Genomics 96 (2010) 259–265

Contents lists available at ScienceDirect

Genomics

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

A comparison between ribo-minus RNA-sequencing and polyA-selected RNA-sequencing

Peng Cui [a,1], Qiang Lin [a,b,1], Feng Ding [a,1], Chengqi Xin [a,b], Wei Gong [a,b], Lingfang Zhang [a,b], Jianing Geng [a], Bing Zhang [a], Xiaomin Yu [a], Jin Yang [a], Songnian Hu [a,*], Jun Yu [a,*]

[a] The CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, 100029 Beijing, China
[b] Graduate School of the Chinese Academy of Sciences, 100029 Beijing, China

ARTICLE INFO

Article history:
Received 16 December 2009
Accepted 28 July 2010
Available online 3 August 2010

Keywords:
Ribominus
RNA-seq
mRNA-seq

ABSTRACT

To compare the two RNA-sequencing protocols, ribo-minus RNA-sequencing (rmRNA-seq) and polyA-selected RNA-sequencing (mRNA-seq), we acquired transcriptomic data–52 and 32 million alignable reads of 35 bases in length–from the mouse cerebrum, respectively. We found that a higher proportion, 44% and 25%, of the uniquely alignable rmRNA-seq reads, is in intergenic and intronic regions, respectively, as compared to 23% and 15% from the mRNA-seq dataset. Further analysis made an additional discovery of transcripts of protein-coding genes (such as *Histone, Heg1, and Dux*), ncRNAs, snoRNAs, snRNAs, and novel ncRNAs as well as repeat elements in rmRNA-seq dataset. This result suggests that rmRNA-seq method should detect more polyA- or bimorphic transcripts. Finally, through comparative analyses of gene expression profiles among multiple datasets, we demonstrated that different RNA sample preparations may result in significant variations in gene expression profiles.

1. Background

Next-generation sequencing technology has been applied successfully to the study of transcriptomics, known as ultra-high-throughput RNA sequencing or RNA-seq [1–8]. This method is advantageous over the existing approaches in dynamic range, sampling depth, and material processing, which include microarrays, expression sequence tags (EST), serial analysis of gene expression (SAGE), cap analysis of gene expression (CAGE), and massively parallel signature sequencing (MPSS). Currently, RNA-seq method is mainly adapted to study polyadenylated-(polyA+)-transcripts [9], and has not fully exploited for identifying non-polyA or NPA-transcripts.

Poly(A+)-transcripts–including mRNAs, microRNAs, and snoRNAs–are synthesized by RNA polymerase II [10] and often isolated by using oligo-dT affinity. Total RNA preparations are often overwhelmed by non-polyA(NpA)-transcripts due to their massive nature, including ribosomal RNAs [11], histone mRNAs (generated by RNA polymerase II [12]), tRNAs and certain small RNAs (generated by RNA polymerase III), and bimorphic RNAs [13]. In fact, recent studies suggested that certain functional non-coding and protein-coding RNAs do not possess polyA tails [14]. Moreover, it was reported that the amount of NpA-transcripts is twice as much as those of polyA+ transcripts among cytosolic RNAs in human cells [15].

To test a new protocol for better defining eukaryotic transcriptomes, we sequenced ribosomal RNA-depleted (ribo-minus or rm) RNA from total RNA through hybridization and biotin–streptavidin binding, using the next-generation sequencing technology (SOLiD, Life Technologies). We also compared the results from both rm-RNA sequencing (rmRNA-seq) protocol and the standard mRNA-seq method, using isolated mRNA based on oligo-dT affinity. Although the RNA samples were from the mouse cerebrum, our analysis was focused on detailed data characteristics between the two methods rather than biological relevance of the RNA source. We report the differences between the two methods and propose that rmRNA-seq has merits in studying eukaryotic transcriptomes in thoroughness.

2. Results

2.1. Sequence acquisition from RNA-seq libraries

From the two libraries constructed from ribo-minus (Fig. 1) and polyA-selected RNAs of the mouse cerebrum, we obtained 140 and 93 millions high-quality reads (35-bp in length), respectively (NCBI short read accession number: SRA009022). We mapped the tags to both the mouse genome (release mm9, July 2007 from UCSC) and our custom-made database containing unique exon-junction sequences (Supplementary Fig. S1). Greater than 19% (~26 million) of the total reads were annotated from the raw rmRNA-seq data to be unique to the genome and similar amount to be multiple (Table 1). We further aligned the unique reads onto Refseq-defined gene models and found

* Corresponding authors.
E-mail addresses: husn@big.ac.cn (S. Hu), junyu@big.ac.cn (J. Yu).
[1] These authors contributed equally to this work.

0888-7543/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2010.07.010

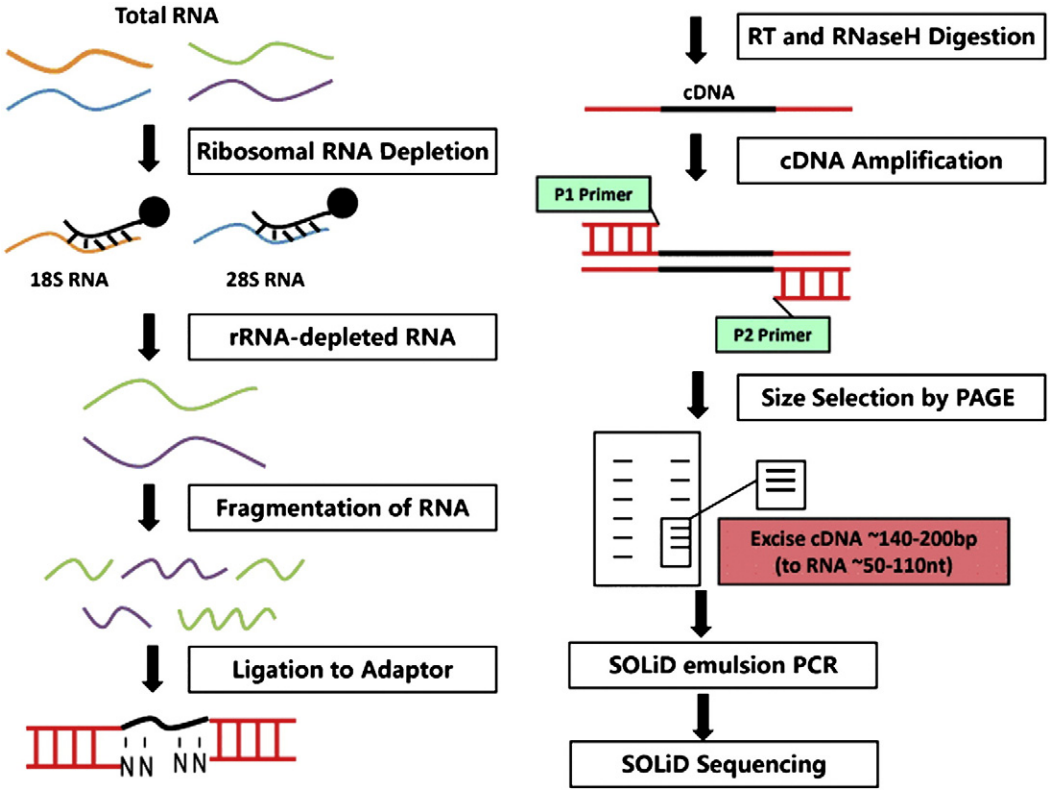

Fig. 1. A flowchart of our rmRNA-seq protocol. Ribosomal RNAs (colored in orange and blue) are depleted with sequence-specific biotin-labeled probes and the remaining mRNA-rich fraction (green and violet) is fragmented with RNase III. After ligation to adaptors (red; NN stands for random oligonucleotide hexamers), the fragments in a size range of ~50 bp are collected and reverse-transcribed into a single-stranded cDNA library. The library is subsequently amplified, size-selected (140 to 200 bp), and sequenced in high coverage.

that 29%, 25%, and 44% of them were within exonic, intronic, and intergenic regions, respectively. In contrast, from the mRNA-seq library we mapped about 28% (~26 millions) and 5% (~4 million) of total reads to unique and multiple loci, respectively. We further aligned the unique reads onto Refseq-defined gene models and found that 59%, 15%, and 23% of them were within exonic, intronic, and intergenic regions, respectively. It became clear that greater ratios of sequence tags from the rmRNA-seq library were mapped onto intronic and intergenic regions than those from the mRNA-seq library. It was further supported by an expression intensity analysis where we placed uniquely mapped reads along the concatenated intergenic region of 21 chromosomes based on both rmRNA-seq and mRNA-seq datasets (Supplementary Fig. S2). 22.5% genomic regions were identified to have significant transcription activities based on the rmRNA-seq data but these regions were lost when we used mRNA-seq data.

Table 1
Summary of read-mapping[a].

	mRNA-seq	rmRNA-seq
Total reads	92,914,107	140,233,818
Ribosomal reads (%)[b]	1.46	10.06
Unique (%)	28.48	18.61
Multiple (%)	4.73	18.31
Exon–exon junction (%)	1.09	0.42
Read distribution[c]		
Exonic region (%)	59.49	29.50
Intronic region (%)	14.77	24.81
Intergenic region (%)	23.26	44.15
Exon–intron junction (%)	2.48	1.54

[a] Raw reads mapped to the mouse genome (mm9, NCBI build 37).
[b] Raw reads removed based on 18S, 28S, and 5S RNA sequences.
[c] The distribution of the uniquely mapped reads based on the RefSeq-defined gene models.

2.2. Abundant non-coding transcripts in the rmRNA-seq dataset

To annotate the non-coding transcripts, we constructed a non-redundant non-coding RNA library derived from ncRNAdb, RNAdb, fRNAdb, FANTOM3, NONCODE, Refseq, and Ensembl. We classified them into 8 categories: misc_RNA, mRNA-like, long ncRNA(>200 bp), short ncRNA(<200 bp), pseudogene, retrotransposed, snoRNA, and snRNA. Using a background of 0.04 hits/kb determined with a Poisson model (detailed in Methods) and five reads cutoff, we were able to identified 20,578 known non-coding transcripts expressed in mouse cerebrum in the rmRNA-seq dataset as compared to 17,358 detected in the mRNA-seq dataset (Supplementary Table S1). The two datasets share 15,319 of these transcripts and each has a few thousands unique to itself—5,261 and 2,040 from the rmRNA-seq and mRNA-seq datasets, respectively. Further analysis revealed that these transcripts unique to rmRNA-seq library are mostly long non-coding RNA, mRNA-like RNA and snoRNAs (Fig. 2 and Supplementary Fig. S3). These transcripts may lack polyA tails, such as certain snoRNAs that are transcribed by Polymerase III, and thus they are not found in the mRNA-seq library. However, as an important note, we considered that most of the rmRNAseq-specific transcripts were resulted from inadequate sampling (based on statistical measures) (Supplementary Fig. S4) since these library-specific transcripts are all expressed at very low levels. We further compared these ncRNAs to the datasets of lincRNAs published by Guttman et al. [16], and found that 101 lincRNAs can be shared by both data.

In additional to those mapped to known loci, we also identified 9428 novel transcribed loci in the rmRNA-seq dataset as compared to 4550 novel transcribed loci detected in the mRNA-seq dataset. Among them, 1218 loci are shared by both datasets and 8210 loci are unique

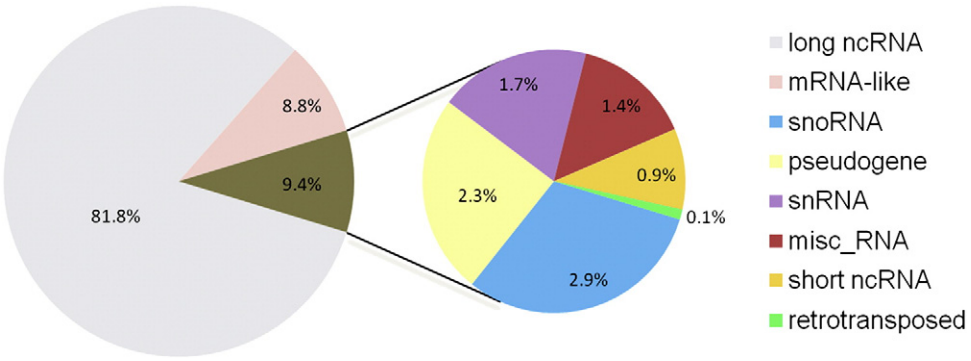

Fig. 2. Classification of the annotated transcripts uniquely identified in rmRNA-seq data. The data are divided into eight categories and color-coded: miRNA, miscRNA, snRNA, snoRNA, mRNA-like, retrotransposed, long ncRNA, and short ncRNA.

to the rmRNA-seq dataset. We went on to validate some of these loci transcribed in the mouse cerebrum by using RT-PCR assays (Supplementary Table S2 and Fig. S5). However, we cannot exclude the possibility that these novel transcripts may contain spliced or unspliced introns from unannotated genes.

2.3. Protein-coding genes

For the analysis of protein-coding genes, we calculated the transcriptional activity of genes by counting the number of reads that are mapped to all exons of individual refseq-defined gene. Using a background of hits–0.04 hit-per-kb determined with the Poisson model–and a minimal hit of five reads per gene locus (see Methods), we identified 16,532 and 16,359 active genes in the rmRNA-seq and mRNA-seq datasets, respectively. The two datasets share 15,809 genes (~96%) (Fig. 3A) and have small numbers of library-specific genes: 723 and 550 in the rmRNA-seq and mRNA-seq libraries, respectively. The limited number of library-specific genes is most likely due to sampling depth (Fig. 3B), similar to the case of the non-coding transcripts.

We also performed a correlation analysis to show that the results from the two methods are comparable (Pearson correlation coefficient is 0.817). However, there are noticeable differences between the two methods when we used a Poisson model [12]. We identified a total of 877 genes that are differentially measured ($P<0.01$), showing 4-fold changes in expression (Fig. 3C). We further looked at these genes and found some of them are attributable to differences in RNA preparations since ribo-minus RNAs contain additional NpA- and biphomic transcripts. For instance, we identify several known genes that are not polyadenylated in the rmRNA-seq library, such as *Histone*, *Heg1* [17], and *Dux* [18], and they were absent in the mRNA-seq dataset. In addition, different RNA preparation protocols are known to contribute to sampling and coverage biases. We found that the mRNA-seq method has a bias in sequence coverage across transcripts, where the sequence coverage is poor at the 5′-end (Fig. 4A and B). This is also demonstrated in a length-dependent analysis (Fig. 4C and D). Obviously, the larger transcripts have a significant bias as compared to the relatively shorter transcripts. This result suggests that the bias may be a result of truncation among polyA+ transcripts due to the use of oligo-dT affinity, and the larger transcripts are relatively more fragile (or by chance in terms of random damage) under gravity or mechanical forces as compared to the short transcripts. As a result, the loss of the 5′ portion of polyA+ transcripts affects the precise measurement of gene expression profiles. We therefore performed a correlation analysis on the expression levels of different length fractions: the first, the middle, and the last exon, and found that the two methods agree reasonably well for the 3′-portion of transcripts (i.e. the last exon) but differ in the 5′-portion (Fig. 5). We further selected 3667 genes with better sequence coverage for another correlation analysis on the expression levels and the result showed that these highly expressed genes have very similar distributions in the two datasets (Supplementary Fig. S6).

We decided to investigate the detailed causative factors as to why such a truncation occurs in the 5′-transcripts and how it influences the determination of gene expression profiles. We turned to the two publicly available mRNA-seq datasets published Nature Method [1,4]; one of them, the Cloonan dataset [1], showed obvious bias of reads coverage and its length-dependence across the transcripts (Supplementary Fig. S7). Closer investigation showed that their experiment protocol for mRNA isolation is slightly different from ours, where they used one round of ribosomal RNA depletion, using the Ribo-minus Eukaryote kit (for RNA-seq, Invitrogen, cat.10837-08) coupled with another round of mRNA isolation by using the Oligotex mRNA Mini Kit (Qiagen, cat.70042) that we used in our control protocol. It becomes clear from the data analysis that the Oligotex-assisted mRNA purification is the responsible step since the bias was not observed in the Mortazavi data [4] where the mRNA sample was prepared by using two rounds of affinity purification with oligo(dT)-Dynabeads (Invitrogen, cat.610). To be more precise, we believe that it is centrifugation (13,000rp/s) in the Oligotex mRNA Mini Kit protocol that causes the truncation of mRNAs.

2.4. Antisense transcription

For antisense transcription analysis, it is important to determine the directionality of the transcripts. By mapping the reads to the exon-junction database, we verified that ~99.99% of the junction reads are in sense orientation. We were surprised by the fact that almost all the expressed genes possess natural antisense transcripts (Supplementary Table S3) and that the poorly expressed genes tend to have more pronounced antisense transcription (Supplementary Fig. S8). The latter feature is in agreement with a recent report that antisense transcripts may be involved in regulating gene expression [19]. We further showed that antisense expression is enriched in the promoter and terminator regions of transcripts (Supplementary Fig. S9). These antisense transcripts within promoter regions have been explained to be the result of divergent transcription initiation of RNA polymerase II [20,21].

2.5. Alternative splicing

Both rmRNA-seq and mRNA-seq can be used for surveying alternative gene splicing [7,22,23]. To assess alternative splicing complexity in the mouse cerebrum, we built a database for splice junction sequences that are generated by pairwise connection of exon sequences from every RefSeq-annotated gene. From the rmRNA-seq dataset, we mapped 584,157 reads onto the junction sequences and identified 51,772 splice junctions (mapping more than two reads for each junction sequence) associated with 10,272 genes. Of these splice junctions, 99% (51,161) were also supported by EST data (known splicing variants) and the

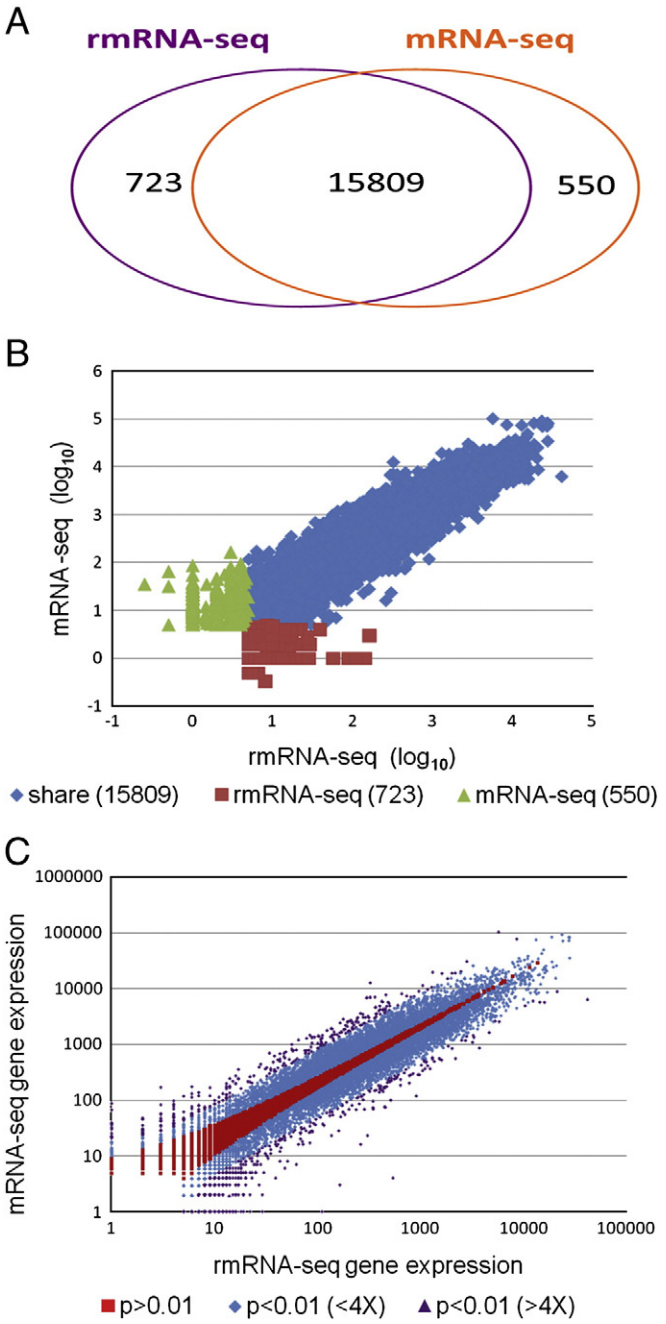

Fig. 3. A comparison of gene expression between rmRNA-seq and mRNA-seq datasets. (A) Most RefSeq defined genes detected are shared by the two datasets albeit a minor difference when >5 tags per locus are considered. (B) Gene expression is highly correlated whereas only genes expressed at the lower level exhibit minor discordance. (C) Differentially expressed genes are defined based on different *P* values.

remaining genes are defined as novel candidate splicing events. From the mRNAs-seq dataset we identified 66,754 splice junctions for 11,124 genes, based on 1,017,187 reads mapped onto junction sequences. Of these splice junctions, 99% and 1% are classified as known and novel splicing events, respectively. We also observed that exon skipping occurs most frequently between adjacent exons, with a sharp declination between distant exons (Supplementary Fig. S10).

2.6. Expression of repetitive sequence

Using our uniquely mapped reads from both libraries, we surveyed the expression of repeat content classified based on Repeatmasker (Supplementary Table S4) .We first removed the repeat elements from intronic sequences since our suspicion is that rmRNA-seq method may detect more degenerated and unspiced introns. In the datasets, we found that 3% (~50,000 repeat units) and 1.8% (~20000 repeat units) of the reads contained repeat elements from the rmRNA-seq and mRNA datasets, respectively. Therefore, the 1.2% difference between the two datasets suggested that the rmRNA-seq data may have slightly more repetitive content due to better coverage of the 5′-UTR and non-protein-coding transcripts since there are about 2 to 4 folds more LINE, LTR, and SINE elements found in the rmRNA-seq dataset.

3. Discussion

In this study, we compared two strategies–the rmRNA-seq and mRNA-seq–for mammalian transcriptomics. We showed, as compared to the traditional mRNA-seq method, that the rmRNA-seq has merits in the discovery of novel ncRNAs. On one hand, we observed higher proportion of reads from rmRNA-seq, about 44% and 25% of uniquely mappable reads, within the intergenic and intronic regions, as compared to the result of mRNA-seq, about 23% and 15% for mRNA-seq reads, respectively. This result suggested that a significant fraction of transcripts, considered to be potential NpA- or bimorphic transcripts, fails to be identified in the mRNA-seq dataset, and these NpA-transcripts are rather abundant in eukaryotic cells as high as about 80% of total transcribed sequences [15]. One the other hand, despite an obvious enrichment for non-protein coding sequences, our analysis support the notion that rmRNA-seq data are comparable to those from mRNA-seq method in transcriptome profiling since both are highly correlated. In summary, we suggest that the rmRNA-seq data are more thorough than those of mRNA-seq for systematic and thorough profiling of transcriptomes when the cost of data acquisition is affordable to the extent that most of NpA-transcripts are to be largely covered.

A frequently asked question for our protocol is whether the extra ncRNAs discovered are due to DNA contamination in the process. We have several reasons to say otherwise. First, after RNase treatment, we fractionate the fragmented RNA with a polyacrylamide gel to isolate 50- to 150-nt RNA segments. At this step, most DNA contaminations, if there is any, especially for large DNA segments, should be filtered out unless there are a large amount of small DNA fragments in the RNA preparation. Second, the ligation reaction of small RNA segments (50–150 nt) to the mosaic DNA/RNA adaptors (instead of reverse transcription with random primers) is used—SOLiD™ Small RNA Expression Kit (Ambion). At this step, only single strand RNAs are able to ligate to the adaptors not single or double strand DNAs. Third, there are also a reverse transcription and a PCR amplification procedure, where the alleged DNA contamination stays at an ignorable level. Last, in the data analysis, we used a Poisson model to assess the background (0.04 hit-per-kb), such as random transcription, and are able to differentiate transcribed genomic loci from DNA contamination. We therefore do not think that DNA contamination is an issue here as the manufacturer's instruction does not suggest DNase treatment. In addition, in this experiment design, we did not set up duplications as we have multiple libraries or tissues to study at the same time. It is why we are going through this painstaking analysis to compare the two methods. Although we used the two libraries as examples, we have done the same analysis for over 30 libraries and have not yet found extraordinary results that contradict our conclusions described in this manuscript.

Another concerned question is why many reads cannot be mapped onto genome. We believe that the remaining reads are either erroneous or low quality below our threshold but the reasons are extremely complex. For instance, our transcriptomic data are from BALB/c mouse but the genomics sequence is from C57BL/6J mouse. The sequence methodology also has a lot of problems in its detailed chemistry where enzyme fidelities and product yields all play major roles. For instance, its multiple-step enzymatic reactions include DNA

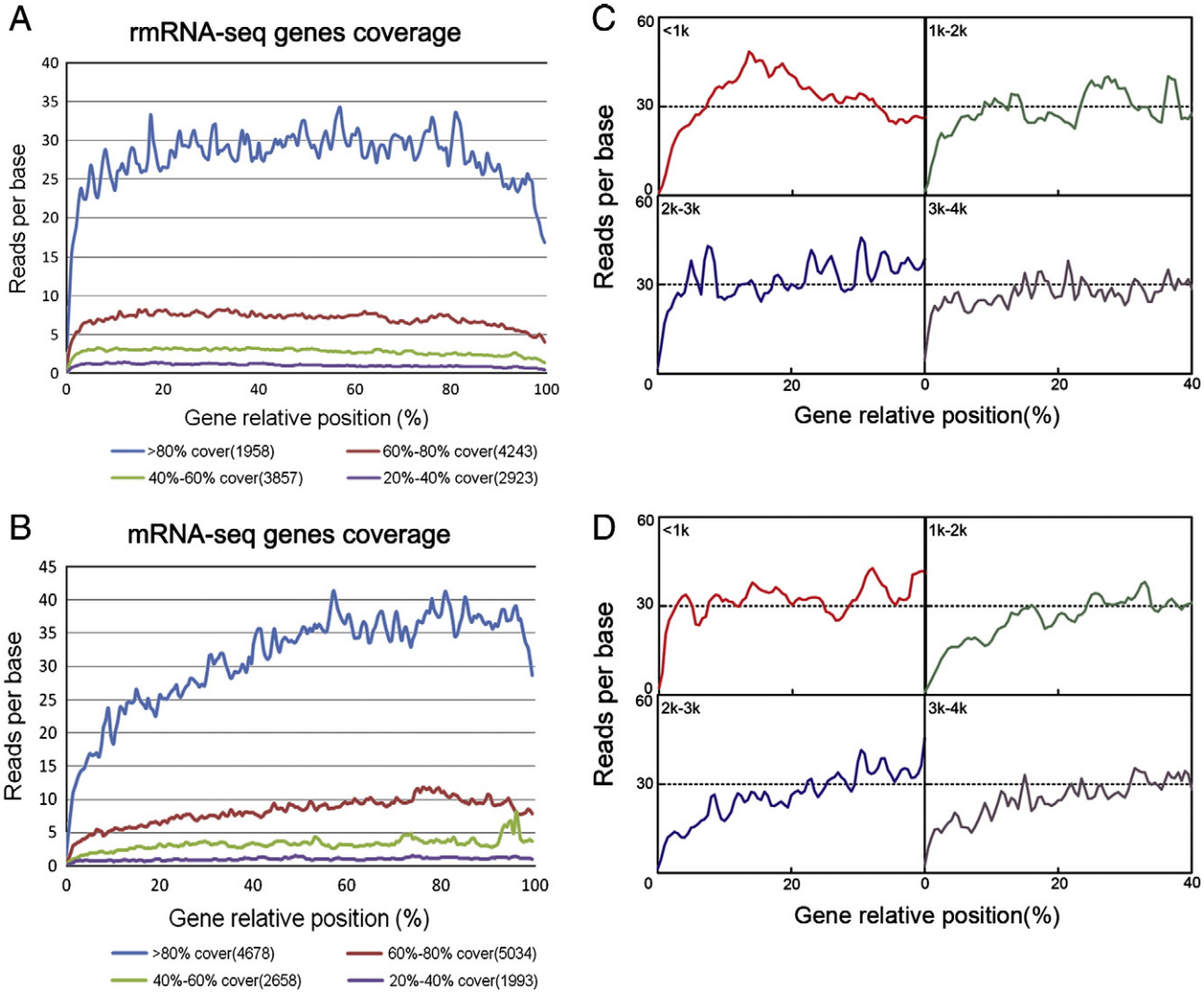

Fig. 4. Reads distribution along gene body. Relative coverage of uniquely mapped tags generated based on the rmRNA-seq (A) and mRNA-seq (B) methods. Genes with different coverage are color-coded and numbers of genes at different coverage are showed in parentheses. The coverage based on the rmRNA-seq method displays better uniformity. We also plotted the length-dependent coverage for the rmRNA-seq (C) and mRNA-seq (D) datasets. Note that larger transcripts show a stronger distribution bias in the latter dataset.

ligation, dephosphorylation and degradation. In addition, the sequencing reaction needs days to complete and the biochemical components are all time- and temperature-sensitive. Therefore, we did not try to recover more nor further analyze the unmapped reads. However, we do have ample data and experiences to estimate if the experiments are successful or not. Our current mapping rate is about 50% of the raw data in average.

4. Conclusions

In this study, we compared the two RNA-sequencing protocols, rmRNA-seq and mRNA-seq, and concluded that rmRNA-seq can detect more transcripts including protein-coding genes, ncRNAs, snoRNAs, and snRNAs. Moreover, we believe that rmRNA-seq protocol avoids 5′-truncation of mRNAs, albeit avoidable when the experiment is carefully designed, and thus gives rise to better gene expression profiles.

5. Materials and methods

5.1. Library construction

For rmRNA-seq library construction, we used Trizol for the isolation of total RNA from the cerebrum of a 10-week old adult male BALB/c mouse and the Ribo-minus Eukaryote kit (for RNA-seq, Invitrogen, cat.10837-08) for depleting ribosomal RNA. The yield is about 1 μg ribo-minus RNA out of 8 μg total RNA. For mRNA-seq library construction, we used a two-round protocol for polyA+ mRNA purification using the Oligotex mRNA Mini Kit (Qiagen, Cat ID70042).

We used the same RNA-seq library construction protocol from SOLiD™ Small RNA Expression Kit (#4397682), starting from 1 μg mRNA or rRNA-depleted RNA. Briefly, we added the following mixture on ice in order: 8 μl RNA (1 μg), 1 μl 10× RNase III buffer, and 1 μl RNase III (Applied Biosystems #AM2290). The mix was incubated at 37 °C for 10 min and at 65 °C for 20 min. We used flashPAGE™ to collect up 50 bp–150 bp fragmented RNA and purified the RNA by flashPAGE Reaction Clean-Up Kit (Applied Biosystems #AM12200). After dehydration, we re-suspended the sample with 3 μl nuclease-free water before ligation. The ligation mix is composed of: sample RNA, Hybridization Solution, and Adaptor Mix (the RNA/DNA oligonucleotides with single-stranded degenerated sequence at one end; only RNA can be linked to the adaptor); the mix was subsequently incubated with RNA ligase and buffer at 65 °C for 10 min and at 16 °C for 5 min. The ligation was carried out at 16 °C for 16 hrs. cDNA was synthesized by adding 20 μl RT Master Mix and incubating at 42 °C for 30 min. RNA residues were removed by RNase H digestion (1 μl to a 10 μl cDNA) at 37 °C for 30 min. The cDNA library was amplified, cleaned with Qiagen MinElute PCR purification Kit

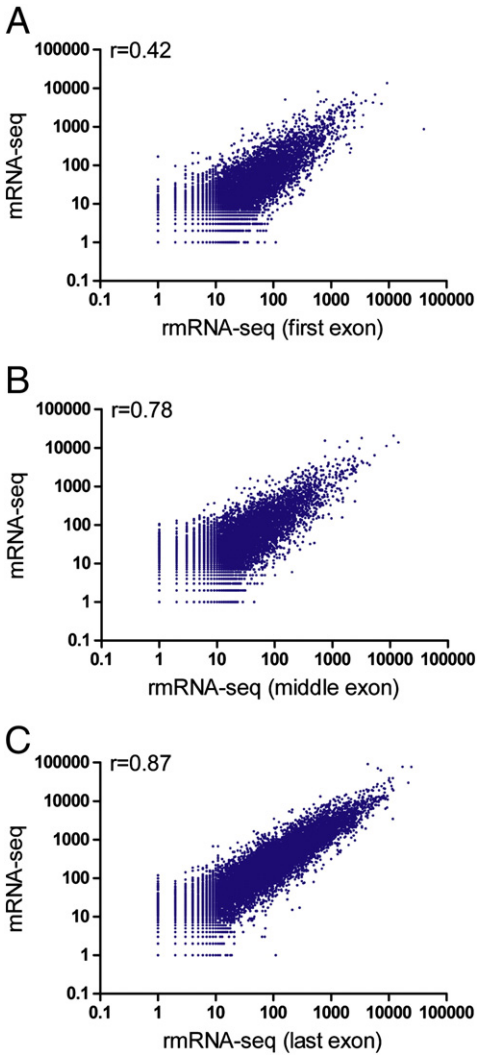

Fig. 5. The expression of different representative exon positions identified from the two datasets. We choose the genes that have more than 5 exons and at least 5 reads for each exon. The results are shown in the order of: (A) first exon, (B) middle exon, and (C) last exon.

(Qiagen #28004, 28006) and purified on a native 6% polyacrylamide gel. Usually, 400 µl reaction product (4×100 µl reactions) is enough for sequencing. A fraction of the library in a size range of 140–200 bp (DNA ladder, Invitrogen #10821-015) was used for SOLiD sequencing according to the manufacturer's instructions.

5.2. RT-PCR

We randomly selected 10 newly identified transcriptional regions and designed a set of primers for validating their actual transcription with an independent method. Total RNA from the mouse cerebrum was extracted based on the Trizol protocol (Invitrogen cat.10837-08), treated with DNAase I, and reverse-transcribed to cDNA (random priming) by using a standard protocol (SuperScript II reverse-transcriptase, Invitrogen). The condition for PCR is as follows: the initial denaturation is at 95 °C for 1.5 min; for 40 cycles the setting is 95 °C 15 s, 60 °C 15 s, and 72 °C 40 s; and the final extension is at 72 °C for 5 min.

5.3. Read alignment to the mouse genome

We retrieved the mouse reference sequence (release mm9, July 2007) from UCSC, which contains 21,896 annotated genes. We also constructed an exon–exon junction database for gene mapping. For each gene, we extracted 25-nt donor and 25-nt acceptor sequences from all possible exon–exon junction combinations so that the data not only includes normal junction sequences but also possible exon–exon junction sequences due to exon-skipping events. We mapped tags to the reference genome using Corona_lite_v4.0 in the following steps (See supplementary Fig. S1). First, we mapped the full-length, 35-bp tags to the reference; second, we analyzed the flow-through against our junction database; third, we repeated the first and the second steps for the first 30- and 25-bp truncated tags (after removals of the tag sequences beyond 30-bp and 25-bp). We used 3- and 2-mismatch options in genome and junction mapping, respectively. Tags are mapped to more than 1,000 positions are believed highly repetitive and discarded from the analysis pipeline. The mapped 35 bp and 30 bp reads were used for alternative splicing analysis. In addition, the RefSeq defined genes were used for assessing the distribution of reads in the genome.

5.4. Defining known ncRNAs and novel transcripts

To define known ncRNAs, we collected mouse ncRNA annotation data from publically available ncRNA databases that include ncRNAdb, RNAdb, fRNAdb, FANTOM3, NONCODE, Refseq, and Ensembl (Supplementary Table S5). Transcripts in the length of <50 bp are excluded from our analysis. We used blat to map these ncRNAs sequence onto the mouse genome for excluding redundancy, using a criterion of match-length/inquiry-length>=0.9 and mismatch-length/inquiry-length<0.1. In addition, transcripts mapped to more than 20 loci are considered to be repeats and discarded. Consequently, we annotated 41,878 ncRNA loci, named as "known ncRNAs", and divided them into following group: miscRNA, snRNA, snoRNA, mRNAlike, retrotransposed, long noncoding RNA (>200 bp), and short non-coding RNA(<200 bp) according to the currently available annotations (Supplementary Fig. S11).

To define novel transcripts, we collected mapped reads beyond protein-coding genes (defined by Aceview database) and known ncRNA regions, and clustered the reads based on their overlaps. The contiguous read-covered regions (more than 5 reads) are defined as candidate transcriptional locus. Furthermore, since the sequencing depth of this experiment is not enough to cover all transcribed regions, we re-clustered the defined candidate transcriptional loci with a criterion that the transcript is contiguous when a single gap between the two neighboring loci is less than 500 bp in length.

5.5. Background calculation based on a density window strategy

Since there are possibilities that random transcription and genomic DNA contamination may exist at transcriptome level, we assess the background read density from RNA-seq data using with the Poisson model. We first divide the genome into 20-kbp windows and calculate the read density in each window independently. The distribution of low-density windows fit very well by placing 9.7% of the total rmRNA-seq reads and 11% of the total mRNA-seq randomly over the mappable portion of the genome (Supplementary Fig. S12). The theoretical curve is described by

$$p(x^*l) = \frac{\lambda^{x^*l} e^{-\lambda}}{(x^*l)!},$$

where x is the read density on both strands in a unit of per base pair. l is window size (20 kb); λ (reads/20 kb) and α (reads/bp) are background read density:

$$\lambda = \alpha^* l$$

$$\alpha = \frac{f^* N_{\text{reads}}}{L_{\text{mappable}}}.$$

Here, f is the fraction of the background reads ($f = 0.09$) (Additional File 14), N_{reads} is the total number of reads aligned to the genome, and L_{mappable} (4,678,398,336) is the total number of mappable 35-bp reads in the genome summed over both strands excluding the repeat regions defined by Repeatmasker.

5.6. Quantization of gene expression

The expressiveness of genes, including ncRNAs is defined as N/L, where N is the number of coding strand reads from the transcriptional start site to the end of each gene and L is the number of mappable bases in this region. The significance of expressiveness for a given gene is determined by the probability of observing at least N reads in an interval of length L from a Poisson distribution of mean α equal to the background density estimation.

$$P = \sum_{n=N}^{\infty} \frac{(\alpha * L)^n e^{-\alpha * L}}{n!}$$

$$\alpha_{\text{mRNAseq}} = 0.00045(\text{reads/bp})$$

$$\alpha_{\text{rmRNAseq}} = 0.00055(\text{reads/bp})$$

All analyses were done with the RefSeq-defined genes and known ncRNAs. Since the density measurement for short genes (or transcripts) is not robust for our position model, we also applied a criterion that the number of reads in a transcriptional region must be more than 5.

5.7. Data availability

NCBI SRA database: SRA009022.

Supplementary materials related to this article can be found online at doi:10.1016/j.ygeno.2010.07.010.

Author contributions

PC, QL, FD, JY and SH designed the experiments. QL, CX, WG, JY performed the experiments. PC, QL, FD, LZ analyzed data. JY, SH, XY and BZ supplied the reagents and materials. PC, QL and JY wrote the manuscript. All authors read and accepted the final manuscript.

Acknowledgments

This study is supported by a grant (2006CB910401, 2006CB910403, 2006CB910404) from the National Basic Research Program (973 Program), the Ministry of Science and Technology of the People's Republic of China. The authors especially thank the LT's experts Hongying Yin, Xin Li, Jiandong Sun, Yangzhou Wang, Bob Nutter, Max Ingman for supporting on the technique and reagents of the experiments.

References

[1] N. Cloonan, A.R.R. Forrest, G. Kolle, B.B.A. Gardiner, G.J. Faulkner, M.K. Brown, D.F. Taylor, A.L. Steptoe, S. Wani, G. Bethel, Stem cell transcriptome profiling via massive-scale mRNA sequencing, Nat. Methods 5 (2008) 613–619.
[2] N. Cloonan, S.M. Grimmond, Transcriptome content and dynamics at single-nucleotide resolution, Genome Biol. 9 (2008) 234.
[3] L.D.W. Hillier, V. Reinke, P. Green, M. Hirst, M.A. Marra, R.H. Waterston, Massively parallel sequencing of the polyadenylated transcriptome of *C. elegans*, Genome Res. 19 (2009) 657.
[4] 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.
[5] U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, M. Snyder, The transcriptional landscape of the yeast genome defined by RNA sequencing, Science 320 (2008) 1344.
[6] R. Rosenkranz, T. Borodina, H. Lehrach, H. Himmelbauer, Characterizing the mouse ES cell transcriptome with Illumina sequencing, Genomics 92 (2008) 187–194.
[7] M. Sultan, M.H. Schulz, H. Richard, A. Magen, A. Klingenhoff, M. Scherf, M. Seifert, T. Borodina, A. Soldatov, D. Parkhomchuk, A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome, Science 321 (2008) 956–959.
[8] F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B.B. Tuch, A. Siddiqui, mRNA-Seq whole-transcriptome analysis of a single cell, Nat. Methods (2009).
[9] Q. Wu, Y.C. Kim, J. Lu, Z. Xuan, J. Chen, Y. Zheng, T. Zhou, M.Q. Zhang, C.I. Wu, S.M. Wang, Poly A-transcripts expressed in HeLa cells, PLoS ONE 3 (2008).
[10] R.D. Kornberg, Eukaryotic transcriptional control, Trends Genet. 15 (1999) 46–49.
[11] I. Grummt, Regulation of mammalian ribosomal gene transcription by RNA polymerase I, Prog. Nucleic Acid Res. Mol. Biol. 62 (1999) 109.
[12] S. Detke, J.L. Stein, G.S. Stein, Synthesis of histone messenger RNAs by RNA polymerase II in nuclei from S phase HeLa S3 cells, Nucleic Acids Res. 5 (1978) 1515.
[13] I.M. Willis, Genes, factors and transcriptional specificity, Eur. J. Biochem. 212 (1993) 1–11.
[14] S. Katayama, Y. Tomaru, T. Kasukawa, K. Waki, M. Nakanishi, M. Nakamura, H. Nishida, C.C. Yap, M. Suzuki, J. Kawai, H. Suzuki, P. Carninci, Y. Hayashizaki, C. Wells, M. Frith, T. Ravasi, K.C. Pang, J. Hallinan, J. Mattick, D.A. Hume, L. Lipovich, S. Batalov, P.G. Engstrom, Y. Mizuno, M.A. Faghihi, A. Sandelin, A.M. Chalk, S. Mottagui-Tabar, Z. Liang, B. Lenhard, C. Wahlestedt, R.G.N.P. Core, G.S. Grp, F. Consortium, Antisense transcription in the mammalian transcriptome, Science 309 (2005) 1564–1566.
[15] J. Cheng, P. Kapranov, J. Drenkow, S. Dike, S. Brubaker, S. Patel, J. Long, D. Stern, H. Tammana, G. Helt, Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution, Science 308 (2005) 1149–1154.
[16] M. Guttman, I. Amit, M. Garber, C. French, M. Lin, D. Feldser, M. Huarte, O. Zuk, B. Carey, J. Cassady, Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals, Nature 458 (2009) 223–227.
[17] N.J. Christensen, G. Habekost, P. Bratholm, A RNA transcript (Heg) in mononuclear cells is negatively correlated with CD14 mRNA and TSH receptor autoantibodies, Clin. Exp. Immunol. 154 (2008) 209–215.
[18] M.C. Beckers, J. Gabriels, S. van der Maarel, A. De Vriese, R.R. Frants, D. Collen, A. Belayew, Active genes in junk DNA? Characterization of DUX genes embedded within 3.3 kb repeated elements, Gene 264 (2001) 51–57.
[19] Y.P. He, B. Vogelstein, V.E. Velculescu, N. Papadopoulos, K.W. Kinzler, The antisense transcriptomes of human cells, Science 322 (2008) 1855–1857.
[20] L.J. Core, J.J. Waterfall, J.T. Lis, Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters, Science 322 (2008) 1845–1848.
[21] P. Preker, J. Nielsen, S. Kammler, S. Lykke-Andersen, M.S. Christensen, C.K. Mapendano, M.H. Schierup, T.H. Jensen, RNA exosome depletion reveals transcription upstream of active human promoters, Science 322 (2008) 1851–1854.
[22] E.T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S.F. Kingsmore, G.P. Schroth, C.B. Burge, Alternative isoform regulation in human tissue transcriptomes, Nature 456 (2008) 470–476.
[23] Q. Pan, O. Shai, L.J. Lee, B.J. Frey, B.J. Blencowe, Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing, Nat. Genet. (2008).