Published in final edited form as:

J Microbiol Methods. 2018 November ; 154: 55–62. doi:10.1016/j.mimet.2018.10.008.

Comparative Evaluation of cDNA Library Construction Approaches for RNA-Seq Analysis from Low RNA-Content Human Specimens

T.L. Masters¹, C.A. Hilker^{2,3}, P.R. Jeraldo^{3,4}, A.V. Bhagwate⁵, K.E. Greenwood-Quaintance¹, B.W. Eckloff², N. Chia^{3,4}, A.D. Hanssen⁶, M.P. Abdel⁶, J.Z. Yao³, J. Jen², and R. Patel^{1,7}

¹Division of Clinical Microbiology, Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, Minnesota

²Genome Analysis Core, Mayo Clinic, Rochester, Minnesota

³Center for Individualized Medicine, Mayo Clinic, Rochester, Minnesota

⁴Department of Surgery, Mayo Clinic, Rochester, Minnesota

⁵Department of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, Minnesota

⁶Department of Orthopedic Surgery, Mayo Clinic, Rochester, Minnesota

⁷Division of Infectious Diseases, Department of Medicine, Mayo Clinic, Rochester, Minnesota

Abstract

With the emergence of RNA sequencing technologies, metatranscriptomic studies are rapidly gaining attention as they simultaneously provide insight into gene expression profiles and therefore disease association pathways of microbial pathogens and their hosts. This approach, therefore, holds promise for applicability in infectious disease diagnostics. A challenge of this approach in the clinical setting is the low amount and quality of RNA, especially microbial RNA in most clinically-infected specimens. Here, we compared two commercially available stranded cDNA library preparation kits, the NuGEN Ovation SoLo RNA-Seq System and the Illumina TruSeq Stranded Total RNA, using RNA extracted from synovial and sonicate fluids from a subject with periprosthetic joint infection. The Ovation SoLo RNA-Seq System provided more useful transcriptomic data for the infecting bacterium, whereas the TruSeq Stranded Total RNA kit provided more useful human transcriptomic data.

Keywords

RNA sequencing; periprosthetic joint infection; cDNA library preparation; gene expression

Corresponding author: Robin Patel, M.D., Division of Clinical Microbiology, Mayo Clinic, 200 1st St SW, Rochester, Minnesota 55905, patel.robin@mayo.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

1. Introduction

Total RNA sequencing, or RNA-Seq, is a sequencing technique that captures the entire transcriptome, including low abundance transcripts, allowing assessment of subtle changes in expression profiles under varying conditions. Compared to other expression assays, such as microarrays, RNA-Seq is unbiased and transcript agnostic, allowing discovery of novel transcripts (Zhao, Fung-Leung, Bittner, Ngo, & Liu, 2014). In clinical specimens, this technique can be used to detect microorganisms and to study their pathogenesis by assaying their gene expression profiles contemporaneously with those of their human hosts. Simultaneous analyses of molecular changes in pathogens and their infected hosts enhance our understanding of their interactions (Westermann et al., 2012; Westermann et al., 2017). RNA-Seq has been slowly implemented for diagnosis of human diseases, including developmental disorders (Gonorazky et al., 2016; Magri et al., 2011), neurodegenerative disorders (Campello & Martín-Nieto, 2013), and cancers (Zhang et al., 2015). Applying RNA-Seq-based analyses in infectious diseases is in its infancy, as efforts are being made to establish and standardize methods (Byron et al., 2016).

One major hurdle as applied to bacterial infections is the tremendous difference in nature, content, and amount of RNA between bacterial and host cells. In a typical infection setting, host RNA is overwhelmingly abundant compared to bacterial RNA (Westermann et al., 2017). In addition, for both prokaryotic and eukaryotic RNA, ribosomal RNA (rRNA) abounds. The challenge can be minimized by depleting bacterial and host rRNA prior to cDNA library construction, thus increasing the information content of messenger RNA (mRNA) transcripts and preventing wasted resources through sequencing of uninformative rRNAs (Humphrys et al., 2013; Westermann et al., 2016). There are several approaches and commercially available kits for rRNA depletion, each yielding different efficiencies (Chen & Duan, 2011; Giannoukos et al., 2012; Yi et al., 2011). Some kits employ a subtractive hybridization method that uses sequence-specific oligonucleotides bound to magnetic beads [e.g., RiboMinus (Thermo Fisher Scientific), Ribozero (Illumina)], whereas others de-emphasize rRNA through cDNA synthesis using a set of short selected primers whose sequences do not match those of rRNA transcripts, thereby selectively enriching non-rRNAs [e.g., Ovation Prokaryotic RNA-Seq System (NuGEN Technologies)] (Armour et al., 2009; Giannoukos et al., 2012; Westermann et al., 2012). Metatranscriptomic sequencing analysis in the clinical setting is challenged by another issue - the low amount of RNA in certain subsets of clinical samples, hindering preparation of sufficient cDNA libraries. To overcome this barrier, whole transcriptome amplification (WTA) may be applied to prepare cDNA libraries from a minute amount of input RNA; this approach is used by the Ovation RNA-Seq System (NuGEN Technologies) (Tariq, Kim, Jejelowo, & Pourmand, 2011), and the TransPlex WTA kit (Rubicon Genomics) (Tomlins et al., 2006). The drawback of such an amplification approach is that it does not necessarily represent the original transcriptome as biased gene expression profiles may occur (Coenen, Liedtke, & Kogler, 2015). More importantly, WTA causes loss of transcript strandedness which informs the orientation of sense and antisense RNAs of genes in overlapping regions, making it difficult to accurately align transcripts and quantify gene expression levels (Zhao et al., 2015).

In this study, two strand-specific cDNA library construction kits, the Ovation SoLo RNA-Seq System (NuGEN Technologies) and the TruSeq Stranded Total RNA (Illumina), were compared for their ability to generate informative cDNA libraries for RNA-Seq from samples from a patient with periprosthetic joint infection, an infection-type with a low microbial burden.

2. Materials and Methods

2.1. RNA-Seq experiment workflow

An overview of the experiments performed is summarized in Figure 1. Clinical specimens were used to isolate total RNA which was treated with DNase and underwent cDNA synthesis and library generation using the two library construction kits. The resulting library sets were separately sequenced and subsequently analyzed for microbial and human RNA.

2.2. Samples

Sonicate and synovial fluid samples from the same subject who was diagnosed as having periprosthetic joint infection were studied; the samples were collected in 2016 under the Mayo Clinic Institutional Review Board protocols 09–000808 and 10–005574, respectively. Sonicate fluid was prepared using a previously described process of vortexing and sonication (Cazanave et al., 2013) of the subject's resected prosthetic knee components in Ringer's solution. Synovial fluid, a fluid in the joint space that helps lubricate the joint, was aspirated 7 days prior to revision total knee arthroplasty. There is a difference in the nature of these two sample types. The sonication process has been shown to dislodge bacteria from biofilms on implant surfaces and disaggregate them; thus sonicate samples potentially harbor more bacteria than does synovial fluid. Further, the organisms in sonicate fluid are more likely to be in a biofilm state than are those in synovial fluid. Cultures revealed *Streptococcus sanguinis* with greater than 100 colony forming units/10 ml in sonicate fluid and growth after one day from a blood culture bottle inoculated with synovial fluid. Bacterial density is not known for synovial fluid, given that culture was performed using a blood culture bottle. The fluids were stored at -80°C without a RNA stabilizer and processed for RNA isolation 13 months later.

2.2 RNA preparation and treatment

Total RNA from human and bacterial cells was extracted using the miRNeasy Serum/Plasma kit (QIAGEN, Germantown, Maryland, USA) according to its manufacturer's instructions and treated with RNase-free DNase I (QIAGEN) following purification. DNase-treated RNA was purified with the RNeasy MinElute Cleanup kit (QIAGEN) to remove the DNase buffer and enzyme and to concentrate RNA. RNA quantity and integrity were determined using a Qubit[®] 2.0 Fluorometer (Thermo Fisher Scientific, Coon Rapids, Minnesota, USA) with the Qubit RNA high-sensitive assay kit, and by the 4200 TapeStation system (Agilent, Santa Clara, California, USA).

2.3 Construction of next generation sequencing (NGS) cDNA libraries and RNA sequencing

Total RNA, including microbial and human RNA, was subjected to cDNA synthesis and NGS library construction using the two library kits, the TruSeq Stranded Total RNA (Illumina, San Diego, California, USA) and the Ovation SoLo RNA-Seq System (NuGEN Technologies, San Carlos, California, USA). The TruSeq Stranded Total RNA kit requires an input of 100 ng – 1 µg and includes rRNA depletion using a set of human-specific probes prior to the first strand cDNA synthesis. The Ovation SoLo RNA-Seq System allows lower input RNA, 10 pg – 10 ng, for initial cDNA synthesis, followed by cleavage of insertdependent adaptors specific for human rRNA transcripts. We used 500 ng total RNA as the starting material for Illumina TruSeq Stranded Total RNA. Illumina libraries were sequenced on an Illumina HiSeq 2500 in rapid mode with 2×250 base paired end reads on one lane. NuGEN Ovation SoLo RNA-Seq System cDNA libraries were prepared from 10 ng input RNA along with a blank negative control and were run on a HiSeq 4000 sequencer with 2×100 base reads, multiplexed with five other samples (unrelated to this study) in one lane.

2.4 Computational analysis of whole transcriptome sequencing data

Paired-end sequencing reads obtained from the tested libraries were analyzed for microbial and human transcriptomes. For human RNA analysis, reads were processed using the Mayo Clinic-developed pipeline MAP-RSeq (version 3.0) which provides gene count, exon count, expressed single nucleotide variants, mapping statistics, and visualization (Kalari et al., 2014). Prior to executing MAP-RSeq, reads were trimmed to remove adaptor sequences. Human mapped reads were also processed through the Ingenuity® pathway analysis (QIAGEN) to predict pathways with which the most expressed genes are associated. For microbial transcriptome analysis, raw sequencing reads underwent a series of processes to identify microbes in the samples as previously described by Thoendel et al. (Thoendel et al., 2016). This process began with adapter trimming and filtering using Trimmomatic v0.32 (Bolger, Lohse, & Usadel, 2014), human read filtering with BioBloom tools v2.0.12 (Chu et al., 2014), and taxonomy calling by the Livermore Metagenomics Analysis Toolkit (LMAT) (Ames et al., 2013). LMAT utilizes a k-merbased approach to classify reads, and we specifically used the marker library kML+Human.v4–14.20.g10.db to create the taxonomy profiles. Taxonomy ID's specificity for a read is determined based on the read score (match score) which is derived from the proportion of k-mers of the read under the taxonomy node normalized by the proportion of k-mers of a random read under the same taxonomy node. Rockhopper 2 (Tjaden, 2015) was employed to analyze gene expression profiles of the reads mapped to *S. sanguinis*. The raw RNA sequencing data has been deposited in the NCBI Sequence Read Archive (SRA) under the BioProject ID PRJNA478235.

3 Results

3.1 Quantity and quality of total RNA and cDNA libraries

Following RNA isolation and DNase treatment, total RNA from the synovial and sonicate fluid samples was checked for quantity and quality (Table 1). As expected, RNA was found in a relatively low amount (total of 1.3 and 1.7 µg RNA for synovial and sonicate fluid,

respectively) and was degraded (RNA Integrity Number < 2.0; DV₂₀₀ < 55%), possibly since the samples had been stored for over a year. Although different amounts of the input RNAs were utilized in each library preparation kit, we observed that both kits produced similar amounts and fragment sizes of DNA libraries (Table 1). Through Illumina paired-end sequencing, we generated an average of 346.5 million reads and 94.5 million reads from the TruSeq Stranded Total RNA and Ovation SoLo RNA-Seq System libraries, respectively (Table 2).

3.2 Microbial RNA analysis

For bacterial expression analysis, cleaned reads underwent filtering of human-derived reads and were assigned taxonomy by LMAT to confirm the presence of the known bacterium. About 3.16 and 1.97 million bacterial reads (less than 1% of all reads) were observed in the TruSeq Stranded Total RNA libraries prepared from sonicate fluid- and synovial fluid-derived RNA, respectively, and 1.45 and 1.93 million bacterial reads (about 2%) were found in the Ovation SoLo RNA-Seq System libraries from sonicate fluid- and synovial fluid-derived RNA, respectively (Table 2). Using these filtered reads, LMAT detected *S. sanguinis* in all four libraries as the most dominant of all microbial reads (Table 2, Table S1). As shown in Table S1, other bacteria and their rRNA sequences were also detected from the libraries. These organisms likely represent contaminants, given the low number of reads and extremely low number of genes detected. Since RNAs were not initially depleted of bacterial rRNA prior to library construction, it was expected that very few non-ribosomal genes would be detected.

Using the Rockhopper tool to identify expressed transcripts from *S. sanguinis*, we observed that the Ovation SoLo RNA-Seq System library preparation kit yielded higher numbers of non-ribosomal RNAs (>150 mRNA transcripts with reads per kilobase per million mapped sequence reads ranging from 1 to 1012) than the TruSeq Stranded Total RNA library preparation kit [7 mRNA transcripts (reads per kilobase per million mapped sequence reads ranging from 11 to 333) using both sonicate and synovial fluids] (Table 2). Bacterial virulence genes were detected among the expressed genes, including *purB* (SSA_0046), *purL* (SSA_0030), *ssaB* (SSA_0260), and *cbpA* (SSA_1063). Table S2 lists the top 10 expressed mRNAs detected from the Ovation SoLo RNA-Seq System libraries and the seven mRNAs from the TruSeq Stranded Total RNA. Comparison of bacterial gene expression profiles showed minimal overlap of the expressed genes detected between the two library construction methods (Figure 2).

There was a disparity in the percentage of microbial reads mapped to the *S. sanguinis* genome, with the TruSeq Stranded Total RNA kit displaying the higher number (Table 2). This could be explained by the different strategies employed by the two kits to produce cDNA. The NuGEN Ovation SoLo RNA-Seq System kit allowed a very small RNA input which underwent extensive amplification to generate sufficient cDNA for sequencing. Amplification could result in the majority of reads being assigned to more abundant RNAs (i.e., human) and unmappable reads (i.e., artifacts), and thus result in a lower percentage of reads mapped to *S. sanguinis*. The microbial reads that did not map to *S. sanguinis* mapped to a variety of environmental microorganisms that likely represent contaminants.

With regard to performance of the cDNA construction kits to produce the stranded libraries, we detected sets of genes arising from both sense and antisense strands of the same regions that are responsible for transcription of ribosomal RNAs and the SSA_0896 gene (encoding a two-component response transcriptional regulator) in the Ovation SoLo RNA-Seq System libraries. This indicates the Ovation kit is effective in retaining strandedness information of expressed genes.

3.3 Human RNA analysis

Human RNA analysis by MAP-RSeq showed differences in transcript expression profiles between the two sets of libraries. The TruSeq Stranded Total RNA kit generated more reads mapped to the human reference genome (90.8% and 90.4% for sonicate and synovial fluid samples, respectively *versus* 49.1% and 46.3% for sonicate and synovial sample, respectively) and more detected transcripts (15.1% and 9.7% for sonicate and synovial fluid samples, respectively *versus* 10.2 % and 7.6% for sonicate and synovial fluid samples, respectively) compared to the Ovation SoLo RNA-Seq System (Table 3), suggesting the TruSeq Stranded Total RNA kit to be more useful for analysis of human transcripts. Read counts of the genes detected in the two libraries prepared from the synovial and sonicate fluids are shown in Table S3. Comparison of the human mRNA genes detected from the TruSeq Stranded Total RNA and the Ovation SoLo RNA-Seq System libraries prepared from sonicate and synovial fluid samples revealed there to be significant overlap of the detected genes between the two library preparation methods [4114 and 1766 genes were found in both the Illumina TruSeq Stranded Total RNA and the NuGEN Ovation SoLo RNA-Seq System libraries from sonicate and synovial fluid samples, respectively (Figure 3a and 3b, respectively)]. When comparing the two sample types (sonicate fluid *versus* synovial fluid) processed by the same library preparation methods, high numbers of genes were found in both libraries [2251 and 2130 genes found in both libraries detected using the Illumina TruSeq Stranded Total RNA and the NuGEN Ovation SoLo RNA-Seq System libraries, respectively (Figure 3c and 3d)]. This comparison also showed sonicate fluid to be an overall better source for RNA-Seq analysis than synovial fluid, generating more transcriptomic information. In addition, the expression levels of the mRNA genes detected from both libraries appeared to correlate with each other, albeit with the genes from the NuGEN kit generally expressed at a higher level than those detected from the Illumina libraries (Table S3).

To further compare expression patterns of the two libraries' data sets, the top 50 expressed non-rRNA genes reported in each library underwent pathway analysis to identify associated pathways and upstream regulators (Table 4). From this analysis, the highly expressed genes detected from both libraries were involved in common canonical pathways, including the oxidative phosphorylation, mitochondrial dysfunction, sirtuin signaling pathway, and caveolar-mediated endocytosis signaling pathways; and shared the same upstream regulators (e.g., DAP3, actinonin, SIRT3, ALKBH1, and NSUN3), suggesting that the two library construction approaches, though yielding different numbers of transcripts, delivered similar overall protein-coding gene expression content.

When comparing expression profiles of micro RNAs (miRNAs) between the two kits, we observed that the Illumina TruSeq approach yielded a larger number of small RNAs, many of which were not detected in the NuGEN Ovation libraries (data not shown). Since miRNAs are believed to play a role in gene regulation, synthesis pathways, and disease etiology (Wahid et al., 2010), the ability of the TruSeq Stranded Total RNA kit to provide more miRNA information adds to evidence of usefulness of the kit for human transcriptome analysis.

With regard to host rRNA removal efficiency, the two library preparation kits employ different strategies to deplete rRNA sequences, and hence, produce libraries with different rRNA contents. The proportion of human rRNA reads was 0.001% on average for the TruSeq Stranded Total RNA libraries, and 7% on average for the Ovation SoLo RNA-Seq System libraries (Table 3), suggesting that the library construction approach from Illumina produces more informative human mRNA transcripts.

4 Discussion

Gene expression-based analysis in clinical settings, especially in infectious diseases, is a developing field that has broad applications, not only improving detection of microorganisms (Byron et al., 2016), but also informing host-pathogen interactions, and potentially using host response as a biomarker for infection (Westermann et al., 2012). In many clinical samples, a challenge is the extremely low quality and quantity of RNA, especially microbial RNA, which potentially impedes effective construction of libraries and subsequent RNA sequencing. Whole transcriptome amplification can increase the amount of nucleic acids for sequencing; yet, this can introduce biases (Coenen et al., 2015), and importantly, lose strand-origin information which allows accurate quantification of expression of genes that share overlapping regions of transcription (Mills, Kawahara, & Janitz, 2013; Zhao et al., 2015). In this study, we compared the ability of two strand-specific cDNA library construction kits to access human and bacterial transcriptomes. We showed that with regard to human gene expression, the Illumina TruSeq Stranded Total RNA kit produced more transcriptome information with more genes detected and fewer rRNA reads. The NuGEN Ovation SoLo RNA-Seq System was better for bacterial RNA analysis, detecting more non-ribosomal transcripts.

RNA sequencing of the bacterium *S. sanguinis* performed directly from clinical samples has not, to the best of our knowledge, been previously reported. Zhu et al. (2017) sequenced the transcriptome of a cultured *S. sanguinis* SK36 isolate. Total RNA was depleted for rRNA using the Ribozero Magnetic Kit for Bacteria, and a cDNA library was prepared with NEBNext Ultra Directional RNA Library Prep kit, and sequenced on an Illumina HiSeq 2000. Since this study utilized RNA of higher quantity (2 µg total RNA) and better quality for the RNA-Seq experiment, a high number of non-ribosomal genes (1,386 genes) was detected, as expected. In our RNA-Seq data obtained from clinical samples infected with *S.*

sanguinis, we detected fewer genes; however, expression of bacterial virulence genes (e.g., the biofilm associated genes) in our data set was observed at higher relative levels than from

the laboratorycultured bacterium. For example, two bacterial biofilm-related genes (SSA_0260 and SSA_1063) were found among the most expressed genes (Table S2).

RNA-Seq is an undeniably powerful tool that provides insights into the transcriptome of organisms in response to intrinsic and extrinsic changes. Yet, it imposes challenges at every step. It is known that different methods for cDNA synthesis and library preparation result in differential outcomes of transcriptomic data (Alberti et al., 2014). Selecting the most appropriate library construction approach for RNA-Seq studies depends on the nature of RNA samples, as each approach has its own strength and drawbacks. The TruSeq Stranded Total RNA kit has been widely used in human gene expression experiments as it produces good transcriptional coverage and gene expression profiles. A major caveat is that it requires µg amounts of starting materials, a requirement hard to meet for many clinical samples. Libraries constructed from 100 ng total RNA by the TruSeq Stranded kit were reported to produce insufficient yield and skew the library content (Alberti et al., 2014). For total RNA with very low quantity, the Ovation SoLo RNA-Seq System library preparation appears to be an option for the formation of a next generation sequencing library which allows one to start with the RNA amount of a single cell. The Ovation SoLo RNA-Seq System is a newer kit developed by NuGEN Technologies in 2016 and has not yet been fully evaluated for its productivity and potential biases. In this study, we observed that of the total reads obtained from the Ovation SoLo RNA-Seq System libraries, only about 50% mapped to the reference human genome (Table 3), 2% mapped to microbial sequences (Table 2), and a very high percentage (46%) were unmappable, likely due to amplification artifact. For reference, the percentage of the unmappable reads observed for the TruSeq Stranded Total RNA kit was 9%. This suggests that this NuGEN strategy used in our study may have over-amplified and/or created nonspecific sequences.

There are limitations to this study. We tested the cDNA synthesis and library preparation kits from RNA extracted from clinical samples that contain both human and microbial cells. We do not know the exact cellular components of the samples or the ratio between the human and microbial cells. This is, however, characteristic of many samples from the sites of human infections. We did not investigate whether cDNA synthesis and amplification methods employed by the two kits have a species bias [e.g., different nucleotide compositions among species may cause amplification bias (Alberti et al., 2014; Armour et al., 2009)]. Another limitation is that in the experimental design, we did not remove bacterial rRNA sequences before or during library construction, resulting in abundant bacterial rRNA reads in both libraries. Given the overwhelmingly predominant host cells in our chosen samples, we omitted depletion of bacterial-specific rRNA hoping to obtain good sequencing coverage of bacterial RNA, as removing rRNA also reduces the yield of mRNA (Westermann et al., 2012). Furthermore, this enabled us to compare bacterial gene expression from the libraries prepared by the two kits. Additional studies are needed to assess libraries from microbial rRNA-diminished RNA pools in a dual RNA sequencing setting.

It has been suggested that analysis of a bacterial transcriptome requires 40,000–60,000 non-ribosomal RNA reads to efficiently detect differential expression (Mandlik et al., 2011; Westermann et al., 2017). This conclusion was drawn from a study using a *Vibrio cholerae*-

infected rabbit model, which had one infecting bacterial species in the sample and an extremely high bacterial load (5×10^8 colony forming unit/ml). This experimental setting resulted in good quality abundant starting RNA materials. Here, we utilized clinical samples which could be infected by multiple organisms, and in which the organisms are present in low quantities. In addition, the samples were collected and stored without using RNA stabilizer, which could affect the quality of extracted RNA. These are limitations of our study as we dealt with a limited amount of pathogen and low quality and quantity input RNA. Our data showed there to be 2,378 non-ribosomal *S. sanguinis*-specific RNA reads mapping to bacterial transcripts. Although this number is not as high as in animal infection studies, it was useful to us to help determine the best method for RNA-Seq analysis of this particular sample type.

The study here aimed to provide proof of concept data comparing two RNA-Seq methods utilizing a small set of samples. While this sample size is a clear limitation, we observed obvious differences in bacterial and human RNA expression between the two methods, for example, the percentage of microbial reads mapped to *S. sanguinis* (44% versus 21%) and the number of *S. sanguinis* mRNAs (7 versus 301). Therefore, for the purpose of comparison of methods for profiling human and pathogen transcriptomes in our sample-types, this work was beneficial in guiding the selection of an approach for generating cDNA libraries from various targeted RNAs. A major drawback in experimental design is that different sequencing read lengths and levels of multiplexing were used, which could impact the RNA-Seq results. While read length can affect mapping ability of reads to particular reference genomes, other studies have also shown that read lengths ≥ 100 bp can accurately quantify gene expression and result in the same level of sensitivity of mapping using several aligners (e.g., 100 bp *versus* 250 bp) (Chhangawala et al., 2015; Thankaswamy-Kosalai et al., 2017). Given this, our data should reflect differences in RNASeq outcomes from the two kits, and not be an artifact of read length. More work may be needed to determine the effect of the read lengths on our data set. The two libraries were multiplexed differently upon sequencing, resulting in different depth of coverage (read number). To normalize the data, the percentage of reads was calculated and used for comparison instead of the number of reads. For instance, although the Illumina TruSeq Stranded Total RNA kit generated a higher number of microbial reads than the NuGEN Ovation SoLo RNA-Seq System kit, the percentage of these reads indicated less microbial RNA in the Illumina libraries.

Choosing a method for cDNA library construction for RNA sequencing depends on many factors, including the nature of the input RNA (e.g., quantity and degradation degree), and the end goals of the experiment (e.g., separate RNA-Seq *versus* dual RNA-Seq, retaining strandedness). Here we present our experience with RNA sequencing and transcriptomic analysis of two cDNA libraries prepared from a mixture of bacterial and human total RNA of relatively low quality and quantity. The NuGEN Ovation SoLo RNA-Seq System library preparation kit, compared to the Illumina TruSeq Stranded Total RNA library preparation kit, allowed very small amounts of starting materials and performed better with respect to bacterial expression; however, it provided less informative transcriptomic data for the human component.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank the Mayo Clinic Medical Genome Facility for assistance with library preparation and sequencing, and the Mayo Clinic Center for Individualized Medicine for assisting with transcriptomic data analysis.

Funding

Research reported in this publication was supported by the National Institute of Arthritis and Musculoskeletal and Skin Diseases of the National Institutes of Health under Award Number R01 AR056647. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

- Alberti A, Belser C, Engelen S, Bertrand L, Orvain C, Brinas L, ... Wincker P (2014). Comparison of library preparation methods reveals their impact on interpretation of metatranscriptomic data. *BMC Genomics*, 15(1), 912. doi:10.1186/1471-2164-15-912 [PubMed: 25331572]
- Ames SK, Hysom DA, Gardner SN, Lloyd GS, Gokhale MB, & Allen JE (2013). Scalable metagenomic taxonomy classification using a reference genome database. *Bioinformatics*, 29(18), 2253–2260. doi:10.1093/bioinformatics/btt389 [PubMed: 23828782]
- Armour CD, Castle JC, Chen R, Babak T, Loerch P, Jackson S, ... Raymond CK (2009). Digital transcriptome profiling using selective hexamer priming for cDNA synthesis. *Nat Methods*, 6(9), 647–649. doi:10.1038/nmeth.1360 [PubMed: 19668204]
- Bolger AM, Lohse M, & Usadel B (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), 2114–2120. doi:10.1093/bioinformatics/btu170 [PubMed: 24695404]
- Byron SA, Van Keuren-Jensen KR, Engelthaler DM, Carpten JD, & Craig DW (2016). Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nature Reviews Genetics*, 17, 257. doi:10.1038/nrg.2016.10
- Campello L, & Martín-Nieto J (2013). RNA-Seq expression profiling of genes related to neurodegenerative disorders affecting the human retina. *EMBnet.journal*; Vol 19: Supplement A.
- Cazanave C, Greenwood-Quaintance KE, Hanssen AD, Karau MJ, Schmidt SM, Gomez Urena EO, ... Patel R (2013). Rapid molecular microbiologic diagnosis of prosthetic joint infection. *J Clin Microbiol*, 51(7), 2280–2287. doi:10.1128/jcm.00335-13 [PubMed: 23658273]
- Chen Z, & Duan X (2011). Ribosomal RNA depletion for massively parallel bacterial RNA-sequencing applications. *Methods Mol Biol*, 733, 93–103. doi:10.1007/978-1-61779-089-8_7 [PubMed: 21431765]
- Chhangawala S, Rudy G, Mason CE, & Rosenfeld JA (2015). The impact of read length on quantification of differentially expressed genes and splice junction detection. *Genome Biology*, 16(1), 131. doi:10.1186/s13059-015-0697-y [PubMed: 26100517]
- Chu J, Sadeghi S, Raymond A, Jackman SD, Nip KM, Mar R, ... Birol I (2014). BioBloom tools: fast, accurate and memory-efficient host species sequence screening using bloom filters. *Bioinformatics*, 30(23), 3402–3404. doi:10.1093/bioinformatics/btu558 [PubMed: 25143290]
- Coenen C, Liedtke S, & Kogler G (2015). RNA amplification protocol leads to biased polymerase chain reaction results especially for low-copy transcripts of human bone marrow-derived stromal cells. *PLOS ONE*, 10(10), e0141070. doi:10.1371/journal.pone.0141070 [PubMed: 26485654]
- Giannoukos G, Ciulla DM, Huang K, Haas BJ, Izard J, Levin JZ, ... Gnirke A (2012). Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. *Genome Biology*, 13(3), r23–r23. doi:10.1186/gb-2012-13-3-r23 [PubMed: 22455878]
- Gonorazky H, Liang M, Cummings B, Lek M, Micallef J, Hawkins C, ... Dowling JJ (2016). RNAseq analysis for the diagnosis of muscular dystrophy. *Annals of Clinical and Translational Neurology*, 3(1), 55–60. doi:10.1002/acn3.267 [PubMed: 26783550]

- Humphrys MS, Creasy T, Sun Y, Shetty AC, Chibucos MC, Drabek EF, ... Myers GSA (2013). Simultaneous transcriptional profiling of bacteria and their host cells. *PLOS ONE*, 8(12), e80597. doi:10.1371/journal.pone.0080597 [PubMed: 24324615]
- Kalari KR, Nair AA, Bhavsar JD, O'Brien DR, Davila JI, Bockol MA, ... Kocher J-PA (2014). MAP-RSeq: Mayo analysis pipeline for RNA sequencing. *BMC Bioinformatics*, 15(1), 224. doi: 10.1186/1471-2105-15-224 [PubMed: 24972667]
- Magri F, Del Bo R, D'Angelo MG, Govoni A, Ghezzi S, Gandossini S, ... Comi GP (2011). Clinical and molecular characterization of a cohort of patients with novel nucleotide alterations of the Dystrophin gene detected by direct sequencing. *BMC Medical Genetics*, 12, 37–37. doi: 10.1186/1471-2350-12-37 [PubMed: 21396098]
- Mandlik A, Livny J, Robins WP, Ritchie JM, Mekalanos JJ, & Waldor MK (2011). RNA-Seq-based monitoring of infection-linked changes in *Vibrio cholerae* gene expression. *Cell Host Microbe*, 10(2), 165–174. doi:10.1016/j.chom.2011.07.007 [PubMed: 21843873]
- Mills JD, Kawahara Y, & Janitz M (2013). Strand-specific RNA-Seq provides greater resolution of transcriptome profiling. *Current Genomics*, 14(3), 173–181. doi:10.2174/1389202911314030003 [PubMed: 24179440]
- Tariq MA, Kim HJ, Jejelowo O, & Pourmand N (2011). Whole-transcriptome RNAseq analysis from minute amount of total RNA. *Nucleic Acids Res*, 39(18), e120–e120. doi:10.1093/nar/gkr547 [PubMed: 21737426]
- Thankaswamy-Kosalai S, Sen P, & Nookaew I (2017). Evaluation and assessment of read-mapping by multiple next-generation sequencing aligners based on genome-wide characteristics. *Genomics*, 109(3), 186–191. doi:https://doi.org/10.1016/j.ygeno.2017.03.001 [PubMed: 28286147]
- Thoendel M, Jeraldo PR, Greenwood-Quaintance KE, Yao JZ, Chia N, Hanssen AD, ... Patel R (2016). Comparison of microbial DNA enrichment tools for metagenomic whole genome sequencing. *J Microbiol Methods*, 127, 141–145. doi:10.1016/j.mimet.2016.05.022 [PubMed: 27237775]
- Tjaden B (2015). De novo assembly of bacterial transcriptomes from RNA-seq data. *Genome Biology*, 16(1), 1. doi:10.1186/s13059-014-0572-2 [PubMed: 25583448]
- Tomlins SA, Mehra R, Rhodes DR, Shah RB, Rubin MA, Bruening E, ... Chinnaiyan AM (2006). Whole transcriptome amplification for gene expression profiling and development of molecular archives. *Neoplasia (New York, N.Y.)*, 8(2), 153–162.
- Wahid F, Shehzad A, Khan T, & Kim YY (2010). MicroRNAs: Synthesis, mechanism, function, and recent clinical trials. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1803(11), 1231–1243. doi:https://doi.org/10.1016/j.bbamcr.2010.06.013 [PubMed: 20619301]
- Westermann AJ, Barquist L, & Vogel J (2017). Resolving host–pathogen interactions by dual RNAseq. *PLOS Pathogens*, 13(2), e1006033. doi:10.1371/journal.ppat.1006033 [PubMed: 28207848]
- Westermann AJ, Forstner KU, Amman F, Barquist L, Chao Y, Schulte LN, ... Vogel J (2016). Dual RNA-seq unveils noncoding RNA functions in host-pathogen interactions. *Nature*, 529(7587), 496–501. doi:10.1038/nature16547 [PubMed: 26789254]
- Westermann AJ, Gorski SA, & Vogel J (2012). Dual RNA-seq of pathogen and host. *Nat Rev Micro*, 10(9), 618–630. doi:http://www.nature.com/nrmicro/journal/v10/n9/supinfo/nrmicro2852_S1.html
- Yi H, Cho Y-J, Won S, Lee J-E, Jin Yu H, Kim S, ... Chun J (2011). Duplex-specific nuclease efficiently removes rRNA for prokaryotic RNA-seq. *Nucleic Acids Res*, 39(20), e140–e140. doi: 10.1093/nar/gkr617 [PubMed: 21880599]
- Zhang W, Yu Y, Hertwig F, Thierry-Mieg J, Zhang W, Thierry-Mieg D, ... Fischer M (2015). Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. *Genome Biology*, 16(1), 133. doi:10.1186/s13059-015-0694-1 [PubMed: 26109056]
- Zhao S, Fung-Leung W-P, Bittner A, Ngo K, & Liu X (2014). Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLOS ONE*, 9(1), e78644. doi:10.1371/journal.pone.0078644 [PubMed: 24454679]
- Zhao S, Zhang Y, Gordon W, Quan J, Xi H, Du S, ... Zhang B (2015). Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap. *BMC Genomics*, 16(1), 675. doi:10.1186/s12864-015-1876-7 [PubMed: 26334759]

Zhu B, Ge X, Stone V, Kong X, El-Rami F, Liu Y, ... Xu P (2017). *ciaR* impacts biofilm formation by regulating an arginine biosynthesis pathway in *Streptococcus sanguinis* SK36. Scientific Reports, 7(1), 17183. doi:10.1038/s41598-017-17383-1 [PubMed: 29215019]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Highlights:

- Compared 2 cDNA library kits for human samples undergoing transcriptomic analysis
- The Ovation SoLo RNA-Seq System was superior for bacterial transcriptomic analysis
- The TruSeq Stranded Total RNA kit did better for human transcriptomic analysis

Figure 1:
RNA-Seq experiment workflow.

Figure 2:

Overlap of the *S. sanguinis* expressed non-rRNA genes between the two library generation methods using the sonicate fluid sample (a) and the synovial fluid sample (b); and between the two sample types processed by the Illumina TruSeq Stranded Total RNA kit (c) and the NuGEN Ovation SoLo RNA-Seq System (d).

Figure 3:

Overlap of the human expressed genes between the two library generation methods using the sonicate fluid sample (a) and the synovial fluid sample (b); and between the two sample types processed by the Illumina TruSeq Stranded Total RNA kit (c) and the NuGEN Ovation SoLo RNA-Seq System (d).

Table 1:

Quantity and quality of total RNA and cDNA libraries prepared by the two kits

		RNA Concentration (ng/μl)	RNA Integrity Number	DV₂₀₀[*]	Library Concentration (ng/μl)	Library average size (bp)
Sonicate fluid	Total RNA	59	1.8	54		
	TruSeq Stranded Total RNA library				15	302
	Ovation SoLo RNA-Seq System library				4	291
Synovial fluid	Total RNA	46	1.9	53		
	TruSeq Stranded Total RNA library				6	282
	Ovation SoLo RNA-Seq System library				5	292

^(*) Percentage of RNA fragments greater than 200 nucleotides

Table 2:

Analysis of microbial RNA

	Sonicate fluid		Synovial fluid	
	TruSeq Stranded Total RNA	Ovation SoLo RNA-Seq System	TruSeq Stranded Total RNA	Ovation SoLo RNA-Seq System
Raw reads (in millions)	406	92	287	97
Microbial reads after filtering in millions (%)	3.16 (0.8)	1.45 (1.6)	1.97 (0.7)	1.93 (2.0)
% Microbial reads mapped to <i>Streptococcus sanguinis</i> genome by LMAT	44.4	21.5	35.7	28.6
Number of <i>S. sanguinis</i> mRNAs by Rockhopper expression analysis	7	151	7	301
RPKM [*] range of <i>S. sanguinis</i> mRNAs	11–333	1–156	19–198	1–1012

* Reads per kilobase per million mapped sequence reads

Table 3:

Analysis of human RNA

	Sonicate fluid		Synovial fluid	
	TruSeq Stranded RNA	Total RNA-Seq System	TruSeq Stranded RNA	Ovation SoLo RNA-Seq System
% Total read mapped to human genome	90.8	49.1	90.4	46.3
% Human genes detected	15.1	10.2	9.7	7.6
% rRNA reads remaining	0.002	1.67	0.0007	12.36

Table 4:

Top canonical pathways and top upstream regulators predicted for the top 50 expressed genes detected from the two libraries

		TruSeq Stranded Total RNA p-value^(*)	Ovation SoLo RNA-Seq System p-value^(*)
Top canonical pathways shared by TruSeq Stranded Total RNA and Ovation SoLo RNA-Seq System libraries	Oxidative phosphorylation	2.52E-13	3.24E-13
	Mitochondrial dysfunction	3.41E-13	4.51E-13
	Sirtuin signaling	7.48E-07	8.97E-07
	Caveolar-mediated endocytosis signaling	7.68E-06	8.51E-06
Uncommon top canonical pathway	OX40 signaling	2.06E-05	
	NRF2-mediated oxidative stress response		2.60E-05
Top upstream regulators shared by TruSeq Stranded Total RNA and Ovation SoLo RNA-Seq System libraries	DAP3	5.59E-23	7.21E-23
	Actinonin	1.56E-22	2.02E-22
	SIRT3	1.18E-19	1.58E-19
	ALKBH1	2.92E-19	3.54E-19
	NSUN3	2.92E-19	3.54E-19

^(*) p-values were calculated based on the detected genes known to be involved in the pathways.