CellPress

Mining diverse small RNA species in the deep transcriptome

Kasey C. Vickers[1], Leslie A. Roteta[1], Holli Hucheson-Dilks[2], Leng Han[3], and Yan Guo[4]

[1] Department of Medicine, Vanderbilt University School of Medicine, Nashville, TN, USA
[2] VANTAGE, Vanderbilt University School of Medicine, Nashville, TN, USA
[3] M.D. Anderson Cancer Center, Houston, TX, USA
[4] Department of Cancer Biology, Vanderbilt University School of Medicine, Nashville, TN, USA

Transcriptomes of many species are proving to be exquisitely diverse, and many investigators are now using high-throughput sequencing to quantify non-protein-coding RNAs, namely small RNAs (sRNA). Unfortunately, most studies are focused solely on microRNA changes, and many investigators are not analyzing the full compendium of sRNA species present in their large datasets. We provide here a rationale to include all types of sRNAs in sRNA sequencing analyses, which will aid in the discovery of their biological functions and physiological relevance.

The emergence of transcriptomic analyses

J. Craig Venter's human expressed sequence tag (EST) database, published in 1991, is considered to be the first human gene expression profiling study and was completed using automated Sanger sequencing methods, a significant advance at the time [1]. By 1995, serial analysis of gene expression (SAGE) was the state-of-the-art method for profiling gene (mRNA) expression; however, hybridization microarrays quickly became the popular choice and remained so until very recently [2]. It was during this time (mid-1990s) that the term 'transcriptome' first appeared, the first of many -omics, derived from the term genomics, that are now popular across science. After a decade of microarrays, sequencing-by-synthesis emerged and set forth the rapid development of DNAseq and RNAseq approaches, which coincided with the availability of short-read massive parallel sequencing platforms, later known as next-generation sequencing (NGS) [3]. Currently, many investigators are using RNAseq approaches to quantify long (e.g., mRNA) or sRNA expression; however, owing to specific barriers they are not fully analyzing the large amount of information provided by these approaches.

miRNA analysis

The study of non-coding sRNAs, particularly miRNAs (19–22 nt), has gained significant attention in recent years as 40% (12 971/32 879) of all miRNA publications in Pubmed have been published during the past 18 months (2013 through June 2014). Currently, there are over 35 000 annotated mature miRNAs in 223 species cataloged in miRBase (v21; http://mirbase.org), >2500 of which are human. Nonetheless, miRNAs are highly abundant and dominant in many non-mammalian species, and researchers from wide-fields of study are investigating miRNAs in yeast, worms, flies, plants, and many other species. Although there are multiple strategies to profile miRNAs, the current state of the art is sRNA-seq, and many investigators are now using this approach on a wide-variety of tissues and fluids. sRNAseq is a class of methods used to perform high-throughput sRNA sequencing on libraries of sRNAs ligated to terminal adapters for reverse transcription and amplification. Although miRNAs are only one of the many sRNA species in sRNAseq datasets, miRNAs remain the most popular class to study, largely because they can arise from autonomous transcriptional units, their processing steps are relatively understood, and the general mechanism for their biological functions is known. Unfortunately, many investigators neglect the copious amounts of non-miRNA sRNA species present in their datasets. A common barrier is often the lack of genomic annotations in alignment tools for non-miRNA species. Many investigators are unsure how to place altered expression values of non-miRNA sRNA species into biological contexts because their biological functions and physiological relevance are largely unknown. Moreover, many of these new sRNAs are not as widely conserved across many species as miRNAs.

Nevertheless, some groups are striving to resolve the functional impact of non-miRNA sRNAs, and there has been an explosion of novel sRNA species reported in literature. Recent advances in library preparation and NGS technologies enable platforms to now generate hundreds of gigabases per run, which allows for tremendous depth of sequencing into the sRNA transcriptome. This has facilitated the identification of many low-abundance species. Most interestingly, a wide-range of sRNA fragments derived from long RNA species has emerged (Figure 1, Table 1) [4]. These sRNAs are not likely to be the result of random degradation because their consistent alignments, specific terminal ends (evidence of RNase III cleavage), high read counts, and sequence characteristics suggest they are instead regulated cleavage products; however, there is bias in different RNAseq approaches, and protective RNA binding proteins may produce specific reads during normal RNA degradation and turnover [5]. Although we know very little about many of these novel sRNAs, they have great potential to regulate gene expression and biological processes similarly to miRNAs. As such,

Corresponding author: Vickers, K.C. (kasey.c.vickers@Vanderbilt.edu).

Keywords: small RNA; RNA sequencing; long-noncoding RNA; transcriptome; microRNA; high-throughput sequencing; data analysis.

0968-0004/

http://dx.doi.org/10.1016/j.tibs.2014.10.009

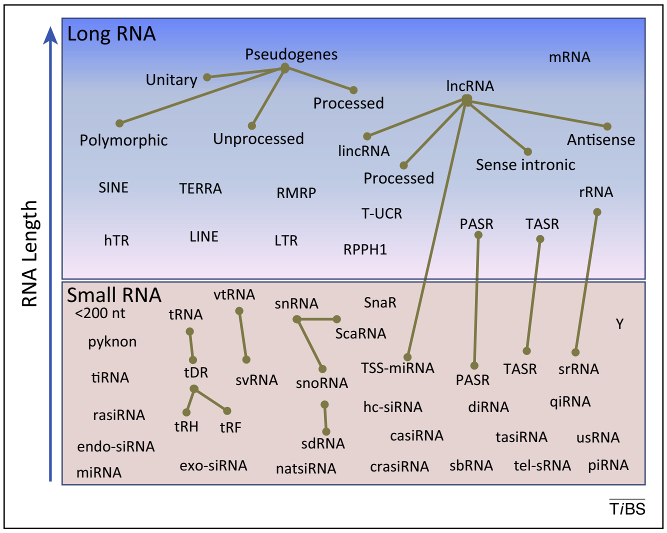

Figure 1. Schematic illustrating the diversity of long and small RNAs. The gold edge represents small RNAs (sRNA) derived from long parent RNA; sRNAs are RNAs ≤200 nt in length whereas long RNA are significantly bigger. **(Long RNAs)** hTR, human telomerase RNA; lincRNA, large intergenic non-coding RNA; LINE, long interspersed element; lncRNA, long non-coding RNA; LTR, long terminal repeat; mRNA; PASR, promoter-associated long RNA; RMPR, RNA component of mitochondrial RNA processing endoribonuclease RPPH1, ribonuclease P RNA component H1; rRNA, ribosomal RNA; SINE, short interspersed element; TASR, termini-associated sRNA; TERRA, telomeric repeat-containing RNA; T-UCR, transcribed ultraconserved region. **(Small RNAs)** casiRNA, *cis*-acting siRNA; crasiRNA, centromere repeat-associated sRNA; diRNA, double-strand break-induced sRNA; endo-siRNA, endogenous small interfering RNA; exo-siRNA, exogeneous small interfering RNA; hc-siRNA, heterochromatic small interfering RNA; miRNA, microRNA; natsiRNA, natural antisense siRNA; piRNA, Piwi-interacting RNA; qiRNA, QDE-2-interacting sRNA; rasiRNA, repeat-associated siRNA; ScaRNA, small Cajal-body RNA; sbRNA, stem-bulge RNA; sdRNA, snoRNA-derived small RNA; SNAR, small NF90-associated RNA; snoRNA, small nucleolar RNA; snRNA, small nuclear RNA; srRNA, sRNAs-derived from rRNA; svRNA, small vault RNA; tasiRNA, *trans*-acting siRNA; tDR, tRNA-derived sRNA; tel-sRNA, telomere-specific sRNA; tiRNA, transcription initiation sRNA; tRH, tRNA-derived halves; tRF, tRNA-derived fragments; TSS-miRNA, transcriptional start-site-microRNA; usRNA, unusually small RNA; vtRNA, vault RNA; Y, Y RNA.

there is a great need to study these sRNAs and the protein binding factors that mediate their biological functions.

miRNAs are only the tip of the iceberg: sRNA diversity

Many sRNAs are named to recognize the parent RNA species from which they were derived. For example, one emerging class that is gaining great interest, the tRNA-derived sRNAs (tDRs), comprise at least four subtypes. tRNA-derived sRNA fragments (tRFs, approximately 20 nt) and tRNA-derived halves (tRHs, approximately 33 nt) are two distinct subclasses with likely different biological functions. Although the physiological roles of tRFs and tRHs are only beginning to be defined, it is clear that they respond to cell stress and their regulatory cleaving enzymes have been elucidated. One subclass of tRFs (3′CCA tRFs) have been reported to act like miRNAs and silence complementary targets [6]. tRHs are generated by angiogenin-mediated cleavage near the anticodon loop in response to starvation, oxidative stress, or other forms of cellular stress. Through a variety of mechanisms, tRHs suppress protein translation [7]. tRHs have also been reported to bind to eukaryotic translation initiation factor 4γ (eIF4G) through a distinct sequence motif on the tRH 5′-terminal end and titrate eIF4G away from the pre-initiation complex [7]. Although the exact mechanisms of 5′ tRH suppression of translation are still emerging, it has been shown that, at least for reporter constructs, they inhibit translation independently of seed-sequence complementarity [8]. tDRs have been reported in multiple cell types and diseases. tRHs are also highly abundant in plasma, where they are found associated with protein complexes, for example, high-density lipoproteins (HDL), but are largely excluded from exosomes [9]. Owing to their robust responses to a wide variety of stresses and their potential value as extracellular RNA biomarkers, the study of tDRs is a rapidly growing area of research. Some sRNAs are processed from long non-coding RNAs, including some interesting species derived from pseudogenes, miscellaneous transcripts, and even coding mRNAs (Table 1, Figure 1). Other sRNAs are processed from various parent transcripts (Table 1), including promoter-associated sRNAs (PASR), transcription initiation sRNAs (tiRNAs), unusually small sRNAs (usRNA), TSS-miRNAs, sno-derived RNA (sdRNA), small vault RNAs (svRNA), and sRNAs-derived from rRNA (srRNA) [10,11]. Other examples of non-coding smRNAs which may be present in a sRNAseq datasets. depending on eukaryotic phylogeny or cell type, include endo-siRNAs, exo-siRNAs, double-stranded RNAs (dsRNA), Piwi-interacting RNAs (piRNA), natural antisense siRNAs (natsiRNA), *cis*-acting siRNAs (casiRNA), *trans*-acting siRNAs (tasiRNA), repeat-associated siRNAs (rasiRNA), centromere repeat-associated sRNAs (crasiRNA), and telomere-specific sRNAs (tel-sRNAs) [12]. Many of these sRNAs have been reported to associate with argonaute-RNA-induced silencing complexes [AGO(1-4)-RISC], and thus probably post-transcriptionally regulate gene expression through partial complementary binding to target mRNAs [13]. sRNAseq datasets may also contain reads representing Y RNAs, double-strand break-induced sRNAs (diRNA), stem bulge RNAs (sbRNA), and endo-siRNA-like sRNAs induced by DNA damage and originating from ribosomal (r)DNA regions (QDE-2-interacting sRNAs, qiRNA) (Table 1) [14]. Moreover, many other novel sRNA species probably remain to be discovered. This is particularly evident because many reads in sRNA datasets do align to the human genome, but to unannotated loci. Although this highlights the problem with incomplete genomic annotations and databases, it also alludes to potential discovery in these datasets. To identify and count sRNAs, investigators can use annotated genomic coordinates to align and count reads, or use a non-genome mapping strategy and simply align reads to canonical long and small RNA sequences [15]. Nevertheless, the crucial barrier to studying non-miRNA sRNAs is the general lack of proper annotations and genomic coordinates (sequence information) for non-coding RNA. Likewise, the field is lacking well-maintained and easily accessible databases for non-miRNA sRNAs. For example, tDR annotations have been problematic for various reasons, including mapping to multiple loci and the organization of tDRs into classes and families. To advance the field, a significant effort is required to further develop and curate sRNA databases to be freely distributed. Nonetheless, a few databases are useful, including the Table Browser from the UCSC Genome Bioinformatics Site (http://www.genome.ucsc.edu) and Ensemble (http://www.ensembl.org) gtf