0000457	3,05E-06	4,09E-05	2,584962501
ACA58	3,06E-06	4,09E-05	2,349584438
MIMAT0003393	3,15E-06	4,14E-05	2,066882865
MIMAT0000423_1	3,42E-06	4,44E-05	-2,72131056
U71d	3,60E-06	4,59E-05	2,304467125
MIMAT0000422_1	3,76E-06	4,73E-05	-3,91753784
MIMAT0004752	4,00E-06	4,90E-05	-4,392317423
MIMAT0004592	4,01E-06	4,90E-05	-2,644848908
MIMAT0004517	4,21E-06	5,07E-05	-2,961180751
MIMAT0000617	4,52E-06	5,36E-05	2,903522357
MIMAT0004925	5,08E-06	5,95E-05	-4,790076931
MIMAT0022709	5,49E-06	6,34E-05	2,450032921

MIMAT0004657	5,89E-06	6,72E-05	3,151309323
ACA31	6,07E-06	6,83E-05	2,812848584
MIMAT0000085	6,41E-06	7,12E-05	-2,238707235
MIMAT0002874	8,63E-06	9,46E-05	2,607667031
MIMAT0000460_1	8,95E-06	9,58E-05	4,010533868
MIMAT0000423	8,96E-06	9,58E-05	-2,902988347
MIMAT0000075	1,03E-05	0,000109173	1,788093537
MIMAT0004680	1,13E-05	0,000117753	1,897011687
MIMAT0000066	1,17E-05	0,000120548	-2,079883022
MIMAT0000074_1	1,23E-05	0,000125177	1,974206381
MIMAT0003260	1,27E-05	0,000128219	3,652076697
MIMAT0026478	1,37E-05	0,000136432	-4,349408831
MIMAT0000449	1,41E-05	0,000138241	3,075625573
MIMAT0004673	1,68E-05	0,000163343	-1,614425802
ts-112	1,80E-05	0,000173395	-2,022507399
MIMAT0004507	2,13E-05	0,000201982	2,077286001
MIMAT0003294	2,16E-05	0,000202986	2,360747344
MIMAT0004928	2,22E-05	0,000205913	3,172099373
MIMAT0000728	2,35E-05	0,000215767	3,173073233
MIMAT0000222	2,54E-05	0,000231084	3,850451278
MIMAT0004543	3,23E-05	0,000290357	3,916806064
U28	3,40E-05	0,00030018	1,602810502
MIMAT0000752	3,41E-05	0,00030018	-2,198897632
MIMAT0004504	3,50E-05	0,00030511	4,273583703
MIMAT0000276	3,67E-05	0,000317144	2,516249751
U68	3,76E-05	0,000321151	2,13258564
MIMAT0000091	3,94E-05	0,000333167	1,964852056
MIMAT0022842	4,13E-05	0,000346563	2,038039361
MIMAT0002875	4,22E-05	0,000350181	-2,533059731
U17b	4,47E-05	0,000367267	2,147295002
MIMAT0001636	4,76E-05	0,000387322	2,790076931
MIMAT0019981	5,06E-05	0,00040585	2,921997488
MIMAT0022727	5,08E-05	0,00040585	1,943956417
MIMAT0004588	5,22E-05	0,000413254	-1,722316808
MIMAT0004515	5,92E-05	0,000464586	-1,642651159
MIMAT0000070	6,08E-05	0,000472543	1,58033653
MIMAT0003888	6,18E-05	0,000476308	-1,955215043
MIMAT0000418	6,95E-05	0,000530323	-1,946789094

MIMAT0000275_1	7,07E-05	0,000534908	-2,61169385
MIMAT0000087	7,39E-05	0,000554273	-2,568156239
MIMAT0000419	7,50E-05	0,000557933	-1,62956557
MIMAT0000094	8,76E-05	0,000646007	1,942460476
U15A	9,26E-05	0,000676611	1,841818332
U17a	0,000102261	0,00073592	2,067057467
MIMAT0000074	0,000103103	0,00073592	2,084064265
MIMAT0003220	0,000103727	0,00073592	2,2410081
MIMAT0004502	0,000104148	0,00073592	-1,781420382
MIMAT0022721	0,000110478	0,000774254	3,843983844
MIMAT0004703	0,000124336	0,000864284	2,465781566
MIMAT0004683	0,000136731	0,00094278	1,630327805
MIMAT0000090	0,000141476	0,000967696	1,693202243
MIMAT0000226_1	0,000149066	0,001011522	2,869777423
MIMAT0004671	0,000152224	0,001020633	3,249196709
U36A	0,000152797	0,001020633	1,774707688
MIMAT0004978	0,000161676	0,001069423	2,902702799
MIMAT0000076	0,000163008	0,001069423	1,751763387
MIMAT0000425	0,000163853	0,001069423	-1,897493391
MIMAT0004485	0,000170682	0,001105557	-1,945086919
MIMAT0000680	0,000176842	0,001127579	1,314448261
MIMAT0000773	0,000177607	0,001127579	-1,881737118
HBII-240	0,000178039	0,001127579	2,125530882
MIMAT0004550	0,000183383	0,001152887	-2,570193041
MIMAT0031175	0,000189967	0,001185558	-4,181897643
MIMAT0019220	0,000202208	0,001252808	2,566346823
MIMAT0001635	0,000218709	0,001345296	2,25947088
MIMAT0003249	0,000227131	0,001387123	2,480525091
MIMAT0000428_1	0,000231065	0,001401137	-3,037089319
MIMAT0000731	0,000233118	0,001403186	-1,996729323
MIMAT0004793	0,000234685	0,001403186	2,554588852
MIMAT0000772	0,000261533	0,00155285	1,441356673
MIMAT0003389	0,00030845	0,001818042	2,494618566
MIMAT0004493	0,000310449	0,001818042	1,435386145
MIMAT0000089	0,000314113	0,001826984	2,625214668
U36B	0,000318282	0,001832429	1,508109582
U71b	0,000319336	0,001832429	1,752213368
MIMAT0026720	0,000338221	0,001920766	-2,44625623

MIMAT0001341	0,000339223	0,001920766	1,724000699
MIMAT0030020	0,000341488	0,001920869	2,529253068
MIMAT0004983	0,000346851	0,001938287	1,972414168
MIMAT0003385	0,00035393	0,001965003	-3,222392421
MIMAT0015053	0,000370406	0,002043207	2,61667136
ACA28	0,00037866	0,002060227	1,994304803
MIMAT0000434	0,000378831	0,002060227	2,056471564
MIMAT0001625	0,00038072	0,002060227	3,129283017
MIMAT0000260	0,000406443	0,002185588	2,276124405
ACA3-2	0,000422336	0,002256857	1,83051911
MIMAT0000461	0,000437991	0,002325977	-1,699073713
MIMAT0004511	0,000443508	0,002340735	-2,620237975
MIMAT0000689	0,000449628	0,002358476	-1,634989303
MIMAT0000077	0,000461493	0,002405953	-1,449365269
MIMAT0003297	0,00049638	0,002563554	-1,587692293
MIMAT0002173	0,000497719	0,002563554	4,834692334
MIMAT0000080	0,000538767	0,00274194	-1,314627839
MIMAT0000097	0,00056041	0,002835208	-2,627655225
MIMAT0018184_1	0,000563833	0,002835748	-4,142957954
MIMAT0022717	0,00058822	0,002941099	-4,475733431
MIMAT0000082_1	0,000593093	0,002948226	-1,203183813
ACA44	0,000614801	0,003038467	1,677167678
MIMAT0000429	0,000642223	0,003155749	-3,663333564
MIMAT0000098	0,000672766	0,003284799	-2,036394902
MIMAT0004503	0,000676169	0,003284799	1,512450001
MIMAT0000438	0,000697506	0,00336931	-1,49562312
MIMAT0004603	0,000705948	0,00339093	-2,369827797
MIMAT0004498	0,000717161	0,003425548	1,816799768
MIMAT0004909	0,000732177	0,003477841	1,948594095
U33	0,000777165	0,003670318	1,157293444
U106	0,000781592	0,003670318	1,459789593
HBII-13	0,000785577	0,003670318	-1,255216268
MIMAT0000065	0,000794853	0,003693472	-1,042143336
U20	0,000832691	0,003841462	1,452409293
MIMAT0000093	0,000839208	0,003841462	1,018581386
MIMAT0000718	0,000844338	0,003841462	-1,345467652
ACA8	0,000844672	0,003841462	1,857120044
MIMAT0000424_1	0,000862572	0,003902113	-2,399930607

U64	0,000875476	0,003934278	1,749565631
U47	0,000878885	0,003934278	1,480968739
MIMAT0000064	0,00089301	0,003976685	-2,372478696
ACA63	0,000903851	0,003986675	-1,498453507
MIMAT0003880	0,000904579	0,003986675	1,40599236
MIMAT0002820	0,00096782	0,004221866	-1,460470924
MIMAT0004505	0,001051484	0,004552086	1,285242071
MIMAT0004951	0,001054779	0,004552086	-1,551871714
MIMAT0018965	0,001059491	0,004552086	2,781359714
MIMAT0000715	0,001100713	0,004705547	-1,299707512
U13	0,001147168	0,004879743	1,808537556
MIMAT0000430_1	0,001187268	0,005025319	-1,993135459
MIMAT0004763	0,001205884	0,005078969	-2,453172628
U43	0,001223454	0,005127711	-1,262935455
MIMAT0004586	0,001245471	0,005175687	1,663152949
MIMAT0023712	0,001247008	0,005175687	-1,215587243
14q(II-3)	0,001356673	0,00560365	-1,424687669
MIMAT0022482	0,001474029	0,006059108	2,184424571
MIMAT0004565	0,001482033	0,006062862	-2,637814383
MIMAT0026482	0,001518524	0,006182563	2,012600037
MIMAT0004761	0,001644678	0,006664452	4,320167637
U95	0,001731671	0,00698386	-1,004914693
MIMAT0010214	0,001767678	0,007095608	-1,5334322
MIMAT0000265	0,00180476	0,007210605	-1,942190732
MIMAT0000082	0,001825996	0,00726152	-1,142439734
MIMAT0004797	0,001869154	0,007398736	-1,601675746
MIMAT0004489	0,002066795	0,008143363	1,855414378
MIMAT0004953	0,002087509	0,008187248	-2,741466986
MIMAT0000683	0,002124559	0,00829451	-1,222968808
MIMAT0000275	0,002157673	0,008385501	-2,186413124
MIMAT0003247	0,002184618	0,008451802	-1,674829701
MIMAT0000062	0,00222594	0,008569955	-1,28804258
ACA32	0,002235205	0,008569955	1,827503141
MIMAT0004587	0,002263271	0,008638823	-1,226562063
MIMAT0015015	0,002332337	0,008862882	2,895302621
HBII-82	0,002348726	0,008885668	1,89077093
MIMAT0004605	0,002386403	0,008976574	-2,251816182
U77	0,002413078	0,008976574	1,822654058

MIMAT0000458	0,002416061	0,008976574	1,693688375
MIMAT0018205	0,002424707	0,008976574	-1,466085105
MIMAT0019820	0,002432581	0,008976574	2,624490865
MIMAT0000067	0,002435749	0,008976574	-0,883173187
U15B	0,002496854	0,009162277	1,16544335
MIMAT0000681	0,002535947	0,00926596	-1,043080091
MIMAT0004985	0,002569202	0,009347522	1,985303489
MIMAT0000705	0,002700154	0,009782337	1,234953711
MIMAT0022838	0,002777077	0,010018569	1,657531127
U14A	0,002821853	0,010137328	1,641344971
MIMAT0005792_1	0,00288929	0,010296656	-1,3264062
MIMAT0003239	0,00289029	0,010296656	-1,165063478
MIMAT0025470	0,002925793	0,010379889	-2,722466024
MIMAT0003257	0,002996684	0,010548946	1,840219556
U79	0,002998121	0,010548946	1,199548678
MIMAT0005923	0,003054208	0,010702244	4,554588852
MIMAT0025477	0,003135292	0,010941528	2,688055994
U101	0,003151445	0,010953192	0,938847362
MIMAT0003250	0,003337355	0,011552381	-2,492396382
ACA50	0,003404895	0,011738651	1,761612601
U49B	0,003515242	0,01207041	1,360452362
MIMAT0018972	0,003531131	0,012076469	-2,554588852
HBII-180C	0,003780037	0,012876221	1,549216839
MIMAT0022471	0,003925527	0,013318751	-2,076350886
MIMAT0000067_1	0,003948613	0,013344127	-1,018910059
MIMAT0000684	0,004041741	0,013605073	-1,236137513
MIMAT0000083	0,004114243	0,013776035	-0,873262426
SNORD121A	0,004124754	0,013776035	1,514573173
MIMAT0004597	0,004211817	0,014012075	-1,107190784
MIMAT0000244	0,004244152	0,014064921	-1,04510803
MIMAT0022258	0,00427022	0,014096671	2,584962501
MIMAT0003258	0,00444705	0,014612903	1,525256255
U24	0,004460781	0,014612903	1,097241646
MIMAT0000510	0,004503599	0,014696858	-1,087862116
MIMAT0000272	0,004820653	0,015671703	3,077244765
U57	0,004954716	0,016046522	1,139953253
MIMAT0001618	0,004995975	0,016119088	1,256013978
MIMAT0003338	0,005042832	0,016209104	1,004063159

14q(II-1)	0,005074022	0,016248274	-1,351763324
MIMAT0000721	0,005145876	0,016416881	1,591095114
MIMAT0002819	0,005183478	0,016475367	-1,342195536
MIMAT0004801	0,005213062	0,01650803	1,215108695
MIMAT0010251	0,005254569	0,016578069	3,211504105
MIMAT0002872	0,005649984	0,01773435	1,499493159
U38B	0,005662547	0,01773435	0,972979076
MIMAT0022726	0,005702972	0,01779577	-1,362983163
MIMAT0000226	0,005892956	0,018321735	2,437063806
snR38C	0,005957472	0,018455213	1,426442263
MIMAT0014996	0,005980242	0,01845887	4,232660757
MIMAT0000440	0,006016955	0,018505384	0,900033028
MIMAT0000063	0,006098796	0,018689858	-0,847030893
MIMAT0000460	0,006390857	0,019514938	3,014873276
MIMAT0002813	0,006537624	0,01982527	1,484024977
MIMAT0000703	0,006538861	0,01982527	-0,754391134
MIMAT0026472	0,007051815	0,021304952	-1,971985624
MIMAT0026479	0,007218847	0,021732797	-1,343954401
HBII-234	0,007377683	0,022133049	1,279115011
MIMAT0002873	0,007484448	0,022374837	1,333423734
MIMAT0000717	0,007664766	0,022834059	-1,254929356
MIMAT0000262	0,008154469	0,02420858	-1,078379293
U16	0,008248012	0,024401559	1,221364128
MIMAT0004602	0,008355606	0,024634633	-1,208430747
MIMAT0004767	0,00861785	0,025320487	-1,639597757
MIMAT0005882	0,008655016	0,025342597	1,33219643
HBII-316	0,008841817	0,025744506	1,215977582
MIMAT0004809	0,008852497	0,025744506	-1,150882554
ACA33	0,009254248	0,026821635	1,38466385
ACA26	0,009507677	0,027463054	1,485426827
U83A	0,00979519	0,028198275	1,088900505
MIMAT0000104	0,010516248	0,030082567	0,961831736
MIMAT0001343	0,010520103	0,030082567	0,862710778
MIMAT0026621	0,010682575	0,030445338	-1,383704292
U51	0,010807329	0,030698558	1,261213793
MIMAT0004491	0,010965262	0,031044036	1,235216462
MIMAT0002888	0,011310151	0,031758957	0,94403957
U44	0,011316141	0,031758957	-0,980612835

MIMAT0005930	0,011329219	0,031758957	2,157541277
MIMAT0000253	0,011417255	0,031813659	-0,971376526
MIMAT0016895	0,01142315	0,031813659	-0,872377497
14q(II-12)	0,011682495	0,032430304	-1,270357747
MIMAT0000092	0,011925965	0,032999028	1,029590319
MIMAT0000693	0,012040346	0,033208052	-0,787110389
tRFdb-3033a	0,012646659	0,033685025	-1,150941898
tRFdb-3033a.1	0,012646659	0,033685025	-1,150941898
tRFdb-3033a.2	0,012646659	0,033685025	-1,150941898
tRFdb-3033a.3	0,012646659	0,033685025	-1,150941898
tRFdb-3033a.4	0,012646659	0,033685025	-1,150941898
tRFdb-3033a.5	0,012646659	0,033685025	-1,150941898
tRFdb-3033a.6	0,012646659	0,033685025	-1,150941898
tRFdb-3033a.7	0,012646659	0,033685025	-1,150941898
tRFdb-3033a.8	0,012646659	0,033685025	-1,150941898
tRFdb-3033a.9	0,012646659	0,033685025	-1,150941898
tRFdb-3033a.10	0,012646659	0,033685025	-1,150941898
MIMAT0004556	0,012881031	0,034147834	-1,411589764
MIMAT0000750	0,012900293	0,034147834	-0,791291163
ACA18	0,013111749	0,034576979	1,308027278
U45C	0,013143296	0,034576979	1,250491124
MIMAT0004555	0,013293835	0,034865733	-1,04075976
HBII-210	0,013381284	0,034987761	-0,938012269
ACA20	0,013532442	0,035184395	1,412125904
U50	0,013554783	0,035184395	0,869727535
MIMAT0003161	0,013579942	0,035184395	1,3408283
MIMAT0000433	0,013888172	0,035818453	1,220353536
U31	0,013908452	0,035818453	-0,847607713
MIMAT0004774	0,013969932	0,035868745	1,038882661
U83B	0,014080748	0,035964668	0,932547818
MIMAT0026476	0,014131581	0,035964668	2,885751829
MIMAT0004693	0,014133484	0,035964668	1,260460462
MIMAT0004613	0,014270361	0,036205218	1,553253641
U42B	0,014598307	0,03692767	-0,880057176
MIMAT0005584	0,014924245	0,037640796	1,761840263
MIMAT0004945	0,016232343	0,040819567	-0,785971052
U48	0,01684336	0,042231886	-0,91799878
MIMAT0004927	0,017075767	0,042618036	1,218534946

MIMAT0000081	0,01709706	0,042618036	0,721333924
MIMAT0021086	0,017634191	0,043829166	2
MIMAT0019693	0,01777677	0,044055473	2,032421478
MIMAT0022692	0,017981841	0,044245923	1,674599713
MIMAT0018204	0,017988969	0,044245923	-1,934904972
HBII-180A	0,018050887	0,044245923	1,476438044
MIMAT0004678	0,018060616	0,044245923	-1,100603941
MIMAT0001541	0,018162208	0,044367681	2,612518223
HBII-108B	0,018773162	0,045729497	1,036246566
MIMAT0000762	0,018926402	0,045971799	-0,880329769
U29	0,019055495	0,046154243	-0,756232442
MIMAT0004584	0,020388249	0,049242804	0,695652891
MIMAT0004495	0,020693441	0,049839133	-1,031197688
MIMAT0003273	0,020776827	0,049899402	-1,86507042
MIMAT0000280	0,020865996	0,049973185	1,526545814
MIMAT0004488	0,021162731	0,050542278	1,352516415
HBII-85-8	0,021380658	0,050920508	-1,117925725
MIMAT0004484	0,021454306	0,050953976	-0,630289395
MIMAT0018186	0,022180642	0,0525331	-1,02727293
MIMAT0000459	0,022422987	0,052960369	-1,115100977
MIMAT0004672	0,022946752	0,054048135	0,757333574
snR38B	0,023418802	0,054799098	1,213955584
MIMAT0000254	0,023457859	0,054799098	-1,251209191
U76	0,023626281	0,055042153	0,920565533
MIMAT0004497	0,024386165	0,056559926	-0,87182379
ACA46	0,024410073	0,056559926	1,459431619
MIMAT0022833	0,024747082	0,057149356	-1,807354922
U82	0,024798142	0,057149356	-0,785231823
MIMAT0005794	0,025015213	0,057494643	-0,90749588
MIMAT0007885	0,025727747	0,058973789	3,475733431
MIMAT0022472	0,026062706	0,059581855	-3,189824559
snR39B	0,027858969	0,06351845	-0,807171772
MIMAT0000454	0,02838443	0,064544382	-2,619896291
MIMAT0003244	0,028524478	0,064690793	1,438573014
ACA57	0,02995947	0,067765468	1,112921513
MIMAT0004949	0,031489456	0,071038219	1,499571009
MIMAT0005825	0,03199737	0,071994083	-0,867053506
MIMAT0018183	0,032135631	0,072115392	-0,83588062

MIMAT0005874	0,032350325	0,072407141	1,080170349
MIMAT0002871	0,032664674	0,072919833	0,789229423
MIMAT0000453	0,032978276	0,073428193	1,401740677
U60	0,034143758	0,075825748	0,843930992
U58C	0,034372033	0,076134944	1,028149942
U45A	0,035090931	0,077526475	0,998029105
MIMAT0009198	0,036279578	0,079945978	1,014451156
MIMAT0000646	0,036380393	0,079962047	1,11451704
MIMAT0027519	0,037474705	0,082156084	-2,736965594
MIMAT0004509	0,037720435	0,082483305	0,615199731
MIMAT0000716	0,037960988	0,082797563	-1,257157839
MIMAT0004950	0,038136901	0,082969594	2,044394119
MIMAT0017988	0,03896194	0,084549388	1,062735755
MIMAT0001532	0,039633267	0,085788463	2,754887502
U78	0,039909116	0,086167409	0,843560766
U73a	0,040045987	0,086245135	-0,67361659
MIMAT0026721	0,040407289	0,086795147	2,115477217
MIMAT0018068	0,040504402	0,086795147	1,525461489
MIMAT0000730	0,041001831	0,087641415	1,21342664
MIMAT0003327	0,041996216	0,089543055	2,289506617
MIMAT0016888	0,042114096	0,089571025	0,879201856
MIMAT0002175	0,043964944	0,093275501	-1,086100737
MIMAT0026475	0,044530323	0,094241153	1,230005605
MIMAT0000450	0,044840764	0,094663836	-0,951825827
ACA48	0,045027455	0,094823829	0,832890014
MIMAT0000759	0,045478533	0,095282302	0,672725835
MIMAT0001627	0,04552291	0,095282302	-1,116343961
MIMAT0027479	0,045579487	0,095282302	1,547487795
MIMAT0011777	0,045774248	0,095321271	0,880039376
HBII-85-16	0,04586406	0,095321271	-0,827441687
MIMAT0019952	0,045932589	0,095321271	1,405256478
HBII-85-2	0,046323198	0,095899115	-0,885995371
HBI-43	0,046599328	0,09614175	0,842540102
U22	0,046665294	0,09614175	0,760150382
U53	0,046949561	0,096424241	1,046833165
MIMAT0000092_1	0,047027963	0,096424241	0,861412221
MIMAT0002818	0,047141521	0,096425839	1,505235308
MIMAT0000062_2	0,047486788	0,096900248	-0,608969156

MIMAT0000764	0,047929899	0,097571579	0,673031229
MIMAT0005901	0,048467706	0,098432039	-0,948925815
MIMAT0004776	0,049127653	0,099535885	-0,987616276
MIMAT0000227	0,049446926	0,099945915	-0,536898508
MIMAT0004674	0,049695365	0,10006205	-0,69366876
MIMAT0027103	0,049738446	0,10006205	-1,167841687
MIMAT0017982	0,049995784	0,100343652	-0,930459067

Supplementary table 6 - Dysregulated pathways found in the case study datasets.

# Pathway Id	Pathway Name	Impact Factor	Corrected Accumulator	pValue	Adjusted pValue
path:hsa04915	Estrogen signaling pathway	141,5919455	29,37020617	0	0
path:hsa04152	AMPK signaling pathway	111,1487884	23,68710986	0	0
path:hsa04630	JAK-STAT signaling pathway	142,6238189	23,3672053	0	0
path:hsa04010	MAPK signaling pathway	175,5042568	20,62252326	0	0
path:hsa04914	Progesterone-mediated oocyte maturation	84,18542026	18,099952	0	0
path:hsa04926	Relaxin signaling pathway	161,7895825	18,06528308	0	0
path:hsa04066	HIF-1 signaling pathway	152,0335168	17,61437649	0	0
path:hsa04151	PI3K-Akt signaling pathway	184,7552123	16,60656147	0	0
path:hsa04530	Tight junction	121,0519833	-16,21177977	0	0
path:hsa04140	Autophagy - animal	114,2792721	15,97382395	0	0
path:hsa04668	TNF signaling pathway	153,6970547	15,15532845	0	0
path:hsa04550	Signaling pathways regulating pluripotency of stem cells	163,5696817	14,93880354	0	0
path:hsa04620	Toll-like receptor signaling pathway	158,3799077	13,83452509	0	0
path:hsa04014	Ras signaling pathway	134,6365584	13,28883801	0	0
path:hsa04150	mTOR signaling pathway	137,2405076	13,08889722	0	0
path:hsa04115	p53 signaling pathway	133,8234914	13,04525948	0	0
path:hsa04810	Regulation of actin cytoskeleton	104,3613291	-12,81017753	0	0
path:hsa04210	Apoptosis	162,3983766	12,72988821	0	0
path:hsa04923	Regulation of lipolysis in adipocytes	38,72952459	11,99015355	0	0
path:hsa04012	ErbB signaling pathway	152,4283396	11,84885009	0	0
path:hsa04662	B cell receptor signaling pathway	149,0742261	11,75392835	0	0
path:hsa04919	Thyroid hormone signaling pathway	182,1227329	11,66511078	0	0
path:hsa04211	Longevity regulating pathway	153,8382119	11,55160155	0	0
path:hsa04380	Osteoclast differentiation	174,9962572	11,44175005	0	0
path:hsa04072	Phospholipase D signaling pathway	81,67978489	10,44888945	0	0
path:hsa04660	T cell receptor signaling pathway	154,9192642	10,01976255	0	0
path:hsa04213	Longevity regulating pathway - multiple species	103,9274246	9,856324468	0	0
path:hsa04935	Growth hormone synthesis, secretion and action	140,6288458	9,602942328	0	0

path:hsa04070	Phosphatidylinositol signaling system	73,21778709	9,083096511	0	0
path:hsa04062	Chemokine signaling pathway	168,7593354	-9,035746942	0	0
path:hsa04725	Cholinergic synapse	79,17120763	8,74392333	0	0
path:hsa04917	Prolactin signaling pathway	160,9709	8,331293401	0	0
path:hsa04218	Cellular senescence	176,8844456	8,285306995	0	0
path:hsa04114	Oocyte meiosis	114,4885472	8,095141091	0	0
path:hsa04024	cAMP signaling pathway	164,252999	7,793042563	0	0
path:hsa04370	VEGF signaling pathway	85,79468148	7,469601017	0	0
path:hsa00562	Inositol phosphate metabolism	67,10262946	7,406562649	0	0
path:hsa04913	Ovarian steroidogenesis	45,01209095	7,219906668	0	0
path:hsa04664	Fc epsilon RI signaling pathway	75,864209	7,034005415	0	0
path:hsa04666	Fc gamma R-mediated phagocytosis	74,38638282	6,938589179	0	0
path:hsa04340	Hedgehog signaling pathway	72,84633834	6,933841452	0	0
path:hsa04020	Calcium signaling pathway	48,81351263	6,743736818	0	0
path:hsa04360	Axon guidance	103,9728822	-6,378179109	0	0
path:hsa04920	Adipocytokine signaling pathway	120,5543523	6,114021289	0	0
path:hsa04670	Leukocyte transendothelial migration	64,80644527	-6,081857123	0	0
path:hsa04022	cGMP-PKG signaling pathway	80,08003082	-5,663207935	0	0
path:hsa04928	Parathyroid hormone synthesis, secretion and action	126,8082061	5,578021111	0	0
path:hsa04621	NOD-like receptor signaling pathway	145,040538	-5,495515236	0	0
path:hsa04650	Natural killer cell mediated cytotoxicity	81,49229544	5,29890327	0	0
path:hsa04110	Cell cycle	152,7933055	5,240928889	0	0
path:hsa04510	Focal adhesion	144,8546582	5,10346321	0	0
path:hsa04921	Oxytocin signaling pathway	101,6671267	-5,08012143	0	0
path:hsa04350	TGF-beta signaling pathway	134,7072061	-4,574809342	0	0
path:hsa04912	GnRH signaling pathway	134,8040672	4,272689886	0	0
path:hsa04722	Neurotrophin signaling pathway	172,4353659	4,217531153	0	0
path:hsa04658	Th1 and Th2 cell differentiation	136,0205813	-4,164477971	0	0
path:hsa04726	Serotonergic synapse	40,8084486	4,080838561	0	0
path:hsa04910	Insulin signaling pathway	118,0627765	3,71438513	0	0

path:hsa04612	Antigen processing and presentation	42,57880102	3,625658289	0	0
path:hsa04520	Adherens junction	117,3158573	-3,550389218	0	0
path:hsa04728	Dopaminergic synapse	80,90481414	3,408395312	0	0
path:hsa00270	Cysteine and methionine metabolism	46,48497164	-3,288103047	0	0
path:hsa04659	Th17 cell differentiation	145,2721715	3,204832073	0	0
path:hsa04215	Apoptosis - multiple species	46,00580444	-3,177956502	0	0
path:hsa04657	IL-17 signaling pathway	123,466414	3,042926077	0	0
path:hsa04922	Glucagon signaling pathway	57,54681471	-2,919485835	0	0
path:hsa04625	C-type lectin receptor signaling pathway	167,4726315	2,536269683	0	0
path:hsa04310	Wnt signaling pathway	174,970501	-2,503860946	0	0
path:hsa04330	Notch signaling pathway	87,63855738	-2,185039634	0	0
path:hsa04060	Cytokine-cytokine receptor interaction	112,6969358	2,145876217	0	0
path:hsa04730	Long-term depression	45,430401	2,122797769	0	0
path:hsa04141	Protein processing in endoplasmic reticulum	54,70120769	2,069953995	0	0
path:hsa04720	Long-term potentiation	46,61956006	-2,027856838	0	0
path:hsa04068	FoxO signaling pathway	151,9102702	-1,958153684	0	0
path:hsa04540	Gap junction	45,9729722	1,934772946	0	0
path:hsa04144	Endocytosis	108,3216583	1,899940429	0	0
path:hsa04750	Inflammatory mediator regulation of TRP channels	58,62048697	1,873131251	0	0
path:hsa04714	Thermogenesis	69,62715886	-1,789894211	0	0
path:hsa04071	Sphingolipid signaling pathway	169,6188416	-1,757352014	0	0
path:hsa04390	Hippo signaling pathway	163,8735933	-1,483978157	0	0
path:hsa04064	NF-kappa B signaling pathway	123,3842217	1,319858015	0	0
path:hsa04961	Endocrine and other factor-regulated calcium reabsorption	72,54005521	1,273041494	0	0
path:hsa04261	Adrenergic signaling in cardiomyocytes	71,06957301	1,164096922	0	0
path:hsa04514	Cell adhesion molecules (CAMs)	40,30515551	-1,05790822	0	0
path:hsa04216	Ferroptosis	90,32069044	0,894824795	0	0
path:hsa01100	Metabolic pathways	131,3145036	0,854597625	0	0
path:hsa04371	Apelin signaling pathway	128,3218226	0,844375016	0	0
path:hsa04061	Viral protein interaction with cytokine and cytokine receptor	58,72914766	-0,735149719	0	0
path:hsa04137	Mitophagy - animal	132,6293247	0,720719511	0	0
path:hsa04925	Aldosterone synthesis and secretion	38,89964494	-0,598856738	0	0

path:hsa04623	Cytosolic DNA-sensing pathway	104,9896366	0,575589168	0	0
path:hsa04622	RIG-I-like receptor signaling pathway	116,7926341	0,558532059	0	0
path:hsa04217	Necroptosis	102,0338687	0,514081741	0	0
path:hsa04512	ECM-receptor interaction	47,3335375	-0,507817399	0	0
path:hsa00310	Lysine degradation	45,49584948	0,450087743	0	0
path:hsa04916	Melanogenesis	83,19349093	-0,285880124	0	0
path:hsa04015	Rap1 signaling pathway	124,5313268	0,280642994	0	0
path:hsa04611	Platelet activation	95,23159445	0,162052851	0	0
path:hsa04927	Cortisol synthesis and secretion	62,78701049	-0,141229637	0	0
path:hsa04973	Carbohydrate digestion and absorption	30,39245226	9,856636904	1,28E-13	3,15E-13
path:hsa04270	Vascular smooth muscle contraction	38,1943697	-7,265439138	1,77E-12	4,31E-12
path:hsa00590	Arachidonic acid metabolism	33,49064979	6,770690191	4,42E-12	1,07E-11
path:hsa04960	Aldosterone-regulated sodium reabsorption	28,07256595	8,253324799	2,62E-11	6,28E-11
path:hsa00020	Citrate cycle (TCA cycle)	15,11002609	-4,240619753	1,41E-08	3,34E-08
path:hsa04723	Retrograde endocannabinoid signaling	26,30113567	-3,124111608	2,34E-08	5,50E-08
path:hsa01200	Carbon metabolism	20,10369964	-4,561766885	5,14E-08	1,20E-07
path:hsa04976	Bile secretion	20,11556408	-3,171535791	6,43E-08	1,48E-07
path:hsa04918	Thyroid hormone synthesis	28,46853769	0,644480421	1,10E-07	2,52E-07
path:hsa03460	Fanconi anemia pathway	26,61821542	0,453914717	1,13E-07	2,56E-07
path:hsa04911	Insulin secretion	28,2853555	-0,50852594	1,87E-07	4,19E-07
path:hsa04713	Circadian entrainment	30,55797388	0,310883763	1,89E-07	4,21E-07
path:hsa04080	Neuroactive ligand-receptor interaction	27,47632754	0,624974902	2,55E-07	5,61E-07
path:hsa04710	Circadian rhythm	27,40799875	0,136224673	3,00E-07	6,56E-07
path:hsa04924	Renin secretion	22,1315126	2,254984244	4,31E-07	9,32E-07
path:hsa00760	Nicotinate and nicotinamide metabolism	22,25952164	0,865403838	5,34E-07	1,15E-06
path:hsa04962	Vasopressin-regulated water reabsorption	19,26198967	1,09841294	5,97E-07	1,27E-06
path:hsa03013	RNA transport	25,52550349	-0,266804086	6,52E-07	1,38E-06
path:hsa03320	PPAR signaling pathway	20,74187806	-1,596393732	6,88E-07	1,42E-06
path:hsa00520	Amino sugar and nucleotide sugar metabolism	5,923676905	0,837348021	6,88E-07	1,42E-06
path:hsa00524	Neomycin, kanamycin and gentamicin biosynthesis	5,912530246	0,722550958	6,88E-07	1,42E-06
path:hsa04971	Gastric acid secretion	18,69608677	-2,229113869	3,14E-06	6,40E-06

path:hsa00982	Drug metabolism - cytochrome P450	13,82588163	1,444947357	3,69E-06	7,47E-06
path:hsa00790	Folate biosynthesis	19,73105288	0,175782042	3,94E-06	7,91E-06
path:hsa03015	mRNA surveillance pathway	14,57940488	-2,278543517	4,76E-06	9,46E-06
path:hsa04610	Complement and coagulation cascades	16,04287161	1,170288645	4,79E-06	9,46E-06
path:hsa00830	Retinol metabolism	9,372323771	1,303391452	8,35E-06	1,63E-05
path:hsa03440	Homologous recombination	16,08041335	-0,808422593	8,38E-06	1,63E-05
path:hsa00330	Arginine and proline metabolism	13,64483314	-3,754164237	8,59E-06	1,66E-05
path:hsa04136	Autophagy - other	22,82485294	0,203823463	1,01E-05	1,93E-05
path:hsa00564	Glycerophospholipid metabolism	13,30880927	1,651635375	1,11E-05	2,11E-05
path:hsa04392	Hippo signaling pathway - multiple species	19,06330353	0,244780466	2,08E-05	3,92E-05
path:hsa04979	Cholesterol metabolism	14,99813657	-0,770899166	2,21E-05	4,13E-05
path:hsa04672	Intestinal immune network for IgA production	13,04809231	2,352593391	2,47E-05	4,59E-05
path:hsa00220	Arginine biosynthesis	9,582466218	-2,806145574	3,35E-05	6,17E-05
path:hsa00230	Purine metabolism	12,34735859	-4,096120955	8,47E-05	1,55E-04
path:hsa00350	Tyrosine metabolism	10,98655126	-1,586018351	1,11E-04	2,01E-04
path:hsa04970	Salivary secretion	14,63862175	-1,03063087	1,14E-04	2,05E-04
path:hsa00980	Metabolism of xenobiotics by cytochrome P450	13,42820958	0,411942911	1,19E-04	2,14E-04
path:hsa00591	Linoleic acid metabolism	8,242482635	0,494892259	1,45E-04	2,59E-04
path:hsa00480	Glutathione metabolism	11,96869331	0,345190339	1,50E-04	2,65E-04
path:hsa00563	Glycosylphosphatidylinositol (GPI)-anchor biosynthesis	5,82666549	1,259513228	2,51E-04	4,40E-04
path:hsa00140	Steroid hormone biosynthesis	10,6544011	-1,041542296	3,62E-04	6,30E-04
path:hsa04978	Mineral absorption	12,00974823	0,733605949	3,84E-04	6,64E-04
path:hsa00620	Pyruvate metabolism	7,514581609	-3,433252716	4,03E-04	6,93E-04
path:hsa00380	Tryptophan metabolism	7,700085338	-2,086353152	4,34E-04	7,39E-04
path:hsa04972	Pancreatic secretion	14,12277648	-0,15774068	4,72E-04	8,00E-04
path:hsa00010	Glycolysis / Gluconeogenesis	8,846136868	-2,141730824	5,45E-04	9,16E-04
path:hsa00600	Sphingolipid metabolism	6,738974375	2,83398622	6,85E-04	0,001143973
path:hsa04724	Glutamatergic synapse	12,50259547	-0,922993889	7,00E-04	0,001162587
path:hsa00250	Alanine, aspartate and glutamate metabolism	4,280692482	-3,596182307	9,62E-04	0,001585869
path:hsa01230	Biosynthesis of amino acids	7,491950149	-3,714356037	0,001175639	0,001925883
path:hsa00260	Glycine, serine and threonine metabolism	7,089264679	0,966554537	0,001546588	0,002516997
path:hsa01210	2-Oxocarboxylic acid metabolism	5,764639832	-0,444960906	0,002263756	0,003660228

path:hsa01212	Fatty acid metabolism	7,101805253	-1,180258444	0,002429858	0,00390345
path:hsa00062	Fatty acid elongation	6,025077579	-1,044796192	0,002565201	0,004094456
path:hsa00051	Fructose and mannose metabolism	5,157076777	0,86826671	0,003509478	0,005565988
path:hsa00730	Thiamine metabolism	5,691658398	0,166464637	0,003723118	0,005867446
path:hsa04975	Fat digestion and absorption	4,699647392	0,329065973	0,0038377	0,006009983
path:hsa00500	Starch and sucrose metabolism	4,453183521	0,673174151	0,003889218	0,006052595
path:hsa04260	Cardiac muscle contraction	8,887718033	-0,205739502	0,004023404	0,006222532
path:hsa00670	One carbon pool by folate	9,336141409	-0,047561585	0,004985854	0,007663442
path:hsa00640	Propanoate metabolism	4,308827951	-2,356880552	0,005642957	0,008620223
path:hsa04977	Vitamin digestion and absorption	0	-0,310734758	0,005997001	0,009105203
path:hsa00052	Galactose metabolism	4,548817841	0,764342437	0,006261492	0,009449161
path:hsa00983	Drug metabolism - other enzymes	8,445067453	0,15811195	0,006317658	0,009476487
path:hsa00561	Glycerolipid metabolism	5,290737357	0,334112188	0,006671613	0,009947495

Supplementary table 7 - Table with feature comparisons of RNAdetector vs other RNA-Seq pipelines.

	Deployment	Supported OS	Offline	GUI	Input files	Aligners	Counting tools	DEA tools	Downstream analysis	Settings for ncRNA analysis	Multi-species supported	Graphical final report
RNAdetector	Docker. No previous dependencies required.	Windows MacOS Linux	X	X	FASTQ, BAM, SAM	STAR HISAT2 BWA SALMON	featureCounts HTSeq SALMON CIRI and CIRIquant for circRNAs	DESeq edgeR LIMMA	miRNA sensitive topological pathway analysis	X	X	X
ArrayExpressHTS	Bioconductor. It can be run locally or remotely at EBI cloud	Linux MacOS	X	-	FASTQ	BOWTIE TOPHAT BWA	Cufflinks MMSEQ	-	-	-	X	-
BioJupies	Web-based application on Jupyter Notebooks.	Windows MacOS Linux	-	X	FASTQ	Kallisto	Kallisto	LIMMA Characteristic Direction	Several enrichment analyses are supported	-	X	X
BioWardrobe	Standalone (it seems to be no longer maintained)	MacOS Linux	X	X	FASTQ SRA	STAR	STAR	DESeq	-	-	X	Only figures
DEWE	Docker (it seems to be no longer maintained)	Windows MacOS Linux	X	X	FASTQ	BOWTIE2 HISAT2	StringTie HTSeq	Ballgown edgeR	GSEA	-	X	X
easyRNASeq	Bioconductor	Windows MacOS Linux	X	-	BAM	-	IRanges GenomicRanges	-	-	-	X	-
ExpressionPlot	Standalone software that runs on a virtual machine (it seems to be no longer maintained)	Windows MacOS Linux	X	X	FASTQ, BAM	BOWTIE	-	DESeq	-	-	X	X
FX	Amazon cloud system or it can be installed on local Hadoop clusters (it seems to be no longer maintained)	Windows MacOS Linux	-	X	FASTQ SAM	GSNAP	-	-	SNP and INDEL detection	-	Only human and mouse	-
GENE-Counter	Standalone. Several dependencies are quired.	MacOS Linux	X	-	FASTQ	CASHX BOWTIE BWA	CASHX	NBPSeq edgeR DESeq	GO analysis	-	X	-
GeneProf	Cloud-based application (it seems to be no longer maintained)	Windows MacOS Linux	-	X	SRA importer tool	BOWTIE TOPHAT	-	DESeq edgeR	GO analysis	-	X	X
Grape RNA-Seq	Standalone. Deployment with Docker or Conda are also available.	MacOS Linux	X	-	FASTQ SAM BAM	GEM	FluxCapacitor	-	-	-	X	X
MAP-RSeq	standalone virtual machine or parallel Sun Grid Engine cluster. Several dependencies are required.	Windows MacOS Linux	X	-	FASTQ	TOPHAT	HTSeq featureCounts	-	SNP calling, Fusion transcript detection	-	Only human	X
RAP	Cloud application	Windows MacOS Linux	-	X	FASTQ SRA BAM SAM	TOPHAT	Cufflinks HTSeq	Cuffdiff2 DESeq	Splicing junction detection, Exon usage analysis, Fusion transcript detection, Differential polyA analysis	-	X	X
RobiNA	Standalone software	Windows MacOS Linux	X	X	FASTQ BAM SAM	BOWTIE	-	DESeq edgeR	-	-	X	X
RSEQtools	Standalone	MacOS Linux	X	-	MRF	-	mrfQuantifier	-	-	-	X	-

RseqFlow	Standalone tool on Pegasus virtual machine (it seems to be no longer maintained)	Windows MacOS Linux	X	-	Single ended reads in FASTQ format	BOWTIE PerM	-	DESeq	SNP calling	-	X	-
S-MART	Standalone	Windows MacOS Linux	X	X	FASTQ SAM	-	-	Independent method developed by the authors	-	-	X	Only figures
TCW	Java desktop application. Several dependencies are required	MacOS Linux	X	X	FASTA	-	-	edgeR DESeq EDASeq DEGseq	GO analysis	-	X	X
TRAPLINE	Galaxy web application	Windows MacOS Linux	-	X	FASTQ	TOPHAT 2	Cufflinks	Cuffdiff2	Splicing junction detection, SNP detection, GO analysis, Protein interaction, miRNA target prediction	miRNAs	X	-
wapRNA	Web application or executable packages for installation on user's local server (it seems to be no longer maintained)	Linux	-	X	FASTA FASTQ	CoronaLite BWA	in-house built Perl module	DEGseq	GO analysis, KEGG pathway functional enrichment, miRNA target prediction	miRNAs	X	Figures and tables

Supplementary table 8 - Table with feature comparisons of RNAdetector vs other ncRNA-Seq pipelines.

	Deployment	***Supported OS***	***GUI***	***Input files***	***Aligners***	***Counting***	***DEA tools***	***Downstream analysis***	***Regulatory ncRNA supported***	***More species supported***	***Graphical final report***
RNAdetector	Docker	Windows MacOS Linux	X	FASTQ BAM SAM	STAR HISAT2 BWA SALMON	featureCounts HTseq SALMON	DESeq edgeR LIMMA	miRNA-sensitive topological pathway analysis (MITHrIL)	miRNAs, snoRNAs, piRNAs, tsRNAs, tUCRs, lncRNAs, circRNAs	X	X
iSmaRT	Standalone (website does not work. Not maintained)	Linux	X	FASTQ	BOWTIE	sRNAbench	DESeq edgeR NOISeq	GO and pathway enrichment analysis. miRNA\piRNA target prediction.	miRNAs, piRNAs	Human, mouse, rat	Only txt files and figures in output
iSRAP	Standalone	MacOS, inux	-	FASTQ BAM	BOWTIE2	BEDTools	DESeq edgeR LIMMA	-	miRNAs, piRNAs, snoRNAs	X	Only txt files and figures in output
miARma-Seq	Docker	Windows MacOS Linux	-	FASTQ BAM	BOWTIE2 , BWA	featureCounts	edgeR NOISeq	GO and pathway enrichment analysis, miRNA target prediction	miRNAs, circRNAs	X	Only txt files and figures in output
Oasis 2	Web-based	Windows MacOS Linux	X	FASTQ	STAR	featureCounts	DESeq	GO and pathway enrichment analysis, miRNA target prediction	miRNAs, piRNAs, snoRNAs	X	Only txt files and figures in output
SPORTS1.0	Standalone	Linux	-	FASTQ	BOWTIE	?	-	-	miRNAs, piRNAs, snoRNAs, tsRNA	X	Only txt files and figures in output
sRNAnalyzer	Standalone	MacOS Linux	-	FASTQ	BOWTIE	?	-	-	miRNAs, piRNAs, snoRNAs, lncRNA	X	-
sRNApipe	Galaxy server installed in user's machine	MacOS Linux	X	Single-end FASTQ with no adaptors	BWA	?	-	-	miRNAs, piRNAs, snoRNAs	X	X

La borsa di dottorato è stata cofinanziata con risorse del
Programma Operativo Nazionale Ricerca e Innovazione 2014-2020 (CCI 2014IT16M2OP005),
Fondo Sociale Europeo, Azione I.1 “Dottorati Innovativi con caratterizzazione Industriale”

UNIONE EUROPEA
Fondo Sociale Europeo