Biochemical and Biophysical Research Communications 422 (2012) 643–646

Contents lists available at SciVerse ScienceDirect

Biochemical and Biophysical Research Communications

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

A simple strand-specific RNA-Seq library preparation protocol combining the Illumina TruSeq RNA and the dUTP methods

Marc Sultan [a,*], Simon Dökel [a], Vyacheslav Amstislavskiy [a], Daniela Wuttig [b], Holger Sültmann [b], Hans Lehrach [a], Marie-Laure Yaspo [a]

[a] Max Planck Institute for Molecular Genetics, Ihnestr. 73, D-14195 Berlin, Germany
[b] German Cancer Research Center, and National Center for Tumor Diseases, Im Neuenheimer Feld 460, D-69120 Heidelberg, Germany

ARTICLE INFO

Article history:
Received 2 May 2012
Available online 15 May 2012

Keywords:
Next generation sequencing
RNA-Seq
TruSeq RNA
Strand-specific sequencing
Transcriptome
Illumina
dUTP

ABSTRACT

Preserving the original RNA orientation information in RNA-Sequencing (RNA-Seq) experiment is essential to the analysis and understanding of the complexity of mammalian transcriptomes. We describe herein a simple, robust, and time-effective protocol for generating strand-specific RNA-seq libraries suited for the Illumina sequencing platform. We modified the Illumina TruSeq RNA sample preparation by implementing the strand specificity feature using the dUTP method. This protocol uses low amounts of starting material and allows a fast processing within two days. It can be easily implemented and requires only few additional reagents to the original Illumina kit.

1. Introduction

RNA-Sequencing is a widely used and powerful method to explore transcriptomes [1,2]. It enables within a single experiment to investigate the expression levels and structure of transcripts without prior knowledge of the transcriptome content [3,4]. To closely reflect the original cellular RNA content of a sample, the library preparation step is crucial. A particularly relevant aspect in transcriptomics is the information of the transcript orientation, which is key for a comprehensive data analysis. It facilitates the detection of overlapping transcripts coded in opposite orientations, and allows an accurate measure of gene activity levels. Among the methods designed for keeping the original RNA strand information, the deoxy-UTP (dUTP) strand-marking protocol [5] has been rated as leading methodology enabling to identify antisense-transcription [5,6] and has been applied in various setups [7–9]. However, the dUTP-based methods published up to now are laborious and difficult to automate given that they include gel purification steps [5,8] or require time-consuming preparation and calibration of several reagents [7,9]. For these reasons, many high-throughput RNAseq production pipelines are currently using commercial kits, such as the Illumina TruSeq RNA protocol optimized for 0.1 to 4 μg total RNA, and allowing the processing in 96 microtiter plates format. However, a major drawback of the Illumina procedure is the loss of the original RNA strand information. In order to circumvent this problem, we combined here the advantages of the Illumina TruSeq protocol with that of the dUTP protocol by introducing the strand specificity feature in the Illumina method. We modified the original protocol to a simple scalable polyA + library preparation method, which is easy to implement in both small and large-scale operations. In brief, we modified a step in the Second Strand Synthesis by incorporation of dUTP instead of dTTP, which is then selectively degraded after the adapter ligation step. All the other steps of the Illumina original protocol were preserved. The library preparation procedure takes only 2 days.

2. Material and methods

We started with 0.5 μg of DNAse-treated human total RNA in the first step of the TruSeq RNA Sample Preparation v2 protocol (Illumina, Part# 15026495 Rev. A). But any amount of total RNA between 0.1 and 4 μg of total RNA, as recommended in the Illumina protocol can be used. Following reagents are required:

- TruSeq RNA sample preparation Kit v2, Set A (Illumina, #RS-122-2001).
- illustraMicroSpin G-50 Columns (GE Healthcare; #27-5330-02).
- 1 mM Tris pH 8.0 (dilution from 1 M Solution, Ambion, #AM9855G).
- Elution Buffer (Qiagen, #19086).
- 10× Reverse Transcription Buffer (Invitrogen, #53032).

* Corresponding author.
E-mail address: sultan@molgen.mpg.de (M. Sultan).

0006-291X/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.bbrc.2012.05.043

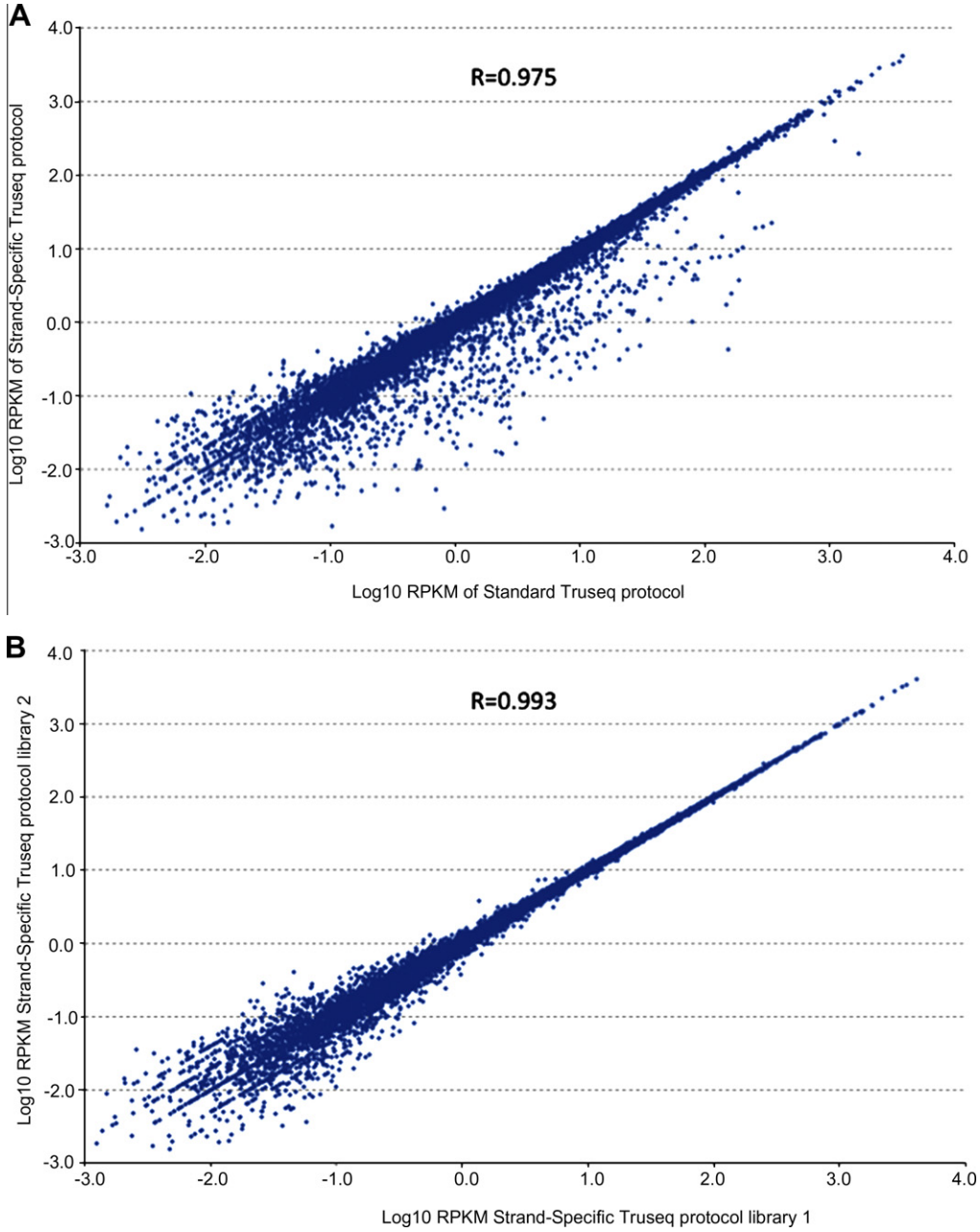

Fig. 1. RPKM scatter plot of 16,470 expressed genes in all three libraries. (A) Scatter plot of the Log10 rpkm values of standard TruSeq protocol (*x*-axis) and Strand specific TruSeq protocol (*y*-axis). (B) Scatter plot of the Log10 rpkm values of two Strand Specific TruSeq libraries. *R* is the correlation coefficient.

- 5× Second Strand Synthesis Buffer (Invitrogen, #11917-010).
- 100 mM DTT (Invitrogen, #11917-010).
- *E.Coli* DNA ligase (10 U/μl) (NEB, #M0205L).
- DNA Polymerase I (10 U/μl) (NEB, #M0209L).
- RNase H (2 U/μl) (Invitrogen, #100004927).
- 100 mM $MgCl_2$ (Dilution from 1 M Solution, Ambion, # AM9530G).
- dUNTP Mix (10 mM each dATP, dCTP, dGTP, dUTP) (Fermentas; #R0146, #R0156,#R0166, #R0133).
- UDGase (1 U/μl) (NEB, #M0280S).
- 10× UDG Buffer (NEB, #B0280S).
- RNase free water

The first steps (1) Purify and Fragment mRNA and (2) First Strand Synthesis were performed as described in the Illumina kit. Step (3) Second Strand cDNA Synthesis was modified as follows:

1. Spin illustra MicroSpin G-50 Columns at 700×*g* for 1 min.
2. Wash Columns three times with 1 mM Tris–HCl pH 8.0:
   i. Add 500 μl Tris–HCl to the column and resuspend the resin by gentle mixing.
   ii. Centrifuge column at 700×*g* for 1 min.
3. Bring the column into a 1.5 ml Low Binding tube.
4. Add 5 μl Elution Buffer to the sample.
5. Add the sample (30 μl) to the G-50 Column and spin the column at 700×*g* for 2 min.
6. Measure the volume of the eluate (should be between 30 and 50 μl).
7. Add RNAse free water to the sample up to a total volume of 52.5 μl.
8. Add Second Strand Mix (22.5 μl) to the sample:

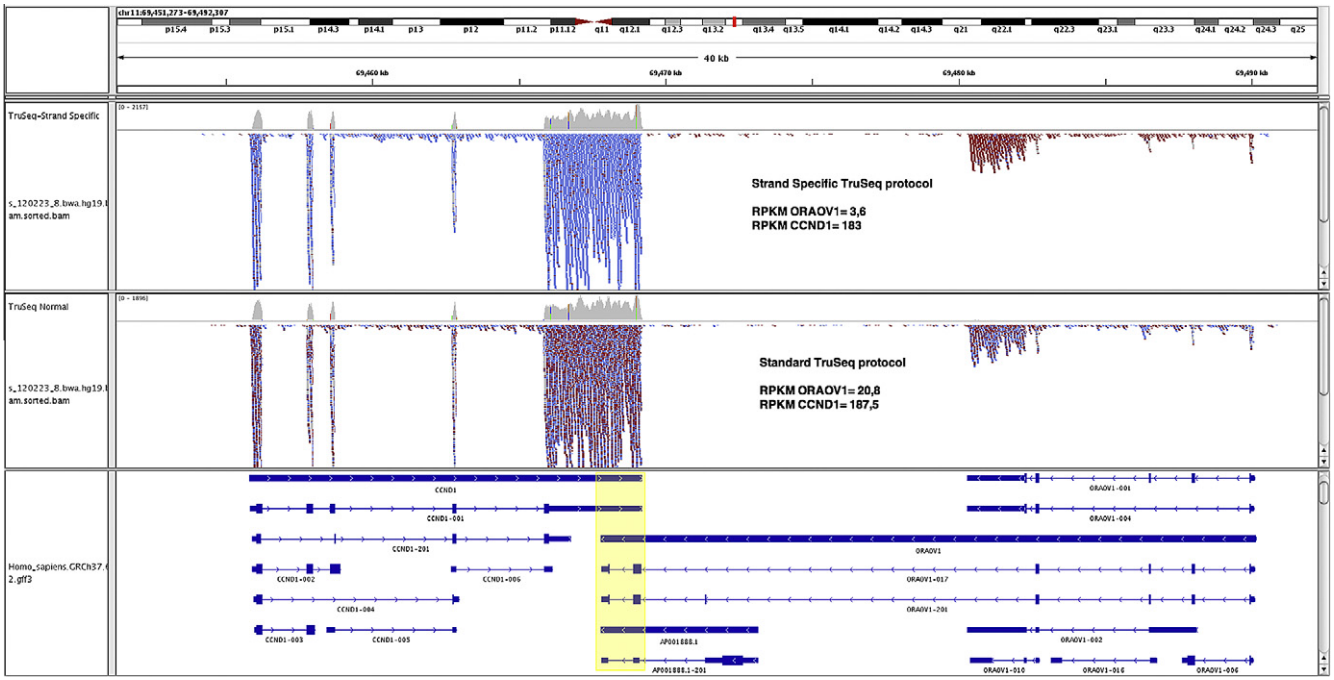

Fig. 2. Snapshot of two overlapping genes in IGV browser. This snapshot illustrates one advantage of the strand-specific method. The picture shows the read coverage of two genes (ORAOV1 and CCND1) encoded on the opposite strand of chromosome 20 and overlapping by their 3′ ends (highlighted region in the annotation track). The first track from top shows the read coverage in this region for the sample prepared with the strand-specific TruSeq protocol, the track below shows the read coverage for the standard non-strand specific TruSeq preparation method. The read alignments are colored by first-in-pair read strand, where blue corresponds to the first-in-pair read aligning to the plus DNA strand and red when the first-in-pair read to the minus DNA strand. This example illustrates that the Strand-Specific protocol enables to clearly resolve whether the sequenced reads belongs to ORAOV1 or to CCND1. Gene expression values are given in each track exemplifying the impact on RPKM values when the mapped reads can't be discriminated in overlapping gene regions.

i. Second Strand Mix: 1 μl of 10× Reverse Transcription Buffer, 15 μl of 5× Second Strand Syntheses Buffer, 0.5 μl of 100 mMMgCl, 1 μl of 100 mM DTT, 2 μl of dUNTP Mix, 0.5 μl of E. Coli DNA ligase, 2 μl of DNA Polymerase, and 0.5 μl of RNase H.

9. Incubate Mix at 16 °C for 2 h.
10. After incubation add 135 μl Ampure XP Beads to the Mix (instead of the 90 μl described in the TruSeq Protocol) and vortex the tube.
11. Continue with the TruSeq Protocol at the incubation step (15 min at RT; point 3 of page 87 of the Illumina protocol) of "Clean Up CDP" of the Second Strand Synthesis.
12. Perform End Repair, Adenylate 3′ Ends and Adapter Ligation steps as described in the TruSeq Protocol.
13. After "Adapter Ligation" and before starting "Enrich DNA Fragment step, perform the the UDGase Treatment by adding 2.3 μl of 10× UDG Buffer and 1 μl of UDGase (1 U/μl) to the sample.
14. Incubate the sample for 30 min at 37 °C.
15. Continue with the "Enrich DNA Fragment" step and follow the TruSeq protocol until the end.

3. Results and discussion

We generated one library using the standard TruSeq RNA sample protocol, and two libraries with the modified version introducing the strand specificity, starting from the same human RNA sample. The three barcoded libraries were pooled and sequenced on one lane of the HiSeq 2000 platform (75 bp, paired-end). Reads were aligned to the Human Reference genome (hg19) with bwa 0.5.9-r16 and normalized expression were calculated using the rpkm method on the Ensembl v.62 annotation [10]. We obtained on average 94 million reads per library uniquely mapped to the genome of which 91% were located in exons. Fig. 1 depicts the normalized expression levels of 16,470 genes expressed (rpkm > 0) in all three libraries compared to each other. The gene expression levels calculated for standard versus strand-specific TruSeq libraries showed very good correlation ($r = 0.975$) (Fig. 1A), but clearly reveals that a significant fraction of genes have overestimated expression values when using the non-strand specific protocol. The comparison of two strand-specific RNA libraries (Fig. 1B) demonstrates the high reproducibility of the protocol ($r = 0.993$). The strand specificity introduced in the TruSeq protocol enabled to resolve the correct expression levels of overlapping transcripts encoded on opposite strands as exemplified in Fig. 2. The calculated expression levels of e.g. ORAVO1 and CCND1, which are coded in opposite directions and overlap by their 3′ ends, is significantly biased without the sequence reads directionality (20.8 and 187.5 rpkm, respectively). When using the strand specific information the expression of both genes is lower (3.6 and 183 rpkm for ORAV1 and CCND1, respectively). This expression level correction is essential when considering that ~16% of the protein coding genes (Ensembl v62) are overlapping and thus potentially affected by biased expression values.

In conclusion we have significantly improved a commercially available RNAseq library generation protocol for polyA + RNA requiring only little amounts of input material, which is easy to use for a highly reproducible, scalable to high-throughput sample processing, and compatible with both single-end and paired-end sequencing.

Acknowledgments

This project was supported by the Max Planck society and by the German Federal Ministry of Education and Science (BMBF) in the framework of the pilotprojekt CancerSys TREAT20 (FKZ:0315852B). The authors are responsible for the content of this publication.

References

[1] M. Sultan, M.H. Schulz, H. Richard, et al., A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome, Science 321 (5891) (2008) 956–960.

[2] F. Ozsolak, P.M. Milos, RNA sequencing: advances, challenges and opportunities, Nat. Rev. Genet. 2 (2010) 87–98.

[3] V. Costa, C. Angelini, I. De Feis, et al., Uncovering the complexity of transcriptomes with RNA-Seq, J. Biomed. Biotechnol. (2010) 853916.

[4] H. Richard, M.H. Schulz, M. Sultan, et al., Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments, Nucleic Acids Res. 38 (10) (2010) e112.

[5] D. Parkhomchuk, T. Borodina, V. Amstislavskiy, et al., Transcriptome analysis by strand-specific sequencing of complementary DNA, Nucleic Acids Res. 37 (18) (2009) e123.

[6] J.Z. Levin, M. Yassour, X. Adiconi, et al., Comprehensive comparative analysis of strand-specific RNA sequencing methods, Nat. Methods 7 (9) (2010) 709–715.

[7] L. Wang, Y. Si, L.K. Dedow, et al., A low-cost library construction protocol and data analysis pipeline for Illumina-based strand-specific multiplex RNA-seq, PLoS One 6 (10) (2011) e26426.

[8] T. Borodina, J. Adjaye, M. Sultan, A strand-specific library preparation protocol for RNA sequencing, Methods Enzymol. 500 (2011) 79–98.

[9] S. Zhong, J.G. Joung, Y. Zheng, et al., High-throughput illumina strand-specific RNA sequencing library preparation, Cold Spring Harb Protoc. 8 (2011) 940–949.

[10] A. Mortazavi, B.A. Williams, K. McCue, et al., Mapping and quantifying mammalian transcriptomes by RNA-Seq, Nat. Methods 5 (7) (2008) 621–628.