SURFR: A Real-Time Platform for Non-Coding RNA Fragmentation Analysis Using Wavelets

Mohan Vamsi Kasukurthi
School of Computing
University of South Alabama
Mobile, AL, U.S.A.
mk1530@jagmail.southalabama.edu

Dominika Houserova
Department of Pharmacology
University of South Alabama
Mobile, AL, U.S.A
dh1001@jagmail.southalabama.edu

Yulong Huang
College of Allied Health Professions
University of South Alabama
Mobile, AL, U.S.A.
yh1623@jagmail.southalabama.edu

Shengyu Li
School of Computing
University of South Alabama
Mobile, AL, U.S.A.
sl1721@jagmail.southalabama.edu

Dongqi Li
School of Computing
University of South Alabama
Mobile, AL, U.S.A.
kevinldq2018@gmail.com

Jingwei Lin
Ocean School
Fuzhou University
Fuzhou, China
549841688@qq.com

Guanhuan Yang
School of Computing
University of South Alabama
Mobile, AL, U.S.A.
euphoriayod@gmail.com

Shaobo Tan
School of Computing
University of South Alabama
Mobile, AL, U.S.A.
tan.shaobo.soc@gmail.com

David Bourrie
School of Computing
University of South Alabama
Mobile, AL, U.S.A.
dbourrie@southalabama.edu

Bin Ma
Qilu University of Technology
(Shandong Academy of Science)
Jinan, Shandong, China
mab@qlu.edu.cn

Glen M. Borchert *
Department of Pharmacology
University of South Alabama
Mobile, AL, U.S.A.
borchert@southalabama.edu

Jingshan Huang *
School of Computing & College of Medicine
University of South Alabama
Mobile, AL, U.S.A.
huang@southalabama.edu

2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 978-1-6654-0126-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/BIBM52615.2021.9669696

***Abstract*— It is well known that microRNAs (miRNAs or miRs) are small (~18-25 nt) yet highly potent non-coding RNA-derived RNAs (ndRNAs), originating from pre-miRNA fragmentation, that have been shown to alter the post-transcriptional functionality of many messenger RNAs (mRNAs). Biologically, the identification and study of miRNAs is very critical due to their increasing significance as biomarkers for many types of cancers and other genetic diseases. While empirical evidence supporting the existence of several novel ndRNAs excised from other longer non coding RNAs (ncRNAs) is growing, recent evidence suggests the full extent of their prevalence is likely underappreciated. Although some computational methods have been designed to help domain experts identify and understand miRNAs by analyzing Next Generation Sequencing (NGS) datasets, there are some crucial challenges, such as efficiency, effectiveness, and generalizability, in the state-of-the-art in-silico methods. To address such problems, our group proposed a new algorithm to mine ndRNAs by applying wavelet-based signal processing techniques as opposed to the current string-based NGS sequence alignment/analysis. However, due to novelty of the approach, our initial version of the algorithm was focused specifically on mining miRNAs, snoRNA-derived RNAs (sdRNAs) and transfer RNA (tRNA) fragments (tRFs) because of their importance in the literature plus the availability of experimentally validated databases to confirm our findings. Despite the computational issues, we still lack a basic understanding of the existence and the range of ndRNA functionalities from a) ndRNAs other than miRs, sdRNAs & tRFs in humans, and b) all ndRNAs in millions of organisms other than humans. Hence, there is an urgent requirement to automate the extraction and experimentation of ndRNAs, especially considering the rate at which NGS data is being produced. Therefore, in the current article, we extended our algorithm to be applicable to ~500 organisms—including eukaryotes, plants, bacteria, fungi, and protists—along with all their ncRNAs available in the current NCBI annotation. We also constructed a real-time user-friendly platform, SURFR, available at salts.soc.southalabama.edu/surfr, to aid domain experts and the aspiring biomedical scientists to perform RNA-Seq experiments to study ndRNAs. Not only our platform is extremely efficient, but we are also capable of allowing the users to identify, analyze, visualize, and compare ndRNAs from up to 30 NGS files to perform rigorous experimentation. Moreover, access to NGS files from public databases like SRA, and ndRNAs from private databases like TCGA are made readily available to the users to further validate their novel findings. Finally, we provide theoretical validation to examine our platform's effectiveness.**

Keywords— microRNAs, small ncRNAs, ncRNA-derived RNAs, Differential Expression Analysis, Gene Expression Visualization, Wavelet

I. INTRODUCTION

Ribonucleic Acids (RNAs) can broadly be classified into two categories: coding, and non-coding RNAs (ncRNAs). The coding RNAs, also commonly known as messenger RNAs (mRNAs), carry and execute the information required for the formation of proteins—hence the name coding/protein-coding RNAs. On the other hand, the RNAs that do not encode the information for protein synthesis are known as ncRNAs.

* Corresponding Authors: borchert@southalabama.edu , huang@southalabama.edu

However, although ncRNAs do not directly code for proteins, they do possess many functionalities within the cell [1]. As a matter of fact, ncRNAs are critical to every cellular function, and how their misexpressions contribute to a wide range of diseases is only now beginning to be appreciated [1-3]. That said, it's now clear that >98% of eukaryotic RNAs are actually non-coding [4] making the study of ncRNAs even more critical. Notably, the functions of thousands of recently discovered ncRNAs remain largely unclear; what's more, recent evidence suggests less than half of functional cellular ncRNAs have yet to be identified [4].

Non coding RNAs (ncRNAs) can be subdivided into two main groups: 1. small ncRNAs (sncRNAs) – shorter than 200 nucleotides (nts) in length, and 2. long ncRNAs (lncRNAs) – longer than 200 nucleotides (nts). MicroRNAs (miRNAs) are a large class of sncRNAs that were discovered only ~2 decades ago [5]. Although overlooked for decades due to their size, miRNAs - only about 18-25 nts long, have turned out to be the best studied group of sncRNAs and are currently the focus of many studies involving billions of research dollars. MiRNAs associate with the RNA-induced silencing complex (RISC) to directly target mRNAs and regulate their expressions in very specific manners [6][7]. Each miRNA is capable of targeting multiple mRNAs, and each mRNA can be targeted by more than one miRNAs, i.e., a many-to-many relation exists between miRNA and mRNA interactions (although the majority of these remain unknown) [8]. What's clear however, is that the expressions and functions of every unique miRNA is highly specific and contributes to a multitude of phenotypic and genotypic outcomes [9-12] (e.g.. cell proliferation, metabolism, drug resistance, tumorigenesis, apoptosis, etc. [4,7,13-15]). Importantly, the identification and mechanistic understanding of miRNAs and novel miRNA-like RNAs will undoubtedly lead to major advances in agriculture, green energy, and medicine in the near future making their characterization a hotly pursued area of research [16-18].

Importantly, small nucleolar RNAs (snoRNAs) and transfer RNAs (tRNAs)—two other widely studied types of ncRNAs—have recently been found to be processed into fully-functional miRNA-like fragments—known as snoRNA-derived RNAs (sdRNAs) and tRNA-Fragments (tRFs) respectively [19-22]. Similar to miRNAs, sdRNAs and tRFs are also exceptionally small in size, i.e., only around 18-35nt, making them very hard to identify and study biologically in the laboratory. That said, despite only being discovered less than a decade ago, numerous studies have now reported specific contributions of sdRNAs and tRFs in a wide range of activities (e.g., RNA silencing, translation regulation, epigenetic regulation, cell invasion, malignant transformation, metastatic progression, etc.) [19-22]. Notably, specific roles for miRNAs in the onset and progression of malignancy is currently one of the hottest areas of oncological and pharmacological research. That said, sdRNAs and tRFs have recently been shown to function almost indistinguishably from miRNAs and to possess clear involvements in various cancers. As such, we believe that they will similarly soon become a research priority to many biomedical and pharmacological studies.

Also of note, several studies have recently reported the existence of miRNA-like RNAs processed from other types of ncRNAs in addition to snoRNAs and tRNAs (e.g rRNAs, Y RNAs, vault RNAs, etc…) [23,24]. As such, the full repertoire of miRNA-like RNAs remains largely unclear in humans and almost entirely unexplored in most other organisms. That said, in a previous study [25], we coined the term, "ncRNA-derived RNAs (ndRNAs)" for any miRNA-like RNA processed from a longer ncRNA [26], and the identification of these ndRNAs constitutes the focus of the current study as we argue that, there is an urgent need to develop highly accurate and efficient new methods for the identification, exploration, and characterization of ndRNAs.

The development and recent advancements in Next Generation Sequencing (NGS) technologies now allow us to investigate the genetic realm computationally [27]. Sequencing refers to the set of technologies that are used to extract data from DNA/RNA samples. Having potential applications within crucial domains such as agriculture, medicine, drug development, microbiology, DNA forensics, health, phylogenetics, etc., NGS technologies have been widely adopted by hundreds to thousands of laboratories across the globe, thus producing daunting amounts of data every year [27][28]. Notably, the prices to perform NGS have reduced dramatically as of late and are still going down. For instance, the Human Genome Project by the National Institutes of Health (NIH) during the 90s, took about 15 years to sequence the human genome, and cost over 2.5 billion dollars [29,30]. With the massively parallelized second-generation sequencing techniques, sequencing a whole genome only takes about a few hours to a day and is priced just over 1000$. That said, the raw data obtained from NGS cannot be interpreted by humans directly as each file is typically several Giga Bytes in size containing 10-100 million lines of DNA/RNA reads making NGS data analysis—and thereby ndRNA analysis—a critical computational challenge. Some of the preeminent issues associated with NGS analyses are efficiency and parallelization.

Although, to date, several computational methods have been proposed to identify and analyze ndRNAs from NGS datasets [31-37], all previous methods carry serious limitations (e.g., human specific, only assessing limited subsets of ncRNAs, poor accuracy/efficiency, etc…). Therefore to address all the above mentioned problems, in our study [25], we developed a highly-efficient, algorithm to comprehensively find, analyze, and visualize the full repertoire of ndRNAs (miRNAs, sdRNAs & tRFs) contained within a RNA-seq NGS dataset. Strikingly, we find that >95% of ndRNA sequences identified by our algorithm that correspond to previously annotated, experimentally validated miRNAs agree within 1-2 nt [25] confirming the accuracy and reproducibility of our method. That said, in [25] we only applied our method to human miRNA, sdRNA, and tRF discovery since the approach was novel and existing experimental evidence in the literature on these three ndRNAs allowed us to rigorously evaluate our findings. Therefore, in the work described herein, we have expanded our resource scope to allow the user to comprehensively screen NGS datasets for all miRNA-like RNAs excised from any known human sncRNAs to likewise identify miRNA-like RNAs in 439 other organisms.

In the current article, we also discuss the development and usage of our new real-time, user-friendly web interface specifically designed for the aspiring biomedical experts to

easily perform complex computational tasks that are significant with respect to ndRNA research such as NGS data transfer, processing, analysis, visualization, and comparison. Such a platform could be extremely helpful even for the experts to explore complex patterns within ndRNA expressions across multiple samples. What makes our platform unique is the usage of specialized, light-weight data structures which are explained further in the paper.

In the next section, the current state of the related literature and the challenges associated are explained. Then in the later section, our algorithm, and the theory on which it is built-upon are described. Most importantly, in the current article, we provide qualitative evidence to evaluate our algorithm's effectiveness and provide some empirical validation. The Materials and Methods section, elaborates on our techniques, tools, and data sources. In the remaining sections, our application and its features are discussed along with example analyses. Finally, we conclude by summarizing our contributions and discussing future research directions.

II. Background and Related Work

NdRNAs are miRNA-like fragments specifically excised from sncRNAs. The first and foremost issue with effective ndRNA mining is that the concept of ndRNAs itself is relatively new in terms of both biology, and computer science. Therefore, there is a huge gap in the literature in terms of the theory behind computationally extracting such RNAs. Moreover, most of the already proposed methods use one of the existing SA methods—such as BLAST or Bowtie — to obtain pre-aligned outputs to roughly estimate the presence of ndRNAs. The problem with almost all the existing SA algorithms is that they are expensive either in terms of time or memory or both. For example, Bowtie is one such method which maps all the NGS reads to a reference genome. However, even though Bowtie is moderately faster, it requires at least 2GB of memory just for the human database. Plus, to date, only a handful of organisms' reference genomes are available to us. On the other hand, pair-wise local alignments such as BLAST require less memory but consume hours of processing time [25].

Furthermore, many of the existing ndRNA algorithms/tools [36][37] focus primarily on miRNAs and are not capable of fully defining ndRNA profiles. That said, the methods that are capable of characterizing novel ndRNAs [31-34] require fairly extensive computational expertise for utilization, and are dependent on pre-aligned file inputs such as BAM. For example, Flaimapper, one such gold-standard method that tries to predict miRNAs by parsing BAM formatted files, can only correctly identify 54% of experimentally validated miRNA end positions [34]. Anyway, such methods provided a means to strengthen our preliminary understanding of miRNAs, but they possess a large number of drawbacks. Hence, there is an urgent requirement for a direct computational approach to:

a) *Aid finding all unknown ndRNAs in any commonly studied organism, including ndRNAs excised from other types of ncRNAs.*

b) *Quickly and accurately identify, and compare the gene expression levels of all the known ndRNAs in millions of NGS files i.e., both existing and the files yet to come.*

c) *Provide a standard means of data analysis for the study of ndRNAs.*

Therefore, to address many of the aforementioned challenges, we proposed an algorithm to retrieve ndRNAs from RNA-Seq files without the necessity to use any intermediate tools. In the next section, we will briefly go through our strategy and explain how and why it works.

III. Materials and Methods

The main objective of the current body of work is to provide a straight forward approach to study ndRNAs to support domain experts explore the complex relation between ncRNAs, ndRNAs, and cancers. This section will describe the materials and methods required in order to provide some of our critical observations to validate our approach. Our proposed method consists of two major steps, 1. Data collection and pre-processing, 2. Sequence Alignment (MoVaK alignment) followed by ndRNA mining using wavelets (SURFR algorithm) [25]. Each of these steps along with theoretical and qualitative validation of our method are elaborated in the sub-sections below.

A. Data Collection and Pre-Processing

Our strategy of sequence alignment makes use of three data structures, namely, Aho-Corasick Automaton (ACA), Similarity Vector (SV), and Differential Expression Vector (DEV). ACA and SVs are necessary to perform the alignment and are required to be pre-processed before-hand. Therefore, in this sub-section, we will briefly discuss pre-processing approaches and the required data sources.

Data Collection: An up-to-date list of all the known ncRNAs of 440 species and the associated sequences are collected from reliable sources such as NCBI [40]. However, NCBI only contains certain types of ncRNAs and do not contain eukaryotic tRNAs. Therefore another database of tRNAs, GtRNAdb [41], has been utilized for obtaining the tRNA sequences.

ACA construction: Aho-Corasick (AC) algorithm is an efficient multi-keyword search strategy that uses the data structure, AC automaton (ACA) [42]. In our method, we make use of AC algorithm to perform a quick initial filtering of all the reads from a file.

SV generation: SV is a data structure that was introduced in a previous study [25] to store and process DNA/RNA sequences efficiently and effectively. SVs represent the information about each character of a genetic sequence stored in the form of binary arrays. Hence, four mandatory SVs for each sequence are constructed, i.e., for each of A, C, T, and G.

B. MoVaK Alignment and Wavelet-based ndRNA mining

The initial step in almost every NGS data analysis is SA, i.e., to match the NGS file's reads to a genome or a known database. Genome alignment methods are used to map all the reads in the file to a reference genome, and pair-wise alignment methods are used to align the reads to known databases. Genome alignment techniques consume high memory because

the entire genome has to be loaded into memory, while pair-wise alignments consume relatively higher time to thoroughly filter all the reads. However, all such traditional SA algorithms strictly consider the problem at hand as a permutation and combination-based string alignment problem and emphasize too much on potential mutations at individual read-level—thereby consuming heavy amounts of resources. Therefore, through our research, we found the necessity to revisit the problem of sequence alignment for NGS data analysis specifically to address the issue of ndRNAs. Therefore, in [25] a new way of SA that produces a data structure called Differential Expression Vector (DEV), as shown in figure 1 below, to identify ncRNA fragmentation. Once the DEVs for all the expressed ncRNAs in a given dataset are calculated using MoVaK alignment, our technique is to employ a wavelet-based approach for identifying differentially fragmented regions with lengths around 18-35 nt within the DEVs to automatically mine ndRNAs [25]. A major part of the current body of work includes showing how the concept of DEVs can be leveraged to perform automated NGS analysis and to construct robust real-time platforms with greater biological significance.

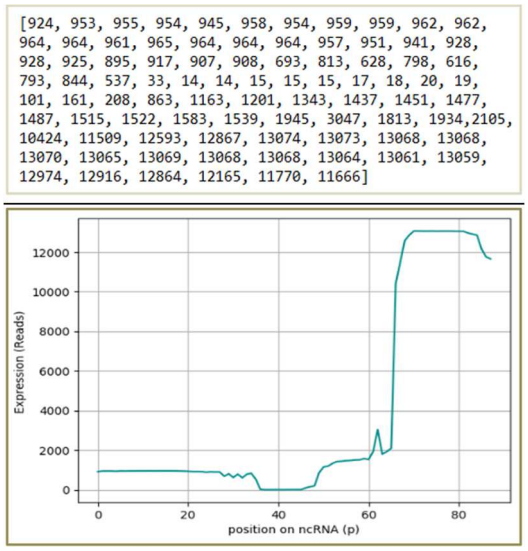

Fig. 1. 1(a) An example of a DEV. 1(b) Visualization of a DEV from 1.A representing the gene expression of one ncRNA.

C. *Theoretical Foundation for SURFR Algorithm*

Biologically, ndRNAs (miRs, sdRNAs, tRFs) are small (18-35nt) yet fully functional RNAs that are derived from larger sncRNAs. Therefore, in our theory, we hypothesized that "being derived into ndRNAs" is one of the characteristics/behaviors possessed by all the ncRNAs thus generalizing the problem statement. Moreover, we aimed to investigate if it is possible to provide a mathematical basis to understand the phenomenon of ncRNAs being processed into ndRNAs? In order to answer that question, it is important to discuss the relationship between ncRNA fragmentation and DEVs in a detailed manner.

Our main idea behind the DEV concept is to capture the entire 'activity' of a given ncRNA into a single construct (i.e., unit/function) representing the current snapshot of its expression within a sample/NGS file. However, instead of considering expression of a ncRNA as mere "read count" or "reads per million count", we use higher dimensional vectors to reformulate the current mathematical understanding of the concept of ncRNA expression.

To elaborate, all the cellular-level activities such as transcription, translation, signaling, etc., happen at real-time speeds. Moreover, some of the transcriptomic activities occur in a continuous/regular basis, while some others occur as a response to some external stimuli. That said, it is reasonable to think that a NGS file is a snapshot of the current genetic activities within a tissue/sample within which the expressed ncRNAs are being processed or already processed. Although we do not know the exact rates at which different ncRNAs are processed, it is reasonable to think that the expression/presence of ncRNAs is directly associated to their processing phenomenon. Therefore, we gathered the information from all the genetic sequences associated to each ncRNA together into a 1d matrix (DEV) to understand the physical process of ncRNA fragmentation. Such an understanding comes from the notion of complex Hilbert spaces representing the current state of a physical phenomenon. As predicted, we were successfully able to view the ncRNA fragmentation using DEVs which is further explained in the next section.

D. *Qualitative validation of our Algorithm: proof for the gene expression curves*

As of now, we explained a new way of quantifying the term "expression" as a 1d matrix as opposed to the common interpretation, i.e., read count or reads per million count or millions of individual strings. That said, in this section, we will reveal some exciting observations within DEVs which we used to layout a mathematical basis for ndRNAs. By comparing DEVs of the same ncRNAs from hundreds of files, we found that one of the direct applications of DEVs is to visualize ndRNAs being processed from ncRNAs. An example of the evidence showing DEVs for a known pre-miRNA (hsa-miR-27a) calculated 17 different RNA-Seq samples is shown in Figure 2 below with the miRNA is being derived/processed highlighted using red start and end lines. Excitingly, miRBase provides experimental validation to confirm that hsa-miR-27a is indeed processed into two miRNA fragments, as shown in Figure 3, one of which is located exactly at the locations as shown in our DEVs Figure 2. Similar phenomenon has also been observed in many sncRNAs instead of just miRNAs. Such example showing a tRF is shown in Figure 4.

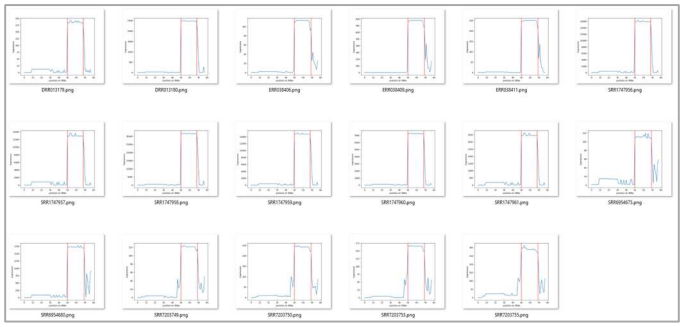

Fig. 2. DEVs and corresponding SURFR-identified miRs originating from known pre-miRNA (hsa-miR-27a) from 17 different samples showing the

mature miRNA fragment being derived from the same positions as validated by miRBase shown in Fig 3.

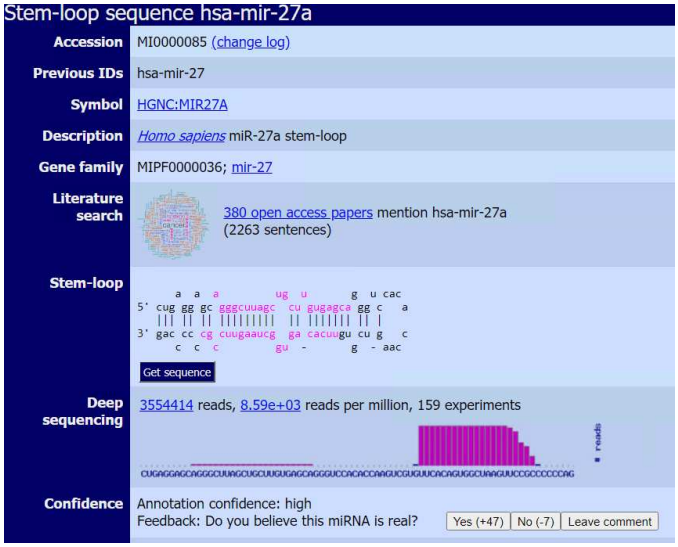

Fig. 3. Experimental evidence from miRBase for hsa-miR-27a.

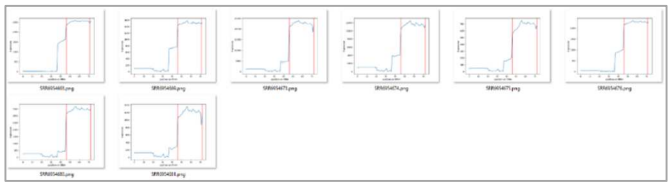

Fig. 4. DEVs of same tRNAs from eight different samples showing a tRF being derived.

To further evaluate that the validity of DEVs in terms of observing the differences in ncRNAs being processed, we performed our analysis on more RNA-Seq samples expecting to find such patterns across 1) the same ncRNAs in different cell types/conditions, 2) same ncRNAs in same cell type/condition, since ndRNA formation depends upon various factors as previously mentioned. Therefore, we chose some publicly available datasets from the NCBI's SRA database [38] to perform our tests. Figure 5 below shows the striking similarity between the DEVs for a misc_RNA in mcf-7 cells belonging to a study with Accession identifier: DRX048619, collected from 5 different samples over a period of 12 weeks. We can clearly notice that, even though the overall expression levels are varying across the samples and across multiple weeks, the gene expression patterns look very similar to each other—which was also apparent for many other ncRNAs from the same analysis. Next, in Figure 6, DEVs for the same snoRNA, SNORD96A, from both the same and different cell types are compared to each other where, SRR4217151 and SRR4217150 correspond to adenocarcinoma, and our DEVs clearly show that SNORD96A does not undergo any processing and is expressed as a full length snoRNA. The other four DEVs in Figure 6 correspond to prostate cancers where, SRR3400536 and SRR3400539 correspond to the same type, and SRR11061161 and SRR10186659 are from different prostate cancers. Overall, apparent differences in SNORD96A processing and ncRNA expression patterns are clearly depicted in the DEV array (Figure 6)—thus providing the basis for our method.

Fig. 5. Comparison of DEVs for 60 samples of same ncRNA across same cell type and condition.

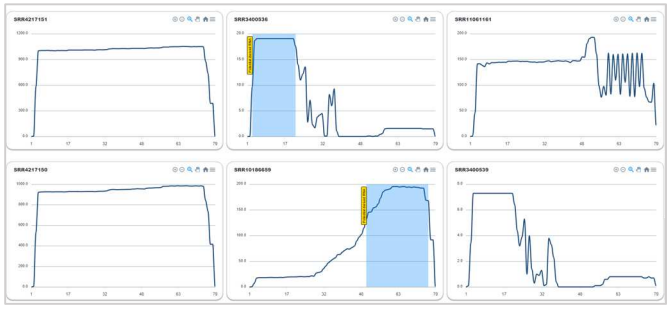

Fig. 6. DEVs for the same ncRNA across different cell types and conditions.

IV. SURFR Platform

SURFR is an interactive web application platform that we developed using the SURFR algorithm specifically to help experts perform automated and visual ncRNA fragmentation analysis to learn and explore new ndRNA functions. Most importantly, through our work we want to provide a standardized ndRNA mining procedure. In this section, the features of our web application along with some instructions on how to use these resources are provided.

A. Features of our Application

Our application is a combination of two portals SURFR, and SURFR-ULTRA designed explicitly to mine novel and annotated ndRNAs. Both are currently available online at salts.soc.southalabama.edu/surfr. Importantly, these resources have been carefully engineered to be capable of handling many of the issues with NGS data analysis including size, structure, format. Notable features are as follows:

Flexible Data Transfer: Our tool provides users with multiple options to transfer data. A total of 10 files can be transferred per each job to be analyzed simultaneously. Users have an option to upload their own files in the raw FASTA/ FASTQ format, or they also have an option to analyze any of the publicly-available files in the NCBI's SRA database by simply entering the SRA file identifier. If needed, users can also mix-and-match between the two options.

Multiple Organisms: As previously noted, we extended our method to identify ndRNAs from a total of 440 organisms including eukaryotes, plants, fungi, protist, and bacteria databases which are readily available in NCBI.

Interactive Results: Since the concept of ndRNAs is relatively new, SURFR is designed for users to explore and

interact with the results. Several filters can be applied to the results to view ndRNAs of interest. Most importantly, interactive visualizations are also included for each ndRNA to reflect gene expression patterns at a single nt level. Figure 7 provides an overview of our dashboard, and Figure 6 provides an example of the visual differential expression analysis portal of our application.

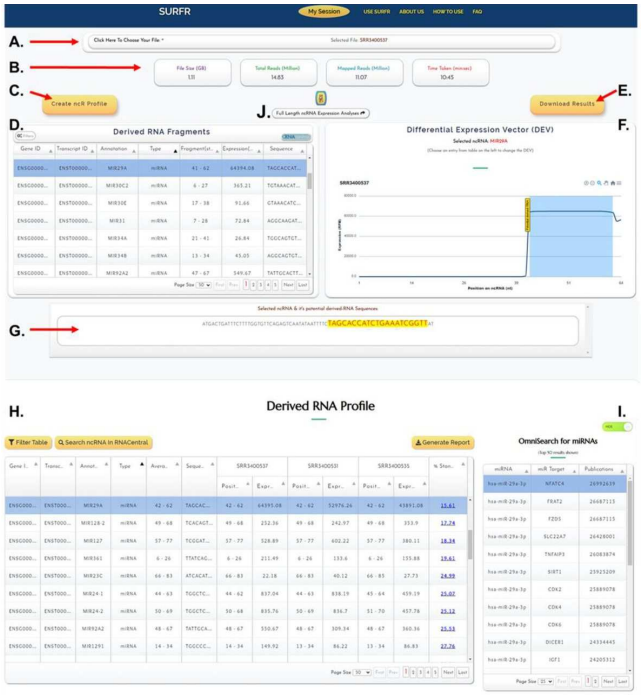

Fig. 7. SURFR dashboard. (A) drop down menu for selecting individual RNA-seq files. (B) A summary of the dataset and time taken. (C) the "Create ncR Profile" button which automatically populates the derived RNA Profile section at the bottom of the page. (D) details about each fragment identified in the individual, selected small RNA-seq dataset. (E) download an excel with all the ndRNAs and associated information (F) The "Differential Expression Vector (DEV)" visualized ndRNAs within the ncRNAs with a blue highlighted area. The x-axis represents the position in the ncRNA selected (e.g. miR-29a) and the y-axis depicts the expression levels of the ncRNA at each position. (G) The full length host ncRNA (miR-29a) highlighting the SURFR-called ndRNA in yellow. (H) Portal to compare nRNA expressions across samples. (I) The "OmniSearch for miRNAs" window lists the top 50 Omnisearch entries (reported targets and PubMed publications) for an individual microRNA selected in the "Derived RNA Profile" window (H). (J) The "Full Length ncRNA Expression Analyses" button in the upper center of the results page redirects the user to SURF-ULTRA platform designed for full-length sncRNA analysis and visual differential expression analysis.

SURFR-ULTRA and visual differential expression analysis:

SURFR-ULTRA is an extended version of our SURFR platform described above. The main goal of SURFR is to provide users with a list of all the ndRNAs obtained after filtering all the DEVs. However, in SURFR, only the ndRNAs with high-confidence levels are provided to the users to avoid any false-positives. But, SURFR-ULTRA allows the users to further explore and compare the full-length ncRNAs including the ncRNAs that are being processed into ndRNAs. An example analysis of SURFR-ULTRA is shown in Figure 6, where, the DEVs of the ncRNAs from multiple files are compared to each other. Such a platform can provide a means to explore the differences between ndRNA processing and how they are affected by different changes/conditions within samples. SURFR-ULTRA is made available to the users using a link on the SURFR platform as shown in Fig 7 (J).

Restorable and Comparable Sessions:

One of the major advantages of our platform compared to that of the others is our concept of restorable and comparable data analysis sessions where each session is protected using state-of-the-art cryptographic algorithms. This feature helps the users to utilize a SURFR-generated session ID/key to retrieve their previously completed data analysis sessions. This feature also enables users to compare files from different sessions simply by entering multiple session IDs/keys in our data retrieval portal, thus aiding the domain experts share their findings in a simple manner. Files from a total of three different sessions can be compared against each other, allowing the users to compare up to 30 files together.

V. Conclusion

This paper presented a novel, generalizable computational algorithm, SURFR, to identify and visualize miRNA-like RNAs in a highly effective and efficient manner. SURFR was designed to automatically extract ndRNAs followed by an intuitive visualization of the processing of these ndRNAs. SURFR is based on (1) a new computational theory that we introduced, where gene expression is considered a multi-dimensional construct and further interpreted as a signal using our data structure DEV; (2) an original pair-wise sequence alignment strategy, MoVaK alignment; and (3) a concept known as the DEVs. The current version of SURFR is able to handle all ncRNAs from about 450 organisms, including animals, plants, fungus, protist, and bacterial species, which are annotated and readily available in NCBI. In addition, a publicly available, interactive NGS data analytics platform has been built for domain experts to conduct RNA-Seq analyses for their ndRNA studies. Not only are users allowed to upload their own datasets, but they can also simply retrieve any publicly available SRA file to be automatically analyzed. Using our visual differential expression analysis, users can now detect single nucleotide level expression changes within ncRNAs in individual files and compare up to 30 files simultaneously allowing domain experts to easily identify complex patterns associated with the ndRNA expressions of interest.

Qualitative evidence was provided to evaluate our theoretical contribution by showing and discussing the proof for different gene expression curves. Altogether a total of three DEV pattern comparisons were used to illustrate ndRNA processing visualization: (1) same RNA same cell, (2) same RNA different cells, and (3) same RNA same cell but with different conditions. Our results show that changes in ncRNA processing are indeed reflected in our DEVs. Most importantly, by comparing DEVs from hundreds of files together, we were able to observe the wavelet-like behavior of the ndRNAs.

One potential direction of our future work is to further improve our algorithm's effectiveness by integrating deep-learning techniques.

References

[1] John S. Mattick and Igor V. Makunin. 2006. Non-coding RNA. Human Molecular Genetics 15, suppl_1 (April 2006), R17–R29. DOI:https://doi.org/10.1093/hmg/ddl046

[2] Rogerio Alves de Almeida, Marcin G. Fraczek, Steven Parker, Daniela Delneri, and Raymond T. O'Keefe. 2016. Non-coding RNAs and disease: the classical ncRNAs make a comeback. Biochem Soc Trans 44, 4 (August 2016), 1073–1078. DOI:https://doi.org/10.1042/BST20160089

[3] Wen-Tao Wang, Cai Han, Yu-Meng Sun, Tian-Qi Chen, and Yue-Qin Chen. 2019. Noncoding RNAs in cancer therapy resistance and targeted drug development. Journal of Hematology & Oncology 12, 1 (June 2019), 55. DOI:https://doi.org/10.1186/s13045-019-0748-z

[4] Giulia Romano, Dario Veneziano, Mario Acunzo, and Carlo M. Croce. 2017. Small non-coding RNA and cancer. Carcinogenesis 38, 5 (May 2017), 485–491. DOI:https://doi.org/10.1093/carcin/bgx026

[5] M. Lagos-Quintana, R. Rauhut, W. Lendeckel, and T. Tuschl. 2001. Identification of novel genes coding for small expressed RNAs. Science 294, 5543 (October 2001), 853–858. DOI:https://doi.org/10.1126/science.1064921

[6] Kenji Nakahara and Richard W. Carthew. 2004. Expanding roles for miRNAs and siRNAs in cell regulation. Curr Opin Cell Biol 16, 2 (April 2004), 127–133. DOI:https://doi.org/10.1016/j.ceb.2004.02.006

[7] Leigh-Ann MacFarlane and Paul R. Murphy. 2010. MicroRNA: Biogenesis, Function and Role in Cancer. Curr Genomics 11, 7 (November 2010), 537–561. DOI:https://doi.org/10.2174/138920210793175895

[8] Olga Plotnikova, Ancha Baranova, and Mikhail Skoblov. 2019. Comprehensive Analysis of Human microRNA–mRNA Interactome. Front. Genet. 10, (2019). DOI:https://doi.org/10.3389/fgene.2019.00933

[9] Maha Abdellatif. 2012. Differential Expression of MicroRNAs in Different Disease States. Circ Res 110, 4 (February 2012), 638–650. DOI:https://doi.org/10.1161/CIRCRESAHA.111.247437

[10] Antonio Marco. 2014. Sex-biased expression of microRNAs in Drosophila melanogaster. Open Biol 4, 4 (April 2014). DOI:https://doi.org/10.1098/rsob.140024.

[11] Barrie S. Bradley, Joseph C. Loftus, Clinton J. Mielke, and Valentin Dinu. 2014. Differential expression of microRNAs as predictors of glioblastoma phenotypes. BMC Bioinformatics 15, (January 2014), 21. DOI:https://doi.org/10.1186/1471-2105-15-21

[12] Farshid Kouhi, Karim Sorkheh, and Sezai Ercisli. 2020. MicroRNA expression patterns unveil differential expression of conserved miRNAs and target genes against abiotic stress in safflower. PLOS ONE 15, 2 (February 2020), e0228850. DOI:https://doi.org/10.1371/journal.pone.0228850

[13] Angie M. Cheng, Mike W. Byrom, Jeffrey Shelton, and Lance P. Ford. 2005. Antisense inhibition of human miRNAs and indications for an involvement of miRNA in cell growth and apoptosis. Nucleic Acids Res 33, 4 (2005), 1290–1297. DOI:https://doi.org/10.1093/nar/gki200

[14] H.-W. Hwang and J. T. Mendell. 2006. MicroRNAs in cell proliferation, cell death, and tumorigenesis. Br J Cancer 94, 6 (March 2006), 776–780. DOI:https://doi.org/10.1038/sj.bjc.6603023

[15] Jacob O'Brien, Heyam Hayder, Yara Zayed, and Chun Peng. 2018. Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation. Front. Endocrinol. 9, (2018). DOI:https://doi.org/10.3389/fendo.2018.00402

[16] Rajesha Rupaimoole and Frank J. Slack. 2017. MicroRNA therapeutics: towards a new era for the management of cancer and other diseases. Nature Reviews Drug Discovery 16, 3 (March 2017), 203–222. DOI:https://doi.org/10.1038/nrd.2016.246

[17] Johora Hanna, Gazi S. Hossain, and Jannet Kocerha. 2019. The Potential for microRNA Therapeutics and Clinical Research. Front Genet 10, (May 2019). DOI:https://doi.org/10.3389/fgene.2019.00478

[18] Julia Alles, Tobias Fehlmann, Ulrike Fischer, Christina Backes, Valentina Galata, Marie Minet, Martin Hart, Masood Abu-Halima, Friedrich A Grässer, Hans-Peter Lenhof, Andreas Keller, and Eckart Meese. 2019. An estimate of the total number of true human miRNAs. Nucleic Acids Research 47, 7 (April 2019), 3353–3364. DOI:https://doi.org/10.1093/nar/gkz097

[19] Mengqian Yu, Bingjian Lu, Jisong Zhang, Jinwang Ding, Pengyuan Liu, and Yan Lu. 2020. tRNA-derived RNA fragments in cancer: current status and future perspectives. Journal of Hematology & Oncology 13, 1 (September 2020), 121. DOI:https://doi.org/10.1186/s13045-020-00955-6

[20] Xin He, Xinxin Chen, Xue Zhang, Xiaobing Duan, Ting Pan, Qifei Hu, Yijun Zhang, Fudi Zhong, Jun Liu, Hong Zhang, Juan Luo, Kang Wu, Gao Peng, Haihua Luo, Lehong Zhang, Xiaoxi Li, and Hui Zhang. 2015. An Lnc RNA (GAS5)/SnoRNA-derived piRNA induces activation of TRAIL gene by site-specifically recruiting MLL/COMPASS-like complexes. Nucleic Acids Research 43, 7 (April 2015), 3712–3725. DOI:https://doi.org/10.1093/nar/gkv214

[21] Elena S. Martens-Uzunova, Michael Olvedy, and Guido Jenster. 2013. Beyond microRNA – Novel RNAs derived from small non-coding RNA and their implication in cancer. Cancer Letters 340, 2 (November 2013), 201–211. DOI:https://doi.org/10.1016/j.canlet.2012.11.058

[22] Dillon G. Patterson, Justin T. Roberts, Valeria M. King, Dominika Houserova, Emmaline C. Barnhill, Aline Crucello, Caroline J. Polska, Lucas W. Brantley, Garrett C. Kaufman, Michael Nguyen, Megann W. Santana, Ian A. Schiller, Julius S. Spicciani, Anastasia K. Zapata, Molly M. Miller, Timothy D. Sherman, Ruixia Ma, Hongyou Zhao, Ritu Arora, Alexander B. Coley, Melody M. Zeidan, Ming Tan, Yaguang Xi, and Glen M. Borchert. 2017. Human snoRNA-93 is processed into a microRNA-like RNA that promotes breast cancer cell invasion. NPJ Breast Cancer 3, (July 2017). DOI:https://doi.org/10.1038/s41523-017-0032-8

[23] Ze Chen, Yu Sun, Xiaojun Yang, Zhenfeng Wu, Kaifei Guo, Xiaoran Niu, Qingsong Wang, Jishou Ruan, Wenjun Bu, and Shan Gao. 2017. Two featured series of rRNA-derived RNA fragments (rRFs) constitute a novel class of small RNAs. PLOS ONE 12, 4 (April 2017), e0176458. DOI:https://doi.org/10.1371/journal.pone.0176458

[24] Tess Cherlin, Rogan Magee, Yi Jing, Venetia Pliatsika, Phillipe Loher, and Isidore Rigoutsos. 2020. Ribosomal RNA fragmentation into short RNAs (rRFs) is modulated in a sex- and population of origin-specific manner. BMC Biology 18, 1 (April 2020), 38. DOI:https://doi.org/10.1186/s12915-020-0763-0

[25] Mohan Vamsi Kasukurthi, Dihua Zhang, Mika Housevera, Yulong Huang, Shaobo Tan, Bin Ma, Dongqi Li, Ryan Benton, Jingwei Lin, Shengyu Li, Glen M. Borchert, and Jingshan Huang. 2019. SURFr: Algorithm for identification and analysis of ncRNA-derived RNAs. In 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 1504–1507. DOI:https://doi.org/10.1109/BIBM47256.2019.8983074

[26] Leslie E. Rogler, Brian Kosmyna, David Moskowitz, Remon Bebawee, Joseph Rahimzadeh, Katrina Kutchko, Alain Laederach, Luigi D. Notarangelo, Silvia Giliani, Eric Bouhassira, Paul Frenette, Jayanta Roy-Chowdhury, and Charles E. Rogler. 2014. Small RNAs derived from lncRNA RNase MRP have gene-silencing activity relevant to human cartilage-hair hypoplasia. Hum Mol Genet 23, 2 (January 2014), 368–382. DOI:https://doi.org/10.1093/hmg/ddt427

[27] Sam Behjati and Patrick S Tarpey. 2013. What is next generation sequencing? Arch Dis Child Educ Pract Ed 98, 6 (December 2013), 236–238. DOI:https://doi.org/10.1136/archdischild-2013-304340

[28] Valerio Bianchi, Arnaud Ceol, Alessandro G. E. Ogier, Stefano de Pretis, Eugenia Galeota, Kamal Kishore, Pranami Bora, Ottavio Croci, Stefano Campaner, Bruno Amati, Marco J. Morelli, and Mattia Pelizzola. 2016. Integrated Systems for NGS Data Management and Analysis: Open Issues and Available Solutions. Front. Genet. 7, (2016). DOI:https://doi.org/10.3389/fgene.2016.00075

[29] Jerzy K. Kulski. 2016. Next-Generation Sequencing — An Overview of the History, Tools, and "Omic" Applications. Next Generation Sequencing - Advances, Applications and Challenges (January 2016). DOI:https://doi.org/10.5772/61964

[30] Human Genome Project FAQ. Genome.gov. Retrieved December 16, 2020 from https://www.genome.gov/human-genome-project/Completion-FAQ

[31] Yuk Yee Leung, Pavel P. Kuksa, Alexandre Amlie-Wolf, Otto Valladares, Lyle H. Ungar, Sampath Kannan, Brian D. Gregory, and Li-San Wang. 2016. DASHR: database of small human noncoding RNAs. Nucleic Acids Res 44, D1 (January 2016), D216-222. DOI:https://doi.org/10.1093/nar/gkv1188

[32] Pavel P. Kuksa, Alexandre Amlie-Wolf, Živadin Katanic, Otto Valladares, Li-San Wang, and Yuk Yee Leung. 2018. SPAR: small RNA-seq portal for analysis of sequencing experiments. Nucleic Acids Res 46, W1 (July 2018), W36–W42. DOI:https://doi.org/10.1093/nar/gky330

[33] Junchao Shi, Eun-A. Ko, Kenton M. Sanders, Qi Chen, and Tong Zhou. 2018. SPORTS1.0: A Tool for Annotating and Profiling Non-coding RNAs Optimized for rRNA- and tRNA-derived Small RNAs. Genomics Proteomics Bioinformatics 16, 2 (April 2018), 144–151. DOI:https://doi.org/10.1016/j.gpb.2018.04.004

[34] Youri Hoogstrate, Guido Jenster, and Elena S. Martens-Uzunova. 2015. FlaiMapper: computational annotation of small ncRNA-derived fragments using RNA-seq high-throughput data. Bioinformatics 31, 5 (March 2015), 665–673. DOI:https://doi.org/10.1093/bioinformatics/btu696

[35] Guillermo Barturen, Antonio Rueda, Maarten Hamberg, Angel Alganza, Ricardo Lebron, Michalis Kotsyfakis, Bu-Jun Shi, Danijela Koppers-Lalic, and Michael Hackenberg. 2014. sRNAbench: profiling of small RNAs and its sequence variants in single or multi-species high-throughput experiments. Methods in Next Generation Sequencing 1, 1 (January 2014). DOI:https://doi.org/10.2478/mngs-2014-0001

[36] Clarissa P. C. Gomes, Ji-Hoon Cho, Leroy Hood, Octávio Luís Franco, Rinaldo Wellerson Pereira, and Kai Wang. 2013. A Review of Computational Tools in microRNA Discovery. In Front. Genet. DOI:https://doi.org/10.3389/fgene.2013.00081

[37] Tobias Fehlmann, Christina Backes, Mustafa Kahraman, Jan Haas, Nicole Ludwig, Andreas E. Posch, Maximilian L. Würstle, Matthias Hübenthal, Andre Franke, Benjamin Meder, Eckart Meese, and Andreas Keller. 2017. Web-based NGS data analysis using miRMaster: a large-scale meta-analysis of human miRNAs. Nucleic Acids Res 45, 15 (September 2017), 8731–8744. DOI:https://doi.org/10.1093/nar/gkx595

[38] Home - SRA - NCBI. Retrieved April 13, 2021 from https://www.ncbi.nlm.nih.gov/sra

[39] BAM File Format. Retrieved April 13, 2021 from https://support.illumina.com/help/BS_App_RNASeq_Alignment_OLH_1000000006112/Content/Source/Informatics/BAM-Format.htm

[40] National Center for Biotechnology Information. Retrieved April 14, 2021 from https://www-ncbi-nlm-nih-gov.libproxy.usouthal.edu/

[41] GtRNAdb: Genomic tRNA Database. Retrieved April 14, 2021 from http://gtrnadb.ucsc.edu/

[42] Alfred V. Aho and Margaret J. Corasick. 1975. Efficient string matching: an aid to bibliographic search. Commun. ACM 18, 6 (June 1975), 333–340. DOI:https://doi.org/10.1145/360825.360855

[43] Harriet Mellenius. 2015. Speed and accuracy in transcription and translation : Modelling of transcript and polypeptide elongation. (2015). Retrieved April 14, 2021 from http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-262698

[44] Paul Kocher and San Francisco. Differential Power Analysis. 10.

[45] Indumathi Saikumar. DES- Data Encryption Standard. 04, 03 , 6.

[46] Jean-sebastien Coron, David Naccache, and Paul Kocher. 2000. Statistics and secret leakage. In In Proceedings of Financial Cryptography, Springer-Verlag, 157.

[47] I. Daubechies. 1990. The wavelet transform, time-frequency localization and signal analysis. IEEE Transactions on Information Theory 36, 5 (September 1990), 961–1005. DOI:https://doi.org/10.1109/18.57199

[48] Pan Du, Warren A. Kibbe, and Simon M. Lin. 2006. Improved Peak Detection in Mass Spectrum by Incorporating Continuous Wavelet Transform-based Pattern Matching. Bioinformatics 22, 17 (August 2006), 2059–2065. DOI:https://doi.org/10.1093/bioinformatics/btl355