scientific reports

OPEN

Poly(A) capture full length cDNA sequencing improves the accuracy and detection ability of transcript quantification and alternative splicing events

Hiroki Ura[1,2]✉, Sumihito Togi[1,2] & Yo Niida[1,2]

The full-length double-strand cDNA sequencing, one of the RNA-Seq methods, is a powerful method used to investigate the transcriptome status of a gene of interest, such as its transcription level and alternative splicing variants. Furthermore, full-length double-strand cDNA sequencing has the advantage that it can create a library from a small amount of sample and the library can be applied to long-read sequencers in addition to short-read sequencers. Nevertheless, one of our previous studies indicated that the full-length double-strand cDNA sequencing yields non-specific genomic DNA amplification, affecting transcriptome analysis, such as transcript quantification and alternative splicing analysis. In this study, it was confirmed that it is possible to produce the RNA-Seq library from only genomic DNA and that the full-length double-strand cDNA sequencing of genomic DNA yielded non-specific genomic DNA amplification. To avoid non-specific genomic DNA amplification, two methods were examined, which are the DNase I-treated full-length double-strand cDNA sequencing and poly(A) capture full-length double-strand cDNA sequencing. Contrary to expectations, the non-specific genomic DNA amplification was increased and the number of the detected expressing genes was reduced in DNase I-treated full-length double-strand cDNA sequencing. On the other hand, in the poly(A) capture full-length double-strand cDNA sequencing, the non-specific genomic DNA amplification was significantly reduced, accordingly the accuracy and the number of detected expressing genes and splicing events were increased. The expression pattern and percentage spliced in index of splicing events were highly correlated. Our results indicate that the poly(A) capture full-length double-strand cDNA sequencing improves transcript quantification accuracy and the detection ability of alternative splicing events. It is also expected to contribute to the determination of the significance of DNA variants to splicing events.

The large majority of human genes are processed at several levels, including transcriptional and post-transcriptional regulation. Alternative splicing of pre-mRNAs that include exons and introns is one of the essential regulatory mechanisms at post-transcriptional regulation[1,2]. Alternative splicing plays an important role in normal cellular and pathogenic processes caused by diverse diseases[3,4]. It has been reported that several alternative splicing events, including alternative 5' or 3' splicing site usage, exon skipping, intron retention, and mutually exclusive exons, occur in abnormal cells in various diseases and normal cells[5–7]. These alternative splicing events produce assorted mRNA that translates to different protein isoforms with different coding sequences. In normal cells, these alternative splicing events are controlled in an appropriate expression pattern[8–10]. Alternatively, an inappropriate expression pattern occurs in some human diseases, including cancers[11–13]. Therefore, it is important to accurately analyze the state of the repertoire of mRNA splicing variants and its changes associated with the pathological condition.

[1]Center for Clinical Genomics, Kanazawa Medical University Hospital, 1-1 Daigaku, Uchinada, Kahoku, Ishikawa 920-0923, Japan. [2]Division of Genomic Medicine, Department of Advanced Medicine, Medical Research Institute, Kanazawa Medical University, 1-1 Daigaku, Uchinada, Kahoku, Ishikawa 920-0923, Japan. ✉email: h-ura@kanazawa-med.ac.jp

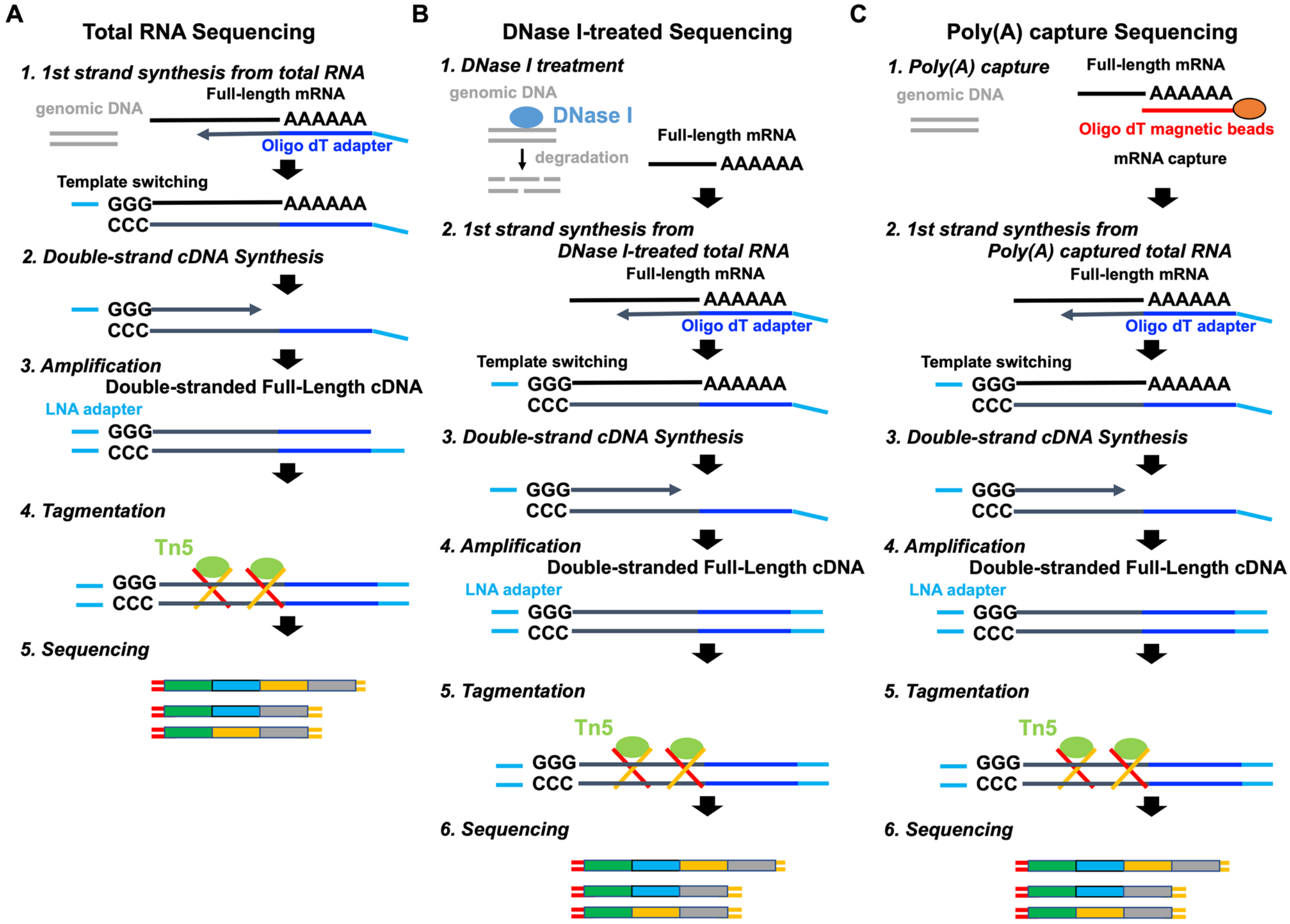

Figure 1. Library preparation workflow. (**A**) Workflow for Total RNA Sequencing. (**B**) Workflow for DNase I-treated Sequencing. (**C**) Workflow for Poly(A) capture Sequencing.

Next-generation sequencing (NGS) is a powerful technology used in the clinical field for genetic diagnosis[14–16]. The use of NGS technologies in the clinical field has led to an unprecedented increase in variants identified in different patients harboring genetic disorders. The 48% of all variants listed on ClinVar are asserted to the variant of uncertain significance (VUS) variants[17]. The current genetic counseling practice almost considers variants that directly affect protein structure. However, the VUS can affect RNA splicing, which causes protein damage. RNA splicing is expected to be disrupted by approximately 62% of all pathogenic variants[18]. It is also reported that aberrant RNA splicing affects the transcription level[19]. Thus, it is also needed to accurately measure RNA splicing and transcription level in precise genetic diagnosis.

RNA sequencing (RNA-Seq) is a powerful technology that can be used to measure not only transcriptional levels but also alternative splicing repertoire[20]. However, so far, RNA-Seq is primarily used to measure the expression level of transcripts but rarely used to detect alternative splicing variants[21,22]. Presently, RNA-Seq is almost performed using short-read sequencers, such as the Illumina NGS. Although RNA-Seq by short-read sequencer can detect alternative splicing events between two exons, short-read RNA-Seq cannot detect the full-length transcript information, including all alternative splicing events. Recently, third-generation sequencers, such as Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio), can be facilitated for alternative splicing analysis of the full-length transcript due to the production of long sequencing reads (> 10 kb)[23]. When creating a library in traditional RNA-Seq, because the captured mRNAs using oligo dT magnet beads are sheared randomly into fragments, then reverse transcribed into cDNAs, it is impossible to apply the library for analysing the full-length transcripts. Alternatively, due to no fragmentation, full-length double-stranded cDNA library can adapt to long-read sequencer for analysing the full-length transcript[24]. In addition, since the full-length cDNA is yielded by PCR amplification, a library can be prepared from a small amount of sample, even as a single cell (Fig. 1A). In principle, novel splicing variants caused by the DNA variants of non-coding regions can be directly clarified, if a DNA sequencing including non-coding regions such as a whole genome sequencing and a full-length cDNA sequencing are performed at the same time. Nevertheless, one of our previous studies indicated that the full-length double-strand cDNA sequencing resulted in non-specific genomic DNA amplification, which affects precise transcriptome analysis, such as alternative splicing and transcript quantification analysis[25]. For that reason, it is needed to eliminate genomic DNA noise when creating a full-length double-stranded cDNA library for more precise mRNA analysis.

To avoid contamination of genome DNA in Total RNA, we tried two possible methods and compared their efficiencies. One method (DNase I-treated full-length double-strand cDNA sequencing) is that genome DNA in Total RNA solution is digested enzymatically using DNase I enzyme[26] (Fig. 1B). Other method (poly(A) capture full-length double-strand cDNA sequencing) is that only messenger RNA which have Poly(A) tail are physically captured using magnetic Oligo(dT) beads[27] (Fig. 1C). It was investigated that the performance of these two methods in transcriptome analysis, such as transcript quantification analysis and alternative splicing analysis.

Methods

Cell culture.

The human induced pluripotent stem cell-line (hiPSC), strain 1383D6, which was provided by RIKEN BioResource Research Center were cultured on iMatrix 511 (Takara)-coated plates (0.5 ug/cm^2) in StemFit medium (REPROCELL) at 37 °C in 5% CO_2[28]. The cells were passaged as clump with TrypLE Select (Life Technologies) at a ratio of 1:6 every 4–5 days.

Genomic DNA extraction.

The genomic DNA sample used in this study was extracted from the whole peripheral blood using a rapid extraction method[29]. The DNA amount and optical density (A260/280 ratio) were measured using Nanodrop (Thermo Fisher Scientific, Waltham, MA, USA).

Total RNA extraction.

The total RNA was extracted from hiPSCs with TRIzol reagent (Thermo Fisher Scientific) following the manufacturer's instructions, as described previously[30,31]. RNA concentration and purity were measured spectrophotometrically (Nanodrop, Thermo Fisher Scientific). The RNA integrity number was determined using a TapeStation 4200 with High Sensitivity RNA ScreenTape (Agilent Technologies, Santa Clara, CA, USA).

Total RNA library synthesis.

Total RNA Library was synthesized from total RNA using a SMART-Seq HT kit (Takara Bio USA, Mountain View, CA, USA) and the Nextera XT DNA Library Kit (Illumina, San Diego, CA, USA), following the manufacturer's standard protocol. The library quality was further evaluated using the TapeStation 4200 with High Sensitivity D1000 ScreenTape (Agilent Technologies, Santa Clara, CA). The library was quantified using the HS Qubit dsDNA assay (Thermo Fisher Scientific, Waltham, MA, USA).

Genomic DNA library synthesis.

Genomic DNA Library was synthesized from genomic DNA using a SMART-Seq HT kit (Takara Bio USA, Mountain View, CA, USA) and the Nextera XT DNA Library Kit (Illumina, San Diego, CA, USA), following the manufacturer's standard protocol. The library quality was further evaluated using the TapeStation 4200 with High Sensitivity D1000 ScreenTape (Agilent Technologies, Santa Clara, CA). The library was quantified using the HS Qubit dsDNA assay (Thermo Fisher Scientific, Waltham, MA, USA).

DNase I-treated full-length double-stranded cDNA.

According to the manufacturer's instruction, the genomic DNA contained in total RNA was digested using TURBO DNA-free Kit (Thermo Fisher Scientific). The RNA integrity number of DNase I-treated total RNA was determined using a TapeStation 4200 with High Sensitivity RNA ScreenTape (Agilent Technologies, Santa Clara, CA, USA). According to the manufacturer's standard protocol, full-length double-stranded cDNA was synthesized from DNase I-treated total RNA using a SMART-Seq HT kit (Takara Bio USA, Mountain View, CA, USA).

Poly(A) capture full-length double-stranded cDNA.

According to the manufacturer's instruction, the mRNA was captured from total RNA with NEBNext Poly(A) mRNA Magnetic Isolation Module (New England BioLabs). Briefly, the mRNA was captured using NEBNext Magnetic Oligo $d(T)_{25}$ Beads and washed. The captured mRNA was resuspended with RNA binding buffer and was denatured at 65 °C for 5 min. Then, captured mRNA was eluted at 80 °C for 2 min. According to the manufacturer's standard protocol, full-length double-stranded cDNA was synthesized from eluted mRNA using a SMART-Seq HT kit (Takara Bio USA, Mountain View, CA, USA).

Library preparation and next-generation sequencing.

The DNase I-treated or Poly(A) capture full-length double-stranded sequencing libraries were prepared using the Nextera XT DNA Library Kit (Illumina, San Diego, CA, USA) for Illumina sequencing following the manufacturer's instructions, as described previously[32]. The library quality was further analyzed using the TapeStation 4200 with High Sensitivity D1000 ScreenTape (Agilent Technologies, Santa Clara, CA, USA). All libraries were quantified using the HS Qubit dsDNA assay (Thermo Fisher Scientific, Waltham, MA, USA). All libraries were sequenced (2 × 75 bp) using Illumina NextSeq 500 (Illumina, San Diego, CA, USA). The FASTQ files were generated using the bcl2fastq software (Illumina, San Diego, CA, USA). The FASTQ data (GSE192928) are deposited in the Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/).

Data analysis.

The FASTQ files were sampled with the same number of reads using Seqkit (version 0.13.2)[33]. The sampled FASTQ files were aligned to the reference human genome (hg38) using HISAT2 (version 2.1.0) and SAMtools (v.1.9)[34,35]. The StringTie algorithm (v.1.3.4d) was then used to assemble RNA-Seq alignment into annotated transcripts to quantify their expression[36]. The transcript expression was normalized using the transcript per million (TPM) algorithm. For differential expression analysis, the R package (edgeR) was used[37]. The

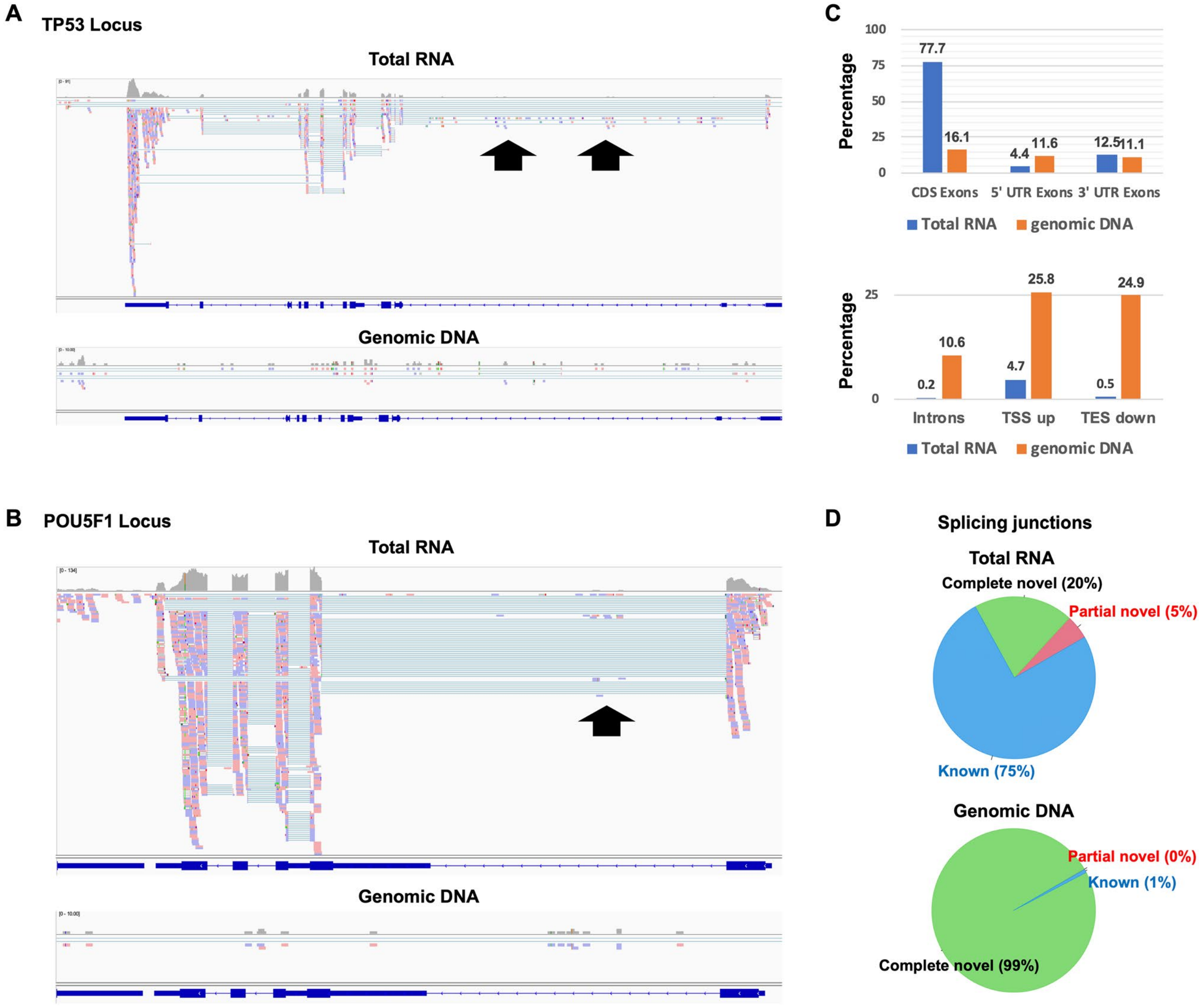

Figure 2. Confirmation of non-specific genomic DNA amplification from genomic DNA. (**A** and **B**) Integrative genomic viewer (IGV) of the TP53 and POU5F1 Locus mapped reads. Black arrows show non-specific genomic DNA amplification. (**C**) The percentage in each region (CDS Exons, 5'UTR Exons, 3'UTR Exons, Introns, TSS upstream (TSS up) and TES downstream (TES down)). (**D**) Pie chart of each splicing junction. The percentage of splicing events were calculated by RSeQC.

mapping rate was measured using HISAT2. For alternative splicing analysis, we used SplAdder software (v.2.4.2) and RSeQC (v.3.0.1)[38,39].

Ethical approval. The study was conducted according to the guidelines of the Declaration of Helsinki, and the Institutional Review Board of Kanazawa Medical University (No. G111, approved November 10, 2015) approved this study. Written informed consent was obtained by Y.N., and the ethics review board of Kanazawa Medical University approved the study design (G111).

Results

Confirmation of Non-specific Genomic DNA Amplification from Genomic DNA. First, the full-length double-stranded cDNA library was generated from only the genomic DNA to verify that non-specific amplifications are yielded from genomic DNA. The library was generated from only the genomic DNA and to be sequenced the same as the library was generated from total RNA. Comparing to Total RNA sequencing (original SMART-Seq library), the reads of genomic DNA sequencing were randomly mapped to the genomic areas in (Fig. 2A,B and Supplementary Table S1). Although, the mapped reads in intron were also seen in Total RNA sequencing as well as genome DNA sequencing (Figs. 1B and 2A , black arrows). Next, the distribution characteristics of mapped reads were investigated (Fig. 2C and Supplementary Table S1). The distribution of coding sequence exons (CDS Exons) was enriched in Total RNA sequencing. In Total RNA sequencing, almost all reads (94.6%) were mapped to exons (CDS, 5'UTR or 3'UTR exons). On the other hand, non-coding regions (Introns,

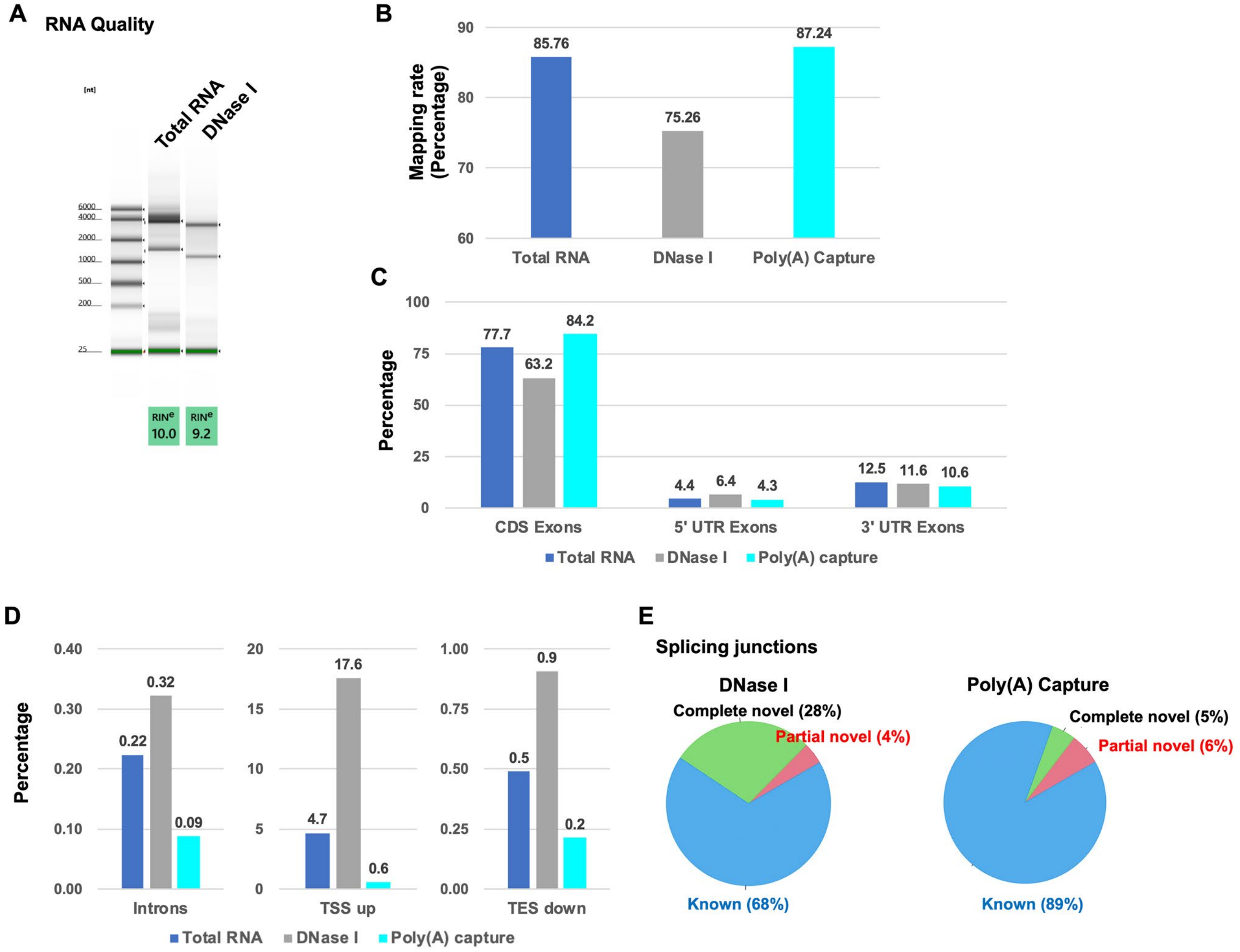

Figure 3. Comparison between Total RNA, DNase I-treated, and Poly(A) capture sequencing for splicing detection accuracy. (**A**) RNA quality in TapeStation. (**B**) Percentage of mapped reads. (**C**) The percentage in each region (CDS Exons, 5'UTR Exons, 3'UTR Exons). (**D**) The percentage in the genomic region (introns, TSS up, and TES down). (E) Pie chart of each splicing junction. The percentage of splicing events were calculated by RSeQC.

TSS upstream and TES downstream) in genomic DNA sequencing were higher than Total RNA sequencing, indicating that non-specific amplification from genomic DNA occurred. Alternatively, the distribution of each region (CDS Exons, 5'UTR Exons, 3'UTR Exons, Introns) was almost the same in genomic DNA sequencing, indicating that non-specific amplification from genomic DNA randomly occurred and was independent of sequence. The splicing junctions in Total RNA sequencing and genomic DNA sequencing were analyzed by RSeQC to assess the effect of non-specific amplification of genome DNA for alternative splicing analysis (Fig. 2D). In genomic DNA sequencing, the junctions were almost complete novel junctions (99%), showing that non-specific amplification of genome DNA may affect the detection of alternative splicing events. Although the complete novel junction in Total RNA sequencing was not as many as genome DNA sequencing (20%), the effect of non-specific amplification of genome DNA is also possible. These results suggested that it is needed to avoid contamination of genomic DNA for alternative splicing analysis.

Comparison between total RNA, DNase I-treated, and Poly(A) capture sequencing for splicing detection accuracy. To evaluate the performance of Total RNA sequencing, DNase I-treated and Poly(A) capture full-length double-stranded cDNA Sequencing, their accuracy of splicing event detection was compared. The quality of DNase I-treated total RNA was the same as DNase I-untreated total RNA (Fig. 3A). The percentage of mapping rate of Poly(A) capture sequencing was slightly higher than that of Total RNA sequencing (Fig. 3B and Supplementary Table S1). In Poly(A) capture sequencing, almost all reads (99.1%) were mapped to exons. The mapping percentage of DNase I-treated sequencing was significantly lower than that of Total RNA sequencing, indicating that non-specific genomic DNA amplification in DNase I-treated sequencing affected mapping efficiency. The distribution of CDS Exons in Poly(A) capture sequencing was also slightly higher than that of Total RNA sequencing (Fig. 3C and Supplementary Table S1). The distribution of CDS Exons in DNase I-treated sequencing was lower than that in Total RNA sequencing. The distribution of 5'UTR Exons and 3'UTR

Exons were almost the same between Total RNA, DNase I, and Poly(A) capture sequencing. The distribution of genomic DNA regions (introns, TSS up, and TES down) in Poly(A) capture sequencing was significantly lower than that in Total RNA sequencing (Fig. 3D). Unexpectedly, genomic DNA regions' distribution in DNase I-treated sequencing was higher than in Total RNA sequencing. The ratio of complete novel splicing junctions of DNase I-treated sequencing (28%) was higher than that of Total RNA sequencing (20%); the ratio in Poly(A) sequencing (5%) was lower than in Total RNA sequencing (Fig. 2D, 3E). These results suggested that Poly(A) capture sequencing improves the accuracy of splicing event detection by removing contamination of genomic DNA amplification from the library, but not DNase I treatment.

Comparison between total RNA, DNase I-treated, and Poly(A) capture sequencing for quantification analysis. To evaluate the performance of Total RNA, DNase I-treated, and Poly(A) capture sequencing, their accuracies of gene detection and expression patterns were compared. The number of detected expressing genes in Poly(A) capture sequencing was slightly higher than that in Total RNA sequencing (Fig. 4A). Alternatively, the number in DNase I-treated sequencing was lower than that in the other two sequencings. However, about 90% of detected expressing genes were commonly detected (Fig. 4B). The expression pattern between these three sequencings showed a relatively high correlation (Fig. 4C). Also, differentially expressed genes (DEG) (FDR < 0.05) were only 90 genes, showing that the expression pattern was highly correlated (Fig. 4D). The hierarchical clustering analysis indicated that the expression pattern of DEG in Total RNA sequencing was more similar to that in DNase I-treated sequencing than that in Poly(A) capture sequencing, indicating that non-specific genomic DNA amplification affected quantification of expressed genes (Fig. 4D). The genes in cluster 1 were more highly expressed in Total RNA and DNase I-treated sequencings than in Poly(A) capture sequencing. The genes in cluster 2 were vice versa. Although the cluster 2 genes were almost coding genes, cluster 1 genes were almost non-coding genes that do not have poly(A) tail (Fig. 4E). Moreover, the cluster 1 genes contained histone genes, such as HIST1H3C, which also do not have poly(A) tail. It seems that cluster 1 genes represent non-specific genomic DNA amplification. Although cluster 2 genes contain non-coding genes, these genes exist in the intragenic regions of another genes, which may represent intron retention of mRNA of another genes. These results suggested that Poly(A) sequencing improves gene detection accuracy and expression patterns.

Comparison between total RNA, DNase I-treated, and Poly(A) capture sequencing for alternative splicing analysis. Next, for alternative splicing analysis, we used SplAdder software, which has been indicated to be superior to some other software such as rMATs, SpliceGrapher and JuncBase, to detect alternative splicing events[38]. To evaluate the performance of Total RNA, DNase I-treated, and Poly(A) capture sequencing, we compared the number and accuracy of the alternative splicing events between these three sequencings. The number of alternative 5' splicing sites, alternative 3' splicing sites, and Exon skipping in Poly(A) capture sequencing was slightly higher than that in Total RNA sequencing (Fig. 5A). Alternatively, the number in DNase I-treated sequencing was fewer than Total RNA sequencing. The number of mutually exclusive exons and multi-exon skip was almost the same between these three sequencings. The number of intron retention in Poly(A) capture sequencing was relatively more than the other two sequencings (Fig. 5B). The intron retention on the *SNHG7* gene was detected using only Poly(A) capture sequencing (Fig. 5C). The percentage spliced in the index (PSI) of the splicing event, which was commonly detected, highly correlated between these three sequencings, indicating that there is no significant difference in the alternative splicing events detected by these three methods as overall transcriptome (Fig. 5D). These results suggested that Poly(A) sequencing enhances the detection number and accuracy of minor alternative splicing events especially in the sense of detecting intron retention.

Discussion

Fluctuations and regulation of mRNA splicing repertoire are associated with all aspects of biological activity. It will be more important to accurately analyze the state of the repertoire of mRNA splicing variants and its changes associated with the pathological condition to understand gene functions and life system. Next generation sequencing (NGS) is an analytical technology that can make this possible.

NGS-based applications for clinical laboratories also have been adopted as a gold standard for diagnostics of Mendelian-inherited diseases and cancers because of its analytic accuracy, high throughput, and cost-effectiveness. The VUS variants are half of all variants on Clinvar[17], and its pathological significance should be determined. Presently, most NGS -based applications in clinical diagnosis target only coding regions, such as whole-exome and gene targeting panel sequencing. Nevertheless, the variants in the intron region, sometimes even in coding region, affect RNA splicing which causes protein damage. Although variants that may change RNA splicing can be computationally predicted, it is difficult to apply to clinical diagnosis because of an incomplete understanding of alternative splicing and normal transcriptome across tissues[4]. For this reason, in addition to the information of the traditional coding regions, it is necessary to determine whether an intron variant alters RNA splicing directly to confirm its pathogenicity in clinical situations, such as an analysis for undiagnosed inherited disease or cancer gene panel.

Current short-read sequencer-based transcriptome analysis cannot characterize the full-length transcripts because of the limitations of read length. Recently, long-read sequencers, such as Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio), have advanced and are gradually used because of their ability to overcome the limitations of read length[40–43]. Although the standard RNA-Seq method for short-read sequencers has been well-established, the standard RNA-Seq method is unavailable for long-read sequencers due to fragmentation at the library preparation step. On the other hand, the full-length double-strand cDNA sequencing method is available for long-read sequencing. Given these points, in order to clarify how DNA variants in deep intron or intragenic regions affect mRNA splicing, it seems useful to perform a DNA sequencing including

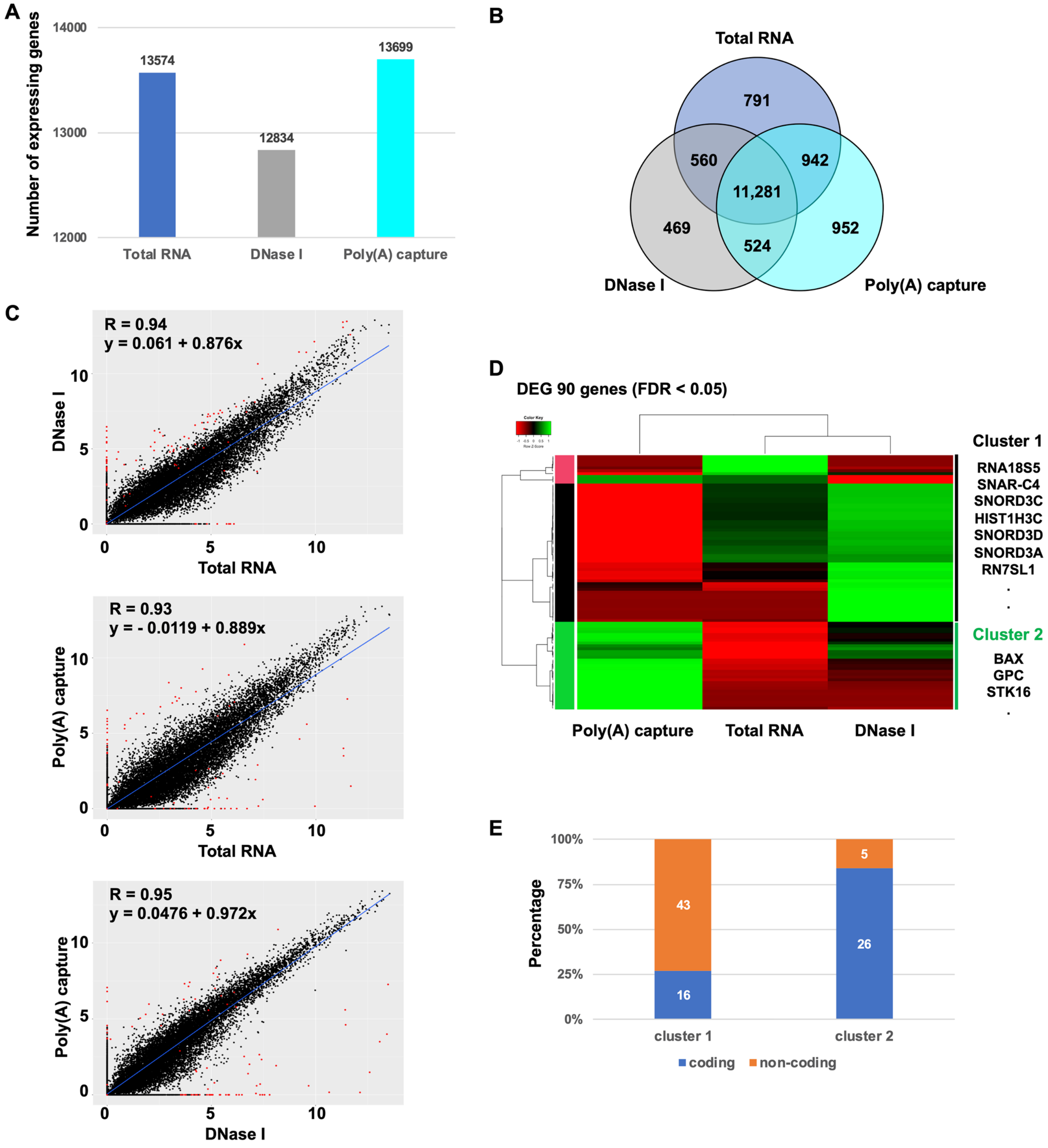

Figure 4. Comparison between Total RNA, DNase I-treated, and Poly(A) capture sequencing for quantification analysis. (**A**) The total number of expressed genes identified by these three methods. (**B**) Venn graph of expressed genes. (**C**) Scatter plot (log2 TPM (Transcripts per million)) of comparing each two methods. (**D**) Heat map of hierarchical clustering between total RNA, DNase I-treated, and Poly(A) capture sequencings. (**E**) The percentage of coding and non-coding genes in cluster 1 and 2.

non-coding regions and a full-length cDNA sequencing simultaneously with the same sample. Nevertheless, a previous study indicated that nonspecific genomic amplifications are yielded at the library preparation step and affect transcriptome analysis, such as transcript quantification and alternative splicing analysis[25]. This study also showed that the same results and non-specific genome amplification were yielded from only the genomic DNA. It is needed to eliminate contamination of genomic DNA for precise transcriptome analysis.

In this study, two methods were employed, which are the DNase I-treated full-length double-strand cDNA sequencing and poly(A) capture full-length double-strand cDNA sequencing to avoid non-specific genomic

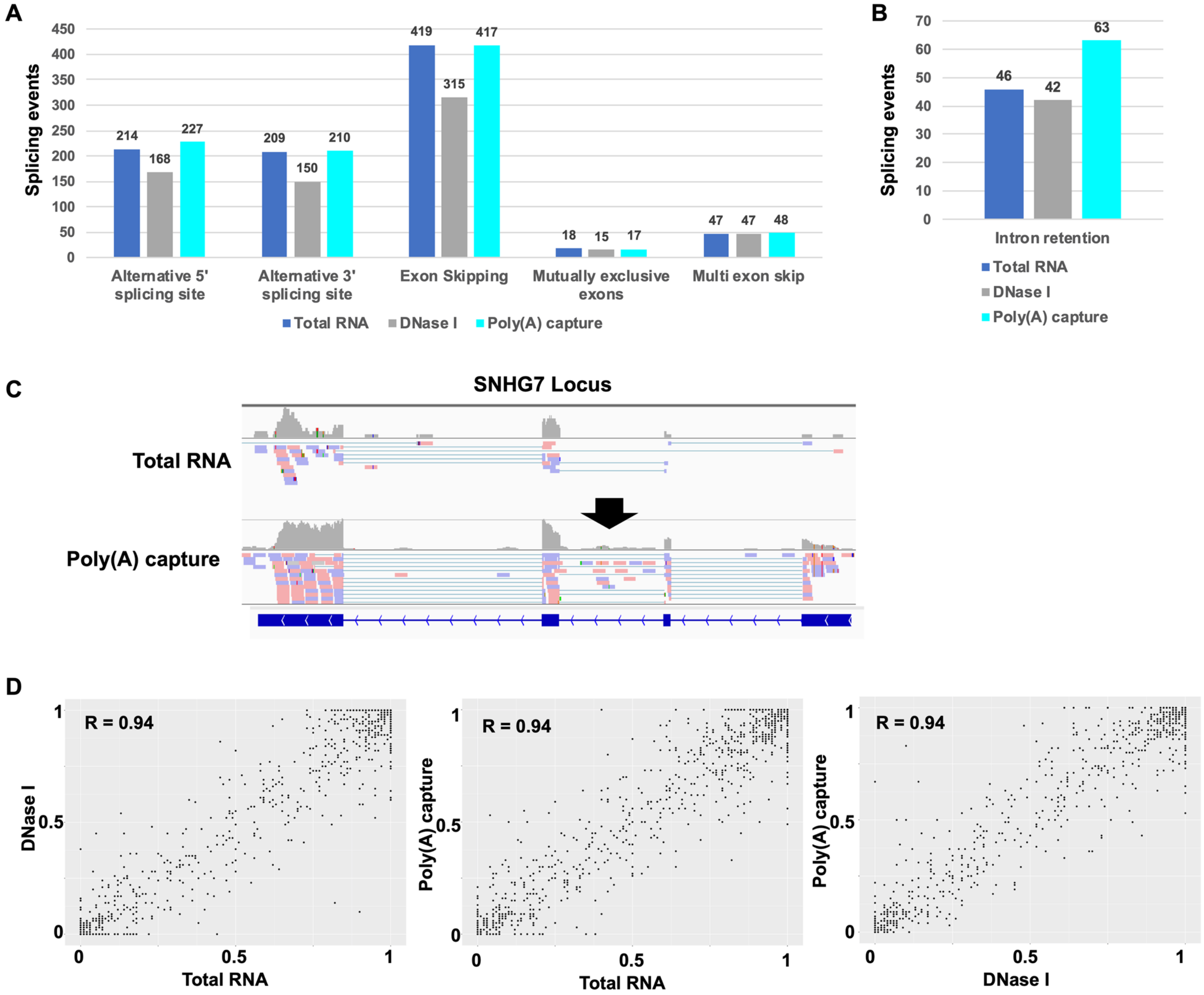

Figure 5. Comparison between Total RNA, DNase I-treated, and Poly(**A**) capture sequencings for alternative splicing analysis using SplAdder. The splicing event number per event patterns (alternative 5' splicing site, alternative 3' splicing site, Exon skipping, mutually exclusive exons, and multi-exon skip). (**B**) The number of intron retention. (**C**) The mapped reads at SNHG7 locus in the integrative genomic viewer (IGV). Black arrow indicated intron retention). (**D**) Scatter plot of the percent splicing index (PSI). The PSI of the splicing events were calculated by SplAdder. The *p*-value were calculated by t-test.

DNA amplification. Unexpectedly, non-specific genomic DNA amplification increased in the DNase I-treated sequencing than original Total RNA sequencing despite DNase I treatment. These non-specific genomic DNA amplifications affected quantification analysis and alternative splicing analysis. These non-specific genomic DNA amplifications might probably yield from the genome DNA that could not be completely digested in the DNase I-treated sequencing. It is possible that the efficiency of non-specific PCR amplification during double-strand cDNA synthesis in the SMARTer method was increased by digesting genomic DNA by DNase I and decreasing its molecular size. Alternatively, non-specific genomic DNA amplification was significantly reduced in poly(A) capture sequencing. The genes that are not actually detected, such as non-coding RNA, which does not have poly(A) tail, were detected in total RNA and DNase I-treated sequencings due to non-specific amplification noise. On the other hand, non-coding RNAs were undetected correctly in poly(A) capture sequencing. Complete novel splicing junctions were dramatically reduced in Poly(A) capture sequencing (Fig. 3E) comparing to the original Total RNA sequencing (Fig. 2D). That is, it was shown that many of the novel splicing variants detected in the total RNA sequence are artifacts. Moreover, alternative splicing events including intron retentions were more accurate and more prevalent in poly(A) capture sequencing.

Our poly(A) capture full-length double-strand cDNA sequencing improves the accuracy and detection ability of transcript quantification and alternative splicing events owing to the elimination of non-specific amplification noise from genomic DNA. Poly (A) capture sequences could be able to more accurately detect minor splicing events that occur in certain genes associated with cell differentiation or pathogenicity. Also, Poly(A) capture sequencing will provide information on all splice events in full-length transcripts because a highly accurate library created by this method can be applied to long-read sequencings, such as ONT and PacBio. Moreover, by combining with whole genome sequencing, this method will contribute to determine the VUS variants' pathological significance that affects splicing events in clinical diagnosis.

Data availability

The datasets have been deposited in the Gene Expression Omnibus (GEO) database. (GSE192928).

Received: 23 February 2022; Accepted: 14 June 2022
Published online: 22 June 2022

References

1. Shi, Y. Mechanistic insights into precursor messenger RNA splicing by the spliceosome. *Nat. Rev. Mol. Cell Biol.* **18**, 655–670. https://doi.org/10.1038/nrm.2017.86 (2017).
2. Kornblihtt, A. R. *et al.* Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nat. Rev. Mol. Cell Biol.* **14**, 153–165. https://doi.org/10.1038/nrm3525 (2013).
3. Gamazon, E. R. & Stranger, B. E. Genomics of alternative splicing: evolution, development and pathophysiology. *Hum. Genet.* **133**, 679–687. https://doi.org/10.1007/s00439-013-1411-3 (2014).
4. Park, E., Pan, Z., Zhang, Z., Lin, L. & Xing, Y. The expanding landscape of alternative splicing variation in human populations. *Am. J. Hum. Genet.* **102**, 11–26. https://doi.org/10.1016/j.ajhg.2017.11.002 (2018).
5. Keren, H., Lev-Maor, G. & Ast, G. Alternative splicing and evolution: diversification, exon definition and function. *Nat. Rev. Genet.* **11**, 345–355. https://doi.org/10.1038/nrg2776 (2010).
6. Alekseyenko, A. V., Kim, N. & Lee, C. J. Global analysis of exon creation versus loss and the role of alternative splicing in 17 vertebrate genomes. *RNA New York N.Y.* **13**, 661–670. https://doi.org/10.1261/rna.325107 (2007).
7. Sugnet, C. W., Kent, W. J., Ares, M. Jr. & Haussler, D. Transcriptome and genome conservation of alternative splicing events in humans and mice. *Pac. Symp. Biocomput.* https://doi.org/10.1142/9789812704856_0007 (2004).
8. Llorian, M. *et al.* The alternative splicing program of differentiated smooth muscle cells involves concerted non-productive splicing of post-transcriptional regulators. *Nucleic Acids Res.* **44**, 8933–8950. https://doi.org/10.1093/nar/gkw560 (2016).
9. Martinez, N. M. *et al.* Alternative splicing networks regulated by signaling in human T cells. *RNA (New York, N.Y.)* **18**, 1029–1040. https://doi.org/10.1261/rna.032243.112 (2012).
10. Giudice, J. *et al.* Alternative splicing regulates vesicular trafficking genes in cardiomyocytes during postnatal heart development. *Nat. Commun.* **5**, 3603. https://doi.org/10.1038/ncomms4603 (2014).
11. Brinkman, B. M. Splice variants as cancer biomarkers. *Clin. Biochem.* **37**, 584–594. https://doi.org/10.1016/j.clinbiochem.2004.05.015 (2004).
12. Srebrow, A. & Kornblihtt, A. R. The connection between splicing and cancer. *J. Cell Sci.* **119**, 2635–2641. https://doi.org/10.1242/jcs.03053 (2006).
13. Venables, J. P. Aberrant and alternative splicing in cancer. *Can. Res.* **64**, 7647–7654. https://doi.org/10.1158/0008-5472.Can-04-1910 (2004).
14. Hartman, P. *et al.* Next generation sequencing for clinical diagnostics: Five year experience of an academic laboratory. *Mol Genet. Metab. Rep.* **19**, 100464. https://doi.org/10.1016/j.ymgmr.2019.100464 (2019).
15. Voelkerding, K. V., Dames, S. & Durtschi, J. D. Next generation sequencing for clinical diagnostics-principles and application to targeted resequencing for hypertrophic cardiomyopathy: a paper from the 2009 William Beaumont Hospital Symposium on Molecular Pathology. *J. Molecul. Diagnost. JMD* **12**, 539–551. https://doi.org/10.2353/jmoldx.2010.100043 (2010).
16. Meldrum, C., Doyle, M. A. & Tothill, R. W. Next-generation sequencing for cancer diagnostics: A practical perspective. *Clin. Biochem. Rev.* **32**, 177–195 (2011).
17. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062-d1067. https://doi.org/10.1093/nar/gkx1153 (2018).
18. López-Bigas, N., Audit, B., Ouzounis, C., Parra, G. & Guigó, R. Are splicing mutations the most frequent cause of hereditary disease?. *FEBS Lett.* **579**, 1900–1903. https://doi.org/10.1016/j.febslet.2005.02.047 (2005).
19. Fackenthal, J. D. & Godley, L. A. Aberrant RNA splicing and its functional consequences in cancer cells. *Dis. Model. Mech.* **1**, 37–42. https://doi.org/10.1242/dmm.000331 (2008).
20. Byron, S. A., Van Keuren-Jensen, K. R., Engelthaler, D. M., Carpten, J. D. & Craig, D. W. Translating RNA sequencing into clinical diagnostics: Opportunities and challenges. *Nat. Rev. Genet.* **17**, 257–271. https://doi.org/10.1038/nrg.2016.10 (2016).
21. Zhao, W. *et al.* Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. *BMC Genomics* **15**, 419. https://doi.org/10.1186/1471-2164-15-419 (2014).
22. Barrett, A. *et al.* (2021). A head-to-head comparison of ribodepletion and polyA selection approaches for C. elegans low input RNA-sequencing libraries. *G3 (Bethesda)*. https://doi.org/10.1093/g3journal/jkab121
23. Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345. https://doi.org/10.1038/nbt.4060 (2018).
24. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181. https://doi.org/10.1038/nprot.2014.006 (2014).
25. Ura, H., Togi, S. & Niida, Y. Target-capture full-length double-strand cDNA sequencing for alternative splicing analysis. *RNA Biol.* https://doi.org/10.1080/15476286.2021.1872961 (2021).
26. Green, M. R. & Sambrook, J. Removing DNA contamination from RNA samples by treatment with RNase-Free DNase I. *Cold Spring Harb. Protoc.* https://doi.org/10.1101/pdb.prot101725 (2019).
27. Green, M. R. & Sambrook, J. Isolation of Poly(A)(+) messenger RNA using magnetic Oligo(dT) beads. *Cold Spring Harb Protoc.* https://doi.org/10.1101/pdb.prot101733 (2019).
28. Nakagawa, M. *et al.* A novel efficient feeder-free culture system for the derivation of human induced pluripotent stem cells. *Sci. Rep.* **4**, 3594. https://doi.org/10.1038/srep03594 (2014).
29. Lahiri, D. K. & Schnabel, B. DNA isolation by a rapid method from human blood samples: effects of MgCl2, EDTA, storage time, and temperature on DNA yield and quality. *Biochem. Genet.* **31**, 321–328. https://doi.org/10.1007/bf02401826 (1993).
30. Ura, H., Togi, S. & Niida, Y. Targeted double-stranded cDNA sequencing-based phase analysis to identify compound heterozygous mutations and differential allelic expression. *Biol. (Basel)* https://doi.org/10.3390/biology10040256 (2021).
31. Togi, S., Ura, H. & Niida, Y. Optimization and validation of multi-modular long-range PCR-based next-generation sequencing assays for comprehensive detection of mutation in tuberous sclerosis complex. *J. Molecul. Diagn.* (In press).
32. Ura, H., Togi, S. & Niida, Y. Dual deep sequencing improves the accuracy of low-frequency somatic mutation detection in cancer gene panel testing. *Int. J. Molecul. Sci.* https://doi.org/10.3390/ijms21103530 (2020).
33. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS ONE* **11**, e0163962. https://doi.org/10.1371/journal.pone.0163962 (2016).
34. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360. https://doi.org/10.1038/nmeth.3317 (2015).

35. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinf. (Oxford, England)* **25**, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352 (2009).
36. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295. https://doi.org/10.1038/nbt.3122 (2015).
37. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinf. (Oxford, England)* **26**, 139–140. https://doi.org/10.1093/bioinformatics/btp616 (2010).
38. Kahles, A., Ong, C. S., Zhong, Y. & Rätsch, G. SplAdder: identification, quantification and testing of alternative splicing events from RNA-Seq data. *Bioinf. (Oxford, England)* **32**, 1840–1847. https://doi.org/10.1093/bioinformatics/btw076 (2016).
39. Wang, L., Wang, S. & Li, W. RSeQC: Quality control of RNA-seq experiments. *Bioinf. (Oxford, England)* **28**, 2184–2185. https://doi.org/10.1093/bioinformatics/bts356 (2012).
40. Rhoads, A. & Au, K. F. PacBio sequencing and its applications. *Genomics Proteom. Bioinf.* **13**, 278–289. https://doi.org/10.1016/j.gpb.2015.08.002 (2015).
41. Pollard, M. O., Gurdasani, D., Mentzer, A. J., Porter, T. & Sandhu, M. S. Long reads: their purpose and place. *Hum. Mol. Genet.* **27**, R234-r241. https://doi.org/10.1093/hmg/ddy177 (2018).
42. Tilgner, H. *et al.* Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. *Nat. Biotechnol.* **33**, 736–742. https://doi.org/10.1038/nbt.3242 (2015).
43. Oikonomopoulos, S., Wang, Y. C., Djambazian, H., Badescu, D. & Ragoussis, J. Benchmarking of the Oxford Nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations. *Sci. Rep.* **6**, 31602. https://doi.org/10.1038/srep31602 (2016).

Acknowledgements

We thank the members of the Center for Clinical Genomics at Kanazawa Medical University Hospital for helpful discussion and feedback on this manuscript. This work is supported by Kanazawa Medical University (No. 11181, 26699).

Author contributions

H.U. designed and planned the experiments; Y.N. supervised the experiments; H.U. S.T. and Y.N. performed the experiments; H.U. analyzed data; H.U prepared figures and H.U. and Y.N. wrote the manuscript. All authors read and approved the final manuscript.

Funding

This study was funded by Kanazawa Medical University (No. 11181, 26699).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-14902-7.

Correspondence and requests for materials should be addressed to H.U.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.