*NAR Genomics and Bioinformatics*, 2024, **6**, 1–11
https://doi.org/10.1093/nargab/lqae008
Advance access publication date: 31 January 2024
**Standard Article**

# Optimization of FFPE preparation and identification of gene attributes associated with RNA degradation

**Yu Lin[1,2,†], Zhou-Huan Dong[3,†], Ting-Yue Ye[4,†], Jing-Min Yang[1], Mei Xie[5], Jian-Cheng Luo[6], Jie Gao[3,*] and An-Yuan Guo[1,2,*]**

[1]Hubei Bioinformatics and Molecular Imaging Key Laboratory, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China
[2]Department of thoracic surgery, West China Biomedical Big Data Center, West China Hospital, Med-X Center for Informatics, Sichuan University, Chengdu 610041, China
[3]The First Medical Center, Chinese People's Liberation Army (PLA) General Hospital, Beijing 100853, China
[4]Nanjing Vazyme Biotech Co., Ltd., Nanjing 210000, China
[5]Department of Respiratory and Critical Care, Chinese People's Liberation Army (PLA) General Hospital, Beijing 100853, China
[6]Aiyi Technology Co., Ltd., Beijing 102609, China

[*]To whom correspondence should be addressed. Tel: +86 02885422819; Fax: +86 02885422819; Email: guoanyuan@wchscu.cn
Correspondence may also be addressed to Jie Gao. Tel: +86 01066937529; Fax: +86 01066937529; Email: 13683691828@139.com
[†]The first three authors should be regard as Joint First Authors.

## Abstract

Formalin-fixed paraffin-embedded (FFPE) tissues are widely available specimens for clinical studies. However, RNA degradation in FFPE tissues often restricts their utility. In this study, we determined optimal FFPE preparation conditions, including tissue ischemia at 4°C (<48 h) or 25°C for a short time (0.5 h), 48-h fixation at 25°C and sampling from FFPE scrolls instead of sections. Notably, we observed an increase in intronic reads and a significant change in gene rank based on expression level in the FFPE as opposed to fresh-frozen (FF) samples. Additionally, we found that more reads were mapped to genes associated with chemical stimulus in FFPE samples. Furthermore, we demonstrated that more degraded genes in FFPE samples were enriched in genes with short transcripts and high free energy. Besides, we found 40 housekeeping genes exhibited stable expression in FF and FFPE samples across various tissues. Moreover, our study showed that FFPE samples yielded comparable results to FF samples in dimensionality reduction and pathway analyses between case and control samples. Our study established the optimal conditions for FFPE preparation and identified gene attributes associated with degradation, which would provide useful clues for the utility of FFPE tissues in clinical practice and research.

## Introduction

Formalin-fixed paraffin-embedded (FFPE) and fresh-frozen (FF) tissues are the common human tissue specimens in clinical practice and medical research (1,2). FF tissues are often preferred, but their availability is often limited due to their laborious collection and expensive preservation. By contrast, as FFPE tissues can be economically stored for long periods and linked to patient clinical data, they are widely used for pathological analysis and molecular testing. Effectively using RNA-seq data from FFPE tissues, along with pertinent patient clinical information, enables the acquisition of large experimental and control cohorts. Additionally, it also facilitates the study of archival specimens (3), making FFPE samples an invaluable resource in clinical research.

In previous studies, DNA from archival FFPE tissue samples has been extensively utilized for next-generation sequencing (NGS) methods such as whole exome sequencing (4,5). Nevertheless, extracting high-quality RNA from FFPE tissue samples is challenging due to the extensive crosslinking and degradation caused by formalin fixation (6). In recent years, advances in FFPE RNA extraction technology could efficiently reverse the crosslinking of FFPE RNA (7). Due to rRNA degradation in FFPE tissues, the RNA integrity number (RIN) is deemed inappropriate for assessing FFPE RNA integrity. The DV200 (percentage of RNA fragments >200 nucleotides in size) was devised to accurately assess the quality of RNA and incorporated into the Illumina protocol (8). For degraded RNA, DV200 is a reliable predictor of the probability of successful library construction (8,9). Notably, RNA integrity in FFPE specimens is influenced by many preanalytical factors, including ischemia time, fixation time, temperature, storage conditions and sampling methods. These preanalytical factors commonly exhibit variations in clinical laboratories, but the analyses of these factors on RNA integrity are still incomplete (10). The RNA quality of FFPE samples also affects the concordance of expression profiles between FFPE and FF specimens (11). The Biospecimen Preanalytical Variables (BPV) program shows that cold ischemia time of up to 12 h has little impact on DV200, and prolonged fixation time (72 h) contributes to RNA fragmentation (12).

While the integrity of RNA in FFPE specimens declines with longer preservation time (13), RNA-seq has been successfully performed on some FFPE specimens after long-term storage (4,13,14). Moreover, many studies have explore the

**Received:** October 31, 2023. **Revised:** December 30, 2023. **Editorial Decision:** January 12, 2024. **Accepted:** January 16, 2024

feasibility of utilizing FFPE tissues in RNA-seq analyses (1,13). When FFPE tissues were adequately prepared and preserved, their expression data can be strongly correlated with those from frozen tissues (5,15–19). The expression data of FFPE specimens have been applied in cancer research (15). The differentially expressed genes (DEGs) between different cancers obtained using FFPE tissues were significantly overlapped with those obtained from FF tissues (12). However, not all FFPE specimens yielded usable RNA-seq data (18,19). Recent studies have introduced methods to normalize FFPE RNA-seq expression data (20,21) and enhanced the availability of FFPE RNA-seq data. Although some studies have analyzed the DEGs between paired FFPE and FF tissues (5,18), the pattern of the changes in FFPE expression data remains unclear.

In this study, we assessed the RNA integrity of FFPE tissues prepared under various conditions and identified the optimal preparation conditions and sampling method. To understand the pattern of changes in FFPE expression data, we explored differences in RNA-seq expression profiling and transcript attributes. We also conducted comparative analysis between tumor and peritumor using FFPE and FF samples, respectively. Our results showed that FFPE samples were available for dimensionality reduction and differential expression analysis in cancer research. Our findings may help to guide the clinical preparation of FFPE tissues, understand the pattern of expression changes in FFPE samples and better harness the power of FFPE tissue resources.

## Materials and methods

### Ethics statement

The research was conducted on lung tissues collected under Institutional Review Board (IRB) approved protocols at Beijing Shijitan Hospital. Informed consents to participate in the study were also obtained according to IRB. We obtained the tissues under IRB approvals (IRB approval number: [sjtkyll-1x-2022(35)]). All experiments were performed in compliance with relevant laws and institutional guidelines. The data were analyzed anonymously to protect the privacy of the patients.

### Specimens preparation and RNA isolation

We collected lung tissues from six patients to prepare the FF and FFPE samples with different preparation conditions. After collecting the tissues, we cut the tissues into 0.3 cm × 0.3 cm × 0.3 cm pieces and preserved the fresh tissues at $-80^\circ$C. As for the FFPE tissues, we used a 4% neutral paraformaldehyde solution to fix the tissues under different fixation conditions. Next, we repeatedly immersed the tissue in ethanol of increasing concentration levels, ending in a 100% ethanol concentration, to dehydrate the tissues at room temperature. Then, we used xylene as a clearing agent to remove all the ethanol in the tissues at room temperature. Finally, tissues were infiltrated with paraffin wax at $63^\circ$C and then left to cool so that they solidified (see Supplementary Table S1).

For FFPE RNA isolation, paraffin scrolls or sections were employed for RNA extraction, quality assessment and sequencing. Prior to sampling, we cut off the outermost layer of paraffin which was exposed to air. Next, paraffin scrolls were obtained by cutting 5 μm thick specimens from FFPE samples. Then, the paraffin scrolls were soaked in water, adhered to the glass slides and dried to get paraffin sections. When investigating the impact of sampling method, we prepared the paraffin sections first and then the paraffin scrolls from the same FFPE blocks for RNA extraction. Total RNA was extracted from the FFPE samples using RNAstorm Kit (celldata, CD501), following the manufacturer's instructions. We used Nanodrop to check the purity of total RNA extracted from FFPE and FF samples. The Qubit® 3.0 Fluorometer was used to assess the RNA concentration. The RNA integrity (DV200 and DV800) was assessed using the Agilent 2100 Bioanalyzer.

### Data composition

This study utilized the RNA-seq data obtained from lung tissues and the public data from the BPV research program. The metadata for the lung tissues was reported in Supplementary Table S2. The data of lung tissues consisted of 15 FF samples and their paired FFPE samples, including three pairs of control-case-matched FF samples and FFPE samples (Supplementary Table S2).

The public data were acquired from the BPV research program developed by National Cancer Institute's (NCI) Biorepositories and Biospecimen Research Branch (BBRB). They consisted of FF samples and FFPE samples from five renal clear cell carcinomas (kidney), seven serous ovarian carcinomas (ovary) and five colon adenocarcinomas (colon). The metadata for public data, including specimen ids, tissue types, cold ischemia time and the percentage of RNA fragments >200 nucleotides (DV200), are reported in Supplementary Table S3. The public data of the BPV research program are accessible through dbGaP (#phs001304). The quality metrics of the individual specimen are available in a previous publication (22).

### Total RNA-sequencing library preparation and sequencing

We prepared the sequencing libraries following the manufacturer's recommendations of Ribo-off rRNA Depletion Kit (Human/Mouse/Rat) (Vazyme Biotech Co., Ltd., Nanjing, China, N406) and VAHTS Universal V6 RNA-seq Library Prep Kit for Illumina (Vazyme Biotech Co., Ltd., Nanjing, China, NR605). The details of library construction are as follows. First, we removed ribosome RNA from 200 ng total RNA using Ribo-off rRNA Depletion Kit (Human/Mouse/Rat) (Vazyme, N406). Then, we fragmented the RNA into small pieces using divalent cations at elevated temperatures. The cleaved RNA fragments were copied into first-strand cDNA using reverse transcriptase and random primers, followed by second-strand cDNA synthesis using DNA Polymerase I, RNase H, deoxyuridine triphosphate (dUTP), deoxyadenosine triphosphate (dATP), deoxyguanosine triphosphate (dGTP) and deoxycytidine triphosphate (dCTP). Then, a single 'A' base and the adapters were subsequently added to these cDNA fragments. In order to select the appropriate cDNA fragment size for sequencing, we selected the library fragments with VAHTSTM DNA Clean Beads (Vazyme, N411). The polymerase chain reaction (PCR) amplification was performed, and the aimed products were finally purified. After cluster generation, the libraries were sequenced on an Illumina novaseq 6000 platform, and the raw fastq files of 150-bp paired-end reads were generated.

### Expression-level quantification

After sequencing, we removed reads with adapters and reads in which unknown bases were >5%. We defined the low-quality base as the base whose sequencing quality was not >10. Next, we removed the reads with over 50% low-quality bases. At the same time, Q20, Q30 and GC content were calculated for clean reads. All downstream analyses were based on the clean reads. Reads were quality-trimmed using fastp (23) with default parameters, which trimmed low-quality bases from the ends of reads, low-quality reads and residual Illumina adapters. RNA-seq reads were aligned to the Homo sapiens reference genome (GRCh38) from the Ensembl database using hisat2-2.1.0 (24) with default parameters. After alignment, the resulting bam file was fed into the RNA-Seq quantification software featureCounts 2.0.3 (25) with paired-end mode to generate counts matrices on the gene level (GTF version: Homo_sapiens.GRCh38.104.chr.gtf). StringTie 2.2.1 (26) was used to generate Transcripts per kilobase of exon model per million mapped reads (TPM) matrices. For this study, we only quantified at the gene level and focused on protein-coding genes. The relevant quality control metrics for RNA-seq were displayed in Supplementary Table S4.

### Post-quantification analysis

We further characterized reads aligned to different regions of genes. The read_distribution.py and geneBody_coverage.py functions from RSeQC 5.0.1 (27) were used to calculate reads distribution over genome features and the RNA-seq reads coverage over the gene body. The bam files output from Hisat2 were fed into rMATS-4.1.0 (28) to analyze alternative splicing with default parameters.

The classic transcript (the earliest released version of transcripts) of each gene was used when analyzing the correlation between the gene attributes and RNA degradation. The genes' transcript length and cDNA sequences were obtained from the Ensembl database using the R package BiomaRt (29). We transformed the cDNA sequences to mRNA sequences and calculated the frequency of nucleobase using the R package Biostrings. The minimum free energy (MFE) of the transcript was calculated using RNAfold in ViennaRNA (30) package with default parameters. The mRNA subcellular localization was predicted by mRNAloc (31) with a support vector machine (SVM) classification threshold of 0.1.

Principal Component Analysis (PCA) was performed using TPM data and the R package FactoMineR (32) with default parameters. The housekeeping gene list of human genes stably expressed across 52 tissues and cell types was obtained from HRT atlas V1.0 (33). The human pan-cancer gene list and the annotation of genes were obtained from the nanoString website: https://nanostring.com/products/ncounter-assays-panels/oncology/ncounter-pancancer-pathways-panel (last accessed date: 13 March 2023).

### Statistical analysis

The concordance correlation coefficient (CCC) function in the R package DescTools was used to evaluate the CCC (34) between paired FFPE and FF samples. To perform paired differential expression analysis, we used the R package DESeq2 (35) and added sample pairwise information as a covariate to the negative binomial regression model. Then, we identified the DEGs with thresholds of adjusted $P$-value $< 0.05$ and absolute ($\log_2$FoldChange) $> 1$. Besides, we identified the genes with small changes between FF and FFPE samples with the criteria of absolute ($\log_2$FoldChange) $< 0.2$, and lfcSE (the standard error of $\log_2$FoldChange) $< 0.1$. R package clusterProfiler (36) was used to perform gene set enrichment analysis (GSEA) with a $q$-value threshold of 0.05. Fisher's exact test was performed with an unadjusted $P$-value threshold of 0.005 and a $|\log_2$ odds ratio$|$ threshold of 1 to analyze the association between the DEGs and transcript attribute. We constructed a generalized linear model using the glm.nb function in the R package MASS and incorporated tissue type as a covariate to investigate the relationship between DV200 and gene expression. An adjusted $P$-value threshold of 0.05 was set to distinguish whether a gene is significantly correlated to DV200.

## Results

### The optimal preparation conditions of FFPE tissues for high-quality RNAs

In order to determine the optimal preparation conditions of FFPE tissues, we compared the RNA quality of FFPE tissues with different fixation times and temperatures, ischemic times, temperatures and sampling methods. Though the study of BPV program found that excessive fixation time up to 72 h resulted in decreased RNA quality (22), we found that adequate fixation duration (48 h) improved the quality of RNA in FFPE samples in our study. Among the fixation times of 12, 24 and 48 h, we observed that FFPE tissues fixed for 48 h exhibited the highest DV200 and DV800 (the percentage of RNA fragments > 800 nucleotides) (Figure 1A and Supplementary Figure S1A).

For the fixation temperature, the DV200 and DV800 of FFPE samples fixed at 4°C were close to FFPE samples fixed at 25°C (Figure 1B; Supplementary Figure S1B and Supplementary Table S5). As 25°C was closer to the FFPE production temperatures in hospitals and produced similar RNA quality as 4°C, we chose the fixation time of 48 h and fixation temperature of 25°C as the optimal fixation conditions. In the range of ischemia time from 0.5, 3, 6 to 12 h at 25°C, no significant decline in FFPE RNA quality was observed, although slightly higher DV200 values were noted for 0.5 h (Figure 1C and Supplementary Figure S1C). Concerning ischemic temperature, when the ischemia time was >0.5 h, the quality of RNA extracted from the sample at 4°C was higher than that at 25°C (Supplementary Table S5).

From the perspective of expression consistency with FF samples, the expression profiles of the samples with ischemic treatment at 4°C were in high agreement with FF samples, and the CCC ranged from 0.79 to 0.9, which was close to FFPE samples without ischemic treatment (Figure 1D and Supplementary Table S6). FFPE samples with 6 and 48 h of ischemia at 25°C were significantly lower in agreement with FF samples (Figure 1D and Supplementary Table S6). For the sampling methods, the quality of RNA extracted from FFPE sections was generally lower than RNA extracted from the FFPE scrolls (see Materials and Methods section). The DV200 of FFPE sections ranged from 30 to 40, while the DV200 of FFPE scrolls was >60 (Supplementary Figure S2). Furthermore, we used the RNA quality data from BPV program (22) and found no significant difference in the DV200 value (RNA quality) among colon, kidney and ovary tissues in the majority of experimental groups (Supplementary Figure S3). This result suggested that out results from the lung tissue can be

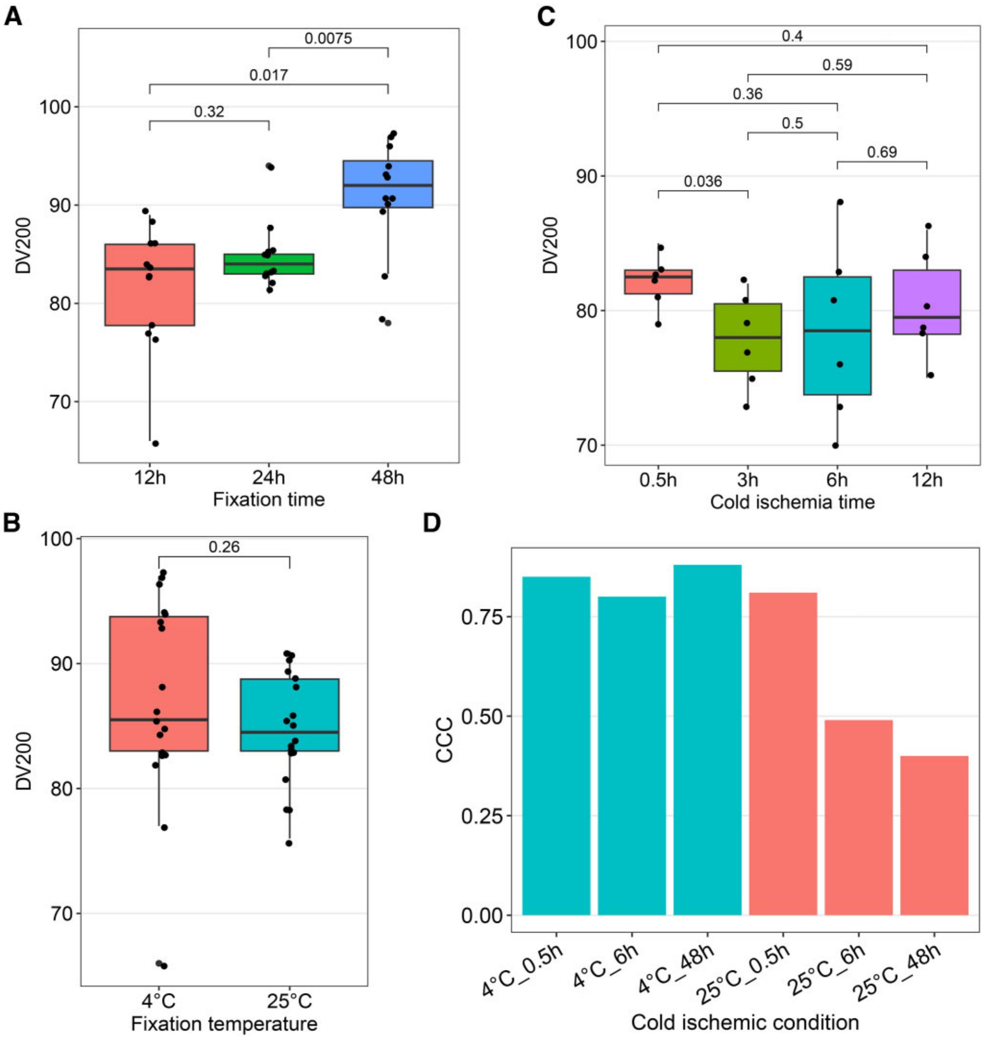


**Figure 1.** Comparison of the RNA quality of FFPE tissues under different preparation conditions. (**A–C**) Boxplot showing DV200 of FFPE samples with different fixation time (A), fixation temperature (B) and ischemia time (C). Six FFPE samples were investigated for each parameter combination (Supplementary Table S5). (**D**) Histogram showing the CCC between paired FFPE and FF samples at different ischemia temperature and time.

generalized to other tissues. In summary, the optimal preparation conditions of FFPE are ischemia of tissues at 4°C (<48 h) or ischemia at 25°C for short time (0.5 h), fixation for 48 h at 25°C and sampling from FFPE scrolls rather than FFPE sections.

## Read distribution and expression profiles differ between FFPE and FF samples

To assess potential systematic differences between FF and FFPE samples, we performed RNA-seq on seven FF samples and their paired FFPE samples from the peritumoral tissues of lung (Supplementary Table S2, sample id: L1-7). We compared the distribution of RNA-seq reads and expression profiles between FF and FFPE samples. First, we found that FFPE and FF samples did not show serious 3′ or 5′ bias on the gene body distribution (Figure 2A). However, the proportion of read distributed in introns was significantly higher in FFPE samples than FF samples, consistent with previous studies (12,17,19). The proportion of intron was generally ∼25% in FF samples, while it was >50% in FFPE samples (Figure 2B and C). Changes in transcript reads distribution and increased intronic reads affected the detection of alternative splicing. Approximately 20–25% of the intron retention events were overrepresented in FFPE samples (Supplementary Figure S4).

In terms of the expression profiles, PCA results showed that the expression profiles of FFPE samples differed from those of FF samples (Figure 2D). When comparing the ranks of genes with TPM $> 10$ ($n = 11745$) ordered by expression level in FFPE samples and their paired FF samples, substantial rank changes were observed in most genes. Only 5–10% of genes exhibited minor change ($|\Delta\text{rank}|<100$). Notably, the expression rank of 4.4% of genes greatly changed ($|\Delta\text{rank}|>1000$) in the same tissue of different sequencing batches. In contrast, over 40% of genes significantly changed in FFPE samples relative to paired FF samples (Figure 2E). Our findings showed systematic differences in read distribution and expression profiles between FFPE and FF samples.

## Genes related to chemical stimulus over-sampled in FFPE samples

To investigate the changes in gene expression between FFPE and FF samples, we performed DEG analysis on our data from lung tissue and the public data from BPV project for three tissues (kidney, ovary and colon). We identified 1001 genes

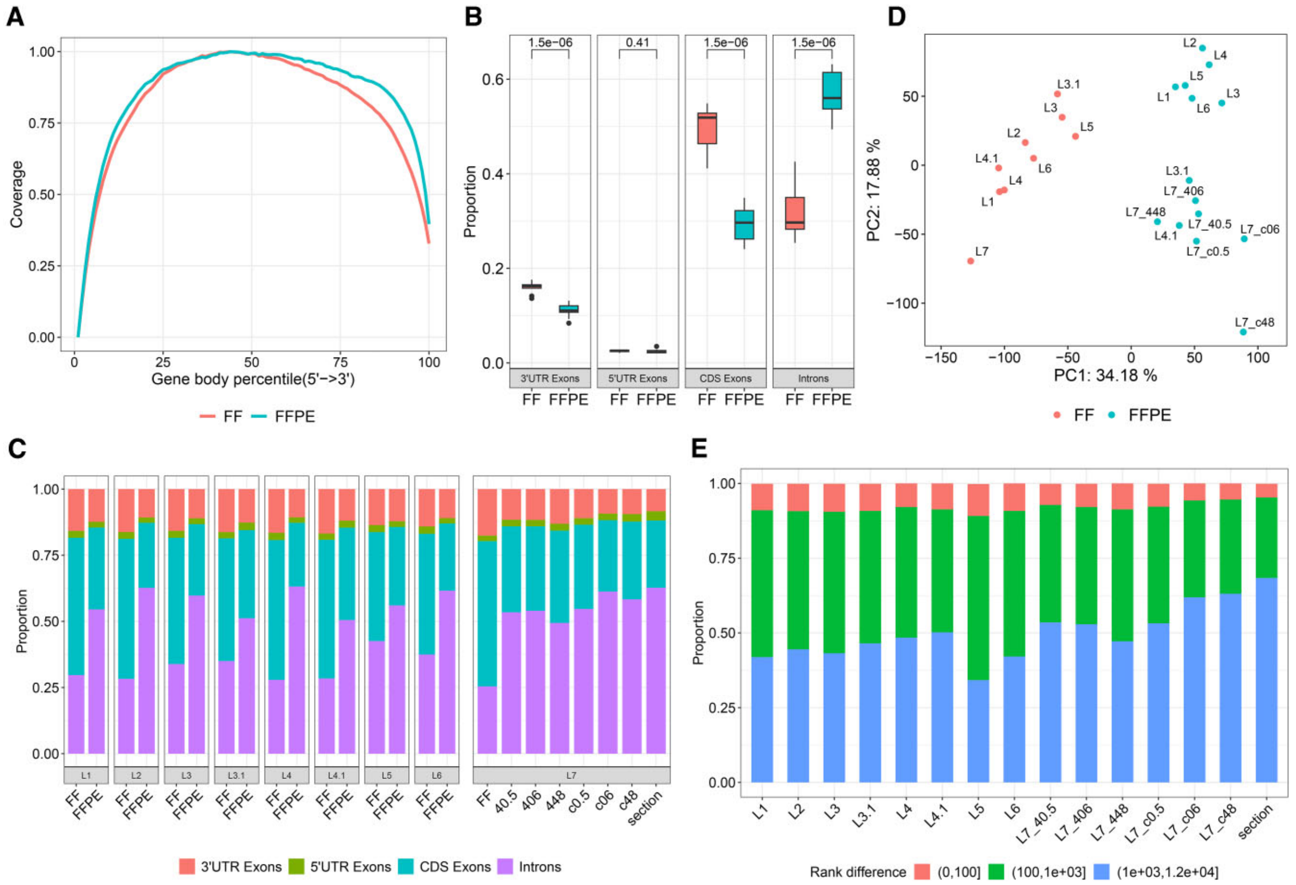


**Figure 2.** Differences in read distribution and expression profiles between FFPE samples and FF samples. (**A**) Average reads coverage distribution from transcript 5′ to 3′, the coverage of 3′ area of the transcript was higher in FFPE samples; (**B, C**) Proportion of read distributed in 3′UTR (untranslated regions), 5′UTR, CDS (coding sequences), introns in paired FFPE and FF samples. (**D**) Dot graph showing PCA result of expression profiles of FFPE and FF samples. (**E**) Difference of gene expression rank between paired FFPE and FF samples.

were consistently over-sampled in at least two kinds of tissues (Figure 3A). There are 420 genes that were consistently more degraded in at least two kinds of tissues (Figure 3B). Next, we performed GSEA. GSEA results showed that the genes with higher expression level in FFPE were significantly enriched in the gene sets of chemical stimulus perception (Figure 3C). We further checked genes in the identified GO terms in our lung samples. After filtering low-expression genes (baseMean > 1), 12 out of 21 (57%) taste receptor genes and 28 out of 91 (31%) olfactory receptor genes were significantly over-sampled in FFPE in lung tissue. In addition, most of the other genes in these two terms had an over-sampled trend in FFPE samples (Figure 3D). We further analyzed the stimulus perception-related genes in the BPV project data and found similar results (Supplementary Figure S5A–C). This finding suggested that reads of genes related to perception of chemical stimulus were overrepresented in FFPE.

## Transcripts with altered expression in FFPE samples significantly correlated with their properties including length, secondary structure and subcellular localization

After analyzing the potential biological processes involved during fixation, we delved into the gene properties of DEGs in FFPE versus FF samples. The classic transcripts ($n = 19\,307$) of protein-coding genes were selected to obtain the transcript length, MFE of secondary structure, nucleobase content and subcellular localization information. Our analysis revealed that the more degraded genes in FFPE samples were overrepresented in genes with short transcript length and high free energy, regardless of tissue types ($P$-value $< 0.05$, $|\log_2\text{odds\_ratio}|>1$) (Figure 4A–D). In other words, these more degraded genes in FFPE samples had significantly shorter transcript lengths and higher MFE than other genes (Figure 4E,F). Regarding nucleobase content, we found that genes over-sampled in FFPE samples were overrepresented in high GC content genes in lung tissues but not in other tissues (Supplementary Figure S6). For the RNA subcellular localization, we found that genes more degraded in FFPE samples were overrepresented in the extracellular regions across all four tissue types (Figure 4A–D). Additionally, almost all mitochondrial genes were consistently overexpressed in FFPE samples across tissue types. To investigate the relationship between gene expression and DV200 values, we constructed a generalized linear model using the FFPE samples in the BPV project. The results showed that only 41 genes were significantly correlated with the DV200 value (padj $< 0.05$), and most genes were not correlated with DV200 (Supplementary Figure S7). In summary, our results suggest that the expression changes of genes/transcripts in FFPE versus FF samples are correlated to transcript length, MFE of secondary structure and subcellular localization.

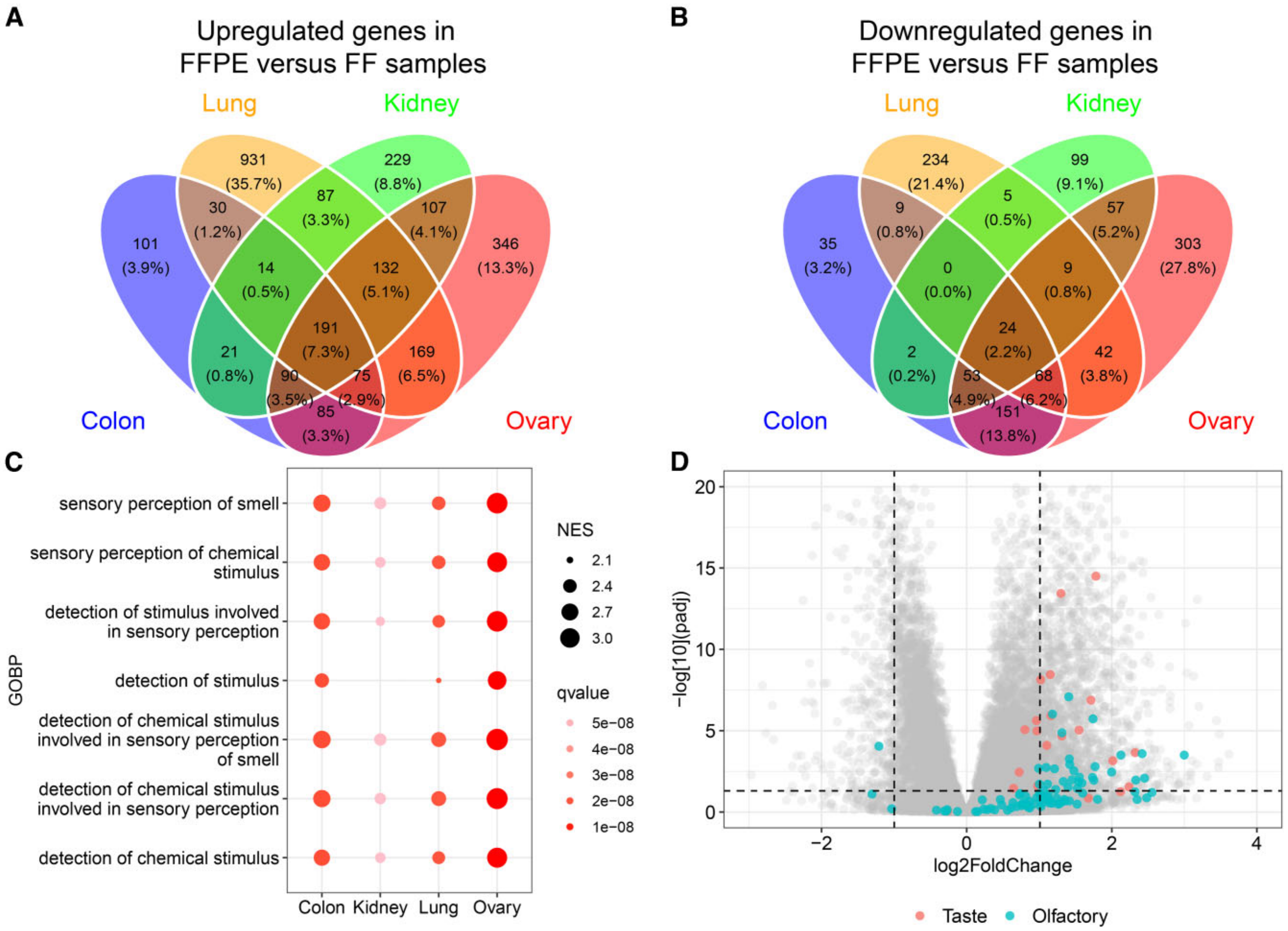


**Figure 3.** DEGs and GO biological processes enriched terms in four kinds of tissues. (**A** and **B**) Venn diagram showing the number of overlapped genes with more (A) or less (B) reads in FFPE across four kinds of tissues. (**C**) Dot graph showing top 5 significantly overrepresented GO terms overlapped in four kinds of tissues. (**D**) Volcano plot showing the DEGs between FFPE and FF samples.

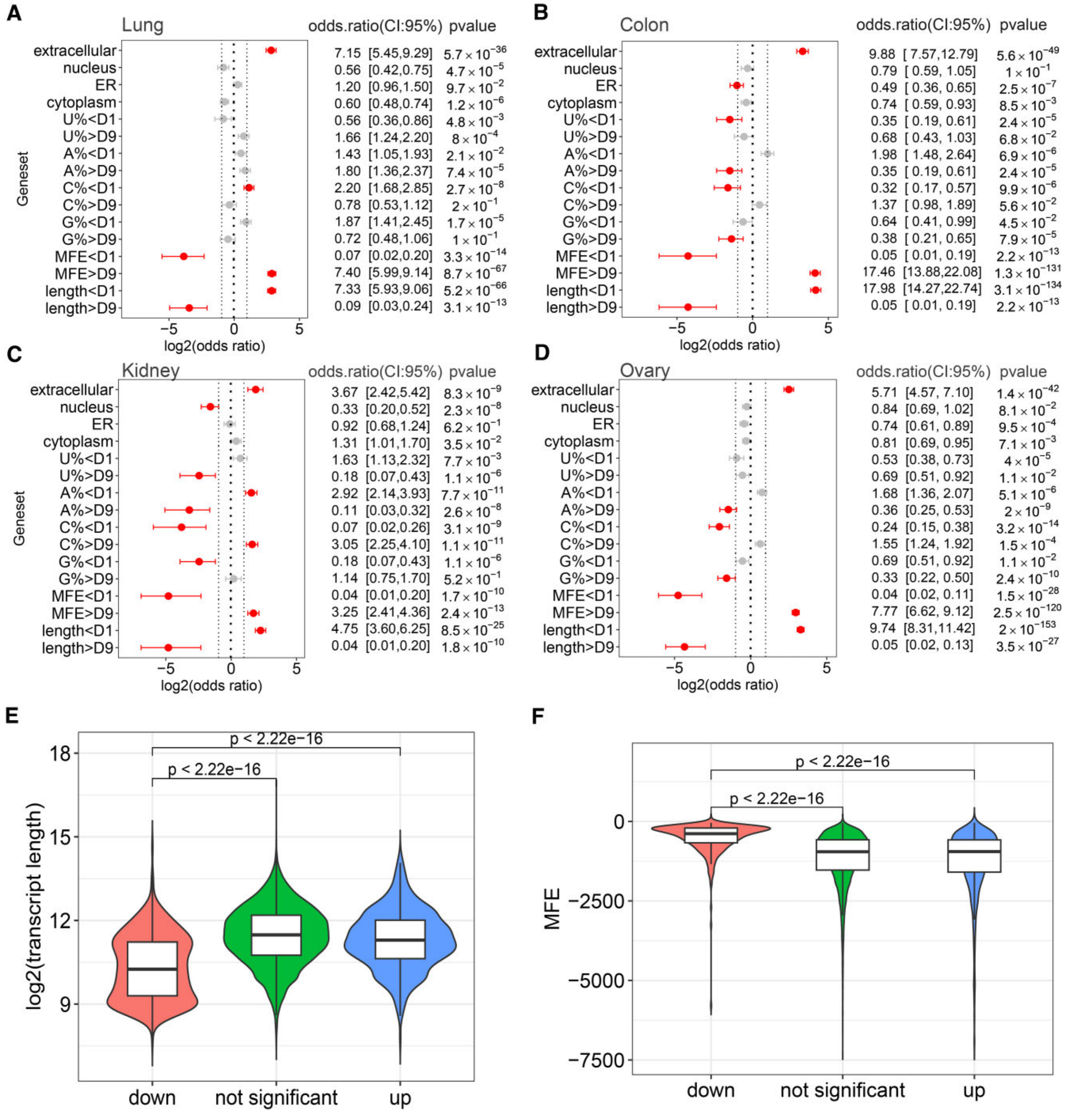
## Housekeeping gene expression alters in FFPE samples

To assess the potential in the expression profile of housekeeping genes (HK genes) in FFPE, we obtained the HK gene list from HRT Atlas v1.0 (33). We found that most HK genes were expressed in both FF and FFPE samples. Among the 2176 HK genes, 85–89% displayed expression levels exceeding 10 TPM in FFPE and FF samples (Figure 5A). Subsequently, we checked the DEG analysis results of HK genes on our lung samples and public data from the BPV project. Among the 2176 HK genes, 4–14% were differentially expressed in FFPE versus FF samples (Figure 5B,C and Supplementary Figure S8). Glyceraldehyde-3-Phosphate Dehydrogenase (GAPDH), the common internal reference gene for qPCR, was more degraded in FFPE samples (Supplementary Figure S8). Applying a criterion for constantly expressed HK genes (see Materials and Methods), we identified a set of 40 HK genes with constant expression across lung, colon, kidney and ovary tissues, which may be suitable as internal reference genes to normalize gene expression data in FFPE (Figure 5D and Supplementary Table S7). Among the 40 stable HK genes, SSU72, SNX1, RNF114, NAP1L4 and VPS26C were the top 5 stably expressed genes. Our results highlight the importance of carefully selecting stable HK genes for internal reference genes, as many HK genes showed significant expression differences between FFPE and FF samples.

## FFPE RNA-seq data can be used for pan-cancer pathway analyses

To assess the utility of FFPE RNA-seq data, we performed transcriptome analysis using FFPE samples and FF samples separately and assessed the consistency of the results. In a previous study (12), a similar analysis has been conducted. The results for FFPE samples were highly consistent with FF samples, with a proportion of overlapped DEGs exceeding 70%. However, the huge variation between different organs (colon, kidney and ovary) may lead to a large number of DEGs, and thousands of DEGs may lead to a high proportion of overlapped DEGs. Thus, we used additional three pairs of samples from tumors and peritumor tissues (Supplementary Table S2) to validate the results.

First, we performed PCA on FF samples and FFPE samples, respectively. The FF samples and FFPE samples had similar distributions on the 2D plane after dimensionality reduction (Figure 6A–C). We further analyze the top 5 principal components in PCA, 4 out of the top 5 principal components correlated between FF and FFPE samples, with a Pearson's correlation coefficient $>0.4$ (Figure 6D). The first principal component obtained from FFPE samples was highly correlated with that in FF samples (Pearson, $r > 0.8$) (Figure 6D), suggesting that the main variation in the data was similar in the FF and FFPE samples.



Lung

| Geneset | odds.ratio(CI:95%) | pvalue |
|---|---|---|
| extracellular | 7.15 [5.45,9.29] | $5.7 \times 10^{-36}$ |
| nucleus | 0.56 [0.42,0.75] | $4.7 \times 10^{-5}$ |
| ER | 1.20 [0.96,1.50] | $9.7 \times 10^{-2}$ |
| cytoplasm | 0.60 [0.48,0.74] | $1.2 \times 10^{-6}$ |
| U%<D1 | 0.56 [0.36,0.86] | $4.8 \times 10^{-3}$ |
| U%>D9 | 1.66 [1.24,2.20] | $8 \times 10^{-4}$ |
| A%<D1 | 1.43 [1.05,1.93] | $2.1 \times 10^{-2}$ |
| A%>D9 | 1.80 [1.36,2.37] | $7.4 \times 10^{-5}$ |
| C%<D1 | 2.20 [1.68,2.85] | $2.7 \times 10^{-8}$ |
| C%>D9 | 0.78 [0.53,1.12] | $2 \times 10^{-1}$ |
| G%<D1 | 1.87 [1.41,2.45] | $1.7 \times 10^{-5}$ |
| G%>D9 | 0.72 [0.48,1.06] | $1 \times 10^{-1}$ |
| MFE<D1 | 0.07 [0.02,0.20] | $3.3 \times 10^{-14}$ |
| MFE>D9 | 7.40 [5.99,9.14] | $8.7 \times 10^{-67}$ |
| length<D1 | 7.33 [5.93,9.06] | $5.2 \times 10^{-66}$ |
| length>D9 | 0.09 [0.03,0.24] | $3.1 \times 10^{-13}$ |

Colon

| Geneset | odds.ratio(CI:95%) | pvalue |
|---|---|---|
| extracellular | 9.88 [ 7.57,12.79] | $5.6 \times 10^{-49}$ |
| nucleus | 0.79 [ 0.59, 1.05] | $1 \times 10^{-1}$ |
| ER | 0.49 [ 0.36, 0.65] | $2.5 \times 10^{-7}$ |
| cytoplasm | 0.74 [ 0.59, 0.93] | $8.5 \times 10^{-3}$ |
| U%<D1 | 0.35 [ 0.19, 0.61] | $2.4 \times 10^{-5}$ |
| U%>D9 | 0.68 [ 0.43, 1.03] | $6.8 \times 10^{-2}$ |
| A%<D1 | 1.98 [ 1.48, 2.64] | $6.9 \times 10^{-6}$ |
| A%>D9 | 0.35 [ 0.19, 0.61] | $2.4 \times 10^{-5}$ |
| C%<D1 | 0.32 [ 0.17, 0.57] | $9.9 \times 10^{-6}$ |
| C%>D9 | 1.37 [ 0.98, 1.89] | $5.6 \times 10^{-2}$ |
| G%<D1 | 0.64 [ 0.41, 0.99] | $4.5 \times 10^{-2}$ |
| G%>D9 | 0.38 [ 0.21, 0.65] | $7.9 \times 10^{-5}$ |
| MFE<D1 | 0.05 [ 0.01, 0.19] | $2.2 \times 10^{-13}$ |
| MFE>D9 | 17.46 [13.88,22.08] | $1.3 \times 10^{-131}$ |
| length<D1 | 17.98 [14.27,22.74] | $3.1 \times 10^{-134}$ |
| length>D9 | 0.05 [ 0.01, 0.19] | $2.2 \times 10^{-13}$ |

Kidney

| Geneset | odds.ratio(CI:95%) | pvalue |
|---|---|---|
| extracellular | 3.67 [2.42,5.42] | $8.3 \times 10^{-9}$ |
| nucleus | 0.33 [0.20,0.52] | $2.3 \times 10^{-8}$ |
| ER | 0.92 [0.68,1.24] | $6.2 \times 10^{-1}$ |
| cytoplasm | 1.31 [1.01,1.70] | $3.5 \times 10^{-2}$ |
| U%<D1 | 1.63 [1.13,2.32] | $7.7 \times 10^{-3}$ |
| U%>D9 | 0.18 [0.07,0.43] | $1.1 \times 10^{-6}$ |
| A%<D1 | 2.92 [2.14,3.93] | $7.7 \times 10^{-11}$ |
| A%>D9 | 0.11 [0.03,0.32] | $2.6 \times 10^{-8}$ |
| C%<D1 | 0.07 [0.02,0.26] | $3.1 \times 10^{-9}$ |
| C%>D9 | 3.05 [2.25,4.10] | $1.1 \times 10^{-11}$ |
| G%<D1 | 0.18 [0.07,0.43] | $1.1 \times 10^{-6}$ |
| G%>D9 | 1.14 [0.75,1.70] | $5.2 \times 10^{-1}$ |
| MFE<D1 | 0.04 [0.01,0.20] | $1.7 \times 10^{-10}$ |
| MFE>D9 | 3.25 [2.41,4.36] | $2.4 \times 10^{-13}$ |
| length<D1 | 4.75 [3.60,6.25] | $8.5 \times 10^{-25}$ |
| length>D9 | 0.04 [0.01,0.20] | $1.8 \times 10^{-10}$ |

Ovary

| Geneset | odds.ratio(CI:95%) | pvalue |
|---|---|---|
| extracellular | 5.71 [4.57, 7.10] | $1.4 \times 10^{-42}$ |
| nucleus | 0.84 [0.69, 1.02] | $8.1 \times 10^{-2}$ |
| ER | 0.74 [0.61, 0.89] | $9.5 \times 10^{-4}$ |
| cytoplasm | 0.81 [0.69, 0.95] | $7.1 \times 10^{-3}$ |
| U%<D1 | 0.53 [0.38, 0.73] | $4 \times 10^{-5}$ |
| U%>D9 | 0.69 [0.51, 0.92] | $1.1 \times 10^{-2}$ |
| A%<D1 | 1.68 [1.36, 2.07] | $5.1 \times 10^{-6}$ |
| A%>D9 | 0.36 [0.25, 0.53] | $2 \times 10^{-9}$ |
| C%<D1 | 0.24 [0.15, 0.38] | $3.2 \times 10^{-14}$ |
| C%>D9 | 1.55 [1.24, 1.92] | $1.5 \times 10^{-4}$ |
| G%<D1 | 0.69 [0.51, 0.92] | $1.1 \times 10^{-2}$ |
| G%>D9 | 0.33 [0.22, 0.50] | $2.4 \times 10^{-10}$ |
| MFE<D1 | 0.04 [0.02, 0.11] | $1.5 \times 10^{-28}$ |
| MFE>D9 | 7.77 [6.62, 9.12] | $2.5 \times 10^{-120}$ |
| length<D1 | 9.74 [8.31,11.42] | $2 \times 10^{-153}$ |
| length>D9 | 0.05 [0.02, 0.13] | $3.5 \times 10^{-27}$ |

**Figure 4.** The gene properties correlated to genes more degraded in FFPE tissues. (**A**–**D**) Fisher's exact test results of more degraded genes in FFPE samples of four different tissues; ER, endoplasmic reticulum; D1, first Deciles; D9, ninth Deciles. (**E**) Violin plot showing the transcript length in different DEG group. The genes significantly more degraded in FFPE are significantly shorter. (**F**) Violin plot of the MFE of secondary structure in different DEG groups. The genes significantly more degraded in FFPE have significantly higher MFE.

Further, we identified the DEGs in tumor versus peritumor in FF samples and FFPE samples, respectively. Among the upregulated genes, 38% (217 genes) were consistently identified in FF and FFPE samples (Figure 6E). For downregulated genes, 31% (96 genes) were consistently identified in FF and FFPE samples (Figure 6F). The consistency between FFPE and FF samples was lower than in previous study (12). We then performed KEGG pathway enrichment analysis using the upregulated/downregulated genes and screened the top 10 significant pathways. The upregulated genes were consistently overrepresented in seven KEGG pathways across FF and FFPE samples (Figure 6G). We also found four KEGG pathways downregulated in tumors across FF and FFPE samples (Figure 6H).

To assess the feasibility of FFPE samples for cancer research, we utilized the pan-cancer gene annotations (see Materials and methods) for GSEA. Notably, we found eight cancer-related pathways significantly upregulated in tumor versus peritumor in FF samples. In FFPE samples, 10 cancer-related pathways were significantly upregulated in tumor tissues, including all 8 pathways identified in FF samples. These results suggested that FFPE samples can be utilized for cancer-related

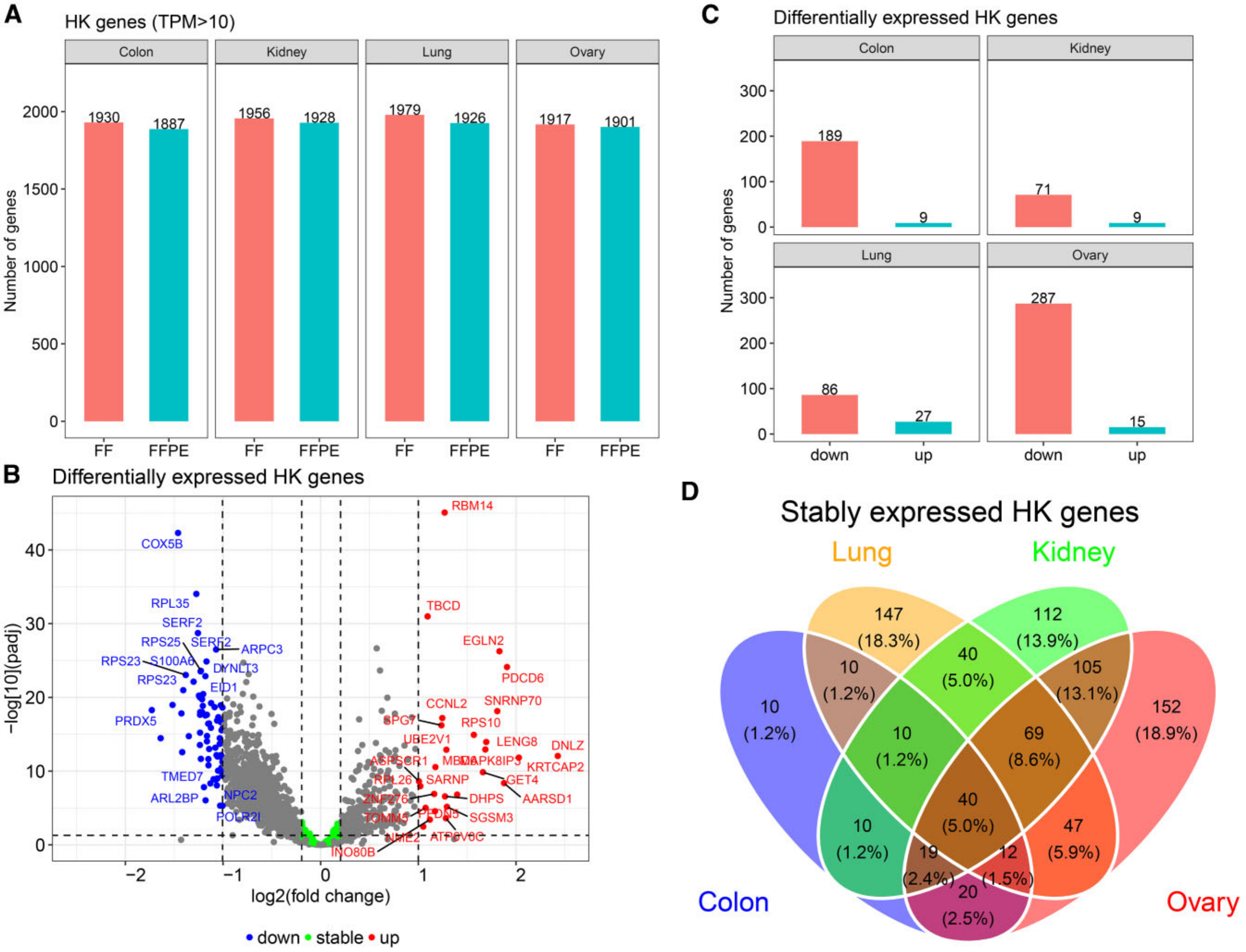


**Figure 5.** The overview of the expression of housekeeping genes in FF and FFPE samples. (**A**) Histogram showing the number of HK genes with minimum expression >10 TPM. (**B**) The volcano plot of differentially expressed HK genes and genes stably expressed in FF and FFPE samples from lung tissues. (**C**) Histogram showing the number of differentially expressed HK genes in four tissues. (**D**) Venn diagram showing the overlap of HK genes with small changes across four tissues.

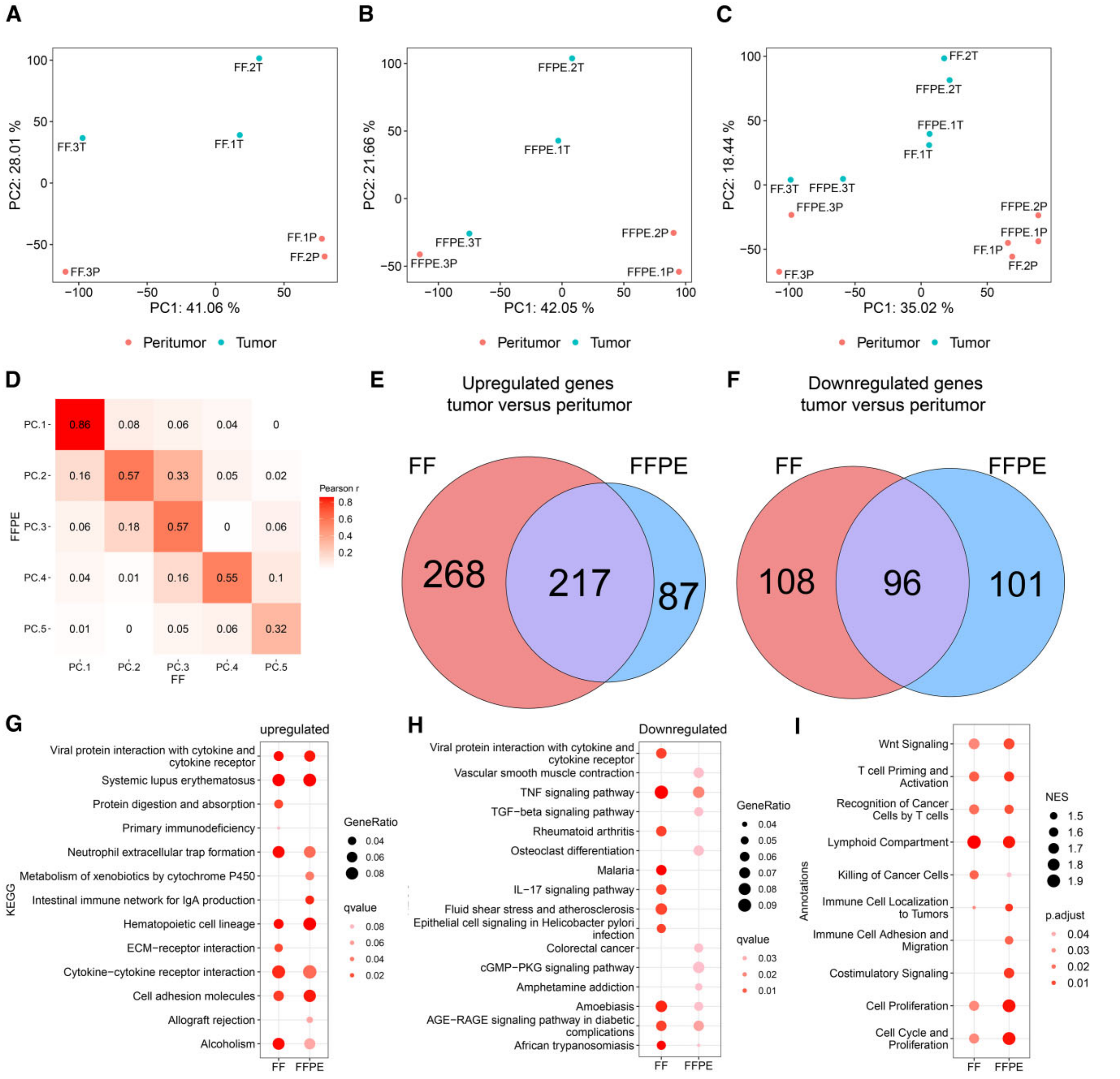
comparative transcriptome analysis and data mining to a considerable extent.

## Discussion

FFPE tissues are the most common human tissue specimens in clinical practice. Preparing FFPE samples with high-quality RNA is essential for successful sequencing and downstream analysis. Understanding the pattern of changes in expression data due to FFPE treatment may also help develop FFPE RNA-seq analysis approaches and better harness the power of FFPE tissue resources.

In previous studies, several factors were found to affect the RNA quality of FFPE tissues, including cold ischemia time (22), specimen size, fixation time, storage time and temperature (37). For the storage condition, a recent study showed that the degradation of nucleic acids was slower when FFPE tissues were stored at 4°C or lower temperatures (38). Our study aimed to determine an optimal condition to prepare FFPE specimens for high-quality RNA. Our finding showed that the quality of RNA extracted from FFPE specimens has little relation to the temperature of fixation but has a positive correlation to the time of fixation ranging from 12 to 48 h. The prolonged ischemia time also distinctly influenced the RNA quality at room temperature. Besides, the sampling method affected the FFPE RNA quality as well. The DV200 of RNA extracted from paraffin scrolls was significantly higher than paraffin sections. Overall, we determined that the optimal preparation conditions of FFPE are ischemia of tissues at 4°C (<48 h) or ischemia at 25°C for short time (0.5 h), fixation for 48 h at 25°C and sampling from FFPE scrolls rather than FFPE sections. In addition, FFPE should be stored at 4°C or lower temperatures to slow down RNA degradation (38).

The systemic differences in the RNA-seq data were observed between FFPE and FF samples, independent of tissue types (5,12,17,19). We found that more reads were mapped to genes associated with the perception of chemical stimulus in FFPE samples, which may be related to biological processes arising during paraformaldehyde fixation. Additionally, we first found that the more degraded gene was more likely to present in the genes with short transcript and high MFE of secondary structure. Moreover, we found that the genes localized to extracellular regions such as extracellular vesicles (31,39) were more likely to downgrade in FFPE samples across four tissues, indicating that these RNAs may be more likely to degrade in FFPE samples. Notably, mitochondrial genes were



| FFPE \ FF | PC.1 | PC.2 | PC.3 | PC.4 | PC.5 |
|---|---|---|---|---|---|
| PC.1 | 0.86 | 0.08 | 0.06 | 0.04 | 0 |
| PC.2 | 0.16 | 0.57 | 0.33 | 0.05 | 0.02 |
| PC.3 | 0.06 | 0.18 | 0.57 | 0 | 0.06 |
| PC.4 | 0.04 | 0.01 | 0.16 | 0.55 | 0.1 |
| PC.5 | 0.01 | 0 | 0.05 | 0.06 | 0.32 |

**Figure 6.** Consistency of analysis results between FF and FFPE samples. (A–C) The dot graph showing the distributions of three pairs of control-case-matched FF (**A**) and FFPE (**B**) samples in the 2D after dimensionality reduction. (**D**) Heatmap showing the correlation between the top 5 principal components from FF and FFPE samples. (**E** and **F**) Venn diagram showing overlap of upregulated (E) and downregulated (F) genes in tumor versus peritumor between FFPE samples and FF samples. (**G** and **H**) Dot graph showing top 5 significantly upregulated (G) and downregulated (H) KEGG pathways overlapped in FF and FFPE samples. (**I**) Dot graph showing GSEA results in cancer-related pathway. The results obtained from FF and FFPE samples are highly consistent.

consistently over-sampled in FFPE samples, which is consistent with a previous study (19). This phenomenon suggested that RNA in mitochondria may be protected from degradation by the mitochondrial membranes and RNA-binding proteins within the mitochondria (40). Unexpectedly, almost no gene correlated with the DV200 values of the FFPE samples, indicating that the DV200 values were only suitable to evaluate the probability of successful library construction (8,9).

Housekeeping genes are involved in fundamental cell biological processes. Thus, they are expected to maintain a constant level of expression in all cells and conditions (41). Owing to this characteristic, housekeeping genes are commonly used as internal reference genes (42,43). However, the systemic differences between FFPE and FF samples altered the expression of many housekeeping genes. We found 40 housekeeping genes with a small change between FF and FFPE samples across four tissues. These genes may be the candidates of internal reference to normalize gene expression data in both FF and FFPE samples. The consistency of DEGs is relatively lower than in the previous study (12). However, from the result of PCA, we found that the major variations in the data were highly consistent across FFPE and FF samples. Although many different genes were detected in differential expression analysis, the enrichment analysis results are similar. In particular, the results of cancer-related gene sets were highly consistent across FFPE and FF samples. Our findings suggested that using the FFPE samples may produce similar results as FF samples in comparative expression analyses between control and case samples.

In summary, we have found an optimal FFPE specimen preparation condition to obtain high-quality RNA. In addition, we discovered that the genes with specific attributes were more prone to alter expression in FFPE samples. We evaluated the feasibility of FFPE RNA-seq data using control-case-matched samples and demonstrated that FFPE samples could be used for cancer-related transcriptome analysis. Our findings provide insights that can inform the clinical preparation of FFPE tissues, enhance our understanding of expression data alteration induced by FFPE treatment and optimize the utilization of FFPE tissue resources.

## Data and code availability

The data that support the findings of this study are available in the Genome Sequence Archive for Human (GSA-Human) at https://ngdc.cncb.ac.cn/gsa-human/s/0AeSL3An; reference number (HRA004941). The public data of the BPV research program are available through dbGaP (#phs001304). The scripts used to reproduce the main results in this study are available in Github repository (https://github.com/liny-suts/FFPE) and Zenodo (https://doi.org/10.5281/zenodo.10516663).

## Supplementary data

Supplementary Data are available at NARGAB Online.

## Acknowledgements

*Author contributions*: A.G., J.G. and J.L. conceived the study. Z.D. was responsible for the preparation and sampling of FFPE samples under different conditions and the acquisition of clinical information. T.Y. contributed RNA quality information and performed RNA sequencing. Y.L. designed and carried out the data analysis. A.G., Y.L and J.Y wrote the manuscript. M.X applied for ethical approval for this project. All authors read, edited and approved the final paper.

## Funding

National Natural Science Foundation of China [32370717].

## Conflict of interest statement

None declared.

## References

1. Talebi,A., Thiery,J.P. and Kerachian,M.A. (2021) Fusion transcript discovery using RNA sequencing in formalin-fixed paraffin-embedded specimen. *Crit. Rev. Oncol. Hematol.*, **160**, 103303.
2. Cazzato,G., Caporusso,C., Arezzo,F., Cimmino,A., Colagrande,A., Loizzi,V., Cormio,G., Lettini,T., Maiorano,E., Scarcella,V.S., *et al.* (2021) Formalin-fixed and paraffin-embedded samples for next generation sequencing: problems and solutions. *Genes (Basel)*, **12**, 1472.
3. Pennock,N.D., Jindal,S., Horton,W., Sun,D., Narasimhan,J., Carbone,L., Fei,S.S., Searles,R., Harrington,C.A., Burchard,J., *et al.* (2019) RNA-seq from archival FFPE breast cancer samples: molecular pathway fidelity and novel discovery. *BMC Med. Genomics*, **12**, 195.
4. Carrick,D.M., Mehaffey,M.G., Sachs,M.C., Altekruse,S., Camalier,C., Chuaqui,R., Cozen,W., Das,B., Hernandez,B.Y., Lih,C.J., *et al.* (2015) Robustness of next generation sequencing on older formalin-fixed paraffin-embedded tissue. *PLoS One*, **10**, e0127353.
5. Hedegaard,J., Thorsen,K., Lund,M.K., Hein,A.M., Hamilton-Dutoit,S.J., Vang,S., Nordentoft,I., Birkenkamp-Demtröder,K., Kruhøffer,M., Hager,H., *et al.* (2014) 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**, e98187.
6. Okello,J.B., Zurek,J., Devault,A.M., Kuch,M., Okwi,A.L., Sewankambo,N.K., Bimenya,G.S., Poinar,D. and Poinar,H.N. (2010) Comparison of methods in the recovery of nucleic acids from archival formalin-fixed paraffin-embedded autopsy tissues. *Anal. Biochem.*, **400**, 110–117.
7. Karmakar,S., Harcourt,E.M., Hewings,D.S., Scherer,F., Lovejoy,A.F., Kurtz,D.M., Ehrenschwender,T., Barandun,L.J., Roost,C., Alizadeh,A.A., *et al.* (2015) Organocatalytic removal of formaldehyde adducts from RNA and DNA bases. *Nat. Chem.*, **7**, 752–758.
8. Illumina (2016) *Evaluating RNA Quality from FFPE Samples.* www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/evaluating-rna-quality-from-ffpe-samples-technical-note-470-2014-001.pdf(18 April 2023, date last accessed).
9. Matsubara,T., Soh,J., Morita,M., Uwabo,T., Tomida,S., Fujiwara,T., Kanazawa,S., Toyooka,S. and Hirasawa,A. (2020) DV200 index for assessing rna integrity in next-generation sequencing. *Biomed. Res. Int.*, **2020**, 9349132.
10. Bass,B.P., Engel,K.B., Greytak,S.R. and Moore,H.M. (2014) A review of preanalytical factors affecting molecular, protein, and morphological analysis of formalin-fixed, paraffin-embedded (FFPE) tissue: how well do you know your FFPE specimen?*Arch. Pathol. Lab. Med.*, **138**, 1520–1530.
11. Lesluyes,T., Pérot,G., Largeau,M.R., Brulard,C., Lagarde,P., Dapremont,V., Lucchesi,C., Neuville,A., Terrier,P., Vince-Ranchère,D., *et al.* (2016) RNA sequencing validation of the Complexity INdex in SARComas prognostic signature. *Eur. J. Cancer*, **57**, 104–111.
12. Jones,W., Greytak,S., Odeh,H., Guan,P., Powers,J., Bavarva,J. and Moore,H.M. (2019) Deleterious effects of formalin-fixation and delays to fixation on RNA and miRNA-Seq profiles. *Sci. Rep.*, **9**, 6980.
13. Zhao,Y., Mehta,M., Walton,A., Talsania,K., Levin,Y., Shetty,J., Gillanders,E.M., Tran,B. and Carrick,D.M. (2019) Robustness of RNA sequencing on older formalin-fixed paraffin-embedded tissue from high-grade ovarian serous adenocarcinomas. *PLoS One*, **14**, e0216050.
14. Huang,W., Goldfischer,M., Babayeva,S., Mao,Y., Volyanskyy,K., Dimitrova,N., Fallon,J.T. and Zhong,M. (2015) Identification of a novel PARP14-TFE3 gene fusion from 10-year-old FFPE tissue by RNA-seq. *Genes Chromosomes Cancer*, **54**, 500–505.
15. Jovanović,B., Sheng,Q., Seitz,R.S., Lawrence,K.D., Morris,S.W., Thomas,L.R., Hout,D.R., Schweitzer,B.L., Guo,Y., Pietenpol,J.A., *et al.* (2017) Comparison of triple-negative breast cancer molecular subtyping using RNA from matched fresh-frozen versus formalin-fixed paraffin-embedded tissue. *BMC Cancer*, **17**, 241.
16. Li,P., Conley,A., Zhang,H. and Kim,H.L. (2014) Whole-Transcriptome profiling of formalin-fixed, paraffin-embedded renal cell carcinoma by RNA-seq. *BMC Genomics*, **15**, 1087.
17. Graw,S., Meier,R., Minn,K., Bloomer,C., Godwin,A.K., Fridley,B., Vlad,A., Beyerlein,P. and Chien,J. (2015) Robust gene expression

and mutation analyses of RNA-sequencing of formalin-fixed diagnostic tumor samples. *Sci. Rep.*, **5**, 12335.
18. Vukmirovic,M., Herazo-Maya,J.D., Blackmon,J., Skodric-Trifunovic,V., Jovanovic,D., Pavlovic,S., Stojsic,J., Zeljkovic,V., Yan,X., Homer,R., *et al.* (2017) 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 Pulm. Med.*, **17**, 15.
19. Esteve-Codina,A., Arpi,O., Martinez-García,M., Pineda,E., Mallo,M., Gut,M., Carrato,C., Rovira,A., Lopez,R., Tortosa,A., *et al.* (2017) A comparison of RNA-Seq results from paired formalin-fixed paraffin-embedded and fresh-frozen glioblastoma tissue samples. *PLoS One*, **12**, e0170632.
20. Yin,S., Wang,X., Jia,G. and Xie,Y. (2020) MIXnorm: normalizing RNA-seq data from formalin-fixed paraffin-embedded samples. *Bioinformatics*, **36**, 3401–3408.
21. Yin,S., Zhan,X., Yao,B., Xiao,G., Wang,X. and Xie,Y. (2021) SMIXnorm: fast and accurate RNA-Seq data normalization for formalin-fixed paraffin-embedded samples. *Front. Genet.*, **12**, 650795.
22. Carithers,L.J., Agarwal,R., Guan,P., Odeh,H., Sachs,M.C., Engel,K.B., Greytak,S.R., Barcus,M., Soria,C., Lih,C.J., *et al.* (2019) The biospecimen preanalytical variables program: a multiassay comparison of effects of delay to fixation and fixation duration on nucleic acid quality. *Arch. Pathol. Lab. Med.*, **143**, 1106–1118.
23. Chen,S., Zhou,Y., Chen,Y. and Gu,J. (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, **34**, i884–i890.
24. Kim,D., Langmead,B. and Salzberg,S.L. (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*, **12**, 357–360.
25. Liao,Y., Smyth,G.K. and Shi,W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.
26. Pertea,M., Pertea,G.M., Antonescu,C.M., Chang,T.C., Mendell,J.T. and Salzberg,S.L. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.*, **33**, 290–295.
27. Wang,L., Wang,S. and Li,W. (2012) RSeQC: quality control of RNA-seq experiments. *Bioinformatics*, **28**, 2184–2185.
28. Shen,S., Park,J.W., Lu,Z.X., Lin,L., Henry,M.D., Wu,Y.N., Zhou,Q. and Xing,Y. (2014) rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl. Acad. Sci. USA*, **111**, E5593–E5601.
29. Durinck,S., Spellman,P.T., Birney,E. and Huber,W. (2009) Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.*, **4**, 1184–1191.
30. Lorenz,R., Bernhart,S.H., Höner Zu Siederdissen,C., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, **6**, 26.
31. Garg,A., Singhal,N., Kumar,R. and Kumar,M. (2020) mRNALoc: a novel machine-learning based in-silico tool to predict mRNA subcellular localization. *Nucleic Acids Res.*, **48**, W239–W243.
32. Lê,S., Josse,J. and Husson,F. (2008) FactoMineR: an R package for multivariate analysis. *J. Stat. Softw*, **25**, 1–18.
33. Hounkpe,B.W., Chenou,F., de Lima,F. and De Paula,E.V. (2021) HRT Atlas v1.0 database: redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive RNA-seq datasets. *Nucleic Acids Res.*, **49**, D947–D955.
34. Lin,L.I. (1989) A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, **45**, 255–268.
35. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
36. Yu,G., Wang,L.G., Han,Y. and He,Q.Y. (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*, **16**, 284–287.
37. von Ahlfen,S., Missel,A., Bendrat,K. and Schlumpberger,M. (2007) Determinants of RNA quality from FFPE samples. *PLoS One*, **2**, e1261.
38. Groelz,D., Viertler,C., Pabst,D., Dettmann,N. and Zatloukal,K. (2018) Impact of storage conditions on the quality of nucleic acids in paraffin embedded tissues. *PLoS One*, **13**, e0203608.
39. Cui,T., Dou,Y., Tan,P., Ni,Z., Liu,T., Wang,D., Huang,Y., Cai,K., Zhao,X., Xu,D., *et al.* (2022) RNALocate v2.0: an updated resource for RNA subcellular localization with increased coverage and annotation. *Nucleic Acids Res.*, **50**, D333–D339.
40. Jedynak-Slyvka,M., Jabczynska,A. and Szczesny,R.J. (2021) Human mitochondrial RNA processing and modifications: overview. *Int. J. Mol. Sci.*, **22**, 7999.
41. Eisenberg,E. and Levanon,E.Y. (2013) Human housekeeping genes, revisited. *Trends Genet.*, **29**, 569–574.
42. Mu,J., Chen,L., Gu,Y., Duan,L., Han,S., Li,Y., Yan,Y. and Li,X. (2019) Genome-wide identification of internal reference genes for normalization of gene expression values during endosperm development in wheat. *J. Appl. Genet.*, **60**, 233–241.
43. Dos Santos,K.C.G., Desgagné-Penix,I. and Germain,H. (2020) Custom selected reference genes outperform pre-defined reference genes in transcriptomic analysis. *BMC Genomics*, **21**, 35.

**Received:** October 31, 2023. **Revised:** December 30, 2023. **Editorial Decision:** January 12, 2024. **Accepted:** January 16, 2024