HHS Public Access

Author manuscript

J Vis Exp. Author manuscript; available in PMC 2022 February 04.

Optimization for Sequencing and Analysis of Degraded FFPE-RNA Samples

Yelena Levin[*,1], **Keyur Talsania**[*,1,2], **Bao Tran**[1], **Jyoti Shetty**[1], **Yongmei Zhao**[1,2], **Monika Mehta**[1]

[1]NCI CCR Sequencing Facility, Frederick National Laboratory for Cancer Research

[2]Advanced Biomedical and Computational Sciences, Frederick National Laboratory for Cancer Research

Abstract

Gene expression analysis by RNA sequencing (RNA-seq) enables unique insights into clinical samples that can potentially lead to mechanistic understanding of the basis of various diseases as well as resistance and/or susceptibility mechanisms. However, FFPE tissues, which represent the most common method for preserving tissue morphology in clinical specimens, are not the best sources for gene expression profiling analysis. The RNA obtained from such samples is often degraded, fragmented, and chemically modified, which leads to suboptimal sequencing libraries. In turn, these generate poor quality sequence data that may not be reliable for gene expression analysis and mutation discovery. In order to make the most of FFPE samples and obtain the best possible data from low quality samples, it is important to take certain precautions while planning experimental design, preparing sequencing libraries, and during data analysis. This includes the use of appropriate metrics for precise sample quality control (QC), identifying the best methods for various steps during the sequencing library generation, and careful library QC. In addition, applying correct software tools and parameters for sequence data analysis is critical in order to identify artifacts in RNA-seq data, filter out contamination and low quality reads, assess uniformity of gene coverage, and measure the reproducibility of gene expression profiles among biological replicates. These steps can ensure high accuracy and reproducibility for profiling of very heterogeneous RNA samples. Here we describe the various steps for sample QC, library preparation and QC, sequencing, and data analysis that can help to increase the amount of useful data obtained from low quality RNA, such as that obtained from FFPE-RNA tissues.

Keywords

Genetics; Issue 160; RNA sequencing; formalin-fixed paraffin embedded; FFPE; next generation sequencing; NGS; RNA-seq analysis

Correspondence to: Jyoti Shetty at jyoti.shetty@nih.gov, Yongmei Zhao at zhaoyong@mail.nih.gov, Monika Mehta at monika.mehta@nih.gov.

*These authors contributed equally

Video Link

The video component of this article can be found at https://www.jove.com/video/61060/

Introduction

Use of next-generation sequencing approaches has enabled us to glean a wealth of information from various types of samples. However, old and poorly preserved samples remain unworkable for the commonly used methods of generating sequence data and often require modifications to well-established protocols. FFPE tissues represent such a sample type that has been widely utilized for clinical specimens[1,2,3]. While FFPE preservation maintains tissue morphology, the nucleic acids in FFPE tissues usually exhibit a wide range of damage and degradation, making it difficult to retrieve the genomic information that may lead to important insights about molecular mechanisms underlying various disorders.

Gene expression data generated by RNA sequencing is often instrumental in studying disease and resistance mechanisms and complements DNA mutation analysis. However, RNA is more susceptible to degradation, making it more challenging to generate accurate gene expression data from FFPE tissues. Furthermore, because the wide availability and affordability of sequencing is relatively recent, older specimens were often not stored in conditions required to preserve RNA integrity. Some of the issues for FFPE samples include degradation of RNA due to embedding in paraffin, chemical modification of RNA leading to fragmentation or refractoriness to enzymatic processes required for sequencing, and loss of the poly-A tails, limiting the applicability of oligo-dT as a primer for reverse transcriptase[4]. Another challenge is the handling/storage of FFPE samples under suboptimal conditions, which may lead to further degradation of labile molecules such as RNA in the tissues[5]. This is especially relevant for older samples that may have been collected at a time when gene expression analysis by RNA sequencing was not anticipated for the samples. All these lead to decreased quality and quantity of the extracted RNA available for generating useful sequence data. The low probability of success, combined with the high cost of sequencing, has dissuaded many researchers from trying to generate and analyze gene expression data from potentially useful FFPE samples. Some studies in recent years have demonstrated the usability of FFPE tissues for gene expression analysis[2,6,7,8,9], albeit for fewer and/or more recent samples.

As a feasibility study, we used RNA extracted from FFPE tumor tissue specimens from three Residual Tissue Repositories from Surveillance, Epidemiology, and End Results (SEER) cancer registries for RNA sequencing and gene expression analysis[10]. Procured from clinical pathology labs, the FFPE tissues from high-grade ovarian serous adenocarcinomas were stored from 7–32 years under varying conditions before RNA extraction. Because in most cases these blocks had been stored in different sites for years without the expectation of any sensitive genetic analysis in the future, not much care had been taken to preserve the nucleic acids. Thus, most of the samples exhibited poor quality RNA, with a large proportion of samples contaminated with bacteria. Nevertheless, we were able to perform gene quantification, measure the uniformity and continuity of gene coverage, and perform the Pearson correlation analysis among biological replicates to measure reproducibility. Based on a set of key signature gene panel, we compared the samples in our study with The Cancer Genome Atlas (TCGA) data and confirmed that approximately 60% of the samples had comparable gene expression profiles[11]. Based on the correlation between various QC

results and sample metadata, we identified key QC metrics that have good predictive value for identifying samples that are more likely to generate usable sequence data[11].

Here we describe the methodology used for FFPE-RNA quality assessment, generation of sequencing libraries starting from extracted RNA samples, and bioinformatic analysis of the sequencing data.

Protocol

1. RNA quantity and quality assessment

1. Select the FFPE samples according to predefined criteria and extract RNA using an appropriate method (e.g., FFPE-nuclei acid extraction kit, Table of Materials). NOTE: There are several different methods available for FFPE-RNA extraction, including the newer microdissection methods that can work with very little tissue and extract good quality RNA[12,13,14].
2. Utmost care should be taken to preserve the integrity of RNA at all stages. This includes working with RNase free deionised water, using RNase free plasticware, and cleaning all instruments that come in contact with the FFPE blocks with RNase decontamination reagents.
3. RNA should always be handled carefully and kept in ice unless otherwise specified to minimize degradation while handling.
4. If enough material is available, extract RNA from more than one region in the FFPE block to generate biological replicates from as many samples as possible. For some of the samples with ample RNA yield, divide the extracted RNA into two to process as technical replicates.
5. If possible, collect a small amount of sample separately after extraction for QC (i.e., a QC aliquot) to avoid repeated handling and freeze-thaw cycles of the sample that will likely lead to degradation of the RNA.
6. Check the quality of the RNA (preferably from the QC aliquot) by running it on an RNA QC system (e.g., Agilent Bioanalyzer system using an RNA Nano chip, Table of Materials) according to the manufacturer's instructions.
7. Analyze the distribution of RNA fragments in the samples (e.g., using the Bioanalyzer 2100 Expert software) by calculating the DV_{200} and DV_{100} values as the percent of fragments larger than 200 nt (DV_{200}) or 100 nt (DV_{100}) in size.
8. Among DV_{200} and DV_{100}, identify the metric that has a larger spread of values for the given sample set, and pick that for grouping the samples according to their degree of intactness. NOTE: For sample sets with more intact RNA molecules (i.e., high DV_{200} values, all or most with $DV_{200} > 40\%$), DV_{200} is likely to be a useful QC metric. However, for sample sets with more degraded transcripts (i.e., low DV_{200} values, all or most with $DV_{200} < 40\%$), DV_{100} is more likely to be useful.

9. Based on the QC metrics, identify the samples that have DV_{100} < 40%. Because this degree of degradation is highly likely to not generate useful sequencing data[11], it is advisable to avoid processing such samples. If replacements for such samples are available, their quality should be checked to ideally only include samples with DV_{100} > 50%.

2. Sequencing library preparation

1. Based on the quality of the samples as assessed in section 1, identify an appropriate method for generating the sequencing libraries.

    1. For sample sets with very low degradation and high DV_{200} values, use mRNA sequencing (i.e., capture of polyadenylated transcripts), targeted RNA sequencing (i.e., use of capture probes for specific genes of interest), RNA exome sequencing (i.e., use of capture probes to enrich for the coding transcriptome), or total RNA sequencing (i.e., use of random primers for reverse transcription to sequence the entire RNA population after removing ribosomal RNA from the samples). However, it is important to note that the fixation process may introduce bias in the extracted RNA. Thus, the capture approaches may not work well in all cases, even with high DV_{200} values.

    2. If the sample set includes samples with high degradation (DV_{200} < 30%), use a total RNA library preparation method and not one that depends on the capture of specific regions of the transcripts, because those specific regions may be missing in degraded samples. The use of random primers for generation of cDNA leads to higher representation of usable RNA in the final library, and is, therefore, more suited for FFPE-RNA samples.

    3. For ribosomal RNA depletion for sample sets with high degradation, use RNaseH-based methods. These are methods where rRNA-specific DNA probes bind to rRNA, double-stranded molecules are digested by RNaseH, and leftover probes are cleaned up by DNase (e.g., NEBNext rRNA depletion kit, Table of Materials). These methods work better for degraded samples than some other methods[8].

2. For generating sequencing libraries, use higher input amounts (if possible) for samples that have more degraded RNA (DV_{100} < 60%). While samples with reasonably good quality RNA (DV_{100} > 60%) may yield good sequence data even at lower input amounts (the lowest tested for this protocol with FFPE-RNA was ~20 ng), for more degraded RNA (DV_{100} < 60%), it is better to start with higher input amounts (e.g., >100 ng). NOTE: If enough (e.g., >500 ng) sample is available, it is advisable to save at least half of the sample for repeating the library preparation, if needed. For low input samples (e.g., <100 ng), it is usually better to use the entire amount and generate a library of sufficient diversity.

3. After selecting a suitable library preparation kit for generating total RNA seq libraries from samples with high degradation (e.g., NEBNext Ultra II RNA

Library Prep Kit for Illumina, see Table of Materials), follow the manufacturer's instructions to generate the libraries. NOTE: During library preparation, it is important to skip the RNA fragmentation step for degraded samples and to ensure the use of random primers for first strand cDNA synthesis.

4. For improving the efficiency and speed, especially for the low-input samples, use appropriate magnetic racks with strong fixed magnets for bead-based purification and size-selection steps (see Table of Materials).

5. For PCR enrichment of adapter ligated DNA, adjust the number of amplification cycles based on the amount of input DNA to ensure maximum representation while avoiding unnecessary duplication of the library molecules. For low input FFPE-RNA samples (<100 ng), we recommend 16–18 amplification cycles, while the high input samples (1,000 ng) usually generate enough library amounts in 12–14 rounds of amplification.

6. Following PCR amplification and cleanup per the manufacturer's instructions, assess the library quality by analyzing library concentration and molecule distribution on an appropriate platform (e.g., Agilent Bioanalyzer DNA Chip, see Table of Materials). For samples with primer peaks (~80 bp) or adapter-dimer peaks (~128 bp), repeat the cleanup to remove those peaks.

7. Calculate the average library size for each library (e.g., using the Bioanalyzer 2100 Expert software).

3. Sequencing library QC

1. Once it has been ascertained that the libraries are free of excess primer and adapter-dimers and have sufficient concentration for subsequent sequencing, quantitate further by qPCR.

   NOTE: Owing to the sensitivity of cluster generation towards library concentration, accurate quantification is vital to prevent costly sequencing runs from underperformance or overloading. Quantitative real-time PCR (qPCR) methods are useful for improving cluster density on Illumina platforms without resulting in overclustering. The qPCR method is more precise and more sensitive than the methods based on qualitative and/or quantitative analysis of all library molecules (e.g., Agilent Bioanalyzer), because it measures the templates that have both adapter sequence so neither end that will form clusters on the flowcell. Library size must, however, be known in advance as a size correction must be applied to all samples so that results can be compared against a standard curve.

   CAUTION: Lab coats and gloves must always be worn when performing qPCR, and the procedure must be performed in a biosafety cabinet following the manufacturer's instructions.

   1. Set up a 96 well plate with three replicates for each sample for error prevention using a suitable kit (e.g., KAPA SYBR FAST qPCR Master Mix for Illumina libraries, a part of Library Quantification kit, see Table of Materials), along with the standards, a positive control (e.g, PhiX

control, see Table of Materials), and a no template control (NTC). The NTC is qPCR mix without DNA library. The positive control can be any library with known concentration and fragment size.

1. Prepare a minimum of six dilutions of the standards following the vendor protocol.

2. After adding all the components (i.e., qPCR master mix, libraries, standards), cover the plate with sealing film and use a squeegee to ensure the film makes even and secure contact with the plate.

3. Vortex and spin down the plate at 1,500 rpm for at least 1 min. Visually inspect the plate to make sure there are no air bubbles at the bottom of the wells.

4. Set up the plate on the thermal cycler (e.g. CFX96 Touch System, see Table of Materials) using the manufacturer's recommended settings.

5. Save the run folder where it can be accessed for data analysis.

6. During data analysis, check that the slope is in the −3.1 to −3.6 range, efficiency from 90% to 110% and the R^2 (coefficient of correlation obtained for the standard curve) no less than 0.98.

2. **Pooling**: Once the qPCR concentration of the sequencing ready libraries is obtained, pool equimolar amounts of each of the libraries, depending on the number of sequencing reads required per sample and the sequencing output of the instrument.

3. **QC of the pools**: Quantitate the library pools again by qPCR following the same protocol as described in step 3.1.

4. Sequencing

1. Depending on the run parameters, pull the sequencing reagent kits and thaw them following the user guide. Please check the Illumina website for the latest versions of all user guides for sequencing on Illumina instruments.

2. Make sure the reagents are completely thawed and place the reagents tray at 4 °C. The run should be started no later than 2 h after the reagents have been defrosted. Not doing that could affect quality of the run results.

3. Invert the cartridge 5x to mix reagents and gently tap on the bench to reduce air bubbles.

4. Set the unwrapped flow cell package aside at room temperature for 30 min.

5. Unwrap the flow cell package and clean the glass surface of the flow cell with a lint-free alcohol wipe. Dry the glass with a low-lint laboratory tissue.

6. Open the Illumina “**Experiment Manager**” application. Choose “**Create Sample Sheet**”, then choose the **Sequencer** and click “**Next**”.

7. Create and upload the sample sheet based on Illumina sequencer criteria (e.g., Illumina Experiment Manager, software guide).

8. At the prompts, scan in the reagent kit barcode and enter the run **Set Up Parameters** (e.g., for a single indexed PE 75 cycle run, enter **76-8-76**).

9. Denature and dilute the library pool based on the sequencer user guide recommendation (e.g., NextSeq 500 System guide from Illumina, see Table of Materials).

10. Denature and dilute the control library PhiX (see Table of Materials) to the appropriate concentration (e.g., 1.8 pM for NextSeq).

11. Mix sample library and PhiX control to result in a 1% PhiX control volume ratio.

12. Load denatured and diluted sample into the reagent cartridge in the designated reservoir.

13. Load the flowcell, buffer cartridge, and the reagent cartridge.

14. Perform an automated check and review to ensure that the run parameters pass the system check.

15. When the automated check is complete, select **Start** to begin the sequencing run.

5. Data analysis and quality assessment

NOTE: A typical RNA-seq data analysis workflow (Figure 1) includes preprocessing and QC, alignment to genome and post alignment QC, gene and transcript quantification, sample correlation analysis, differential analysis between different sample groups, treatment conditions, and gene set enrichment and pathway analysis.

The RNA-seq data may have quality issues that can affect the accuracy of gene profiling and lead to erroneous conclusions. Therefore, initial QC checks for sequencing quality, contamination, sequencing coverage bias, and other sources of artifacts are very important. Applying an RNA-Seq QC pipeline similar to the workflow described here is recommended to detect artifacts and apply filtering or correction before downstream analysis.

1. Preprocessing

   NOTE: This includes demultiplexing, assessment of sequence read quality, GC content, presence of sequencing adapters, overrepresented *k*-mers, and PCR duplicated reads. This information helps to detect sequencing errors, PCR artifacts, or contamination.

   1. Demultiplex Illumina sequencing run using the Illumina software tool **bcl2fastq2** to generate raw FASTQ files for each sample defined in the sample sheet. Allow one mismatch in the sample index barcodes to tolerate sequencing errors if there is no barcode collision.

   2. Run the **FASTQC** [15] software tool to perform a quality check on raw FASTQ files to detect any poor quality or abnormalities in sequencing reads.

3. For adapter and low-quality bases trimming, trim the sequencing adapters and low quality bases using **Cutadapt**[16] or **Trimmomatic**[17] software tools. Save the trimmed reads in the pair-end fastq files.

4. Contamination screen

    1. Run **FASTQ_screen**[18] to detect possible cross contamination with other species.

    2. Run **miniKraken** of Kraken2[19] to identify the taxonomies of contaminating species.

2. Alignment to reference genome and post alignment QC

    1. The trimmed reads can be aligned to a reference genome sequence (GRCh Build hg19 or hg38) using STAR aligner[20]. Apply the Gencode annotation GTF file to guide the spliced transcript alignment. It is recommended to run **STAR 2-pass** to increase sensitivity to novel splice junctions. In the second pass, all reads will be remapped using annotated gene and transcripts and novel junctions from the first pass.

    2. Perform post-alignment QC.

        1. Run Picard's[21] **MarkDuplicates** to evaluate the library complexity by determining the amount of unique or nonduplicated reads in the samples.

        2. Run Picard's **CollectRnaSeqMetrics** program to collect mapping percentages on coding, intronic, intergenic, UTR regions, and gene body coverage.

        3. Run **RSeQC**[22] to determine the read pair inner distance, read distribution among CDS exons, 5'UTR, 3'UTR, intron, TSS_up_1kb, TSS_up_5kb, TSS_up_10kb, TES_down_1kb, TES_down_5kb, TES_down_10kb, read GC content, junction saturation, and library strand information.

        4. Run **multi-QC**[23] to generate an aggregated report in HTML format.

3. Gene quantification and correction analysis

    1. Run **RSEM**[24] to get raw count as well as normalized read count on genes and transcripts. The read count measurement such as RPKM (reads per kilobase of exon model per million reads), FPKM (fragments per kilobase of exon model per million mapped reads), and TPM (transcripts per million) are the most often reported RNA-seq gene expression values. Genes expressed below a noised threshold (such as TPM < 1 or raw count <5) can be filtered.

2. Perform transcript quantification to aggregate raw counts of mapped reads to each transcript sequences using programs such as HTSeq-count or featureCounts.

3. Run **Principal Components Analysis** (PCA) using an **R script** to determine batch effects and assess a quality map of the given dataset[25]. Sample correlation analysis can be carried out using the Pearson correlation between different metrics.

4. Differential gene expression analysis

   1. Perform gene differential analysis between sample conditions using the program **edgeR**[26,27] and/or **limma-Voom**[28] and use normalization methods including **TPM**, **TMM**, **DESeq**, or **UpperQuartile**.

   2. It is recommended to run at least two differential analysis software tools in order to call two set of DEGs lists for comparison and get the final DEGs to improve detection sensitivity and accuracy.

5. Gene set enrichment and pathway analysis

   1. Perform **Gene Set Enrichment Analysis** (GSEA)[29,30] based on ranking of transcripts according to a measurement of differentially expressed genes (DEGs) list to determine if the DEGs show statistically significant, concordant differences between biological conditions.

   2. Perform function analysis using resources such as **Gene Ontology**[31], **DAVID**[32,33], or other available software tools.

Representative Results

The methodology described above was applied to 67 FFPE samples that had been stored under a variety of different conditions for 7–32 years (the median sample storage time was 17.5 years). The dataset and analysis results presented here were previously described and published in Zhao et al.[11]. On checking the sample quality as described earlier (i.e., example traces in Figure 2), DV_{100} was found to be more useful than DV_{200} because it is more sensitive to accurately measure the proportion of smaller fragment sizes for highly degraded RNA samples.

In the given sample set, fewer than 10% of the samples (7 of 67) were above the DV_{200} cut off of 30%, as recommended by Illumina[34]. About 26% of the samples (19 of 67) had a DV_{100} > 60% (i.e., higher likelihood of generating good sequence data), 40% (27 of 67) were in the 40%-60% range for DV_{100} (i.e., acceptable, but with a lower likelihood of generating good sequence data), and about 10% (7 of 67) had a DV_{100} of <40% (i.e., very low likelihood of resulting in good sequence data). For 14 of 67 samples, the software was unable to determine the DV values. Table 1 shows a summary of QC metrics for the samples in different DV_{100} categories. For detailed QC analysis and data correlation for all 67 samples, please see Zhao et al.[11].

Given the high degree of degradation in the sample set, a ‘total RNA’ library preparation method was chosen, and sequencing libraries were prepared using the NEBNext Ultra II RNA Library Prep Kit for Illumina (Table of Materials). In order to improve the representation of the sequencing libraries in spite of the high degree of sample degradation, the maximum possible amount of RNA (1,000 ng when available) was used as input for library preparation. Additionally, the high degradation of the FFPE-RNA samples necessitated the rRNA depletion method, because the degraded transcripts were likely to not have the poly-A tails for mRNA capture. Following the depletion of ribosomal RNA by hybridization to specific probes and digestion of the hybridized transcripts using RNaseH, the remaining transcripts were converted into cDNA using random primers. Size selection was also avoided for libraries prepared from lower input samples. Example traces of final libraries are shown in Figure 3.

Highly degraded FFPE samples represent a great challenge for gene expression profiling in tumor samples. Thus, applying correct bioinformatics analysis methods and software tools is critical to detect artifacts or abnormalities in datasets to ensure high accuracy and reproducibility of gene quantification. The software tools used in this study are listed in the Supplementary Table. In the given sample set, we performed sequencing and library quality assessment, with some example metrics shown in Figure 4. An overview of raw fastq file sequencing quality and sample adapter content are shown in Figure 4A and Figure 4B, respectively. Fastqc screen can help detect contamination, such as bacterial and mouse contamination, in the samples as shown in Figure 4C. In the given sample set, 41 of 67 samples had 5%–48% bacterial contamination, and six samples had 4%–11% mouse contamination (Figure 4C). STAR alignment results (Figure 4D) showed the proportion of reads mapped to the reference genome, percentage of reads uniquely mapped to the reference genome, and proportion of reads that were not mapped or mapped to multiple loci. Picard CollectRNAStatistics was used to determine the percent mRNA, intronic, and intergenic bases present in the alignment files (Figure 4E). In order to assess the uniformity of read coverage on gene and transcripts, we used the Picard software tool to generate a gene body coverage plot, which measures the percentage of reads that cover each nucleotide position of all genes scaled into bins from 5' UTR to 3' UTR. Figure 4F shows that some degraded libraries had 3’ bias, where more reads are mapped closer to 3’ end than to the 5’ end.

FFPE samples usually have large variability in gene expression profiles that may arise due to variable degradation during sample storage, RNA extraction, or sample processing. It is important to use appropriate statistical methods to uncover the underlying patterns and measure the variation and correlation among samples. We applied Principal Component Analysis (PCA) for six pairs of biological replicates from a subset of the 67 FFPE samples. A PCA plot showed that 26% of total variation was captured by the first principal component and 19% from the second and third components combined (Figure 5). Among the six pairs of replicates, two pairs of replicates had higher variations (correlations below 0.22) than the last four samples (correlation values between 0.7–0.8) when comparing gene expression values between the replicate pairs. Because the replicates were generated by extracting RNA from two different tissue curls cut from the same FFPE blocks, the tissue age was not a factor in the higher variance here, and it was likely caused by the different

amount of bacterial contamination (1%–55%) as well as different mRNA content (2–3 fold difference) between the replicates. The randomness of mRNA degradation after extraction could also contribute to the higher variance between samples of similar origin.

Discussion

The method described here outlines the main steps required to obtain good sequence data from FFPE-RNA samples. The main points to consider with this method are: (1) Ensure that the RNA is preserved as best as possible after extraction by minimizing the sample handling and freezing and thawing cycles. Separate QC aliquots are very helpful. (2) Use a QC metric that is best for the given sample set. RIN values and DV_{200} are often not useful for degraded samples, and DV_{100} may be the metric of choice to assess the quality in a given sample set. (3) For more degraded samples, it is best to use a high sample input. Higher input amounts lead to better diversity and lower duplication in the final library, leading to improved data quality. Because not all RNA in FFPE-RNA samples is usable due to high degradation and refractoriness to enzymatic processes, these effects are more pronounced in FFPE-RNA compared to fresh frozen RNA. (4) Use random priming for the reverse transcription step as opposed to the use of oligo-dT or specific sequences as primers. Unless the set of specific probes is able to cover as much sequence as possible for all transcripts of interest, random primers are a safe bet to ensure the conversion of a maximum number of transcripts (or fragments thereof) into cDNA. Thus, total RNA library prep methods are more useful for degraded samples than mRNA methods, which rely on the presence of poly-A tails. (5) Accurate quantification of libraries by quantitative real-time PCR (qPCR) is important to avoid underperformance or overloading of the sequencers. (6) Assess potential contamination of the RNA as part of the standard post sequencing RNA-Seq QC protocols. Bacterial contamination and genomic DNA contamination are common for FFPE samples due to storage conditions and sample preparation procedures. Samples contaminated with foreign species can waste sequencing coverage, depending on the extent of contamination. In addition, internal contamination can arise from incomplete rRNA depletion, leading to a high percentage of reads mapping to rRNAs. Inefficient genomic DNA removal during DNase digestion could lead to false positive expression detection of transcripts or erroneous de novo assembly of transcripts. Adapter contamination introduced during library preparation is also a common problem for highly degraded RNAs with very short RNA fragments. Contamination can affect the gene and transcript profiling accuracy and lead to false discovery. Therefore, it is important to accurately identify the contamination sources and remove the contamination, if possible, during the sample or library preparation steps, or filter the contaminating reads during the data processing step. (7) Preprocessing and post-alignment quality control are important to detect bad quality and low mRNA content samples. Those samples should be eliminated from further analysis. Gene expression data from samples that generate low gene counts, poor coverage should be used with caution. (8) It is good practice to include biological replicates in order to measure samples variance and correlation to ensure data reproducibility.

FFPE samples represent a very valuable resource for a large number of diseases. The ability to obtain reliable sequence information from such samples would aid a lot of studies aimed at understanding the molecular mechanisms behind various disorders, resistance,

and susceptibility. Though the limitations imposed by the frequently suboptimal quality of RNA extracted from such samples do hamper such efforts, the steps described here help to mitigate those limitations to some extent and enable us to make the most of FFPE-RNA to obtain reliable gene expression information.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We are thankful to Dr. Danielle Carrick (Division of Cancer Control and Population Sciences, National Cancer Institute) for continued help, especially for initiating this study, providing us with the samples, and for helpful suggestions during data analysis. We sincerely thank all members of the CCR Sequencing Facility at the Frederick National Laboratory for Cancer Research for their help during sample preparation and sequencing, especially Brenda Ho for assistance in sample QC, Oksana German for library QC, Tatyana Smirnova for running the sequencers. We also would like to thank Tsai-wei Shen and Ashley Walton at Sequencing Facility Bioinformatics Group for helping with data analysis and RNA-seq pipeline implementation. We also thank CCBR and NCBR for assistance with RNaseq analysis pipeline and best practices development.

Disclosures

This work was funded by the National Cancer Institute (NCI), National Institutes of Health (NIH). Leidos Biomedical Research, Inc. is the operations and technical support contractor for the Frederick National Laboratory for Cancer Research which is fully funded by NIH. Several authors (YZ, MM, KT, YL, JS, BT) are affiliated with Leidos Biomedical Research, Inc., but all of the authors are fully funded by the National Cancer Institute including authors' salaries and research materials. Leidos Biomedical Research, Inc. did not provide salary for the authors (YZ, MM, KT, YL, JS, BT) or material for the study, nor did it have any role in the study design, data collection, analysis, decision to publish, or preparation of the manuscript.

References

1. Carrick DM et al. Robustness of Next Generation Sequencing on Older Formalin-Fixed Paraffin-Embedded Tissue. PLoS One. 10 (7), e0127353 (2015). [PubMed: 26222067]
2. Hedegaard J et al. Next-generation sequencing of RNA and DNA isolated from paired fresh-frozen and formalin-fixed paraffin-embedded samples of human cancer and normal tissue. PLoS One. 9 (5), e98187 (2014). [PubMed: 24878701]
3. Zhang P, Lehmann BD, Shyr Y, Guo Y The Utilization of Formalin Fixed-Paraffin-Embedded Specimens in High Throughput Genomic Studies. International Journal of Genomics. 2017, 1926304 (2017). [PubMed: 28246590]
4. Srinivasan M, Sedmak D, Jewell S Effect of fixatives and tissue processing on the content and integrity of nucleic acids. American Journal of Pathology. 161 (6), 1961–1971 (2002).
5. von Ahlfen S, Missel A, Bendrat K, Schlumpberger M Determinants of RNA quality from FFPE samples. PLoS One. 2 (12), e1261 (2007). [PubMed: 18060057]
6. Esteve-Codina A et al. A Comparison of RNA-Seq Results from Paired Formalin-Fixed Paraffin-Embedded and Fresh-Frozen Glioblastoma Tissue Samples. PLoS One. 12 (1), e0170632 (2017). [PubMed: 28122052]
7. Vukmirovic M et al. Identification and validation of differentially expressed transcripts by RNA-sequencing of formalin-fixed, paraffin-embedded (FFPE) lung tissue from patients with Idiopathic Pulmonary Fibrosis. BMC Pulmonary Medicine. 17 (1), 15 (2017). [PubMed: 28081703]
8. Adiconis X et al. Comparative analysis of RNA sequencing methods for degraded or low-input samples. Nature Methods. 10 (7), 623–629 (2013). [PubMed: 23685885]
9. Sinicropi D et al. Whole transcriptome RNA-Seq analysis of breast cancer recurrence risk using formalin-fixed paraffin-embedded tumor tissue. PLoS One. 7 (7), e40092 (2012). [PubMed: 22808097]

10. Altekruse SF et al. SEER cancer registry biospecimen research: yesterday and tomorrow. Cancer Epidemiology, Biomarkers & Prevention. 23 (12), 2681–2687 (2014).
11. Zhao Y et al. Robustness of RNA sequencing on older formalin-fixed paraffin-embedded tissue from high-grade ovarian serous adenocarcinomas. PLoS One. 14 (5), e0216050 (2019). [PubMed: 31059554]
12. Amini P et al. An optimised protocol for isolation of RNA from small sections of laser-capture microdissected FFPE tissue amenable for next-generation sequencing. BMC Molecular Biology. 18 (1), 22 (2017). [PubMed: 28835206]
13. Amini P, Nassiri S, Ettlin J, Malbon A, Markkanen E Next-generation RNA sequencing of FFPE subsections reveals highly conserved stromal reprogramming between canine and human mammary carcinoma. Disease Models and Mechanisms. 12 (8) (2019).
14. Wimmer I et al. Systematic evaluation of RNA quality, microarray data reliability and pathway analysis in fresh, fresh frozen and formalin-fixed paraffin-embedded tissue samples. Scientific Reports. 8 (1), 6351 (2018). [PubMed: 29679021]
15. Babraham Bioinformatics. <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/> (2019).
16. Martin M Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17 (1), 10–12 (2011).
17. Bolger AM, Lohse M, Usadel B Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 30 (15), 2114–2120 (2014). [PubMed: 24695404]
18. Babraham Bioinformatics. <https://www.bioinformatics.babraham.ac.uk/projects/fastq_screen/> (2019).
19. Wood DE, Salzberg SL Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biology. 15 (3), R46 (2014). [PubMed: 24580807]
20. Dobin A et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 29 (1), 15–21 (2013). [PubMed: 23104886]
21. Broad Institute. <http://broadinstitute.github.io/picard/> (2019).
22. Wang L, Wang S, Li W RSeQC: quality control of RNA-seq experiments. Bioinformatics. 28 (16), 2184–2185 (2012). [PubMed: 22743226]
23. Ewels P, Magnusson M, Lundin S, Kaller M MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 32 (19), 3047–3048 (2016). [PubMed: 27312411]
24. Li B, Dewey CN RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 12 323 (2011). [PubMed: 21816040]
25. Son K, Yu S, Shin W, Han K, Kang K A Simple Guideline to Assess the Characteristics of RNA-Seq Data. BioMed Research International. 2018 2906292 (2018). [PubMed: 30519573]
26. McCarthy DJ, Chen Y, Smyth GK Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Research. 40 (10), 4288–4297 (2012). [PubMed: 22287627]
27. Robinson MD, McCarthy DJ, Smyth GK edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 26 (1), 139–140 (2010). [PubMed: 19910308]
28. Ritchie ME et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research. 43 (7), e47 (2015). [PubMed: 25605792]
29. Subramanian A et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America U S A. 102 (43), 15545–15550 (2005).
30. Mootha VK et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature Genetics. 34 (3), 267–273 (2003). [PubMed: 12808457]
31. Ashburner M et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics. 25 (1), 25–29 (2000). [PubMed: 10802651]

32. Huang da W, Sherman BT, Lempicki RA Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols. 4 (1), 44–57 (2009). [PubMed: 19131956]

33. Huang da W, Sherman BT, Lempicki RA Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Research. 37 (1), 1–13 (2009). [PubMed: 19033363]

34. Illumina. Evaluating RNA Quality from FFPE Samples. <https://www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/evaluating-rna-quality-from-ffpe-samples-technical-note-470-2014-001.pdf> (2016).

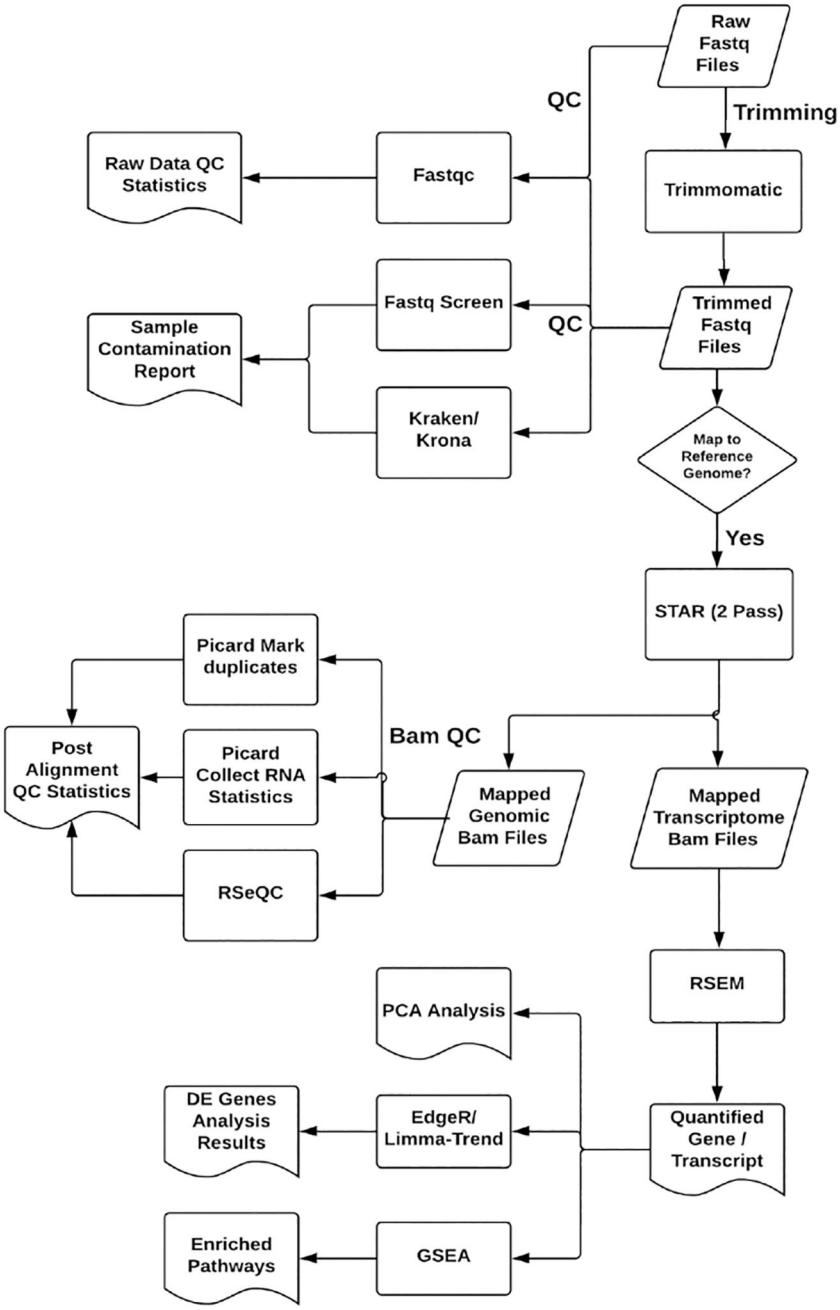

Figure 1: RNaseq analysis workflow.

The flowchart describes the analysis steps for preprocessing, quality assessment, mapping to reference, gene quantification, and differential analysis between different sample groups.

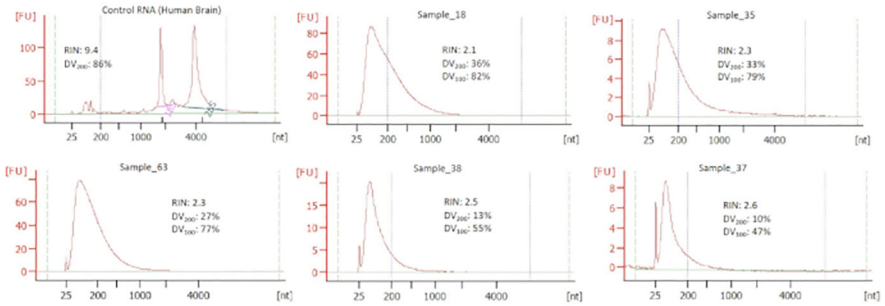

Figure 2: Example Bioanalyzer traces of six different FFPE-RNA samples.

The horizontal axis denotes the molecular weight (bp) and fluorescence units (FU) and the vertical axis shows the concentration of different sized fragments. The RNA Integrity Numbers (RIN), DV_{200} (i.e., percent of fragments >200 bp), and DV_{100} (i.e., percent of fragments >100 bp) values are indicated on each profile. A 25 bp peak in each profile indicates the molecular weight marker.

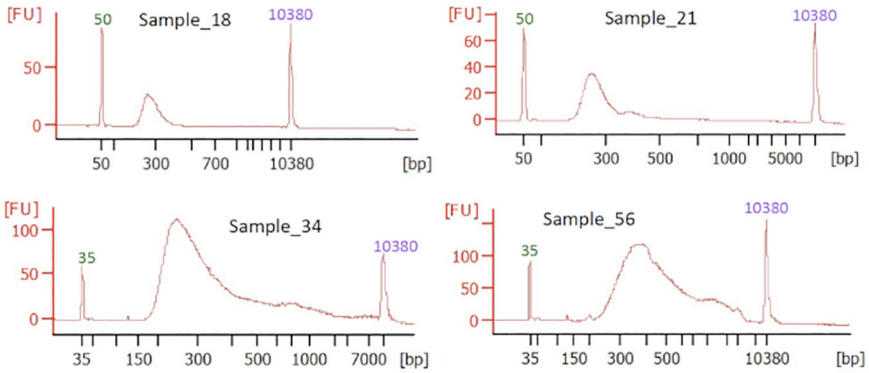

Figure 3: Example Bioanalyzer traces of final libraries prepared from four different samples.
The horizontal axis denotes the molecular weight (bp) and fluorescence units (FU) on the vertical axis indicate the concentration of different sized fragments. The lower (35 bp or 50 bp) and upper (10,380 bp) marker peaks are labeled in green and purple, respectively.

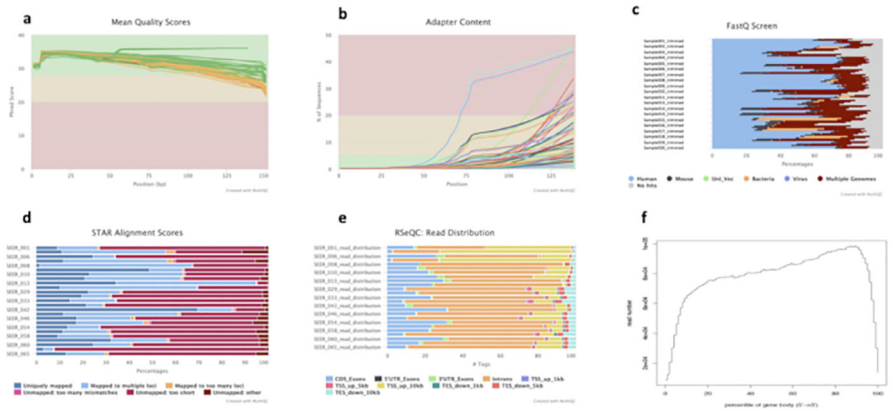

Figure 4: Example multi-QC report for preprocessing QC results.

(**A**) Line chart showing the percentages of Q30 bases of all sequencing reads in each sample. (**B**) Sequencing adapter content in raw fastq files. (**C**) Contamination screen to check closely matched species. (**D**) Genome mapping statistics. (E) Read distribution based on Gencode gene annotation. (**F**) Gene body/transcript coverage

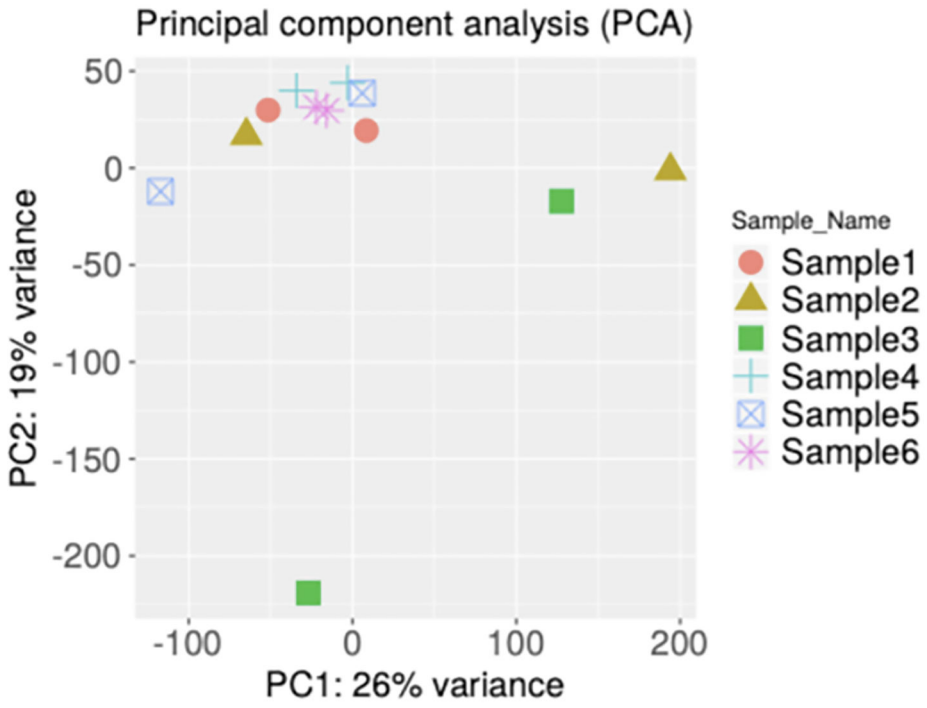

Figure 5: Example PCA analysis to show sample group concordance.

PCA analysis for biological replicates. PCA plot with samples plotted in two dimensions using their projections onto the first two principal components. Biological replicates are shown in the same color.

Table 1:

Summary of sample set QC metrics.

	Number of samples	Median Input for lib prep (ng)	Median RIN	Median DV_{200}	Median DV_{100}	Median Lib size (bp)	Median Lib yield (ng)	Median Lib Molarity (nM)	Median Specimen storage time (Years)	Median % contamination	Median Gene Count
DV100 <40%	7	237.6	2.5	6	34	445	24.5	7	22	27.4	14,759
DV100 40–60%	27	1000	2.5	12	51	408	19.8	5.9	18	9.9	10,202
DV100 >60%	19	1000	2.3	26	73	355	84.9	24	13	3.2	9,993

The table shows the QC metrics of the samples, grouped according to their DV_{100} values. The number of samples in each group is listed, and median values for each metric are shown.

Materials

Name	Company	Catalog Number	Comments
2100 Bioanalyzer	Agilent	G2939BA	
Agilent DNA 7500 Kit	Agilent	5067–1506	
Agilent High Sensitivity DNA Kit	Agilent	5067–4626	
Agilent RNA 6000 Nano Kit	Agilent	5067–1511	
AllPrep DNA/RNA FFPE Kit	Qiagen	80234	
CFX96 Touch System	Bio-Rad	1855195	
Library Quantification kit v2-Illumina	KapaBiosystems	KK4824	
NEBNext Ultra II Directional RNA Library Prep Kit for Illumina	New England Biolabs	E7765S	https://www.neb.com/protocols/2017/02/07/protocol-for-use-with-ffpe-rna-nebnext-rrna-depletion-kit
NEBNext rRNA Depletion Kit (Human/Mouse/Rat)	New England Biolabs	E6310L	
NextSeq 500 Sequencing System	Illumina	SY-415–1001	NextSeq 500 System guide: https://support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/nextseq/nextseq-500-system-guide-15046563-06.pdf
NextSeq PhiX Control Kit	Illumina	FC-110–3002	
NSQ 500/550 Hi Output KT v2.5 (150 CYS)	Illumina	20024907	
10X Genomics Magnetic Separator	10X Genomics	120250	
Rotator Multimixer	VWR	13916–822	
C1000 Touch Thermal Cycler	Bio-Rad	1851197	
Sequencing reagent kit	Illumina	20024907	
Flow cell package	Illumina	20024907	
Buffer cartridge and the reagent cartridge	Illumina	20024907	
Sodium hydroxide solution (0.2N)	Millipore Sigma	SX0607D-6	
TRIS-HCL Buffer 1.0M, pH 7.0	Fisher Scientific	50–151–871	