Federal Aviation Administration

DOT/FAA/AM-23/01
Office of Aerospace Medicine
Washington, D.C. 20591

An Evaluation of the Downstream Effects of Purification Methods on RNA-Seq Differential Expression

Susan K. Munster
Hilary A. Uyhelji
Scott J. Nicholson

Civil Aerospace Medical Institute (CAMI)
Federal Aviation Administration
Oklahoma City, OK 73169

January 2023

NOTICE

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The United States Government assumes no liability for the contents thereof.

This publication and all Office of Aerospace Medicine technical reports are available in full-text from the Civil Aerospace Medical Institute's publications website (www.faa.gov/go/oamtechreports) and at the National Transportation Library's Repository & Open Science Access Portal (https://rosap.ntl.bts.gov/)

Technical Documentation Page

1. Report No. DOT/FAA/AM-23/01	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle An Evaluation of the Downstream Effects of Purification Methods on RNA-Seq Differential Expression		5. Report Date January 2023 6. Performing Organization Code AAM-612
7. Author(s) Munster SK, Uyhelji HA, Nicholson SJ		8. Performing Organization Report No. DOT/FAA/AM-23/01
9. Performing Organization Name and Address Civil Aerospace Medical Institute (CAMI) Federal Aviation Administration Oklahoma City, OK 73169		10. Work Unit No. (TRAIS) NA? 11. Contract or Grant No.
12. Sponsoring Agency Name and Address Office of Aerospace Medicine Federal Aviation Administration 800 Independence Ave., S.W. Washington, DC 20591		13. Type of Report and Period Covered Technical Report 14. Sponsoring Agency Code

15. Supplementary Notes

Author ORCIDs:
Munster (0000-0002-3223-0076), Uyhelji (0000-0002-3433-8320), Nicholson (0000-0002-2201-744X)
Technical report DOI: https://doi.org/10.21949/1524442

16. Abstract

Ribonucleic acid sequencing (RNA-Seq) is a valuable and commonly used technique to quantify the number of individual RNA transcripts within a sample. RNA-Seq typically requires a small amount of pure and concentrated RNA, which can necessitate additional concentration or purification of previously isolated RNA samples. Magnetic beads and silica-based columns are often used to concentrate and/or purify RNA samples, but little is known about how these techniques influence downstream analyses. In this study, we collected blood from volunteer human subjects and pooled those samples during RNA extraction to minimize variance due to input material. We then purified aliquots of that sample pool to evaluate how sample purification and concentration influenced gene expression observations. Extracted RNA was sequenced, and the resulting RNA-Seq files were evaluated to determine the degree of differential expression between methods. Differential expression was detected in roughly half of the comparisons made and appeared attributable at least partly to differences in sample concentration and purification techniques.

17. Key Word RNA-Seq, sample purification, differential expression	18. Distribution Statement Document is available to the public through the National Transportation Library: https://ntl.bts.gov/ntl		
19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 27	22. Price NA

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

Author Note

Funding: This research was funded by the Federal Aviation Administration.

Conflicts of Interest: The authors declare that they have no competing interests.

Author Contributions:

SKM, SJN, and HAU designed the study. SKM performed sample processing, analysis, and writing with input from SJN and HAU. All authors have read and approved this manuscript.

Data Availability: Genetics datasets are stored at NIH dbGaP accession (phs003001.v1.p1).

Acknowledgements

The authors wish to acknowledge the feedback and guidance provided by our reviewers, Alynn A. Martin, Ph.D., and Christopher Tracy, Ph.D.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

DE	Differential expression
FDR	False discovery rate
FPKM	Fragments per kilobase of exon model per million reads map
HC	High concentration
HC-BR	BR5-eluted unpurified high concentration
lncRNA	Long non-coding RNA*
miRNA	MicroRNA*
miscRNA	Miscellaneous RNA*
PCA	Principal components analysis
RNA-Seq	Ribonucleic acid sequencing
snRNA	Small nuclear RNA*
snoRNA	Small nucleolar RNA*

*as described by Ensembl West (2022c)

Abstract

Ribonucleic acid sequencing (RNA-Seq) is a valuable and commonly used technique to quantify the number of individual RNA transcripts within a sample. RNA-Seq typically requires a small amount of pure and concentrated RNA, which can necessitate additional concentration or purification of previously isolated RNA samples. Magnetic beads and silica-based columns are often used to concentrate and/or purify RNA samples, but little is known about how these techniques influence downstream analyses. In this study, we collected blood from volunteer human subjects and pooled those samples during RNA extraction to minimize variance due to input material. We then purified aliquots of that sample pool to evaluate how sample purification and concentration influenced gene expression observations. Extracted RNA was sequenced, and the resulting RNA-Seq files were evaluated to determine the degree of differential expression between methods. Differential expression was detected in roughly half of the comparisons made and appeared attributable at least partly to differences in sample concentration and purification techniques.

Introduction

Ribonucleic acid sequencing (RNA-Seq) has become the prevalent technique used in gene expression analyses. The approach is invaluable for biomarker discovery, including efforts to discover molecular signatures associated with human impairment and aviation safety risks. As RNA-Seq processing requires a small amount of pure and concentrated RNA, increasing the concentration of low-yield samples is occasionally necessary. Even when RNA concentrations are adequate, a post-extraction purification step may be necessary to remove residual salts or other contaminants. Column and bead-based purification are well-established methods that simultaneously concentrate and purify samples. However, little is known regarding whether either of these methods could have a downstream effect on differential expression (DE) that may obscure the detection of meaningful biological differences in biomarker discovery research.

Magnetic bead purification uses the physical properties of RNA to reversibly bind it to magnetic beads while in the presence of a buffer, wash the RNA to remove impurities, and then elute it using an appropriate volume of water or elution buffer (Alderton et al., 1992). By eluting most of the retained RNA in a reduced volume relative to the original sample, this approach allows the sample to be both purified and concentrated in a single procedure. According to Beckman Coulter Life Sciences, their AMPure magnetic beads are widely used in more than 200 library preparation kits (Beckman Coulter, 2020). Column purification similarly uses selective binding to adhere RNA to a silica membrane column in the presence of a binding buffer containing a guanidine compound, wash contaminants through the column, and elute the purified and concentrated RNA (Boom et al., 1990; Wen et al., 2008). Both methods are preferred over previously developed methods for their simplicity, the resulting product purity, and because they

use fewer hazardous chemicals (Berensmeier, 2006; Boom et al., 1990; Gonzalez-Perez et al., 2007).

Previous studies evaluating these techniques demonstrate conflicting results, with some finding that column purification provides greater yields of nucleic acid (Diefenbach et al., 2018; Kim et al., 2009; Riemann et al., 2007), some finding no statistically significant difference between the two methods (Lee et al., 2010), and others finding that magnetic bead purification resulted in higher yields (Jorgez et al., 2006). Both approaches are widely used in manual and automated purification procedures. In the present study, the use of column and magnetic bead purification prior to RNA sample library preparation and sequencing were tested for their impact on differential gene expression. Human blood samples were collected and pooled to create a homogeneous sample source, and RNA was eluted in two different solvents to test the impact of the elution buffer. Samples eluted in water were either directly sequenced or diluted and subjected to an additional cleanup step using MinElute silica columns or AMPure magnetic beads prior to RNA sequencing. Additionally, different concentrations of RNA were used for library preparation and sequencing to test the effect of RNA quantity. The null hypothesis is that none of these processing variances will affect differential expression, and no DE will be observed among samples due to elution solvent choice, input sample concentration, or purification approach.

Materials & Methods

Sample collection and extraction

Samples were homogenized to minimize technical variation other than the tested factors of solvent, purification method, or quantity. To this end, multiple human blood samples were collected from three individuals (with informed consent and Federal Aviation Administration Institutional Review Board approval) in PAXgene Blood RNA Tubes (BD Biosciences, P/N: 762165). Total RNA was extracted with QIAGEN's PAXgene Blood miRNA kit (P/N: 763134) using a QIAcube Connect (QIAGEN, P/N: 9002864). Following pellet suspension in BM1 buffer, the BM1 suspensions were combined and thoroughly mixed. The combined suspension was then divided into aliquots to complete the extraction process. Aliquoted samples proceeded through the manufacturer's protocol until the final elution step, where the elution buffer effect was assayed. One-third of the samples were eluted in BR5, the elution buffer supplied with the PAXgene Blood miRNA kit, and two-thirds were eluted in RNAse-free water (PreAnalytiX), referred to hereafter as water, for comparison.

QA/QC

After extraction, the RNA integrity number (RIN^e) for each sample was measured using Tapestation 4200 (Agilent, P/N: G2991BA) RNA screentapes (Agilent, P/N: 5067-5576). Samples with $RIN^e \geq 7.5$ were retained for further analysis. RNA concentration was measured using Qubit 3.0 (Invitrogen Life Technologies, P/N: Q33216) and the Qubit RNA BR Assay kit.

All water-eluted samples were pooled, and BR5-eluted samples were separately pooled. Concentrations were measured again, and samples were diluted for use in testing.

Dilutions and aliquot preparation for purification

BR5 and water-eluted samples were diluted using BR5 and water, respectively, to 80 ng/µL to provide standardized high-concentration unpurified samples. Three aliquots of the high-concentration unpurified BR5-eluted sample, designated HC-BR1, HC-BR2, and HC-BR3, were set aside for RNA-Seq. Three aliquots of the high-concentration unpurified water-eluted sample, designated as HC-1, HC-2, and HC-3, were also set aside for RNA-Seq analysis. Unpurified samples of medium (~30 ng/µL) and low concentration (~15 ng/µL) were produced by diluting the HC sample with water and designated as LOW-30A, LOW-30B, LOW-15A, and LOW-15B. Aliquots for purification and concentration using AMPure beads were made by diluting some of the HC sample to 20 ng/µL (LC-A1, LC-A2, LC-AP1, LC-AP2, LC-AW1, LC-AW2, LC-AB1, LC-AB2, LC-AT1, LC-AT2, LC-ATB1, and LC-ATB2). Aliquots for purification and concentration using MinElute columns were made by diluting the HC sample to approximately 60 ng/µL, 30 ng/µL, and 10 ng/µL to make high-concentration samples HC-ME1 and HC-ME2, medium-concentration samples MC-ME1 and MC-ME2, and low-concentration samples LC-ME1 and LC-ME2.

AMPure purification

Samples LC-A1, LC-A2, LC-AP1, LC-AP2, LC-AW1, LC-AW2, LC-AB1, LC-AB2, LC-AT1, LC-AT2, LC-ATB1, and LC-ATB2 were prepared to test variations in the magnetic bead purification procedure. AMPure bead purification/concentration was carried out using AMPure XP beads (Beckman Coulter, P/N: NC9959336); in the final step, all samples were eluted in water. In previous studies conducted in our facility, AMPure purification efficiency varied. Therefore, multiple modifications to the AMPure concentration procedure were tested (Table 1). Two samples (LC-A1 and LC-A2) were prepared according to the manufacturer's recommended procedure (Beckman Coulter, 2016). Two samples (LC-AP1 and LC-AP2) were prepared with added mixing by pipetting 10 times while samples underwent the initial bead incubation before being placed on the magnet. Two more samples (LC-AW1 and LC-AW2) were prepared with added pipette mixing (10 times) while being incubated on the magnet prior to ethanol washes. Two samples (LC-AB1 and LC-AB2) were prepared with added pipette mixing (10 times) both before and while the sample was incubated on the magnet. Two samples (LC-AT1 and LC-AT2) did not have any additional mixing but did have an added 10 minutes of incubation while samples were on the magnet prior to the ethanol washes. Lastly, two samples (LC-ATB1 and LC-ATB2) were prepared with pipette mixing 10 times both prior to and while on the magnet and with an added 10 minutes of incubation while on the magnet, prior to the ethanol washes.

Table 1.
Sample Preparation Table

Sample	Eluted in	Purified with	Added Purification Processing Steps	Target Concentration Prior to Purification (ng/μL)	Concentration After Purification (ng/μL)
HC-1 HC-2 HC-3	RNase free water	N/A	None	80	74.2 74.2 74.2 (mean=74.2)
LOW-30A LOW-30B	RNase free water	N/A	None	30	37.4 37.4 (mean = 37.4)
LOW-15A LOW-15B	RNase free water	N/A	None	15	19.9 19.9 (mean = 19.9)
HC-BR1 HC-BR2 HC-BR3	BR5	N/A	None	80	74.5 74.5 74.5 (mean = 74.5)
LC-A1 LC-A2	RNase free water	AMPure beads	As per manufacturer's instructions	20	21.8 30.9 (mean = 26.4)
LC-AP1 LC-AP2	RNase free water	AMPure beads	As per manufacturer's instructions, with added mixing before placed on magnet*	20	32.3 59.1 (mean = 45.7)
LC-AW1 LC-AW2	RNase free water	AMPure beads	As per manufacturer's instructions, with added mixing while on magnet*	20	51.3 24.3 (mean = 37.8)
LC-AB1 LC-AB2	RNase free water	AMPure beads	As per manufacturer's instructions with added mixing before and while on magnet*	20	60.1 51.2 (mean = 55.7)
LC-AT1 LC-AT2	RNase free water	AMPure beads	As per manufacturer's instructions with added time while on magnet*	20	18.0 32.7 (mean = 25.4)
LC-ATB1 LC-ATB2	RNase free water	AMPure beads	As per manufacturer's instructions, with added mixing before and while on magnet and added time while on magnet*	20	36.1 35.7 (mean = 35.9)
HC-ME1 HC-ME2	RNase free water	MinElute column	As per manufacturer's instructions	60	58.5 62.3 (mean = 60.4)
MC-ME1 MC-ME2	RNase free water	MinElute column	As per manufacturer's instructions	30	48.4 46.7 (mean = 47.6)
LC-ME1 LC-ME2	RNase free water	MinElute Column	As per manufacturer's instructions	9.6	10.0 10.3 (mean = 10.2)

Note. N/A = not applicable.

*All added mixing occurred prior to ethanol washes.

MinElute purification

The Rneasy MinElute Cleanup Kit (QIAGEN, P/N: 74204) was used for column concentration and purification of samples at low (LC-ME1 and LC-ME2), medium (MC-ME1 and MC-ME2), and high (HC-ME1 and HC-ME2) concentrations. Dilution factors were chosen to model typical-to-low sample RNA concentrations often encountered with blood samples. The MinElute procedure was performed, and samples were eluted in water as directed by the manufacturer without deviation. Descriptions of all samples prepared are given in Table 1.

RNA-Seq, data QA/QC, alignment, and differential expression analysis

After all samples were prepared and any sample purifications/concentrations indicated above were performed, sample concentrations were again measured prior to submission to Baylor College of Medicine for library preparation and sequencing (Table 2). Total RNA-Seq was performed using 9 μL volume per sample for library preparation with Illumina's TruSeq Stranded Total RNA with Ribo-Zero Globin Kit (P/N: 20020612) to produce paired-end reads of 150 bp in length, followed by sequencing on a NovaSeq 6000 (Illumina). Sequences are stored at NIH dbGaP (accession phs003001.v1.p1). The effects of the purification and concentration were assessed by comparing final sequencing results to detect differential gene expression.

Table 2.
Input Concentrations, Mass, Raw Reads, and Trimmed Reads for Each Sample Before Alignment and Differential Expression Analysis

Sample Name	Input Concentration, ng/µL (Qubit)	Input Mass Provided, ng (Qubit)	Raw Reads (n, millions)	Total Trimmed Reads (n, millions)
HC-1	74.2	667.8	117.76	117.38
HC-2	74.2	667.8	91.89	91.72
HC-3	74.2	667.8	84.61	84.44
HC-BR1	74.5	670.5	96.75	96.5
HC-BR2	74.5	670.5	123.74	123.48
HC-BR3	74.5	670.5	94.43	94.31
LC-A1	21.8	196.2	83.42	83.17
LC-A2	30.9	278.1	64.75	64.6
LC-AB1	60.1	540.9	81.25	80.84
LC-AB2	51.2	460.8	88.56	88.43
LC-AP1	32.3	290.7	92.75	92.56
LC-AP2	59.1	531.9	70.52	70.4
LC-AT1	18.0	162	68.96	68.81
LC-AT2	32.7	294.3	60.25	60.17
LC-ATB1	36.1	324.9	76.52	76.37
LC-ATB2	35.7	321.3	55.45	55.33
LC-AW1	51.3	461.7	101.63	101.14
LC-AW2	24.3	218.7	54.27	54.17
LOW-15-1	19.9	179.1	68.94	68.61
LOW-15-2	19.9	179.1	65.37	65.19
LOW-30-1	37.4	336.6	122.82	122.49
LOW-30-2	37.4	336.6	92.88	92.73
LC-ME1	10.0	90	58.43	58.14
LC-ME2	10.3	92.7	55.91	55.82
MC-ME1	48.4	435.6	54.51	54.38
MC-ME2	46.7	420.3	104.66	104.52
HC-ME1	58.5	526.5	151.2	150.93
HC-ME2	62.3	560.7	62.86	62.78

The initial quality review was performed on .fastq files using FastQC Version 0.11.8 (Andrews, 2010), with all samples passing per base sequence quality, per sequence quality scores, per base N content, and sequence length distribution. Fastq files were then trimmed with Trimmomatic Version 0.39 using the built-in adapters file TruSeq3-PE-2.fa, with settings LEADING:30, TRAILING:20, MAXINFO:50:0.5, MINLEN: 50, and remaining settings at default (Bolger et al., 2014) followed again by FastQC quality checks. Samples were then aligned using the 'align' command in the Rsubread (Liao et al., 2019) package, Version 2.4.2, with default settings, using R version 4.0.3 (R Core Team, 2020). Gene counts were tabulated using Rsubread 'featureCounts' command. The GRCh38.primary_assembly.genome.fa.gz and gencode.v36.primary_assembly.annotation.gtf.gz (Gencode Project, 2020a; 2020b) files were used during alignment and annotation.

Differential expression analysis was performed with edgeR version 3.36.0 (Robinson et al., 2010) in R version 4.1.2 (R Core Team, 2021). Counts tables generated by Rsubread's

'featureCounts' command were filtered to remove genes with expression below edgeR's recommended threshold, using 'filterByExpr', and trimmed mean of M-values normalized using edgeR. Principal components analysis (PCA) was performed using normalized and filtered counts comparing all samples (Figure 1). Contrast matrices were constructed as recommended by edgeR documentation, and DE was assessed using 'glmQLFit'. Differential expression counts for each comparison were totaled and compared to input concentrations sent for RNA sequencing to determine if DE was related to actual sample differences, input concentration, or other factors (Table 3). Genes shown to be differentially expressed were tabulated to determine if the same gene was shown to be differentially expressed in different comparisons, and checked for functionality using DAVID (Huang et al., 2009; Sherman et al., 2022; Supplementary Table 1) or biomaRt version 2.50.3 (Durinck et al., 2005; 2009; Table 4; Supplementary Table 2; Table 5; Supplementary Table 3; Supplementary Table 4), and data on gene biotype and expression levels within blood and other tissues were collected from Ensembl's Gene Expression tool (Ensembl West, 2022b; Table 6). Differentially expressed genes in the LC-A vs. LC-AB, LC-A vs. LC-AW, and LC-A vs. LC-ATB comparisons were also examined to determine if there was an overlap. Gene functionality and gene biotype were collected using biomaRt (Supplementary Table 5).

Differentially expressed gene lists were also compiled from the HC-ME vs. LC-ME and MC-ME vs. LC-ME comparisons and from the HC vs. LOW-15 and HC vs. LOW-30 comparisons to determine if any genes overlapped. BiomaRt was used to retrieve gene biotypes and functionality data for both the MinElute comparisons (HC-ME vs. LC-ME, MC-ME vs. LC-ME) and the unpurified comparisons (HC vs. LOW-15, HC vs. LOW-30) and listed in Supplementary Tables 6 and 7, respectively. Pearson correlations were determined using 'rcorr' from the R package Hmisc version 4.7-1 (Harrell & Dupont, 2021) between concentration ratio and the number of DE genes detected and between input concentration ratio and the number of raw reads. The normalized expression counts for the top 10 differentially expressed genes across all samples were plotted using ggplot2 v. 3.3.6 (Figure 2; Wickham, 2016). Differential expression comparisons between different purification methods with similar input concentrations (LC-A vs. LOW-30, LC-A vs. LOW-15, LC-A vs. MC-ME, HC vs. HC-ME, LOW-30 vs. MC-ME) were also tabulated and used to plot a five-way Venn diagram using the R package 'venn' (Dusa, 2021) to look for any overlap of genes found to demonstrate DE (Figure 3). Data from the five-way Venn diagram was used to produce an UpSet plot (Gehlenborg, 2019; Figure 3). A false discovery rate (FDR) threshold of less than 0.05 and a $\log_2$ fold change greater than $|\pm 1|$ were chosen as the criteria for determining differential expression in all comparisons.

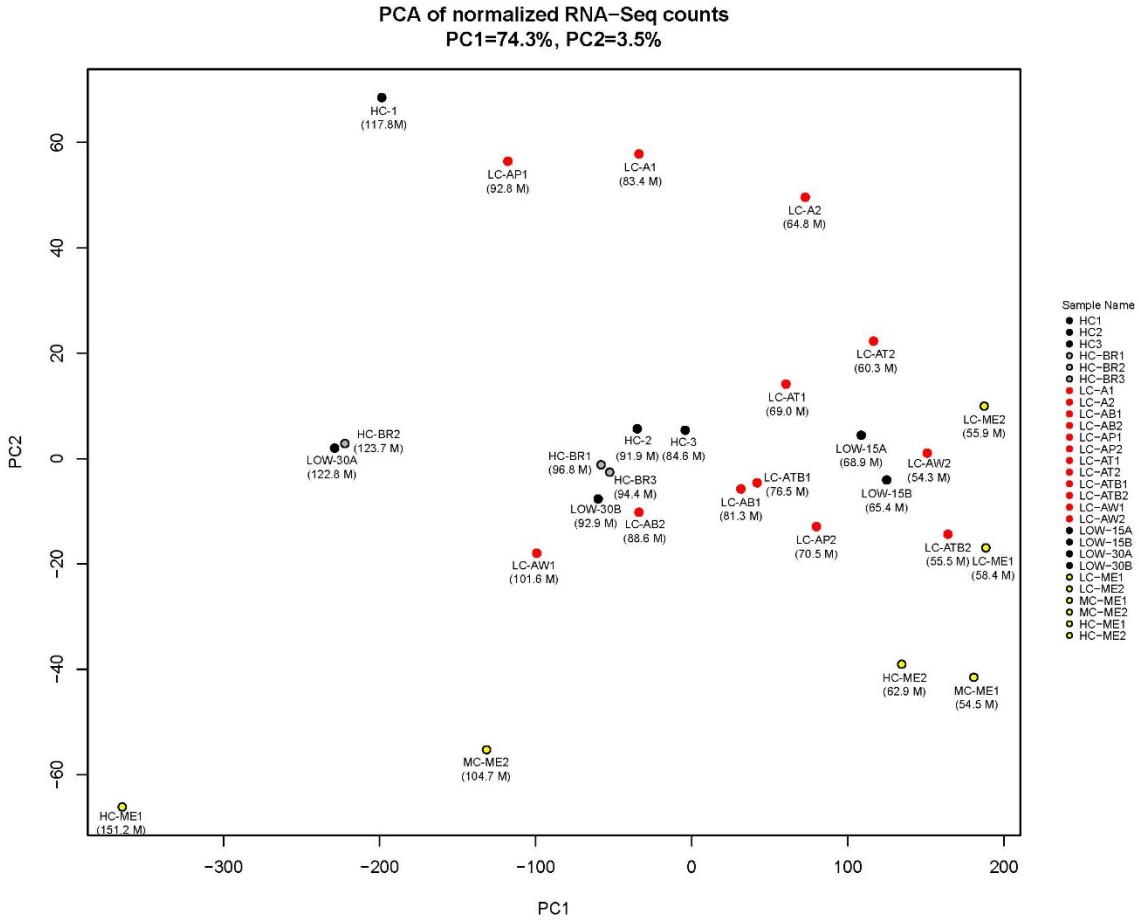

Figure 1.
Principal components analysis (PCA) of Normalized RNA-Seq Counts for All Samples. Raw library sizes given in parenthesis below each sample.

Table 3.
Differential Expression Counts and Library Preparation Input Concentrations for Each Comparison

Comparison (27,855 genes present after filtering)	Differentially expressed genes (LFC > |±1|, FDR < 0.05)	Sample Concentration Average After Extraction and Purification (ng/µL)	Input Concentration Ratio
Water-Eluted Unpurified High (HC) vs. BR5-Eluted Unpurified High (HC-BR)	0	74.2 vs. 74.5	0.996
Water-Eluted Unpurified High (HC) vs. Water-Eluted Unpurified Low 30 (LOW-30)	169	**74.2 vs. 37.4***	**0.504***
Water-Eluted Unpurified High (HC) vs. Water-Eluted Unpurified Low 15 (LOW-15)	91	**74.2 vs. 19.9***	**0.268***
Water-Eluted Unpurified Low 30 (LOW-30) vs. Water-Eluted Unpurified Low 15 (LOW-15)	0	37.4 vs. 19.9	0.532
Unpurified High (HC) vs. AMPure concentrated (LC-A)	57	**74.2 vs. 26.35***	**0.355***
AMPure (LC-A) vs. Unpurified Low (LOW-30)	250	26.35 vs. 37.4	0.705
AMPure (LC-A) vs. Unpurified Low (LOW-15)	55	26.35 vs. 19.9	0.755
AMPure (LC-A) vs. AMPure Time (LC-AT)	0	26.35 vs. 25.35	0.962
AMPure (LC-A) vs. AMPure Prior (LC-AP)	0	26.35 vs. 45.7	0.577
AMPure (LC-A) vs. AMPure Both (LC-AB)	148	**26.35 vs. 55.65***	**0.473***
AMPure (LC-A) vs. AMPure While (LC-AW)	101	26.35 vs. 37.8	0.697
AMPure (LC-A) vs. AMPure Time Both (LC-ATB)	91	26.35 vs. 35.9	0.734
AMPure (LC-A) vs. MinElute High (HC-ME)	577	**26.35 vs. 60.4***	**0.436***
AMPure (LC-A) vs. MinElute Medium (MC-ME)	675	26.35 vs. 47.55	0.554
AMPure (LC-A) vs. MinElute Low (LC-ME)	73	**26.35 vs. 10.15***	**0.385***
MinElute High (HC-ME) vs MinElute Medium (MC-ME)	0	60.4 vs. 47.55	0.787
MinElute High (HC-ME) vs. MinElute Low (LC-ME)	96	**60.4 vs. 10.15***	**0.168***
MinElute Medium (MC-ME) vs. MinElute Low (LC-ME)	183	**47.55 vs. 10.15***	**0.213***
MinElute High (HC-ME) vs. Unpurified High (HC)	181	60.4 vs. 74.2	0.814
MinElute Medium (MC-ME) vs. Unpurified Low (LOW-30)	58	47.55 vs. 37.4	0.787
MinElute Low (LC-ME) vs. Unpurified Low (LOW-15)	0	10.15 vs. 19.9	0.510

Note. Abbreviations: LFC = $\log_2$ fold change, FDR = false discovery rate.

*Input concentration ratios (lower conc. / higher conc.) in bold are < 0.51.

Table 4

Overlapping Differentially Expressed Genes Occurring in All AMPure (LC-A) vs. Unpurified (LOW-15, LOW-30, HC) Comparisons

Gene Designation	LC-A vs. LOW-15		LC-A vs. LOW-30		LC-A vs. HC		Gene Description
	LFC	FDR	LFC	FDR	LFC	FDR	
ENSG00000275219	-2.06	1.77E-05	-1.39	9.91E-04	1.21	3.54E-04	U2 spliceosomal RNA [Source:RFAM;Acc:RF00004]
ENSG00000274585	-2.00	6.35E-05	-1.86	1.41E-05	2.27	1.21E-06	RNA, U2 small nuclear 1 [Source:HGNC Symbol; Acc:HGNC:10142]
ENSG00000200959	-1.63	1.19E-04	-1.43	7.09E-05	1.78	2.04E-11	small nucleolar RNA, H/ACA box 74A [Source:HGNC Symbol; Acc:HGNC:10119]
ENSG00000274432	-2.27	2.47E-04	-2.19	5.06E-05	2.71	1.94E-09	U2 spliceosomal RNA [Source:RFAM; Acc:RF00004]
ENSG00000212402	-1.27	2.50E-04	-1.23	3.51E-05	1.20	3.54E-04	small nucleolar RNA, H/ACA box 74B [Source:HGNC Symbol ; Acc:HGNC:32660]
ENSG00000273709	-2.36	2.79E-04	-2.24	7.09E-05	2.96	1.69E-09	U2 spliceosomal RNA [Source:RFAM; Acc:RF00004]
ENSG00000278048	-1.82	3.40E-04	-1.88	5.06E-05	2.30	3.07E-09	U2 spliceosomal RNA [Source:RFAM; Acc:RF00004]
ENSG00000212443	-2.08	4.22E-04	-1.60	7.18E-04	2.03	5.23E-08	small nucleolar RNA, H/ACA box 53 [Source:HGNC Symbol; Acc:HGNC:32646]
ENSG00000277903	-2.16	4.50E-04	-2.00	1.84E-04	2.58	3.07E-09	U2 spliceosomal RNA [Source:RFAM; Acc:RF00004]
ENSG00000278774	-2.09	3.10E-03	-2.14	3.99E-04	2.80	3.07E-09	U2 spliceosomal RNA [Source:RFAM; Acc:RF00004]
ENSG00000212232	-1.47	1.69E-02	-1.61	2.32E-03	1.53	5.19E-04	small nucleolar RNA, C/D box 17 [Source:HGNC Symbol; Acc:HGNC:32713]

Note. LFC = $\log_2$ fold change, FDR = false discovery rate.

Table 5.

Overlapping Differentially Expressed Genes Occurring in All in AMPure (LC-A) vs. MinElute (LC-ME, MC-ME, HC-ME) Comparisons

Gene Designation	LC-A vs. LC-ME		LC-A vs. MC-ME		LC-A vs. HC-ME		Gene Description
	LFC	FDR	LFC	FDR	LFC	FDR	
ENSG00000280287	-1.10	2.22E-03	-1.25	1.62E-06	-1.44	5.57E-07	novel transcript
ENSG00000229422	-1.29	3.51E-03	-1.13	1.47E-04	-1.51	1.77E-06	novel transcript
ENSG00000276368	-1.36	3.64E-03	-1.27	7.94E-05	-1.40	1.80E-05	H2A clustered histone 14 [Source:HGNC Symbol;Acc:HGNC:4727]
ENSG00000212443	-1.70	3.64E-03	-1.18	1.39E-03	-1.28	7.42E-04	small nucleolar RNA, H/ACA box 53 [Source:HGNC Symbol;Acc:HGNC:32646]
ENSG00000277775	-1.50	6.94E-03	-1.29	4.04E-04	-1.42	1.85E-04	H3 clustered histone 7 [Source:HGNC Symbol;Acc:HGNC:4773]
ENSG00000007516	-1.20	8.14E-03	-1.32	2.26E-05	-1.35	2.23E-05	BAI1 associated protein 3 [Source:HGNC Symbol;Acc:HGNC:948]
ENSG00000278048	-1.29	8.19E-03	-1.89	3.18E-06	-1.64	2.14E-05	U2 spliceosomal RNA [Source:RFAM;Acc:RF00004]
ENSG00000275708	-3.62	9.95E-03	-4.09	5.89E-06	-3.65	8.03E-05	microRNA 3648-1 [Source:HGNC Symbol;Acc:HGNC:38941]
ENSG00000173801	-1.54	1.09E-02	-1.73	1.20E-05	-1.64	3.52E-05	junction plakoglobin [Source:HGNC Symbol;Acc:HGNC:6207]
ENSG00000277903	-1.44	1.34E-02	-2.03	1.28E-05	-1.70	9.71E-05	U2 spliceosomal RNA [Source:RFAM;Acc:RF00004]
ENSG00000281571	-1.21	1.34E-02	-1.39	1.43E-05	-1.50	5.03E-06	TEC
ENSG00000126895	-1.20	1.45E-02	-1.45	6.47E-06	-1.44	1.03E-05	arginine vasopressin receptor 2 [Source:HGNC Symbol;Acc:HGNC:897]
ENSG00000273709	-1.52	1.50E-02	-2.41	2.79E-06	-1.91	4.44E-05	U2 spliceosomal RNA [Source:RFAM;Acc:RF00004]
ENSG00000160446	-1.16	1.54E-02	-1.49	2.91E-06	-1.30	3.61E-05	zinc finger DHHC-type palmitoyltransferase 12 [Source:HGNC Symbol;Acc:HGNC:19159]
ENSG00000274432	-1.43	1.60E-02	-2.15	4.25E-06	-1.91	2.38E-05	U2 spliceosomal RNA [Source:RFAM;Acc:RF00004]
ENSG00000167513	-1.45	1.60E-02	-1.17	3.17E-03	-1.22	1.71E-03	chromatin licensing and DNA replication factor 1 [Source:HGNC Symbol;Acc:HGNC:24576]
ENSG00000065268	-1.00	1.60E-02	-1.15	3.03E-05	-1.07	8.56E-05	WD repeat domain 18 [Source:HGNC Symbol;Acc:HGNC:17956]
ENSG00000139725	-1.25	1.70E-02	-1.35	7.87E-05	-1.30	1.35E-04	ras homolog family member F, filopodia associated [Source:HGNC Symbol;Acc:HGNC:15703]
ENSG00000276116	-1.94	1.72E-02	-2.07	8.42E-05	-1.64	2.11E-03	FUT8 antisense RNA 1 [Source:HGNC Symbol;Acc:HGNC:44294]
ENSG00000226981	-1.71	2.03E-02	-1.90	7.13E-05	-1.36	4.26E-03	ABHD17A pseudogene 6 [Source:HGNC Symbol;Acc:HGNC:34044]

Gene Designation	LC-A vs. LC-ME		LC-A vs. MC-ME		LC-A vs. HC-ME		Gene Description
	LFC	FDR	LFC	FDR	LFC	FDR	
ENSG00000182685	-1.10	2.08E-02	-1.38	7.87E-06	-1.14	1.78E-04	BRICHOS domain containing 5 [Source:HGNC Symbol;Acc:HGNC:28309]
ENSG00000197238	-1.74	2.37E-02	-1.79	6.78E-04	-1.79	8.90E-04	H4 clustered histone 11 [Source:HGNC Symbol;Acc:HGNC:4785]
ENSG00000212232	-1.33	2.79E-02	-1.99	4.47E-05	-1.90	1.03E-04	small nucleolar RNA, C/D box 17 [Source:HGNC Symbol;Acc:HGNC:32713]
ENSG00000164880	-1.01	2.79E-02	-1.07	5.50E-04	-1.16	3.22E-04	integrator complex subunit 1 [Source:HGNC Symbol;Acc:HGNC:24555]
ENSG00000229873	-1.76	2.79E-02	-1.43	6.93E-03	-1.62	1.50E-03	OGFR antisense RNA 1 [Source:HGNC Symbol;Acc:HGNC:40724]
ENSG00000180535	-1.34	2.80E-02	-1.42	2.62E-04	-1.51	1.07E-04	basic helix-loop-helix family member a15 [Source:HGNC Symbol;Acc:HGNC:22265]
ENSG00000186940	-1.33	2.84E-02	-1.27	1.28E-03	-1.17	2.69E-03	coiled-coil-helix-coiled-coil-helix domain containing 2 pseudogene 9 [Source:HGNC Symbol;Acc:HGNC:23676]
ENSG00000104892	-1.24	2.84E-02	-1.17	1.46E-03	-1.20	1.02E-03	kinesin light chain 3 [Source:HGNC Symbol;Acc:HGNC:20717]
ENSG00000203852	-1.42	2.84E-02	-1.33	1.68E-03	-1.38	1.22E-03	H3 clustered histone 15 [Source:HGNC Symbol;Acc:HGNC:20505]
ENSG00000283206	-2.42	2.84E-02	-2.13	3.98E-03	-2.08	4.40E-03	microRNA 941-1 [Source:HGNC Symbol;Acc:HGNC:33684]
ENSG00000263969	-1.89	2.84E-02	-1.44	1.43E-02	-1.55	6.14E-03	RNA, 7SL, cytoplasmic 678, pseudogene [Source:HGNC Symbol;Acc:HGNC:46694]
ENSG00000249884	-2.02	3.06E-02	-1.85	2.62E-03	-1.67	6.86E-03	RNF103-CHMP3 readthrough [Source:HGNC Symbol;Acc:HGNC:38847]
ENSG00000271303	-2.34	3.14E-02	-1.73	2.56E-02	-1.67	2.88E-02	sulfiredoxin 1 [Source:HGNC Symbol;Acc:HGNC:16132]
ENSG00000196747	-1.06	3.17E-02	-1.01	1.98E-03	-1.11	9.52E-04	H2A clustered histone 13 [Source:HGNC Symbol;Acc:HGNC:4725]
ENSG00000276168	-1.22	3.17E-02	-1.07	4.97E-03	-1.13	4.14E-03	RNA component of signal recognition particle 7SL1 [Source:HGNC Symbol;Acc:HGNC:10038]
ENSG00000278463	-1.20	3.64E-02	-1.34	3.03E-04	-1.15	1.57E-03	H2A clustered histone 4 [Source:HGNC Symbol;Acc:HGNC:4734]
ENSG00000278774	-1.44	3.69E-02	-2.26	2.22E-05	-1.93	1.45E-04	U2 spliceosomal RNA [Source:RFAM;Acc:RF00004]
ENSG00000138080	-1.20	3.70E-02	-1.28	4.45E-04	-1.23	6.67E-04	elastin microfibril interfacer 1 [Source:HGNC Symbol;Acc:HGNC:19880]
ENSG00000241162	-1.44	3.71E-02	-1.00	3.51E-02	-1.62	2.15E-04	RNA, 7SL, cytoplasmic 617, pseudogene [Source:HGNC Symbol;Acc:HGNC:46633]

Gene Designation	LC-A vs. LC-ME		LC-A vs. MC-ME		LC-A vs. HC-ME		Gene Description
	LFC	FDR	LFC	FDR	LFC	FDR	
ENSG00000239224	-1.76	3.85E-02	-1.45	8.39E-03	-2.06	3.95E-04	RNA, 7SL, cytoplasmic 546, pseudogene [Source:HGNC Symbol;Acc:HGNC:46562]
ENSG00000063660	-1.10	4.03E-02	-1.10	1.34E-03	-1.12	9.18E-04	glypican 1 [Source:HGNC Symbol;Acc:HGNC:4449]
ENSG00000264916	-1.35	4.16E-02	-1.02	1.68E-02	-1.20	5.12E-03	RNA, 7SL, cytoplasmic 230, pseudogene [Source:HGNC Symbol;Acc:HGNC:46246]
ENSG00000265735	-1.20	4.21E-02	-1.23	2.38E-03	-1.34	1.56E-03	RNA, 7SL, cytoplasmic 5, pseudogene [Source:HGNC Symbol;Acc:HGNC:10040]
ENSG00000277483	-1.76	4.27E-02	-1.43	1.39E-02	-1.75	1.41E-03	RNA, 7SL, cytoplasmic 321, pseudogene [Source:HGNC Symbol;Acc:HGNC:46337]
ENSG00000274618	-1.02	4.43E-02	-1.07	1.86E-03	-1.06	2.09E-03	H4 clustered histone 6 [Source:HGNC Symbol;Acc:HGNC:4783]
ENSG00000183598	-1.46	4.48E-02	-1.32	5.32E-03	-1.32	5.11E-03	H3 clustered histone 13 [Source:HGNC Symbol;Acc:HGNC:25311]
ENSG00000263740	-1.43	4.52E-02	-1.88	4.56E-04	-1.71	1.19E-03	RNA, 7SL, cytoplasmic 4, pseudogene [Source:HGNC Symbol;Acc:HGNC:10039]
ENSG00000169962	-1.56	4.59E-02	-1.59	1.32E-03	-1.93	9.57E-05	taste 1 receptor member 3 [Source:HGNC Symbol;Acc:HGNC:15661]

Note. LFC = $\log_2$ fold change, FDR = false discovery rate.

Table 6.
Ensembl Gene Expression Results for Genes found to be DE in 11, 10, 9, or 8 Comparisons

Number of Comparisons with DE Results	Ensembl Gene ID	Ensembl Expression in Blood, TPM/FPKM (No. Studies With Data)	Ensembl Expression Information Across All Tissues
11	ENSG00000186940	no data	Low to none across all tissues
11	ENSG00000226981	0 (1)	Low to none across all tissues
10	ENSG00000158106	4 (1)	None to medium across all tissues
10	ENSG00000159884	5 (1)	None to medium across all tissues
10	ENSG00000201955	no data	None to medium across all tissues
10	ENSG00000274585	no data	None to low across all tissues
10	ENSG00000278774	no data	None to low across all tissues
9	ENSG00000201098	0 (1)	None to high across tissues
9	ENSG00000202538	2 (1)	None to medium across all tissues
9	ENSG00000207513	no data	None to low across all tissues
9	ENSG00000239039	no data	None to medium across all tissues
9	ENSG00000259932	0 (1)	None to low across all tissues
9	ENSG00000273768	0 (1)	None to low across all tissues
8	ENSG00000104892	2 (1)	None to medium across all tissues
8	ENSG00000183598	0.5 (1)	None to low across all tissues
8	ENSG00000186827	6 (1)	None to medium across all tissues
8	ENSG00000200795	0 (1)	None to medium across all tissues
8	ENSG00000201998	0 (1)	None to medium across all tissues
8	ENSG00000203852	no data	None to low across all tissues
8	ENSG00000206596	no data	None to below cutoff across all tissues
8	ENSG00000207005	no data	None to medium across all tissues
8	ENSG00000207142	0 (1)	None to medium across all tissues
8	ENSG00000208892	no data	None to medium across all tissues
8	ENSG00000215914	no data	None to low across all tissues
8	ENSG00000260035	0 (1)	None to low across all tissues
8	ENSG00000273709	no data	None to low across all tissues
8	ENSG00000275708	no data	None to low across all tissues
8	ENSG00000276368	0 (1)	None to medium across all tissues

Note. TPM = transcripts per million; FPKM = fragments per kilobase of exon model per million reads map.

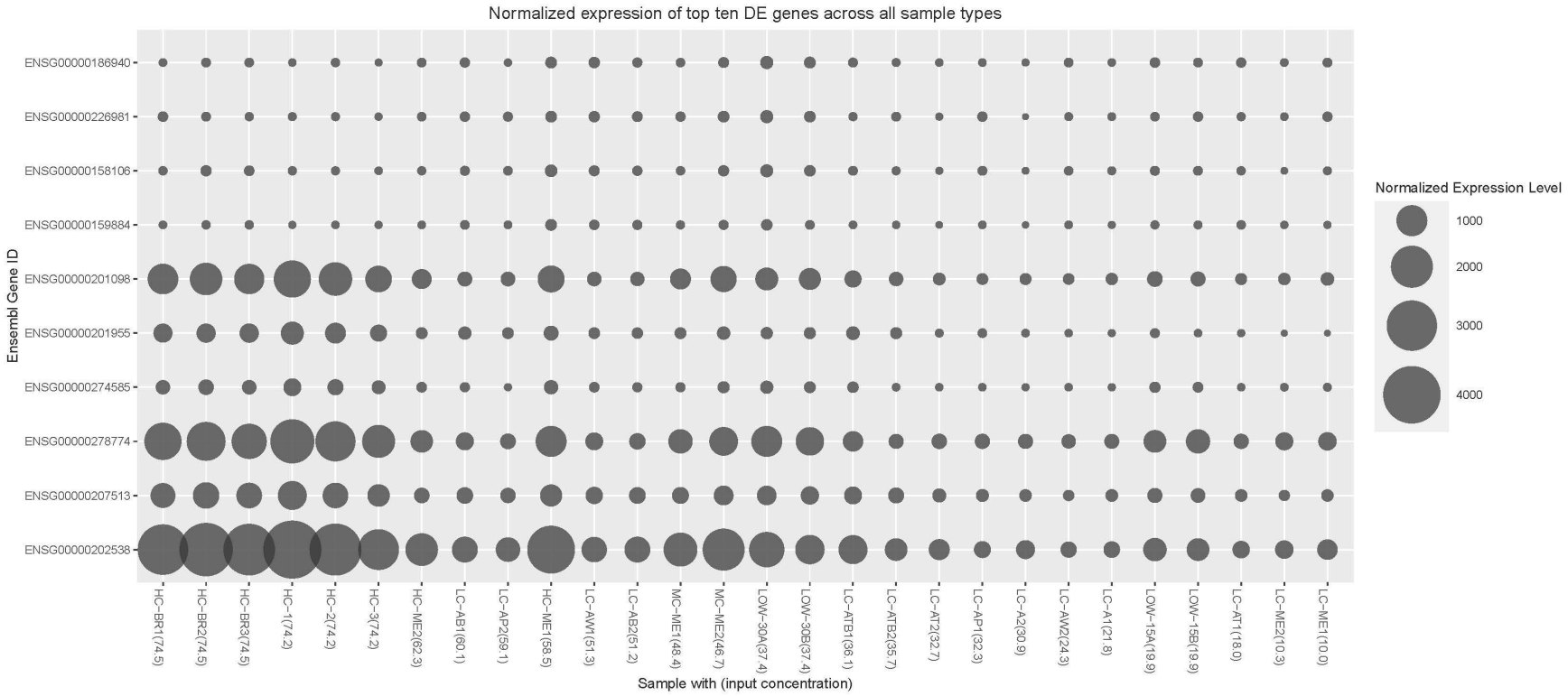

Figure 2.
Sample vs. Normalized Expression Count for the Top Ten Genes Most Often Differentially Expressed

Note. Samples on the x-axis are in descending order, from greatest input concentration to lowest input concentration.

Overlap of differentially expressed genes between differing purification methods

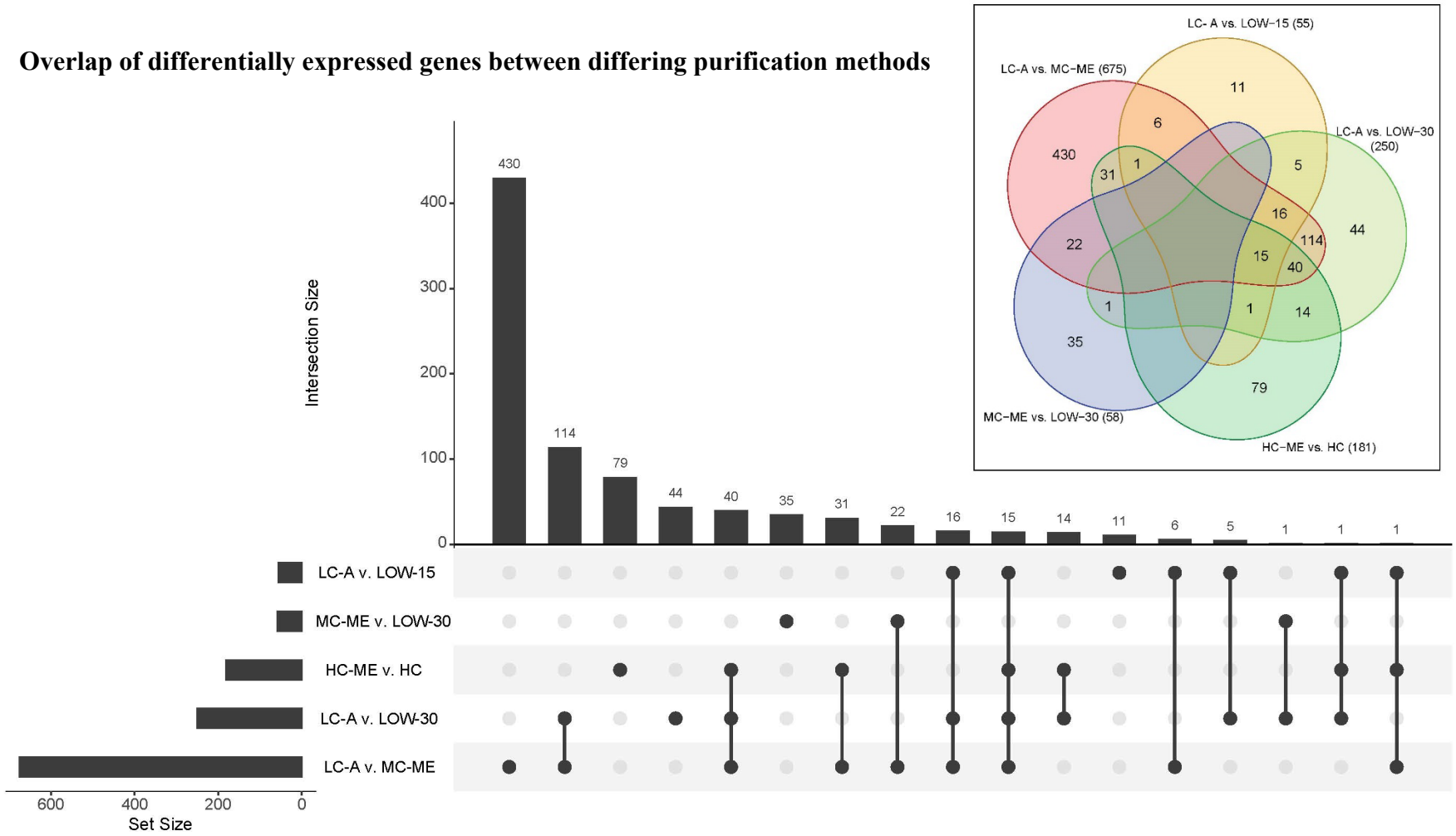

Figure 3.
Five-way Venn Diagram and UpSet Plot Between Unpurified, AMPure Purified, and MinElute Purified Samples Demonstrating Differential Expression

Note. Total number of differentially expressed genes is given with each comparison in parentheses. The LC-A vs. LC-ME, and LC-ME vs. LOW-15, are not included because neither comparison yielded any differentially expressed genes.

Results

Differential expression and input concentration

Differential expression counts are listed for each comparison (Table 3). Fifteen of 21 total comparisons demonstrated DE. Of the 15 comparisons that showed differential expression, there were eight pairwise comparisons (HC vs. LOW-30, HC vs. LOW-15, HC vs. LC-A, LC-A vs. LC-AB, LC-A vs. HC-ME, LC-A vs. LC-ME, HC-ME vs. LC-ME, and MC-ME vs. LC-ME) where the RNA sample concentration after purification showed the more dilute sample having half or less of the concentration value of the more concentrated sample (input RNA concentration for library preparation with a ratio < 0.51; Table 3). The relationship between input concentration ratio and DE count was not linear, as there were instances (HC vs. LOW-30) with greater counts of differential expression between samples with a low input concentration ratio than between samples with a higher ratio. Indeed, a simple Pearson correlation analysis across all 21 comparisons relating the number of differentially expressed genes to the concentration ratio was not significant (P=0.36). There were thirteen comparisons where the input concentration ratio was 0.51 or greater. Six did not show DE, and seven comparisons (LC-A vs. LOW-30, LC-A vs. LOW-15, LC-A vs. AW, LC-A vs. LC-ATB, and LC-A vs. MC-ME, HC-ME vs. HC, MC-ME vs. LOW-30) showed differential expression.

Principal components analysis (PCA) was run on normalized counts for all samples (Figure 1). Principal components 1 and 2 were responsible for 74.3% and 3.5% of the observed variation, respectively. Approximately three-fourths of the observed variation does occur along the first principal component axis, and distribution along PC1 roughly appears to correspond with the number of raw sequenced reads (Table 2). Moreover, the number of raw (and trimmed) sequencing reads significantly correlates to the input mass concentration sent for sequencing (Pearson correlation coefficient = 0.54; P=0.003). Counts were tabulated for all comparisons demonstrating DE to determine how frequently the same genes were shown to be differentially expressed. The top 10 genes most frequently detected as differentially expressed across comparisons were plotted according to sample vs. normalized expression counts. Samples on the x-axis were placed in descending order of input concentration and normalized read counts were placed on the y-axis. While expression varies from gene to gene, it generally follows a trend with greatest input concentration to lowest also having the greatest expression count to the least expression count (Figure 2).

Characteristics of differentially expressed genes

Results of DAVID searches for genes (Supplementary Table 1) from all 15 comparisons showing DE in eight or more contrasts do not indicate an overall biological role in cellular function or pathway. The RNA functions, or biotypes, described in Ensembl West (2022c) will be used throughout this publication for clarity and consistency. Twenty-eight genes were differentially expressed in at least eight comparisons. The function of those genes varied, with 13 of them being small nuclear RNAs (snRNAs), microRNAs (miRNAs), or spliceosomes, five

having no known function, three were involved in protein binding, three were involved in histone functions, two were pseudogenes, and the remaining four genes had varying cellular functions. Seven comparisons considered in the DAVID search are between different concentrations within the same method, such as LC-A vs. LC-AB or HC vs. LOW-30. The other eight comparisons are between samples prepared using different methods, such as HC vs. LC-A, MC-ME vs. LOW-30, and LC-A vs. LC-ME. Only one of the genes considered in the DAVID search, ENSG00000273709, is found to occur almost entirely in comparisons between methods, including HC vs. LC-A, HC-ME vs. HC, LC-A vs. HC-ME, LC-A vs. LC-ME, LC-A vs. LOW-15, LC-A vs. LOW-30, and LC-A vs. MC-ME. The gene ENSG00000273709 was also differentially expressed in the same method comparison HC vs. LOW-30. Ensembl notes that gene is a U2 spliceosomal RNA (Ensembl West, 2022a). Ensembl's gene expression tool (Ensembl West, 2022b) was also used (Table 6) to estimate relative gene expression in blood and across all tissues. Of 28 DE genes found in eight or more comparisons in the present study, 13 were found in the Ensembl search to have no blood expression data but were found to be expressed in other tissues. The expression level of the remaining fifteen genes ranged from zero to six TPM/FPKM (transcripts per million / fragments per kilobase of exon model per million reads mapped) in blood which is considered a low expression level in the Ensembl gene expression atlas.

Comparisons between AMPure and MinElute purification methods

The comparisons of greatest interest in determining the effects of sample concentration are between differing methods, such as comparing AMPure bead purification and MinElute column purification. Lists of differentially expressed genes from contrasts that compared two different preparations were combined to determine if there was an overlap between the different preparation methods. Comparisons to AMPure purification were made using data from LC-A samples so that comparisons between procedures, such as comparing MinElute column purification to AMPure bead purification, could be made between unmodified procedures. LC-A was compared to LOW-15, LOW-30, and HC (Table 4; Supplemental Table 2) to determine if the same group of genes were differentially expressed when comparing methods. LC-A was also compared to LC-ME, MC-ME, and HC-ME (Table 5; Supplemental Table 3). The varying concentrations of starting material used for MinElute purifications allowed for comparisons which were more similar in input concentration (Table 1), so HC-ME was compared to HC, MC-ME was compared to LOW-30, and LC-ME was compared to LOW-15 (Supplemental Table 4). For each of these groupings, lists of differentially expressed genes were tabulated to determine which genes, if any, were determined to be differentially expressed in one, two, or all three comparisons made between each pair of methods.

Comparisons between similar input concentrations and differing methods

Differential expression was observed between AMPure-purified samples and those without purification. The comparisons between LC-A vs. LOW-15, LC-A vs. LOW-30, and LC-A vs. HC did produce 11 differentially expressed genes common to all three contrasts

(Supplemental Table 2). Six genes were U2 spliceosomal RNAs, and the remaining five were snRNAs or small nucleolar RNAs (snoRNAs). Thirty-seven DE genes were common to two of three of the AMPure vs. unpurified comparisons, including eight snRNAs or snoRNAs and one U2 spliceosomal RNA. The remaining 28 differentially expressed genes had varied gene biotype descriptions. There were also 255 genes differentially expressed in only one of the comparisons made between LC-A vs. LOW-15, LC-A vs. LOW-30, or LC-A vs. HC, including 19 snRNA or snoRNAs and three additional U2 spliceosomal RNAs.

Likewise, differences were observed in AMPure vs. MinElute purification tests, with LC-A vs. HC-ME and LC-A vs. MC-ME showing the most DE of all 21 comparisons. The comparisons between LC-A vs. LC-ME, LC-A vs. MC-ME, and LC-A vs. HC-ME produced 49 differentially expressed genes common to all three contrasts (Supplemental Table 3). There is considerably more variation in the gene descriptions for the 49 genes in common. Eight genes are varying clustered histone genes. Seven are 7SL cytoplasmic RNAs. Five are U2 spliceosomal RNAs, and two are snRNAs, like those 11 genes found to be differentially expressed in all the AMPure vs. unpurified comparisons. The remaining 27 genes commonly detected in all three comparisons varied in their biotypes. There were 315 genes differentially expressed in two of three comparisons, including 27 snRNAs or snoRNAs and three spliceosomal RNAs. A total of 548 genes were differentially expressed in a single comparison. Thirty-five genes were either snRNAs or snoRNA genes. Two of the genes were also U2 spliceosomal RNAs.

DE was again detected in tests of MinElute vs. no purification. There were no differentially expressed genes detected in the LC-ME vs. LOW-15 comparison. The comparisons between HC-ME vs. HC and MC-ME vs. LOW-30 did produce 182 and 58 differentially expressed genes, respectively. No differentially expressed genes were found in common between the HC-ME vs. HC and MC-ME vs. LOW-30 comparisons (Supplemental Table 4). In the HC-ME vs. HC comparison, 15 of the 182 genes were snRNAs or snoRNAs, and six were U2 spliceosomal RNAs. Eight differentially expressed genes detected in the MC-ME vs. LOW-30 comparison were snRNA or snoRNAs.

Differential expression comparisons within AMPure variations

Three comparisons using the AMPure purification procedure variations demonstrated differential expression. Tests of LC-A vs. LC-AT and LC-A vs. LC-AP showed no DE. However, LC-A vs. LC-AB, LC-A vs. LC-AW, and LC-A vs. LC-ATB had 148, 101, and 91 differentially expressed genes, respectively (Table 3). The LC-A vs. LC-AB had an input concentration ratio of 0.473, below the 0.51 threshold. LC-A vs. LC-AW and LC-A vs. LC-ATB had input concentration ratios of 0.697 and 0.734, respectively, both above the 0.51 threshold. There were six genes found to be differentially expressed in all three comparisons (Supplementary Table 5). Three were protein-coding genes, and three were pseudogenes. Thirty-one differentially expressed genes overlap the LC-A vs. LC-AB and the LC-A vs. LC-AW comparisons. There are six additional differentially expressed genes that overlap the LC-A vs. LC-AB and LC-A vs. LC-ATB comparisons. There are not any differentially expressed genes

that overlap, only LC-A vs. LC-AW and LC-A vs. LC-ATB. Of the 37 differentially expressed genes overlapping two comparisons, 24 are protein-coding genes, five are long non-coding RNA (lncRNA), two are miscellaneous RNA (miscRNA), two are unprocessed pseudogenes, and one each of miRNA, processed pseudogene, snoRNA, and snRNA.

Differential expression comparisons within MinElute variations

Differential expression was observed within some of the MinElute comparisons. MinElute comparisons of HC-ME vs. LC-ME and MC-ME vs. LC-ME did demonstrate DE with 96 and 183 DE genes, respectively (Table 3). The HC-ME vs. MC-ME comparison did not produce any differentially expressed genes. The comparisons HC-ME vs. LC-ME and MC-ME vs. LC-ME had input concentration ratios of 0.168 and 0.213, respectively. The comparison that did not produce differentially expressed genes, HC-ME vs. MC-ME, had an input concentration ratio of 0.787. Twenty-five overlapping DE genes were found in the HC-ME vs. LC-ME and the MC-ME vs. LC-ME comparisons (Supplementary Table 6). The overlapping genes were six lncRNA, one miRNA, two miscRNA, four processed pseudogene, six protein coding, one ribosomal RNA (rRNA) pseudogene, three snoRNA, one transcribed unprocessed pseudogene, and one unprocessed pseudogene biotype.

Differential expression comparisons within unpurified variations

Differential expression was demonstrated between unpurified sample contrasts HC vs. LOW-15, with 91 DE genes, and HC vs. LOW-30, with 169 DE genes, but not in the LOW-15 vs. LOW-30 comparison (Table 3). Both comparisons showing DE had input concentration ratios lower than the 0.51 threshold. There were 46 genes found to overlap between the two sets of comparisons, HC vs. LOW-15 and HC vs. LOW-30 (Supplementary Table 7). Roughly 30% of the overlapping genes, or 14 out of 46, were snRNA biotype genes, and seven were snoRNA biotype genes. There were also five lncRNA, six miscRNA, one processed pseudogene, nine protein coding, one rRNA pseudogene, one transcribed unprocessed pseudogene, and two unprocessed pseudogenes.

Differential expression comparisons across all purification methods

To determine if there were differentially expressed genes common to all three preparation methods, lists of DE genes from those comparisons between methods were combined and evaluated to determine which differentially expressed genes overlapped between comparisons. To reduce the factors potentially affecting samples chosen for analysis, the comparisons selected for this analysis were restricted to only include those with similar input concentrations. The comparison HC vs. LC-A was excluded because of the difference in input concentration between the AMPure purified samples (LC-A) and the unpurified high-concentration samples (HC). LOW-30 and LOW-15 were more similar, in terms of their input concentrations, to the AMPure purified sample (LC-A), so these were deemed to make a better comparison to the AMPure purified sample. The AMPure purified sample (LC-A) vs. MinElute High (HC-ME) and the MinElute Medium (MC-ME) vs. unpurified low 15 (LOW-15) comparisons were also excluded

because of differences in input concentration. The AMPure purified sample (LC-A) vs. MinElute Low (LC-ME) and MinElute Low (LC-ME) vs. unpurified Low 15 (LOW-15) did not show DE and could not be included. Gene lists from the remaining comparisons of different methods were used to construct a five-way Venn diagram to show overlap between the comparisons LC-A vs. MC-ME, LC-A vs. LOW-30, LC-A vs. LOW-15, MC-ME vs. LOW-30, and HC-ME vs. HC. There are not any genes that were found to overlap between all five comparisons. There are, however, 15 genes found to overlap between LC-A vs. MC-ME, LC-A vs. LOW-15, LC-A vs. LOW-30, and HC-ME vs. HC. These four comparisons include an AMPure vs. MinElute (LC-A vs. MC-ME), two AMPure vs. unpurified methods (LC-A vs. LOW-15, LC-A vs. LOW-30), and a MinElute vs. unpurified method (HC-ME vs. HC), which are all the possible comparisons between methods. Additionally, a total of 865 genes were found to be differentially expressed in at least one comparison, with the majority of those (n=430) occurring solely in the LC-A vs. MC-ME comparison (Figure 3).

Discussion

Input concentration

Despite analytical normalization steps, the concentration of material used for library preparation and sequencing appeared to influence downstream results. The water-eluted unpurified high-concentration (HC) vs. BR5-eluted unpurified high concentration (HC-BR) were very similar in their input concentration and showed no DE for any genes tested. Comparisons with distinct differences in initial concentration but prepared with the same purification method, such as HC vs. LOW-30, HC vs. LOW-15, HC-ME vs. LC-ME, and MC-ME vs. LC-ME did show differential expression (Table 3, Supplementary Table 6, Supplementary Table 7). These differences occurred despite sample concentrations generally falling within the recommended library preparation threshold (100 to 1000 ng). Only the LC-ME1 and LC-ME2 samples were outside of this threshold, with input masses of 90 and 92.7 ng, respectively (Table 2). The HC vs. LOW-30 and HC vs. LOW-15 were handled identically, other than the added dilution of the LOW-15 and LOW-30 samples. LC-ME and MC-ME only differed from HC-ME by dilution factor. Previous reports have noted that differences in sample quality and methods designed to work with low-input samples can introduce bias to RNA-Seq studies (Adiconis et al., 2013; Parkinson et al., 2012; Shi et al., 2021). In the present study, despite normalization to account for library size differences, samples largely appeared separated by sample concentration in principle component analysis (Figure 1). Notably, DE is generally detected in comparisons (HC vs. LOW-30, HC vs. LOW-15, HC vs. LC-A, LC-A vs. LC-AB, LC-A vs. HC-ME, LC-A vs. LC-ME, HC-ME vs. LC-ME, and MC-ME vs. LC-ME) where the average input concentration of one sample is half, or less than half, the other sample (Table 3).

All comparisons where the concentration input ratio is < 0.51 detect DE. This supports the inference that large variation in input concentration risks the introduction of technical variation in DE analyses. For samples with input concentration ratios ≥ 0.51, DE was detected in

some cases but not all. This may suggest that factors other than input concentration become more important when RNA quantities are similar. Two identical samples may produce different expression levels simply through stochastic differences that occur during library preparation and sequencing. In principle component analysis (Figure 1), there was no readily apparent clustering of sample types or technical replicates. Alignment along axis 1, which explains approximately 75% of the variance, appears related to library size and input concentration. Based on the PCA analysis, one possible interpretation is that the purification method had less effect on the sample than on the effect of library size. Stochastic differences are more likely to occur as the sample concentration drops, particularly at low sequencing depths.

While differential expression was detected amongst some of the comparisons made (Table 3), it is important to note that these comparisons were made across 27,855 genes retained after normalization and filtering. A false positive threshold of 5% was used as a cutoff for FDR, which means that some of the genes 'detected' as differentially expressed may not be genuinely differentially expressed. It has been recently suggested that both edgeR and DESeq2 may have false discovery rates greater than the expected 5% (Li et al., 2022). Many packages, including edgeR, implement the Benjamini-Hochberg method of determining FDR. Although it may be an improvement over previous methods utilizing family-wise error rate, using a 5% FDR cutoff does not eliminate all false positives from an analysis (Benjamini & Hochberg, 1995).

Comparisons between purification methods

Comparisons between preparation methods demonstrated DE, even when it may not have been expected. In the AMPure purification vs. unpurified sample comparisons between LC-A vs. LOW-15, LC-A vs. LOW-30, and LC-A vs. HC, the input concentration ratios are 0.755, 0.705, and 0.355, respectively (Table 3). While in the LC-A vs. HC comparison, differences in their input concentrations may have contributed to findings of differential expression, for the LC-A vs. LOW-15 and LC-A vs. LOW-30 comparisons with similar input concentrations, it appears likely that differences reflect the impact of adding an AMPure purification step. All three of these comparisons detected DE. The LC-A samples were diluted from the HC sample and then purified and concentrated using AMPure magnetic beads. One way in which AMPure purification may have impacted results is via size selection. At the ratio used in this study, AMPure magnetic beads are known to preferentially bind nucleic acid greater than 100 bp long (Beckman Coulter, 2016; 2022); thus, transcripts < 100 bp may be reduced during purification. Magnetic bead purification is also used multiple times in typical TruSeq kit library preparation, including the processing of samples in the current study. Therefore, the additional bead purification performed on the LC-A samples may not have been anticipated to have any discernable effect on DE, but the comparisons between LC-A and HC, LOW-30, and LOW-15 all demonstrated differential expression. Each of the genes common to all three AMPure purification vs. unpurified sample comparisons are either snRNA or snoRNA molecules (Supplemental Table 2). When comparing samples of similar concentration, LC-A vs. LOW-30, where one was AMPure purified, and the other was not, we see the greatest number of

differentially expressed genes. This indicates that AMPure purification may alter transcript levels sufficiently to detect DE, potentially influencing study findings.

Although AMPure vs. no purification comparisons typically resulted in DE for small molecules, the types of genes differentially expressed in comparisons of MinElute vs. AMPure were more varied. Forty-nine genes were differentially expressed in all three comparisons between AMPure bead purification and MinElute column purification of varied biotypes (Supplemental Table 3). Roughly half, or 25 of 49 genes, were protein-coding genes. Smaller RNA molecules, including miRNA, snoRNA, and snRNA make up approximately 35%, or 17 of 49, genes identified as differentially expressed in all three MinElute vs. AMPure comparisons. Considering that all three MinElute vs. AMPure comparisons demonstrated DE and two had low ratios while another had a ratio above 0.51, it is possible to conclude that both input concentration differences and sample handling differences contributed to the DE observed.

Differential expression was observed with some of the AMPure purification mixing and incubation variations. The two comparisons, LC-A vs. LC-AT and LC-A vs. LC-AP, showed no DE, and both had input concentration ratios > 0.51 (Table 3). Three comparisons (LC-A vs. LC-AB, LC-A vs. LC-AW, and LC-ATB) demonstrated DE. The common factor in all comparisons where DE was detected was mixing while on the magnet. Sample mixing, as tested, increased yields. Mixing prior to placing the sample on the magnet increased yields without evidence of DE, suggesting that mixing prior to placement on the magnet is advisable.

Conclusion

This study aimed to evaluate the impact of purification approaches on the results of RNA sequencing. The use of the elution buffer BR5, as provided with the QIAGEN's PAXgene Blood miRNA kit vs. water, did not substantially impact findings, suggesting they may be used interchangeably. However, input concentration and selection of a purification approach had larger effects. The quantity of RNA used for library preparation was seen here to correlate with the quantity of sequenced reads. While analytical approaches to normalize based on library size may mitigate such effects, differences related to input concentration still appeared to affect the DE produced. Figure 1 demonstrates that normalized data from the samples tested in this experiment still display separation appearing to associate with input concentration. Previous studies have also noted differing input concentrations as a factor affecting RNA-Seq results (Adiconis et al., 2013; Parkinson et al., 2012; Shi et al., 2021). When differences in sample concentration cause variations in ratios between genes, researchers may find DE where it is not anticipated, as with the HC vs. LOW-30 and HC vs. LOW-15 comparisons (Table 3). In future biomarker discovery projects where DE is detected between treatment groups with widely varying input concentrations, care may be needed when attributing differences to a treatment effect due to the potential confounding influence of sample concentration.

Purification protocols that permit elution in a reduced volume provide a means of concentrating samples but may introduce additional sources of variation. In this study, there were

detectable differences between the AMPure purified samples and the unpurified samples, the AMPure purified samples and the MinElute purified samples, and between the MinElute purified samples and the unpurified samples. Smaller transcripts were commonly differentially expressed in purification methods, but larger and protein-coding genes were also DE in some comparisons. Nonetheless, relating the purification method to DE in this study was confounded by the differences in input concentration between many of the samples.

With the potential for DE findings from even slight variations in the implementation of a purification approach, such as use of additional mixing or incubation steps in bead purification, this study supports the generally recognized principle of the importance of consistency in sample handling. Meta-analyses seeking to compare results across RNA-Seq studies may benefit from a review of sample processing approaches and consideration of the potential biases introduced by variations in RNA processing. In the present study, both AMPure and MinElute purifications were associated with DE based on comparison with unpurified samples from the same source RNA. This study does not find a strong advantage of one approach over the other. Because the largest numbers of differentially expressed genes were seen in comparisons of MinElute vs. AMPure purifications (LC-A vs. MC-ME and LC-A vs. MC-ME), as opposed to comparisons of unpurified samples with either AMPure or MinElute concentrations, it may be inferred that only a single purification approach should be used in a given project to avoid technical bias.

Levels of DE between unpurified samples of varying concentrations (comparisons of HC vs. LOW-30 and LOW-15 with 169 and 91 differentially expressed genes) were roughly comparable to that observed between unpurified samples, and AMPure purified (55 to 250 differentially expressed genes), or MinElute purified (0 to 181 differentially expressed genes). Hence, introducing a concentration step to reduce input quantity variation may simply represent a trade-off of one source of processing bias for another. Pending time and budget considerations when large numbers of samples must be processed, such concentration procedures may not add a considerable benefit. In samples where the source RNA varies in quality, the ability of either silica or bead-based purification to remove contaminants such as salts may improve results and warrant the extra effort. The impact of sample quality was not tested in the current study and would be worth evaluation in follow-on methodological investigations.

References

Adiconis, X., Borges-Rivera, D., Satija, R., DeLuca, D. S., Busby, M. A., Berlin, A. M., Sivachenko, A., Thompson, D. A., Wysoker, A., Fennell, T., Gnirke, A., Pochet, N., Regev, A., & Levin, J. Z. (2013). Comprehensive comparative analysis of RNA sequencing methods for degraded or low input samples. *Nature Methods, 10*(7), 623-629. https://doi.org/10.1038/nmeth.2483

Alderton, R. P., Eccleston, L. M., Howe, R. P., Read, C. A., Reeve, M. A., & Beck, S. (1992). Magnetic bead purification of M13 DNA sequencing templates. *Analytical Biochemistry*, *201*(1), 166-169. https://doi.org/10.1016/0003-2697(92)90190-I

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. [Computer software]. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Beckman Coulter (2016). Agencourt AMPure XP PCR purification manual. https://www.beckmancoulter.com/wsrportal/techdocs?docname=B37419

Beckman Coulter (2020). *AMPure XP Reagent for PCR Purification*. https://www.beckman.com/reagents/genomic/cleanup-and-size-selection/pcr?utm_medium=cpc&utm_source=google&utm_campaign=ecommerce-lead-gen-ab-test-amp-fp-exact&utm_content=geno_ecommerce_bottom&creative=438890571804&keyword=agencourt%20ampure%20xp%20beads&matchtype=p&network=g&device=c&gclid=EAIaIQobChMIlrzonN6W9wIVhv_jBx20WgszEAAYASAAEgJuxfD_BwE#techdocs-wrapper

Beckman Coulter (2022) *AMPure XP Capabilities*. https://www.beckman.com/reagents/genomic/cleanup-and-size-selection/pcr/ampure-xp-capabilities

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Berensmeier, S. (2006). Magnetic particles for the separation and purification of nucleic acids. *Applied Microbiology and Biotechnology*, *73*(3), 495-504. https://doi.org/10.1007/s00253-006-0675-0

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114-2120. https://doi.org/10.1093/bioinformatics/btu170

Boom, R., Sol, C. J., Salimans, M. M., Jansen, C. L., Wertheim-van Dillen, P. M., & van der Noordaa, J. (1990). Rapid and simple method for purification of nucleic acids. *Journal of Clinical Microbiology*, *28*(3), 495-503. https://doi.org/10.1128/jcm.28.3.495-503.1990

Diefenbach, R. J., Lee, J. H., Kefford, R. F., & Rizos, H. (2018). Evaluation of commercial kits for purification of circulating free DNA. *Cancer Genetics*, *228*, 21-27. https://doi.org/10.1016/j.cancergen.2018.08.005

Durinck, S., Spellman, P., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols, 4*(8), 1184-1191. https://doi.org/10.1038/nprot.2009.97

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., & Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics, 21*(16), 3439-3440. https://doi.org/10.1093/bioinformatics/bti525

Dusa, A. (2021). Venn: draw Venn diagrams. R package version 1.10.

Ensembl West. (Jul. 2022a). *Gene: U2 ENSG00000273709.* https://uswest.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000273709;r=17:43241180-43241370;t=ENST00000616535

Ensembl West. (Jul. 2022b). *Search.* https://uswest.ensembl.org/index.html

Ensembl West. (Oct. 2022c). *Biotypes*. https://uswest.ensembl.org/info/genome/genebuild/biotypes.html

Gehlenborg, N. (2019). UpSetR: a more scalable alternative to Venn and Euler diagrams for visualizing intersected sets. R package version 1.4.0.

Gencode Project. (2020a, November). *GRCh38.primary_assembly.genome.fa.gz* [Data file]. http://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_36/

Gencode Project. (2020b, November). *gencode.v36.primary_assembly.annotation.gtf.gz* [Data file]. http://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_36/

Gonzalez-Perez, I., Armas Cayarga A., García de la Rosa, I., & González González, Y. J. (2007). Homemade viral RNA isolation protocol using silica columns: a comparison of four protocols. *Analytical Biochemistry*, *360*(1), 148-150. https://doi.org/10.1016/j.ab.2006.10.022

Harrell, F. E. & Dupont, C. (2021). Hmisc: Harrell Miscellaneous. R package version 4.7.1.

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols, 4*(1), 44-57. https://doi.org/10.1038/nprot.2008.211

Jorgez, C. J., Dang, D. D., Simpson, J. L., Lewis, D. E., & Bischoff, F. Z. (2006). Quantity versus quality: optimal methods for cell-free DNA isolation from plasma of pregnant women. *Genetics in Medicine*, *8*(10), 615-619. https://doi.org/10.1097/01.gim.0000241904.32039.6f

Kim, J., Johnson, M., Hill, P., & Gale, B. K. (2009). Microfluidic sample preparation: cell lysis and nucleic acid purification. *Integrative Biology*, *1*(10), 574-586. https://doi.org/10.1039/b905844c

Lee, J. H., Park, Y., Choi, R. J., Lee, E. K., & Kim, H. S. (2010). Comparisons of three automated systems for genomic DNA extraction in a clinical diagnostic laboratory. *Yonsai Medical Journal*, *51*(1), 104-110. https://doi.org/10.3349/ymj.2010.51.1.104

Li, Y., Ge, X., Peng, F., Li, W., & Li, J. J. (2022). Exaggerated false positives by popular differential expression methods when analyzing human population samples. *Genome Biology, 23*(1), 1-13. https://doi.org/10.1186/s13059-022-02648-4

Liao, Y., Smyth, G. K., & Shi, W. (2019). The R package *Rsubread* is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Research*, *47*(8), e47. https://doi.org/10.1093/nar/gkz114

Parkinson, N. J., Maslau, S., Ferneyhough, B., Zhang, G., Gregory, L., Buck, D., Ragoussis, J., Ponting, C. P., & Fischer, M. D. (2012). Preparation of high-quality next-generation sequencing libraries from pictogram quantities of target DNA. *Genome research, 22*(1), 125-133. https://doi.org/10.1101/gr.124016.111

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Riemann, K., Adamzik, M., Frauenrath, S., Egensperger, R., Schmid, K. W., Brockmeyer, N. H., & Siffert, W. (2007). Comparison of manual and automated nucleic acid extraction from whole-blood samples. *Journal of Clinical Laboratory Analysis*, *21*(4), 244-248. https://doi.org/10.1002/jcla.20174

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139-140. https://doi.org/10.1093/bioinformatics/btp616

Sherman, B. T., Hao, M., Qiu, J., Jiao, X., Baseler, M. W., Lane, H. C., Imamichi, T., & Chang, W. (2022). DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Research, 50*(W1), W216-W221. https://doi.org/10.1093/nar/gkac194

Shi, H., Zhou, Y., Jia, E., Pan, M., Bai, Y., & Ge, Q. (2021). Bias in RNA-Seq library preparation: current challenges and solutions. *BioMed Research International, 2021,* 6647597. https://doi.org/10.1155/2021/6647597

Wen, J., Legendre, L. A., Bienvenue, J. M., & Landers, J. P. (2008). Purification of nucleic acids in microfluidic devices. *Analytical Chemistry*, *80*(17), 6472-6479. https://doi.org/10.1021/ac8014998

Wickham, H. (2016). Package 'ggplot2': elegant graphics for data analysis. *Springer-Verlag, New York*.