frontiers in
GENETICS

REVIEW ARTICLE
published: 17 December 2012
doi: 10.3389/fgene.2012.00287

Bioinformatics of cancer ncRNA in high throughput sequencing: present state and challenges

Natasha Andressa Nogueira Jorge[1], Carlos Gil Ferreira[2] and Fabio Passetti[1]*

[1] *Bioinformatics Unit, Clinical Research Coordination, Instituto Nacional de Câncer, Rio de Janeiro, Brazil*
[2] *Clinical Research Coordination, Instituto Nacional de Câncer, Rio de Janeiro, Brazil*

Edited by:
Peng Jin, Emory University School of Medicine, USA

Reviewed by:
Peng Jin, Emory University School of Medicine, USA
Hongyan Xu, Georgia Health Sciences University, USA

Correspondence:
Fabio Passetti, Bioinformatics Unit, Clinical Research Coordination, Instituto Nacional de Câncer, Rua André Cavalcanti, 37 – Centro, Rio de Janeiro 20231-050, Brazil.
e-mail: passetti@inca.gov.br

The numerous genome sequencing projects produced unprecedented amount of data providing significant information to the discovery of novel non-coding RNA (ncRNA). Several ncRNAs have been described to control gene expression and display important role during cell differentiation and homeostasis. In the last decade, high throughput methods in conjunction with approaches in bioinformatics have been used to identify, classify, and evaluate the expression of hundreds of ncRNA in normal and pathological states, such as cancer. Patient outcomes have been already associated with differential expression of ncRNAs in normal and tumoral tissues, providing new insights in the development of innovative therapeutic strategies in oncology. In this review, we present and discuss bioinformatics advances in the development of computational approaches to analyze and discover ncRNA data in oncology using high throughput sequencing technologies.

Keywords: bioinformatics, high throughput sequencing, cancer, non-coding RNA, gene expression

INTRODUCTION

The ENCODE project discovered that most of the human genome is transcribed, but only a tiny fraction of human DNA encode for proteins (ENCODE Project Consortium et al., 2007; Elgar and Vavouri, 2008). The remaining transcriptome is defined as non-coding RNA (ncRNA) and is divided into distinct classes, each of them with its own three-dimensional folding and presenting a specific function. Some ncRNA classes are known for years, such as ribosomal and transport RNAs (essential to translation); small nucleolar RNAs (snoRNAs; biogenesis and control of ribosome activity); and small nuclear RNAs (to promote splicing of pre-mRNAs). Recently, additional ncRNA classes have been described and shown to be able to repress gene expression (microRNAs, miRNA); to regulate cellular proliferation, apoptosis (small interfering RNAs, siRNAs), and imprinting (long non-coding RNAs, lncRNA); and also to inhibit transposon and DNA methylation (PIWI-interacting RNAs, piRNA; for a detailed description of the known ncRNAs, see Eddy, 2001; Mitra et al., 2012).

The most studied ncRNA class in oncology is miRNA. These small RNAs have on average 22 nucleotides in length and mediate gene silencing by partially paring with specific regions of messenger RNAs (mRNA) to prevent its translation (Wu et al., 2012). The miRNA target genes are usually related to fundamental cellular processes like proliferation, differentiation, apoptosis, and development (Schulte et al., 2010). Aberrations in miRNAs expression levels have been extensively studied in several types of cancer as they may act as tumor suppressor genes or oncogenes (Meiri et al., 2010).

Additionally, two ncRNA classes with special attention in studies in oncology are lncRNA and piRNA. The lncRNAs are more than 200 nucleotides long and although most of them have not been fully characterized, they have been related to the regulation of several cellular processes such as epigenetics, differentiation, proliferation, and nuclear import (Tahira et al., 2011). Recent studies reported alterations in different lncRNAs in several types of cancer (Reis et al., 2004; Guffanti et al., 2009; Cheng et al., 2011; Cui et al., 2011; Esposito et al., 2011; Prensner et al., 2011; Tahira et al., 2011; Yang et al., 2012a,b). The piRNA class has also been related to have a possible involvement in the biogenesis of cancer. The piRNAs interact with PIWI proteins in order to promote silencing of transposable elements and maintain DNA integrity (Cheng et al., 2011).

Since 1977, when the first genome was sequenced, the DNA sequencing technology has been evolving to higher throughput and lower cost (Kircher and Kelso, 2010). Current high throughput sequencing (HTS), also known as next-generation sequencing, provides the opportunity to obtain a more accurate profiling with higher resolution, increased throughput, sequencing depth, and low experimental complexity (Prensner et al., 2011; Zhou et al., 2011). One characteristic of this technology is the amount of data produced, making methods in bioinformatics essential for its analysis.

Bioinformatics emerged as a multidisciplinary discipline which aimed to analyze biological data using programming techniques and the computational processing power. The first studies in Bioinformatics were performed in the early 1960s, when the first computational approaches were used to address gene and protein sequences (for a time line review, see Hagen, 2000). The term bioinformatics was coined by Hesper and Hogeweg (1970) as "the study of informatics processes in biotic systems" (Hogeweg, 2011). However, after the emergence of high throughput methods in molecular biology and the establishment of the Human Genome Program in 1990, the definition of bioinformatics has shifted to assist in the management, storage, visualization, and analysis of large amounts of data. In conjunction to the development of bioinformatics tools, many molecular biology techniques

were created in the last two decades such as qPCR, microarray, tilling array and SAGE, which permitted to quantify gene expression. A large number of studies have been taken using molecular biology techniques to produce large amounts of raw data and bioinformatics tools to assist the biological interpretation of the findings. An example of the importance of bioinformatics to the science was the announcement of the draft of the human genome in 2001, which was presented after the development of a computational tool to assemble the unsorted fragments of the human genome (Kent and Haussler, 2001; Lander et al., 2001).

As depicted in **Figure 1**, bioinformatics can assist two types of research: disease-oriented (e.g., cancer) and methodologically driven (e.g., HTS). In the former, several technologies can be used to study distinct biological patterns and then a systems biology approach is taken to assist in the comprehension of cancer. In the latter, an unique molecular biology technique is used to answer a specific interrogation, for example, the expression pattern of human genes after a group of patients received a standard treatment against a specific cancer type.

In this review, we present some examples of ncRNA discovered, its potential to be used as cancer biomarkers and the role and challenges in bioinformatics to analyze HTS data.

WHY STUDYING NON-CODING RNAs IN CANCER?

Calin et al. (2002) documented the first differentially expressed ncRNA in cancer samples. The small RNAs *miR-15* and *miR-16* were described to be deleted or down regulated in more than half of the patients with Chronic Lymphocytic Leukemia (CLL) and B-cell CLL. The absence of those genes led to an over expression of the *Bcl-2* gene, preventing apoptosis. Two years later, additional data revealed that some miRNAs genes are located at fragile and frequently altered sites in cancer, including regions with amplifications, loss of heterozygosity, or breakpoints (Calin et al., 2004). Since then, several other reports have presented alterations related to ncRNAs in different cancer samples.

One of the first approaches to associate ncRNA and oncology was performed by Mishra et al. (2007). The authors evaluated polymorphisms in the human dihydrofolate reductase (DHFR) mRNA binding site for *miR-24*. As result, the polymorphism led to the loss of *miR-24* function and resulted in *DHFR* overexpression, increasing resistance to chemotherapy. Among miRNAs, the oncogene *miR-21* has been extensively studied (Dillhoff et al., 2008; Frankel et al., 2008; Krichevsky and Gabriely, 2009; Li et al., 2009a,b; Rabinowits et al., 2009; Ribas et al., 2009; Seike et al., 2009; Wickramasinghe et al., 2009; Iliopoulos et al., 2010). This miRNA appears over expressed in different tumor samples and targets *PTEN*, *PDCD4*, *TPM1*, and *Maspin* human genes, promoting growth, migration, and invasion in different tumor types (Zhu et al., 2008).

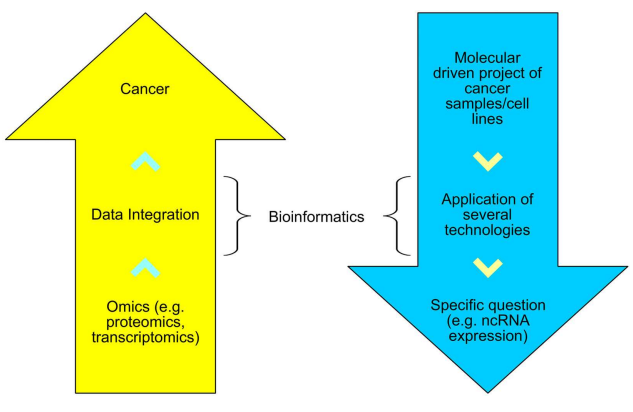

FIGURE 1 | The disease-oriented and methodological-driven types of research assisted by bioinformatics.

Regarding lncRNAs, recently, a single nucleotide polymorphism located in the *ANRIL* gene was associated with the number of plexiform neurofibromas in neurofibromatosis type 1 patients. Moreover, one of its allele was associated with low levels of *ANRIL*, suggesting a relation between the *ANRIL* and the susceptibility to plexiform neurofibromas (Pasmant et al., 2011). In addition, in a recent review, Gustschner and Diederichs (2012) were able to link cellular processes influenced by lncRNAs to the hallmarks of cancer.

Several studies associating cancer and ncRNA aim to discover molecular signatures for diagnosis and prognosis. In this direction, cancer biomarkers are molecular features that are produced either by the tumor or by the host as a response due to the change of the default cell metabolism. Examples of possible biomarkers are mutations and alterations in gene expression and epigenetics (for a deep view of cancer epigenetics, see Brait and Sidransky, 2011). The identification of specific cancer biomarkers may provide parameters for cancer early detection, diagnosis, prognosis, prediction of response to anticancer treatments, prediction of recurrence, and identification of putative drug targets. However, due to cancer complexity, it has been recently suggested that single biomarker may not be adequate for clinical practice and it is suggested to use a set of biomarkers in a panel (Tainsky, 2009). The study of Hennessey et al. (2012) compared the miRNA expression profile in the serum of non-small cell lung cancer (NSLC) patients and healthy individuals. The authors proposed the combination of the expression levels of *miR-15b* and *miR-27b* would be able to discriminate the healthy and the sick individuals. Another study in NSLC was performed by Chen et al. (2012) in which it is suggested a 10 miRNA panel to differentiate tumor types. Wu et al. (2012) analyzed the serum of 42 breast cancer patients and were able to detect more than 800 circulating miRNAs and associate them with tumor status. The low levels of miRNA *miR-375* and high levels of miRNA *miR-122* have been suggested as biomarkers for predicting metastasis in early patients. In this direction, Liu et al. (2011) compared the expression of miRNAs in the serum of 20 patients with gastric cancer against 20 normal samples. Among the 19 over expressed miRNA identified, the *miR-1*, *miR-20a*, *miR-27a*, *miR-34*, and *miR-423-5p* have been identified as potential biomarkers for gastric cancer diagnostics and tumor profiling.

Another aspect of ncRNA and cancer is the possibility to associate them with drug resistance. A very large effort to comprehend the role of drug activity and resistance in cancer cell lines was performed by Liu et al. (2010). The microarray technology has been used to evaluate the mRNA and miRNA expression profiling of

the 60 cancer cell lines of the National Cancer Institute Developmental Therapeutics Program, also known as the NCI-60 panel. The authors used bioinformatics approaches to analyze and cluster some cell groups according to their tissue of origin and to associate the levels of mRNAs and miRNAs with sensitivity or resistance to many drugs routinely used in the clinic. To facilitate the visualization of the data produced, the authors developed the CellMiner, a web based tool very useful to clinicians and researchers from basic to applied research (Reinhold et al., 2012).

The aforementioned studies exemplify how miRNA are involved in cancer development and progression. Another advantage of analyzing small ncRNA profile in cancer regards the distinct types of samples may be use to study it, from fresh tissues, body fluids (including blood, urine, and saliva), and formalin-fixed, paraffin-embedded (FFPE) tissues (Lussier et al., 2012). Therefore, the study of ncRNAs and its expression profiling in cancer cells may help understand the mechanisms of the disease and improve diagnostics and prognostics by personalizing cancer treatment (Hu et al., 2010).

WHY USING HTS FOR ncRNA PROFILING IN CANCER?

The most common approach used to study ncRNA is to first produce large-scale profiling on microarray followed by validation by more specific techniques such as microarray with fewer probes or multiplexed RT-PCR. Regarding ncRNAs, miRNA microarrays provide an overview of the set of miRNAs in a sample and can be further validated by northen blot, Rnase protection assay, primer extension assay, quantitative RT-PCR, and *in situ* hybridization (Tainsky, 2009). However, with the advent of HTS technology, it is possible not only to infer the expression level of ncRNA, but also to detect uncharacterized ones. Another advantage of HTS over other existing expression profiling technologies is the fact that the process requires no previous information about the transcripts that will have its expression quantified (Isakov et al., 2012). This characteristic of HTS is suggestive for its use in the quantification of the heterogeneous transcriptome of cancer (Meyerson et al., 2010). Distinct from other techniques, HTS does not use specific or random probes, instead, the RNA molecules from the sample are linked to adaptors and amplified by PCR (McCormick et al., 2011), permitting the sequencing of the exact transcript on a single nucleotide resolution (Zhou et al., 2011). This step allows the identification of variations in length or composition, deletions, duplications, low abundant, and novel transcripts present in cancer samples (Meyerson et al., 2010). **Figure 2** depicts some advantages of HTS over other techniques and how bioinformatics is essential to analyze them.

A comparison between the expression profile using HTS and microarray was performed by Weng et al. (2010). The authors used HTS technology to evaluate the profile of small RNAs in three paired clear cell renal cell carcinoma (ccRCC) FFPE samples and performed miRNA microarray and RT-PCR to validate the results from the former. Besides the known miRNA genes, the HTS experiments were able to reveal million of short sequences that included sequences from snoRNAs, sRNA, snRNA, tRNAs, rRNAs, introns, exons, and several others, including unknown nucleotide sequences. Bioinformatics techniques were used to cluster the miRNA detected and to distinguish between tumor and normal

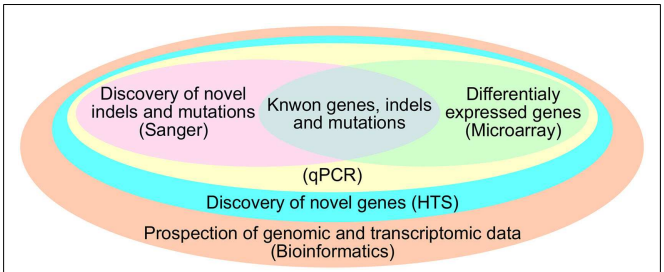

FIGURE 2 | Advantages of Bioinformatics and HTS over other techniques.

samples. The miRNA microarray were able to detect up to 453 miRNAs, while the HTS could identify up to 598 miRNAs and both platforms showed correlated expression levels that were validated by RT-PCR in seven randomly chosen altered miRNAs. As can be observed, HTS let to the quantification of 145 additional ncRNAs not present in the microarray experiment.

Several ncRNA HTS studies revealed putative novel ncRNAs (Jima et al., 2010; Keller et al., 2011; Prensner et al., 2011). Deep sequencing of the enriched Poly(A) transcriptome was used to evaluate the expression of both protein coding and lncRNAs in cancer samples by Prensner et al. (2011) in 102 prostate tissues and cell lines, including normal samples and benign, localized, and metastatic samples. The authors were able to describe the novel lncRNA *PCAT-1*, over expressed in metastatic samples. Further experiments pointed it as a prostate specific regulator of cell proliferation that targets the Policomb Repressive Complex 2 (*PRC2*). Jima et al. (2010) evaluated small ncRNAs in normal and malignant B cells. The authors proposed a panel of known and novel miRNAs to distinguish between the subgroups of lymphoma and found that one previously annotated miRNA cluster has its expression levels inversely correlated with its putative targets *SMAD2* and *SMAD3*, known mediators of the transforming growth factor-β (TGF-β) signaling pathway. Keller et al. (2011) evaluated the miRNAs differentially expressed in the blood of NSLC patients and found some unknown miRNAs, including novel mature forms from known precursors.

Another example of HTS as tool to the identification of novel small ncRNA class is found in the study of Meiri et al. (2010). The authors used HTS to evaluate the miRNA transcriptome of 23 solid tumor samples, including breast, bladder, colon, and lung. They discovered 49 novel miRNA and sequence variants with different expression patterns among the samples and identified a novel class of small ncRNAs derived from Y-RNAs and endogenous siRNAs.

Most of the HTS studies published so far have tried to identify miRNA to use as diagnostic or prognostic biomarkers in solid tumors or in circulation. The two studies by Wu et al. (2012) and Liu et al. (2011) referred to in the previous section used HTS to infer their candidate biomarkers. Martens-Uzunova et al. (2012) and Ryu et al. (2011) went further. Martens-Uzunova et al. (2012) used the miRNA expression found in one organ-confined and one metastatic lymph node tumor samples of prostate cancer to create a miR-classifier that was able to correctly distinguish 89% of the prostate cancer cell samples. Besides miRNA, the

experiment was able to find snoRNAs and tRNAs with altered expression levels and novel miRNA with very low counts. Ryu et al. (2011) applied a bioinformatics approach to validate the novel miRNAs in breast cancer cell lines. The authors obtained 189 putative novel miRNAs, considering thermodynamics stability, presence of complementary sequences, and phylogenetic conservation.

There are several HTS platforms commercially available, each with its own characteristics such as data throughput, read length, error rate, and price (Zhou et al., 2011). Therefore, the choice of the platform to be used must be according to its characteristics and the needs of the experiment. Kircher and Kelso (2010) reviewed the sequencing technologies of some HTS platforms and Toedling et al. (2012) present the comparison of different sequencing protocols and the results obtained. The authors recommend comparing data generated only by the same protocol.

HOW COMPUTATIONAL PROCEDURES CAN AID ncRNA HTS PROFILING?

High throughput sequencing experiments generate a large amount of data, hence bioinformatics methods are necessary for the proper storage, visualization, and analysis. After sequencing, one or more text files are produced in the fasta, fastq or csfasta, and qual formats, depending on the equipment settings and platform used. These files contain the nucleotides sequenced for each read and a quality score for each base/color call (Isakov and Shomron, 2011). Usually, the sequencer manufacturer provides software able to process this data in the very beginning steps toward publication. In this section, we will discuss available independent tools for each step of the downstream analysis. **Figure 3** shows some of the steps for HTS analysis.

Among the sequenced data, it is common to find reads with miscalled bases, unidentified bases, poor quality, and adaptor contamination. Those artifacts must be removed before alignment to avoid wrong mapping and also to save computational time (Patel and Jain, 2012).

For the removal of low quality reads and unidentified bases, some authors use their own script as described, for example, by Meiri et al. (2010). However, other studies use public available toolkits, like Fastx-toolkit (Gordon and Hannon, unpublished) and QC Toolkit (Patel and Jain, 2012). The aforementioned tools are a collection of programs for processing short reads fastq and fasta files and reporting the quality of sequencing run, filtering reads for their quality, and removing unknown nucleotides.

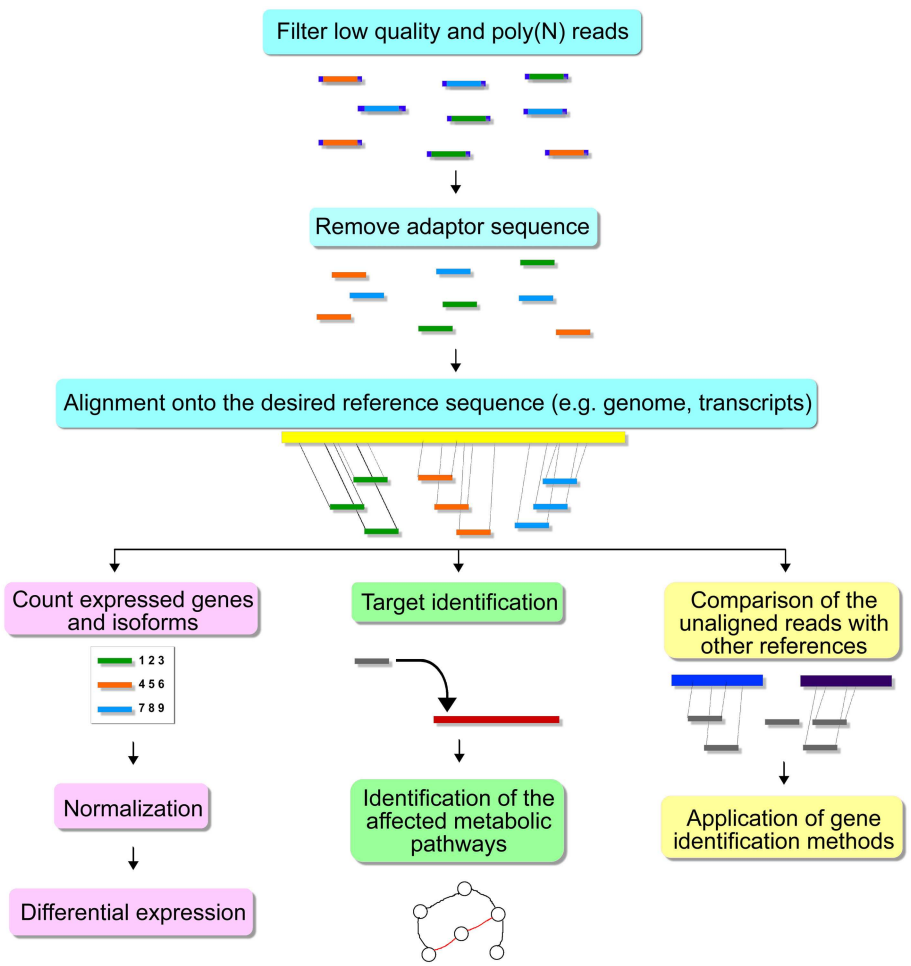

FIGURE 3 | Steps for HTS analysis.

If the aim is to sequence short RNAs (sRNAs), most probably the size of the desired sRNA is smaller than the read's length (Martin, 2011). In this case, a subsequence of the adaptor used in the sequencing process will be present in the final result and, because it does not belong to the sequenced genome, it must be removed (McCormick et al., 2011). Both of the toolkits mentioned above can remove those sequences. Other tools include the Cutadapt program (Martin, 2011), the Bioconductor's package for short read processing called Biostrings (Pàges et al., 2012) and the aligners Novoalign (Hercus, 2008) and SOAP (version 1; Li et al., 2008). **Table 1** presents some preprocessing alignment tools. The Bioconductor's packages Biostrings (Pàges et al., 2012) and ShortRead (Morgan et al., 2009) together can assess the quality and remove adaptor sequence from fasta and fastq files, but they require user knowledge of programming language R and Bioconductor. The Cutadapt algorithm can remove the adaptor sequence from the reads obtained by the major sequencing platforms, but, differently from the aforementioned algorithms, it cannot access or filter low quality reads (Martin, 2011). Regarding the mentioned aligners, its adaptor removal propriety is linked to the alignment algorithm; therefore they cannot be applied if the user wishes to use another alignment tool.

The next step of the analysis is aligning sequence reads onto the genome of the reference organism. This can be a computationally demanding task due to the great volume of short sequences produced and also nucleotide and structural variance, sequencing errors, RNA editing, and epigenetic modifications (Isakov and Shomron, 2011), especially for the traditional alignment tools (Lee et al., 2011). Hence, a new generation of short read aligners has been developed, saving computational time by indexing the read sequences, or the genome prior alignment (Lee et al., 2011). Several aspects of the aligner must be considered: memory and time requirements and limitations, and how the tool is adequate to the task (Isakov and Shomron, 2011). For instance, many short read aligners can be programmed to return the results of the reads whose first part perfectly matches the reference genome, which allows to search for potential isoforms of miRNA (Motameny et al., 2010). In this direction, the Novoalign software has a special option to align miRNA in which it searches for regions complementary to the reads near the mapped loci (Hercus, 2008). Most sequence aligners generate results in the sam file format which can be processed by the SAMtools kit (Li et al., 2009c; Isakov and Shomron, 2011). One thing worth noticing is that when a short sequence is aligned to a large and complex genome with repetitive regions, such as the human genome, is expected to find reads mapped in multiple locations in the genome (McCormick et al., 2011). Most software does not report such results as default, resulting in the loss of some sequences (Motameny et al., 2010). Other strategies to manipulate such reads are to divide their count by all putative loci and their estimate a proportion according to the levels of uniquely mapped reads in neighbor loci (McCormick et al., 2011). Some alignment tools for HTS data are shown in **Table 2** and were evaluated by Ruffalo et al. (2011).

As important as the aligner, is the database to map the processed reads. There are several genome and ncRNA databases available, but the most commonly used sequence databases for studying cancer are the following: the human genome hg18 assembly provided by the UCSC Genome Bioinformatics group (Dreszer et al., 2012), miRBase (Kozomara and Griffiths-Jones, 2011) and Rfam (Gardner et al., 2011). It is important to notice that the human genome sequence in the hg18 version provided through the UCSC Genome Browser website is identical to the NCBI36 version. **Table 3** exemplifies some of these databases.

Regarding ncRNA analysis, it is important to use annotation databases having information regarding the annotation of prediction and experimentally defined ncRNAs. The UCSC Table Browser provides open accesses to high quality human genome annotation including alignment of RefSeq genes, mRNAs and EST from GenBank and also other gene and gene prediction tracks such as Ensembl Genes (Karolchik et al., 2004). Currently, this tool is under migration to the latest version of the human genome sequence (hg19/NCBI37; Dreszer et al., 2012). One another important source of annotation files for studying ncRNA is ncRNA.org, which is part of the Functional RNA database and is an extended mirror of the UCSC Genome Browser. NcRNA.org displays information about functional ncRNAs and associated elements in the hg17 and hg18 versions of the human genome (Mituyama et al., 2009). Another frequently database used in studies in oncology and HTS is the miRBase (Kozomara and Griffiths-Jones, 2011). This database is the primary source for miRNA sequence and annotation. The miRBase effort has the objective to provide curated nomenclature scheme for known and novel miRNAs, to act as central repository for mature and precursor miRNA sequence and also to provide access to the primary evidence that supports miRNA annotations. Another database used in researches that go beyond the miRNA family is named Rfam. This database maintains

Table 1 | Preprocessing alignment tools.

Name	Site	Description	Authors
Fastx-toolkit	http://hannonlab.cshl.edu/fastx_toolkit/	FASTA/FASTQ file processing	Gordon and Hannon (unpublished)
QC tools	http://www.nipgr.res.in/ngsqctoolkit.html	Ilumina and Roche 454 FASTQ file processing	Patel and Jain (2012)
Cutadapt	http://code.google.com/p/cutadapt/	Removes adapter sequence	Martin (2011)
ShortRead	http://bioconductor.org/packages/2.10/bioc/html/ShortRead.html	FASTA/FASTQ file processing	Morgan et al. (2009)
Biostrings	http://bioconductor.org/packages/2.10/bioc/html/Biostrings.html	String objects representing biological sequences, and matching algorithms	Pàges et al. (2012); R package version 2.24.1

Table 2 | Alignment tools.

Name	Site	Authors
Soap	http://soap.genomics.org.cn/soapaligner.html	Li et al. (2008)
Bwa	http://bio-bwa.sourceforge.net/	Li and Durbin (2009)
Bowtie	http://bowtie-bio.sourceforge.net/index.shtml	Langmead et al. (2009)
Novoalign	http://www.novocraft.com/main/index.php	Hercus (2008)

Table 3 | Sequence databases.

Name	Site	Description	Authors
UCSC hg18/NCBI36	http://genome.ucsc.edu/	Human genome sequence	International Human Genome Sequencing Consortium
ncRNA.org	http://www.ncrna.org/	ncRNA database sequence	Mituyama et al. (2009)
miRBase	http://www.mirbase.org/	miRNA database sequence	Kozomara and Griffiths-Jones (2011)
Rfam	http://rfam.sanger.ac.uk/	ncRNA database sequence	Gardner et al. (2011)

automated and curated sequences, alignments, secondary structure, and annotations of several ncRNAs families. Each family represents a set of RNA sequences that share a common ancestral (Gardner et al., 2011).

All the aforementioned tools require Linux and programming knowledge from the end user. Aiming to assist small to medium bioinformatics research groups to analyze miRNA HTS, several pipelines have been developed for processing raw files, identify novel transcripts, calculate differential expression, and provide fast annotation of genomic coordinates and single nucleotide variations (revised by Li et al., 2012; **Table 4**). One exception is the RandA pipeline (Isakov et al., 2012), that uses the whole Rfam database, and can be applied to different ncRNAs. Segtor (Renaud et al., 2011) is another tool that works to assist in one important step in the biological interpretation effort of every HTS experiment. Segtor allows the fast annotation of sequences from a given HTS experiment and provide a list of ncRNA genes affected by multiple types of nucleotide polymorphisms.

One of the advantages of HTS over other profile techniques resides in the fact that its quantification is based on how many reads were mapped in the same region/transcript. However, the read count is subject to sample and experimental variation, therefore, they must be normalized to be compared to other samples (Datta et al., 2010). There are several normalization methods, like linear total count scaling, quantile-based, trimmed mean of M value, two-step non-linear regression and others, each with its own advantages and disadvantages (McCormick et al., 2011). One of the most common normalization methods is to compute the RPKM (reads per kilobase per million) of each unique reads (Motameny et al., 2010). Some of the mentioned methods can be applied using the Bioconductor's package easyRNASeq (Delhomme et al., 2012). This process must not include the sequencing errors that passed the initial filters and it is also recommended to remove reads with low counts (Motameny et al., 2010).

After normalization, the appropriate statistical method can be applied to find differentially expressed ncRNAs. Microarray is a method widely used for large-scale quantification of gene expression. However, raw data from microarray and HTS differ because the former provides continuous values and the latter discrete values for measuring gene expression. Hence, well-established statistical methods used for the detection of differentially expressed genes in microarray data cannot be applied for HTS studies. Some examples of packages and softwares for HTS analysis are the Bioconductor's packages DESeq (Anders and Huber, 2010), EdgeR (Robinson et al., 2010), based on the negative binomial distribution, and baySeq (Hardcastle and Kelly, 2010), which uses a statistical Bayesian approach. Some authors also prefer to use variations of the Poison's distribution like the Two-Stage Poison Model (Auer and Doerge, 2011). Recently, some articles were published comparing the performance of some of the aforementioned differential expression Bioconductor packages and other softwares based on simulated and real data (Kvam et al., 2012; Robles et al., 2012; Vijay et al., 2012). **Table 5** presents some Bioconductor's packages for normalization or differential expression analysis of HTS data.

It is interesting to further validate any novel transcripts discovered. Computational and experimental techniques for gene finding are difficult to be applied to ncRNAs, due to their specific function and the fact that they do not have the same characteristics as the well known protein coding genes (Mendes et al., 2009). Concerning ncRNAs, most of the gene finding tools is directed to miRNA genes (revised by Oulas et al., 2011). A tool constructed specially to validate novel miRNAs found by HTS experiments is mirDeep (Friedländer et al., 2008; **Table 6**). This tool searches for reads that form the precursor miRNA and uses the folding algorithm of the Vienna package to evaluate the possibility of a hairpin structure (Friedländer et al., 2008). As mentioned, the structure of ncRNA families is well conserved and is usually used to assist as an additional step toward confirming a new or a known ncRNA.

There are several folding algorithms to predict RNA secondary structure (**Table 7**). Among the most well known are the ViennaRNA package (Lorenz et al., 2011), Mfold (Zuker, 2003) and Rfold (Kiryu et al., 2008). The ViennaRNA package uses thermodynamic parameters and dynamic programming to predict the secondary structure. It also provides information about centroid and maximum expected accuracy structures derived from base paring probabilities (Lorenz et al., 2011). The web version contains the most used tools and can be applied to obtain a putative secondary structure of a specific sequence or the consensus structure of a group of sequences (Hofacker, 2003). The Mfold algorithm uses free energy data to predict the minimum free energy for different foldings based on several user defined parameters. The output of Mfold includes structure plots, single strand frequency plots, and energy plots (Zuker, 2003). Another tool to predict secondary

Table 4 | Pipelines for HTS analysis.

Name	Site	Description	Authors
miRExpress	http://mirexpress.mbc.nctu.edu.tw/	miRNA profiling	Wang et al. (2009)
RandA	http://ibis.tau.ac.il/RandA/	ncRNA profiling and differential expression	Isakov et al. (2012)
mirAnalyzer	http://bioinfo2.ugr.es/miRanalyzer/miRanalyzer.php	miRNA profiling and gene discovery	Hackenberg et al. (2009)
miRNAkey	http://ibis.tau.ac.il/miRNAkey/	miRNA profiling and differential expression	Ronen et al. (2010)

Table 5 | Bioconductor's packages for normalization and differential expression of HTS data.

Name	Site	Description	Authors
easyRNASeq	http://bioconductor.org/packages/2.10/bioc/html/easyRNASeq.html	Count summarization and normalization for RNA-seq data	Delhomme et al. (2012)
DESeq	http://bioconductor.org/packages/2.10/bioc/html/DESeq.html	Differential gene expression analysis based on the negative binomial distribution	Anders and Huber (2010)
edgeR	http://bioconductor.org/packages/2.10/bioc/html/edgeR.html	Empirical analysis of digital gene expression data in R	Robinson et al. (2010)
baySeq	http://www.bioconductor.org/packages/release/bioc/html/baySeq.html	Normalization and differential gene expression by Bayesian methods	Hardcastle and Kelly (2010)

Table 6 | miRNA gene discovery for HTS.

Name	Site	Authors
miRDeep	http://www.mdc-berlin.de/en/research/research_teams/systems_biology_of_gene_regulatory_elements/projects/miRDeep/index.html	Friedländer et al. (2008)

Table 7 | Secondary structure prediction tools.

Name	Site	Authors
Mfold	http://www.bioinfo.rpi.edu/applications/mfold	Zuker (2003)
ViennaRNA package	http://www.tbi.univie.ac.at/∼ivo/RNA/	Lorenz et al. (2011)
Rfold	http://www.ncrna.org/software/Rfold/	Kiryu et al. (2008)

structure of RNAs is the Rfold algorithm which performs base paring probabilities (Kiryu et al., 2008).

Other additional step in the interpretation of HTS ncRNA experiments includes finding the protein coding genes targeted by the detected ncRNAs. Even for the most studied ncRNA class, miRNAs, this is a complex task, due to their small size and few base pairing to their targets. The currently available tools rely on known properties like paring pattern, thermodynamic stability, and conservation to predict putative targets (Min and Yoon, 2010). There are several databases and software for miRNA target recognition (**Table 8**). Among them, may be cited Miranda (John et al., 2004), Pictar (Krek et al., 2005), and Diana-microT (Maragkakis et al., 2009; for a complete view of such databases, see, Yousef et al., 2009). The Miranda algorithm was used to predict miRNA targets presented in the microRNAs.org database (Betel et al., 2008). This algorithm uses the binding energy, complementary pattern, evolutionary conservation, and position of the binding site in the mRNA. Also, is the unique program which is available for download (John et al., 2004). The Pictar algorithm uses the type of paring between miRNA and mRNA, the free energy of the paring and target site conservation to generate a probability and a score of the putative target site (Krek et al., 2005). The DIANA-microT algorithm uses the type of paring and the conservation to calculate a score for each predicted binding site. This score is compared to the score obtained by using random miRNAs to calculate a signal-to-noise ratio (Maragkakis et al., 2009).

The visualization of the reads aligned to the reference genome is another important set of tools for projects working with HTS. Data visualization permits to researchers to investigate HTS experiments in a user friendly way (Zhou et al., 2011). Several tools were developed for visualization of HTS experiments, some of them were listed by Lee et al. (2011), among them are Integrated Genomics Viewer (IGV; Thorvaldsdottir et al., 2012), Artemis (Carver et al., 2012), and Tablet (Milne et al., 2010). Also, the UCSC and Ensembl genome browsers have been updated to support HTS data. The downside of using a web viewer is uploading large amount of data (Fiume et al., 2010). **Table 9** shows some bioinformatics tools for visualization of HTS experiments.

CHALLENGES IN BIOINFORMATICS OF ncRNA AND HTS

The management of the data produced by HTS methods is the first challenge in bioinformatics. Many gigabytes of raw data may be

Table 8 | miRNA target prediction tools and databases.

Name	Site	Description	Authors
TargetScan	http://www.targetscan.org/	miRNA target prediction algorithm	Lewis et al. (2003)
DIANA-microT	http://diana.cslab.ece.ntua.gr/microT/	miRNA target prediction algorithm	Maragkakis et al. (2009)
RNA Hybrid	http://bibiserv.techfak.uni-bielefeld.de/rnahybrid/	Tool for finding the minimum free energy hybridization of a long and a short RNA	Rehmsmeier et al. (2004)
miRDB	http://mirdb.org/miRDB/	Database for miRNA target prediction by MirTarget2 and functional annotation	Wang and El Naqa (2008)
microRNA.org	http://www.microrna.org/microrna/home.do	Database of miRNA target prediction by the miRanda algorithm	Betel et al. (2008)
TarBase	http://diana.cslab.ece.ntua.gr/tarbase/	Manually curated database of experimentally supported microRNA targets	Papadopoulos et al. (2009)
miR2Disease	http://www.mir2disease.org/	Manually curated database of miRNA deregulation in various human diseases	Jiang et al. (2009)
miRecords	http://mirecords.biolead.org/index.php	Database of experimentally validated miRNA targets and integration of predicted miRNA targets produced by 11 miRNA target prediction programs	Xiao et al. (2009)

Table 9 | Tools for visualizations of HTS experiments.

Name	Site	Authors
BamView	http://bamview.sourceforge.net/	Carver et al. (2012)
IGV	http://www.broadinstitute.org/igv/	Thorvaldsdottir et al. (2012)
Artemis	http://www.sanger.ac.uk/resources/software/artemis/	Carver et al. (2012)
Savant	http://genomesavant.com/savant/	Fiume et al. (2010)
Tablet	http://bioinf.scri.ac.uk/tablet/	Milne et al. (2010)

produced during a regular project aiming to detect the expression profile of ncRNAs in oncology and this amount may increase if it is considered data of mapped reads and all annotation databases used to analyze them. Furthermore, the hardware and network speed may be taken into account for appropriate analysis prior starting a HTS project. Other important challenge in Bioinformatics is to create protocols to assist in the analysis of ncRNA data. There are some efforts to assist protein coding genes in HTS data, but none was taken to ncRNA genes (Trapnell et al., 2012). Almost every article analyzing ncRNA expression profile using HTS methods present distinct normalization and statistical approaches. Finally, since Bioinformatics is still an emerging field of knowledge, there is few groups with graduate students developing innovative projects in bioinformatics and ncRNAs. In conclusion, there are three major limitations in bioinformatics of HTS projects: data management, analysis, and visualization; definition of protocols to data analysis; and professionals with expertise in ncRNA analysis.

ACKNOWLEDGMENTS

The authors acknowledge Nicole Scherer and Gabriel Lira Espindola Mendes for critical reading. Natasha Andressa Nogueira Jorge is supported by Vice-Presidência de Ensino, Informação e Comunicação/Pró-Reitoria – IOC/FIOCRUZ and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). Fabio Passetti is supported by CNPq (#312733/2009-7). Fabio Passetti and Carlos Gil Ferreira acknowledge the support of Fundação do Câncer.

REFERENCES

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* 11, 1–12.

Auer, P. L., and Doerge, R. W. (2011). A two-stage Poison model for testing RNA-seq data. *Stat. Appl. Genet. Mol. Biol.* 10, 1–26.

Betel, D., Wilson, M., Gabow, A., Marks, D. S., and Sander, C. (2008). The microRNA.org resource: targets and expression. *Nucleic Acids Res.* 36, D149–D153.

Brait, M., and Sidransky, D. (2011). Cancer epigenetics: above and beyond. *Toxicol. Mech. Methods* 21, 275–288.

Calin, G. A., Dumitru, C. D., Shimizu, M., Bichi, R., Zupo, S., Noch, E., et al. (2002). Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proc. Natl. Acad. Sci. U.S.A.* 99, 15524–15529.

Calin, G. A., Sevignani, C., Dumitru, C. D., Hyslop, T., Noch, E., Yendamuri, S., et al. (2004). Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. *Proc. Natl. Acad. Sci. U.S.A.* 101, 2999–3004.

Carver, T., Harris, S. R., Berriman, M., Parkhill, J., and McQuillan, J. A. (2012). Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* 28, 464–469.

Chen, X., Hu, Z., Wang, W., Ba, Y., Ma, L., Zhang, C., et al. (2012). Identification of ten serum microRNAs from a genome-wide serum microRNA expression profile as novel noninvasive biomarkers for nonsmall cell lung cancer diagnosis. *Int. J. Cancer* 130, 1620–1628.

Cheng, J., Guo, J. M., Xiao, B. X., Miao, Y., Jiang, Z., Zhou, H., et al. (2011). piRNA, the new non-coding RNA, is aberrantly expressed in human cancer cells. *Clin. Chim. Acta* 412, 1621–1625.

Cui, L., Lou, Y., Zhang, X., Zhou, H., Deng, H., Song, H., et al. (2011). Detection of circulating tumor cells in peripheral blood from patients with gastric cancer using piRNAs as markers. *Clin. Biochem.* 44, 1050–1057.

Datta, S., Datta, S., Kim, S., Chakraborty, S., and Gill, R. (2010). Statistical analyses of next generation sequence data: a partial overview. *J. Proteomics Bioinform.* 3, 183–190.

Delhomme, N., Padioleau, I., Furlong, E. E., and Steinmetz, L. M. (2012). easyRNASeq: a bioconductor package for processing RNA-Seq data. *Bioinformatics* 28, 2532–2533.

Dillhoff, M., Liu, J., Frankel, W., Croce, C., and Bloomston, M. (2008). MicroRNA-21 is overexpressed in pancreatic cancer and a potential predictor of survival. *J. Gastrointest. Surg.* 12, 2171–2176.

Dreszer, T. R., Karolchik, D., Zweig, A. S., Hinrichs, A. S., Raney, B. J., Kuhn, R. M., et al. (2012). The UCSC Genome Browser database: extensions and updates 2011. *Nucleic Acids Res.* 40, D918–D923.

Eddy, S. R. (2001). Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet.* 2, 919–929.

Elgar, G., and Vavouri, T. (2008). Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. *Trends Genet.* 24, 344–352.

ENCODE Project Consortium, Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R., Gingeras, T. R., et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447, 799–816.

Esposito, T., Magliocca, S., Formicola, D., and Gianfrancesco, F. (2011). piR_015520 belongs to Piwi-associated RNAs regulates expression of the human melatonin receptor 1A gene. *PLoS ONE* 6:e22727. doi:10.1371/journal.pone.0022727

Fiume, M., Williams, V., Brook, A., and Brudno, M. (2010). Savant: genome browser for high-throughput sequencing data. *Bioinformatics* 26, 1938–1944.

Frankel, L. B., Christoffersen, N. R., Jacobsen, A., Lindow, M., Krogh, A., and Lund, A. H. (2008). Programmed cell death 4 (PDCD4) is an important functional target of the microRNA miR-21 in breast cancer cells. *J. Biol. Chem.* 283, 1026–1033.

Friedländer, M. R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S., et al. (2008). Discovering microRNAs from deep sequencing data using miRDeep. *Nat. Biotechnol.* 26, 407–415.

Gardner, P. P., Daub, J., Tate, J., Moore, B. L., Osuch, I. H., Griffiths-Jones, S., et al. (2011). Rfam: wikipedia, clans and the "decimal" release. *Nucleic Acids Res.* 39, D141–D145.

Guffanti, A., Iacono, M., Pelucchi, P., Kim, N., Soldà, G., Croft, L. J., et al. (2009). A transcriptional sketch of a primary human breast cancer by 454 deep sequencing. *BMC Genomics* 10: 163. doi:10.1186/1471-2164-10-163

Gustschner, T., and Diederichs, S. (2012). The hallmarks of cancer: a long non-coding RNA point of view. *RNA Biol.* 9, 703–719.

Hackenberg, M., Sturm, M., Langenberger, D., Falcón-Pérez, J. M., and Aransay, A. M. (2009). miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments. *Nucleic Acids Res.* 37, W68–W76.

Hagen, J. H. (2000). The origins of bioinformatics. *Nat. Rev. Genet.* 1, 231–236.

Hardcastle, T. J., and Kelly, K. A. (2010). baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics* 11:422. doi:10.1186/1471-2105-11-422

Hennessey, P. T., Sanford, T., Choudhary, A., Mydlarz, W. W., Brown, D., Adai, A. T., et al. (2012). Serum microRNA biomarkers for detection of non-small cell lung cancer. *PLoS ONE* 7:e32307. doi:10.1371/journal.pone.0032307

Hercus, C. (2008). *Novoalign: A Short Read Aligner with Qualities.* Available at: www.novocraft.com

Hesper, B., and Hogeweg, P. (1970). Bioinformatica: een werkconcept. *Kameleou* 1, 28–29.

Hofacker, I. L. (2003). Vienna RNA secondary structure server. *Nucleic Acids Res.* 31, 3429–3431.

Hogeweg, P. (2011). The roots of bioinformatics in theoretical biology. *PLoS Comput. Biol.* 7:e1002021. doi:10.1371/journal.pcbi.1002021

Hu, Z., Chen, X., Zhao, Y., Tian, T., Jin, G., Shu, Y., et al. (2010). Serum microRNA signatures identified in a genome-wide serum MicroRNA expression profiling predict survival of non–small-cell lung Cancer. *J. Clin. Oncol.* 28, 1721–1726.

Iliopoulos, D., Jaeger, S. A., Hirsch, H. A., Bulyk, M. L., and Struhl, K. (2010). STAT3 activation of miR-21 and miR-181b-1 via PTEN and CYLD are part of the epigenetic switch linking inflammation to cancer. *Mol. Cell* 39, 493–506.

Isakov, O., Ronen, R., Kovarsky, J., Gabay, A., Gan, I., Modai, S., et al. (2012). Novel insight into the non-coding repertoire through deep sequencing analysis. *Nucleic Acids Res.* 40, 1–6.

Isakov, O., and Shomron, N. (2011). "Deep sequencing data analysis: challenges and solutions," in *Bioinformatics – Trends and Methodologies*, ed. M. A. Mahdavi (Rijeka: InTech Press), 655–676.

Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., et al. (2009). miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res.* 37, D98–D104.

Jima, D. D., Zhang, J., Jacobs, C., Richards, K. L., Dunphy, C. H., Choi, W. W., et al. (2010). Deep sequencing of the small RNA transcriptome of normal and malignant human B cells identifies hundreds of novel microRNAs. *Blood* 116, e118–e127.

John, B., Enright, A. J., Aravin, A., Tuschl, T., Sander, C., and Marks, D. S. (2004). Human MicroRNA targets. *PLoS Biol.* 2:e363. doi:10.1371/journal.pbio.0020363

Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., et al. (2004). The UCSC table browser data retrieval tool. *Nucleic Acids Res.* 32, D493–D496.

Keller, A., Backes, C., Leidinger, P., Kefer, N., Boisguerin, V., Barbacioru, C., et al. (2011). Next-generation sequencing identifies novel microRNAs in peripheral blood of lung cancer patients. *Mol. Biosyst.* 7, 3187–3199.

Kent, W. J., and Haussler, D. (2001). Assembly of the working draft of the human genome with GigAssembler. *Genome Res.* 11, 1541–1548.

Kircher, M., and Kelso, J. (2010). High-throughput DNA sequencing – concepts and limitations. *Bioassays* 32, 524–536.

Kiryu, H., Kin, T., and Asai, K. (2008). Rfold: an exact algorithm for computing local base pairing probabilities. *Bioinformatics* 24, 367–373.

Kozomara, A., and Griffiths-Jones, S. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.* 39, D152–D157.

Krek, A., Grün, D., Poy, M. N., Wolf, R., Rosenberg, L., Epstein, E. J., et al. (2005). Combinatorial microRNA target predictions. *Nat. Genet.* 37, 495–500.

Krichevsky, A. M., and Gabriely, G. (2009). miR-21: a small multi-faceted RNA. *J. Cell. Mol. Med.* 13, 39–53.

Kvam, V. M., Liu, P., and Si, Y. (2012). A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. *Am. J. Bot.* 99, 248–256.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10, 1–10.

Lee, H. C., Lai, K., Lorenc, M. T., Imelfort, M., Duran, C., and Edwards, D. (2011). Bioinformatics tools and databases for analysis of next-generation sequence data. *Brief. Funct. Genomics* 11, 12–24.

Lewis, B. P., Shih, I. H., Jones-Rhoades, M. W., Bartel, D. P., and Burge, C. B. (2003). Prediction of mammalian microRNA targets. *Cell* 115, 787–798.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760.

Li, J., Huang, H., Sun, L., Yang, M., Pan, C., Chen, W., et al. (2009a). MiR-21 indicates poor prognosis in tongue squamous cell carcinomas as an apoptosis inhibitor. *Clin. Cancer Res.* 15, 3998–4008.

Li, T., Li, D., Sha, J., Sun, P., and Huang, Y. (2009b). MicroRNA-21 directly targets MARCKS and promotes apoptosis resistance and invasion in prostate cancer cells. *Biochem. Biophys. Res. Commun.* 383, 280–285.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009c). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079.

Li, R., Li, Y., Kristiansen, K., and Wang, J. (2008). SOAP: short oligonucleotide alignment program. *Bioinformatics* 24, 713–714.

Li, Y., Zhang, Z., Liu, F., Vongsangnak, W., Jing, Q., and Shen, B. (2012). Performance comparison and evaluation of software tools for microRNA deep-sequencing data analysis. *Nucleic Acids Res.* 40, 4298–4305.

Liu, H., D'Andrade, P., Fulmer-Smentek, S., Lorenzi, P., Kohn, K. W., Weinstein, J. N., et al. (2010). mRNA and microRNA expression profiles of the NCI-60 integrated with drug activities. *Mol. Cancer Ther.* 9, 1080–1091.

Liu, H., Yin, J., Wang, S., Zen, K., Ba, K., and Zhang, C. Y. (2011). A five-microRNA signature identified

from genome wide serum microRNAs expression profiling serves as a fingerprint for gastric cancer diagnosis. *Eur. J. Cancer* 47, 784–791.

Lorenz, R., Bernhart, S. H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., et al. (2011). ViennaRNA Package 2.0. *Algorithms Mol. Biol.* 6, 1–14.

Lussier, Y. A., Stadler, W. M., and Chen, J. L. (2012). Advantages of genomic complexity: bioinformatics opportunities in microRNA cancer signatures. *J. Am. Med. Inform. Assoc.* 19, 156–160.

Maragkakis, M., Reczko, M., Simossis, V. A., Alexiou, P., Papadopoulos, G. L., Dalamagas, T., et al. (2009). DIANA-microT web server: elucidating microRNA functions through target prediction. *Nucleic Acids Res.* 37, W273–W276.

Martens-Uzunova, E. S., Jalava, S. E., Dits, N. F., van Leenders, G. J. L. H., Møller, S., Trapman, J., et al. (2012). Diagnostic and prognostic signatures from the small non-coding RNA transcriptome in prostate cancer. *Oncogene* 31, 978–991.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* 17, 10–12.

McCormick, K. P., Willmann, M. R., and Meyers, B. C. (2011). Experimental design, preprocessing, normalization and differential expression analysis of small RNA sequencing experiments. *Silence* 2, 1–19.

Meiri, E., Levy, A., Benjamin, H., Ben-David, M., Cohen, L., Dov, A., et al. (2010). Discovery of microRNAs and other small RNAs in solid tumors. *Nucleic Acids Res.* 38, 6234–6246.

Mendes, N. D., Freitas, A. T., and Sagot, M. F. (2009). Current tools for the identification of miRNA genes and their targets. *Nucleic Acids Res.* 37, 2419–2433.

Meyerson, M., Gabriel, S., and Getz, G. (2010). Advances in understanding cancer genomes through second-generation sequencing. *Nat. Rev. Genet.* 11, 685–696.

Milne, I., Bayer, M., Cardle, L., Shaw, P., Stephen, G., Wright, F., et al. (2010). Tablet – next generation sequence assembly visualization. *Bioinformatics* 26, 401–402.

Min, H., and Yoon, S. (2010). Got target? Computational methods for microRNA target prediction and their extension. *Exp. Mol. Med.* 42, 233–244.

Mishra, P. J., Humeniuk, R., Mishra, P. J., Longo-Sorbello, G. S., Banerjee, D., and Bertino, J. R. (2007). A miR-24 microRNA binding-site polymorphism in dihydrofolate reductase gene leads to methotrexate resistance. *Proc. Natl. Acad. Sci. U.S.A.* 104, 13513–13518.

Mitra, S. A., Mitra, A. P., and Triche, T. J. (2012). A central role for long non-coding RNA in cancer. *Front. Genet.* 3:17. doi:10.3389/fgene.2012.00017

Mituyama, T., Yamada, K., Hattori, E., Okida, H., Ono, Y., Terai, G., et al. (2009). The Functional RNA Database 3.0: databases to support mining and annotation of functional RNAs. *Nucleic Acids Res.* 37, D89–D92.

Morgan, M., Anders, S., Lawrence, M., Aboyoun, P., Pàges, H., and Gentleman, R. (2009). ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics* 25, 2607–2608.

Motameny, S., Wolfers, S., Nürberg, P., and Schumacher, B. (2010). Next generation sequencing of miRNAs – strategies, resources and methods. *Genes* 1, 70–84.

Oulas, A., Karathanasis, N., Louloupi, A., and Poirazi, P. (2011). Finding cancer-associated miRNAs: methods and tools. *Mol. Biotechnol.* 49, 97–107.

Pàges, H., Aboyoun, P., Gentleman, R., and DebRoy, S. (2012). *Biostrings: String Objects Representing Biological Sequences, and Matching Algorithms.* R Package Version 2.25.4, Seattle.

Papadopoulos, G. L., Reczko, M., Simossis, V. A., Sethupathy, P., and Hatzigeorgiou, A. G. (2009). The database of experimentally supported targets: a functional update of TarBase. *Nucleic Acids Res.* 37, D155–D158.

Pasmant, E., Sabbagh, A., Masliah-Planchon, J., Ortonne, N., Laurendeau, I., Melin, L., et al. (2011). Role of noncoding RNA ANRIL in genesis of plexiform neurofibromas in neurofibromatosis type 1. *J. Natl. Cancer Inst.* 103, 1713–1722.

Patel, R. K., and Jain, M. (2012). NGS QC toolkit: a toolkit for quality control of next generation sequencing data. *PLoS ONE* 7:e30619. doi:10.1371/journal.pone.0030619

Prensner, J. R., Iyer, M. K., Balbin, O. A., Dhanasekaran, S. M., Cao, Q., Brenner, J. C., et al. (2011). Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nat. Biotechnol.* 29, 742–749.

Rabinowits, G., Gerçel-Taylor, C., Day, J. M., Taylor, D. D., and Kloecker, G. H. (2009). Exosomal microRNA: a diagnostic marker for lung cancer. *Clin. Lung Cancer* 10, 42–46.

Rehmsmeier, M., Steffen, P., Hochsmann, M., and Giegerich, R. (2004). Fast and effective prediction of microRNA/target duplexes. *RNA* 10, 1507–1517.

Reinhold, W. C., Sunshine, M., Liu, H., Varma, S., Kohn, K. W., Morris, J., et al. (2012). 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.

Reis, E. M., Nakaya, H. I., Louro, R., Canavez, F. C., Flatschart, A. V., Almeida, G. T., et al. (2004). Antisense intronic non-coding RNA levels correlate to the degree of tumor differentiation in prostate cancer. *Oncogene* 23, 6684–6692.

Renaud, G., Neves, P., Folador, E. L., Ferreira, C. G., and Passetti, F. (2011). Segtor: rapid annotation of genomic coordinates and single nucleotide variations using segment trees. *PLoS ONE* 6:e26715. doi:10.1371/journal.pone.0026715

Ribas, J., Ni, X., Haffner, M., Wentzel, E. A., Salmasi, A. H., Chowdhury, W. H., et al. (2009). miR-21: an androgen receptor-regulated microRNA that promotes hormone-dependent and hormone-independent prostate cancer growth. *Cancer Res.* 69, 7165–7169.

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

Robles, J. A., Quereshi, S. E., Stephen, S. T., Wilson, S. R., Burden, C. J., and Taylor, J. M. (2012). Efficient experimental design and analysis strategies for the detection of differential expression using RNA-sequencing. *BMC Genomics* 13:484. doi:10.1186/1471-2164-13-484

Ronen, R., Gan, I., Modai, S., Sukacheov, A., Dror, G., Halperin, E., et al. (2010). miRNAkey: a software for microRNA deep sequencing analysis. *Bioinformatics* 26, 2615–2616.

Ruffalo, M., LaFramboise, T., and Koyutürk, M. (2011). Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics* 27, 2790–2796.

Ryu, S., Joshi, N., McDonnell, K., Woo, J., Choi, H., Gao, D., et al. (2011). Discovery of novel human breast cancer microRNAs from deep sequencing data by analysis of pri-microRNA secondary structures. *PLoS ONE* 6:e16403. doi:10.1371/journal.pone.0016403

Schulte, J. H., Marschall, T., Martin, M., Rosenstiel, P., Mestdagh, P., Schlierf, S., et al. (2010). Deep sequencing reveals differential expression of microRNAs in favorable versus unfavorable neuroblastoma. *Nucleic Acids Res.* 38, 5919–5928.

Seike, M., Goto, A., Okano, T., Bowman, E. D., Schetter, A. J., Horikawa, I., et al. (2009). MiR-21 is an EGFR-regulated anti-apoptotic factor in lung cancer in never-smokers. *Proc. Natl. Acad. Sci. U.S.A.* 106, 12085–12090.

Tahira, A. C., Kubrusly, M. S., Faria, M. F., Dazzani, B., Fonseca, R. S., Maracaja-Coutinho, V., et al. (2011). Long noncoding intronic RNAs are differentially expressed in primary and metastatic pancreatic cancer. *Mol. Cancer* 10, 1–19.

Tainsky, M. A. (2009). Genomic and proteomic biomarkers for cancer: a multitude of opportunities. *Biochim. Biophys. Acta*, 1796, 176–193.

Thorvaldsdottir, H., Robinson, J. T., and Mesirov, J. P. (2012). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinformatics.* PMID:22517427. [Epub ahead of print].

Toedling, J., Servant, N., Ciaudo, C., Farinelli, L., Voinnet, O., Heard, E., et al. (2012). Deep-sequencing protocols influence the results obtained in small-RNA sequencing. *PLoS ONE* 7:e32724. doi:10.1371/journal.pone.0032724

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7, 562–578.

Vijay, N., Poelstra, J. W., Künstner, A., and Wolf, J. B. W. (2012). Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive *in silico* assessment of RNA-seq experiments. *Mol. Ecol.* PMID:22998089. [Epub ahead of print].

Wang, W. C., Lin, F. M., Chang, W. C., Lin, K. Y., Huang, H. D., and Lin, N. S. (2009). miRExpress: analyzing high-throughput sequencing data for profiling microRNA expression. *BMC Bioinformatics* 10:328. doi:10.1186/1471-2105-10-328

Wang, X., and El Naqa, I. M. (2008). Prediction of both conserved and nonconserved microRNA targets in animals. *Bioinformatics* 24, 325–332.

Weng, L., Wu, X., Gao, H., Mu, B., Li, X., Wang, J. H., et al. (2010). MicroRNA profiling of clear cell renal cell carcinoma by whole-genome small RNA deep sequencing of paired frozen and formalin-fixed, paraffin-embedded tissue specimens. *J. Pathol.* 222, 41–51.

Wickramasinghe, N. S., Manavalan, T. T., Dougherty, S. M., Riggs, K. A., Li, Y., and Klinge, C. M. (2009). Estradiol downregulates miR-21 expression and increases miR-21 target gene expression in MCF-7 breast cancer cells. *Nucleic Acids Res.* 37, 2584–2595.

Wu, X., Somlo, G., Yu, Y., Palomares, M. R., Li, A. X., Zhou, W., et al. (2012). De novo sequencing of circulating miRNAs identifies novel markers predicting clinical outcome of locally advanced breast cancer. *J. Transl. Med.* 8, 1–10.

Xiao, F., Zuo, Z., Cai, G., Kang, S., Gao, X., and Li1, T. (2009). miRecords: an integrated resource for microRNA–target interactions. *Nucleic Acids Res.* 37, D105–D110.

Yang, F., Bi, J., Xue, X., Zheng, L., Zhi, K., Hua, J., et al. (2012a). Upregulated long non-coding RNA H19 contributed to proliferation of gastric cancer cell. *FEBS J.* 279, 3159–3165.

Yang, F., Yi, F., Zheng, Z., Ling, Z., Ding, J., Guo, J., et al. (2012b). Characterization of a carcinogenesis-associated long non-coding RNA. *RNA Biol.* 9, 110–116.

Yousef, M., Showe, L., and Showe, M. (2009). A study of microRNAs in silico and in vivo: bioinformatics approaches to microRNA discovery and target identification. *FEBS J.* 276, 2150–2156.

Zhou, L., Li, X., Liu, Q., Zhao, F., and Wu, J. (2011). Small RNA transcriptome investigation based on next-generation sequencing technology. *J. Genet. Genomics* 38, 505–513.

Zhu, S., Wu, H., Wu, F., Nie, D., Sheng, S., and Mo, Y.-Y. (2008). MicroRNA-21 targets tumor suppressor genes in invasion and metastasis. *Cell Res.* 18, 350–359.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* 31, 3406–3415.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 04 September 2012; accepted: 22 November 2012; published online: 17 December 2012.

Citation: Jorge NAN, Ferreira CG and Passetti F (2012) Bioinformatics of cancer ncRNA in high throughput sequencing: present state and challenges. Front. Gene. ***3****:287. doi: 10.3389/fgene.2012.00287*

This article was submitted to Frontiers in Non-Coding RNA, a specialty of Frontiers in Genetics.