Published in final edited form as:

Methods Mol Biol. 2012 ; 822: 219–232. doi:10.1007/978-1-61779-427-8_15.

Profiling of Short RNAs Using Helicos Single-Molecule Sequencing

Philipp Kapranov, Fatih Ozsolak, and Patrice M. Milos

Abstract

The importance of short (<200 nt) RNAs in cell biogenesis has been well documented. These short RNAs include crucial classes of molecules such as transfer RNAs, small nuclear RNA, microRNAs, and many others (reviewed in Storz et al., *Annu Rev Biochem* 74:199–217, 2005; Ghildiyal and Zamore, *Nat Rev Genet* 10:94–108, 2009). Furthermore, the realm of functional RNAs that fall within this size range is growing to include less well-characterized RNAs such as short RNAs found at the promoters and 3' termini of genes (Affymetrix ENCODE Transcriptome Project et al., *Nature* 457:1028–1032, 2009; Davis and Ares, *Proc Natl Acad Sci USA* 103:3262–3267, 2006; Kapranov et al., *Science* 316:1484–1488, 2007; Taft et al., *Nat Genet* 41:572–578, 2009; Kapranov et al., *Nature* 466:642–646, 2010), short RNAs involved in paramutation (Rassoulzadegan et al., *Nature* 441:469–474, 2006), and others (reviewed in Kawaji and Hayashizaki, *PLoS Genet* 4:e22, 2008). Discovery and accurate quantification of these RNA molecules, less than 200 bases in size, is thus an important and also challenging aspect of understanding the full repertoire of cellular and extracellular RNAs. Here, we describe the strategies and procedures we developed to profile short RNA species using single-molecule sequencing (SMS) and the advantages SMS offers.

Keywords

Single-molecule sequencing; Short RNAs; Promoter-associate short RNAs; Polyadenylated short RNAs

1. Introduction

In a cell, a final functional product of a precursor RNA species is often a shorter RNA derived from it. The most obvious examples of this event include splicing and 3' end RNA processing to generate mature long mRNAs. However, a large number of other classes of functional RNAs that are not protein coding are made from longer precursors. Many of such RNAs are in the realm of less than 200 nt, even though this range is somewhat arbitrary and simply based on methods for fractionation. Most commonly biochemical column-based methods are used to fractionate RNA into species that are shorter or longer than 200 nt. Such RNAs are represented by tRNAs, small nuclear (sn)RNAs, small nucleolar (sno) RNAs, micro (mi)RNAs, and others (1–9). The latter miRNA class is probably one of the extreme examples in which very long (kbs) precursor RNA is cleaved by two enzymatic steps into a final functional product of 21–23 nt (2). Therefore, a full understanding of the cellular repertoire and functional cellular products must include the study of complex, short RNA (sRNA) population that cannot be achieved by profiling only long RNAs (3–9). In addition, extracellular sRNAs appear to be a promising class of biomarker molecules (10, 11). The

following chapter describes the methods to enable successful profiling of this important class of short RNAs.

2. Materials

2.1. RNA Isolation

1. Purification of sRNA from total RNA or cultured cells could be achieved using these kits:

*mirVana*TM miRNA Isolation Kit (Ambion).

miRNeasy Mini Kit (Qiagen).

RNA/DNA kit (Qiagen) – suitable for preparation of large quantities of sRNAs from cultured cells.

2. Alternatively, sRNA of a desired fraction could be purified using TBE-Urea polyacrylamide gel electrophoresis and overnight elution (12).

2.2. cDNA Synthesis

2.2.1. General Method

1. *Escherichia coli* PolyA polymerase (Ambion).
2. 100 mM CTP (Roche).
3. ThermoScript reverse transcriptase (15 U/μL Invitrogen) – also, (see Note 5).
4. Phenol/chloroform/isoamyl alcohol (Ambion).
5. 5 M Ammonium acetate (Ambion).
6. cDNA synthesis primer – custom made from Integrated DNA Technologies sequence: TCG CGA GCG GCC GCG GGG GGG GGG GGrG rGrG. Important – last three bases are ribonucleotides.
7. RNase A (Ambion).

2.2.2. Method for Profiling 3' polyA sRNAs

1. SuperScript III reverse transcriptase (Invitrogen).
2. USER enzyme (New England Biolabs).
3. dTU-V cDNA synthesis primer – custom made from Integrated DNA Technologies sequence: TTTTUTTUTUTTTUTTTTUTTTTUTTV.
4. RNase H (Invitrogen).
5. RNase 1f (New England Laboratories).

2.2.3. Reagents Common to Both Methods

1. 100 and 70% ethanol (Sigma).
2. RNase inhibitors: ANTI-RNase (Ambion) or RNaseOUT (Invitrogen).
3. 10 mM dNTPs (Invitrogen).

⁵We have not tested different reverse transcriptases and the ones listed in this report may or may not be optimal for the detection of certain RNA species.

4. AMPure® beads (Agencourt).
5. Magnetic stand for 1.5-mL tubes.
6. PCR machine. The protocol below has been tested on the BioRad Tetrad 2 Thermal Cycler.

2.3. Sequencing of cDNA

1. 20 U/μL Terminal Transferase (New England Biolabs).
2. dATP (Helicos BioSciences).
3. 1 mM Biotin-ddATP (Perkin Elmer).
4. 10 mg/mL Bovine serum albumin (BSA) (New England Biolabs).
5. Quant-iT™ OliGreen® ssDNA Reagent (Invitrogen).
6. NanoDrop 3300 (Thermo Scientific).
7. HeliScope™ Single Molecule Sequencer (Helicos BioSciences Corporation).
8. Helicos® Flow Cells (Helicos BioSciences Corporation).

2.4. Data Analysis

1. A suite of unix-based Helicos processing tools that could be freely downloaded from here: <http://open.helicosbio.com/mwiki/index.php/Releases>.
2. Computer hardware. In general, it is suggested to have at least 5 GB per CPU core for alignments to the human genome using the Helicos aligner indexDPgenomic (13).
 - a. A cluster, desirable if processing of multiple channels is required. An example of specification would be a combination of dual and quad-core 3 GHz CPUs (i.e., Dell 1950) with memory ranging from 8 to 32 GB.
 - b. A standalone unix system if processing of only a few channels is required with a similar specs to the ones listed above.

3. Methods

Profiling of short RNAs has several unique challenges that the researcher should consider. First, the sRNA fraction of the desired length range must be isolated, otherwise, the signal could be derived from a long RNA that overlaps a short RNA. Second, most (not all) of the sRNAs of interest lack an easy molecular handle such as the 3' polyA tail of the mRNAs that is needed for conversion into cDNAs. Third, they are often too short for efficient conversion into cDNA using random hexamers. Fourth, certain sRNAs have modifications at their 5' (and 3') ends that interfere with subsequent molecular manipulation and thus can go undetected by certain methods. Fifth, some classes of sRNAs have strong secondary structures (see Note 4) that preclude them from being detected by enzymatic methods that are typically conducted under mild (nondenaturing) conditions. Sixth, some sRNAs like miRNAs have lengths that are too short for efficient mapping to very complex genomes, thus making discovery work challenging. Seventh, most of the methods used rely on ligation and PCR amplification that can skew both the composition of the population and the quantification of the RNA species. Eighth, sRNA fraction could be dominated by relatively

⁴We typically do not detect the pre-miRNAs species (the products of the Drosha cleavage) most likely due to their very stable secondary structure.

few, very highly abundant RNA classes, such as rRNA, snRNA, and snoRNAs which will require significant depth for complete characterization of the sRNA population. This can be avoided by further selections such as selection of a specific size range enriched in miRNAs (~18–25 bases) or selection of those with 3' polyA tails (Subheading 3.3).

Below, we provide two general methods for the detection of sRNAs using single molecule sequencing: one can be used for detecting any sRNA that has a 3' OH and the second for detecting 3' polyA sRNAs. To circumvent some of the challenges presented above, we start with an isolated sRNA fraction followed by (1) the addition of 3' polyC tail using polyA polymerase and cDNA synthesis using a polyG-containing oligonucleotide or (2) cDNA synthesis using a polyU oligonucleotide without tailing of RNA. The resulting cDNA could then be purified and sequenced directly after polyA tailing and 3' blocking. The diagram of the first method is shown in Fig. 1.

Both methods do not require ligation or amplification and are also ambiguous as to the status of the 5' end of an sRNA – thus sRNAs with any modification at their 5' end can be detected. We also show the results one could expect when profiling a population of sRNAs less than 200 nt in the human HeLaS3 cells and pros and cons of each method.

3.1. Isolation of sRNA Fraction

1. The sRNA fraction could be isolated using a variety of methods (see Notes 1 and 2). If a general <200 nt fraction is required, commercially available kits like mirVana (Ambion); miRNeasy, or RNA/DNA from Qiagen could be used. We have used these kits following the manufacturer's guidelines with satisfactory outcomes.
2. If it is desired to isolate sRNAs in a specific size range, a TBE-Urea denaturing polyacrylamide gel-electrophoresis could be considered as an alternative purification method. We have followed the protocols described in (12).

3.2. General Method for Profiling of sRNAs

3.2.1. Tailing of RNA with 3' polyC

1. Bring RNA to 30 μ L in water in a PCR tube: 10 ng to 5 μ g of short RNAs can be used with this protocol.
2. Incubate for 2 min at 85°C in a PCR machine and then put on ice for at least 2 min.
3. Add the following reagents: 10 μ L of 5 $\times$ *E. coli* PolyA polymerase buffer; 5 μ L of 25 mM MnCl₂; 1 μ L of 100 mM CTP; 1 μ L of Anti-RNase or RNaseOUT, and 3 μ L of 2 U/ μ L *E. coli* PolyA polymerase. Mix well (do not vortex) and incubate for 3 h at 37°C.
4. Add 40 μ L of water and 10 μ L of 5 M ammonium acetate.
5. Extract twice with phenol/chloroform/isoamyl by vortexing vigorously for 30 s.

¹The RNA should be free from genomic DNA contamination and, therefore, DNase I treatment is recommended. The RNA should be pure before the DNase I digestion so that the enzymatic step is not inhibited. Thus, an ethanol precipitation step before the DNase I treatment is recommended. Otherwise, incomplete DNase I digestion could create oligonucleotides that are themselves a better substrate for sequencing than the genomic DNA.

²Isolation of sRNA fraction is an important step. Care should be taken to remove the long RNA fraction otherwise the signal from the sRNA fraction could be contaminated by the reads coming from the overlapping long RNA. The sRNA fraction should be checked on the BioAnalyzer to ensure absence of the long RNA fraction. See Fig. 2, for example, of pure [part (a)] small RNA fraction and the one contaminated with long RNAs [part (b)].

6. Precipitate by adding 3× volumes of 100% EtOH for a minimum of 30 min at -80°C .
7. Centrifuge at 4°C for 30 min at top speed. Wash once with 70% EtOH. Vacuum-dry and resuspend in 30.5 μL water.

3.2.2. cDNA Synthesis

1. Add 1 μL of 100 μM cDNA synthesis primer: TCG CGA GCG GCC GCG GGG GGG GGG GGrG rGrG (*the last three bases – RNA*) to the 30.5 μL of RNA from the step above and mix. The presence of three RNA bases at the 3' end is to ensure that the oligo is not tailed by terminal transferase, as TdT does not utilize RNA as a substrate.
2. Incubate for 2 min at 70°C in a PCR machine: fast ramp to 4°C and incubate for 5 min at 4°C .
3. While the samples are at 4°C , add the following: 10 μL of 5× ThermoScript cDNA Synthesis buffer; 5 μL of 0.1 M DTT; 2.5 μL of 10 mM dNTPs, and 1 μL of ThermoScript reverse transcriptase. If you need to scale up the protocol to 3–5 μg of RNA, you could use more ThermoScript at the cDNA synthesis step: $\sim 1 \mu\text{L}/1 \mu\text{g}$ of RNA.
4. Slow Ramp (30 min) to 60°C followed by 1.5-h incubation. The following conditions were used on the BioRad cyclor: $0.1^{\circ}\text{C}/\text{s}$ ramp from 4 to 42°C , incubate for 15 min followed by $0.1^{\circ}\text{C}/\text{s}$ ramp from 42 to 50°C , incubate for 15 min and followed by $0.1^{\circ}\text{C}/\text{s}$ ramp from 4 to 60°C and incubation at 60°C for 90 min.
5. Fast ramp to 75°C followed by incubation for 15 min to inactivate the reverse transcriptase.

3.2.3. Purification of cDNA

1. Treat the cDNA synthesis reaction with 1 μL of RNase A for 30 min at 37°C .
2. Mix the AMPure beads suspension really well to ensure that the magnetic beads are resuspended. Add 150 μL of the suspended AMPure beads and incubate for 30 min with mixing at room temperature.
3. Capture the beads using the magnetic stand for 5 min.
4. Carefully remove and discard the supernatant.
5. Wash twice with 200 μL of 70% EtOH.
6. Dry the pellet for 30–45 min at room temperature or 30 min at 37°C .
7. Elute cDNA twice with 20 μL of water (pipet up and down at least 20 times) (see Note 3).

3.3. Profiling of sRNAs with polyA Tails

3.3.1. cDNA Synthesis

1. Combine 1 μL of 50 μM dTU-V primer with (1–5 μg) of small RNA and 1 μL of 10 mM dNTPs and water if required in a total volume of 10 μL . Mix and incubate at 65°C for 5 min in a thermocycler, followed by rapid cooling on a prechilled

³It is hard to estimate the yield of cDNA after purification since the oligonucleotide used to prime cDNA synthesis will co-purify with the small RNA cDNAs. We typically tail ~ 200 ng of cDNA assuming that $\sim 50\%$ of that is the oligonucleotide. Due to the presence of three ribonucleotides at its 3' terminus, the oligonucleotide will not be tailed by TdT.

aluminum block kept in an ice and water slurry (~0°C). Let the samples sit on the cold aluminum block for 1 min before proceeding with the next step.

2. Add to the above reaction 2 μ L of 10 $\times$ SuperScript III reaction buffer, 4 μ L of 25 mM MgCl₂, 2 μ L of 0.1 M DTT, 1 μ L of SuperScript III, and 1 μ L of RNaseOut while keeping the samples on the cold aluminum block. Mix the solution well after adding the reagents. Total volume is now 20 μ L.
3. Incubate the 20- μ L cDNA synthesis mix with the following temperature and time conditions: fast ramp to 50°C, incubate for 50 min followed by fast ramp to 85°C, incubate for 5 min and keep the samples at 4°C.
4. To remove the dTU-V primer sequences, add 1 μ L USER enzyme to the reaction, mix and incubate at 37°C for 15 min. Total volume should now be 21 μ L.
5. To digest away the RNA, add 1 μ L of RNase H and 1 μ L of RNase If, mix and incubate for 20 min at 37°C. The total volume is now 23 μ L.

3.3.2. Purification of cDNA

1. While USER/RNase treatment is in progress, warm up AMPure beads (180 μ L per sample) to room temperature.
2. After the USER/RNase treatment step, add 180 μ L of warmed-up AMPure beads to the 23 μ L cDNA reaction. Incubate with shaking at room temperature for 40 min to 1 h to bind cDNA to the beads.
3. Following the binding step, capture the beads using the magnetic stand for 5 min. Remove the supernatant with a pipettor. Wash the beads twice with 500 μ L of 70% EtOH (no need to resuspend the beads in ethanol during the ethanol washes).
4. After the 70% ethanol washes, dry the pellet for 45 min at RT or 30 min at 37°C in a clean warm room or oven. At the end of the drying step, ensure that there is no liquid is visible in the tube (the pellet often assumes a cracked appearance).
5. To elute the cDNA from the beads, add 20 μ L of nuclease-free water, and pipet up and down at least 10–20 times. Keep the sample at room temperature or 37°C for 5 min.
6. Place the tube on the magnet for 5 min and collect the eluate (~18 μ L).
7. Repeat the elution step with 20 μ L nuclease-free water and incubate at room temperature or 37°C for 5 min. Place the tube on the magnet for 5 min, collect the eluate again (~19 μ L), and combine with the first eluate (step 6 above). Total volume of the cDNA is now ~37 μ L.

3.4. Preparation of cDNA for Sequencing

The cDNA has to be tailed at the 3' end with polyA residues using terminal transferase (TdT) and blocked at the 3' end so that it can bind to the oligo-dT present at the surface of a flow cell (14). This step is common for the general protocol and for the polyA protocol. Amounts of cDNA are quantified using either the regular NanoDrop if the expected concentration is at least 5–10 ng/ μ L or Quant-iTTM OliGreen[®] ssDNA kit and NanoDrop 3300 (Thermo Scientific). Two major differences are dictated by the amounts of cDNA.

3.4.1. Tailing of Small Amounts of cDNA—If only small amounts (<10 ng) of cDNA are available or expected, use the following:

1. Prepare cDNA to be tailed (<10 ng) in 10.8 μ L of water. If cDNA is in larger volume – dry it down using speed-vac. Add 2 μ L of 10 $\times$ TdT buffer and 2 μ L of 2.5 mM CoCl₂. Incubate at 95°C for 5 min in a thermocycler for denaturation, followed by rapid cooling on a prechilled aluminum block kept in an ice and water slurry (~0°C).
2. Add the following: 4 μ L of 50 μ M dATP; 0.2 μ L of BSA, and 1 μ L of TdT diluted 1:4 (final concentration 5 U/ μ L) in 1 $\times$ TdT buffer. Final volume should be 20 μ L.
3. Mix well and incubate in a PCR machined at 37°C for 60 min followed by 70°C for 10 min and 4°C forever (or ice).
4. Heat to 95°C for 5 min, transfer on ice for a minimum of 2 min.
5. Add the following: 1 μ L of 10 $\times$ TdT buffer; 1 μ L of 2.5 mM CoCl₂; 0.5 μ L of 200 μ M Biotin-ddATP; 6.5 μ L of water, and 1 μ L of TdT diluted 1:4 (final concentration 5 U/ μ L) in 1 $\times$ TdT buffer. Final volume should be 30 μ L.
6. Mix well and incubate in a PCR machined at 37°C for 30 min followed by 70°C for 20 min and 4°C forever (or ice). Use directly for sequencing or plate assay to measure the concentration of tailed material (Subheading 3.5).

3.4.2. Tailing of Regular Amounts of cDNA—If larger ($\geq$ 50 ng) of cDNA are available, use the following:

1. Prepare cDNA to be tailed (50–200 ng or 2–3 pmol, see Note 3) in 33.8 μ L of water. Incubate at 95°C for 5 min in a thermocycler for denaturation, followed by rapid cooling on a prechilled aluminum block kept in an ice and water slurry (~0°C).
2. Add the following: 5 μ L of 10 $\times$ TdT buffer; 5 μ L of 2.5 mM CoCl₂; 5 μ L of 50 μ M dATP, and 1.2 μ L of TdT. Final volume should be 50 μ L.
3. Mix well and incubate in a PCR machined at 42°C for 60 min followed by 70°C for 10 min and 4°C forever (or ice).
4. Add 0.6 μ L of 1 mM biotin-ddATP to the tailing reaction from above.
5. Heat to 95°C for 5 min, transfer on ice for a minimum of 2 min.
6. Add 1.2 μ L of TdT.
7. Mix well and incubate in a PCR machined at 37°C for 30 min followed by 70°C for 10 min and 4°C forever (or ice). Use directly for sequencing or plate assay to measure the concentration of tailed material (Subheading 3.5).

3.5. Estimation of Concentration of polyA Tailed cDNA

Helicos scientists have developed the OptiHybTM Assay to determine the concentration of polyA-tailed templates and to allow the loading of the samples at optimal loading densities for single molecule sequencing on the Helicos[®] Genetic Analysis System. The protocols and reagents for this assay can be obtained from Helicos BioSciences Corporation (<http://www.helicosbio.com>).

3.6. Mapping of Reads to the Genome or Sequences of Known sRNAs

The sequence information obtained from the HeliScope Sequencer could be used for two general purposes: counting abundances of known sRNAs (digital gene expression) and discovery of new sRNAs. We have found that with an error rate of 3–5% that at present accompanies single-molecule sequencing, a read has to be at least 25 bases so that it can be

reliably aligned to a genome as complex as human even though shorter reads with this error rate could be aligned to less complex genomes. Thus, discovery of novel miRNAs via mapping of *individual* reads to the genome may not be possible with a current state of the technology. However, one can envision that algorithms that perform clustering of reads prior to alignments could alleviate this problem. Also, the decrease in the error rate of SMS that will inevitably happen as the technology matures will likely alleviate this problem as well. However, currently digital gene expression of known sRNAs by aligning reads only to the sequences of the sRNA whose expression is desired to be quantified rather than to the whole genome may be more appropriate because accurate alignments to a smaller reference set would be possible with a smaller length of reads – 18 or 20 bases. Below we describe different steps for the two analytical approaches. The various unix-based programs and pipelines for the analysis of the HeliScope data including the ones referenced below could be downloaded freely here: <http://open.helicosbio.com/mwiki/index.php/Releases>. Their descriptions could be found here: http://open.helicosbio.com/helisphere_user_guide/.

3.6.1. Filtering of the Reads

1. Output from a single-molecule sequencing experiment using the Helicos platform is represented by a distribution of the read lengths, typically ranging from 6 to 70 bases (14, 15). The reads have to be filtered to remove (Fig. 1) (1) very short reads; (2) to trim the 5' polyT sequence that corresponds to the polyA-tail added by TdT to the 3' end of cDNA; (3) to trim the sequence that corresponds to trailing polyC-tail added by the polyA-polymerase to the 3' end of a nascent RNA prior to the cDNA synthesis in the general method or the sequence that corresponds to 3' polyA if the method for profiling of sRNAs with 3' polyA tails was used; (4) artifactual reads. The filtering is done by a program called “filterSMS” using different parameters.
2. Depending on the goals of the experiments and the genomes that the sequences are to be aligned to later, one can change to filter the sequences based on different minimal lengths. The standard default minimal length is 25 bases, however, one may choose to use less stringent minimal lengths such as 20 bases, for example, if one is interested in detecting miRNAs and/or if the alignments are to be done against smaller references such as sequences of known sRNAs or *E.coli* genome.
3. Filtering of preceding and/or trailing homopolymeric tails (e.g., polyT or polyC) could be done with different parameters, such as specifying the minimal homopolymeric content of either tail. The default parameter is 75%.
4. Removing artifactual reads is done by filtering out reads that have repetition of a base-addition order sequence (CTAG); removing reads with high AT content and reads having high fraction of repeats of certain dinucleotides.
5. Usage of any of these parameters is optional, for example, filtering only by read length could be employed if the entire raw sequence is desirable as an output.

3.6.2. Sequence Alignment

1. The unique property of single molecule sequencing is that the error profile is dominated by indels rather the substitutions as in other sequencing platforms (14). Thus, an aligner that is tolerant to these types of errors should be employed. We recommend using the indexDPgenomic aligner (13) developed at Helicos and freely available for download (<http://open.helicosbio.com/mwiki/index.php/Releases>). Other aligners such as Mosaik (<http://bioinformatics.bc.edu/marthlab/Mosaik>), BWA (16), and Shrimp (17) could be used as well.

2. One of the basic parameters used by the indexDPgenomic is the normalized score computed as in the example below:

Tag sequence CCTCCGTGTTGTTCCAGCC-CAGTGCTCGCAGG

Ref sequence C-TCCGTGTTGTTCCAGCCACAGTGCTCGCAGG

Length of alignment block: 33

Length of tag sequence: 32

Number of matches: 31

Number of errors: 2

Score: $(31 \times 5) - (2 \times 4) = 155 - 8 = 147$

Normalized score = $147/32 = 4.59375$

3. Output of the indexDPgenomic is a binary alignment file that could be further filtered using a filterAlign program to select only, for example, for reads that could only be aligned to the genome once (uniquely aligning reads), reads that align with a certain minimal normalized score, etc. The binary alignment files could also be converted into the various text formats, including the standard BED or SAM/BAM formats using the following programs: printAlignmentFile; align2txt or align2sam.
4. Filtering of the reads, sequence alignment, and filtering of the alignment files could be combined by running the basic pipeline. Various specifications for the filterSMS, indexDPgenomic, and filterAlign could be incorporated into a configuration file that could then be invoked by the basic pipeline.
5. Expected results based on mapping of the reads obtained using both the general protocol and the protocol for profiling of 3' polyA sRNA can be seen in Table 1. Note that the 3' poly protocol is heavily enriched in novel sRNAs, specifically in PASRs and TASRs.

3.6.3. Digital Gene Expression

1. This analysis requires the same tools as above. Just like in the example of the basic pipeline above the DGE pipeline, combines the various processing steps and, as one of its outputs generates a text *.count.txt file that contains counts for each reference.
2. There are three different types of counts generated for each reference in the “Min,” “Frac,” or “RMC” columns. The difference is in how they use the nonuniquely mapping reads. The “Min” column lists only counts based only on the reads that align uniquely to each reference. The “Frac” and “RMC,” on the other hand, use the nonunique reads as well. The “Frac” evenly splits the nonunique reads among the different references they align to, while “RMC” used Bayesian-based approach to assign nonunique reads to different mapping positions (15).

Acknowledgments

We wish to thank Sharon Bleakney for the help with manuscript editing.

References

1. Storz G, Altuvia S, Wassarman KM. An abundance of RNA regulators. Annu Rev Biochem. 2005; 74:199–217. [PubMed: 15952886]

2. Ghildiyal M, Zamore PD. Small silencing RNAs: an expanding universe. *Nat Rev Genet.* 2009; 10:94–108. [PubMed: 19148191]
3. Affymetrix ENCODE, Transcriptome Project, Cold Spring, et al. Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. *Nature.* 2009; 457:1028–32. [PubMed: 19169241]
4. Davis CA, Ares M Jr. Accumulation of unstable promoter-associated transcripts upon loss of the nuclear exosome subunit Rrp6p in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci USA.* 2006; 103:3262–7. [PubMed: 16484372]
5. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, et al. RNA maps reveal new RNA classes and a possible function for pervasive. transcription. *Science.* 2007; 316:1484–8. [PubMed: 17510325]
6. Taft RJ, Glazov EA, Cloonan N, Simons C, Stephen S, Faulkner GJ, et al. Tiny RNAs associated with transcription start sites in animals. *Nat Genet.* 2009; 41:572–8. [PubMed: 19377478]
7. Kapranov P, Ozsolak F, Kim SW, Foissac F, Lipson D, Hart C, et al. Novel class of human RNAs associated with gene termini suggests an uncharacterized RNA copying mechanism. *Nature.* 2010; 466:642–6. [PubMed: 20671709]
8. Rassoulzadegan M, Grandjean V, Gounon P, Vincent S, Gillot I, Cuzin F. RNA-mediated non-mendelian inheritance of an epigenetic change in the mouse. *Nature.* 2006; 441:469–74. [PubMed: 16724059]
9. Kawaji H, Hayashizaki Y. Exploration of small RNAs. *PLoS Genet.* 2008; 4:e22. [PubMed: 18225959]
10. Mitchell PS, Parkin RK, Kroh EM, Fritz BR, Wyman SK, Pogosova-Agadjanyan EL, et al. Circulating microRNAs as stable blood-based markers for cancer detection. *Proc Natl Acad Sci USA.* 2008; 105:10513–8. [PubMed: 18663219]
11. Chen X, Ba Y, Ma L, Cai X, Yin Y, Wang K, et al. Characterization of microR-NAs in serum: a novel class of biomarkers for diagnosis of cancer and other diseases. *Cell Res.* 2008; 18:997–1006. [PubMed: 18766170]
12. Sambrook, J.; Russell, DW. *Molecular cloning: a laboratory manual.* Cold Spring Harbor Laboratory Press; Cold Spring Harbor, N.Y.: 2001.
13. Giladi E, Healy J, Myers G, Hart C, Kapranov P, Lipson D, et al. Error tolerant indexing and alignment of short reads with covering template families. *J Comput Biol.* 2010; 17:1397–411. [PubMed: 20937014]
14. Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Braslavsky I, et al. Single-molecule DNA sequencing of a viral genome. *Science.* 2008; 320:106–9. [PubMed: 18388294]
15. Lipson D, Raz T, Kieu A, Jones DR, Giladi E, Thayer E, et al. Quantification of the yeast transcriptome by single-molecule sequencing. *Nat Biotechnol.* 2009; 27:652–8. [PubMed: 19581875]
16. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009; 25:1754–60. [PubMed: 19451168]
17. Rumble SM, Lacroute P, Dalca AV, Fiume M, Sidow A, Brudno M. SHRiMP: accurate mapping of short color-space reads. *PLoS Comput Biol.* 2009; 5:e1000386. [PubMed: 19461883]

[illegible]

Fig. 1.
A diagram depicting the Helicos small RNA method for the preparation of cDNA for sequencing to the downstream processing of raw sequence reads.

Fig. 2. BioAnalyzer profiles of (a) pure small RNA fraction and (b) the fraction contaminated with long RNAs.

Results of the mapping of HeLaS3 sRNAs to the human genome with a minimal length of a trimmed read = 25 bases and minimal normalized score = 4.5

Table 1

Category	General method		Method to profile 3' polyA sRNAs			
	Number of reads: three channels	Fraction of uniquely mapping reads (%)	Fraction of uniquely mapping novel reads (%)	Number of reads: two channels	Fraction of uniquely mapping reads (%)	Fraction of uniquely mapping novel reads (%)
Total filtered reads	54,183,628			32,719,870		
All mapping reads	17,938,108			10,103,423		
Uniquely mapping	10,353,110	100		7,131,693	100	
Complete ribosomal repeat unit	5,974,541	57.7		4,593,222	64.4	
Mitochondrial	380,982	3.7		101,060	1.4	
chrY	8,855	0.1		1,345	0	
Repeats ^a	2,600,485	25.1		1,031,027	14.5	
Sno-miRNAs	997,801	9.6		49,408	0.7	
Selected group of small, non-coding RNAs (RNAse P RNA, U12, etc.)	105,204	1		57,845	0.8	
RNA genes	100,178	1		22,500	0.3	
Predicted sno	26,610	0.3		243	0	
Novel	158,454	1.5		1,275,043	17.9	
PASRs ^b	30,762	0.3	19.4	583,907	8.2	45.8
TASRs ^c	16,152	0.2	10.2	322,220	4.5	25.3

^a As annotated by the RepeatMasker on the UCSC Browser track

^b PASRs and TASRs were defined as regions ±500 bp from the respectively 5' and 3' end of a transcript annotated by UCSC Genes track

^c The reads were first overlapped with PASR regions and then those that do not overlap the PASR regions were overlapped with the TASR regions