Title

Dissociation protocols used for sarcoma tissues bias the transcriptome observed in single-cell and single-nucleus RNA sequencing

Danh D. Truong, Salah-Eddine Lamhamedi-Cherradi, Robert W. Porter, Sandhya Krishnan, Jyothishmathi Swaminathan, Amber L. Gibson, Alexander J. Lazar, John A. Livingston, Vidya Gopalakrishnan, Nancy Gordon, Najat C. Daw, Richard Gorlick, and Joseph A. Ludwig.

Keywords

RNA-seq, scRNA-seq, snRNA-seq, single-cell, Ewing's sarcoma, desmoplastic small round cell tumor, osteosarcoma

Abstract

Background

Single-cell RNA-seq has emerged as an innovative technology used to study complex tissues and characterize cell types, states, and lineages at a single-cell level. Rapid adoption of this technology has led to a flurry of research creating single-cell atlases for many organs, cancers, and developmental models. Despite the tremendous success of this technology, sarcomas, rare cancers of mesenchymal origin, have not yet widely benefited from the adoption of single-cell RNA-seq. Due to their rarity, obtaining fresh clinical specimens is challenging. Single-nucleus RNA-seq could remove limitations from obtaining fresh tissue and enable immediate processing of archival tissue. However, biases may exist during dissociation of either fresh or frozen specimens that can hinder reproducible results and introduce technical or biological artifacts.

Results

To address these questions, we systematically assessed dissociation methods across different sarcoma subtypes. Here, we compare gene expression from single-cell and single-nucleus RNA-sequencing of 125,831 whole-cells and nuclei from Ewing's sarcoma, desmoplastic small round cell tumor, and osteosarcoma patient-derived xenografts. We detected warm dissociation artifacts from single-cell samples and gene length bias in single-nucleus samples. Classic sarcoma gene signatures were observed regardless of dissociation method. Finally, we show that dissociation method biases can be computationally corrected.

Conclusions

We highlighted transcriptional biases, including warm dissociation and gene-length biases, introduced by any dissociation method for various sarcoma subtypes. This work will inform researchers on choice of dissociation method and careful interpretation of data due to dissociation biases.

Background

Tumors are composed of a diverse multicellular microenvironment that dictate cancer progression and response to therapy. While cells share an identical genome, their phenotype and behavior are driven by their transcriptome[1]. Cellular heterogeneity within the tumor ecosystem has precluded the ability to fully understand the cell biology and interactions that drive cancer progression[1]. Recently, single-cell RNA-seq (scRNA-seq) has emerged as an innovative technology used to break down complex tissues into individual cells in order to characterize cell types, states, and lineages[2]. Rapid adoption of this technology has led to a flurry of research generating single-cell atlases for many organs, cancers, and developmental models pushing the boundaries of cell biology to the next level[3].

Despite the tremendous success of this technology when applied to different cancer types, sarcomas, which are cancers of mesenchymal origin, have not yet widely benefited from the adoption of scRNA-seq. Differences in tissue origin may require optimized dissociation to capture accurate *in vivo* gene expression and cellular composition. Recent work on different methods of tissue dissociation, be it enzymatic or mechanical, have been shown to bias cellular composition and reduce cellular quality. Many gold standard dissociation protocols require extended incubation at 37 °C, where cellular transcription is still active and may introduce gene expression artifacts[4]. Cold-active protease is a recent alternative to dissociation at 37 °C, which may limit and minimize transcriptional activity and environmental stresses on cells[4,5].

Challenges in obtaining fresh clinical specimens and the lack of ability to immediately process specimens have also hindered workflows[6]. While cancer models for sarcoma, including cell lines, xenografts, and PDXs, are readily accessible for scRNA-seq, the extent that these models represent the original cancer specimen have not yet been adequately evaluated. Single-nucleus RNA-seq (snRNA-seq) of accessible frozen tissue has demonstrated concordance with scRNA-seq[6-10]. SnRNA-seq can remove limitations from obtaining fresh tissue and immediate processing by enabling access to archival tissue and ease the coordination of tissue acquisition by allowing sequencing of snap-frozen tissue. Furthermore, difficulties with cell fragility or size when considering scRNA-seq can be circumvented using snRNA-seq.

The biases introduced by different methods have been studied between single-cell and single-nucleus as well as dissociation using cold-active proteases and standard digestion at 37 °C[4]. However, these studies did not include sarcoma specimens, which may not replicate the findings observed in studies using carcinomas. Importantly, to fully realize the potential of scRNA-seq and snRNA-seq, further analysis would be required on sarcoma samples. To address these challenges, we systematically assessed dissociation methods across different sarcoma subtypes. We compared tissue dissociation methods for single-cell and single-nucleus workflows. We focused our analysis on well-controlled PDX specimens of different and rare sarcomas to enable accessibility to samples since fresh sarcoma specimens are difficult to acquire. This further enabled our group to explore multiple dissociation methods on the same sample.

Overall, we envision our work will provide both a technical and analytical framework for scRNA-seq and snRNA-seq for sarcoma samples and to an extent other cancer subtypes as well.

Results

Single-cell and single-nucleus RNA sequencing of various sarcoma subtypes

In this work, we studied sarcomas from varying tissue origins, including osteosarcoma (OS), Ewing's sarcoma (ES), and desmoplastic small round cell tumor (DSRCT). We used different dissociation protocols: Miltenyi Tumor Dissociation Kit, cold-active protease derived from *Bacillus licheniformis*, and Nuclei EZ Prep. These three protocols are described herein as Warm, Cold, and Nuclei protocols. For the OS specimens, we used the same Nuclei protocol and a different Warm protocol optimized for OS. For DSRCT and ES specimens, we performed the additional Cold protocol, using the cold-active protease, as we had more specimens available. Each sarcoma subtype included three PDX specimens derived from different patients. In total, we analyzed 125,831 whole-cells and nuclei across the three sarcoma subtypes and three dissociation protocols.

Evaluation of quality control metrics for tissue dissociation protocols

We evaluated and compared the protocols based on cell/nucleus quality and transcriptomic signatures. For cell/nucleus quality, we measured the percent of reads mapping to the transcriptome, number of genes, unique molecular identifiers, and percent of mitochondrial genes for each cell or nucleus. We tested the protocols on multiple PDXs from the same sarcoma subtype to evaluate for trends and biases that are consistent across different PDXs. Previous work has shown that dying and dead cells can influence the transcriptome and introduce artifacts that preclude useful biological insight[4]. Therefore, we optimized our strategy by incorporating fluorescently activated cell sorting (FACS) or fluorescently activated nuclei sorting (FANS) to enrich for live single cells and single nucleus respectively (**Fig. S1**).

We evaluated common quality control (QC) metrics across all samples to assess the effect of each dissociation protocol (**Fig. 1A**). We observed some variations in QC metrics for the number of genes and unique molecular identifiers when comparing between protocols while limiting the comparisons to between PDXs of each sarcoma subtype. However, some of the variations could be explained by number of cells sequenced and sequencing depth since there is an inverse relationship between these two metrics when total reads are kept constant (**Fig. 1B**). Notably, nucleus samples demonstrated little to no percentage of mitochondrial genes. This is as expected since high quality nucleus would not contain mitochondrial transcripts. With respect to each protocol, we did not discern a positive or negative influence on the QC metrics.

Dissociation protocol biases the transcriptome

Given that we utilized different protocols, we asked if we would observe larger differences between PDXs and minor differences between protocols of the same PDX specimen. To explore this question, we visualized the UMAP embeddings of all whole-cells and nuclei without batch or technical corrections. When labeled by fresh specimens (whole-cells) or frozen specimens (nuclei), It is clear that there is a distinct delineation between fresh and frozen (**Fig. 2A**). We suspect that there is either a biological artifact from fresh tissue dissociation or a set of genes that is enriched only in the nucleus that separated whole-cells and nuclei within the UMAP embedding. When the UMAP embedding was colored by dissociation type, we can clearly observe that Warm and Cold overlap when they come from the same PDX but the Nuclei clusters

further away. When colored by sarcoma subtype, the same sarcoma subtype still cluster together but there are two distinct clusters based on nuclei and whole-cells for each sarcoma subtype except for OS. OS appeared to cluster together regardless of dissociation type. Recent works have shown that the enzymatic dissociation of fresh specimens at 37 °C from normal and carcinoma tissues induced a 'Warm Dissociation Signature', which included genes such as, FOS, FOSB, and JUN. To investigate this, we selected some of the genes within the Warm Dissociation signature and compared their expression. Averaged gene expression from each sarcoma subtype showed that these genes are elevated in the Warm protocol as opposed to the others (**Fig. 2B**). Furthermore, since prior literature have stated that long non-coding RNAs (lncRNAs) are localized to the nucleus, we also explored their gene expression in these specimens[11]. Consistent with previous results, we observed that lncRNAs were elevated in the Nuclei protocol. Together, these results illustrated that there are differences in the gene expression data due to the choice of dissociation method.

SnRNA-seq could enable access to a host of archival fresh-frozen tissues. Since Warm is the conventional method for dissociation, we asked if there were enriched differentially expressed genes (DEGs) that could bias either protocol. The DEG analysis between Warm and Nuclei protocols demonstrated a consistent trend shared for each sarcoma type. We observed that genes with the largest fold-change in the Warm method included mitochondrial and ribosomal protein genes as expected (**Fig. 3, Table S1-3**). On the other hand, genes enriched in the Nuclei protocol did not have a clear consensus or overlap between sarcoma types. For the Warm protocol, we found a common set of 325 genes enriched after filtering for log2 fold change over 1.5. Similarly, we found 117 genes enriched in the Nuclei protocol (**Fig. S2A, B**). We performed a pathway enrichment of the MSigDB hallmark gene set using Enrichr. We observed stress-associated pathways in each sarcoma type that was enriched in the Warm protocol including Hypoxia, Apoptosis, DNA repair, and TNF-alpha Signaling via NF-kB, which is consistent with prior work. On the other hand, we observed for the Nuclei protocol enrichment in Mitotic Spindle. When comparing the Warm and Cold protocols for only ES and DSRCT, we again observed an increase in several of the commonly identified stress-related pathways like previous results (**Fig. S3, Table S4, 5**). The UMAP embedding suggested that the differences in Warm and Cold are minimal due to the two data sets overlapping when accounting for each PDX (**Fig. 2A**).

Furthermore, we found in total that there were 24 commonly enriched pathways suggesting a core set of conserved genes enriched in the Warm protocol for sarcoma samples **Fig. S2C**. Interestingly, we did not observe any common pathways enriched between sarcomas for the Nuclei protocol. However, further analysis did reveal that genes enriched in the Nuclei protocol are coded by transcripts longer in length compared to those enriched in the Warm protocol. To confirm this, we compared the gene lengths of commonly enriched genes in Warm versus Nuclei protocol for all sarcomas. We found that genes enriched in the Nuclei protocol had significantly longer genes (Wilcoxon test, p-value < 2.2e-16) (**Fig. S2D**). This suggests that there is a possible gene length-associated bias in snRNA-seq. Recent work indicated that hybridization of the polyT RT-primer to intronic polyA stretches of nascent transcripts results in the gene length bias[12]. In fact, our analysis showed that 52% of reads for Nuclei mapped to intronic regions whereas 23% of reads were mapped for Warm protocol (**Fig. S4A**). We binned the genes into quartiles based on the gene length termed as Short (0 – 8077 nt), Short Med. (8077 – 24399 nt), Long. Med. (24399 – 66502 nt), and Long (> 66502 nt). We observed on average 55.6 % of total genes greater

than 66,502 nt (Long) were enriched in the Nuclei protocol compared to 28.2 % and 26.6 % in the Warm and Cold protocols respectively. Interestingly, we also observed an opposite effect in the short genes (0 – 8077 nt) with 20.7 % and 21.4 % of the total genes in Warm and Cold protocols respectively as opposed to 4.1 % in the Nuclei protocol. To answer if there were indeed a bias due to polyA stretches, we counted the number of polyA regions, defined as greater than 15 A repeats, within the full-length cDNA including both intronic and exonic regions for every gene. We observed a significant correlation between increasing gene length and polyA regions (R = 0.7. p-value < 2.2e16). While we saw a positive correlation between scRNA and snRNA expression for each sarcoma type, there is a skew toward higher expression of genes that are longer and with many polyA regions for snRNA data (**Fig. S4B-D).**

Single-nucleus RNA sequencing enriches for genes with long transcripts

To evaluate the enrichment of long transcripts, we generated a length bias score by taking the top 200 genes with the highest number of polyA regions and combining them into a signature gene set (**Table S6**). Note that these genes were chosen purely by length and is agnostic to the underlying biology. In addition, we evaluated the expression of a previously generated warm dissociation signature from O'Flanagan et al.[4] Together, these signatures robustly delineated the biases imparted by scRNA- and snRNA-seq for the three sarcoma types. To further illustrate this, we evaluated logistic regression models using the length bias and warm dissociation signatures to classify affected cells. We randomly split the ES data set into training and test groups. Using the logistic regression model, we could accurately predict samples that underwent the Nuclei protocol (AUC = 1.00) and whole-cells that displayed stress from the Warm protocol (AUC = 0.92) (**Fig. S5).** We applied the same model to the OS and DSRCT data set and observed the same findings (**Fig. S6).** To test if we could extrapolate this classifier to single-cell and single-nucleus libraries processed outside our lab, we used data from a recent paper[6]. In this work, the authors used collagenase 4 at 37 °C to dissociate a neuroblastoma PDX (O-PDX) into single cells and Tween with salts and Tris to dissociate O-PDX into single nuclei[6]. The authors demonstrated that the same dissociation signature we evaluated was elevated in cells when compared to nuclei. Like our data, we applied the length bias and warm dissociation signatures and was able to accurately predict samples that were nuclei (AUC = 1). Importantly, we observed the same findings in a neuroblastoma resection from a patient specimen to rule out if this was limited to only PDX specimens. Our results suggest that single nuclei could readily be classified just by the length bias score and that this may not be limited to just PDX samples (**Fig. S7)**. This implies that snRNA-seq enriches for longer transcript when compared to scRNA-seq from paired samples.

Sarcoma signatures are highly expressed in the respective sarcoma subtypes regardless of dissociation method

Next, to evaluate if any of the protocols influenced signatures associated with a particular sarcoma type, we analyzed expression of gene sets curated from literature (**Table S7**). For ES, we used a set of genes that are direct targets of the EWS-FLI1 fusion protein, which included KDSR, CAV1 and FCGRT[13]. Likewise, a gene set for EWS-WT1 targets, generated from cell lines, was used to evaluate the effect of each protocol in DSRCT[14]. Since OS does not have a clearly defined gene set, we utilized curated genes associated with osteoblastic and chondroblastic signatures classically associated with the tissue origin of OS. Strikingly for each protocol, each of the sarcoma

signature are highly expressed in the respective sarcoma subtypes (**Fig. 5**). This suggests that regardless of dissociation protocol the cells still exhibited the classic signatures for each sarcoma studied. For instance, the EWS-WT1 gene targets are expressed only in the DSRCT PDX specimens. Also, the EWS-FLI1 target genes are only expressed in ES irrespective of which protocol was used. However, this gene set interestingly appears to be overexpressed in the Nuclei protocol for ES. We suspect that this may be due to the gene length-associated bias in snRNA-seq. For each sarcoma signature, we divided the genes into four bins based on quartiles of gene length (**Fig. S8A)**. As we expected, for the EWS-FLI1 target gene set, over 40% of genes in this specific set were considered long (> 66502 nt). This enrichment was not observed in the other gene sets. To further explore the effect of longer genes, we split the EWS-FLI1 target gene set into two groups – short (< 65502 nt) and long (> 65502 nt) genes. Next, we evaluated the resulting expression in ES and indeed observed a gene length-associated bias but only in the group that included long genes (**Fig. S8B)**.

Data integration recovers conserved markers and matching cell-states

Comprehensive analysis will require joint analysis of both scRNA-seq and snRNA-seq as both protocols will provide different and complementary information of the same specimen. As sarcomas are rare and difficult to collect prospectively fresh, having the ability to both sequence fresh-frozen archival tissue and integrating the data with fresh samples is important. However, as demonstrated, there are technical biases imparted by both scRNA-seq and snRNA-seq. As shown in the UMAP embedding, the nucleus samples clearly separate from the cell samples, which suggests that a cluster analysis would not be able to group these two samples even if they are derived from a common tissue specimen. This makes it challenging to reliably perform downstream analyses. To address this challenge, many bioinformatic tools have been developed to remove covariates such as technical batches or different technologies (i.e., CEL-Seq, Fluidigm, 10x Chromium). Utilizing the integration workflow for Seurat v3, we corrected for protocol biases and integrated each dataset of matched PDX specimens[15]. Each sarcoma subtype dataset was split by dissociation method and then integrated. As can be seen after applying integration, technical biases from dissociation protocols no longer strongly affected the clustering and that instead matched PDX specimens with similar cell states clustered together (**Fig. 6A**). Furthermore, expression of the Warm dissociation and length bias scores were clearly demarcating the separation of fresh and frozen specimens, but after integration, these scores were normalized across the UMAP (**Fig. 6B**). When observing conserved markers for OS PDXs, we predicted that both fresh and frozen from the same PDX would exhibit the same DEGs when compared to other PDXs. Based on the figure, we observed that the DEGs are conserved between fresh and frozen specimens (**Fig. 6C**). As we expected, the unintegrated UMAP does not neatly align the conserved markers, which is most likely due to the technical biases that we have shown influencing the algorithm (**Fig. 6D**, left column). Upon integration of the samples, we can clearly observe the conserved DEGs localizing to each PDX (**Fig. 6D**, right column). We extended this same analysis to both ES and DSRCT, and we observed the same effects (**Fig. S9, 10)**. Like batch corrections, this illustrates the need for properly integrating diverse datasets, where in this work paired specimens underwent different dissociation protocols, to reliably perform downstream analyses.

Discussion

In this work, we sought to understand the inherent biases of different dissociation protocols on the transcriptome of various sarcoma types. We avoided logistical difficulties for obtaining extremely rare sarcoma specimens by limiting the scope of the study to PDX sarcomas available to us at our institution. All protocol types generated useful data that are in line with the best practices for single-cell technologies. Furthermore, regardless of protocol type, the data recapitulated the classic transcriptional signatures associated with each sarcoma subtype. However, we observed different technical biases depending on which protocol was utilized, which led us to explore how these technical artifacts may have affected sarcoma biology.

Our work builds upon prior studies analyzing the biases of different protocols used in scRNA-seq and snRNA-seq[4-6,15-17]. Previous works have indicated a stress-response related gene expression set associated with enzymatic digestion at 37 °C[4]. While these works were primarily in carcinomas and normal tissues, our findings for various sarcoma subtypes are in line with their results. Immediate early genes (IEGs), including FOS, JUN, MYC, were highly expressed in the Warm protocol, which were minimized in the Cold protocol and almost absent in the Nuclei dissociation. While these gene expression artifacts are present, we also addressed if there was an effect on classic gene signatures for each sarcoma. While we expected that the stress-response would affect the expression of EWS-FLI1 target genes, when instead, we observed that snRNA-seq had a greater impact due to enriching for genes with longer transcript. This led to EWS-FLI1 target gene set to be overexpressed in snRNA-seq as opposed to scRNA-seq. We removed longer genes from the gene set, which minimized the enrichment of longer transcripts affecting the expression of the overall gene set. However, this may not be ideal as some informative genes may have been removed. As to why the EWS-FLI1 target genes contain an overabundance of long genes, we explored a few possibilities. The EWS-FLI1 transcription factor is known to bind to GGAA microsatellite repeats of 9 or more[18,19]. This may be influenced with transcription length like the increase of polyA region with increasing length. However, many of the microsatellite repeats that enable EWS-FLI1 binding were found within the first intron or the promoter region, which could be located as far as 1 Mb upstream of the transcription start sites[19]. A more likely explanation may be found in the broader analysis of long genes. A review of the effect of gene length found positive correlations with intron number, protein size, and SNPs[20]. Interestingly, gene length is also associated with cancer, heart diseases, and neuronal development[20,21]. Given that a portion of EWS-FLI1 targets are known to be neural genes, we suspect that some of the long genes in the EWS-FLI1 gene set are neural related[22].

Overall, special care must be considered when comparing data between whole-cell and nucleus. In that regard, we did generate a length bias signature using genes with long transcripts, which may potentially be used to regress the technical bias introduced by snRNA-seq. Other have shown that technical biases or batch-to-batch effects can be regressed from snRNA- or scRNA-seq data[16]. Regression of the length bias from the snRNA-seq could produce comparable results to scRNA-seq[12]. However, comparing whole-cell and nucleus transcriptomes between specimens of different tissue origin or disease may need to be interpreted with caution. As noted already, gene length is associated with cancer, heart diseases, and neuronal development and correlated with SNPs[20,21]. On the other hand, our data and others have demonstrated that snRNA-seq data is enriched with lncRNA as compared to scRNA-seq[11]. While this may seem like a confounding variable when trying to compare the two different modalities (i.e., scRNA- and snRNA-seq), it

may be beneficial to utilize snRNA-seq if the intent is to enrich and study lncRNA that are regulating cell biology.

After applying Seurat v3 integration, matched PDX specimens with similar cell states clustered together on the UMAP embedding. This is to be expected since Seurat v3 jointly reduces the dimensionality of datasets using a diagonalized CCA to identify shared biological markers and conserved gene expression signatures[15]. The algorithm then finds mutual nearest neighbors in this low-dimensional representation to recover matching cell states between datasets[23]. In addition, feature selection for integration is limited to variable features within each dissociation protocol so differences between protocols, such as the warm dissociation signature, will play a smaller role.

Not performed in this study but an important concept to highlight when using different dissociation protocols is the effect on cellular composition bias. While scRNA-seq and scRNA-seq adequately represent the original cell populations, others have noted some differences, especially for immune cells[6,7]. Due to the scope of this study being limited to PDXs, we did not have the opportunity to assess the differences in cellular composition. Our work did utilize integration techniques to compare cell states that exist in any of the dissociation methods. In other works, snRNA-seq has advantages when utilized for tissues that are difficult to dissociate such as brain, adipose, or skeletal muscle. Additionally, future work may even compare to spatial transcriptomics, which takes a snapshot of the tissue like snRNA-seq but preserves the spatial architecture and cellular composition[24]. Current resolutions do not yet detect transcriptomes at a single-cell level but future iterations with single-cell resolution may prove to be a suitable benchmark for dissociation methods.

Another interesting area of research to explore is single-cell ATAC-seq (scATAC-seq), which captures open chromatin regions at a single level. To our knowledge, dissociation protocols have not been extensively tested for scATAC-seq[16]. Typically, scATAC-seq is performed on the nucleus of cells that are freshly dissociated but can also be done on fresh-frozen samples. Comparisons of bulk ATAC-seq on flash frozen cells or tissues compared to fresh or cryopreserved samples have yielded lower quality data[25]. The open chromatin peaks were not as easily identifiable in flash frozen samples as opposed to fresh or cryopreserved samples[25]. Whether scATAC-seq suffers from similar changes in quality due to flash freezing have yet to be determined. Work by Massoni-Badosa *et al.* compared storage time at room temperature prior to scATAC-seq and showed increasing accessibility of transcription factor binding sites for *JUNB* and *FOSL1* with longer storage time. This is similar to what has been shown in our work and others with gene expression for scRNA-seq[4].

Taken together, the protocols tested here in this study were able to recapitulate the transcriptional signature of various sarcomas. However, it is critical to identify and understand the biological and technical biases of each protocol and how they may affect expression of target gene sets. Interpretations of biological processes must consider whether the protocol had an influence on the interpretation.

Materials and Methods

Collection of fresh tissue for scRNA-seq

All experiments were conducted per protocols and conditions approved by the University of Texas MD Anderson Cancer Center (MDACC; Houston, TX) Institutional Animal Care and Use Committee (eACUF Protocols #00000712-RN02). Male NOD (SCID)-IL-2Rg^{null} mice (The Jackson Laboratory; Farmington, CT) were subcutaneously injected with PDX explants (2 mm) to generate xenografts. All mice were maintained under barrier conditions and treated using protocols approved by The University of Texas MD Anderson Cancer Center's Institutional Animal Care and Use Committee. MDA-SA98-TIS02, OS1, and OS31, are PDX lines maintained by the Pediatric Solid Tumors Comprehensive Data Resource Core[26]. Once their tumors reached a volume of 150 mm^3, tumors were explanted and a portion was flash-frozen for snRNA-seq, while the remainder underwent dissociation.

Dissociation workflow from fresh solid tumor samples

Samples were collected and immediately placed into MACS® Tissue Storage Solution (Miltenyi Biotec) and kept on ice during transport. On arrival to the laboratory, samples were minced using a scalpel into fragments under ~ 0.5 mm under aseptic conditions. Next, samples were evenly split for either warm or cold enzymatic dissociation.

For warm dissociation of ES and DSRCT PDX specimens, the human Tumor Dissociation Kit (Miltenyi Biotec) was used. The dissociation was performed under manufacturer's protocol using the gentleMACS™ Dissociator (Miltenyi Biotec), a benchtop instrument for the semi-automated dissociation of tissues into single-cell suspensions. The gentleMACS Program sequenced followed the suggestion for 'Soft' Tumor type. Following completion of the program, 2x volume of media was added to the samples. This was followed by filtration through a MACS SmartStrainer (70 µm, Miltenyi Biotec) and centrifugation at 300g for 5 min. Cells were resuspended in 90% FBS and 10% DMSO at a concentration of 1 million cells per mL and placed in a Thermo Scientific™ Mr. Frosty™ Freezing Container in a -80 °C freezer.

For warm dissociation of OS PDX specimens, tissue was minced into 3 – 4 mm pieces with sterile a scalpel or scissor. The tissues were washed several times with Hank's Balanced Salt Solution (HBSS). HBSS was next aspirated and dissociation buffer (HBSS, 1 mg/mL collagenase, 3mM $CaCl_2$, 1 µg/mL DNase) was added to submerge the tissue. Tissue is then incubated at 37 °C for up to 12 hours. The cell suspension was then filtered using a 40 µm cell strainer. The filtrate is pelleted using a cenfriifugation at 400g for 5 min. Cells were resuspended freezing meddium and placed in a Thermo Scientific™ Mr. Frosty™ Freezing Container in a -80 °C freezer.

For cold dissociation, protocol was adapted from *Adam et al*[5]. Cold protease solution was prepared from 5 mM $CaCl_2$, 10 mg/mL *B. Licheniformis* protease, and 125 U/mL DNase I in 1x PBS. Tissue was minced using a scalpel into fragments under 0.5 mm. Fragments were placed in a MACS C-tube and 5 mL of ice-cold cold protease solution was added. The samples were incubated for 10 min at 4 °C with rocking. This was followed by placing the samples in a gentleMACS™ Dissociator (Miltenyi Biotec) and running the m_brain_03 program twice in succession. Afterwards, the samples were centrifuged at 300g for 5 min and resuspended in 3 mL of trypsin-EDTA for 1 min at room temperature. The trypsin-EDTA was then neutralized using ice-cold 10% FBS in 1x PBS and triturated. This was followed by filtration through a MACS SmartStrainer (70 µm, Miltenyi Biotec) and centrifugation at 300g for 5 min. Cells were resuspended freezing

medium at a concentration of 1 million cells per mL and placed in a Thermo Scientific™ Mr. Frosty™ Freezing Container in a -80 °C freezer. Cryovials were moved to LN2 storage for long-term.

Thawing cryopreserved cells

The cells were removed from LN_2 or -80 °C freezer, if they were reccently cryopreserved, and placed into a 37 °C water bath for 3 min. The contents were then transferred to a 15 mL centrifuge tube. 1 mL of complete medium was used to wash the cryovial and added drop-wise into the centrifuge tube. Next, 8 mL of complete medium was added drop-wise to reduce osmotic shock. Cells were then centrifuged at 300g for 5 min and resupsneded in 1x PBS supplemented with 0.04% BSA. This was followed by live cell enrichment using FACS. Single-cell suspensions were stained with Caclein AM live cell stain and SYTOX™ Red dead cell stain.

Nuclei Isolation workflow

The protocol was adapted from Habib *et al*[9]. We isolated nuclei from fresh-frozen tissue using the Nuclei EZ Prep Kit (Sigma-Aldrich). Fresh-frozen tissue specimen were cut into piecces < 5 mm over dry ice and then placed in 0.5 mL ice-cold EZ lysis buffer. This was followed by homogenizing using a Chemglass Life SciencesSupplier BioVortexer Mixer (Fisher Scientific) attached with a plastic microcentrifuge pestle on ice. Then 1 mL of ice-cold EZ lysis buffer was added and samples were incubated on ice for 5 min. Debris was filtered out using a pluriStrainer Mini 70 µm into a new tube. This was followed by centrifugation at 500g for 5 min. Samples were then incubated with 1 mL of ice-cold EZ lysis buffer on ice for 5 min followed by centrifugation. Afterwards, the supernatant was aspirated and 0.5 mL of Nuclei Wash and Resuspension Buffer (NWRB, 1X PBS supplemented with 1.0 % BSA and 0.2U/µl RNase Inhibitor) was carefully added without disrupting the pellet, which was followed by 5 min of incubation. Next, we added 0.5 mL of NWRB and centrifuged at 500g for 5 min. We repeated the wash and incubation once more, followed by centrifugation. The supernatant was aspirated and the nuclei were resuspended in NWRB. A portion was visualized with Trypan blue under the microscope to inspect for debris and nuclei integrity.

To sort nuclei, single-nucleus suspensions were stained with 7-AAD in NWRB for 5 min on ice. Then a BD cell sorter was used to sort up to 100,000 7-AAD positive events. Quality control of post-sort nuclei concentration was evaluated under a microscope to ensure adequate count. This was followed by loading nuclei onto a 10X chip.

Library preparation and sequencing

We followed the standard protocol set by 10x Genomics for single-cell/single-nucleus capture. A targeted capture of 5000 single cell or single nucleus were loaded onto each channel of a Chromium single-cell 3' Chip. The single cells and single nuclei were partitioned using the gel beads within the Chromium Controller. Afterwards, we performed cDNA amplification and fragmentation. This was followed by index and adapter attachment. Samples were pooled and sequenced on a NovaSeq 6000 with targeted sequence depth at 100,000 reads/cell or nucleus.

sc/snRNA-seq data preprocessing

We used Cell Ranger mkfastq to generate demultiplexed FASTQ files. Reads were aligned to human GRCh38 genome and reads were then quantified as UMIs buy Cell Ranger count. For snRNA-seq, reads were mapped with both introns and exons in Cell Ranger 5.0 using the include-introns option for counting intronic reads[10].

We performed QC and normalization separately for each sarcoma PDX. We followed the guidelines for QC from OSCA and others[17]. We inspected UMIs, gene counts, and percentage of mitochondrial genes and identified outliers based on median absolute deviation (MAD). We used a strict value of 2 or more MADs from the median while also using generic cut offs. Cells that did not meet the criteria were removed from analysis. Scrublet was used to predict and detect doublets within the data[27]. While doublets were flagged, there was not a single cluster of doublets, which would be evident as an artifact, so no cells were removed.

Data normalization, dimensional reduction, and comparisons

Seurat v3 was used for sample normalization, dimensional reduction, scaling, and differential expression analysis[15]. We used Wilcoxon test to compare gene expression between protocols. Enrichr was used for pathway enrichment. We set a log2 fold change threshold of log2(1.5) or greater. This will result in genes that are 50% greater than the baseline. The AddModuleScore function in Seurat v3 was used to observe the averaged gene expression of different gene set. We used curated gene sets of a warm dissociation signature from O'Flanagan et al.[4], EWS-FLI1 gene targets[13], EWS-WT1 gene targets[14] and osteoblastic and chondroblastic signatures classically associated with the tissue origin of OS (**Table S7**). The osteoblastic and chondroblastic signatures were found on Harmonizome (https://maayanlab.cloud/Harmonizome/). The osteoblastic signature was specifically found in the GeneRIF Biological Term Annotations under 'Osteoblastic'. The chondroblastic signature was specifically found in the TISSUES Text-mining Tissue Protein Expression Evidence Scores under 'Chondroblasts'. To find conserved markers between dissociation methods, we used the function FindConservedMarkers in Seurat v3. We performed integration using the integration functions within Seurat v3. The datasets were integrated by dissociation protocol.

Predicting sample type by bias scores

To classify nuclei and cell using the length bias and warm dissociation scores, data sets were randomly split into a training and test set. To prevent data leakage, scaled data was not used. We then calculated the gene set scores separately on the training and test sets. A logistic regression model was fit to the training set on either the warm dissociation or length bias score to predict for cells and nuclei respectively. We calculated the probabilities and the area under the curve using the pROC v1.18.0 package. This was compared to a random gene signature equal in number of genes of either length bias or warm dissociation gene sets.

Declarations

Ethics Approval

All experiments were conducted per protocols and conditions approved by the University of Texas MD Anderson Cancer Center (MDACC; Houston, TX) Institutional Animal Care and Use Committee (eACUF Protocols #00000712-RN02).

Consent for Publication

N/A

Availability of data and materials

The datasets generated analyzed in this study will be available from the Gene Expression Omnibus (GEO) repository at time of publication. Additional datasets analyzed in this study can be found at GSE140819.

Competing interests

The authors declare that they have no competing interests.

Funding

The University of Texas MD Anderson Cancer Center is supported by the National Institutes of Health through Cancer Center Support Grant CA016672. The ATGC is supported by the Core grant CA016672 (ATGC) and NIH 1S10OD024977-01. JAL is supported by R01-CA180279-01A1. and the generous philanthropy of the Cory Monzingo Foundation. DDT and JAL are supported by the Moeller Foundation.

Affiliations

Sarcoma Medical Oncology department, The University of Texas MD Anderson Cancer Center, Houston, Texas 77005

Danh D. Truong, Salah-Eddine Lamhamedi-Cherradi, Robert W. Porter, Sandhya Krishnan, John A. Livingston & Joseph A. Ludwig.

Division of Pediatrics, The University of Texas MD Anderson Cancer Center, Houston, Texas 77005

Jyothishmathi Swaminathan, Amber L. Gibson, Vidya Gopalakrishnan, Nancy Gordon, Najat C. Daw & Richard Gorlick

Division of Pathology, The University of Texas MD Anderson Cancer Center, Houston, Texas 77005

Alexander J. Lazar

Authors' contributions

DDT contributed to the conception, design, analysis, and interpretation of data. SC, RWP, SK, and JS contributed to sample acquisition and dissociation. SC, NG, NCW, VG, and ALG contributed to specimen database and clinical information. VG, NG, NCW contributed to interpretation and review of data. SC, AJL, and JA Livingston contributed conception and design. DDT and JA Ludwig contributed to the writing, review, and revision of the manuscript with input from all authors. JA Ludwig and SC contributed to study supervision. All authors read and approved the final manuscript.

Corresponding author

Correspondence to Joseph A. Ludwig.

Acknowledgements

The authors acknowledge the support of the High-Performance Computing for research facility at the University of Texas MD Anderson Cancer Center for providing computational resources that have contributed to the research results reported in this paper. The authors acknowledge the support of the Cancer Prevention Research Institute of Texas (CPRIT) Pediatric Solid Tumors Comprehensive Data Resource Core for providing tissue specimens.

Figures

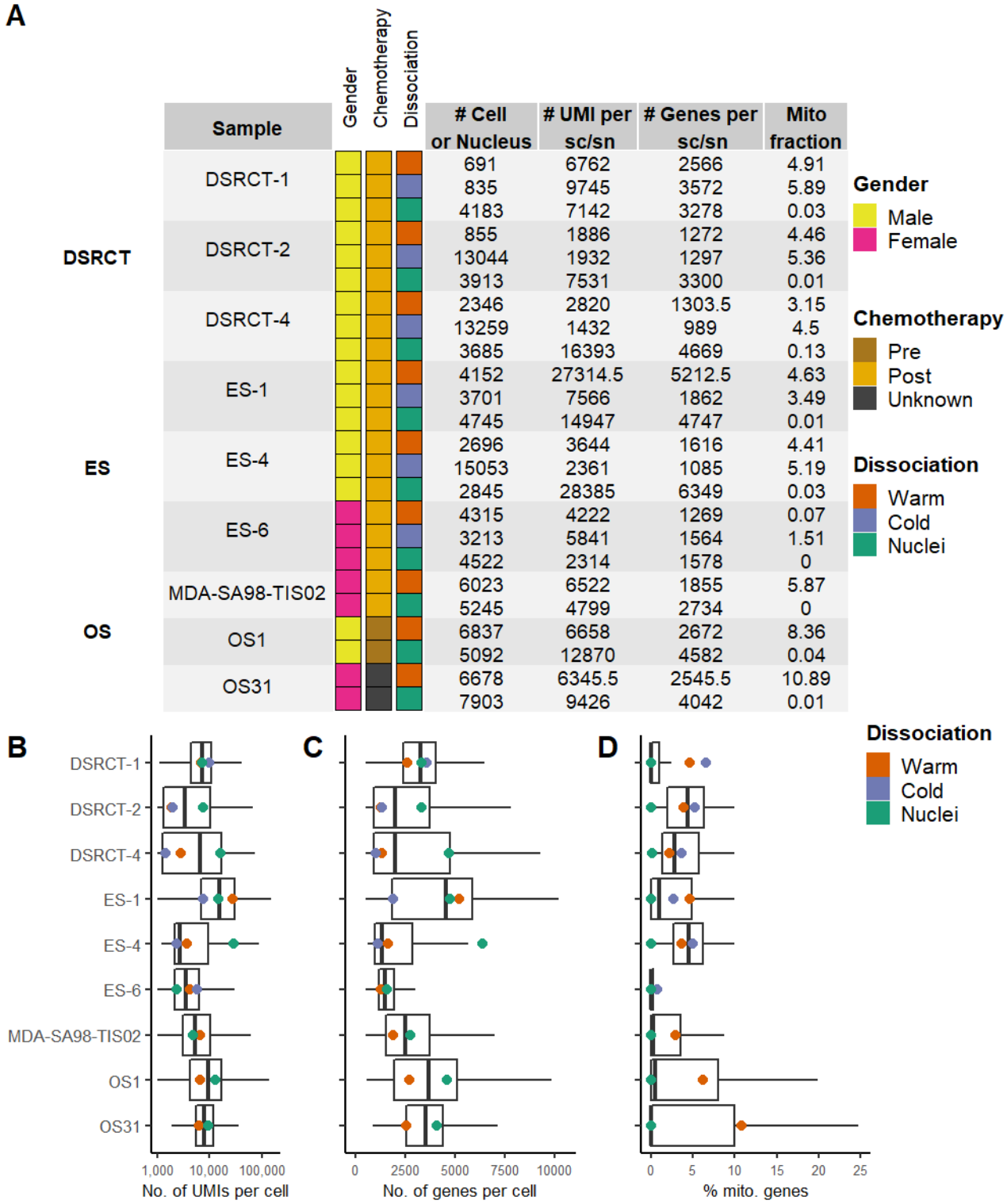

Figure 1. Patient information and quality control metrics. a Overview of all sarcoma subtypes that were processed and evaluated. For each sample, the number of cells or nuclei passing QC thresholds, number of sequencing reads per cell/nuc, number of genes per cell/nuc, and the

median percentage of UMIs mapping to mitochondrial genes are displayed in the table. All samples had less than 0.01 doublet fraction. **b** Distributions (median and first and third quartiles) for number of UMIs per cells/nuc, number of genes per cell/nuc, and percentage of UMIs mapping to mitochondrial genes vary across sarcoma subtypes and choice of protocol.

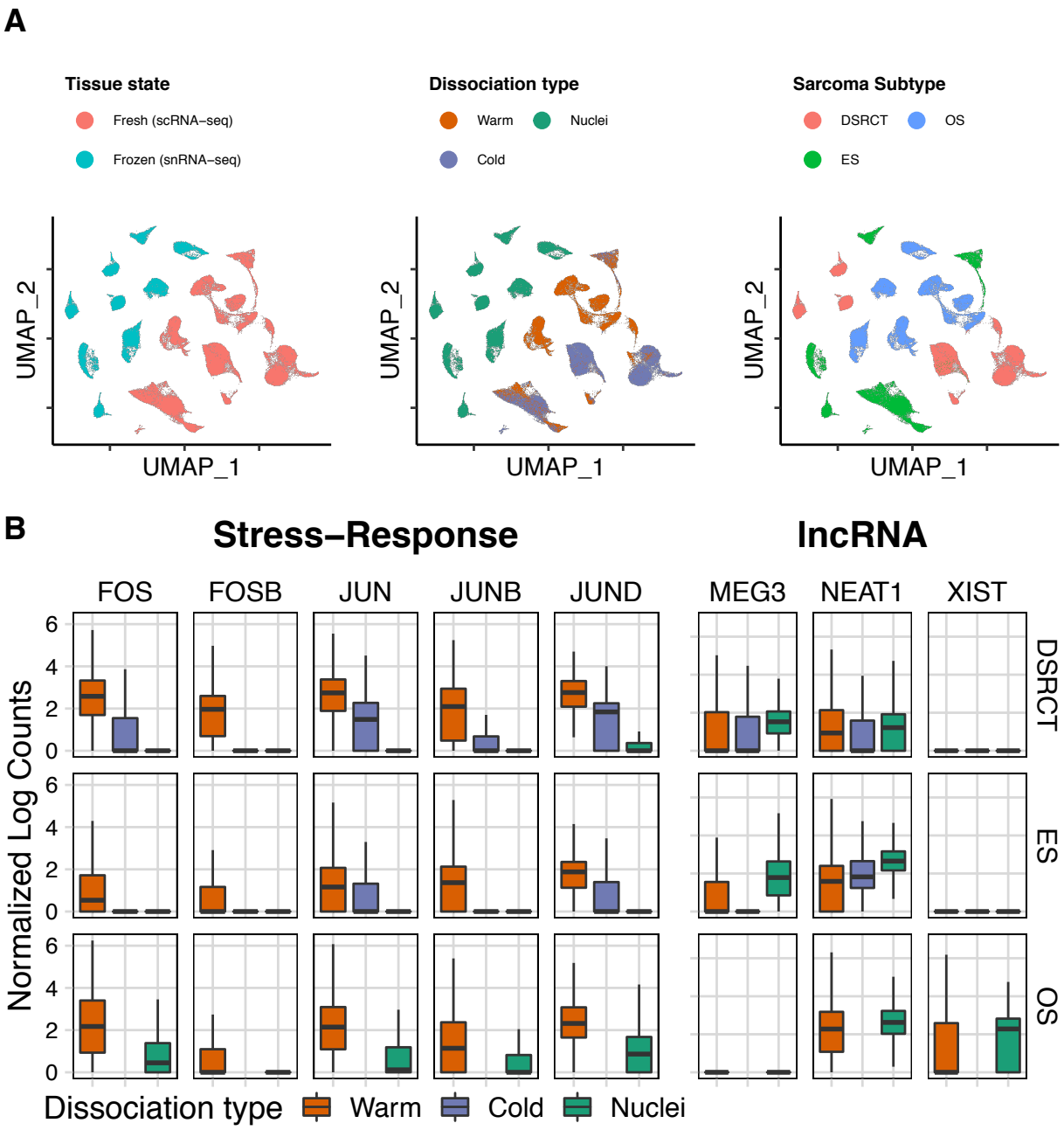

Figure 2. Performance of each protocol on various sarcoma subtypes. a UMAP embedding of all cells labeled by Tissue state, Dissociation type, and Sarcoma subtype. **b** Enrichment of select stress-response genes in the Warm dissociation protocol and select lncRNA in the Nuclei isolation protocol.

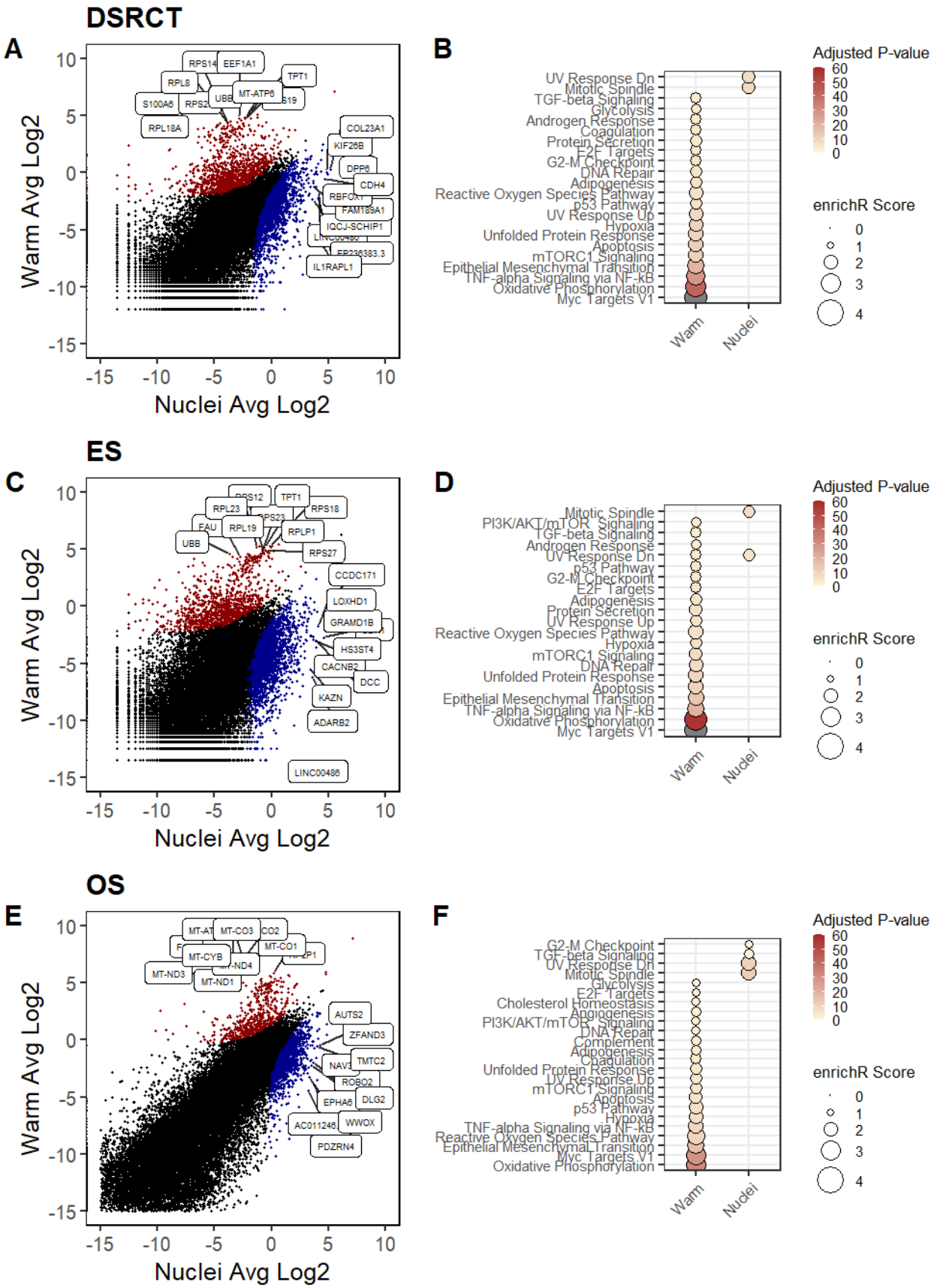
DSRCT
A
Warm Avg Log2
Nuclei Avg Log2
RPS14
EEF1A1
TPT1
RPL8
MT-ATP6
S100A6
UBB
RPL18A
COL23A1
KIF26B
DPP6
CDH4
FAM189A1
IQCJ-SCHIP1
IL1RAPL1
B
UV Response Dn
Mitotic Spindle
TGF-beta Signaling
Glycolysis
Androgen Response
Coagulation
Protein Secretion
E2F Targets
G2-M Checkpoint
DNA Repair
Adipogenesis
Reactive Oxygen Species Pathway
p53 Pathway
UV Response Up
Hypoxia
Unfolded Protein Response
Apoptosis
mTORC1 Signaling
Epithelial Mesenchymal Transition
TNF-alpha Signaling via NF-kB
Oxidative Phosphorylation
Myc Targets V1
Warm
Nuclei
Adjusted P-value
enrichR Score
ES
C
Warm Avg Log2
Nuclei Avg Log2
TPT1
RPS18
RPL23
RPL19
RPLP1
UBB
RPS27
CCDC171
LOXHD1
GRAMD1B
HS3ST4
CACNB2
DCC
KAZN
ADARB2
LINC00486
D
Mitotic Spindle
PI3K/AKT/mTOR Signaling
TGF-beta Signaling
Androgen Response
UV Response Dn
p53 Pathway
G2-M Checkpoint
E2F Targets
Adipogenesis
Protein Secretion
UV Response Up
Reactive Oxygen Species Pathway
Hypoxia
mTORC1 Signaling
DNA Repair
Unfolded Protein Response
Apoptosis
Epithelial Mesenchymal Transition
TNF-alpha Signaling via NF-kB
Oxidative Phosphorylation
Myc Targets V1
Warm
Nuclei
Adjusted P-value
enrichR Score
OS
E
Warm Avg Log2
Nuclei Avg Log2
MT-CYB
MT-ND3
MT-ND1
MT-CO1
AUTS2
ZFAND3
TMTC2
ROBO2
EPHA6
DLG2
WWOX
PDZRN4
F
G2-M Checkpoint
TGF-beta Signaling
UV Response Dn
Mitotic Spindle
Glycolysis
E2F Targets
Cholesterol Homeostasis
Angiogenesis
PI3K/AKT/mTOR Signaling
DNA Repair
Complement
Adipogenesis
Coagulation
Unfolded Protein Response
UV Response Up
mTORC1 Signaling
Apoptosis
p53 Pathway
Hypoxia
TNF-alpha Signaling via NF-kB
Reactive Oxygen Species Pathway
Epithelial Mesenchymal Transition
Myc Targets V1
Oxidative Phosphorylation
Warm
Nuclei
Adjusted P-value
enrichR Score

Figure 3. DEG Biases introduced by Warm and Nuclei Protocols. Scatter plot of log transformed gene expression levels between warm and nuclei on the left column. Red indicates up-regulated in warm, and blue indicates up-regulated in nuclei with p-value < 0.05. Black is non-significant. Dot plot of enrichR scores of the Hallmark gene sets from MSigDB are on the right column. Plots are shown for DSRCT **a, b**; ES **c, d**; and OS **e, f**.

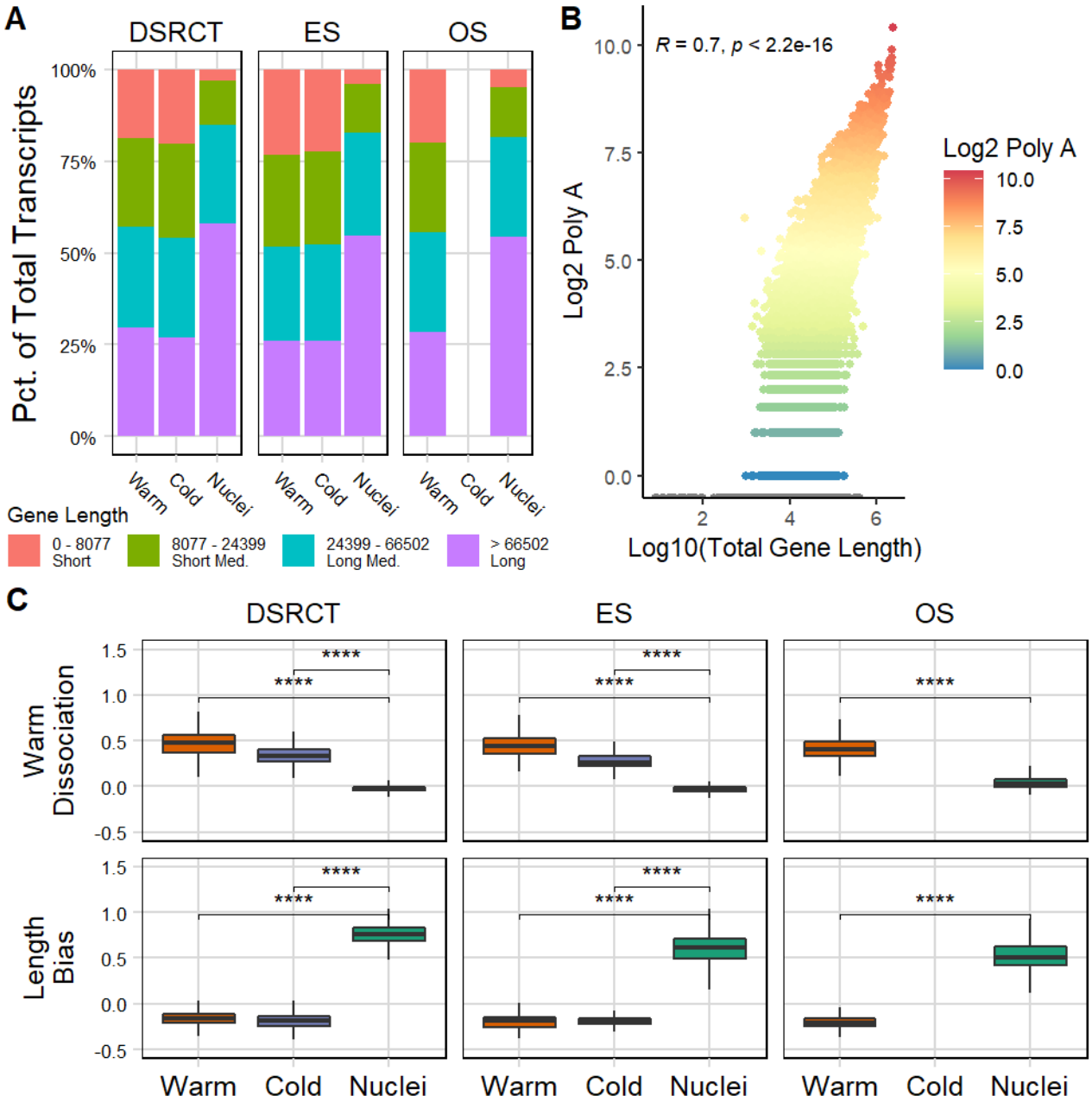

Figure 4. Gene-length associated bias in snRNA-seq. a snRNA-seq samples were enriched in genes with longer gene length. Genes were split into quartiles based on gene length. **b** Increased PolyA regions were associated with longer gene length. **c** Warm Dissociation score was significantly higher in the Warm and Cold protocol as opposed to the Nuclei protocol. Length Bias score was significantly higher in the Nuclei protocol (Wilcoxon test, **** denotes p <= 0.0001)

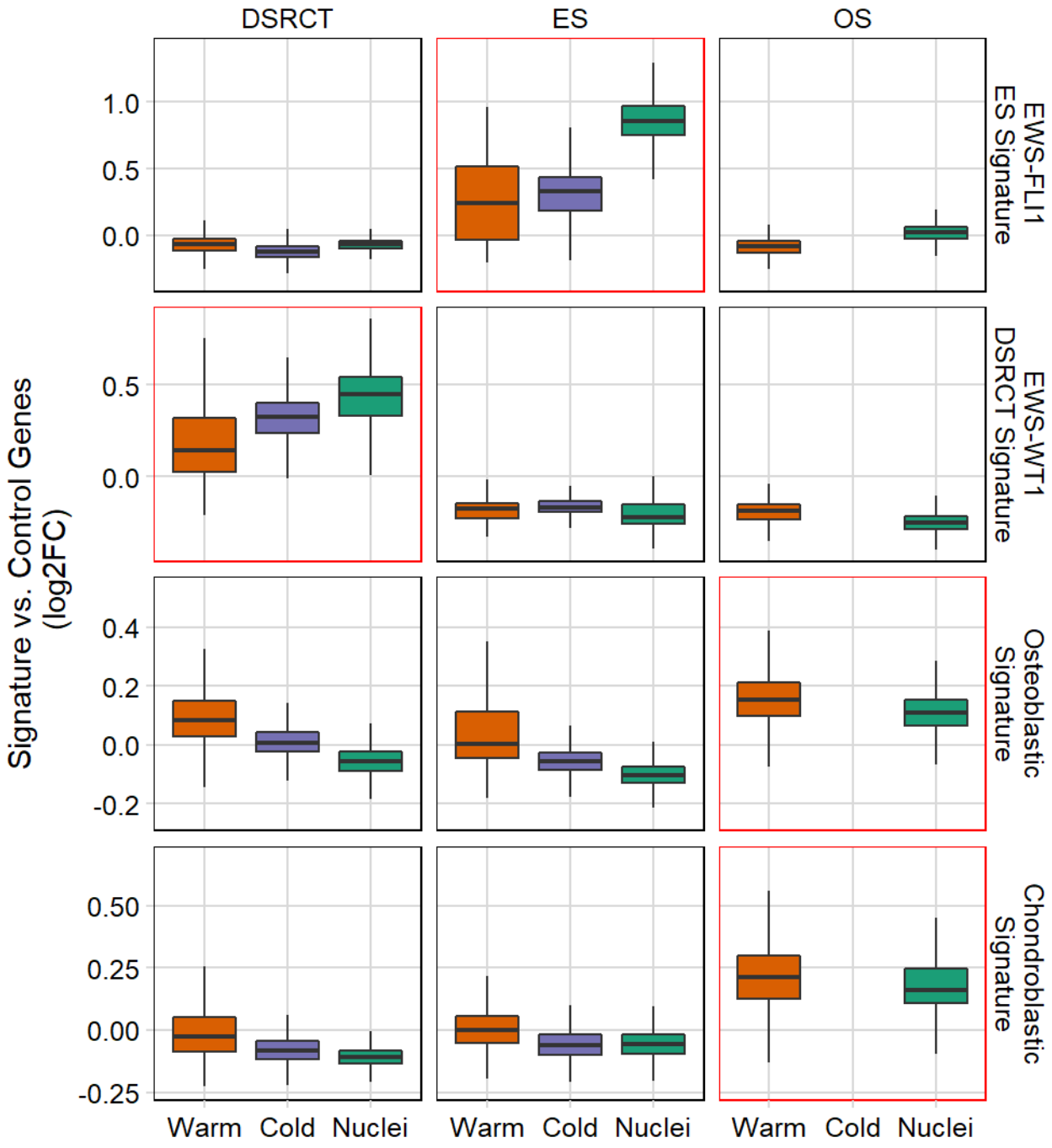

Figure 5. Biases associated with protocol on Sarcoma specific signatures. Sarcoma specific signatures were only detected in the respective sarcoma subtypes (red box). EWS-FLI1 ES signature is derived from a gene set of EWS-FLI1 target genes. EWS-WT1 DSRCT signature is derived from a gene set of EWS-WT1 target genes. Both osteoblastic and chondroblastic signatures are gene sets derived from reference databases.

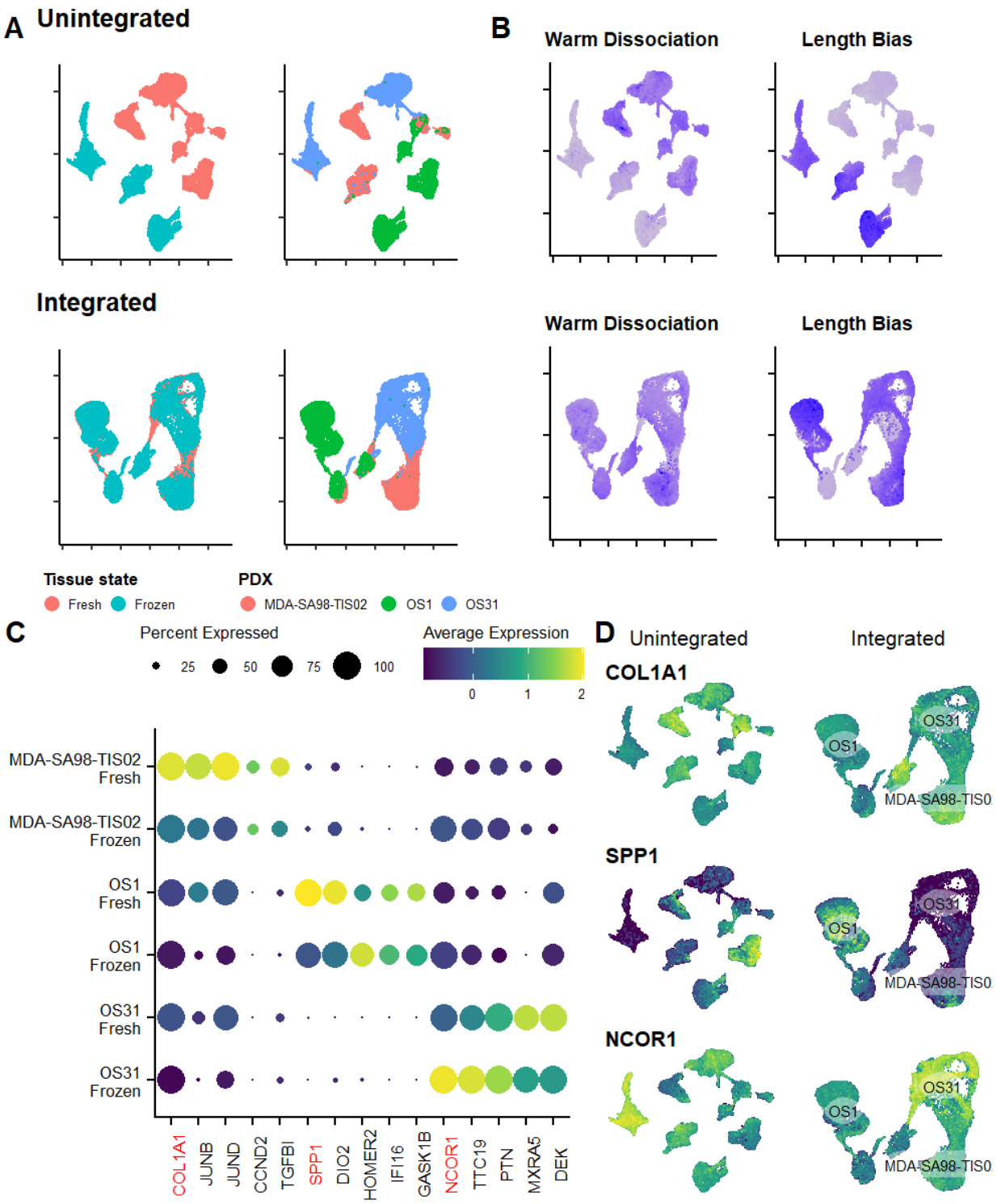

Figure 6. Integration recovers matching cell states from different dissociation methods for OS. **a** UMAP embeddings after integration (bottom row) showed alignment of matching PDX specimens. **b** After integration (bottom row), clusters on the UMAP were no longer affected by the identified biases. Darker blue indicates higher expression. **c** Dot plot of conserved markers

show unique gene expression for each PDX. Top marker is in red. **d** After integration (right column), top conserved markers correctly aligned with the associated PDX cluster.

References

1. Alizadeh AA, Aranda V, Bardelli A, et al. Toward understanding and exploiting tumor heterogeneity. Nat Med 2015;21:846-53.
2. Method of the Year 2019: Single-cell multimodal omics. Nat Methods 2020;17:1.
3. Wang Y, Navin NE. Advances and applications of single-cell sequencing technologies. Mol Cell 2015;58:598-609.
4. O'Flanagan CH, Campbell KR, Zhang AW, et al. Dissociation of solid tumor tissues with cold active protease for single-cell RNA-seq minimizes conserved collagenase-associated stress responses. Genome Biol 2019;20:210.
5. Adam M, Potter AS, Potter SS. Psychrophilic proteases dramatically reduce single-cell RNA-seq artifacts: a molecular atlas of kidney development. Development 2017;144:3625-32.
6. Slyper M, Porter CBM, Ashenberg O, et al. A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human tumors. Nat Med 2020;26:792-802.
7. Denisenko E, Guo BB, Jones M, et al. Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. Genome Biol 2020;21:130.
8. Ding J, Adiconis X, Simmons SK, et al. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. Nat Biotechnol 2020;38:737-46.
9. Habib N, Avraham-Davidi I, Basu A, et al. Massively parallel single-nucleus RNA-seq with DroNc-seq. Nat Methods 2017;14:955-8.
10. Bakken TE, Hodge RD, Miller JA, et al. Single-nucleus and single-cell transcriptomes compared in matched cortical cell types. PLoS One 2018;13:e0209648.
11. Sun Q, Hao Q, Prasanth KV. Nuclear Long Noncoding RNAs: Key Regulators of Gene Expression. Trends Genet 2018;34:142-57.
12. Lake BB, Codeluppi S, Yung YC, et al. A comparative strategy for single-nucleus and single-cell transcriptomes confirms accuracy in predicted cell-type expression from nuclear RNA. Sci Rep 2017;7:6031.
13. Aynaud MM, Mirabeau O, Gruel N, et al. Transcriptional Programs Define Intratumoral Heterogeneity of Ewing Sarcoma at Single-Cell Resolution. Cell Rep 2020;30:1767-79 e6.
14. Gedminas JM, Chasse MH, McBrairty M, Beddows I, Kitchen-Goosen SM, Grohar PJ. Desmoplastic small round cell tumor is dependent on the EWS-WT1 transcription factor. Oncogenesis 2020;9:41.
15. Stuart T, Butler A, Hoffman P, et al. Comprehensive Integration of Single-Cell Data. Cell 2019;177:1888-902 e21.
16. Massoni-Badosa R, Iacono G, Moutinho C, et al. Sampling time-dependent artifacts in single-cell genomics studies. Genome Biol 2020;21:112.
17. Amezquita RA, Lun ATL, Becht E, et al. Orchestrating single-cell analysis with Bioconductor. Nat Methods 2020;17:137-45.
18. Gangwal K, Sankar S, Hollenhorst PC, et al. Microsatellites as EWS/FLI response elements in Ewing's sarcoma. Proc Natl Acad Sci U S A 2008;105:10149-54.

19. Guillon N, Tirode F, Boeva V, Zynovyev A, Barillot E, Delattre O. The oncogenic EWS-FLI1 protein binds in vivo GGAA microsatellite sequences with potential transcriptional activation function. PLoS One 2009;4:e4932.
20. Lopes I, Altab G, Raina P, de Magalhaes JP. Gene Size Matters: An Analysis of Gene Length in the Human Genome. Front Genet 2021;12:559998.
21. Sahakyan AB, Balasubramanian S. Long genes and genes with multiple splice variants are enriched in pathways linked to cancer and other multigenic diseases. BMC Genomics 2016;17:225.
22. Kovar H. Downstream EWS/FLI1 - upstream Ewing's sarcoma. Genome Med 2010;2:8.
23. Haghverdi L, Lun ATL, Morgan MD, Marioni JC. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol 2018;36:421-7.
24. Method of the Year 2020: spatially resolved transcriptomics. Nat Methods 2021;18:1.
25. Fujiwara S, Baek S, Varticovski L, Kim S, Hager GL. High Quality ATAC-Seq Data Recovered from Cryopreserved Breast Cell Lines and Tissue. Sci Rep 2019;9:516.
26. Houghton PJ, Morton CL, Tucker C, et al. The pediatric preclinical testing program: description of models and early testing results. Pediatr Blood Cancer 2007;49:928-40.
27. Wolock SL, Lopez R, Klein AM. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. Cell Syst 2019;8:281-91 e9.