Reliable RNA-seq analysis from FFPE specimens as a means to accelerate cancer-related health disparities research

Frederick M.J.,[1] Perez-Bello D.,[1] Yadollahi P.,[1] Castro P.,[2] Frederick A.,[3] Frederick A.,[3] Osman R.A., Essien F.,[1] Yebra I., [1] Hamlin A.,[1] Ow, T.J.,[5,6] Skinner H.D.,[7] Sandulache V.C. [1, 8, 9]

Author's Affiliations:
1- Bobby R. Alford Department of Otolaryngology Head and Neck Surgery, Baylor College of Medicine, Houston, TX
2- Department of Pathology & Immunology, Baylor College of Medicine, Houston, TX
3- Informagen, Wilmington, DE
4- Department of Biological Sciences, Vanderbilt University College of Arts and Science, Nashville, TN
5- Department of Otorhinolaryngology-Head and Neck Surgery, Montefiore Medical Center, Bronx, NY
6- Department of Pathology, Montefiore Medical Center, Bronx, NY
7- Department of Radiation Oncology, UPMC Hilman Cancer Center, Pittsburgh PA
8- ENT Section, Operative CareLine, Michael E. DeBakey VAMC, Houston, TX
9- Center for Translational Research on Inflammatory Diseases, Michael E. DeBakey Veterans Affairs Medical Center, Houston, TX

Co-corresponding authors:
Vlad C. Sandulache MD PhD
Bobby R. Alford Department of Otolaryngology – Head and Neck Surgery
Baylor College of Medicine
One Baylor Plaza, MS: NA102
Houston, TX 77030
Phone: (713) 798-7218
Fax: (713) 798-5078
Email: vlad.sandulache@bcm.edu

Mitchell J. Frederick PhD
Bobby R. Alford Department of Otolaryngology – Head and Neck Surgery
Baylor College of Medicine
One Baylor Plaza, MS: NA102
Houston, TX 77030
Phone: (713) 798-7218
Fax: (713) 798-5078
Email: Mitchell. Frederick@bcm.edu

Running title: RNA and DNA extraction from archival tissue

Keywords: RNA, DNA, ancestry, FFPE

Data availability: The data generated in this study are available within the article and its supplementary data files.

Conflict of interest statement: The authors report no conflicts of interests relevant to the work summarized in the current manuscript. V.C.S. is a consultant and equity holder in Femtovox Inc. and a consultant for PDS Biotechnology.

Funding: This work was supported by the National Institute of Dental and Craniofacial Research (R21DE032344), the National Institutes of Health (R01DE0323337, P50 CA097190, R01 DE028061), the Dan L Duncan Comprehensive Cancer Center (P30-CA125123), the Human Tissue Acquisition & Pathology Core Baylor College of Medicine, a Price Family Foundation Pilot Project to Develop Experimental Cancer

Therapeutics for Underrepresented Populations, awarded by the Montefiore Einstein Comprehensive Cancer Center and an Administrative Supplement to Support Diversity Research awarded under NIH-NCI. 3U54CA274321-02S1.

Abstract

Whole transcriptome sequencing (WTS/ RNA-Seq) is a ubiquitous tool for investigating cancer biology. RNA isolated from frozen sources limits possible studies for analysis of associations with phenotypes or clinical variables requiring long-term follow-up. Although good correlations are reported in RNA-Seq data from paired frozen and formalin fixed paraffin embedded (FFPE) samples, uncertainties regarding RNA quality, methods of extraction, and data reliability are hurdles to utilization of archival samples. We compared three different platforms for performing RNA-seq using archival FFPE oropharyngeal squamous carcinoma (OPSCC) specimens stored up to 20 years, as part of an investigation of transcriptional profiles related to health disparities. We developed guidelines to purify DNA and RNA from FFPE tissue and perform downstream RNA-seq and DNA SNP arrays. RNA was extracted from 150 specimens, with an average yield of 401.8 ng/cm^2 of tissue. Most samples yielded sufficient RNA reads >13,000 protein coding genes which could be used to differentiate HPV-associated from HPV-independent OPSCCs. Co-isolated DNA was used to identify patient ancestry. Utilizing the methods described in this study provides a robust, reliable, and standardized means of DNA & RNA extraction from FFPE as well as a means by which to assure the quality of the data generated.

Introduction

Owing to technical advances in next generation sequencing (NGS) and widespread optimization of protocols and commercial kits for RNA extraction, whole transcriptome RNA expression profiling (i.e, RNA-seq) has become a ubiquitous research tool. Often, RNA-seq is performed using RNA isolated from fresh or snap-frozen samples.[1-5] While such investigations have provided insight into the biology of diseases, the availability of fresh or frozen samples can constrain study designs, create selection bias, and limit research endeavors that require long-term follow-up. This is unfortunate because many research institutions and laboratories have amassed large biobanks of archival specimens collected over decades during routine diagnosis, medical practice, or drug studies, which far outnumber collections of fresh/frozen tissue.[6-8] Therefore, the ability to reliably perform bulk RNA-seq profiling from archival formalin fixed paraffin embedded (FFPE) specimens would greatly expand scientific research.[9] Despite evidence that RNA-seq can successfully be performed using FFPE specimens of varying age,[10-12] uncertainties regarding optimal RNA extraction procedures and the variability/ reliability of whole transcriptome expression data generated remain the biggest obstacles preventing more widespread acceptance and application.[13, 14]

One area that would benefit immensely from routine RNA-seq analysis of archival FFPE samples is health disparities research. In addition to socio-economic and cultural differences, there is evidence that disparities in clinical disease and outcome among minority populations could also have a biological basis.[15-17] If so, it is likely that known differences in single nucleotide polymorphisms (SNPs) underpinning ancestral differences at the genomic level could exert their phenotypic impact by influencing RNA expression levels of specific genes.[15, 18] Moreover, demographic data collected through medical records can be imperfect[19] or altogether missing, making genomic ancestry a potentially more accurate method.[20] Due to this need, we describe optimized procedures to extract both RNA and DNA from FFPE tumor sections, guidelines for performing quality control (QC) of extracted nucleic acid, recommendations for specific RNA-seq and DNA SNP platforms, and suggest approaches for standardized data processing and data QC.

Results

Isolation and quality control of RNA from FFPE specimens. Various publications have compared in-house and commercial kit-based procedures for optimally extracting RNA, DNA, or both from FFPE tissue.[14, 21-25] Here, rather than re-investigate all the variables, our goal was to provide researchers with a detailed set of guidelines and instructions—all in one place— to successfully purify DNA and RNA from FFPE tissue and for downstream analyses like RNA-seq and DNA SNP arrays. Our procedures were extensively vetted and yielded robust results when applied by multiple individuals in our group, regardless of experience level. Although we make no claim that our methods produce superior results, we present them as best practices that could enhance the likelihood of experimental success using FFPE prepared tissue.

We opted to first isolate RNA using the AllPrep DNA/RNA FFPE kit from Qiagen and further process residual DNA pellets using the COBAS kit from Roche (see below), with some modifications to the manufacturers' instructions (**Supplementary Methods**). All RNA/DNA isolations began from pre-cut slides of oropharyngeal squamous cell carcinomas (OPSCC) prepared from archival blocks that ranged in storage time from 1 to 20 years. Alternatively, sections could be sliced directly into Eppendorf tubes and stored at room temperature in a desiccator prior to processing. FFPE slide sections were cut to 6 µm thickness; **Supplementary Table 1** summarizes the physical properties from 43 FFPE samples processed for RNA isolation and commercially sequenced, along with an additional 150 samples also analyzed for RNA yield but not sequenced. A breakdown of yields for the 43 sequenced specimens and for all 193 samples processed is provided in **Supplementary Table 2**. For the 43 sequenced samples, the average total RNA yield was 3629 ng (range of 485 ng to 13,247 ng), average area of tissue per slide was 2 cm, average number of slides used was ~9, average total area of tissue processed was 18 cm, and average yield per total area of tissue was 402

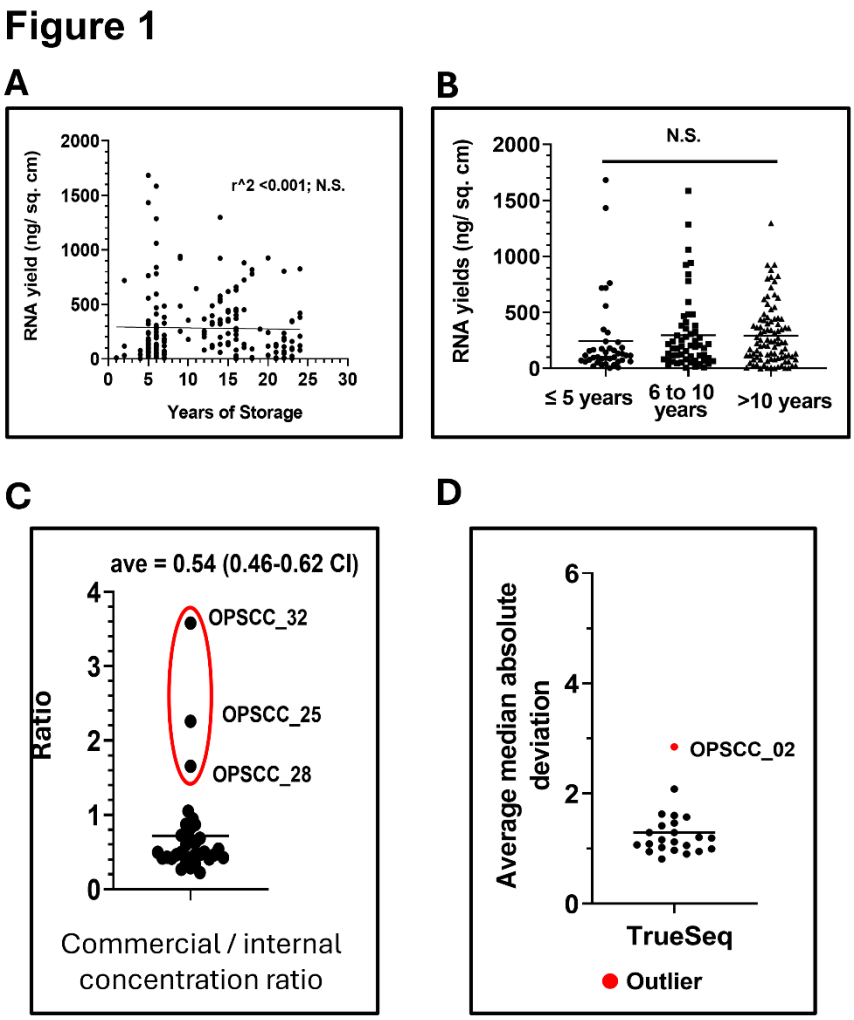

Figure 1. RNA yields. There was no correlation between RNA yields and relative age of the individual clinical specimens (A,B). C) Laboratory concentrations were approximately 50% of concentrations measured at the commercial vendor. D) Technical outlier detection using the MAD method, showing the average MAD values from the TrueSeq cohort, which cluster well except for outlier sample OPSCC_02 (Adj P <0.0001, Supplementary Table 5).

ng/cm^2 (median = 274 ng/cm). Most standard commercial RNA sequencing platforms can work with 200-500 ng RNA, albeit the true yields measured by fluorescent methods were usually half that determined by absorbance as discussed further below. Summary sample metrics for the entire cohort of 193 specimens were not fundamentally different from the subset of 43 selected for sequencing (**Supplementary Table 2**). We did not find a statistically significant relationship between how long tissue blocks had been stored and the RNA yields per square centimeter of tissue (**Fig. 1A**).

We qualitatively and quantitatively compared the reliability of three different library/platforms for performing whole transcriptomic RNA-Seq from archival FFPE specimens (**Supplementary Table 1**). RNA extracted from a subset of FFPE specimens was processed through one commercial sequencing company that used either the TruSeq RNA exome kit from Illumina (recommended specifically for FFPE samples) for library preparation or the Illumina TruSeq Stranded kit followed by removal of noncoding RNA with the Ribo-zero kit. Another set of RNA samples was processed through a different vendor for massive analysis of cDNA ends (MACE-seq) that captures 3' ends of polyadenylated RNA, but samples were unavoidably stored at 4°C for several days during transit. However, MACE-seq potentially performs better than other platforms with partially degraded RNA. [26, 27]

We compared the RNA concentration measured in-house by absorbance to that obtained through fluorescence methods performed commercially (**Supplementary Table 3, Fig. 1B**, **Supplementary Fig. S1**), excluding samples that were improperly stored during transit. On average, in-house concentrations measured by absorbance were about half that obtained through fluorescence techniques (**Fig. 1B**). Not surprisingly for FFPE specimens, RIN values were under the recommended value of 7 or higher and ranged from 1.2 to 2.5 (**Supplementary Table 3**). A more common metric for fragmented RNA is the DV200 (i.e., percentage of fragments ≥ 200 bp) which varied widely from 1.48% to 71.47% (**Supplementary Table 3**), with a median value of just 18.65, excluding the samples used for MACE-seq that were stored at warmer temperatures for an extended time.

Recommendations for processing RNA-Seq data. There are many publications examining approaches for processing RNA-Seq data,[28-32] each with its merits. Here, we present a user-friendly pipeline (**Supplementary Fig. 2**) and Python code (**Supplementary Methods**) that normalizes RNA expression data in a standardized fashion from raw reads, puts gene expression values on the same scale regardless of the sequencing platform or depth of coverage, and controls for inter-sample variability. Furthermore, we incorporate procedures for consistent identification of technical outlier samples and minimizing the influence of low expression or zero values. These steps are not intended to substitute for batch corrections but do allow some gross comparisons and inferences to be made across experiments spanning different platforms. Nonetheless, our normalization methods and their outputs should be compatible with batch adjustment methods like the frequently used Combat algorithm.[33, 34] Popular RNA-seq normalization methods include calculating fragments per kilobase of transcript per million reads (FPKM), or transcripts per million (TPM). We use an upper quartile (UQ) normalization approach known as FPKM-UQ, frequently adopted by The Cancer Genome Atlas (TCGA) project.[35-37] Each of these methods have been used successfully and they all control for gene size—which while not strictly necessary for analyzing differences in gene expression—should be employed whenever single sample gene set enrichment (ssGSEA) is a downstream analysis.[9, 38] The UQ normalization in our pipeline divides gene expression values by a sample specific UQ value (i.e., 75th percentile excluding zeros), rather than a sample total, and has been shown by others to control well for inter-sample variability.[39, 40]

According to our data processing pipeline shown in **Supplementary Fig. 2** the starting input file should contain meta data (i.e., combined sample data) with individual gene sizes and either raw counts (e.g., HTSeq-Count) or expected counts if RNA-Seq by Expectation Maximization (RSEM) was used. Gene sizes are constant across samples and derived purely through reference annotation when HTSeq-Counts are employed, but for RSEM the effective gene sizes will vary slightly from sample to sample. It's worth

noting that many sequencing cores and companies already provide FPKM and TPM normalizations with the data.[41, 42] However, such values should be used cautiously, because these normalizations frequently do not filter out non-protein coding genes in the process. It is not uncommon for the gene expression data provided by sequencing centers to include tens of thousands of non-protein coding genes with 40,000 to 60,000 "genes" listed when in fact there the true number of protein coding human genes is in the range of 19,000-20,000 (depending on whether immune specific genes are included). For easy reference, we provide our comprehensive list of 19,709 human protein coding genes (**Supplementary Table 4**) derived from the HGNC website (**Supplementary Methods**) along with their official symbols, crossmatched to their Entrez gene ID's, Ensembl IDs, and HGNC ID's.

A consequence of our normalization approach is that some downstream software packages for differential analysis of gene expression (DEG) that require raw data input and perform their own internal normalization, such as DESeq2[43] or EdgeR,[44] are not appropriate. DEG packages like Limma[45] or Voom[46] which accept pre-normalized data for linear modeling can be used. For simpler analyses between groups, we typically combine multiple T-testing with the Benjamini Hochberg procedure for controlling false discovery rate (FDR). The main advantage with our data processing pipeline is better control over what is happening during normalization.

Analysis and comparison of RNA-Seq data sets generated from FFPE specimens using different sequencing approaches. The methods described here were developed to aid in comparing RNA-Seq data from multiple platforms, but they are appropriate for routine analysis of RNA-Seq experiments. The total mapped reads or sequencing depth varied greatly between the different library preparations/platforms we examined, with an average of roughly 17.3 million, 6.6 million, and 0.67 million for the TruSeq, TruSeq stranded, and MACE-seq approaches, respectively (**Supplementary Table 3, Supplementary Fig. 3**). We examined the relationship between total mapped reads and reads mapped to protein coding genes (**Supplementary Table 3, Supplementary Fig. 3**) and found best results with MACE-Seq (average = 88%), followed by TruSeq (average = 82%), compared to Stranded/Ribo-zero (average = 55%). Differences in sequencing depth between methods were reflected in the UQ values used during normalization (**Supplementary Table 3**).

The distributions of gene expression (i.e., median UQ normalized log 2 values) for each platform/cohort, along with the TCGA OPSCC RNA-Seq dataset similarly normalized, are shown in **Supplementary Fig. 4A,C,E, and G** before global rescaling. To directly compare the behavior of RNA values it is helpful to re-scale each cohort to have a common global median. However, samples that represent technical outliers should first be removed to avoid data distortion. Outliers were identified using a two-step procedure based on the Median Absolute Deviation (MAD) approach. First, for every gene, MAD values were computed by taking the absolute value of a sample's gene expression minus the cohort median and delta values were then averaged across all genes to calculate a sample specific average MAD value. In the second step, a linear model was fit across all the sample average MAD values to generate residuals, which were normalized to calculate the probability (i.e., p-value) of observing each sample's average MAD score using the cumulative distribution function with an FDR =0.01 for multiple testing. Samples OPSCC_02 (TrueSeq RNA exome), OPSCC_32 and 33 (Stranded RNA/Ribo-zero), OPSCC_34 and 35 (MACE-Seq) were all identified as outliers (**Supplementary Table 5**, FDR <0.0001, **Fig. 1C**, **Supplementary Fig. 5**) and tended to have some of the lowest 75th PCTL raw counts relative to other samples in their respective cohorts.

The global medians of each data set (excluding outliers) were then used to re-scale cohorts to obtain distributions with a new and common global median of 7. Essentially, each normalized log2 gene expression value was modified by adding (7- initial global median) to obtain rescaled or adjusted values, except that gene expression values corresponding to true zeros were not adjusted. Histograms for the newly rescaled cohorts were generated (**Supplementary Fig. 4B,D,F,H**). After this global adjustment, low

expression values could be treated uniformly across cohorts. To minimize data distortion and noise around low expression genes, all negative log2 expression values were set to zero. For downstream applications like DEG that require testing multiple genes and p-value corrections it is helpful to minimize the number of statistical tests performed or genes analyzed. One advantage of our normalization procedure that rescales all cohorts to the same global median of 7 is that a universal low expression threshold can be set. We recommend filtering out genes where the maximum average log2 expression of any group ≤ 2, which typically results in anywhere from 12,000 to 16,000 genes being tested (e.g., **Supplementary Table 5**). Finally, all re-scaled log2 values <0 were uniformly substituted with zeros, to minimize analysis artifacts associated with small values.

Figure 2

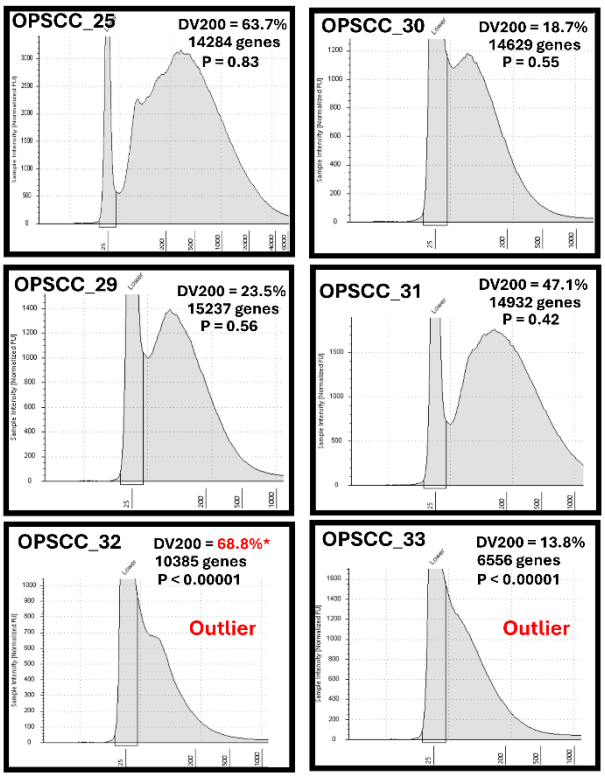

Figure 2. Relationship between DV200 and biological richness of gene expression. DV200 values and RNA fragment size are shown for representative samples along with the relative number of usable genes and outlier associated p-values from supplementary table 5. * denotes potentially erroneous DV200 reading.

RNA sample quality metrics driving successful sequencing. As mentioned, RNA RIN values are of little use for assessing RNA isolated from FFPE samples. After examining the different sequencing metrics described above, we re-considered which, if any, RNA sample QC measurements might be useful to predict good quality sequencing results. TapeStation readouts for representative samples sequenced with the Stranded library plus Ribozero approach (**Fig. 2**) or the TruSeq library (**Supplementary Fig. 6**) illustrated several differences among samples. Samples with large DV200 (e.g. percent of RNA molecules ≥ 200 bp) values, such as OPSCC_25 (DV200 = 63.7%) and OPSCC_31 (DV200 = 47.1%) yielded over 14,000 genes having log2 expression values ≥ 2 and were not among the outliers. Even samples with considerably lower DV200 values like OPSCC_09 (DV200 = 13.7%) or OPSCC_13 (DV200 =19.9%) also yielded a substantial number of genes expressed above the low threshold cut-off. Outliers like OPSCC_02, OPSCC_32, and OPSCC_33, on the other hand, visibly had some of the lowest DV200 values (<14%). However, several samples like OPSCC_01 and OPSCC_16 also had DV200 values below 10%, but performed reasonably well with over 13,000 genes each expressed above the low threshold. These latter samples had very subtle shoulders visible in their TapeStation curves that distinguished them from the outlier OPSCC_02 sample in the TruSeq cohort. Generally, the samples used in the Stranded/Ribo-zero library were of better quality based on TapeStation curves, with more a greater proportion of RNA fragments in the 100 bp range. Collectively, our experience suggests that although a high DV200 percentage is very likely to produce good reads, samples with peaks or shoulders in the 100 bp range may also generate acceptable sequencing data.

Biological Reliability of RNA-seq data from FFPE specimens. We found that wholesale examination of correlations among all protein coding genes between the TCGA cohort and our datasets was not

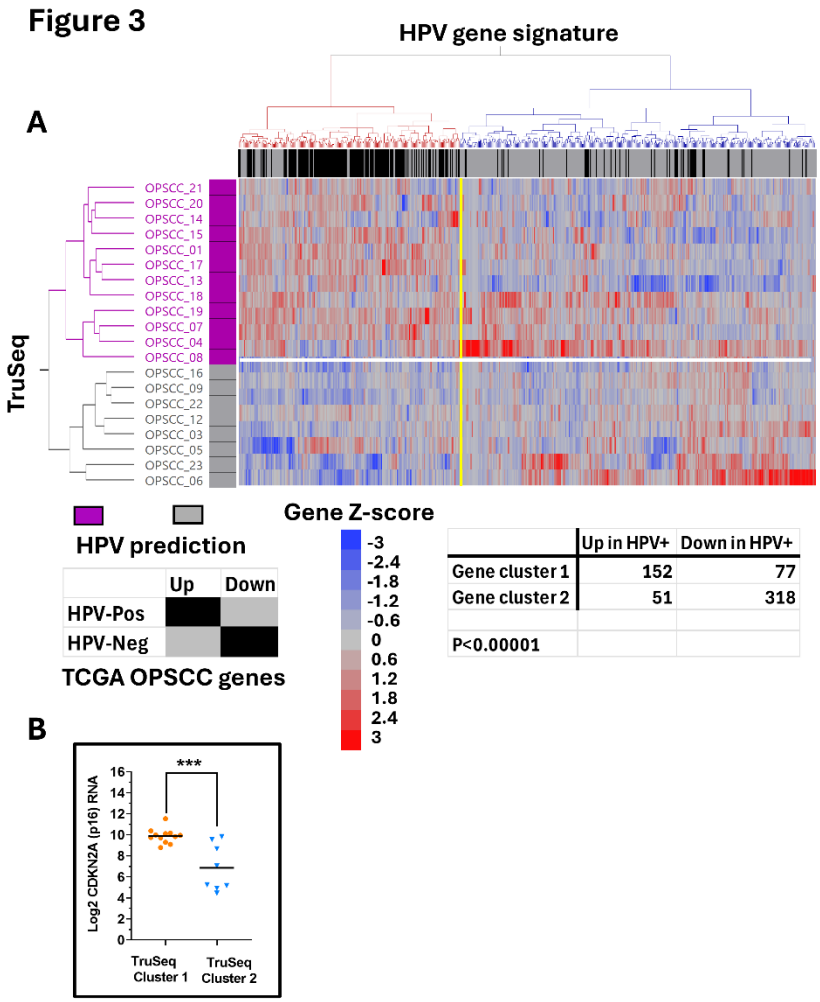

Figure 3. Correlation of HPV associated genes in TruSeq cohort. Expression of HPV-associated DEGs (identified from the TCGA OPSCC reference cohort) was used for unsupervised 2-way clustering of TruSeq samples (i.e., Ward's agglomerative hierarchical clustering) to predict HPV status (A). Samples predicted to be HPV-associated (i.e., HPV-Pos) are annotated with purple boxes. The regulation status of genes is annotated across the top of the heatmap with black boxes if they were also upregulated in HPV-associated TCGA samples or grey boxes if they were upregulated in HPV-independent (i.e., HPV-Neg) TCGA samples. Significant enrichment of genes upregulated in TCGA HPV-associated cancers was found in gene cluster 1 (red cluster) and enrichment of genes upregulated in HPV-independent TCGA samples is found in gene cluster 2 (blue cluster), demonstrating significant separation (e.g., P <0.00001 by Chi-square testing). Specimens from sample cluster 1 (purple) predicted to be HPV-associated based on their gene expression pattern had significantly higher CDKN2A (p16) expression (P) < 0.0005 than specimens from sample cluster 2 (grey) predicted to be HPV-independent (B), validating predictions.

particularly informative and seemingly good correlations existed even when tested against highly unrelated cancer types like pancreatic tumors (data not shown). Very likely, the majority of genes do not contribute to biological variation and serve as a buffer when included in such analyses. Furthermore, if subsets of genes expected to be more highly expressed in a tissue specific manner were chosen for cross-dataset correlations using cohort medians, results were disappointing. By randomly modeling the cohort makeups, we concluded that this likely stems from the high variability introduced to cohort medians when analyzing small sample sets.

We theorized a more robust approach for examining biological fidelity of RNA-Seq datasets would be to look for known or expected patterns of gene co-expression, considering both positive and negative correlations. Clinically, OPSCCs can be divided by the presence or absence of high-risk human papillomavirus (HPV) as a cancer driver, which is reflected in very different biology and treatment outcomes.[47-52] The incidence of HPV-associated OPSCC has been steadily increasing in the last few decades and as of 2021, ~75% of cancers occurring in this anatomical subsite are associated with HPV.[6, 7] We previously identified that 51 out of 78 TCGA OPSCC were HPV-associated based on re-analyzing publicly available

whole exome sequencing data, and used this to define genes differentially expressed between HPV-associated and HPV-independent OPSCC after normalizing gene expression for the TCGA cohort (**Supplementary Table 6**). Next, we selected genes that showed ≥ 3-fold significant (FDR <0.1) difference up or downregulation based on HPV status in the OPSCC TCGA cohort (**Supplementary Fig. 7A**, **Supplementary Table 6**) and filtered out immune genes to focus on differences in tumor biology, leaving 855 differentially regulated genes. Theoretically, these genes should show a similar pattern of expression in our cohorts, based on the assumption that both HPV-associated and HPV-independent tumors are expected. In reality, some amount of data noise is expected in our datasets due to platform differences or smaller cohort sizes. Therefore, we examined the co-correlation of these 855 differentially expressed genes in our largest FFPE cohort, the TruSeq library data set that included 20 samples.

Co-correlation coefficients from the 855 differentially expressed genes calculated using the TruSeq cohort are listed in **Supplementary Table 7** and hierarchical clustering was used to identify modules of genes with patterns of co-correlation or anti-correlation (**Supplementary Fig. 8**), with directionality of fold change expected from the TCGA cohort annotated. It is worth noting that genes with increased or decreased expression in the TCGA cohort according to HPV status were clearly segregating in the TruSeq FFPE cohort (**Supplementary Fig. 8**), which is already indicative that biology was being strongly preserved. We selected genes in clusters 1 and 3 (collectively 598 out of 855 genes) to see if they could predict HPV status of our FFPE specimens. Because we do not have the HPV status of the samples used in our study and the p16 (CDKN2A) immunohistochemistry (IHC) sometimes used as surrogate was not available for all samples, we relied on CDKN2A mRNA levels. Overexpression of CDK2NA RNA, which also drives p16 protein detection by IHC, results from a negative feedback loop in HPV-associated cancers with diminished Rb and highly active E2F transcription factor. Using the OPSCC TCGA cohort we demonstrated that HPV-associated tumors abundantly overexpress CDKN2A RNA with an average of 34-fold elevation (**Supplementary Table 6**) that is highly significant ($p < 0.0001$, **Supplementary Fig. 7B**).

We used the 598 DEGs to cluster samples from all 3 of our FFPE cohorts (**Supplementary Tables 8-10**) to predict their HPV status (**Fig. 3-4, Supplementary Fig. 9**) based on the patterns of their gene expression. In the TruSeq cohort (**Fig. 3**), two main gene clusters were found which were enriched or depleted for genes previously determined to be differentially expressed in the TCGA cohort based on HPV status. The segregation of these genes was highly significant ($p<0.00001$) indicating strong preservation of tumor biology. This was additionally validated by significant elevation of CDKN2A RNA among the FFPE specimens predicted to be HPV-associated ($p<0.005$, **Fig. 3B**), despite the fact that CDKN2A had been purposely removed from the list of genes for clustering. Highly similar results were found for the Stranded/Ribo-zero and Mace-Seq datasets, regarding segregation of the HPV-signature genes, although genes associated with HPV-independent tumors formed two subclusters. Predicted differences in CDKN2A RNA similarly validated the HPV status predictions, albeit the p-values were not as robust because of the smaller sample numbers in these cohorts.

Assessing Gene Dropout. The phenomena whereby low sequencing depth randomly leads to genes not being represented is well described for single cell RNA-Seq experiments.[53, 54] We wondered whether gene dropout might also occur in FFPE samples for different biological or technical reasons, such as selective RNA instability. Put in mathematical terms, gene dropout means observing more samples with zero gene expression than expected by chance. We modeled the expected probability of observing samples with zero expression individually for every gene using the existing OPSCC TCGA cohort (N= 71) as a gold standard because the dataset was comparatively large and derived using frozen tissue. Because our cohort of samples were predicted to contain roughly 50% HPV-associated tumors, we used an average of the observed number of zero reads between HPV-associated and HPV-independent sample in the TCGA (**Supplementary Table 11**) to derive an estimate for each gene of the probability of getting zero reads. For every gene (excluding variable immunoglobulin or T cell receptor genes), the binomial probability

Figure 4

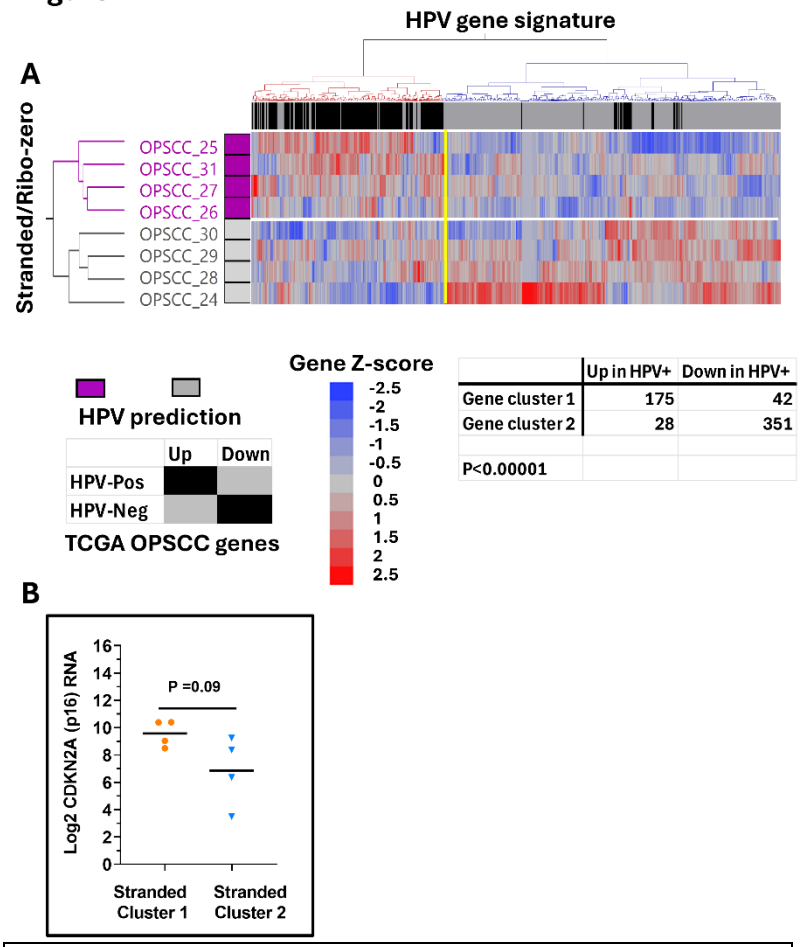

Figure 4. Correlation of HPV associated genes in the Stranded Cohort. Expression of HPV-associated DEGs (identified from the TCGA OPSCC reference cohort) was used for unsupervised 2-way clustering of Stranded cohort samples (i.e., Ward's agglomerative hierarchical clustering) to predict HPV status (A). Samples predicted to be HPV-associated (i.e., HPV-Pos) are annotated with purple boxes. The regulation status of genes is annotated across the top of the heatmap with black boxes if they were also upregulated in HPV-associated TCGA samples or grey boxes if they were upregulated in HPV-independent (i.e., HPV-Neg) TCGA samples. Significant enrichment of genes upregulated in TCGA HPV-associated cancers was found in gene cluster 1 (red cluster) and enrichment of genes upregulated in HPV-independent TCGA samples is found in gene cluster 2 (blue cluster), demonstrating significant separation (e.g., P <0.00001 by Chi-square testing). Specimens from sample cluster 1 (purple) predicted to be HPV-associated based on their gene expression pattern had higher CDKN2A (p16) expression than specimens from sample cluster 2 (grey) predicted to be HPV-independent (B), which nearly reached significance.

distribution was applied to approximate the individual probability (p-value) and adjusted p-value (for multiple testing) of observing at least as many samples found in our cohorts with zero counts. The gene-wise adjusted p-values for all 3 of our RNA-seq cohorts are provided in the supplementary tables (**Supplementary Table 12-14).** To compare gene dropout between sequencing platforms, we plotted the frequency distribution of genes verses the log Adj p-values (**Fig. 5A**). Counting the number of genes with adjusted p-values <0.01, dropout was highest in the TruSeq cohort (N= 3,100), lowest for the Stranded/Ribo-zero dataset (N = 583), and in between with the MACE-Seq samples (N= 2,181). A Venn diagram identifying overlap of gene dropout is shown in **Fig. 5B**.

To better understand the study limitations or biological bias potentially introduced from gene dropout we performed a Gene Ontology Enrichment Analysis on the 283 common genes , and then for each cohort's dropout after removing these common genes. The 283 genes commonly missing from all cohorts were enriched for genes related to sexual reproduction and actually depleted for genes involved in DNA damage or metabolism (**Supplementary Table 15**). Implicit from the analysis of common gene dropout, genes associated with metabolism or nucleic acid regulation (e.g., synthesis, repair) were also statistically missing or not among the lists of gene dropout for all three sequencing platforms (**Supplementary Table 16-18**). Genes with higher-than-expected

loss from both the TruSeq and MACE-Seq cohorts were enriched for neuro-regulation pathways (**Supplementary Tables 16-17**), which might be attributed to a higher inclusion of adjacent nerve cells in the TCGA cohort—possibly due to larger average size of specimens in the latter. Although far fewer in number, genes with dropout from the Stranded/Ribo-zero dataset were enriched for keratinization, which could be attributed to differences in the cohort makeup given the small sample size and highl variability of differentiation markers naturally present among OPSCC tumors. To provide insight into the differences in gene dropout between samples prepared using the Stranded/Riob-zero kit, we plotted the cohort medians for genes with dropout in the TruSeq or MACE-seq dataset against their medians in the Stranded/Ribo-zero dataset (**Fig. 5C**). Genes that dropped out had a bias towards lower expression in the Stranded/Ribo-zero dataset and this was most prominent when examining gene dropout in the TruSeq cohort.

Figure 5

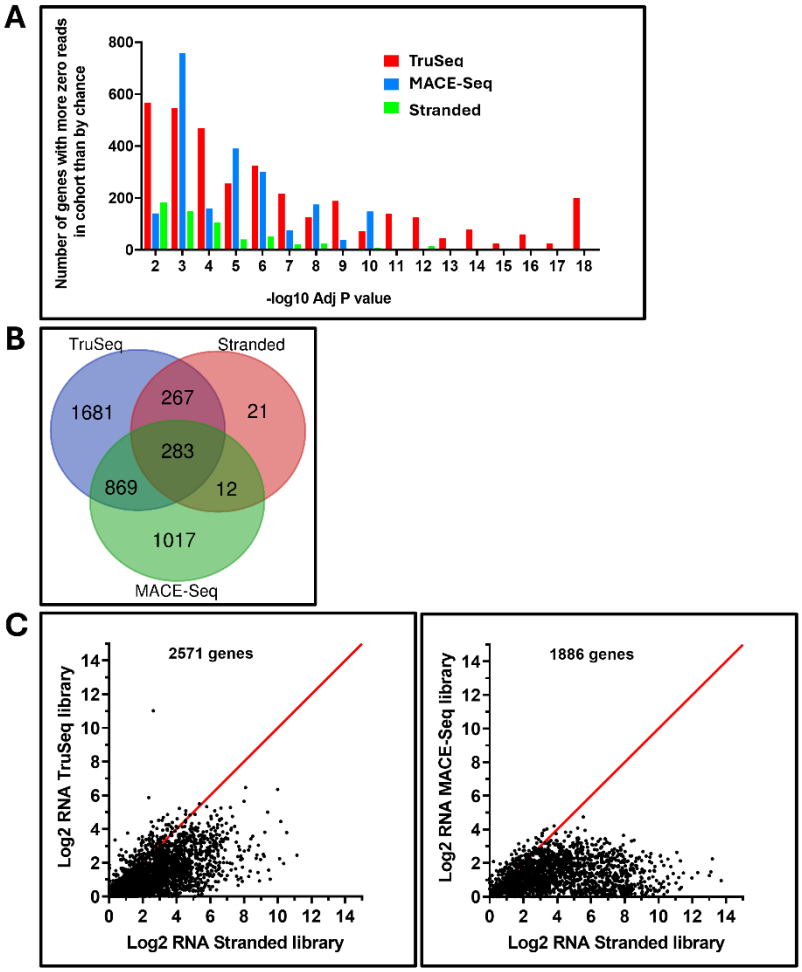

Figure 5. Gene dropout. Frequency distribution for number of genes verses the -log 10 Adj p-values derived from binomial probability calculations (A) demonstrated dropout was highest in the TruSeq cohort (N= 3,100), lowest for the Stranded/Ribo-zero dataset (N = 583), and in between with the MACE-Seq samples (N= 2,181); overlapping dropout genes (B). C) Cohort medians for genes with dropout in the TruSeq or MACE-seq dataset plotted against their medians in the Stranded/Ribo-zero dataset demonstrate a bias towards dropout for low expression genes.

Co-extraction of DNA and SNP profiling patient ancestry. We analyzed DNA pellets left over during RNA extraction from a subset of samples using a modified COBAS protocol for ancestry using the Illumina Global Diversity SNP array. Our average, DNA yield was 245 ng or 24 ng/cm^2 of tissue (**Supplementary Table 19**), which was sufficient for the SNP platform where 200 ng is the recommended amount. For comparison we isolated DNA from additional specimens that were not processed for RNA and ran them on the same SNP array. We found that the leftover DNA extracted during the RNA isolations performed satisfactorily but had slightly diminished SNP call rates which averaged 83.5% compared to 92.0% for samples where only DNA was isolated (P =0.05, **Fig. 6A**). In all cases, we were able to use the SNP data generated to predict ancestry (**Supplementary Table 20**) using software freely available (**Supplementary Methods**). Using the K18M model of SNPS, we identified a variety of ethnic groups at various percentages, but 9 out of 14 (e.g., 64%) were of Atlantic/Northern European

descent, which was consistent with patient demographics for OPSCC. Two patients were found to be African American, and another was of Western Mediterranean ancestry. In samples from patients #1 (OPSCC_03/OPSCC_26) and patient # 8 (e.g., OPSCC_24), we isolated DNA directly or after RNA purification and compared their ancestry prediction. For both patients, (**Fig. 6B**, **6C**), the predicted ancestries matched accordingly, with one showing strong descent from Equatorial Africa and the other a mixed ancestry predominately of Atlantic/Northern European origin.

Figure 6

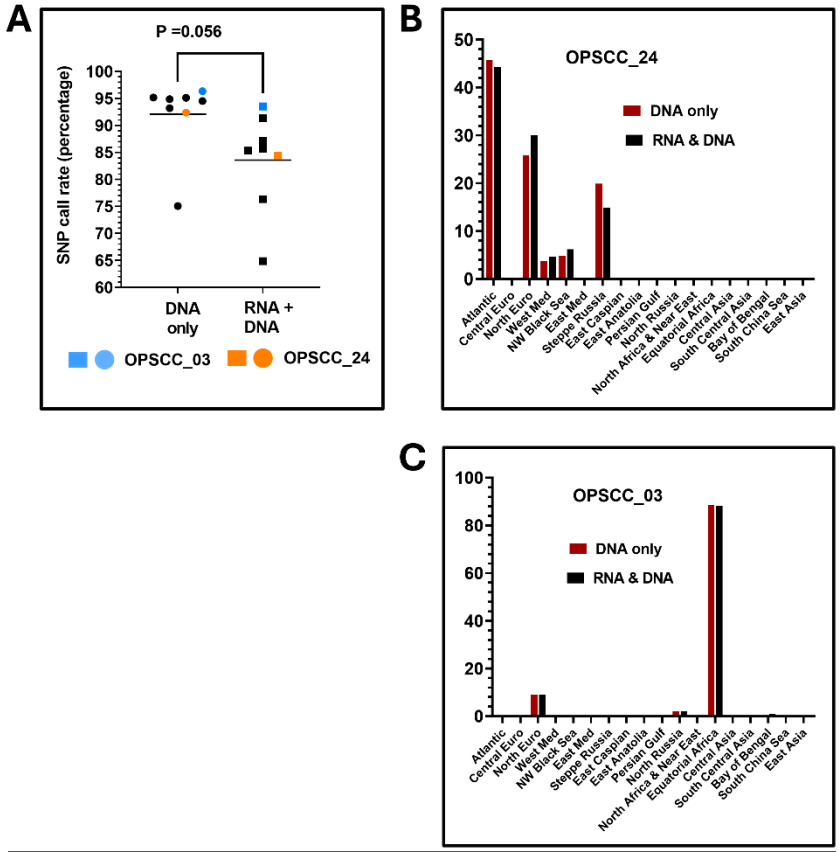

Figure 6. Ancestry measurements. A) SNP call rate was calculated for each DNA extraction approach for each individual sample. B, C) Comparison of ancestry results from two unique samples where DNA was extracted twice using either the COBAS method alone (red) or from residual pellets following RNA isolation in a combined procedure (black).

DISCUSSION

In the present work we demonstrate the feasibility of utilizing archival FFPE specimens of varying age for whole transcriptome sequencing (i.e., RNA-Seq) to obtain biologically meaningful data. Further, we provide technical details for: 1) isolating both RNA and DNA, 2) insight into the most meaningful quality control metrics, 3) methods for standardizing data analysis, and 4) suggestions for data quality control increasing confidence in the data. Lastly, we show that residual DNA recovered from specimens during RNA isolation is of sufficient quality for SNP array platforms, which can be used to confirm ancestry or racial identity of study participants. We recommend the AllPrep DNA/RNA FFPE kit to isolate RNA because in our hands it performed robustly with respect to yield for nearly 200 specimens and led to high quality downstream RNA-seq reads using three different library prep methods. On average, a total of 18cm^2 of tissue (e.g., less than 10 slides) cut at standard thickness (6 µM) generated more than sufficient RNA for the different sequencing options available. If RNA yields are less than 246 ng/cm^2 (**Supplementary Table 2**, lower 90% confidence interval of mean) based on A260 measurements, then appropriate trouble shooting steps are warranted. FFPE RNA is known to be fragmented, so typical metrics like RIN are not appropriate. DV200 measurements should be interpreted cautiously, because it is still possible to generate reliable RNA-Seq data from samples with a DV200 < 20%. Rather the shape of the TapeStation curve closer to 100 bp may be a better indicator.

After filtering out non-protein coding genes (e.g., **Supplementary Table 4**) our quality control process consists of five main steps easily applied, followed by some type of analysis for biological fidelity

that by necessity will be context specific. We have incorporated an UQ normalization into our ready to use Python code (**Supplementary Methods**), but other methods could work as well. The percentage of mapped reads belonging to protein coding genes should be examined for each sample with the caveat that stranded library approaches may lower the values because of the additional requirement that captured fragmented DNA also be from a specific strand. Our method of detecting technical outliers based on MAD residuals works differently than principal component analysis (PCA) and should be applied upstream of the latter.[55] Unlike PCA, this method makes use of information from every gene rather than a few thousand most variable genes. Therefore, the MAD method may have better sensitivity for detecting just a single outlier in larger cohorts where the most variable genes selected for PCA may be driven largely by biological variability. In fact, it is probably more consistent than PCA when it comes to separating technical rather than biologic variation. However, the method is not a substitute for PCA that could be applied afterwards to resolve differences based on biologic variation that may reveal further heterogeneity to be considered when comparing gene expression.

The number of protein coding genes with log2 RNA expression ≥ 2 should be calculated only after re-adjusting the global median to equal 7. This re-centering allows the same standard threshold for low gene expression to be more consistently applied, regardless of sequencing depth, batch affects, or platforms used. By way of comparison, the OPSCC TCGA cohort samples average 14,565 genes (Range:13,383 to 15,389) with log2 expression ≥ 2. Typically, samples with too few protein coding genes above the low expression threshold are easily identified as technical outliers. Using the MAD residuals method avoids having to set a precise cut-off value *for number of genes* above the threshold to consider acceptable, provided most samples in the cohort express reasonable numbers of genes. For our datasets, excluding outliers, the average number of genes above the threshold of 2 was ~14,000. It is useful to compare the distribution (i.e., histogram) of normalized log2 RNA expression values to a standard like the TCGA data, which follows a quasi-log normal distribution. In our experience, the re-centered RNA-seq data from the TCGA has a similar appearance regardless of the cancer chosen. In the last simple QC step, we recommend identifying genes with an abnormal degree of sample dropout, if the expected frequency of samples with zero expression can be appropriately modeled with an orthogonal data set. We observed the least amount of gene dropout in the cohort prepared using the Stranded library kit, however, it is likely that the average quality of RNA was better for these specimens judging by the TapeStation curves. Regardless, for our datasets the genes lost seemed unrelated to tumor biology and likely represented low expression genes. Genes related to metabolism or nucleic acid processing were depleted from the lists of dropout and therefore reasonably represented.

If possible, the biological fidelity of new data should be evaluated independent of the scientific questions being explored. For our data set, we leveraged the fact that a substantial subset of OPSCCs are linked to high risk HPV, which leads to programs of genes that are both up and downregulated. We then used hierarchical clustering to examine if these genes defined by an independent cohort (i.e., TCGA data) behaved similarly in our datasets. This same approach could be applied using just about any set of genes from pathways known to be co-regulated in the tissue under study. Co-regulation or anti-correlations among gene clusters can be examined statistically using simple Chi-square tests. Lastly, we showed that SNP data from residual DNA, co-isolated during RNA purification, is of sufficient quality for SNP arrays and can be expected to have call rates that average above 80%, which may be sufficient for many applications and gives comparable results to DNA directly isolated from FFPE. Our data, methods, and recommended analytical approaches should provide investigators with helpful guidelines and tools to better leverage archival specimens to answer important biological questions.

Methods

All specimen collection, retrieval and analysis performed followed approval from Baylor College of Medicine and Michael E. DeBakey Veterans Affairs Medical Center (MEDVAMC) Institutional Review

Boards. Basic steps in our RNA data processing pipeline include: 1) Filtering out non-protein coding genes; 2) Excluding zero count data to calculate a 75th percentile (UQ) read value for each sample; 3) Calculating the cohort median count upper quartile (MUQ) value; 4) Replacing any gene sizes ≤ 252 bp with a minimum value of 252 (i.e., size of smallest known human gene) to prevent data distortion or a zero division error; 4) Calculating the global cohort gene size median (GMGS); 5) Normalizing each count value through dividing by the sample specific 75th quartile and the individual gene size (either constant or sample specific); 6) Multiplying the quotient by the product of the MUQ x GMGS to bring numbers back to the scale of counts; 7) Performing a log2 transformation after adding back 0.01 to every value (Log2 [X + 0.01]); 8) Identifying and removing "technical" outlier samples using the average median absolute deviation (MAD) approach coupled with a downstream statistical method to detect MAD outliers; 9) Re-scaling the data so the distribution has a global median set to 7.0; and 10) Replacing negative log2 values with a zero, to avoid the pitfalls of artificially inflating standard deviations and fold changes associated with very low expression values. Importantly, because the data are rescaled to median of 7 before replacing low expression values, zero replacements are handled consistently across data with different sequencing depth. Furthermore, it becomes feasible to select a single low expression threshold cut-off (e.g., Log2 ≤ 2) to suffice for every experiment across time and different sequencing depths. **Supplementary Methods** include the specific isolation protocols and coding used in the analysis, which incorporates the above steps.

References

1. Yu, W., et al., *Evolution of cisplatin resistance through coordinated metabolic reprogramming of the cellular reductive state.* Br J Cancer, 2023. **128**(11): p. 2013-2024.
2. Kazi, M.A., et al., *Glutathione peroxidase 2 is a metabolic driver of the tumor immune microenvironment and immune checkpoint inhibitor response.* Journal for ImmunoTherapy of Cancer, 2022.
3. Frederick, M., et al., *High expression of oxidative phosphorylation genes predicts improved survival in squamous cell carcinomas of the head and neck and lung.* Sci Rep, 2020. **10**(1): p. 6380.
4. Yu, W., et al., *Cisplatin generates oxidative stress which is accompanied by rapid shifts in central carbon metabolism.* Sci Rep, 2018. **8**(1): p. 4306.
5. Jacobsen, S.B., et al., *Comparison of whole transcriptome sequencing of fresh, frozen, and formalin-fixed, paraffin-embedded cardiac tissue.* PLoS One, 2023. **18**(3): p. e0283159.
6. Wilde, D.C., et al., *Oropharyngeal cancer outcomes correlate with p16 status, multinucleation and immune infiltration.* Mod Pathol, 2022. **35**(8): p. 1045-1054.
7. Elhalawani, H., et al., *Tobacco exposure as a major modifier of oncologic outcomes in human papillomavirus (HPV) associated oropharyngeal squamous cell carcinoma.* BMC Cancer, 2020. **20**(1): p. 912.
8. Gao, X.H., et al., *Comparison of Fresh Frozen Tissue With Formalin-Fixed Paraffin-Embedded Tissue for Mutation Analysis Using a Multi-Gene Panel in Patients With Colorectal Cancer.* Front Oncol, 2020. **10**: p. 310.
9. Zhao, Y., et al., *Robustness of RNA sequencing on older formalin-fixed paraffin-embedded tissue from high-grade ovarian serous adenocarcinomas.* PLoS One, 2019. **14**(5): p. e0216050.
10. Choi, Y., et al., *Optimization of RNA Extraction from Formalin-Fixed Paraffin-Embedded Blocks for Targeted Next-Generation Sequencing.* J Breast Cancer, 2017. **20**(4): p. 393-399.
11. Cannizzo, M.D., et al., *Case study: Targeted RNA-sequencing of aged formalin-fixed paraffin-embedded samples for understanding chemical mode of action.* Toxicol Rep, 2022. **9**: p. 883-894.
12. Matsunaga, H., et al., *Reproducible and sensitive micro-tissue RNA sequencing from formalin-fixed paraffin-embedded tissues for spatial gene expression analysis.* Sci Rep, 2022. **12**(1): p. 19511.

13. Eikrem, O., 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): p. e0149743.
14. Marczyk, M., et al., *The impact of RNA extraction method on accurate RNA sequencing from formalin-fixed paraffin-embedded tissues.* BMC Cancer, 2019. **19**(1): p. 1189.
15. Liu, Y., et al., *Immunogenetic Determinants of Susceptibility to Head and Neck Cancer in the Million Veteran Program Cohort.* Cancer Res, 2022.
16. Deshmukh, S.K., et al., *Biological basis of cancer health disparities: resources and challenges for research.* Am J Cancer Res, 2017. **7**(1): p. 1-12.
17. Neagu, A.N., et al., *Biological Basis of Breast Cancer-Related Disparities in Precision Oncology Era.* Int J Mol Sci, 2024. **25**(7).
18. Bachtiar, M., et al., *Architecture of population-differentiated polymorphisms in the human genome.* PLoS One, 2019. **14**(10): p. e0224089.
19. Polubriaginof, F.C.G., et al., *Challenges with quality of race and ethnicity data in observational databases.* J Am Med Inform Assoc, 2019. **26**(8-9): p. 730-736.
20. Plichta, J.K., et al., *Implications of missing data on reported breast cancer mortality.* Breast Cancer Res Treat, 2023. **197**(1): p. 177-187.
21. Patel, P.G., et al., *Reliability and performance of commercial RNA and DNA extraction kits for FFPE tissue cores.* PLoS One, 2017. **12**(6): p. e0179732.
22. Kresse, S.H., et al., *Evaluation of commercial DNA and RNA extraction methods for high-throughput sequencing of FFPE samples.* PLoS One, 2018. **13**(5): p. e0197456.
23. Sarnecka, A.K., et al., *DNA extraction from FFPE tissue samples - a comparison of three procedures.* Contemp Oncol (Pozn), 2019. **23**(1): p. 52-58.
24. Ondracek, R.P., et al., *Results and lessons from dual extraction of DNA and RNA from formalin-fixed paraffin-embedded breast tumor tissues for a large Cancer epidemiologic study.* BMC Genomics, 2022. **23**(1): p. 614.
25. Atanesyan, L., et al., *Optimal Fixation Conditions and DNA Extraction Methods for MLPA Analysis on FFPE Tissue-Derived DNA.* Am J Clin Pathol, 2017. **147**(1): p. 60-68.
26. Boneva, S., et al., *3' MACE RNA-sequencing allows for transcriptome profiling in human tissue samples after long-term storage.* Lab Invest, 2020. **100**(10): p. 1345-1355.
27. Jang, J.S., et al., *Application of the 3' mRNA-Seq using unique molecular identifiers in highly degraded RNA derived from formalin-fixed, paraffin-embedded tissue.* BMC Genomics, 2021. **22**(1): p. 759.
28. Koch, C.M., et al., *A Beginner's Guide to Analysis of RNA Sequencing Data.* Am J Respir Cell Mol Biol, 2018. **59**(2): p. 145-157.
29. Tan, Z., et al., *Comprehensive analysis of scRNA-Seq and bulk RNA-Seq reveals dynamic changes in the tumor immune microenvironment of bladder cancer and establishes a prognostic model.* J Transl Med, 2023. **21**(1): p. 223.
30. Zoabi, Y. and N. Shomron, *Processing and Analysis of RNA-seq Data from Public Resources.* Methods Mol Biol, 2021. **2243**: p. 81-94.
31. Deshpande, D., et al., *RNA-seq data science: From raw data to effective interpretation.* Front Genet, 2023. **14**: p. 997383.
32. Lee, G.Y., S. Ham, and S.V. Lee, *Brief guide to RNA sequencing analysis for nonexperts in bioinformatics.* Mol Cells, 2024. **47**(5): p. 100060.
33. Chen, A.A., et al., *Privacy-preserving harmonization via distributed ComBat.* Neuroimage, 2022. **248**: p. 118822.
34. Zhang, Y., G. Parmigiani, and W.E. Johnson, *ComBat-seq: batch effect adjustment for RNA-seq count data.* NAR Genom Bioinform, 2020. **2**(3): p. lqaa078.

35. Lehrer, S. and P.H. Rheinstein, *EARS2 significantly coexpresses with PALB2 in breast and pancreatic cancer.* Cancer Treat Res Commun, 2022. **32**: p. 100595.
36. Padegal, G., et al., *Analysis of RNA-Seq data using self-supervised learning for vital status prediction of colorectal cancer patients.* BMC Bioinformatics, 2023. **24**(1): p. 241.
37. Molania, R., et al., *Removing unwanted variation from large-scale RNA sequencing data with PRPS.* Nat Biotechnol, 2023. **41**(1): p. 82-95.
38. Hanzelmann, S., R. Castelo, and J. Guinney, *GSVA: gene set variation analysis for microarray and RNA-seq data.* BMC Bioinformatics, 2013. **14**: p. 7.
39. Evans, C., J. Hardin, and D.M. Stoebel, *Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions.* Brief Bioinform, 2018. **19**(5): p. 776-792.
40. Xia, Y., *Statistical normalization methods in microbiome data with application to microbiome cancer research.* Gut Microbes, 2023. **15**(2): p. 2244139.
41. Zhao, Y., et al., *TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository.* J Transl Med, 2021. **19**(1): p. 269.
42. Ilgisonis, E.V., et al., *Gene-centric coverage of the human liver transcriptome: QPCR, Illumina, and Oxford Nanopore RNA-Seq.* Front Mol Biosci, 2022. **9**: p. 944639.
43. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biol, 2014. **15**(12): p. 550.
44. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-40.
45. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic Acids Res, 2015. **43**(7): p. e47.
46. Law, C.W., et al., *voom: Precision weights unlock linear model analysis tools for RNA-seq read counts.* Genome Biol, 2014. **15**(2): p. R29.
47. Dahlstrom, K.R., et al., *An evolution in demographics, treatment, and outcomes of oropharyngeal cancer at a major cancer center: a staging system in need of repair.* Cancer, 2013. **119**(1): p. 81-9.
48. Fakhry, C., et al., *Association of Oral Human Papillomavirus DNA Persistence With Cancer Progression After Primary Treatment for Oral Cavity and Oropharyngeal Squamous Cell Carcinoma.* JAMA Oncol, 2019. **5**(7): p. 985-992.
49. Fakhry, C., et al., *Validation of NRG oncology/RTOG-0129 risk groups for HPV-positive and HPV-negative oropharyngeal squamous cell cancer: Implications for risk-based therapeutic intensity trials.* Cancer, 2019. **125**(12): p. 2027-2038.
50. Gleber-Netto, F.O., et al., *Variations in HPV function are associated with survival in squamous cell carcinoma.* JCI Insight, 2019. **4**(1).
51. Masterson, L., et al., *CD8(+) T cell response to human papillomavirus 16 E7 is able to predict survival outcome in oropharyngeal cancer.* Eur J Cancer, 2016. **67**: p. 141-151.
52. Welters, M.J.P., et al., *Intratumoral HPV16-Specific T Cells Constitute a Type I-Oriented Tumor Microenvironment to Improve Survival in HPV16-Driven Oropharyngeal Cancer.* Clin Cancer Res, 2018. **24**(3): p. 634-647.
53. Saliba, A.E., et al., *Single-cell RNA-seq: advances and future challenges.* Nucleic Acids Res, 2014. **42**(14): p. 8845-60.
54. Hwang, B., J.H. Lee, and D. Bang, *Single-cell RNA sequencing technologies and bioinformatics pipelines.* Exp Mol Med, 2018. **50**(8): p. 1-14.
55. Chen, X., et al., *Robust principal component analysis for accurate outlier sample detection in RNA-Seq data.* BMC Bioinformatics, 2020. **21**(1): p. 269.