BIOINFORMATIC IDENTIFICATION OF SMALL RNA TARGETS IN THE GLAUCOPHYTE ALGA *CYANOPHORA PARADOXA*: ANALYSIS OF ANCIENT EVOLUTIONARY MECHANISMS THAT AROSE IN PRIMORDIAL ALGAE

Sana Wajid

Rutgers University, School of Environmental and Biological Sciences

Undergraduate Research in Biotechnology

May 30, 2014

swajid@eden.rutgers.edu

ABBREVIATIONS

Shorthand		Notes
siRNA	small interfering RNA	Exogenous sources such as viral, dsRNA is usually perfectly base-paired
miRNA (not used in this paper because of misleading shorthand however this is used in the references)	microRNA	Endogenous (found in the genome) dsRNA base-pairing contains mismatches
RNAi	RNA interference	
AGO	argonaute	Protein family which can bind short sequences of
piRNA	PIWI-interactingRNA	Type of sRNA found in germ-line cells
RISC	RNA-*induced*-silencing complex	
dsRNA	double-stranded RNA	
ssRNA	single-stranded RNA	
CDS	predicted protein coding sequences	
GC	genomic contig	
EST	EST contig	

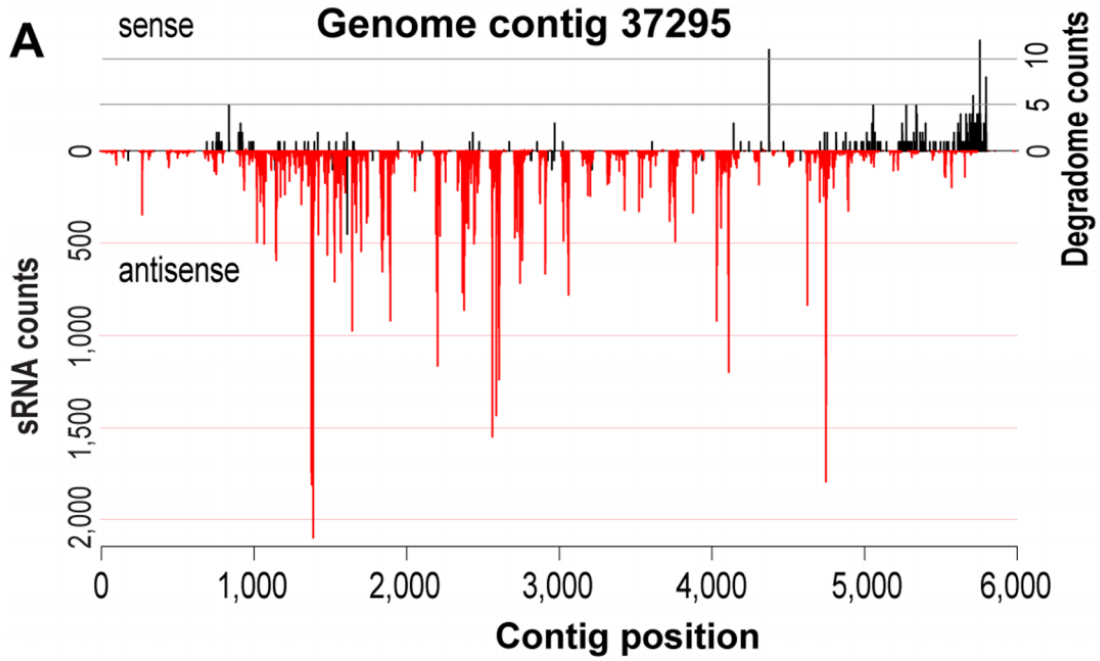
TABLES

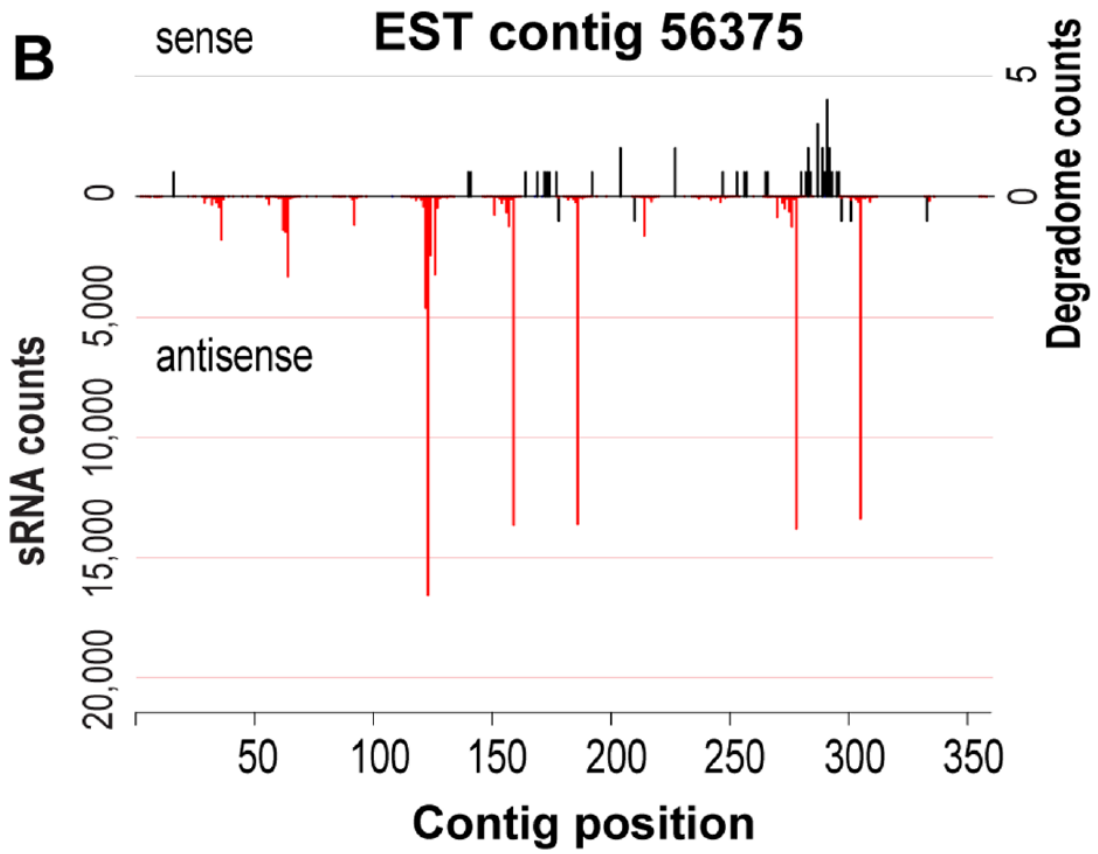
Table 1. Reads mapped

	CDS data	
Sense reads (+)	36,7709	20.86%
Antisense reads (-)	1,394,966	79.14%
Total reads	1,762,675	
Contigs	31,895	

FIGURES

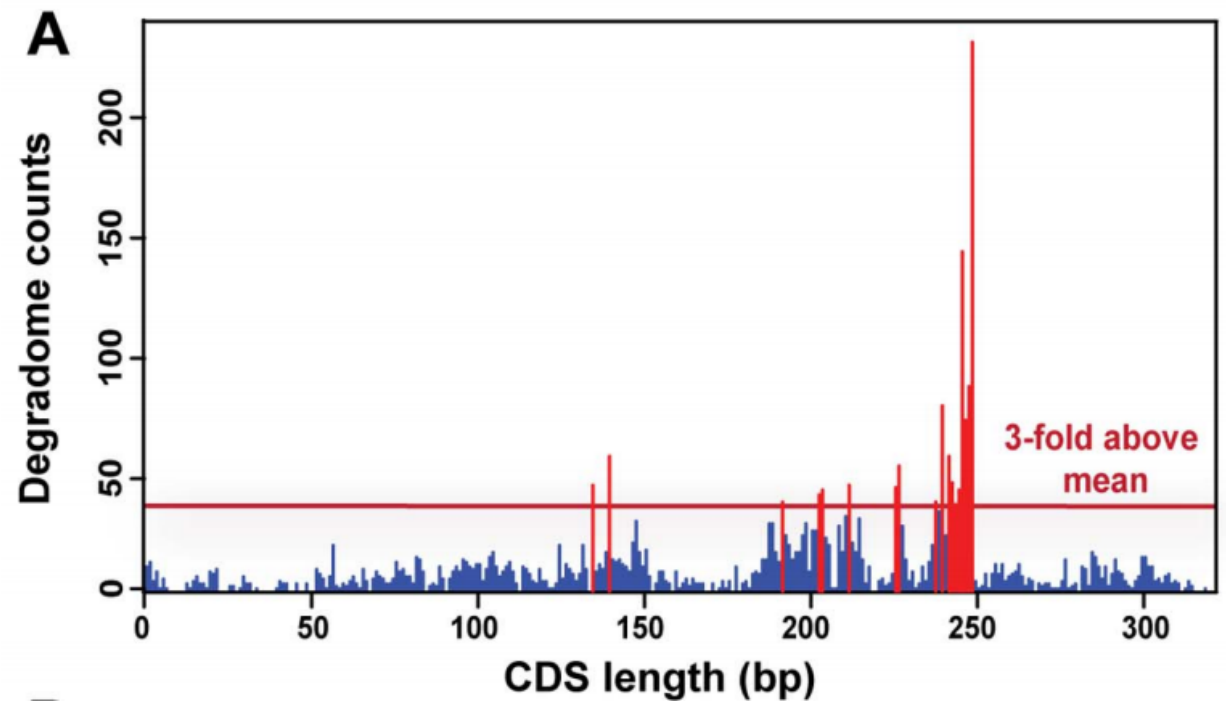
Figure 1. Mapped 5' position of degradome tags on (A). Genomic Contig (B) CDS contig





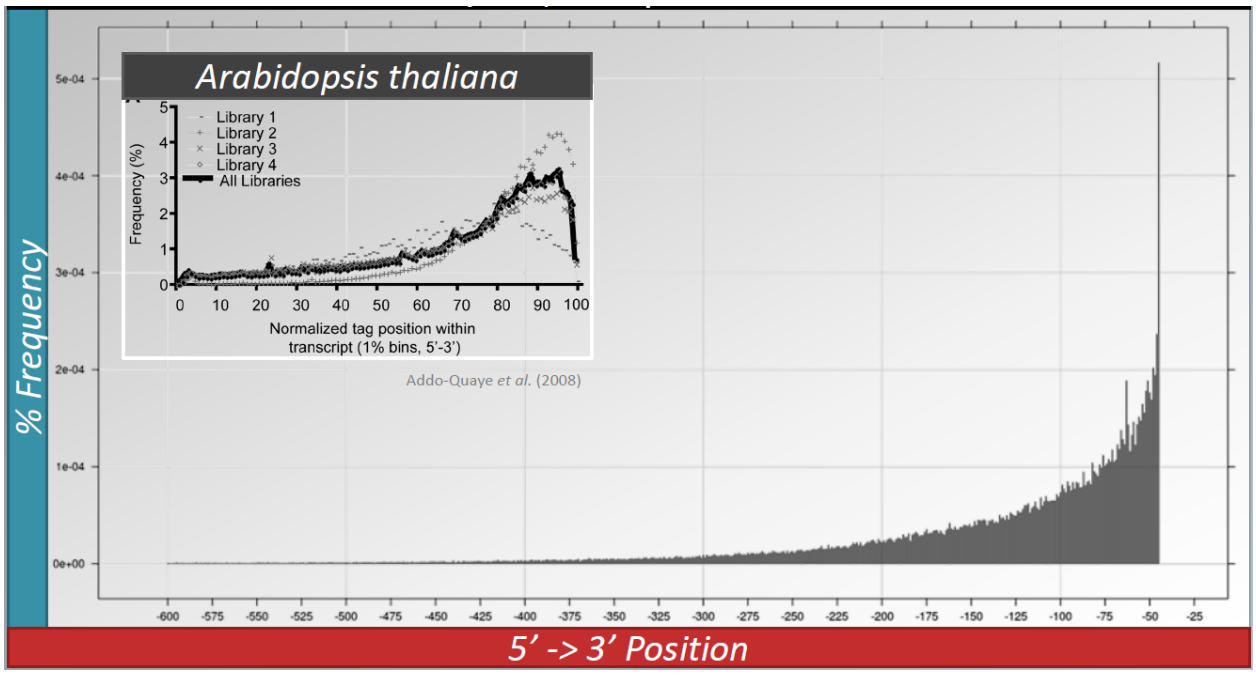
(A,B) Precise capture of degradome tags where the first position of the 5'-ends is mapped to their location on the genomic contig 37295 from *Cyanophora paradoxa.* Genome contig, x37295 with mapped sense [degradome tags (black)] and antisense [sRNAs (red)] plotted with relative position within contig.

Figure 2. Degradome profile of Contig 7658.3



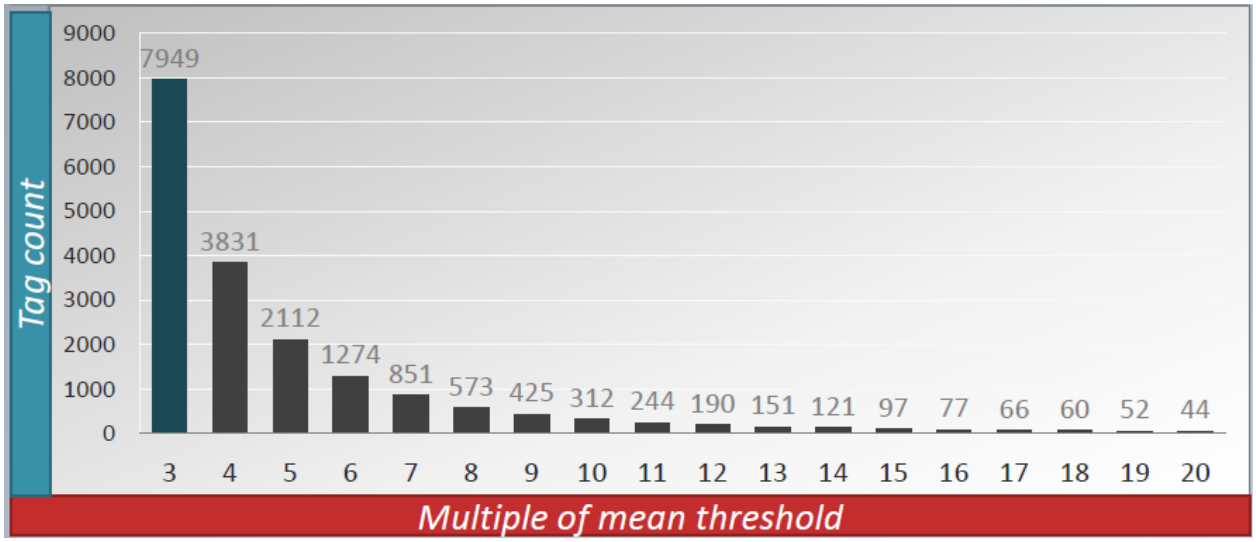
Within a CDS, a mean value was calculated regarding the frequency of tag counts. Degradome tag counts that are 3-fold above the mean (e.g. 3x the mean) are displayed in red.

Figure 3. Portrait of an aggregate exonucleolytic profile: 5' -> 3' Degradome ends signifying mRNA decay



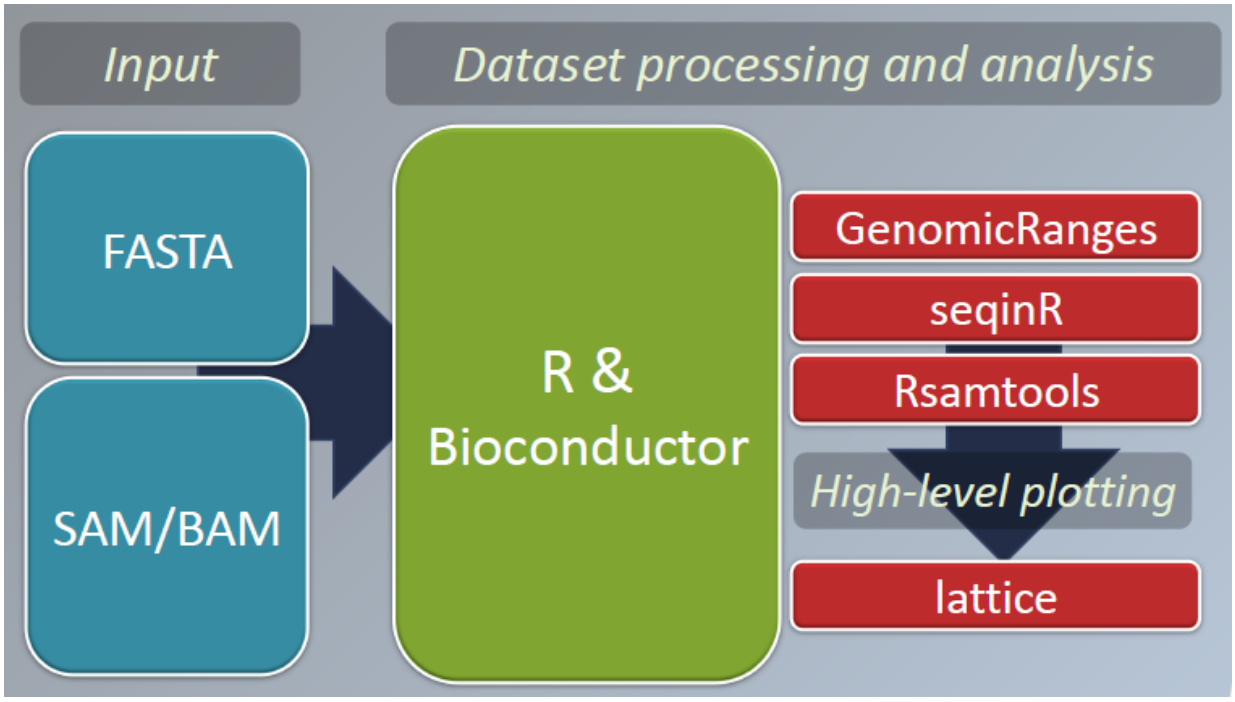
The portrait of the exonucleolytic profile of *Cyanophara paradoxa* is compared with a similar profile from *Arabidopsis thaliana*. The plot is a histogram where the x-axis is the numbers of nucleotides away from the end of the contig which is also the location of the putative stop codon. The percent frequency is normalized against all the contigs.

Figure 4. Abundance of cleavage tags in Cyanophora paradoxa



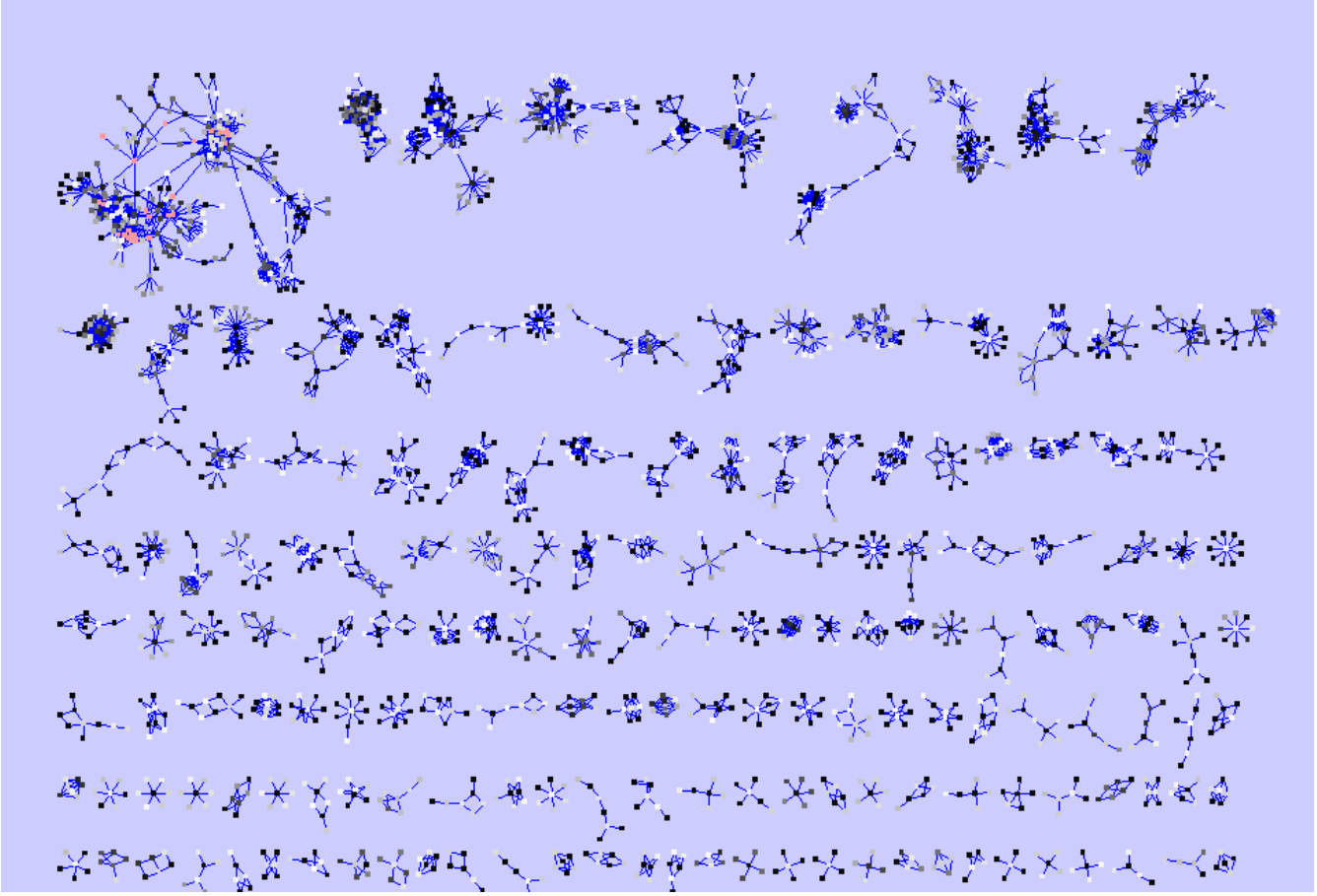
A histogram of "NX" vs. Tag count of cleavage tags where N = {3:20}, and X signifies that multiples of tag count that were found.

Figure 5. Pipeline for aligned small RNA seq data



The pipeline I developed for this research utilizes these steps for the input of very large sequence alignment files then uses R/Bioconductor for dissemination and reorganization.

Figure 6. A small RNA seq network



Clusters of sequence similarity of sRNAs. First, I selected top sequences using R, generated a .fasta file from the table which I submitted to BLASTClust and then visualized the results in Cytoscape.

INTRODUCTION AND BACKGROUND

Tracking photosynthetic organelles: cyanelle, muroplast and plastid; their genomes and other "omes" from the model organism Cyanophora paradoxa to elucidate primary endosymbiosis

Photosynthetic eukaryotes (e.g. algae, land plants), which contain the cellular metabolic and photosynthetic organelles mitochondria and plastids (eg. chloroplasts) are products of endosymbiosis events. The plastid originated at least once > 1.5 billion years ago, where a cyanobacterial ancestor that was captured by a single-celled protist as food-stuff but remained as an endosymbiont [1-10]. Following many photoautotrophic cyanobacterial capture events, a novel metabolic-toolkit, which involved photosynthesis, emerged among microbiological life. The geochemically active Earth and its readily adapting inhabitants cyclically co-evolved through various experimentation of genomic fine-tuning and gave rise to algae and plants [1-10]. Autonomous genomic replication was a pivotal stage in evolution and parasites which could utilize this mechanism propagated their genetic material [25]. Plastids are semiautonomous metabolic organelles, found in all alga and plant cells that are dependent on their nuclear genomes to support their functions [7,8].

Cyanophora paradoxa is a model organism for the endosymbiosis because it contains two blue-greenish vesicles called cyanelles in its protoplasm which were are thought to be acquired endosymbiotically from cyanobacteria [5]. *Cyanophora paradoxa* is a Glaucophyte, an anciently diverged group that split before its sister groups, the red

(*Rhodophyta*) and green (*Viridiplantae*) lineages within the nonophyletic eukaryotic supergroup Plantae [1,5]. Found only in Glaucophytes, muroplasts or cyanoplasts are chloroplasts with double-membranes that sandwich cyanobacterial-derived peptidoglycan (PG) much like those of the prokaryotes [8,9].

There exist combinations of singular to multiple endosymbiotic events, selection and gene transfers (e.g. protein translocation machinery from muroplast) [10]. At the very least it may be concluded of the current survivors in the tree of life that they possess the "core" metabolic processes which allowed for photosynthesis and other features found in plastids which were then transferred to the nuclear genome [8,10,11]. Relics of these genes are found in nuclear genome while plastids autonomously produce 100 of their approximately 2,500 proteins [8]. Molecular phylogenetics and comparative genomics are toolkits that are utilized to make inferences about the evolutionary past and has been bolstered by next generation sequencing platforms [31].

RNA turnover mechanisms utilize RNA interference

The central dogma of molecular biology reserves the role of RNA as an intermediary between DNA and its encoded protein product thus linking genotype and phenotype. The eukaryotic genome contains multiple layers of information from which a small portion forms the transcriptome consisting of either housekeeping RNA (e.g. rRNA, tRNA) or diverse regulatory RNAs which encompasses a large array of newly discovered species of RNA and are classified based on biochemistry of biogenesis, function and structure

[21,26]. Non-protein-coding RNA is a large component of the genome and may better correlate to complexity of an organism than the total size of genomes [20,29].

First discovered in the *Caenorhabditis elegans* as endogenously expressed 22-nt long miRNA *lin-4*, RNAi is a new member in the gene-regulation toolkit, controlling cell growth, differentiation and proliferation through sequence-specific transcript up and down regulation [29,32]. RNAi is primarily used for gene silencing (i.e. inhibition of gene expression) and also through siRNA, whereby double-stranded RNA intermediates are tagged then undergo degradation [24]. Although RNA interference (RNAi) is eukaryote specific, it is an ancient process that likely existed before the split of prokaryotes and eukaryotes where RNAi may have served primarily as an anti-viral and transposon defense mechanism [24,25]. For example, in early evolution, genomic defense against self and non-self RNA became an important tool to abate pervasive and parasitic self-replicating autonomous genomes [25]. Moreover, numerous types of endogenous small RNAs have been discovered. These include small RNAs anywhere from 20-30 nt in length and are catalogued in online databases [27].

Outcomes of the RNAi gene-protein network

Canonical RNAi mechanism is triggered by the formation of dsRNA duplex precursor which is cleaved by Ribonuclease III enzyme Dicer at both termini and then loaded onto the Argonaute complex [24]. Products of Dicer cleavage are small interfering RNAs (siRNAs) of variable sizes (21-25 nt) with siRNA 3'-2 nt overhang pairing and 5'-monophosphates [24]. The RNA-induced silencing complex (RISC) core component

protein Argonaute incorporates the formed siRNA duplex and cleaves one of the paired strands (passenger strand) through strand selection [24]. The remaining guide strand loaded Argonaute uses the sequence-specific ribonucleoprotein complex to perform subcellular surveillance by binding to complementary ssRNA then silence at two fronts: posttranscriptional to complementary target mRNA or cotranscriptional to cDNA produced by genomic loci [24,27,29].

A form of negative regulation, post-transcriptional silencing can occur through either: translational repression then RNA decay or endonucleolytic cleavage through argonaute slicing mechanisms

Major hallmarks of diverse small RNAs (e.g. microRNA, siRNA) are the following proteins signifying each of the three main stages in RNAi. First, a Dicer-like protein which varies between organisms with a RNAse III active center used for recognition (3' Dicer pocket to 3' end of dsRNA) and cleavage of dsRNA (via phosphodiester hydrolysis) and N-terminus helicase domain (also used for recognition) [17,29]. Second, a highly conserved Argonaute, subsets of RNase H family of enzymes and core component of the effector complex RNA-induced silencing complex (RISC) [25,28]. Last, RNA-dependent RNA polymerase (RdRP) produces from template ssRNA, dsRNA which may be targeted by Dicer-like endonucleases [20, 25, 34]. Double stranded RNA can also be constructed using RdRPs through either without primers or using siRNAs as primers [22]. Efficacy of transcript silencing with depends also on a seed sequence of 2-6 nt designed from the guide strand which is used to initialize complementary binding and may involve

imperfect base-pairing as well (e.g. microRNAs and siRNAs) [29]. The process is further complicated with double stranded RNA binding proteins (dsRBPs), that may not only mediate the handing off of dsRNA from Dicer to RISC but are also involved in strand selection, ribonucleoprotein complex stabilization and other fine-tuning [29].

Small RNA pathways in plants are diverse and perhaps reflect the importance of defense against stresses (biotic and abiotic) for sessile organisms to thrive [16]. Post-transcriptional gene silencing (PTGS) through distinct genomic loci is involved in the plant's self defence system which inactivates, degrades or represses mRNA. In one example, plants utilize natural antisense transcript-derived siRNAs (natsiRNAs) from constitutively expressed transcripts whose cDNA (antisense) is transcribed under stress (e.g. salt stress) to generate dsRNA utilized for AGO targeted cleavage [16]. Other examples includes siRNAs silencing via *cis* or *trans* after a dsRNA is sliced by Dicer-like endonucleases into 21-24 nt sized double stranded fragments. Silencing in *cis*, casiRNAs (size: 24 nt) are triggered by transposons and repeats that are used endogenous as DNA methylation at homologous loci [16]. Secondary regulators silencing, in *trans* with tasiRNas (size: 21 nt) is triggered by miRNAs that cleave mRNA from TAS loci for phased processing and currently 8 *TAS* loci have been identified in Arabidopsis [16,18].

Evolutionary origins of complex small RNA pathways in eukaryotes

Recent studies in small RNA regulation reveal complex regulatory networks within cells which are products of diverse yet fundamental RNAi pathways [1,25]. The RNAi

mechanism among different members of the tree of life is diverse, directly referring to the protein components of the pathways which have been diversified due to multiple paralogous events [20]. According to Shabalina and Koonin (2008), the Last Eukaryotic Common Ancestor, or LECA may have contained at least the following three key RNAi proteins as primordial peptides: Ago-Piwi, Dicer or Dicer-like analogues and RDRP which means at least silencing at the translational and transcriptional level could have been possible with thorough Ago and Piwi pathways, respectively [20].

Comparative genomic studies based on existence of the three main components of RNAi by Cerutti and Mollano (2006) and Shabalina and Koonin (2008) using also parsimonious evaluation reveal several unicellular eukaryotes which have lost RNAi functionalities, perhaps independently and multiple times: *Saccharomyces cerevisiae* (Opisthokonta), *Trypanosoma cruzi* and *Leishmania major* (Excavata), *Cyanidioschyzon merolae* (Archaeplastida), and *Plasmodium falciparum* (Chromalveolata) [20, 22]. It is likely that the RNAi mechanism is not necessary for unicellular eukaryotic life whereas it remains an integral part of multicellular eukaryotic developmental gene regulation (e.g. complex body pattern formation) [20].

Next generation sequencing of non-coding RNAs and genome wide analysis of small RNAs in Cyanophora paradoxa

Next generation sequencing methods allow for a clearer picture of the composition of genomes and transcriptomes (i.e., RNA expression). A wider density of small RNA sequences allows for a varied and more opaque visual where before RNA-seq methods

were developed, microarrays were utilized to study and profile the transcriptome and gene expression [34].

Next generation sequencing is simply put, a massive parallelization of the original Sanger sequencing method. In particular, the Illumina Platform utilizes reversible termination. Bridge amplification followings adaptor ligation of the (c)DNA fragments and is the key step for massive parallelization in the Illumina platform [42]. A cyclic addition of fluorescent primer then its removal identified the sequence of each nucleotide base and each sequencing cluster located on the flow cell is imaged then the sequence is outputted as a raw, encoded FASTQ format which is ready to be disseminated through a quality checking (QC) bioinformatics pipeline [42].

Eukaryotic RNA turnover was an important development to remove aberrant transcripts through various degradative pathways

Post-transcriptional gene regulation selects from a cellular pool of mRNA, transcripts which are significant for translation while others are go through a garbage collection process of RNA degradation and RNA silencing [37]. Mature mRNA molecules consists of distinct parts: 5'-7-methyl guanosine cap and 3' poly-A tail which together protect mRNA molecules from RNases [37].These molecular identifiers are signals for cellular processes such as RNA turnover but also serve as key genome features in the field of bioinformatics to study the degradome [37].

The 5' monophosphate is an important biochemical marker

In RNAi, the product of Argonaute-initiated cleavage of target dsRNA with perfect base complementarity results in two products. Degradation of mRNA or RNA turnover is a specific event that can begin at the 5' or 3' end of mature mRNAs and is highly conserved within the eukaryotes [34]. Capped 5' cleavage product contains: the 5' m7Gpp end and 3' without a poly-A tail where the capped mRNA is signaled for degradation by the 5'->3' RNA degradation pathway [34, 37]. The other half of sliced mRNA is the 3' cleavage product which contains the 5' monophosphate end and 3' poly-A tail and this species persists [34]. Thereafter, both cleavage products are then susceptible to 3'->5' or 5'->3' degradation: in 5'->3' degradation, the fundamental 5' cap is removed and exonucleolytic cleavage continues in the 5' -> 3' direction while in 3' -> 5' degradation, deadenylation then 3'->5' exonuclease activity [34, 37]. Furthermore, Nonsense-mediated RNA decay (NMD) is a process that if an incorrect stop codon is placed earlier in the transcript, NMD proceeds with 5' cap removal then 5'->3' exonuclease degradation without deadenylation [15]. The breadth of silencing is dependent on the argonaute protein properties such as its own ribonuclease activity, associated interacting proteins and localization within the cell [27].

The 5' monophosphate is fundamental to bioinformatic studies because it allows us to trace the degradome: decapping prevents mRNA from translation process

Uncapped/cleaved mRNA molecules undergo the RACE protocol to rapidly amplify downstream cleavage fragments uncapped 5’ monophosphate with polyadenylated tail fragments that are first ligated to RNA adaptors at the 3’ end [1, 37, 38]. The 5' ends of uncapped mRNAs are RNA degradation intermediates and genomic-specific targets which are degraded through small RNA-directed protein complexes and sequencing of cleaved key fragments, such as the 3’ ends of mRNA cleavage products reveal the genome-wide mapping of uncapped and cleaved transcripts [1, 37]. A snapshot of the transcriptome is achieved using methods that exist for degradome sequencing and identification of si/miRNA cleavage sites by locating the 5’ monophosphate and study 5’->3 decay. Genome-wide mapping of uncapped transcripts (GMUCT) is a protocol which samples the 5’-ends of uncapped mRNAs but not 3’-5’ decay [27]. Since T4 RNA ligase 1 requires uncapped 5’ monophosphate, sequenced fragments by GMUCT libraries are specific.

HYPOTHESIS, METHODS, MATERIALS & RESULTS

Experimental Rationale

Deep next-generation sequencing allows identification of different types of putative small RNA families and pathways enough so that potentially each genome nucleotide may be saturated with small RNA target for the first time [25]. Understanding how small RNAs regulate post-transcriptional gene expression in organisms is a grand challenge in biology and bioinformatics. For this study, a means of visually analyzing RNA-seq degradome data from *Cyanophora paradoxa* is proposed.

Our lab has recently sequenced genome of *Cyanophora paradoxa* and this was used as a template for preliminary evolutionary bioinformatic checks to indicate presence of the RNAi molecular toolkit [1,7]. According to Gross et. al (2013), *Cyanophora paradoxa* nuclear genome contains gene model matches with high BLASTp e-values to three RNAi components: Dicer, Ago and RDRp [1]. In this experiment to understand RNAi mechanism within basal alga species, a genome-wide analysis of sRNAs in the glaucophyte alga *Cyanophora paradoxa* was conducted to elucidate the ancestral features of the RNAi system [1].

Growth and subsequent library preparation of Cyanophora paradoxa small RNA

Four different conditions were set up for *C. paradoxa* cultures to grow on: normal, salt, cold, and continuous light then each condition's culture was processed by total RNA extraction using TRIzol method [1]. Size selection of total RNA was reduced from total RNA to those RNA species, signifying sRNAs, of 15-35 nt via gel extraction [1]. A small RNA amplification procedure was used to first, attach 3' and 5' cloning linkers via ligation to the size-selected small RNA species then use reverse transcriptase for cDNA synthesis and amplification [1]. Illumina GAIIx library prep involved adding Illumina adaptors to size-selected total RNA [1].

Degradome library prep involved first, the extraction of mRNA from total RNA, of which 5'P of the cleaved mRNA was ligated to Illumina adaptor [1]. RNA-seq library prep involved taking cDNA from salt and normal stressed libraries converted from

mRNA then using *Nextera* Tagmentation Reaction protocol to fragment and tag, generate RNA-seq libraries appropriate for use with *Illumina GAIIx* sequencing platform [1].

Bioinformatic Pipeline I: Read profiling and trimming raw sequence data

Information outputting from the sequencer is multiplexed in many ways. Making sense of the information and deriving necessary conclusions involves a careful bioinformatics pipeline which aims to fundamentally: 1. demultiplex sequencing data and 2. summarize massive information using visuals, tables the ultimately bolster experimental conclusions. Filtering small RNAs involved: removing rRNA, tRNA, plastid RNA species, trimming *Illumina* adaptors then filtering in small RNAs of 16-32nt length which represent nuclear encoded sRNAs [1]. Filtering degradome tags involved an extra step of removing poly-A tails and no size-specific filtering was performed [1]. Three template sequence libraries of *C. paradoxa* exist: genomic contigs, EST contigs (nt >= 200) and and CDS. Mapping filtered sRNAs against these three sequence libraries with 100% identity produced the small RNA and degradome library data used for this analysis [1].

Bioinformatic Pipeline II: Expression analysis of small RNAs

Using CLC Genomics, reads were mapped to *C. paradoxa* contigs, ESTS and CDS. Expression analysis of reads mapped to CDS of different conditions [1].

Bioinformatic Pipeline III: Building a database of filtered and mapped putative small RNA sequences to a consensus

To understand with a big picture analysis of the mapped small RNA sequences, their spatial location to nucleotide-level clarity, a series of scripts were developed to process small RNA data using R and Bioconductor libraries. Data from degradome sequencing provides a means for empirically detecting various cleaved mRNA targets without predictions [1]. These data on target cleavages and their products provide insights on degradation pathways in *C. paradoxa*.

Bioconductor libraries utilized in my research and analysis include the following: seqinr for sequence extraction and proceeding exploratory data analysis, `Rsamtools` for importing and parsing BAM files, `GenomicRanges` for representation and manipulation of genomic intervals. Database manipulation utilized, `reshape` and `plyr` libraries while plotting of data explicitly utilized `lattice`. Outside of R, the following tools are utilized as a backbone of this bioinformatics study: CLC Genomics Workbench, GUI for analyzing and visualizing NGS data and StatET, an Eclipse based IDE for R. High-throughput transcriptome sequencing data is outputted to an aligned file such as SAM or BAM [41]. Sequence alignment/map (SAM) file, represent a generic alignment between reads against reference sequences and are converted to a compacted Binary Alignment/Map (BAM) format (S Table 1) [41].

In this program, CDS and degradome data in BAM format is imported through a readGappedAlignmentsobject. From here, all possible contig names (based on

Q/RNAME) are extracted along with their respective start, stop and strand information (based on POS, SIZE and FLAG; see Table S1). The new condensed database is stored as a data frame object, the two dimensional array data type in R. Further processing involves separation of all reads based on their sense and antisense directionalities. A frequency table of each read mapped to a start position per contig is constructed. Relative frequency normalization consists of taking read frequency counts divided by total frequency counts (per contig). Small RNA database is then collapsed into a nonredundant set, based on frequency counts that are 3-fold above the mean per contig.

RESULTS & DISCUSSION

Small RNA producing genomic hotspots are also exonic

According to Gross et. al (2012), reads mapped to the *Cyanophora paradoxa* datasets (CDS, EST, Genomic) are exonic [1]. For more on this, see Gross et. al (2012).

Bias for antisense small RNAs

In the CDS data, 31,895 contigs had at least one read mapped from a total pool of 1,762,675 reads where 1,394,966 (79.14%) were antisense and 367,709 (20.86%) were sense. Genomic contigs that map small RNAs constituted more than half of the database, indicating pool of small RNAs target parts of genomic more than others and may signify small RNA hotspots [1].

Degradome tags are biased to the 3' end of transcripts

Small RNA production is abundantly found within transcribed regions, where mapped regions comprised of over 70% ESTs while many small RNAs mapped to protein coding regions [1]. Similar to *Arabidopsis*, degradome tags were 3' biased (Figure 3) [1].

Cleavage tag abundance and noise reduction of dataset

As shown in Figure 2, each contig to which small RNAs were mapped showed peaks. To reduce the dataset and remove noise, Jef and I devised a plan to remove all peaks where were less than 3x the mean of all peaks in that contig then to all contigs (Figure 2).

FUTURE OBJECTIVES

Observed patterns in the Cyanophora paradoxa small RNA transcriptome network

Network biology is an emerging field that utilizes systems biology and bioinformatics. A network approach to understanding proteomics, transcriptomics offers in general a useful means of understanding interaction among nodes. Open to users, the mathematical libraries associated with graph theory and therefore its application to represent genomic data, networks offer a dynamic model that can be augmented and pruned. Visual networks bridge a gap to promote understanding of complex phenomena by biologists (Figure 6). One way of disseminating the small RNA pool hot spots is to generate an orthogonal data set such as genetic interaction.

CONTRIBUTIONS AND ACKNOWLEDGEMENTS

I joined the Bhattacharya Lab in 2012 and began to work on the small RNA project with my Postdoc adviser Dr. Jef Gross. I wrote the scripts utilized in this paper in R. A poster presentation discussing the results of this research was presented by me in 2012 at the Phycological Society of America (PSA) Annual Meeting, in Charleston, South Carolina.

I acknowledge that over the course of this project and others, the DB Lab has been an unparalleled learning experience in phylogenomics, bioinformatics and collaborative interdisciplinary research. I also am grateful for Professor Bhattacharya's mentorship and revision of this paper.

SUPPLEMENTARY TABLES

ST1. Fields in the SAM format [41]

Index	Name	Description
1	QNAME	Query NAME of the read or read pair
2	FLAG	Bitwise FLAG (pairing, strand, mate strand, etc.)
3	RNAME	Reference sequence NAME
4	POS	1-Based leftmost Position of the clipped alignment
5	MAPQ	Mappning quality (Phred-scaled)
6	CIGAR	Extended CIGAR string (operations: MIDNSHP)
7	MRNM	Mate Reference NaMe
8	MPOS	1-Based leftmost Mate Position
9	ISIZE	Inferred insert size
10	SEQ	Query Sequence on the same strand as the reference
11	QUAL	Query Quality

REFERENCES

1. Gross, J., Wajid, S., Price, D. C., Zelzion, E., Li, J., Chan, C. X., & Bhattacharya, D. (2013). Evidence for Widespread Exonic Small RNAs in the Glaucophyte Alga *Cyanophora paradoxa*. PloS one, 8(7), e67669.

2. Bhattacharya, D., Price, D. C., Chan, C. X., Gross, J., Steiner, J. M., & Löffelhardt, W. (2014). Analysis of the Genome of *Cyanophora paradoxa*: An Algal Model for Understanding Primary Endosymbiosis. In Endosymbiosis (pp. 135-148). Springer Vienna. Chicago

3. Reyes-Prieto, A., Weber, A. P., & Bhattacharya, D. (2007). The origin and establishment of the plastid in algae and plants. Annu. Rev. Genet., 41, 147-168.

4. Price, D. C., Chan, C. X., Yoon, H. S., Yang, E. C., Qiu, H., Weber, A. P., Schwacke, R. & Bhattacharya, D. (2012). *Cyanophora paradoxa* genome elucidates origin of photosynthesis in algae and plants. Science, 335(6070), 843-847.

5. Lee, R. E. (2008). Phycology. Cambridge University Press.

6. Seckbach, J. (Ed.). (2002). Symbiosis: mechanisms and model systems (Vol. 4). Springer.

7. Gross, J., & Bhattacharya, D. (2011). Endosymbiont or host: who drove mitochondrial and plastid evolution. Biol Direct, 6, 12.

8. Wise, R. R. (2006). The diversity of plastid form and function. In The structure and function of plastids (pp. 3-26). Springer Netherlands.

9. Facchinelli, F., Pribil, M., Oster, U., Ebert, N. J., Bhattacharya, D., Leister, D., & Weber, A. P. (2013). Proteomic analysis of the *Cyanophora paradoxa* muroplast provides clues on early events in plastid endosymbiosis. Planta, 237(2), 637-651.

10. Falkowski, P. G., Fenchel, T., & Delong, E. F. (2008). The microbial engines that drive Earth's biogeochemical cycles. Science, 320(5879), 1034-1039.

11. Kim, J. D., Senn, S., Harel, A., Jelen, B. I., & Falkowski, P. G. (2013). Discovering the electronic circuit diagram of life: structural relationships among transition metal binding sites in oxidoreductases. Philosophical Transactions of the Royal Society B: Biological Sciences, 368(1622).

12. Mathews, K., van Holde and Ahern. (2000) Biochemistry Third Edition. Benjamin Cummings.

13 Floener, L., Danneberg, G., & Bothe, H. (1982). Metabolic activities in *Cyanophora paradoxa* and its cyanelles. Planta, 156(1), 70-77.

14. Zvelebil, M. J., & Baum, J. O. (2008). Understanding bioinformatics. Garland Science.

15. Brown, T. A. (2006). Genomes 3. Wiley-Liss, Oxford.

16. Ghildiyal, M., & Zamore, P. D. (2009). Small silencing RNAs: an expanding universe. Nature Reviews Genetics, 10(2), 94-108.

17. Gao, Z., Wang, M., Blair, D., Zheng, Y., & Dou, Y. (2014). Phylogenetic Analysis of the Endoribonuclease Dicer Family. PloS one, 9(4), e95350.

18 - Zhang, X., Xia, J., Lii, Y. E., Barrera-Figueroa, B. E., Zhou, X., Gao, S., Lu, L., & Jin, H. (2012). Genome-wide analysis of plant nat-siRNAs reveals insights into their distribution, biogenesis and function. Genome Biol, 13(3), R20.

19. -

20. Shabalina, S. A., & Koonin, E. V. (2008). Origins and evolution of eukaryotic RNA interference. Trends in Ecology & Evolution, 23(10), 578-587.

21. Jackowiak, P., Nowacka, M., Strozycki, P. M., & Figlerowicz, M. (2011). RNA degradome—its biogenesis and functions. Nucleic acids research, 39(17), 7361-7370.

22. Cerutti, H., & Casas-Mollano, J. A. (2006). On the origin and functions of RNA-mediated silencing: from protists to man. Current genetics, 50(2), 81-99.

23. Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X. & Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. Nature methods, 6(5), 377-382.

24. Nakanishi, K., Weinberg, D. E., Bartel, D. P., & Patel, D. J. (2012). Structure of yeast Argonaute with guide RNA. Nature, 486(7403), 368-374.

25. Joshua-Tor, L., & Hannon, G. J. (2011). Ancestral roles of small RNAs: an Ago-centric perspective. Cold Spring Harbor perspectives in biology, 3(10), a003772.

26. Tuck, A. C., & Tollervey, D. (2011). RNA in pieces. Trends in genetics, 27(10), 422-432.

27. Poulsen, C., Vaucheret, H., & Brodersen, P. (2013). Lessons on RNA silencing mechanisms in plants from eukaryotic argonaute structures. The Plant Cell Online, 25(1), 22-37.

28. Kuhn, C. D., & Joshua-Tor, L. (2013). Eukaryotic Argonautes come into focus. Trends in biochemical sciences, 38(5), 263-271.

29. -

30. Perrineau, M. M., Zelzion, E., Gross, J., Price, D. C., Boyd, J., & Bhattacharya, D. (2014). Evolution of salt tolerance in a laboratory reared population of Chlamydomonas reinhardtii. Environmental microbiology.

31. Baldauf, S. L. (2003). Phylogeny for the faint of heart: a tutorial. TRENDS in Genetics, 19(6), 345-351.

32. Llave, C., Xie, Z., Kasschau, K. D., & Carrington, J. C. (2002). Cleavage of Scarecrow-like mRNA targets directed by a class of Arabidopsis miRNA. Science, 297(5589), 2053-2056.

33. Brodersen, P., Sakvarelidze-Achard, L., Bruun-Rasmussen, M., Dunoyer, P., Yamamoto, Y. Y., Sieburth, L., & Voinnet, O. (2008). Widespread translational inhibition by plant miRNAs and siRNAs. Science, 320(5880), 1185-1190.

34. Willmann, M. R., Berkowitz, N. D., & Gregory, B. D. (2013). Improved genome-wide mapping of uncapped and cleaved transcripts in eukaryotes—GMUCT 2.0. Methods.

35. Cech, T. R., & Steitz, J. A. (2014). The Noncoding RNA Revolution—Trashing Old Rules to Forge New Ones. Cell, 157(1), 77-94.

36. Addo-Quaye, C., Miller, W., & Axtell, M. J. (2009). CleaveLand: a pipeline for using degradome data to find cleaved small RNA targets. Bioinformatics, 25(1), 130-131.

37. Endres, M. W., Cook, R. T., & Gregory, B. D. (2011). A high-throughput sequencing-based methodology to identify all uncapped and cleaved RNA molecules in eukaryotic genomes. In MicroRNAs in Development (pp. 209-223). Humana Press.

38. Frohman, M. A. (1990). RACE: rapid amplification of cDNA ends. PCR protocols: A guide to methods and applications, 28.

40. Wu, A. R., Neff, N. F., Kalisky, T., Dalerba, P., Treutlein, B., Rothenberg, M. E., ... & Quake, S. R. (2014). Quantitative assessment of single-cell RNA-sequencing methods. Nature methods, 11(1), 41-46.

41. SAM Format Specication Working Group. The sam format specification (v1. 4-r985).

42. Perdomo, C., Campbell, J., & Schembri, F. (2014). Detecting Noncoding RNA Expression: From Arrays to Next-Generation Sequencing. In Non-coding RNAs and Cancer (pp. 25-44). Springer New York.

43. Addo-Quaye, C., Eshoo, T. W., Bartel, D. P., & Axtell, M. J. (2008). Endogenous siRNA and miRNA Targets Identified by Sequencing of the Arabidopsis Degradome. Current Biology, 18(10), 758-762.