Received: 10 August 2021 | Revised: 6 October 2021 | Accepted: 6 October 2021

DOI: 10.1002/gcc.23002

RESEARCH ARTICLE

WILEY

Transcriptome sequencing of archived lymphoma specimens is feasible and clinically relevant using exome capture technology

Aron Skaftason[1] | Ying Qu[1] | Maysaa Abdulla[2] | Jessica Nordlund[3] | Mattias Berglund[4] | Susanne Bram Ednersson[5,6] | Per-Ola Andersson[7,8] | Gunilla Enblad[4] | Rose-Marie Amini[2] | Richard Rosenquist[1,9] | Larry Mansouri[1]

[1]Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden

[2]Clinical and Experimental Pathology, Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala, Sweden

[3]Department of Medical Sciences and Science for Life Laboratory, Uppsala University, Uppsala, Sweden

[4]Experimental and Clinical Oncology, Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala, Sweden

[5]Department of Pathology, Sahlgrenska University Hospital, Gothenburg, Sweden

[6]Institute of Biomedicine, Sahlgrenska Academy at the University of Gothenburg, Gothenburg, Sweden

[7]Department of Medicine, Section of Hematology, South Älvsborg Hospital, Borås, Sweden

[8]Institute of Medicine, Sahlgrenska Academy at the University of Gothenburg, Gothenburg, Sweden

[9]Clinical Genetics, Karolinska University Laboratory, Karolinska University Hospital, Stockholm, Sweden

Correspondence
Richard Rosenquist, Department of Molecular Medicine and Surgery, Karolinska Institutet, 171 76 Stockholm, Sweden.
Email: richard.rosenquist@ki.se

Funding information
Cancerfonden; Knut och Alice Wallenbergs Stiftelse; Radiumhemmets Forskningsfonder; Vetenskapsrådet; Swedish Research Council; Uppsala Universitet; Science for Life Laboratory

Abstract

Formalin-fixed, paraffin-embedded (FFPE) specimens are an underutilized resource in medical research, particularly in the setting of transcriptome sequencing, as RNA from these samples is often degraded. We took advantage of an exome capture-based RNA-sequencing protocol to explore global gene expression in paired fresh–frozen (FF) and FFPE samples from 16 diffuse large B-cell lymphoma (DLBCL) patients. While FFPE samples generated fewer mapped reads compared to their FF counterparts, these reads captured the same library complexity and had a similar number of genes expressed on average. Furthermore, gene expression demonstrated a high correlation when comparing housekeeping genes only or across the entire transcriptome ($r = 0.99$ for both comparisons). Differences in gene expression were primarily seen in lowly expressed genes and genes with small or large coding sequences. Using cell-of-origin classifiers and clinically relevant gene expression signatures for DLBCL, FF, and FFPE samples from the same biopsy paired nearly perfectly in clustering analysis. This was further confirmed in a validation cohort of 50 FFPE DLBCL samples. In summary, we found the biological differences between tumors to be far greater than artifacts created as a result of degraded RNA. We conclude that exome capture transcriptome sequencing data from archival samples can confidently be used for cell-of-origin classification of DLBCL samples.

Richard Rosenquist and Larry Mansouri contributed equally to the study as last authors.

KEYWORDS
capture-based RNA-sequencing, degraded RNA, DLBCL, gene expression profiling

1 | INTRODUCTION

Biopsies taken for morphological assessment and molecular investigation in clinical diagnostics are commonly stored as formalin-fixed, paraffin-embedded (FFPE) tissue. It is estimated that across governmental biobanks, universities, and hospitals, there exist over one billion FFPE samples worldwide.[1] These samples offer well-documented retrospective cohorts, often with lifetime follow-up data, and constitute a potentially invaluable resource in medical research where sample availability is often a major limitation. That said, several factors including handling and processing, preservation conditions, and duration of storage affect sample nucleic acid integrity. This typically results in poor-quality RNA, which has limited the usage of FFPE samples, especially in transcriptome sequencing settings. Recently, a novel RNA capture approach was developed for profiling samples from degraded tissue using next-generation sequencing.[2] This method takes advantage of exon-targeting probes to enrich for coding RNA fragments and covers as much as 97% of the coding genome with high read mappability (94%–96%) from degraded RNA,[3] providing an attractive option for investigating FFPE samples.

Diffuse large B-cell lymphoma (DLBCL) is the most common lymphoid malignancy in adults and accounts for about 35% of all non-Hodgkin lymphomas.[4] The disease originates from mature B-cells and, based on gene expression profiling (GEP), is subclassified in to clinically relevant subtypes, that is, germinal center B-cell (GC), activated B-cell (ABC), and an unclassified subtype, with distinct outcomes.[5–9] In clinical settings, various alternative approaches have been developed that are primarily based on immunohistochemistry to subclassify DLBCL patients as defined by GEP. Among these, one of the most commonly employed algorithms, proposed by Hans et al., uses staining of CD10, BCL6, and MUM1 and subsequently subgroups the patients into GC or non-GC (NGC) subgroups.[10] More recently, the lymphoma/leukemia molecular profiling project (LLMPP) developed a digital 20 gene expression-based test, the Lymph2Cx assay using NanoString technology, which has been shown to assign robustly DLBCL subtypes in FFPE tissue.[11,12]

Although a number of studies have investigated different RNA-sequencing (RNA-seq) protocols for degraded and low-quantity samples,[3,13–20] there is none that, to the best of our knowledge, has applied the coding RNA enrichment assay and investigated clinically relevant gene expression signatures to assess performance in paired FFPE/fresh–frozen (FF) routine samples. Therefore, in this study, we used DLBCL as a model disorder to evaluate the Illumina TruSeq RNA Exome (formerly TruSeq RNA Access, Illumina, San Diego, CA) protocol which is specifically designed for degraded RNA, in paired archival samples from the same biopsy to evaluate the feasibility and validity of FFPE RNA-seq.

2 | MATERIALS AND METHODS

2.1 | Patient samples

Paired FF and FFPE samples (stored at room temperature) from the same excisional biopsy were collected prior to treatment from 16 DLBCL patients (9 GC and 7 NGC based on the Hans classification[10]) from the biobank at the Department of Pathology, Uppsala University Hospital, Sweden. Median time from sampling to RNA extraction was 6.0 years (range, 1–11 years). A second independent cohort of FFPE samples (stored at room temperature) from 50 Swedish DLBCL patients, mainly from Sahlgrenska University Hospital, Gothenburg, Sweden, was also used for validation. All samples were evaluated by experienced hematopathologists (MA, SBE, and RMA) both at diagnosis and upon inclusion in the study according to the WHO classification.[21] The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the local Ethics Review Committee.

2.2 | RNA extraction and quality assessment

Extraction of total RNA was performed on FF samples using the Allprep DNA/RNA/miRNA Universal Kit (Qiagen, Valencia, CA), and FFPE samples from ≥ three 10-μm tissue sections using the AllPrep DNA/RNA FFPE Kit (Qiagen). For all samples, on-column DNase digestion (Qiagen) was carried out to remove genomic DNA. Total RNA was eluted in 30 μL RNase-free water. RNA quantity was measured by fluorometric quantitation using the Qubit RNA HS assay kit (ThermoFisher Scientific, Waltham, MA) while RNA integrity was assessed with the 4200 TapeStation System (Agilent Technologies, Waldbronn, Germany) using the DV_{200} metric (the percentage of fragments > 200 nucleotides). Samples were stored in −80°C directly after extraction.

2.3 | Library construction and sequencing

The Illumina TruSeq RNA Exome protocol (Illumina) was employed to prepare RNA-seq libraries for both FF and FFPE samples according to manufacturer's protocol. For samples with DV_{200} values of < 50%, 100 ng input RNA was used for library construction, while for samples with DV_{200} values of 50%–70% or > 70%, 40 and 20 ng input RNA was used, respectively, as indicated by the manufacturer's protocol. For the paired FF/FFPE samples, libraries were constructed in four separate batches and sequenced in pools of eight samples, each run in one lane of a HiSeq Flow Cell v4 and using HiSeq SBS Kit v4 chemistry on a HiSeq 2500 instrument in paired-end 2 × 125 bp mode. For the validation cohort, the libraries were constructed in two separate batches and sequenced using the same protocol.

2.4 | Bioinformatics

Raw sequencing reads were aligned using the nf-core/RNA-seq (1.0), a pipeline written in Nextflow.[22] Briefly, raw reads were adapter trimmed with the help of Trim Galore (0.5.0)[23] using standard parameters and aligned to the reference genome (hg19) using STAR (2.6.1).[24] Duplicate reads were estimated with Picard's MarkDuplicates (2.18.14)[25] and Dupradar (1.8.0).[26] Quality of reads was determined using FastQC (0.11.7)[27] and RSeQC (2.6.4)[28] and quality metrics were summarized with the help of MultiQC (1.6).[29] Counts were retrieved with FeatureCounts[30] and counts for protein-coding genes were normalized with DESeq.[31] Counts were transformed using variance stabilizing transformation[32] or regularized logarithmic transformation[31] for visual representation. Transcript lengths were retrieved from the Ensembl database.[33] Downstream statistical analysis was performed, and all images were generated, in R (3.4.4).

2.5 | Assessing cell-of-origin using gene expression signatures and the NanoString assay

We used three different published signatures to assess our dataset. The RNA-seq-based classifier (RNA-seq classification) initially developed by Wright et al. and applied in Reddy et al. was used for cell-of-origin (COO) classification in paired FF and FFPE samples (Table 1S).[6,34] A DLBCL automatic classifier developed by Barrans et al. and based on published data from the LLMPP, consisting of 20 classifier genes combined with ABC/GC signature genes and totaling 143 genes (133 genes in our dataset, Tables 1S and 2S) was used to compare gene expression patterns in paired FF and FFPE samples.[6,8,11] In addition, a B-cell–associated gene set defined by Dybkær et al. based on subset-specific B-cell-associated gene signatures (BAGS) for naive, centrocyte, centroblast, memory, and plasmablast B cells comprising 223 genes (180 genes in our dataset, Table 3S) was employed for the same purpose.[35]

In addition to these signatures, the Lymph2Cx assay and Research Use Only version of the NanoString Lymphoma Subtyping Test (NanoString Technologies, Seattle, WA, USA) was utilized to classify COO subtype for all FFPE samples according to manufacturer's protocol.[11,36] Briefly, gene expression data from five housekeeping genes were initially analyzed to assess input RNA quality, while gene expression results from the gene panel included in the Lymphoma Subtyping Test were applied to create a linear predictor score to classify cases as ABC, GC, or unclassified subtypes (Supporting Information Table 1S).

3 | RESULTS

3.1 | RNA isolation and integrity

To evaluate the performance of the Illumina TruSeq RNA Exome protocol in degraded samples, we performed RNA-seq on matched FF and FFPE samples from the same biopsy in 16 DLBCL cases. Sample integrity was assessed prior to library preparation using the DV_{200} metric. We confirmed that FFPE samples displayed significantly lower RNA integrity compared to their FF counterparts (median DV_{200} value of 39% vs. 83%, $p < 0.001$, Figure 1A and Table 4S). Furthermore, we found no correlation between storage time and RNA integrity for the FFPE samples (Figure 1B). For two cases, the FFPE samples did not meet the recommended RNA quality measurements for the RNA Exome protocol ($DV_{200} \geq 30\%$); these cases had instead DV_{200} values of 21% (sample FFPE_15) and 27% (sample FFPE_4). Post alignment quality control assessment did, however, indicate sufficient quality for further downstream analysis and these samples were included in the study for comparative purposes and are highlighted in the analyses and further discussed below. All samples for the validation cohort displayed a DV_{200} value > 30%.

3.2 | Sequencing of RNA-seq libraries

The RNA libraries were sequenced on an Illumina HiSeq 2500 instrument in paired-end 2 × 125 bp mode. The runs generated an average of 32.1 M mapped reads (range, 17.1–67.5 M) with average insert read lengths of 247 and 233 bases for the FF and FFPE samples, respectively ($p < 0.001$, Figure 2A and Table 4S). We observed differences in proportion of all mapped reads (97.4% ± 0.5 vs. 96.7% ± 0.6, average ± SD, $p < 0.01$) and in reads mapping to coding exon sequences (96.6% ± 0.7 vs. 94.7% ± 0.8, average ± SD, $p < 0.001$) when comparing FF to FFPE tissue, respectively (Figures 2B, C), values highly similar to previously published data.[3] All samples displayed high base sequence quality indicated by the Phred quality score with no difference between FF and FFPE samples (Figure 1S). The sequencing runs achieved highly comparable library complexity estimated as the percentage of non-duplicate (unique) reads among all counted fragments (84.1% ± 4.0 vs. 84.4% ± 2.1, average ± SD) for FF

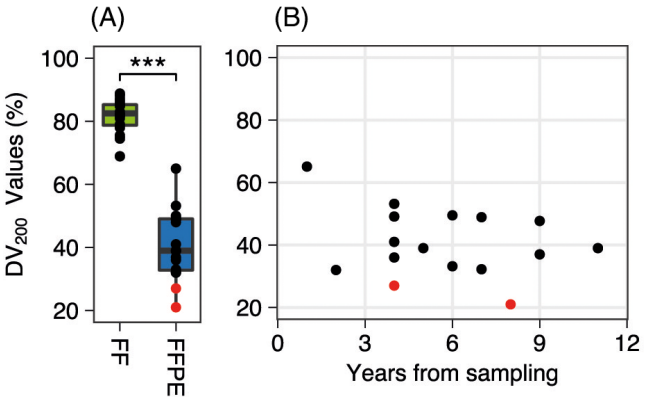

FIGURE 1 Assessment of RNA integrity. (A) Boxplot indicating median and inte-quartile DV_{200} values for paired fresh–frozen and formalin-fixed, paraffin-embedded (FFPE) samples included for Illumina TruSeq RNA Access protocol. (B) DV_{200} values versus time from sampling to RNA isolation for FFPE samples. Red dots represent samples with $DV_{200} < 30\%$. FF indicates fresh–frozen tissue. ***$p < 0.001$

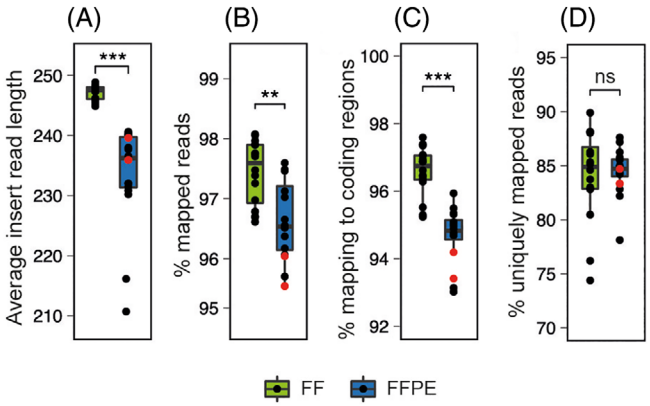

FIGURE 2 Sequencing reads metrics for paired fresh–frozen and formalin-fixed, paraffin-embedded (FFPE) tissue. (A) Average insert read length, (B) % mapped reads, (C) % reads mapping to exon coding sequence, and (D) average library complexity defined as uniquely mapped reads to coding RNA sequences in FF and FFPE samples. Red dots represent samples with DV_{200} < 30%. Boxplots indicate median and interquartile values. $^{**}p < 0.01$ while $^{***}p < 0.001$

and FFPE tissue (Figure 2D). Although the two aforementioned samples with RNA integrity values lower than recommended (marked in red in Figure 2) displayed less than the average percent mapped reads as well as percent reads mapping to coding sequences (Figures 2B, C), they were not found to deviate from other FFPE samples when calculating the proportion of uniquely mapped reads (Figure 2D) and were thus included for further analyses. The sequencing read metrics are summarized in Table 3S. Finally, in order to investigate potential batch effects among the different batches of library preparations and sequencing runs, we evaluated all samples using principal component analysis and noted no difference in global expression from the different sequencing runs (Figure 2S).

3.3 | Transcript coverage and expression

We compared the variation in gene body coverage in FF versus FFPE tissue measured as number of reads for each coding base along the length of each gene in 5′–3′ direction. We detected a highly similar

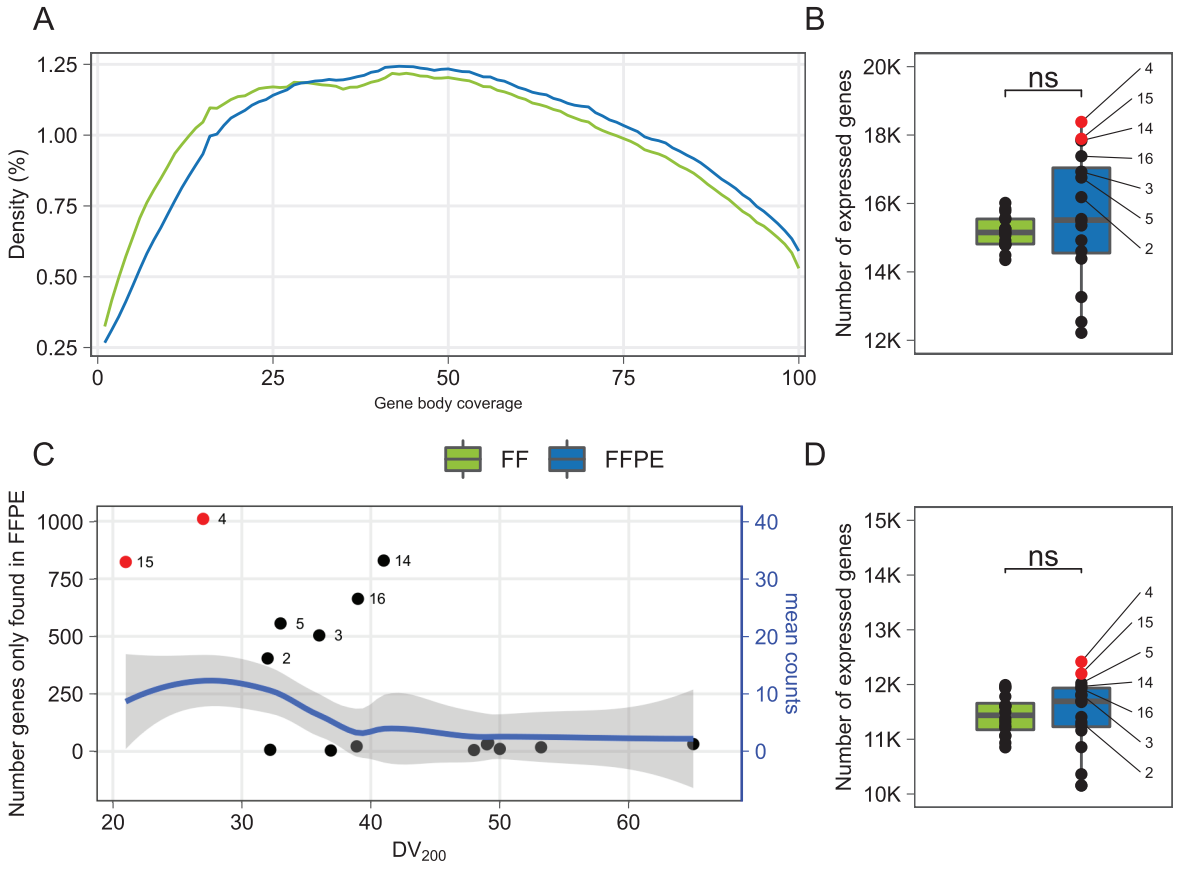

FIGURE 3 Gene expression analysis in paired fresh–frozen and formalin-fixed, paraffin-embedded (FFPE) samples. (A) Relative read density along mapped transcripts in 5′–3′ direction. Green and blue lines represent FF and FFPE samples, respectively. (B) Boxplot indicating median and interquartile number of detected genes in the two tissue types. (C) Number of detected genes exclusively found in FFPE tissue as a function of RNA DV_{200} value. The blue line represents their mean expression per sample using locally weighted smoothing and shaded areas 95% confidence interval. (D) Boxplot indicating median and interquartile number of detected genes excluding genes with less than 10 mapped reads. Red dots represent samples with DV_{200} < 30%. FF indicates fresh–frozen tissue while ns indicates not significant

pattern in read distribution with both tissue types showing a bias towards coding bases in the second and third quartiles along the length of all transcripts when compared to either 5′ or 3′ regions (Figure 3A). We then investigated the number of annotated coding genes expressed in our datasets. In order to get an unbiased view of the data, we included all uniquely mapped transcripts. This resulted in an average of 15.2 K (range, 14.4–16.0 K) expressed genes in FF tissue and 15.6 K (range, 12.2–18.4 K) expressed genes for FFPE samples (Figure 3B).

We next explored potential differences in genes expressed and their expression levels when comparing FF and FFPE tissue. In our dataset, 17 937 genes were detected to be co-expressed in at least one patient and both tissue types. These genes had an average expression of 1307 counts across all samples. In addition, a total of 143 genes were found exclusively expressed in FF tissue, while in contrast, 1097 genes were only detected in FFPE samples. However, these tissue-specific genes were expressed at far lower levels; on average 0.5 counts per gene and sample exclusively expressed in FF tissue, and 2.4 counts per gene and sample expressed only in FFPE tissue. The majority of genes in the latter category were detected in samples displaying the lowest DV_{200} values, including the two cases with DV_{200} values < 30% (Figure 3C). Excluding these two cases, the average counts per gene and sample expressed only in FFPE tissue was instead 1.4 indicating that these transcripts might represent artifacts in the dataset. To explore further the potential effect of these lowly expressed genes in the dataset, we employed an absolute cutoff of at least 10 mapped reads per gene. This resulted in a significantly reduced number of expressed genes in both FF and FFPE tissues and, more importantly, greatly reduced the inter-tissue variability among the FFPE samples (Figure 3D).

In order to get a view of gene size distribution in our sequencing libraries, we plotted the gene density of all expressed genes as a function of gene length. Among these genes, 95% displayed a length of 750–4250 coding bases with the most frequent gene length found around 1600 bases (Figure 4A). We then plotted mean normalized expression of uniquely mapped coding transcripts for both FF and FFPE samples as a function of gene length. Here, genes within the mentioned range had uniformly high expression and displayed the low relative differences in expression when comparing tissue types and low variance within each tissue type (Figure 4A). In contrast, we observed notable differences for small and large genes in gene expression when comparing FF and FFPE samples outside the mentioned size range. Based on these findings, we excluded low expressing genes defined as having fewer than 10 mapped reads on average for the dataset or genes having extremely small or large size as described above and found a notable reduction in variability in the number of genes expressed when comparing FF (11.7 K, range 11.0–12.5 K) and FFPE (12.0 K, range 10.3–13.6 K) samples (statistically not significant, Figure 4B).

Regardless, Pearson's product–moment correlation analysis of expressed genes in all samples plotted as FFPE versus FF tissue revealed a very high correlation when looking at reference genes[37] only (Figure 5A, $r = 0.99$, $p < 0.0001$) or when including all uniquely mapped coding transcripts, including those having low expression or the ones exclusively expressed in one tissue type (Figure 5B, $r = 0.99$, $p < 0.0001$).

3.4 | Evaluation of COO and gene signatures

In order to evaluate the feasibility of RNA-seq in FFPE tissue, we first applied the RNA-seq-based classifier developed by Reddy et al.[34] (RNA-seq classification) for COO classification. Using this algorithm, both the FF and the FFPE sample within each sample

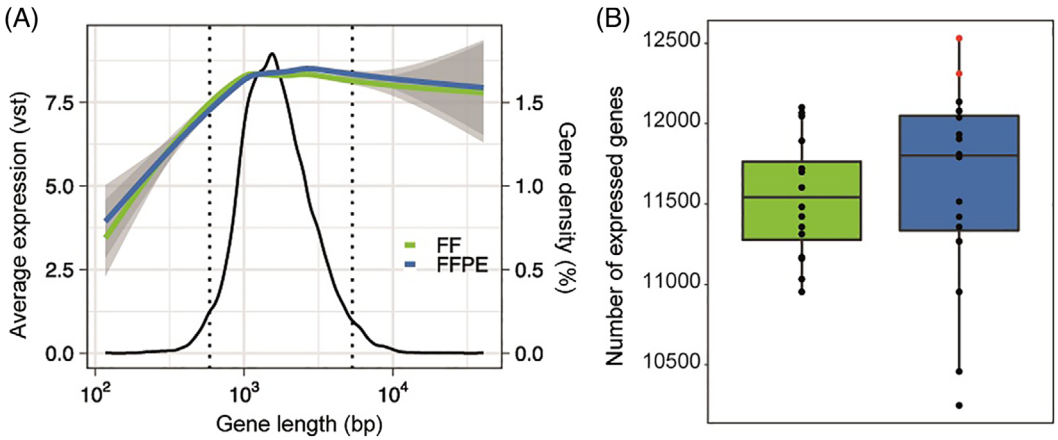

FIGURE 4 Assessing similarities in gene expression in paired fresh–frozen and formalin-fixed, paraffin-embedded (FFPE) samples (A) Gene expression in relation to gene length in paired fresh–frozen and FFPE tissue. The black line represents relative gene density versus gene length. Green and blue lines show average expression displayed using variance stabilizing transformation (VST) for FF and FFPE tissue, respectively, while the shaded areas indicate 95% confidence interval. Dotted vertical lines represent the lower (750 bp) and the upper (4250 bp) limits for the 95% most common gene lengths in the dataset. (B) Boxplot indicating median and interquartile number of detected genes in the two tissue types filtered for minimum 10 average counts and gene length. FF indicates fresh–frozen tissue

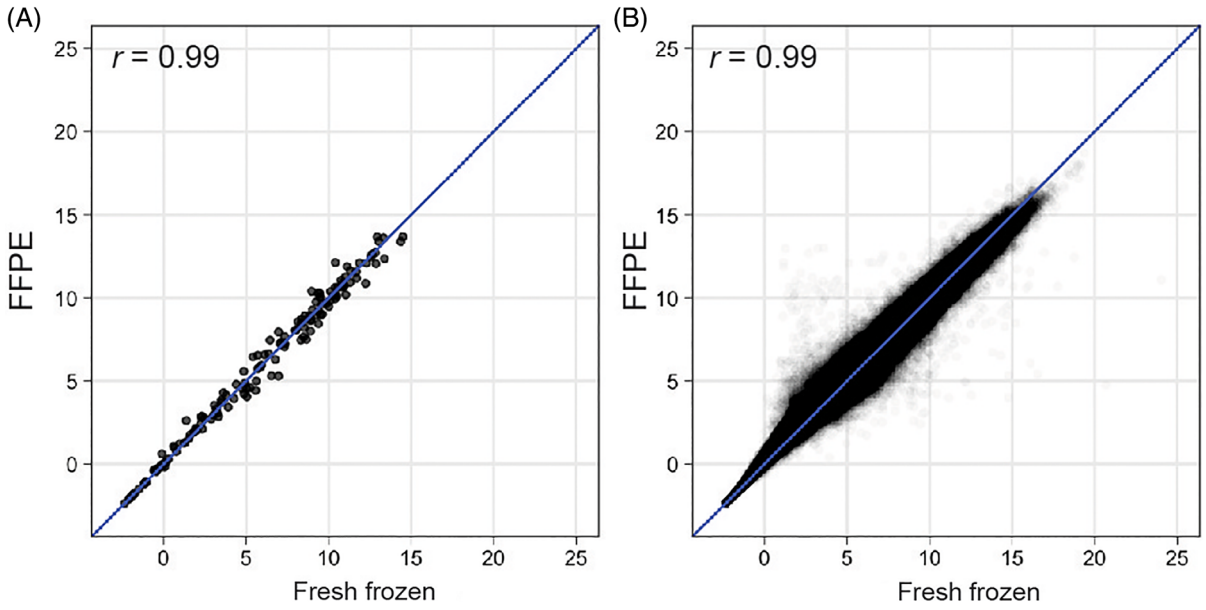

FIGURE 5 Gene expression correlation for all samples comparing paired fresh–frozen (FF) and formalin-fixed, paraffin-embedded (FFPE) tissue. (A) Housekeeping genes (*C1orf43*, *CHMP2A*, *EMC7*, *GPI*, *PSMB2*, *PSMB4*, *RAB7A*, *REEP5*, *SNRPD3*, *VCP*, and *VPS29*)[37] only and (B), all mapped coding reads. R_{log} transformed expression values are displayed

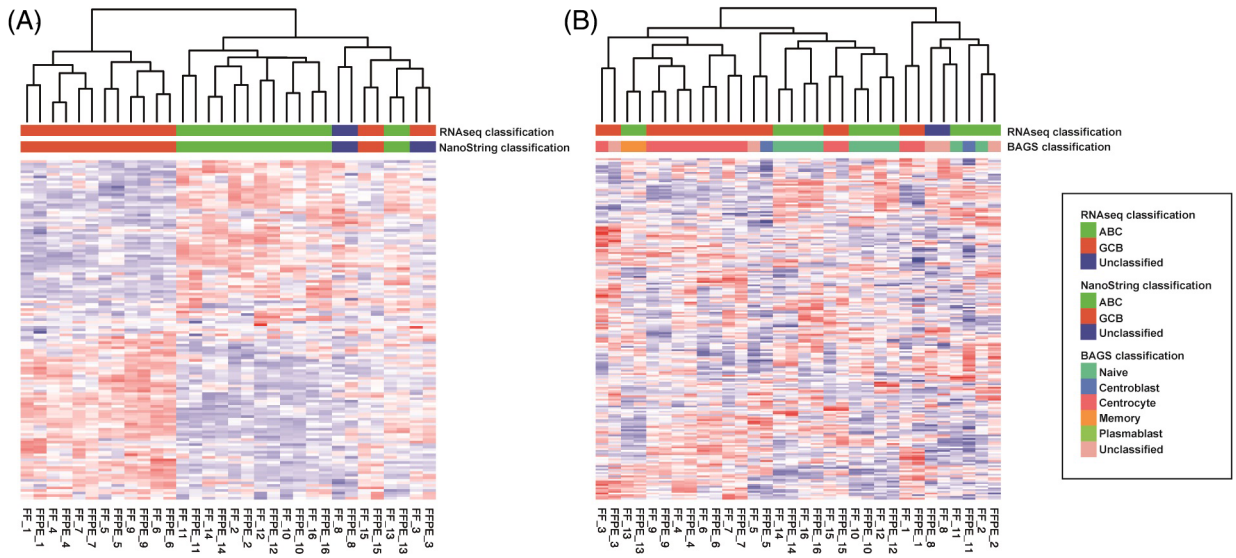

FIGURE 6 Hierarchical clustering of all samples based on the gene lists defined by (A) Barrans et al.[8] and (B) Dybkaer et al.[35] The RNA-seq classification based on the Reddy et al.[34] and the nanostring classification[11,36] are also provided

pair were classified as having identical COO without exception (Figure 6A). For the most part, subtype classification based on RNA-seq data confirmed the results from the Hans algorithm except for three cases. Samples 2 and 8, which were classified as GC subtype samples using the Hans algorithm, were instead categorized as ABC and unclassified subtypes with the RNA-seq classification, while Sample 15, an NGC subtype sample using the Hans algorithm, was instead categorized as GC using the RNA-seq classifier (Table 1).

Next, we applied hierarchical clustering and previously published gene signatures that have been shown to classify DLBCL samples into biologically and clinically relevant subgroups to investigate similarity in expression within paired samples and across patients.[8,35] We first employed a gene set published by Barrans et al. based on data from the LLMPP.[8,11] Here, all FFPE samples paired together with their FF counterparts from the same biopsy (Figure 6A). We repeated the same exercise, this time using the B-cell–associated gene signature defined by Dybkær et al.[35] (Figure 6B). Similar to the Barrans et al.

TABLE 1 Cell-of-origin subtype classification of the formalin-fixed, paraffin-embedded (FFPE) diffuse large B-cell lymphoma (DLBCL) samples using the Hans,[10] Reddy et al.[34] (RNA-seq classification) and NanoString algorithms[11,36]

Sample	Hans classification	RNA-seq classification	RNA-seq subtype score	NanoString classification	NanoString quality score	Nanostring LPS
FFPE_9	GC	GC	−2.249	GC	Pass	−227
FFPE_6	GC	GC	−1.932	GC	Pass	333
FFPE_1	GC	GC	−1.886	GC	Pass	334
FFPE_4[a]	GC	GC	−1.78	GC	Pass	146
FFPE_5	GC	GC	−1.582	GC	Pass	352
FFPE_7	GC	GC	−1.131	GC	Pass	657
FFPE_15[b]	NGC	GC	−0.388	GC	Borderline	1247
FFPE_3	GC	GC	−0.255	Unclassified	Pass	1921
FFPE_8	GC	Unclassified	0.048	Unclassified	Pass	2270
FFPE_13	NGC	ABC	0.324	ABC	Pass	2507
FFPE_10	NGC	ABC	0.748	ABC	Pass	3063
FFPE_16	NGC	ABC	1.153	ABC	Pass	3006
FFPE_2	GC	ABC	1.283	ABC	Pass	3157
FFPE_14	NGC	ABC	1.623	ABC	Pass	3511
FFPE_11	NGC	ABC	1.869	ABC	Pass	3669
FFPE_12	NGC	ABC	2.397	ABC	Pass	3877

Note: FFPE samples with poor RNA integrity. Samples have been sorted in ascending order based on the RNA-seq subtype score.
Abbreviations: ABC, activated B-cell; GC, germinal center B-cell; LPS, linear predictor score; NGC, non-GC.
[a]DV200 = 27%.
[b]DV200 = 21%.

gene set, most FFPE samples paired together with their FF counterparts with two exceptions (Samples 2 and 8) where the sample pairs instead clustered closely within the same branch (Figure 6B).

3.5 | NanoString assay

All samples were analyzed using the Lymph2Cx assay to determine COO. One sample (FFPE_15, DV200 value of 21%) returned a borderline quality score but was included in the analysis. Altogether, data from the NanoString assay correlated very strongly with the findings from RNA-seq analysis using the Reddy et al. algorithm for COO classification[34] (Table 1), with the only exception being Sample 3 which was classified as a GC subtype using RNA-seq data, while the NanoString assay identified the same sample as belonging to the unclassified subtype.

Finally, we also investigated COO in a validation cohort consisting of 50 FFPE DLBCL samples using the Hans, RNA-seq (Reddy et al.), and NanoString algorithms. Overall, the results correlated well when comparing the different methods. Interestingly, the highest concordance was found between the RNA-seq and Hans algorithms where 21/21 ABC samples and 44/50 samples (88%) were classified as having the same COO (Table 5S). This was followed by the comparison between RNA-seq and the NanoString assay, which showed 44/50 (88%) concordance, while the Hans and NanoString algorithms displayed 41/50 (82%) concordance (Table 5S).

4 | DISCUSSION

The primary objective of this study was to assess the performance of the Illumina Truseq RNA Exome protocol for the analysis of global gene expression from degraded RNA in archival samples. To this end, we performed RNA-seq on paired FF and FFPE samples from the same biopsy in 16 DLBCL samples. As expected, RNA extracted from FFPE tissue displayed significantly lower sample integrity when compared to RNA isolated from FF tissue. However, there was no correlation between storage time and FFPE sample RNA integrity. The latter observation is also in line with previous findings that for FFPE samples stored at room temperature, there is a significant and rapid decline in RNA quality within the first 6 months after which time the samples become stable.[38] The median time to sample isolation for our FFPE cohort in this study was 6 years (ranging up to 11 years) implying that sample age is not necessarily a limiting factor when choosing cases for analysis.

Mapped read metrics for the paired samples revealed that FF samples had longer average read lengths after adapter trimming and more frequently mapped to coding sequences. In addition, for several samples, sequencing data from FF tissue generated significantly higher number of total reads (Table 4S). Having said that, the FFPE tissue samples averaged 27.5 million mapped reads with 94.7% mappability to coding sequence and showed high base quality. Given that RNA-seq experiments using intact RNA to study global gene expression often make use of ribosomal RNA depletion protocols with far lower

mappability to the coding exome (commonly around 60% of reads) and a typical targeted sequencing depth of around 30–40 million reads, the sequencing depth obtained in our experiments with FFPE RNA appears at the very least to be on a comparable level. Additionally, in our experiments, among the reads that did map to coding sequences, no difference was detected in the proportion of uniquely mapped reads when comparing the two tissue types, suggesting that library complexity is not diminished in FFPE tissue. This was also true for the two samples having RNA integrity values below the recommended threshold for the protocol (DV_{200} > 30%). We next explored the read distribution along the length of all expressed genes in our dataset and found a striking similarity in the two tissue types. We also noted that mapped reads were predominantly distributed in the second and third quartiles when compared to the 5′ or 3′ regions (Figure 3A) as previously reported.[3,15,18]

We detected on average a similar number of expressed genes for both tissue types (15.2 K for FF vs. 15.6 K for FFPE), which again is in line with what previously has been reported.[3] However, the FFPE samples displayed a much greater spread with the two samples with the lowest RNA integrity values having the highest number of expressed genes. Based on this finding, we investigated differences in tissue-specific gene expression both in terms of number of expressed genes and levels of gene expression. Genes that were found expressed in both tissue types in at least one patient (17 937 genes) had significantly higher expression (on average 1307 reads) compared to those found exclusively in FF (143 genes with an average of 0.5 read/gene) or FFPE (1097 genes with an average of 2.4 reads/gene) tissue. The latter finding appears to be directly related to input RNA quality as the majority of these genes were found in samples displaying the poorest RNA integrity values and are most likely sequencing artifacts and should be excluded from downstream analysis. As it is difficult to estimate confidently actual gene expression levels in genes with few reads in relative expression estimates, many studies commonly enforce an absolute cutoff of at least 10 mapped reads for global expression analysis. Applying this strategy here would also efficiently remove false-positive data due to degraded RNA. This is evident when comparing the number of detected genes in paired samples before and after implementing the mentioned 10 read cutoff (Figure 3D and Table 4S).

In order to explore further differences in gene expression between the tissue types, we first investigated the frequency at which expressed genes of various lengths are detected. In our dataset, 95% of all expressed genes displayed a length of 750–4250 coding bases. Within this range, we noted only marginal inter- and intra-tissue variability in gene expression while genes on either side of this size interval instead showed significant differences both within and across tissue types. It could be argued that genes with extremely large coding sequences might be more prone to, and that small genes with relatively low expression are more sensitive to, degradation, in particular in FFPE tissue. Interestingly, in both these instances, we noted a higher expression in FFPE samples compared to FF counterparts, which points towards technical artifacts as the main reason for the variation seen in the dataset. However, to explore fully the underlying cause of the observed variation for these genes would be beyond the scope of this study. Regardless and based on our data, it is difficult to estimate accurately true expression for genes that are either very small or large in their coding sequence and these genes should therefore be either removed from subsequent analysis or at the very least be interpreted with caution. Nonetheless and despite these apparent differences between tissue types, we found in unfiltered data a near-perfect correlation in expression levels when comparing housekeeping genes alone or investigating all expressed transcripts indicating that global similarities in expression greatly outweigh the tissue-specific differences.

Ultimately, the best measure of FFPE RNA-seq performance is how well the data generated can identify known biological differences in samples in comparison to data obtained from FF tissue. This is one area in which previous research evaluating RNA-seq in FFPE tissue is lacking.[3,13–20] In the present study, we first applied the classifier developed by Reddy et al.[34] in our RNA-seq DLBCL dataset to subclassify the tumors based on COO (RNA-seq classification) and found the results to be highly concordant with data obtained using the Hans algorithm (Table 1). We next applied previously published gene signatures that have been shown to classify DLBCL samples into biologically and clinically relevant subgroups to investigate biological similarities within paired samples and across patients. In a study by Barrans et al.,[8] the authors developed a DLBCL classifier gene signature based on previous work by the LLMPP to assess routinely processed clinical biopsies stored as FFPE tissue. They performed GEP using the Illumina WG-DASL assay and used their DLBCL classifier to evaluate the samples. However, they found an overall poor correlation between their GEP and the GC/NGC classification by the Hans algorithm. As no FF tissue was analyzed, the authors instead attempted to validate their findings by correlating their results with clinical characteristics and outcome.[5] Using the DLBCL classifier by Barrans et al.,[8] our analysis revealed distinct patterns of gene expression which correlated very well with data obtained for COO using the Reddy and the Hans algorithms. In addition, these results were in agreement and highly concordant with data from the NanoString Lymph2Cx assay to determine COO validating our RNA-seq data (Table 1 and Figure 6A). Most notably, all FFPE samples paired with their FF counterparts underscoring that the biological differences between samples appear to be far greater than artifacts created as a result of RNA degradation (Figure 6A). We also used our dataset to test a refined DLBCL classification system, recently proposed by Dybkaer et al.[35] which is based on subset-specific BAGS for naive, centrocyte, centroblast, memory, and plasmablast B cells. The authors found that BAGS assignment was significantly associated with overall survival and progression-free survival within the GC subclass, where the centrocyte subtype had a superior prognosis compared with the centroblast subtype.[35] Here, samples within four pairs classified differently when comparing FF to FFPE tissue. However, in clustering analysis, three of the four FFPE samples paired correctly with their corresponding FF counterpart, while the final sample was found in the nearest branch. All FFPE cases displayed a highly similar gene expression pattern compared to their FF counterparts and, although

classified differently, did not cluster with other samples of the same BAGS classification. This is an indication that the paired samples are still biologically highly similar, and that the classification mismatch is more likely due to the algorithm which in our study only contains 180 genes from the original list of 223 genes (Figure 6B). Finally, we also investigated a validation cohort of 50 FFPE DLBCL samples where we compared COO classification using the Hans, RNA-seq, and Nanostring algorithms. Overall, the three different methods were highly concordant again confirming the feasibility of using capture-based RNA-sequencing for gene expression analysis.

In conclusion, our data suggest that archived routine FFPE samples can reliably be used for differential expression analysis using the Illumina Truseq RNA Exome protocol and RNA-seq. In addition, we recommend gene expression data to be filtered for both lowly expressed genes as well as those having extremely low or high number of coding bases as it may be difficult to accurately estimate true expression for these genes. Having said that, these recommendations do not solely apply for RNA-seq data generated from FFPE tissue but are also true for FF samples.

ACKNOWLEDGMENTS

The computations were performed on resources provided by Swedish National Infrastructure for Computing through Uppsala Multidisciplinary Center for Advanced Computational Science. Lymph2Cx assay was run by Clinical Genomics Uppsala, Science for Life Laboratory, Department of Immunology, Genetics and Pathology, Uppsala Universitet, Sweden and data were analyzed by Nanostring Technologies. The library preparation and sequencing was performed by the SNP&SEQ Technology Platform, National Genomics Infrastructure (NGI), hosted by Science for Life Laboratory. The SNP&SEQ Platform is supported by Science for Life Laboratory, the Swedish Research Council and the Knut och Alice Wallenbergs Stiftelse. This work was supported by grants from the Swedish Cancer Society, the Swedish Research Council, the Knut and Alice Wallenberg Foundation, Karolinska Institutet, Stockholm, Radiumhemmets Forskningsfonder, Stockholm, Borås Cancer Foundation, Lion's Cancer Research Foundation, Marcus Borgström's Foundation, Uppsala, and Erik, Karin and Gösta Selander's Foundation, Uppsala.

CONFLICT OF INTEREST

POA joined the speakers' bureau of Roche, Gilead and Janssen; has been a consultant for Abbvie, Gilead, Janssen and Roche, and received a research grant from Gilead Nordic. RR received honoraria from Abbvie, AstraZeneca, Illumina and Roche. The other authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

RNA-seq data is available upon request.

REFERENCES

1. Blow N. Tissue preparation: Tissue issues. *Nature*. 2007;448(7156):959-963.
2. Cieslik M, Chugh R, Wu YM, et al. The use of exome capture RNA-seq for highly degraded RNA with application to clinical cancer sequencing. *Genome Res*. 2015;25(9):1372-1381.
3. Schuierer S, Carbone W, Knehr J, et al. A comprehensive assessment of RNA-seq protocols for degraded and low-quantity samples. *BMC Genomics*. 2017;18(1):442.
4. Chapuy B, Stewart C, Dunford AJ, et al. Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes. *Nat Med*. 2018;24(5):679-690.
5. Rosenwald A, Wright G, Chan WC, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N Engl J Med*. 2002;346(25):1937-1947.
6. Wright G, Tan B, Rosenwald A, Hurt EH, Wiestner A, Staudt LM. A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma. *Proc Natl Acad Sci U S A*. 2003;100(17):9991-9996.
7. Cai YD, Huang T, Feng KY, Hu L, Xie L. A unified 35-gene signature for both subtype classification and survival prediction in diffuse large B-cell lymphomas. *PLoS One*. 2010;5(9):e12726.
8. Barrans SL, Crouch S, Care MA, et al. Whole genome expression profiling based on paraffin embedded tissue can be used to classify diffuse large B-cell lymphoma and predict clinical outcome. *Br J Haematol*. 2012;159(4):441-453.
9. Alizadeh AA, Eisen MB, Davis RE, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*. 2000;403(6769):503-511.
10. Hans CP, Weisenburger DD, Greiner TC, et al. Confirmation of the molecular classification of diffuse large B-cell lymphoma by immunohistochemistry using a tissue microarray. *Blood*. 2004;103(1):275-282.
11. Scott DW, Wright GW, Williams PM, et al. Determining cell-of-origin subtypes of diffuse large B-cell lymphoma using gene expression in formalin-fixed paraffin-embedded tissue. *Blood*. 2014;123(8):1214-1217.
12. Abdulla M, Hollander P, Pandzic T, et al. Cell-of-origin determined by both gene expression profiling and immunohistochemistry is the strongest predictor of survival in patients with diffuse large B-cell lymphoma. *Am J Hematol*. 2020;95(1):57-67.
13. Waddell N, Cocciardi S, Johnson J, et al. Gene expression profiling of formalin-fixed, paraffin-embedded familial breast tumours using the whole genome-DASL assay. *J Pathol*. 2010;221(4):452-461.
14. Sinicropi D, Qu K, Collin F, et al. Whole transcriptome RNA-Seq analysis of breast cancer recurrence risk using formalin-fixed paraffin-embedded tumor tissue. *PLoS One*. 2012;7(7):e40092.
15. Adiconis X, Borges-Rivera D, Satija R, et al. Comparative analysis of RNA sequencing methods for degraded or low-input samples. *Nat Methods*. 2013;10(7):623-629.
16. Norton N, Sun Z, Asmann YW, et al. Gene expression, single nucleotide variant and fusion transcript discovery in archival material from breast tumors. *PLoS One*. 2013;8(11):e81925.
17. Hedegaard J, Thorsen K, Lund MK, 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*. 2014;9(5):e98187.
18. Graw S, Meier R, Minn K, et al. Robust gene expression and mutation analyses of RNA-sequencing of formalin-fixed diagnostic tumor samples. *Sci Rep*. 2015;5:12335.
19. Eikrem O, Beisland C, Hjelle K, et al. Transcriptome sequencing (RNAseq) enables utilization of formalin-fixed, paraffin-embedded biopsies with clear cell renal cell carcinoma for exploration of disease biology and biomarker development. *PLoS One*. 2016;11(2):e0149743.
20. Esteve-Codina A, Arpi O, Martinez-Garcia M, et al. A comparison of RNA-seq results from paired formalin-fixed paraffin-embedded and fresh-frozen glioblastoma tissue samples. *PLoS One*. 2017;12(1):e0170632.

21. Swerdlow SH, Campo E, Harris NL, et al. *WHO Classification of Tumors of Haematopoietic and Lymphoid Tissues*. Lyon; 2008.
22. Ewels PA, Peltzer A, Fillinger S, et al. The nf-core framework for community-curated bioinformatics pipelines. *Nat Biotechnol*. 2020; 38(3):276-278.
23. Krueger F. *Trim Galore*. https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
24. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15-21.
25. Broad Institute. *Picard Tools*. http://broadinstitute.github.io/picard/
26. Sayols S, Scherzinger D, Klein H. dupRadar: a Bioconductor package for the assessment of PCR artifacts in RNA-Seq data. *BMC Bioinformatics*. 2016;17(1):428-428.
27. Andrews S. *FastQC: A Quality Control Tool for High Throughput Sequence Data*. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
28. Wang L, Wang S, Li W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics*. 2012;28(16):2184-2185.
29. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*. 2016;32(19):3047-3048.
30. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*. 2014;30(7):923-930.
31. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014; 15(12):550-550.
32. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol*. 2010;11(10):R106-R106.
33. Zerbino DR, Achuthan P, Akanni W, et al. Ensembl. *Nucleic Acids Res*. 2018;46(D1):D754-D761.
34. Reddy A, Zhang J, Davis NS, et al. Genetic and functional drivers of diffuse large B cell lymphoma. *Cell*. 2017;171(2):481-494. e415.
35. Dybkaer K, Bogsted M, Falgreen S, et al. Diffuse large B-cell lymphoma classification system that associates normal B-cell subset phenotypes with prognosis. *J Clin Oncol*. 2015;33(12):1379-1388.
36. Wallden B, Ferree S, Ravi H, et al. Development of the molecular diagnostic (MDx) DLBCL lymphoma subtyping test (LST) on the nCounter analysis system. *J Clin Oncol*. 2015;33(15):8536.
37. Scarlet D, Ertl R, Aurich C, Steinborn R. The orthology clause in the next generation sequencing era: novel reference genes identified by RNA-seq in humans improve normalization of neonatal equine ovary RT-qPCR data. *PLoS One*. 2015;10(11):e0142122.
38. Groelz D, Viertler C, Pabst D, Dettmann N, Zatloukal K. Impact of storage conditions on the quality of nucleic acids in paraffin embedded tissues. *PLoS One*. 2018;13(9):e0203608.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Skaftason A, Qu Y, Abdulla M, et al. Transcriptome sequencing of archived lymphoma specimens is feasible and clinically relevant using exome capture technology. *Genes Chromosomes Cancer*. 2022;61(1):27-36. doi:10.1002/gcc.23002