Analytical Biochemistry 419 (2011) 317–322

Contents lists available at SciVerse ScienceDirect

Analytical Biochemistry

journal homepage: www.elsevier.com/locate/yabio

Local and global factors affecting RNA sequencing analysis

Edward Sendler [a], Graham D. Johnson [b], Stephen A. Krawetz [a,b,*]

[a] Department of Obstetrics and Gynecology, C.S. Mott Center for Human Growth and Development, Wayne State University School of Medicine, Detroit, MI 48201, USA
[b] Center for Molecular Medicine and Genetics, C.S. Mott Center for Human Growth and Development, Wayne State University School of Medicine, Detroit, MI 48201, USA

ARTICLE INFO

Article history:
Received 12 July 2011
Received in revised form 8 August 2011
Accepted 8 August 2011
Available online 16 August 2011

Keywords:
RNA
RNA sequencing
Sequence analysis
High-throughput RNA sequencing

ABSTRACT

High-throughput RNA sequencing (RNA-seq) continues to provide unparalleled insight into transcriptome complexity. Now the "gold standard" for assessing global transcript levels, RNA-seq is poised to revolutionize our understanding of transcription and posttranscriptional regulation of RNA. Despite significant advantages over prior experimental strategies, RNA-seq is not without pitfalls. We have identified a number of confounding factors that significantly affect sequencing coverage. These include regional GC content, preferential sites of fragmentation, and read "pile-up" due to primer affinity and transcript end effects. Independent of cell type and laboratory, when ignored, these factors can bias analyses. Understanding the underlying principles responsible for producing these artifacts is key to recognizing both their presence and how their effects may be controlled. Here we outline the causes of and strategies to avoid several previously unreported complicating factors common to RNA-seq experiments.

The advantages of using high-throughput RNA sequencing (RNA-seq)[1] technologies for analysis of a population of RNAs and their composition are many-fold [1–3]. Unlike microarrays and reverse transcription polymerase chain reaction (PCR), massively parallel sequencing requires no a priori knowledge of content while enabling quantitative interrogation of the transcriptome [4,5]. This has led to the discovery of many novel transcripts, which in turn has spurred an explosion in the number of accepted classes of RNAs [6,7]. Novel noncoding RNAs such as micro RNAs (miRNAs) and Piwi-interacting RNAs (piRNAs), which are intrinsic to global gene regulation, may be inferred by computational analysis of RNA-seq results [8,9]. Posttranscriptional modification of individual RNAs stemming from specific cleavage or degradation pathways may be detected from individual RNA profiles [10]. In addition, RNA-seq is able to provide an accurate assessment of the relative levels of individual transcripts, including their related isoforms [11–13].

To fully capitalize on the opportunities afforded by RNA-seq, one must be aware of its limitations. Common pitfalls in sample preparation and analysis can bias sequence representation in a template-dependent manner. The net effect of multiple artifacts on read distribution may skew both inferred transcript structure, quantification of relative abundance, and mask transcriptome complexity. This is quite perilous to those who may review only the final "black box" summary of the sequencing results [14].

RNA-seq initiates from the extraction of total RNA, removal of the majority of ribosomal RNAs (rRNAs), fragmentation to approximately twice the expected read size (usually by metal ion hydrolysis), adapter ligation, followed by PCR amplification and finally sequencing. Examination of RNA-seq datasets generated by our group and others has uncovered several factors intrinsic to this process that can influence interpretation. Here we present a summary of the artifacts we have observed, including their probable causes and (where possible) resolution. Recognition is essential to accurately assessing complex transcriptional networks.

Materials and methods

RNA library preparation and sequencing

Sequencing datasets from spermatozoal RNAs extracted from three human ejaculates were prepared essentially as described [14]. All samples were subject to paired-end sequencing using the Illumina Genome Analyzer GAIIx for 36 cycles. Image analysis and base calling were performed using the Firecrest and Bustard modules of the genome analyzer pipeline software (Illumina Pipeline software, version 1.3.0). For supplementary comparison, two somatic RNA-seq datasets, GSM424320 (a B-cell lymphoblastoid cell line) [11] and SRR002321 (a liver cell line) [15], were obtained from NCBI Gene Expression Omnibus (GEO) and Sequence Read Archives (SRA), respectively. In addition, to compare expression characteristics from alternative sequencing methods, human placental

* Corresponding author at: Department of Obstetrics and Gynecology, C.S. Mott Center for Human Growth and Development, Wayne State University School of Medicine, Detroit, MI 48201, USA. Fax: +1 313 577 8554.
E-mail address: steve@compbio.med.wayne.edu (S.A. Krawetz).

[1] *Abbreviations used:* RNA-seq, RNA sequencing; PCR, polymerase chain reaction; rRNA, ribosomal RNA; FRT-seq, amplification-free, strand-specific transcriptome sequencing; SNORD, small nucleolar.

0003-2697/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ab.2011.08.013

sample 2664_1 sequenced by amplification-free, strand-specific transcriptome sequencing (FRT-seq) was downloaded from European Nucleotide Archive (ERA000183).

Data assembly and analysis

The RNA-seq reads for all samples were aligned separately to rRNAs 18S (NR_003286.2, 1869 bp length) and 28S (NR_003287.2, 5070 bp length). Alignments were carried out using Novoalign (version 2.05.43, Novocraft Technologies). Samples with paired-end reads were aligned as separate ends to maximize the number of alignments. Default Novoalign parameters for alignment threshold and acceptable quality control of sequence reads were used. Alignment results were confirmed independently using GERALD (Illumina). Selected samples were aligned to the complete human mitochondrion genome (HQ231912.1; 12S region: 649–1602 nt; 16S region: 1672–3229 nt) to check for the presence of mitochondrial rRNA. The 5S and 5.8S RNAs corresponded to less than 0.1% of the total number of sequence reads and were not considered further.

Results and discussion

Initial observations showing significant nonuniformity of sequencing reads when progenitor and mature cell types were compared [16] prompted the analysis of RNAs isolated from other cell lines and sequenced by independent laboratories. rRNAs were selected for comparison because they are the most abundant cellular RNAs, typically representing at least 80% and up to 98% of the transcripts in total fraction of isolated RNAs, and also are comparatively resistant to degradation. Accordingly, comparison of the distribution of the sequencing reads across the length of the ribosomal transcripts should provide an unbiased survey of sequencing uniformity.

Ideally, alignment of the sequencing reads across a given transcript should produce uniform coverage of that transcript barring degradation or the expression of isoforms and/or overlapping RNAs. As shown in Fig. 1, this is usually not the case when rRNA sequencing reads from several studies [11,14,15] were compared across 28S rRNA (NR_003287.2, 5040 bp) and 18S rRNA (NR_003286.2, 1840 bp) (not shown). The profile of 28S and 18S transcripts shows similar read-sparse and read-dense regions across four different cell types (multiple R^2 correlation = .68). The number of reads varies significantly across the sequence as a function of nucleotide position among samples (multiple R^2 correlation across complete 28S length for three sperm samples = .96). As described below, the reiterative global and local variations in sequencing coverage observed for the rRNAs have highlighted several factors that affect RNA-seq analyses.

Substrate

The most noticeable features of the intact 28S RNA-seq read profile are the large regions that are virtually absent. It is known that Illumina sequencing presents a read bias at GC residues compared with AT residues [17]. Subregions with high GC content are significantly reduced from sequencing libraries due to a high degree of secondary structure impeding amplification [18], which is alleviated when composition changes. As shown in Fig. 1, this effect is highlighted at sites where a local decrease in GC content (e.g., positions 3000–3500) in high-GC areas coincides directly with a significant increase in the number of sequence reads at that position. This negative correlation of GC content as a function of sequence reads along the length of 28S rRNA is shown in Fig. 1. Interestingly, when PCR-free FRT-seq is employed, the same reduced coverage of high GC-rich regions in rRNA is observed [19]. This modified sequencing library protocol eliminates biases introduced at the amplification step by undertaking the reverse transcription reaction directly on the flow cell following ligation of the adaptors to the RNAs. Thus, it appears that in addition to PCR amplification [17], high GC content also significantly impairs the efficiency of reverse transcription. Hence, the underrepresentation of GC-rich RNAs in sequencing libraries is due to being less amenable to reverse transcription [20] and their corresponding complementary DNAs (cDNAs) being more resistant to denaturation, thereby reducing PCR amplification [14,21].

To examine how each of these factors may play a role in the final population of individual regions within a library, intramolecular and duplex second-strand binding energies for 60-nt segments along the 28S sequence were calculated using OligoWalk (version 5.1) [22,23]. Intramolecular energy measures the proclivity of single-stranded RNA segments to undergo self-structure formation, whereas duplex energy reflects how strongly an RNA segment binds to a strand complement. As shown in Fig. 2, both intramolecular and duplex binding energy are correlated with amplification bias. In addition, this shows that significant self-structure impedes both reverse transcription and further PCR amplification. The level of impediment is illustrated by the R^2 of 0.20 when intramolecular versus log(read count) is considered and 0.43 when duplex energy

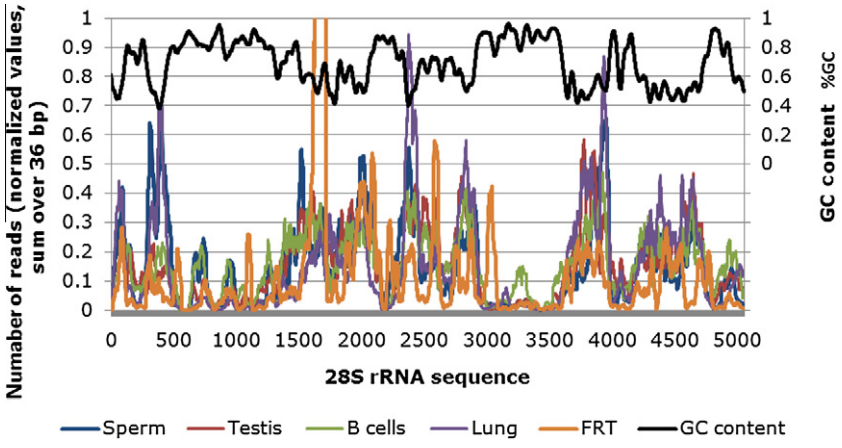

Fig.1. Elevated GC content negatively influences sequencing coverage. (A) Despite poly(A) enrichment, the 28S rRNA exhibits significant but nonuniform representation across sequencing libraries from multiple cell types. Across all samples, sequencing coverage of this transcript exhibited a strong negative correlation with GC content. Regions exceeding 70% GC were read sparse. Conversely, dips in GC content correspond directly to local increases in coverage. FRT sample was sequenced by amplification-free, strand-specific sequencing (human placental).

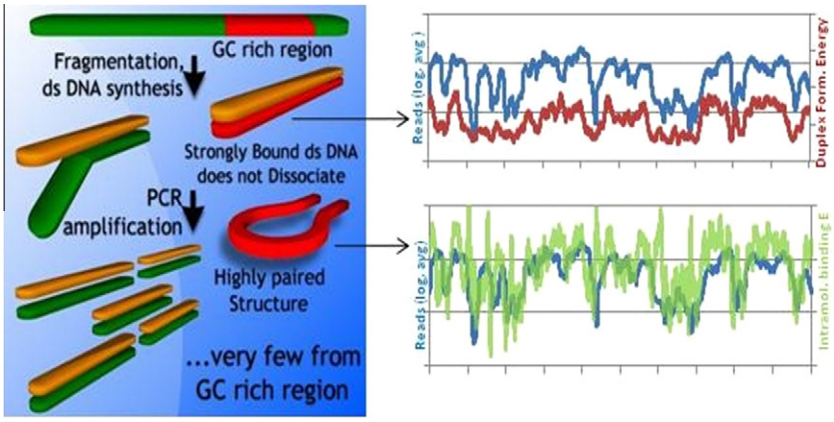

Fig.2. Reverse transcription PCR amplification exclusion of GC-rich regions. GC-rich fragments undergo base pairing, forming double-stranded and/or highly paired secondary structures that can impede both reverse transcription and subsequent PCR amplification. Energies of highly stable duplexes (top) and self-structures (bottom) are compared with sequencing coverage across the 28S rRNA (more negative energy reflects stronger annealing). Log (total counts) is shown because it more clearly shows the correlation to energy calculations across range of sequenced read values. ds, double-stranded.

versus log(read count) is considered. Although both factors may affect representation of specific fragments within the final population, characteristics of each may differ across a given region. The structure of a given fragment as indicated by intramolecular energy may change from one base position to the next. In comparison, duplex binding energy shows variation only over a larger region. This is clearly illustrated by using the RNAstructure (version 5.3) software package to predict the structure of 60-nt RNA fragments initiating at nucleotides 272 through 276 of the 28S rRNA (Fig. 3A) [24]. Shifting the start position by a single nucleotide (from position 274 to 275) markedly reduced folding stability, as indicated by the open conformation. A concomitant increase in sequencing coverage of this subregion was observed after nucleotide 274 (Fig. 3B). This may reflect increased efficiency of both second-strand synthesis and PCR amplification.

Primer

Creating an RNA-seq library often employs the use of random hexamer primers. However, each primer within this heterogeneous mixture has a unique binding affinity [25]. Fig. 4 clearly shows the effect of 5′ primer bias on the relative distribution of initial 3-mers of all reads across the 18S and 28S rRNA and 12S mitochondrial RNA.

Fragmentation and sequence pile-up

The susceptibility of specific regions of an RNA to chemical fragmentation is largely dependent on secondary structure, with single-stranded or loop regions being more susceptible to cleavage than double-stranded regions. In addition, even within regions of similar structure, cleavage occurs preferentially at specific sites as dictated by local nucleotide composition [26]. The increased susceptibility of specific sites to fragmentation can exaggerate the representation of that sequence within a library. Preferential fragmentation resolves as a pile-up of reads in the forward direction at nucleotide n, and the corresponding pile-up of reverse reads starting (with mapped read aligning in the reverse direction) at position $n - 1$. As illustrated in Fig. 5A, many sites of such preferential fragmentation can be observed across the 28S rRNA and can be discerned from typical background nonuniformity. Overall preference for cleavage between specific dinucleotides is summarized in Fig. 5C. For example, the presence of bidirectional read pile-up at nucleotide positions 2984 and 2985 shows an accumulation of reads originating from the same site but with no clear apparent end location (Fig. 5B). This corresponds to a domain with a single clearly defined site of fragmentation. In comparison, the 3050–3100 range shows sequence reads starting at both ends, extending in a single direction toward the inner part of the region. This pattern is indicative of a broader GC-rich region that is generally resistant to sequencing but in which an "island" of reduced GC content exists, enabling that sequence to be read.

Read pile-up may be observed at very specific positions of the extreme ends of transcripts. Any clearly defined fragment end will act as an anchor for read start and end sites. It is expected that a diminution of reads will be observed at the 5′ and 3′ ends of RNA due to the biologically expected modes of RNA degradation [12]. This effect differs when poly(A)-enriched RNAs serve as templates. Only transcripts with intact 3′ poly(A) tails are selected, masking 3′ degradation while the reduction in 5′ sequence reads is maintained. Ends of transcripts are perhaps the most clearly defined example of such fragment demarcations given that newly synthesized full-length RNAs—without further fragmentation—contain these precise termination sites. Hence, there will be increased clustering of read starts at the 5′ and 3′ transcript ends, with the exact start site of these reads situated n nucleotides from the terminus, where n is sequence read length. SNORD (small nucleolar) RNAs are prime examples of this RNA-seq artifact. Typically 60–100 nt in length, sequence read pile-ups at both ends are readily observed (Fig. 6). There is a clear concentration of read starts exactly at the start position of SNORD transcripts (relative to transcript orientation), with the second cluster of reads anchored so that the read termination falls at the transcript end.

A series of factors contributing to the nonuniform distribution of sequencing reads will affect interpretation. It should be noted that although the above focused on the nonuniformity arising from sequencing and associated sample preparation, mapping of sequencing reads to a given genome may introduce additional confounders. Consider repetitive regions across the genome. Reads stemming from repeat-rich segments would likely not be mapped because mapping algorithms typically discard (or at least classify separately) any reads that map with similar accuracy to multiple sites.

Large regional variations in expression determination appear primarily in transcripts or transcript regions with a GC content greater than 0.7. Such regions include portions of rRNAs, collagen-type genes, and the 5′ and 3′ ends of insulin receptor (IR) and cSRC kinase. Many regulatory and housekeeping genes may also be resistant to accurate measurement because they are relatively GC rich (>0.6) or within CpG islands. This effect may be

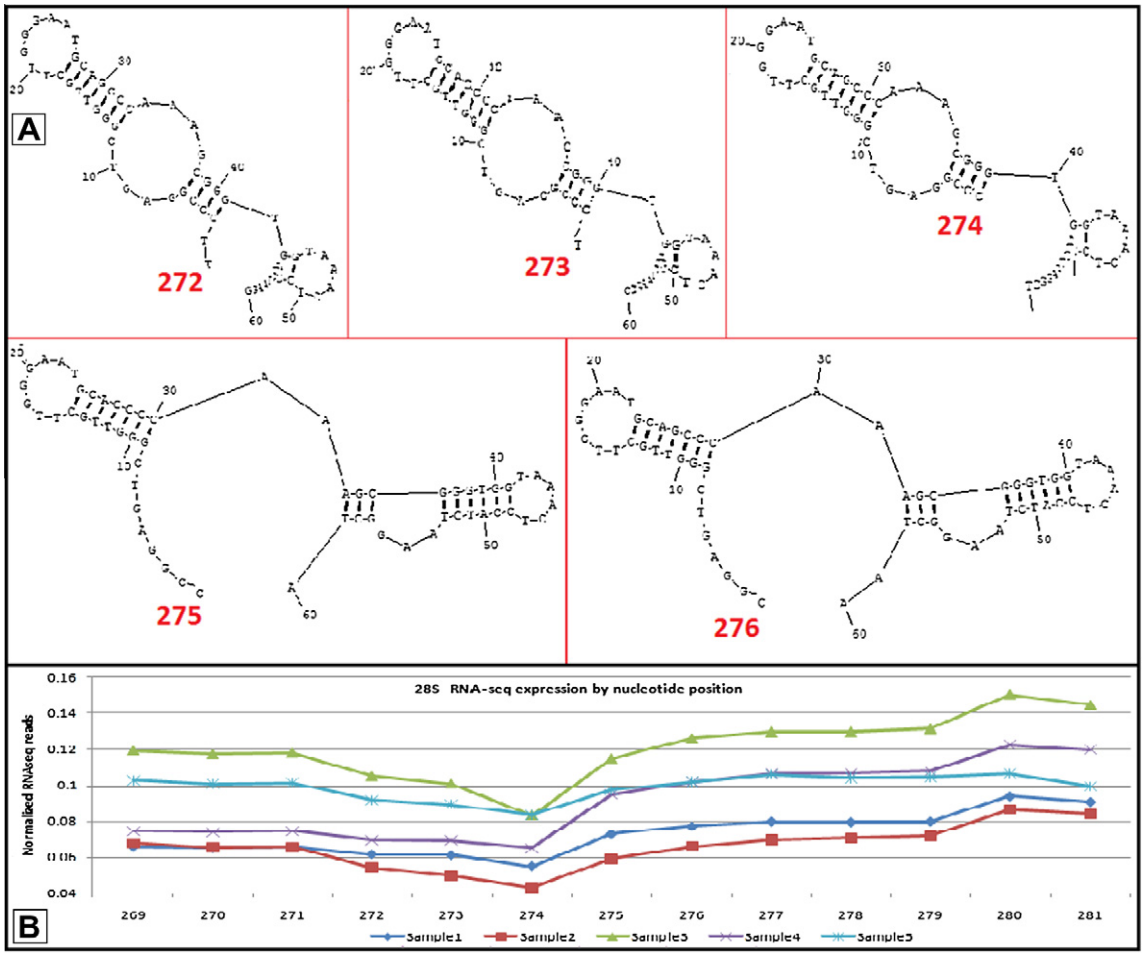

Fig.3. Structural dependence of RNA fragments on local expression. (A) Predicted secondary structures of 60-nt fragments (average library length) initiating at nucleotides 272 through 276 of the 28S rRNA. Changing the fragment start by a single nucleotide can have a dramatic effect on intramolecular (self-binding) structure. (B) Normalized sequencing coverage of 60-nt fragments initiating from the region highlighted above as well as from flanking nucleotides. Values from different sperm sequencing libraries are shown. Note the more open conformation of those fragments starting after nucleotide 274 and the increase in coverage observed for this subregion following that position.

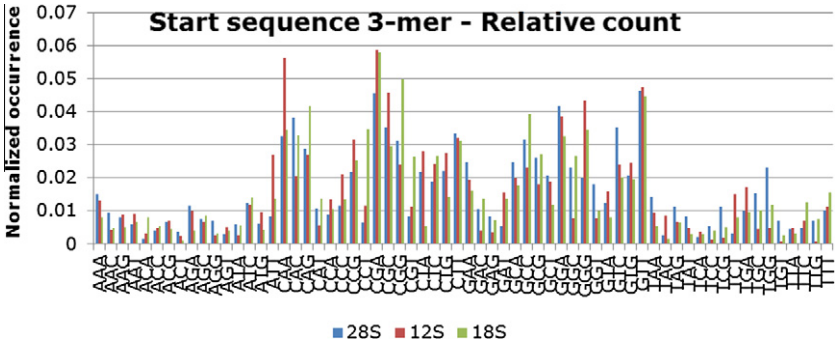

Fig.4. Average number of reads for read sequences with different 3-mer start sequences. For reads aligning to each of 18S and 28S rRNA and 12S mitochondrial RNA, normalized (by total 3-mer occurrence) number of reads for each possible 3-mer start sequence is shown. Bias due to nonrandom primers for second-strand synthesis exhibits a significant preference for specific starting sequences common to all transcripts examined.

attenuated somewhat for nonextreme GC domains by adjusting the amplification protocol [18].

Although considerable, local biases reflected in priming or sites of fragmentation may largely average out over a longer transcript and have little effect on the overall quantification of gene expression. However, disparities do arise from secondary structure and duplex stability, preferential sites of chemical fragmentation, PCR primer bias, and transcript end pile-ups. Some of these effects may be minimized with changes in RNA-seq sample preparation or sequencing protocol. Alternative methods of fragmentation, such as nebulization and adaptive focused acoustics, have been suggested as a means of improving fragment size resolution [27]. To what degree these methods might influence, lessen relative preference for fragmentation, or improve uniformity of coverage

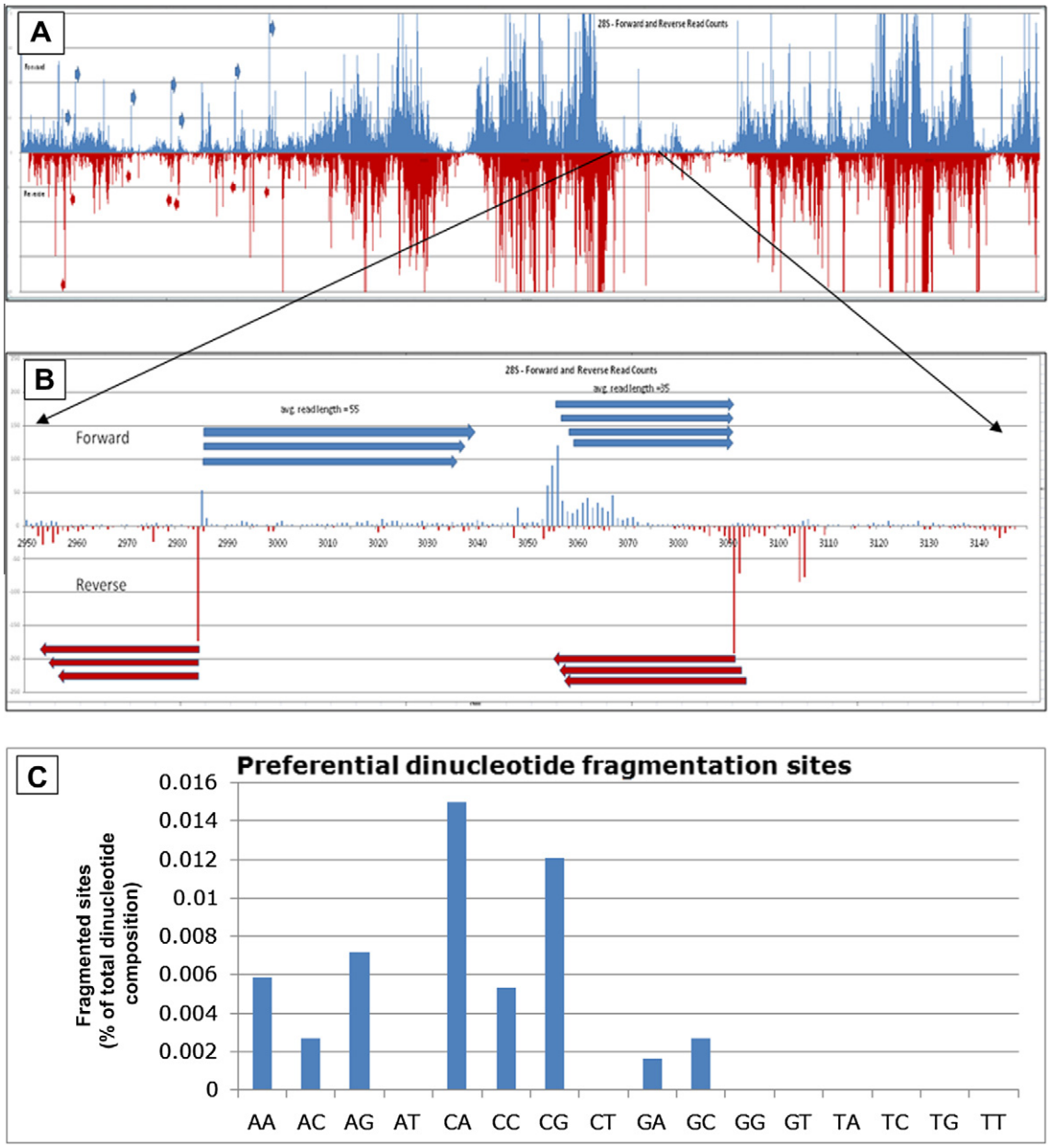

Fig.5. Read pile-ups at sites of preferential RNA fragmentation. (A) Paired-end, strand-specific (blue: forward; maroon: reverse) reads from full length (5040 nt) of 28S rRNA (FRT sample). Arrows denote locations where preferred sites of fragmentation are observed, with pile-up of reads in both forward direction (at position n) and reverse direction (at position $n-1$). Arrows show region magnified in panel B. (B) Blow-up of 2950–3150 region of 28S. Preferential cleavage site at position 2985 is seen by pile-up of bidirectional reads starting at this position. Forward reads starting at position 2985 have an average length of 55 and no preferred end site. Forward reads starting at position 3054 have an average length of 35 and extend primarily to positions 3088 to 3091—evidence that fragments from this start site are preferentially terminated at this position. (Read lengths of each read are obtained from position of reverse read of forward paired end.) (C) Dinucleotide distribution of sites of 28S rRNA fragmentation. A total of 36 clearly identifiable preferential fragmentation sites along the length of 28S rRNA were examined for nucleotide composition at the start site and one prior base. Fragment cleavage with resultant start of RNA-seq read occurs between these dinucleotide pairs. The y axis shows fraction of sites for each dinucleotide that shows evidence of cleavage; of 1244 CG dimers in 28S, 15 (=.012) are identified as cleavage sites. Chemical fragmentation appears to show bias toward particular base pairings. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

was not examined. Similarly, PCR primer bias may be reduced by taking advantage of newer methodologies that have no PCR amplification such as FRT-seq [19].

Recent studies have examined whether these effects may be normalized by considering local sequence [28,29]. Global levels of individual nucleotide sequences over every occurrence in the mapping library were examined and used as a training set. A weighting factor was calculated based on overall level of each sequence in this set and then applied as normalization to each corresponding sequence in the test library. This simple statistical approach can resolve up to 50% of the irregularity. It is likely that some factors of bias identified in this study are amenable to normalization. Primer bias, which is largely determined by the initial 8 to 13-mer of the target and GC content, may be reasonably estimated by local sequence. Preferential sites of chemical fragmentation that lead to read pile-ups may be somewhat influenced and identifiable by local dinucleotide pairing. Local proclivity to RNA cleavage, which in turn dictates favored sequence start sites, is largely influenced by RNA secondary structure [26]. Secondary structure also dictates the efficacy of reverse transcription and subsequent fragment amplification. These present formidable challenges and have yet to be addressed. Numerical normalization may

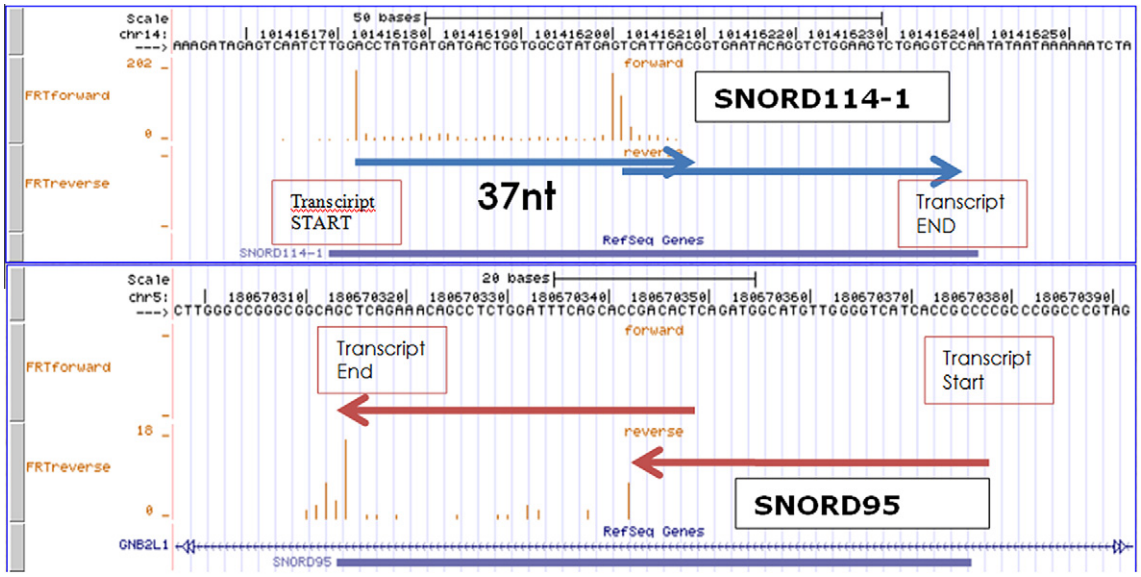

Fig.6. Read pile-ups at transcript start and end termini. Shown are forward and reverse reads from strand-specific FRT sample across length of two SNORD RNAs. Reads aligning in reverse direction start at the 5′ terminal position of read. For both examples, reads are piled up at two positions: at the transcript start according to direction of transcription and in the middle of the transcript at a site corresponding to the end of the transcript end read length (37 nt).

be helpful in reducing some RNA-seq irregularities, but currently it cannot account for a significant fraction of underlying nonuniformity in expression.

Acknowledgments

We thank Yitzchok Sendler for his assistance with artwork. This work was supported in part by the Presidential Research Enhancement Program in Computational Biology and was supported by the Charlotte B. Failing Professorship to S.A.K. This article is subject to National Institutes of Health (NIH) public access policy.

References

[1] Z. Wang, M. Gerstein, M. Snyder, RNA-seq: a revolutionary tool for transcriptomics, Nat. Rev. Genet. 10 (2009) 57–63.
[2] B.T. Wilhelm, S. Marguerat, S. Watt, F. Schubert, V. Wood, I. Goodhead, C.J. Penkett, J. Rogers, J. Bahler, Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution, Nature 453 (2008) 1239–1243.
[3] M. Garber, M.G. Grabherr, M. Guttman, C. Trapnell, Computational methods for transcriptome annotation and quantification using RNA-seq, Nat. Methods 8 (2011) 469–477.
[4] M. Yassour, T. Kaplan, H.B. Fraser, J.Z. Levin, J. Pfiffner, X. Adiconis, G. Schroth, S. Luo, I. Khrebtukova, A. Gnirke, C. Nusbaum, D.A. Thompson, N. Friedman, A. Regev, Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing, Proc. Natl. Acad. Sci. USA 106 (2009) 3264–3269.
[5] F. Denoeud, J.M. Aury, C. Da Silva, B. Noel, O. Rogier, M. Delledonne, M. Morgante, G. Valle, P. Wincker, C. Scarpelli, O. Jaillon, F. Artiguenave, Annotating genomes with massive-scale RNA sequencing, Genome Biol. 9 (2008) R175.
[6] H. Siomi, M.C. Siomi, On the road to reading the RNA-interference code, Nature 457 (2009) 396–404.
[7] T.R. Mercer, M.E. Dinger, J.S. Mattick, Long non-coding RNAs: insights into functions, Nat. Rev. Genet. 10 (2009) 155–159.
[8] M.R. Friedlander, W. Chen, C. Adamidi, J. Maaskola, R. Einspanier, S. Knespel, N. Rajewsky, Discovering microRNAs from deep sequencing data using miRDeep, Nat. Biotechnol. 26 (2008) 407–415.
[9] A. Kozomara, S. Griffiths-Jones, miRBase: integrating microRNA annotation and deep-sequencing data, Nucleic Acids Res. 39 (2011) D152–D157.
[10] S. Marguerat, J. Bahler, RNA-seq: from technology to biology, Cell. Mol. Life Sci. 67 (2010) 569–579.
[11] J.M. Toung, M. Morley, M. Li, V.G. Cheung, RNA-sequence analysis of human B-cells, Genome Res. 21 (2011) 991–998.
[12] A. Mortazavi, B.A. Williams, K. McCue, L. Schaeffer, B. Wold, Mapping and quantifying mammalian transcriptomes by RNA-seq, Nat. Methods 5 (2008) 621–628.
[13] C. Trapnell, B.A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M.J. van Baren, S.L. Salzberg, B.J. Wold, L. Pachter, Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation, Nat. Biotechnol. 28 (2010) 511–515.
[14] G.D. Johnson, E. Sendler, C. Lalancette, R. Hauser, M.P. Diamond, S.A. Krawetz, Cleavage of rRNA in ensures translational cessation to prepare sperm for fertilization, Mol. Hum. Reprod. (2011) PMID:21831882.
[15] J.C. Marioni, C.E. Mason, S.M. Mane, M. Stephens, Y. Gilad, RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays, Genome Res. 18 (2008) 1509–1517.
[16] G.D. Johnson, C. Lalancette, A.K. Linnemann, F. Leduc, G. Boissonneault, S.A. Krawetz, The sperm nucleus: chromatin, RNA, and the nuclear matrix, Reproduction 141 (2011) 21–36.
[17] J.C. Dohm, C. Lottaz, T. Borodina, H. Himmelbauer, Substantial biases in ultra-short read data sets from high-throughput DNA sequencing, Nucleic Acids Res. 36 (2008) e105.
[18] D. Aird, M.G. Ross, W.S. Chen, M. Danielsson, T. Fennell, C. Russ, D.B. Jaffe, C. Nusbaum, A. Gnirke, Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries, Genome Biol. 12 (2011) R18.
[19] L. Mamanova, R.M. Andrews, K.D. James, E.M. Sheridan, P.D. Ellis, C.F. Langford, T.W. Ost, J.E. Collins, D.J. Turner, FRT-seq: amplification-free, strand-specific transcriptome sequencing, Nat. Methods 7 (2010) 130–132.
[20] Y.J. Zhang, H.Y. Pan, S.J. Gao, Reverse transcription slippage over the mRNA secondary structure of the LIP1 gene, BioTechniques 31 (2001) 1286–1290.
[21] F. Payvar, R.T. Schimke, Methylmercury hydroxide enhancement of translation and transcription of ovalbumin and conalbumin mRNA's, J. Biol. Chem. 254 (1979) 7636–7642.
[22] N. Sugimoto, S. Nakano, M. Katoh, A. Matsumura, H. Nakamuta, T. Ohmichi, M. Yoneyama, M. Sasaki, Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes, Biochemistry 34 (1995) 11211–11216.
[23] D.H. Mathews, M.E. Burkard, S.M. Freier, J.R. Wyatt, D.H. Turner, Predicting oligonucleotide affinity to nucleic acid targets, RNA 5 (1999) 1458–1469.
[24] J.S. Reuter, D.H. Mathews, RNAstructure: software for RNA secondary structure prediction and analysis, BMC Bioinf. 11 (2010) 129.
[25] K.D. Hansen, S.E. Brenner, S. Dudoit, Biases in Illumina transcriptome sequencing caused by random hexamer priming, Nucleic Acids Res. 38 (2010) e131.
[26] J. Ciesiolka, D. Michalowski, J. Wrzesinski, J. Krajewski, W.J. Krzyzosiak, Patterns of cleavages induced by lead ions in defined RNA secondary structure motifs, J. Mol. Biol. 275 (1998) 211–220.
[27] M.A. Quail, I. Kozarewa, F. Smith, A. Scally, P.J. Stephens, R. Durbin, H. Swerdlow, D.J. Turner, A large genome center's improvements to the Illumina sequencing system, Nat. Methods 5 (2008) 1005–1010.
[28] J. Li, H. Jiang, W.H. Wong, Modeling non-uniformity in short-read rates in RNA-seq data, Genome Biol. 11 (2010) R50.
[29] A. Roberts, C. Trapnell, J. Donaghey, J.L. Rinn, L. Pachter, Improving RNA-seq expression estimates by correcting for fragment bias, Genome Biol. 12 (2011) R22.